**“Clustering Hirarki Facebook Live Sellers in Thailand Data Set”**

**Pembelajaran Mesin**



**DISUSUN OLEH:**

|  |  |
| --- | --- |
| 11418003 | Miranti Sinaga |
| 11418007 | Valentine Trihandayani |
| 11418025  11418035  11418047 | Elsa Pitalita Sihombing  Frans Z Siregar  Medianto Saragih |
|  |  |

PROGRAM STUDI DIV TEKNOLOGI

REKAYASA PERANGKAT LUNAK

FAKULTAS INFORMATIKA DAN ELEKTRO

T.A 2020/2021

Institut Teknologi Del

**DAFTAR ISI**

[**BAB I DASAR TEORI** 3](#_Toc60427112)

[**BAB II PREPROCESSING DATA** 4](#_Toc60427113)

[**BAB III EKSPLORASI DATA** 6](#_Toc60427114)

[**BAB IV DESIGN EKSPERIMEN** 10](#_Toc60427115)

[**4.1 Algoritma Clustering** 10](#_Toc60427116)

[**4.2 K-Means** 14](#_Toc60427117)

[**BAB V CODE** 16](#_Toc60427118)

[**BAB VI ANALISIS OUTPUT DAN HASIL** 30](#_Toc60427119)

[**REFERENSI** 36](#_Toc60427120)

# **BAB I DASAR TEORI**

Clustering adalah proses pengelompokan kumpulan dari objek-objek yang anggotanya memiliki kesamaan menjadi class-class dari objek yang sama dan merupakan algoritma dari unsupervised learning. Salah satu metode pada clustering adalah Hierarchical Clustering Method. Dengan menggunakan metode ini, penyajian data facebook live seller dalam jumlah yang besar dapat lebih mudah dipahami. Oleh karena itu, tujuan utama adalah melakukan implementasi clustering pada data pengguna facebook untuk menjelaskan facebook live terhadap pola, struktur statistic, dan tingkat keterlibatan pengguna serta viralitas kiriman. Dimana secara khusus video secara langsung menghasilkan banyak komentar, reaksi emosional, share, dan waktu yang terlibat dalam kognitif, afektif, dan perilaku yang lebih tinggi dengan seller dibandingan dengan jenis posting facebook lainnya.

Metoda Hierarchial clustering diklasifikasikan dalam dua jenis yaitu Agglomerative dan Divisible Hierarchial Clustering tergantung pada komposisi hirarkis dilakukan dari bentuk bottom-up atau top-up. Metode yang digunakan pada kajian ini adalah Metoda agglomerative Hierarchical Clustering pada metode linkage. Metode ini dilakukan dengan serangkaian pengelompokan objek (banyak objek menjadi satu bagian). Dimulai dengan menempatkan setiap objek pada clustering dan kemudian menggabungkan struktur clustering tersebut menjadi clustering yang lebih besar sampai semua objek berada pada satu clustering yang sama atau sampai memenuhi kondisi yang telah ditentukan. Metode linkage terdiri dari 3 yaitu single linkage (mencari jarak terkecil), complete linkage (mencari jarak terjauh), dan average linkage (jarak rata-rata). Hasil-hasil dari metode agglomerative maupun metode divivise ditampilkan dalam bentuk diagram yang disebut dendrogram yang menggambarkan penggabungan atau pembagian yang akan dibuat pada tingkat yang berurutan.

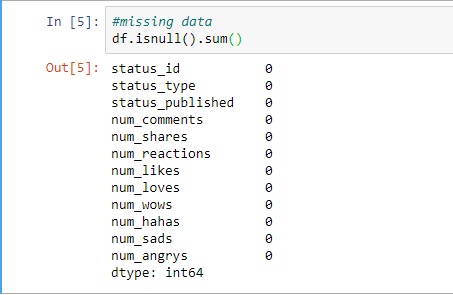
# **BAB II PREPROCESSING DATA**

Pada bab ini menjelaskan beberapa cara untuk preprocessing data dalam machine learning. Preprocessing data merupakan merupakan tahapan cleaning yang bertujuan membuang duplikasi data, memeriksa data yang inkonsisten serta memperbaiki kesalahan pada data sebelum data mining dilakukan. Berikut beberapa cara dalam preprocessing data:

1. Missing data

Missing data merupakan kehilangan data baik sepenuhnya maupun secara acak yang terjadi didalam dataset. Tujuan dari missing data dalam dataset membuat data analisa data menjadi rapi dan harus ditangani sebelum dataset diproses dengan machine learning.

Untuk mencari dan menangani missing data, terdapat dalam library phyton pandas pada fungsi isnull(), isna(). Seperti code dibawah ini fungsi ini digunakan untuk mengembalikan nilai boolean.



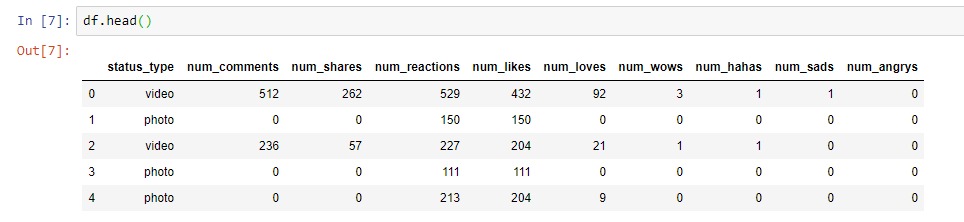
1. Normalisasi

Penggunaan normalisasi data digunakan jika terdapat data yang terlalu besar sampai terlalu kecil, yang bertujuan untuk menghindari dimensi data yang terlalu besar atau terlalu kecil. Proses penskalaan nilai atribut dari suatu data terletak pada rentang skala tertentu. Dari data yang telah dikumpulkan itu, akan dilakukan normalisasi data dengan menggunakan min-max normalization. Pada data Live Sellers in Thailand Data Set tidak terjadi normalisasi data.

1. Outlier

Outlier merupakan data yang berada diluar yang mengganggu perhitungan data. Untuk mengatasi itu, terdapat cara membuang, memberikan tanda, atau memberiarkan saya. Tetapi pada penggunaan code dibawah ini dilakukan dengan cara mendrop data status id, status published.

C:\Users\Valentina Sinaga\Downloads\WhatsApp Image 2021-01-01 at 15.10.04.jpeg

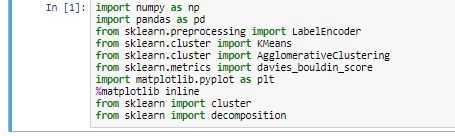


# **BAB III EKSPLORASI DATA**

Pada bab ini menjelaskan beberapa eksplorasi data yang digunakan untuk menjelaskan isi dan komponen penyusun data. Eksplorasi data dilakukan setelah tahap preprocessing dan sebelum melakukan proses feature engineering dan modeling.

Berikut eksplorasi data yang dilakukan yaitu sebagai berikut:

1. Library awal atau yang biasa digunakan untuk melakukan ekplorasi data dengan phyton, yaitu pandas untuk data frame, numpy untuk array, matplotlib untuk menciptakan visualisasi dan sklearn.

. 

1. Selanjutnya untuk membaca dan menampilkan data 5 teratas dataset Live Sellers in Thailand Data Set digunakan syntax dibawah ini.



1. Untuk mengecek atau melihat dataset terdapat fungsi sebagai berikut:



fungsi describe bertujuan untuk menunjukkan rangkuman statistic seperti rata-rata, median, pada kolom.



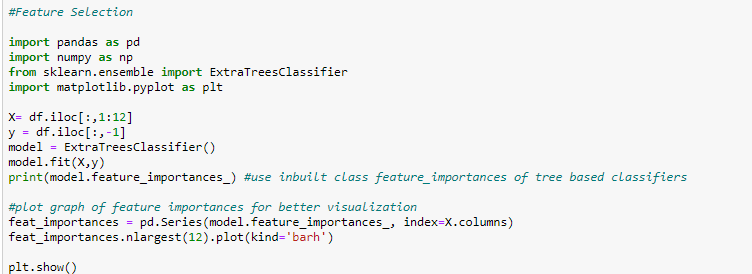
fungsi info() digunakan untuk melihat nomor index beserta tipe datanya





fungsi head() berfungsi untuk melihat data secara random.

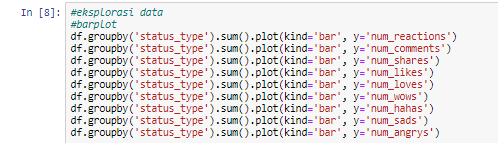
1. Kemudian, perintah untuk memilih baris dan kolom, dimana pada index yang berisi “:” dapat memilih semua urutan pada index, yang digunakan yaitu iloc() untuk memilih data berdasarkan index.



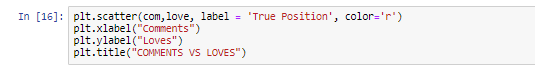
1. Selanjutnya data cleaning isnull() yang bertujuan mengecek apakah ada nilai null dengan keluaran Boolean.



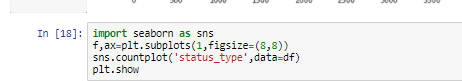
1. Visualisasi data dialkukan untuk melihat bentuk data dengan lebih jelas dengan cara memvisualisasikan, dengan melihat distribusi masing-masing atribut dengan menampilkan dalam bentuk boxplot.

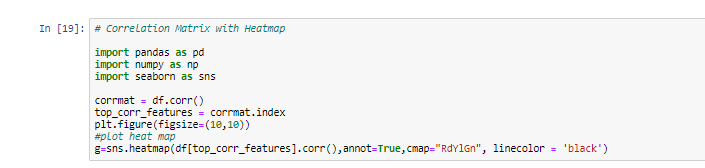


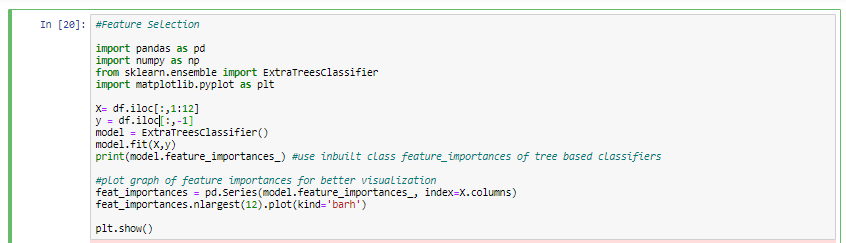












Kemudian, dendrogram bertujuan menampilkan pengelompokkan data dari penggabungan maupun pemisahan yang ditentukan secara acak dan hasil dari hierarchical clustering, seperti contoh pada code program dibawah ini.





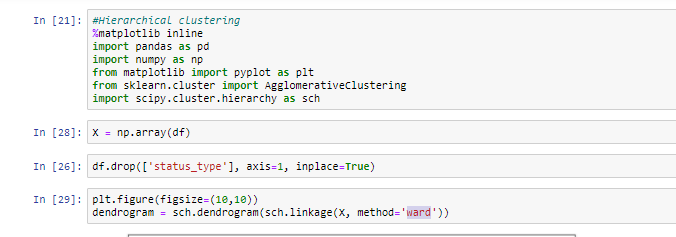




# **BAB IV DESIGN EKSPERIMEN**

Pada bab ini menjelaskan design eksperimen clustering suatu studi yang melibatkan keterlibatan memanipulasi beberapa variabel, mengamati, dan mengobservasi. Pada design eksperimen dibedakan menjadi 2 dimana design dari non hirarki yaitu kmeans sedangkan hirarki yaitu clustering hirarki

# **4.1 Algoritma Clustering**

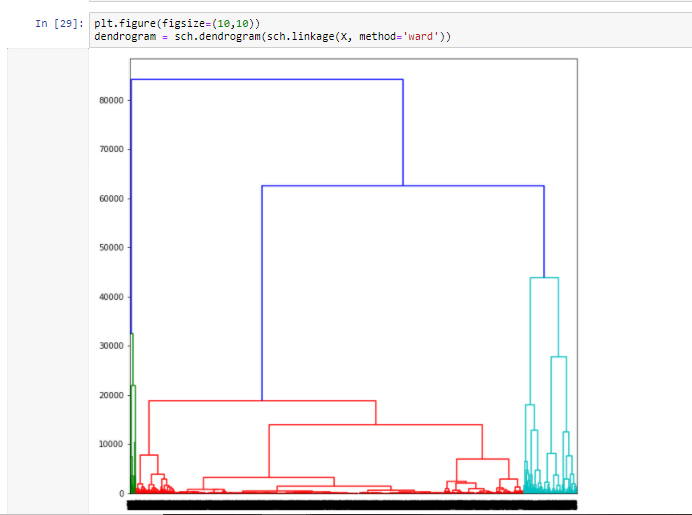


1. **Mengimport Library**

Pada line 21 kita akan mengimport library yang digunakan dalam proses clustering hirarki. Kemudian pada line 28 kita melakukan konversi pada dataframe (df) untuk numpy np.array

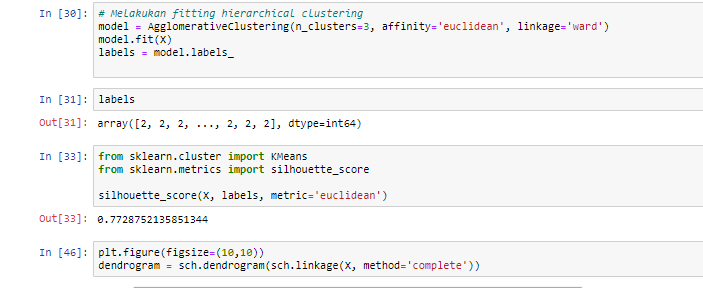
1. **Clustering menggunakan metode Ward Linkage**

Pada code dibawah ini kita akan membuat dendogram untuk mengelompokkan data menggunakan metode Ward Linkage, sehingga didapatkan hasil sebagai berikut:



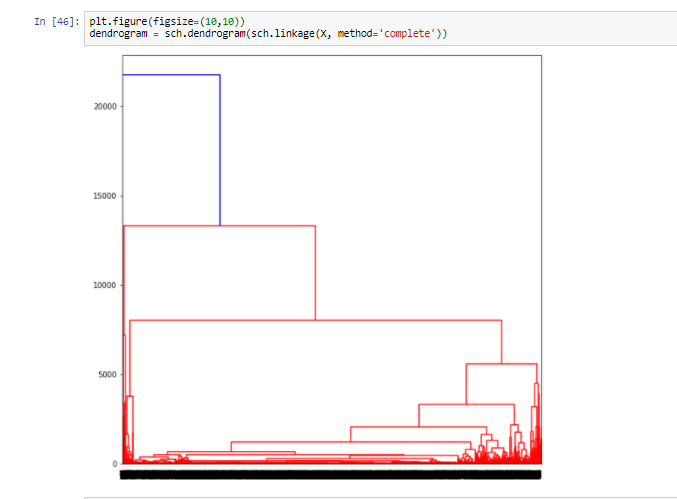
Pada line 30 kita melakukan fitting hierarchical clustering dengan membuat model Agglomerative Clustering menggunakan n\_clustes = 3, affinity = Euclidean dan linkage = ward. Kemudian kita membuat label dan ditampilkan label dari metode ini pada line 31.

Kemudian kita melihat hasil silhouette\_score yang terbentuk dari metode ini yaitu 0.728752135851344



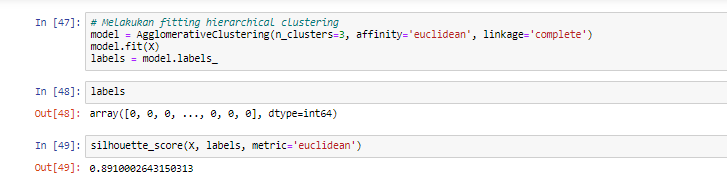
1. **Clustering menggunakan metode Complete Linkage**

Pada code dibawah ini kita akan membuat dendogram untuk mengelompokkan data menggunakan metode Complete Linkage, sehingga didapatkan hasil sebagai berikut:



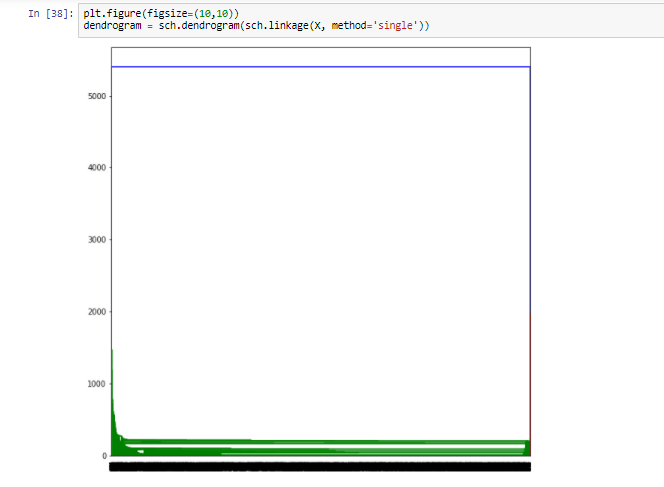
Pada line 47 kita melakukan fitting hierarchical clustering dengan membuat model Agglomerative Clustering menggunakan n\_clustes = 3, affinity = Euclidean dan linkage = complete. Kemudian kita membuat label dan ditampilkan label dari metode ini pada line 48.

Kemudian kita melihat hasil silhouette\_score yang terbentuk dari metode ini yaitu 0.8910002643150313



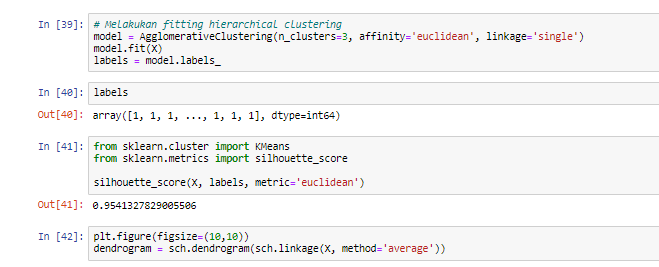
1. **Clustering menggunakan metode Single Linkage**

Pada code dibawah ini kita akan membuat dendogram untuk mengelompokkan data menggunakan metode Single Linkage, sehingga didapatkan hasil sebagai berikut:

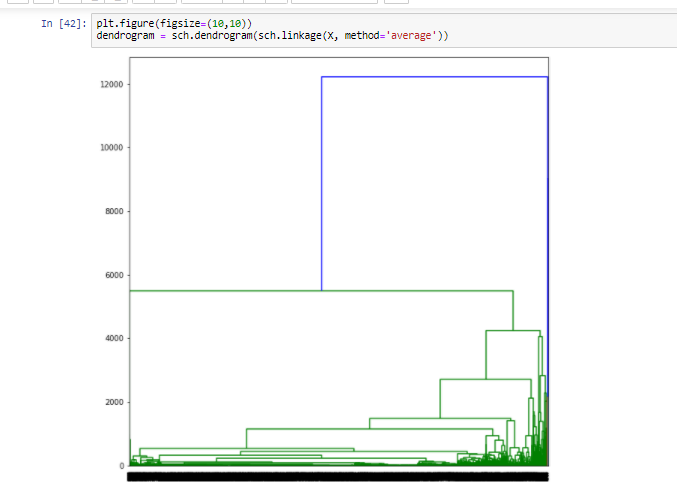


Pada line 39 kita melakukan fitting hierarchical clustering dengan membuat model Agglomerative Clustering menggunakan n\_clustes = 3, affinity = Euclidean dan linkage = single. Kemudian kita membuat label dan ditampilkan label dari metode ini pada line 40.

Kemudian kita melihat hasil silhouette\_score yang terbentuk dari metode ini yaitu 0.9541327829005506

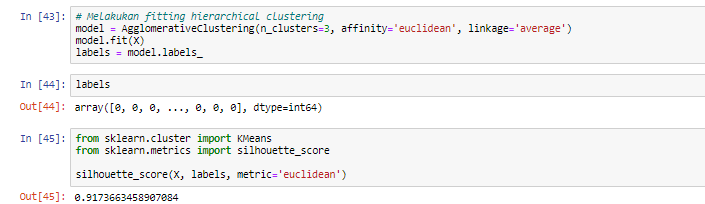


1. **Clustering menggunakan metode Average Linkage**

Pada code dibawah ini kita akan membuat dendogram untuk mengelompokkan data menggunakan metode Average Linkage, sehingga didapatkan hasil sebagai berikut:

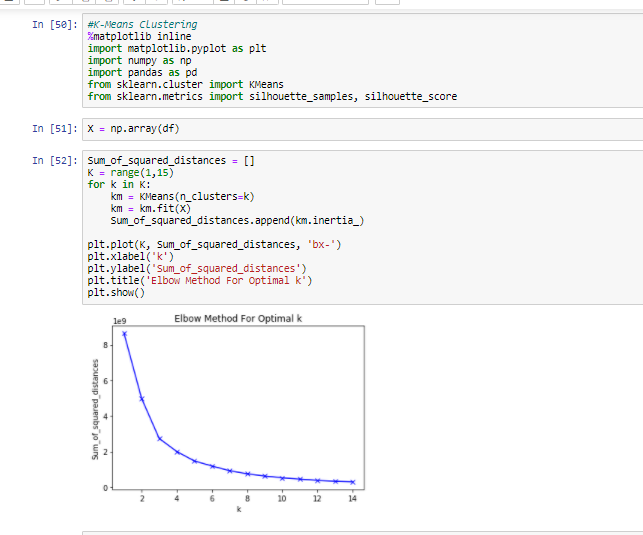
Pada line 43 kita melakukan fitting hierarchical clustering dengan membuat model Agglomerative Clustering menggunakan n\_clustes = 3, affinity = Euclidean dan linkage = average. Kemudian kita membuat label dan ditampilkan label dari metode ini pada line 44.

Kemudian kita melihat hasil silhouette\_score yang terbentuk dari metode ini yaitu 0.9173663458907084



# **K-Means**

K-Means Clustering adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan tanpa supervise (*unsupervised)* dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. K-Means merupakan salah satu metode data *clustering Non-Hierarchical* atau  *Pertitional Clustering.*

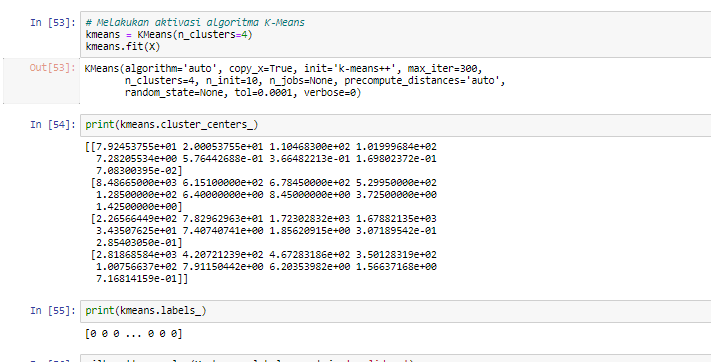


Pada algoritma k-means clustering, kita menggunakan sum\_of\_squared\_distance (SSE) untuk mendapatkan hasil perbandingan dari masing-masing nilai cluster. Dimana semakin besar jumlah cluster K maka nilai SSE akan semakin kecil.

Pada line 52 dapat dilihat bahwa semakin besar nilai K dari 0-14 maka SSE yang terbentuk akan semakin kecil.

Pada line 53 kita menggunakan n\_cluster = 4 sehingga didapati hasil seperti gambar dibawah.

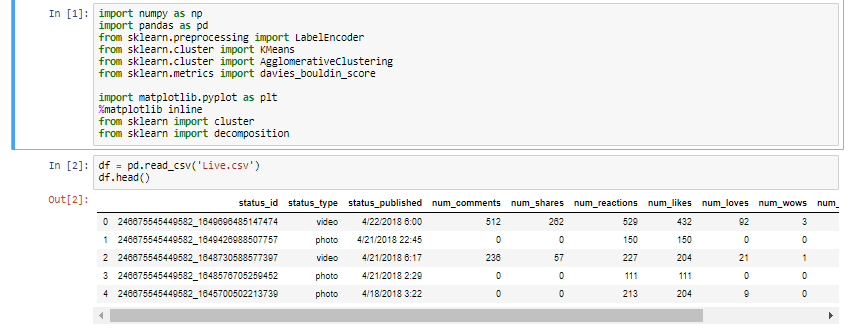
Pada line 54 dilakukan print kmeans.cluster\_centers\_ yang berfungsi untuk mencari nilai pusat dari masing-masing cluster, kemudian pada line 55 menampilkan hasil kluster dan pada line 56-57 dihitung menggunakan rata-rata jarak antara cluster terdekat untuk setiap sampel, dan score.



# **BAB V CODE**

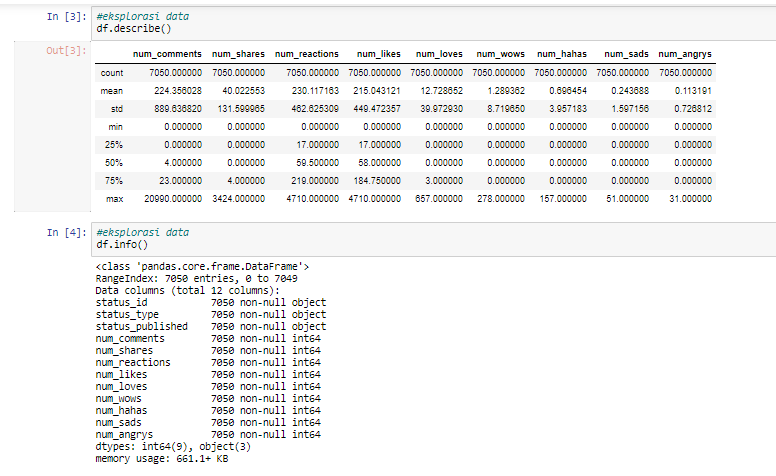
Pada line 1 digunakan untuk melakukan ekplorasi data dengan phyton, yaitu pandas untuk data frame, numpy untuk array, matplotlib untuk menciptakan visualisasi dan sklearn.

Pada line 2 untuk membaca dan menampilkan data 5 teratas dataset Live Sellers in Thailand Data Set digunakan syntax dibawah ini.



Pada line 3 digunakan untuk mengecek atau melihat dataset.

Pada line 4 fungsi info() digunakan untuk melihat nomor index beserta tipe datanya

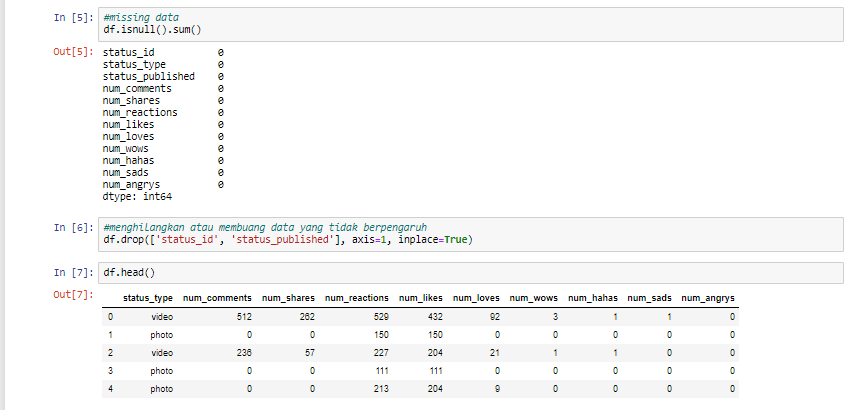


Pada line 5 dilakukan untuk mencari dan menangani missing data, terdapat dalam library phyton pandas pada fungsi isnull(), isna(). Seperti code dibawah ini fungsi ini digunakan untuk mengembalikan nilai boolean.

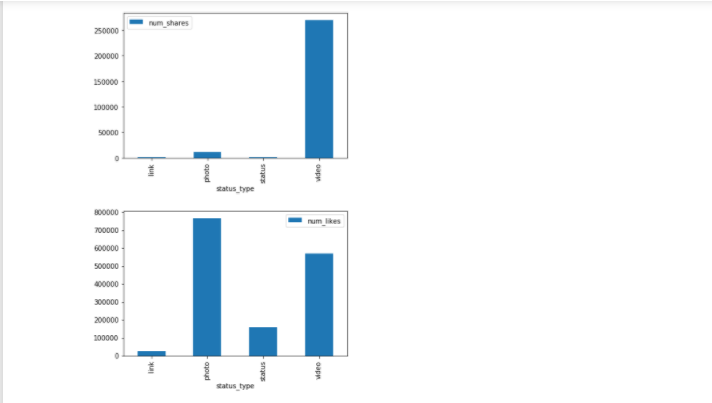
Pada line 6 digunakan untuk mengatasi itu, terdapat cara membuang, memberikan tanda, atau memberiarkan saya. Tetapi pada penggunaan code dibawah ini dilakukan dengan cara mendrop data status id, status published.

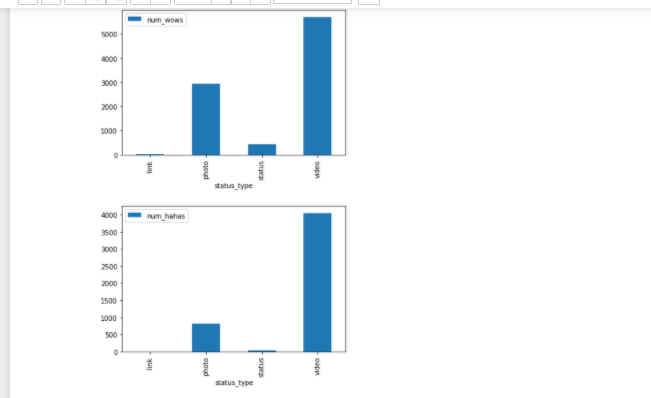
Pada line 7 digunakan untuk melihat data secara random.

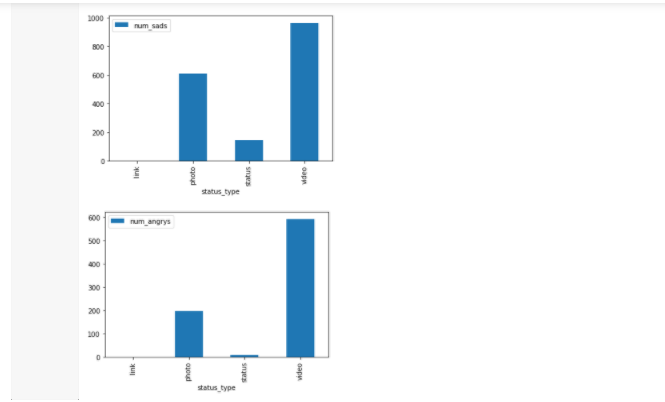
Pada line 8 digunakan untuk melihat bentuk data dengan lebih jelas dengan cara memvisualisasikan, dengan melihat distribusi masing-masing atribut dengan menampilkan dalam bentuk boxplot

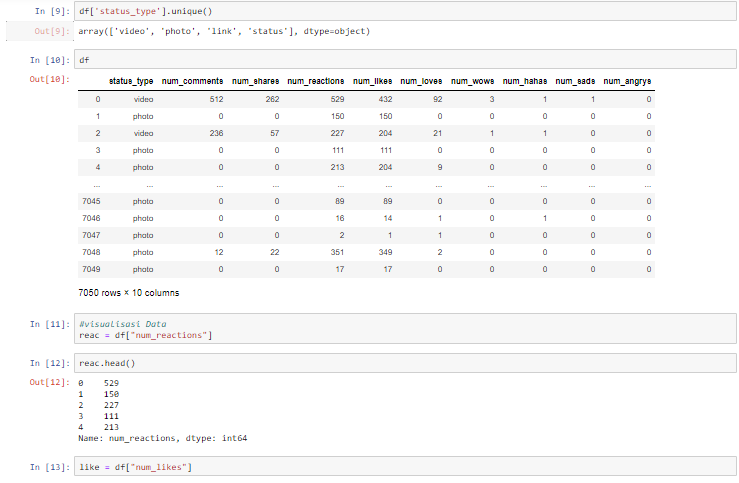


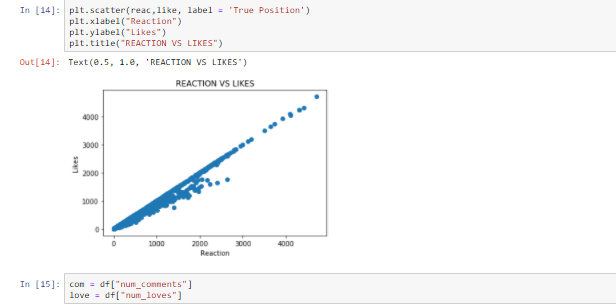


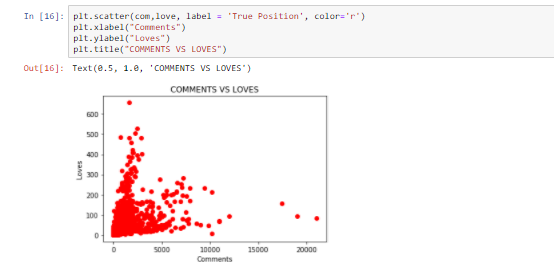


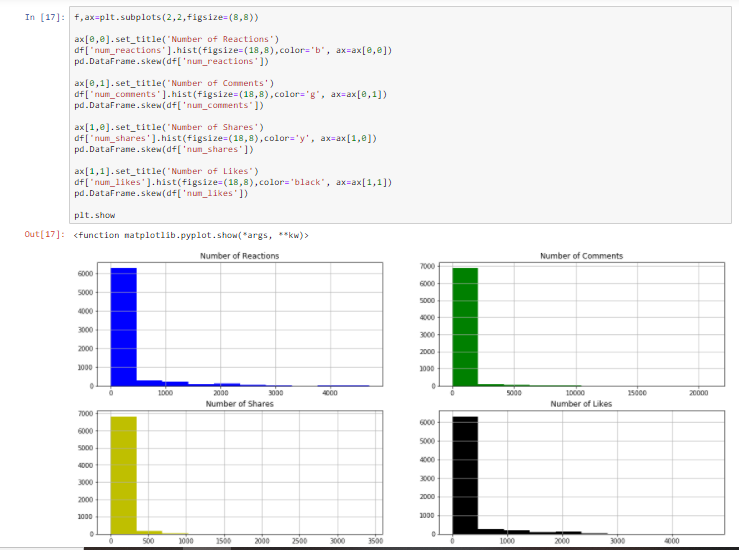


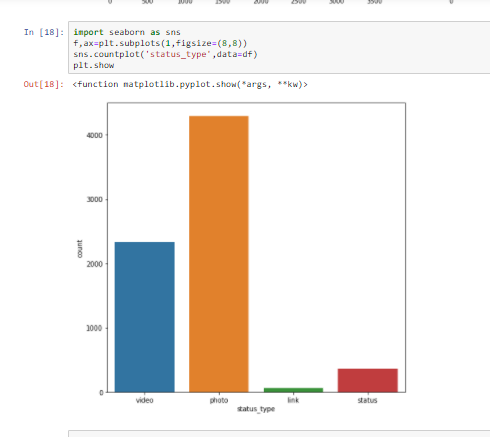


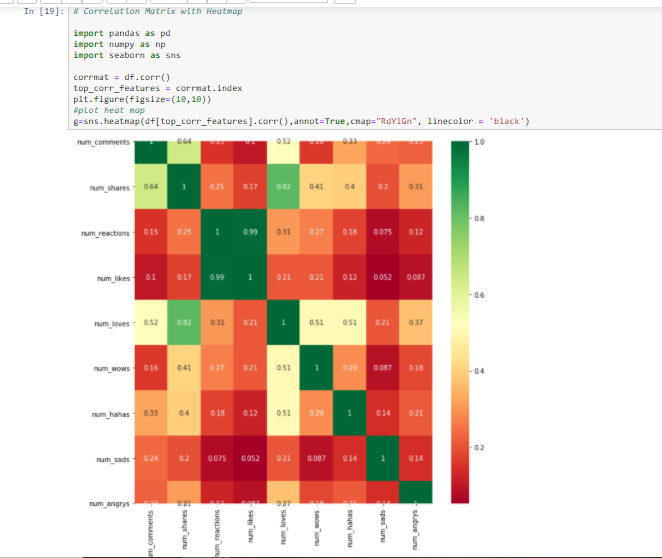




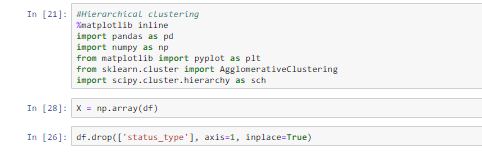


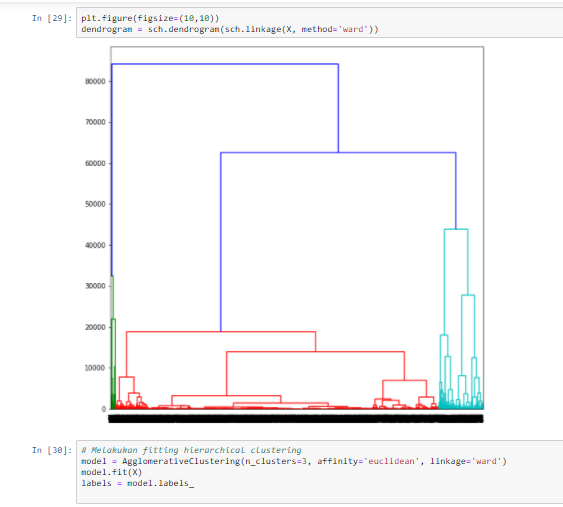




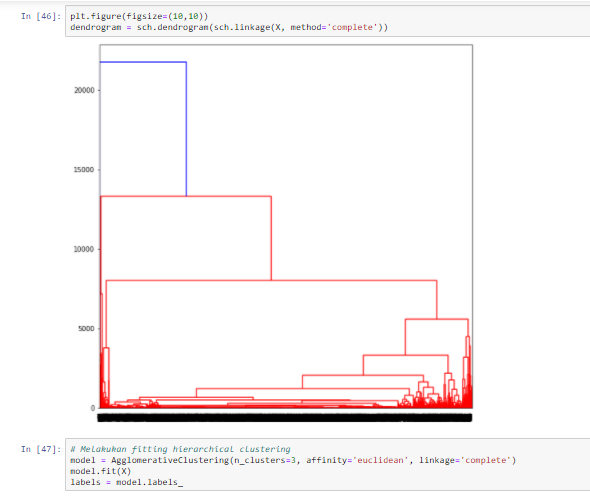


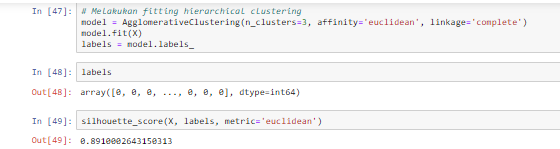


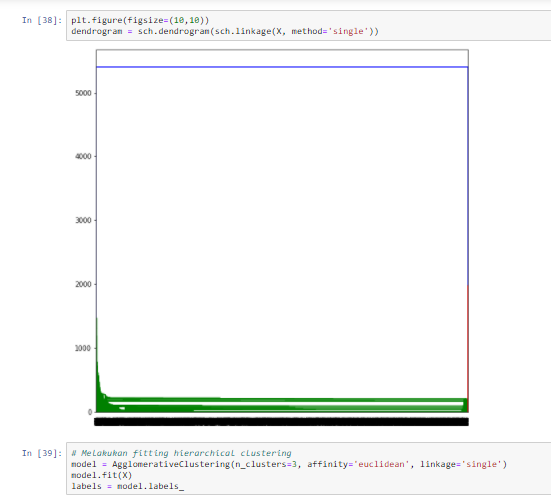


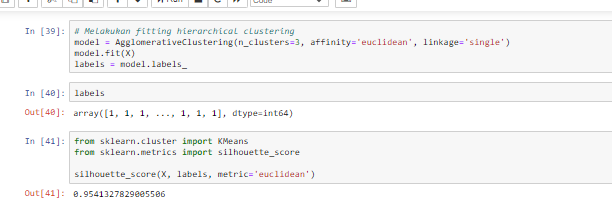


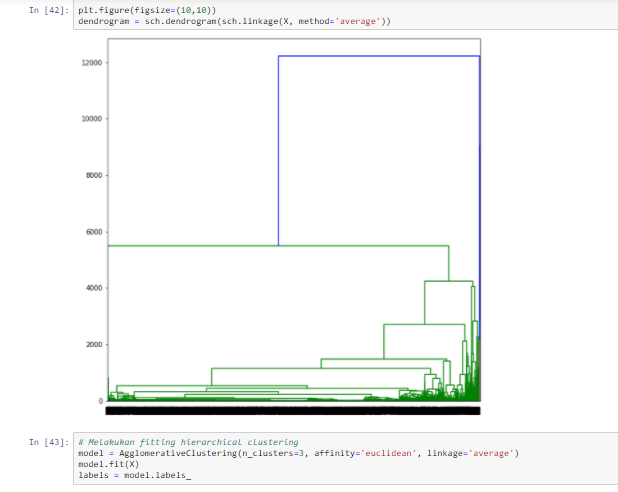


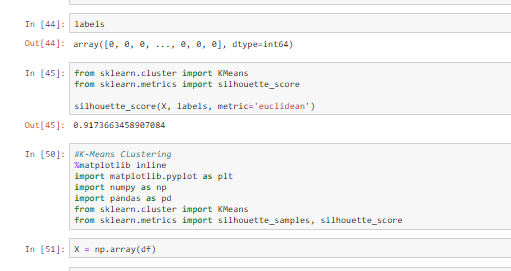


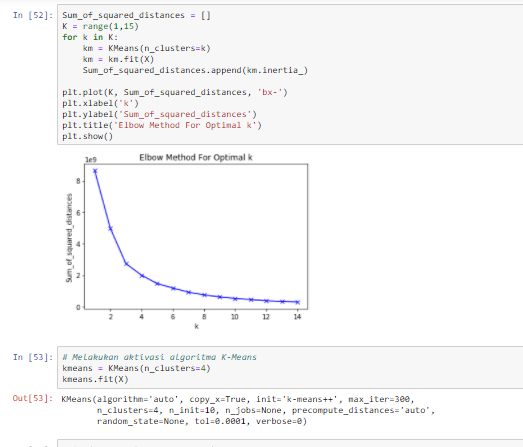


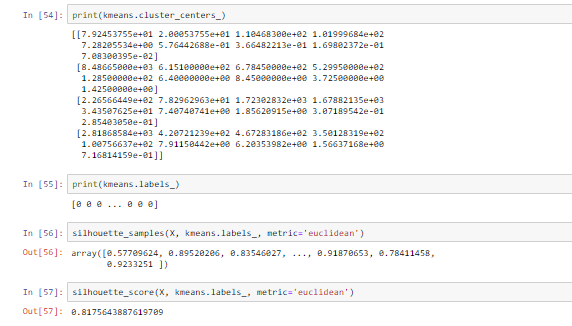








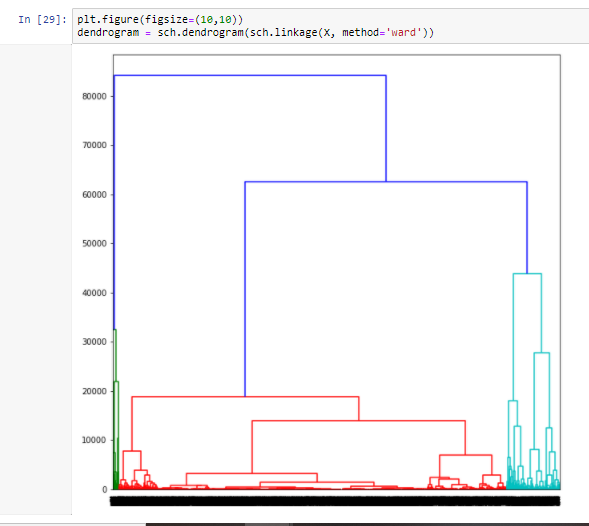


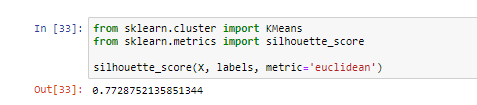


# **BAB VI ANALISIS OUTPUT DAN HASIL**

Berikut hasil analisis output dan hasil dari dataset menggunakan algoritma clustering hirarki dan kmeans. Pada algoritma clustering hirarki kami mengelompokkan data menggunakan 4 metode, yaitu:

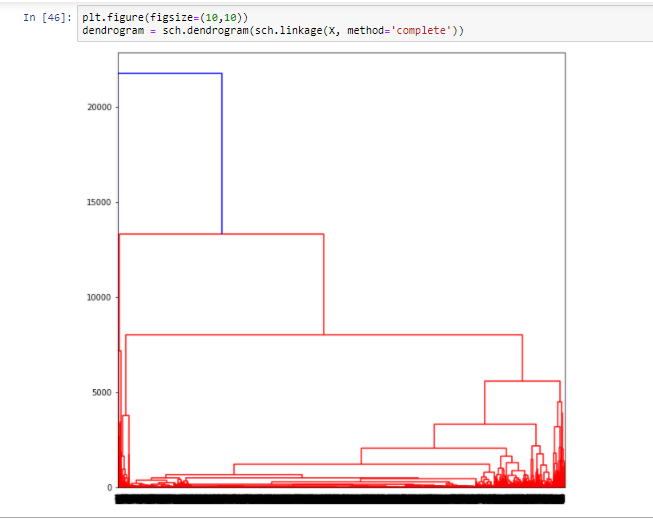
1. ward linkage



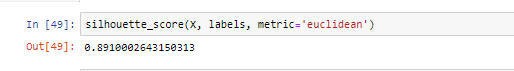


Pada metode ini ward linkage dilakukan dengan memperhatikan N cluster yang mempunyai satu responden cluster (dimana responend dianggap sebagai cluster), cluster pertama dibentuk dengan memilih dua dari cluster untuk meminumkan heterogen. Hasil pengelompokkan data dendogram yang terbentuk pada tipe data status berada pada jarak 0 sampai 32000, Video berada 0-18000, photo berada pada jarak 0-45000, status berada 0 sampai 32000, link berada pada jarak 18000 sampai > 80000

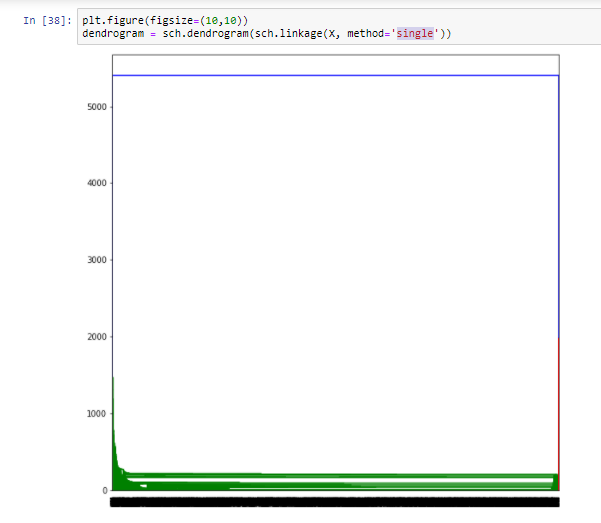
1. complete linkage



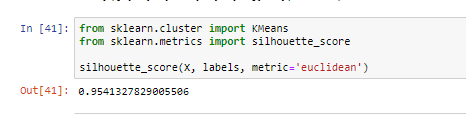
Method complete linkage dilakukan dengan mencari jarak maksimum atau jarak terjauh antar kelompok dan proses pengelompokkan sama dengan single linkage. Hasil pengelompokkan data dendogram yang terbentuk pada tipe data video berada pada jarak 0 sampai 13000, dan link berada 13000 sampai 23000



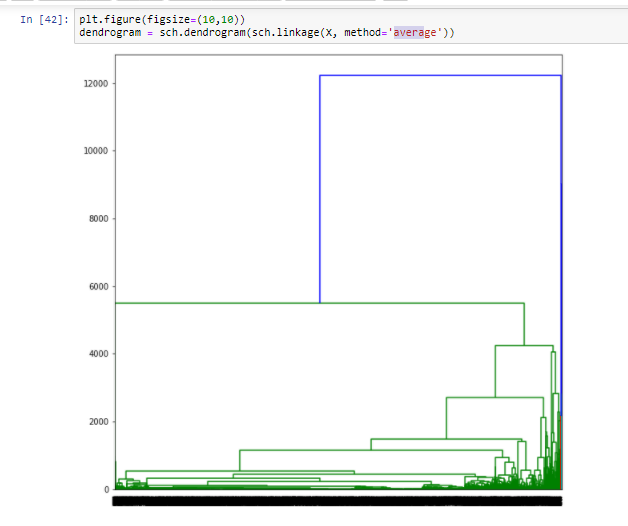
1. Single linkage



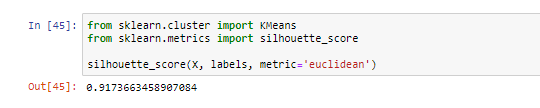
Method single linkage dilakukan dengan mencari jarak terdekat dan menggabungkan cluster cluster menurut jarak antara anggota-anggota terdekat diantara dua cluster. Hasil pengelompokkan data dendogram yang terbentuk pada tipe data status berada pada jarak 0-1300, dan video berada 0-2000



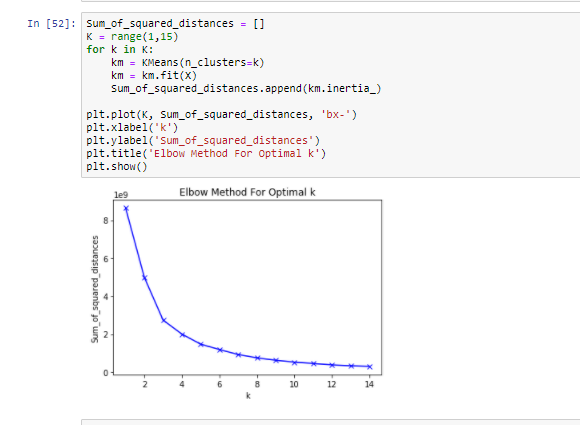
1. Average Linkage



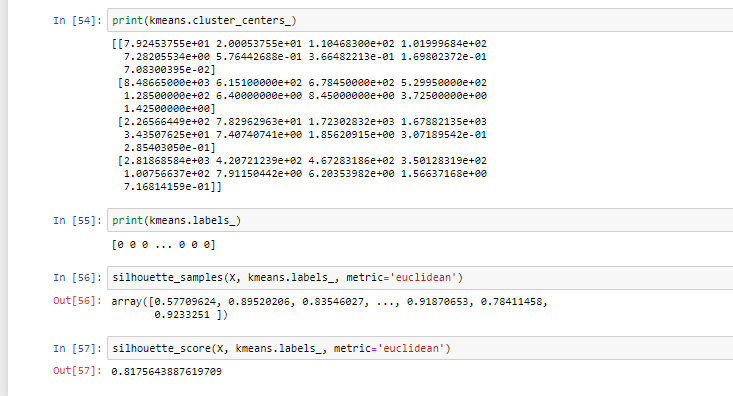
Method average linkage dilakukan dengan menggabungkan cluster-cluster menurut jarak rata-rata pasangan anggota masing-masing pada himpunan dua cluster. Hasil pengelompokkan data dendogram yang terbentuk pada tipe data status berada pada jarak 0 sampai 5000, link berada pada jarak 5000 sampai13000, dan video berada pada jarak 0-2200



Kemudian pada algoritma k-means clustering, kita menggunakan sum\_of\_squared\_distance (SSE) untuk mendapatkan hasil perbandingan dari masing-masing nilai cluster. Dimana semakin besar jumlah cluster K maka nilai SSE akan semakin kecil. Dapat dilihat pada gambar, semakin besar nilai K dari 0-14 dapat disimpulkan SSE yang terbentuk akan semakin kecil.



Pada syntax dibawah inline 54 bertujuan untuk mencari nilai pusat dari masing-masing cluster, kemudian pada inline 55 menampilkan hasil kluster tersebut, selanjutnya pada inline 56-57 dihitung menggunakan rata-rata jarak antara cluster terdekat untuk setiap sampel, dan score.



# **REFERENSI**

[1] <https://www.megabagus.id/machine-learning-hierarchical-clustering/3/> Diakses pada 30 Desember 2020.

[2] <https://gifadn.medium.com/k-means-clustering-menggunakan-python-deeb0881333c> Diakses pada 30 Desember 2020.