**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**

**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**ЛАБОРАТОРНАЯ РАБОТА**

**по дисциплине «Машинное обучение»**

**Тема: “Подготовка и нормализация данных”**

| Студент |  |  |
| --- | --- | --- |
| Преподаватель |  | Гатауллин Р. И. |

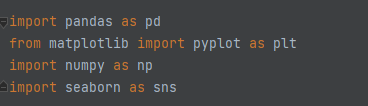
Санкт-Петербург

2023

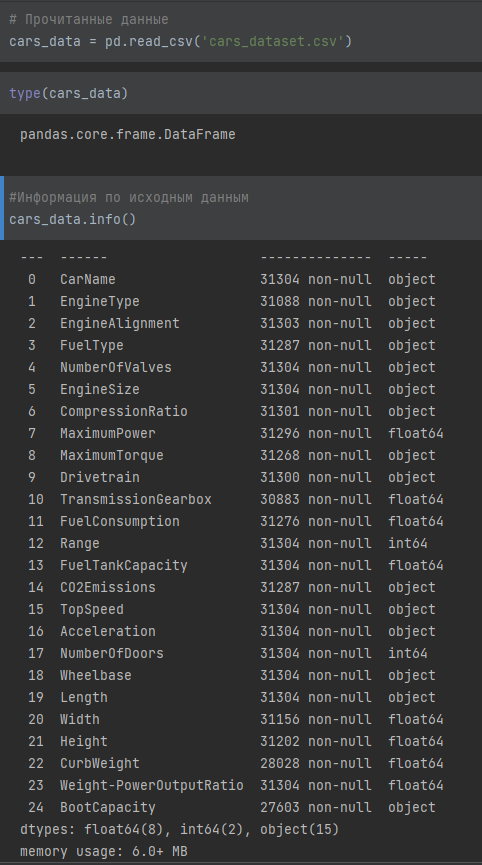
Датасет: <https://www.kaggle.com/datasets/shrirangmhalgi/world-cars-data>

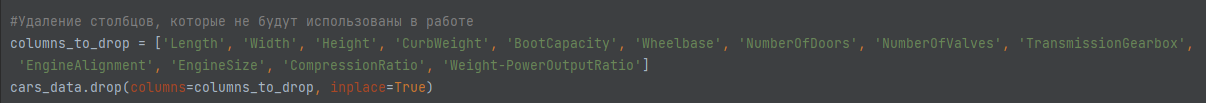
Код программы: <https://drive.google.com/file/d/1NwxzOkst6-VTqK7MFvsoT_aztz6s4RWY/view?usp=sharing>

