# Assignment 4 (Part 1)

Created by: Aitpayev Arlan, Akhanov Askar, Mussanova Anel, Shaken Nurassyl

BD-2001

1. **Explore the dataset. Do the descriptive statistics**

Firs of all, import all the libraries we used. It is convenient to write them out at the very beginning. And we also load data from files into our compiler.

Изображение выглядит как текст

Автоматически созданное описание

**Dataset exploring.**

Transactions dataset:

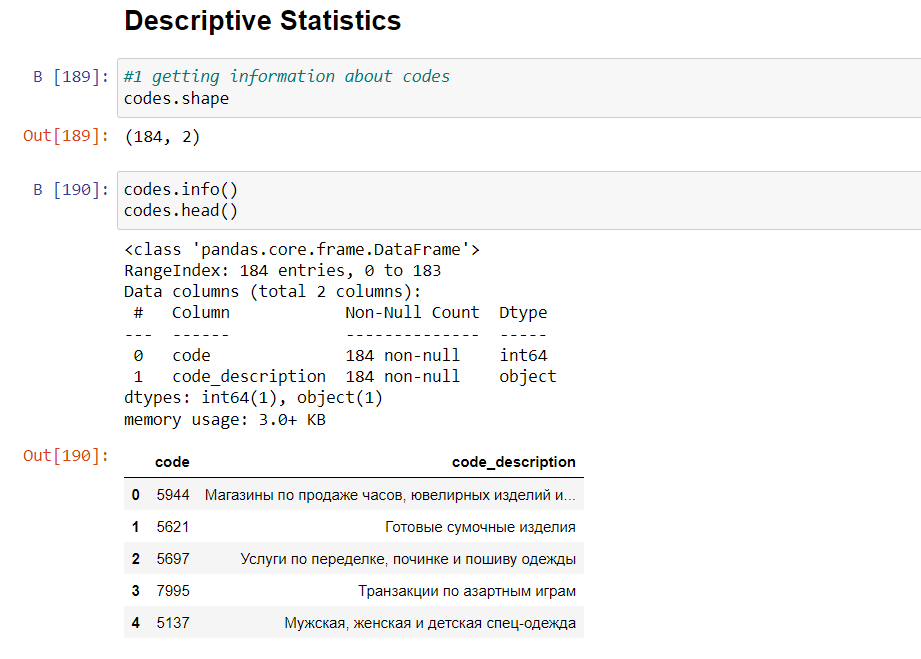
Find the median, mode, measures of spread, IQR, variance, standard division for the dataset "transactions". To do this, we use default formulas.

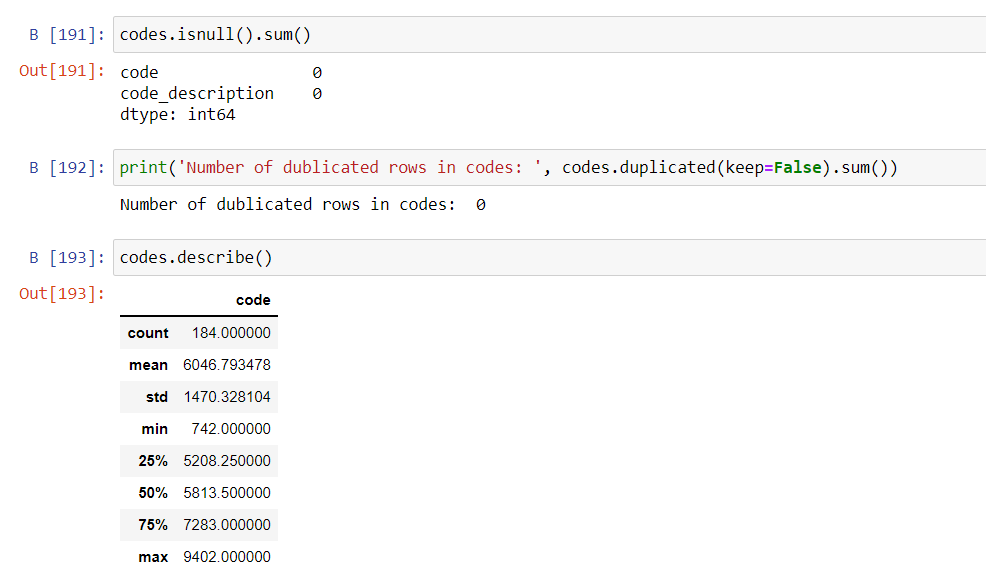
Изображение выглядит как текст

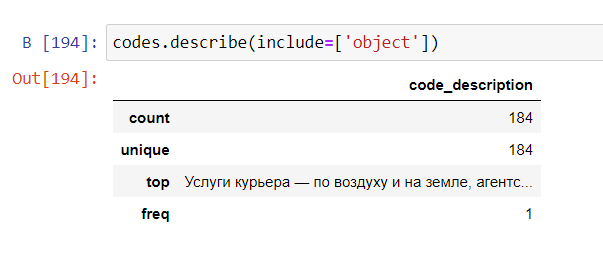
Автоматически созданное описаниеИзображение выглядит как текст

Автоматически созданное описание

Codes dataset:

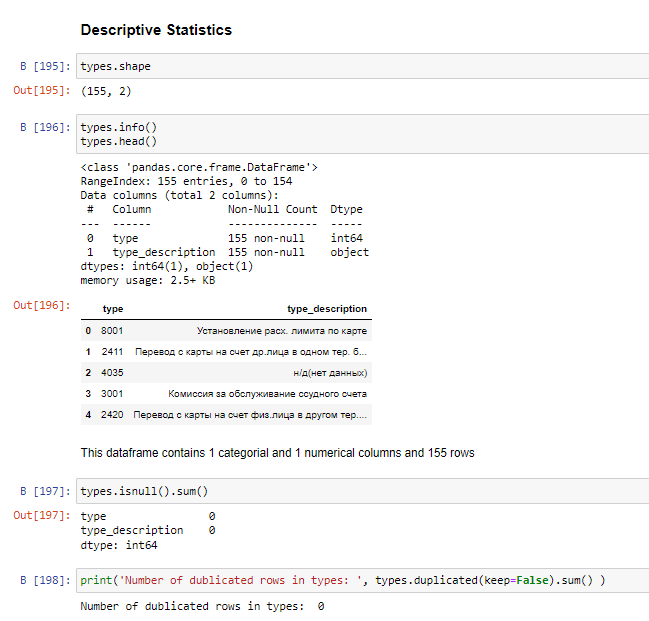


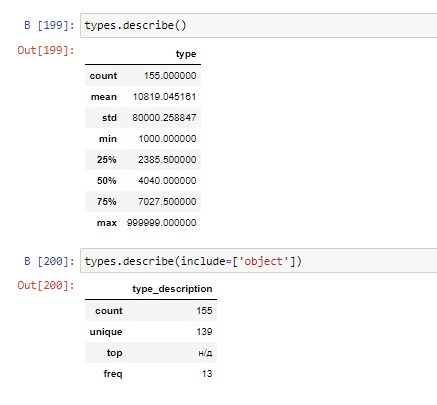




Codes dataset contains 184 rows and 2 columns. One of the columns is numerical and second one is categorical. Number of dublicated rows is 0.

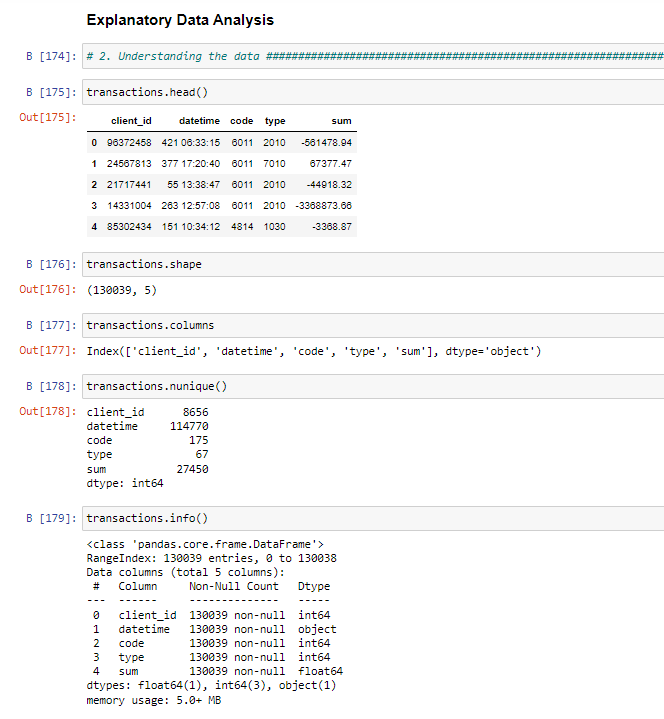
Types dataset:



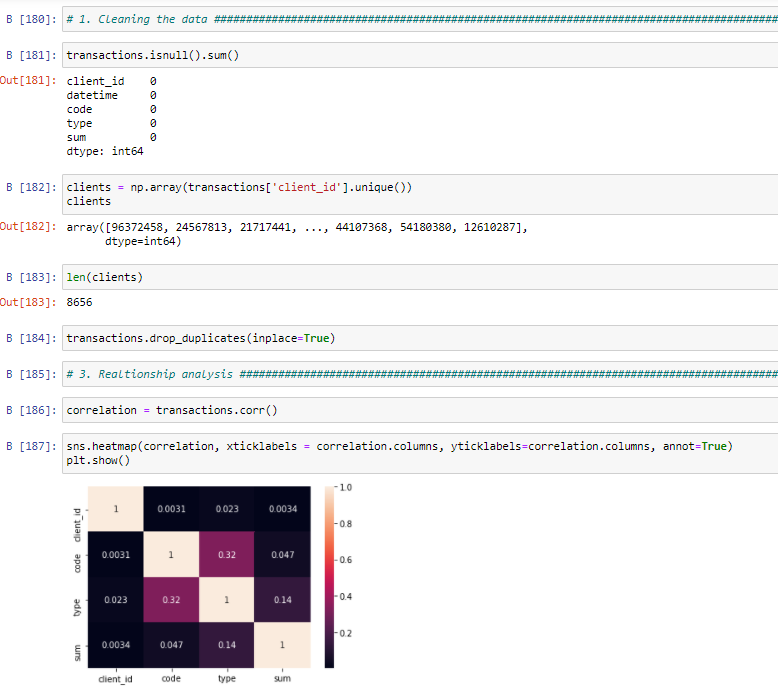


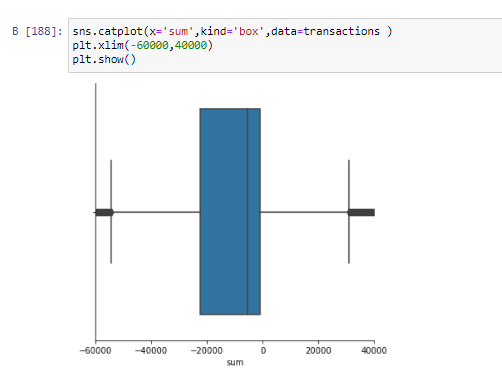
Types dataset contains 155 rows and 2 columns. One of the columns is numerical and second one is categorical. Number of dublicated rows is 0.

1. **Explanatory data analysis. Exploring the features, visualizations etc.**



Transactions dataset contains 130039 rows and 5 columns. It contains columns like ‘client\_id’, ’datetime’, ’code’, ’type’, ’sum’. Also in this dataset a lot of non-unique values.



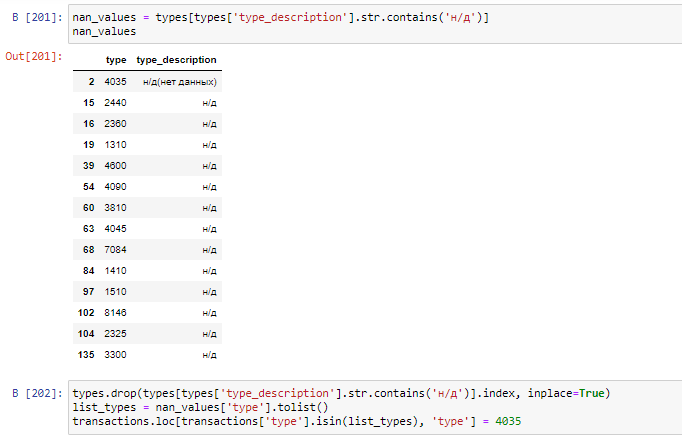


Dataset doesn’t have null values. There are 8656 unique clients here. And we are dropped duplicated values. To understand the relationship between columns in this dataset we drawn correlation heatmap. Also to represent the mean of the table and visualize the data we drawn catplot.

