**Documentation for Netflix Data Analysis Project**

**Table of Contents**

1. Introduction
2. Data Loading
3. Data Cleaning
4. Exploratory Data Analysis (EDA)
5. Filtering and Querying Data
6. Visualizations
7. Conclusion
8. Future Work

**1. Introduction**

This project analyzes a dataset of Netflix shows and movies. The objective is to understand various aspects of the content available on Netflix, including categories, genres, countries of production, and release dates.

**2. Data Loading**

The first step is to load the dataset into a pandas DataFrame. The dataset is in a CSV file format.

import pandas as pd

# Load the dataset

df = pd.read\_csv('path/to/netflix\_titles.csv')

df.head()

**3. Data Cleaning**

Data cleaning involves handling missing values, converting data types, and ensuring the dataset is ready for analysis.

* Handling missing values:

df.isnull().sum()

df.dropna(inplace=True) # Dropping rows with missing values

* Converting data types:

df['Date\_N'] = pd.to\_datetime(df['Date\_N'])

**4. Exploratory Data Analysis (EDA)**

EDA involves understanding the distribution of various features and identifying patterns or trends in the data.

* Summary statistics:

df.describe()

* Distribution of content by category:

df['Category'].value\_counts().plot(kind='bar')

**5. Filtering and Querying Data**

This section includes examples of filtering and querying the dataset to extract meaningful insights.

* Filtering by category and type:

df[(df['Category'] == 'Movie') & (df['Type'] == 'Drama')]

* Extracting kids' TV shows:

df[(df['Category'] == 'TV Show') & (df['Type'] == "Kids' TV")]

**6. Visualizations**

Visualizations help in understanding the data better through graphical representations.

* Bar chart of content categories:

import matplotlib.pyplot as plt

df['Category'].value\_counts().plot(kind='bar')

plt.title('Distribution of Content by Category')

plt.xlabel('Category')

plt.ylabel('Count')

plt.show()

* Time series of content release dates:

df['Release\_Date'].value\_counts().sort\_index().plot(kind='line')

plt.title('Number of Releases Over Time')

plt.xlabel('Year')

plt.ylabel('Number of Releases')

plt.show()

**>.These Functions have been used in this Project to analyse the Data**

\*head () - It shows the first N rows in the data (by default, N=5).

\* tail () - It shows the last N rows in the data (by default, N=5).

\* shape - It shows the total no. of rows and no. of columns of the dataframe.

\* size - To show No. of total values(elements) in the dataset.

\* columns - To show each Column Name.

\* dtypes - To show the data-type of each column.

\* info() - To show indexes, columns, data-types of each column, memory at once.

\* value\_counts - In a column, it shows all the unique values with their count. It can be applied on a single column only.

\* unique() - It shows the all unique values of the series.

\* nunique() - It shows the total no. of unique values in the series.

\* duplicated( ) - To check row wise and detect the Duplicate rows.

\* isnull( ) - To show where Null value is present.

\* dropna( ) - It drops the rows that contains all missing values.

\* isin( ) - To show all records including particular elements.

\* str.contains( ) - To get all records that contains a given string.

\* str.split( ) - It splits a column's string into different columns.

\* to\_datetime( ) - Converts the data-type of Date-Time Column into datetime[ns] datatype.

\* dt.year.value\_counts( ) - It counts the occurrence of all individual years in Time column.

\* groupby( ) - Groupby is used to split the data into groups based on some criteria.

\* sns.countplot(df['Col\_name']) - To show the count of all unique values of any column in the form of bar graph.

\* max( ), min( ) - It shows the maximum/minimum value of the series.

\* mean( ) - It shows the mean value of the series.

**TASKS**

* Task. 1) Is there any Duplicate Record in this dataset ? If yes, then remove the duplicate records.
* Task. 2) Is there any Null Value present in any column ? Show with Heat-map.

Q. 1) For 'House of Cards', what is the Show Id and Who is the Director of this show ?

Q. 2) In which year the highest number of the TV Shows & Movies were released ? Show with Bar Graph.

Q. 3) How many Movies & TV Shows are in the dataset ? Show with Bar Graph.

Q. 4) Show all the Movies that were released in year 2000.

Q. 5) Show only the Titles of all TV Shows that were released in India only.

Q. 6) Show Top 10 Directors, who gave the highest number of TV Shows & Movies to Netflix ?

Q. 7) Show all the Records, where "Category is Movie and Type is Comedies" or "Country is United Kingdom".

Q. 8) In how many movies/shows, Tom Cruise was cast ?

Q. 9) What are the different Ratings defined by Netflix ?

Q. 9.1) How many Movies got the 'TV-14' rating, in Canada ?

Q. 9.2) How many TV Shows got the 'R' rating, after year 2018 ?

Q. 10) What is the maximum duration of a Movie/Show on Netflix ?

Q. 11) Which individual country has the Highest No. of TV Shows ?

Q. 12) How can we sort the dataset by Year ?

Q. 13) Find all the instances where: Category is 'Movie' and Type is 'Dramas' or Category is 'TV Show' & Type is 'Kids' TV'.

**7. Conclusion**

In this analysis of the Netflix dataset, we explored various aspects of the content available on the platform. Our primary focus was on understanding the distribution of movies and TV shows, their genres, countries of origin, and release dates.