y-tube Channels Analysis Using Python

The y-tube set has the information about the Channels.

The Data set available from Flexible which is a Third Party y-tube which engine , and available on Kaggle dataset for free.

Import Library

In [1]:
```
import pandas as pd
```

In [2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as sns
```

```
C:\Users\Syed Arif\anaconda3\lib\site-packages\scipy\__init__.py:146: UserWar
ning: A NumPy version >=1.16.5 and <1.23.0 is required for this version of Sc
iPy (detected version 1.25.1
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
```

Uploading Csv fle

In [3]:
```
df = pd.read_excel(r"C:\Users\Syed Arif\Downloads\Y-tube-Channels.xlsx")
```

Data Preprocessing

.head()

head is used show to the By default = 5 rows in the dataset

In [4]:
```
df.head()
```

Out[4]:

	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

.tail()

tail is used to show rows by Descending order

In [5]:
```
df.tail()
```

Out[5]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
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

.shape

It show the total no of rows & Column in the dataset

In [6]:
```
df.shape
```

Out[6]:
```
(5000, 6)
```

.Columns

It show the no of each Column

In [7]:
```
df.columns
```

Out[7]:
```
Index(['Rank', 'Grade', 'Channel name', 'Video Uploads', 'Subscribers',
       'Video views'],
      dtype='object')
```

.dtypes

This Attribute show the data type of each column

In [8]:
```
df.dtypes
```

Out[8]:
```
Rank             object
Grade            object
Channel name     object
Video Uploads    object
Subscribers      object
Video views       int64
dtype: object
```

.unique()

In a column, It show the unique value of specific column.

In [9]:
```
df["Channel name"].unique()
```

Out[9]:
```
array(['Zee TV', 'T-Series', 'Cocomelon - Nursery Rhymes', ...,
       'Mastersaint', 'Bruce McIntosh', 'SehatAQUA'], dtype=object)
```

.nuique()

It will show the total no of unque value from whole data frame

In [10]:
```
df.nunique()
```

Out[10]:
```
Rank             5000
Grade               6
Channel name     4993
Video Uploads    2286
Subscribers      4612
Video views      5000
dtype: int64
```

.describe()

It show the Count, mean , median etc

In [11]:
```
df.describe()
```

Out[11]:

	Video views
count	5.000000e+03
mean	1.071449e+09
std	2.003844e+09
min	7.500000e+01
25%	1.862329e+08
50%	4.820548e+08
75%	1.124368e+09
max	4.754884e+10

.value_counts

It Shows all the unique values with their count

In [12]:
```
df["Channel name"].value_counts()
```

Out[12]:
```
Thơ Nguyễn                 2
Various Artists - Topic    2
Learn Colors For Kids      2
Super Kids                 2
Funny Vines                2
                          ..
MeLlamanFredy              1
Soosloli PoP               1
SBS 뉴스                     1
酷酷的文                       1
SehatAQUA                  1
Name: Channel name, Length: 4993, dtype: int64
```

Get Overall Statistics about the datafram

In [13]:
```
df.describe()
```

Out[13]:

	Video views
count	5.000000e+03
mean	1.071449e+09
std	2.003844e+09
min	7.500000e+01
25%	1.862329e+08
50%	4.820548e+08
75%	1.124368e+09
max	4.754884e+10

Convert the Exponential part into Decimal +03 , +09 etc

In [14]:
```
pd.options.display.float_format = "{:.2f}".format
```

In [15]:
```
df.describe()
```

Out[15]:

	Video views
count	5000.00
mean	1071449400.15
std	2003843972.12
min	75.00
25%	186232945.75
50%	482054780.00
75%	1124367826.75
max	47548839843.00

Replace "--" to "Nan"

In [16]:
```
df.head(20)
```

Out[16]:

	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
5	6th	A++	Movieclips	30243	17149705	16618094724
6	7th	A++	netd müzik	8500	11373567	23898730764
7	8th	A++	ABS-CBN Entertainment	100147	12149206	17202609850
8	9th	A++	Ryan ToysReview	1140	16082927	24518098041
9	10th	A++	Zee Marathi	74607	2841811	2591830307
10	11th	A+	5-Minute Crafts	2085	33492951	8587520379
11	12th	A+	Canal KondZilla	822	39409726	19291034467
12	13th	A+	Like Nastya Vlog	150	7662886	2540099931
13	14th	A+	Ozuna	50	18824912	8727783225
14	15th	A+	Wave Music	16119	15899764	10989179147
15	16th	A+	Ch3Thailand	49239	11569723	9388600275
16	17th	A+	WORLDSTARHIPHOP	4778	15830098	11102158475
17	18th	A+	Vlad and Nikita	53	--	1428274554
18	19th	A+	Badabun	3060	23603062	5860444053
19	20th	A+	WorkpointOfficial	24287	17687229	14022189654

In [17]:
```
df=df.replace("--", np.nan, regex=True)
```

In [18]:
```
df.head(20)
```

Out[18]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
0	1st	A++	Zee TV	82757.00	18752951.00	20869786591
1	2nd	A++	T-Series	12661.00	61196302.00	47548839843
2	3rd	A++	Cocomelon - Nursery Rhymes	373.00	19238251.00	9793305082
3	4th	A++	SET India	27323.00	31180559.00	22675948293
4	5th	A++	WWE	36756.00	32852346.00	26273668433
5	6th	A++	Movieclips	30243.00	17149705.00	16618094724
6	7th	A++	netd müzik	8500.00	11373567.00	23898730764
7	8th	A++	ABS-CBN Entertainment	100147.00	12149206.00	17202609850
8	9th	A++	Ryan ToysReview	1140.00	16082927.00	24518098041
9	10th	A++	Zee Marathi	74607.00	2841811.00	2591830307
10	11th	A+	5-Minute Crafts	2085.00	33492951.00	8587520379
11	12th	A+	Canal KondZilla	822.00	39409726.00	19291034467
12	13th	A+	Like Nastya Vlog	150.00	7662886.00	2540099931
13	14th	A+	Ozuna	50.00	18824912.00	8727783225
14	15th	A+	Wave Music	16119.00	15899764.00	10989179147
15	16th	A+	Ch3Thailand	49239.00	11569723.00	9388600275
16	17th	A+	WORLDSTARHIPHOP	4778.00	15830098.00	11102158475
17	18th	A+	Vlad and Nikita	53.00	NaN	1428274554
18	19th	A+	Badabun	3060.00	23603062.00	5860444053
19	20th	A+	WorkpointOfficial	24287.00	17687229.00	14022189654

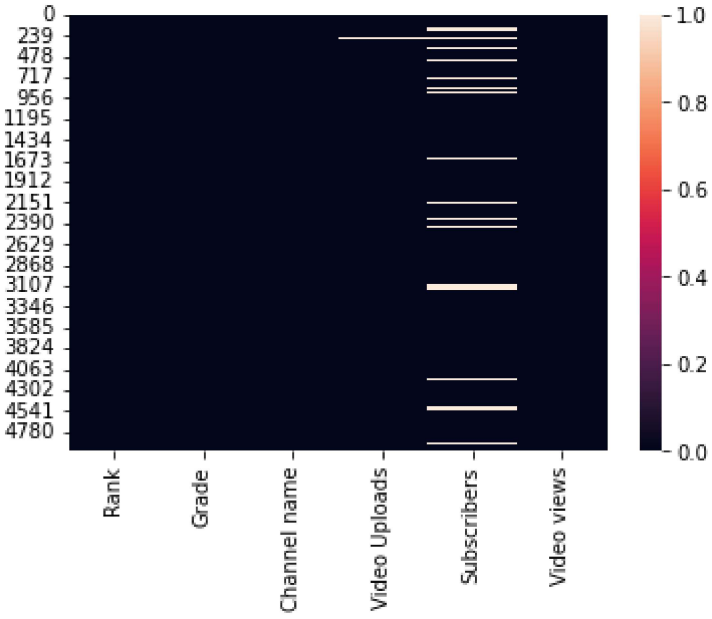
check the Missing Values in our dataset

In [19]:
```
df.isnull().sum()
```

Out[19]:
```
Rank              0
Grade             0
Channel name      0
Video Uploads     6
Subscribers     387
Video views       0
dtype: int64
```

In [20]:
```
sns.heatmap(df.isnull())
```

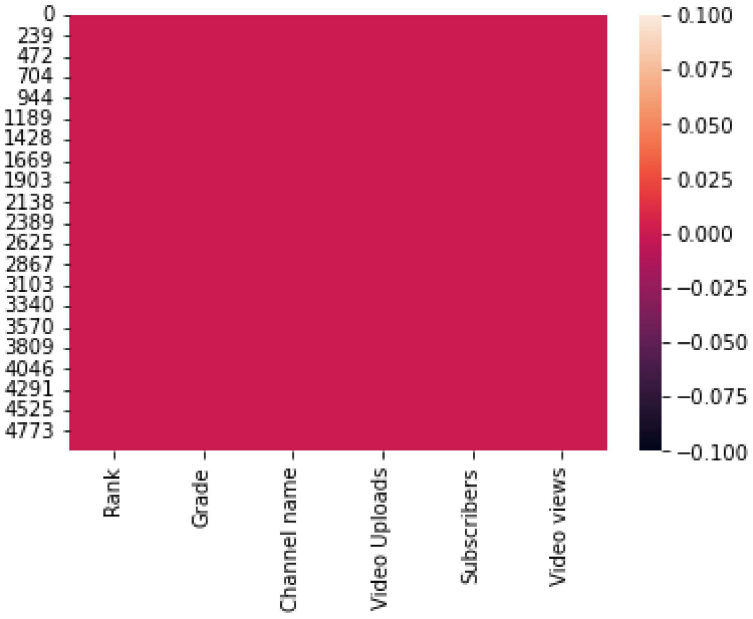
Out[20]: <AxesSubplot:>



In [21]:
```
df.dropna(axis = 0, inplace =True)
```

In [22]:
```
sns.heatmap(df.isnull())
```

Out[22]: <AxesSubplot:>



Remove the string values from Rank Column

In [23]:
```
df
```

Out[23]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
0	1st	A++	Zee TV	82757.00	18752951.00	20869786591
1	2nd	A++	T-Series	12661.00	61196302.00	47548839843
2	3rd	A++	Cocomelon - Nursery Rhymes	373.00	19238251.00	9793305082
3	4th	A++	SET India	27323.00	31180559.00	22675948293
4	5th	A++	WWE	36756.00	32852346.00	26273668433
...	...	...	...	...	...	...
4995	4,996th	B+	Uras Benlioğlu	706.00	2072942.00	441202795
4996	4,997th	B+	HI-TECH MUSIC LTD	797.00	1055091.00	377331722
4997	4,998th	B+	Mastersaint	110.00	3265735.00	311758426
4998	4,999th	B+	Bruce McIntosh	3475.00	32990.00	14563764
4999	5,000th	B+	SehatAQUA	254.00	21172.00	73312511

4610 rows × 6 columns

In [24]:
```
df["Rank"] = df["Rank"].str[0:-2]
```

In [25]:
```
df.tail()
```

Out[25]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
4995	4,996	B+	Uras Benlioğlu	706.00	2072942.00	441202795
4996	4,997	B+	HI-TECH MUSIC LTD	797.00	1055091.00	377331722
4997	4,998	B+	Mastersaint	110.00	3265735.00	311758426
4998	4,999	B+	Bruce McIntosh	3475.00	32990.00	14563764
4999	5,000	B+	SehatAQUA	254.00	21172.00	73312511

We Want To remove Commas from Rank Columns

In [26]:
```
df["Rank"] = df["Rank"].str.replace(",",'')
```

In [27]:
```
df.tail()
```

Out[27]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
4995	4996	B+	Uras Benlioğlu	706.00	2072942.00	441202795
4996	4997	B+	HI-TECH MUSIC LTD	797.00	1055091.00	377331722
4997	4998	B+	Mastersaint	110.00	3265735.00	311758426
4998	4999	B+	Bruce McIntosh	3475.00	32990.00	14563764
4999	5000	B+	SehatAQUA	254.00	21172.00	73312511

In [28]:
```
df.dtypes
```

