**Introduction to Data Management - Project Report**

**Netflix Show Analysis**

Nikhil Gudipally, Prajakta Ghumatkar, Sindhya Balasubramanian, Vishnu Rohan Surapaneni (Section 2)

**1.0 Summary**

Technological boom today has changed the way each of us have lived our lives. At the end of a long tiring day we come home and switch on entertainment in the form of Netflix, Hulu, Prime Video. These OTT platforms have become an intrinsic part of our lives.

With great power however comes great responsibility.

Being the people behind the scenes operating these OTT platforms, gives them the responsibility to keep the consumer or audience engaged with the content being produced. Key here is to note that the average attention span of a consumer watching content on an OTT platform is about 90 seconds. It is hence vital for movies and shows being produced to catch the attention of the viewer from the very start.

The problem that we hence wish to solve is to be able to understand the consumer best to be able to provide the right recommendations to increase viewership across shows/movies on netflix. To achieve this goal, we would be using a dataset that contains information on close to 15.5K Netflix shows and movies consisting of 26 descriptive attributes such as rating, tags, genre, writer, director, actor etc.

The goal here was to be able to take in a few movies that a user has seen and provide a list of recommended movies that the user would enjoy based on his/her previous choices. To achieve this goal, we made use of unsupervised learning to cluster movies and shows to understand how they related to each other in terms of their ratings and their popularity. Based on this, we then moved ahead to use the user provided content input to recommend other content within the same cluster that the user would enjoy.

(Netflix currently employs clustering as well to provide recommendations to each of its consumers based on the watching behavior and taste in movies and TV shows - [Link](https://medium.com/@springboard_ind/how-netflixs-recommendation-engine-works-bd1ee381bf81))

**2.0 Methods & Results**

The following were the steps taken to achieve the goal on hand as given below -

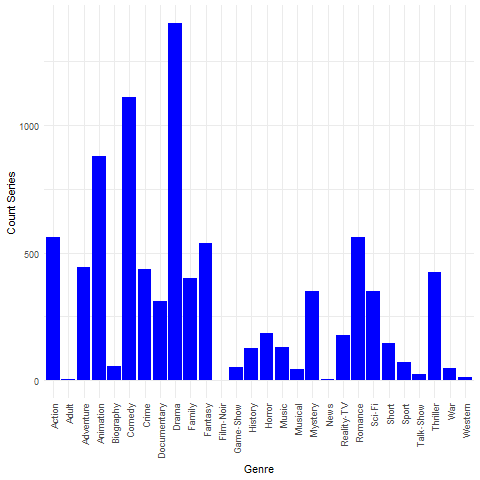
1. Processing (ETL)

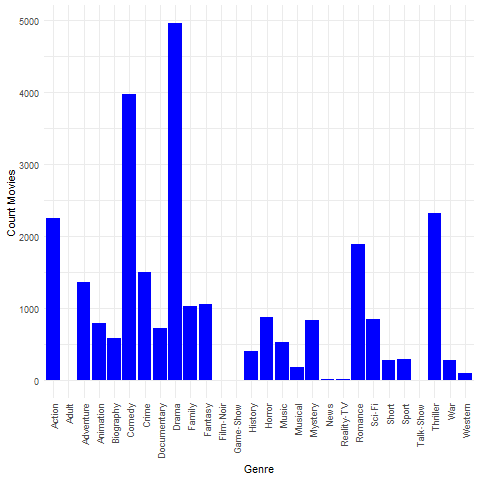
We started our project work with data processing and tidying the data on hand. We had about 15.5K rows of data on which attribute-wise the following processing was performed -

| SNo. | Attribute | Processing Step |
| --- | --- | --- |
| 1 | Title Series or Movie Hidden Gem Score Runtime View Rating Metacritic Score | - |
| 2 | Genre Tags Languages Country Availability  Director  Writer  Actors | Separated into different rows |
| 3 | IMDB Score Rotten Tomatoes Score | Combined to create Score Attribute |
| 4 | Awards Received  Awards Nominated For  Boxoffice  Release Date  Netflix Release Date  Production House  Netflix Link  IMDb Link  Summary  IMDb Votes  Image  Poster  TMDb Trailer  Trailer Site | Attribute Removed |

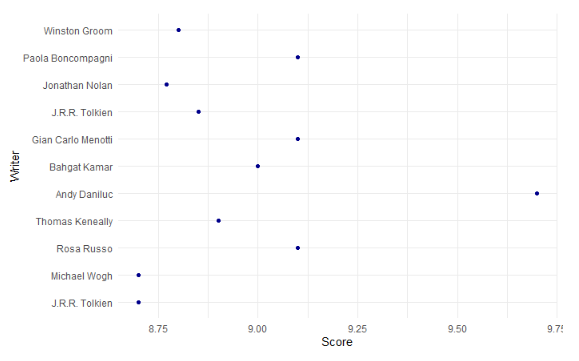
1. Visualization (EDA)

In this second step, we performed exploratory analysis on our data to understand the different attributes and the data distribution better. Below are a few key analysis outputs that we received -

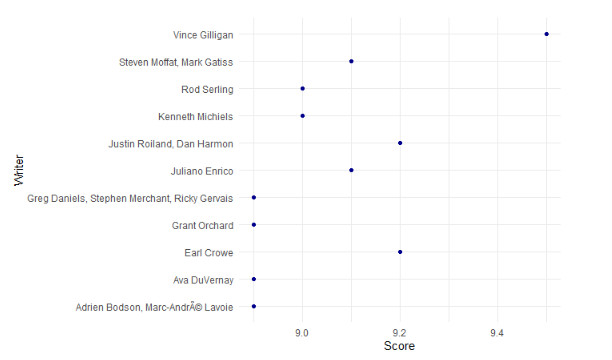
* Distribution of shows and movies basis genre



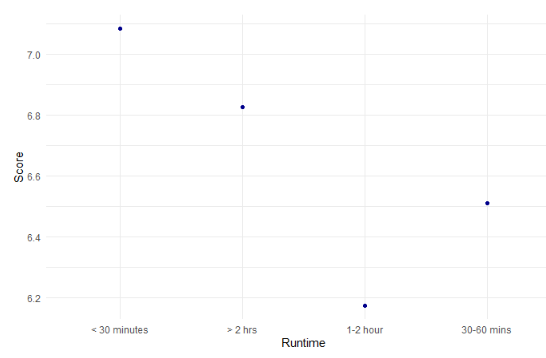
* Writer Ratings: Movies:



* Series:



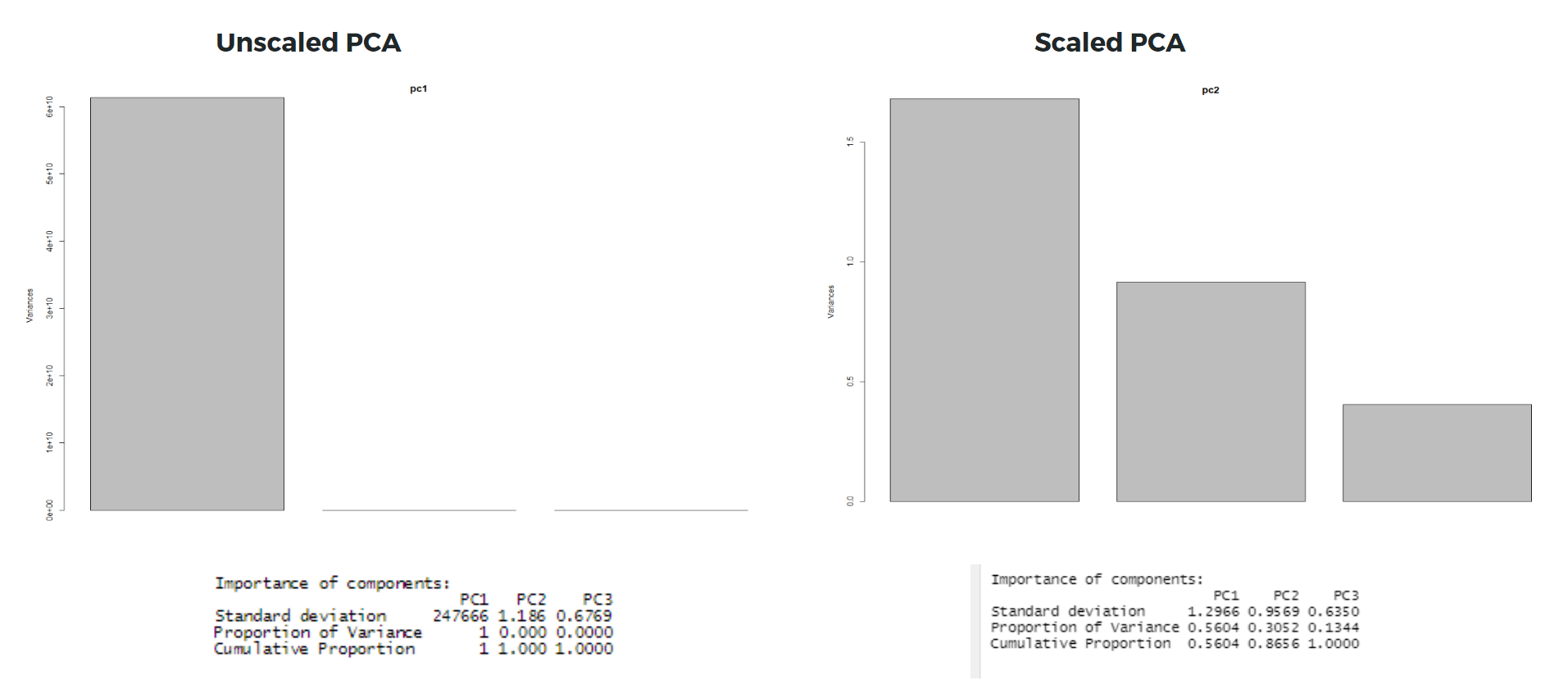
* Duration of tv show or movie vs rating:



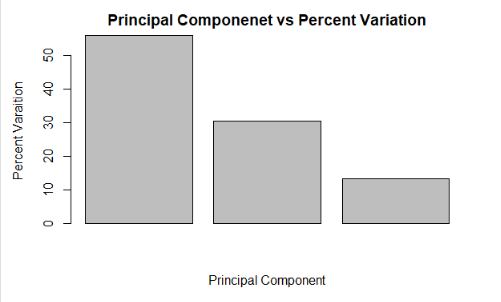
1. Dimension Reduction

We tried the following dimension reduction methods basis our analysis to see which attributes will be impactful for our analysis -

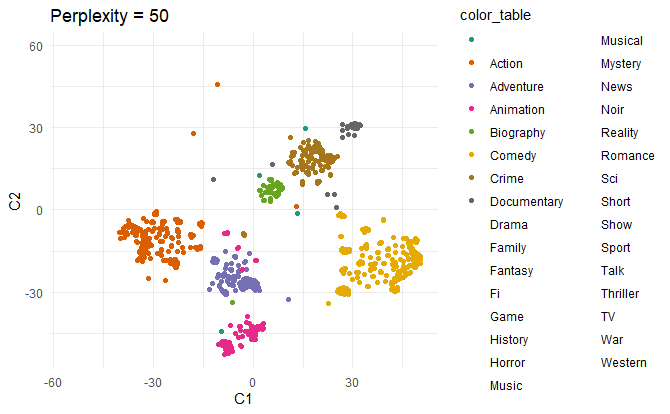
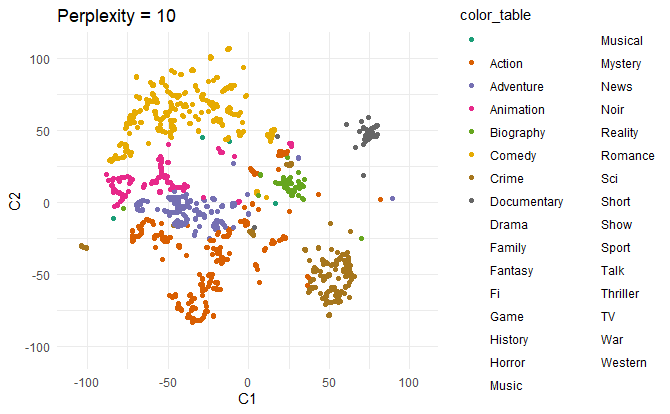
1. Principal Component Analysis (For linear relationships) -



* Principal component analysis (PCA) is the process of identifying components which are crucial in the dataset provided
* In the PCA analysis, the idea is to go with the principal components that account for the most variation in the data
* For our analysis, we used the scaled PCA and PRCOMP() functions
* From the cumulative proportion we saw that the first two components represented more than **85% of data.**
* From the Screen plot we determined that we only need the first two i.e., Score and Hidden.gem.score for our analysis.
* From the PCA Analysis, we determined how much variation each principal component accounts for and used the loading scores to determine the effect of variables on the PCA graph



1. t-Distributed Stochastic Neighbourhood Embedding (For non-linear relationships)

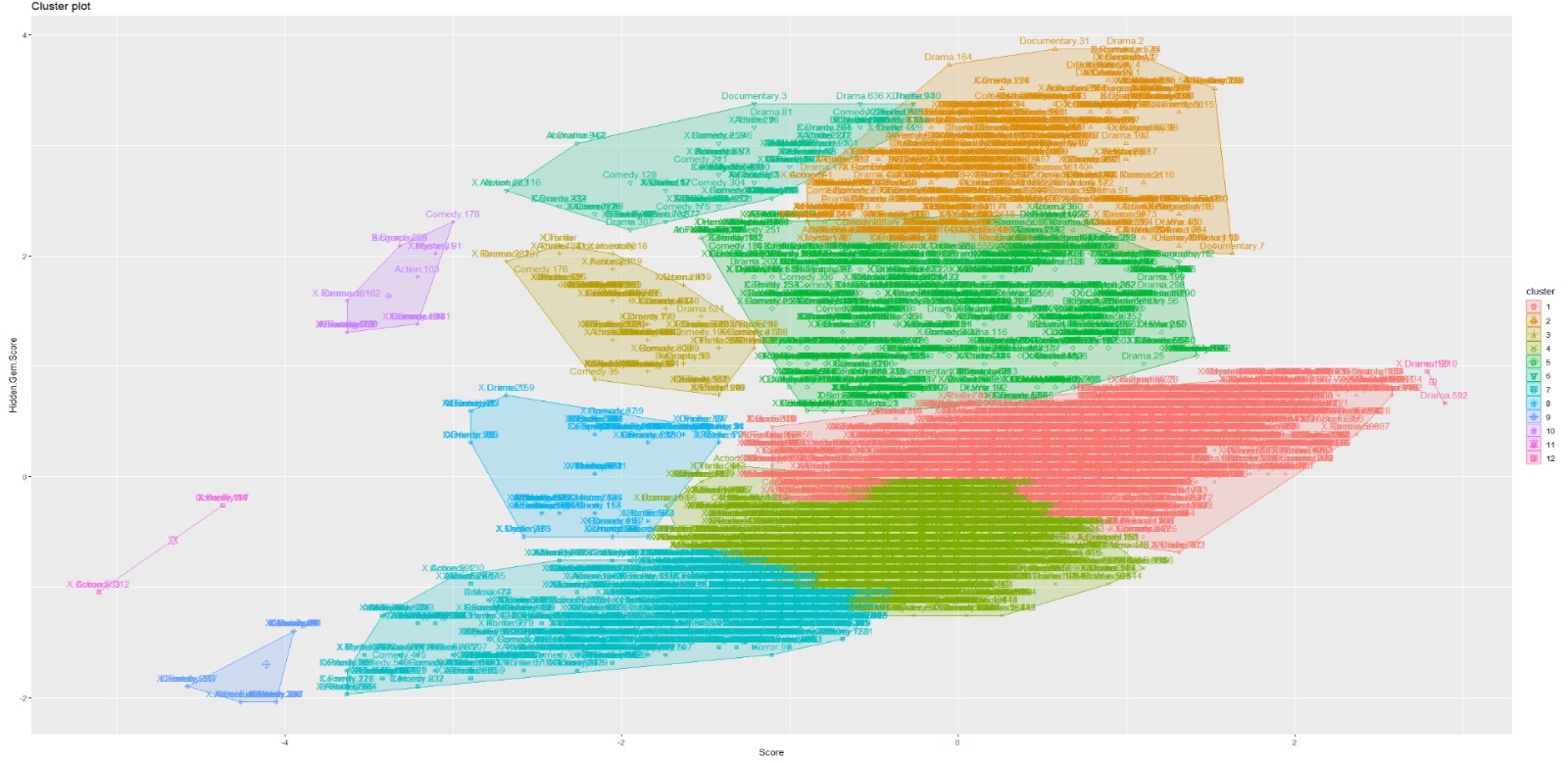


* t-SNE is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map
* We used categorical variables inclusive of Languages, Genre and found the distance between them using Gower distance from the Daisy library
* First, t-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects are assigned a higher probability while dissimilar points are assigned a lower probability.
* Second, t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence (KL divergence) between the two distributions with respect to the locations of the points in the map.
* Initially the overlapping was high, so we have increased the perplexity to have better defined clusters and the sample size was taken as 5000 for all the clusters

1. Clustering Techniques

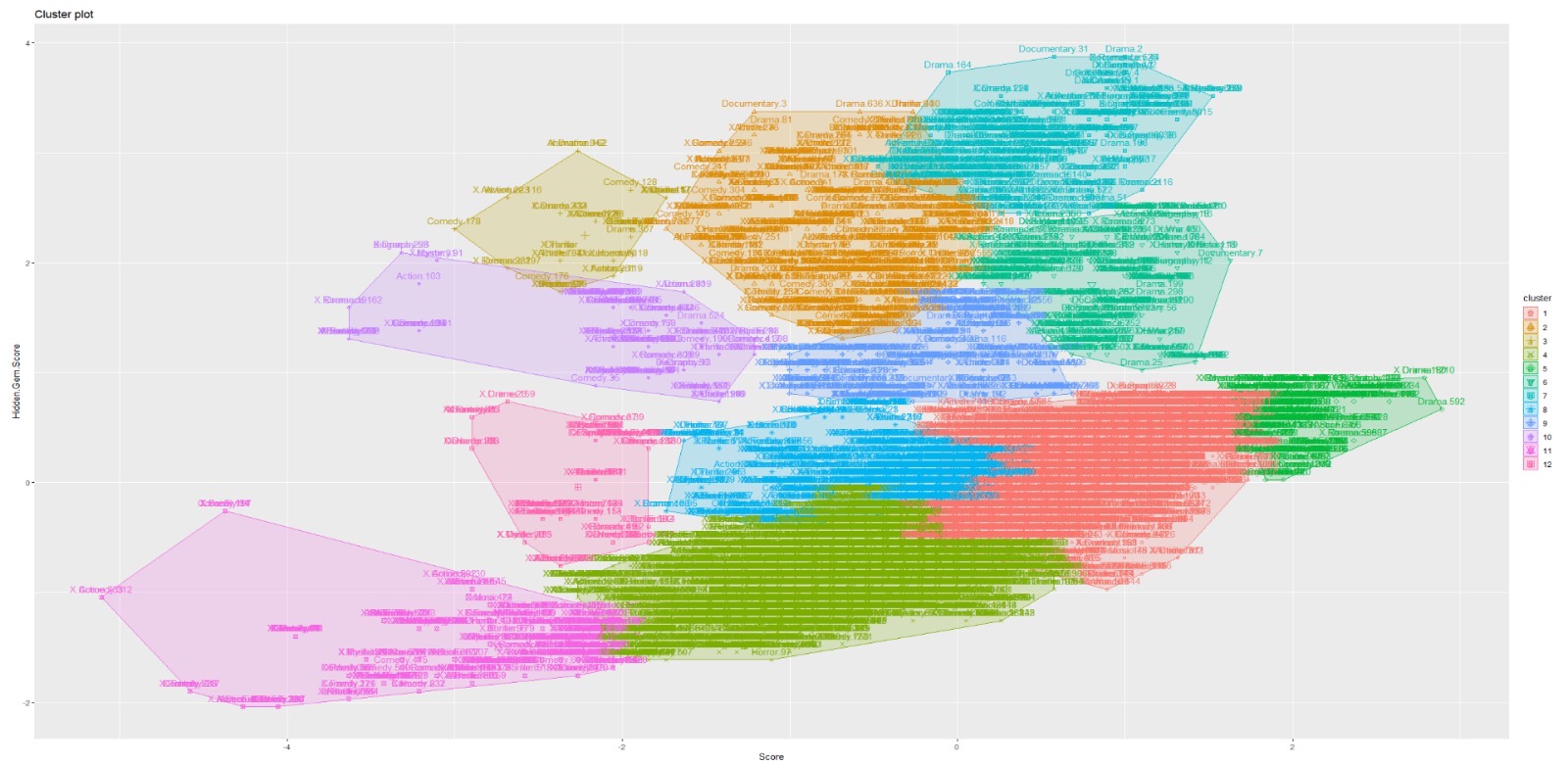
We tried the following clustering techniques to see which techniques give us the best results for our analysis -

1. Hierarchical Clustering



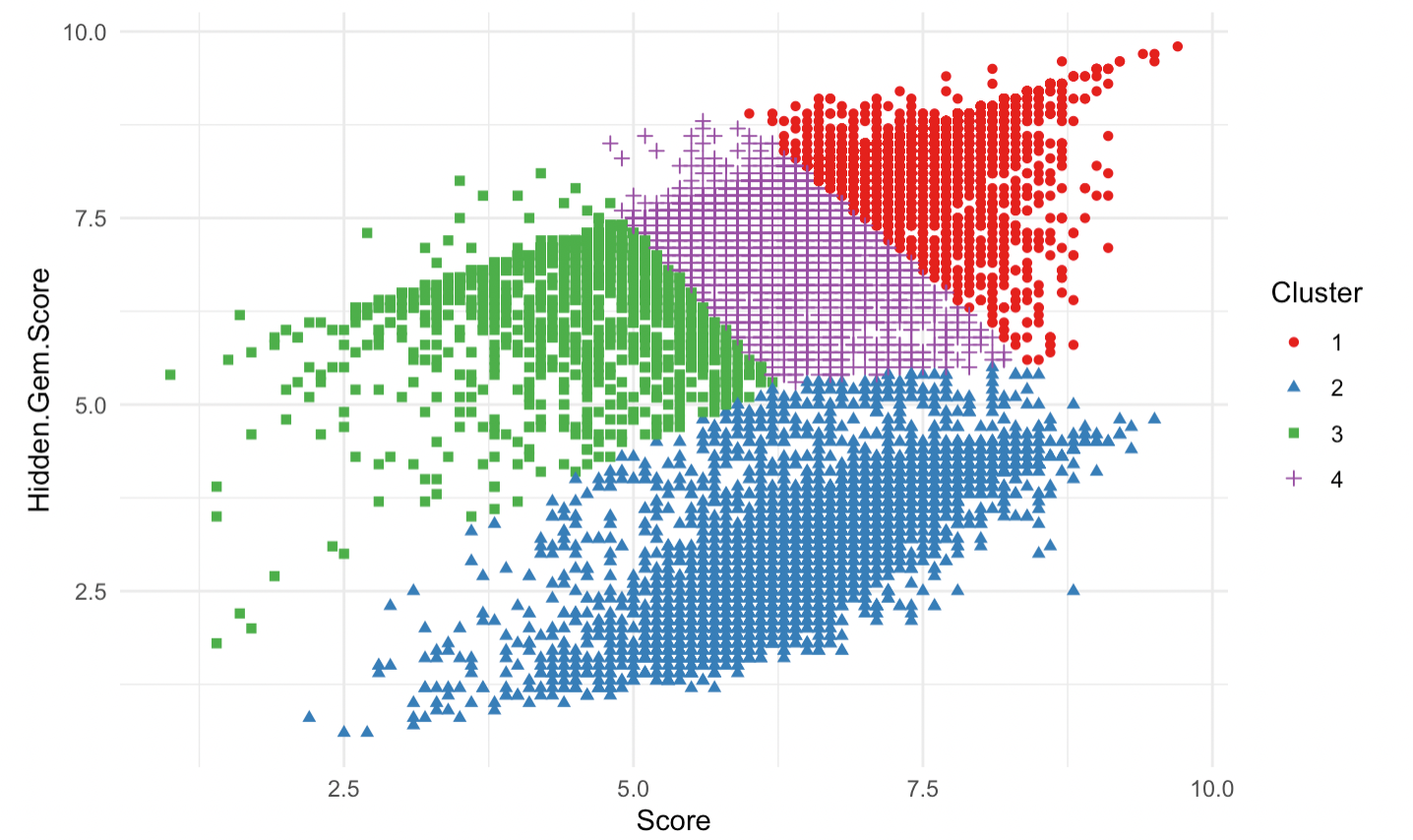
*Figure - Complete Linkage*

* Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. This can be looked at as a "bottom-up" or “top-down” approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
* Complete-linkage clustering is one of several methods of agglomerative hierarchical clustering. At the beginning of the process, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster. The method is also known as farthest neighbor clustering.
* The result of the clustering can be visualized as a dendrogram, which shows the sequence of cluster fusion and the distance at which each fusion took place



*Figure - Average Linkage*

* Similarly, the average linkage clustering is a method of calculating distance between clusters in hierarchical cluster analysis. The linkage function specifying the distance between two clusters is computed as the average distance between objects from the first cluster and objects from the second cluster.

1. k-Means Cluster

* K Means has identified centroids with data points clustered all around the centroid.
* K-Means doesn’t allow nearby data points to share the same cluster no matter how obvious the between them might be.

**3.0 Discussion**

These results can be broken down into different categories inclusive of Visualization results, Dimension reductionality results and finally modelling results. Below are the relevance of each of the results found -

* Visualization results - We note that categorical variables such as the Writer who wrote the script, the Genre of the content and Language of production are highly indicative of the type of movies and shows tailored to engage the audience currently on Netflix. Such as the Director Christian has the highest average rating for his movie over time
* The PCA results indicate that the principal components from the continuous variables found have explained more than 85% of the variance noted in the data
* The modelling results noted from k-Means and Hierarchical clustering showcases a clear logical separation among the different clusters specially noted in k-means clustering

This project would benefit OTT platforms such as Netflix, Hulu and Prime to understand their consumers better and accordingly improve their overall recommendation systems. These clusters can be used to understand how different consumer viewing behavior and taste can be used to recommend similar movies and shows to the associated consumers.

This project can be further improved in the future to include categorical variables and visual information such as aesthetics to better cluster the viewership behavior and improve the type of movies and shows being recommended to the end user.

**4.0 Statement of Contribution**

Below are the names and contributions of each member of the team involved in the project -

* Nikhil Gudipally - The contributions are as given below -
  + k-Means clustering and associated results visualization
* Prajakta Ghumatkar - The contributions are as given below -
  + Data cleaning and tidying as dictated in Section above
  + Visualizations of various continuous and categorical variables
  + Gower distance for t-SNE calculation for dimension reductionality
  + t-SNE dimension reductionality and plotting of associated results
* Sindhya Balasubramanian - The contributions are as given below -
  + Data cleaning and tidying as dictated in Section above
  + Visualizations of various continuous and categorical variables
  + Gower distance for t-SNE calculation for dimension reductionality
  + t-SNE dimension reductionality and plotting of associated results
* Vishnu Rohan Surapaneni- The contributions are as given below -
  + Data cleaning and tidying as dictated in Section above
  + Visualizations of various continuous and categorical variables
  + PCA dimension reduction and associated results visualization
  + Hierarchical clustering and associated results visualization

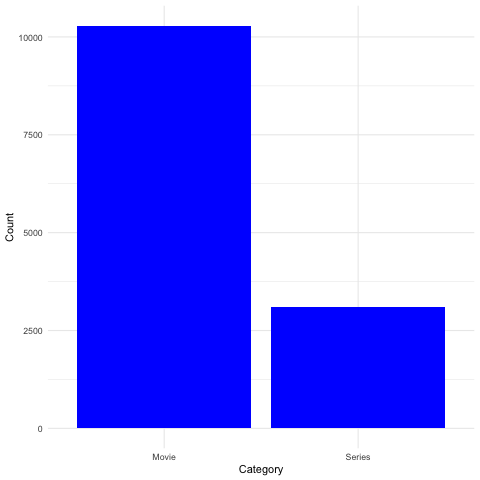
**5.0 References**

1. Dataset : <https://www.kaggle.com/ashishgup/netflix-rotten-tomatoes-metacritic-imdb>
2. Understanding Gower Distance : <https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3>

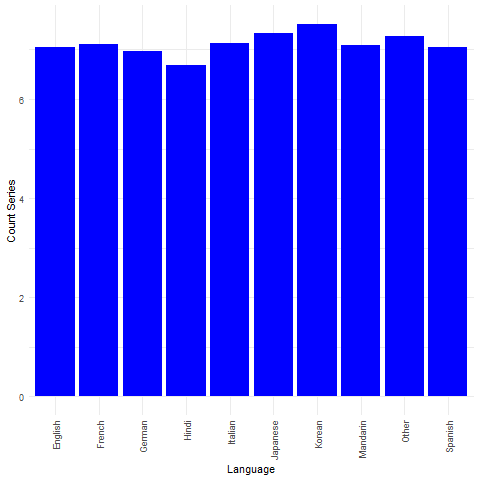
**6.0 Appendix**

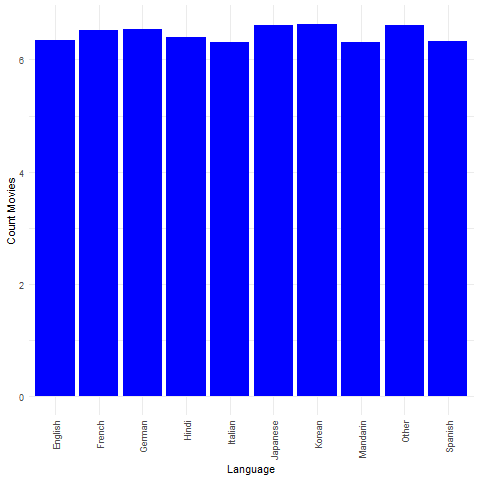
1. Exploratory Data Analysis Outputs -

* Distribution of shows and movies as individual categories

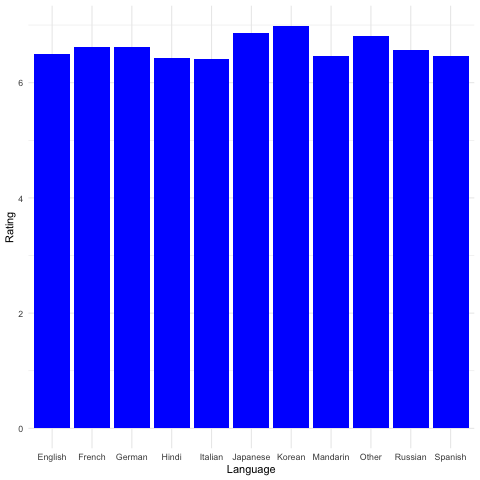


* Distribution of shows and movies across languages

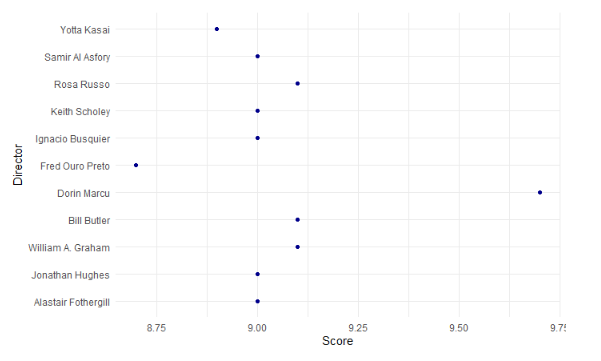




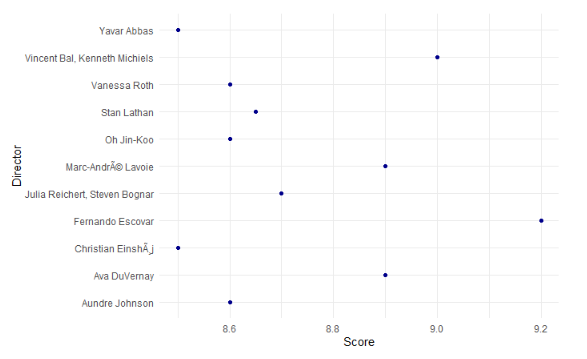
* Language level analysis of ratings



* Director Ratings (top 11): Movies:



Series:



* Duration of tv show or movie vs rating:

