LAPORAN

PEMBELAJARAN MESIN



Laporan ini dibuat untuk memenuhi Tugas Pemrograman Tahap 1 Pembelajaran Mesin IF-42-03

Disusun Oleh :

Adabi Raihan Muhammad 1301180379

Telkom University

Bandung

2021

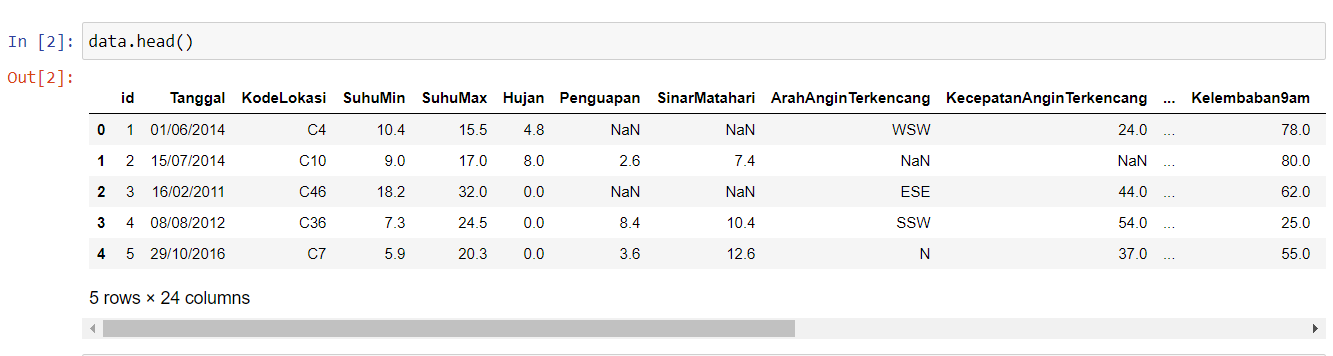
**Formulasi Masalah**

Pada tugas machine learning tahap 1 ini, saya diharuskan membuat model clustering dari dataset salju yang nantinya akan memprediksi apakah di hari tersebut akan turun salju / tidak.

**Data Exploration**

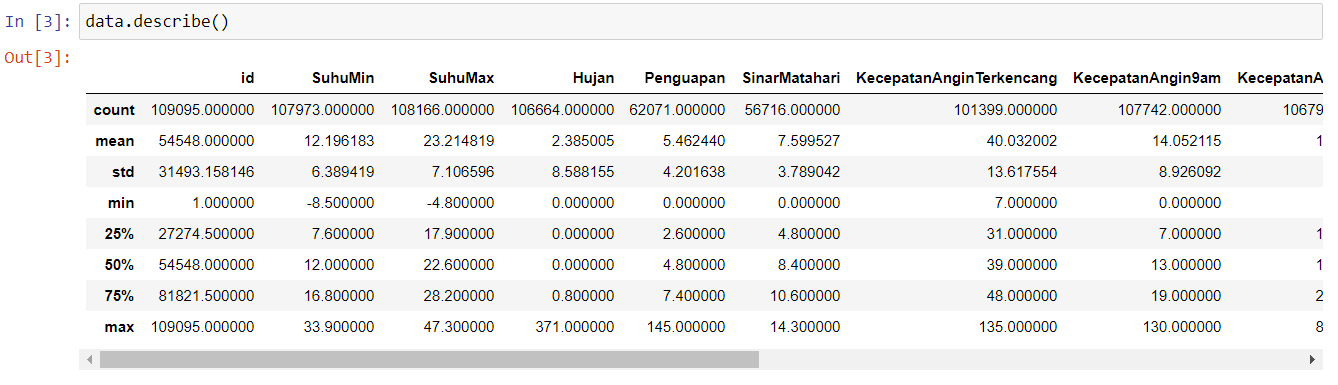
Tahap pertama dalam tugas ini yaitu Data Exploration, berguna untuk melihat kriteria dari dataset yang ingin diolah (clustering) berikut metode Data Exploration yang saya gunakan dalam dataset salju :

* data.head()

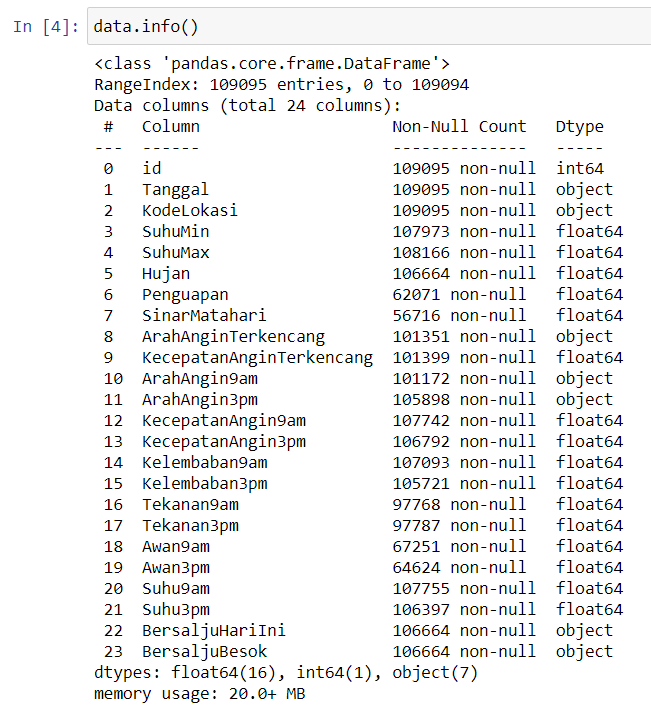


Berfungsi untuk melihat 5 data pertama, saya menggunakan metode data.head() karena ingin melihat apa saja isi dari dataset salju dengan hanya menampilkan 5 buat data.

* data.describe()

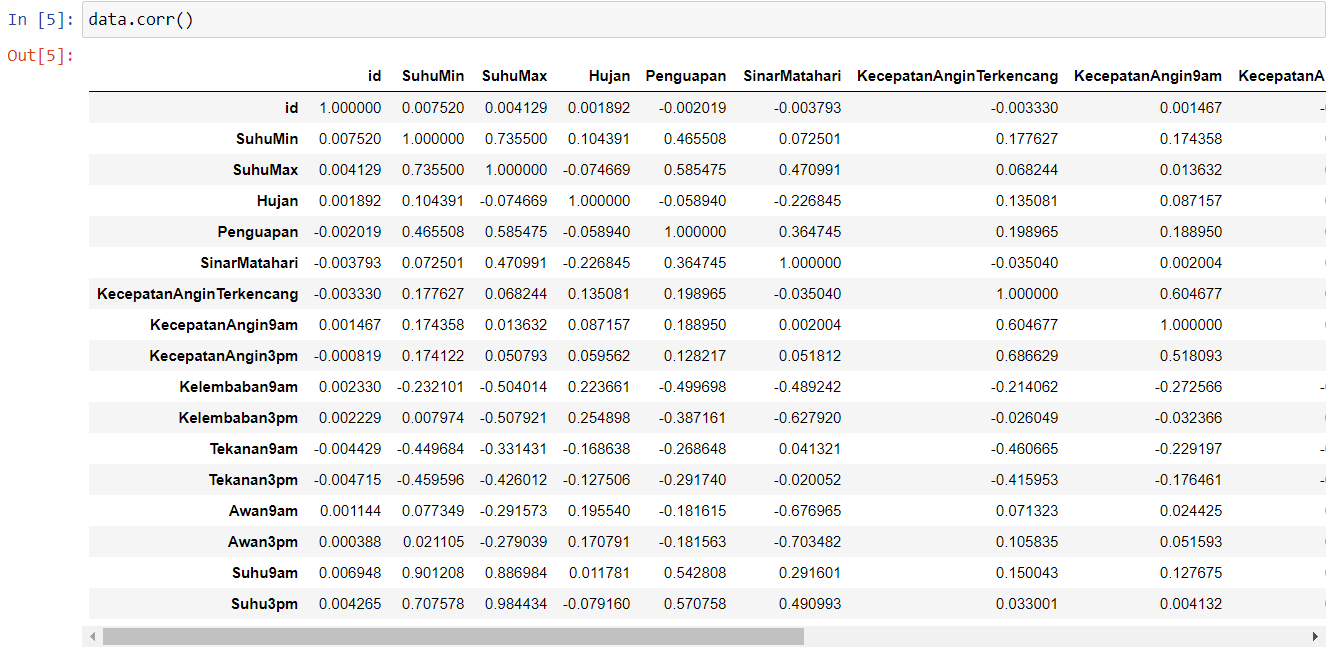


Berfungsi untuk melihat count, mean, dll dari dataset salju.

* data.info()

Berfungsi untuk melihat tipe data dari tiap column dan juga melihat data mana yang mempunyai nilai null paling tinggi.

* data.corr()

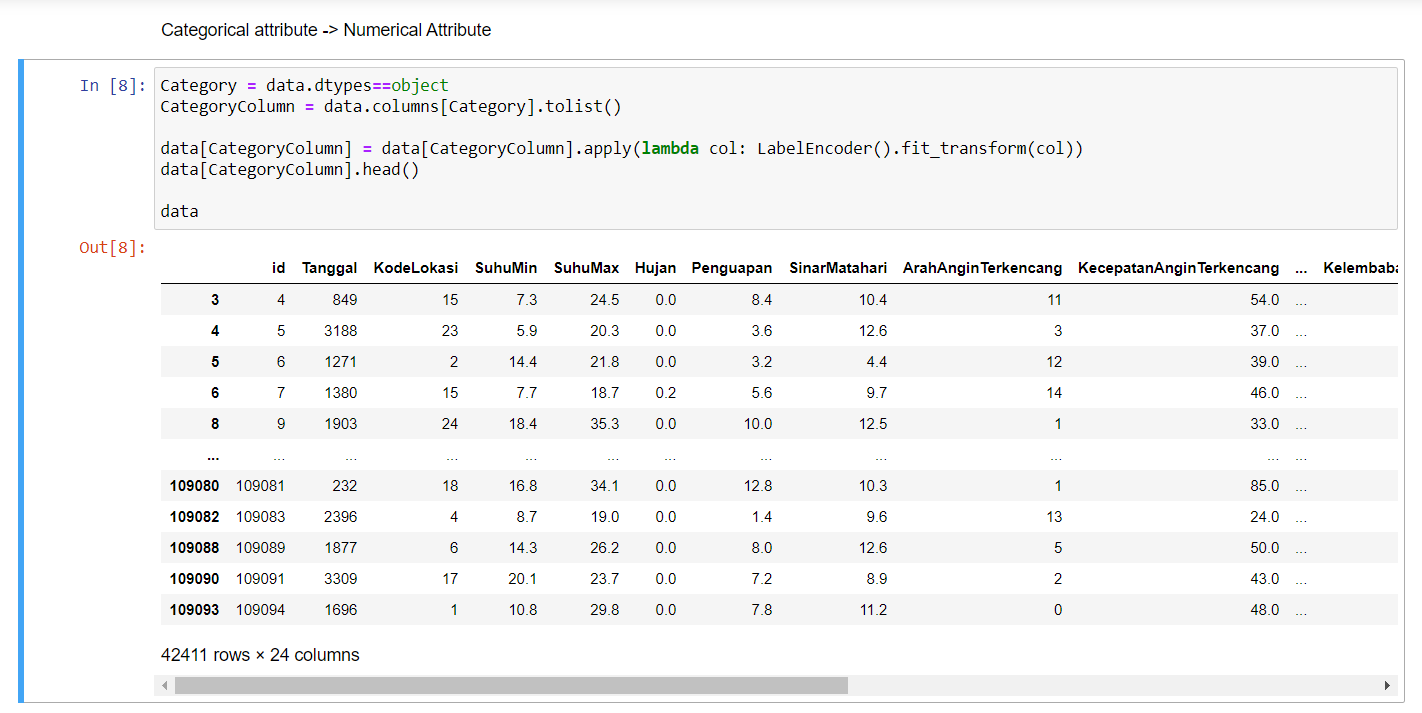
  
Berfungsi untuk melihat korelasi antar data.

**Data Pre-Processing**

Tahap kedua ada Data Pre-Processing yang bertujuan untuk memilah data mana saja yang penting / tidak, guna saat melakukan olah data (clustering) akan menghasilkan data cluster yang optimal, berikut metode yang saya gunakan untuk Data Pre-Processing :



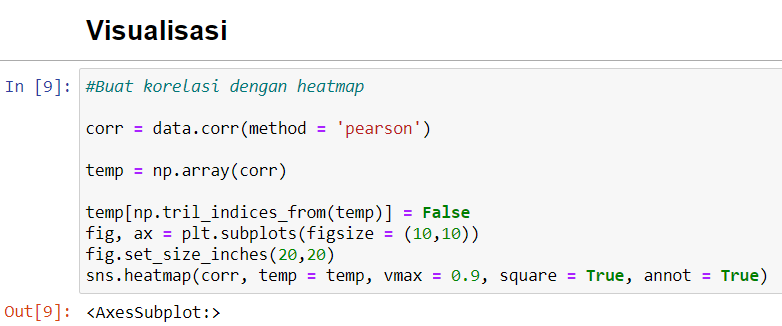
Drop yang memilki nilai null secara automatis, lalu cek apakah masih ada/tidak yang memiliki nilai null pada setiap atribut.

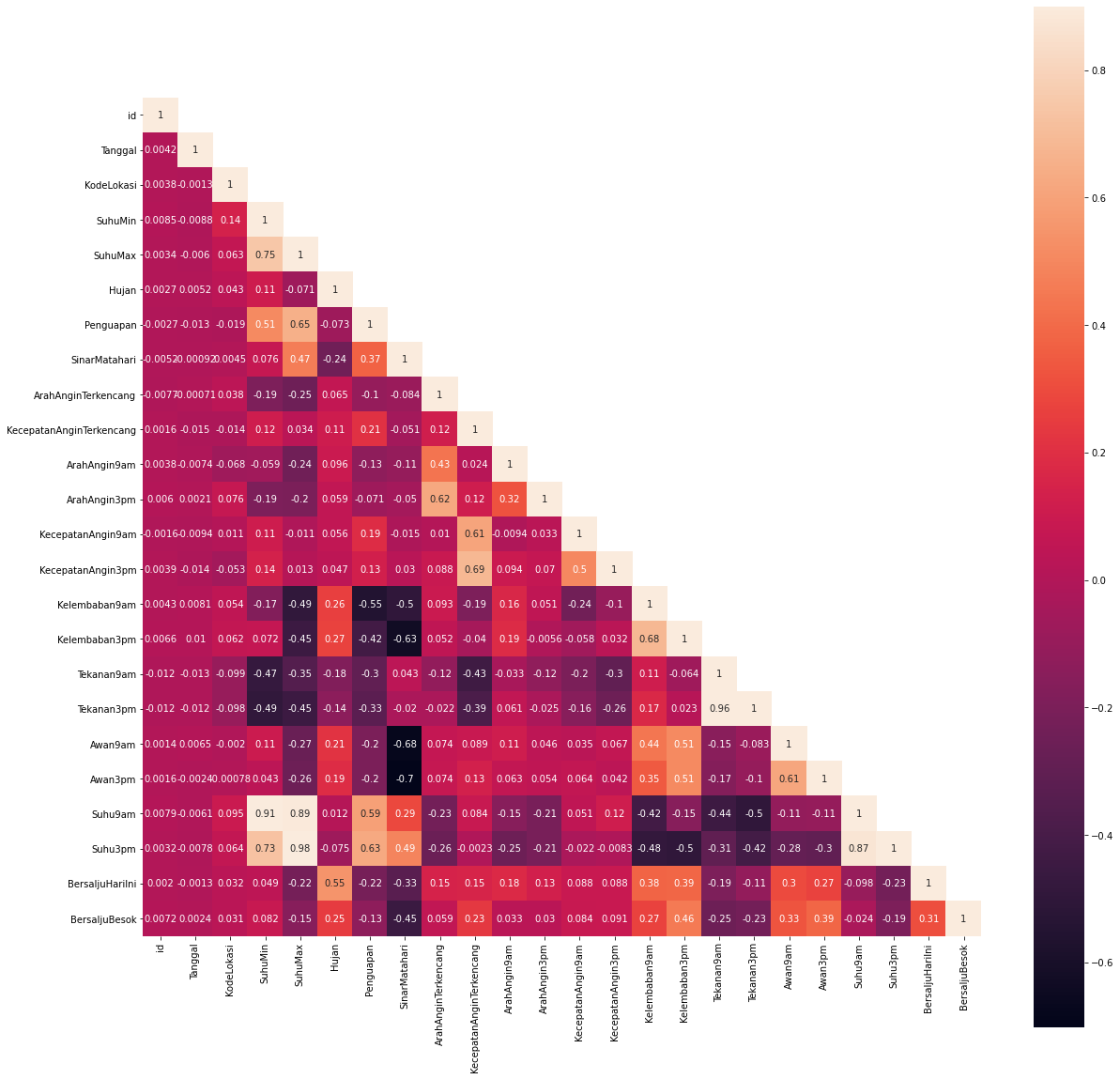


Men-konversi atribut category 🡪 numerical agar data dapat diolah.

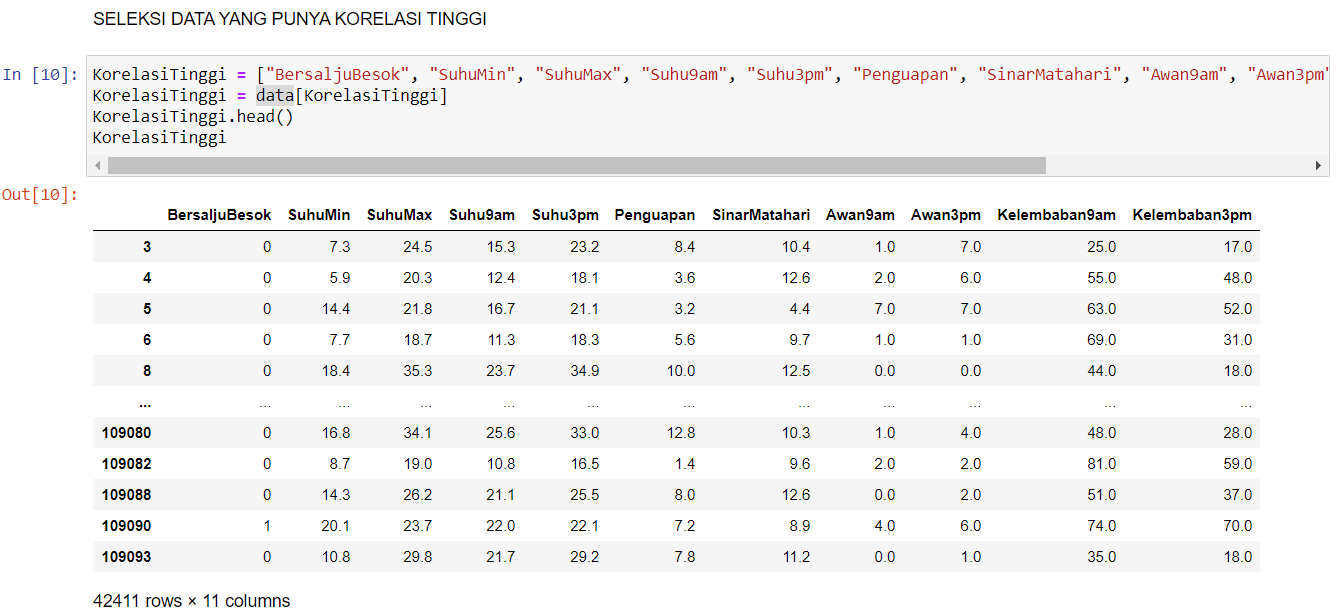
**Visualisasi Data**

Tahapan ketiga ada visualisasi data, bertujuan untuk mempermudah melihat data mana saja yang memiliki tingkat korelasi tinggi, lalu saya juga membuat sebuah array yang isinya merupakan data yang memiliki korelasi tinggi, lalu saya juga menggunakan boxplot untuk melihat apakah data tersebut (korelasi tinggi) memiliki outlier / tidak, jika memiliki outlier akan di drop.



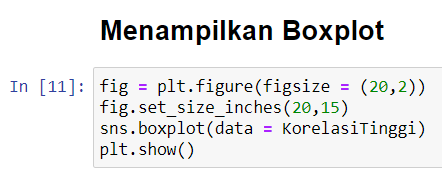


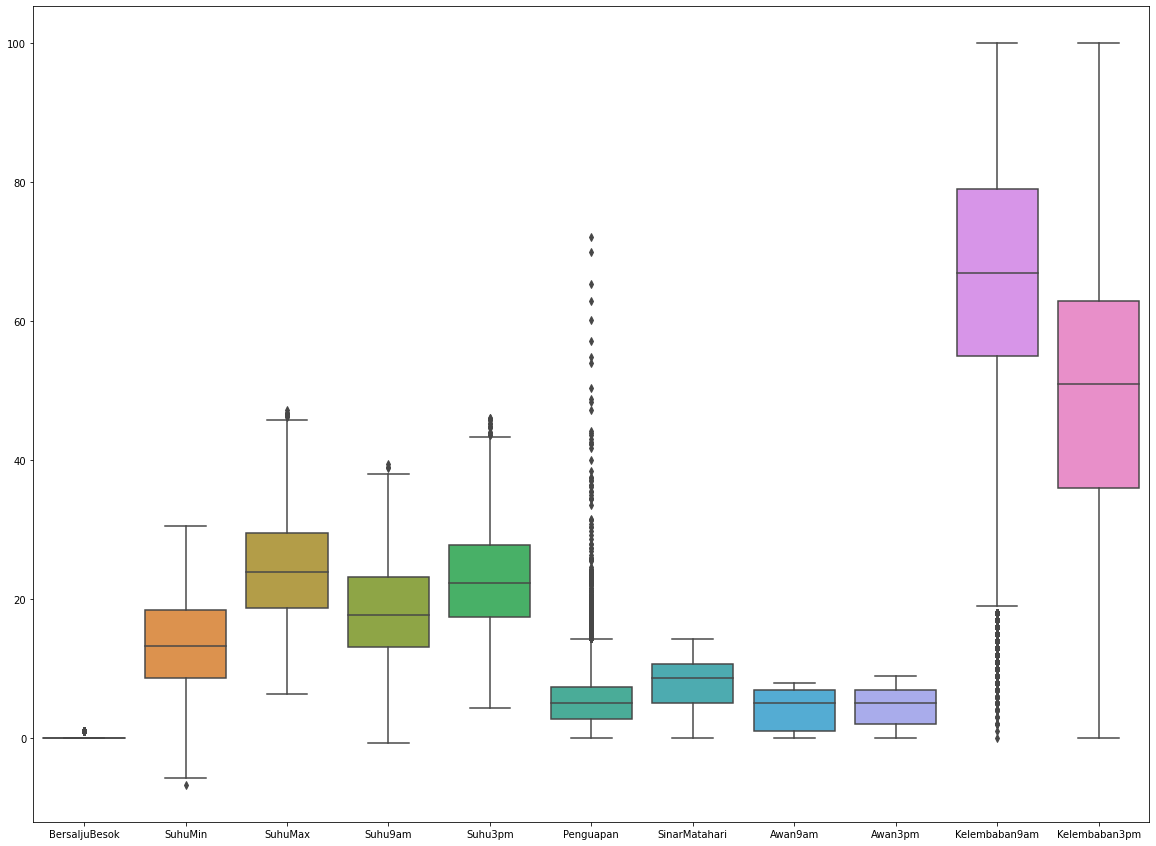
Saya menggunakan visualisasi data berupa Heat Map agar dapat dengan mudah melihat korelasi data yang nantinya akan saya olah.



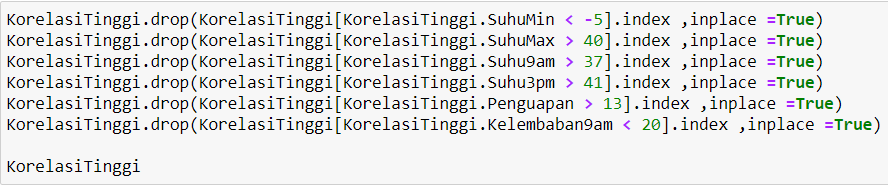
Setelah melihat dengan menggunakan heat map, saya mendapatkan bahwa data-data berikut memiliki korelasi paling tinggi diantara data-data lainnya.

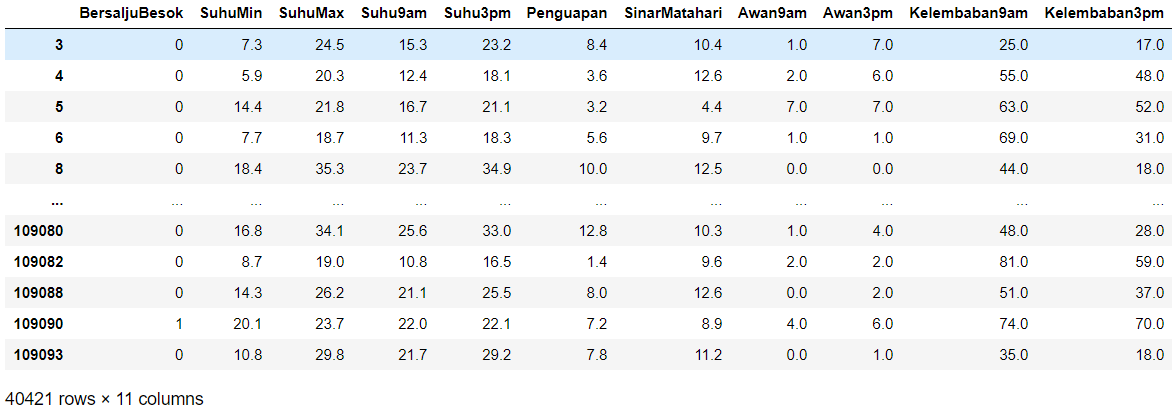
**KorelasiTinggi = ["BersaljuBesok", "SuhuMin", "SuhuMax", "Suhu9am", "Suhu3pm", "Penguapan", "SinarMatahari", "Awan9am", "Awan3pm", "Kelembaban9am", "Kelembaban3pm"]**



****

Saya menggunakan Boxplot sebagai visualisasi data selanjutnya untuk melihat outlier yang terdapat pada setiap data yang nantinya akan di drop (jika memiliki outlier).





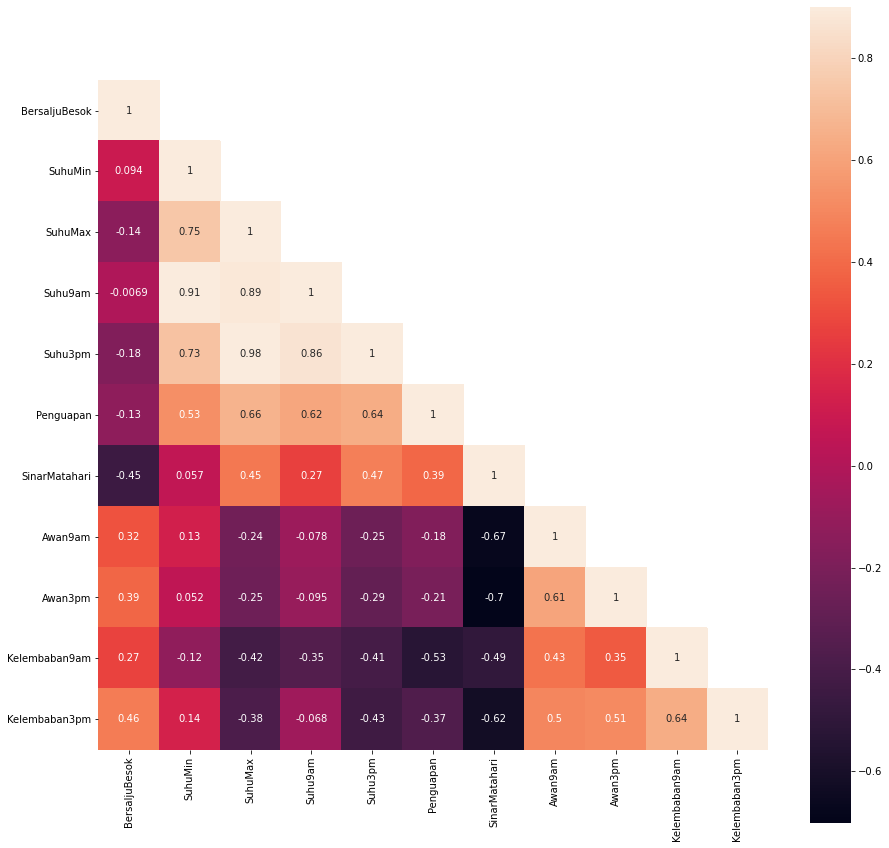
Setelah mendapatkan outlier dari boxplot sebelumnya, saya melakukan drop data sesuai nilai outlier dari setiap data.

**Scaling**

Tahapan kelima yaitu melakukan scaling terhadap data yang sebelumnya, scaling bertujuan agar setiap atribut dari data tesebut memiliki skala minimum dan maksimum yang sama nilai nya ( 0 sampai 1 ), dan juga agar mempermudah pemrosesan data di tahap selanjutnya.



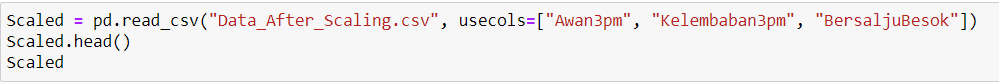
Di tahap scaling saya menggunakan fungsi MinMaxScaler() untuk melakukan scaling terhadap data data yang sudah ada, setelah di scaling dapat dilihat bahwa semua atribut sudah mempunyai nilai 0 sampai 1.



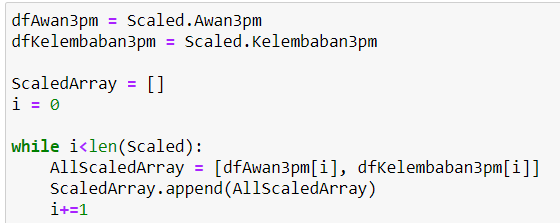
Setelah data di scaling saya menggunakan heatmap lagi untuk memastikan apa data benar sudah di scaling / belum.



Lalu saya memasukan data yang sudah di scaling menjadi file baru agar dapat mudah di proses di tahap selanjutnya tanpa ada nya data ambigu.



Saya juga melakukan perubahan dari dataframe menjadi sebuah array agar memudahkan pemroresan data.

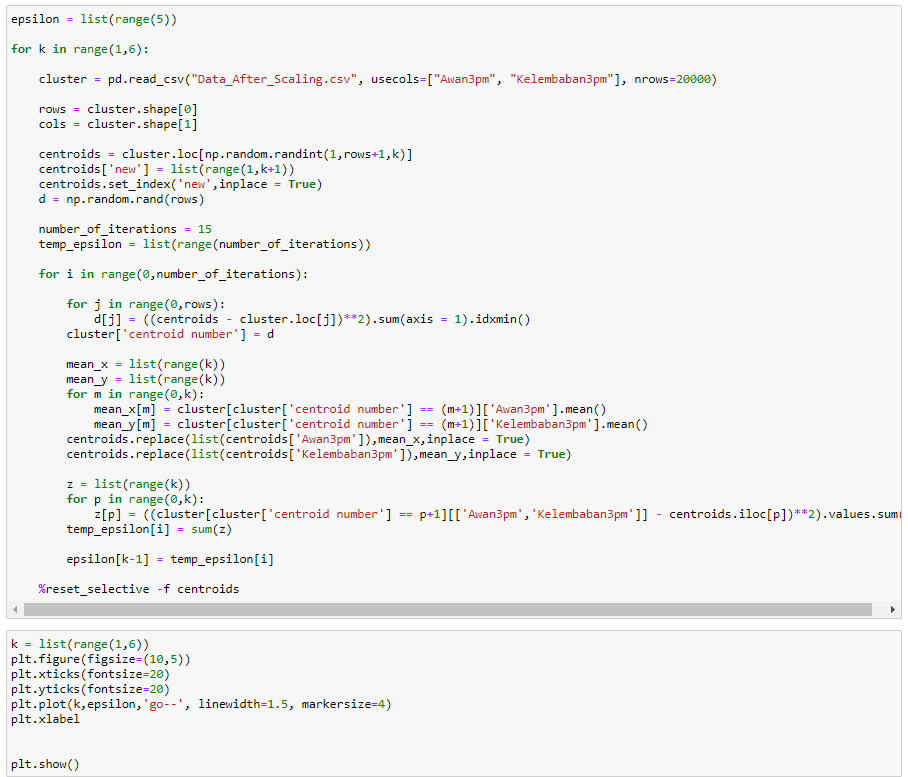


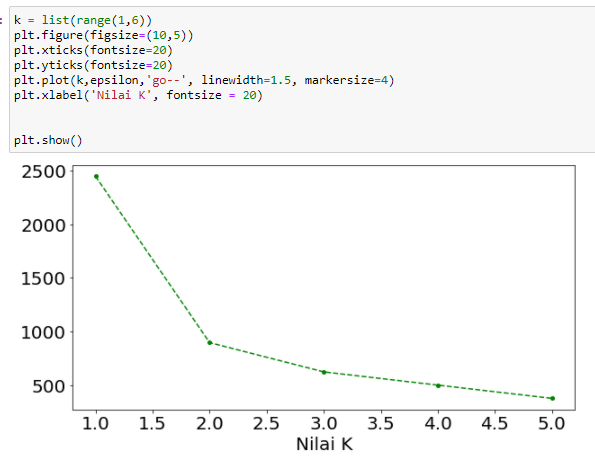
**Mengevaluasi dataset dengan menentukan nilai K dengan Elbow Method**

Di tahapan ini, saya menggunakan elbow method untuk mengetahui berapa nilai K optimal untuk nanti saat clustering.

**NOTE : Dalam melakukan Elbow Method ini saya mengambil referensi dari**

[**https://medium.com/analytics-vidhya/elbow-method-of-k-means-clustering-algorithm-a0c916adc540**](https://medium.com/analytics-vidhya/elbow-method-of-k-means-clustering-algorithm-a0c916adc540)



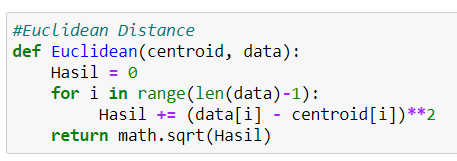


Setelah menggunakan Elbow Method didapati bahwa tekukan pertama ada di angka 2 yang menjadikan angka 2 tersebut adalah nilai K yang paling optimal diantara nilai lainnya.

**Clustering**

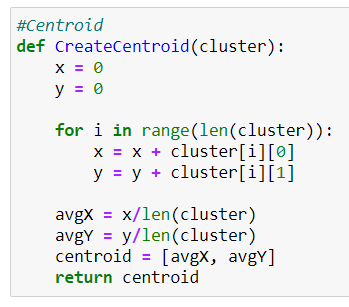
Di proses Clustering saya menggunakan 3 buah fungsi yaitu :

* Euclidean Distance



Fungsi Euclidean Distance bertujuan untuk mencari jarak antara dua titik.

* Create Centroid



Fungsi Create Centroid bertujuan untuk membuat centroid baru di proses clustering.

* KMeans



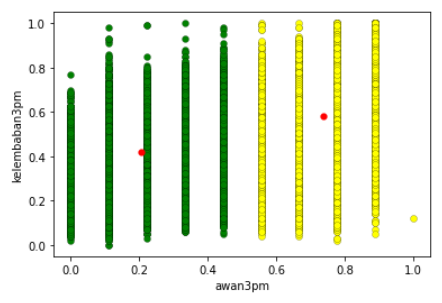


Fungsi KMeans bertujuan untuk melakukan tujuan utama yaitu Clustering.

**Hasil**

Di tahapan terakhir ini saya akan menampilkan penyebaran cluster nya pada data yang sebelumnya telah melalui beberapa tahapan.





Dari hasil visualisasi data clustering data berikut, dapat dilihat bahwa ada peralihan cluster yang cukup jelas antara cluster1 dan cluster2, tetapi ini belum 100% pasti dikarenakan dibutuhkan tahapan lebih lanjut yaitu klasifikasi agar dapat menambah keakuratan apakah benar penyebaran cluster tersebut akurat / tidak.



Di akhir saya konversi kan kembali array yang sebelumnya adalah dataframe menjadi balik ke asalnya menjadi dataframe lagi agar dapat dibuat ke dalam file .csv

**Kesimpulan**

Saya dapat menarik kesimpulan bahwa dengan menggunakan metode mencari antar atribut dengan korelasi tinggi, scaling,elbow dan clustering dapat membuat suatu prediksi akan suatu hal ( sesuai atribut di dataset tersebut ) tetapi prediksi tersebut belum tentu PASTI dikarenakan dibutuhkan tahapan selanjutnya yaitu klasifikasi untuk lebih menambah keakuratan-nya ( penyebaran cluster-nya ).

Link Youtube :

<https://www.youtube.com/watch?v=lWj0u1TMruc&t=166s>