Final Project

**Report**

Teacher: Zhanat Karashbaeva

Students: Alisher Alibekov, Miras Beketaev, Amina Tasbulatova, Zhaniya Bekzhan

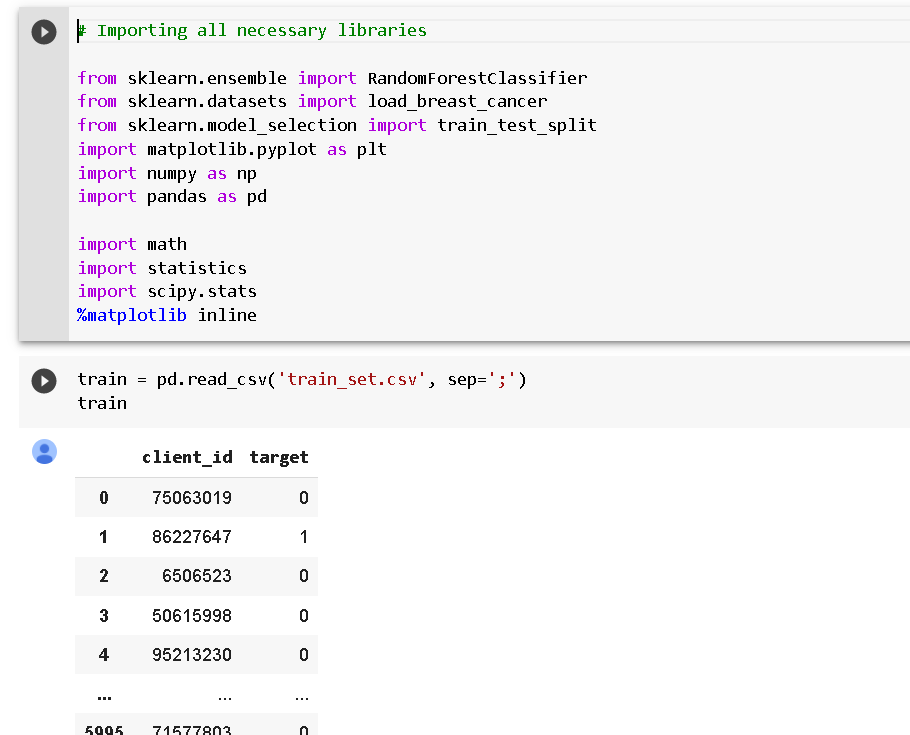
**INTRODUCTION**

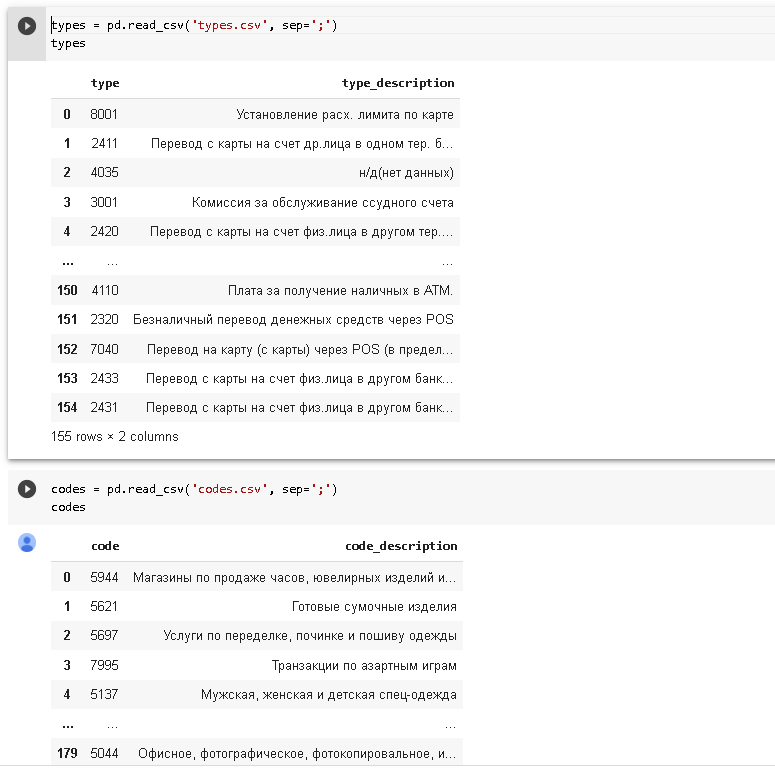
In this case, we have to use an online e-commerce database, which contains a year's worth of customer transactions from an e-commerce website. The goal here is to categorize the customer as quickly as possible so that the marketing team can send the right offers to the right customers.

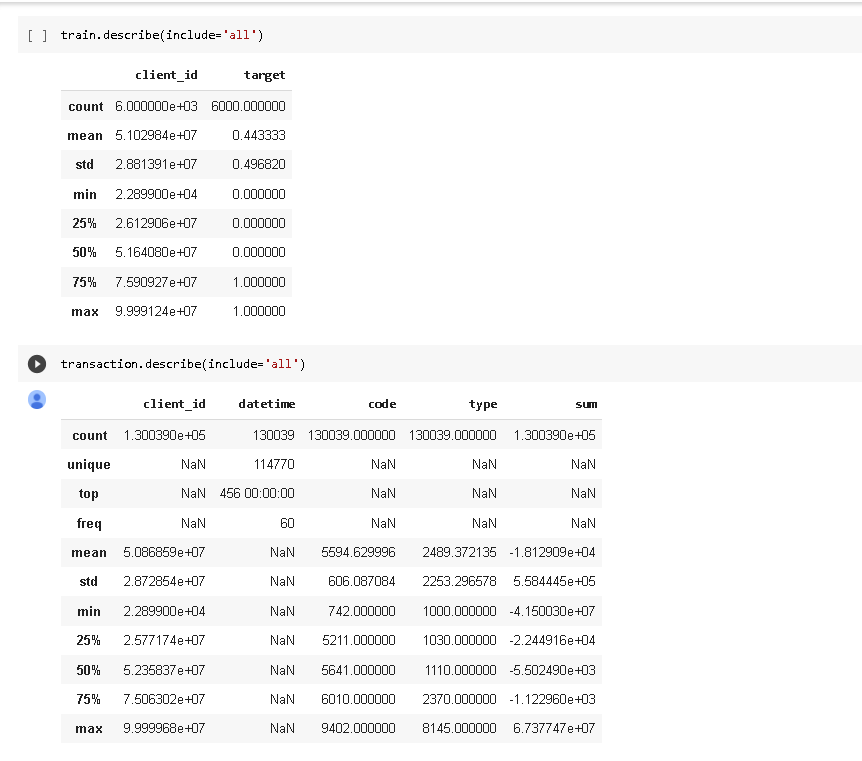
**MAIN PART**

# Part 1. Explore the dataset. Do the descriptive statistics.

The first step was to perform data preparation, which consists of processing null and duplicate values, as well as performing some descriptive statistics to get a basic understanding of our data. Then we did a preliminary analysis of the data using function visualization and created new functions for further unsupervised learning.

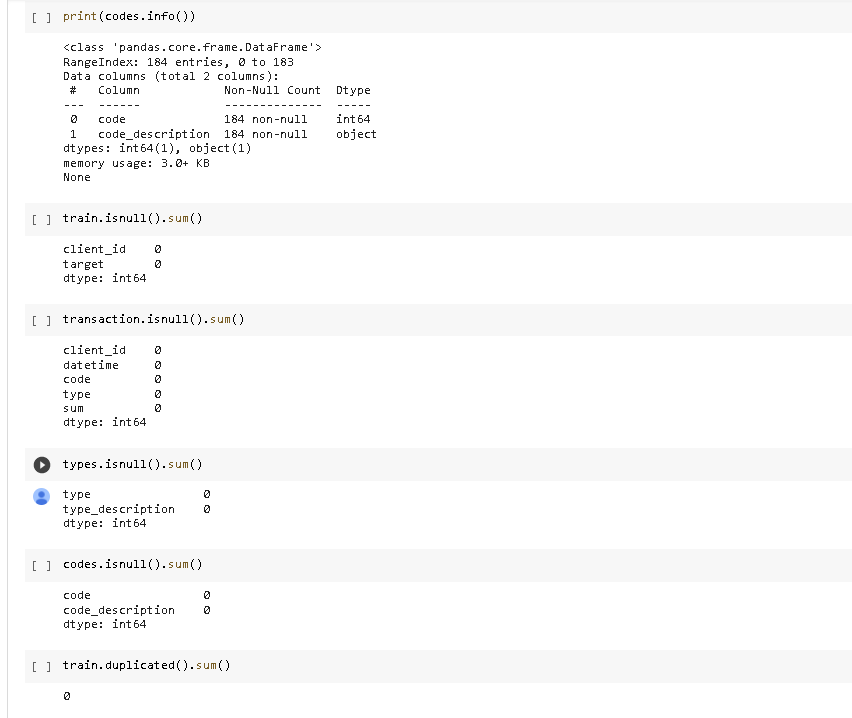
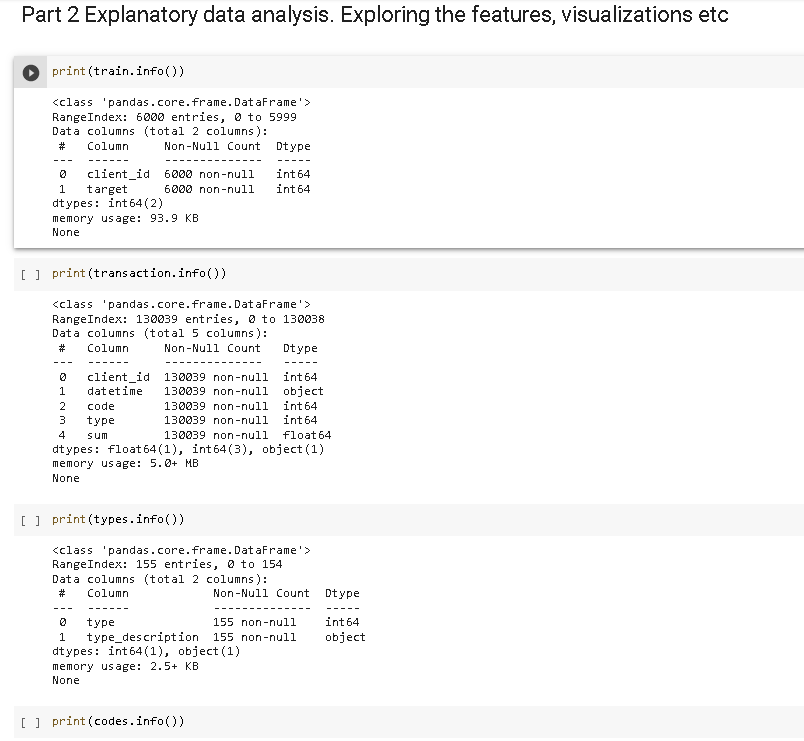


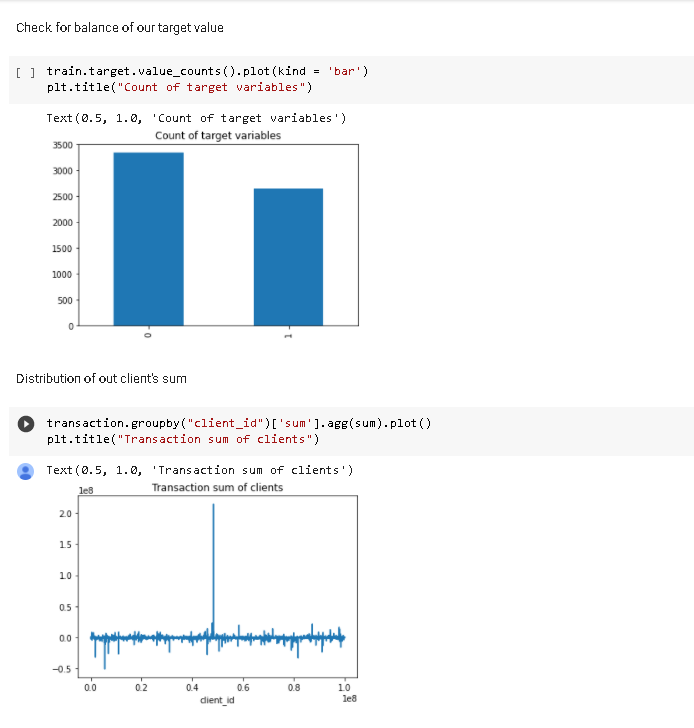
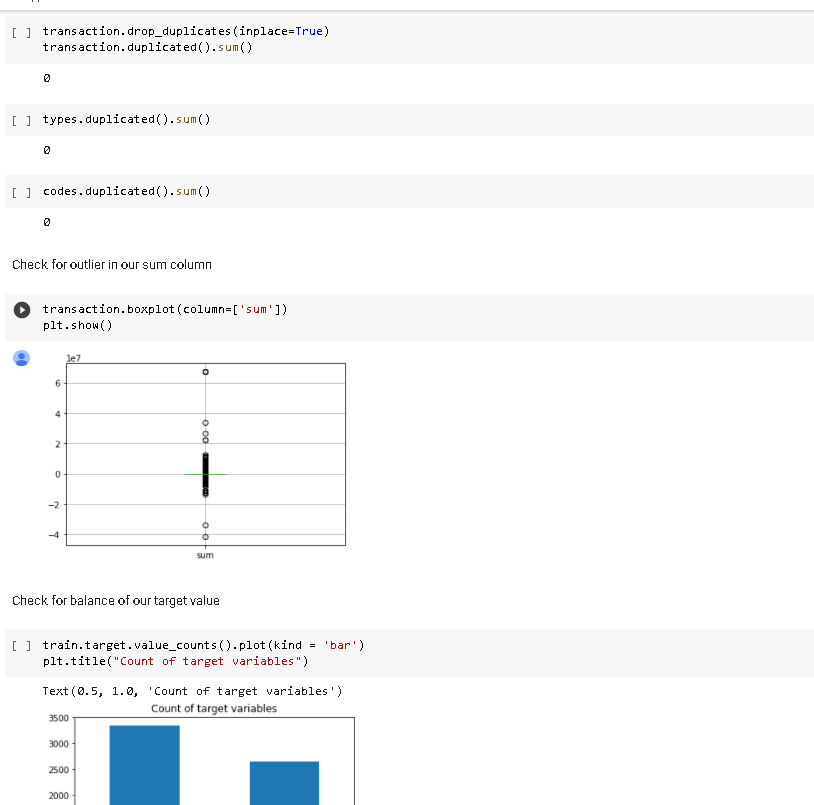
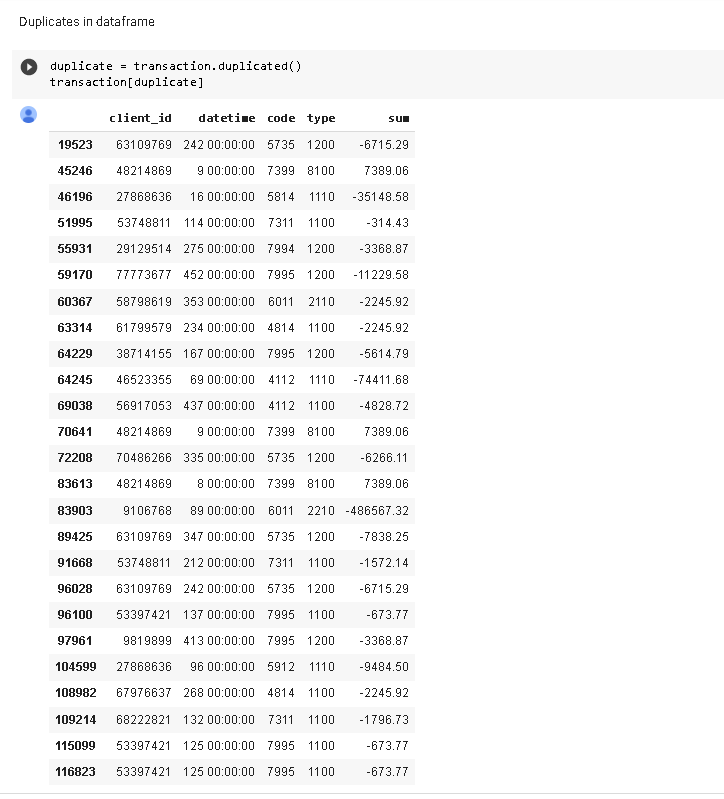


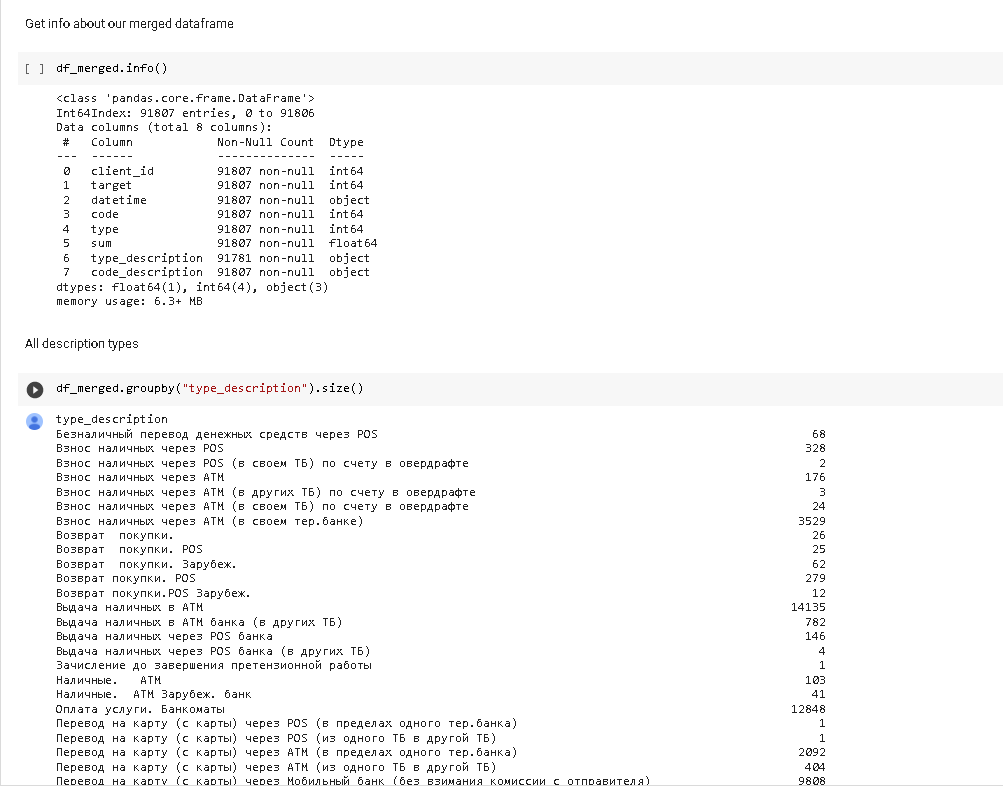


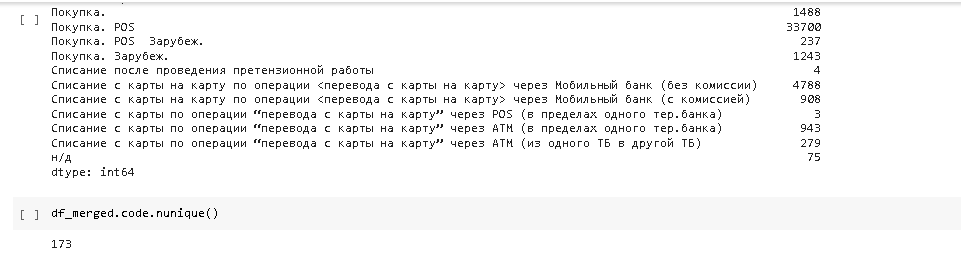
And so on…

# Part 2. Explanatory data analysis. Exploring the features, visualizations etc.



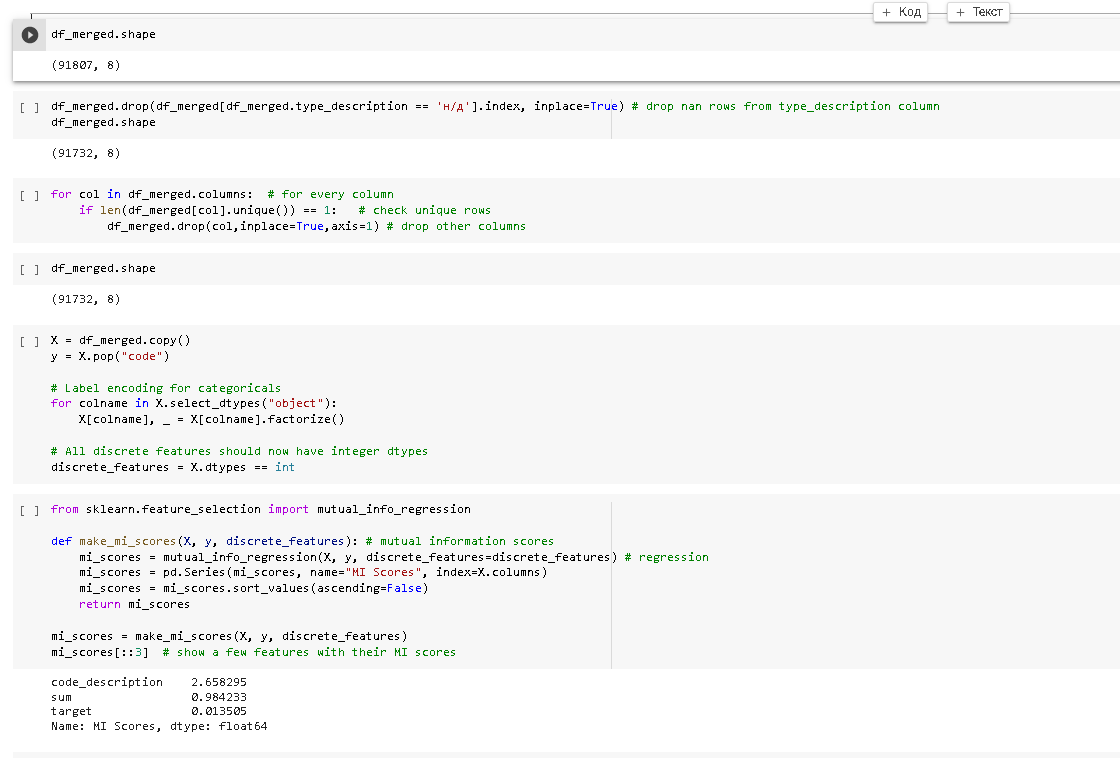


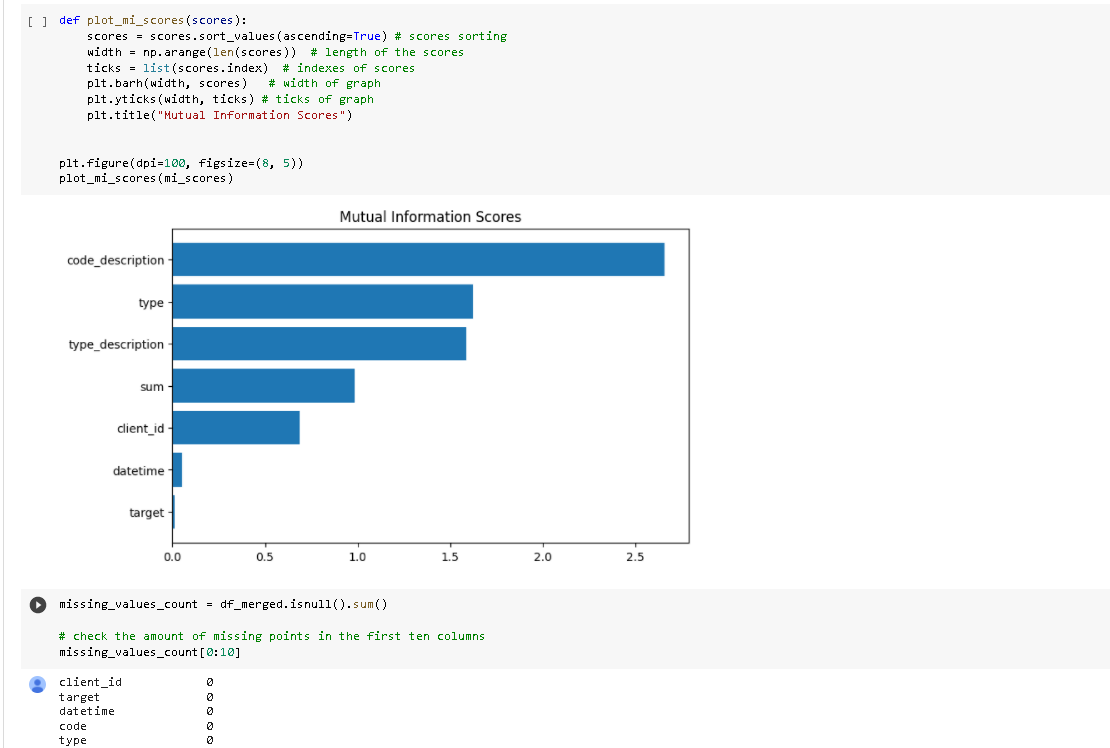




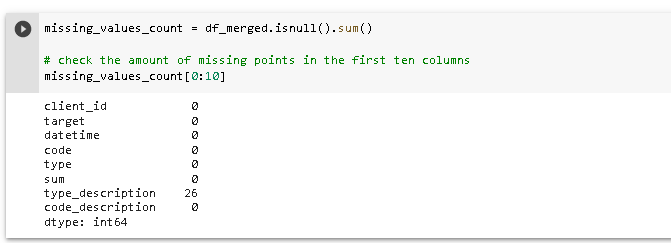
# Part 3. Feature engineering. Encodings, generating the features from date-time, sum and from other columns.

**Mutual information** describes relationships in terms of uncertainty. The mutual information (MI) between two quantities is a measure of the extent to which knowledge of one quantity reduces uncertainty about the other. In probability theory and information theory, the mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" obtained about one random variable by observing the other random variable.



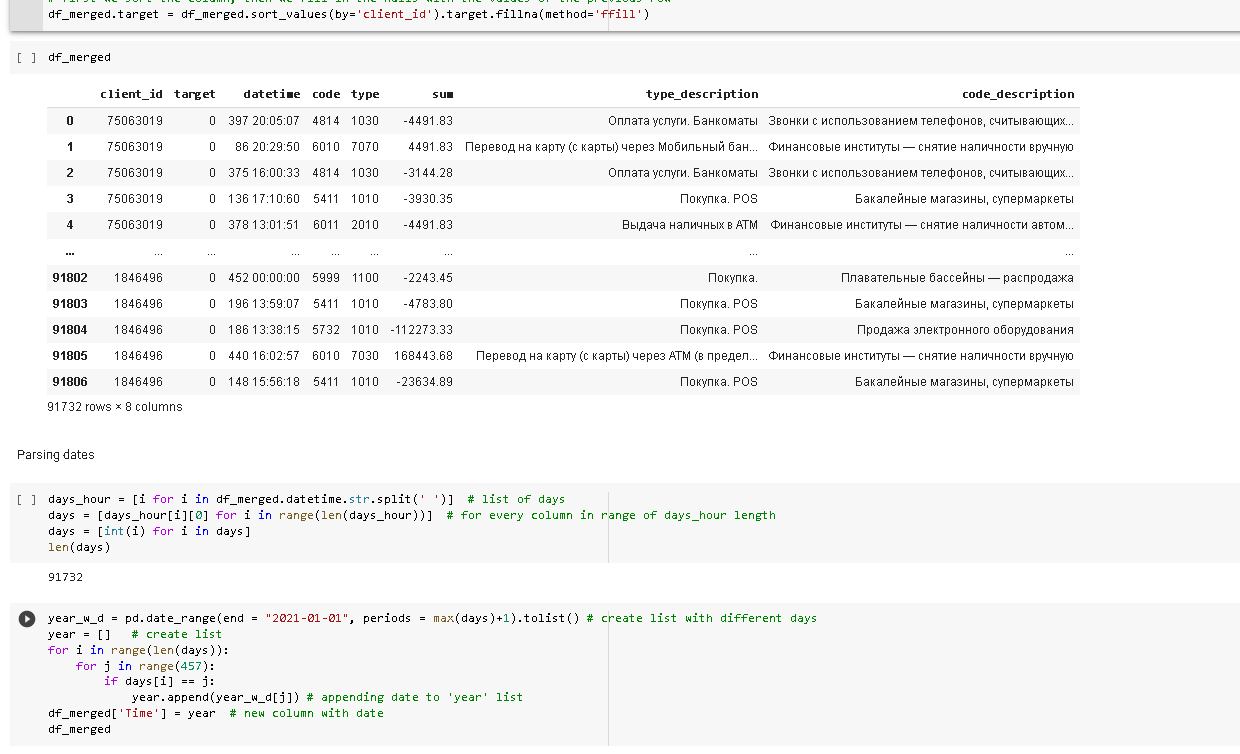


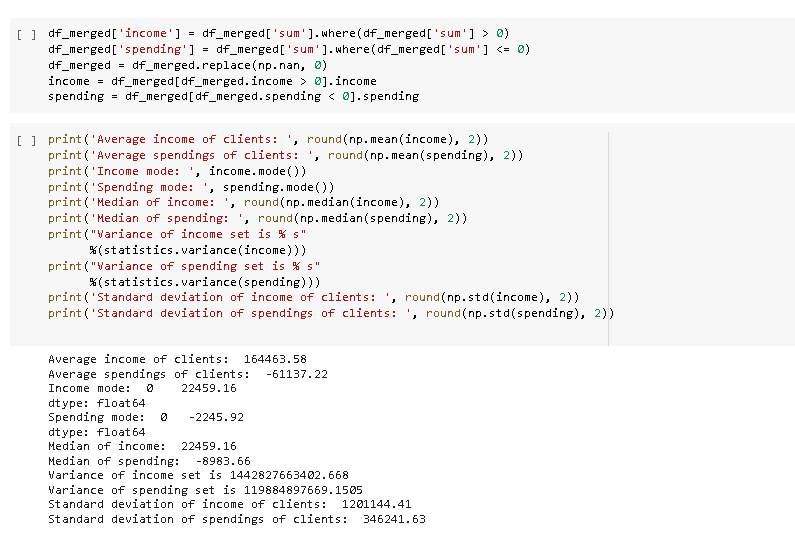
Check the amount of missing points in the first ten columns:



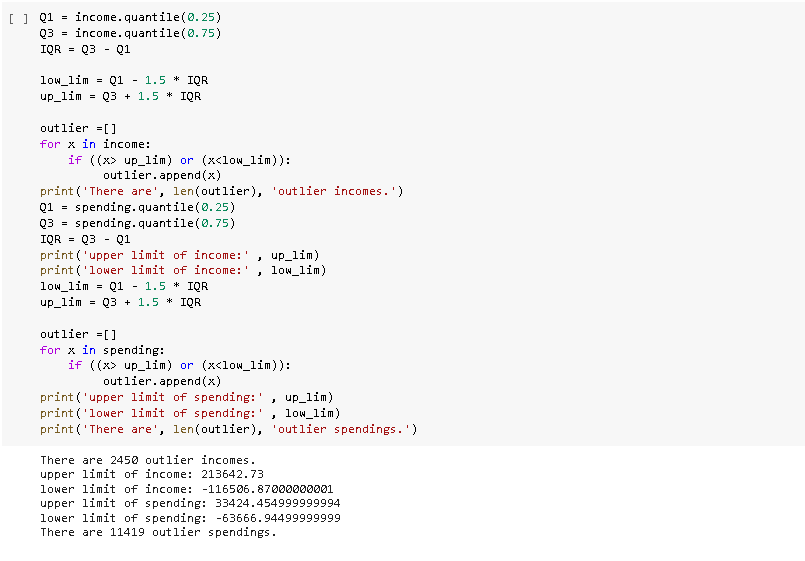
Fill NaN values in the target column.

First we sort the column, then we fill in the nulls with the values of the previous row:





# Outliers defining



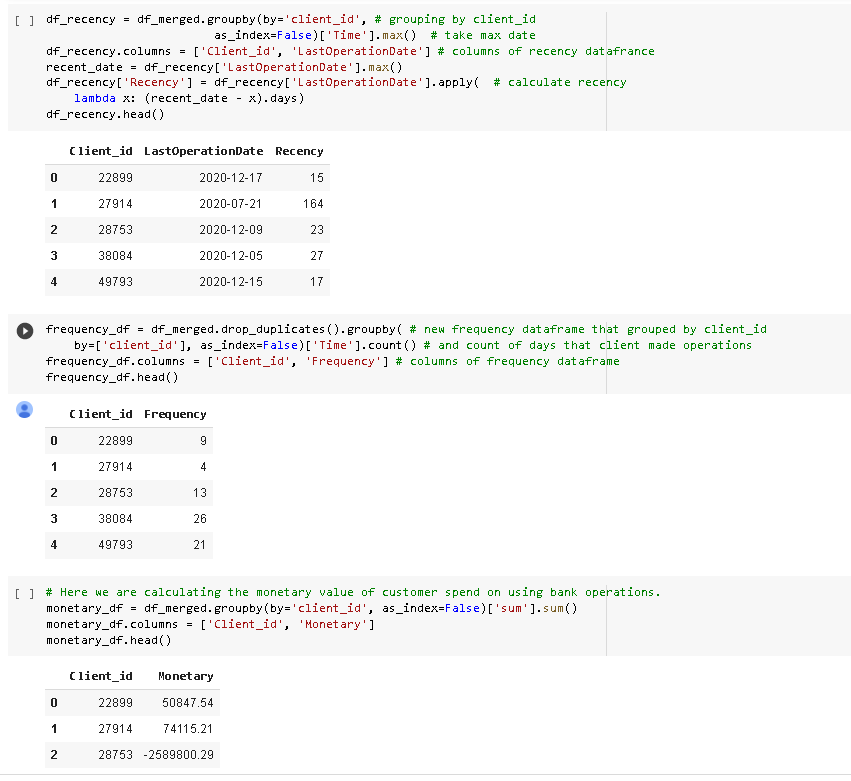
# RFM analysis

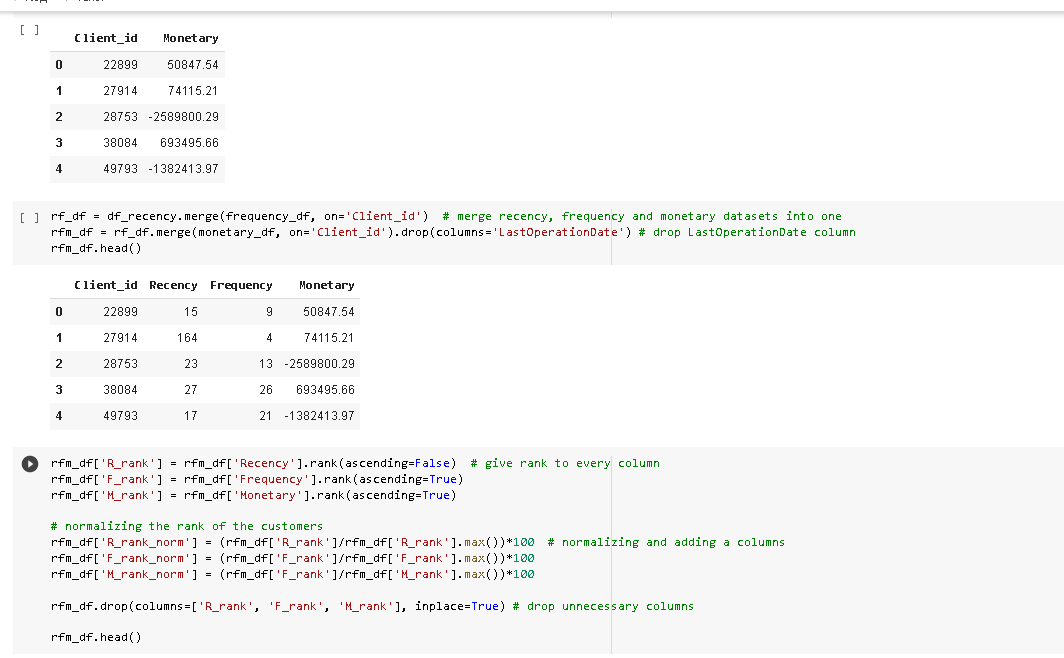
The “RFM” in RFM analysis stands for recency, frequency and monetary value. RFM analysis is a way to use data based on existing customer behavior to predict how a new customer is likely to act in the future. An RFM model is built using three key factors:

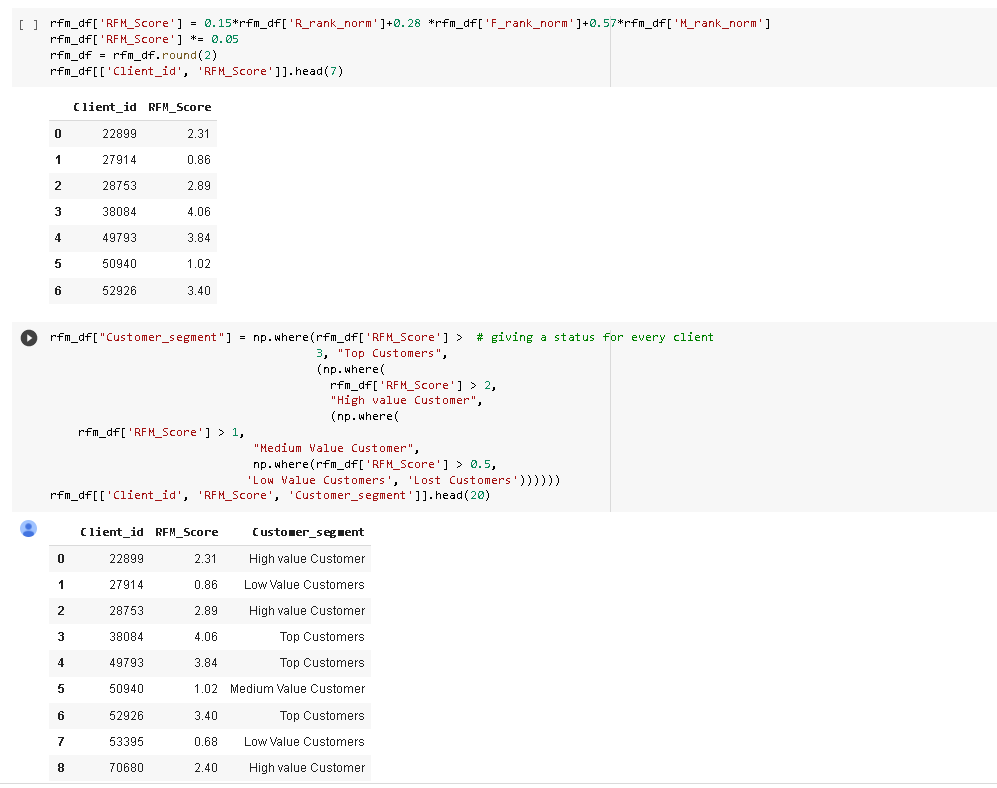
1) how recently a customer has transacted with a brand

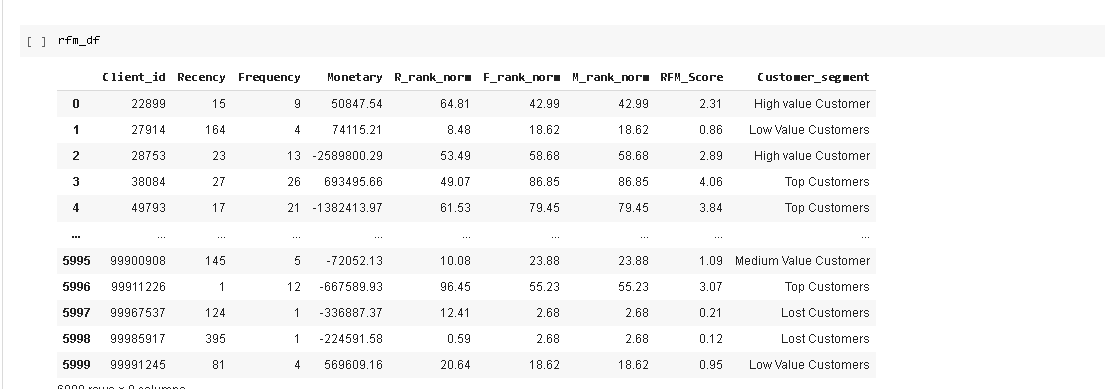
2) how frequently they’ve engaged with a brand

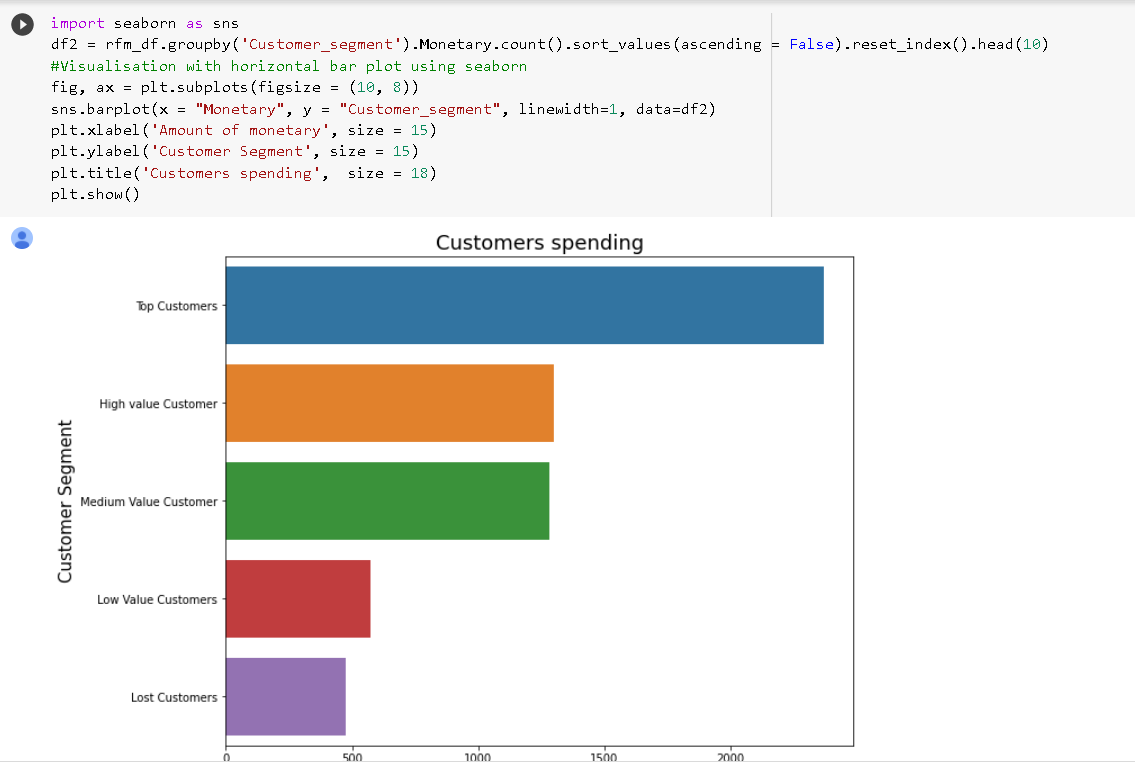
3) how much money they’ve spent on a brand’s products and services

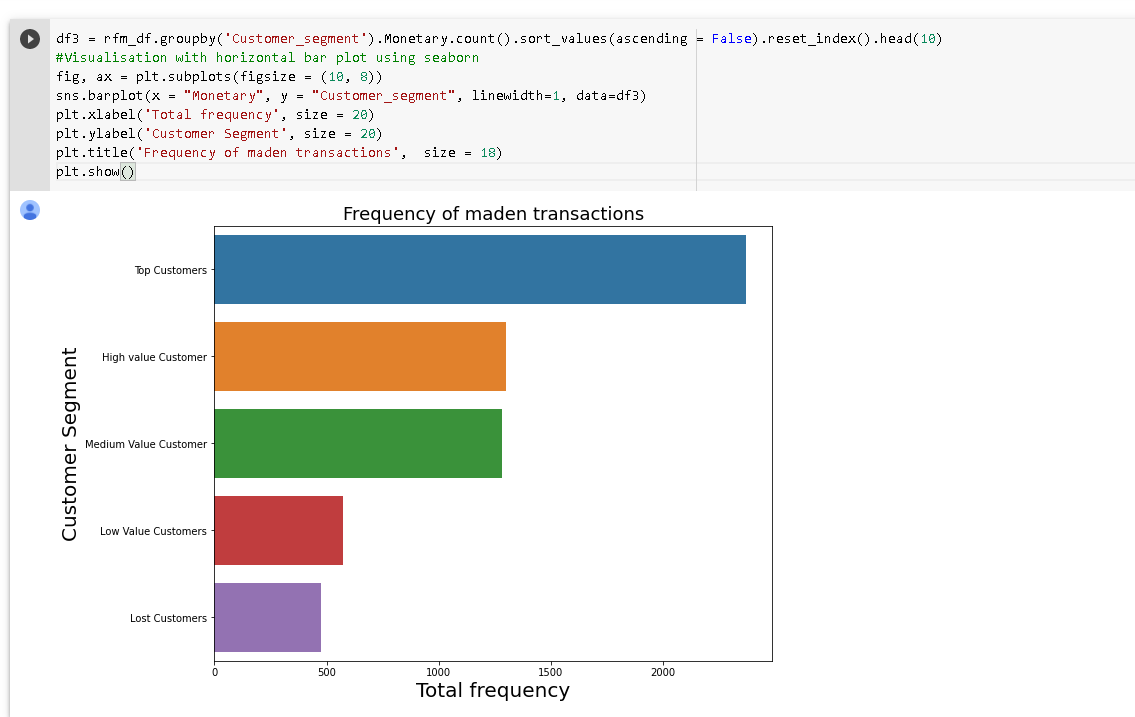


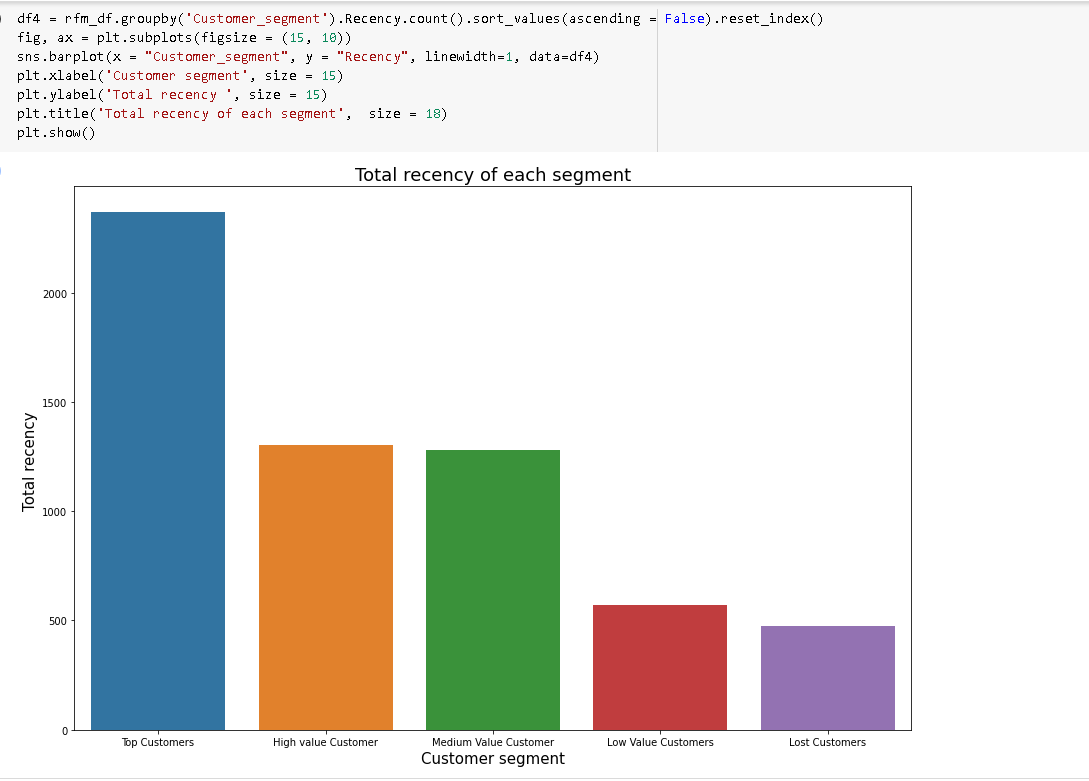






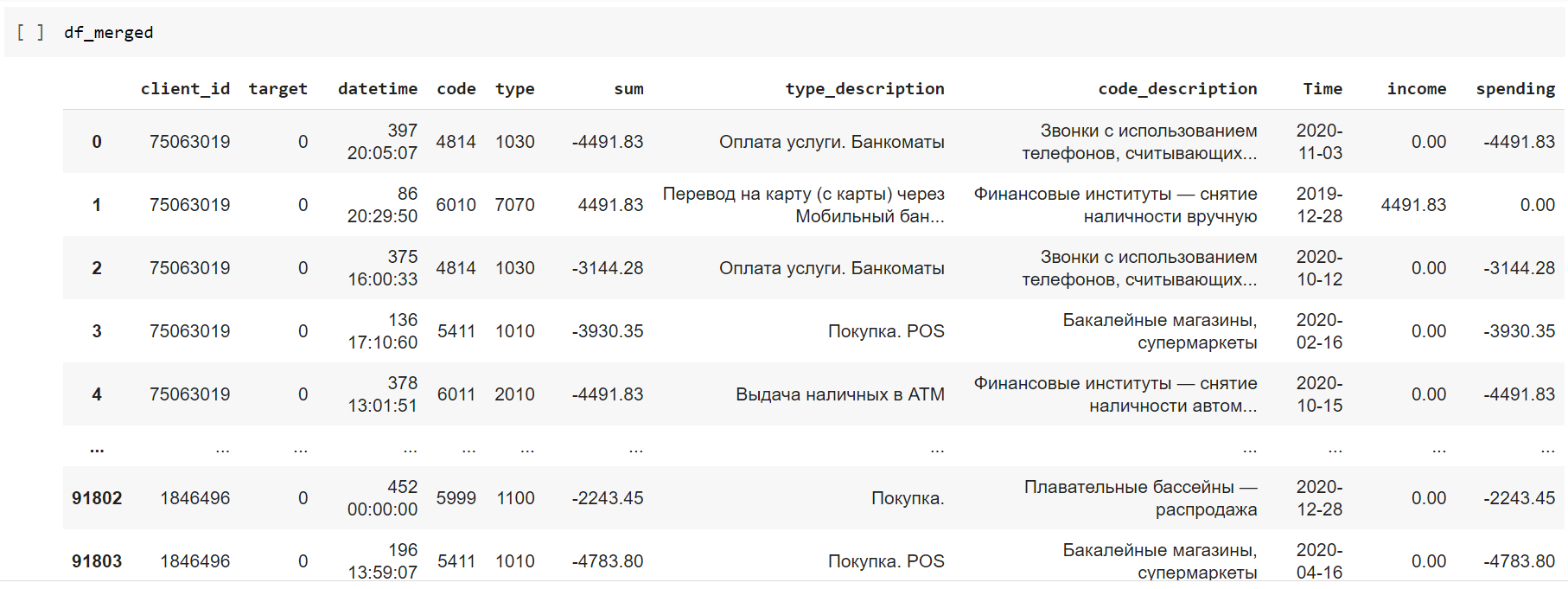


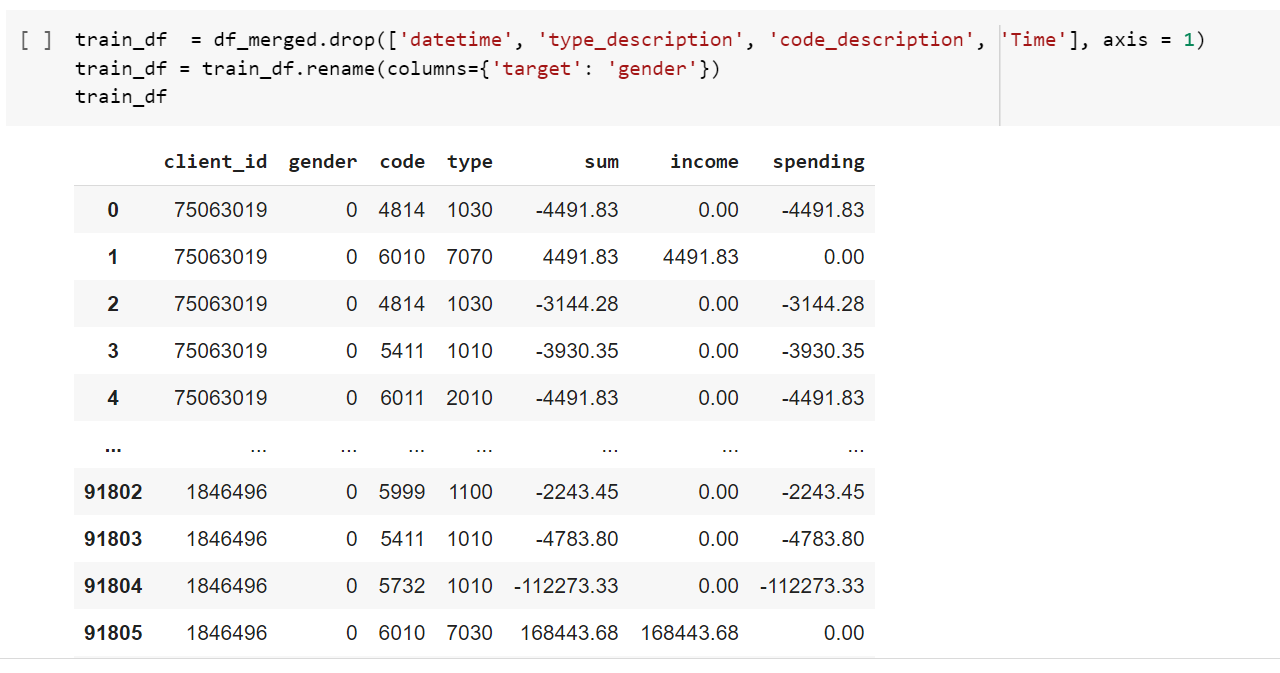


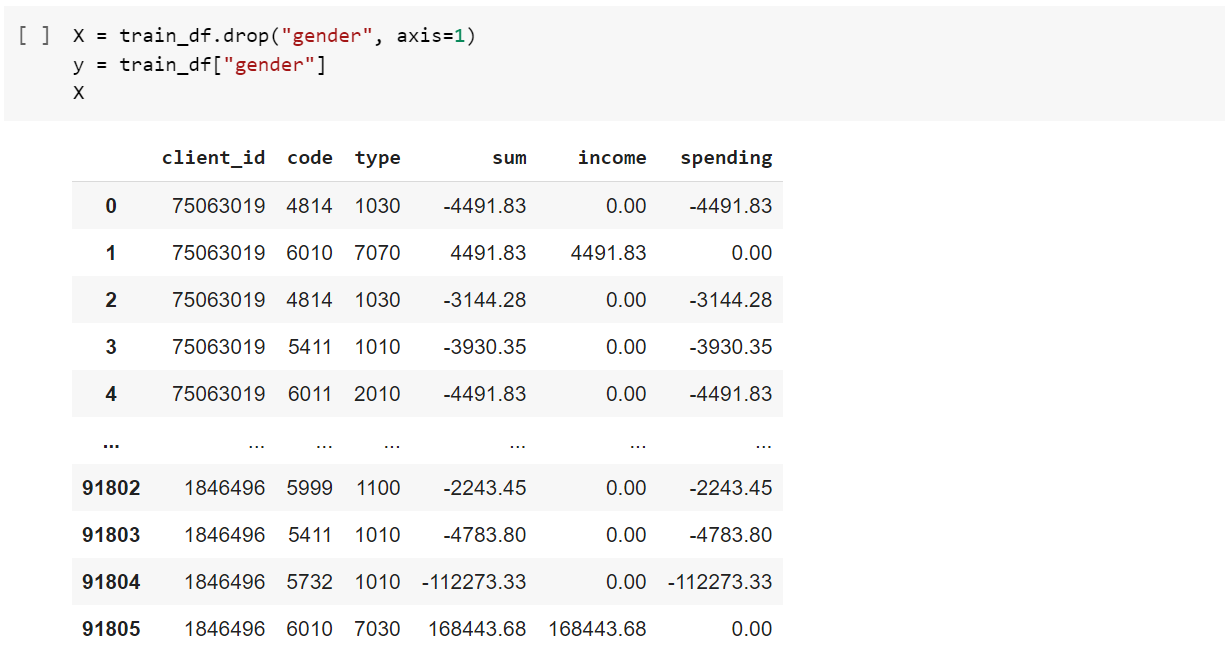


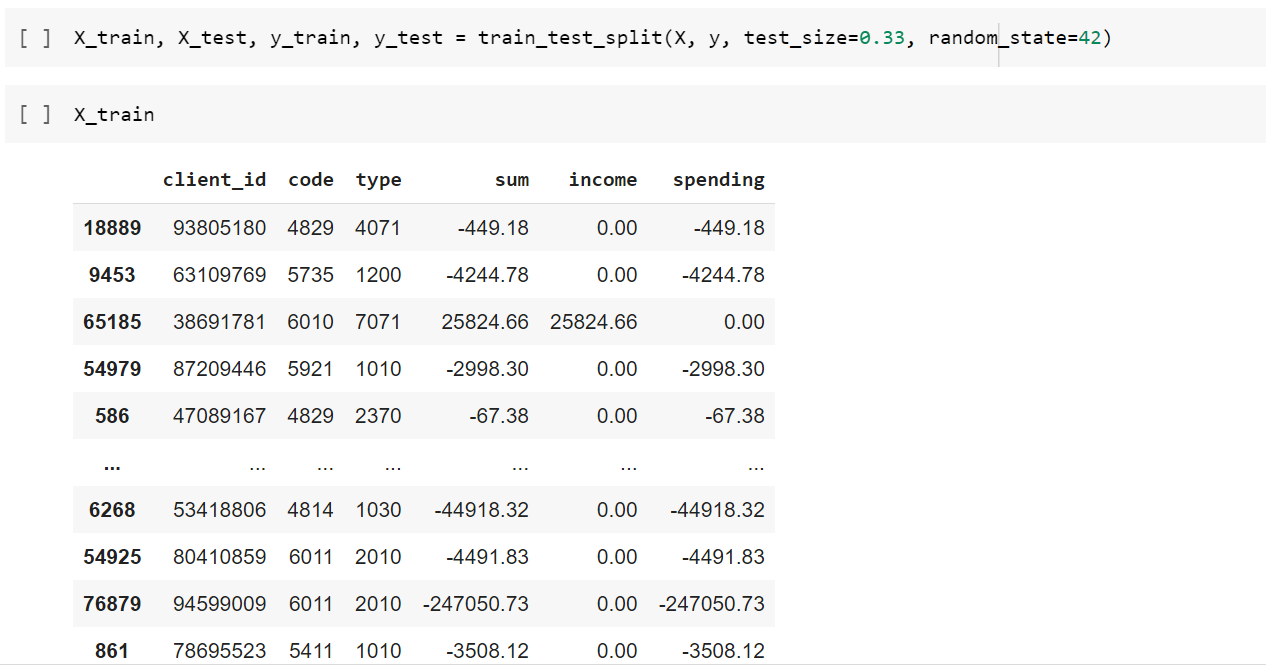
# Part 4. Supervised learning. Build a model for predicting the gender of the clients. Decision Trees, KNN, Random Forest. Tune the hyper parameters, grid search, cross validation etc. Visualization of the models etc.

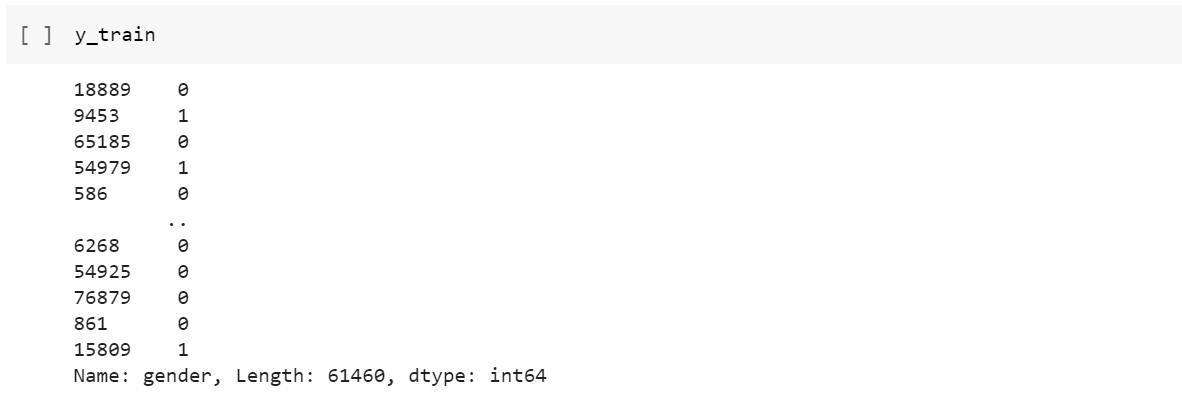
### Preparing datasets

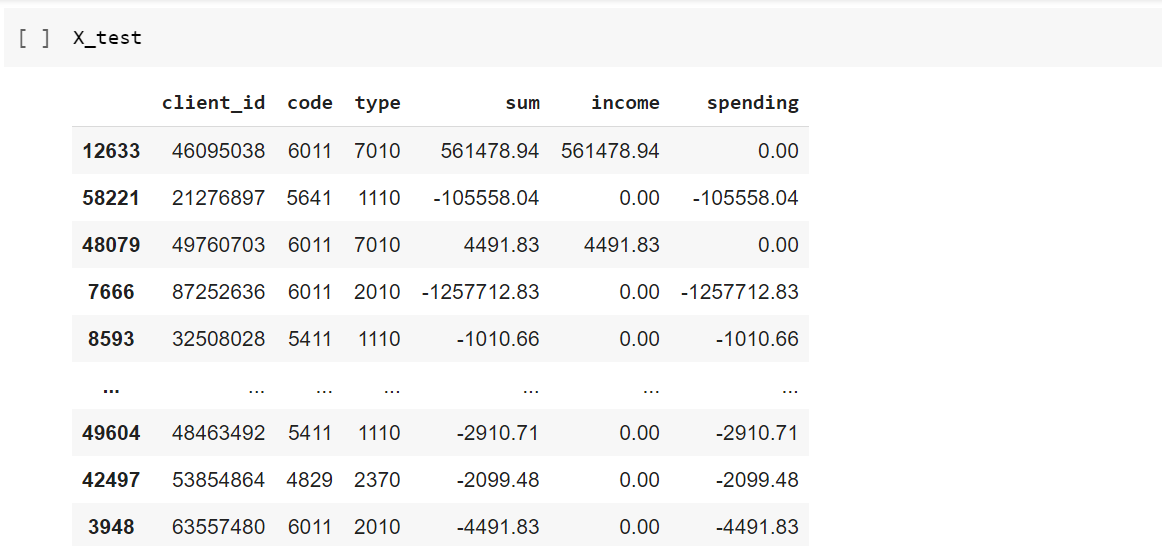












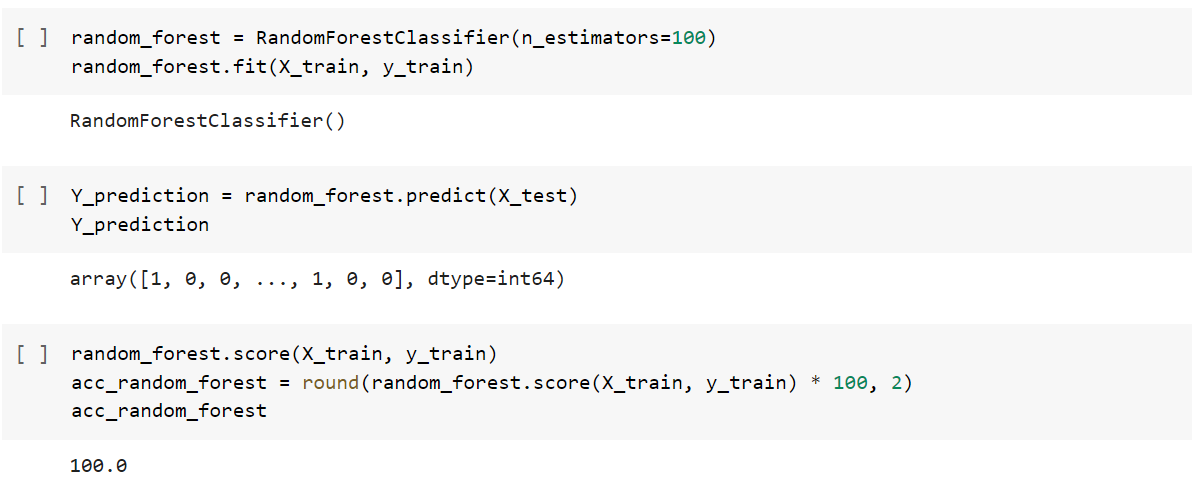


### 

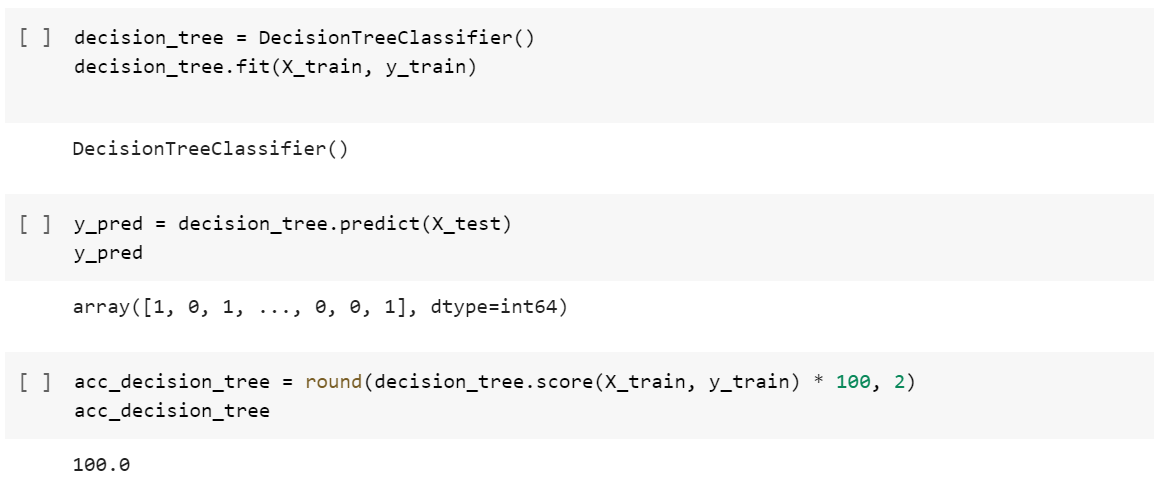
### K-Nearest Neighbors



### Random Forest

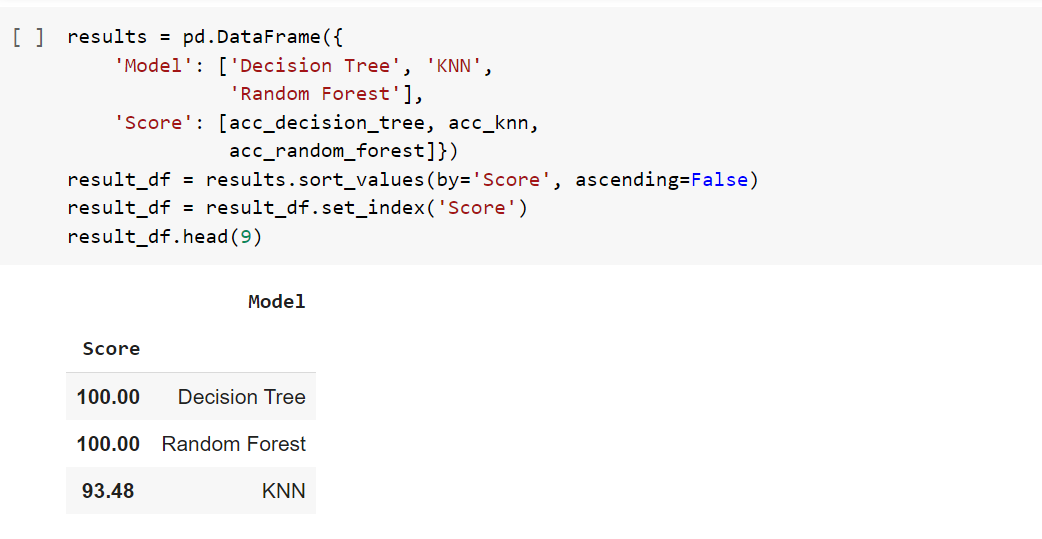


### Decision Tree

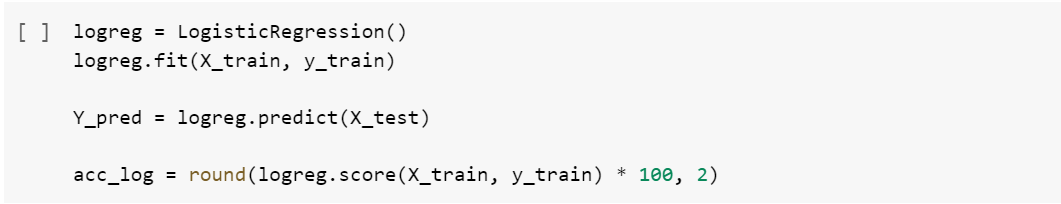




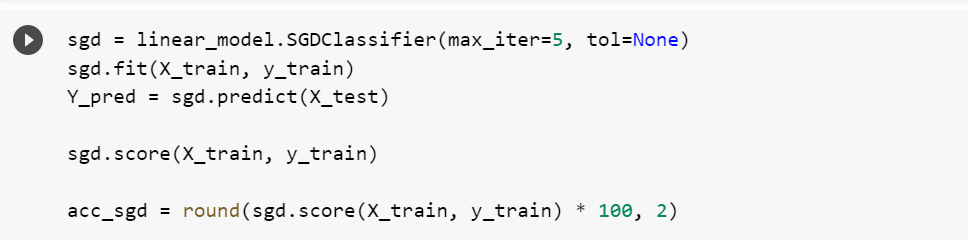
### Defining best model



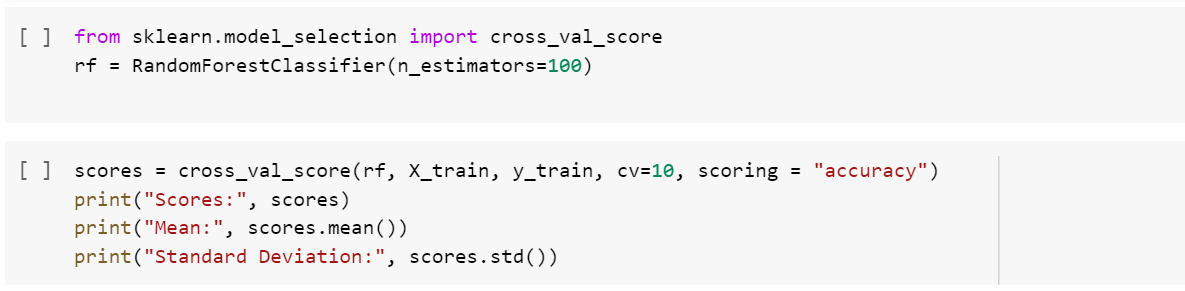
### Logistic Regression



### Stochastic Gradient Descent (SGD)

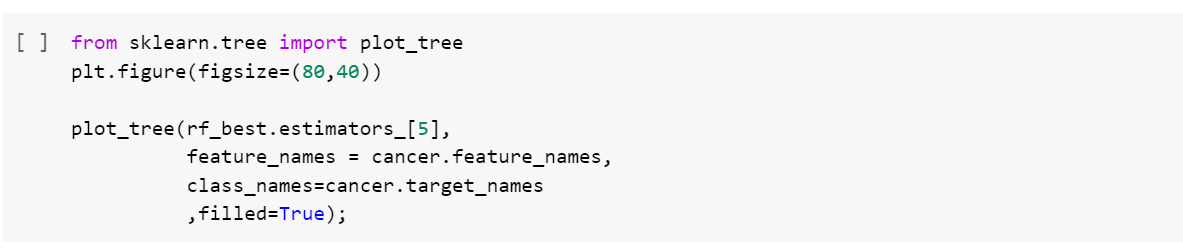


### K Fold cross validation



### Hyper Parameter Tuning



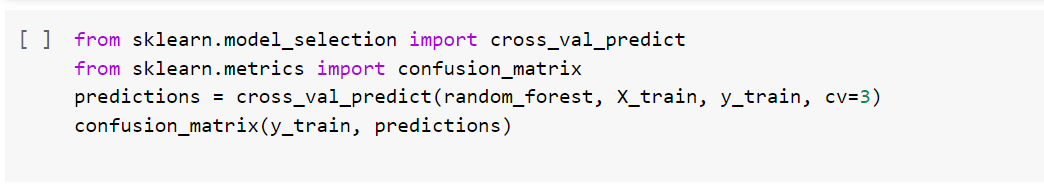




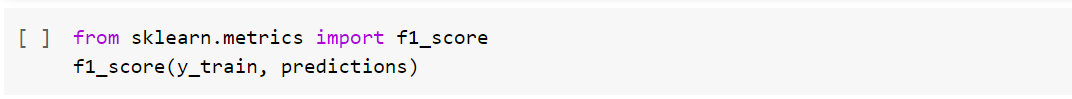
# Part 5. Analyze models, Result comparison, ROC/AUC, precision and recall curves, deep analyzing.

### Confusion Matrix

**Confusion matrix:** shows the actual and predicted labels from a classification problem

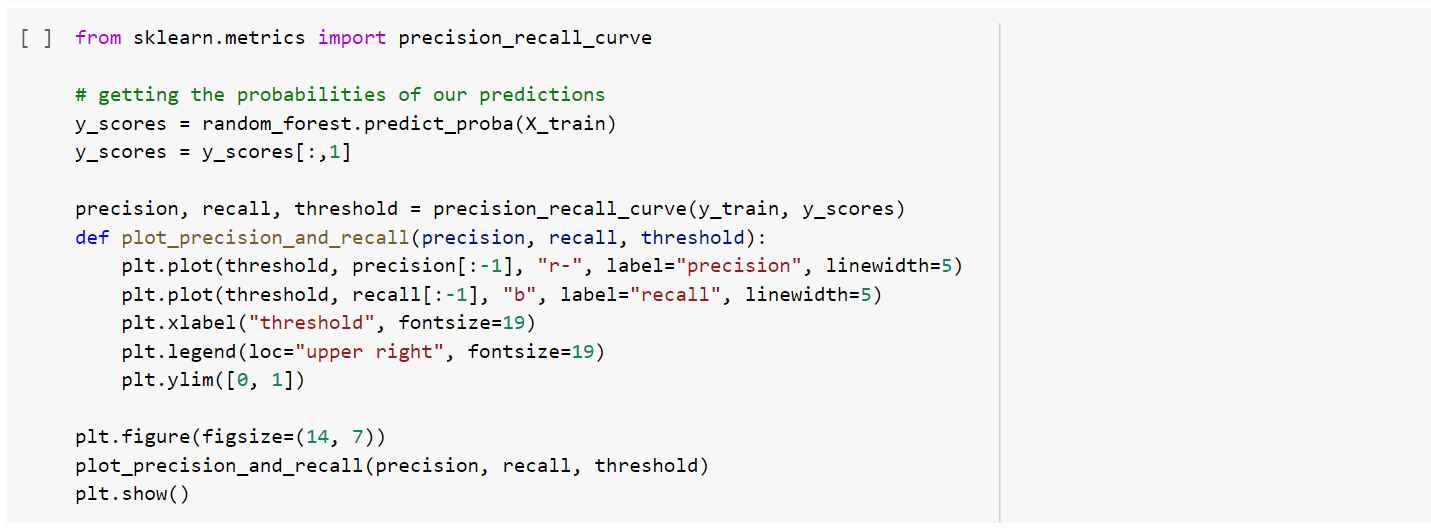


**F1 score:** single metric that combines recall and precision using the harmonic mean

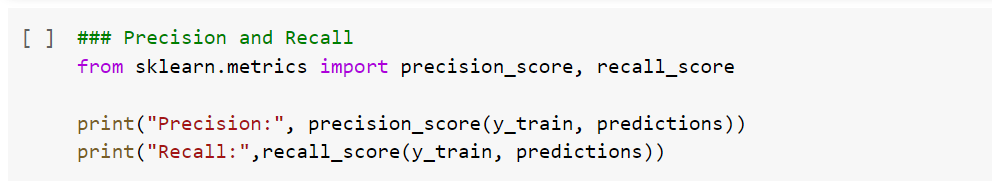


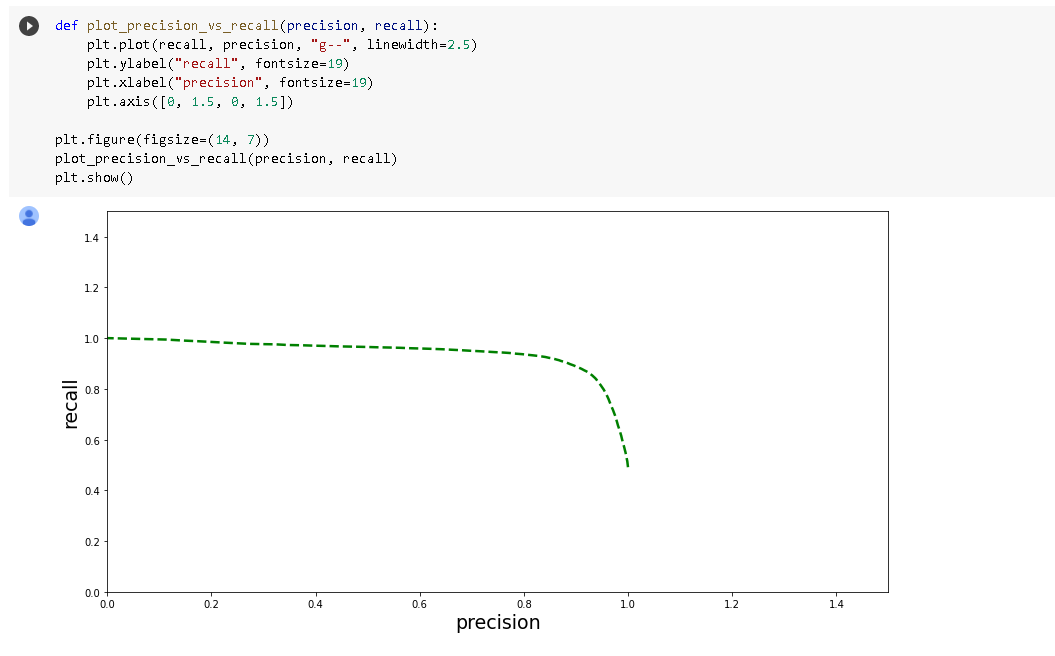
### Precision Recall Curve

#### Recall: ability of a classification model to identify all relevant instances



**Precision:** ability of a classification model to return only relevant instances

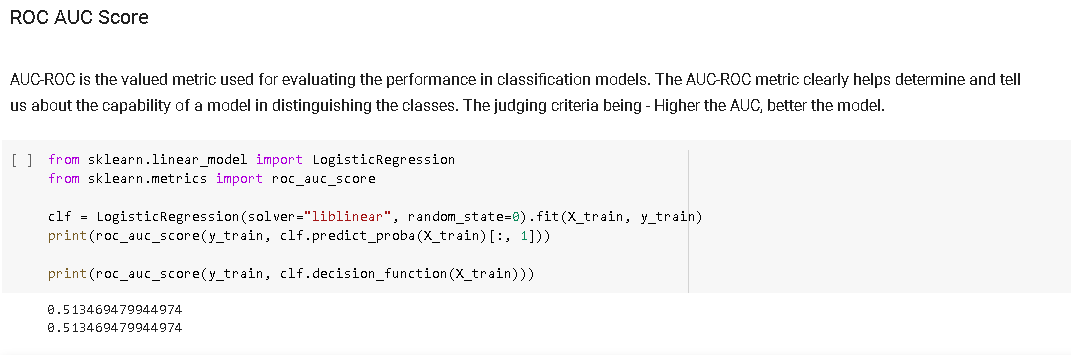




### ROC AUC Score

**Receiver operating characteristic (ROC) curve:** plots the true positive rate (TPR) versus the false positive rate (FPR) as a function of the model’s threshold for classifying a positive

**Area under the curve (AUC):** metric to calculate the overall performance of a classification model based on area under the ROC curve



### ROC AUC Curve

We analyzed 4 datasets that were combined into one merged dataset. We checked each dataset for duplicates and Nulls and removed them if there were any. We determined the outliers of sum value , checked for balance of target value, determined the distribution of our client's sum and plotted them. Next, we added feature engineering, defined Mutual information , their scores , then we made date parsing with our merged dataframe . Using the RFM principle we get 4 separated types of clients 'Top Value', 'High value', 'Medium Value' and 'Low value' clients.

We tend to use better-suited metrics such as recall and precision may seem foreign, we already have an intuitive sense of why they work better for some problems such as imbalanced classification tasks. Statistics provides us with the formal definitions and the equations to calculate these measures.

**CONCLUSION**

We have used different methods to explore and get information about transactions in bank accounts. With Python functionality and other modules we analyzed 4 datasets that were combined into one merged dataset. We have also visualized the transactions.

And hierarchical clustering was implemented in order to segment customers based on the similarity of their transactions.

Data science is about knowing the right tools to use for a job, and often we need to go beyond accuracy when developing classification models. Knowing about recall, precision, F1, and the ROC curve allows us to assess classification models and should make us think skeptically about anyone touting only the accuracy of a model, especially for imbalanced problems.