**ETS KAPITA SELEKTA ILMU KOMPUTASI**

**Menganalisa Data**



**Disusun oleh:**

**Adrian Maulana Muhammad** **(06111540000099)**

**Departemen Matematika**

**Fakultas Sains dan Analitika Data**

**Institut Teknologi Sepuluh Nopember**

**Surabaya**

**2020**

1. **Analisa Data**

Analisa data adalah sebuah proses untuk memeriksa, membersihkan, mengubah, dan membuat permodelan data dengan maksud untuk menemukan informasi atau permasalahan yang berguna sehingga dapat memberikan petunjuk bagi peniliti untuk mengambil keputusan. Secara umum, tujuan analisis data adalah untuk menjelaskan suatu data agar lebih mudah dipahami, kemudian dibuat kesimpulan.

1. **Langkah-langkah dari Analisa Data**

* **Pengumpulan Data**

Tahap awal kegiatan analisa data adalah pengumpulan data untuk dianalisa.

* **Tahap Penyuntingan**

Proses pengecekan kejelasan dan kelengkapan terkait pengisian instrument pengumpulan data.

* **Tahap Pengkodean**

Proses identifikasi dan klasifikasi semua pernyataan pada instrument untuk mengumpulkan data berdasarkan variabel yang sedang dipelajari.

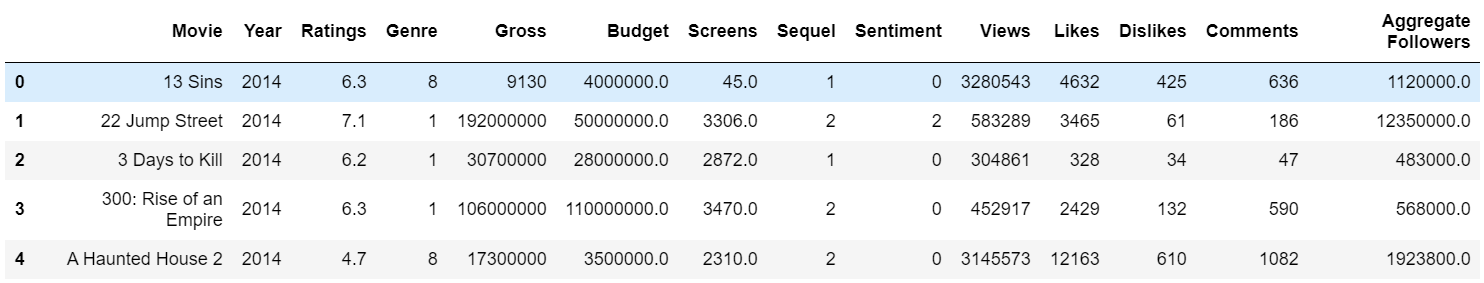
* **Tahap Mendeskripsi Data**

Proses menggambarkan data dengan menyajikannya dalam bentuk table frekuensi atau diagram dengan berbagai ukurannya. Tujuannya adalah untuk memahami karakteristik data.

* **Tahap Pengujian Hipotesis**

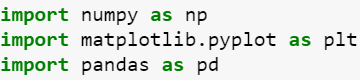
Proses pengujian proposisi apakah dapat diterima atau ditolak, apakah memiliki makna atau tidak. Berdasarkan tahap ini, nanti kesimpulan akan dibuat.

1. **Importing Data**

Data yang digunakan adalah bertipe *xlsx* yang merupakan list dari judul film Hollywood dari tahun 2014-2015 beserta fiturnya, seperti *Ratings, Genre, Gross, Budget, Screens, Sequel, Sentiment, Views, Likes, Dislikes, Comments*, dan *Aggregate Followers.* Berikut 4 baris awal dari data:

**II. Importing The Libraries**

Terdapat beberapa langkah pada proses ini. Langkah yang pertama adalah *Importing the Libraries.*



Library dalam konteks computer programing bisa diibaratkan menjadi sebuah kotak perkakas yang isinya adalah kumpulan perkakas yang bisa anda gunakan untuk membantu anda dalam mengerjakan projek. Dan yang menarik pada penggunaan library adalah anda hanya perlu memberikan input saja, library akan mengerjakannya dan menghasilkan output. Library yang digunakan pada projek ini adalah Numpy, Matplotlib, dan Pandas. Numpy akan membantu kita dalam konteks mathematical processing, Matplotlib akan membantu kita membuat plotting pada chart, dan Pandas akan membantu kita dalam hal import dan manage data pada projek ini.