Используемые библиотеки

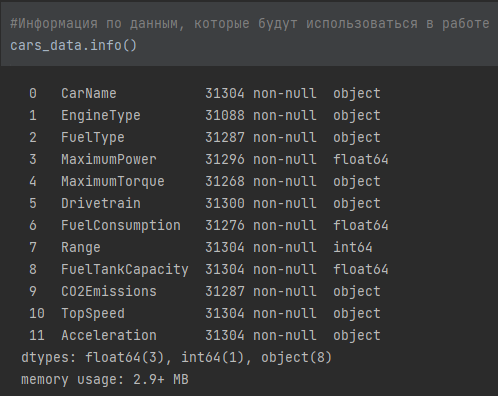


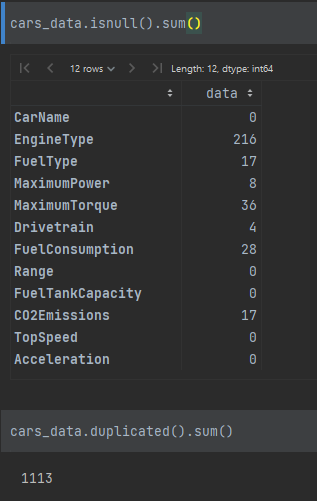
Входные данные



Удаление столбцов входных данных, которые не будут использованы в работе

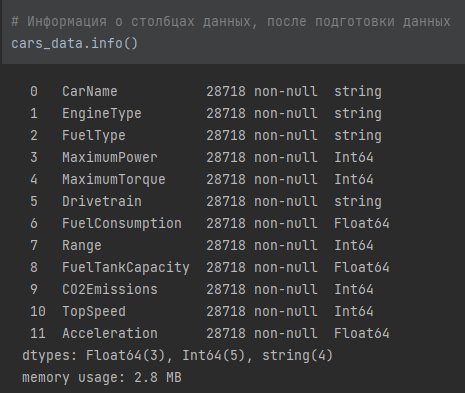
Данные используемые в работе

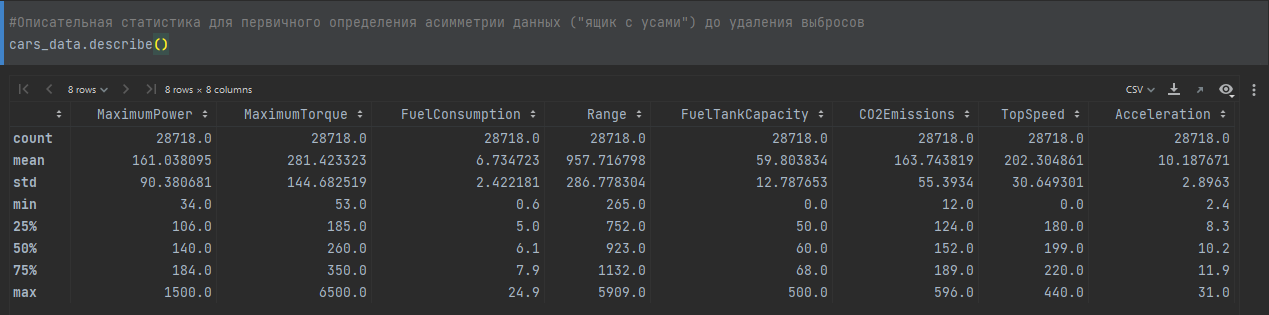


Кол-во строк с незаполненными данными и дубликатами

Подготовка входных данных для исследования



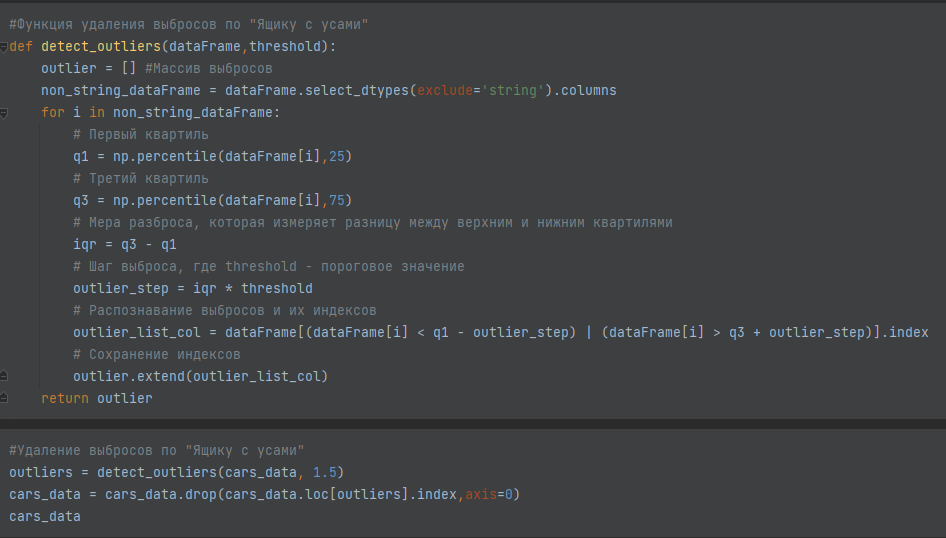


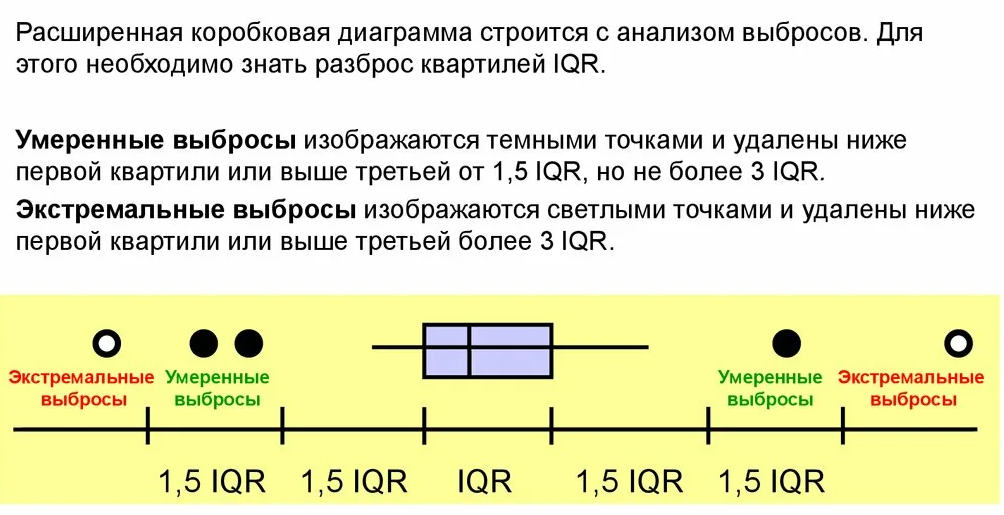
Описательная статистика 

Описательная статистика показывает, что поля *MaximumPower*, *MaximumTorque*, *FuelConsumption*, *Range*, *CO2Emissions*, *TopSpeed*, вероятно, имеют правую асимметрию, поскольку средние значения (строка - mean) этих атрибутов больше, чем медианные (строка - 50%).

С другой стороны, *FuelTankCapacity*, *Acceleration* вероятно, симметричен с небольшим отклонением.

Удаление выбросов





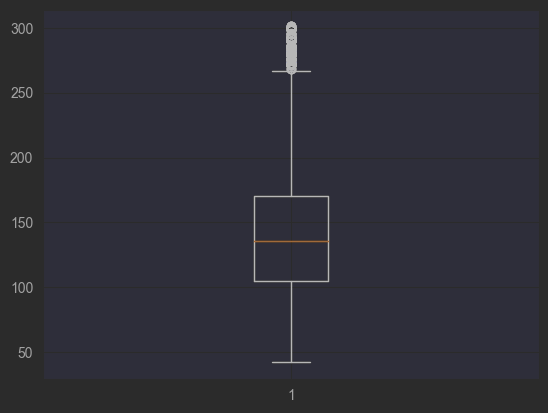
“Ящики с усами” до и после удаления выбросов

