Assignment 2

Cloud computing

Name :Youstena Malak Tadros.

ID: 2206199.

Department of cyper security(level 2).

Steps:

1) I create a directory called dockerimages, then I download the data set and store it in this file.

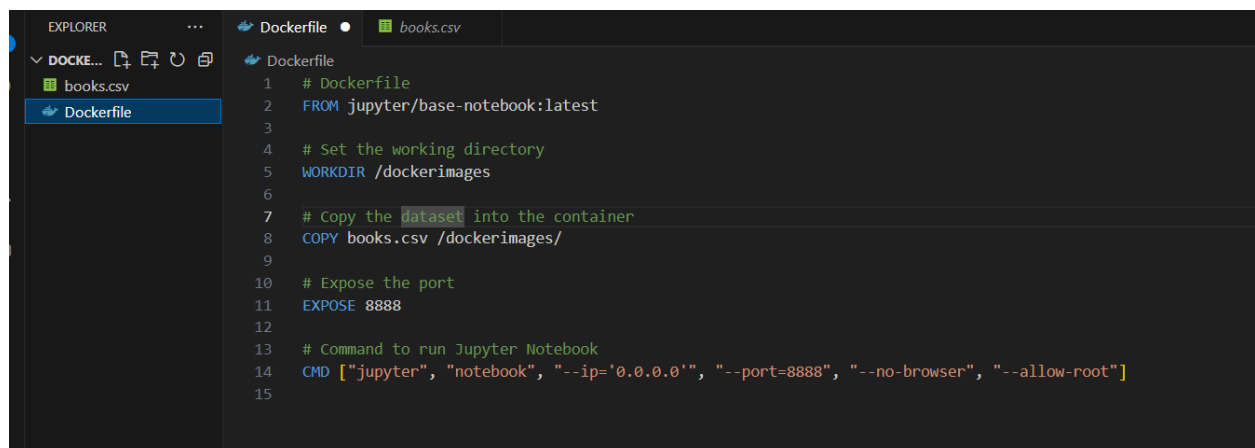2) Pull the base image from docker hub by this written command in the power shell

3) Open vs code, then write these commands to make a Dockerfile (the explanation of each command is in the above comment)

4) In CMD, I wrote this command to build the image

```
C:\Users\LENOVO>docker build -t harrybook C:\dockerimages
2024/04/22 20:07:41 http2: server: error reading preface from client //./pipe/docker_engine: file has already been closed
[+] Building 2.7s (8/8) FINISHED                                                                          docker:default
 => [internal] load build definition from Dockerfile                                                               0.0s
 => => transferring dockerfile: 429B                                                                               0.0s
 => [internal] load metadata for docker.io/jupyter/base-notebook:latest                                            2.5s
 => [internal] load .dockerignore                                                                                  0.0s
 => => transferring context: 2B                                                                                    0.0s
 => [1/3] FROM docker.io/jupyter/base-notebook:latest@sha256:8c903974902b0e9d45d9823c2234411de0614c5c98c4bb782b3d  0.0s
 => [internal] load build context                                                                                  0.0s
 => => transferring context: 32B                                                                                   0.0s
 => CACHED [2/3] WORKDIR /dockerimages                                                                             0.0s
 => CACHED [3/3] COPY books.csv /dockerimages/                                                                     0.0s
 => exporting to image                                                                                             0.0s
 => => exporting layers                                                                                            0.0s
 => => writing image sha256:45aaf0a585262a413ed4ec2b3264f26673e7db56fb14871c6749918bd80b0dd8                       0.0s
 => => naming to docker.io/library/harrybook                                                                       0.0s
```

5) Running container

6) Open docker to make sure the images and container are exist

	Name	Tag	Status	Created	Size	Actions		
☐	**harrybook** 45aaf0a58526 📋	latest	In use	52 minutes ag	1.06 GB	▶	⋮	🗑
☐	**jupyter/datascience-notebook** f78a42f3bc9a 📋	latest	Unused	6 months ago	5.92 GB	▶	⋮	🗑

	Name	Image	Status	CPU (%)	Port(s)	Last started	Actions
☐	great_dewd 3690d74149e	harrybook	Running	0%	8888:8888 ⧉	2 days ago	▪ ⋮ 🗑

7) Enter to this link

127.0.0.1:8888/tree

8) Go to this notebook to write the code

▪ Files	⬤ Running			
Open	Download	Rename	Duplicate	**Delete**

📁 /

☐ Name	Last Modified	File Size
✓ • 📓 Untitled.ipynb	1 minute ago	4.3 KB
☐ 📓 Untitled2.ipynb	2 minutes ago	3.2 KB
☐ ⊞ books.csv	yesterday	422.5 KB
☐ ⊞ harry_potter_best_selling_books.csv	1 minute ago	1.2 KB

9) Write the code

```python
import pandas as pd

# Load the dataset
df = pd.read_csv("books.csv")

# Perform data cleaning and preprocessing
# Handle missing values
df.dropna(inplace=True)

# Remove unnecessary columns
df.drop(['book_id', 'goodreads_book_id', 'best_book_id', 'work_id', 'isbn', 'isbn13', 'image_url', 'small_image_url'], axis=1, inplace=True)

# Filter dataset for Harry Potter series
harry_potter_df = df[df['authors'].str.contains('J.K. Rowling') & df['title'].str.contains('Harry Potter')]

# Check the filtered dataset
harry_potter_df.head()

# Find the most selling books within the Harry Potter series
most_selling_books = harry_potter_df.sort_values(by='ratings_count', ascending=False).head()

# Display the most selling books
print(most_selling_books[['title', 'authors', 'ratings_count', 'average_rating']])

# Save the most selling books to a CSV file
most_selling_books.to_csv("harry_potter_best_selling_books.csv", index=False)

# Calculate the average rating of the Harry Potter books
average_rating_hp = harry_potter_df['average_rating'].mean()

# Display the average rating
print(f"Average rating of Harry Potter books: {average_rating_hp}")
```

10) Output

```
                                                       title  \
1     Harry Potter and the Sorcerer's Stone (Harry P...
6     Harry Potter and the Prisoner of Azkaban (Harr...
9     Harry Potter and the Chamber of Secrets (Harry...
10    Harry Potter and the Goblet of Fire (Harry Pot...
11    Harry Potter and the Deathly Hallows (Harry Po...

                                 authors  ratings_count  average_rating
1             J.K. Rowling, Mary GrandPré        4602479            4.44
6    J.K. Rowling, Mary GrandPré, Rufus Beck       1832823            4.53
9             J.K. Rowling, Mary GrandPré        1779331            4.37
10            J.K. Rowling, Mary GrandPré        1753043            4.53
11            J.K. Rowling, Mary GrandPré        1746574            4.61
Average rating of Harry Potter books: 4.550000000000001
```

11) Explain of this code: Importing pandas and read a CSV file, then Removing missing values and drop Unnecessary columns like 'book_id', 'goodreads_book_id', A new Data Frame is created by filtering rows where the 'authors' column contains 'J.K. Rowling' and the 'title' column contains 'Harry Potter'. After that Finding the most selling books to save the most selling books we create another csv file to store it then the code prints the details (title, authors, ratings count, and average rating) of the most selling Harry Potter books. Finally, Calculating average rating and print it.