

**Project Report**

Student Name: **Piyush Jaloniya**  UID: **24MCI10042**

Student Name: **Khushi** UID: **24MCI10059**

Branch: MCA (AIML) Section/Group:24MAM 1-B

Semester: 1st Date: 25/10/2024

Subject Name:R Programming Subject Code: 24CAP-614

1. **Title**

Data Analysis and Visualization

1. **Aim/Overview of the project:**

The primary aim of this project is to perform **Exploratory Data Analysis (EDA)** on a chosen dataset from Kaggle or the UCI Machine Learning Repository using **R** programming. The project involves understanding the dataset, cleaning it, exploring the distribution of variables, identifying outliers, and visualizing relationships between different variables. The objective is to gain insights into the data by using various data visualization techniques like histograms, scatter plots, and boxplots.

1. **Task to be done:**

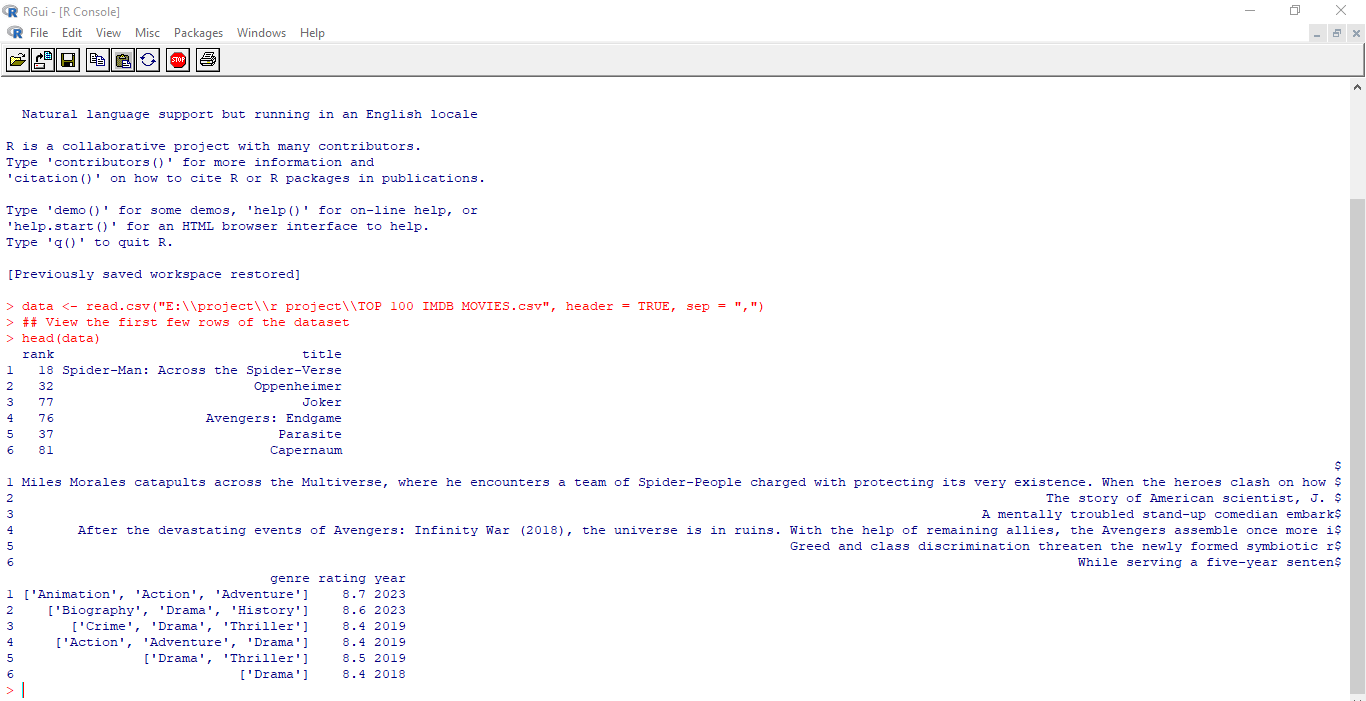
* **Dataset Selection**:
  + Select a suitable dataset from Kaggle or UCI Machine Learning Repository for performing EDA.
* **Data Loading and Cleaning**:
  + Load the dataset into R.
  + Handle missing values, if any, using appropriate methods (e.g., imputation).
  + Perform necessary data cleaning tasks such as removing duplicates, handling invalid data, and feature engineering if required.
* **Exploratory Data Analysis (EDA)**:
  + **Descriptive Statistics**: Compute summary statistics such as mean, median, mode, standard deviation, and range for numerical variables, and frequency distribution for categorical variables.
  + **Univariate Analysis**: Explore individual variables in the dataset using histograms, boxplots, and summary statistics.
  + **Bivariate Analysis**: Explore relationships between two variables using scatter plots, boxplots, and correlation analysis.
  + **Outlier Detection**: Identify outliers in continuous variables using techniques like boxplots.
* **Data Visualization**:
  + Create visualizations to understand the distribution of the data and the relationships between variables.
  + Plot histograms for continuous variables to examine their distribution.
  + Create boxplots to identify outliers and visualize the spread of data.
  + Plot scatter plots to observe relationships between two continuous variables.
  + Analyze patterns or anomalies in the data using visual insights.
* **Conclusion**:
  + Summarize the key insights gained from the data through EDA.
  + Discuss the patterns, trends, or relationships observed between the variables.
  + Highlight any significant findings such as the presence of outliers or correlations between variables.

