



Phân tích Top 1000 Youtube Channels Dataset

Lời giới thiệu về YouTube và lí do chọn bộ dữ liệu này:

YouTube là một trong những nền tảng chia sẻ video lớn nhất trên Internet, thu hút hàng tỷ người dùng hàng ngày. Nó không chỉ là một nơi để xem video giải trí, mà còn là một kênh quảng cáo mạnh mẽ và một nền tảng sáng tạo cho các nhà sản xuất nội dung.

Bộ dữ liệu Top 1000 YouTube Channels là một tập hợp dữ liệu quan trọng cho phân tích và khám phá các kênh YouTube phổ biến. Bộ dữ liệu này cung cấp thông tin về các kênh hàng đầu trên YouTube, bao gồm thông tin về lượt đăng ký, lượt xem, số lượng video và nhiều thuộc tính khác. Điều này cho phép chúng ta nắm bắt được xu hướng, thị phần và mối quan hệ giữa các yếu tố khác nhau trong các kênh YouTube.

Lý do chọn bộ dữ liệu này là để:

Nghiên cứu các yếu tố ảnh hưởng đến sự phát triển và thành công của các kênh YouTube. Hiểu rõ hơn về sự tương tác giữa số lượng lượt đăng ký, lượt xem và số lượng video của mỗi kênh. Phân tích xu hướng phát triển của các lĩnh vực chủ đạo trên YouTube. Tìm hiểu các kênh YouTube nổi tiếng và đánh giá các yếu tố thành công của họ. Xây dựng mô hình dự đoán hoặc gợi ý cho sự phát triển kênh YouTube. Bằng cách phân tích bộ dữ liệu Top 1000 YouTube Channels, chúng ta có thể tìm ra những chiến lược thành công, xu hướng phát triển và những yếu tố quan trọng trong việc xây dựng và quản lý kênh YouTube.

Nguồn: https://us.youtubers.me/global/all/top-1000-most_subscribed-youtube-channels
https://us.youtubers.me/global/all/top-1000-most_subscribed-youtube-channels

Thu thập dữ liệu

```
In [1]: 1 # Nhập các thư viện cần thiết
2 import numpy as np
3 import pandas as pd
4 import seaborn as sns
5 import matplotlib.pyplot as plt
6
7 # visualization
8 import matplotlib
9 import matplotlib.pyplot as plt
10 import seaborn as sns
11 import plotly.express as px
12 import plotly.graph_objects as go
13 import plotly.subplots as sp
14
```

```
In [2]: 1 # Sau khi dùng Excel Lấy dữ liệu từ trang web về tôi đã tạo file dữ liệu là "top1000Subscribed.csv"
2 df = pd.read_csv("top1000Subscribed.csv", sep = ";")
3 df.head(10)
```

Out[2]:

	Rank	Youtube Channel	Subscribers	Video Views	Video Count	Category	Started
0	1	T-Series	243,000,000	224,881,706,554	19,757	Music	2006
1	2	Cocomelon - Nursery Rhymes	160,000,000	161,435,301,255	927	Education	2006
2	3	SET India	157,000,000	146,246,275,148	113,723	Shows	2006
3	4	Sony SAB	81,600,000	98,603,821,529	69,353	Shows	2007
4	5	✿ Kids Diana Show	111,000,000	92,321,006,440	1,097	People & Blogs	2015
5	6	Like Nastya	106,000,000	89,956,905,855	800	People & Blogs	2016
6	7	WWE	95,200,000	76,590,809,077	69,232	Sports	2007
7	8	Vlad and Niki	97,700,000	76,255,951,806	561	Entertainment	2018
8	9	Zee TV	69,400,000	71,539,733,613	126,034	Entertainment	2005
9	10	Colors TV	63,200,000	60,190,628,294	110,59	Shows	2008

Bộ dữ liệu "Top 1000 Kênh Youtube có số lượng đăng ký cao nhất" bao gồm 6 biến.

1. Youtube Channel: Tên của kênh Youtube.
2. Subscribers: Tổng số lượng người đăng ký.
3. Video Views: Tổng số lượt xem video.
4. Video Count: Tổng số video đã được đăng.
5. Category: Danh mục của kênh.
6. Started: Thời điểm bắt đầu hoạt động của kênh.

Tiền xử lý dữ liệu

```
In [3]: 1 df.shape
```

Out[3]: (1000, 7)

In [4]:

```
1 df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Rank              1000 non-null    int64  
 1   Youtube Channel  1000 non-null    object  
 2   Subscribers      1000 non-null    object  
 3   Video Views      1000 non-null    object  
 4   Video Count       1000 non-null    object  
 5   Category          964 non-null    object  
 6   Started           1000 non-null    int64  
dtypes: int64(2), object(5)
memory usage: 54.8+ KB
```

Bộ dữ liệu có 1000 dòng và 7 cột. Cột Category có 36 missing values. Dưới đây là một số nhận xét về các cột trong bộ dữ liệu:

- Cột "Rank" (Xếp hạng): Đây là cột chứa xếp hạng của các kênh YouTube trong bộ dữ liệu. Dữ liệu trong cột này có kiểu dữ liệu là số nguyên.
- Cột "Youtube Channel" (Kênh YouTube): Đây là cột chứa tên của các kênh YouTube. Dữ liệu trong cột này có kiểu dữ liệu là chuỗi (object).
- Cột "Subscribers" (Người đăng ký): Đây là cột chứa số lượng người đăng ký của các kênh YouTube. Dữ liệu trong cột này có kiểu dữ liệu là chuỗi (object).
- Cột "Video Views" (Lượt xem video): Đây là cột chứa số lượt xem video của các kênh YouTube. Dữ liệu trong cột này có kiểu dữ liệu là chuỗi (object).
- Cột "Video Count" (Số lượng video): Đây là cột chứa số lượng video đã được tải lên của các kênh YouTube. Dữ liệu trong cột này có kiểu dữ liệu là chuỗi (object).
- Cột "Category" (Thể loại): Đây là cột chứa thông tin về thể loại của các kênh YouTube. Dữ liệu trong cột này có kiểu dữ liệu là chuỗi (object).
- Cột "Started" (Ngày bắt đầu): Đây là cột chứa thông tin về ngày bắt đầu hoạt động của các kênh YouTube. Dữ liệu trong cột này có kiểu dữ liệu là số nguyên.

Tuy nhiên, cần lưu ý rằng các cột "Subscribers", "Video Views", "Video Count" nên có kiểu dữ liệu là số thay vì chuỗi để phân tích dữ liệu hiệu quả hơn. Cho nên ta sẽ xử lý chúng.

In [5]:

```
1 col = ['Subscribers', 'Video Views', 'Video Count']
2 for i in col:
3     df[i] = df[i].str.replace(","," ")
4
5 col = ['Subscribers', 'Video Views', 'Video Count']
6 for i in col:
7     df[i] = df[i].astype('int64')
8 df.head()
```

Out[5]:

	Rank	Youtube Channel	Subscribers	Video Views	Video Count	Category	Started
0	1	T-Series	243000000	224881706554	19757	Music	2006
1	2	Cocomelon - Nursery Rhymes	160000000	161435301255	927	Education	2006
2	3	SET India	157000000	146246275148	113723	Shows	2006
3	4	Sony SAB	81600000	98603821529	69353	Shows	2007
4	5	✿ Kids Diana Show	111000000	92321006440	1097	People & Blogs	2015

```
In [6]: 1 df.dtypes
```

```
Out[6]: Rank           int64
Youtube Channel    object
Subscribers        int64
Video Views         int64
Video Count          int64
Category            object
Started             int64
dtype: object
```

```
In [7]: 1 df.Category.unique()
```

```
Out[7]: array(['Music', 'Education', 'Shows', 'People & Blogs', 'Sports',
   'Entertainment', 'Film & Animation', 'Comedy', nan, 'Gaming',
   'Howto & Style', 'News & Politics', 'Pets & Animals', 'Trailers',
   'Science & Technology', 'Movies', 'Autos & Vehicles',
   'Nonprofits & Activism', 'Travel & Events'], dtype=object)
```

Kết quả của lệnh df.Category.unique() cho thấy các giá trị duy nhất trong cột "Category" là:

- 'Music': Danh mục âm nhạc.
- 'Film & Animation': Danh mục phim và hoạt hình.
- 'Education': Danh mục giáo dục.
- 'Shows': Danh mục chương trình truyền hình.
- 'Entertainment': Danh mục giải trí.
- 'Gaming': Danh mục trò chơi.
- 'People & Blogs': Danh mục về con người và blog.
- 'Sports': Danh mục thể thao.
- 'Howto & Style': Danh mục hướng dẫn và phong cách.
- 'News & Politics': Danh mục tin tức và chính trị.
- 'Comedy': Danh mục hài kịch.
- 'Trailers': Danh mục trailer.
- 'Nonprofits & Activism': Danh mục phi lợi nhuận và hoạt động chính trị.
- 'Science & Technology': Danh mục khoa học và công nghệ.
- 'Movies': Danh mục phim.
- 'Pets & Animals': Danh mục về thú cưng và động vật.
- 'Autos & Vehicles': Danh mục ô tô và phương tiện.
- 'Travel & Events': Danh mục du lịch và sự kiện.

Tuy nhiên, cần lưu ý rằng trong danh sách các giá trị duy nhất này, có một giá trị là "nan" không phải là một danh mục hợp lệ. Có thể có sự lỗi trong quá trình xử lý dữ liệu hoặc các giá trị bị nhiễu. Ta không thể thay thế bằng các giá trị như mode vì như thế tính đúng đắn của bộ dữ liệu của ta sẽ giảm. Ta chỉ có thể loại bỏ nó hoặc đặt một giá trị là "Other". Vì dữ liệu là rất quan trọng ta hạn chế loại bỏ dữ liệu một cách tối đa nên ta sẽ đặt nó là giá trị other.

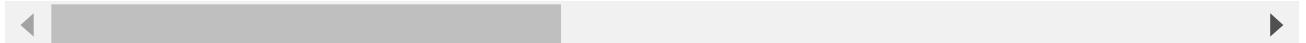
```
In [8]: 1 df['Category'] = df['Category'].fillna('Other')
```

```
In [9]: 1 # cột "Rank" Lúc này không còn sử dụng nữa nên ta sẽ bỏ cột này
2 df = df.drop(["Rank"], axis = 1 )
```

In [10]: 1 df.groupby("Category").describe().T

Out[10]:

	Category	Autos & Vehicles	Comedy	Education	Entertainment	Film & Animation	Gaming	Howto & Style
Subscribers	count	1.000000e+00	5.100000e+01	3.600000e+01	2.570000e+02	5.400000e+01	6.300000e+01	1.200000e+01
	mean	2.130000e+07	1.661137e+07	2.720667e+07	1.729740e+07	1.963611e+07	2.068190e+07	2.606167e+07
	std	NaN	9.059651e+06	2.779335e+07	1.457883e+07	1.517202e+07	1.562250e+07	1.962924e+07
	min	2.130000e+07	2.380000e+06	4.070000e+06	2.180000e+06	3.450000e+06	2.760000e+06	7.940000e+06
	25%	2.130000e+07	9.170000e+06	1.215000e+07	9.120000e+06	9.677500e+06	1.275000e+07	1.422500e+07
	50%	2.130000e+07	1.420000e+07	1.835000e+07	1.420000e+07	1.565000e+07	1.710000e+07	2.095000e+07
	75%	2.130000e+07	2.320000e+07	3.275000e+07	2.030000e+07	2.542500e+07	2.360000e+07	2.707500e+07
	max	2.130000e+07	4.040000e+07	1.600000e+08	1.580000e+08	8.590000e+07	1.110000e+08	7.990000e+07
Video Views	count	1.000000e+00	5.100000e+01	3.600000e+01	2.570000e+02	5.400000e+01	6.300000e+01	1.200000e+01
	mean	9.357260e+09	1.088914e+10	1.936549e+10	1.185446e+10	1.203686e+10	1.052161e+10	9.555500e+09
	std	NaN	5.263578e+09	2.683225e+10	8.765790e+09	9.451738e+09	5.213323e+09	5.555031e+09
	min	9.357260e+09	5.945097e+09	5.923633e+09	5.927849e+09	5.943941e+09	6.005230e+09	6.035069e+09
	25%	9.357260e+09	7.364105e+09	7.616730e+09	7.067249e+09	7.271233e+09	6.858226e+09	6.907470e+09
	50%	9.357260e+09	9.149291e+09	1.059708e+10	8.877930e+09	8.464669e+09	8.260931e+09	7.188726e+09
	75%	9.357260e+09	1.239141e+10	2.070356e+10	1.282614e+10	1.253336e+10	1.331341e+10	1.041344e+10
	max	9.357260e+09	3.344836e+10	1.614353e+11	7.625595e+10	5.917688e+10	2.901104e+10	2.612156e+10
Video Count	count	1.000000e+00	5.100000e+01	3.600000e+01	2.570000e+02	5.400000e+01	6.300000e+01	1.200000e+01
	mean	2.934000e+03	9.244784e+03	1.114250e+03	2.081012e+04	3.366111e+03	3.478048e+03	3.105167e+03
	std	NaN	4.748221e+04	8.116400e+02	4.417286e+04	5.818682e+03	3.053365e+03	2.257266e+03
	min	2.934000e+03	8.200000e+01	1.060000e+02	2.200000e+01	2.000000e+01	1.710000e+02	5.290000e+02
	25%	2.934000e+03	5.770000e+02	5.932500e+02	7.900000e+02	1.027000e+03	1.425500e+03	1.239000e+03
	50%	2.934000e+03	8.330000e+02	8.265000e+02	2.773000e+03	1.534500e+03	2.666000e+03	2.336500e+03
	75%	2.934000e+03	2.391000e+03	1.424500e+03	1.386200e+04	3.736500e+03	4.698500e+03	5.478750e+03
	max	2.934000e+03	3.384830e+05	3.397000e+03	3.458230e+05	3.887100e+04	1.551600e+04	6.527000e+03
Started	count	1.000000e+00	5.100000e+01	3.600000e+01	2.570000e+02	5.400000e+01	6.300000e+01	1.200000e+01
	mean	2.013000e+03	2.014784e+03	2.014083e+03	2.013241e+03	2.012000e+03	2.012302e+03	2.016083e+03
	std	NaN	5.231878e+00	3.516289e+00	4.426943e+00	3.776791e+00	3.504076e+00	2.843120e+00
	min	2.013000e+03	2.005000e+03	2.006000e+03	2.005000e+03	2.005000e+03	2.006000e+03	2.009000e+03
	25%	2.013000e+03	2.011000e+03	2.012750e+03	2.010000e+03	2.009250e+03	2.011000e+03	2.015000e+03
	50%	2.013000e+03	2.016000e+03	2.014000e+03	2.013000e+03	2.012500e+03	2.012000e+03	2.016500e+03
	75%	2.013000e+03	2.020000e+03	2.016000e+03	2.016000e+03	2.014750e+03	2.014000e+03	2.017250e+03
	max	2.013000e+03	2.021000e+03	2.020000e+03	2.021000e+03	2.020000e+03	2.021000e+03	2.021000e+03



Trực quan hóa và phân tích dữ liệu

Tạo boxplot cho các cột số trong DataFrame và hiển thị chúng

In [11]:

```
1 # Tạo danh sách các cột dữ liệu số
2 numerical_cols = [feature for feature in df.columns if df[feature].dtype != "O"]
3
4 # Tạo figure với số hàng bằng số cột dữ liệu số và chỉ có 1 cột
5 fig = sp.make_subplots(rows=len(numerical_cols), cols=1)
6
7 # Mảng màu sắc cho các boxplot
8 colors = px.colors.qualitative.Plotly
9
10 # Với mỗi cột dữ liệu số, thêm một trace boxplot vào figure
11 for i, column in enumerate(numerical_cols):
12     fig.add_trace(px.box(x=df[column]).data[0], row=i+1, col=1)
13
14     # Cập nhật màu sắc cho trace boxplot
15     fig.update_traces(marker=dict(color=colors[i]), row=i+1, col=1)
16
17     # Cập nhật trục x của từng trace boxplot
18     for i, column in enumerate(numerical_cols):
19         fig.update_xaxes(title_text=column, row=i+1, col=1)
20
21     # Cập nhật layout của figure
22 fig.update_layout(title="Biểu đồ Boxplot cho các cột số",
23                   title_font_size=16,
24                   height=200*len(numerical_cols),
25                   showlegend=False)
26
27 # Hiển thị figure
28 fig.show()
29
```

Biểu đồ Boxplot cho các cột số

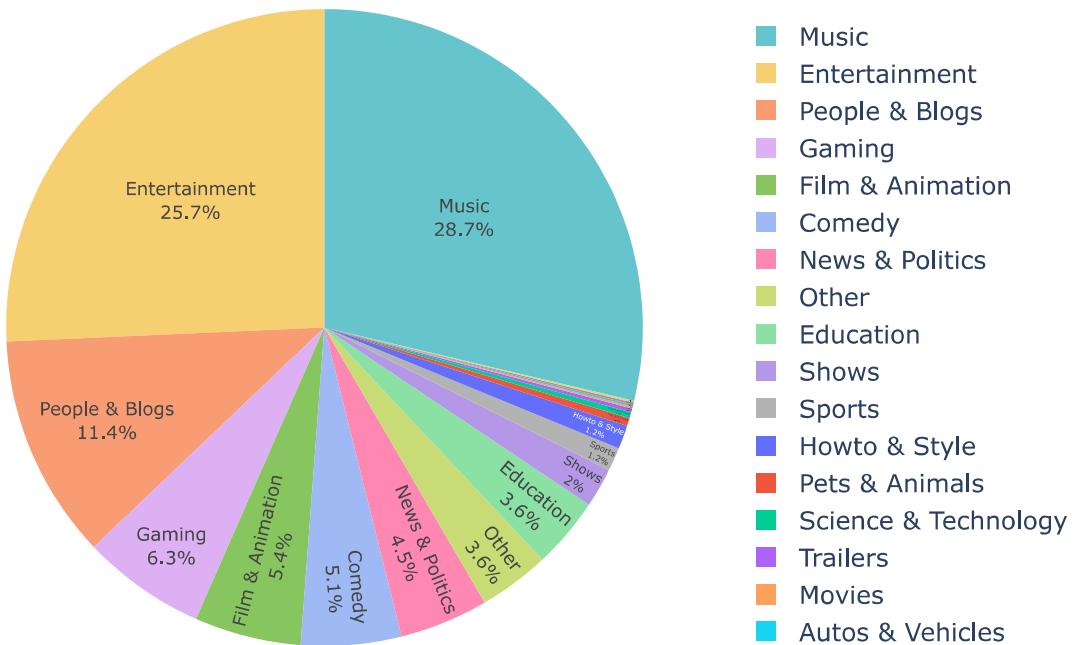


Mục đích của bước này là tạo các boxplot để trực quan hóa phân phối và biểu diễn thông tin thống kê của các cột số trong DataFrame. Boxplot cung cấp thông tin về phạm vi giá trị, giá trị trung vị (median), các phần vị (quartiles) và các giá trị ngoại lệ (outliers) của dữ liệu. Việc hiển thị các boxplot cho các cột số giúp chúng ta có cái nhìn tổng quan về phân phối và sự biến động của dữ liệu trong từng cột.

Tạo biểu đồ pie chart về tỉ lệ các kênh YouTube theo danh mục

```
In [12]: 1 # Đếm số Lượng kênh YouTube theo danh mục
2 categories = df['Category'].value_counts()
3
4 # Tạo biểu đồ pie chart (biểu đồ hình tròn) với giá trị là số Lượng kênh và tên là tên danh mục
5 fig = px.pie(values=categories.values, names=categories.index)
6
7 # Cập nhật giao diện biểu đồ
8 fig.update_layout(
9     title='Tỉ Lệ Các Kênh Youtube Theo Danh Mục', # Tiêu đề của biểu đồ
10    font_size=15, # Cỡ chữ
11    title_x=0.5 # Vị trí tiêu đề theo trục x
12 )
13
14 # Cập nhật hiển thị dữ liệu trên biểu đồ
15 fig.update_traces(
16     textposition='inside', # Hiển thị nhãn văn bản bên trong phần hình tròn
17     textfont_size=11, # Cỡ chữ của nhãn văn bản
18     textinfo='percent+label', # Hiển thị thông tin bao gồm tỷ lệ phần trăm và nhãn tên
19     marker=dict(colors=px.colors.qualitative.Pastel)
20 )
21
22 # Hiển thị biểu đồ
23 fig.show()
24
```

Tỉ Lệ Các Kênh Youtube Theo Danh Mục



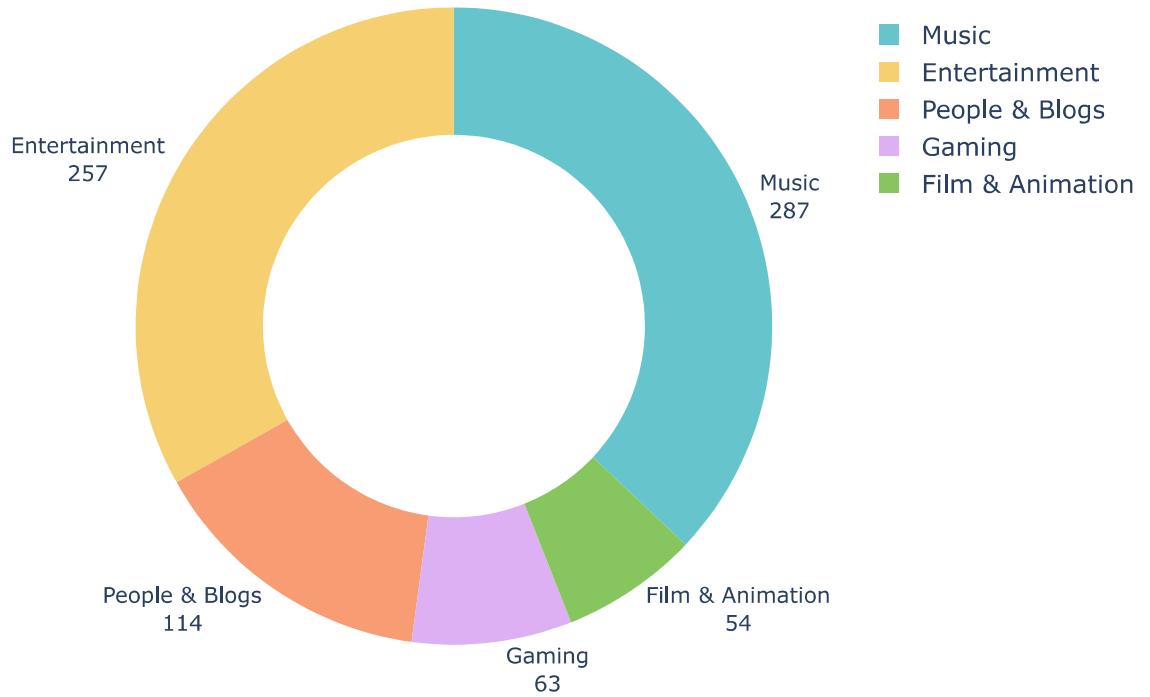
Mã trên sẽ đếm số lượng kênh YouTube theo danh mục, sau đó tạo một biểu đồ pie chart để trực quan hóa dữ liệu. Biểu đồ sẽ hiển thị tỉ lệ phần trăm của mỗi danh mục và tên của từng danh mục.

Ta thấy được các kênh thuộc loại Entertainment và Music chiếm số lượng nhiều nhất.

In [13]:

```
1 # Chọn chỉ 5 danh mục hàng đầu để hiển thị trên biểu đồ pie
2 fig = px.pie(values=categories[:5].values, names=categories[:5].index, hole=.6)
3 # Tạo biểu đồ pie với giá trị là số Lượng kênh của 5 danh mục hàng đầu và tên là tên của các danh mục
4
5 fig.update_layout(title='Top 5 Kênh Youtube Theo Danh Mục', font_size=15, title_x=0.5)
6 # Cập nhật tiêu đề cho biểu đồ là "Top 5 Kênh Youtube Theo Danh Mục", cỡ chữ là 15 và tiêu đề được
7
8 fig.update_traces(textposition='outside', textfont_size=13, textinfo='label + value', marker=dic
9 # Cập nhật vị trí văn bản là ngoài hình tròn, cỡ chữ của văn bản là 13 và hiển thị nhãn và giá trị
10
11 fig.show()
12 # Hiển thị biểu đồ.
13
```

Top 5 Kênh Youtube Theo Danh Mục



Mục đích của đoạn code này là tạo một biểu đồ pie cho danh sách Top 5 danh mục YouTube dựa trên số lượng kênh.

Có một số lập luận có thể được rút ra từ việc Entertainment, Music, ... chiếm số lượng kênh nhiều nhất trên YouTube:

Nhu cầu giải trí của con người là một yếu tố quan trọng trong việc sử dụng YouTube. Với sự phát triển của công nghệ và internet, người dùng thường xem YouTube như một nguồn giải trí chính, và do đó, nhu cầu xem các nội dung giải trí như video hài, phim, chương trình truyền hình, ca nhạc, v.v. tăng cao.

Entertainment và Music là hai lĩnh vực phổ biến và có sức hút lớn đối với công chúng. Người dùng thường tìm kiếm các video giải trí, nhạc, MV, live show, v.v. trên YouTube để giải trí và thư giãn. Các kênh trong danh mục này thường cung cấp nội dung hấp dẫn và mang tính giải trí cao, thu hút một lượng lớn người xem và đăng ký.

Có sự đa dạng về danh mục trên YouTube, nhưng Entertainment và Music được đầu tư và phát triển mạnh mẽ. Do tiềm năng kinh doanh và sự quan tâm của người dùng, nhiều người sáng tạo và các doanh nghiệp đã tạo ra nhiều kênh chuyên về giải trí và âm nhạc để khai thác thị trường này. Điều này giải thích tại sao Entertainment và Music chiếm tỉ lệ lớn trong số lượng kênh trên YouTube.

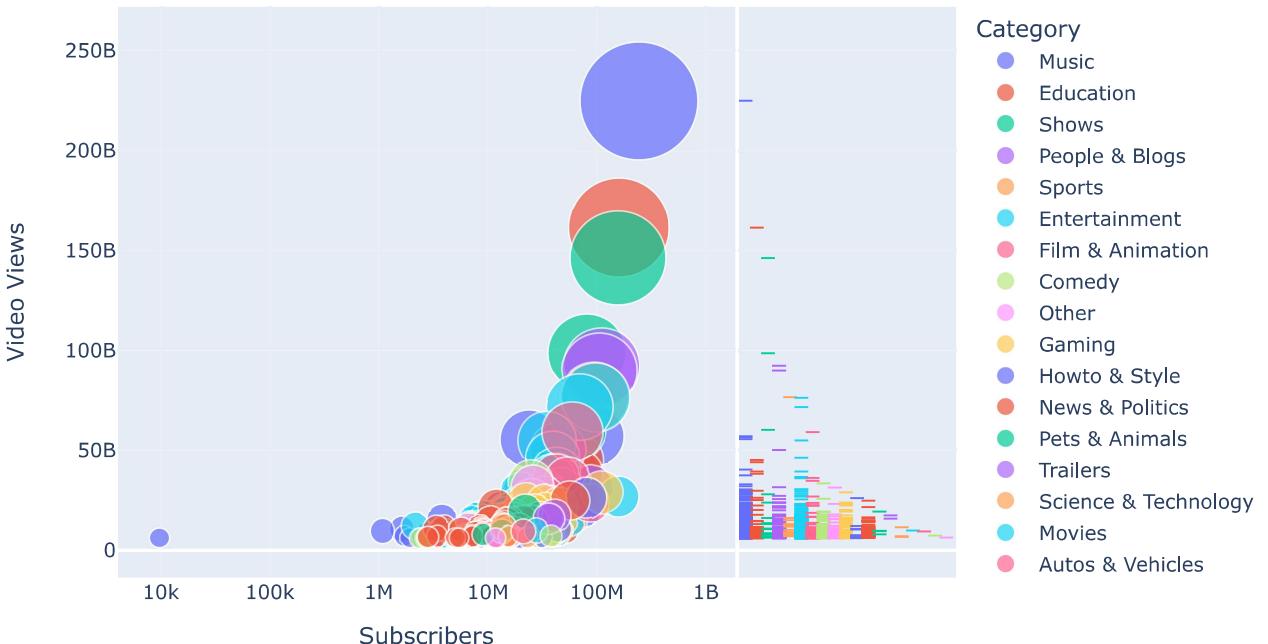
Tuy nhiên, lưu ý rằng lập luận trên chỉ dựa trên việc phân tích tỉ lệ phần trăm của mỗi danh mục trên biểu đồ pie chart và có thể không áp dụng cho tất cả các trường hợp hoặc không phản ánh toàn bộ thực tế của thị trường YouTube.

Biểu đồ Scatter: Số lượng người đăng ký và lượt xem video theo danh mục

In [14]:

```
1 fig = px.scatter(df, x="Subscribers", y="Video Views",
2                   size="Video Views", color="Category",
3                   log_x=True, size_max=50,
4                   title="Lượt Xem Và Lượng Người Đăng Ký Kênh Theo Danh Mục",
5                   marginal_y='rug')
6 fig.show()
```

Lượt Xem Và Lượng Người Đăng Ký Kênh Theo Danh Mục



Bước tạo biểu đồ scatter plot nhằm mục đích hiển thị mối quan hệ giữa số lượng người đăng ký kênh và lượt xem video trên YouTube. Khi xét cả hai yếu tố này, ta nhận thấy có sự thay đổi và xuất hiện các danh mục phổ biến khác như Education và Shows, mặc dù chúng không chiếm tỉ lệ cao trong số lượng kênh.

Sự xuất hiện của các danh mục khác như Education và Shows có thể được giải thích bằng một số lý do sau:

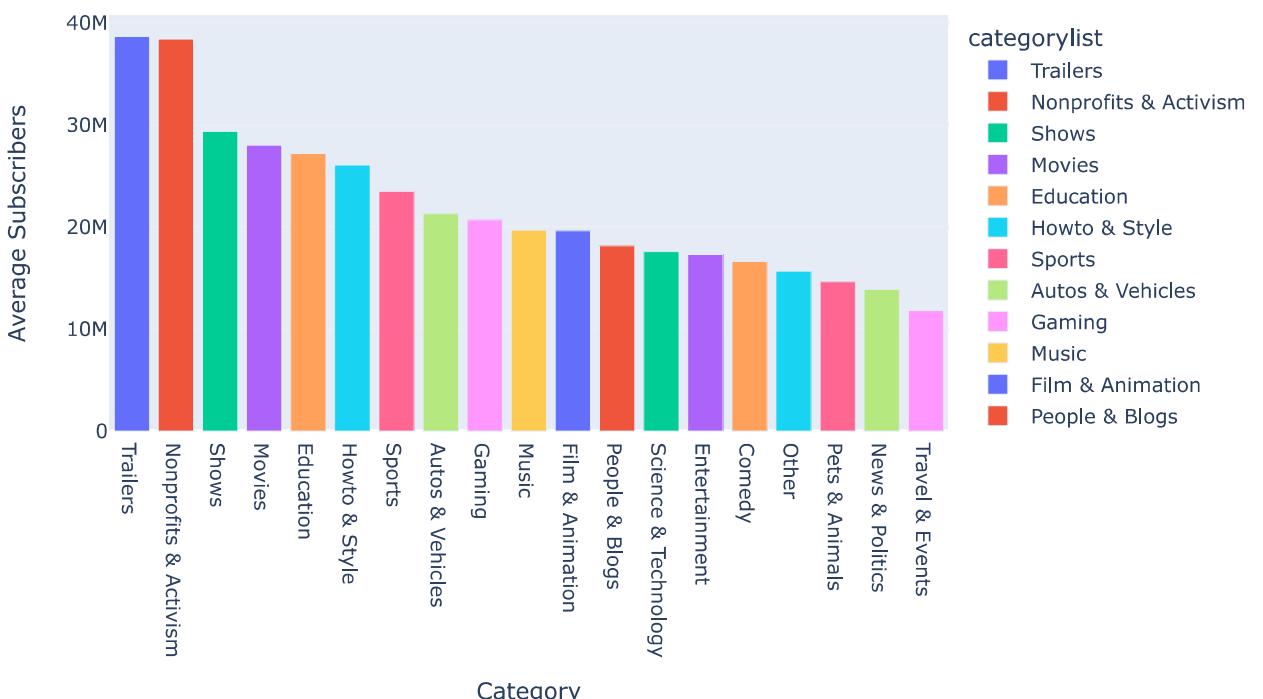
- Các nội dung giáo dục: Các kênh giáo dục trên YouTube thường cung cấp nội dung bổ ích, hướng dẫn, khóa học và tư vấn trong nhiều lĩnh vực khác nhau. Người xem có thể đăng ký các kênh này để theo dõi và học hỏi, dẫn đến việc có một lượng lớn người đăng ký và lượt xem cao.
- Các chương trình truyền hình và sự kiện trực tiếp: Shows và các sự kiện trực tiếp trên YouTube như livestream, concert, gameshow, v.v. thu hút sự quan tâm của đông đảo khán giả. Điều này dẫn đến việc có một lượng lớn lượt xem và người đăng ký trên các kênh chuyên về shows và sự kiện này.

Lập luận trên cho thấy rằng, trong số lượng kênh ít nhưng có lượt xem và người đăng ký cao, có sự đa dạng về danh mục và nguyên nhân là do những kênh có nội dung đặc biệt như ca sĩ được fan hâm mộ xem đi xem lại nhiều lần hoặc các video giáo dục và shows có lượng xem lớn. Điều này giải thích sự chuyển dịch và sự đa dạng của top danh mục kênh trên YouTube.

Số Người Đăng Ký Kênh Youtube Trung Bình Theo Danh Mục

```
In [15]: 1 category_list = list(df.Category.unique())
2 sub = []
3
4 for i in category_list:
5     x = df[df.Category == i]
6     mean_sub = x["Subscribers"].mean()
7     sub.append(mean_sub)
8 df_subs = pd.DataFrame({'categorylist': category_list, 'subs': sub})
9 sorted_df_subs = df_subs.sort_values('subs', ascending=False)
10
11 fig = px.bar(sorted_df_subs, x='categorylist', y='subs', color='categorylist',
12               title='Số Người Đăng Ký Trung Bình Theo Danh Mục')
13 fig.update_layout(xaxis_title='Category', yaxis_title='Average Subscribers', title_font_size=20)
14
15 fig.show()
16
```

Số Người Đăng Ký Trung Bình Theo Danh Mục



Bước vẽ biểu đồ scatter plot đã giúp ta nhìn thấy mối quan hệ giữa số lượng người đăng ký kênh và lượt xem video trên YouTube. Khi xem xét cả hai yếu tố này, chúng ta đã nhận thấy xuất hiện các danh mục kênh phổ biến khác như Education và Shows.

Tuy nhiên, để kiểm tra lập luận mới, cần đưa ra bằng chứng từ biểu đồ scatter plot. Theo biểu đồ, ta nhìn thấy các danh mục kênh có số lượng đăng ký cao nhất là 'Trailers' và 'Nonprofits & Activism'. Điều này cho thấy những kênh trong các danh mục này thường đăng các video Trailer và nội dung liên quan đến các tổ chức phi lợi nhuận và hoạt động xã hội.

Có một số giải thích có thể được áp dụng để lý giải sự phổ biến của các danh mục này:

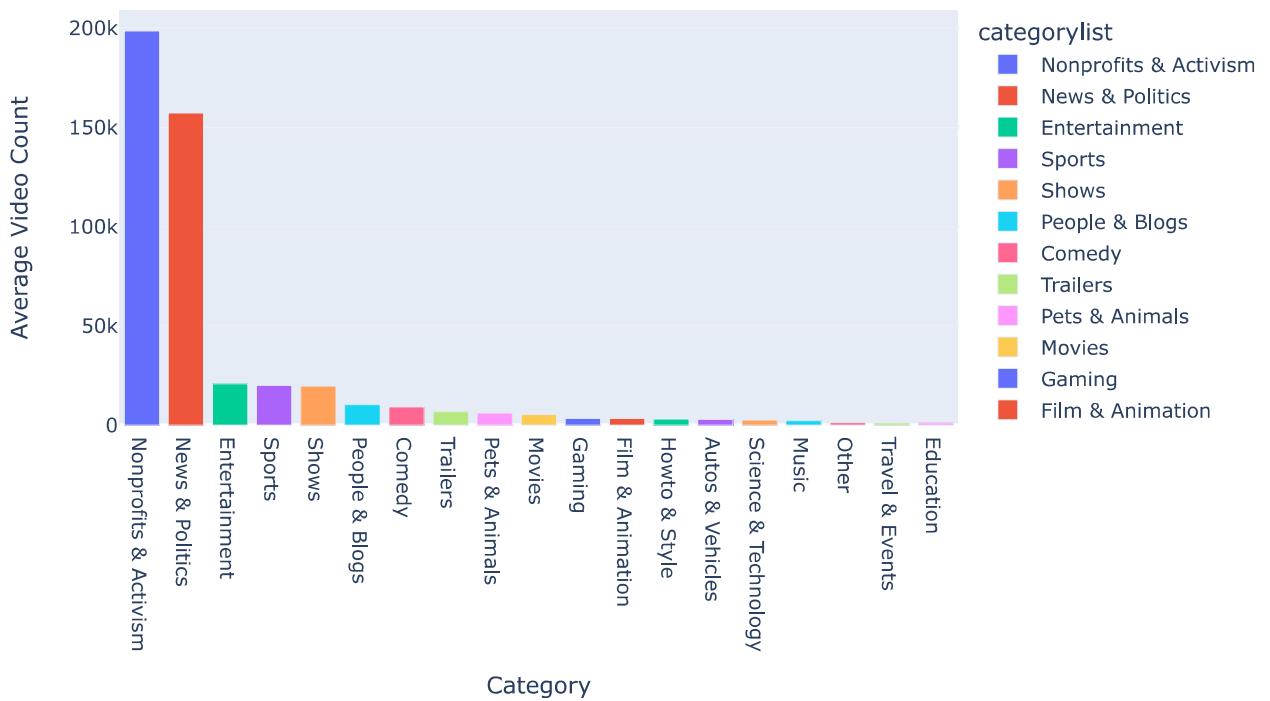
- Kênh 'Trailers': Các kênh này thường chuyên đăng tải các trailer của phim, chương trình truyền hình, hoặc video quảng cáo. Việc đăng ký kênh này sẽ giúp người dùng không bỏ lỡ bất kỳ trailer nào của các tác phẩm mà họ quan tâm. Điều này đặc biệt hấp dẫn đối với những người yêu điện ảnh và muốn nắm bắt thông tin về các bộ phim mới.
- Kênh 'Nonprofits & Activism': Các kênh này thường liên quan đến các tổ chức phi lợi nhuận và hoạt động xã hội, nhằm tăng cường nhận thức và thu hút sự quan tâm từ cộng đồng. Những kênh này thường đăng tải video về hoạt động từ thiện, nỗ lực giúp đỡ cộng đồng, hoặc chia sẻ thông tin về các vấn đề xã hội nhạy cảm. Người dùng có thể quan tâm đến các vấn đề này và đăng ký kênh để được cập nhật thông tin và tham gia vào các hoạt động xã hội.

Tổng quan, việc phân tích biểu đồ scatter plot cùng với sự thay đổi của các danh mục kênh trên YouTube cho thấy sự đa dạng và lý do phổ biến của các danh mục kênh. Những kênh có số lượng đăng ký cao như 'Trailers' và 'Nonprofits & Activism' phản ánh xu hướng người dùng quan tâm đến việc không bỏ lỡ các thông tin mới về điện ảnh và các hoạt động xã hội.

Số lượng video trung bình của kênh Youtube theo danh mục

```
In [16]: 1 category_list = list(df.Category.unique())
2 v_count = []
3
4 for i in category_list:
5     x = df[df.Category == i]
6     mean_count = x["Video Count"].mean()
7     v_count.append(mean_count)
8 df_count = pd.DataFrame({'categorylist': category_list, 'VideoCount': v_count})
9 sorted_df_count = df_count.sort_values('VideoCount', ascending=False)
10
11 fig = px.bar(sorted_df_count, x='categorylist', y='VideoCount', color='categorylist',
12                 title='Số Lượng Video Trung Bình Theo Danh Mục')
13 fig.update_layout(xaxis_title='Category', yaxis_title='Average Video Count', title_font_size=20)
14
15 fig.show()
16
```

Số Lượng Video Trung Bình Theo Danh Mục



Dựa vào đồ thị, chúng ta nhận thấy có hai danh mục kênh là 'Nonprofits & Activism' và 'News & Politics' có số lượng video trung bình vượt trội so với các danh mục khác. Điều này cho thấy các kênh thuộc hai danh mục này thường đăng tải một lượng lớn video hơn so với các danh mục khác.

Lời giải thích cho hiện tượng này có thể được áp dụng như sau:

- Kênh 'Nonprofits & Activism': Đây là danh mục chứa các kênh liên quan đến các tổ chức phi lợi nhuận và hoạt động xã hội, nhằm tăng cường nhận thức và gây quỹ cho các vấn đề xã hội. Các kênh trong danh mục này có xu hướng đăng tải nhiều video nhằm chia sẻ thông tin về các hoạt động từ thiện, nỗ lực giúp đỡ cộng đồng và chia sẻ câu chuyện của những người cần sự giúp đỡ. Số lượng video lớn có thể phản ánh sự đa dạng của các hoạt động từ thiện và nhu cầu chia sẻ thông tin với cộng đồng.
- Kênh 'News & Politics': Đây là danh mục chứa các kênh liên quan đến tin tức và chính trị. Các kênh trong danh mục này thường đăng tải nhiều video nhằm cung cấp thông tin, bình luận và phân tích về các sự kiện, tin tức và chính sách đang diễn ra trong lĩnh vực này. Với tính chất nhanh chóng và thường xuyên cập nhật của lĩnh vực tin tức và chính trị,

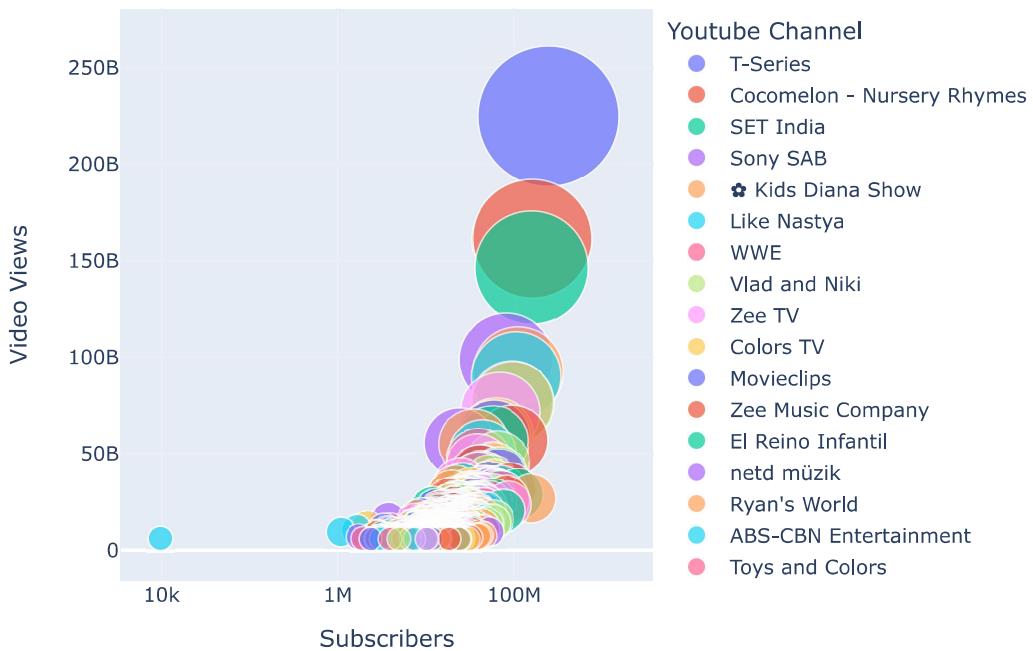
số lượng video lớn có thể phản ánh nhu cầu cung cấp thông tin nhanh chóng và đa dạng về các vấn đề quan trọng đang xảy ra trong xã hội.

Tổng quan, việc phân tích đồ thị số lượng video trung bình của các danh mục kênh trên YouTube cho thấy sự khác biệt về số lượng video đăng tải giữa các danh mục. Các danh mục 'Nonprofits & Activism' và 'News & Politics' đặc biệt nổi bật với số lượng video trung bình cao hơn, thể hiện sự đa dạng và nhu cầu cung cấp thông tin của các kênh thuộc danh mục này.

Lượt xem và đăng ký video theo kênh Youtube

```
In [17]: 1 fig = px.scatter(df, x="Subscribers", y="Video Views", size="Video Views", color="Youtube Channel",  
2                      title="Lượt Xem Và Lượng Người Đăng Ký Kênh Theo Từng Kênh")  
3 fig.show()
```

Lượt Xem Và Lượng Người Đăng Ký Kênh Theo Từng Kênh

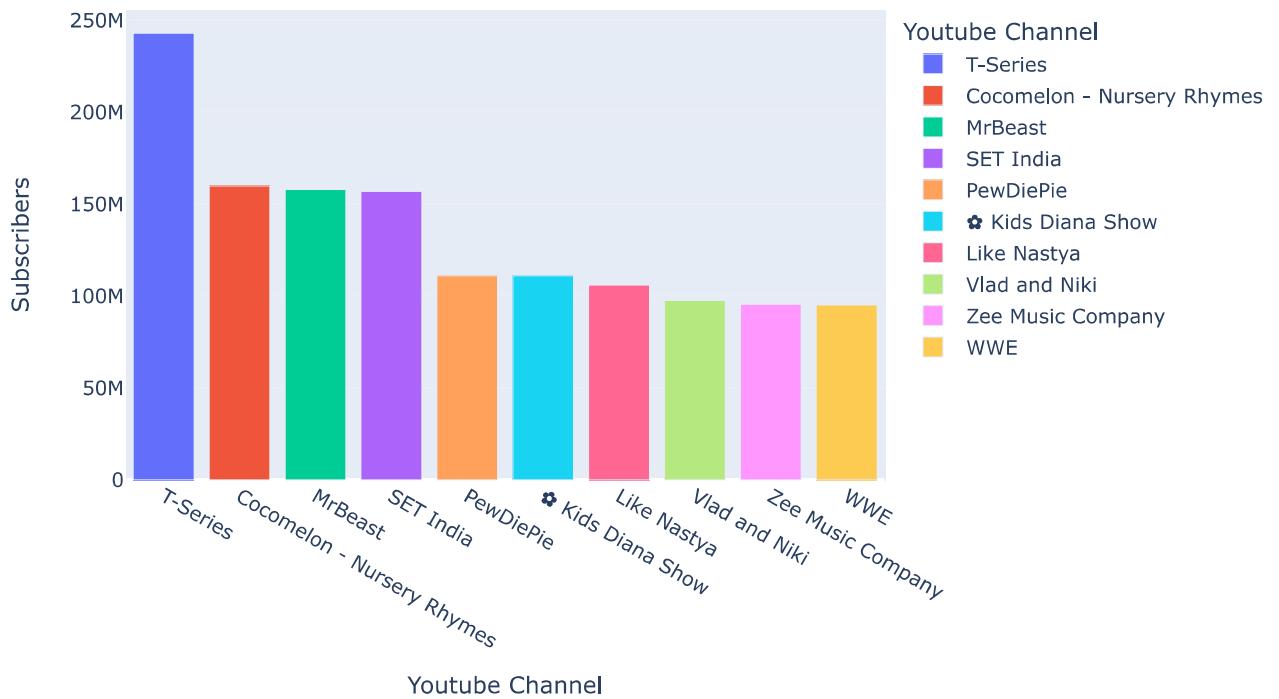


Biểu đồ trên cho ta thấy top các kênh có lượt đăng ký kênh và lượt xem nhiều nhất.

Top 10 kênh có số đăng ký và lượt xem video nhiều nhất

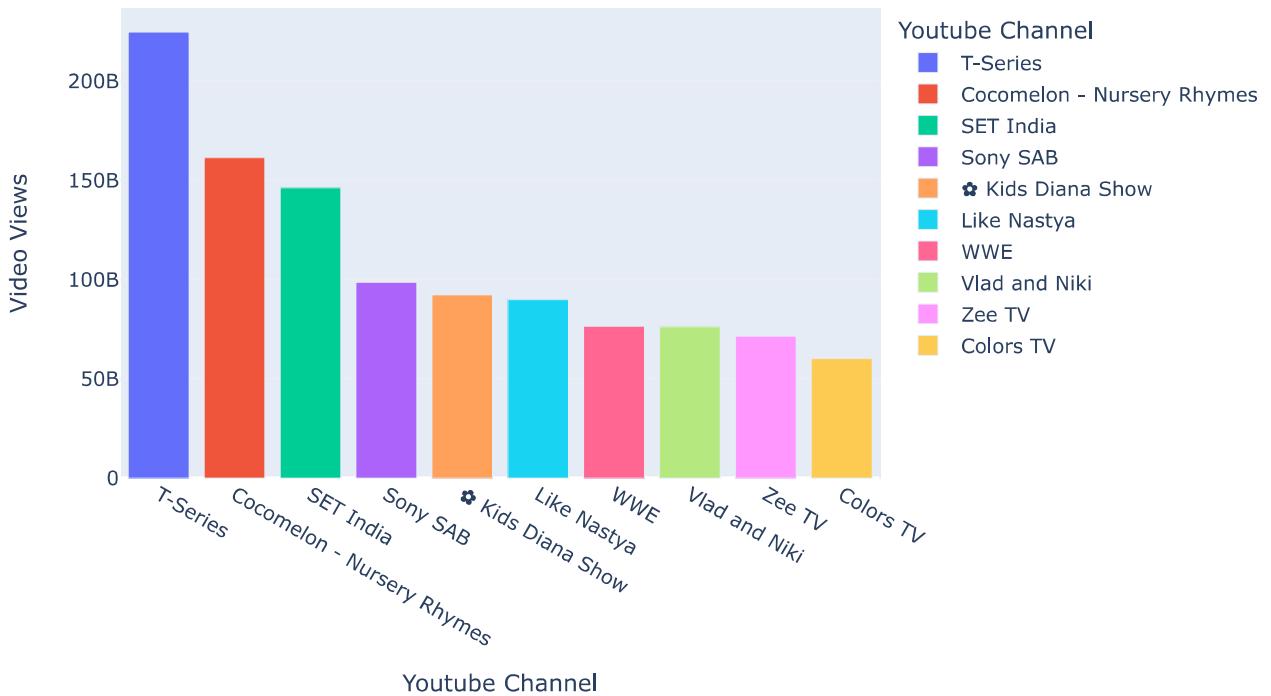
```
In [18]: 1 var = df.groupby(['Youtube Channel'])['Subscribers'].sum().sort_values(ascending=False).head(10)
2 var = var.reset_index()
3
4 fig = px.bar(var, x='Youtube Channel', y='Subscribers', color='Youtube Channel',
5               title='Top 10 Kênh Có Số Lượt Đăng Ký Kênh Nhiều Nhất')
6 fig.update_layout(xaxis_title='Youtube Channel', yaxis_title='Subscribers', title_font_size=20)
7
8 fig.show()
9
```

Top 10 Kênh Có Số Lượt Đăng Ký Kênh Nhiều Nhất



```
In [19]: 1 var = df.groupby(['Youtube Channel'])['Video Views'].sum().sort_values(ascending=False).head(10)
2 var = var.reset_index()
3
4 fig = px.bar(var, x='Youtube Channel', y='Video Views', color='Youtube Channel',
5               title='Top 10 Kênh Có Số Lượt Xem Video Nhiều Nhất')
6 fig.update_layout(xaxis_title='Youtube Channel', yaxis_title='Video Views', title_font_size=20)
7
8 fig.show()
9
```

Top 10 Kênh Có Số Lượt Xem Video Nhiều Nhất



Dựa vào 2 biểu đồ là "Top 10 Kênh Có Số Lượt Đăng Ký Kênh Nhiều Nhất" và "Top 10 Kênh Có Số Lượt Xem Video Nhiều Nhất", ta bắt gặp nhiều tên quen thuộc. Các kênh này thuộc top các kênh nổi tiếng toàn thế giới về số lượng người đăng ký và lượt xem thu về một khoản thu nhập rất lớn từ nền tảng YouTube. Ta bắt gặp một kênh là "T-Series".

"T-Series" là một kênh YouTube nổi tiếng và đứng đầu trong danh sách các kênh có số lượng người đăng ký nhiều nhất trên toàn cầu. Super Cassettes Industries Private Limited, được biết với tên kinh doanh T-Series, là một công ty, hãng thu âm sản xuất phim nhạc Ấn Độ trong thể loại nhạc Bollywood và nhạc pop Ấn Độ, do Bhushan Kumar thành lập năm 1983. Điều này có thể được giải thích bằng sự ảnh hưởng của thị trường đông dân số của Ấn Độ. Với dân số đông đúc, Ấn Độ có một lượng lớn người dùng YouTube và lượng đăng ký của "T-Series" có thể được hiểu là một phần phản ánh sự quan tâm và ủng hộ từ cộng đồng người dùng Ấn Độ.

Tuy nhiên có sự khác biệt giữa số lượng người đăng ký và số lượt xem: Một điểm quan trọng cần lưu ý là không phải lúc nào các kênh có số lượng người đăng ký nhiều cũng có số lượt xem cao tương ứng. Điều này phụ thuộc vào rất nhiều yếu tố khác nhau, bao gồm:

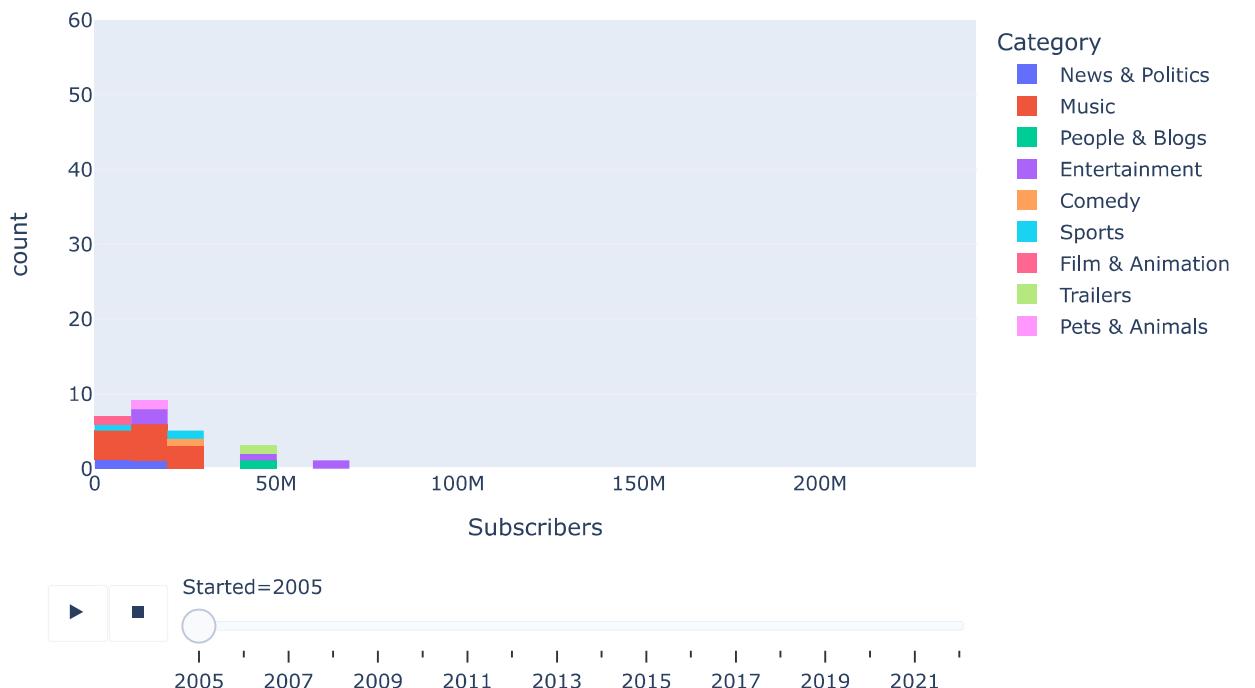
- Nội dung: Chất lượng nội dung và sự hấp dẫn của video có thể ảnh hưởng đáng kể đến lượt xem. Một kênh có số lượng đăng ký lớn không đảm bảo rằng mọi video của họ đều được người dùng quan tâm và xem.
- Chiến lược tiếp thị: Các kênh có chiến lược tiếp thị hiệu quả và khả năng tương tác với khán giả có thể thu hút được lượng lớn lượt xem, ngay cả khi số lượng đăng ký không quá cao.
- Thời gian hoạt động: Thời gian mà kênh đã tồn tại và tích lũy lượng người xem theo thời gian có thể ảnh hưởng đến số lượng lượt xem.
- Ngoài ra còn có thể do lượng người dùng đăng ký kênh : do lượng người dùng ảo không dùng nick nữa, ...

Do đó, để hiểu rõ hơn về sự khác biệt giữa số lượng đăng ký và lượt xem của các kênh YouTube, cần xem xét các yếu tố nêu trên và các yếu tố khác như chất lượng nội dung, phân phối video và sự tương tác với khán giả.

Biểu đồ động giúp ta quan sát sự biến động về lượng người đăng ký theo thời gian

```
In [20]: 1 px.histogram(df.query('Started > 1970').sort_values('Started', ascending=True), x="Subscribers",
2           range_x=[df['Subscribers'].min(), df['Subscribers'].max()],
3           range_y=[0, 60],
4           animation_frame="Started", title="Lượng Đăng Ký Theo Thời Gian Tạo Kênh")
```

Lượng Đăng Ký Theo Thời Gian Tạo Kênh

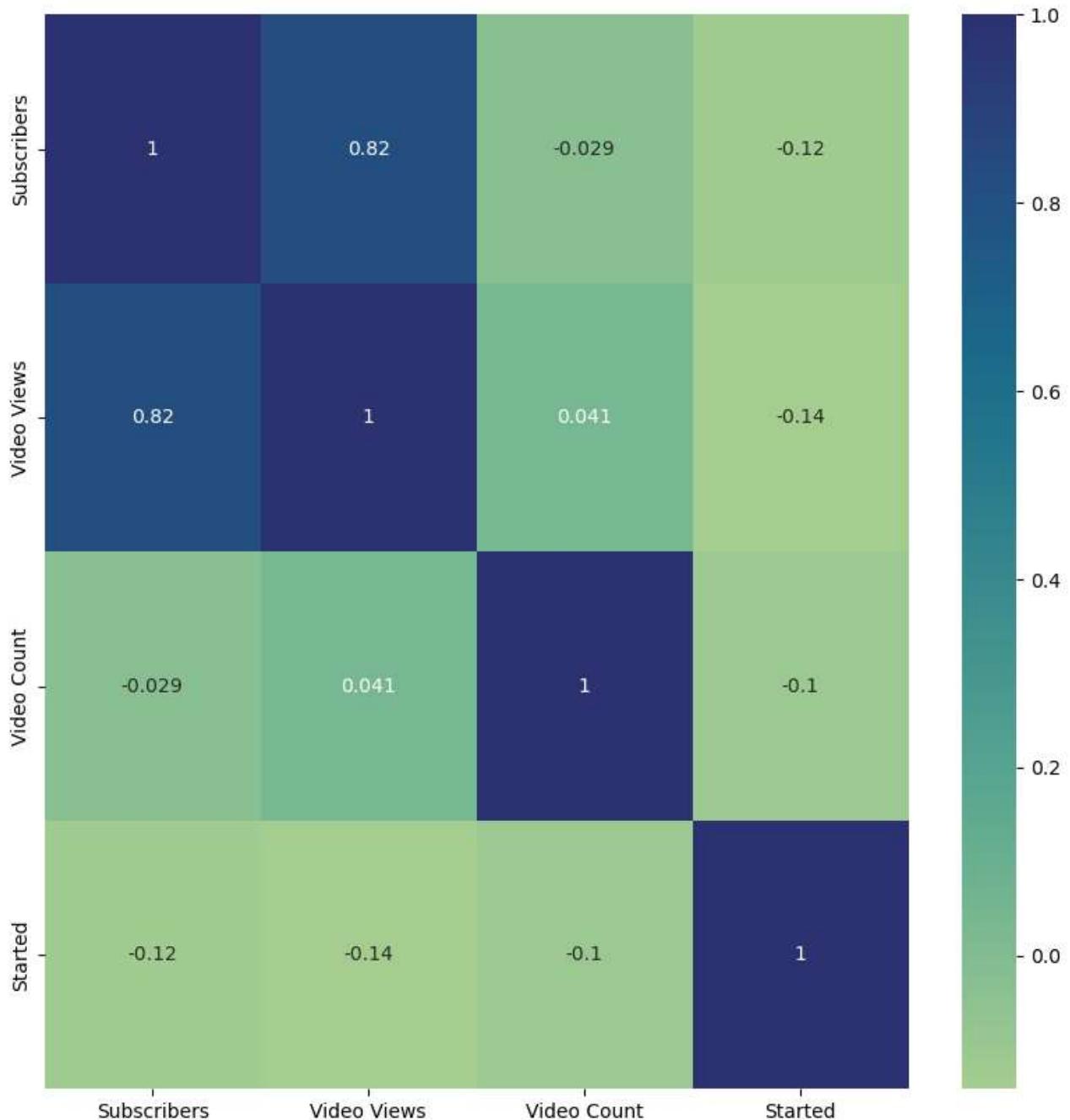


Mối quan hệ giữa các cột trong bộ dữ liệu

Heatmap

```
In [21]: 1 plt.figure(figsize=(10, 10))
2 corr = df.corr(numeric_only=True)
3 sns.heatmap(corr, annot=True, cmap="crest")
```

Out[21]: <Axes: >



Biểu đồ heatmap cho phép ta nhìn thấy mức độ tương quan giữa các cặp biến số trong DataFrame. Các ô có màu sáng hơn đại diện cho tương quan mạnh hơn, trong khi các ô có màu tối hơn đại diện cho tương quan yếu hơn. Giá trị tương quan cụ thể của mỗi cặp biến số được hiển thị trên biểu đồ, cho phép ta dễ dàng đọc và hiểu sự tương quan giữa các biến số.

Trong trường hợp này, biểu đồ heatmap cho thấy một số mối tương quan quan trọng giữa các biến số. Cụ thể:

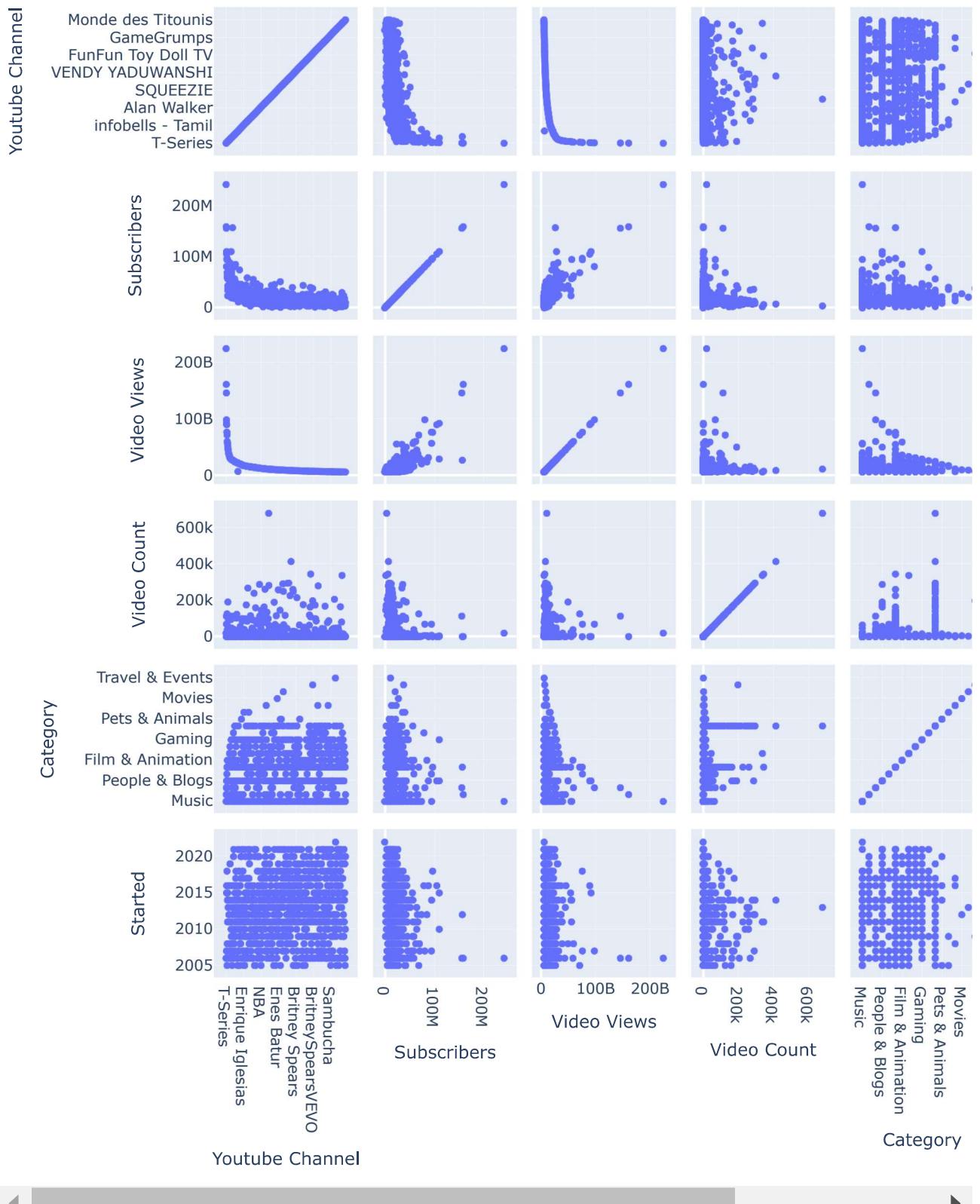
- Số lượng người đăng ký (Subscribers) có mối tương quan chặt chẽ với Lượt xem video (Video Views). Các kênh có nhiều người đăng ký thường có lượt xem video cao và ngược lại. Điều này cho thấy sự ảnh hưởng của số lượng người đăng ký đến lượng lượt xem của một kênh.
- Số lượng người đăng ký không có mối tương quan đáng kể với Số lượng video (Video Count). Điều này có nghĩa là một kênh có nhiều video không đồng nghĩa với việc có nhiều người đăng ký. Có thể có các kênh có số lượng video lớn nhưng không thu hút được nhiều người đăng ký.
- Có mối tương quan yếu giữa Lượt xem video và Số lượng video. Tuy nhiên, tương quan này không đủ mạnh để có thể kết luận rằng một kênh có nhiều video sẽ có lượt xem cao. Có thể có các kênh có số lượng video ít nhưng có lượt xem lớn, hoặc ngược lại.

Tóm lại, biểu đồ heatmap giúp chúng ta hiểu được mức độ tương quan giữa các biến số trong bộ dữ liệu. Nó cung cấp cái nhìn toàn diện về mối quan hệ giữa số lượng người đăng ký, lượt xem video và số lượng video trên các kênh YouTube.

In [22]:

```
1 fig = px.scatter_matrix(df)
2 fig.update_layout(
3     width=1000,
4     height=1000,
5 )
```

D:\Anaconda\lib\site-packages\plotly\express_core.py:279: FutureWarning:
iteritems is deprecated and will be removed in a future version. Use .items instead.



Đoạn mã trên tạo ra một scatter matrix plot bằng cách sử dụng thư viện Plotly Express (px). Scatter matrix plot là một biểu đồ hiển thị tất cả các cặp biến số trong DataFrame dưới dạng các scatter plot. Mỗi scatter plot biểu thị mối quan hệ giữa hai biến số.

Cụ thể, biểu đồ này hiển thị một ma trận các scatter plot với các biến số trong DataFrame. Trên đường chéo chính của ma trận, chúng ta thấy các histogram của từng biến số để hiểu phân phối của chúng. Các ô trên đường chéo chính không chứa scatter plot mà chỉ có một đường thẳng thể hiện mối quan hệ tương đương của cùng một biến với chính nó.

Scatter matrix plot giúp ta dễ dàng nhận ra mối quan hệ giữa các biến số trong DataFrame, xem xét sự tương quan và phân bố dữ liệu. Nó là một công cụ hữu ích trong việc khám phá dữ liệu và tìm hiểu mối quan hệ giữa các biến số.

Thể loại Music có số lượt xem cao nhất so với các thể loại khác.

Rõ ràng rằng số lượt xem video và số lượng người đăng ký có mối quan hệ trực tiếp. Điều này có nghĩa là khi số lượng người đăng ký tăng, số lượt xem video cũng sẽ tăng theo.

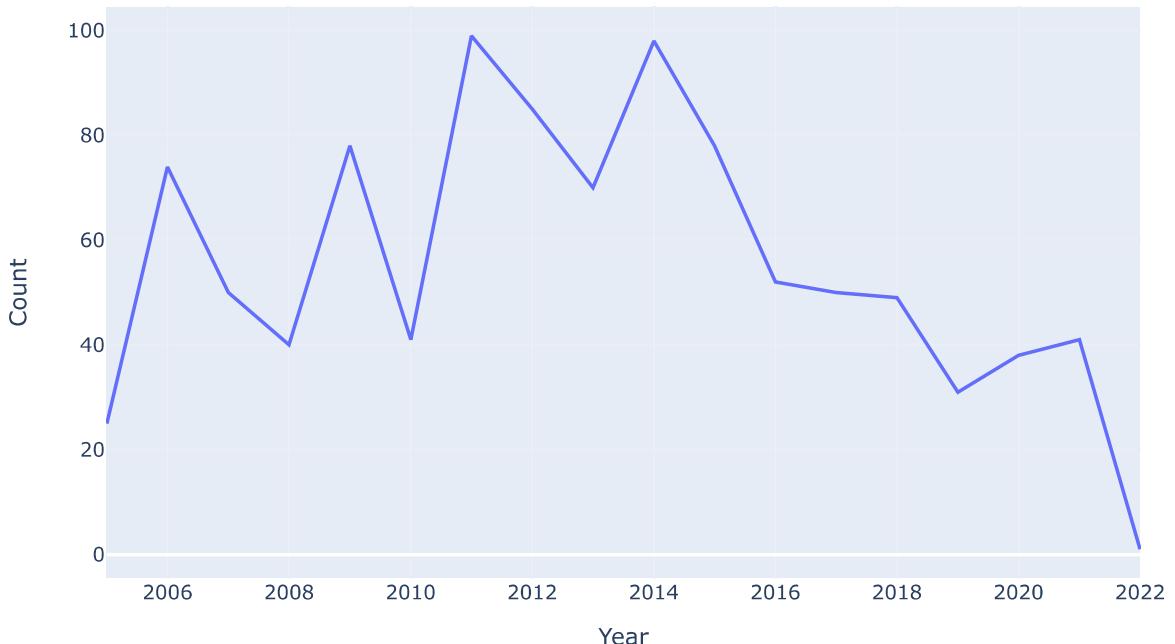
Các nhận định trên cho thấy một số xu hướng và mối quan hệ giữa các biến số trong bộ dữ liệu. Chẳng hạn, thể loại Music và Entertainment đều có xu hướng và có số lượng người đăng ký và lượt xem cao. Việc hiểu và nhận ra các mối quan hệ

Xu hướng và sự bùng nổ các kênh Youtube

In [23]:

```
1 year = df['Started'].value_counts().reset_index()
2 year.columns = ['Year', 'Count']
3 year = year.sort_values('Year') # Sắp xếp dữ liệu theo năm
4
5 fig = px.line(x=year['Year'], y=year['Count'], title='Xu Hướng Các Kênh YouTube Được Tạo Hàng Năm')
6 fig.update_layout(xaxis_title='Year', yaxis_title='Count')
7
8 fig.show()
```

Xu Hướng Các Kênh YouTube Được Tạo Hàng Năm



Dựa trên các đồ thị và kết quả trên, chúng ta có thể đưa ra các nhận định sau:

- Từ đồ thị "Xu Hướng Các Kênh YouTube Được Tạo Hàng Năm", ta nhận thấy rằng số lượng kênh YouTube nổi tiếng đa số được tạo ra trong giai đoạn từ năm 2011 đến 2014. Từ đó, số lượng kênh này dần giảm và ít hơn trong các năm tiếp theo. Điều này cho thấy rằng việc xây dựng và phát triển một kênh YouTube thành công không phải là điều xảy ra ngay lập tức, mà đòi hỏi thời gian và công sức để đạt được sự ổn định và tăng trưởng.

- Bên cạnh đó, thông qua việc quan sát các thống kê và xu hướng trên YouTube, có thể nhận thấy chất lượng các kênh hiện nay đang có xu hướng giảm dần và khó thu hút được nhiều người xem. Điều này có thể được giải thích bởi sự cạnh tranh khốc liệt từ nhiều nền tảng truyền thông xã hội và sự đa dạng của các nhà sáng tạo nội dung trên YouTube. Để thu hút người xem và tạo sự khác biệt, các nhà sáng tạo nội dung phải liên tục tìm kiếm và sản xuất nội dung mới, hấp dẫn và sáng tạo hơn. Tuy nhiên, đôi khi điều này dẫn đến tình trạng nội dung bẩn hoặc không đáng xem xuất hiện, khi mà một số người tạo nội dung không ngần ngại sử dụng các phương pháp hoặc nội dung gây chú ý để thu hút lượng người xem.

Các nhận định trên cho thấy sự phức tạp và thay đổi của môi trường YouTube, cũng như sự cạnh tranh và áp lực mà các nhà sáng tạo nội dung phải đối mặt. Để thành công trên nền tảng này, nhà sáng tạo nội dung cần phải cải thiện chất lượng nội dung, tạo sự khác biệt và đáp ứng nhu cầu của người xem để tiếp tục thu hút và duy trì sự quan tâm của khán giả.

Kết luận

Dựa trên quá trình phân tích dữ liệu top 1000 YouTube Channels thông qua các bước thu thập, tiền xử lý, trực quan hóa và phân tích, chúng ta có thể đưa ra một kết luận tổng quát về bộ dữ liệu như sau:

- Phân tích danh mục kênh:** Các danh mục phổ biến nhất trên YouTube là Music, Entertainment, và People & Blogs. Điều này cho thấy rằng đa số người dùng YouTube tìm kiếm nội dung giải trí và liên quan đến con người. Ngoài ra, các danh mục như Education và Shows cũng có sự quan trọng và ảnh hưởng đáng kể.
- Sự tương quan giữa số lượng đăng ký, lượt xem và số lượng video:** Số lượng người đăng ký có mối tương quan chặt chẽ với lượt xem video, trong khi không có mối tương quan đáng kể với số lượng video. Điều này cho thấy rằng một kênh có nhiều video không nhất thiết sẽ có nhiều người đăng ký. Tuy nhiên, có một mối tương quan yếu giữa lượt xem video và số lượng video, cho thấy rằng sự đa dạng và chất lượng nội dung có thể ảnh hưởng đến lượt xem.
- Các kênh nổi tiếng nhất:** T-Series, Cocomelon Nursery Rhymes và MrBeats là ba kênh có số lượng người đăng ký nhiều nhất trên YouTube. Điều này chỉ ra sự ảnh hưởng của các kênh này và việc tạo ra thu nhập lớn từ YouTube. Đặc biệt, kênh T-Series đạt được số lượng người đăng ký cao nhất, điều này có thể được giải thích bởi dân số đông đúc của Ấn Độ.
- Phân tích các kênh theo năm thành lập:** Qua biểu đồ "Xu Hướng Các Kênh YouTube Được Tạo Hàng Năm", ta nhận thấy rằng số lượng kênh YouTube nổi tiếng được tạo ra nhiều nhất trong giai đoạn từ năm 2011 đến 2014. Điều này cho thấy rằng việc xây dựng một kênh thành công trên YouTube đòi hỏi thời gian và công sức để đạt được sự ổn định và tăng trưởng.
- Sự thay đổi và sự cạnh tranh trên YouTube:** Có những sự thay đổi và sự cạnh tranh trong các danh mục kênh và chất lượng nội dung trên YouTube. Các nhà sáng tạo nội dung phải liên tục tìm kiếm nội dung mới, hấp dẫn và sáng tạo để thu hút và duy trì sự quan tâm của khán giả. Điều này có thể dẫn đến việc xuất hiện các nội dung bẩn hoặc không đáng xem, khi mà một số người tạo nội dung không ngần ngại sử dụng các phương pháp gây chú ý để thu hút lượng người xem.
- Sự tương quan giữa lượt xem và số lượng người đăng ký:** Có sự tương quan trực tiếp giữa lượt xem và số lượng người đăng ký trên YouTube. Điều này cho thấy rằng khi số lượng người đăng ký tăng, lượt xem cũng tăng theo. Điều này có thể được giải thích bởi sự tương tác và quan tâm của khán giả đối với nội dung từ các kênh đã được họ đăng ký.

Tổng quát lại, bài phân tích dữ liệu trên top 1000 YouTube Channels đã cung cấp cho chúng ta một cái nhìn tổng quan về xu hướng và đặc điểm của các kênh YouTube nổi tiếng. Hiểu rõ các yếu tố ảnh hưởng đến sự thành công của một kênh trên YouTube có thể giúp các nhà sáng tạo nội dung và người dùng YouTube hiểu rõ hơn về cách tăng cường tương tác và tăng trưởng kênh của mình trên nền tảng này.

Tác giả: Huỳnh Bảo Khang - VNU-HCMUS

In []:

1