Capstone Project

Netflix Movies & TV Shows

Kiran Mamtani

Content

- Defining problem statement
- Exploratory Data Analysis and Feature Engineering
- Unsupervised Machine Learning Models
- Conclusion
- Challenges

Source: https://www.freepik.com/

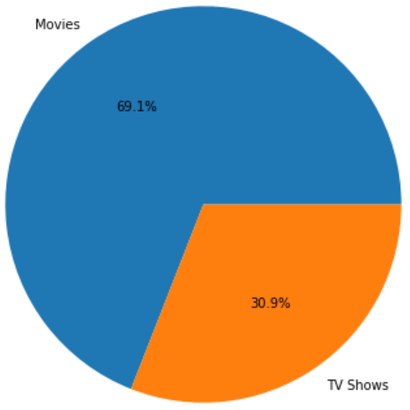
Defining Problem Statement

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. We have 7787 rows and 12 columns.

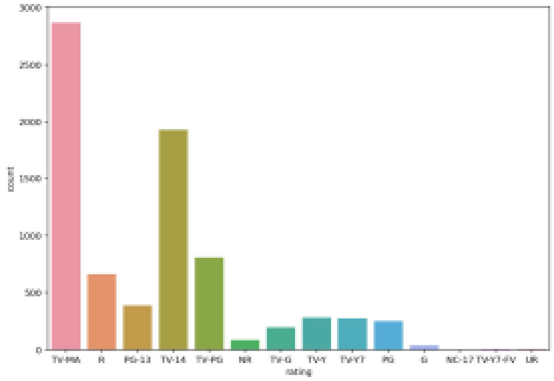
In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service’s number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore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

In this project we have to understand what type content is available, is Netflix has increasingly focusing on TV rather than movies in recent years, clustering similar content by matching text-based features; etc.

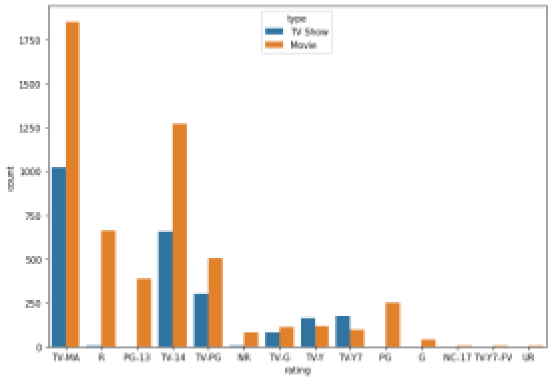
EDA & Feature Engineering



The data has more number of movies compared to TV shows. 69% - Movies and 31% - TV Shows

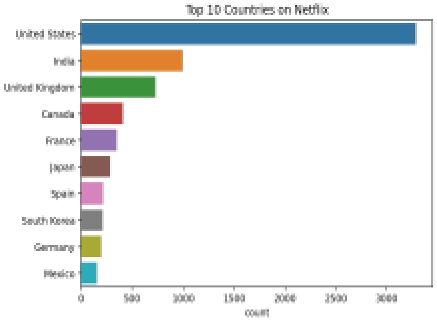


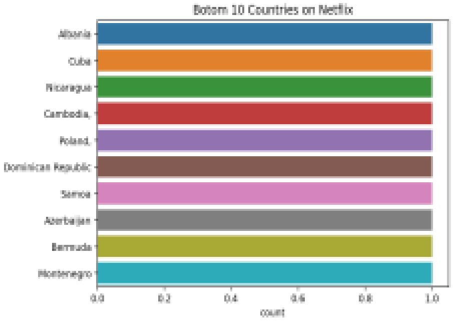
Our data has TV-MA (TV Mature Audience Only) rating as a highest numbers followed by TV-14 (unsuitable for children under 14 years of age) , TV-PG (Parental Guidance)

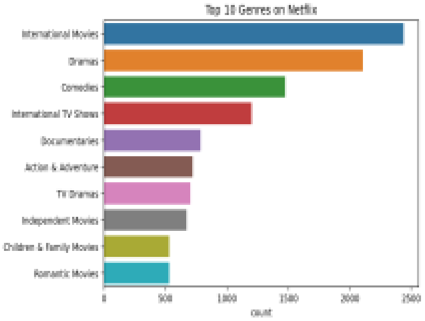


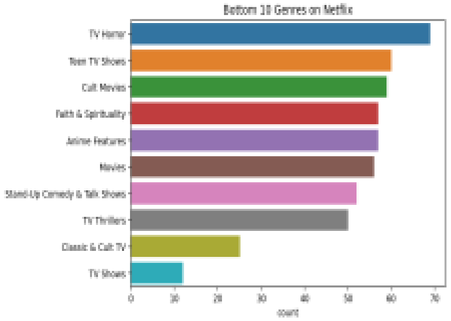
In TV shows, the top three ratings are TV-MA, TV-14 and TV-PG (which is same as overall data) In Movies, the top three ratings are TV-MA, TV-14 and R

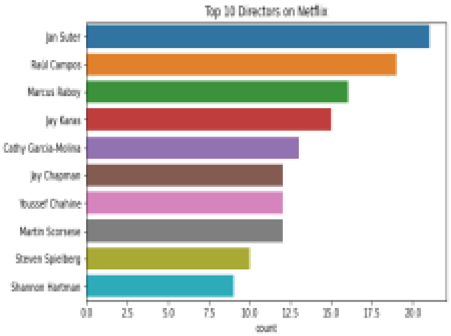
EDA & Feature Engineering (continued)...

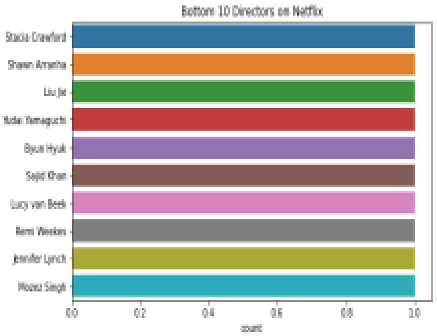


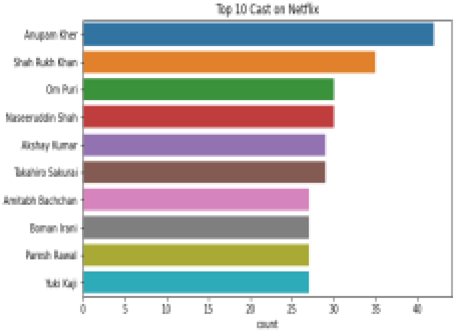


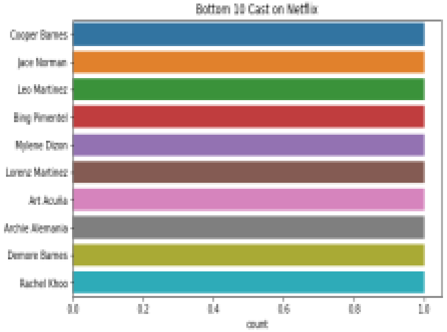




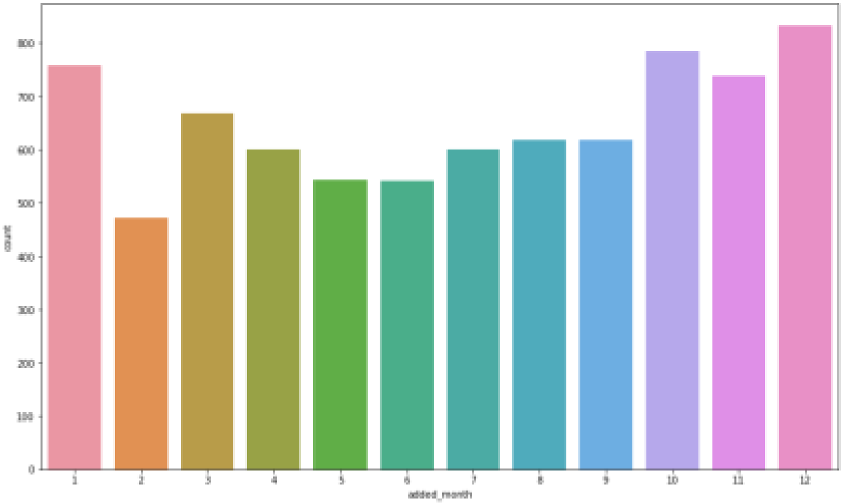


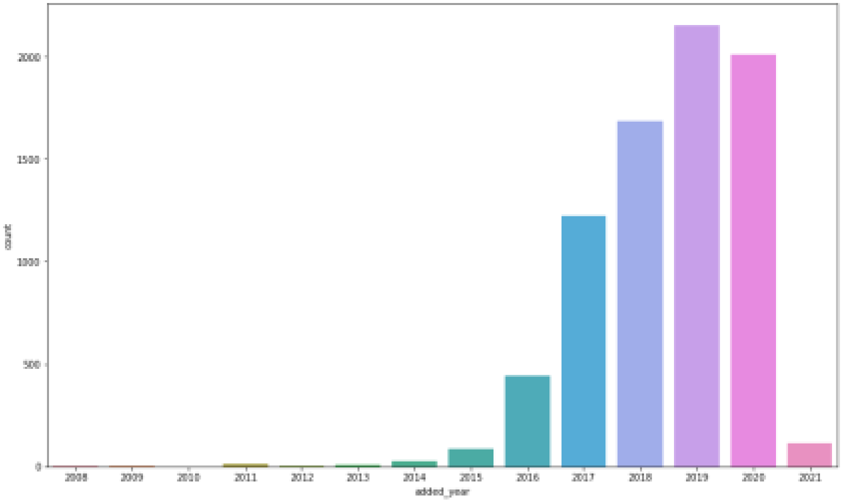






EDA & Feature Engineering (continued)...

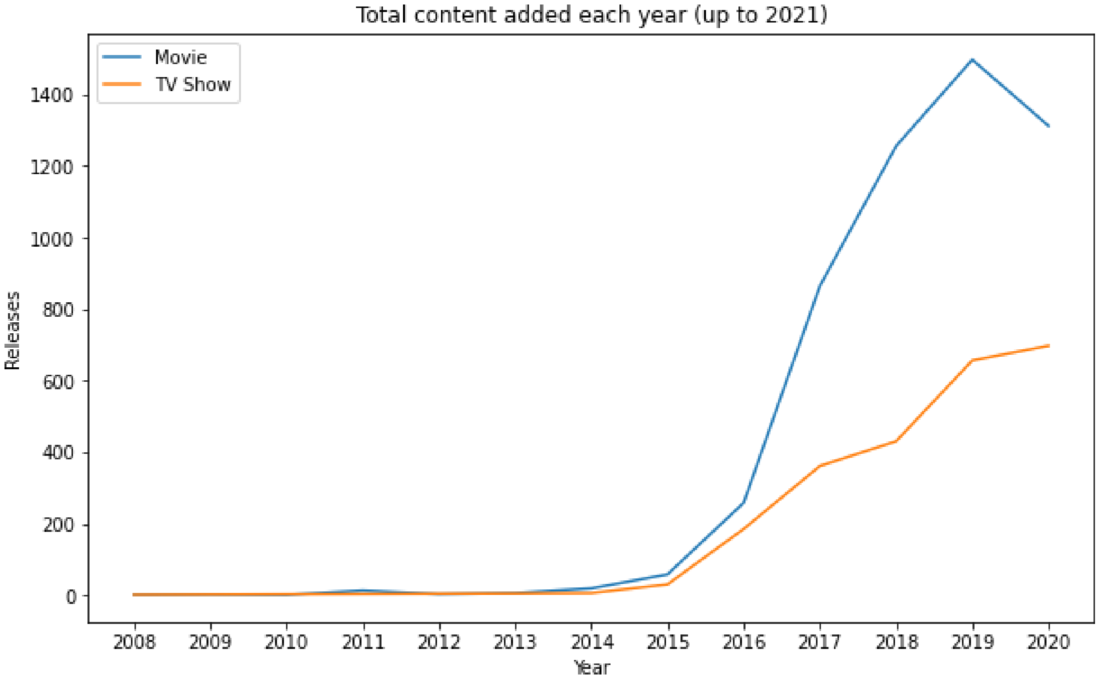




So the highest video content was added in month of December followed by September and January. Here we can simply assume that these are those months where holidays are more that too in almost every countries...

In year 2019, maximum video content was added on Netflix followed by year 2020 and year 2018. And as seen in graph is has gradually increase from 2015. So we can say that Netflix got famous in world after 2014.

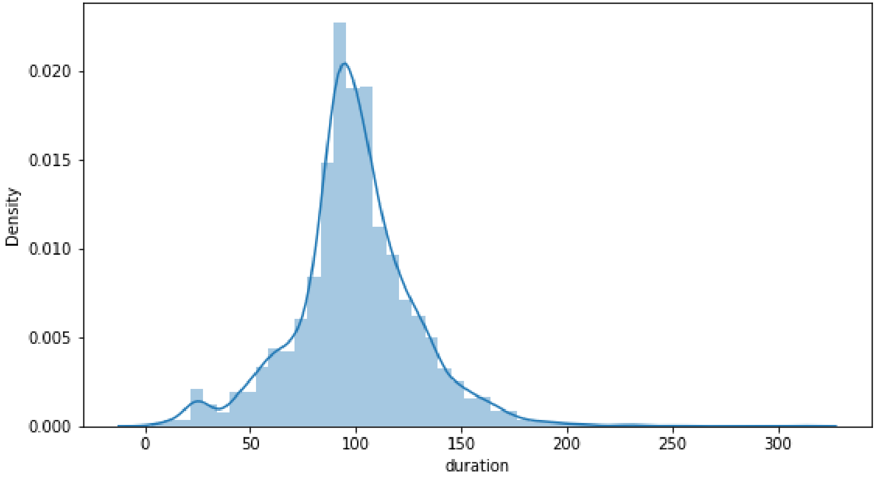
EDA & Feature Engineering (continued)...

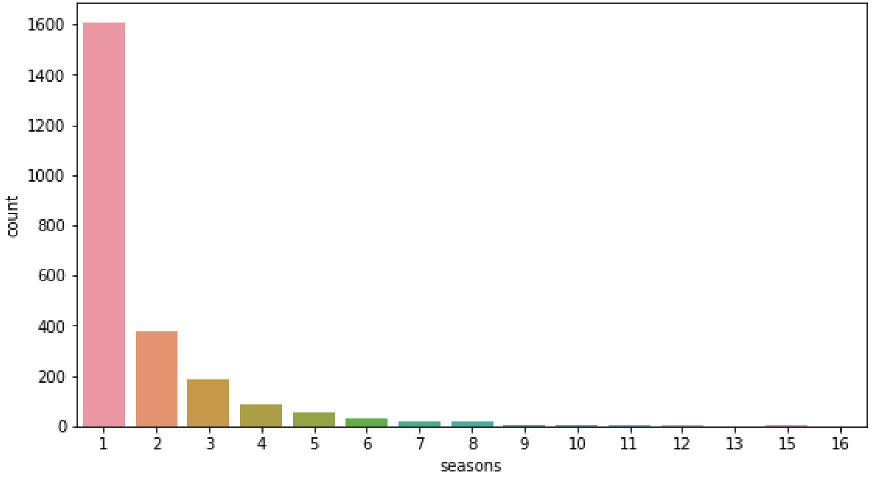


Based on the timeline, we can see that the popular streaming platform started gaining traction after 2014. Since then, the amount of content added has been tremendous. We decided to exclude content added during 2021 since the data does not include a full years worth of data. We can see that there has been a consistent growth in the number of movies on Netflix compared to shows.

EDA & Feature Engineering (continued)...

Now we will look into the duration of Netflix films. Since movies are measured in time and shows are measured by seasons, we had split the dataset between movies and TV shows



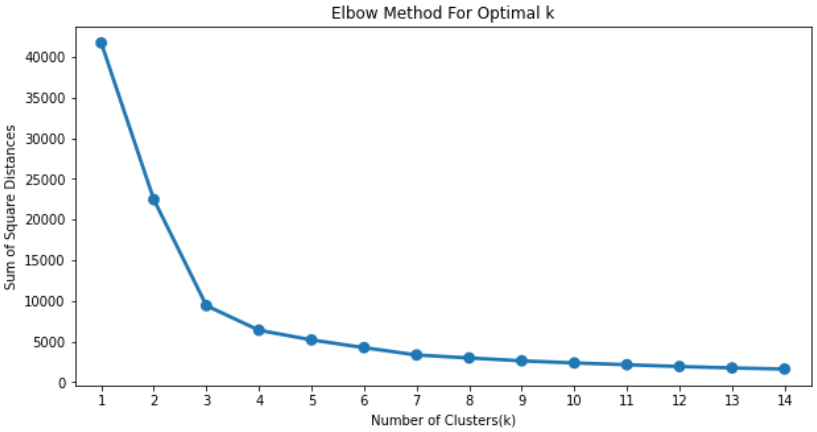


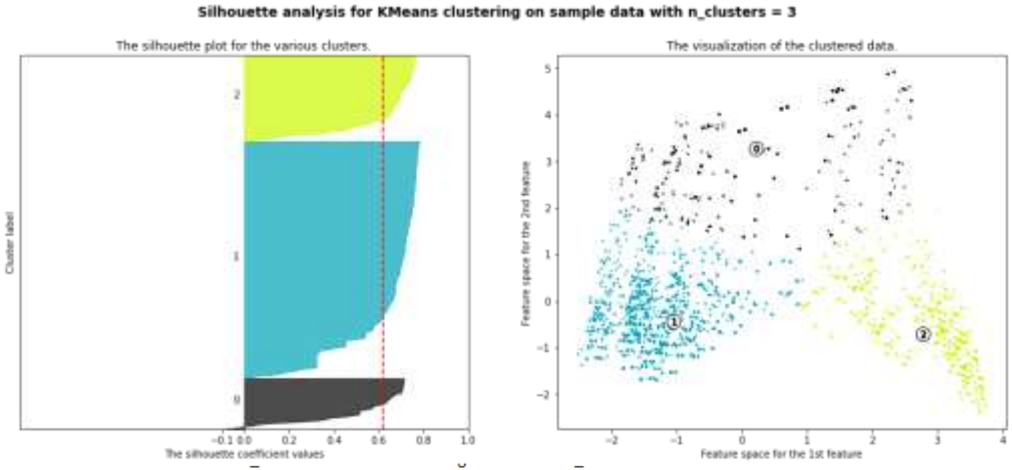
Above on the left, we can see that the duration for Netflix movies closely resembles a normal distribution with the average viewing time spanning about 90 minutes which seems to make sense. Netflix TV shows on the other hand seems to be heavily skewed to the right where the majority of shows only have 1 season

Unsupervised ML Techniques

K-Means Clustering

Applied Elbow method and Silhouette Score method for finding best number of clusters





```
For n_clusters = 3 The average silhouette_score is : 0.6221758037723033
```

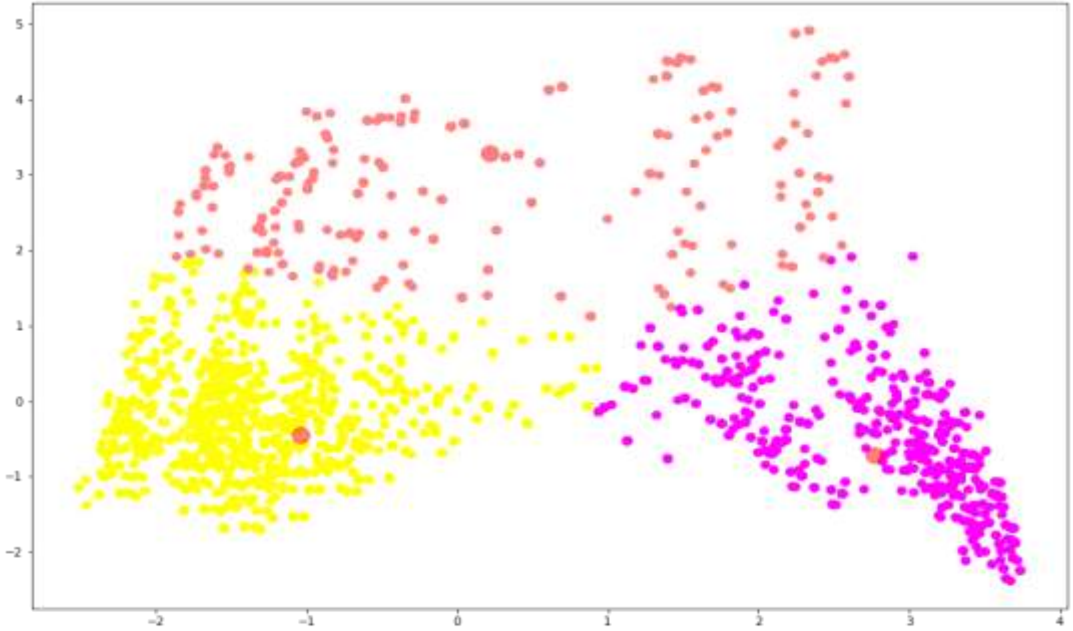
In Elbow method, the optimal k value is 3 since from 3 it's decreasing

The highest Silhouette score is of cluster 3 as there is least outliers

Unsupervised ML Techniques

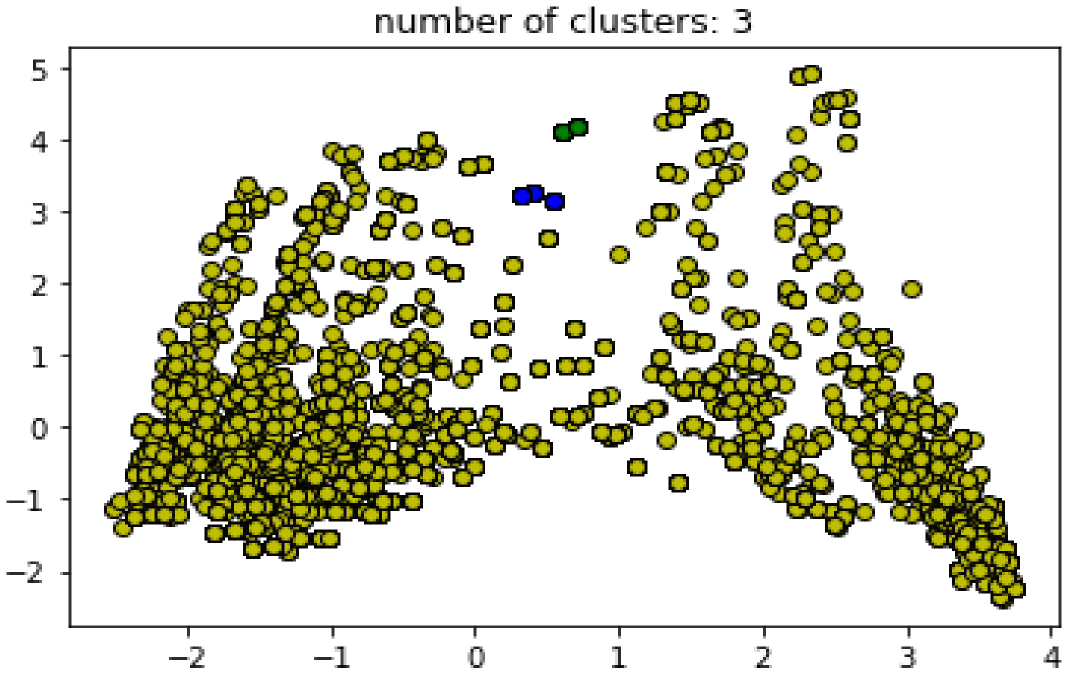
K-Means Clustering

In both Elbow method and Silhouette method we got k = 3



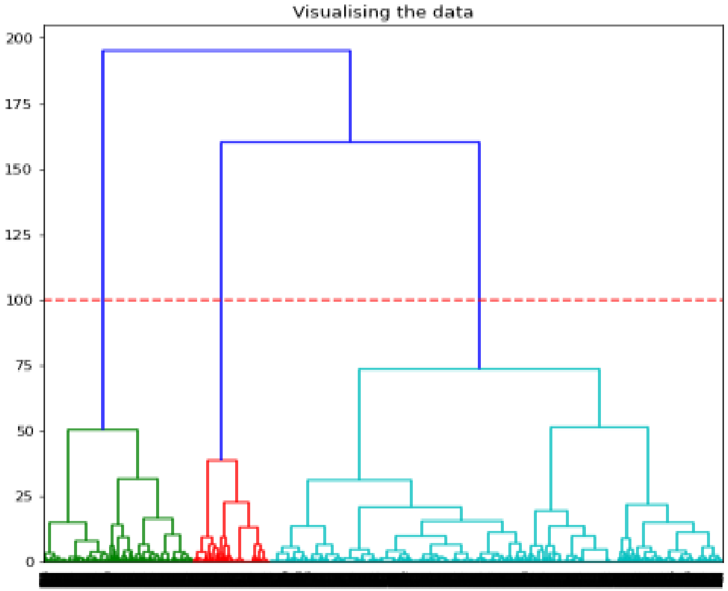
Unsupervised ML Techniques

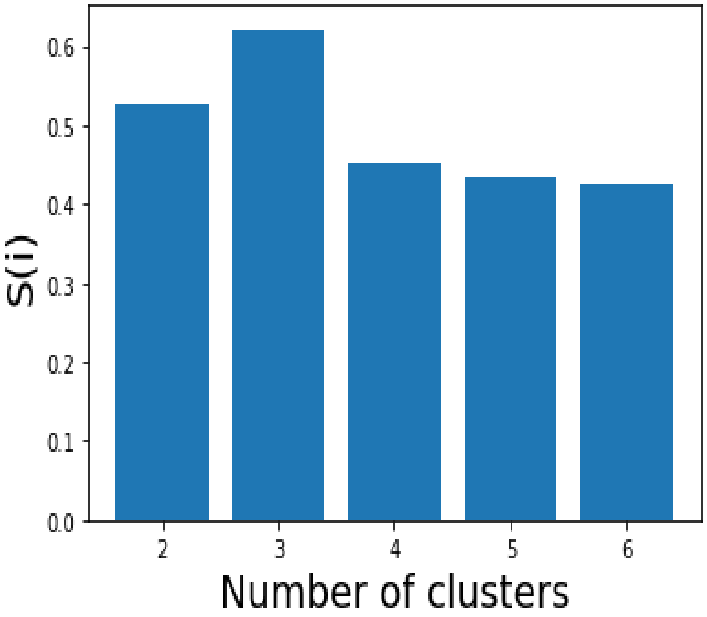
DBSCAN – Density based spatial clustering of application with noise



Unsupervised ML Techniques

Agglomerative Clustering

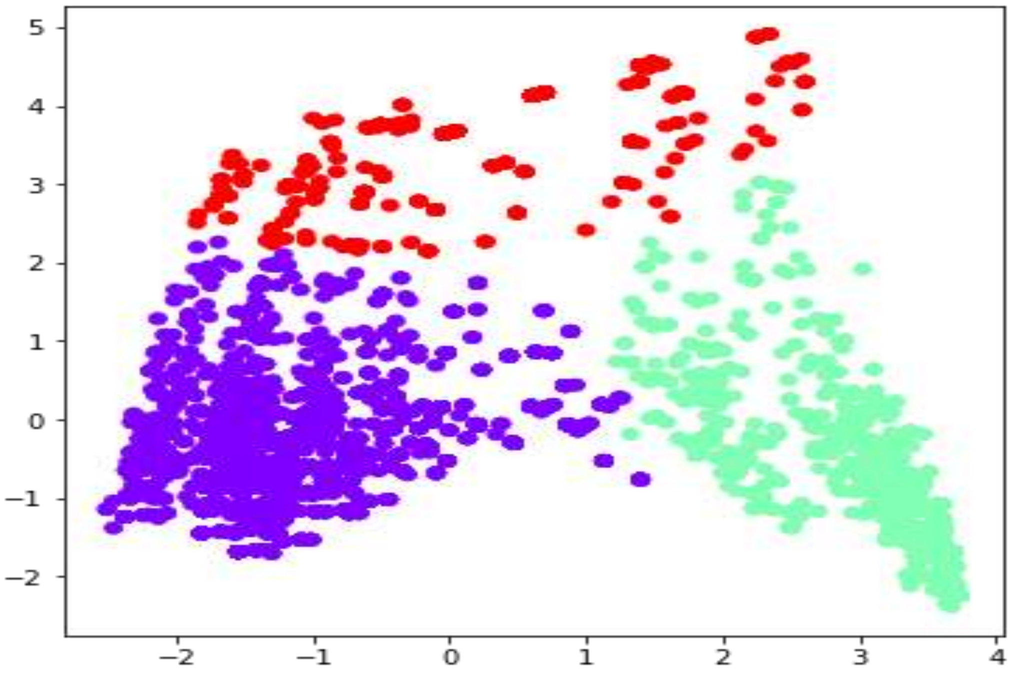




So here also we are getting 3 clusters

Unsupervised ML Techniques

Agglomerative Clustering



Unsupervised ML Techniques

Text based Clustering Techniques

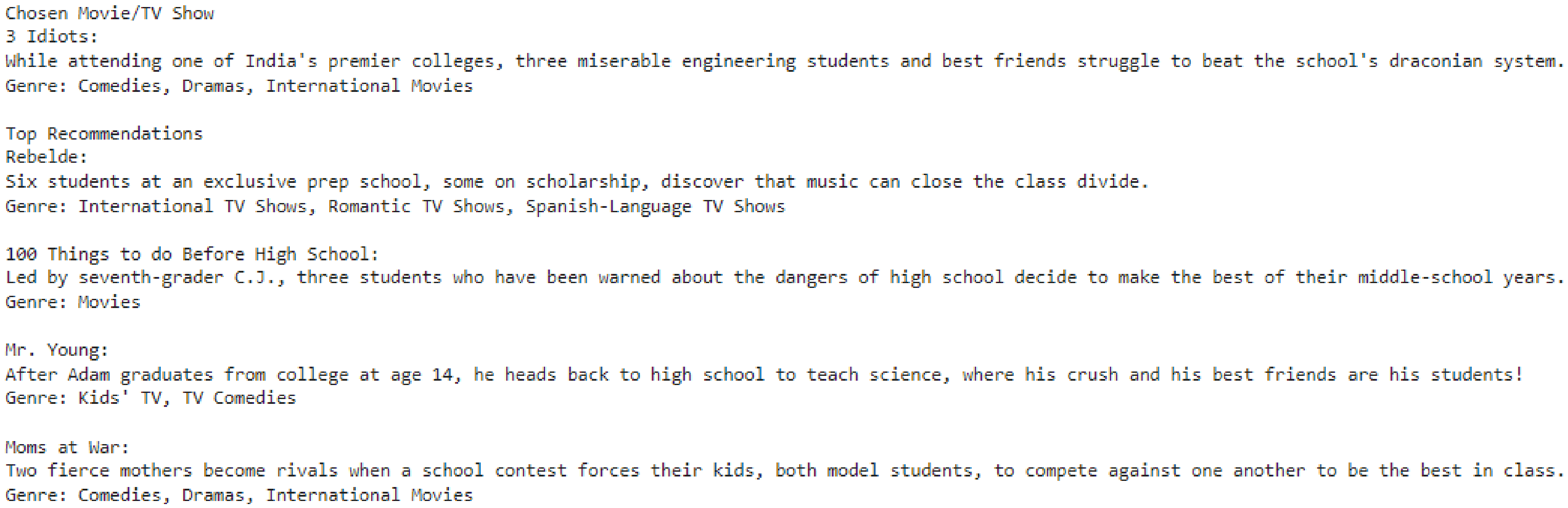
For text based clustering, three techniques were applied.

- NLP
- LSA & LDA under topic modeling

Before applying the text based model, the data was
cleaning and removed all punctuations & stop words

Unsupervised ML Techniques

Natural Language Processing (NLP)

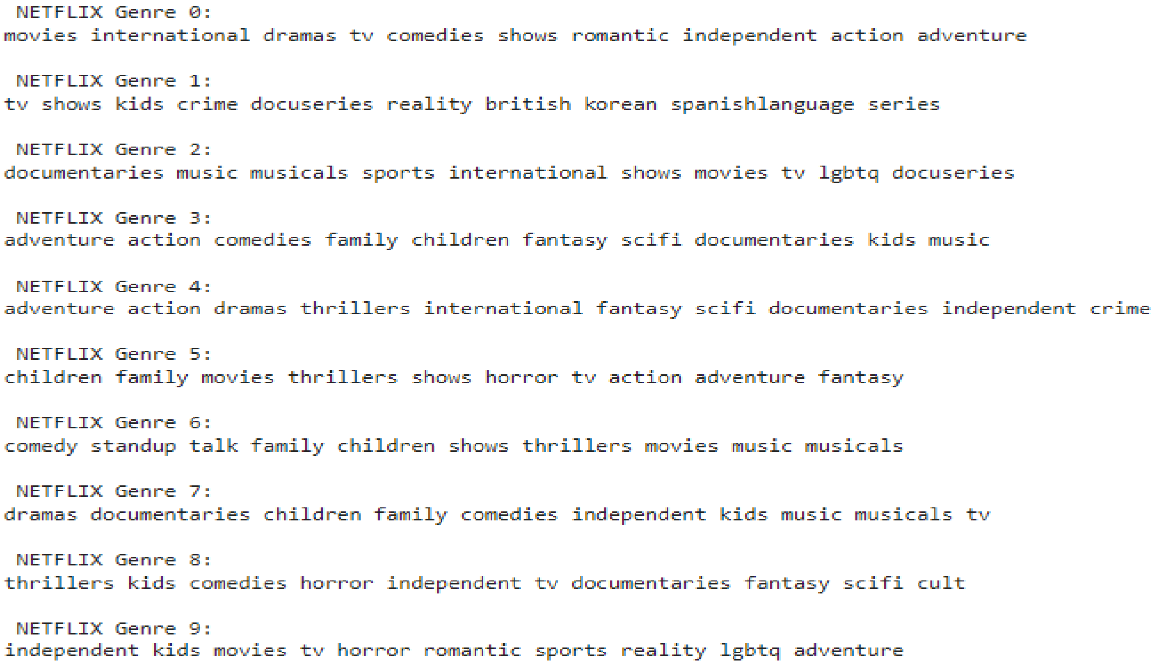


```
Chosen Movie/TV Show
3 Idiots:
While attending one of India's premier colleges, three miserable engineering students and best friends struggle to beat the school's draconian system.
Genre: Comedies, Dramas, International Movies

Top Recommendations
Rebelde:
Six students at an exclusive prep school, some on scholarship, discover that music can close the class divide.
Genre: International TV Shows, Romantic TV Shows, Spanish-Language TV Shows

100 Things to do Before High School:
Led by seventh-grader C.J., three students who have been warned about the dangers of high school decide to make the best of their middle-school years.
Genre: Movies

Mr. Young:
After Adam graduates from college at age 14, he heads back to high school to teach science, where his crush and his best friends are his students!
Genre: Kids' TV, TV Comedies

Moms at War:
Two fierce mothers become rivals when a school contest forces their kids, both model students, to compete against one another to be the best in class.
Genre: Comedies, Dramas, International Movies
```

Unsupervised ML Techniques

Topic Modeling

Latent semantic analysis

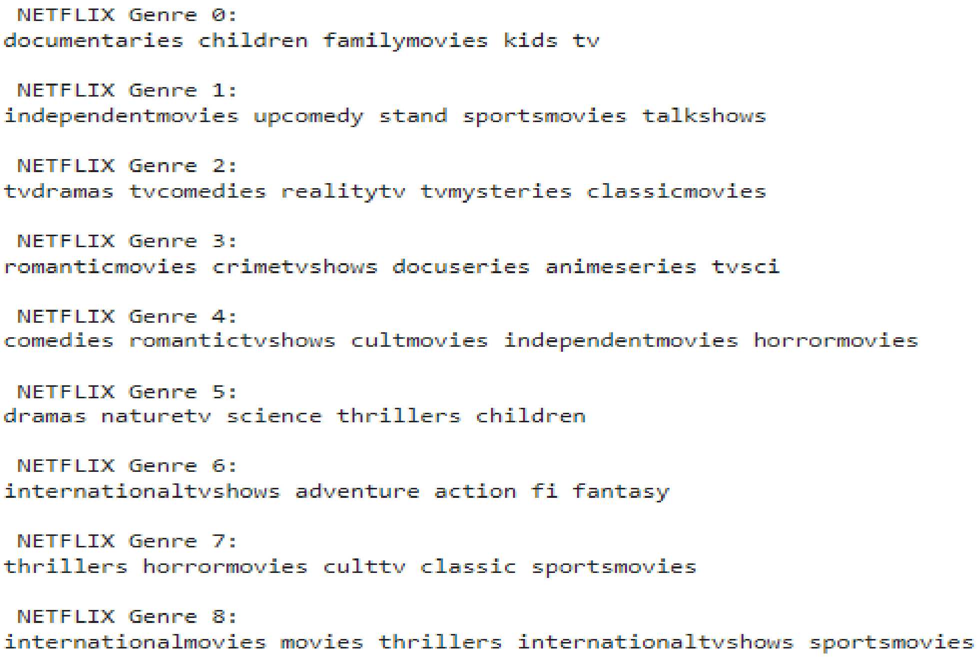


```
 NETFLIX Genre 0:
movies international dramas tv comedies shows romantic independent action adventure

 NETFLIX Genre 1:
tv shows kids crime docuseries reality british korean spanishlanguage series

 NETFLIX Genre 2:
documentaries music musicals sports international shows movies tv lgbtq docuseries

 NETFLIX Genre 3:
adventure action comedies family children fantasy scifi documentaries kids music

 NETFLIX Genre 4:
adventure action dramas thrillers international fantasy scifi documentaries independent crime

 NETFLIX Genre 5:
children family movies thrillers shows horror tv action adventure fantasy

 NETFLIX Genre 6:
comedy standup talk family children shows thrillers movies music musicals

 NETFLIX Genre 7:
dramas documentaries children family comedies independent kids music musicals tv

 NETFLIX Genre 8:
thrillers kids comedies horror independent tv documentaries fantasy scifi cult

 NETFLIX Genre 9:
independent kids movies tv horror romantic sports reality lgbtq adventure
```

Unsupervised ML Techniques

Topic Modeling

Latent Dirichlet Allocation



```
 NETFLIX Genre 0:
documentaries children familymovies kids tv

 NETFLIX Genre 1:
independentmovies upcomedy stand sportsmovies talkshows

 NETFLIX Genre 2:
tvdramas tvcomedies realitytv tvmysteries classicmovies

 NETFLIX Genre 3:
romanticmovies crimetvshows docuseries animeseries tvsci

 NETFLIX Genre 4:
comedies romantictvshows cultmovies independentmovies horrormovies

 NETFLIX Genre 5:
dramas naturetv science thrillers children

 NETFLIX Genre 6:
internationaltvshows adventure action fi fantasy

 NETFLIX Genre 7:
thrillers horrormovies culttv classic sportsmovies

 NETFLIX Genre 8:
internationalmovies movies thrillers internationaltvshows sportsmovies
```

Conclusion

1.All the necessary libraries were imported and data was loaded in the first step.

2.We have 7787 rows and 12 columns in our data frame.

3.After performing EDA we have the below outcome:

- There are about 69% of movies and 31% of TV shows on Netflix
- There are 4 features with null values and that are director - 2389, cast - 718, country - 507, date_added - 10 and rating - 7
- Have cleaned the data by replacing the highest rating (TV-MA) with null values. For the date_added feature, we removed the 10 null values because we can’t predict it, and also it's less in number
- Now other three remaining features director, cast, and the country have a high number of null values so removing null values will impact our readings. So here the null value is replaced with No Director, No Cast, and Country Unavailable respectively
- Our data has TV-MA (TV Mature Audience Only) rating as the highest number followed by TV-14 (unsuitable for children under 14 years of age), TV-PG (Parental Guidance)
- Compared rating of Movies and TV shows differently. In TV shows, the top three ratings are TV-MA, TV-14, and TV-PG (which is the same as overall data) In Movies, the top three ratings are TV-MA, TV-14, and R
- The top three countries for content are US, India, and UK whereas the bottom three are Azerbaijan, Bermuda, and Montenegro

Conclusion

- ➢ The top three genres for content are International Movies, Dramas, and Comedies whereas the bottom three are TV Thrillers, Classics & Cult TV, and TV shows
- ➢ The top three directors are Jan Suter, Raul Campos, and Marcus Raboy whereas the bottom three are Remi Weekers, Jennifer Lynch, and Mozez Singh
- ➢ The top three casts are Anupam Kher, Shah Rukh Khan, and Om Puri whereas the bottom three are Archie Alemania, Demore Barnes, and Rachel Khoo

4. Created three new features called added_day, added_month, and added_year from the date_added feature which is in DateTime format.
5. The highest video content was added in the month of December followed by September and January. Here we can simply assume that these are those months where holidays are more, that too in almost every country.
6. In 2019, maximum video content was added followed by 2020 and 2018.
7. The average viewing time span of movies is about 90 minutes and for TV Shows majority of shows only have 1 season.
8. Applied one hot encoding to convert our text data into numerical data which will make our rescaling process easy.
9. By Using Principal Component Analysis, reduced high dimensional data into 2 dimensional.

Conclusion

10.Next applied different unsupervised techniques. 3 clustering techniques and 2 text-based clustering techniques

a. K-Means Clustering

Applied Silhouette Score Method and highest Silhouette score (0.6) is of cluster k=3.

b. DBSCAN - Density-based spatial clustering of applications with noise

In DBSCAN, three clusters were formed

c. Agglomerative Clustering

Analyzed the data for different numbers of clusters and the highest Silhouette score is for cluster k=3

d. Natural Language Processing (NLP)

Converted each text into vectors and applied the NLP method. Grouped all the similar content and formed a recommendation system that worked properly.

e. Topic Modeling

To sort similar genres, applied two methods below:

- Latent semantic analysis
- Latent Dirichlet Allocation

Got similar and related genres that can be recommended to the viewers

So here we conclude that 3 clusters are optimal clusters for our data

Challenges

- The most difficult thing I faced in this project was applying one hot encoding. I was getting error again and again
- Also I got afraid if I do not miss any other important clustering technique
- The coding of NLP was also difficult

Thank You!