Movies Reviews

علي يحيي زكريا فهمي	2021170820
احمد سامح احمد مختار	2021170802
فادي مجدي زكي ابراهيم	2021170824
احمد طارق محمد كمال	2021170804
كريم فؤاد شهب محمد	2021170826

Year 2 SWE Section 1,2

Introduction

- *Statistical analysis for Movies reviews dataset.*
- *Project goals are to analyze the relation between the movies rating , the movie runtime,release year and number of votes.*
- *The direction of the project is to compare between new movies and old ones.*

Research question

Does the IMDB rating varies with the year of release and the runtime of the movie and is there a relation between the runtime of the movie and the year it got released in?

Exploratory data analysis

Number of votes

- *Mean is 273697.4*
- *Sd is 327536.6*

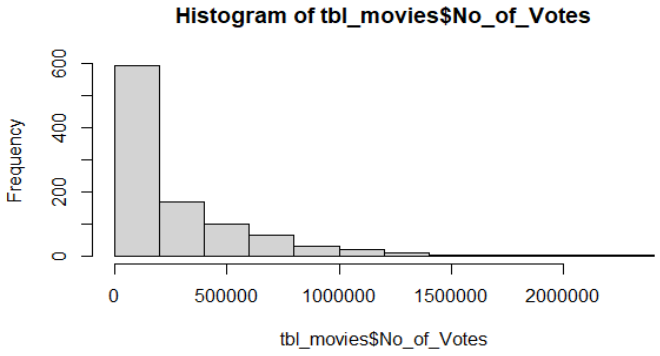

Figure 1 Histogram of number of votes

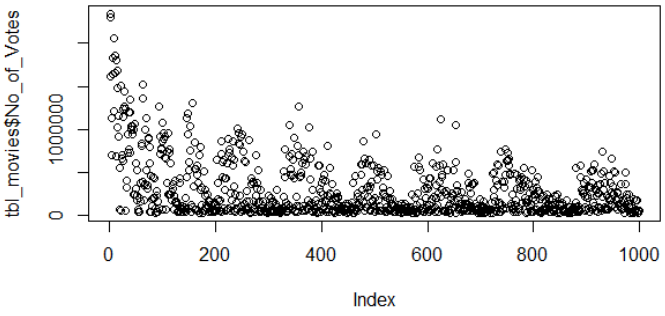

Figure 2 Dot plot of number of votes

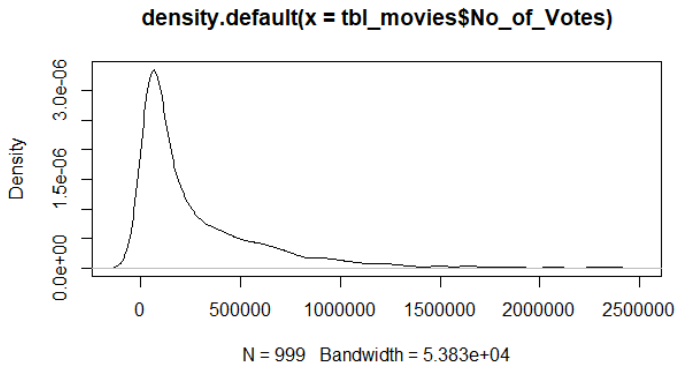

Figure 3 Density plot of number of votes

IMDB Rating

- *Mean is 7.94965*
- *SD is 0.2754071*

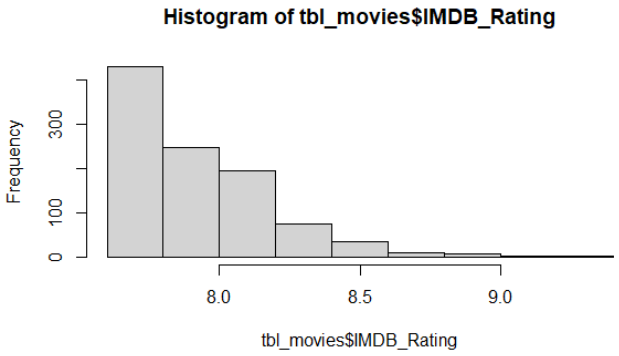

Figure 4 Histogram of IMDB Ratings

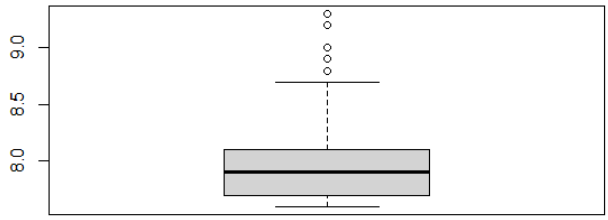

Figure 5 Boxplot of IMDB Rating (with outliers)

