YouTube Streamer Analysis

Introduction

This dataset contains valuable information about the Top 1000 YouTube streamers.

The dataset contains 9 columns namely:

1. Rank: Ranking of each account
2. Username: Account name of each Youtuber
3. Categories: Category of content created
4. Subscribers: Number of subscribers
5. Country: Country of the Youtubers
6. Visits: Number of Views
7. Likes: Numbers of Likes
8. Comments: Total numbers of comment
9. Links: link to the account

Task

The task is to perform a comprehensive analysis to extract insights about the top YouTube content creators.

In [1]:
```
# Import the necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

pd.options.display.float_format = '{:.0f}'.format # Suppress scientific notation
```

Data Exploration:

- Start by exploring the dataset to understand its structure and identify key variables.

- Check for missing data and outliers.

In [2]:
```
# Import the dataset
data = pd.read_csv("youtubers_df.csv")

# made a copy of the data
df = data.copy()

#check information about the data
df.info()

# view the data
df
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Rank        1000 non-null   int64  
 1   Username    1000 non-null   object 
 2   Categories  694 non-null    object 
 3   Suscribers  1000 non-null   float64
 4   Country     1000 non-null   object 
 5   Visits      1000 non-null   float64
 6   Likes       1000 non-null   float64
 7   Comments    1000 non-null   float64
 8   Links       1000 non-null   object 
dtypes: float64(4), int64(1), object(4)
memory usage: 70.4+ KB
```

Out[2]:

	Rank	Username	Categories	Suscribers	Country	Visits	Likes	Comments	Links
0	1	tseries	Música y baile	249500000	India	86200	2700	78	http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy...
1	2	MrBeast	Videojuegos, Humor	183500000	Estados Unidos	117400000	5300000	18500	http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
2	3	CoComelon	Educación	165500000	Unknown	7000000	24700	0	http://youtube.com/channel/UCbCmjCuTUZos6Inko4...
3	4	SETIndia	NaN	162600000	India	15600	166	9	http://youtube.com/channel/UCpEhnqL0y41EpW2TvW...
4	5	KidsDianaShow	Animación, Juguetes	113500000	Unknown	3900000	12400	0	http://youtube.com/channel/UCk8GzjMOrta8yxDcKf...
...	...	...	...	...	...	...	...	...	...
995	996	hamzymukbang	NaN	11700000	Estados Unidos	397400	14000	124	http://youtube.com/channel/UCPKNKldggioffXPkSm...
996	997	Adaahqueen	NaN	11700000	India	1100000	92500	164	http://youtube.com/channel/UCk3fFpqI5kDMf__mUP...
997	998	LittleAngelIndonesia	Música y baile	11700000	Unknown	211400	745	0	http://youtube.com/channel/UCdrHrQf0o0TO8YDntX...
998	999	PenMultiplex	NaN	11700000	India	14000	81	1	http://youtube.com/channel/UCObyBrdrtQ20BU9PxH...
999	1000	OneindiaHindi	Noticias y Política	11700000	India	2200	31	1	http://youtube.com/channel/UCOjgc1p2hJ4GZi6pQQ...

1000 rows × 9 columns

the Categories and Country values are not in english so i translate them to english

i thought of using the googletrans library but i keep getting a timeout error due to the large dataset so i used an alternaive approach: getting the unique data in each columns and using ChatGPT to translate and convert them into dictionary

In [3]:
```
categories_dict = {
    'Música y baile': 'Music and Dance',
    'Videojuegos, Humor': 'Video Games, Humor',
    'Educación': 'Education',
    'Unknown': 'Unknown',
    'Animación, Juguetes': 'Animation, Toys',
    'Películas, Videojuegos': 'Movies, Video Games',
```

```
    'Juguetes': 'Toys',
    'Videojuegos': 'Video Games',
    'Películas, Animación': 'Movies, Animation',
    'Películas': 'Movies',
    'Noticias y Política': 'News and Politics',
    'Animación, Humor': 'Animation, Humor',
    'Música y baile, Animación': 'Music and Dance, Animation',
    'Música y baile, Películas': 'Music and Dance, Movies',
    'Películas, Juguetes': 'Movies, Toys',
    'Películas, Humor': 'Movies, Humor',
    'Vlogs diarios': 'Daily Vlogs',
    'Videojuegos, Juguetes': 'Video Games, Toys',
    'Animación, Videojuegos': 'Animation, Video Games',
    'Animación': 'Animation',
    'Música y baile, Humor': 'Music and Dance, Humor',
    'Diseño/arte, DIY y Life Hacks': 'Design/Art, DIY and Life Hacks',
    'Ciencia y tecnología': 'Science and Technology',
    'Fitness, Salud y autoayuda': 'Fitness, Health and Self-help',
    'Belleza, Moda': 'Beauty, Fashion',
    'Humor': 'Humor',
    'Comida y bebida': 'Food and Drink',
    'Deportes': 'Sports',
    'Fitness': 'Fitness',
    'Viajes, Espectáculos': 'Travel, Entertainment',
    'Comida y bebida, Salud y autoayuda': 'Food and Drink, Health and Self-help',
    'Diseño/arte': 'Design/Art',
    'DIY y Life Hacks, Juguetes': 'DIY and Life Hacks, Toys',
    'Educación, Juguetes': 'Education, Toys',
    'Juguetes, Coches y vehículos': 'Toys, Cars and Vehicles',
    'Música y baile, Juguetes': 'Music and Dance, Toys',
    'Animales y mascotas': 'Animals and Pets',
    'ASMR': 'ASMR',
    'Moda': 'Fashion',
    'DIY y Life Hacks': 'DIY and Life Hacks',
    'Diseño/arte, Belleza': 'Design/Art, Beauty',
    'Coches y vehículos': 'Cars and Vehicles',
    'Animación, Humor, Juguetes': 'Animation, Humor, Toys',
    'ASMR, Comida y bebida': 'ASMR, Food and Drink',
    'Comida y bebida, Juguetes': 'Food and Drink, Toys',
    'Juguetes, DIY y Life Hacks': 'Toys, DIY and Life Hacks'
}


# Function to translate Spanish words to English
```

```
def categories_to_english(word):
    return categories_dict.get(word, word) # Return the translated word if it exists,


# Apply the translation function to the 'Category' column
df['Categories'] = df['Categories'].apply(categories_to_english)
```

In [4]:

```
countries_dict = {
    'India': 'India',
    'Estados Unidos': 'United States',
    'Unknown': 'Unknown',
    'Brasil': 'Brazil',
    'México': 'Mexico',
    'Rusia': 'Russia',
    'Pakistán': 'Pakistan',
    'Filipinas': 'Philippines',
    'Indonesia': 'Indonesia',
    'Tailandia': 'Thailand',
    'Francia': 'France',
    'Colombia': 'Colombia',
    'Iraq': 'Iraq',
    'Japón': 'Japan',
    'Ecuador': 'Ecuador',
    'Argentina': 'Argentina',
    'Turquía': 'Turkey',
    'Arabia Saudita': 'Saudi Arabia',
    'El Salvador': 'El Salvador',
    'Bangladesh': 'Bangladesh',
    'Reino Unido': 'United Kingdom',
    'Argelia': 'Algeria',
    'España': 'Spain',
    'Perú': 'Peru',
    'Egipto': 'Egypt',
    'Jordania': 'Jordan',
    'Marruecos': 'Morocco',
    'Singapur': 'Singapore',
    'Somalia': 'Somalia'
}

# Function to translate Spanish words to English
def countries_to_english(word):
    return countries_dict.get(word, word) # Return the translated word if it exist
```

```
# Apply the translation function to the 'Category' column
df['Country'] = df['Country'].apply(countries_to_english)
```

In [5]:
```
# View the first 5 rows of the now translated dataset
df.head()
```

Out[5]:

	Rank	Username	Categories	Suscribers	Country	Visits	Likes	Comments	Links
0	1	tseries	Music and Dance	249500000	India	86200	2700	78	http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy...
1	2	MrBeast	Video Games, Humor	183500000	United States	117400000	5300000	18500	http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
2	3	CoComelon	Education	165500000	Unknown	7000000	24700	0	http://youtube.com/channel/UCbCmjCuTUZos6Inko4...
3	4	SETIndia	NaN	162600000	India	15600	166	9	http://youtube.com/channel/UCpEhnqL0y41EpW2TvW...
4	5	KidsDianaShow	Animation, Toys	113500000	Unknown	3900000	12400	0	http://youtube.com/channel/UCk8GzjMOrta8yxDcKf...

In [6]:
```
# checking the numbers of unique values in each columns
df.nunique()
```

Out[6]:
```
Rank          1000
Username       994
Categories      45
Suscribers     274
Country         29
Visits         713
Likes          648
Comments       389
Links          994
dtype: int64
```

the username and links columns has some duplicated data.

There should be a unique link for each youtubers so we remove all duplicates

In [7]:
```
# check for duplicates using the username column
df[df['Username'].duplicated(keep=False)]
```

Out[7]:

	Rank	Username	Categories	Suscribers	Country	Visits	Likes	Comments	Links
249	250	NickyJamTV	Music and Dance, Movies	23700000	Colombia	15800	1200	58	http://youtube.com/channel/UCpb_iJuhFe8V6rQdbN...
250	251	NickyJamTV	Music and Dance, Movies	23700000	Colombia	15800	1200	58	http://youtube.com/channel/UCpb_iJuhFe8V6rQdbN...
447	448	mgcplayhouse	Toys, Cars and Vehicles	17800000	Unknown	56300	96	0	http://youtube.com/channel/UC6zPzUJo8hu-5TzUk8...
449	450	thexoteam	NaN	17800000	United States	797600	50400	179	http://youtube.com/channel/UClZAOlfhJQJRym39Wo...
450	451	thexoteam	NaN	17900000	United States	772800	45000	185	http://youtube.com/channel/UClZAOlfhJQJRym39Wo...
451	452	mgcplayhouse	Toys, Cars and Vehicles	17800000	Unknown	63600	75	0	http://youtube.com/channel/UC6zPzUJo8hu-5TzUk8...
946	947	Super_Senya_RU	Animation, Toys	12100000	Unknown	47200	192	0	http://youtube.com/channel/UCTn9Vyy-3fzLIr0bqh...
947	948	HiTechIslamic	Music and Dance	12100000	Pakistan	62200	810	59	http://youtube.com/channel/UCtKKyuORzErSd7TWfk...
949	950	Family-Box	Movies	12000000	Russia	173600	6600	105	http://youtube.com/channel/UC-jHNWViReG6R_kJ6b...
952	953	Super_Senya_RU	Animation, Toys	12100000	Unknown	47200	192	0	http://youtube.com/channel/UCTn9Vyy-3fzLIr0bqh...
953	954	HiTechIslamic	Music and Dance	12100000	Pakistan	62200	810	59	http://youtube.com/channel/UCtKKyuORzErSd7TWfk...
956	957	Family-Box	Movies	12000000	Russia	177400	6300	86	http://youtube.com/channel/UC-jHNWViReG6R_kJ6b...

In [8]:
```
# drop all duplicates
df.drop_duplicates(subset=['Username'], inplace=True, ignore_index=True)

# check again to confirm the dataset is no more having duplicate data
df['Username'].duplicated().sum()
```

Out[8]:
```
0
```

In [9]:
```
# the rank column is needed for the analysis so i'll drop it
df.drop(columns='Rank', inplace=True)

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 994 entries, 0 to 993
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Username    994 non-null    object
 1   Categories  689 non-null    object
 2   Suscribers  994 non-null    float64
 3   Country     994 non-null    object
 4   Visits      994 non-null    float64
 5   Likes       994 non-null    float64
 6   Comments    994 non-null    float64
 7   Links       994 non-null    object
dtypes: float64(4), object(4)
memory usage: 62.2+ KB
```

we are left with 994rows and 7cols, notice only the categories column is having missing values so i deal with the missing data and also correct the subscribers column name

In [10]:
```
# check the total missing values
df['Categories'].isna().sum()
```

Out[10]:
```
305
```

305 missing values, about 30% of the dataset, dropping them will reduce the dataset drastically and this may affect the accuracy of our analysis so i'll fill it with "Unknown"

In [11]:
```
df['Categories'].fillna('Unknown', inplace=True)

#check to confirm no more missing values in the dataset
df.isna().sum().sum()
```

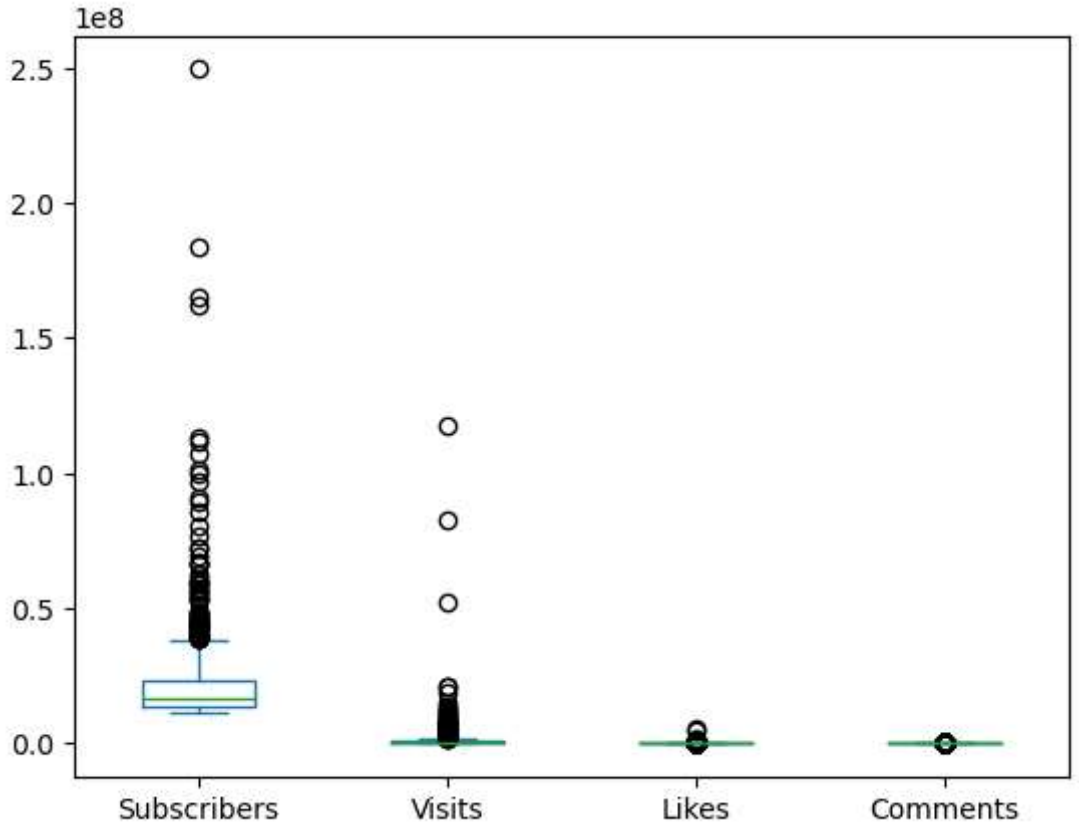
Out[11]:
```
0
```

In [12]:
```
df.rename(columns={'Suscribers': 'Subscribers'}, inplace=True)

df.head()
```

Out[12]:

	Username	Categories	Subscribers	Country	Visits	Likes	Comments	Links
0	tseries	Music and Dance	249500000	India	86200	2700	78	http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy...
1	MrBeast	Video Games, Humor	183500000	United States	117400000	5300000	18500	http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
2	CoComelon	Education	165500000	Unknown	7000000	24700	0	http://youtube.com/channel/UCbCmjCuTUZos6Inko4...
3	SETIndia	Unknown	162600000	India	15600	166	9	http://youtube.com/channel/UCpEhnqL0y41EpW2TvW...
4	KidsDianaShow	Animation, Toys	113500000	Unknown	3900000	12400	0	http://youtube.com/channel/UCk8GzjMOrta8yxDcKf...

In [13]:
```
#checking for outliers we construct a boxplot

df.plot(kind='box');
```



there are hugh amount of outliers, this is understandable because content creator followers varies based on how their contents gains attraction.

Trend Analysis:

- Identify trends among the top YouTube streamers. Which categories are the most popular?
- Is there a correlation between the number of subscribers and the number of likes or comments?

In [14]:
```
# we identify trends among the top youtube streamers by the number of subscribers, visit, and likes

# top 10 by numbers of subscribers
display(df.groupby('Categories')[['Subscribers']].sum().sort_values(by='Subscribers', ascending=False).head(10))

# top 10 by numbers of visits
display(df.groupby('Categories')[['Visits']].sum().sort_values(by='Visits', ascending=False).head(10))

# top 10 by numbers of likes
df.groupby('Categories')[['Likes']].sum().sort_values(by='Likes', ascending=False).head(10)
```

Categories	Subscribers
Unknown	6320100000
Music and Dance	4281800000
Movies, Animation	1384300000
Animation, Toys	839800000
Music and Dance, Movies	774800000
News and Politics	676100000
Animation, Video Games	659400000
Daily Vlogs	654900000
Movies, Humor	622100000
Education	600300000

Categories	Visits
Unknown	368427116
Video Games, Humor	174074500
Daily Vlogs	126330500
Animation, Humor	101523400
Music and Dance	59839900
Animation, Video Games	40802000
Movies, Animation	33631100
Food and Drink	32669400
Movies, Humor	31916600
Education	26545000

Out[14]:

Categories	Likes
Unknown	16476915
Video Games, Humor	7148700
Daily Vlogs	6928063
Animation, Humor	3935745
Music and Dance	2784099
Animation, Video Games	2695997
Humor	1699900
Movies, Animation	1565932
Food and Drink	1543977
Movies, Humor	1383277

Since the "Unknown" is the missing values,

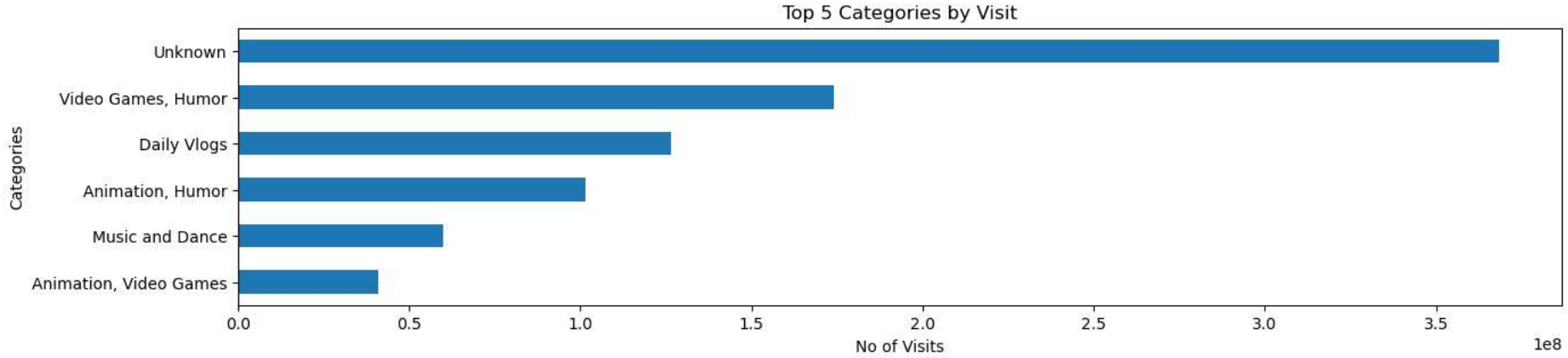
the top 5 most popular categories are similar by numbers of Visits and Likes which are:

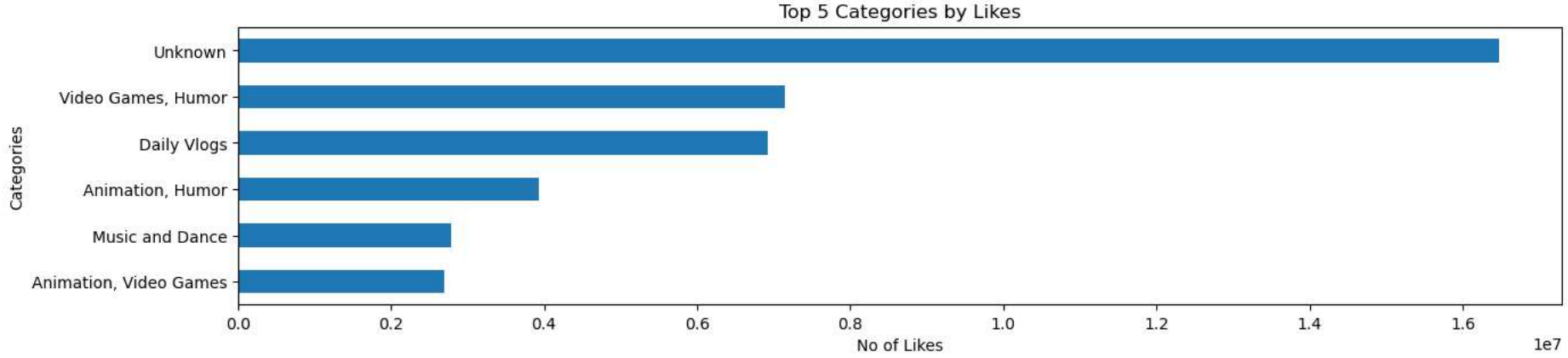
1. Video Games, Humor
2. Daily Vlogs
3. Animation, Humor
4. Music and Dance
5. Animation, Video Games

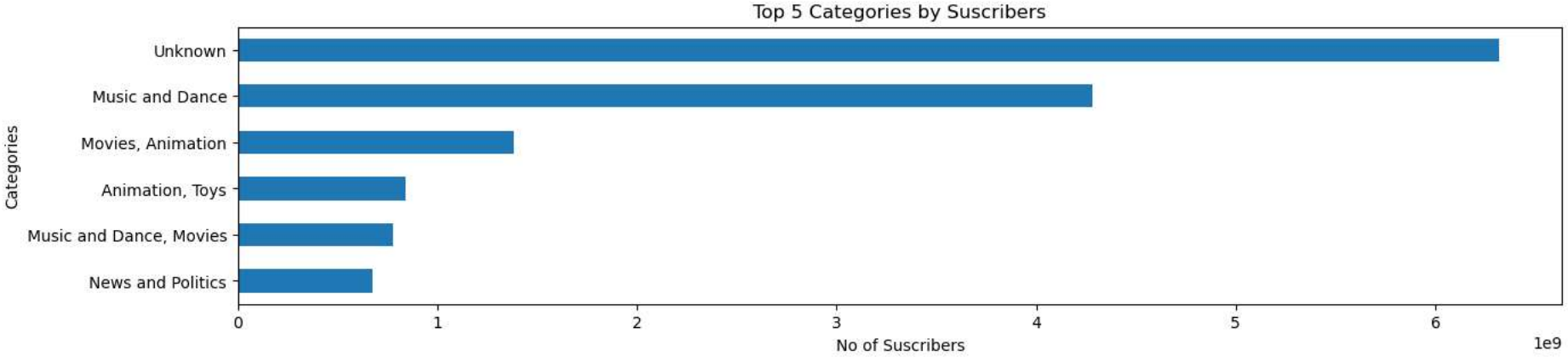
but the case is different for the top categories by numbers of subscribers which are:

1. Music and Dance
2. Movies, Animation
3. Animation, Toys
4. Music and Dance, Movies
5. News and Politics

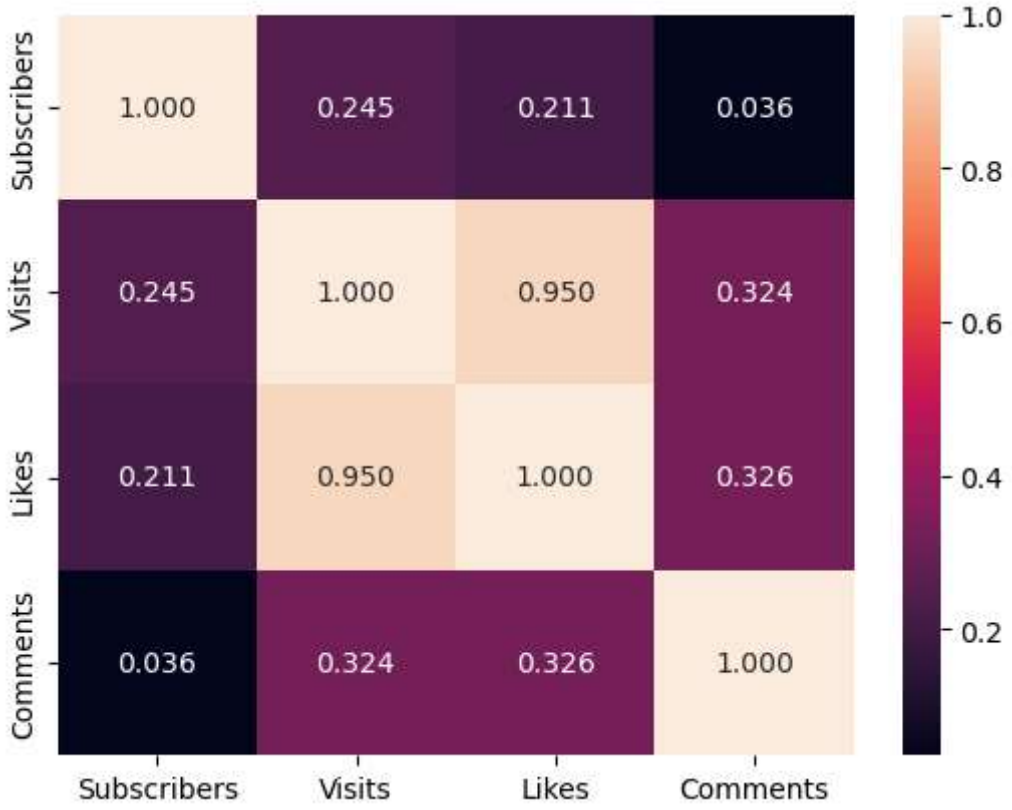
In [15]:
```
fig, axs = plt.subplots(figsize=(15,12), nrows=3, ncols=1)
plt.subplots_adjust(hspace=0.4)
df.groupby('Categories')['Visits'].sum().sort_values().tail(6).plot(kind='barh', ax=axs[0])
axs[0].set_title("Top 5 Categories by Visit")
axs[0].set_xlabel("No of Visits")

df.groupby('Categories')['Likes'].sum().sort_values().tail(6).plot(kind='barh', ax=axs[1])
axs[1].set_title("Top 5 Categories by Likes")
axs[1].set_xlabel("No of Likes")

df.groupby('Categories')['Subscribers'].sum().sort_values().tail(6).plot(kind='barh', ax=axs[2])
axs[2].set_title("Top 5 Categories by Suscribers")
axs[2].set_xlabel("No of Suscribers")

plt.show()
```


Top 5 Categories by Visit
Unknown
Video Games, Humor
Daily Vlogs
Animation, Humor
Music and Dance
Animation, Video Games
Categories
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
No of Visits
1e8


Top 5 Categories by Likes
Unknown
Video Games, Humor
Daily Vlogs
Animation, Humor
Music and Dance
Animation, Video Games
Categories
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
No of Likes
1e7


Top 5 Categories by Suscribers
Unknown
Music and Dance
Movies, Animation
Animation, Toys
Music and Dance, Movies
News and Politics
Categories
0
1
2
3
4
5
6
No of Suscribers
1e9

In [16]:
```
# to check for correlation consider the heatmap:
sns.heatmap(df.select_dtypes('number').corr(), annot=True, fmt='.3f');
```



Observations:

There is no relationship between the Subscribers, Likes and Comment but there is a strong relationship between Visits and Likes

Audience Study:

- Analyze the distribution of streamers' audiences by country. Are there regional preferences for specific content categories?

In [17]:
```
df.head()
```

Out[17]:

	Username	Categories	Subscribers	Country	Visits	Likes	Comments	Links
0	tseries	Music and Dance	249500000	India	86200	2700	78	http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy...
1	MrBeast	Video Games, Humor	183500000	United States	117400000	5300000	18500	http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
2	CoComelon	Education	165500000	Unknown	7000000	24700	0	http://youtube.com/channel/UCbCmjCuTUZos6Inko4...
3	SETIndia	Unknown	162600000	India	15600	166	9	http://youtube.com/channel/UCpEhnqL0y41EpW2TvW...
4	KidsDianaShow	Animation, Toys	113500000	Unknown	3900000	12400	0	http://youtube.com/channel/UCk8GzjMOrta8yxDcKf...

Notice "Unknown" Values in both the Categories and Countries columns i filter them out

In [18]:
```
# removing the rows with "Unknown" values in both the Categories and Country column
df2 = df[(df['Categories'].str.contains("Unknown") != True) & (df['Country'].str.contains("Unknown") != True)]

df2.head()
```

Out[18]:

	Username	Categories	Subscribers	Country	Visits	Likes	Comments	Links
0	tseries	Music and Dance	249500000	India	86200	2700	78	http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy...
1	MrBeast	Video Games, Humor	183500000	United States	117400000	5300000	18500	http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
5	PewDiePie	Movies, Video Games	111500000	United States	2400000	197300	4900	http://youtube.com/channel/UC-IHJZR3Gqxm24_Vd_...
8	zeemusiccompany	Music and Dance	99700000	India	74300	2600	32	http://youtube.com/channel/UCFFbwnve3yF62-tVXk...
9	WWE	Video Games	97200000	United States	184500	6300	214	http://youtube.com/channel/UCJ5v_MCY6GNUBTO8-D...

In [19]:
```
# region preference by numbers of subscribers
display(df2.groupby(['Country', 'Categories'])[['Subscribers']].sum().sort_values(by=['Country', 'Subscribers'], ascending=False))

# region preference by numbers of visits
display(df2.groupby(['Country', 'Categories'])[['Visits']].sum().sort_values(by=['Country', 'Visits'], ascending=False))

# region preference by numbers of Likes
df2.groupby(['Country', 'Categories'])[['Likes']].sum().sort_values(by=['Country', 'Likes'], ascending=False)
```

		Subscribers
Country	**Categories**	
United States	**Music and Dance**	1425800000
	Video Games, Humor	437300000
	Animation, Video Games	393200000
	Animation, Humor	334000000
	Video Games	302100000
...	**...**	...
Argentina	**Movies, Animation**	34400000
	Animation	25700000
	Movies, Humor	18800000
	Movies	14200000
Algeria	**Education**	12200000

151 rows × 1 columns

		Visits
Country	**Categories**	
United States	**Video Games, Humor**	159509700
	Daily Vlogs	86443400
	Animation, Humor	75702600
	Music and Dance	31250600
	Food and Drink	30352400
...	**...**	...
Argentina	**Animation**	5700000
	Movies, Humor	5600000
	Movies, Animation	594800
	Movies	76600
Algeria	**Education**	333500

151 rows × 1 columns

Out[19]:

Country	Categories	Likes
United States	**Video Games, Humor**	6450500
	Daily Vlogs	5126400
	Animation, Humor	2584706
	Animation, Video Games	1625202
	Food and Drink	1449177
...	**...**	...
Argentina	**Movies, Humor**	401700
	Animation	208400
	Movies, Animation	28100
	Movies	1200
Algeria	**Education**	16000

151 rows × 1 columns

There are regional preferences for specific content categories and it's different based on the Numbers of Subscribers, Visits and Likes

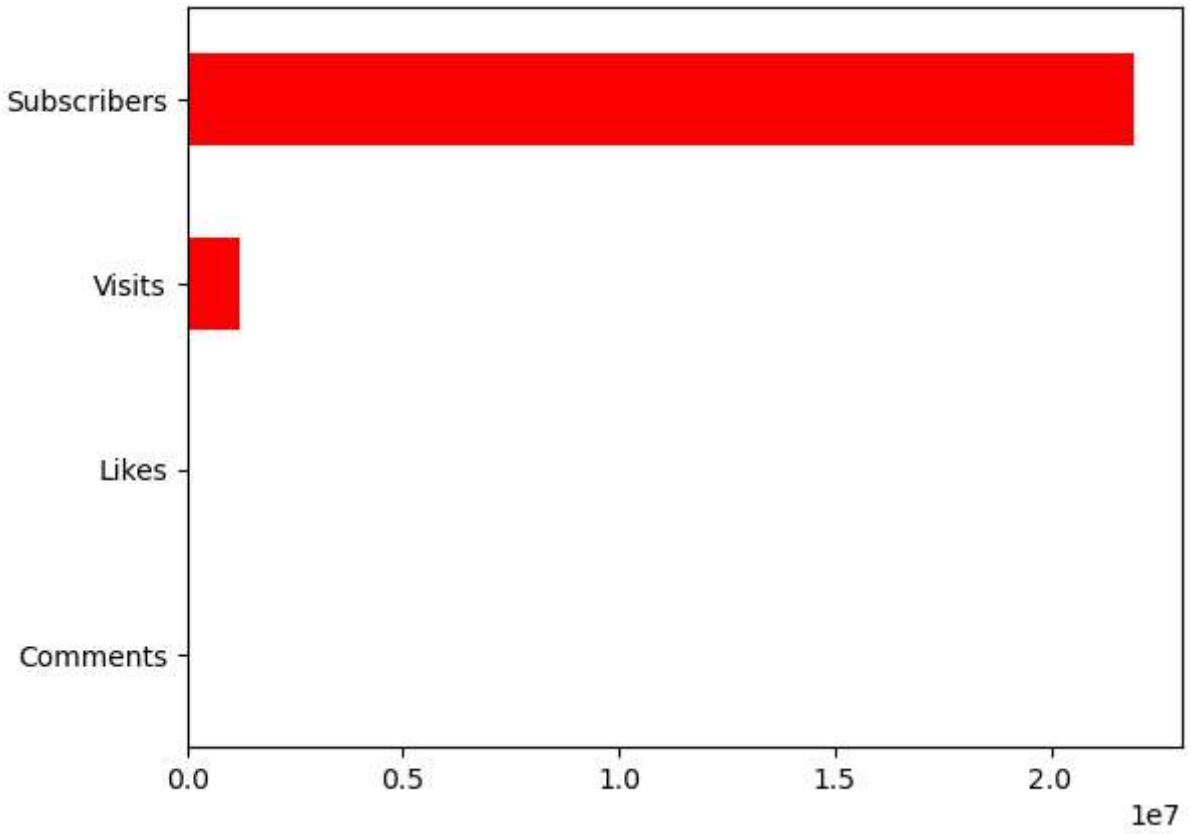
Performance Metrics:

- Calculate and visualize the average number of subscribers, visits, likes, and comments.
- Are there patterns or anomalies in these metrics?

In [20]:
```
df.select_dtypes('number').mean()
```

Out[20]:
```
Subscribers   21930382
Visits         1215601
Likes            53902
Comments          1296
dtype: float64
```

In [21]:
```
df.select_dtypes('number').mean().sort_values().plot(kind='barh', color='r');
```



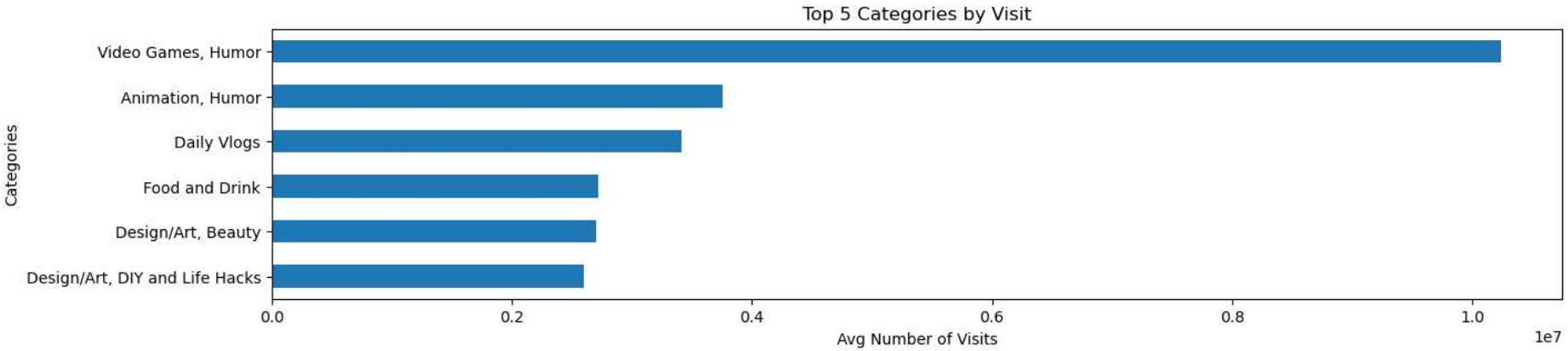
There is an anomalies, the average number of Subscribers is greater than the average number of Visits and Likes.

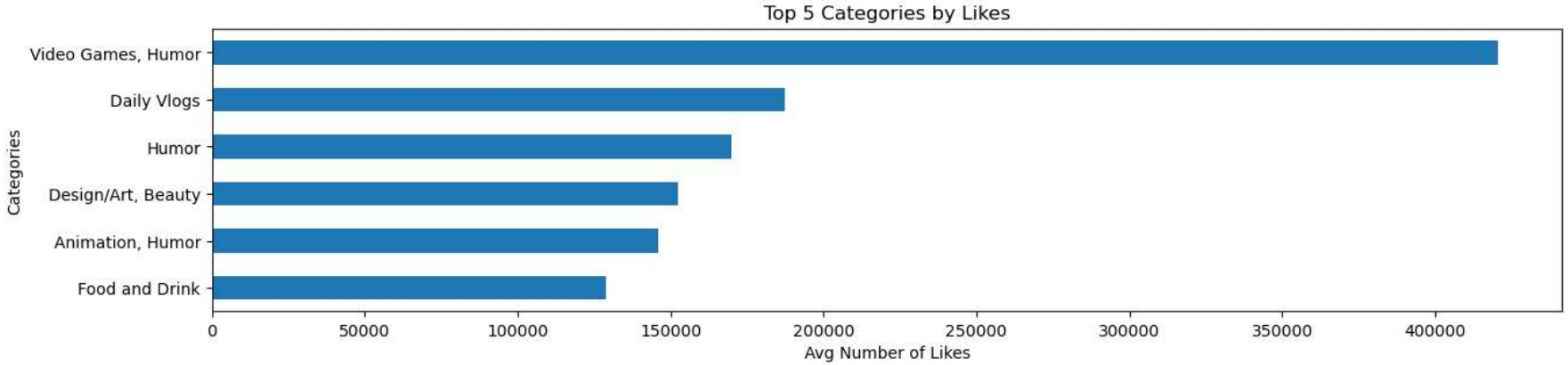
Content Categories:

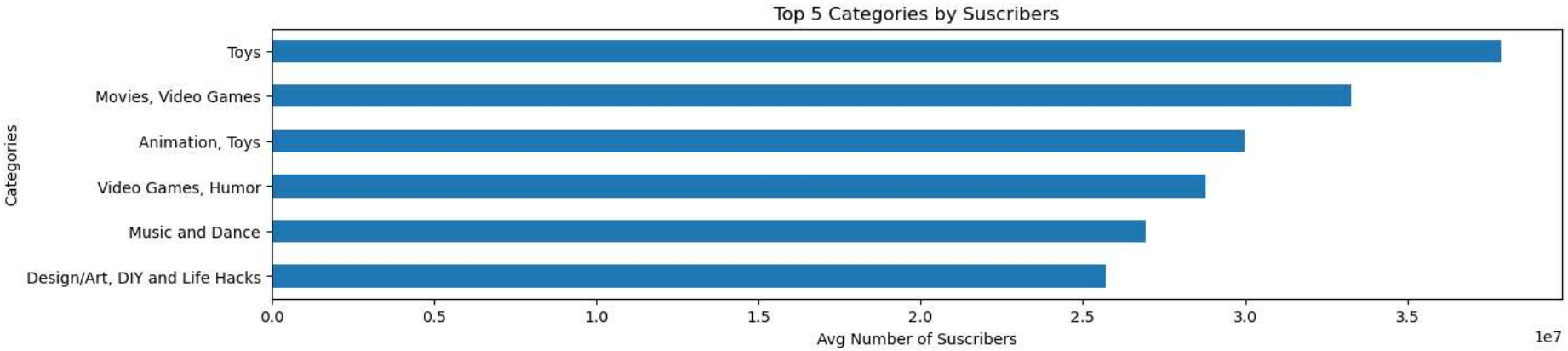
- Explore the distribution of content categories. Which categories have the highest number of streamers?
- Are there specific categories with exceptional performance metrics?

In [22]:
```
fig, axs = plt.subplots(figsize=(15,12), nrows=3, ncols=1)
plt.subplots_adjust(hspace=0.4)
df.groupby('Categories')['Visits'].mean().sort_values().tail(6).plot(kind='barh', ax=axs[0])
axs[0].set_title("Top 5 Categories by Visit")
```

```
axs[0].set_xlabel("Avg Number of Visits")

df.groupby('Categories')['Likes'].mean().sort_values().tail(6).plot(kind='barh', ax=axs[1])
axs[1].set_title("Top 5 Categories by Likes")
axs[1].set_xlabel("Avg Number of Likes")

df.groupby('Categories')['Subscribers'].mean().sort_values().tail(6).plot(kind='barh', ax=axs[2])
axs[2].set_title("Top 5 Categories by Suscribers")
axs[2].set_xlabel("Avg Number of Suscribers")

plt.show()
```



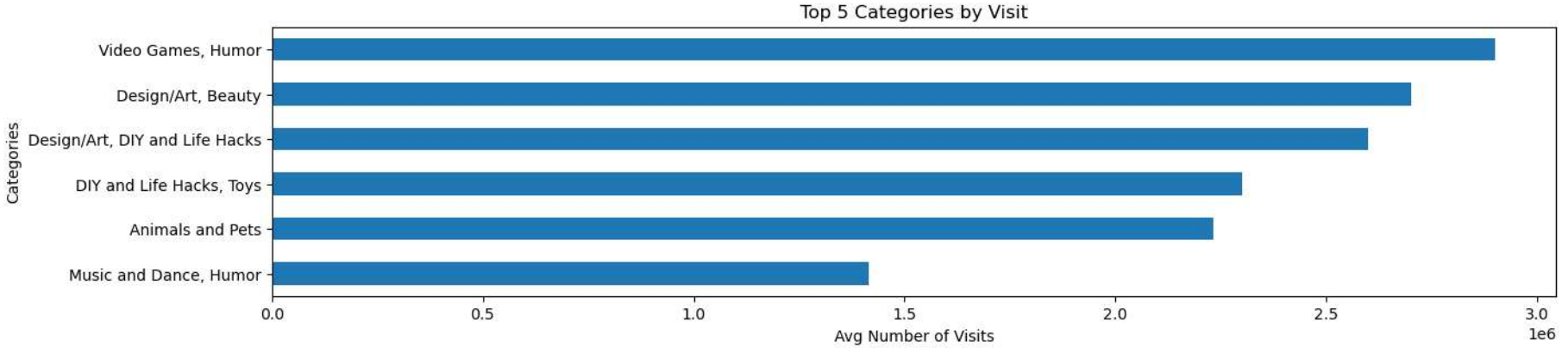


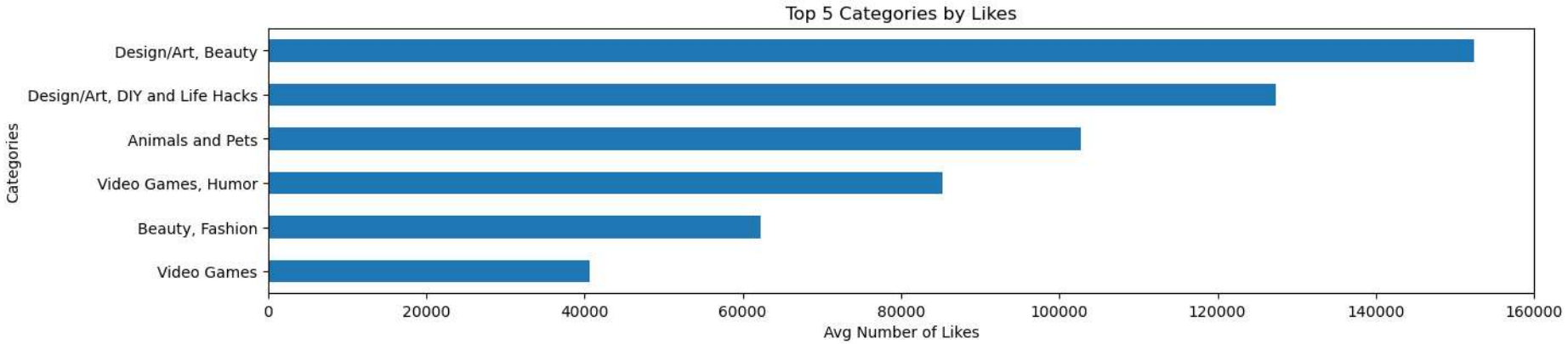


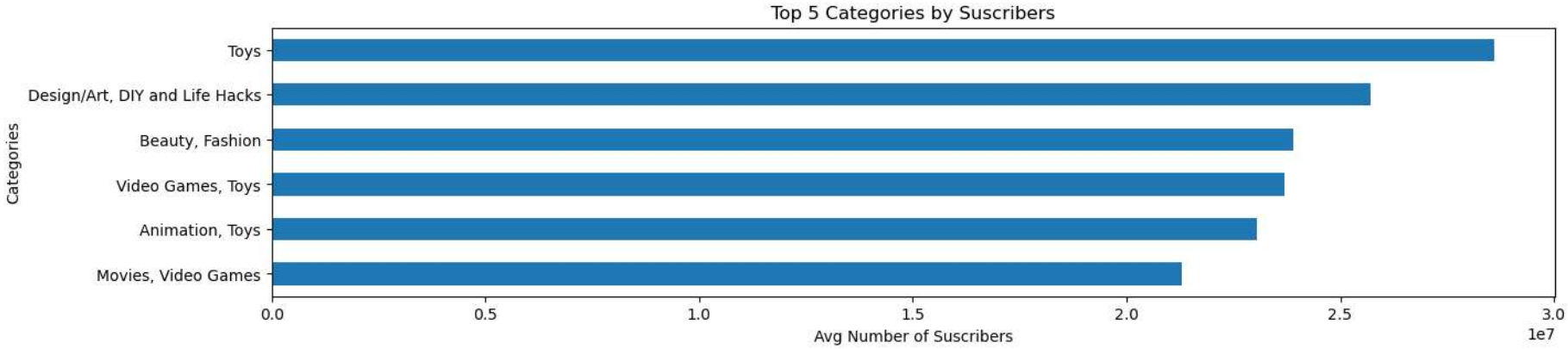
Notice how the Top 6 Categories varies by number of Visits, Likes and Subscribers. The "Video Gaames, Humor" is consistent at the top by number of Visits and Likes.

But from our data exploration there are outliers in the data. So i'll visualize the Top 6 using the "Median"

In [23]:
```
fig, axs = plt.subplots(figsize=(15,12), nrows=3, ncols=1)
plt.subplots_adjust(hspace=0.4)
df.groupby('Categories')['Visits'].median().sort_values().tail(6).plot(kind='barh', ax=axs[0])
axs[0].set_title("Top 5 Categories by Visit")
axs[0].set_xlabel("Avg Number of Visits")

df.groupby('Categories')['Likes'].median().sort_values().tail(6).plot(kind='barh', ax=axs[1])
axs[1].set_title("Top 5 Categories by Likes")
axs[1].set_xlabel("Avg Number of Likes")

df.groupby('Categories')['Subscribers'].median().sort_values().tail(6).plot(kind='barh', ax=axs[2])
axs[2].set_title("Top 5 Categories by Suscribers")
axs[2].set_xlabel("Avg Number of Suscribers")

plt.show()
```







using the Median i discover that the Top Categories varies also but taking a closer look, it all contains the same Categories but ranking differently except for the numbers of Subscribers where the Toys appears for the first time among the Top 6, eliminating: Animal, Pets and Humor from the

list.

In [24]:
```
# To know the specific categories with exceptional performance metrics i obtain the top 10% across all metrics
df2.select_dtypes('number').quantile(0.90)
```

Out[24]:
```
Subscribers   34200000
Visits         2800000
Likes           122820
Comments          3920
Name: 0.9, dtype: float64
```

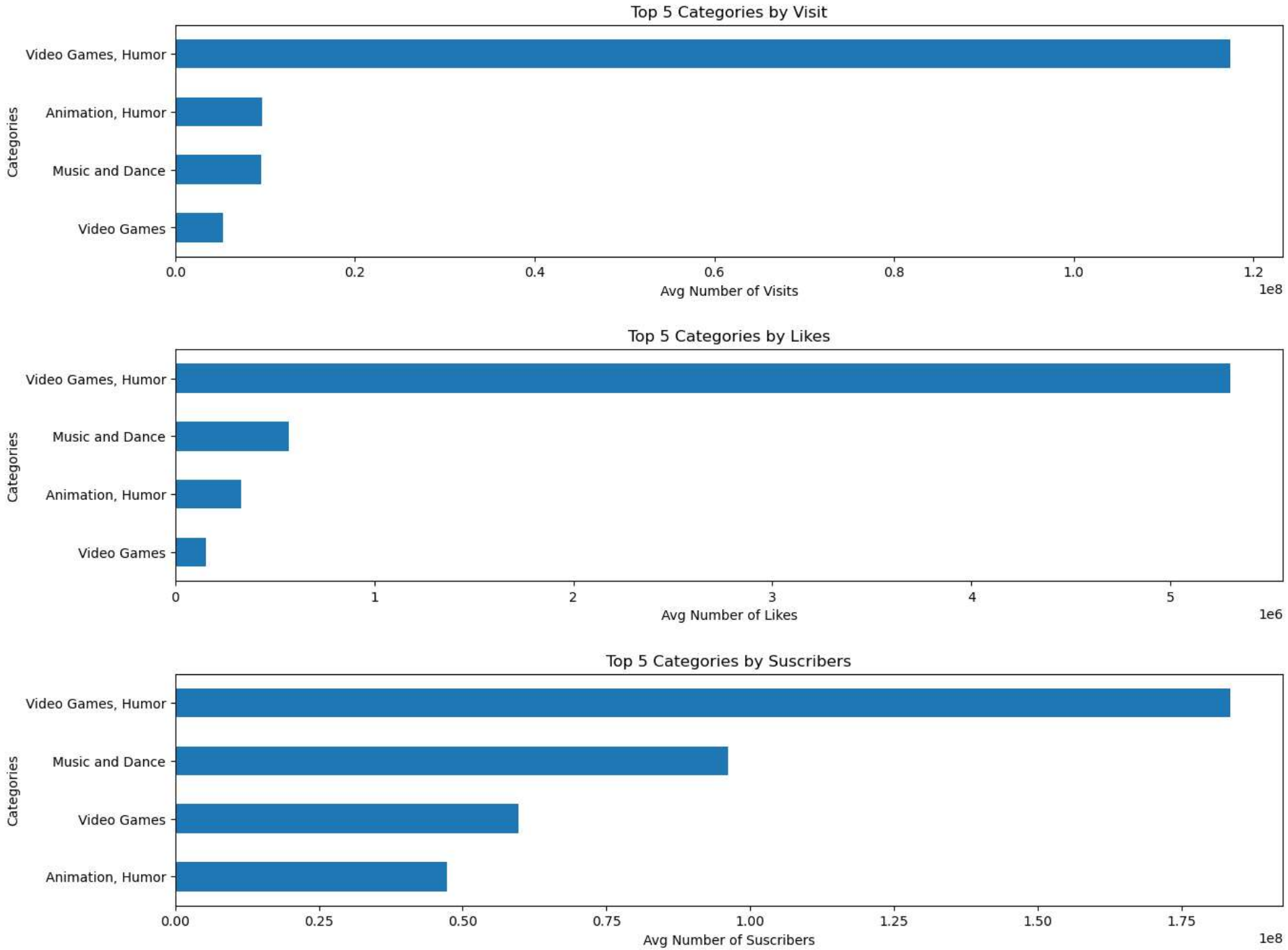
In [25]:
```
subscribers, visit, likes, comments = round(df2.select_dtypes('number').quantile(0.90))

df_top_cat = df2[(df2['Subscribers'] > subscribers) & (df2['Visits'] > visit) & (df2['Likes'] > likes) & (df2['Comments'] > comme


display(df_top_cat)

print("===" * 38)
print(f"Specific Categories with exceptional performance metrics are: \n\n           {df_top_cat['Categories'].unique()}")
print("===" * 38)
```

	Username	Categories	Subscribers	Country	Visits	Likes	Comments	Links
1	MrBeast	Video Games, Humor	183500000	United States	117400000	5300000	18500	http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
26	dudeperfect	Video Games	59700000	United States	5300000	156500	4200	http://youtube.com/channel/UCRijo3ddMTht_IHyNS...
34	TaylorSwift	Music and Dance	54100000	United States	4300000	300400	15000	http://youtube.com/channel/UCqECaJ8Gagnn7YCbPE...
43	A4a4a4a4	Animation, Humor	47300000	Russia	9700000	330400	22000	http://youtube.com/channel/UC2tsySbe9TNrI-xh2I...
62	KimberlyLoaiza	Music and Dance	42100000	Mexico	5300000	271300	16000	http://youtube.com/channel/UCQZfFRohQ7UX-0CdXI...

```
==========================================================================================================================
Specific Categories with exceptional performance metrics are:

           ['Video Games, Humor' 'Video Games' 'Music and Dance' 'Animation, Humor']
==========================================================================================================================
```

In [26]:

```
fig, axs = plt.subplots(figsize=(15,12), nrows=3, ncols=1)
plt.subplots_adjust(hspace=0.4)
df_top_cat.groupby('Categories')['Visits'].sum().sort_values().plot(kind='barh', ax=axs[0])
axs[0].set_title("Top 5 Categories by Visit")
axs[0].set_xlabel("Avg Number of Visits")

df_top_cat.groupby('Categories')['Likes'].sum().sort_values().plot(kind='barh', ax=axs[1])
axs[1].set_title("Top 5 Categories by Likes")
axs[1].set_xlabel("Avg Number of Likes")

df_top_cat.groupby('Categories')['Subscribers'].sum().sort_values().plot(kind='barh', ax=axs[2])
axs[2].set_title("Top 5 Categories by Suscribers")
axs[2].set_xlabel("Avg Number of Suscribers")

plt.show()
```


Top 5 Categories by Visit
Video Games, Humor
Animation, Humor
Music and Dance
Video Games
Categories
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Avg Number of Visits
1e8
Top 5 Categories by Likes
Video Games, Humor
Music and Dance
Animation, Humor
Video Games
Categories
0
1
2
3
4
5
Avg Number of Likes
1e6
Top 5 Categories by Suscribers
Video Games, Humor
Music and Dance
Video Games
Animation, Humor
Categories
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Avg Number of Suscribers
1e8

Benchmarking:

- Identify streamers with above-average performance in terms of subscribers, visits, likes, and comments.
- Who are the top-performing content creators?

In [27]:
```
df.select_dtypes('number').mean()
```

Out[27]:
```
Subscribers   21930382
Visits         1215601
Likes            53902
Comments          1296
dtype: float64
```

In [28]:
```
sub_mean, visit_mean, likes_mean, comments_mean = round(df.select_dtypes('number').mean())

df_top = df[(df['Subscribers'] > sub_mean) & (df['Visits'] > visit_mean) & (df['Likes'] > likes_mean) & (df['Comments'] > comment

# Top performing content creators by Subscribers
df_top_sub = df_top[['Username', 'Subscribers']].nlargest(6, 'Subscribers')
display(df_top_sub)
print("===" * 38)
print(f"Top 6 content creators by numbers of subscribers are: \n      {df_top_sub['Username'].unique().tolist()}", end=(' '))

# Top performing content creators by Visits
df_top_visit = df_top[['Username', 'Visits']].nlargest(6, 'Visits')
display(df_top_visit)
print("===" * 38)
print(f"Top 6 content creators are by numbers of visit are: \n      {df_top_visit['Username'].unique().tolist()}", end=(' '))

# Top performing content creators by Likes
df_top_likes = df_top[['Username', 'Likes']].nlargest(6, 'Likes')
display(df_top_likes)
print("===" * 38)
print(f"Top 6 content creators are by numbers of likes are: \n      {df_top_likes['Username'].unique().tolist()}", end=(' '))
```

	Username	Subscribers
1	MrBeast	183500000
5	PewDiePie	111500000
26	dudeperfect	59700000
34	TaylorSwift	54100000
39	JuegaGerman	48600000
43	A4a4a4a4	47300000

```
==========================================================================================================================
Top 6 content creators by numbers of subscribers are:
      ['MrBeast', 'PewDiePie', 'dudeperfect', 'TaylorSwift', 'JuegaGerman', 'A4a4a4a4']
```

	Username	Visits
1	MrBeast	117400000
136	MrBeast2	83100000
153	DaFuqBoom	52700000
287	VillageCookingChannel	21500000
277	StokesTwins	11700000
43	A4a4a4a4	9700000

```
==========================================================================================================================
Top 6 content creators are by numbers of visit are:
      ['MrBeast', 'MrBeast2', 'DaFuqBoom', 'VillageCookingChannel', 'StokesTwins', 'A4a4a4a4']
```

	Username	Likes
1	MrBeast	5300000
136	MrBeast2	5000000
153	DaFuqBoom	1700000
123	MRINDIANHACKER	617400
238	alanbecker	582600
131	fedevigevani	412200

```
================================================================================================================================
Top 6 content creators are by numbers of likes are:
      ['MrBeast', 'MrBeast2', 'DaFuqBoom', 'MRINDIANHACKER', 'alanbecker', 'fedevigevani']
```

In [29]:
```
df_top
```

Out[29]:

	Username	Categories	Subscribers	Country	Visits	Likes	Comments	Links
1	MrBeast	Video Games, Humor	183500000	United States	117400000	5300000	18500	http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
5	PewDiePie	Movies, Video Games	111500000	United States	2400000	197300	4900	http://youtube.com/channel/UC-lHJZR3Gqxm24_Vd_...
26	dudeperfect	Video Games	59700000	United States	5300000	156500	4200	http://youtube.com/channel/UCRijo3ddMTht_IHyNS...
34	TaylorSwift	Music and Dance	54100000	United States	4300000	300400	15000	http://youtube.com/channel/UCqECaJ8Gagnn7YCbPE...
39	JuegaGerman	Movies, Animation	48600000	Mexico	2000000	117100	3000	http://youtube.com/channel/UCYiGq8XF7YQD00x7wA...
43	A4a4a4a4	Animation, Humor	47300000	Russia	9700000	330400	22000	http://youtube.com/channel/UC2tsySbe9TNrI-xh2I...
58	Mikecrack	Movies, Animation	43400000	Mexico	2200000	183400	1800	http://youtube.com/channel/UCqJ5zFEED1hWs0KNQC...
62	KimberlyLoaiza	Music and Dance	42100000	Mexico	5300000	271300	16000	http://youtube.com/channel/UCQZfFRohQ7UX-0CdXI...
64	luisitocomunica	Unknown	41100000	Mexico	2500000	128900	1800	http://youtube.com/channel/UCECJDeK0MNapZbpaOz...
70	JessNoLimit	Movies, Animation	39600000	Indonesia	1300000	73500	1600	http://youtube.com/channel/UCvh1at6xpV1ytYOAzx...
96	TotalGaming093	Movies, Video Games	36300000	India	1500000	129400	4900	http://youtube.com/channel/UC5c9VIYTSvBSCaoMu_...
98	TechnoGamerzOfficial	Unknown	35600000	India	6200000	341800	16500	http://youtube.com/channel/UCX8pnu3DYUnx8qy8V_...
100	markiplier	Animation, Video Games	35500000	United States	2100000	126500	3800	http://youtube.com/channel/UC7_YxT-KID8kRbqZo7...
122	AboFlah	Animation, Video Games	32700000	Iraq	3300000	382000	11400	http://youtube.com/channel/UCqq5n-Oe-r1EEHI3yv...

	Username	Categories	Subscribers	Country	Visits	Likes	Comments	Links
123	MRINDIANHACKER	Unknown	32600000	India	6500000	617400	26000	http://youtube.com/channel/UCSiDGb0MnHFGjs4E2W...
131	fedevigevani	Animation, Humor	32000000	Mexico	7700000	412200	17000	http://youtube.com/channel/UCoQm-PeHC-cbJcIKJY...
132	dream	Animation, Video Games	31900000	United States	3300000	309200	19000	http://youtube.com/channel/UCTkXRDQl0IuXxVQrRQ...
136	MrBeast2	Daily Vlogs	31300000	United States	83100000	5000000	11600	http://youtube.com/channel/UC4-79UOlP48-QNGgCk...
145	jacksepticeye	Animation, Humor	30400000	United States	1600000	83400	2300	http://youtube.com/channel/UCYzPXprvl5Y-Sf0g4v...
153	DaFuqBoom	Animation, Humor	29800000	United States	52700000	1700000	82800	http://youtube.com/channel/UCsSsgPaZ2GSmO6iI8C...
176	CrazyXYZ	Unknown	27800000	India	4200000	284100	8600	http://youtube.com/channel/UCebC4x5I2-PQxg46Uc...
177	DanTDM	Animation, Video Games	27800000	United States	3500000	285000	52500	http://youtube.com/channel/UCS5Oz6CHmeoF7vSad0...
179	brentrivera	Video Games, Humor	27600000	United States	6400000	154100	5000	http://youtube.com/channel/UC56D-IHcUvLVFTX_8N...
180	NichLmao	Daily Vlogs	27500000	United States	1500000	85800	1600	http://youtube.com/channel/UC6VAIqNQBc7ggiqvsO...
195	nickiminaj	Music and Dance	26100000	United States	1600000	98300	7600	http://youtube.com/channel/UC3jOd7GUMhpgJRBhiL...
206	AlejoIgoa	Animation	25700000	Argentina	5700000	208400	1700	http://youtube.com/channel/UCZs0WwC0Dn_noiQE2B...
207	ZHCYT	Design/Art, DIY and Life Hacks	25700000	United States	2600000	127300	2200	http://youtube.com/channel/UClQubH2NeMmGLTLgNd...
234	rug	Video Games, Humor	24300000	United States	3200000	85300	5100	http://youtube.com/channel/UCilwZiBBfI9X6yiZRz...

	Username	Categories	Subscribers	Country	Visits	Likes	Comments	Links
238	alanbecker	Animation, Video Games	24300000	United States	7600000	582600	5900	http://youtube.com/channel/UCbKWv2x9t6u8yZoB3K...
241	juandediospantojaa	Music and Dance, Movies	24000000	Mexico	3000000	133200	3600	http://youtube.com/channel/UCzoUWqjCbcfWFdOMvo...
265	DrossRotzank	Unknown	23100000	Mexico	1700000	105900	3900	http://youtube.com/channel/UCNYW2vfGrUE6R5mIJY...
271	AmiRodrigueZZ	Animation, Humor	22900000	Colombia	4300000	294400	1300	http://youtube.com/channel/UCwohKbRK2QnojtLh3p...
277	StokesTwins	Video Games, Humor	22700000	United States	11700000	235000	10000	http://youtube.com/channel/UCbp9MyKCTEww4CxEzc...
280	SSundee	Animation, Video Games	22700000	United States	1700000	59800	1800	http://youtube.com/channel/UCke6I9N4KfC968-yRc...
281	souravjoshivlogs7028	Daily Vlogs	22700000	India	5600000	382300	8900	http://youtube.com/channel/UCjvgGbPPn-FgYeguc5...
287	VillageCookingChannel	Unknown	22500000	India	21500000	321500	5900	http://youtube.com/channel/UCk3JZr7eS3pg5AGEvB...

In [ ]: