Implementasi *Exploratory Data Analysis* Pada *Dataset* Video Trending Harian YouTube

Abi Vegari[1], Setia Budi[2]

Program Studi S1 Sistem Informasi, Universitas Kristen Maranatha
Jl. Surya Sumantri No. 65 Bandung
[1]abivegari@gmail.com
[2]setia.budi@it.maranatha.edu

Abstract — **YouTube is a video sharing website that allows its users to interact through videos created by video creators (YouTubers). Videos on YouTube can go to the 'Trending' tab that shows videos that are considered trending by YouTube. The YouTube Help website says that they use many parameters to determine trends. However, YouTube does not specify exact parameters and numbers. Therefore, data analysis was performed on video datasets in three countries namely Canada, the United Kingdom and the United States using the Exploratory Data Analysis method. Data processing was carried out with Pandas and data was visualized with the Matplotlib, Seaborn, Bokeh, and WordCloud libraries. Work starts from normalizing categorical data, changing the shape of the data into the desired form, visualizing the data, and taking meaning from the information generated from exploration and visualization results. The results of exploration and visualization of data in the form of boxplots, bar charts, line plots, and word clouds show patterns in the categories and tags contained in videos that discuss trends in the three countries.**

Keywords— ***data analysis, exploratory data analysis, python, trend, youtube*.**

I. PENDAHULUAN

Video-sharing website memungkinkan penggunanya menonton, berkomentar, menyukai atau tidak menyukai, hingga membagikan video yang dibuat oleh video creator. YouTube adalah salah satu video-sharing website gratis dengan lebih dari 1,9 Miliar pengguna terdaftar yang berkunjung setiap bulan dan layanan yang tersedia di lebih dari 91 negara di seluruh dunia [1]. YouTube memiliki beragam konten orisinal seperti seri vlog, film, konten pendidikan, video musik, hingga bantuan untuk melakukan pekerjaan rumah

Video di YouTube dapat masuk ke tab yang dinamakan ‘Trending’ yang dimaksudkan untuk membantu pemirsa melihat apa yang sedang terjadi di YouTube maupun di dunia. Beberapa video yang akan masuk Trending dapat diprediksi, seperti trailer film baru atau lagu baru dari artis populer. Tetapi, di antaranya tidak dapat diduga, seperti video amatir yang menjadi viral [2].

Website Bantuan YouTube mengatakan bahwa mereka menggunakan banyak parameter untuk menentukan tren yang walaupun tidak terbatas pada parameter itu sendiri (banyaknya penonton, seberapa cepat video menghasilkan penonton, dari mana asal penontonnya, usia video). Umumnya, video yang masuk ke dalam tab Trending akan mengalami kenaikan jumlah penonto. Kenaikan jumlah penonton dapat membantu pembuat konten menghasilkan uang. Pembuat konten dapat mendapatkan uang dari YouTube maupun dari pengiklan yang bekerja sama dengan YouTube. Tetapi, cara YouTube untuk menentukan video mana yang dapat masuk ke dalam tab Trending tidak disebutkan dengan angka-angka atau parameter yang pasti [2].

Penelitian *Exploratory Data Analysis* (EDA) ini dilakukan pada *dataset* dari seorang pengguna situs Kaggle bernama Mitchell J. dengan judul “Trending YouTube Video Statistics” yang berisi data video yang mengalami tren harian atau video yang masuk ke dalam tab trending di YouTube. Data tren harian tersebut diambil dari beberapa negara di dunia yang diambil sejak tanggal 14 November 2017 hingga 14 Juni 2018. Tetapi, penelitian dilakukan pada Negara Canada, Great Britain, dan United States dengan pertimbangan persamaan bahasa yang digunakan. Penelitian dilakukan dengan harapan pola-pola video yang mengalami tren di YouTube dapat sedikit diketahui dan dipahami sehingga dapat dijadikan referensi ataupun sekadar pengetahuan bagi orang yang membutuhkan.

II. KAJIAN TEORI

A. Exploratory Data Analysis

Exploratory Data Analysis (EDA) diperkenalkan oleh ahli statistik bernama John W. Tukey pada 1977. EDA didefinisikan sebagai pekerjaan detektif karena peneliti melakukan eksplorasi terhadap suatu data tanpa memiliki ide atau prasangka terhadap informasi apa yang akan didapatkan dari data yang dianalisis [3]. EDA adalah pendekatan analisis data yang menggunakan berbagai teknik terutama grafis untuk mengekstrak variabel penting dari *dataset* untuk memaksimalkan wawasan di *dataset* tersebut [4]. EDA diterapkan pada data yang dikumpulkan tanpa hipotesis yang jelas dengan tujuan menemukan suatu petunjuk yang dapat menginspirasi ide dan hipotesis lainnya [5].

Langkah utama dalam EDA adalah :

1. Menampilkan data;
2. Mengidentifikasi fitur data yang terlihat mencolok;
3. Menafsirkan fitur data yang terlihat mencolok [5]

B. Pandas

Pandas adalah pustaka Python yang berisi struktur data dan alat manipulasi data yang dirancang untuk membuat pembersihan dan analisis data dapat dilakukan dengan cepat dan mudah dengan Python [6]. Pandas, menyediakan seperangkat alat tingkat tinggi, fleksibel, dan cepat untuk memungkinkan Anda memanipulasi data ke dalam bentuk yang tepat.

C. Dataframe

Dataframe adalah Struktur data tabel dua dimensi yang dapat diubah dengan sumbu berlabel (baris dan kolom) [7]. *Dataframe* mewakili tabel data persegi panjang dan berisi kumpulan kolom yang masing-masing dapat berupa tipe nilai yang berbeda (numerik, string, boolean, dll.). Kolom pada dataframe adalah objek dari series, sedangkan baris pada *dataframe* adalah elemen dari series [8].

D. Visualisasi Data

Data dapat ditampilkan dalam bentuk visualisasi data. Visualisasi yang informatif adalah salah satu yang terpenting dalam analisis data. Contohnya, visualisasi data dapat membantu mengidentifikasi outlier atau sebagai cara untuk menghasilkan ide untuk mentransformasi data agar muncul suatu fitur data yang mencolok [9]. Visualisasi data melibatkan penyajian data dalam bentuk grafik atau gambar yang membuat informasi mudah dimengerti untuk membantu menjelaskan fakta dan menentukan tindakan apa yang harus dilakukan selanjutnya. [10].

E. Matplotlib

Matplotlib adalah paket untuk melakukan plotting yang dirancang untuk membuat plot yang dapat dipublikasi. John Hunter memulai proyek ini pada 2002. Matplotlib mendukung berbagai sistem pengerjaan dan hasil visualisasinya dapat diekspor ke semua format grafik vektor seperti PDF, SVG, JPG, PNG, BMP, GIF, dll. [9].

F. Seaborn

Seaborn adalah pustaka visualisasi data Python yang dibuat berdasarkan matplotlib. Seaborn menyediakan antarmuka tingkat tinggi untuk menggambar grafik statistik yang menarik dan informatif. Seaborn menyederhanakan menciptakan banyak jenis visualisasi yang umum [11]. Seaborn dapat memodifikasi skema warna dan gaya plotting default dari matplotlib untuk meningkatkan keterbacaan dan estetika

G. Bokeh

Bokeh adalah pustaka visualisasi interaktif yang dapat dijalankan di browser web. Bokeh memberikan konstruksi grafis serbaguna yang elegan dan ringkas, dan memberikan interaktivitas kinerja tinggi terhadap kumpulan data yang besar. Bokeh dapat membantu siapa saja yang ingin dengan cepat dan mudah membuat plot interaktif, dashboard, dan lain-lain [12].

H. WordCloud

WordCloud adalah pustaka visualisasi data yang digunakan untuk mewakili data dalam bentuk teks di mana ukuran setiap kata menunjukkan frekuensi atau tingkat kepentingannya [13].

I. YouTube

YouTube adalah platform berbagi video atau video-sharing website yang populer di mana pemirsa dapat menonton video yang dibuat oleh pembuat konten yang disebut YouTuber. Diluncurkan pada tahun 2005, YouTube adalah platform berbagi video yang populer dengan lebih dari satu miliar pengguna dan satu miliar jam video ditonton setiap hari [14].

