Bilenkin550Week2_Exercise_2.2

March 21, 2025

1. Using a data set of your choice, write an introduction explaining the data set.

For this exercise, I selected a Netflix movies dataset that contains 16,000 records from 2010 to 2025. The dataset includes 16 descriptive attributes, such as movie titles, year of release, country, genre, revenue, budget, rating, popularity and vote count. Using this dataset, I aim to analyze trends in Netflix's movie collection and viewer preferences. To achieve this, I will use Matplotlib to create visual representation of key insights.

2. Identify a question or question(s) that you would like to explore in your data set.

Research Questions

Using this dataset, I aim to answer the following questions:

1) Which genre is watched the most?
2) Which movie earned the highest revenue?
3) What country watches the most Netflix movies?
4) Which movie had the highest budget?
5) Which movie received the highest rating?
6) Which movie was the most popular?
7) Which movie had the highest vote count?

[146]:
```
import pandas as pd

# Loading dataset and assigning to a variable for a future use
dataset_file_path = r"C:\Users\maxim\OneDrive\Desktop\BU\DSC␣
 ↪550\netflix_movies_detailed_up_to_2025.csv"
df = pd.read_csv(dataset_file_path)
```

3. Create at least three graphs that help answer these questions. Make sure your graphs are clearly readable and are labeled appropriately and professionally.

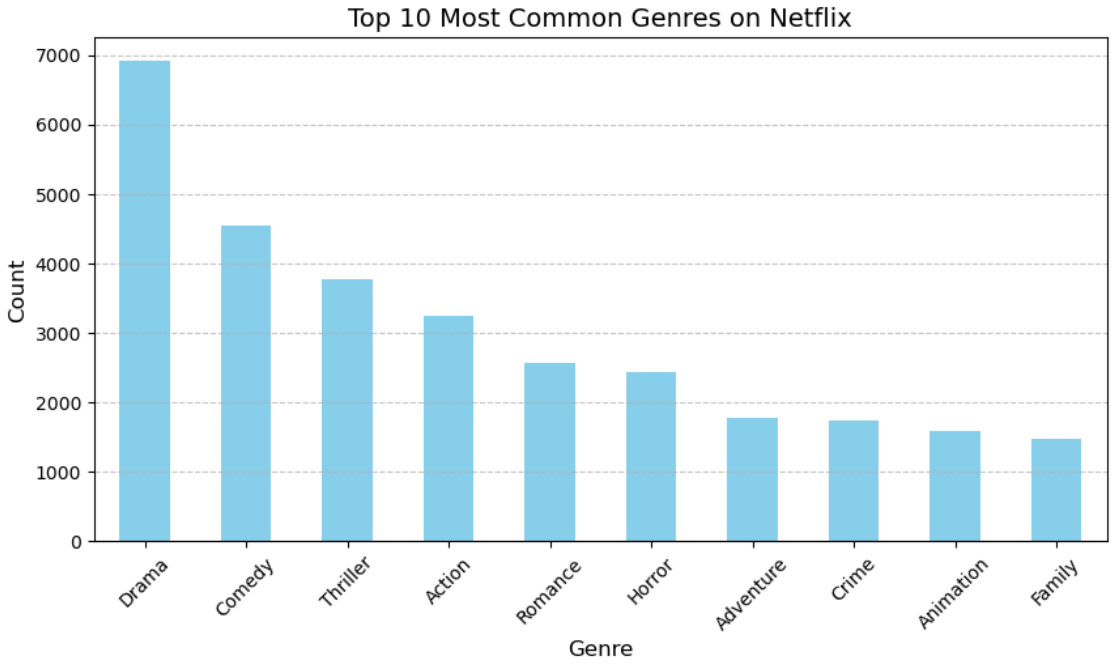
1) Which genre is watched the most?

[147]:
```
import matplotlib.pyplot as plt

# Splitting genres, removing spaces, and counting occurrences
genre_list = df['genres'].dropna().str.split(',').explode().str.strip()  #␣
 ↪Clean whitespace
genre_counts = genre_list.value_counts().head(10)  # Get top 10 genres

# Plotting
```

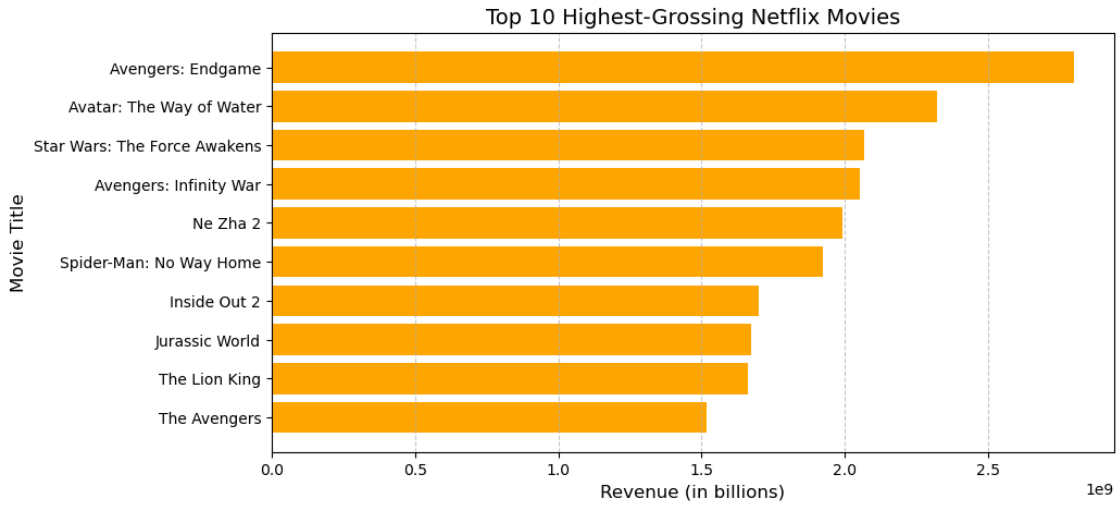
```
plt.figure(figsize=(10, 5))
genre_counts.plot(kind='bar', color='skyblue')
plt.title("Top 10 Most Common Genres on Netflix", fontsize=14)
plt.xlabel("Genre", fontsize=12)
plt.ylabel("Count", fontsize=12)
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



2) Which movie earned the highest revenue?

[148]:
```
# Selecting top 10 movies by revenue
top_revenue_movies = df[['title', 'revenue']].sort_values(by='revenue',␣
 ↪ascending=False).head(10)

# Rerunning your plotting code
plt.figure(figsize=(10, 5))
plt.barh(top_revenue_movies['title'], top_revenue_movies['revenue'],␣
 ↪color='orange')
plt.xlabel("Revenue (in billions)", fontsize=12)
plt.ylabel("Movie Title", fontsize=12)
plt.title("Top 10 Highest-Grossing Netflix Movies", fontsize=14)
plt.gca().invert_yaxis()  # Inverting y-axis to have the highest at the top
plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.show()
```



3) What country watches the most Netflix movies?

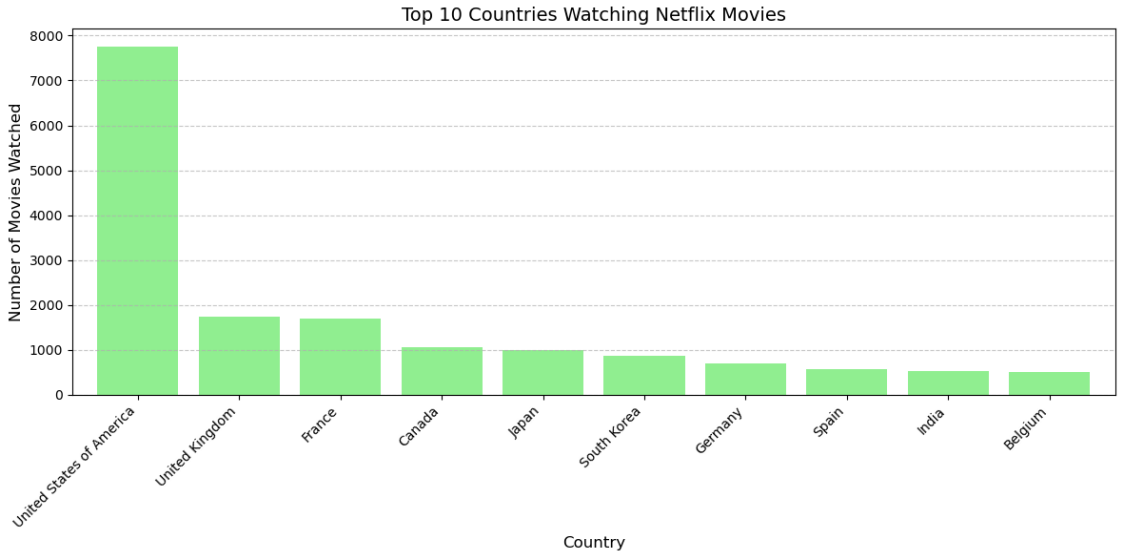
[149]:
```
# Splitting country column where multiple countries are listed
df['country'] = df['country'].str.split(', ')

# Exploding the list so each country gets its own row
df_exploded = df.explode('country')

# Counting movies per country and selecting the top 10
top_countries = df_exploded['country'].value_counts().head(10)

# Rerunning plotting code
plt.figure(figsize=(12, 6))
top_countries.plot(kind='bar', color='lightgreen', width=0.8)  # Adjusting␣
↪width if needed
plt.title("Top 10 Countries Watching Netflix Movies", fontsize=14)
plt.xlabel("Country", fontsize=12)
plt.ylabel("Number of Movies Watched", fontsize=12)
plt.xticks(rotation=45, ha='right')  # Rotating and aligning text to the right
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Ensuring labels fit properly
plt.tight_layout()
plt.show()
```



4) Which movie had the highest budget?

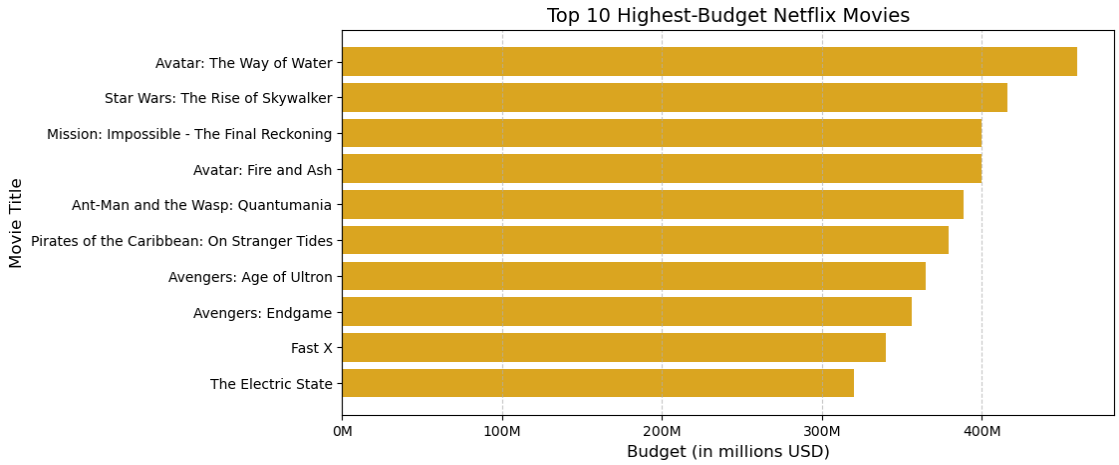
[150]:
```
import matplotlib.ticker as ticker

# Finding the Top 10 movies with the highest budget
top_budget_movies = df.nlargest(10, 'budget')[['title', 'budget']]

# Plotting
plt.figure(figsize=(10, 5))
plt.barh(top_budget_movies['title'], top_budget_movies['budget'],␣
 ↪color='goldenrod')
plt.xlabel("Budget (in millions USD)", fontsize=12)
plt.ylabel("Movie Title", fontsize=12)
plt.title("Top 10 Highest-Budget Netflix Movies", fontsize=14)
plt.gca().invert_yaxis()  # Inverting y-axis so highest budget is on top

# Formating x-axis to display budget in millions
plt.gca().xaxis.set_major_formatter(ticker.FuncFormatter(lambda x, _: f'{x/1e6:.
 ↪0f}M'))

plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.show()
```



5) Which movie received the highest rating?

[151]:
```
# Finding the highest rating
highest_rating = df['rating'].max()

# Filtering movies that have the highest rating
highest_rated_movies = df[df['rating'] == highest_rating][['title', 'rating']]

# Converting rating to an integer (removing decimals)
highest_rated_movies['rating'] = highest_rated_movies['rating'].astype(int)

# Printing the DataFrame to ensure it appears in PDF output
print(highest_rated_movies.to_string(index=False))  # Removing index for␣
 ↪cleaner output

# Applying styling for Jupyter Notebook (won't affect PDF)
styled_df = highest_rated_movies.style.set_properties(**{'text-align':␣
 ↪'center'})
styled_df.set_table_styles([{
    'selector': 'th',
    'props': [('text-align', 'center')]
}])
```

title	rating
Inácio Garapa, Um Matuto Sonhador	10
The Photographer	10
An Unholy Affair: A Younger Man and a Busty Wife	10
It	10
Swapping Guest House	10
Youthful Mother-in-law	10
Nice Sister-In-Law	10
Three Sisters Swapping	10

Actresses: Sex Audition	10
The Shepherd	10
Sincheon Station Exit 3	10
Secret Night Of Mother And Daughter	10
Dangerous Younger Cousin	10
Underpants Thief	10
Family Matters	10
TOGEFILM - Mei Mei	10
Salome	10
El Apocalipsis de san Juan	10
Jailbreak Affair	10
The Williams	10
Balota	10
Marco	10
Queen of the Ring	10
Dragon	10
The American Backyard	10
Mere Husband Ki Biwi	10
The Kite	10
Close To Me	10
Red Silk	10
Hiram na Sandali	10
Lore Of The Ring Light	10
The Crucifix	10
Butterfly	10
F1 75 Live at The O2	10
The Wrong Obsession	10

[151]: <pandas.io.formats.style.Styler at 0x212f1a39af0>

6) Which movie was the most popular?

[152]:
```
# Finding the highest popularity score
highest_popularity = df['popularity'].max()

# Filtering movies that have the highest popularity score
most_popular_movie = df[df['popularity'] == highest_popularity][['title',␣
 ↪'popularity']]

# Formatting popularity to remove decimals and add commas
most_popular_movie['popularity'] = most_popular_movie['popularity'].astype(int).
 ↪apply(lambda x: f"{x:,}")

# Printing the DataFrame (ensuring it appears in PDF)
print(most_popular_movie.to_string(index=False))  # Removing index for cleaner␣
 ↪output
```

```
# Optional: Displaying styled DataFrame (only for Jupyter Notebook, won't␣
 ↪affect PDF)
styled_df = most_popular_movie.style.set_properties(**{'text-align': 'center'})
styled_df.set_table_styles([{
    'selector': 'th',
    'props': [('text-align', 'center')]
}])

# Displaying for Jupyter Notebook
styled_df
```

```
    title popularity
The Gorge      3,876
```

[152]: <pandas.io.formats.style.Styler at 0x212f863b140>

7) Which movie had the highest vote count?

[153]:
```
# Finding the highest vote count
highest_vote_count = df['vote_count'].max()

# Filtering movies that have the highest vote count
most_voted_movie = df[df['vote_count'] == highest_vote_count][['title',␣
 ↪'vote_count']]

# Formatting vote count with comma separator
most_voted_movie['vote_count'] = most_voted_movie['vote_count'].apply(lambda x:␣
 ↪f"{x:,}")

# Displaying the result
print("Most Voted Movie(s):")
print(most_voted_movie.to_string(index=False))  # Converting to string format␣
 ↪for proper PDF export
```

```
Most Voted Movie(s):
    title vote_count
Inception     37,119
```

4. Explain what you have learned from each of your graphs.

The Power of Visualization in Data Analysis

It is amazing how much insight can be gained through data visualization. By generating graphs, I have uncovered surprising facts that challenge common assumptions.

For example, I initially believed that action movies would be the most-watched genre on Netflix. However, the data revealed that drama is the most popular genre, with almost 7,000 different drama movies watched by customers. This demonstrates how personal beliefs can be misleading and how big data helps uncover objective truths.

Another common misconception I had was that Avatar: The Way of Water earned the highest

revenue in movie history, including streaming revenue. Many people believe this to be true, but the data shows otherwise. In reality, Avengers: Endgame holds the record for the highest earnings at approximately $2.8 billion, compared to $2.3 billion for Avatar: The Way of Water.

It is not surprising that audiences in the United States watch the most Netflix movies, given that Netflix originated in the U.S. This aligns with expectations, as local audiences tend to drive the success of homegrown platforms.

One surprising fact was that Avatar: The Way of Water had the highest production budget at approximately $460 million. I always thought Titanic had the highest budget, but it did not even make it into the top 10 highest-budget movies.

I also discovered that 35 movies received the highest rating on Netflix. Some of these highly rated films include The Shepherd, The Williams, Queen of the Ring, Butterfly, and many more.

Surprisingly, The Gorge was the most popular movie on Netflix, with an engagement score of 3,876.

Finally, Inception had the highest vote count at 37,119 votes. This high number suggests that the rating for Inception is likely highly reliable, as a larger sample size generally leads to more accurate and representative feedback.

5. Write a conclusion that summarizes your findings.

To summarize all the findings, I generated graphs to visualize the data. Using the Netflix dataset, I uncovered several insights that challenged common assumptions. Drama emerged as the most-watched genre, substantially surpassing action, which I initially thought would be the leader. This underscores how access to big data can correct personal biases.

I discovered that Avengers: Endgame earned the highest box office revenue at approximately $2.8 billion, contradicting to the common belief that Avatar: The Way of Water was the highest-grossing film. Additionally, while Avatar: The Way of Water had the highest production budget of around $460 million, Titanic – which many assume to be the most expensive film-did not even make the top 10 list.

The data also revealed that the United States has the largest Netflix audience, which was not surprising to me. This aligns with my expectations, given that Netflix originated in the U.S. The most popular movie in terms of engagement was The Gorge, with a popularity score of 3,876. Furthermore, Inception had the highest vote count, with 37,119 people rating the movie, making its rating highly reliable due to the large sample size.

Overall, this analysis demonstrates the power of data visualization in uncovering trends, correcting misconceptions, and providing objective insights. What we assume to be true is not always supported by data. This highlights the importance of data-driven decision-making in understanding consumer behavior and industry trends.