UTM SCHOOL OF COMPUTING
Faculty of Engineering
UNIVERSITI TEKNOLOGI MALAYSIA

SECI2143 – 04 PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT REPORT 2

SECTION : 04 – 1SECR

COURSE NAME : BACHELOR OF COMPUTER SCIENCE – COMPUTER NETWORK & SECURITY

NO.	NAME	STUDENT ID
1	AZ MUKHLIS ISKANDAR BIN AZLI	A19EC0183

LECTURER : DR. SUHAILA MOHAMAD YUSUF

DATE OF SUBMISSION : 27/6/2020

I. ABSTRACT

The entertainment industry consists of film, print, radio, and television. These particular segments can be diluted and simplified into movies, TV shows, radio shows. news, music, newspapers, books and magazines. Movies, in general, have always been the behemoth, or the giant that carries the entertainment industry into the successful industry that we see today. Throughout the increasing development in technology, the accessibility of this industry has also risen quite exponentially. We can see and study the factors of how movies or films helped the entertainment industry to become as successful as they are in this day and age.

II. INTRODUCTION

Great movies often offer a memorable experience for the watchers. At the end of the day, the production of a movie does not matter much in terms of how successful it will be, because the reason why a movie is successful is depending on the people that watched it. Undoubtedly, movies and the way that it is made are evolving as time passes, and different times call for different measures. Generally, this study is being done to see the successful rate of a movie, particularly in this modern age, and to help the people that will make them to cater to these results in their future projects.

III. OBJECTIVE

The purpose of this data set is to give a better understanding and see the pattern of movies in terms of how successful it is. To be precise, the purpose of this study is to provide further information in regards to the aspects that make a movie successful from users' or profit perspective, and it can be combined with other movie datasets publicly available (RottenTomatoes, FilmTV, etc.). These aspects can be divided into:

- To test if there are any differences between the mean of the number of user reviews and the number of critic reviews for a movie.
- To see the relationship between the number of votes and the average vote.
- To study the relationship between the duration of a movie and its metascore.
- To see the dependency of the average vote towards the year that the movie was released.

IV. DATA DESCRIPTION

This dataset was collected by a user on kaggle named Stefano Leone. This data has been scraped from the publicly available website https://www.imdb.com. The secondary data that I have obtained was from a website called Kaggle:

https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset/data

The dataset can be obtained from the link above but it is a zipped file that contains 4 datasets which are IMDb movies, IMDb names, IMDb ratings and IMDb title_principals with all of them are .csv type files. However, I decided to choose IMDb movies.csv as my primary dataset and make an analysis towards this dataset only. For the dataset to work properly in Excel, I have changed the type file into .xlsx and I also removed some of the data inside the dataset for it to work properly in R Studio and for it to be analyzed effectively. The table below shows the list of all 22 variables and their measurement levels.

Population : Movies released throughout the start of the film industry
Sample : 941 movies (released from 2015 – 2019)

Target Population : Film industry

Selected Variables :

- Genre
- Average Vote
- Budget
- Worldwide Gross Income
- No. of User Reviews
- No. of Critic Reviews

Inferential statistics that will be carried out are hypothesis testing two sample, correlation, regression and chi-square test of independence.

Expected outcome :

- Hypothesis testing
  The mean of the number of users' and critics' reviews are different

- Correlation

  There is a relationship between the number of votes and the average vote

- Regression

  The longer the duration of a movie, the higher its metascore.

- Chi-Square Test of Independence

  The average vote for a movie is independent on which year it came out.

Variables	Type
Title ID	Nominal
Title	Nominal
Original title	Nominal
Year	Interval
Date	Interval
Genre	Nominal
Duration	Ratio
Movie Country	Nominal
Movie Language	Nominal
Director’s Name	Nominal
Writer’s Name	Nominal
Production Company	Nominal
Actors	Nominal
Description	Nominal
Average Vote	Ratio
No. of Votes	Ratio
Budget	Ratio
USA Gross Income	Ratio
Worldwide Gross Income	Ratio
Metascore	Ratio
No. of User Reviews	Ratio
No. of Critic Reviews	Ratio

V. RESULTS

a) Hypothesis Testing for 2 Samples

This test is done to see if there are any differences between the mean number of user reviews and the mean number of critic reviews for a movie. The population variances are unknown and are assumed to be different for both samples.

Let μ_U : Sample mean of number of user reviews

Let μ_C : Sample mean of number of critic reviews

$H_0 : \mu_U = \mu_C$

$H_1 : \mu_U \neq \mu_C$

Significance level, α = 0.05

```
> #T statistics
> t0 <- (xbar1-xbar2-0)/(sqrt((v1/n1)+(v2/n2)))
> #T Critical Value
> v <- ((v1/n1)+(v2/n2))^2/((((v1/n1)^2)/(n1-1))+(((v2/n2)^2)/(n2-1)))
> alpha <- 0.05
> t.alpha = qt(alpha/2, floor(v))
> t0
[1] 8.581587
> t.alpha
[1] -1.962263
>
```

The coding above was taken from R Studio where I did the hypothesis testing to calculate the T statistics value and T critical value. The variables xbar1 and v1 are the mean and the variance for users' reviews respectively while xbar2 and v2 are the mean and the variance for critics' reviews respectively. Furthermore, n1 and n2 are the number of movies in the dataset.

Based on the calculation in R Studio above, we can see that the test is two tailed and the value of T statistics value is t0 = 8.581587 while T critical value is t.alpha = ±1.962263.

Since the value of t0 > t.alpha, we reject null hypothesis. There is sufficient evidence to prove that the sample mean of the number of user reviews is different than the sample mean of the number of critic reviews.

b) Correlation Between The Number of Votes and The Average Vote

```
> ##Correlation Between Amount of Votes and The Average Vote
> Amount_of_Votes <- IMDb_movies$votes
> Average_Votes <- IMDb_movies$avg_vote
> cor(Average_Votes,Amount_of_Votes)
[1] 0.4285873
> 
```

To study the correlation between the number of votes and the average vote, I used the cor() function that is available in R Studio with variables Amount_of_Votes and Average_Votes as its arguments. Based on the correlation test above taken from R Studio, the correlation coefficient is 0.4285873. This means that there is relationship between the number of votes for a movie and the average vote although it is a considered as a weak relationship.

The function cor.test() was used in R Studio in order to do the correlation significance test. This test is done to see whether there is any linear correlation between the two samples stated above.

H_0 : $\rho = 0$ (no linear correlation)

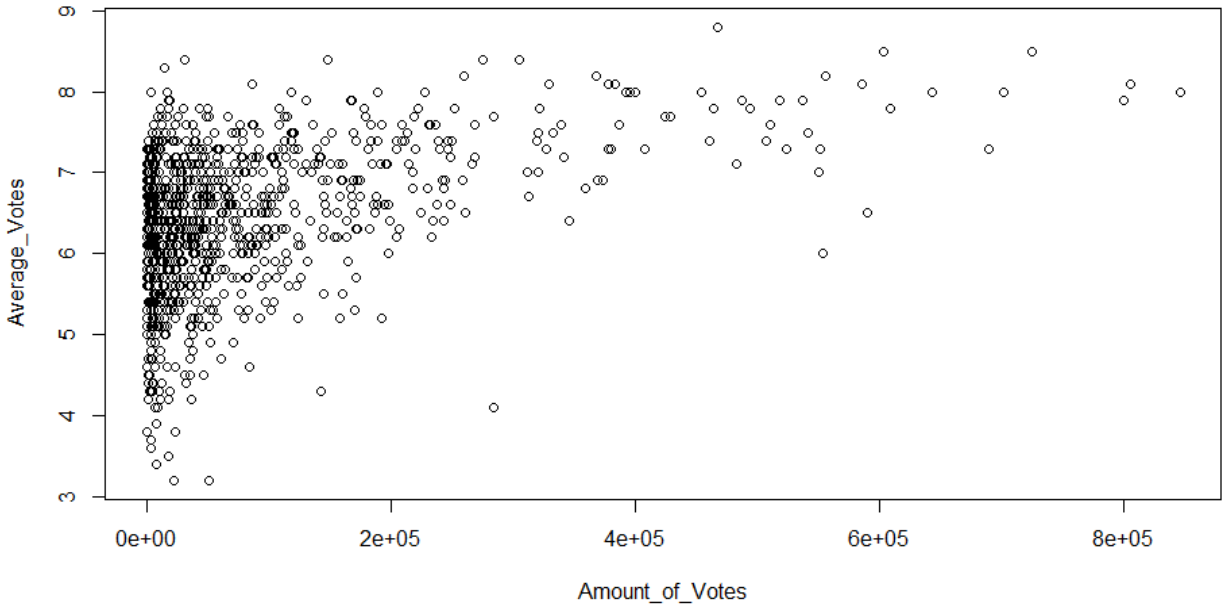
H_1 : $\rho \neq 0$ (linear correlation exists)

Significance level, $\alpha = 0.05$

```
> cor.test(Average_Votes,Amount_of_Votes)

        Pearson's product-moment correlation

data:  Average_Votes and Amount_of_Votes
t = 14.536, df = 939, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3749494 0.4793653
sample estimates:
      cor
0.4285873

> df <- 940
> alpha <- 0.05
> qt(alpha,df)
[1] -1.646476
> 
```

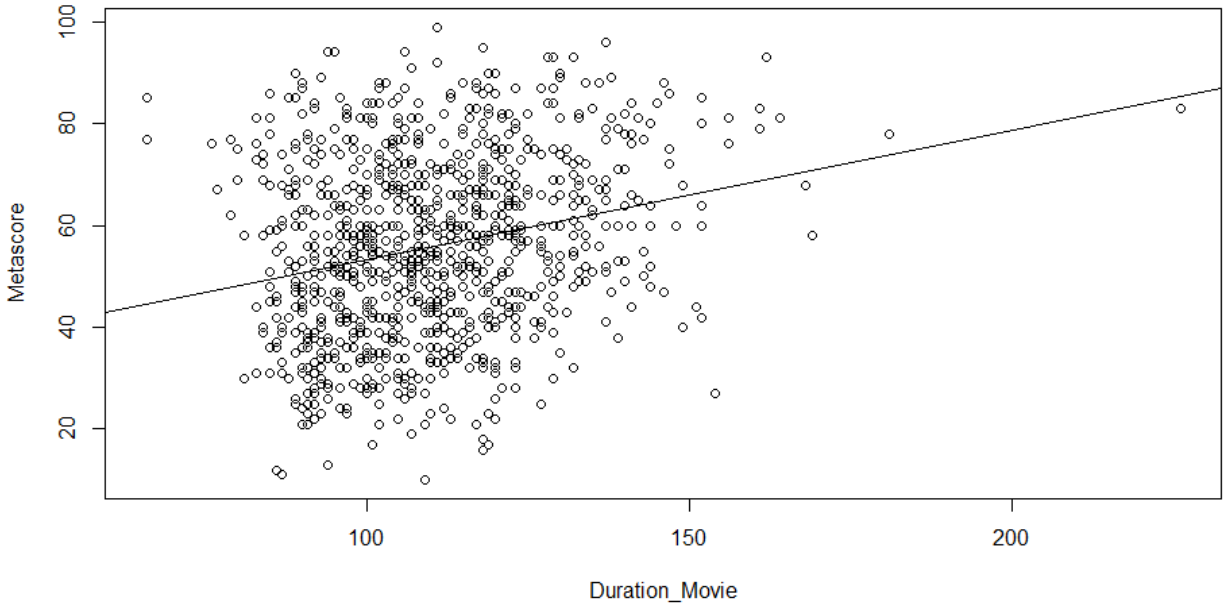
The T statistics value that was obtained from the cor.test() function is 14.536 and the critical value obtained from the qt() function is -1.646476. Since the value of T statistics value > T critical value, we reject the null hypothesis. There is sufficient evidence of a linear correlation between the number of votes for a movie and the average vote for the movie.

We can clearly see the correlation between the two samples by looking at the scatter plot for the samples.



Based on the scatter plot above, the variables Amount_of_Votes and Average_Votes clearly have a relationship and depend on each other. The points on the scatter plot seems to be following a pattern of a low positive correlation where the more the number of votes for a movie, the higher the average vote of the movie.

c) Regression between Duration of a Movie and The Metascore



The figure above shows the scatter plot and the regression line between the duration of a movie and the metascore. The regression model involves only one independent variable and it is called simple linear regression. The regression model is a positive linear model that has a straight-line relationship.

Through regression analysis, we can predict that the metascore for a movie will increase with the increase of the duration of the movie. We can find the linear equation for the best-fit line using the summary() function in R Studio.

```
> summary(model)

Call:
lm(formula = Metascore ~ Duration_Movie)

Residuals:
    Min      1Q  Median      3Q     Max
-45.567 -13.000  -0.403  12.656  42.925

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    27.91197    3.69606   7.552 1.02e-13 ***
Duration_Movie  0.25372    0.03319   7.645 5.17e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.19 on 939 degrees of freedom
Multiple R-squared:  0.05859,   Adjusted R-squared:  0.05759
F-statistic: 58.44 on 1 and 939 DF,  p-value: 5.168e-14

> |
```

Based on the result from the summary() function, we can find the regression equation to be:

$$\hat{y} = 27.91197 + 0.25372x$$

From the equation, we know that the estimated change in average metascore for a movie increase by 0.25372 point, on average, for each additional one minute of the movie. The value of intersection coefficient (27.91197) indicates the range of the metascore for a movie if the duration of the movie is 0 minutes. This can be discarded because there is no existing movie with a duration of 0 minutes.

Coefficient of determination, $R^2 = 0.05859$

From R2, we can say that only a small percentage (5.86%) of the variation in the metascore is explained by variation in the duration of a movie. This percentage also shows that the linear relationship between the metascore and the duration of a movie is weak.

H_0 : $\beta_1 = 0$ (no linear relationship)

H_1 : $\beta_1 \neq 0$ (linear relationship exists)

Significance level, α = 0.05

In addition, since the p-value obtained from the summary function (5.168e-14) is less than α (0.05), therefore we reject the null hypothesis. There is sufficient evidence to suggest that there is a linear relationship between the metascore of a movie and the duration of the movie.

d) Chi-square Test of Independence

For chi-square test of independence, two variables are tested to see whether one of them is independent on the other. Those variables are the average vote and the year the movie was released. The chisq.test() function was used in R Studio in order to find the χ^2 and the p-value.

H_0 : The average vote is independent of year.

H_1 : The average vote is not independent of year.

Significance, level α = 0.05

```
> chisq.test(d, correct = FALSE)

        Pearson's Chi-squared test

data:  d
X-squared = 210.26, df = 208, p-value = 0.443
```

From the figure above, the value of χ^2 is 210.26 and the p-value is 0.443. The critical value for χ^2 can be seen in the figure below.

```
> x2.alpha <- qchisq(alpha, df=208, lower.tail = FALSE)
> x2.alpha
[1] 242.6465
>
```

Since χ^2 statistics value = 210.26 < Critical value of χ^2 = 242.6465 at significance level of 0.05, we fail to reject the null hypothesis. We can also interpret from the p-value = 0.443 which is greater than 0.05 that indicates strong evidence for the null hypothesis. Therefore, there is enough evidence to suggest that the average vote is independent of the year the movie was released.

VI. CONCLUSION

The dataset that I had used contains a different amount of data from the original dataset that were collected by Stefano Leone. However, I would argue that my data was taken in order to keep the technicalities and calculations in R Studio to be done steadily, minimizing errors that would pop out because of certain data that could not be collected as some of the movies were released too long ago and their data is unusable. In spite of that, I am satisfied with my results as it aligns with my original objective which was to study the user's perspective and the perspective of profit on movies.

Speaking of my results, the hypothesis testing that were done towards the number of users' and critics' reviews achieved the outcome that was expected which was the mean of the samples are different. The T statistics value that was collected was greater than the positive T critical value shows that the number of users' reviews are greater than the number of critics' reviews.

As for the correlation between the number of votes and the average vote, I found that there is a weak relationship between those two variables. A conclusion can be made from this relationship where the more people watch the movie, the more people vote for it therefore making the average vote higher as it is liked by a lot of people.

Furthermore, the linear regression model displayed a relationship between the duration of a movie and its metascore. However, as we can see in the best fit line in the scatter plot, the relationship is considered as a weak relationship because the slope line is close to zero.

Lastly, the chi-square test that were done to see the dependency of the average vote of a movie towards the year when it was released resulted in the independency of the former variable towards the latter. The χ^2 that was collected was smaller than the critical value of χ^2 therefore making the average vote for a movie does not depend on the year the movie was released.

In conclusion, the reasons that make a movie successful can vary greatly when seen in different perspectives. Before making a movie, filmmakers should consider the past and see how moviegoers respond to certain types of movies in order to create a successful one. I believe that the objectives of this study were successfully achieved and I hope that more research and innovation will be done to maintain the entertainment industry as it is today.