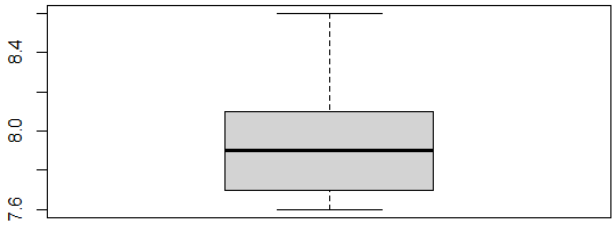

Figure 6 Boxplot of IMDB Rating (without outliers)

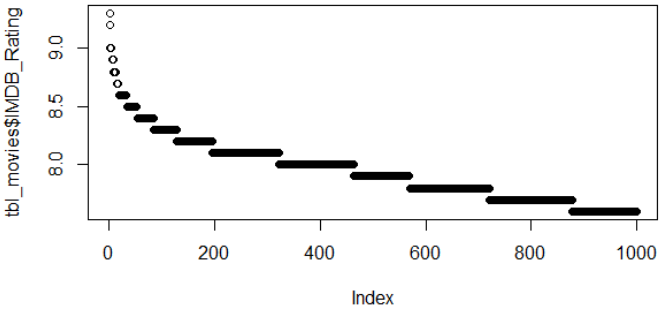

Figure 7 Dot plot of IMDB Rating

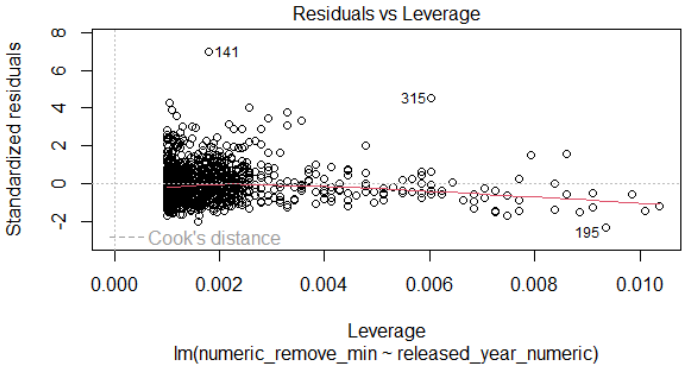

Figure 8 Line regression between IMDB Ratings and Year of Release

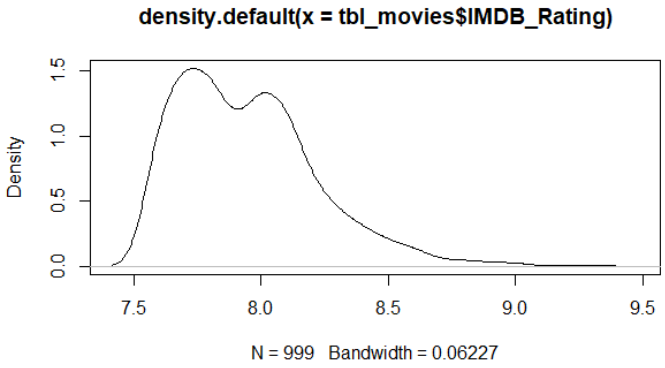

Figure 9 Density plot of IMDB Ratings

Released Year

- *Mean is 1991.217*
- *Sd is 23.29702*

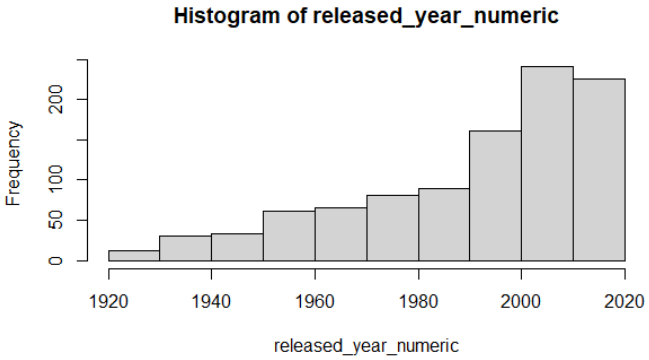

Figure 10 Histogram of movies released year

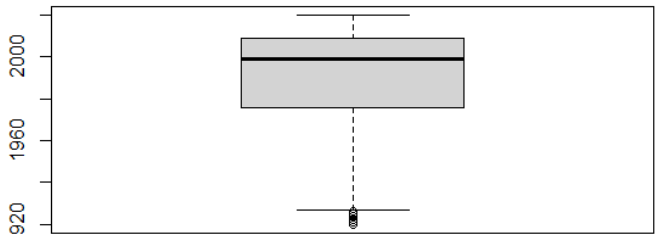

Figure 11 Boxplot of movies released year (with outliers)

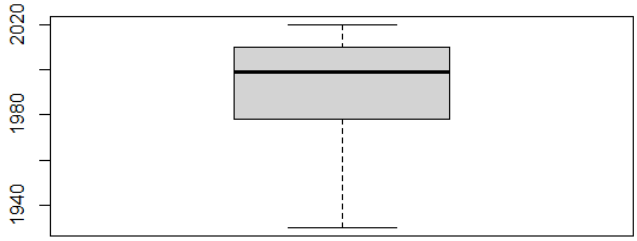

Figure 12 Boxplot of movies released year (without outliers)

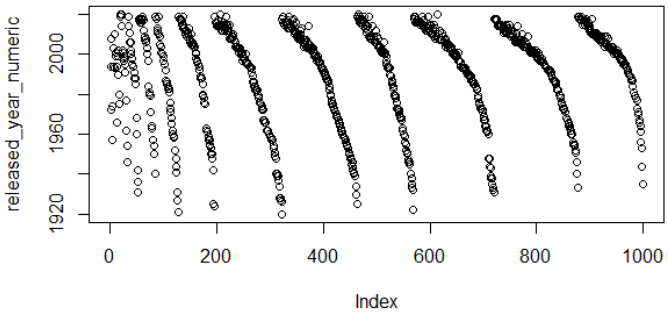

Figure 13 Dot plot of movies released year

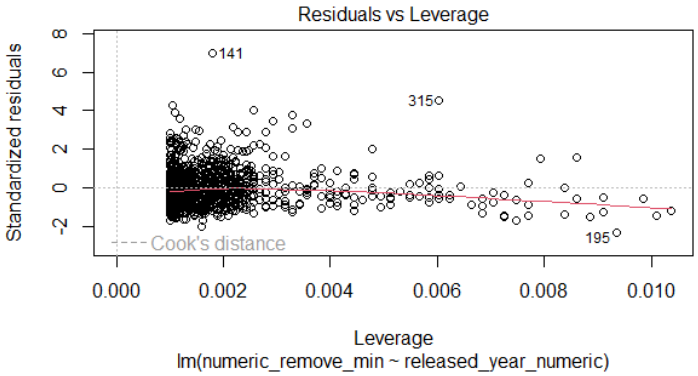

Figure 8 Line regression between IMDB Ratings and Year of Release

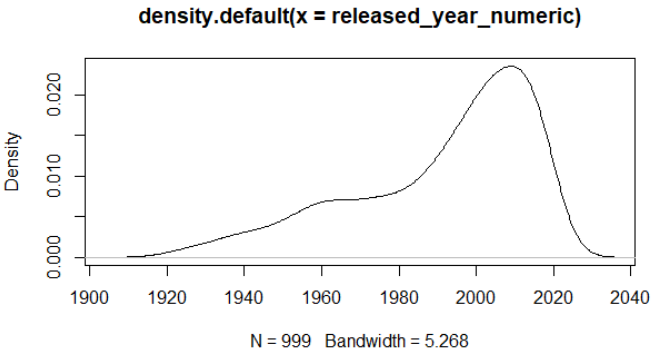

Figure 14 Density plot of movies released year

Runtime

- *Mean is 122.8739*
- *Sd is 28.10252*

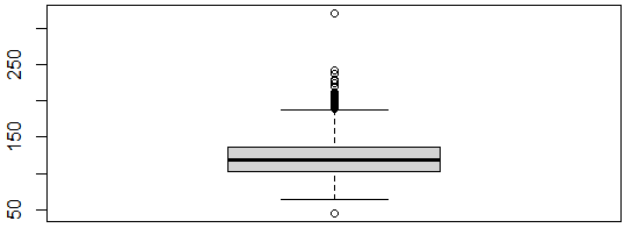

Figure 15 Boxplot of movies runtime (with outliers)

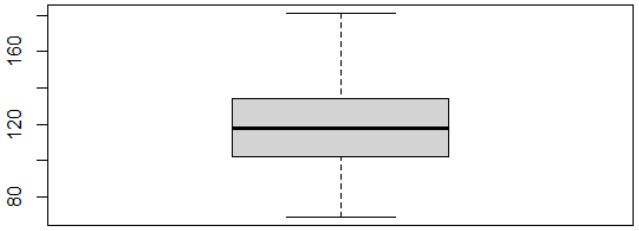

Figure 16 Boxplot of movies runtime (without outliers)

Z-score

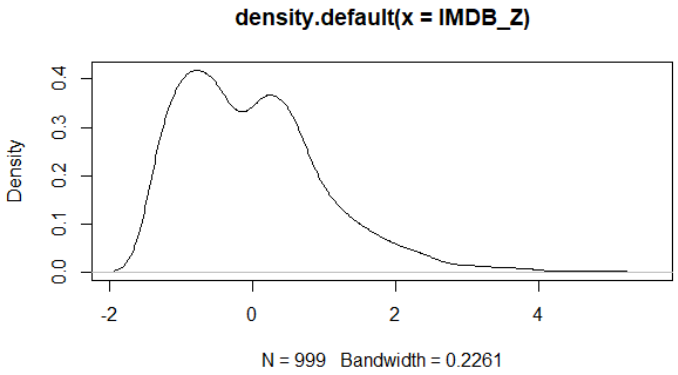

Figure 17 Density plot of movies IMDB Rating (with outliers)

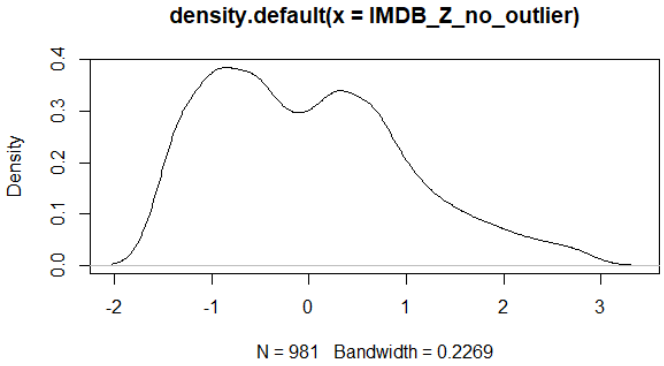

Figure 18 Density plot of movies IMDB Rating (without outliers)

- *At figure 17, there are some z-score values in the table that exceed Z=3 therefore, they are outliers.*
- *At figure 18, the z-score values are lower than Z=3 due to the removal of the outliers.*

Correlations

- *The correlation between IMDB rating and the movie runtime is 0.2441116 which shows that they have a weak positive relation.*
- *The correlation between the movie runtime and the released year is 0.1658067 which shows that they have a weak positive relation.*
- *The correlation between IMDB rating and the released year is -0.1310527 which shows that they have a weak negative relation.*

Final Report

The previous statistical data and graphs helps us in analyzing the movies rating dataset:

- *When the movie is longer it mostly get better IMBD rating.*
- *Newer movies have longer runtime mostly.*
- *Newer movies have less IMDB rating than older ones.*
- *At figure 4 , 7 and 9 , most movies tend to be distributed around the 7.5 and 8.5.*
- *At figure 10 , most of the movies at the dataset are newer.*
- *At figure 14 , most of the movies are released between 1980 and 2020.*

Improvements and Limitations

Improvements

- Use a bigger dataset to lower our margin error

- Search for a dataset that contains the least number of nulls possible and has appropriate values to work on a relation.

Limitations

- *Column runtime from table movies had the minutes included in as characters. To use it in calculations, we had to remove the mins from the column and then convert it to numeric values*

- *In the dataset, there was a lot of null values. It was necessary to remove them to use it in calculations.*

- *We couldn't calculate the Z-score using qnorm function due to NaN produced error. So, we used the naïve solution of using the mean and S.D functions in the Z-score formula.*