МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

ОТЧЕТ
АНАЛИЗ НАБОРА ДАННЫХ «GERMAN CREDIT DATA»

Выполнил: Большаков Артём Викторович
Группа: 4
Учебное заведение: Белорусский государственный университет
Факультет: прикладной математики и информатики
Лабораторная работа №: 3
Дата выполнения: 02.12.2025

Минск, 2025

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Целью данной работы является закрепление навыков работы с библиотеками python (pandas, matplotlib, Seaborn, sqlite3), путем проведения анализа набора данных "German Credit Data" для выявления ключевых факторов, влияющих на статус кредита, а также моделирования данных и их визуализации.

План выполнения работы:

1. Загрузка и подготовка исходных данных
2. Анализ переменных и их распределений
3. Кодирование категориальных признаков
4. Построение графиков и диаграмм
5. Формирование SQL-запросов для работы с базой данных

Ключевыми вопросами в данном анализе данных являются:

- Влияние цели кредита на сумму запрашиваемого кредита
- Отличие “хороших” и “плохих” заемщиков

ЗАГРУЗКА И ПОДГОТОВКА ДАННЫХ

Источником данных является текстовый файл `german.data` (полученный по ссылке https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data), формата CSV и разделителем пробелом. Файл состоит из 1000 строк и 21 столбца-атрибута. Исходный файл не содержал заголовков, поэтому именования столбцов были назначены вручную используя данные из файла `german.doc` со спецификацией датасета.

Для обработки данных использовался метод `df.isnull().sum()`. В результате пропущенных значений не было обнаружено. Данные уже были чистыми.

Кодирование категориальных признаков выполнялось с помощью метода LabelEncoder из библиотеки sklearn.

АНАЛИЗ ДАННЫХ

Анализ числовых признаков был произведен путем использования метода `describe()`, для получения статистики (среднее, медиана, минимум, максимум, стандартное отклонение).

Анализ категориальных признаков был произведен путем использования метода `df.describe(include=["object"])`, для получения статистики (частота, наиболее часто встречающиеся категории).

Распределение признаков асимметрично, присутствует много выбросов данных.

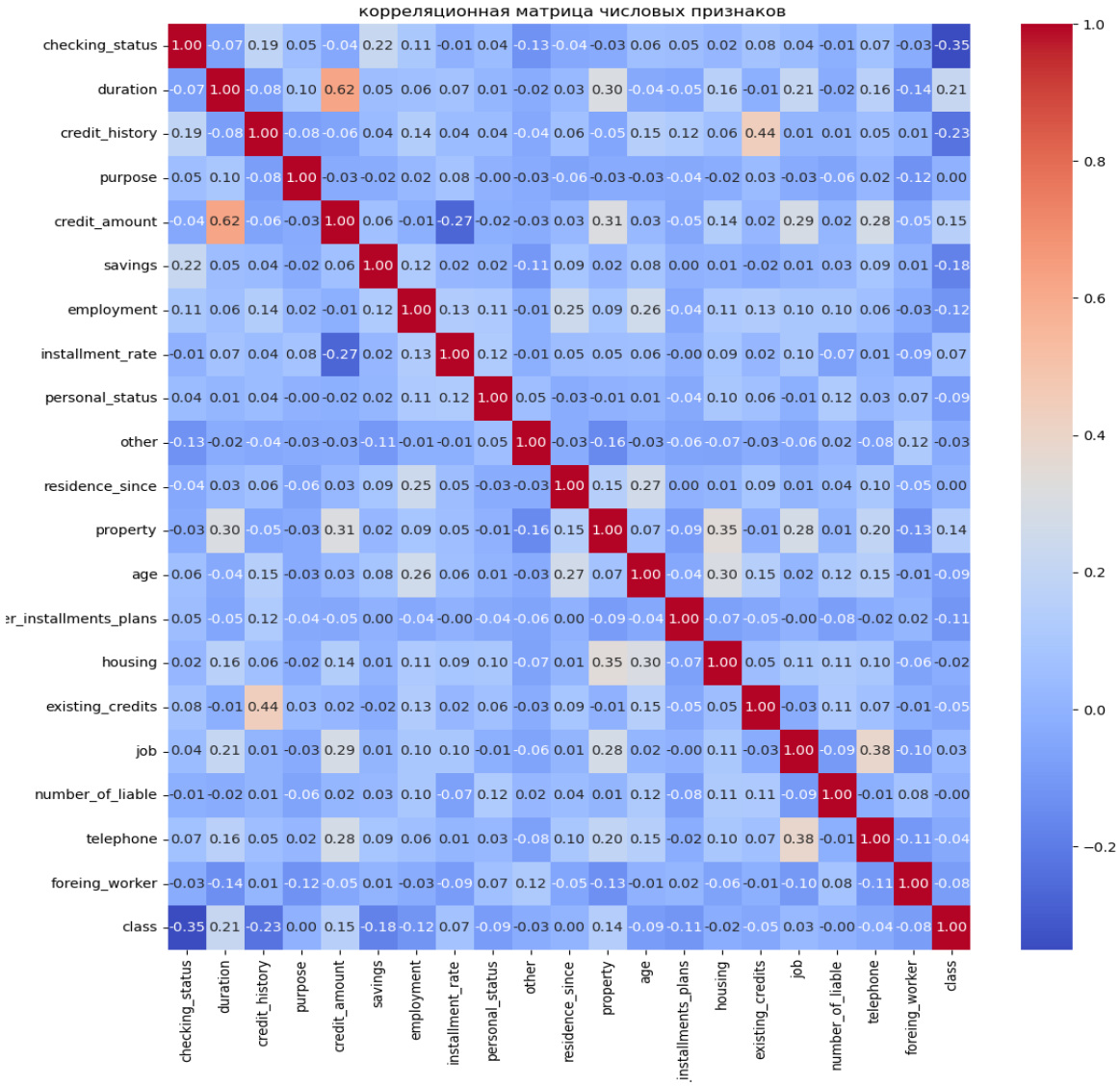
Ключевые выводы:

- Наиболее сильная корреляция между суммой кредита и его длительностью.
- Присутствует обратная зависимость между процентом рассрочки и суммой кредита (чем выше процент выплаты от зарплаты, тем меньше кредит).

ВИЗУАЛЬНЫЙ АНАЛИЗ

Для построения визуального анализа были использованы библиотеки matplotlib и seaborn. Были получены следующие графики:

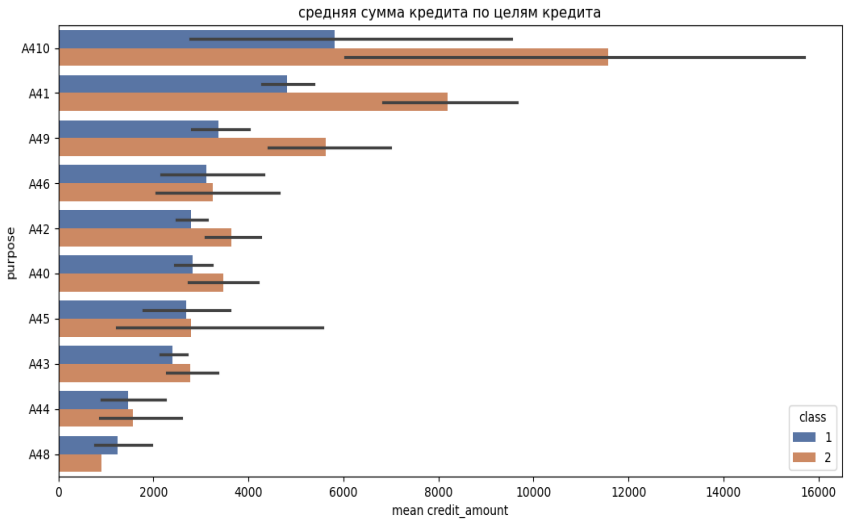
- Данный график визуализирует матрицу коэффициентов корреляции (красный — прямая, синий — обратная корреляция). Из графика видно, что самая сильная связь между суммой кредита и его длительностью, но так же есть корреляция между суммой кредита и процентом рассрочки и суммой кредита.



(Тепловая карта корреляций)

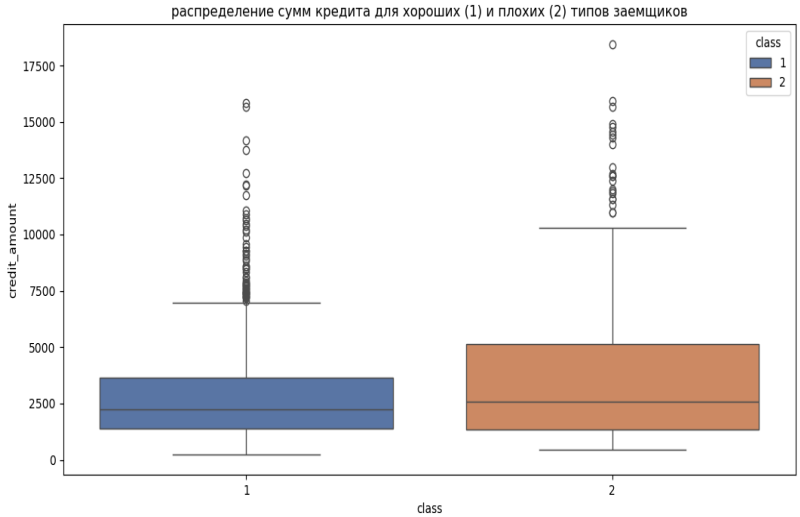
- График отображает среднее значение суммы кредита в зависимости от цели с разделением на “хороших” (1) и “плохих” (2) заемщиков. Черные линии на столбцах

(доверительные интервалы) показывают разброс данных и степень уверенности в среднем значении. Из графика видно, что самые дорогие кредиты в категории “прочее”, а самые дешевые кредиты в категориях “переобучение” и “бытовая техника”.



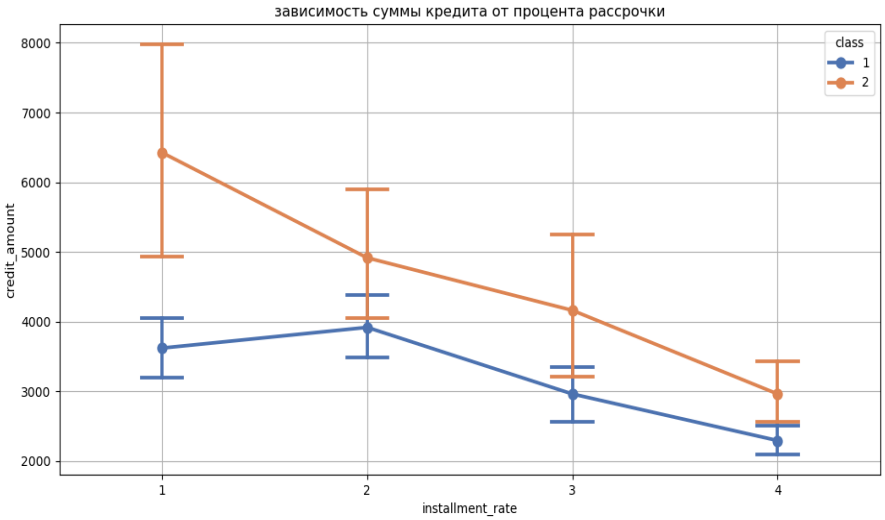
Столбчатая диаграмма (Barplot): Средняя сумма кредита по целям

- График сравнивает распределение сумм кредита для двух групп заемщиков: «хороших» (1) и «плохих» (2) с разделением на “хороших” (1) и “плохих” (2) заемщиков. Из графика видно, что у группы “плохих” заемщиков медиана ниже, чем у “хороших”, а также разброс результатов у “плохих” значительно больше.



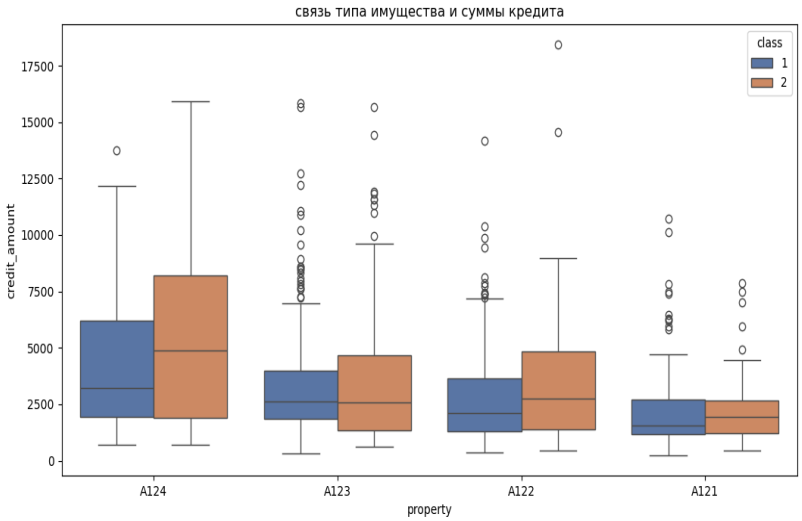
Boxplot: Распределение сумм кредита по классам заемщиков

- График демонстрирует изменение средней суммы кредита в зависимости от ставки рассрочки, которая измеряется как процент от располагаемого дохода (от 1 до 4) с разделением на “хороших” (1) и “плохих” (2) заемщиков. Из графика видно, что есть обратная зависимость. С увеличением процента ежемесячных выплат средняя сумма кредита падает.



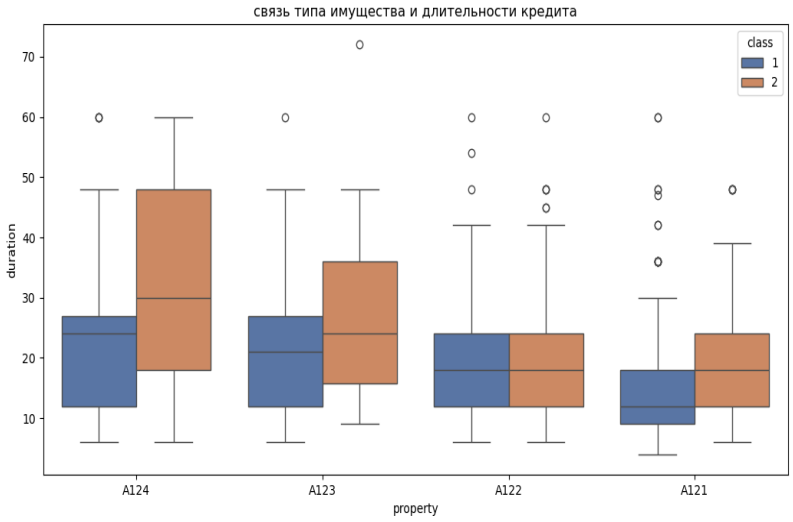
Точечный график (Pointplot): Зависимость суммы кредита от процента рассрочки

- График показывает распределение сумм кредита в зависимости от типа имущества, которым владеет заемщик с разделением на “хороших” (1) и “плохих” (2) заемщиков. Из графика видно, что клиенты категории без имущества/не известно имеют самую высокую медианную сумму кредита, а клиенты категории с недвижимостью запрашивают самые маленькие суммы.



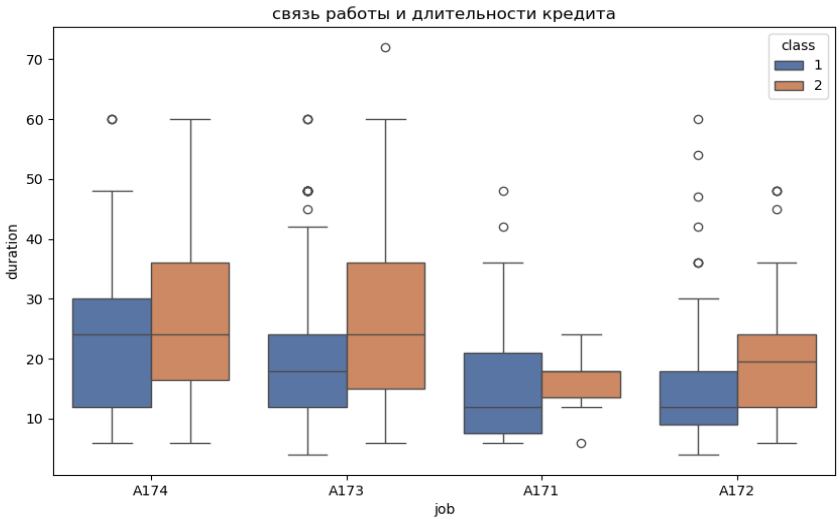
Boxplot: Связь типа имущества и суммы кредита

- График показывает распределение срока кредита в зависимости от типа имущества с разделением на “хороших” (1) и “плохих” (2) заемщиков. Из графика видно, что во всех категориях имущества медианная длительность кредита у «плохих» заемщиков выше, чем у «хороших». Это подтверждает, что более длительные кредиты чаще становятся проблемными.