**III. Data Characteristics**

Untuk mendapatkan ukuran dari data:



Didapatkan ukuran dari data:



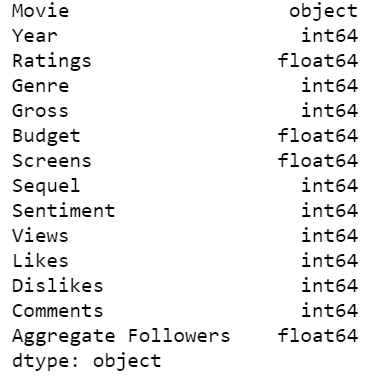
Ukuran dari data adalah terdapat 234 baris dan 14 kolom. Lalu, untuk mendapatkan dimensi dari data:



Didapatkan dimensi dari data adalah dua, yaitu baris (axis = 0) dan kolom (axis = 1). Selanjutnya, untuk mendapatkan *data types* pada elemen per kolom:



Didapatkan *data type*s per kolom:



**IV. Data Preprocessing**

*Data Preprocessing* adalah langkah-langkah dimana data awal mengalami proses *transformed* dan *encoded*, dimana hal tersebut bertujuan untuk menyiapkan data untuk proses selanjutnya. Langkah-langkahnya:

* Penanganan *Missing Values*
* Penanganan *Outliers*
* Penanganan *Duplicate Values*

1. **Missing Values**

Sering terjadi terdapat elemen berupa *null* (merepresentasikan *no value*) atau *NaN* (merepresentasikan “Not a Number”) pada data. Hal ini biasa terjadi pada proses pengumpulan data atau validasi data. Penanganan *missing values* diperlukan karena jika terdapat data yang hilang akan berpengaruh terhadap *noise* pada data. Terdapat dua jenis penanganan:

* **Eliminasi Baris pada Data yang Hilang**

Merupakan strategi yang simple dan terkadang efektif. Tidak dapat diterapkan jika mempunyai banyak *missing values*. Jika suatu *feature* mempunyai elemen yang sebagian besar adalah *missing values,* maka *feature* tersebut dapat dieliminasi.

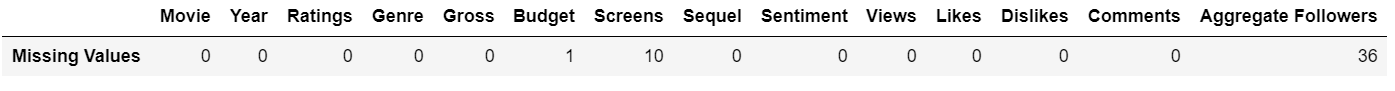
* **Estimasi dari Data yang Hilang**

Jika terdapat beberapa persen *missing values,* kita dapat melakukan metode interpolasi untuk mengisi data yang hilang tersebut. Metode yang biasa digunakan adalah mengisi data yang hilang dengan *mean, median,* atau *mode.*

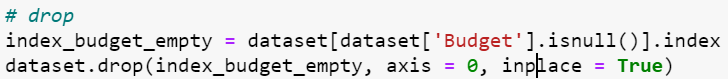
Pertama kita akan cek apakah ada *missing value* pada baris tiap kolom.



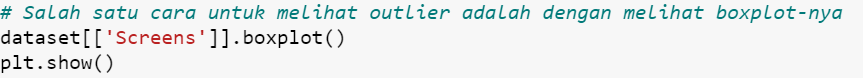
Didapatkan:



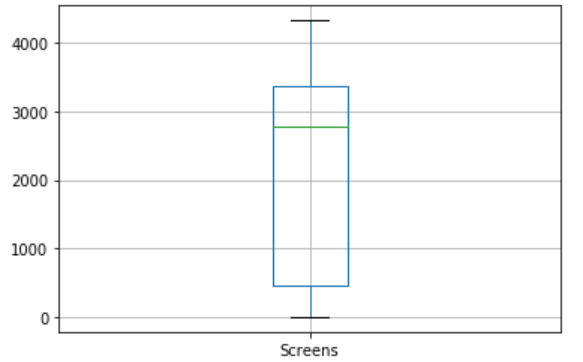
Dari *Data Frame* di atas bisa dilihat bahwa *feature* yang mempunyai *missing value* adalah *budget, screens*, dan *aggregate followers*. *Missing value* pada *feature budget* hanya 1 oleh karena itu, kita bisa eliminasi missing value tersebut.



Terlihat bahwa *missing value* pada *feature screens* dan *aggregate followers* cukup banyak, oleh karena itu kita tidak bisa melakukan eliminasi. Untuk mengisi data kosong, kita harus mengetahui apakah terdapat *outlier*. Jika ada, maka kita bisa mengisi data kosong tersebut dengan *median*, karena nilai *median* tidak berpengaruh terhadap *outlier*. Jika tidak ada, maka bisa kita isi dengan nilai *mean*. Untuk mengetahui apakah ada outlier atau tidak, kita menggunakan *boxplot*. *Boxplot* dapat mengukur seberapa baik data terdistribusi, ia membagi data menjadi tiga kuartil. Grafiknya merepresentasikan nilai minimum, maksimum, median, kuartil pertama, dan kuartil ketiga. Kita akan melihat *boxplot* dari *feature screens*.



Didapatkan:



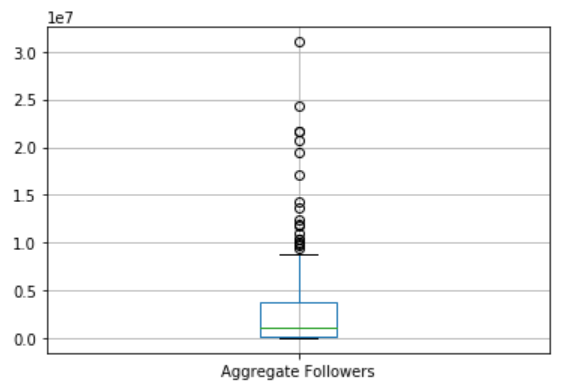
Bisa dilihat bahwa data tidak mempunyai nilai di atas kuartil ketiga, maka tidak mempunyai *outliers*. Oleh karena itu kita bisa mengisi missing value dengan nilai *mean.*



Lalu, kita akan melihat *boxplot* dari *feature aggregate followers*.



Didapatkan:



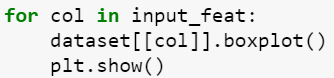
Bisa dilihat bahwa terdapat banyak data di atas kuartil ketiga, maka terdapat banyak *outliers.* Oleh karena itu kita akan mengisi *missing value* pada *feature aggregate followers* dengan nilai median.



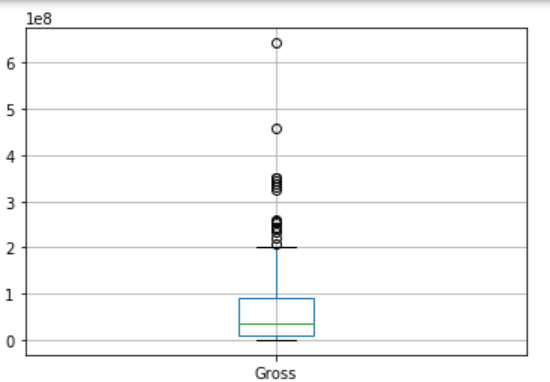
1. **Outliers**

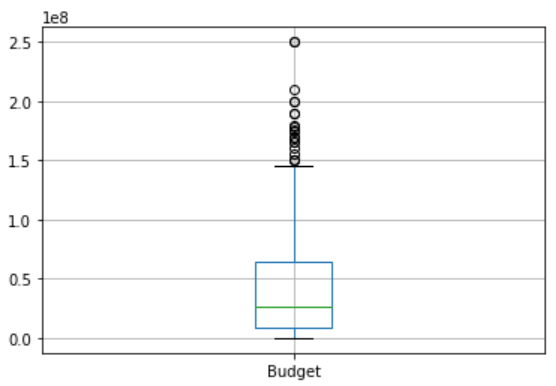
*Outliers* adalah titik data yang terletak jauh di antara titik-titik data yang lain. Hal tersebut bisa terjadi karena variasi pada pengukuran atau error pada percobaan. Jika memungkinkan, outliers bisa disingkirkan dari data set. Kita akan mengecek grafik *boxplot* pada setiap *feature*.

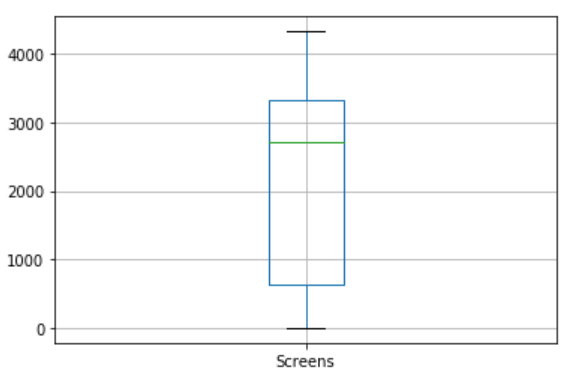


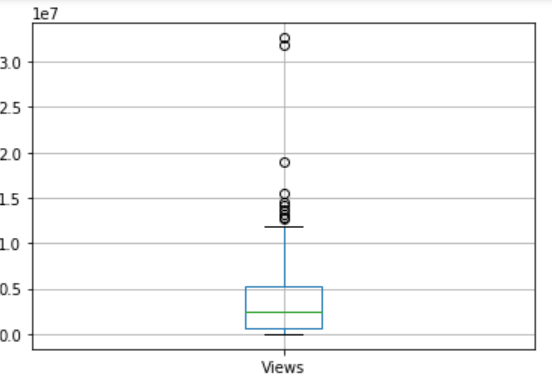


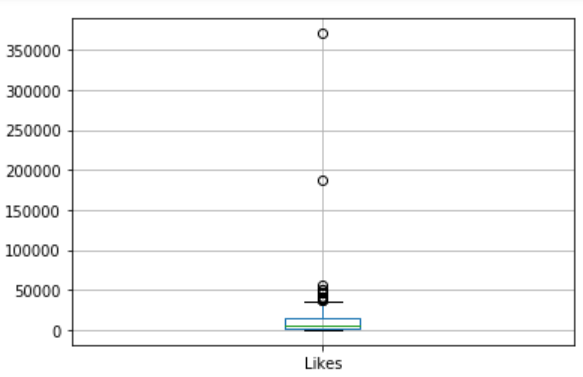
Didapatkan:

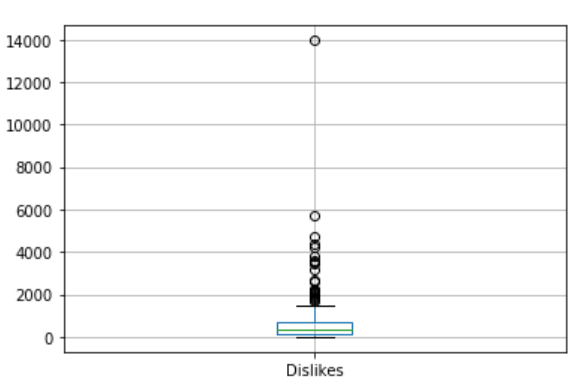


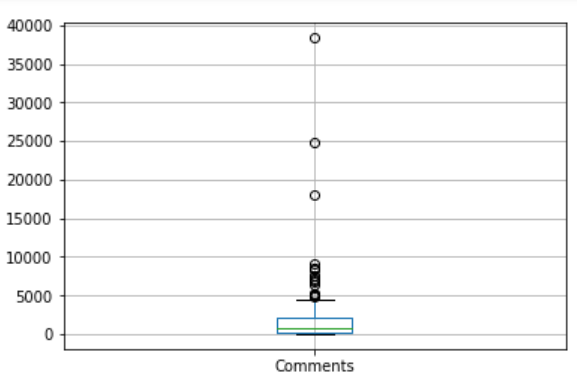


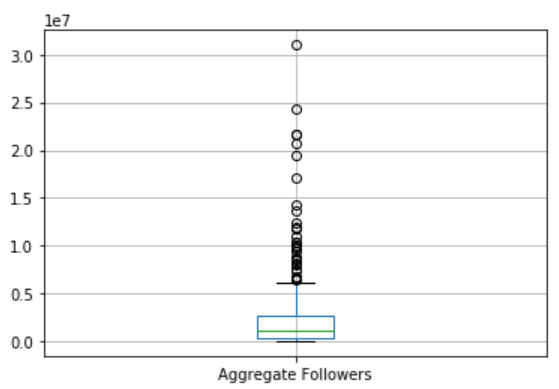




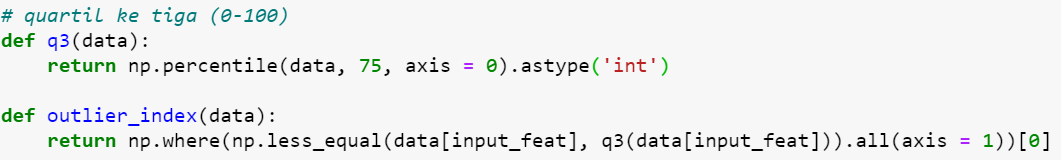








Bisa dilihat dari grafik-grafik *boxplot* di atas, bahwa hampir semua *feature* (kecuali *screens*) mempunyai *outliers*. Untuk itu, kita akan singkirkan *outliers* tersebut.





Setelah menghilangkan *outliers,* kita lihat ukuran dari data.



Didapatkan:



Proses-proses di atas mengurangi baris dari 234 menjadi 142.

1. **Data Duplikat**

Suatu data set memungkinkan untuk mempunyai nilai data yang duplikat, hal ini bisa terjadi disebabkan error saat pengumpulan data. Secara umum, data duplikat disingkirkan dari data, agar tidak memberikan *bias*.



Kita akan coba lihat ukuran data setelah menyingkirkan data duplikat.



Didapatkan:



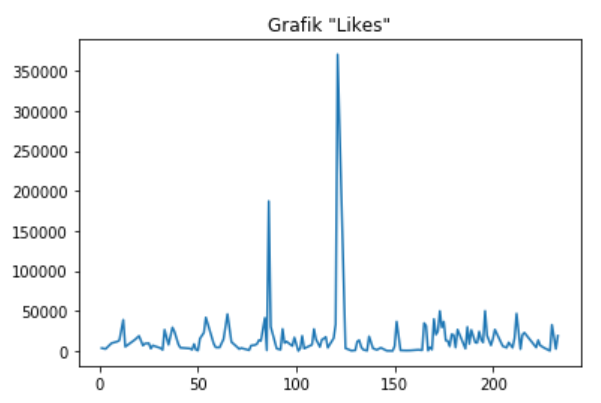
Dapat dilihat bahwa ukuran data tidak berubah, maka tidak ada data duplikat.

V. **Analisa Data**

Sering kali kita melihat kualitas suatu film hanya pada suatu sudut pandang, dalam hal ini kita akan membandingkan dengan memperlihatkan “likes” dan “dislikes”.



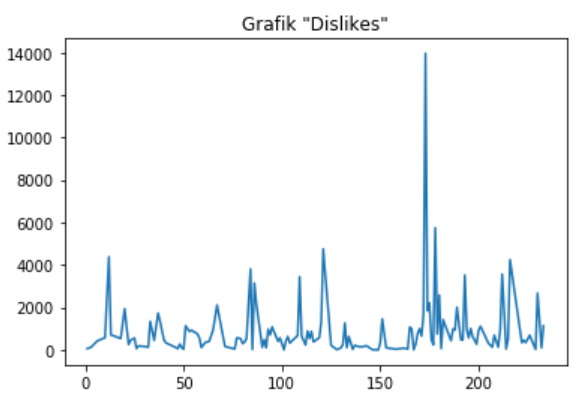
Didapatkan:



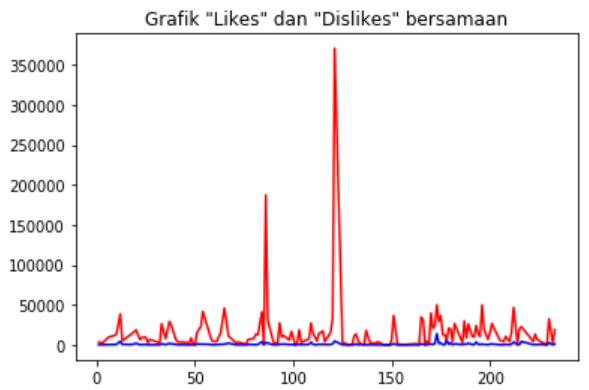
Lalu untuk mendapatkan grafik “dislikes”.



Didapatkan:



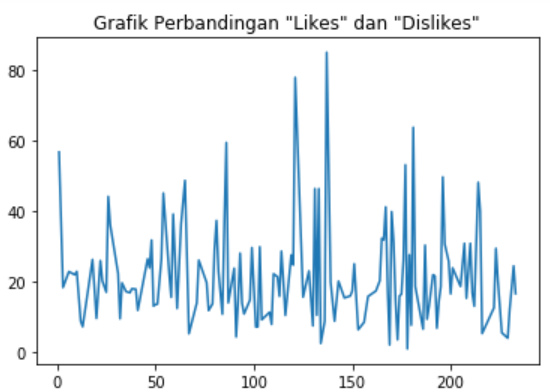
Lalu kita buat grafik “likes” dan “dislikes” secara bersamaan.

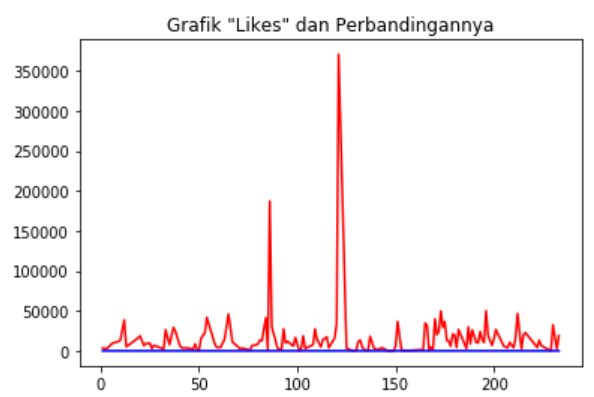


Selanjutnya, kita buat perbandingan pada “likes” dan “dislikes”.



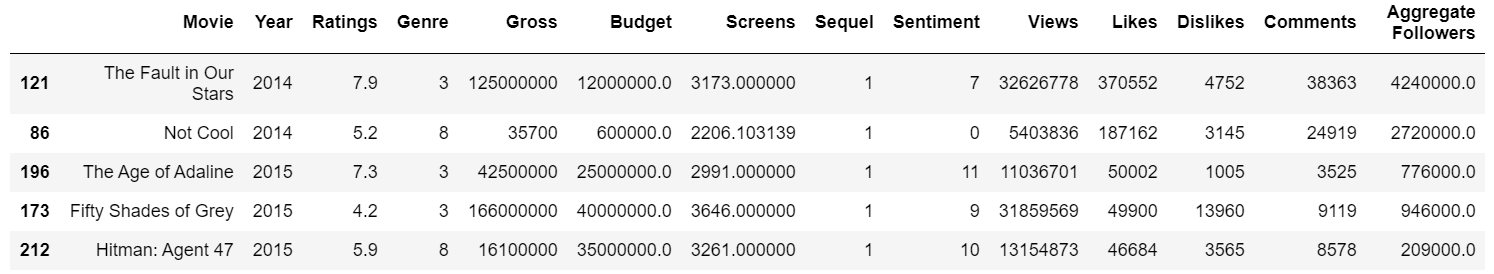
Didapatkan:





Film terbaik menurut “likes”:





Movie terbaik menurut perbandingan “like” dan “dislikes”:

