|  |  |
| --- | --- |
|  | |
| МИНистерство Науки и высшего образования российской федерации | |
| Федеральное государственное бюджетное образовательное учреждениевысшего образования"МИРЭА - Российский технологический университет"РТУ МИРЭА | |
| Институт информационных технологий (ИТ) | |
| Кафедра прикладной математики (ПМ) | |

|  |  |
| --- | --- |
| **ОТЧЕТ**  **ПО ПРАКТИЧЕСКИМ РАБОТАМ** | |
| **по дисциплине** | |
| «Технологии и инструментарий анализа больших данных» | |
|  | |
| Выполнили студенты группы ИВБО-07-19 | Гридасов Е.А. |
| Принял ассистент кафедры ПМ | Горячев А.А. |

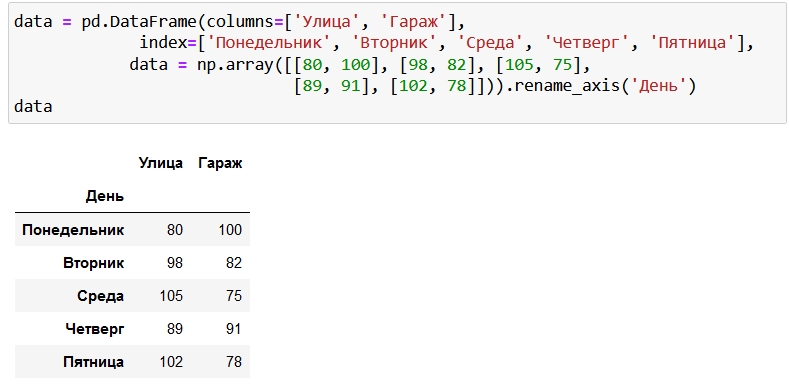
|  |  |  |
| --- | --- | --- |
| Практическая работа выполнена | «\_\_»\_\_\_\_\_\_\_\_\_2022 г. |  |
|  |  |  |
| «Зачтено» | «\_\_»\_\_\_\_\_\_\_\_\_2022 г. |  |

Москва 2022

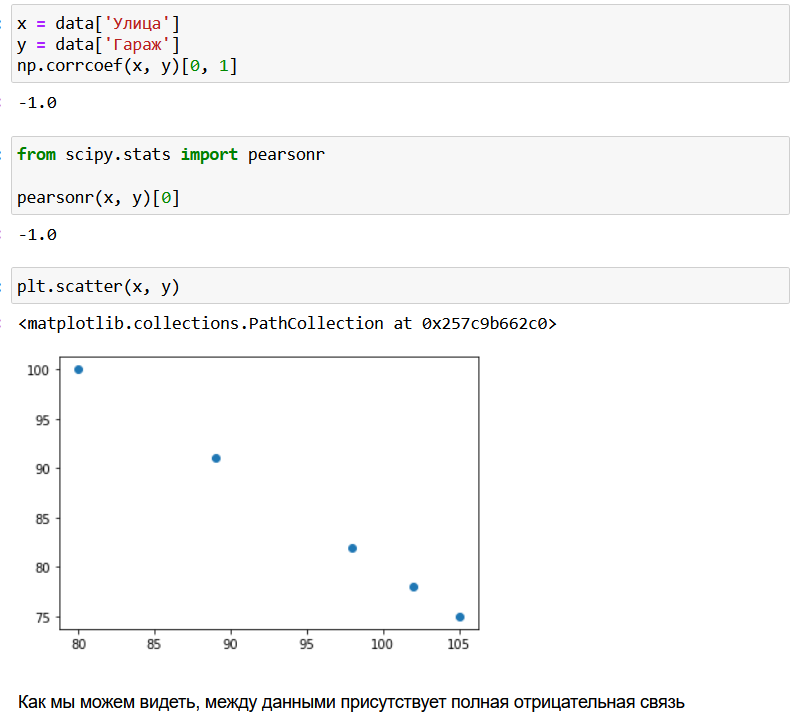
# ВЫПОЛНЕНИЕ ЗАДАЧ

Целью работы было изучить линейную регрессию.

С целью выполнения задачи был создан датафрейм, вычислена корреляция между переменными «Улица» и «Гараж» с помощью библиотеки numpy и stats по Пирсону.



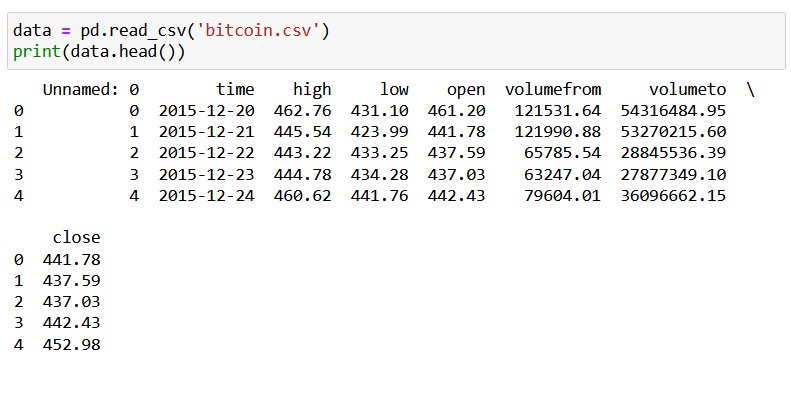
**Рисунок 1 – Создание датасета**



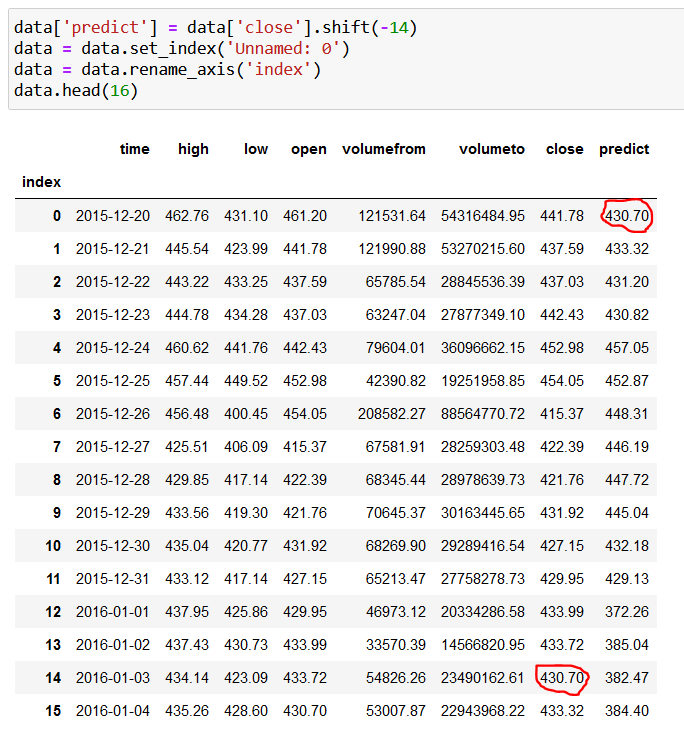
**Рисунок 2 — Вычисление корреляции**

Корреляция была вычислена с помощью numpy и stats.pearsonr. Видно, что полученные значения равны, а по графику можем убедиться в правильности предположения, что между данными — полная отрицательная связь.

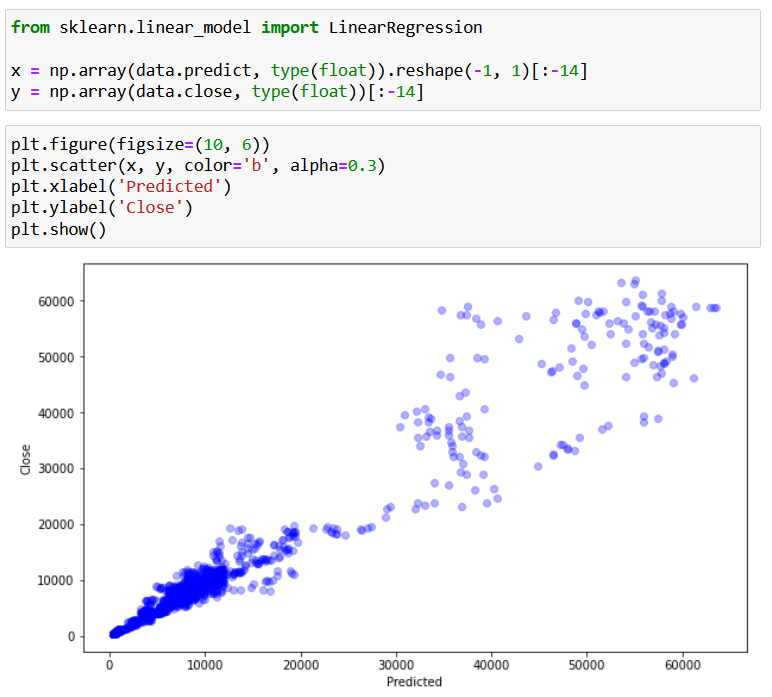
После нужно было загрузить файл bitcoin.csv, скрыть последние 14 дней, вывести угол наклона и y-перехвата, построить диаграмму.



