Московский авиационный институт
(национальный исследовательский университет)

Институт информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Искусственный интеллект»

Студентка: Кочуйкова Д.В.
Группа: М8О-307Б

Москва, 2020

Постановка задачи:

Необходимо сформировать два набора данных для приложений машинного обучения. Первый датасет должен представлять из себя табличный набор данных для задачи классификации. Второй датасет должен быть отличен от первого, и может представлять из себя набор изображений, корпус документов, другой табличный датасет или датасет из соревнования Kaggle, предназначенный для решения интересующей вас задачи машинного обучения. Необходимо провести анализ обоих наборов данных, поставить решаемую вами задачу, определить признаки необходимые для решения задачи, в случае необходимости заняться генерацией новых признаков, устранением проблем в данных, визуализировать распределение и зависимость целевого признака от выбранных признаков. В отчете описать все проблемы, с которыми вы столкнулись и выбранные подходы к их решению

Датасеты:

Для первой части лабораторной работы был взят датасет из соревнования Kaggle, он представляет собой табличный набор данных. Датасет состоит из данных о тестировании людей разных национальностей, полов и т.д. по математике, чтению и письму.

Для второй части лабораторной был также использован датасет из соревнования Kaggle, в нем показана история акций компании с момента ее основания. В таблице представлены дата торгового дня, цена при открытии, максимальная, минимальная цены во время торгов, цена при закрытии, скорректированная цена закрытия и объемы акций.

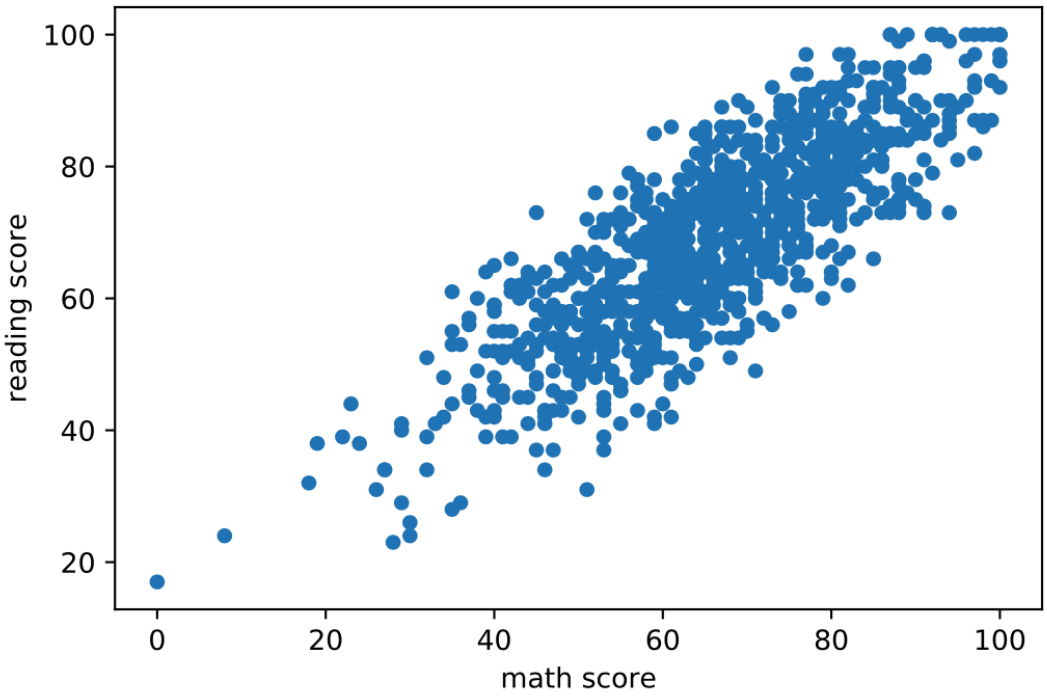
1 задача:

