abdelrahman abdullah 2203189 cloud assign report

April 24, 2024

[ ]:
```
#importing libraries and data
import numpy as np
import pandas as pd
books=pd.read_csv('books.csv')
```

[ ]:
```
# Check for missing values in the dataset
missing_values = books.isnull().sum()
missing_percentage = (missing_values / len(books)) * 100
missing_data_info = pd.DataFrame({'Missing_Values': missing_values,␣
 ↪'Missing_Percentage': missing_percentage})
missing_data_info = missing_data_info.sort_values(by='Missing_Percentage',␣
 ↪ascending=False)
print("Columns with the highest percentage of missing values:")
print(missing_data_info.head())
```

```
Columns with the highest percentage of missing values:
                           Missing_Values  Missing_Percentage
language_code                         109            8.050222
isbn                                   52            3.840473
original_title                         52            3.840473
isbn13                                 44            3.249631
original_publication_year               3            0.221566
```

[ ]:
```
# Fill missing values in numerical columns with median and categorical columns␣
 ↪with mode
numerical_cols = books.select_dtypes(include=[np.number]).columns
books[numerical_cols] = books[numerical_cols].fillna(books[numerical_cols].
 ↪median())
categorical_cols = books.select_dtypes(exclude=[np.number]).columns
books[categorical_cols] = books[categorical_cols].
 ↪fillna(books[categorical_cols].mode().iloc[0])
```

[ ]:
```
# Recheck for missing values in the dataset
missing_values = books.isnull().sum()
missing_percentage = (missing_values / len(books)) * 100
missing_data_info = pd.DataFrame({'Missing_Values': missing_values,␣
 ↪'Missing_Percentage': missing_percentage})
```

```
missing_data_info = missing_data_info.sort_values(by='Missing_Percentage',␣
↪ascending=False)
print("Columns with the highest percentage of missing values:")
print(missing_data_info.head())
```

```
Columns with the highest percentage of missing values:
               Missing_Values  Missing_Percentage
book_id                     0                 0.0
average_rating              0                 0.0
image_url                   0                 0.0
ratings_5                   0                 0.0
ratings_4                   0                 0.0
```

[ ]:
```
#Filter harry potter books
harry_potter_df = books[books['title'].str.contains('Harry Potter', case=False)]
```

[ ]:
```
#sort harry potter books by ratings count to get most selled books
sorted_harry_potter_df = harry_potter_df.sort_values(by='ratings_count',␣
↪ascending=False)
print("most selled harry potter books are:",sorted_harry_potter_df["title"].
↪head())
```

```
most selled harry potter books are: 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…
Name: title, dtype: object
```

[ ]:
```
#calculate average rating of all harry potter books
average_rating = harry_potter_df['average_rating'].mean()
print("Average rating of Harry Potter books:", average_rating)
```

```
Average rating of Harry Potter books: 4.482727272727273
```

Dockerfile:

```
Dockerfile > ...
1   FROM jupyter/datascience-notebook
2   WORKDIR /prog
3   COPY books.csv /prog
4   COPY notebook.ipynb /prog
5   EXPOSE 8888
6   CMD ["jupyter", "notebook", "--ip='0.0.0.0'", "--port=8888", "--no-browser", "--allow-root"]
```

Docker image:

```
PS C:\Users\Abdelrahman Abdullah\Desktop\cloud assignment> docker build -t harry_potter_analysis .
[+] Building 1.8s (9/9) FINISHED
 => [internal] load build definition from Dockerfile
 => => transferring dockerfile: 243B
 => [internal] load metadata for docker.io/jupyter/datascience-notebook:latest
 => [internal] load .dockerignore
 => => transferring context: 2B
 => [1/4] FROM docker.io/jupyter/datascience-notebook:latest
 => [internal] load build context
```

```
PS C:\Users\Abdelrahman Abdullah\Desktop\cloud assignment> docker images
REPOSITORY                    TAG       IMAGE ID       CREATED          SIZE
harry_potter_analysis         latest    c19ea98148b7   34 minutes ago   5.92GB
jupyter/datascience-notebook  latest    f78a42f3bc9a   6 months ago     5.92GB
PS C:\Users\Abdelrahman Abdullah\Desktop\cloud assignment>
```