Департамент образования и науки города Москвы

Государственное автономное образовательное учреждение высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики, управления и технологий

ДИСЦИПЛИНА:

Инструменты для хранения и обработки больших данных

**Лабораторная работа №5.**

**Тема:**

**«**Обработка данных с использованием Apache Spark**»**

Выполнил(а): Сачкова Г.Г, группа: АДЭУ-211

Преподаватель: Босенко Т.М.

Москва

2024

**Цель работы:**

* Познакомиться с понятием «большие данные» и способами их обработки;
* Познакомиться с инструментом Apache Spark и возможностями, которые он предоставляет для обработки больших данных.
* Получить навыки выполнения разведочного анализа данных использованием pyspark.

**Оборудование и программное обеспечение:**

* Компьютер с операционной системой Windows.
* Docker compose
* Python 3.10

**Теоретическая часть**

Apache Spark – это распределенный фреймворк обработки данных, предоставляющий высокую производительность и масштабируемость для анализа больших объемов данных в реальном времени.

Обработка данных с использованием Apache Spark является ключевым аспектом современного анализа данных в больших масштабах. Apache Spark представляет собой мощный инструмент для обработки и анализа больших объемов данных параллельно и эффективно. Вот некоторые ключевые аспекты теоретической части работы на эту тему:

Ключевые компоненты Apache Spark:

* Resilient Distributed Dataset (RDD): Основная абстракция данных в Apache Spark, обеспечивающая устойчивое распределенное хранилище данных для обработки.
* Structured Data Processing: Apache Spark SQL обеспечивает поддержку структурированных данных с использованием DataFrame API для выполнения SQL-подобных запросов.
* Machine Learning Library (MLlib): Библиотека машинного обучения для реализации алгоритмов машинного обучения на больших объемах данных.
* Streaming Data Processing: Apache Spark Streaming позволяет обрабатывать потоковые данные в реальном времени.

**Практическая часть**

В первую очередь был запущен Docker Desktop. На рисунке 1 видно, что мы добились его работоспособности.

