Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

|  |
| --- |
| Институт информационных технологий и анализа данных |

наименование института

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 1

по дисциплине:

|  |
| --- |
| **Методы анализа данных** |
| **«Подготовка данных для анализа»** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Выполнил | АСУб-20-2 |  |  |  | Арбакова А.В. |
|  | шифр группы |  | подпись |  | Фамилия И.О. |
| Проверил |  |  |  |  | Бучнев О.С. |
|  | должность |  | подпись |  | Фамилия И.О. |

Иркутск 2022 г.

**Содержание**

[Введение 3](#_Toc116638193)

[Ход выполнения работы 4](#_Toc116638194)

[Заключение 8](#_Toc116638195)

[Список литературы 9](#_Toc116638196)

# **Введение**

**Цель лабораторной работы:**

1. Выбрать среду программирования для языка Python. Используя в качестве источника данных портал https://www.kaggle.com/datasets, загрузить набор данных.

2. С использованием Python выполнить подготовку данных для анализа. При этом необходимо:

* описать исходные данные;
* выполнить анализ типов шкал, в которых измерены данные (количественные, качественные /номинальные, дихотомические, порядковые/);
* определить группирующие переменные и переменные отклика;
* выявить наличие пропущенных значений. В случае если пропущенные значения найдены, для заполнения применить методы п. 1.6;
* для выявления выбросов для каждой пары числовых признаков построить диаграммы Кливленда (достаточно 4 признаков). Выделить наблюдения, похожие на выбросы;
* получить значения основных показателей описательной статистики для признаков, измеренных в количественной шкале, построить для них гистограммы частот;
* построить диаграмму размаха для каждого признака, измеренного в количественной шкале;
* выполнить стандартизацию данных, измеренных в количественной шкале, и повторно построить диаграмму размаха для каждого признака;
* для каждой пары числовых признаков построить диаграмму рассеяния.

3. Оформить отчет, в который включить все результаты анализа с описанием результатов.

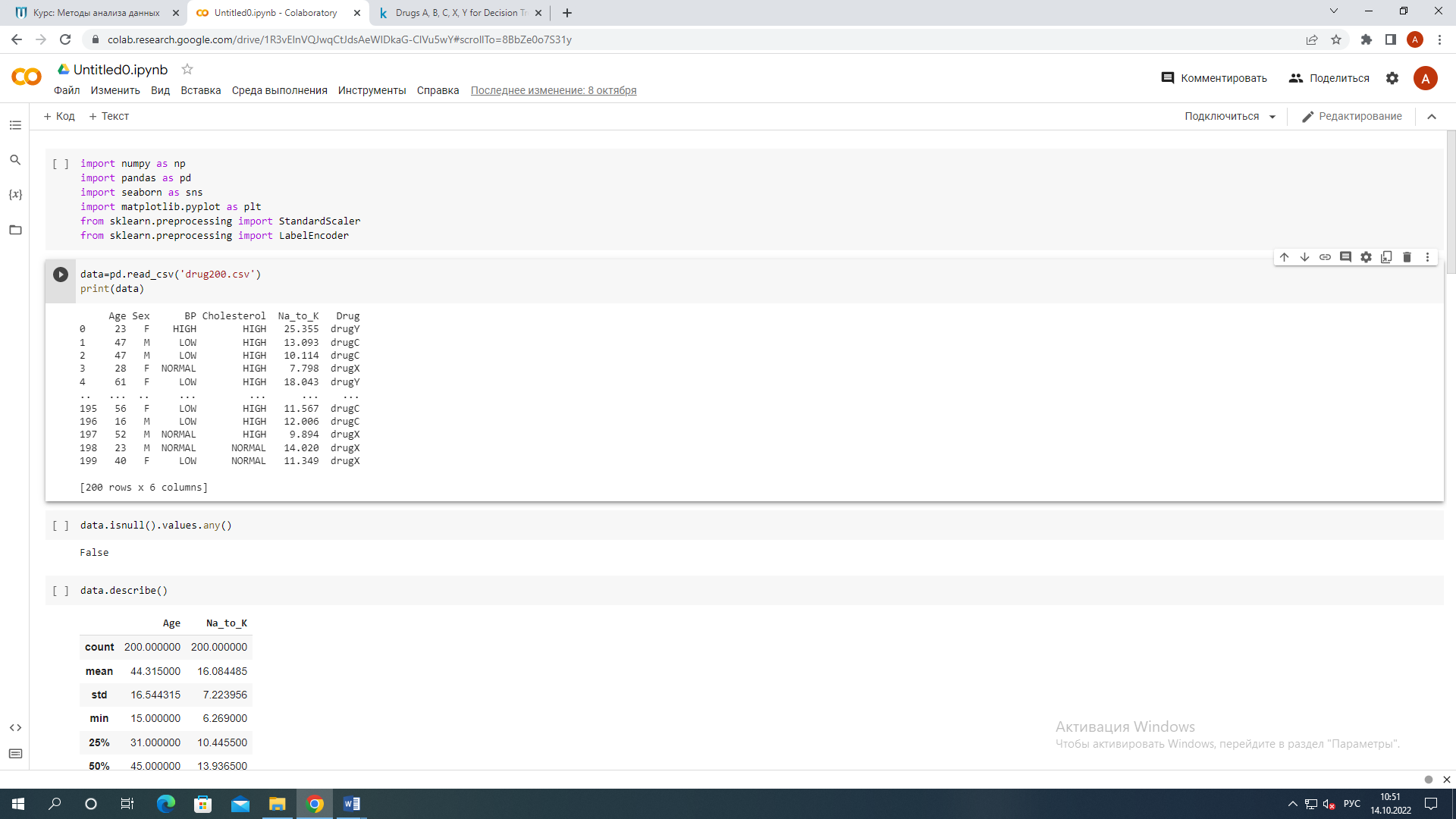
**Язык программирования:** Python

# **Ход выполнения работы**

Перед выполнением лабораторной работывыоплним подготовку данных. Подготовка данных – это процесс подготовки необработанных данных для их дальнейшей обработки и анализа.