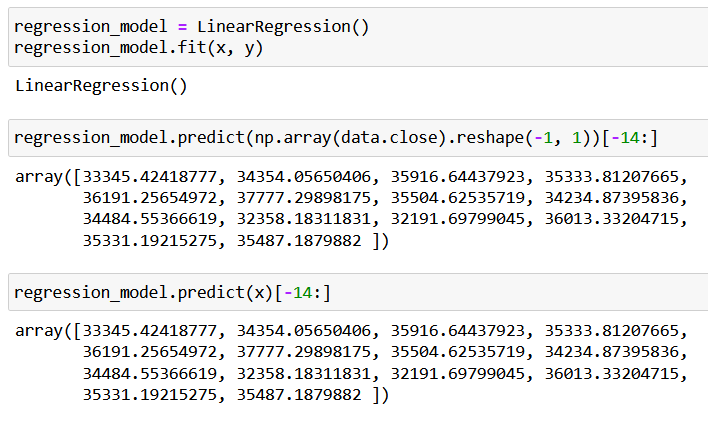
**Рисунок 3 — Загрузка датасета bitcoin.csv**



**Рисунок 4 — Смещение значений и исправление нейминга индексов для удобства**

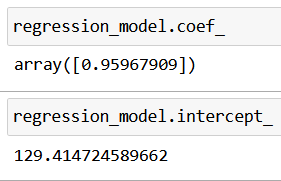


**Рисунок 5 — Создание данных для обучения модели**



**Рисунок 6 — Обучение модели и предсказание**

Как мы видим, значения, вычисленные на изначальных данных и данных, полученных смещением, равны.



**Рисунок 7 — Угол наклона и y-перехват**



**Рисунок 8 — Построение диаграммы рассеивания с регрессией**

Последняя часть задания заключалась в скачивании датасета housePrice.csv, предобработка данных, реализация линейной регрессии вручную, вывод угла наклона и y-перехвата, построение диаграммы.

Для начала был изменен нейминг колонок, чтобы было легче ориентироваться в будущем.

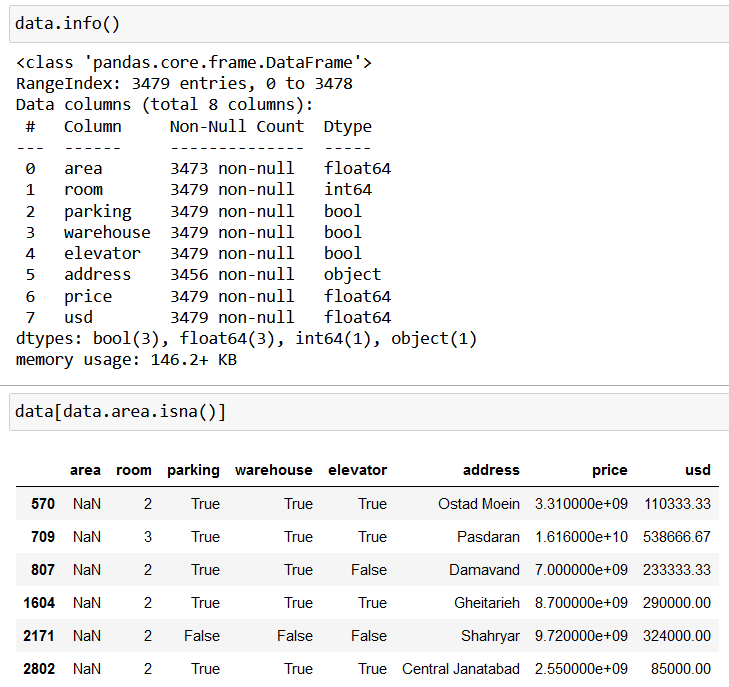
При изучении было подмечено две вещи: корреляция наиболее очевидна между значениями usd и area — на них решено было построить линейную регрессию, а также то, что часть значений area – пропущена. Ввиду их малого количества, было решено заменить недостающие данные на среднее значение данных колонки.



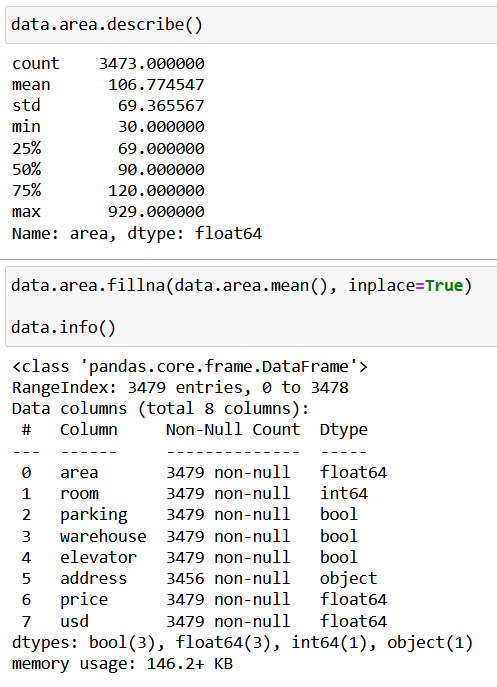
**Рисунок 9 — Скачивание датасета и изменение названий колонок**



