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:
Проектный практикум по разработке ETL-решений

Вебинар №2

Тема:

«Бизнес кейс «Umbrella»

Выполнил(а): Морозова Валерия АДЭУ-211

Преподаватель:

Москва

2025

4.1.1. Развернуть ВМ ubuntu_mgpu.ova в VirtualBox.

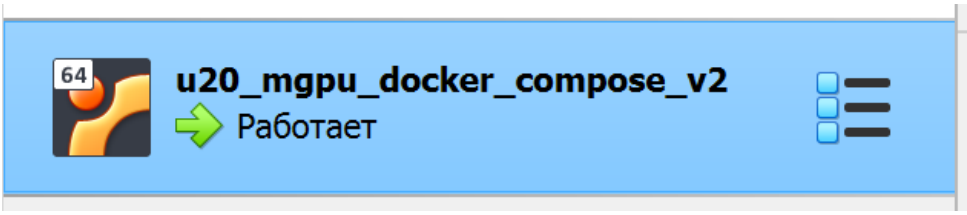


Рисунок 1. Образ развернут

4.1.2. Клонировать на ПК задание **Бизнес кейс Umbrella** в домашний каталог ВМ.

```
git clone https://github.com/BosenkoTM/workshop-on-ETL.git
```

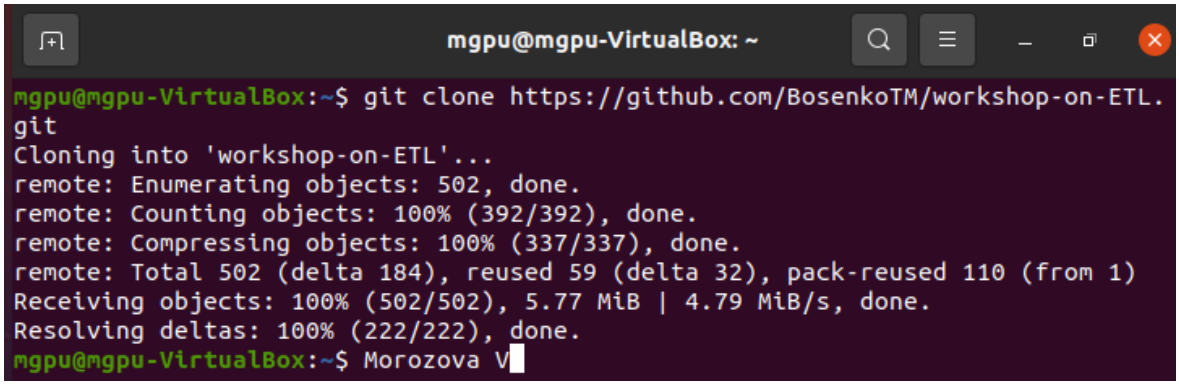


Рисунок 2. Задание клонировано в домашний каталог

4.1.3. Запустить контейнер с кейсом, изучить и описать основные элементы интерфейса `Apache Airflow`.

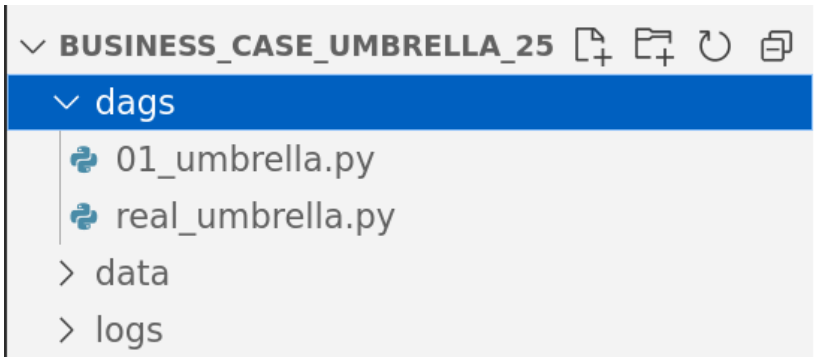


Рисунок 3. Создание папки для записи данных и логов

В качестве основных элементов выступают файлы docker-compose.yaml, Dockerfile и, конечно, файлы dags. В data выгружаются сгенерированные данные после запуска DAG.

4.1.4. Спроектировать верхнеуровневую архитектуру аналитического решения задания **Бизнес кейс Umbrella** в `draw.io`. Необходимо использовать:

`Source Layer` - слой источников данных.

`Storage Layer` - слой хранения данных.

`Business Layer` - слой для доступа к данным бизнес пользователей.

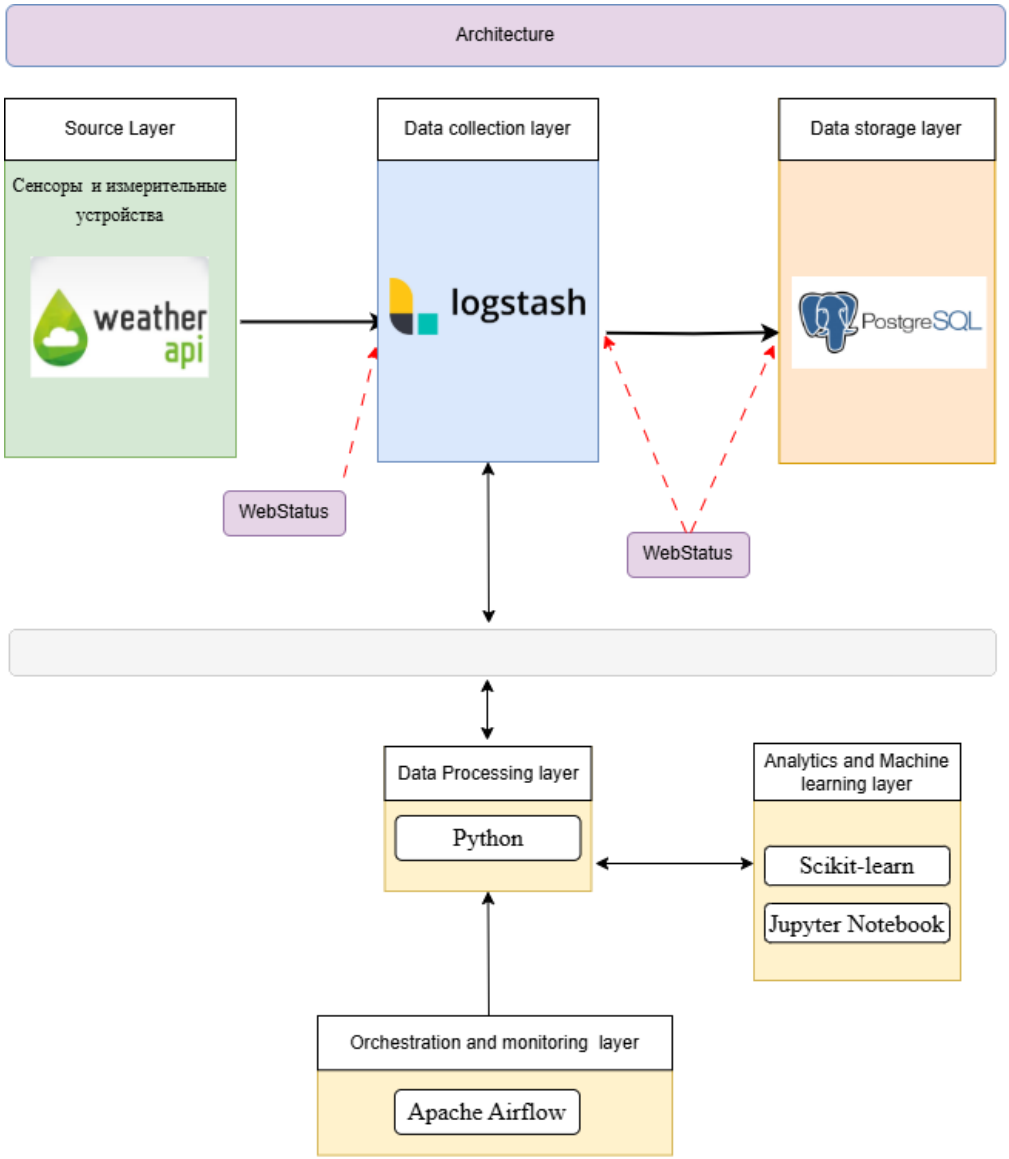


Рисунок 4. Архитектура работы приложения

```
mgpu@mgpu-VirtualBox:~/workshop-on-ETL/business_case_umbrella_25$ sudo docker ps
[sudo] password for mgpu:
CONTAINER ID   IMAGE     COMMAND   CREATED   STATUS    PORTS     NAMES
mgpu@mgpu-VirtualBox:~/workshop-on-ETL/business_case_umbrella_25$
```

Рисунок 5. Проверка запущенных контейнеров

Как видно, запущенных контейнеров нет.

```
mgpu@mgpu-VirtualBox:~/workshop-on-ETL/business_case_umbrella_25$ sudo chown -R 50000:5
0000 ./data
mgpu@mgpu-VirtualBox:~/workshop-on-ETL/business_case_umbrella_25$ Valeria M
```

Рисунок 6. Установление прав к данным

```
● mgpu@mgpu-VirtualBox:~/workshop-on-ETL/business_case_umbrella_25$ sudo docker build -t custom-airflo
w:slim-2.8.1-python3.11 .
[+] Building 127.2s (7/7) FINISHED                                                    docker:default
 => [internal] load build definition from Dockerfile                                          0.1s
 => => transferring dockerfile: 568B                                                          0.0s
 => [internal] load metadata for docker.io/apache/airflow:slim-2.8.1-python3.11               2.3s
 => [internal] load .dockerignore                                                             0.1s
 => => transferring context: 2B                                                               0.0s
 => [1/3] FROM docker.io/apache/airflow:slim-2.8.1-python3.11@sha256:751babd58a83e44ae23c393  55.8s
 => => resolve docker.io/apache/airflow:slim-2.8.1-python3.11@sha256:751babd58a83e44ae23c393f  0.1s
 => => sha256:e1caac4eb9d2ec24aa3618e5992208321a92492aef5fef5eb9e470895f771 29.12MB / 29.12MB  5.7s
 => => sha256:3ee88b8d122ebb0fbb9be864918a05a7621f1b4e1801154b2a0bd64e9476c33 4.47kB / 4.47kB  0.0s
 => => sha256:a205efa96734ac8633bf8d388ed9b6cd527835d31ebee070ba1cedfb880b4 25.59kB / 25.59kB  0.0s
```

Рисунок 7. Собираем Docker-образ из текущей директории (.) и присваивает ему тег

```
webserver-1  | 172.18.0.1 - - [21/Mar/2025:10:37:36 +0000] "GET /static/dist/main.9645e1e98ff7a669af
f7.css HTTP/1.1" 304 0 "http://localhost:8080/home" "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:123.
0) Gecko/20100101 Firefox/123.0"
webserver-1  | 172.18.0.1 - - [21/Mar/2025:10:37:36 +0000] "GET /static/dist/loadingDots.84963375c34
df3f17aab.css HTTP/1.1" 200 0 "http://localhost:8080/home" "Mozilla/5.0 (X11; Ubuntu; Linux x86_64;
rv:123.0) Gecko/20100101 Firefox/123.0"
webserver-1  | 172.18.0.1 - - [21/Mar/2025:10:37:38 +0000] "POST /blocked HTTP/1.1" 200 2 "http://lo
calhost:8080/home" "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:123.0) Gecko/20100101 Firefox/123.0"
webserver-1  | 172.18.0.1 - - [21/Mar/2025:10:37:38 +0000] "POST /dag_stats HTTP/1.1" 200 318 "http:
//localhost:8080/home" "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:123.0) Gecko/20100101 Firefox/123
.0"
webserver-1  | 172.18.0.1 - - [21/Mar/2025:10:37:38 +0000] "POST /last_dagruns HTTP/1.1" 200 2 "http
://localhost:8080/home" "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:123.0) Gecko/20100101 Firefox/12
3.0"
webserver-1  | 172.18.0.1 - - [21/Mar/2025:10:37:38 +0000] "POST /task_stats HTTP/1.1" 200 1048 "htt
p://localhost:8080/home" "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:123.0) Gecko/20100101 Firefox/1
```

Рисунок 8. Сборка и запуск контейнеров

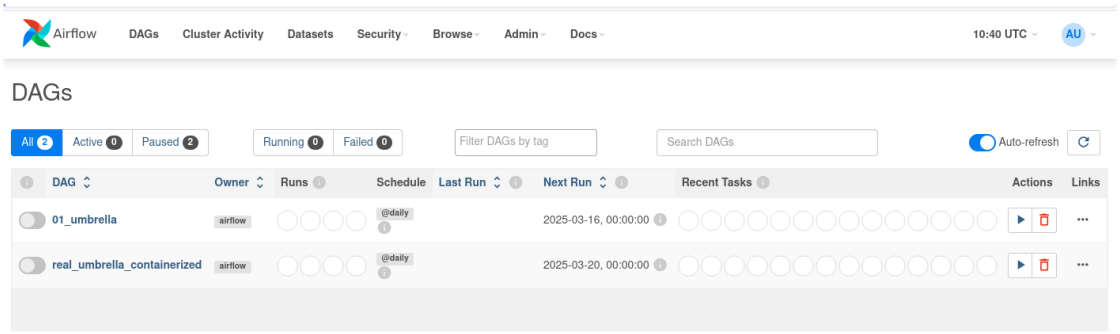


Рисунок 9. Проверка доступа Airflow

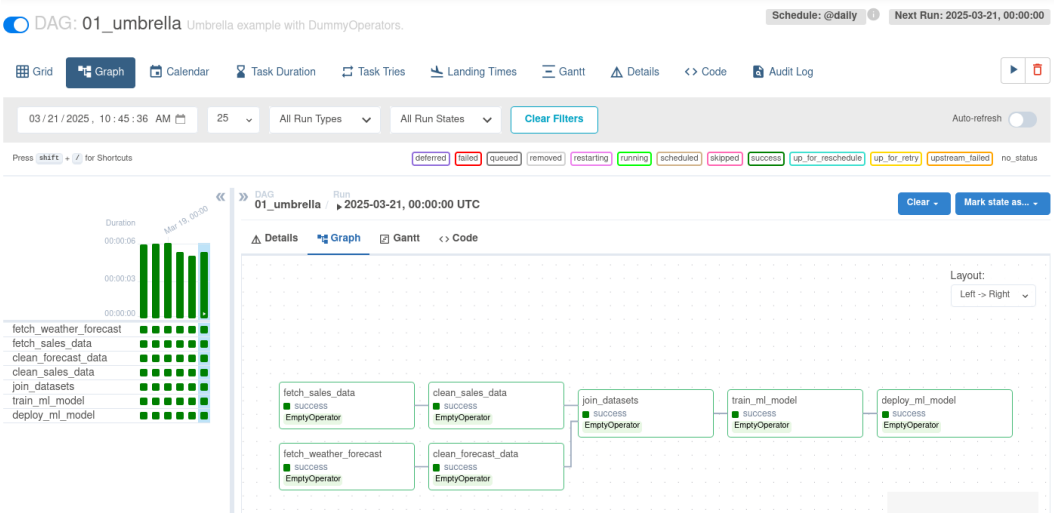


Рисунок 10. Запуск DAG 01_umbrella

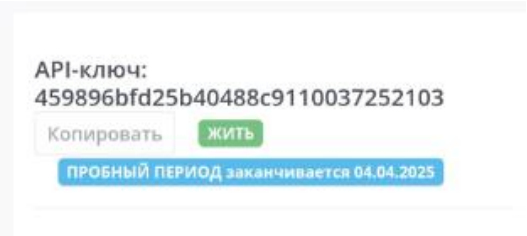


Рисунок 11. Генерация своего ключа API

```
key={api_key}&q=New York&days=3"
```

Рисунок 12. Внесение изменений согласно варианту №10

Получить прогноз в Нью-Йорке на 3 дня	Удалить строки с пропусками	Построить таблицу: дата и температура

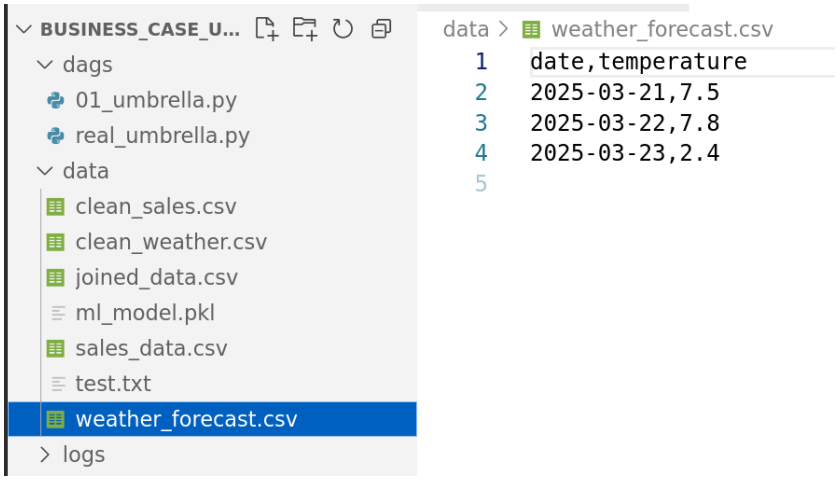


Рисунок 13. Прогноз температуры на 3 дня в Нью Йорке

После триггера второго DAG файла были выгружены файлы, в которых содержатся сгенерированные и очищенные данные по продажам и погоде, файл, где данные объединены и модель машинного обучения.

```
mgpu@mgpu-VirtualBox:~/workshop-on-ETL/business_case_umbrella_25$ sudo docker cp b
usiness_case_umbrella_25-webserver-1:/opt/airflow/data/ml_model.pkl ./ml_model.pkl
Successfully copied 2.56kB to /home/mgpu/workshop-on-ETL/business_case_umbrella_25
/ml_model.pkl
mgpu@mgpu-VirtualBox:~/workshop-on-ETL/business_case_umbrella_25$ Valeria Morozova
```

Рисунок 14. Перенесение файла с моделью

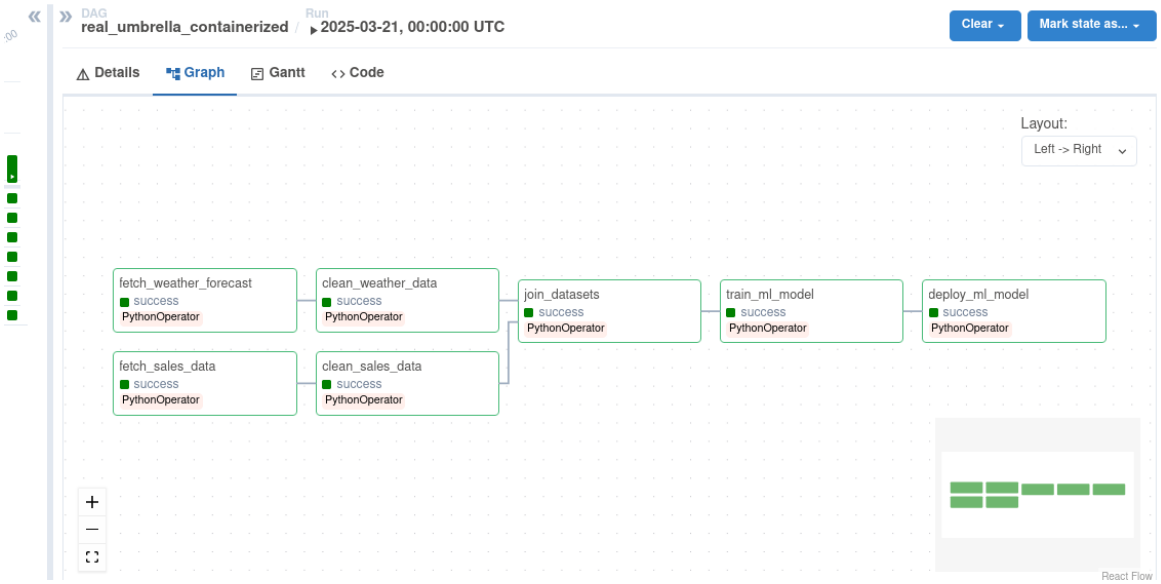


Рисунок 15. Повторный запуск второго DAG с добавлением генерации 500 строк продаж и погоды за 30 дней

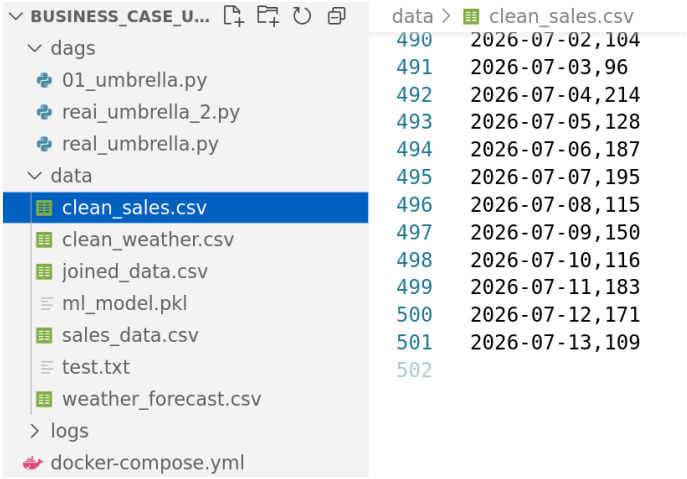


Рисунок 16. Сгенерированы 500 строк в таблице продаж

Все файлы выгружены с виртуальной машины, модель, обученная на 500 значениях запущена в Google Collab.

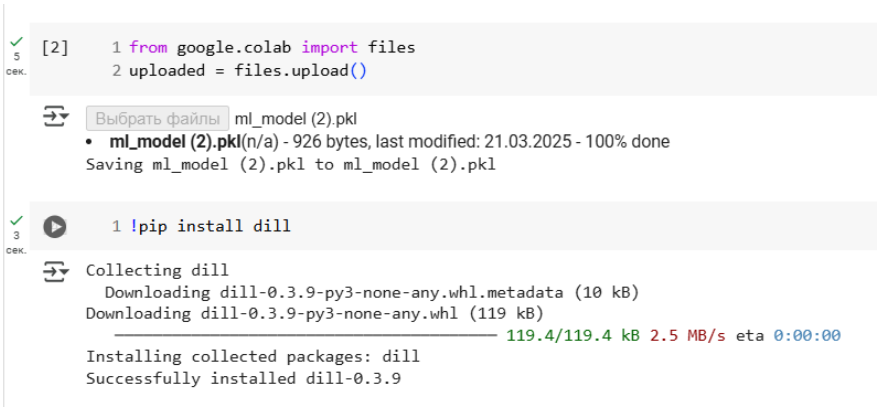
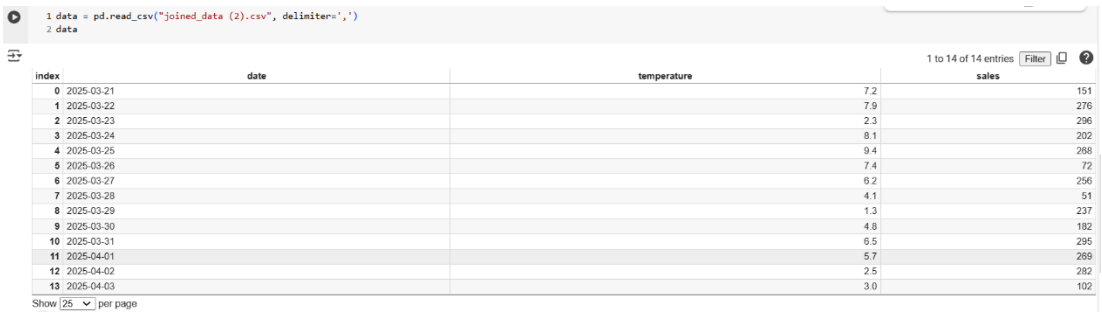


Рисунок 17. Загрузка файла модели и установка пакетов dill

Рисунок 18. Запуск и тестирование модели

Как видно, значения довольно детализированы, что говорит о неплохой обучаемости модели.

Рисунок 19. Загрузка двух файлов (объединенного и прогнозирующего)



1 data = pd.read_csv("joined_data (2).csv", delimiter=',')
2 data

1 to 14 of 14 entries Filter

index	date	temperature	sales
0	2025-03-21	7.2	151
1	2025-03-22	7.9	276
2	2025-03-23	2.3	296
3	2025-03-24	8.1	202
4	2025-03-25	9.4	268
5	2025-03-26	7.4	72
6	2025-03-27	6.2	256
7	2025-03-28	4.1	51
8	2025-03-29	1.3	237
9	2025-03-30	4.8	182
10	2025-03-31	6.5	295
11	2025-04-01	5.7	269
12	2025-04-02	2.5	282
13	2025-04-03	3.0	102

Show 25 per page

Рисунок 20. Отображение в виде таблицы

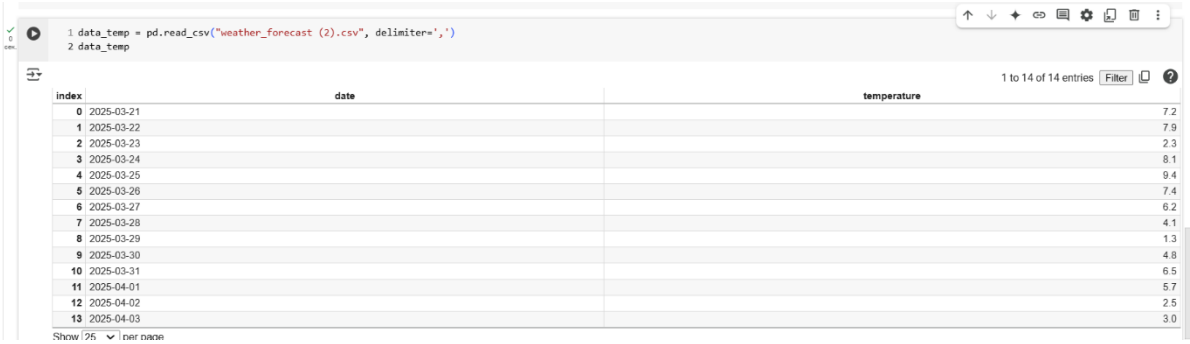
```
[12]  1 #проверка на наличие нулевых значений
      2 data.isna().sum().sum()/len(data)

np.float64(0.0)

      1 #отбрасывание нулевых значений
      2 data.dropna(inplace=True)

[14]  1 data_temp = pd.read_csv("weather_forecast (2).csv", delimiter=',')
      2 data_temp.head(3)
```

Рисунок 21. Проверка на наличие нулевых значений



```
1 data_temp = pd.read_csv("weather_forecast (2).csv", delimiter=',')
2 data_temp
```

1 to 14 of 14 entries Filter

index	date	temperature
0	2025-03-21	7.2
1	2025-03-22	7.9
2	2025-03-23	2.3
3	2025-03-24	8.1
4	2025-03-25	9.4
5	2025-03-26	7.4
6	2025-03-27	6.2
7	2025-03-28	4.1
8	2025-03-29	1.3
9	2025-03-30	4.8
10	2025-03-31	6.5
11	2025-04-01	5.7
12	2025-04-02	2.5
13	2025-04-03	3.0

Show 25 per page

Рисунок 22. Выведение таблицы Дата, температура

Выводы:

1.1. Развернута ВМ ubuntu_mgpu.ova в VirtualBox.

1.2. Клонирована на ПК задание Бизнес кейс Umbrella в домашний каталог ВМ.

1.3. Запущен контейнер 01_umbrella.py с кейсом, изучены и описаны основные элементы интерфейса Apache Airflow.

1.4. Спроектирована верхнеуровневая архитектура для real_umbrella.py аналитического решения задания Бизнес кейс Umbrella в draw.io.

1.5. Запущены dags и сгенерированы 500 значений для качественного обучения модели

1.6. Модель обучена и протестирована

1.7. Выполнено индивидуальное задание

Получен прогноз в Нью-Йорке на 3 дня	Удалены строки с пропусками	Построена таблица: дата и температура