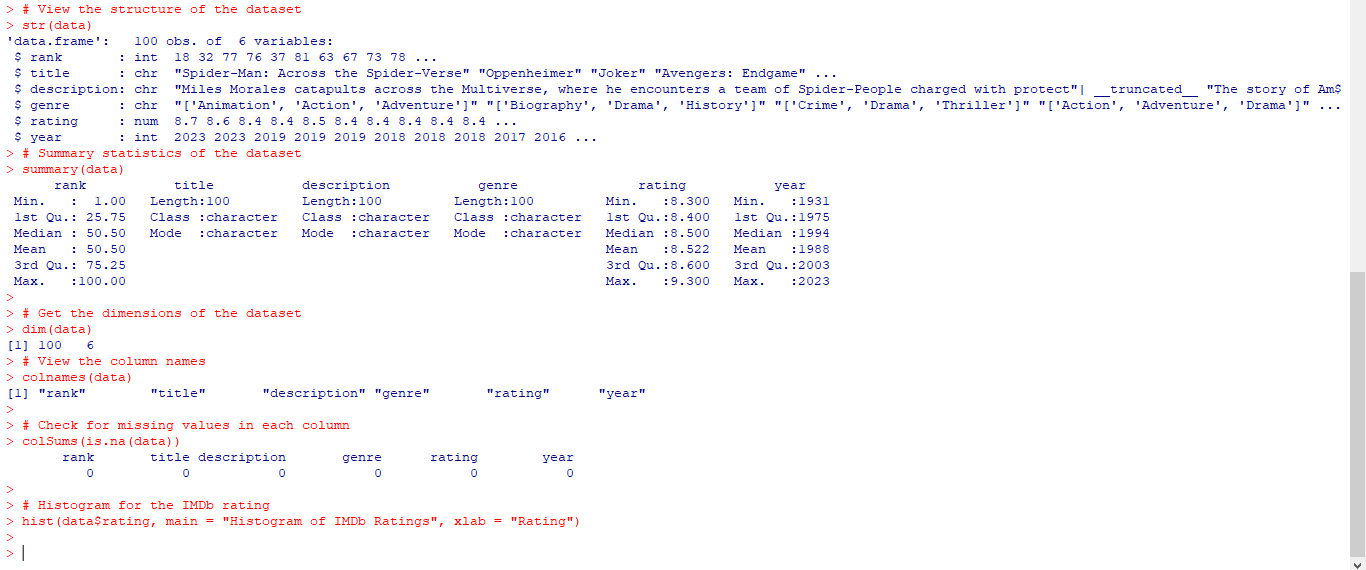
1. **Dataset**

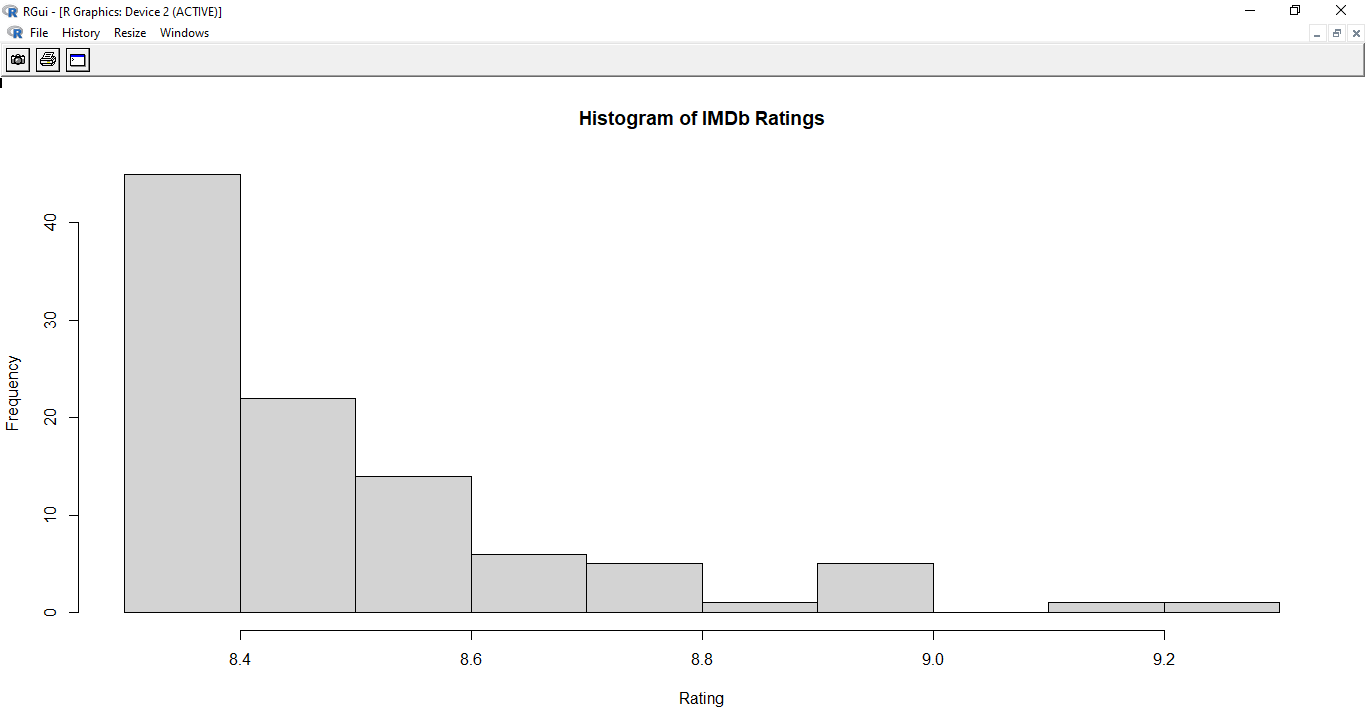
Top 100 IMDB movies dataset from Kaggle

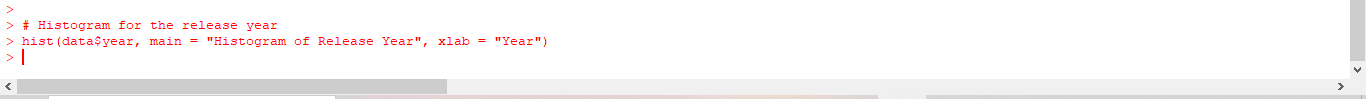
<https://www.kaggle.com/datasets/mayurkadam9833/top-100-imdb-movies>

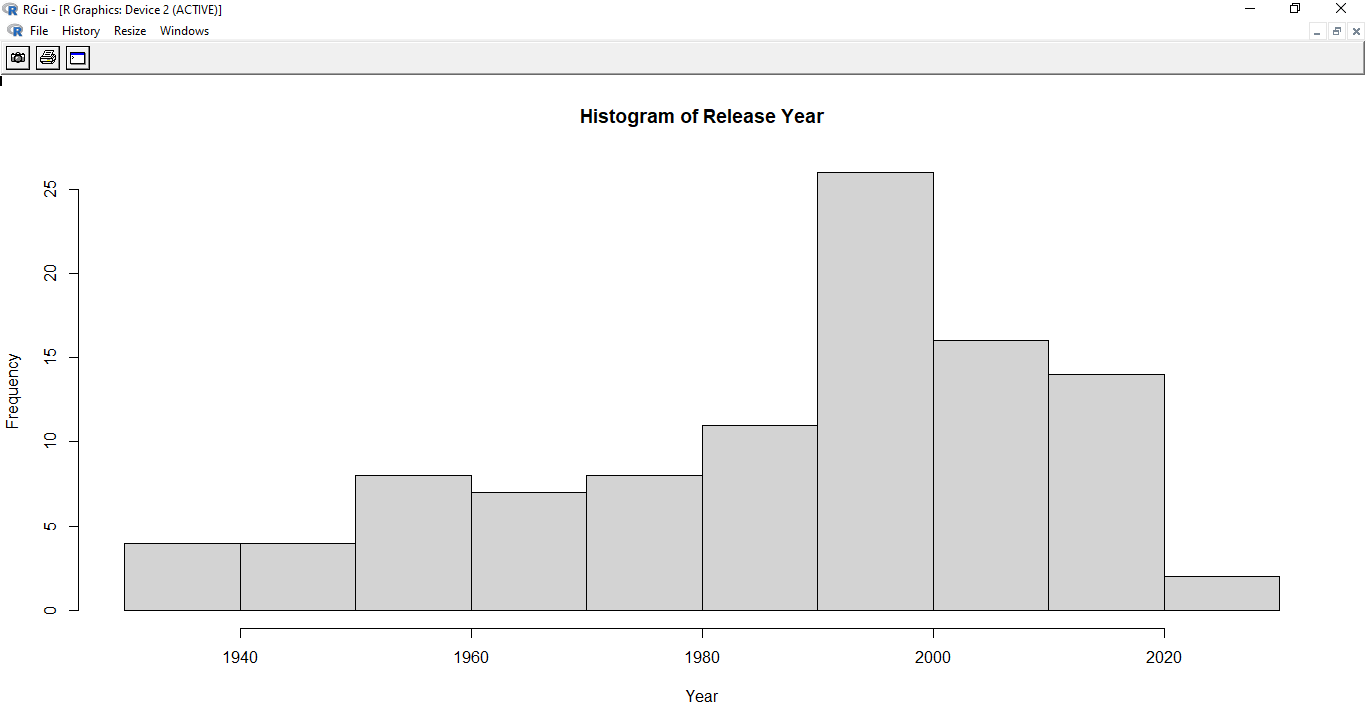
1. **Steps/Commands involved to perform project:**

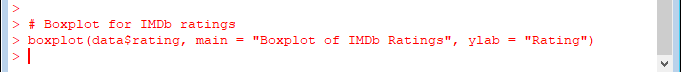


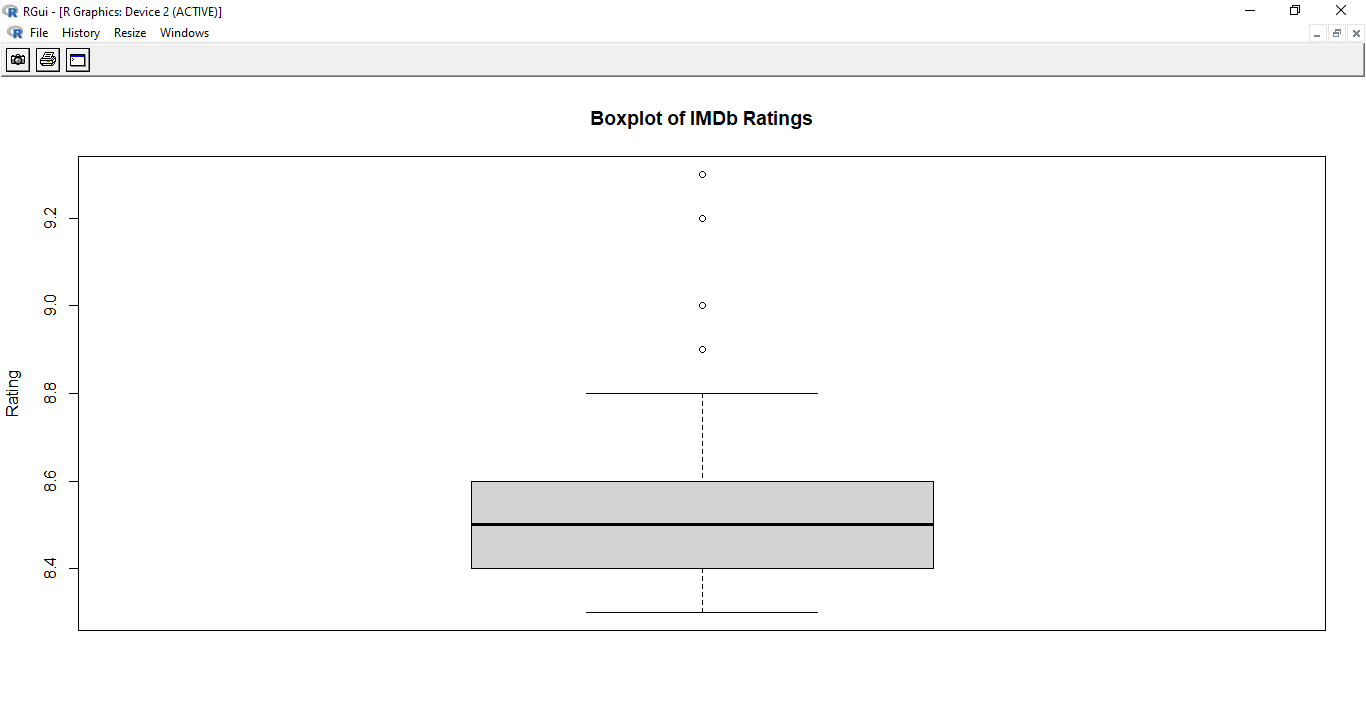
****

****

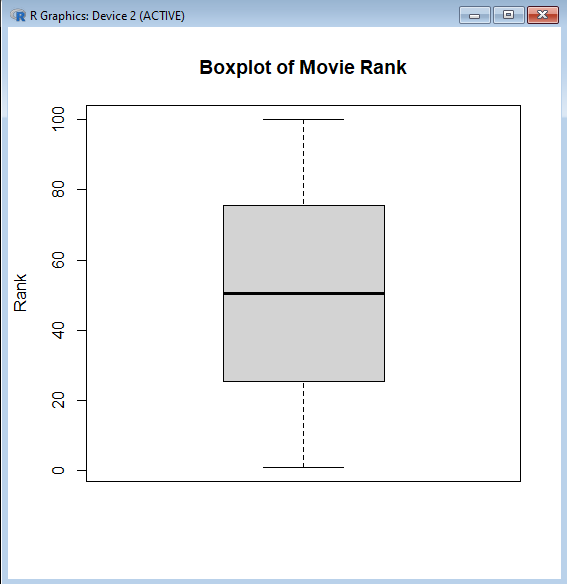
****

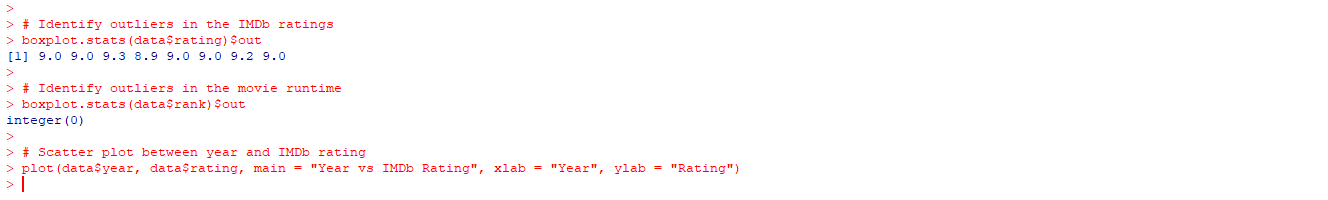
****

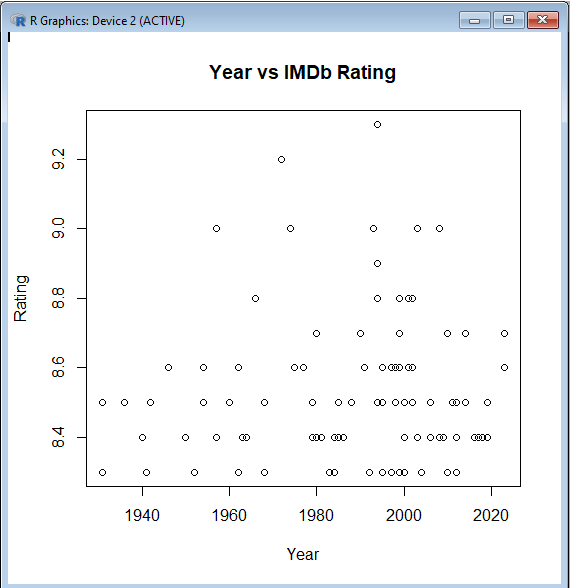
****

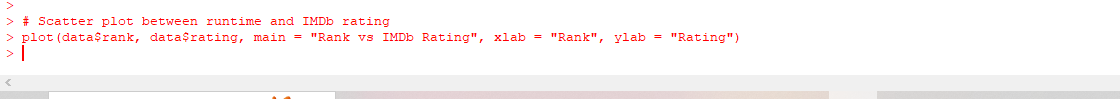
****

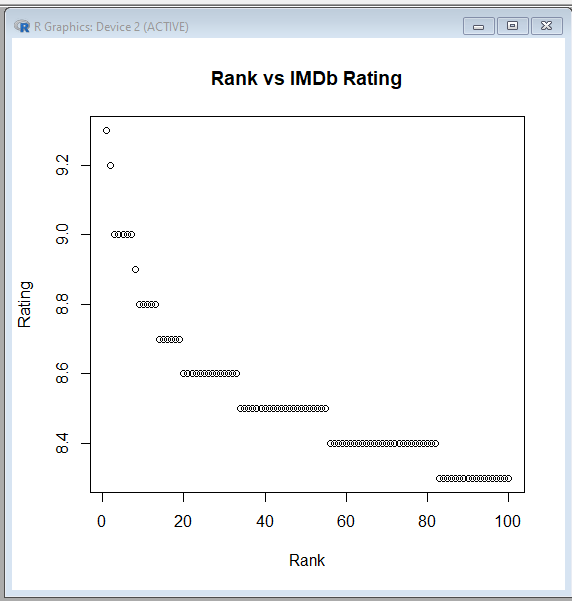
****

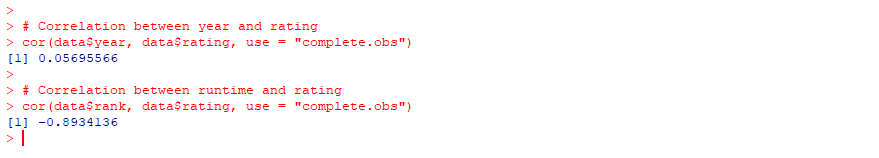
****

****

****

****

****

****

**3.) Learning outcomes(What I have learnt):**

Upon successful completion of this project, learning outcomes are :

* **Understanding of EDA Process**: The project will provide a strong foundation in performing exploratory data analysis in R. You will learn how to approach a new dataset, clean it, and explore its structure.
* **Hands-on Experience with R for EDA**: By the end of the project, you will gain practical experience in using R programming to perform data cleaning, compute descriptive statistics, and create various plots for data visualization.
* **Proficiency in Visualizing Data**: You will be able to visualize data distributions and relationships between variables using histograms, scatter plots, and boxplots. This will help in deriving meaningful insights from the data.
* **Ability to Detect Outliers**: You will develop the skills to identify outliers and anomalies in data using visualization techniques and statistical summaries.
* **Data-Driven Insights**: The project will enhance your ability to make data-driven decisions by uncovering patterns, trends, and relationships in the data through EDA.