**LAPORAN PROYEK** **PEMBELAJARAN MESIN**

**Uji Online Retail dengan Algoritma Clustering K – Means**



**DIUSULKAN OLEH:**

|  |  |
| --- | --- |
| Rafelli Simangunsong | 11S19003 |
| Albert S. Sormin | 11S19039 |
|  |  |
|  |  |
|  |  |

**S1 INFORMATIKA**

**INSTITUT TEKNOLOGI DEL**

**LAGUBOTI**

**2022**

**Daftar Isi**

[**BAB I DASAR TEORI 3**](#_Toc60426632)

[**1.1 Data Mining 3**](#_Toc60426633)

[**1.2 K-Means 3**](#_Toc60426634)

[**1.3 K-Means Clustering 4**](#_Toc60426635)

[**BAB II PREPROCESSING DATA 6**](#_Toc60426636)

[**2.1 Missing Data 6**](#_Toc60426638)

[**2.2 Normalisasi 8**](#_Toc60426639)

[**2.3 Outlier 9**](#_Toc60426640)

[**2.3.1 Mengecek Outlier 9**](#_Toc60426641)

[**2.3.2 Menghandle Outlier 10**](#_Toc60426642)

[**BAB III EKSPLORASI DATA 11**](#_Toc60426643)

[**3.1 Melihat / mengecek data 11**](#_Toc60426644)

[**BAB IV CODE 15**](#_Toc60426645)

[**BAB V PEMBAHASAN 40**](#_Toc60426646)

[**5.1 Analisis Output 40**](#_Toc60426650)

[**5.2 Hasil 44**](#_Toc60426651)

[**REFERENSI 45**](#_Toc60426652)

# BAB I DASAR TEORI

## Data Mining

Data mining merupakan suatu metode pengolahan data untuk menemukan pola yang tersembunyi dari data tersebut. Hasil dari pengolahan data dengan metode data mining ini dapat digunakan untuk mengambil keputusan di masa depan. Data mining adalah pengolahan data dengan skala besar, sehingga data mining memiliki peranan penting dalam bidang industri, keuangan, cuaca, ilmu dan teknologi (Ong, 2013). Data mining juga dapat dilakukan pada berbagai jenis database dan penyimpanan informasi, namun jenis pola yang akan ditemukan ditentukan oleh berbagai fungsi data mining seperti deskripsi class atau konsep, asosiasi, analisis korelasi, klasifikasi, prediksi, analisis cluster dan lain-lain (Dash et al., 2010).

Data mining menggunakan penerapan algoritma tertentu untuk mengekstrak pola dari data. Dimana proses ini secara otomatis akan mencari pola yang sederhana terhadap data yang besar menggunakan analisa tertentu. Data mining juga menggunakan algoritma matematika yang canggih untuk mensegmen data dan mengevaluasi kemungkinan beberapa hasil yang ditetapkan oleh pengguna (Ndehedehe et al., 2013).

Ada lima tahapan yang terdapat pada tahapan data mining yaitu:

1. Pre-processing / Cleaning

Proses data mining bisa dilakukan dalam proses cleaning pada data yang menjadi fokus KDD.

1. Transformation

Transformasi pada data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung kepada jenis, atau pola informasi yang akan dicari

1. Data Selection

Data yang perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai dari data hasil seleksi yang akan digunakan untuk proses data mining disimpan dalam suatu berkas, terpisah dari basis data operasional.

1. Data Mining

Data mining merupakan proses untuk mencari pola dan informasi yang menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.

## K-Means

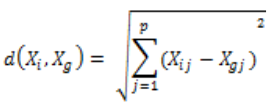
K-Means merupakan salah satu algoritma *clustering*, dimana pada algoritma ini, komputer akan mengelompokkan sendiri data-data yang menjadi masukannya tanpa mengetahui terlebih dulu target kelasnya. Pembelajaran ini termasuk dalam *Unsupervised Learning.* Input yang diterima berupa data (objek) dan *k* buah kelompok (*cluster*) yang diinginkan. Algoritma ini akan mengelompokkan data (objek) ke dalam *k* buah kelompok tersebut. Pada setiap *cluster* terdapat titik pusat (*centroid*) yang merepresentasikan *cluster* tersebut .

Langkah-langkah algoritma K-Means

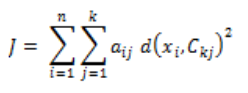
1. Menentukan k sebagai jumlah cluster yang ingin dibentuk.
2. Mengalokasikan data ke dalam cluster secara acak.
3. Menentukan pusat cluster (centroid) dari data yang ada pada masing-masing cluster dengan persamaan :

*https://lh6.googleusercontent.com/SnN5vGEQbH8EJ-3uwtMe31mW41sZnebG1x4uNzumlJOraNVraWxLTsRo9winUcV606vrg2NyDwstOyJkAGrWVmwg8yRrff9O39QKvGyfnPu8wAf4wwknaknO3b5VADXOp1V6xVRd.*

1. Menentukan jarak setiap objek dengan setiap centroid dengan perhitungan jarak setiap objek dengan setiap centroid menggunakan jarak Euclidean.



1. Menghitung fungsi objektif dengan formula:



1. Mengalokasikan masing-masing data ke centroid/rata-rata terdekat

## K-Means Clustering

K-Means Clustering merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik (similarity) antara satu data dengan data yang lain. Algoritma K-Means Clustering adalah proses pengelompokan benda serupa ke dalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah data set kedalam subset, sehingga data dalam setiap subset memiliki arti yang bermanfaat. Dimana sebuah cluster terdiri dari kumpulan benda-benda yang mirip antara satu dengan yang lainnya dan berbeda dengan benda yang terdapat pada cluster lainnya.

Algoritma clustering terdiri dari dua bagian yaitu secara hirarkis dan secara partitional. Algoritma hirarkis menemukan cluster secara berurutan dimana cluster ditetapkan sebelumnya, sedangkan algoritma partitional menentukan semua kelompok pada waktu tertentu (Madhu Latha, 2012). Clustering juga bisa dikatakan suatu proses dimana mengelompokan dan membagi pola data menjadi beberapa jumlah data set sehingga akan membentuk pola yang serupa dan dikelompokan pada cluster yang sama dan memisahkan diri dengan membentuk pola yang berbeda ke cluster yang berbeda (HUNG et al., 2005).

Berikut ini langkah-langkah dari algoritma untuk melakukan *K-Means clustering* :

1. Pilih *K* buah titik *centroid* secara acak.
2. Kelompokkan data sehingga terbentuk *K* buah *cluster* dengan titik *centroid* dari setiap *cluster* merupakan titik *centroid* yang telah dipilih sebelumnya.
3. Perbaharui nilai titik *centroid.*
4. Ulangi langkah 2 dan 3 sampai nilai dari titik *centroid* tidak lagi berubah.

Proses pengelompokkan data ke dalam suatu *cluster* dapat dilakukan dengan cara menghitung jarak terdekat dari suatu data ke sebuah titik *centroid.* Rumus untuk menghitung jarak tersebut adalah  :

https://lh4.googleusercontent.com/wVdf95Z368X3EONy6RcPNAtQizeph_gWYHBqodP8oLUJNnP9BFrcvld9HMpgo-CmN8YELMRrFJZMg8KAAVqs4YwiI_Sokhs6s3glFUrQ0Ev5eupdWEmvLOT0j9vCEmdEqlNSF8IV

dengan:

g = 1, untuk menghitung jarak Manhattan

g = 2, untuk menghitung jarak Euclidean

g = ~, untuk menghitung jarak Chebyshev

# BAB II PREPROCESSING DATA



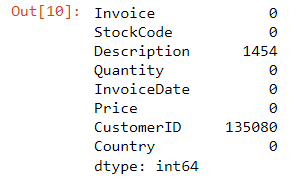
## Missing Data

Missing data adalah informasi yang hilang atau tidak tersedia untuk sebuah obyek. Missing data merupakan masalah yang sering dijumpai dalam penelitian, keberadaan missing data dapat mengganggu analisis yang akan dilakukan.

Untuk mengecek adanya missing data pada dataset online\_retail dapat dilihat pada code di bawah ini.

https://lh3.googleusercontent.com/GUZ2ajYWFx-NyWLiKQM9D9sDli_HUZv2FapFZ_M2Npyrt84U_c_7sONgTO3lKU-a38KHWDCmlKMsA4IyeQt_GsML8r3NZqiTQnOYQow3uuL_wt8wJvrxjcG8RwAB1cNPnMtYua4S

Hasil:



Pada hasil pengecekan data di atas diketahui bahwa terdapat missing value pada feature **Description** dan **CustomerID**. Berdasarkan hasil di atas data kosong yang terdapat pada kolom Description yaitu ada 1454 baris, dan pada kolom CustomerID ada 135080 baris data yg kosong. Oleh karena itu solusi yang dilakukan untuk menangani missing data yaitu dengan mengganti kategori unik. Data todak hilang akan tetapi diganti dengan kategori unik.

**Penanganan Data**

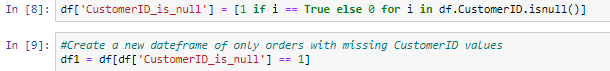
1. Mengecek persentase customer yang kosong

https://lh4.googleusercontent.com/krBypxGxeO0ROYpBqxw3AqROGbGLGuu9cEESomFPb-NkddJGYDBlbEtaSwkPaWVqyk92DO4PX5ooVRU0U6YGUsLD5JccHq3080nkjjV6Thsui8IvZqSXHgjIwWpuwle7tdqjfAKC

1. Menambahkan kolom CustomerID\_is\_null pada dataset. Kolom ini berisi value 1 jika

customerID kosong. Kemudian membuat dataframe yang terbaru yang telah berisi kolom

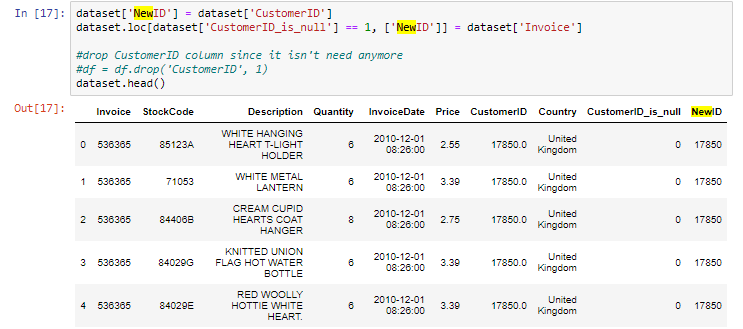
**CustomerID\_is\_null**.



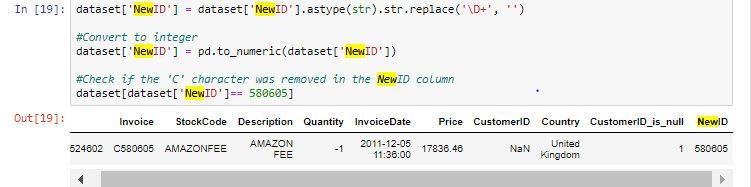
1. Menambahkan kolom NewID. Diasumsikan bahwa customer sudah memiliki id. Sehingga

ditambahi lah kolom NewID bertindak sebagai customer baru untuk setiap pesanan yang

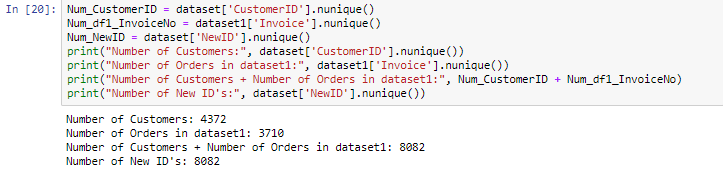
unik berdasarkan feature invoice.



1. Meng-*convert* data string yang ada pada invoice menjadi tipe data integer di kolom NewID



1. Mengetahui jumlah customerID keseluruhan



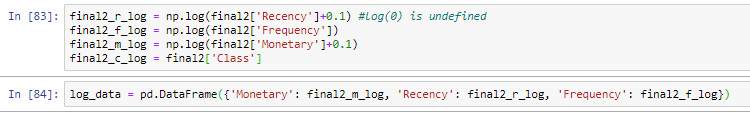
## Normalisasi

Normalisasi data adalah proses membuat beberapa variabel memiliki rentang nilai yang sama, tidak ada yang terlalu besar maupun terlalu kecil sehingga dapat membuat analisis statistik menjadi lebih mudah. Terdapat 3 metode untuk melakukan normalisasi data. Pada proyek ini kami menggunakan metode Min-Max.

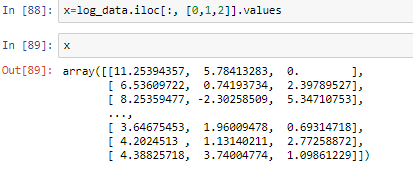
Cara kerjanya setiap nilai pada sebuah fitur dikurangi dengan nilai minimum fitur tersebut, kemudian dibagi dengan rentang nilai atau nilai maksimum dikurangi nilai minimum dari fitur tersebut.

1. Menggunakan 3 faktor  yaitu recency (Jumlah hari sejak pembelian terakhir), frequency

(Jumlah ‘proses transaksi), dan monetary (Jumlah total transaksi (kontribusi pendapatan).



1. Mengubah bentuk tiap value pada setiap faktor menjadi array agar dapat dinormalisasi.



1. Normalisasi data siap digunakan untuk di-clustering.

https://lh5.googleusercontent.com/seWlFf-xweS_DnzhqBe-dwxR4YiejF27yzpMcMrA0R6IV3KsgjcVOo2w9U5SFTBmDZ0wXnZhdKTkGqktdYHNebo_Pz-vucQ9BEbiZkmPU5kUBD1oJkrO6WXLMk7AhG49W8kEHaSp

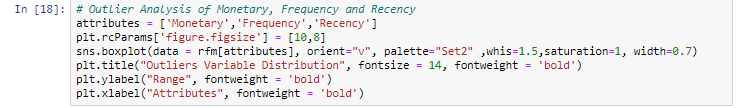
## Outlier

Outlier adalah kasus atau data yang memiliki karakteristik unik yang terlihat sangat berbeda jauh dari observasi-observasi lainnya dan muncul dalam bentuk nilai ekstrim baik untuk sebuah variabel tunggal atau kombinasi (Ghozali, 2011 : 41). Menurut (Ghozali. 2011 : 41) Ada empat penyebab timbulnya data outlier (1) kesalahan dalam meng-entri data, (2) gagal menspesifikasi adanya missing value dalam program komputer, (3) outlier bukan merupakan anggota populasi yang kita ambil sebagai sampel, tetapi (4) outlier berasal dari populasi yang kita ambil sebagai sampel, tetapi distribusi dari variabel dalam populasi tersebut memiliki nilai ekstrim dan tidak berdistribusi secara normal.

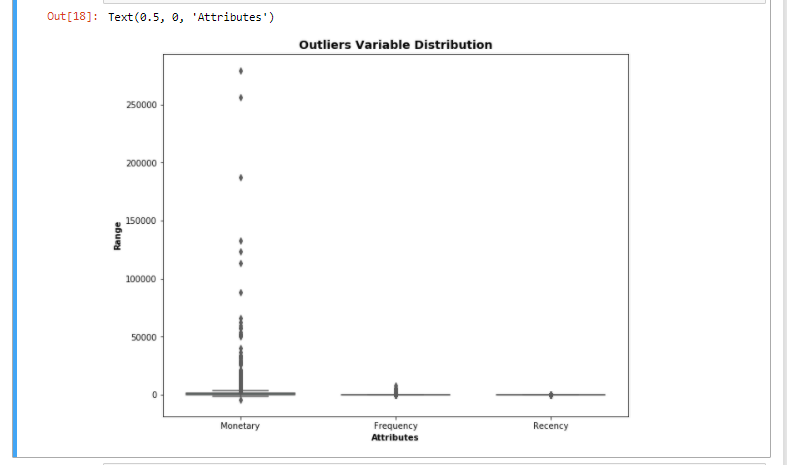
Boxplot merupakan ringkasan distribusi sampel yang disajikan secara grafis sebagai penggambaran bentuk distribusi data, ukuran tendensi sentral dan ukuran penyebaran (keragaman) data pengamatan. Selain itu boxplot juga bisa menunjukkan ada tidaknya nilai outlier dan nilai ekstrem di data pengamatan.

### **Mengecek Outlier**

Boxplot untuk memvisualisasikan Monetary, Frequency, dan Recency



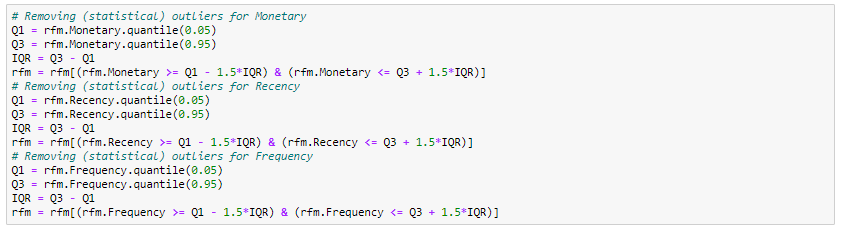
Output:



### **Menghandle Outlier**

Menghandle outlier pada masing-masing faktor.

* Jika data lebih kecil daripada data pada batas bawah maka data akan digantikan menjadi data pada batas bawah.
* Jika data lebih besar daripada data pada batas atas maka data akan digantikan menjadi data pada batas atas.

****

# BAB III EKSPLORASI DATA

Eksplorasi data adalah langkah pertama dalam analisis data dan biasanya melibatkan rangkuman karakteristik utama dari set data, termasuk ukurannya, pola awal dalam data dan atribut lainnya. Sebelum melakukan analisis pada data, Anda harus tahu berapa banyak kasus dalam kumpulan data, variabel apa yang dimasukkan, berapa banyak nilai yang hilang, dan hipotesis umum apa yang mungkin terjadi pada data tersebut. Eksplorasi awal dari kumpulan data dapat membantu menjawab pertanyaan-pertanyaan tersebut.

Eksplorasi data merupakan teknik yang cukup sederhana tetapi sangat powerfull dari segi manfaat. Dalam dunia bisnis, eksplorasi data merupakan hal yang sangat perlu dilakukan karena dapat membantu melihat demografi konsumen, demografi karyawan, dan demografi penjualan produk di berbagai daerah. Eksplorasi data tidak hanya dilakukan dengan mengukur frekuensi, ukuran pemusatan (mean,modus,median, max,min,dll), ukuran penyebaran ( standar deviasi, varians, dll) tetapi dapat juga dilakukan dengan melihat visualisasi dari data tersebut.

Pada explorasi data beberapa hal yang akan dilakukan yaitu:

* Melihat dimensi dataset
* Contoh datanya
* Ringkasan statistik semua atribut
* Mengelompokkan data berdasarkan variabel target

**3.1 Melihat / mengecek data**

* Yang pertama menampilkan dimensi datanya(jumlah row dan kolom) dengan menggunakan atribut **shape. Attribute.shape digunakan untuk mengetahui berapa baris dan kolom, hasilnya dalam format tuple (baris, kolom).**

https://lh4.googleusercontent.com/l-EiusGIbEKmgj3wyqe_MvHfqS1pcU9uRtP92UrjPyTUU5QCXSFA6oma5anek4c-4xsWJRdZqewgBdEkcAtglZWvTGboBSWi0QAI4q8ICxYYtMc7kXmj3QlEkbwMl2GJhCex25fy

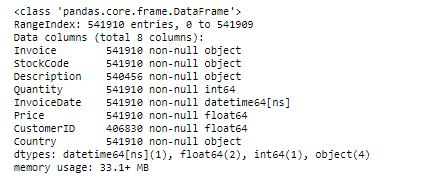
Output:

https://lh4.googleusercontent.com/X6zs60dgXcuahbZyvqhDcdwWYGVd1sUg6Ty0fX4dRvket9Fg5EuyBWavym1IbYoEvwGCi2X8VkQLcYE8e_t2LMwhUhkun-oH5u9qh828G0mD-Cqu4QxvcnwwSm6h0DQk7vGXJ4lK

Selanjutnya menampilkan nomor index beserta tipe datanya dengan menggunakan atribut **info().** Attribute **.info()** digunakan untuk mengecek kolom apa yang membentuk dataframe itu, data types, berapa yang non null, dll. Attribute ini tidak dapat digunakan pada series, hanya pada data frame saja**.** Berikut code yang digunakan untuk menampilkan informasi mengenai dataset tersebut

**https://lh3.googleusercontent.com/y0cLfKcGovuvsOvFP5a6_-R49mI2c8JnvoBQzg49_61EnvrOwidBH1RpjZ7_JXXgoclSnKCtHru2ZChBuVkeMx_3r4ytZXVabpAv9mUyo4BYMrfTSv8s4J45UwETtw1BUdbhcU3H**

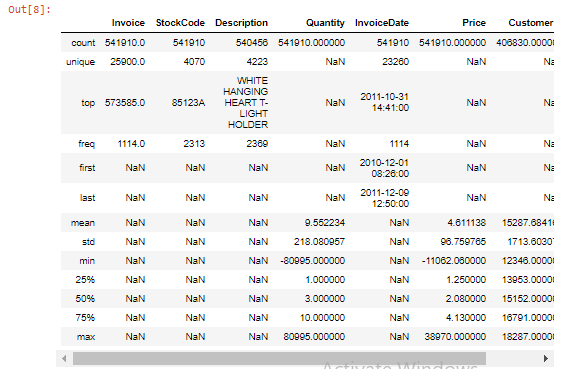
**output:**

****

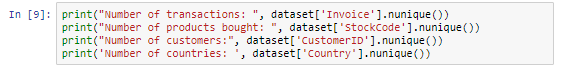
* Menampilkan atau menunjukkan rangkuman statistik seperti rata-rata, mean, median, dan lain-lain pada kolom tersebut. Atribut yang digunakan adalah **describe()**

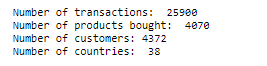
**https://lh6.googleusercontent.com/Aas0qxPzI8oKB01MBO9Na0HjoD3ZgE8-KgtvwMzLVxFsJEDd8kjJd59Lb_k8kVruih3N8XEYY8gm87kY8p5JcpFR6OKx6AZH6ZSX3B0pJh6F8Y5CWZlD-x45lP31C8XYiA9QAVr9**

**output:**

****

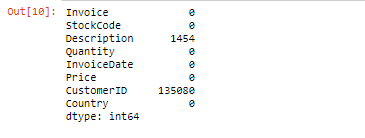
* Menampilkan jumlah dari nilai yang unik atau berbeda pada dataset tersebut. Atribut yang digunakan adalah .**nunique().** Berikut code yang digunakan dalam menampilkan jumlah nilai yang berbeda dari setiap atribut

****

****

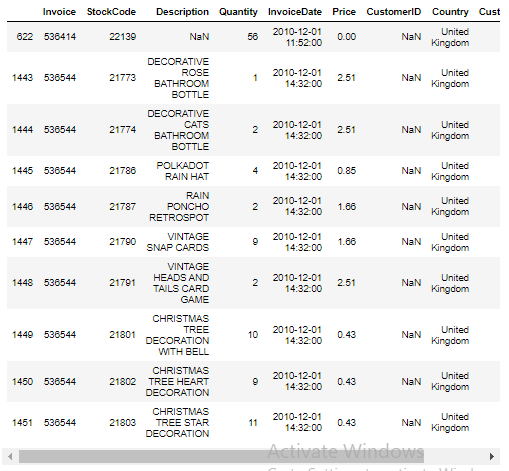
* Menampilkan jumlah nilai yang kosong(missing value) atau berbeda pada setiap atribut. Atribut yang digunakan adalah .**isnull()**

**https://lh3.googleusercontent.com/o9QDNohB0IV3Jl-is6rpzobmYm_pbUjzCR-JHq5S8EzFnkP-Lmyb-lmkA_6Y3qJY4jb7fdo0KwB1EW9SPWlBMrfPeJRUpIk__ZHjEpHM7C9FuaNGWPD5GQODppclAlqzPmQw41MW**

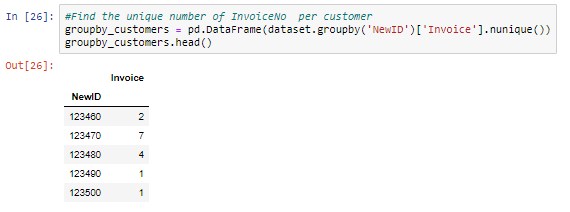
****

* Menampilkan n data pertama yaitu menggunakan atribut **.head(n).**

**https://lh4.googleusercontent.com/S9z5MJRo4NBorTn9sPq98I3knwZTEVIhDMO8mXCuFq6LtBEBcuHvB6VVw0b44EoEgOg3NgfnujHQnGp40gyGLpxwUlfTjhN26UkU8Iu892t4HI5jgc0GPR3vv6Rc915Kf2dUsV7o**

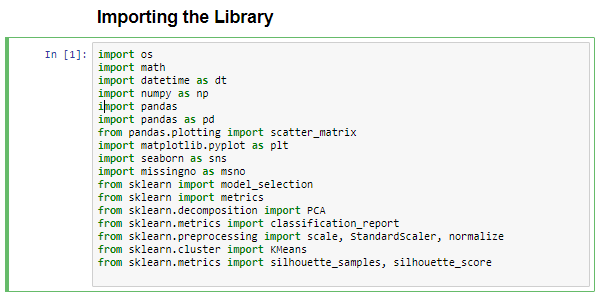
****

* Menggabungkan beberapa atribut yaitu menggunakan atribut **.groupby()**.

****

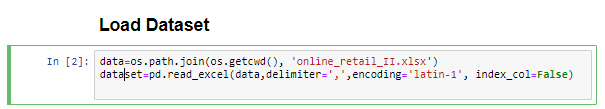
# BAB IV CODE

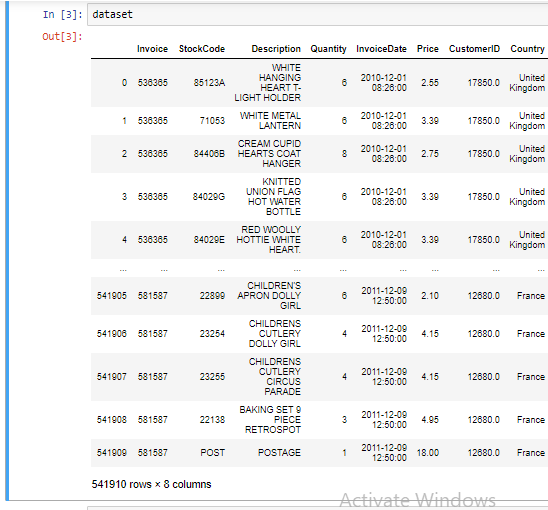
Langkah 1: Importing the Library   Hal pertama yang harus dilakukan adalah meng-*import* beberapa Library Python untuk kebutuhan dataframe, visualisasi dan clustering, dengan langkah-langkah sebagai berikut:



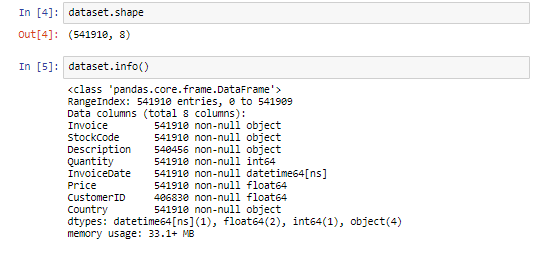
Step 2: Reading and Understanding the Data

Langkah selanjutnya, yaitu meng-input dataset. Dataset yang akan digunakan pada tulisan ini adalah dataset online\_retail\_II. Dataset OnlineRetail merupakan sekumpulan data transaksional dari toko-toko online/retail di UK yang terdaftar di suatu perusahaan retail online, dengan rentang waktu periode 1 Desember 2010 sampai dengan tanggal 9 Desember 2011. Berikut adalah langkah untuk meng-input dataset dan melihat detail informasi tentang dataset:





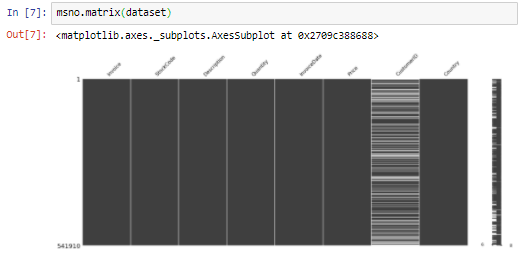
Berdasarkan output diatas, dapat disimpulkan bahwa dataset tersebut memiliki 8 atribut, diantaranya: Invoice, StockCode, Description, Quantity, InvoiceDate, Price, CustomerID dan Country. Untuk detail info dan statistika deskriptif terkait dataset tersebut dapat diperoleh pada output-output dibawah ini:



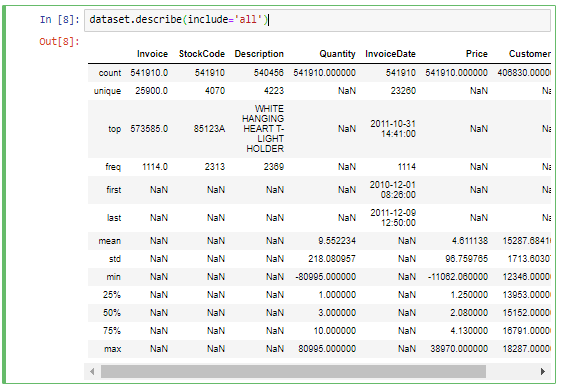
Langkah 3: Data Cleansing

Langkah selanjutnya yaitu akan dilihat komposisi persentase missing value yang terdapat pada dataset, dengan langkah sebagai berikut:

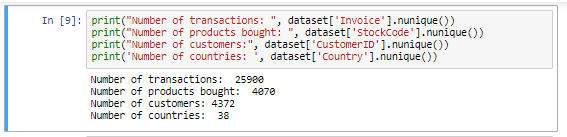
* Pertama akan menampilkan grafik pada setiap atribut yang mengalami missing value. Pada grafik tersebut dapat dilihat yang mengalami missing value yaitu pada atribut Customer Id



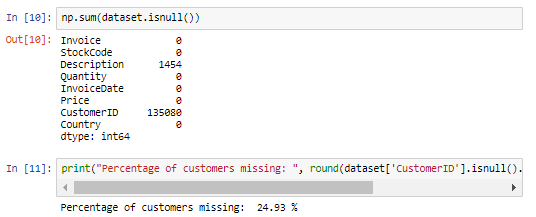
* Kemudian menampilkan secara keseluruhan rangkuman statistik



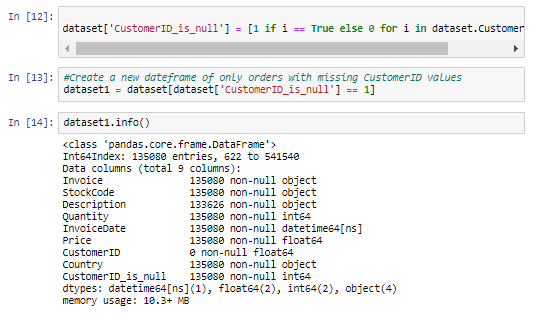
* Menampilkan jumlah nilai yang berbeda dari setiap atribut



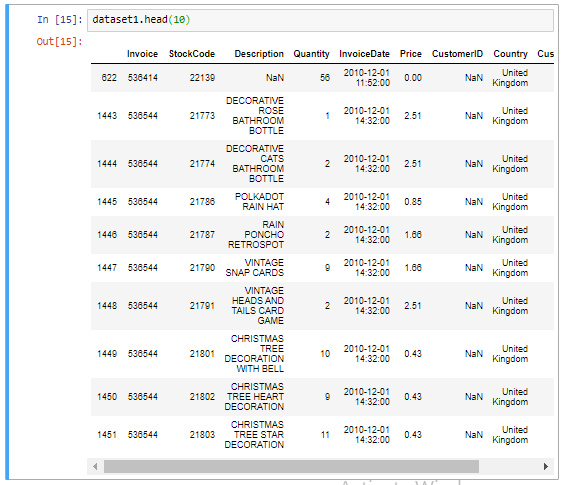
* Untuk mengetahui atribut yang mengalami *missing value* dapat juga menggunakan atribut .isnull() selain menggunakan grafik. Kode di bawah ini juga dapat menampilkan jumlah nilai yang mengalami *missing value*. Kemudian akan menampilkan jumlah persentase atribut yang tidak memiliki nilai dan *missing value*.



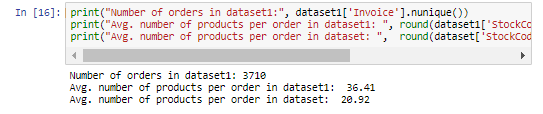
* Untuk mengatasi *missing value* maka data pada dataset online\_retail\_II akan dihapus dengan menggunakan if… else..

**

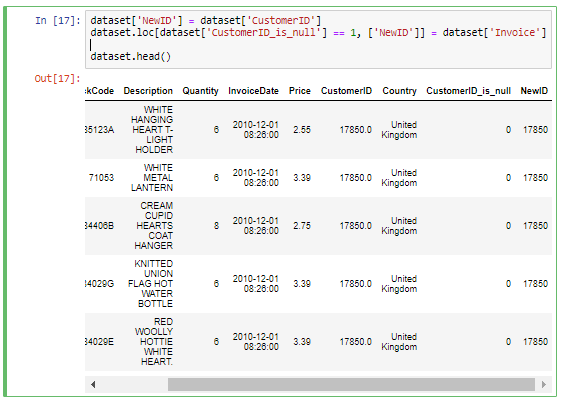
* Menampilkan data online-retail\_II 10 data paling atas setelah yang disimpan di dalam variabel yang baru yaitu dataset1



* Menampilkan jumlah nilai yang berbeda atau unik dari atribut invoice setelah missing valuenya sudah dihapus. Kemudian menampilkan nilai rata-rata dari atribut StockCode dibagi dengan nilai atribut invoice sebelum dan setelah penghapusan *missing value*. Pada code di bawah ini dapat dilihat nilai nilai rata-rata dataset setelah penghapusan *missing value* lebih tinggi dibandingkan dengan nilai rata-rata dataset ketika terdapat *missing value*.

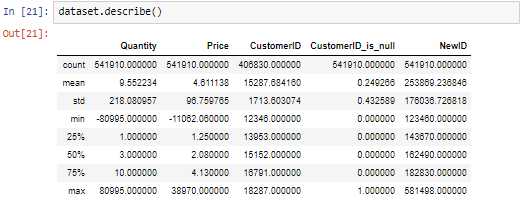


* Menggunakan nilai di kolom Invoice akan menjadi pendekatan yang paling mudah. Membuat kolom ID pelanggan baru bernama NewID dengan nomor faktur mengisi nilai yang hilang. Kemudian kami akan menambahkan jumlah pesanan unik di dataset1 dan ke jumlah nilai unik di ID Pelanggan dan melihat apakah itu sama dengan jumlah nilai unik di NewID. Ini akan memeriksa apakah ada nilai baru yang cocok dengan nilai yang ada di kolom dan memastikan kami tidak menambahkan lebih banyak pesanan ke pelanggan yang sudah ada.

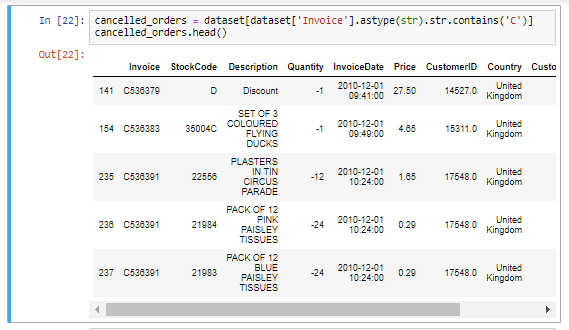


Langkah 4: Analisis Eksplorasi Data

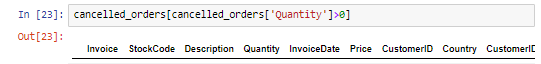
* Sebelumnya kita akan mengeksplorasi data menggunakan atribut **.describe()** dimana akan menampilkan ringkasan statistika pada dataset tersebut seperti mean, median, dan lain sebagainya. Dapat dilihat bahwa terdapat nilai yang negatif pada kolom Quantity dan Price. terdapat nilai negatif sepertiny terdapat pesanan yang dikembalikan ataupun dibatalkan



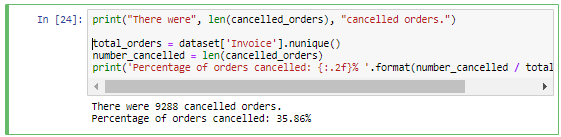
* Seperti yang dijelaskan di atas jika terdapat awalan nomor invoice “C” maka pesan tersebut dibatalkan



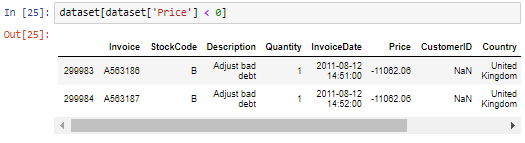
* Karena tidak ada barang yang dikembalikan  setelah di filter berarti nilai negatif tersebut adalah pesanan yang dibatalkan



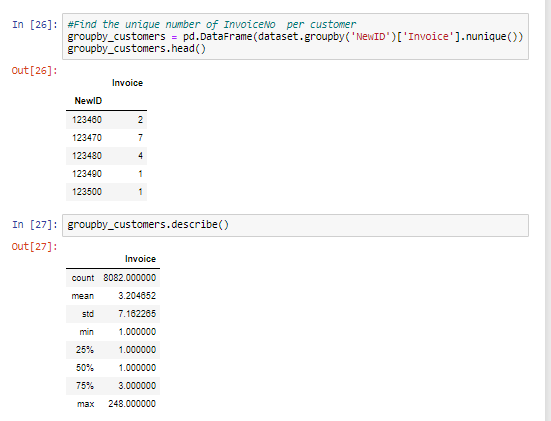
* Terdapat 9288 pesanan dibatalkan



* Selanjutnya maksud dari nilai negatif dari kolom price adalah  customer yang tidak pernah membayar pesanan atau dana yang tidak mencukupi



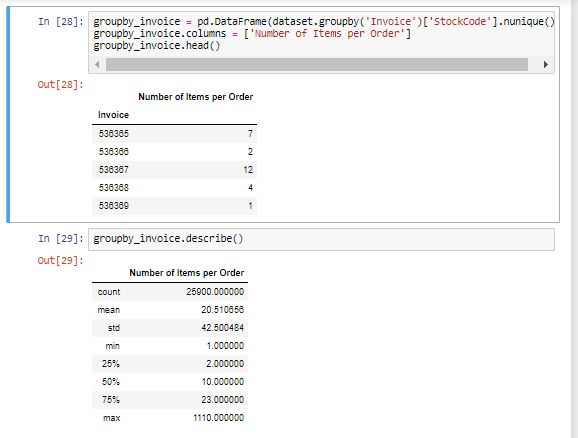
* Kode di bawah ini berfungsi untuk menampilkan jumlah rata-rata pesanan setiap customer



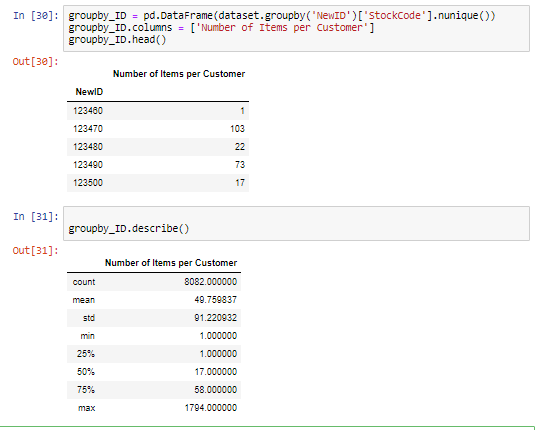
Jumlah rata-rata pesanan per pelanggan adalah 3.

Kami menemukan dalam statistik deskriptif di atas bahwa pelanggan membeli jumlah rata-rata sekitar 10 per produk.

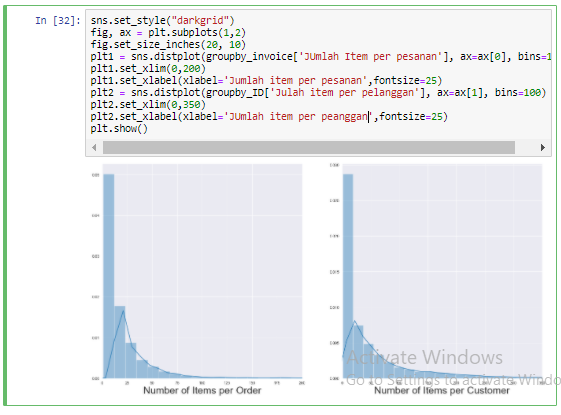
* Menampilkan jumlah produk unik per pelanggan



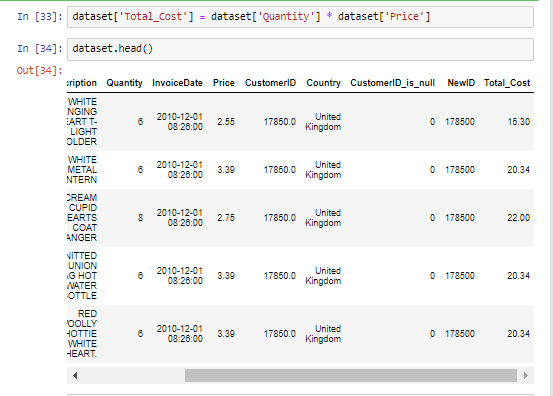
* menampilkan jumlah item per pelanggan



* Visualisasi data yaitu jumlah item per pesanan dan jumlah item per pelanggan. Jumlah rata-rata barang per pesanan adalah 20,5 dan jumlah rata-rata barang per pelanggan adalah 50.



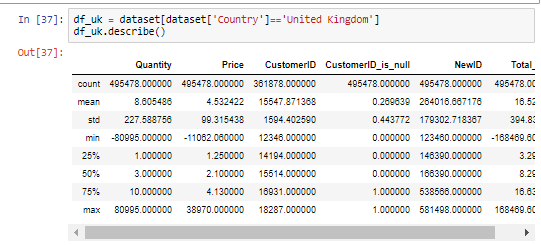
* Jumlah pesanan per negara
* Kode di bawah ini menghitung total biaya yaitu perkalian antara Quantity dengan price



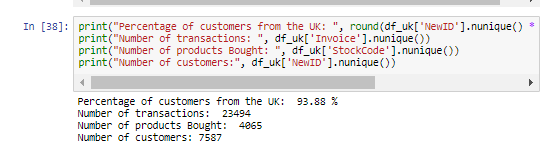
* Menampilkan grafik antara jumlah pelanggan dengan county. Dapat dilihat negara United kingdom merupakan negara yang paling banya pelanggannya



* Eksplorasi Pasar United Kingdom
* Menampilkan rangkuman statistik dari  negara United Kingdom

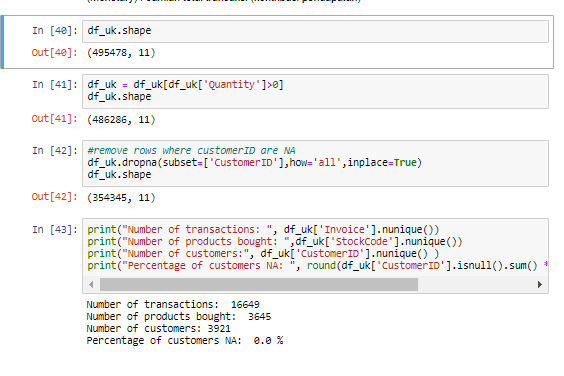


* Menampilkan jumlah persentasi pelanggan di negara United kingdom dan jumlah nilai atribut invoice, StockCode, dan NewId.

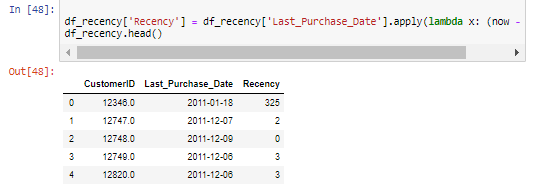
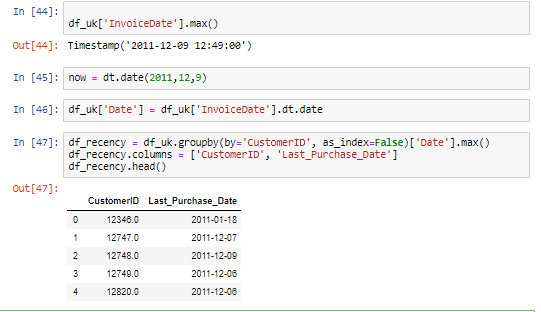


* Langkah 5 : Mempersiapkan Data United Kingdom

Untuk memperoleh segmentasi dari atribut Customers, yang didasarkan pada 3 faktor berikut: R (Recency) : Jumlah hari sejak pembelian terakhir F (Frequency) : Jumlah ‘proses transaksi’ M (Monetary) : Jumlah total transaksi (kontribusi pendapatan)

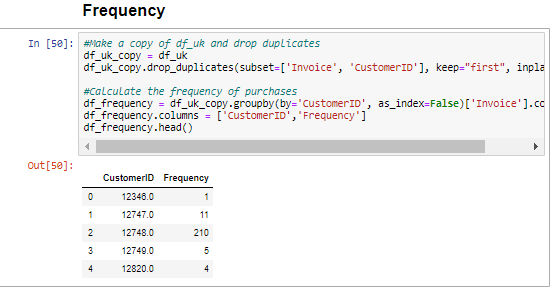


* Pertama kita akan menghitung recency, kita perlu memilih tanggal sebagai acuan untuk mengevaluasi berapa hari yang lalu pelanggan melakukan pembelian terakhir.



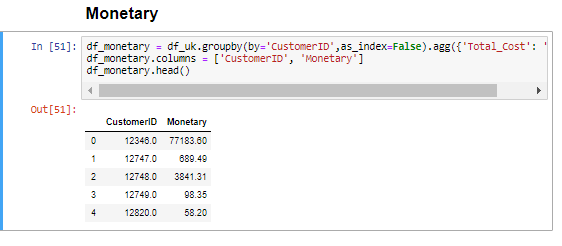
* Menghitung Frequency

Untuk menghitung berapa kali pelanggan membeli sesuatu, kita perlu menghitung berapa banyak faktur yang dimiliki setiap pelanggan.

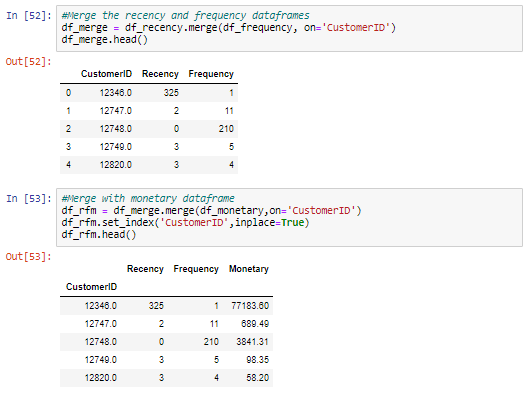


* Menghitung Monetary

Nilai moneter dihitung dengan menjumlahkan biaya pembelian pelanggan. Kami sudah melakukan ini ketika kami membuat kolom Total\_Cost



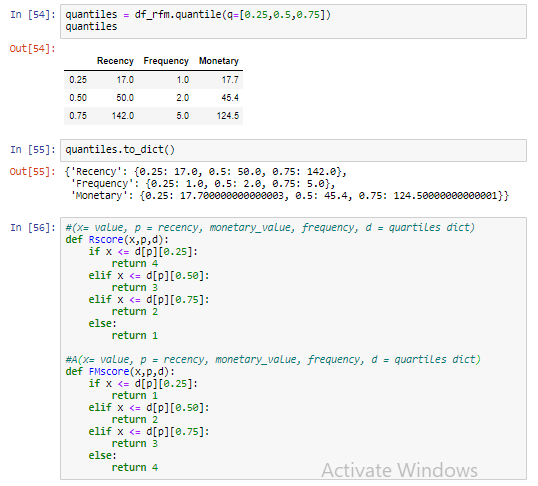
* Menggabungkan tabel Recency, frequency, dan moneter



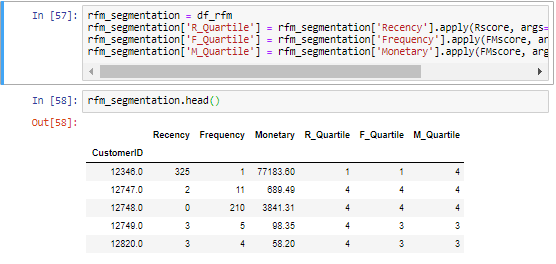
Jika kita melihat Customer ID: 123460, pembelian terakhirnya adalah 325 hari yang lalu (dari tanggal referensi), mereka hanya membeli dari kami satu kali, dan mereka menghabiskan total $ 77.183.60

* Menghitung nilai RFM

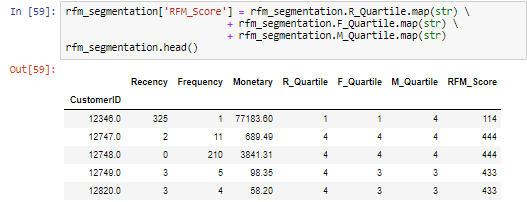
Cara termudah untuk membuat segmen pelanggan dari model RFM adalah dengan menggunakan Kuartil. Kami akan memberikan skor dari 1 hingga 4 untuk setiap kategori (Keterkinian, Frekuensi, dan Moneter) dengan 4 sebagai nilai tertinggi / terbaik. Skor RFM akhir dihitung dengan menggabungkan nilai RFM individu.



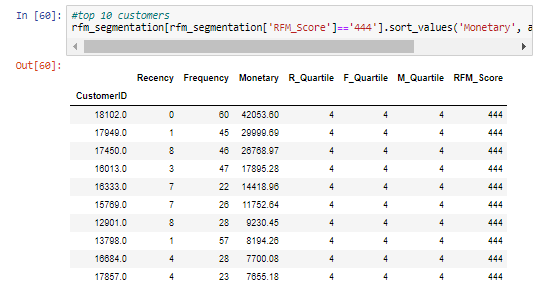
Kami akan membuat dua kelas segmentasi karena recency tinggi (lebih banyak hari sejak pembelian terakhir) buruk, sedangkan frekuensi tinggi dan nilai bulanan bagus.



Sekarang setelah kami menilai setiap pelanggan, kami akan menggabungkan skor untuk segmentasi.

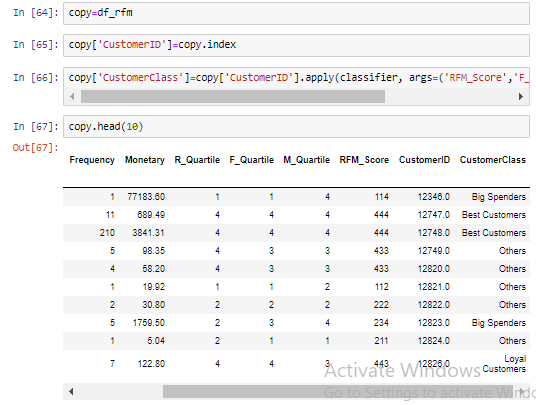


Skor 4 menunjukkan pelanggan berada di persentil ke-75 untuk kategori itu. Mari kita lihat siapa pelanggan terbaik kita.

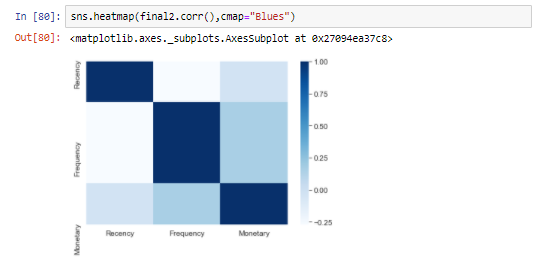
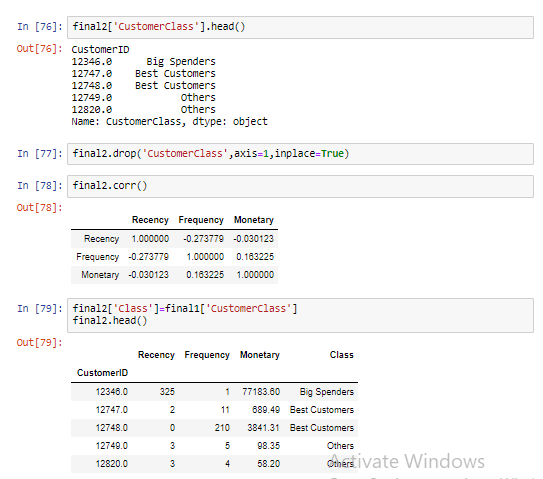


* Sekarang kami dapat memfilter kumpulan data kami berdasarkan setiap kategori dan menargetkan setiap grup secara berbeda.

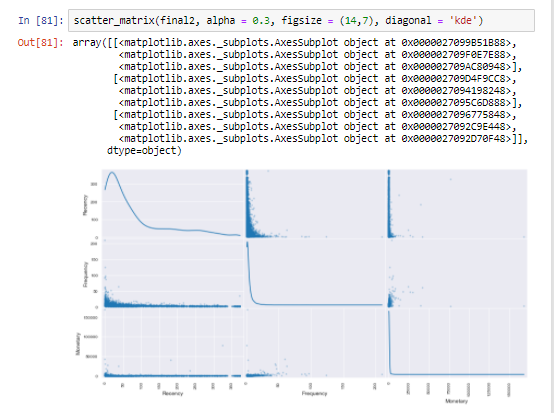




* Langkah 6: Mengimplementasikan K-Means Clustering pada variabel RFM

Melihat peta panas ini, kita melihat bahwa ada korelasi negatif antara Recency: Frekuensi dan Recency: Moneter, tetapi ada korelasi positif antara Frekuensi: Moneter

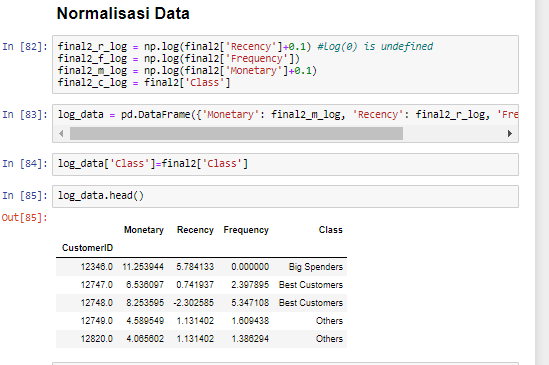
* Memvisualisasikan Distribusi Fitur



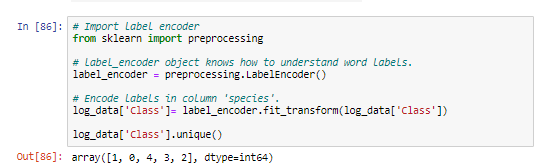
Perhatikan bahwa ada distribusi miring untuk 3 variabel dan ada outliers.

Karena algoritme pengelompokan memerlukan distribusi normal, diperlukan normalisasi data.

* Normalisasi Data



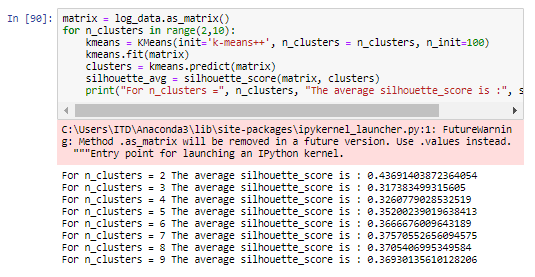
* Label Encoder





* Silhouette Analysis

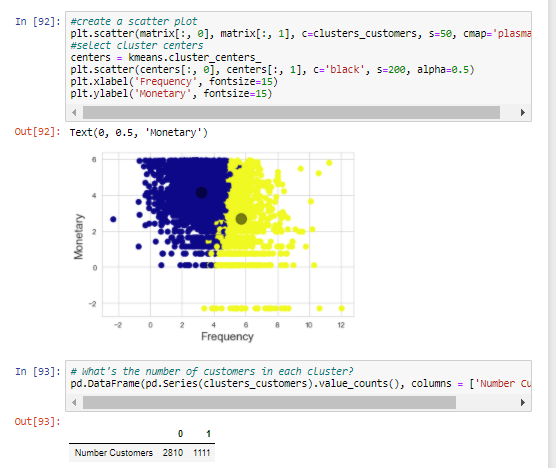
Secara umum, algoritma dari Silhouette analysis yaitu mengukur seberapa dekat (baik) setiap titik pada sebuah cluster dengan titik-titik data lain di clusternya. Semakin tinggi nilai rata-rata dari silhouette, menunjukkan suatu peng-cluster-an yang baik.



Berdasarkan output tersebut, dapat disimpulkan bahwa untuk n\_clusters=2 menghasilkan nilai silhouette yang tinggi.

* Visualisasi Cluster

Kluster kuning memiliki pusat massa di sekitar (5.9, 2.8) dan mewakili "pelanggan bernilai rendah". Kluster biru tua memiliki pusat di sekitar (3,8, 4.2) dan mewakili "pelanggan bernilai tinggi".



# BAB V PEMBAHASAN

Tujuan dari penyelesaian permasalahan ini, yaitu untuk melakukan segmentasi customer berdasarkan RFM (Recency, Frequency dan Monetary), sehingga dapat menargetkan segmentasi customer secara lebih efisien. Permasalahan ini akan dimulai dengan analisis RFM dan kemudian melengkapi dengan analisis prediktif menggunakan K-Means Clustering Algoritma. Analisis RFM (Recency, Frequency, Monetary) adalah teknik segmentasi pelanggan untuk menganalisis nilai pelanggan berdasarkan perilaku pembelian sebelumnya.



## Analisis Output

Berdasarkan hasil analisis dan pembahasan pengelompokan data, maka Analisis cluster membuat pengelompokan objek berdasarkan jarak antara pasangan objek. Jarak merupakan ukuran yang digunakan untuk mengukur kemiripan dari suatu objek.  Seperti yang telah disebutkan sebelumnya bahwa, tujuan analisis yang dilakukan pada case ini adalah untuk memperoleh segmentasi dari atribut **Customers,**yang didasarkan pada 3 faktor berikut:

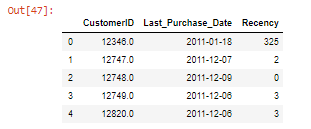
* R (Recency) : Jumlah hari sejak pembelian terakhir
* F (Frequency) : Jumlah ‘proses transaksi’
* M (Monetary) : Jumlah total transaksi (kontribusi pendapatan)

Case yang digunakan untuk menyelesaikan masalah ini menggunakan tahapan KDD.

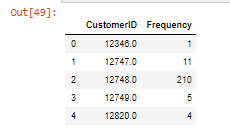
1. **Data selection**

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional. Atribut yang akan diseleksi tidak secara langsung dari dataset tersebut tetapi hasil penggabungan dari bebrapa atribut yang dolah untuk menghasilkan atribut yang baru

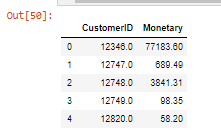
Pertama kita akan menghitung recency, kita perlu memilih tanggal sebagai acuan untuk mengevaluasi berapa hari yang lalu pelanggan melakukan pembelian terakhir.



Untuk menghitung berapa kali pelanggan membeli sesuatu, kita perlu menghitung berapa banyak faktur yang dimiliki setiap pelanggan.

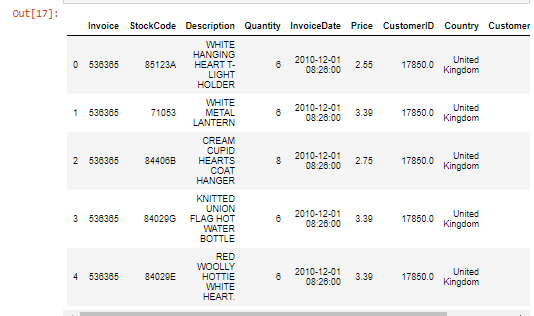


Nilai moneter dihitung dengan menjumlahkan biaya pembelian pelanggan. Kami sudah melakukan ini ketika kami membuat kolom Total\_Cost



**2. Pre-processing / cleaning**

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data. Pada data tersebut terdapat *missing value* yaitu pada atribut Customer Id dan Price. Maka nilai yang mengalami *missing value* akan dihapus.



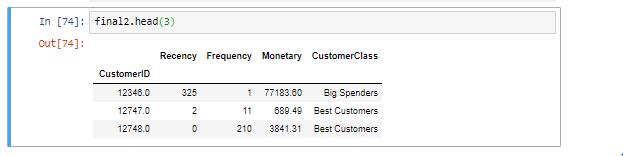
**3. Transformation**

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

Setelah melakukan analisis terhadap jumlah RFM maka kami dapat memfilter kumpulan data kami berdasarkan setiap kategori dan menargetkan setiap grup secara berbeda.

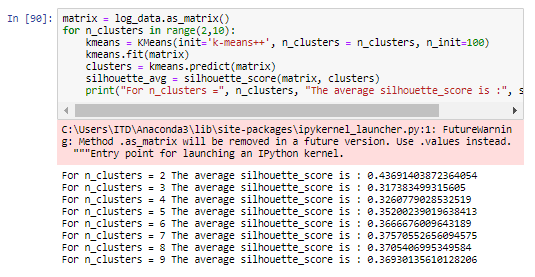
Berikut output yang dihasilkan dan dikelompokkan setiap target. Terdapat 6 kategori dalam menentukan jenis pelanggan yaitu:

* Best Customers
* Loyal Customers
* Big Spenders
* Almost Lost
* Lost Customers
* Lost Cheap Customers



**4. Data mining**

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan. Pada case ini algoritma yang digunakan yaitu menggunakan algoritma K-Means Clustering dan untuk menghitung jaraknya menggunakan *Silhouette analysis*yaitu mengukur seberapa dekat (baik) setiap titik pada sebuah cluster dengan titik-titik data lain di clusternya. Semakin tinggi nilai rata-rata dari silhouette, menunjukkan suatu peng-cluster-an yang baik.



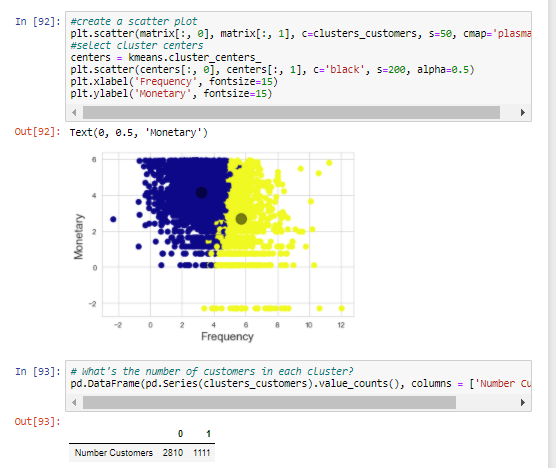
Berdasarkan output tersebut, dapat disimpulkan bahwa untuk n\_clusters=2 menghasilkan nilai silhouette yang tinggi.

**5. Interpretation / evalution**

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya

## Hasil

Kluster kuning memiliki pusat massa di sekitar (5.9, 2.8) dan mewakili "pelanggan bernilai rendah". Kluster biru tua memiliki pusat di sekitar (3,8, 4.2) dan mewakili "pelanggan bernilai tinggi".



Meskipun kami tidak mendapatkan dua kluster yang dipisahkan dengan jelas, kami dapat membuat model yang dapat mengklasifikasikan pelanggan baru ke dalam grup "bernilai rendah" dan "bernilai tinggi". Umumnya, jika pelanggan hanya bertransaksi dengan kami beberapa kali, mereka setidaknya harus berada di persentil ke-50 teratas dalam pengeluaran moneter agar dianggap sebagai "pelanggan bernilai tinggi". Penugasan cluster berantakan, yang mungkin disebabkan oleh outliers yang tidak dihapus.

# REFERENSI

Tukar), Y. (2015). Ukuran Perusahaan (UP), Profitabilitas (ROE), Kinerja Lingkungan (KL), Leverage (DER), dan Islamic Social Reporting (ISR) perusahaan manufaktur yang terdaftar di Jakarta Islamic Index (JII) selama periode 2012–2015. 44(6), 36–50.

<https://iniakunhuda.medium.com/data-preprocessing-menangani-data-kosong-di-dataset-786a2d8c9fa7>

<https://www.kajianpustaka.com/2017/09/data-mining.html>

<https://www.datacamp.com/community/tutorials/k-means-clustering-python?utm_source=adwords_ppc&utm_campaignid=1455363063&utm_adgroupid=65083631748&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=278443377092&utm_targetid=aud-299261629614:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=20450&gclid=CjwKCAiArbv_BRA8EiwAYGs23De3YSXOjjNpXUh_X70DiMzw5uL5vQQyXlh9lbbz0kUjkKVfI-Rc7hoCKsIQAvD_BwE>