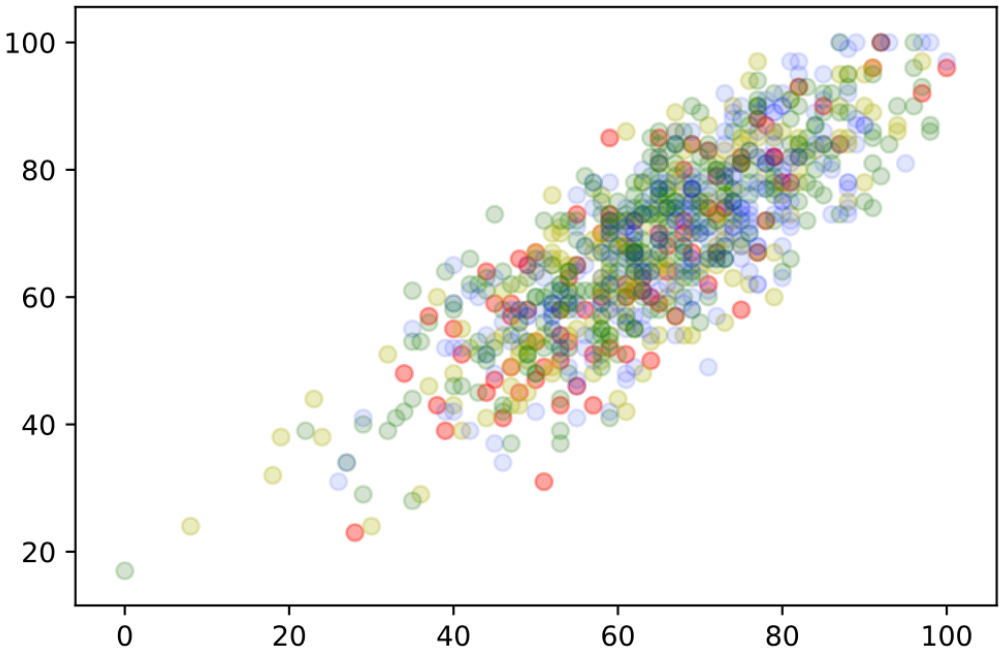
Для задачи классификации был выбран датасет из данных о тестировании людей. Составляющие таблицы: gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, writing score

Для начала подгружаем датасет и преобразуем наши категориальные признаки, чтобы с ними можно было работать.

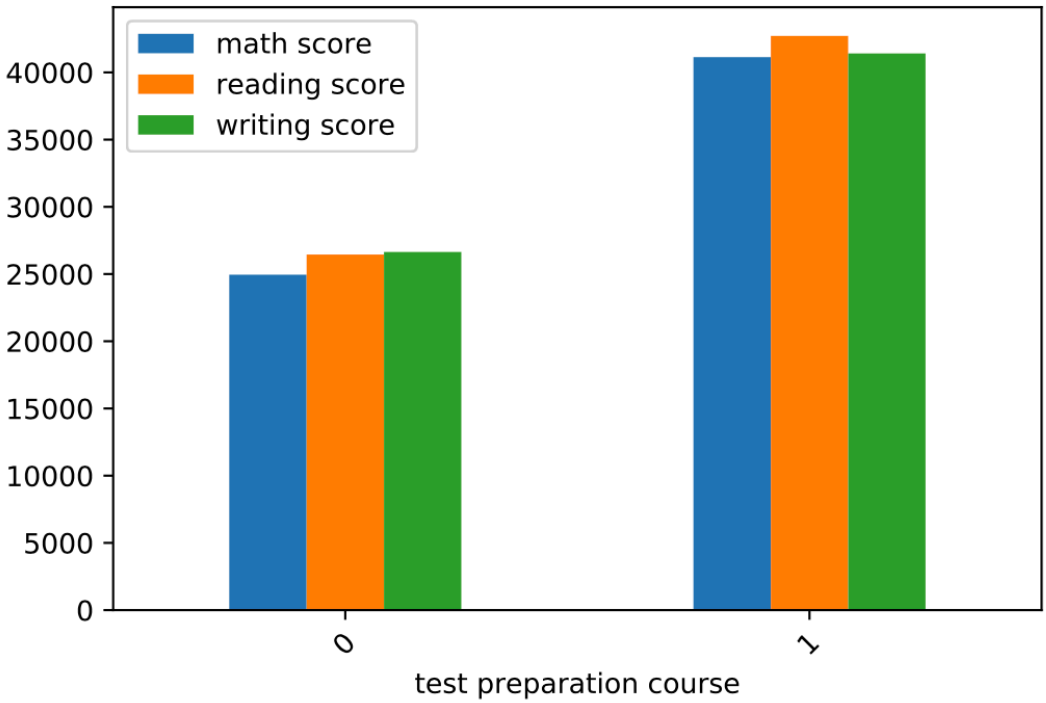
	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	0	1	1	1	1	72	72	74
1	0	2	4	1	0	69	90	88
2	0	1	3	1	1	90	95	93
3	1	0	0	0	1	47	57	44
4	1	2	4	1	1	76	78	75

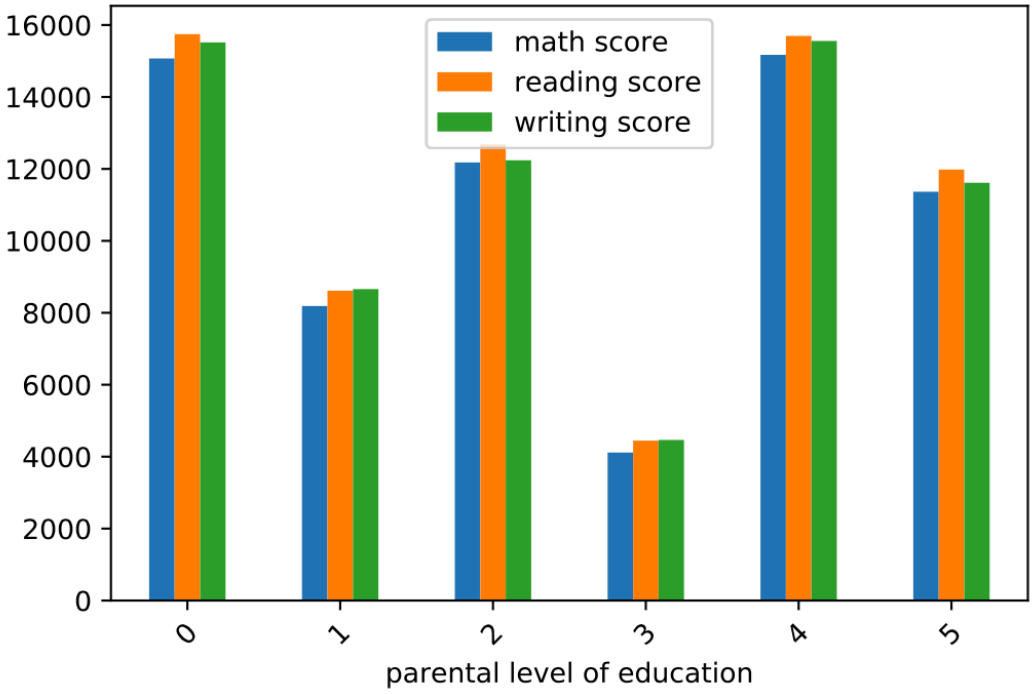
Зависимость результатов друг от друга, с чтением ситуация аналогичная