*Максимальная мощность*

*До*

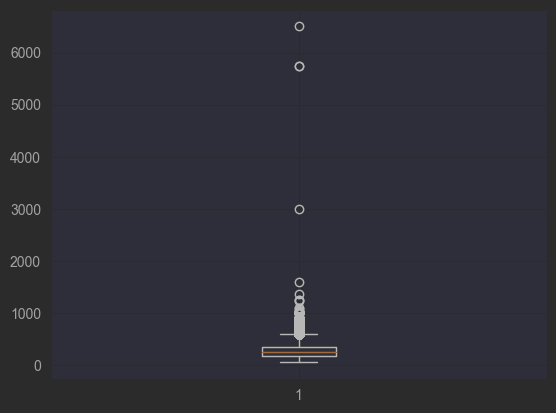


*После*

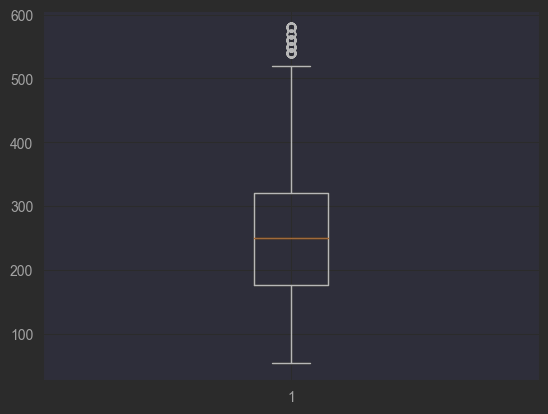
**

*Максимальный крутящий момент*

*До*

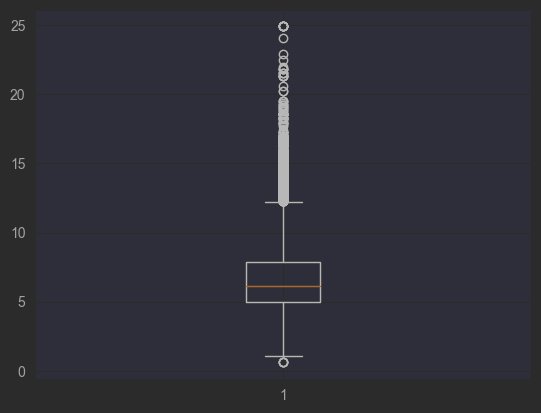
**

*После*

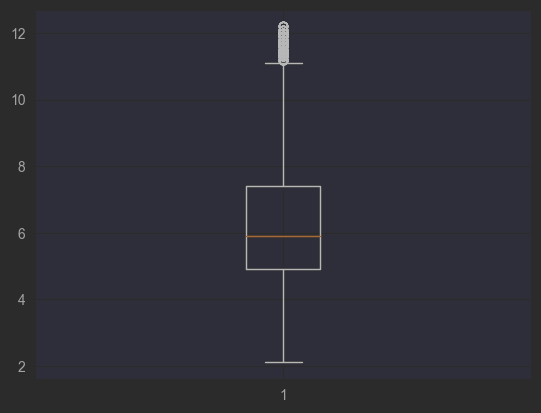
**

*Расход топлива*

*До*

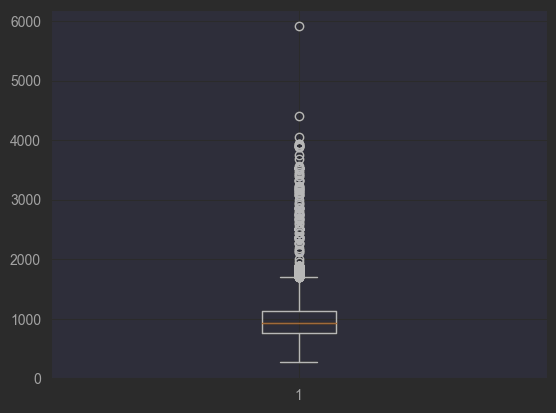
**

*После*

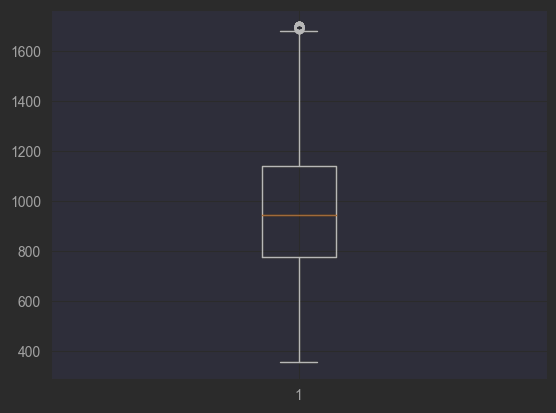
**

*Максимальный километраж на полном баке*

*До*

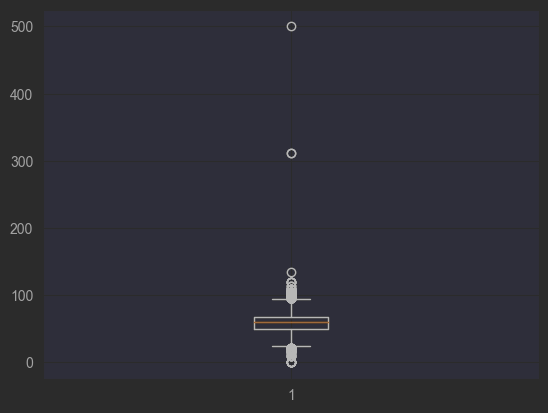
**

*После*

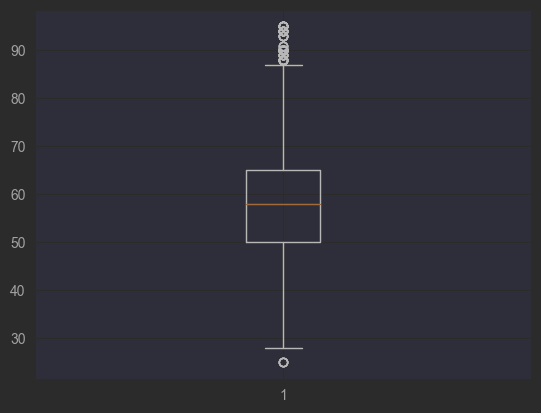
**

*Вместимость бака*

*До*

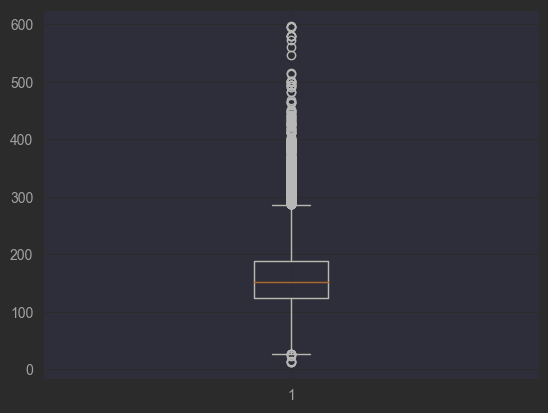
