STAT - 650
PROJECT - 1

Introduction and Objective

Purpose

Netflix, is one of the world's leading streaming platforms and it provides a vast variety of TV shows and movies, which is helping users to find content that aligns with their taste and preference, and it is very much important for user satisfaction and retention. The main aim of this project is to build a movie recommendation system using a dataset that contains various Netflix titles, which can be used to make the user experience better by suggesting them the relevant content.

Primary Objective:

- The primary objective is to conduct good exploratory data analysis of the Netflix movies, understanding the patterns, seeing the trends and insights that can make our recommendation system better.

- To understand the key features and attributes that influence the viewing habits of the user and user preferences

- To develop a basic movie recommendation system according to the content similarity and leveraging insights which we gained from our EDA

- To evaluate the effectiveness of our recommendation system and propose if there are any potential future improvements

- To analyse the temporal trends in content creation and addition to the platform, including how genre, ratings and content type have evolved over time.

By understanding the data and achieving these objectives, our aim is to create a data - driven approach in the field of content recommendation that can potentially improve the user satisfaction and engagement on netflix. This project will showcase the application of statistical analysis and machine learning techniques to take care and solve real-world problems in the streaming industry , furthermore by using the insights from comprehensive analysis, we can create more personalised recommendation algorithms which adapts to evolving user experience but also gives valuable insights on content acquisition and production strategies, which ultimately contributes to the platforms competitive edge in the rapidly growing steaming market

Dataset Selection and Description

Dataset Selection

The dataset which I chose for this project is the Netflix Movies and TV Shows dataset, which is needed to create the movie and TV show recommendation system, this dataset contains information about the content which is available on netflix.

The dataset is publicly available and has been downloaded from Kaggle, which is a famous platform for machine learning and data science projects. It was created to list out all the shows and movies that are available on netflix

Dataset Description

The dataset contains information on over 8,800 unique titles, which includes both TV shows and movies available on netflix, with so many entries we can gain rich foundation for our analysis and system development

Variable

The dataset which I chose has a diverse range of variables some of the key variables are:

Type: Indicates whether it is a TV show or a movie

- **Title:** Describes the name of the movie or the TV show
- **Director**: Describes the information about the director of the movie and TV show
- **Cast**: contains the list of all the actors who acted in the movie or the TV show
- **Rating**: Describes the information of the age certification of the movie (e.g , PG-13, TV-MA)
- **Listed_in**: Describes the information on the genre of the movie

Quantitative Variables:

- **Release_Year:** Describes the information on when the movie or the show was released
- **Duration:** Tells us about how long the movie or the show is.
- **IMDB Scores**: Describes on how much is the content rated on IMDB
- **IMDB Votes**: Describes the number of votes the content has received
- **TMDB popularity**: Describes the popularity of the content on TMDB

The rich dataset provides us with good opportunity to conduct in-depth analysis and development of a good movie recommendation system, the combination of both quantitative and categorical variables will allow for diverse analytical approaches which includes content based filtering and collaborative filtering

Data Pre-Processing

Import the data:

```
import pandas as pd
import numpy as np

# Load the dataset
netflix_df = pd.read_csv('netflix_titles.csv')
```

Data Cleaning:

Handling Missing Values:

1)

```
print(netflix_df.isnull().sum())
```

```
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
description        0
dtype: int64
```

2)

```python
import pandas as pd
import numpy as np

# Load the dataset
netflix_df = pd.read_csv('netflix_titles.csv')

# Handle missing values in categorical columns
categorical_cols = ['type', 'title', 'director', 'cast', 'country', 'rating', 'listed_in', 'description']
for col in categorical_cols:
    netflix_df[col] = netflix_df[col].fillna('Unknown')

# Handle missing values in date_added column
netflix_df['date_added'] = pd.to_datetime(netflix_df['date_added'], errors='coerce')
netflix_df = netflix_df.dropna(subset=['date_added'])

# Handle missing values in release_year (numeric column)
netflix_df['release_year'] = netflix_df['release_year'].fillna(netflix_df['release_year'].median())

# Handle duration column (mixed numeric and categorical)
netflix_df[['duration_value', 'duration_unit']] = netflix_df['duration'].str.split(' ', n=1, expand=True)
netflix_df['duration_value'] = pd.to_numeric(netflix_df['duration_value'], errors='coerce')
netflix_df['duration_value'] = netflix_df['duration_value'].fillna(netflix_df['duration_value'].median())
netflix_df['duration_unit'] = netflix_df['duration_unit'].fillna(netflix_df['duration_unit'].mode()[0])

# Reconstruct the duration column
netflix_df['duration'] = netflix_df['duration_value'].astype(str) + ' ' + netflix_df['duration_unit']

# Drop temporary columns
netflix_df = netflix_df.drop(columns=['duration_value', 'duration_unit'])

# Verify the changes
print(netflix_df.isnull().sum())
```

3)

```
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
description     0
dtype: int64
```

Remove Duplicates:

```python
# Check for duplicates
print(netflix_df.duplicated().sum())

# Remove duplicates if any
netflix_df = netflix_df.drop_duplicates()
```

```
0
```

Correct Any Inconsistencies or errors in data:

```
# Convert text to lowercase
text_columns = ['title', 'director', 'cast', 'country', 'listed_in', 'description']
for col in text_columns:
    netflix_df[col] = netflix_df[col].str.lower()

# Trim whitespace
netflix_df = netflix_df.apply(lambda x: x.str.strip() if x.dtype == "object" else x)
```

```
# Convert 'release_year' to integer
netflix_df['release_year'] = pd.to_numeric(netflix_df['release_year'], errors='coerce').astype('Int64')

# Convert 'date_added' to datetime
netflix_df['date_added'] = pd.to_datetime(netflix_df['date_added'], errors='coerce')
```

```
# Split duration into value and unit
netflix_df[['duration_value', 'duration_unit']] = netflix_df['duration'].str.split(' ', expand=True)
netflix_df['duration_value'] = pd.to_numeric(netflix_df['duration_value'], errors='coerce')

# Convert movie durations to minutes if in hours
mask = netflix_df['duration_unit'] == 'h'
netflix_df.loc[mask, 'duration_value'] *= 60
netflix_df.loc[mask, 'duration_unit'] = 'min'
```

```
# Split 'listed_in' into separate genres
netflix_df['genres'] = netflix_df['listed_in'].str.split(',')

# If you want to clean up the genre names (remove leading/trailing whitespace)
netflix_df['genres'] = netflix_df['genres'].apply(lambda x: [genre.strip() for genre in x] if isinstance(x, list) else x)

# Print the first few rows to verify the changes
print(netflix_df[['title', 'listed_in', 'genres']].head())
```

Conclusion:

After performing these steps we can conclude that all the missing values which were present were handled successfully by imputing and removing, also all kinds of duplicate values were removed if there were any, and there are no kinds of errors and inconsistencies in the data. We can now begin the next step which is exploratory data analysis

Exploratory Data Analysis (EDA)

Univariate Analysis:

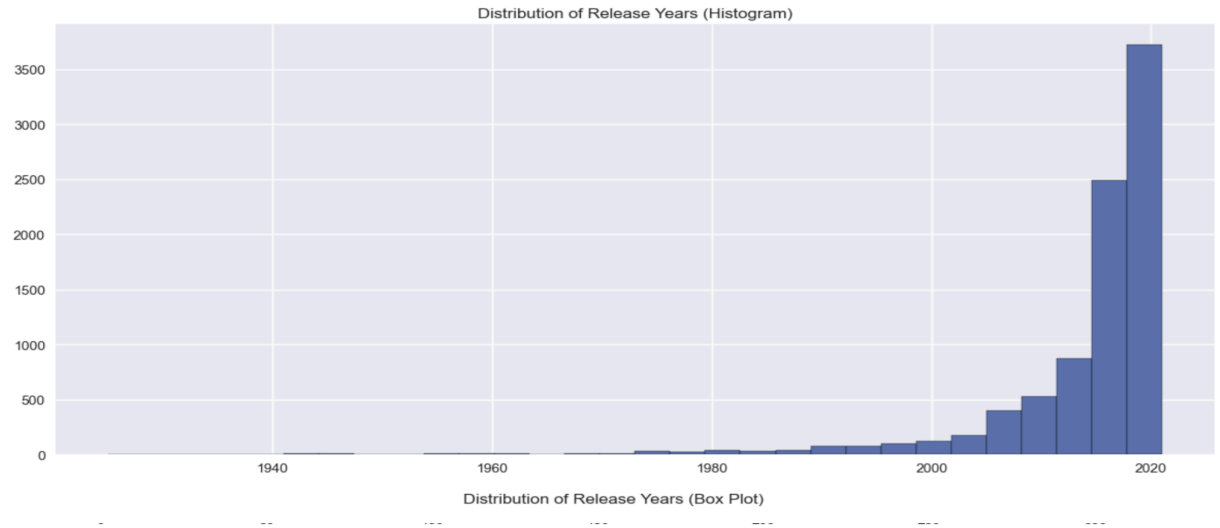
Summary: For my dataset my main focus will be on two categories “release_year” and “duartion_value”, these will be my key indicators

After getting to know about the summary we found out the following details:

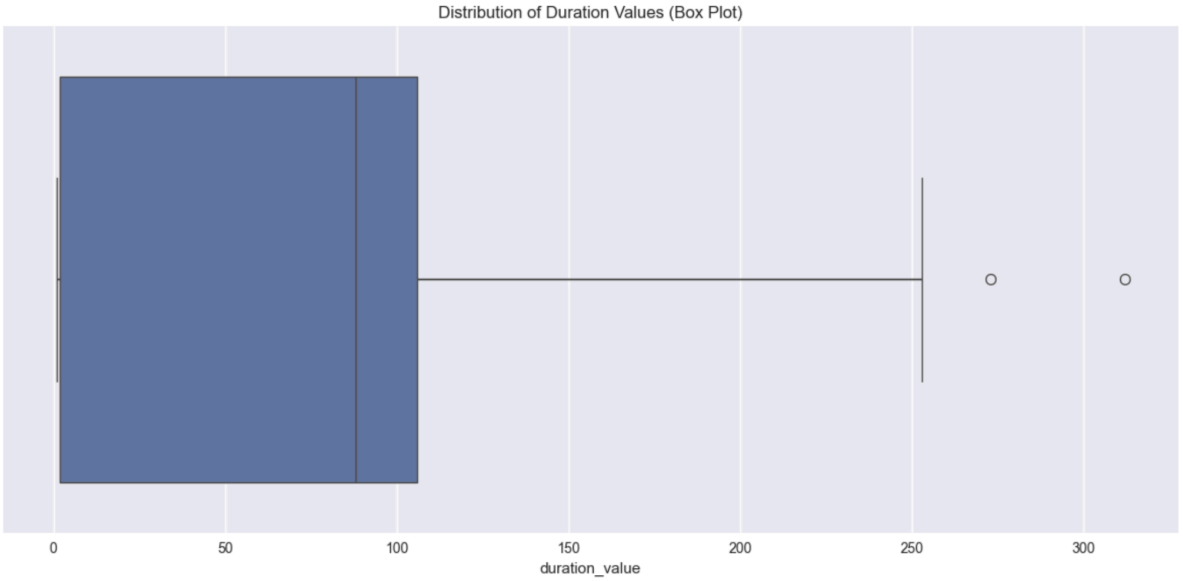
```
Summary Statistics for release_year:
Mean: 2014.18
Median: 2017.00
Mode: 2018.00
Standard Deviation: 8.82
Minimum: 1925.00
Maximum: 2021.00
25th Percentile: 2013.00
75th Percentile: 2019.00
Skewness: -3.45
Kurtosis: 16.23

Summary Statistics for duration_value:
Mean: 69.85
Median: 88.00
Mode: 1.00
Standard Deviation: 50.81
Minimum: 1.00
Maximum: 312.00
25th Percentile: 2.00
75th Percentile: 106.00
Skewness: -0.19
Kurtosis: -1.08
```

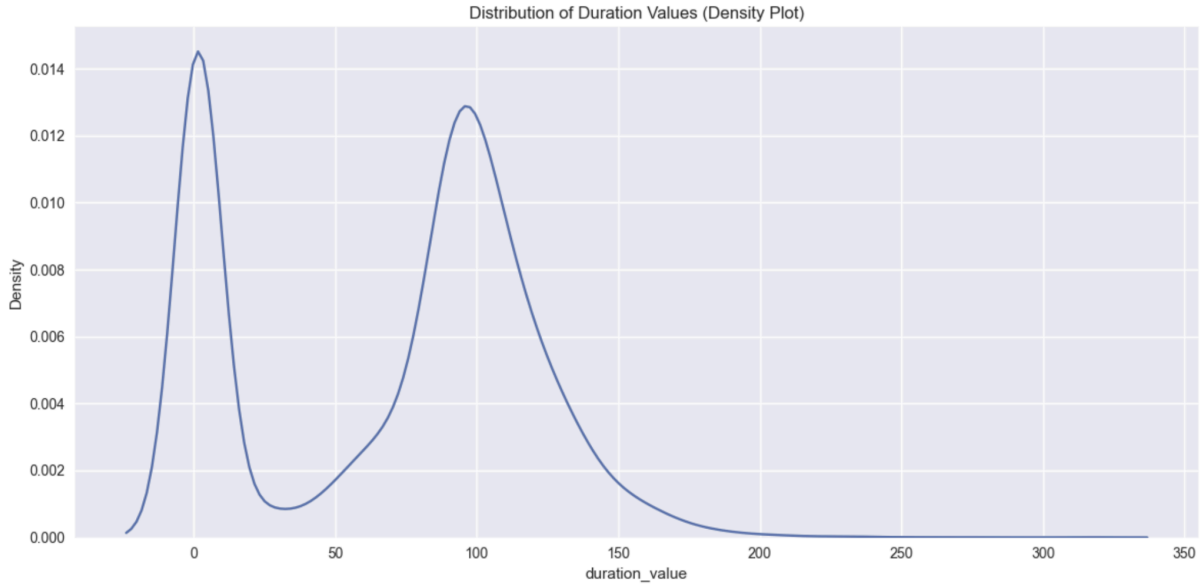
```
Top 5 categories in type:
type
movie      6131
tv show    2676
Name: count, dtype: int64

Top 5 categories in rating:
rating
TV-MA    3207
TV-14    2160
TV-PG     863
R         799
PG-13     490
Name: count, dtype: int64

Top 5 categories in country:
country
united states     2818
india              972
nan                831
united kingdom     419
japan              245
Name: count, dtype: int64
```

Visual Plot Examples:



Distribution of Release Years (Box Plot)





Conclusion:

From the above experiment we found out the summary statistics of all the columns present in the data set, which included Count of non-null values, Unique Values, Frequency of top values, mean, minimum, maximum and standard deviation of the numeric columns and most frequent values,

Key insights from summary:

- The dataset has a total of 8,807 entries
- Release year ranges from 1925 to 2021, with a mean of 2014 and a median of 2017.
- The duration value ranges from 1 to 312, with a mean of 69.85 and a median of 88.

