|  |  |
| --- | --- |
|  | |
| 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. |

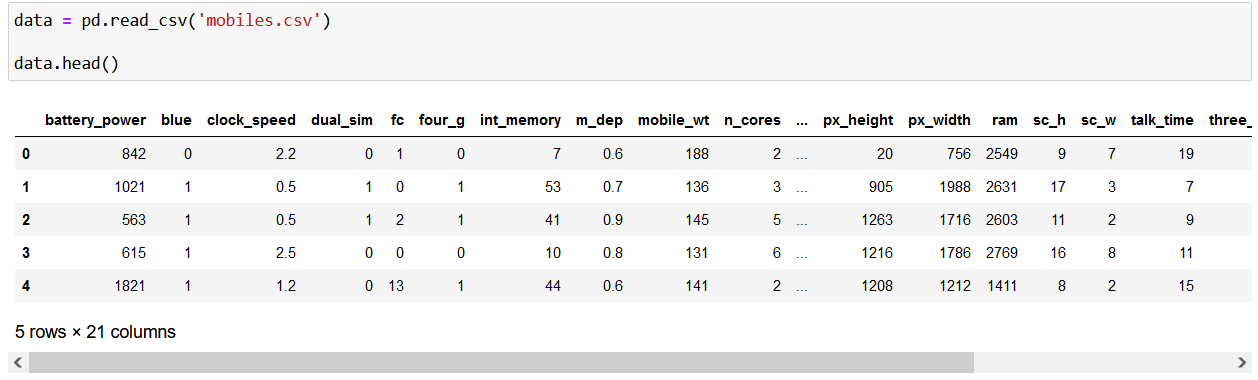
|  |  |  |
| --- | --- | --- |
| Практическая работа выполнена | «\_\_»\_\_\_\_\_\_\_\_\_2022 г. |  |
|  |  |  |
| «Зачтено» | «\_\_»\_\_\_\_\_\_\_\_\_2022 г. |  |

Москва 2022

# ВЫПОЛНЕНИЕ ЗАДАЧ

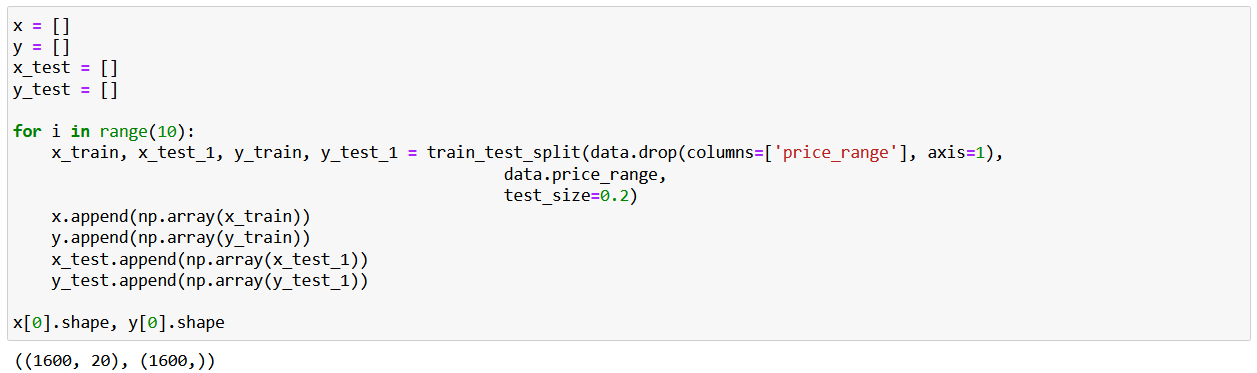
Целью работы было изучение дерева решений, а также методов баггинга, бустинга.

Для выполнения задания был скачан датасет для классификации мобильных телефонов на ценовые категории по параметрам: 0 — низкая цена, 1 — средняя, 2 — высокая цена, 3 — дорого.



**Рисунок 1 – Датасет**

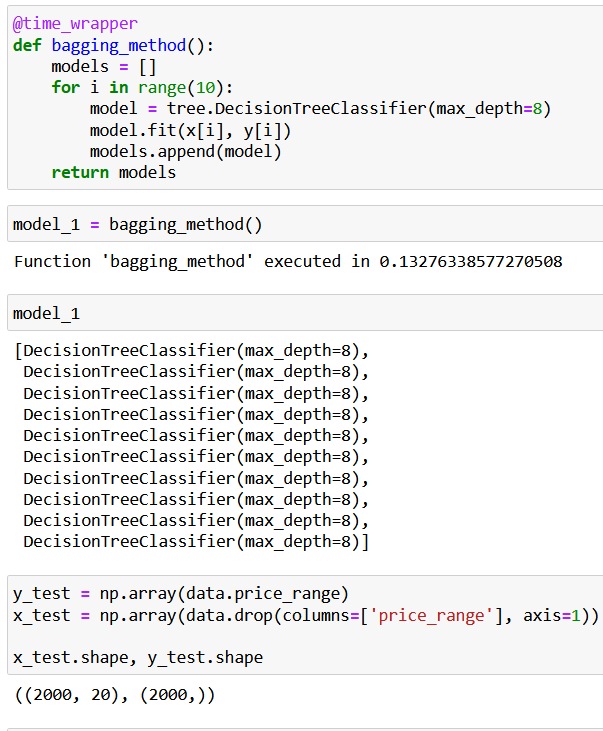
Для начала был реализован баггинг на основе моделей деревьев решений.



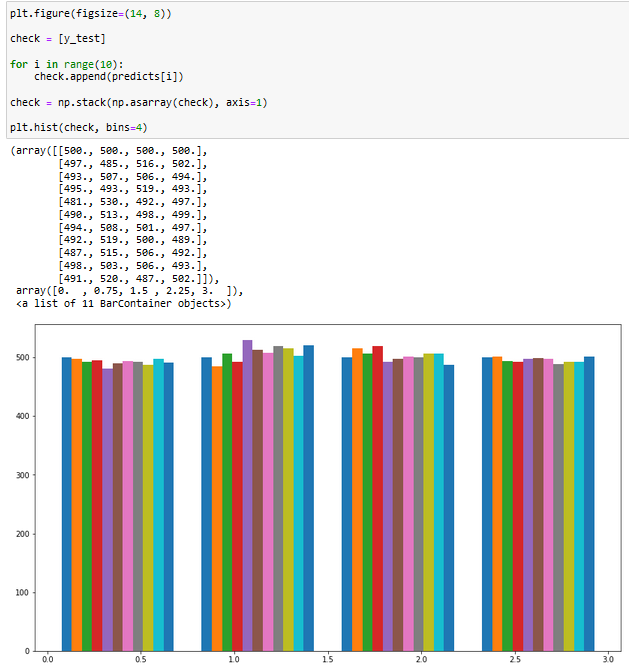
**Рисунок 2 — Подготовка данных для обучени**



