

Assignment 2 | Dockerfile and Data Analysis with Popular Books Dataset

Name : Abdelrhman Ahmed Fathy

ID : 20221441784 Ai 3rd year

1. Docker Setup:

- Use the official Jupyter Docker image.
- Refer to the documentation for guidance.

2. Dockerfile:

- Create a Dockerfile in my project file Name → data_analysis_project
- Use the Official Jupyter Docker image from the following link: [[Jupyter Docker Image](#)].
- Add instructions to install any additional libraries or dependencies required for your analysis.
- Configure the Dockerfile to run Jupyter Notebook using the provided command:

CMD ["jupyter", "notebook", "--ip='0.0.0.0'", "--port=8888", "--no-browser", "--allow-root"]

3. Notebook Development

- Create a Jupyter Notebook to my project directory.
- Load the Popular Books Dataset.
- Perform data cleaning and preprocessing steps to prepare the dataset.
- Focus your analysis on the Harry Potter series:
 - Filter the dataset to include only Harry Potter books.
 - Analyze sales data to find the most selling books within the series.
 - Calculate the average rating of the Harry Potter books.

4. GitHub Repository:

- Create a GitHub repository.
- upload the following items:
- Dockerfile
- Jupyter Notebook
- Dataset (downloaded from Kaggle)
- PDF file containing analysis results.

Results

After we make Data cleaning and preprocessing

Checking the missing values and Drop any duplicate entries

```
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
```

Drop rows of the missing value

```
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
```

Filter and Check the unique book titles in the Harry Potter series Bec, it help us later of the analysis

Unique Harry Potter Books:

```
["Harry Potter and the Philosopher's Stone"
 'Harry Potter and the Prisoner of Azkaban'
 'Harry Potter and the Order of the Phoenix'
 'Harry Potter and the Chamber of Secrets'
 'Harry Potter and the Goblet of Fire'
 'Harry Potter and the Deathly Hallows'
 'Harry Potter and the Half-Blood Prince'
 'Complete Harry Potter Boxed Set'
 'Harry Potter Collection (Harry Potter, #1-6)'
 'The Magical Worlds of Harry Potter: A Treasury of Myths, Legends, and Fascinating Facts']
```

Discover the Best-Selling Books from the Magical Series

In this part, we find out which books from the magical series are the most popular based on the number of people who have rated them. The number beside each book title shows how many people have rated that book. This number can give us an idea of how well-liked and widely sold the book is.

Most Selling Books in Harry Potter Series:

original_title

Harry Potter and the Philosopher's Stone 4602479

Harry Potter and the Prisoner of Azkaban 1832823

Harry Potter and the Chamber of Secrets 1779331

Harry Potter and the Goblet of Fire 1753043

Harry Potter and the Deathly Hallows 1746574

Name: ratings_count, dtype: int64

- The first book in the series has the most ratings, with 4,602,479 people having rated it.
- The third book comes next with 1,832,823 ratings.
- Following that, the second book has 1,779,331 ratings.
- The fourth book has 1,753,043 ratings.
- Lastly, the fifth book has 1,746,574 ratings, completing the top five.

Average Rating of the Magic Series Books

The average rating for all the magic series books in the data set is 4.49. This average rating gives a general idea of how readers perceive the quality of the books. A higher average rating suggests that readers generally like the books.

Average rating of the Harry Potter books: 4.49