Netflix Movies &TV Shows Recommender System

MC 594 Seminar

Group 1
Garima Bansal *(2211MC05)*
Konika Mandal *(2211MC17)*
Nikunj Pansari *(2211MC21)*

Indian Institute of Technology Patna

Contents

- Introduction
- Objective
- Data Visualization (using Python)
- Model Implementation : Bagging
- Analysis
- Conclusion
- Further Scope
- References

Introduction

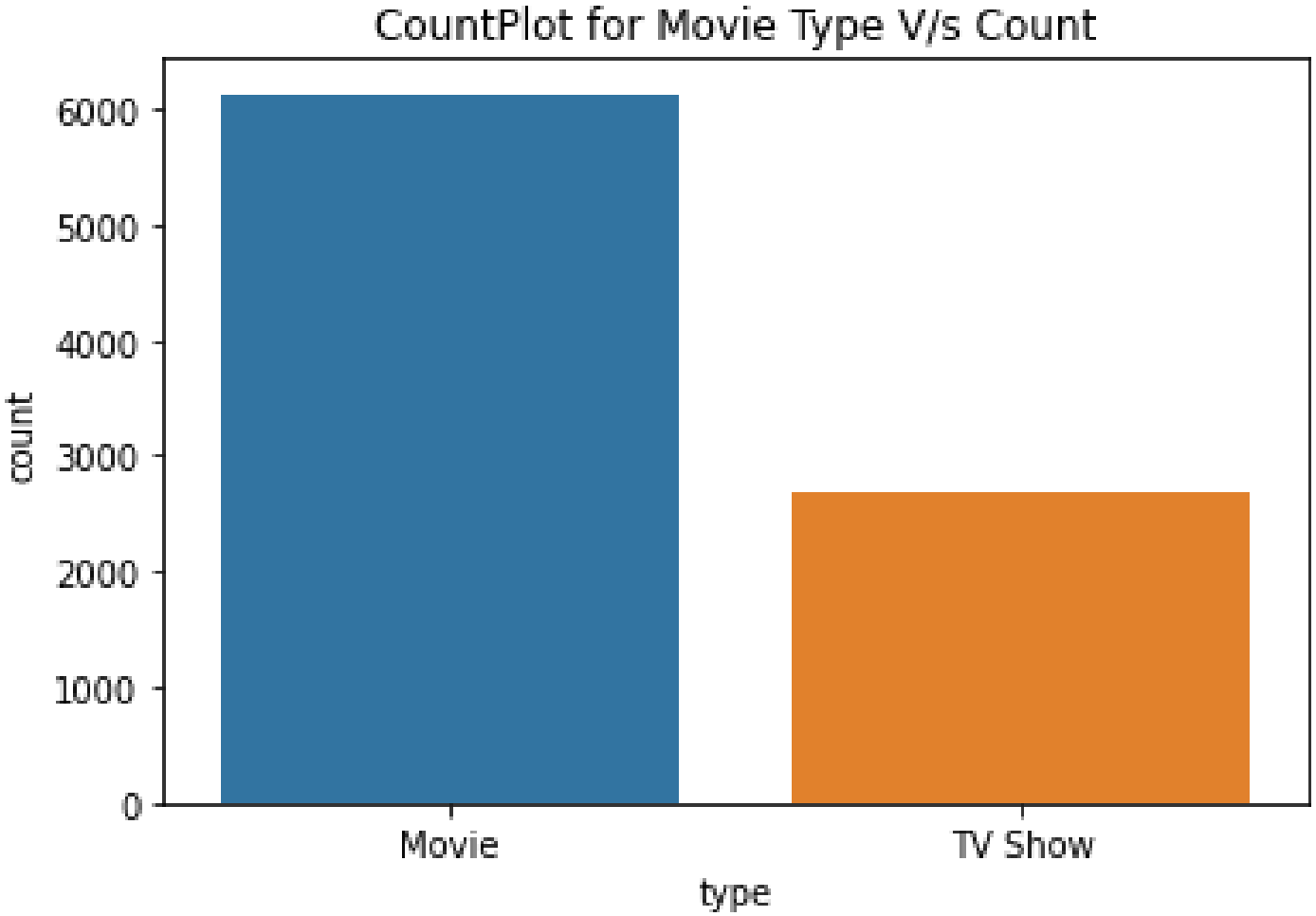
- Choosing a movie or TV show to watch can be difficult at times, when a similar category of movie or shows are available
- Thus, there comes the filter of rating
- Data analysis of Netflix dataset would be beneficial in building a movie or show recommender system based on some metrics like ratings and few others.

Objective

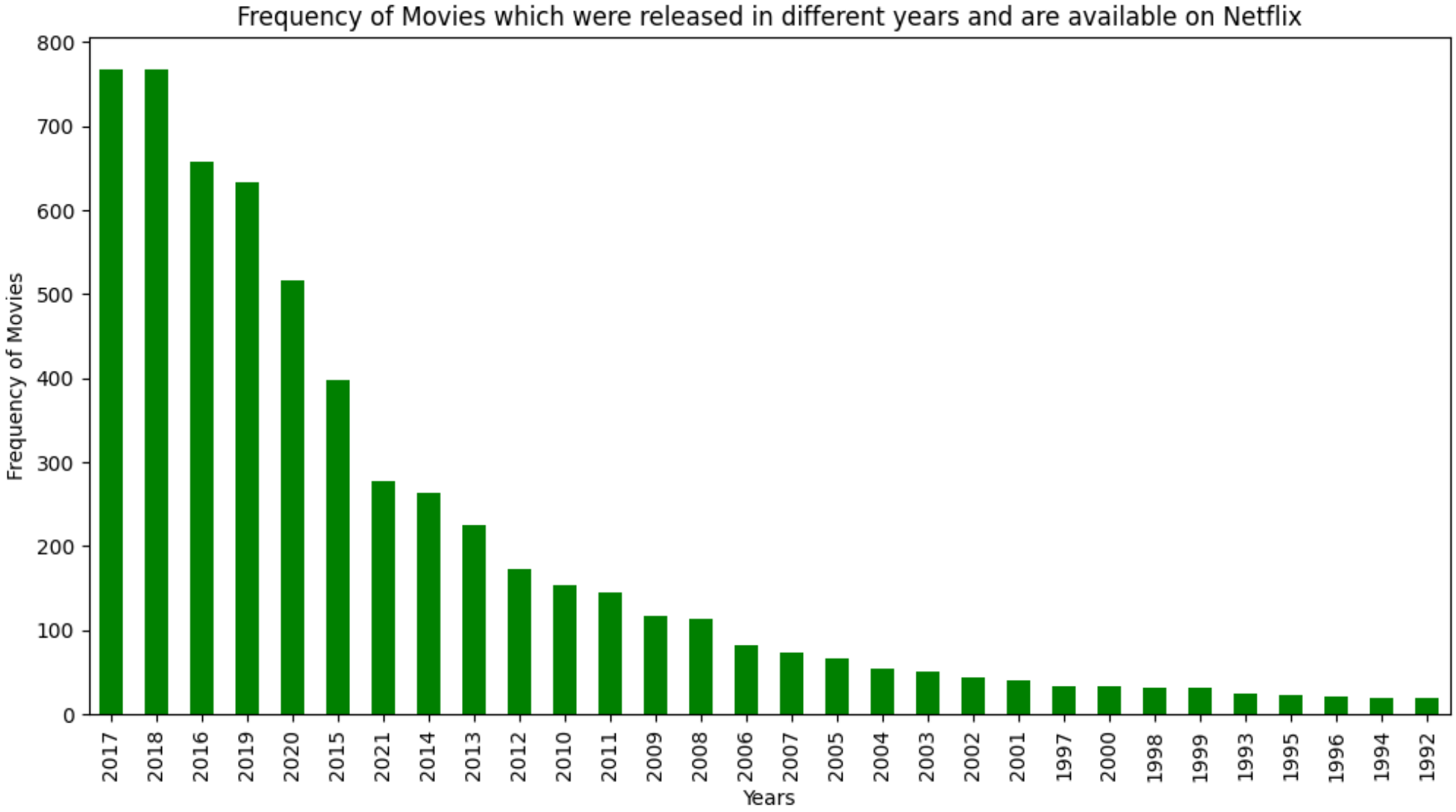
- To analyze the content released by Netflix and gain some insights into the way forward
- To identify which genres are most popular and find content that fits their personal preferences
- Help users to discover new content according to different age groups

Data Visualization

- To show relationship between different attributes, data visualization is important for proper analysis.
- Movies are more than TV Shows

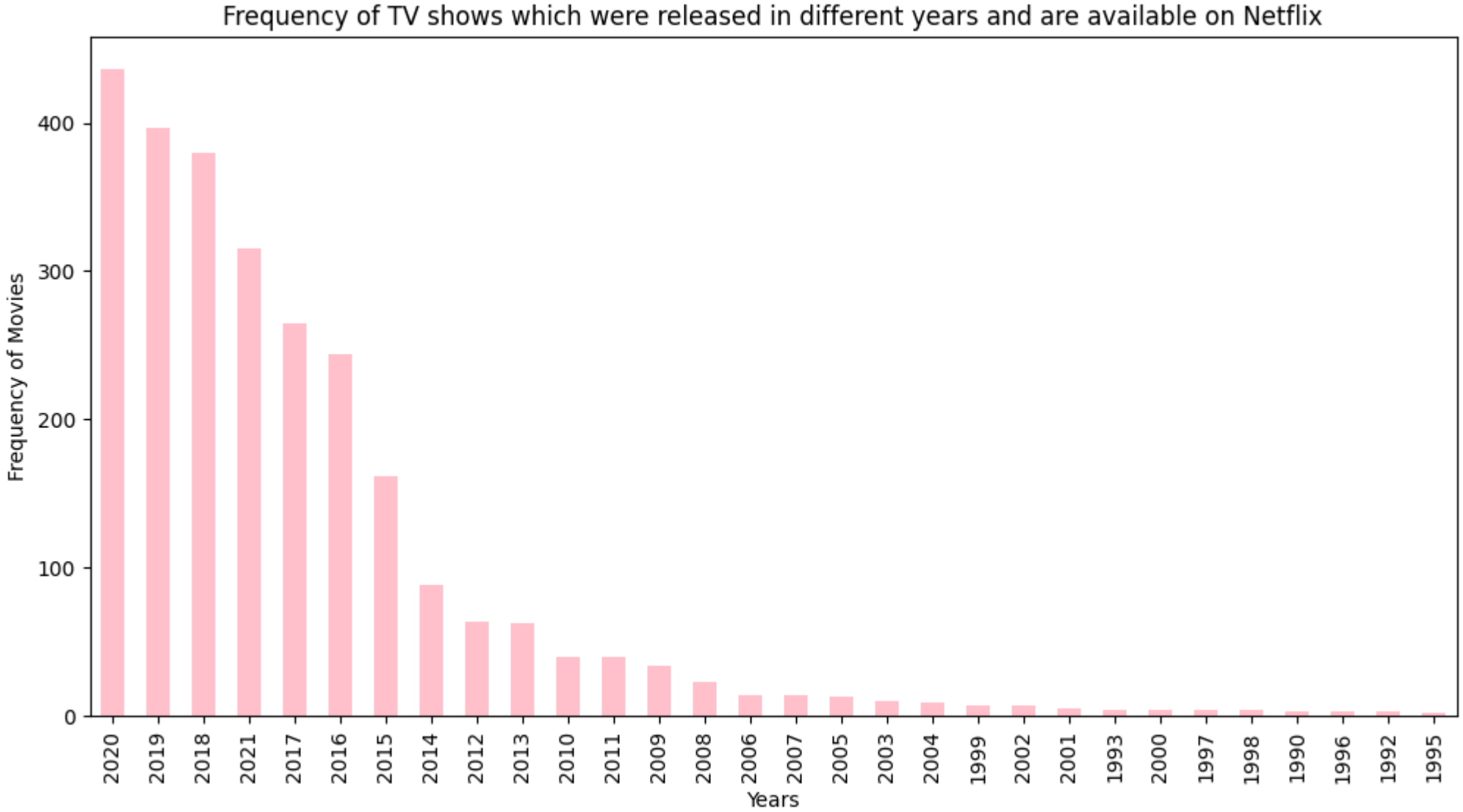


Bar Plot



- Highest number of movies were added on Netflix in year 2017-18
- Growth rate is increasing slowly every year.

Bar Plot



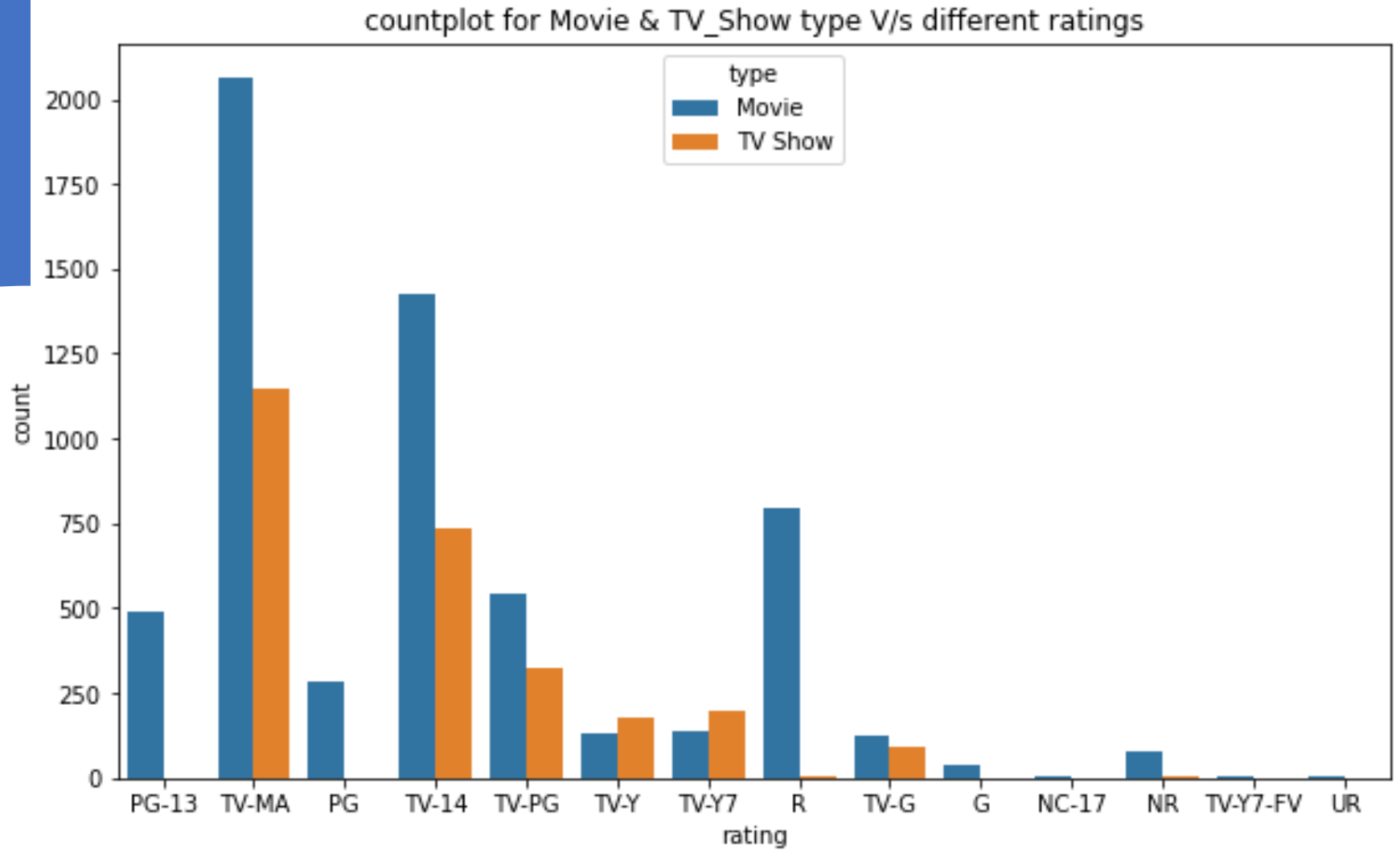
- Highest number of TV Shows were added on Netflix in year 2020
- From year 1995-2011 very less TV shows were added each year, after that it increased rapidly

Rating

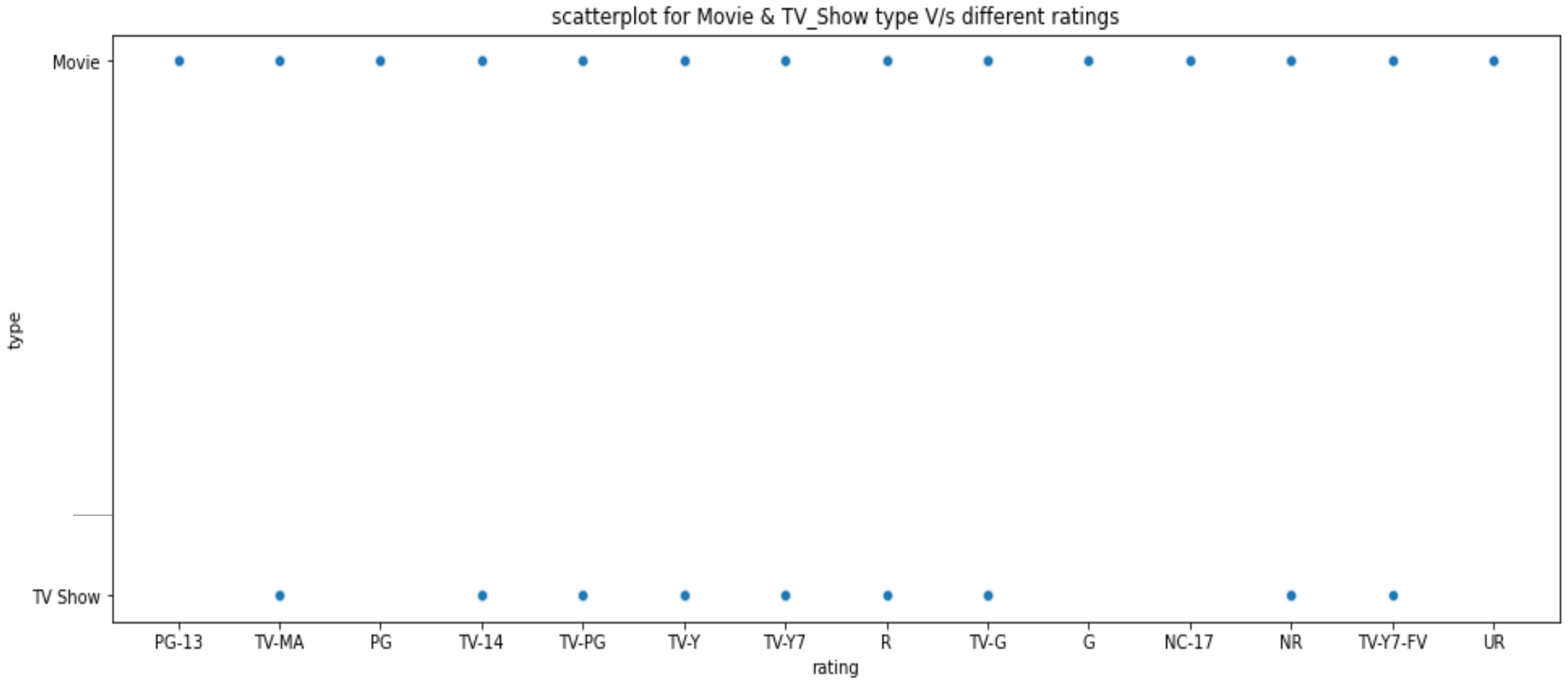
PG-13 Inappropriate For Children Under 13	**TV-MA** Mature Audience Only	**PG** Parental Guidance under age 15	**TV-14** Unsuitable for children under 14	**TV-PG** Parental guidance under age 14
TV-Y Aimed at age 2-6	**TV-Y7** Most appropriate for age 7 and up	**R** Under 17 can watch under PG	**TV-G** Suitable for general audience contains little or no violence	**G** Appropriate for people of all ages
NC-17 No children 17 and under can watch	**NR** Not Rated , reviewed but no rating assigned	**TV-Y7-FV** Fantasy Violence more than TV-Y7	**UR** Unrated , not rated at all	

Count Plot

- TV- MA rating has highest count for both movies & TV Shows
- Mature Audience content is highly demanded

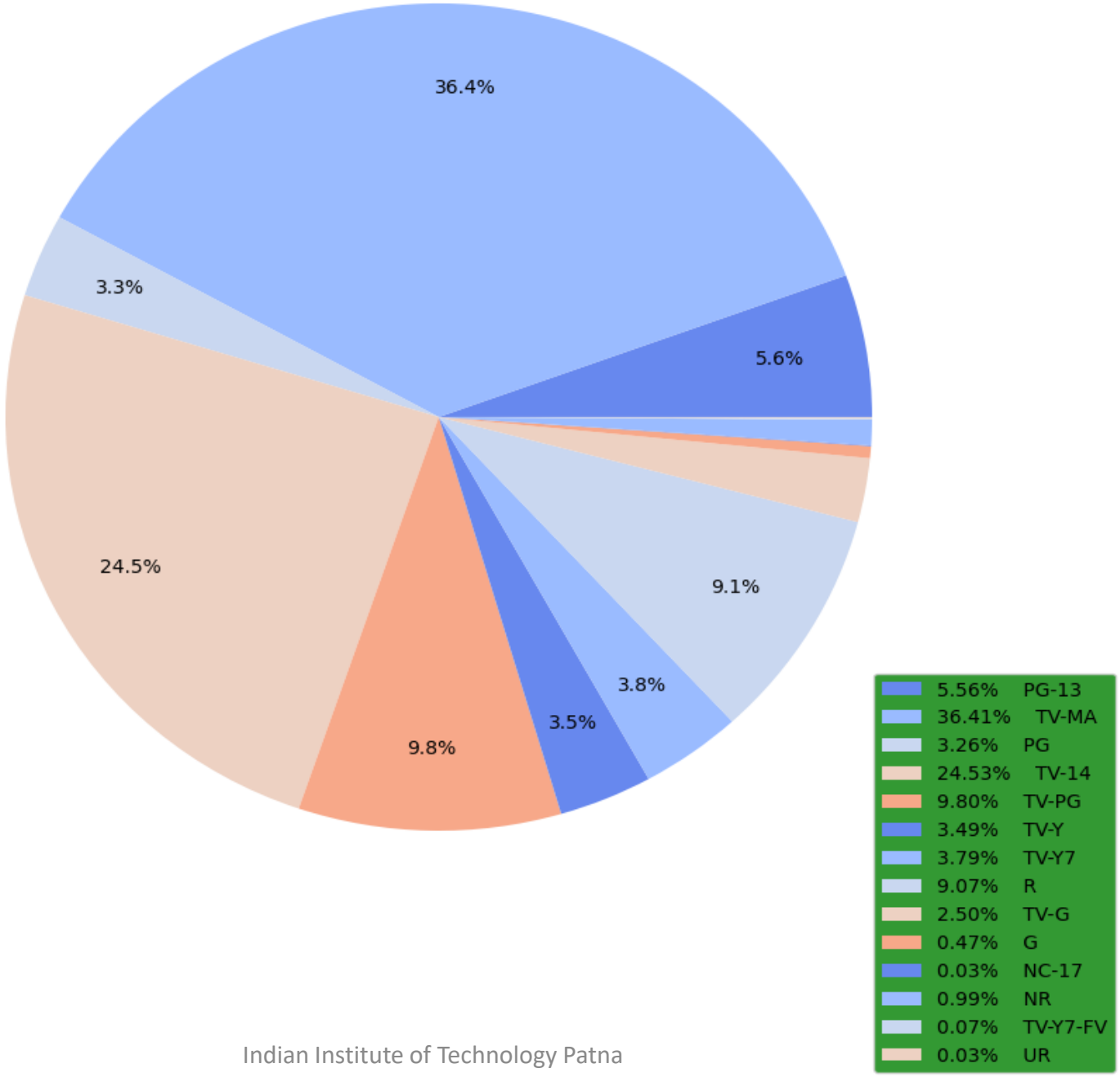


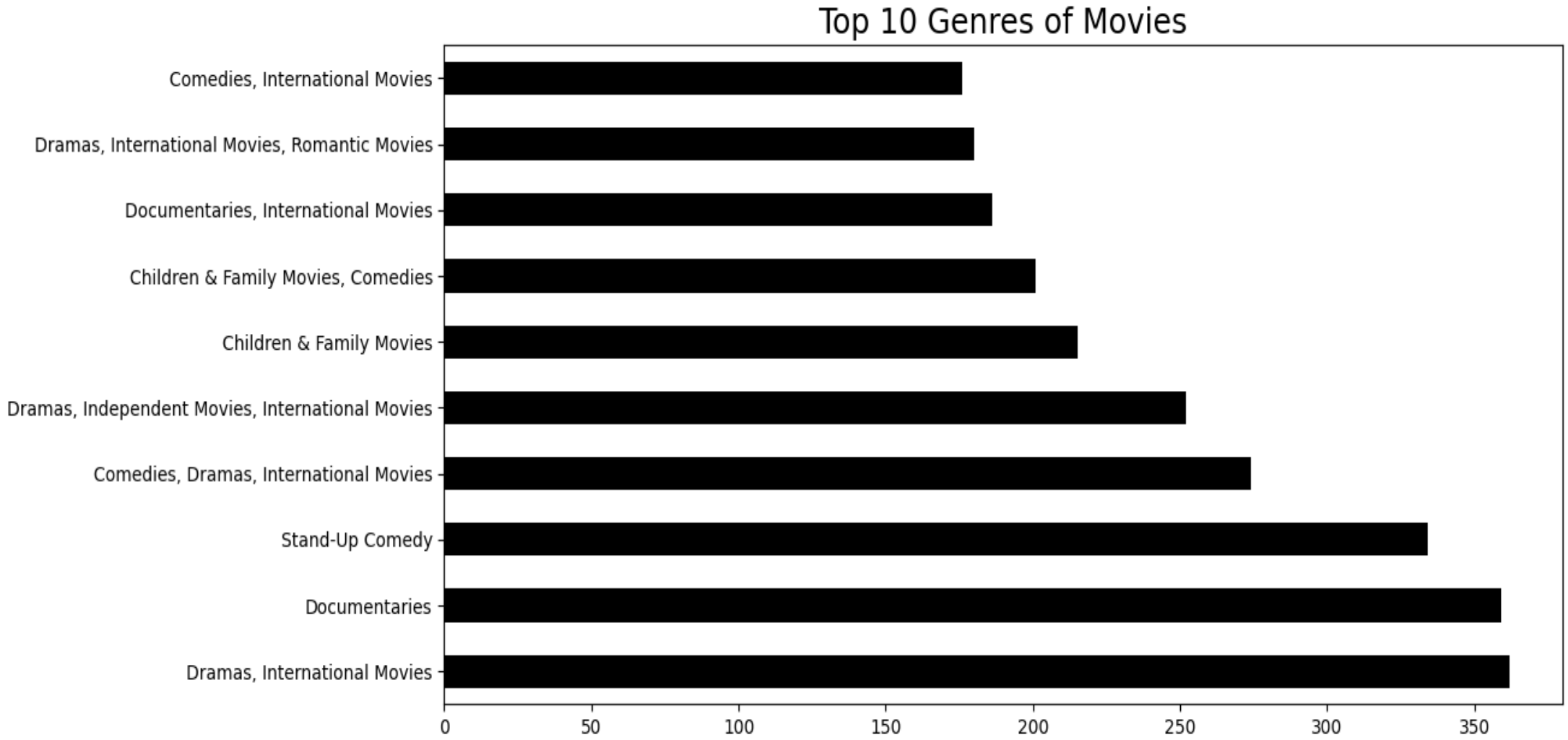
Scatter Plot



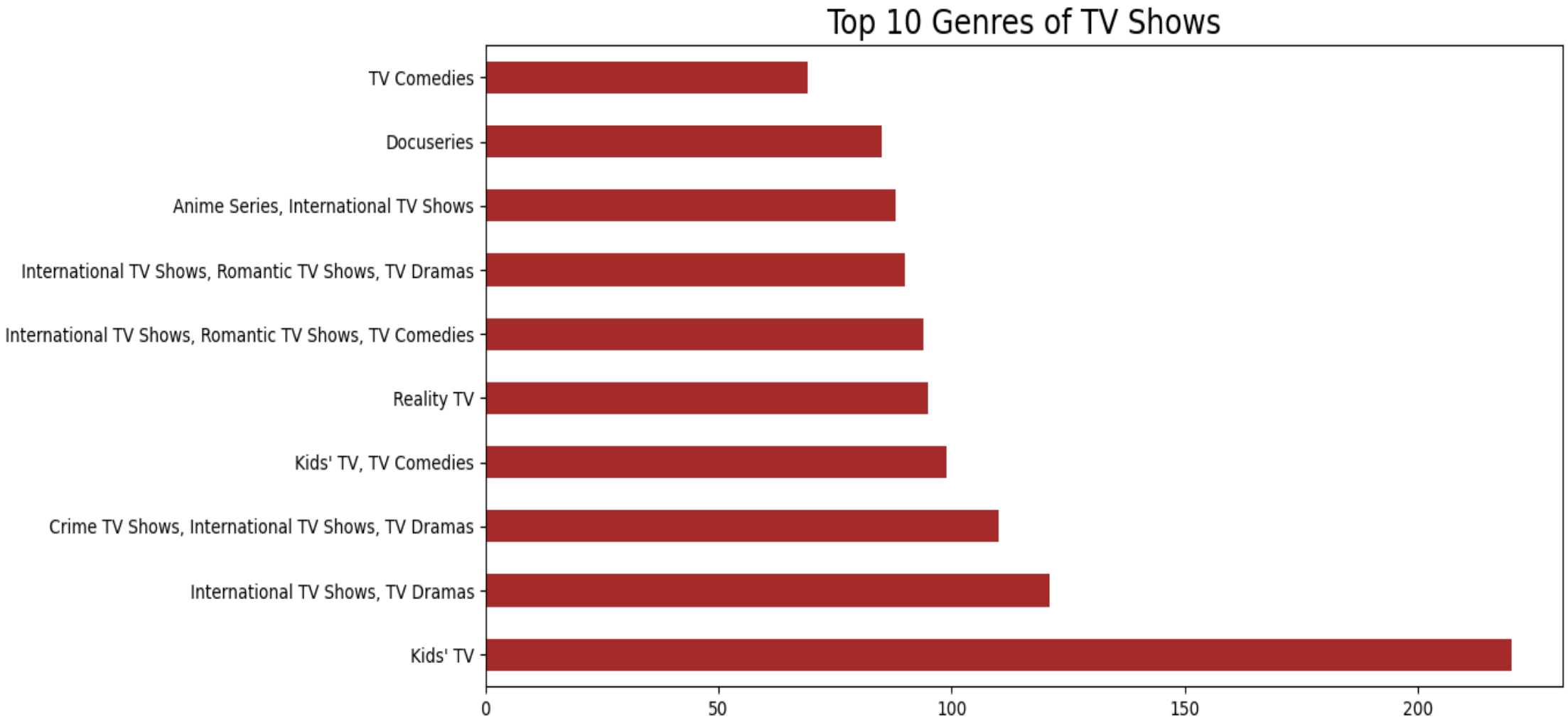
- Content for movies are available in all types of rating
- TV Shows are not available with PG-13, PG, G, NC-17 and UR ratings

Pie Chart


36.4%
3.3%
5.6%
24.5%
9.1%
3.8%
3.5%
9.8%
5.56% PG-13
36.41% TV-MA
3.26% PG
24.53% TV-14
9.80% TV-PG
3.49% TV-Y
3.79% TV-Y7
9.07% R
2.50% TV-G
0.47% G
0.03% NC-17
0.99% NR
0.07% TV-Y7-FV
0.03% UR



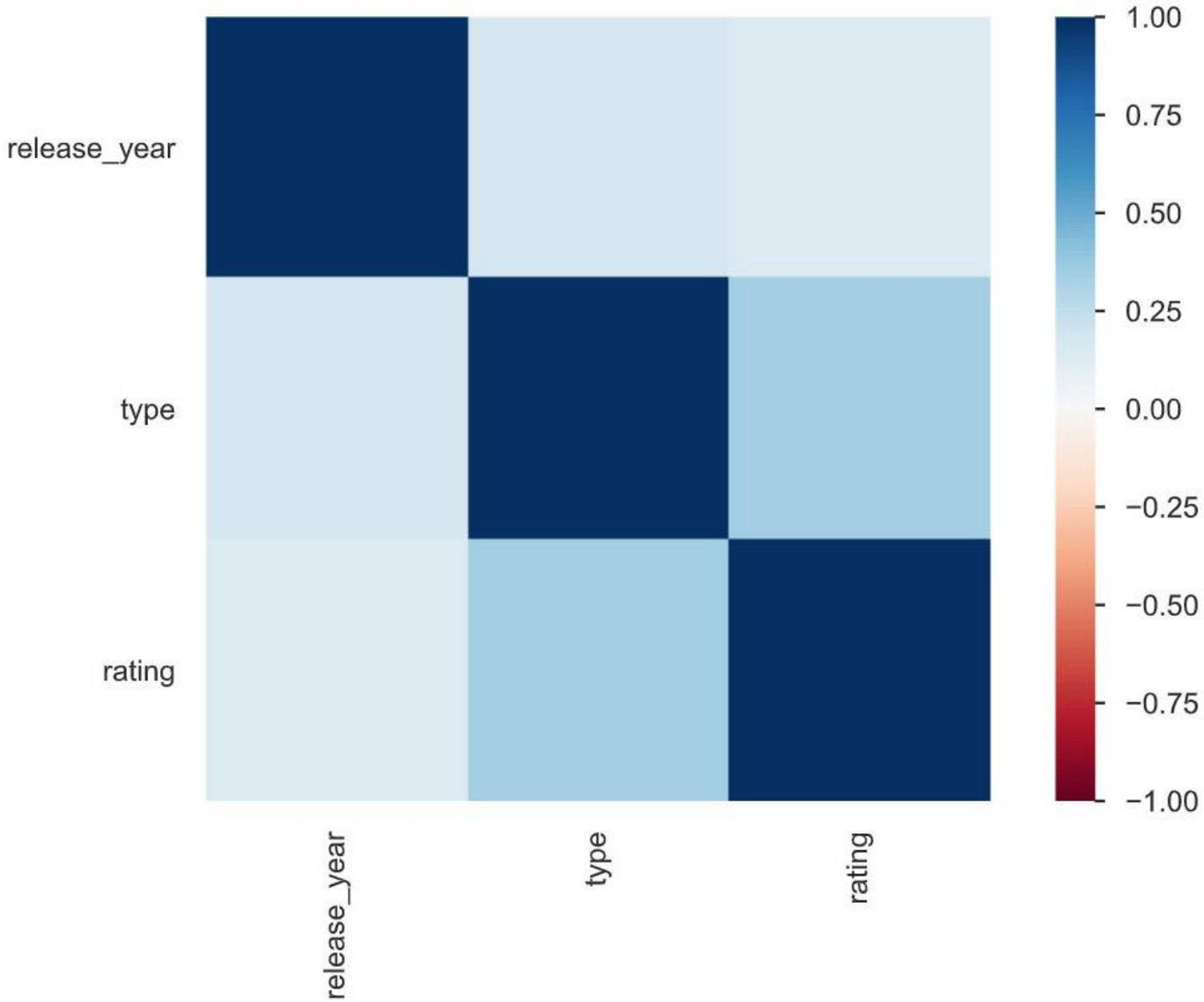
- Most of the users are interested in Dramas, International Movies among movies available on Netflix



- Most of the users are interested in kid's TV among TV shows available on Netflix

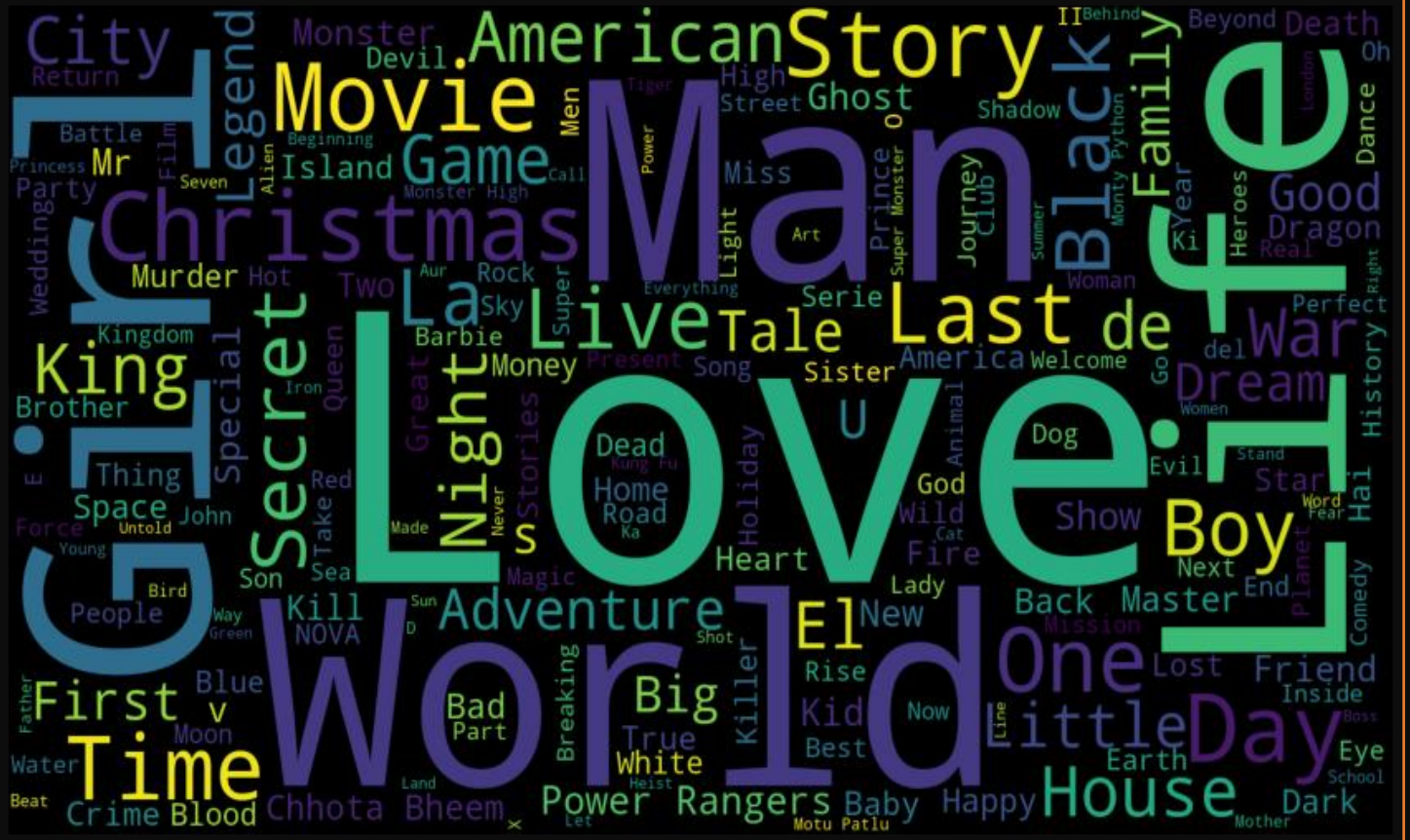
Correlation Matrix (Heat Map)

- The release year directly doesn’t imply the type of the movie or the show
- The dark blue color shows a high dependency, light blue shows less dependency between the attributes



WordCloud

- **Here the size of each word indicates its importance or frequency in the whole dataset**
- **Frequency of the word “LOVE” is maximum in our dataset**



Ensemble Learning

Multiple models, often called base models, are combined to produce an effective optimal prediction model

Bagging

- Ensemble learning technique
- Bootstrap Aggregation
- Involves randomly sampling the training data with replacement
- Used for both regression & classification models

Decision Tree

- Tree-structured classifier
- Internal nodes : features
- Branches : decision rules
- Each leaf node : outcome

Steps of Bagging Classifier

- Pre-processing the data
- Splitting the data
- Building the Bagging Classifier
- Tuning the Hyper parameters
- Evaluating the model performance

WHY WE USED THIS ML MODEL?

Handling multi-class classification problem

To minimize overfitting of data

To improve accuracy of the model

High-dimensionality of the input data

Data Before Encoding

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Encoded Data

	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	relevant	show_id
0	0	1975	2295	7677	603	1713	72	4	210	274	2577	0	0
1	1	1091	4516	409	426	1712	73	8	110	414	1762	0	1111
2	1	2651	2105	6296	748	1712	73	8	0	242	7341	0	2222
3	1	3506	4516	7677	748	1712	73	8	0	297	3617	0	3333
4	1	3861	4516	4815	251	1712	73	8	110	393	4416	0	4444
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8802	0	8770	979	4677	603	1108	59	5	70	269	895	0	8671
8803	1	8773	4516	7677	748	987	70	11	110	424	8483	1	8672
8804	0	8774	3631	3231	603	1092	61	5	206	207	5228	0	8673
8805	0	8777	3247	7061	603	1155	58	3	206	125	3315	0	8674
8806	0	8781	2926	7297	251	870	67	6	16	328	1004	0	8675

8807 rows × 13 columns

Results & Analysis

```
Training Accuracy : 0.981
Test Accuracy : 0.514
```

```
precision_recall_fscore_support(y_test,y_pred,average='macro')
```

```
(0.4336404658537637, 0.3911153988920781, 0.404165924991073, None)
```

Using Bagging classifier on Netflix Movies & TV Shows dataset we can observe that our model's accuracy on training set is 98% and 51% on test set. (F1- score : 0.4)

Formulas

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Conclusion

- The Bagging Classifier is simple & effective for the selection of the movies or TV shows for different age group of people .

- The model is not fitting the dataset as is, we need to choose a different model .

FUTURE SCOPE

- Try a different ensemble method like boosting, or stacking
- Hyperparameter tuning
- Different type of algorithms can be used such as matrix factorization or Deep Learning models which are specifically designed for collaborative filtering

Contributions

Garima Bansal *(2211MC05) : 33.33%*

Konika Mandal *(2211MC17) : 33.33%*

Nikunj Pansari *(2211MC21) : 33.33%*

References

- https://www.kaggle.com/datasets/shivamb/netflixshowsdatasetId=434238&sortBy=voteCount
- https://github.com/shinanna/Netflix_MachineLearning
- Vybhav Achar Bhargav, Seongwoo Choi, and David Haddad. Dataanalysis on netflix datasets. 03 2022
- Karthik Babu Vadloori and Shriya Madhavi Sanghishetty. Exploratory and sentiment analysis of netflix data. International Journal of Engineering Research & Technology (IJERT), 10(9):213–217, 2021

Thank You!!

Any Questions !!