III. METODOLOGI PENELITIAN

A. Platform Penelitian

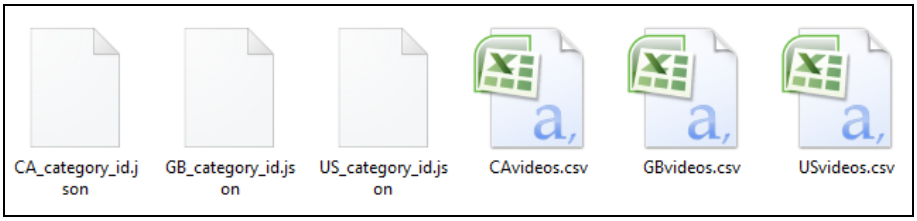
Platform yang digunakan dalam penelitian ini adalah Jupyter Notebook dengan bantuan library Pandas untuk mengolah dataset penelitian. Jupyter Notebook digunakan karena Jupyter Notebook tepat digunakan sebagai platform analisis data dengan bahasa Python seperti pada penelitian ini. Jupyter Notebook dapat menjalankan baris kode satu per satu sehingga jika terjadi suatu kesalahan hal yang perlu diperbaiki adalah kesalahan kode pada suatu baris yang dianggap bermasalah. Hasil analisis data akan divisualisasikan dengan menggunakan bantuan pustaka matplotlib, seaborn, bokeh, dan wordcloud.

B. Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari Kaggle.com dengan judul "Trending YouTube Video Statistics ". Dataset ini bersifat publik dan dapat dipergunakan dengan bebas untuk keperluan penelitian dan pengembangan pengetahuan tanpa dibatasi oleh lisensi. Dataset ini berisi rekaman data video-video YouTube yang masuk ke dalam tab Trending harian (*daily trending videos*) dari beberapa negara (United States, Great Britain, Germany, Canada, dan France). Tiap hari terdapat sampai dengan 200 video yang masuk ke dalam tab trending (untuk tiap negaranya). Penelitian ini berfokus pada data trending harian di tiga negara, yaitu United States (Amerika Serikat), Great Britain (Britania Raya), dan Canada (Kanada). Batasan ini diambil dengan pertimbangan ketiga negara ini menggunakan bahasa yang sama (Bahasa Inggris). Dataset ini berisi rekaman trending harian sejak tanggal 14 November 2017 hingga 14 Juni 2018 (7 bulan atau 205 hari).

C. Karakteristik Dataset

Data dari setiap negara tersedia dalam dua jenis file terpisah. Jenis data pertama adalah tiga data berformat CSV (Comma Separated Values) yang berisikan data tentang video yang masuk ke dalam tab trending per harinya (Data Video). Data kedua berisi nama-nama kategori video dari setiap negara dan disimpan dalam format JSON (Javascript Object Notation).



Gambar 1. File Dataset

Ketiga Data video berformat CSV memiliki susunan data yang sama sebagai berikut :

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

Gambar 2. Susunan Data Video

Ketiga Data kategori berformat JSON memiliki susunan data bersarang yang sama per kategorinya sebagai berikut :

```
{'kind': 'youtube#videoCategoryListResponse',
 'etag': '"ld9biNPKjAjgjV7EZ4EKeEGrhao/1v2mrzYSYG6onNLt2qTj13hkQZk"',
 'items': [{'kind': 'youtube#videoCategory',
   'etag': '"ld9biNPKjAjgjV7EZ4EKeEGrhao/Xy1mB4_yLrHy_BmKmPBggty2mZQ"',
   'id': '1',
   'snippet': {'channelId': 'UCBR8-60-B28hp2BmDPdntcQ',
    'title': 'Film & Animation',
    'assignable': True}},
```


Gambar 3. Susunan Data Kategori

Data video CSV diimport ke Jupyter Notebook dengan menggunakan Pandas agar dapat dibentuk ke dalam bentuk *dataframe*. Data kategori JSON akan dikenakan proses normalisasi dan dimasukkan ke dalam *dataframe* baru agar dapat mudah dilihat.

D. Eksplorasi Data

Berikut ini adalah kegitan eksplorasi yang dilakukan dalam penelitian ini. Hasil eksplorasi akan dijelaskan pada bab IV di dalam penelitian ini.

1) Menormalisasi Data Kategori : Ketiga data kategori JSON yang berisi objek bersarang diolah terlebih dahulu. Berdasarkan Gambar 3, yang dibutuhkan dalam penelitian adalah objek yang terdapat di dalam 'items'. Pustaka json pada Jupyter Notebook memiliki fungsi json_normalize yang dapat digunakan untuk menormalisasikan atau mengambil objek yang terdapat pada objek yang bersarang kemudian dijadikan dataframe. Setelah itu hasil dari normalisasi ditampung menjadi dataframe kategori sementara. Setelah ditelaah ternyata kolom yang dibutuhkan hanyalah kolom 'id' yaitu nomor unik/ID untuk kategori dan 'snippet.title' yaitu nama kategori video. Dataset video CSV dan dataset kategori JSON saling berhubungan. Dataset video memuat ID kategori dari tiap video dan dataset kategori memiliki ID kategori beserta nama kategorinya. Tetapi, ID kategori yang ada pada dataset kategori tidak semuanya ada pada ID kategori di dataset video. Maka, data kategori pada dataframe kategori baru yang telah dibuat tadi perlu disesuaikan dengan ID kategori dari video-video yang dimuat dalam dataset video agar dataframe kategori lebih akurat (sesuai dengan dataset video).

2) Identifikasi Macam-Macam Top-5 : Eksplorasi identifikasi macam-macam Lima Teratas / *Top-5* yang dilakukan di tiap negara adalah sebagai berikut :

TABEL I
IDENTIFIKASI MACAM-MACAM TOP-5

No.	Objek Eksplorasi
1	Top-5 Kategori dengan Jumlah Video Tertinggi & Terendah Yang Masuk Trending
2	Top-5 Channel dengan Jumlah Penonton Terbanyak
3	Top-5 Channel yang Paling Sering Masuk Trending
4	Top-5 Video yang Paling Sering Masuk Trending

Untuk menemukan Kategori dengan jumlah video tertinggi & terendah dilakukan dengan mengelompokkan ID kategori (category_id) di *dataframe* video dengan fungsi .group_by() kemudian dihitung berdasarkan frekuensi kemunculannya dengan fungsi .count(). Kemudian disortir ulang berdasarkan angka tertinggi dan terendah dengan sort_values(). Untuk menemukan channel dengan jumlah penonton terbanyak dilakukan pengelompokkan berdasarkan nama channel (channel_title) di *dataframe* video kemudian kolom jumlah penonton (views) dijumlahkan dengan fungsi .sum() dan disortir berdasarkan jumlah tertinggi. Untuk Channel yang paling sering masuk trending ditemukan dengan cara mengelompokkan nama channel (channel_title), lalu dihitung frekuensinya dengan value_count() dan disortir berdasarkan yang tertinggi. Untuk menemukan video yang paling sering masuk trending ditemukan dengan cara mengelompokkan kolom ID video (video_id) dan nama video (title) kemudian menghitungnya dan diurutkan berdasarkan jumlah tertinggi.

3) Identifikasi Bentuk Distribusi Data Tiap Kategori : Berikut adalah eksplorasi yang dilakukan untuk mengidentifikasi bentuk distribusi data yang dilakukan pada *dataframe* video dari tiap negara:

TABEL III
IDENTIFIKASI BENTUK DISTRIBUSI DATA TIAP KATEGORI

No.	Objek Eksplorasi	Nama Kolom Eksplorasi
1	Identifikasi Bentuk Distribusi Data Penonton di Tiap Kategori	*views*
2	Identifikasi Bentuk Distribusi Data Likes di Tiap Kategori	*likes*
3	Identifikasi Bentuk Distribusi Data Dislikes di Tiap Kategori	*dislikes*
4	Identifikasi Bentuk Distribusi Data Komentar di Tiap Kategori	*comments*

Distribusi data divisualisasikan dengan bantuan pustaka data matplotlib dan seaborn berdasarkan kategori di tiap negara. Maka dari itu perlu dibuat dataframe baru yang memuat data penonton, likes, dislikes, dan komentar berdasarkan kategori video dengan menggunakan list comprehension pada masing-masing kolom yang diinginkan pada *dataframe* video berdasarkan ID kategori di *dataframe* kategori. *Dataframe* baru tersebut divisualisasikan ke dalam bentuk boxplot dengan fungsi .sns.boxplot().

4) Identifikasi Data Statistik pada Dataframe Video : Berikut adalah eksplorasi yang dilakukan untuk mengidentifikasi data statistik pada *dataframe* video tiap negara :

TABEL IIIII
IDENTIFIKASI DATA STATISTIK DATAFRAME VIDEO

No.	Objek Eksplorasi	Nama Kolom Eksplorasi
1	Identifikasi Jumlah Tertinggi, Terendah, Rata-Rata, dan Median Penonton dari Tiap Kategori	*views*
2	Identifikasi Jumlah Tertinggi, Terendah, Rata-Rata, dan Median Likes dari Tiap Kategori	*likes*
3	Identifikasi Jumlah Tertinggi, Terendah, Rata-Rata, dan Median Komentar dari Tiap Kategori	*comments*

Untuk menemukan data penonton, likes, dan komentar dari tiap kategori di masing-masing negara dapat dilakukan dengan membuat tiga *dataframe* baru di masing-masing negara yang berasal dari pengelompokkan berdasarkan ID kategori di *dataframe* video. Kemudian, *dataframe* diisi dengan jumlah tertinggi, terendah, rata-rata, dan median dari kolom views, likes, dan comments. Fungsi-fungsinya telah disediakan oleh Pandas. Jumlah tertinggi dapat ditemukan dengan fungsi .max(), jumlah terendah dengan fungsi .min(), jumlah rata-rata dengan fungsi .mean(), dan median dengan .median(). Masing-masing *dataframe* stastik ini kemudian divisualisasikan dalam bentuk *bar chart* dengan bantuan pustaka Bokeh.

5) Identifikasi Jumlah Video yang Trending Per Hari Berdasarkan Kategori di Tiap Negara : Untuk menemukan jumlah video yang trending per hari berdasarkan kategori di tiap Negara dibutuhkan *dataframe* baru yang memuat frekuensi video yang trending berdasarkan masing-masing kategori dari mulai 14 November 2017 hingga 14 Juni 2018. *Dataframe* baru dibuat berdasarkan *dataframe* video yang ditransformasi sehingga indeksnya adalah tanggal trending atau kolom 'trending_date' dengan kolom yang berisi jumlah video yang trend pada tanggal tersebut. Hal ini dapat dicapai dengan memanfaatkan fungsi Pandas yaitu pd.crosstab(['sumbu x', 'sumbu y']) di mana sumbu x adalah indeks ('trending_date') dan sumbu y adalah jumlah video yang trend dari masing-masing kategori. *Dataframe* ini akan divisualisasikan dalam bentuk *line plot* dengan bantuan pustaka Bokeh dengan fungsi .line(). Fungsi .line() hanya dapat membuat satu line plot saja, maka dari itu dilakukan perulangan berdasarkan jumlah kolom nama kategori yang ada untuk membuat plot dari semua kategori tersebut.

6) Identifikasi Tag Video yang Paling Banyak Digunakan : Pengidentifikasian diharapkan dapat memberi wawasan mengenai tag-tag yang berpotensi menyebabkan video masuk ke dalam tab Trending di Youtube. Untuk mendapatkan datanya, data di kolom 'tags' dari tiap *dataframe* video dimasukkan ke dalam list dengan fungsi .tolist(). String pada list tersebut diolah ulang dengan menghilangkan karakter yang tidak perlu dan menyamakan semua hurufnya menjadi non-kapital. Kemudian, list dijadikan *dataframe* dan masing-masing tag dihitung frekuensi kemunculannya dengan fungsi .value_count(). Hasil penghitungan divisualisasikan ke dalam bentuk visualisasi teks dengan bantuan pustaka WordCloud.

IV. HASIL PENELITIAN

Di bab ini akan dijelaskan mengenai hasil eksplorasi yang dilakukan berdasarkan subbab Eksplorasi Data di bab tiga. Tetapi, tidak semua hasil eksplorasi akan ditampilkan di sini. Yang akan ditampilkan hanyalah hasil eksplorasi yang dianggap memiliki makna yang cukup signifikan.

A. Hasil Normalisasi Data Kategori

Gambar berikut adalah hasil data kategori berformat JSON yang telah dinormalisasi dan dibentuk ulang ke dalam *dataframe* berdasarkan ID kategori yang ada pada data video.

Canada

	id	category
0	1	Film & Animation
1	2	Autos & Vehicles
2	10	Music
3	15	Pets & Animals
4	17	Sports
5	18	Short Movies
6	19	Travel & Events
7	20	Gaming
8	21	Videoblogging
9	22	People & Blogs
10	23	Comedy
11	24	Entertainment
12	25	News & Politics
13	26	Howto & Style
14	27	Education
15	28	Science & Technology
16	30	Movies
17	31	Anime/Animation
18	32	Action/Adventure
19	33	Classics
20	34	Comedy
21	35	Documentary
22	36	Drama
23	37	Family
24	38	Foreign
25	39	Horror
26	40	Sci-Fi/Fantasy
27	41	Thriller
28	42	Shorts
29	43	Shows
30	44	Trailers

Great Britain

	id	category
0	1	Film & Animation
1	2	Autos & Vehicles
2	10	Music
3	15	Pets & Animals
4	17	Sports
5	18	Short Movies
6	19	Travel & Events
7	20	Gaming
8	21	Videoblogging
9	22	People & Blogs
10	23	Comedy
11	24	Entertainment
12	25	News & Politics
13	26	Howto & Style
14	27	Education
15	28	Science & Technology
16	30	Movies
17	31	Anime/Animation
18	32	Action/Adventure
19	33	Classics
20	34	Comedy
21	35	Documentary
22	36	Drama
23	37	Family
24	38	Foreign
25	39	Horror
26	40	Sci-Fi/Fantasy
27	41	Thriller
28	42	Shorts
29	43	Shows
30	44	Trailers

United States

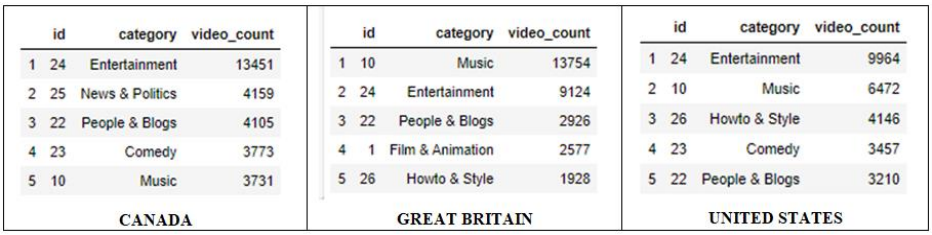
	id	category
0	1	Film & Animation
1	2	Autos & Vehicles
2	10	Music
3	15	Pets & Animals
4	17	Sports
5	18	Short Movies
6	19	Travel & Events
7	20	Gaming
8	21	Videoblogging
9	22	People & Blogs
10	23	Comedy
11	24	Entertainment
12	25	News & Politics
13	26	Howto & Style
14	27	Education
15	28	Science & Technology
16	29	Nonprofits & Activism
17	30	Movies
18	31	Anime/Animation
19	32	Action/Adventure
20	33	Classics
21	34	Comedy
22	35	Documentary
23	36	Drama
24	37	Family
25	38	Foreign
26	39	Horror
27	40	Sci-Fi/Fantasy
28	41	Thriller
29	42	Shorts
30	43	Shows
31	44	Trailers

Gambar 4. Data Kategori Setelah Normalisasi

Dapat dilihat bahwa Negara Canada dan Great Britain memiliki jumlah kategori yang sama yaitu 31 kategori, sedangkan United States memiliki 32 kategori. Perbedaannya adalah United States memiliki kategori Nonprofits & Activism.

B. Hasil Identifikasi Macam-Macam Top-5

Penemuan yang cukup penting ada pada eksplorasi mengenai "*Top-5* Kategori dengan Jumlah Video Tertinggi & Terendah Yang Masuk Trending" dan yang dicoba dibahas di sini adalah jumlah video tertingginya. Berikut adalah *Top-5* kategori dengan jumlah video tertinggi yang masuk trending YouTube di ketiga negara.



