|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

**Факультет «Информатика и системы управления»**

**Кафедра «Системы обработки информации и управления»**

Отчет по лабораторной работе №6

**«Анализ и прогнозирование временного ряда»**

по дисциплине «Технологии машинного обучения»

Выполнил:

студент группы ИУ5Ц-84Б

Перевощиков Н.Д.

подпись, дата

Проверил:

к.т.н., доц., Гапанюк Ю.Е.

подпись, дата

2024 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Цель лабораторной работы 3](#_Toc161683608)

[2. Задание 3](#_Toc161683609)

[3. Основные характеристики датасета 3](#_Toc161683610)

[4. Листинг 4](#_Toc161683611)

[5. Вывод 28](#_Toc161683612)

## Цель лабораторной работы

Изучение основных методов анализа и прогнозирование временных рядов.

## Задание

1. Выберите набор данных (датасет) для решения задачи прогнозирования временного ряда.
2. Визуализируйте временной ряд и его основные характеристики.
3. Разделите временной ряд на обучающую и тестовую выборку.
4. Произведите прогнозирование временного ряда с использованием как минимум двух методов.
5. Визуализируйте тестовую выборку и каждый из прогнозов.
6. Оцените качество прогноза в каждом случае с помощью метрик.

## Основные характеристики датасета

Название датасета: Набор данных о временных рядов акций DJIA 30

Ссылка: https://www.kaggle.com/datasets/szrlee/stock-time-series-20050101-to-20171231?select=GOOGL\_2006-01-01\_to\_2018-01-01.csv

**О датасетах**

Данные фондового рынка могут быть интересны для анализа, и в качестве дополнительного стимула надежные прогностические модели могут принести большую финансовую отдачу. Количество финансовых данных в Интернете, по-видимому, бесконечно. Получить большой и хорошо структурированный набор данных по широкому кругу компаний может быть непросто. Здесь представлен набор данных с историческими ценами акций (за последние 12 лет) для 29 из 30 компаний DJIA (исключая 'V', поскольку в нем нет полных данных за 12 лет).

**Структура данных**

Здесь указаны 5 параметров:

open - цена акций в момент открытия рынка

high - самая высокая цена акций, достигнутая в день

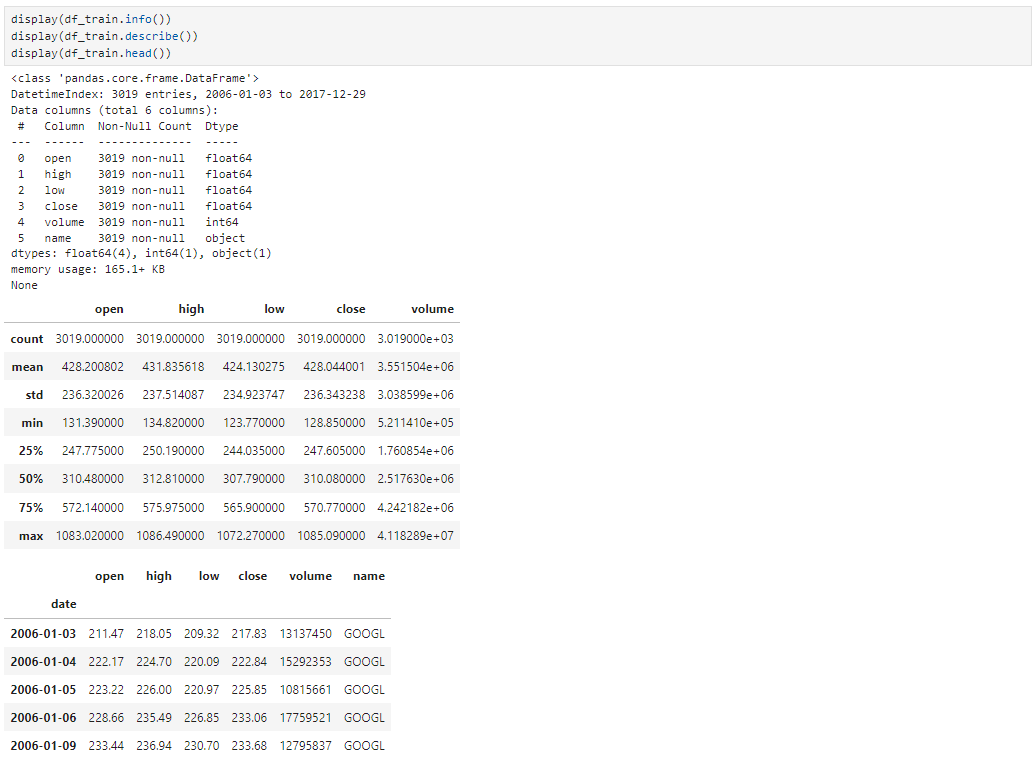
low - самая низкая цена акций, достигнутая в день

close - цена акций в момент закрытия рынка

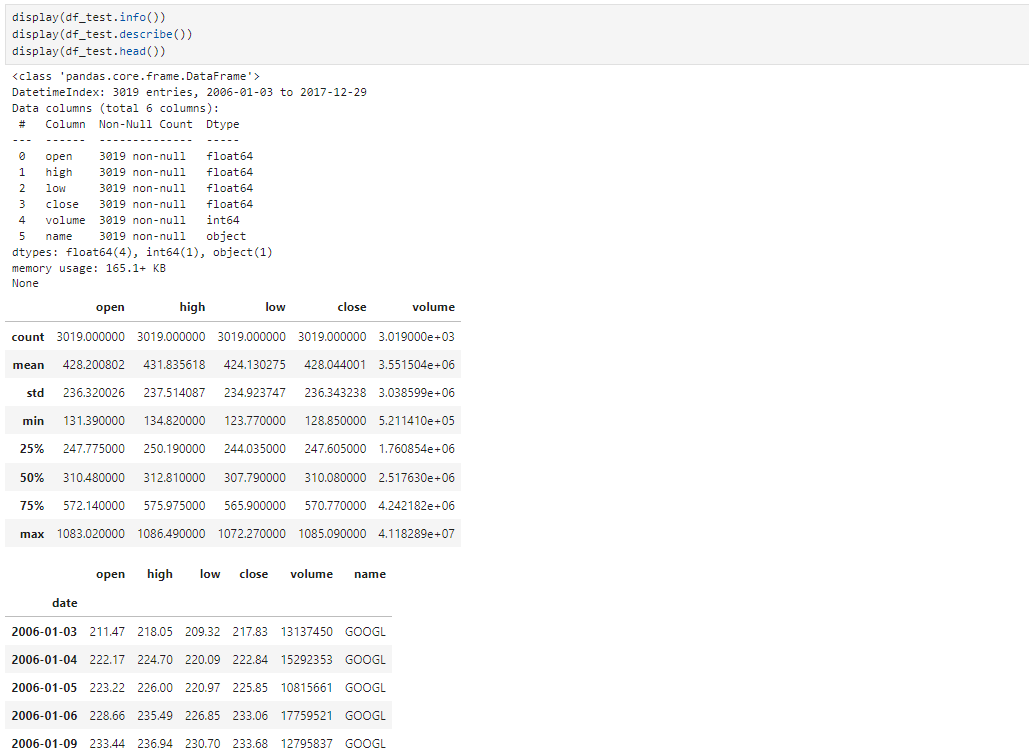
volume - количество проданных акций

## Листинг

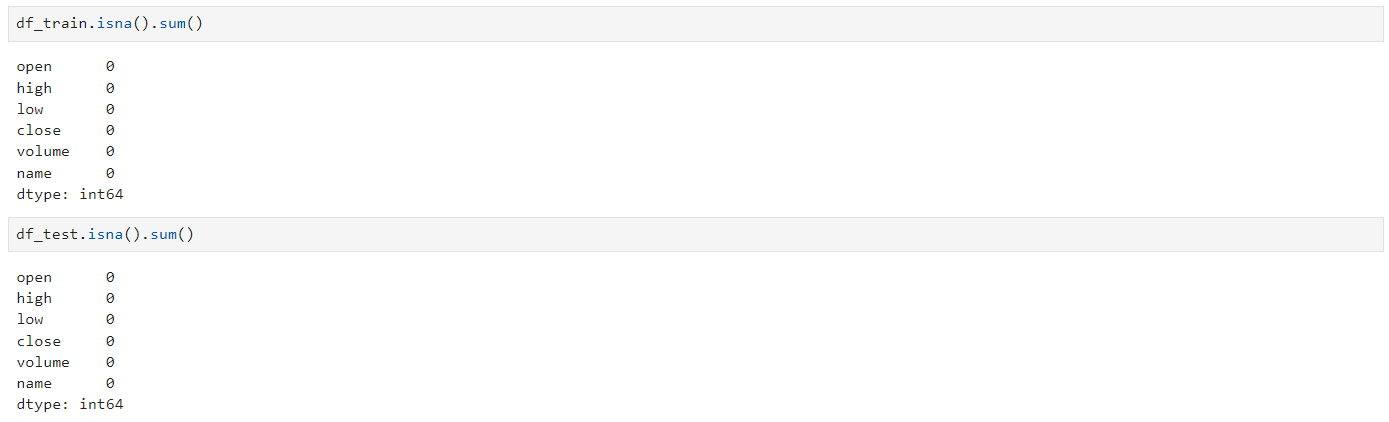
* 1. **Изучение данных**
     1. **Обучающий датасет**

****

* + 1. **Тестовый датасет**

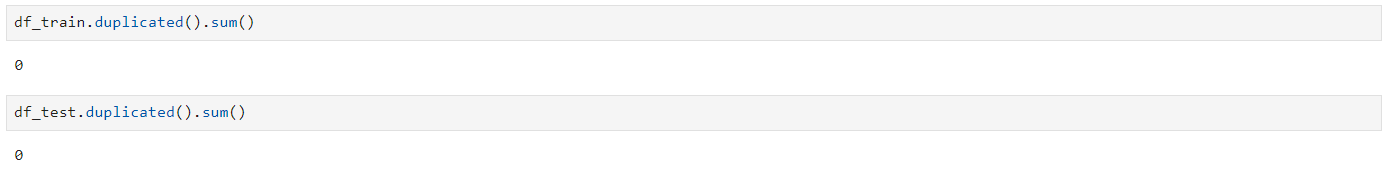
****

* 1. **Предобработка данных**
     1. **Пропуски**

****

Пропусков нет, это говорится о том, что датасет был сделан хорошо.

* + 1. **Дубликаты**

****

Дубликатов нет, это говорится о том, что датасет был сделан хорошо.

* + 1. **Удаление не числового типа**

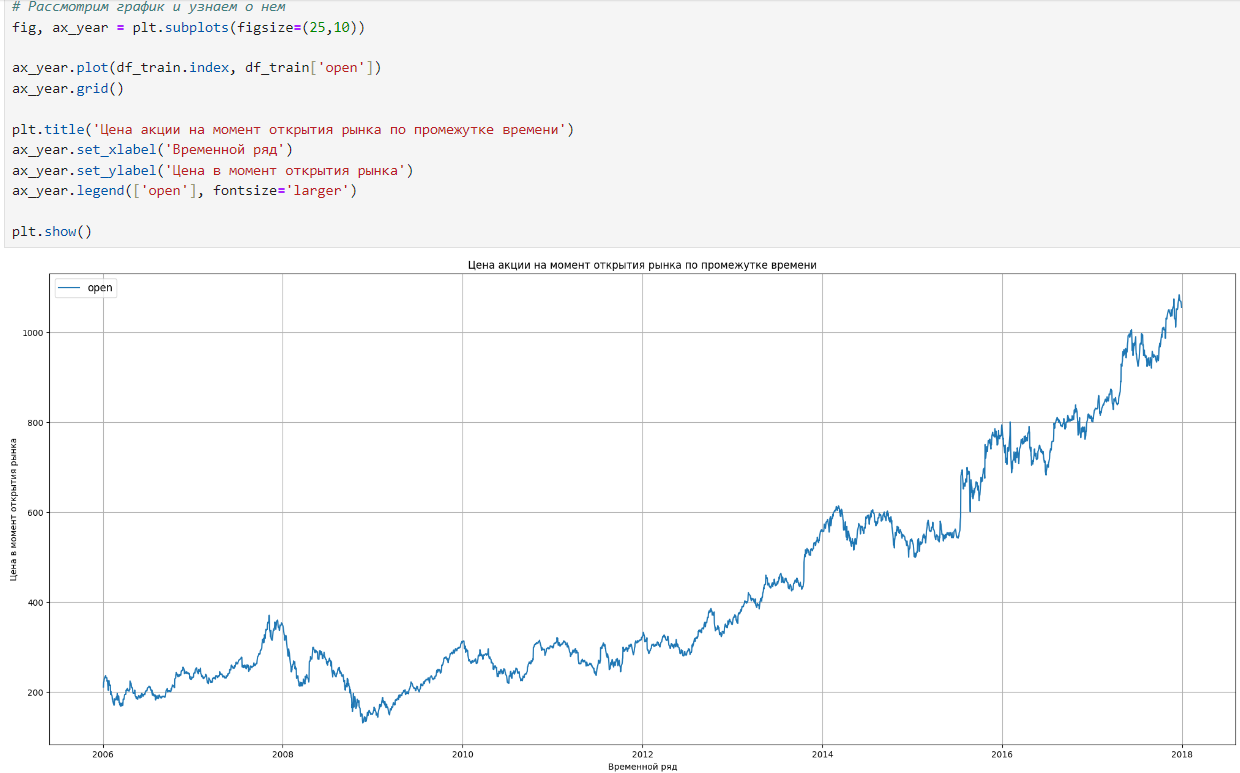
У нас есть значение в столбце name, мы его удалим, т.к. чтобы не было конфликта в обучающей и тестовой выборке.

****

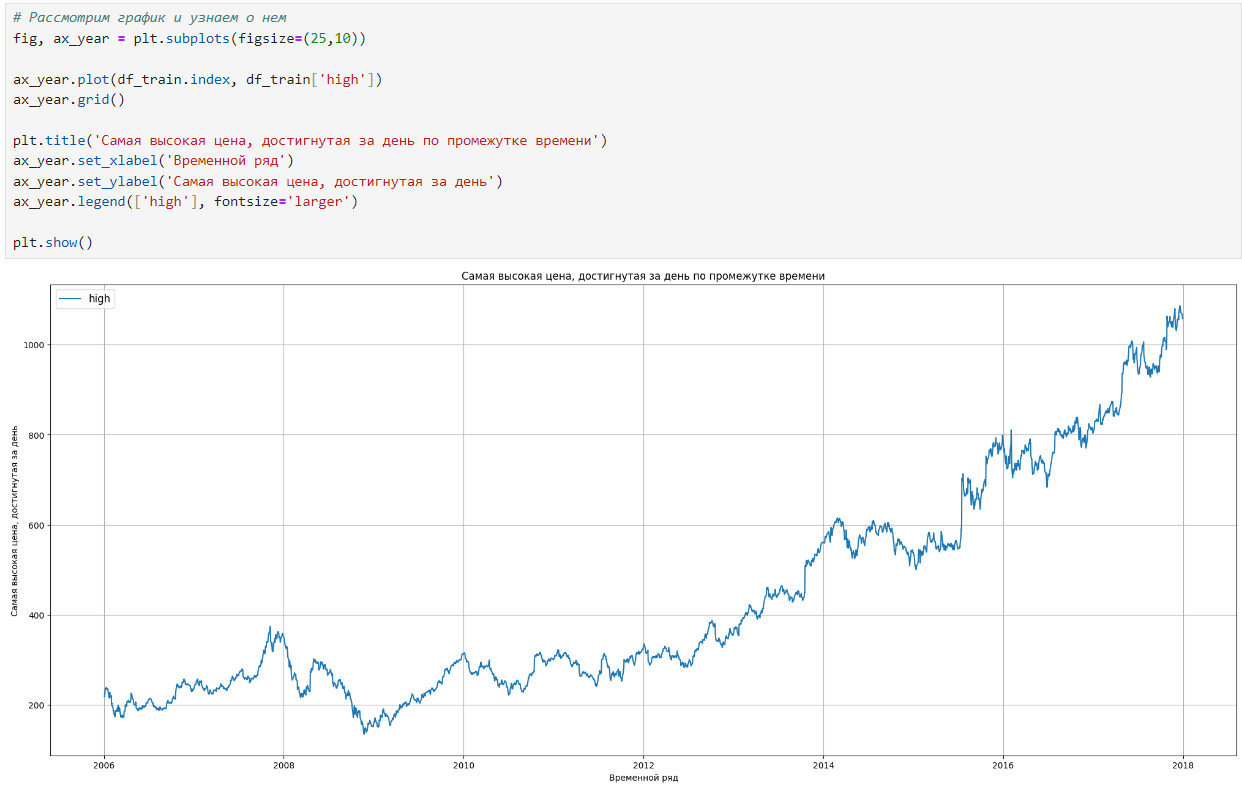
* + 1. **Добавление среднего значения**

****

* 1. **Анализ графики временного ряда**
     1. **Цена акций в момент открытия рынка**

Цена акций на момент открытия рынка увеличивается с каждым годом.

* + 1. **Самая высокая цена акций, достигнутая в день**

****

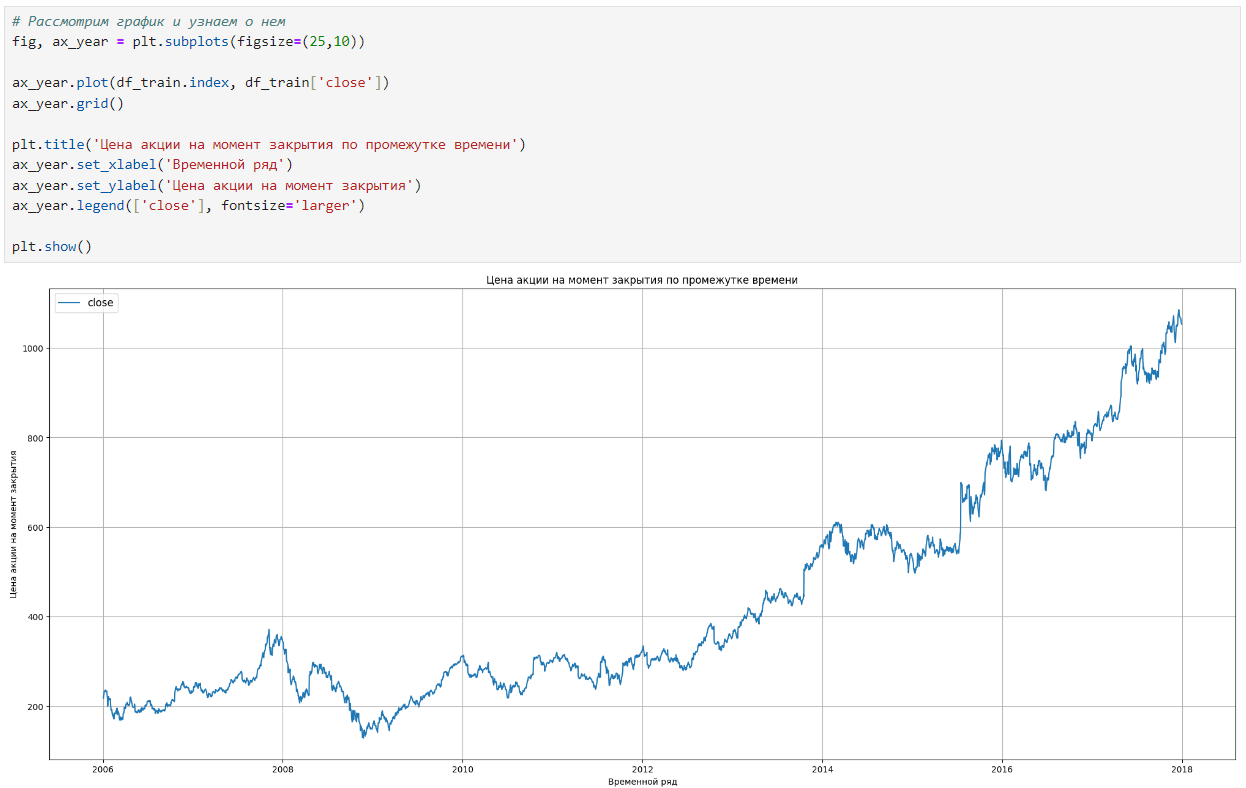
Самая высокая цена акций, достигнутая в день, увеличивается с каждым годом.

* + 1. **Самая низкая цена акций, достигнутая в день**

****

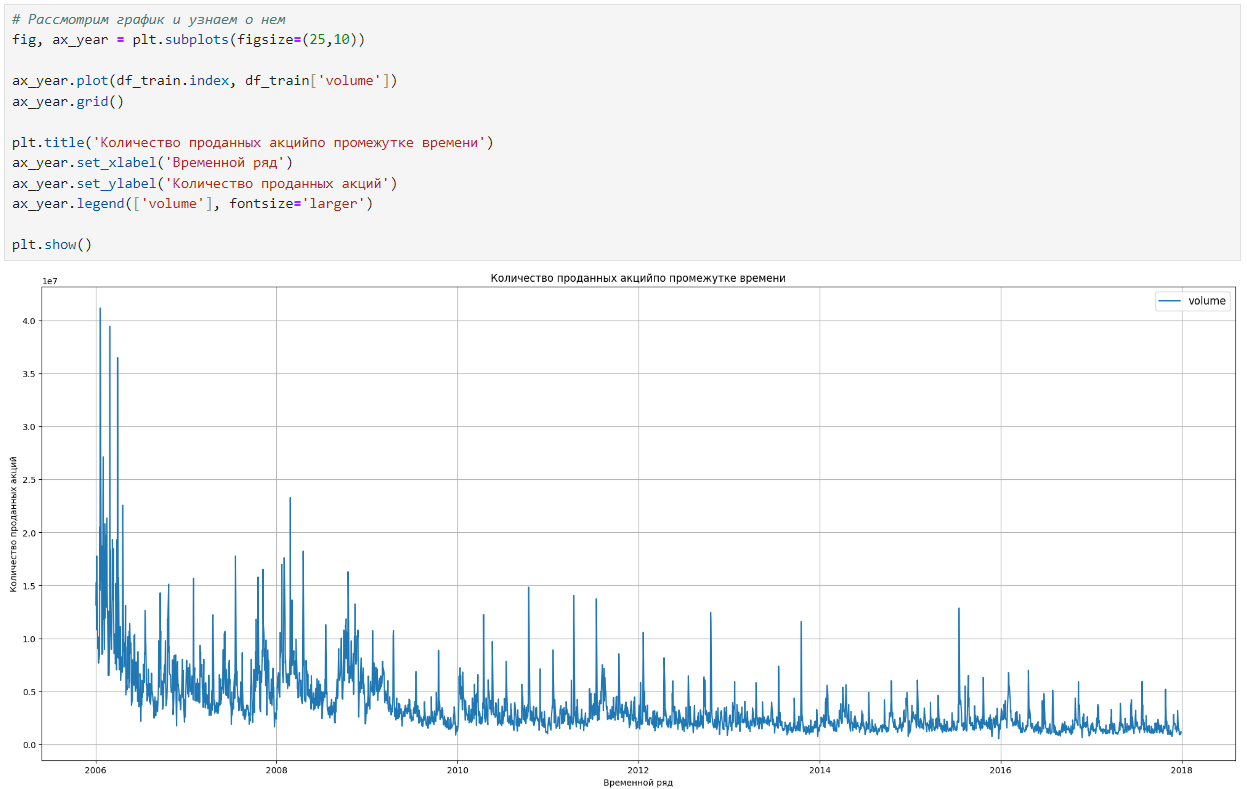
Самая низкая цена акций, достигнутая в день, увеличивается с каждым годом.

* + 1. **Цена акций в момент закрытия рынка**

****

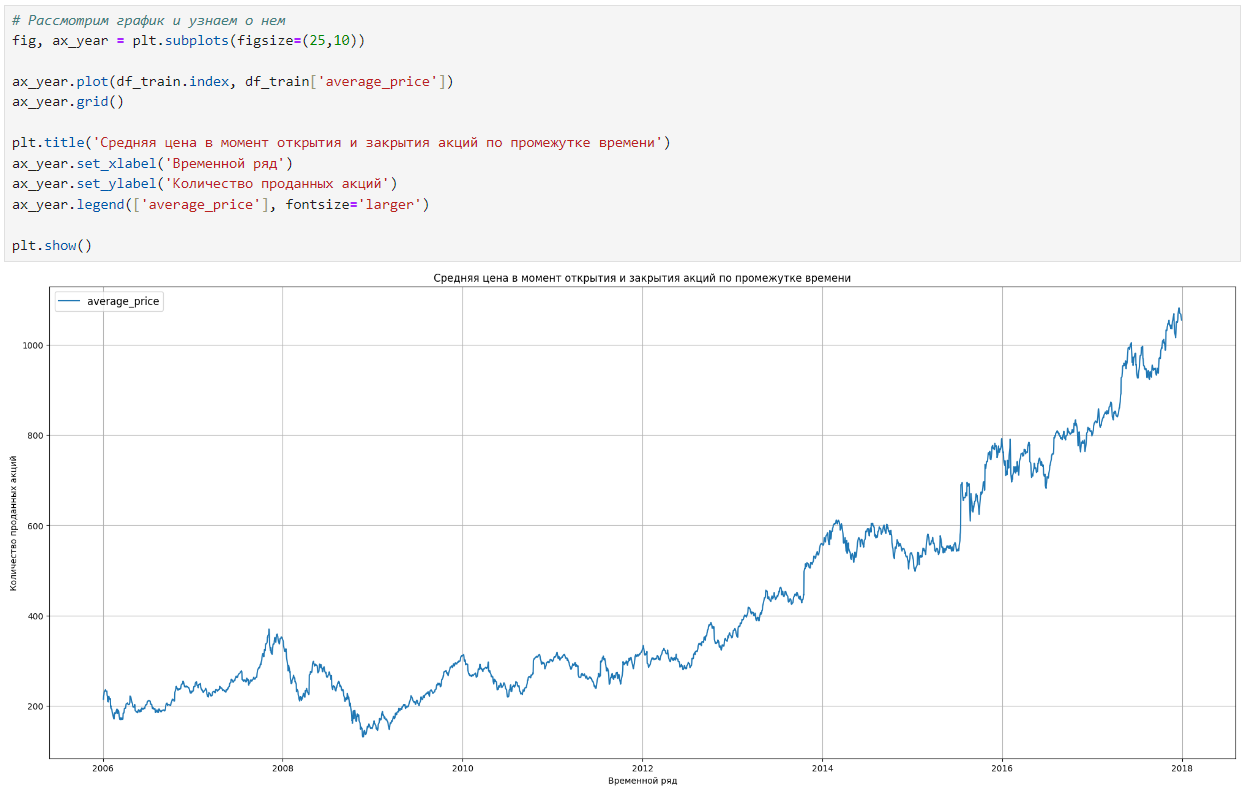
Цена акций на момент закрытия рынка увеличивается с каждым годом.

* + 1. **Количество проданных акций**

****

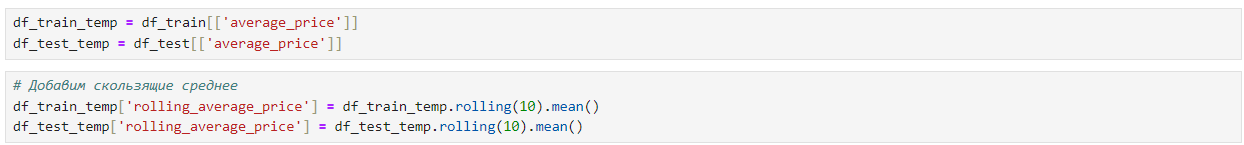
Количество проданных акций уменьшается с каждым годом.

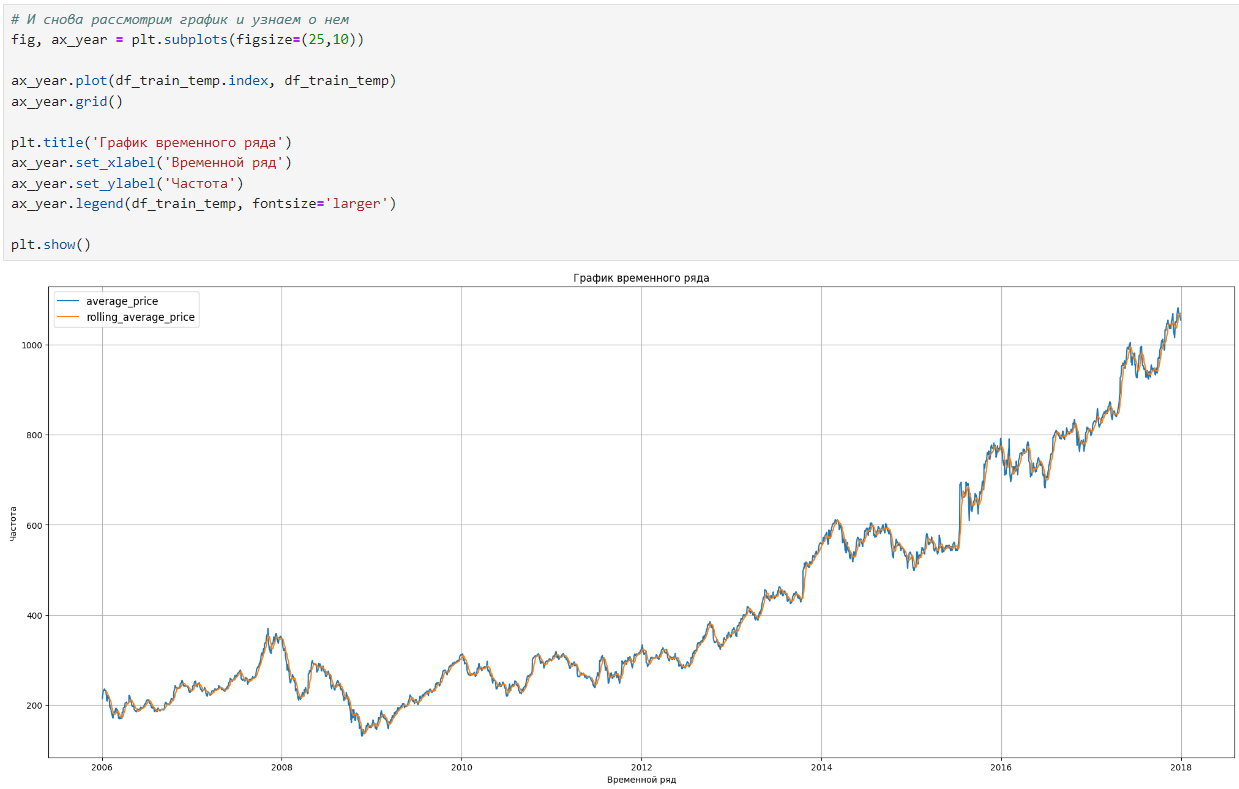
* + 1. **Средняя цена акций в момент открытия и закрытия рынка**

****

Средняя цена акций в момент открытия и закрытия рынка увеличивается с каждым годом.

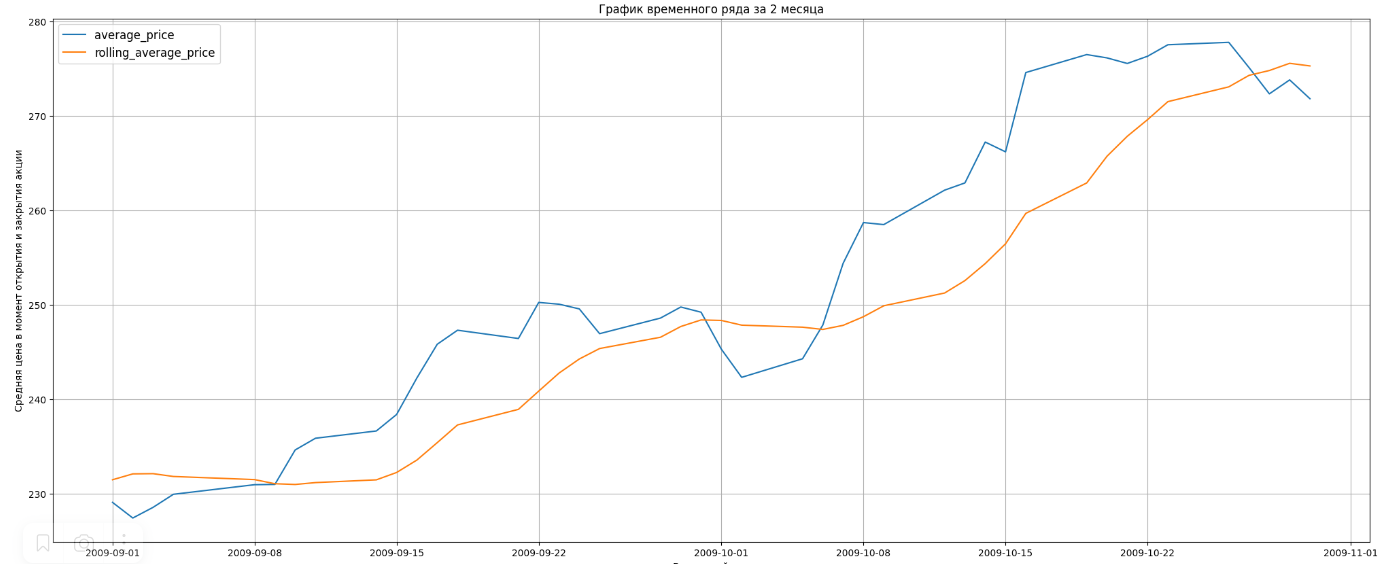
* + 1. **График со скользящими среднее**

****

****

Здесь получилось неудобной и неинформативной, рассмотрим определенный промежуток времени.

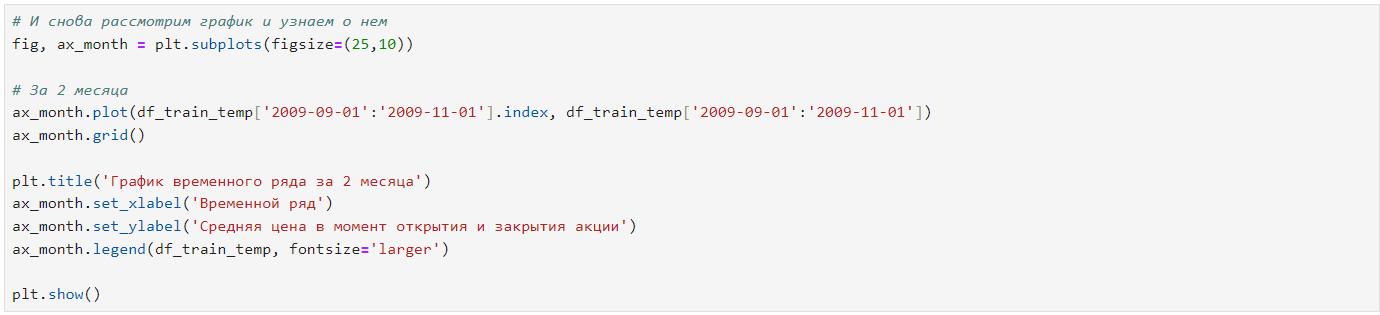
****

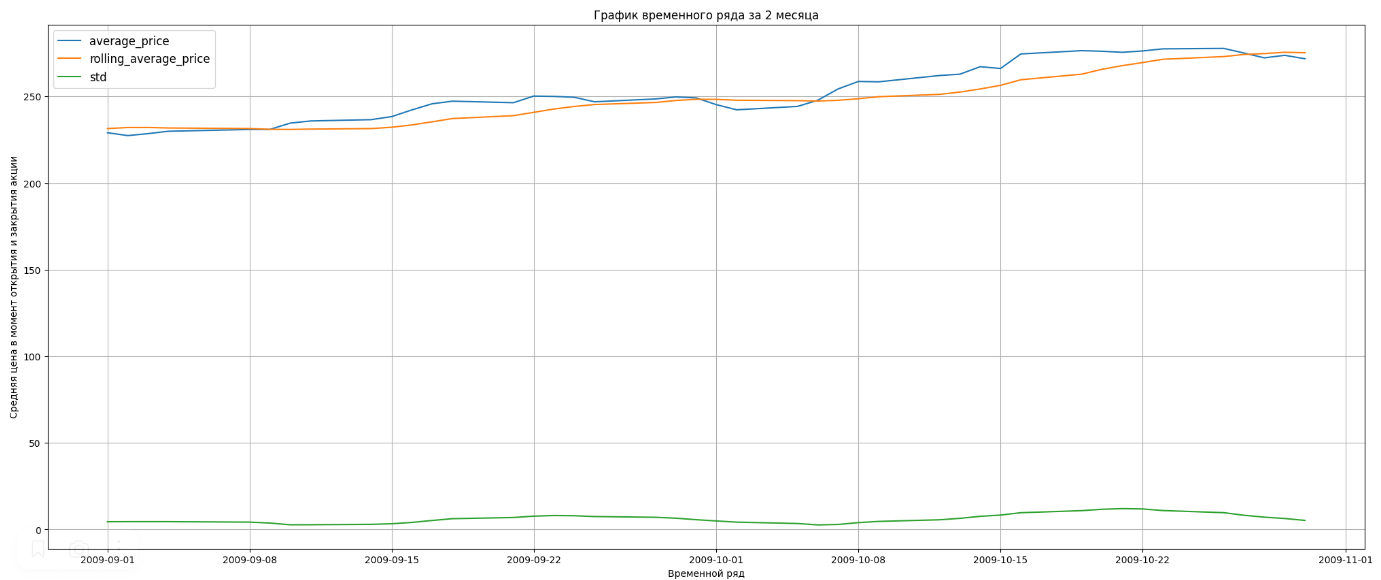
****

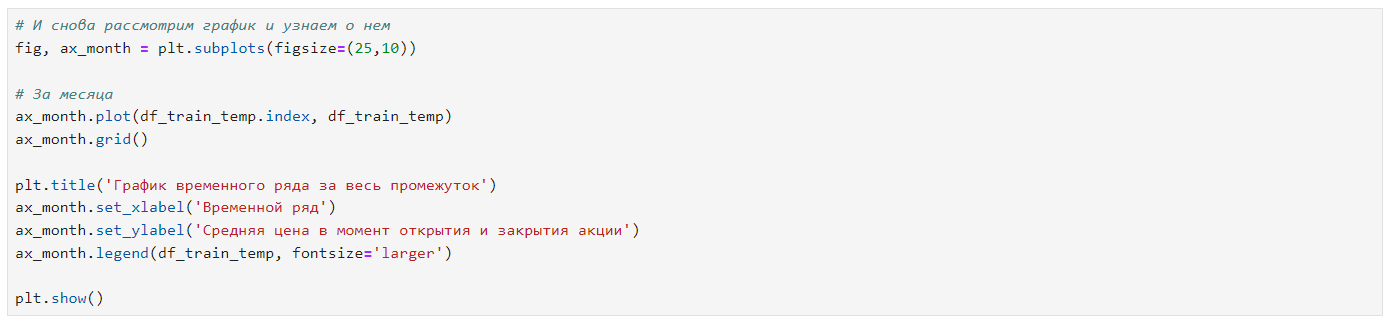
Рост средней цены акции за осень 2009 года

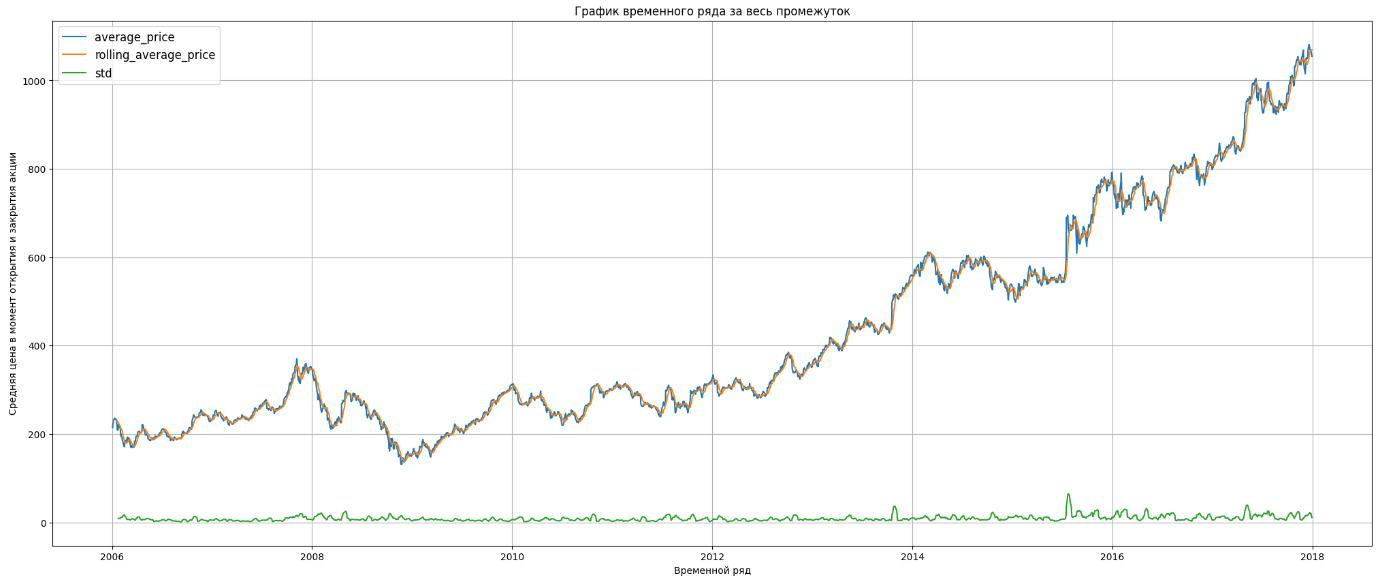
* + 1. **Исследование стационарных рядов**

****

****

****

****

****

Примечание:

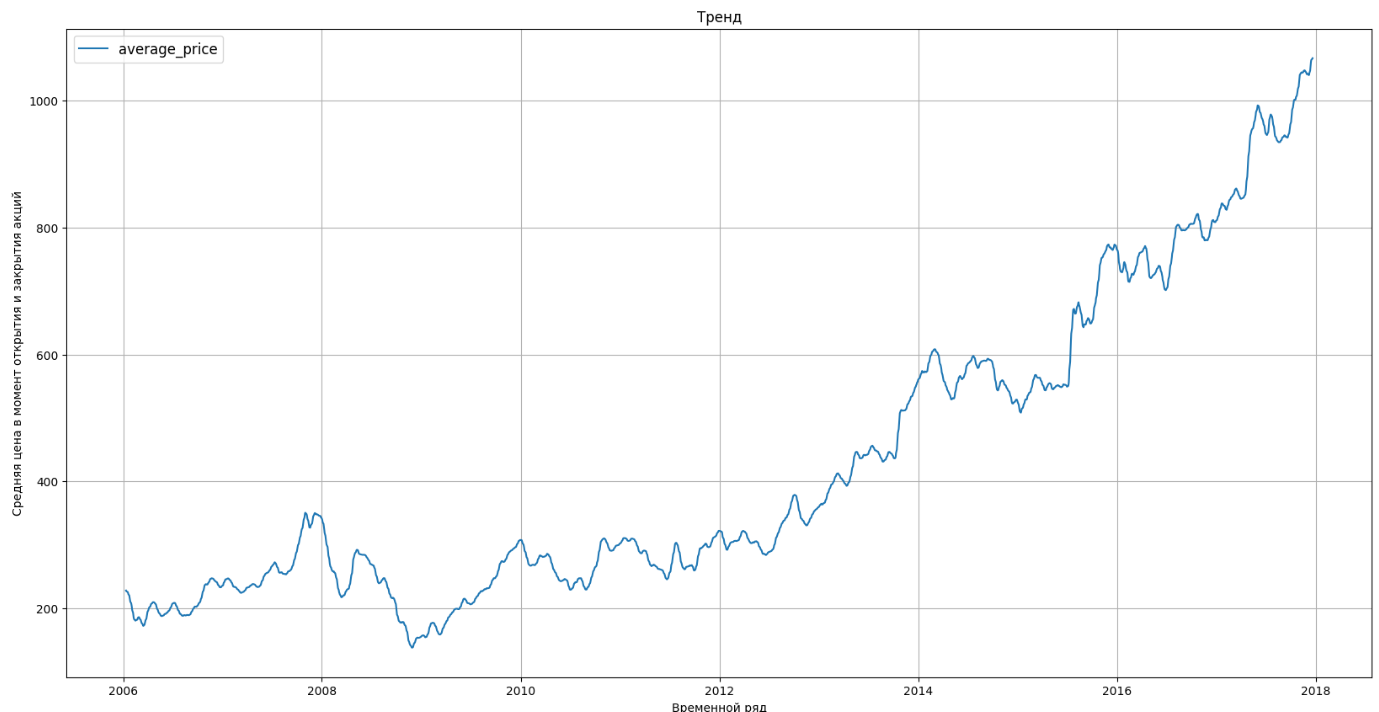
Стохастический процесс - это случайная величина, у которой со временем меняется её распределение. У этой величины есть среднее и дисперсия, которые тоже меняются Стохастический процесс стационарный (англ. stationary stochastic process), если его распределение со временем не меняется. Например, к такому процессу относятся периодические колебания значений.

Если распределение меняется, то процесс называется нестационарным.

Насчет временного ряда, его можно назвать стационарным или нет - нельзя, пока не проверим на тренды, сезонность и остаток декомпозиции.

* + - 1. **Тренды и сезонность**
         1. **Тренды**

****

****

Тренд в последнее время увеличивается.

* + - * 1. **Сезонность**

Рассмотрим сезонность в коротком промежутке, т.к. на длинном он не приносит мало полезных информаций. Рассмотрим за неделю.

****

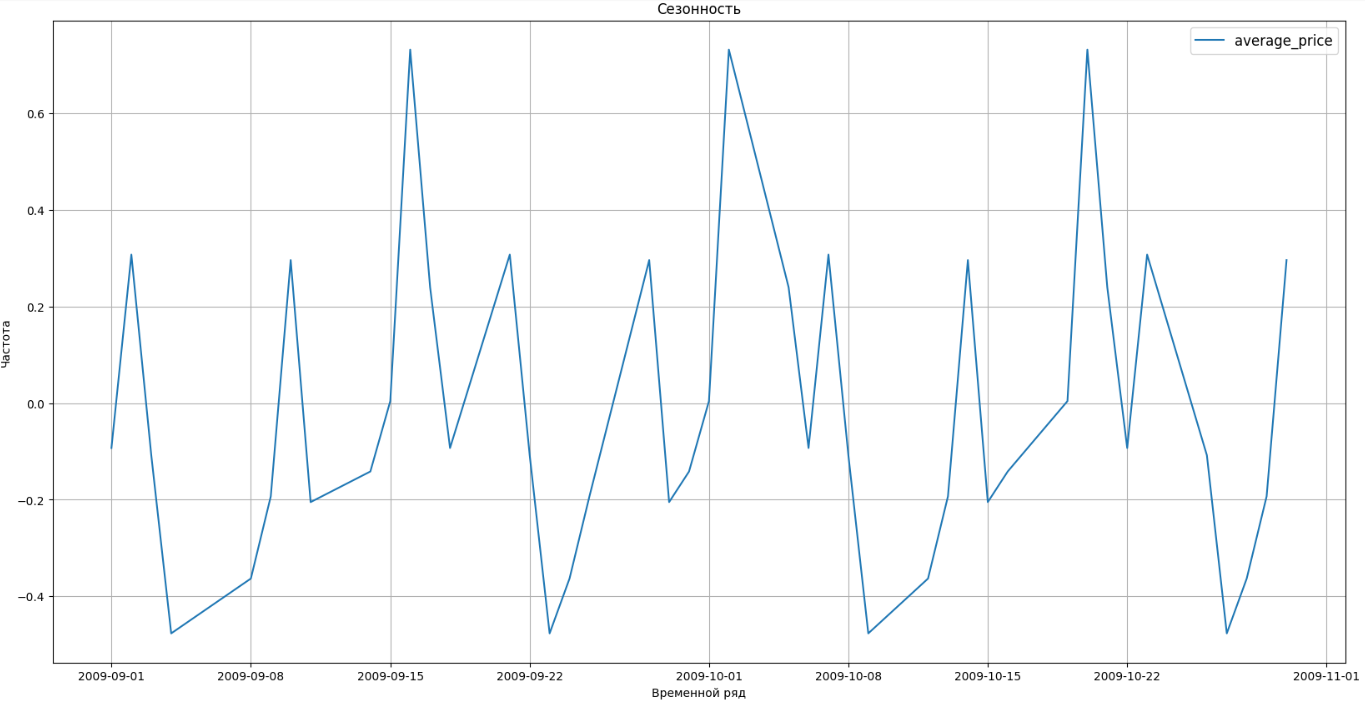
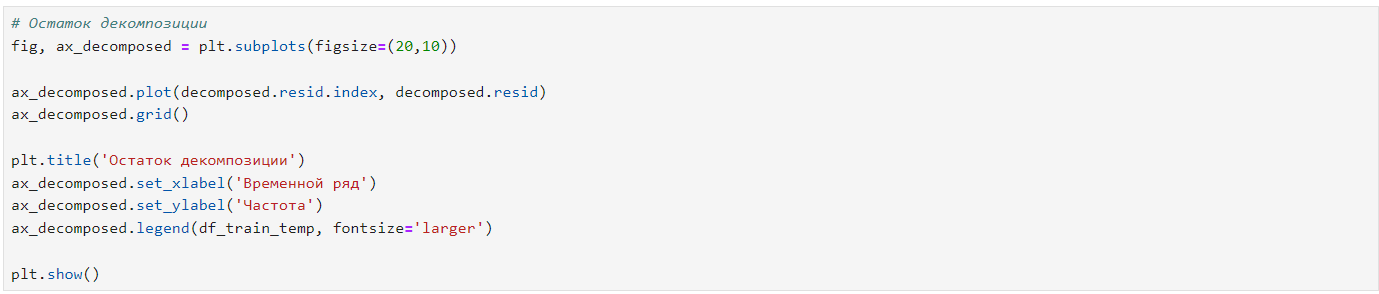
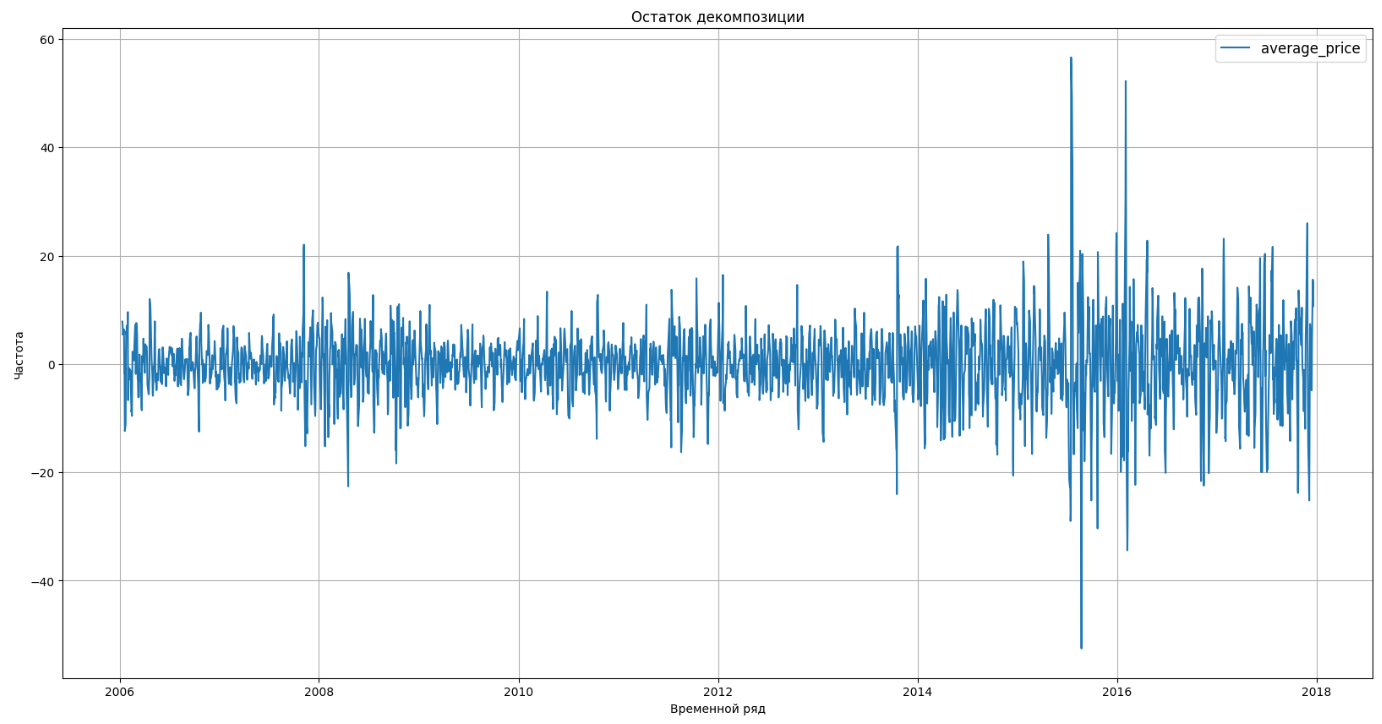
****

График представляет собой цикличным способом (выглядит почти как кардиограмма сердца). Можно предполагать, что этот временной ряд стационарен, однако все равно лучше убедиться математическим способом, который приведен ниже.

* + - * 1. **Остаток декомпозиции**

****

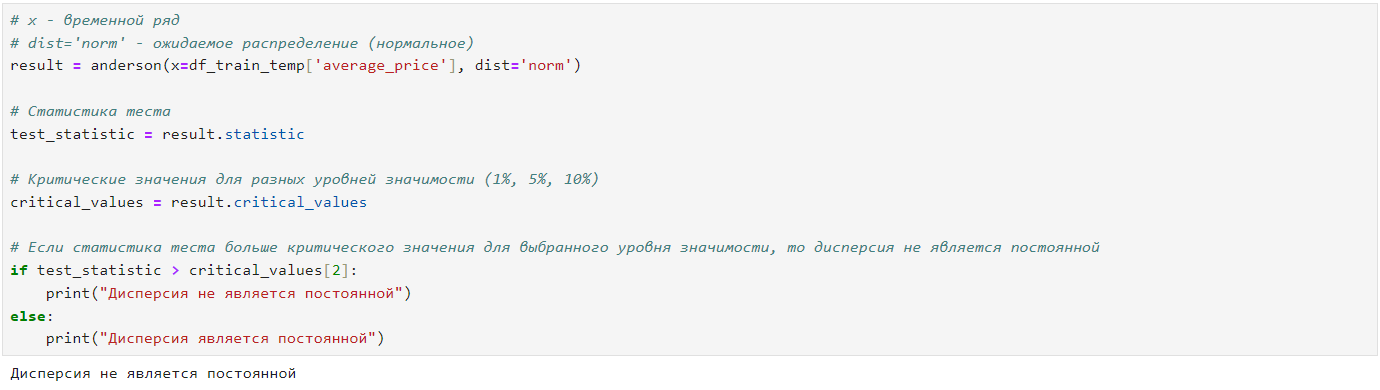
****

По поводу остатка декомпозиции нужно провести исследование отдельно, причем детально, поскольку декомпозиции остатков интрепетировать сложнее, чем сезонность и тренды - в идеале, график остатков должен содержать только шум без систематических компонентов.

Если остатки систематически распределены (например, отрицательны в первой части ряда и примерно равны нуля во второй) или включают некоторую периодическую компоненту, то это свидетельствует о неадекватности модели. Анализ остатков чрезвычайно важен и необходим при анализе временных рядов. Процедура оценивания предполагает, что остатки не коррелированы и нормально распределены.

* + - 1. **Проверка на дисперсию математическим способом**

Можно проверить дисперсию, применяя теста Андерсона-Дарлинга.

****

Этот промежуток времени, т.е. с 2006 до конца, является нестанционарным, т.к. присутствует тренд, который стремится увеличиваться. А также дисперсия не является постоянной. Проверим на станционарность математическим способом.

* + - 1. **Проверка на стационарность математическим способом**
         1. **Тест Дики-Фуллера**

****

* + - * 1. **Тест KPSS – Квятковский-Филлипс-Шмидт-ШинаДики-Фуллер**

****

ADF-тест не обнаруживает статистически значимого доказательства о стационарности временного ряда, в то время как KPSS-тест отвергает нулевую гипотезу о стационарности. Это противоречие может свидетельствовать о том, что ряд не является полностью стационарным, но также не является сильно нестационарным. Возможно, в ряде присутствует некоторая форма нестационарности, которая может быть устранена с помощью дополнительной обработки данных или моделирования.

* + 1. **Промежуточный вывод**

По графику видно, что с каждым годом цена акций увеличивается, по тренду это доказано.

Также на математическом языке были доказаны следующие тесты:

* На дисперсию
* На стационарность
* Тест Дики-Фуллера
* Тест KPSS - Квятковский-Филлипс-Шмидт-ШинаДики-Фуллер

Тест на дисперсию говорит, что не является постоянной, т.е. здесь получилось, что статистика теста больше критического значения для выбранного уровня значимости

А тесты на стационарность дали разные результаты - в одном отсутствует стационарность, а в другом имеется станционарность. Исходя из этих факторов, различные результаты тестов могут быть объяснены разными свойствами и характеристиками временного ряда. Важно учитывать контекст и особенности данных при интерпретации результатов тестов на стационарность.

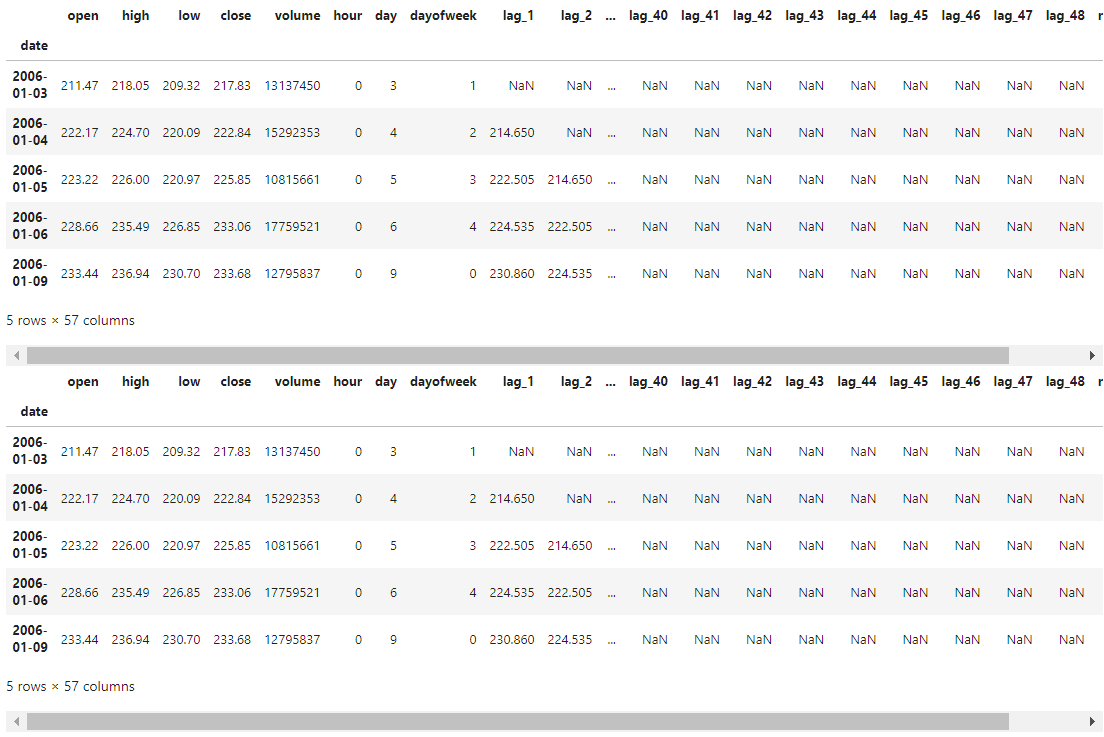
* 1. **Обучение и тестирование**

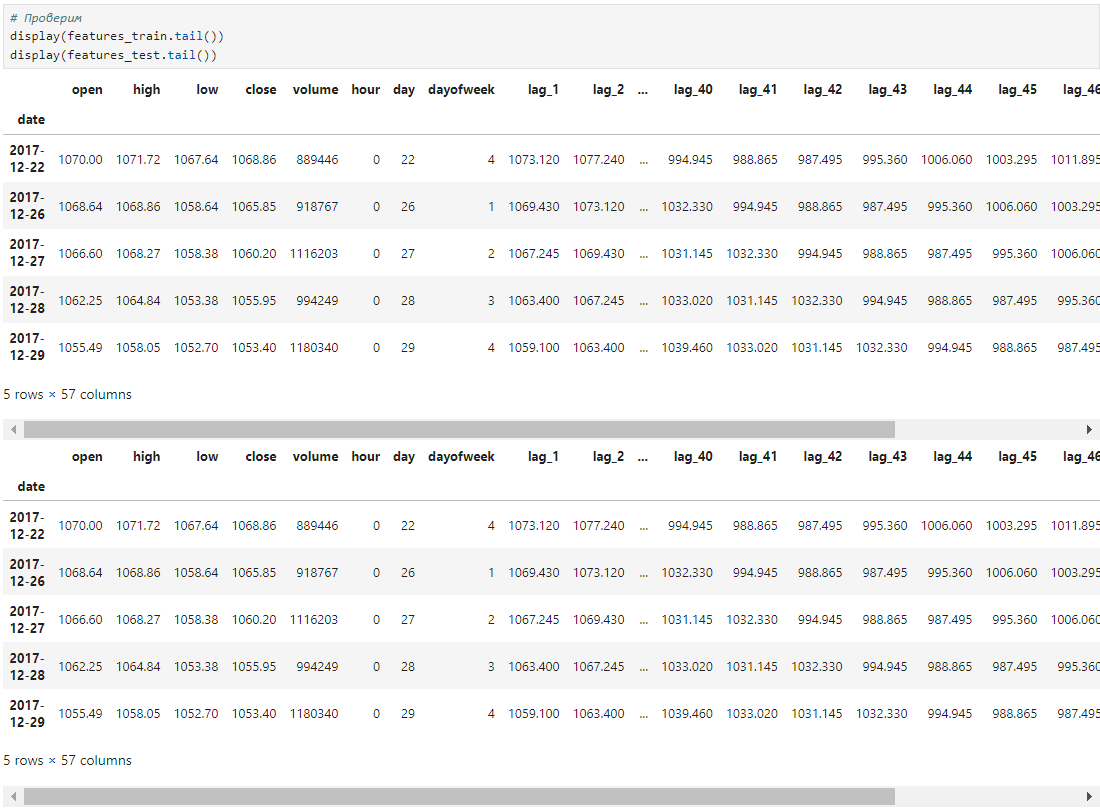
****

* + 1. **Деление на обучающий и тестовой выборки**

Имеется 2 датасета: обучающий и тестовый.

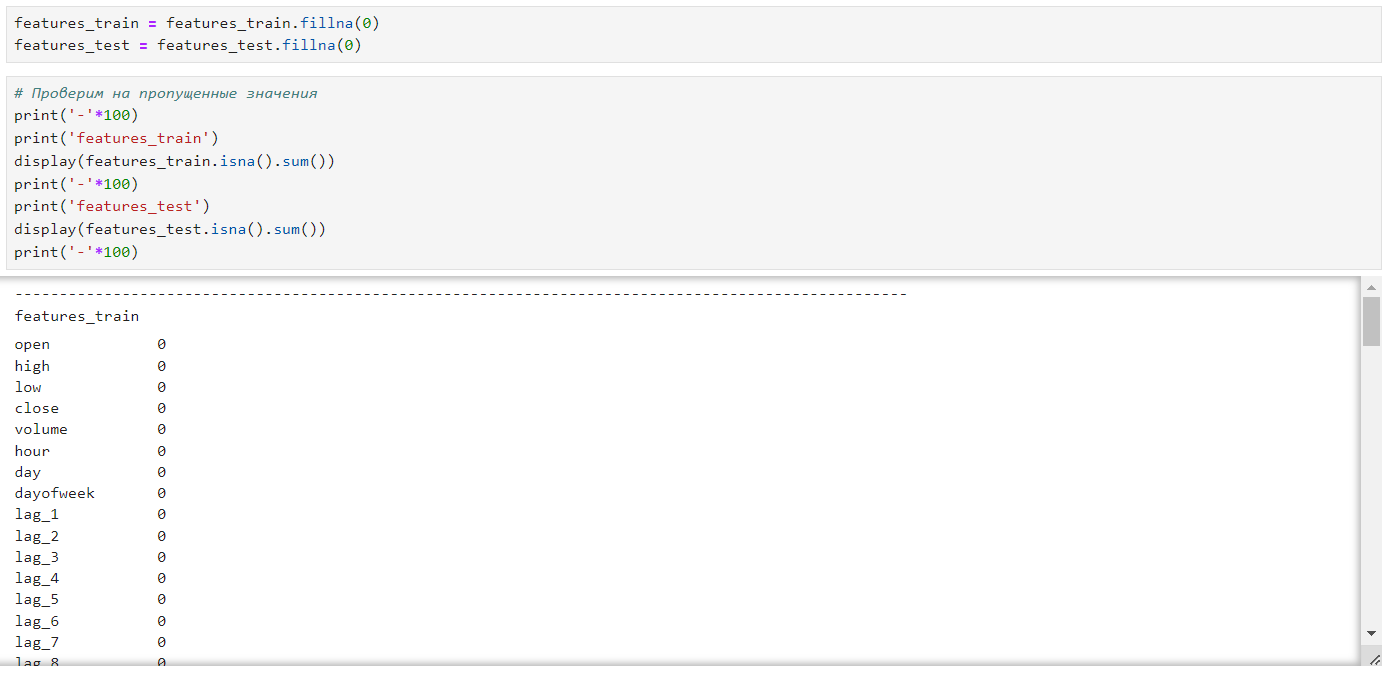
****

****

****

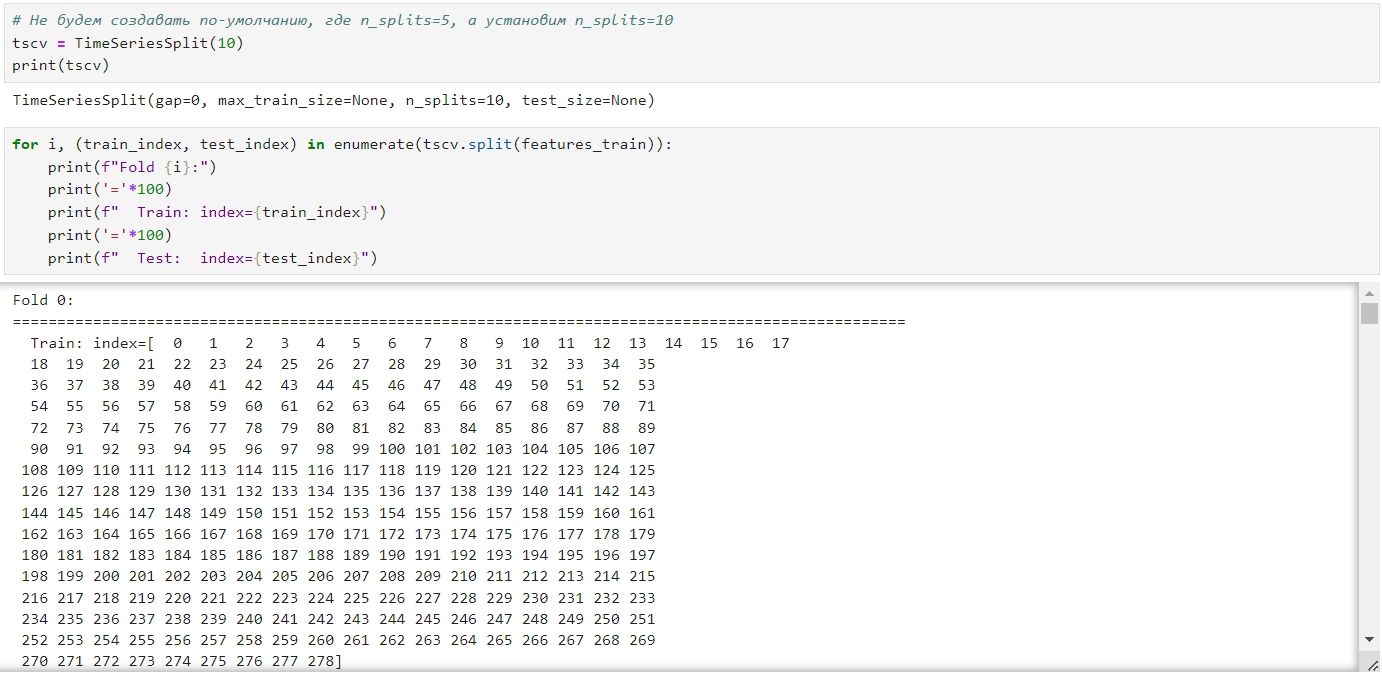
****

Есть пропущенных значений, заполним их значением 0.

****

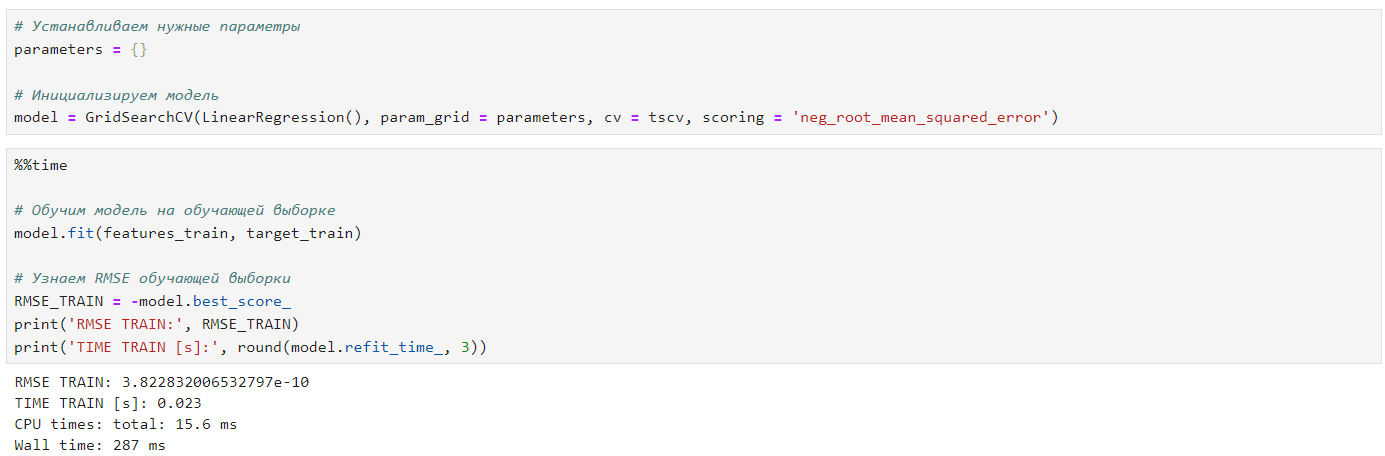
**Обучение моделей с применением TimeSeriesSplit.**

Мы обучаем модель **TimeSeriesSplit**, у которой по индексу стоит временной ряд. А чтобы они не пересекались, это поможет инструмент TimeSeriesSplit. Как работает, рассмотрим ниже.

****

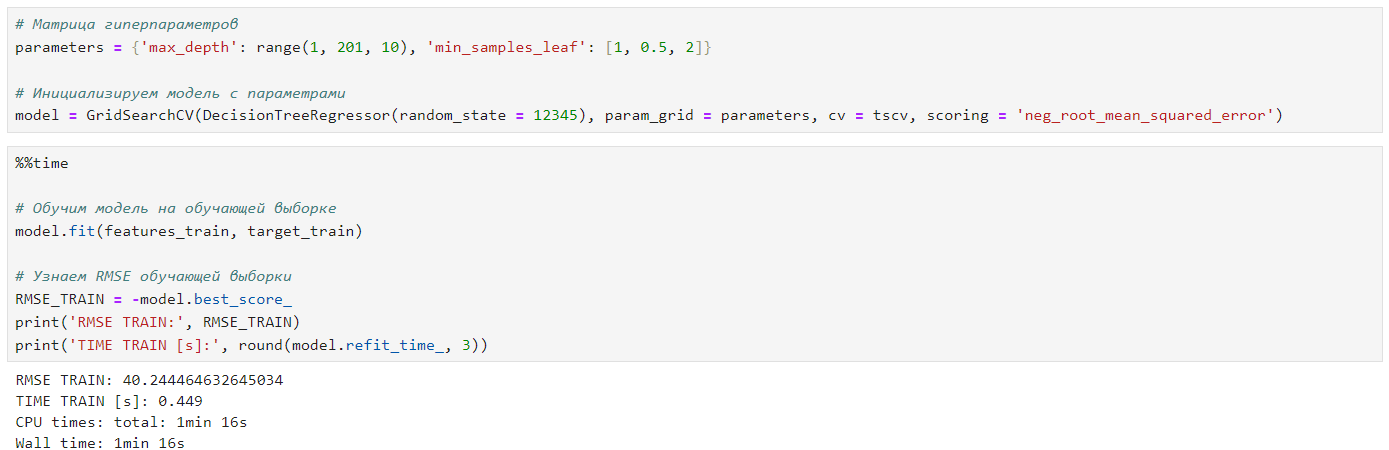
Как и видим, что разделен на несколько промежутков времени, по 10 блоков времени.

* + 1. **Машинное обучение**
       1. **LinearRegression**

****

****

* + - 1. **DecisionTreeRegressor**

****

****

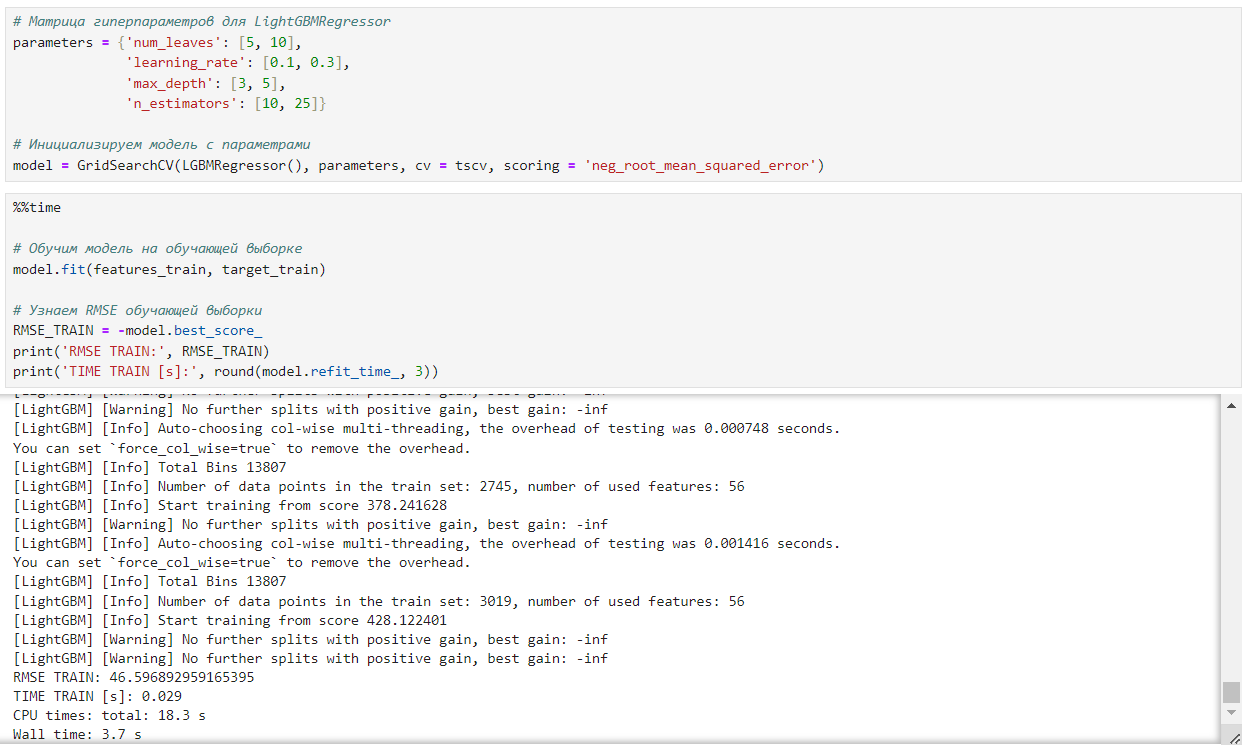
* + - 1. **RandomForestRegressor**

****

****

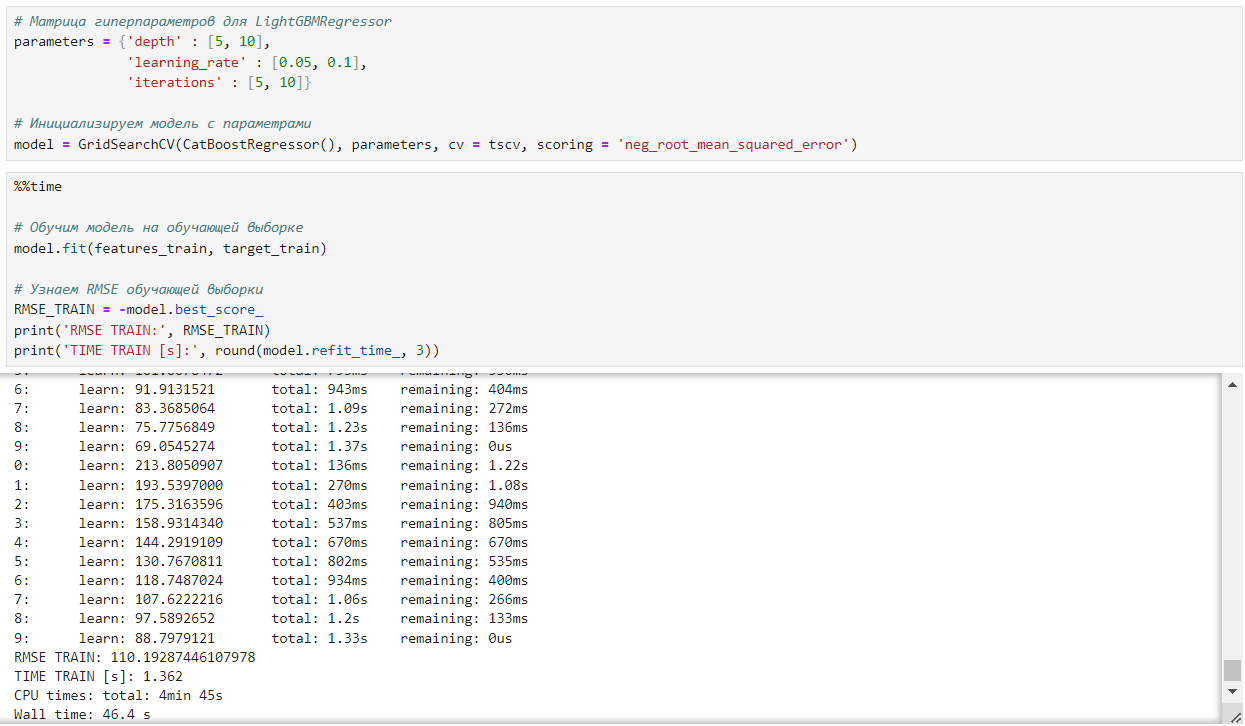
****

* + - 1. **LightCBMRegressor**

****

****

* + - 1. **CatBoostRegressor**

****

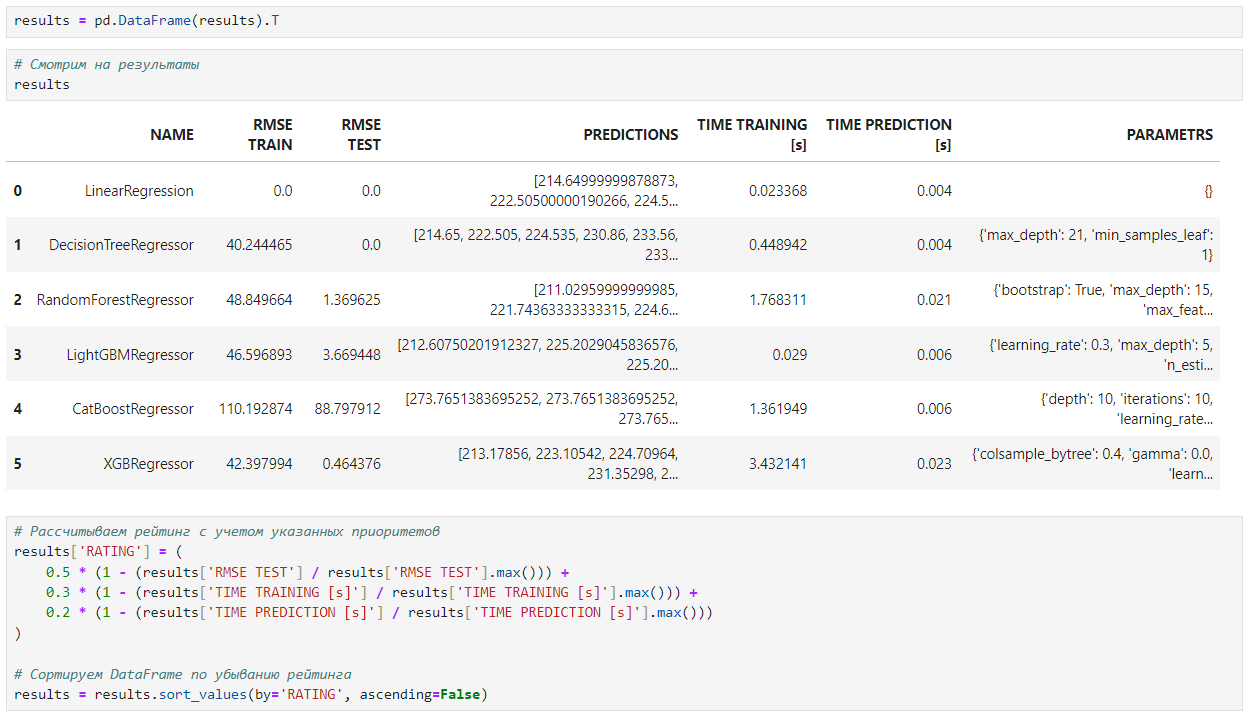
****

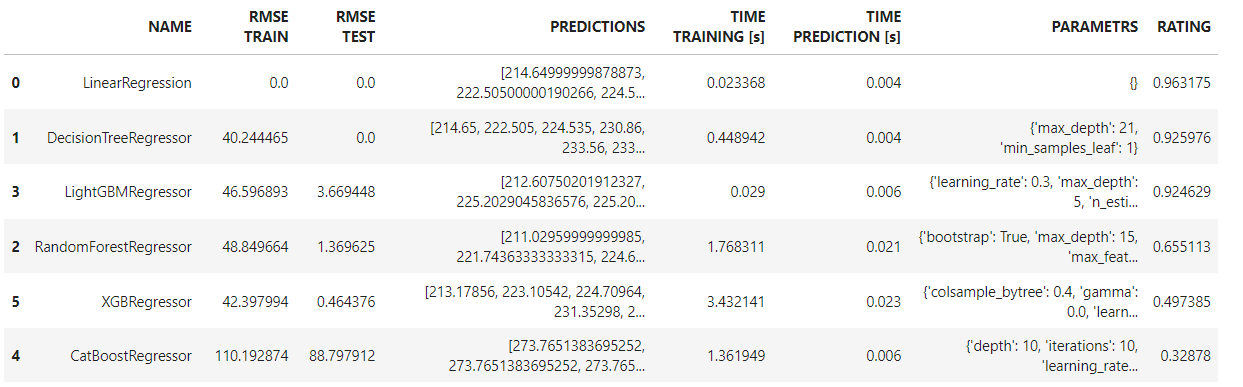
* + - 1. **XGBRegressor**

****

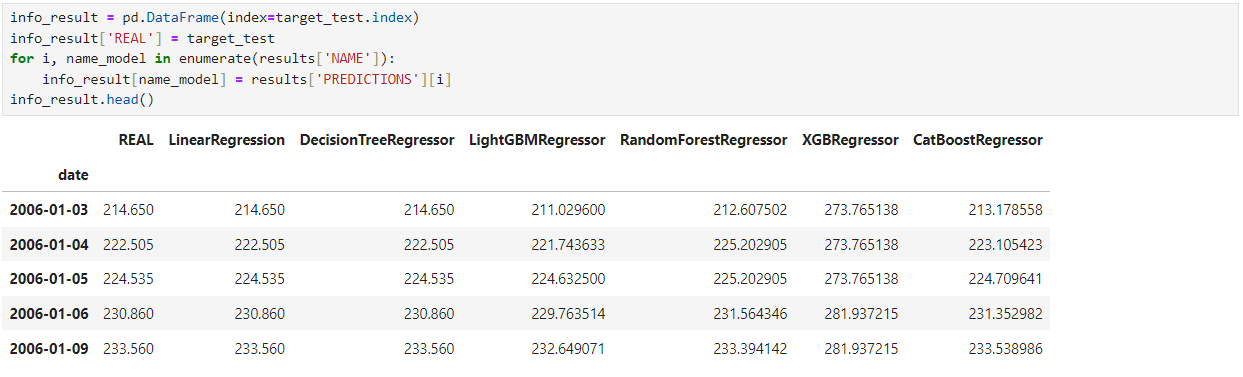
****

* + 1. **Результаты**

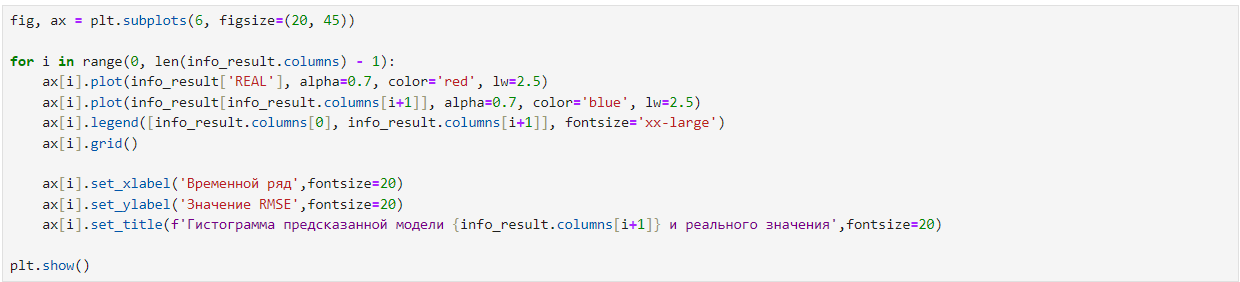
****

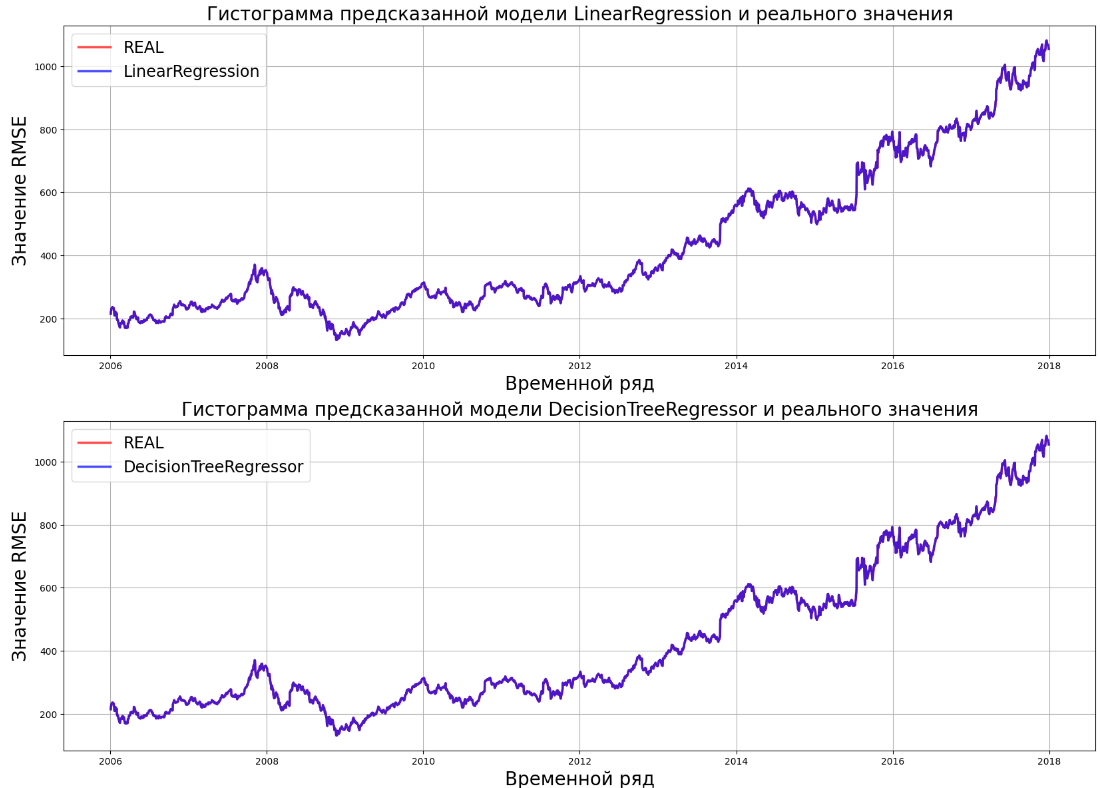
****

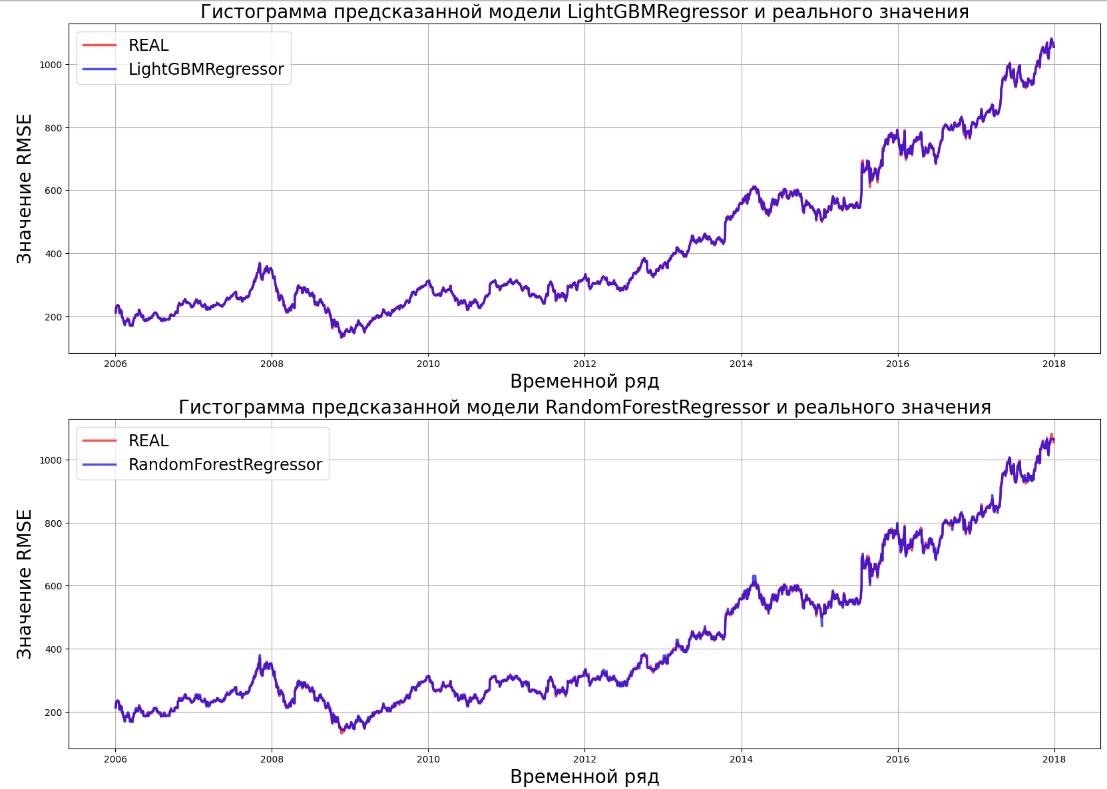
Создадим отдельный датасет, в котором будет храниться реальное значение и предсказания модели.

****

Сразу все данные в одном графике нежелательно уложиться, т.к. не очень информативно, поэтому разобъем на несколько блоков.









Предсказания модели в некотором степени отличаются от реальных значений.

* 1. **Вывод машинного обучения**

Лучшие модели - LightGBMRegressor и Linear Regression, с низкой ошибкой RMSE и быстрым временем обучения и предсказания. RandomForestRegressor и XGBRegressor также показывают хорошие результаты, но требуют больше времени на обучение. DecisionTreeRegressor имеет нулевую ошибку на обучающем наборе, что может указывать на переобучение. CatBoostRegressor показывает наихудшие результаты.

## Вывод

В ходе лабораторной работы изучили основные методы анализа и прогнозирование временных рядов.