Министерство науки и высшего образования Российской Федерации  
Пензенский государственный университет  
Кафедра вычислительная техника

**ОТЧЕТ**

по лабораторной работе №5

по дисциплине «Программные средства кибертехнических систем»

на тему «Изучение набора данных»

выполнили: ст -ты гр. 22ВВП1

Захаров А. С.

Беляев Д. И.

Демин М. С

Проверили:

Карамышева Н.С.

Зинкин С.А.

Пенза 2025

**Цель работы**

Изучить набор данных

**Задание**

1. Для полученного набора данных и сделанного в лабораторной работе 1 описания провести предварительный анализ данных в соответствии с этапами 2-и и 3-й фазы методологии CRISP.

2. Описать имеющиеся данные.

3. Проверить гипотезы о наличии пропусков и выбросов в данных.

4. Выделить наиболее релевантные решаемой задаче признаки.

5. Сформировать предложения о дополнительном сборе данных (если это необходимо и возможно).

6. Оформить отчёт, содержащий информацию о выполнении пунктов 1-5 задания.

Вариант 6. Задача кредитного скоринга (15-й столбец набора данных {1 – кредит выдан, 2 – кредит не выдан}) по 14 признакам.

**Ход работы**

Так как данные были взяты с открытого источника они относятся к категории сторонних. Количество экземпляров: 690. Частотные характеристики получены с помощью Wavelet Transform.

Таблица содержит 15 столбцов:

A1 – Бинарный категориальный признак (0,1).

A2 – Непрерывный числовой признак.

A3 – Непрерывный числовой признак.

A4 – Категориальный признак с 3 значениями (1,2,3).

A5 – Категориальный признак с 14 значениями (1–14).

A6 – Категориальный признак с 9 значениями (1–9).

A7 – Непрерывный числовой признак.

A8 – Бинарный категориальный признак (0,1).

A9 – Бинарный категориальный признак (0,1).

A10 – Непрерывный числовой признак.

A11 – Бинарный категориальный признак (0,1).

A12 – Категориальный признак с 3 значениями (1,2,3).

A13 – Непрерывный числовой признак.

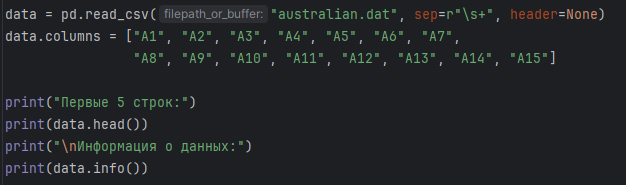
A14 – Непрерывный числовой признак.

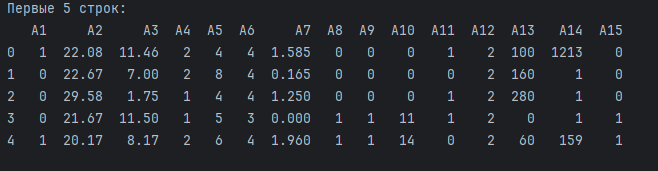
A15 (Класс) – Целевая переменная (1 – одобрение заявки, 2 – отказ).

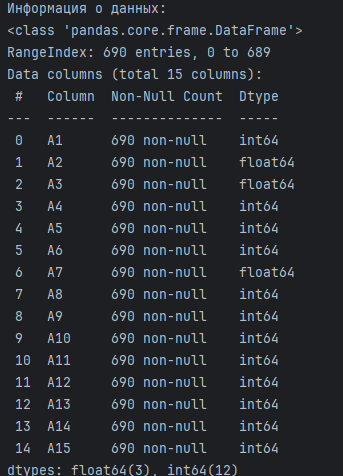
Нашей задачей будет выявить закономерности, позволяющие предсказать одобрение или отклонение кредитной заявки.

**Анализ данных**

Рассмотрим данные и общие сведения о них.

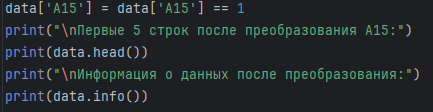


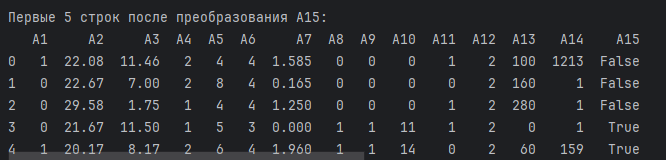


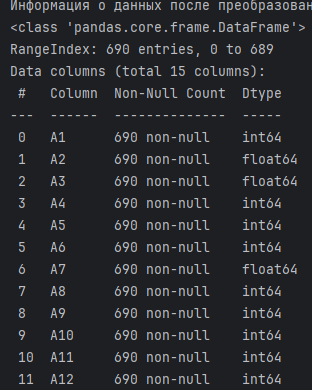


Видим, что количество записей в каждой столбце равно 690, а значит пропусков в данных нет. Помимо этого, в таблице нет категориальных признаков.

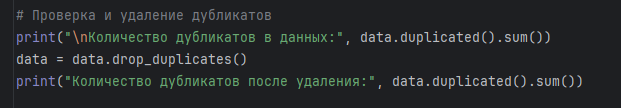
Приведём целевой параметр к булевой переменной и уменьшим размер датасета.

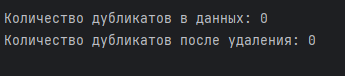




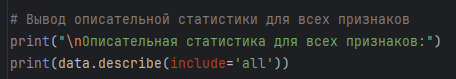


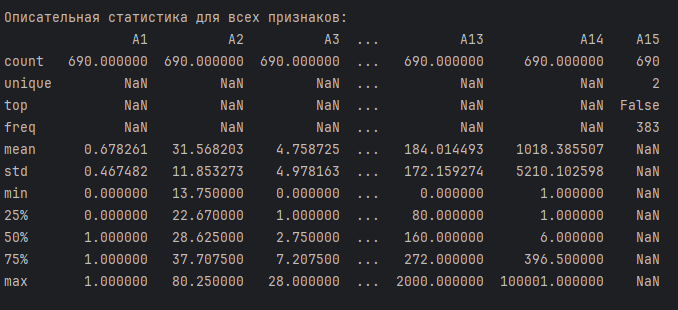
Мы уже знаем, что пропусков в данных нет, а есть ли дубликаты? Проверим это с помощью duplicated.





Теперь рассмотрим общие сведения о каждом параметре по отдельности.

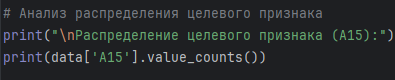


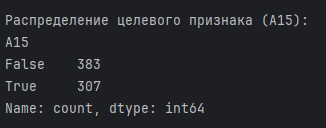


Мы видим средние значения, стандартное отклонение, минимальные и максимальные значения для числовых признаков.

По этим данным сложно определить, какие параметры влияют на определение принадлежности класса, поэтому рассмотрим параметры для каждого класса по отдельности.

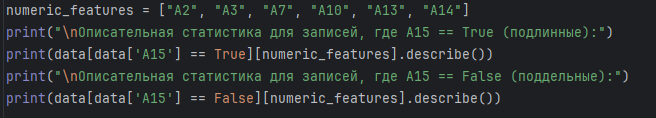
Сначала определим отношение одобренных кредитных карт к неодобренным.

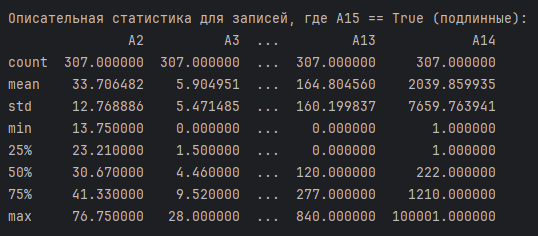


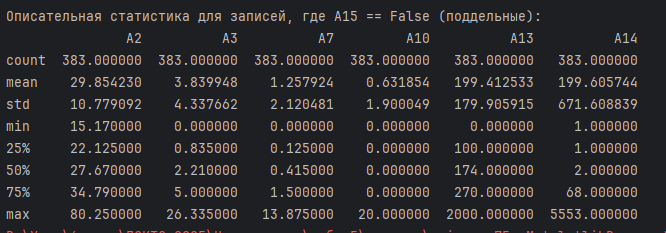


Отношение близко к ½, что очень хорошо.

Детальнее исследуем каждый класс по отдельности.







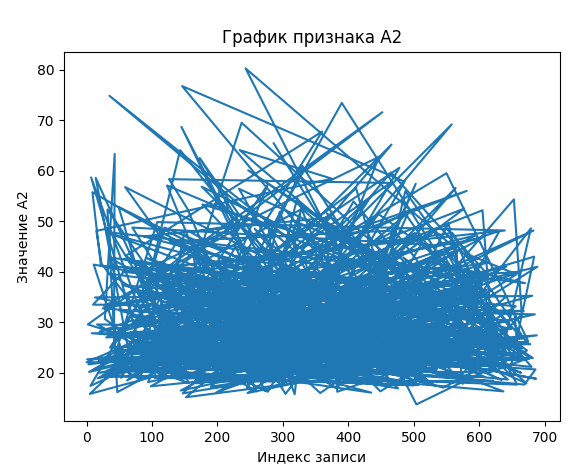
False означает, что кредитная карта одобрена.

True означает, что кредитная карта неодобрена.

Теперь лучше видно, что, например, медиана для неодобренных заявок равна 222, а среднее значение достигает около 2039.86, что значительно превышает показатели одобренных заявок (медиана – 2, среднее – примерно 199.61)..

Теперь построим различные графики для лучшего понимания картины.

Построим график одного из признаков



Представим некоторые признаки на графике в виде точек.

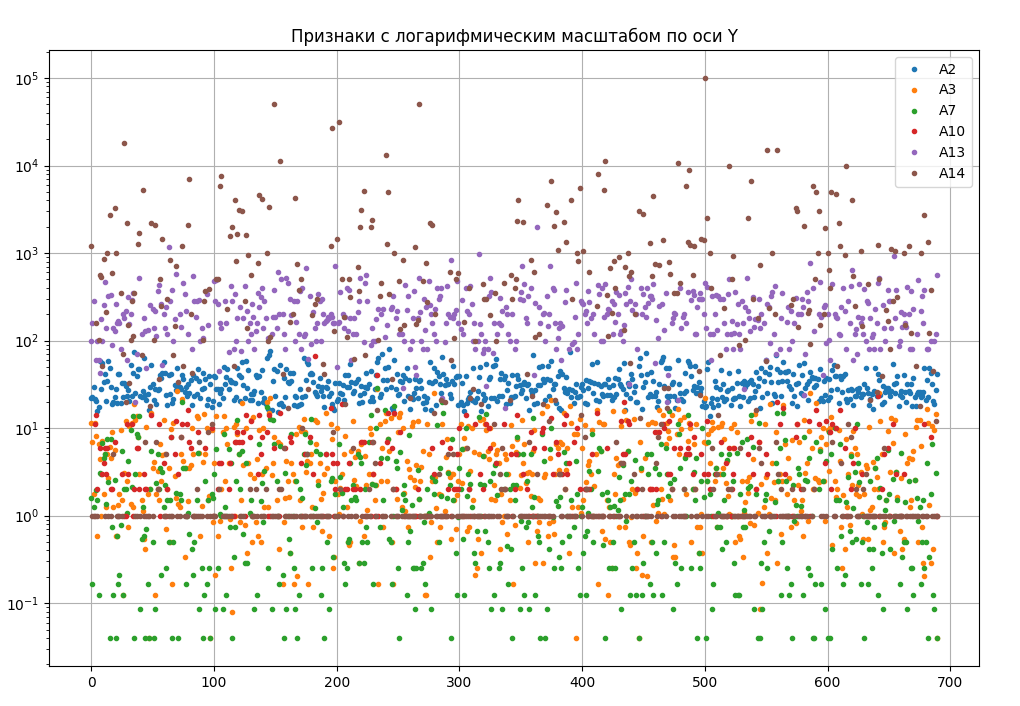
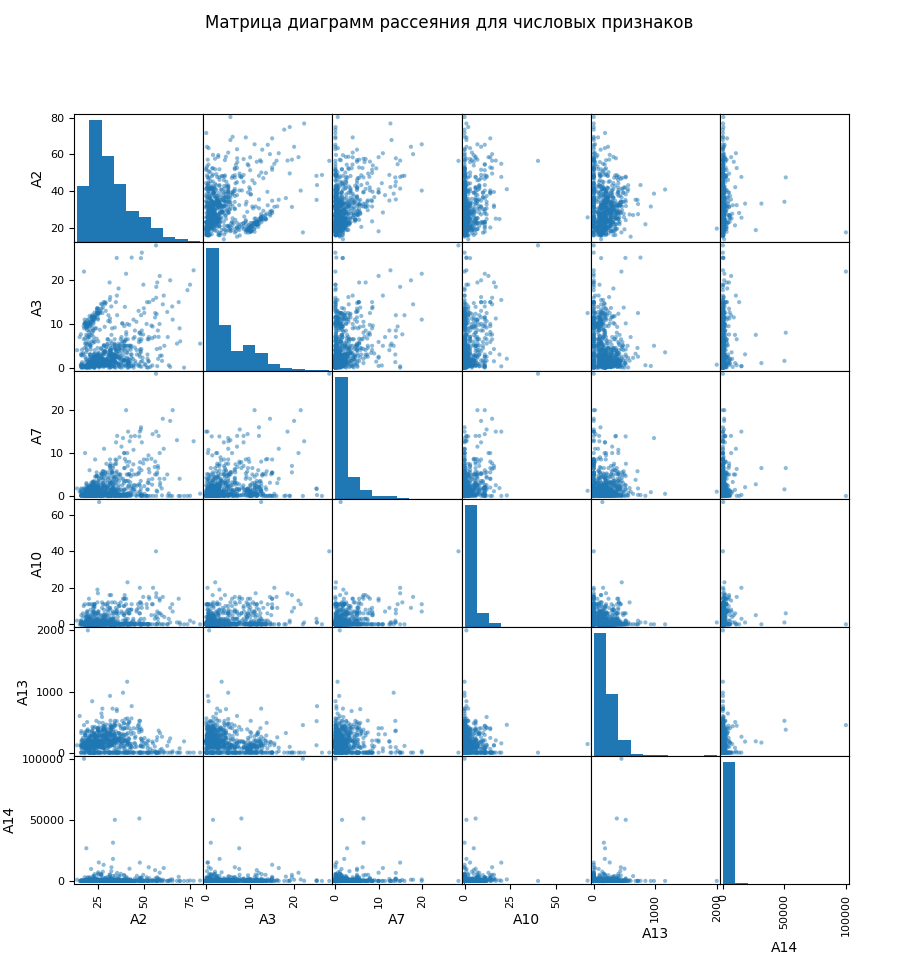
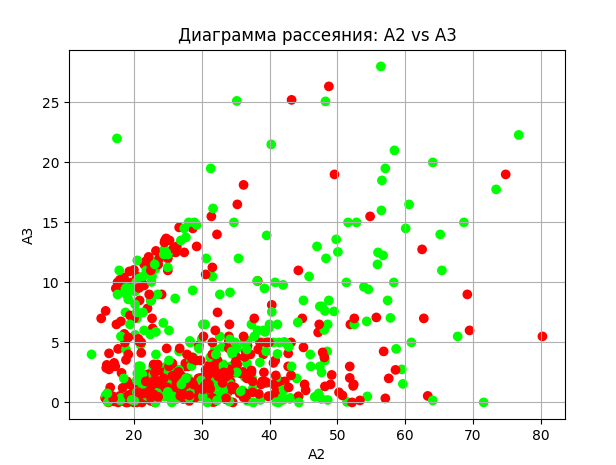
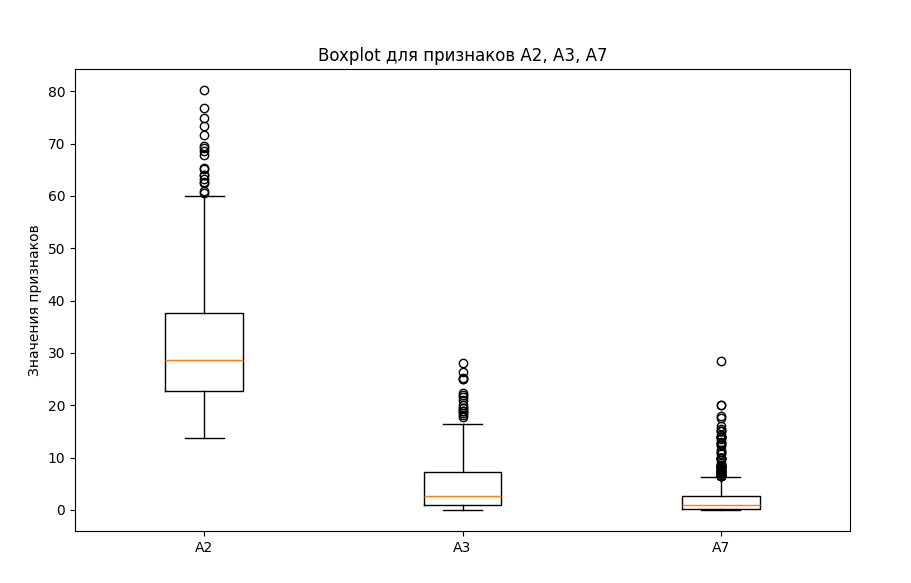


Диаграмма рассеяния, или точечная диаграмма:



Точечная диаграмма, на которой каждый класс объектов раскрашен в свой цвет.

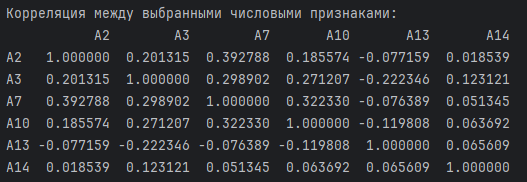
  
Диаграмма размаха



A2 варьируется значительно сильнее, чем A3 и A7, и имеет множество выбросов, что говорит о том, что часть объектов сильно выделяется по данному признаку. A3 и A7 в целом принимают меньшие значения и более «сжаты». При этом A3 имеет более выраженный «хвост» вверх, а A7 обычно находится около 0 или 1, что может означать скошенность распределения.

Наличие выбросов во всех признаках (особенно в A2) указывает на необходимость нормализации или логарифмического преобразования данных.

Напоследок рассмотрим корреляцию между признаками.



A2 и A14 (0.018), A7 и A14 (0.051) – практически нулевая корреляция, эти признаки слабо связаны.

Нет признаков с очень высокой корреляцией (> 0.8), что означает отсутствие явных дублирующих параметров.

Признаки A2, A3, A7 и A10 проявляют умеренную связь между собой, что может указывать на потенциальную зависимость возрастных и финансовых показателей.

Признак A13 имеет слабые отрицательные связи с несколькими другими признаками, что может говорить о его особой роли (например, как показатель долга, который снижается с опытом или возрастом).

A14 почти ни с чем не коррелирует, значит, он может быть независимым фактором.

**Вывод**

В результате выполнения лабораторной работы изучили набор данных