ИНТЕЛЛЕКТУАЛЬНЫЙ МЕГАПОЛИС

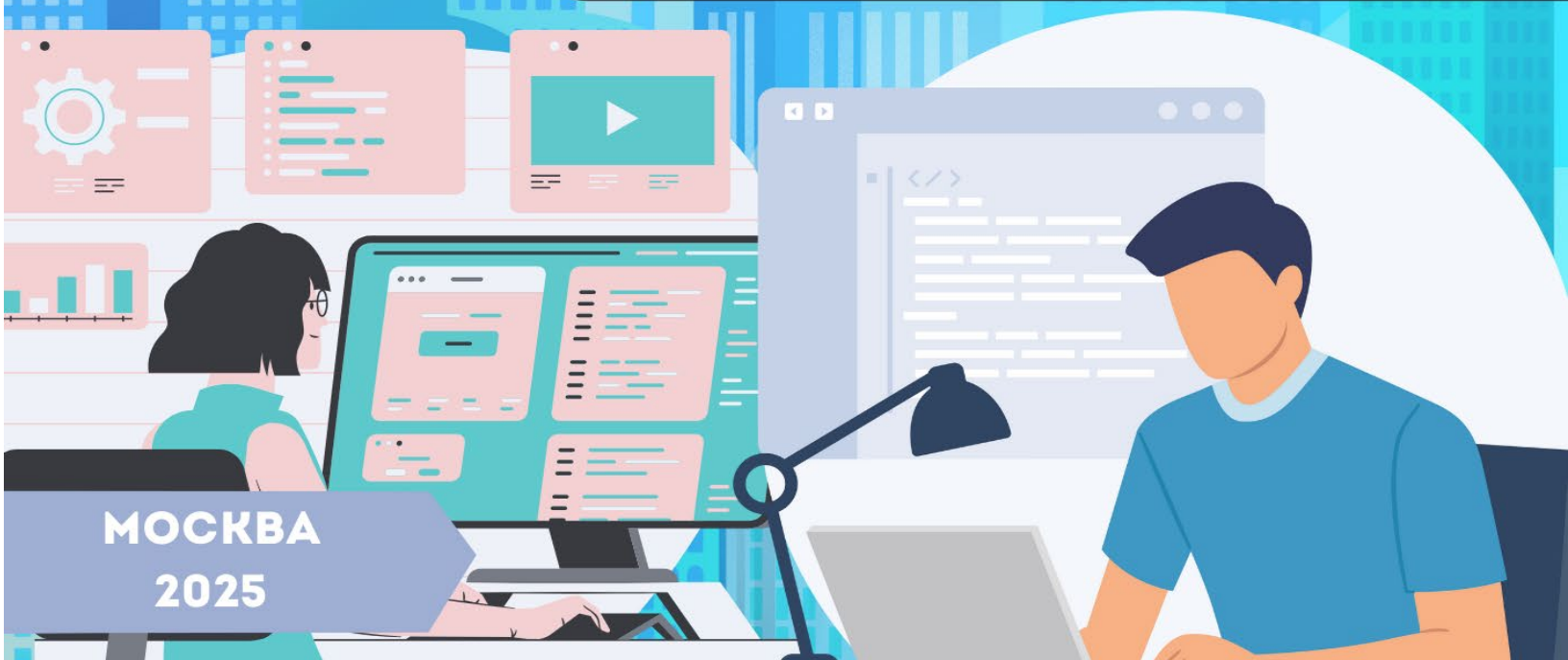
ЗАДАЧНИК

ИТ-класс

в московской школе

НАПРАВЛЕНИЕ БОЛЬШИЕ ДАННЫЕ И ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ПРАКТИЧЕСКИЙ ЭТАП



МОСКВА
2025

ИНТЕЛЛЕКТУАЛЬНЫЙ МЕГАПОЛИС

ЗАДАЧНИК РАЗРАБОТАН:

РТУ МИРЭА

Титоренко Екатерина Александровна
Заместитель начальника организационно-методического отдела

МОСКВА
2025

Оглавление

Критерии оценки задач ***8***

Вариант 1 ..... ***11***

Задача 1. ..... **11**

Задача 2. ..... **12**

Задача 3. ..... **14**

Задача 4 ..... **15**

Вариант 2 ..... ***17***

Задача 1. ..... **17**

Задача 2. ..... **18**

Задача 3. ..... **21**

Задача 4 ..... **22**

Вариант 3 ..... ***23***

Задача 1. ..... **23**

Задача 2. ..... **25**

Задача 3. ..... **27**

Задача 4 ..... **28**

Вариант 4 ..... ***29***

Задача 1. ..... **29**

Задача 2. ..... **30**

Задача 3. ..... **33**

Задача 4 ..... **34**

Вариант 5 ..... ***36***

Задача 1. ..... **36**

Задача 2. ..... **37**

Задача 3. ..... **40**

Задача 4 ..... **41**

Вариант 6 ..... ***43***

Задача 1. ..... **43**

Задача 2. ..... **44**

Задача 3. ..... **47**

Задача 4. ..... **47**

Вариант 7 ..... *50*
Задача 1. 50
Задача 2. 51
Задача 3. 54
Задача 4. 55
Вариант 8 ..... *57*
Задача 1. 57
Задача 2. 58
Задача 3. 60
Задача 4. 61
Вариант 9 ..... *63*
Задача 1. 63
Задача 2. 64
Задача 3. 67
Задача 4. 69
Вариант 10 ..... *71*
Задача 1. 71
Задача 2. 72
Задача 3. 75
Задача 4. 76
Вариант 11 ..... *77*
Задача 1. 77
Задача 2. 78
Задача 3. 82
Задача 4. 82
Вариант 12 ..... *83*
Задача 1. 83
Задача 2. 84
Задача 3. 88
Задача 4. 90
Вариант 13 ..... *91*
Задача 1. 91
Задача 2. 92

Задача 3. ..... **96**
Задача 4. ..... **97**
Вариант 14 ..... ***98***
Задача 1. ..... **98**
Задача 2. ..... **99**
Задача 3 100
Задача 3. ..... **102**
Задача 4. ..... **103**
Вариант 15 ..... ***104***
Задача 1. ..... **104**
Задача 2. ..... **105**
Задача 3. ..... **108**
Задача 4. ..... **109**
Вариант 16 ..... ***110***
Задача 1. ..... **110**
Задача 2. ..... **111**
Задача 3. ..... **115**
Задача 4. ..... **116**
Вариант 17 ..... ***117***
Задача 1. ..... **117**
Задача 2. ..... **118**
Задача 3. ..... **122**
Задача 4. ..... **122**
Вариант 18 ..... ***123***
Задача 1. ..... **123**
Задача 2. ..... **124**
Задача 3. ..... **127**
Задача 4. ..... **128**
Вариант 19 ..... ***129***
Задача 1. ..... **129**
Задача 2. ..... **130**
Задача 3. ..... **134**
Задача 4. ..... **135**

Вариант 20 ..... *136*
Задача 1. 136
Задача 2. 137
Задача 3. 140
Задача 4. 141
Вариант 21 ..... *142*
Задача 1. 142
Задача 2. 143
Задача 3. 146
Задача 4. 147
Вариант 22 ..... *148*
Задача 1. 148
Задача 2. 149
Задача 3. 152
Задача 4. 153
Вариант 23 ..... *154*
Задача 1. 154
Задача 2. 155
Задача 3. 158
Задача 4. 159
Вариант 24 ..... *160*
Задача 1. 160
Задача 2. 161
Задача 3. 165
Задача 4. 166
Вариант 25 ..... *167*
Задача 1. 167
Задача 2. 168
Задача 3. 172
Задача 4. 172
Вариант 26 ..... *174*
Задача 1. 174
Задача 2 175

Задача 3. ..... **178**
Задача 4 ..... **179**
Вариант 27 ..... ***181***
Задача 1. ..... **181**
Задача 2. ..... **183**
Задача 3. ..... **186**
Задача 4. ..... **187**
Вариант 28 ..... ***188***
Задача 1. ..... **188**
Задача 2. ..... **189**
Задача 3. ..... **193**
Задача 4. ..... **194**
Вариант 29 ..... ***196***
Задача 1. ..... **196**
Задача 2. ..... **198**
Задача 3. ..... **200**
Задача 4. ..... **201**
Вариант 30 ..... ***203***
Задача 1. ..... **203**
Задача 2. ..... **205**
Задача 3. ..... **207**
Задача 4. ..... **208**

Критерии оценки задач

Задача 1.

Корректный синтаксис		1
Запрос выполняется без ошибок в PostgreSQL	1	
Есть незначительные ошибки (например, пропущена запятая, неверный алиас), но запрос работает после исправления	0.5	
Синтаксические ошибки, из-за которых запрос не выполняется (например, неправильное использование ключевых слов, некорректные имена таблиц)	0	
Логика		7
Результат полностью соответствует условию (учтены фильтры, сортировка, группировка, исключены дубли)	7	
Частичные ошибки в логике (например, не учтена сортировка по убыванию, пропущен год в фильтре)	3	
Логика нарушена (например, вывод книг, изданных до 2015 года вместо после)	0	
Оптимальность и дополнительные требования		4
Использованы оптимальные методы (например, JOIN вместо подзапросов). Выполнены все дополнительные условия. Нет избыточных операций (например, SELECT * вместо конкретных колонок).	4	
Частично выполнено (например, NOT EXISTS заменен на LEFT JOIN, но результат верный)	2	
Грубые нарушения (например, запрос с CROSS JOIN, ведущим к декартову произведению)	0	
ИТОГО		12

Максимальная оценка за каждый запрос составляет 12 балла. Общее количество баллов за задание – 24.

Задача 2.

Загрузка и предобработка данных		4
Данные загружены корректно Проведена проверка на пропуски (если требуется)	4	
Загрузка выполнена, но не учтены особенности данных (кодировка, разделители)	2	
Данные загружены с ошибками (например, неверный путь к файлу)	1	
Данные не загружены	0	
Расчет показателей		10
Все метрики (динамика, средние, медианы и т.д.) рассчитаны верно Учтены краевые случаи (деление на ноль, отрицательные значения)	10	
Метрики есть, но есть ошибки в формулах или группировках	5	
Расчеты проведены, но не соответствуют условию задачи	1	
Показатели не рассчитаны	0	
Добавление новых столбцов		4
Столбцы созданы через векторные операции (без циклов) Категоризация выполнена по условиям задачи	4	
Логика верна, но код не оптимален (например, использование apply)	2	
Столбцы добавлены, но с ошибками в условиях	1	
Новые столбцы отсутствуют	0	
Визуализация		5
Построен график требуемого типа (heatmap, line, bar и т.д.). Наличие подписей осей, заголовка, легенды. График сохранен в файл	5	
График построен, но: Неверный тип визуализации, Отсутствуют подписи/легенда,	2	
Есть график, но не соответствует данным	1	
Визуализация отсутствует	0	
Оформление кода и вывод		1
Код читаем, есть комментарии (где необходимо). Результаты выведены в требуемом формате (например, print или to_markdown).	1	
Код не структурирован, вывод отсутствует	0	
ИТОГО		24

Максимальная количество баллов за задание – 24

Задача 3.

Логичность и отсутствие ошибок		3
Последовательность рассуждений, пояснение каждого шага	1	
Отсутствие арифметических, логических ошибок или противоречий	2	
Разделение на случаи (например, центральная цифра совпадает с первой/второй)		2
Выявление особенностей объектов, соответствующих условию (например, палиндромы, номера с повторяющимися цифрами)	1	
Правильное применение комбинаторных методов (правила умножения, сочетания, размещения)	1	
Предоставлен верный ответ		1
ИТОГО		6

Максимальная количество баллов за задание – 6.

Задача 4.

Определение типа распределения		1
Правильно указан тип распределения	1	
Тип распределения указан с неуверенностью или допущена несущественная ошибка (например, «похоже на нормальное» без точного названия)	0.5	
Тип распределения указан неверно или не указан вовсе	0	
Указание параметров распределения		2
Все параметры корректны и соответствуют графику	2	
Указан только один параметр правильно, или оба с частичной точностью	1	
Параметры не указаны или указаны с грубыми ошибками	0	
Обоснование выбора типа распределения		2
Обоснование подробное, логичное и основано на анализе формы графика (симметрия, пик, хвосты, размах и т.д.)	2	
Обоснование частично верное или не до конца раскрыто (общие фразы, без привязки к графику)	1	
Обоснование отсутствует или ошибочно	0	
ИТОГО		5

Максимальная оценка за каждый график составляет 5 баллов. Общее количество баллов за задание – 10.

Вариант 1

Задача 1.

Образовательная онлайн-платформа анализирует активность пользователей и популярность курсов для улучшения качества обучения и обновления контента. Вам, как аналитику, поручили подготовить выборки из базы PostgreSQL.

Структура базы данных:

- instructors (id, name, expertise) — преподаватели.
- courses (id, title, year, instructor_id, duration_hours) — курсы.
- students (id, name, email, registration_date) — студенты.
- course_enrollments (id, course_id, student_id, enrollment_date, completion_date) — записи о прохождении курсов.

Подготовьте SQL-запросы к базе данных для следующих целей:

1. Преподаватели с короткими курсами

Найти имена преподавателей, у которых все курсы длятся меньше 10 часов

2. Незавершённые курсы

Найти имена студентов, которые записались на курс, но не завершили его

и названия этих курсов

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT i.name FROM instructors i JOIN courses c ON i.id = c.instructor_id GROUP BY i.id, i.name HAVING MAX(c.duration_hours) < 10;
2	SELECT s.name, c.title FROM students s JOIN course_enrollments ce ON s.id = ce.student_id JOIN courses c ON ce.course_id = c.id WHERE ce.completion_date IS NULL;

Задача 2.

Вы работаете аналитиком в магазине электроники и вам предстоит проанализировать продажи холодильников у конкурентов.

У вас есть файл fridge_prices.csv, содержащий информацию о магазинах, холодильниках и их ценах за определенный период времени.

Описание данных:

Поле	Описание
store	Магазин (Эльдорадо, М.Видео, Ситилинк, DNS)
brand	Бренд (Bosch, LG, Samsung, Indesit, Haier)
model	Модель холодильника
capacity	Объем (литры, целое число 200-600)
price	Цена (20000-150000)
month	Месяц (1-12)
year	Год (2022, 2023)

Требуется

1. Загрузить данные из файла fridge_prices.csv - файл должен находиться в том же проекте, что и программа.

2. Создайте столбец price_category:

"Бюджетные" (< 50000 руб)

"Средние" (50000-100000 руб)

"Премиум" (> 100000 руб) Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

store
brand
model
capacity
price
month
year
price_category

Проверьте, что порядок и формат соответствуют.

3. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке убывания по полю Price.

Используйте sort_values()

4. Постройте гистограмму (hist) распределения цен:

Ось X — Цена

Ось Y — Количество моделей

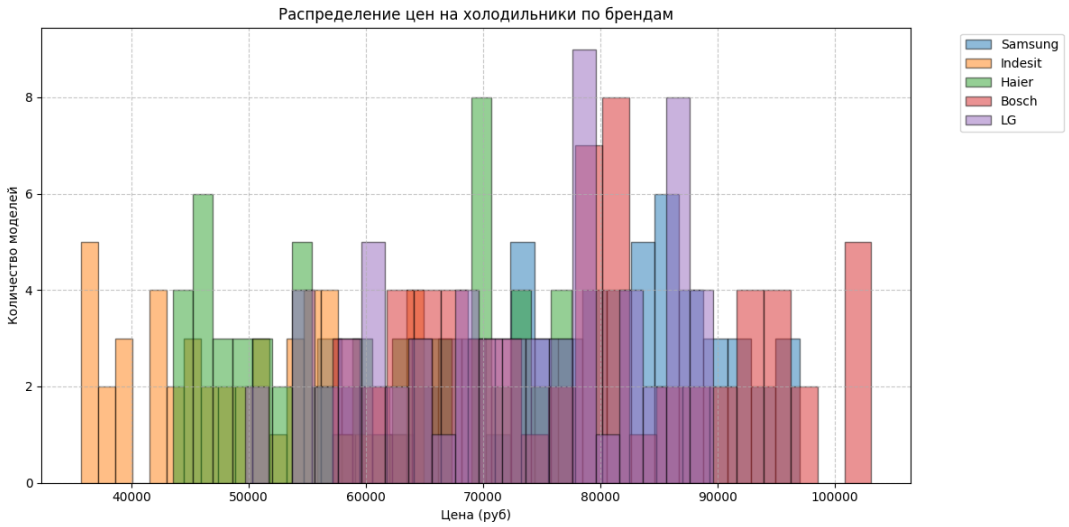
Разделите по брендам (разные цвета). Подпишите оси и добавьте заголовок, выведите легенду.

Используйте matplotlib.pyplot.hist(), plt.xlabel(), plt.ylabel(), plt.title().

Сохранить изображение как fridge_prices_hist.png

Решение

```
import pandas as pd
import matplotlib.pyplot as plt_

df = pd.read_csv("fridge_prices.csv")

def get_price_category(price):
    if price < 50000:
        return "Бюджетные"
    elif 50000 <= price <= 100000:
        return "Средние"
    else:
        return "Премиум"

df["price_category"] = df["price"].apply(get_price_category)

top_expensive = df.sort_values("price", ascending=False).head(10)
print(top_expensive.to_string(index=False))

plt.figure(figsize=(12, 6))

for brand in df["brand"].unique():
    brand_data = df[df["brand"] == brand]["price"]
    plt.hist(brand_data, bins=20, alpha=0.5, label=brand, edgecolor="black")

plt.title("Распределение цен на холодильники по брендам")
plt.xlabel("Цена (руб)")
plt.ylabel("Количество моделей")
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.grid(True, linestyle="--", alpha=0.7)

plt.tight_layout()
plt.savefig("fridge_prices_hist.png", bbox_inches="tight")
plt.show()
```

Вывод 10 строк	store brand model capacity price month year price_category М.Видео Bosch Bosch-3414 442 103116 2 2023 Премиум Эльдорадо Bosch Bosch-4695 536 102105 8 2022 Премиум Ситилинк Bosch Bosch-2366 534 101931 3 2023 Премиум DNS Bosch Bosch-6058 256 101239 2 2023 Премиум Эльдорадо Bosch Bosch-8699 393 101155 10 2023 Премиум DNS Samsung Samsung-6495 506 96997 12 2022 Средние М.Видео Samsung Samsung-7399 333 96665 12 2023 Средние Ситилинк Bosch Bosch-5398 215 96626 2 2023 Средние DNS Bosch Bosch-6907 357 96512 12 2023 Средние М.Видео Bosch Bosch-1228 450 95965 1 2023 Средние
График	

Задача 3.

В зоопарке 60% животных — млекопитающие, 30% — птицы и 10% — рептилии. Среди них:

70% млекопитающих — хищники,

50% птиц — хищные,

40% рептилий — ядовитые (считаются хищниками).

Если выбранное животное — не хищник, то с вероятностью 20% оно всё равно может быть опасно (например, из-за агрессивного поведения).

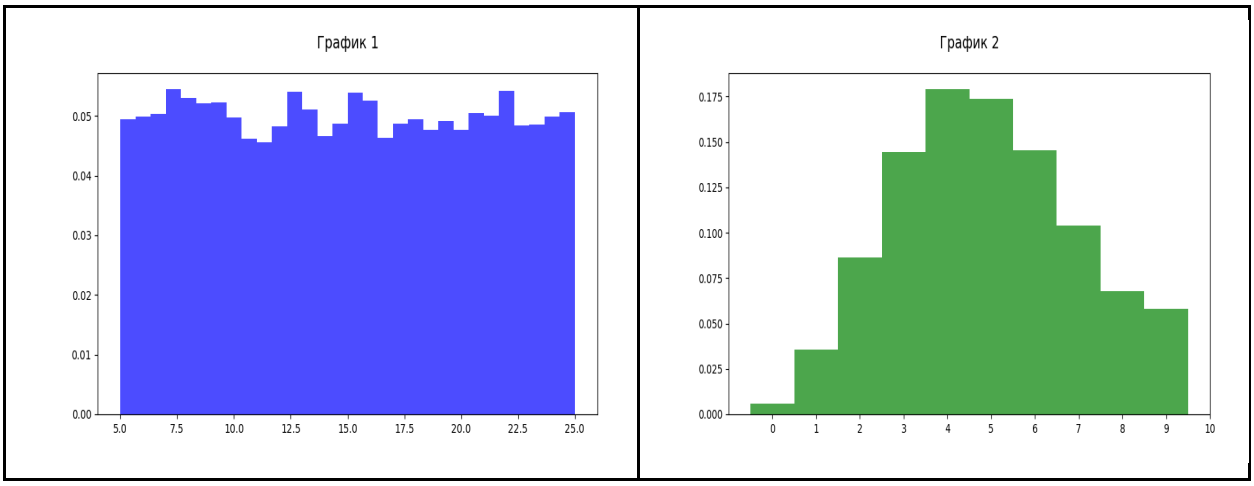
Какова вероятность, что случайно выбранное животное опасно (либо является хищником, либо не хищником, но агрессивным)?

Ответ округлите до сотых и переведите в проценты.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 69%

Задача 4



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределений:
   - Для Графика 1: границы интервала (a и b).
   - Для Графика 2: параметр λ (лямбда).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Решение:

	График 1. Равномерное распределение График 2. Распределение Пуассона

	График 1: a ≈ 5, b ≈ 25 (значения равномерно распределены от 5 до 25). График 2: λ ≈ 5 (пик около 5, среднее = λ)
3.	*График 1:* гистограмма плоская, все значения в диапазоне [5, 25] встречаются примерно одинаково часто. *График 2:* дискретные значения, пик около 5, асимметрия вправо (характерно для Пуассона).

Допустимые погрешности при учете ответа:

a и b: ±2–3% от диапазона (например, данные от 5 до 25 → a ≈ 5 ± 0.5, b ≈ 25 ± 0.5).

λ: ±1–2 единицы (например, λ = 5 → ответ 4–6)

Вариант 2

Задача 1.

Образовательная онлайн-платформа анализирует активность пользователей и популярность курсов для улучшения качества обучения и обновления контента. Вам, как аналитику, поручили подготовить выборки из базы PostgreSQL.

Структура базы данных:

- instructors (id, name, expertise) — преподаватели.
- courses (id, title, year, instructor_id, duration_hours) — курсы.
- students (id, name, email, registration_date) — студенты.
- course_enrollments (id, course_id, student_id, enrollment_date, completion_date) — записи о прохождении курсов.

Подготовьте SQL-запросы к базе данных для следующих целей:

1. Преподаватели с самыми популярными курсами

Вывести имена топ-3 преподавателей, чьи курсы имеют наибольшее количество записей на их курсы и количество записей на их курсы.

2. Анализ успеваемости студентов

Найти имена студентов, которые завершили более 2 курсов, и количество пройденных курсов, а также среднее время прохождения в днях.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

1	SELECT
	  i.name,
	  COUNT(ce.id) AS total_enrollments
	FROM
	  instructors i
	JOIN
	  courses c ON i.id = c.instructor_id
	JOIN
	  course_enrollments ce ON c.id = ce.course_id
	GROUP BY
	  i.id, i.name
	ORDER BY
	  total_enrollments DESC
	LIMIT 3;

2	SELECT
	  s.name,
	  COUNT(ce.id) AS completed_courses,
	  AVG(ce.completion_date - ce.enrollment_date) AS avg_completion_days
	FROM
	  students s
	JOIN
	  course_enrollments ce ON s.id = ce.student_id
	WHERE
	  ce.completion_date IS NOT NULL
	GROUP BY
	  s.id, s.name
	HAVING
	  COUNT(ce.id) > 2
	ORDER BY
	  avg_completion_days;

Задача 2.

Вы аналитик данных в книжном магазине. Вам нужно проанализировать динамку изменения цен у конкурентов.

Файл book_sales.csv содержит информацию о продажах книг в различных интернет-магазинах.

Описание данных:

Поле	Описание

store	Название магазина (строка)
category	Жанр книги (строка)
price	Цена книги (дробное число)
sales	Количество проданных книг (целое число)
date	Дата продажи (в формате YYYY-MM-DD)

Требуется

1. Загрузить данные из файла book_sales.csv - файл должен находиться в том же проекте, что и программа.
2. Преобразуйте дату из формата YYYY-MM-DD в YYYY (новый столбец year). Создайте столбец median_price с медианной ценой для каждого магазина по годам.
3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

store
category
price
sales
date
year
median_price

Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке убывания по полю sales.

   Используйте sort_values()
5. Постройте график изменения медианной цены по магазинам

   Ось X — Год

   Ось Y — Медианная цена

   Укажите подписи для осей и название график.

   Используйте plt.plot(), plt.xlabel(), plt.ylabel(), plt.title().
6. Сохранить изображение как book_sales_plot.png

Запишите свое решение в файл: task_2.py.

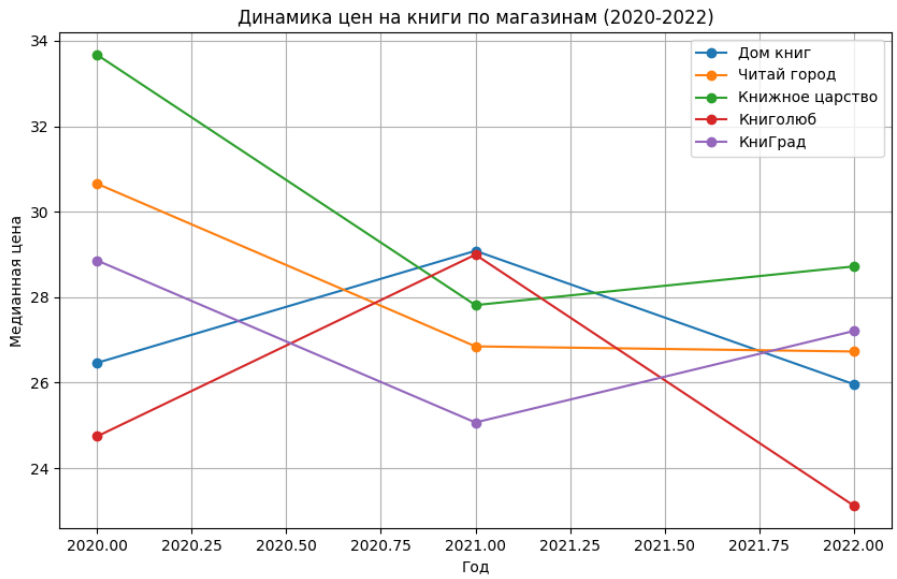
Документацию по приведенным методам можно получить с помощью функции help().

Решение

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("book_sales.csv")

df["year"] = pd.to_datetime(df["date"]).dt.year
df["median_price"] = df.groupby(["store", "year"])["price"].transform("median")

result_df = df[["store", "category", "price", "sales", "date", "year",
"median_price"]]

print(result_df.sort_values("sales", ascending=False).head(10))

plt.figure(figsize=(10, 6))
for store in df["store"].unique():
    store_data = df[df["store"] == store]
    yearly_median = store_data.groupby("year")["price"].median()
    plt.plot(yearly_median.index, yearly_median.values, label=store, marker="o")

plt.xlabel("Год")
plt.ylabel("Медианная цена")
plt.title("Динамика цен на книги по магазинам (2020-2022)")
plt.legend()
plt.grid()
plt.savefig("book_sales_plot.png")
plt.show()
```

Вывод 10 строк

	store	category	price	...	date	year	median_price
122	Читай город	Биография	48.46	...	2021-07-24	2021	26.850
415	Читай город	Утопия	11.09	...	2022-01-07	2022	26.730
755	Книжное царство	Роман	35.90	...	2022-02-17	2022	28.720
636	КниГрад	Роман	37.70	...	2022-01-30	2022	27.210
118	КниГрад	Биография	6.16	...	2020-11-09	2020	28.860
477	КниГрад	Биография	35.11	...	2022-07-14	2022	27.210
335	Книголюб	Научная фантастика	21.57	...	2022-02-21	2022	23.125
451	Дом книг	Научная фантастика	34.44	...	2020-11-10	2020	26.465
157	Читай город	Научная фантастика	44.90	...	2021-01-18	2021	26.850
733	Книголюб	Биография	44.34	...	2021-09-11	2021	28.995

График	

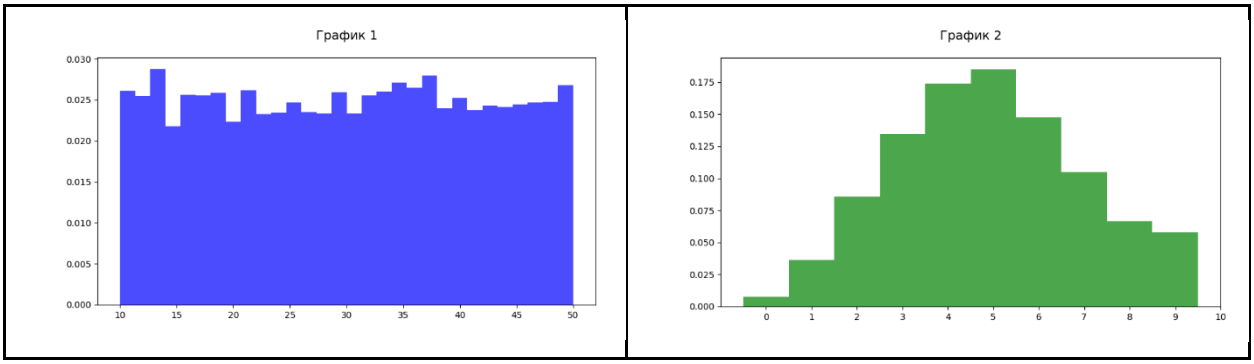
Задача 3.

В университете 90% студентов — бакалавры, а 10% — магистры. 50% бакалавров получают стипендию, а 20% магистров — нет. Какова вероятность, что случайно выбранный студент получает стипендию?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 53%

Задача 4



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределений:
   - Для Графика 1: границы интервала (a и b).
   - Для Графика 2: параметр λ (лямбда).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Решение:

	График 1 — равномерное, График 2 — Пуассона
	График 1: a ≈ 10, b ≈ 50 (значения равномерно распределены от 10 до 50). График 2: λ ≈ 5 (пик около 5, среднее = λ)
3.	График 1: гистограмма плоская, все значения в диапазоне [10, 50] встречаются примерно одинаково часто. График 2: дискретные значения, пик около 5, асимметрия вправо (характерно для Пуассона).

Допустимые погрешности при учете ответа:
a и b: ±2–3% от диапазона (например, данные от 5 до 25 → a ≈ 5 ± 0.5, b ≈ 25 ± 0.5).
λ: ±1–2 единицы (например, λ = 5 → ответ 4–6)

Вариант 3

Задача 1.

Образовательная онлайн-платформа анализирует активность пользователей и популярность курсов для улучшения качества обучения и обновления контента. Вам, как аналитику, поручили подготовить выборки из базы PostgreSQL.

Структура базы данных:

- instructors (id, name, expertise) — преподаватели.
- courses (id, title, year, instructor_id, duration_hours) — курсы.
- students (id, name, email, registration_date) — студенты.
- course_enrollments (id, course_id, student_id, enrollment_date, completion_date) — записи о прохождении курсов.

Подготовьте SQL-запросы к базе данных для следующих целей:

1. Курсы с самым высоким процентом завершения

Вывести названия курсов, на которые сделано более 10 записей, количество прохождений этих курсов, количество записей на них, частоту их успешного завершения (отношение завершенных к общему числу записей), отсортировать по убыванию частоты завершения.

2. Сезонная активность студентов

Определить, в каком месяце чаще всего записываются на курсы.

Вывести месяцы, количество записей в каждый месяц, отсортировать по убыванию количества записей.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT
	  c.title,
	  COUNT(CASE WHEN ce.completion_date IS NOT NULL THEN 1 END) AS completed,
	  COUNT(ce.id) AS total_enrollments,
	  ROUND(COUNT(CASE WHEN ce.completion_date IS NOT NULL THEN 1 END) * 100.0 / COUNT(ce.id), 2) AS completion_rate
	FROM
	  courses c
	JOIN
	  course_enrollments ce ON c.id = ce.course_id
	GROUP BY
	  c.id, c.title
	HAVING
	  COUNT(ce.id) > 10
	ORDER BY
	  completion_rate DESC;

2	SELECT
	  EXTRACT(MONTH FROM enrollment_date) AS month,
	  COUNT(*) AS enrollments
	FROM
	  course_enrollments
	GROUP BY
	  month
	ORDER BY
	  enrollments DESC;

Задача 2.

Вы работаете аналитиком в учебном отделе образовательного учреждении, вам нужно определить как время, уделенное обучению, влияет на успеваемость студентов.

У вас есть файл marks_data.csv, содержащий информацию о количестве курсов, оценках и времени затраченным на учебу.

Описание данных:

Поле	Описание
number_courses	Количество курсов, выбранных студентом
time_study	Среднее время, затрачиваемое студентом на учебу в день
Marks	Оценки, полученные студентом

Требуется

1. Загрузить данные из файла marks_data.csv - файл должен находиться в том же проекте, что и программа.
2. Создать новый столбец xcof – коэффициент времени проведенным за обучением.
   Считать по формуле: number_course / time_study
   Этот столбец поможет определить какое количество времени студент тратит на курс.
3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

xcof
number_courses
time_study
Marks

Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке возрастания по полю Marks.

Используйте sort_values()

5. Для оценки успеваемости студентов постройте график распределения оценок.

Ось X — коэффициент времени

Ось Y — успеваемость

Укажите подписи для осей и название график.

Используйте matplotlib.pyplot.hist(), plt.xlabel(), plt.ylabel(), plt.title().

6. Сохранить изображение как marks.png

Запишите свое решение в файл: task_2.py.

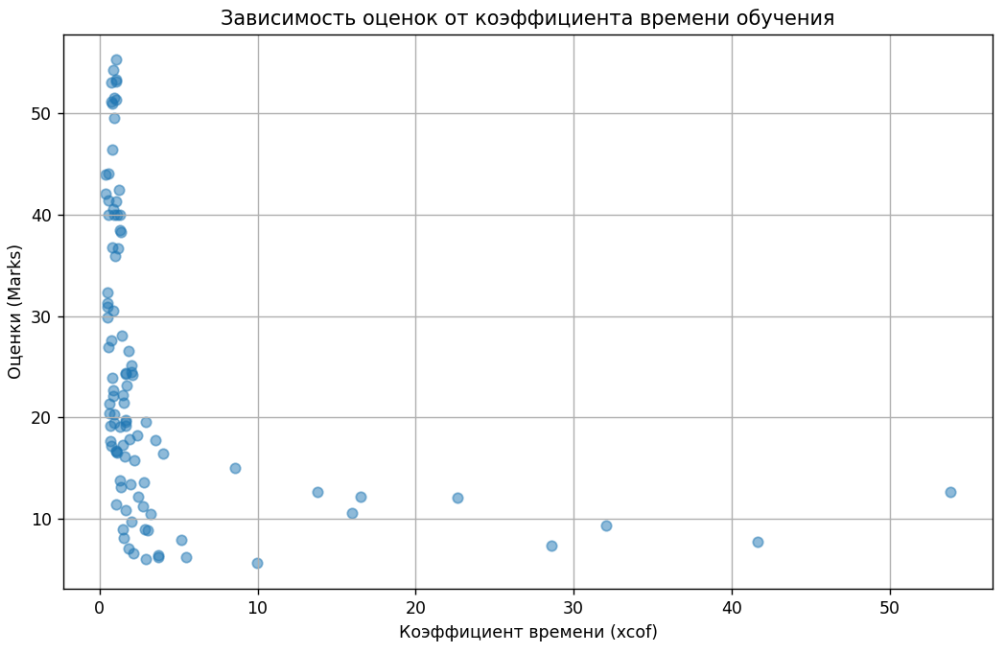
Документацию по приведенным методам можно получить с помощью функции ***help()***.

Решение

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('marks_data.csv')

df['xcof'] = df['number_courses'] / df['time_study']

df = df[['xcof', 'number_courses', 'time_study', 'Marks']]

df_sorted = df.sort_values('Marks')
print(df_sorted.head(10))

plt.figure(figsize=(10, 6))
plt.scatter(df['xcof'], df['Marks'], alpha=0.5)
plt.xlabel('Коэффициент времени (xcof)')
plt.ylabel('Оценки (Marks)')
plt.title('Зависимость оценок от коэффициента времени обучения')
plt.grid(True)
plt.savefig('marks.png')
plt.show()
```

Вывод
10 строк

	xcof	number_courses	time_study	Marks
96	9.966777	3	0.301	5.609
87	2.904163	3	1.033	6.053
67	5.454545	3	0.550	6.185
75	3.735990	3	0.803	6.217
79	3.726708	3	0.805	6.349
25	2.132196	3	1.407	6.623
73	1.841621	3	1.629	7.014
40	28.571429	4	0.140	7.336
1	41.666667	4	0.096	7.734
52	5.188067	4	0.771	7.892

График	

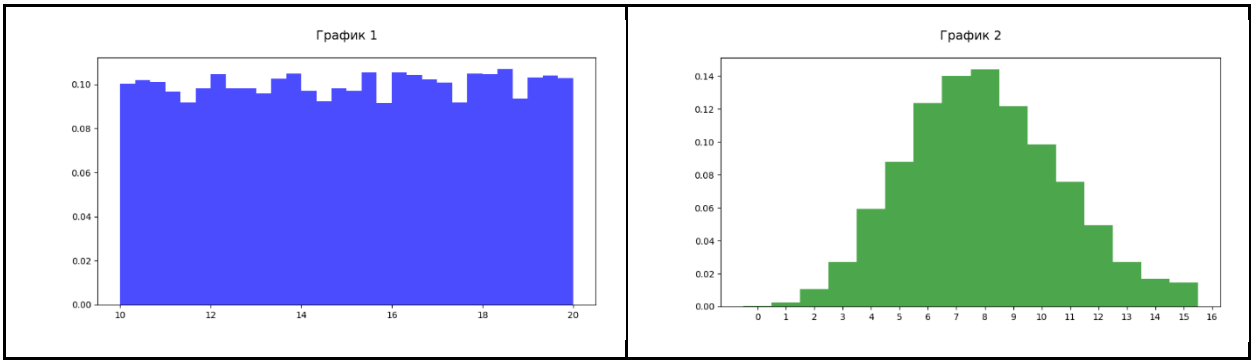
Задача 3.

В классе 15 учеников, среди которых 7 девочек и 8 мальчиков. Сколькими способами можно выбрать 3 дежурных так, чтобы среди выбранных была хотя бы одна девочка.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 399

Задача 4



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределений:
   - Для Графика 1: границы интервала (a и b).
   - Для Графика 2: параметр λ (лямбда).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

	График 1 — равномерное, График 2 — Пуассона
	График 1: a ≈ 10, b ≈ 20 (значения равномерно распределены от 10 до 20). График 2: λ ≈ 8 (пик около 8, среднее = λ)
3.	График 1: гистограмма плоская, все значения в диапазоне [10, 20] встречаются примерно одинаково часто. График 2: дискретные значения, пик около 8, асимметрия вправо (характерно для Пуассона).

Допустимые погрешности при учете ответа:
a и b: ±2–3% от диапазона (например, данные от 5 до 25 → a ≈ 5 ± 0.5, b ≈

25 ± 0.5).

λ: ±1–2 единицы (например, λ = 5 → ответ 4–6)

Вариант 4

Задача 1.

Задача 1.

Образовательная онлайн-платформа анализирует активность пользователей и популярность курсов для улучшения качества обучения и обновления контента. Вам, как аналитику, поручили подготовить выборки из базы PostgreSQL.

Структура базы данных:

- instructors (id, name, expertise) — преподаватели.
- courses (id, title, year, instructor_id, duration_hours) — курсы.
- students (id, name, email, registration_date) — студенты.
- course_enrollments (id, course_id, student_id, enrollment_date, completion_date) — записи о прохождении курсов.

Подготовьте SQL-запросы к базе данных для следующих целей:

1. Курсы с самым долгим временем прохождения

Найти 5 курсов, на прохождение которых у студентов ушло больше всего времени, выведите их названия, среднее время прохождения.

2. Преподаватели с самым разносторонним опытом

Найти преподавателей, которые ведут курсы в 3-х и более разных областях (поле expertise) выведите их имена, количество областей.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT
	  c.title,
	  AVG(ce.completion_date - ce.enrollment_date) AS avg_days_to_complete
	FROM
	  courses c
	JOIN
	  course_enrollments ce ON c.id = ce.course_id
	WHERE
	  ce.completion_date IS NOT NULL
	GROUP BY
	  c.id, c.title
	ORDER BY
	  avg_days_to_complete DESC
	LIMIT 5;

2	SELECT
	  i.name,
	  COUNT(DISTINCT i.expertise) AS unique_expertise_count
	FROM
	  instructors i
	JOIN
	  courses c ON i.id = c.instructor_id
	GROUP BY
	  i.id, i.name
	HAVING
	  COUNT(DISTINCT i.expertise) >= 3;

Задача 2.

Вы работаете аналитиком в автосалоне. Вам нужно посмотреть, какие машины в вашем салоне имеют самую высокую надежность.

У вас есть файл cars_data.csv, в котором хранится вся нужная информация,

Описание данных:

Поле	Описание
model	Марка машины
mileage	Пробег
year	Год выпуска
wear	Степень износа

Требуется

1. Загрузить данные из файла cars_data.csv - файл должен находиться в том же проекте, что и программа.
2. Создать новый столбец safety – коэффициент пригодности транспортного средства для дальнейшей эксплуатации. Считать по формуле: mileage * wear
3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

model
mileage
year
wear
safety

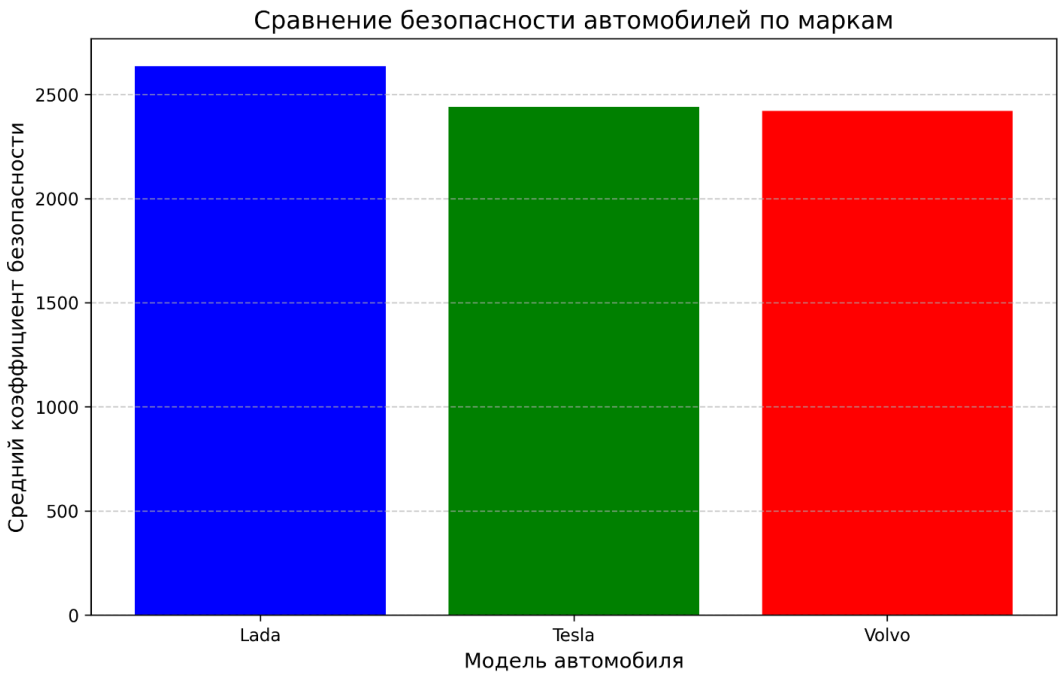
Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке не убывания по полю safety.
   Используйте sort_values()
5. Для оценки самой надежной модели постройте гистограмму безопасности от модели машины
   Ось X — модель
   Ось Y — безопасность
   Укажите подписи для осей и название график.
   Используйте matplotlib.pyplot.hist(), plt.xlabel(), plt.ylabel(), plt.title().
6. Сохранить изображение как cars_plot.png

Запишите свое решение в файл: task_2.py.

*Документацию по приведенным методам можно получить с помощью функции **help()**.*

Решение	import pandas as pd import matplotlib.pyplot as plt df = pd.read_csv('cars_data.csv') df['safety'] = df['mileage'] * df['wear'] df = df[['model', 'mileage', 'year', 'wear', 'safety']] sorted_df = df.sort_values('safety') print(sorted_df.head(10)) plt.figure(figsize=(10, 6)) safety_by_model = df.groupby('model')['safety'].mean() plt.bar(safety_by_model.index, safety_by_model.values, color=['blue', 'green', 'red']) plt.xlabel('Модель автомобиля', fontsize=12) plt.ylabel('Средний коэффициент безопасности', fontsize=12) plt.title('Сравнение безопасности автомобилей по маркам', fontsize=14) plt.grid(axis='y', linestyle='--', alpha=0.7) plt.savefig('cars_plot.png', dpi=300, bbox_inches='tight') plt.show()

Вывод 10 строк	model mileage year wear safety 688 Tesla 86 2006 0.066 5.676 616 Lada 32 2023 0.185 5.920 837 Tesla 522 2011 0.013 6.786 828 Volvo 113 1992 0.066 7.458 297 Lada 1269 2003 0.007 8.883 316 Lada 19 2016 0.472 8.968 954 Tesla 9028 2017 0.001 9.028 947 Lada 9703 2001 0.001 9.703 385 Volvo 303 1994 0.033 9.999 69 Tesla 18 2000 0.588 10.584
График	

Задача 3.

В розыгрыше участвуют билеты с 6-значными номерами (от 000000 до 999999). Есть возможность выиграть как главный приз, так и дополнительный. Условия для их получения:

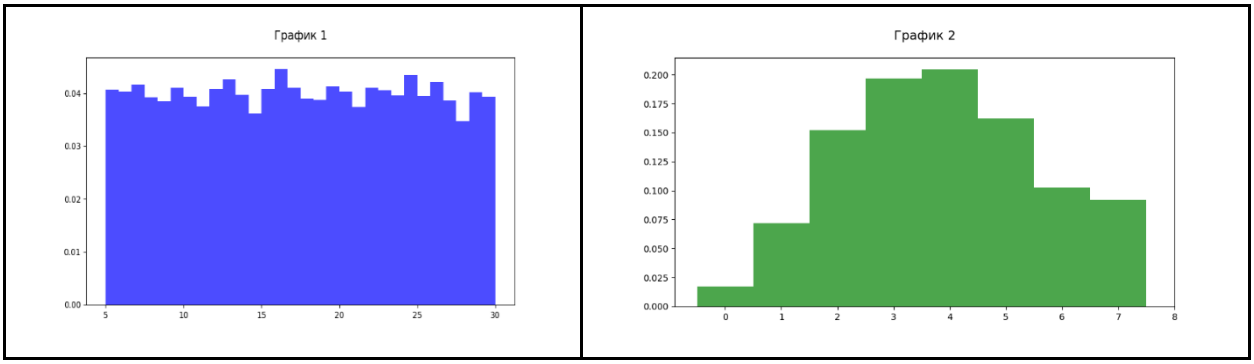
1. Главный приз: номер содержит ровно четыре одинаковые цифры (например, 111100, 223332).
2. Дополнительный приз: номер — палиндром (например, 123321, 456654).

Какова вероятность, что случайно выбранный номер одновременно претендует на главный приз и является палиндромом?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.00017

Задача 4



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределений:
   - Для Графика 1: границы интервала (a и b).
   - Для Графика 2: параметр λ (лямбда).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

	График 1 — равномерное, График 2 — Пуассона
	График 1: a ≈ 5, b ≈ 30 (значения равномерно распределены от 5 до 30). График 2: λ ≈ 4 (пик около 4, среднее = λ)

3.	График 1: гистограмма плоская, все значения в диапазоне [5, 30] встречаются примерно одинаково часто. График 2: дискретные значения, пик около 4, асимметрия вправо (характерно для Пуассона).

Допустимые погрешности при учете ответа:
a и b: ±2–3% от диапазона (например, данные от 5 до 25 → a ≈ 5 ± 0.5, b ≈

25 ± 0.5).

λ: ±1–2 единицы (например, λ = 5 → ответ 4–6)

Вариант 5

Задача 1.

Образовательная онлайн-платформа анализирует активность пользователей и популярность курсов для улучшения качества обучения и обновления контента. Вам, как аналитику, поручили подготовить выборки из базы PostgreSQL.

Структура базы данных:

- instructors (id, name, expertise) — преподаватели.
- courses (id, title, year, instructor_id, duration_hours) — курсы.
- students (id, name, email, registration_date) — студенты.
- course_enrollments (id, course_id, student_id, enrollment_date, completion_date) — записи о прохождении курсов.

Подготовьте SQL-запросы к базе данных для следующих целей:

1. Анализ оттока студентов

Найти студентов, которые записались на курс, но не завершили ни одного, выведите их имена, количество записей.

2. Анализ повторной вовлеченности студентов

Найти студентов, которые записывались на новые курсы после успешного завершения предыдущих, выведите их имена, электронные почты, количество взятых курсов.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT
	  s.name,
	  COUNT(ce.id) AS enrollments,
	  COUNT(CASE WHEN ce.completion_date IS NOT NULL THEN 1 END) AS completed
	FROM
	  students s
	JOIN
	  course_enrollments ce ON s.id = ce.student_id
	GROUP BY
	  s.id, s.name
	HAVING
	  COUNT(ce.id) > 0 AND COUNT(CASE WHEN ce.completion_date IS NOT NULL THEN 1 END) = 0;

2	SELECT
	  s.name,
	  s.email,
	  COUNT(DISTINCT ce2.course_id) AS additional_courses_taken
	FROM
	  students s
	JOIN
	  course_enrollments ce1 ON s.id = ce1.student_id
	JOIN
	  course_enrollments ce2 ON s.id = ce2.student_id
	WHERE
	  ce1.completion_date IS NOT NULL
	  AND ce2.enrollment_date > ce1.completion_date
	GROUP BY
	  s.id, s.name, s.email
	HAVING
	  COUNT(DISTINCT ce2.course_id) > 1
	ORDER BY
	  additional_courses_taken DESC;

Задача 2.

Вы работаете аналитиком в отделе продаж электроники. Вам нужно определить какие смартфоны имеют лучшую оценку цена/качество.

У вас есть файл phones_data.csv, в котором хранится вся нужная информация,

Описание данных:

Поле	Описание
model	Марка телефона (строка)
sim	Количество слотов для сим карт (целое число от 1 до 2)
display	Диагональ дисплея (дробное число от 4.4 до 7.7)
battery	Емкость аккумулятора (число от 2000 до 5000)
price	Цена телефона (число от 30000 до 100000)

Требуется

1. Загрузить данные из файла phones_data.csv - файл должен находиться в том же проекте, что и программа.
2. Создать новый столбец apply – коэффициент удачности соотношения цены и качества. Считать по формуле: (display + battery + sim) / price
3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

model
sim
display
battery
price
apply

Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке убывания по полю apply.

Используйте sort_values()

5. Для оценки самой подходящей модели постройте гистограмму отношения модели и коэффициента качества

Ось X — Модель телефона

Ось Y — Коэффициент цена качество

Укажите подписи для осей и название график.

Используйте matplotlib.pyplot.hist(), plt.xlabel(), plt.ylabel(), plt.title().

6. Сохранить изображение как phone_plot.png

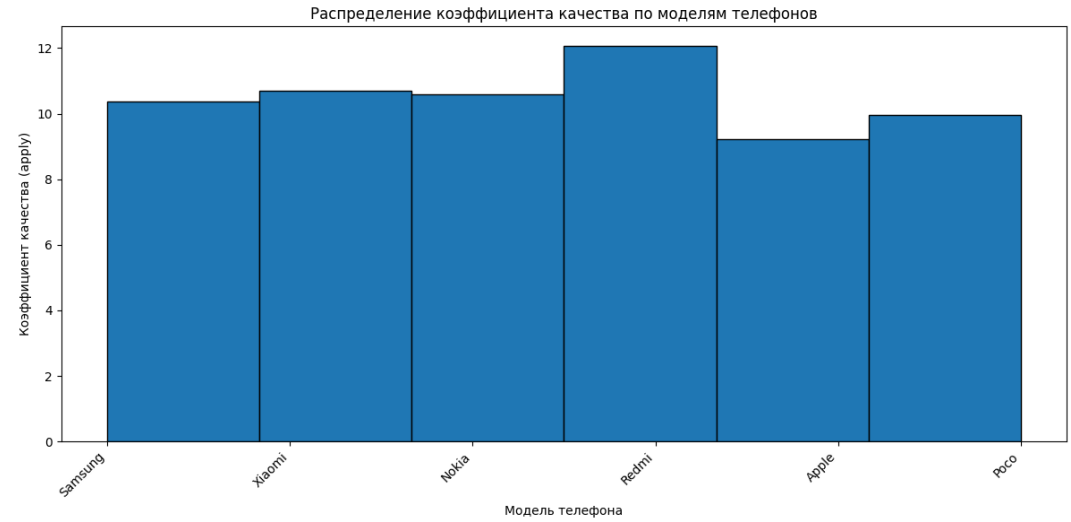
Запишите свое решение в файл: task_2.py.

Документацию по приведенным методам можно получить с помощью функции help().

Решение

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('phones_data.csv')

df['apply'] = (df['display'] + df['battery'] + df['sim']) / df['price']
```

	df = df[['model', 'sim', 'display', 'battery', 'price', 'apply']] sorted_df = df.sort_values('apply', ascending=False) print(sorted_df.head(10)) plt.figure(figsize=(12, 6)) plt.hist(df['model'], weights=df['apply'], bins=len(df['model'].unique()), edgecolor='black') plt.xlabel('Модель телефона') plt.ylabel('Коэффициент качества (apply)') plt.title('Распределение коэффициента качества по моделям телефонов') plt.xticks(rotation=45, ha='right') plt.tight_layout() plt.savefig('phone_plot.png') plt.show()
Вывод 10 строк	model sim display battery price apply 207 Poco 1 7.2 4961 30074 0.165232 397 Apple 2 6.5 4870 31756 0.153625 155 Redmi 2 7.1 4739 31163 0.152363 908 Xiaomi 2 6.7 4812 32276 0.149359 28 Samsung 2 7.6 4530 30797 0.147404 514 Apple 2 4.7 4478 30667 0.146239 111 Apple 1 6.4 4563 31716 0.144104 737 Redmi 1 7.6 4825 33629 0.143733 910 Apple 2 6.5 4939 34586 0.143049 602 Nokia 2 6.0 4838 33972 0.142647

График	

Задача 3.

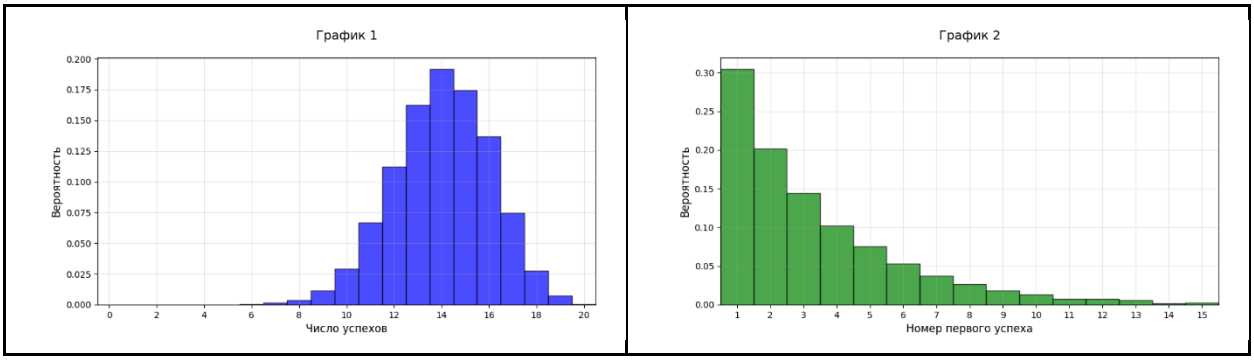
Вы составляете буднее меню (за исключением пятницы) для кафе. У вас есть 10 десертов. Нужно выбрать 4 для акции «Десерт дня» с условиями:

- В среду — обязательно клубничный чизкейк
- Ровно один десерт должен быть шоколадным (в списке таких 5)
- Десерты не повторяются

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 180

Задача 4



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:

- Для Графика 1: вероятность успеха (p).
- Для Графика 2: укажите вероятность успеха (p).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt

Решение:

	График 1. Биномиальное распределение График 2. Геометрическое распределение
	График 1. $p \approx 0.7$ График 2. $p \approx 0.3$
3.	*График 1:* дискретные значения (целые числа от 0 до 20), пик около 14, характерная форма колокола — соответствует биномиальному распределению. *График 2:* убывающая ступенчатая форма с максимумом на 1, значения — только натуральные числа. Это соответствует геометрическому распределению, описывающему номер первого успеха.

Допустимые погрешности при учете ответа:

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)

σ: ±10–15% (например, σ = 10 → ответ 9–11)

n: точное значение (например, n = 20 → только 20, так как это целое число).

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)

Вариант 6

Задача 1.

Образовательная онлайн-платформа анализирует активность пользователей и популярность курсов для улучшения качества обучения и обновления контента. Вам, как аналитику, поручили подготовить выборки из базы PostgreSQL.

Структура базы данных:

- instructors (id, name, expertise) — преподаватели.
- courses (id, title, year, instructor_id, duration_hours) — курсы.
- students (id, name, email, registration_date) — студенты.
- course_enrollments (id, course_id, student_id, enrollment_date, completion_date) — записи о прохождении курсов.

Подготовьте SQL-запросы к базе данных для следующих целей:

1. Выявление самых сложных курсов

Определить 5 курсов с наибольшим разрывом между средней продолжительностью курса и фактическим временем прохождения. Выведите их названия, продолжительность, среднее время прохождения, разрыв.

2. Анализ сезонности создания курсов

Выявить, в какие месяцы преподаватели чаще всего создают новые курсы. Выведите месяца, отсортированные по убыванию количества созданных курсов.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT
	  c.title,
	  c.duration_hours AS expected_hours,
	  AVG(EXTRACT(EPOCH FROM (ce.completion_date - ce.enrollment_date))/3600)::numeric(10,2) AS actual_hours,
	  (AVG(EXTRACT(EPOCH FROM (ce.completion_date - ce.enrollment_date))/3600) - c.duration_hours)::numeric(10,2) AS difficulty_gap
	FROM
	  courses c
	JOIN
	  course_enrollments ce ON c.id = ce.course_id
	WHERE
	  ce.completion_date IS NOT NULL
	GROUP BY
	  c.id, c.title, c.duration_hours
	ORDER BY
	  difficulty_gap DESC
	LIMIT 5;

2	SELECT
	  EXTRACT(MONTH FROM make_date(c.year, 1, 1) + (c.id % 12) * interval '1 month') AS creation_month,
	  COUNT(*) AS courses_created
	FROM
	  courses c
	GROUP BY
	  creation_month
	ORDER BY
	  courses_created DESC;

Задача 2.

Вы аналитик продаж в компании, которая занимается продажей квартир. Вам нужно определить самый дорогой район по продажам.

У вас есть файл flat_data.csv, в котором содержится вся нужная вам информация,

Описание данных:

Поле	Описание
location	Расположение квартиры (строка)
rooms	Количество комнат в квартире (целое число от 1 до 5)
ceiling	Высота потолков (дробное число от 2.7 до 3.3)
balcony	Количество балконов (число от 0 до 4)
floor	Номер этажа (число от 1 до 20)
area	Площадь квартиры (дробное число от 30.0 до 144.52)
price	Цена квартиры (число от 5 000 000 до 100 000 000)

Требуется

1. Загрузить данные из файла flat_data.csv - файл должен находиться в том же проекте, что и программа.
2. Создать новый столбец price_meter – цена за квадратный метр. Считать по формуле: price / area
3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

location
rooms
ceiling
balcony
floor
area
price
price_meter

Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке не убывания по полю price_meter.

Используйте sort_values()

5. Для оценки самой подходящей модели постройте гистограмму распределения цены квадратного метра от локации

Ось X — Цена за квадратный метр

Ось Y — Локация

Укажите подписи для осей и название график.

Используйте boxplot(), plt.xlabel(), plt.ylabel(), plt.title().

6. Сохранить изображение как flat_price.png

Запишите свое решение в файл: task_2.py.

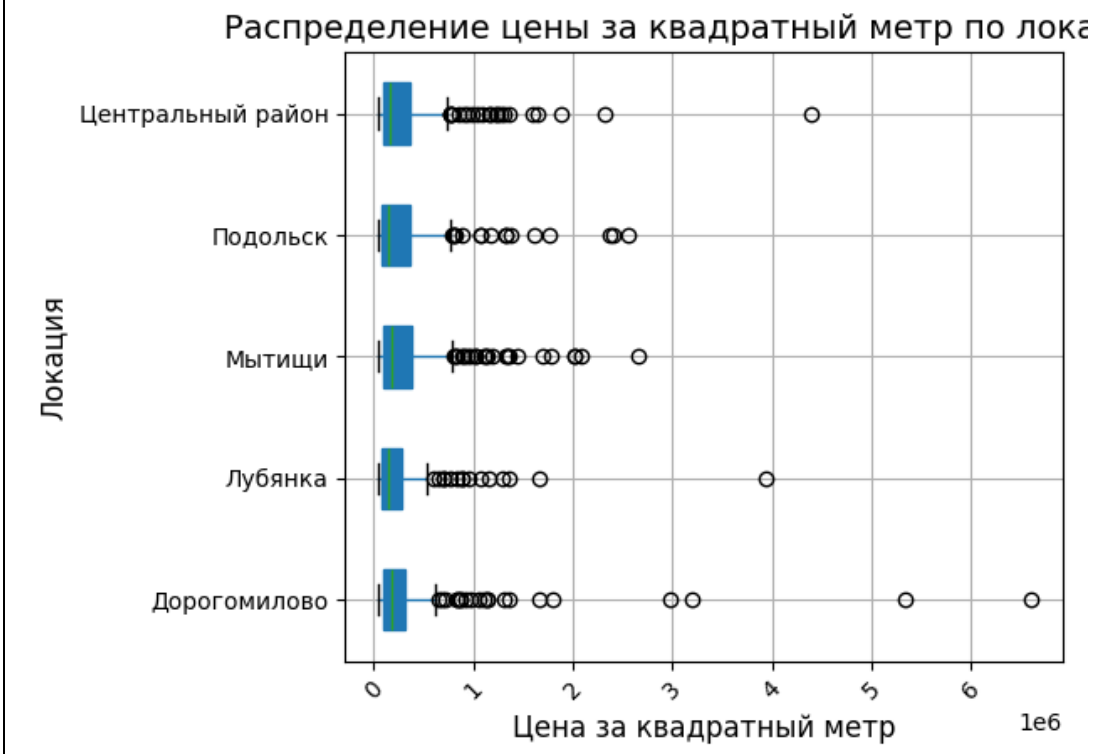
Документацию по приведенным методам можно получить с помощью функции help().

Решение

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('flat_data.csv')

df['price_meter'] = df['price'] / df['area']

df = df[['location', 'rooms', 'ceiling', 'balcony', 'floor', 'area', 'price', 'price_meter']]

sorted_df = df.sort_values('price_meter')
print(sorted_df.head(10))

plt.figure(figsize=(12, 8))

df.boxplot(column='price_meter', by='location', vert=False, grid=True,
patch_artist=True)

plt.xlabel('Цена за квадратный метр', fontsize=12)
plt.ylabel('Локация', fontsize=12)
plt.title('Распределение цены за квадратный метр по локациям', fontsize=14)
plt.suptitle('')
plt.xticks(rotation=45)

plt.tight_layout()
plt.savefig('flat_price.png')
plt.show()
```

Вывод 10 строк

	location	rooms	ceiling	...	area	price	price_meter
544	Лубянка	4	2.3	...	140.26	5160071	36789.326964
863	Центральный район	4	2.5	...	138.43	5104523	36874.398613
150	Мытищи	3	2.2	...	142.82	5680135	39771.285534
184	Лубянка	1	2.5	...	138.73	5720962	41238.102790
691	Мытищи	3	2.5	...	128.06	5307570	41445.962830
606	Лубянка	4	3.0	...	121.29	5046594	41607.667574
404	Лубянка	4	2.9	...	142.74	5959471	41750.532437
28	Подольск	3	3.0	...	131.92	5515686	41810.839903
144	Лубянка	1	2.4	...	119.82	5238744	43721.782674
954	Мытищи	2	2.9	...	124.77	5583951	44753.955278

График	

Задача 3.

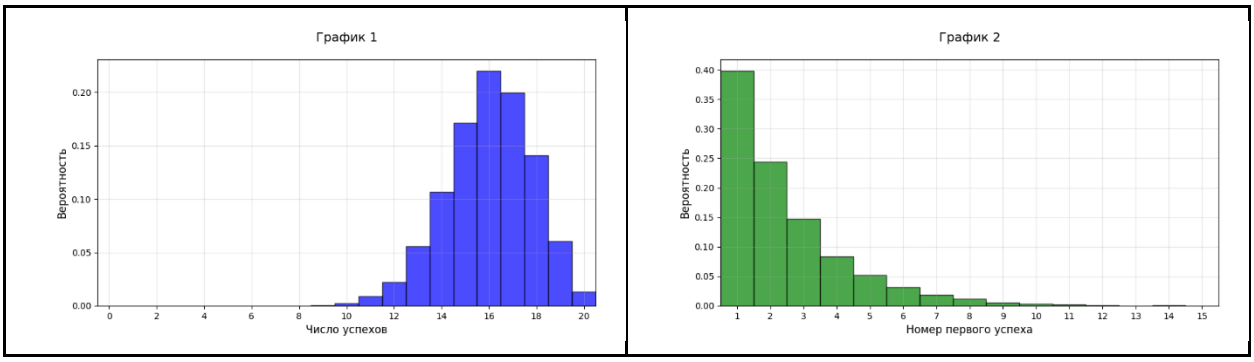
В университете 8 разных предметов. Нужно составить расписание из 5 пар с условиями:

1. Первая пара — всегда математика.
2. Ровно одна пара должна быть гуманитарной (их 3 в списке)

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 2 520

Задача 4.



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:

- Для Графика 1: вероятность успеха (p).
- Для Графика 2: укажите вероятность успеха (p).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt

Решение:

	График 1. Биномиальное распределение График 2. Геометрическое распределение
	График 1. $p \approx 0.8$ График 2. $p \approx 0.4$
3.	*График 1:* дискретные значения (целые числа от 0 до 20), пик около 16, характерная форма колокола — соответствует биномиальному распределению. *График 2:* убывающая ступенчатая форма с максимумом на 1, значения — только натуральные числа. Это соответствует геометрическому распределению, описывающему номер первого успеха.

Допустимые погрешности при учете ответа:

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)

σ: ±10–15% (например, σ = 10 → ответ 9–11)

n: точное значение (например, n = 20 → только 20, так как это целое число).

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)

Вариант 7

Задача 1.

Образовательная онлайн-платформа анализирует активность пользователей и популярность курсов для улучшения качества обучения и обновления контента. Вам, как аналитику, поручили подготовить выборки из базы PostgreSQL.

Структура базы данных:

- instructors (id, name, expertise) — преподаватели.
- courses (id, title, year, instructor_id, duration_hours) — курсы.
- students (id, name, email, registration_date) — студенты.
- course_enrollments (id, course_id, student_id, enrollment_date, completion_date) — записи о прохождении курсов.

Подготовьте SQL-запросы к базе данных для следующих целей:

1. Выявление самых активных периодов обучения

Определить дни недели и время суток с наибольшей активностью студентов, вывести день, время суток, число записей, отсортируйте по убыванию записей (Время суток - утро, если enrollment_date - с 6 до 12, день - с 13 до 18, вечер - с 19 до 23 часов)

2. Самый популярный год выпуска курсов

Определить в какой год было больше всего выпусков курсов, выведите год, количество курсов в этот год.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT
	  EXTRACT(DOW FROM ce.enrollment_date) AS day_of_week,
	  CASE
	  WHEN EXTRACT(HOUR FROM ce.enrollment_date) BETWEEN 6 AND 12 THEN 'Утро'
	  WHEN EXTRACT(HOUR FROM ce.enrollment_date) BETWEEN 13 AND 18 THEN 'День'
	  WHEN EXTRACT(HOUR FROM ce.enrollment_date) BETWEEN 19 AND 23 THEN 'Вечер'
	  ELSE 'Ночь'
	  END AS time_of_day,
	  COUNT(*) AS enrollments
	FROM
	  course_enrollments ce
	GROUP BY
	  day_of_week, time_of_day
	ORDER BY
	  enrollments DESC;

2	SELECT year, COUNT(*) as courses_count FROM courses GROUP BY year ORDER BY courses_count DESC LIMIT 1;

Задача 2.

Вы работаете в экологической компании. Вам нужно посмотреть динамику изменения качества воды за последние 10 лет.

У вас есть файл water_data.csv, в котором есть вся нужная для вас информация,

Описание данных:

Поле	Описание
location	Расположение водохранилища (строка)
trace_elements	Содержание микроэлементов (дробное число от 0.1 до 99.9)
purity	Чистота воды (дробное число от 0.1 до 99.9)
date	Дата сбора материала (в формате YYYY-MM)

Требуется

1. Загрузить данные из файла water_data.csv - файл должен находиться в том же проекте, что и программа.

2. Преобразуйте дату из формата YYYY-MM в YYYY и назовите столбец year. Создайте столбец медианных значений количества микроэлементов по этому году и назовите его median
3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

location
trace_elements
purity
Date
Year
median

Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке убывания по полю purity.

Используйте sort_values()

5. Для оценки изменения качества воды за последние 10 лет, постройте график, где видно, как с каждым годом менялась концентрация микроэлементов (по медианным значениям)

Ось X — Год

Ось Y — Медианное значение содержания микроэлементов в воде в этот год

Укажите подписи для осей и название график.

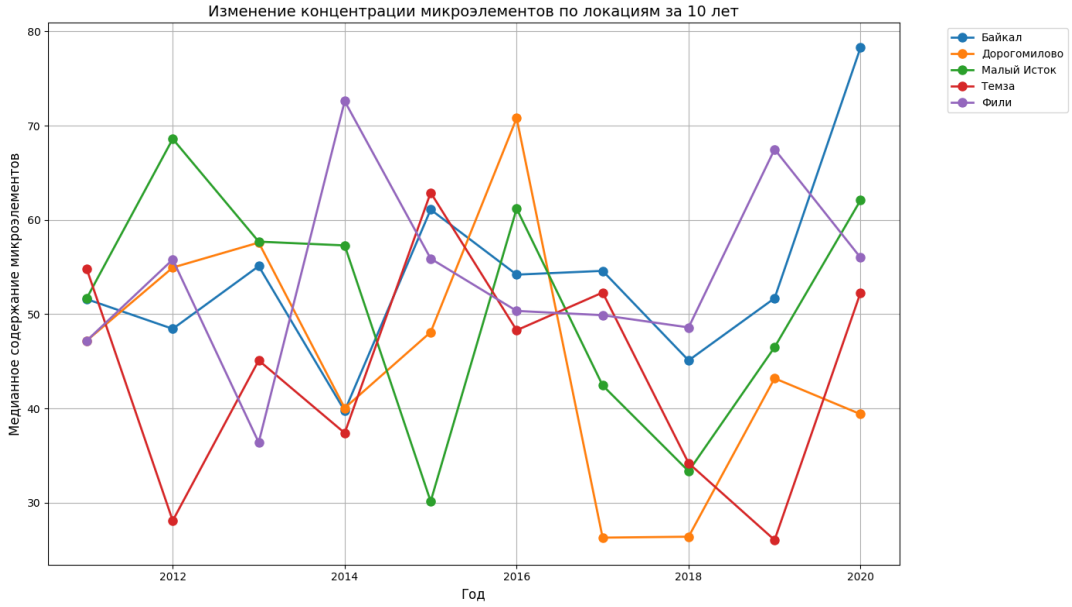
Используйте scatterplott(), plt.xlabel(), plt.ylabel(), plt.title().

6. Сохранить изображение как water_plot.png

Запишите свое решение в файл: task_2.py.

Документацию по приведенным методам можно получить с помощью функции ***help()***.

Решение	import pandas as pd
	import matplotlib.pyplot as plt
	
	df = pd.read_csv('water_data.csv')
	
	df['year'] = pd.to_datetime(df['date']).dt.year
	median_by_year = df.groupby(['year',
	'location'])['trace_elements'].median().reset_index()
	
	df['median'] = df.groupby(['year', 'location'])['trace_elements'].transform('median')
	df = df[['location', 'trace_elements', 'purity', 'date', 'year', 'median']]
	
	print(df.sort_values('purity', ascending=False).head(10))
	
	plt.figure(figsize=(14, 8))
	
	recent_years = sorted(df['year'].unique())[-10:]
	recent_data = median_by_year[median_by_year['year'].isin(recent_years)]
	
	for location in recent_data['location'].unique():
	  loc_data = recent_data[recent_data['location'] == location].sort_values('year')
	  plt.plot(loc_data['year'], loc_data['trace_elements'],
	  marker='o', markersize=8, linewidth=2,
	  label=location)
	
	plt.xlabel('Год', fontsize=12)
	plt.ylabel('Медианное содержание микроэлементов', fontsize=12)
	plt.title('Изменение концентрации микроэлементов по локациям за 10 лет',
	fontsize=14)
	plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
	plt.grid(True)
	
	plt.tight_layout()
	plt.savefig('water_plot.png', bbox_inches='tight')
	plt.show()

Вывод 10 строк	location trace_elements purity date year median 726 Байкал 74.5 9.9 2015-12 2015 61.10 205 Байкал 89.8 9.9 2019-01 2019 51.70 6 Фили 23.3 9.9 2010-07 2010 58.30 973 Темза 73.6 9.9 2013-02 2013 45.10 37 Темза 90.3 9.9 2012-08 2012 28.10 642 Малый Исток 25.6 9.9 2018-04 2018 33.35 913 Малый Исток 65.2 9.8 2016-05 2016 61.20 621 Дорогомилово 31.9 9.8 2013-02 2013 57.60 989 Малый Исток 11.0 9.8 2015-07 2015 30.15 454 Малый Исток 86.6 9.8 2014-06 2014 57.30
График	

Задача 3.

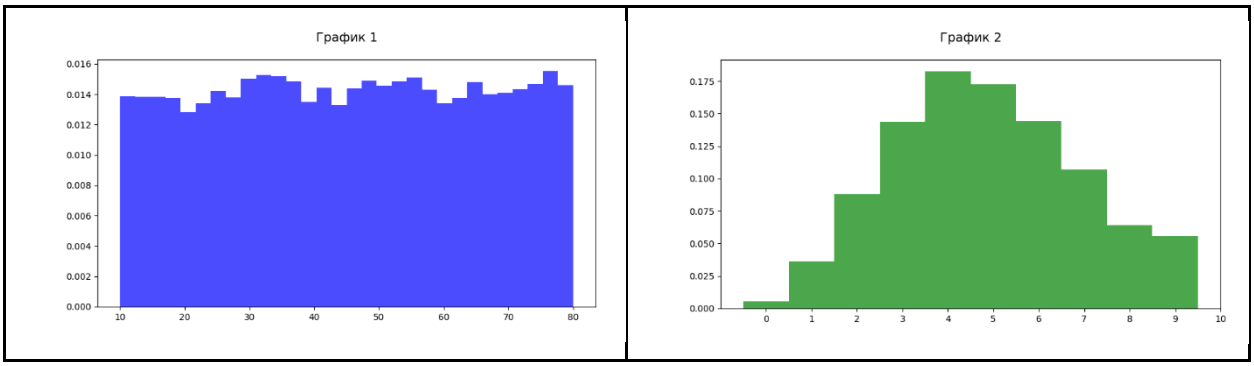
В магазине 7 красных, 3 синих и 5 зелёных шаров. Случайно берут 4 шара.

Какова вероятность того, что среди выбранных шаров будет ровно два красных? Ответ округлите до сотых и запишите в виде процентов.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 43%

Задача 4.



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределений:
   - Для Графика 1: границы интервала (a и b).
   - Для Графика 2: параметр λ (лямбда).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Решение:

1. График 1 — равномерное, График 2 — Пуассона
2. График 1: a ≈ 10, b ≈ 80 (значения равномерно распределены от 10 до 80).

График 2: λ ≈ 5 (пик около 5, среднее = λ)

Допустимые погрешности:

a и b: ±2–3% от диапазона (например, данные от 10 до 80 → a ≈ 10 ± 0.5, b ≈

80 ± 0.5).

λ: ±1–2 единицы (например, λ = 5 → ответ 4–6)

3. График 1: гистограмма плоская, все значения в диапазоне [10, 80] встречаются примерно одинаково часто.

График 2: дискретные значения, пик около 5, асимметрия вправо (характерно для Пуассона).

Вариант 8

Задача 1.

Образовательная онлайн-платформа анализирует активность пользователей и популярность курсов для улучшения качества обучения и обновления контента. Вам, как аналитику, поручили подготовить выборки из базы PostgreSQL.

Структура базы данных:

- instructors (id, name, expertise) — преподаватели.
- courses (id, title, year, instructor_id, duration_hours) — курсы.
- students (id, name, email, registration_date) — студенты.
- course_enrollments (id, course_id, student_id, enrollment_date, completion_date) — записи о прохождении курсов.

Подготовьте SQL-запросы к базе данных для следующих целей:

1. Количество курсов каждого преподавателя

Выведите имя преподатвателя, количество его курсов, отсортируйте по уменьшению количества курсов.

2. Самые активные студенты

Определить студентов с наибольшим количеством записей на курсы, выведите их имена, количество записей на курс, отсортируйте по убыванию записей.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

```
1  SELECT i.name, COUNT(c.id) as courses_count
   FROM instructors i
   LEFT JOIN courses c ON i.id = c.instructor_id
   GROUP BY i.name
   ORDER BY courses_count DESC;
```

2	SELECT s.name, COUNT(ce.id) as enrollments_count FROM students s JOIN course_enrollments ce ON s.id = ce.student_id GROUP BY s.name ORDER BY enrollments_count DESC

Задача 2.

Вы преподаватель, которому поручили выяснить как учатся в разных университетах. Вам дали базу данных, в которой содержится успеваемость и посещаемость учеников в разных высших учебных заведениях.

У вас есть файл student_performance.csv в котором хранится вся нужная информация,

Описание данных:

Поле	Описание
university	Название университета (строка)
subject	Название предмета (строка)
gpa	Средний балл (дробное число от 1.0 до 5.0)
attendance	Посещаемость в % (дробное число от 0.1 до 100.0)
exam_date	Дата экзамена (в формате YYYY-MM)

Требуется

1. Загрузить данные из файла student_performance.csv - файл должен находиться в том же проекте, что и программа.
2. Преобразуйте дату из формата YYYY-MM в YYYY (новый столбец year). Создайте столбец median_gpa с медианным баллом (gpa) по каждому году для каждого университета.
3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

university
subject
gpa
attendance
exam_date
year
median_gpa

Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке убывания по полю attendance.

Используйте sort_values()

5. Постройте график изменения медианного балла за последние 5 лет по университетам:

Ось X — Год
Ось Y — Медианный GPA

Укажите подписи для осей и название график.

Используйте plt.plott(), plt.xlabel(), plt.ylabel(), plt.title().

6. Сохранить изображение как student_performance_plot.png

Запишите свое решение в файл: task_2.py.

*Документацию по приведенным методам можно получить с помощью функции **help()**.*

Решение

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("student_performance.csv")

df["year"] = pd.to_datetime(df["exam_date"]).dt.year
df["median_gpa"] = df.groupby(["university", "year"])["gpa"].transform("median")

df = df[["university", "subject", "gpa", "attendance", "exam_date", "year",
"median_gpa"]]

print(df.sort_values("attendance", ascending=False).head(10))

plt.figure(figsize=(12, 6))

last_5_years = sorted(df["year"].unique())[-5:]

for uni in df["university"].unique():
    uni_data = df[(df["university"] == uni) & (df["year"].isin(last_5_years))]
    yearly_median = uni_data.groupby("year")["gpa"].median().reset_index()
    plt.plot(yearly_median["year"], yearly_median["gpa"], marker="o", label=uni)

plt.xlabel("Год")
plt.ylabel("Медианный GPA")
plt.title("Динамика успеваемости студентов по университетам")
plt.grid(True)
plt.legend(bbox_to_anchor=(1.05, 1), loc="upper left")

plt.tight_layout()
plt.savefig("student_performance_plot.png", bbox_inches="tight")
plt.show()
```

Вывод 10 строк	university subject gpa attendance exam_date year median_gpa 500 МФТИ Биология 5.0 100.0 2018-05 2018 4.35 156 СПбГУ Математика 4.5 100.0 2021-03 2021 4.30 800 МГУ Химия 4.1 100.0 2022-02 2022 4.40 580 МГУ Биология 3.2 100.0 2023-09 2023 4.30 582 СПбГУ Экономика 4.7 100.0 2021-05 2021 4.30 161 НГУ Физика 3.4 100.0 2023-05 2023 3.95 798 КФУ Биология 4.2 100.0 2022-09 2022 3.90 163 НГУ Биология 4.7 100.0 2018-12 2018 4.00 583 КФУ Математика 4.5 100.0 2023-01 2023 3.80 796 НГУ Биология 4.3 100.0 2020-05 2020 4.00
График	

Задача 3.

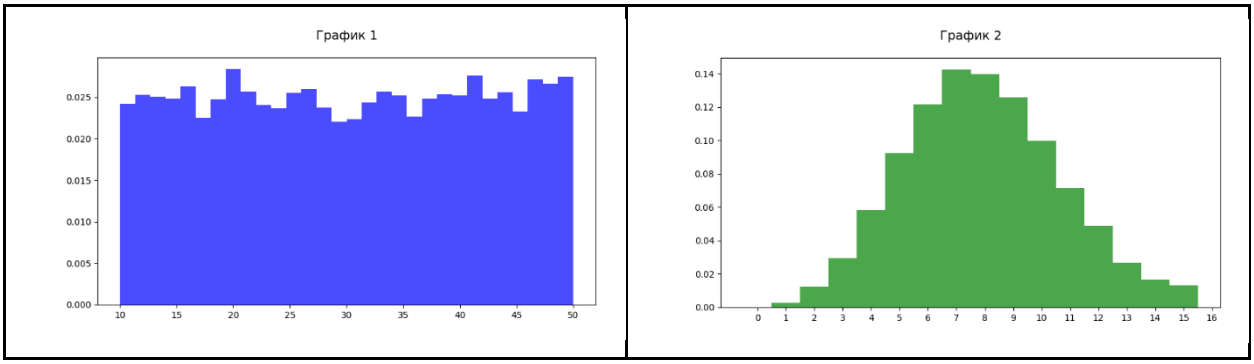
В коробке 5 ручек, 4 карандаша и 3 маркера. Случайно выбирают 4 предмета.

Какова вероятность того, что среди случайно вытащенных канцелярских принадлежностей окажется ровно два карандаша? Ответ округлите до сотых и запишите в процентах.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 34%

Задача 4.



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределений:
   - Для Графика 1: границы интервала (a и b).
   - Для Графика 2: параметр λ (лямбда).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Решение:

1. График 1 — равномерное, График 2 — Пуассона

2. График 1: a ≈ 10, b ≈ 50 (значения равномерно распределены от 10 до 50).

График 2: λ ≈ 8 (пик около 8, среднее = λ)

Допустимые погрешности:

a и b: ±2–3% от диапазона (например, данные от 10 до 50 → a ≈ 10 ± 0.5, b ≈

50 ± 0.5).

λ: ±1–2 единицы (например, λ = 8 → ответ 7–9)

3. График 1: гистограмма плоская, все значения в диапазоне [10, 50] встречаются примерно одинаково часто.

График 2: дискретные значения, пик около 8, асимметрия вправо (характерно для Пуассона).

Вариант 9

Задача 1.

Образовательная онлайн-платформа анализирует активность пользователей и популярность курсов для улучшения качества обучения и обновления контента. Вам, как аналитику, поручили подготовить выборки из базы PostgreSQL.

Структура базы данных:

- instructors (id, name, expertise) — преподаватели.
- courses (id, title, year, instructor_id, duration_hours) — курсы.
- students (id, name, email, registration_date) — студенты.
- course_enrollments (id, course_id, student_id, enrollment_date, completion_date) — записи о прохождении курсов.

Подготовьте SQL-запросы к базе данных для следующих целей:

1. Топ 5 самых популярных курсов

Найти топ 5 курсов с наибольшим количеством записей студентов, выведите их название, количество записей, отсортируйте по убыванию записей.

2. Преподаватели без курсов

Найти преподавателей, которые еще не создали ни одного курса, выведите их имена.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

```
1  SELECT c.title, COUNT(ce.id) as enrollments_count
   FROM courses c
   JOIN course_enrollments ce ON c.id = ce.course_id
   GROUP BY c.title
   ORDER BY enrollments_count DESC
   LIMIT 5;
```

2	SELECT i.name FROM instructors i LEFT JOIN courses c ON i.id = c.instructor_id WHERE c.id IS NULL;

Задача 2.

Вы аналитик, которому поручили выяснить как сдают нормативы в школе разные группы учеников. Вам дали базу данных, в которой содержится вся информация о сданных упражнений.

У вас есть файл sport_results.csv, в котором содержится вся нужная информация,

Описание данных:

Поле	Описание
student_id	Уникальный номер ученика (1-200)
class	Класс (5-11)
sport_type	Вид спорта (Бег, Плавание, Прыжки)
result	Результат (в секундах для бега/плавания, в метрах для прыжков)
year	Год проведения теста (2021, 2022, 2023)

Требуется

1. Загрузить данные из файла sport_results.csv - файл должен находиться в том же проекте, что и программа.
2. Создайте столбец result_category:
   "Отлично" (лучшие 25% результатов)
   "Хорошо" (средние 50%)
   "Нужно тренироваться" (нижние 25%)

3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

student_id
class
sport_type
result
year
result_category

Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке убывания по полю result.

Используйте sort_values()

5. Постройте столбчатую диаграмму:

Ось X — Вид спорта
Ось Y — Средний результат

Укажите подписи для осей и название графика, для удобства восприятия, выделите столбцы разными цветами.

Используйте bar plot, plt.xlabel(), plt.ylabel(), plt.title().

6. Сохранить изображение как sport_results.png

Запишите свое решение в файл: task_2.py.

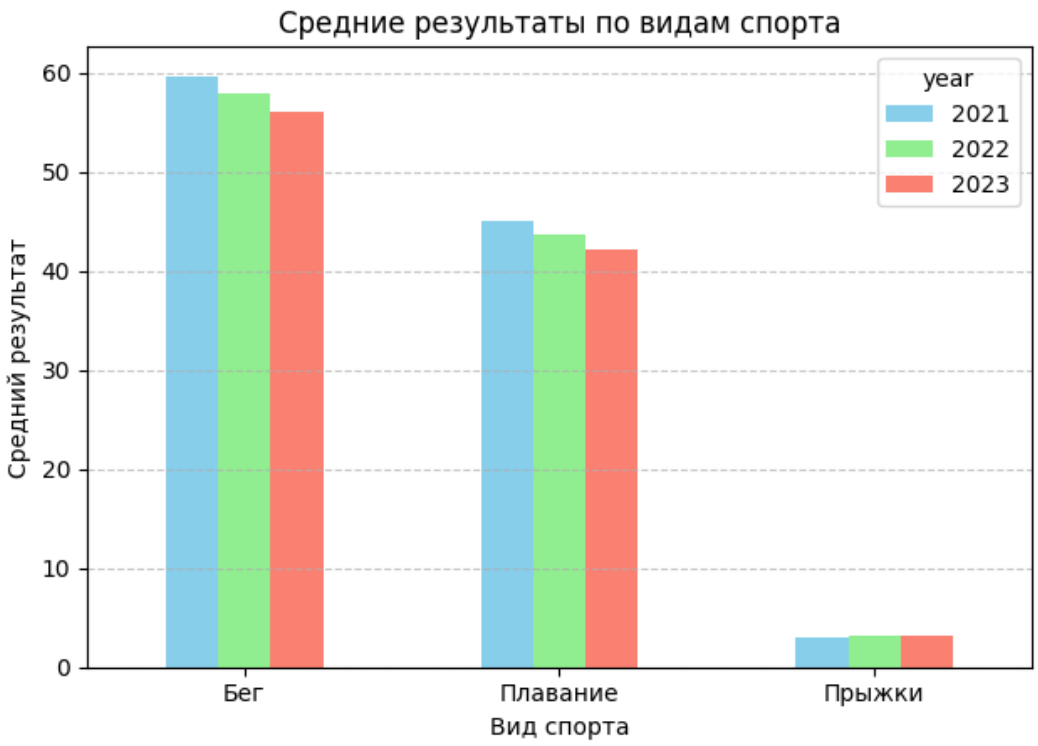
Документацию по приведенным методам можно получить с помощью функции ***help()***.

Решение

Задание 2

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("sport_results.csv")

for sport in df['sport_type'].unique():
    sport_df = df[df['sport_type'] == sport]
    df.loc[df['sport_type'] == sport, 'result_category'] = pd.qcut(
        sport_df['result'],
        q=[0, 0.25, 0.75, 1],
        labels=["Отлично", "Хорошо", "Нужно тренироваться"]
    )

top_10_results = df.sort_values('result', ascending=False).head(10)
print(top_10_results.to_string(index=False))

plt.figure(figsize=(10, 6))

grouped = df.groupby(['sport_type', 'year'])['result'].mean().unstack()

ax = grouped.plot(kind='bar', color=['skyblue', 'lightgreen', 'salmon'])
plt.suptitle('')

ax.set_title("Средние результаты по видам спорта")
ax.set_xlabel("Вид спорта")
ax.set_ylabel("Средний результат")
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.tight_layout()
```

	plt.savefig("sport_results.png") plt.show()

Вывод 10 строк	student_id class sport_type result year result_category 147 7 Бег 64.89 2021 Нужно тренироваться 101 11 Бег 64.81 2021 Нужно тренироваться 200 7 Бег 64.72 2021 Нужно тренироваться 172 9 Бег 64.71 2021 Нужно тренироваться 83 11 Бег 64.65 2021 Нужно тренироваться 21 6 Бег 64.64 2021 Нужно тренироваться 168 6 Бег 64.63 2021 Нужно тренироваться 163 7 Бег 64.62 2021 Нужно тренироваться 62 10 Бег 64.56 2021 Нужно тренироваться 152 5 Бег 64.50 2021 Нужно тренироваться
График	

Задача 3.

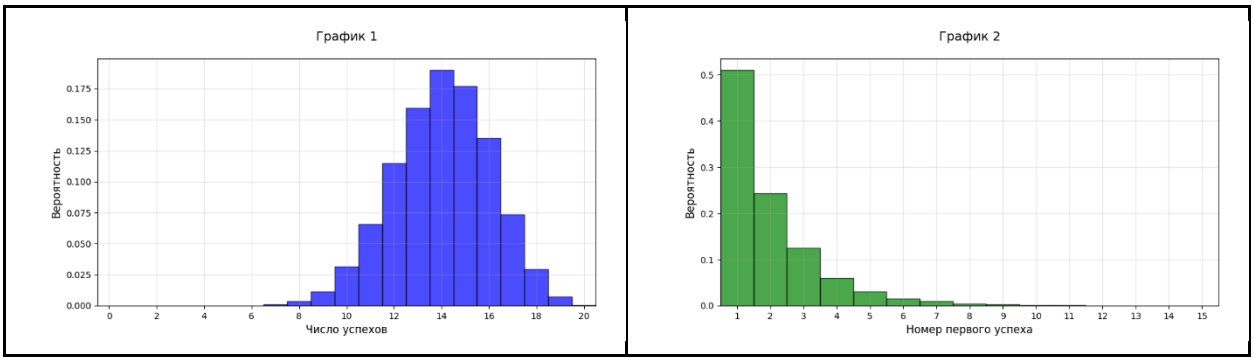
В спортзале для тренировки есть спортивный инвентарь: 7 мячей, 4 скакалки, 2 гантели

Случайно берут 5 предметов. Какова вероятность, что среди них будет ровно 3 мяча и 1 скакалка? Результат округлите до сотых

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.22

Задача 4.



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:

- Для Графика 1: вероятность успеха (p).
- Для Графика 2: укажите вероятность успеха (p).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt

Решение:

	График 1. Биномиальное распределение График 2. Геометрическое распределение
	График 1. $p \approx 0.7$ График 2. $p \approx 0.5$
3.	*График 1:* дискретные значения (целые числа от 0 до 20), пик около 14, характерная форма колокола — соответствует биномиальному распределению. *График 2:* убывающая ступенчатая форма с максимумом на 1, значения — только натуральные числа. Это соответствует геометрическому распределению, описывающему номер первого успеха.

Допустимые погрешности при учете ответа:

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)

σ: ±10–15% (например, σ = 10 → ответ 9–11)

n: точное значение (например, n = 20 → только 20, так как это целое число).

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)

Вариант 10

Задача 1.

Образовательная онлайн-платформа анализирует активность пользователей и популярность курсов для улучшения качества обучения и обновления контента. Вам, как аналитику, поручили подготовить выборки из базы PostgreSQL.

Структура базы данных:

- instructors (id, name, expertise) — преподаватели.
- courses (id, title, year, instructor_id, duration_hours) — курсы.
- students (id, name, email, registration_date) — студенты.
- course_enrollments (id, course_id, student_id, enrollment_date, completion_date) — записи о прохождении курсов.

Подготовьте SQL-запросы к базе данных для следующих целей:

1. Активность преподавателей по годам

Показать динамику создания курсов преподавателями, вывести их имена, год, количество созданных курсов за этот год.

2. Среднее количество курсов на студента

Рассчитать среднюю вовлеченность студентов (среднее число записей на курсы на студента)

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

```
1  SELECT i.name, c.year, COUNT(c.id) as courses_created
   FROM instructors i
   JOIN courses c ON i.id = c.instructor_id
   GROUP BY i.name, c.year
   ORDER BY i.name, c.year;
```

2	SELECT AVG(course_count) as avg_courses_per_student
	FROM (
	  SELECT student_id, COUNT(*) as course_count
	  FROM course_enrollments
	  GROUP BY student_id
	) as student_courses;

Задача 2.

Вы аналитик данных, которому предстоит определить в каком из предприятий была самая большая скидка и на какую продукцию.

У вас есть файл pizza_discounts.csv,

Описание данных:

Поле	Описание
restaurant	Название ресторана (ПиццаМаркет, ДодоПицца, ПапаДжонс, ПиццаХат)
pizza_type	Вид пиццы (Пепперони, Маргарита, Гавайская, ЧетыреСыра)
original_price	Оригинальная цена (целое число)
discount_price	Цена со скидкой (целое число)
month	Месяц (1-12)
year	Год (2022, 2023)

Требуется

1. Загрузить данные из файла pizza_discounts.csv - файл должен находиться в том же проекте, что и программа.
2. Создайте столбец discount_percent (процент скидки) по формуле:
   ((original_price - discount_price) / original_price) * 100
3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

restaurant
pizza_type
original_price
discount_price
discount_percent
month
year

Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке убывания по полю discount_percent.

Используйте sort_values()

5. Постройте линейный график:

Ось X — Месяц
Ось Y — Средний процент скидки

Укажите подписи для осей и название графика, для удобства восприятия, выведите легенду.

Используйте plt.plot, plt.xlabel(), plt.ylabel(), plt.title().

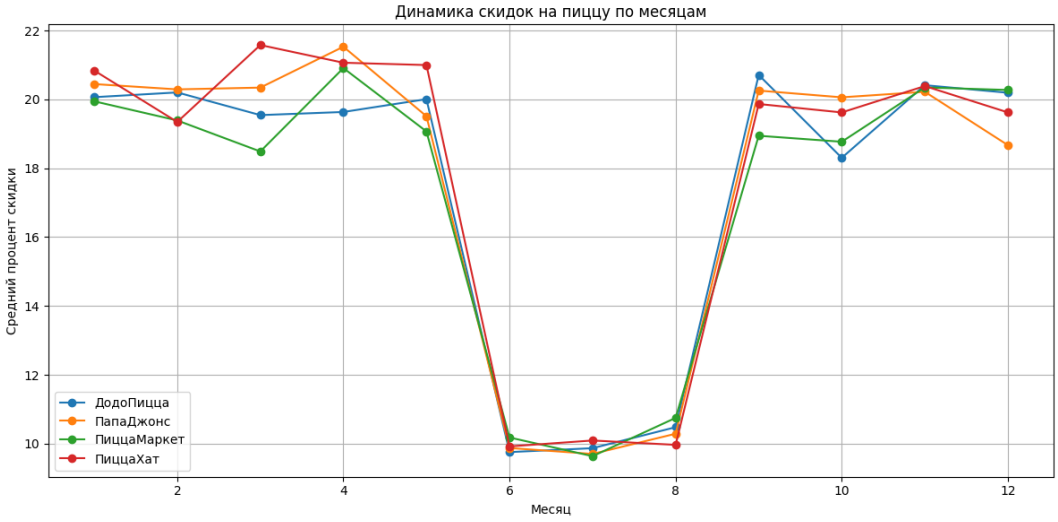
6. Сохранить изображение как pizza_discounts_plot.png

Запишите свое решение в файл: task_2.py.

Документацию по приведенным методам можно получить с помощью функции ***help()***.

Решение

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("pizza_discounts.csv")

df["discount_percent"] = ((df["original_price"] - df["discount_price"]) /
df["original_price"]) * 100

top_discounts = df.sort_values("discount_percent", ascending=False).head(10)
print(top_discounts[["restaurant", "pizza_type", "original_price", "discount_price",
"discount_percent", "month", "year"]].to_string(index=False))

grouped = df.groupby(["month",
"restaurant"])["discount_percent"].mean().unstack()

plt.figure(figsize=(12, 6))
for restaurant in grouped.columns:
    plt.plot(grouped.index, grouped[restaurant], marker="o", label=restaurant)

plt.title("Динамика скидок на пиццу по месяцам")
plt.xlabel("Месяц")
plt.ylabel("Средний процент скидки")
plt.legend()
plt.grid(True)

plt.tight_layout()
plt.savefig("pizza_discounts_plot.png")
plt.show()
```

Вывод 10 строк	restaurant pizza_type original_price discount_price discount_percent month year ПапаДжонс Гавайская 591 449 24.027073 2 2023 ПиццаМаркет ЧетыреСыра 616 468 24.025974 12 2022 ПапаДжонс Маргарита 606 461 23.927393 11 2022 ДодоПицца Пепперони 590 449 23.898305 9 2023 ПапаДжонс ЧетыреСыра 653 497 23.889740 11 2022 ПиццаХат ЧетыреСыра 721 549 23.855756 5 2022 ПапаДжонс Маргарита 588 448 23.809524 4 2023 ПиццаХат Гавайская 689 525 23.802612 3 2023 ПапаДжонс Пепперони 647 493 23.802164 9 2023 ПиццаМаркет ЧетыреСыра 601 458 23.793677 4 2023
График	

Задача 3.

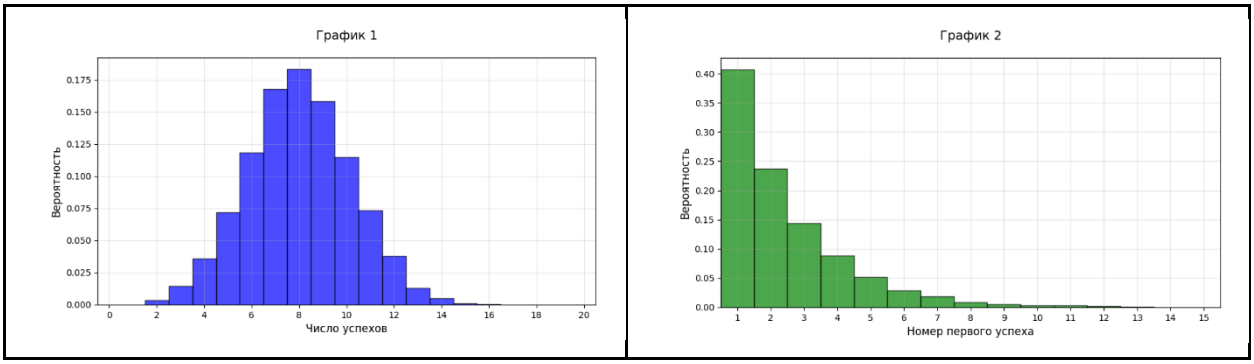
В библиотеке 6 детективов, 5 романов и 4 фантастики. Случайно берут 5 книг.

Какова вероятность, что среди них будет ровно 3 детектива и 1 роман? Ответ, округлите до сотых и запишите в процентах.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 13%

Задача 4.



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:

- Для Графика 1: вероятность успеха (p).
- Для Графика 2: укажите вероятность успеха (p).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt

Решение:

	График 1. Биномиальное распределение График 2. Геометрическое распределение
	График 1. $p \approx 0.4$ График 2. $p \approx 0.4$
3.	*График 1:* дискретные значения (целые числа от 0 до 20), пик около 8, характерная форма колокола — соответствует биномиальному распределению. *График 2:* убывающая ступенчатая форма с максимумом на 1, значения — только натуральные числа. Это соответствует геометрическому распределению, описывающему номер первого успеха.

Допустимые погрешности при учете ответа:

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)

σ: ±10–15% (например, σ = 10 → ответ 9–11)

n: точное значение (например, n = 20 → только 20, так как это целое число).

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)

Вариант 11

Задача 1.

Вы работаете в качестве аналитика в компании, занимающейся публикацией научных работ в онлайн-журналах. В данный момент происходит подведение итогов работы компании за текущий фискальный год.

В вашем распоряжении есть таблицы базы данных с описаниями:

authors (id, name, institution)– основная информация об авторах статей.

articles (id, title, field, author_id, publish_year) – информация о статьях/работах.

downloads (id, article_id, download_date) – информация о загрузках статей.

Вашей задачей является формирование запросов, удовлетворяющих следующим условиям:

1. Найти авторов, чьи имена начинаются на «Ол» или содержат «др». Вывести имя и учреждение.

2. Найти области знаний (field), по которым было более 30 загрузок. Вывести область и количество загрузок.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT name, institution FROM authors WHERE name LIKE 'Ол%' OR name LIKE '%др%';
2	SELECT a.field, COUNT(d.id) AS download_count FROM articles a JOIN downloads d ON a.id = d.article_id GROUP BY a.field HAVING COUNT(d.id) > 30;

Задача 2.

Вы работаете аналитиком в магазине продуктов крупной сети, и вам предстоит проанализировать продажи прохладительных напитков у конкурентов.

У вас есть файл 11_drinks.csv, содержащий информацию о магазинах, напитках, их ценах и продажах за определенный период времени.

Описание данных:

Поле	Описание
item	Товар (Добрый Кола, Тоник...)
store	Магазин (Пятерочка, Перекресток...)
cost	Цена (84-350)
volume	Объем (1-10)
city	Город (Москва, Волгоград...)
amount	Количество (1-1000)

Требуется:

1. Загрузить данные из файла 11_drinks.csv - файл должен находиться в том же проекте, что и программа.

2. Создайте столбец price_category:

"Дешевые" (< 100 руб)

"Сбалансированные" (100-300 руб)

"Дорогие" (> 300 руб)

Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

item
store
cost
volume
city
amount
price_category

Проверьте, что порядок и формат соответствуют.

3. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке не убывания по полю cost.

Используйте sort_values().

4. Постройте гистограмму (hist) распределения цен:

Ось X — Цена

Ось Y — Количество напитков

Разделите по товарам (разные цвета). Подпишите оси и добавьте заголовок, выведите легенду.

Используйте matplotlib.pyplot.hist(), plt.xlabel(), plt.ylabel(), plt.title().

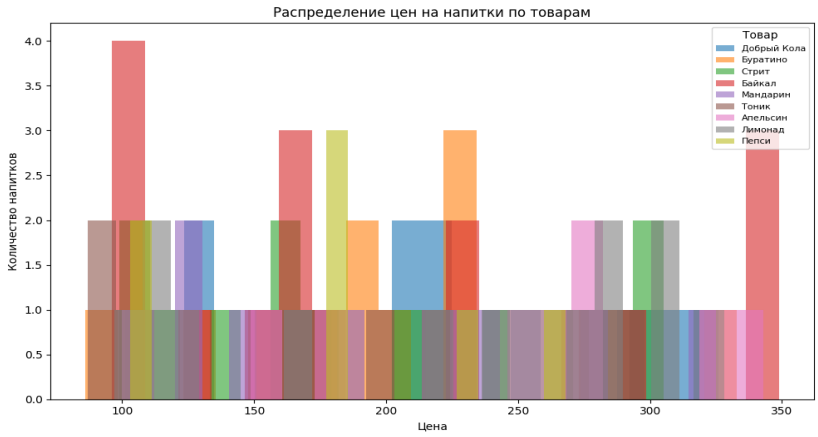
Сохранить изображение как drinks_prices_hist.png

Запишите свое решение в файл: task_2.py.

Документацию по приведенным методам можно получить с помощью функции help().

Решение	import pandas as pd import matplotlib.pyplot as plt

```
df = pd.read_csv("11_drinks.csv")


def categorize_price(price):
if price < 100:
return "Дешевые"
elif 100 <= price <= 300:
return "Сбалансированные"
else:
return "Дорогие"


df["price_category"] = df["cost"].apply(categorize_price)


df = df[["item", "store", "cost", "volume", "city", "amount",
"price_category"]]


df_sorted = df.sort_values(by="cost")
print(df_sorted.head(10))


plt.figure(figsize=(10, 6))


items = df["item"].unique()


for item in items:
    subset = df[df["item"] == item]
    plt.hist(subset["cost"], bins=20, alpha=0.6, label=item)


plt.xlabel("Цена")
plt.ylabel("Количество напитков")
```

	plt.title("Распределение цен на напитки по товарам") plt.legend(title="Товар", fontsize=8) plt.tight_layout() plt.savefig("drinks_prices_hist.png") plt.show()
Вывод 10 строк	item store cost volume city amount price_category 8 Буратино Магнит 86 2 Волгоград 934 Дешевые 53 Тоник Дикси 87 1 Уфа 403 Дешевые 14 Тоник Перекресток 94 2 Уфа 794 Дешевые 66 Байкал Перекресток 96 6 Екатеринбург 803 Дешевые 62 Лимонад Магнит 97 7 Волгоград 234 Дешевые 74 Стрит Пятерочка 99 6 Санкт-Петербург 436 Дешевые 76 Апельсин Дикси 100 3 Екатеринбург 658 Сбалансированные 60 Добрый Кола Магнит 101 9 Волгоград 291 Сбалансированные 24 Тоник Дикси 102 7 Санкт-Петербург 32 Сбалансированные 39 Байкал Пятерочка 102 8 Москва 416 Сбалансированные
График	

Задача 3.

Коллекционер антиквариата увлекается сбором старинных автобусных бумажных проездных билетов. Каждый билет содержит серию (одну букву от А до Я), и номер (пять цифр, начиная от 00001 до 99999).

Билет считается редким, если:

а. Серия — буквы от А до Г;

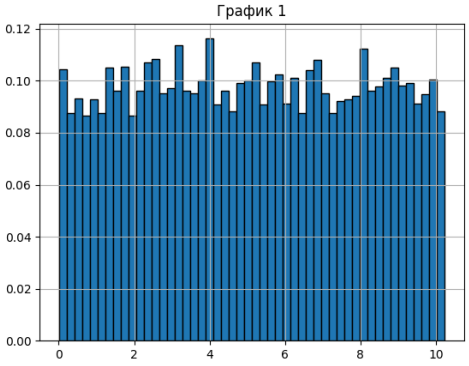
б. Номер — число в промежутке 00001-05001 включительно.

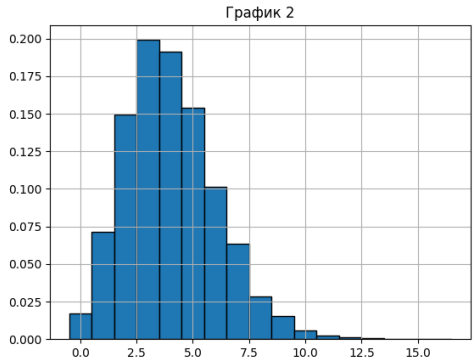
Найдите вероятность того, что любой случайно выбранный билет окажется редким.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.00606

Задача 4.





1. Определите тип распределения для каждого графика.

2. Оцените параметры распределения:

- Для графика 1: укажите границы интервала (a, b).
- Для графика 2: параметр лямбда (λ).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Равномерное, Гр.2 - Пуассона.
2	Для гр.1: **a ≈ 0.0, b ≈ 10.2**. Для гр.2: **λ ≈ 4.0.**
3	Гр.1: плоская гистограмма, значения в [0.0, 10.2] встречаются равномерно. Гр.2: дискретные значения, пик около 4.0, асимметрия вправо — характерно для Пуассона.

Вариант 12

Задача 1.

Вы работаете аналитиком в национальном музее искусств. Вам необходимо проанализировать данные для подготовки отчетов об экспонатах и их посещениях.

В вашем распоряжении есть таблицы базы данных с описаниями:

visitors (id, name, city, visit_date) — информация о посетителях;

exhibits (id, title, artist, year_created) — информация об экспонатах;

visits (id, visitor_id, exhibit_id, rating) — информация о посещениях.

Вам необходимо:

1. Найти всех посетителей, чьи имена начинаются на «Ан» или содержат буквосочетание «ов», которые посещали музей в 2024 году. Вывести имя и город.

2. Определить, какие экспонаты получили в среднем рейтинг выше 4.0. Вывести названия экспонатов, имена художников и средние рейтинги.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT DISTINCT v.name, v.city FROM visitors v JOIN visits vs ON v.id = vs.visitor_id WHERE (v.name LIKE 'Ан%' OR v.name LIKE '%ов%') AND EXTRACT(YEAR FROM v.visit_date) = 2024;
2	SELECT e.title, e.artist, AVG(vs.rating) AS avg_rating FROM exhibits e JOIN visits vs ON e.id = vs.exhibit_id GROUP BY e.id, e.title, e.artist HAVING AVG(vs.rating) > 4.0;

Задача 2.

Вы работаете аналитиком в магазине стройматериалов крупной сети, и вам предстоит проанализировать динамику изменения цен у конкурентов.

У вас есть файл 12_sales.csv, содержащий информацию о магазинах, напитках, их ценах и продажах за определенный период времени.

Описание данных:

Поле	Описание
store	Магазин (Лемана Про, Ашан...)
cost	Цена (1000-50000)
type	Тип товара (Лопата, цемент...)
city	Город (Москва, Волгоград...)
sales	Количество (1-1000)

date	Дата продажи (2023-2025) (YYYY-MM-DD)

Требуется:

1. Загрузить данные из файла 12_sales.csv - файл должен находиться в том же проекте, что и программа.

2. Преобразуйте дату из формата YYYY-MM-DD в YYYY (новый столбец year). Создайте столбец median_price с медианной ценой для каждого магазина по годам.

Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

store
cost
type
city
sales
date
year
median_price

Проверьте, что порядок и формат соответствуют.

3. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке не убывания по полю sales.

Используйте sort_values().

4. Постройте график изменения медианной цены по магазинам:

Ось X — Год

Ось Y — Медианная цена

Подпишите оси и добавьте заголовок, выведите легенду.

Используйте plt.plot(), plt.xlabel(), plt.ylabel(), plt.title().

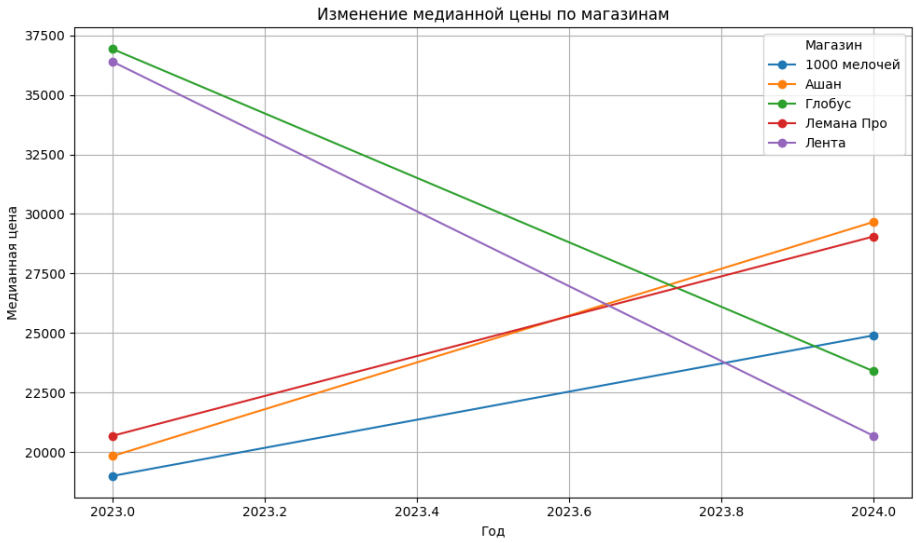
5. Сохранить изображение как item_sales_plot.png

6. Запишите свое решение в файл: task_2.py.

Документацию по приведенным методам можно получить с помощью функции help().

Решение	import pandas as pd
	import matplotlib.pyplot as plt
	
	df = pd.read_csv("12_sales.csv")
	
	df["date"] = pd.to_datetime(df["date"])
	df["year"] = df["date"].dt.year
	
	median_prices = df.groupby(["store", "year"])["cost"].median().reset_index()
	median_prices.rename(columns={"cost": "median_price"}, inplace=True)
	
	df = pd.merge(df, median_prices, on=["store", "year"])
	
	df = df[["store", "cost", "type", "city", "sales", "date", "year",
	"median_price"]]
	
	df_sorted = df.sort_values(by="sales")
	print(df_sorted.head(10))
	
	plt.figure(figsize=(10, 6))
	
	for store in median_prices["store"].unique():
	  subset = median_prices[median_prices["store"] == store]
	  plt.plot(subset["year"], subset["median_price"], marker='o', label=store)
	
	plt.xlabel("Год")
	plt.ylabel("Медианная цена")
	plt.title("Изменение медианной цены по магазинам")

	plt.legend(title="Магазин") plt.grid(True) plt.tight_layout() plt.savefig("item_sales_plot.png") plt.show()
Вывод 10 строк	store cost type city sales date year median_price 19 Ашан 17139 Лампа Санкт-Петербург 8 2023-05-30 2023 19827.0 47 Глобус 44870 Коврик Екатеринбург 19 2023-09-14 2023 36932.0 20 Лемана Про 32756 Лопата Владимир 29 2024-08-28 2024 29053.5 90 Глобус 39384 Лопата Ростов-на-Дону 30 2023-04-02 2023 36932.0 57 Ашан 48874 Бочка Самара 62 2023-12-15 2023 19827.0 58 Ашан 22414 Кровать Уфа 72 2024-06-24 2024 29667.0 91 Лемана Про 20901 Мангал Самара 88 2024-02-21 2024 29053.5 29 Лемана Про 25124 Лампа Волгоград 121 2023-07-04 2023 20678.0 40 Лемана Про 21907 Коврик Санкт-Петербург 124 2023-07-24 2023 20678.0 51 Глобус 14702 Бочка Казань 131 2024-11-26 2024 23392.0

График	

Задача 3.

Студенту высшего учебного заведения предстоит сдать 3 экзамена. Каждый экзамен студент сдаёт с вероятностью 50%.

В случае, если студент не сдаёт экзамен с первого раза, он может попробовать пересдать его два раза, при этом каждый экзамен будет принимать один из трёх преподавателей - Иван, Сергей или Роман, каждый из которых по-разному оценивает работу.

Вероятность того, что студент будет писать экзамен:

у Ивана — 20%;

у Сергея — 30%;

у Романа — 50%.

Вероятность того, что студент успешно сдаст экзамен:

у Ивана — 80%;

у Сергея — 60%;

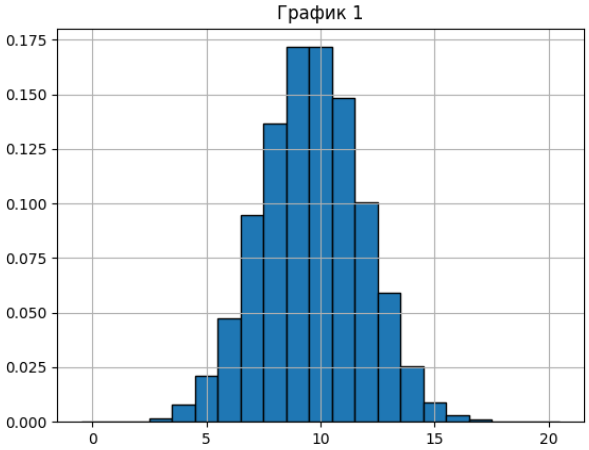
у Романа — 40%.

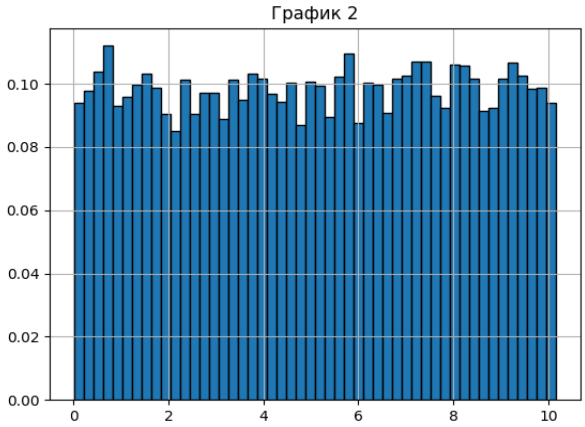
Найдите вероятность того, что студент не сдаст хотя бы один экзамен (то есть, его отчислят).

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 28,5%

Задача 4.





1. Определите тип распределения для каждого графика.

2. Оцените параметры распределения:

- Для графика 1: число испытаний (n) и вероятность успеха (p).
- Для графика 2: укажите минимальное (a) и максимальное (b) значения.

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	1 - Биномиальное, 2 - Равномерное.
2	Для гр.1: n = 20, p ≈ 0.48. Для гр.2: a ≈ 0.0, b ≈ 10.2.
3	Гр.1: дискретные значения, пик около 9.6 — характерно для биномиального распределения. Гр.2: плоская гистограмма, значения в [0.0, 10.2] встречаются равномерно.

Вариант 13

Задача 1.

Вы анализируете поведение клиентов в сети кинотеатров и подготавливаете отчеты для рекламных кампаний.

Вам предоставлены описания таблиц базы данных:

customers (id, name, email) — информация о покупателях;

movies (id, title, genre, release_year) — информация о фильмах;

tickets (id, customer_id, movie_id, purchase_date, price) — информация о проданных билетах.

Вам необходимо:

1. Найти клиентов, которые купили билеты на фильмы, где название содержит «звезда» или начинается на «Вел». Вывести имя, email.

2. Найти фильмы, на которые было куплено более 20 билетов. Вывести название фильма, жанр, количество покупок.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT DISTINCT c.name, c.email FROM customers c JOIN tickets t ON c.id = t.customer_id JOIN movies m ON t.movie_id = m.id WHERE m.title ILIKE '%звезда%' OR m.title ILIKE 'Вел%';
2	SELECT m.title, m.genre, COUNT(t.id) AS ticket_count FROM movies m JOIN tickets t ON m.id = t.movie_id GROUP BY m.id, m.title, m.genre HAVING COUNT(t.id) > 20;

Задача 2.

Вы работаете аналитиком в сети магазинов бытовой техники, и вам нужно проанализировать распределение цен на пылесосы.
У вас есть файл 13_vacuums.csv, содержащий информацию о модели, магазине, цене, объеме пылесборника, рейтинге и количестве продаж.

Поле	Описание
model	Модель
store	Магазин
cost	Цена

volume	Объем пылесборника
rating	Рейтинг (1-5)
sales	Кол-во проданных единиц

Требуется:

Загрузить данные из файла 13_vacuums.csv.

Добавить столбец price_segment:

"Дешевые" (< 10000 руб)

"Средние" (10000–50000 руб)

"Дорогие" (> 50000 руб)

Итоговая таблица должна содержать столбцы (в указанном порядке): model, store, cost, volume, rating, sales, price_segment

Вывести первые 10 строк, отсортировав по убыванию поля sales.

Построить гистограмму цен с разными цветами по категориям:

Ось X — Цена

Ось Y — Количество товаров

Подпишите оси, добавьте заголовок, выведите легенду.

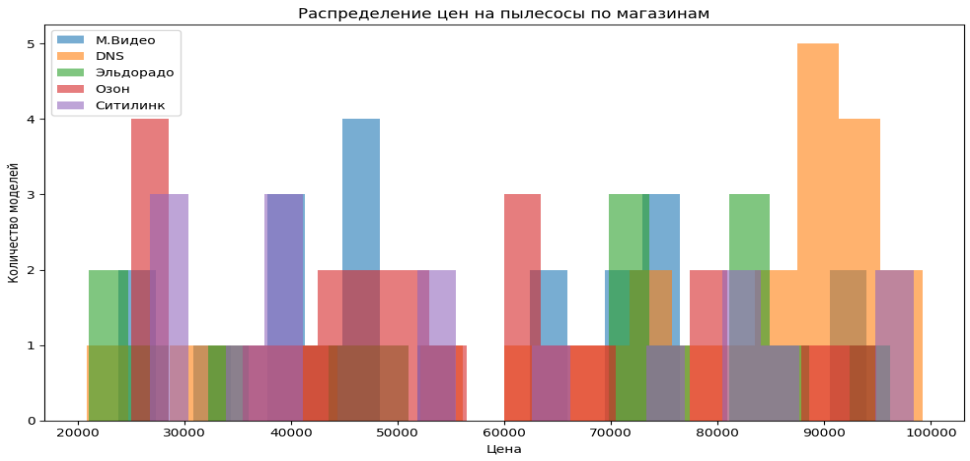
Сохраните изображение как vacuums_price_hist.png.

Запишите решение в файл task_2.py.

Решение	import pandas as pd
	import matplotlib.pyplot as plt
	
	df = pd.read_csv("13_vacuums.csv")
	
	def segment_price(price):
	  if price < 10000:
	  return "Дешевые"
	  elif 10000 <= price <= 50000:

	return "Средние"
	else:
	return "Дорогие"
	
	df["price_segment"] = df["cost"].apply(segment_price)
	
	df = df[["model", "store", "cost", "volume", "rating", "sales", "price_segment"]]
	
	df_sorted = df.sort_values(by="sales", ascending=False)
	print(df_sorted.head(10))
	
	plt.figure(figsize=(10, 6))
	for store in df["store"].unique():
	plt.hist(df[df["store"] == store]["cost"], bins=20, alpha=0.6, label=store)
	
	plt.xlabel("Цена")
	plt.ylabel("Количество моделей")
	plt.title("Распределение цен на пылесосы по магазинам")
	plt.legend()
	plt.tight_layout()
	plt.savefig("vacuums_price_hist.png")
	plt.show()

Вывод 10 строк	model store cost volume rating sales price_segment 11 Гик Ситилинк 65139 10 4 1954 Дорогие 27 Бизон DNS 89205 16 2 1951 Дорогие 77 Вектор DNS 38777 11 4 1951 Средние

	2 Вега DNS 88877 15 2 1941 Дорогие 13 Мастер Озон 77586 14 3 1937 Дорогие 34 Вега Эльдорадо 32484 10 2 1906 Средние 60 Вега Эльдорадо 23822 12 2 1882 Средние 64 Атлант М.Видео 65164 10 1 1864 Дорогие 89 Гик Озон 91374 8 1 1839 Дорогие 5 Атлант М.Видео 94003 14 3 1819 Дорогие
График	

Задача 3.

Владимиру нужно добраться к школе от своего дома, и для этого он совершает поездку на метро. При этом, он может выбрать один из четырех маршрутов, проходящих через различные линии:

Маршрут 1 — через линии метро 1 и 2,

Маршрут 2 — через линии метро 1 и 3,

Маршрут 3 — через линии метро 2 и 4.

Маршрут 4 — через линии метро 1 и 4.

Вероятности, что он выберет какой-либо маршрут, равны.

На каждой из линий ходят поезда различных моделей, при этом количество поездов различных моделей на одной линии одинаково, и очередность поездов на линии случайна:

Линия 1 — модель А,

Линия 2 — модель Б,

Линия 3 — модели А и В,

Линия 4 — модели Б и Г.

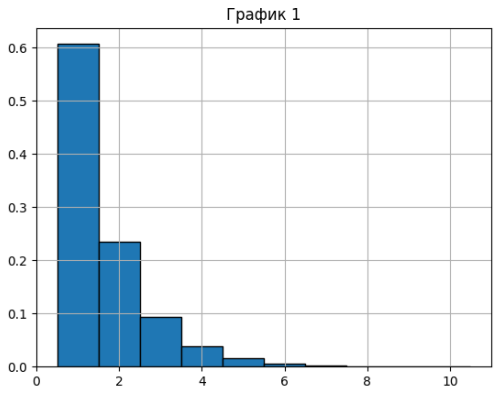
Владимир предпочитает ездить на поездах моделей Б или В, но садится в любой первый попавшийся поезд.

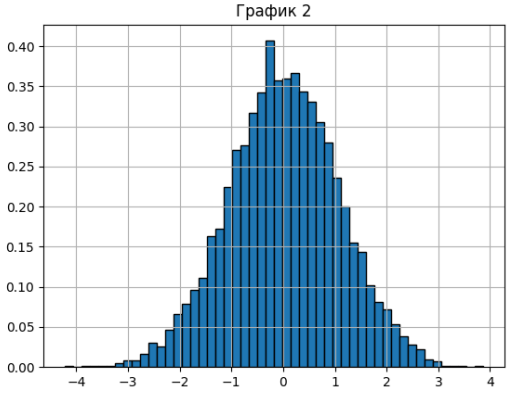
Найдите вероятность того, что во время своей поездки он хотя бы один раз прокатится на поезде своей любимой модели. Ответ округлите до тысячных.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается

Ответ: 0.75

Задача 4.





1. Определите тип распределения для каждого графика.

2. Оцените параметры распределения:

- Для графика 1: укажите вероятность успеха (p).
- Для графика 2: укажите математическое ожидание (μ) и стандартное отклонение (σ).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Геометрическое, Гр.2 - Нормальное.
2	Для гр.1: $p \approx 0.61$, $\sigma \approx 1.04$. Для гр.2: $\mu \approx 0.01$, $\sigma \approx 1.07$.
3	Гр.1: убывающее распределение с пиком на 1, характерно для геометрического распределения. Гр.2: симметричный колоколообразный график с центром около μ — характерно для нормального распределения.

Вариант 14

Задача 1.

Вы работаете с данными о турах и клиентах в туристической компании, чтобы определять постоянных клиентов и выявлять популярные направления для отдыха.

В вашем распоряжении есть следующие описания таблиц:

customers (id, name, surname, patronymic, phone) — информация о клиентах фирмы;

tours (id, destination, price, departure_date) — информация о турах;

bookings (id, client_id, tour_id, booking_date) — информация о бронях.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти информацию о клиентах, чьи имена содержат «ил» или начинаются на «Ма». Вывести имя, фамилию, и номер телефона.

2. Вывести направления, которые бронировались более 5 раз. Показать направление и количество бронирований.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT name, surname, phone FROM customers WHERE name ILIKE '%ил%' OR name ILIKE 'Ма%';

2	SELECT t.destination, COUNT(b.id) AS booking_count FROM tours t JOIN bookings b ON t.id = b.tour_id GROUP BY t.destination HAVING COUNT(b.id) > 5;

Задача 2.

Вы работаете аналитиком в сети заправок. Необходимо проанализировать динамику изменения средней цены на бензин.
У вас есть файл 14_fuel.csv, содержащий информацию о цене, дате, заправке, городе и объеме продаж.

Описание данных:

Поле	Описание
station	Название заправки (Газпром, Роснефть...)
cost	Цена литра (30–90)
type	Тип топлива (АИ-92, АИ-95, ДТ)
city	Город (Москва, Волгоград...)
volume	Объем продаж в литрах (100-10000)
date	Дата (YYYY-MM-DD)

Требуется:

1. Загрузить данные из файла 14_fuel.csv.
2. Добавить столбец year — год (из поля date).
3. Добавить столбец mean_price — средняя цена на каждый тип топлива по годам.

4. Итоговая таблица должна содержать столбцы:
   station, cost, type, city, volume, date, year, mean_price
5. Вывести 10 строк, отсортировав по убыванию volume.
6. Построить график изменения средней цены топлива:
   - Ось X — Год
   - Ось Y — Средняя цена
   - Разные линии — разные типы топлива
   - Добавить заголовок, легенду
   - Сохранить изображение как fuel_price_trend.png
7. Сохраните код в файл task_14.py

Задача 3.

На складе электронной техники 50% устройств — ноутбуки, 30% — смартфоны, 20% — планшеты. Среди них:

60% ноутбуков — с SSD;

40% смартфонов — водонепроницаемые;

50% планшетов — с поддержкой SIM-карты.

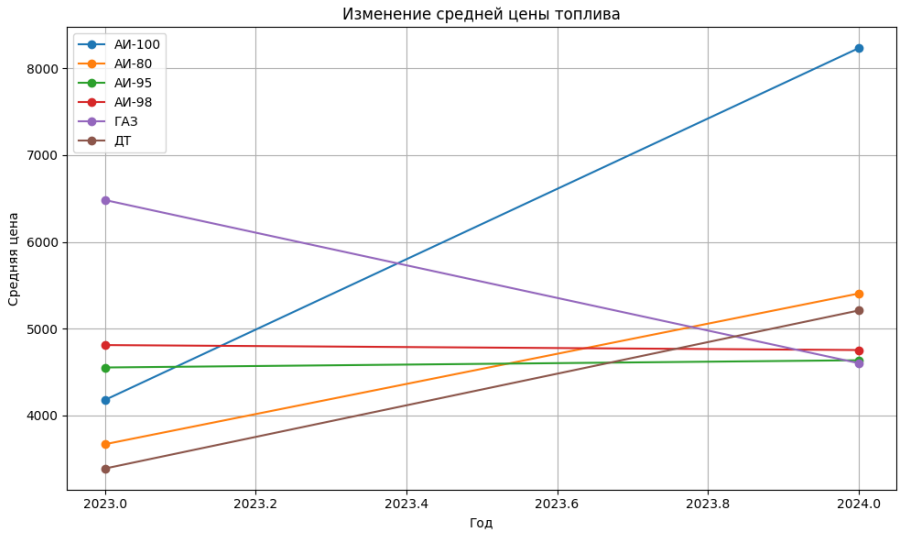
Какова вероятность, что случайно выбранное устройство будет иметь дополнительную функцию?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Решение	import pandas as pd import matplotlib.pyplot as plt df = pd.read_csv("14_fuel.csv") df["date"] = pd.to_datetime(df["date"]) df["year"] = df["date"].dt.year mean_price = df.groupby(["type", "year"])["cost"].mean().reset_index()

	mean_price.rename(columns={"cost": "mean_price"}, inplace=True)
	
	df = pd.merge(df, mean_price, on=["type", "year"])
	df = df[["station", "cost", "type", "city", "volume", "date", "year", "mean_price"]]
	
	df_sorted = df.sort_values(by="volume", ascending=False)
	print(df_sorted.head(10))
	
	plt.figure(figsize=(10, 6))
	for fuel in mean_price["type"].unique():
	  subset = mean_price[mean_price["type"] == fuel]
	  plt.plot(subset["year"], subset["mean_price"], marker="o", label=fuel)
	
	plt.xlabel("Год")
	plt.ylabel("Средняя цена")
	plt.title("Изменение средней цены топлива")
	plt.legend()
	plt.grid(True)
	plt.tight_layout()
	plt.savefig("fuel_price_trend.png")
	plt.show()

Вывод 10 строк	station cost type city volume date year mean_price 61 Шелл 4082 ДТ Москва 39 2024-06-19 2024 5209.142857 40 Шелл 8752 АИ-80 Новгород 39 2023-07-06 2023 3668.000000 60 Ростнефть 2209 ДТ Новгород 39 2023-03-27 2023 3387.875000 76 Ростнефть 8750 АИ-98 Псков 39 2023-02-07 2023 4810.888889 23 Ростнефть 1005 АИ-100 Ростов-на-Дону 38 2023-12-05 2023 4177.636364

	96 Шелл 1226 АИ-95 Самара 38 2023-05-22 2023 4552.625000 88 Лукоил 6150 ДТ Санкт-Петербург 38 2023-10-28 2023 3387.875000 58 Татнефть 2092 АИ-80 Казань 37 2024-11-26 2024 5404.583333 55 Шелл 6310 АИ-95 Самара 37 2023-02-23 2023 4552.625000 90 Роснефть 3412 АИ-95 Мурманск 37 2024-03-08 2024 4635.941176
График	

Задача 3.

На складе электронной техники 50% устройств — ноутбуки, 30% — смартфоны, 20% — планшеты. Среди них:

60% ноутбуков — с SSD;

40% смартфонов — водонепроницаемые;

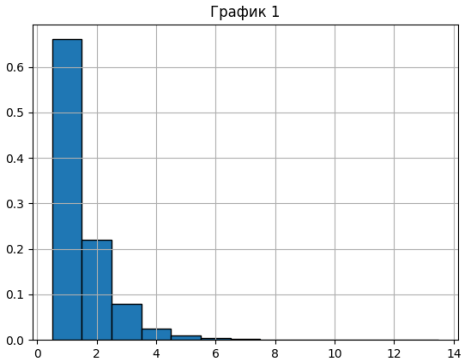
50% планшетов — с поддержкой SIM-карты.

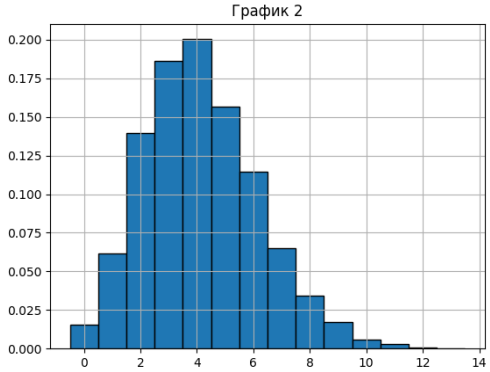
Какова вероятность, что случайно выбранное устройство будет иметь дополнительную функцию?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 52%

Задача 4.





1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:

- Для графика 1: укажите вероятность успеха (p) и стандартное отклонение (σ).
- Для графика 2: укажите среднее число событий (λ).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt

Решение:

1	Гр.1 - Геометрическое, Гр.2 - Пуассона.
2	Для гр.1: $p \approx 0.66$, $\sigma \approx 0.92$. Для гр.2: $\lambda \approx 4.1$.

3	Гр.1: убывающее распределение с пиком на 1, характерно для геометрического распределения. Гр.2: дискретные значения, пик около 4.1, асимметрия вправо — характерно для Пуассона.

Вариант 15

Задача 1.

Вы работаете аналитиком данных в фирме, занимающейся проведением образовательных курсов для школьников. В данный момент в вашем отделе идёт формирование отчетов об активности учащихся и эффективности программы курсов.

В вашем распоряжении есть следующие описания таблиц:

students (id, name, surname, phone, email) — информация об учащихся;

courses (id, title, category) — информация о курсах;

enrollments (id, student_id, course_id, enrollment_date, progress) — информация о прохождении курсов.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти тех учащихся, чьи имена начинаются на «Ел» или содержат «са». Вывести их имя, email.

2. Найти курсы, на которых более 10 учащихся завершили более 80% прогресса. Вывести название курса и количество таких учащихся.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

```
1  SELECT name, email
       FROM students
       WHERE name LIKE 'Ел%' OR name LIKE '%са%';
```

2	SELECT c.title, COUNT(*) AS student_count FROM courses c JOIN enrollments e ON c.id = e.course_id WHERE e.progress > 80 GROUP BY c.id, c.title HAVING COUNT(*) > 10;

Задача 2.

Вы работаете аналитиком в сети автосервисов. Нужно проанализировать распределение цен на услуги по типу. У вас есть файл 15_services.csv.

Описание данных:

Поле	Описание
service	Тип услуги (Замена масла, Диагностика...)
cost	Стоимость (500–20000)
mechanic	Имя механика
city	Город
rating	Оценка клиента (1-5)
visits	Количество посещений (1-100)

Требуется:

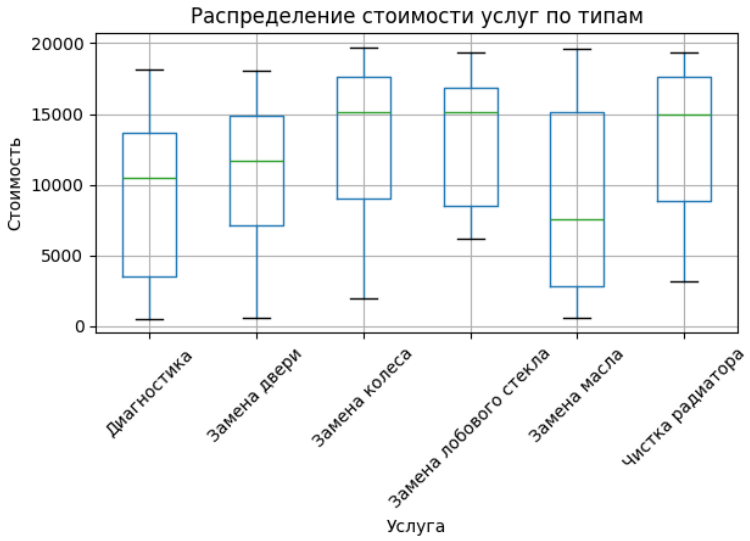
1. Загрузить файл 15_services.csv.
2. Добавить столбец cost_level:
   - "Недорогие" (< 3000)
   - "Средние" (3000–10000)
   - "Дорогие" (> 10000)
3. Итоговая таблица должна содержать поля:
   service, cost, mechanic, city, rating, visits, cost_level
4. Вывести 10 строк, отсортировав по убыванию rating.
5. Построить boxplot цен по типу услуги:
   - X — Услуга
   - Y — Стоимость

- Сохранить как services_boxplot.png

6. Сохраните код в файл task_15.py

Решение	import pandas as pd
	import matplotlib.pyplot as plt
	
	df = pd.read_csv("15_services.csv")
	
	def cost_label(x):
	  if x < 3000:
	  return "Недорогие"
	  elif 3000 <= x <= 10000:
	  return "Средние"
	  else:
	  return "Дорогие"
	
	df["cost_level"] = df["cost"].apply(cost_label)
	
	df = df[["service", "cost", "mechanic", "city", "rating", "visits", "cost_level"]]
	
	df_sorted = df.sort_values(by="rating", ascending=False)
	print(df_sorted.head(10))
	
	plt.figure(figsize=(10, 6))
	df.boxplot(column="cost", by="service", rot=45)
	plt.title("Распределение стоимости услуг по типам")
	plt.suptitle("")

	plt.xlabel("Услуга") plt.ylabel("Стоимость") plt.tight_layout() plt.savefig("services_boxplot.png") plt.show()
Вывод 10 строк	service cost mechanic city rating visits cost_level 52 Замена двери 15418 Борис Санкт-Петербург 4 9 Дорогие 90 Чистка радиатора 15710 Ефим Казань 4 13 Дорогие 75 Диагностика 15589 Алексей Казань 4 19 Дорогие 71 Замена лобового стекла 15018 Фома Мурманск 4 10 Дорогие 23 Замена лобового стекла 15188 Ефим Уфа 4 17 Дорогие 39 Замена двери 6394 Ефим Волгоград 4 9 Средние 21 Диагностика 17110 Ефим Уфа 4 8 Дорогие 83 Чистка радиатора 5063 Фома Ростов-на-Дону 4 9 Средние 84 Замена двери 14704 Борис Волгоград 4 2 Дорогие 64 Замена масла 15201 Алексей Санкт-Петербург 4 9 Дорогие
График	



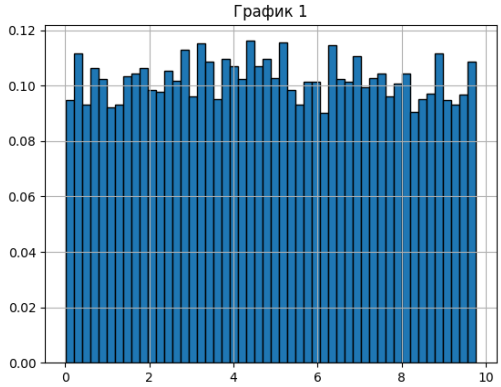
Задача 3.

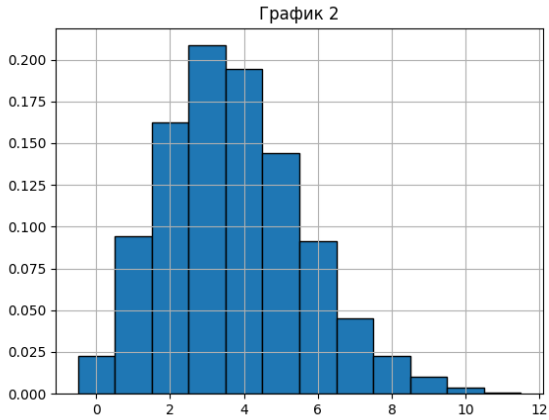
На курсе в университете 80% студентов — очники, 20% — заочники. 70% очников посещают лекции. 40% заочников — не посещают лекции. Какова вероятность, что случайный студент посещает лекции?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0,68

Задача 4.





1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:

- Для графика 1: укажите минимальное (a) и максимальное (b) значения.
- Для графика 2: укажите среднее число событий (λ).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Равномерное, Гр.2 - Пуассона.
2	Для гр.1: $a \approx 0.0$, $b \approx 9.8$. Для гр.2: $\lambda \approx 3.7$.
3	Гр.1: плоская гистограмма, значения в [0.0, 9.8] встречаются равномерно. Гр.2: дискретные значения, пик около 3.7, асимметрия вправо — характерно для Пуассона.

Вариант 16

Задача 1.

Вы работаете аналитиком данных в компании, занимающейся арендой транспортных средств. В данный момент в вашем отделе идёт формирование отчетов об использовании автомобилей в парке компании и о предпочтениях клиентов.

В вашем распоряжении есть следующие описания таблиц:

cars (id, model, brand, manufacture_date, vin)— информация об автомобилях в парке;

clients (id, name, surname, phone, city) — информация о клиентах компании;

rentals (id, car_id, client_id, rental_date, return_date, cost) — информация об оформленных арендах.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти клиентов, имена которых начинаются на «А» или содержат «ик». Вывести имя и город.

2. Найти модели автомобилей, которые сдавались в аренду более 5 раз. Вывести модель и количество аренд.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

```
1  SELECT name, city
       FROM clients
       WHERE name LIKE 'A%' OR name LIKE '%ик%';
```

2	SELECT c.model, COUNT(r.id) AS rental_count FROM rentals r JOIN cars c ON r.car_id = c.id GROUP BY c.model HAVING COUNT(r.id) > 5;

Задача 2.

Вы работаете аналитиком в фитнес-сети. Анализируете связь между возрастом клиента, количеством посещений и потраченной суммой. Вам предоставлен файл 16_clients.csv с данными.

Описание данных:

Поле	Описание
name	Имя клиента
age	Возраст (18–65)
visits	Количество посещений в месяц (1–30)
total_spent	Потрачено (500–10000)
city	Город

Требуется:

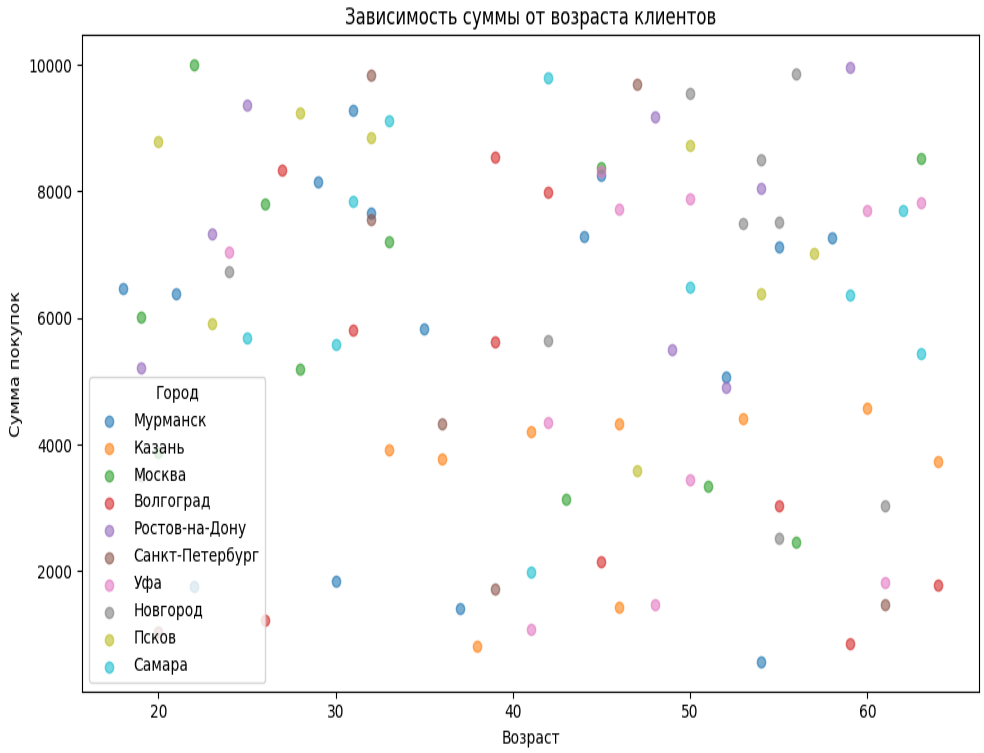
1. Загрузить файл 16_clients.csv.
2. Добавить столбец activity_group:
   - "Редкие" (< 5 посещений)
   - "Средние" (5–15)
   - "Частые" (> 15)

3. Итоговая таблица:
   name, age, visits, total_spent, city, activity_group
4. Вывести 10 строк, отсортировав по убыванию total_spent.
5. Построить scatter plot:
   - X — Возраст
   - Y — Сумма покупок
   - Цвет — город
   - Сохранить как client_spending_scatter.png
6. Сохраните код в файл task_16.py

Решение	import pandas as pd
	import matplotlib.pyplot as plt
	
	df = pd.read_csv("16_clients.csv")
	
	def group_visits(visits):
	  if visits < 5:
	  return "Редкие"
	  elif 5 <= visits <= 15:
	  return "Средние"
	  else:
	  return "Частые"
	
	df["activity_group"] = df["visits"].apply(group_visits)
	
	df = df[["name", "age", "visits", "total_spent", "city", "activity_group"]]
	
	df_sorted = df.sort_values(by="total_spent", ascending=False)
	print(df_sorted.head(10))

	plt.figure(figsize=(10, 6))
	for city in df["city"].unique():
	  subset = df[df["city"] == city]
	  plt.scatter(subset["age"], subset["total_spent"], label=city, alpha=0.6)
	
	plt.xlabel("Возраст")
	plt.ylabel("Сумма покупок")
	plt.title("Зависимость суммы от возраста клиентов")
	plt.legend(title="Город")
	plt.tight_layout()
	plt.savefig("client_spending_scatter.png")
	plt.show()

Вывод 10 строк	name age visits total_spent city activity_group 38 Борис 22 12 9998 Москва Средние 14 Борис 59 3 9961 Ростов-на-Дону Редкие 62 Георгий 56 6 9864 Новгород Средние 89 Денис 32 7 9832 Санкт-Петербург Средние 72 Георгий 42 29 9802 Самара Частые 64 Георгий 47 11 9698 Санкт-Петербург Средние 59 Денис 50 12 9553 Новгород Средние 44 Георгий 25 2 9365 Ростов-на-Дону Редкие 17 Георгий 31 1 9283 Мурманск Редкие 86 Фома 28 23 9239 Псков Частые

График	 Зависимость суммы от возраста клиентов Сумма покупок Возраст 10000 8000 6000 4000 2000 20 30 40 50 60 Город Мурманск Казань Москва Волгоград Ростов-на-Дону Санкт-Петербург Уфа Новгород Псков Самара

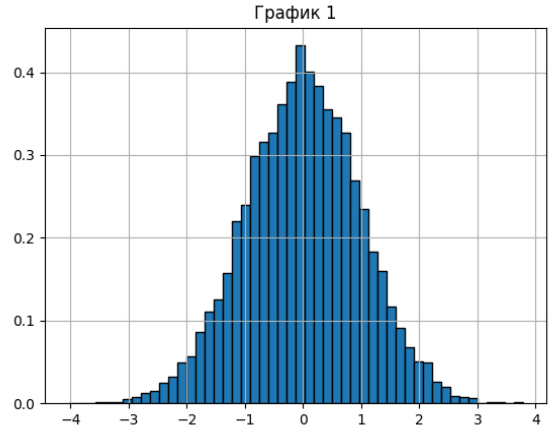
Задача 3.

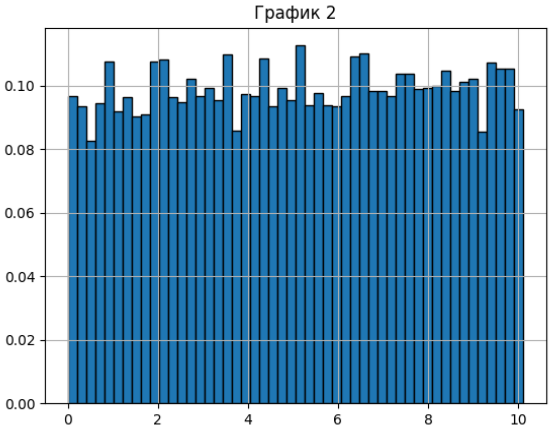
В библиотеке есть 4 книги по математике, 2 книги по информатике и 3 книг по физике. Они расставлены случайным образом на полке в один ряд. Какова вероятность, что книги по ровно одному предмету будут стоять вместе в начале полки (все подряд)?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.0429

Задача 4.





1. Определите тип распределения для каждого графика.

2. Оцените параметры распределения:

- Для графика 1: укажите математическое ожидание (μ) и стандартное отклонение (σ).
- Для графика 2: укажите минимальное (a) и максимальное (b) значения.

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Нормальное, Гр.2 - Равномерное.
2	Для гр.1: $\mu = 0$, $\sigma \approx 1$. Для гр.2: $a \approx 0.0$, $b \approx 10.0$.
3	Гр.1: Непрерывное распределение, симметричная колоколообразная форма — характерно для нормального распределения. Гр.2: плоская гистограмма, значения в [0.0, 10.0] встречаются равномерно.

Вариант 17

Задача 1.

Вы работаете аналитиком данных в интернет-магазине, который специализируется на продаже книг. В данный момент в магазине идёт формирование отчетов о предпочтениях клиентов.

В вашем распоряжении есть следующие описания таблиц:

users (id, name, surname, phone, email)— информация о покупателях;

books (id, title, author, genre) — информация о книгах;

orders (id, user_id, book_id, order_date, price) — информация об оформленных заказах.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти пользователей, имена которых начинаются на «Нат» или содержат «ми». Вывести имя, email.

2. Найти жанры литературы, по которым было продано более 15 книг. Вывести жанр и количество продаж.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT name, email FROM users WHERE name LIKE 'Нат%' OR name LIKE '%ми%';
2	SELECT b.genre, COUNT(*) AS sales_count FROM orders o JOIN books b ON o.book_id = b.id GROUP BY b.genre HAVING COUNT(*) > 15;

Задача 2.

Вы аналитик в сети книжных магазинов. Нужно сравнить количество проданных книг по жанрам.
Файл: 17_books.csv

Описание данных:

Поле	Описание
title	Название книги
genre	Жанр (Фантастика, Детектив...)
cost	Цена (100–2000)
author	Автор
store	Магазин

Поле	Описание
sold	Кол-во проданных экземпляров (1–1000)

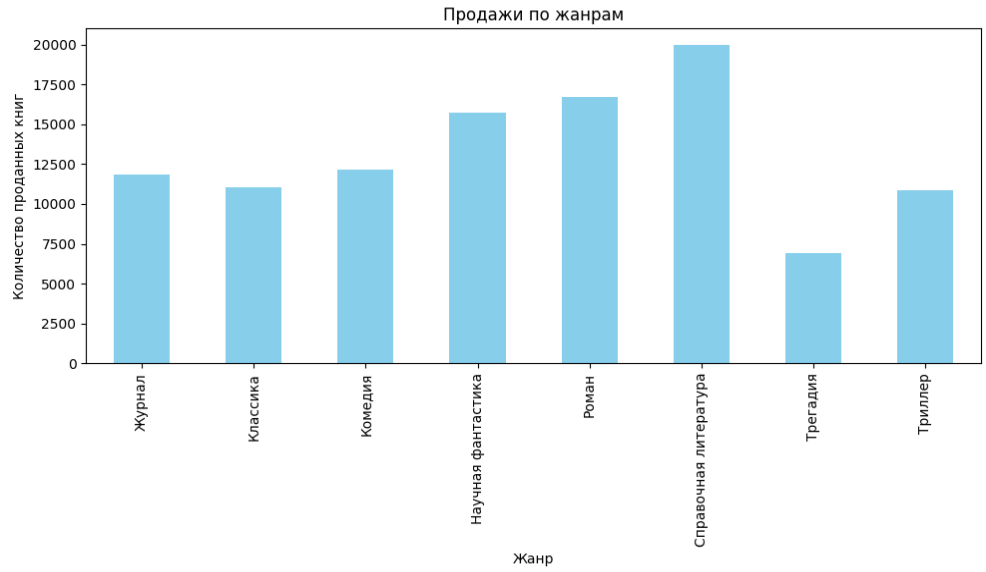
Требуется:

1. Загрузить 17_books.csv.
2. Добавить столбец price_tier:
   - "Бюджетные" (< 500)
   - "Средние" (500–1500)
   - "Дорогие" (> 1500)
3. Итоговая таблица:
   title, genre, cost, author, store, sold, price_tier
4. Вывести 10 строк, отсортировав по убыванию sold.
5. Построить столбчатую диаграмму по жанрам:
   - X — Жанр
   - Y — Общее количество продаж
   - Сохранить как books_genre_bar.png
6. Сохраните код в файл task_17.py

Решение

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv("17_books.csv")


def tier(price):
    if price < 500:
        return "Бюджетные"
    elif 500 <= price <= 1500:
        return "Средние"
    else:
        return "Дорогие"
```

	df["price_tier"] = df["cost"].apply(tier) df = df[["title", "genre", "cost", "author", "store", "sold", "price_tier"]] df_sorted = df.sort_values(by="sold", ascending=False) print(df_sorted.head(10)) sales_by_genre = df.groupby("genre")["sold"].sum() plt.figure(figsize=(10, 6)) sales_by_genre.plot(kind="bar", color="skyblue") plt.xlabel("Жанр") plt.ylabel("Количество проданных книг") plt.title("Продажи по жанрам") plt.tight_layout() plt.savefig("books_genre_bar.png") plt.show()
Вывод 10 строк	title genre cost author store sold price_tier 82 З Классика 705 Ефим Твоя Книга 1966 Средние 18 К Справочная литература 631 Фома Читай-город 1965 Средние 93 З Научная фантастика 1505 Константин Твоя Книга 1964 Дорогие 51 Г Комедия 1853 Фома Читай-город 1925 Дорогие 33 Л Справочная литература 809 Алексей Букстор 1923 Средние 27 Б Научная фантастика 1391 Алексей Читай-город 1921 Средние 87 И Роман 786 Денис Букстор 1907 Средние

	98 Л Журнал 1030 Владимир Мир Книг 1864 Средние 70 Ж Роман 774 Георгий Твоя Книга 1812 Средние 14 Б Классика 1657 Борис Твоя Книга 1797 Дорогие
График	

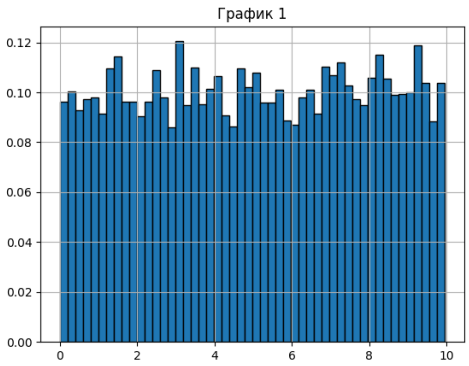
Задача 3.

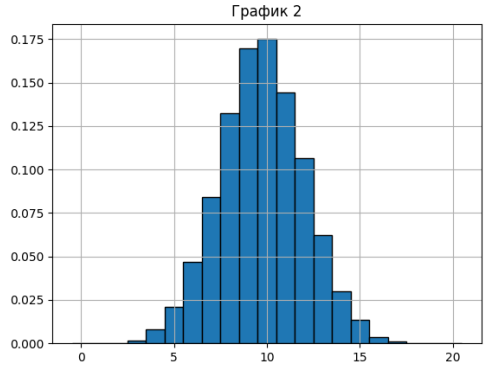
На вечеринке в рамках конкурса участники предлагают сыграть в 2 карточные игры, 4 настольные игры, 3 устные игры и 2 спортивные игры. Случайным образом посредством жребия выбирается 3 игры. Найдите вероятность, что среди выбранных будет хотя бы одна спортивная игра.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.4909

Задача 4.





1. Определите тип распределения для каждого графика.

2. Оцените параметры распределения:

- Для графика 1: укажите минимальное (a) и максимальное (b) значения.
- Для графика 2: число испытаний (n) и вероятность успеха (p).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Равномерное, Гр.2 - Биномиальное.
2	Для гр.1: a ≈ 0.0, b ≈ 10.0. Для гр.2: n = 20, p ≈ 0.49.
3	Гр.1: плоская гистограмма, значения в [0.0, 10.0] встречаются равномерно. Гр.2: дискретные значения, пик около 9.7 — характерно для биномиального распределения.

Вариант 18

Задача 1.

Вы работаете аналитиком данных в частном спортивном зале, предлагающий услуги тренеров. В данный момент в вашем отделе идёт формирование отчетов о предпочтениях клиентов.

В вашем распоряжении есть следующие описания таблиц:

athletes (id, name, surname, phone, sport)— информация о посетителях спортзала;

coaches (id, name, surname, specialization)— информация о тренерах;

sessions (id, athlete_id, coach_id, date, duration)— информация о проведенных занятиях.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти спортсменов, имена которых начинаются на «Д» или содержат «ен». Вывести имя, вид спорта.

2. Найти тренеров, которые провели в среднем более 2 часов тренировок. Вывести имя, специализацию, среднюю продолжительность.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT name, sport FROM athletes WHERE name LIKE 'Д%' OR name LIKE '%ен%';
2	SELECT c.name, c.specialization, AVG(s.duration) AS avg_duration FROM coaches c JOIN sessions s ON c.id = s.coach_id GROUP BY c.id, c.name, c.specialization HAVING AVG(s.duration) > 2;

Задача 2.

Вы — аналитик в магазине электроники. Необходимо исследовать распределение продаж планшетов. Файл: 18_tablets.csv

Описание данных:

Поле	Описание
model	Модель планшета
price	Цена за штуку (5000–90000)
sold	Количество проданных устройств
brand	Бренд (Samsung, Apple, Xiaomi...)

Поле	Описание
city	Город

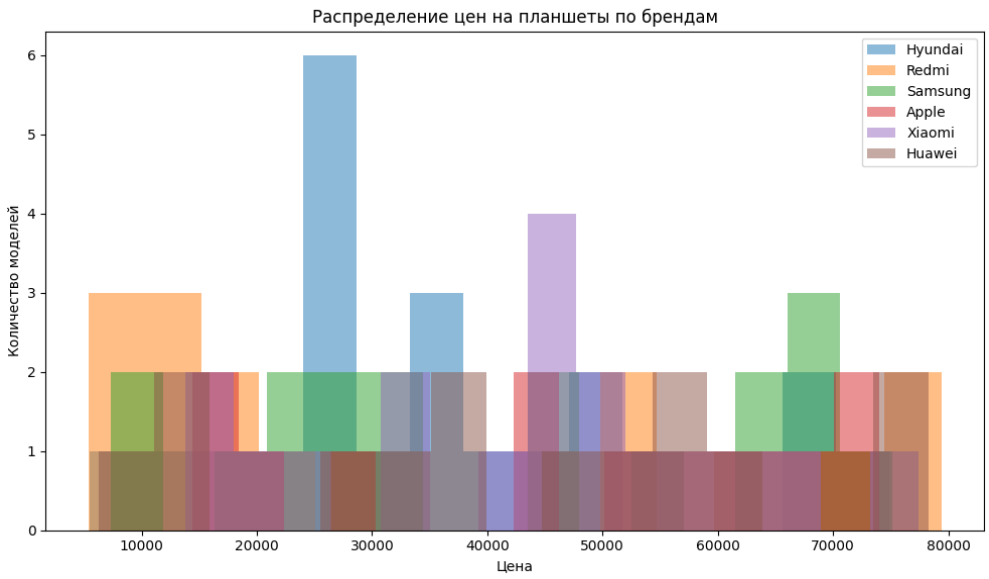
Требуется:

1. Загрузить данные из 18_tablets.csv
2. Добавить столбец revenue = price * sold
3. Добавить столбец price_segment:
   - "<20000" - Бюджетные
   - "20000-50000" - Средние
   - ">50000" - Флагманские
4. Итоговая таблица: model, price, sold, brand, city, revenue, price_segment
5. Вывести 10 строк, отсортированных по убыванию revenue
6. Построить гистограмму цен с разделением по брендам:
   - Сохранить как tablets_histogram.png
7. Сохраните код в файл task_17.py

Решение

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv("18_tablets.csv")


df["revenue"] = df["price"] * df["sold"]


def segment(p):
    if p < 20000:
        return "Бюджетные"
    elif p <= 50000:
        return "Средние"
    else:
        return "Флагманские"
```

	df["price_segment"] = df["price"].apply(segment) df = df[["model", "price", "sold", "brand", "city", "revenue", "price_segment"]] df_sorted = df.sort_values(by="revenue", ascending=False) print(df_sorted.head(10)) plt.figure(figsize=(10, 6)) for brand in df["brand"].unique(): plt.hist(df[df["brand"] == brand]["price"], bins=15, alpha=0.5, label=brand) plt.xlabel("Цена") plt.ylabel("Количество моделей") plt.title("Распределение цен на планшеты по брендам") plt.legend() plt.tight_layout() plt.savefig("tablets_histogram.png") plt.show()
Вывод 10 строк	model price sold brand city revenue price_segment 94 A 74908 1996 Hyundai Новгород 149516368 Флагманские 44 E 66908 1882 Redmi Казань 125920856 Флагманские 4 Ж 77433 1593 Xiaomi Псков 123350769 Флагманские 86 Д 62439 1973 Samsung Волгоград 123192147 Флагманские 88 Б 66197 1845 Xiaomi Волгоград 122133465 Флагманские 9 Н 61354 1921 Apple Самара 117861034 Флагманские

	22 A 79396 1394 Redmi Уфа 110678024 Флагманские 82 Г 74037 1444 Apple Самара 106909428 Флагманские 60 Д 66089 1612 Hyundai Волгоград 106535468 Флагманские 68 Е 52997 1835 Redmi Уфа 97249495 Флагманские
График	

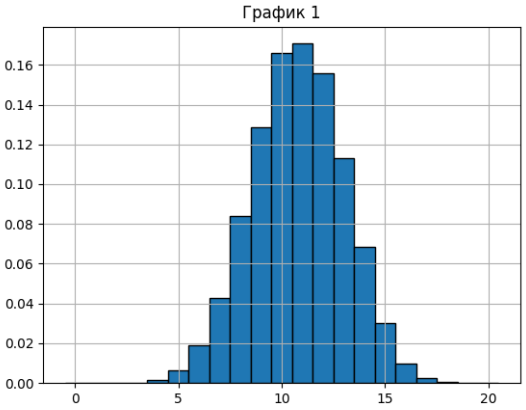
Задача 3.

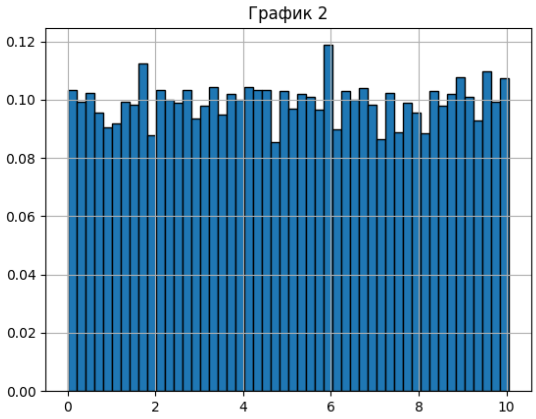
В команде разработчиков коммерческого проекта 10 человек: 4 программиста, 3 дизайнера и 3 тестировщика. Нужно выбрать 4-х участников проекта. Какова вероятность, что в команде после случайного распределения окажется хотя бы один тестировщик? Ответ округлить до сотых.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0,833

Задача 4.





1. Определите тип распределения для каждого графика.

2. Оцените параметры распределения:

- Для графика 1: число испытаний (n) и вероятность успеха (p).
- Для графика 2: укажите минимальное (a) и максимальное (b) значения.

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Биномиальное, Гр.2 - Пуассона.
2	Для гр.1: n = 20, p ≈ 0.54. Для гр.2: a = 0, b = 10.
3	Гр.1: дискретные значения, пик около 10.8 — характерно для биномиального распределения. Гр.2: плоская гистограмма, значения в [0.0, 10.0] встречаются равномерно.

Вариант 19

Задача 1.

Вы работаете аналитиком данных в благотворительном фонде. В данный момент в фонде идёт формирование отчетов об эффективности работы фонда.

В вашем распоряжении есть следующие описания таблиц:

donors (id, name, surname, patronymic, city)— информация о благотворителях;

campaigns (id, title, start_date, end_date)— информация о кампаниях;

donations (id, donor_id, campaign_id, amount, donation_date)— информация о пожертвованиях.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти благотворителей, имена которых начинаются на «Вла» или содержат «ри». Вывести имя и город.

2. Найти кампании, которые собрали более 100.000 рублей суммарно. Вывести название и сумму.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT name, city FROM donors WHERE name LIKE 'Вла%' OR name LIKE '%ри%';
2	SELECT c.title, SUM(d.amount) AS total_amount FROM campaigns c JOIN donations d ON c.id = d.campaign_id GROUP BY c.id, c.title HAVING SUM(d.amount) > 100000;

Задача 2.

Вы — аналитик в онлайн-сервисе просмотра видео. Необходимо исследовать распределение пользователей по городам. Вам предоставлен файл: 19_users.csv — база подписчиков стримингового сервиса.

Описание данных:

Поле	Описание
user_id	Уникальный ID
age	Возраст (12–75)

Поле	Описание
city	Город
watch_time	Кол-во часов в неделю

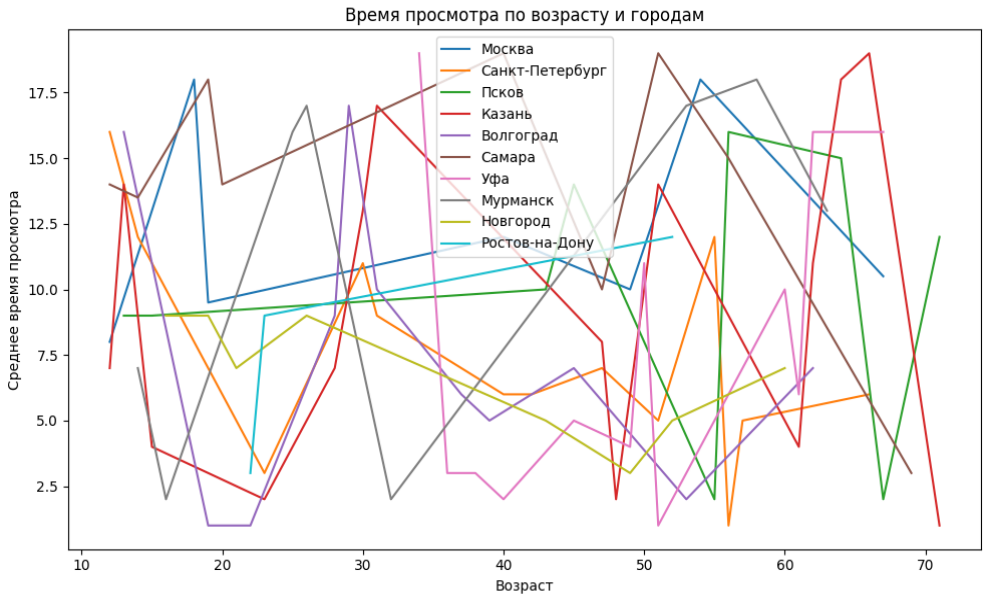
Требуется:

1. Загрузить файл 19_users.csv
2. Добавить столбец age_group:
   - "<18"
   - "18–35"
   - "36–60"
   - ">60"
3. Добавить столбец intensity:
   - "Низкая" (<5 ч)
   - "Средняя" (5–15 ч)
   - "Высокая" (>15 ч)
4. Итоговая таблица: user_id, age, city, watch_time, age_group, intensity
5. Вывести 10 строк, отсортированных по убыванию watch_time
6. Построить линейный график:
   - X — возраст
   - Y — среднее время просмотра
   - Цвет — город
   - Сохранить как user_watch_city.png
7. Сохраните код в файл task_2.py.

Решение	import pandas as pd
	import matplotlib.pyplot as plt
	
	df = pd.read_csv("19_users.csv")
	
	def age_group(a):
	  if a < 18:
	  return "<18"
	  elif a <= 35:
	  return "18–35"

```
    elif a <= 60:
        return "36–60"
    else:
        return ">60"

def watch_level(h):
    if h < 5:
        return "Низкая"
    elif h <= 15:
        return "Средняя"
    else:
        return "Высокая"

df["age_group"] = df["age"].apply(age_group)
df["intensity"] = df["watch_time"].apply(watch_level)

df = df[["user_id", "age", "city", "watch_time", "age_group", "intensity"]]
df_sorted = df.sort_values(by="watch_time", ascending=False)
print(df_sorted.head(10))

plt.figure(figsize=(10, 6))
for city in df["city"].unique():
    grouped = df[df["city"] == city].groupby("age")["watch_time"].mean()
    plt.plot(grouped.index, grouped.values, label=city)

plt.xlabel("Возраст")
plt.ylabel("Среднее время просмотра")
```

	plt.title("Время просмотра по возрасту и городам") plt.legend() plt.tight_layout() plt.savefig("user_watch_city.png") plt.show()
Вывод 10 строк	user_id age city watch_time age_group intensity 25 37 40 Самара 19 36–60 Высокая 45 75 66 Казань 19 >60 Высокая 26 30 51 Самара 19 36–60 Высокая 86 59 34 Уфа 19 18–35 Высокая 0 8 18 Москва 18 18–35 Высокая 72 38 19 Самара 18 18–35 Высокая 15 37 64 Казань 18 >60 Высокая 27 25 58 Мурманск 18 36–60 Высокая 64 46 54 Москва 18 36–60 Высокая 1 44 55 Санкт-Петербург 18 36–60 Высокая

График	

Задача 3.

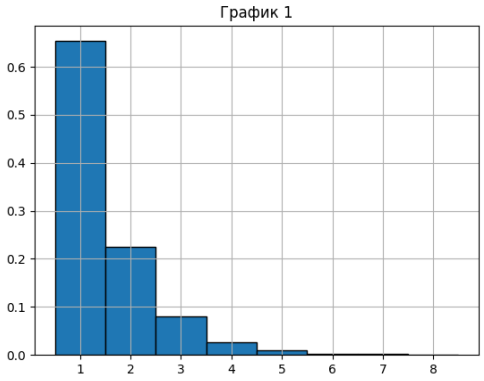
В самолете есть 24 ряда по 6 мест. За дополнительную плату пассажир может выбрать любое доступное место, а для пассажиров, от этой услуги отказавшихся, распределение мест происходит случайным образом. Какова вероятность, что пассажир, место которому досталось случайным образом, сядет у окна, если количество пассажиров,

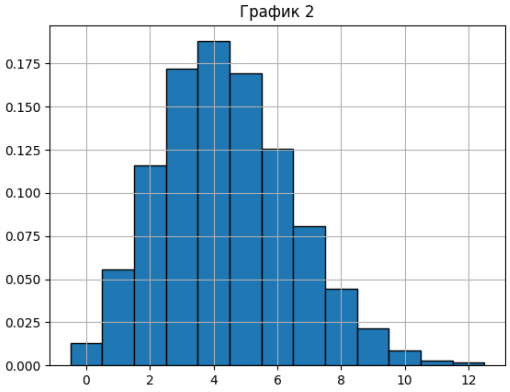
самостоятельно выбравших место, равно 24, а вероятность того, что каждый из этих пассажиров выберет место у окна, равно 75%?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.25

Задача 4.





1. Определите тип распределения для каждого графика.

2. Оцените параметры распределения:

- Для графика 1: укажите вероятность успеха (p) и стандартное отклонение (σ).
- Для графика 2: укажите среднее число событий (λ).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Геометрическое, Гр.2 - Пуассона.

2	Для гр.1: $p \approx 0.66$, $\sigma \approx 0.89$. Для гр.2: $\lambda \approx 4.4$.
3	Гр.1: убывающее распределение с пиком на 1, характерно для геометрического распределения. Гр.2: дискретные значения, пик около 4.4, асимметрия вправо — характерно для Пуассона.

Вариант 20

Задача 1.

Вы работаете бизнес-аналитиком в сети кофеен «Кофеман». В данный момент в компании идёт формирование отчетов об эффективности работы сети и предпочтениях клиентов.

В вашем распоряжении есть следующие описания таблиц:

customers (id, name, phone, email)— информация о клиентах;

branches (id, city, branch_name) — информация о филиалах сети;

orders (id, customer_id, branch_id, order_date, total_amount)— информация о заказах.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти клиентов, чьи имена начинаются на «Лю» или содержат «на». Вывести имя, email.

2. Найти города, в которых сделано более 20 заказов. Вывести город и количество заказов.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT name, email FROM customers WHERE name LIKE 'Лю%' OR name LIKE '%на%';
2	SELECT b.city, COUNT(o.id) AS orders_count FROM orders o JOIN branches b ON o.branch_id = b.id GROUP BY b.city HAVING COUNT(o.id) > 20;

Задача 2.

Вы работаете аналитиком в университете. В данный момент в учреждении идёт формирование отчетов об эффективности работы студентов. Предоставленный вам файл 20_students.csv содержит информацию об успеваемости студентов по экзамену.

Описание данных:

Поле	Описание
student_id	Уникальный номер студента
score	Балл за экзамен (от 0 до 100)

Поле	Описание
subject	Предмет (Математика, Физика...)
group	Учебная группа (A1, A2, B1...)
date	Дата проведения экзамена

Требуется:

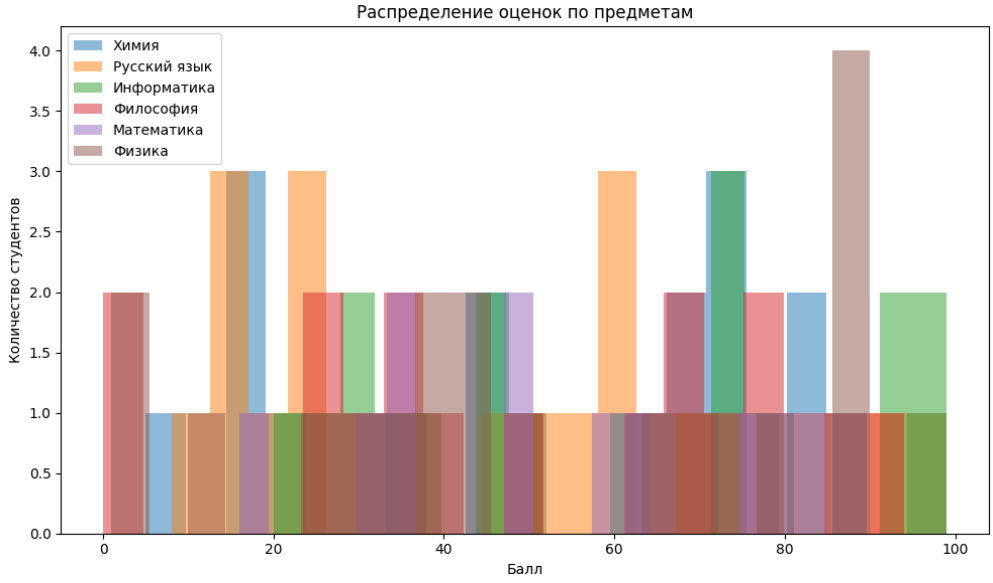
1. Загрузить данные из файла 20_students.csv.
2. Добавить столбец grade_level, в зависимости от значения score:
   - "<60"
   - "60–79"
   - "80–89"
   - "90–100"
3. Итоговая таблица должна включать поля: student_id, score, subject, group, date, grade_level
4. Вывести 10 строк с наивысшими оценками (по убыванию score)
5. Построить гистограмму оценок:
   - Ось X — баллы
   - Ось Y — количество студентов
   - Цвета по предмету (subject)
   - Сохранить график как students_scores_hist.png
6. Сохраните код в файл task_2.py.

Решение	

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv("20_students.csv")


def grade(score):
    if score < 60:
        return "<60"
    elif score <= 79:
        return "60–79"
    elif score <= 89:
```

	return "80–89"
	else:
	return "90–100"
	
	df["grade_level"] = df["score"].apply(grade)
	
	df = df[["student_id", "score", "subject", "group", "date", "grade_level"]]
	df_sorted = df.sort_values(by="score", ascending=False)
	print(df_sorted.head(10))
	
	plt.figure(figsize=(10, 6))
	for subject in df["subject"].unique():
	plt.hist(df[df["subject"] == subject]["score"], bins=20, alpha=0.5, label=subject)
	
	plt.xlabel("Балл")
	plt.ylabel("Количество студентов")
	plt.title("Распределение оценок по предметам")
	plt.legend()
	plt.tight_layout()
	plt.savefig("students_scores_hist.png")
	plt.show()

Вывод 10 строк	student_id score subject group date grade_level 29 84 99 Информатика B3 2023-12-16 90–100 75 54 99 Русский язык C2 2023-04-26 90–100 14 72 99 Химия B1 2023-10-21 90–100

	89 89 96 Информатика A3 2024-05-02 90–100 86 48 94 Философия B1 2024-02-13 90–100 24 50 94 Русский язык A2 2023-10-20 90–100 84 43 92 Информатика A3 2023-01-29 90–100 13 34 92 Информатика A1 2023-07-21 90–100 41 39 90 Физика A1 2023-09-08 90–100 50 87 89 Химия A1 2023-05-11 80–89
График	

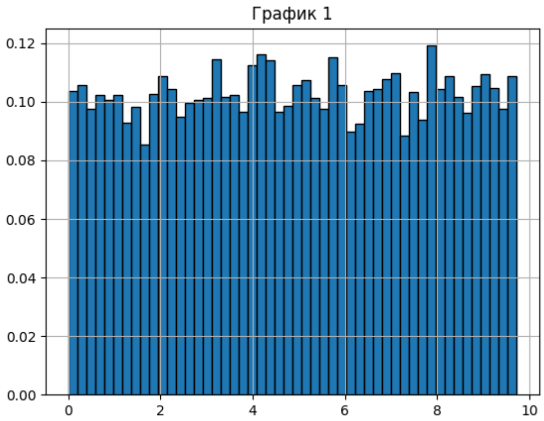
Задача 3.

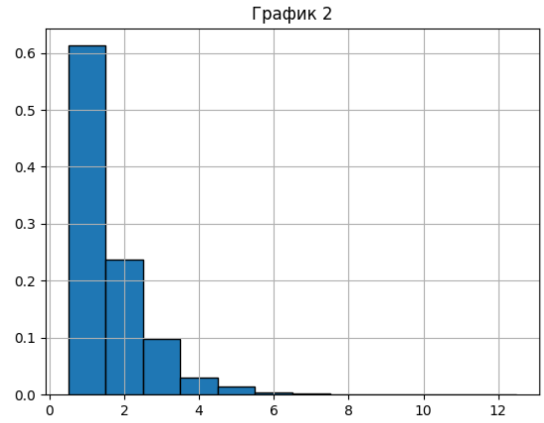
В коробке лежит 6 синих, 5 белых, 4 жёлтых кубиков равного размера и одинаковых на ощупь. Случайным образом из коробки достают 4 кубика. Какова вероятность, что среди них будет ровно 2 белых? Ответ округлить до сотых.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0,33

Задача 4.





1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:

- Для графика 1: укажите минимальное (a) и максимальное (b) значения.
- Для графика 2: укажите вероятность успеха (p) и стандартное отклонение (σ).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Равномерное, Гр.2 - Геометрическое.
2	Для гр.1: a ≈ 0.0, b ≈ 9.7. Для гр.2: p ≈ 0.62, σ ≈ 1.01.
3	Гр.1: плоская гистограмма, значения в [0.0, 9.7] встречаются равномерно. Гр.2: убывающее распределение с пиком на 1, характерно для геометрического распределения.

Вариант 21

Задача 1.

Вы работаете бизнес-аналитиком в частной сети медицинских клиник. В данный момент в сети идёт формирование отчетов об эффективности работы сети и предпочтениях пациентов.

В вашем распоряжении есть следующие описания таблиц:

patients (id, name, city, birth_date) — информация о пациентах;

doctors (id, name, specialization) — информация о врачах;

appointments (id, patient_id, doctor_id, appointment_date) — записи клентов ко врачам;

diagnoses (id, appointment_id, diagnosis_code, severity) — информация о диагнозах.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти пациентов, имена которых начинаются на «Ал» или содержат «ей». Вывести имя, город, дату рождения.

2. Найти врачей, у которых было более 15 приёмов. Вывести имя врача, специализацию, количество приёмов.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT name, city, birth_date FROM patients WHERE name LIKE 'Ал%' OR name LIKE '%ей%';
2	SELECT d.name, d.specialization, COUNT(a.id) AS appointment_count FROM doctors d JOIN appointments a ON d.id = a.doctor_id GROUP BY d.id, d.name, d.specialization HAVING COUNT(a.id) > 15;

Задача 2.

Вы — аналитик в онлайн-магазине по продаже книг. Необходимо исследовать распределение продаж книг. В предоставленном вам файле 21_books.csv содержатся данные о продажах книг.

Описание данных:

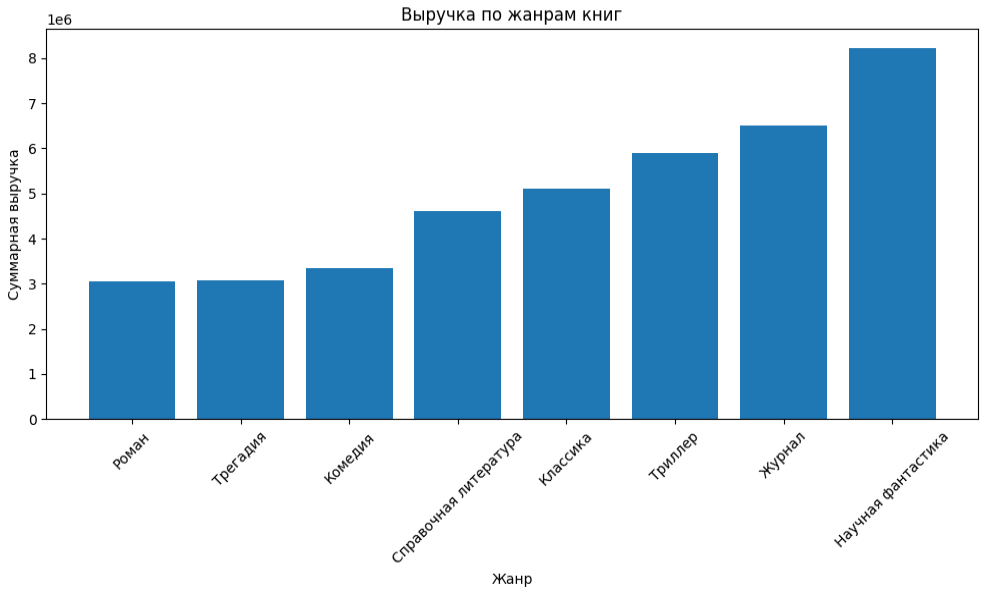
Поле	Описание
book_id	Уникальный код книги
title	Название книги
genre	Жанр (Фантастика, Детектив...)
price	Цена книги (150–3000)
copies	Продано экземпляров (0–500)
store	Название магазина

Требуется:

1. Загрузить данные из 21_books.csv
2. Добавить столбец revenue = price * copies
3. Добавить столбец price_range, в котором цены должны быть распределены на следующие промежутки:
   - "<500"
   - "500–1500"
   - ">1500"
4. Итоговая таблица: book_id, title, genre, price, copies, store, revenue, price_range
5. Вывести 10 строк с наибольшей выручкой
6. Построить **столбчатую диаграмму**:
   - Ось X — жанры (genre)
   - Ось Y — суммарная выручка по жанрам
7. Сохранить как books_revenue_bar.png

Решение	import pandas as pd
	import matplotlib.pyplot as plt
	
	df = pd.read_csv("21_books.csv")
	
	df["revenue"] = df["price"] * df["copies"]
	
	def price_range(p):
	  if p < 500:
	  return "<500"
	  elif p <= 1500:
	  return "500–1500"
	  else:
	  return ">1500"
	
	df["price_range"] = df["price"].apply(price_range)

	df = df[["book_id", "title", "genre", "price", "copies", "store", "revenue", "price_range"]] df_sorted = df.sort_values(by="revenue", ascending=False) print(df_sorted.head(10)) genre_revenue = df.groupby("genre")["revenue"].sum().sort_values() plt.figure(figsize=(10, 6)) plt.bar(genre_revenue.index, genre_revenue.values) plt.xlabel("Жанр") plt.ylabel("Суммарная выручка") plt.title("Выручка по жанрам книг") plt.xticks(rotation=45) plt.tight_layout() plt.savefig("books_revenue_bar.png") plt.show()
Вывод 10 строк	**book_id title genre price copies store revenue price_range** **81 85 Д Научная фантастика 2816 498 КнигМаг 1402368 >1500** **91 17 Н Роман 2785 481 МояКнига 1339585 >1500** **85 73 О Классика 2950 437 МояКнига 1289150 >1500** **5 17 Б Журнал 2704 471 ЛучшаяКнига 1273584 >1500** **2 65 А Триллер 2840 425 МояКнига 1207000 >1500** **9 64 Д Комедия 2652 421 БукСтор 1116492 >1500** **87 37 Г Триллер 2376 464 КнигМаг 1102464 >1500** **26 47 И Справочная литература 2889 370 КнигМаг 1068930 >1500** **31 89 И Журнал 2456 415 КнигМаг 1019240 >1500** **89 75 Д Журнал 2195 460 БукСтор 1009700 >1500**

График	

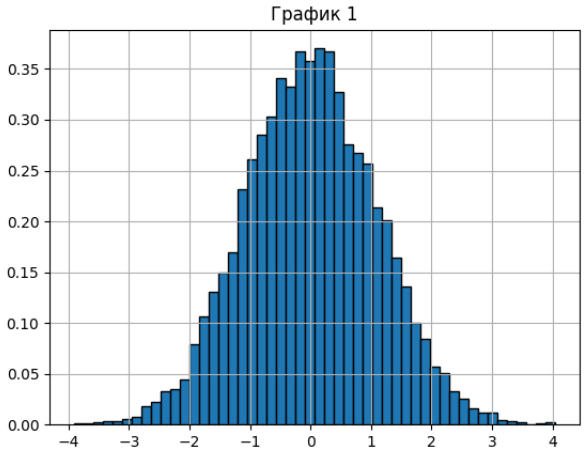
Задача 3.

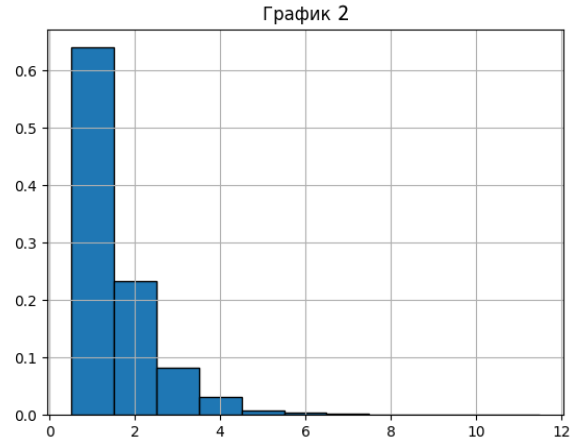
В лаборатории работают 12 учёных: 6 физиков, 4 биолога, 2 химика. Нужно выбрать команду из 5 человек. Какова вероятность, что после случайного выбора сотрудников в команде окажутся представители всех трёх специальностей? Ответ округлить до тысячных.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0,611

Задача 4.





1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:

Для графика 1: укажите математическое ожидание (μ) и стандартное отклонение (σ).

Для графика 2: укажите вероятность успеха (p) и стандартное отклонение (σ).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Нормальное, Гр.2 - Геометрическое.
2	Для гр.1: $\mu \approx 0.01$, $\sigma \approx 1.09$. Для гр.2: $p \approx 0.65$, $\sigma \approx 0.92$.
3	Гр.1: симметричный колоколообразный график с центром около μ — характерно для нормального распределения. Гр.2: убывающее распределение с пиком на 1, характерно для геометрического распределения.

Вариант 22

Задача 1.

Вы работаете бизнес-аналитиком в компании, организующая корпоративные мероприятия. В данный момент в вашем отделе идёт формирование отчетов об эффективности работы компании и предпочтениях корпоративных клиентов.

В вашем распоряжении есть следующие описания таблиц:

participants (id, name, email) — информация об участниках мероприятий;

events (id, title, theme, event_date) — информация о мероприятиях;

speakers (id, name, topic) — информация о ведущих, участвующих в мероприятии;

registrations (id, participant_id, event_id, speaker_id) — информация о регистрациях.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти участников, имена которых начинаются на «Ма» или содержат «ил». Вывести имя и email.

2. Определить темы мероприятий, на которые зарегистрировались более 30 участников. Вывести тему и количество регистраций.

В качестве ответа необходимо предоставить два sql-запроса.
Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT name, email FROM participants WHERE name LIKE 'Ма%' OR name LIKE '%ил%';
2	SELECT e.theme, COUNT(r.id) AS registrations_count FROM events e JOIN registrations r ON e.id = r.event_id GROUP BY e.theme HAVING COUNT(r.id) > 30;

Задача 2.

Вы работаете аналитиком в компании, занимающейся продажей смартфонов. Вам необходимо проанализировать данные о продажах по регионам и выявить распределение цен.

У вас есть файл 22_phones.csv, содержащий информацию о моделях смартфонов, цене, объёме продаж и магазинах.

Описание данных:

Поле	Описание
model	Модель смартфона (iPhone 13, Galaxy A50...)
brand	Бренд (Apple, Samsung, Xiaomi...)

Поле	Описание
price	Цена устройства (от 10 000 до 150 000)
region	Регион продаж (Москва, СПб, Урал...)
sold	Продано штук (1–500)

Требуется:

- Загрузить данные из файла 22_phones.csv — файл должен находиться в том же проекте, что и программа.
- Добавить столбец price_segment:
  - "Бюджетный" (< 30000)
  - "Средний" (30000–70000)
  - "Премиум" (> 70000)
- Итоговая таблица должна содержать следующие столбцы (в указанном порядке):
  model, brand, price, region, sold, price_segment
- Вывести первые 10 строк, отсортированных по убыванию sold
- Построить гистограмму цен смартфонов:
  - Ось X — цена
  - Ось Y — количество моделей
  - Разделить по брендам (разные цвета)
  - Подпишите оси и добавьте заголовок, выведите легенду
  - Сохраните изображение как phones_price_hist.png

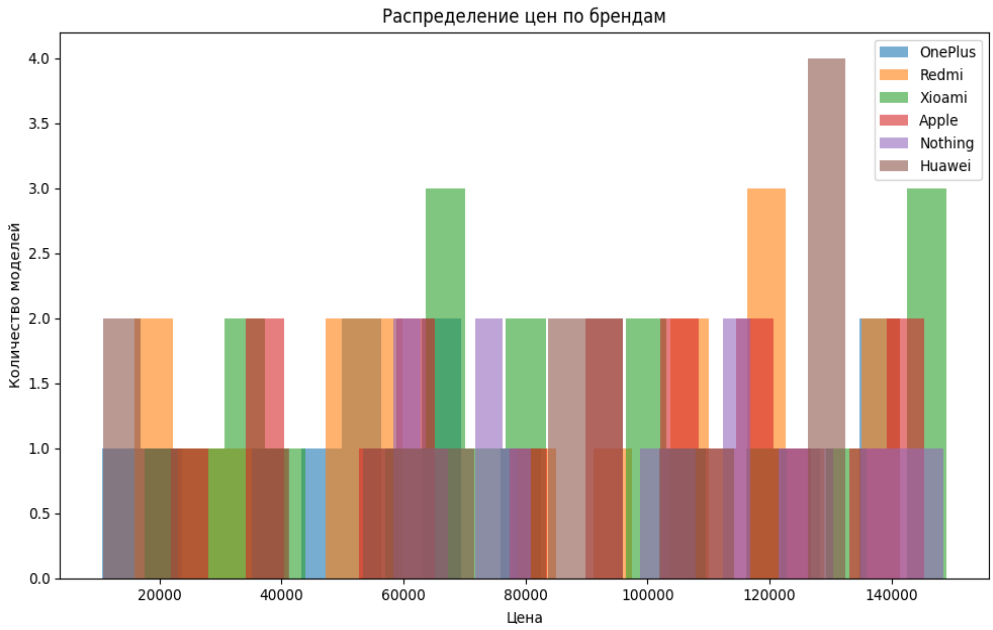
6. Записать ваш ответ в файл task_2.py

Решение

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv("22_phones.csv")


def categorize_price(p):
    if p < 30000:
        return "Бюджетный"
    elif p <= 70000:
        return "Средний"
    else:
        return "Премиум"
```

	df["price_segment"] = df["price"].apply(categorize_price)
	
	df = df[["model", "brand", "price", "region", "sold", "price_segment"]]
	
	df_sorted = df.sort_values(by="sold", ascending=False)
	print(df_sorted.head(10))
	
	plt.figure(figsize=(10, 6))
	for brand in df["brand"].unique():
	  subset = df[df["brand"] == brand]
	  plt.hist(subset["price"], bins=20, alpha=0.6, label=brand)
	
	plt.xlabel("Цена")
	plt.ylabel("Количество моделей")
	plt.title("Распределение цен по брендам")
	plt.legend()
	plt.tight_layout()
	plt.savefig("phones_price_hist.png")
	plt.show()

Вывод 10 строк	model brand price region sold price_segment 62 Ж Nothing 78948 Ростов-на-Дону 496 Премиум 24 Л Apple 123220 Новгород 495 Премиум 13 Н Huawei 112253 Самара 488 Премиум 44 Е Redmi 22416 Ростов-на-Дону 488 Бюджетный 6 Н Redmi 80857 Волгоград 477 Премиум 30 Л Redmi 58589 Самара 473 Средний 50 И Apple 38802 Казань 472 Средний

	46 Ж Nothing 148436 Новгород 453 Премиум 7 И Nothing 113110 Санкт-Петербург 447 Премиум 0 А OnePlus 52189 Москва 438 Средний
График	

Задача 3.

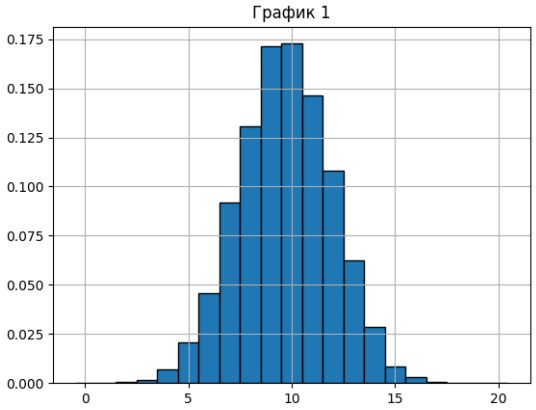
В ресторане клиент заказывает ужин из трёх блюд: закуска, основное и десерт.

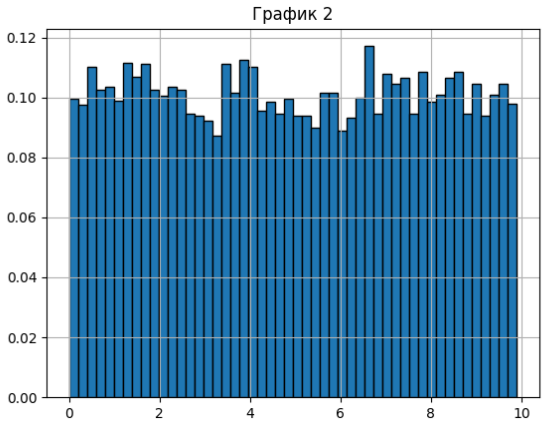
На выбор — 5 закусок (2 из них острые), 6 основных блюд (3 из них с морепродуктами), 4 десерта (1 из них содержит карамель). Повар случайным образом выбирает по одному блюду из каждой категории. Найдите вероятность того, что в заказе окажется ровно одно «интересное» блюдо — с острой закуской, с основным блюдом с морепродуктами или с десертом с карамелью.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0,45

Задача 4.





1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:
   - Для графика 1: число испытаний (n) и вероятность успеха (p).
   - Для графика 2: укажите минимальное (a) и максимальное (b) значения.
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Биномиальное, Гр.2 - Равномерное.
2	Для гр.1: n = 20, p ≈ 0.48.

	Для гр.2: a ≈ 0.0, b ≈ 9.9.
3	Гр.1: дискретные значения, пик около 9.7 — характерно для биномиального распределения. Гр.2: плоская гистограмма, значения в [0.0, 9.9] встречаются равномерно.

Вариант 23

Задача 1.

Вы работаете бизнес-аналитиком в компании, организующая доставку еды. В данный момент в вашем отделе идёт формирование отчетов об эффективности работы компании и предпочтениях заказчиков.

В вашем распоряжении есть следующие описания таблиц:

customers (id, name, address) — информация о заказчиках;

couriers (id, name, region) — информация о курьерах;

restaurants (id, name, cuisine) — информация о ресторанах;

orders (id, customer_id, courier_id, restaurant_id, order_date, total_price) — информация о заказах.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти клиентов, чьи имена начинаются на «Са» или содержат «ра». Вывести имя, адрес.

2. Найти рестораны, с которых было сделано более 20 заказов. Вывести название и кухню.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

1	SELECT name, address FROM customers WHERE name LIKE 'Ca%' OR name LIKE '%pa%';
2	SELECT r.name, r.cuisine FROM restaurants r JOIN orders o ON r.id = o.restaurant_id GROUP BY r.id, r.name, r.cuisine HAVING COUNT(o.id) > 20;

Задача 2.

Вы работаете аналитиком в крупной логистической компании. Ваша задача — проанализировать стоимость доставки грузов и определить закономерности в зависимости от типа груза и дальности маршрута.

У вас есть файл **23_deliveries.csv**, содержащий данные по доставке грузов.

Описание данных:

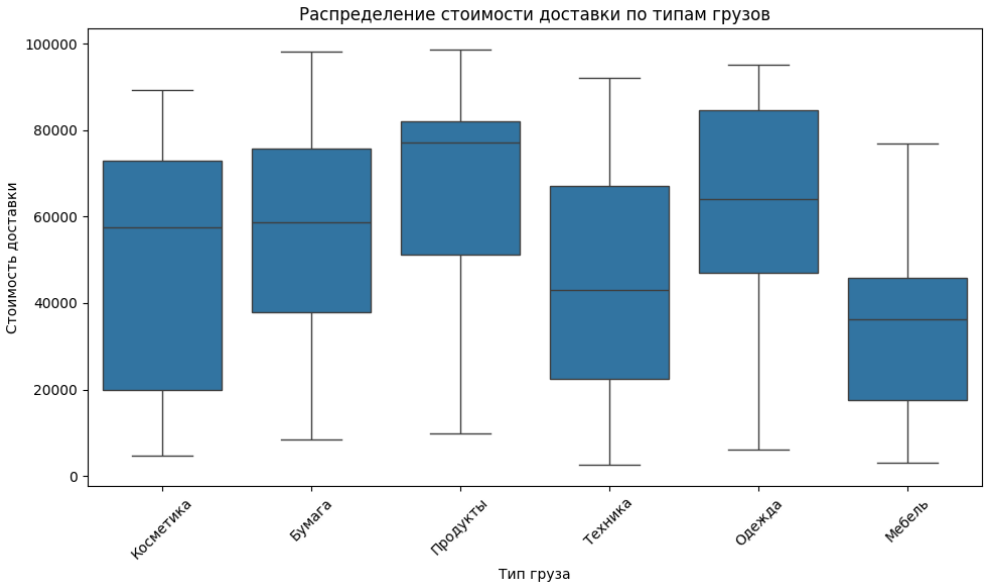
Поле	Описание
cargo_id	Идентификатор груза
cargo_type	Тип груза (Техника, Одежда, Продукты...)
cost	Стоимость доставки (от 1000 до 100000)
distance	Длина маршрута (км, от 10 до 5000)
region	Регион доставки (Центр, Север, Юг...)

Требуется:

- Загрузить данные из файла **23_deliveries.csv**.
- Добавить столбец distance_range:
  - "Близко" (< 100 км)
  - "Средне" (100–1000 км)
  - "Далеко" (> 1000 км)
- Итоговая таблица должна содержать столбцы: cargo_id, cargo_type, cost, distance, region, distance_range
- Вывести 10 строк с **наивысшей стоимостью доставки**
- Построить **boxplot (можете использовать sns.boxplot(data, x, y))**:
  - Ось X — тип груза (cargo_type)
  - Ось Y — стоимость (cost)
  - Сравните распределения стоимости доставки по типам
  - Сохранить изображение как **deliveries_boxplot.png**

Решение	import pandas as pd
	import matplotlib.pyplot as plt
	
	df = pd.read_csv("23_deliveries.csv")
	
	def categorize_distance(d):
	  if d < 100:
	  return "Близко"
	  elif d <= 1000:
	  return "Средне"
	  else:
	  return "Далеко"
	
	df["distance_range"] = df["distance"].apply(categorize_distance)
	
	df = df[["cargo_id", "cargo_type", "cost", "distance", "region", "distance_range"]]

	df_sorted = df.sort_values(by="cost", ascending=False) print(df_sorted.head(10)) cargo_by_cost = [df[df["cargo_type"] == ct]["cost"] for ct in df["cargo_type"].drop_duplicates()] plt.figure(figsize=(10, 6)) plt.boxplot(cargo_by_cost, labels=df["cargo_type"].drop_duplicates()) plt.xlabel("Тип груза") plt.ylabel("Стоимость доставки") plt.title("Распределение стоимости доставки по типам грузов") plt.xticks(rotation=45) plt.tight_layout() plt.savefig("deliveries_boxplot.png") plt.show()
Вывод 10 строк	cargo_id cargo_type cost distance region distance_range 97 8 Продукты 98631 1337 Новгород Далеко 27 29 Бумага 98223 4337 Москва Далеко 9 55 Одежда 95186 3494 Москва Далеко 41 70 Одежда 94950 2660 Псков Далеко 32 53 Бумага 94840 575 Мурманск Средне 50 1 Одежда 93526 621 Казань Средне 38 68 Бумага 92470 528 Уфа Средне 52 2 Техника 91996 4306 Москва Далеко 92 66 Техника 90755 1780 Уфа Далеко

	36 45 Косметика 89307 2752 Уфа Далеко
График	

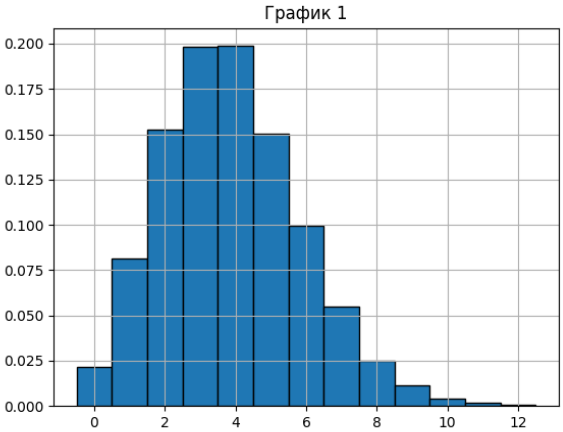
Задача 3.

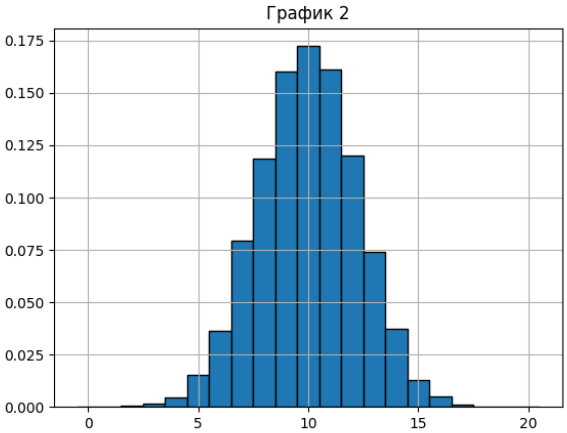
У Анны 12 значков: 5 — с животными, 4 — с эмблемами, и 3 — с надписями. Она выбирает 4 случайных значка для выставки. Какова вероятность, что среди них будут значки всех трёх видов?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0,5455

Задача 4.





1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:
   - Для графика 1: укажите среднее число событий (λ).
   - Для графика 2: число испытаний (n) и вероятность успеха (p).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Пуассона, Гр.2 - Биномиальное.

2	Для гр.1: $\lambda \approx 3.9$. Для гр.2: $n = 20$, $p \approx 0.50$.
3	Гр.1: дискретные значения, пик около 3.9, асимметрия вправо — характерно для Пуассона. Гр.2: дискретные значения, пик около 10.0 — характерно для биномиального распределения.

Вариант 24

Задача 1.

Вы работаете в отделе аналитиков в компании, организующей крупные музыкальные концерты. В данный момент в вашем отделе идёт формирование отчетов об эффективности работы компании и предпочтениях заказчиков.

В вашем распоряжении есть следующие описания таблиц:

visitors (id, name, city)— информация о заказчиках;

artists (id, name, genre)— информация о курьерах;

performances (id, visitor_id, artist_id, performance_date)— информация о концертах;

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти посетителей, чьи имена начинаются на «И» или содержат «ел». Вывести имя и город.

2. Найти жанры, у которых было более 15 посещений. Вывести жанр и количество посещений.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT name, city FROM visitors WHERE name LIKE 'И%' OR name LIKE '%ел%';
2	SELECT a.genre, COUNT(*) AS visit_count FROM performances p JOIN artists a ON p.artist_id = a.id GROUP BY a.genre HAVING COUNT(*) > 15;

Задача 2.

Вы работаете аналитиком в компании, специализирующейся на аренде техники для строительства. Ваша задача — проанализировать арендные ставки и выявить закономерности в зависимости от длительности аренды и типа техники.

У вас есть файл 24_rentals.csv, содержащий данные по аренде техники различными компаниями.

Описание данных:

Поле	Описание
machine	Название техники (Экскаватор, Бульдозер...)
type	Тип техники (Строительная, Погрузочная...)

Поле	Описание
price	Стоимость аренды за день (от 1000 до 25000)
days	Количество дней аренды (1 - 90)
location	Локация аренды (Москва, Новосибирск...)

Требуется:

- Загрузить данные из файла **24_rentals.csv** — файл должен находиться в том же проекте, что и программа.
- Добавить столбец rental_level:
  - "Краткосрочная" — до 7 дней включительно
  - "Средняя" — от 8 до 30 дней включительно
  - "Долгосрочная" — больше 30 дней
- Итоговая таблица должна содержать следующие столбцы (в указанном порядке):
  machine, type, price, days, location, rental_level
- Вывести первые **10 строк**, отсортированных по убыванию price.
- Построить **bar chart** (столбчатую диаграмму) средней стоимости аренды по типу техники:
  - Ось X — тип техники
  - Ось Y — средняя цена
  - Добавьте подписи осей, заголовок, сохраните изображение как **rental_price_bar.png**

6. Записать ваш ответ в файл task_2.py

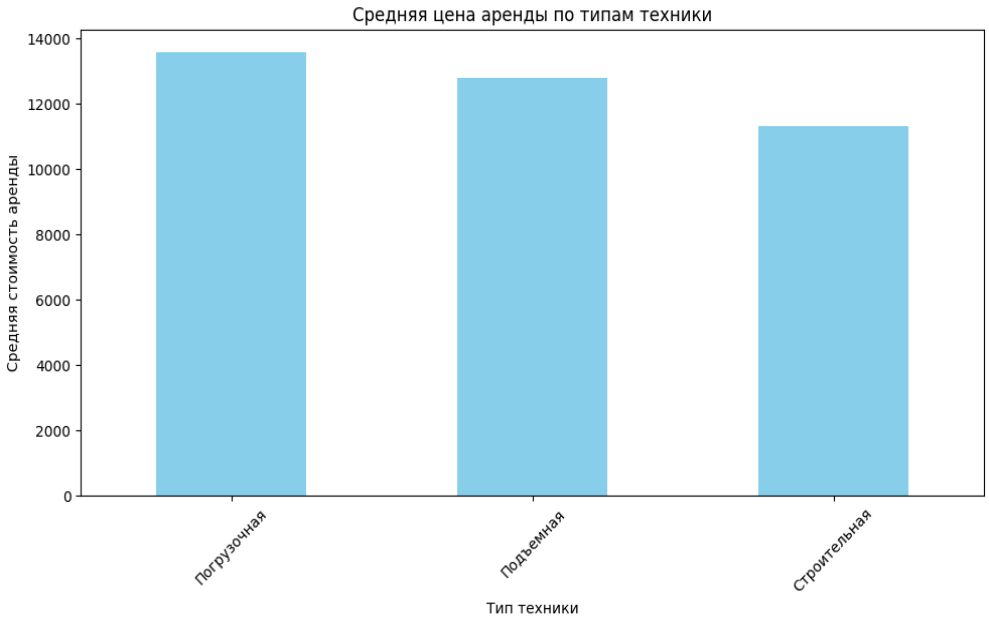
Решение

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv("24_rentals.csv")


def categorize_days(d):
    if d <= 7:
```

```
        return "Краткосрочная"
    elif d <= 30:
        return "Средняя"
    else:
        return "Долгосрочная"

df["rental_level"] = df["days"].apply(categorize_days)

df = df[["machine", "type", "price", "days", "location",
"rental_level"]]

df_sorted = df.sort_values(by="price", ascending=False)
print(df_sorted.head(10))

avg_prices = df.groupby("type")["price"].mean()

plt.figure(figsize=(10, 6))
avg_prices.plot(kind="bar", color="skyblue")
plt.xlabel("Тип техники")
plt.ylabel("Средняя стоимость аренды")
plt.title("Средняя цена аренды по типам техники")
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig("rental_price_bar.png")
plt.show()
```

Вывод 10 строк

```
    machine      type  price  days  location  rental_level
39      Каток  Погрузочная  24916    59     Псков  Долгосрочная
```

	73 Бульдозер Погрузочная 24862 35 Волгоград Долгосрочная 35 Автовышка Строительная 24612 17 Казань Средняя 34 Каток Погрузочная 24057 9 Псков Средняя 76 Экскаватор Подъемная 24002 38 Казань Долгосрочная 81 Кран Подъемная 23838 68 Псков Долгосрочная 7 Каток Строительная 23817 35 Москва Долгосрочная 36 Фронтальный погрузчик Строительная 23601 39 Казань Долгосрочная 85 Фронтальный погрузчик Подъемная 23187 57 Псков Долгосрочная 98 Кран Подъемная 22901 6 Мурманск Краткосрочная
График	

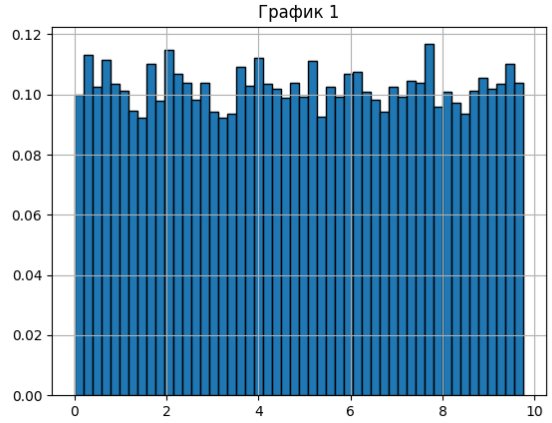
Задача 3.

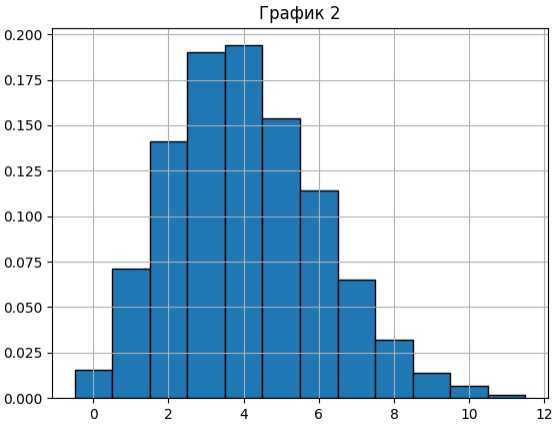
В школе проходит конкурс, где 40% участников — ученики младших классов, а 60% — старшеклассники. Среди младших 25% выбрали тему по биологии, остальные — по физике. Среди старших 70% выбрали тему по физике, остальные — по биологии. Найдите вероятность того, что случайно выбранный участник выбрал тему по физике.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.72

Задача 4.





1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:

   Для графика 1: укажите минимальное (a) и максимальное (b) значения.
   Для графика 2: укажите среднее число событий (λ).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

Решение:

1	Гр.1 - Равномерное,

	Гр.2 - Пуассона.
2	Для гр.1: a ≈ 0.0, b ≈ 9.8. Для гр.2: λ ≈ 4.1.
3	Гр.1: плоская гистограмма, значения в [0.0, 9.8] встречаются равномерно. Гр.2: дискретные значения, пик около 4.1, асимметрия вправо — характерно для Пуассона.

Вариант 25

Задача 1.

Вы работаете в отделе аналитиков в компании, владеющей известным музеем. В данный момент в вашем отделе идёт формирование отчетов об эффективности работы компании и предпочтениях посетителей.

В вашем распоряжении есть следующие описания таблиц:

visitors (id, name, age) — информация о посетителях;

exhibitions (id, title, period) — информация об экспонатах;

tickets (id, visitor_id, exhibition_id, visit_date, price) — информация о проданных билетах.

Вам, как аналитику, необходимо решить следующие задачи:

1. Найти посетителей, имена которых начинаются на «Ан» или содержат «ва». Вывести имя и возраст.

2. Найти выставки, на которые было продано более 20 билетов. Вывести название и количество билетов.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT name, age FROM visitors WHERE name LIKE 'Ан%' OR name LIKE '%ва%';
2	SELECT e.title, COUNT(t.id) AS ticket_count FROM exhibitions e JOIN tickets t ON e.id = t.exhibition_id GROUP BY e.id, e.title HAVING COUNT(t.id) > 20;

Задача 2.

Вы работаете аналитиком в сервисе каршеринга. Ваша задача — исследовать закономерности в использовании автомобилей разного типа в разных городах.

У вас есть файл **25_carsharing.csv**, содержащий данные по поездкам.

Описание данных:

Поле	Описание
car_id	Идентификатор автомобиля
car_type	Тип автомобиля (Седан, Кроссовер...)
duration	Время поездки в минутах (от 5 до 300)

Поле	Описание
distance	Расстояние поездки в км (от 1 до 100)
city	Город поездки (Москва, Казань...)

Требуется:

- Загрузить данные из файла **25_carsharing.csv** — файл должен находиться в том же проекте, что и программа.
- Добавить столбец trip_length:
  - "Короткая" — до 20 минут включительно
  - "Средняя" — от 21 до 60 минут
  - "Длинная" — больше 60 минут
- Итоговая таблица должна содержать следующие столбцы (в указанном порядке):
  car_id, car_type, duration, distance, city, trip_length
- Вывести первые **10 строк**, отсортированных по убыванию distance
- Построить **scatter plot** (диаграмму рассеяния), где:
  - Ось X — duration
  - Ось Y — distance
  - Цвета — разные типы автомобилей
  - Добавьте подписи осей, заголовок, сохраните изображение как **carsharing_scatter.png**

6. Записать ваш ответ в файл task_2.py

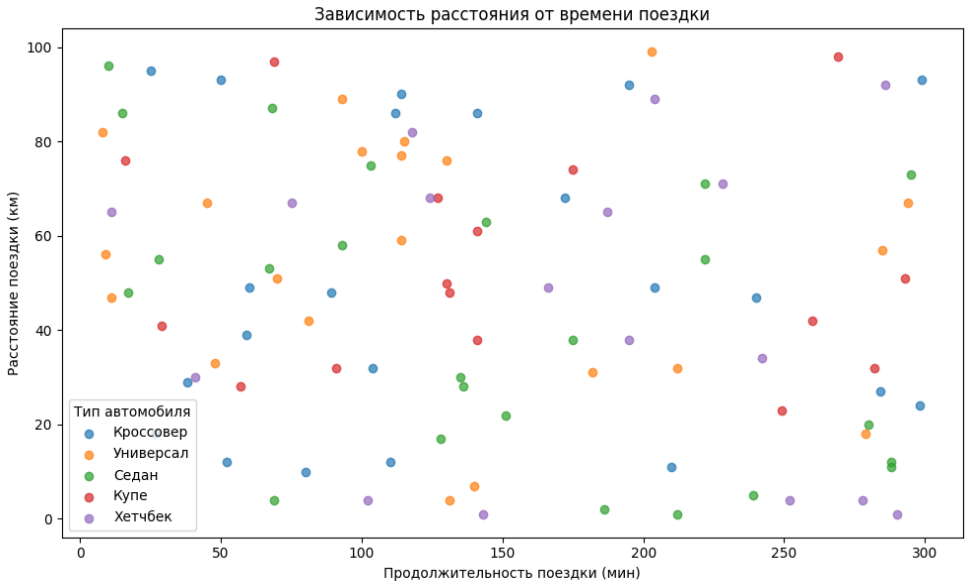
Решение	import pandas as pd
	import matplotlib.pyplot as plt
	
	df = pd.read_csv("25_carsharing.csv")
	
	def categorize_duration(d):
	  if d <= 20:

```
        return "Короткая"
    elif d <= 60:
        return "Средняя"
    else:
        return "Длинная"

df["trip_length"] = df["duration"].apply(categorize_duration)

df = df[["car_id", "car_type", "duration", "distance", "city", "trip_length"]]

df_sorted = df.sort_values(by="distance", ascending=False)
print(df_sorted.head(10))

plt.figure(figsize=(10, 6))

for car_type in df["car_type"].unique():
    subset = df[df["car_type"] == car_type]
    plt.scatter(subset["duration"], subset["distance"], label=car_type,
alpha=0.7)

plt.xlabel("Продолжительность поездки (мин)")
plt.ylabel("Расстояние поездки (км)")
plt.title("Зависимость расстояния от времени поездки")
plt.legend(title="Тип автомобиля")
plt.tight_layout()
plt.savefig("carsharing_scatter.png")
plt.show()
```

Вывод 10 строк

car_id	car_type	duration	distance	city	trip_length
24	60	Универсал	203	99	Уфа	Длинная
68	32	Купе	269	98	Санкт-Петербург	Длинная
5	72	Купе	69	97	Мурманск	Длинная
58	39	Седан	10	96	Новгород	Короткая
84	89	Кроссовер	25	95	Москва	Средняя
77	61	Кроссовер	299	93	Новгород	Длинная
81	86	Кроссовер	50	93	Ростов-на-Дону	Средняя
35	81	Хетчбек	286	92	Казань	Длинная
66	2	Кроссовер	195	92	Волгоград	Длинная
19	8	Кроссовер	114	90	Мурманск	Длинная

График



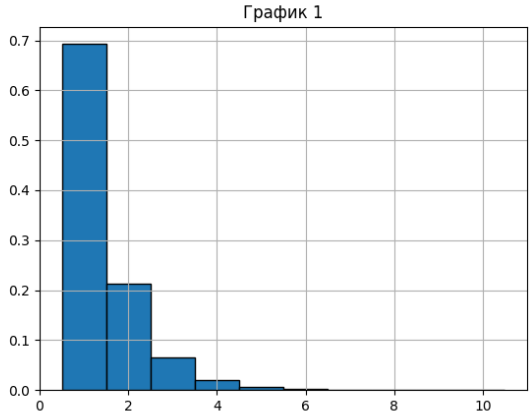
Задача 3.

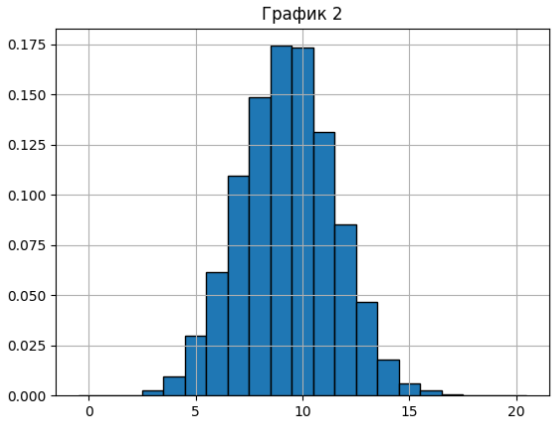
В системе умного дома установлено 6 различных датчиков: 3 датчика движения, 2 датчика дыма и 1 датчик температуры. Сбой системы может отключить случайные 2 датчика. Найдите вероятность, что после сбоя останется работать хотя бы один датчик каждого типа.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0,6

Задача 4.





1. Определите тип распределения для каждого графика.

2. Оцените параметры распределения:

Для графика 1: укажите вероятность успеха (p) и стандартное отклонение (σ).

Для графика 2: укажите математическое ожидание (μ) и стандартное отклонение (σ).

3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt.

1	Гр.1 - Геометрическое, Гр.2 - Нормальное.
2	Для гр.1: $p \approx 0.70$, $\sigma \approx 0.81$. Для гр.2: $\mu \approx -0.01$, $\sigma \approx 0.96$.

3	Гр.1: убывающее распределение с пиком на 1, характерно для геометрического распределения. Гр.2: симметричный колоколообразный график с центром около μ — характерно для нормального распределения.

Вариант 26

Задача 1.

Вы работаете аналитиком в крупном банке. Ваша задача – помогать бизнесу принимать решения на основе данных. Сейчас отделы формируют финальные отчеты о работе с клиентами.

Вам предоставлено описание таблиц базы данных:

clients (id, name, email, city, registration_date) — информация о клиентах.

accounts (id, client_id, account_type, balance) — банковские счета клиентов.

transactions (id, account_id, transaction_type, amount, transaction_date) — транзакции по счетам.

Для составления отчетов сформируйте запросы, удовлетворяющие следующим условиям:

1. Банк готовит персонализированные предложения для клиентов, имена которых соответствуют определённым паттернам. Особенно интересуют те, чьи имена начинаются на «Сер» или содержат «на» — часто это важные клиенты, имеющие долгую историю сотрудничества.

   В качестве результата выведите следующую информацию: имя, город и почта

2. Финансовый отдел хочет понять, какие клиенты наиболее активны. Цель — выделить клиентов, у которых 10 и более транзакций, чтобы предложить им премиальные услуги.

   В качестве результата выведите следующую информацию: имя, почта, количество транзакций.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

1	SELECT
	  name,
	  city,
	  email
	FROM
	  clients
	WHERE
	  name LIKE 'Сер%'
	  OR name LIKE '%на%';

2	SELECT
	  c.name,
	  c.email,
	  COUNT(t.id) AS transaction_count
	FROM
	  clients c
	JOIN
	  accounts a ON c.id = a.client_id
	JOIN
	  transactions t ON a.id = t.account_id
	GROUP BY
	  c.id, c.name, c.email
	HAVING
	  COUNT(t.id) >= 10;

Задача 2

Вы работаете аналитиком в крупном банке с данными транзакций клиентов. Руководству важно понять, какие клиенты проявляют высокую активность, а также как распределена транзакционная нагрузка по сегментам клиентов (например, VIP, бизнес и т. д.).

У вас есть файл `transactions_data.csv`, содержащий данные за 6 месяцев.

Формат таблицы:

client_id — идентификационный номер клиента (целое число)
month — месяц в формате YYYY-MM (строка)
transaction_count — количество транзакций в месяце (число, может быть ноль)
segment – сегмент клиента (строка - retail, vip, business)

Требуется

1. Загрузить данные из файла `transactions_data.csv` - файл должен находиться в том же проекте, что и программа.
   Используйте pd.to_datetime(), sort_values()

Решение можно оформить так: загрузка → преобразование месяца → сортировка.

2. Для каждого сегмента клиентов (segment) определить общее количество операций за весь период: Используйте groupby() и agg().
3. Создайте новый столбец total_transactions, который принимает значение общего количества транзакций за весь период. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

   client_id
   month
   transaction_count
   total_transactions
   segment
   Приведите дату обратно к формату 'YYYY-MM' через .dt.strftime()
   Проверьте, что порядок и формат соответствуют.
4. Вывести первые 10 строк итоговой таблицы.

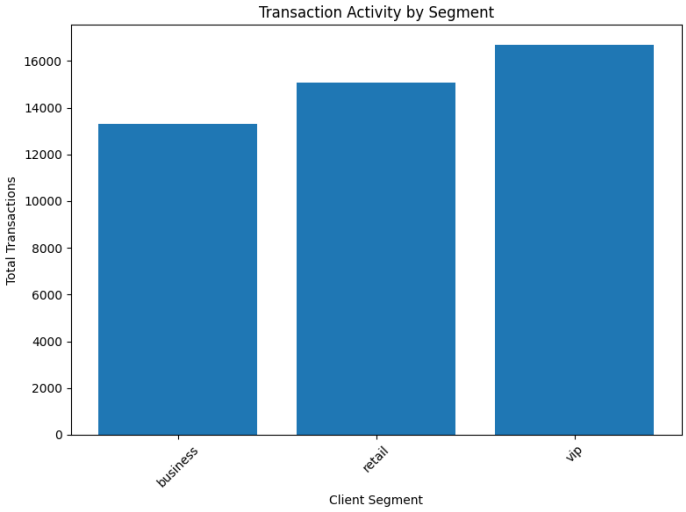
   Для оценки распределения транзакционной активности по сегментам
   Построить столбчатую диаграмму (bar plot), где:

   Ось X — segment
   Ось Y — total_transactions за весь период
   Каждый столбец отражает общее число транзакций для соответствующего сегмента
   Используйте drop_duplicates() для удаления дубликатов из таблицы.

   Используйте matplotlib.pyplot (plot(), xticks(), legend(), boxplot(), bar() и т. д.)
5. Сохранить изображение как activity_barplot.png

Запишите свое решение в файл: task_2.py.
*Документацию по приведенным методам можно получить с помощью функции **help()**.*

Решение	import pandas as pd import matplotlib.pyplot as plt file_path = 'transactions_data.csv' df = pd.read_csv(file_path) df['month'] = pd.to_datetime(df['month'], format='%Y-%m') df = df.sort_values(by=['client_id', 'month']) segment_totals = df.groupby('segment')['transaction_count'].agg(total_transactions='sum').reset_inde x() client_totals = df.groupby('client_id')['transaction_count'].transform('sum') df['total_transactions'] = client_totals df['month'] = df['month'].dt.strftime('%Y-%m') final_df = df[['client_id', 'month', 'transaction_count', 'total_transactions', 'segment']] print(final_df.head(10)) unique_clients = df.drop_duplicates(subset='client_id') segment_summary = unique_clients.groupby('segment')['total_transactions'].sum().reset_index() plt.figure(figsize=(8, 6)) plt.bar(segment_summary['segment'], segment_summary['total_transactions']) plt.xlabel('Client Segment') plt.ylabel('Total Transactions') plt.title('Transaction Activity by Segment') plt.xticks(rotation=45) plt.tight_layout() plt.savefig('activity_barplot.png') plt.close()

Вывод 10 строк	client_id month ... total_transactions segment 0 1 2024-01 ... 56 business 1 1 2024-02 ... 56 business 2 1 2024-03 ... 56 business 3 1 2024-04 ... 56 business 4 1 2024-05 ... 56 business 5 1 2024-06 ... 56 business 6 2 2024-01 ... 112 vip 7 2 2024-02 ... 112 vip 8 2 2024-03 ... 112 vip 9 2 2024-04 ... 112 vip
График	

Задача 3.

На школьной ярмарке проводится игра с кубиками. Игрок бросает три одинаковых шестигранных игральных кубика. Выигрыш определяется по следующим правилам:

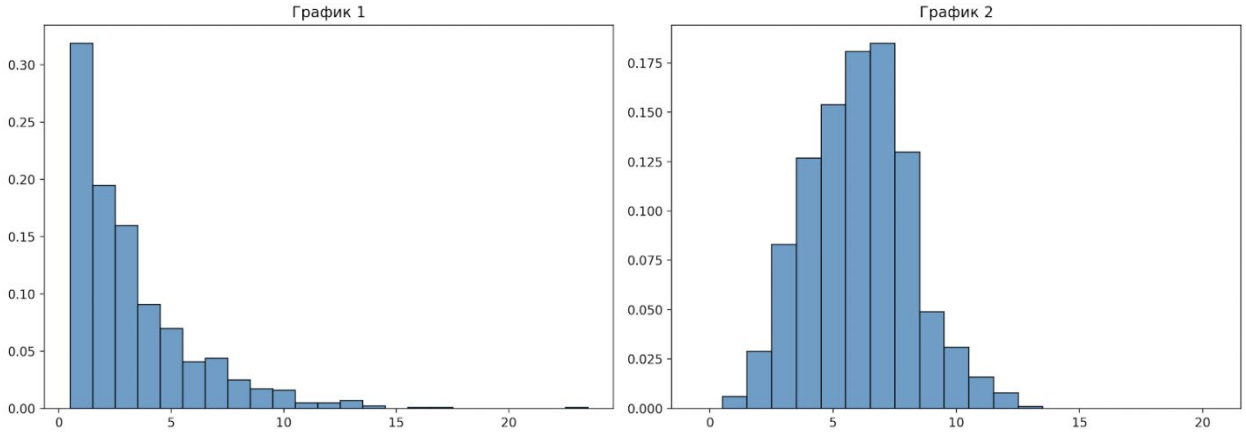
1. Главный приз разыгрывается, если выпавшие значения образовывают неубывающую последовательность из трёх последовательных чисел (например, 2–3–4 или 4–5–6).
2. Дополнительный приз: сумма всех трёх выпавших чисел делится на 5.

Какова вероятность того, что один случайный бросок приведёт одновременно к главному и дополнительному призу?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 1/216

Задача 4



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:
   - Для Графика 1: укажите вероятность успеха (p).
   - Для Графика 2: число испытаний (n) и вероятность успеха (p).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt

Решение:

	График 1. Геометрическое распределение График 2. Биномиальное распределение
	График 1. $p \approx 0.3$. График 2. $n = 20, p = \frac{\mu}{n} \approx \frac{6}{20} \approx 0.3$
	График 1: убывающая ступенчатая форма с максимумом на 1, значения — только натуральные числа. Это соответствует геометрическому распределению, описывающему номер первого успеха. *График 2:* дискретные значения (целые числа от 0 до 20), пик около 6, характерная форма колокола — соответствует биномиальному распределению.

Допустимые погрешности при учете ответа:

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)σ: ±10–15% (например, σ = 10 → ответ 9–11)
n: точное значение (например, n = 20 → только 20, так как это целое число).
p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)

Вариант 27

Задача 1.

Вы работаете аналитиком в транспортной компании. Руководству требуется отчёт на основе базы данных, содержащей информацию о пассажирах, кораблях и типах пассажиров.
.

Вам предоставлено описание таблиц базы данных:

passengers (id, name, age, sex, ship_id, passenger_type_id, fare, survived (1- выжил или 0 – не выжил)) — информация о пассажирах.

ships (id, name, departure_port, arrival_port, departure_date (дата отправления)) — информация о кораблях.

passenger_types (id, type_name) — типы пассажиров (например, взрослый, ребёнок, экипаж).

Для составления отчетов сформируйте запросы, удовлетворяющие следующим условиям:

1. Отдел перевозок просит выделить те рейсы, где доля выживших была ниже 50%. Выведите id, название корабля, порт отправления, порт прибытия, дата отправления, среднее число выживших на каждом маршруте, отсортируйте такие рейсы по дате отправления (от самых ранних).

2. Маркетинговый отдел просит рассчитать по каждому типу пассажиров: средний возраст и среднюю стоимость билета.

В качестве ответа необходимо предоставить два sql-запроса.
Запишите свое решение в файл: задача_1.txt.

Решения:

1	SELECT
	  s.id AS ship_id,
	  s.name AS ship_name,
	  s.departure_port,
	  s.arrival_port,
	  s.departure_date,
	  AVG(p.survived) AS survival_rate
	FROM
	  passengers p
	JOIN
	  ships s ON p.ship_id = s.id
	GROUP BY
	  s.id, s.name, s.departure_port, s.arrival_port, s.departure_date
	HAVING
	  AVG(p.survived) < 0.5
	ORDER BY
	  s.departure_date ASC;

2	SELECT
	  pt.type_name,
	  AVG(p.age) AS avg_age,
	  AVG(p.fare) AS avg_fare
	FROM
	  passengers p
	JOIN
	  passenger_types pt ON p.passenger_type_id = pt.id
	GROUP BY
	  pt.type_name;

Задача 2.

Вы работаете аналитиком в исследовательской группе, изучающей поведение пассажиров на титанике. Цель анализа – понять, какие факторы (класс, пол, возраст) могли влиять на выживаемость пассажиров, а также визуализировать распределение возрастов выживших в разных классах.

У вас есть файл titanic_data.csv, содержащий данные о пассажирах.
Формат таблицы:

PassengerId — уникальный идентификатор пассажира (целое число)
Survived — признак выживания (0 — не выжил, 1 — выжил)
Pclass — класс билета (1, 2, 3)
Name — имя пассажира
Sex — пол
Age — возраст (может быть пропущен)
SibSp — количество братьев/сестер или супругов на борту
Parch — количество родителей или детей на борту
Ticket — номер билета
Fare — стоимость билета
Cabin — каюта (может быть пропущена)
Embarked — порт посадки (C, Q, S)

Требуется

1. Загрузить данные из файла titanic_data.csv - файл должен находиться в том же проекте, что и программа.
   Используйте sort_values()
   Решение можно оформить так: загрузка → преобразование месяца → сортировка.
2. Создайте новый столбец familySize, который является суммой SibSp и Parch. Это значение отражает размер семьи на борту.
3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):
   PassengerId
   Pclass
   Sex
   Age
   Fare
   FamilySize
   Survived
   Проверьте, что порядок и формат соответствуют.
4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав данные по не возрастанию возраста.
5. Для оценки распределения возрастов выживших пассажиров по классам - построить диаграмму размаха (boxplot), где:
   Ось X — PClass
   Ось Y — Age

Отображаются только пассажиры, которые выжили (Survived == 1)
Используйте matplotlib.pyplot (boxplot(), xticks(), xlabel(), ylabel(), title() и т. д.).
Для построения графика:

Отфильтруйте данные по условию Survived == 1

Разделите возраст по классам (Pclass) — можно использовать groupby() или списки по классам вручную

Постройте boxplot по трем классам

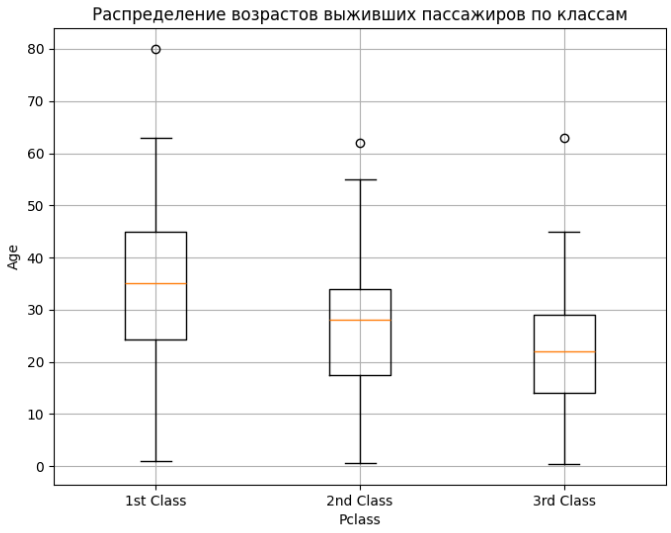
6. Сохранить изображение как **survival_age_boxplot.png**

Запишите свое решение в файл: task_2.py.

*Документацию по приведенным методам можно получить с помощью функции **help()**.*

Решение

```
import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv("titanic_data.csv")

data['FamilySize'] = data['SibSp'] + data['Parch']

final_columns = ['PassengerId', 'Pclass', 'Sex', 'Age', 'Fare', 'FamilySize',
'Survived']
final_data = data[final_columns]

final_data_sorted = final_data.sort_values(by='Age', ascending=False)
print(final_data_sorted.head(10))

survivors = final_data[final_data['Survived'] == 1]

ages_by_class = [
   survivors[survivors['Pclass'] == pclass]['Age'].dropna()
   for pclass in [1, 2, 3]
]

plt.figure(figsize=(8, 6))
plt.boxplot(ages_by_class, labels=["1st Class", "2nd Class", "3rd Class"])
plt.xlabel("Pclass")
plt.ylabel("Age")
plt.title("Распределение возрастов выживших пассажиров по классам")
plt.grid(True)
plt.savefig("survival_age_boxplot.png")
plt.show()
```

Вывод 10 строк	PassengerId Pclass Sex Age Fare FamilySize Survived 630 631 1 male 80.0 30.0000 0 1 851 852 3 male 74.0 7.7750 0 0 493 494 1 male 71.0 49.5042 0 0 96 97 1 male 71.0 34.6542 0 0 116 117 3 male 70.5 7.7500 0 0 672 673 2 male 70.0 10.5000 0 0 745 746 1 male 70.0 71.0000 2 0 33 34 2 male 66.0 10.5000 0 0 54 55 1 male 65.0 61.9792 1 0 280 281 3 male 65.0 7.7500 0 0
График	

Задача 3.

На пристани стоят 3 одинаковых корабля, каждый из которых должен взять на борт ровно 5 пассажиров. На пристани ожидают 15 пассажиров, среди которых 3 шпиона, а остальные – обычные граждане.

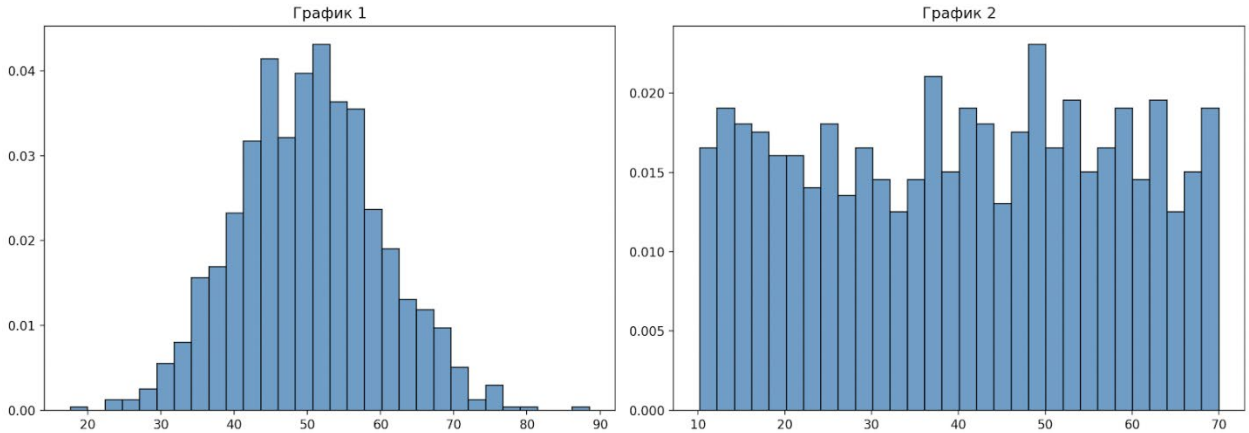
Пассажиры распределяются случайным образом, без предпочтений. Никто не остается на берегу.

Какова вероятность, что на каждом корабле окажется ровно один шпион?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.2747

Задача 4.



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:
   - Для Графика 1: укажите среднее (μ) и стандартное отклонение (σ).
   - Для Графика 2: укажите границы a и b.
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt

Решение:

1	*График 1:* нормальное распределение. *График 2:* равномерное распределение.
2	*График 1*: $\mu \approx 50$, $\sigma \approx \frac{\lvert 40-60 \rvert}{2} \approx 10$ *График 2: a = 10, b = 70*
3	*График 1*: характерной симметричной форме, пик в центре и постепенное спадание по краям. *График 2:* характерная «плоская» форма, почти одинаковая высота по всему интервалу — указывает на равномерное распределение значений в заданном диапазоне.

Допустимые погрешности при учете ответа:

μ: ±5% от истинного значения (например, если μ = 50, допустим ответ 48–52).
σ: ±10–15% (например, σ = 10 → ответ 9–11)

a и b: ±2–3% от диапазона (например, данные от 10 до 30 → a ≈ 10 ± 0.5, b ≈ 30 ± 0.5).

Вариант 28

Задача 1.

Вы работаете аналитиком в министерстве образования. Руководству требуется отчёт на основе базы данных, содержащей информацию о университетах, их показателей деятельности, студентах и прочих характеристиках.

Вам предоставлено описание таблиц базы данных:

unis (id, name, country, type, score) — университеты (с итоговым баллом).

metrics (uni_id, teach, research, cite, income, intl_score) — показатели (обучение, исследования, цитирования, доход, международность).

students (uni_id, total, ratio, intl_pct) — студенты (общее число, соотношение студент/преподаватель, % иностранных).

Для составления отчетов сформулируйте SQL-запросы, удовлетворяющие следующим условиям:

1. Отдел международного сотрудничества просит выделить университеты, где доля иностранных студентов выше 30%, а международный показатель выше 70.
   Выведите название, страну, общий балл (score) таких университетов,

   упорядочите по убыванию общего балла.

2. Аналитический отдел просит для каждого типа университета рассчитать средние значения обучения, исследований и цитирования.
   Выведите тип, средний балл обучения, исследования, цитирования, отсортируйте по среднему баллу обучения в порядке убывания

В качестве ответа необходимо предоставить два sql-запроса.
Запишите свое решение в файл: задача_1.txt.

Решение:

```
1  SELECT u.name, u.country, u.score
   FROM unis u
   JOIN metrics m ON u.id = m.uni_id
   JOIN students s ON u.id = s.uni_id
   WHERE s.intl_pct > 30 AND m.intl_score > 70
   ORDER BY u.score DESC;
```

2	SELECT u.type, AVG(m.teach) AS avg_teach, AVG(m.research) AS avg_research, AVG(m.cite) AS avg_cite FROM unis u JOIN metrics m ON u.id = m.uni_id GROUP BY u.type ORDER BY avg_teach DESC;

Задача 2.

Вы работаете аналитиком в команде, изучающей академическую эффективность университетов по всему миру. Цель анализа – понять, какие ключевые показатели (преподавание, исследования, международная активность) влияют на общий рейтинг университета, а также визуализировать их связь с итоговой оценкой.

У вас есть файл university_data.csv, содержащий информацию о различных университетах.
Формат таблицы:

name — Название университета
scores_teaching — Балл за преподавание (0–100)
scores_research — Балл за научные исследования (0–100)
scores_citations — Балл за цитирования (0–100)
scores_international_outlook — Международная вовлечённость (0–100)
overall_score — Общий рейтинг университета (0–100)

Требуется

1. Загрузить данные из файла university_data.csv - файл должен находиться в том же проекте, что и программа. Отсортируйте строки по убыванию overall_score. Используйте sort_values()
   Решение можно оформить так: загрузка → сортировка.
2. Создайте новый столбец **engagement_score**, который представляет собой среднее значение из: **scores_teaching**, **scores_research**, **scores_citations**.
3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):
   - name
   - scores_teaching
   - scores_research
   - scores_international_outlook
   - engagement_score
   - overall_score

Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав данные по не возрастанию overall_score.
5. Для выявления зависимости между академической вовлечённостью и общим рейтингом университета. Используйте только 100 первых значений.

   Ось X — engagement_score

   Ось Y — overall_score

   Используйте matplotlib.pyplot (scatter(), xlabel(), ylabel(), title() и т. д.)

   Настройте подписи осей и заголовок:

   Подпись оси X — Индекс вовлечённости

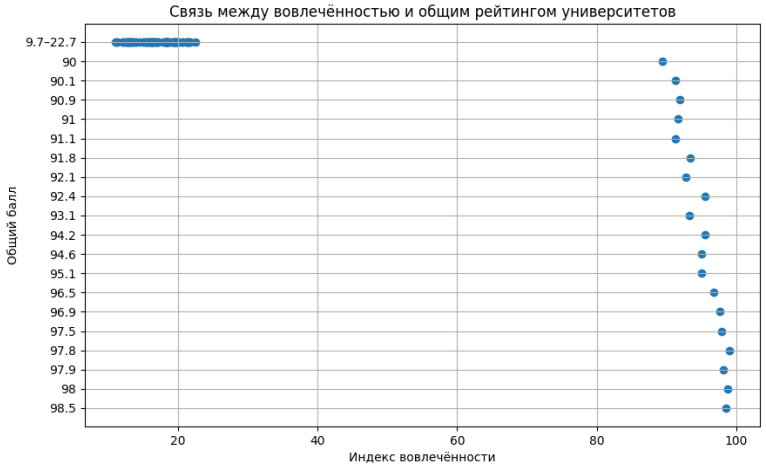
   Подпись оси Y — Общий балл

   Заголовок — Связь между вовлечённостью и общим рейтингом университетов
6. Сохранить изображение как **engagement_vs_overall_scatter.png**

Запишите свое решение в файл: task_2.py.

Документацию по приведенным методам можно получить с помощью функции ***help()***.

Решение	```
	import pandas as pd
	import matplotlib.pyplot as plt
	
	file_path = 'university_data.csv'
	df = pd.read_csv(file_path)
	df_sorted = df.sort_values(by='overall_score', ascending=False)
	
	df_sorted['engagement_score'] = df_sorted[
	  ['scores_teaching', 'scores_research', 'scores_citations']
	].mean(axis=1)
	
	final_df = df_sorted[[
	  'name',
	  'scores_teaching',
	  'scores_research',
	  'scores_international_outlook',
	  'engagement_score',
	  'overall_score'
	]]
	
	print(final_df.head(10))
	
	x = final_df['engagement_score'].head(100)
	y = final_df['overall_score'].head(100)
	
	plt.figure(figsize=(10, 6))
	plt.scatter(x, y)
	plt.xlabel('Индекс вовлечённости')
	plt.ylabel('Общий балл')
	plt.title('Связь между вовлечённостью и общим рейтингом университетов')
	plt.grid(True)
	
	plt.savefig('engagement_vs_overall_scatter.png')
	plt.close()
	``

Вывод 10 строк	name ... overall_score 0 University of Oxford ... 98.5 1 Stanford University ... 98 2 Massachusetts Institute of Technology ... 97.9 3 Harvard University ... 97.8 4 University of Cambridge ... 97.5 5 Princeton University ... 96.9 6 California Institute of Technology ... 96.5 7 Imperial College London ... 95.1 8 University of California, Berkeley ... 94.6 9 Yale University ... 94.2
График	

Задача 3.

В группе из 15 студентов:
5 учатся на «отлично»,
6 — на «хорошо»,
4 — на «удовлетворительно».

Для участия в стажировке случайным образом выбирают 3 студентов. Стажировку могут пройти только те трое, у кого средний балл группы — не ниже “хорошо”. Считается, что:

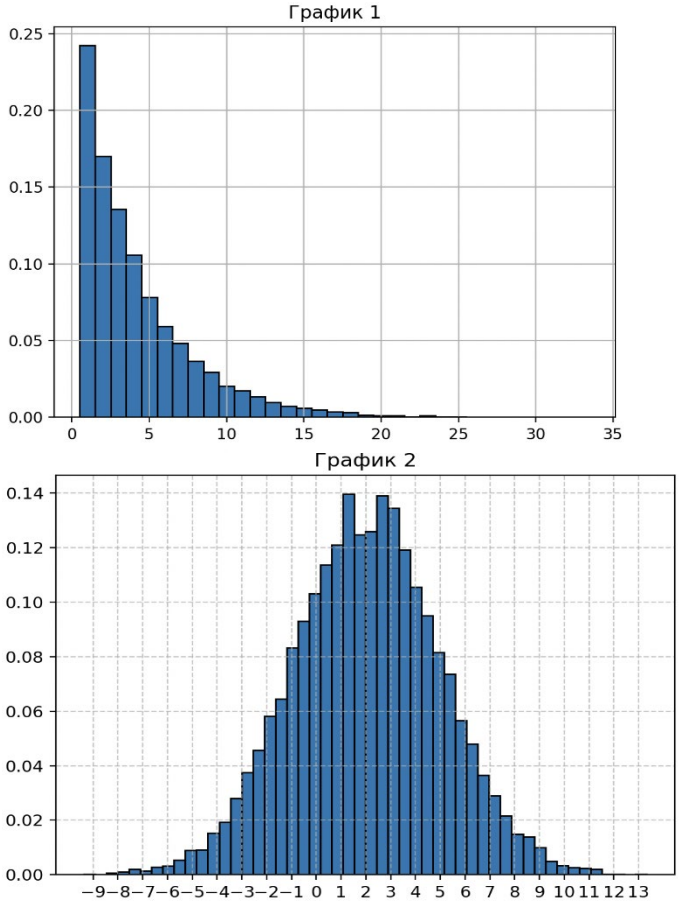
«Отлично» — это 5 баллов,
«Хорошо» — 4 балла,
«Удовлетворительно» — 3 балла.

Какова вероятность того, что средний балл троих выбранных студентов будет не ниже 4?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.714

Задача 4.



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:
   - Для Графика 1: укажите вероятность успеха (p).
   - Для Графика 2: укажите математическое ожидание (μ) и стандартное отклонение (σ)
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt

Решение:

1	*График 1:* геометрическое распределение. *График 2:* нормальное распределение.
2	*График 1*$p \approx 0.23$. *График 2*: $\mu \approx 2$, $\sigma \approx \frac{\lvert -1-5 \rvert}{2} \approx 3$

3	*График 1.* Гистограмма состоит из отдельных столбцов — это дискретное распределение. Столбцы быстро убывают вправо, начиная с максимума в самом начале. Наиболее вероятный результат — 1 или 2, затем вероятность резко падает. Распределение асимметрично — “хвост” вытянут вправо (положительная асимметрия). Эти признаки характерны для геометрического распределения, которое описывает количество испытаний до первого успеха. *График 2*. Гистограмма имеет симметричную колоколообразную форму. Центр графика — это максимум вероятности (математическое ожидание). Значения равномерно убывают от центра в обе стороны. Отдалённые от центра значения встречаются реже, но постепенно — без резких спадов. Всё это типично для нормального распределения, описывающего множество природных и социальных явлений.

Допустимые погрешности при учете ответа:

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)σ: ±10–15% (например, σ = 10 → ответ 9–11)

μ: ±5% от истинного значения (например, если μ = 50, допустим ответ 48–52).
σ: ±10–15% (например, σ = 10 → ответ 9–11)

Вариант 29

Задача 1.

Вы работаете аналитиком в приюте для животных. Ваша задача – помогать руководству принимать решения на основе данных. Сейчас сотрудники формируют финальные отчеты о работе с животными.

Вам предоставлено описание таблиц базы данных:

animals (id, name, species, arrival_date, shelter_id) — информация о животных, поступивших в приют.

shelters (id, name, city) — информация о приютах.

adoptions (id, animal_id, adopter_name, adoption_date) — информация об усыновлениях.

Для составления отчетов сформируйте запросы, удовлетворяющие следующим условиям:

1. Руководство приюта хочет запустить акцию по продвижению животных с особыми именами. Интересуют животные, чьи имена начинаются на «Ма» или содержат «ша». Необходимо получить список таких животных с указанием их имени, вида и города приюта, в котором они находятся.
2. Волонтёры планируют предложить специальную программу для приютов, где происходит особенно много усыновлений. Необходимо определить приюты, где было 5 и более усыновлений. Для отчёта нужно вывести название приюта, город и количество усыновлений.

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

```sql
SELECT
    a.name AS animal_name,
    a.species,
    s.city
FROM
    animals a
JOIN
    shelters s ON a.shelter_id = s.id
WHERE
    a.name LIKE 'Ма%' OR a.name LIKE '%ша%';
```

2	SELECT
	  s.name AS shelter_name,
	  s.city,
	  COUNT(*) AS adoption_count
	FROM
	  adoptions ad
	JOIN
	  animals a ON ad.animal_id = a.id
	JOIN
	  shelters s ON a.shelter_id = s.id
	GROUP BY
	  s.id, s.name, s.city
	HAVING
	  COUNT(*) >= 5;

Задача 2.

Вы работаете аналитиком в городском приюте для животных. Руководство хочет понять, как меняется активность усыновлений со временем, а также насколько популярны те или иные виды животных (например, кошки, собаки и т.д.).

У вас есть файл pet_adoptions.csv, содержащий данные о питомцах и их усыновлении.

Формат таблицы:

Поле	Описание
pet_id	Уникальный идентификатор питомца (строка)
pet_name	Имя питомца (строка)
species	Вид животного (например, Dog, Cat, Rabbit)
breed	Порода животного
gender	Пол (Male или Female)
color	Цвет животного
adopted	Признак, был ли усыновлен (TRUE/FALSE)
adoption_date	Дата усыновления (формат YYYY-MM-DD)

Требуется

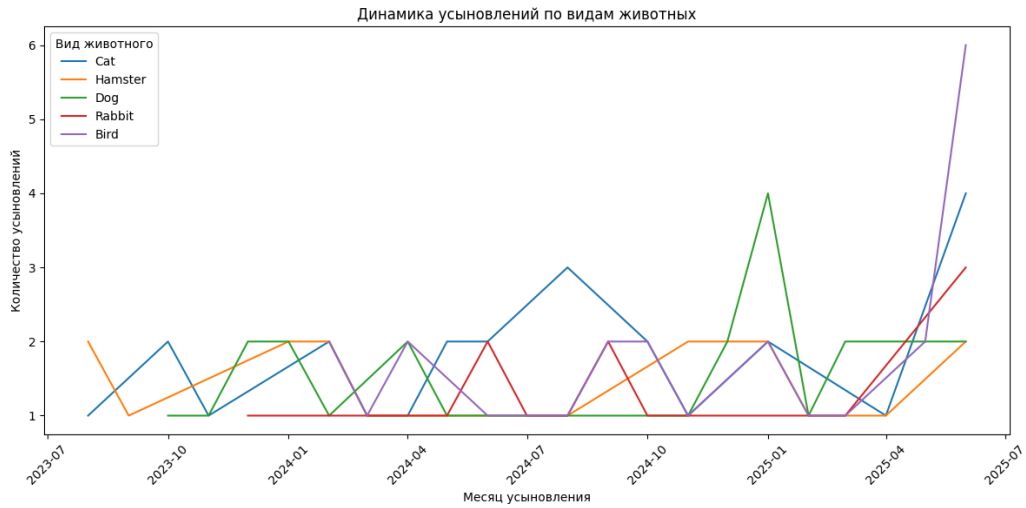
1. Загрузить данные из файла pet_adoptions.csv - файл должен находиться в том же проекте, что и программа.
   Используйте pd.to_datetime
   Решение можно оформить так: загрузка → преобразование месяца
2. Оставьте только усыновленных питомцев (у которых adopted == TRUE). Убедитесь, что значения даты усыновления не содержат пропусков или ошибок.
3. Преобразуйте дату усыновления в формат YYYY-MM. Используйте dt.to_period('M') и dt.to_timestamp().
4. Для каждого месяца и вида животного (species) определите общее количество усыновлений. Используйте groupby() и size(). Запишите это значение в новый столбец adoptions.
5. Выведите первые 10 строк итоговой агрегированной таблицы (с указанием adoption_date (YYYY-MM), species, adoptions) для проверки.
6. Постройте линейный график (line plot), отображающий динамику усыновлений по видам животных:
   Ось X — Месяц усыновления (в формате YYYY-MM)
   Ось Y — Количество усыновлений
   Каждая линия отражает вид животного (species)
   Добавьте подписи осей, заголовок, легенду, поверните подписи дат
   Используйте matplotlib.pyplot (plot(), xticks(), legend(), bar() и т. д.)
7. Сохранить изображение как *adoption_lineplot.png*

Запишите свое решение в файл: task_2.py.

Документацию по приведенным методам можно получить с помощью функции ***help()***.

Решение	

```
import pandas as pd
import matplotlib.pyplot as plt

file_path = 'pet_adoptions.csv'
df = pd.read_csv(file_path)
df['adoption_date'] = pd.to_datetime(df['adoption_date'], errors='coerce')

adopted_df = df[df['adopted'] == True].copy()
adopted_df = adopted_df.dropna(subset=['adoption_date'])

adopted_df['adoption_month'] =
adopted_df['adoption_date'].dt.to_period('M').dt.to_timestamp()

grouped = adopted_df.groupby(['adoption_month',
'species']).size().reset_index(name='adoptions')

print(grouped.head(10))

plt.figure(figsize=(12, 6))
for species in grouped['species'].unique():
    species_data = grouped[grouped['species'] == species]
    plt.plot(species_data['adoption_month'], species_data['adoptions'],
label=species)

plt.xlabel('Месяц усыновления')
plt.ylabel('Количество усыновлений')
plt.title('Динамика усыновлений по видам животных')
plt.xticks(rotation=45)
plt.legend(title='Вид животного')
plt.tight_layout()

plt.savefig('adoption_lineplot.png')
plt.show()
```

Вывод 10 строк	adoption_month species adoptions 0 2023-08-01 Cat 1 1 2023-08-01 Hamster 2 2 2023-09-01 Hamster 1 3 2023-10-01 Cat 2 4 2023-10-01 Dog 1 5 2023-11-01 Cat 1 6 2023-11-01 Dog 1 7 2023-12-01 Dog 2 8 2023-12-01 Rabbit 1 9 2024-01-01 Dog 2
График	

Задача 3.

В приюте для животных «Сказочный островок» живут 10 животных:

4 кошки,

3 собаки,

3 кролика.

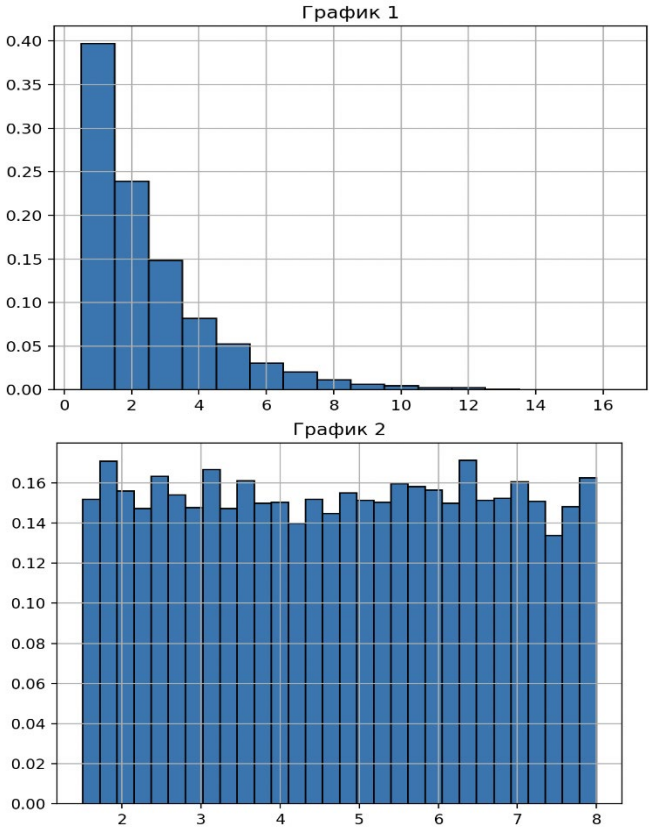
К выставке планируется выбрать ровно трёх животных. Выбор происходит случайно и одновременно, порядок не имеет значения.

Найдите вероятность того, что в тройке окажутся животные разных видов.

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.3

Задача 4.



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:
    - Для Графика 1: укажите вероятность успеха (p).
    - Для Графика 2: укажите границы равномерного распределения (a и b).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt

Решение:

1	*График 1:* геометрическое распределение. *График 2:* равномерное распределение.

2	*График 1*$p \approx 0.4$. *График 2*: $a \approx 2, b \approx 8$
3	*График 1.* Гистограмма состоит из отдельных столбцов — это дискретное распределение. Столбцы быстро убывают вправо, начиная с максимума в самом начале. Наиболее вероятный результат — 1 или 2, затем вероятность резко падает. Распределение асимметрично — “хвост” вытянут вправо (положительная асимметрия). Эти признаки характерны для геометрического распределения, которое описывает количество испытаний до первого успеха. *График 2*. Гистограмма имеет прямоугольную форму — высота столбцов примерно одинаковая во всём диапазоне от 2 до 8. Это говорит о том, что вероятность всех значений в этом интервале одинакова, а за пределами интервала — равна нулю. Такие характеристики соответствуют непрерывному равномерному распределению.

Допустимые погрешности при учете ответа:

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)σ: ±10–15% (например, σ = 10 → ответ 9–11)

a и b: ±2–3% от диапазона (например, данные от 10 до 30 → a ≈ 10 ± 0.5, b ≈ 30 ± 0.5).

Вариант 30

Задача 1.

Вы работаете аналитиком в исследовательском центре, изучающем поведение покемонов и их тренеров. Руководство просит вас подготовить отчёты на основе базы данных, содержащей информацию о покемонах, тренерах и типах покемонов.

Вам предоставлено описание таблиц базы данных:

pokemons (id, name, level, type_id, trainer_id, hp, captured) — информация о покемонах. Поле captured содержит 1, если покемон был пойман, и 0 — если нет.

trainers (id, name, region, experience_years) — информация о тренерах.

pokemon_types (id, type_name) — типы покемонов (например, огненный, водный, электрический и т.д.)

Для составления отчетов сформируйте запросы, удовлетворяющие следующим условиям:

1. Аналитический отдел просит найти имена тренеров, у которых меньше половины покемонов были пойманы. Отсортируйте таких тренеров по опыту (experience_years) по возрастанию.
2. Отдел классификации просит рассчитать по каждому типу покемонов: средний уровень (level) и среднее количество очков здоровья (hp).

В качестве ответа необходимо предоставить два sql-запроса.

Запишите свое решение в файл: задача_1.txt.

Решение:

Запрос 1.

```sql
SELECT
    t.name,
    t.experience_years
FROM
    trainers t
JOIN
    pokemons p ON t.id = p.trainer_id
GROUP BY
    t.id, t.name, t.experience_years
HAVING
    AVG(p.captured) < 0.5
ORDER BY
    t.experience_years ASC;
```

Запрос 2.	SELECT pt.type_name, AVG(p.level) AS average_level, AVG(p.hp) AS average_hp FROM pokemons p JOIN pokemon_types pt ON p.type_id = pt.id GROUP BY pt.type_name;

Задача 2.

Вы работаете аналитиком в биоинформатической лаборатории. Команде важно понять, как распределены физические характеристики различных покемонов, чтобы использовать эти данные в будущих исследованиях.

У вас есть файл pokemon_data.csv, содержащий информацию о базовом опыте, росте и весе покемонов.

Описание данных:

Поле	**Описание**
id	Уникальный идентификационный номер покемона (целое число)
name	Имя покемона (строка)
base_experience	Базовое количество опыта, получаемое при победе над этим покемоном (целое число)
height	Рост покемона, измеренный в условных единицах (целое число) – всегда больше 0
weight	Вес покемона, измеренный в условных единицах (целое число) – всегда больше 0

Требуется

1. Загрузить данные из файла pokemon_data.csv - файл должен находиться в том же проекте, что и программа.
2. Создать новый столбец power_ratio, рассчитанный как: base_experienceheight ×weight. Этот показатель покажет насколько “мощен” покемон с учетом его физических размеров

3. Итоговая таблица должна содержать следующие столбцы (в указанном порядке):

id
name
base_experience
height
weight
power_ratio

Проверьте, что порядок и формат соответствуют.

4. Вывести первые 10 строк итоговой таблицы. Предварительно отсортировав в порядке не убывания по полю power_ratio.

Используйте sort_values()

5. Для оценки роста покемонов постройте гистограмму распределения роста.

Ось X — значение роста

Ось Y — количество покемонов с таким ростом.

Укажите подписи для осей и название гистограммы.

Используйте matplotlib.pyplot.hist(), plt.xlabel(), plt.ylabel(), plt.title().

6. Сохранить изображение как height_histogram.png

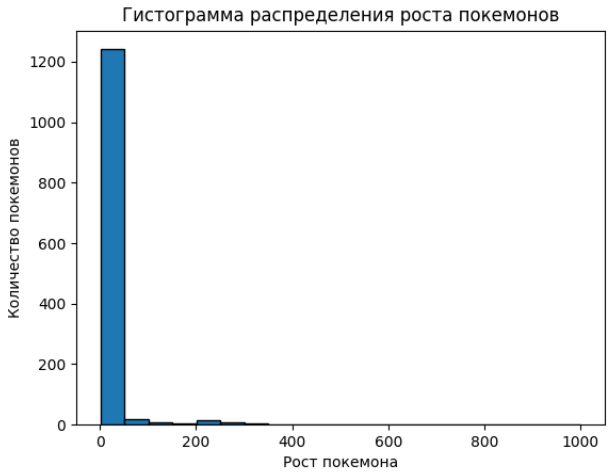
Запишите свое решение в файл: task_2.py.
*Документацию по приведенным методам можно получить с помощью функции **help()**.*

Решение

```
import pandas as pd
import matplotlib.pyplot as plt

file_path = 'pokemon_data.csv'
data = pd.read_csv(file_path)

data['power_ratio'] = data['base_experience'] / (data['height'] * data['weight'])

final_columns = ['id', 'name', 'base_experience', 'height', 'weight', 'power_ratio']
final_data = data[final_columns]

sorted_data = final_data.sort_values(by='power_ratio', ascending=True)

print(sorted_data.head(10))

plt.hist(data['height'], bins=20, edgecolor='black')
plt.xlabel('Рост покемона')
plt.ylabel('Количество покемонов')
plt.title('Гистограмма распределения роста покемонов')
plt.savefig('height_histogram.png')
plt.close()
```

Вывод строк

	id	name	base_experience	height	weight	power_ratio
1224	10200	meowth-gmax	58	330	10000	0.000018
1244	10220	centiskorch-gmax	184	750	10000	0.000025
1229	10205	eevee-gmax	65	180	10000	0.000036
1249	10225	duraludon-gmax	187	430	10000	0.000043
1223	10199	pikachu-gmax	112	210	10000	0.000053
1230	10206	snorlax-gmax	189	350	10000	0.000054
1247	10223	alcremie-gmax	173	300	10000	0.000058
1239	10215	coalossal-gmax	255	420	10000	0.000061
1235	10211	inteleon-gmax	265	400	10000	0.000066
1241	10217	appletun-gmax	170	240	10000	0.000071

График	

Задача 3.

В регионе Кантон обитает 10 различных покемонов: 9 обычных (их назовём A_1, A_2, …, A_9) и 1 редкий — Мью (обозначим R).

Когда тренер заходит в высокую траву, с равной вероятностью может появиться один из обычных покемонов, а редкий Мью появляется с вероятностью вдвое меньше, чем любой обычный.

То есть:

- Вероятность встретить любого обычного покемона: p
- Вероятность встретить редкого покемона: p/2
- Всего: $9p + p/2 = 19p/2 = 1 \Rightarrow p = 2/19$, а $p(R) = 1/19$

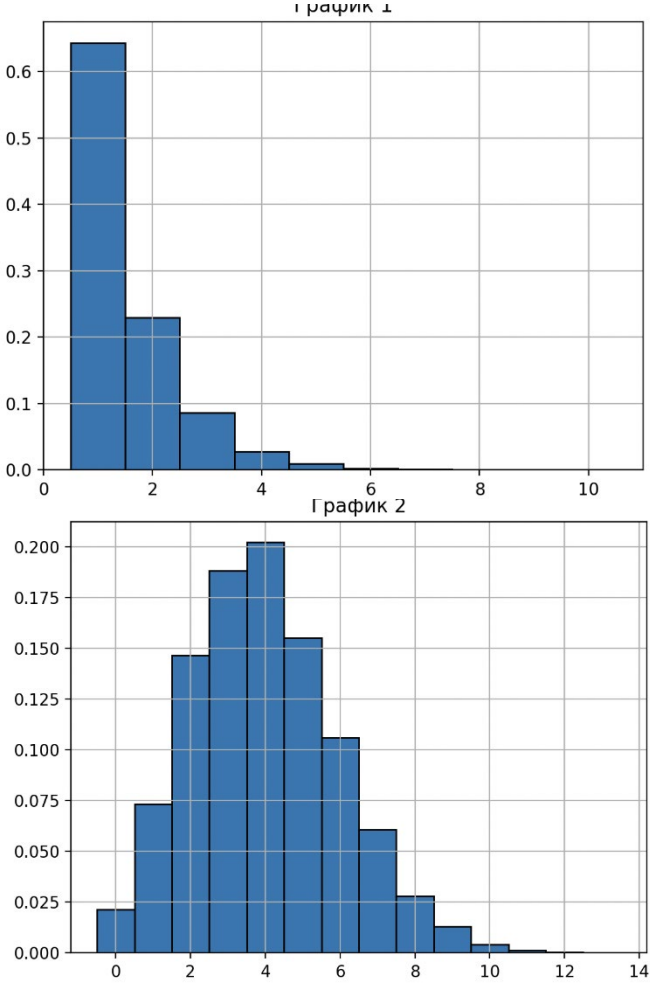
Тренер встретил четырёх покемонов, все они оказались разными.

Какова вероятность, что среди них был редкий покемон R?

Запишите свое решение в файл: задача_3.txt. Ваше решение должно основываться на математических вычислениях - использование сторонних программ не допускается.

Ответ: 0.25

Задача 4.



1. Определите тип распределения для каждого графика.
2. Оцените параметры распределения:
   Для Графика 1: укажите вероятность успеха (p).
   Для Графика 2: укажите параметр λ (интенсивность события).
3. Объясните, по каким визуальным признакам вы определили тип распределения.

Запишите свое решение в файл: задача_4.txt

Решение:

1	*График 1:* геометрическое распределение. *График 2:* распределение Пуассона.
2	*График 1*$p \approx 0.65$. *График 2*: $\lambda \approx 4$

3	*График 1.* Гистограмма состоит из отдельных столбцов — это дискретное распределение. Столбцы быстро убывают вправо, начиная с максимума в самом начале. Наиболее вероятный результат — 1 или 2, затем вероятность резко падает. Распределение асимметрично — “хвост” вытянут вправо (положительная асимметрия). Эти признаки характерны для геометрического распределения, которое описывает количество испытаний до первого успеха. *График 2* Гистограмма также дискретная, но в отличие от графика 1, сначала значения возрастают, достигают максимума около x = 4, а затем убывают. Такая форма характерна для **распределения Пуассона**, которое описывает число событий, происходящих за фиксированный интервал времени или пространства. Визуально оно напоминает “горб” и имеет правостороннюю асимметрию, особенно при малых λ.

Допустимые погрешности при учете ответа:

p: ±0.05 (например, p = 0.7 → ответ 0.65–0.75)σ: ±10–15% (например, σ = 10 → ответ 9–11)

- λ:±10–15% от истинного значения