MOVIES

A FINAL PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR COURSE STAT 291 –
STATISTICAL COMPUTING I
DEPARTMENT OF STATISTICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GROUP 4

AYLİN ÇELİK
DENİZCAN BOZKURT
MISRA SARGIN
TUNCA AKÇAOĞLU

January 2022

ABSTRACT

IMDB is the world's most popular source for movies. Before choosing to watch it, people look at how qualified the film is and wonder about themselves. IMDB offers a rating scale that allows users to rate films on a scale of one to ten. In this study, by using descriptive statistics and some statistical tests, namely, ANOVA—Kruskal-Wallis, T-test, and linear regression, the relationships between movies' properties and their ratings were examined. At the end of these analyses, results founded in below.

- Positive correlation between IMDB score and the number of users for review and voted users. However, there is a negative correlation between IMDB scores and movie release years.
- Language is not a significant factor affecting IMDB scores.
- Christian Bale starring movies have a greater IMDB score than the average score.
- IMDB score increases as the duration increases.

1. Introduction

In this study, five thousand movies from the years 1924-2016 basic pieces of information are collected in the data set are used.

The dataset includes 14 different variables:

Director name, Movie title, Movie IMDB link, Duration in minutes, Genres

Actor name1(Leading actor), Actor name2(Leading actor), Actor name3(Leading actor)

Number of voted users (5 to 1,689,764), Number of users for reviews (1 to 5,060)

Language, Country, Release year, IMDB score (1 to 10)

This study aims to find the logical relationship between some variables with ratings.

1.1 Data Description

The dataset is from data.world, and it was last updated in 2017. It was a refined movie list by a user and included 5043 rows and 14 columns in the data frame structure. In analysis, seven variables out of 14 were used: year of release, duration in minutes, users voted film, reviews number, language, actor, and IMDB score. Also, rows that include NA values were removed, and some arrangements happened in some variables because of the language sample size difference. The levels for categorical variables were adjusted after the structure was verified to confirm if the classes were in proper form.

The link is https://data.world/himan/imdb-movie-dataset/

1.2 Research questions

In this study, factors that might associate with IMDB score analyzed. In order to show these associations, 4 research questions were asked.

1. What is the relationship between movie release year, number of users voted film, number of users for reviews and IMDB score?

2. Are IMDB scores different for the 5 languages of movies?

3. Is the IMDB rating of the movies starring Christian Bale higher than the average score?

4. Is there any relationship between the duration of movies and their IMDB scores?

1.3 Aim of the study

The main purpose of the study is to find factors that affect IMDB scores. To find these, this analysis interests six variable associations with IMDB scores investigated. Finding meaningful relations that positively or negatively affect IMDB scores between six variables is expected. Some tests were conducted to find the relationships, the models were displayed using plots, and the models yielded several conclusions.

2. Methodology/Analysis

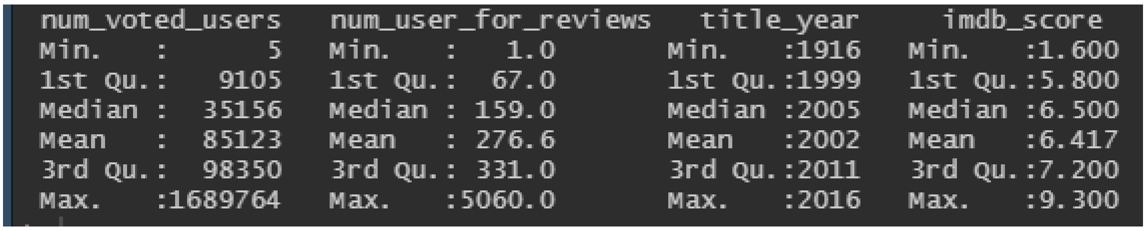
We examined the relationship between the IMDB score from our movie data set and the other observations we chose with descriptive statistics. We tried to find answers to the research questions we determined using summary statistics, spearman's test, t-test, ANOVA—Kruskal-

Wallis, and we reached specific results. We also visualized some of these results, making the data more understandable.

3. Results and Findings

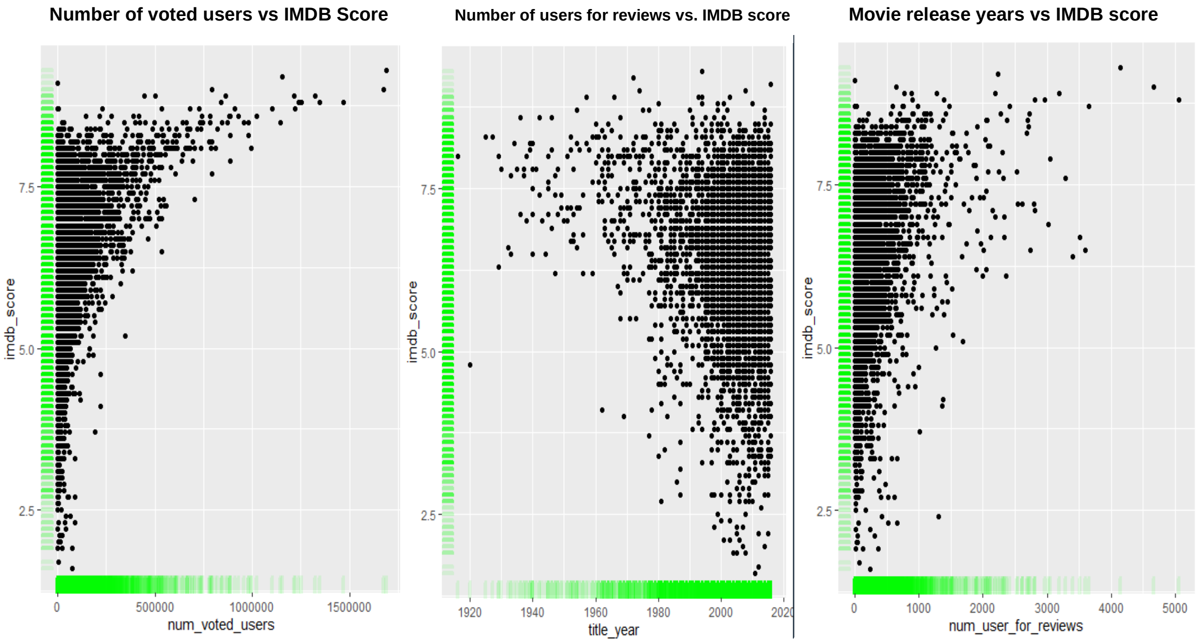
- **What is the relationship between movie release year, number of users voted film, number of users for reviews and IMDB score?**

First, we applied the summary() function to recognize the movie dataset and examine the relationship between the IMDB score and other continuous variables such as the number of voted users, number of users for reviews, movie release year.



```
 num_voted_users    num_user_for_reviews   title_year     imdb_score   
 Min.   :      5    Min.   :   1.0       Min.   :1916   Min.   :1.600  
 1st Qu.:   9105    1st Qu.:  67.0       1st Qu.:1999   1st Qu.:5.800  
 Median :  35156    Median : 159.0       Median :2005   Median :6.500  
 Mean   :  85123    Mean   : 276.6       Mean   :2002   Mean   :6.417  
 3rd Qu.:  98350    3rd Qu.: 331.0       3rd Qu.:2011   3rd Qu.:7.200  
 Max.   :1689764    Max.   :5060.0       Max.   :2016   Max.   :9.300  
```

Then, we tried to see the correlation more clearly by visualizing the scatter plot and IMDB score vs. the other three continuous variables that we determined.



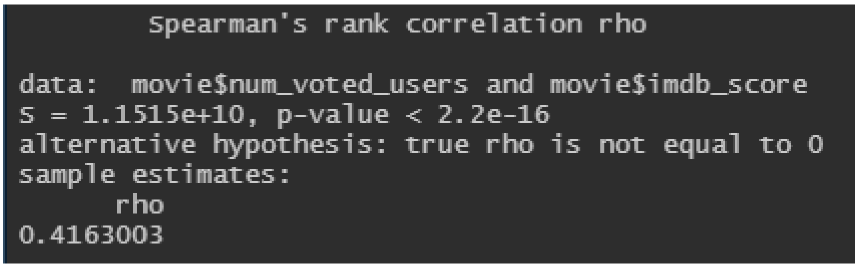
Number of voted users vs IMDB Score

Number of users for reviews vs. IMDB score

Movie release years vs IMDB score

IMDB vs. Voted, as seen from the visualized data, there is a positive linear correlation between IMDB vs. voted users. Also, there is a positive correlation between IMDB and the number of users for review. On the other hand, there is a negative correlation between IMDB vs. Movie Release Years. We examined the relationship between IMDB score and the number of voted users more closely with the correlation test, as can be seen from the graphs that there is a strong linear relationship between them.

Number of Voted Users vs. IMDB

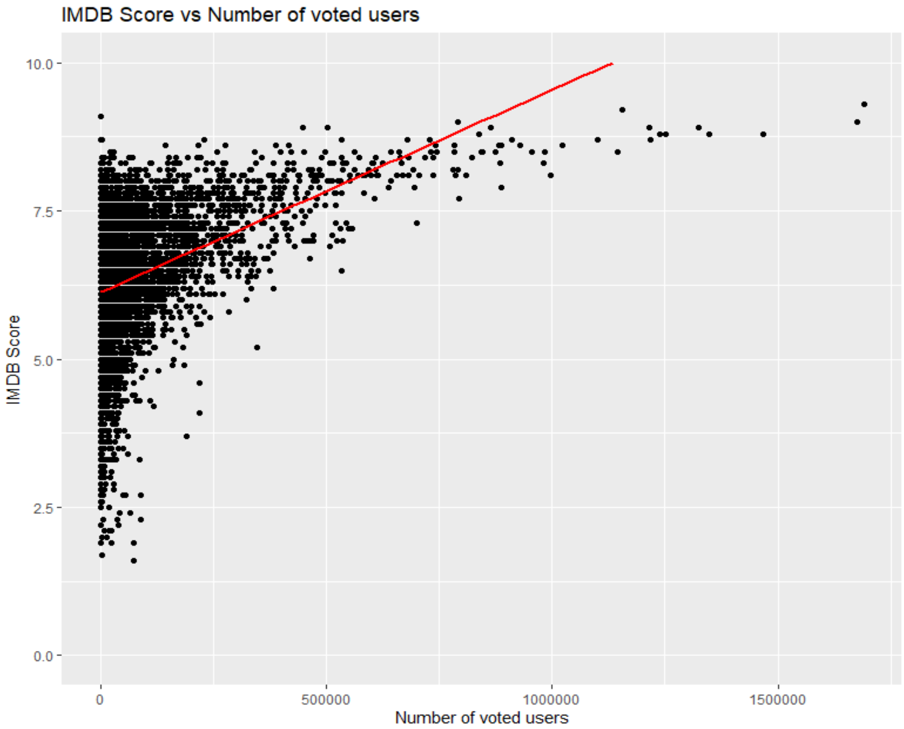


```
        Spearman's rank correlation rho

data:  movie$num_voted_users and movie$imdb_score
S = 1.1515e+10, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4163003
```

Test statistics is equal to 1.15. and alternative hypothesis tells us that Spearman rank correlation is not equal to zero. A correlation coefficient is a value between -1 and 1 and according to the sample, the correlation coefficient is equal to positive 0.42. That means there is a positive linear correlation between the number of voted users and IMDB score.

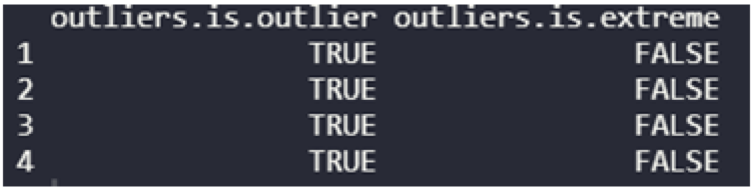
As a result of our charts and correlation tests, there is a positive correlation between the number of voted users-IMDB.



- **Are IMDB scores different for the 5 languages of movies?**

In this research question, hypothesis is different languages film IMBD scores are different. To check this, ANOVA was used. The first step is check assumption for ANOVA test which are significant outliers, independency, homogeneity and normality.

First, extreme outliers were checked.

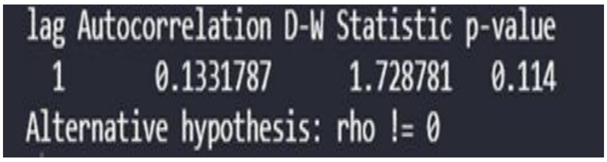


```
  outliers.is.outlier outliers.is.extreme
1                TRUE               FALSE
2                TRUE               FALSE
3                TRUE               FALSE
4                TRUE               FALSE
```

Then independency was checked with Durbin-Watson Test and hypothesis was stated like below.

H_O: There is no correlation among the residuals.

H_A: The residuals are autocorrelated.



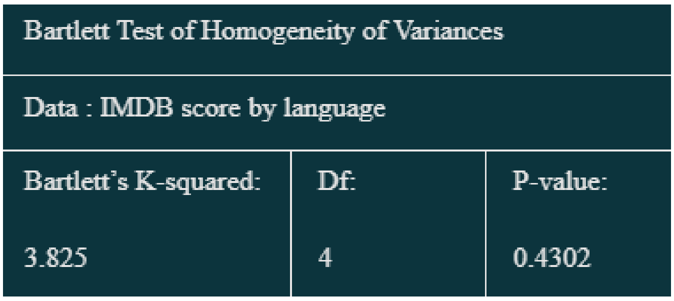
```
lag Autocorrelation D-W Statistic p-value
  1       0.1331787      1.728781   0.114
Alternative hypothesis: rho != 0
```

Then find p-value is 0.114 greater than 0.05. So, the null hypothesis was not rejected.

After it was concluded that they are independent, the homogeneity was checked with Bartlett Test and visualized. The hypothesis stated like below.

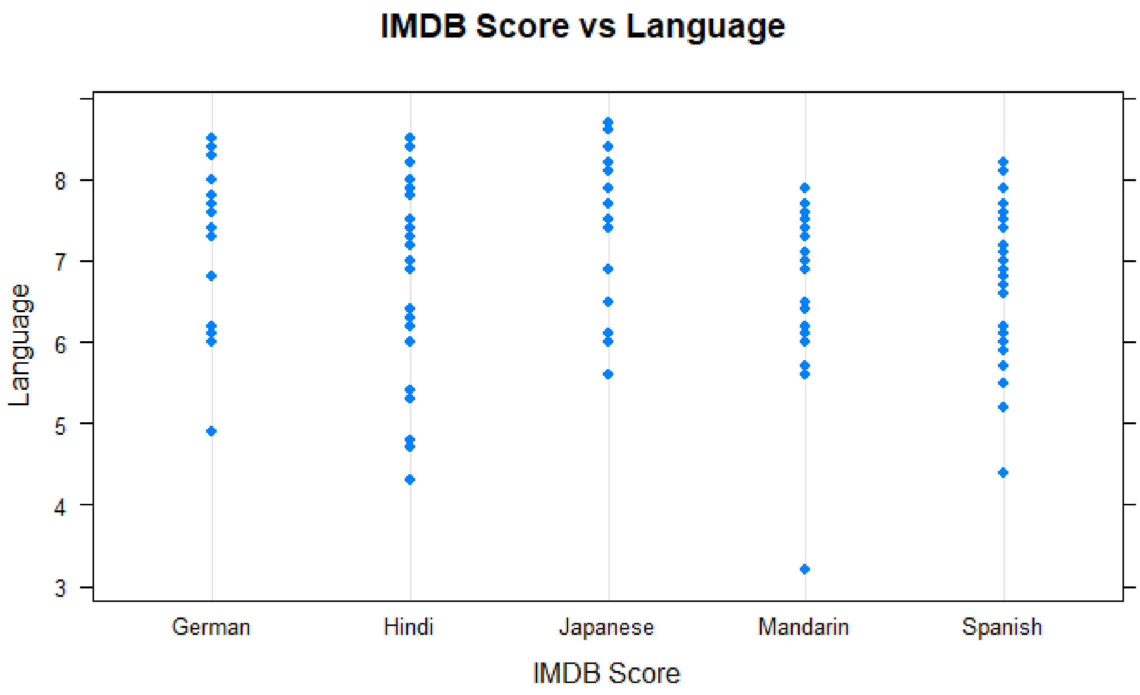
H_O: all samples' variances are equal.

H_A: at least two of them differ.



Bartlett Test of Homogeneity of Variances		
Data : IMDB score by language		
Bartlett's K-squared: 3.825	Df: 4	P-value: 0.4302

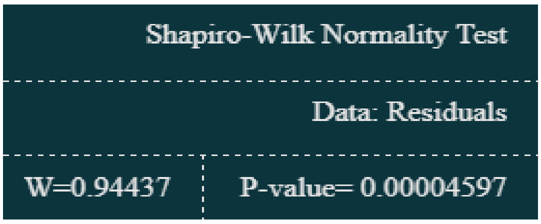
The p-value is 0.4302 greater than 0.05. The null hypothesis was not rejected. Also, homogeneity seen on the dot plot below.



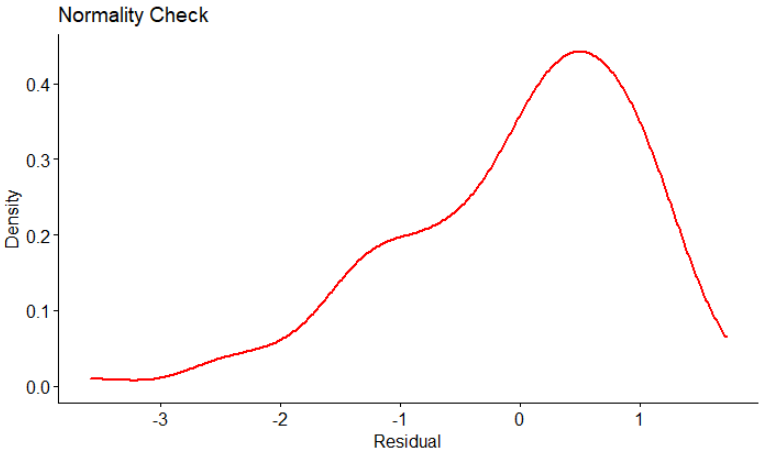
Then normality was checked with Shapiro-Wilk Test.

H_O: data come from a normal distribution.

H_A: data do not come from a normal distribution.



Shapiro-Wilk Normality Test	
Data: Residuals	
W=0.94437	P-value= 0.00004597

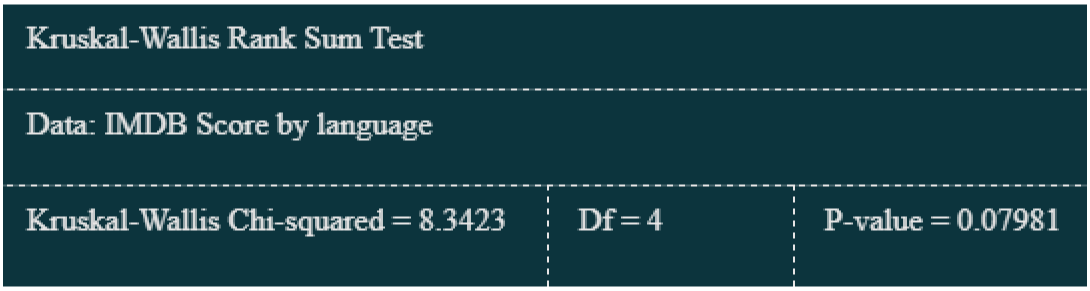


The p-value is 0.00004597 lower than 0.05. Thus, the null hypothesis was rejected. Also, data not come from a normal distribution is seen on density plot.

This means normality assumption was not reached, so the non-parametric version of the ANOVA—the Kruskal-Wallis test is needed to check this hypothesis.

H_O: The sample medians are all equal.

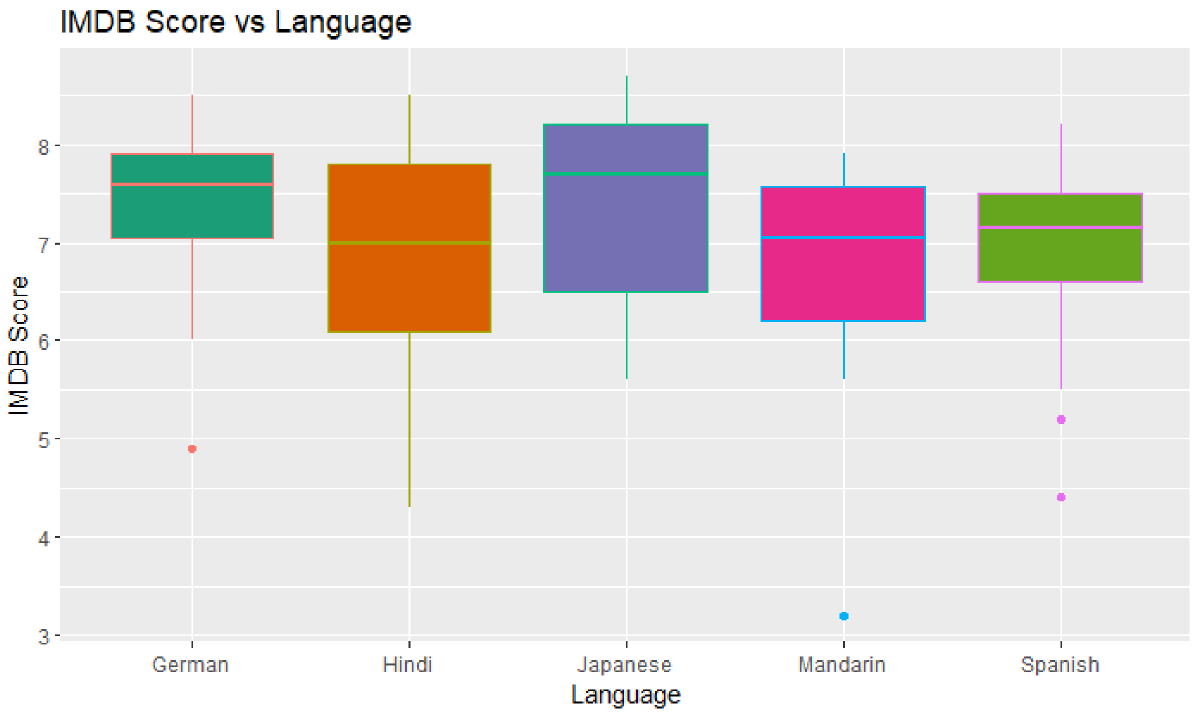
H_A: At least 1 sample median is different to others.



Kruskal-Wallis Rank Sum Test		
Data: IMDB Score by language		
Kruskal-Wallis Chi-squared = 8.3423	Df = 4	P-value = 0.07981

The p-value is 0.07981 greater than 0.05. So, the null hypothesis was not rejected.

Result reveals that the differences between the medians are not statistically significant.

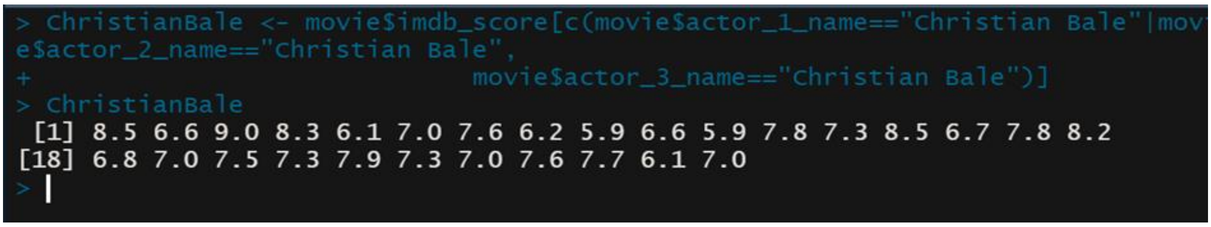


As a result, we can say that IMDB scores average distributions are not significantly different and languages is not a significant factor to affect IMDB score.

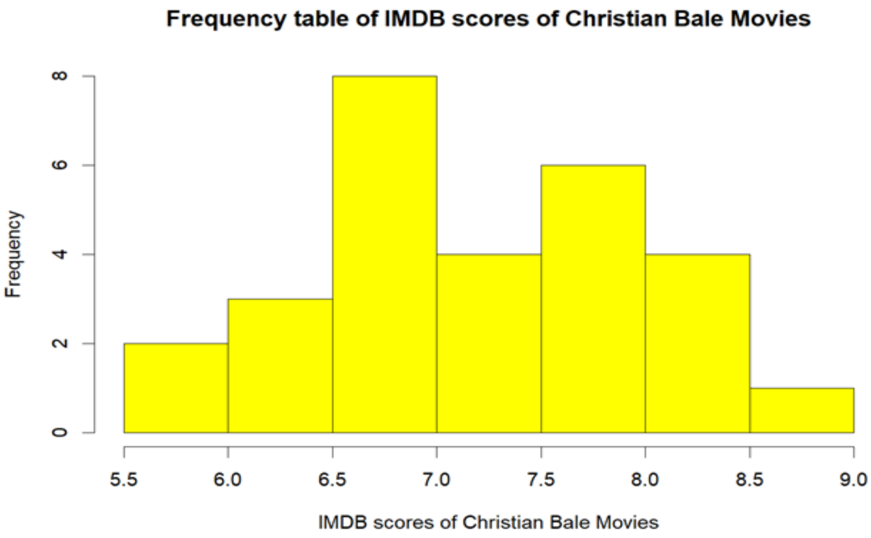
- **Is the IMDB rating of the movies starring Christian Bale higher than the average score?**

We use the t-test to compare the IMDB scores of the movies starring Christian Bale with the average IMDB scores. In our data, there are three categories of actors/actresses.

Firstly, we take IMDB scores of movies that Christian Bale is an actor in each of these three categories.



```
> ChristianBale <- movie$imdb_score[c(movie$actor_1_name=="Christian Bale"|mov
e$actor_2_name=="Christian Bale",
+                               movie$actor_3_name=="Christian Bale")]
> ChristianBale
 [1] 8.5 6.6 9.0 8.3 6.1 7.0 7.6 6.2 5.9 6.6 5.9 7.8 7.3 8.5 6.7 7.8 8.2
[18] 6.8 7.0 7.5 7.3 7.9 7.3 7.0 7.6 7.7 6.1 7.0
> |
```

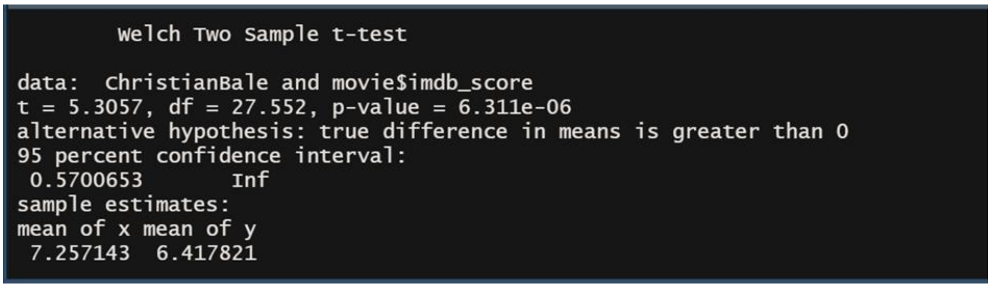


We state the hypothesis,

H_O: Christian Bale starring movies IMDB score is not significantly different from the average IMDB score.

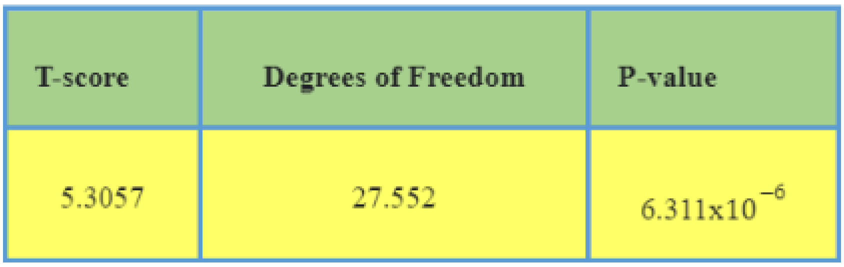
H_A: Christian Bale starring movies have higher IMDB scores than the average IMDB score.

Then we applied t-test our two samples, IMDB scores of movies that Christian Bale starred in and IMDB scores of all the movies that in our data.



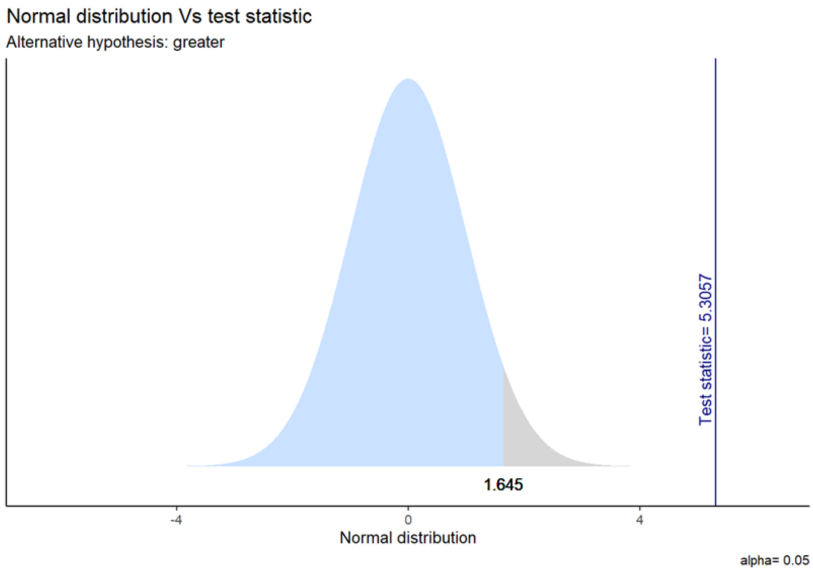
```
        Welch Two Sample t-test

data:  ChristianBale and movie$imdb_score
t = 5.3057, df = 27.552, p-value = 6.311e-06
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.5700653       Inf
sample estimates:
mean of x mean of y
 7.257143  6.417821
```



T-score	Degrees of Freedom	P-value
5.3057	27.552	$6.311\text{x}10^{-6}$

With 95% confidence interval, and critical value is 1.645 and our test value is 5.3057.



As can be seen in the table above, our test value is much higher than our critical value, (5.3037>1.645) then, we reject the null hypothesis. As a result, we can say that Christian Bale starring movies have greater IMDB scores than the average score.

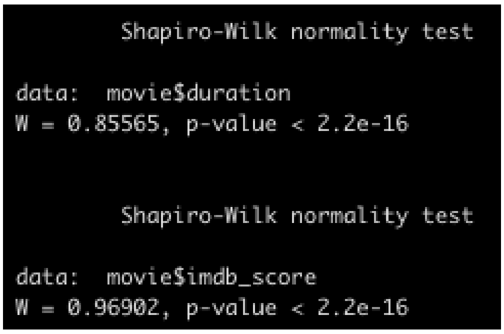
- **Is there any relationship between the duration of movies and their IMDB scores?**

In order to find out correlation between two variables, we should check the durations and normally to apply correlation test between.

We will use Shapiro-Wilk test to check their normality.

H_O: data is distributed as normal

H_A: data is not distributed as normal



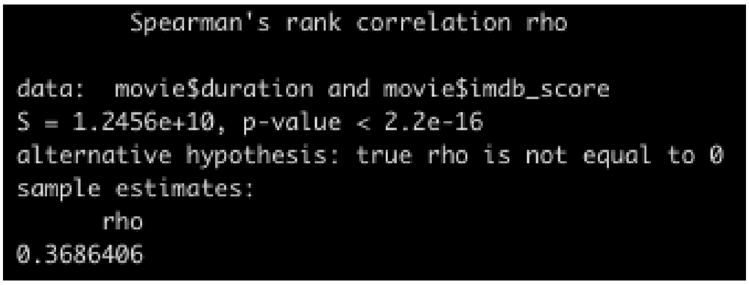
```
        Shapiro-Wilk normality test

data:  movie$duration
W = 0.85565, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  movie$imdb_score
W = 0.96902, p-value < 2.2e-16
```

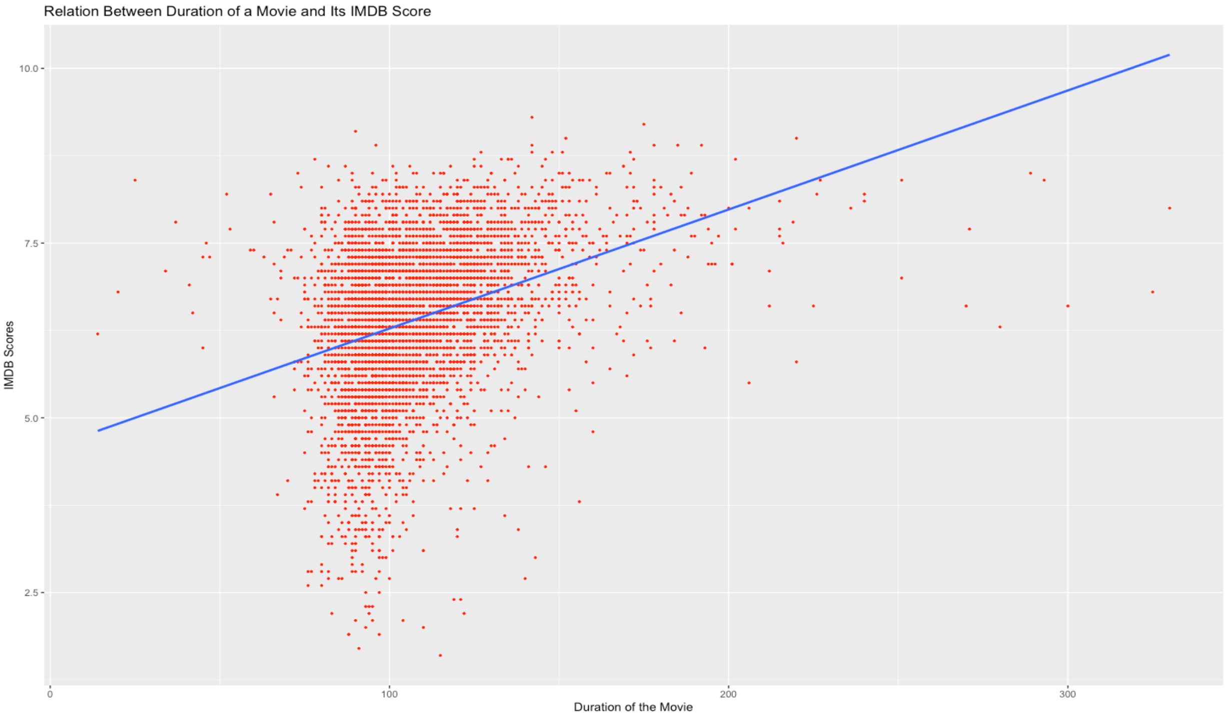
Both variable's p-values are less than 0.05 significance level. So, we reject null hypothesis which states data is distributed normal. Then we must apply Spearman correlation test, which is a nonparametric method, to see if these two variables are correlated or not.



```
        Spearman's rank correlation rho

data:  movie$duration and movie$imdb_score
S = 1.2456e+10, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3686406
```

As you can see in the graph, blue line is the regression line of the duration and the IMDB score. Although we see a stacked data of around 100 minutes, but we can visually notice that the IMDB score increases as the duration increases.



4. Discussion/Conclusion

In summary, 4 different subjects were focused on in this analysis namely: release year, language, Christian Bale movies and duration, and how they affected IMDB Score. Different statistical methods such that Correlation coefficient, Kruskal-Wallis, and T-test were applied on suitable cases. At first the properties of year variable were checked and found the descriptive statistics such that range, mean, median. In addition, with the correlation test, it was found that the IMDB score was positively correlated with the voting users, while the release date and the IMDB score were negatively correlated. Then we examined whether the IMDB score of the movies changed according to the languages. For this, we decided to apply the ANOVA test. First, our dataset must fulfill the ANOVA conditions. To apply ANOVA our dataset shouldn't have any outlier, should be independent, have equal variances and finally be normally distributed. Since our dataset is not normally distributed, we could not apply ANOVA. Using a parametric test, the Kruskal Wallis test, we found that the IMDB scores of the films did not change according to the languages. Then we investigated whether Christian Bale movies were rated higher than the average movie. As a result of the t-test, we concluded that Christian Bale movies' IMDB scores are significantly higher than the average movie. Finally, we examined the relationship between movie duration and IMDB score. We first applied a test of normality to decide which correlation test to use. After seeing that our dataset is not normally distributed, we applied the spearman correlation test, which is a nonparametric test, and as a result, we found that the movie duration and IMDB scores were positively correlated. Following the research, certain aspects of the research questions were answered, and the rest were left to be explained by further studies.

References

1. IMDB movie dataset - dataset by Himan. data.world. (2017, June 6). Retrieved January 31, 2022, from https://data.world/himan/imdb-movie-dataset/
2. https://statsandr.com/blog/anova-in-r/
3. https://ncu.libguides.com/statsresources/Spearmans

Appendices

```
library("ggplot2")
library("dplyr")
library("plotly")
library("ggpubr")
library("car")
library("rstatix")
library("lattice")
library("gginference")
movie <- read.csv("movie_data.csv")
movie <- movie[,c(7,10,13,14)]
movie <- na.omit(movie)
summary(movie)
ggplot(data=movie, aes(x=title_year, imdb_score)) +
  geom_point() +
  geom_rug(col="green",alpha=0.1, size=1.5)
ggplot(data=movie,aes(x=num_voted_users, imdb_score))
geom_point() +
  geom_rug(col="green",alpha=0.1, size=1.5)

ggplot(data=movie, aes(x=num_user_for_reviews, imdb_score))
geom_point() +
  geom_rug(col="green",alpha=0.1, size=1.5)
shapiro.test(movie$num_voted_users)
shapiro.test(movie$imdb_score)
cor.test(movie$num_voted_users,movie$imdb_score, method = 'spearman', exact=F)
```

```
ggplot(movie, aes(x=num_voted_users, y=imdb_score)) +
  geom_point()+
  geom_smooth(method=lm, se=FALSE,color="red")+ylim(0,10)+xlab("Number of voted users")+
  ylab("IMDB Score")+ggtitle("IMDB Score vs Number of voted users")
movie <- read.csv("movie_data.csv")
movie <- movie[-c(1,9)]
movie <- na.omit(movie)
ggplot(movie, aes(duration, imdb_score)) +geom_point(colour = "red", size = 0.5) +
  geom_smooth(method = "lm", se = F) + xlab("Duration of the Movie")+
  ylab("IMDB Scores")+ ggtitle("Relation Between Duration of a Movie and Its IMDB Score")
shapiro.test(movie$duration)
shapiro.test(movie$imdb_score)
cor.test(movie$duration,movie$imdb_score, method = 'spearman')
languages <- unique(movie$language)
vec_languages <- c()
for(i in 1:length(languages)){
  if(sum(movie$language==languages[i])>10){
    vec_languages <- c(vec_languages,languages[i])}}
vec_languages <- vec_languages[-c(1,3,7,8)]
movie1 <- movie[movie$language %in% vec_languages,]
durbinWatsonTest(lm(imdb_score~language,data = movie1))
outliers <- movie1 %>% group_by(language) %>% identify_outliers(imdb_score)
data.frame(outliers$is.outlier,outliers$is.extreme)
dotplot(imdb_score~language,data=movie1,xlab="IMDB Score",ylab="Language",
        main="IMDB Score vs Language")
bartlett.test(imdb_score~language,data=movie1)


res_aov <- aov(imdb_score ~ language,data = movie1)
```

```
shapiro.test(res_aov$residuals)

ggdensity(res_aov$residuals,color = "red",xlab="Residual",ylab = "Density",

          size=1,title = "Normality Check")

kruskal.test(imdb_score~language,data=movie1)

ggplot(movie1, aes(x=language, y=imdb_score,fill=language)) +

  geom_boxplot(aes(color = language))+theme(legend.position="none")+

  scale_fill_brewer(palette="Dark2")+xlab("Language")+ylab("IMDB Score")+

  ggtitle("IMDB Score vs Language")



ChristianBale<-movie$imdb_score[c(movie$actor_1_name=="Christian
Bale"|movie$actor_2_name=="Christian Bale",movie$actor_3_name=="Christian Bale")]

ChristianBale

t.test(ChristianBale,movie$imdb_score,alternative = "greater")

ggttest(t.test(ChristianBale,movie$imdb_score,alternative = "greater"))

hist(ChristianBale,main="Frequency table of IMDB scores of Christian Bale Movies",

     col="yellow", border="black", xlab = "IMDB scores of Christian Bale Movies")
```