INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

WEB SCRAPING AND EXPLORATORY DATA ON
'YOUTUBE CHANNELS'

TEAM MEMBERS:

Y. Mohana…………..Qualification(B.Tech, CE)
A. Maheswari………Qualification(B.sc
Statistics)

What is Data Science?

- Data science is widely used in understanding and analysing huge data.

- As data science deals with Statistics, SQL, and Python program, it allows me to find a career not just in IT firms but also in Agri based companies.

Why Innomatic's ?

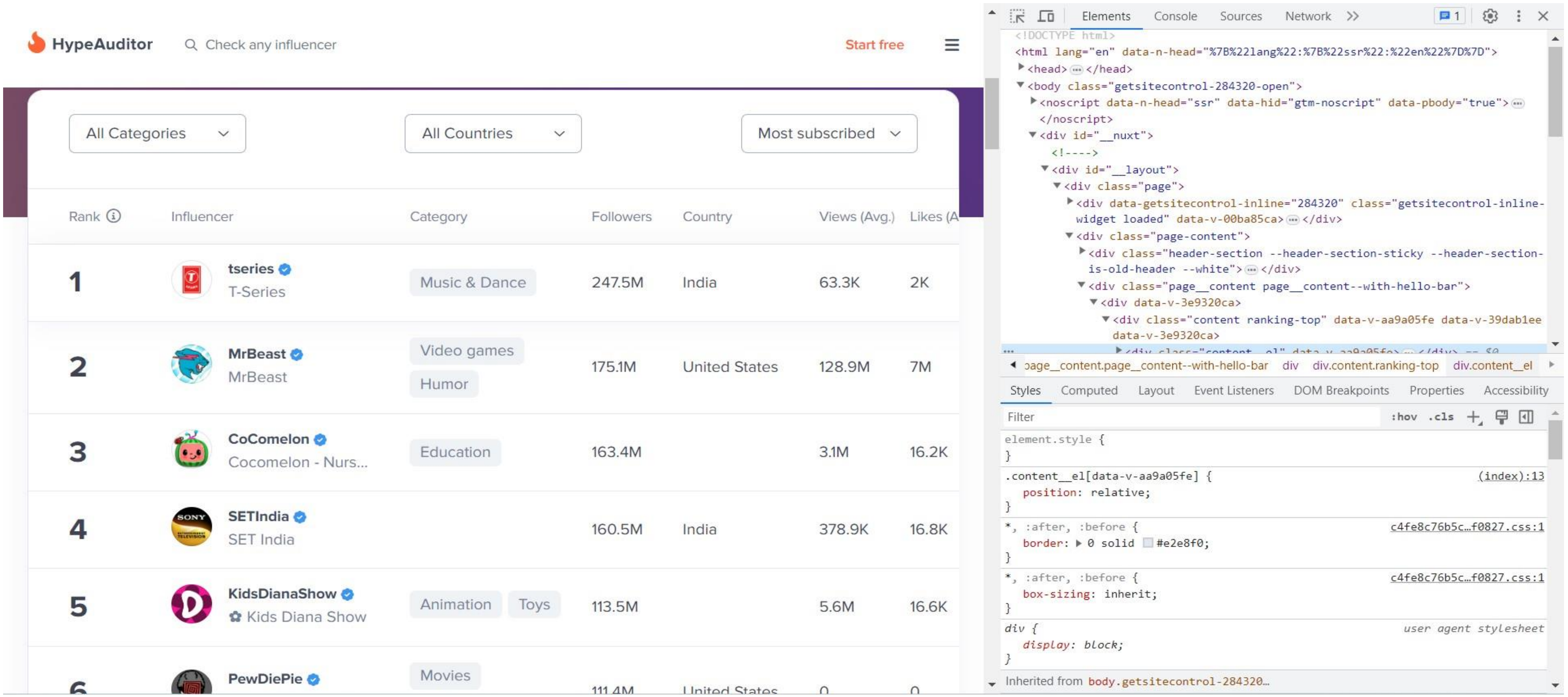
- Highly reputed Institute with great career opportunities for students of any educational background.
- Good teaching staff and also provide mentoring sessions.

Contents:

- Selection of website
- Problem Statement
- Libraries used
- Data frame Creation
- Data Cleaning
- Descriptive Statistics
- Univariant Analysis
- Bivariant / Multi-variant analysis
- Final Conclusion

Websites:

https://hypeauditor.com/top-youtube/



Rank	Influencer	Category	Followers	Country	Views (Avg.)	Likes (A
1	tseries T-Series	Music & Dance	247.5M	India	63.3K	2K
2	MrBeast MrBeast	Video games Humor	175.1M	United States	128.9M	7M
3	CoComelon Cocomelon - Nurs...	Education	163.4M		3.1M	16.2K
4	SETIndia SET India		160.5M	India	378.9K	16.8K
5	KidsDianaShow Kids Diana Show	Animation Toys	113.5M		5.6M	16.6K
6	PewDiePie	Movies	111.4M	United States	0	0

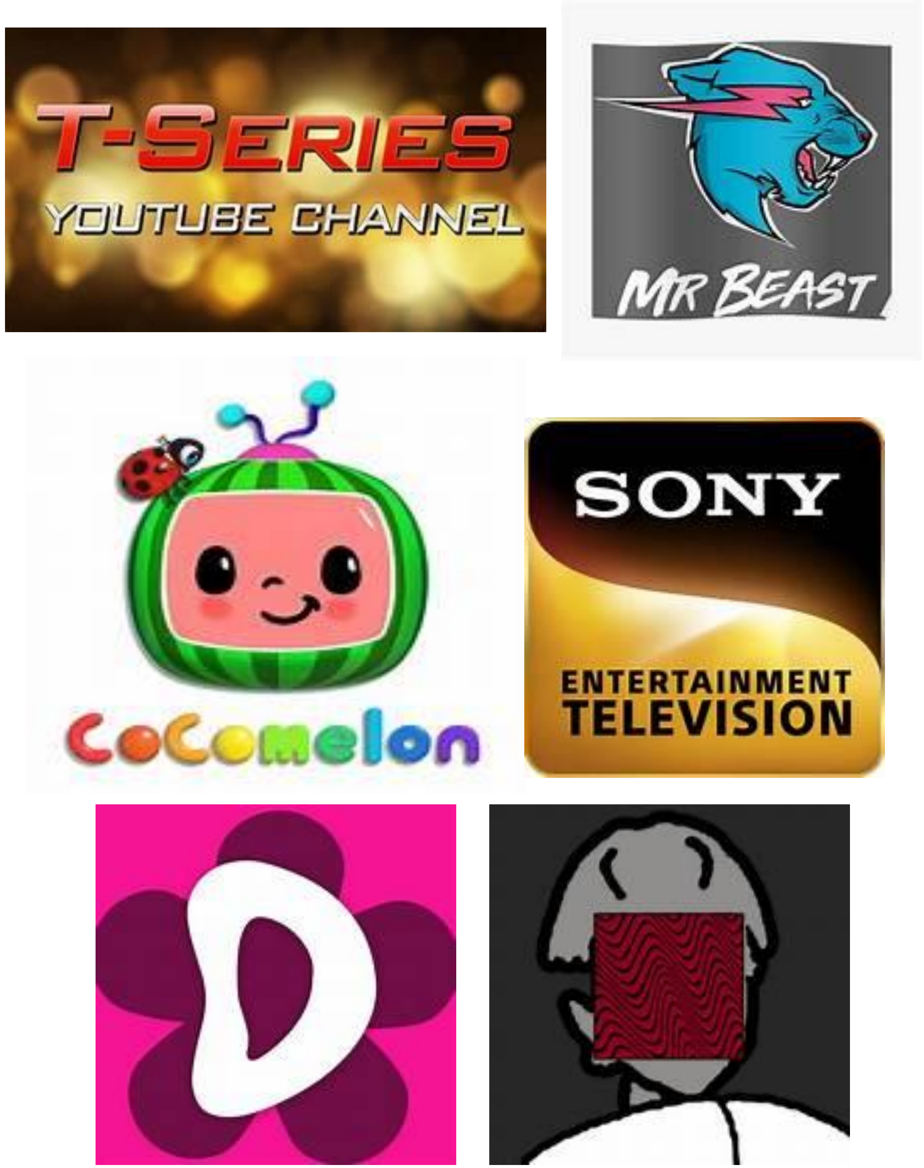
```
<!DOCTYPE html>
<html lang="en" data-n-head="%7B%22lang%22:%7B%22ssr%22:%22en%22%7D%7D">
 ▶<head>…</head>
 ▼<body class="getsitecontrol-284320-open">
   ▶<noscript data-n-head="ssr" data-hid="gtm-noscript" data-pbody="true">…
    </noscript>
   ▼<div id="__nuxt">
      <!---->
     ▼<div id="__layout">
       ▼<div class="page">
         ▶<div data-getsitecontrol-inline="284320" class="getsitecontrol-inline-
          widget loaded" data-v-00ba85ca>…</div>
         ▼<div class="page-content">
           ▶<div class="header-section --header-section-sticky --header-section-
            is-old-header --white">…</div>
           ▼<div class="page__content page__content--with-hello-bar">
             ▼<div data-v-3e9320ca>
               ▼<div class="content ranking-top" data-v-aa9a05fe data-v-39dab1ee
                data-v-3e9320ca>
...              ▶<div class="content__el" data-v-aa9a05fe>…</div> == $0
```

```
◀ page__content.page__content--with-hello-bar   div   div.content.ranking-top   div.content__el ▶
```

```
element.style {
}
.content__el[data-v-aa9a05fe] {                                  (index):13
  position: relative;
}
*, :after, :before {                              c4fe8c76b5c…f0827.css:1
  border: ▶ 0 solid □ #e2e8f0;
}
*, :after, :before {                              c4fe8c76b5c…f0827.css:1
  box-sizing: inherit;
}
div {                                               user agent stylesheet
  display: block;
}
Inherited from body.getsitecontrol-284320…
```

Problem Statement:

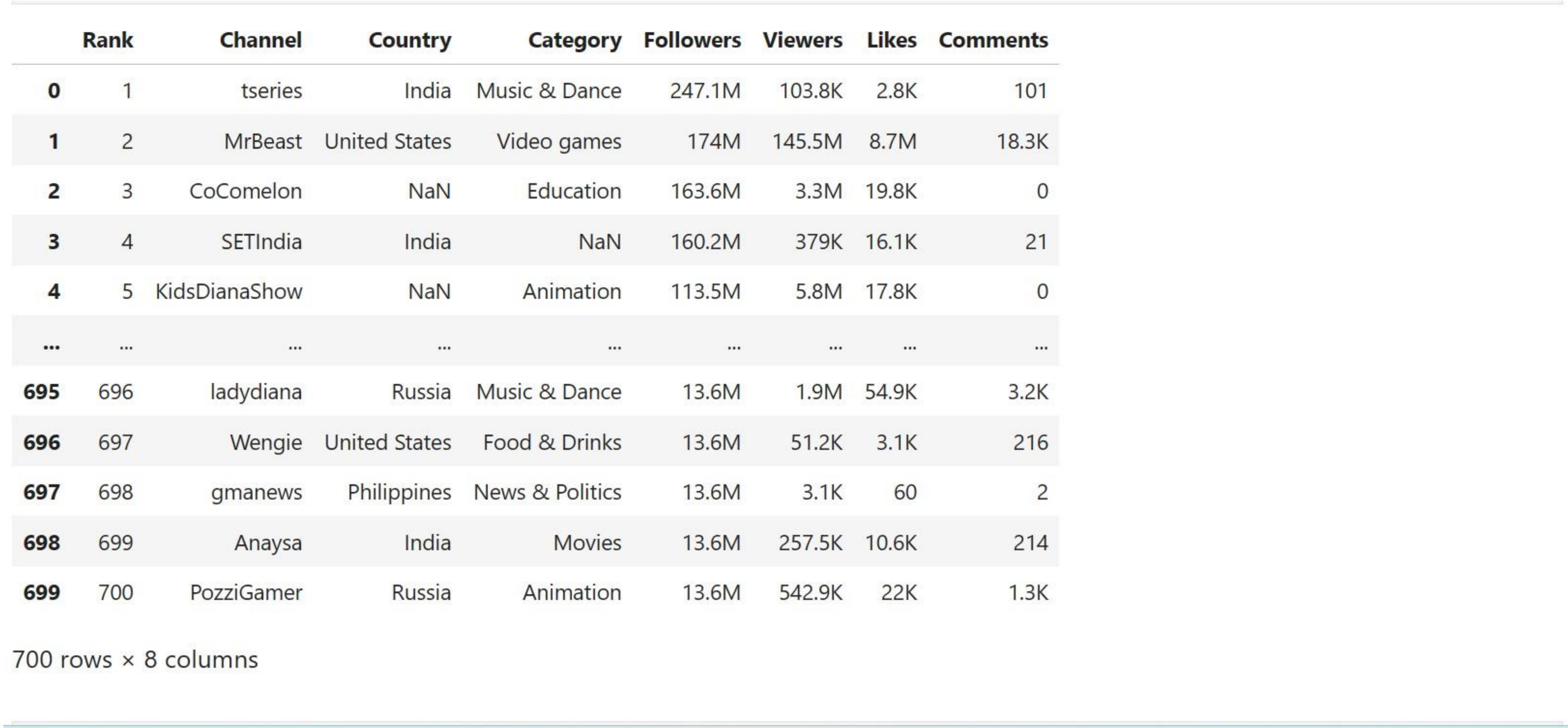
- To analyze which YouTube channel was highly viewed in this Hype auditor website of YouTube .
- To know the highest value of Followers and views.
- To identify the Country, Category, and Followers having outliers in every variable/column.



Libraries used:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Requests
- Beautiful Soup

Data Frame Creation:



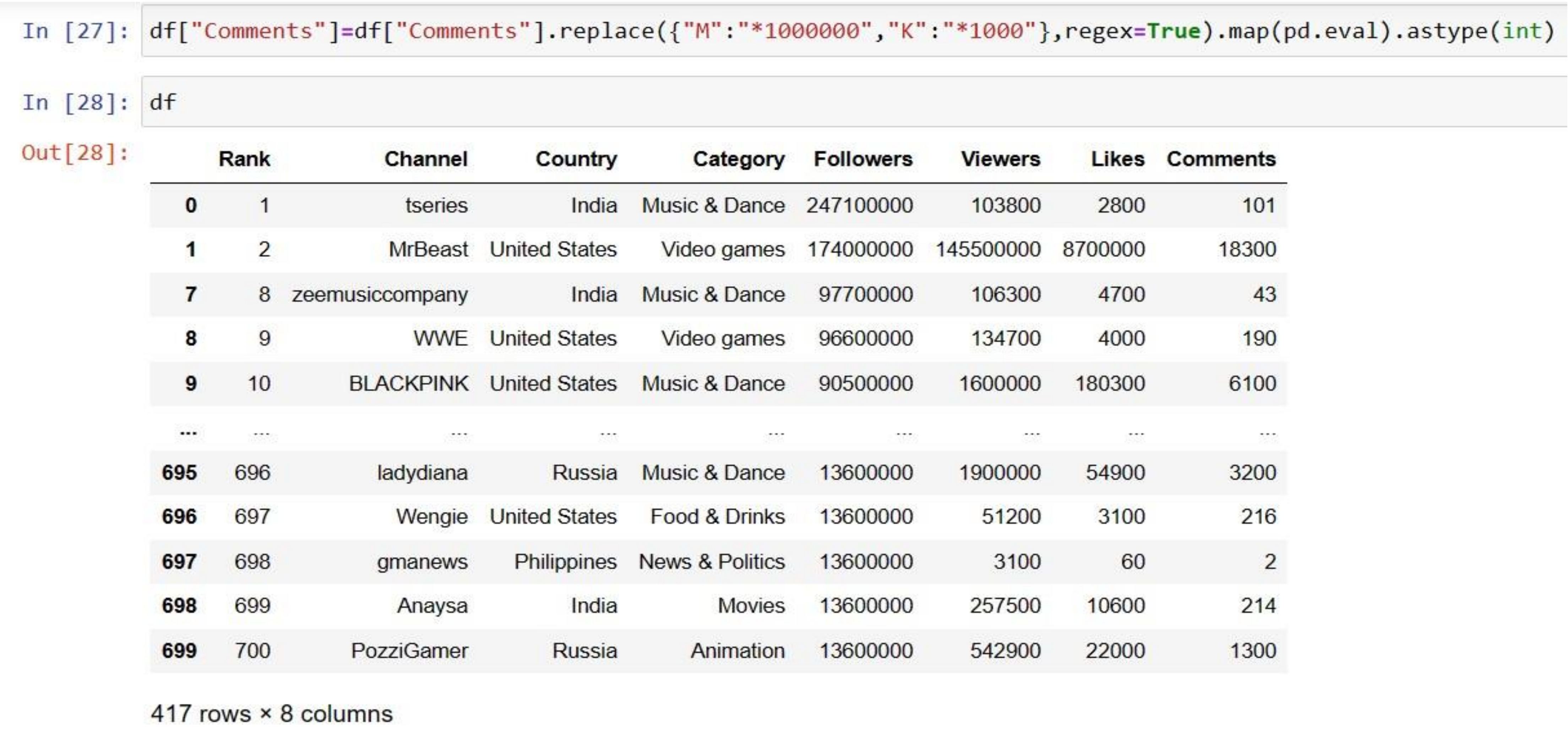
	Rank	Channel	Country	Category	Followers	Viewers	Likes	Comments
0	1	tseries	India	Music & Dance	247.1M	103.8K	2.8K	101
1	2	MrBeast	United States	Video games	174M	145.5M	8.7M	18.3K
2	3	CoComelon	NaN	Education	163.6M	3.3M	19.8K	0
3	4	SETIndia	India	NaN	160.2M	379K	16.1K	21
4	5	KidsDianaShow	NaN	Animation	113.5M	5.8M	17.8K	0
...	...	...	...	...	...	...	...	...
695	696	ladydiana	Russia	Music & Dance	13.6M	1.9M	54.9K	3.2K
696	697	Wengie	United States	Food & Drinks	13.6M	51.2K	3.1K	216
697	698	gmanews	Philippines	News & Politics	13.6M	3.1K	60	2
698	699	Anaysa	India	Movies	13.6M	257.5K	10.6K	214
699	700	PozziGamer	Russia	Animation	13.6M	542.9K	22K	1.3K

700 rows × 8 columns

Data Cleaning :

- Identifying Null values
- Identifying Missing values
- Treating rows and columns with null values and missing values
- Extracting required data from columns to create another columns using regular expressions
- Treating columns and converting required columns from object to int or float datatype

Final Data Frame:



```
In [27]: df["Comments"]=df["Comments"].replace({"M":"*1000000","K":"*1000"},regex=True).map(pd.eval).astype(int)
```

```
In [28]: df
```

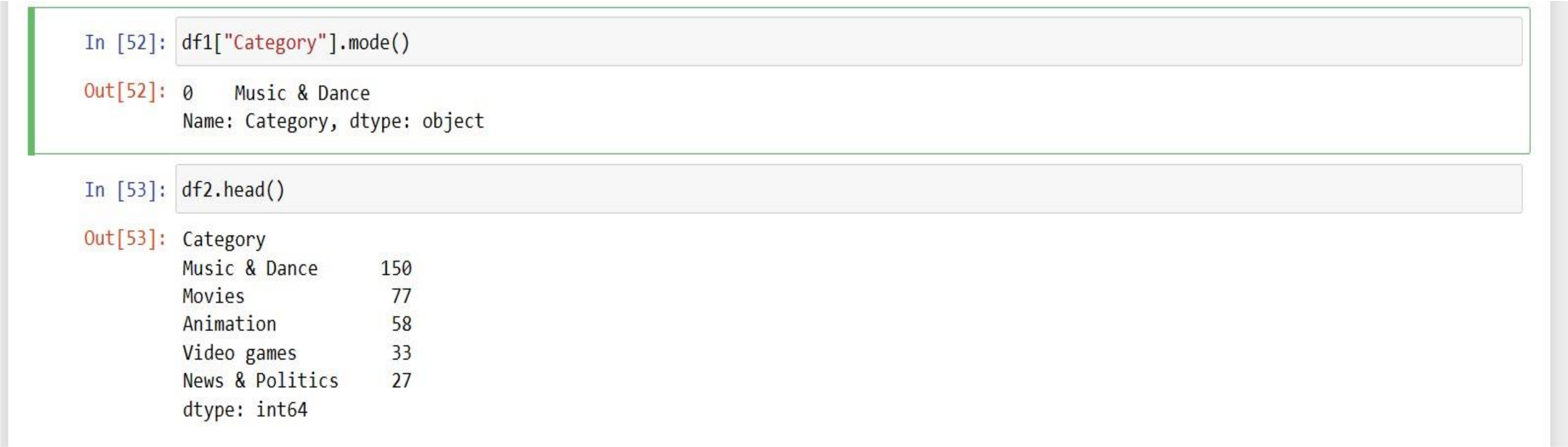
Out[28]:

	Rank	Channel	Country	Category	Followers	Viewers	Likes	Comments
0	1	tseries	India	Music & Dance	247100000	103800	2800	101
1	2	MrBeast	United States	Video games	174000000	145500000	8700000	18300
7	8	zeemusiccompany	India	Music & Dance	97700000	106300	4700	43
8	9	WWE	United States	Video games	96600000	134700	4000	190
9	10	BLACKPINK	United States	Music & Dance	90500000	1600000	180300	6100
...	...	...	...	...	...	...	...	...
695	696	ladydiana	Russia	Music & Dance	13600000	1900000	54900	3200
696	697	Wengie	United States	Food & Drinks	13600000	51200	3100	216
697	698	gmanews	Philippines	News & Politics	13600000	3100	60	2
698	699	Anaysa	India	Movies	13600000	257500	10600	214
699	700	PozziGamer	Russia	Animation	13600000	542900	22000	1300

417 rows × 8 columns

Descriptive Statistics:

Category which is highly Viewed in HypeAuditor website



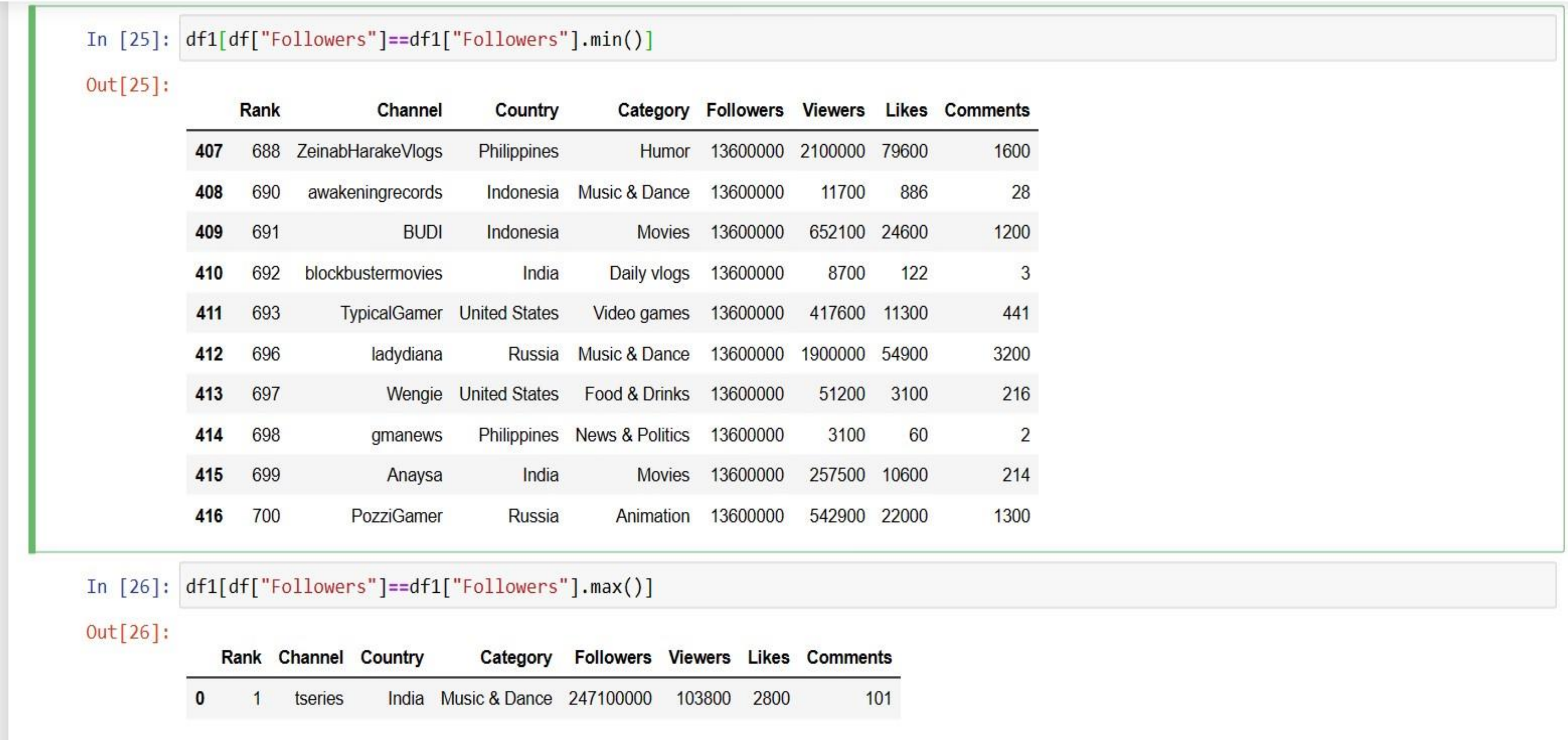
```
In [52]: df1["Category"].mode()

Out[52]: 0    Music & Dance
         Name: Category, dtype: object

In [53]: df2.head()

Out[53]: Category
         Music & Dance      150
         Movies              77
         Animation           58
         Video games         33
         News & Politics     27
         dtype: int64
```

Maximum and Minimum number of Followers:



```
In [25]: df1[df["Followers"]==df1["Followers"].min()]
```

Out[25]:

	Rank	Channel	Country	Category	Followers	Viewers	Likes	Comments
407	688	ZeinabHarakeVlogs	Philippines	Humor	13600000	2100000	79600	1600
408	690	awakeningrecords	Indonesia	Music & Dance	13600000	11700	886	28
409	691	BUDI	Indonesia	Movies	13600000	652100	24600	1200
410	692	blockbustermovies	India	Daily vlogs	13600000	8700	122	3
411	693	TypicalGamer	United States	Video games	13600000	417600	11300	441
412	696	ladydiana	Russia	Music & Dance	13600000	1900000	54900	3200
413	697	Wengie	United States	Food & Drinks	13600000	51200	3100	216
414	698	gmanews	Philippines	News & Politics	13600000	3100	60	2
415	699	Anaysa	India	Movies	13600000	257500	10600	214
416	700	PozziGamer	Russia	Animation	13600000	542900	22000	1300

```
In [26]: df1[df["Followers"]==df1["Followers"].max()]
```

Out[26]:

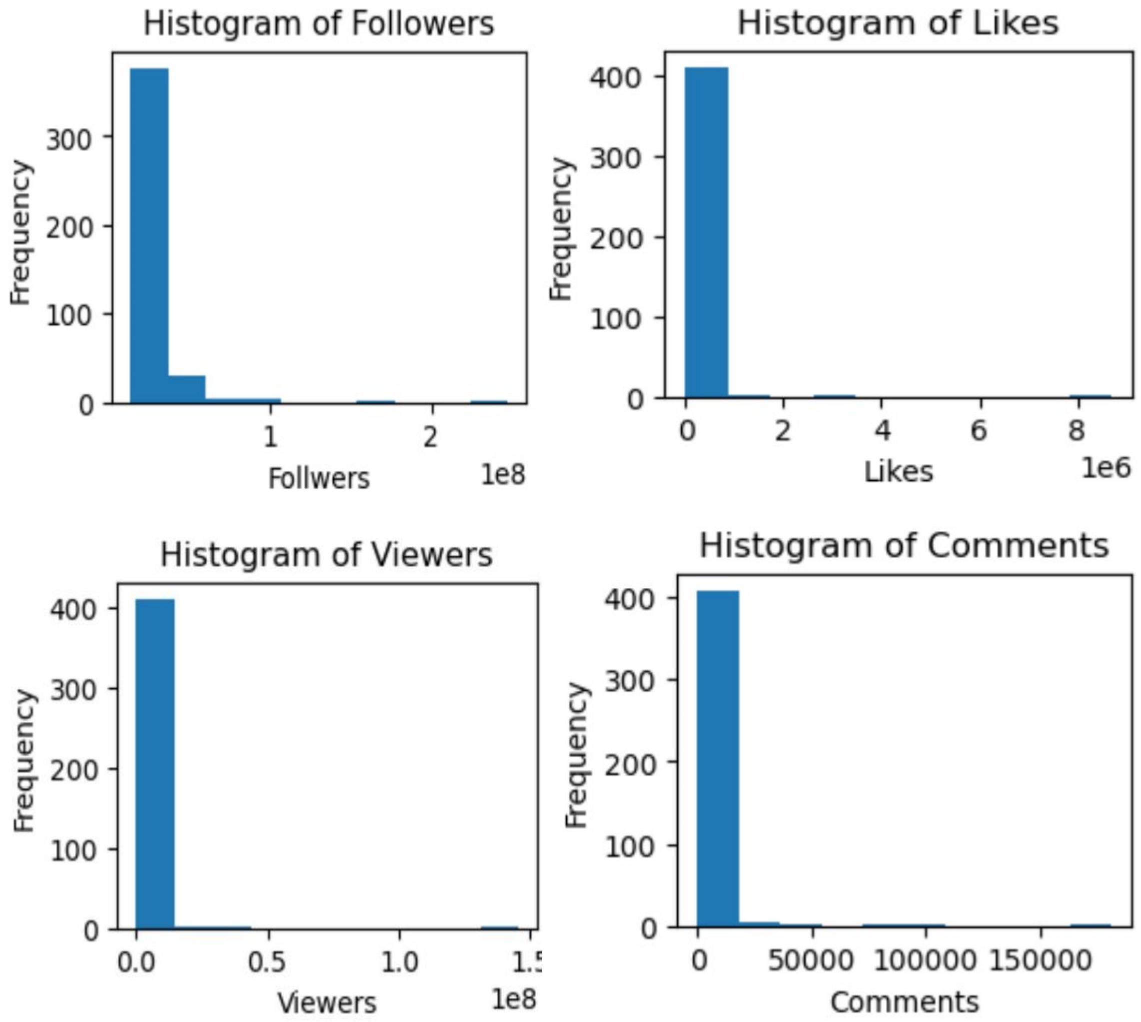
	Rank	Channel	Country	Category	Followers	Viewers	Likes	Comments
0	1	tseries	India	Music & Dance	247100000	103800	2800	101

DATA VISUALIZATION

Data Visualization:

- **Univariant Analysis:**

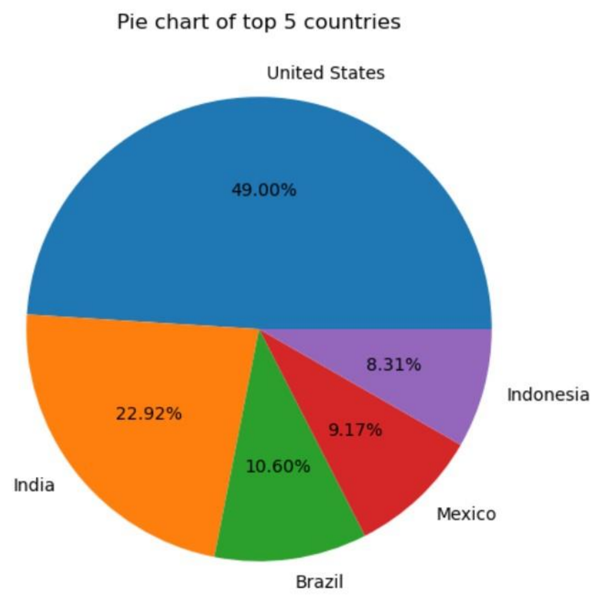
Histogram:

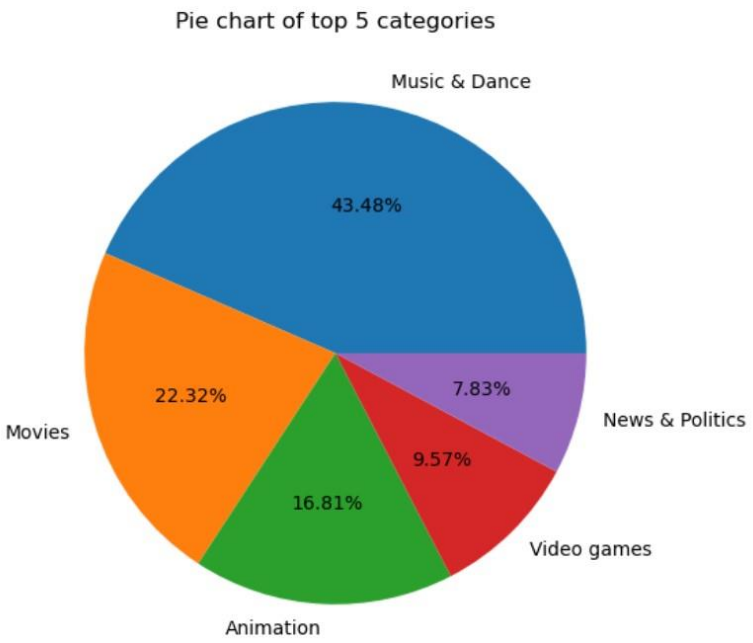


Insights:

By looking into this graph , we clearly understand the highest ranges of Channels for single variable/column (Followers, Viewers, Likes, comments.)

***PIE** CHART*:





Insights:

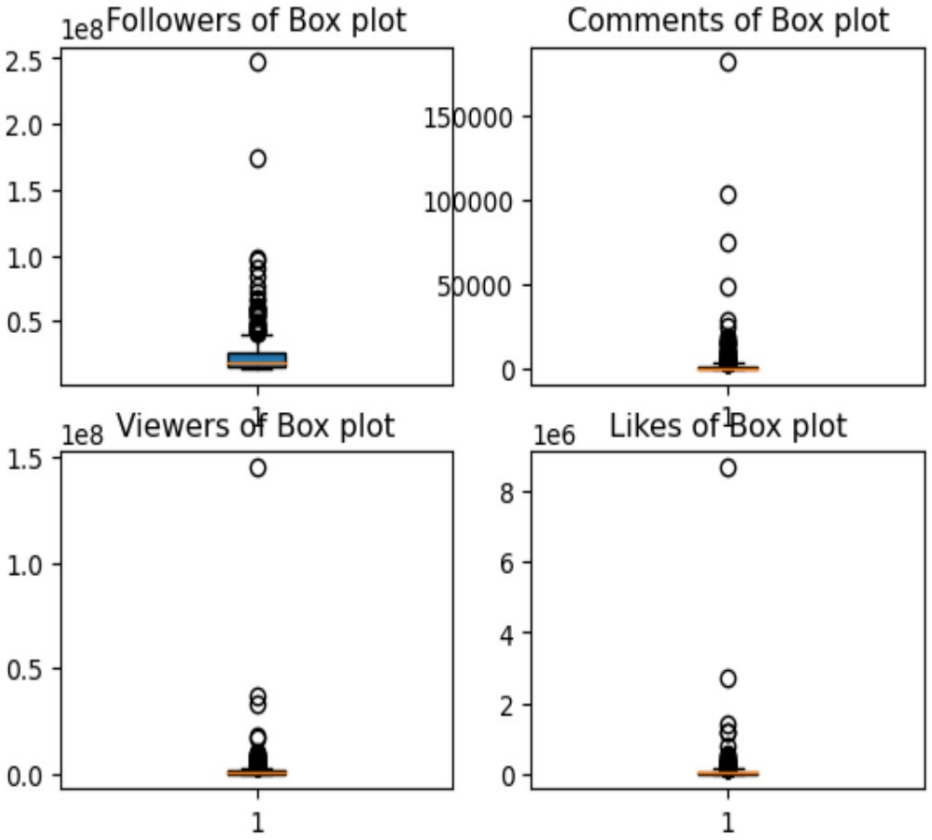
Top Ten Countries which are high in percentage are:

- United States (49.00)
- India (22.92)
- Brazil (10.60)
- Mexico (9.17)
- Indonesia (8.31)

Top Ten Categories which are high in percentage are:

- Music & Dance (43.48)
- Movies (22.32)
- Animation (16.81)
- Video games (9.57)
- News & Politics (7.83)

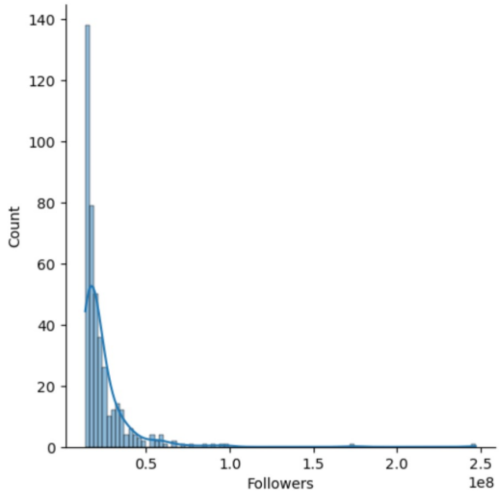
BOX PLOT:

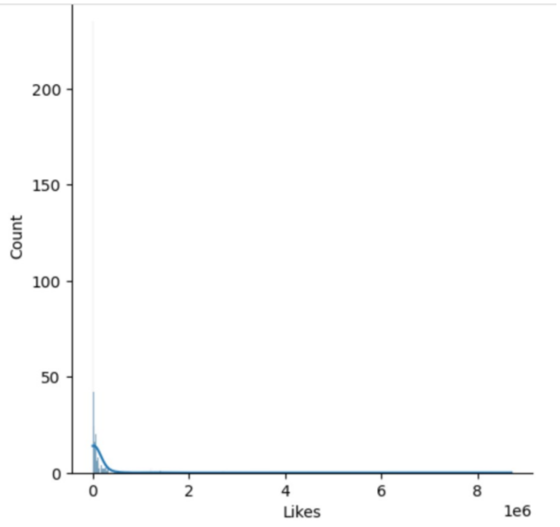


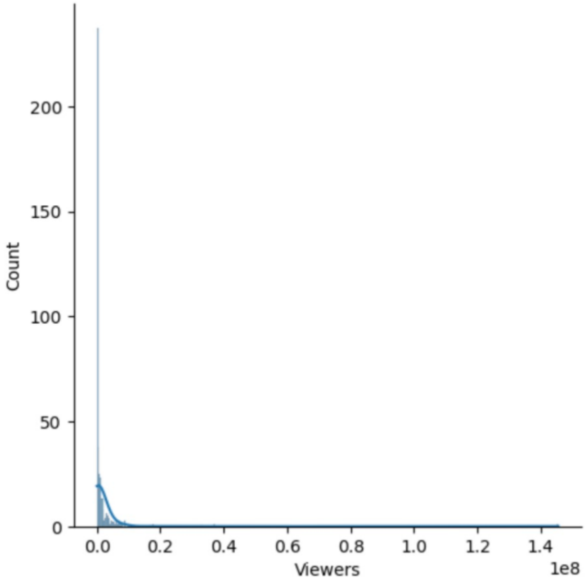
Insights:

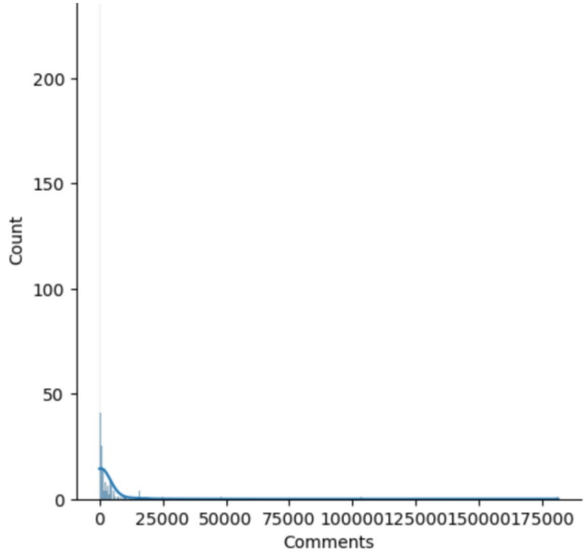
- ➢ By this Box Plot, we determine the outliers in each numerical columns which are Maximum and Minimum which are far away from mean or median.

Density Plot On Numerical Values:









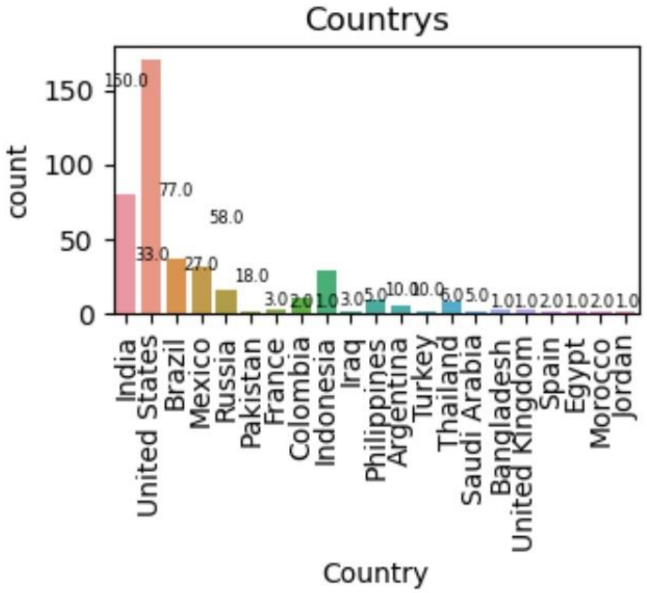
Insights:

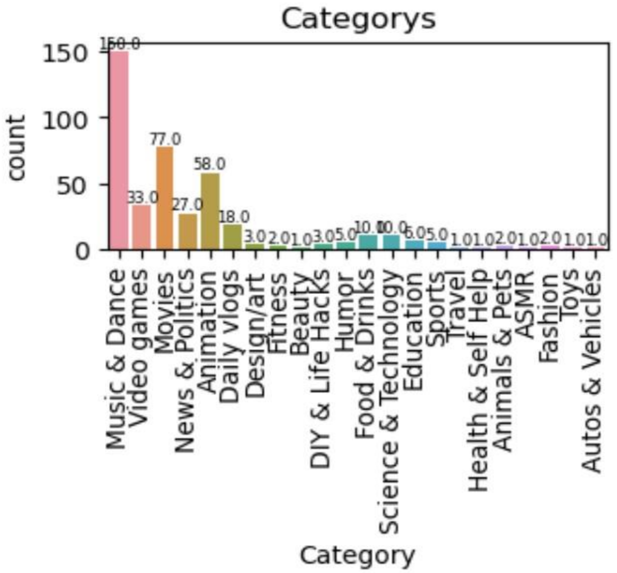
. From the two graphs we know that the density range of Followers are high around 130 for T-series channel when compared to the likes for the same Channel is around 45.

. From these two graphs we observe that the density range of viewers are high around 250 for T-series channel when compared to the density range of comments for the same channel is around 48.

.

Country and Category Counts:



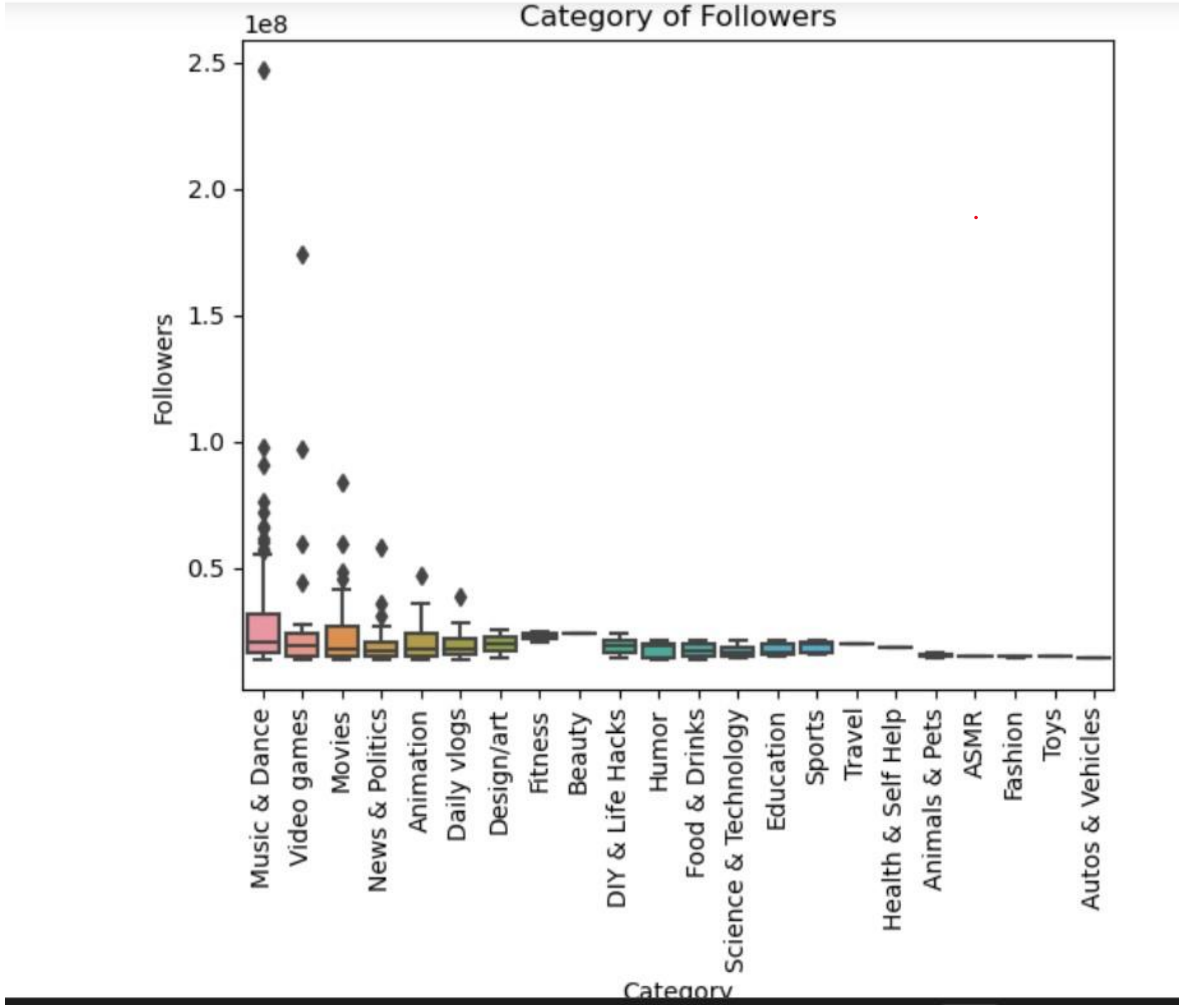


Insights:

. By data visualization graphs we observe that more number of Followers, Viewers, Likes are from the country united states which in first place is around 157 and the second place is taken by India is 77.0

. By these graphs we determine that only three Categories are having high count in which first place is for Music & Dance is 150.0 and second place is for Movies i.e; 77.0 and the third one is for Animation which is 50.0

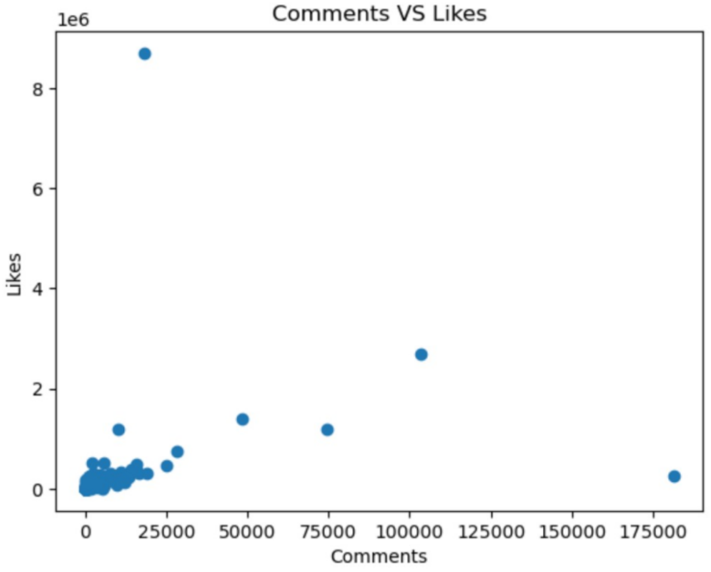
Category vs Followers:

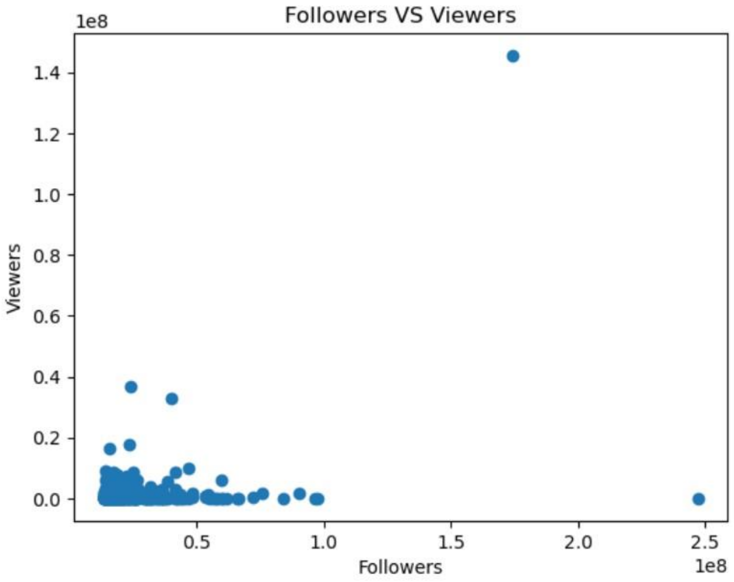


Insights:

- By this Box Plot we observe that the outliers of Category vs Followers in each numerical columns which are Maximum and Minimum which are far away from mean and median.
- In the Boxplot Music & Dance , Video Games, Movies, News & Politics, Animation and Daily vlogs have outliers.

BI- VARIANT ANALYSIS:

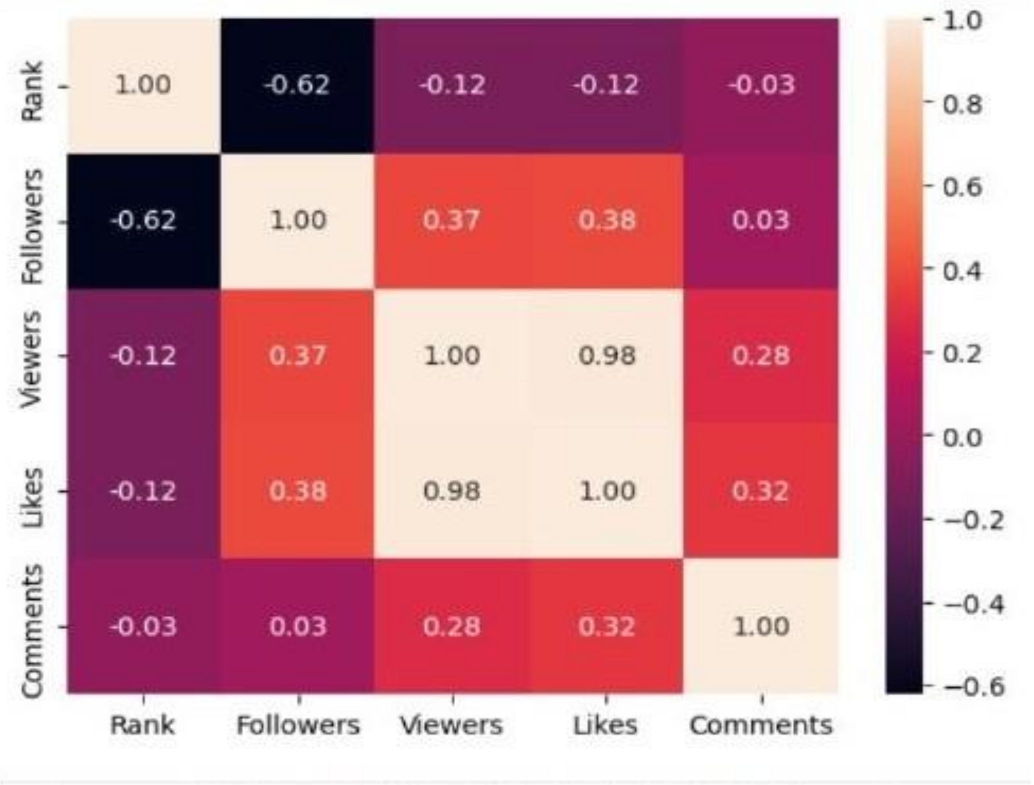




Insights:

- In both the graphs there is relation between (Comments vs Likes) and (Followers and Viewers).

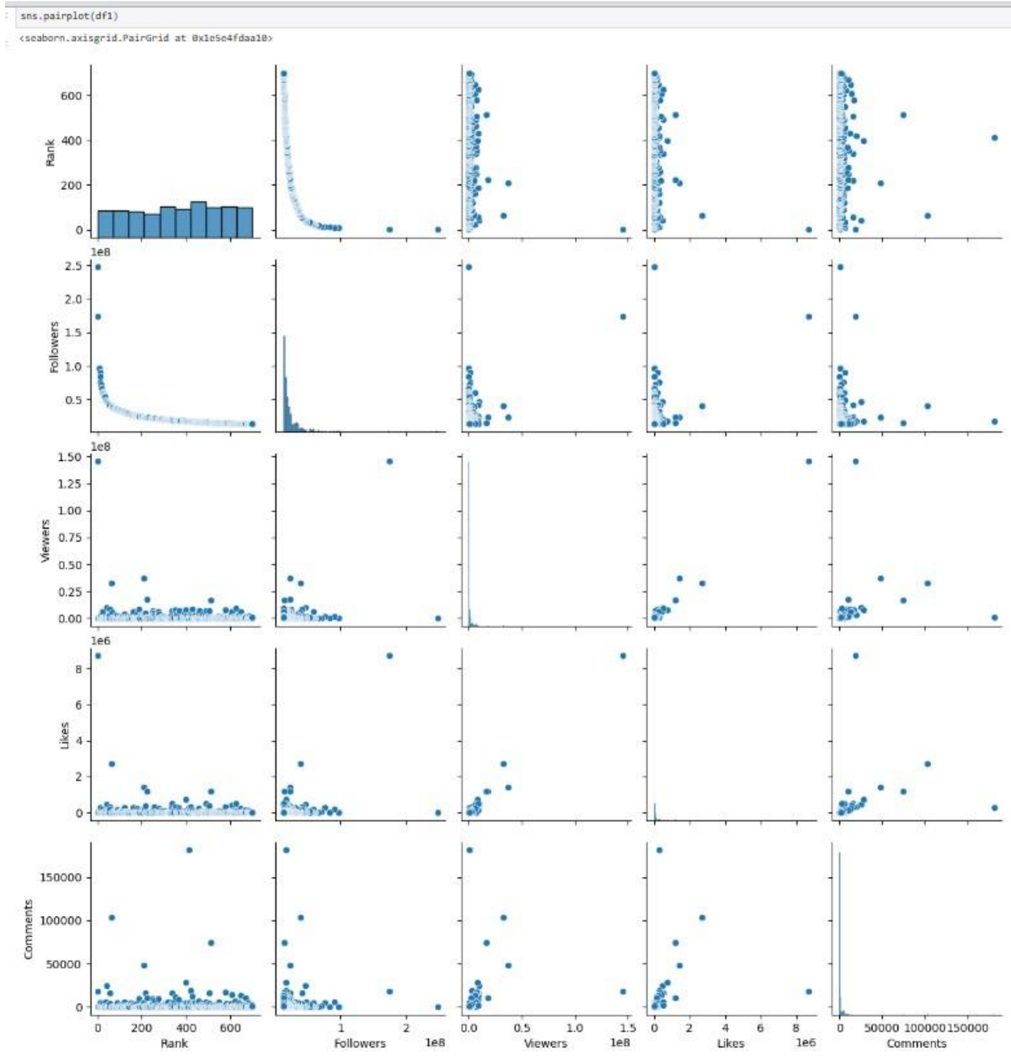
HEAT MAP:



Insights:

- By this Heat map we can describe that the co-relation between Rank, Followers, Viewers ,Likes ,Comments.
- Rank has max Negative co-relation , where as Followers, Viewers, Likes, Comments has max Positive co-relation.

PAIR PLOT :



```
sns.pairplot(df1)
```

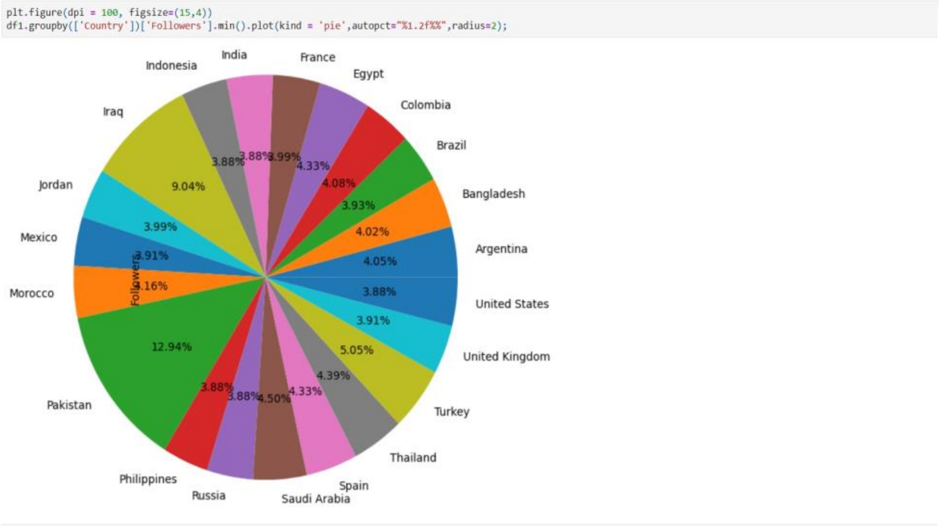
```
<seaborn.axisgrid.PairGrid at 0x1e5c4fdaa10>
```

- ❖ To pair plot multiple pair wise bivariate distributions in a data set, you can use the pair plot()function
- ❖ The diagonal plots are the univariate plots, and this displays the relationship for the (n,2) combination of variables in a Data Frame as a matrix of plots

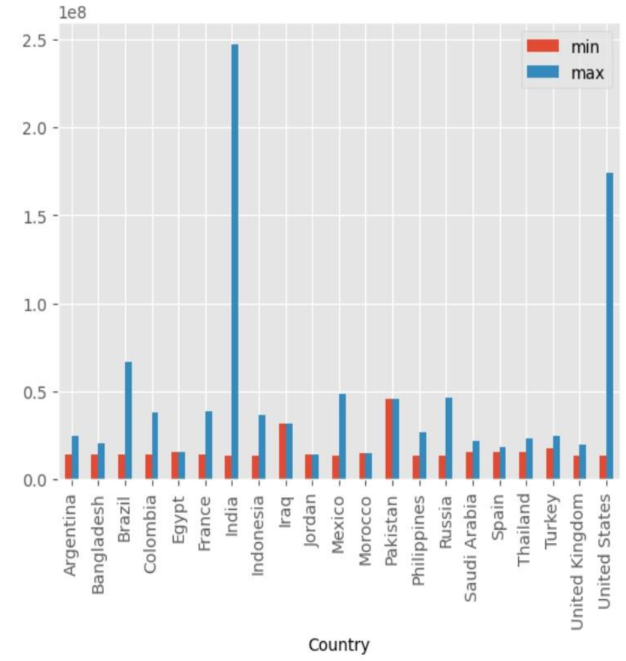
Insights:

- ❑ From the pair plot we can describe the completely about Followers, Comments, Rank, Viewers, Likes.

COUNTRY vs FOLLOWERS:

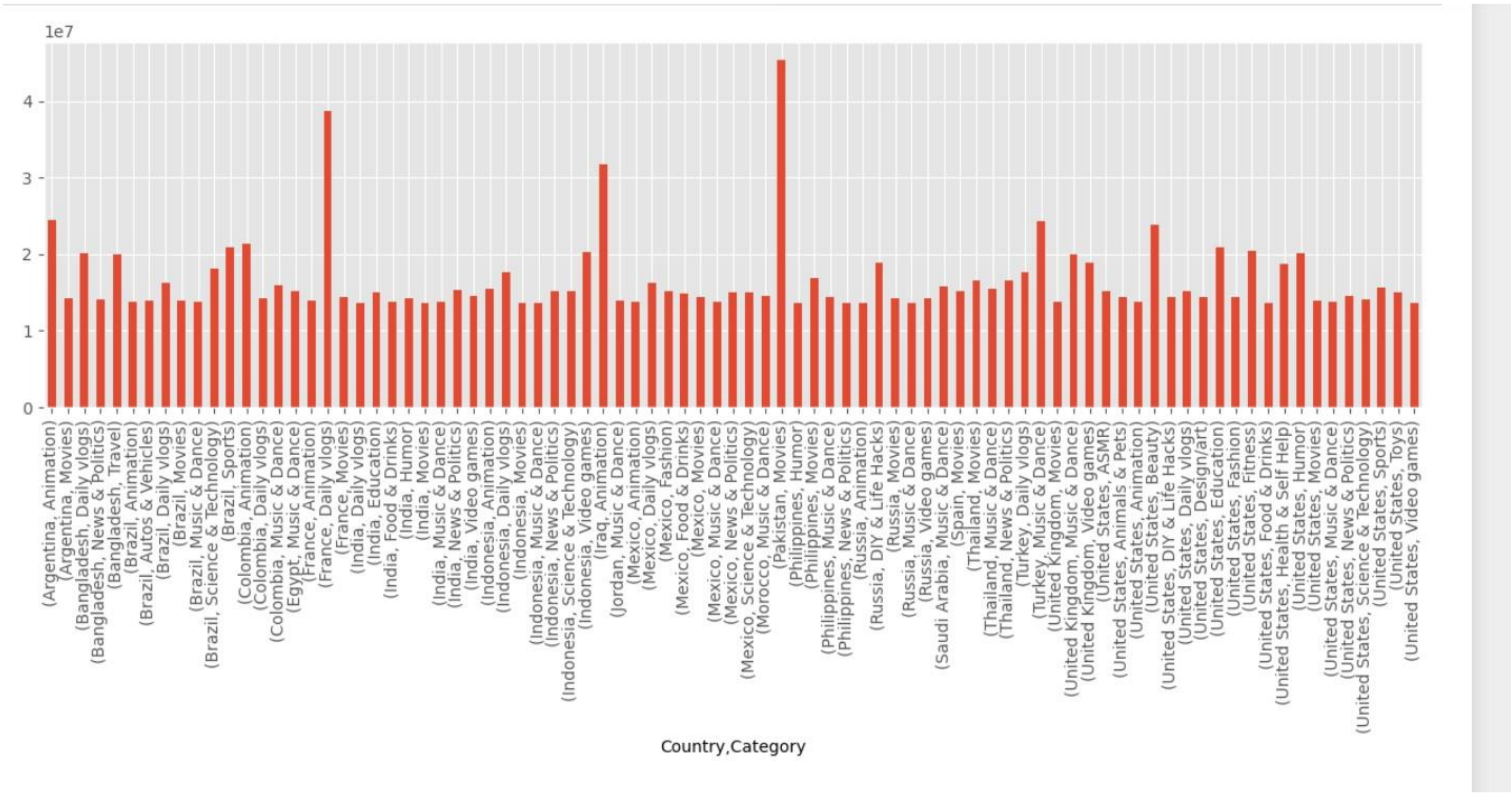


```
plt.figure(dpi = 100, figsize=(15,4))
df1.groupby(['Country'])['Followers'].min().plot(kind = 'pie',autopct="%1.2f%%",radius=2);
```



Insights : In this Charts we find the country wise followers percentage of and maximum and minimum values in Followers in Country wise.

COUNTRY.CATEGORY:



CHALLENGES FACED:

- While I was collecting data from the websites I had to go through each page and do separately.
- Every page has Null values and Data missing even through it is presenting web page .
- I had gone through each page and checked for any data present but is missing when extracted and added them manually using Python.
- If I had to replace more data I just have done any operation on that particular column.

Conclusion:

I here by conclude by sharing my observations in this particular project is:

- Here my problem statement was recommending highest Followers, Comments, viewers, Likes from different YouTube channels based on Ranking .

- Considering that to my observations I say that T-series channel Highest Ranking(1) as it consists of more number of Followers(24710000) and Likes(79600).

- By this we conclude that the Highest country based on channels is United States(157) have first place with category of Music & Dance(150).

- When we compared with Country vs Category vs Followers Pakistan took First with Category of Movies.

Thank you