Project Capstone Modul 2
Data Analyst

Disusun oleh : Adha Ozy Prima Dewangga

US You tube Trending Data / (Data Youtube yang Trending di US)

Latar Belakang

BRAND AMBASSADOR

Google AdSense

YouTube

CONTENT CREATOR

Pernyataan Masalah

- Youtube ingin membuat kampanye yang akan melibatkan konten kreator yang akan membantu youtube dalam meningkatkan performa dan pendapatan youtube melalui program yang telah direncanakan.

- Youtube ingin mencari cara mengoptimasi iklan pada setiap jenis video pada youtube yang trending.

- Yang menjadi pertanyaan apakah jumlah tags , likes, describsi, publish_time, dan comments menjadi pengaruh besar banyak tidaknya views sehingga menentukan trending tidak suatu video.

- Apa yang menjadi faktor - faktor tolak ukur viewer banyak pada video yang trending sehingga pengiklanan menjadi lebih maksimal dan cara menentukan apakah konten kreator itu memenuhi syarat menjadi brand ambasador ?

- Apakah tags , likes, publish_time, and comments memiliki korelasi satu sama lain ?
- Apakah ada tidaknya describsi mempengaruhi viewer pada video yang di publish di Negara US ?
- Apakah ada tidaknya tags mempengaruhi viewer pada video yang di publish di Negara US ?
- Berapa lama waktu yang diperluhkan agar video menjadi trending di Negara US?
- Apakah banyak atau sedikitnya comments mempengaruhi viewer pada video yang di publish di Negara US sehingga menjadi trending ?
- Pada Negara US konten dengan kategori apa yang paling trending dan yang tidak ?
- Apakah tindakan disabled pada comments, ratings, dan video_error_or_removed mempengaruhi rank video yang di publish di Negara US?
- Cara menentukan apakah seorang konten kreator cocok menjadi brand ambasador di Negara US ?

Data

JSON File (US_kategori_id.json)

Load Datset JSON

```
df_json = pd.read_json('E:/Video_Purwadika/Capstone P 2/drive-download-20230320T123409Z-001/US_category_id.j
df_json.head(5)
```

	kind	etag	items
0	youtube#videoCategoryListResponse	"m2yskBQFythfE4irbTleOgYYfBU/S730Ilt-Fi-emsQJv...	{'kind': 'youtube#videoCategory', 'etag': '"m2...
1	youtube#videoCategoryListResponse	"m2yskBQFythfE4irbTleOgYYfBU/S730Ilt-Fi-emsQJv...	{'kind': 'youtube#videoCategory', 'etag': '"m2...
2	youtube#videoCategoryListResponse	"m2yskBQFythfE4irbTleOgYYfBU/S730Ilt-Fi-emsQJv...	{'kind': 'youtube#videoCategory', 'etag': '"m2...
3	youtube#videoCategoryListResponse	"m2yskBQFythfE4irbTleOgYYfBU/S730Ilt-Fi-emsQJv...	{'kind': 'youtube#videoCategory', 'etag': '"m2...
4	youtube#videoCategoryListResponse	"m2yskBQFythfE4irbTleOgYYfBU/S730Ilt-Fi-emsQJv...	{'kind': 'youtube#videoCategory', 'etag': '"m2...

CSV File (USvideos.csv)

Load CSV dataset

```
df = pd.read_csv("E:/Video_Purwadika/Capstone P 2/drive-download-20230320T123409Z-001/USvideos.csv")
df.head()
```

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	v
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13T17:13:01.000Z	SHANtell martin	74
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13T07:30:00.000Z	last week tonight trump presidency|"last week ...	24
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	31
3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017-11-13T11:00:04.000Z	rhett and link|"gmm"|"good mythical morning"|"...	34
4	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12T18:01:41.000Z	ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...	20

Data Understanding And Cleaning

Data Understanding

- Menggabungkan 2 data set json dan csv

```
df = pd.merge(df, df_json_new, how = 'inner', left_on = 'category_id', right_on = 'id_category')
```

- Data

```
# Dataset ini sudah memiliki category dari konten atau video yang dibuat
df
```

	video_id	trending_date	title	channel_title	category_id	publish_time	tags
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13T17:13:01.000Z	SHANtell martin
1	0mlNzVSJrT0	17.14.11	Me-O Cats Commercial	Nobrand	22	2017-04-21T06:47:32.000Z	cute|"cats"|"thai"|"eggs"
2	STI2fI7sKMo	17.14.11	AFFAIRS, EX BOYFRIENDS, $18MILLION NET WORTH -...	Shawn Johnson East	22	2017-11-11T15:00:03.000Z	shawn johnson|"andrew east"|"shawn east"|"shaw...
3	KODzih-pYIU	17.14.11	BLIND(folded) CAKE DECORATING CONTEST (with Mo...	Grace Helbig	22	2017-11-11T18:08:04.000Z	itsgrace|"funny"|"comedy"|"vlog"|"grace"|"
4	8mhTWqWlQzU	17.14.11	Wearing Online Dollar Store Makeup For A Week	Safiya Nygaard	22	2017-11-11T01:19:33.000Z	wearing online dollar store makeup for a week|...

- Melihat tipe data untuk setiap kolum

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 40949 entries, 0 to 40948
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   video_id                40949 non-null  object
 1   trending_date           40949 non-null  object
 2   title                   40949 non-null  object
 3   channel_title           40949 non-null  object
 4   category_id             40949 non-null  int64
 5   publish_time            40949 non-null  object
 6   tags                    40949 non-null  object
 7   views                   40949 non-null  int64
 8   likes                   40949 non-null  int64
 9   dislikes                40949 non-null  int64
 10  comment_count           40949 non-null  int64
 11  thumbnail_link          40949 non-null  object
 12  comments_disabled       40949 non-null  bool
 13  ratings_disabled        40949 non-null  bool
 14  video_error_or_removed  40949 non-null  bool
 15  description             40379 non-null  object
 16  id_category             40949 non-null  int64
 17  category                40949 non-null  object
dtypes: bool(3), int64(6), object(9)
memory usage: 5.1+ MB
```

- Melakukan pengecekan nilai unik pada tiap kolum

```
df.describe()
```

	category_id	views	likes	dislikes	comment_count	id_category
count	40949.000000	4.094900e+04	4.094900e+04	4.094900e+04	4.094900e+04	40949.000000
mean	19.972429	2.360785e+06	7.426670e+04	3.711401e+03	8.446804e+03	19.972429
std	7.568327	7.394114e+06	2.288853e+05	2.902971e+04	3.743049e+04	7.568327
min	1.000000	5.490000e+02	0.000000e+00	0.000000e+00	0.000000e+00	1.000000
25%	17.000000	2.423290e+05	5.424000e+03	2.020000e+02	6.140000e+02	17.000000
50%	24.000000	6.818610e+05	1.809100e+04	6.310000e+02	1.856000e+03	24.000000
75%	25.000000	1.823157e+06	5.541700e+04	1.938000e+03	5.755000e+03	25.000000
max	43.000000	2.252119e+08	5.613827e+06	1.674420e+06	1.361580e+06	43.000000

Data Cleaning

- Mengecek ada tidaknya NaN pada tiap kolum

```
df_1.isna().sum()
```

```
video_id                  0
trending_date             0
title                     0
channel_title             0
category_id               0
publish_time              0
tags                      0
views                     0
likes                     0
dislikes                  0
comment_count             0
thumbnail_link            0
comments_disabled         0
ratings_disabled          0
video_error_or_removed    0
description             570
id_category               0
category                  0
dtype: int64
```

```
df_1.loc[df_1['description'].isna()]
```

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	di
71	aujUI3yt6nM	17.19.11	Quad9 How To Install with Windows	Quad9 DNS	22	2017-11-16T01:56:43.000Z	DNS|"privacy"|"security"	4759	40	2
84	aujUI3yt6nM	17.20.11	Quad9 How To Install with Windows	Quad9 DNS	22	2017-11-16T01:56:43.000Z	DNS|"privacy"|"security"	5218	42	2
99	aujUI3yt6nM	17.21.11	Quad9 How To Install with Windows	Quad9 DNS	22	2017-11-16T01:56:43.000Z	DNS|"privacy"|"security"	5963	47	2

- Merubah Data Nan Menjadi data isi dengan tulisan kosong

```
# Menggisi Missing Value dengan String
df_1['description'] = df_1['description'].replace(np.nan, 'Kosong', regex=True)
```

```
df_1.isna().sum()
```

```
video_id                0
trending_date           0
title                   0
channel_title           0
category_id             0
publish_time            0
tags                    0
views                   0
likes                   0
dislikes                0
comment_count           0
thumbnail_link          0
comments_disabled       0
ratings_disabled        0
video_error_or_removed  0
description             0
id_category             0
category                0
dtype: int64
```

- Mencari string dalam bentuk None pada seluruh data set

```
substring = 'none'
df_1[df_1.apply(lambda row: row.astype(str).str.contains(substring, case=False).any(), axis=1)]
```

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	co
12	1640fZpYBSY	17.14.11	I love the Price is Right! Wooo! -Kevin	Anaki Abo	22	2017-11-07T18:54:39.000Z	[none]	358597	1211	72	59
15	wRGldR_SQAA	17.14.11	Apple Clips sample	Steve Kovach	22	2017-11-09T18:01:04.000Z	[none]	2259	0	0	0
24	Fyyua5JzD9w	17.15.11	Baby loves Jeopardy!	Daelric	22	2017-11-10T17:57:59.000Z	[none]	48372	382	11	29
33	Jidk0O6uu-0	17.16.11	Granulated Sugar From Honey	Cody'sLab Backup	22	2017-11-15T07:25:03.000Z	[none]	52607	3835	32	35
41	Fyyua5JzD9w	17.16.11	Baby loves Jeopardy!	Daelric	22	2017-11-10T17:57:59.000Z	[none]	49915	386	11	24
...	...	...	...	...	...	...	...	...	...	...	...
40800	sBMOXHrXJtI	18.13.06	Cuphead DLC Announcement Trailer | Xbox	Studio MDHR	20	2018-06-10T21:14:22.000Z	[none]	1340584	65720	660	13

- Merubah column 'comments_disabled', 'ratings_disabled','video_error_or_removed' menjadi nilai bolean

```
df_1.replace({False: 0, True: 1}, inplace=True)
```

```
df_1.head()
```

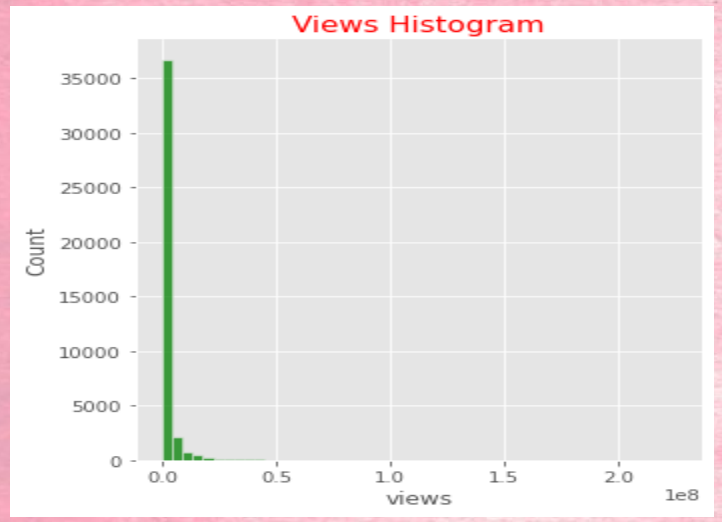
_link	comments_disabled	ratings_disabled	video_error_or_removed	description	id_category	category	tag_1	Desc
.jpg	0	0	0	SHANTELL'S CHANNEL - https://www.youtube.com/s...	22	People & Blogs	1	1
jpg	0	0	0	Kittens come out of the eggs in a Thai commerc...	22	People & Blogs	1	1
)g	0	0	0	Subscribe for weekly videos ▶ http://bit.ly/sj...	22	People & Blogs	1	1
pg	0	0	0	Molly is an god damn amazing human and she cha...	22	People & Blogs	1	1
lt.jpg	0	0	0	I found this online dollar store called ShopMi...	22	People & Blogs	1	1

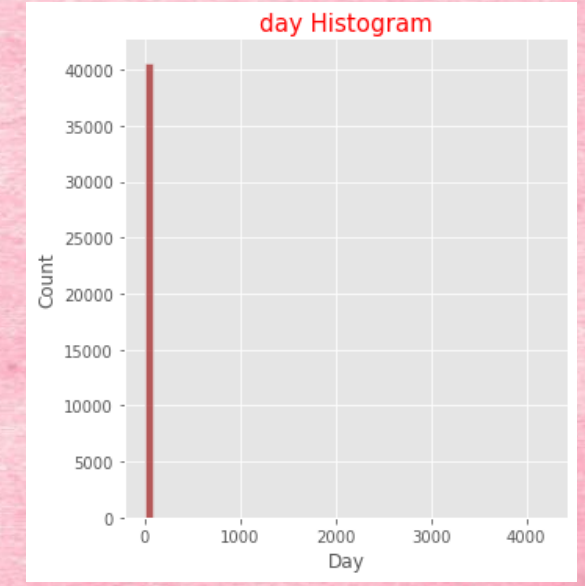
- Ditambahkan 4 kolum tambahan yaitu kategori, tag_1, Description_1, avg_channel_views, dan label_trending

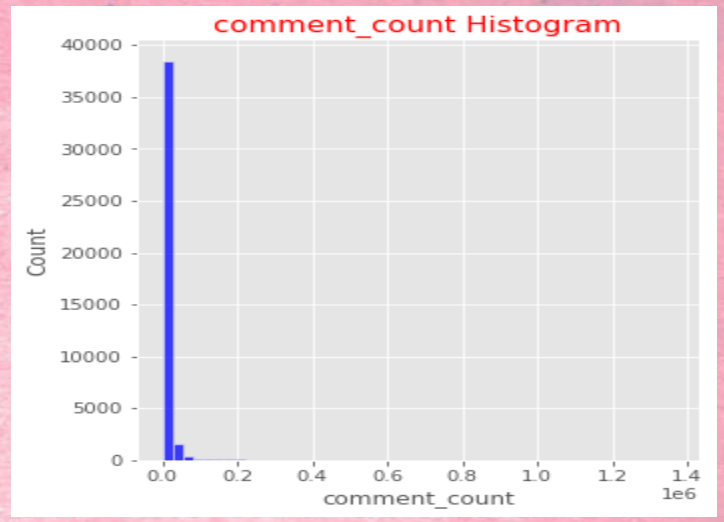
id_category	category	tag_1	Description_1	Day	avg_channel_views	label_trending
22	People & Blogs	1	1	1	2449950	1
22	People & Blogs	1	1	207	116099	1
22	People & Blogs	1	1	3	386685	1
22	People & Blogs	1	1	3	231687	1
22	People & Blogs	1	1	3	3799480	1

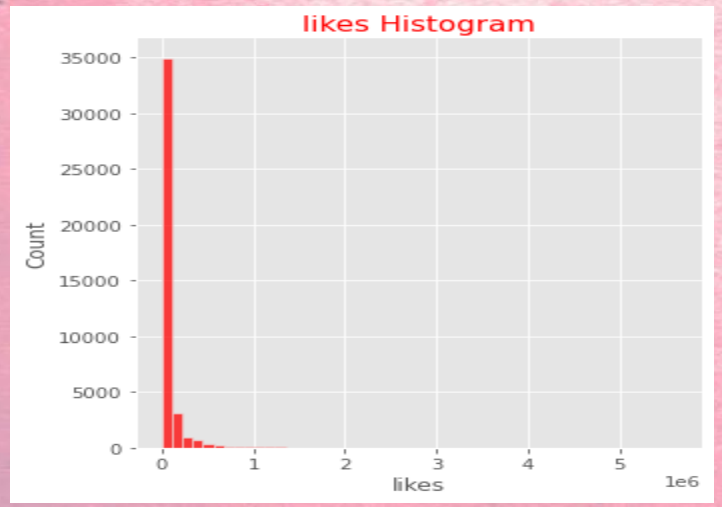
Data Analysis

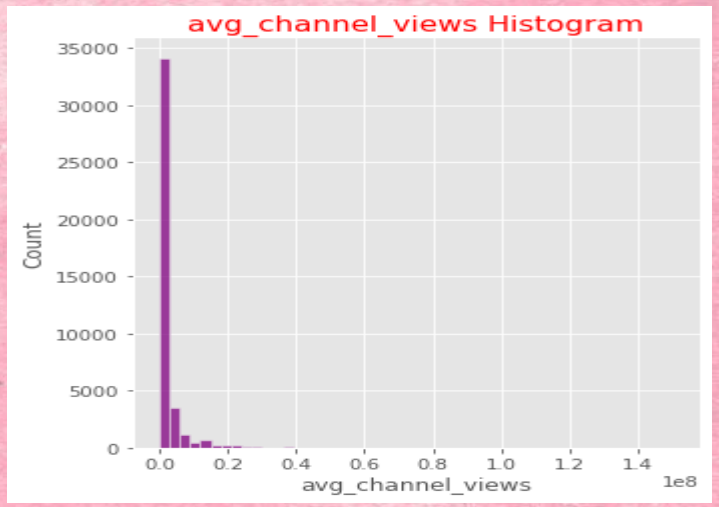
- Penilaian Normalitas
  - Grapical Methods











- Penilaian Normalitas
  - Frequentist Test

```
P-Value: 0.0. Jadi, kami tidak menganggap distribusi normal pada views
```

Views

```
P-Value: 0.0. Jadi, kami tidak menganggap distribusi normal pada likes
```

Likes

```
P-Value: 0.0. Jadi, kami tidak menganggap distribusi normal pada day
```

Day

```
P-Value: 0.0. Jadi, kami tidak menganggap distribusi normal pada comment_count
```

Comment_count

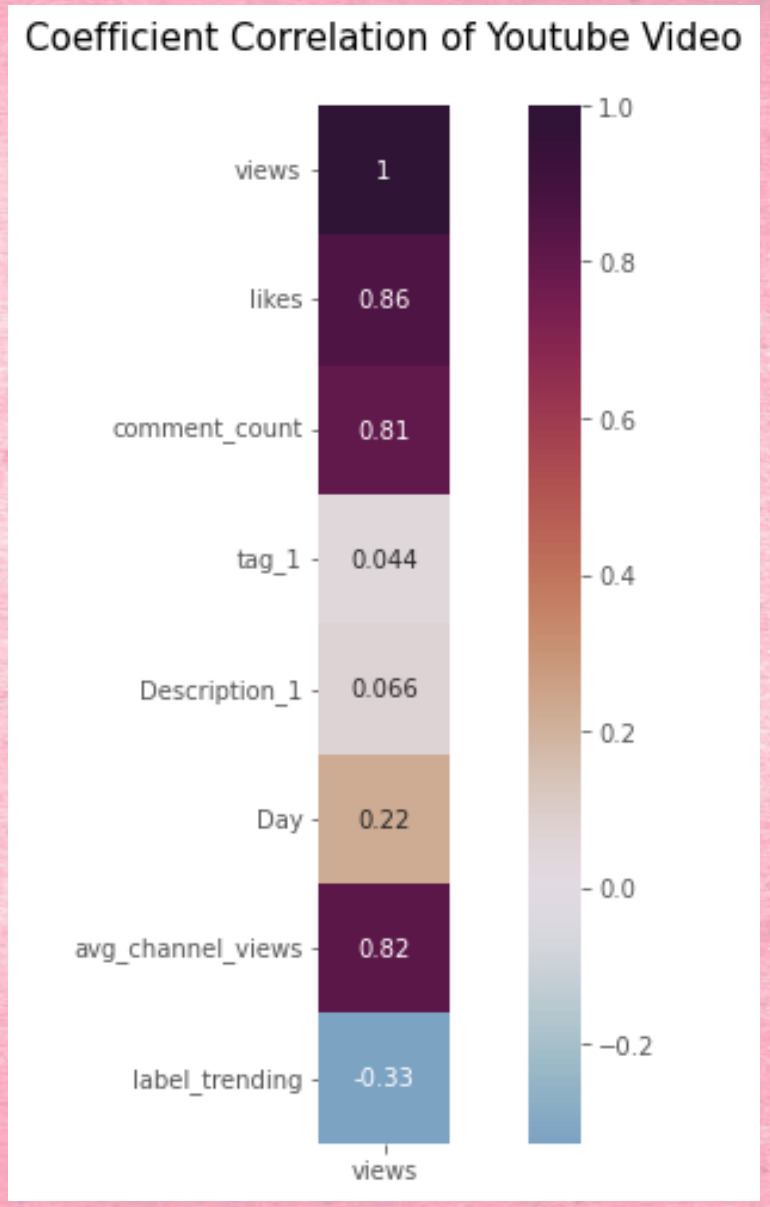
```
P-Value: 0.0. Jadi, kami tidak menganggap distribusi normal pada avg_channel_views
```

Avg_channel_views

Untuk mengetahui karakteristik dari 'views', 'likes', 'dislikes', 'comment_count', 'avg_channel_views' maka kita akan melihat distribusi datanya menggunakan uji Normalitas D'Agostino and Pearson's Test.

Apakah tags , likes, publish_time, and comments memiliki korelasi satu sama lain ?

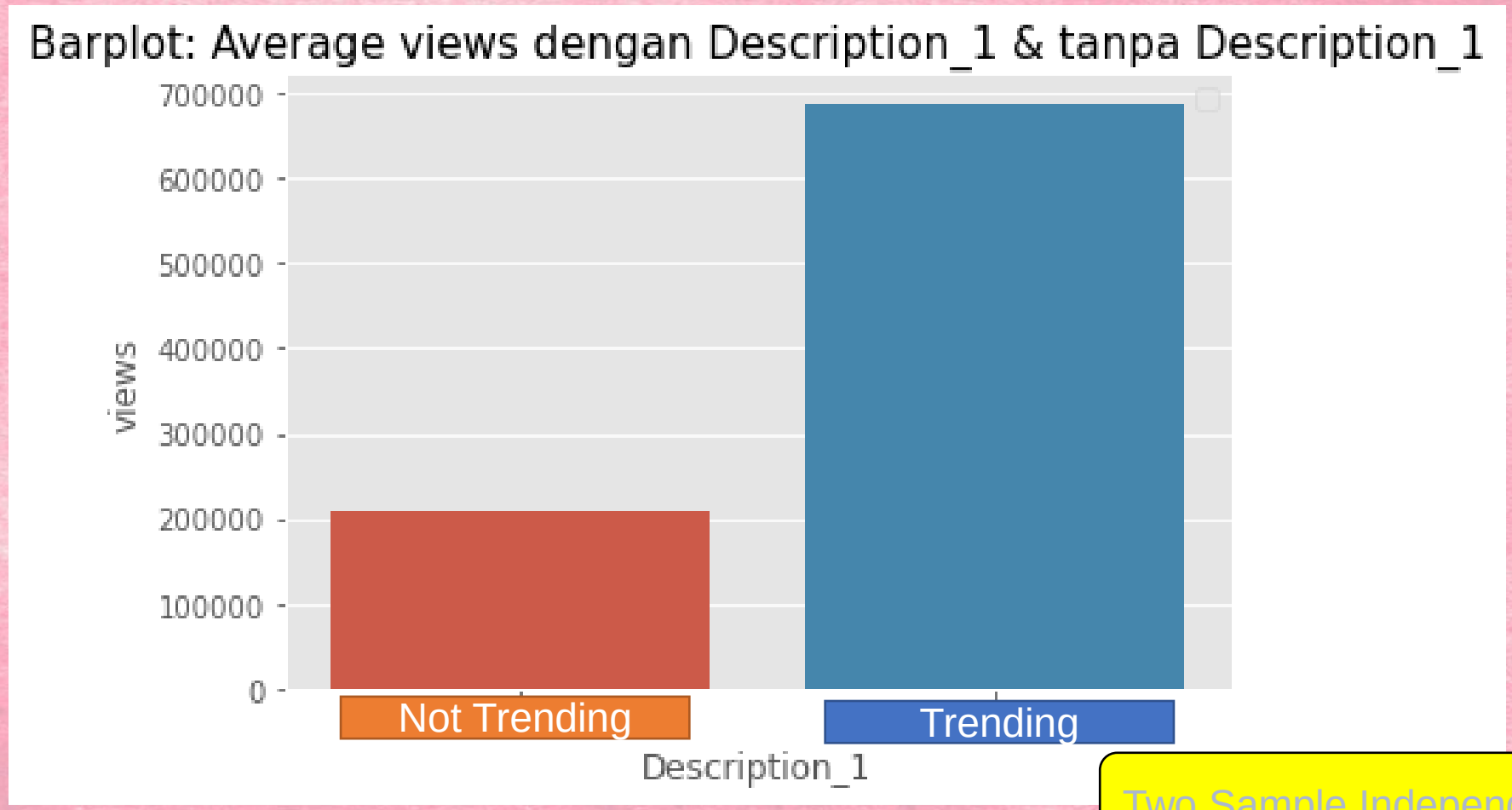
Grafik : Coefficient Correlation of Youtube Video



method spearman

Grafik : Barplot Average views dengan Description_1 & tanpa Description_1

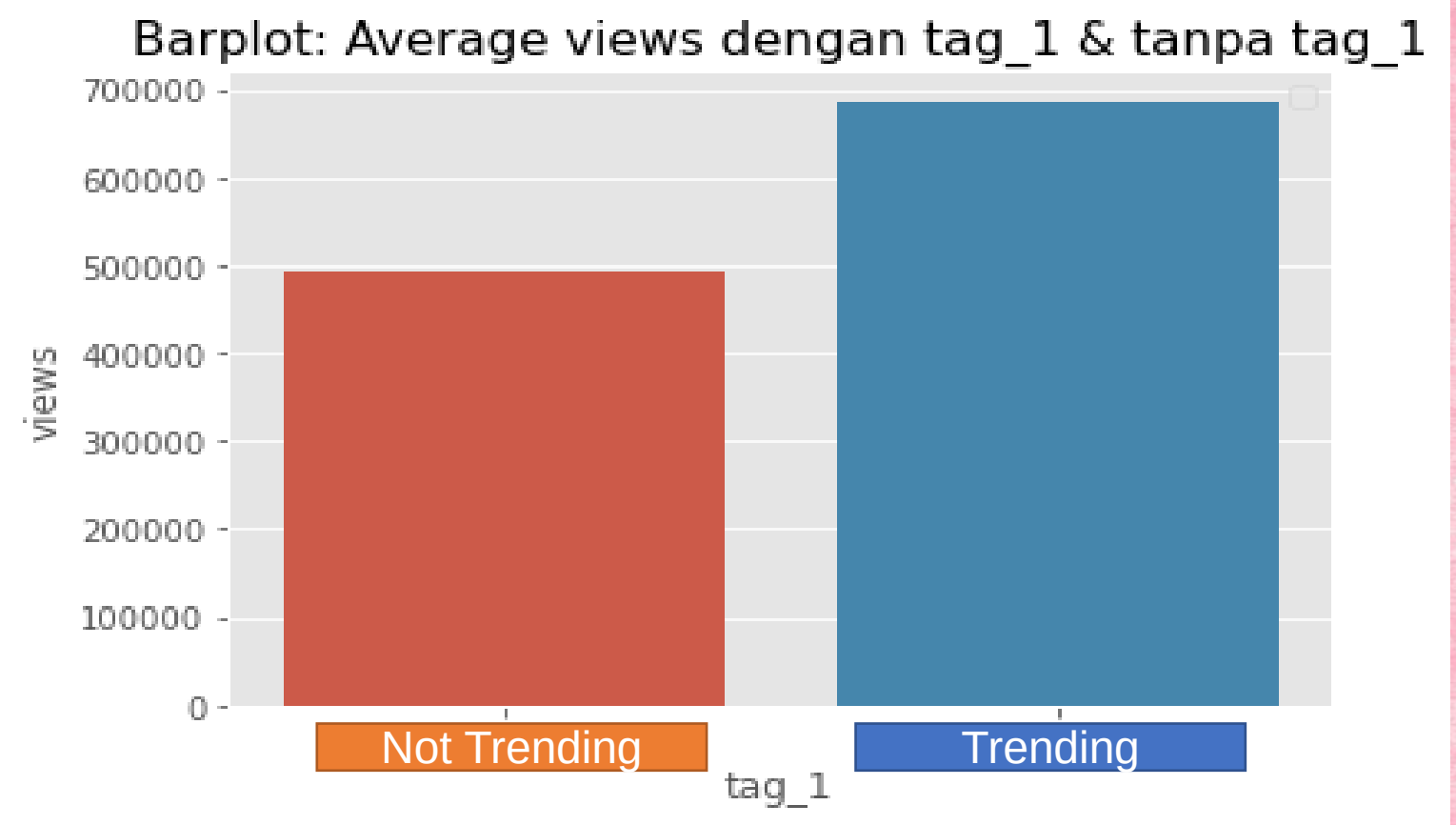
Apakah ada tidaknya describsi mempengaruhi viewer pada video yang di publish di Negara US ?



Two Sample Independent T-Test

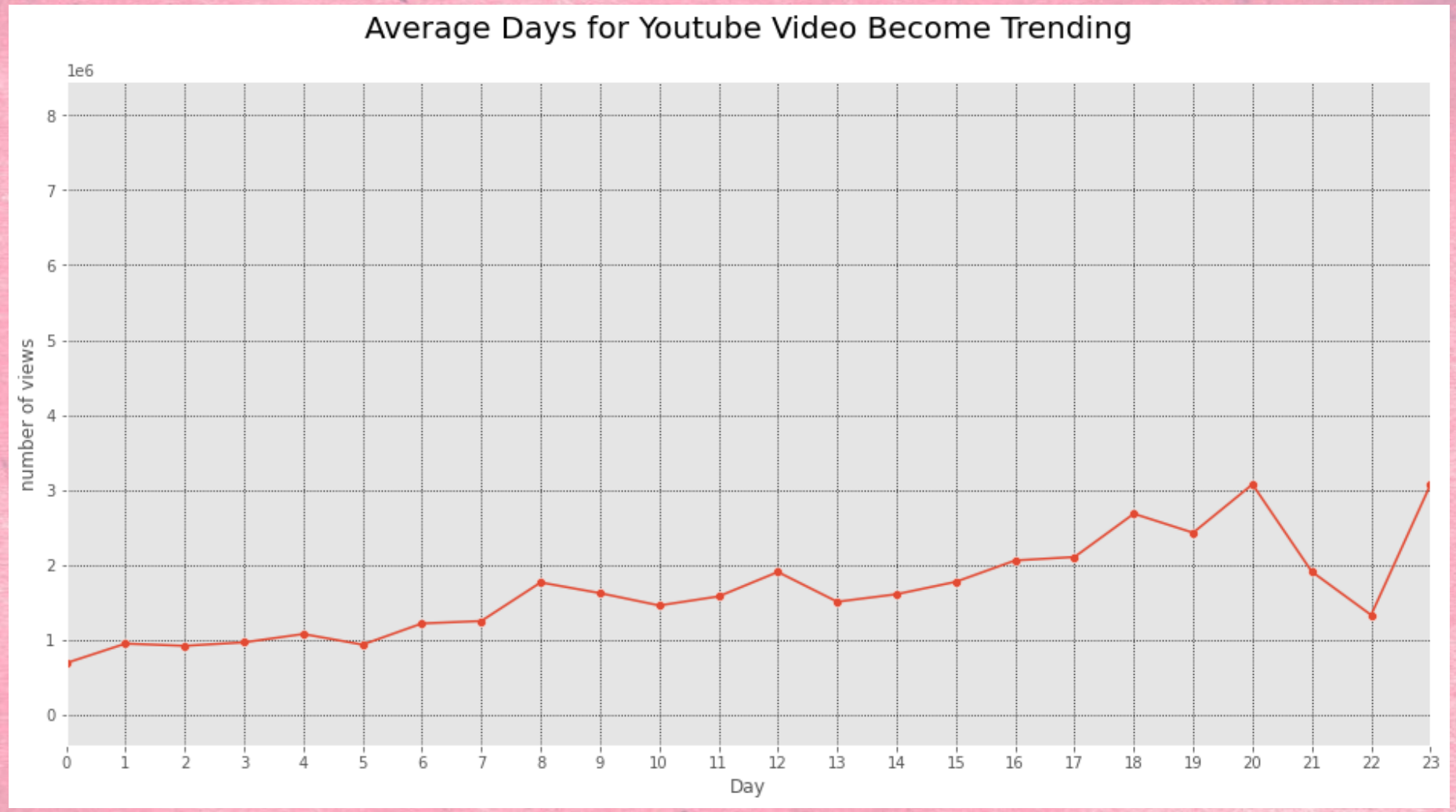
Grafik : Barplot Average Views dengan tag_1 & tanpa tag_1

Apakah ada tidaknya tags mempengaruhi viewer pada video video yang di publish di Negara US ?



Two Sample Independent T-Test

Berapa lama waktu yang diperluhkan agar video menjadi trending di Negara US?

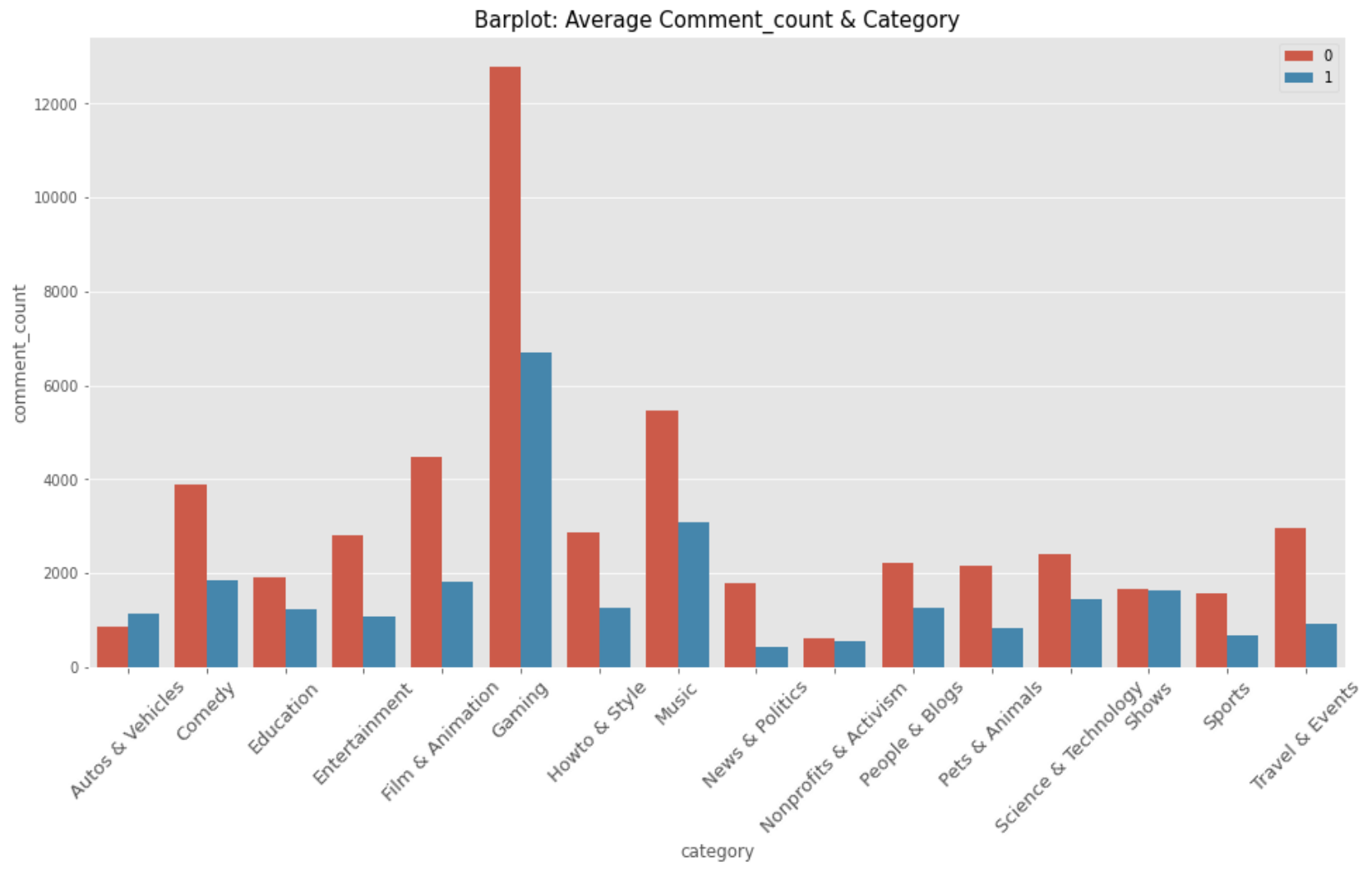


Grafik : Lineplot Average Days for Youtube Video Become Trending

Mann Whitney

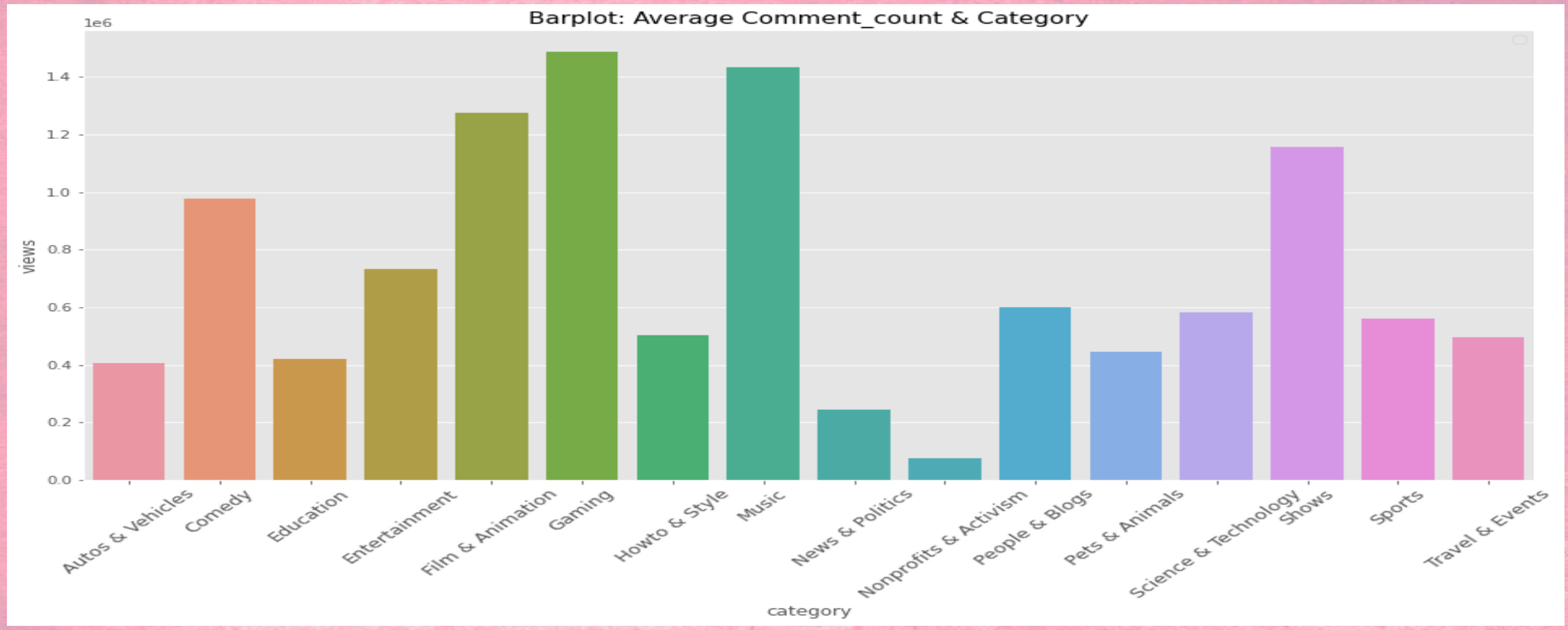
Apakah banyak atau sedikitnya comments mempengaruhi viewer pada video yang di publish di Negara US sehingga menjadi trending ?

Grafik : Barplot Average Comment_count & Category



Mann Whitney

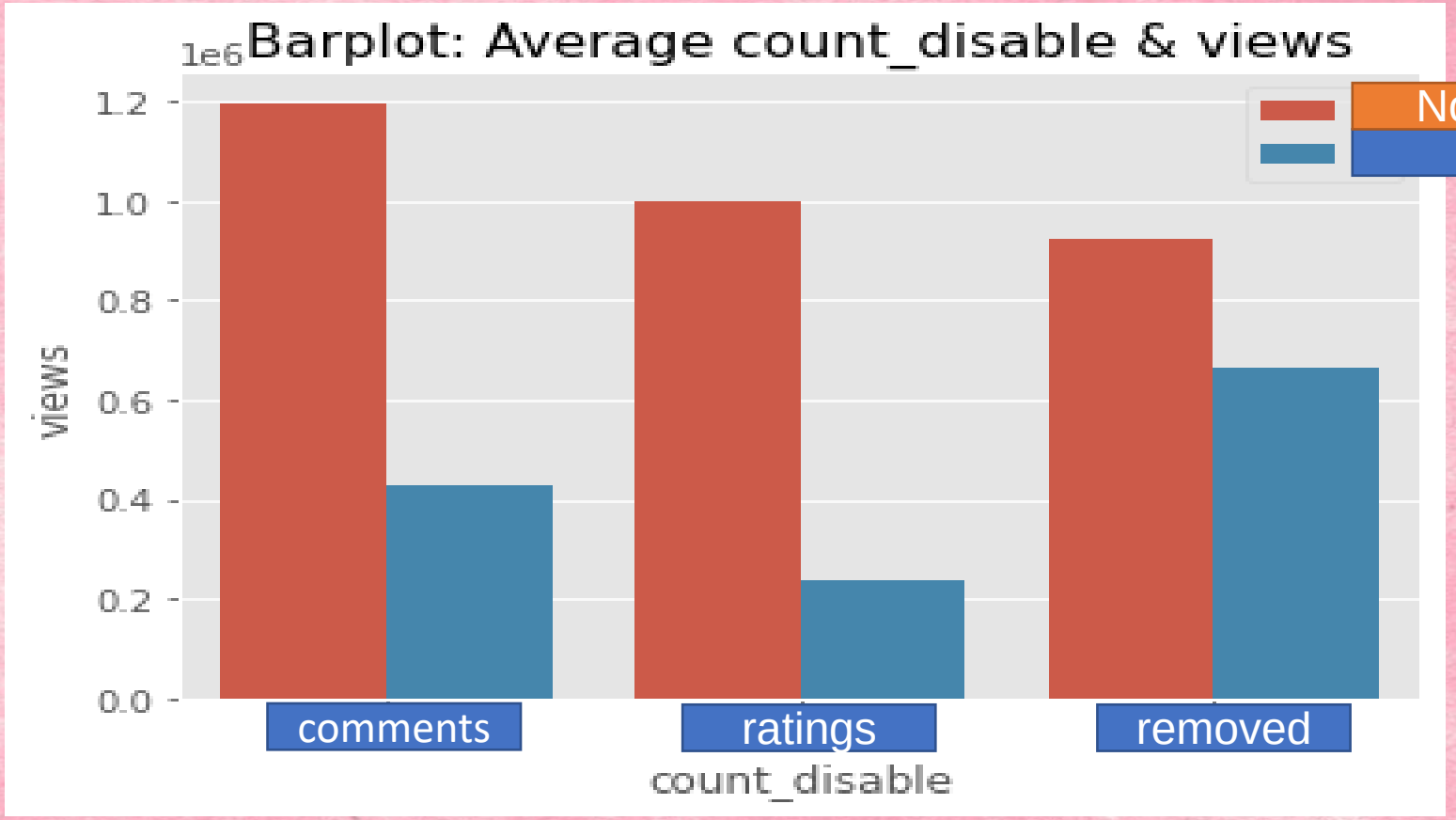
Pada Negara US konten dengan kategori apa yang paling trending dan yang tidak ?



Grafik : Barplot Average Comment_count & Category

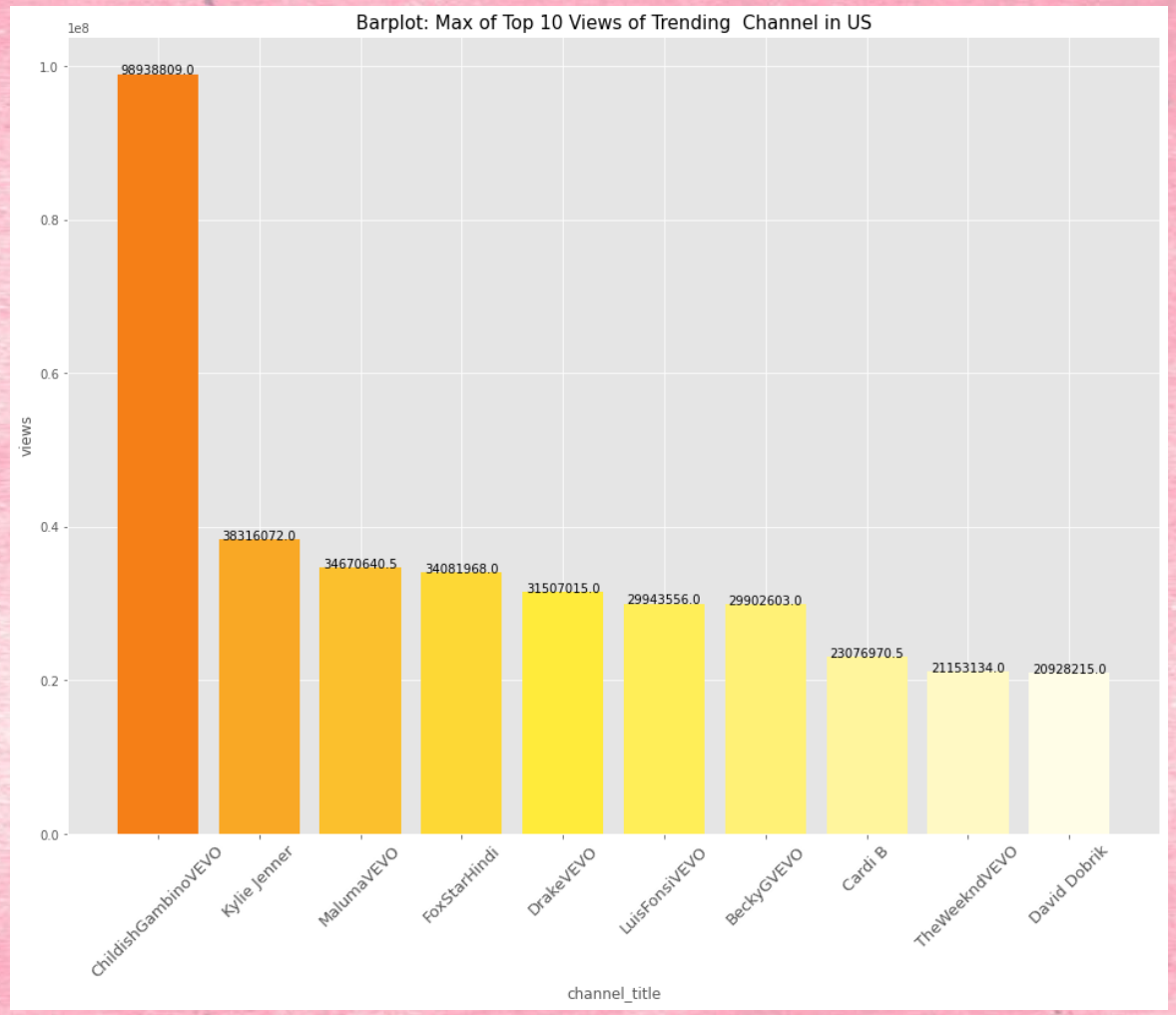
Kruskal Wallis

Apakah tindakan disabled pada comments, ratings, dan video_error_or_removed mempengaruhi rank video yang di publish di Negara US?



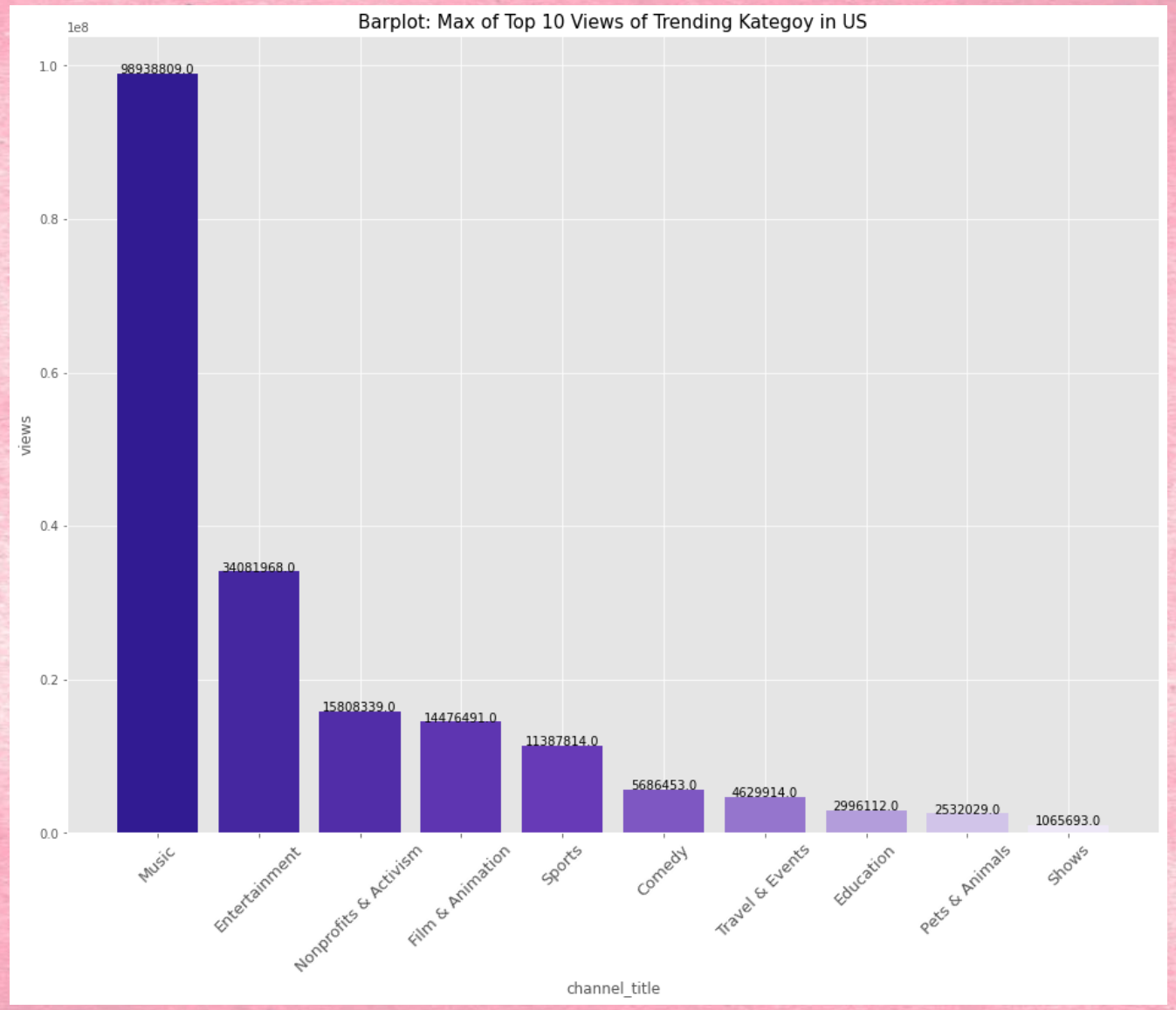
Grafik : Barplot Average count_disable & views

Cara menentukan apakah seorang konten kreator cocok menjadi brand ambasador di Negara US ?

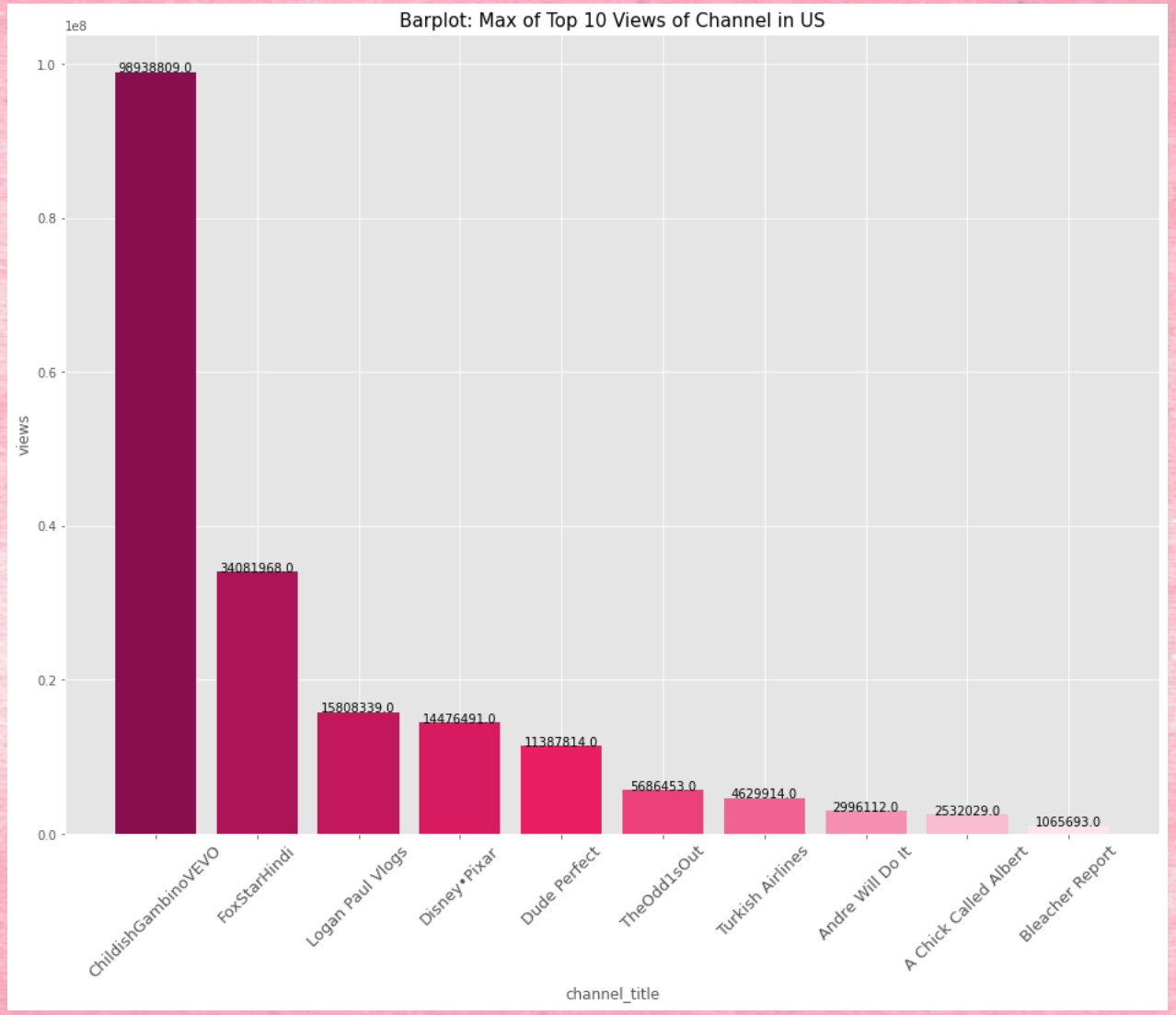


Grafik : Barplot Maksimal 10 Penayangan Teratas dari Saluran Tren AS

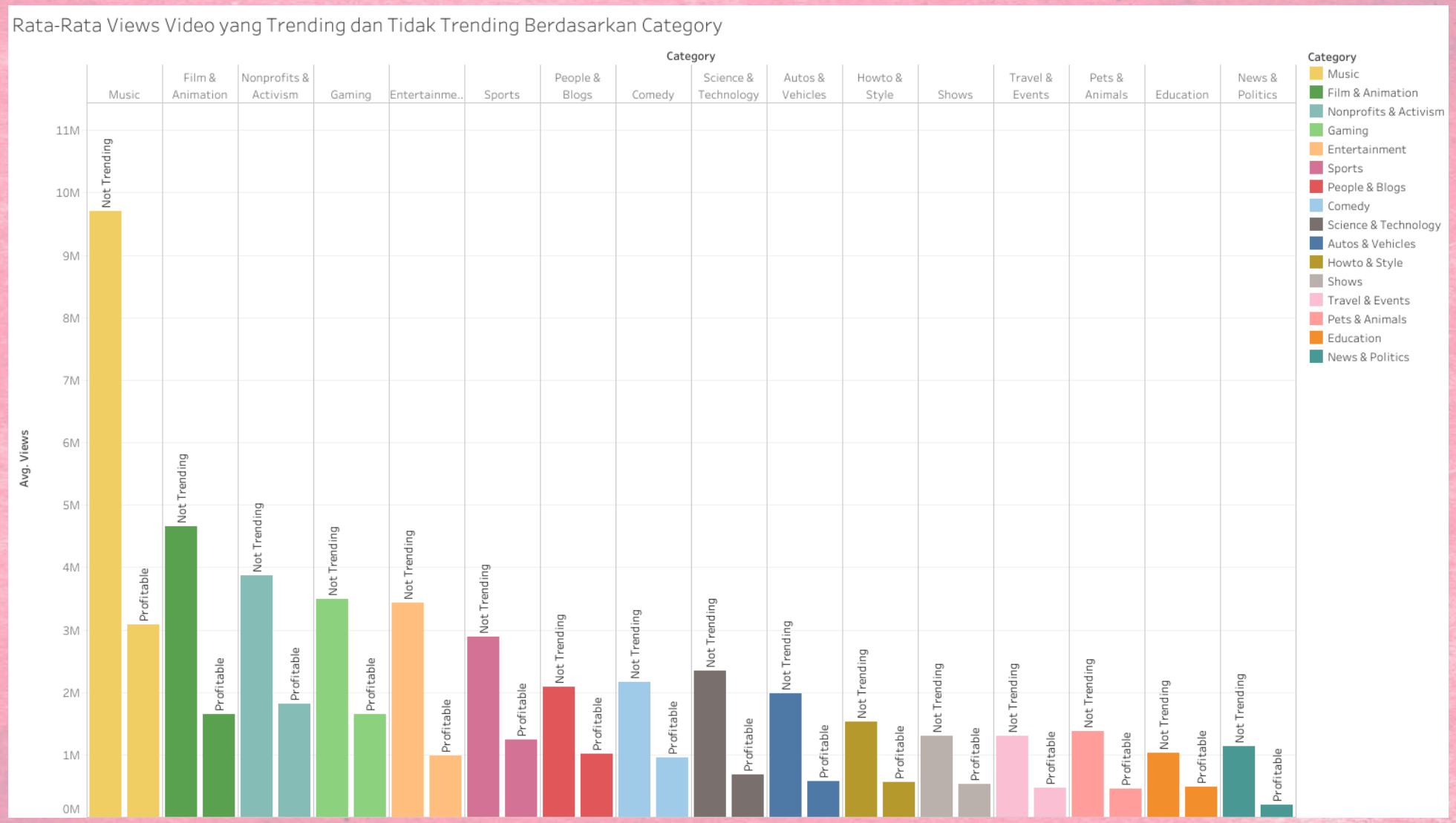
Grafik : Barplot Maksimal 10 Penayangan Teratas dari Kategori Tren US



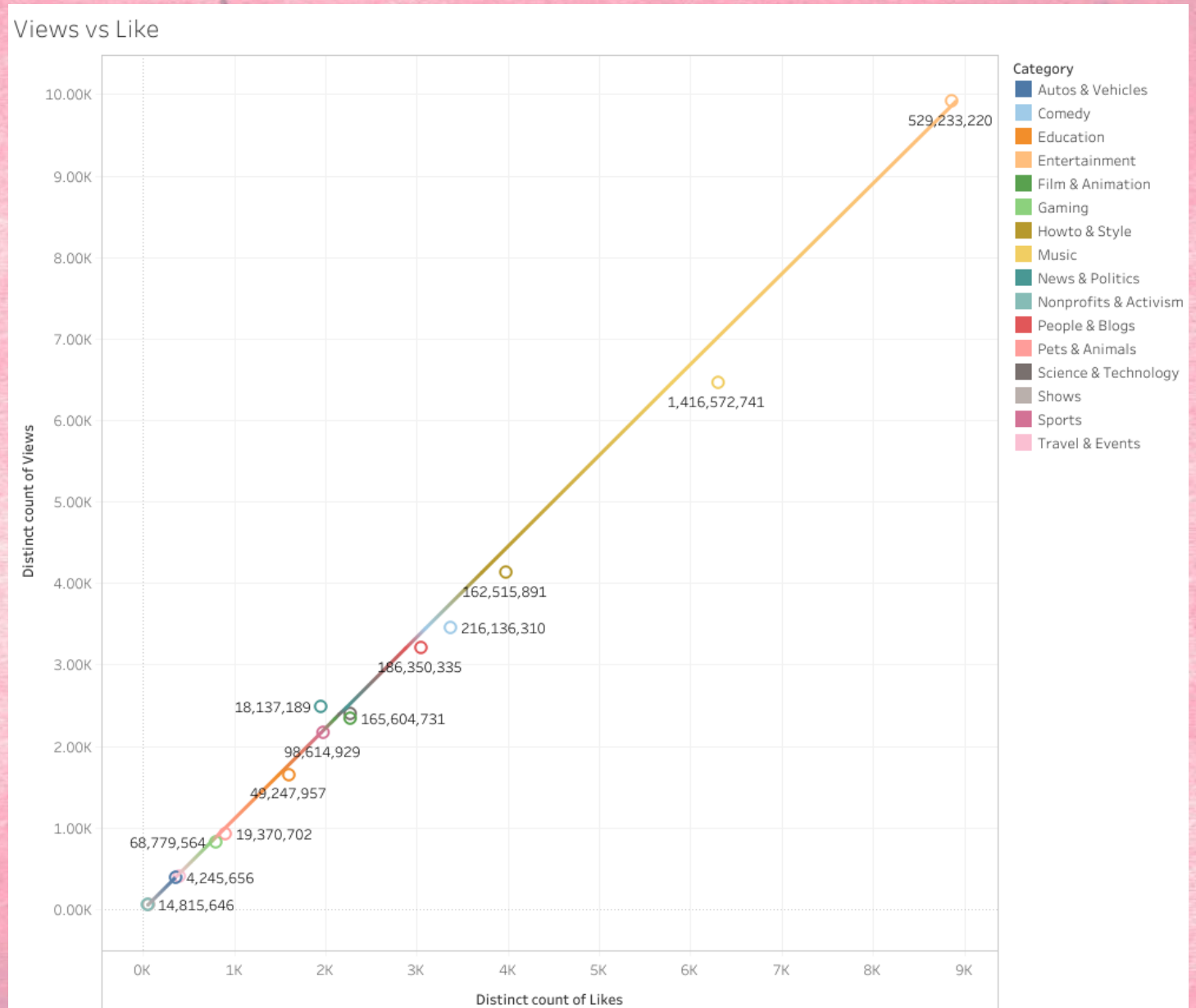
Grafik : Barplot Maksimal 10 Penayangan Saluran Teratas di US



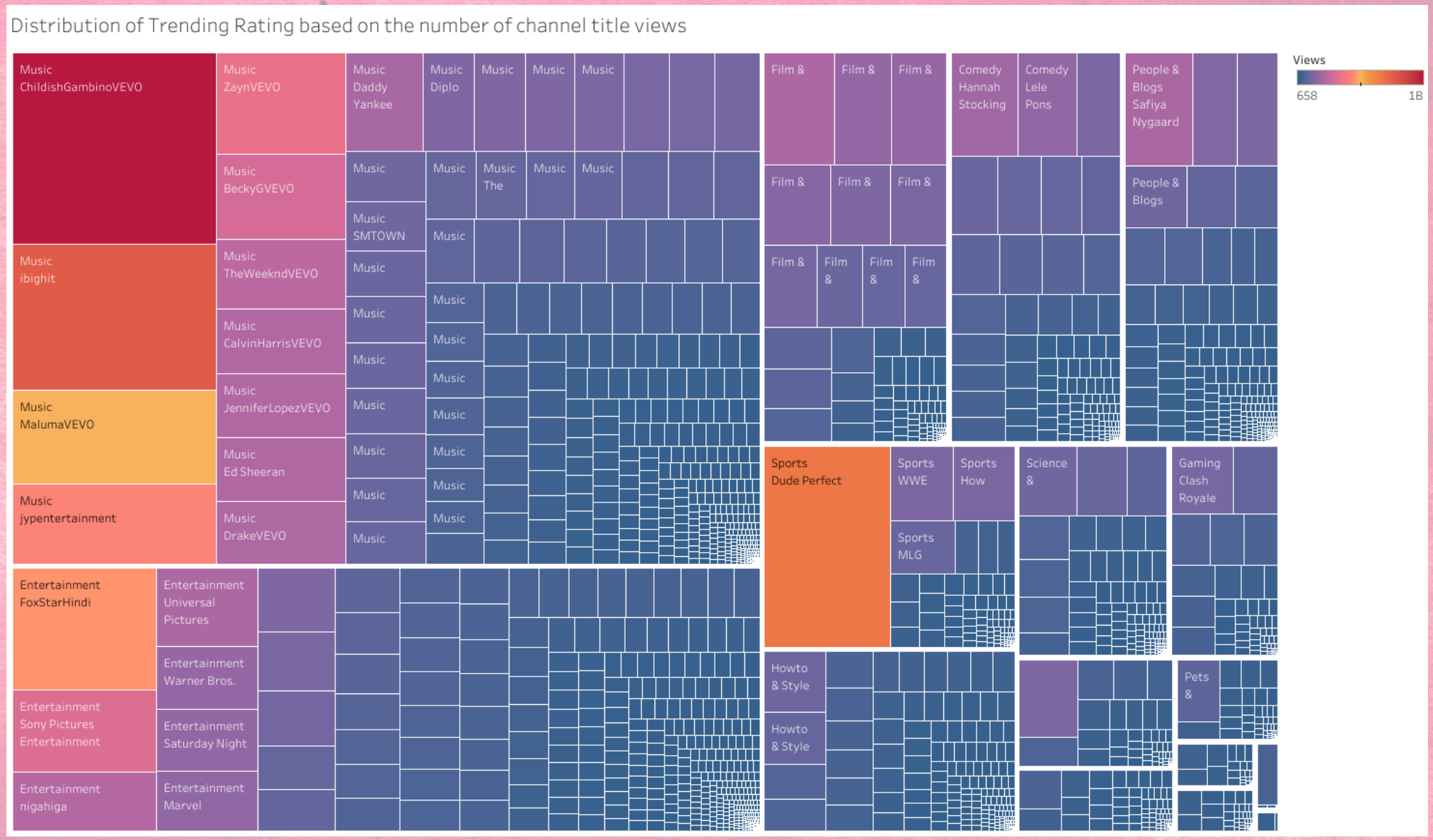
Grafik : Rata-Rata Views Video yang Trending dan Tidak Trending Berdasarkan Kategory



Grafik : Ditonton vs Suka



Grafik Pohon : Pembagian Rating Berdasarkan Jumlah Penayangan Judul Channel



Kesimpulan

Dari analisis yang telah dilakukan, kita bisa membuat kesimpulan tentang kriteria pemilihan brand ambasador pada Negara US:

- Dari semua konten kreator yang trending dan memiliki viewers yang banyak terdapat channel yang menempati posisi 3 besar yaitu : **ChildishGambinoVEVO (969 juta views), FoxStarHindi (340 juta views), Logan Paul Vlogs (158 juta views)**.
- Dari 16 kategory yang trending dan memiliki viewers yang banyak posisi 3 besar yaitu : **Music (969 juta views), Entertainment (340 juta views), Nonprofits & Activism (158 juta views).**

- Kategory konten yang paling trending pada Negara US berdasarkan average Comment_count adalah **Gaming** rata rata comment berkisar 6000 yang menduduki peringkat 1 disusul **Music** rata rata comment berkisar 3000 pada peringkat 2 dan **Film & Animation** rata rata comment berkisar 2000 pada peringkat 3.
- Mayoritas konten kreator yang melakukan disable pada comments, ratings, dan video_error_or_removed akan mendapatkan rata-rata viewers lebih sedikit yaitu berkisar dibawah 1 juta views.
- Kategory konten yang paling tidak trending pada Negara US adalah News & Politics dengan average views sebesar 200 ribuan.

Pola video yang trending dan tidak trending pada video yang di publish, terutama pada negara US, adalah sebagai berikut:¶

- Mayoritas video yang trending memiliki views, comment_count, dan avg_channel_views yang banyak berdasarkan Coefficient Correlation dengan nilai diatas 0,81.
- Video dengan deskripsi memiliki jumlah rata-rata views berkisar 700000 yang lebih tinggi dari pada tanpa deskripsi dengan jumlah rata-rata views berkisar 200000.
- Video dengan tags memiliki jumlah rata-rata views berkisar 700000 yang lebih tinggi dari pada tanpa tags dengan jumlah rata-rata views berkisar 500000 walau terpaut sedikit perbedaan yang tidak terlalu signifikan akan tetapi jumlah rata_rata views berkisar 200000 masih banyak yang dengan tags.

- Biasanya video yang trending pada waktu 5 hari pertama di rentang 1 juta views dimana ini mengalami stucknan, pada hari selanjutnya akan naik secara bertahap di mulai dari hari ke-6.
- Banyak tidaknya koment pada video yang trending kurang menjadi pengaruh karena banyak video yang tidak trending justru memiliki comment_count yang lebih banyak dengan nilai tertinggi video tidak trending memiliki 13000 lebih rata-rata Comment_count.

Rekomendasi

1. Gaming merupakan kategori yang paling banyak mendapatkan jumlah perhitungan koment. Sehingga sektor game dapat memberikan kontribusi kepada sector edukasi. Dimana tidak hanya cara bermain game tapi juga cara membuat game itu seperti apa, sehingga akan ada generasi selanjutnya yang tidak beracu pada satu acuan saja kedepanya.
2. Meminta para youtuber yang menduduki peringat atas seperit kategori music untuk saling membahu membantu meningkatkan performa views pada kategori yang tertinggal seperti Education, dan Shows.
3. Youtube harus memberikan edukasi kepada para konten kreator pada negara US dengan cara memberikan notifikasi pada page content creator di youtube atau melewati email, supaya mengaktifkan kolom koment agar peluang video mereka mendapatkan views lebih banyak.

4. Penghargaan seperti Youtube Award harus dilakukan untuk memberikan penghargaan pada konten kreator di Negara US baik video yang trending atau tidak, agar menambah semangat konten kreato lainya supaya terus berkreasi.

5. Pemilihan Brand Ambasador dapat dipilih berdasarkan 10 Channel Youtube yang trending sebagai berikut :

1. ChildishGambinoVEVO
2. FoxStarHindi
3. Logan Paul Vlogs
4. Disney•Pixar
5. Dude Perfect
6. TheOdd1sOut
7. Turkish Airlines
8. Andre Will Do It
9. A Chick Called Albert
10. Bleacher Report

6. Gaming adalah kategory yang paling viral di US maka akan lebih baik jika pihak youtube mengundang para publish dan penyedia game untuk melakukan iklan di Youtube mengingat Youtube menjadi platfrom yang populer dikalangan gamers.
7. Edukasi diperluhkan untuk semua konten kreator di negara US baik yang pemula dan yang lama. Edukasi yang dimaksud adalah himbauan agar mereka memaksimalkan fitur seperti tags, dan deskripsi pada video yang akan mereka publish.
8. Diperluhkan kajian kenapa pada negara US topik kategory News & Politics kurang diminati oleh user youtube di negara tersebut.

6. Melakukan kegiatan seperti sharing antara konten kreator yang mendatangkan konten kreator dari channel yang sering videonya viral kepada konten kreator lain di negara US.
7. Penyusaian Proporsi iklan pada 16 Kategory guna mengoptimalkan penguna jasa iklan agar mendapatkan benefit yang lebih baik.
8. Diberlakukan pengawasan ketat agar kategory satu dengan yang lain yang tidak berhubungan seperti "iklan tentang game tidak muncul pada kategory edukasi"

`

- Dengan optimalisasi yang tepat pada beberapa hal yang telah kita rekomendasikan diatas seperti pemilihan Brand Ambasador dan pengoptimalan para penguna jasa iklan, diharapkan analisis ini bisa membantu youtube menjadi platfrom nomer 1 sharing video di negara US.

TERIMA KASIH