**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 6, 2022

**Table of Contents**

[**Abstract**](#_unlc5r1l9qx9) **3**

[**Literature Review**](#_n49r74vifxw5) **5**

[**Data Description**](#_d8cf2wotanlu) **13**

[**Data Approach**](#_ra7nscpkhux9) **17**

[Define Problem](#_ythmfyph1wdp) 17

[Data Understanding](#_jjvc4jtnkzw2) 17

[Data Preparation & Cleaning](#_3f57mplxgfse) 18

[Data Exploration](#_4fy44q9e089z) 20

[Design & Modeling](#_yjvsgxpw2u05) 23

[Evaluation](#_e77tk4v7g4re) 23

[Conclusion](#_2h926ml4yt) 24

[**Reference**](#_z3czp52j83tw) **25**

[**Appendix A**](#_8kcy4wqb3uwd) **28**

[**Appendix B**](#_6hyda5fa1p7k) **29**

# 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 and their corresponding techniques are listed below:

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

* The paper will use KNN classification/prediction to understand which platform has more variety of content, region, language, genres, etc., to target different audiences. The Cross Validation can be conducted to predict the trend year over year.

1. Is Netflix able to continue to provide the most new content?

* Using Naive Bayes Algorithm and Logistic Regression to predict which service provider has more and new content than others. The accuracy will be compared between these two models.

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

* The paper will explore the tweeter’s data to perform Sentiment Analysis to compare user’s comments of Netflix against its competitors by using Natural Language Processing (NLP). NLP 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 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 proposed to be used for this project are: Python, RStudio, ggplot2 and Tableau for data visualization.

# Literature Review

**Paper 1: Learning to Predict Movie Ratings from the Netflix Dataset.** [**[05]**](https://www.researchgate.net/publication/266457208_Learning_to_Predict_Movie_Ratings_from_the_Netflix_Dataset)

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]**](https://www.researchgate.net/publication/359747031_Data_Analysis_on_Netflix_datasets_Data_Analysis_on_the_Netflix_Datasets_Motivation)

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

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]**](https://www.theseus.fi/handle/10024/503990)

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]**](https://www.researchgate.net/profile/Abhijay-Paliwal/publication/360069067_Analysis_of_Different_American_Streaming_Services_and_Shows/links/626005b0ee24725b3eb8747f/Analysis-of-Different-American-Streaming-Services-and-Shows.pdf)

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]**](https://ucladatares.medium.com/streaming-wars-netflix-prime-video-hulu-and-disney-c568a77a36ff)

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. Netflix could release 2020 shows but added in 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 use regions and language tied to the version of the title streaming service offered. The more language Netflix provided in each title, the more diverse Netflix was.

**Paper 7: In-depth study of Netflix’s original content of fictional series. Forms, styles and trends in the new streaming scene.** [**[11]**](https://www.researchgate.net/profile/Jesus-Segarra-Saavedra/publication/352020628_Radiografia_de_los_contenidos_originales_de_ficcion_seriada_de_Netflix_Formas_estilos_y_tendencias_en_el_nuevo_escenario_in_streaming/links/60b60f7a299bf106f6ee5048/Radiografia-de-los-contenidos-originales-de-ficcion-seriada-de-Netflix-Formas-estilos-y-tendencias-en-el-nuevo-escenario-in-streaming.pdf)

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]**](https://link.springer.com/article/10.1365/s42681-021-00020-x)

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]**](https://www.atlantis-press.com/article/125961774.pdf)

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 language/IMDB rating/genre to see if Netflix has diverse targeted consumers. Regarding the place, we can analyze the region where the versions of titles were released 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]**](https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012002/pdf)

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 Natural Language Toolkit (NLTK) 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, country, duration, description) that will not be used as they are not related to the paper’s problem statement. There are two columns (e.g. release\_year, listed\_in) that should be selected from the IMDB dataset instead of from the streaming table. The explanation will be provided in the feature selection and attribution selection in the Dimensionality Reduction section below. In addition, three fields (e.g. month\_added, 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 “title\_certificate” and in the IMDB table will be renamed to “imdb\_rating”.

Beside the streaming table, some columns will be selected from the three IMDB tables (title.akas.tsv.gz, title.basic.tsv.gz, title.ratings.tsv.gz) because IMDB has consistent and accurate data. For example, the release\_year of title in Amazon Prime dataset is not matching to IMDB nor Netflix dataset even though the name of the title is the same. Genre description is used differently depending on the streaming services. As a result, these two fields will be selected from the IMDB table in order to have consistent comparative analysis. In order to check the diversity of language, the number of released language titles will be calculated and added into the table. All selected fields in the IMDB dataset will be joined by titleId and then joined with the streaming dataset by title, except imdb\_akas because multiple regions and languages can be offered in the same title. This is a different structure than other IMDB tables that is one id per record. Once the streaming table and IMDB table are joined, those titles without services and titleId will be dropped as a list of movies /TV shows in IMDB may include other companies’ content while this project only focuses on 4 streaming services. Now, the data increases to 178,262 observations of 12 variables. This observation will decrease more when we do complete.case for removing all missing values. The third dataset is sentiment data taken from Twitter using Tweepy library. The last dataset is country name based on region symbol which is easier to identify the distribution of titles. The relationship of tables with appropriate data type and names is identified in [[Appendix A]](#_8kcy4wqb3uwd). The description of the streaming table is in [[Appendix B]](#_6hyda5fa1p7k).

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. Based on the result of Pearson’s correlation, the correlation coefficient (-0.00697) is low and the R-square is even less than 1% meaning there is no linear relation between year added and IMDB rating. Same as the correlation coefficient between year added and release language, there is positive relationship but still low (0.0614) with adjusted R-squared (0.00374).

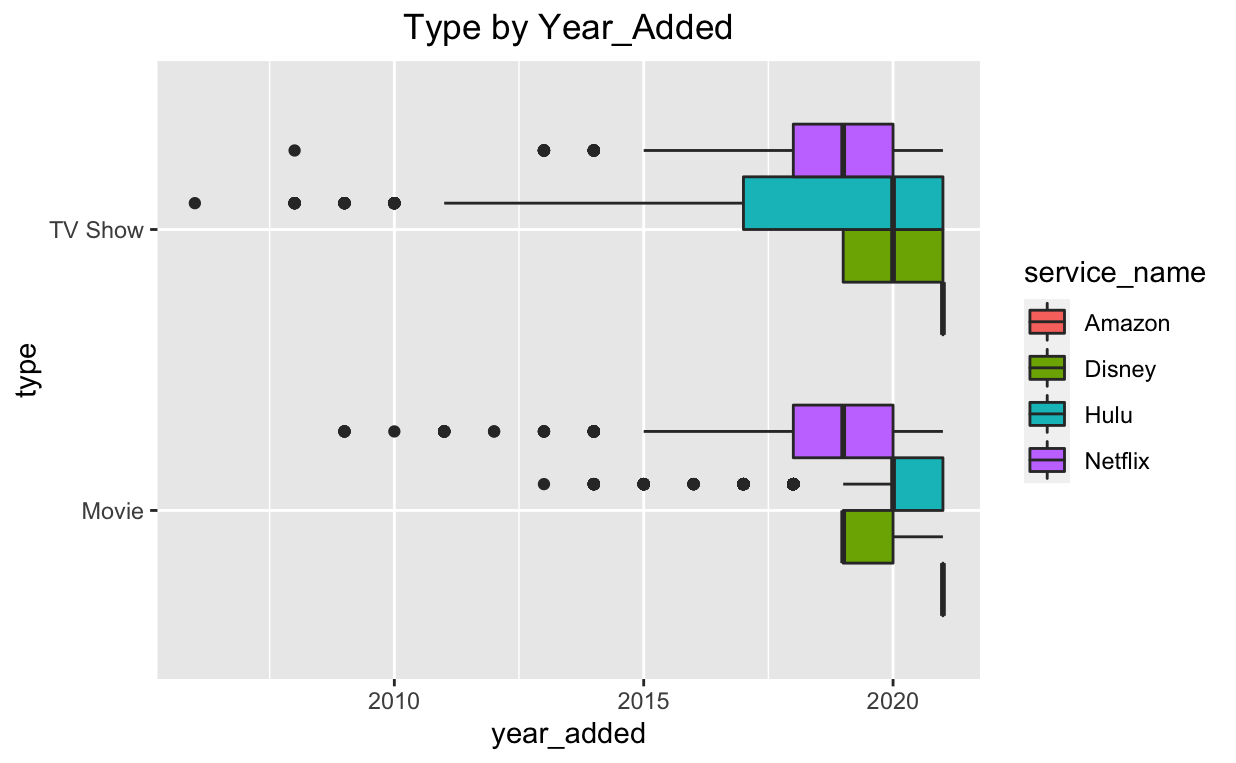
Below is the observation and the count of movies / TV shows by service:

| type | Netflix | Hulu | Disney | Amazon |
| --- | --- | --- | --- | --- |
| Movie | 50,688 | 15,333 | 7,556 | 59,856 |
| TV Show | 21,528 | 12,931 | 972 | 9,398 |
| Observation | 72,216 | 28,264 | 8,528 | 69,254 |

Below is the mean of language and IMDB\_rating by service:

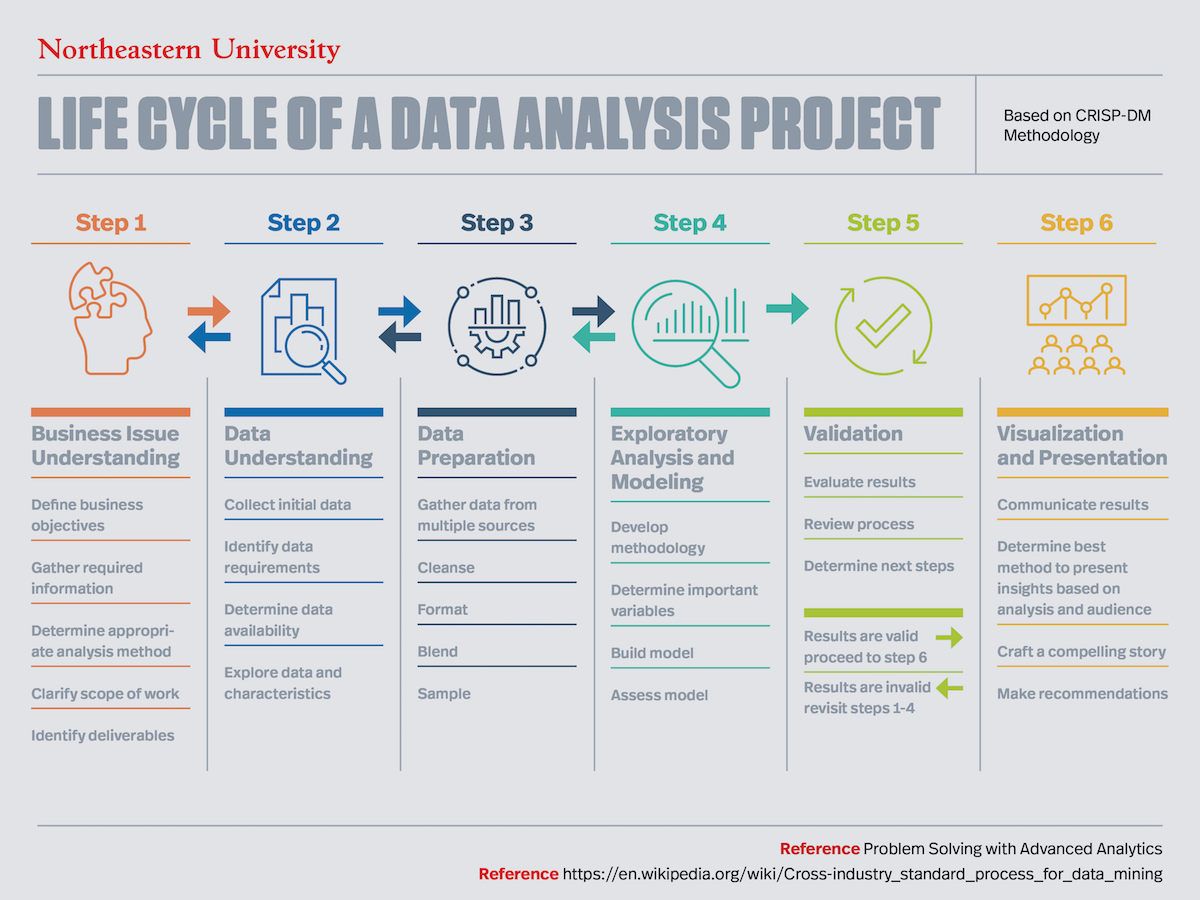
|  | Netflix | Hulu | Disney | Amazon |
| --- | --- | --- | --- | --- |
| Mean of language | 8.753 | 9.084 | 12.8 | 7.603 |
| Mean of imdb\_rating | 6.729 | 6.797 | 6.671 | 6.913 |

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 is not working [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 as the problem statement is if Netflix can still maintain its leadership after COVID-19.



In order to return a logical vector with complete cases, all missing values will be removed. The observation decreases from 178,262 to 32,191 which is 82% overall. In addition, the missing value of title\_certificate can be replaced by mode and re-organized the levels that decrease from 98 levels to 7 levels with better analysis.

# 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 project, 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 is the details of missing value in each table, and also the average of missing values:

|  | Netflix | Hulu | Disney | Amazon | Netflix | Hulu | Disney | Amazon |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| title | 0 | 0 | 0 | 0 | 0.00% | 0.00% | 0.00% | 0.00% |
| type | 0 | 0 | 0 | 0 | 0.00% | 0.00% | 0.00% | 0.00% |
| title\_certificate | 3 | 3029 | 1 | 2311 | 0.00% | 10.72% | 0.01% | 3.34% |
| genre | 1543 | 539 | 175 | 1442 | 2.14% | 1.91% | 2.05% | 2.08% |
| date\_added | 250 | 324 | 2 | 68537 | 0.35% | 1.15% | 0.02% | 98.96% |
| release\_year | 4117 | 1586 | 367 | 4159 | 5.70% | 5.61% | 4.30% | 6.01% |
| release\_language | 25220 | 10113 | 3983 | 23316 | 34.92% | 35.78% | 46.70% | 33.67% |
| imdb\_rating | 44927 | 17701 | 4831 | 43411 | 62.21% | 62.63% | 56.65% | 62.68% |
|  |  |  |  |  |  |  |  |  |
| observation | 72216 | 28264 | 8528 | 69254 |  |  |  |  |
| average of missing value |  |  |  |  | 11.70% | 13.09% | 12.19% | 22.97% |

1. “Date\_added” - The “date\_added” field has space in front of the month (e.g. “ October 3, 2021 “) and is trimmed before converting from character format (e.g. “October 3, 2021”) to date format (e.g. “2021-10-03”).
2. Add a new column named “year\_added” that will be used to compare “release\_year”. Also add “month\_added” to deep dive if the quantity of content increases a lot during the COVID-19 lockdown.
3. Add a calculation to get the number of languages for each title so that the paper can see how diverse they are.
4. There are 98 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.
5. Hulu has about 11% missing value and Amazon has about 3% missing value in the title\_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 |

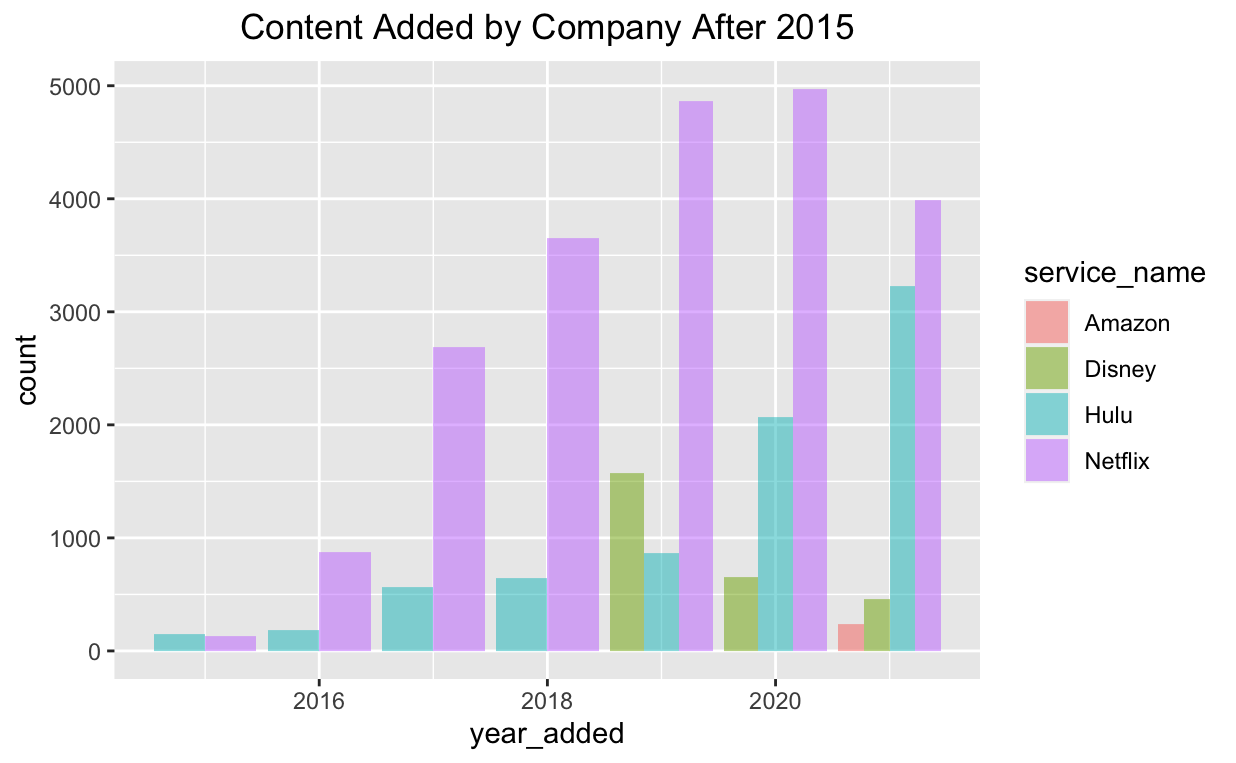
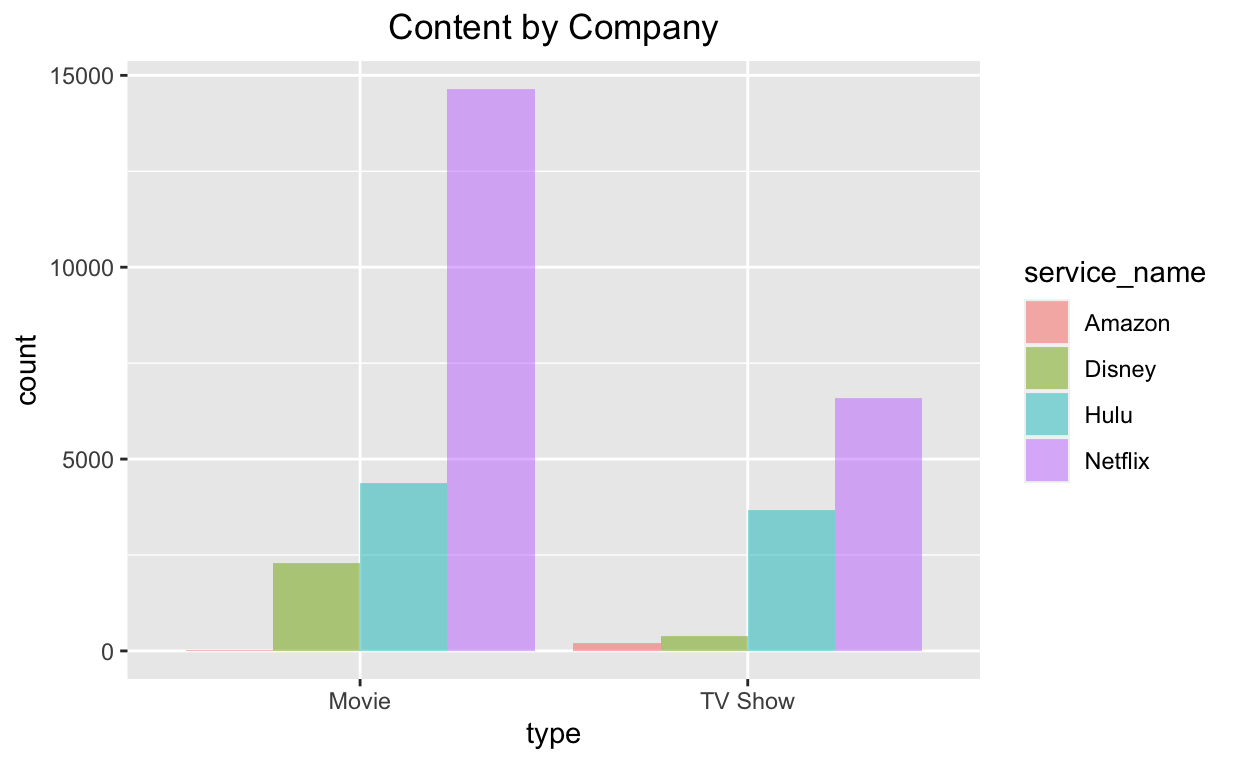
1. In order to have more control over what the output we can get, these fields’ data type will be efficiently changed from character to factor such as “type”, “genre”, “title\_certificate”, “service\_name”.
2. Remove all missing values to only return a logical vector with cases which are completed. The total observation decreases 82% but it still has 32,191 observations for analysis. Due to the high missing value ratio in Amazon Prime, the data may be imbalanced for doing the comparative analysis.

| type | Netflix | Hulu | Disney | Amazon | Total |
| --- | --- | --- | --- | --- | --- |
| Observation | 72,216 | 28,264 | 8,528 | 69,254 | 178,262 |
| New Observation | 21,237 | 8,037 | 2,678 | 239 | 32,191 |
| % Different | -71% | -72% | -69% | -100% | -82% |

1. Text Preprocessing for sentiment analysis such as removing punctuation, lowercase, stop words, stemming / lemmatization.

## Data Exploration

1. The bar graph below shows Netflix added 6 times more on Movies than other competitors, also double on TV shows. Netflix is still leading in the amount of content. However, Hulu is increasing their content while Netflix is slowing down adding new content. The leader of streaming service is possible to be taken by Hulu if Netflix is not able to reverse the trend.



1. Anderson-Darling Normality Test for normality is used to test if a sample of data came from a population with a specific distribution. Since the p-value < 0.05, it means the average of IMDB rating and the year added are not following a normal distribution. For an IMDB rating of 4 services, they are a bit left skewed both on movies / TV shows. For the year added to the title, the distribution is not even distributed for Disney+ and Amazon Prime. It may be because of a lack of “year\_added” data, especially because there is more than 98% missing value in “year\_added” in Amazon Prime.
2. Subsetting multiple variables and clustering the data based on different factors can be used to efficiently analyze by group.
3. Visualization is suggested to be worked on in this section for easier understanding.
4. Dimensionality reduction

Feature Selection

In order to improve performance of the model, the feature selection approach will be done to reduce the number of input variables while they are not related to our problem statement.

* “Show\_id” - is not unique after union as companies are using their own id
* “Country” - country where the movie / show is produced. The more production in one country does not explain which company is doing better.
* “Director” - director of the movie. 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.
* “Cast” - actors involved in the movie / show. Same as the reason as “director” above. It does not explain how the cast will be impacted by the diverse content.
* “Duration” - is total duration in minutes or number of sessions. 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.
* “Description” - summary of description of movies / TV shows. This is provided by the studio with standard format. Therefore, analysis cannot be very effective using this field.

Attribution Selection

Attribute selection process is needed as there are some concerns on the following fields:

1. “Release\_year” - Amazon Prime does not have the actual release\_year of the movie / show correctly after validating with the information in IMDB site which is the world’s most popular and authoritative source for movie, TV, and celebrity content. Thus, this data will be selected from IMDB.
2. “Listed\_in” - is a list of genres but different companies named differently. It is hard to do comparative analysis. Thus, this genre field will be selected from IMDB.

## Design & Modeling

1. design which algorithm is best fit to the model. The best practice is to split the data into training and test sets which is 70% / 30% respectively in this project. Cross validation will be used to get ROC.
2. Few classification models will be used in this project such as K-NN, Naive Bayes, Decision Trees, and/or Logistic Regression.
3. Bag of Words (BOW) Model will be used in sentiment.

## Evaluation

* The evaluation of the model consists in comparing the predicted value with the actual outcome. The loss function measures the variance between both values such as the mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE). The lower error is better. This project will also evaluate results by looking at the confusion matrix, receiver operating characteristic (ROC), recall, precision and accuracy.

Cross Validation

This paper will apply the partition rule to identify how many training sets and test sets. In this case, the training set will be set 70% and the test set will be set 30%. The split rule is to set 10 groups in the training set (e.g. 10 fold), repeat cross validation 3 times, and calculate ROC value. The “gbm” test is used to predict this model vs new data. Confusion matrix will also be used to show the decision model accuracy.

## Conclusion

* Provide a brief summary of the report such as its SWOT (strengths, weakness, opportunities, threats) about Netflix and 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-40360b31732>

[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_Ratings_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_datasets_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_of_Different_American_Streaming_Services_and_Shows/links/626005b0ee24725b3eb8747f/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-c568a77a36ff>

[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_Radiografia_de_los_contenidos_originales_de_ficcion_seriada_de_Netflix_Formas_estilos_y_tendencias_en_el_nuevo_escenario_in_streaming/links/60b60f7a299bf106f6ee5048/Radiografia-de-los-contenidos-originales-de-ficcion-seriada-de-Netflix-Formas-estilos-y-tendencias-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>

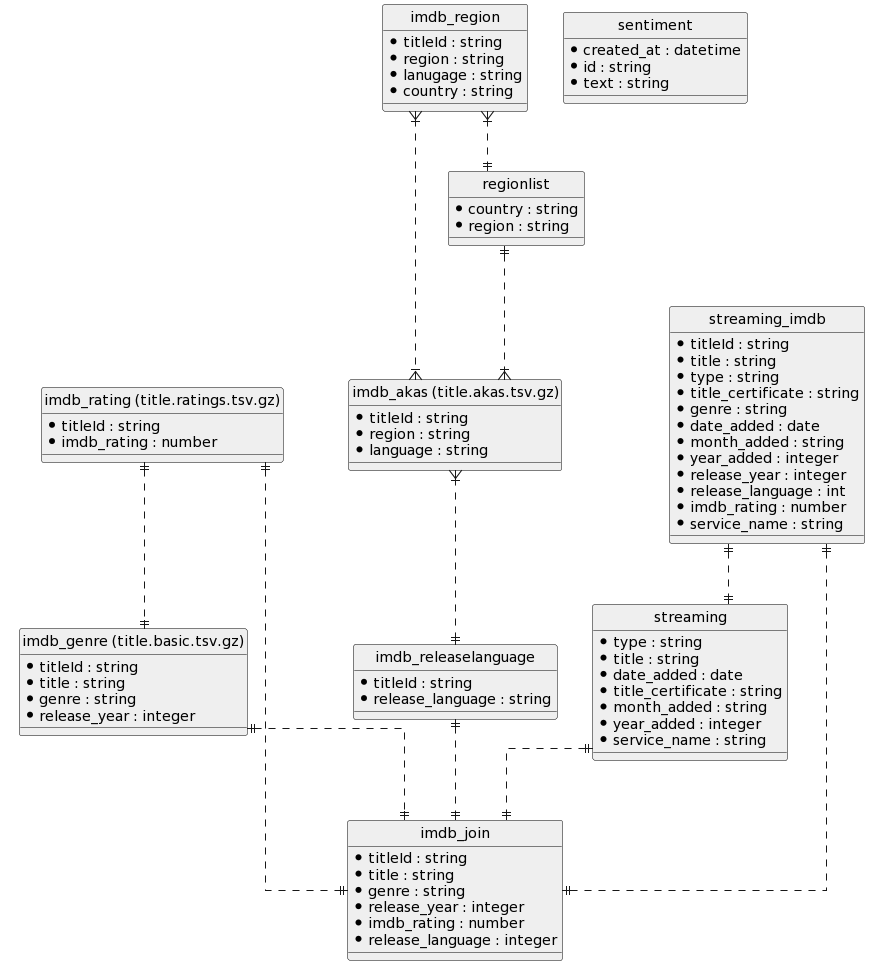
[13] Wu, Xiaoxue. Zhou, Jiatong (2021). The Influential Factors of Developing Better in the Global Stream Media Market: An Analysis of Netflix. Retrieved from [The Influential Factors of Developing Better in the Global Stream Media Market: An Analysis of Netflix](https://www.atlantis-press.com/article/125961774.pdf)

[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#>

# Appendix A



# Appendix B

