Table of Contents

[**A.** **Initial Exploratory Data Analysis** 2](#_Toc159991991)

[1. Installing Tidyverse Package 2](#_Toc159991992)

[2. Loading and Viewing CVS File 2](#_Toc159991993)

[3. Checking Data Type 2](#_Toc159991994)

[**A.** **Data Cleaning** 3](#_Toc159991995)

[**B.** **Exploratory Data Analysis** 3](#_Toc159991996)

[1. Summary Statistics 3](#_Toc159991997)

[2. Data Visualisation - Scatterplot 4](#_Toc159991998)

[3. Data Visualisation – Bar Chart 4](#_Toc159991999)

[**C.** **Exporting Clean Data** 5](#_Toc159992000)

[**D.** **Power BI Data Visualisation** 5](#_Toc159992001)

[1. Number of Movies Each Year 6](#_Toc159992002)

[2. Average Rotten Tomatoes Ratings of Each Genre 6](#_Toc159992003)

[3. Audience Score for Each Film 7](#_Toc159992004)

[4. Profitability Per Studio 7](#_Toc159992005)

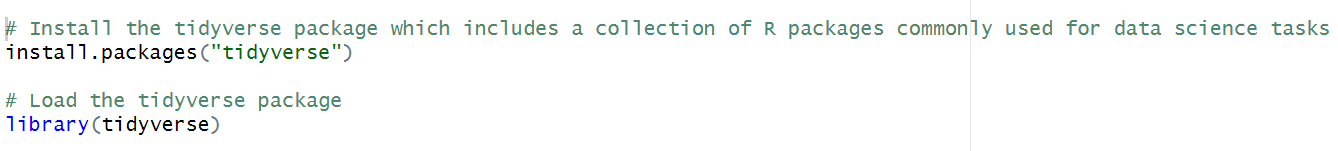
[5. Worldwide Gross Per Genre 8](#_Toc159992006)

# **Initial Exploratory Data Analysis**

## **Installing Tidyverse Package**

In the code below, the install.packages("tidyverse") command was used to install the 'tidyverse' package in R. The 'tidyverse' is a collection of R packages designed for data science tasks, including data manipulation, visualization, and analysis. It includes popular packages like 'dplyr', 'ggplot2', 'tidyr', and others, which are widely used for their consistent syntax and powerful functionality in data analysis workflows.

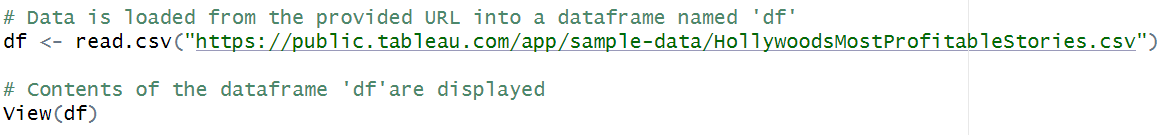
library(tidyverse) command loads the tidyverse package, making all its functions and datasets available for use in the current R session.



Top of Form

## **Loading and Viewing CVS File**

The read.csv() function was used to load data from the provided URL into a dataframe named df in the code below.



View(df) opens a new window displaying the contents of the dataframe df. This allows for visual inspection of the data, including its structure, variables, and values, which can be helpful for understanding the dataset before performing any analysis or manipulation.

A screenshot of a computer

Description automatically generated

## **Checking Data Type**

The str() function has been used to display the structure of the dataframe df, including information about its columns, data types, and the first few values.

As can be seen there are total 8 variables in this dataset and 74 observations/rows.

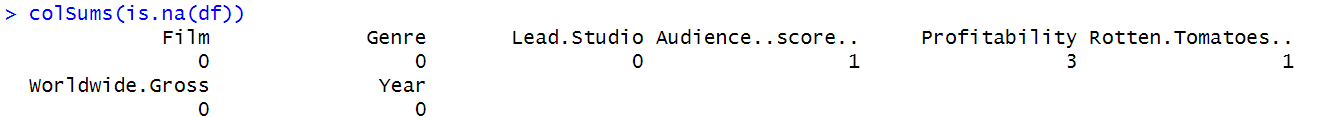
A close-up of a number

Description automatically generated

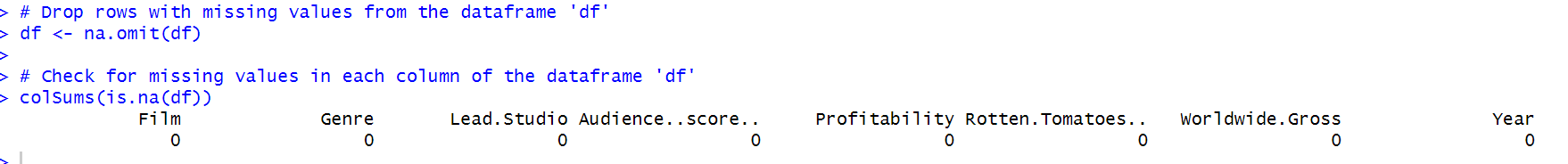
# **Data Cleaning**

To calculate the total number of missing values in each column of the dataframe, colSums(is.na(df)) command has been used , with it returning a vector with the sums of missing values for each column.

Currently, as can be seen, the dataset has 3 missing values in the Profitability column and 1 missing value in the Rotten Tomatoes % column.



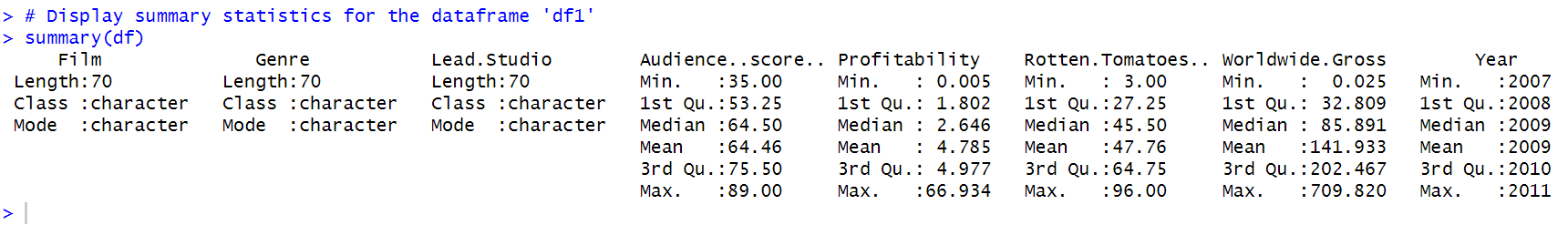
Then we have used na.omit(df to removes rows with any missing values (NA) from the dataframe and return a new dataframe without those missing value rows.

****

# **Exploratory Data Analysis**

## **Summary Statistics**

We have generated summary statistics using summary(df), which provides a quick overview of the distribution and central tendency of the data.



For numerical data such as audience score, profitability, rotten tomatoes, etc., statistics including minimum, maximum, standard deviation, and mean have been generated.

For non-numeric data such as Film, Genre, and Lead Studio, statistics such as length, class, and mode have been generated.

## **Data Visualisation - Scatterplot**

We utilized ggplot for visualizing the relationship between Lead Studio and Rotten Tomatoes ratings using scatterplot.

ggplot(df, aes(x = Lead.Studio, y = Rotten.Tomatoes Rating %)) was used to specify the dataframe and mappings for x (Lead Studio) and y (Rotten Tomatoes ratings) aesthetics and geom\_point() added points to create a scatterplot.

We also formatted y-axis labels using commas for better readability and rotated x-axis labels by 90 degrees for better readability.

The scatterplot generated can be seen below:

**A graph with black dots

Description automatically generated**

There is considerable variability in Rotten Tomatoes ratings across different lead studios. Some studios have movies with high ratings (e.g., Disney with ratings as high as 96%), while others have lower ratings (e.g., Lionsgate with ratings as low as 3%).

