Capstone Project 04

Netflix Movies and TV Shows Clustering

By: Prashant Gaikwad
(Individual Project)

Points for Discussion

- ❑ Introduction
- ❑ Problem Statement
- ❑ Netflix Movies and TV Shows Dataset
- ❑ Data Exploration
- ❑ Data Preparation & Cleaning
- ❑ Exploratory Data Analysis (EDA)
- ❑ Feature Reduction & Feature Engineering
- ❑ Clustering Model Building
- ❑ Word Cloud
- ❑ Conclusion

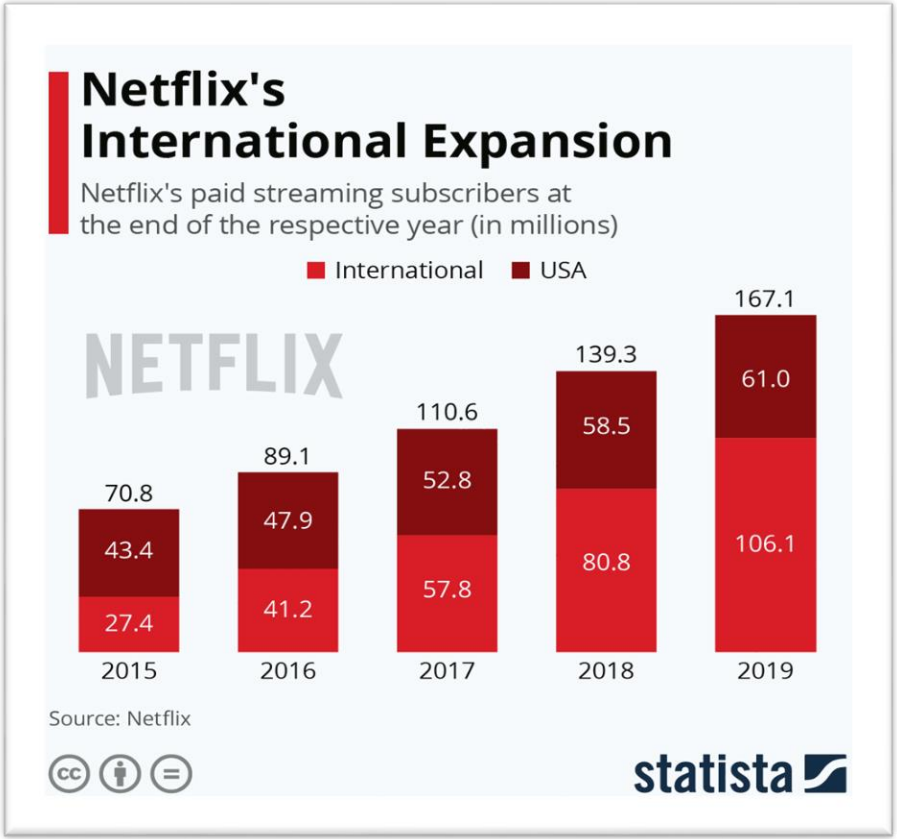
INTRODUCTION

- Many online streaming services offer a large number of TV shows, which are at our disposal to watch, at the price of a subscription cost.

- Netflix is a popular entertainment service used by people around the world.

- We will explore the Netflix dataset through visualizations, graphs using python libraries and clustering model building.

- We will explore the Netflix dataset through visualizations like graphs using python libraries
- Understanding what type content is available in different countries
- Is Netflix has increasingly focusing on TV rather than movies in recent years.
- Clustering similar content by matching text-based features



PROBLEM STATEMENT

Netflix is an American technology and media services provider and production company headquartered in Los Gatos, California. Netflix was founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California. The company's primary business is its subscription-based streaming service, which offers online streaming of a library of films and television series, including those produced in-house.

Their most successful algorithm, Netflix Recommendation Engine (NRE), is made up of algorithms which filter content based on each individual user profile. The engine filters over 3,000 titles at a time using 1,300 recommendation clusters based on user preferences.

It's so accurate that 80% of Netflix viewer activity is driven by personalized recommendations from the engine.

Netflix Movies and TV Shows Dataset

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

5. cast : Actors involved in the movie / show

6. country : Country where the movie / show was produced

7. date_added : Date it was added on Netflix

8. release_year : Actual Releaseyear of the movie / show

9. rating : TV Rating of the movie / show

10. duration : Total Duration - in minutes or number of seasons

11. listed_in : Genere

12. description: The Summary description

DATA EXPLORATION

- Data exploration refers to a data user being able to find his or her way through large amounts of data in order to gather necessary information.

- Explored and analyzed the data to discover key factors responsible for app engagement and success.

DATA PREPARATION & CLEANING

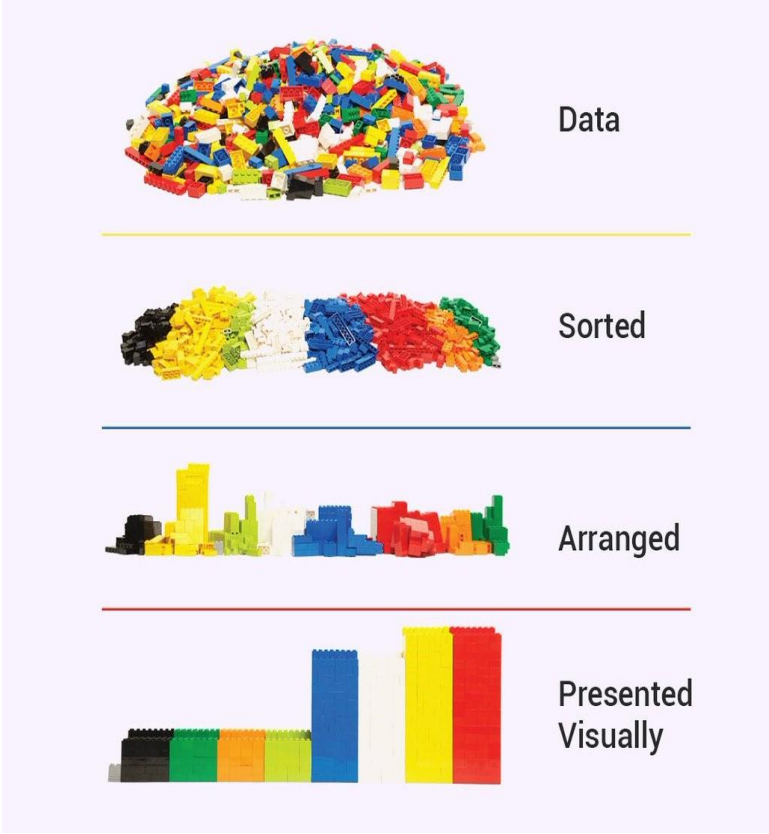
- Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data, and the combining of data sets to enrich data.

- Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

- Dataset contains a large number of null values which might tend to disturb our analysis hence we dropped some unnecessary variables.
- Used “fillna()” method to filled some required variable null values from them at the beginning of project in order to get a better result.
- Encoded all categorical variable to numeric values

EXPLORATORY DATA ANALYSIS

- Exploratory Data Analysis is a process of examining or understanding the data and extracting insights or main characteristics of the data.
- **EDA** is very essential because it is a good practice to first understand the problem statement and the various relationships between the data features before getting hands dirty.



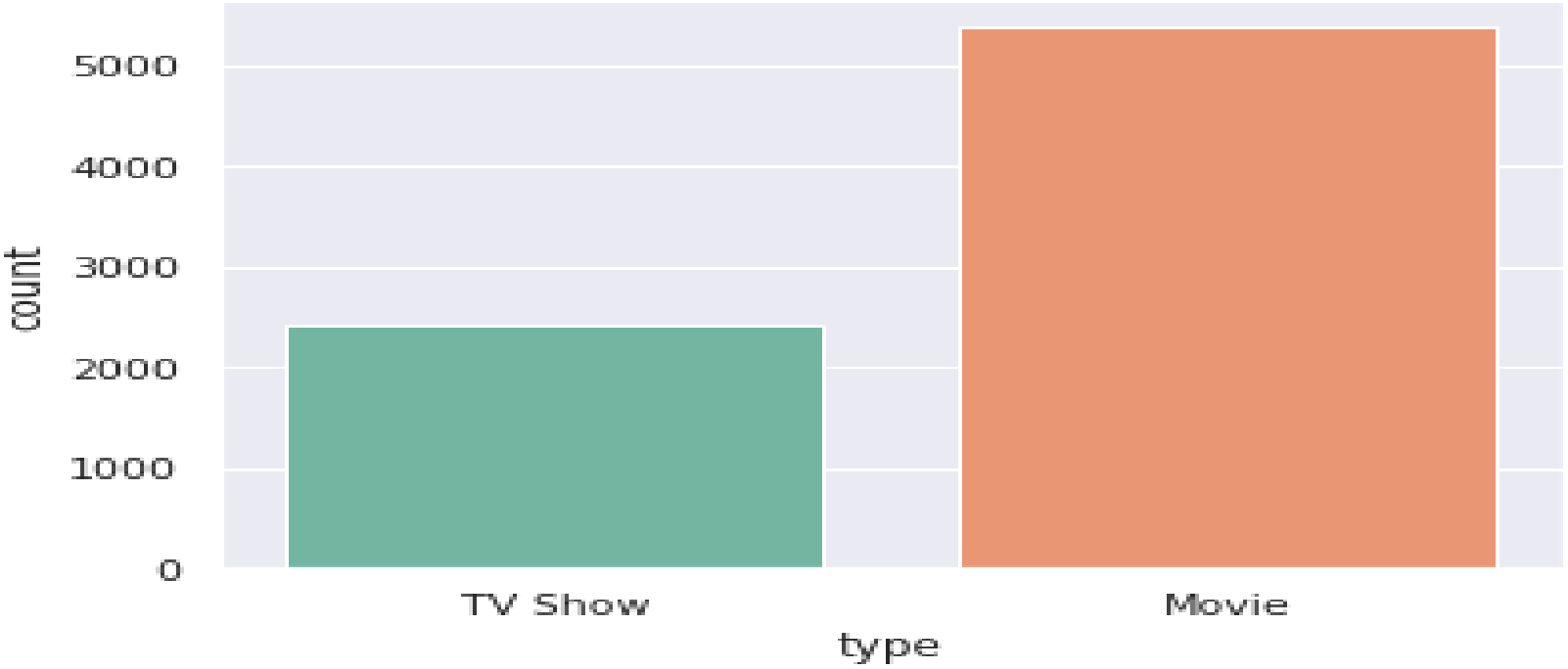
Types of EDA are:

- Univariate Analysis - analysis of a single variable
- Bivariate Analysis - analysis of exactly two variables
- Multivariate Analysis - analysis of dependent variable and multiple independent variables

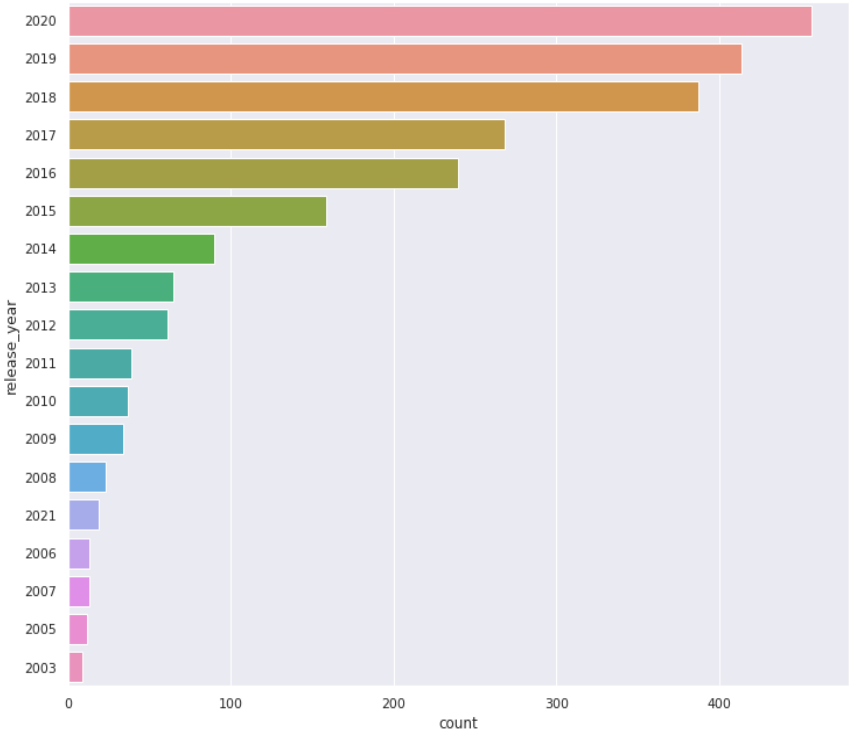
The primary motive of EDA is to

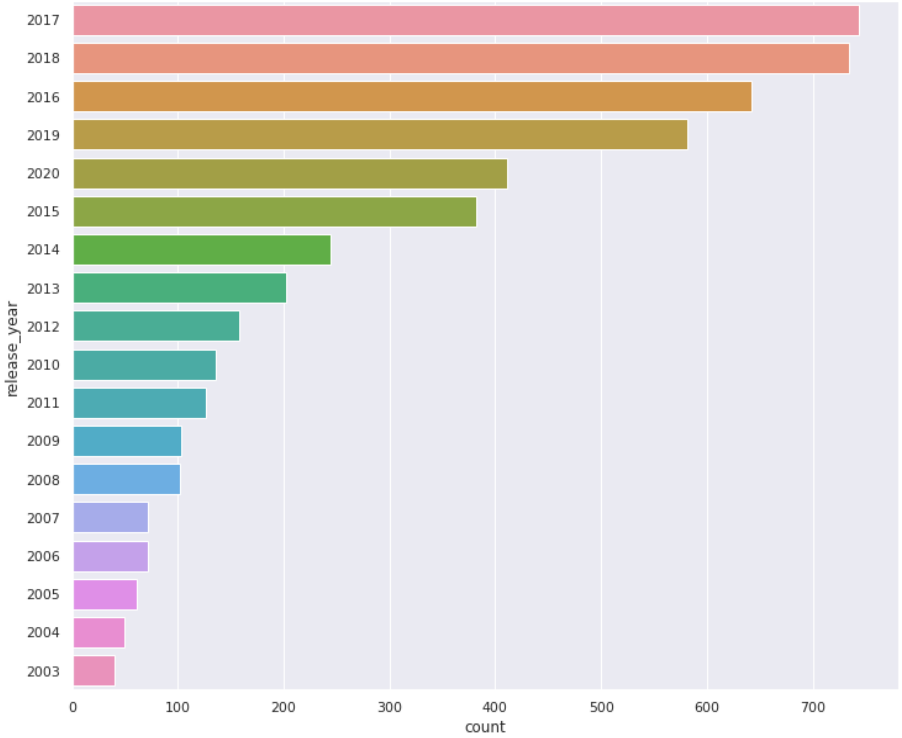
- Examine the data distribution
- Handling missing values of the dataset (a most common issue with every dataset)
- Handling the outliers
- Removing duplicate data
- Encoding the categorical variables
- Normalizing and Scaling

Type of content

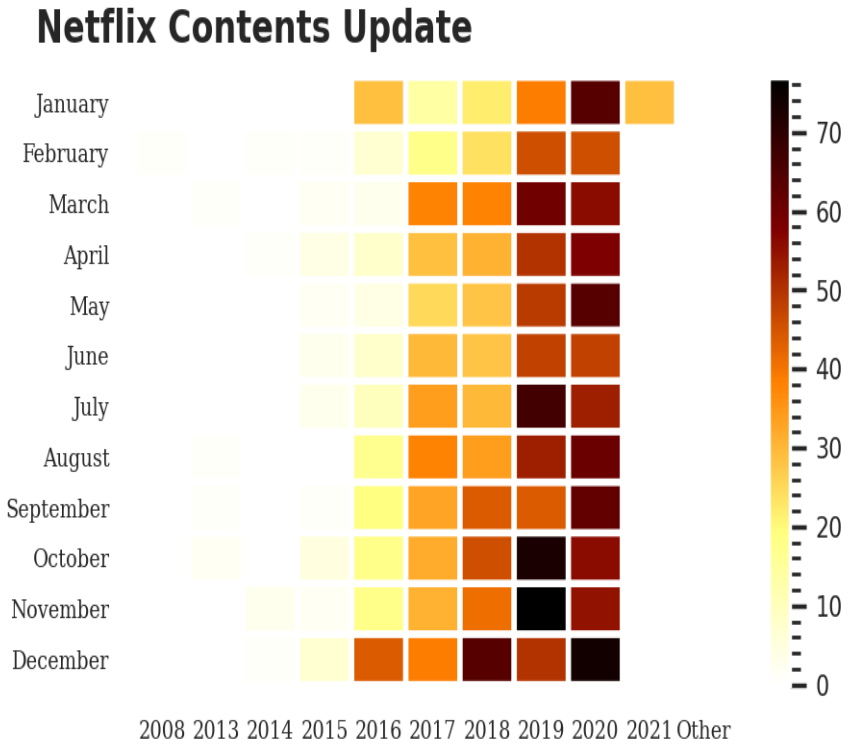


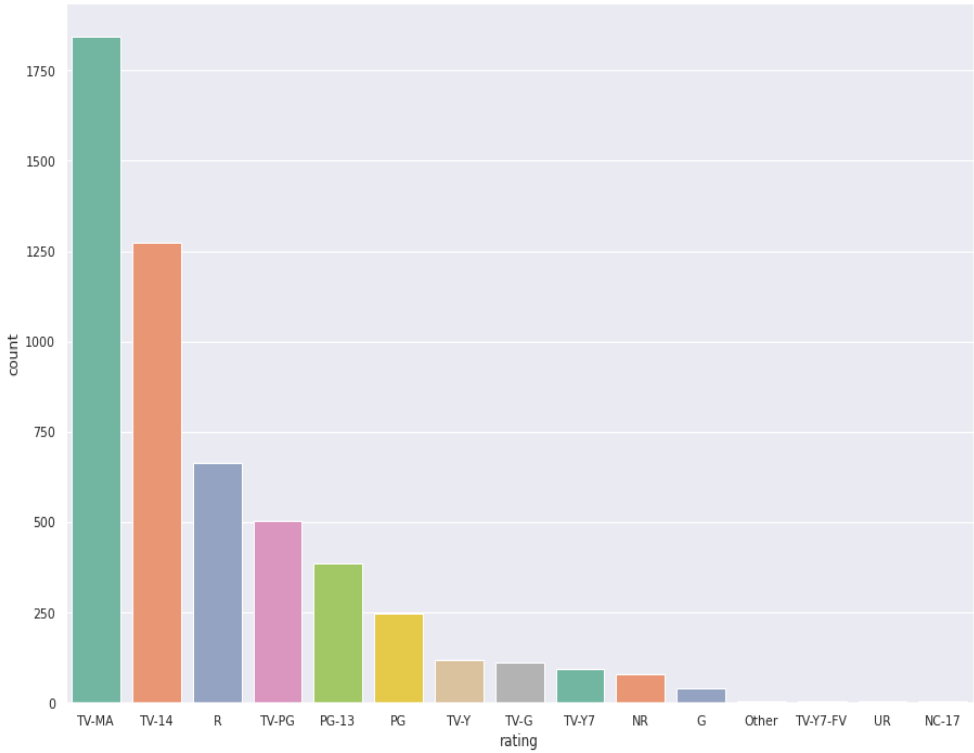
Release year of Shows and Movies



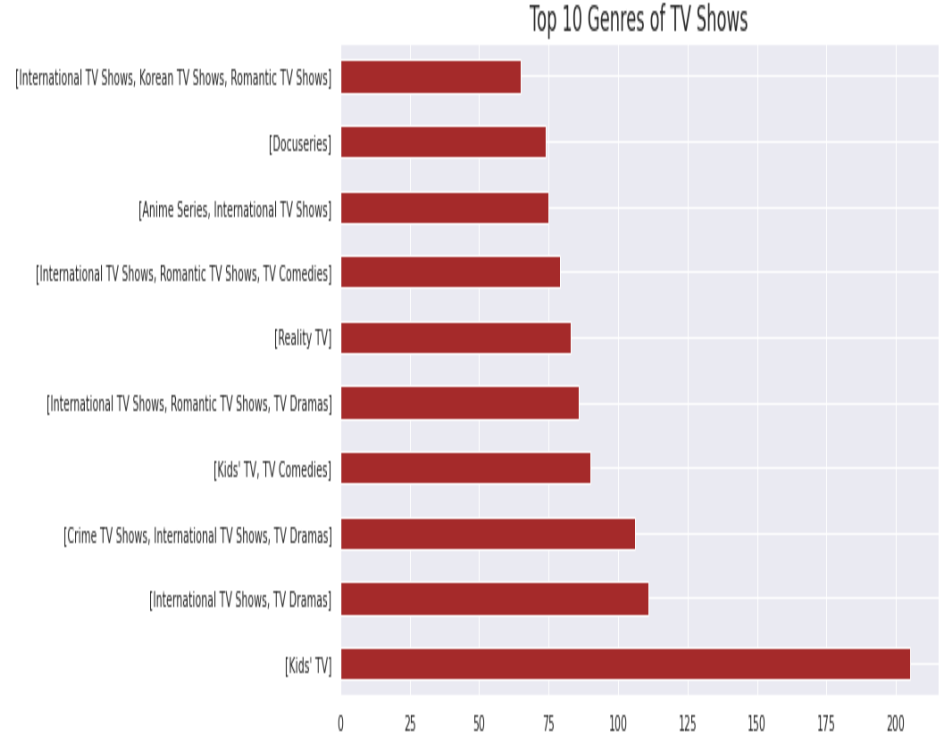


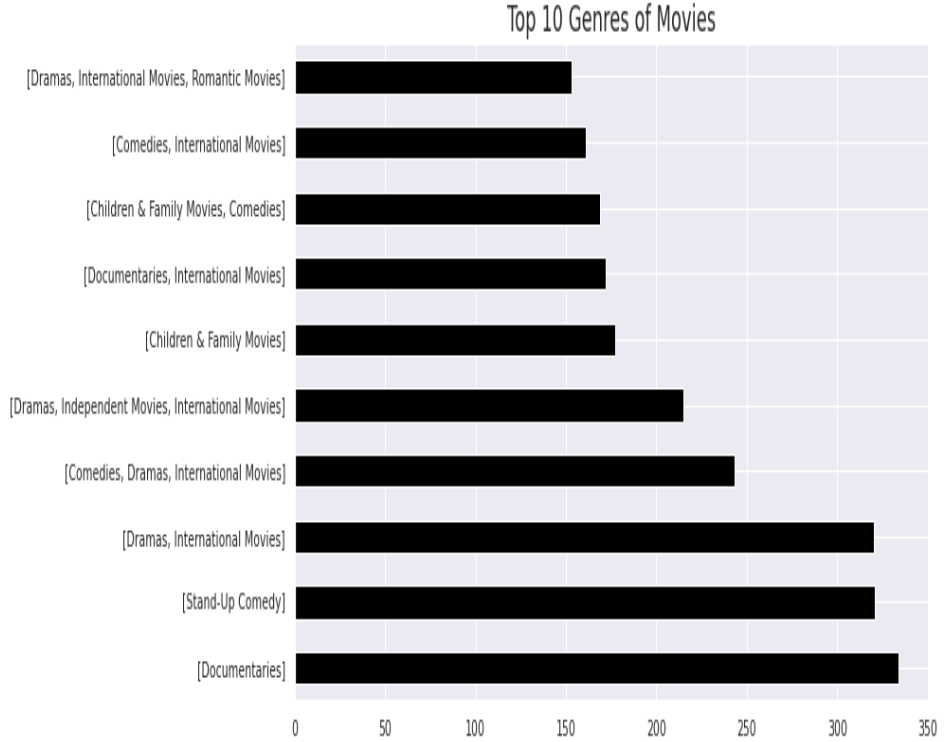
EDA (continued)





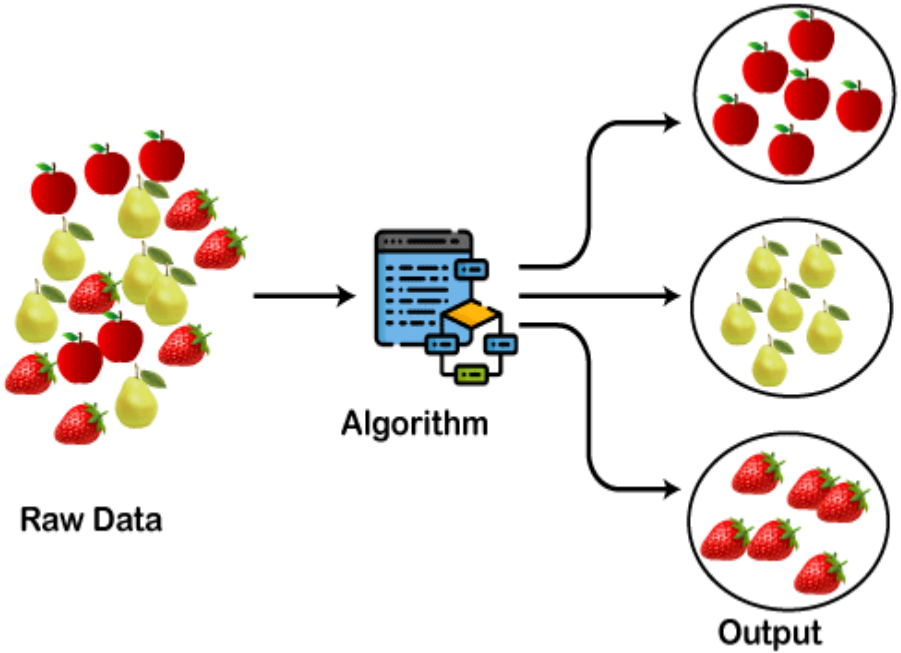
TOP 10 GENRES





Cluster Model Building

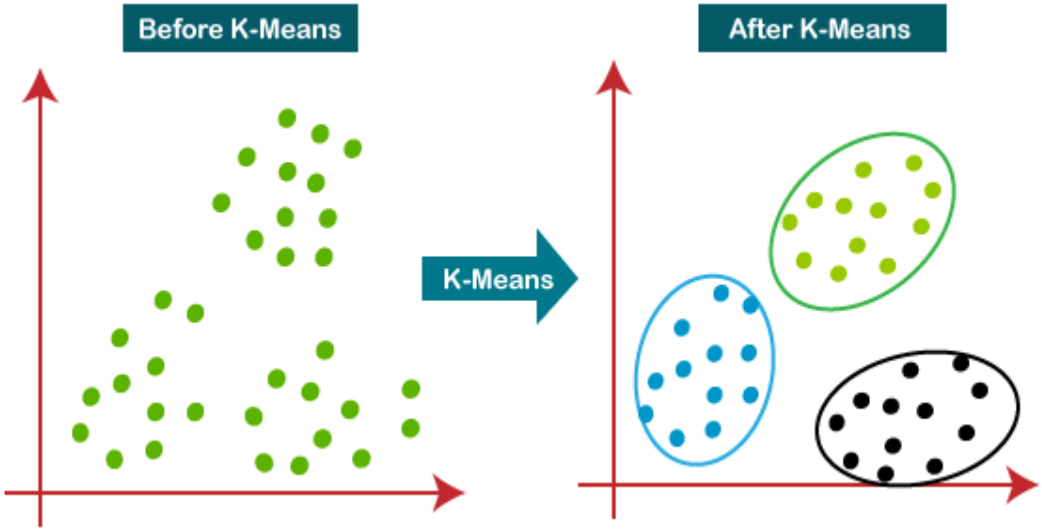
- **Clustering :**
- **Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.**
- **In simple words, the aim is to segregate groups with similar traits and assign them into clusters.**



K-Means Clustering:

The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.



Advantages

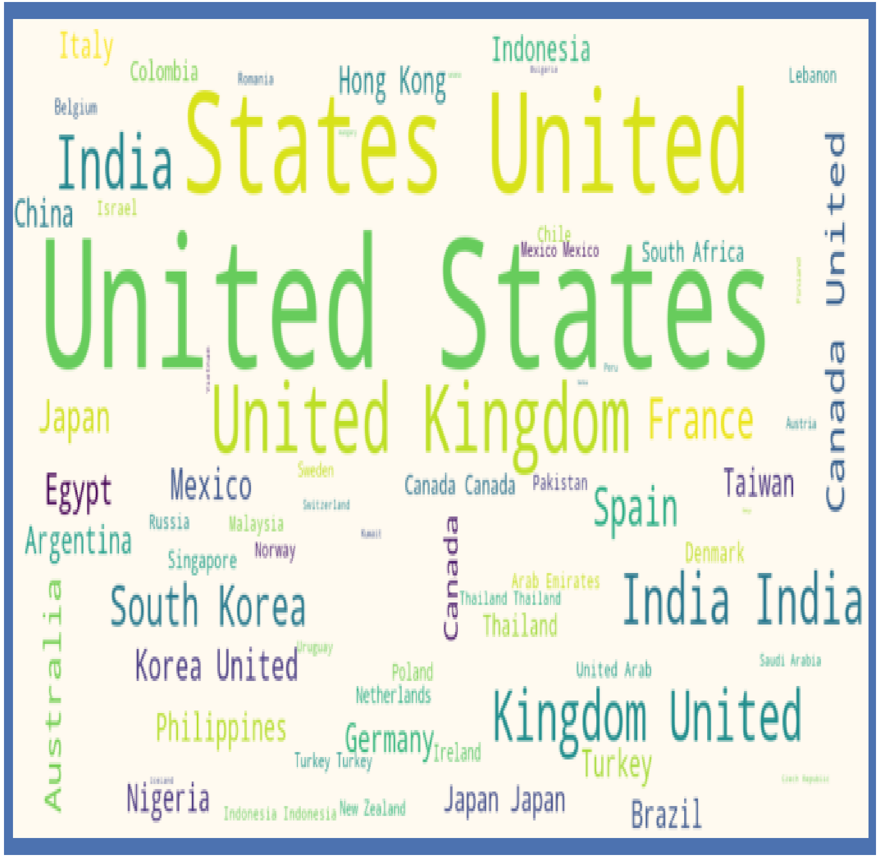
Recommendation engines

Market segmentation

Social network analysis

WORD CLOUDS

travel
meet
lead
case
couple
student
son
find
parent
trying
daughter
live
take
different
local
street
show
husband
gang
child
gets
sister
heart
go
dark
everything
must
evil
return
drama
way
head
pal
year
follow
help
behind
documentary
big
school
year old
sex
change
save
tale
marriage
comic
night
long
journey
move
people
end
form
mother
thing
come
music
life
new
turn
comedy
murder
career
adventure
film
now
force
keep
will
share
ex
determined
one
police
friendship
friend
play
power
crime
living
powerful
college
mysterious
children
member
look
seek
girl
boy
goes
day
man
begin
mission
human
whose
learn
history
leave
India
join
young man
fight
cop
home
wife
dream
teen
two
business
high
school
past
fall
use
story
Christmas
culture
brother
detective
crew
discover
animal
set
father
become
three
hit
town
American
special
team
together
drug
personal
investigate
four
romance
comedian
host
game
face
soon
put
young
family
back
forced
job
bring
stories
event
former
star
women
future
girlfriend
kid
land
struggle
relationship
navigate
plan
city
death
make
Earth
wealthy
time
may
start
search
old
work
bond
challenge
run
revenge
group
battle
attempt
dangerous
try
explore
love
woman
world
five
first
young woman
secret
series
tries
doctor
writer
village
struggling
best friend
New York
criminal



Italy
Colombia
Romania
Hong Kong
Indonesia
Lebanon
Belgium
India
States
United
China
Israel
Chile
Mexico Mexico
South Africa
United
United States
Canada
Japan
United Kingdom
France
Austria
Sweden
Egypt
Mexico
Canada Canada
Pakistan
Taiwan
Russia
Malaysia
Spain
Argentina
Singapore
Norway
Denmark
Canada
Arab Emirates
Thailand Thailand
Thailand
India India
South Korea
Australia
Korea United
Uruguay
Poland
United Arab
Saudi Arabia
Netherlands
Philippines
Germany
Ireland
Kingdom United
Turkey Turkey
Turkey
Nigeria
Indonesia Indonesia
New Zealand
Japan Japan
Brazil

CONCLUSION

- 1) The largest count of movies are made with the 'TV-MA' rating. "TV-MA" is a rating assigned by the TV Parental Guidelines to a television program that was designed for mature audiences only.
- 2) Highest number of movies released in Netflix was in the year 2017 and 2018
- 3) Highest number of Tv shows released in Netflix was in the years 2018, 2019, and 2020.
- 4) The months of October, November, December and January had the largest number of films and television series released.
- 5) In the given years we have seen in the month of Feb, May less movies was released so producer can release on this months
- 6) There are about 70% movies and 30% TV shows on Netflix.

7) The United States has the highest number of content on Netflix by a huge margin followed by India.

8) Raul Campos and Jan Sulter collectively have directed the most content on Netflix.

9) James, Lee, Michael, Daniel, John, etc have casted in most of the content on Netflix.

10) **Anupam Kher** has acted in the highest number of films on Netflix. Drama is the most popular genre followed by comedy.

11) More of the content is released in holiday season - October, November, December and January.

12) The number of releases have significantly increased after 2015 and have dropped in 2021 because of Covid 19.

13) The most frequent words used in description are LIFE, FAMILY, WOMEN, FATHER, LOVE, FRIEND, etc.

THANK YOU