Assignment for second attestation

Work was done by: [Alisher Onggar](https://moodle.astanait.edu.kz/user/view.php?id=1457&course=1472), [Rakhat Shalgynbai](https://moodle.astanait.edu.kz/user/view.php?id=945&course=1472), [Konyrtaev Zein](https://moodle.astanait.edu.kz/user/view.php?id=1522&course=1472), Aldiyar Kenzhebayev

According to the tasks we should have done, we divided our project into 5 parts:

* Cleaning Data
* Exploratory Analysis
* Clustering
* Visualization

**Cleaning Data**

First of all, we imported the needed libraries.

Изображение выглядит как текст

Автоматически созданное описание

Next, we use the read\_csv function to read our files as a dataframe and save them.

Изображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как текст

Автоматически созданное описаниеИзображение выглядит как текст

Автоматически созданное описаниеAfter which we get a quick overview of the datasets by using the info() function

Изображение выглядит как текст

Автоматически созданное описаниеИзображение выглядит как текст

Автоматически созданное описание

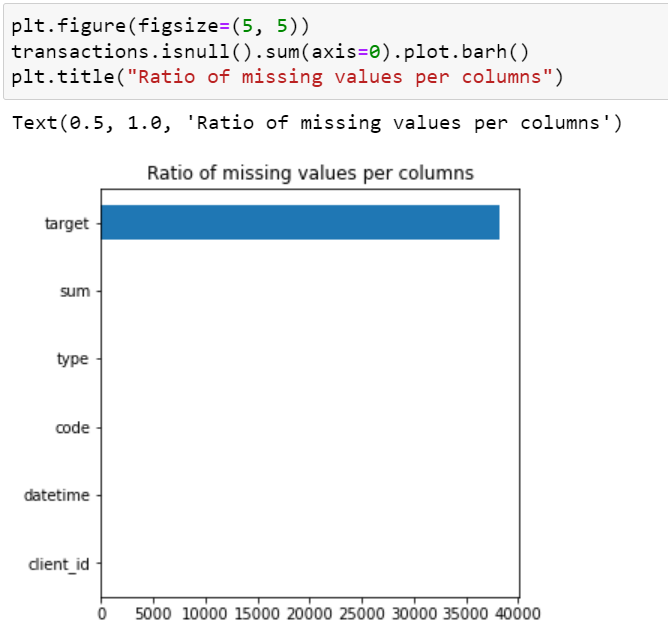
Types Codes Transactions Train

Adding csv’s data into one dataframe which is called transaction

Изображение выглядит как стол

Автоматически созданное описание

As we can see here, there are some missing values inside our dataframe

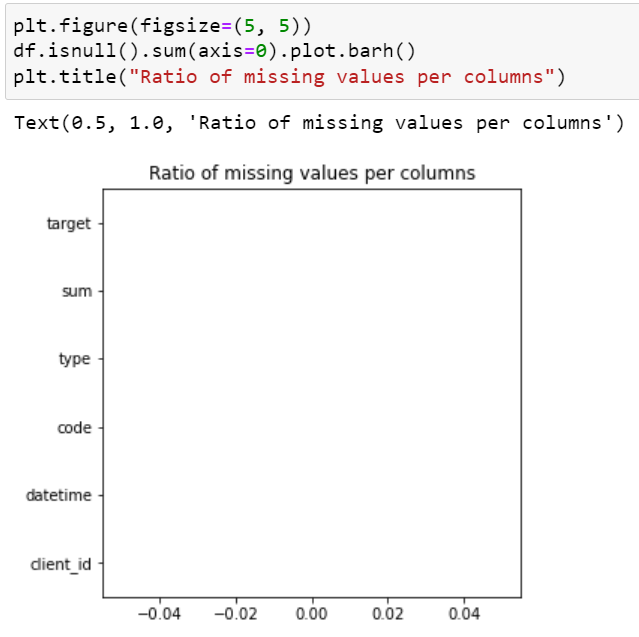


Here, we saving only those rows, where column target in not null

Изображение выглядит как стол

Автоматически созданное описание

Now, there is no missing values inside dataframe



**Exploratory Analysis**

Here, we show transaction’s type’s identification number and times it was made in descending order. As we see, people use at most ‘Покупка POS’ which is 21427 times.

**Изображение выглядит как текст

Автоматически созданное описание**

The same thing here, but with transaction’s code. The most used transaction code wad made for ‘Финансовые институты – снятие наличности автоматом’

Изображение выглядит как текст

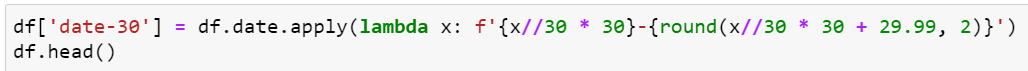
Автоматически созданное описание

Here, we wanted to divide the column date into parts of 30 days, so we would be able to calculate in which range the most transactions were made, identify the correlation and maybe predict the flood of transactions and prevent it



Изображение выглядит как текст

Автоматически созданное описание



Изображение выглядит как стол

Автоматически созданное описание

We see that most transactions were made in range of 420th and 450th day of the year

Изображение выглядит как стол

Автоматически созданное описание

Here, we want to calculate if after making all transactions person gone to plus or minus. To do that, we sum all transactions made by one person and give a status 0 – he’s in minus, 1 – in plus.

It could help us to predict a person, whether it is male or female according to number of transactions, its type and spend/gained money.

Изображение выглядит как стол

Автоматически созданное описание

We see that majority of people actually are in minus, but it could making a lot of transactions or just a few but which were expensive

Изображение выглядит как текст

Автоматически созданное описание

Here we output how many unique variables we have according to each column

6000 – persons

Изображение выглядит как текст, квитанция, снимок экрана

Автоматически созданное описание



While cleaning data, we found that column code in dataframe codes has some decency. Exactly, code between 1000 and 2000 has same category as ‘подрядчики’. Next one is rent hotel, cars, holidays, so we give category name as ‘Турисм’. And so on…

Изображение выглядит как текст

Автоматически созданное описание

Here is quantity of each category.

Изображение выглядит как текст

Автоматически созданное описание

In dataframe types, we did same thing as before. We categorize each gap with key words.

Then, according to the steps above, we merge each data frame to create a big and complete data set for continued work.Изображение выглядит как стол

Автоматически созданное описание

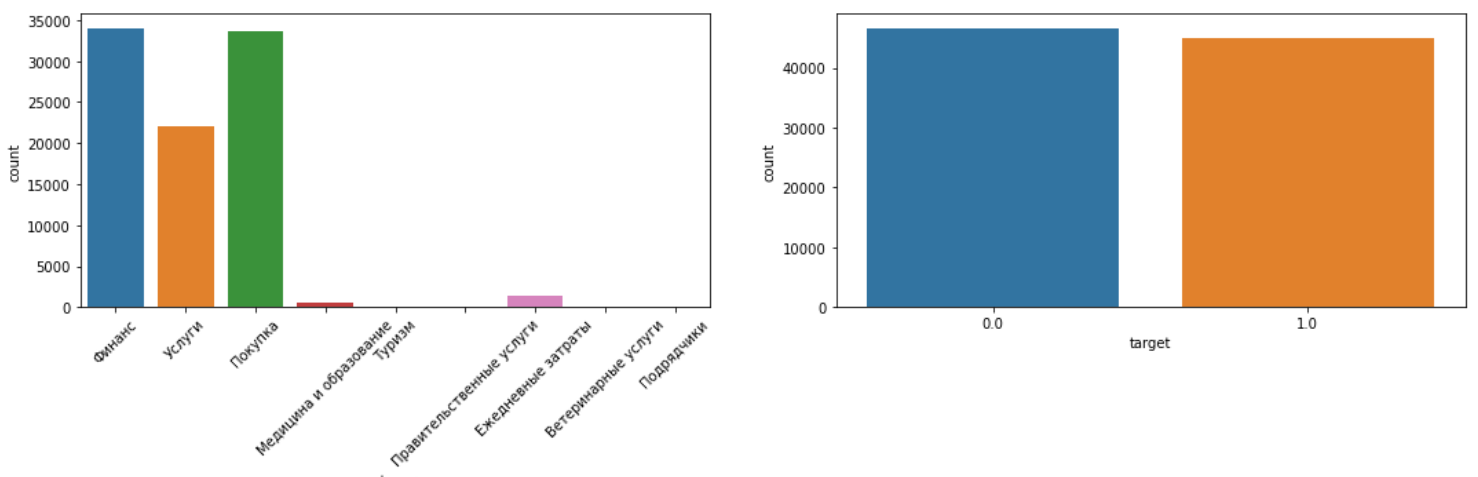
**Visualization**

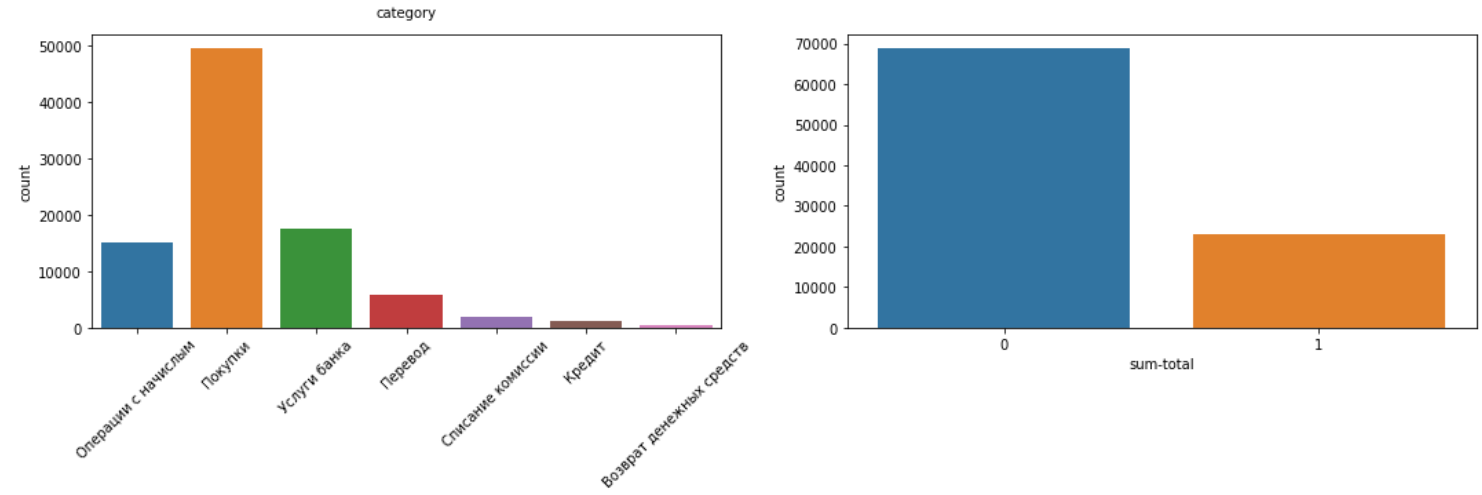
Now, let’s go through some visualization

**Изображение выглядит как текст

Автоматически созданное описание**

Here, we see which category people used at most.

****

****

Here we see that финансы are in plus which means that people at most take money. Услуги и покупка in minus zone so that means transaction of that group were the people needed to spend money. Others are +- the same, in balance