**

*После*

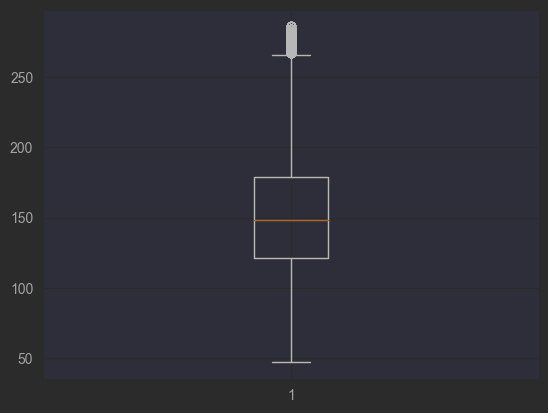
**

*Выбросы CO2*

*До*

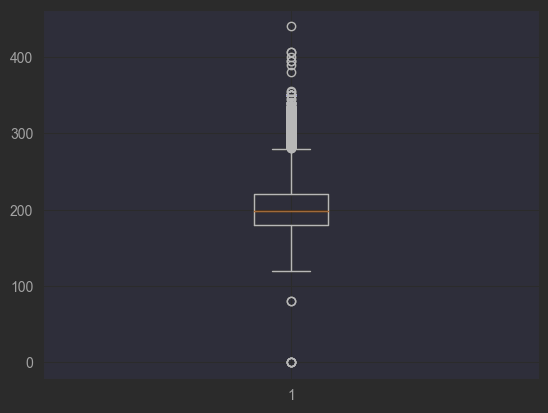
**

*После*

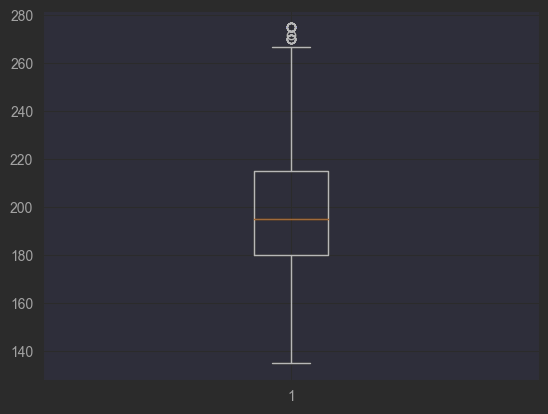
**

*Максимальная скорость*

*До*

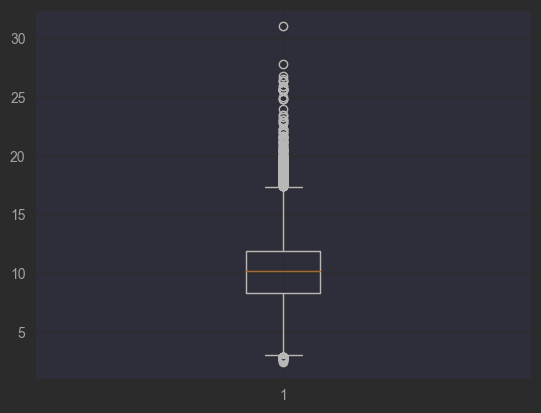
**

*После*

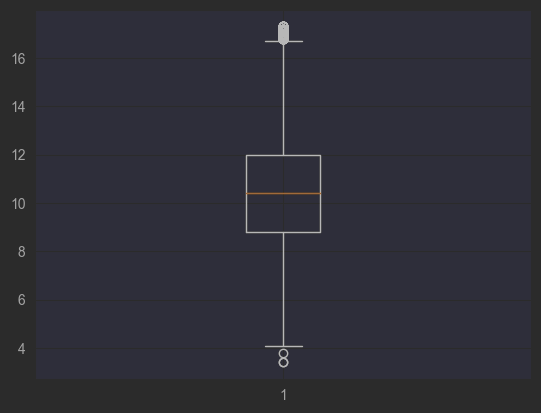
