DAI 101 : Assignment 1

Shubham Kataria, 23114092

1. **Objectives**
   The objective of this report is to conduct an efficient Exploratory Data Analysis on a given dataset, using the following python modules : numpy, pandas, matplotlib and seaborn. This dataset includes data of various movies on netflix platform which includes their budget, genres, revenue, language, title, popularity etc. The following analysis focuses on data cleaning and univariate, bivariate and multivariate analysis on the dataset.

2. **Difficulties with the data**
   The dataset contains various inconsistencies which include :
   - Missing values
   - Duplicate Values
   - Zero values for numerical data
   - Outliers

3. **Data cleaning**
   The data was cleaned using the functions given in numpy and pandas. The following steps were performed for the same :
   - Filling the missing values for non-important fields like genres with “unknown”
   - Dropping the rows with missing values for important fields like title.
   - Removing the data with zero budget or revenue
   - Finding and removing the outliers using the IQR method

4. **Univariate Analysis**
   Univariate analysis was performed on both categorical and numerical data. The following were the observations :
   - Categorical analysis
     - Bar graphs of the movies on the basis of language and popularity were plotted
     - A very large chunk of the data consisted of english movies while french and spanish movies were highest among the non-english ones.
     - Low and Medium rated movies were contributing to the data while Blockbuster ones were quite rare.

- Numerical Analysis
    - Histogram and Boxplots were drawn for the budget, revenue and the average runtime.
    - Budget and Revenue followed a similar pattern, both following an exponentially decaying graph.
    - The runtime followed normal distribution centered around the standard 90-110 min time.

5. **Bivariate Analysis**
   Various techniques were incorporated while representing the relation between the columns which include :
    - Scatterplot
        - Graph between budget and revenue showed that a majority of the data consisted of low budget low revenue movies.
        - Too short and too long movies were often followed by a low rating while the 80 - 120 minute ones got the highest vote average
    - Bar Graph
        - Average Revenue vs Languages showed that japanese and chinese movies were quite successful while french and italian movies didn't perform quite well
    - Violin Plots
        - Graphs showed that the low rated movies didn't bring much revenue while the high and blockbuster movies were a huge success.
    - Box Plots
        - Japanese and Chinese movies were highly rated while the english ones were normally rated.

6. **Multivariate Analysis**
   Multiple columns were taken into consideration while forming the following graphs :
    - Pairplot
        - Results show that the majority of the data consists of low budget low revenue and 2hr movies. A longer movie required a longer budget.
    - Correlation Heatmap
        - Budgets and Revenue were highly correlated to each other. A high revenue movie often required a high budget.
        - Also, a high revenue movie often generated a high vote average.

- Bar Graph
    - Japanese gave the highest grossed blockbusters while highly rated french movies also generated a great amount of revenue.

7. Conclusion

The following insights can be derived from this analysis :

- Data cleaning enhanced data quality
- Budget and Revenue were strongly related
- Most of the dataset consisted of english movies
- The 1.5 - 2hr segment was the most optimal time for a movie.