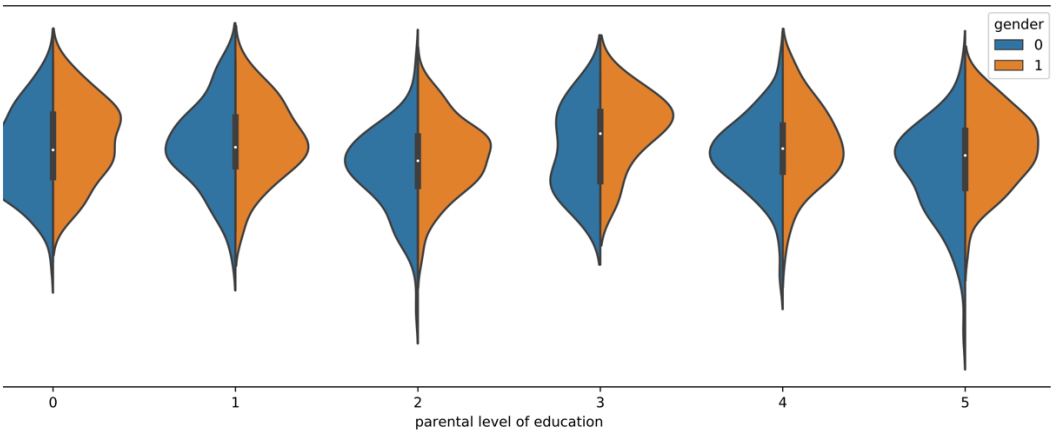
Исследуем зависимости от категориальных признаков



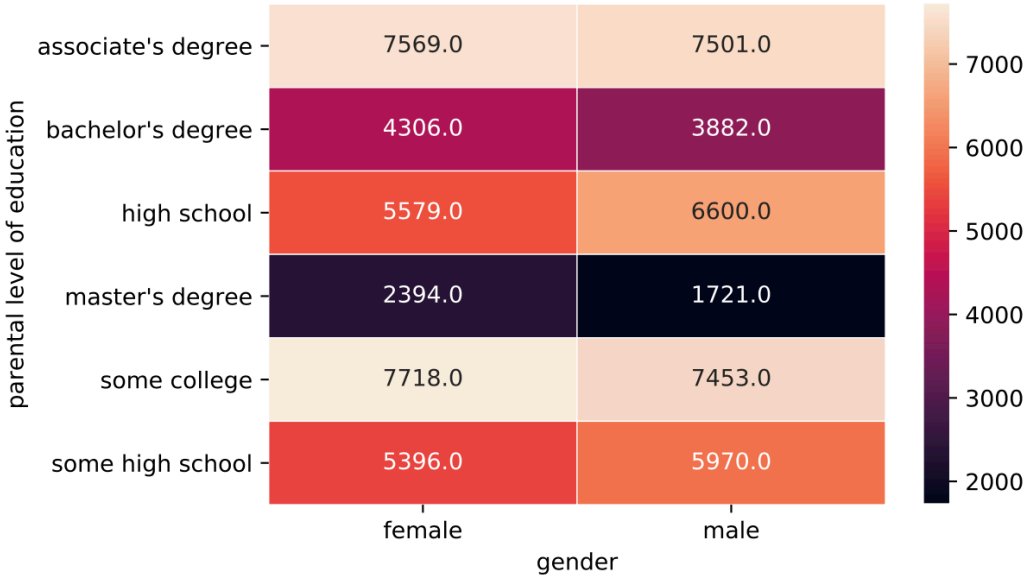


Я исследовала распределения результатов по признакам уровня образования и прохождения подготовительных курсов.

Также было исследовано распределение по половому признаку, результаты показали примерно одинаковые результаты и для женского и для мужского пола



Ниже показана таблица распределения баллов по 2 категориальным признакам, где была обнаружена зависимость.

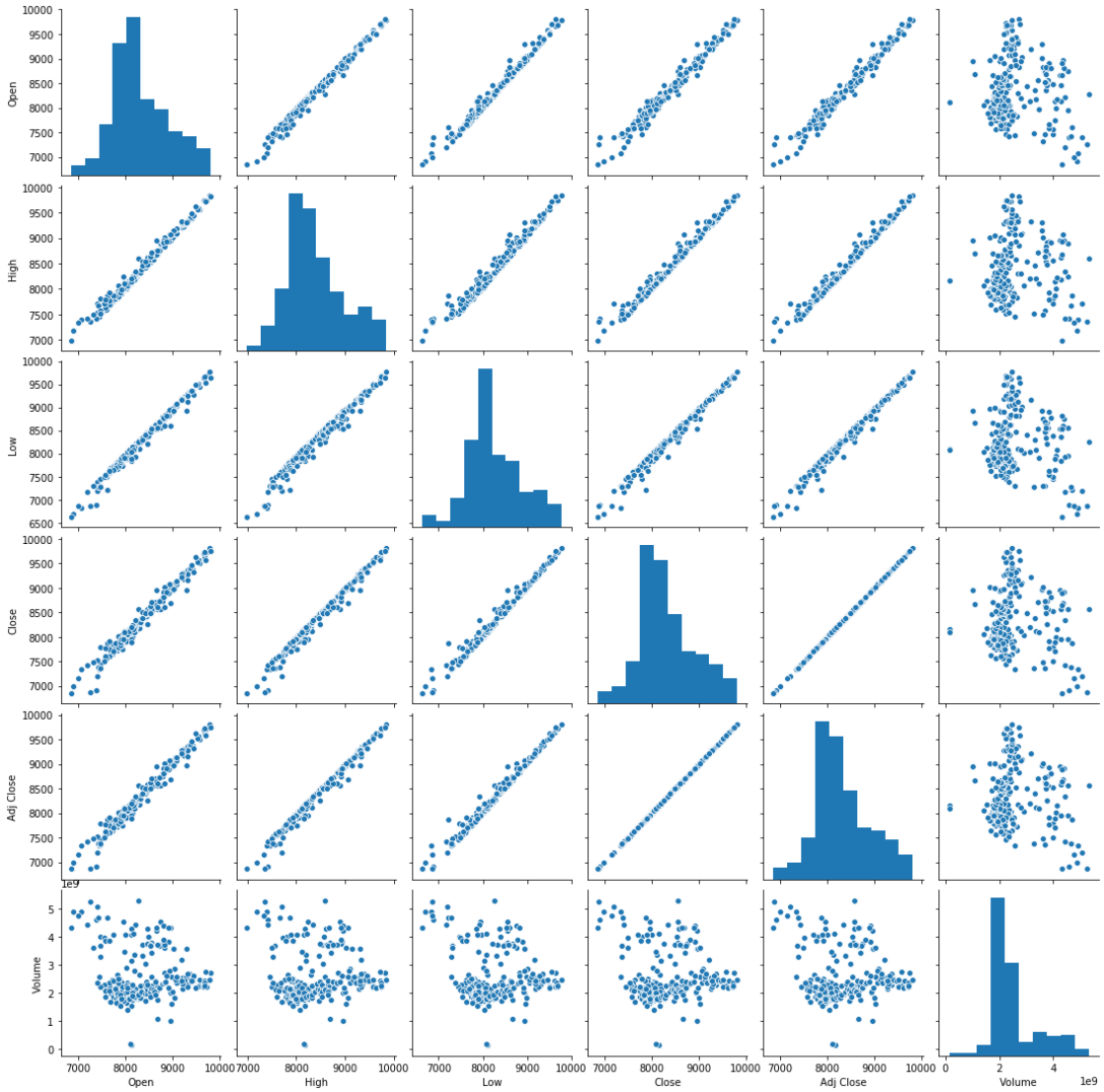


На основании изучения данных была поставлена задача классификации, целью которой является определение уровня обучения человека. При выполнении этого задания возникла проблема визуализации зависимости от категориального признака. В данных проблем на данном этапе не было обнаружено.

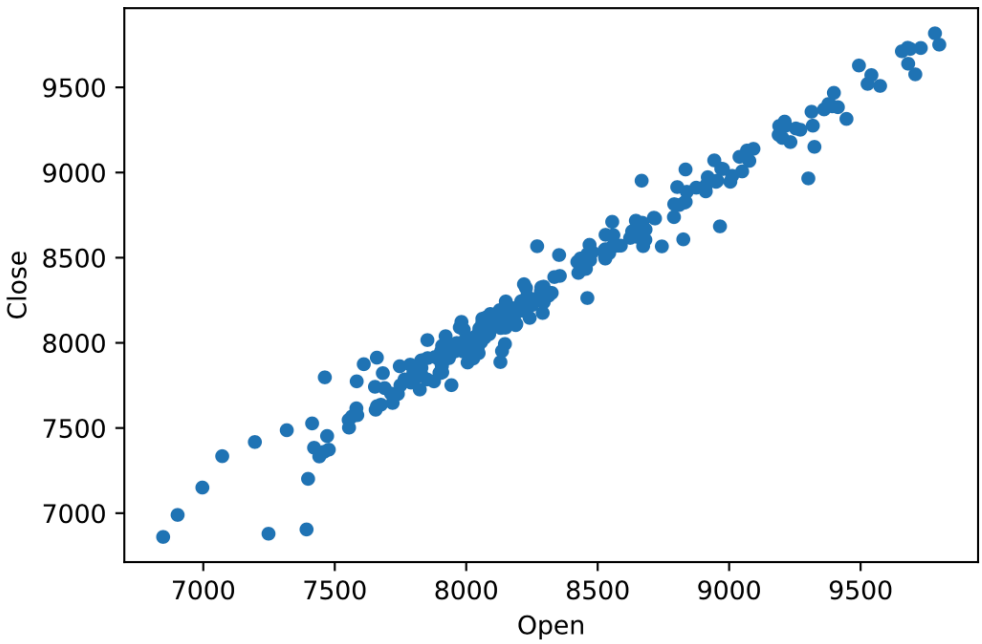
2 задача

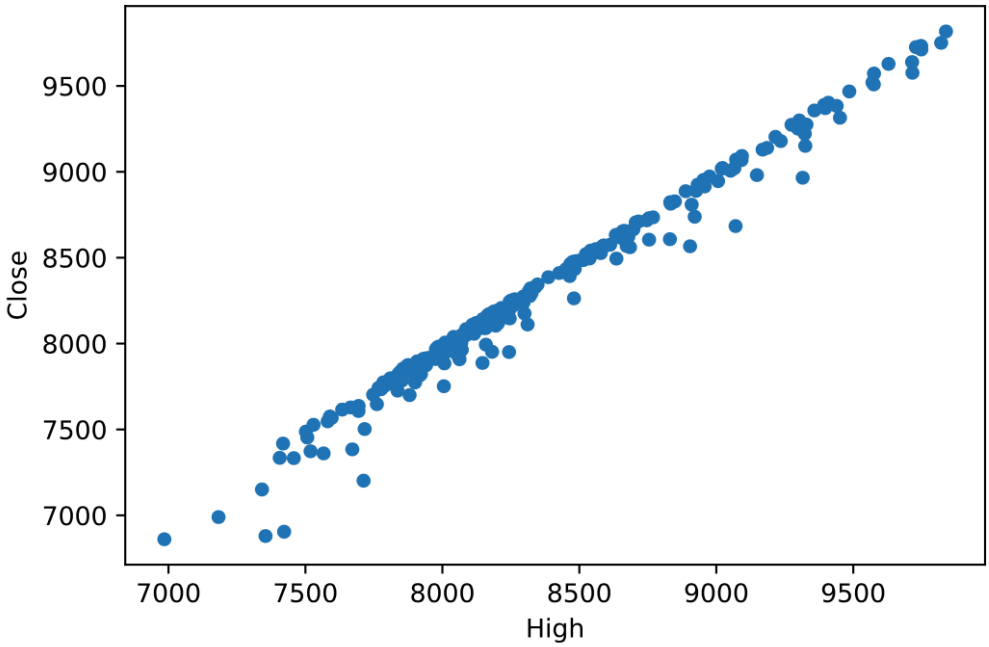
Для 2 задачи был выбран датасет из данных о стоимости акций. Составляющие таблицы: Date, Open, High, Low, Close, Adj Close, Volume

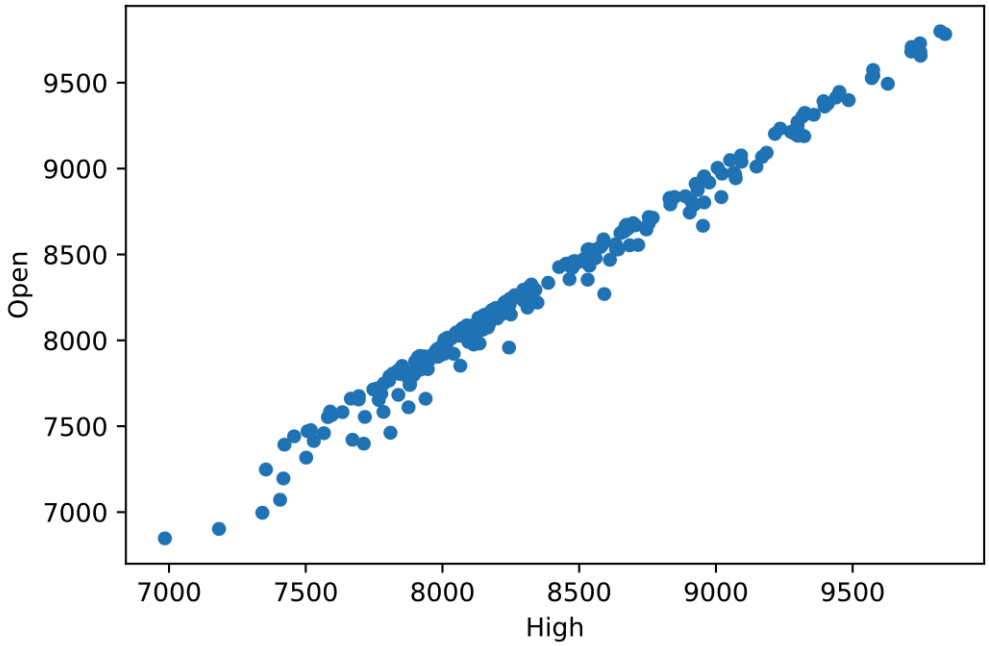
Анализ данных я начала с визуализации зависимости всех признаков таблицы. Видно, что некоторые из них линейно зависимы.


Open
High
Low
Close
Adj Close
Volume
1e9

А конкретно:


Close
Open
9500
9000
8500
8000
7500
7000






```
Avg: 8322.171471346455
Median: 8184.044921999999
Max: 9817.179688
Min: 6860.669922
Delta: 2956.509766
Div: 598.3041272468345
Dispersion: 357967.8286805963
```

Максимальная цена почти всегда больше цены открытия – это связано с тем, что цена акций совершает довольно большие колебания в течение дня.

Цена на момент закрытия имеет несущественное отличие от максимальной цены за день.

Цена закрытия в многих случаях больше цены открытия, это связано с тем, что в такой большой выборке компания вела дела довольно успешно.

На основе анализа данных были выбраны признаки, которые требуется предсказать: максимальная/минимальная цена за день. Проблем в данных на данном этапе также не было обнаружено

Вывод.

Выполнив лабораторную работу, я ознакомилась с возможностями визуализации данных в python. Изучила способы нахождения информации и датасетов. Научилась ставить задачи для имеющихся данных.