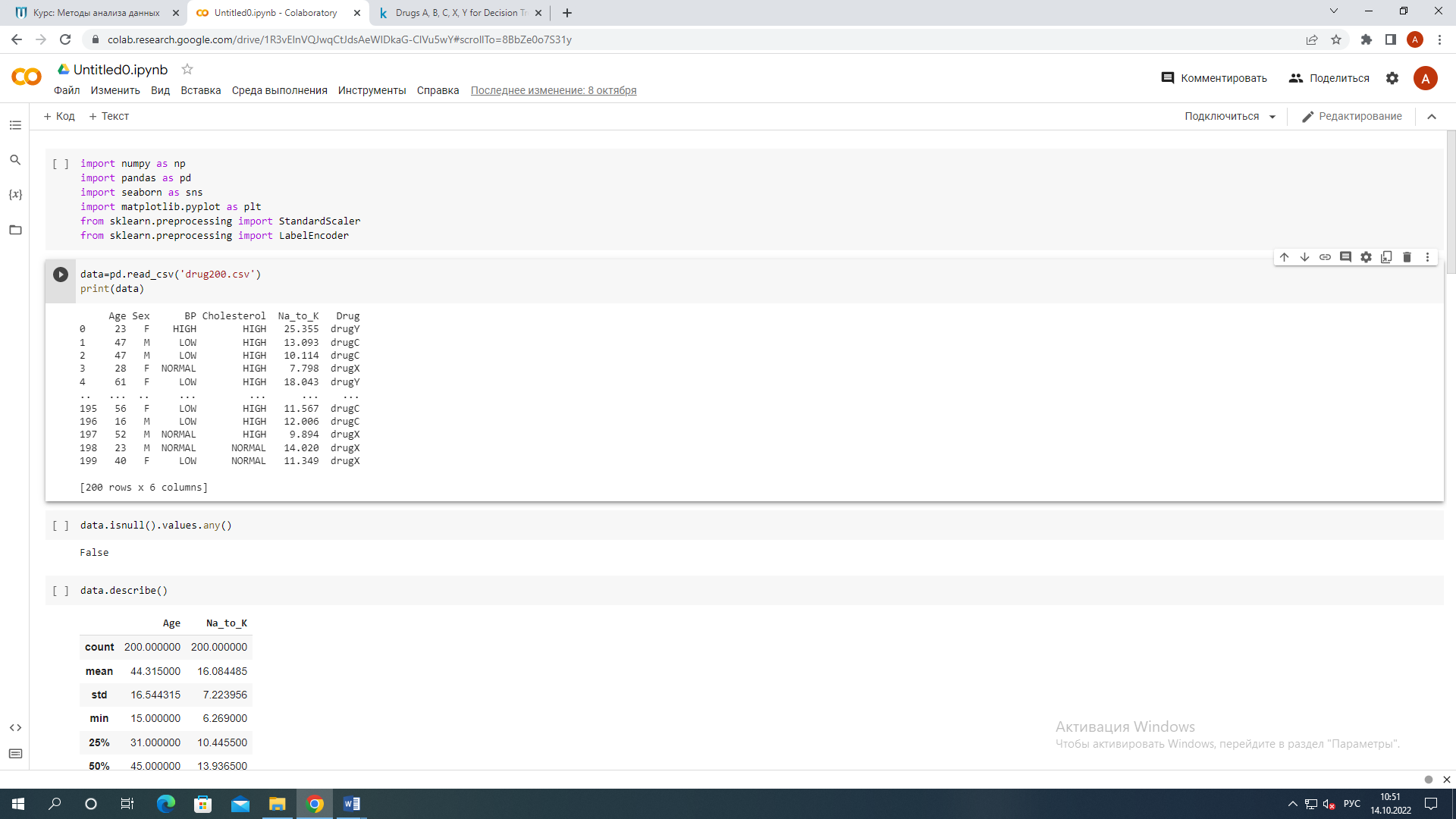
Для выполнения лабораторной работы были использованы данных из Excel-файла drug200.crv. В файле 200 наблюдений о группе пациентов, все из которых страдали одним и тем же заболеванием. Во время курса лечения каждый пациент реагировал на одно из 5 лекарств: препарат А, препарат В, препарат C, препарат X и Y. Особенностями этого набора данных являются возраст, пол, артериальное давление и уровень холестерина пациентов.

Перед выполнением лабораторной работы подключим необходимые библиотеки.



Изображение 1 – Библиотеки языка Python

Загрузим данные о пациентах.



