Name: Atharva Shinde Div: D15C Roll No:54

EXP 2

Aim:

Data Visualization/ Exploratory data Analysis using Matplotlib and Seaborn.

1. Create bar graph, contingency table using any 2 features.
2. Plot Scatter plot, box plot, Heatmap using seaborn.
3. Create histogram and normalized Histogram.
4. Describe what this graph and table indicates.
5. Handle outlier using box plot and Inter quartile range.

Introduction:

Exploratory Data Analysis (EDA), introduced by John Tukey in the 1970s, is the first step in analyzing datasets to summarize their key characteristics using statistical and visual techniques. It helps understand data patterns, detect anomalies, and prepare the data for machine learning models.

Why Perform EDA?

EDA is essential for:

- Identifying key features and trends in the data.
- Detecting correlations between variables.
- Assessing data quality and handling missing values.
- Determining the need for data preprocessing.
- Communicating insights effectively using visual tools.

Common EDA Techniques:

- Histograms and frequency distributions to analyze data distribution.
- Box plots to identify outliers and data spread.
- Scatter plots to observe relationships between variables.
- Heatmaps to visualize correlations between features.
- Bar charts and pie charts for categorical data analysis

Name: Atharva Shinde Div: D15C Roll No:54

Data visualization is crucial for analyzing the top 10,000 movies dataset because it helps in the following ways:

1. **Identifying Trends and Patterns:**
    - Visualizations like line charts and histograms can reveal trends over time, such as the rise or fall in movie production, changing genre popularity, or variations in box office revenues.
2. **Comparative Analysis:**
    - Bar charts and box plots enable comparisons between different categories like genres, production studios, or directors, helping identify which ones are more successful or popular.
3. **Correlation Analysis:**
    - Scatter plots help examine relationships between variables, such as budget vs. revenue or rating vs. box office performance.
4. **Distribution Insights:**
    - Histograms and density plots illustrate the distribution of continuous variables like movie ratings, durations, or revenues, showing if they are normally distributed or skewed.
5. **Outlier Detection:**
    - Box plots or scatter plots can highlight outliers, such as unusually high-grossing movies or extremely low-rated ones, which could be interesting for further investigation.
6. **Audience Preferences and Trends:**
    - Word clouds or bar charts showing the frequency of keywords in movie titles or genres can reveal audience interests and emerging themes.
7. **Data Storytelling:**
    - Interactive dashboards make the data more engaging, helping stakeholders or audiences understand complex information easily.

Name: Atharva Shinde Div: D15C Roll No:54

1) Bar Graph (genere vs vote Count)

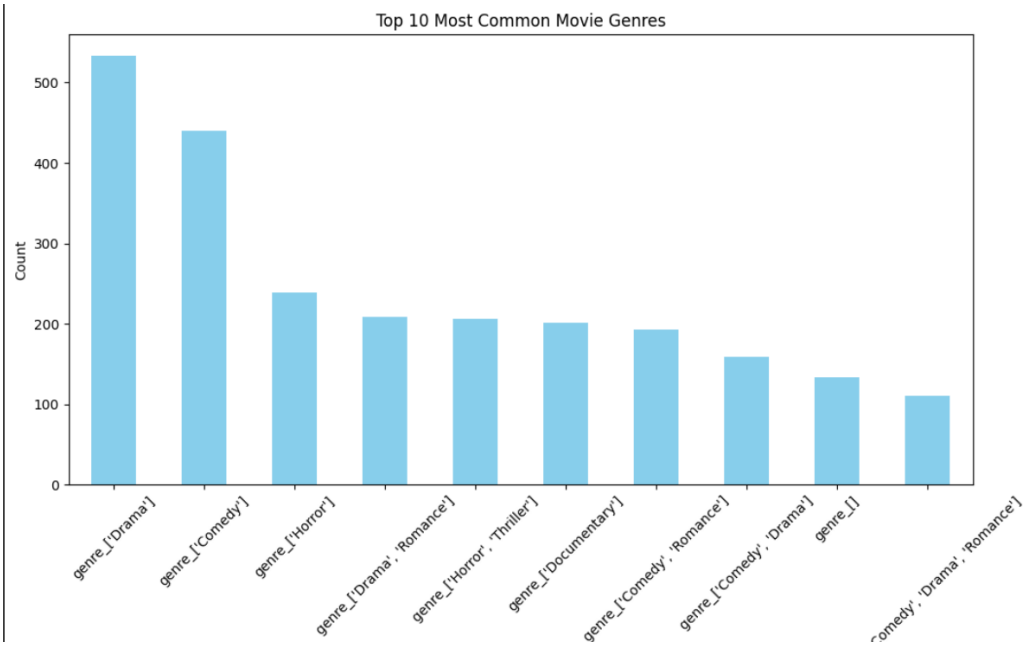
Inference:

1. Drama and Comedy are the most popular genres, showing high audience interest in storytelling and humor.

2. Hybrid genres like Drama-Romance and Horror-Thriller are also common, reflecting a demand for blended narratives.

3. Empty genre entries indicate possible data inconsistencies that could be cleaned for better analysis.

```
import matplotlib.pyplot as plt

genre_columns = [col for col in df.columns if col.startswith("genre_")]

genre_counts = df[genre_columns].sum().sort_values(ascending=False)

plt.figure(figsize=(12, 6))
genre_counts.head(10).plot(kind="bar", color="skyblue")

plt.title("Top 10 Most Common Movie Genres")
plt.xlabel("Genre")
plt.ylabel("Count")
plt.xticks(rotation=45)
plt.show()
```



2)Inference: Box Plot

1. Highly Skewed Distribution:

- The revenue data is highly skewed to the right, with the majority of movies earning relatively low revenue. This is indicated by the box being compressed near the lower end of the axis.
- A small number of movies have extremely high revenues, as shown by the large number of outliers. These outliers represent blockbuster hits that earned significantly more than the average movie.

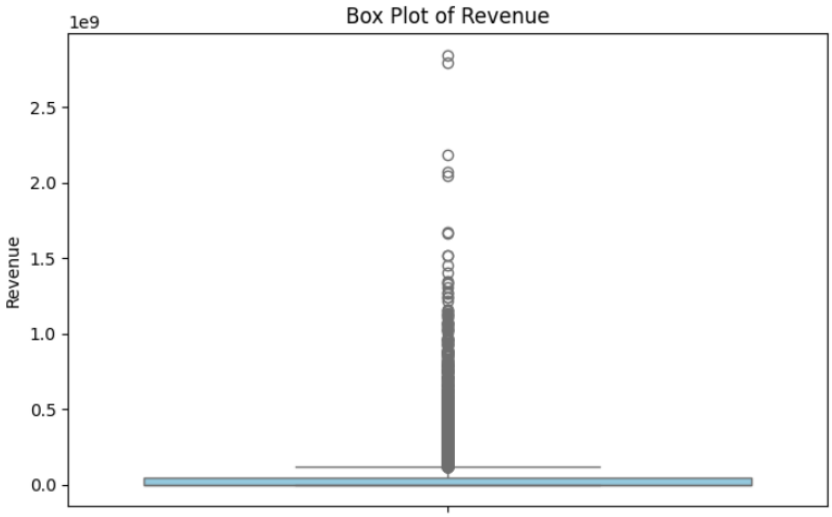
2. Outliers and Blockbusters:

- Numerous outliers are present above the upper whisker, extending beyond $1 billion, and a few even surpass $2.5 billion. These represent major box office successes, likely including franchises or highly anticipated releases.
- The concentration of outliers suggests that while most movies earn modest amounts, a few exceptional hits dominate total box office revenue.

3. Central Tendency and Spread:

- The interquartile range (IQR) is narrow and positioned near the lower end, highlighting that 50% of the movies have relatively low earnings.
- The median (central line within the box) is close to the bottom, confirming that half of the movies earn less than the average revenue, reinforcing the right-skewed distribution.

```
plt.figure(figsize=(8, 5))
sns.boxplot(y=df["revenue"], color="skyblue")

plt.title("Box Plot of Revenue")
plt.ylabel("Revenue")
plt.show()
```



5) Heatmap of Numerical Features Correlation

Purpose:

This heatmap visually represents the correlation between numerical features in the movies dataset. The values range from -1 to 1, where:

- +1 → Strong positive correlation (when one factor increases, the other also increases).
- 0 → No correlation (factors do not impact each other).
- -1 → Strong negative correlation (when one factor increases, the other decreases).

Key Observations from the movies Dataset:

Strong Positive Correlation:

- **Revenue** and **Vote Count** have a high positive correlation (0.77), indicating that movies with more votes typically generate higher revenue. This suggests popular movies (in terms of audience engagement) tend to be more profitable.
- **Revenue** also strongly correlates with its normalized and standardized versions, as expected.

Moderate Positive Correlation:

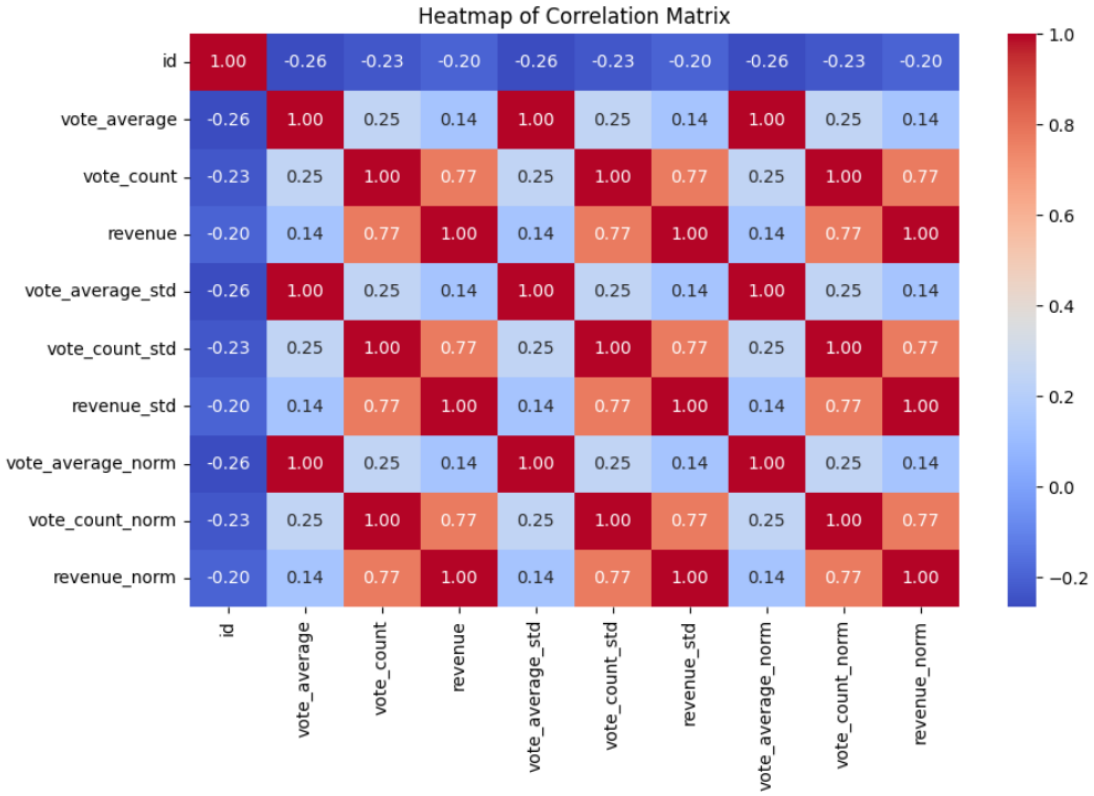
- **Vote Count** and **Vote Average** show a moderate positive correlation (0.25), implying that higher audience engagement slightly influences better ratings.
- **Revenue** and **Vote Average** have a low positive correlation (0.14), suggesting that higher-rated movies might earn more, but the relationship is not very strong.

Low or Negative Correlation:

- **ID** shows negative correlations with most variables, which is normal since it's just an identifier and has no direct relationship with movie attributes.
- **Vote Average** has a weak correlation with **Revenue** and **Vote Count**, indicating that high ratings don't necessarily translate to high earnings or more votes

Name: Atharva Shinde Div: D15C Roll No:54

```
import numpy as np

numeric_columns = df.select_dtypes(include=np.number)

plt.figure(figsize=(10, 6))
sns.heatmap(numeric_columns.corr(), annot=True, cmap="coolwarm", fmt=".2f")

plt.title("Heatmap of Correlation Matrix")
plt.show()
```



6) Histogram

Inference (From Histogram)

Highly Skewed Distribution:

- The histogram shows a **right-skewed distribution**, indicating that most movies earn relatively low revenue, while a few outliers make extremely high amounts.

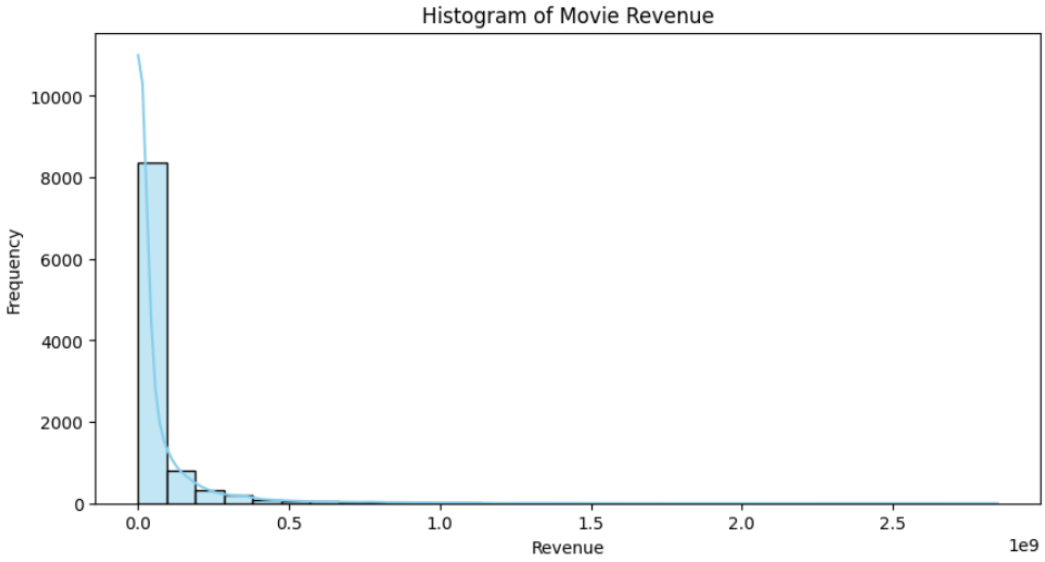
Majority in Low Revenue Range:

- A significant majority of movies have revenues clustered near zero, suggesting that high-grossing films are rare compared to lower-earning ones.

Presence of Outliers:

- A long tail extends towards the right, showing the presence of blockbuster movies with revenues exceeding **$1 billion**, which significantly skew the distribution.

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 5))
sns.histplot(df["revenue"], bins=30, kde=True, color="skyblue")