The statistics revealed that:

The distribution of release year is negatively skewed (-3.45) with a high kurtosis (16.23), which indicates a concentration of more new releases.

The duration value distribution is slightly negatively skewed (-0.19) with low kurtosis (-1.08), suggesting flat distribution

Categorical Variable Summaries:

The analysis we did told us about top 5 categories for each categorical value:

- Tv shows are less common compared to movies
- ‘TV-MA’ is the most common rating
- The two content-producing countries are “United States” and “India”

- International movies and drama are the most watched genres.

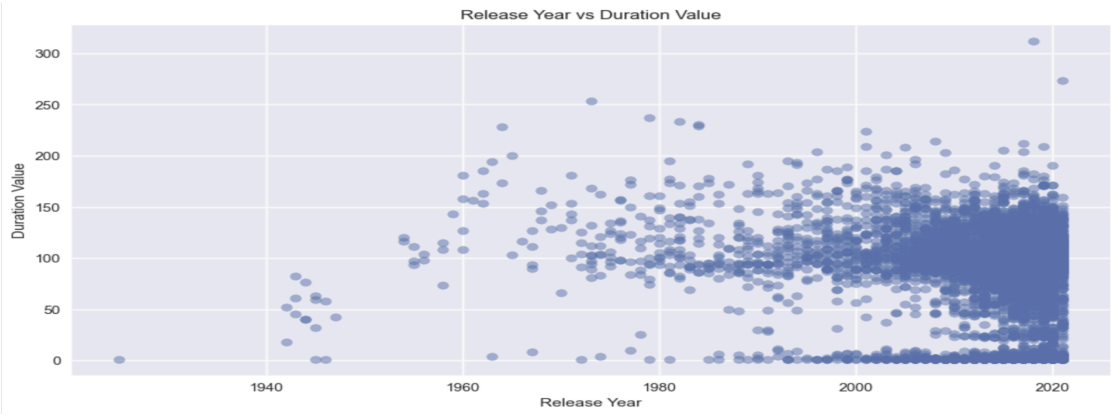
The univariate analysis steps showcase a comprehensive overview of each dataset and give us insights on distribution of content types, durations, ratings, release years and countries. The combination of visualisation and summary statistics allows us to gain through understanding of the individual variables, by setting a base for further analysis and development of movie recommendation systems.

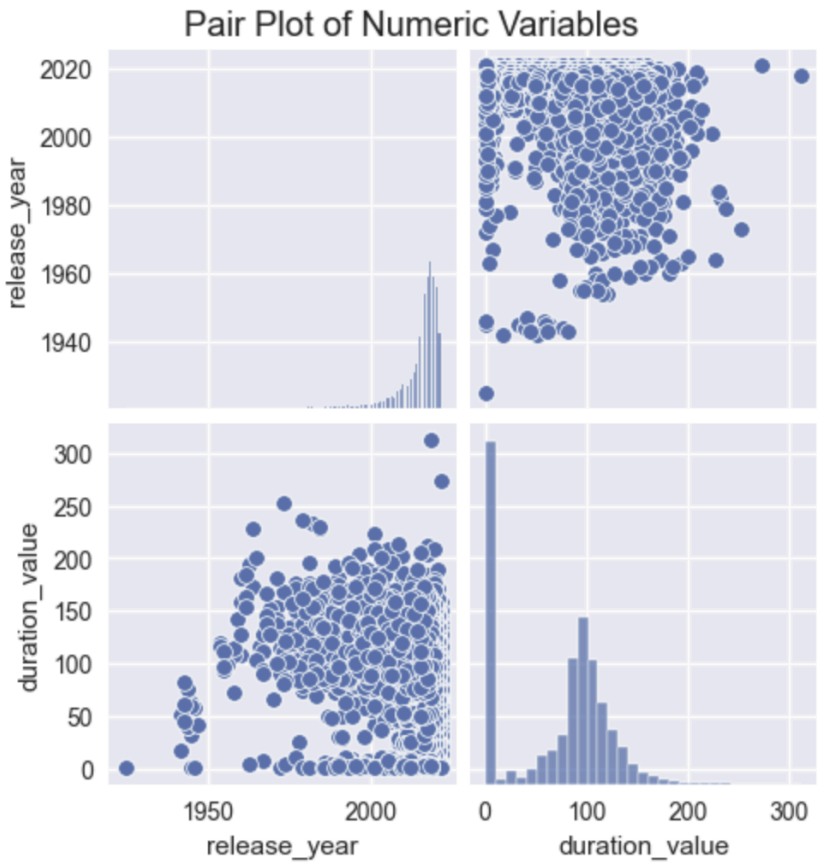
Bivariate Analysis:

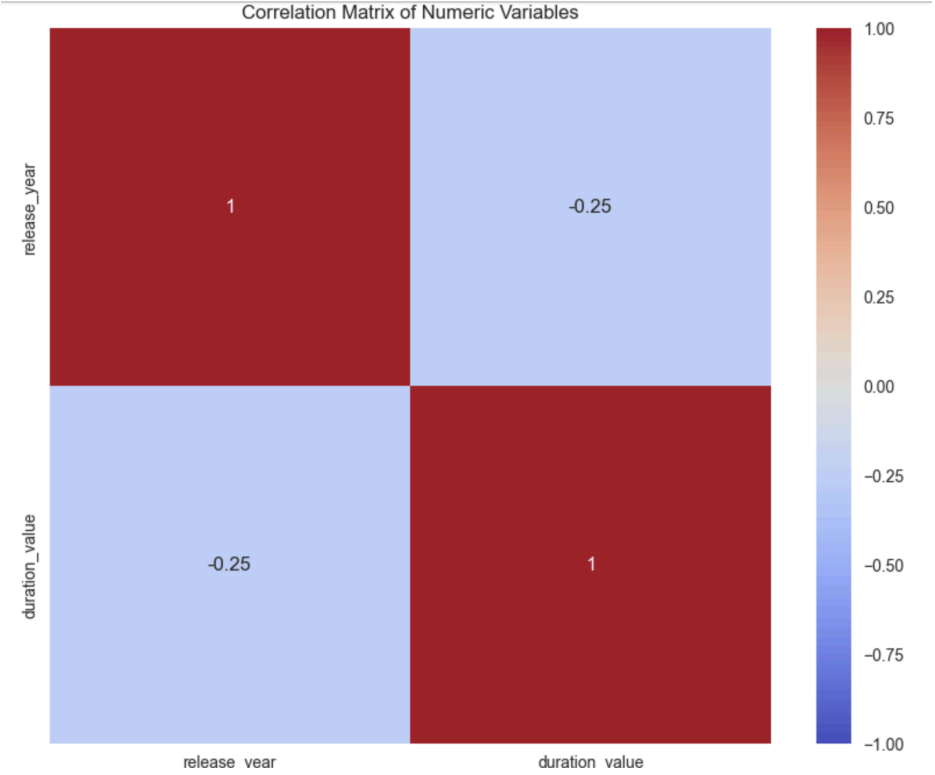
To showcase the Bivariate analysis by focusing on analysing relationships between two variables using various diagrams, In my dataset I chose the two variables to be “Release year” and “Duration”, I chose this because it will help me get an understanding of how the duration of movies and TV shows has been changing over the past few years, one more category which is chose is “Release Year” vs “Type”, to understand the trend of how the shows and movies have been performing across all years

To showcase this trend I implemented many figures to properly understand the data and do the further analysis to create a good movie recommendation system

Examples:




Pair Plot of Numeric Variables
release_year
duration_value
2020
2000
1980
1960
1940
300
250
200
150
100
50
0
1950
2000
0
100
200
300
release_year
duration_value


Correlation Matrix of Numeric Variables
release_year
duration_value
1
-0.25
-0.25
1
release_year
duration_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Multivariate Analysis:

I conducted Multivariate analysis on multiple variables, below I have listed my findings after conducting this analysis and also few examples of how the plots look like

Correlation Analysis:

- I could see that there is a moderate negative correlation between 'duration_value' and 'release_year' (-0.19), which showcases that newer content trends have slightly shorter duration.

- 'Type_encoded' (1 for movies and 0 for TV shows), this showcases a positive correlation with 'duration_value' (0.72), indicating that movies generally have longer duration that TV shows

- 'Rating_encoded' has weak correlation with other variables, showing that the duration and the release are not much dependent on content rating.

Pairwise Relationships:

- The plot of 'release_year' is heavily skewed towards recent years, with a good increase in content from the year 2015 onwards.

- 'Duration_value' showcases bimodal distribution, which proves that the movie duration is between (90-120 minutes) and TV show episode duration is (20-60 minutes).

PCA Analysis: The cumulative explained variance plot showcases a small number of principal components, and can explain a large amount of variance in the data, this suggests that there is a significant correlation between the variables, and the dimensionality of the dataset can be reduced without losing too much information.

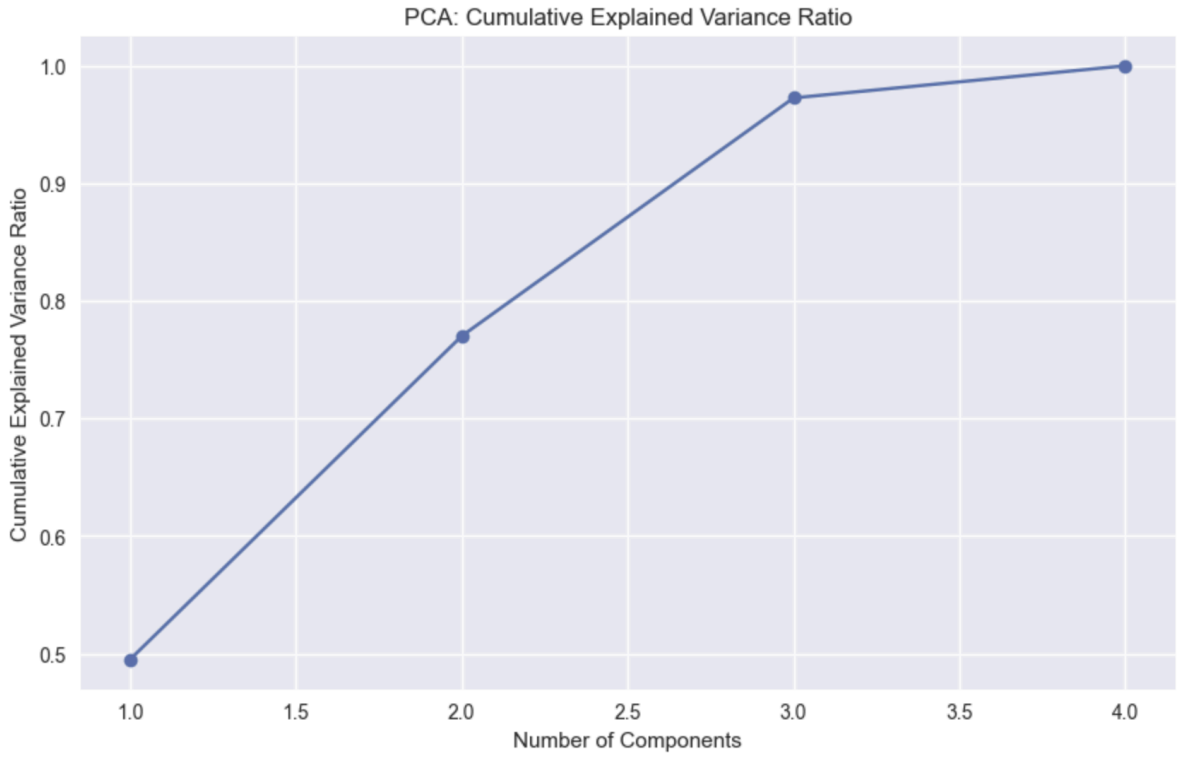
Some findings:

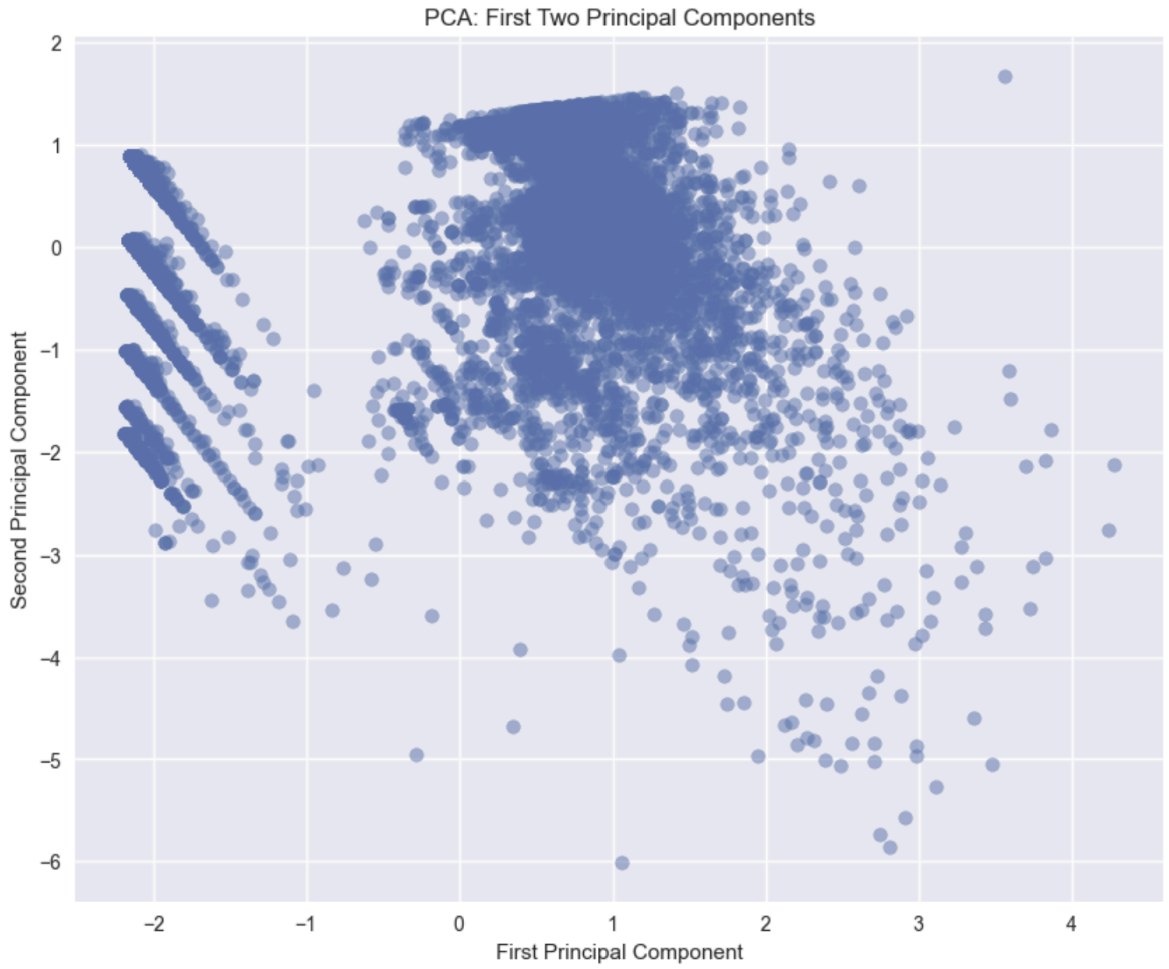
- The spread of points along PC1 should represent the variation in release years or durations

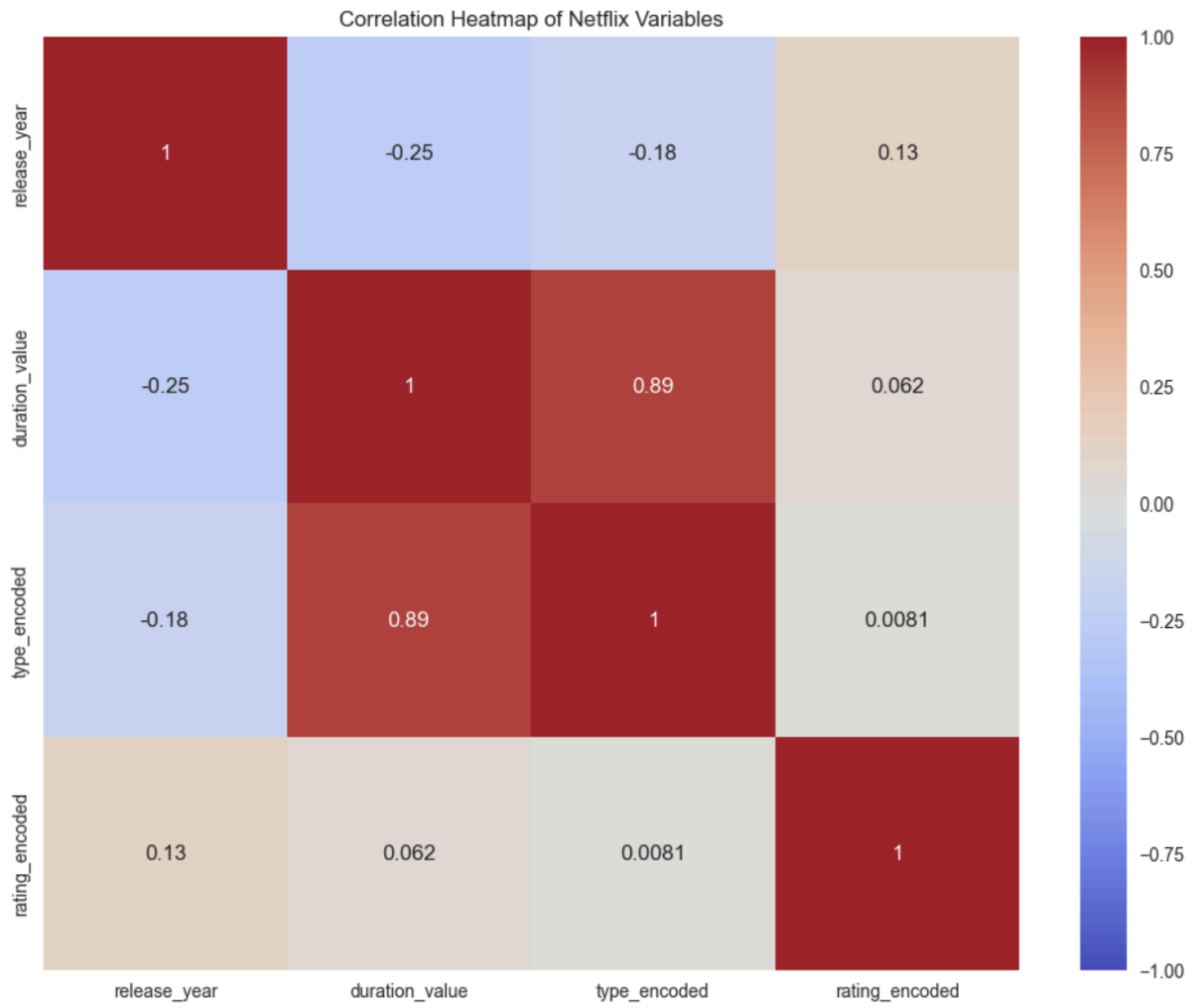
- Any kind of outliers present in the plots could represent a unique content in the netflix catalog

- There might be two main clusters, corresponding to our two main variables which are "Movies" and "Tv shows"

This multivariate analysis reveals complex relationships between duration, content type, release year and rating in the netflix catalog. It showcases how the platform is recently focused on mature content, with a mix of shows and movies catering to a diverse international audience. These insights could be valuable for developing a good recommendation system that takes care of multiple factors when suggesting content to users




PCA: First Two Principal Components
Second Principal Component
First Principal Component
2
1
0
−1
−2
−3
−4
−5
−6
−2
−1
0
1
2
3
4



Python Implementation:

Function Usage:

Pandas:

- The code mostly uses pandas for data analysis and manipulation
- Dataframe consists of operations like .describe(), .info(), and .value_counts(), which is useful in providing the summary of the data
- The pivot_table() function gives us a idea about reshaping capabilities and data aggregation

- Groupby operations in pandas is helpful in showcasing pandas ability to split-apply-combine data much efficiently

Numpy:

- Numpy is a function that is used for numerical computing on arrays.
- It consists of functions like np.percentile(), np.std(), np.mean() and np.median() which are used to calculate various statistical measures
- The .to_numpy() method is something which was used in the code and it is used to convert series to NumPy arrays for efficient numerical operations

Matplotlib:

- This is the main foundation of creating visualisations
- This was used to create all the visualisations in the code above, be it histogram, scatterplot, box plot etc. the function matplotlib_demo() shows us how to create basic plots
- This is also used in setting the size of the figure, tables and titles and is helpful in customising the plots

Seaborn:

- Seaborn is something that is built on matplotlib and is used to showcase advanced statistical visualisations.
- The boxplot in seaborn_demo() shows on how the data is distributed across various categories
- The visuals such as heatmap are used to correlate between numeric variables and demonstrate seaborn's statistical plotting capabilities.

Results and implementation

Summary of Findings:

Content Distribution:

- The dataset contains 8,807 titles which is divided into 6,131 movies (69.6%) and 2,676 TV shows (30.4%)

- Release of the content varies from 1925 to 2021, with a mean of 2014 and median of 2017

Duration:

- The average duration of the content is 69.85 minutes and it has a median of 88 minutes.

- There is wide range of durations of content with minimum being 1 minute of content to 312 minutes of content

- TV shows are mostly measured on the basis of seasons, with season 1 being the most common season (1,792) shows.

Ratings:

- The most common rating is TV-MA (3207 titles), followed by TV-14 (2,160 titles).

- From this we found out that customers are more inclined towards watching mature content

Content Origin:

- United States is the primary source of content in netflix with releasing over (2818 titles) and then they are followed by india who released over (972 titles) until now

- There is also a significant international representation with 749 unique countries listed in the content market

Genres:

Dramas and International movies are the most common genres among all the other present genres

- Following to these genres we also have documentaries and stand-up comedies being featured prominently

Release Year Distribution:

- The distribution of release years is highly peaked and heavily skewed towards recent years

- Also we could notice a sharp increase in content release from 2015.

Insights

Content Strategy: Netflix appears to be more focused on mature content with a strong emphasis on movies. This gives us an idea to focus more on adult audiences and staying current with new releases

Global Research: The increase in international content, especially from India, showcases Netflix's global expansion strategy and effort to make their content available for all the regions.

Content Length: The wide range of durations, mostly for movies, shows Netflix's flexibility in content format and potentially taking care of different viewing habits

Genere Focus: The impact of international movies and dramas suggests that these genres are the most popular ones among Netflix users, the platform might be using this data to make good production decisions and informed content acquisition.

Recent Content Emphasis: The skew which was seen in the graph shows that Netflix is constantly updating their library with new content, and this could be a key factor in retaining subscribers

Mature Content preference: The viewing of content which is rated TV-MA and TV-14 suggests that Netflix's core audience may be older teens or adults

Limitations:

Time Sensitivity: The dataset only has data till the year 2021, so it might not tell us about most recent trends on the platform

Lack of Viewing Data: The dataset does not have much information on popularity and viewership, which could be good to provide more insights into content performance

Genre Classification: The 'listed_in' column has information on multiple genres, making it difficult for genre-based analysis

Content Changes: Netflix's catalog changes very frequently, so this static dataset might not be that accurate and represent current offering.