In [1]:

```
import pandas as pd
```

In [2]:

```
import seaborn as sns
```

In [4]:

```
#Loading the Dataset
data = pd.read_csv(r'C:\Users\KIIT\Desktop\pandas project\Youtube Analysis\top-5000-youtube-channels.csv
```

In [5]:

```
data
```

Out[5]:

	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

1. Display all the rows except the last 5 rows using the Head method .

In [6]:

```
data.head(-5)
```

Out[6]:

	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
...	...	...	...	...	...	...
4990	4,991st	B+	Ho Ngoc Ha's Official Channel	208	--	127185704
4991	4,992nd	B+	Toys to Learn Colors	11	663114	141933264
4992	4,993rd	B+	KAZKA	25	131766	74304638
4993	4,994th	B+	United CUBE (CUBE Entertainment…	1055	1586835	371299166
4994	4,995th	B+	Wings Marathi	1735	1099659	346175699

4995 rows × 6 columns

2. Display All rows except the first 5 rows using tail method.

In [7]:

```
data.tail(-5)
```

Out[7]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
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
...	...	...	...	...	...	...
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

4995 rows × 6 columns

3. Find the shape of our dataset (Number of rows and coloumns)

In [8]:

```
data.shape
```

Out[8]:

```
(5000, 6)
```

```
print("Number of Rows :",data.shape[0])
print("Number of Columns : ",data.shape[1])
```

4. Find Information about our dataset like total number of rows,total number of columns,datatypes of each column and memory requirement.

In [11]:

```
data.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
```

5. Find overall statistics of our dataset

In [15]:

```
pd.options.display.float_format = '{:.2f}'.format
```

In [16]:

```
data.describe()
```

Out[16]:

	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

6. Data Cleaning (Replace '--' with Nan)

In [18]:

```
import numpy as np
```

In [19]:

```
data.head(20)
```

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

In [20]:

```
data= data.replace('--',np.nan,regex=True)
```

In [21]:

```
data.head(20)
```

Out[21]:

	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	NaN	1428274554
18	19th	A+	Badabun	3060	23603062	5860444053
19	20th	A+	WorkpointOfficial	24287	17687229	14022189654

7. Check Null values in the dataset .

In [23]:

```
data.isnull().sum()
```

Out[23]:

```
Rank              0
Grade             0
Channel name      0
Video Uploads     6
Subscribers     387
Video views       0
dtype: int64
```

In [25]:

```
percentage_missing = data.isnull().sum() * 100 / len(data)
```

In [26]:

```
percentage_missing
```

Out[26]:

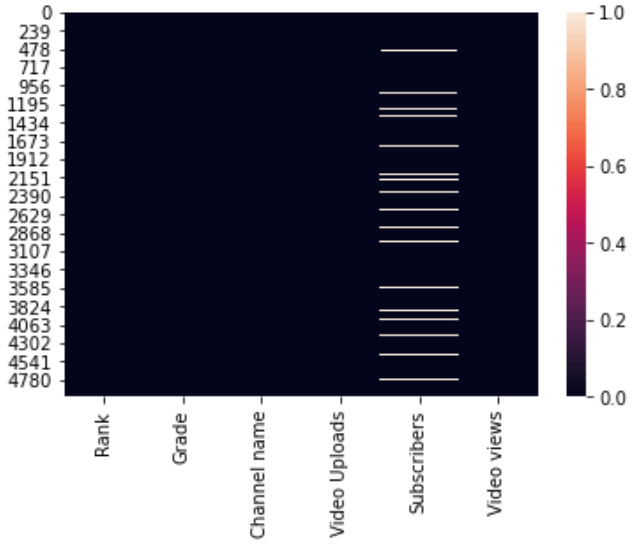
```
Rank            0.00
Grade           0.00
Channel name    0.00
Video Uploads   0.12
Subscribers     7.74
Video views     0.00
dtype: float64
```