**

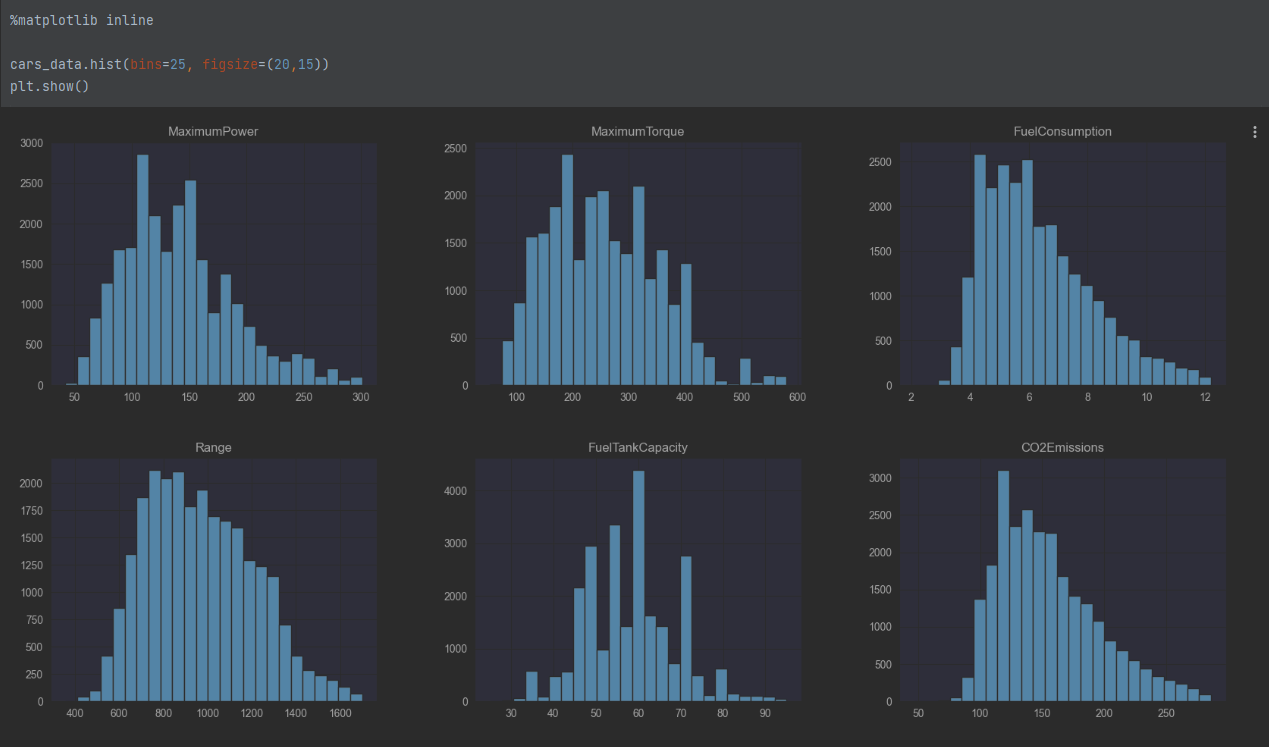
*Ускорение 0-100км/ч*

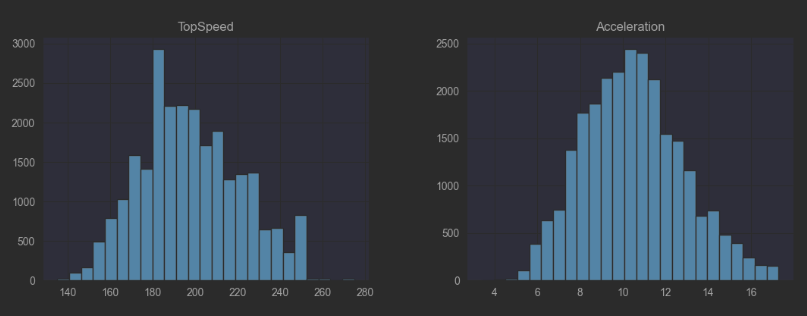
*До*

**

*После*

**

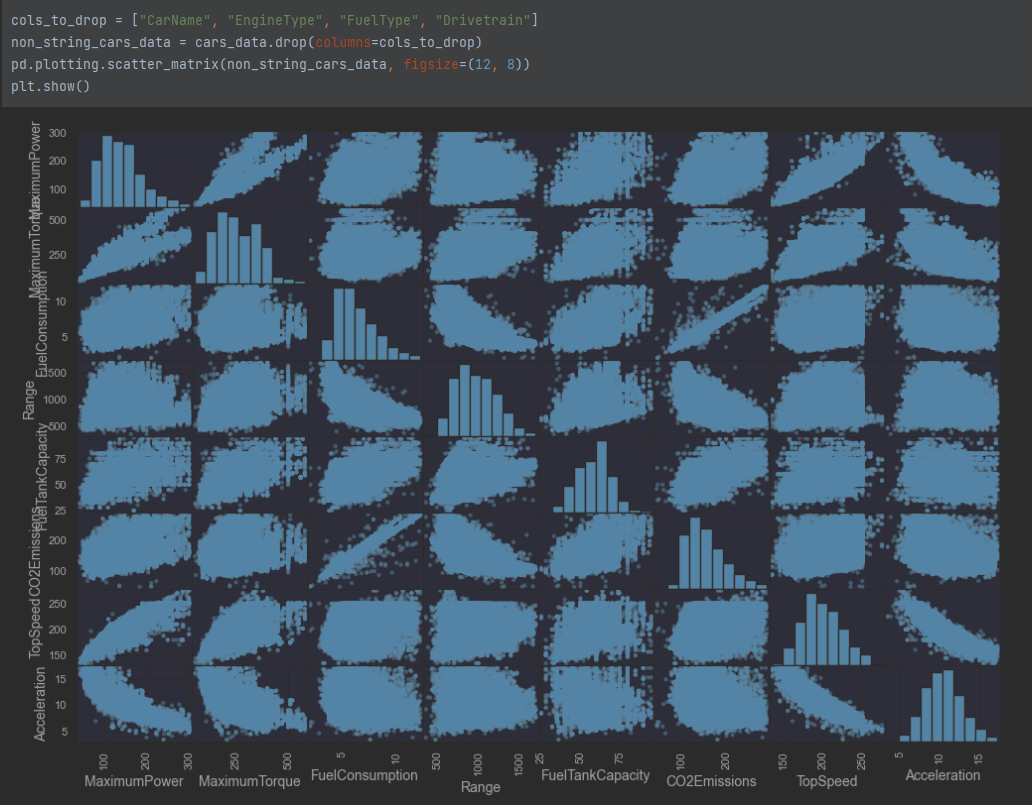
Гистограммы 



По данным диаграммам можно сделать выводы и подтвердить предположения по описательной статистике.

Как и предполагалось поля *MaximumPower*, *MaximumTorque*, *FuelConsumption*, *Range*, *CO2Emissions*, *TopSpeed* имеют правую асимметрию.

А поля *FuelTankCapacity*, *Acceleration -* симметричны, однако есть небольшое отклонение.

Диаграммы рассеяния 

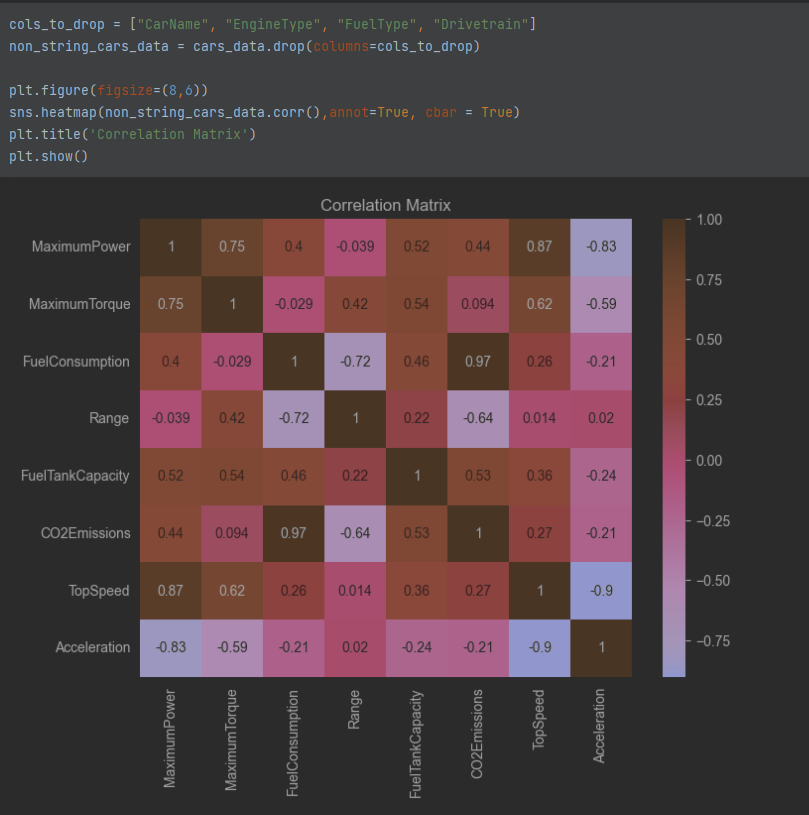
На данных диаграммах рассеяния можно наблюдать много разных корреляций, как положительных так и отрицательных, однако можно точно выделить несколько зависимостей:

*CO2Emissions/FuelConsumption* - в этом случае, корреляция положительна, как и в случае *MaximumTorque/MaximumPower* и *TopSpeed/MaximumPower,* действительно: чем больше расход топлива, тем больше выбросы CO2 и чем больше мощность машины, тем больше крутящего момента она имеет и тем больше ее максимальная скорость (в общем случае).

Также можно выделить несколько отрицательных корреляций:

*Acceleretion/MaximumPower и Acceleretion/TopSpeed -* т.к. ускорение 0-100 измеряется в секундах, то чем мощнее машина, тем меньше времени потребуется, чтобы набрать скорость 100 км/ч, аналогично для параметра *TopSpeed.* (т.к. он непосредственно зависит от MaximumPower)

Матрица корреляции



Данная матрица корреляции подтверждает все, выше описанное, и объясняет зависимости всех параметров машин между собой.

**Вывод:** в ходе лабораторной работы, был подготовлен датасет про машины,

изучены основные зависимости параметров машины. Датасет готов для дальнейшей работы.