**Exploratory Data Analysis Report**

**Dataset – Netflix**

|  |  |
| --- | --- |
| TASK | Methods used |
| Show top 5 records of the dataset | df.head(5) |
| Show bottom 10 records of the dataset | df.tail(5) |
| Show no. of rows and columns | df.shape |
| Show no of total values(elements) in the dataset | df.size |
| Show the column names of the dataset | df.columns |
| Show the data type of the columns of the dataset | df.dtypes |
| Show the data type of the columns of the dataset | df.info() |
| TASK 1 | |
| To find duplicate records in dataset | df.duplicated() |
| To drop duplicate records in a dataset | df.drop\_duplicates() |
| TASK 2 | |
| To show null values | df1.isnull() |
| To count of null values | df1.isnull()  sum() |
| To create heatmap for null values | sns.heatmap(df.isnull()) |
| TASK 3 | |
| To show specified rows details | df[df[‘Title’]==[’House of Cards’][[‘Show\_Id’,’Director’]] |
| TASK 4 | |
| To Convert 'Release\_Date' to datetime and extract the year | to\_datetime()  dt.year |
| To Drop rows with missing year | dropna() |
| To Count number of Movies and TV Shows per year | groupby()  size()  unstack()  fillna(0) |
| To Find the year with the highest total releases | sum()  idxmax() |
| To Plot the bar chart | plot(kind='bar', figsize=(12,6), colormap='Pastel1') |
| TASK 5 | |
| To find total movies and TV shows in dataset | sum() |
| To Count number of Movies and TV Shows | value\_counts() |
| To Plot using seaborn countplot | sns.countplot(data=df1, x='Category', hue='Category', palette='Set2', legend=False) |
| TASK 6 | |
| To find Movies released in the year 2000 | df1[  (df1['Category'] == 'Movie') &  (df1['year'] ==2000)  ] |
| To find titles of TV Shows released only in India | unique() |
| To show the Top 10 directors who gave the highest number of TV shows and movies to Netflix. | value\_counts()  head(10) |
| To show all the records where “category is movie and type is comedies” or “county is united kingdom” | str.contains() |
| To find that, In how many movies/shows, Tom Cruise were cast? | astype()  str.contains() |
| Task 7 | |
| To find all ratings defined by Netflix? | unique() |
| To find distribution of content ratings in a graph | sns.countplot(  data=df1,  y='Rating',  hue='Rating',  palette='coolwarm',  order=df1['Rating'].value\_counts().index,  legend=False  ) |
| To find how many movies got “TV-14” rating in Canada? | df1[  (df1['Category'] == 'Movie') &  (df1['Rating'] == 'TV-14') &  (df1['Country'] == 'Canada')  ] |
| To show the ratings and no. of unique ratings | unique()-displays all ratings  nunique()- gives number of unique ratings |
| To find how many Tv Shows got rating “R” after year 2018? | df1[  (df1['Category'] == 'TV Show') &  (df1['Rating'] == 'R') &  (df1['year'] > 2018)  ] |
| TASK 8 | |
| To find the maximum duration of a movie/tv show in Netflix?  Also Show the lowest as well | str.extract(r'(\d+)\s\*(\w+)')-split season and minutes in duration  idxmin()  idxmax() |
| TASK 9 | |
| To find which country has the highest no. of .Tv Shows | value\_counts()  idxmax()  max() |
| To show all the countries and no .of tv shows in a graph as well | astype()  split()  strip()  explode()  value\_counts()  sns.barplot(  data=top\_countries,  x='TV\_Show\_Count',  y='Country',  hue='Country',  palette='rocket',  legend=False  ) |
| To compare each country tv show and movie count and show them in a graph as well | size()  unstack()  sum()  melt()  sns.barplot(data=melted, x='Country', y='Count', hue='Type', palette='Set2') |
| TASK 10 | |
| To sort by year | sort\_values() |
| To show the count of movies/tv shows released in a year in line chart | sort\_index()  sns.lineplot(x=yearly\_counts.index, y=yearly\_counts.values, marker='o', color='teal') |
| To show the Top 5 directors in the Netflix based on the count of movies/tv shows directed by them. | value\_counts().head(5)  sns.barplot(  data=top\_directors,  x='Count',  y='Director',  hue='Director', # Set hue to match y-axis  palette='viridis',  legend=False # Hide the redundant legend  ) |