**Московский государственный технический**

**университет им. Н.Э. Баумана**

Факультет "Информатика и системы управления"

Кафедра ИУ5 "Системы обработки информации и управления"

Технологии машинного обучения

Отчет по лабораторной работе №1

Выполнила:

Студентка группы ИУ5-63Б

Желтова Александра

Москва, 2021

**Задание**

Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

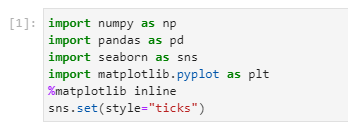
**Выполнение лабораторной работы:**

1.Текстовое описание данных

Набор данных о выставленных на продажу подержанных автомобилях Skoda, который содержит следующие колонки:

* model - модель автомобиля
* year - год выпуска
* price - цена (в фунтах стерлингах)
* transmission - тип коробки передач
* mileage - пробег (в милях)
* fuelType - тип используемого топлива
* tax - налог на автомобиль
* mpg - расход топлива (миль на галон)
* engineSize - объем двигателя

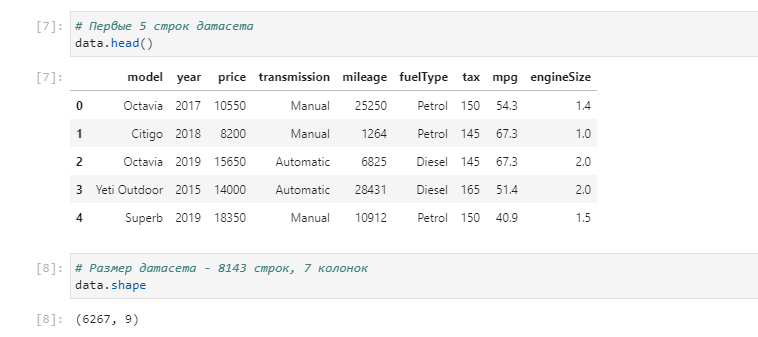
Импорт библиотек

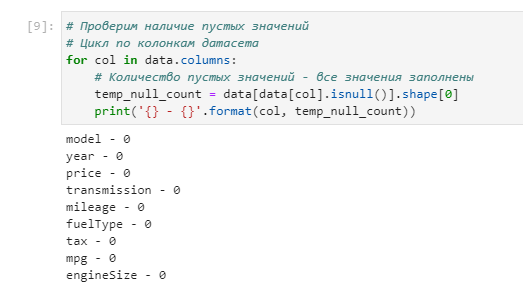


Загрузка данных



2.Основные характеристики датасета



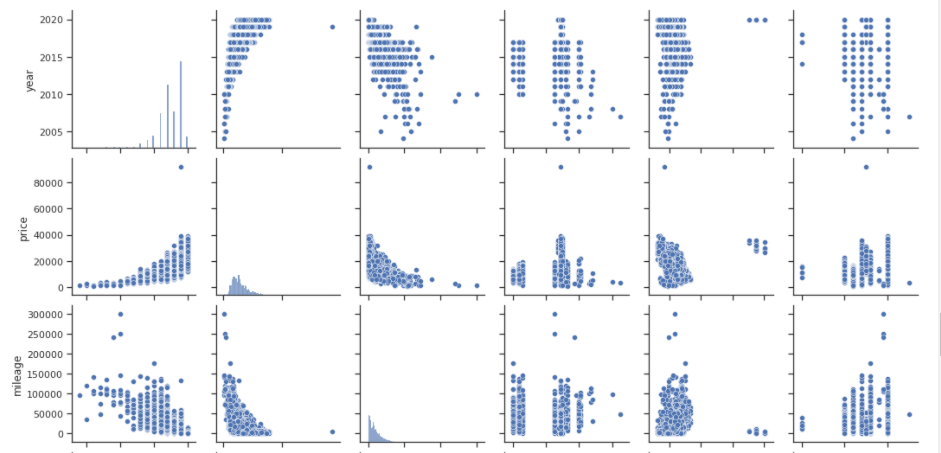


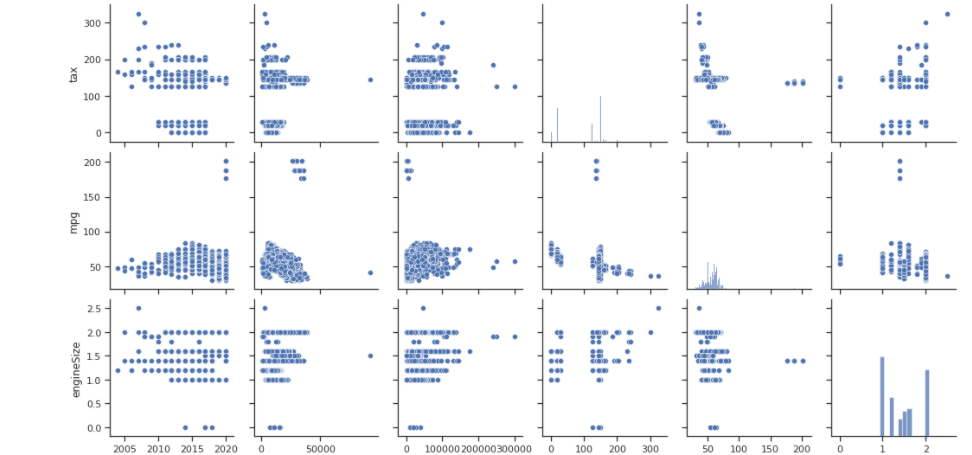


3.Визуальное исследование датасета

Для более наглядного примера выведем «Парные диаграммы», которые являются комбинацией гистограмм и диаграмм рассеивания для всего набора данных.

Выводится матрица графиков. На пересечении строки и столбца, которые соответствуют двум показателям, строится диаграмма рассеивания. В главной диагонали матрицы строятся гистограммы распределения соответствующих показателей.



4.Информация о корреляции признаков

Проверка корреляции признаков позволяет решить две задачи:

1. Понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком (в нашем примере это колонка "Occupancy"). Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Нужно отметить, что некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели.
2. Понять какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.