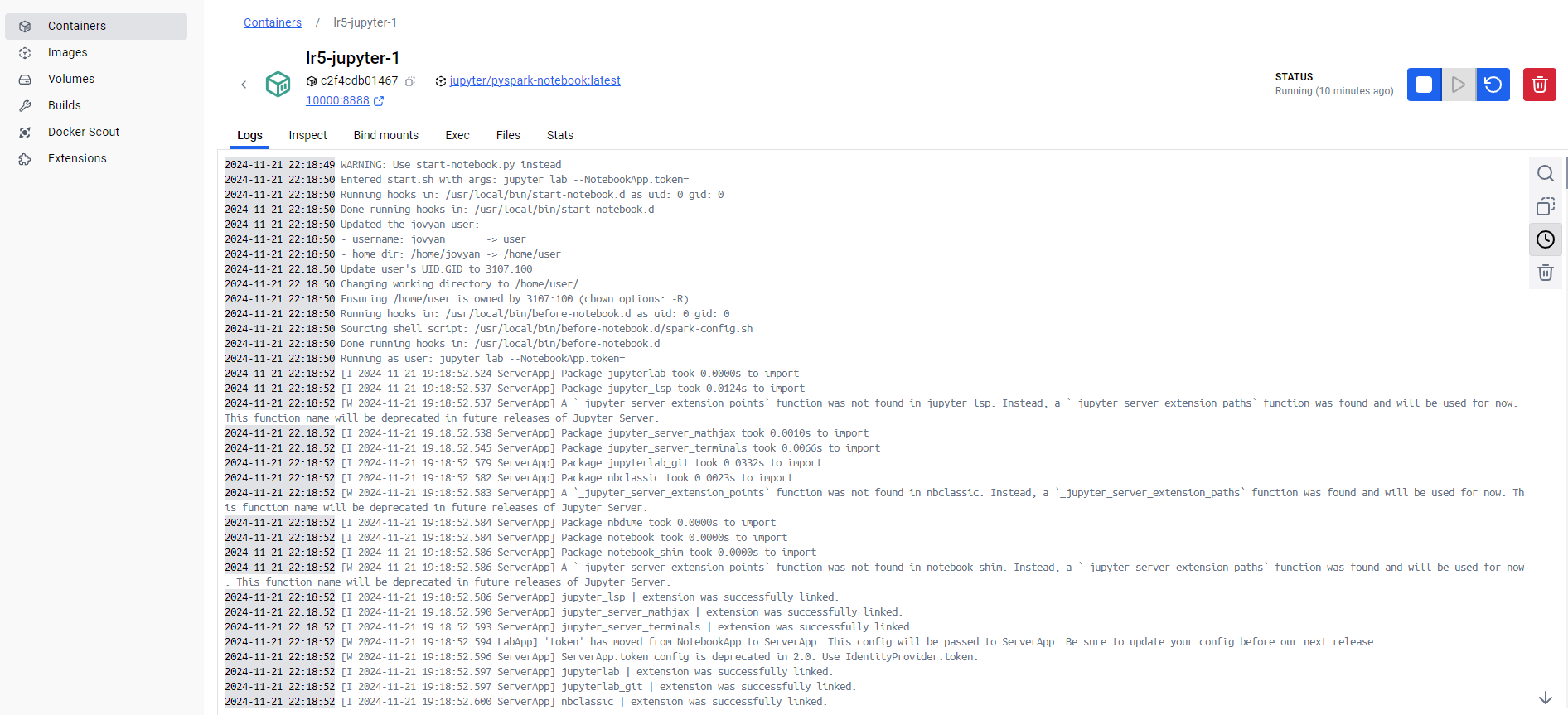


Рисунок 1 – Docker Desktop

Далее был скачен датасет, расположенный по адресу: <https://drive.google.com/file/d/1yiAp1fFDy3wSqUR0X_btCZPtuczbLwCe/view?usp=drive_link>. Файл был распакован и помещен в директорию “data”. Прежде чем приступить к выполнению самостоятельной работы, был запущен и изучен файл с готовым кодом. Мы ознакомились с различными функциями и результатами работы. Один из выполненных кодов с выводом представлен на рисунке 2.

