LAPORAN TUGAS BESAR PENGENALAN KOMPUTASI
KU1102
LAPORAN DATA ANALYSIS TRENDING YOUTUBE

DISUSUN OLEH
KELOMPOK 8 - KELAS 23

GAGAS PRAHARSA BAHAR	**16520289**
MALIK AKBAR HASHEMI R.	**16520299**
ALIFIA RAHMAH	**16520309**
NG KYLE	**16520319**

DAFTAR ISI

I. LAPORAN INTI

1.1 Tugas 1 Deskripsi Data dan File

Dalam tugas analisis ini, kami menggunakan data set dari kaggle.com berupa data video trending di berbagai region. Untuk analisis ini kita menggunakan data region US (Amerika Serikat). Untuk sumber data yang digunakan adalah "USvideos.csv" pada https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv. Ukuran dari file csv tersebut adalah 40949 baris dan 16 kolom / atribut berformat csv. Selain itu sebagai file penunjang ada pula file json "US_category_id.json" pada https://www.kaggle.com/datasnaek/youtube-new?select=US_category_id.json sebagai referensi category_id dengan category_name yang akan membantu pada pengolahan data selanjutnya. Ukuran file data csv 59.85 MB dan file json 8.3 KB.

Dalam pengolahan data ini, kami menggunakan Google Colab dan menyimpan data pada Google Drive. Berikut blok program import data ke Google Colab:
(Untuk file code bisa diakses : https://colab.research.google.com/drive/1ffgTK6ESgxDT7FDgtacAvldfR5lkGyiN?usp=sharing)

```
# import module yang dibutuhkan
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from pandas.plotting import scatter_matrix

# link drive file csv
url_path =
'https://drive.google.com/uc?export=download&id=1_WDXGuFrDad9a5z0B9NvhbF
9Ft_yqsY2'

# Membaca file csv
df = pd.read_csv(url_path, index_col=0)
```

1.2 Tugas 2 Karakteristik Data

Pada data yang digunakan, berikut karakteristik masing-masing kolom:

Kolom	Penjelasan	Kategori	Range / Atribut Kategori	Unique Values
video_id	ID unik tiap video yang terdaftarkan pada database youtube	Categorical	-	6351
trending_date	Tanggal video tersebut termasuk dalam video trending / populer	Categorical (belum Time Series)	17-4-11 -- 18-14-06	205
title	Judul / nama video yang terdaftarkan pada database	Categorical	-	6455
channel_title	Nama channel yang mengupload video trending	Categorical	-	2207
category_id	Key category (masing-masing category memiliki key yang berbeda)	Categorical (Nominal)	22, 24, 23, 28, 1, 25, 17, 10, 15, 27, 26, 2, 19, 20, 29, 43	16
publish_time	Waktu mengunggah video	Categorical (belum Time Series)	2017-11-13T17:13:01.000Z, 2017-11-13T07:30:00.000Z, ..., 2018-06-13T09:00:06.000Z	6269
tags	Merupakan kumpulan kata kunci (Sebagai keyword pencarian)	Categorical	-	6055
views	Merupakan banyaknya video tersebut ditonton	Kuantitatif	549 -- 225 juta	40478
likes	Merupakan banyaknya likes setiap video	Kuantitatif	57527 -- 357079	29850
dislikes	Merupakan	Kuantitatif	2966 -- 212976	8516

	banyaknya dislikes setiap video			
thumbnail_link	Link unik tiap video untuk mengakses/menonton video	Categorical	-	6352
comments_disabled	Menentukan apakah kolom comment pada video dinyalakan atau tidak	Categorical (Binary)	True/False	2
ratings_disabled	Menentukan apakah opsi untuk menilai (like/dislike) dinyalakan atau tidak	Categorical (Binary)	True/False	2
video_error_or_removed	Menentukan apakah video tersebut masih ada/atau	Categorical (Binary)	True/False	2
description	Deskripsi pengupload video	Categorical	-	6901
*category_name	Kategori video berdasarkan key category_id	Categorical (Nominal)	'People & Blogs', 'Entertainment', 'Comedy', 'Science & Technology', 'Film & Animation', 'News & Politics', 'Sports', 'Music', 'Pets & Animals', 'Education', 'Howto & Style', 'Autos & Vehicles', 'Travel & Events', 'Gaming', 'Nonprofits & Activism', 'Shows'	16

Keterangan :
* data json yang telah di-append dan disesuaikan dengan key category_id.

Dari data.info() dapat dilihat kolom description terdapat data yang kosong. (Data lain 40949 sedangkan description hanya 40379)

```
#Ukuran Data
print("Banyak Baris :", len(df))
print("Banyak Kolom :", len(df.columns))

#Mencari data unik
for i in range (len(df.columns)):
  print("Data unik kolom", df.columns[i],"ada sebanyak",
len(df.iloc[:,i].unique()))

#Mencari range atau member category tiap kolom time series dan numerikal
num_date = ["trending_date","publish_time","views","likes","dislikes"]
for i in num_date:
  print("Kolom", i, "memiliki range :", df[i].min(), "--",df[i].max())
print()

#Mencari nilai unik kategorikal
kategorikal = ["category_id", "category_name"]
for i in kategorikal:
  print("Member pada kolom", i, "adalah:")
  for j in df[i].unique():
    print(j)
  print()
df.info()
```

1.3 Tugas 3 Data Preprocessing / Cleaning

Untuk data cleaning, kami menghapus kolom yang tidak akan digunakan untuk analisis data. Kolom yang kami hapus yaitu:

- video_id
  Kolom ini kami putuskan untuk dihapus karena nilainya berbeda untuk setiap data dan tidak dibutuhkan untuk analisis data.
- thumbnail_link
  Kolom ini kami putuskan untuk dihapus karena nilainya berbeda untuk setiap data dan tidak dibutuhkan untuk analisis data.
- description
  Kolom ini kami putuskan untuk dihapus karena nilainya berbeda untuk setiap data dan tidak dibutuhkan untuk analisis data.
- category_id, yang digantikan dengan category_name
  Kolom category_id kami gantikan dengan category_name dari data json yang telah di-append sehingga analisis dan pembacaan data menjadi lebih mudah.
- tags
  Kolom ini kami putuskan untuk dihapus karena nilainya berbeda untuk setiap data dan tidak dibutuhkan untuk analisis data.

```
# delete kolom yang tidak diperlukan
df.drop('video_id', inplace=True, axis=1)
df.drop('thumbnail_link', inplace=True, axis=1)
df.drop('description', inplace=True, axis=1)
```

```
df.drop('category_id', inplace=True, axis=1)
df.drop('tags', inplace=True, axis=1)
```

Selain itu, format publish_time dan trending_date kami ubah menjadi time series sehingga kami dapat mengolah data berdasarkan publish_time dan trending_date.

```
# mengubah tipe data publish_time dan trending_date menjadi time series
df['publish_time'] = pd.to_datetime(df['publish_time'],
format='%Y-%m-%dT%H:%M:%S.%fZ')
df["trending_date"]=pd.to_datetime(df["trending_date"],format="%y.%d.%m"
)
```

1.4 Tugas 4 Statistik

1.4.1 Statistik Umum

Berikut sampel data dari 5 baris pertama yang dapat ditampilkan menggunakan `df.head()`

	trending_date	title	channel_title	publish_time	views	likes	dislikes	comment_count	comments_disabled	ratings_disabled	video_error_or_removed	category_name
0	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	2017-11-13 17:13:01	748374	57527	2966	15954	False	False	False	People & Blogs
1	2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	2017-11-13 07:30:00	2418783	97185	6146	12703	False	False	False	Entertainment
2	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	2017-11-12 19:05:24	3191434	146033	5339	8181	False	False	False	Comedy
3	2017-11-14	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	2017-11-13 11:00:04	343168	10172	666	2146	False	False	False	Entertainment
4	2017-11-14	I Dare You: GOING BALD!?	nigahiga	2017-11-12 18:01:41	2095731	132235	1989	17518	False	False	False	Entertainment

Berikut statistik umum pada setiap atribut data numerik

Nama Kolom	Rata-rata	Standar Deviasi	minimum	maksimum
views	2.360785e+06	7.394114e+06	549	225211923
likes	7.426670e+04	2.288853e+05	0	5613827
dislikes	3.711401e+03	2.902971e+04	0	1674420
comment_count	8.446804e+03	3.743049e+04	0	1361580
comments_disabled	1.545825e-02	1.233680e-01	0	1
ratings_disabled	4.127085e-03	6.411047e-02	0	1
video_error_or_removed	5.616743e-04	2.369330e-02	0	1

Nama Kolom	Persentil				
	10%	25%	50%	75%	90%
views	70680.0	242329.0	681861.0	1823157.0	4602002.2

likes	1118.0	5424.0	18091.0	55417.0	160315.0
dislikes	58.0	202.0	631.0	1938.0	6033.2
comment_count	159.0	614.0	1856.0	5755.0	16959.2
comments_disabled	0.0	0.0	0.0	0.0	0.0
ratings_disabled	0.0	0.0	0.0	0.0	0.0
video_error_or_removed	0.0	0.0	0.0	0.0	0.0

Dapat dilihat dari comments disabled, ratings disabled, video error or disabled memiliki mean jauh kurang dari 0.5 yang mengindikasikan kebanyakan video dinyalakan komennya, dinyalakan ratingnya, dan videonya tetap dinyalakan maupun tidak error.

```
# informasi singkat data numerik
df.describe()

# Rata-rata
df.mean()


# Standar deviasi
df.std()


#Ekstremum minimum
df.min(numeric_only=True)


#Ekstremum maksimum
df.max(numeric_only=True)


#persentil
persentil = [0.1,0.25,0.5,0.75,0.9]
for i in persentil:
  print("Persentil", i*100,"%")
  print(df.quantile(i))
  print()
```

1.4.2 Statistik trending_date

Kolom trending_date merupakan kolom dengan tipe time series. Berikut merupakan statistik untuk trending date berdasarkan tiap bulan dan tahunnya.

tahun trending_date	bulan trending_date	views	likes	dislikes	comment_count	comment_disabled	ratings_disabled	video_error_or_removed
2017	11	42067981	1590323	5412376	1762021	62	18	1

		40	78		2			
	12	8237204070	297367071	23602890	3811686 9	123	27	0
2018	1	6441668933	251882440	22243761	3820591 4	69	33	0
	2	8046104893	217896307	10478352	2511255 2	108	53	3
	3	11548564376	402809262	14429156	4636690 2	81	29	0
	4	14590073499	442508556	16760897	4548006 8	50	9	8
	5	29181106586	828146533	36880602	8993414 0	111	0	11
	6	14420249655	441504651	22170121	4505150 7	29	0	0

```
# pengelompokan data berdasar bulan trending
df.groupby([(df.trending_date.dt.year),(df.trending_date.dt.month)]).sum
()
```

1.4.3 Statistik category_name

Dari pengelompokan berdasar masing-masing kategori, didapatkan tabel frekuensinya untuk data kuantitatif sebagai berikut.

category_name	views	likes	dislikes	comment_count
Auto & Vehicles	520690717	4245656	243010	784447
Comedy	5117426208	216346746	7230391	22545582
Education	1180629990	49257772	1351972	5442242
Entertainment	20604388195	530516491	42987663	73566498
Film & Animation	7284156721	165997476	6075148	17887060
Gaming	2141218625	69038284	9184466	14740713
Howto & Style	4078545064	162880075	5473899	23149550
Music	40132892190	1416838584	51179008	125296396
News & Politics	1473765704	18151033	4180049	6039433
Nonprofits & Activism	168941392	14815646	3310381	4808797

People & Blogs	4917191726	186615999	10187901	24778032
Pets & Animals	764651989	19370702	527379	2660705
Science & Technology	3487756816	82532638	4548402	11989926
Shows	51501058	1082639	24508	95117
Sports	4404456673	98621211	5133551	11192155
Travel & Events	343557084	4836246	340427	911511

category_name	comments_disabled	ratings_disabled	video_error_or_removed	frekuensi
Auto & Vehicles	5	10	0	384
Comedy	2	0	0	3457
Education	8	5	0	1656
Entertainment	196	30	8	9964
Film & Animation	28	14	13	2345
Gaming	8	0	0	817
Howto & Style	11	11	0	4146
Music	9	24	0	6472
News & Politics	174	1	0	2487
Nonprofits & Activism	4	4	0	57
People & Blogs	66	37	0	3210
Pets & Animals	4	0	0	920
Science & Technology	90	16	0	2401
Shows	0	0	0	57
Sports	28	17	2	2174
Travel & Events	0	0	0	402

```
# statistik yang dikelompokkan berdasarkan category_name
frekuensi_kategori = df.groupby("category_name").sum()
frekuensi_kategori
```

Tabel frekuensi yang berisikan category-category yang pernah masuk trending dan seberapa sering video dengan kategori tersebut masuk sebagai berikut.

category_name	frekuensi
Entertainment	9964
Music	6472
Howto & Style	4146
Comedy	3457
People & Blogs	3210
News & Politics	2487
Science & Technology	2401
Film & Animation	2345
Sports	2174
Education	1656
Pets & Animals	920
Gaming	817
Travel & Events	402
Autos & Vehicles	384
Shows	57
Nonprofits & Activism	57

```
# tabel frekuensi category-category yang pernah masuk trending
# dan seberapa sering video dengan kategori tersebut masuk
df['category_name'].value_counts()
```

1.4.4 Statistik channel_title

Dari kolom channel_title pada dataset ini, dapat dilihat seberapa sering channel tersebut masuk ke trending list. Berikut tabel frekuensi 10 besar channel yang sering masuk ke trending list.

channel_title	frekuensi
ESPN	203
The Tonight Show Starring Jimmy Fallon	197
Vox	193

TheEllenShow	193
Netflix	193
The Late Show with Stephen Colbert	187
Jimmy Kimmel Live	186
Late Night with Seth Meyers	183
Screen Junkies	182
NBA	181

```
# tabel frekuensi daftar 10 channel yang pernah masuk trending (paling sering)
df['channel_title'].value_counts()[:10]
```

1.4.5 Statistik Title (Ekstremum)

Berikut merupakan kumpulan title dari nilai ekstremum tiap data numerik. Untuk beberapa data yang memiliki beberapa atribut numerik lebih dari satu, diambil salah satu data saja.

Kategori	Minimum (Title)	Maksimum (Title)
views	1 dead, others injured after Ky. school shooting	Childish Gambino - This Is America (Official Video)
likes	Apple Clips sample	BTS (방탄소년단) 'FAKE LOVE' Official MV
dislikes	The Oak Beams of New College, Oxford	So Sorry.
comment_count	Amazon Christmas Advert 2017 - Toys & Games	So Sorry.

```
# Video dengan banyak views terbesar
df.loc[df['views'] == df['views'].max()]
# video dengan banyak likes terbesar
df.loc[df['likes'] == df['likes'].max()]
# video dengan dislike terbanyak
df.loc[df['dislikes'] == df['dislikes'].max()]
# video dengan komentar terbanyak
df.loc[df['comment_count'] == df['comment_count'].max()]

# Video dengan banyak views terkecil
df.loc[df['views'] == df['views'].min()]['title']
# video dengan banyak likes terkecil
df.loc[df['likes'] == df['likes'].min()]
# video dengan dislike tersedikit
df.loc[df['dislikes'] == df['dislikes'].min()]
```

```
# video dengan komentar tersedikit
df.loc[df['comment_count'] == df['comment_count'].min()]
# video dengan komentar tersedikit, meskipun comments_disabled false
df.loc[(df['comment_count'] == df['comment_count'].min()) &
(df['comments_disabled'] == False)]
```

1.4.5 Statistik publish_time

Untuk publish_time, kita melihat pengelompokan berdasar hari dan bulan apa dari suatu video di-publish ke Youtube.

Berikut pengelompokan berdasar hari dari waktu video di-publish ke Youtube.

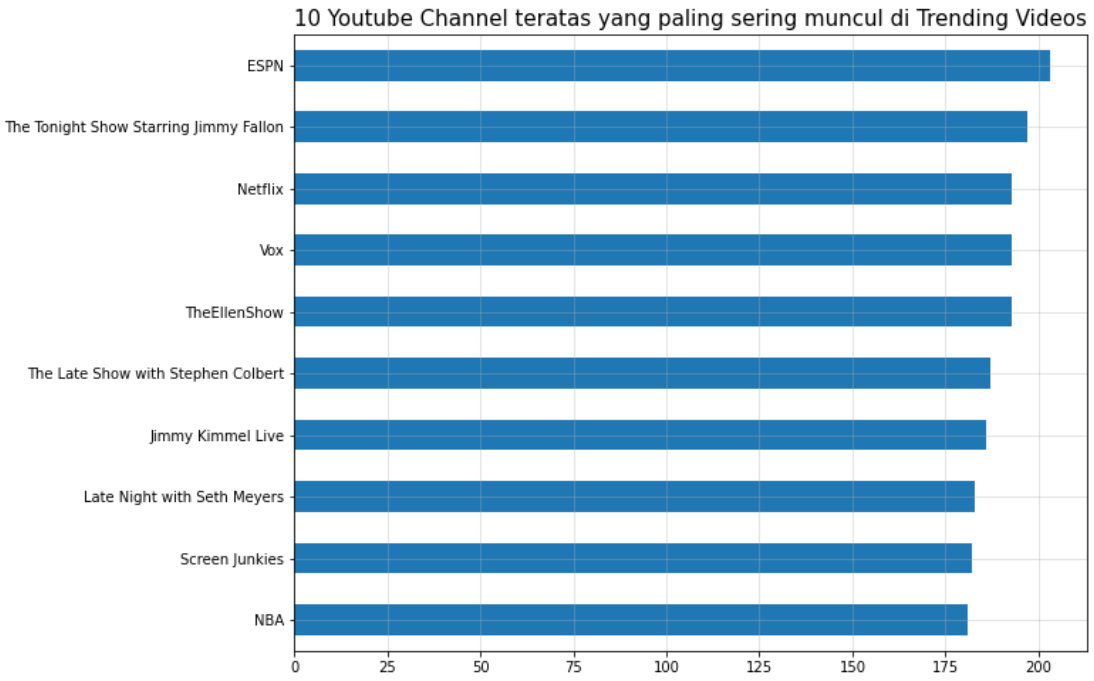
Hari	Video
Friday	7002
Thursday	6950
Tuesday	6786
Wednesday	6762
Monday	6177
Sunday	3679
Saturday	3593

```
# pengelompokan data berdasar hari di-publish
df["publish_time"].apply(lambda x:
x.strftime('%A')).value_counts().to_frame().reset_index()
```

1.5 Tugas 5 Visualisasi

1.5.1 Perbandingan Kategori

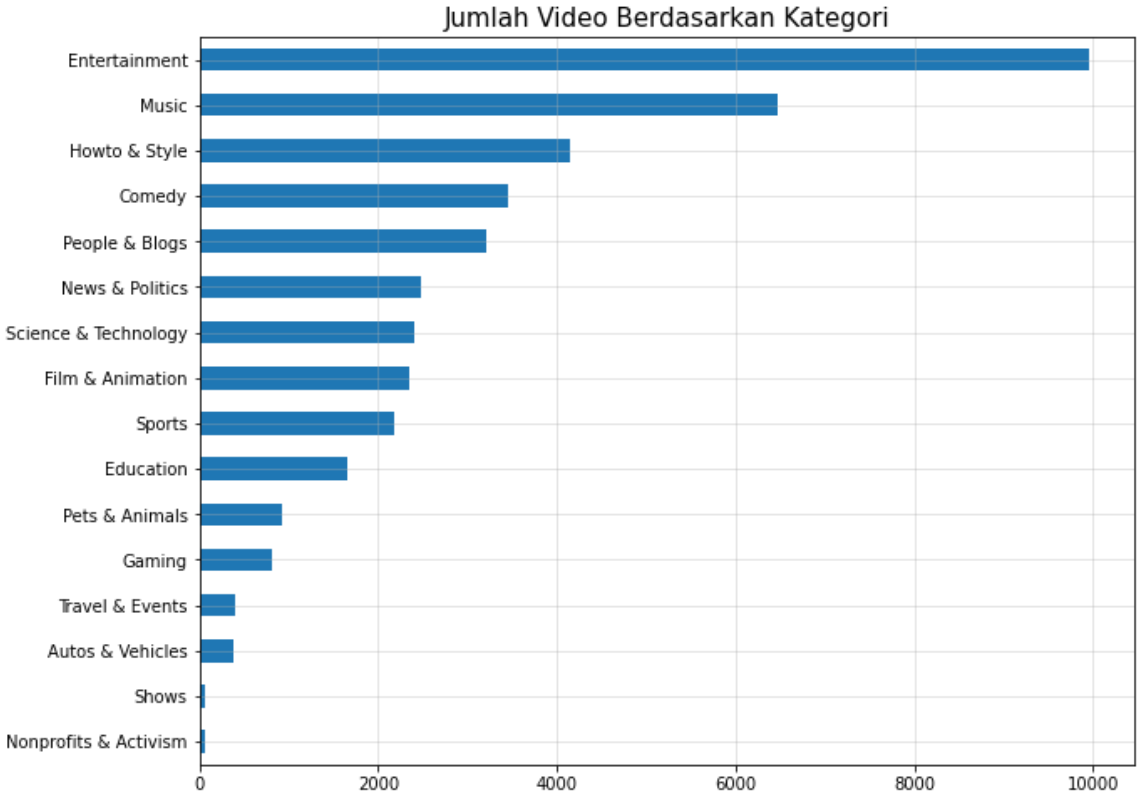
1.5.1.1 10 Channel Youtube Teratas Frekuensi pada Trending Video




```
# horizontal bar plot dari 10 youtube channel teratas yang paling sering
muncul di trending
df['channel_title'].value_counts().head(10).sort_values().plot(kind='bar
h', figsize=(10,8))
plt.title("10 Youtube Channel teratas yang paling sering muncul di
Trending Videos", size=(15))
plt.grid(alpha=0.4)
plt.show()
df['channel_title'].value_counts().head(10).sort_values().plot(kind='bar
h', figsize=(10,8))
plt.title("10 Youtube Channel teratas yang paling sering muncul di
Trending Videos", size=(15))
plt.grid(alpha=0.4)
plt.show()
```

Grafik tersebut menunjukkan 10 Youtube Channel teratas yang paling sering muncul di Trending Videos, dengan channel nomor 1 adalah ESPN. Berdasarkan grafik diatas, channel yang paling sering masuk Trending biasanya adalah channel yang kontennya rutin dan banyak penggemar, seperti basket dan *reality show*. Dari 10 besar semuanya merupakan channel komersil.

1.5.1.2 Jumlah Video Berdasarkan Kategori

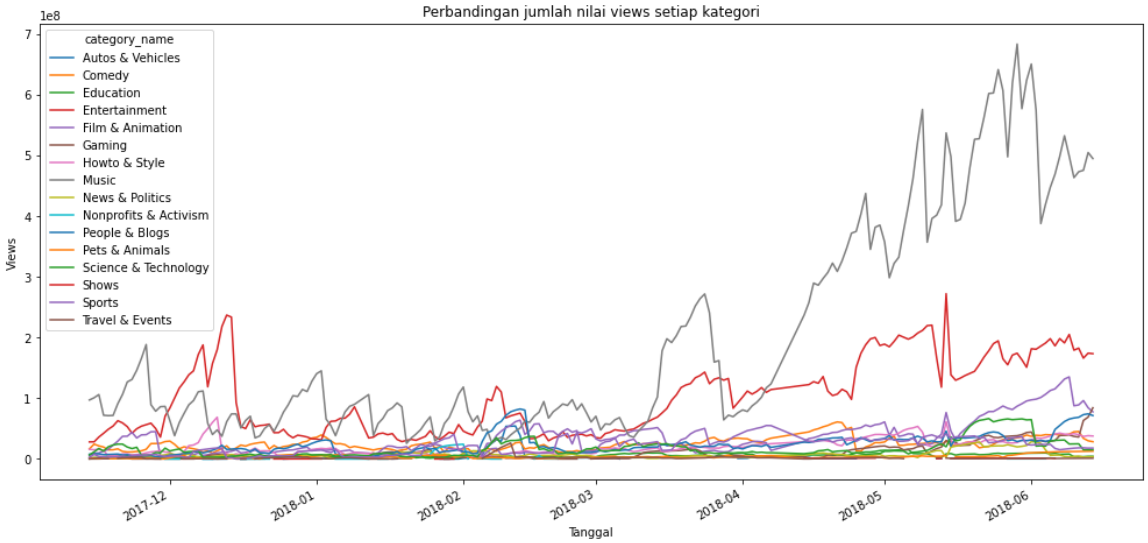



```
df['category_name'].value_counts().sort_values().plot(kind='barh',
figsize=(10,8))
plt.title("Jumlah Video Berdasarkan Kategori", size=15)
plt.grid(alpha=0.4)
plt.show()
```

Dari grafik ini, diketahui kategori dengan jumlah video terbanyak adalah Entertainment. Insight yang didapatkan dari grafik ini adalah kategori entertainment paling sering masuk trending, disusul dengan Music. Ini artinya banyak pengguna YouTube yang menonton untuk konten hiburan/mendengar musik.

1.5.2 Penampilan Perubahan terhadap Waktu

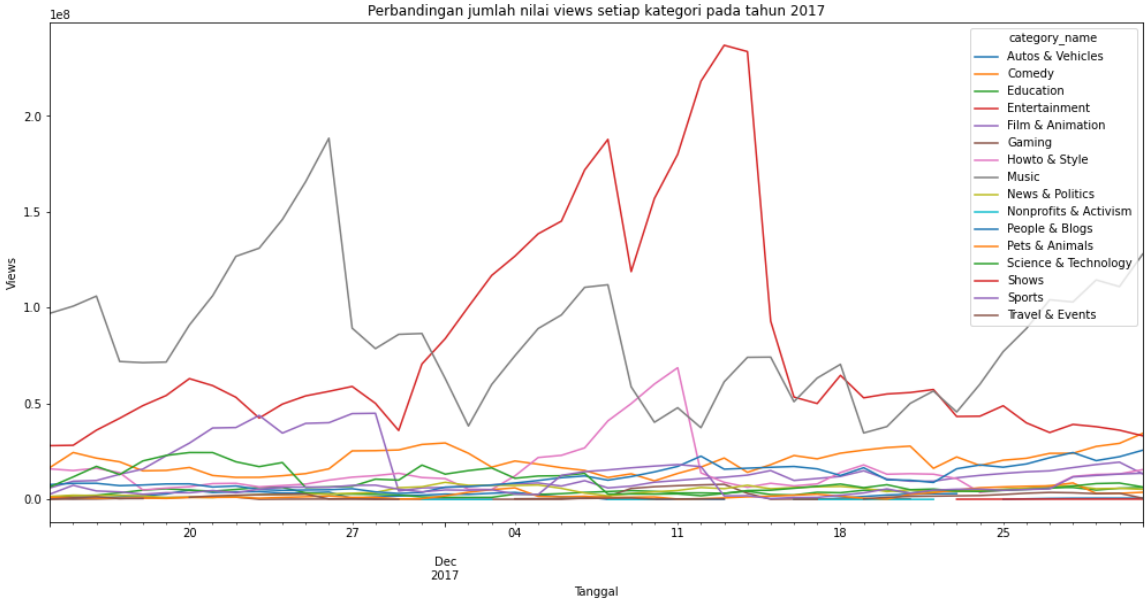
1.5.2.1 Perbandingan jumlah views tiap kategori

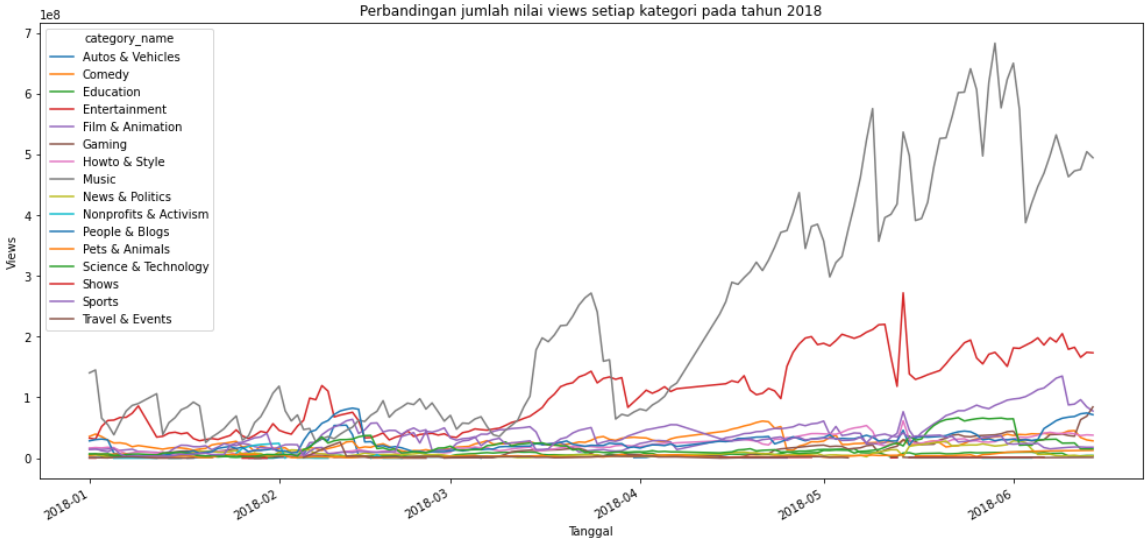



```
df.groupby(["trending_date","category_name"])["views"].sum().unstack().plot(kind="line",figsize=(17,8),title="Perbandingan jumlah nilai views setiap kategori")
plt.xlabel("Tanggal")
plt.ylabel("Views")
plt.show()
```

Dari grafik diatas, diketahui bahwa kategori Music naik mendominasi Trending dimulai dari bulan April 2018. Sebelum April 2018, kategori Music dan Entertainment bersaing untuk mendapatkan views terbanyak pada video yang trending. Berbeda dengan keseluruhan data, pada tahun 2017 kategori Entertainment lebih mendominasi. Dapat dilihat pula, semakin lama Music memiliki peminat yang semakin

tinggi.





Bila kita membagi menjadi 2 interval berdasarkan tahun, dari grafik diatas, diketahui bahwa kategori Music naik mendominasi Trending dimulai dari bulan April 2018. Sebelum April 2018, kategori Music dan Entertainment bersaing untuk mendapatkan views terbanyak pada video yang trending.

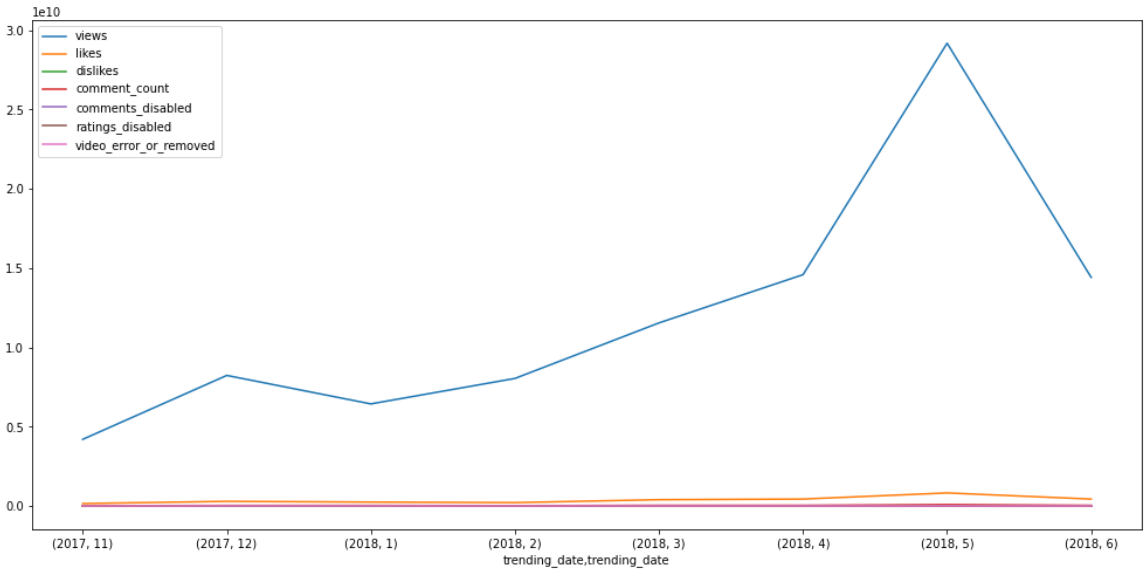
Berbeda dengan keseluruhan data, pada tahun 2017 kategori Entertainment lebih mendominasi atau dapat dikatakan kategori Music mulai mendominasi sejak tahun 2018, lebih tepatnya April 2018.

```
df.loc[df["trending_date"].dt.year ==
2017].groupby(["trending_date","category_name"])["views"].sum().unstack(
).plot(kind="line",figsize=(17,8),title="Perbandingan jumlah nilai views
setiap kategori pada tahun 2017")
plt.xlabel("Tanggal")
plt.ylabel("Views")
```

```
plt.show()

df.loc[df["trending_date"].dt.year ==
2018].groupby(["trending_date","category_name"])["views"].sum().unstack(
).plot(kind="line",figsize=(17,8),title="Perbandingan jumlah nilai views
setiap kategori pada tahun 2018")
plt.xlabel("Tanggal")
plt.ylabel("Views")
plt.show()
```

1.5.2.2 Perbandingan data tiap bulannya



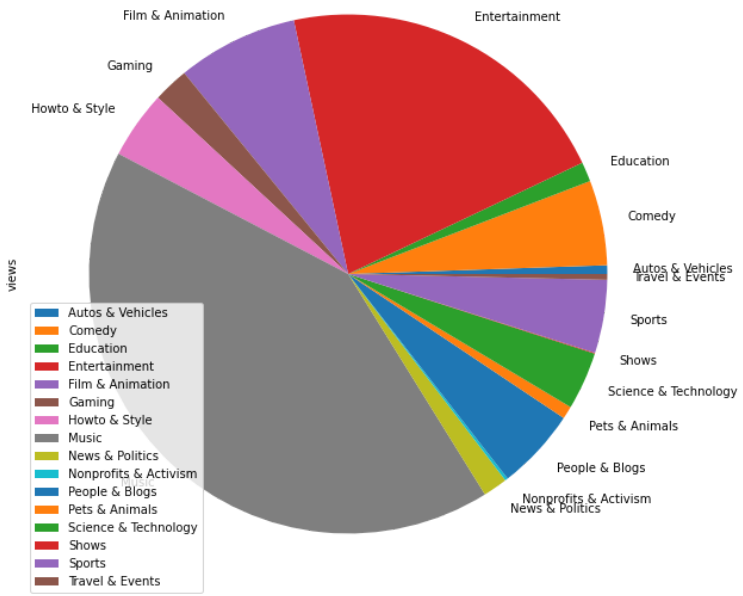
Dari grafik tersebut dapat dilihat nilai views selalu jauh dari nilai lainnya, ini dapat mengindikasikan kurangnya inisiatif penonton video untuk memberikan rating (like maupun dislike) maupun komen. Selain itu dapat dilihat secara garis besar jumlah penonton video youtube paling banyak (khususnya untuk trending) adalah pada Mei 2018.

```
trending.plot(kind="line", figsize = (17,8))
```

1.5.3 Penampilan Hierarki dan Hubungan Keseluruhan-Bagian

1.5.3.1 Perbandingan Jumlah Views Setiap Kategori

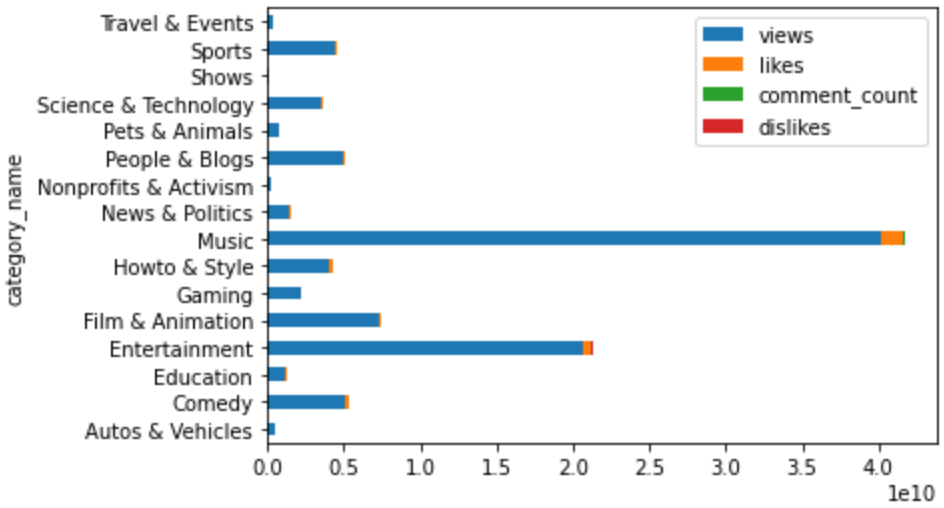
Berikut merupakan perbandingan jumlah views setiap kategori yang ditampilkan dengan pie chart.




```
df.groupby('category_name').sum()['views'].plot(kind='pie', legend=True,
figsize=(11,10))
```

Dari data ini, ditemukan bahwa kategori dengan jumlah views paling banyak yaitu kategori Music, diikuti dengan kategori Entertainment.

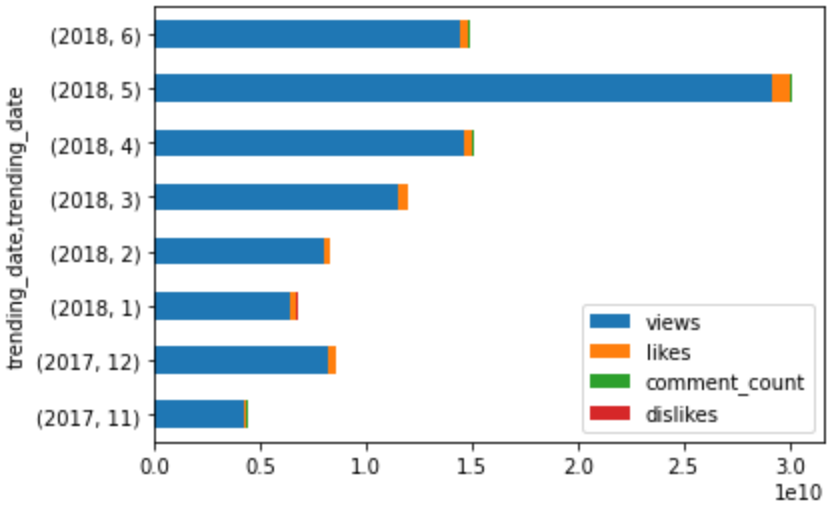
1.5.3.2 Perbandingan Jumlah Views, Likes, dan Comments Setiap Kategori




```
df.groupby(['category_name']).sum().plot(kind='barh', y=['views',
'likes', 'comment_count'], stacked=True)
```

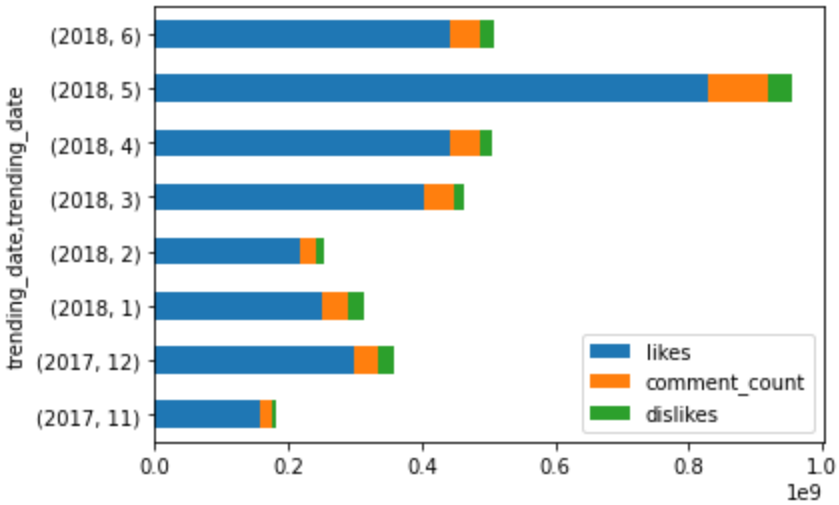
Pada grafik ini, terlihat perbandingan banyak views, likes, dan comment_count pada setiap kategori berbeda jauh. Dengan demikian, ditemukan insight bahwa penonton Youtube dominan hanya menonton video saja, sangat jarang reaksi berupa like dan comment.
Selain itu, ditemukan pula pada kategori "Entertainment", meskipun sedikit, terlihat dislikes yang cukup dominan dibandingkan dislikes dari kategori lain, yang sekilas tidak terlihat. Dengan demikian, ditemukan insight bahwa pada kolom "Entertainment" cenderung terdapat aktivitas dislike yang menonjol daripada kategori lain. Hal ini dapat disebabkan karena pada kategori entertainment terdapat berbagai video kontroversial yang menyebabkan banyak dislikes. Ini bisa juga disebabkan oleh haters suatu video yang suka men-dislike video dari selebriti tertentu.

1.5.3.3 Perbandingan Data Setiap Bulan




```
trending.plot(kind = "barh", stacked = True, y = ["views", "likes",
"comment_count","dislikes"])
```

Dapat dilihat bahwa proporsi views selalu jauh lebih besar dari gabungan likes, dislikes, dan comment count seperti halnya tiap kategori.



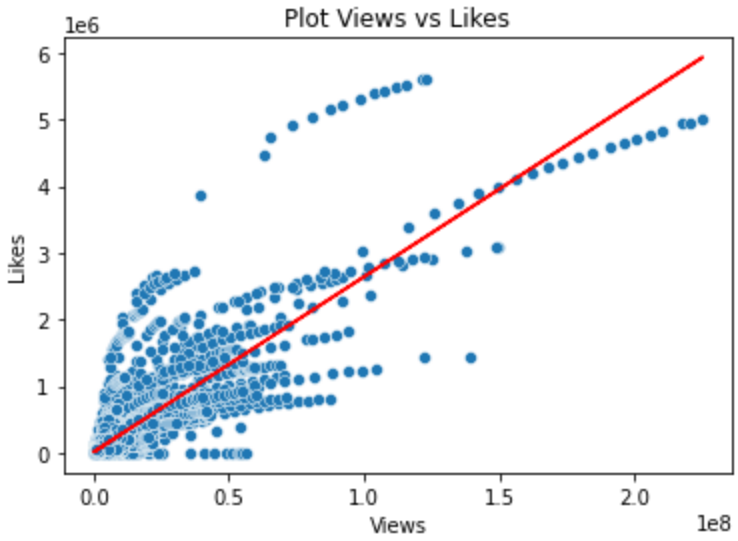
Jika kita lihat tanpa views, proporsi like melebihi comment count dan dislikes digabungkan. Bisa dikatakan seorang lebih menilai suatu video bila merasa video itu bagus dibandingkan merasa ingin dislike maupun memberikan komentar.

```
trending.plot(kind = "barh", stacked = True, y = ["likes",
"comment_count","dislikes"])
```

1.5.4 Plotting Relationship

1.5.4.1 Hubungan Antara Views dan Like (Keseluruhan)

Berikut hubungan antara views terhadap likes beserta regresi liniernya.

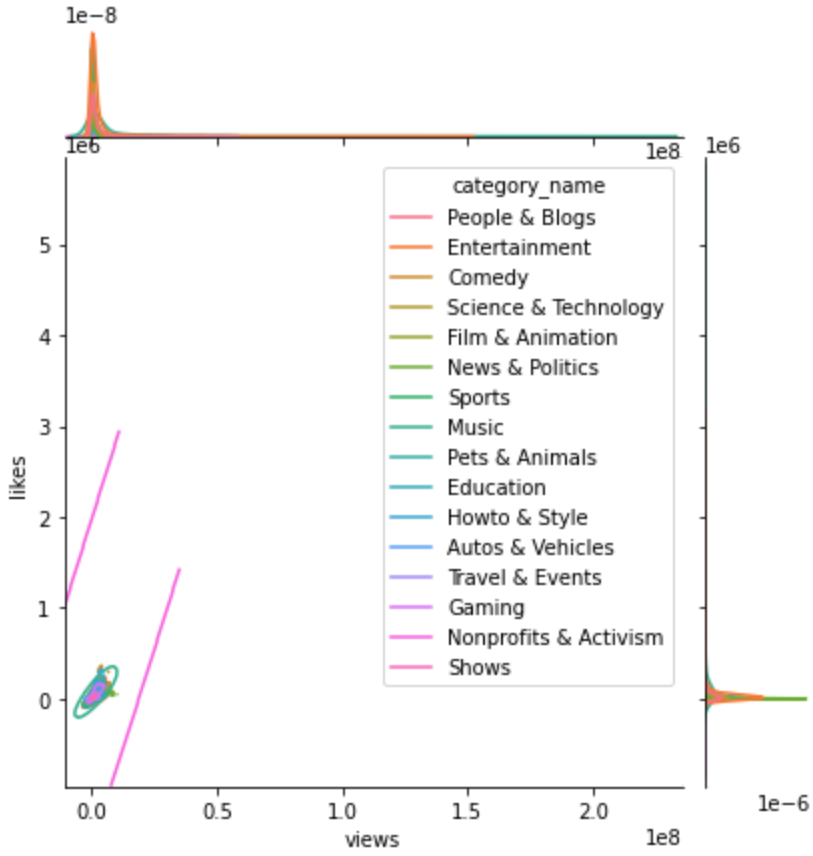


Dari data terlihat bahwa hubungan antara views dan likes berbanding lurus cukup kuat, didukung dengan gradien garis regresi yang mendekati 1 dan persebaran plot data kebanyakan searah dengan garis regresi.

```
x = df["views"].values.reshape(-1,1)
y = df["likes"].values.reshape(-1,1)

regresi = LinearRegression().fit(x,y)
hasil = regresi.predict(x)
sns.scatterplot(x=df["views"], y=df["likes"])
plt.plot(x,hasil, color="red")
plt.xlabel("Views")
plt.ylabel("Likes")
plt.title("Plot Views vs Likes")
```

1.5.4.2 Hubungan Views dan Likes (Category)



Dari data dapat dilihat semua category memiliki kepadatan antara hubungan views dengan likes yang realtif sama kecuali anomali pada Nonprofits & Activism yang dimungkinkan karena data yang cukup menyebar karena jumlah sample data yang sedikit (dapat dilihat dari tabel frekuensi Nonprofits & Activism paling rendah).

```
sns.jointplot(data=df, x="views", y="likes",hue = "category_name",
kind="kde", figsize = (18,8))
```

1.6 Tugas 6 Korelasi

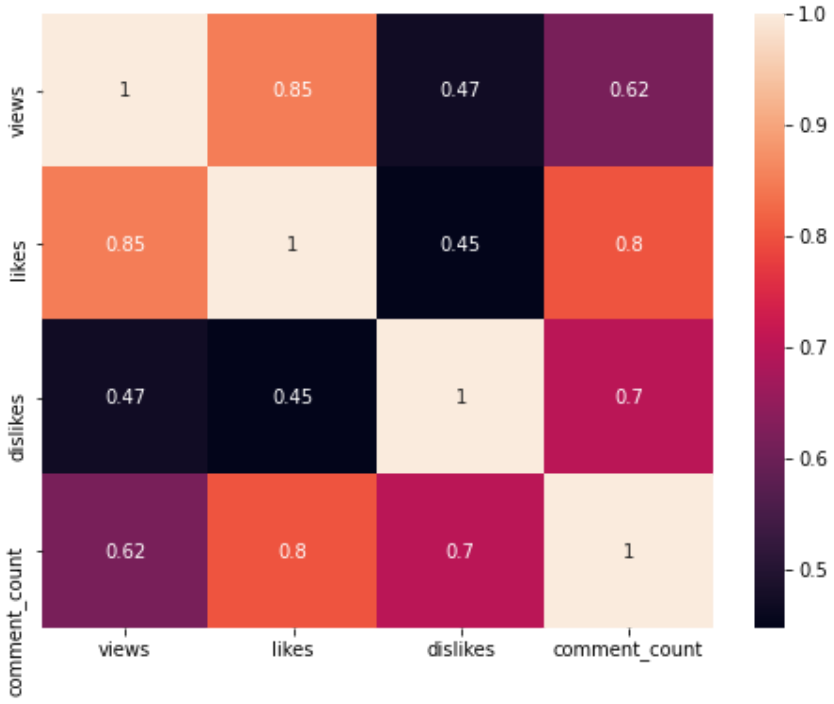
Berikut korelasi dari setiap kolom data kuantitatif pada dataset ini.

	views	likes	dislikes	comment_count
views	0.617621	0.849177	0.472213	0.617621
likes	0.849177	1.000000	0.447186	0.803057
dislikes	0.472213	0.447186	1.000000	0.700184
comment_count	0.617621	0.803057	0.700184	1.000000

```
correlation_list = ['views', 'likes', 'dislikes', 'comment_count']
hm_data = df[correlation_list].corr()
display(hm_data)
```

Dari tabel ini, didapatkan korelasi terbesar ditunjukkan oleh like-view dengan nilai R 0.849. Insight yang didapatkan dari tabel korelasi ini adalah apabila views tinggi, maka kecenderungan likes dan comment akan naik pula. Dan apabila likes atau dislikes tinggi, kecenderungan untuk mendapatkan comment naik juga. Sehingga dapat disimpulkan video trending biasanya viral dengan banyak likes yang berbanding lurus dengan views.

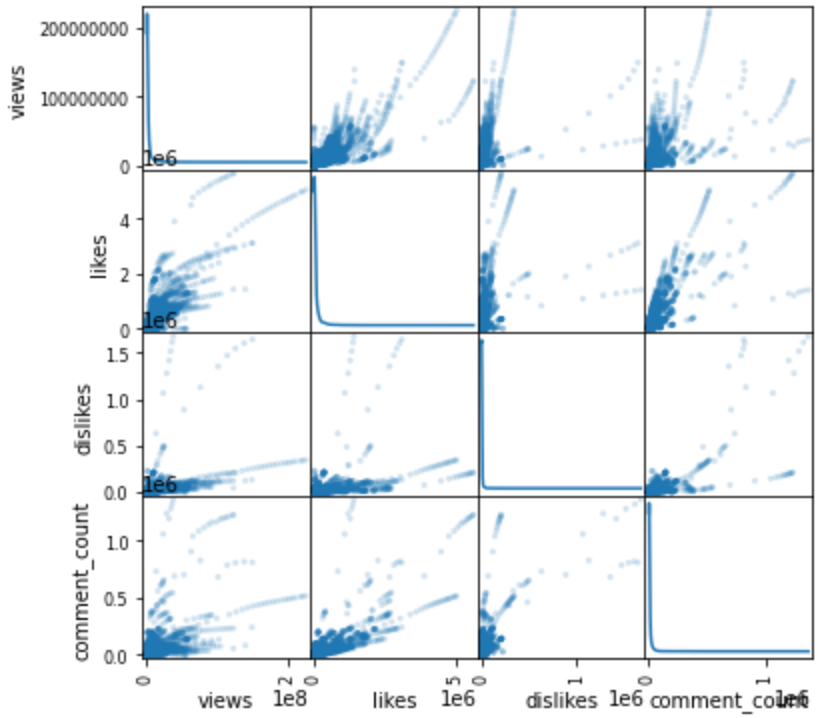
Berikut heatmap dari data korelasi yang telah dipaparkan diatas, agar lebih mudah dilihat dalam bentuk visual.



Warna yang semakin terang berarti semakin berkolerasi positif, sedangkan semakin gelap semakin tidak berkolerasi. Untuk diagonal utama, dapat dilihat bernilai 1 dan warna terang dikarenakan merupakan data kolom yang sama. Dapat dilihat pula kolom dislikes dengan likes dan views memiliki korelasi yang rendah (dapat dilihat nilai <0.5 dan warna yang gelap). Dapat dilihat yang paling berkorelasi satu sama lain adalaha likes dengan views yang korelasi paling tinggi yang diikuti antara likes dengan comment count, bisa diktakan kedua set tersebut bila semakin tinggi maka yang lainnya semakin tinggi pula (views tinggi, likes tinggi, comment count tinggi).

```
plt.figure(figsize=(8,6))
sns.heatmap(hm_data, annot=True);
```

Berikut scatter matrix dari korelasi yang telah dipaparkan di atas, agar dapat menampilkan seluruh grafik korelasi dari keseluruhan kolom data kuantitatif pada dataset ini.




```
scatter_matrix(df[correlation_list], alpha=0.2, figsize=(6, 6),
diagonal='kde')
```

II. KESIMPULAN DAN LESSON LEARNED

2.1 Kesimpulan

Untuk menganalisa data, pada hal ini data trending YouTube, diperlukan beberapa tahapan. Dimulai dari pencarian dataset, dilanjutkan dengan pembersihan data (membuang yang tidak diperlukan), lalu bisa mulai dianalisis.

Pada proses analisis data, statistik sangat penting untuk mengetahui karakteristik dari data yang dianalisis. Sehingga, pada awalnya, perlu ditampilkan statistik dari data yang akan diproses. Lalu kemudian dengan bantuan python dan module pandas serta pyplot, data mentah dapat divisualisasikan ke dalam bentuk yang dapat dicerna manusia.

Dengan data yang telah divisualisasikan, data akan memunculkan *insight-insight* baru yang awalnya tenggelam dalam data. Pengetahuan ini dapat digunakan untuk berbagai hal. Pada analisis data YouTube sendiri, kami dapat melihat karakteristik video yang trending.

2.2 Lesson Learned

Dari proses analisis data, banyak sekali pelajaran yang dapat diambil. Untuk mempermudah pengerjaan, kami menggunakan Google Colab yang memudahkan kolaborasi secara online.

Untuk membersihkan data, diperlukan analisis yang cukup dalam. Data-data yang tidak berguna dan dapat mengotori proses analisis harus dibuang. Pada perjalanannya, kami mempelajari banyak *syntax* baru yang sebelumnya belum kami ketahui, yang membuka lebih banyak kemungkinan dalam analisis.

Pada proses analisis data, diperlukan alur kerja yang sistematis. Tanpa pengerjaan yang sistematis dan berurut, maka akan lebih sulit untuk mengambil hasil dari analisis data dan akan mempersulit proses analisis data itu sendiri. Selain itu, penggunaan komentar juga sangat penting, khususnya apabila bekerja sama dengan orang lain. Penambahan komentar dapat menyampaikan maksud dari suatu baris program sehingga *workflow* atau alur kerja antar tim akan menjadi lebih lancar.

III. PEMBAGIAN TUGAS DALAM KELOMPOK

Tugas dikerjakan bersama-sama secara sinkron dan dinamis.
Selain itu, kami juga saling membantu satu sama lain dalam mengerjakan bagian masing-masing.

IV. DAFTAR REFERENSI

Trending YouTube Video Statistics - Kaggle.com. (2019, June 03). Retrieved December 09, 2020, from https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv