Name : Bhoomi Mangesh Naik
Class : D15C
Roll No. : 60
Practical No. : 3

Aim : To perform Exploratory Data Analysis and visualization using python

Introduction:

Exploratory Data Analysis (EDA) is an essential step in the data analysis process that involves summarizing and visualizing the dataset to understand its underlying structure, detect patterns, identify anomalies, and generate insights. It helps in making informed decisions about data preprocessing, feature engineering, and model selection in later stages of analysis. EDA consists of descriptive statistics, data visualization, and correlation analysis, which provide a comprehensive understanding of the dataset. Since our dataset consists of the top 2000 global leading companies, EDA will help us uncover trends related to company revenue, profit, market value, industry distribution, and geographical insights.

Descriptive analysis - Central tendency

Definition:

Central tendency refers to the measure that represents the center or typical value of a dataset. The three common measures of central tendency are:

- Mean (Average) – The sum of all values divided by the number of observations.
- Median – The middle value when the dataset is sorted.
- Mode – The most frequently occurring value.

Execution:

- For our dataset, we can calculate the mean, median, and mode for numerical variables such as:
- Revenue (Total revenue of a company)
- Profit (Total profit of a company)
- Market Value (Market capitalization)
- Assets (Total assets owned by a company)

By analyzing these measures, we can determine whether the data is normally distributed or skewed.

Inference:

A high mean revenue compared to the median suggests the presence of outliers (large companies dominating the dataset).

If the median profit is lower than the mean, it indicates a right-skewed distribution where a few companies have exceptionally high profits.

Descriptive analysis - Dispersion

Definition:

Dispersion measures how spread out the data is. The key dispersion measures include:

- Range – The difference between the maximum and minimum values.
- Variance – The average squared deviation from the mean.
- Standard Deviation – The square root of variance, indicating how much data deviates from the mean.
- Interquartile Range (IQR) – The difference between the 75th percentile (Q3) and 25th percentile (Q1), highlighting the spread of the middle 50% of the data.

Execution:

We compute dispersion metrics for variables such as revenue, profit, and market value to observe variability among companies.

Inference:

A high standard deviation in market value suggests significant variations between large corporations and smaller companies. Outliers in revenue and profit can indicate the dominance of multinational corporations.

Correlation

Definition:

Correlation measures the relationship between two numerical variables. It indicates how changes in one attribute affect another. The common correlation metrics are:

- Pearson Correlation (measures linear relationship, ranges from -1 to +1)
- Spearman Correlation (measures monotonic relationships)

Execution:

We calculate correlation between:

- Revenue and Profit (to check if higher revenue leads to higher profit)
- Market Value and Assets (to see if companies with more assets have a higher market valuation)

Inference:

A high correlation between revenue and profit indicates that larger companies tend to be more profitable. A weak correlation between market value and assets suggests that brand reputation and stock performance may also influence market capitalization.

Data Visualization

1. Histogram (Distribution of Revenue)

A histogram is used to visualize the frequency distribution of numerical data.

Inference:

If the histogram of revenue is right-skewed, it indicates that most companies generate moderate revenue while a few generate exceptionally high revenue.

2. Box Plot (Profit by Industry)

A box plot helps in detecting outliers and median values across different industries.

Inference:

If the box plot of profit by industry shows long whiskers, it suggests high profit variability within

industries like technology and finance.

3. Scatter Plot (Revenue vs. Profit)

A scatter plot helps visualize the relationship between two numerical variables.

Inference:

A positive linear trend between revenue and profit indicates that higher revenue generally leads to higher profit.

Outliers might represent companies with high revenue but low profit margins (e.g., heavy R&D investments).

4. Bar Chart (Top 10 Industries by Market Value)

A bar chart is useful for comparing categorical variables like industries.

Inference:

If the technology industry dominates the chart, it suggests a major shift toward digital businesses. The financial and healthcare sectors might also have a significant presence.

5. Heatmap (Correlation Matrix)

A heatmap visualizes the correlation between multiple variables.

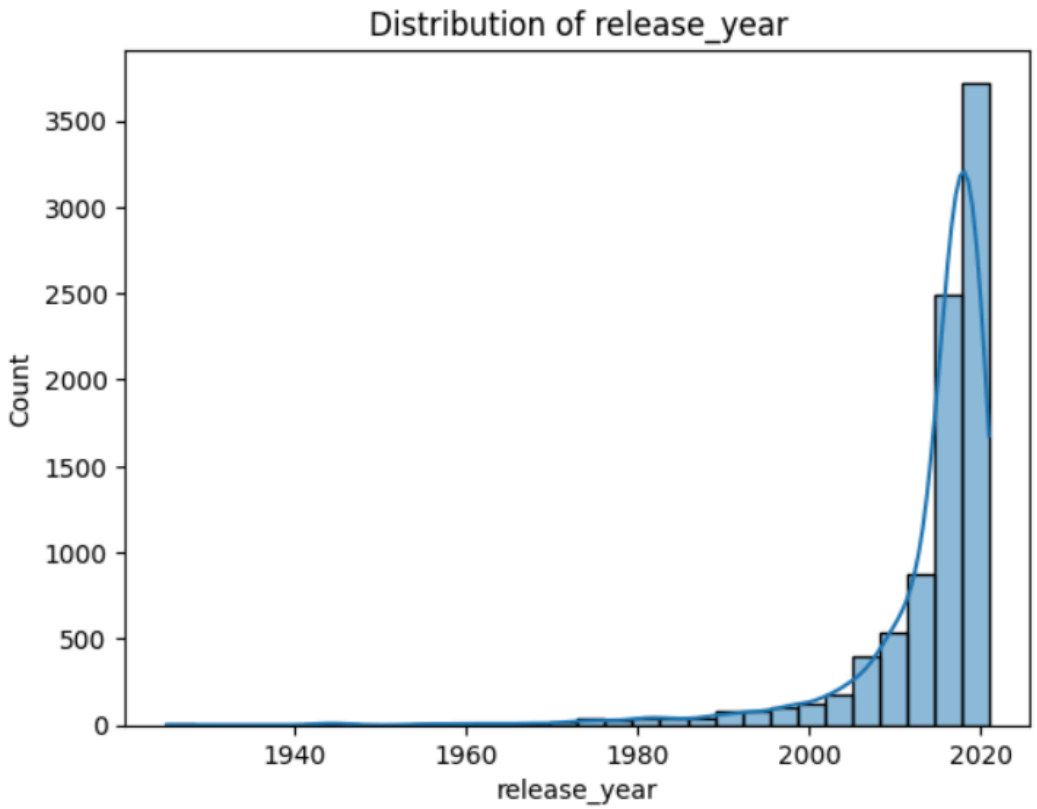
Inference:

If market value is highly correlated with revenue and profit, it indicates that financial strength influences a company's valuation. Weak correlation between assets and profit suggests that some companies generate significant profit despite fewer assets.

Code and Output :

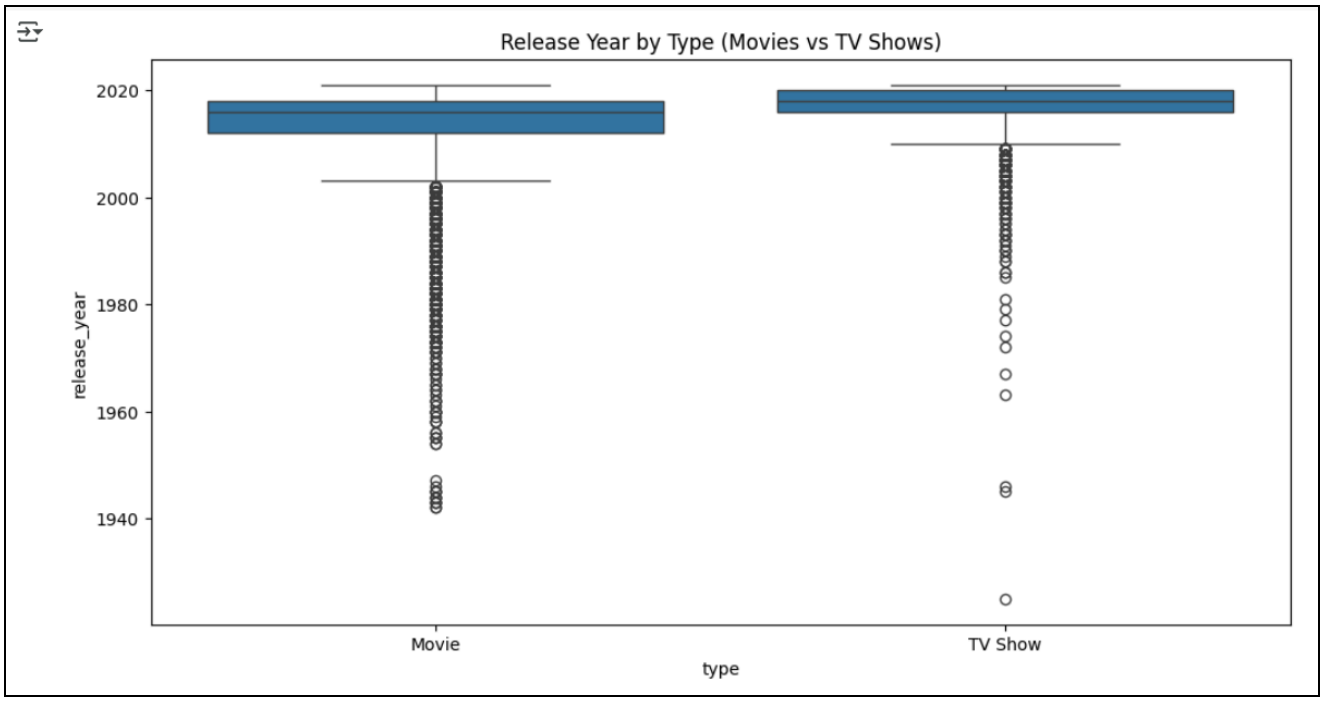
1. Histogram

```
sns.histplot(df["release_year"], bins=30, kde=True)
plt.title("Distribution of release_year")
plt.show()
```



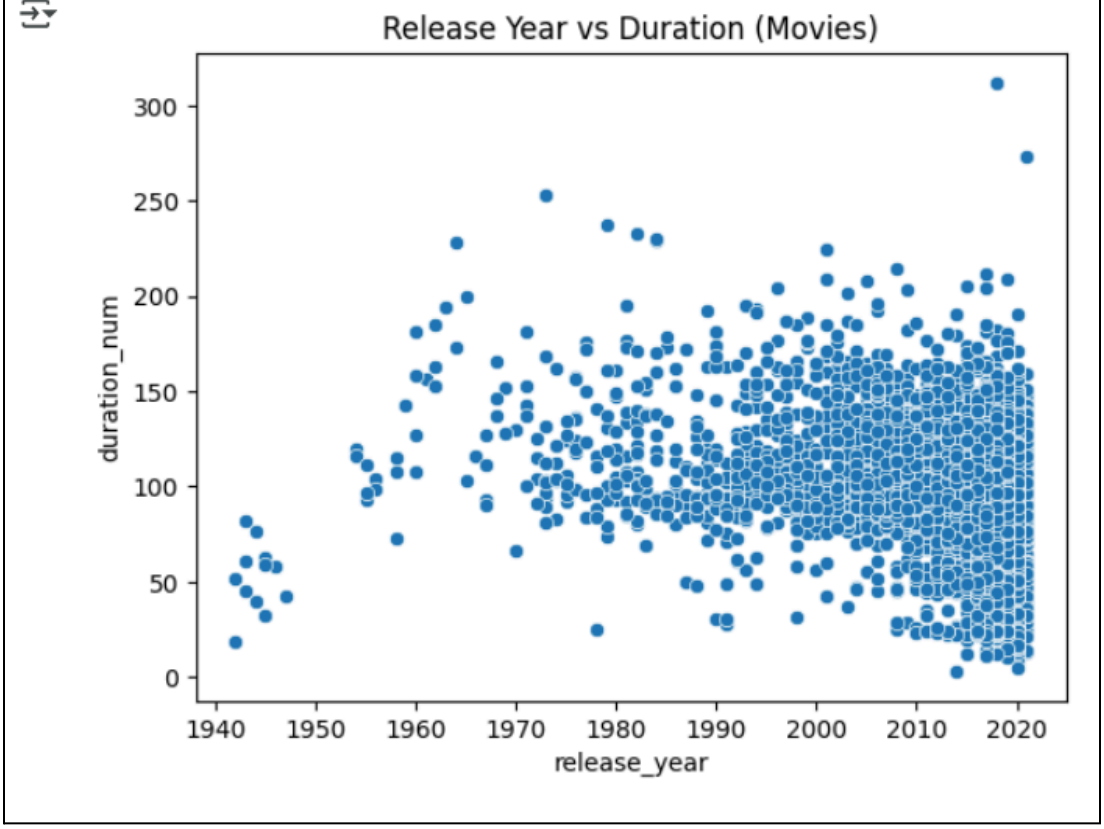
2. Box plot

```
plt.figure(figsize=(12,6))
sns.boxplot(x="type", y="release_year", data=df)   # Movie vs TV Show
plt.title("Release Year by Type (Movies vs TV Shows)")
plt.show()
```



3. Scatter plot

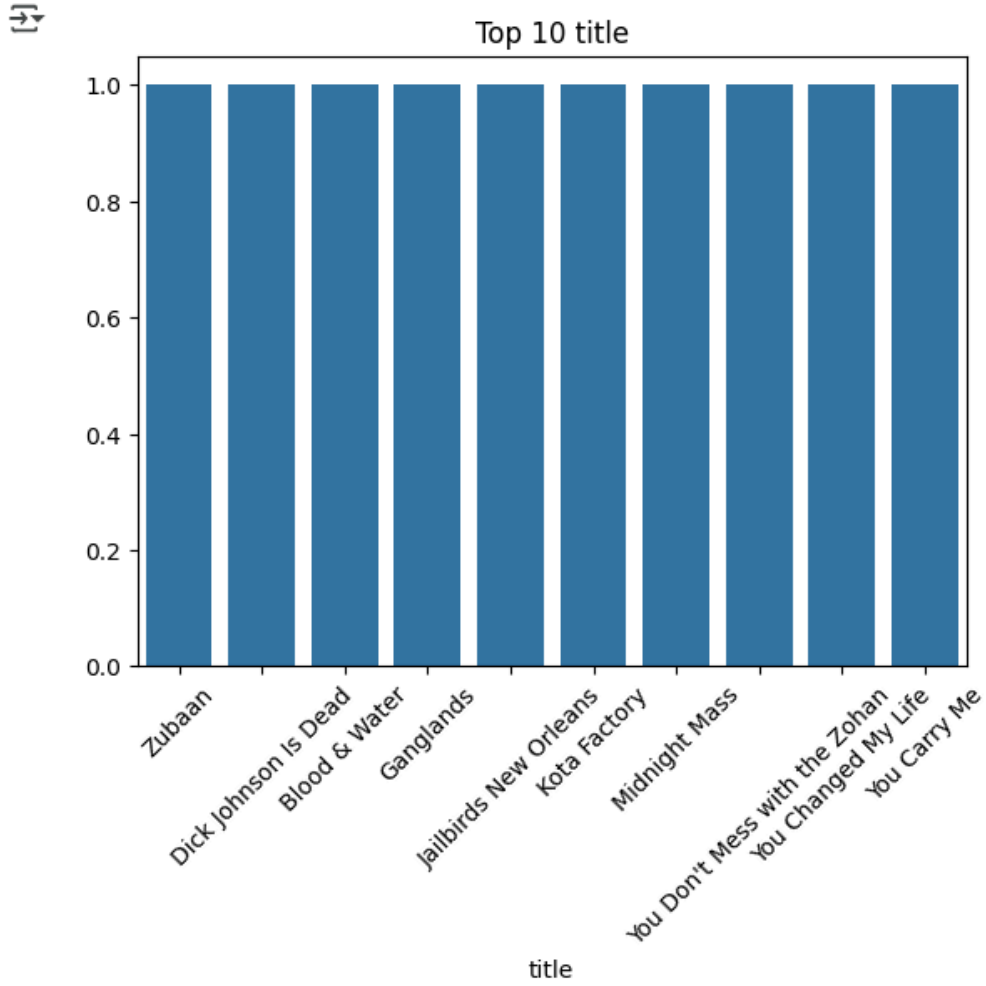
```
sns.scatterplot(x="release_year", y="duration_num", data=df_movies)
plt.title("Release Year vs Duration (Movies)")
plt.show()
```