Out[28]:
```
Rank              object
Grade             object
Channel name      object
Video Uploads    float64
Subscribers      float64
Video views        int64
dtype: object
```

In [31]:
```
df["Rank"] = df["Rank"].astype("int")
```

In [33]:
```
df["Subscribers"] = df["Subscribers"].astype("int")
```

In [34]:
```
df.dtypes
```

Out[34]:
```
Rank               int32
Grade             object
Channel name      object
Video Uploads    float64
Subscribers        int32
Video views        int64
dtype: object
```

Data Cleaning "Grade" Column

In [35]:
```
df["Grade"].unique()
```

Out[35]:
```
array(['A++ ', 'A+ ', 'A ', 'A- ', 'B+ '], dtype=object)
```

In [37]:
```
df["Grade"]=df["Grade"].map({'A++ ': 5,'A+ ': 4 ,'A ': 3,'A- ': 2,'A- ': 1})
```

In [38]:
```
df.head()
```

Out[38]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
0	1	5.00	Zee TV	82757.00	18752951	20869786591
1	2	5.00	T-Series	12661.00	61196302	47548839843
2	3	5.00	Cocomelon - Nursery Rhymes	373.00	19238251	9793305082
3	4	5.00	SET India	27323.00	31180559	22675948293
4	5	5.00	WWE	36756.00	32852346	26273668433

Find Out the Maximum Number of "Videos Upload"

In [39]:
```
df.columns
```

Out[39]:
```
Index(['Rank', 'Grade', 'Channel name', 'Video Uploads', 'Subscribers',
       'Video views'],
      dtype='object')
```

In [43]:
```
df.sort_values(by = 'Video Uploads', ascending = False).head(5)
```

Out[43]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
3453	3454	NaN	AP Archive	422326.00	746325	548619569
1149	1150	1.00	YTN NEWS	355996.00	820108	1640347646
2223	2224	NaN	SBS Drama	335521.00	1418619	1565758044
323	324	3.00	GMA News	269065.00	2599175	2786949164
2956	2957	NaN	MLB	267649.00	1434206	1329206392

Find the Corelation

In [44]:
```
df.corr()
```

Out[44]:

	Rank	Grade	Video Uploads	Subscribers	Video views
Rank	1.00	-0.88	-0.07	-0.38	-0.40
Grade	-0.88	1.00	0.08	0.31	0.38
Video Uploads	-0.07	0.08	1.00	0.01	0.09
Subscribers	-0.38	0.31	0.01	1.00	0.79
Video views	-0.40	0.38	0.09	0.79	1.00

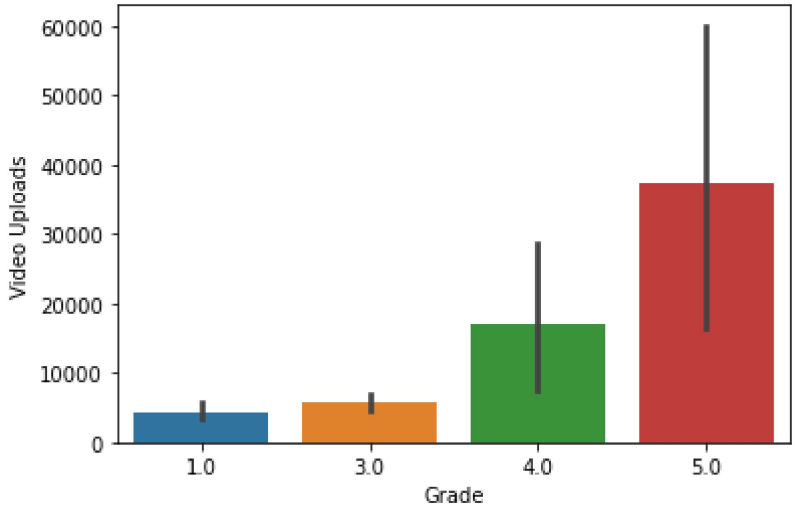
Whcih Grade the Maximum Number of Video_Upload

In [45]:
```
df.columns
```

Out[45]:
```
Index(['Rank', 'Grade', 'Channel name', 'Video Uploads', 'Subscribers',
       'Video views'],
      dtype='object')
```

In [49]:
```
sns.barplot(x = "Grade", y = "Video Uploads", data = df)
```

Out[49]:
```
<AxesSubplot:xlabel='Grade', ylabel='Video Uploads'>
```



Which Grade has Heighest Number of Views

In [50]:
```
sns.barplot(x = "Grade", y = 'Video views', data = df)
```

Out[50]:
```
<AxesSubplot:xlabel='Grade', ylabel='Video views'>
```