Изображение 2 – Загруженные данные из файла Excel

Разведочный анализ данных ­(РАД) – анализ основных свойств данных, нахождение в них закономерностей, распределений и аномалий, построение начальных моделей. Основная задача РАД ­ преход к компактному описанию при более полном сохранении существенных аспектов информации, содержащихся в исходных данных. Разведочный анализ данных ­(РАД) применяется для нахождения связей между переменными в ситуациях, когда отсутсвуют (или недостаточны) априорные представления о природе этих связей.

РАД включает в себя применение различных процедур: вычисление основных статистических характеристи, преобразование переменных, сокращение размерности с использованием главных компонент, метода класификации и т.д.

Этапы РАД:

1. Формулировка исследовательской гипотезы. Выполнение экспериментов/наблюдений анализа данных.
2. Разведочный анализ данных:
   1. Выявление точек выбросов
   2. Проверка однородности дисперсий
   3. Проверка нормальности распределения данных
   4. Выявление избыточного числа нулевых значений
   5. Выявление коллинеарных переменных
   6. Выявление характера связи между анализируемыми переменными
   7. Выявление взаимодействий между переменными
   8. Выявление пространственно-временных корреляций среди значений зависимой переменной.
3. Применение статистического метода, соотвествующего выявленным закономерностям в структуре экспериментальных данных.

Типы шкал, в которых измерены значения признаков:

1. Номинальная (категориальная) задается для переменных, значения для которых являются условными именами.

Разновидностью номинальной шкалы является дихотомическая шкала – переменных могут принимать только два заначения.

1. Порядковая (ранговая) – значения можно сравнивать между собой, но величина разности значения не имеет.
2. Количественная шкала позволяет отражать количественные характеристики исследуемых объектов.

Номинальная и порядковая шкалы являются качественнми шкалами.

В подготовленной двумерной таблице с данными в строках содержатся объекты (наблюдения), а в столбцах значения признаков каждого объекта выделяют зависимые и независимые признаки.

Независимые признаки — это переменные, которые используются для объяснения или предсказания значений других признаков, а именно независимых переменных.

Зависимые признаки — это признаки, значения которых зависят от значения независимых признаков, и даже могут определяться ими.

Признаки делят на группирующие переменные и переменные отклика:

Группирующие переменные являются независимыми и считаются факторами, значения которых влияют на значения зависимых переменных — отклика.

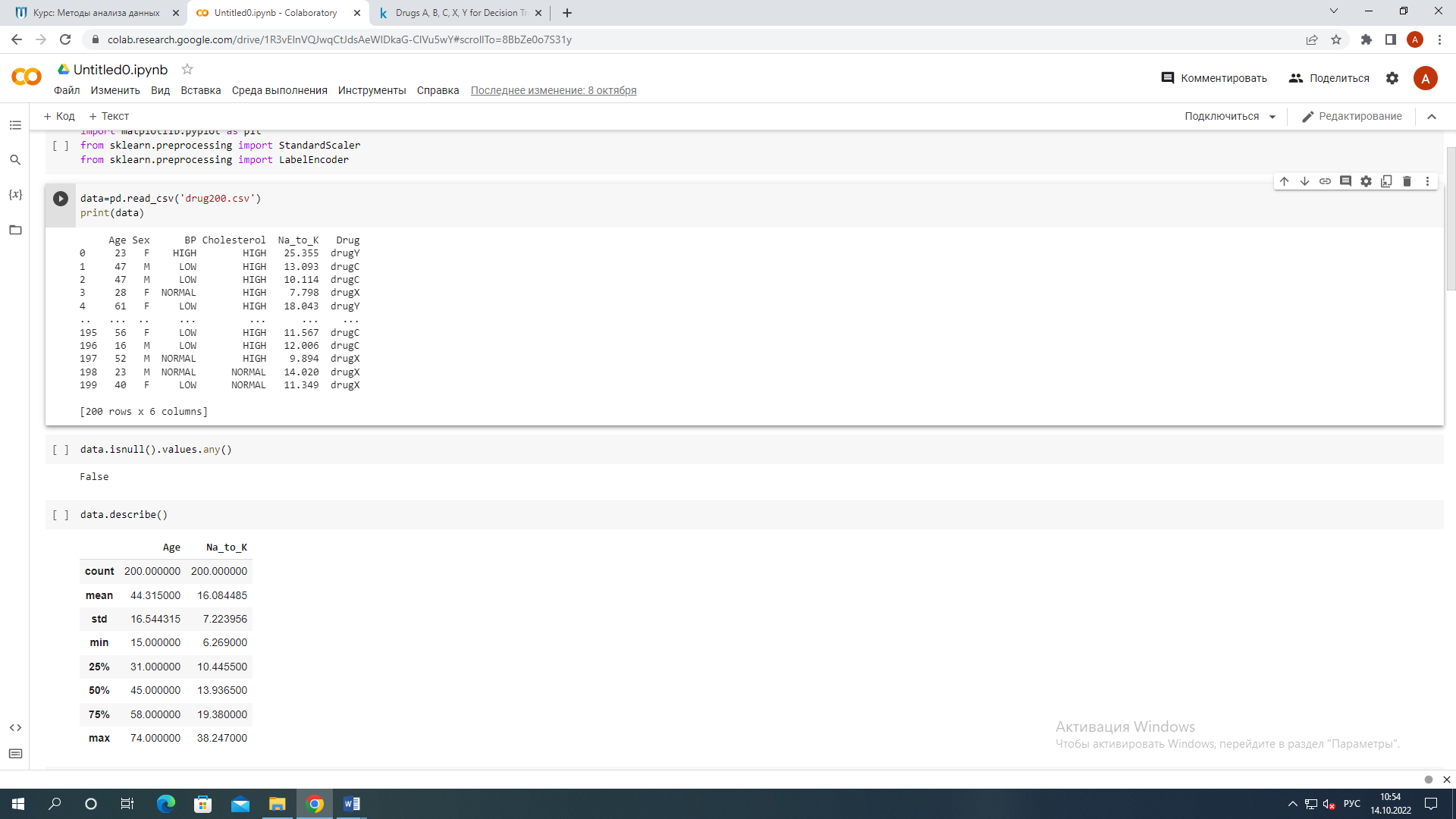
После того как данные сведены в таблицу, необходимо убедиться в том, что содержимое таблицы не исказит результаты анализа. Такое может

произойти, например, если в каком-либо столбце с числовыми данными

есть выбросы (как результат так называемого «человеческого фактора»),

или в силу каких-либо объективных причин в данных появились пропущенные значения.

Проверим, есть ли в имеющихся данных пропущенные значения.

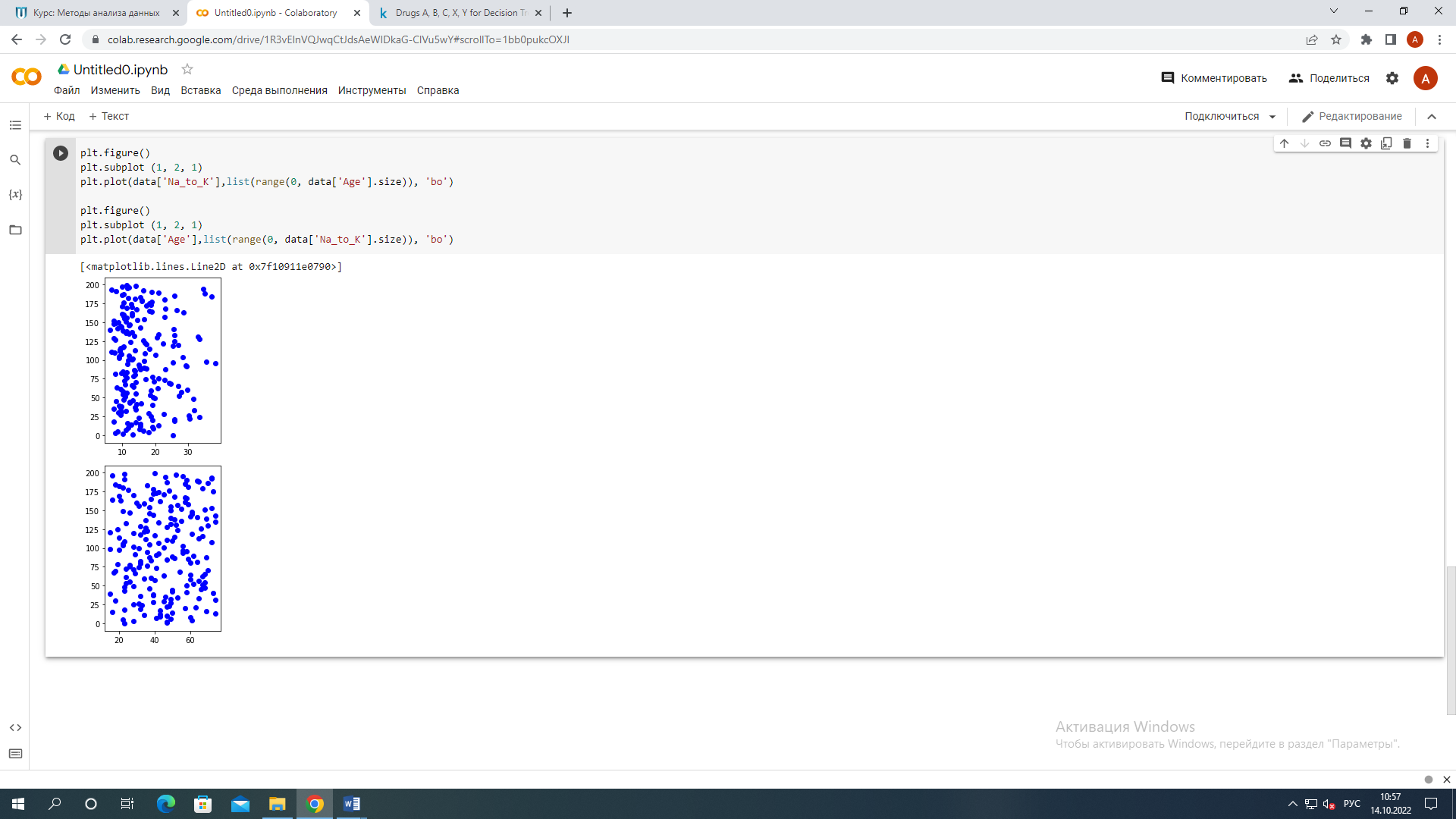


Изображение 3 – Проверка на пропущенные значения

Как мы видим, пропущенные значения отсутствуют.

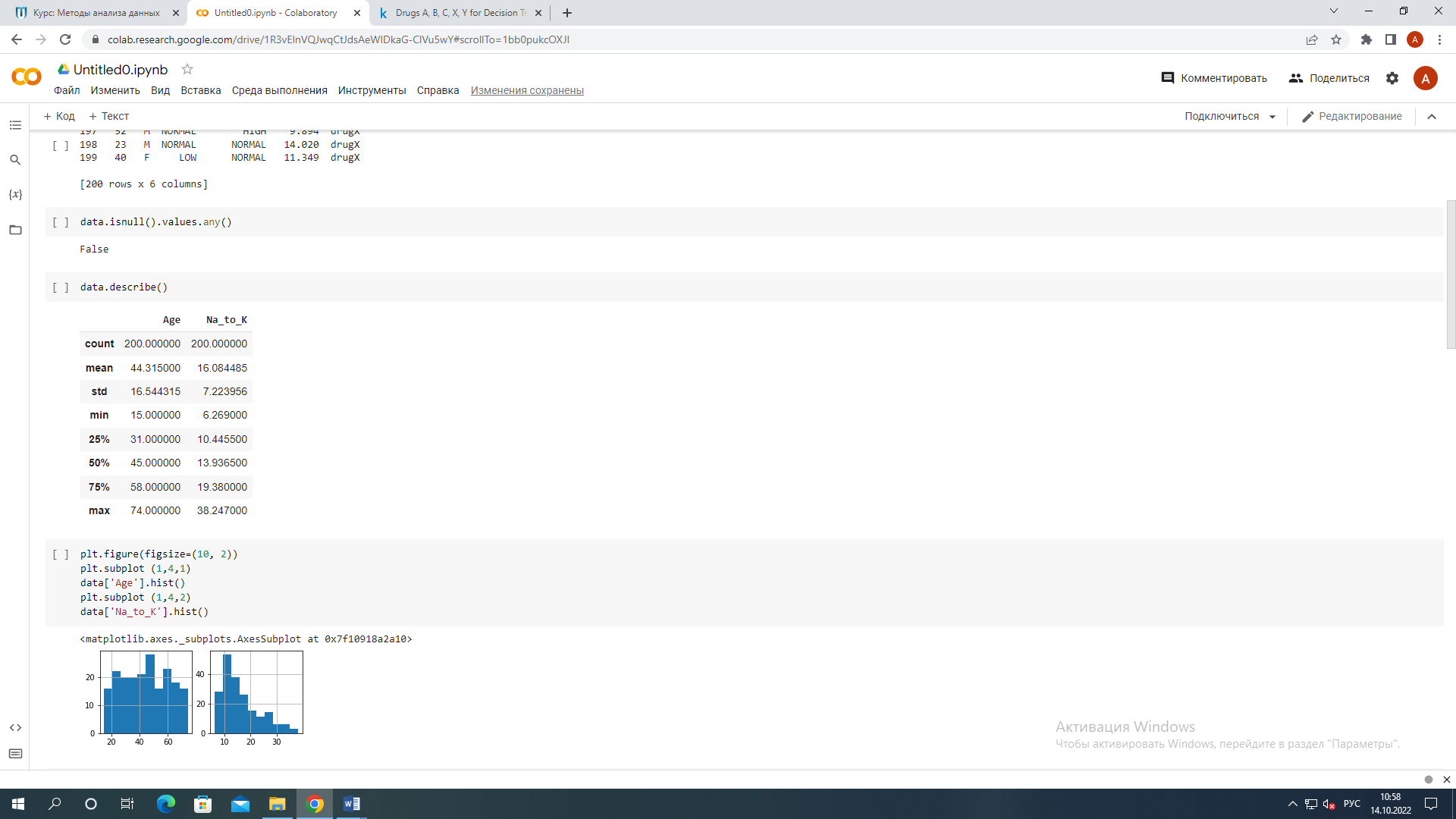
Точечные диаграммы Кливленда представляют собой графики, на которых точки используются для отображения значений некоторой количественной переменной (или переменных), разбитых на группы в соответствии с уровнями некоторой номинальной переменной (или переменных). Этот инструмент графического анализа данных получил свое название в честь предложившего его У. Кливленда.

Далее для обнаружения выбросов построим диаграммы Кливленда для пар признаков: Na\_to\_K – Age и Age – Na\_to\_K.



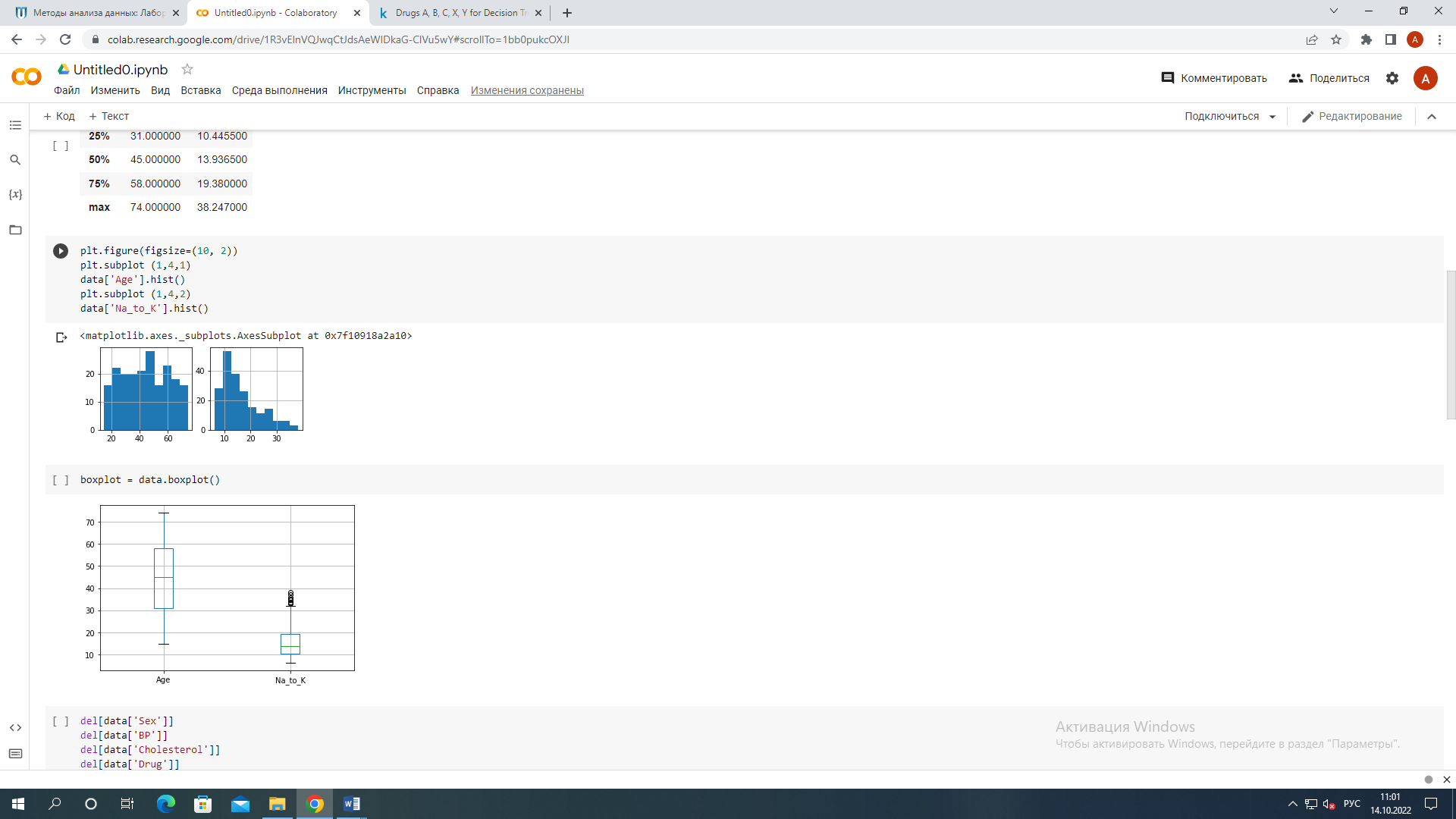
Изображение 4 – Диаграммы Кливленда

Следом получим значения основных показателей описательной статистики.



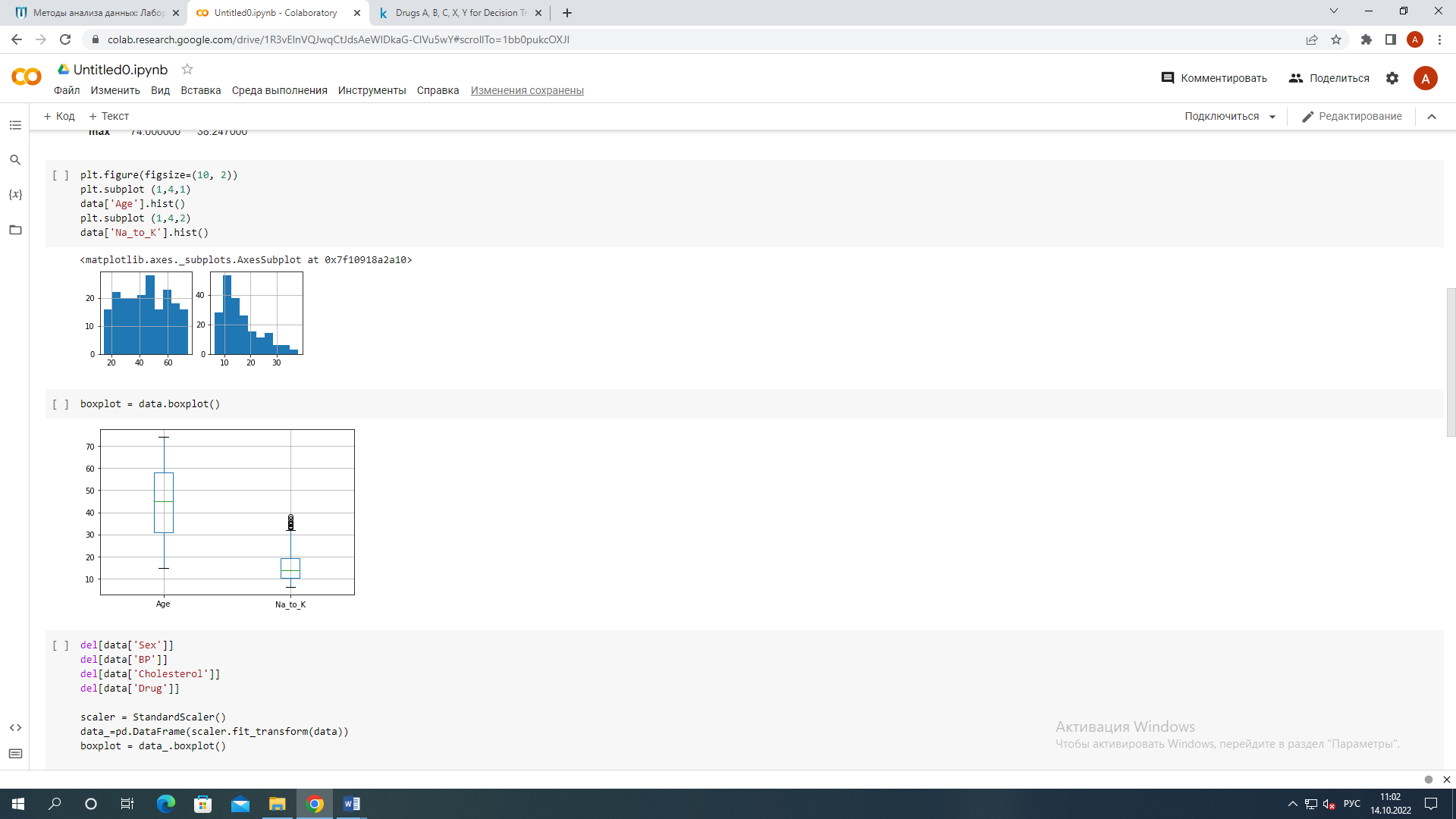
Изображение 5 – Основные показатели описательной статистики

Построим гистограмму частот.



Изображение 6 – Гистограмма частот

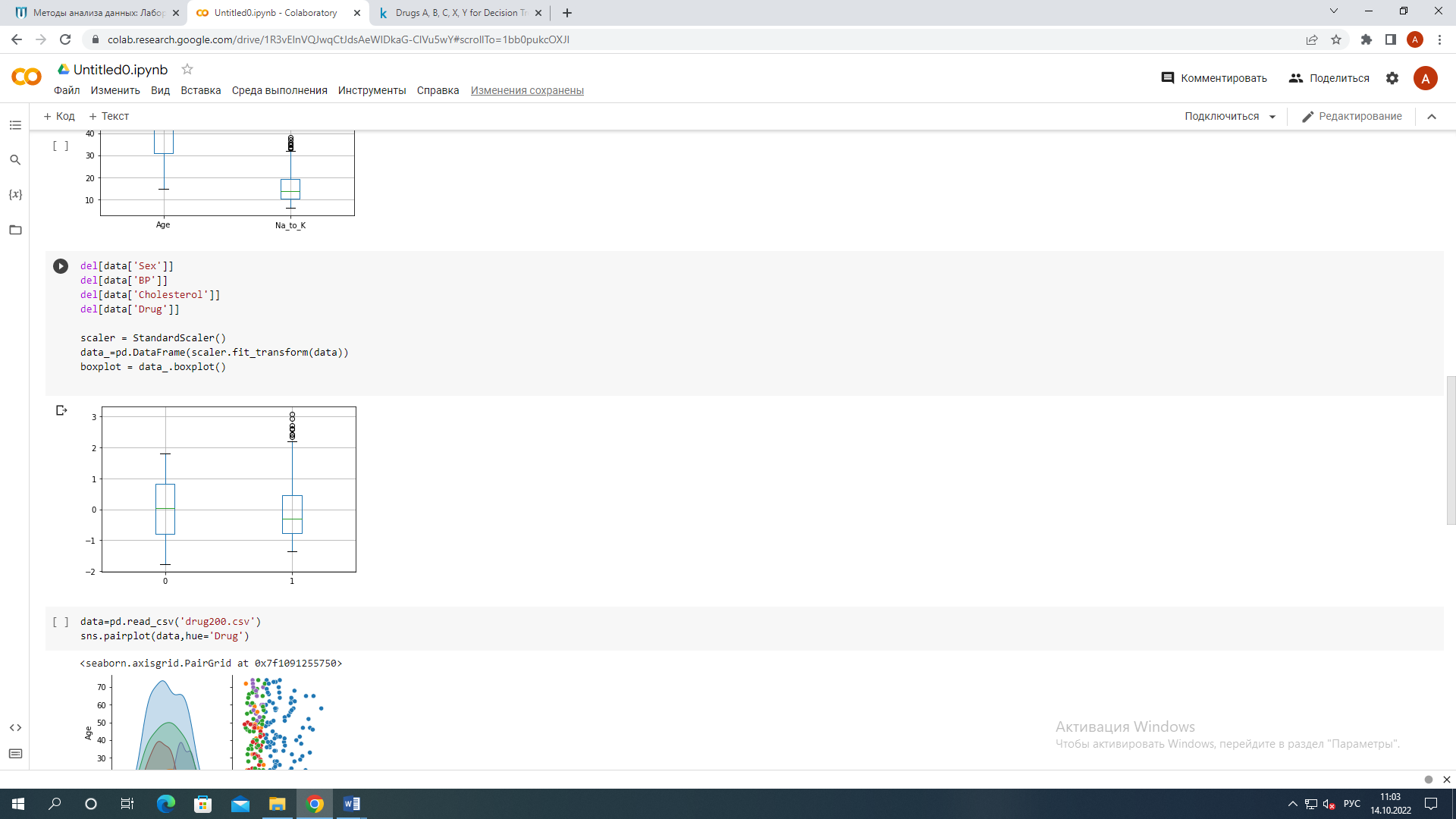
Построим диаграммы размаха для признаков, измеренных в количественной шкале.



Изображение 7 – Диаграмма размаха

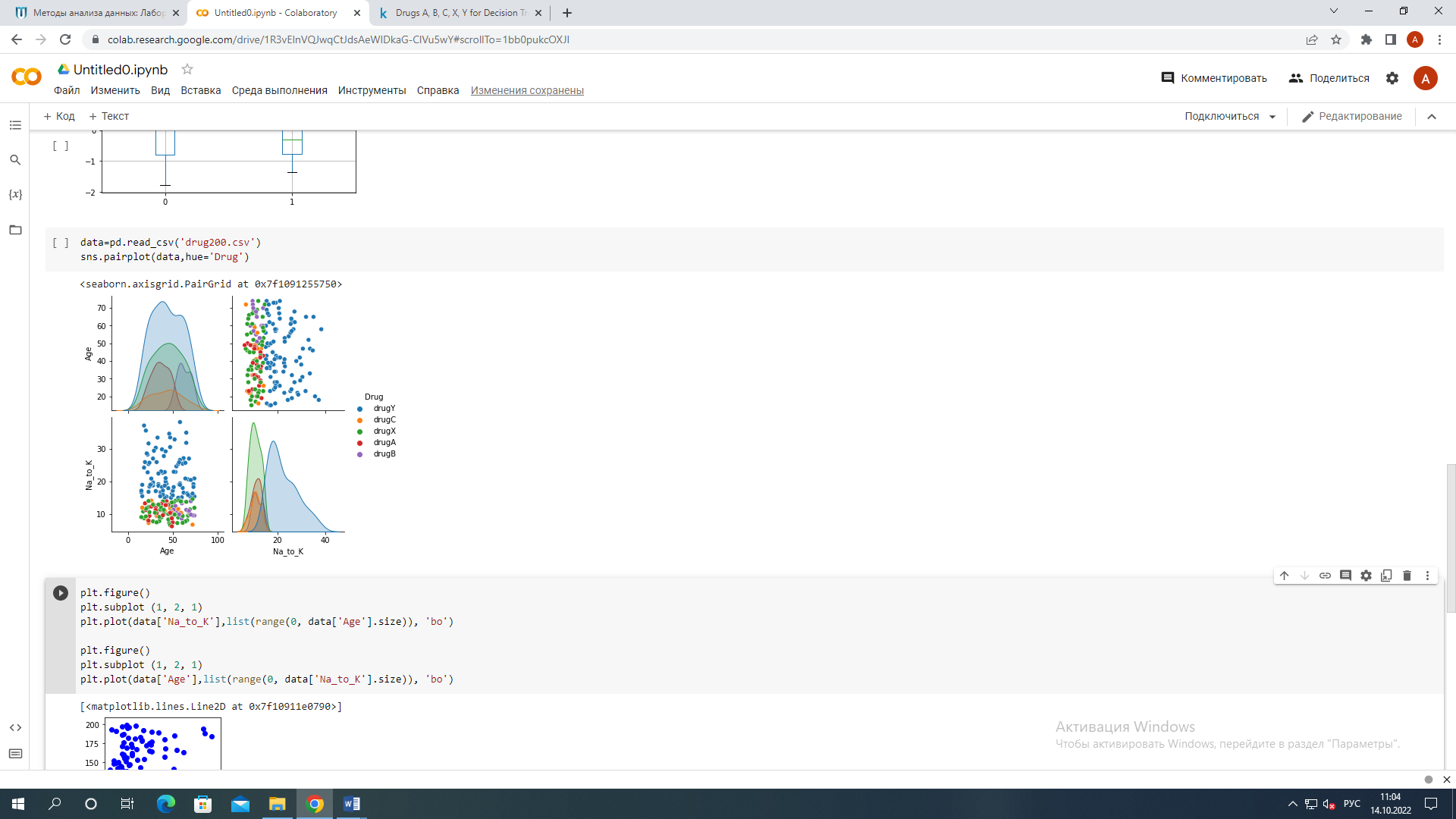
По диаграмме размаха можно определить наличие или отсутствие выбросов.

Выполним нормализацию данных и повторно построим диаграмму размаха.



Изображение 8 – Диаграмма размаха после нормализации данных

Для каждой пары количественных признаков построим диаграмму рассеяния. Она позволит оценить меру и вид зависимости между признаками, а так же выделить подмножества.



Изображение 9 – Диаграмма рассеяния

# **Заключение**

В ходе выполнения лабораторной работы по теме «Подготовка данных для анализа» была выполнена цель работы по ознакомлению с программными средствами статистического анализа и обработке наблюдений.

Было изучены такие темы как: сбор и подготовка данных, источники данных, разведочный анализ данных, основные этапы, структура данных, запись в таблице, объекты и признаки, группирующие переменные и переменные отклика, типы шкал, в которых могут быть измерены признаки, особенности каждой шкалы, актуальность проблемы пропущенных значений, способы решения, реализованные в библиотеке Pandas, выбросы, диаграмма размаха, диаграмма Кливленда, описательная статистика, основные показатели, диаграмма рассеяния.

Знания были закреплены во время использования среды разработки Google Colab и использования языка программирования Python, в котором было выполнено задание лабораторной работы.

# **Список литературы**

1. Бучнев О. С. Методы анализа данных: лабораторный практикум – Иркутск, 2022. – 115 с
2. <https://colab.research.google.com>
3. С.Э. Мастицкий, В.К. Шитиков Статистический анализ и визуализация данных с помощью языка R, Хайдельберг – Лондон – Тольятти, 2014 г. – 401 с
4. https://www.kaggle.com/datasets