Predicting imdb score

Introduction

This Python script is designed to analyze a dataset of Netflix original movies. It uses various data visualization techniques to gain insights into the dataset. The code uses libraries such as pandas, numpy, seaborn, matplotlib, plotly, and statsmodels.

Data Loading and Preprocessing

1. Import required libraries, including pandas, numpy, seaborn, matplotlib, plotly, and statsmodels.
2. Load the dataset from the file "NetflixOriginals.csv" using pandas.
3. Explore the dataset with functions like head(), info(), describe(), and check for missing values.

Data Preprocessing

1. Process the "Premiere" column to convert dates into a consistent format.
2. Extract the release year from the "Premiere" column.
3. Convert the "PremiereDate" column to a datetime object.
4. Calculate the average runtime of Hindi movies.
5. Generate a histogram and a boxplot for runtime and IMDB scores to analyze their distributions.

Data Visualization

Bar Chart

- Create a horizontal bar chart to display the count of movies in different languages with a runtime of at least 2 hours.
- Customize the appearance of the chart, including spines, tick positions, and grid lines.

Pie Chart

- Create a pie chart to show the distribution of languages in movies with a runtime of at least 2 hours.
- Customize the appearance of the chart, including title and legend.

Line Chart

- Create a line chart to display the count of movies in different languages with a runtime of at least 2 hours.
- Rotate x-axis labels for readability.

Bar Chart

- Create a bar chart to display the 10 longest-running movies.
- Customize the appearance of the chart, including title and axis labels.

Pie Chart

- Create a pie chart to visualize the total number of movies released in different years.

- Customize the appearance of the chart, including title.

Bar Chart

- Create a bar chart to show the average IMDB scores of movies in different languages.
- Customize the appearance of the chart, including title and axis labels.

Line Chart

- Create a line chart to visualize the total runtime of movies by release year.
- Customize the appearance of the chart, including title and axis labels.

Box Plot

- Create box plots to analyze the distribution of runtime and IMDB scores, identifying potential outliers.

QQ Plot

- Create quantile-quantile (QQ) plots to assess the normality of runtime and IMDB scores.

Outlier Detection

- Implement functions to detect outliers in the dataset using the IQR method and double MAD (Median Absolute Deviation).
- Identify outliers in the “Runtime” and “IMDB Score” columns and provide the count of outliers found.

Conclusion

- Summarize the key findings and insights from the data analysis, including the most used languages, the top-rated movies, and outlier detection result