Cloud Computing
Assignment 2

Name: Zainab Mohamed Abdallah
ID: 20221310251
Department: AI level 3

For the collab notebook:

I first uploaded the dataset onto collab using the pandas read_csv command.

Then I did the data preprocessing and data cleaning:
First I identified the missing values:

Data Preprocessing

```
[ ]  #identify missing values
     print("\nMissing values:")
     print(df.isnull().sum())
```

```
Missing values:
book_id                        0
goodreads_book_id              0
best_book_id                   0
work_id                        0
books_count                    0
isbn                          52
isbn13                        44
authors                        0
original_publication_year      3
original_title                52
title                          0
language_code                109
average_rating                 0
ratings_count                  0
work_ratings_count             0
work_text_reviews_count        0
ratings_1                      0
ratings_2                      0
ratings_3                      0
ratings_4                      0
ratings_5                      0
image_url                      0
small_image_url                0
dtype: int64
```

After that, I handled the missing values by filling in the NA values by the mode of the columns:

```
#Fill missing values in this column with the most frequent value
df['isbn'].fillna(df['isbn'].mode()[0], inplace=True)
```

```
[5] #Fill missing values in this column with the most frequent value
df['isbn13'].fillna(df['isbn13'].mode()[0], inplace=True)
```

```
[6] #Fill missing values in this column with the most frequent value
df['original_publication_year'].fillna(df['original_publication_year'].mode()[0], inplace=True)
```

```
[7] #Fill missing values in this column with the most frequent value
df['original_title'].fillna(df['original_title'].mode()[0], inplace=True)
```

```
[8] #Fill missing values in this column with the most frequent value
df['language_code'].fillna(df['language_code'].mode()[0], inplace=True)
```

Missing values after handling them:

```
[9] #identify missing values after filling NA
print("\nMissing values:")
print(df.isnull().sum())
```

```
Missing values:
book_id                    0
goodreads_book_id          0
best_book_id               0
work_id                    0
books_count                0
isbn                       0
isbn13                     0
authors                    0
original_publication_year  0
original_title             0
title                      0
language_code              0
average_rating             0
ratings_count              0
work_ratings_count         0
work_text_reviews_count    0
ratings_1                  0
ratings_2                  0
ratings_3                  0
ratings_4                  0
ratings_5                  0
image_url                  0
small_image_url            0
dtype: int64
```

Then I cleaned the data by dropping the duplicates:

```
# Remove duplicates
df.drop_duplicates(inplace=True)
df
```

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publicatio
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	
3	6	11870085	11870085	16827462	226	525478817	9.780525e+12	John Green	
4	12	13335037	13335037	13155899	210	62024035	9.780062e+12	Veronica Roth	
...	...	...	...	...	...	...	...	...	
1349	9925	86737	86737	3877968	52	1582349177	9.781582e+12	Mary Hoffman	
1350	9937	13010211	13010211	18171867	22	1596435712	9.781596e+12	Caragh M. O'Brien	
1351	9942	16074758	16074758	21869436	18	1442486597	9.781442e+12	Abigail Haas, Abby McDonald	
1352	9947	21393526	21393526	40690062	19	62320521	9.780062e+12	Maria Dahvana Headley	

Then I saved the cleaned and preprocessed dataframe.

As for the harry potter data:

I filtered the harry potter book series from the rest of the books:

Harry Potter books analysis

```
[12] # Filter the dataset for titles containing "Harry Potter"
     harry_potter_books = df[df['title'].str.contains('Harry Potter', case=False)]
```