**Рисунок 3 — Декоратор времени**

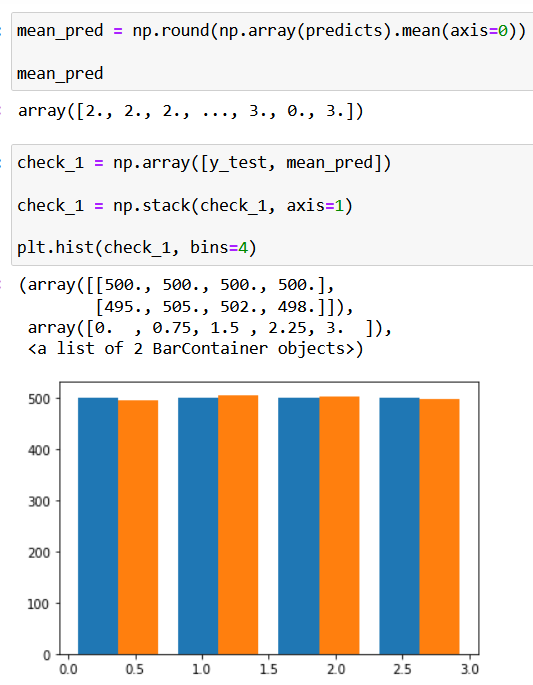


**Рисунок 4 — Баггинг и подготовка данных для тестирования**

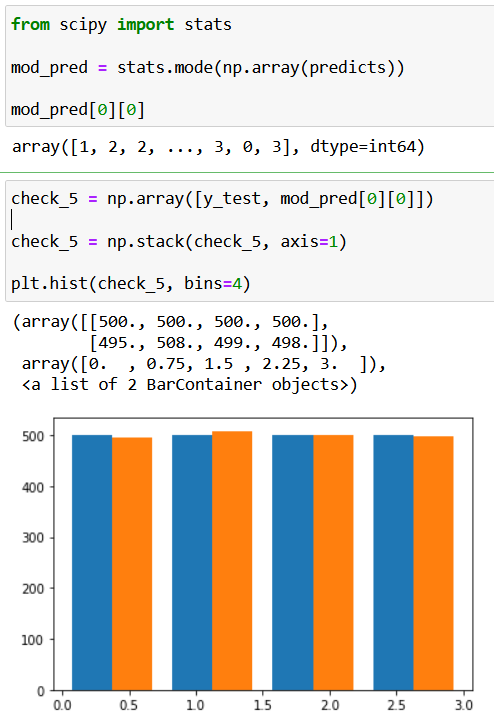


**Рисунок 5 — Результат визуализации баггинга**

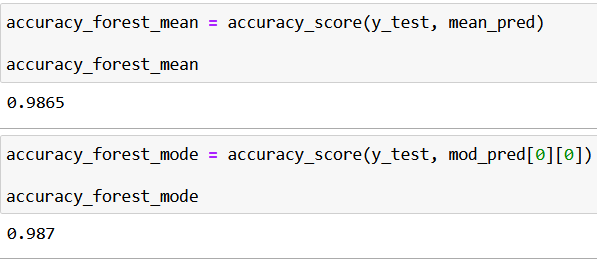
Видно, что деревья дают разный результат. Попробуем сделать предсказание по среднему и моде.



**Рисунок 6 — Результат предсказания среднего моделей**

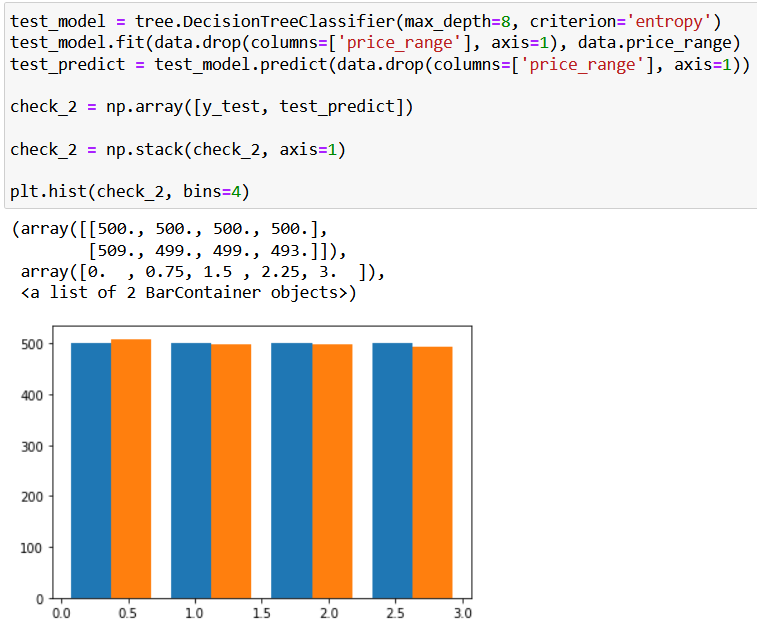


**Рисунок 7 — Результат предсказаний моды моделей**

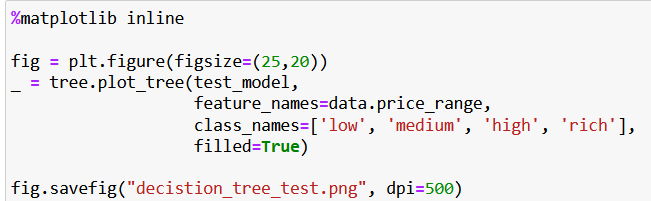


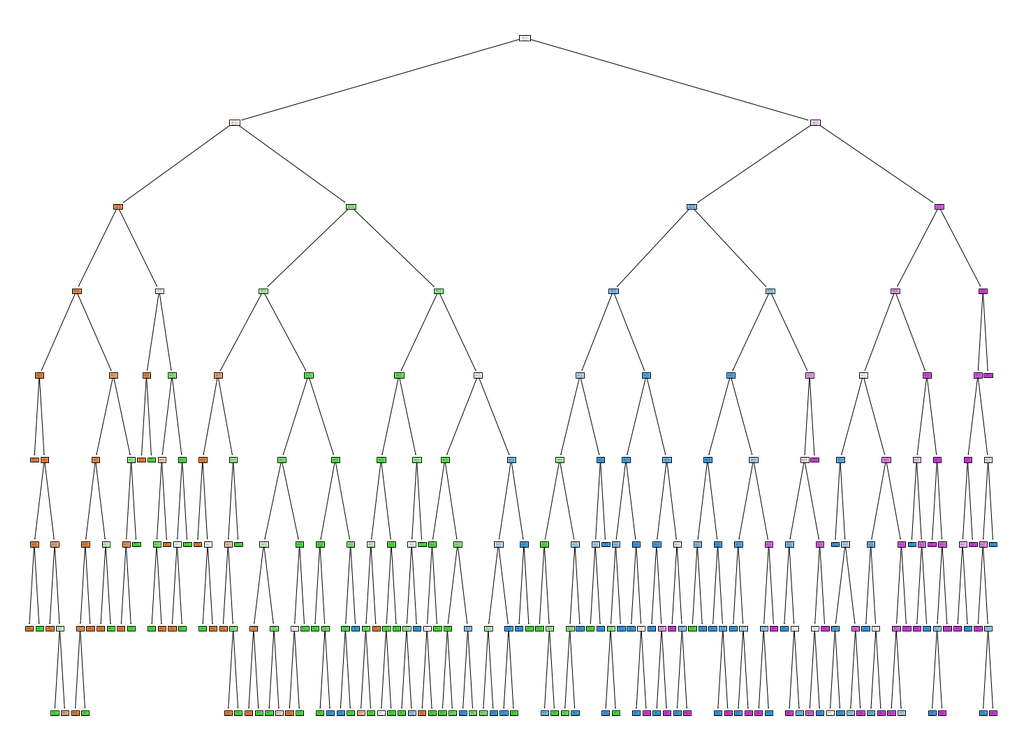
**Рисунок 8 — Тест точности**

После была реализована классификация с помощью одного дерева решений. Само построенное дерево будет приложено в виде png картинки.

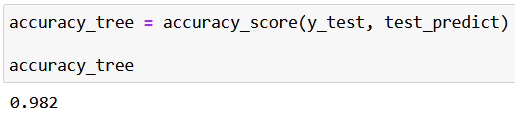


**Рисунок 9 — Визуализация предсказаний обученной модели дерева решений**



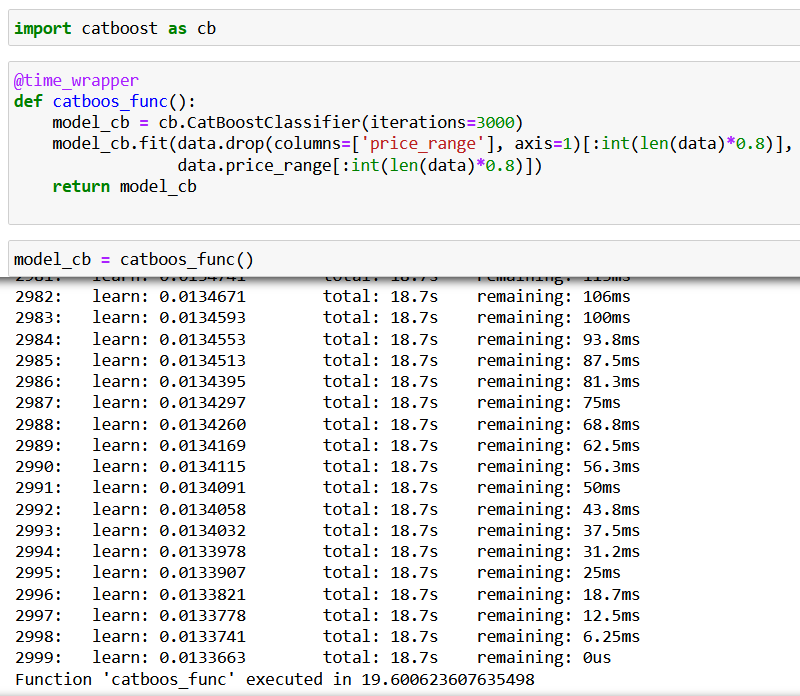


**Рисунок 10 — Построенное дерево**

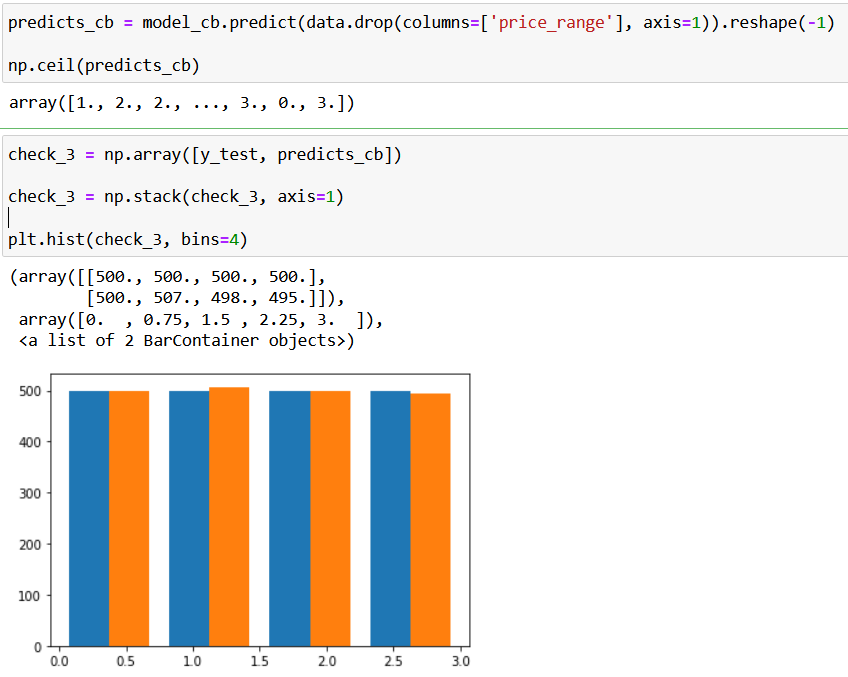


**Рисунок 11 — Тест точности**

Для бустинга были выбраны catboost и XGBoost. Сначала catboost.

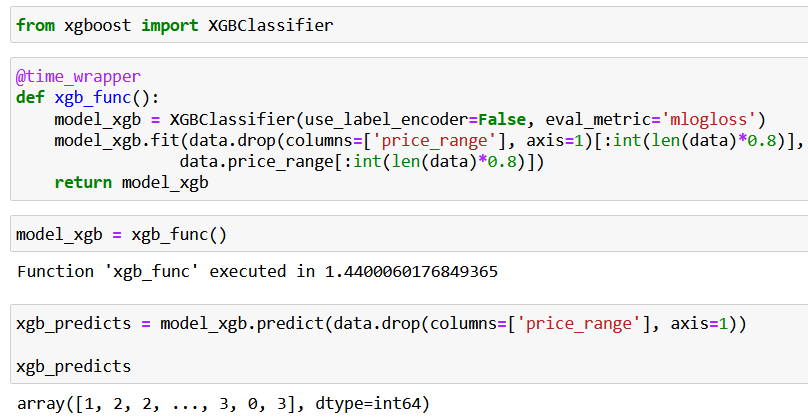


**Рисунок 12 — Модель catboost классификатор**

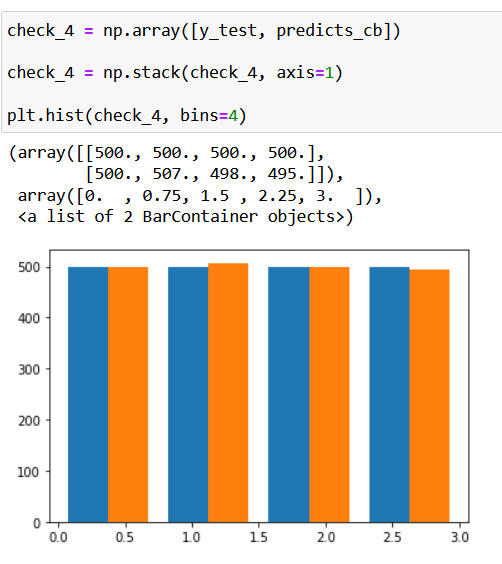


**Рисунок 13 — Визуализирование предсказаний**

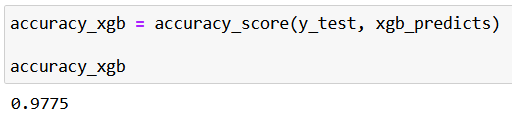
Последним методом бустинга стал XGBoost классификатор.



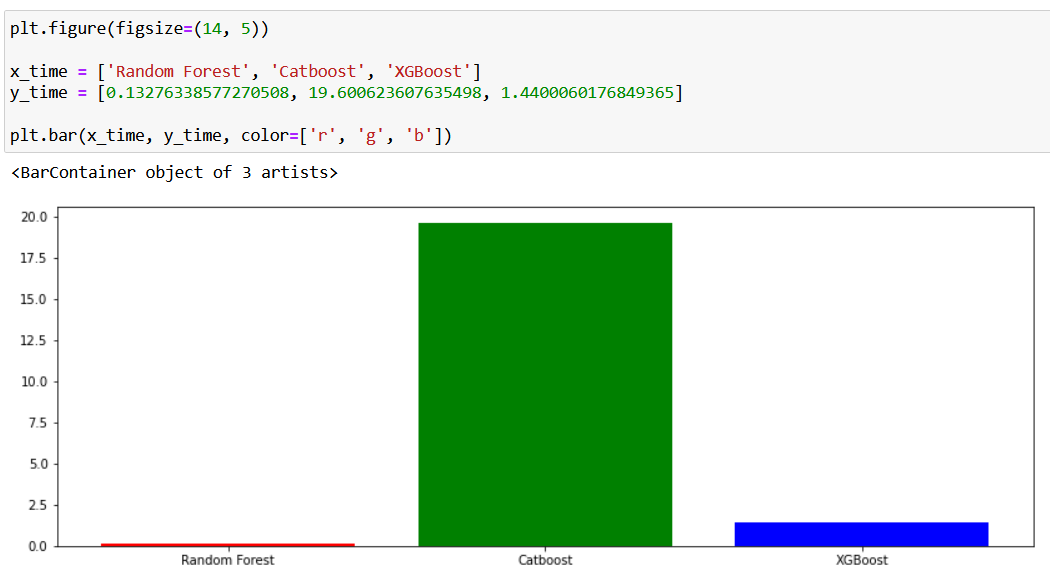
**Рисунок 14 — XGBoost классификатор и предсказания**



**Рисунок 15 — Визуализация предсказаний XGBoost**



**Рисунок 16 — Тест точности**



**Рисунок 17 — Сравнение методов по времени**



**Рисунок 18 — Сравнение точности моделей**

По результату были сделаны выводы: метод catboost требует самое большое количество времени за счет количества итераций, а рандомный лес – самое малое. По точности хуже всего показало себя дерево решений, а лучше всего – catboost. К тому же XGBoost показал худший результат по точности, уступая случайному лесу. Из двух способов реализации случайного леса для этого датасета лучше показало себя предсказание по моде от всех предсказаний деревевьев.

# ВЫВОДЫ

В ходе выполнения работы были изучены библиотеки matplotlib, plotly, sklearn, scipy. Освоена тема баггинга, бустинга.

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>