Analysis of Netflix Ability to Maintain Leadership

Carol Wan

Student ID: 501092321

Toronto Metropolitan University

CIND 820: Big Data Analytics Project

Instructor: Dr. Ceni Babaoglu

June 27, 2022

Table of Contents

Abstract

During COVID-19 lockdown, there were more people streaming online even with younger kids. It pushed up the Internet usage 70% higher and streaming more than 12%. The screen time overall was up almost a third (31%) in 2021 [01].

The social distancing and lockdown due to COVID-19 also impacted individuals' daily habits. Binge-watching, meaning users are watching multiple episodes of TV series in a single session, has increased during COVID-19 [02]. Before COVID-19, users might prefer watching movies on weekdays and TV shows on weekends. However, during COVID-19 lockdown, users excessively involved in watching TV series and emerged form of addictive behavior. In order to maintain itself as a top streaming service provider, Netflix spent billions of dollars on content to keep viewers interested [03]. However, there was news stating Netflix was going to slow down after lockdown. Recently, Netflix has also stated the shared accounts issue has impacted the number of subscribers. The problem statement in this project is if Netflix can still maintain itself as industry leader after COVID-19.

The comparative analysis and predictive analytics will be conducted to analyze if Netflix overall is doing better than its competitors in streaming service (such as Hulu, Disney+ and Amazon Prime). The ANOVA analysis will look into the diversity in selection offered, amount of content, and more popular movies/TV-shows. It will also use Sentiment Analysis to analyze the user's opinion of Netflix. The dependent variable in this paper is the service provider to analyze which service is doing better than others.
The three research questions are listed below:

1. Is Netflix providing the most diversified contents to different target audiences?

2. Is Netflix able to continue to provide the most new contents?

3. Is Netflix able to provide better user experience compared to its competitors?

For question 1 & 2, the paper explores KNN, Logistic Regression, Naive Bayes, Random Forest with 10-fold Cross Validation, and Decision Tree Model for accuracy, precision, recall, and ROC and compared among them.

The paper will explore the tweeter's data to perform Sentiment Analysis to compare user's comments of Netflix against its competitors by using TextBlob. TextBlob is like Natural Language Processing (NLP) that helps machines process and understand the human language so that they can automatically perform repetitive tasks and predict tasks based on a sentence or a series of words. The polarity and subjectivity is measured to understand which streaming service has better performance based on consumers' feedback in Twitter.

The datasets can be found on Kaggle and IMDB site:

Netflix - https://www.kaggle.com/datasets/shivamb/netflix-shows

Amazon Prime - https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows

Hulu - https://www.kaggle.com/datasets/shivamb/hulu-movies-and-tv-shows

Disney+ - https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows

IMDB - https://www.imdb.com/interfaces/

Twitter - Developer Platform - https://developer.twitter.com/en

The link to a repository on GitHub website - https://github.com/carolkkw/CIND820_Project

The tools used for this project are: Python, RStudio, ggplot2 for data visualization.

Literature Review

Paper 1: Learning to Predict Movie Ratings from the Netflix Dataset. [05]

Summary

In this paper, the two main approaches were made to recommendation - collaborative filtering and content-based classification. The task of a recommender system was to predict the ratings for items that users had not already seen and then to produce a ranked list of these items. Without regard for the item representation, it based on a recommendation to identify users with similar viewing patterns. Users who viewed the same items and rated them similarly are neighbors and their future ratings will likely be similar. Thus, when making a recommendation for a particular item-user pair, they looked at the ratings the user's neighbors gave to that item and computed an average of their scores. They got RMSE error for pure collaborative filtering was 0.43 and pure content-based classification was 1.68. In addition, they used Pearson correlation to calculate a similarity score for each pair of users.

Related to Project

Since the dataset in this project does not provide users preferences, the collaborative filtering cannot be used as it is a method of making automatic predictions about the interests of a user by collecting preferences information from users. This project will predict the IMDB ratings of movies or TV shows by each service provider. The Naive Bayes Classification will be used instead of a task of a recommender system. The Naive Bayes Model will be conducted to predict the IMDB rating of items. If Netflix has lower RMSE error and higher accuracy ratings, it means Netflix is having better content with higher IMDB rating than other competitors.

Paper 2: Data Analysis on Netflix datasets Data Analysis on the Netflix Datasets Motivation. [06]

Summary

In this paper, they explored Netflix data to find out how long the Netflix platform took a movie or a TV show to release on its platform; how many movies and TV shows were related in specific time frame; how many movies and TV shows were released in the recent ten years on the platform; what were the top 10 genres that the audience of the Netflix platform liked the most; who were the top 10 directors; who were the top 10 cast. They also rolled the sum of movies and found out most of the movies were released after the year 2000s. Also, they checked the difference between movie / TV show added year and released year to see if Netflix had updated movies.

Related to Project

Regarding this project problem statement, the top 10 genres / directors / cast will not help to determine if Netflix can provide the most diversified content to different target audiences. However, it is meaningful for finding which service providers are able to add content in a short period of time after the released year. The comparative analysis will tell us which service provider can attract new users if the service provider can provide most-up-to-update content. Release year means the movies / TV shows were produced. Their graph was not really showing Netflix had dramatically grown after the year 2000s and had that movies during that period. The added year of title will be used instead in order to determine if Netflix provides the most new content.

Paper 3: Impacts of Binge-Watching on Netflix during the COVID-19 pandemic. [07]

Summary

This paper deep dive into various binge-watching habits of Netflix users amidst the COVID-19 pandemic using a semi structured questionnaire. The sample size was 105 participants

successfully filled up the survey from Dhaka city, the capital of Bangladesh. The evidence

proved that these individuals were potential Netflix binge-watchers during the pandemic. The

finding was that the consumers spent over 70 hours per month binge-watching on Netflix. The

demographic distribution of respondents are 61% in 20-24 years old, 31.4% in 25-29 years old,

7.6% in 30-34 years old.

Related to Project

It is kind of understandable that people spent more time on watching movies / TV shows online

during COVID-19 lockdown. The dataset did not provide the age group to directly analyze if

Netflix is on the right track to target younger audiences or not. However, it can be analyzed using

the rating of the content to determine if the streaming platform's content is tailored toward the

specific age group, like teens, adults, or kids.

Paper 4: Value Proposition at NETFLIX. Retrieved from. [08]

Summary

This paper analyzed whether Netflix could be considered as a forerunner on the global market

based on the survey sent out to different social media platforms. Theoretical research was used to

introduce Netflix as a company and its company culture and also to compare Netflix to its

competitors on the global level. Compared to competitors' share, content, subscribers, Netflix

had produced 461 by far the most original TV shows in the market. The second was Amazon

Prime and the third was Hulu. Disney+ was a new service but still produced 17 originals.

Related to Project

Based on the paper, Netflix seems to still have potential leading the streaming service industry.

Since my dataset does not have any data related to shares, user preference and number of

subscribers, this project will focus on the content if Netflix is still able to add more new movies

or TV shows to gain new users and retain existing subscribers. Classification will be used to understand if Netflix is really able to have diverse content to diverse audiences.

Paper 5: Analysis of Different American Streaming Services and Shows. [09]

Summary

This paper stated users chose a platform which has high rated shows, popularity etc. Netflix, Amazon Prime, Hulu and Disney+ are a number of the various over-the-top (OTT) offerings which might be famous to the public. Scikit-learn methodology was used to identify the diverse classification, regression and clustering algorithms together with support-vector machines, random forests, gradient boosting, and k-approach. They concluded that Netflix was the best OTT platform among all its competitors till date, having high show choices and high rating movies across genres. It also showed that IMDB and Rotten Tomatoes ratings were not related.

Related to Project

Without other scores/ratings, the Pearson correlation coefficient cannot be used in this project as it is not possible to calculate the covariance since the categorical variable by definition cannot yield a mean. In our project, the plan is to check if Netflix has diversified genres and shows which are matching the approach in this paper.

Paper 6: Streaming Wars: Netflix, Prime Video, Hulu, and Disney+. [10]

Summary

This paper looked at diversity in the selection offered, amount of content, popular and highly-rated movies, and exclusive or original streaming content to determine which streaming service was doing better. The paper showed each distribution of years of production, the trend of each diversity category (e.g. year, country, genre, language, age rating, runtime), and amount of content by service. At the end, they confirmed Netflix was the best overall streaming service.

Related to Project

This paper is related to my project statement. However, my project will not use the release year to analyze if Netflix provided update-to-date content to users. The title might have already been released in 2020 but Netflix did not add until 2022 compared to other competitors who might add in the same year of release. In order to analyze properly, the suggestion is to find the variance of added year and released year instead. Regarding their diversity analysis, my project will analyze if Netflix can provide multiple levels of genre, certificate and production country.

Paper 7: In-depth study of Netflix's original content of fictional series. Forms, styles and trends in the new streaming scene. [11]

Summary

This paper presented an analysis of the original content of fictional series created by Netflix. This classified these contents according to their strategic nature, and offered a formal overview on their forms, formats, languages, genres and description. Its sample was made up of 490 series available on the Spanish version of the platform from its beginning in 2013 - 2019. This paper found the original production was dramatically increased from 2018 to 2019. The predominant language in Netflix's content was English. The genre was leading to drama.

Related to Project

The dataset in this project does not have anything related to original content. The trend of original content may not be able to be done. However, Netflix is likely to produce original content based on the most famous trend of shows/movies voted by users. As a result, we could compare how many shows/movies with higher IMDB ratings that Netflix had and compared to other competitors.

Paper 8: Executing a business model change: identifying key characteristics to succeed in

volatile markets. [12]

Summary

The change of the business model and the change of the organization were needed to succeed in the market. Netflix had successfully changed their business model twice: first, from an online-DVD-rental service to a streaming provider and then, to a content provider. Netflix changes possess certain triggers and environmental dynamics which could be classified into three triggers. The article discussed the change of Netflix's main competitor over time, and listed Amazon Prime, Hulu, HBO, Apple TV+, and Disney+ as Netflix's main competitors.

Related to Project

It is good to know that Netflix changed their business model twice successfully in order to maintain its leadership position compared to upcoming competitors or existing competitors. Although this paper did not mention any algorithms or models from data science perspective, it concluded that Netflix's content is still the top of streaming service. This project will further analyze Netflix's content characteristics to see if Netflix has any competitive advantages over its competitors discussed by the paper.

Paper 9: The Influential Factors of Developing Better in the Global Stream Media Market: An Analysis of Netflix. [13]

Summary

This paper took the Marketing Mix theory as the theoretical support, which consisted of four main factors: product, price, promotion, place, to analyze the relationship between 4Ps and consumer behavior by combining the 4Ps with SWOT analysis, and proved the hypothesis of the study by analyzing the specific market data of Netflix. The results of the paper stated that Netflix's strategies had promoted consumers' purchase of Netflix to varying degrees and

expanded Netflix's global market share. Also, it concluded that Netflix had used its own localized global expansion strategy to obtain its high-quality targeted consumer in the streaming media market. However, it still faced different competitors, and there were deficiencies in evaluating quality and price.

Related to Project

My dataset does not have anything related to the price. However, in order to check if the quality of content is really deficient in Netflix, we can compare the IMDB rating with other competitors to see if Netflix can provide better quality with higher ratings than others. Regarding the targeted consumer, we can analyze title certificate/IMDB rating/genre to see if Netflix has diverse targeted consumers. Regarding the place, we can analyze the country where the versions of titles were produced in order to confirm if Netflix launches different content and unique content preferences in different countries and different markets.

Paper 10: Sentiment Analysis, Tweet Analysis and Visualization on Big Data Using Apache Spark and Hadoop. [14]

Summary

This paper mentioned where and how to download data from Tweets using two methods which were search by a hashtag and keyword "#Netflix". Also, they used Hadoop for storage purposes and Spark to do cluster control. Logistic Regression and Random Forest Classifier were conducted to predict and classify where the output variable was dichotomous in nature, to calculate the accuracy of these two models with comparison. While the Random Forest Classifier showed the accuracy of prediction was 48.07%, the Logistic Regression stated the accuracy was 98.88%. The difference of almost 50% was because they had a "Categorical variable dataset".

Related to Project

This project will try to get the latest sentiment data from twitter to analyze and compare between these streaming services - Netflix, Hulu, Disney+, and Amazon Prime. However, this project will use the Textblot to analyze unstructured data and contain human-readable text which the paper did not do. Random Forecast Classifier will be used to calculate the accuracy of the model. Depending on the data we can get from Twitter, it is good to do Logistic Regression to compare the accuracy with the Random Forecast Classifier.

Data Description

After collecting data, Movies/TV shows data from Netflix, Amazon Prime, Hulu, and Disney+ are all in the same schema and data type. As a result, all files will be unioned first before doing further Initial Data Analysis. There are 22,998 observations of 13 variables. The source is mentioned in the abstract above. There are few columns (e.g. show_id, director, cast, duration, description) that will not be used as they are not related to the paper's problem statement.

In addition, two fields (e.g. year_added, service_name) will be added for better analysis and grouping. Some of the columns will be renamed to appropriate column names. For example, there are two rating fields in the streaming table and IMDB table. The rating in the streaming table will be renamed to "certificate" and in the IMDB table will be renamed to "imdb_rating".

Beside the streaming table, data (e.g. genre, rating) will be selected in IMDB tables (title.basic.tsv.gz, title.ratings.tsv.gz). The IMDB rating will be a dependent variable for predicting if Netflix is doing better than other competitors. The genre description will replace "listed_in" in the streaming table as it is inconsistently used by the streaming services. Examples are listed in [Appendix C].

All selected fields in the IMDB dataset will be joined by titleId and then joined with the streaming dataset by title and release_year. The reason to join by title and release year is because one title might be released in different years. This will reduce duplicate titles in each streaming service. Once the streaming table and IMDB table are joined, those titles without services will be dropped as a list of movies /TV shows in IMDB may include other companies' content while this paper only focuses on 4 streaming services. Now, the data increases 10% from 22,998 to 25,225 observations of 10 columns. However, this observation will decrease 60% from the original

dataset after we do complete.case for removing all missing values. There are 35% in average of "imdb_rating" missing as there are no more updated IMDB ratings after 2020. Also, some titles are not voted/shown in IMDB. The relationship of tables with appropriate data type and names is identified in [Appendix A]. The description of the streaming table is in [Appendix B].

The last dataset is sentiment data taken from Twitter using Tweepy library started collecting data from June 12, 2022. It explains the sentiment of polarity and subjectivity based on text/ feedback given by consumers. There are no missing values and the correlation between them is about 0.236 with positive relationship. Consumers might mention multiple streaming services in the same text. In order to declare which streaming service is related to that polarity and subjectivity, the rule of thumb is to take the first service name.

The Initial Analysis starts from checking observation, summary() (e.g. min, max, mean, median, missing value), str() for checking data type, head() for reviewing tables, checking outliers, and finding correlation between variables. The missing value handling is listed in the Data Cleaning section below. For the normalization, Anderson-Darling Normality Test for normality is conducted to test p-value and found out the IMDB rating is left skewed. Pearson's correlation shows that there is no linear relation between year added and IMDB rating.

According to the boxplot below, the outlier for Netflix TV Shows and Movies are both the same prior to 2015 because one of the Netflix servers in its farm did not work [04]. Since the outlier is not due to incorrectly entered or measured data, the paper will not drop the outlier. However, the paper will focus more on the data after 2015, especially the problem statement is if Netflix can still maintain its leadership after COVID-19. The missing value of title_certificate can be replaced by mode and re-organized the levels that decrease from 106 levels to 7 levels with better analysis.

Type by Year_Added
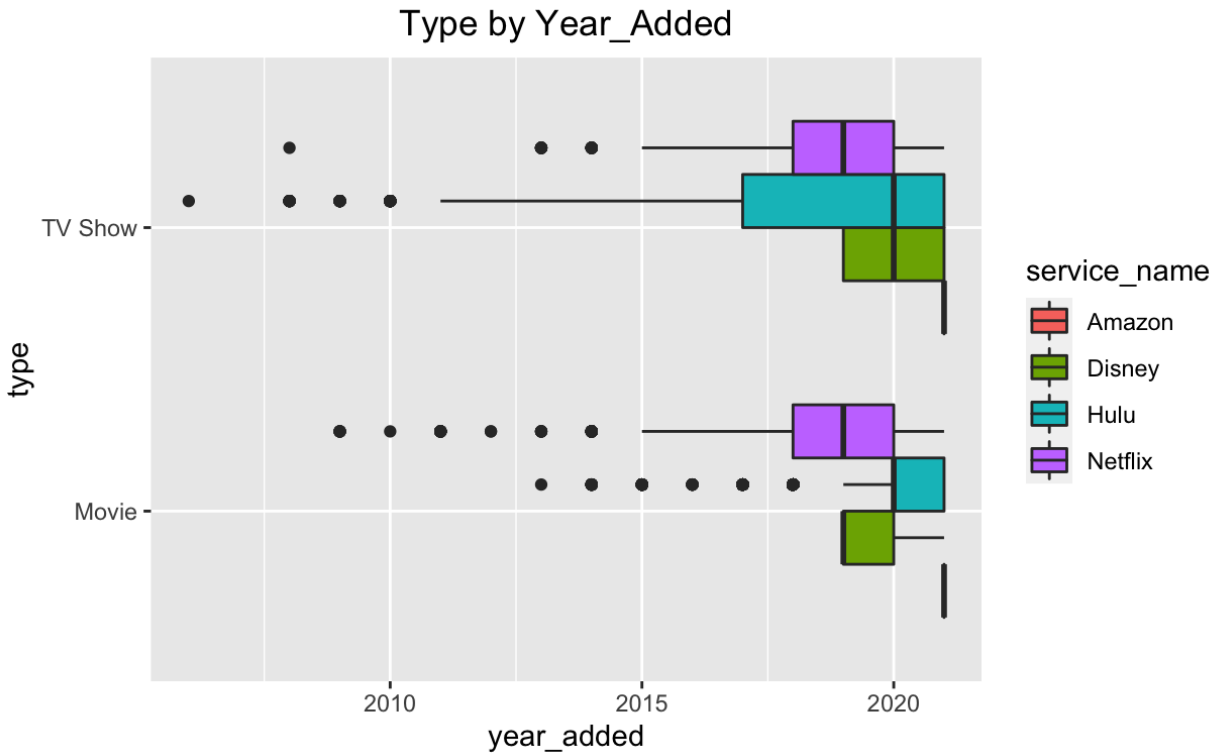
Through the table below, it stated Amazon Prime has less population size after removing missing values and has the highest standard deviation (1.282369) that the data is less reliable. Disney+ has the lowest standard deviation (1.074363) that the data are clustered closely around the mean and more reliable. Hulu and Netflix have about the same standard deviation but the mean of imdb_rating of Hulu is doing a bit better than Netflix. This also tells us that this data is skewed to the right.

service_name <fctr>	size <int>	mean <dbl>	median <dbl>	sd <dbl>
Amazon	12	6.141667	6.3	1.282369
Disney	1081	6.595560	6.7	1.074321
Hulu	1479	6.771602	6.9	1.210438
Netflix	6491	6.447651	6.5	1.216121

Data Approach
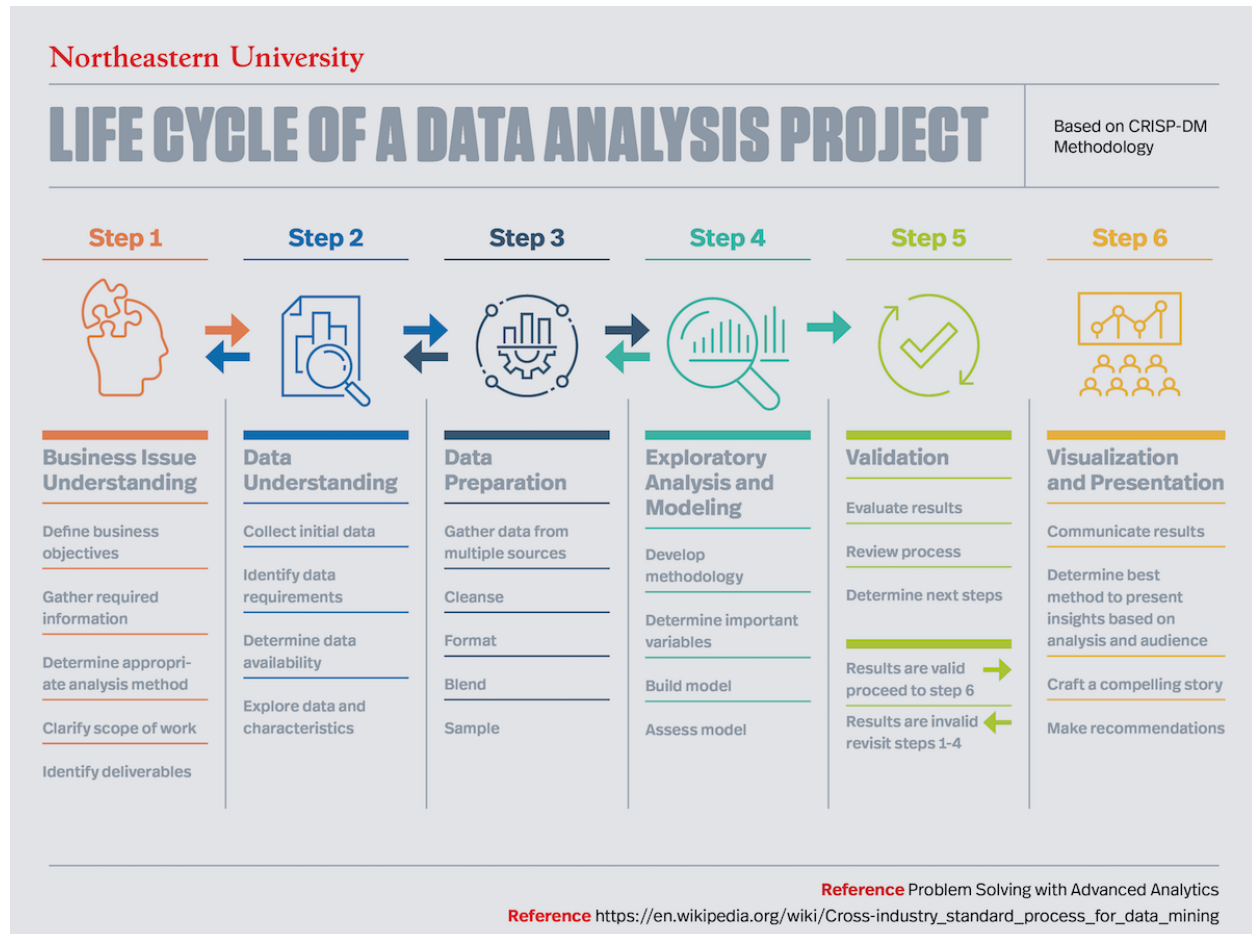
The image above is telling us the main steps for the data analytics process [15] that can go back and forth depending on the needs and requirements. Below is the main steps for data analytics process:

Define Problem

Define the problem statement for the paper, gather information and identify deliverables. For example, the three questions and the main problem statement are mentioned in the abstract.

Data Understanding

Collect and gather dataset(s) from data sources, parse, import and prepare the data to be processed and analyzed. This is like the data description mentioned above.

Data Preparation & Cleaning

1. Below are the removed columns.

   a. "Show_id" - is not unique after union as companies are using their own id

   b. "Director" - Although a company has more movies/TV shows from specific directors, it does not mean it is good quality of content which can attract new subscribers or retain existing ones.

   c. "Cast" - Same as the reason as "director" above. It does not explain how the cast will be impacted by the diverse content.

   d. "Duration" - The length of movies or TV shows cannot explain if Netflix is able to provide the most new content nor if Netflix is providing the most diversified content.

   e. "Description" - This is provided by the studio with standard format. Therefore, analysis cannot be very effective using this field.

2. The average of missing values are 6.9% in Netflix, 11.9% in Hulu, 7.5% in Disney+, 40% in Amazon Prime [Appendix D]. There are more than 92% missing values in the fields of the country, date_added in Amazon Prime. The country field is missing close to 50% in Hulu. This paper will remove all missing values.

3. The "date_added" field has space in front of the month (e.g. " October 3, 2021 "). It is trimmed before converting from character format (e.g. "October 3, 2021") to date format (e.g. "2021-10-03").

4. Add a new column named "year_added" that will be used to compare "release_year". This will convert from character to integer for calculation purposes.

5. There are 105 levels in the "rating" categorical variable, including incorrect data such as

minutes of movies and number of seasons. This paper will organize and replace those

with NA. In addition, Amazon Prime uses different descriptions of rating which will be

replaced by equivalent names so that it can align with other providers (e.g. "13+" in

Amazon Prime is equivalent to "PG-13" in Netflix). The name of rating will be

re-assigned based on IMDB Certificates by Country [15]. The level decreases to 7 levels.

6. Hulu has about 16% missing value and Amazon has about 3% missing value in the

    certificate field. Those missing values will be replaced by mode for better analysis.

    Below is the mode by company.

	Netflix	Hulu	Disney	Amazon
mode	TV-MA	TV-14	PG	PG

7. In order to have more control over what the output we can get, these fields' data type will

    be efficiently changed from character to factor such as "type", "genre", "certificate",

    "country", "service_name".

8. Text Preprocessing for sentiment analysis such as removing punctuation, lowercase, stop

    words, stemming / words inflection and lemmatization, spelling correction, etc using

    TextBlob that is kind of doing the same thing as Natural Language Processing.

Data Exploration

1. Based on the titles voted in IMDB, the bar graphs below shows Netflix is leading the

    streaming services. Due to COVID-19 lockdown, all streaming services did not have any

    new released year of the title. As a result, both movies and TV shows dramatically

    decreased about 50%.

Movie Released by Company After 2015

TV Shows Released by Company After 2015
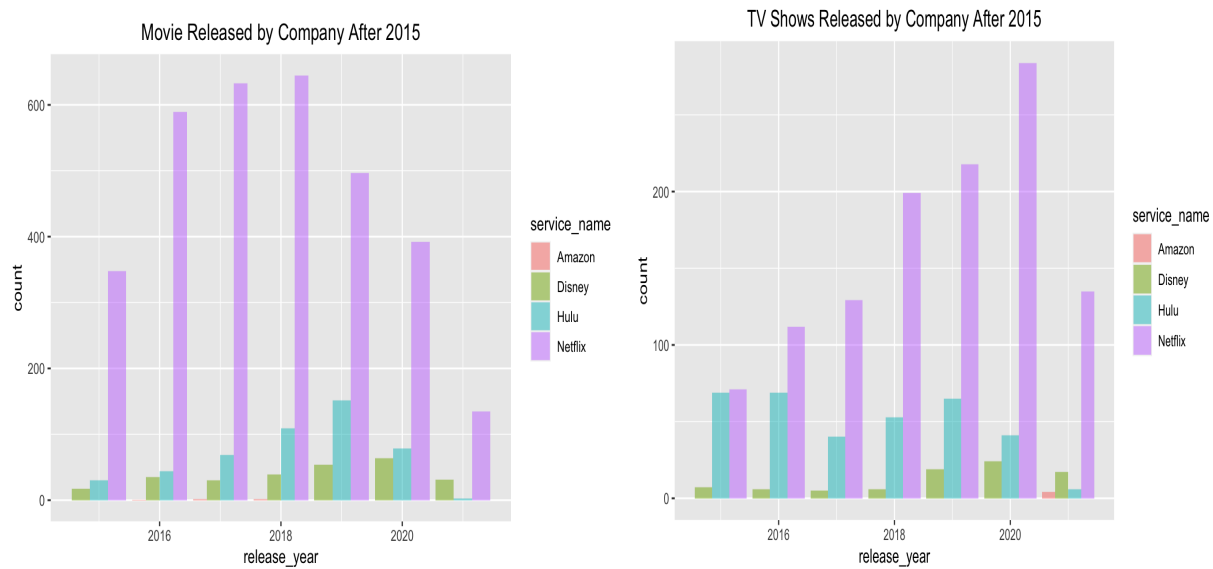
2. In order to understand if Netflix offered contents in a timely manner or not. The analysis is based on the difference between the year added and year released of the title. It is observed that Hulu is increasingly adding new TV contents while Netflix is dramatically down due to COVID-19. Hulu will be the leader of providing new content in streaming services if Netflix is not able to reverse this trend. Due to 98% missing values in Amazon's "year_added" field, the paper cannot analyze Amazon's trend in this criteria.

Content Added by Company After 2015

Movie Added by Company After 2015

TV Show Added by Company After 2015
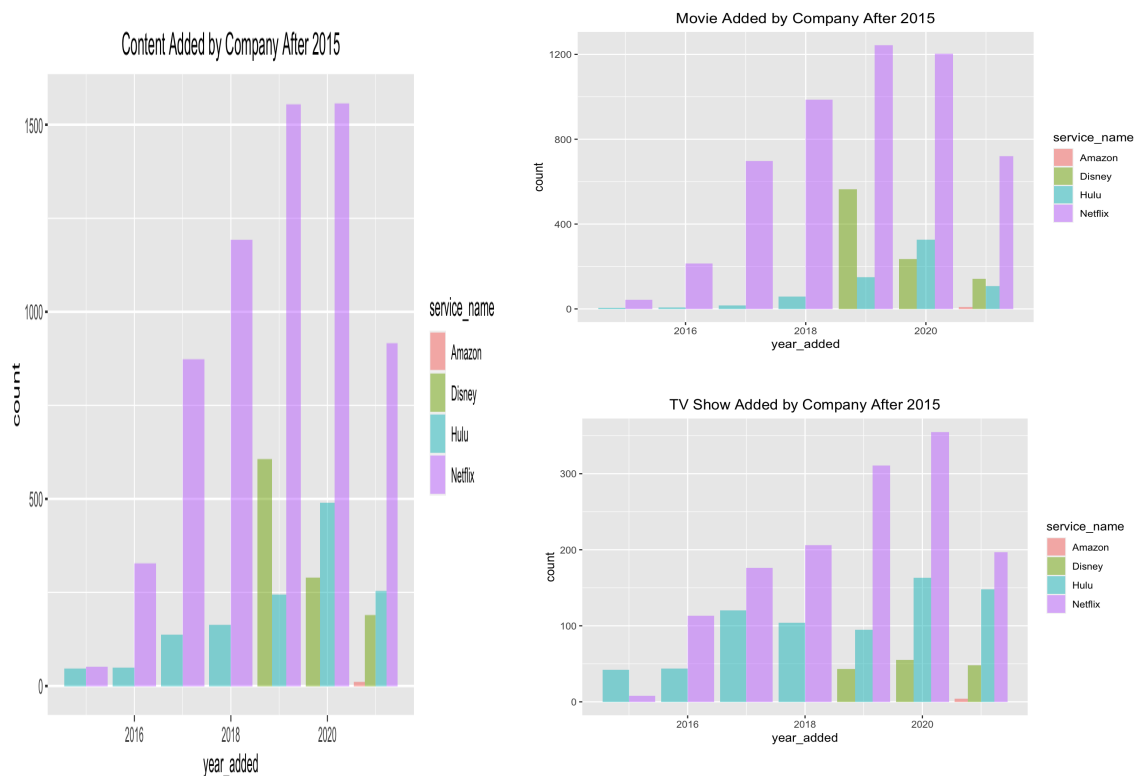
3. Anderson-Darling Normality Test for normality is used to test if a sample of data came from a population with a specific distribution. The p-value is less than 0.05, which means the average of IMDB rating and the year added are not following a normal distribution. As per bar chart [Appendix E], they are left skewed both on movies / TV shows. TV Shows at Amazon Prime cannot be analyzed due to the sample size being less than 7. In order to normalize the left skewed data, Min-Max normalization is done before doing any modeling.

4. Correlation is very low overall. The highest positive correlation (0.2645) is between imdb_rating and type. Due to the weak positive/negative relationship, the variables are hardly related [Appendix F].

5. Pearson's correlation coefficient is conducted with the result of negative correlation (-0.0415) between year added and imdb_rating for the 4 streaming services. The R-squared (0.0699) means there is no linear relation between variables. [Appendix G] The correlation of year_added and release_year for Netflix and Disney+ are about the same, which is 0.033. It is a positive correlation which means they both are able to add new content once the titles are released. Due to less sample size, the correlation coefficient cannot be done for Amazon Prime. The result aligns to point 4 above.

6. By checking ANOVA analysis as table below, genre and year_added do not have a statistically significant effect on IMDB ratings. Certificate and country are statistically significant while others are highly significant.

```
ANOVA:Analysis of Variance Table

Response: imdb_rating
              Df  Sum Sq Mean Sq  F value    Pr(>F)
type           1   921.5  921.55 692.6771 < 2.2e-16 ***
genre          1    14.9   14.90  11.1998 0.0008214 ***
certificate    1     3.5    3.55   2.6668 0.1024974
country        1    13.7   13.70  10.2954 0.0013383 **
year_added     1    15.7   15.68  11.7827 0.0006005 ***
release_year   1   146.3  146.31 109.9750 < 2.2e-16 ***
Residuals   9056 12048.3    1.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. Using ANOVA models, p-value (<2e-16) is smaller than 0.05, we reject the null

   hypothesis that all means are equal. Therefore, we can conclude that at least one

   streaming service is different than the others in terms of imdb_rating.

```
ANOVA:Analysis of Variance Table

Response: imdb_rating
               Df  Sum Sq Mean Sq F value    Pr(>F)
service_name    3   135.4  45.149  31.393 < 2.2e-16 ***
Residuals    9059 13028.5   1.438
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is less than 0.05, the Post-hoc test [17] (after obtaining statistically

significant ANOVA results) will be used to find out which streaming service(s) is/are

different. In the output of the Tukey HSD test, the first column shows the comparisons

which have been made and the last column shows the adjusted p-values for each

comparison. The output looks like any comparisons with Amazon are different. It may be

because of a lack of sample size. [Appendix H]. Below is the comparison of

Kruskal-Wallis model and Anova model [18]. Their p-values are the same.
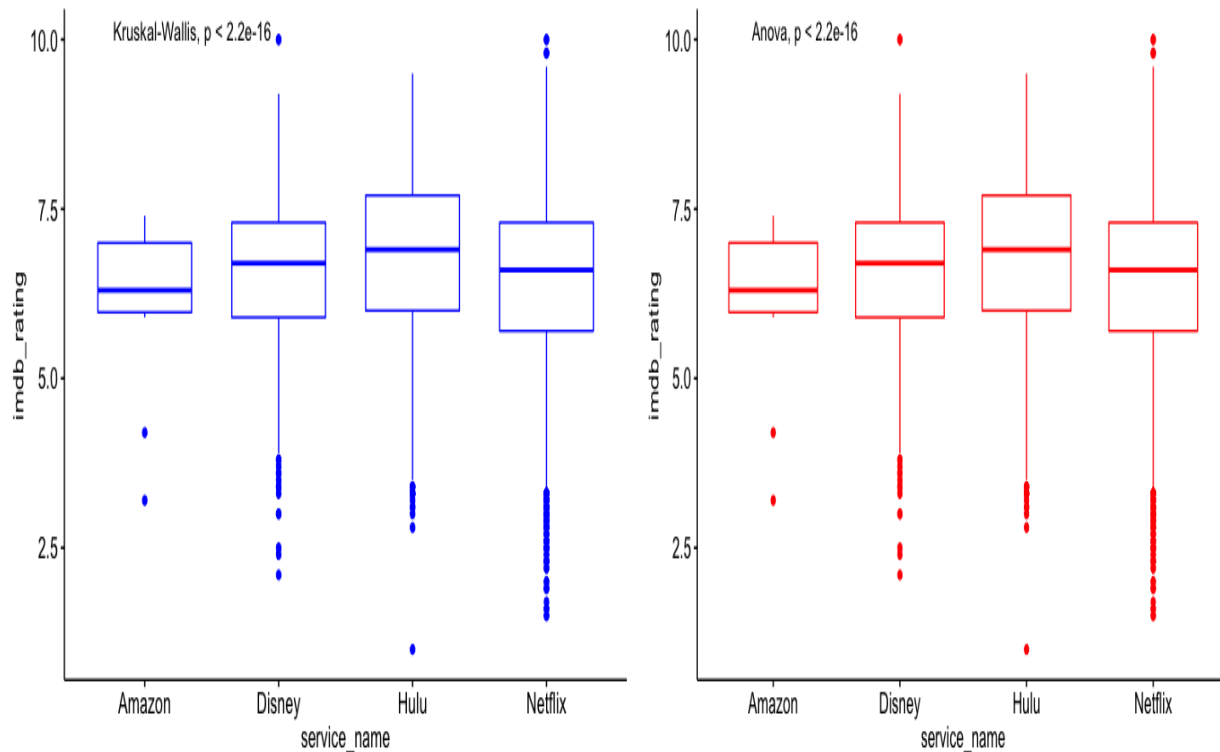
8. Based on the summary of ANOVA analysis, the p-value (2.26e-07) is less than 0.05.

Therefore, the null hypothesis is rejected. It means that the imdb_ratings are not the same

among the countries. The difference of imdb_rating across the countries selected is

statistically significant. The boxplot chart states that Netflix has production in all top 10

countries with a stable IMDB rating among other competitors. Unlike Hulu, its

imdb_rating is up and down depending on the countries. It is indirectly telling that the

quality of content is not as good/stable as Netflix even though Hulu and Netflix may

produce in the same country. In the top 10 countries, Disney+ shows in 4 countries and

Amazon Prime shows in 2 countries. All streaming services are producing movies and

TV shows in the United States. Only Amazon Prime does not produce TV shows in the

United Kingdom. Only Netflix produced movies and TV shows in Canada while the other

3 competitors only did for movies. This explains Netflix acquires from different sources

for targeting global. The breakdown by movie and by TV show can be reviewed in

[Appendix J].

Country vs IMDB Rating Top 10
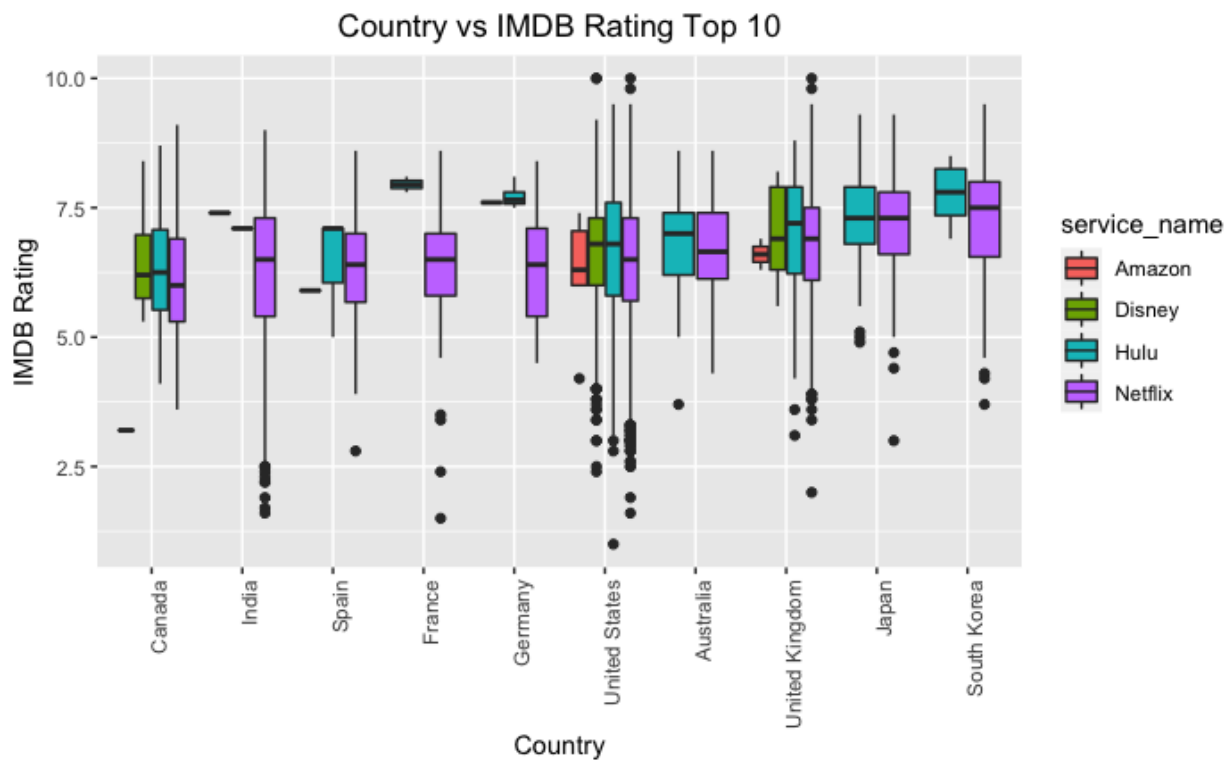
9. Netflix provides more content with different certificates to consumers than the

competitors. Netflix targets movies more for Adults and Teens, and then 18+ and Kids,

and more on TV shows for Adults, Kids, and Teens while Hulu focuses more on Teens

only. Disney+ in general leads on Kids movies and TV shows. The paper cannot analyze

which certificate is Amazon Prime targeting because its movies / TV shows are not in the

list of IMDB. [Appendix I]

10. According to the boxplot below, Netflix shows up in the Top 10 genre vs imdb_rating.

"Documentary" has higher ratings at Netflix, Hulu, and Disney+. Amazon Prime is not

visible in the chart due to missing data and/or its movies/TV shows are not in the IMDB

dataset. Disney+ has the higher rating in the genre of "Family" compared to Netflix. This

is normal as Disney+ is more leading on kids and family shows. Also, Disney+ has a

large amount of animation which is its main production line.
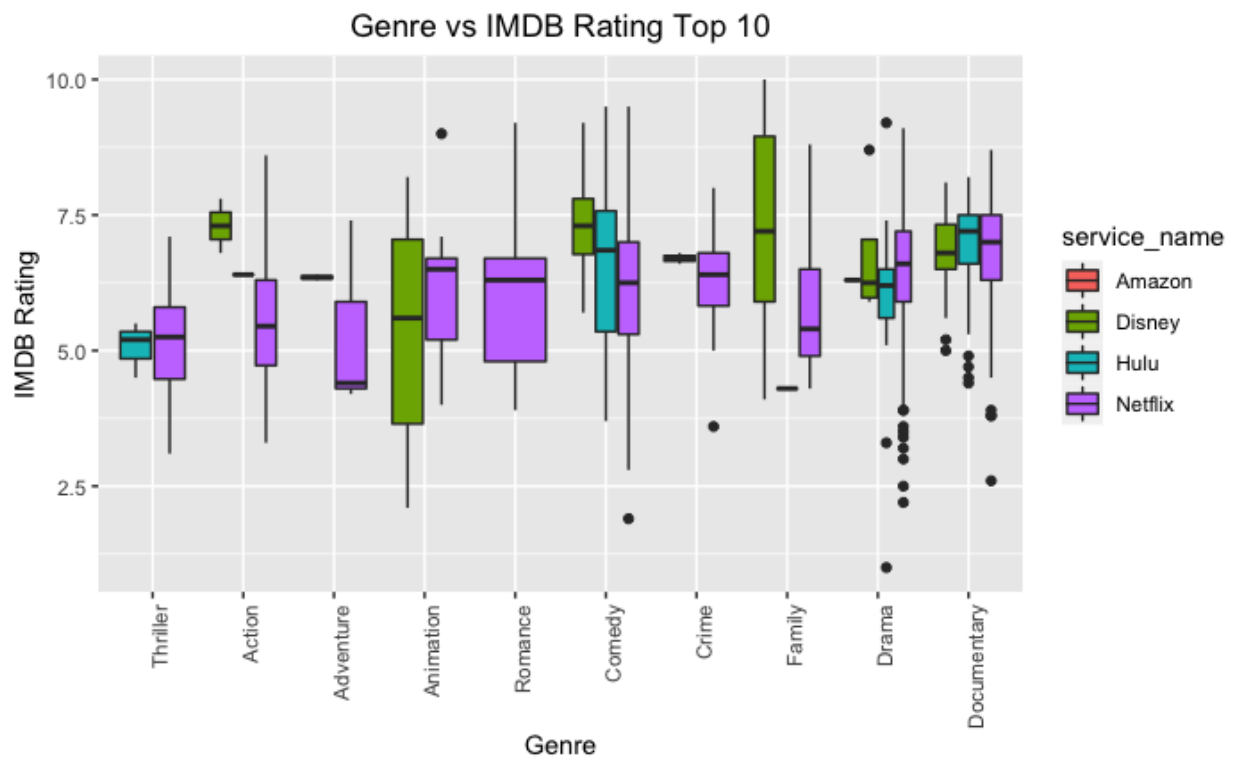
11. Sentiment data will continue to be collected until the end of the project. From June 12th

till now, only Hulu has a negative score on polarity and highest subjectivity among all

streaming services. Netflix, Disney+, and Amazon Prime seem about the same.

Design -> Modeling -> Evaluation

In order to handle imbalance data, the paper tests two methods - Combinations of Resampling

Method, and the Synthetic Minority Oversampling Technique (SMOTE). The combination of

oversampling and undersampling provides about highest 0.5 accuracy with about highest 0.5

error rate [Appendix K]. It is because oversampling involves duplicating examples in the

minority class while under sampling involves randomly deleting examples from the majority

class. SMOTE oversampling is used for increasing the number of minority classes in the dataset

in a balanced way. The highest accuracy is about 0.78 while the error is about 0.34 [Appendix

L]. Since it is not suggested to remove useful data, the paper will choose SMOTE oversampling

to handle imbalance data. SMOTE fits into 10-fold Cross Validation which is implemented to

split train and test sets into 70% and 30% respectively and then apply to the following Models..

1. KNN Model - using the optimal chart of "Error Rate vs K Value" and "Accuracy vs K

    Value" [Appendix L] without Cross-Validation applied, k=14 seems better fit into the

    model with 0.78 accuracy. The accuracy is the same for both 5-fold and 10-fold.

    However, after running the optimal value of K for KNN using Cross-Validation, selecting

    3 for k has a better accuracy of 0.79.
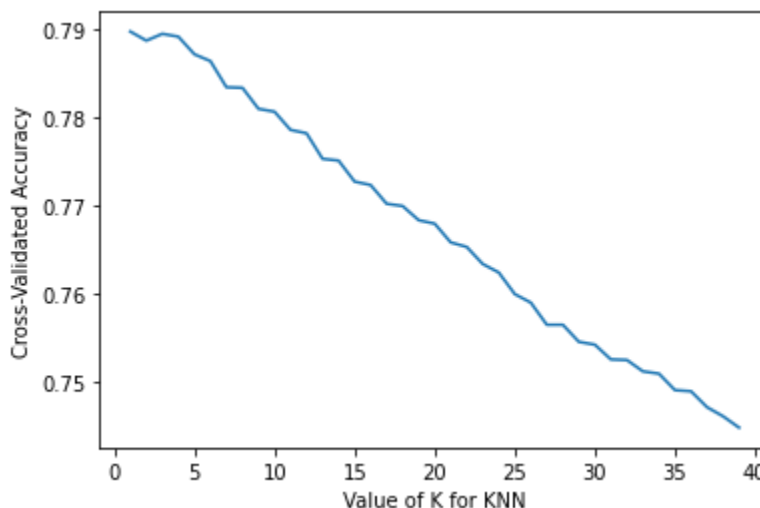
2. Logistic Regression Model - with 10-fold Cross-Validation, the accuracy is 0.465.

3. Naive Bayes Model - with 10-fold Cross-Validation, the accuracy is 0.418.

4. Random Forest Model - with 10-fold Cross-Validation, the accuracy is 0.0.797.

Compared to 4 Models above, the paper concluded KNN and Random Forest Model are about

the same. The paper will look at ROC by streaming services later.

Decision Tree Model

- The maximum depth allows the tree to grow with less training error. The paper found the

    max depth is 24 and minimum sample leaf is 28 in order to reach the highest 0.768

    accuracy rate.

Sentiment Analysis

- The paper does not have enough data collected over time so the Time Series Analysis

    may not be suitable. The SVM and Multinomial Naive Bayes Model will provide better

    accuracy in this case.

Conclusion

The conclusion will provide a brief summary of the report such as its SWOT (strengths,

weakness, opportunities, threats) about Netflix. Based on the result from the analysis, it will try

to conclude if Netflix can still maintain its leadership within the streaming service.

Reference

[01] Rajan, Amol (2020). TV watching and online streaming surge during lockdown. Retrieved

from https://www.bbc.com/news/entertainment-arts-53637305

[02] Aghababian, Anahys H (2021). Binge Watching during COVID-19: Associations with

Stress and Body Weight. Retrieved from https://pubmed.ncbi.nlm.nih.gov/34684420/

[03] BCC (2022). Netflix faces rocky road after pandemic wins. Retrieved from

https://www.bbc.com/news/business-60077485

[04] Fisher-Ogden, Philip. Burrell, Greg. Sanden, Chris. Rioux, Cody (2015). Tracking down the

Villains: Outlier Detection at Netflix.

https://netflixtechblog.com/tracking-down-the-villains-outlier-detection-at-netflix-40360b

31732

[05] Kim, Donginn (2022). Learning to Predict Movie Ratings from the Netflix Dataset.

Retrieved from

https://www.researchgate.net/publication/266457208_Learning_to_Predict_Movie_Ratin

gs_from_the_Netflix_Dataset

[06] Bhargav, Vybhav Achar (2022). Data Analysis on Netflix datasets Data Analysis on the

Netflix Datasets Motivation. Retrieved from

https://www.researchgate.net/publication/359747031_Data_Analysis_on_Netflix_dataset

s_Data_Analysis_on_the_Netflix_Datasets_Motivation

[07] Rahman, Kazi Turin (2021). Impacts of Binge-Watching on Netflix during the COVID-19

pandemic. Retrieved from

https://www.emerald.com/insight/content/doi/10.1108/SAJM-05-2021-0070/full/html

[08] Csalló, Rebeka (2021). Value Proposition at NETFLIX. Retrieved from

https://www.theseus.fi/handle/10024/503990

[09] Paliwal, Abhijay. Sangai, Lobhas. Khare, Vikas (2022). Analysis of Different American

Streaming Services and Shows. Retrieved from

https://www.researchgate.net/profile/Abhijay-Paliwal/publication/360069067_Analysis_o
f_Different_American_Streaming_Services_and_Shows/links/626005b0ee24725b3eb874
7f/Analysis-of-Different-American-Streaming-Services-and-Shows.pdf

[10] Chang, Zachary. Pae, Eleano. Xu, Kevin. Li, Annie (2021). Streaming Wars: Netflix, Prime

Video, Hulu, and Disney+. Retrieved from

https://ucladatares.medium.com/streaming-wars-netflix-prime-video-hulu-and-disney-c56
8a77a36ff

[11] Hidalgo-Mari, Tatiana. Segarra-Saavedra, Jesus. Palomares-Sanchez, Patricia (2021).

In-depth study of Netflix's original content of fictional series. Forms, styles and trends in

the new streaming scene. Retrieved from

https://www.researchgate.net/profile/Jesus-Segarra-Saavedra/publication/352020628_Rad
iografia_de_los_contenidos_originales_de_ficcion_seriada_de_Netflix_Formas_estilos_y
_tendencias_en_el_nuevo_escenario_in_streaming/links/60b60f7a299bf106f6ee5048/Ra
diografia-de-los-contenidos-originales-de-ficcion-seriada-de-Netflix-Formas-estilos-y-ten
dencias-en-el-nuevo-escenario-in-streaming.pdf

[12] Allegretti, Sebastian. Seidenstricker, Sven. Fischer, Heiko. Arslan, Sefkan (2021).

Executing a business model change: identifying key characteristics to succeed in volatile

markets. Retrieved from https://link.springer.com/article/10.1365/s42681-021-00020-x

[13] Wu, Xiaoxue. Zhou, Jiatong (2021). The Influential Factors of Developing Better in the

Global Stream Media Market: An Analysis of Netflix. Retrieved from The Influential

Factors of Developing Better in the Global Stream Media Market: An Analysis of Netflix

[14] Shetty, Sujala (2021). Sentiment Analysis, Tweet Analysis and Visualization on Big Data

Using Apache Spark and Haddop. Retrieved from

https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012002/pdf

[15] Graduate Programs Staff (2020). How Do I Start a Data Project: Understanding a Project

Lifecycle. Retrieved from

https://www.northeastern.edu/graduate/blog/data-analysis-project-lifecycle/

[16] IMDB Certificates by Country.

https://help.imdb.com/article/contribution/titles/certificates/GU757M8ZJ9ZPXB39?ref_=

helpart_nav_27#

[17] Soetewey, Antoine (2020). Stats and R. Retrieved from

https://statsandr.com/blog/anova-in-r/#anova-in-r

[18] Kassambara (2017). Articles - ggpubr: Publication Ready Plots. Retrieve From

http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/76-add-p-value

s-and-significance-levels-to-ggplots/

Appendix A

imdb_rating (title.ratings.tsv.gz)
- titleId : string
- imdb_rating : numeric

imdb_genre (title.basic.tsv.gz)
- titleId : string
- title : string
- genre : string
- release_year : integer

sentiment
- created_at : string
- id : numeric
- text : string
- query : string
- polarity : numeric
- subjectivity : numeric

streaming
- type : string
- title : string
- country : string
- date_added : date
- certificate : string
- service_name : string
- release_year : integer

imdb_join
- titleId : string
- title : string
- genre : string
- release_year : integer
- imdb_rating : numeric

streaming_imdb
- type : string
- title : string
- country : string
- certificate : string
- genre : string
- date_added : date
- year_added : integer
- release_year : integer
- imdb_rating : numeric
- service_name : string

Appendix B

E streaming_imdb
• title : string «Title of the Movie / Tv Show» • type : string «e.g. Movie / TV Show» • certificate : string «certificate of tile» • genre : string «includes up to three genres associated with the title» • date_added : date «Date title was added on» • year_added : integer «the year title is added» • release_year : integer «represents the release year of a title. In the case of TV Series, it is the series start year.» • country : string «Country where the movie / show was produced» • imdb_rating : numeric «weighted average of all the individual user ratings» • service_name : string «name of streaming service»

E sentiment
• created_at : datetime «datetime tweet created» • id : string «id of the tweet» • text : string «text of the tweet» • query : string «search blob provided to twitter API» • polarity : float «sentiment score» • subjectivity : float «identify sentiment subjective»

Appendix C

title <chr>	type <chr>	listed_in <chr>	service_name <chr>
10 Things I Hate About You	Movie	Comedy, Coming of Age, Romance	Disney
10 Things I Hate About You	Movie	Comedy, Drama, Romance	Amazon

title <chr>	type <chr>	listed_in <chr>	service_name <chr>
100 Streets	Movie	Drama	Hulu
100 Streets	Movie	Action, Drama, Suspense	Amazon

title <chr>	type <chr>	listed_in <chr>	service_name <chr>
21	Movie	Dramas	Netflix
21	Movie	Drama	Hulu
21	Movie	Drama, Suspense	Amazon

Appendix D

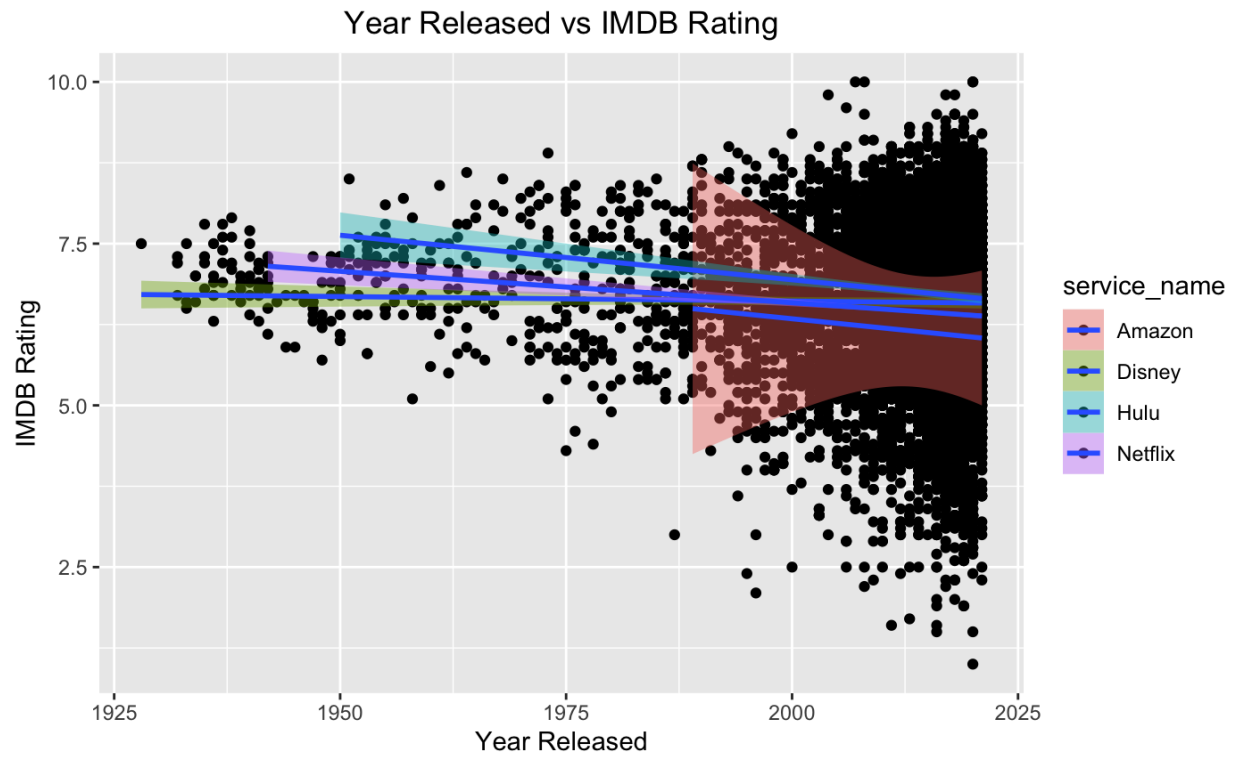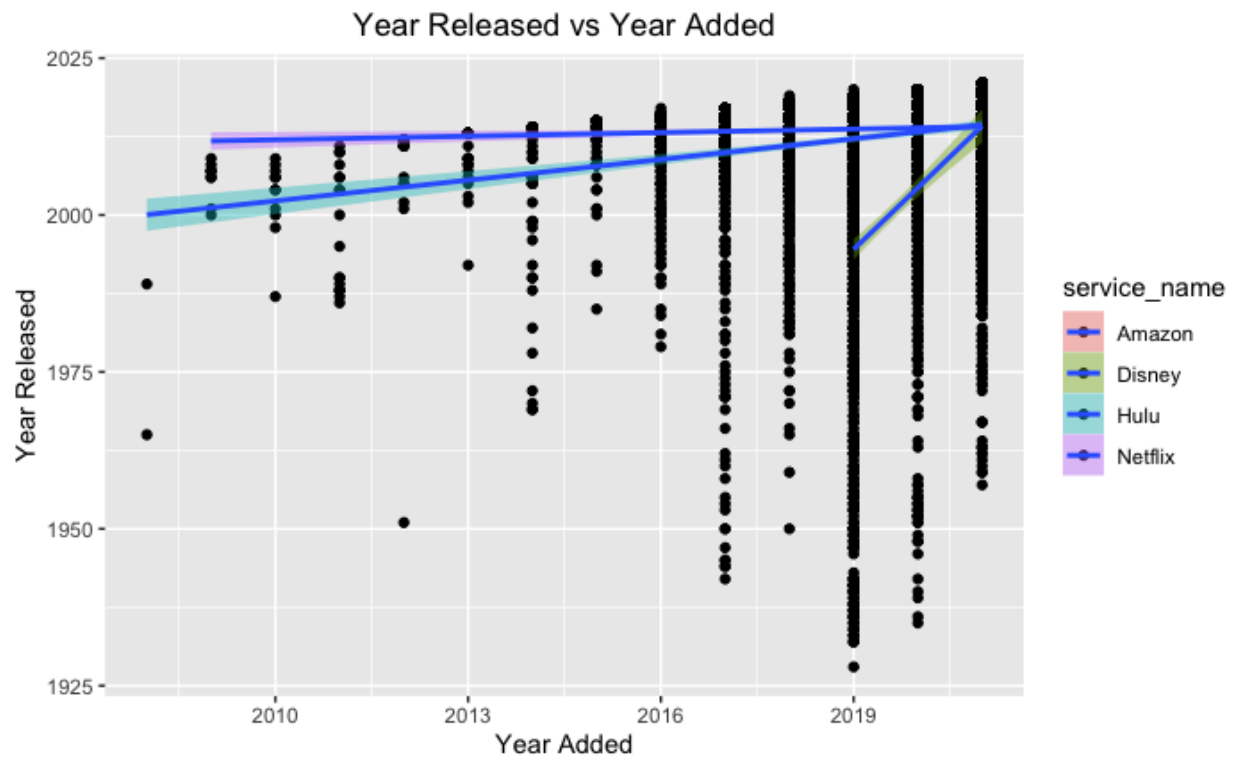| streaming_imdb	Netflix	Hulu	Disney	Amazon	Total	Netflix	Hulu	Disney	Amazon
type	0	0	0	0	0	0.00%	0.00%	0.00%	0.00%
title	0	0	0	0	0	0.00%	0.00%	0.00%	0.00%
country	883	1647	228	9459	12217	8.99%	47.42%	13.42%	92.41%
date_added	10	33	3	10079	10125	0.10%	0.95%	0.18%	98.47%
release_year	0	0	0	0	0	0.00%	0.00%	0.00%	0.00%
certificate	4	543	3	349	899	0.04%	15.63%	0.18%	3.41%
service_name	0	0	0	0	0	0.00%	0.00%	0.00%	0.00%
year_added	10	33	3	10079	10125	0.10%	0.95%	0.18%	98.47%
imdb_rating	2925	941	516	5485	9867	29.80%	27.09%	30.37%	53.59%
genre	2936	948	520	5512	9916	29.91%	27.30%	30.61%	53.85%
observation	9817	3473	1699	10236	25225				
average of missing value						6.89%	11.93%	7.49%	40.02%

Appendix E

Disney+ IMDB Rating on Movie

Disney+ IMDB Rating on TV Show

Amazon IMDB Rating on Movie

Amazon IMDB Rating on TV Show

Appendix F

Year Added vs IMDB Rating

Year Released vs IMDB Rating

Appendix G

Year Released vs Year Added
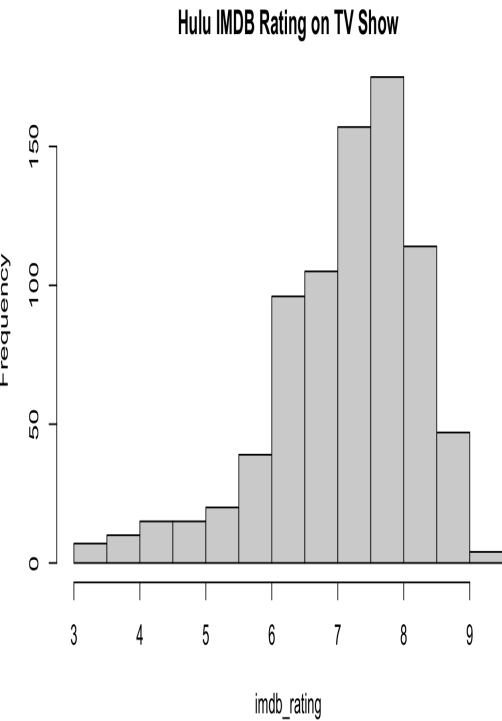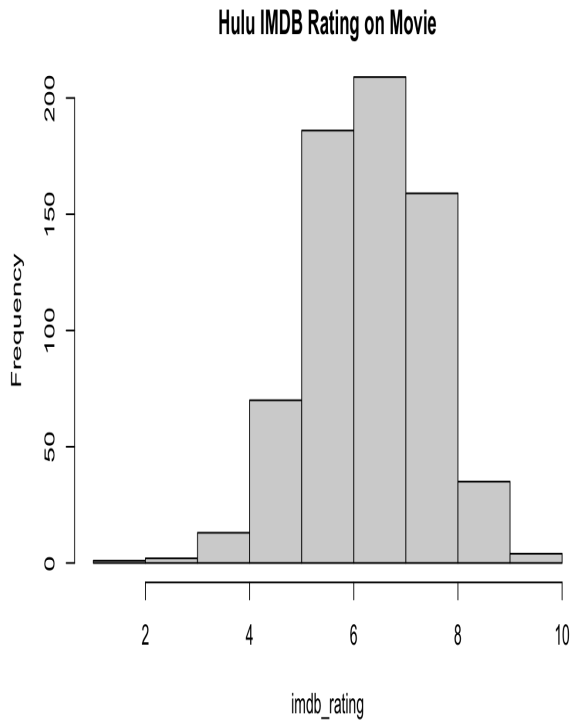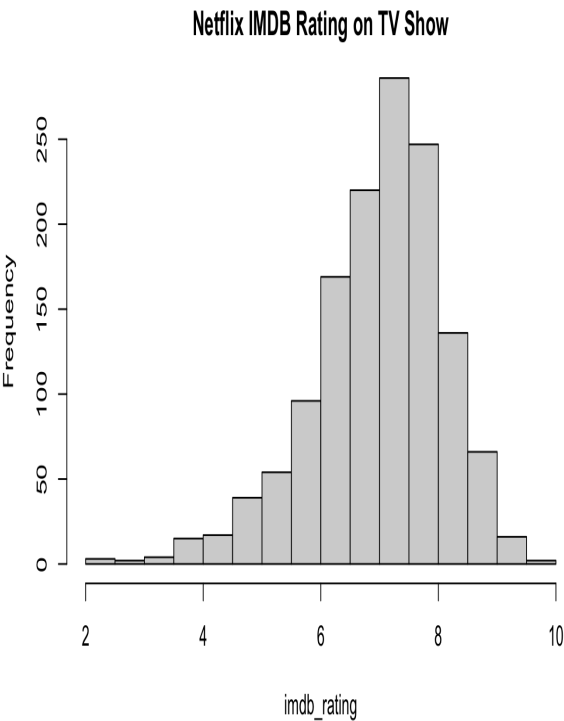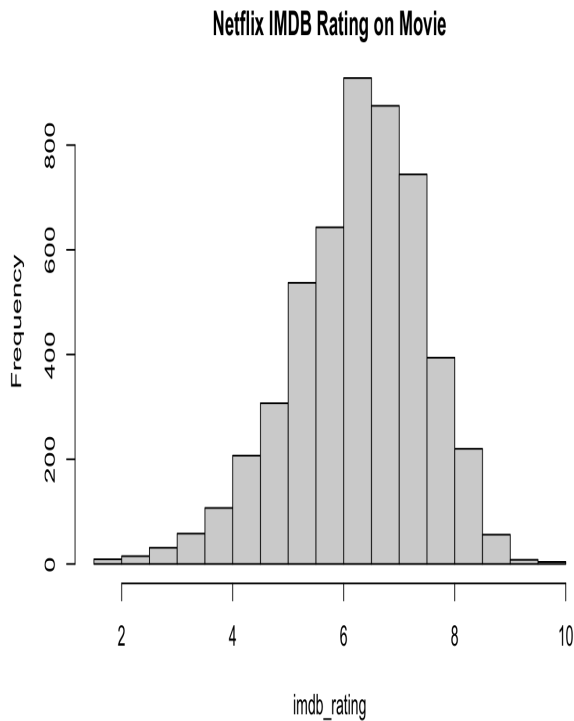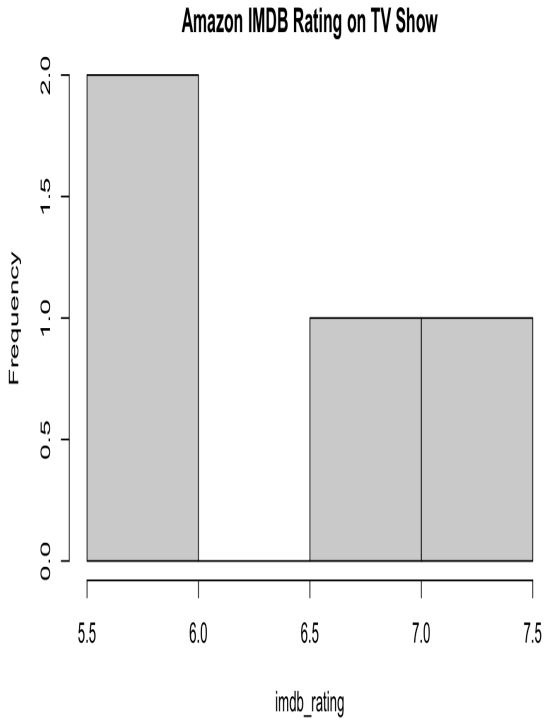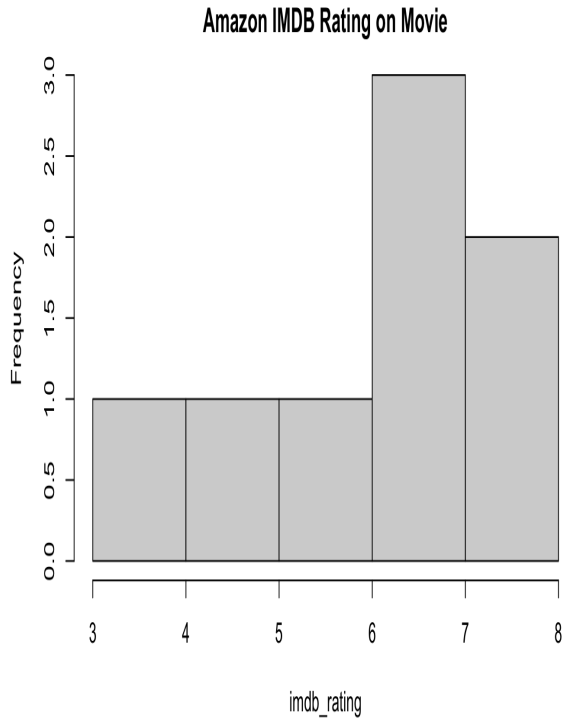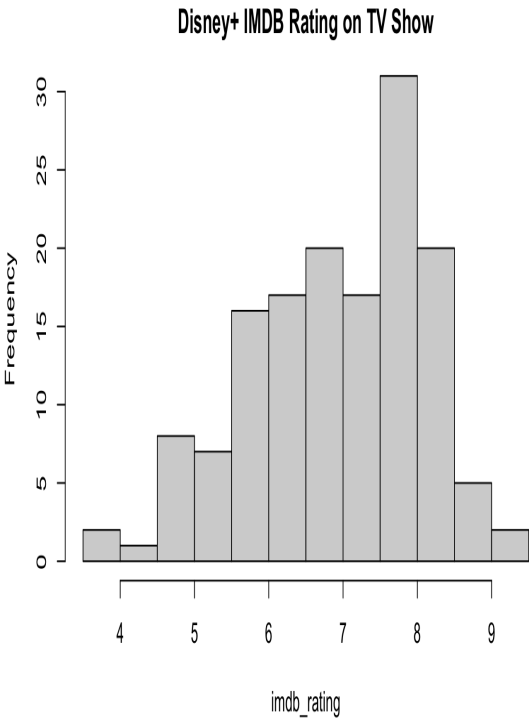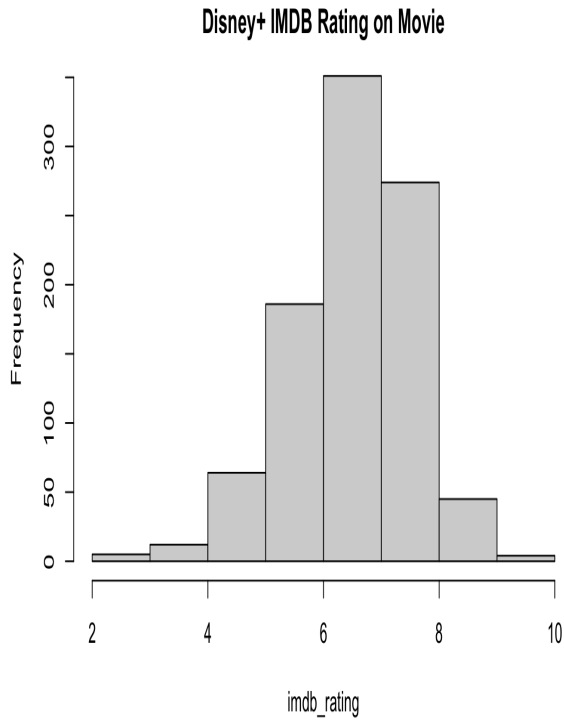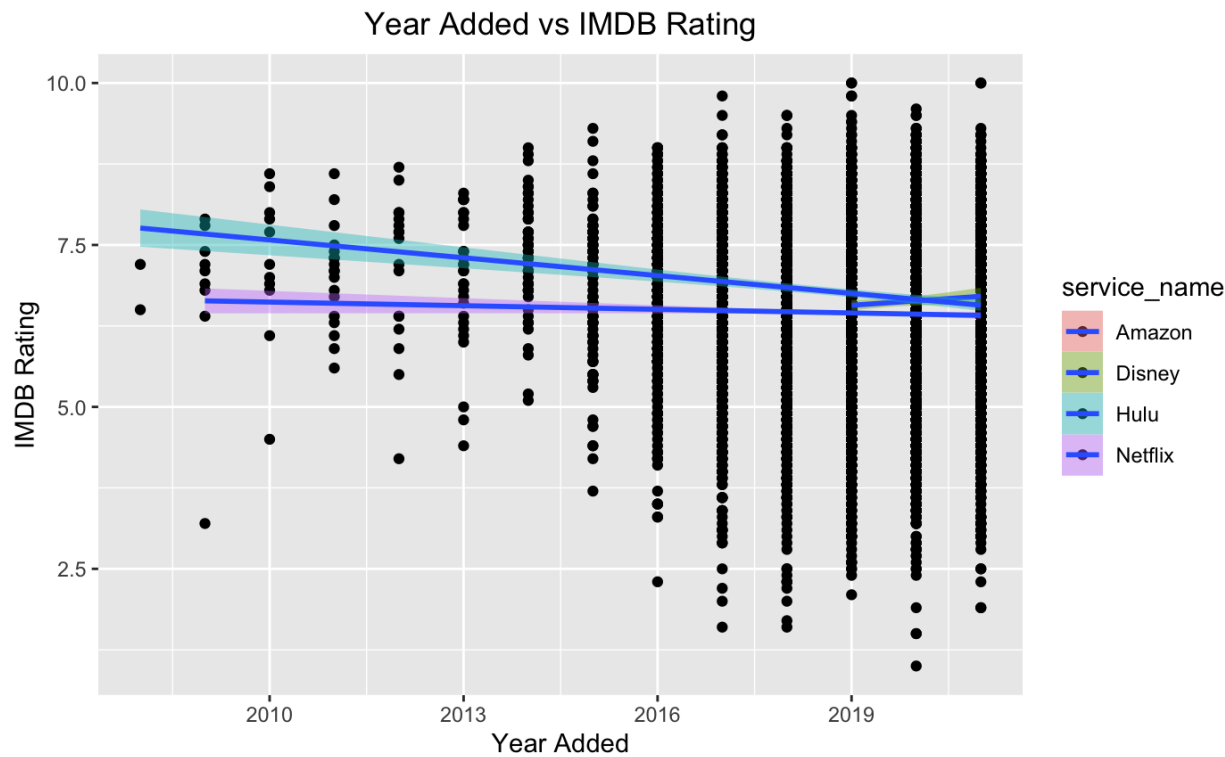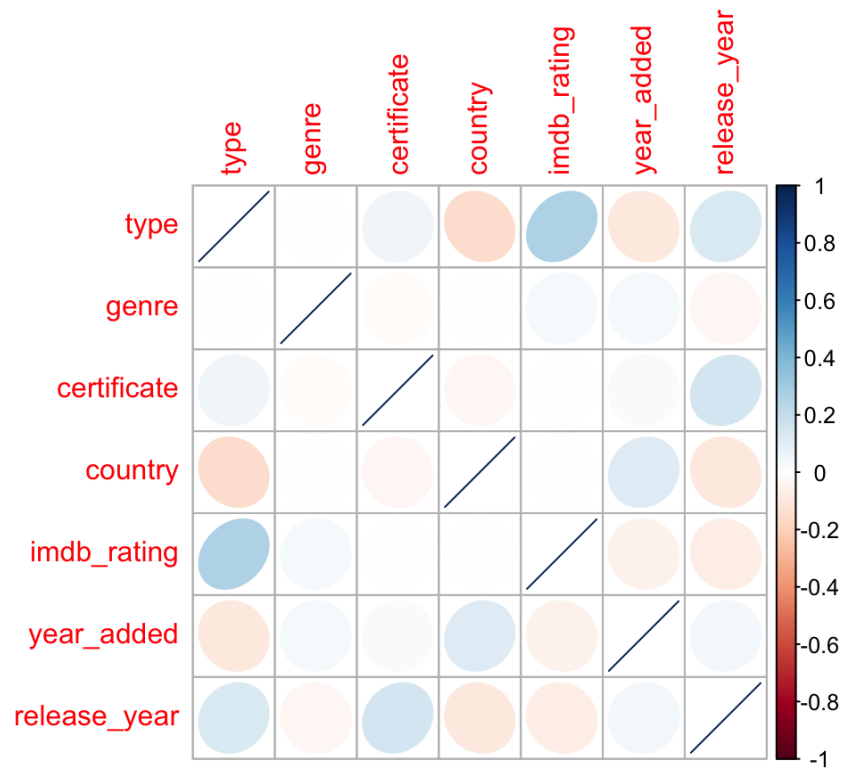
Appendix H

```
        Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = imdb_rating ~ service_name, data = df_number)

Linear Hypotheses:
                    Estimate Std. Error t value Pr(>|t|)
Disney - Amazon == 0   0.45389    0.34811   1.304  0.51310
Hulu - Amazon == 0     0.62994    0.34759   1.812  0.22672
Netflix - Amazon == 0  0.30598    0.34651   0.883  0.78469
Hulu - Disney == 0     0.17604    0.04799   3.668  0.00111 **
Netflix - Disney == 0 -0.14791    0.03940  -3.754  < 0.001 ***
Netflix - Hulu == 0   -0.32395    0.03455  -9.375  < 0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```
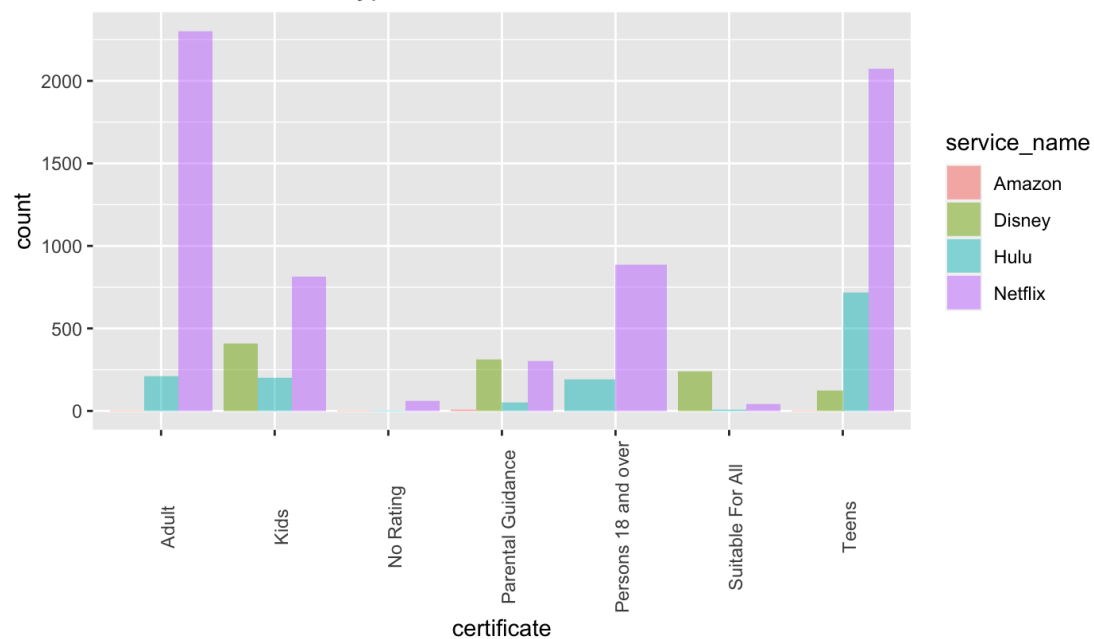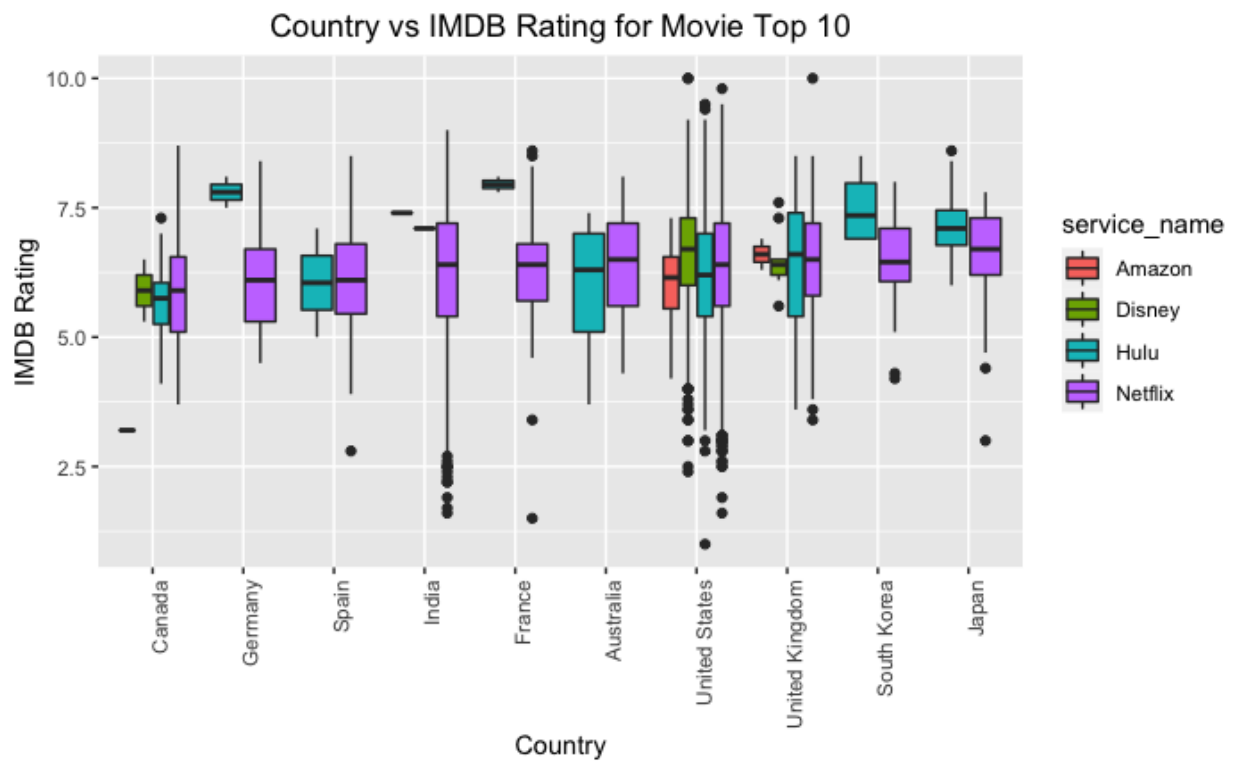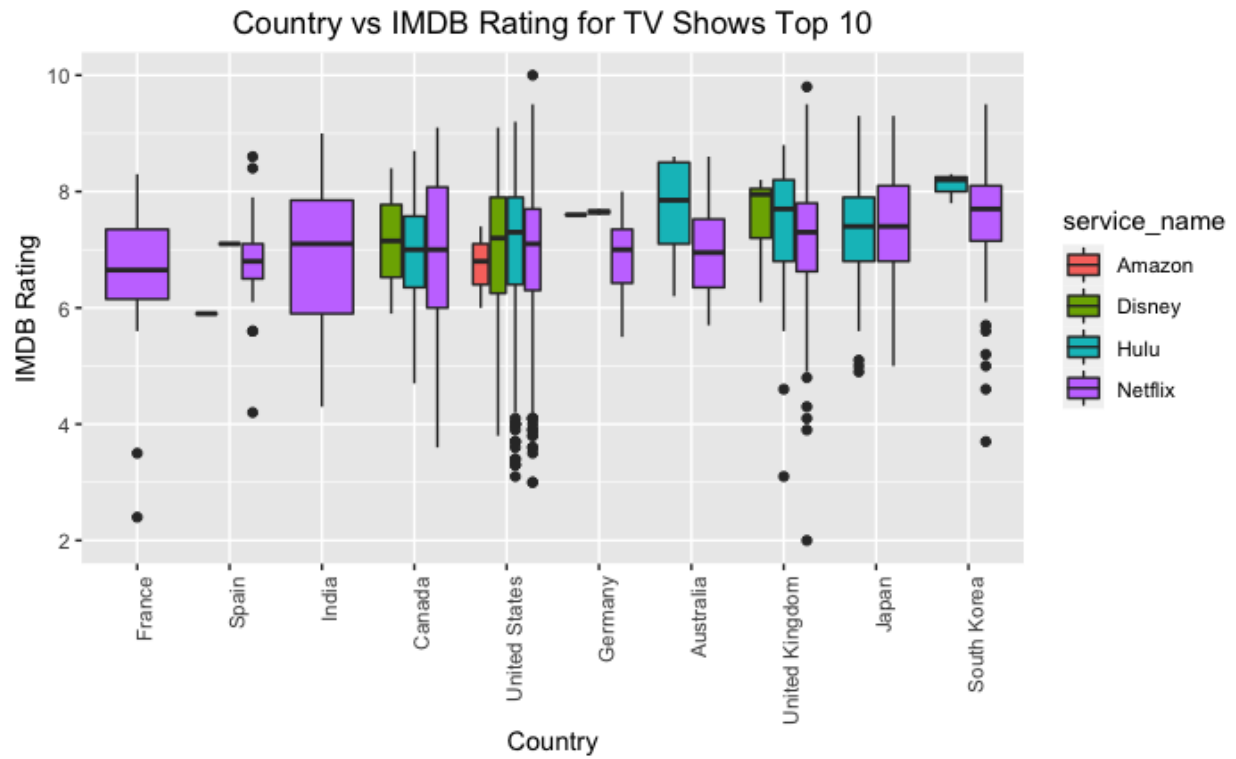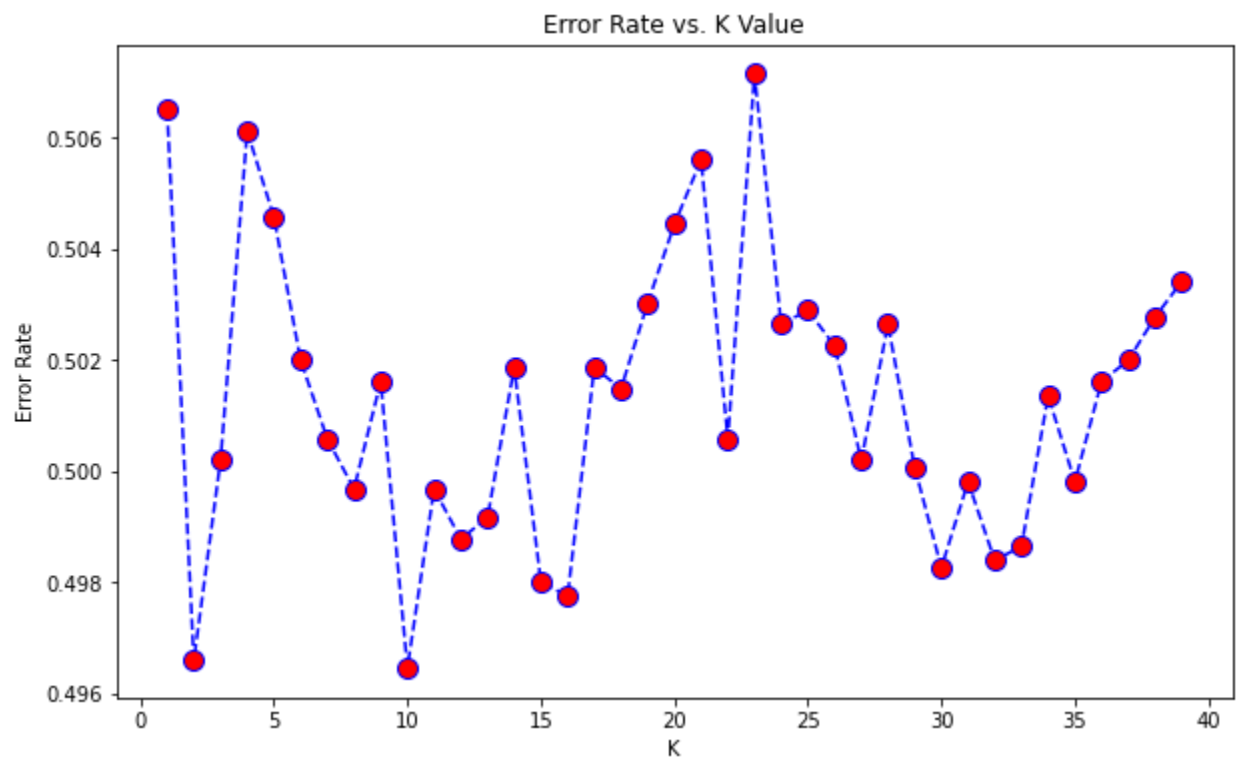
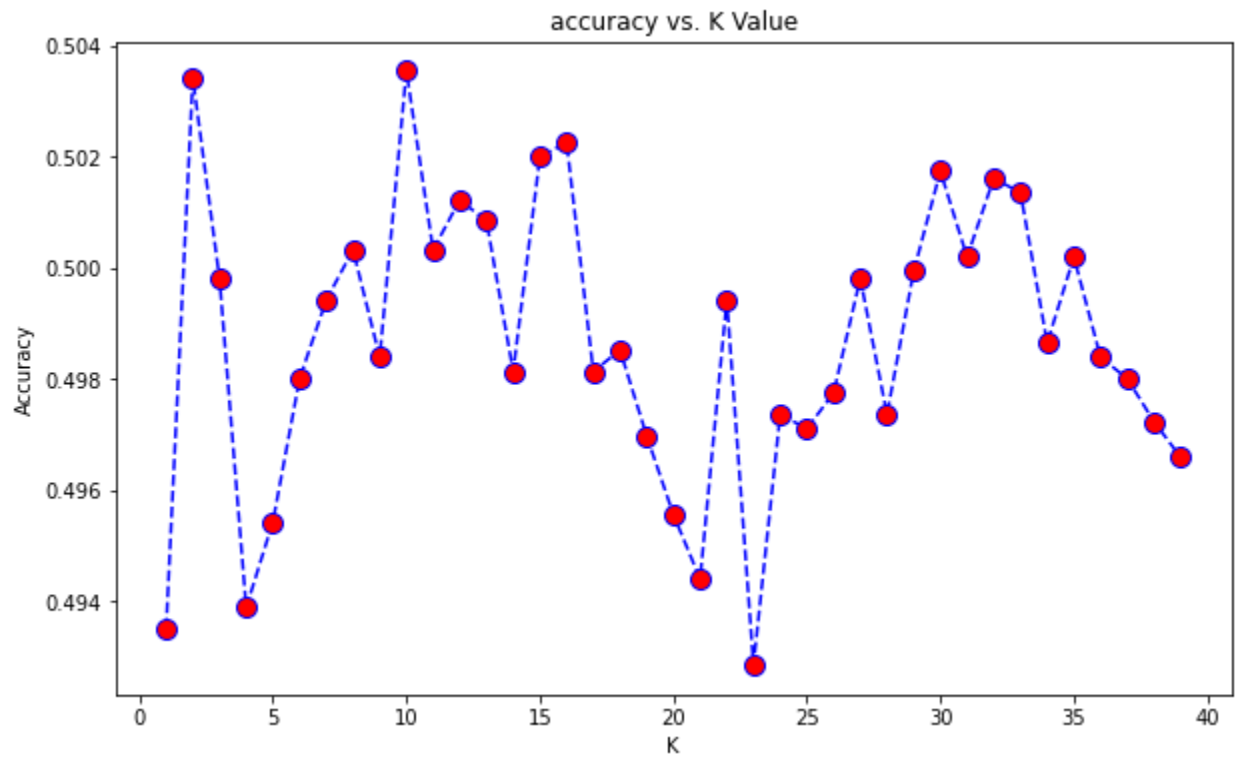**Appendix I**

All Type of Certificate After 2015

Appendix J

Country vs IMDB Rating for TV Shows Top 10

Appendix K

Error Rate vs. K Value

accuracy vs. K Value

Appendix L

Error Rate vs. K Value

Appendix M

Subjectivity