Yohanes Marakub Efruan - 94072

Knowledge

1. Apa yang Anda ketahui tentang Big Data Analytics dan kenapa disebut BigData?

Big data analytics merupakan proses mengumpulkan, mengorganisasikan dan menganalisa sekumpulan besar data (big data) untuk mendapatkan pola-pola dan informasi yang bermanfaat.

Big data analytics tidak hanya membantu untuk memahami informasi yang terkandung di dalam data tapi juga membantu untuk mengidentifikasi data yang paling penting untuk keputusan bisnis saat ini dan masa datang.

Ada beberapa tantangan dalam melakukan Big Data Analytics :

1. Tantangan pertama adalah Bagaimana memecah data untuk dapat memungkinkan mengakses semua data organisasi yang disimpan di tempat penyimpanan yang berbeda dan bahkan juga disimpan pada sistem yang berbeda.
2. Tantangan besar kedua adalah Membuat platform yang dapat menarik unstructured data semudah menarik structured data. Volume data ini begitu besar sehingga sulit untuk memprosesnya menggunakan database dan metode perancangan software yang tradisional

Terdapat dua teknik utama untuk menganalisis big data : *the store and analyze approach*, dan *the analyze and store approach*. Sesuai Namanya, Kedua Teknik tersebut dibedakan berdasarkan tahap eksekusinya.

Suatu data disebut sebagai Big data karena merupakan kumpulan dari data - data dalam skala besar sehingga tidak dapat dikelola dengan cara biasa. Suatu Big Data juga mengandung 5V, yaitu :

- Volume : memiliki jumlah data yang besar
- Velocity : memiliki waktu penerimaan dan pemrosesan data yang sangat cepat
- Variety : datanya variative, dapat berupa *structured*, *semi-structured* maupun *unstructured* data
- Value : seberapa besar impact insight yang bisa kita ambil dari Big Data tersebut
- Veracity : mengacu pada data *quality, integrity, credibility* dan *accuracy*

**note :

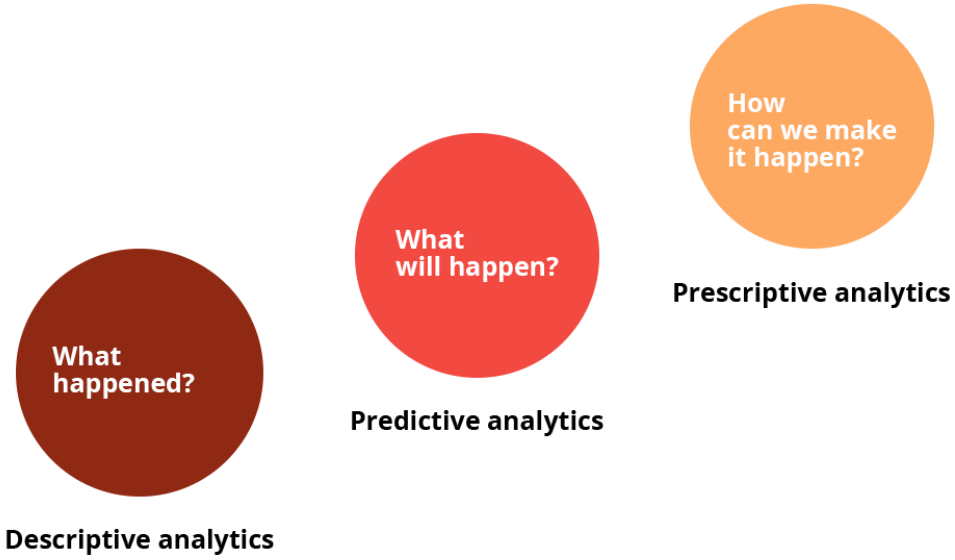
hukum V's berkembang seiring waktu dari 3V,5V,7V sampai 10V. Hanya saja yang umum dijadikan acuan untuk bigdata yaitu 5V

2. Apa perbedaan antara descriptive analytics, predictive analytice &prescriptive analytics

- **Descriptive** Analytics, merupakan tipe data analytics yang menggunakan data aggregation dan data mining untuk memberikan insight tentang peristiwa di masa lalu, dan digunakan untuk menjawab pertanyaan “What has happened?”. Contohnya yaitu Google Analytics

- **Predictive** Analytics, merupakan tipe data analytics yang menggunakan model statistik dan Teknik Teknik forecasting untuk memahami serta memperkirakan peristiwa di masa depan, dan digunakan untuk menjawab pertanyaan “What could happen?”. Contohnya yaitu situs e-commerce Amazon yang memprediksi barang-barang yang kira kira akan dibeli customer.

- **Prescriptive** Analytics, merupakan tipe data analytics yang mengguanak algoritma optimisasi dan simulasi untuk memberikan advice/masukan terhadap kemungkinan kemungkinan yang akan terjadi, dan digunakan untuk menjawab pertanyaan “What should we do?”



3. Jelaskan tahapan tahapan dalam membuat propensity model lengkap dengan key activities di setiap tahapan tersebut dan bagaimana cara untuk memperbaiki accuracy dari sebuah model.

Tahapan-tahapan dalam membuat pronpensity model yaitu :

Ada beberapa cara untuk memperbaiki accuracy dari sebuah model :

1. **Add more Data**
   Menambahkan data tentunya memberikan lebih banyak referensi bagi alogoritma untuk mencari dan mencocokan pola. Lakukan hal ini apabila memungkinkan.
2. **Treat missing values dan Outlier values.**
   Lakukan impute or omit apabila ditemukan missing value. Untuk outlier dapat dengan di transform, delete, binning , etc
3. **Feature Engineering.**
   Dibagi menjadi 2 tahap, yaitu Feature Transformation (contohnya Normalization/ Scaling data menjadi bilangan antara 0 – 1) dan Feature Creation.
4. **Feature Selection.**
   Dapat dilakukan dengan mengvisualisasikan korelasi antar variables untuk mencari variable mana yang berpengaruh besar ke hasil modeling/prediksi dan mana yang kurang berpengaruh dan dapat di buang.
5. **Alogirthm Tuning.**
   Dilakukan dengan merubah nilai hyper-parameter saat pembuatan model. Contoh nya parameter “n.tree”, “laplace”, “oob_score”, dll
6. **Ensemble Methods.**
   Ada 2 ensemble methods yang umum digunakan yaitu Bagging dan Boosting. Ensemble Methods merupakan Teknik dimana menggabungkan multiple weak model untuk menghasilkan result yang lebih baik

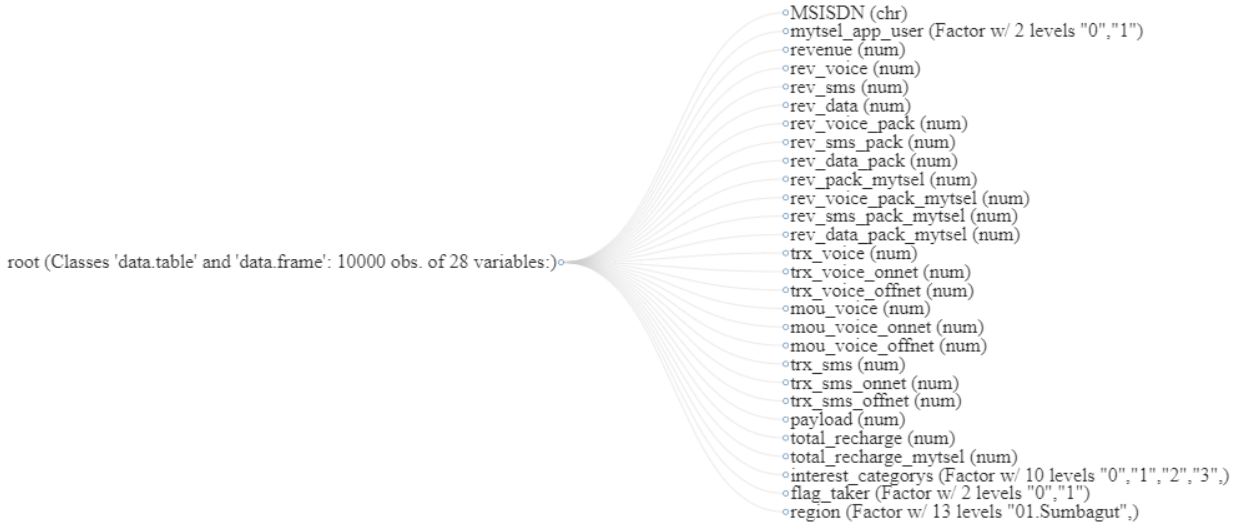
   *note : cross validation / split data needed untuk mencegah overfiting

Analysis Data Set

Note : Proses merapihkan awal data saya lakukan di **Mircosroft Excel**, kemudian di import ke **Rstudio**

Exploratory Data Analysis (EDA)

Sebelum lanjut ke proses yang lain, terlebih dahulu saya rapihkan type dari setiap variabel agar sesuai dengan yang seharusnya (categorical/numerical) di Rstudio :



Kita dapat melihat summary dari dataset sebagai berikut :

```
   MSISDN          mytsel_app_user    revenue          rev_voice          rev_sms          rev_data       rev_voice_pack    rev_sms_pack       rev_data_pack
Length:10000       0:6333          Min.   :      0   Min.   :      0   Min.   :      0   Min.   :      0   Min.   :      0   Min.   :    0.000   Min.   :      0
Class :character   1:3667          1st Qu.:  16067   1st Qu.:      0   1st Qu.:      0   1st Qu.:      0   1st Qu.:      0   1st Qu.:    0.000   1st Qu.:      0
Mode  :character                   Median :  75662   Median :      0   Median :      0   Median : 75000   Median :      0   Median :    0.000   Median : 75000
                                   Mean   :  85862   Mean   : 10843   Mean   :    970   Mean   : 66384   Mean   :   4249   Mean   :    8.976   Mean   : 63504
                                   3rd Qu.: 119457   3rd Qu.:  8842   3rd Qu.:    660   3rd Qu.: 90010   3rd Qu.:      0   3rd Qu.:    0.000   3rd Qu.: 80000
                                   Max.   :1637445   Max.   :430059   Max.   :103000   Max.   :997562   Max.   :351350   Max.   :27500.000   Max.   :914030

rev_pack_mytsel  rev_voice_pack_mytsel rev_sms_pack_mytsel rev_data_pack_mytsel   trx_voice       trx_voice_onnet   trx_voice_offnet   mou_voice
Min.   :     0   Min.   :     0.0      Min.   :    0.000   Min.   :     0       Min.   :   0.00   Min.   :   0.00   Min.   :  0.00   Min.   :    0.0
1st Qu.:     0   1st Qu.:     0.0      1st Qu.:    0.000   1st Qu.:     0       1st Qu.:   3.00   1st Qu.:   1.00   1st Qu.:  0.00   1st Qu.:    1.0
Median :     0   Median :     0.0      Median :    0.000   Median :     0       Median :  26.00   Median :  15.00   Median :  3.00   Median :   28.0
Mean   :  8048   Mean   :   305.4      Mean   :    3.675   Mean   :  7722       Mean   :  84.75   Mean   :  63.66   Mean   : 18.62   Mean   :  126.4
3rd Qu.:     0   3rd Qu.:     0.0      3rd Qu.:    0.000   3rd Qu.:     0       3rd Qu.:  97.00   3rd Qu.:  66.00   3rd Qu.: 19.25   3rd Qu.:  125.0
Max.   :914000   Max.   :184000.0      Max.   :16250.000   Max.   :914000       Max.   :1750.00   Max.   :1741.00   Max.   :525.00   Max.   :10776.0

mou_voice_onnet   mou_voice_offnet     trx_sms        trx_sms_onnet      trx_sms_offnet       payload         total_recharge    total_recharge_mytsel
Min.   :    0.0   Min.   :    0.00   Min.   :   0.00   Min.   :    0.000   Min.   :  0.000   Min.   :       0   Min.   :     0   Min.   :     0.0
1st Qu.:    0.0   1st Qu.:    0.00   1st Qu.:   0.00   1st Qu.:    0.000   1st Qu.:  0.000   1st Qu.:     146   1st Qu.:  10000   1st Qu.:     0.0
Median :   15.0   Median :    1.00   Median :   3.00   Median :    1.000   Median :  0.000   Median :  7211296   Median :  75000   Median :     0.0
Mean   :  107.8   Mean   :   15.11   Mean   :  11.84   Mean   :    8.996   Mean   :  2.337   Mean   : 10832037   Mean   :  85252   Mean   :   355.5
3rd Qu.:   89.0   3rd Qu.:   14.25   3rd Qu.:  10.00   3rd Qu.:    6.000   3rd Qu.:  1.000   3rd Qu.: 17073051   3rd Qu.: 120000   3rd Qu.:     0.0
Max.   :10765.0   Max.   :1309.00   Max.   :1057.00   Max.   :1057.000   Max.   :442.000   Max.   :192172798   Max.   :2375000   Max.   :300000.0

total_recharge_URP interest_categorys flag_taker             region
Min.   :0          0      :2274       0:6615     05.Central Jabotabek:1567
1st Qu.:0          4      :1368       1:3385     06.Eastern Jabotabek:1529
Median :0          6      :1252                  04.Western Jabotabek:1298
Mean   :0          5      :1143                  09.Jatim            :1018
3rd Qu.:0          3      :1093                  08.Jateng           : 906
Max.   :0          1      :1071                  07.Jabar            : 873
                   (other):1799                  (Other)             :2809
```

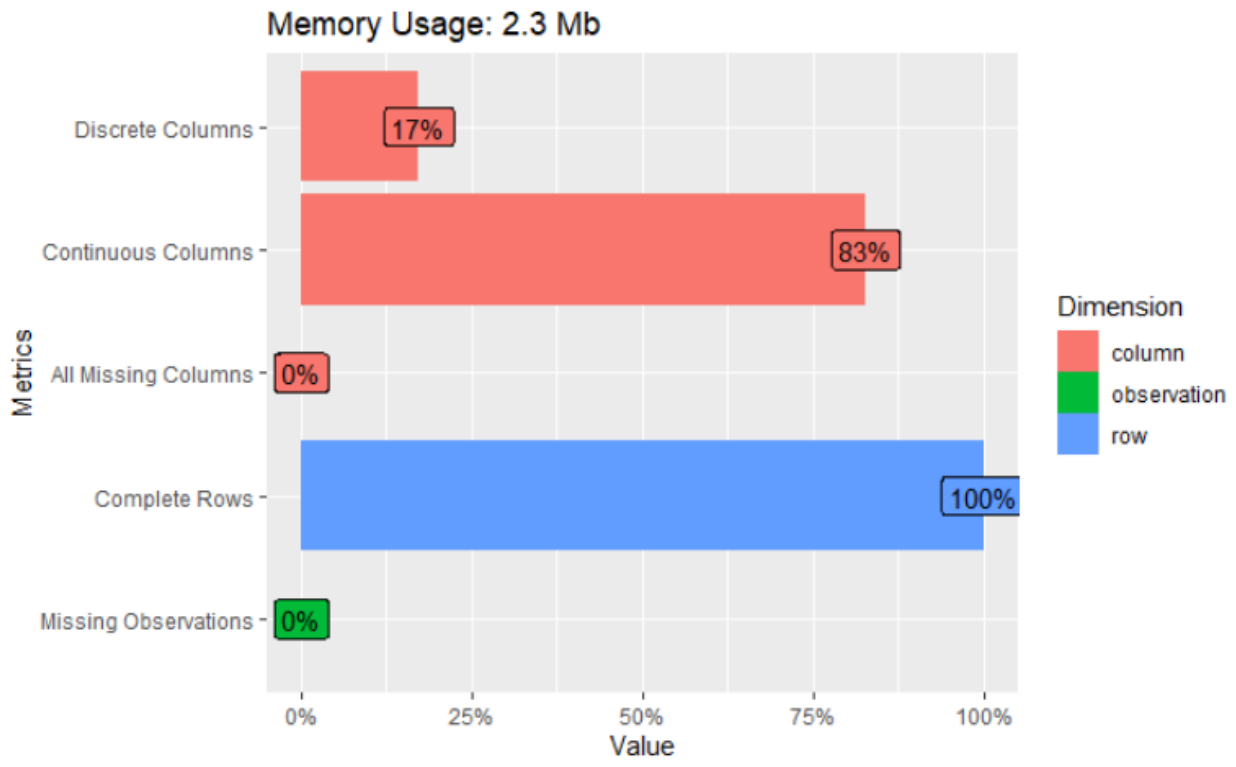
Variable “total_recharge_urp” hanya memiliki 1 value, oleh karena itu kita early exclude saja agar tidak menghalangi proses modeling

Berikut basic statistic untuk melihat gambaran data secara keseluruhan, sekaligus identify missing value. Apabila memang ditemukan missing value/ NA, maka bisa 2 opsi :

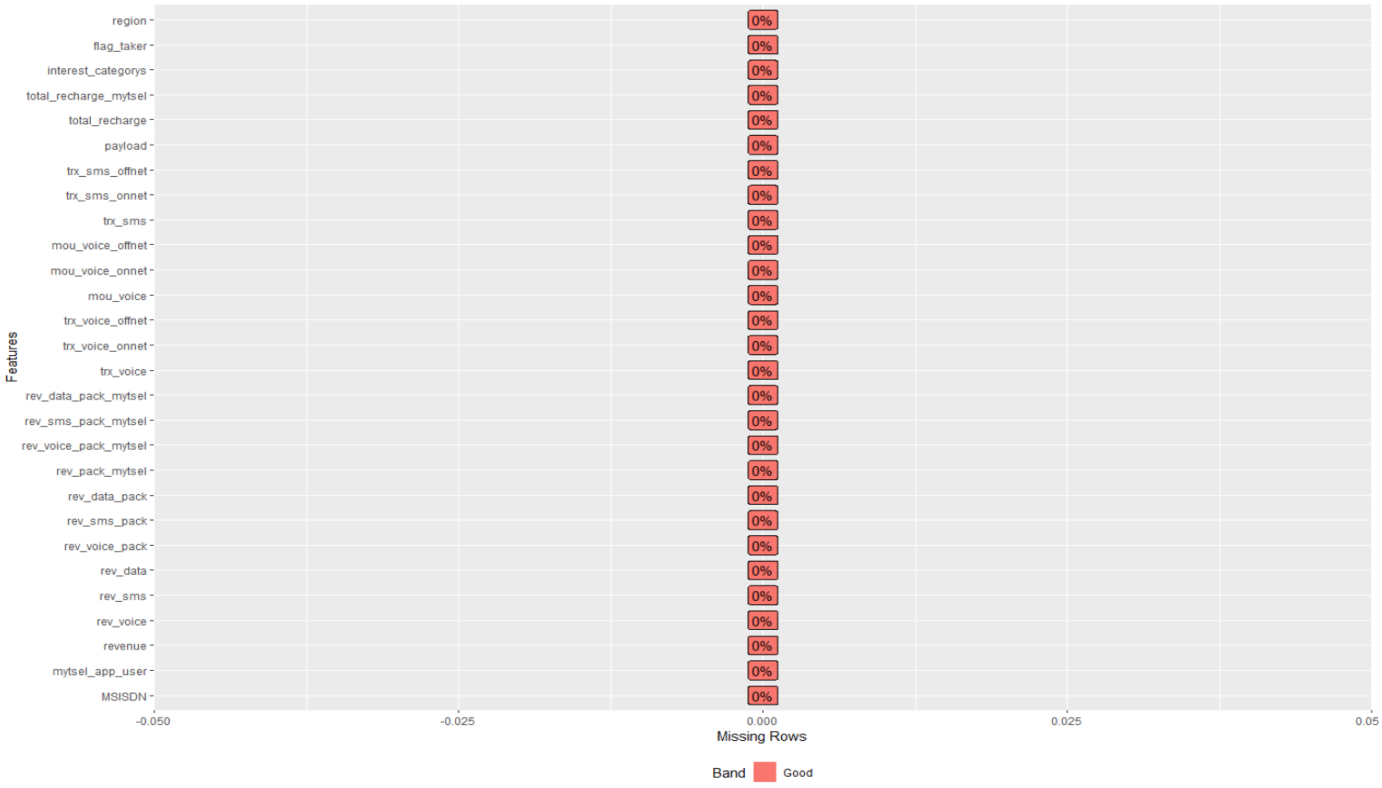
1. na.omit (row yang mengandung NA dibuang dari dataset)
2. impute, dapat dengan na.roughfix (menggunakan median dari data) atau randomforest (menggunakan library missForest)

Name	Value
Rows	10,000
Columns	29
Discrete columns	5
Continuous columns	24
All missing columns	0
Missing observations	0
Complete Rows	10,000
Total observations	290,000
Memory allocation	2.3 Mb

Graph interpretation :

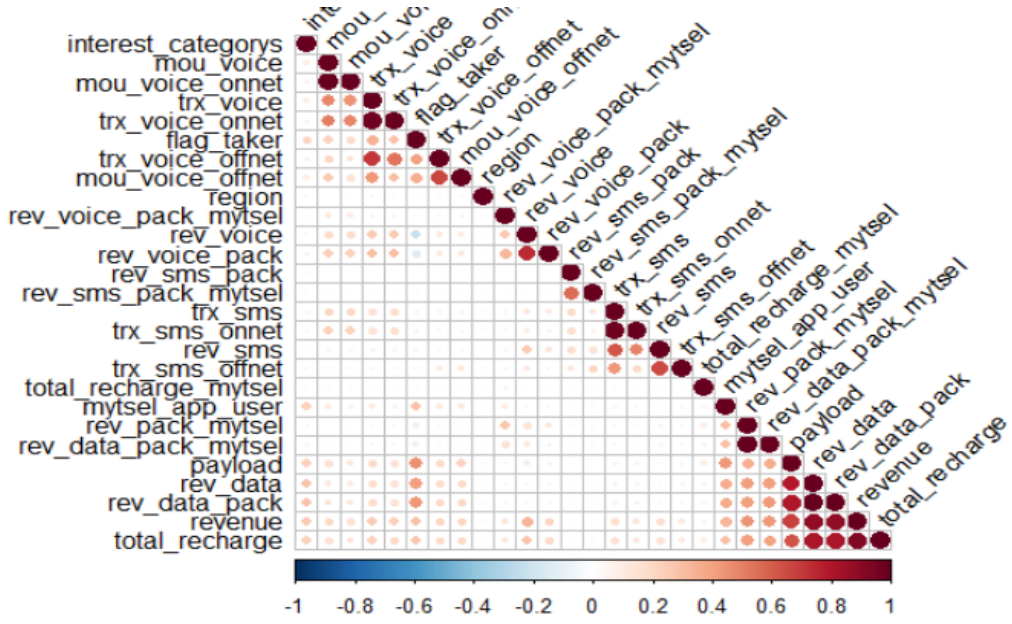


Dataset terdiri dari 10000 obs dan 29 variable. 24 Diantaranya (83%) merupakan continuous column yang berisi numerical value , 5 sisanya (17%) merupakan discrete column yang berisi categorical value.



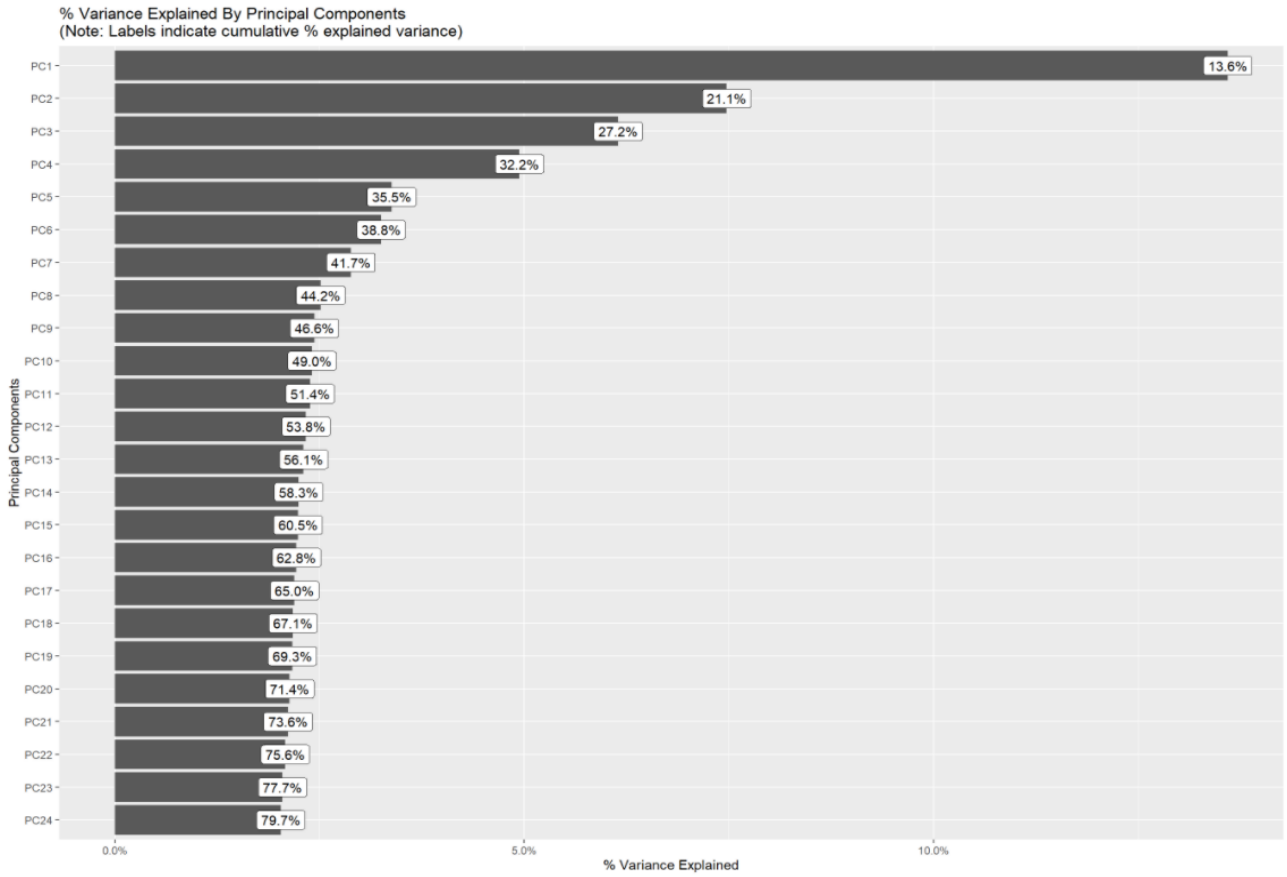
No missing value found in dataset. Dataset sudah cukup clean.

Berikutnya saya coba cek korelasi antar variable pada dataset, dan show dalam bentuk heat-map :

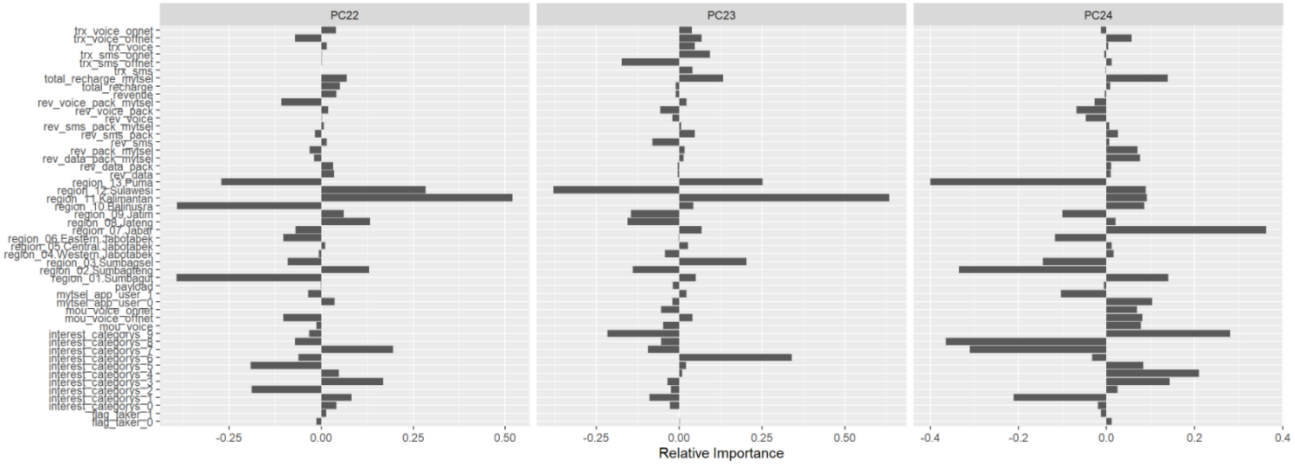


Dari graph dapat dilihat bahwa variabel “rev_data”, “rev_data_pack”, “revenue” dan “total_recharge” memiliki korelasi yang tinggi. Variabel (trx_sms,trx_sms_offnet) , (trx_voice,trx_voice_onnet), (mou_voice, mou voiceonnet) juga memiliki korelasi tinggi. Kedepannya tergantung dari DS apakah variabel yang berkorelasi mau di exclude atau tidak. Untuk case kali ini tidak saya exclude.

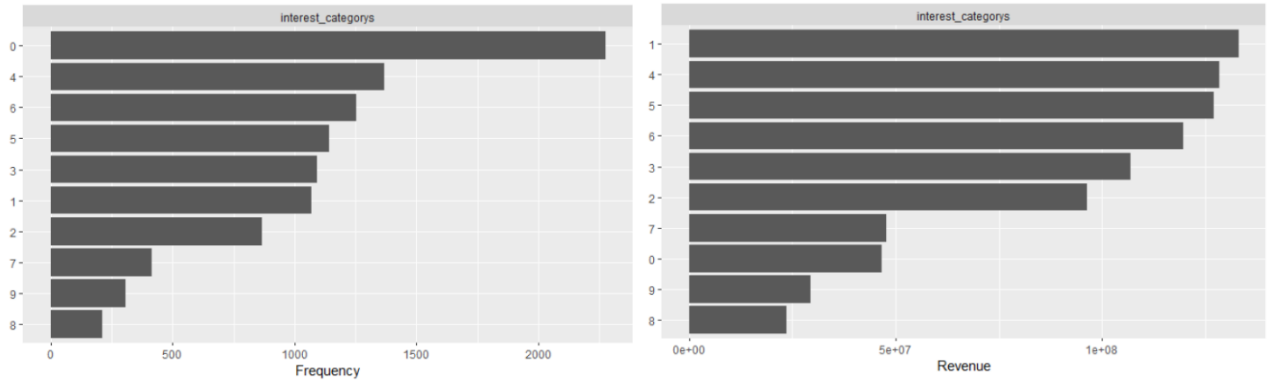
Untuk mengetahui *feature importance* dapat dilakukan dengan PCA :



Rekomendasi variance explained yang baik tidak boleh < 60%, oleh karena itu saya coba pilih PC1-PC24 dimana sudah menyentuh 79.7%



Beberapa insight lainnya juga dapat di tarik untuk lebih memahami data, sesuai kebutuhan demi menjawab defined problem yang ada, contohnya antara lain sbb :



interest category 1 (Video) meskipun secara frequency menempati posisi 5 di sekitar angka 1071 obs, namun secara revenue menjadi penyumbang terbesar bagi Telkomsel. Mengalahkan Games, eCommerce, socialnetwork, dan lainnya.

Answer Sheet

Jawaban nomor 1 :

Row Labels	Count of MSISDN	Sum of revenue
01.Sumbagut	**611**	**50524873**
Non-User MyTsel	431	28399276
User MyTsel	180	22125597
02.Sumbagteng	**415**	**34844882**
Non-User MyTsel	281	19428732
User MyTsel	134	15416150
03.Sumbagsel	**543**	**45366304**
Non-User MyTsel	347	20279788
User MyTsel	196	25086516
04.Western Jabotabek	**1298**	**118554903**
Non-User MyTsel	763	52355648
User MyTsel	535	66199255
05.Central Jabotabek	**1567**	**144773099**
Non-User MyTsel	953	68937753
User MyTsel	614	75835346
06.Eastern Jabotabek	**1529**	**130538657**
Non-User MyTsel	974	60957310
User MyTsel	555	69581347
07.Jabar	**873**	**69961974**
Non-User MyTsel	549	32511058
User MyTsel	324	37450916
08.Jateng	**906**	**73806717**
Non-User MyTsel	558	34912564
User MyTsel	348	38894153
09.Jatim	**1018**	**82144482**
Non-User MyTsel	661	40585798
User MyTsel	357	41558684
10.Balinusra	**206**	**19160613**
Non-User MyTsel	133	9730400
User MyTsel	73	9430213
11.Kalimantan	**334**	**28086739**
Non-User MyTsel	219	13910332
User MyTsel	115	14176407
12.Sulawesi	**643**	**54934716**
Non-User MyTsel	425	26088653
User MyTsel	218	28846063
13.Puma	**57**	**5922117**
Non-User MyTsel	39	3434600
User MyTsel	18	2487517