plt.title("Histogram of Movie Revenue")
plt.xlabel("Revenue")
plt.ylabel("Frequency")
plt.show()
```



7) Normalized Histogram

Right-Skewed Distribution Remains:

- Even after normalization, the distribution is highly right-skewed, indicating that most movies have relatively low normalized revenue, with a few outliers achieving significantly higher values.

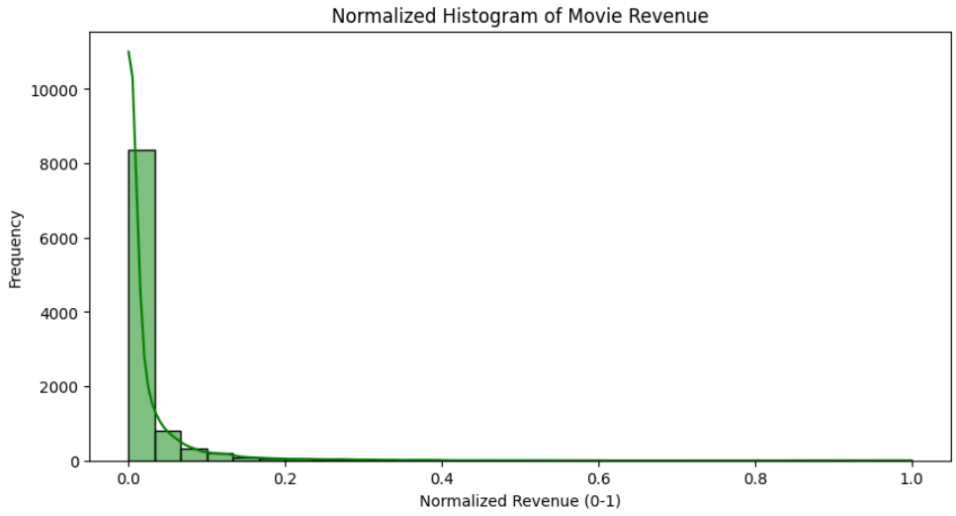
Concentration Near Zero:

- The majority of movies are concentrated near the lower end of the normalized scale (close to 0), showing that most films earn a small fraction of the maximum revenue observed.

Outliers Still Present:

- A long tail persists towards the right, suggesting that a few movies generate disproportionately high revenue even when scaled between 0 and 1.

```
plt.figure(figsize=(10, 5))
sns.histplot(df["revenue_norm"], bins=30, kde=True, color="green")

plt.title("Normalized Histogram of Movie Revenue")
plt.xlabel("Normalized Revenue (0-1)")
plt.ylabel("Frequency")
plt.show()
```

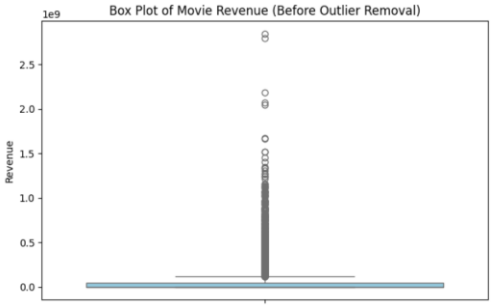


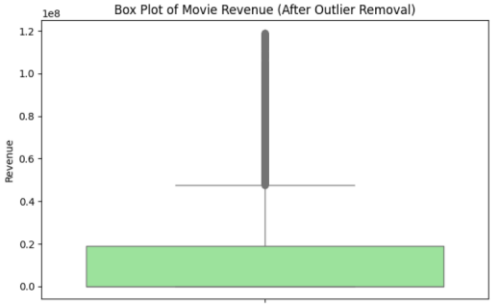
Name: Atharva Shinde Div: D15C Roll No:54

8) **Handle outlier using box plot**

Inference: Box Plot

```
# Compute Q1 (25th percentile) and Q3 (75th percentile)
Q1 = df["revenue"].quantile(0.25)
Q3 = df["revenue"].quantile(0.75)

# Compute Interquartile Range (IQR)
IQR = Q3 - Q1

# Define lower and upper bounds for outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Remove outliers
df_cleaned = df[(df["revenue"] >= lower_bound) & (df["revenue"] <= upper_bound)]

print(f"Original dataset size: {df.shape[0]}")
print(f"Dataset size after outlier removal: {df_cleaned.shape[0]}")
```





Conclusion:

Hence we learned about exploratory data analysis and various types of statistical measures of data along with correlation. We also learnt about visualization and applied these concepts with hands-on experience on our chosen dataset.