```
# -*- coding: utf-8 -*-
"""EDA project on IMDB.ipynb

Automatically generated by Colab.

Original file is located at

https://colab.research.google.com/drive/1mmT1eAtbKSul7obXnooqWK7O6zCD7p1D

Understand how movie genres, runtime, and release years affect IMDb
ratings and the number of votes. Identify trends in movie production and
viewer preferences.
"""

import pandas as pd

# Load IMDb basics
try:
      df_basics = pd.read_csv("C:\\Users\\Adon
Banker\\Downloads\\title.basics.tsv.gz", sep="\t", na_values='\\N',
compression='gzip')
      df_ratings = pd.read_csv("C:\\Users\\Adon
Banker\\Downloads\\title.ratings.tsv.gz", sep="\t", na_values='\\N',
compression='gzip')
except FileNotFoundError as e:
      print(f"File not found: {e.filename}. Please check the file path.")
      df_basics = None
      df_ratings = None

df = pd.merge(df_basics, df_ratings, on='tconst')

print(df.head())

print(df.shape)

print(df.info())

df = df[["primaryTitle", "startYear", "runtimeMinutes", "genres",
"averageRating", "numVotes","titleType"]]# I only want these columns
df = df[df["titleType"] == "movie"]# I want only movies coz imdb has tv
shows too

df["startYear"] = pd.to_numeric(df["startYear"], errors='coerce')
df["runtimeMinutes"] = pd.to_numeric(df["runtimeMinutes"],
errors='coerce')

df.dropna(subset=['startYear', 'runtimeMinutes', 'genres',
'averageRating'], inplace=True)

df = df[df['runtimeMinutes'] < 300]

df = df[df['numVotes'] > 100]

import matplotlib.pyplot as plt
import seaborn as sns

sns.histplot(data=df, x='averageRating', bins=20,)
plt.title('Distribution of Average Ratings for Movies')
```

```
plt.xlabel('Average Rating')
plt.ylabel('Frequency')
plt.show()
#Most movies are rated between 5 and 7. Very few movies have extremely
high or low ratings.

sns.scatterplot(data=df, x='runtimeMinutes', y='averageRating')
plt.title('Average Rating vs Runtime Minutes for Movies')
plt.xlabel('Runtime Minutes')
plt.ylabel('Average Rating')
plt.show()
# There is no clear correlation between runtime and average rating. Most
movies have a runtime between 60 and 180 minutes, with average ratings
mostly between 5 and 7.

df["genres"] = df["genres"].str.split(",")

df_genre = df.explode('genres')

df.genres

sns.boxplot(data=df_genre, x='genres', y='averageRating')
plt.title('Average Rating by Genre')
plt.xlabel('Genre')
plt.ylabel('Average Rating')
plt.xticks(rotation=90)
plt.show()
# The boxplot shows that genres like "Documentary", "History", and
"Biography" tend to have higher average ratings, while genres like
"Horror" and "Western" tend to have lower average ratings.

plt.figure(figsize=(10,5))
sns.countplot(x='startYear', data=df[df['startYear'] >= 2000])
plt.xticks(rotation=90)
plt.title('Movies Released per Year (2000 onwards)')
plt.show()
# The countplot shows that there has been a steady increase in the number
of movies released each year since 2000, with a significant spike in 2016
and 2017.

df.to_csv("cleaned_imdb_movies.csv", index=False)

"""Conclusion
most movies fall under the range of 5-7 rating
history and biography have a higher rating than other genres
runtime doent seem to have a significant impact on rating
there is a steady increce in movies over the years
"""
```