|  |  |
| --- | --- |
|  | **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   
Папин А.В.

подпись, дата

Проверил:

к.т.н., доц., Ю.Е. Гапанюк

подпись, дата

2024 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Цель лабораторной работы 3](#_Toc5)

[2. Описание задание 3](#_Toc6)

[3. Основные характеристики датасета 4](#_Toc7)

[4. Листинг 5](#_Toc8)

# **Цель лабораторной работы**

Изучение основных методов анализа и прогнозирование временных рядов.

# **Описание 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: Daily Climate time series data (Ежедневные данные о временных рядах изменения климата)

Ссылка: https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data

**О датасетах**

Набор данных полностью предназначен для разработчиков, которые хотят обучить модель прогнозированию погоды для климата Индии. Этот набор данных предоставляет данные за период с 1 января 2013 года по 24 апреля 2017 года в городе Дели, Индия.

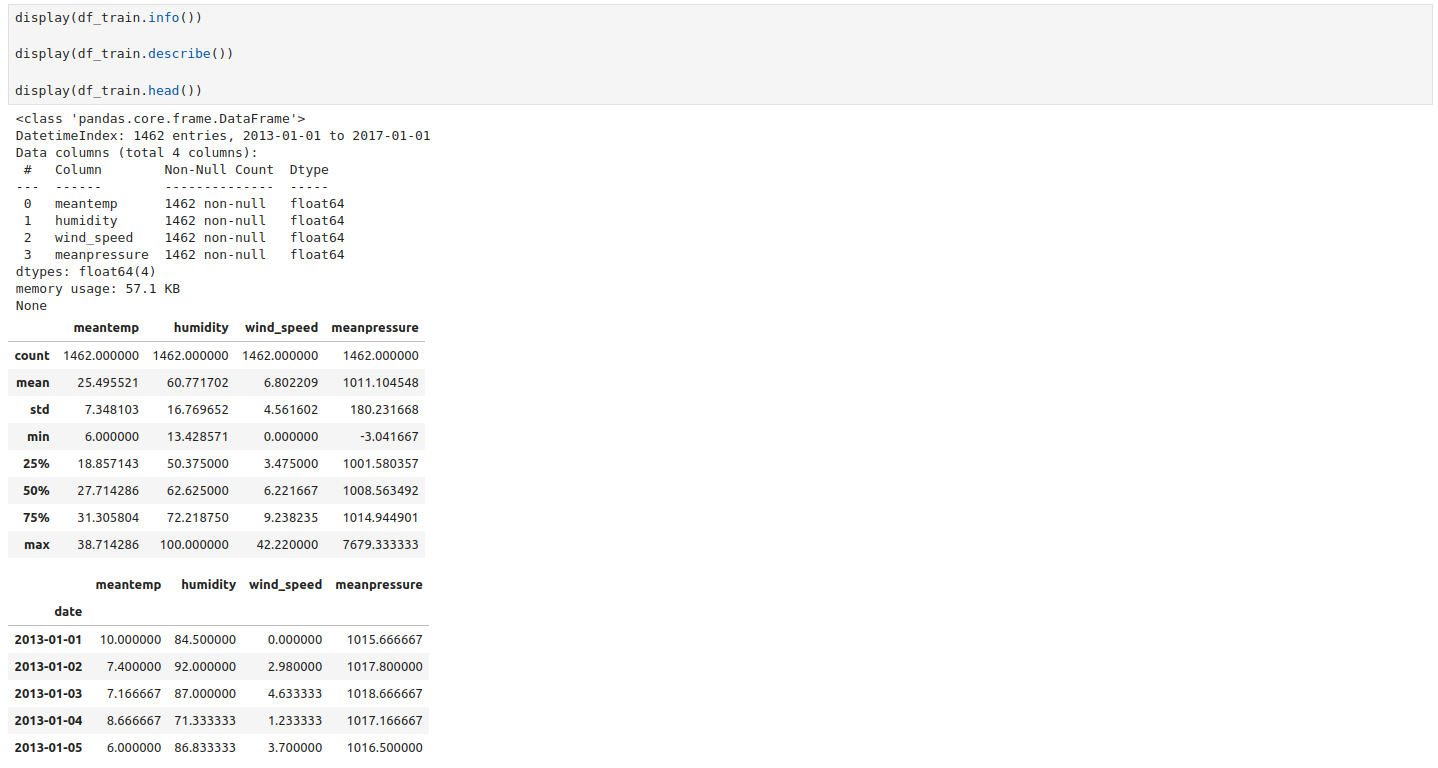
**Структура данных**

Здесь указаны 4 параметра:

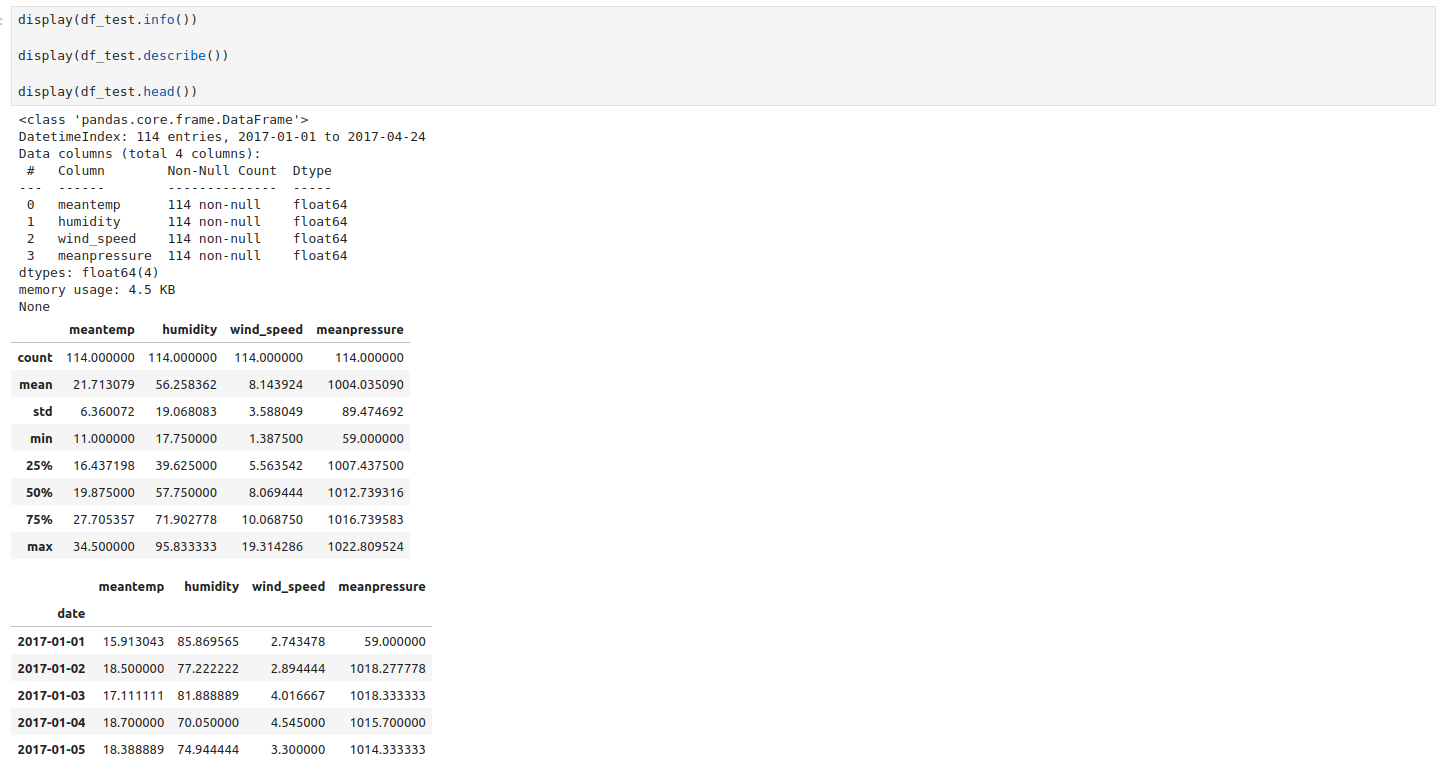
* meantemp - средняя температура [усредненный по часовому интервалу 3 часа в течении дня]
* humidity - влажность [грамм ^ 3]
* wind\_speed - скорость ветра [км/ч]
* meanpressure - среднее давление [атм]

# **Листинг**

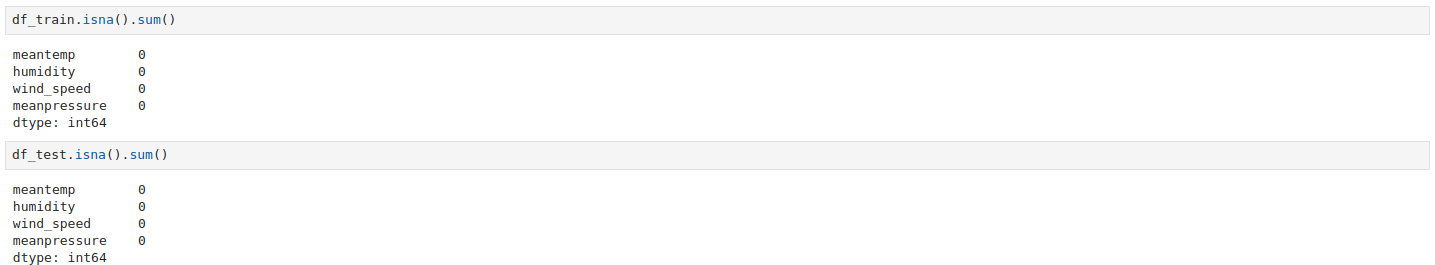
* 1. **Изучение данных**
     1. **Обучающий датасет**



* + 1. **Тестовый датасет**



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



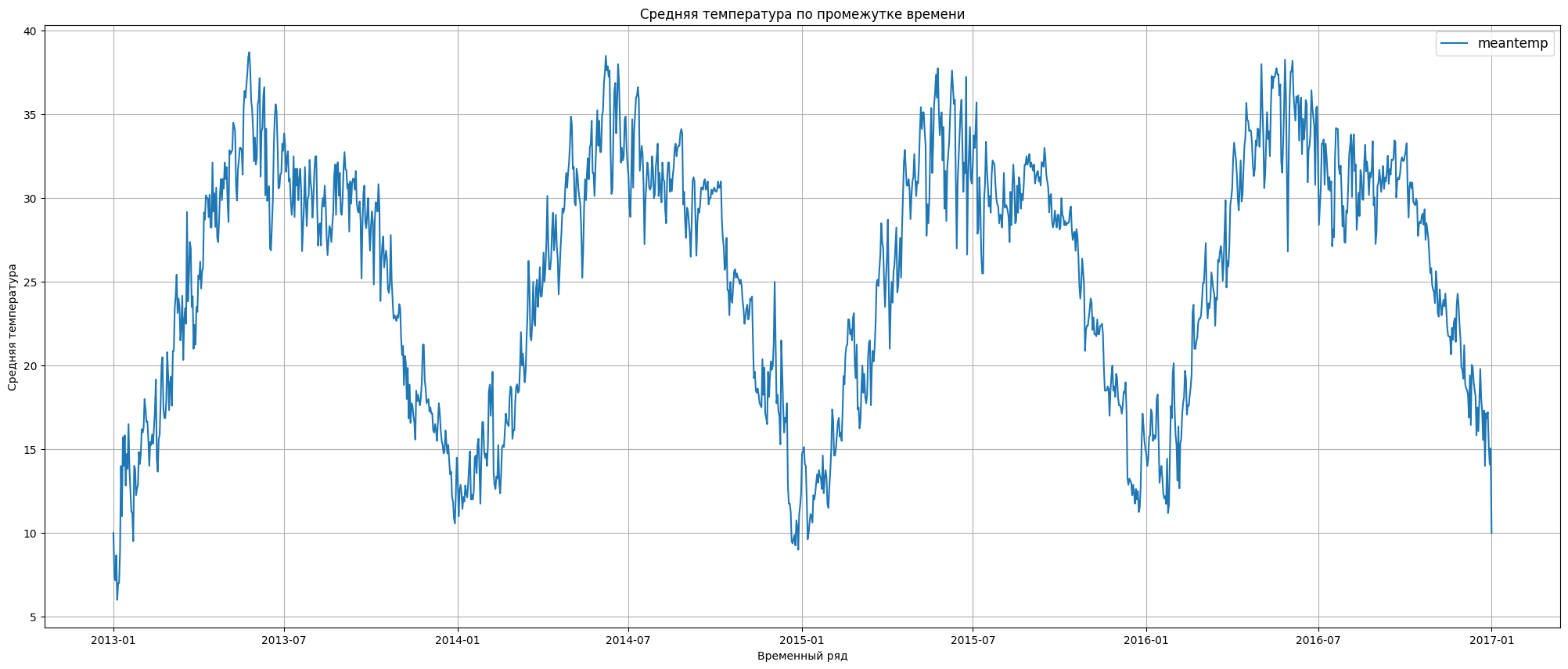
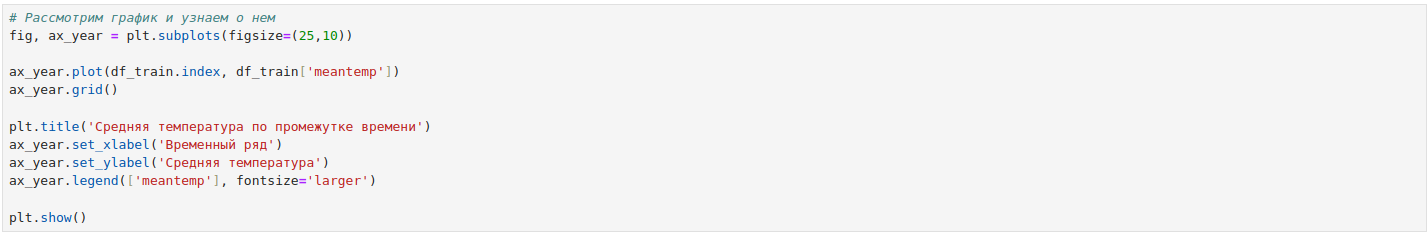
Отсутствует пропуски, это и хорошо

* 1. **Дублирование**



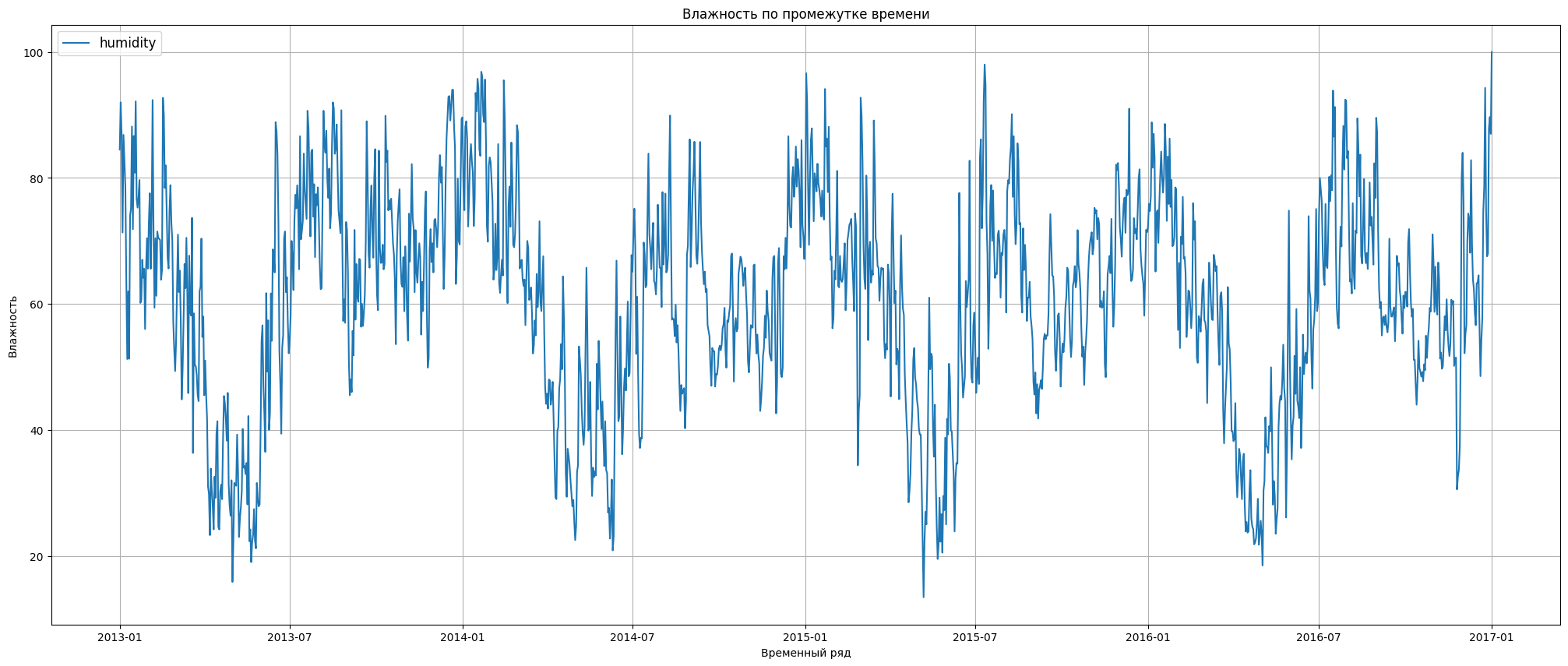
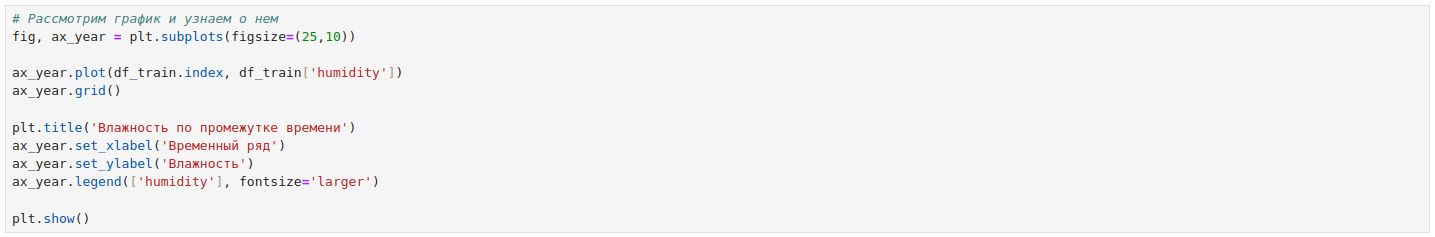
Аналогично, отсутствует дублирование, это и хорошо

* 1. **Анализ графика временного ряда**
     1. **Средняя температура**



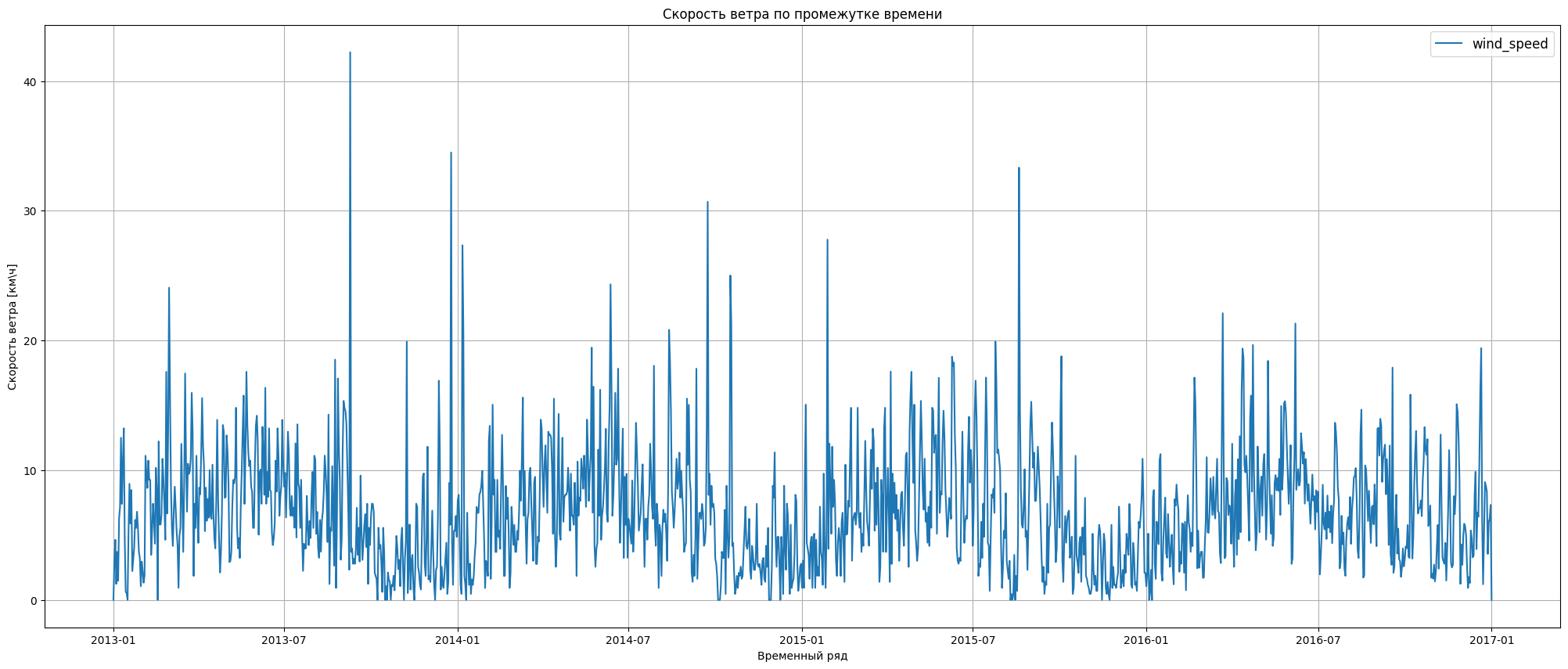
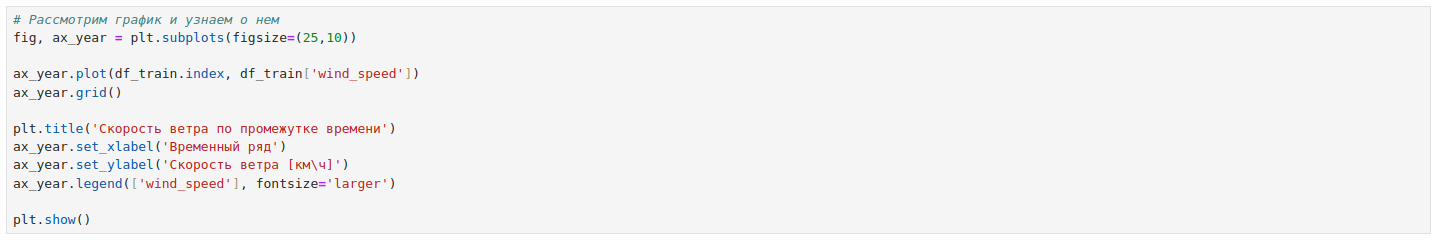
Наблюдается тенденция увеличения температуры в летний сезон и спадения в зимний сезон

* + 1. **Влажность**



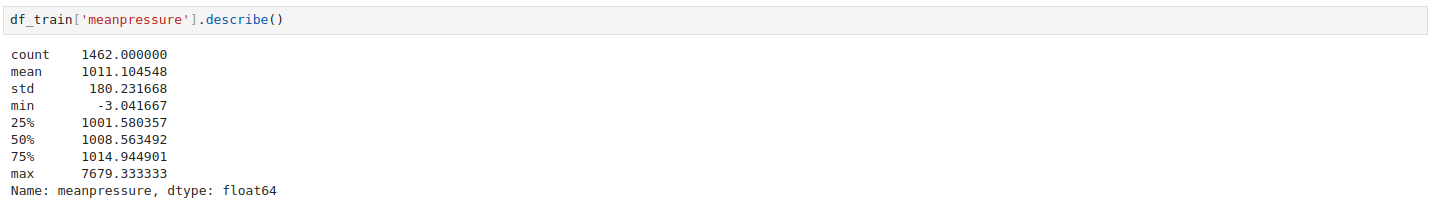
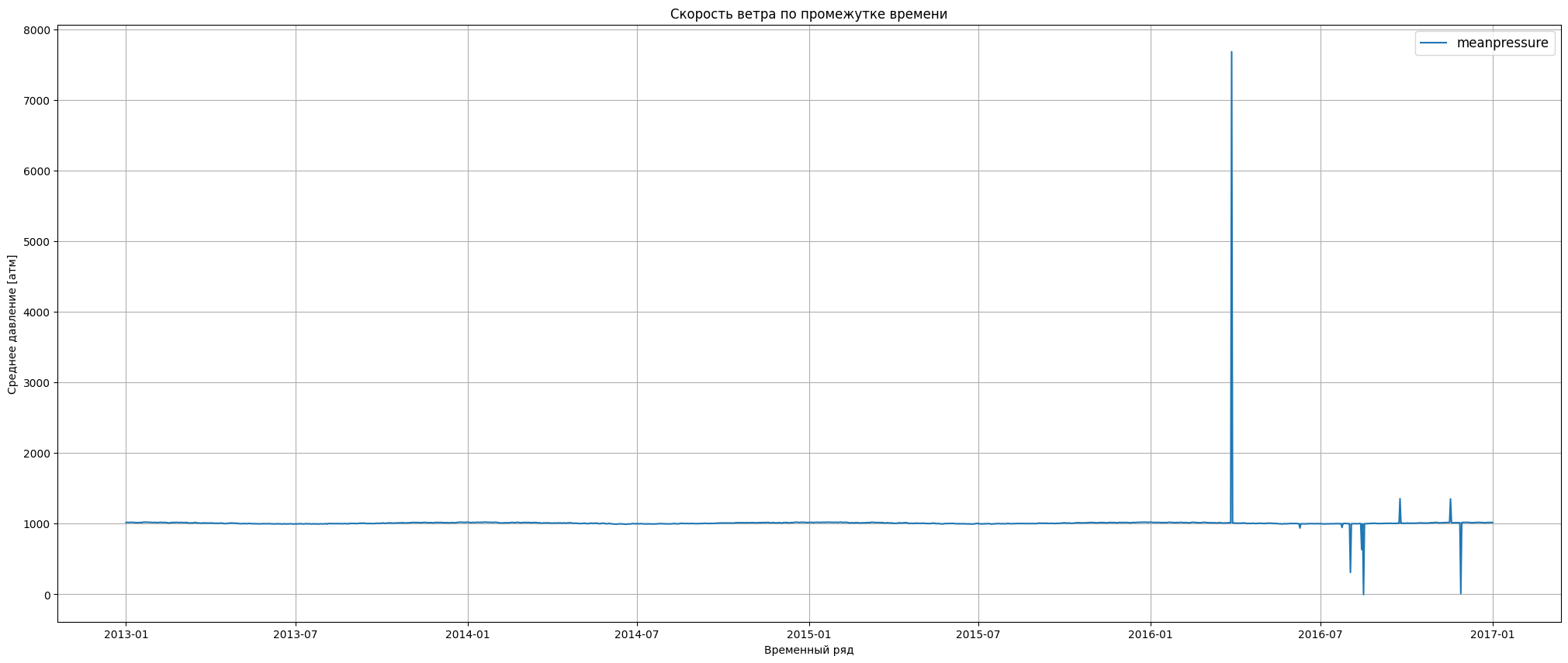
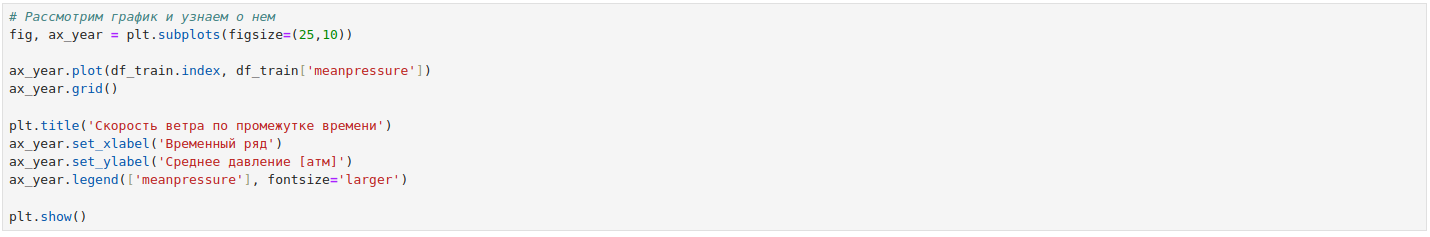
Здесь наблюдается противоположность по отношению к температуры, т.е. температура увеличивается, влажность уменьшается

* + 1. **Скорость ветра**

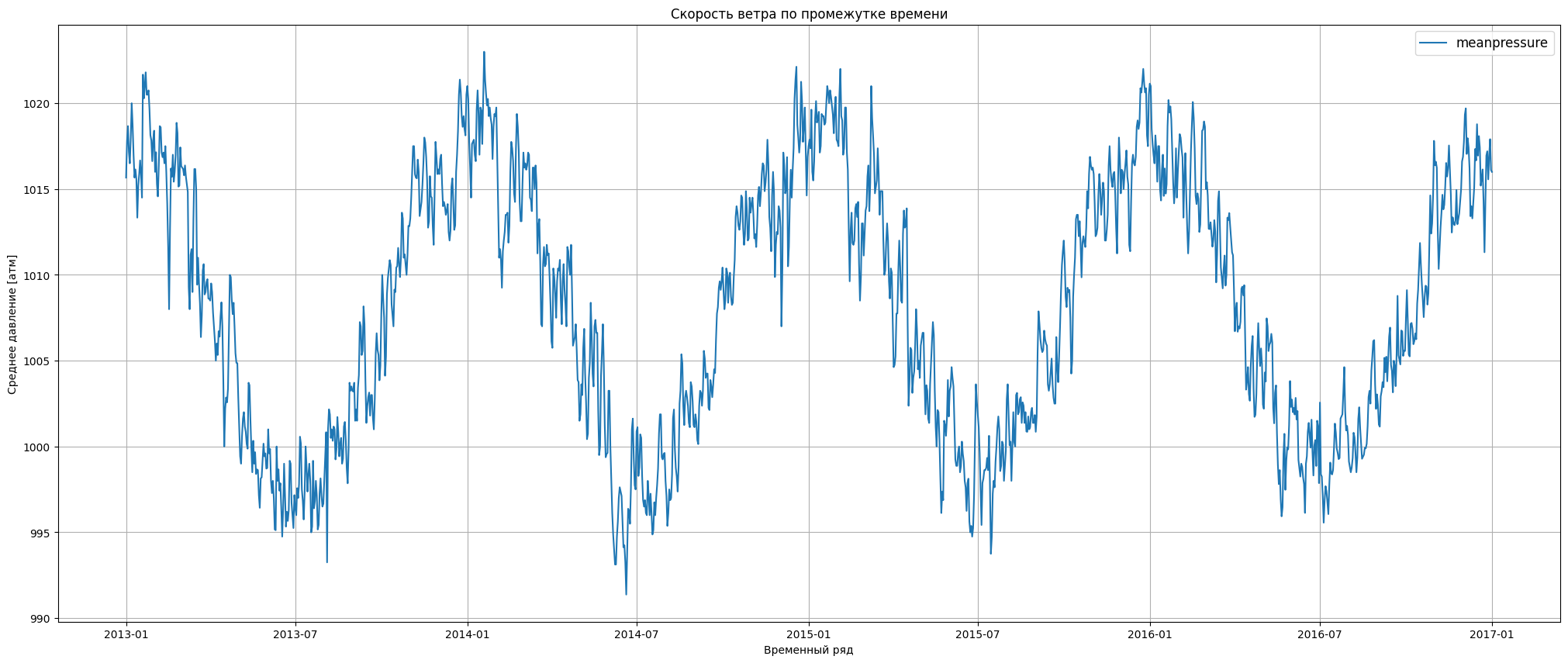


Скорость ветра не сильно привязан к вышеперичсленному признаку, она сама по себе, но иногда встречается всплеск огромной скорости

* + 1. **Среднее давление**

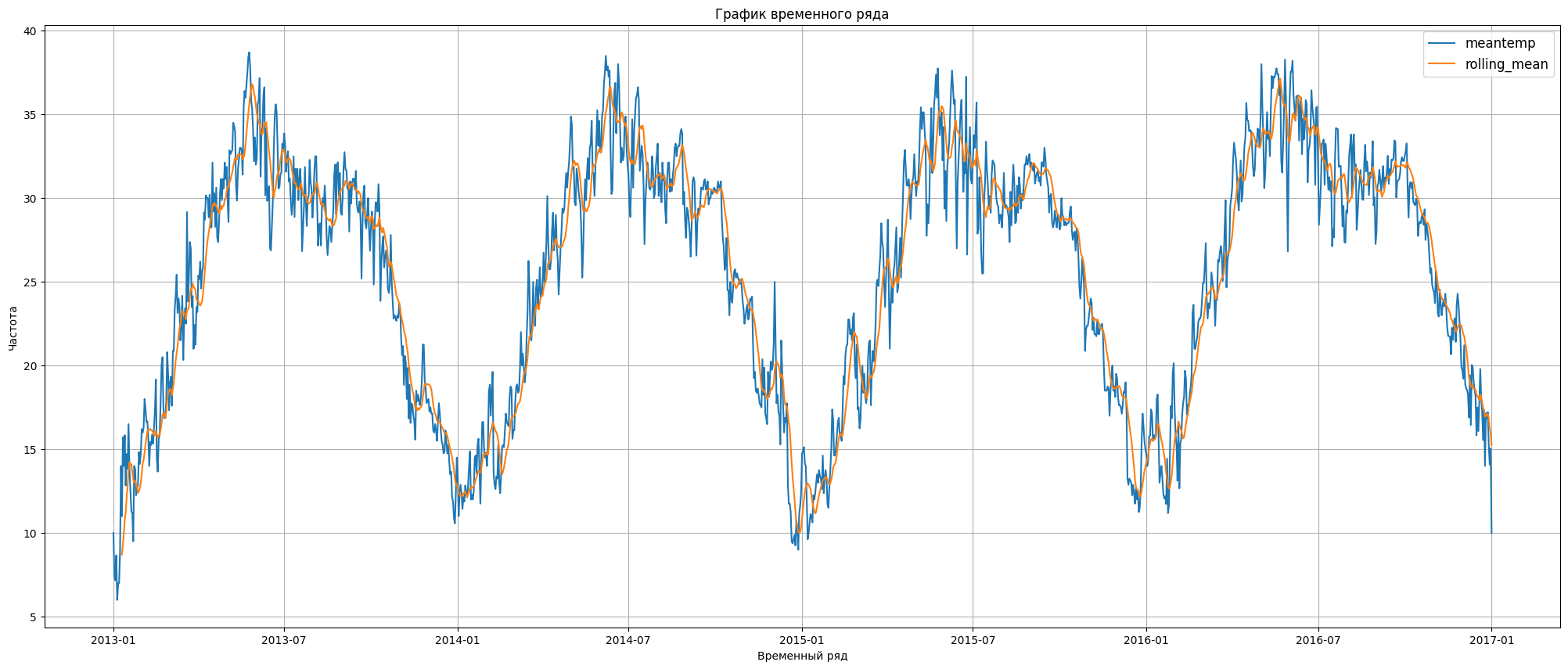
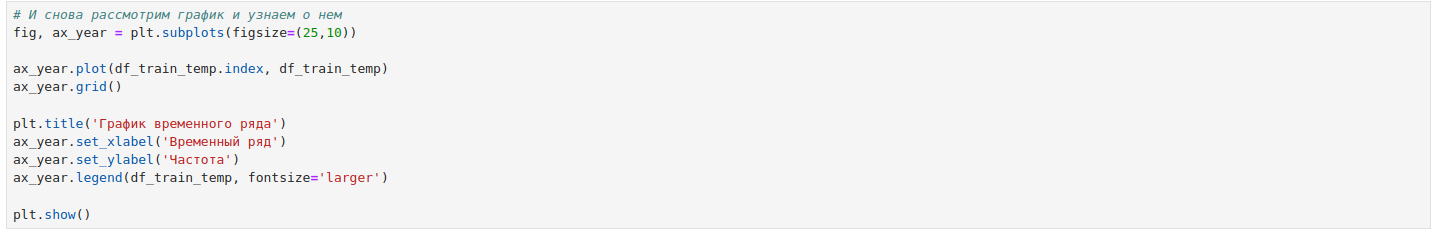


Виден огромный выброс давления, устраняем его, поставим на нормальное ограничение - от 980 по 1100

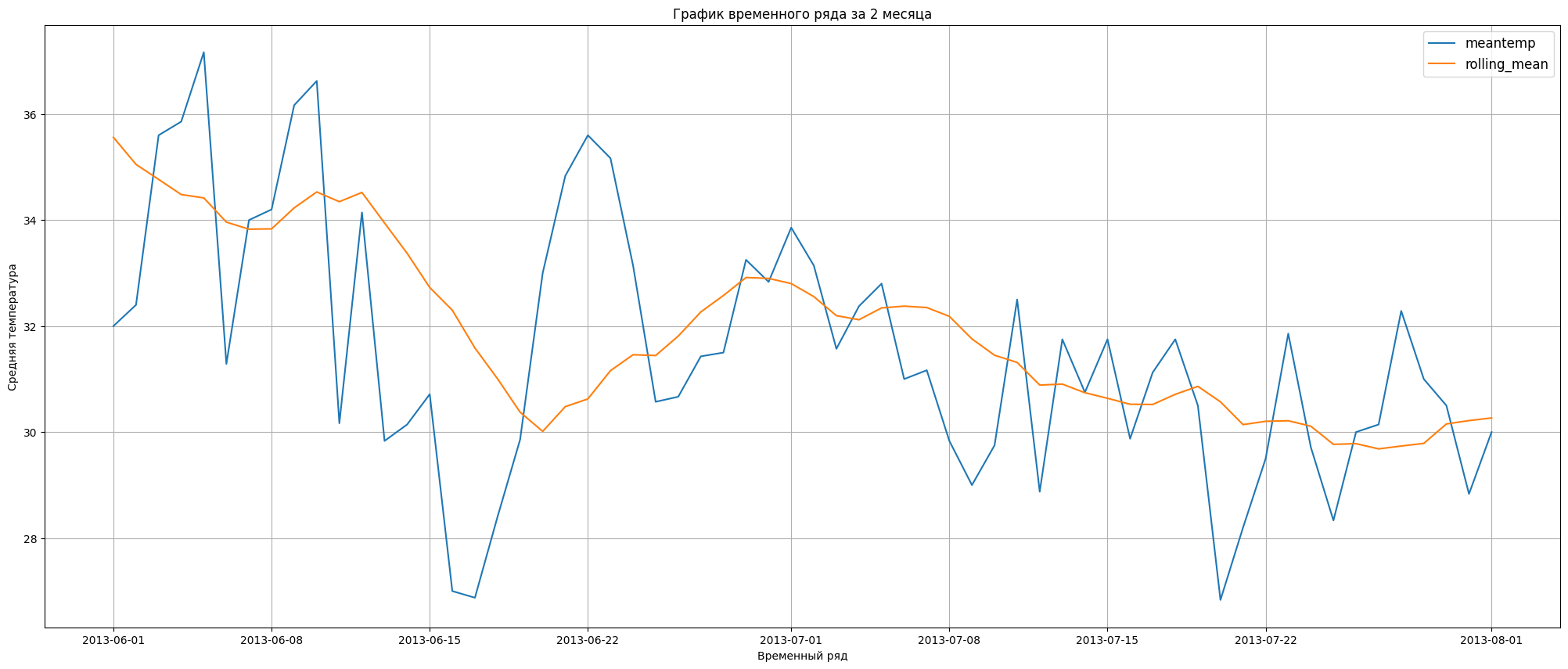


Наблюдается корреляция с уровнем влажности - чем выше влажность, тем выше давление.

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

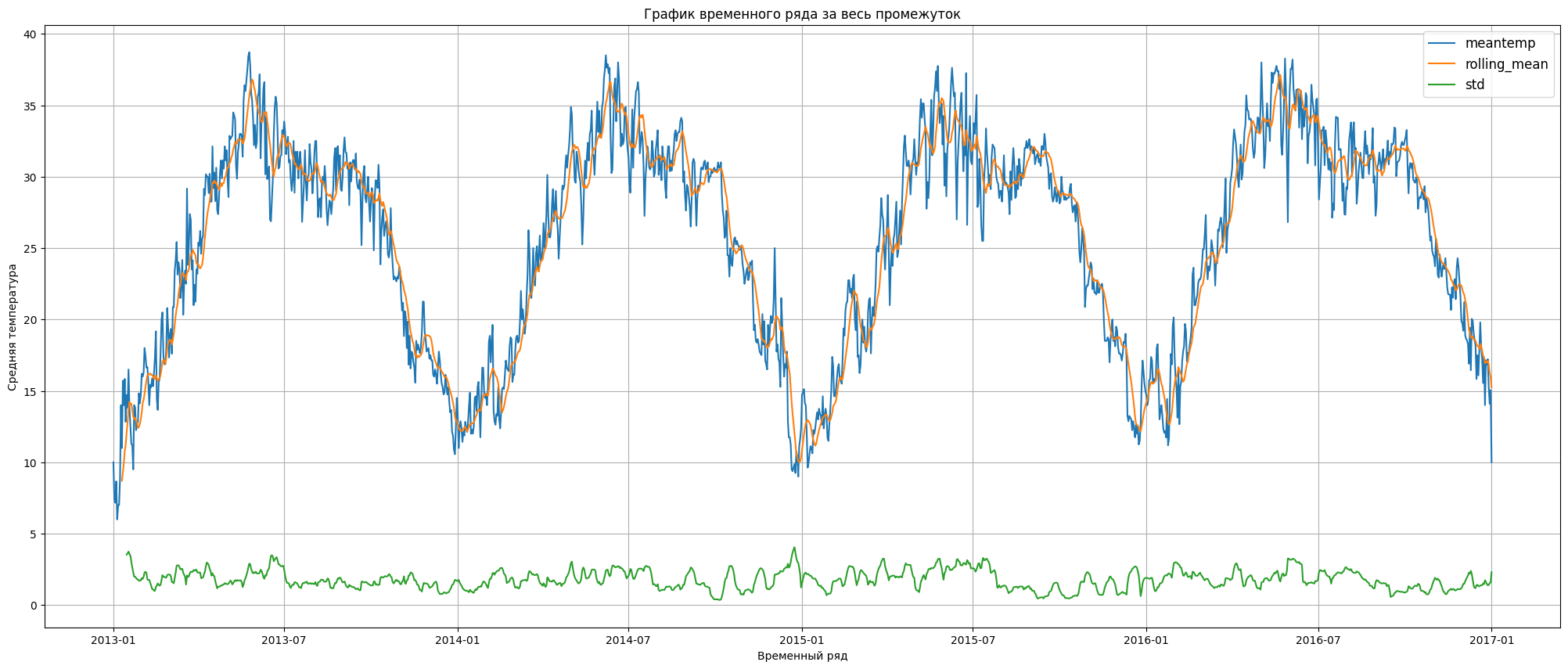
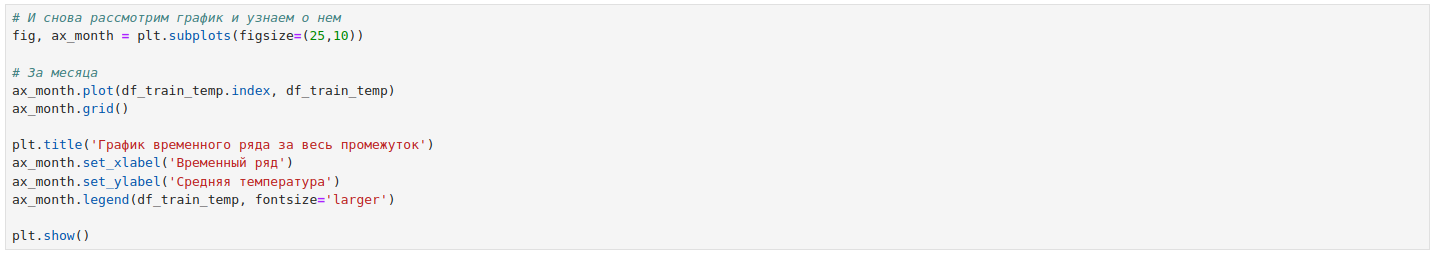
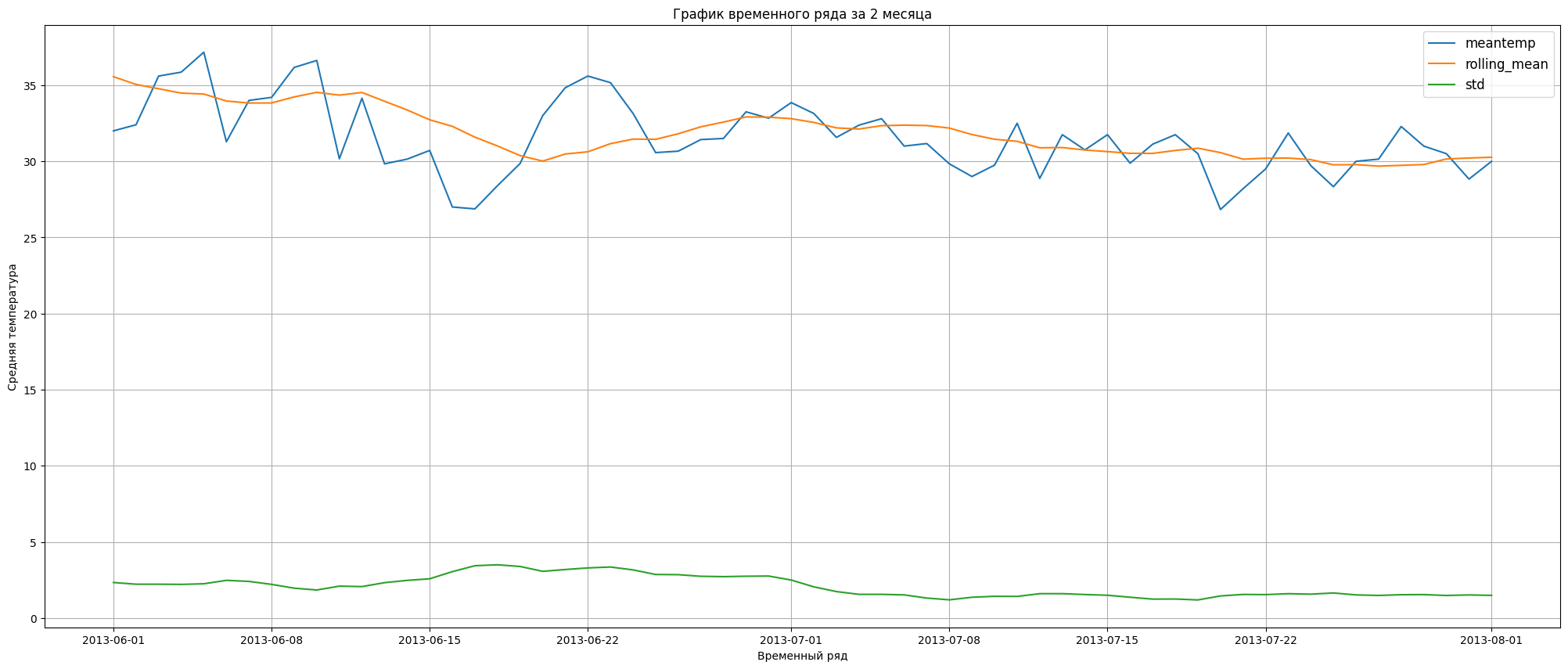


Здесь получилось неудобной и неинформативной, рассмотрим определенный промежуток времени



Видно, что идет спад температуры к началу осени

* + 1. **Исследование стационарных рядов**



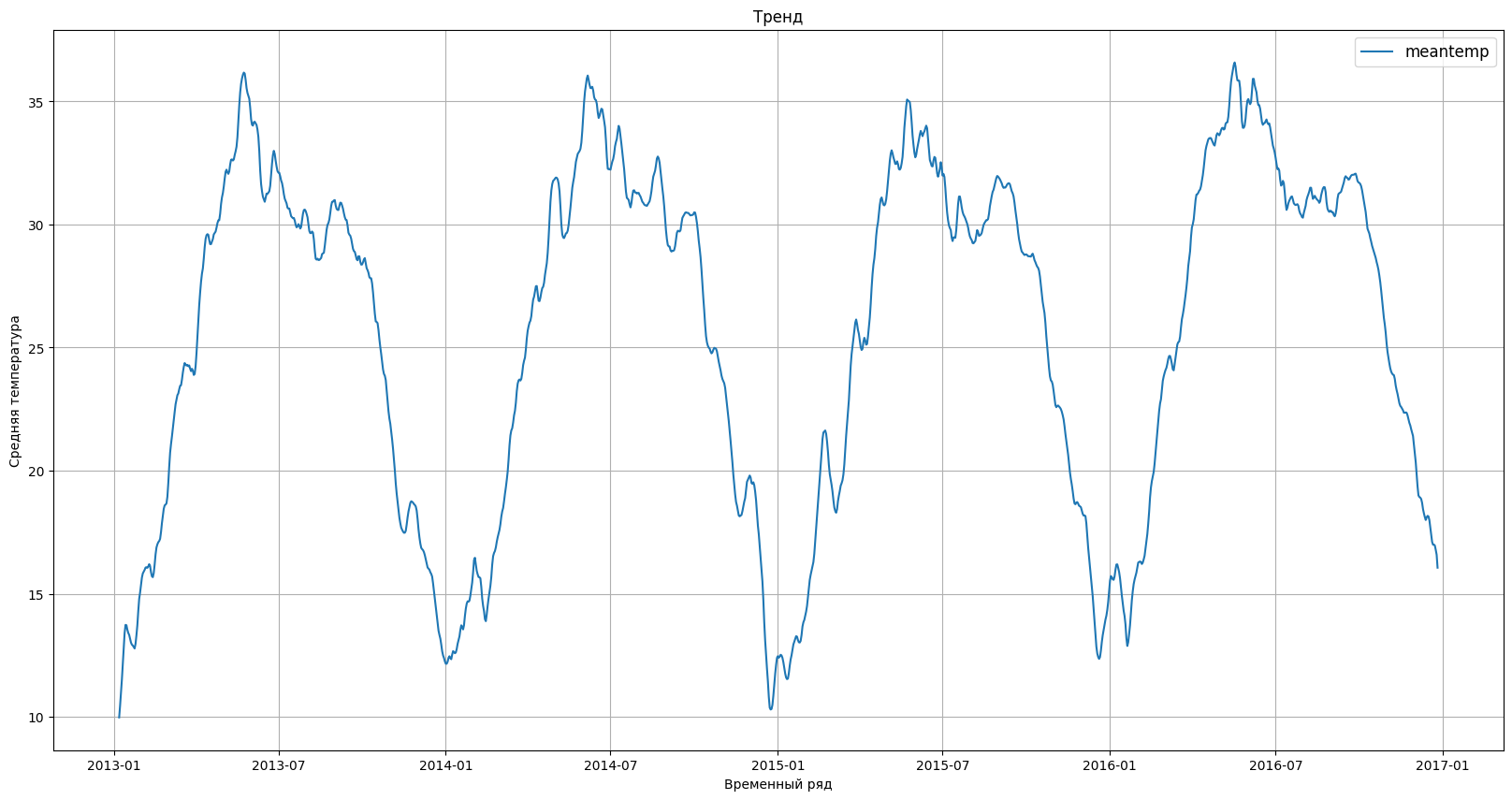
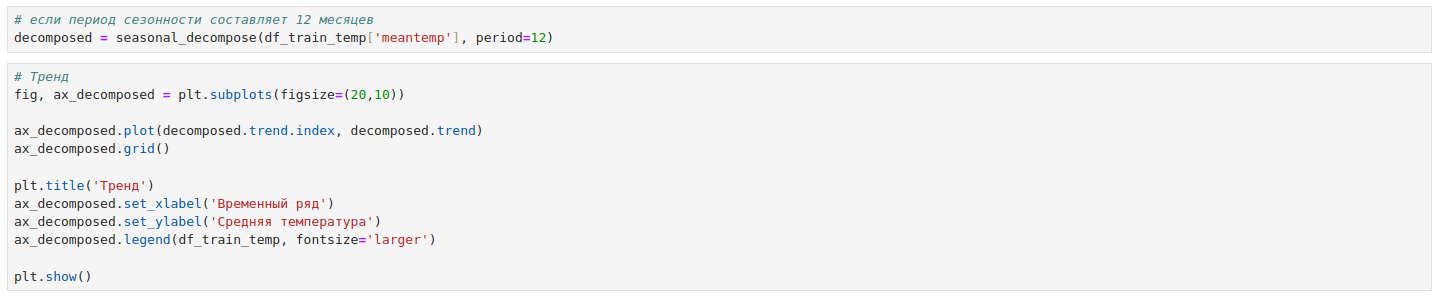
Примечание:

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

А что насчет временный ряд, его можно назвать стационарным или нет - нельзя, пока не проверим на тренды, сезонность и остаток декомпозиции

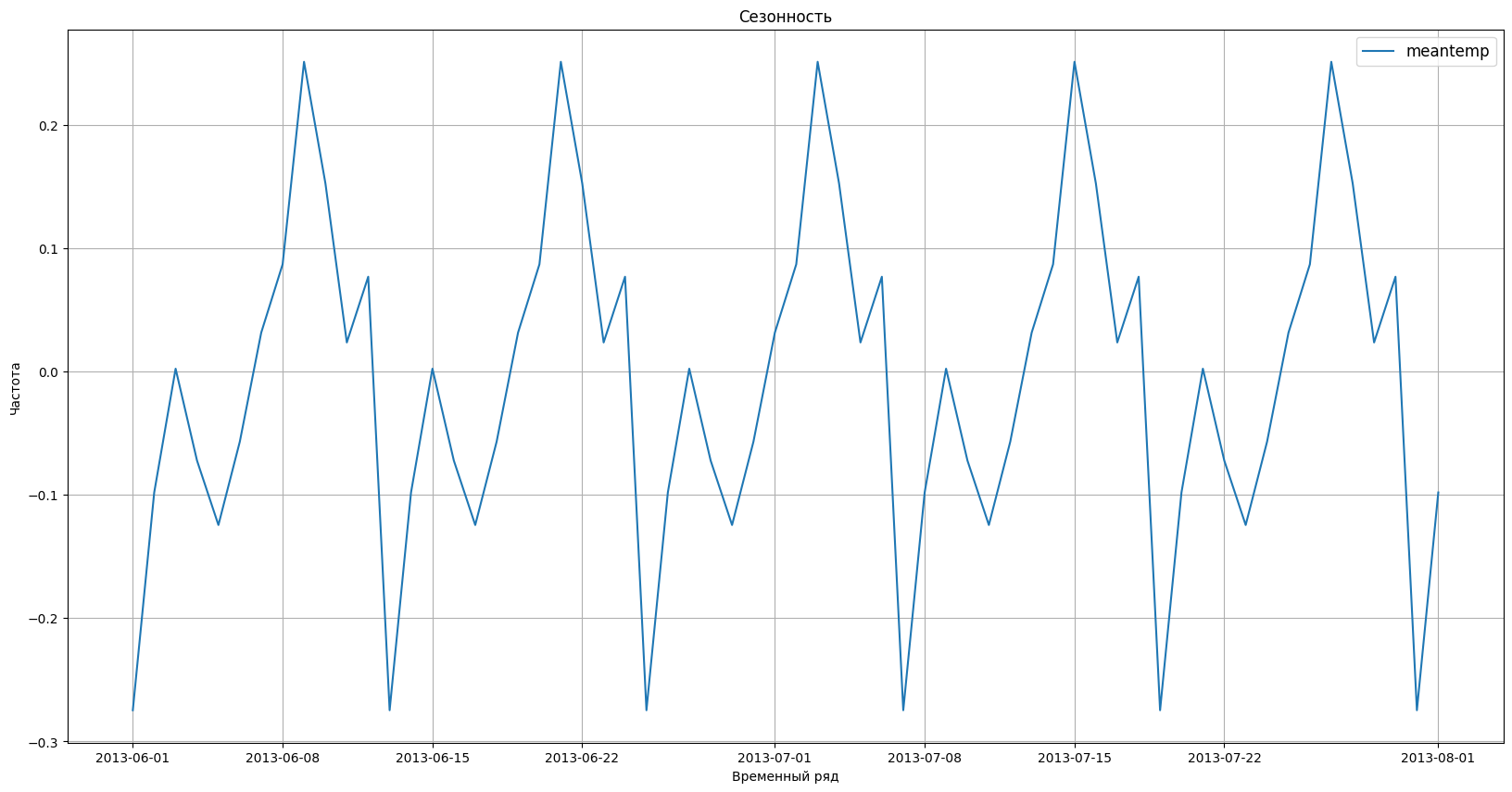
* + - 1. **Тренды и сезонность**
         1. **Тренды**



Видно, что тренд в последнее время увеличивается, т.е. с июня больше заказов на такси поступают нежели остальных

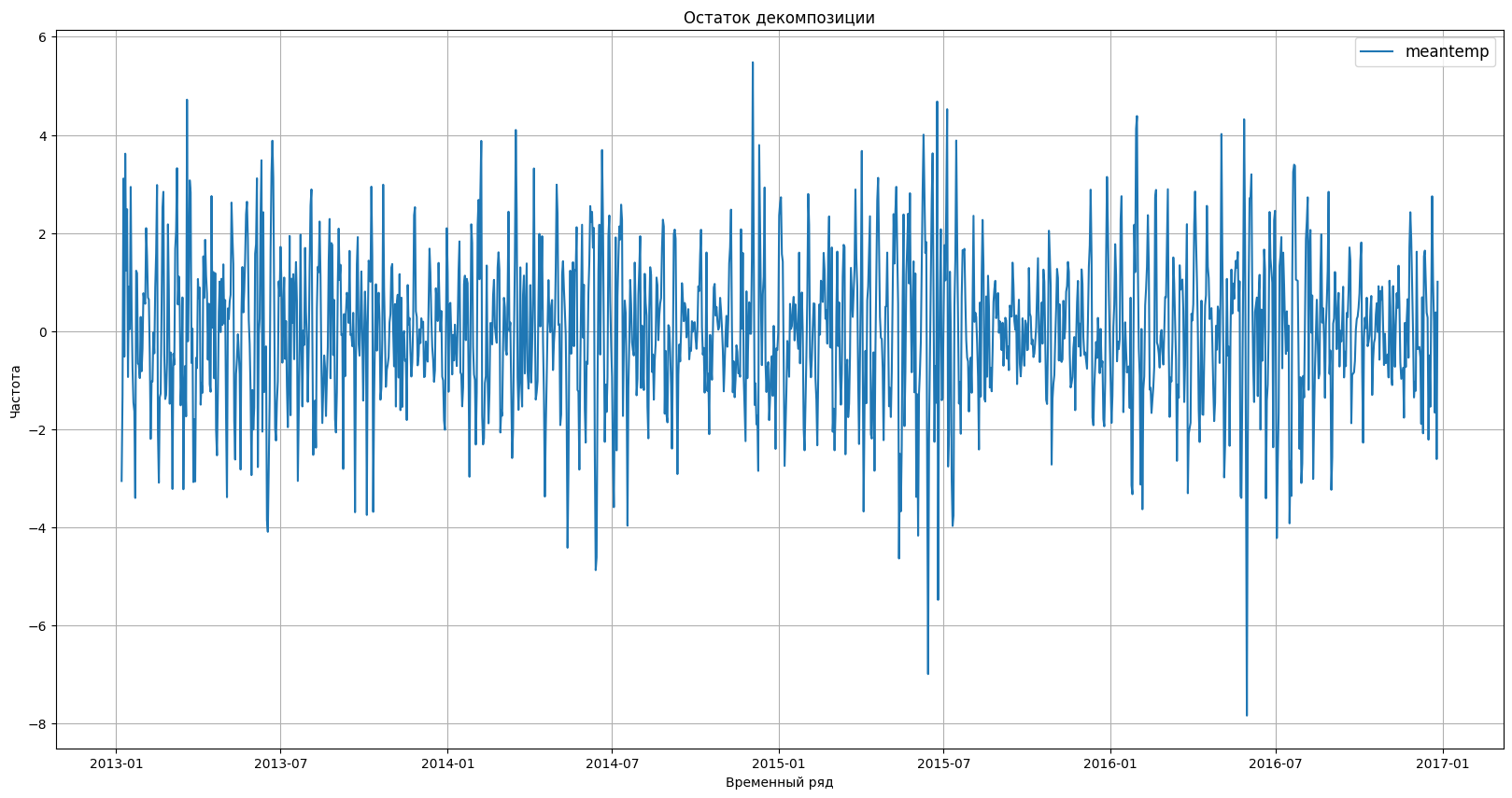
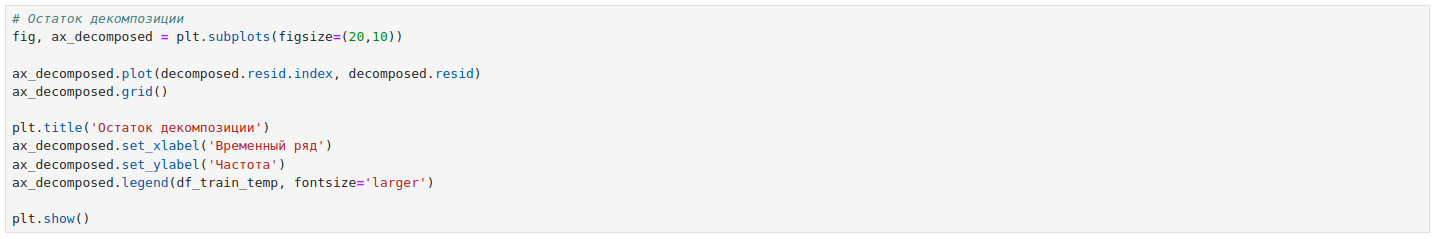
* + - * 1. **Сезонность**

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



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

* + - * 1. **Остаток декомпозиции**

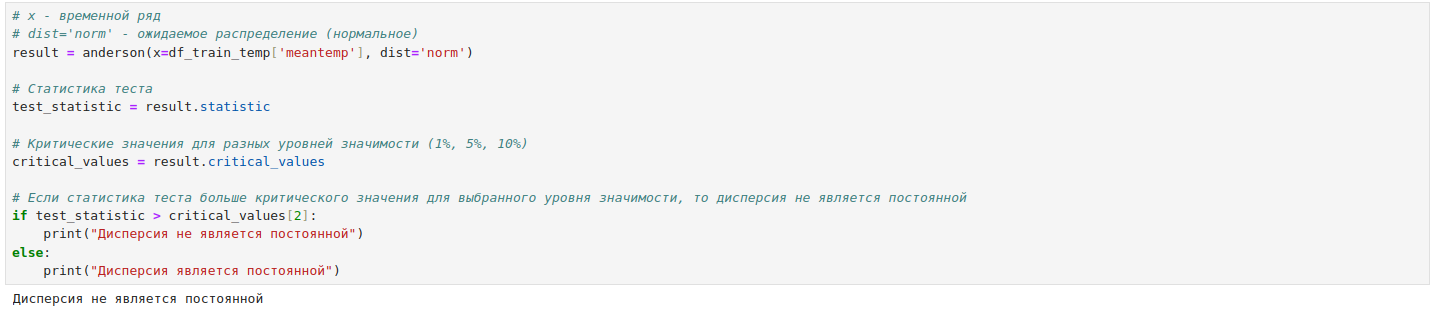


По поводу остатка декомпозиции нужно провести исследование отдельно, причем детально, поскольку декомпозиции остатков интрепетировать сложнее, чем сезонность и тренды - в идеале, график остатков должен содержать только шум без систематических компонентов.

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.

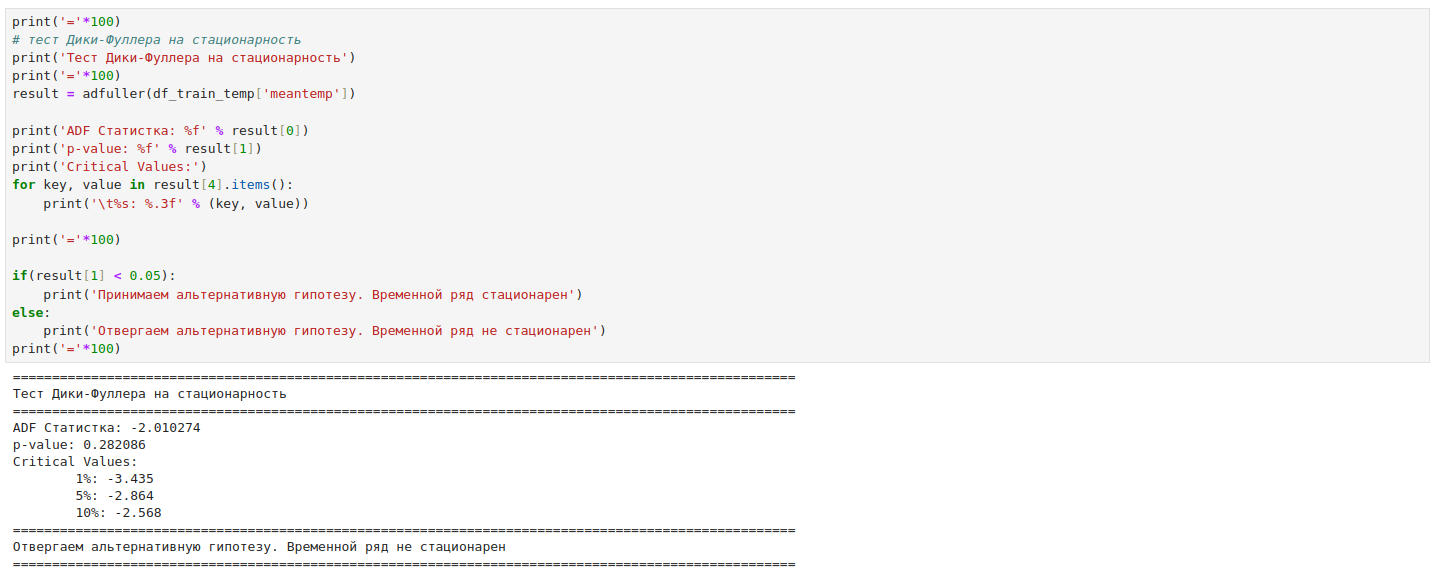
* + - 1. **Проверка на дисперсию математическим способом**

Можно проверить дисперсию, применяя теста Андерсона-Дарлинга



Здесь можно сказать, что этот промежуток времени, т.е. с 2013 до конца, является нестанционарным, т.к. присутствует тренд, который стремится увеличиваться. А также дисперсия не является постоянной. Проверим на станционарность математическим способом, что она скажет.

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



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



Два разные теста одно и то же говорит, а по графику глазами не сможем судить, что является стационарен или не стационарен. Даже если приглядеться, то можно увидеть небольшой рост, хоть капельку.

* + 1. **Промежуточный вывод**

По графику видно, что, как обычно, перед началом сезоном осени, начинается спад температуры, а уровень влажности повышается.

А также не стоит забывать, что заметен тренд и на математическом языке были доказаны следующие тесты:

На дисперсию и 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

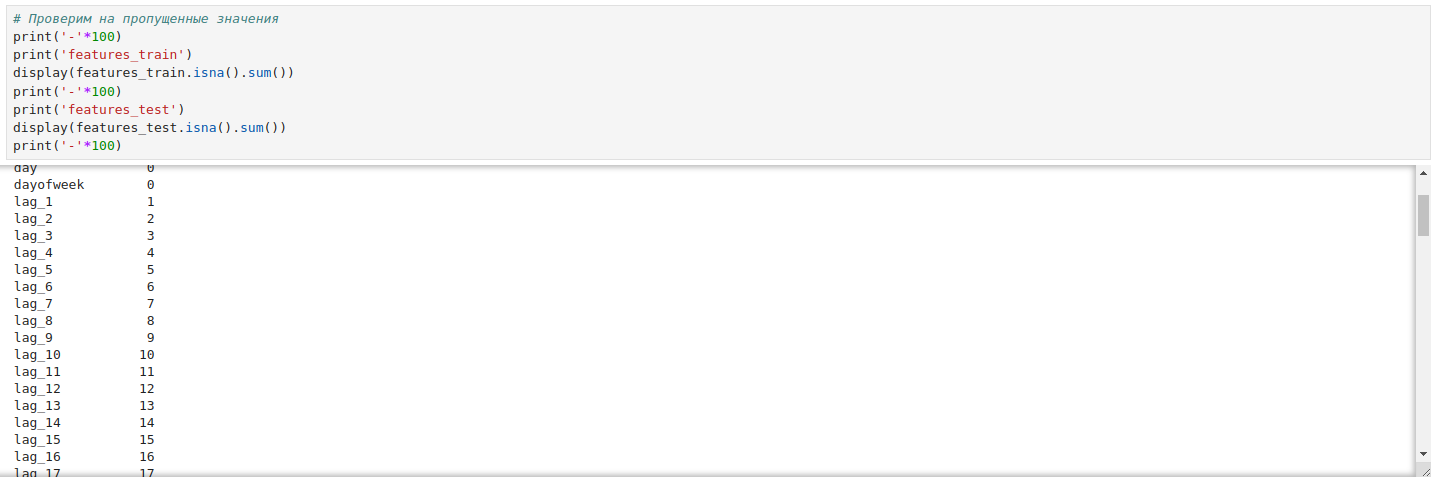
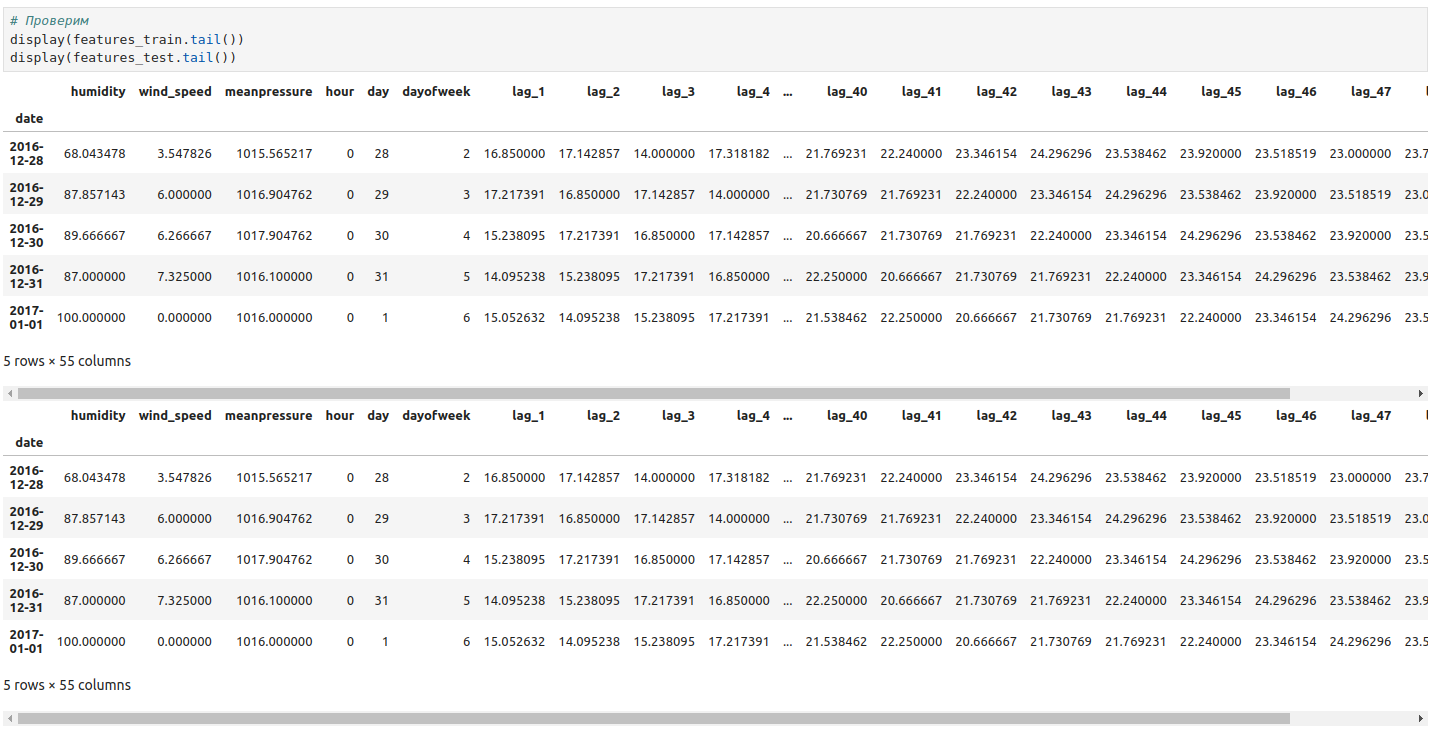
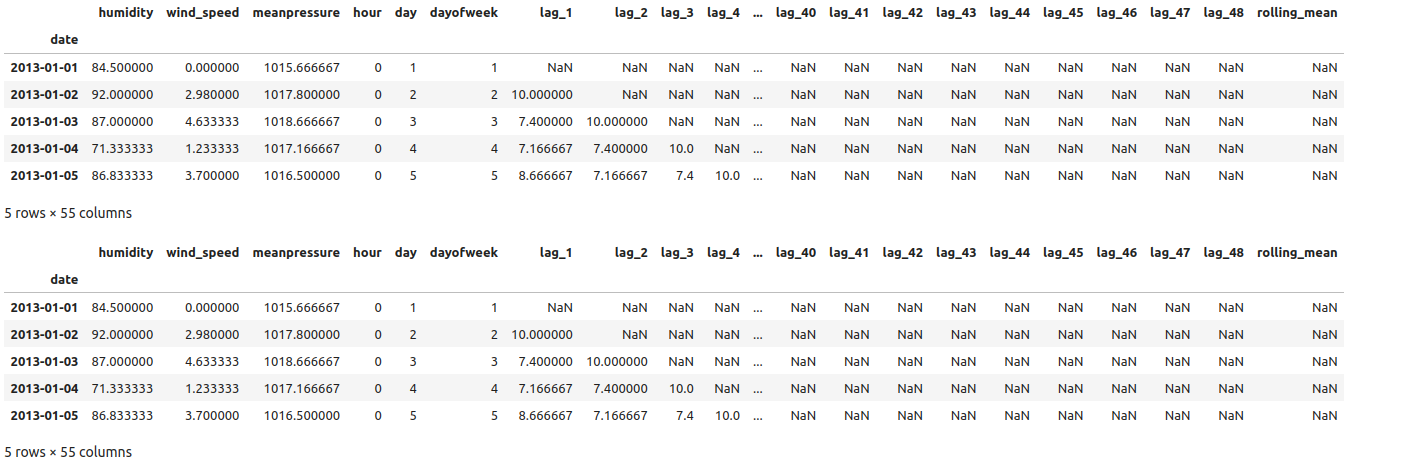
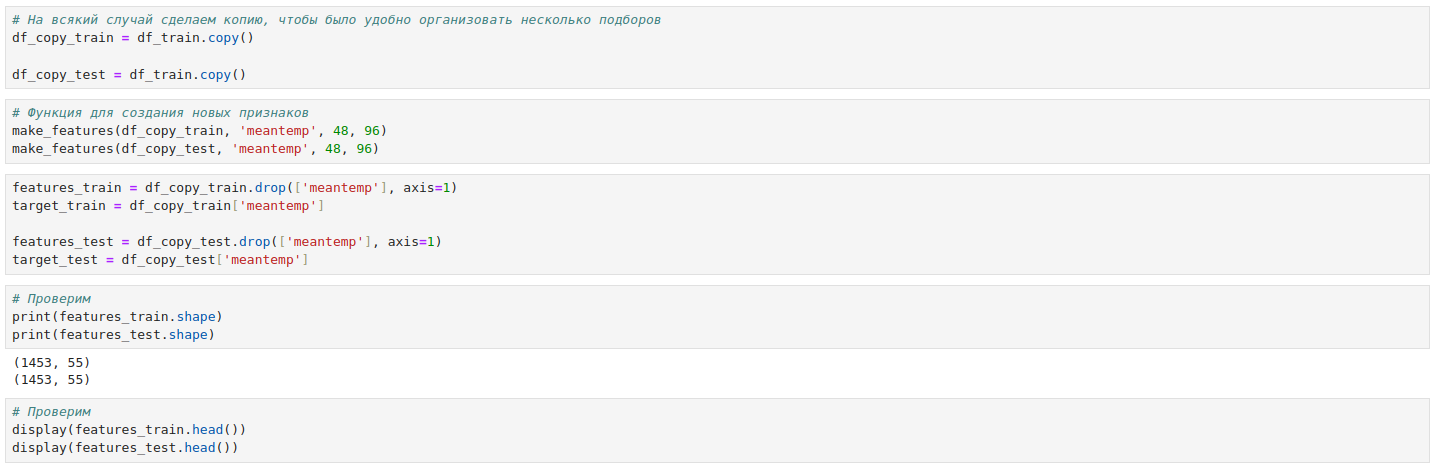
А тесты на стационарность дали одинаковые результаты - отсутствует стационарность, значит, что распределение со временем меняется.

* 1. **Обучение и тестирование**



* + 1. **Деление на обучающие и тестовые выборки**

У нас есть 2 датасет: обучающий и тестовый

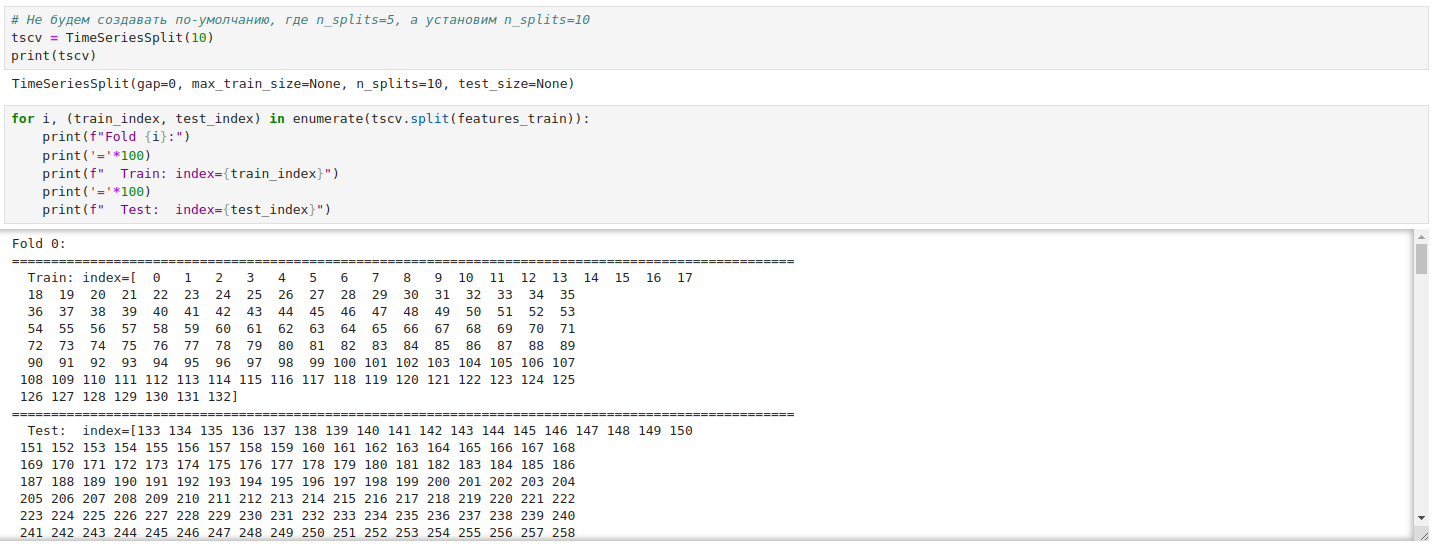


Есть пропущенных значений, заполним их значением 0



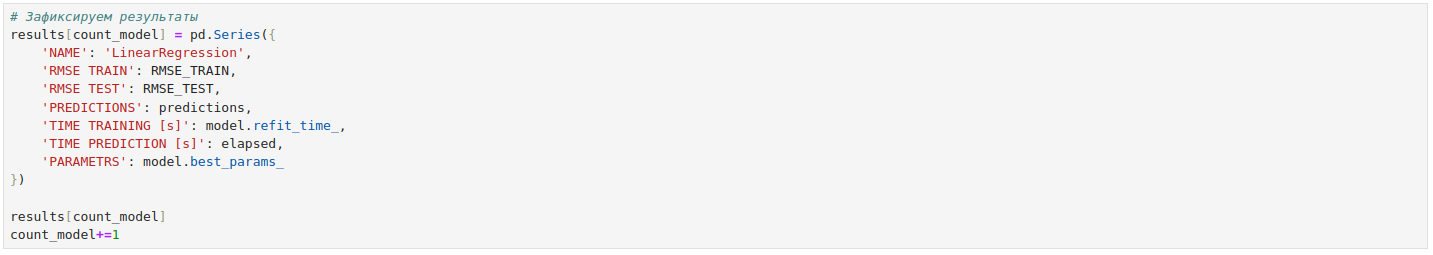
Обучение моделей с применением TimeSeriesSplit

Почему обучаем с применением TimeSeriesSplit, потому что мы обучаем модель, у которой по индексу стоит временный ряд. А чтобы они не пересекались, это поможет инструмент TimeSeriesSplit. Как работает, рассмотрим ниже.

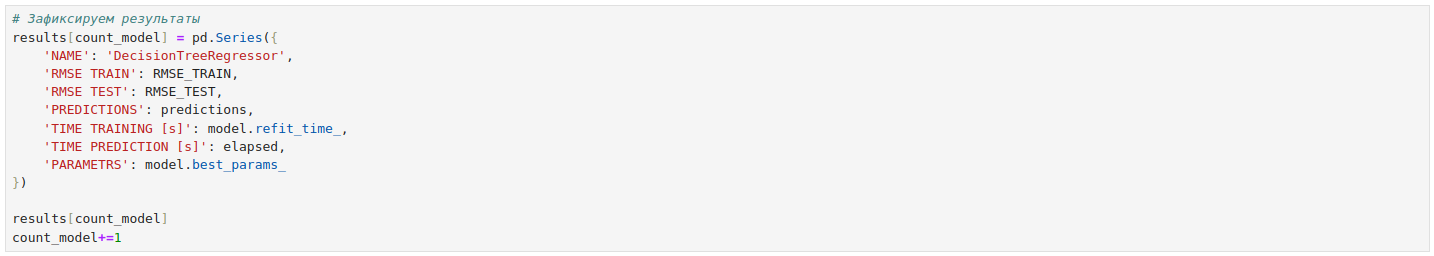


Как и видим, что разделен на несколько промежутков времени, по 10 блоков времени

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



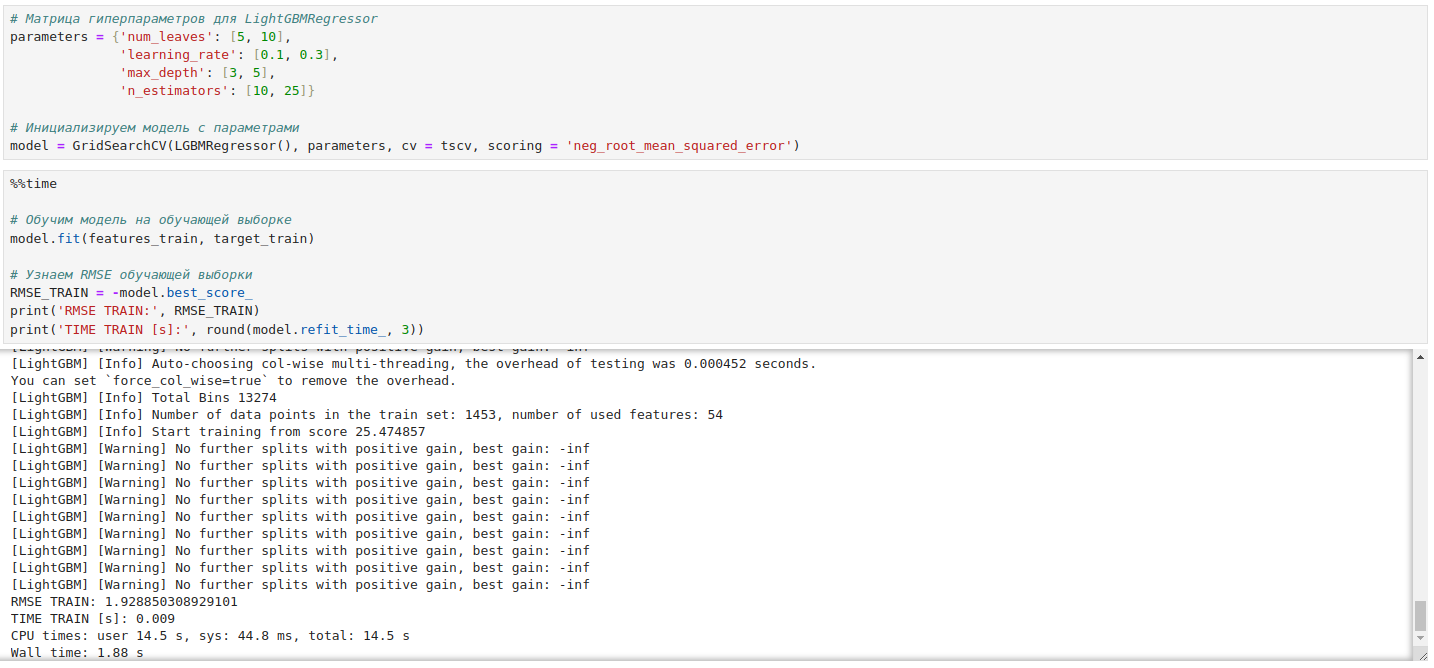
* + - 1. **DecisionTreeRegressor**



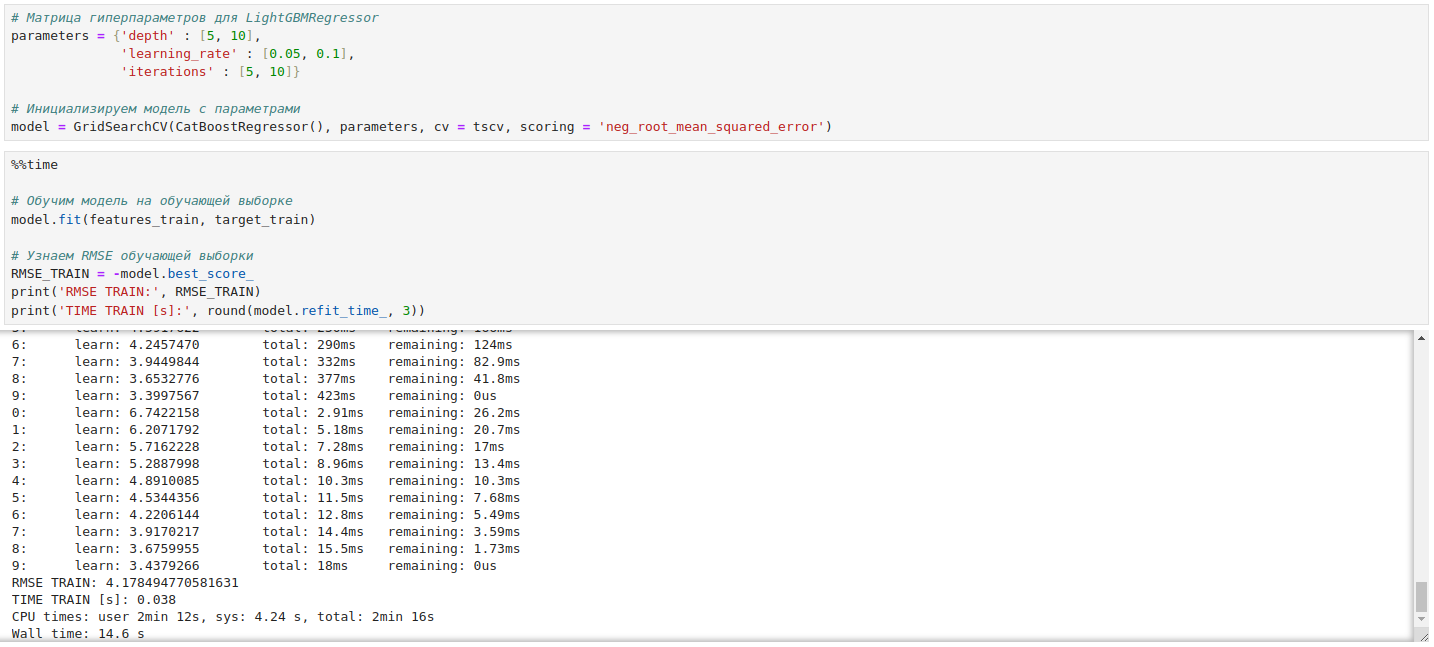
* + - 1. **RandomForestRegressor**



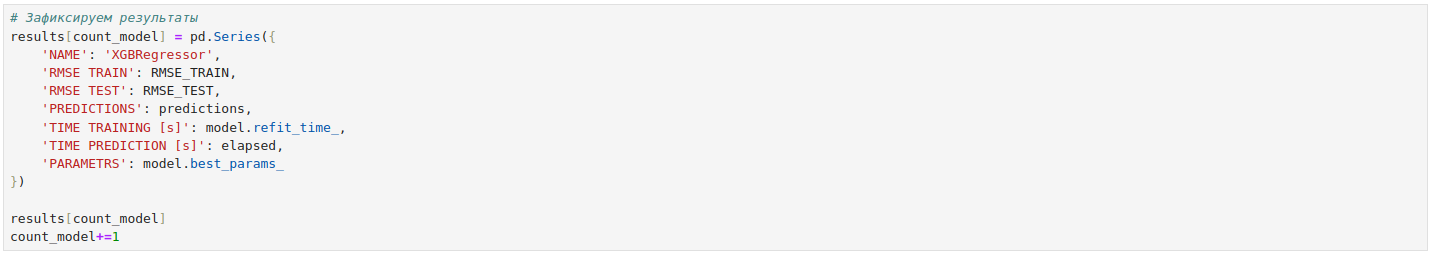
* + - 1. **LightGBMRegressor**



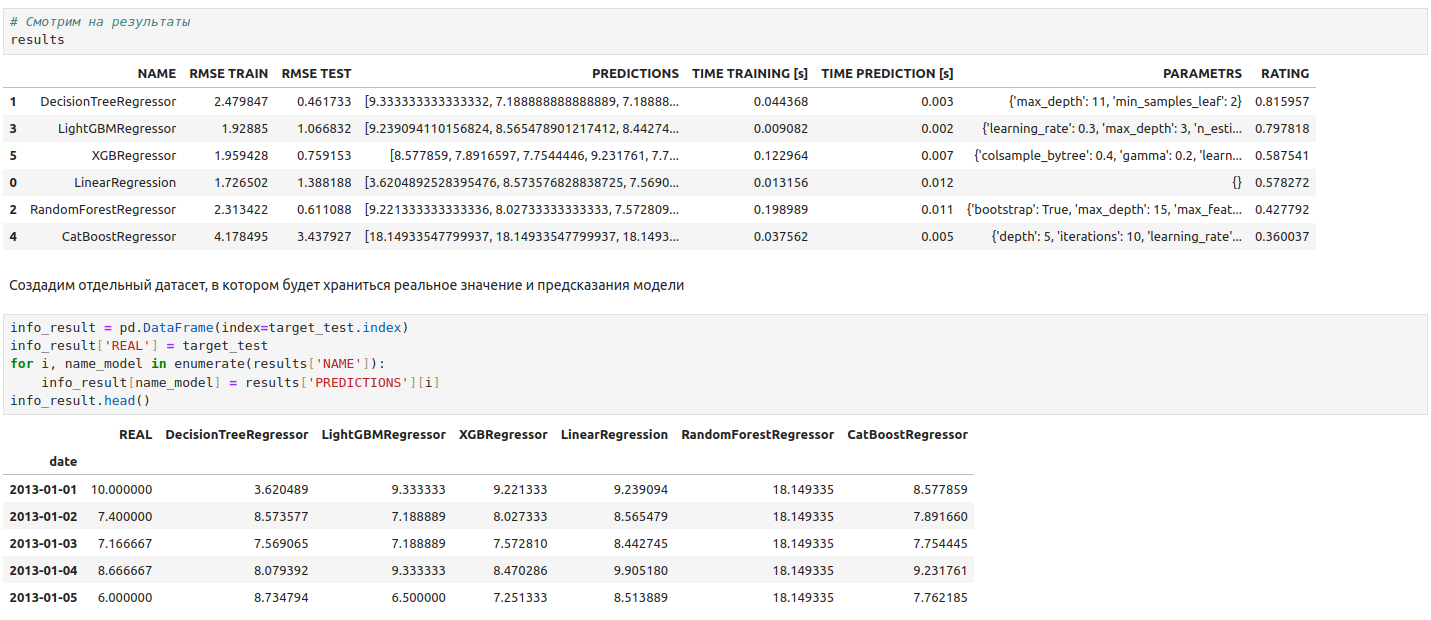
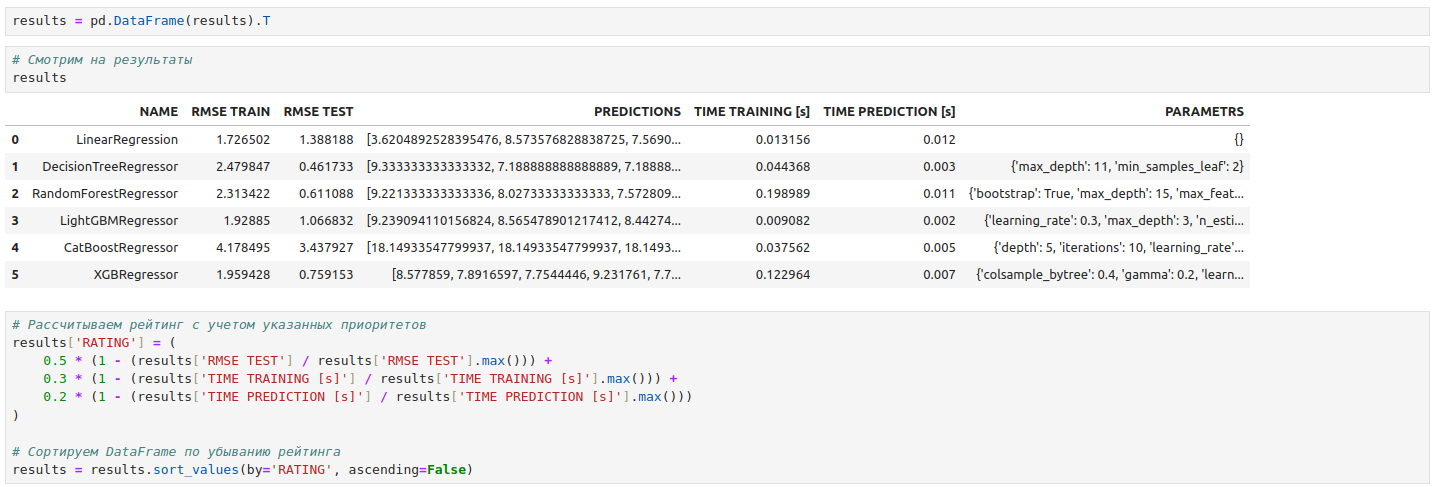
* + - 1. **CatBoostRegressor**



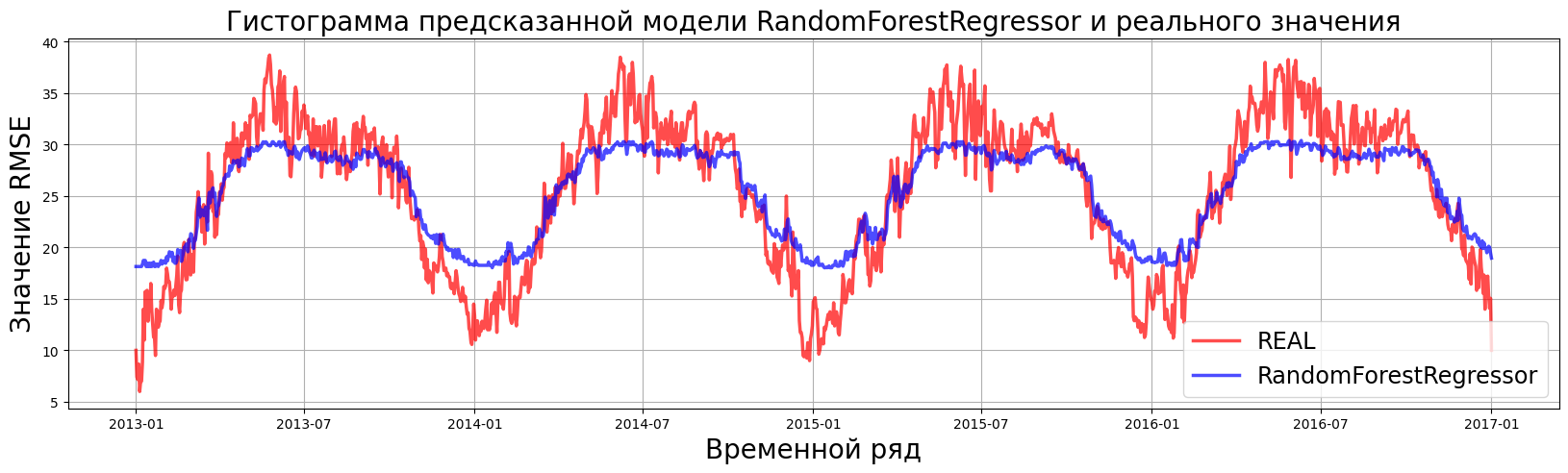
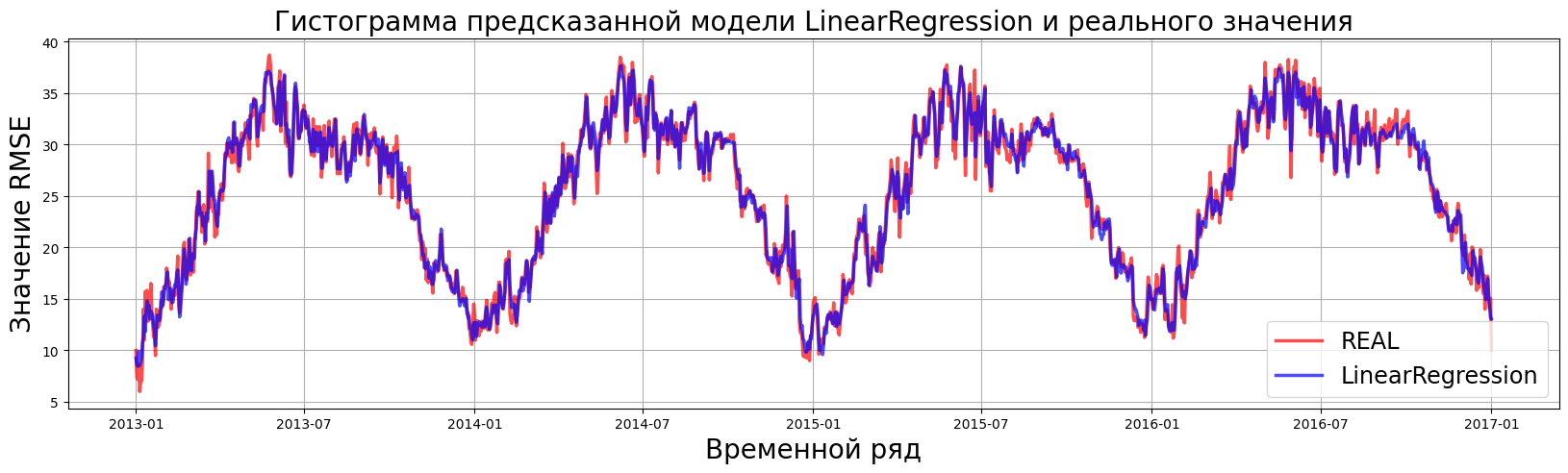
* + - 1. **XGBRegressor**



* + 1. **Результаты**



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



По графику видно, что предсказания модели в некотором степени отличаются от реальных значений

* 1. **Вывод**

Самой лучшей моделей является DecisionTreeRegressor - RMSE составляет 0.461733, а скорость обучения составляет почти 0.045 секунда, время предсказания - 0.003 секунда. Вполне очень акдеватный результат, чтобы выбрать эту модель для реализации.

Есть другая модель - LightGBMRegressor, результаты очень неплохие, идеальный кандидат, чтобы конкурироваться с XGBRegressor

А четвертое место заслужила простая модель - LinearRegression, однако эта модель может похвастаться скоростью обучения и предсказанией