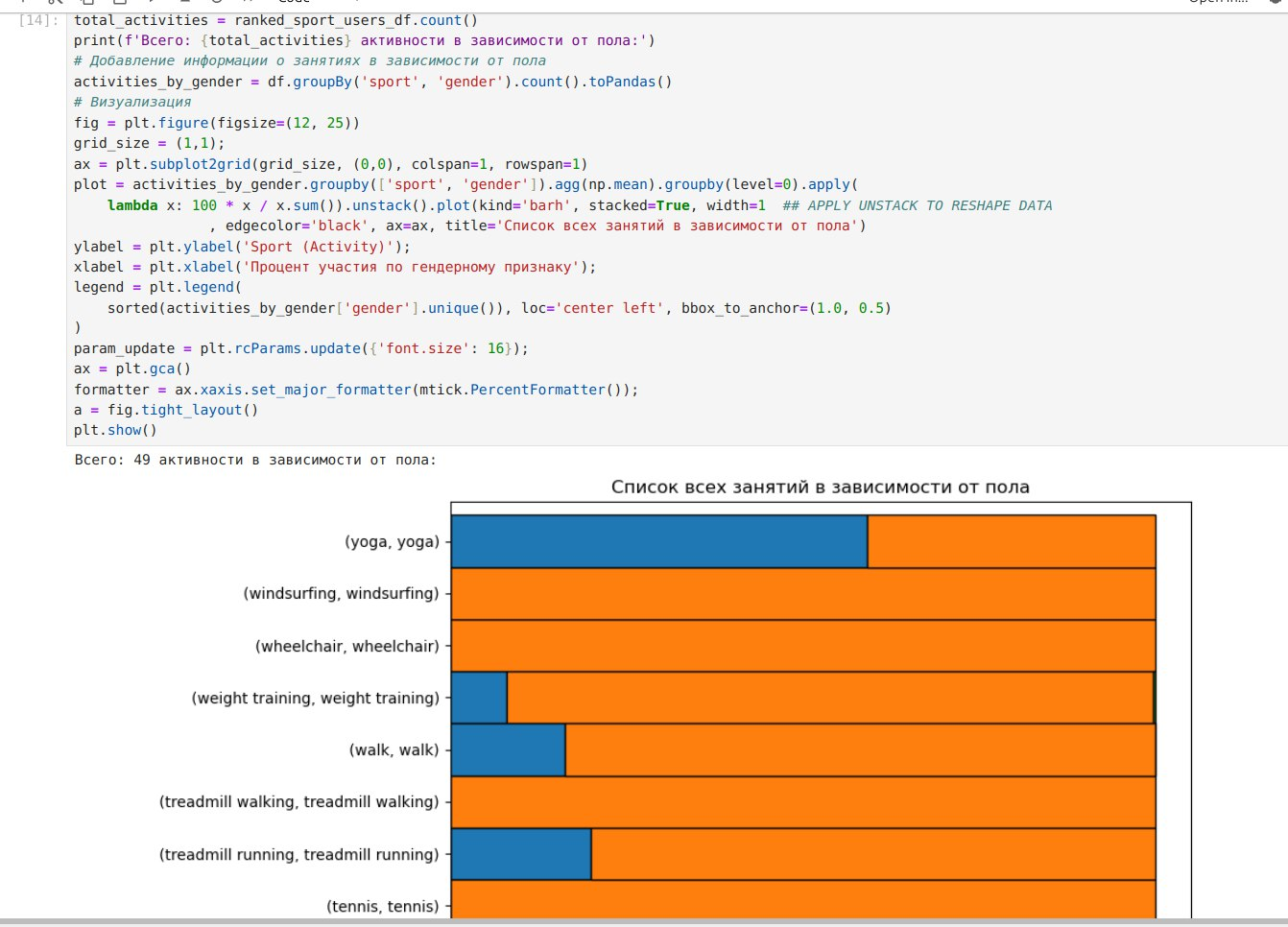


Рисунок 2 – Ознакомительные данные

Следующим этапом мы переходим к выполнению самостоятельной работы. Для этого был найден подходящий датасет на сайте Kaggle. Были выбраны метеорологические данные Бразилии, которые были получены при помощи автоматических станциях. Вес файла составляет более 6 Гб. Ознакомиться с данными более подробно можно по ссылке:

<https://www.kaggle.com/datasets/saraivaufc/automatic-weather-stations-brazil?select=automatic_weather_stations_inmet_brazil_2000_2021.csv>

Датасет был загружен, итог загрузки можно посмотреть на рисунке 3. Для этого были выведены колонки данного датасета.

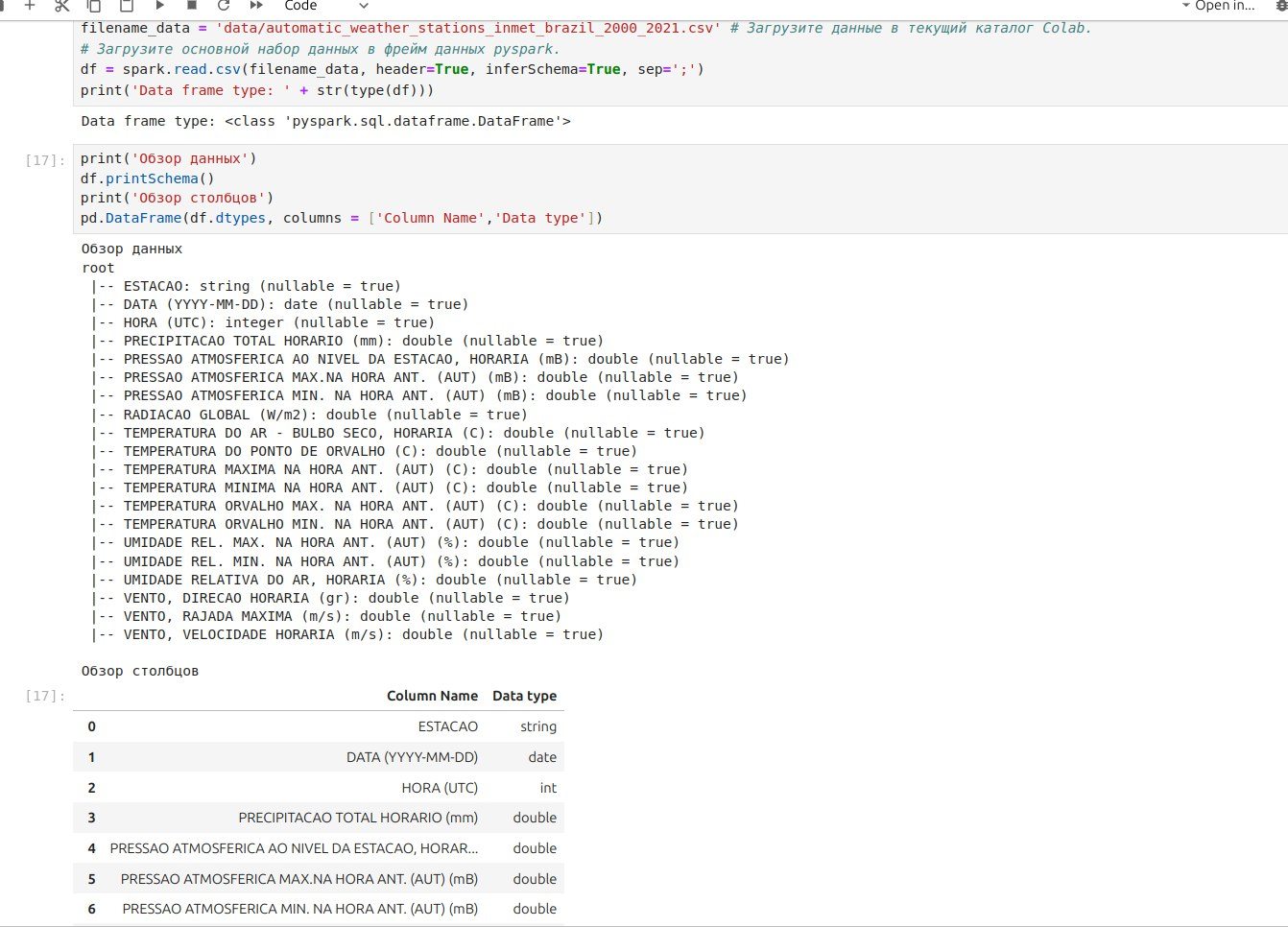


Рисунок 3 – Данные загруженного датасета.

Для удобства и сбережения времени, которое затрачивается на обработку, были оставлены только выборочные колонки. Приведем их описание: 1) ESTACAO – Код станции, фиксировавшей данные 2) DATA (YYYY-MM-DD) – Дата фиксации показателей 3) HORA (UTC) – Час, в которой фиксировались показатели 4) PRECIPITACAO TOTAL HORARIO (mm) – Количество осадков, выпавших за час 5) PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA (mB) – Атмосферное давление на уровне станции 6) TEMPERATURA DO AR - BULBO SECO, HORARIA (C) – Температура воздуха. Описанные данные представлены на рисунке 4. Также на рисунке представлен, отработавший код для удаления пропусков в данных.

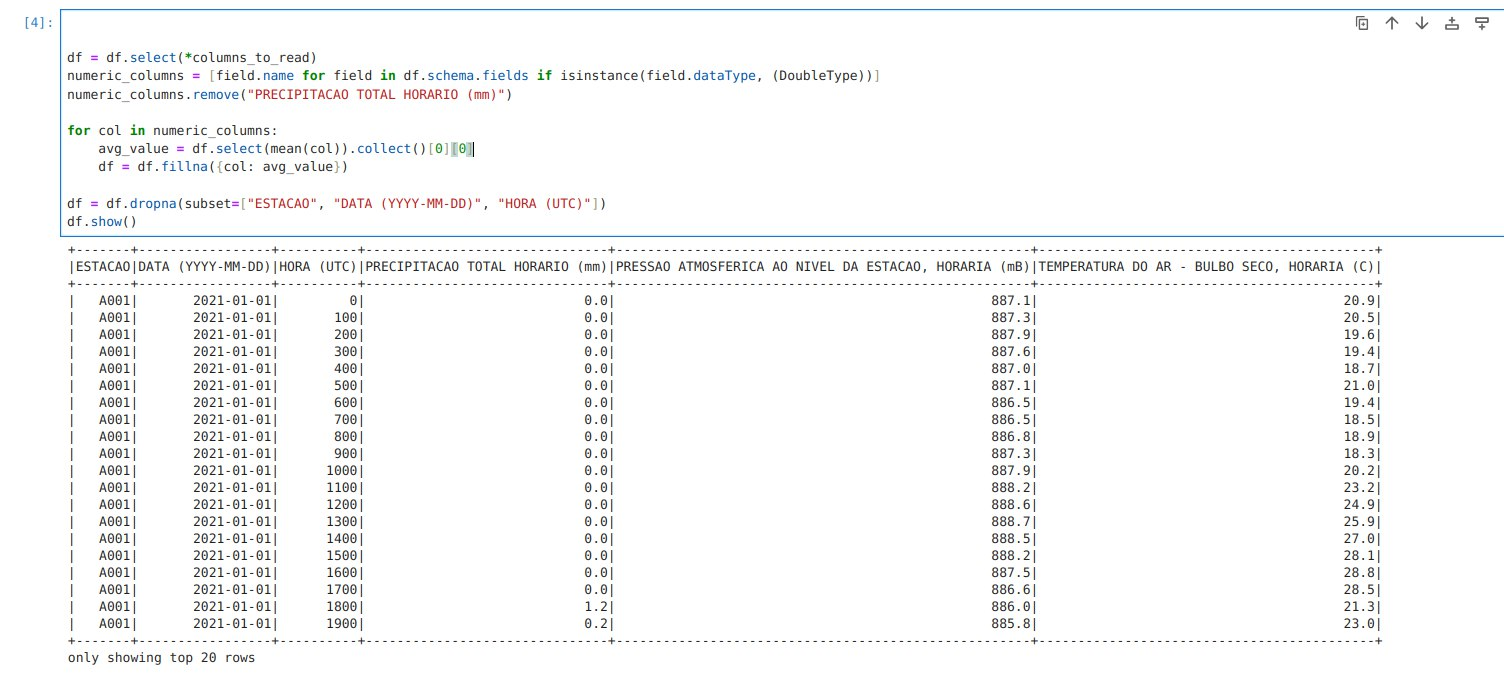


Рисунок 4 – Данные для анализа

С помощью квантилей и медианных значений были заменены выбросы на границы квантилей. Рисунок 5 показывает и результат его выполнения.



Рисунок 5 – Замена выбросов

Следующим этапом мы рассчитали основные статистические показатели (рисунок 6).

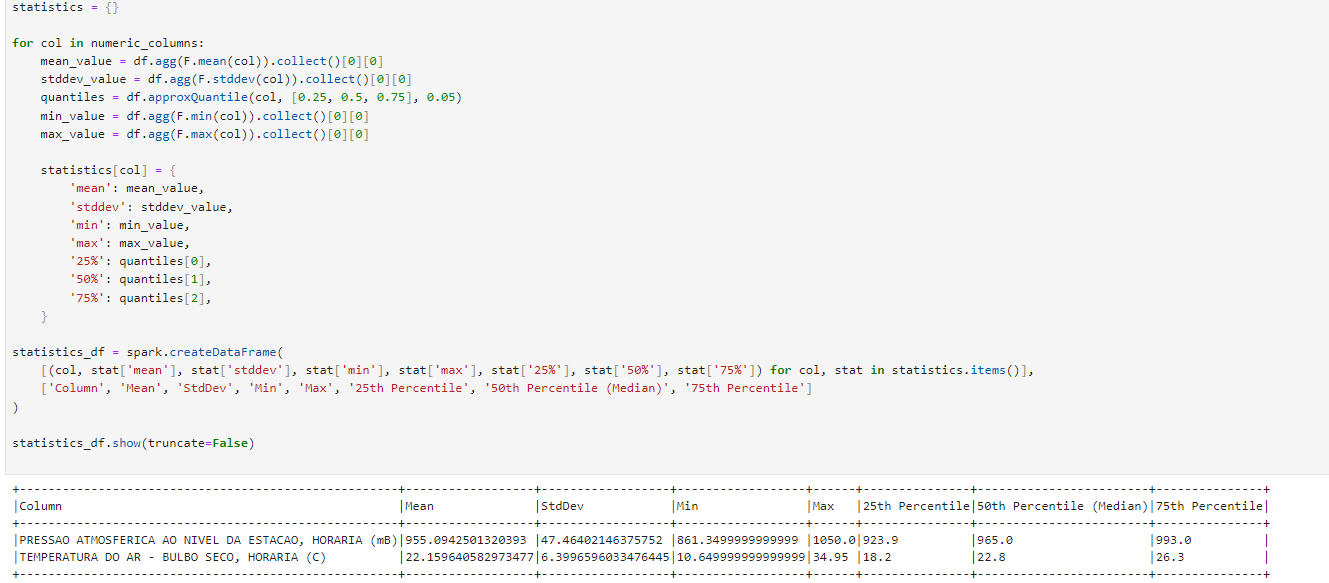


Рисунок 6 – Статистические показатели

Следующим пунктом была получена матрица корреляции между признаками. Код и матрица представлены на рисунке 7.

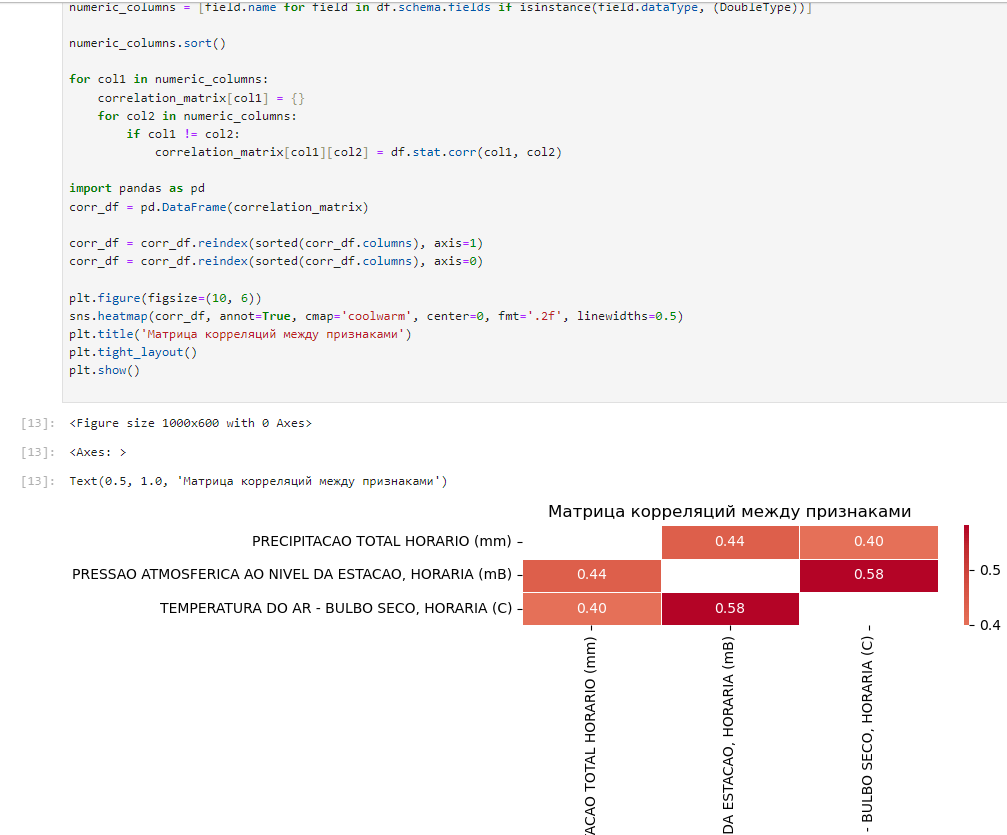


Рисунок 7 – Матрица корреляций между признаками

Между осадками и атмосферным давлением имеется положительная корреляция в размере 0.44. Это означает, что изменения в одном признаке сопровождаются изменениями в другом признаке на умеренном уровне. Между осадками и температурой также существует положительная корреляция в размере 0.4. Это говорит о том, что изменения в этих признаках идут в одном направлении и не сильно связаны между собой. Между давлением и температурой самый большой уровень корреляции – 0.58. Это указывает на наличие довольно сильной связи между этими признаками.

Заключающим этапом стало построение двух графиков, средних значений параметров по датам. Первый график средней температуры (рисунок 8), а второй среднего давления (рисунок 9).

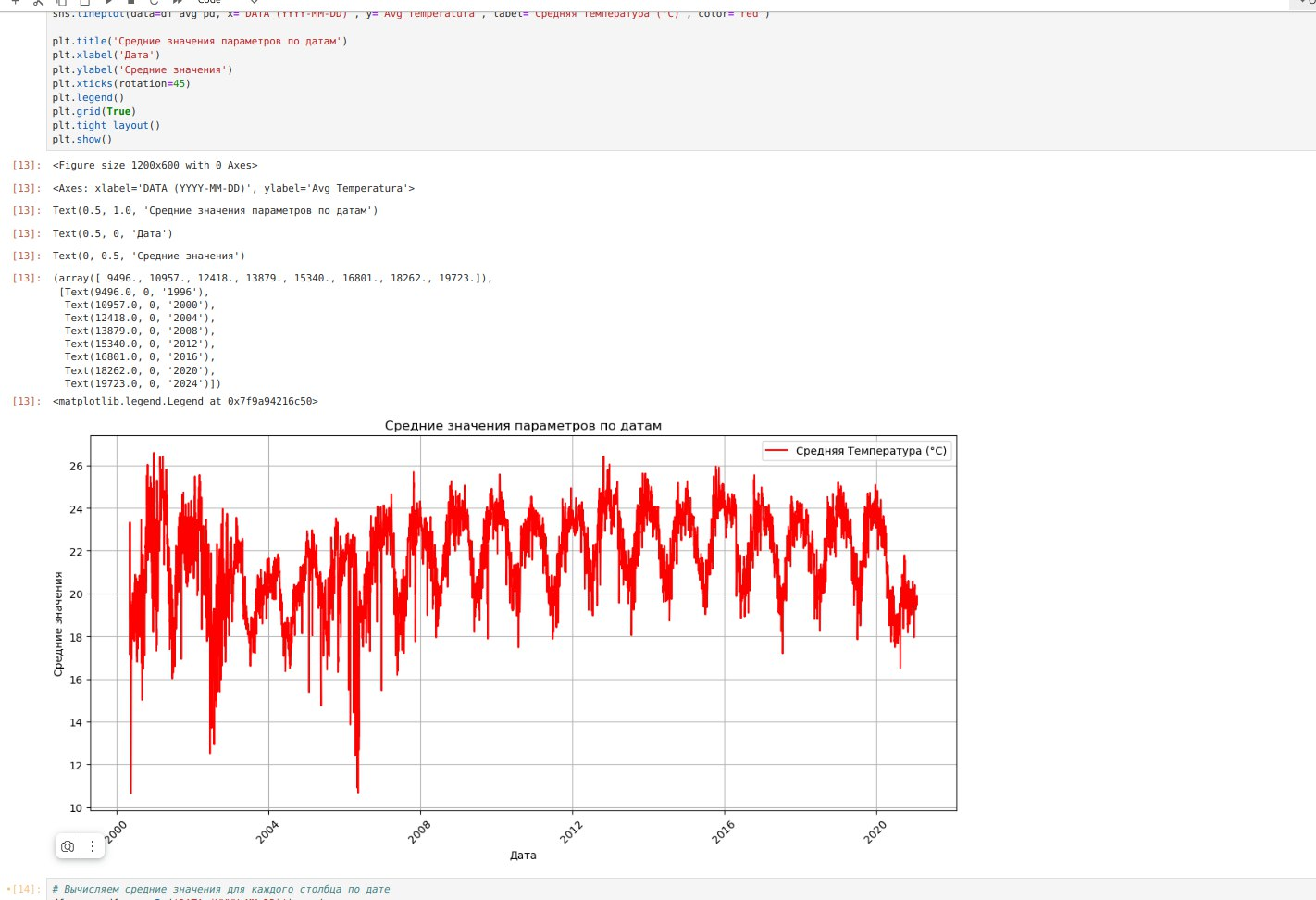


Рисунок 8 – Средняя температура относительно даты



Рисунок 9 – Среднее давление относительно даты

По графику средней температуры видно, что температура циклична и по графику можно наблюдать явную сезонность, чего нельзя явно сказать про давление. Давление не зависит от даты, поэтому график достаточно хаотичен.