**Изображение выглядит как текст

Автоматически созданное описание**

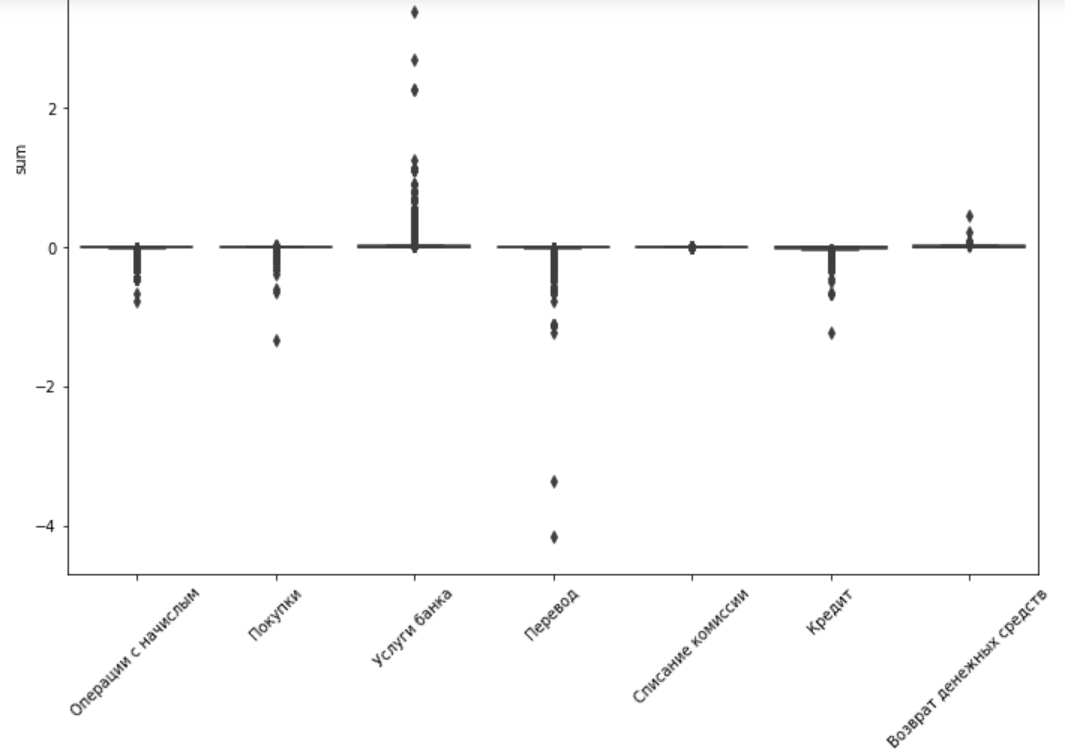
**Изображение выглядит как антенна

Автоматически созданное описание**

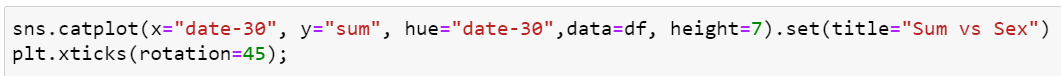
This graph shows which transaction is in plus or minus

