ALY 6010: Probability Theory and Introduction to Statistics

Final Project

Name: Yijun Wang

NUID: 002950159

2021 Fall

Introduction

There should be several crucial factors for a data analyst to grasp any dataset. Is the data descriptive of a connection and easy to grasp for the user? We may try to assess the query and see if there is any link between it and the dataset.

I got this dataset from the Kaggle website, and I picked the data set that included movies from Netflix, Prime Video, Hulu, and Disney+. This topic was chosen because of covid-19. The long-term decline of movie theaters has created an opportunity for enormous expansion for different streaming media providers. The reason I'd like to study this data is because I'm curious about which streaming service individuals should use.

```
> #string
> str(df)
'data.frame':   6664 obs. of  17 variables:
 $ X              : int  0 1 2 3 4 5 6 7 8 9 ...
 $ ID             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Title          : Factor w/ 6664 levels "'71","#Alive",..: 5492 1403 1432 2985 4377 6123 5872 3825 5131 5942 ...
 $ Year           : int  2019 2016 2020 2001 2018 2018 2020 2017 2018 2020 ...
 $ Age            : Factor w/ 6 levels "","13+","16+",..: 4 5 5 5 4 2 2 2 3 4 ...
 $ IMDb           : Factor w/ 75 levels "","1.5/10","1.7/10",..: 62 68 74 65 61 55 60 57 57 62 ...
 $ Rotten.Tomatoes: Factor w/ 83 levels "","10/100","12/100",..: 83 82 81 80 80 80 79 78 78 78 ...
 $ Netflix        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Hulu           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Prime.Video    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Disney.        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Type           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Directors      : Factor w/ 4793 levels "","A-mer Faruk Sorak",..: 2844 3240 135 432 1 4313 1970 599 1331 30 ...
 $ Genres         : Factor w/ 1078 levels "","Action","Action,Adventure",..: 548 118 809 965 207 666 845 71 660 939 ...
 $ Country        : Factor w/ 781 levels "","Afghanistan,United States",..: 626 281 516 276 611 626 626 449 626 759 ...
 $ Language       : Factor w/ 727 levels "","Afrikaans",..: 247 477 65 477 65 65 65 264 136 267 ...
 $ Runtime        : int  209 161 83 224 52 99 94 120 133 129 ...
>
```

This dataset has 6664 observations and 17 columns. All of dataset are talking around four different platforms: Netflix, Hulu, Disney, and Prime. Video. The goal of this study is to examine the dataset statistically and provide relevant, in-depth findings about the dataset, including data levels, descriptive statistics, and data distribution. Rest of dataset are Movies' title, publish year, Age's level, Rotten.Tomatoes's scores, IMDb's score, Directors, Genres, Country, Language and run time. These data are composed of int and factors.

From the column of "IMDb", the number is followed by the symbol "/10." Although we recognize that this is a scoring system, it is quite cumbersome when undertaking data processing. As a result, I eliminated the "/10" and changed these as number.

```
> #Remove /100 from Rotten Tomatoes and /10 from IMDb
> df$Rotten.Tomatoes <- substring(df$Rotten.Tomatoes,0,2)
> df$IMDb <- substring(df$IMDb,0,3)
> head(df)
  X ID                                Title Year Age IMDb Rotten.Tomatoes Netflix Hulu Prime.Video Disney. Type
1 0  1                   The Irishman 2019 18+  7.8             98       1    0           0       0    0
2 1  2                        Dangal 2016  7+  8.4             97       1    0           0       0    0
3 2  3 David Attenborough: A Life on Our Planet 2020  7+    9             95       1    0           0       0    0
4 3  4        Lagaan: Once Upon a Time in India 2001  7+  8.1             94       1    0           0       0    0
5 4  5                         Roma 2018 18+  7.7             94       1    0           0       0    0
6 5  6        To All the Boys I've Loved Before 2018 13+  7.1             94       1    0           0       0    0
                                Directors                          Genres
1                          Martin Scorsese           Biography,Crime,Drama
2                            Nitesh Tiwari     Action,Biography,Drama,Sport
3 Alastair Fothergill,Jonathan Hughes,Keith Scholey           Documentary,Biography
4                       Ashutosh Gowariker             Drama,Musical,Sport
5                                           Action,Drama,History,Romance,War
6                             Susan Johnson            Comedy,Drama,Romance
                                  Country                          Language Runtime
1                           United States English,Italian,Latin,Spanish,German     209
2 India,United States,United Kingdom,Australia,Kenya,Namibia                   Hindi,English     161
3                                  United Kingdom                         English      83
```

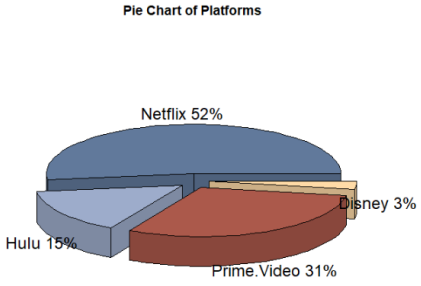
And I did same thing to “Rotten.Tomatoes”.

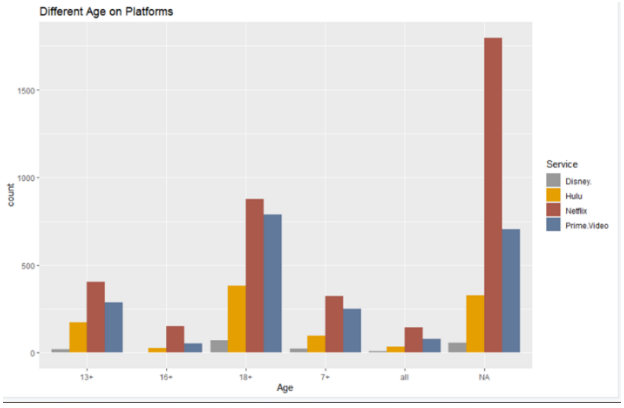
```
6                                                    United States
> #conver to number
> df$IMDb= as.numeric(df$IMDb)
> df$Rotten.Tomatoes= as.numeric(df$Rotten.Tomatoes)
> #conver to number
> df$IMDb= as.numeric(df$IMDb)
> df$Rotten.Tomatoes= as.numeric(df$Rotten.Tomatoes)
> class(df$IMDb)
[1] "numeric"
> class(df$Rotten.Tomatoes )
[1] "numeric"
>
```

Here is the summary of this dataset.

```
> #summary
> summary(df)
       X                ID                                  Title          Year          Age            IMDb      Rotten.Tomatoes
 Min.   :   0   Min.   :   1   '71                               :   1   Min.   :1915      :2747   Min.   :1.50   Min.   :10.00
 1st Qu.:1666   1st Qu.:1667   #Alive                            :   1   1st Qu.:2010   13+: 822   1st Qu.:5.60   1st Qu.:49.00
 Median :3332   Median :3332   #AnneFrank. Parallel Stories      :   1   Median :2016   16+: 226   Median :6.40   Median :56.00
 Mean   :3205   Mean   :3206   #cats_the_mewvie                  :   1   Mean   :2011   18+:1972   Mean   :6.29   Mean   :56.62
 3rd Qu.:4997   3rd Qu.:4998   #FriendButMarried                 :   1   3rd Qu.:2019   7+ : 646   3rd Qu.:7.10   3rd Qu.:64.00
 Max.   :5366   Max.   :5367   #FriendButMarried 2               :   1   Max.   :2021   all: 251   Max.   :9.20   Max.   :98.00
                               (Other)                           :6658                             NA's   :113    NA's   :7
    Netflix           Hulu          Prime.Video       Disney.           Type                    Directors
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0                       : 246
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0   RaÃºl Campos,Jan Suter:  22
 Median :1.0000   Median :0.0000   Median :0.0000   Median :0.00000   Median :0   Jay Chapman           :  19
 Mean   :0.5545   Mean   :0.1571   Mean   :0.3244   Mean   :0.02746   Mean   :0   Jay Karas              :  17
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0   Marcus Raboy          :  17
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :0   Cathy Garcia-Molina   :  11
                                                                                  (Other)               :6332
                Genres                Country                Language       Runtime
 Comedy             : 575   United States :2779   English        :3489   Min.   :  3.0
 Drama              : 444   India         : 710   Hindi          : 344   1st Qu.: 89.0
 Documentary        : 344   United Kingdom: 253                   : 186   Median : 98.0
 Comedy,Drama       : 275                 : 159   Spanish        : 185   Mean   :100.1
 Drama,Romance      : 224   Canada        : 142   English,Spanish: 154   3rd Qu.:112.0
 Comedy,Drama,Romance: 218   Japan         : 105   Japanese       :  98   Max.   :359.0
 (Other)            :4584   (Other)       :2516   (Other)        :2208   NA's   :195
>
```

Data Analysis:



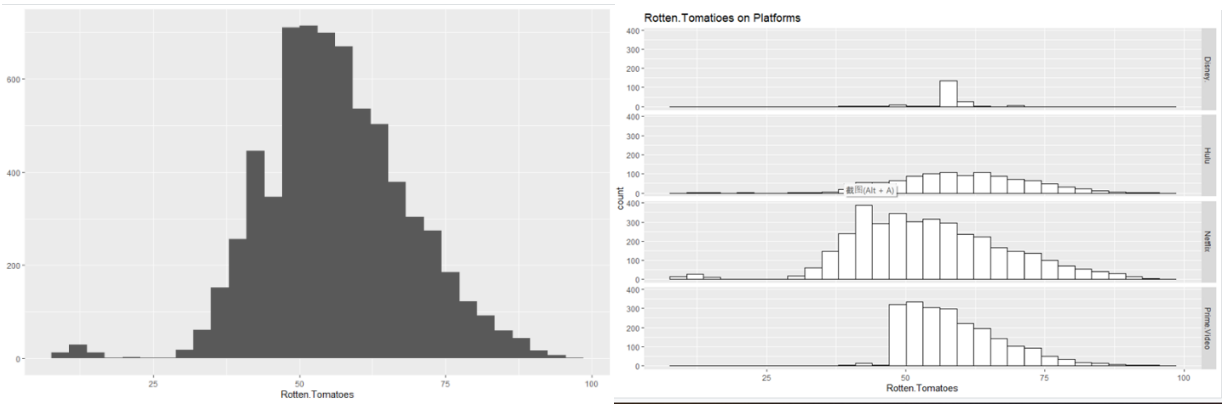


We can see from the above two perspectives the proportion of movies played on each platform and the distribution of the number of movies played by different age groups in this dataset.

Question 1: The mean of Rotten.Tomatoes is greater equal to 55 or less than 55?

(Ho) **Null Hypothesis**: The mean of Rotten.Tomatoes >= 55

(Ha) **Alternate Hypothesis**: The mean of Rotten.Tomatoes < 55



The one on the left is the bar chart of Rotten.Tomatoes, and the one on the right is the bar chart of Rotten.Tomatoes for each platform. From the graph we can see that the main distributions are all between 50 and 75. Then I am going to use t-test to test the relationship between them.

```
> t.test(df$Rotten.Tomatoes,alternative="two.sided", Var.equal=TRUE,conf.level = 0.95)

        One Sample t-test

data:  df$Rotten.Tomatoes
t = 391.48, df = 7079, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 56.52514 57.09407
sample estimates:
mean of x
  56.8096

>
```

95% Confidence interval for difference is 56.52514,57.09407, difference is 7079. T-value is 391.48, P-value <2.2e-16.

My null hypothesis is the mean of Rotten.Tomatoes >=55, and the alternative hypothesis is less than 55. From t-test, I got t-value 391.48, the p-value is <2.2e-16 which is less than the alpha value 0.05. So, we can reject the null hypothesis. In other words, although the ratings of many movies are distributed in the 50-75 range, it may be because some movies are rated so low that the average score is pulled down. Therefore, the average score of Rotten Tomatoes is less than 55.

Question 2: "16+"&"all “has any difference on Rotten. Tomatoes score?

(Ho) **Null Hypothesis**: The Difference is equal to 0.

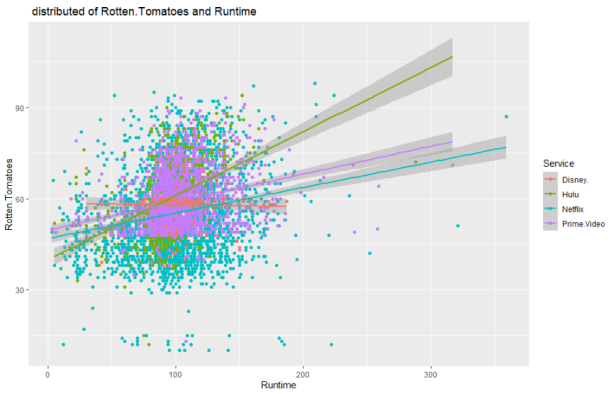
(Ha) **Alternate Hypothesis**: The Difference is not equal to 0.

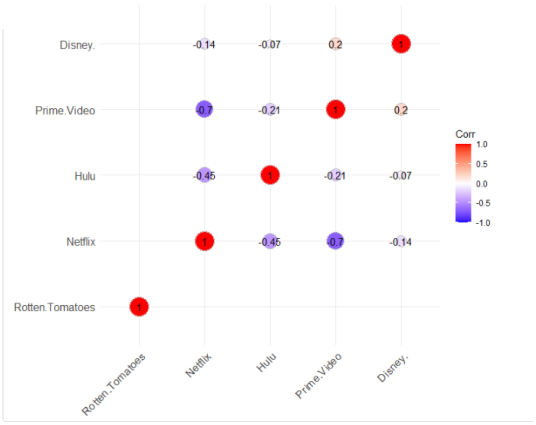
```
> t.test(A$Rotten.Tomatoes, B$Rotten.Tomatoes,alternative="two.sided", Var.equal=TRUE,conf.level = 0.95)

        Welch Two Sample t-test

data:  A$Rotten.Tomatoes and B$Rotten.Tomatoes
t = 3.6587, df = 487.2, p-value = 0.0002812
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.707981 5.670490
sample estimates:
mean of x mean of y
 57.85652  54.16729

>
```

I use two sample t-tests to check whether there is a score difference between the "16+" and "all" movies. From t-test, I got t-value 3.6587, the p-value is 0.0002812which is less than the alpha value 0.05. So, we can reject the null hypothesis. In other words, the difference between “16+” and “all” is not equal to zero.

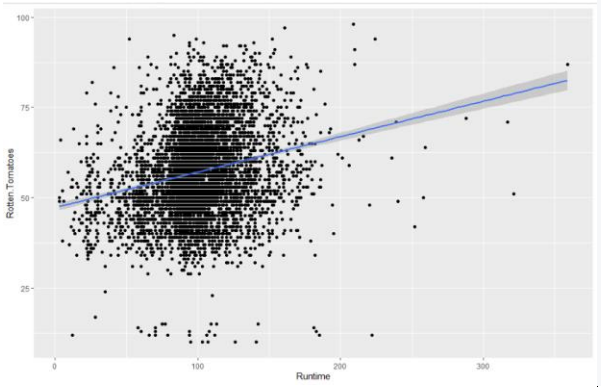
Question 3: Is there a correlation between Runtime and Rotten Tomatoes scores?

I found that the highest Rotten Tomatoes score was 98 out of 100, while the lowest was 10 out of 100. The highest number of runtimes was 359, while the lowest was only 3. My guess is that the more runtimes, the higher the Rotten Tomatoes score. So, I made a scatter plot to show the relationship between Rotten Tomatoes and runtime, and used different colors to express the different platforms.





The two graphs above show the distribution of Rotten Tomatoes scores and runtime for different platforms. In order to test my hypothesis, I performed regression analysis on rotten tomatoes and runtime.



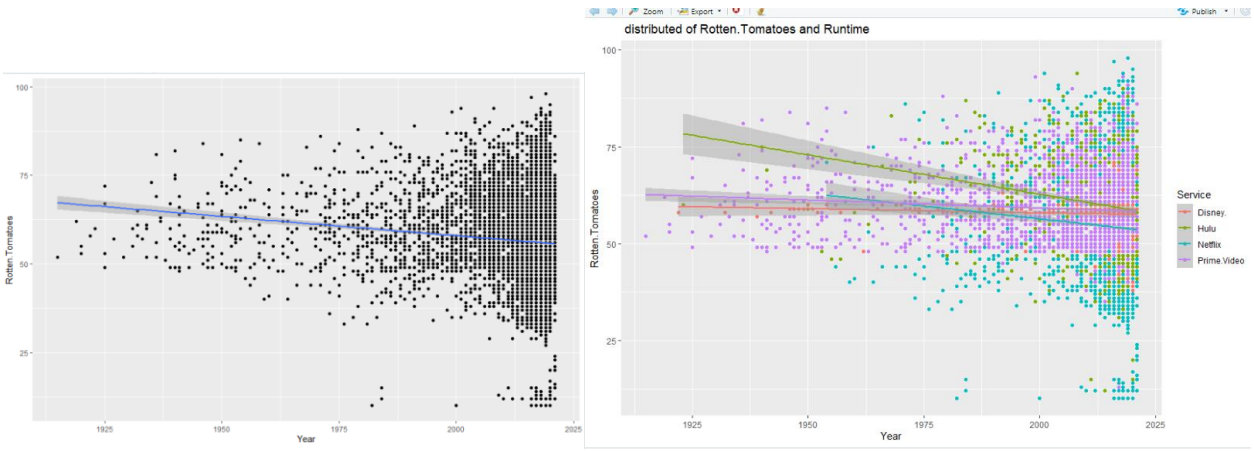

```
> #regression table of Rotten.Tomatoes and runtime
> runtime_Lm=lm(df$Rotten.Tomatoes~df$Runtime,data=df)
> runtime_Lm

Call:
lm(formula = df$Rotten.Tomatoes ~ df$Runtime, data = df)

Coefficients:
(Intercept)   df$Runtime
   47.29389      0.09817

> summary(runtime_Lm)

Call:
lm(formula = df$Rotten.Tomatoes ~ df$Runtime, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-57.087  -7.307  -0.540   7.216  41.601

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.293894   0.554395   85.31   <2e-16 ***
df$Runtime   0.098167   0.005347   18.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.6 on 6464 degrees of freedom
  (198 observations deleted due to missingness)
Multiple R-squared:  0.04955,   Adjusted R-squared:  0.04941
F-statistic:   337 on 1 and 6464 DF,  p-value: < 2.2e-16

>
```

The red color indicates a positive correlation, while the purple color indicates a negative correlation. The closer the color is to red, the more closed the relationship is. From the graph we can see that the colors of Prime Video and NETFLIX are closer to the negative correlation, which means that Prime Video and NETFLIX have the least overlapping movies. The color of Disney and Prime Video is lighter, but still red, which means that the relationship between Disney and Prime Video is closer than the others.

This regression equation's P- value is less than 2.2e-16, which is less than the alpha value 0.05. So, we can reject the null hypothesis. Also, the R-squared is 0.04941, indicating that there is a weak variation between the variables. Multiple R-squared is bigger than Adjusted R-squared 0.00014. Because the relative error is modest, the equation is considered accurate and may be used to generate predictions about future scenarios. In summary, I think there is a positive correlation between Rotten Tomatoes and runtime. That is, the more viewers may find the movie good, so they are willing to watch it more times, and thus the higher the score.

Question 4: Is there a correlation between Year and Rotten Tomatoes scores?

While compiling the data, I wondered if there could be a correlation between year and Rotten Tomatoes score.



The two charts above, the one on the left represents the year of the movie and the Rotten Tomatoes score, while the one on the right is a little more detailed, comparing the year of the movie from each platform to the Rotten Tomatoes score.

```
> #regression table of Rotten.Tomatoes and year
> year_Lm=lm(df$Rotten.Tomatoes~df$Year,data=df)
> year_Lm

Call:
lm(formula = df$Rotten.Tomatoes ~ df$Year, data = df)

Coefficients:
(Intercept)      df$Year
   276.6497      -0.1093

> summary(year_Lm)

Call:
lm(formula = df$Rotten.Tomatoes ~ df$Year, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-49.953  -7.454  -0.673   7.453  42.093

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 276.649695  20.043198   13.80   <2e-16 ***
df$Year      -0.109333   0.009968  -10.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.11 on 7078 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.01671,   Adjusted R-squared:  0.01657
F-statistic: 120.3 on 1 and 7078 DF,  p-value: < 2.2e-16

> |
```

This regression equation’s P- value is less than 2.2e-16, which is less than the alpha value 0.05. So, we can reject the null hypothesis. The R squared for this graph is 0.01657. R squared can be called the coefficient of determination. It can measure the fitness of our model because R squared is analyzing the differences in one variable that a difference in another variable can explain.

There will be a negative correlation between the two variables for this question. We can speculate that in these days of rapidly developed networks, the pace of movie releases has

become rapid, leading to varying quality of content and increasingly low ratings. The movies of long years are concerned with quality not pursuing the speed of filming, so they are almost always classics with high ratings. In another words, the rating will decrease year by year along with the increasing number of movies.

Conclusion:

Overall, this is a very good dataset. Through this dataset, I can understand that if I need to buy service membership, Netflix will be the first choice. On the one hand, it is because Netflix shows a lot more movies than other platforms, regardless of the number of movies from any age group. On the other hand, Netflix's movie ratings are relatively average and stable, with no very low scores and no very high scores. The second point is that Rotten Tomatoes' ratings are more comprehensive and unbiased. This data set has a total of 6664 observations, and Rotten Tomatoes only has 7 data misses, so it is informative. The third point is that, through the positive correlation between runtime and Rotten Tomatoes, the more runtime a movie has, the higher the rating, and the more worthy it is for people to watch.

Bibliography

Schork, J. (2020, 11 20). *Change Colors of Axis Labels & Values of Base R Plot (2 Examples)*. Retrieved from statisticsglobe: https://statisticsglobe.com/r-change-colors-axis-labels-values-of-plot

Zach. (2021, 02 04). *Format Numbers as Percentages in R (With Examples)*. Retrieved from statology: https://www.statology.org/percentage-in-r/

ZACH. (2021, 04 21). *How to Create Relative Frequency Tables in R*. Retrieved from statology: https://www.statology.org/relative-frequency-table-in-r/