Project: Investigating a Dataset- IMDB Movie Dataset from Kaggle

Questions posed:

- The number of movies released each year
- Get the total revenue made from movies for each year
- Top genres according to number of movies released
- What are the top 10 most popular genres throughout the years?
- What are the top genres with the most revenue?
- Which genres are most popular from year to year?
- Movies with a vote average greater or equal to a certain number
- Movies with runtimes greater or equal to a certain number
- Get the movies released each year
- What are the top 10 Production companies with the highest revenue?
- Who are the top Directors with the highest movie count?
- Who are the top Actors with the highest movie count?
- Who are the top 10 most popular actors?
- What are the relationships between features?

Description of the analysis:

The first step carried in the analysis was data wrangling. Duplicate rows were removed, less important columns were deleted. Some columns such as cast, director, genres are separated by '|'. To analyze such column, we used the split method to separate the string by '|'. Then used the explode method to explode the data. Then carried out suitable analysis using groupby method, value_counts() method, exploratory analysis techniques, and several visualizations to show patterns, and relationship.

Data Wrangling made:

- Deleted duplicate rows.
- Deleted columns such as budget, revenue, homepage, overview, tagline, and keywords
- Normalize the popularity column to correct outliers
- Create several copy of the data, to explode them (for columns having values separated by '|")

Summary Statistics and Plots:

The describe method was used to generate the summary statistics for numeric features in the data.

Several plots such as histogram, bar chart were used to visualize analysis. The heatmap method from seaborn module was used to visualize correlation matrix to show the level of association between features.