Then I calculated the average rating for each harry potter series book:

```
# Calculate the average rating for each Harry Potter book
average_ratings = harry_potter_books.groupby('title')[['ratings_1', 'ratings_2', 'ratings_3', 'ratings_4', 'ratings_5']]

# Display the average rating for each Harry Potter book
print("Average Ratings of Harry Potter Books:")
print(average_ratings)
```

```
Average Ratings of Harry Potter Books:
                                                ratings_1  ratings_2  \
title
Harry Potter Boxset (Harry Potter, #1-7)           1105.0     1285.0
Harry Potter Collection (Harry Potter, #1-6)        203.0      186.0
Harry Potter Schoolbooks Box Set: Two Classic B...  106.0      304.0
Harry Potter and the Chamber of Secrets (Harry ...  8253.0    42251.0
Harry Potter and the Deathly Hallows (Harry Pot...  9363.0    22245.0
Harry Potter and the Goblet of Fire (Harry Pott...  6676.0    20210.0
Harry Potter and the Half-Blood Prince (Harry P...  7308.0    21516.0
Harry Potter and the Order of the Phoenix (Harr...  9528.0    31577.0
Harry Potter and the Prisoner of Azkaban (Harry...  6716.0    20413.0
Harry Potter and the Sorcerer's Stone (Harry Po...  75504.0  101676.0
The Magical Worlds of Harry Potter: A Treasury ...   329.0     1125.0

                                                ratings_3  ratings_4  \
title
Harry Potter Boxset (Harry Potter, #1-7)           7020.0    30666.0
Harry Potter Collection (Harry Potter, #1-6)        946.0     3891.0
Harry Potter Schoolbooks Box Set: Two Classic B...  1548.0     2595.0
Harry Potter and the Chamber of Secrets (Harry ...  242345.0  548266.0
Harry Potter and the Deathly Hallows (Harry Pot...  113646.0  383914.0
Harry Potter and the Goblet of Fire (Harry Pott...  151785.0  494926.0
Harry Potter and the Half-Blood Prince (Harry P...  136333.0  459028.0
Harry Potter and the Order of the Phoenix (Harr...  180210.0  494427.0
Harry Potter and the Prisoner of Azkaban (Harry...  166129.0  509447.0
Harry Potter and the Sorcerer's Stone (Harry Po...  455024.0  1156318.0
The Magical Worlds of Harry Potter: A Treasury ...   3766.0     3593.0

                                                ratings_5
title
Harry Potter Boxset (Harry Potter, #1-7)           164049.0
Harry Potter Collection (Harry Potter, #1-6)        21048.0
Harry Potter Schoolbooks Box Set: Two Classic B...   7179.0
Harry Potter and the Chamber of Secrets (Harry ...  1065084.0
Harry Potter and the Deathly Hallows (Harry Pot...  1318227.0
Harry Potter and the Goblet of Fire (Harry Pott...  1195045.0
Harry Potter and the Half-Blood Prince (Harry P...  1161491.0
Harry Potter and the Order of the Phoenix (Harr...  1124806.0
Harry Potter and the Prisoner of Azkaban (Harry...  1266670.0
Harry Potter and the Sorcerer's Stone (Harry Po...  3011543.0
The Magical Worlds of Harry Potter: A Treasury ...     6332.0
```

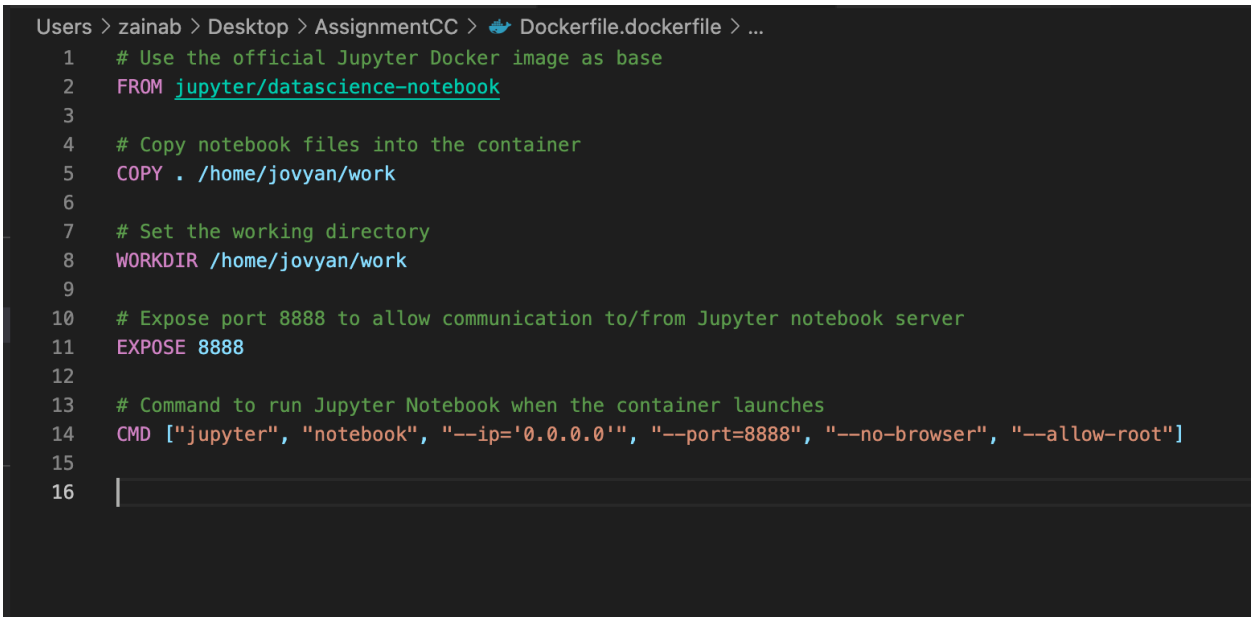
Then I calculated the top selling harry potter book depending on the average rating:

```
# Find the book with the highest sales (prints all info about the book)
top_selling_book = harry_potter_books.loc[harry_potter_books['average_rating'].idxmax()]

# Display the top-selling book
print("Top Selling Book within the Harry Potter Series:")
print(top_selling_book)
```

```
Top Selling Book within the Harry Potter Series:
book_id                                                                422
goodreads_book_id                                                   862041
best_book_id                                                        862041
work_id                                                            2962492
books_count                                                             76
isbn                                                             545044251
isbn13                                                    9780545044260.0
authors                                                       J.K. Rowling
original_publication_year                                           1998.0
original_title                                  Complete Harry Potter Boxed Set
title                                  Harry Potter Boxset (Harry Potter, #1-7)
language_code                                                           eng
average_rating                                                         4.74
ratings_count                                                        190050
work_ratings_count                                                   204125
work_text_reviews_count                                                6508
ratings_1                                                              1105
ratings_2                                                              1285
ratings_3                                                              7020
ratings_4                                                             30666
ratings_5                                                            164049
image_url                  https://images.gr-assets.com/books/1392579059m...
small_image_url            https://images.gr-assets.com/books/1392579059s...
Name: 96, dtype: object
```

As for the docker file, I wrote the following code:



And saved it as a docker file.
After that, I opened docker and then I wrote the following commands in the terminal:

- `docker pull jupyter/datascience-notebook`
- `cd /Users/zainab/Desktop/AssignmentCC`
- `docker build -t my-notebook -f /Users/zainab/Desktop/AssignmentCC/Dockerfile.dockerfile .`
- `docker run -p 8888:8888 my-notebook`

The output of the last command was the following:

```
(base) zainab@192 AssignmentCC % docker run -p 8888:8888 my-notebook
[I 2024-04-23 20:56:29.714 ServerApp] Package notebook took 0.0000s to import
[I 2024-04-23 20:56:29.724 ServerApp] Package jupyter_lsp took 0.0098s to import
[W 2024-04-23 20:56:29.724 ServerApp] A `_jupyter_server_extension_points` function was not found in jupyter_lsp. Instead, a `_jupyter_server_extension_paths` function was found and will be used for now.
This function name will be deprecated in future releases of Jupyter Server.
[I 2024-04-23 20:56:29.726 ServerApp] Package jupyter_server_mathjax took 0.0013s to import
[I 2024-04-23 20:56:29.769 ServerApp] Package jupyter_server_proxy took 0.0427s to import
[I 2024-04-23 20:56:29.775 ServerApp] Package jupyter_server_terminals took 0.0057s to import
[I 2024-04-23 20:56:29.775 ServerApp] Package jupyterlab took 0.0000s to import
[I 2024-04-23 20:56:30.220 ServerApp] Package jupyterlab_git took 0.0280s to import
[I 2024-04-23 20:56:30.223 ServerApp] Package nbclassic took 0.0033s to import
[W 2024-04-23 20:56:30.225 ServerApp] A `_jupyter_server_extension_points` function was not found in nbclassic. Instead, a `_jupyter_server_extension_paths` function was found and will be used for now. Th
is function name will be deprecated in future releases of Jupyter Server.
[I 2024-04-23 20:56:30.225 ServerApp] Package nbdime took 0.0000s to import
[I 2024-04-23 20:56:30.225 ServerApp] Package notebook_shim took 0.0000s to import
[W 2024-04-23 20:56:30.225 ServerApp] A `_jupyter_server_extension_points` function was not found in notebook_shim. Instead, a `_jupyter_server_extension_paths` function was found and will be used for now
. This function name will be deprecated in future releases of Jupyter Server.
[I 2024-04-23 20:56:30.225 ServerApp] jupyter_lsp | extension was successfully linked.
[I 2024-04-23 20:56:30.228 ServerApp] jupyter_server_mathjax | extension was successfully linked.
[I 2024-04-23 20:56:30.228 ServerApp] jupyter_server_proxy | extension was successfully linked.
[I 2024-04-23 20:56:30.230 ServerApp] jupyter_server_terminals | extension was successfully linked.
[I 2024-04-23 20:56:30.232 ServerApp] jupyterlab | extension was successfully linked.
[I 2024-04-23 20:56:30.233 ServerApp] jupyterlab_git | extension was successfully linked.
[I 2024-04-23 20:56:30.235 ServerApp] nbclassic | extension was successfully linked.
[I 2024-04-23 20:56:30.235 ServerApp] nbdime | extension was successfully linked.
[I 2024-04-23 20:56:30.237 ServerApp] notebook | extension was successfully linked.
[I 2024-04-23 20:56:30.238 ServerApp] Writing Jupyter server cookie secret to /home/jovyan/.local/share/jupyter/runtime/jupyter_cookie_secret
[I 2024-04-23 20:56:30.426 ServerApp] notebook_shim | extension was successfully linked.
[I 2024-04-23 20:56:30.437 ServerApp] notebook_shim | extension was successfully loaded.
[I 2024-04-23 20:56:30.438 ServerApp] jupyter_lsp | extension was successfully loaded.
[I 2024-04-23 20:56:30.439 ServerApp] jupyter_server_mathjax | extension was successfully loaded.
[I 2024-04-23 20:56:30.445 ServerApp] jupyter_server_proxy | extension was successfully loaded.
[I 2024-04-23 20:56:30.445 ServerApp] jupyter_server_terminals | extension was successfully loaded.
[I 2024-04-23 20:56:30.451 LabApp] JupyterLab extension loaded from /opt/conda/lib/python3.11/site-packages/jupyterlab
[I 2024-04-23 20:56:30.451 LabApp] JupyterLab application directory is /opt/conda/share/jupyter/lab
[I 2024-04-23 20:56:30.452 LabApp] Extension Manager is 'pypi'.
[I 2024-04-23 20:56:30.453 ServerApp] jupyterlab | extension was successfully loaded.
[I 2024-04-23 20:56:30.455 ServerApp] jupyterlab_git | extension was successfully loaded.
[I 2024-04-23 20:56:30.458 ServerApp] nbclassic | extension was successfully loaded.
[I 2024-04-23 20:56:30.501 ServerApp] nbdime | extension was successfully loaded.
[I 2024-04-23 20:56:30.502 ServerApp] notebook | extension was successfully loaded.
[I 2024-04-23 20:56:30.502 ServerApp] Serving notebooks from local directory: /home/jovyan/work
[I 2024-04-23 20:56:30.502 ServerApp] Jupyter Server 2.8.0 is running at:
[I 2024-04-23 20:56:30.502 ServerApp] http://0e1e2f0f507b:8888/tree?token=1ca56b8e0650e876b3e6b821d515a22b7ee7f309da4fc5eb
[I 2024-04-23 20:56:30.502 ServerApp]     http://127.0.0.1:8888/tree?token=1ca56b8e0650e876b3e6b821d515a22b7ee7f309da4fc5eb
[I 2024-04-23 20:56:30.503 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 2024-04-23 20:56:30.504 ServerApp]

    To access the server, open this file in a browser:
        file:///home/jovyan/.local/share/jupyter/runtime/jpserver-7-open.html
    Or copy and paste one of these URLs:
        http://0e1e2f0f507b:8888/tree?token=1ca56b8e0650e876b3e6b821d515a22b7ee7f309da4fc5eb
        http://127.0.0.1:8888/tree?token=1ca56b8e0650e876b3e6b821d515a22b7ee7f309da4fc5eb
[I 2024-04-23 20:56:31.425 ServerApp] Skipped non-installed server(s): bash-language-server, dockerfile-language-server-nodejs, javascript-typescript-langserver, jedi-language-server, julia-language-serve
r, pyright, python-language-server, python-lsp-server, r-languageserver, sql-language-server, texlab, typescript-language-server, unified-language-server, vscode-css-languageserver-bin, vscode-html-langua
geserver-bin, vscode-json-languageserver-bin, yaml-language-server
0.00s - Debugger warning: It seems that frozen modules are being used, which may
0.00s - make the debugger miss breakpoints. Please pass -Xfrozen_modules=off
0.00s - to python to disable frozen modules.
0.00s - Note: Debugging will proceed. Set PYDEVD_DISABLE_FILE_VALIDATION=1 to disable this validation.
```

After putting the highlighted link in a browsers it opens the notebook and dockerfile like the following: