# Introduction:

In this report, we will discuss the R code provided and explain the reasoning behind each step. The goal of this code is to clean and analyse a data set in R. The data set in question includes information on various movies, such as the year they got released, the profitability, lead studio, and more.

# Data Cleaning:

Before trying to analyse the data, we must first perform some data cleaning commands. The first step in data cleaning is to import the necessary libraries. In this case, we will be using the “tidyverse” and “ggplot2” libraries, which provide a wide range of functions for data cleaning and visualisation.

Next, we want to check the data types of our variables to ensure that they are correct. This can be done using the “str” function, which can be seen in Figure 4. After that, we check for missing values in our data set using the “colSums” and “is.na” functions. In this case, we have missing values, so we use “na.omit” function to remove them from our data set.

Moreover, we want to check for duplicates in our data set using the “duplicated” function. If we find duplicates, we can remove them by using the “subset” function.

After we are done cleaning the data set, we can round off our numeric variables to two decimal places using the “round” function.

# Data Visualisation:

Visualising our data set is a very important step as it allows us to better understand what we are working with. Additionally, we can use data visualisation techniques to identify outliers in our data, which could skew our results. To check for outliers, we use a boxplot to visualise the distribution of our data (Figure 9). In this case, we are interested in the relationship between profitability and worldwide gross, so we can create a boxplot using the “ggplot” function. We also highlight the outliers using the “outlier.color” and “outlier.shape” arguments.

# Outlier Removal:

After identifying outliers in our data, we want to remove them to revent them from skewing our results. For this we use the interquartile range (IQR) method to identify outliers and remove them using the “subset” function.

# Data analysis:

Once we have cleaned our data and removed outliers, we can perform a univariate analysis to summarise our data. We can use the “summary” function to get summary statistics for our data set, such as the minimum and maximum values, the mean, and the class of our data.

We also perform a bivariate analysis to explore the relationship between Lead Studio and Rotten Tomatoes ratings. In this case we created a scatterplot using the “ggplot” function. Additionally we created a bar chart using the same function in order to see the count of films in each year.

# Data export

Finally, we export our clean data set to a CSV file using the “write.csv” function, and perform further analysis using PowerBi.

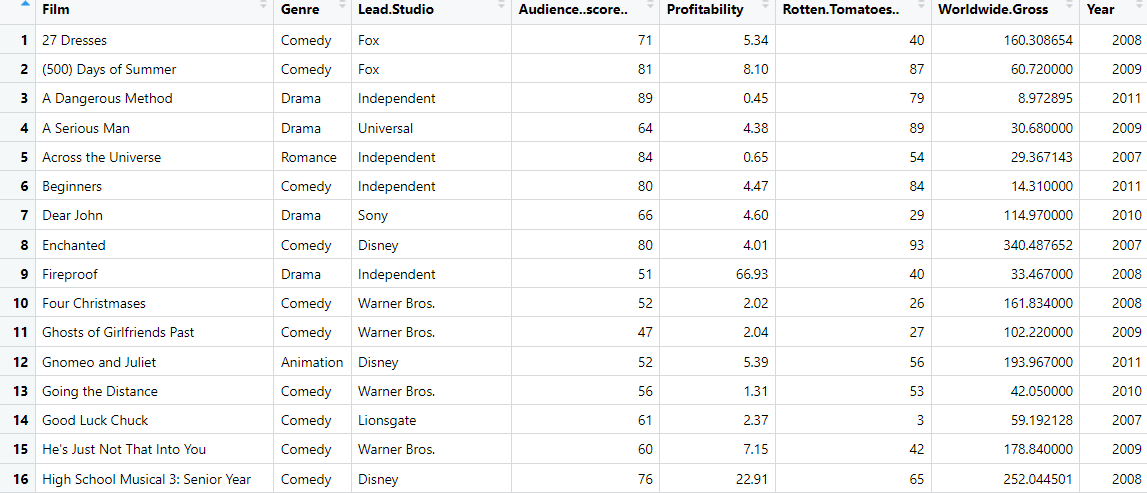
# Conclusion:

To conclude, the R code provided allows us to clean and analyse a data set in R. By performing various data cleaning techniques and visualising our data, we are able to identify and remove outliers, as well as summarise and analyse our data. This process if important for ensuring that our data is accurate and reliable and for providing insights that be used for further analysis or decision-making.

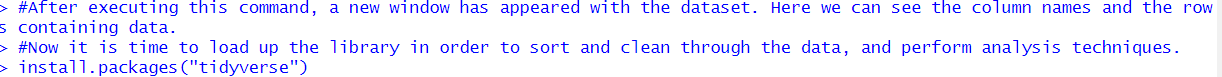
Graphical user interface

Description automatically generated with medium confidence

Figure



Figure



Figure

A picture containing text

Description automatically generated

Figure

Graphical user interface, text, application

Description automatically generated

Figure



Figure

Text

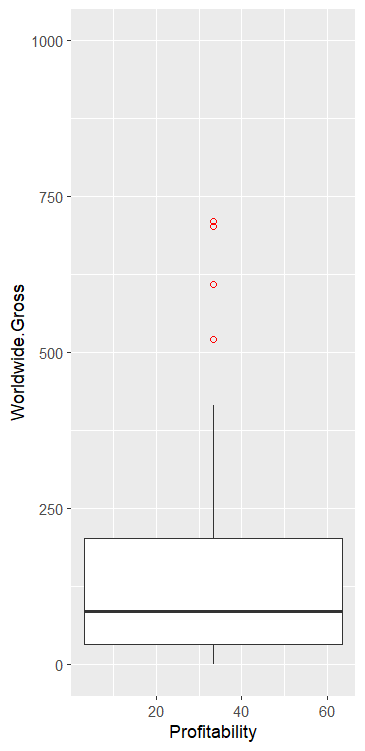
Description automatically generated

Figure

Text

Description automatically generated

Figure



Figure

Graphical user interface, text

Description automatically generated

Figure

Text, letter

Description automatically generated

Figure

Text

Description automatically generated

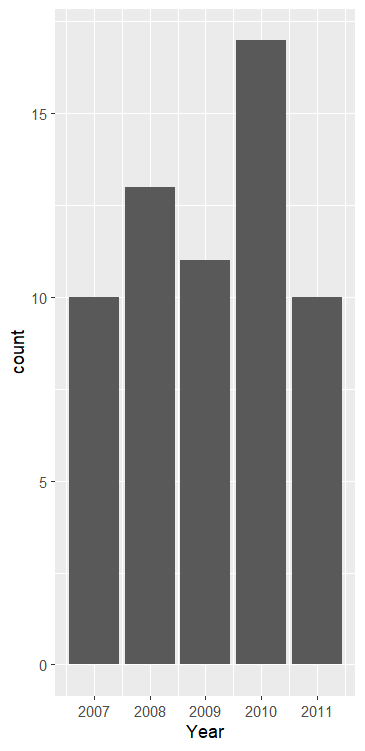
Figure



Figure



Figure



Figure



Figure

Text

Description automatically generated with medium confidence

Figure