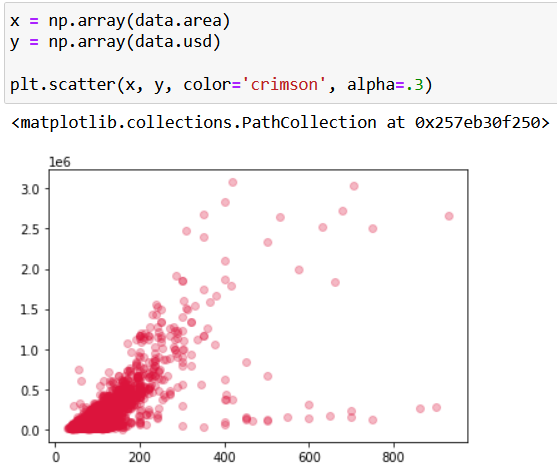
**Рисунок 10 — Просмотр матрицы корреляции**



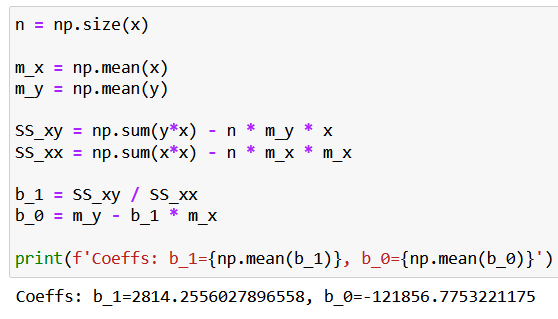
**Рисунок 11 — Проверка датасета**



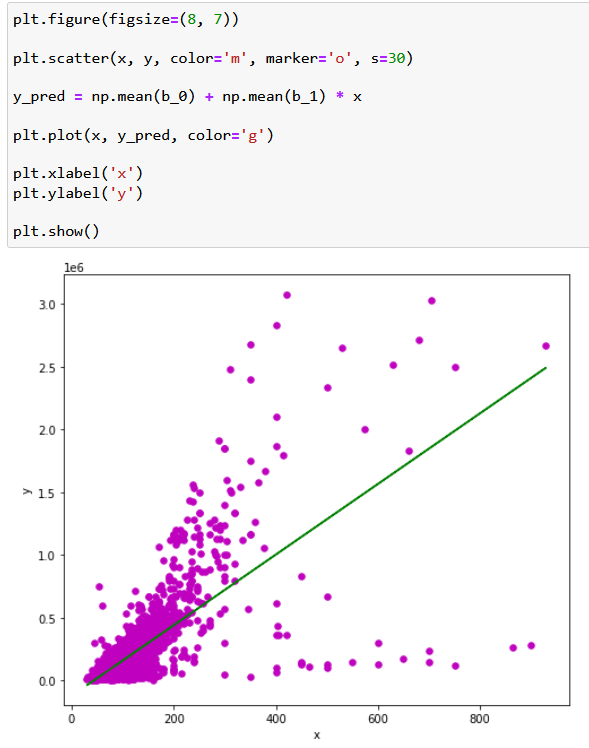
**Рисунок 12 — Заполнение пустых строк**



**Рисунок 13 — Построение диаграммы рассеивания**



**Рисунок 14 — Вычисление и вывод наклона и y-перехвата**



**Рисунок 15 — Построение диаграммы рассеивания и линейной регрессии**

# ВЫВОДЫ

В ходе выполнения работы были изучены библиотеки matplotlib, seaborn, sklearn, stats. Освоена тема линейной регрессии, корреляции данных, метода наименьших квадратов. Вычислена линейная регрессия вручную и с использованием библиотеки sklearn.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Python Documentation [Электронный ресурс] - <https://www.python.org/>
2. Jupyter Documentation [Электронный ресурс] - <https://docs.jupyter.org/en/latest/index.html>
3. Pandas Documentation [Электронный ресурс] - <https://pandas.pydata.org/docs/getting_started/index.html>
4. Matplotlib Documentation [Электронный ресурс] - <https://matplotlib.org/>
5. Plotly Documentation [Электронный ресурс] - <https://plotly.com/python/>
6. Numpy Documentation [Электронный ресурс] - <https://numpy.org/doc/>
7. Stats Documentation [Электронный ресурс] - <https://docs.scipy.org/doc/scipy/reference/stats.html>