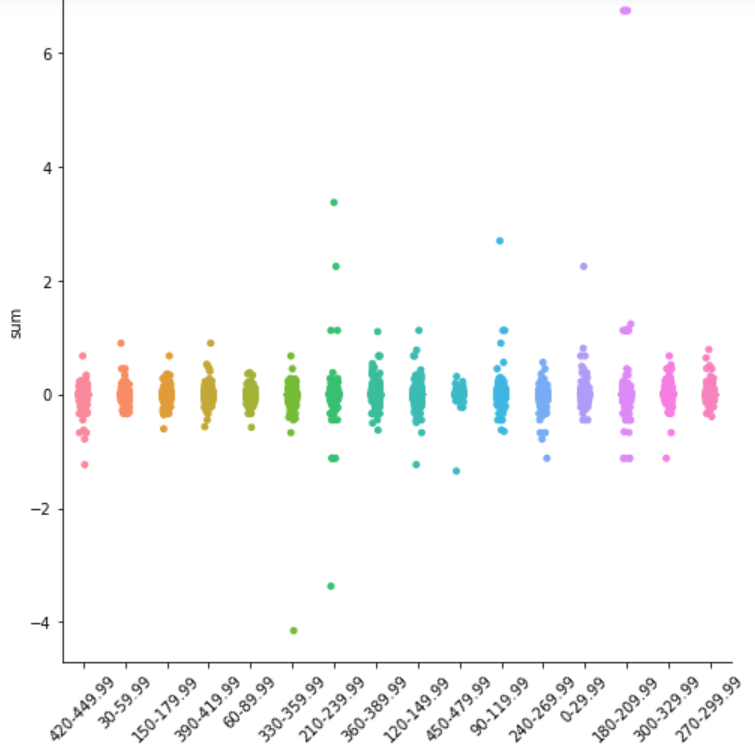
**Изображение выглядит как текст

Автоматически созданное описание**

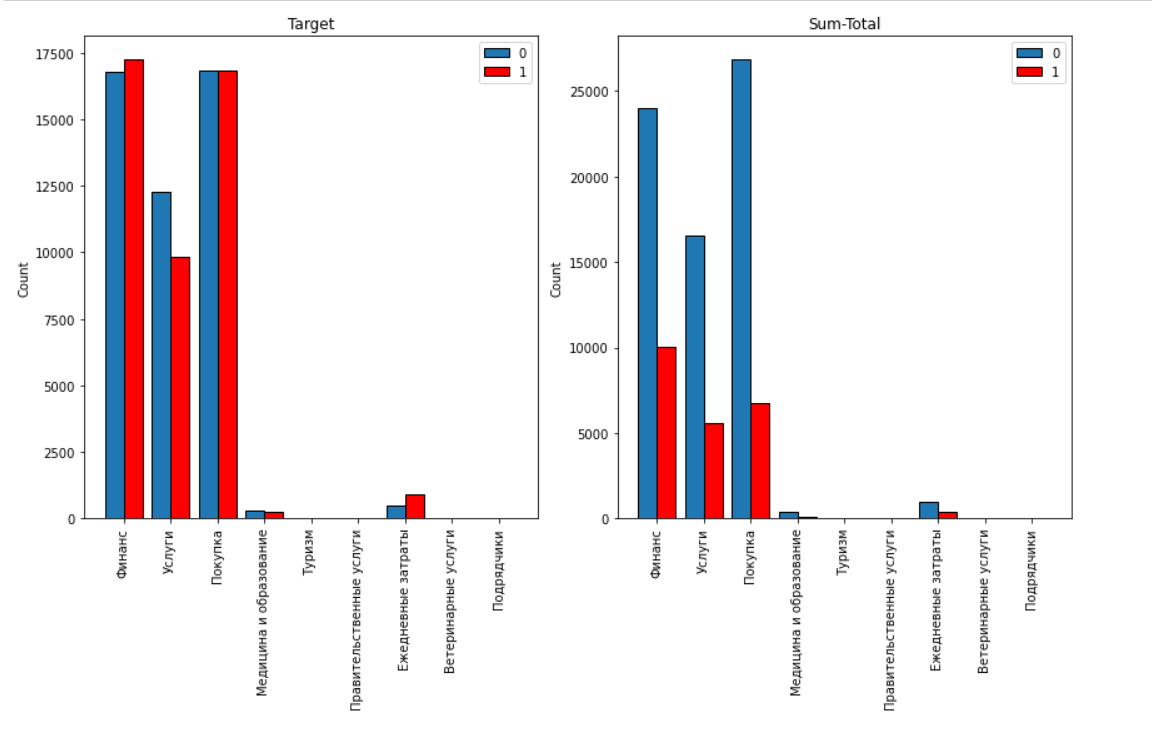
****

Here we see that most operations were made in range 420-450. Also we could see whether people spend or took money of that period, so this helps us to predict the gender somehow.

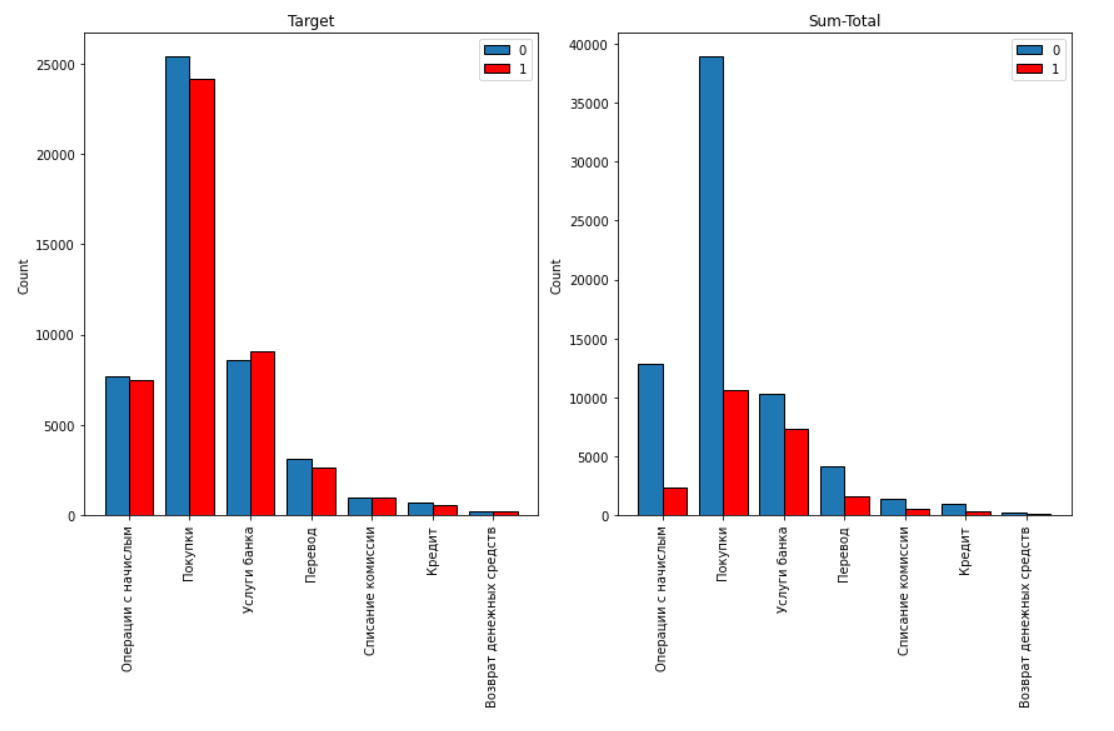
****

****

In our dataframe we had target which is 0 and 1, but we don’t know which one is male or female, however we according to the graph, we can make assumption. Here we consider group of transactions

****

the same is here, but we consider the transactions itself

****