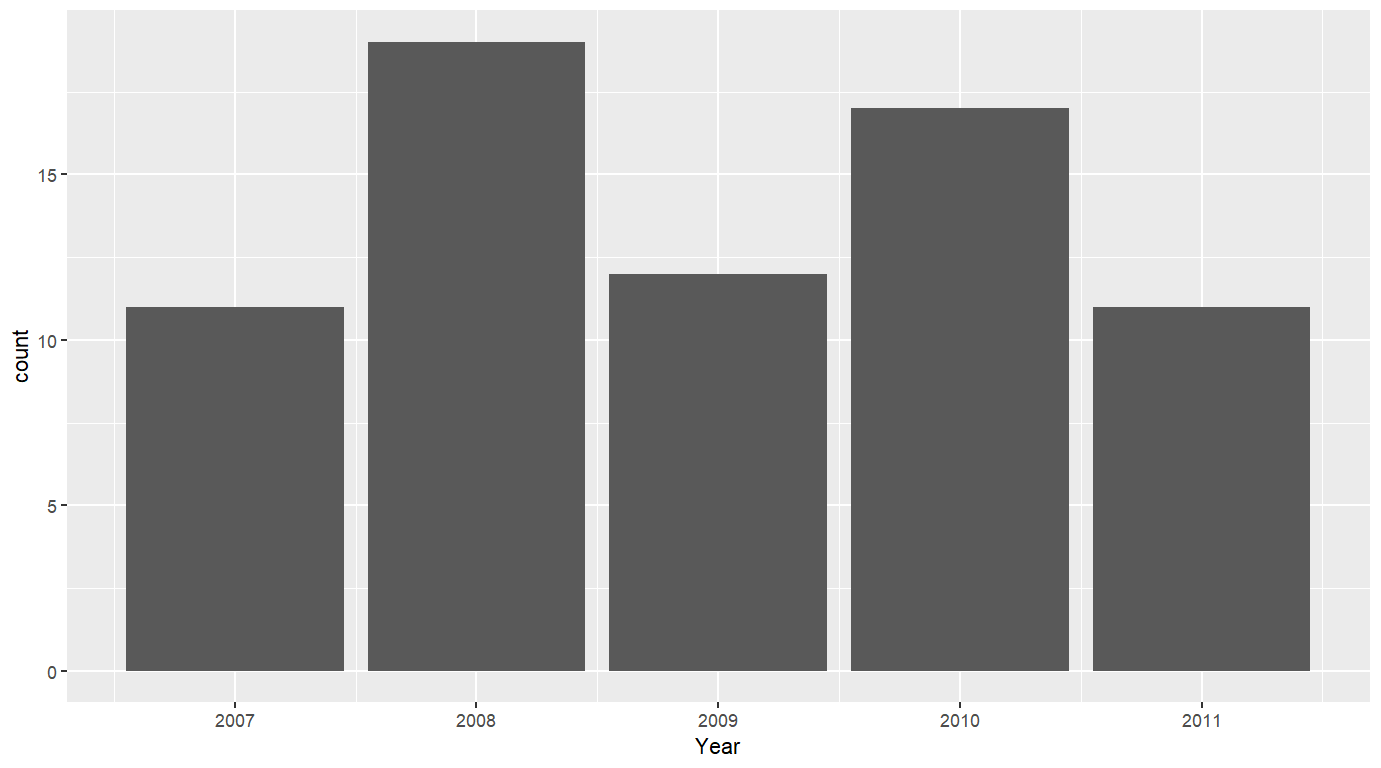
Independent studios produce movies with a wide range of Rotten Tomatoes ratings, indicating variability in audience reception. Ratings range from as low to high as can be seen in the scatterplot.

## **Data Visualisation – Bar Chart**

A bar chart was created to illustrates the distribution of movies over various years. The x-axis displays the years, while the height of each bar corresponds to the number of movies released in that particular year.

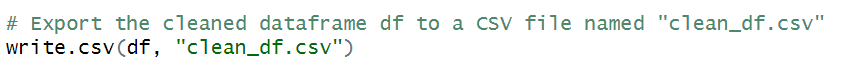


Years 2008 and 2010 has produced the most movies, as can be seen below.

****

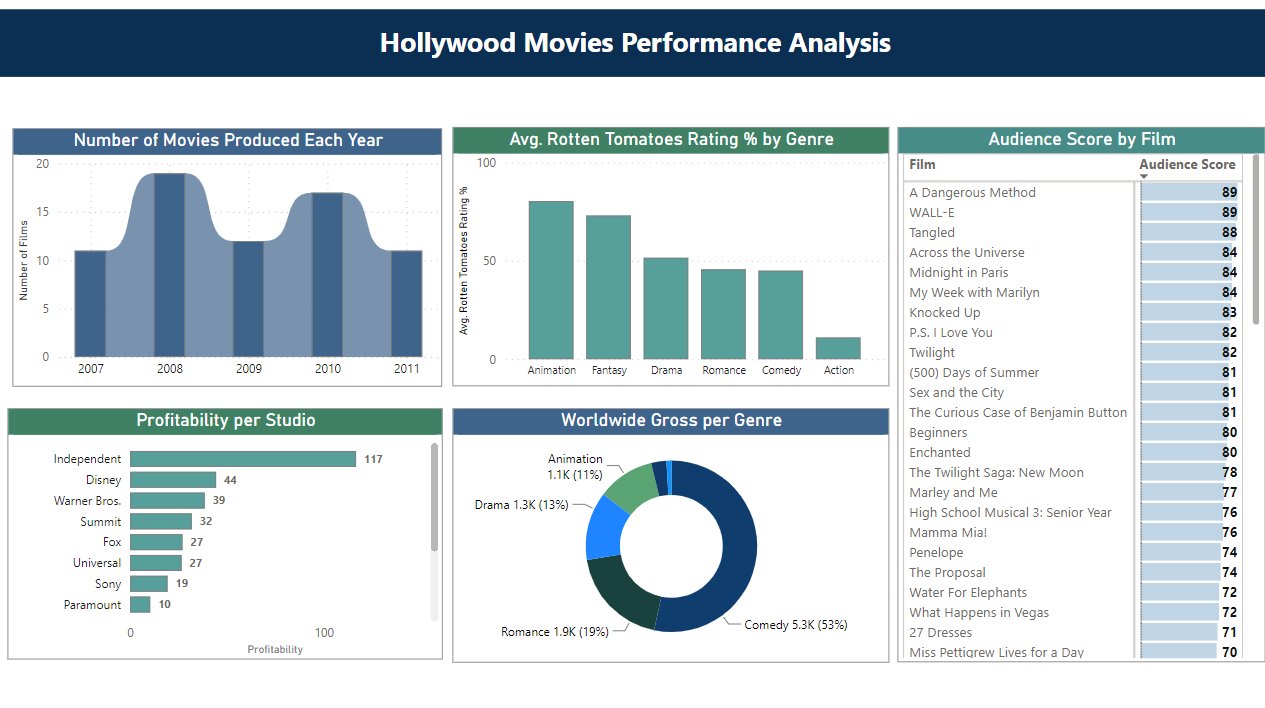
# **Exporting Clean Data**

After conducting data cleaning and exploratory data analysis we have then exported the data to a CSV file named "clean\_df.csv". This CSV file contains the cleaned data without any missing values or modifications made during data processing. We will be using this file for further data visualisation in Power BI.

****

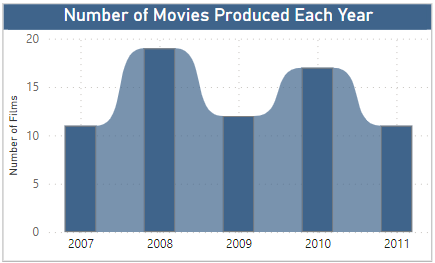
# **Power BI Data Visualisation**

Power BI was employed to create a comprehensive and visually compelling dashboard, tailored to the client's specifications. The dashboard comprised five key visuals, each strategically designed to best represent the dataset, with a predominant use of blue and green shades in accordance with the client/assignment’s colour requirements.

****

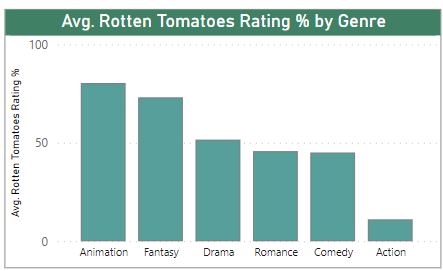
## **Number of Movies Each Year**

The ribbon chart was used to effectively visualise Hollywood's annual movie production trends. Notably, 2008 and 2010 stood out with the highest number of films released, highlighting dynamic shifts or significant industry events during those years.



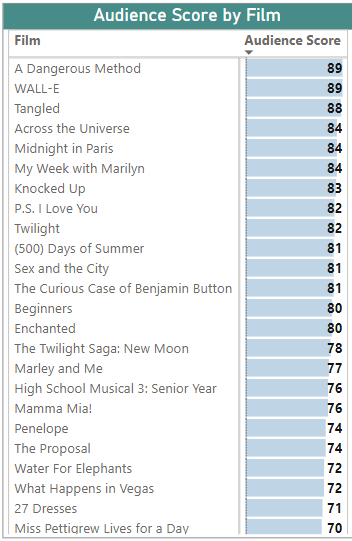
## **Average Rotten Tomatoes Ratings of Each Genre**

Utilizing a bar chart, the analysis of average Rotten Tomatoes Ratings % across genres revealed Animation and Fantasy as the top performers with the highest ratings. Conversely, the Action genre recorded the lowest average ratings, providing valuable insights into audience preferences within the dataset.



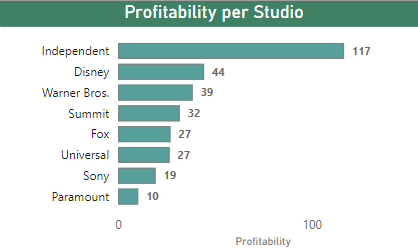
## **Audience Score for Each Film**

Leveraging a matrix visualization for audience scores, the data was effectively presented with data bars to visualize scores. The matrix, thoughtfully sorted, showcased films with the highest ratings at the top, offering a clear and concise overview of audience preferences. Movie ‘Dangerous Method’ and ‘WALL-E’ had the highest audience scores.



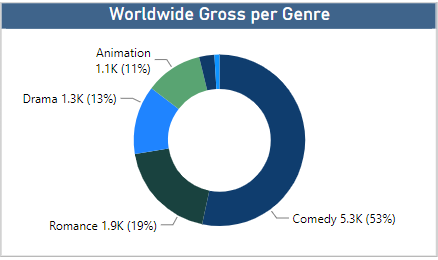
## **Profitability Per Studio**

The stacked bar chart was used to effectively portray profitability per studio, with Independent studios emerging as the most profitable



## **Worldwide Gross Per Genre**

Pie Chart was used to showcase the worldwide gross per genre, revealing Comedy as the highest grossing genre with $5300 million, while Action occupied the lowest share. This visual presentation facilitated a quick and intuitive understanding of the distribution of global earnings across different movie genres.



Top of Form