CANADA

	id	category	video_count
1	24	Entertainment	13451
2	25	News & Politics	4159
3	22	People & Blogs	4105
4	23	Comedy	3773
5	10	Music	3731

GREAT BRITAIN

	id	category	video_count
1	10	Music	13754
2	24	Entertainment	9124
3	22	People & Blogs	2926
4	1	Film & Animation	2577
5	26	Howto & Style	1928

UNITED STATES

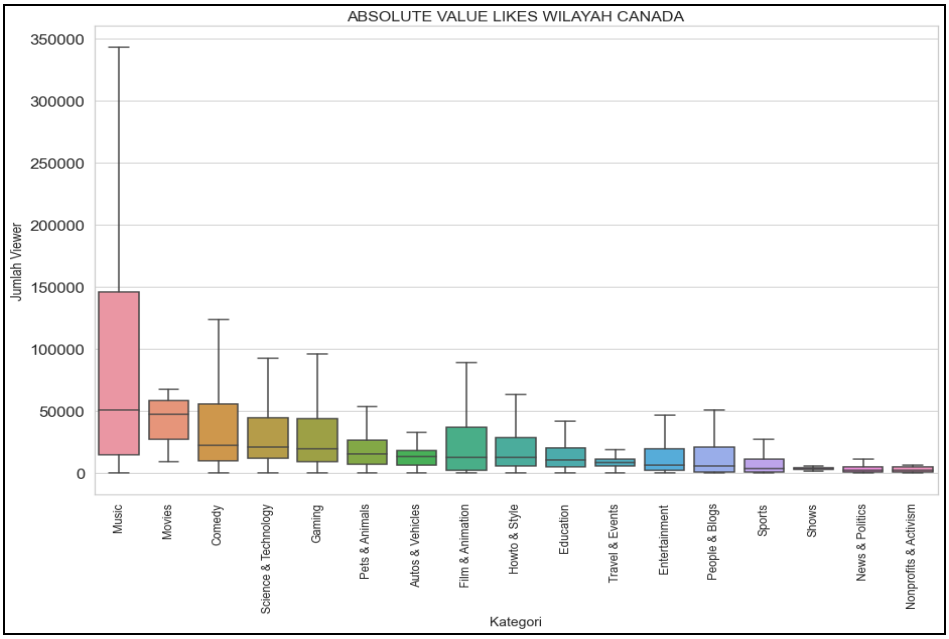
	id	category	video_count
1	24	Entertainment	9964
2	10	Music	6472
3	26	Howto & Style	4146
4	23	Comedy	3457
5	22	People & Blogs	3210

Gambar 5. Top-5 Kategori dengan Jumlah Video Tertinggi

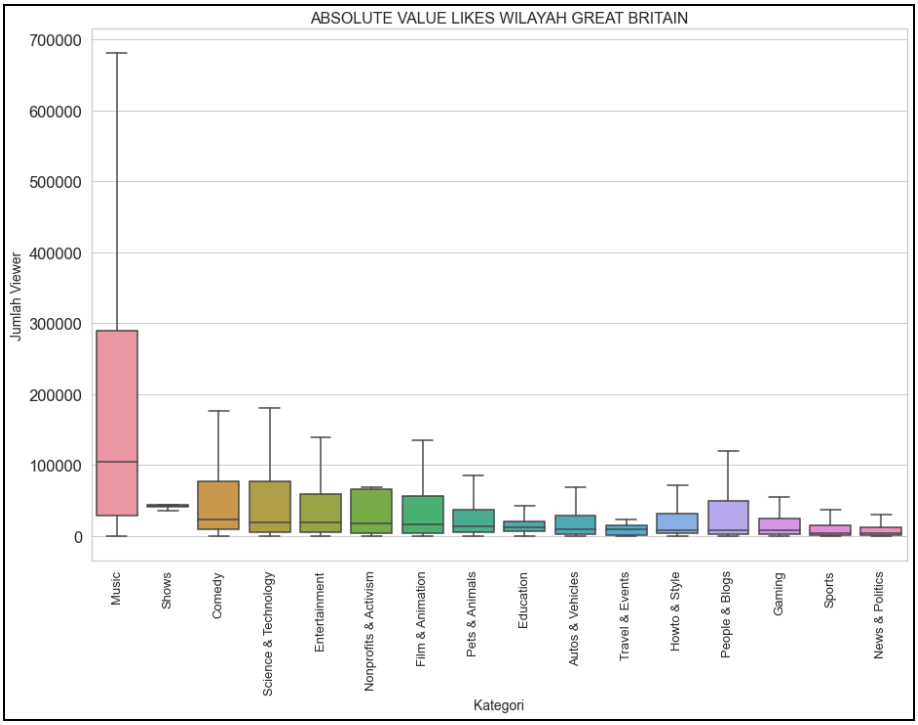
Berdasarkan gambar di atas, Kategori Entertainment menjadi kategori dengan jumlah video tertinggi yg masuk ke dalam tab trending di YouTube. Sedangkan di Great Britain Kategori Music menjadi yang tertinggi dan Entertainment sebagai ke-2 tertinggi. Di United States kategori Entertainment mnejadi yang tertinggi dan Music menjadi tertinggi ke-2. Tetapi, di Canada kategori Music berada di peringkat 5 kategori dengan jumlah video tertinggi yang masuk trending.

C. Visualisasi Bentuk Distribusi Data Tiap Kategori

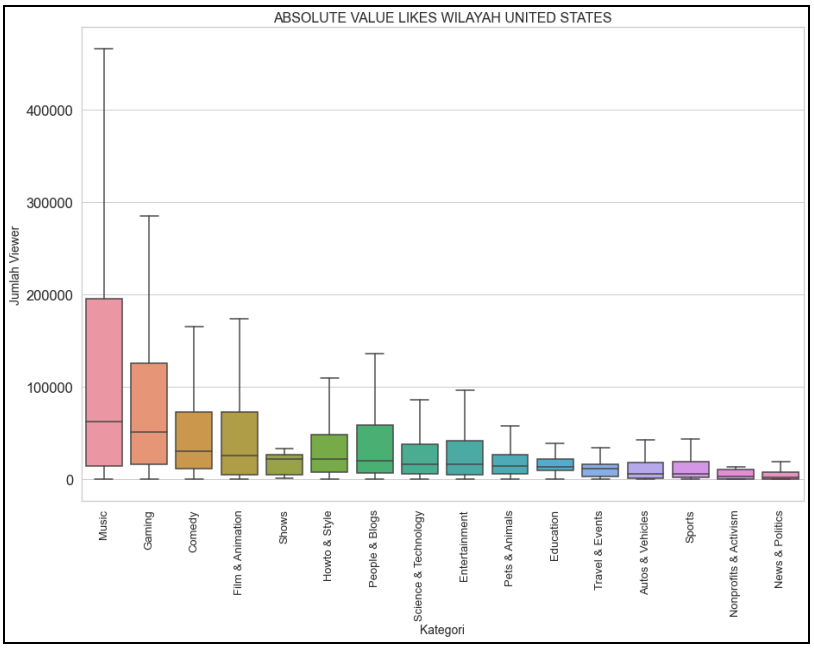
Setelah dilakukan eksplorasi bentuk distribusi data tiap kategori pada data penonton, likes, dislikes, dan komentar dengan *boxplot* hasilnya menunjukkan bahwa bentuk persebaran datanya sangat beragam dan tidak simetris sehingga sulit untuk mendeskripsikan *boxplot* dengan baik. Tetapi, ada kesamaan yang ditemukan pada boxplot data likes ketiga negara yang dapat dilihat di gambar berikut.



Gambar 6. Bentuk Distribusi Data Likes Tiap Kategori di Canada



Gambar 7. Bentuk Distribusi Data Likes Tiap Kategori di Great Britain

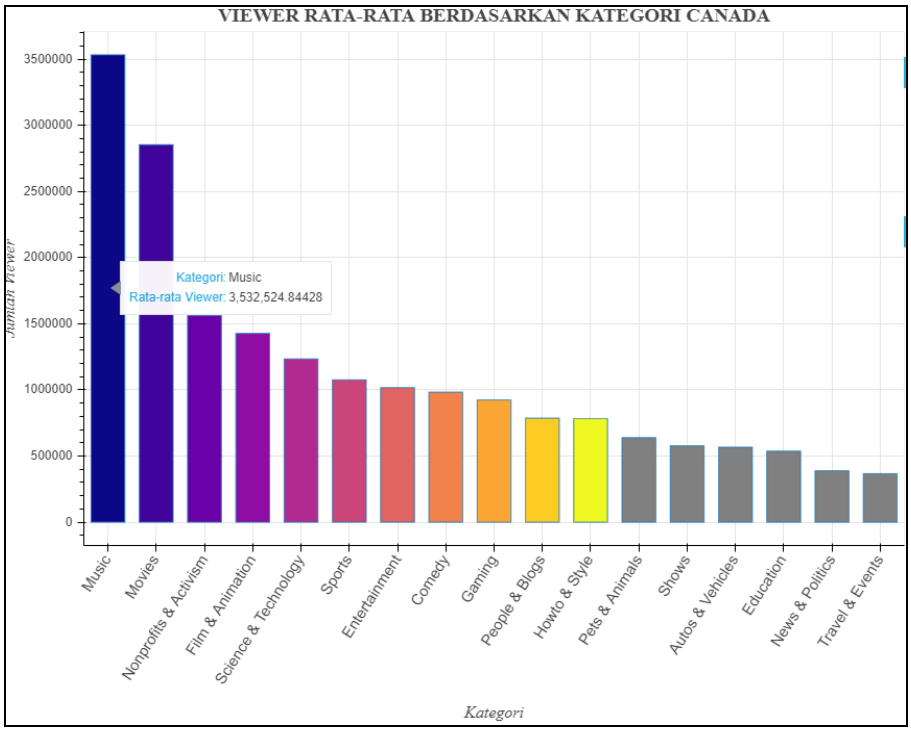


Gambar 8. Bentuk Distribusi Data Likes Tiap Kategori di United States

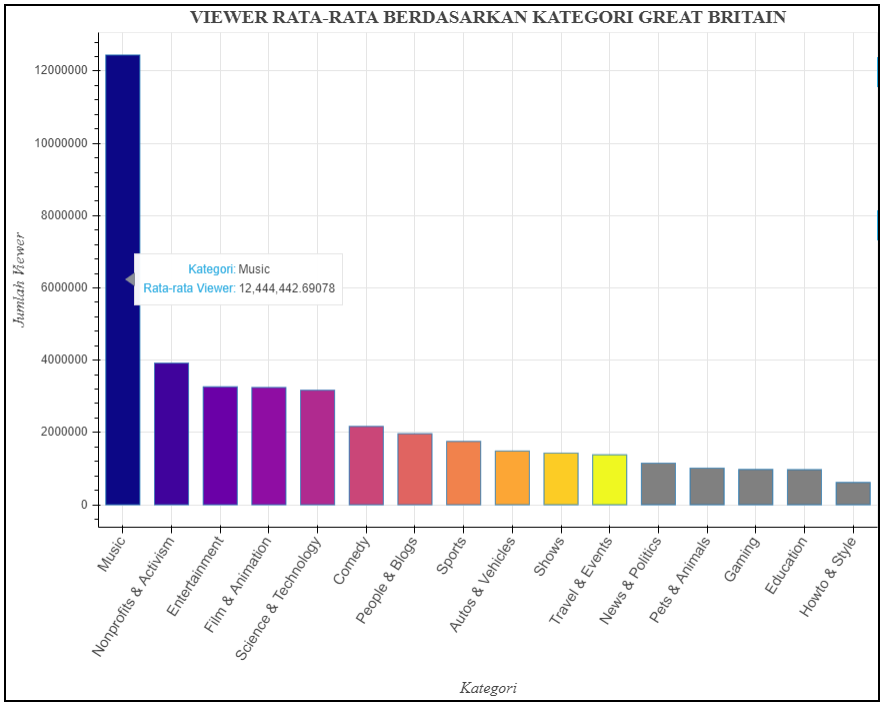
Dari ketiga gambar di atas dapat dipahami bahwa persebaran data likes pada kategori Music di ketiga negara telihat cukup luas berdasarkan besar dari boxnya. Median data likes pada kategori Music juga menjadi yang tertinggi di Canada dan Great Britain, sedangkan di United States mediannya tertinggi kedua setelah kategori Gaming.

D. Visualisasi Data Statistik pada Dataframe Video

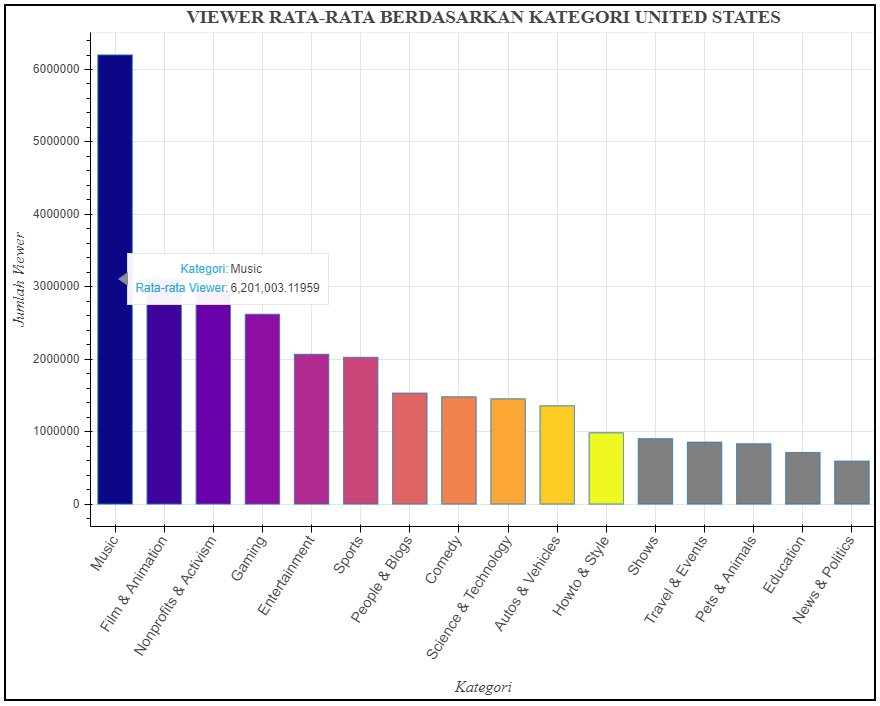
Eksplorasi Data Statistik pada Dataframe Video yang dilakukan pada data penonton, likes, dan komentar menghasilkan beberapa informasi menarik. Tetapi, kesamaan yang ditemukan ada pada rata-rata penonton dan likes. Berikut adalah tiga gambar yang menunjukan bar chart data penonton rata-rata dari ketiga negara.



Gambar 9. Bar Chart Data Penonton Rata-Rata di Canada

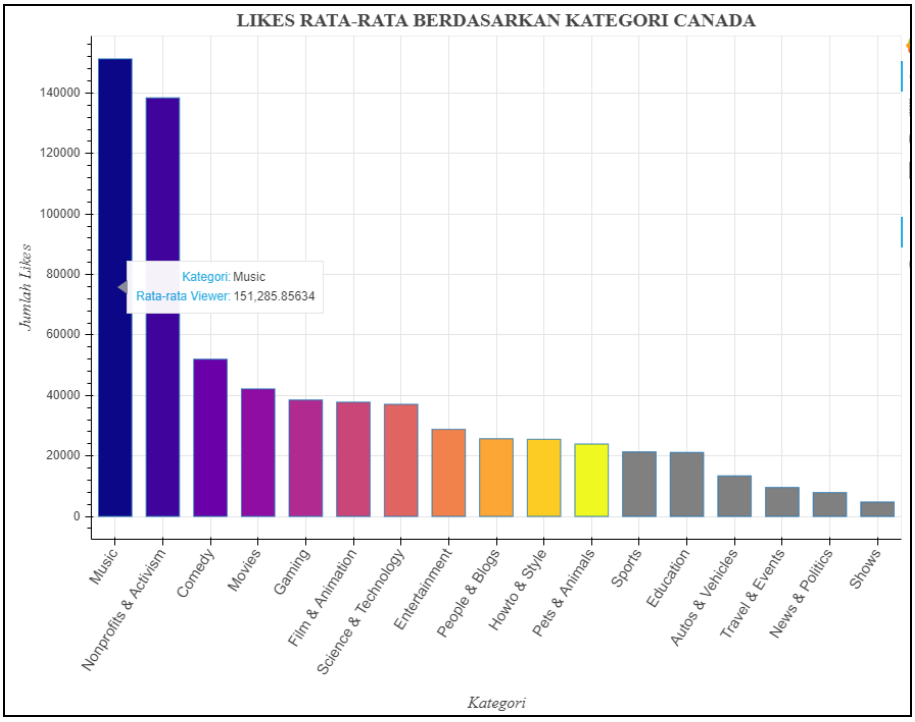


Gambar 10. Bar Chart Data Penonton Rata-Rata di Great Britain

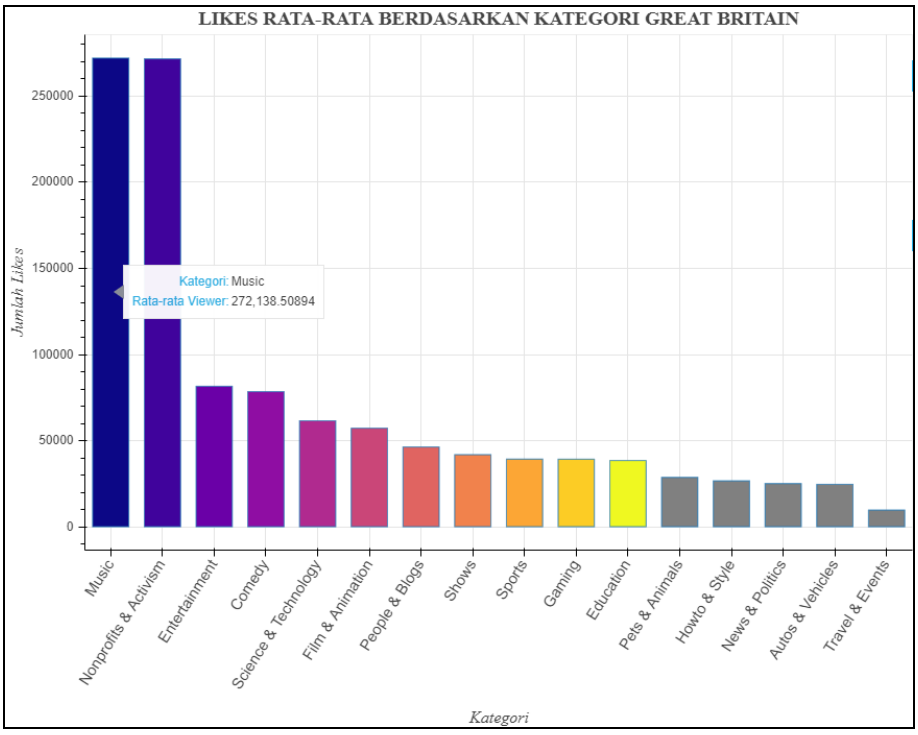


Gambar 11. Bar Chart Data Penonton Rata-Rata di United States

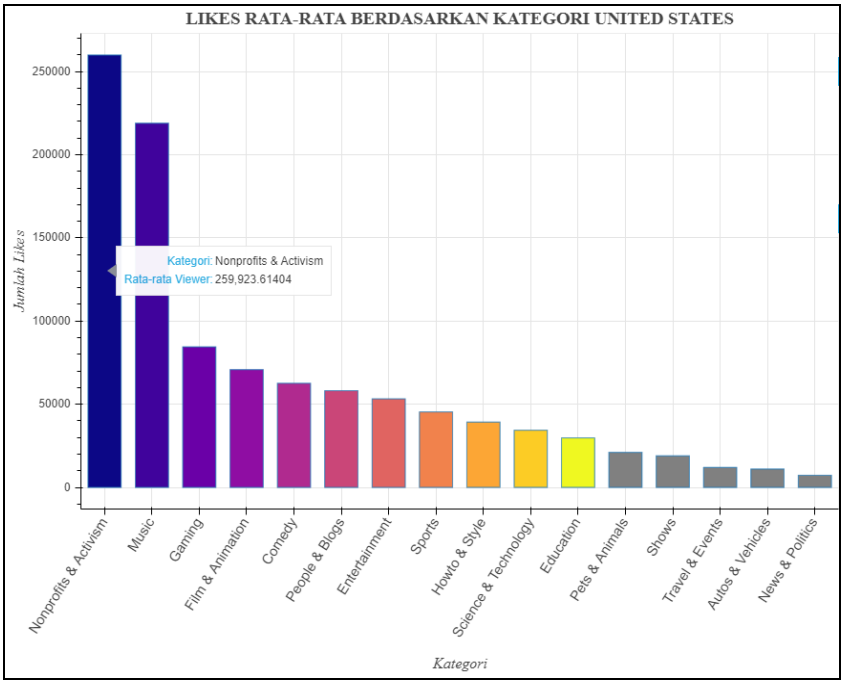
Ketiga gambar di atas menunjukkan bahwa kategori Music memiliki jumlah rata-rata penonton tertinggi pada ketiga negara, bahkan jaraknya terpaut cukup jauh dibandingkan dengan kategori dengan jumlah rata-rata penonton tertinggi ke-2 di Great Britain dan United States. Di Canada rata-rata penonton kategori Music adalah 3.532.524,84 penonton. Di Great Britain 12.444.442,691 penonton, dan di United States 6.201.003,11 penonton. Sedangkan berikut ini adalah tiga gambar yang menunjukkan jumlah rata-rata likes di tiap kategori pada ketiga negara.



Gambar 12. Bar Chart Likes Rata-Rata di Canada



Gambar 13. Bar Chart Likes Rata-Rata di Great Britain

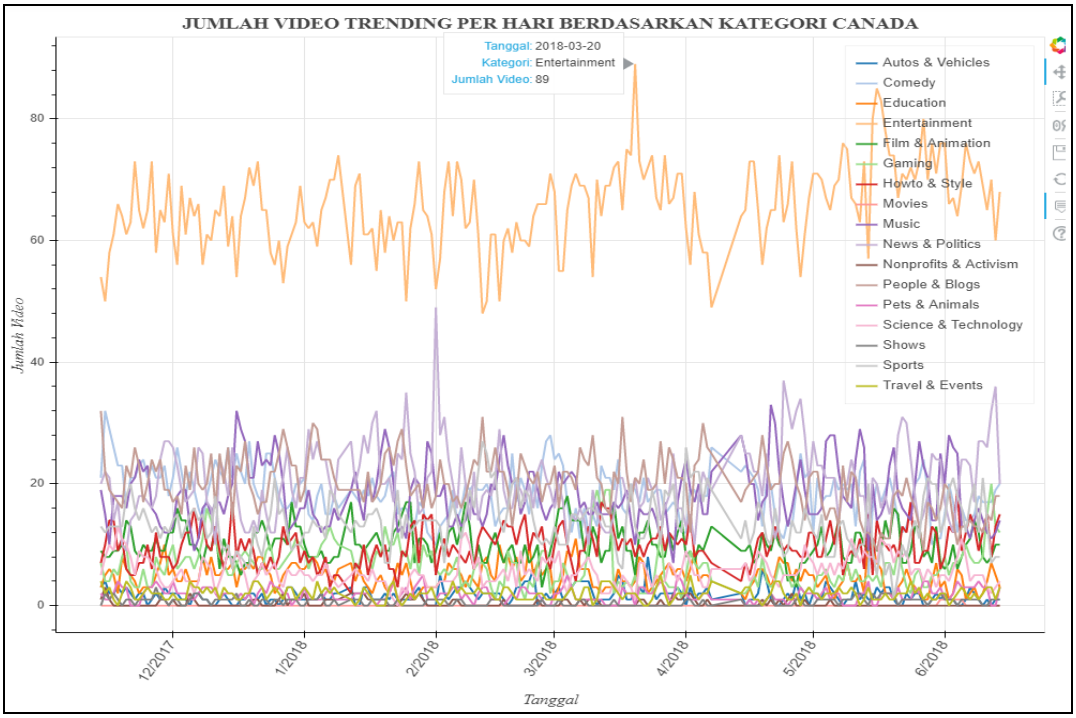


Gambar 14. Bar Chart Likes Rata-Rata di United States

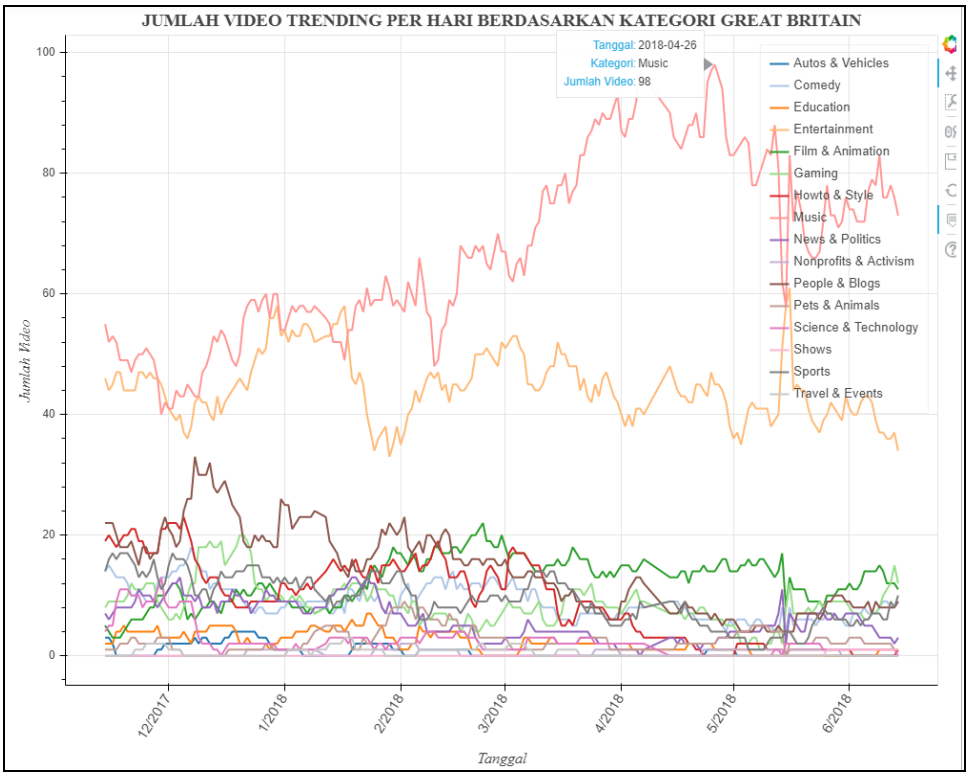
Ketiga gambar di atas menunjukkan bahwa Kategori Music di Canada memiliki rata-rata likes tertinggi sebanyak 151.285,85 likes. Di Great Britain kategori Music memiliki rata-rata likes tertinggi sebanyak 272.138,5 likes. Sedangkan di United States kategori Nonprofits & Activism dengan rata-rata likes tertinggi yaitu 259.923,61 likes.

E. Visualisasi Jumlah Video yang Trending Per Hari Berdasarkan Kategori di Tiap Negara

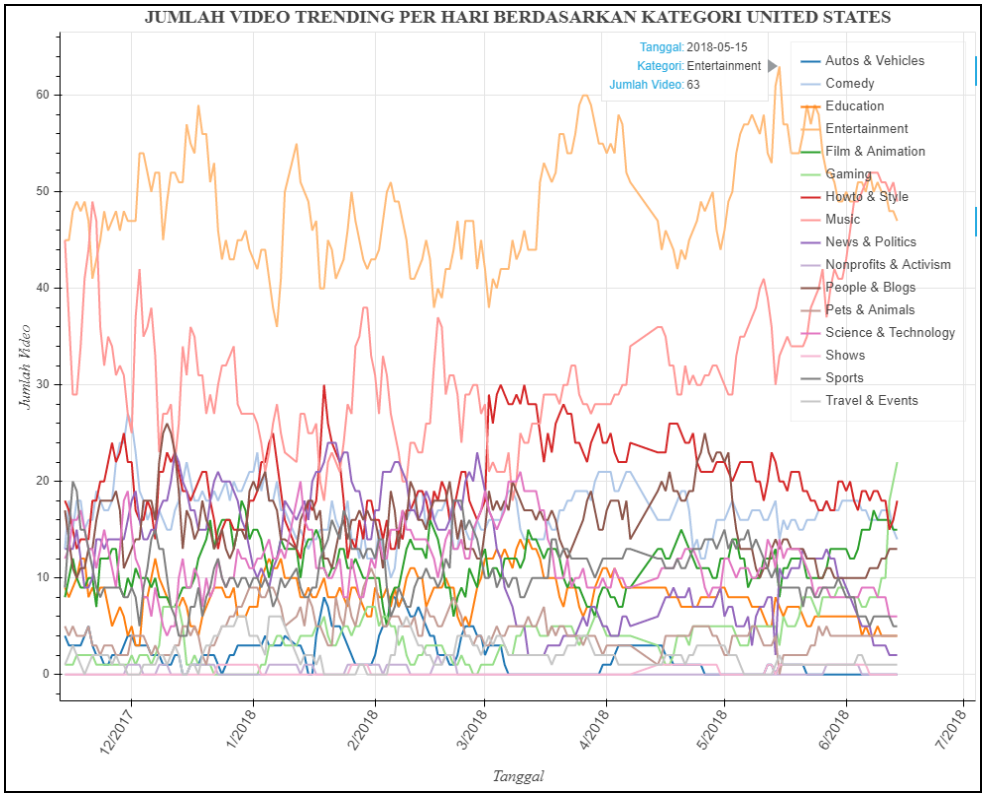
Berikut ini adalah visualisasi jumlah video yang trending per hari berdasarkan kategori di tiap negara dengan menggunakan *line plot* dari pustaka Bokeh. Visualisasi garis mewakili jumlah video yang trending per harinya.



Gambar 15. Line Plot Jumlah Video Trending Per Hari Berdasarkan Kategori di Canada



Gambar 16. Line Plot Jumlah Video Trending Per Hari Berdasarkan Kategori di Great Britain

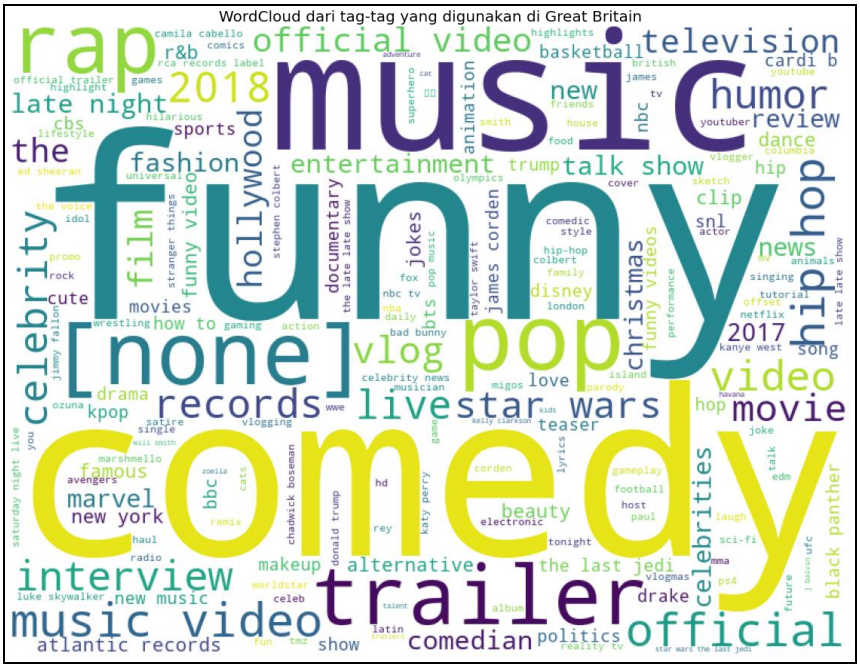


Gambar 17. Line Plot Jumlah Video Trending Per Hari Berdasarkan Kategori di United States

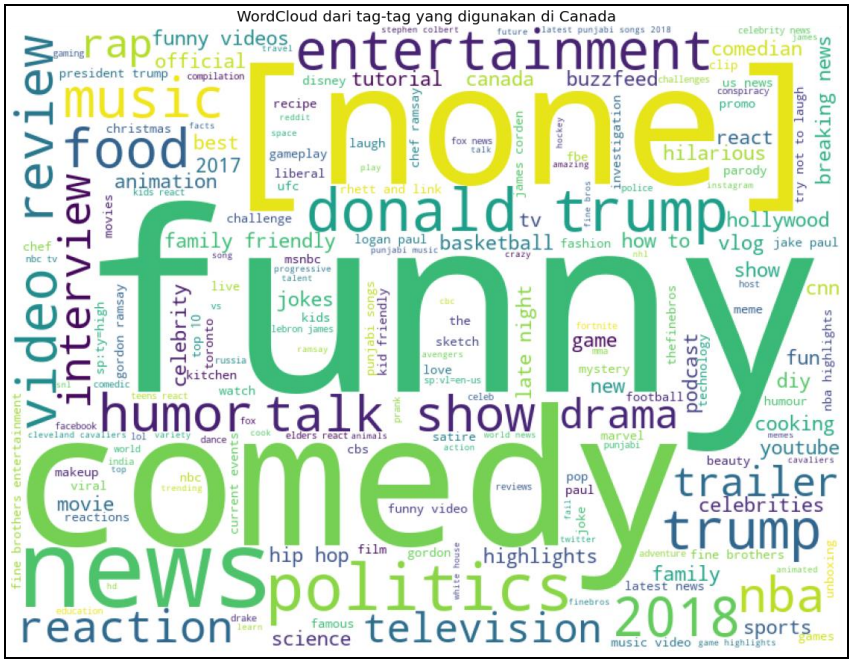
Kategori dengan jumlah video trending tertinggi di Canada adalah kategori Entertainment di tanggal 20 Maret 2018 dengan 89 video. Puncak jumlah video tertinggi yang masuk trend di Great Britain adalah kategori Music di tanggal 26 April 2018 dengan 98 video. Sedangkan Kategori dengan jumlah video trending tertinggi di United States adalah kategori Entertainment pada tanggal 15 Mei 2018 dengan 63 video.

F. Visualisasi Tag Video yang Paling Banyak Digunakan

Berikut adalah visualisasi tag video yang paling banyak digunakan pada ketiga negara dalam bentuk awan kata yang dibuat dengan bantuan pustaka WordCloud.



Gambar 18. Visualisasi Tag-Tag yang Digunakan Pada Tiap Video di Canada



Gambar 19. Visualisasi Tag-Tag yang Digunakan Pada Tiap Video di Great Britain

Gambar 20. Visualisasi Tag-Tag yang Digunakan Pada Tiap Video di United States

Berdasarkan tiga gambar di atas dapat dilihat bahwa tag-tag yang digunakan pada video yang Trending di tiap negara memiliki kemiripan yaitu tag ‘funny’ dan ‘comedy’ menjadi tag terbanyak yang dipakai di video tiap negara. Di Canada tag ‘funny’ digunakan sebanyak 3.891 kali, Di Great Britain tag ‘funny’ digunakan sebanyak 3.010 kali. Sedangkan di United States tag ‘funny’ digunakan sebanyak 4.142 kali. Sedangkan tag ‘comedy’ di Canada digunakan sebanyak 2.954, di Great Britain digunakan sebanyak 2.920 kali, dan di United States di Selain itu ternyata banyak juga video yang masuk Trending tanpa menggunakan tag, dapat dilihat dari ditemukannya tag ‘[none]’ yang ada di peringkat 3 di Canada dan peringkat 5 di Great Britain.

V. KESIMPULAN

Penelitian ini membuktikan bahwa *Exporatory data analysis* akan menghasilkan informasi yang beragam. Informasi yang dihasilkan akan bergantung pada sudut pandang pelaku eksplorasi. Informasinya dapat memiliki makna ataupun tidak memiliki makna sama sekali bergantung pada sudut pandang pelaku eksplorasi. Kesimpulan yang dapat diambil pun dapat beragam. Apabila dikaitkan dengan latar belakang penelitian ini, pola terjelas yang berhasil ditemukan di eksplorasi ini adalah bahwa video dengan kategori Music memiliki jumlah rata-rata penonton tertinggi pada ketiga negara. Begitu pula rata-rata likesnya walaupun rata-rata likes video kategori Music ada di nomor dua setelah Nonprofits & Activism di United States. Kemudian, kategori Entertainment adalah kategori dengan jumlah video terbanyak yang masuk ke dalam trend (per hari dan secara keseluruhan) di dua negara yaitu Canada dan United States. Sedangkan di Great Britain kategori Music menjadi kategori dengan jumlah video terbanyak yang masuk trend. Kemudian tag yang paling sering ditemukan di video Trending pada ketiga Negara adalah Funny dan Comedy.

Maka dari itu, kesimpulan yang didapatkan adalah video dengan kategori Music memiliki potensi untuk masuk ke trending berdasarkan jumlah rata-rata penonton dan jumlah rata-rata likes. Video dengan kategori Entertainment memiliki potensi untuk masuk ke trending berdasarkan jumlah video yang masuk ke dalam trending YouTube. Sedangkan, video yang memiliki tag Funny dan Comedy memiliki potensi untuk masuk ke trending berdasarkan frekuensi ditemukannya kedua tag tersebut.

DAFTAR PUSTAKA

[1] YouTube, "YouTube For Press," 2019. [Online]. Available: https://www.youtube.com/intl/en-GB/about/press/. [Accessed: 20-Oct-2019].
[2] YouTube, "Trending on YouTube," 2019. [Online]. Available: https://support.google.com/youtube/answer/7239739?hl=en. [Accessed: 20-Oct-2019].
[3] W. L. Martinez, A. R. Martinez, and J. L. Solka, *Exploratory Data Analysis with MATLAB Third Edition*, Third. Boca Raton: Taylor & Francis Group, 2017.
[4] J. J. Filliben, A. Heckert, C. Croarkin, B. Hembree, and W. Guthrie, *NIST/SEMATECH Engineering Statistics Handbook*. NIST, 2012.
[5] J. de Mast and B. P. H. Kemper, "Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn from a Well-Known Case?," *Res. Gate*, 2009.
[6] pandas 0.23.4 documentation, "Intro to Data Structures," 2018. [Online]. Available: https://pandas.pydata.org/pandas-docs/version/0.23.4/dsintro.html. [Accessed: 02-Oct-2019].
[7] Pandas, "pandas: powerful Python data analysis toolkit version 0.25.1," 2019. [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html. [Accessed: 02-Oct-2019].
[8] Aan Wahyu, "Pengenalan Pandas dan Series," 2019. [Online]. Available: https://petruknisme.com/2019/04/15/pengenalan-pandas-dan-series/. [Accessed: 12-May-2019].
[9] W. McKinney, "Python for Data Analysis. Data Wrangling with Pandas, NumPy, and IPython," 2nd ed., O'Reilly, 2018.
[10] A. E. Shadare and C. Akujuobi, "Data visualization," no. December, 2016.
[11] M. Waskom, "seaborn: statistical data visualization," 2018. [Online]. Available: https://seaborn.pydata.org/. [Accessed: 03-Oct-2019].
[12] Bokeh, "Bokeh Documentation," 2019. [Online]. Available: https://docs.bokeh.org/en/latest/index.html. [Accessed: 20-Mar-2020].
[13] D. Vu, "Generating WordClouds in Python," 2019. [Online]. Available: https://www.datacamp.com/community/tutorials/wordcloud-python. [Accessed: 13-Jun-2020].
[14] P. De Bérail, M. Guillon, C. Bungener, H. Processes-ea, U. Paris, and D. P. Cité, "Computers in Human Behavior The relations between YouTube addiction , social anxiety and parasocial relationships with YouTubers : A moderated-mediation model based on a cognitive-behavioral framework," *Comput. Human Behav.*, vol. 99, no. April, pp. 190–204, 2019.