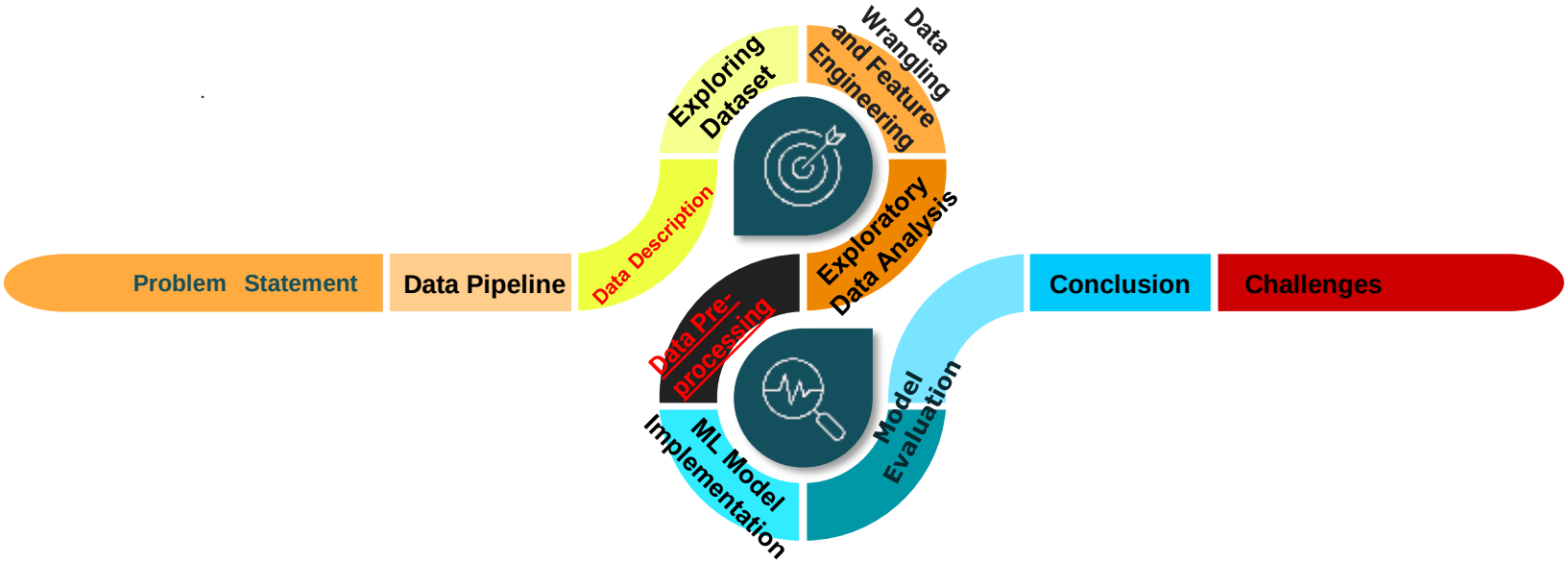
Capstone Project

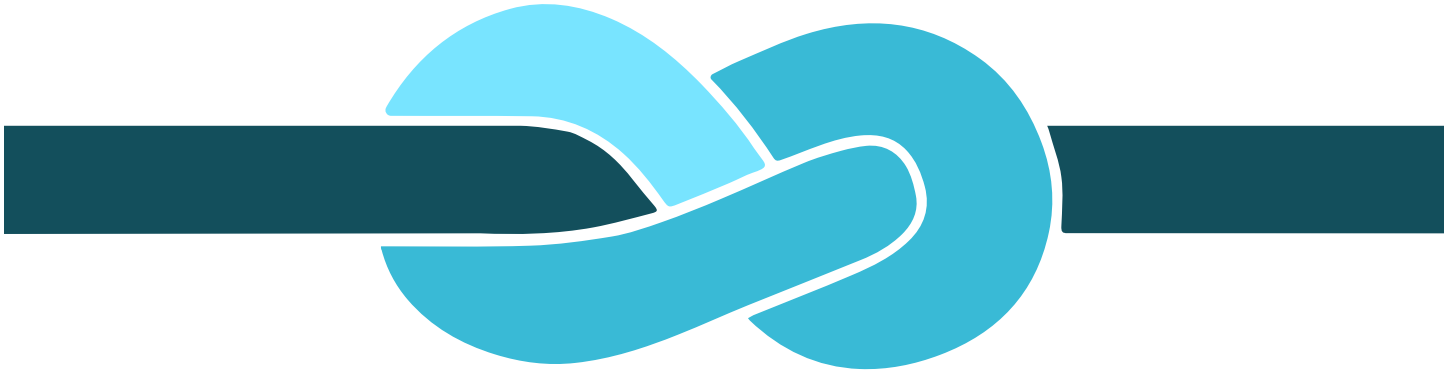
Netflix Movies and Tv Shows Clustering

Ajit Sharad Mane
(ajitmane36@gmail.com/+918888288752)

Index


Problem Statement
Data Pipeline
Data Description
Exploring Dataset
Data Wrangling and Feature Engineering
Exploratory Data Analysis
Data Pre-processing
ML Model Implementation
Model Evaluation
Conclusion
Challenges

Problem Statement



This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

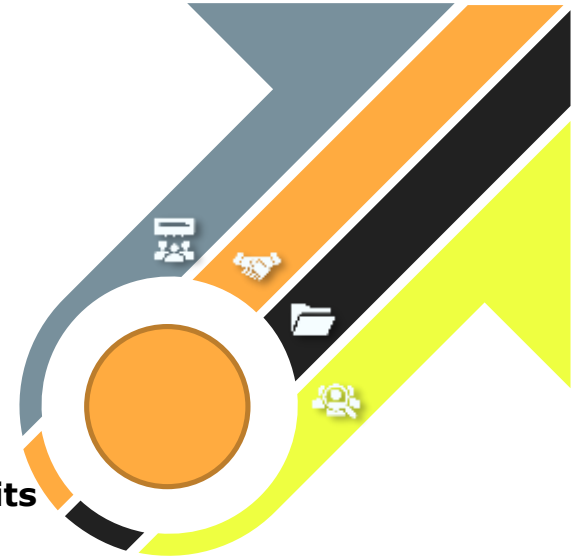
Data Pipeline

Data Description

Sr.No.	Feature Name	Description
1	**show_id**	Unique ID for every Movie / Tv Show
2	**type**	Identifier - A Movie or TV Show
3	**title**	Title of the Movie / Tv Show
4	**director**	Director of the Movie / TV Show
5	**cast**	Actors involved
6	**country**	Country of production
7	**date_added**	Date it was added on Netflix
8	**release_year**	Actual Release year of the movie / TV show
9	**rating**	TV Rating of the movie / TV show
10	**duration**	Total Duration in minutes or number of seasons
11	**listed_in**	Geners
12	**description**	The Summary description

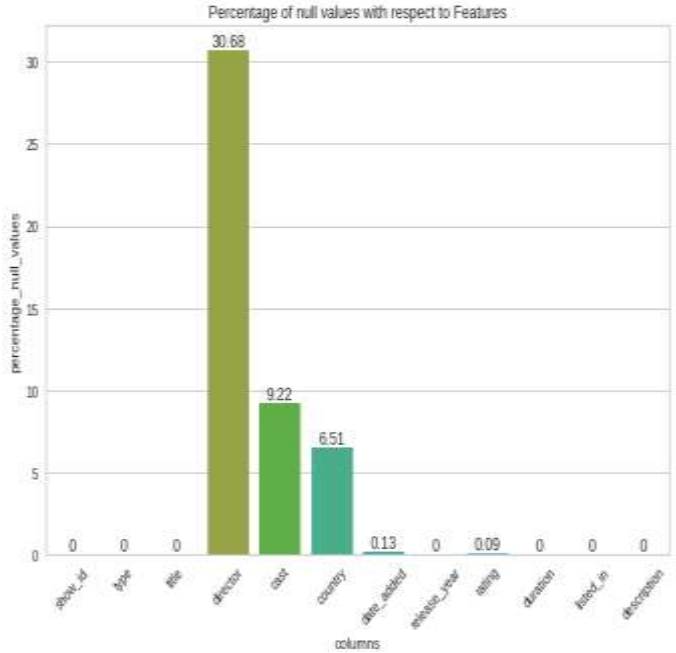
Data Exploration

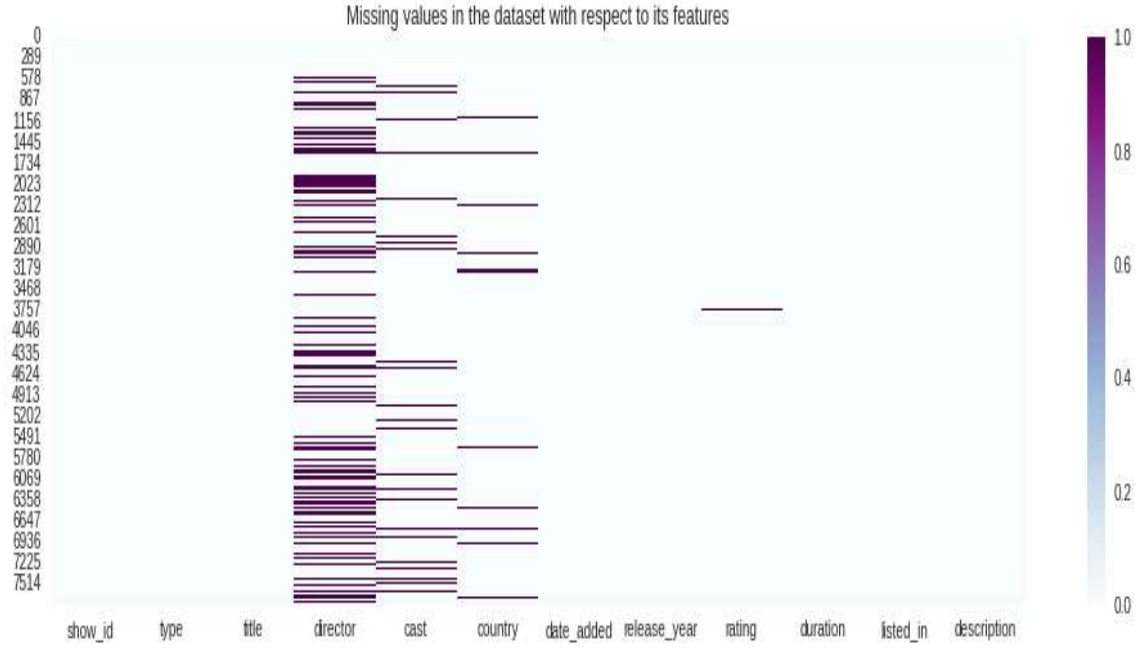
- dataset having 7787 rows and 12 columns
- There are no duplicate values in this dataset.
- Director, cast, country, date_added, and rating all have null values.
- The feature release year is numerical, and everything else is categorical.
- The date_added feature contains dates, but its datatype incorrectly associates an object.



Handling Null Values :

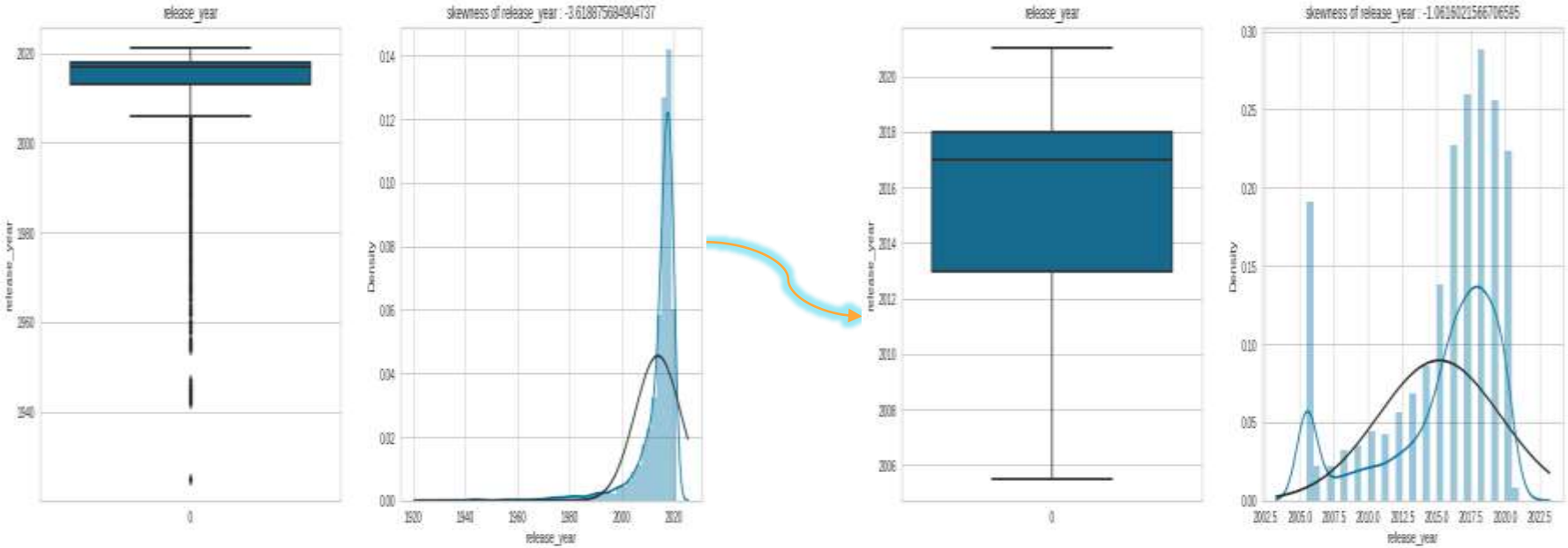
Data Wrangling and Feature Engineering





- Director, cast, country, date_added, and rating have null values in 30.68%, 9.22%, 6.51%, 0.13%, and 0.09% of their respective features
- Since there are many null values for features like director, cast, and country, those null values cannot be dropped; instead, they have been substituted with director Unavailable, Cast Unavailability, and Country Unavailable, accordingly.
- Features such as date_added and rating have a very low number of null values, so we dropped those null values.

Handling Outliers



- Outliers from variable release_year are successfully treated using the interquartile range.

Feature Engineering

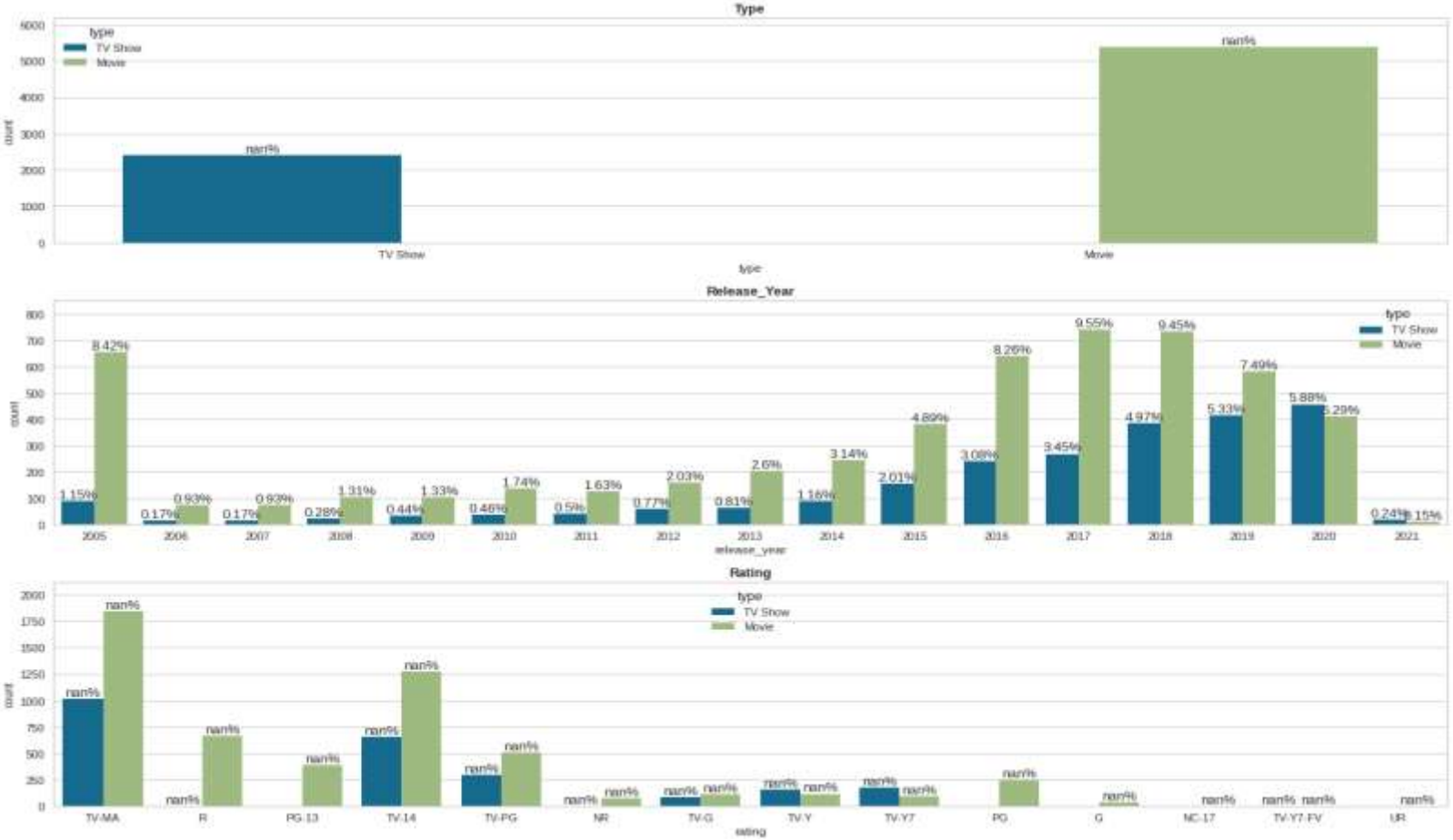
- **Converted feature date_added to datetime and created new features from it, such as year_added, month_added, and day_added, before removing the feature date_added.**

- **The listed_in feature has been renamed to geners.**

- **Because year cannot be a float, the feature release_year data type changed from float64 to int64.**

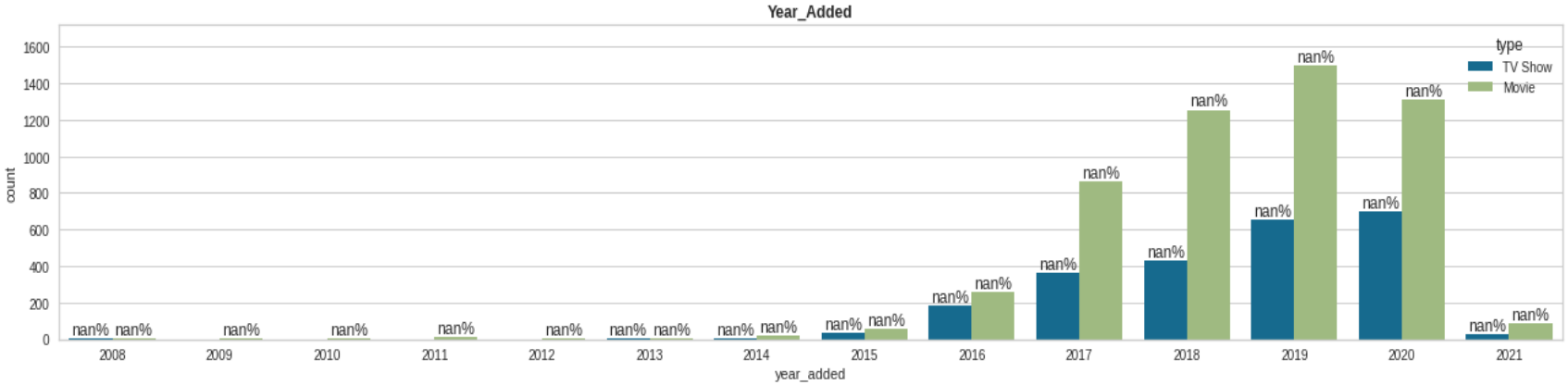
Exploratory Data Analysis

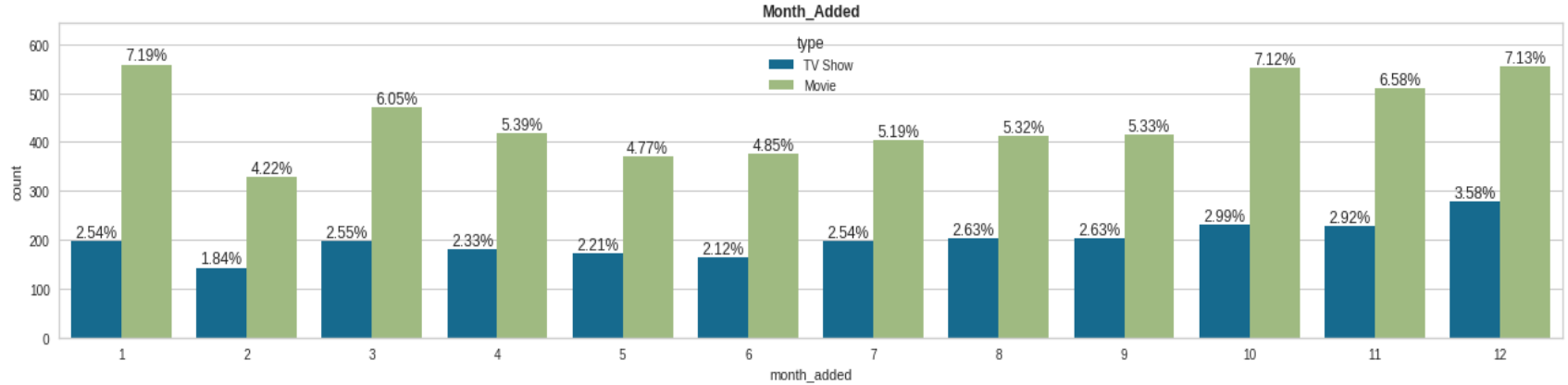
Univariate Analysis



Exploratory Data Analysis

Univariate Analysis



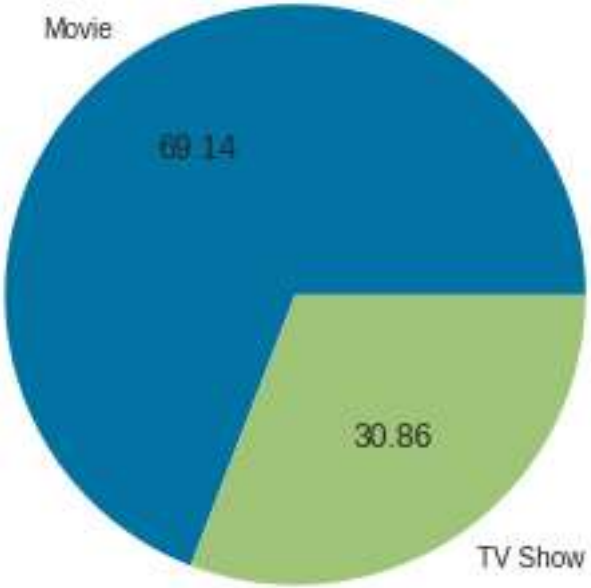


Univariate Analysis

Observations :

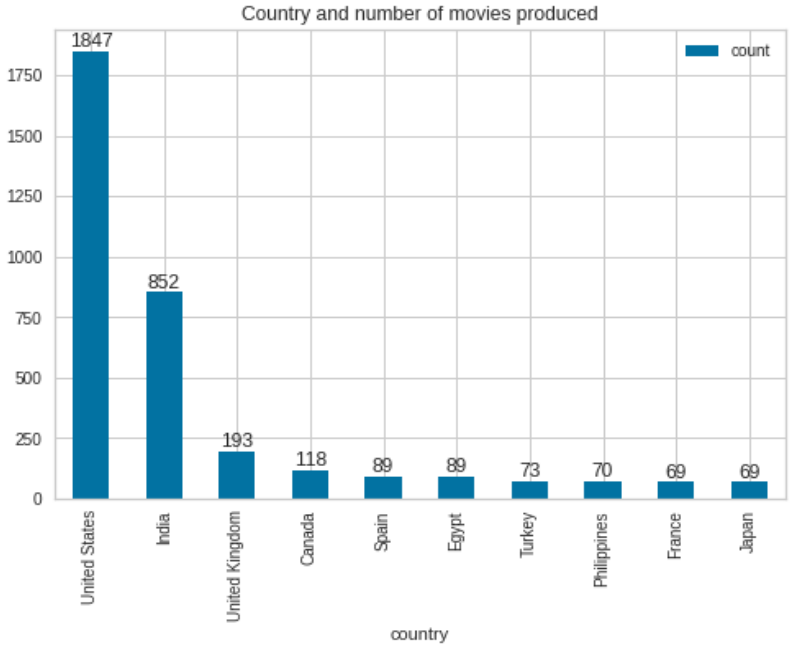
- More movies (69.14%) than TV shows (30.86%) are available on Netflix.
- The majority of Netflix movies were released between 2015 and 2020, and the majority of Netflix TV shows were released between 2018 and 2020.
- The most movies and TV shows were released for public viewing on Netflix in 2017 and 2020, respectively, out of all released years.
- From 2006 to 2019 Netflix is constantly releasing more new movies than TV shows, but in 2020, it released more TV shows than new movies, indicating that Netflix has been increasingly focusing on TV rather than movies in recent years.
- More TV shows will be released for public viewing in 2020 and 2021 than at any other time in the history of Netflix.
- The majority of TV shows and movies available on Netflix have a TV-MA rating, with a TV-14 rating coming in second.
- The majority of movies were added to Netflix in 2019 and the majority of TV shows were added to Netflix in 2020.
- In 2019, Netflix added nearly one-fourth (27.71%) of all content (TV shows and movies).
- The majority of the content added to Netflix was in October and January, respectively, but almost all months throughout the year saw Netflix adding content to its platform.

Bivariate Analysis



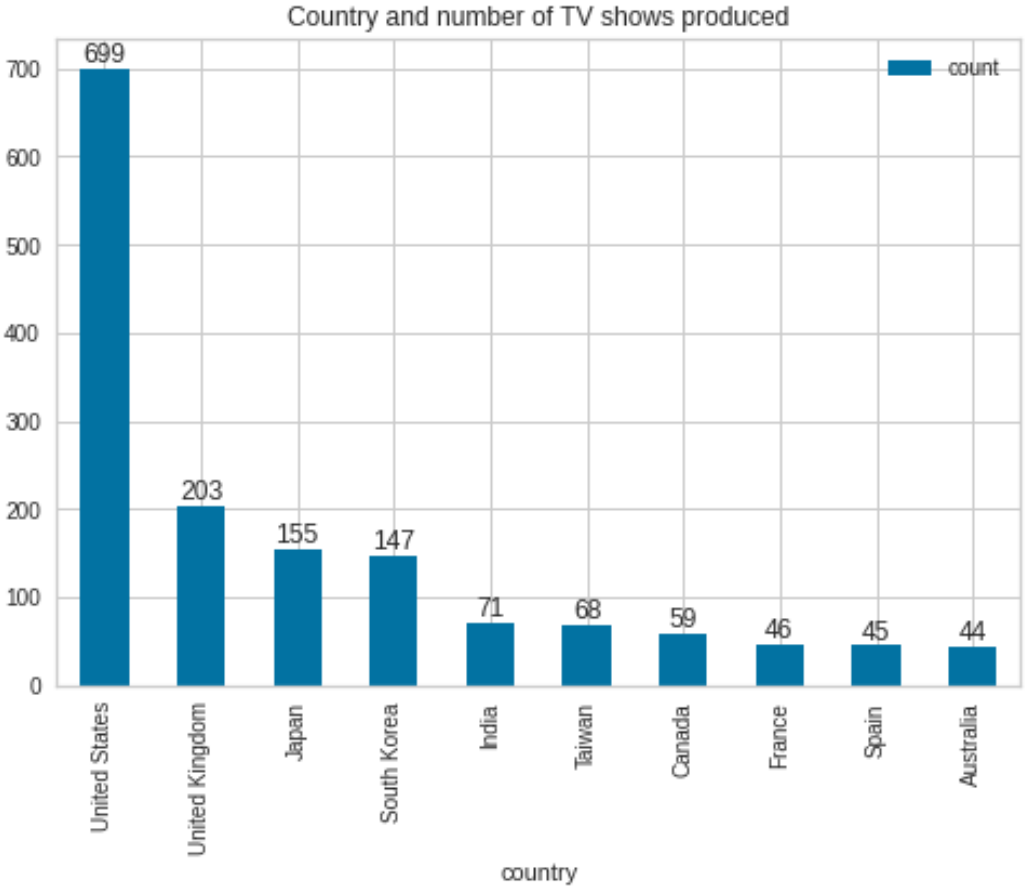
- **Netflix has more movies (69.14%) than TV shows (30.86%).**

Bivariate Analysis



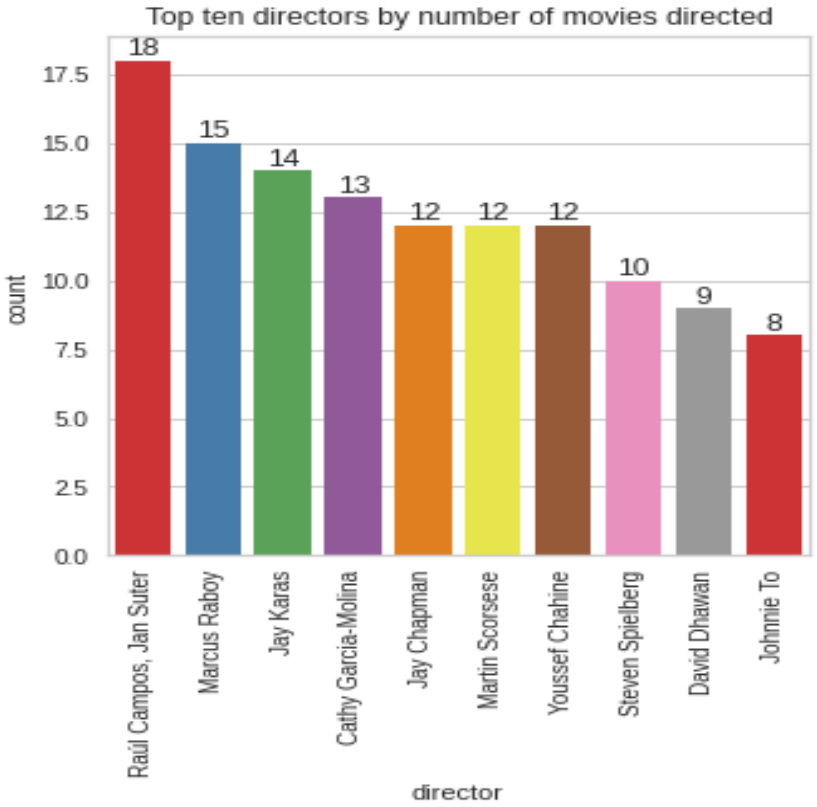
- The majority of movies available on Netflix are produced in the United States, with India coming in second

Bivariate Analysis



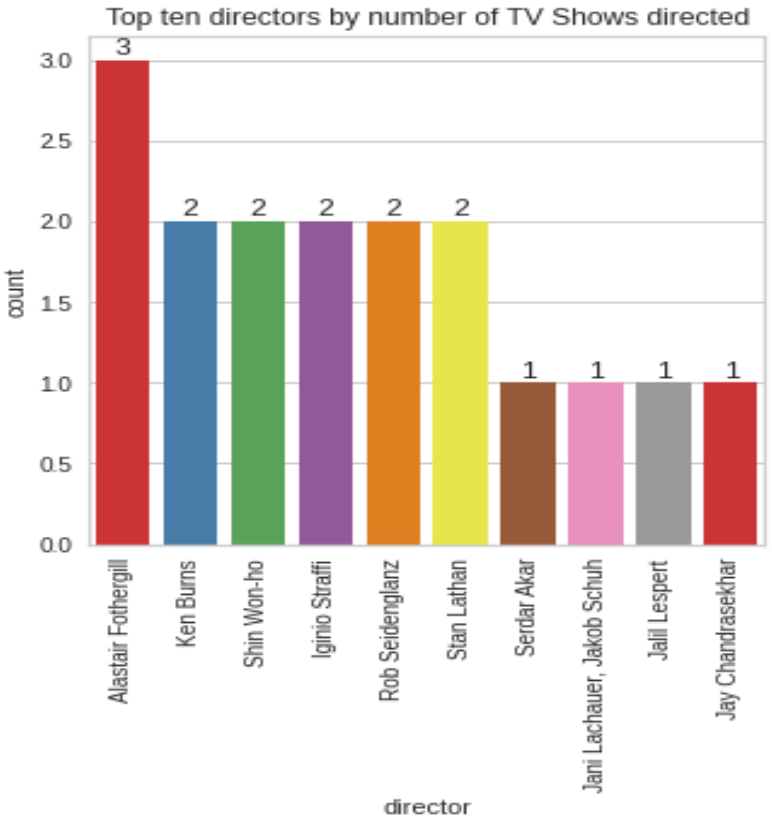
- The United States and the United Kingdom are the two countries that produced most of the TV shows that are available on Netflix.

Bivariate Analysis



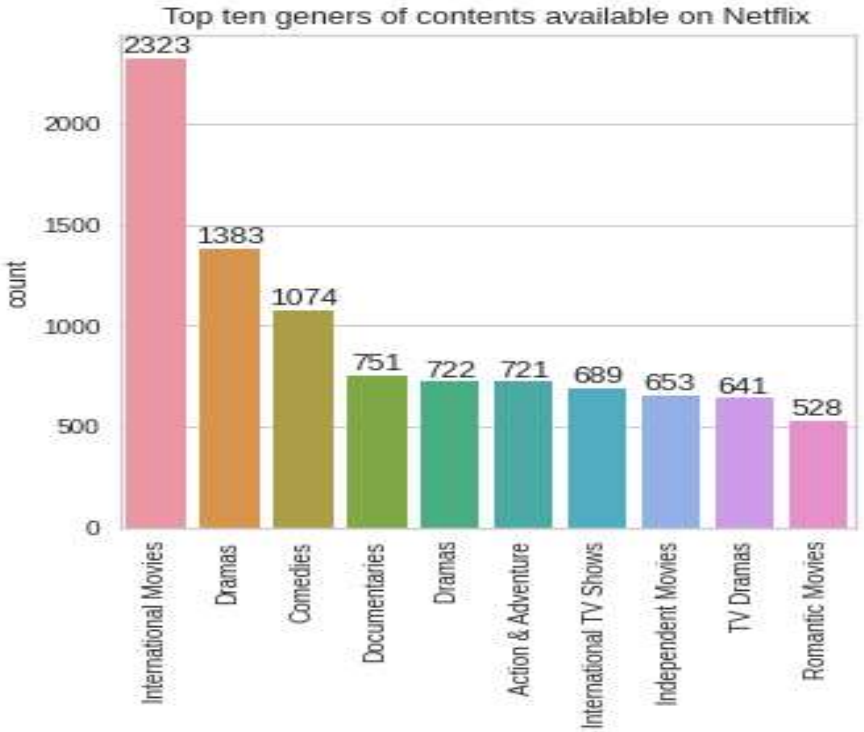
- Raul Campos and Jan Suter directed most of the movies available on Netflix for public viewing.

Bivariate Analysis



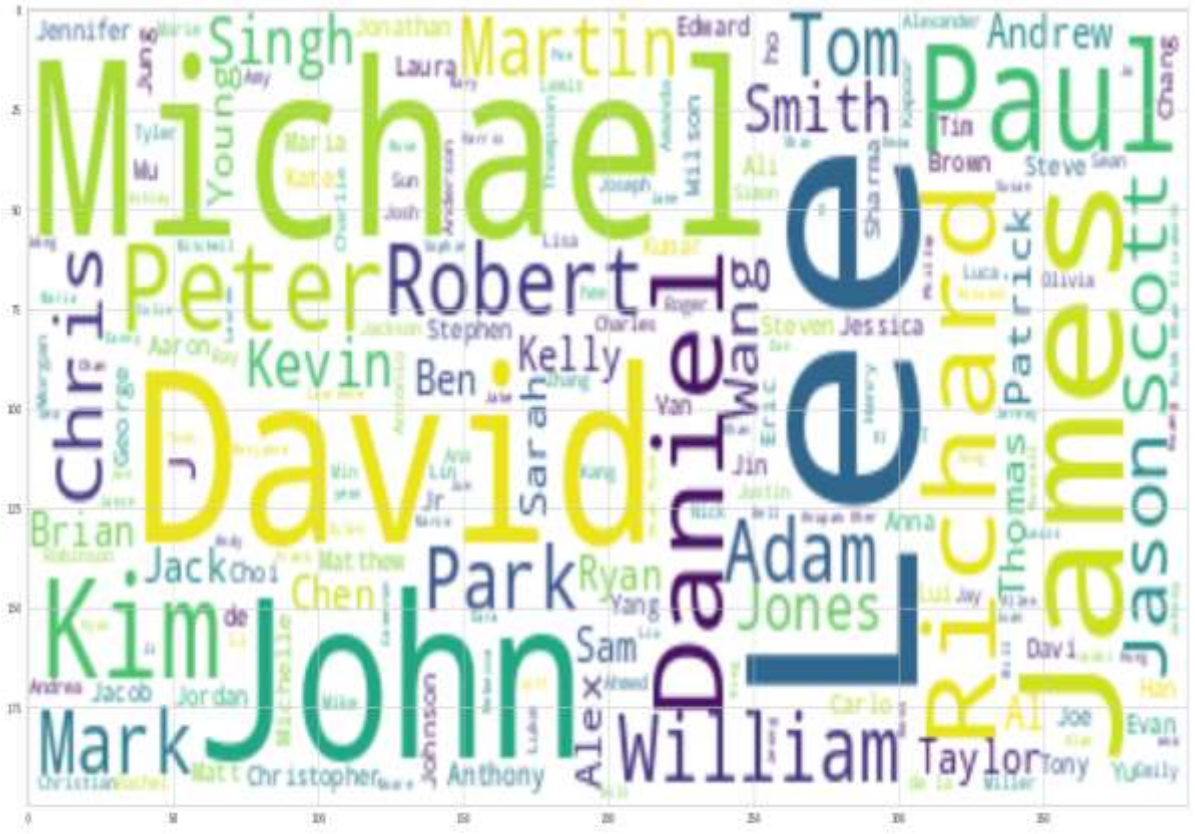
- Alastair Fothergill directed most of the TV shows available on Netflix for public viewing.

Bivariate Analysis



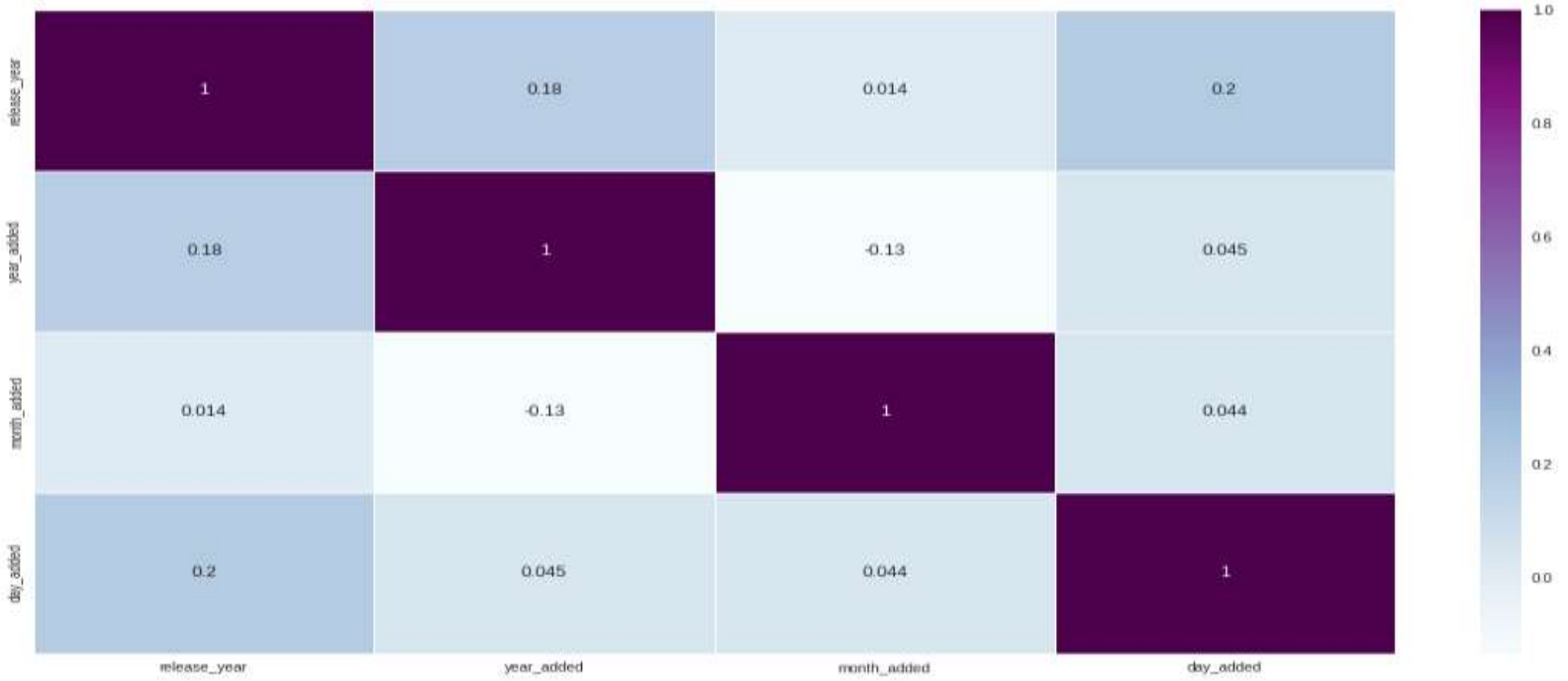
- International movies and the second-most popular dramas are available on Netflix as content.

Bivariate Analysis



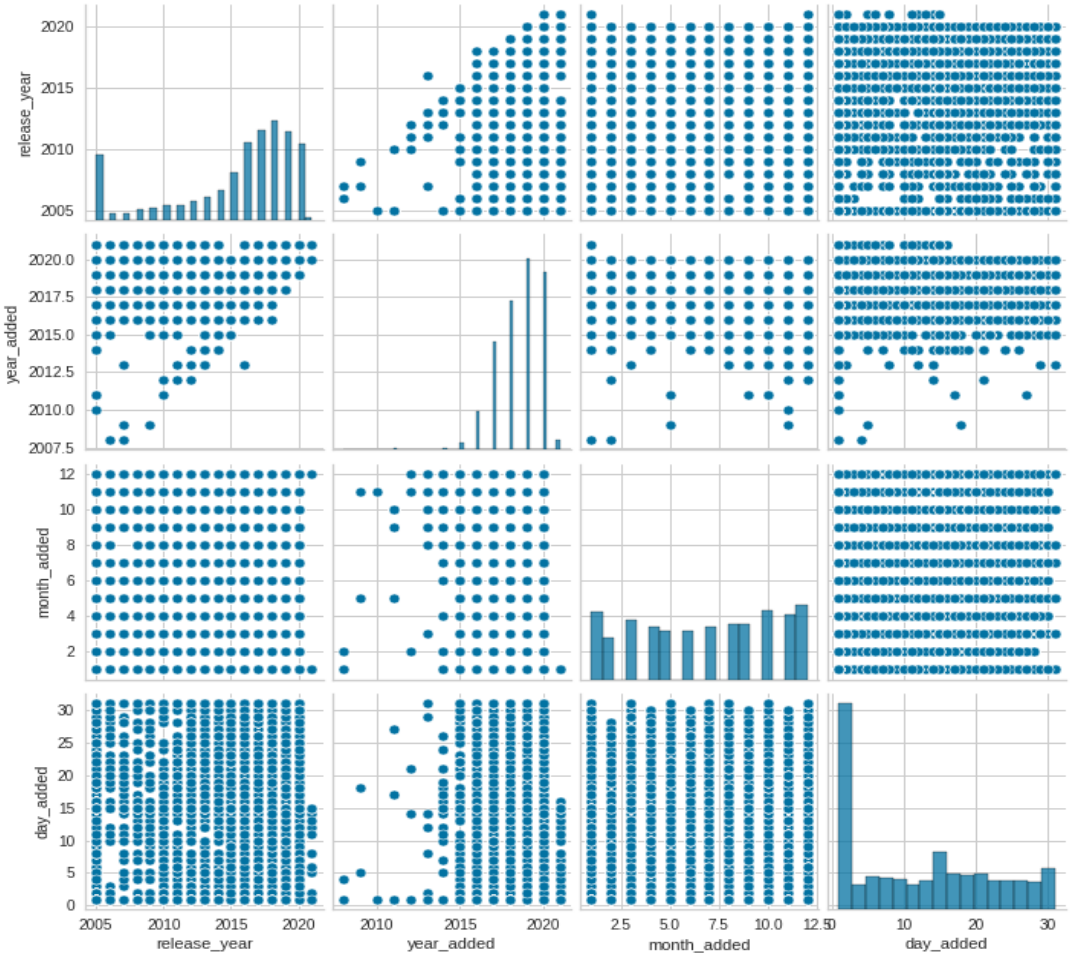
- Actors who have appeared in films and TV shows that are most available on Netflix are Lee, Michel, David, Jhon, and James.

Multivariate Analysis



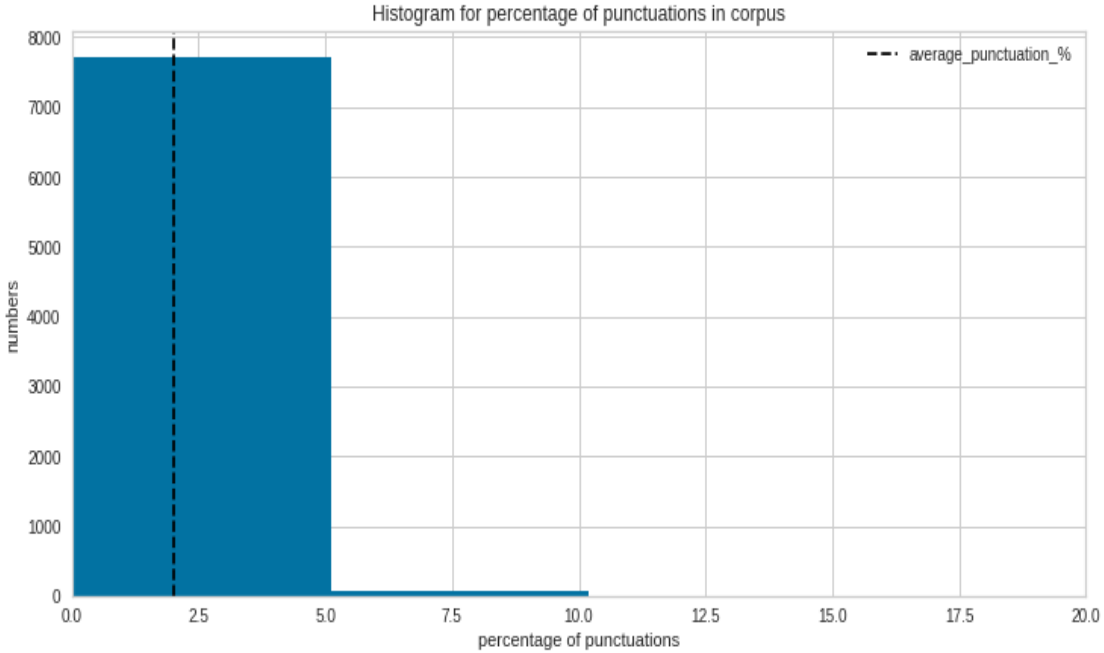
- We see that the movie or TV show release year and day of the month on movies or TV shows added to Netflix are slightly correlated with each other.

Multivariate_Analysis



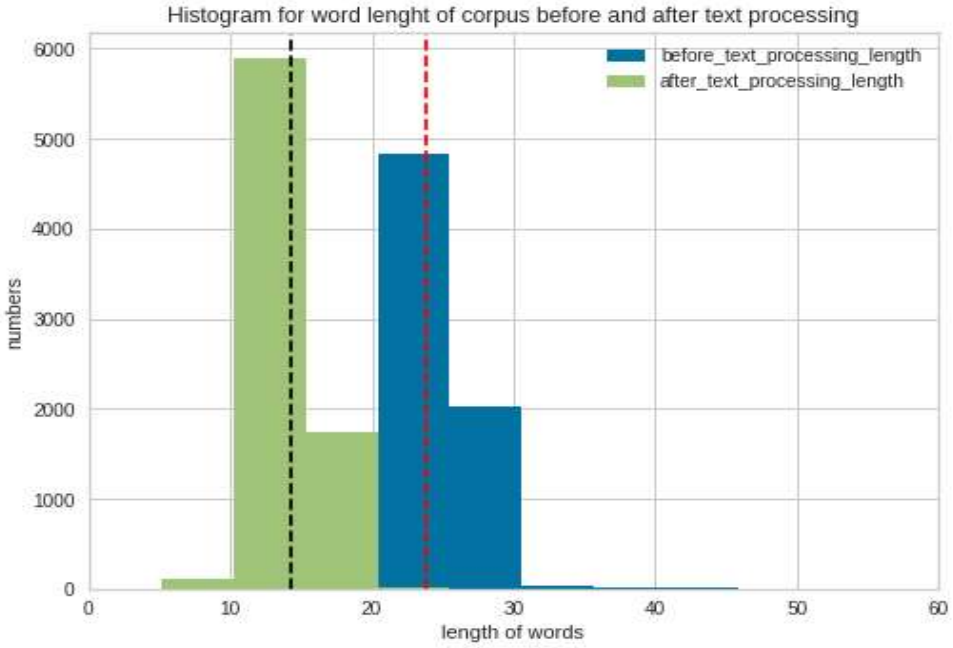
- Based on the plot of release_year and year_added, we can conclude that Netflix is increasingly adding and releasing movies and TV shows over time.
- We can conclude from plot release_year and month_added that Netflix releases movies and TV shows throughout all months of the year.

Data Pre-processing



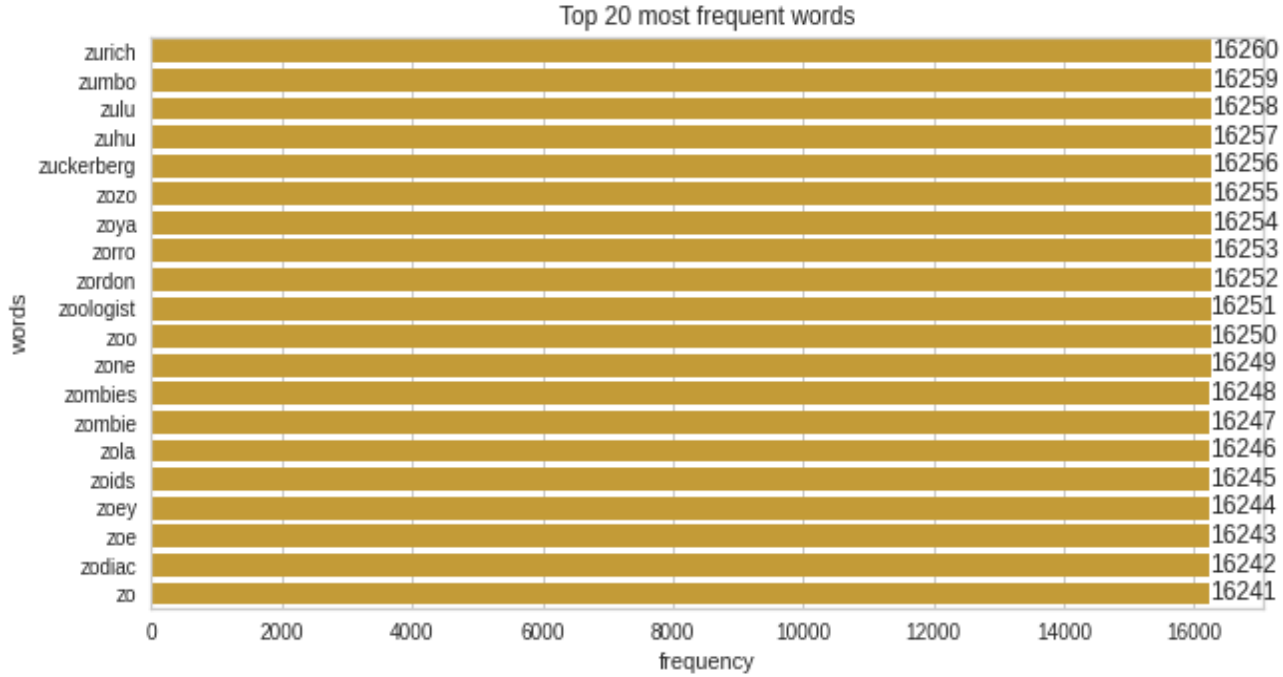
- The majority of the corpus contains punctuation that accounts for less than 5% of the total corpus.

Data Pre-processing



- After text processing, each corpus has, on average, 14 words, but before text processing, each corpus contains, on average, 24 words.

Data Pre-processing



- Zurich, Zumbo, Zulu, uhu, and Zuckerberg are the top 5 most frequent words in the corpus.

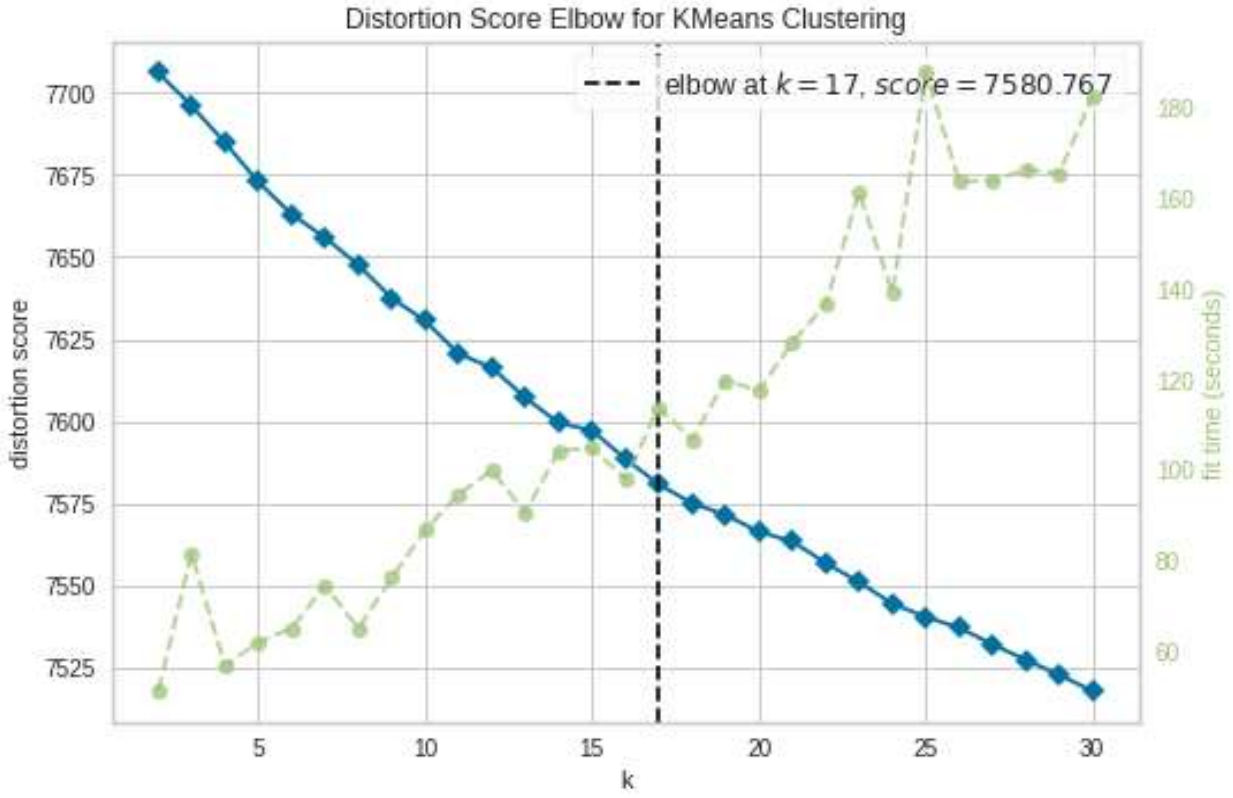
Machine Learning Model Implementation

[1] K-Means Clustering

[2] Hierarchical Clustering

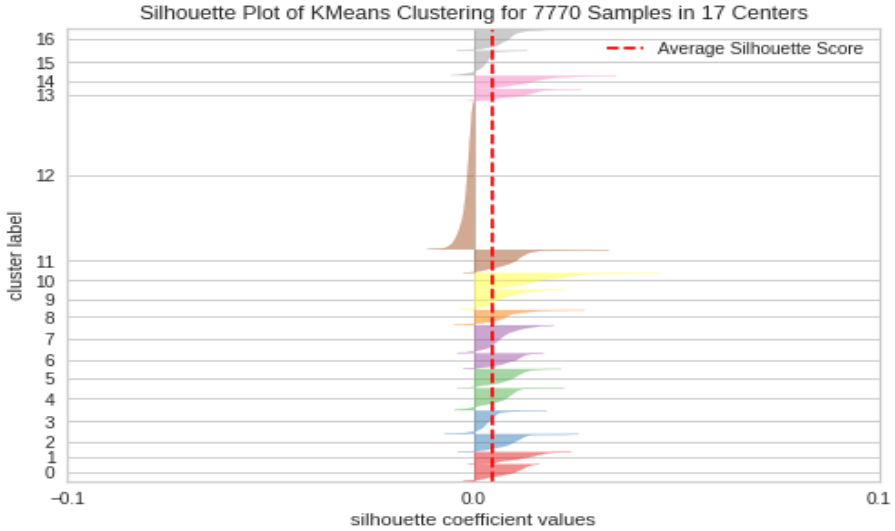
[3] DBSCAN Clustering

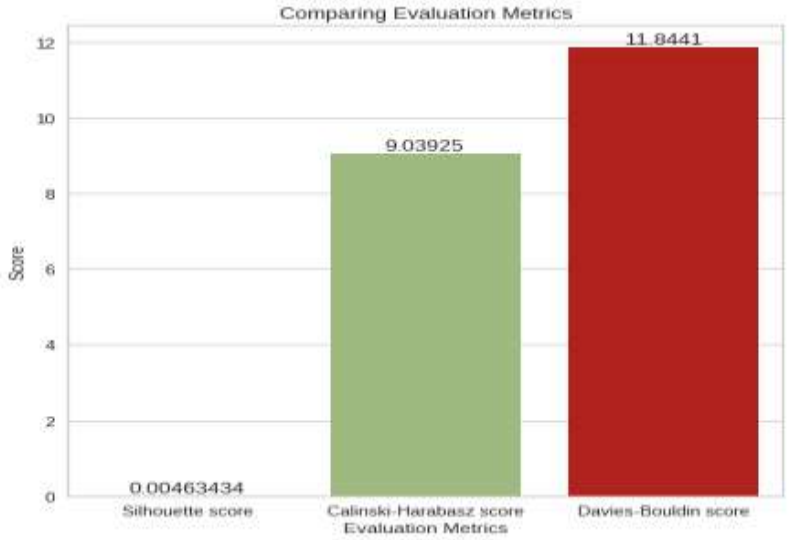
K-Means Clustering



- Optimal number of cluster using Elbow method : 17

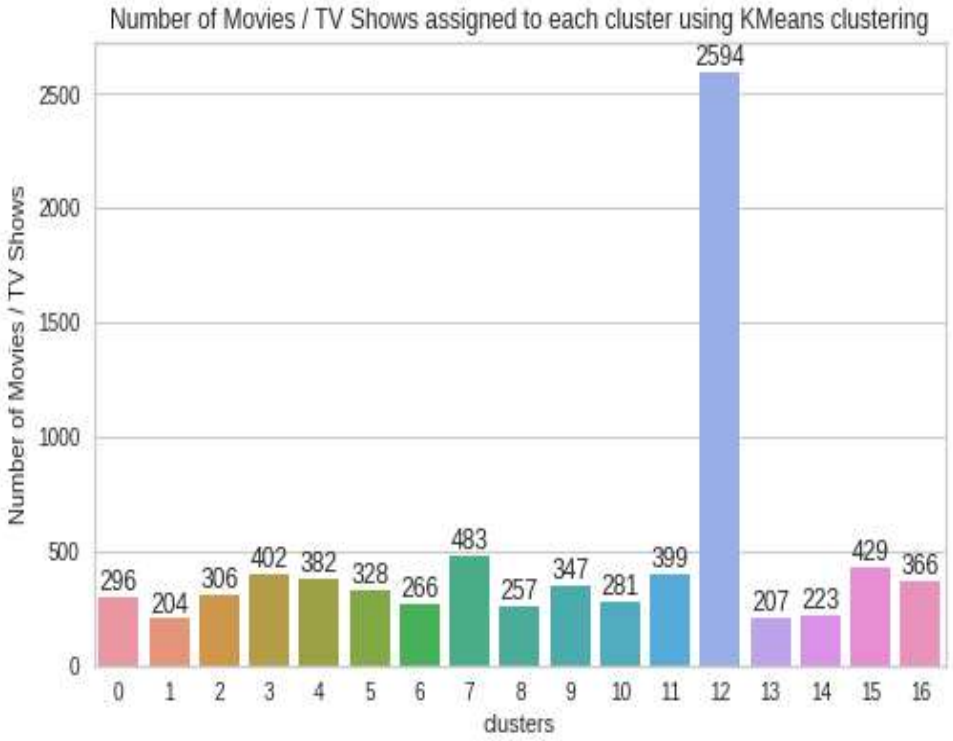
Comparing Evaluation Metrics for K-Means Clustering





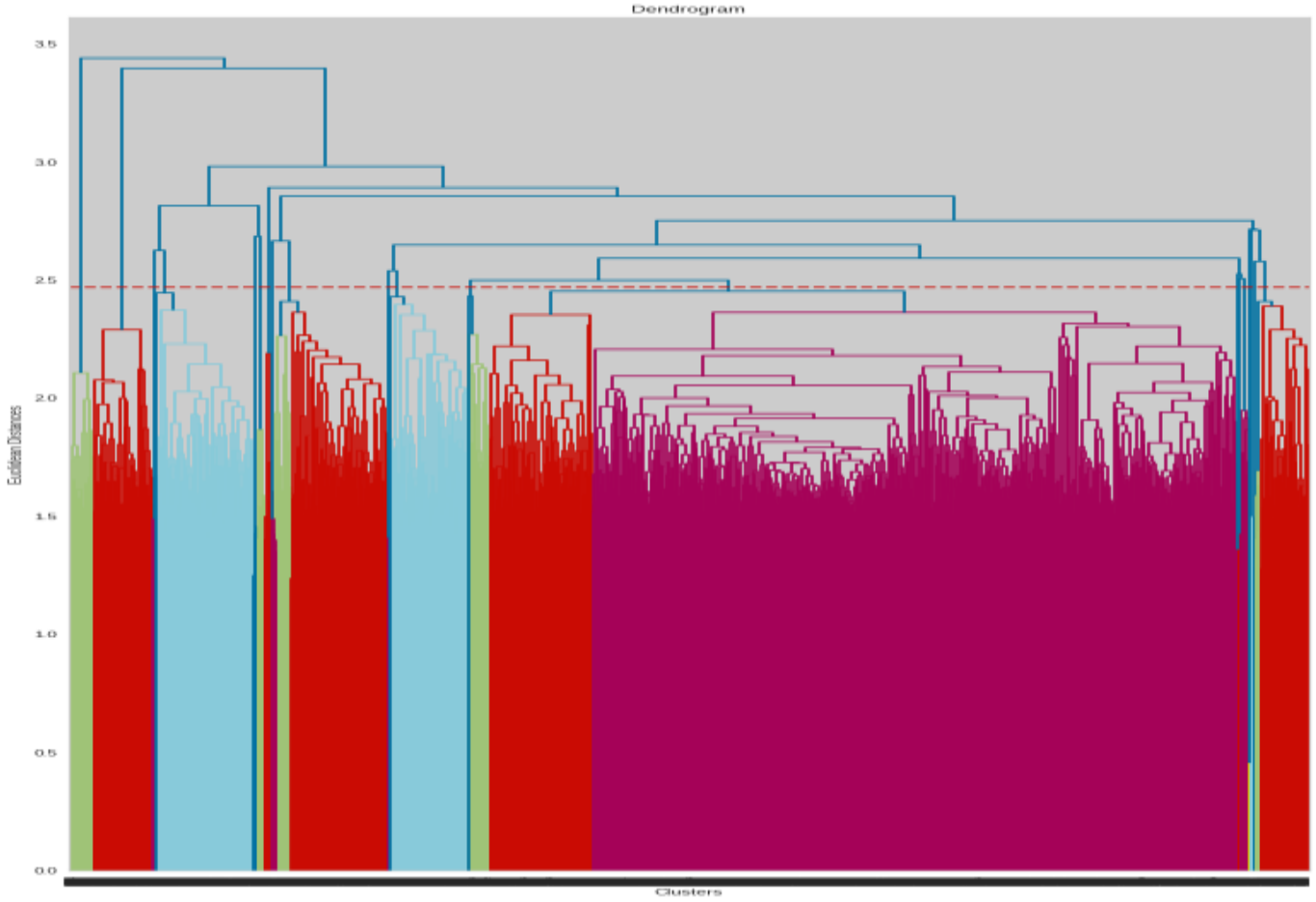
- We got a Silhouette score of 0.00463434, a Calinski-Harabasz score of 9.03925, and a Davies-Bouldin score of 11.8441 after evaluation of the model.

K-Means Clustering



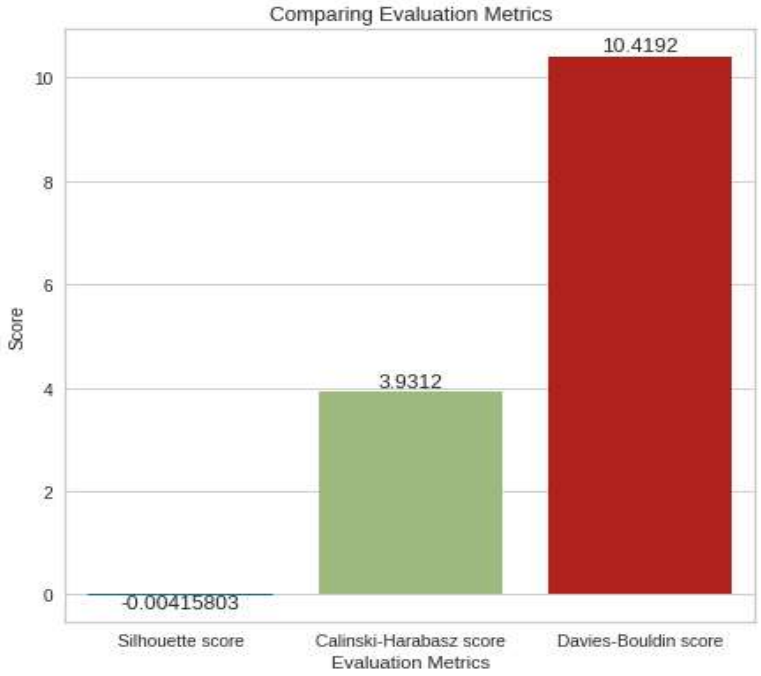
- Cluster 12 has the most number of movies and TV shows, followed by clusters 7 and 15.

Hierarchical Clustering



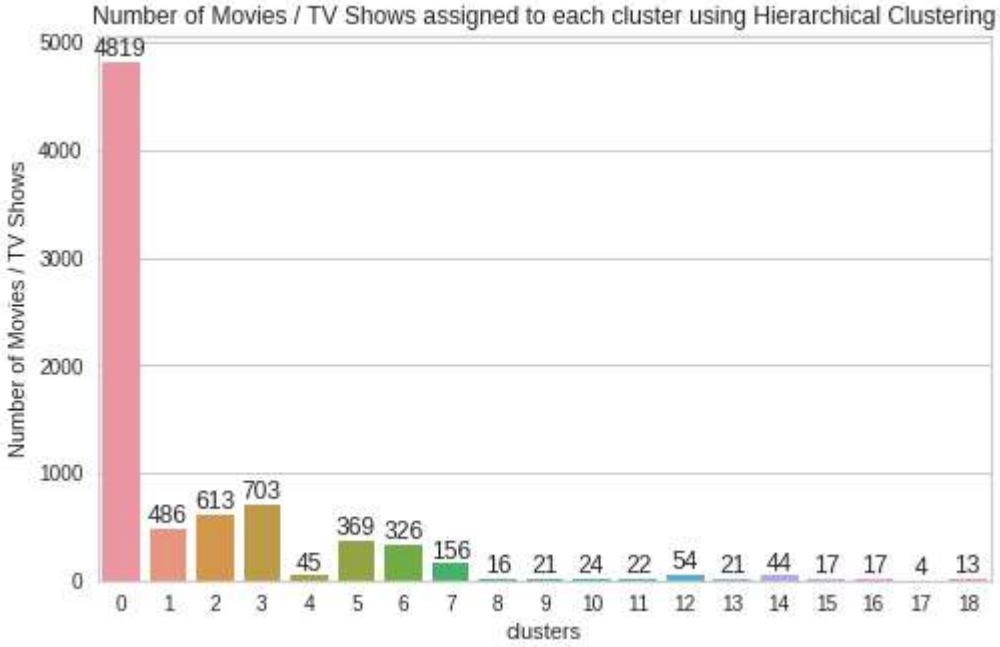
- After cutting the horizontally tallest vertical line, 19 vertical lines are intersected, and we get the optimal number of clusters: 19

Comparing Evaluation Metrics for Hierarchical Clustering



- We got a Silhouette score of -0.00415803, a Calinski-Harabasz score of 3.9312, and a Davies-Bouldin score of 10.4192 after evaluation of the model.

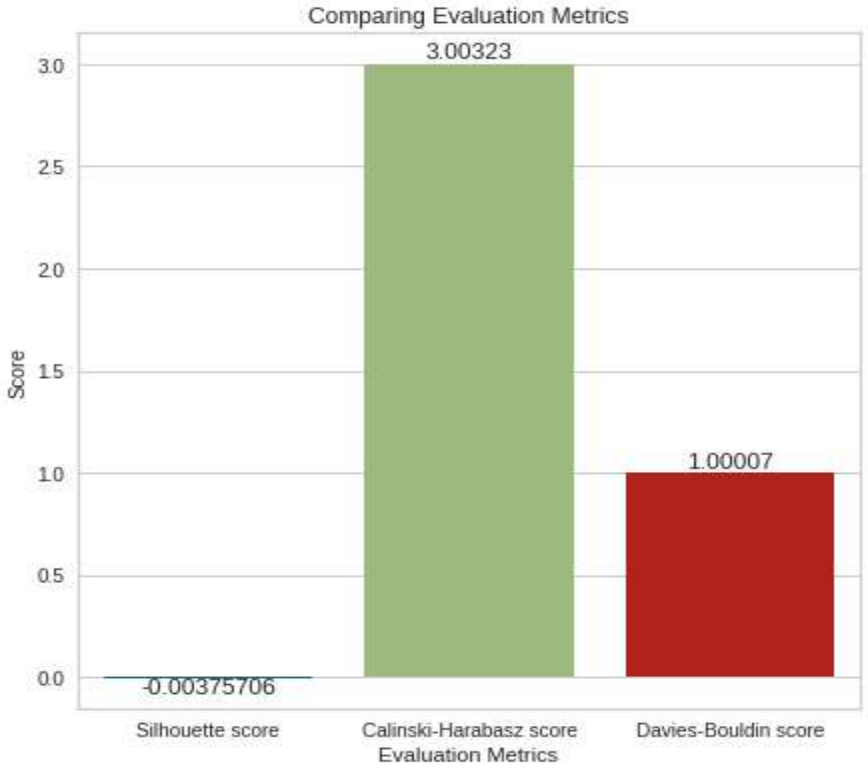
Hierarchical Clustering



- Cluster 0 has the most number of movies and TV shows, followed by clusters 3 and 2.

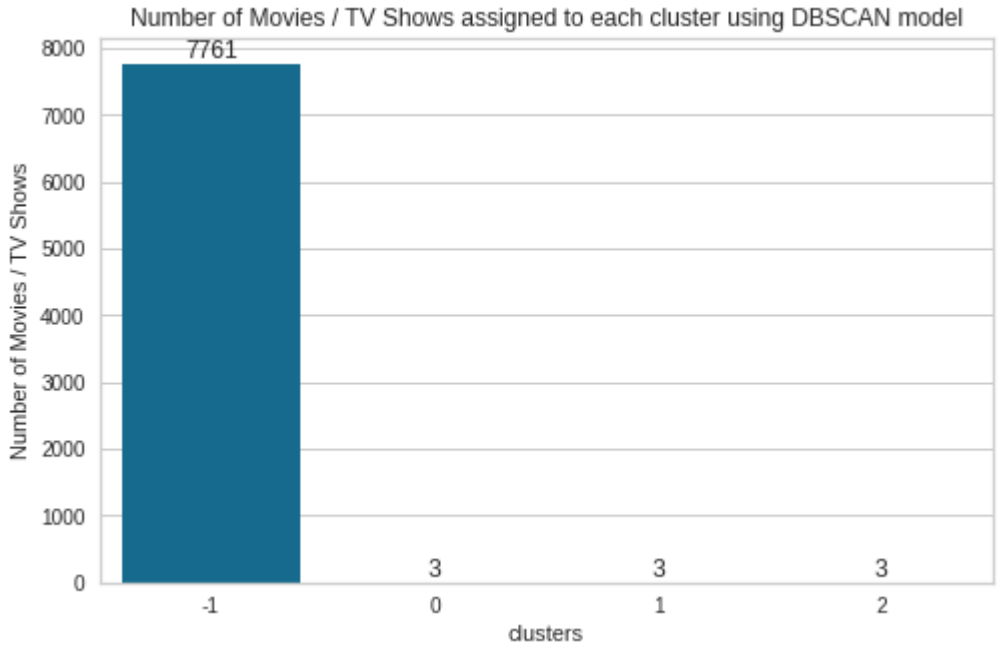
DBSCAN Clustering

Comparing Evaluation Metrics for DBSCAN Clustering



- We got a Silhouette score of -0.00375706, a Calinski-Harabasz score of 3.00323, and a Davies-Bouldin score of 1.00007 after evaluation of the model

DBSCAN Clustering

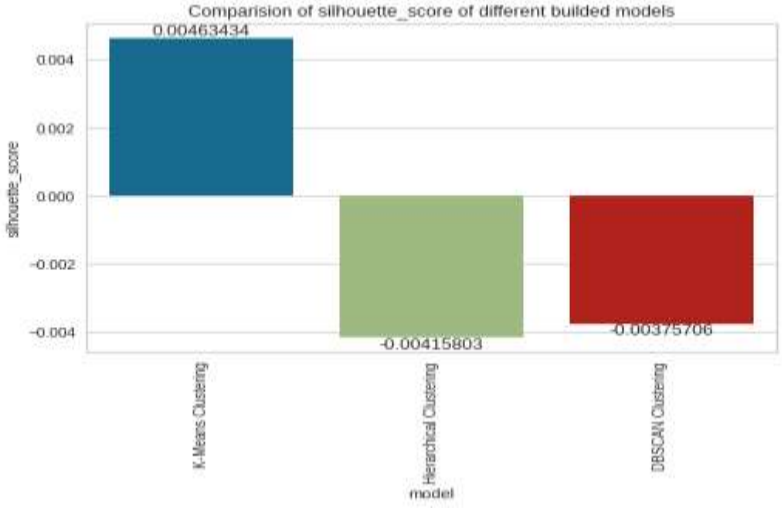


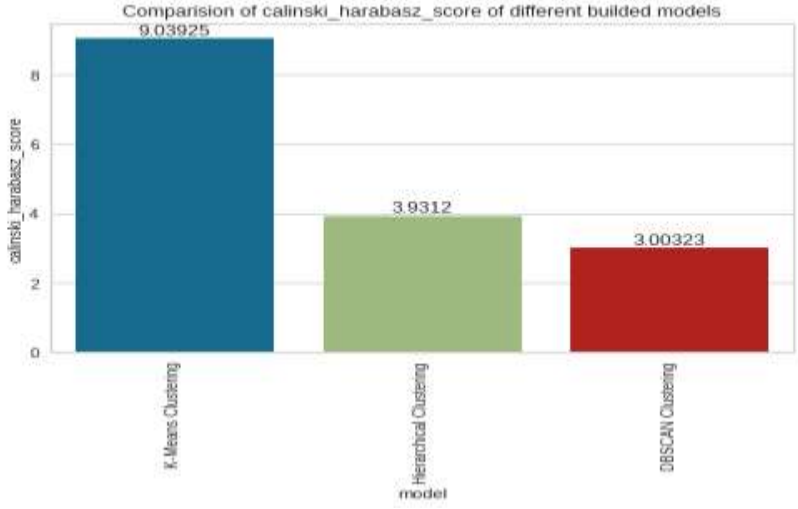
- Cluster -1 has assigned the most number of movies and TV shows.

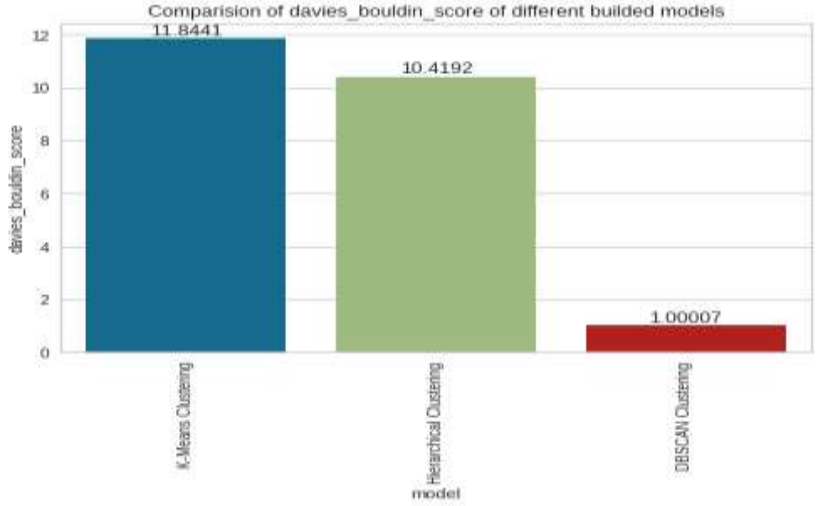
Model Evaluation

ML Model / Metrics	K-Means Clustering	Hierarchical Clustering	DBSCAN Clustering
Silhouette score	0.004634	- 0.004158	- 0.003757
Calinski - Harabasz score	9.039247	3.931202	3.003227
Davies bouldin score	11.844054	10.419206	1.000067

Comparing Evaluation Metrics







- Among all models, the K-Means Clustering model has the highest Calinski-Harabasz score (9.039247). Also, K-Means Clustering model has a silhouette_score of 0.004634, which is close to 1 than other models, which means the K-Means Clustering model is capable of perfectly clustering items.

Conclusion

- ✓ **The K-Means Clustering model has the highest Calinski-Harabasz score out of all the models (9.039247). Also, the silhouette score for the K-Means Clustering model is 0.004634, which is close to one compared to other models, indicating that it can cluster items perfectly.**

- ✓ **The K-Means Clustering model is the optimal model and well-trained for clustering TV shows and movies based on the content due to its high Calinski-Harabasz score (9.039247) and silhouette score (0.004634), which are close to 1 than other builded models.**

Challenges

AI

- **Data Preprocessing:**

Data preprocessing is an essential step in any machine learning project, and we faced difficulties identifying and fixing errors in the data.

- **Feature Engineering:**

we faced difficulties choosing the right features, but it was difficult to determine which ones were most important.

- **Algorithm Selection:**

Choosing the right algorithm is critical to the success of the model. It was difficult to determine which algorithm would be most effective for a particular problem.

- **Model Training:**

Training the model requires a lot of resources and can take a long time. It was important to choose the right parameters for the model in order to achieve the best performance.

- **Model Evaluation:**

Evaluating the performance of the model is essential to determining how well it is performing. It was important to choose the right metrics for evaluating the model and to interpret the results correctly.

Thank You !