**ALY 6010: Probability Theory and Introduction to Statistics**

**Final Project**

Name: Yijun Wang

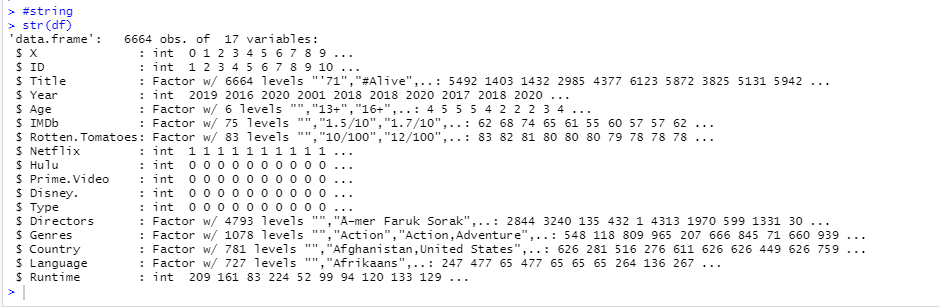
NUID: 002950159

2021 Fall

**Introduction**

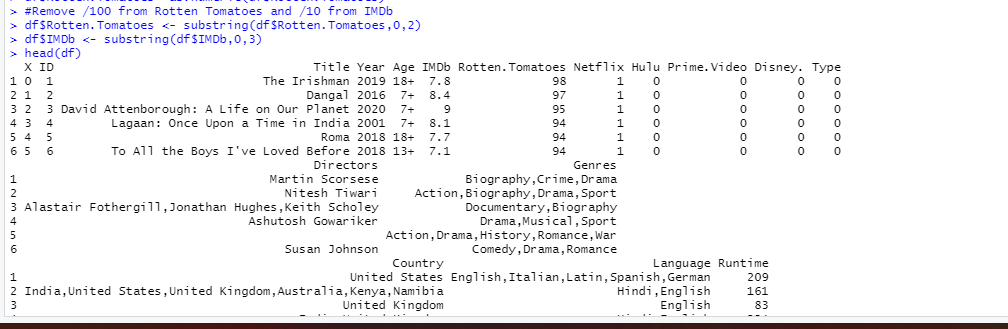
There should be several crucial factors for a data analyst to grasp any dataset. Is the data descriptive of a connection and easy to grasp for the user? We may try to assess the query and see if there is any link between it and the dataset.

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.

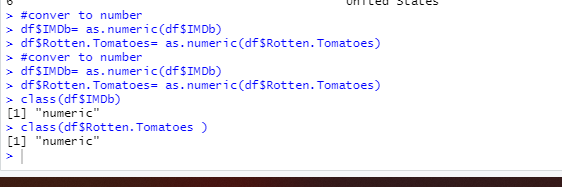


This dataset has 6664 observations and 17 columns. All of dataset are talking around four different platforms: Netflix, Hulu, Disney, and Prime. Video. The goal of this study is to examine the dataset statistically and provide relevant, in-depth findings about the dataset, including data levels, descriptive 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.

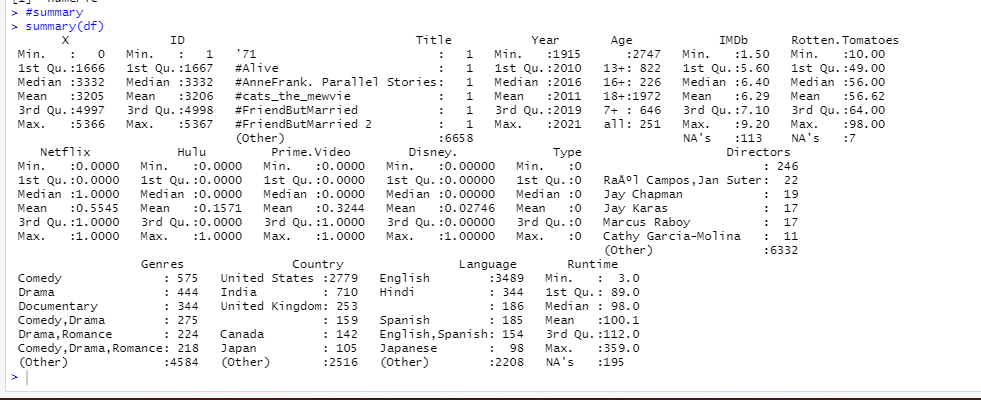
From the column of “IMDb”, the number is followed by the symbol "/10." Although we recognize that this is a scoring system, it is quite cumbersome when undertaking data processing. As a result, I eliminated the "/10" and changed these as number.



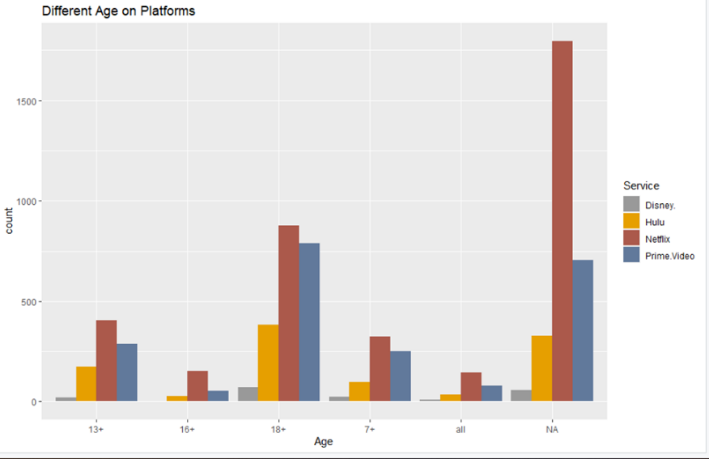
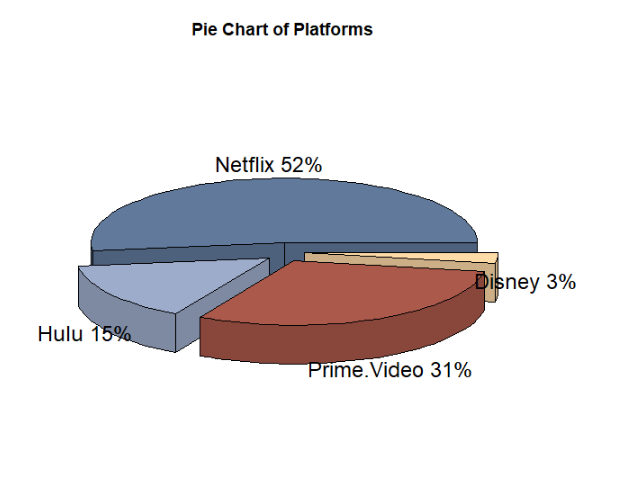
And I did same thing to “Rotten.Tomatoes”.



Here is the summary of this dataset.



**Data Analysis:**

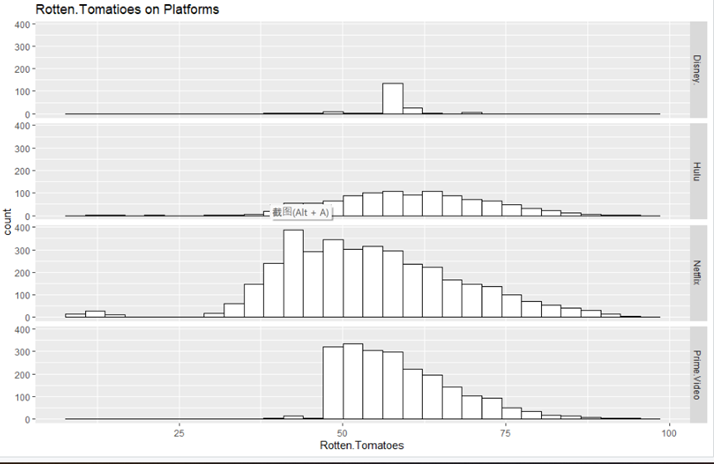
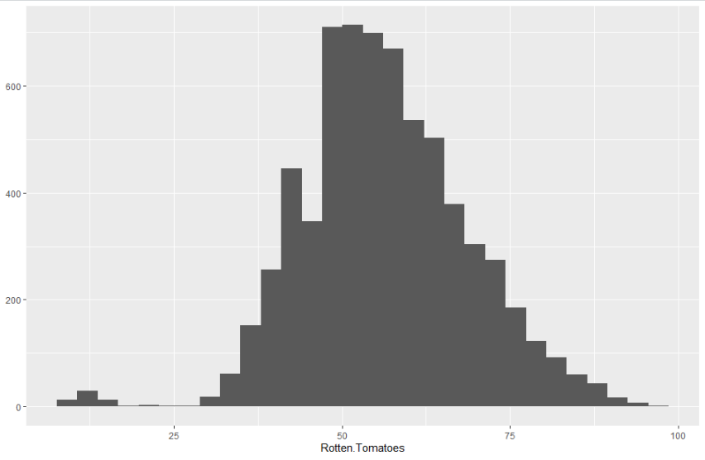


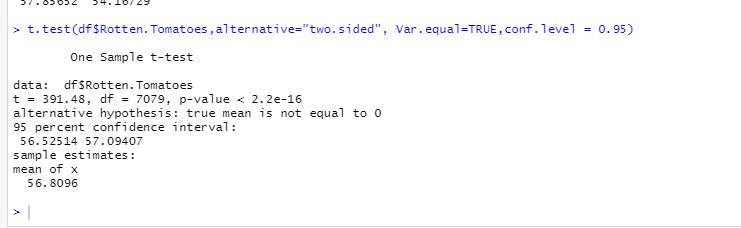
We can see from the above two perspectives the proportion of movies played on each platform and the distribution of the number of movies played by different age groups in this dataset.

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.

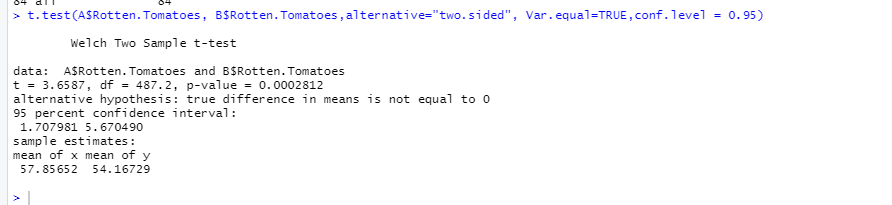
95% Confidence interval for difference is 56.52514,57.09407, difference is 7079. T-value is 391.48, P-value <2.2e-16.

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.

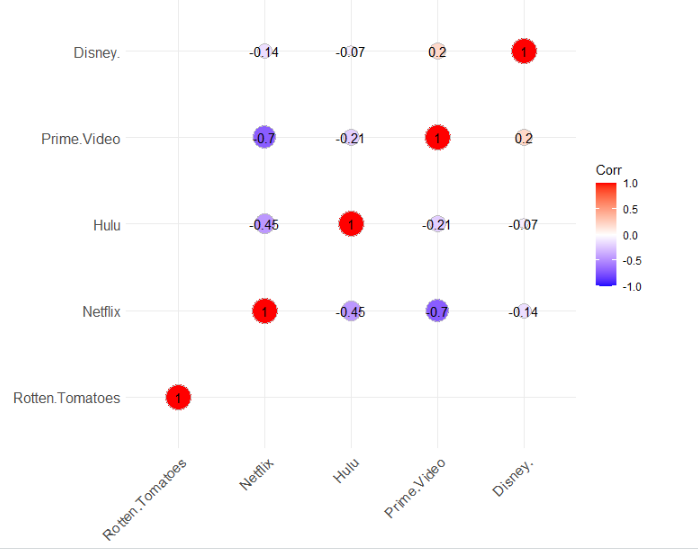
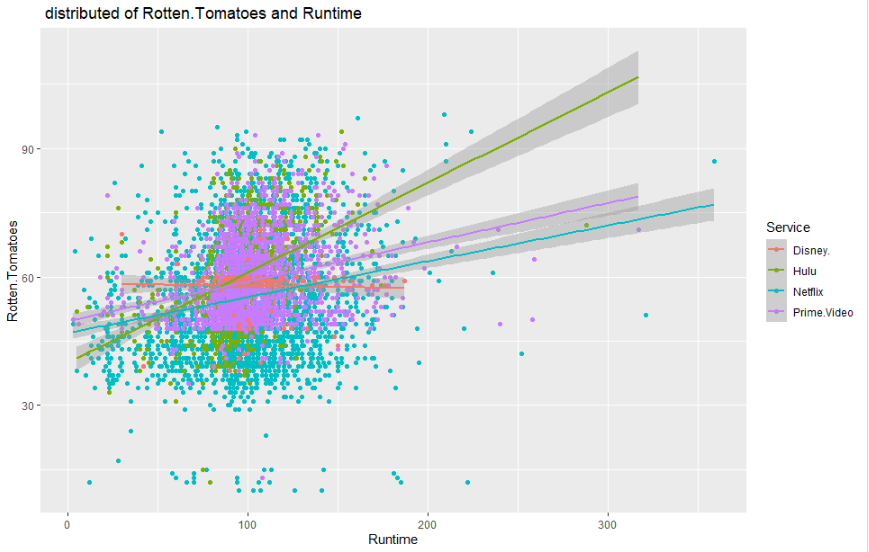
(Ha) **Alternate Hypothesis**: The Difference is not equal to 0.



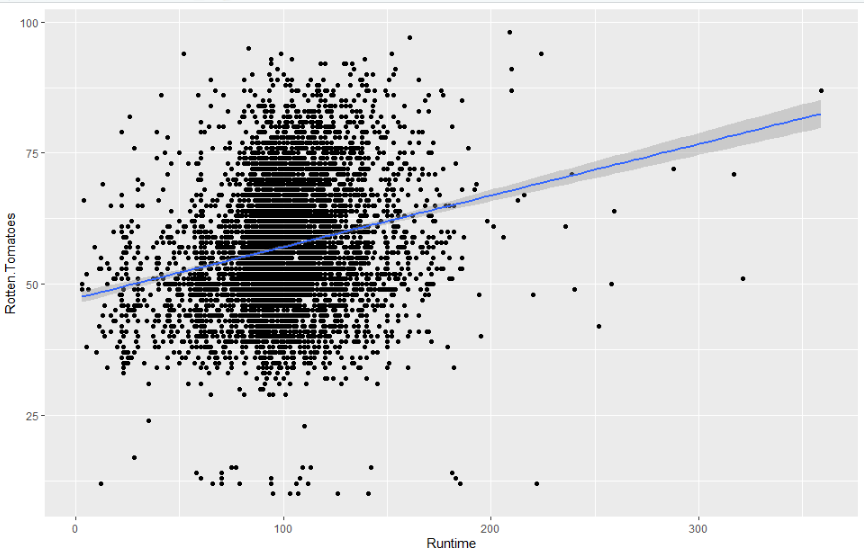
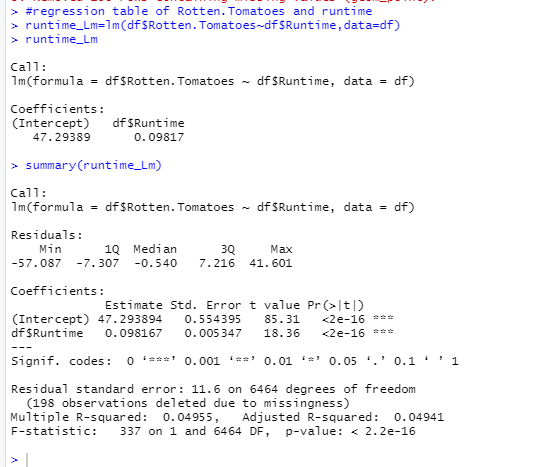
I use two sample t-tests to check whether there is a score difference between the "16+" and "all" movies. From t-test, I got t-value 3.6587, the p-value is 0.0002812which is less than the alpha value 0.05. So, we can reject the null hypothesis. In other words, the difference between “16+” and “all” is 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.

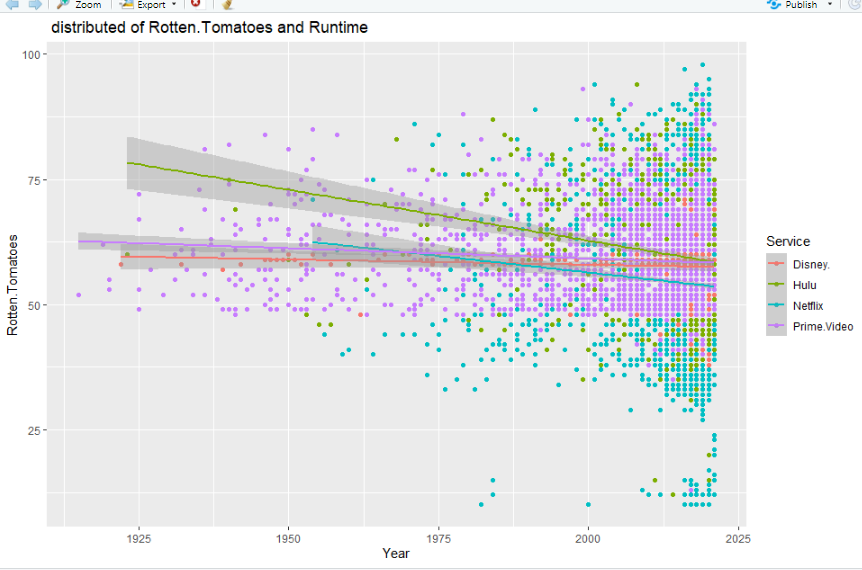
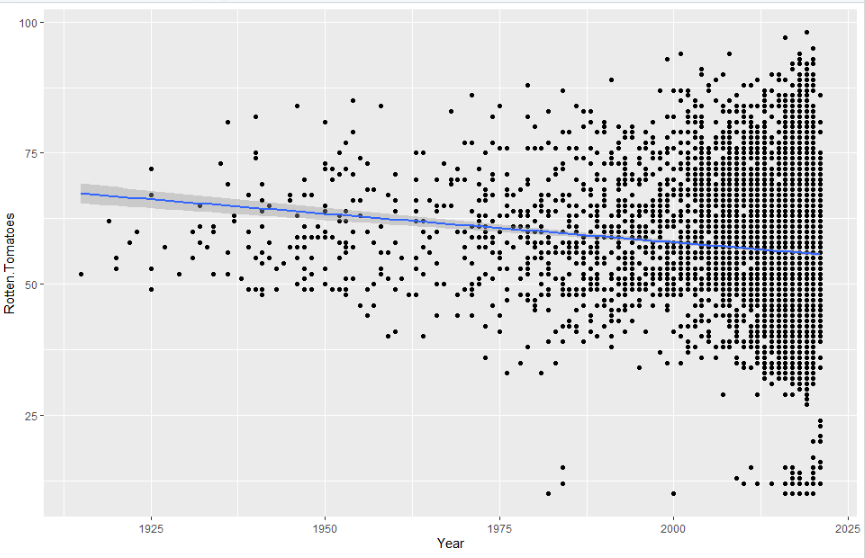
 

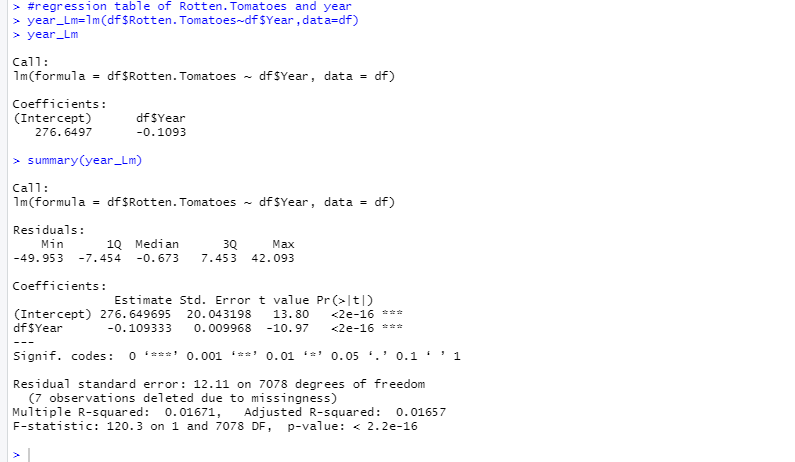
The red color indicates a positive correlation, while the purple color indicates a negative correlation. The closer the color is to red, the more closed the relationship is. From the graph we can see that the colors of Prime Video and NETFLIX are closer to the negative correlation, which means that Prime 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.

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.

