akhil-eda-week2

December 14, 2025

[215]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

[216]:
```
df = pd.read_csv("/movies.csv")
df.columns
df.shape
```

[216]:
```
(4803, 24)
```

[217]:
```
# df.head()
```

[218]:
```
df.tail()
```

[218]:

	index	budget	genres	homepage	id	keywords	original_language	original_title
4798	4798	220000	Action Crime Thriller	NaN	9367	united states\u2013mexico barrier legs arms pa…	es	
4799	4799	9000	Comedy Romance	NaN	72766	NaN	en	
4800	4800	0	Comedy Drama Romance TV Movie	http://www.hallmarkchannel.com/signedsealeddel…	231617	date love at first sight narration investigati…	en	
4801	4801	0	NaN	http://shanghaicalling.com/	126186	NaN	en	
4802	4802	0	Documentary	NaN	25975	obsession camcorder crush dream girl	en	

```
4798                El Mariachi
4799                  Newlyweds
4800  Signed, Sealed, Delivered
4801           Shanghai Calling
4802          My Date with Drew

                                                 overview  popularity  …  \
4798  El Mariachi just wants to play his guitar and …   14.269792  …
4799  A newlywed couple's honeymoon is upended by th…    0.642552  …
4800  "Signed, Sealed, Delivered" introduces a dedic…    1.444476  …
4801  When ambitious New York attorney Sam is sent t…    0.857008  …
4802  Ever since the second grade when he first saw …    1.929883  …

     runtime                                spoken_languages    status  \
4798    81.0  [{"iso_639_1": "es", "name": "Espa\u00f1ol"}]  Released
4799    85.0                                              []  Released
4800   120.0       [{"iso_639_1": "en", "name": "English"}]  Released
4801    98.0       [{"iso_639_1": "en", "name": "English"}]  Released
4802    90.0       [{"iso_639_1": "en", "name": "English"}]  Released

                                             tagline  \
4798  He didn't come looking for trouble, but troubl…
4799  A newlywed couple's honeymoon is upended by th…
4800                                              NaN
4801                          A New Yorker in Shanghai
4802                                              NaN

                      title vote_average vote_count  \
4798               El Mariachi          6.6        238
4799                 Newlyweds          5.9          5
4800  Signed, Sealed, Delivered          7.0          6
4801           Shanghai Calling          5.7          7
4802          My Date with Drew          6.3         16

                                                 cast  \
4798  Carlos Gallardo Jaime de Hoyos Peter Marquardt…
4799  Edward Burns Kerry Bish\u00e9 Marsha Dietlein …
4800  Eric Mabius Kristin Booth Crystal Lowe Geoff G…
4801  Daniel Henney Eliza Coupe Bill Paxton Alan Ruc…
4802  Drew Barrymore Brian Herzlinger Corey Feldman …

                                                 crew           director
4798  [{'name': 'Robert Rodriguez', 'gender': 0, 'de…  Robert Rodriguez
4799  [{'name': 'Edward Burns', 'gender': 2, 'depart…       Edward Burns
4800  [{'name': 'Carla Hetland', 'gender': 0, 'depar…        Scott Smith
4801  [{'name': 'Daniel Hsia', 'gender': 2, 'departm…        Daniel Hsia
4802  [{'name': 'Clark Peterson', 'gender': 2, 'depa…  Brian Herzlinger
```

```
[5 rows x 24 columns]
```

[219]:
```
df['title']
```

[219]:
```
0                                       Avatar
1     Pirates of the Caribbean: At World's End
2                                      Spectre
3                        The Dark Knight Rises
4                                  John Carter
                          …
4798                               El Mariachi
4799                                 Newlyweds
4800                 Signed, Sealed, Delivered
4801                          Shanghai Calling
4802                         My Date with Drew
Name: title, Length: 4803, dtype: object
```

[220]:
```
df['budget']
```

[220]:
```
0       237000000
1       300000000
2       245000000
3       250000000
4       260000000
          …
4798       220000
4799         9000
4800            0
4801            0
4802            0
Name: budget, Length: 4803, dtype: int64
```

[221]:
```
df['genres']
```

[221]:
```
0       Action Adventure Fantasy Science Fiction
1                         Adventure Fantasy Action
2                           Action Adventure Crime
3                    Action Crime Drama Thriller
4                 Action Adventure Science Fiction
                            …
4798                       Action Crime Thriller
4799                              Comedy Romance
4800                 Comedy Drama Romance TV Movie
4801                                           NaN
4802                                   Documentary
Name: genres, Length: 4803, dtype: object
```

[222]:
```
df.isnull()
```

[222]:

	index	budget	genres	homepage	id	keywords	original_language	\
0	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	
…	…	…	…	…	…	…	…	
4798	False	False	False	True	False	False	False	
4799	False	False	False	True	False	True	False	
4800	False	False	False	False	False	False	False	
4801	False	False	True	False	False	True	False	
4802	False	False	False	True	False	False	False	

	original_title	overview	popularity	…	runtime	spoken_languages	\
0	False	False	False	…	False	False	
1	False	False	False	…	False	False	
2	False	False	False	…	False	False	
3	False	False	False	…	False	False	
4	False	False	False	…	False	False	
…	…	…	…	…	…	…	
4798	False	False	False	…	False	False	
4799	False	False	False	…	False	False	
4800	False	False	False	…	False	False	
4801	False	False	False	…	False	False	
4802	False	False	False	…	False	False	

	status	tagline	title	vote_average	vote_count	cast	crew	director
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
…	…	…	…	…	…	…	…	…
4798	False	False	False	False	False	False	False	False
4799	False	False	False	False	False	False	False	False
4800	False	True	False	False	False	False	False	False
4801	False	False	False	False	False	False	False	False
4802	False	True	False	False	False	False	False	False

[4803 rows x 24 columns]

[223]:
```
df.duplicated()
```

[223]:
```
0       False
1       False
```

```
2       False
3       False
4       False
        …
4798    False
4799    False
4800    False
4801    False
4802    False
Length: 4803, dtype: bool
```

[224]:
```
df['popularity'] = df['popularity'].astype(int)
df['popularity']
```

[224]:
```
0       150
1       139
2       107
3       112
4        43
       …
4798     14
4799      0
4800      1
4801      0
4802      1
Name: popularity, Length: 4803, dtype: int64
```

[225]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4803 entries, 0 to 4802
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   index                 4803 non-null   int64
 1   budget                4803 non-null   int64
 2   genres                4775 non-null   object
 3   homepage              1712 non-null   object
 4   id                    4803 non-null   int64
 5   keywords              4391 non-null   object
 6   original_language     4803 non-null   object
 7   original_title        4803 non-null   object
 8   overview              4800 non-null   object
 9   popularity            4803 non-null   int64
 10  production_companies  4803 non-null   object
 11  production_countries  4803 non-null   object
 12  release_date          4802 non-null   object
```

```
 13  revenue               4803 non-null   int64
 14  runtime               4801 non-null   float64
 15  spoken_languages      4803 non-null   object
 16  status                4803 non-null   object
 17  tagline               3959 non-null   object
 18  title                 4803 non-null   object
 19  vote_average          4803 non-null   float64
 20  vote_count            4803 non-null   int64
 21  cast                  4760 non-null   object
 22  crew                  4803 non-null   object
 23  director              4773 non-null   object
dtypes: float64(2), int64(6), object(16)
memory usage: 900.7+ KB
```

[226]:
```
df.describe()
```

[226]:

	index	budget	id	popularity	revenue	\
count	4803.000000	4.803000e+03	4803.000000	4803.000000	4.803000e+03	
mean	2401.000000	2.904504e+07	57165.484281	21.005205	8.226064e+07	
std	1386.651002	4.072239e+07	88694.614033	31.807675	1.628571e+08	
min	0.000000	0.000000e+00	5.000000	0.000000	0.000000e+00	
25%	1200.500000	7.900000e+05	9014.500000	4.000000	0.000000e+00	
50%	2401.000000	1.500000e+07	14629.000000	12.000000	1.917000e+07	
75%	3601.500000	4.000000e+07	58610.500000	28.000000	9.291719e+07	
max	4802.000000	3.800000e+08	459488.000000	875.000000	2.787965e+09	

	runtime	vote_average	vote_count
count	4801.000000	4803.000000	4803.000000
mean	106.875859	6.092172	690.217989
std	22.611935	1.194612	1234.585891
min	0.000000	0.000000	0.000000
25%	94.000000	5.600000	54.000000
50%	103.000000	6.200000	235.000000
75%	118.000000	6.800000	737.000000
max	338.000000	10.000000	13752.000000

[227]:
```
df['genres'].value_counts()
```

[227]:
```
genres
Drama                                      370
Comedy                                     282
Drama Romance                              164
Comedy Romance                             144
Comedy Drama                               142
                                          ...
Drama Fantasy Horror Mystery Romance         1
Fantasy Family Action                        1
```

```
Thriller Crime Romance                          1
Drama War Romance Western                       1
Adventure Comedy Crime Science Fiction          1
Name: count, Length: 1168, dtype: int64
```

[228]:
```
df['director'].value_counts()
```

[228]:
```
director
Steven Spielberg      27
Woody Allen           21
Clint Eastwood        20
Martin Scorsese       20
Spike Lee             16
                      ..
Bradley Rust Gray      1
Collin Joseph Neal     1
Kirk Loudon            1
Kevin Jordan           1
Malcolm Goodwin        1
Name: count, Length: 2349, dtype: int64
```

[229]:
```
Popular_director = df.groupby("director")["popularity"].mean().
  ↪sort_values(ascending=False).head(5)
print(Popular_director)
```

```
director
Kyle Balda           875.0
Tim Miller           514.0
Colin Trevorrow      221.5
Damien Chazelle      192.0
Christopher Nolan    185.0
Name: popularity, dtype: float64
```

[230]:
```
Leastpopular_director = df.groupby("director")["popularity"].mean().
  ↪sort_values(ascending=False).tail(5)
print(Leastpopular_director)
```

```
director
Jonathan Parker    0.0
Jonathan Meyers    0.0
Deryck Broom       0.0
Robert M. Young    0.0
Stephen Kijak      0.0
Name: popularity, dtype: float64
```

[231]:
```
Popular_genres = df.groupby("genres")["popularity"].mean().
  ↪sort_values(ascending=False).head(10)
```

```
print(Popular_genres)
```

```
genres
Family Animation Adventure Comedy                    256.75
Science Fiction Adventure Thriller                   206.00
Adventure Family Animation Action Comedy             203.00
Science Fiction Action Thriller Adventure            202.00
Adventure Drama Science Fiction                      194.00
Drama Action Crime Thriller                          187.00
Action Thriller Science Fiction Mystery Adventure    167.00
Drama Adventure Science Fiction                      167.00
History Drama Thriller War                           145.00
Science Fiction Action Adventure Fantasy Comedy      143.00
Name: popularity, dtype: float64
```

[232]:
```
Leastpopular_genres = df.groupby("genres")["popularity"].mean().
  ↪sort_values(ascending=False).tail(10)
print(Leastpopular_genres)
```

```
genres
Thriller Comedy Mystery                          0.0
Animation Family Foreign                         0.0
Adventure Drama Foreign                          0.0
Adventure Mystery Thriller                       0.0
Action Crime Drama Romance                       0.0
Thriller Horror Comedy                           0.0
Action Comedy Foreign                            0.0
Action Comedy Drama Western                      0.0
Action Comedy Romance Science Fiction Thriller   0.0
Action Crime Comedy Thriller                     0.0
Name: popularity, dtype: float64
```

[233]:
```
AverageBudgetbygenres = df.groupby("genres")["budget"].mean().
  ↪sort_values(ascending=False).head(10)
print(AverageBudgetbygenres)
```

```
genres
Adventure Fantasy Action Science Fiction     270000000.0
Action Adventure Western                     255000000.0
Thriller Action Adventure Science Fiction    209000000.0
Family Fantasy Adventure                     200000000.0
Action Family Fantasy                        195000000.0
Adventure Family Mystery Science Fiction     190000000.0
Animation Adventure Comedy Family Action     185000000.0
Drama Action Crime Thriller                  185000000.0
Fantasy Adventure Action Family Romance      180000000.0
Science Fiction Fantasy Action Adventure     176000003.0
Name: budget, dtype: float64
```

[234]:
```
Budgetbygenres = df.groupby("genres")["budget"].sum().
  ↪sort_values(ascending=False).head(10)
print(Budgetbygenres)
```

```
genres
Comedy                                5729327015
Drama                                 4804834923
Comedy Romance                        2969946486
Action Adventure Science Fiction      2340805523
Drama Romance                         2167477000
Animation Family                      1924777699
Action Thriller                       1808200000
Comedy Drama Romance                  1760634549
Comedy Drama                          1739814000
Adventure Fantasy Action              1646900000
Name: budget, dtype: int64
```

[235]:
```
Revenuebydirector = df.groupby("director")["revenue"].mean().
  ↪sort_values(ascending=False)
print(Revenuebydirector)
```

```
director
Chris Buck        1.274219e+09
Kyle Balda        1.156731e+09
Lee Unkrich       1.066970e+09
Joss Whedon       9.879437e+08
Chris Renaud      8.759583e+08
                      …
Orson Welles      0.000000e+00
Ossie Davis       0.000000e+00
Panos Cosmatos    0.000000e+00
Paolo Monico      0.000000e+00
Dan Zukovic       0.000000e+00
Name: revenue, Length: 2349, dtype: float64
```

[236]:
```
Revenuebytitle = df.groupby("title")["revenue"].sum().
  ↪sort_values(ascending=False)
print(Revenuebytitle)
```

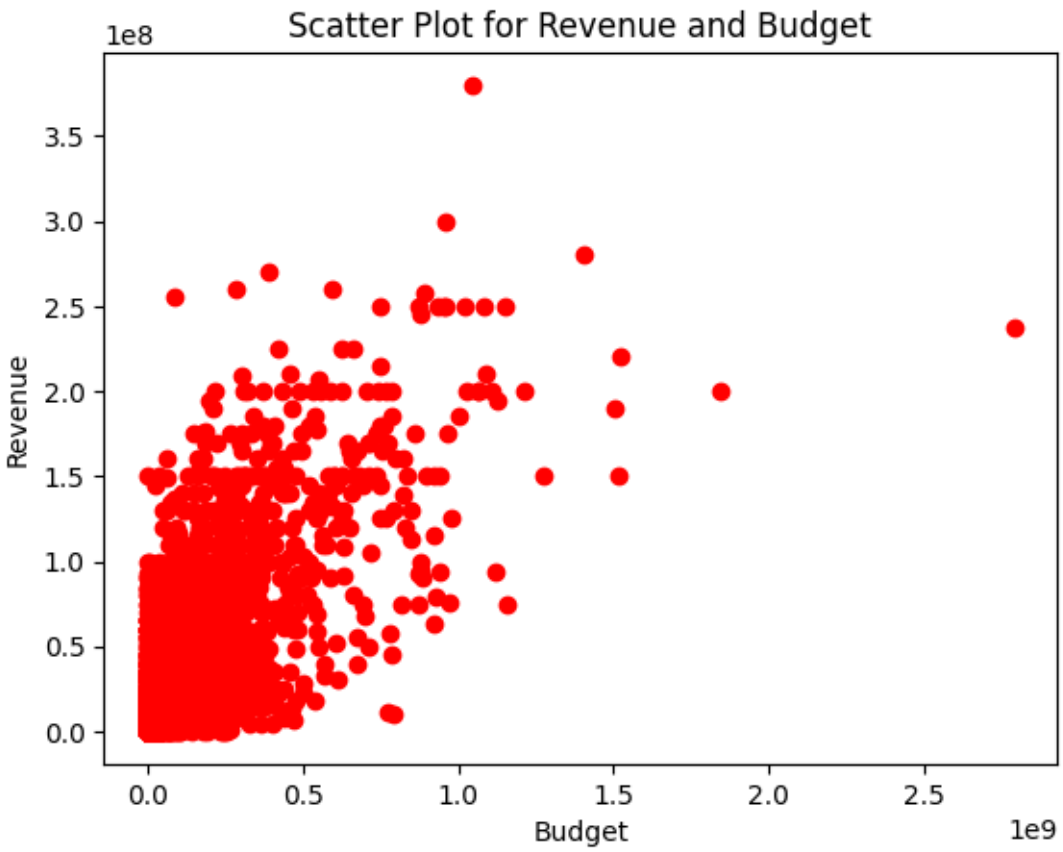
```
title
Avatar                    2787965087
Titanic                   1845034188
The Avengers              1519557910
Jurassic World            1513528810
Furious 7                 1506249360
                                …
The Hudsucker Proxy               0
The Helpers                       0
```

```
The Hills Have Eyes 2             0
The Hit List                      0
Forget Me Not                     0
Name: revenue, Length: 4800, dtype: int64
```

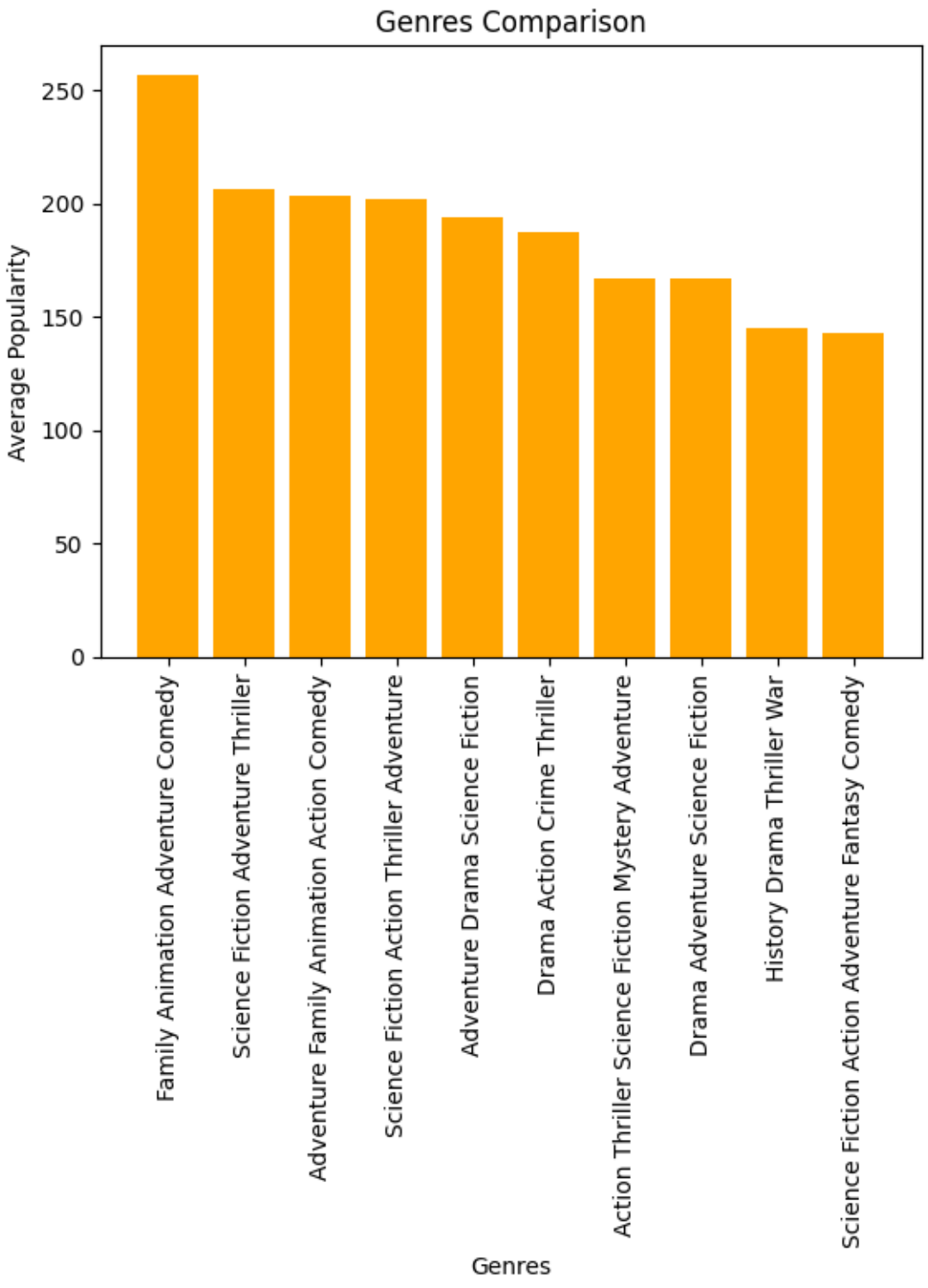
[237]:
```
Budgetbytitle = df.groupby("title")["budget"].sum().sort_values(ascending=False)
print(Budgetbytitle)
```

```
title
Pirates of the Caribbean: On Stranger Tides    380000000
Pirates of the Caribbean: At World's End       300000000
Avengers: Age of Ultron                        280000000
Superman Returns                               270000000
Tangled                                        260000000
                                                   …
Broken Horses                                         0
The Man from Earth                                    0
Slacker                                               0
Sisters in Law                                        0
Mr. Turner                                            0
Name: budget, Length: 4800, dtype: int64
```

[238]:
```
plt.title("Scatter Plot for Revenue and Budget")
plt.scatter(df['revenue'], df['budget'], color='red')
plt.xlabel("Budget")
plt.ylabel("Revenue")
plt.show()
```



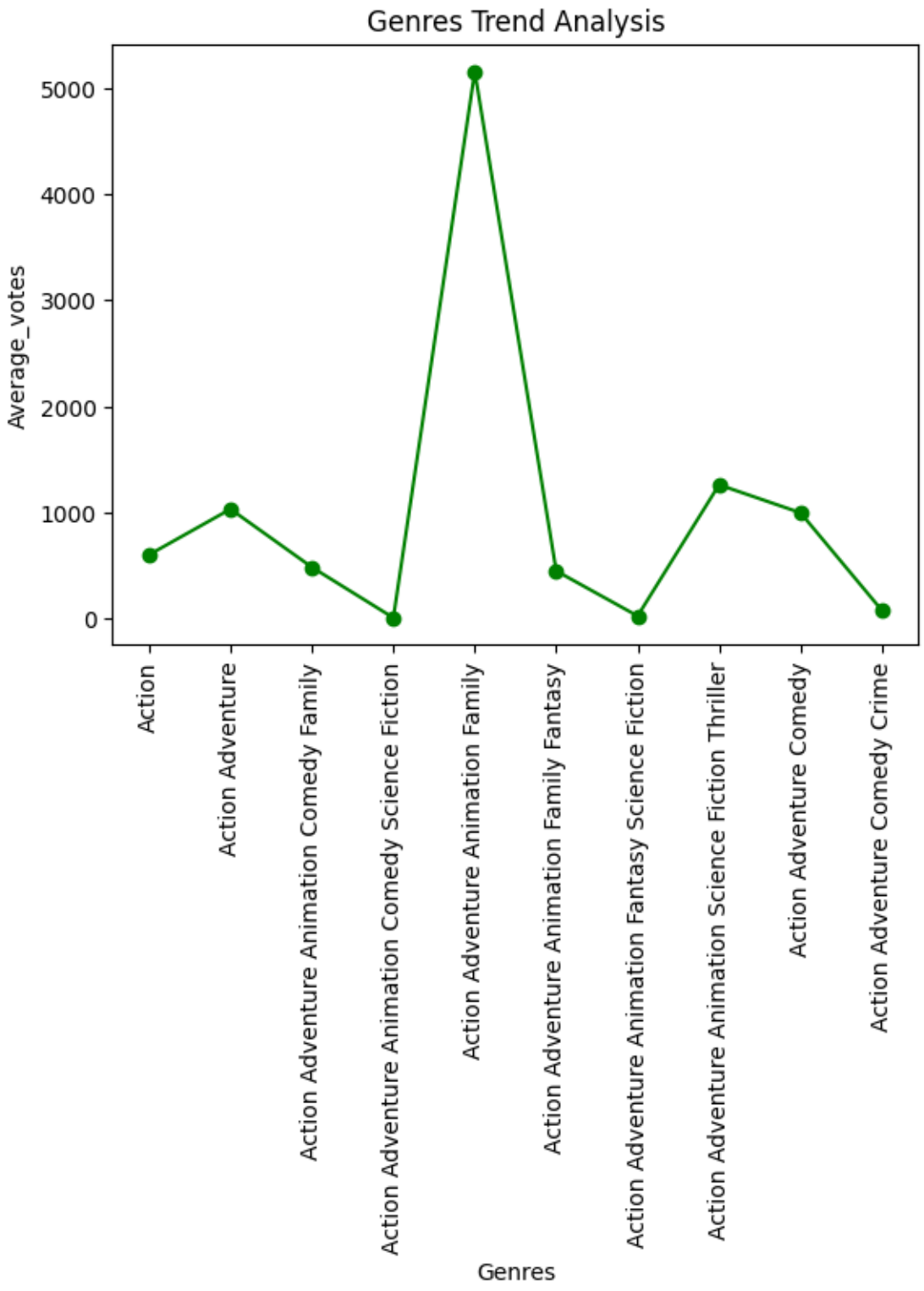
[239]:
```
plt.title("Genres Comparison")
plt.bar(Popular_genres.index, Popular_genres.values, color='orange')
plt.xlabel("Genres")
plt.ylabel("Average Popularity")
plt.xticks(rotation=90)
plt.show()
```



[240]:
```
genre_trend = df.groupby('genres')['vote_count'].mean().head(10)
print(genre_trend)
```

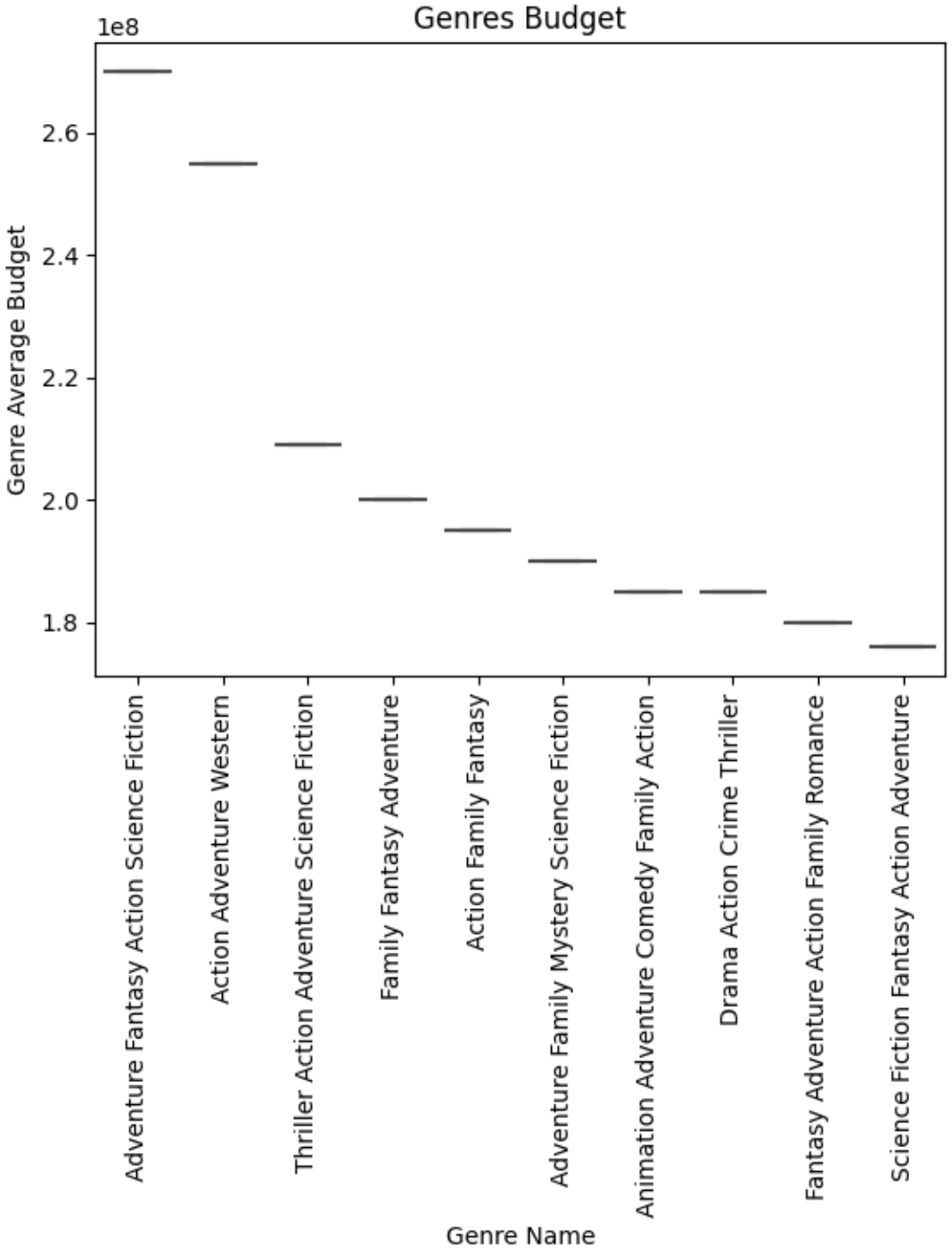
```
genres
Action                                                603.476190
Action Adventure                                     1033.250000
Action Adventure Animation Comedy Family              485.750000
Action Adventure Animation Comedy Science Fiction      10.000000
Action Adventure Animation Family                    5152.000000
Action Adventure Animation Family Fantasy             451.000000
Action Adventure Animation Fantasy Science Fiction     27.000000
Action Adventure Animation Science Fiction Thriller   1262.000000
Action Adventure Comedy                               998.400000
Action Adventure Comedy Crime                          79.333333
Name: vote_count, dtype: float64
```

[241]:
```
plt.title("Genres Trend Analysis")
plt.plot(genre_trend.index,genre_trend.values, marker='o', color='green')
plt.xlabel("Genres")
plt.ylabel("Average_votes")
plt.xticks(rotation=90)
plt.figure(figsize=(16, 12))
plt.show()
```




```
<Figure size 1600x1200 with 0 Axes>
```

[242]:
```
plt.title('Genres Budget')
sns.boxplot(x=AverageBudgetbygenres.index, y=AverageBudgetbygenres.values,␣
 ↪color="#FFD700", linewidth=1.5)
plt.xlabel('Genre Name')
plt.ylabel('Genre Average Budget')
plt.xticks(rotation=90)
plt.show()
```



[243]:
```
df.to_csv("output.csv", index=False)
```

[243]:

[244]:
```
print("EDA Complete Indsight Given Below in Code File")
```

```
EDA Complete Indsight Given Below in Code File
```

INSIGHTS FROM THE ABOVE DATASET

---

1. Director Popularity

The dataset enables the identification of popular directors based on their recent work and audience response. Directors who have consistently produced successful and well-received movies in recent years stand out in terms of popularity and impact within the film industry.

---

2. Genre Analysis

This dataset provides detailed information about different genres in the film industry. Although audiences often watch movies without explicitly recognizing their genre classifications, the analysis highlights a wide range of genres such as Drama, Action, Comedy, Horror, Science Fiction, and others.

---

3. Popular Genres

By analyzing audience votes and trends over recent years, the dataset reveals the most popular genres among viewers. Genres such as Action, Action-Adventure, and Science Fiction have consistently received higher ratings and greater audience engagement, indicating strong viewer preference.

---

4. Movie Budget vs. Revenue

The dataset offers insights into the relationship between movie budgets and box office revenue. The analysis shows that higher budgets do not always guarantee higher revenue. Several low-budget movies achieved high financial success, while some high-budget films underperformed. This suggests that factors such as storyline, plot quality, and audience appeal play a critical role in a movie's success.

---

5. Popular Movies

The analysis identifies several globally popular movies that have achieved long-term audience appreciation and commercial success. Notable examples include *Avatar*, *Titanic*, *The Avengers*, and *Jurassic World*.