Business Case: Netflix Intro

1)

```
#Importing Libraries
import numpy as np #helps in working on Matrices and Arrays
import pandas as pd # helps in to read the dataset/working on dataset/manipulating the dataset
import matplotlib.pyplot as plt
import seaborn as sns
```

2)

```
#Reading the Dataet
df = pd.read_csv("Dataset.csv")
df.head()
```

3)

```
#Help us to see the rows & columns in the dataset.
df.shape
```

4)

```
#Help us to see some basic stats of dataset,
#it is showing only in release_year column because only release_year have numerical or integer values
#and other columns have only string values.
df.describe()
```

5)

```
#Helps us to see the no. of columns as well as their datatype & non-null count,also shows the memory usage.
df.info()
```

Missing Values

6)

```
df.isna().sum()
```

Adjust Data Type and fill the missing values

Correcting and verifying the dataset make sense so that we can analyse the data in a good manner.

The following columns which does not require any changes as well as filling.

- show_id
- type
- title
- release_year
- listed_in
- description

The following columns which require changes.

- director
- cast
- country
- date_added
- rating
- duration

After seeing all the columns in the dataset, we have to update the data type of date_added from object/string to datetime.

First we update date_added to datetime and check

7)

```
# converting the datatype from object to datetime64
df['date_added'] = pd.to_datetime(df['date_added'])
df.head()
```

How to handle missing values?

We can handle missing values by fiiling 'Unavailable' in all nulls

8)

```
df.fillna({'director':'Unavailable', 'cast':'Unavailable', 'country':'Unavailable', 'rating':'Unavailable'},
inplace = True
df.isna().sum()
```

For nulls in date_added, missing date_added is to be replaced with the most recent date from date addded because Netflix has a tendency to addmore content over time

9)

```
#First we see the null values in date_added column
df[df.date_added.isnull()]
```

10)

```
#Replacing all null values with most recent date in date_added column
most_recent_date = df['date_added'].max()
df.fillna({'date_added': most_recent_date}, inplace = True)
df.head()
```

After executing the code above we can see all null values in date_added column is replaced with mmost recent date values.

Now we do data cleaning of duration column

11)

```
#First we see the null values in duration column
df[df.duration.isnull()]
```

12)

```
#First we see movies of the director "Louis C.K."
df[df.director == 'Louis C.K.']
```

13)

```
#First we replace the values of rating with duration and write 'Unavailable' in rating column because the rating
column
#is incorrect and it is come in this column by human mistake
#we use loc because it helps us easily to get the columns by name
df.loc[df['director'] == 'Louis C.K.', 'duration'] = df['rating']df[df.director == 'Louis C.K.']
```

14)

```
#Now we put 'Unavailable' in rating column
df.loc[df['director'] == 'Louis C.K.', 'rating'] = 'Unavailable'df[df.director == 'Louis C.K.']
```

Visualizations

Comparison of tv shows vs. movies.

15)

```
df.type.value_counts() #value_counts helps to show the count of different movies and Tv shows
```

16)

```
#countplot helps to show the count of each category
sns.countplot(x = 'type', data = df)
plt.title('Movies vs TV Shows')
```

As you can see the above countplot chart,Netflix has more movies as compared to Tv shows

Country Analysis

17)

```
df['country'].value_counts()
```

18)

```
plt.figure(figsize = (12,6))
sns.countplot(order = df['country'].value_counts().index[0:12], y = 'country', data = df)
plt.title('Country wise content on netflix')
```

19)

```
# Now checking the type of content based on country
movie_countries = df[df['type'] == 'Movie']
Tv_show_countries = df[df['type'] == 'TV Show']
```

20)

```
plt.figure(figsize = (12,6))
sns.countplot(y = 'country', order = df['country'].value_counts().index[0:10], data = movie_countries)
plt.title('Top 10 countries producing movies in Netflix')

plt.figure(figsize = (12,6))
sns.countplot(y = 'country', order = df['country'].value_counts().index[0:10], data = Tv_show_countries)
plt.title('Top 10 countries producing Tv shows in Netflix')
```

As you can see the above two charts Netflix has produce more Movies than Tv shows and United States produce most no. of Tv Shows and Movies for Netflix, India is second in this list

Ratings of shows in Netflix

21)

```
df.rating.value_counts()
```

22)

```
plt.figure(figsize = (20,6))
sns.countplot(x = 'rating', order = df['rating'].value_counts().index[0:], data = df)
plt.title('Rating of shows in Netflix')
```

As you can see the above chart it shows the Netflix has produced more content for mature viewers where as the Netflix produce second highest content for the age of 14 and above

Content release per year in Netflix

23)

```
df.release_year.value_counts()[0:20]
```

24)

```
plt.figure(figsize = (12,6))
sns.countplot(x = 'release_year', order = df.release_year.value_counts().index[0:20], data = df)
plt.title('Content release per year in Netflix')
```

As you can see Netflix has released more content in 2018 and second highest in 2017

Popular Genres Analysis

25)

```
plt.figure(figsize = (12,10))
sns.countplot(y = 'listed_in', order = df['listed_in'].value_counts().index[0:20], data = df)
plt.title('Most popular Genres on Netflix')
```

As you can see the most popular genre in Netflix is Dramas, International Movies after that Documentaries and so on

Thank you