**Data Analysis and Classification of Netflix Movie Dataset**

Name

Institution

Professor

Date

**Breakdown of this Notebook:**

Importing Libraries Loading dataset Data Cleaning:

* Deleting redundant columns.
* Dropping duplicates.
* Cleaning individual columns.
* Remove the NaN values from the dataset
* Some Transformations

**Data Visualization**: Using plots to find relations between the features. Type: Movie and TV Shows Rating Relation between Type and Rating Word Cloud genre stars Classification Analysis This dataset contains information concerning TV Shows and Movies added to the Netflix catalog, including:

**General information**: id, title, type (TV Show or Movie), stars and a brief description. Date fields: When the show was released and when it was added to the catalog. Categorization: Rating and votes in which the show is listed.

**Introduction**

Netflix! What is believed to have started in 1997 as a DVD rental service has since exploded into the largest entertainment/media company by market capitalization, boasting over 200 million subscribers as of January 2021. It needs no introduction; it is a streaming service that allows members to watch a wide variety of contents like movies, TV shows, documentaries and more. Netflix allows us to enjoy unlimited ad-free viewing of the contents. It was founded in 1997 and has become $100 billion company in 20 years. The number of movies has decreased by over 2,000 titles since 2003 while the number of TV shows has nearly tripled. Some important variables will be analyzed to perform exploratory data analysis and also predict the number of Movies and TV shows that will be added in 2020.

Given the large number of movies and series available on the platform, it is a perfect opportunity to flex our data manipulation skills and dive into the entertainment industry. Our friend has also been brushing up on their Python skills and has taken a first crack at a CSV file containing Netflix data. For their first order of business, they have been performing some analyses, and they believe that the average duration of movies has been declining.

If we're going to be working with this data, we know a good place to start would be to probably start working with pandas.

**About the Dataset**

**Context**

This data is all about Movies That are available on Netflix Website movies title, stars, ,desc of movies, duration, rating on IMDB, voted by people, year, genre, certificate

**Source**

This dataset originates from the IMDB website data and is collected using web scraping

**Importing libraries and Loading the dataset**

Alright, we now have a pandas, the most common way to work with tabular data in Python. Now back to the task at hand. We want to follow up on our friend's assertion that movie lengths have been decreasing over time. A great place to start will be a visualization of the data. Given that the data is continuous, a line plot would be a good choice, with the dates represented along the x-axis and the average length in minutes along the y-axis. This will allow us to easily spot any trends in movie durations. There are many ways to visualize data in Python, but matploblib.pyplot is one of the most common packages to do so.

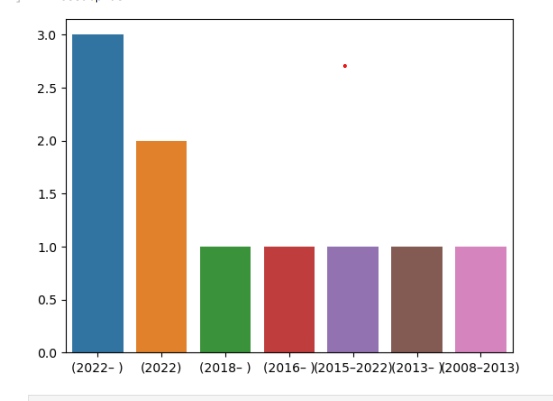
**Note:** In order for us to correctly test your plot, you will need to initialize a matplotlib.pyplot Figure object, which we have already provided in the cell below. You can continue to create your plot as you have learned in Intermediate Python.

**Statistical or Hypothetical questions**

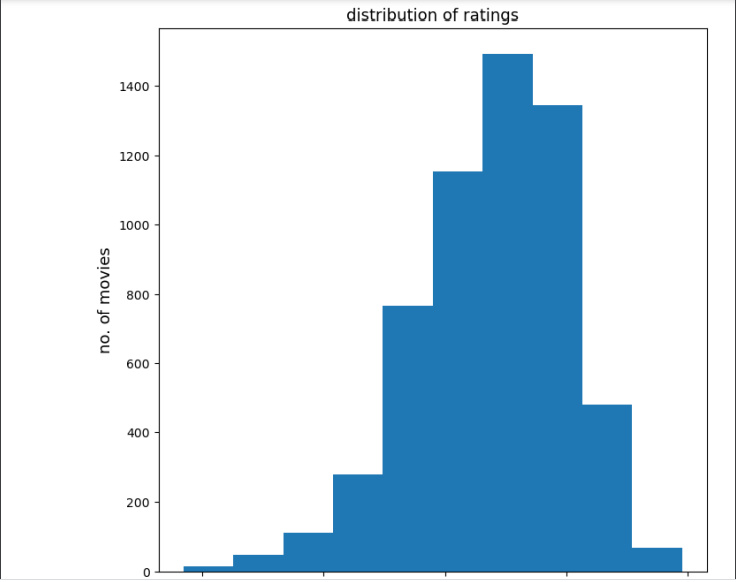
* What is the duration of the contents?
* What are the top categories or genre?
* What are the types of contents and their rating?

**Outcome of EDA**

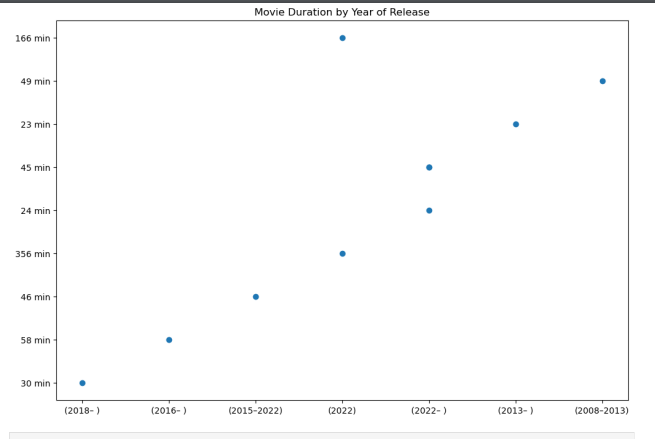
* Below shows content added over the years. The highest content was added in 2022



* Number of movies uploaded has a normal distribution against the ratings.



* The movie release against duration has a linear distribution.



* Below are what top categories mean:

**TV-MA:** This program is specifically designed to be viewed by adults and therefore may be unsuitable for children under 17.

**TV-14:** This program contains some material that many parents would find unsuitable for children under 14 years of age.

**TV-PG:** This program contains material that parents may find unsuitable for younger children.

* Based on the prediction, 2,014 Movies and 933 TV Shows will be added in 2020.

One variable that I felt missing in IMDB ratings value. It would be good to analyze the IMDB ratings of Netflix contents. I found it little challenging to plot hypothesis test with this data. Did some additional research about the variables and came up with CDF plot (Dinesh, 2020).

**Conclusion**

* The average ratings for the movies is 6.79 scores.
* The model used to train and predict has a prediction accuracy of 0.80.

The number of movies produced over the years had a high trend of increasing in now and during the start of 2022.

Well, as we suspected, non-typical genres such as children's movies and documentaries are all clustered around the bottom half of the plot. But we can't know for certain until we perform additional analyses. We've performed an exploratory analysis and classification of some entertainment data, and there are lots of fun ways to develop your skills as a Pythonic data scientist. These include learning how to analyze data further with statistics, creating more advanced visualizations, and perhaps most importantly, learning more advanced ways of working with data in pandas.

**Limitations**

Finally, we must add that all the above results are limited to the given database, adding that during the past two years and according to the COVID-19 pandemic, there has been a significant change in all production industries which naturally affects the film industry.

References:

Dinesh, M. (2020). Netflix Ranking by Combination of K-Nearest Neighbour and Singular Value Decomposition. *International Journal of Computational Science and Engineering*, *10*(1), 1-10.