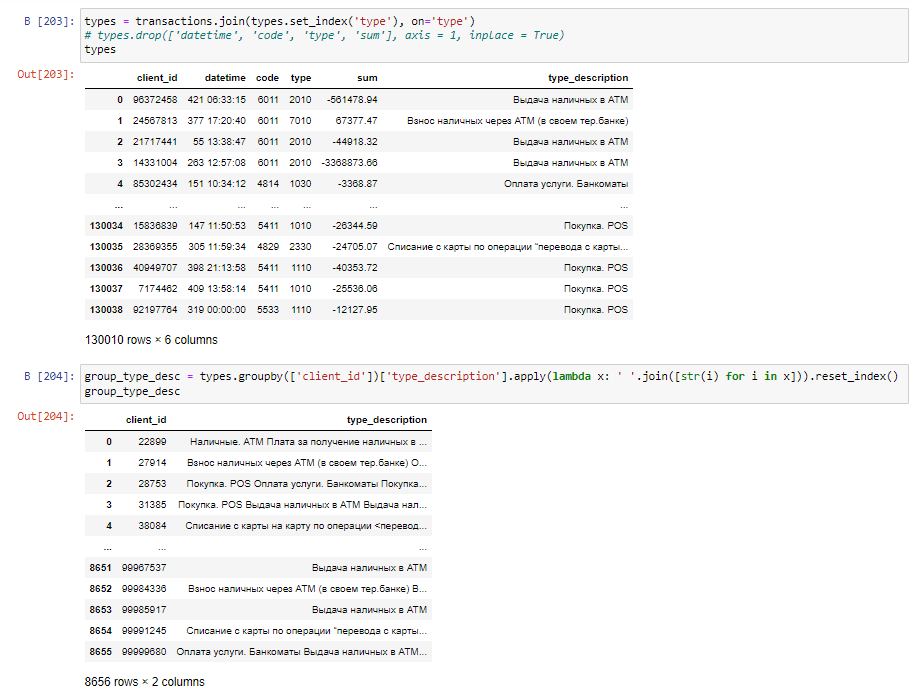
1. **Feature engineering**

Types dataset:

After studying the dataset, we found out that there are rows in types that do not carry information significant, namely, there were several rows without a description. Therefore, we decided to combine them into one type.

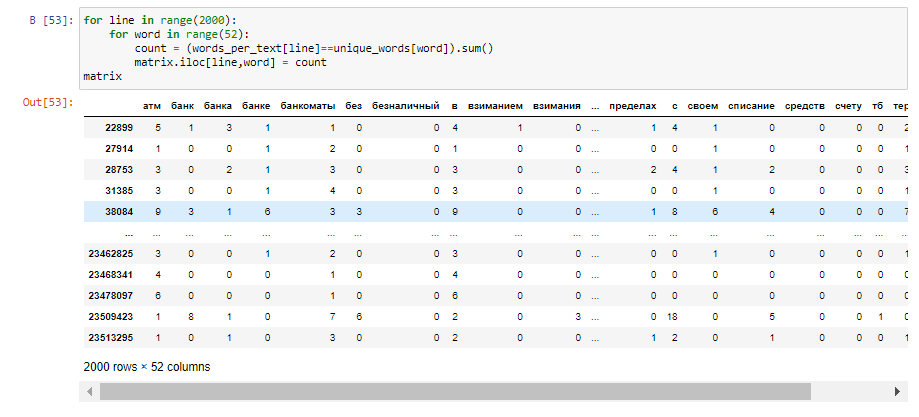


We dropped them from types dataset and from transactions dataset.



After dropping we joined two datasets, transactions and types, by types columns. Created new table named group\_type\_desc, which contains client\_id and type\_description columns.

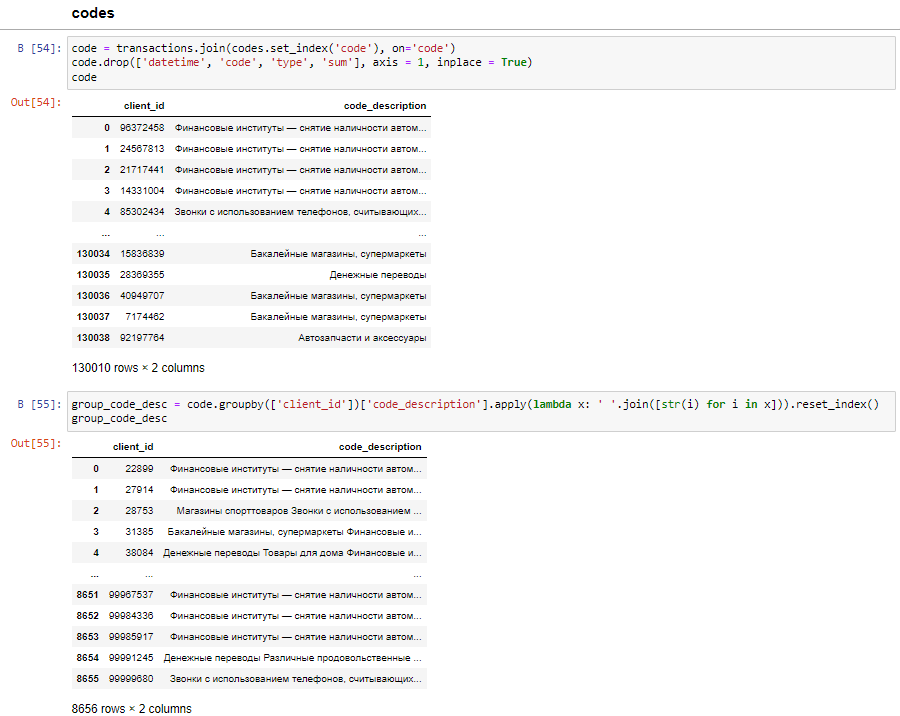


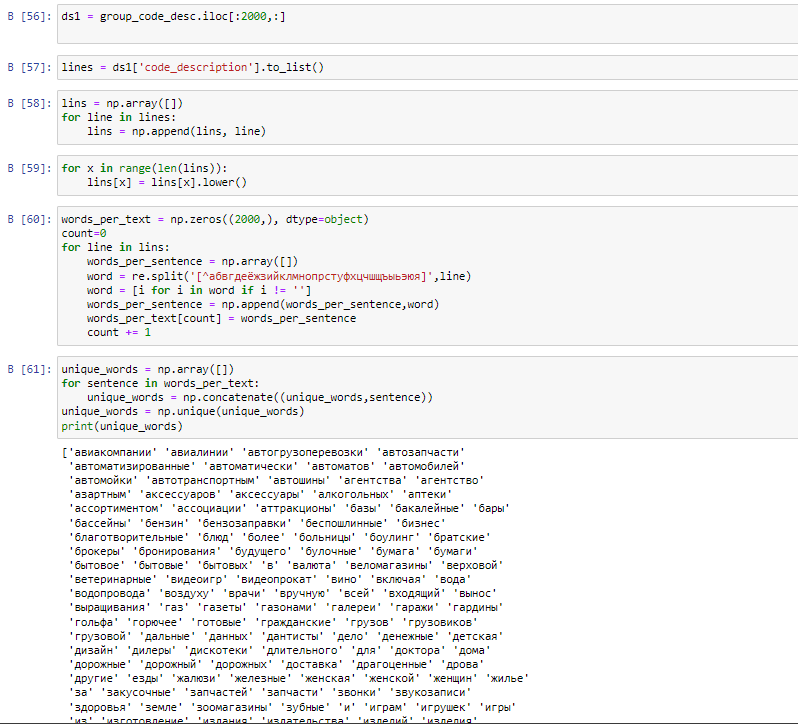


Then, we created tf table by first 2000 rows of this dataset.

Codes dataset:

We perform similar actions for codes dataset.







We will use this table in the future.

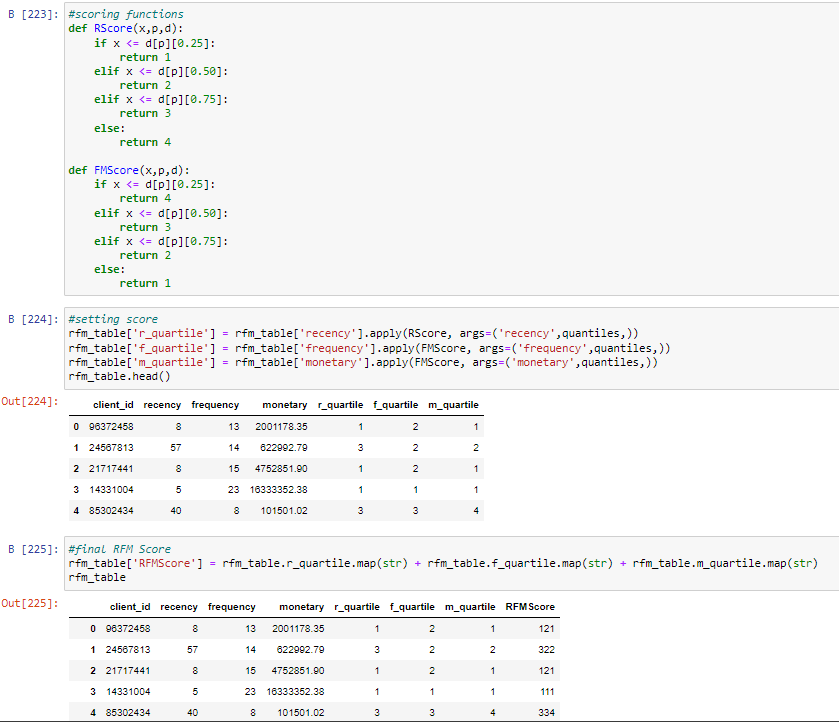
Transactions dataset:

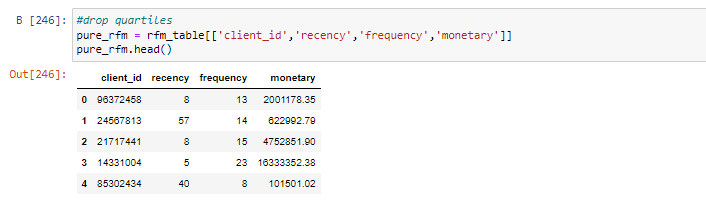


In our transactions table, firstly we making absolute sum for our ‘sum’ column in transactions. Secondly, we divided our ‘datetime’ column into date and time, separately.

We'll implement here the RFM principle to classify the customers in this database. This part is inspired by the work of Susan Li. RFM stands for Recency, Frequency and Monetary. It is a customer segmentation technique that uses past purchase behavior to divide customers into groups.





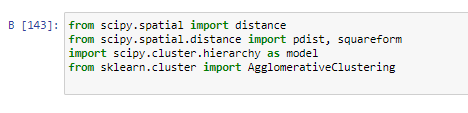


# **Unsupervised learning. Cluster analysis. Segment the customers. K-means,Hierarchical Clustering. With different metrics, linkages. Visualize the clusters etc.**

### **Hierarchical Clustering**

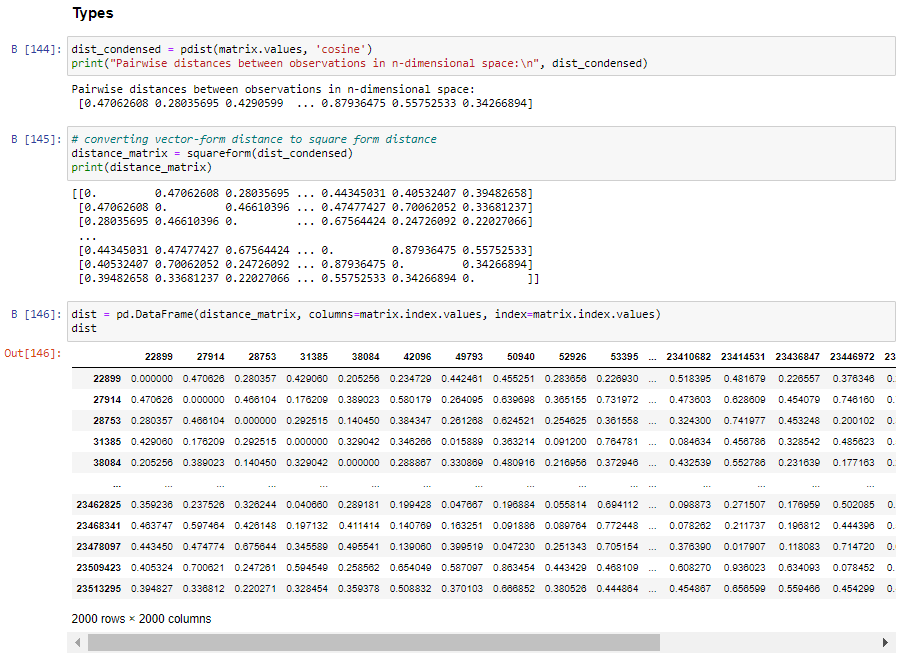
Find similar transactions by cosine distance

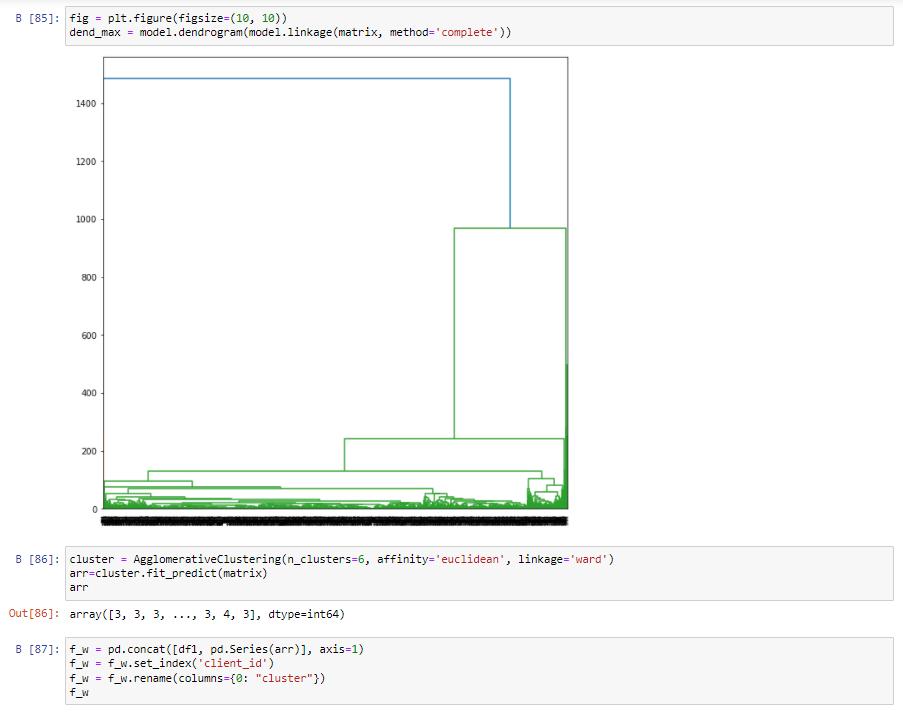
Importing necessary libraries:



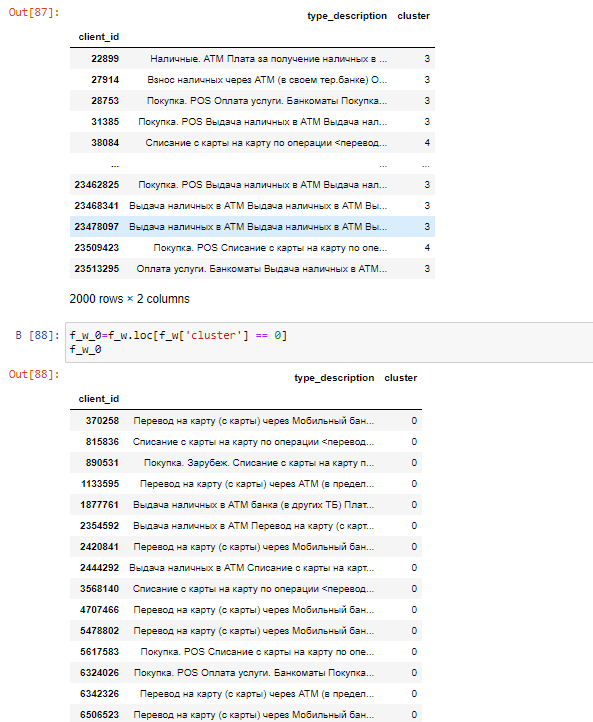
Types dataset:

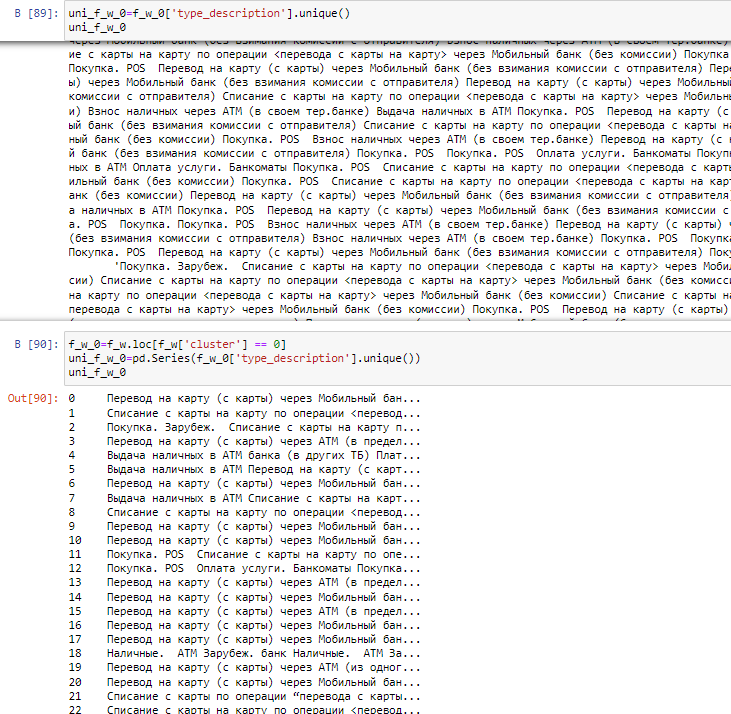
Firstly, we calculating cosine distance for our previous ‘matrix‘. After calculating we created index dataframe with the distance between each indexes. Then by given distances we clasterize them, and create hierarchical clustering.

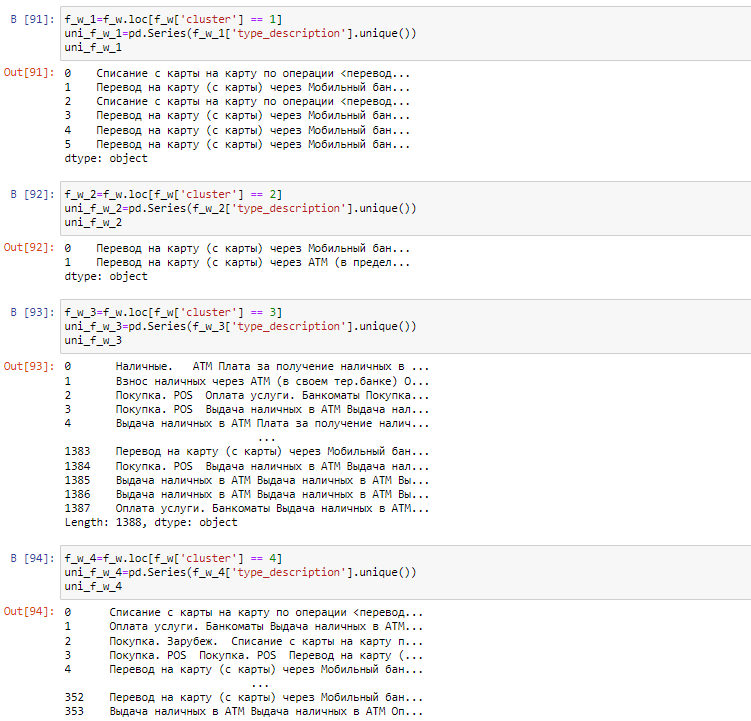


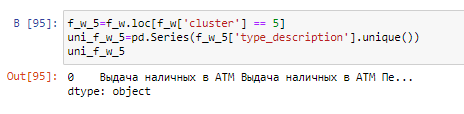


After, drawing dendogram we decided to clusterize our data into 6 clusters



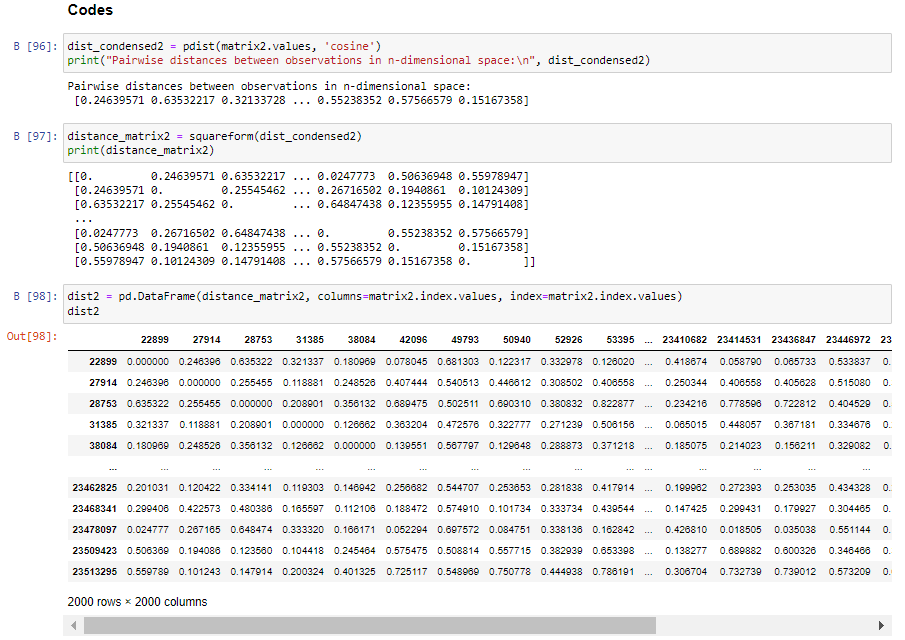




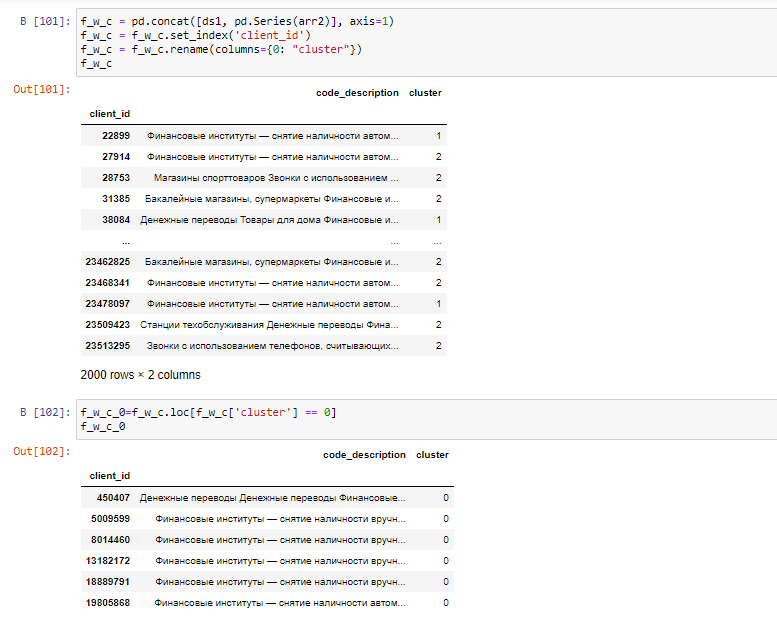


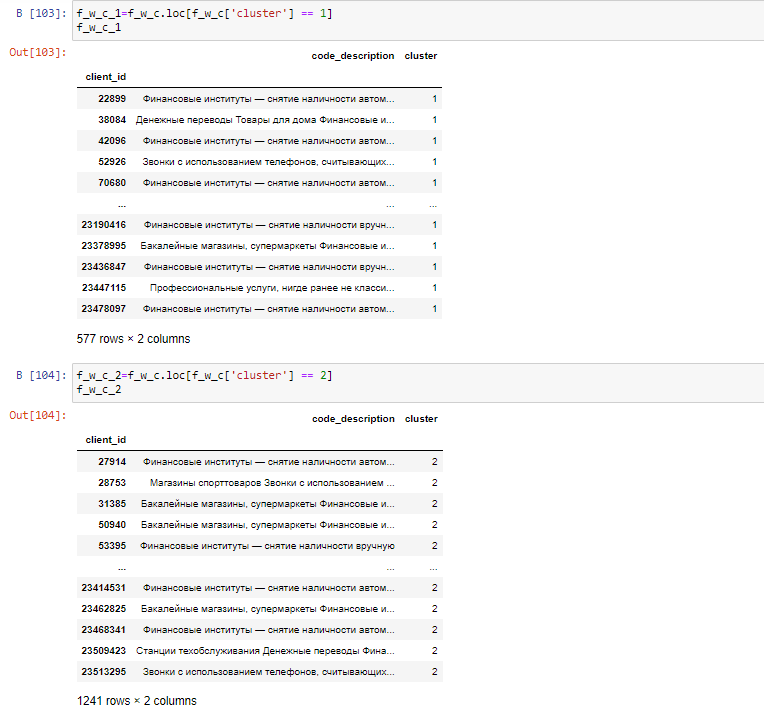
Codes dataset:

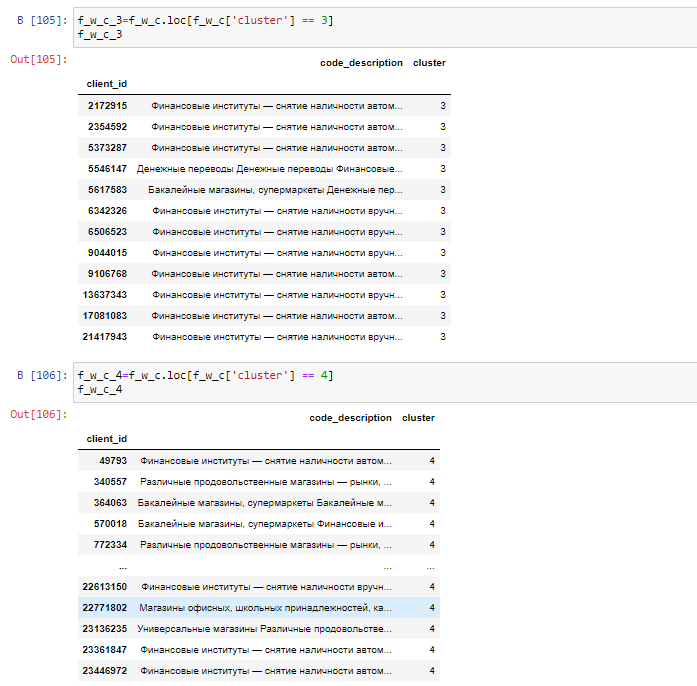
We perform similar actions for codes dataset.

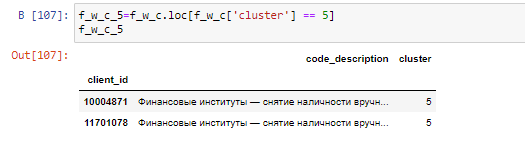












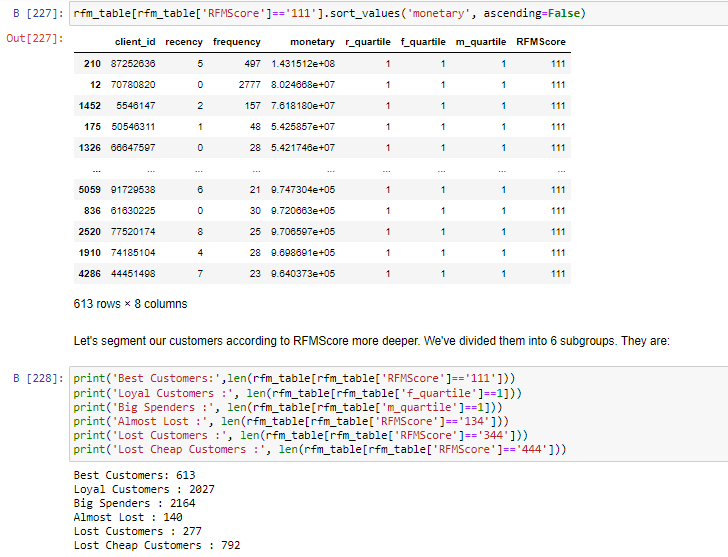
## **Segmenting the customer**

RFM check

Now we understand that r\_quartile being equal to 1 means that the client has done his transaction recently and if it equals to 2, it means that recency is less than when it was equal to 1. The logic goes on same.

But in frequency and monetary we have opposite logic. Scoring big value means that client has transacted a big sum of money and respectively less score associated with smaller sums.

Let's get our best clients(recency = 1, frequency=1, monetary =1)



**K-means clustering**

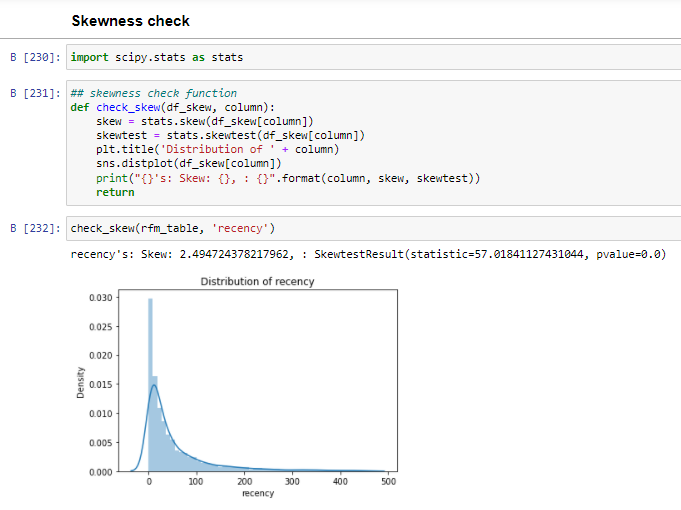
Importing necessary libraries:

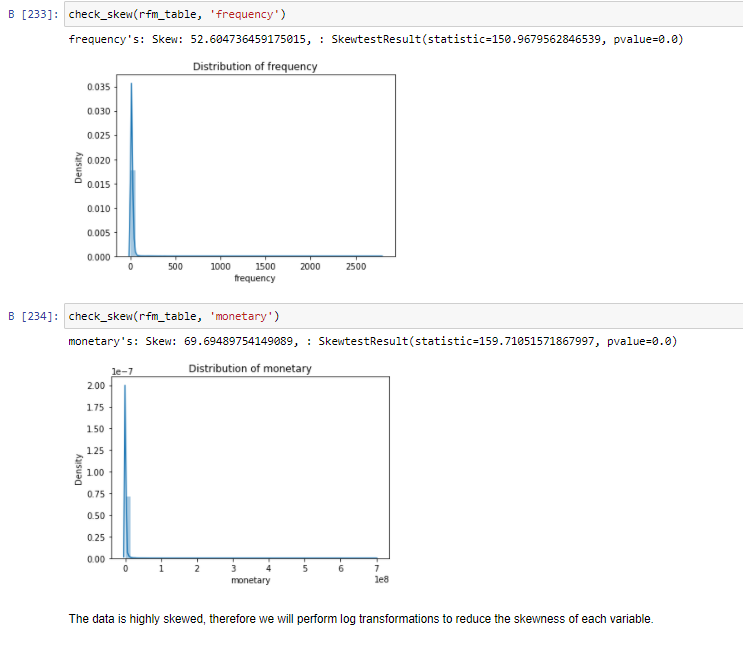
****

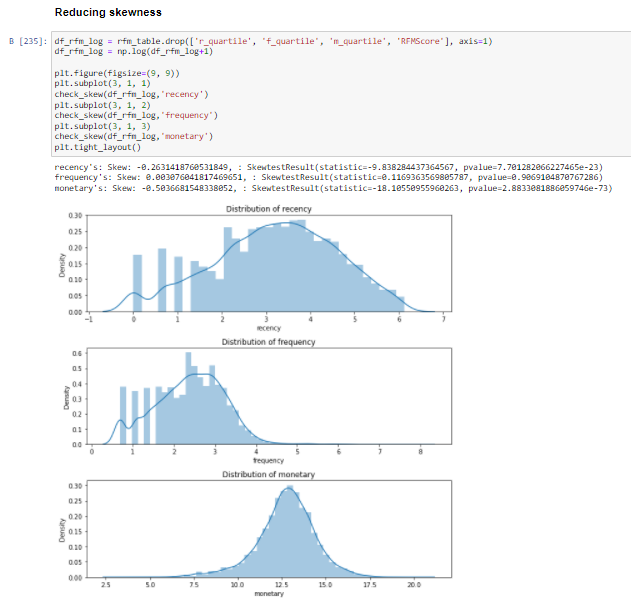
K-means gives the best result under the following conditions:

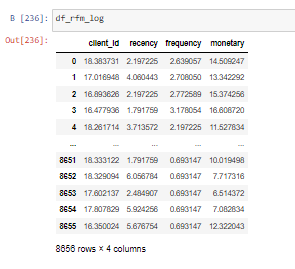
1) Data’s distribution is not skewed.

2) Data is standardised (i.e. mean of 0 and standard deviation of 1).



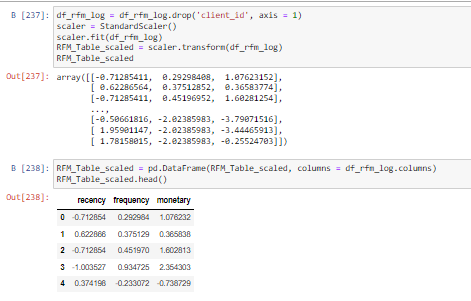


After looking for the skewnees, we need to reduce them.  


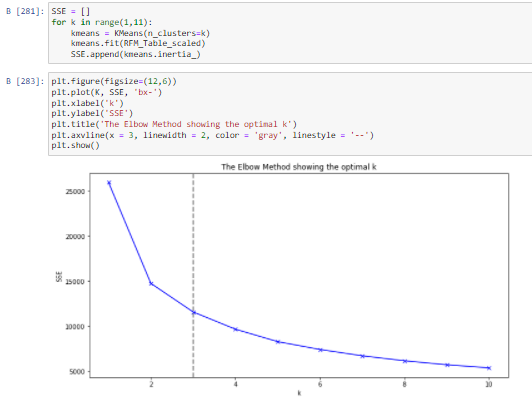


### **Data scaling**

Since we avoided skewness, we can start standardizing data by scaling and centring



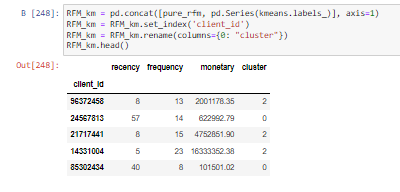
### Computing cluster amount



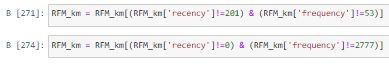
### Data fitting



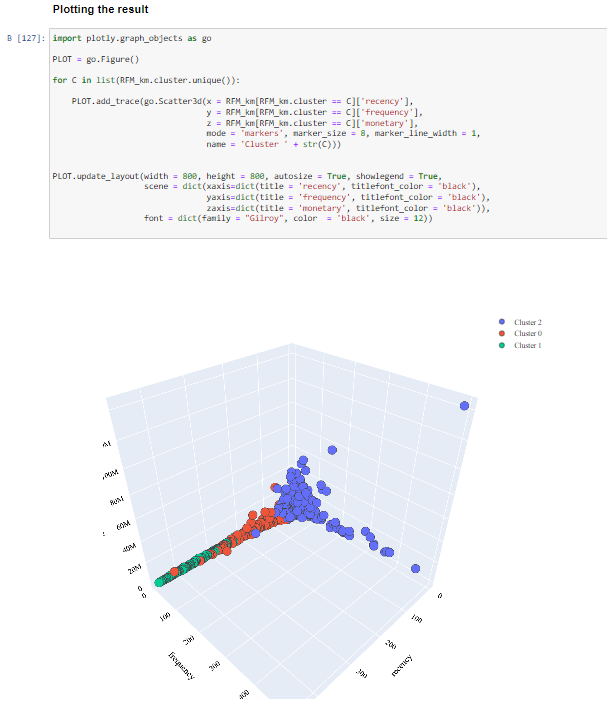
### Data labeling



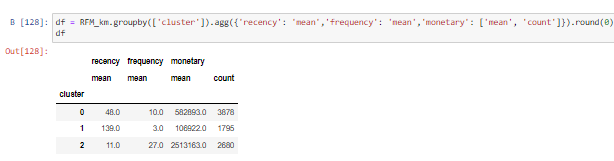
### Deleting savage outliers



### Plotting the result



# Conclusion



What does each cluster represent?

1) The first cluster belongs to the “Best Customers” segment which we saw earlier as they purchase recently, frequent buyers, and spent the most.

2) Customers in the second cluster are somewhere in between first and last class.

3) The third cluster is more related to the “Lost Cheap Customers” segment as they haven’t purchased for a long time, they have small frequency number and spent very little.