Homework Solution - Regression

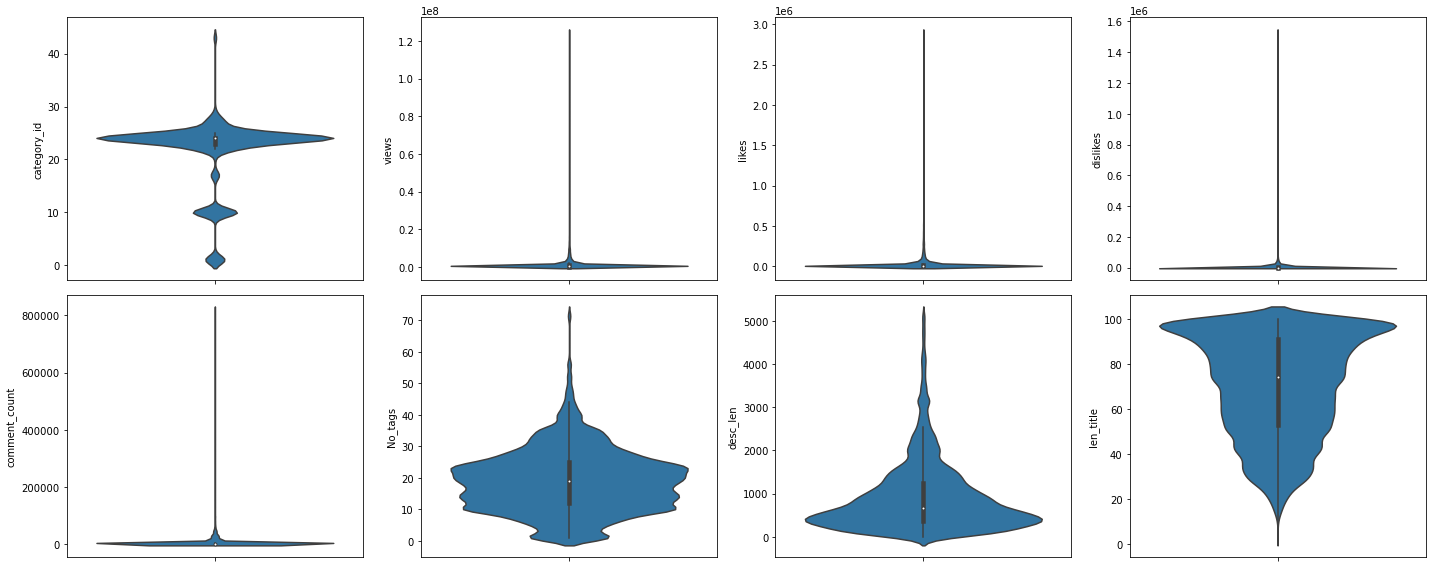
Team:

1. Attariq Muhammad Kasfilla
2. Nizar Mochammad Yusuf

# Exploratory Data Analysis

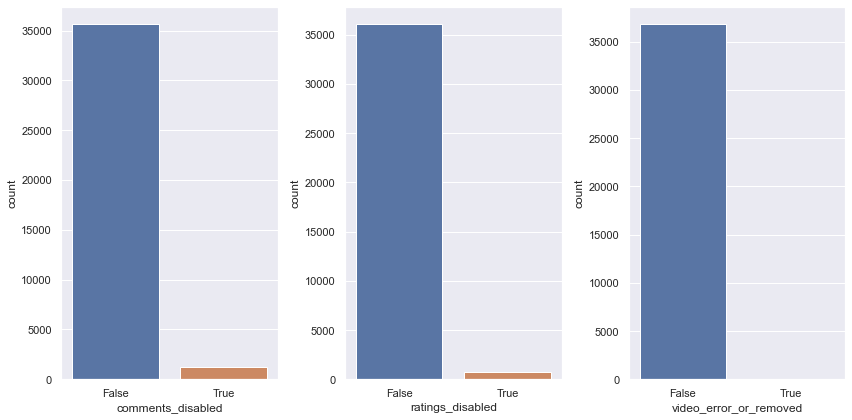
1. Persebebaran Distribusi Pada Dataset

Berikut merupakan distibusi data numerik dengan menggunakan sns.violinplot

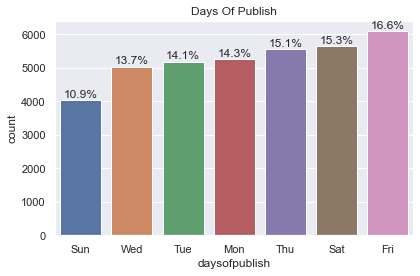


# Dapat dilihat bahwa distibusi pada dataset memiliki banyak ouliers didalamnya yang harus di prepocessing terlebih dahulu sebelum diolah.

1. Distribusi Data Tipe ‘*bool’*

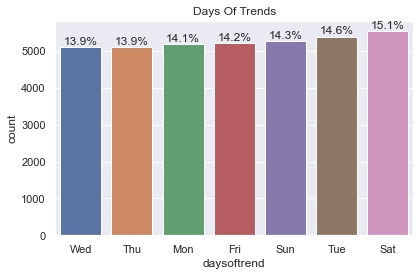


1. Distirbusi Data Day Of Publish



Mayoritas para *youtubers* mengunggah video mereka pada hari jumat sebanyak 16.6%. Hal ini mungkin dikarenakan banyak dari views mereka yang aktif menonton video nya pada hari jumat sampai sabtu karena pada hari tersebut mereka memiliki waktu senggang lebih untuk menonton.

1. Distribusi Data Day Of Trends



Distribusi terbanyak *day of trends* pada hari sabtu mungkin karena para views atau para penonton lebiih cenderung untuk menonton konten mereka pada hari tersebut.

1. Distribusi Data No\_tags

Count 36791.000000

mean 18.938463

std 9.843531

min 1.000000

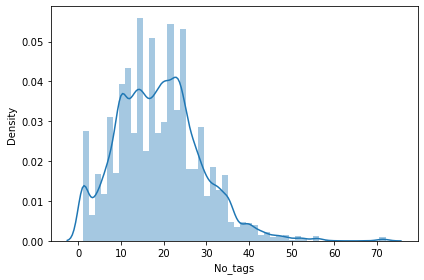
25% 12.000000

50% 19.000000

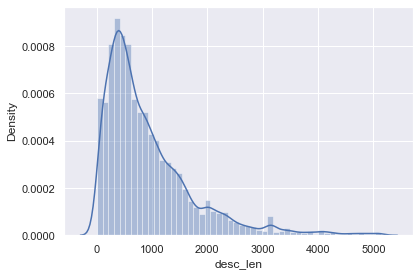
75% 25.000000

max 72.000000

Name: No\_tags, dtype: float64

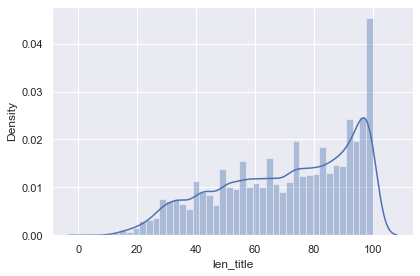
No\_tags memiliki persebaran rata-rata 18 tags dan median 19 tags. Hal ini mungkin karena jumlah tags yang mereka mempengaruhi pada pencarian video oleh para views mereka.

1. Distribusi Data desc\_len



Desc\_len memiliki rata-rata 923 huruf dan median 677 huruf pada deskripsi yang dicantumkan pada konten video mereka.

1. Distribusi Data len\_title



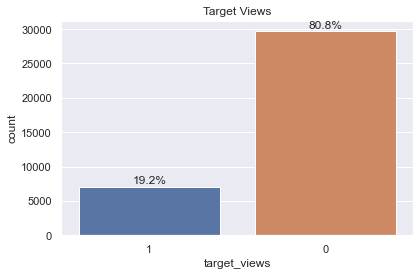
Jumlah huruf judul video yang mereka cantumkan rata-rata berjumlah 70 huruf sedangkan mediannya berjumlah 74 huruf. Jumlah huruf yang mereka cantumkan mungkin berdasarkan konten video yang mereka tampilkan.

1. Distribusi Data views dan target\_views



Views mereka rata-rata berjumlah kurang lebih 1juta views sedangkan mediannya berjumlah 300 ribu views. Jumlah views yang mereka hasilkan mungkin diakibatkan oleh faktor video mereka menarik atau tidak.

Berikut merupakan distribusi target views yang diambil berdasarkan rata-rata keseluruhan populasi views yang mereka peroleh.



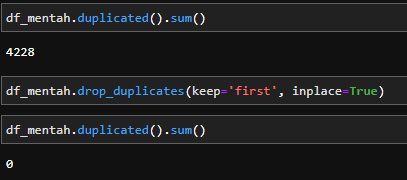
80.8% konten mereka tidak mencapai taget views dari keseluruhan populasi dan hanya 19.2% saja dari keseluhuran konten yang memenuhi target views.

# Data Preprocessing

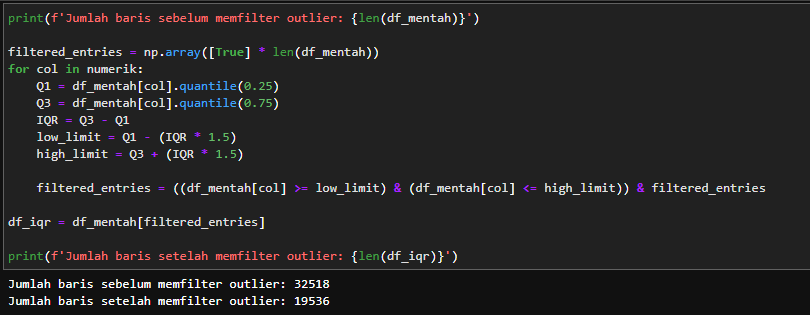
* Missing Data



* Duplicated



* Outliers

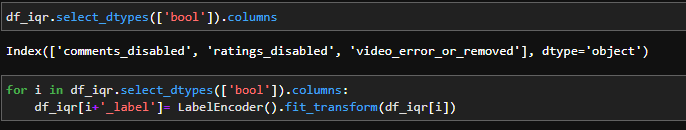


# Feature Engineering

* Menambahkan Features Baru yang Akan Digunakan
* Disini kita menambahkan features ‘*target\_views*’ dengan berdasarkan data ‘*views*’ mana saya yang melebihi median diberi index 1 sedangkan data yang dibawah median diberi index 0.

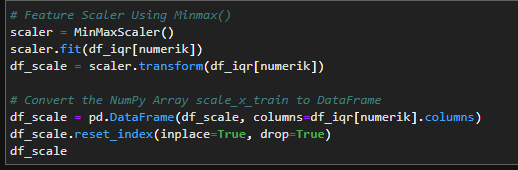


* Selanjutnya mengubah data bertipe ‘*bool*’ menjadi data denga tipe numerik agar dapat dilakukannya proses modeling pada data tersebut.



* Normalization

Kita melakukan proses *normalization* karena data tersebut tidak memenuhi syarat untuk dilakukannya *standarization*. Proses *normalization* dilakukan dengan *MinMaxScaler*.

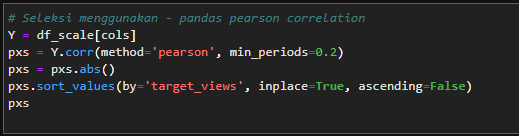


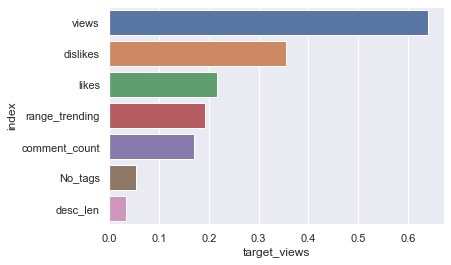
* Imbalance Class

Imbalance Class yang kita gunakan dengan cara *SMOTE.*

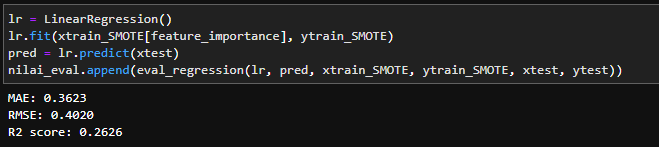
* Feature Importance

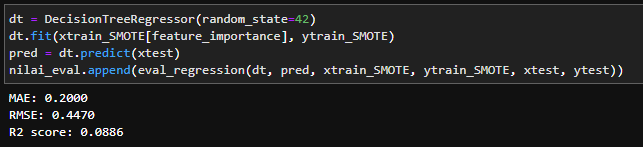
Kita menseleksi feature apasaja yang akan digunakan untuk proses selanjutnya dengan metode *pearson correlation*.



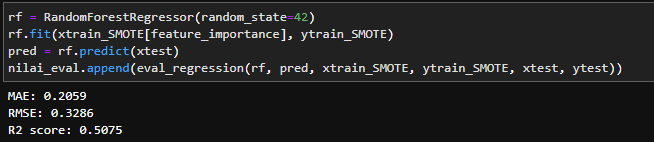


# Modeling

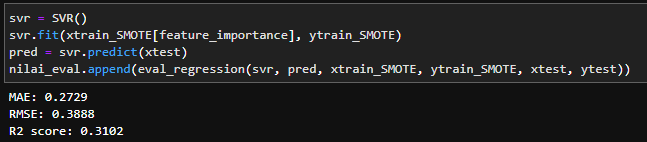
* Linear Regression
* Decision Tree



* Random Forest

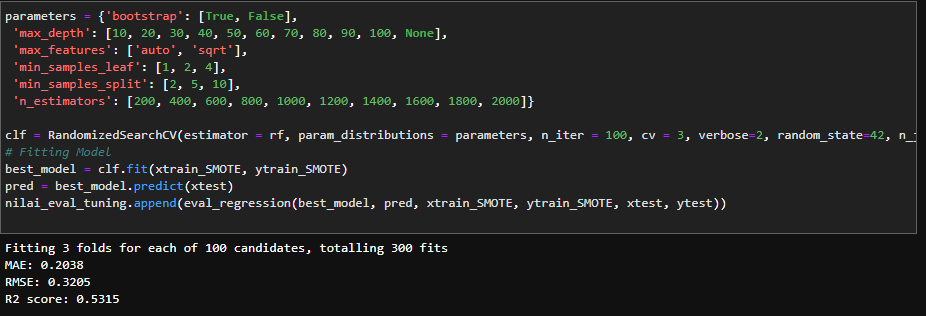


* Support Vector Regressor



Berdasarkan keempat algoritma yang jalankan, algoritma terbaik berdasarkan RMSE dan R2 adalah Random Forest karena RMSE yang dihasilkan 0.32 lebih kecil dibandingkan dengan ketiga lainnya dan R2 yang dihasilkan 0.50 lebih besar dibandingkan dengan algoritma yang lainnya. Hasil tersebut didasari karena bila RMSE nilainya mendekati 0 maka feature tersebut minim kesalahan dan R2 nilainya harus mendekati 1 karena feature tersebut akan semakin akurat.

Setelah dilakukannya tunning pada Random Forest.



# Summary

Kesimpulan yang kita dapat berdasarkan insight, feature dan machine learning bahwa feature *dislikes, likes, range\_trending, comment\_count, No\_tags* dan *desc\_len* berhubungan langsung dengan target yaitu *target\_views/views.* Rekomendasi bisnis yang kita sarankan pada para youtubers adalah dengan cara membuat konten negatif atau membuat konten berita viral yang membuat para penonton semakin penasaran dan akan membuat feature tersebut semakin tinggi.

# Appendix

Tuliskan kontribusi pengerjaan masing-masing anggota tim di bagian ini

* Nizar : bisnis insight dan EDA
* Attariq : teknikal percodingan

Tuliskan juga kesulitan-kesulitan dalam pengerjaan tugas ini di bagian ini (jika ada)

* Teknik coding pada hyperparametic algoritma yang digunakan
* Kekurangan sudut pandang pada insight