In [27]:

```
sns.heatmap(data.isnull())
```

Out[27]:

```
<AxesSubplot:>
```



In [29]:

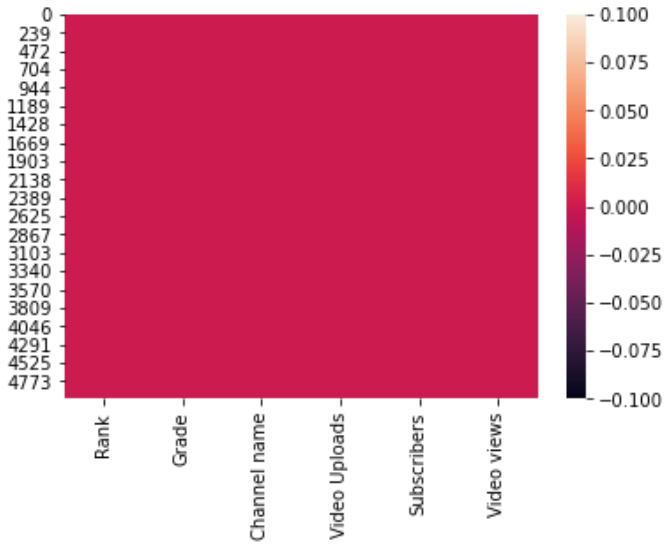
```
data.dropna(axis=0,inplace=True)
```

In [30]:

```
sns.heatmap(data.isnull())
```

Out[30]:

```
<AxesSubplot:>
```



8. Data cleaning (rank column)

In [31]:

```
data.head()
```

Out[31]:

	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

In [32]:

```
data.dtypes
```

Out[32]:

```
Rank            object
Grade           object
Channel name    object
Video Uploads   object
Subscribers     object
Video views      int64
dtype: object
```

In [33]:

```
data['Rank']= data['Rank'].str[:-2]
```

In [34]:

```
data.head()
```

Out[34]:

	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

In [36]:

```
data['Rank'] = data['Rank'].str.replace(',','').astype('int')
```

In [37]:

```
data.tail()
```

Out[37]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views
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

9. Data cleaning (Video uploads and subscribers)

In [38]:

```
data.head()
```

Out[38]:

	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

In [39]:

```
data.dtypes
```

Out[39]:

```
Rank             int32
Grade           object
Channel name    object
Video Uploads   object
Subscribers     object
Video views      int64
dtype: object
```

In [40]:

```
data['Video Uploads']= data['Video Uploads'].astype('int')
```

In [41]:

```
data['Subscribers']= data['Subscribers'].astype('int')
```

In [42]:

```
data.dtypes
```

Out[42]:

```
Rank             int32
Grade           object
Channel name    object
Video Uploads    int32
Subscribers      int32
Video views      int64
dtype: object
```

10 . Data cleaning (Grade column)

In [43]:

```
data.head()
```

Out[43]:

	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

In [44]:

```
data['Grade'].unique()
```

Out[44]:

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

In [45]:

```
data['Grade'] = data['Grade'].map({'A++ ':5,'A+ ':4,'A ':3,'A- ':2,'B+ ':1})
```

In [46]:

```
data.head()
```

Out[46]:

	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

In [47]:

```
data.dtypes
```

Out[47]:

```
Rank              int32
Grade             int64
Channel name     object
Video Uploads     int32
Subscribers       int32
Video views       int64
dtype: object
```

11. Find Average views for each column

In [48]:

```
data.columns
```

Out[48]:

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

In [50]:

```
data['Avg Views'] = data['Video views'] / data['Video Uploads']
```

In [51]:

```
data.head()
```

Out[51]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views	Avg Views
0	1	5	Zee TV	82757	18752951	20869786591	252181.53
1	2	5	T-Series	12661	61196302	47548839843	3755535.89
2	3	5	Cocomelon - Nursery Rhymes	373	19238251	9793305082	26255509.60
3	4	5	SET India	27323	31180559	22675948293	829921.62
4	5	5	WWE	36756	32852346	26273668433	714813.05

12.Find out the top 5 channels with maximum number of Video Uploads.

In [52]:

```
data.columns
```

Out[52]:

```
Index(['Rank', 'Grade', 'Channel name', 'Video Uploads', 'Subscribers',
       'Video views', 'Avg Views'],
      dtype='object')
```

In [57]:

```
data.sort_values('Video Uploads',ascending = False).head(5)
```

Out[57]:

	Rank	Grade	Channel name	Video Uploads	Subscribers	Video views	Avg Views
3453	3454	1	AP Archive	422326	746325	548619569	1299.04
1149	1150	2	YTN NEWS	355996	820108	1640347646	4607.77
2223	2224	1	SBS Drama	335521	1418619	1565758044	4666.65
323	324	3	GMA News	269065	2599175	2786949164	10357.90
2956	2957	1	MLB	267649	1434206	1329206392	4966.23

13. Find Correlation Matrix

In [58]:

```
data.corr()
```

Out[58]:

	Rank	Grade	Video Uploads	Subscribers	Video views	Avg Views
Rank	1.00	-0.87	-0.07	-0.38	-0.40	-0.15
Grade	-0.87	1.00	0.09	0.43	0.48	0.16
Video Uploads	-0.07	0.09	1.00	0.01	0.09	-0.06
Subscribers	-0.38	0.43	0.01	1.00	0.79	0.29
Video views	-0.40	0.48	0.09	0.79	1.00	0.29
Avg Views	-0.15	0.16	-0.06	0.29	0.29	1.00

14. Which Grade has maximum number of Video Uploads ?

In [60]:

```
data.columns
```

Out[60]:

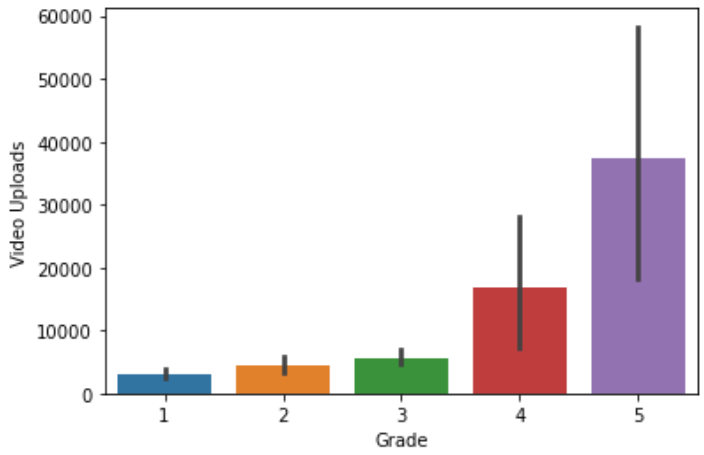
```
Index(['Rank', 'Grade', 'Channel name', 'Video Uploads', 'Subscribers',
       'Video views', 'Avg Views'],
      dtype='object')
```

In [66]:

```
sns.barplot(x='Grade',y='Video Uploads',data=data)
```

Out[66]:

```
<AxesSubplot:xlabel='Grade', ylabel='Video Uploads'>
```



15. Which Grade has the Highest Average Views ?

In [68]:

```
data.columns
```

Out[68]:

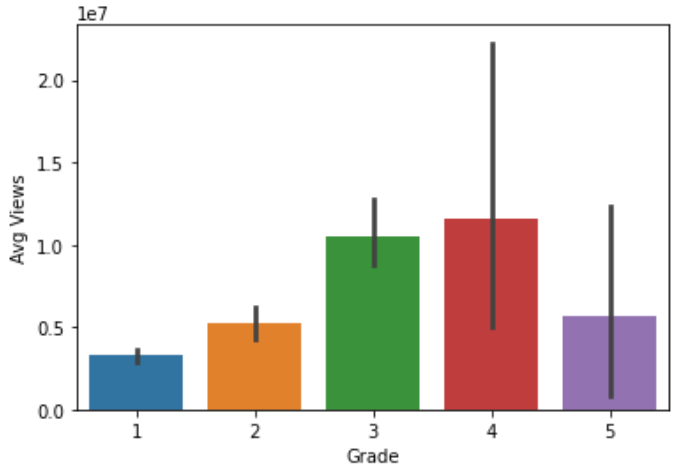
```
Index(['Rank', 'Grade', 'Channel name', 'Video Uploads', 'Subscribers',
       'Video views', 'Avg Views'],
      dtype='object')
```

In [69]:

```
sns.barplot(x='Grade',y='Avg Views',data=data)
```

Out[69]:

```
<AxesSubplot:xlabel='Grade', ylabel='Avg Views'>
```



16. Which grade has the highest number of subscribers ?

In [70]:

```
data.columns
```

Out[70]:

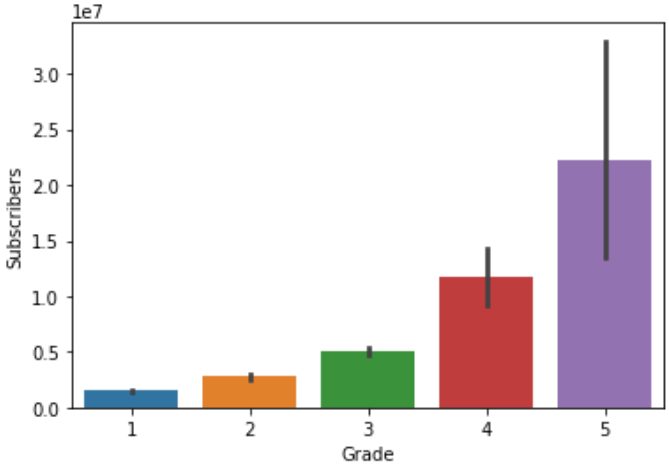
```
Index(['Rank', 'Grade', 'Channel name', 'Video Uploads', 'Subscribers',
       'Video views', 'Avg Views'],
      dtype='object')
```

In [71]:

```
sns.barplot(x='Grade',y='Subscribers',data=data)
```

Out[71]:

```
<AxesSubplot:xlabel='Grade', ylabel='Subscribers'>
```



17. Which Grade has the Highest Video Views ?

In [72]:

```
data.columns
```

Out[72]:

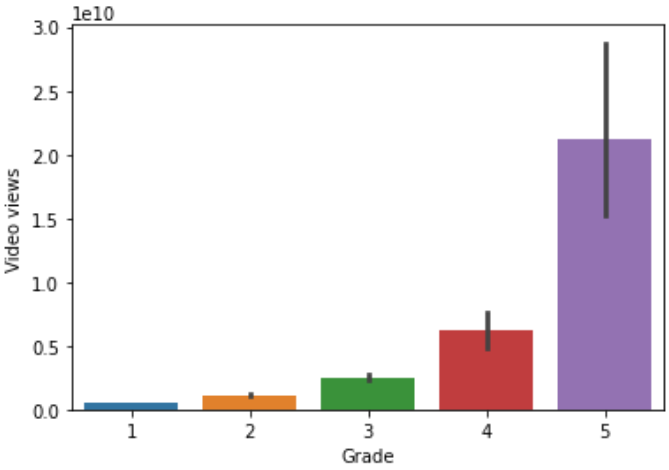
```
Index(['Rank', 'Grade', 'Channel name', 'Video Uploads', 'Subscribers',
       'Video views', 'Avg Views'],
      dtype='object')
```

In [74]:

```
sns.barplot(x='Grade',y='Video views',data=data)
```

Out[74]:

```
<AxesSubplot:xlabel='Grade', ylabel='Video views'>
```



In [ ]: