Vaibhav kumar

Rollno 19

Day 4 _ DATA CLEANING YOUTUBE

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [2]:
```
ytube=pd.read_csv('D:\\vk\\TRIM 3\\ML\\DATASET\\top-5000-youtube-channels.csv')
```

In [3]:
```
ytube
```

Out[3]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
0	1st	A++	Zee TV	82757	18752951	20869786591
1	2nd	A++	T-Series	12661	61196302	47548839843
2	3rd	A++	Cocomelon - Nursery Rhymes	373	19238251	9793305082
3	4th	A++	SET India	27323	31180559	22675948293
4	5th	A++	WWE	36756	32852346	26273668433
...	...	...	...	...	...	...
4995	4,996th	B+	Uras Benlioğlu	706	2072942	441202795
4996	4,997th	B+	HI-TECH MUSIC LTD	797	1055091	377331722
4997	4,998th	B+	Mastersaint	110	3265735	311758426
4998	4,999th	B+	Bruce McIntosh	3475	32990	14563764
4999	5,000th	B+	SehatAQUA	254	21172	73312511

5000 rows × 6 columns

In [4]:
```
ytube.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Rank           5000 non-null   object
 1   Grade          5000 non-null   object
 2   Channel name   5000 non-null   object
 3   Video Uploads  5000 non-null   object
 4   Subscribers    5000 non-null   object
 5   Video views    5000 non-null   int64 
dtypes: int64(1), object(5)
memory usage: 234.5+ KB
```

In [5]:
```
ytube['Rank']=ytube['Rank'].str[0:-2].str.replace(',','').astype('int')
```

In [6]:
```
ytube
```

Out[6]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
0	1	A++	Zee TV	82757	18752951	20869786591
1	2	A++	T-Series	12661	61196302	47548839843
2	3	A++	Cocomelon - Nursery Rhymes	373	19238251	9793305082
3	4	A++	SET India	27323	31180559	22675948293
4	5	A++	WWE	36756	32852346	26273668433
...	...	...	...	...	...	...
4995	4996	B+	Uras Benlioğlu	706	2072942	441202795
4996	4997	B+	HI-TECH MUSIC LTD	797	1055091	377331722
4997	4998	B+	Mastersaint	110	3265735	311758426
4998	4999	B+	Bruce McIntosh	3475	32990	14563764
4999	5000	B+	SehatAQUA	254	21172	73312511

5000 rows × 6 columns

In [7]:
```
ytube.dropna()
```

Out[7]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
0	1	A++	Zee TV	82757	18752951	20869786591
1	2	A++	T-Series	12661	61196302	47548839843
2	3	A++	Cocomelon - Nursery Rhymes	373	19238251	9793305082
3	4	A++	SET India	27323	31180559	22675948293
4	5	A++	WWE	36756	32852346	26273668433
...	...	...	...	...	...	...
4995	4996	B+	Uras Benlioğlu	706	2072942	441202795
4996	4997	B+	HI-TECH MUSIC LTD	797	1055091	377331722
4997	4998	B+	Mastersaint	110	3265735	311758426
4998	4999	B+	Bruce McIntosh	3475	32990	14563764
4999	5000	B+	SehatAQUA	254	21172	73312511

5000 rows × 6 columns

In [8]:
```
ytube.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Rank           5000 non-null   int32 
 1   Grade          5000 non-null   object
 2   Channel name   5000 non-null   object
 3   Video Uploads  5000 non-null   object
 4   Subscribers    5000 non-null   object
 5   Video views    5000 non-null   int64 
dtypes: int32(1), int64(1), object(4)
memory usage: 215.0+ KB
```

In [9]:
```
#we need to remove the - from the data
mask1=ytube[ytube['Subscribers'].str.contains('-')].index
ytube.drop(labels=mask1,axis=0,inplace=True)
```

In [10]:
```
#we have to convert the object datatype as INTEGER !
ytube['Subscribers']=ytube['Subscribers'].astype('int')
```

In [11]:
```
mask1
```

Out[11]:
```
Int64Index([  17,  108,  115,  142,  143,  152,  156,  175,  180,  189,
            ...
            4892, 4893, 4895, 4912, 4936, 4941, 4948, 4956, 4961, 4990],
           dtype='int64', length=387)
```

In [12]:
```
ytube.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4613 entries, 0 to 4999
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Rank           4613 non-null   int32 
 1   Grade          4613 non-null   object
 2   Channel name   4613 non-null   object
 3   Video Uploads  4613 non-null   object
 4   Subscribers    4613 non-null   int32 
 5   Video views    4613 non-null   int64 
dtypes: int32(2), int64(1), object(3)
memory usage: 216.2+ KB
```

In [13]:
```
mask2=ytube[ytube['Video Uploads'].str.contains('-')].index
ytube.drop(labels=mask2,axis=0,inplace=True)
```

In [14]:
```
ytube['Video Uploads']=ytube['Video Uploads'].astype('int')
```

In [15]:
```
mask2
```

Out[15]:
```
Int64Index([2323, 3072, 4898], dtype='int64')
```

In [16]:
```
ytube.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4610 entries, 0 to 4999
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Rank           4610 non-null   int32 
 1   Grade          4610 non-null   object
 2   Channel name   4610 non-null   object
 3   Video Uploads  4610 non-null   int32 
 4   Subscribers    4610 non-null   int32 
 5   Video views    4610 non-null   int64 
dtypes: int32(3), int64(1), object(2)
memory usage: 198.1+ KB
```

In [17]:
```
ytube['Grade'].unique()
```

Out[17]:
```
array(['A++ ', 'A+ ', 'A ', 'A- ', 'B+ '], dtype=object)
```

In [18]:
```
channel_map={'A++ ':5,'A+ ':4,'A ':3,'A- ':2,'B+ ':1}
```

In [19]:
```
ytube['Grade']=ytube['Grade'].map(channel_map)
```

In [20]:
```
ytube
```

Out[20]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
0	1	5	Zee TV	82757	18752951	20869786591
1	2	5	T-Series	12661	61196302	47548839843
2	3	5	Cocomelon - Nursery Rhymes	373	19238251	9793305082
3	4	5	SET India	27323	31180559	22675948293
4	5	5	WWE	36756	32852346	26273668433
...	...	...	...	...	...	...
4995	4996	1	Uras Benlioğlu	706	2072942	441202795
4996	4997	1	HI-TECH MUSIC LTD	797	1055091	377331722
4997	4998	1	Mastersaint	110	3265735	311758426
4998	4999	1	Bruce McIntosh	3475	32990	14563764
4999	5000	1	SehatAQUA	254	21172	73312511

4610 rows × 6 columns