Unsupervised Machine learning With Clustering Method

Cases Study Online Retail Clean

By: Ferdiansyah

Business Understanding

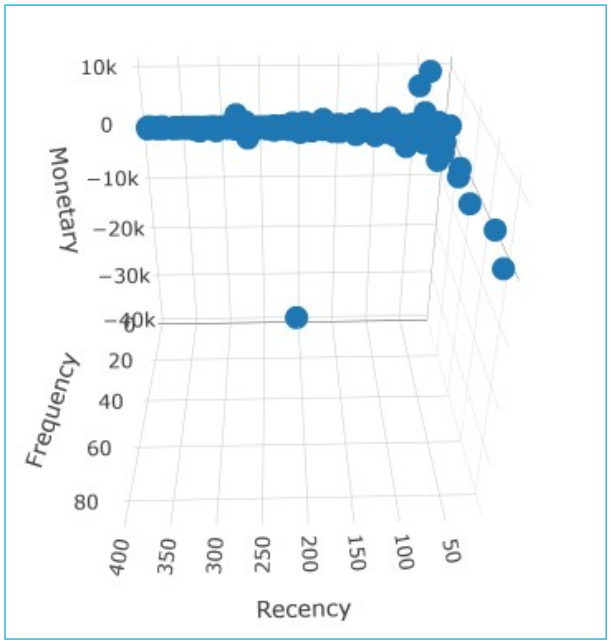
- Online Retail Clean merupakan sebuah perusahaan jasa retail
- Data set dari data transaksi berisi data customer ID, Frequency dan Monetary

Data Understanding

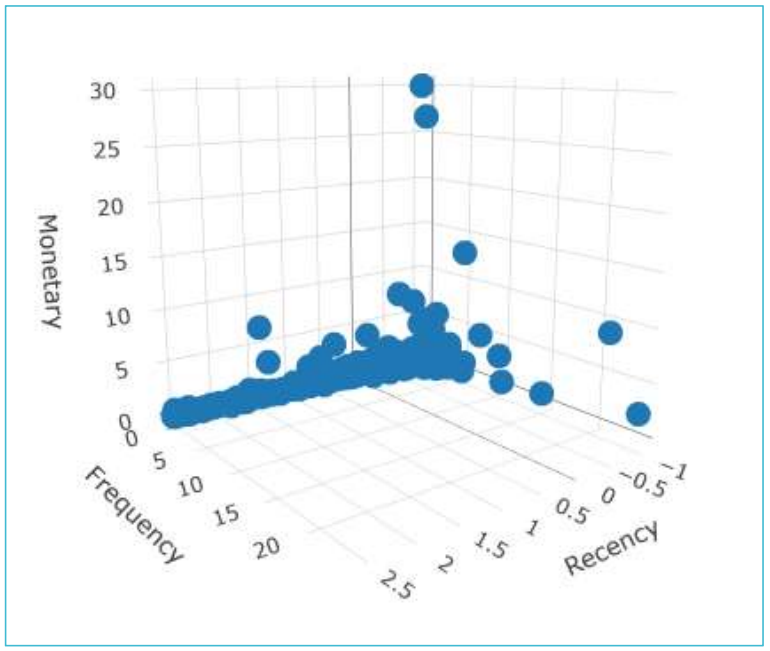
- Cutomer ID : adalah ID unik yang dimiliki oleh masing-masing pelanggan
- Recency : adalah jumlah hari dari hari terakhir customer membeli (satuan hari)
- Frequency : adalah jumlah pembelian yang dilakukan oleh customer (satuan kali)
- Monetary : adalah total nilai pembelian dari customer (satuan Dollar)

Data Preparation

- Cek Data Awal
- Membuang data monetary negative
- Scale Data
  - Pada metode cluster sangat dipengaruhi jarak antar point pada setiap variable. Dikarenakan setiap variable memiliki satuan dan skala yang berbeda, maka perlu dilakukan scaling dari setiap nilai pada variable agar memiliki skala yang sama
  - scale dilakukan dengan menghitung z score dari masing2 nilai variable



Sebelum Data Preparation



Sesudah Data Preparation

Modeling

Mencari Jumlah Kluster dan Jenis Algoritma Terbaik

- Dilakukan pengujian untuk 3 Jenis Algoritma dengan variasi jumlah kluster dari 2-6 kluster dengan fungsi clValid().
- Dengan menggunakan pengukuran internal, kualitas kluster di uji berdasarkan Dunn Index, Connectivity dan Silhoutte coeffisien

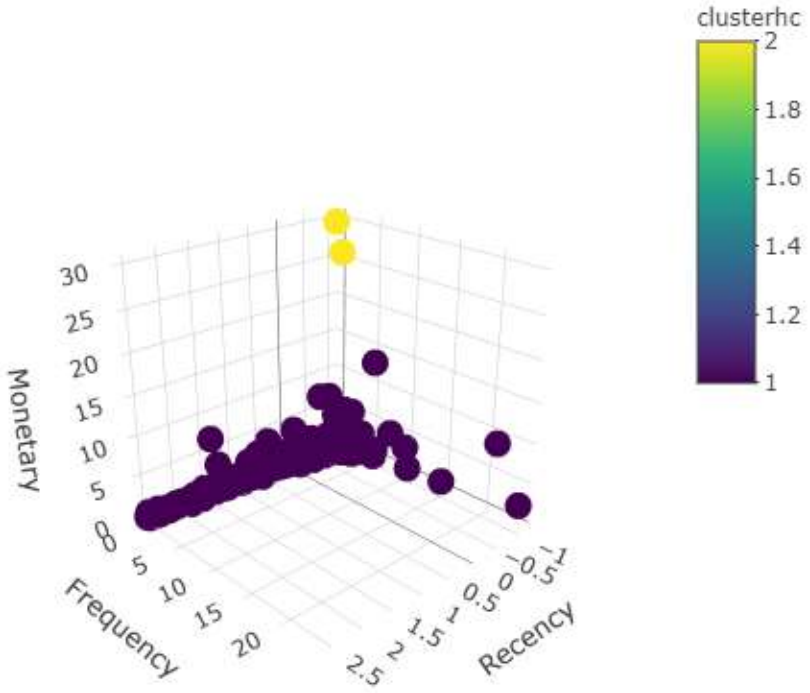
```
> clmethods <- c("hierarchical","kmeans","pam")
> intr <- clValid(x[,5:7], nClust = 2:6, clMethods = clmethods,validation = "internal", maxitems =
2350 ,metric = "euclidean",method = "complete")
> summary(intr)
```

Modeling

```
Clustering Methods:
 hierarchical kmeans pam

Cluster sizes:
 2 3 4 5 6

Validation Measures:
                                  2        3        4        5        6

hierarchical Connectivity    3.8579  10.4567  13.3151  22.8099  25.0266
             Dunn            0.5741   0.2713   0.2808   0.2056   0.2504
             Silhouette      0.9478   0.9077   0.8763   0.8311   0.8039
kmeans       Connectivity   14.4048  12.6524  65.1853  25.7480 102.0024
             Dunn            0.0611   0.1926   0.0005   0.0891   0.0017
             Silhouette      0.9111   0.9066   0.5582   0.7778   0.5400
pam          Connectivity   48.7032  79.6806  99.9222 156.5710 161.3635
             Dunn            0.0003   0.0005   0.0003   0.0003   0.0002
             Silhouette      0.5291   0.3844   0.4448   0.3302   0.3058

Optimal Scores:

             Score  Method       Clusters
Connectivity 3.8579 hierarchical 2
Dunn         0.5741 hierarchical 2
Silhouette   0.9478 hierarchical 2
```

Modeling

- Dari data tersebut maka diketahui bahwa algoritma terbaik adalah menggunakan algoritma Hierarki dengan jumlah kluster 2



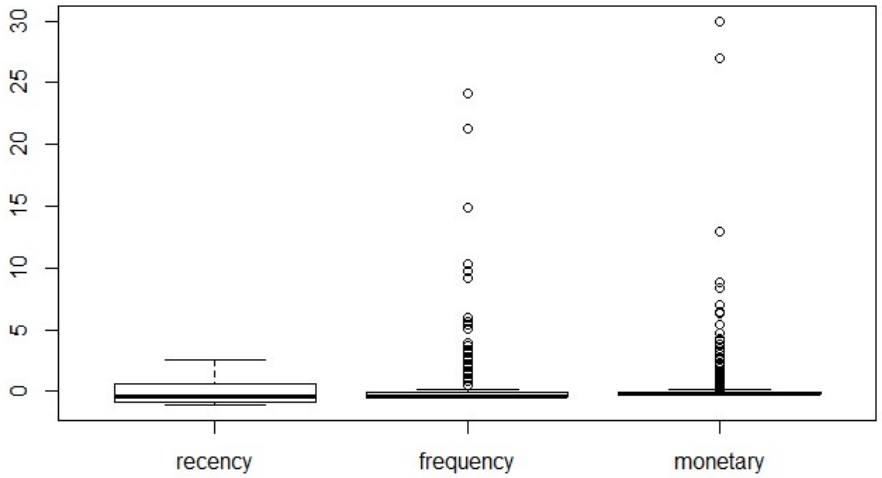
Plot 3D hasil data clustering

Evaluation

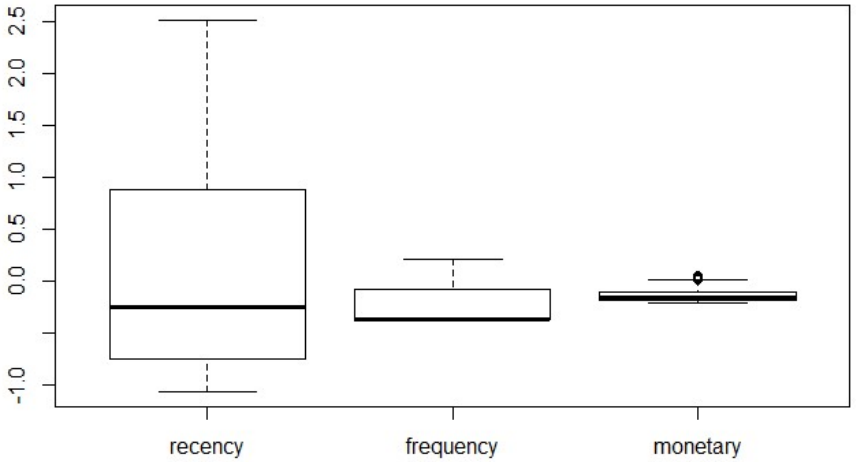
- Dari Model yang dihasilkan diketahui hanya terdapat dua data pada kluster kedua
- Dari model tersebut juga tidak dapat ditarik kesimpulan yang menggambarkan kondisi customer dari bisnis online retail clean
- Perlu dilakukan pengolaan ulang agar model dapat memberikan gambaran kelompok customer dari bisnis online retail clean

Data Preparation II

- Cek Outlier
- Membuang Outlier



Data sebelum data preparation II. Banyak terdapat outlier pada data frequency dan data monetary



Data setelah data preparation II. Outlier pada data frequency dan data monetary telah dihilangkan

Modeling II

Mencari Jumlah Kluster dan Jenis Algoritma Terbaik dengan data yang telah dibersihkan dari ourlier

- Dilakukan pengujian untuk 3 Jenis Algoritma dengan variasi jumlah kluster dari 2-6 kluster dengan fungsi clValid().
- Dengan menggunakan pengukuran internal, kualitas kluster di uji berdasarkan Dunn Index, Connectivity dan Silhoutte coeffisien

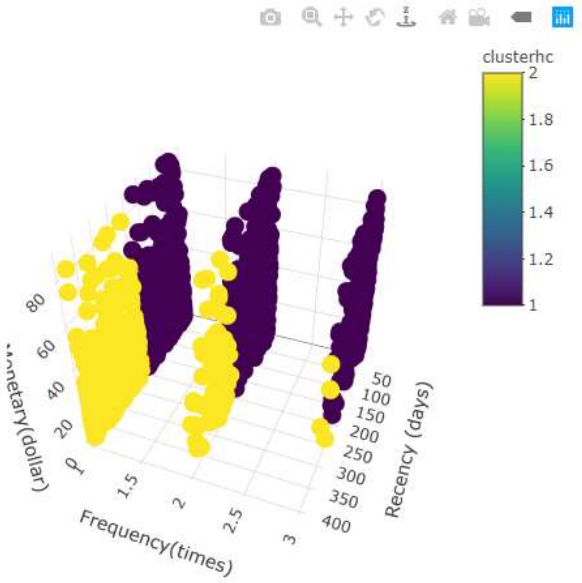
```
> clmethods <- c("hierarchical","kmeans","pam")
> intr <- clValid(xclean2[,5:7], nClust = 2:6, clMethods = clmethods,validation = "internal",
maxitems = 2350 ,metric = "euclidean",method = "complete")
> summary(intr)
```

Modeling II

```
Clustering Methods:
 hierarchical kmeans pam

Cluster sizes:
 2 3 4 5 6

Validation Measures:
                                   2       3       4       5       6

hierarchical Connectivity  11.6921 19.2964 30.0897 45.0107 52.6056
             Dunn           0.0087  0.0146  0.0146  0.0085  0.0098
             Silhouette     0.6365  0.5868  0.5004  0.4313  0.4104
kmeans       Connectivity  13.4782 33.8901 52.7087 67.4234 65.0996
             Dunn           0.0132  0.0081  0.0074  0.0084  0.0119
             Silhouette     0.6476  0.5757  0.4814  0.4439  0.4232
pam          Connectivity  20.6921 34.9595 51.4635 62.3433 50.1964
             Dunn           0.0019  0.0061  0.0021  0.0021  0.0091
             Silhouette     0.6457  0.5740  0.4503  0.4367  0.4481

Optimal Scores:

             Score   Method       Clusters
Connectivity 11.6921 hierarchical 2
Dunn          0.0146 hierarchical 4
Silhouette    0.6476 kmeans       2
```

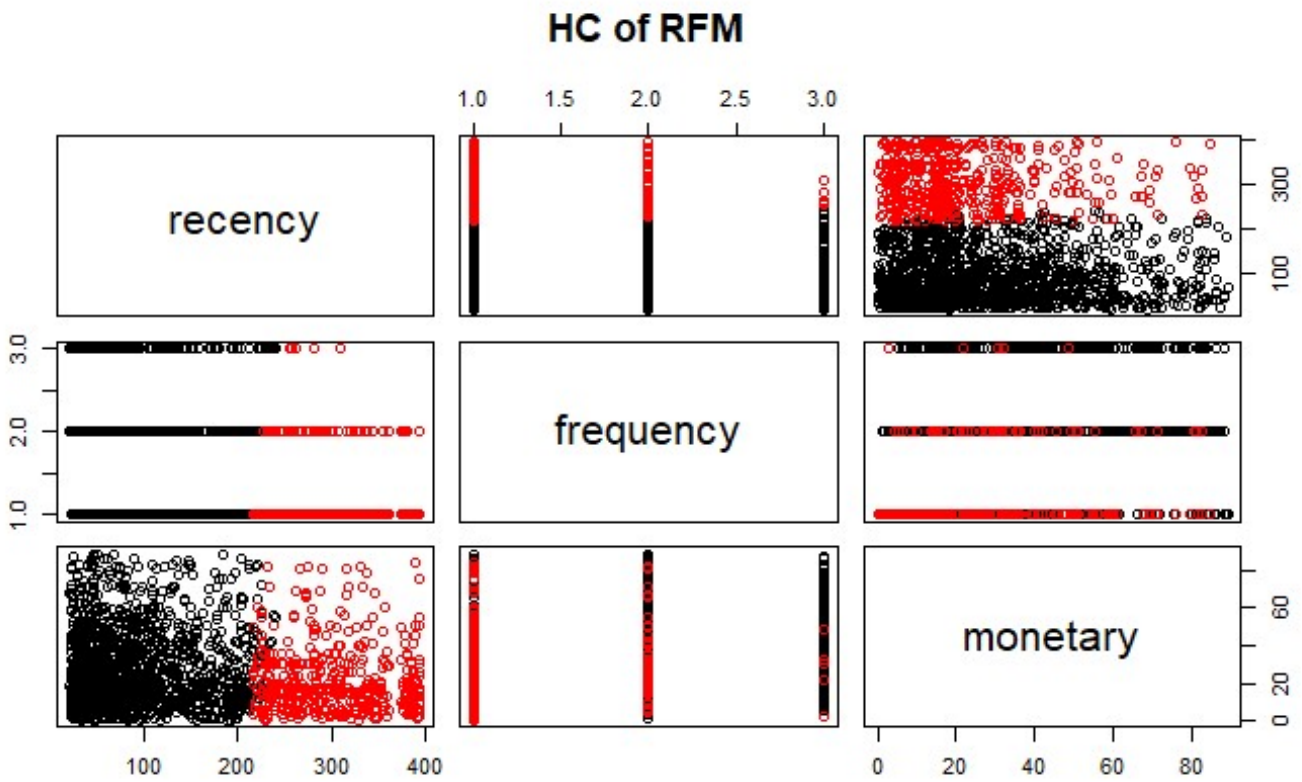
Modeling II

- Dari data tersebut maka diketahui bahwa algoritma terbaik adalah menggunakan algoritma Hierarki dengan jumlah kluster 4



Plot 3D hasil data clustering

Modeling II


HC of RFM
recency
frequency
monetary
1.0
1.5
2.0
2.5
3.0
100
200
300
400
0
20
40
60
80

Kesimpulan

- Dari data transaksi pelanggan diketahui terdapat 2 kelompok pelanggan
- 2 kelompok pelanggan lebih di bagi berdasarkan kekinian dari mereka melakukan transaksi
- nilai Monetary terdistribsi rapat sehingga dianggap satu kelompok. sedangkan nilai frequency distribusinya terlalu kecil hanya terdiri 3 nilai sehingga dianggap satu kelompok. Pembagian kelompok lebih kepada mempertimbangkan nilai recency
    - Kelompok 1 adalah pembeli yang membeli kurang dari 250 hari
    - KeLompok 2 adalah pembeli yang membeli lebih dari 250 hari