Customers who use your platform have different needs and they have their own different profile. You should adapt your actions depending on that.

You can do many different segmentations according to what you are trying to achieve. If you want to increase retention rate, you can do a segmentation based on churn probability and take actions. But there are very common and useful segmentation methods as well. Now we are going to implement one of them to our business: **“RFM”**

**RFM**stands for Recency - Frequency - Monetary Value.

**Recency**

To calculate recency, we need to find out most recent visit date of each customer and see how many days they are inactive for.

**Frequency**

To calculate frequency, we need to calculate the number of visits for each customer.

**Monetary**

To calculate monetary, we need to calculate the total amount of usage (MG) for each customer.

For this method, we will use a dataset about the customers who visit and use our application. This dataset includes 6 variables and 3752013 records.

The variables are ‘User\_ID’, ‘DateKey’, ‘Usage\_MB’, ‘age’, ‘gender\_v’ and ‘device\_type’.

In preprocessing step, first a General description of data set is needed and finding missing values and outlier are second step.

For example, unreasonable values are observed in age column (refer to describe function to understand that).

For executing the RFM, creating these three factors is necessary:

* **Recency**
* **Frequency**
* **Monetary**

Recency

First, the last recorded date of the whole data must be found, which is equal to ‘2021/03/01’. Then we find the last date of each customer visiting the application and subtract the mentioned date from it.

Frequency

By counting the number of visits per customer, frequency is obtained.

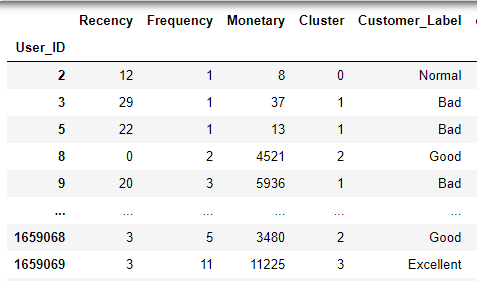
Monetary  
By summing the amount of usage of each customer, monetary is obtained.

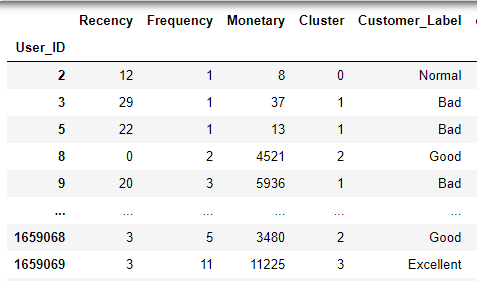
Now using these three factors, the RFM data frame is created based on each customer information that is given a label to each customer. This labeling is done by clustering algorithm (K-means), which indicates the type of customer behavior.

By comparing the R, F and M values in each cluster, we choose the appropriate name for each cluster.

Loyal customers (excellent) are placed in a cluster with higher F, M values and lower R value and in the same way, other clusters are determined behavior.

For example, Consider the following two customers:





As can be seen, the customer, which is in the excellent cluster, has both high usage and visits, and the number of days he is inactive is only 3 days.

But the customer, which is in the normal cluster, has less usage and number of visits and more days of inactivity compared to the previous customer.

Now you can add age, gender and device type variables to the RFM data frame.

As mentioned earlier, the age variable has unreasonable values. Instead of deleting the rows that have this problem, the average age of the cluster can be replaced by looking at the label of each irrational row.

Using IQR, the range of unreasonable ages can be identified.

On the other hand, the average age of each cluster is equal to the average total age of the data frame, which is thirty-seven. So instead of unreasonable ages, we can put the average age of the whole data frame.

In this project, the number of clusters has been determined using the elbow method. But the number of clusters can be determined according to organizational conditions.