4. Bar chart

```
[ ]  top_categories = df["title"].value_counts().head(10)

     sns.barplot(x=top_categories.index, y=top_categories.values)
     plt.xticks(rotation=45)
     plt.title("Top 10 title")
     plt.show()
```



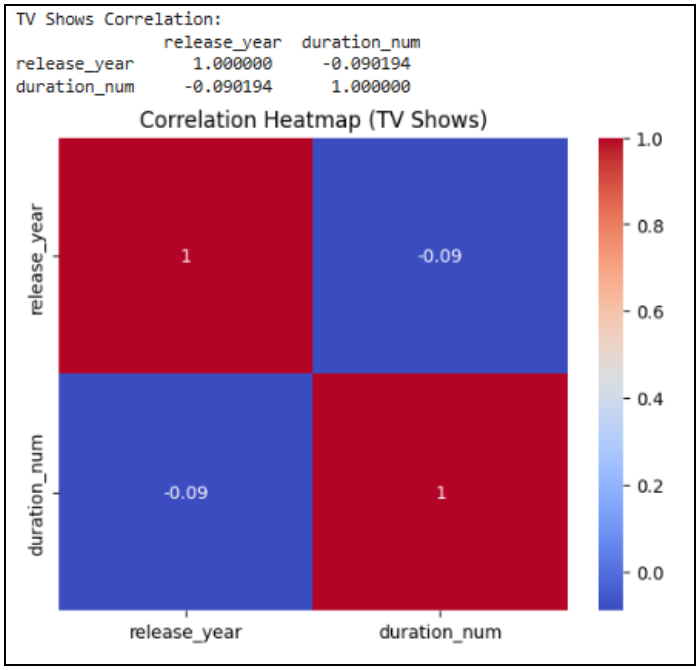
5. Heatmap

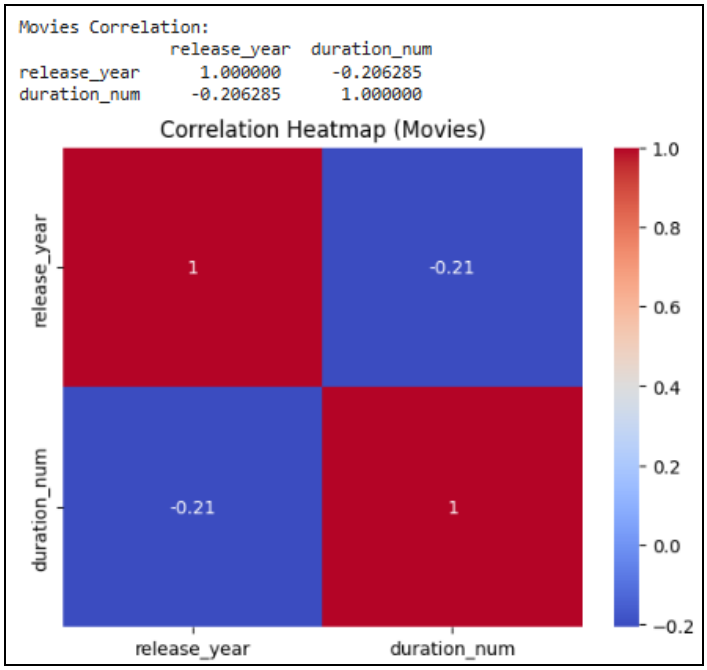
```
print("Movies Correlation:\n", df_movies[["release_year","duration_num"]].corr())

sns.heatmap(df_movies[["release_year","duration_num"]].corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap (Movies)")
plt.show()

print("TV Shows Correlation:\n", df_shows[["release_year","duration_num"]].corr())

sns.heatmap(df_shows[["release_year","duration_num"]].corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap (TV Shows)")
plt.show()
```





Conclusion :

Exploratory Data Analysis (EDA) and visualization using Python provide valuable insights into the structure, quality, and patterns within a dataset. By leveraging libraries such as Pandas, NumPy, Matplotlib, and Seaborn, we can clean and preprocess the data, identify trends, correlations, and anomalies, and represent information visually for better interpretation. Visualization helps in simplifying complex data and supports decision-making by highlighting meaningful patterns. Overall, EDA serves as a crucial step before applying machine learning or statistical modeling, ensuring data-driven approaches are accurate and reliable.