```
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
import ydata_profiling as pp

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

#load the dataset
youtube= pd.read_csv("C:\\Users\\elmaf\\Desktop\\DS Data Sets\\
youtubers_df.csv")
youtube.head()
```

```
   Rank       Username           Categories   Suscribers
Country  \
0     1        tseries       Música y baile  249500000.0
India
1     2        MrBeast   Videojuegos, Humor  183500000.0  Estados
Unidos
2     3      CoComelon            Educación  165500000.0
Unknown
3     4       SETIndia                  NaN  162600000.0
India
4     5  KidsDianaShow  Animación, Juguetes  113500000.0
Unknown

        Visits      Likes  Comments  \
0      86200.0     2700.0      78.0
1  117400000.0  5300000.0   18500.0
2    7000000.0    24700.0       0.0
3      15600.0      166.0       9.0
4    3900000.0    12400.0       0.0

                                                Links
0  http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy...
1  http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
2  http://youtube.com/channel/UCbCmjCuTUZos6Inko4...
3  http://youtube.com/channel/UCpEhnqL0y41EpW2TvW...
4  http://youtube.com/channel/UCk8GzjMOrta8yxDcKf...
```

```
pp.ProfileReport(youtube)
```


```
{"model_id":"187a05d306a0422bb8227a5761d5fe2c","version_major":2,"vers
ion_minor":0}

{"model_id":"519c534820cd4e92870675d6a304947d","version_major":2,"vers
ion_minor":0}
```

{"model_id":"c8971abb9824492bbb97b4cfda0949e1","version_major":2,"version_minor":0}

```
<IPython.core.display.HTML object>
```

```
#basic structure of the dataset
youtube.info()
```

```
<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
```

```
#Statistical summary
youtube.describe()
```

```
              Rank    Suscribers        Visits         Likes  \
count  1000.000000  1.000000e+03  1.000000e+03  1.000000e+03   
mean    500.500000  2.189440e+07  1.209446e+06  5.363259e+04   
std     288.819436  1.682775e+07  5.229942e+06  2.580457e+05   
min       1.000000  1.170000e+07  0.000000e+00  0.000000e+00   
25%     250.750000  1.380000e+07  3.197500e+04  4.717500e+02   
50%     500.500000  1.675000e+07  1.744500e+05  3.500000e+03   
75%     750.250000  2.370000e+07  8.654750e+05  2.865000e+04   
max    1000.000000  2.495000e+08  1.174000e+08  5.300000e+06   

           Comments  
count   1000.000000  
mean    1288.768000  
std     6778.188308  
min        0.000000  
25%        2.000000  
50%       67.000000  
75%      472.000000  
max   154000.000000  
```

```
#Checking for missing data
youtube.isnull()
```

```
      Rank  Username  Categories  Suscribers  Country  Visits
Likes  \
0    False     False       False       False    False   False  False

1    False     False       False       False    False   False  False

2    False     False       False       False    False   False  False

4    False     False       False       False    False   False  False

5    False     False       False       False    False   False  False

..     ...       ...         ...         ...      ...     ...    ...

989  False     False       False       False    False   False  False

990  False     False       False       False    False   False  False

991  False     False       False       False    False   False  False

997  False     False       False       False    False   False  False

999  False     False       False       False    False   False  False


     Comments  Links
0       False  False
1       False  False
2       False  False
4       False  False
5       False  False
..        ...    ...
989     False  False
990     False  False
991     False  False
997     False  False
999     False  False

[694 rows x 9 columns]
```

```
#Null value count
youtube.isnull().sum().sort_values(ascending = False)
```

```
Categories    306
Rank            0
Username        0
Suscribers      0
Country         0
Visits          0
Likes           0
Comments        0
```

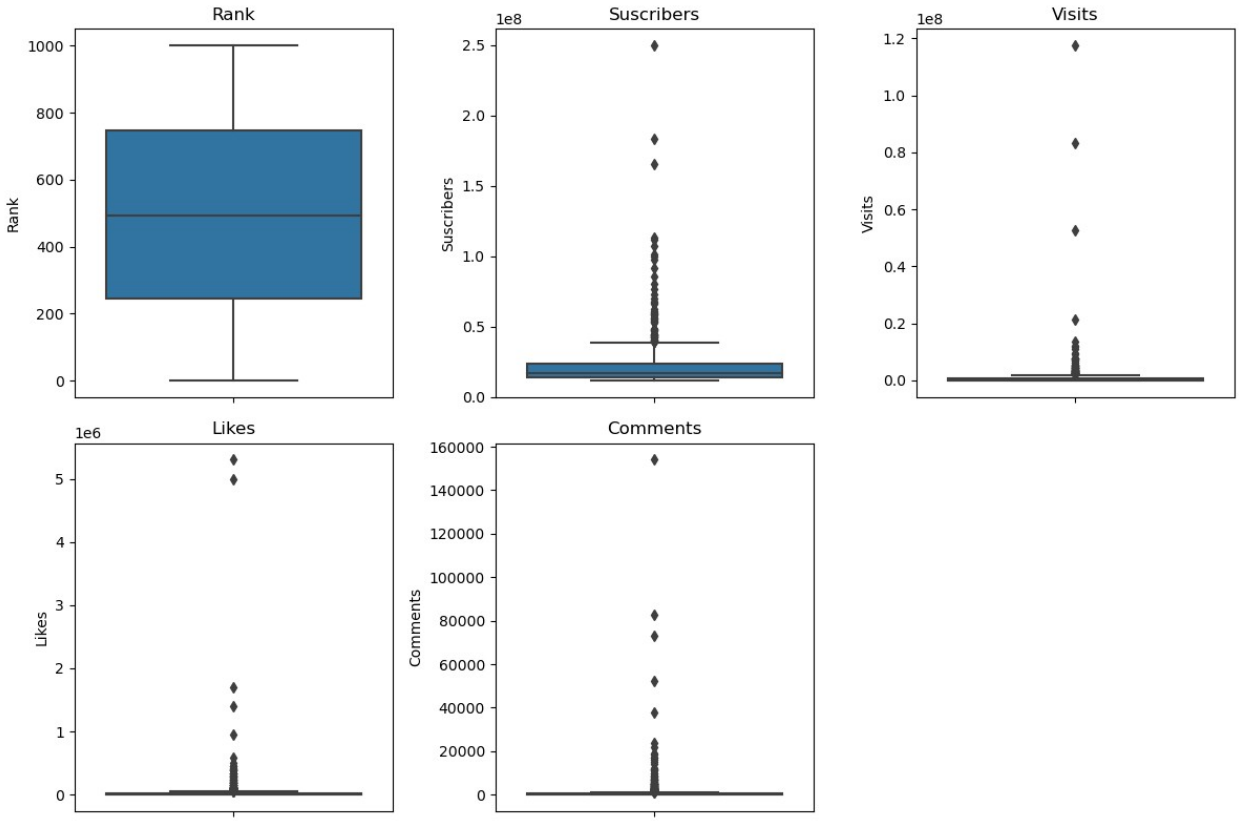
```
Links            0
dtype: int64
```

```
#Visualize missing data using heatmap
#import matplotlib
#matplotlib.use('TkAgg')  # Use a GUI backend (e.g., Tkinter)
#import matplotlib.pyplot as plt
# Visualize missing data using a heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(youtube.isnull(),cbar = False, cmap='coolwarm')
plt.title("Missing Data Heatmap")
plt.show()
```

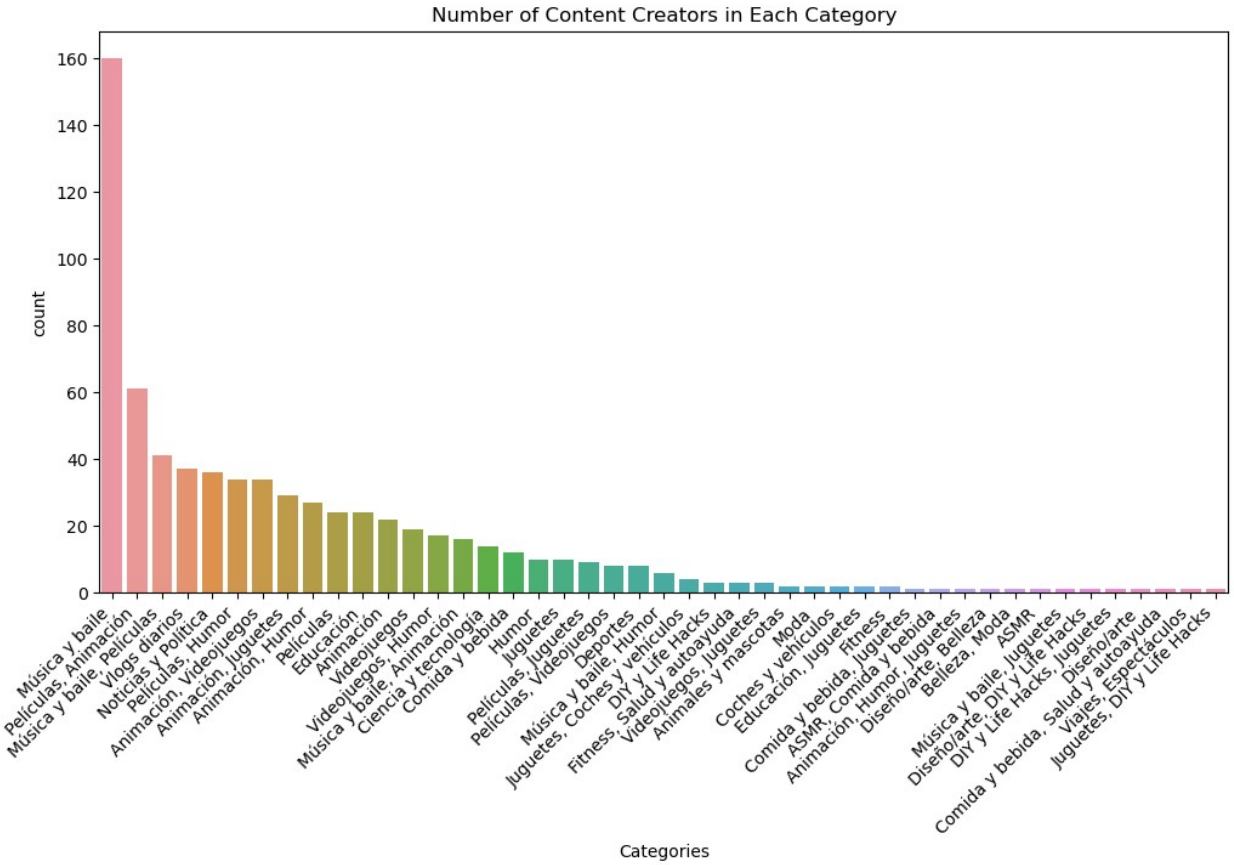
```
#Drop na values
youtube.dropna(inplace = True)
youtube.isnull().sum()
```

```
Rank          0
Username      0
Categories    0
Suscribers    0
Country       0
Visits        0
Likes         0
Comments      0
Links         0
dtype: int64
```

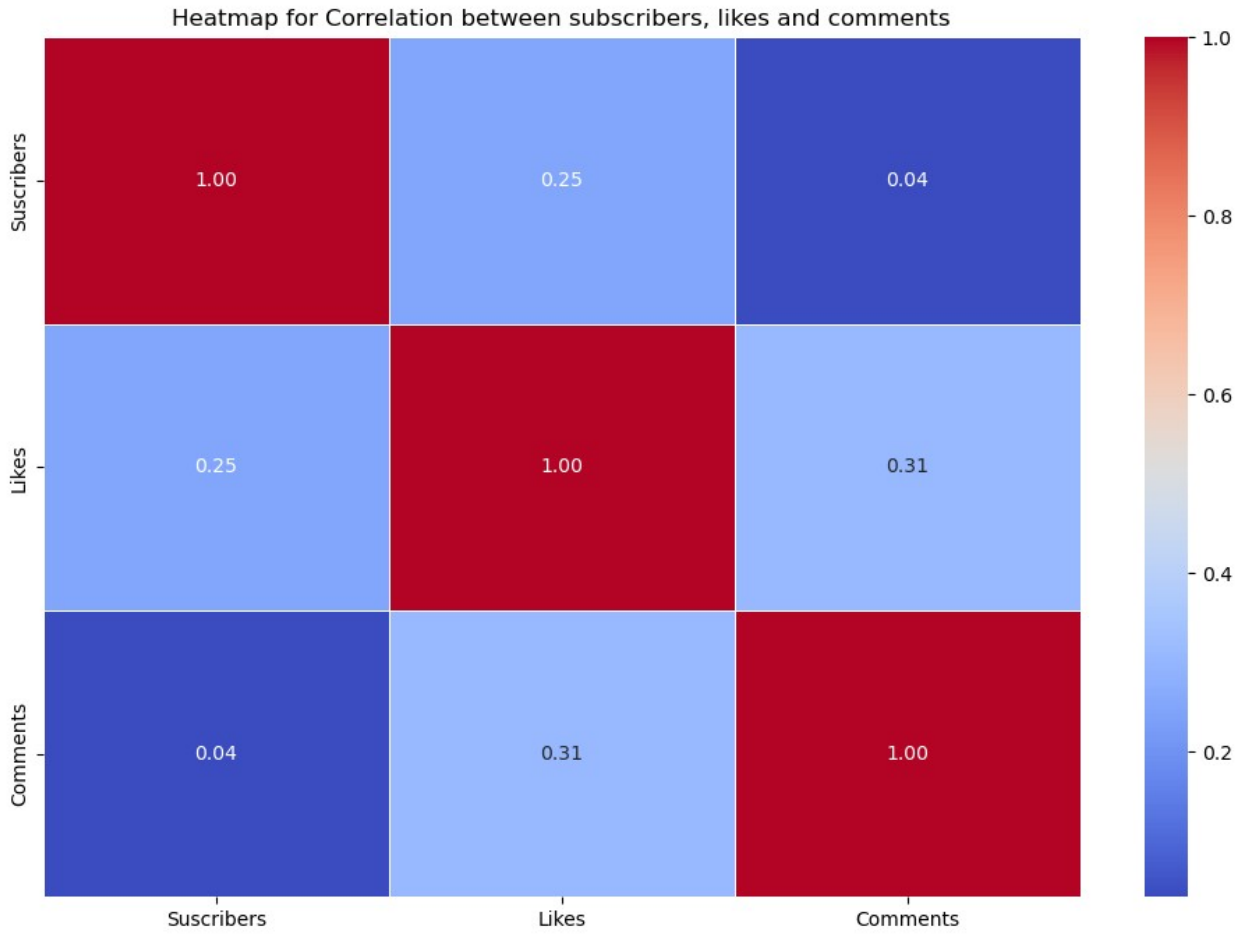
```
#Checking for outliers using boxplots
numeric_columns = ['Rank', 'Suscribers', 'Visits', 'Likes',
'Comments']
plt.figure(figsize=(12, 8))
for i, column in enumerate(numeric_columns):
    plt.subplot(2, 3, i+1)
    sns.boxplot(data=youtube, y=column)
    plt.title(column)
plt.tight_layout()
plt.show()
```



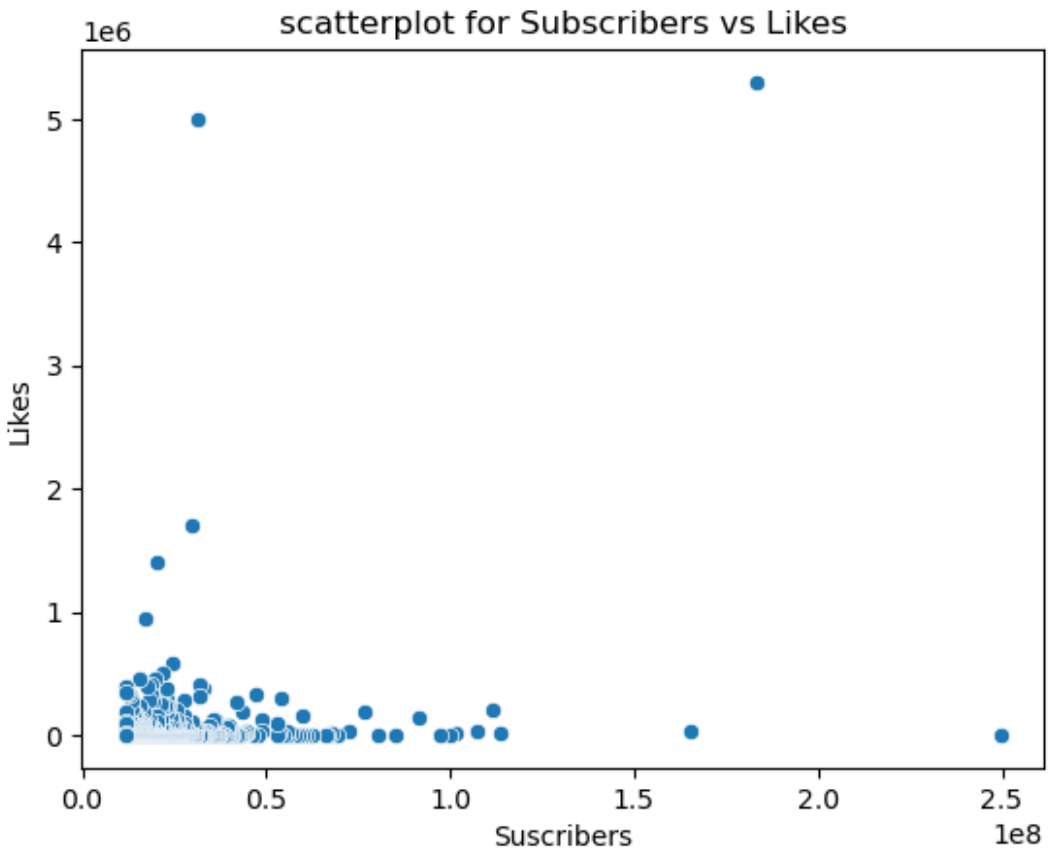

```
#Identify popular categories
plt.figure(figsize=(12, 6))
sns.countplot(x='Categories', data=youtube,
order=youtube['Categories'].value_counts().index)
plt.title("Number of Content Creators in Each Category")
plt.xticks(rotation=45, ha='right')
plt.show()
```



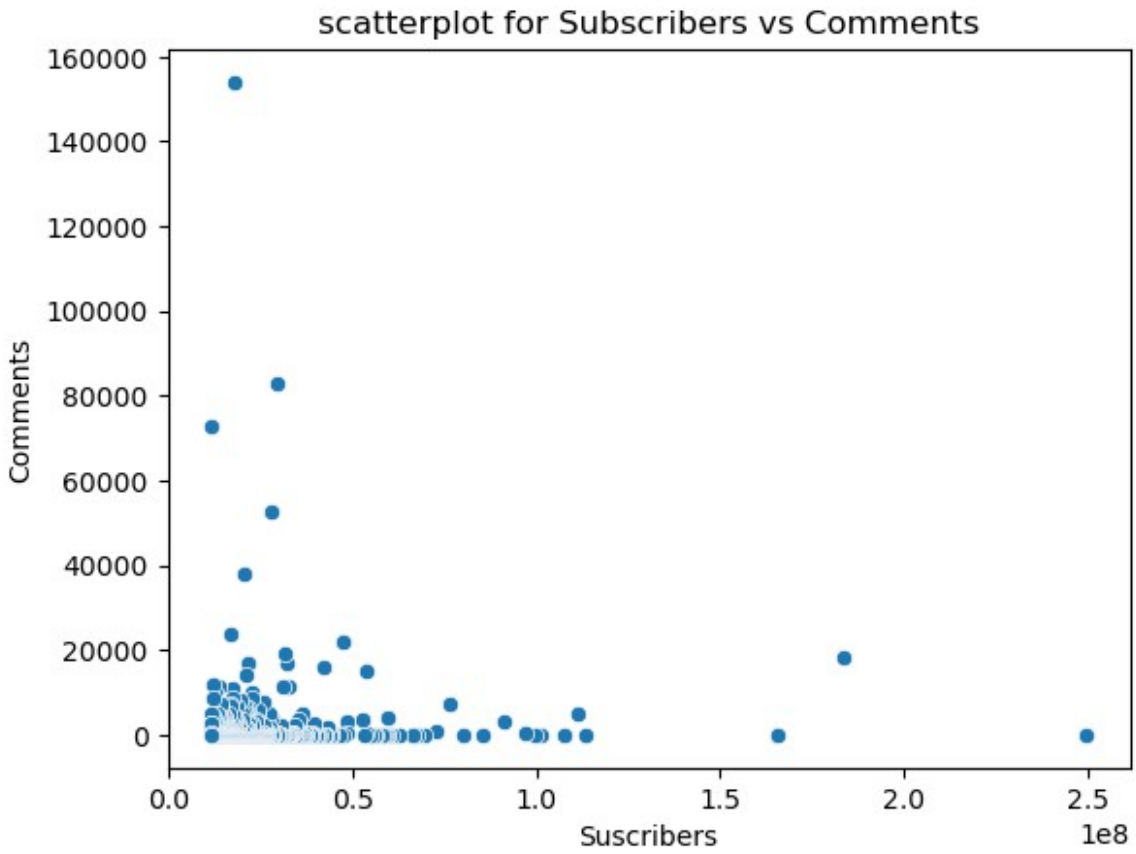

```
#Analyze the correlation between subscribers, likes and comments
correlation_yt = youtube[['Suscribers','Likes','Comments']].corr()
plt.figure(figsize=(12,8))
sns.heatmap(correlation_yt, annot=True, cmap='coolwarm', fmt='.2f',
linewidths=.5)
plt.title("Heatmap for Correlation between subscribers, likes and
comments")
plt.show()
```



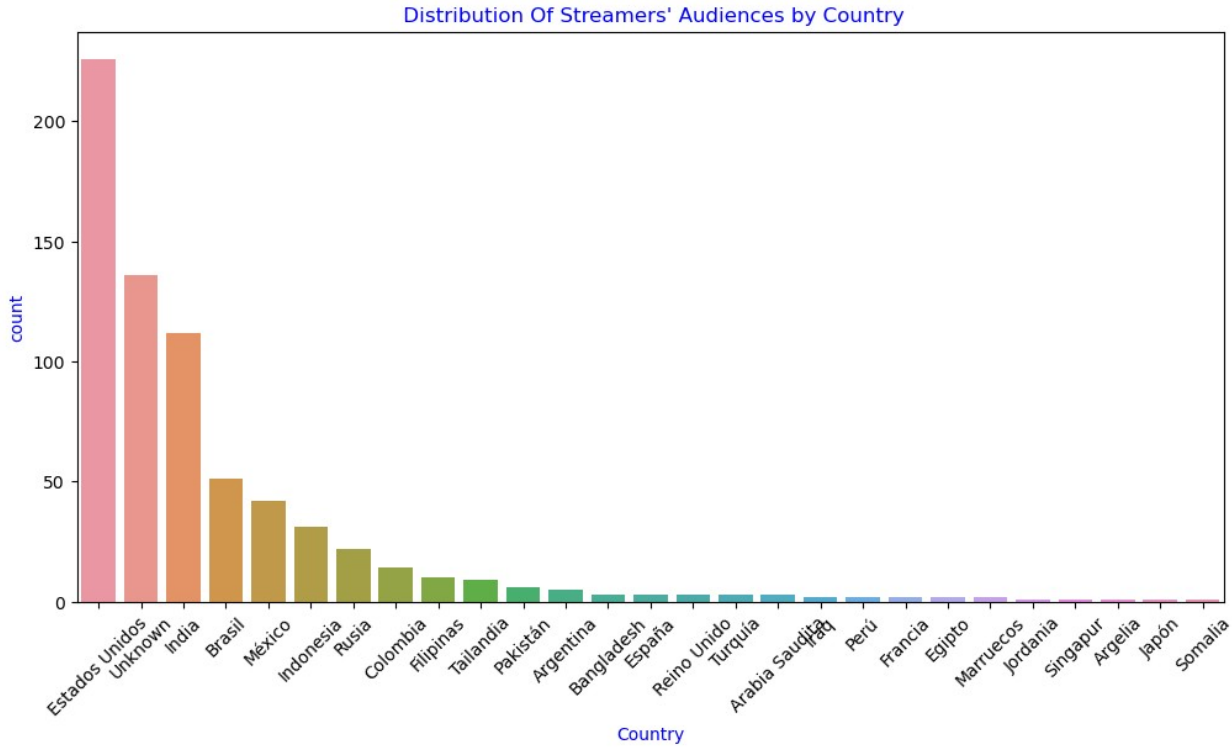

```
#scatterplot for Subscribers vs Likes
sns.scatterplot(x='Suscribers', y='Likes', data= youtube )
plt.title("scatterplot for Subscribers vs Likes")
plt.show()
```



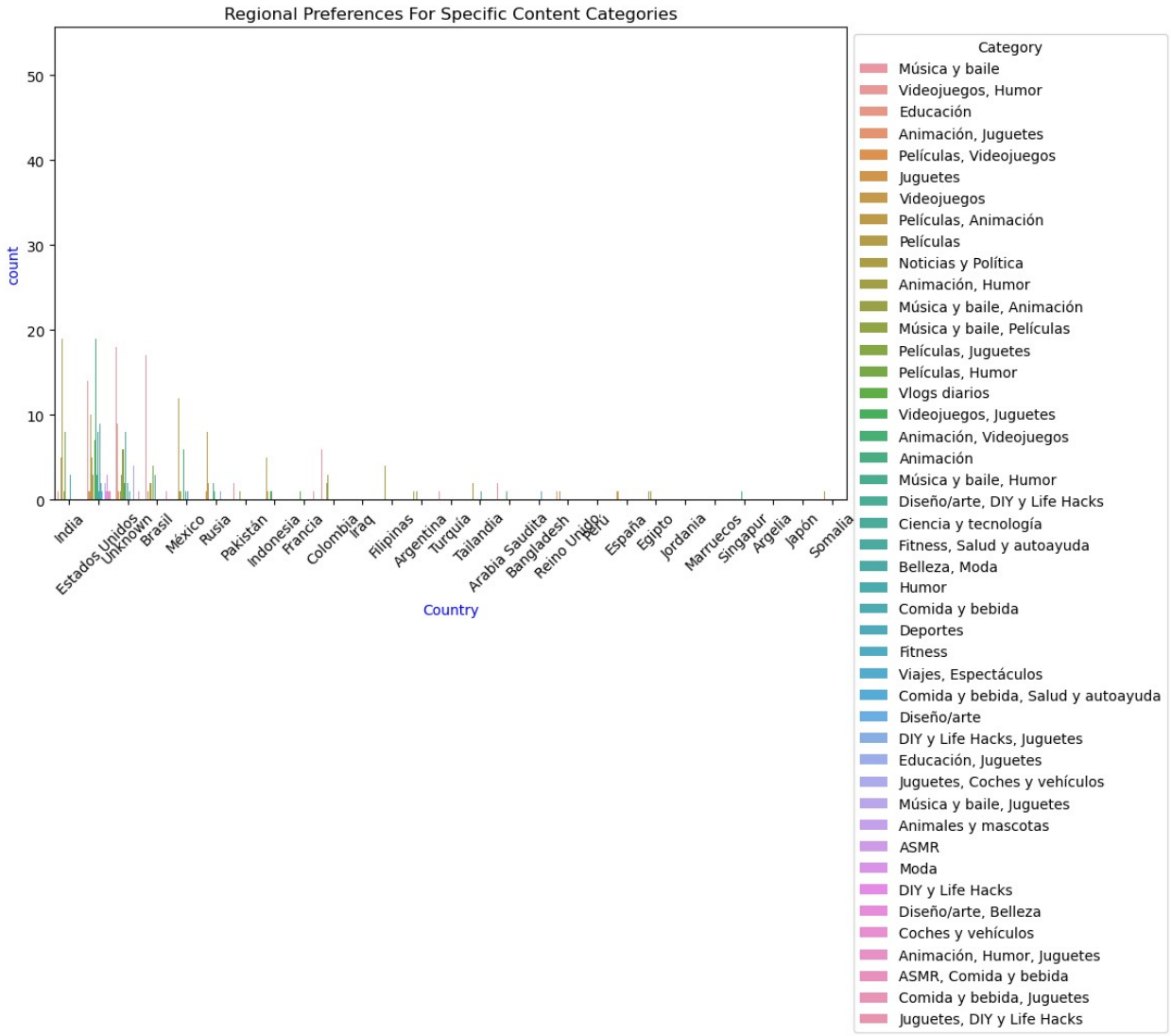

```
#scatterplot for Subscribers vs comments
sns.scatterplot(x='Suscribers', y='Comments', data= youtube )
plt.title("scatterplot for Subscribers vs Comments")
plt.show()
```




```
#Analyze the distribution of streamers' audiences by country
country_count = youtube['Country'].value_counts()
plt.figure(figsize=(12,6))
sns.barplot(x= country_count.index, y = country_count.values)
plt.title(" Distribution Of Streamers' Audiences by Country", color=
"Blue")
plt.xlabel('Country', color = "blue")
plt.ylabel('count', color = "blue")
plt.xticks(rotation=45)
plt.show()
```




```
#Are there regional preferences for specific content categories
plt.figure(figsize=(10, 6))
sns.countplot(data= youtube, x = 'Country', hue= 'Categories')
plt.title("Regional Preferences For Specific Content Categories")
plt.xlabel('Country', color = "blue")
plt.ylabel('count', color = "blue")
plt.xticks(rotation=45)
plt.legend(title='Category', bbox_to_anchor=(1, 1))
plt.show()
```



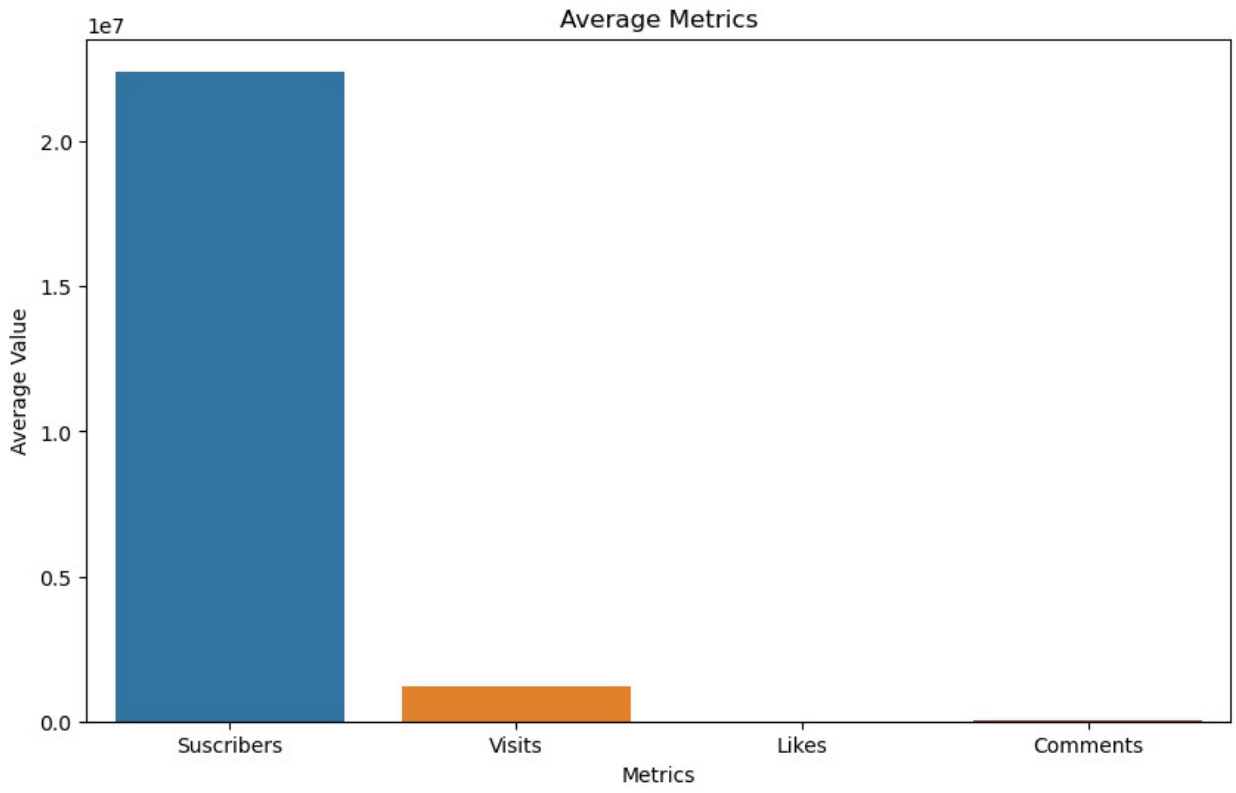

```
#Performance Metrics:
#Calculate the average number of subscribers, visits, likes, and
comments
average_subscribers= youtube['Suscribers'].mean()
average_visits= youtube['Visits'].mean()
average_likes= youtube['Likes'].mean()
average_comments= youtube['Comments'].mean()

print("Average subscribers: ", average_subscribers)
print("Average visits: ", average_visits)
print("Average comments: ", average_comments)
print("Average likes: ", average_likes)
```

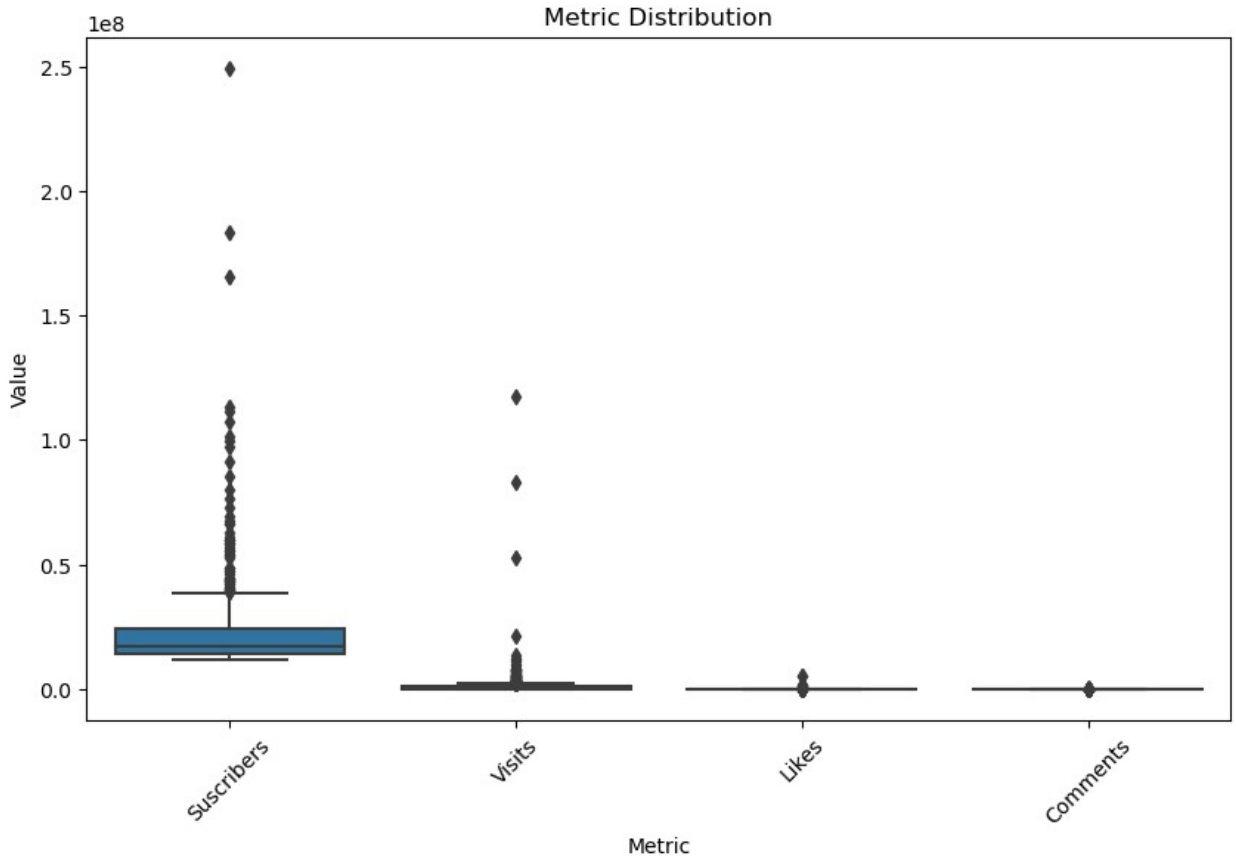
```
Average subscribers:  22415561.95965418
Average visits:  1210729.6829971182
Average comments:  1558.793948126801
Average likes:  53473.59798270893
```

```
# Visualise the average number of subscribers, visits, likes, and
comments
metrics = ['Suscribers', 'Visits','Likes','Comments']
average_metrics = [average_subscribers,average_visits,
average_comments, average_likes]

plt.figure(figsize=(10,6))
sns.barplot(x=metrics, y=average_metrics)
plt.xlabel("Metrics")
plt.ylabel("Average Value")
plt.title("Average Metrics")
plt.show()
```




```
#Are there patterns or anomalies in these metrics
plt.figure(figsize= (10,6))
sns.boxplot(data= youtube[metrics])
plt.xlabel("Metric")
plt.ylabel("Value")
plt.title("Metric Distribution")
plt.xticks(rotation= 45)
plt.show()
```




```
#the distribution of content categories
#  categories with the highest number of streamers
top_categories = youtube['Categories'].value_counts().head()
print("Top Categories with the Highest Number of Streamers:")
print(top_categories)
```

```
Top Categories with the Highest Number of Streamers:
Música y baile               160
Películas, Animación          61
Música y baile, Películas     41
Vlogs diarios                 37
Noticias y Política           36
Name: Categories, dtype: int64
```

```
# Calculate performance metrics by category
category_metrics = youtube.groupby('Categories')[['Suscribers',
'Visits', 'Likes', 'Comments']].mean().reset_index()

#Identify streamers with above-average performance in terms of
subscribers, visits, likes, and comments.
above_average_suscribers= youtube[youtube['Suscribers'] >
average_subscribers]
above_average_visits= youtube[youtube['Visits'] > average_visits]
above_average_likes= youtube[youtube['Likes'] > average_likes]
```

```
above_average_comments= youtube[youtube['Comments'] >
average_comments]
print("Streamers with Above-Average Performance")
print("Above Average Subscribers ", above_average_suscribers)
print("Above Average Visits ", above_average_visits)
print("Above Average Likes ", above_average_likes)
print("Above Average Comments ", above_average_comments)
```

```
Streamers with Above-Average Performance
Above Average Subscribers       Rank              Username
Categories   Suscribers  \
0       1               tseries             Música y baile
249500000.0
1       2               MrBeast         Videojuegos, Humor
183500000.0
2       3             CoComelon                   Educación
165500000.0
4       5         KidsDianaShow        Animación, Juguetes
113500000.0
5       6             PewDiePie     Películas, Videojuegos
111500000.0
..    ...                   ...                         ...           ..
.
282   283  souravjoshivlogs7028              Vlogs diarios
22700000.0
283   284       THECHAINSMOKERS  Música y baile, Películas
22700000.0
284   285              ZachKing           Películas, Humor
22600000.0
285   286            BenAzelart         Videojuegos, Humor
22500000.0
287   288                fgteev     Películas, Videojuegos
22500000.0

            Country       Visits      Likes  Comments  \
0             India      86200.0     2700.0      78.0
1    Estados Unidos  117400000.0  5300000.0   18500.0
2           Unknown    7000000.0    24700.0       0.0
4           Unknown    3900000.0    12400.0       0.0
5    Estados Unidos    2400000.0   197300.0    4900.0
..              ...          ...        ...       ...
282           India    5600000.0   382300.0    8900.0
283  Estados Unidos      38900.0     1800.0      76.0
284  Estados Unidos    4900000.0   238500.0     522.0
285  Estados Unidos    3700000.0    44900.0    2700.0
287  Estados Unidos     826800.0    10300.0    2100.0

                                                 Links
0    http://youtube.com/channel/UCq-Fj5jknLsUf-MWSy...
1    http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
```

```
2    http://youtube.com/channel/UCbCmjCuTUZos6Inko4...
4    http://youtube.com/channel/UCk8GzjMOrta8yxDcKf...
5    http://youtube.com/channel/UC-lHJZR3Gqxm24_Vd_...
..                                                 ...
282  http://youtube.com/channel/UCjvgGbPPn-FgYeguc5...
283  http://youtube.com/channel/UCq3Ci-h945sbEYXpVl...
284  http://youtube.com/channel/UCq8DICunczvLuJJq41...
285  http://youtube.com/channel/UCwVg9btOceLQuNCdoQ...
287  http://youtube.com/channel/UCC-RHF_77zQdKcA75h...

[203 rows x 9 columns]
Above Average Visits       Rank            Username
Categories   Suscribers  \
1       2             MrBeast         Videojuegos, Humor  183500000.0

2       3           CoComelon                  Educación  165500000.0

4       5       KidsDianaShow        Animación, Juguetes  113500000.0

5       6           PewDiePie     Películas, Videojuegos  111500000.0

6       7  LikeNastyaofficial                   Juguetes  107500000.0

..    ...                 ...                        ...          ...

976   977      NickDiGiovanni            Comida y bebida   11900000.0

978   979           HikakinTV                       Humor   11900000.0

983   984          mussoumano  Música y baile, Animación   11900000.0

985   986      lukedavidson81           Animación, Humor   11800000.0

990   991           JoeHattab                  Películas   11700000.0


             Country       Visits      Likes  Comments  \
1    Estados Unidos  117400000.0  5300000.0   18500.0
2           Unknown    7000000.0    24700.0       0.0
4           Unknown    3900000.0    12400.0       0.0
5    Estados Unidos    2400000.0   197300.0    4900.0
6           Unknown    2600000.0    28000.0       0.0
..              ...          ...        ...       ...
976  Estados Unidos    4300000.0   339800.0    2700.0
978           Japón    2100000.0    51000.0    2300.0
983          Brasil    1600000.0    68700.0    2400.0
985  Estados Unidos    1700000.0   109200.0    1000.0
990         Somalia    1900000.0    98500.0    2900.0

                                                  Links
1    http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
```

```
2    http://youtube.com/channel/UCbCmjCuTUZos6Inko4...
4    http://youtube.com/channel/UCk8GzjMOrta8yxDcKf...
5    http://youtube.com/channel/UC-lHJZR3Gqxm24_Vd_...
6    http://youtube.com/channel/UCJplp5SjeGSdVdwsfb...
..                                                 ...
976  http://youtube.com/channel/UCMyOj6fhvKFMjxUCp3...
978  http://youtube.com/channel/UCZf__ehlCEBPop-_sl...
983  http://youtube.com/channel/UC607Gii1uo93qDqiGR...
985  http://youtube.com/channel/UCuwJoiGWRxPYStBp0l...
990  http://youtube.com/channel/UCe6eisvsctSPvBhmin...

[134 rows x 9 columns]
Above Average Likes       Rank        Username
Categories   Suscribers  \
1       2         MrBeast         Videojuegos, Humor  183500000.0
5       6       PewDiePie     Películas, Videojuegos  111500000.0
10     11       BLACKPINK             Música y baile   91300000.0
14     15             BTS             Música y baile   76500000.0
26     27     dudeperfect                Videojuegos   59700000.0
..    ...             ...                        ...          ...
965   966         mmdcrew  Música y baile, Películas   11900000.0
976   977  NickDiGiovanni            Comida y bebida   11900000.0
983   984      mussoumano  Música y baile, Animación   11900000.0
985   986  lukedavidson81           Animación, Humor   11800000.0
990   991       JoeHattab                  Películas   11700000.0

            Country       Visits      Likes  Comments  \
1    Estados Unidos  117400000.0  5300000.0   18500.0
5    Estados Unidos    2400000.0   197300.0    4900.0
10   Estados Unidos     863200.0   146900.0    3400.0
14            India     969700.0   180300.0    7400.0
26   Estados Unidos    5300000.0   156500.0    4200.0
..              ...          ...        ...       ...
965           Rusia    3100000.0   189900.0    4800.0
976  Estados Unidos    4300000.0   339800.0    2700.0
983          Brasil    1600000.0    68700.0    2400.0
985  Estados Unidos    1700000.0   109200.0    1000.0
990         Somalia    1900000.0    98500.0    2900.0

                                                  Links
1    http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
5    http://youtube.com/channel/UC-lHJZR3Gqxm24_Vd_...
10   http://youtube.com/channel/UCOmHUn--16B90oW2L6...
14   http://youtube.com/channel/UCLkAepWjdylmXSltof...
26   http://youtube.com/channel/UCRijo3ddMTht_IHyNS...
..                                                 ...
965  http://youtube.com/channel/UCWnqnojAgMdN0fQpr_...
976  http://youtube.com/channel/UCMyOj6fhvKFMjxUCp3...
983  http://youtube.com/channel/UC607Gii1uo93qDqiGR...
985  http://youtube.com/channel/UCuwJoiGWRxPYStBp0l...
```

```
990  http://youtube.com/channel/UCe6eisvsctSPvBhmin...

[119 rows x 9 columns]
Above Average Comments       Rank        Username
Categories   Suscribers  \
1       2         MrBeast         Videojuegos, Humor  183500000.0
5       6       PewDiePie     Películas, Videojuegos  111500000.0
10     11       BLACKPINK             Música y baile   91300000.0
14     15             BTS             Música y baile   76500000.0
26     27     dudeperfect                Videojuegos   59700000.0
..    ...             ...                        ...          ...
965   966         mmdcrew  Música y baile, Películas   11900000.0
976   977  NickDiGiovanni            Comida y bebida   11900000.0
978   979       HikakinTV                      Humor   11900000.0
983   984      mussoumano  Música y baile, Animación   11900000.0
990   991       JoeHattab                  Películas   11700000.0

             Country       Visits      Likes  Comments  \
1    Estados Unidos  117400000.0  5300000.0   18500.0
5    Estados Unidos    2400000.0   197300.0    4900.0
10   Estados Unidos     863200.0   146900.0    3400.0
14            India     969700.0   180300.0    7400.0
26   Estados Unidos    5300000.0   156500.0    4200.0
..              ...          ...        ...       ...
965           Rusia    3100000.0   189900.0    4800.0
976  Estados Unidos    4300000.0   339800.0    2700.0
978           Japón    2100000.0    51000.0    2300.0
983          Brasil    1600000.0    68700.0    2400.0
990         Somalia    1900000.0    98500.0    2900.0

                                                 Links
1    http://youtube.com/channel/UCX6OQ3DkcsbYNE6H8u...
5    http://youtube.com/channel/UC-lHJZR3Gqxm24_Vd_...
10   http://youtube.com/channel/UCOmHUn--16B90oW2L6...
14   http://youtube.com/channel/UCLkAepWjdylmXSltof...
26   http://youtube.com/channel/UCRijo3ddMTht_IHyNS...
..                                                 ...
965  http://youtube.com/channel/UCWnqnojAgMdN0fQpr_...
976  http://youtube.com/channel/UCMyOj6fhvKFMjxUCp3...
978  http://youtube.com/channel/UCZf__ehlCEBPop-_sl...
983  http://youtube.com/channel/UC607Gii1uo93qDqiGR...
990  http://youtube.com/channel/UCe6eisvsctSPvBhmin...

[112 rows x 9 columns]
```

```
# Who are the top-performing content creators
top_subscribers= above_average_suscribers.sort_values('Suscribers',
ascending = False).head()
top_visits= above_average_visits.sort_values('Visits', ascending =
False). head()
```

```
top_likes= above_average_likes.sort_values('Likes', ascending =
False). head()
top_comments= above_average_comments.sort_values('Comments', ascending
= False). head()

print("Top-Performing Content Creators by Subscribers")
print(top_subscribers[['Username', 'Suscribers']])

print("Top-Performing Content Creators by Visits")
print(top_visits[['Username', 'Visits']])

print("Top-Performing Content Creators by Likes")
print(top_likes[['Username', 'Likes']])

print("Top-Performing Content Creators by Comments")
print(top_comments[['Username', 'Comments']])
```

```
Top-Performing Content Creators by Subscribers
        Username   Suscribers
0        tseries  249500000.0
1        MrBeast  183500000.0
2      CoComelon  165500000.0
4  KidsDianaShow  113500000.0
5      PewDiePie  111500000.0
Top-Performing Content Creators by Visits
              Username       Visits
1              MrBeast  117400000.0
136           MrBeast2   83100000.0
153          DaFuqBoom   52700000.0
488  BeastPhilanthropy   21500000.0
958            dojacat   13600000.0
Top-Performing Content Creators by Likes
              Username      Likes
1              MrBeast  5300000.0
136           MrBeast2  5000000.0
153          DaFuqBoom  1700000.0
341    triggeredinsaan  1400000.0
488  BeastPhilanthropy   952100.0
Top-Performing Content Creators by Comments
               Username  Comments
436  BispoBrunoLeonardo  154000.0
153           DaFuqBoom   82800.0
958             dojacat   73000.0
177              DanTDM   52500.0
341     triggeredinsaan   38000.0
```

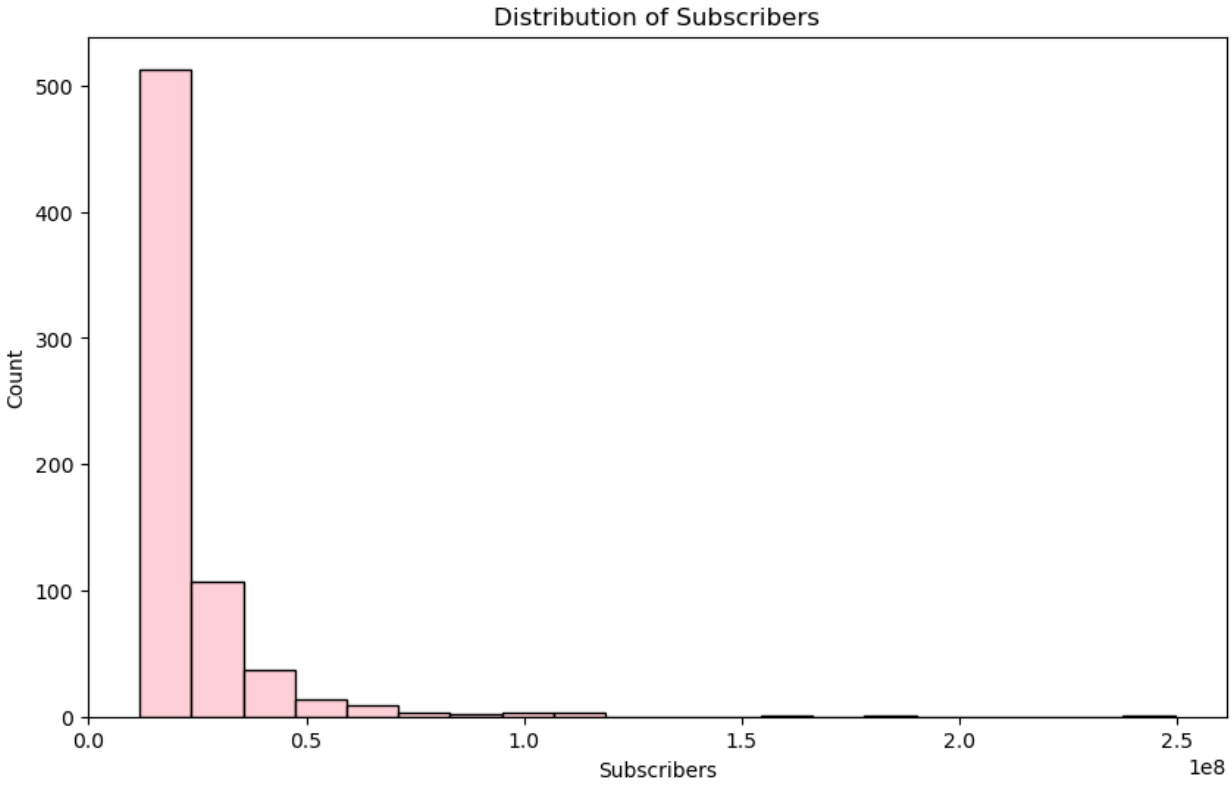
```
# streamers with above-average performance
above_average_streamers = youtube[
    (youtube['Suscribers'] > average_subscribers) &
    (youtube['Visits'] > average_visits) &
```

```
    (youtube['Likes'] > average_likes) &
    (youtube['Comments'] > average_comments)
]

# Display the top-performing content creators
top_performers = above_average_streamers.sort_values(by='Suscribers',
ascending=False).head(10)
print("Top-Performing Content Creators:")
print(top_performers[['Rank', 'Suscribers', 'Visits', 'Likes',
'Comments']])
```

```
Top-Performing Content Creators:
    Rank   Suscribers       Visits      Likes  Comments
1      2  183500000.0  117400000.0  5300000.0   18500.0
5      6  111500000.0    2400000.0   197300.0    4900.0
26    27   59700000.0    5300000.0   156500.0    4200.0
34    35   54100000.0    4300000.0   300400.0   15000.0
39    40   48600000.0    2000000.0   117100.0    3000.0
43    44   47300000.0    9700000.0   330400.0   22000.0
58    59   43400000.0    2200000.0   183400.0    1800.0
62    63   42100000.0    5300000.0   271300.0   16000.0
70    71   39600000.0    1300000.0    73500.0    1600.0
96    97   36300000.0    1500000.0   129400.0    4900.0
```

```
# Analyze the distribution of subscribers
plt.figure(figsize=(10, 6))
sns.histplot(data=youtube, x='Suscribers', bins=20, color= 'pink')
plt.title('Distribution of Subscribers')
plt.xlabel('Subscribers')
plt.ylabel('Count')
plt.show()
```


Distribution of Subscribers
Count
Subscribers
1e8
0
100
200
300
400
500
0.0
0.5
1.0
1.5
2.0
2.5