Boxplot: Связь типа имущества и длительности кредита

- График показывает распределение срока кредита в зависимости от типа занятости с разделением на “хороших” и “плохих” заемщиков. В категориях высококвалифицированных специалистов и управленцев (A174 и A173) наблюдается наибольший разброс по длительности кредита. При этом медианный срок кредита у “плохих” заемщиков в этих группах заметно выше, чем у “хороших”.



Boxplot: Связь типа работы и длительности кредита

РАБОТА С БАЗОЙ ДАННЫХ SQLITE

Для взаимодействия с базой данных была использована библиотека sqlite3. Была создана локальная база данных `german_data.db` с таблицей `credits`, структура которой соответствует исходным данным (количество столбцов-атрибутов 21, а типы данных INTEGER и CHAR[] для соответствующих столбцов).

Данные были вставлены следующим методом: `df.to_sql("credits", conn, if_exists="replace", index=False).`

Выполненные SQL-запросы:

- запрос: назначение, размер и длительность кредита, возраст и наличие работы

```
запрос: назначение, размер и длительность кредита, возраст и наличие работы:
  purpose  credit_amount  duration  age   job  class
0    A410          18424        48   32  A174      2
1     A49          15945        54   58  A173      2
2    A410          15857        36   43  A174      1
3     A49          15672        48   23  A173      2
4     A43          15653        60   21  A173      1
```

```
SELECT purpose, credit_amount, duration, age, job, class
FROM credits
ORDER BY credit_amount DESC
LIMIT 5
```

- запрос: количество и средний размер кредита по целям кредита

```
запрос: количество и средний размер кредита по целям кредита
  purpose  count  avg_amount
0    A410     12      8209.0
1     A41    103      5370.0
2     A49     97      4158.0
3     A46     50      3180.0
4     A42    181      3067.0
5     A40    234      3063.0
6     A45     22      2728.0
7     A43    280      2488.0
8     A44     12      1498.0
9     A48      9      1206.0
```

```
SELECT purpose,
       COUNT(*) as count,
       ROUND(AVG(credit_amount), 0) as avg_amount
FROM credits
GROUP BY purpose
ORDER BY avg_amount DESC
```

- запрос: возраст кредитора, размер и длительности кредита, при условии, что размер кредита не меньше 4000, у людей без имущества

```
запрос: возраст кредитора, размер и длительности кредита, при условии что размер кредита не меньше 4000
   age  credit_amount  duration
0   58          15945        54
1   68          14896         6
2   60          14782        60
3   57          14318        36
4   27          14027        60
```

```
SELECT age, credit_amount, duration
FROM credits
WHERE class = 2 AND property = "A124" AND credit_amount
>= 4000
ORDER BY credit_amount DESC
LIMIT 5
```

ИТОГОВЫЕ ВЫВОДЫ

Наиболее значимыми признаками поведения и категории заемщика оказались: сумма, длительность кредита. Тип имущества сильно влияет на поведение двух рассматриваемых категорий заемщиков.

Обнаруженные взаимосвязи: “плохие” заемщики в среднем запрашивают более крупные суммы, наиболее сильная корреляция между суммой кредита и его длительностью, заемщики без имущества берут большие суммы, а люди с недвижимостью берут меньшие суммы.

Основные рекомендации: с повышенной осторожностью выдавать кредиты на сумму от 10000 людям, подходящим под категорию “плохих” заемщиков, а также людям, запрашивающим крупные суммы на длительные сроки.

ЗАКЛЮЧЕНИЕ

В ходе лабораторной работы были успешно закреплены навыки обработки и анализа данных на языке Python. Применение библиотек pandas, matplotlib и seaborn позволило выявить зависимости атрибутов данных, а использование sqlite3 продемонстрировало возможности интеграции Python с SQL для выполнения выборок. И были выполнени поставленные задачи.

Направление для дальнейшего анализа: в качестве возможного направления можно назвать построение ML-модели, для предсказания поведения заемщика на основе этих критериев.