HOMEWORK

FINAL PROJECT

Theofilus Arifin

Christofer Bryan N. K.

Ramlan Apriyansyah

Muhammad Iqbal

Hanifah Arrasyidah

Christopher Stephen

Muhammad Rizq N. A.

Ujang Pian

1. EDA & Preprocessing

Statistic Descriptive

Statistic Descriptive akan menjelaskan data secara statistik.

Data Info

```
0   trending_date           36791 non-null  object
1   title                   36791 non-null  object
2   channel_title           36791 non-null  object
3   category_id             36791 non-null  int64
4   publish_time            36791 non-null  object
5   tags                    36791 non-null  object
6   likes                   36791 non-null  int64
7   dislikes                36791 non-null  int64
8   comment_count           36791 non-null  int64
9   comments_disabled       36791 non-null  bool
10  ratings_disabled        36791 non-null  bool
11  video_error_or_removed  36791 non-null  bool
12  description             36746 non-null  object
13  No_tags                 36791 non-null  int64
14  desc_len                36791 non-null  int64
15  len_title               36791 non-null  int64
16  publish_date            36791 non-null  datetime64[ns]
17  views                   36791 non-null  int64
```

Menggunakan function info dapat diketahui bahwa terdapat kesalahan tipe data pada trending date dan publish time. Kedua data tersebut seharusnya adalah datetime namun bertipe object

Null Data

```
trending_date             0
title                     0
channel_title             0
category_id               0
publish_time              0
tags                      0
likes                     0
dislikes                  0
comment_count             0
comments_disabled         0
ratings_disabled          0
video_error_or_removed    0
description              45
No_tags                   0
desc_len                  0
len_title                 0
publish_date              0
views                     0
dtype: int64
```

.

Mmenggunakan isna dapat diketahui bahwa terdapat 45 data kosong pada description, data ini perlu diberi suatu perlakuan khusus agar bisa digunakan.

Category Describe

	title	category_id	channel_title	tags	comments_disabled	ratings_disabled	video_error_or_removed	description
count	36791	36791	36791	36791	36791	36791	36791	36746
unique	16431	17	1390	12463	2	2	2	13979
top	Mission: Impossible - Fallout (2018) - Officia...	24	VikatanTV	[none]	False	False	False	Subscribers Link: http://bit.ly/2qb69dZ\n\nCon...
freq	19	16462	284	1120	35611	36034	36780	166

Kolom comments_disabled, ratings_disabled, video_error_or_removed merupakan kategorikal biner dan terlalu dominan pada satu kategori saja. Maka dari itu, kolom ini kemungkinan besar tidak memberikan dampak yang signifikan terhadap data dan layak untuk di drop

Numeric Describe

	views	likes	dislikes	comment_count	No_tags	desc_len	len_title
count	3.679100e+04	3.679100e+04	3.679100e+04	36791.000000	36791.000000	36791.000000	36791.000000
mean	1.071490e+06	2.745069e+04	1.685363e+03	2714.022043	18.938463	923.079123	70.609361
std	3.207149e+06	9.783129e+04	1.619732e+04	14978.114328	9.843531	815.038867	22.409174
min	4.024000e+03	0.000000e+00	0.000000e+00	0.000000	1.000000	3.000000	5.000000
25%	1.256040e+05	8.790000e+02	1.090000e+02	83.000000	12.000000	368.000000	53.000000
50%	3.078360e+05	3.126000e+03	3.310000e+02	336.000000	19.000000	677.000000	74.000000
75%	8.066315e+05	1.409500e+04	1.032000e+03	1314.500000	25.000000	1237.000000	91.000000
max	1.254322e+08	2.912710e+06	1.545017e+06	827755.000000	72.000000	5136.000000	100.000000

1. Likes
Rata-rata likes per video adalah sekitar 27,450, dengan variasi yang signifikan (standar deviasi sekitar 97,831). Video yang tidak mendapatkan likes memiliki nilai minimum 0, sementara video dengan likes tertinggi mencapai 2,912,710.

2. Dislikes
Jumlah dislikes rata-rata per video adalah sekitar 1,685, dengan variasi yang cukup besar (standar deviasi sekitar 16,197). Video yang tidak mendapatkan dislikes memiliki nilai minimum 0, sementara video dengan dislikes tertinggi mencapai 1,545,017.

3. Comment Count:
Rata-rata comment count per video adalah sekitar 2,714, dengan variasi yang signifikan (standar deviasi sekitar 14,978). Video tanpa komentar memiliki nilai minimum 0, sementara video dengan comment count tertinggi mencapai 827,755.

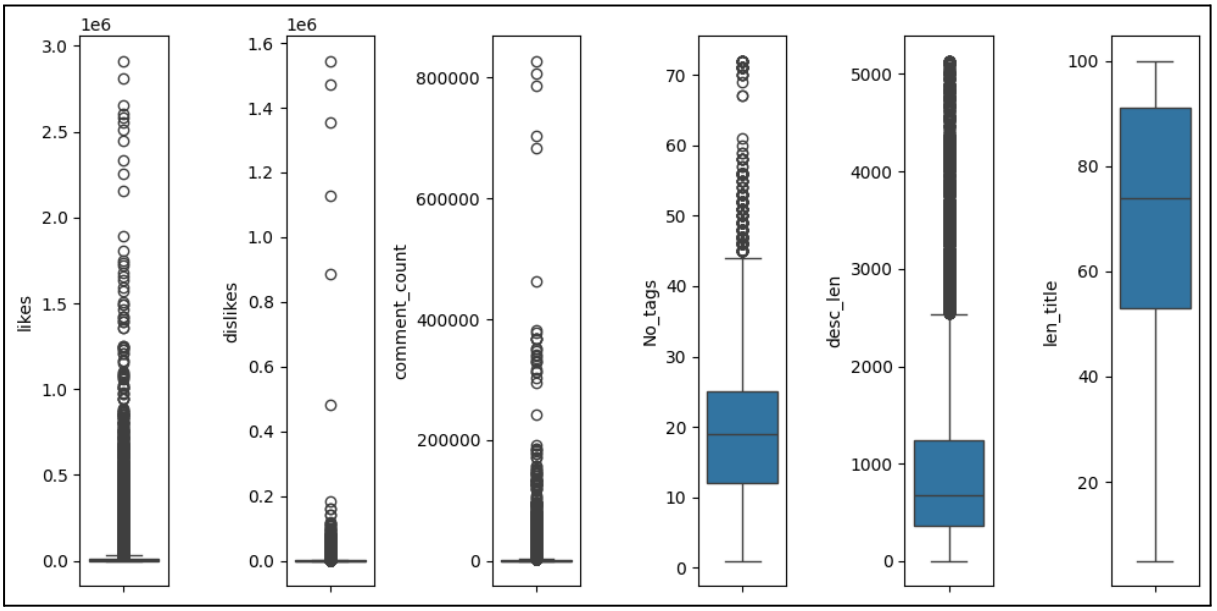
4. No_tags
Rata-rata jumlah tags per video adalah sekitar 18.94, dengan variasi sekitar 9.84. Video dengan jumlah tags paling sedikit memiliki 1 tag, sementara video dengan jumlah tags tertinggi memiliki 72 tags.

5. Desc_len
Panjang deskripsi rata-rata per video adalah sekitar 923.08 karakter, dengan variasi yang cukup besar (standar deviasi sekitar 815.04). Deskripsi terpendek memiliki 3 karakter, sementara deskripsi terpanjang mencapai 5,136 karakter.

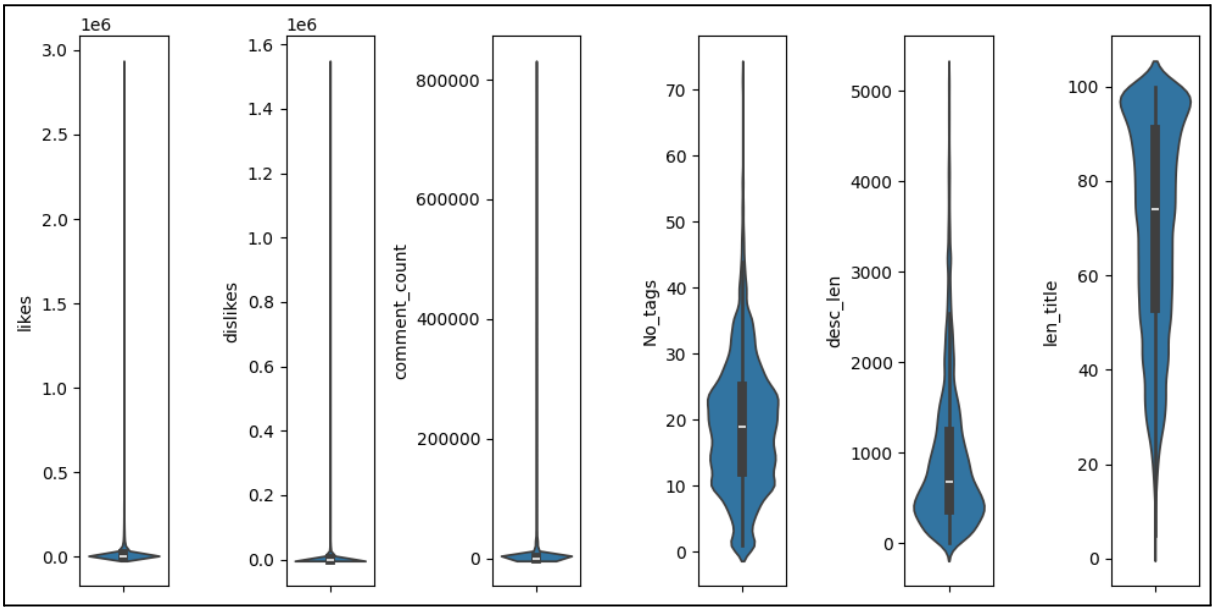
6. Len_title
Panjang judul rata-rata per video adalah sekitar 70.61 karakter, dengan variasi sekitar 22.41. Judul video terpendek memiliki 5 karakter, sementara judul video terpanjang memiliki 100 karakter.

Univariate Analysis

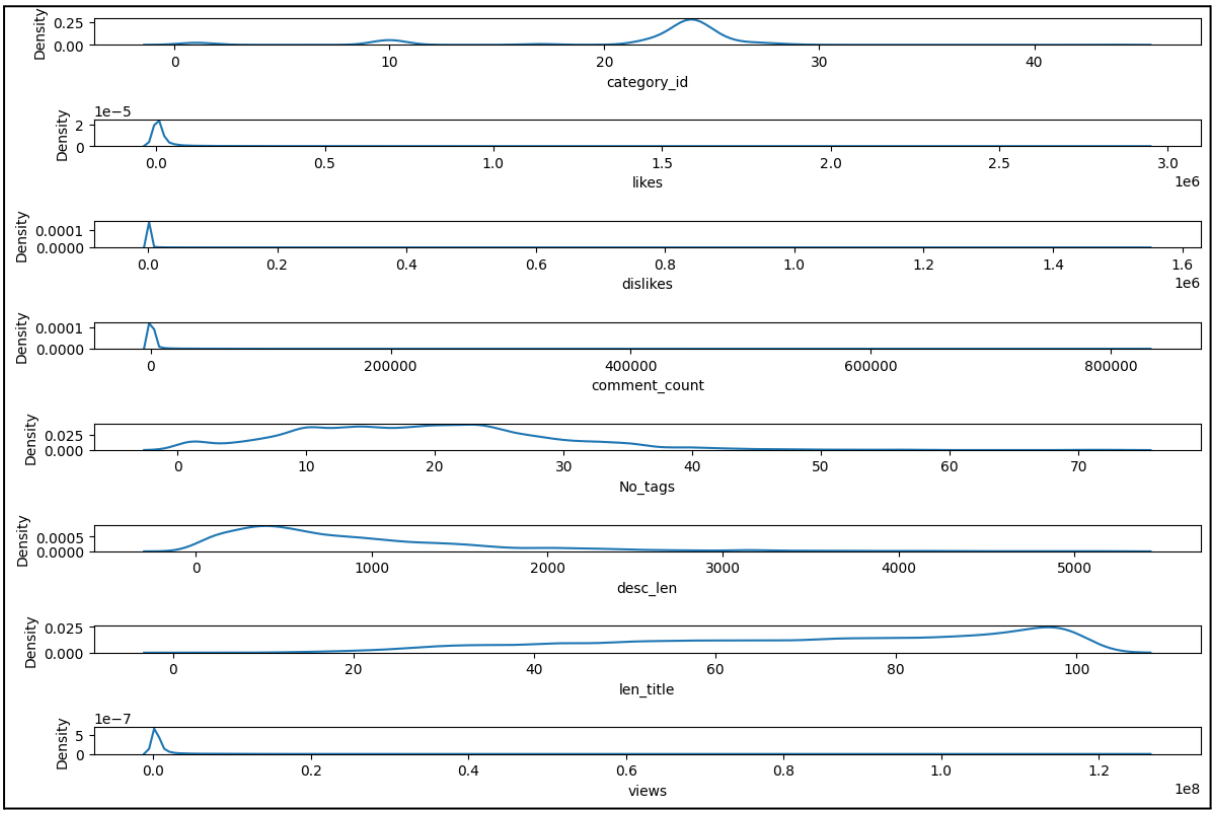
Boxplot



Violin Plot



Kdeplot



Dari kdeplot dapat ditarik beberapa kesimpulan sebagai berikut.

- Kolom likes, dislikes, comment_count memiliki jumlah outlier yg sangat besar.
- Kolom No_tags, desc_len cenderung positive skewed.
- Kolom len_title lebih ke negative skewed.

Multivariate Analysis

Preprocessing

a. Handling Missing Value

```
[4]: # Menangani missing value
     missing_value = df.isnull().sum()
     missing_value
```

```
[4]: 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
     comments_disabled         0
     ratings_disabled          0
     video_error_or_removed    0
     description              45
     No_tags                   0
     desc_len                  0
     len_title                 0
     publish_date              0
     dtype: int64
```

```
[5]: df['description'].fillna('Unknown', inplace=True)
     df.info()
```

Terdapat missing value pada feature description sebanyak 45 data, sehingga data tersebut tidak di drop namun diisi dengan ‘Unknown’. Karena data type pada feature description berupa object atau string

b. Handling Duplicates

```
[12]: df.duplicated().sum()
```

```
[12]: 4229
```

```
[13]: df = df.drop_duplicates()
      df
```

Terdapat sekitar 4229 data duplikat. Sehingga data duplikat tersebut harus di drop dan data yang akan di analisa ada sekitar 32562 data.

c. Handling Outliers

[15]:
```
from scipy import stats
```

[16]:
```
print(f'Jumlah baris sebelum memfilter outlier: {len(df)}')

filtered_entries = np.array([True] * len(df))

for nums in ['likes', 'dislikes', 'comment_count', 'No_tags', 'desc_len', 'len_title']:
    zscore = abs(stats.zscore(df[nums])) # hitung absolute z-scorenya
    filtered_entries = (zscore < 3) & filtered_entries # keep yang kurang dari 3 absolute z-scorenya

df = df[filtered_entries] # filter, cuma ambil yang z-scorenya dibawah 3

print(f'Jumlah baris setelah memfilter outlier: {len(df)}')
```

```
Jumlah baris sebelum memfilter outlier: 32562
Jumlah baris setelah memfilter outlier: 31084
```

Dilakukan pengecekan apakah adanya outliers pada feature yang dibutuhkan selain feature targetnya. Terdapat sekitar 1500 data outlier sehingga perlu di filter data yang akan digunakan.

d. Distribusi

Dilakukan pengecekan distribusinya dari setiap feature yang akan digunakan, apakah distribusinya normal atau skewed. Dan diketahui bahwa hanya feature No_tags yang distribusinya normal, sehingga feature lain perlu dilakukan log transformation.

2. Feature Engineering

Datetime Extraction

Kita dapat melakukan ekstraksi hari, bulan, dan tahun dari Date waktu trending dan waktu publish. Selain itu pada publish hour, kita dapat melakukan ekstraksi terhadap waktu jamnya saja dengan kategori pagi atau malam hari. Hal ini dilakukan dengan asumsi perbedaan menit dan detik tidak terlalu signifikan. Setelah melakukan ekstraksi, kolom lama akan dihapus karena bertipe datetime dan sudah tidak dipakai lagi

Ranking Channel Title

Channel title merupakan column kategorikal yang harus di preprocess untuk menjadi numerik agar bisa dimengerti oleh model. Terdapat 1390 jenis channel title sehingga one hot encodding tidak bisa dilakukan karena akan membuat dimensi fitur menjadi sangat besar. Oleh karena itu feature engineering akan dilakukan dengan menggunakan label encoding berdasarkan urutan rata-rata views tiap channel dari yang terbesar hingga terkecil.

Hasil Korelasi

Ternyata hour, day, month, dan year tidak ada yang memiliki signifikan yang tinggi terhadap views. Maka dari itu kolom tersebut akan di drop agar dimensi tidak terlalu besar

Scallling

Scalling akan dilakukan agar tidak ada feature yang lebih berpengaruh daripada yang lain. Berikut adalah hasil akhir dari feature engineering

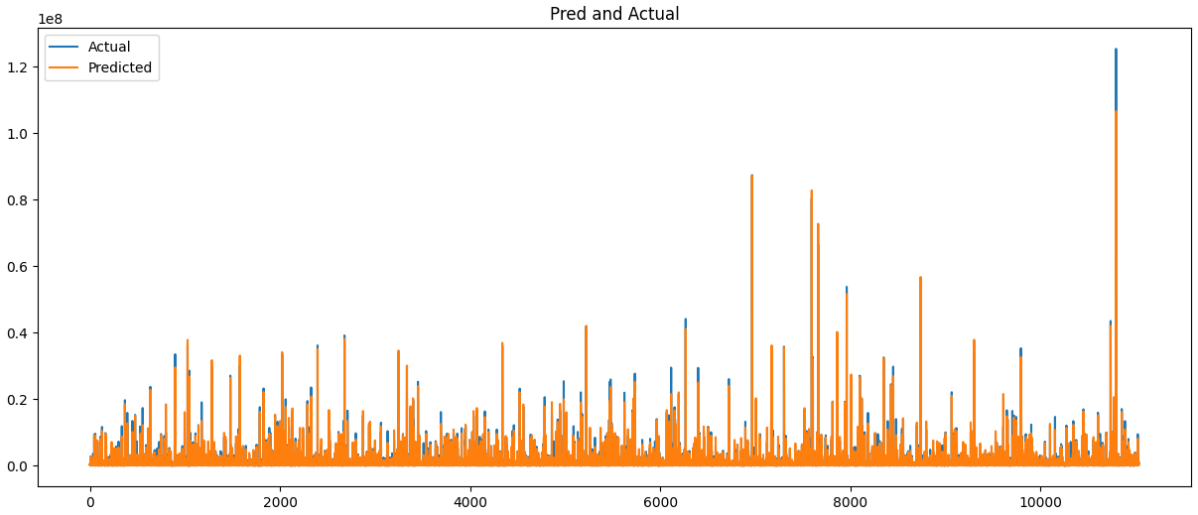
3. Modeling

	Model Supervised			
	Linear Regression	Decision Tree	Random Forest	SVR
RMSE(test)	1599975	728368	631312	3342507.66719669
RMSE(train)	1553522	10929	235124	3277677.8276602305
R2(test)	0.56	0.74	0.83	-0.0708937522206587
R2(train)	0.76	0.99	0.99	-0.057411651959714316

Dari tabel diatas dapat terlihat bahwa model supervised yang lebih baik digunakan adalah **Random forest**, sebab model tersebut memiliki nilai selisih RMSE test dan trainnya cukup kecil sedangkan untuk nilai R2 selisihnya adalah yang cukup besar.

4. Evaluation

Dari hasil modelling yang dilakukan sebelumnya, kami sepakat untuk menggunakan model **Random Forest**. Random Forest terbukti memiliki hasil yang terbaik. Hal tersebut dapat dilihat dari hasil r2-train dan r2-test berturut-turut sebesar 99% dan 84%. Artinya fitur memiliki kemampuan yang sangat baik dalam mempengaruhi target.



Dari plot yang ditampilkan, dapat dilihat bahwa jarak antara data yang diprediksi dengan data aktual cukup dekat. Artinya model sudah cukup baik dalam melakukan prediksi terhadap jumlah views (target).

Adapun fitur yang memiliki pengaruh paling signifikan terhadap target adalah likes dan channel_title, dengan nilai importance masing-masing sebesar 58% dan 13%. Oleh karena itu, rekomendasi yang dapat kami berikan kepada perusahaan agar mendapat jumlah views yang banyak adalah dengan cara meng-*encourage* penonton untuk memberikan like sebanyak mungkin. Serta memilih channel youtube yang sekiranya dapat memberikan banyak views.