The Popularity of TV Shows on Streaming Services

Jenny Chin, Vivianne Huang, Sayaka Inatome

AMS 325: Computing and Programming Fundamentals

Dr. Xiangmin Jiao

Project Objectives

Introduction

For our project, we chose to analyze data of TV shows on various streaming services. Streaming services have been really popular in the past couple of years, especially during the pandemic. We all had a mutual interest in watching different TV shows, so we focused on this interest. Our dataset from Kaggle shows a wide variety of TV shows and includes four popular streaming services: Netflix, Hulu, Prime Video, and Disney+.

We all collaborated on the reports and wrote them together, combined our findings to form a conclusion, and collaborated to achieve our goals in our project. For further reference in the report, each person wrote about their set of tasks. Each of these questions in the rest of our report will be referenced by the number listed below in the goals.

Goals

For our project, we are trying to answer the following question: which streaming service is best overall to subscribe to in terms of TV shows, and whether or not there are patterns in Netflix, Hulu, Prime Video, and Disney+ with what they include? To answer our main question, we have created some tasks to guide us. Here are the following tasks:

1. Is there a correlation between IMDb ratings and Rotten Tomatoes ratings?
2. What, if any, is the correlation between Rotten Tomatoes ratings with whether they are on a streaming service?
3. Which streaming service offers the most variety of TV age ratings?
4. What shows are on multiple streaming services? Two or more? Three or more?
5. Does the date of release have an effect on whether or not the TV show is on a streaming service?
6. Display information about a particular show based on the information in our dataset.

We divided each of these tasks evenly and wrote about our own respective tasks in this report; Vivianne worked on tasks 1 and 2, Jenny worked on tasks 3 and 4, and Sayaka worked on tasks 5 and 6.

Significance & Expected Outcomes

Task 1

Significance: We want to see whether these two sites have similar opinions on the same TV shows.

Expected Outcomes: There is a positive moderate to strong correlation.

Task 2

Significance: We want to see whether streaming services take ratings into consideration.

Expected Outcomes: Netflix, Hulu, and Prime Video will have more highly rated shows. For Disney+, there will most likely not be a correlation as it matters more about branding for them.

Task 3

Significance: We are attempting to find out which streaming service is most appropriate for each household (whether they are a family with kids or a couple etc.).

Expected Outcomes: The expectation is that Netflix will have shows appropriate to the widest range of age groups.

Task 4

Significance: Potential customers can decide which streaming platforms to subscribe to based on whether or not the platform has their favorite shows available and compare them to each other.

Expected Outcomes: We expect multiple shows to be on various streaming platforms, especially if they are popular with a large fan base.

Task 5

Significance: We want to see how many streaming services include shows from certain years to see if there are any patterns.

Expected Outcomes: We expect there to be more recent shows on streaming services. (2020, 2019…) Lately, more “original” shows have been coming out and due to licensing, there should be fewer older shows as back then, the norm used to be watching shows on television instead of streaming services.

Techniques & Tools

Software & Packages

Our project is written using Python, specifically on Jupyter Notebook. We also used Github, linked in the references, which was the easiest platform for us to collaborate. Packages that we are using in this project include Pandas for dataframes, Matplotlib for plotting, Scipy, and Numpy.

Methods

To complete each task, the data was downloaded from Kaggle. The dataset, shown in Figure 0, has 5,368 unique data points and utilizes nine variables: Title, Year, Age, IMDb, Rotten Tomatoes, Netflix, Hulu, Prime Video, and Disney+. Title is the title of the TV show. Year is the year in which the TV show was first produced. Age represents the target age group audience for the show. IMDb shows the IMDb rating of the show; the ratings are out of ten. Rotten Tomatoes shows the Rotten Tomatoes rating of the show; this rating is out of 100. The Netflix, Hulu, Prime Video, and Disney+ columns consist of binary variables. A one means that the show is on that streaming service; a zero means that the show is not on that streaming service.

Task 1

We first imported our dataset as a dataframe using Pandas and named it "df_data". We checked to see if there were any missing values in the dataset and used `df_data.dropna()`. Because the rating values are inputted as strings, the IMDb and Rotten Tomatoes columns were converted into floats and integers, respectively. After that, we continued to use Pandas function of `.corr()` to find the correlation between the IMDb ratings and the Rotten Tomatoes ratings. To have a visualization of the ratings, we plotted the ratings against each other using `matplotlib.pyplot()` where IMDb ratings are on the x-axis and Rotten Tomatoes ratings are on the y-axis as seen in Figure 1.

Task 2

For the second task, we went back to our original dataset without dropping the missing values. We found that none of the Rotten Tomatoes ratings column contains missing values so we decided to only work with Rotten Tomatoes ratings when talking about streaming services. I first looked into Netflix, so I created a subset that only included the Rotten Tomatoes and Netflix columns and named it "df_netflix". The first thing we did was, like for task 1, converting the strings of ratings and Netflix values to floats and integers, respectively. For the Netflix column, the only possible values are 1, which means that the show is on Netflix, or 0, which means that the show is not on Netflix. Next, because the Rotten Tomatoes ratings are out of 100, we grouped the ratings of TV shows on Netflix by tens using `arange()` from the package Numpy. A pie chart was then created with `df_netflix.plot.pie()` to show the proportions of the ratings by tens. We also created a boxplot of the ratings grouped by whether the TV show is or is not on Netflix. This way we can visualize the values to prepare our t-test to determine if there is a difference in means of ratings between shows on and not on Netflix. We created two sets of data, one that contains all the shows on Netflix and one that contains all the shows not on Netflix. We then took a sample of 100 from each of these datasets to ensure that it satisfies the normality assumption in a t-test. After that, we used the function `stats.levene()` from the Scipy package to test whether the variances of each sample were equal to each other. If the p-value was greater than or equal to 0.05, then a variable we created called “var” would be equal to True. If not, “var” would be equal to False. With all this information, we were now able to conduct a t-test using `stats.ttest_ind()`. The test took into account the arguments `equal_var` and `alternative`. `equal_var` would be based on our variable “var”, and alternative would be based off of the boxplot. Since the boxplot of the ratings of Netflix showed that the median of ratings of shows on Netflix was higher than the median of ratings of shows not on Netflix, we set the alternative in this case to `alternative = “greater”`. Based on the p-value of the t-test, we would then be able to make conclusions about whether or not the mean of ratings of shows on Netflix is greater than the mean of ratings of shows not on Netflix. I continued this process with the three other streaming services. For Hulu and Disney+, we set the `alternative = “greater”`, but for Prime Video, we set the alternative to “less” which means we would be testing whether or not the mean ratings of shows on Prime Video is less than the mean of ratings of shows not on Prime Video.

Task 3

To analyze which platform contained the widest TV age ratings, I first counted the instances of each age occurrence in the dataset using `.value_counts()` to get a wider view of the age data. I used the `.sort_values().plot()` attribute to plot a bar graph to easily compare the number of each TV age rating, seen in Figure 3.1. I found that some shows were missing a TV age rating so those data points were ignored.

To specify the number of instances of a specific age rating for each platform, I used `df_tv.loc[df_tv['Age'] == '7+','Netflix'].sum()` and repeated for each streaming platform. I repeated this for each of the age ratings offered in the data set (all, 7+, 13+, 16+, 18+) to find the sum of each age rating on each individual platform. After finding the sums of the various age ratings on each platform, I used `matplotlib.pyplot` to plot a histogram and compare the sums with the streaming service on the x-axis and the TV age rating sums on the y-axis as seen in Figure 3.2. I grouped the histogram by platform and displayed the sum of each TV age rating. As seen in Figure 3.1, there were not many TV shows rated “13+” so it is not depicted in the histogram in Figure 3.2.

Task 4

We wanted to analyze which shows are on multiple streaming platforms. To analyze which shows were on two or more platforms, I utilized `.sum(axis = 1)` and created a new column named “Sum.” I used `.loc[df_tv['Sum'] >= 2]` to check the condition whether the value in the “Sum” column was greater than or equal to 2. Considering the Netflix, Hulu, Prime

Video, and Disney+ columns are binary variables, summing each row displays whether each show is on multiple platforms. Then I plotted a bar graph of the sum of the instances of TV shows that overlapped on different platforms, seen in Figure 4.1.

I also compared different combinations of three platforms to display the specific shows that were on three or more platforms. Considering the four total platforms, there are four different combinations of three platforms we want to find overlap. I found instances where all of the platforms displayed a 1 and created a new column that contained the result "True" or "False" using`.apply(lambda x: x.Hulu == x['Prime Video'] == x['Disney+'] == 1)`. I then used `.loc[df_tv['HPD'] == True, 'Title']` to display the specific titles of the TV shows that were on the three platforms. I repeated this process with the other possible combinations of three platforms.

Task 5

We want to see if there is any relation between the year of release of a show to a show being on a streaming service. A new dataframe was created which counts the number of shows released in a certain year. This was done by using `df.groupby('Year').count()`. Although this does count the number of shows released in a year, it shows the total rather than separating by each column. This was fixed by replacing all the nan values with zeros, but we need to check if there are any nan values in the 'Year' column before proceeding as it will affect our results. Thankfully there were no nan values in 'Year' so we can move on. Lastly, we drop all the unnecessary columns and the final result is seen in Figure 5.1.

Now that we have our new dataframe, we can sort it by highest to lowest so we can see which year has the most shows in each streaming platform. This is shown in Figure 5.2.

Task 6

We decided to add a function to our project where the user can input a TV show and it will give the information we have to the user. By using the `input` function, it will allow the user to type in a TV show name. We can then proceed with an if statement so that if the show received is in our dataset, it will give all the information we have to the user. The example used in Figure 6.1 shows all the information after the user asked for 'Breaking Bad'. If not, then it will return 'not found' as seen in Figure 6.2.

Observations and Conclusions

Observation

Task 1

The correlation coefficient between IMDb and Rotten Tomatoes ratings is 0.517, which means that there is a positive moderate correlation. The scatter plot (Figure 1) reiterates this result since a somewhat positive linear relationship can be seen. There are, however, way too many outliers that affect the correlation, making the correlation weaker.

Task 2

For Netflix, after doing a t-test with the `alternative = "greater"`, it returned a p-value of 0.00014 which is less than 0.05, so we can reject the null hypothesis that there is no difference in means of ratings of shows on and not on Netflix. We can conclude that the mean rating of shows on Netflix is higher than the mean rating of shows not on Netflix. For Hulu, we conducted the same test and came to the same conclusion as Netflix since the p-value for Hulu's test is 0.0233. For Prime Video where we set the `alternative = "less"`, we received a p-value of 8.215×10^{-10}, which is less than 0.05. We reject the null hypothesis and conclude that the mean rating of shows on Prime Video is less than the mean rating of shows not on Prime Video. For Disney+, we can see the p-value is 0.1863, which is greater than 0.05. In this case, we fail to reject the null hypothesis and conclude that the difference in mean ratings between shows on and not on Disney+ is not statistically significant.

When looking at and comparing the pie charts between the four streaming services, we can see that 17.1% of the shows on Hulu are rated 80 or higher on Rotten Tomatoes. This is the highest percentage of the four streaming services. Prime Video has the lowest percentage of the four at 7.38%. The pie charts also show that Netflix, Hulu, and Disney+ have similar distributions of Rotten Tomatoes ratings of shows that are on their platform. Prime Video, in general, has a more even distribution of Rotten Tomatoes ratings than the other three streaming services.

The results are below in Figures 2.1, 2.2, 2.3, and 2.4 with each of them representing Netflix, Hulu, Prime Video, and Disney+, respectively. The left side of each figure shows the boxplot and the right side shows the pie chart. The bottom of the figures shows the t-test results.

Task 3

Comparing the age ratings within Prime Video, we see that they include around the same number of each TV show's age rating. As seen in Figure 3.2, Prime Video has the most proportional TV age ratings between the various ages making it most appropriate for households with multiple different ages. We can also conclude that although Netflix and Hulu have more data points overall, they have a larger proportion of TV shows rated 18+ and 16+ respectively; Netflix and Hulu are more suitable for adults and couples, and Disney+ is most suitable for households with younger children.

Task 4

Analyzing which shows were found on multiple platforms, we found Hulu to have the most shows overlapping with at least one other platform. As seen in Figure 4.1, Hulu had the most shows appearing on two or more platforms, and Prime Video and Netflix were slightly lower than Hulu. Assuming the shows on multiple platforms are more popular and in higher demand, Hulu is the best in this regard.

When analyzing shows on three or more platforms, Netflix, Hulu, and Prime Video are seen to have the most overlap as opposed to any other combination of streaming services that includes Disney+. TV shows on Disney+ are least likely to overlap on any other platform.

Task 5

We can see in Figure 5.2, Netflix has the most shows from more recent years along with Disney+. We can assume these two platforms will release even more shows every year. As for Hulu and Prime Video, there are older shows compared to the other two services. After sorting by oldest to newest, Prime Video has the oldest show from the dataset being from 1904, and Hulu has the second oldest from 1931.

<u>Conclusions</u>

When deciding what streaming service to use, our analysis guides consumers on what each streaming service has to offer. We found that each streaming service had some positive attributes for various households.

Netflix was seen to have the greatest number of shows. We also found that it has the greatest number of new shows being released each year. It has a large variety of new shows so consumers would never run out of shows to watch.

Hulu has a greater proportion of shows that are highly rated compared to the other platforms, as seen on Rotten Tomatoes. It also contains the most shows that overlap with other streaming services. Hulu is suitable for those looking to watch popular shows that are in high demand with a fandom following

Prime Video was found to have the most evenly distributed Rotten Tomatoes rated shows. It also includes around similar amounts of each TV show's age rating, making it appropriate for households with a variety of ages.

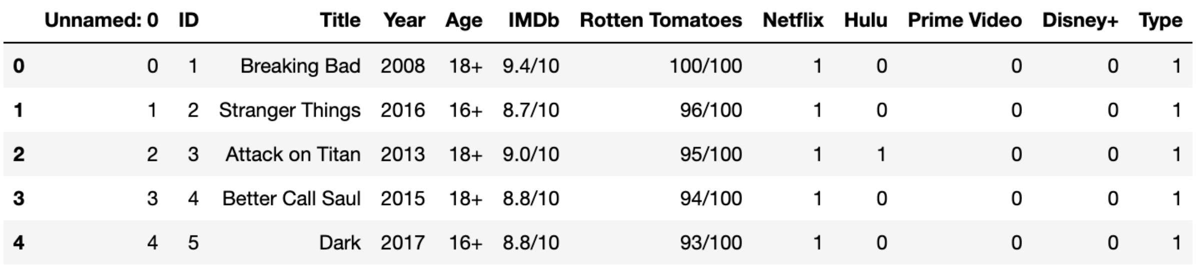
Disney+ was seen to be the most family oriented considering the age ratings are mostly rated for “all.” It also has the most exclusive content since its shows rarely overlap with any other platforms.

Considering the conclusions for each streaming service, consumers can use our analysis to choose which streaming service best caters to them and their needs.

Graphics & References

Graphics

Figure 0



	Unnamed: 0	ID	Title	Year	Age	IMDb	Rotten Tomatoes	Netflix	Hulu	Prime Video	Disney+	Type
0	0	1	Breaking Bad	2008	18+	9.4/10	100/100	1	0	0	0	1
1	1	2	Stranger Things	2016	16+	8.7/10	96/100	1	0	0	0	1
2	2	3	Attack on Titan	2013	18+	9.0/10	95/100	1	1	0	0	1
3	3	4	Better Call Saul	2015	18+	8.8/10	94/100	1	0	0	0	1
4	4	5	Dark	2017	16+	8.8/10	93/100	1	0	0	0	1

Figure 1.1

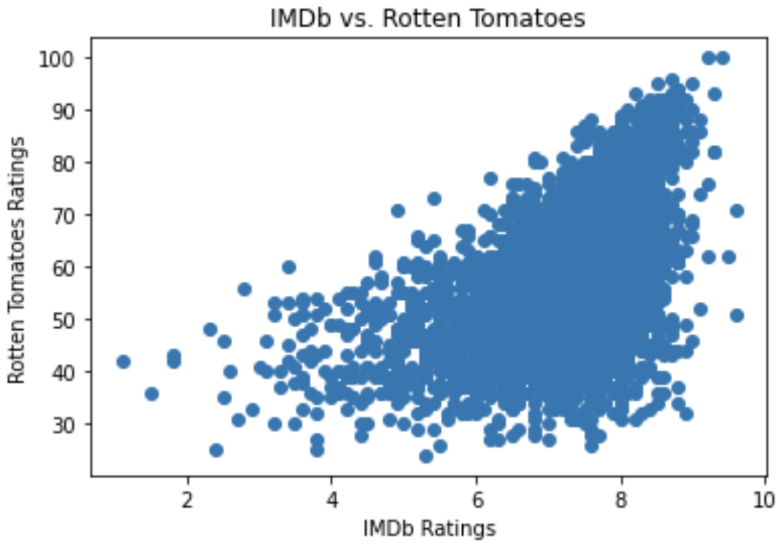
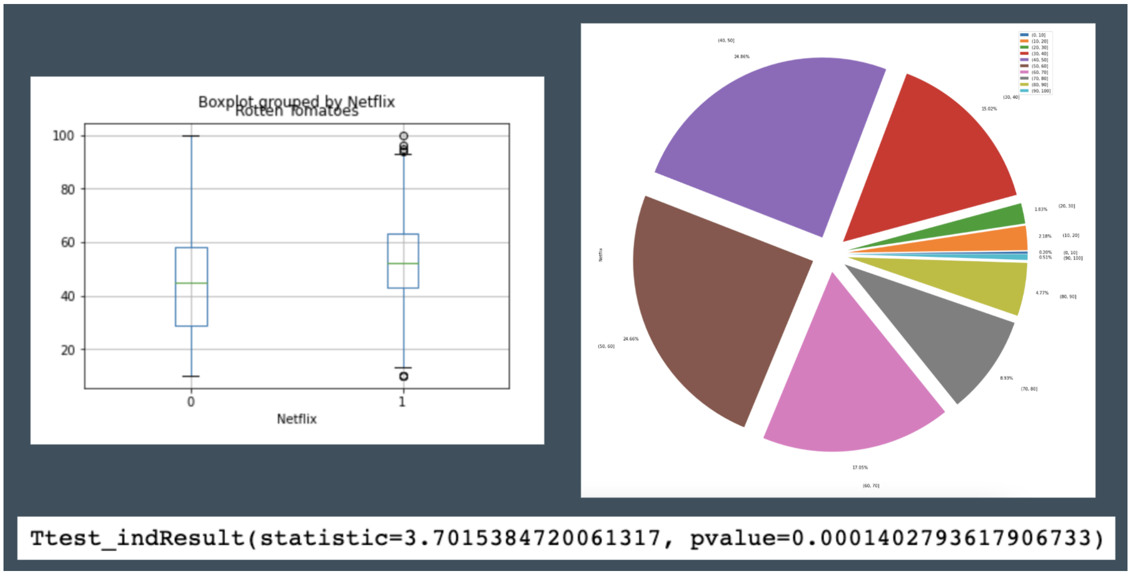
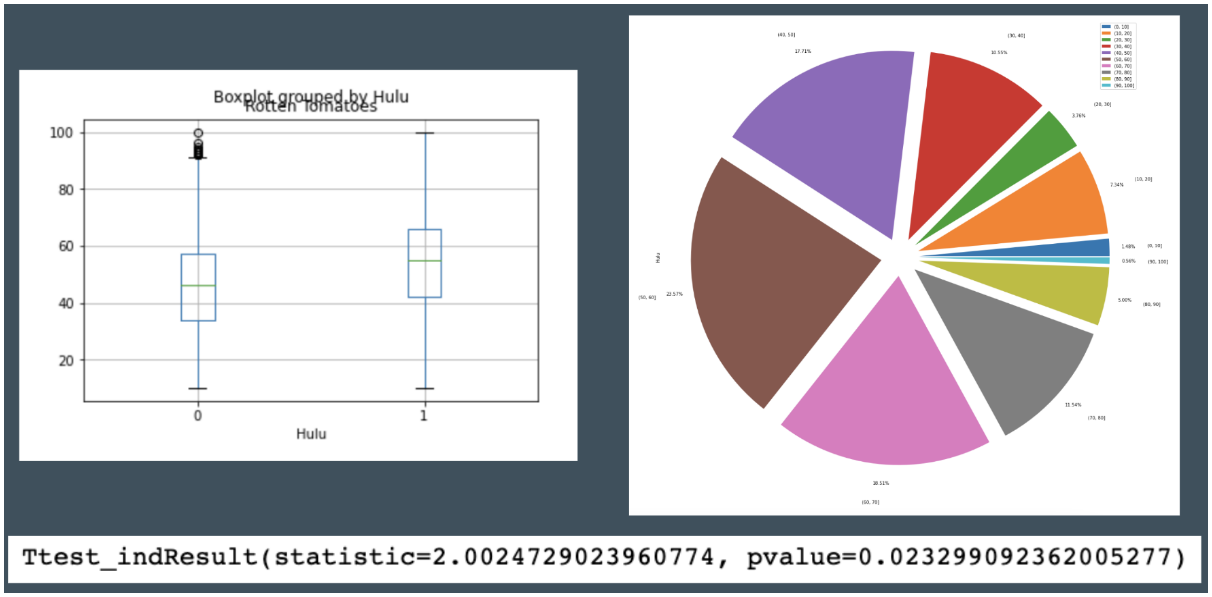


Figure 2.1



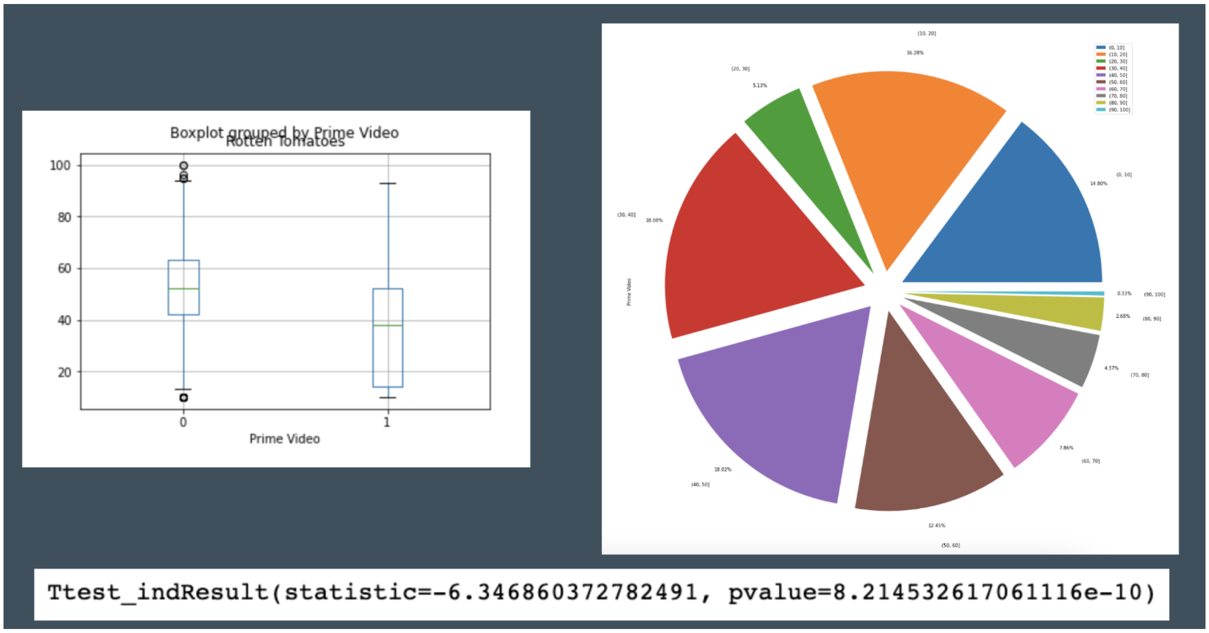
```
Ttest_indResult(statistic=3.7015384720061317, pvalue=0.0001402793617906733)
```

Figure 2.2



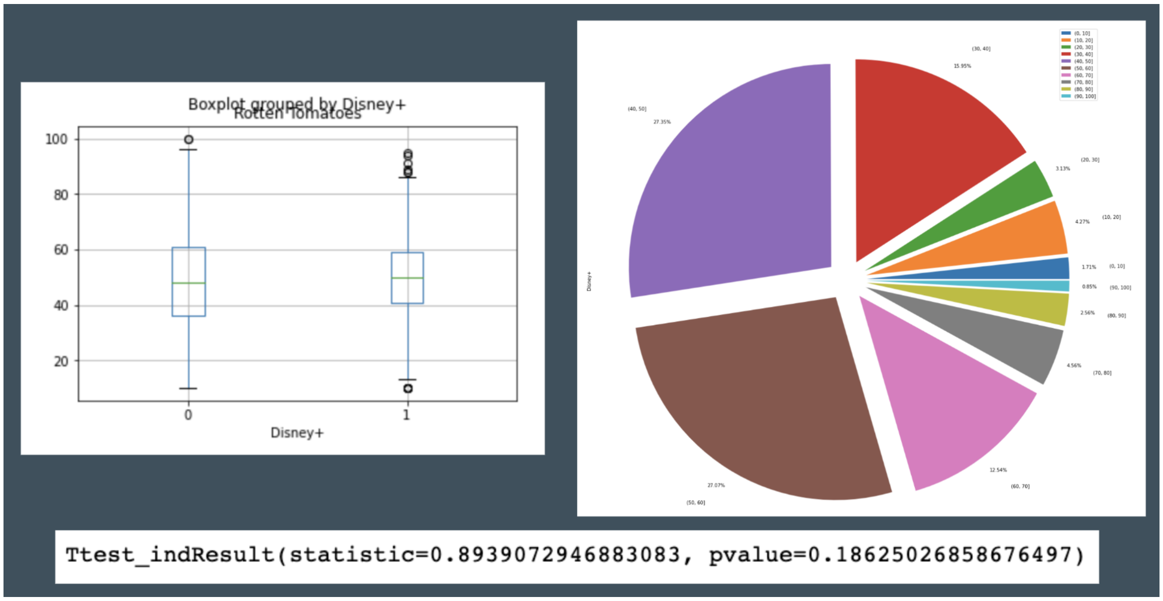
```
Ttest_indResult(statistic=2.0024729023960774, pvalue=0.023299092362005277)
```

Figure 2.3



```
Ttest_indResult(statistic=-6.346860372782491, pvalue=8.214532617061116e-10)
```

Figure 2.4



```
Ttest_indResult(statistic=0.8939072946883083, pvalue=0.18625026858676497)
```

Figure 3.1

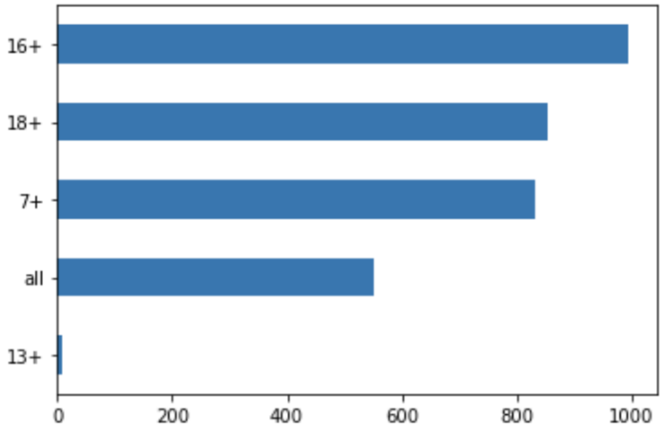


Figure 3.2

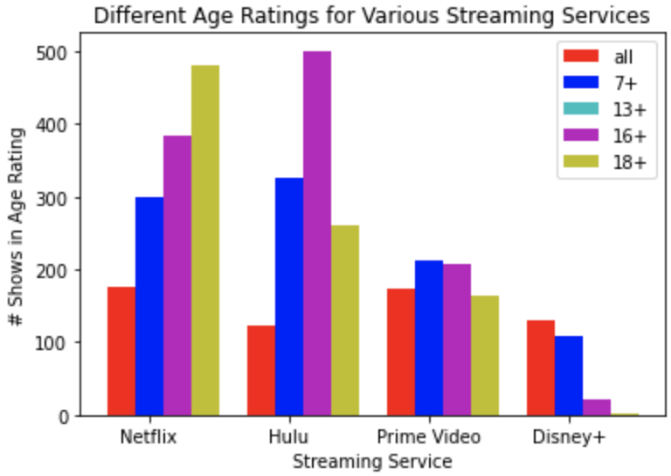


Figure 4.1

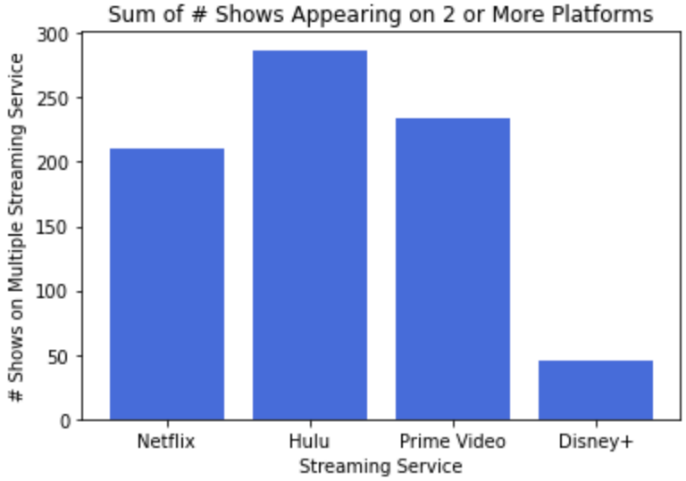
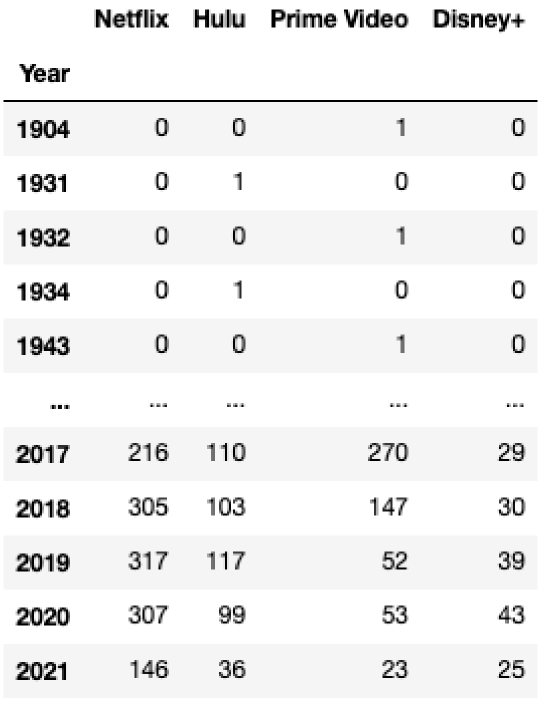
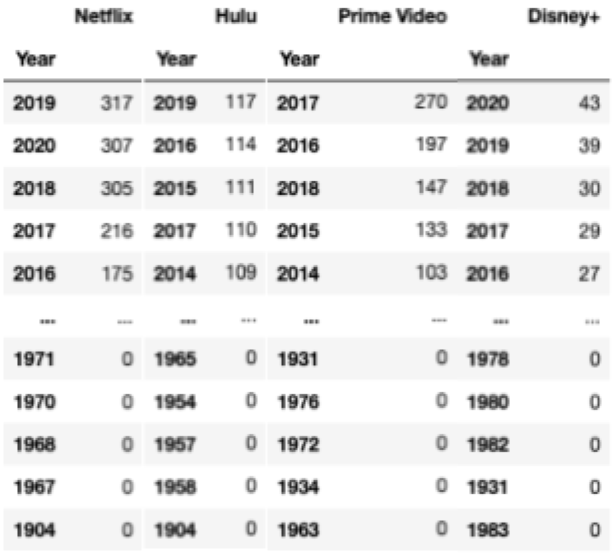


Figure 5.1



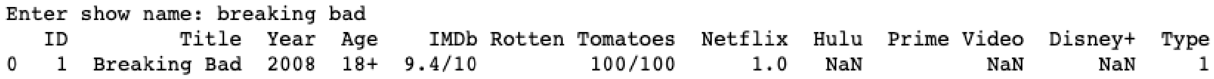
	Netflix	Hulu	Prime Video	Disney+
Year				
1904	0	0	1	0
1931	0	1	0	0
1932	0	0	1	0
1934	0	1	0	0
1943	0	0	1	0
...	...	...	...	...
2017	216	110	270	29
2018	305	103	147	30
2019	317	117	52	39
2020	307	99	53	43
2021	146	36	23	25

Figure 5.2



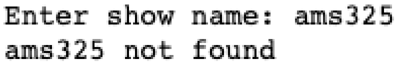
	Netflix		Hulu		Prime Video		Disney+
Year		Year		Year		Year	
2019	317	2019	117	2017	270	2020	43
2020	307	2016	114	2016	197	2019	39
2018	305	2015	111	2018	147	2018	30
2017	216	2017	110	2015	133	2017	29
2016	175	2014	109	2014	103	2016	27
...	...	...	...	...	...	...	...
1971	0	1965	0	1931	0	1978	0
1970	0	1954	0	1976	0	1980	0
1968	0	1957	0	1972	0	1982	0
1967	0	1958	0	1934	0	1931	0
1904	0	1904	0	1963	0	1983	0

Figure 6.1



```
Enter show name: breaking bad
   ID         Title  Year  Age    IMDb Rotten Tomatoes  Netflix  Hulu  Prime Video  Disney+  Type
0   1  Breaking Bad  2008  18+  9.4/10         100/100      1.0   NaN          NaN      NaN     1
```

Figure 6.2



```
Enter show name: ams325
ams325 not found
```

References

Dataset:
https://www.kaggle.com/datasets/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney
Github: https://github.com/vivyhuang/AMS-325---Group-Project