FINANCIAL PROGRAMMING

Group Project – Financial Base Table

By SHKURIN Aleksandr, AYESHA Noor & RANA Inderpreet

Overview of the Project:

Create base table with the information extracted from financial dataset Berka. And finding the correlations between the various features useful for understanding the business trends to identify the customers who are eligible are not eligible to grant loan and credit card.

Data Exploration:

We used Customer data that was been provided to us in the form of various tables like credit card, daily transactions, account, loan, demographics, disposition, orders, and client information to create the data mart. There are 5,369 unique clients' observations with 86 columns in our Base Table.

Data Preparation:

As part of data preparation, we have followed bellow steps.

1. Each record of the dataset read describes the static characteristics of an account.
2. The "Age" and "Age Group" variables were calculated assuming current year as 1997.
3. For Client Table new columns "birthyear", "birth month", and "birthday" were extracted from "birth number" and a new column "Gender" was extracted from "birth month".
4. For Trans table –
    - Converted "trans_id" to integer so that we can index it
    - Added flags for null values
    - Added the columns transaction day, month, and year
    - Checked that the column doesn't have values that should be in Operations
    - Changed the values in type column to English for better understanding
    - Imputed missing operations value based on trans type combination
    - Filled missing values with "cc withdrawal" and "cash withdrawal" as per the customer % contribution in operation type
    - Renamed the k_symbols to more understandable names while making sure there are no wrong values.
    - Imputed missing K-symbol and null values for bank
    - account is blank only where operation is not through another bank hence dropped that column later as it has low significance
    - Identified outliers in the amount and replaced 2 * 3rd Quartile's outliers with 2 * 3rd quartile, and for 1st Quartiles / 2, we have replaced it with the outliers by 1st quartile/2.
    - Identified total withdrawal and credit per account.
    - Used group by to analyse the missing values and thereby assigning them the suitable k_symbols
    - Checked for the outliers in Amount and replaced them as per their quartile

5. For Card Table we have extracted columns like card type, issued_card, issued_year_card, issued_month_card, issued_day_card.
6. Disp table is used to identify type of owner.
7. From order table we have extracted columns such as account_id_order, order_count_order, recipient_bank_count_order, recipient_account_count_order, total_payment_amount_order,total_household_payments_order, total_loan_payments_order,total_insurance_payments_order, total_lease_payments_order, total_unknown_payment_order.
8. In Orders table fixed the names for all k_symbols.
9. For Loan table checked outliers and replaced them as per their quartile in terms of amounts. Further grouped the loan table as per customer.
10. For District table renamed the columns named as “A1 – A16” to more meaningful names

Our Final base table has 2230 rows and 77 columns.

```
base_table.head()
```

✓ 0.5s Python

	account_id	bank_district_id	frequency	acc_open_year	acc_open_month	acc_open_day	LOR	client_id	type	client_district_id	...	status_desc_loan	loan_issued_in_97	card_id_card	type_card
0	576	55	MONTHLY FEE	1993	1	1	3	692	OWNER	74	...	0	0.0	0.0	0
1	3818	74	MONTHLY FEE	1993	1	1	3	4601	OWNER	1	...	0	0.0	0.0	0
2	704	55	MONTHLY FEE	1993	1	1	3	844	OWNER	22	...	0	0.0	0.0	0
3	2378	16	MONTHLY FEE	1993	1	1	3	2873	OWNER	16	...	0	0.0	0.0	0
4	2632	24	MONTHLY FEE	1993	1	2	3	3177	OWNER	24	...	0	0.0	0.0	0

5 rows × 77 columns

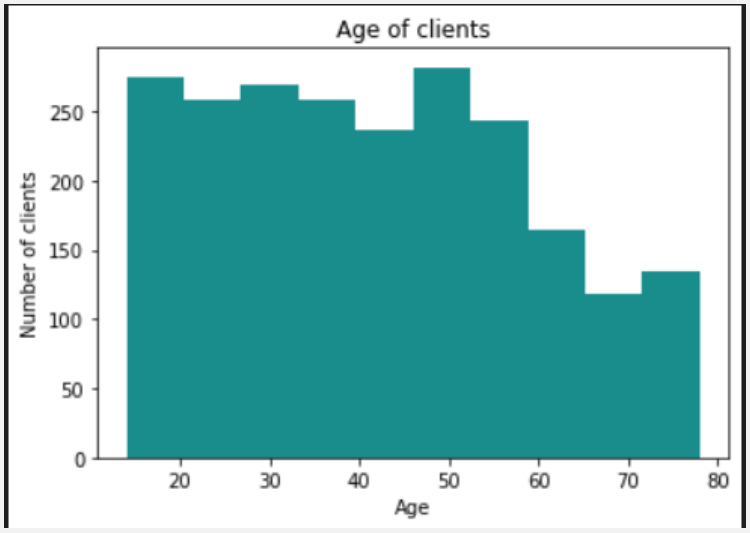
```
#base_table.isna().sum().sum()
base_table.shape
```

✓ 0.3s

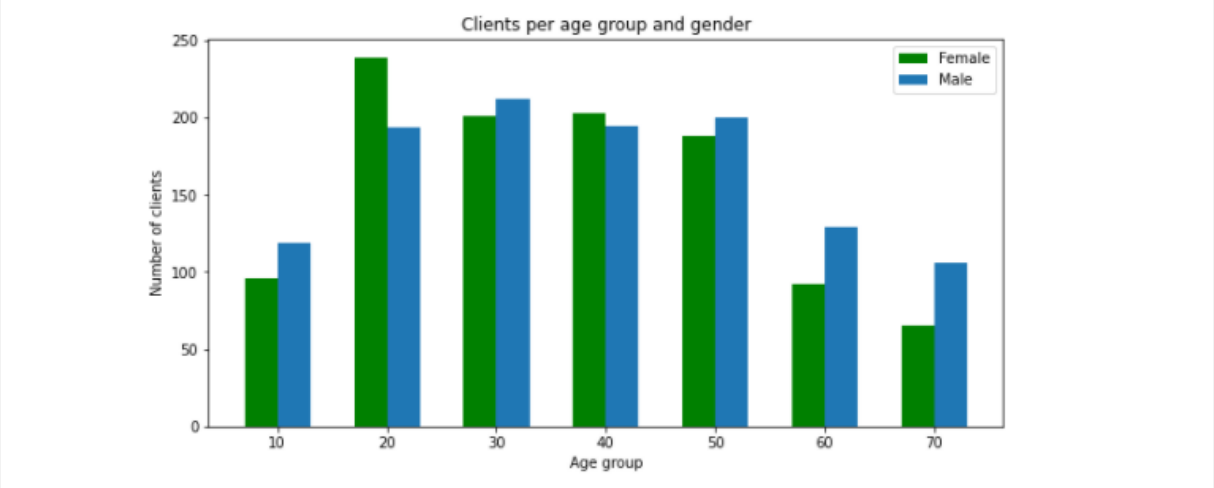
```
(2239, 77)
```

Data Visualization (Insights):

Age group of clients:



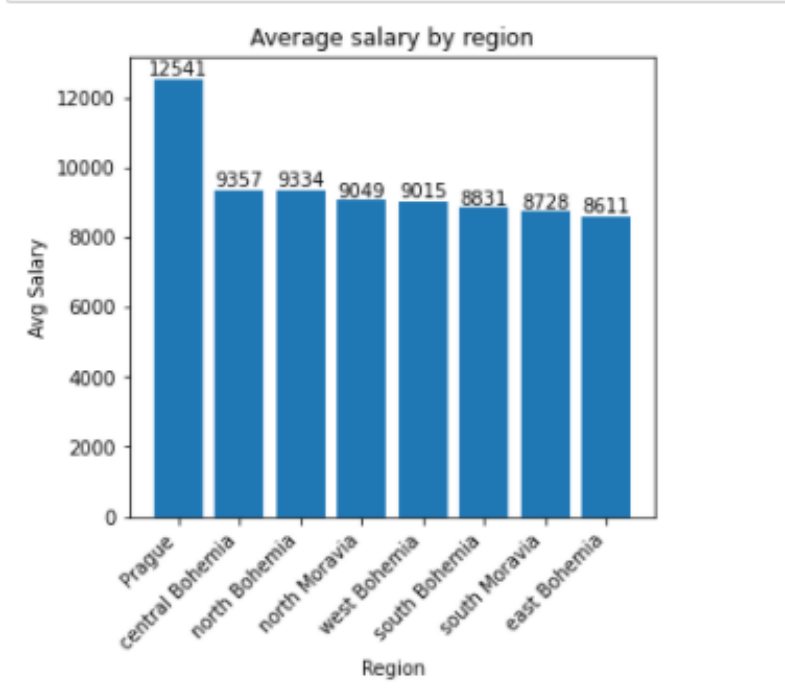
Number of clients per age group and gender:



As can be seen in the graph that the low number of clients in age group 10, 60 and 70. We can observe that the age groups from 20 to 50 contribute maximum to the bank's number of clients. As the age further increases, a downward trend is observed for age groups 60 and 70 and the minimum amount is reached with the age group of 70's, as expected.

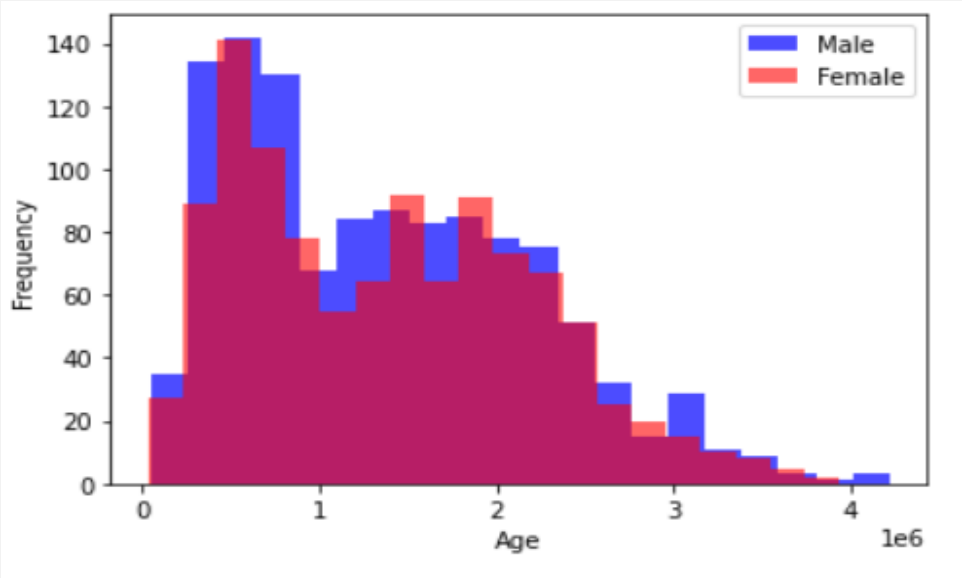
On the other hand, there is no obvious trend seen among the gender groups to see an inclination towards a specific gender. Across the age groups, the no. of female to male clients are almost comparable with the only exception for the 20's and 70's age group where the total no. of male clients is almost 133% of female clients.

Average Salary by region:



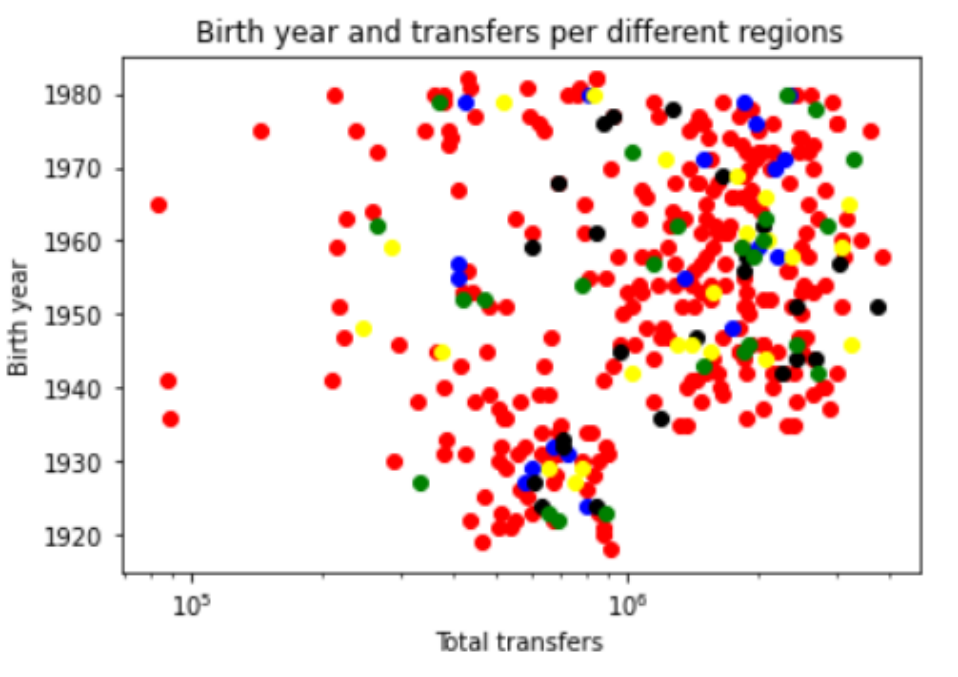
As we can easily distinguish from this bar graph that the 'Prague' region is offering much lucrative salaries as compared to any of the other regions, whose values are all very similar to each other. One may even say that the region has more well-established businesses as compared to others and that the companies are having a vision of rewarding people with good pay scales. Hence, clients applying for loans and credit cards from 'Prague' will have high potential of paying loans and card debts on time.

Histogram of total transfers per gender:



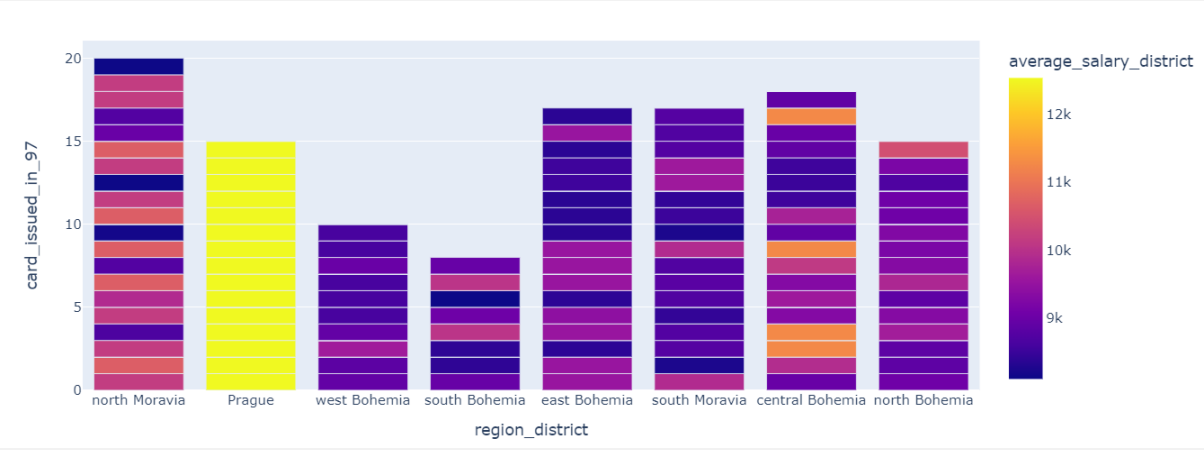
Based on this histogram, we can conclude that the number of transfers made by men is slightly higher than that of women. This, in turn, can tell us that it is quite possible that the male part of clients more often uses banking services, which in turn can give us some help in analysing and choosing target clients for certain companies, and so on.

Birth year and transfers per different regions:



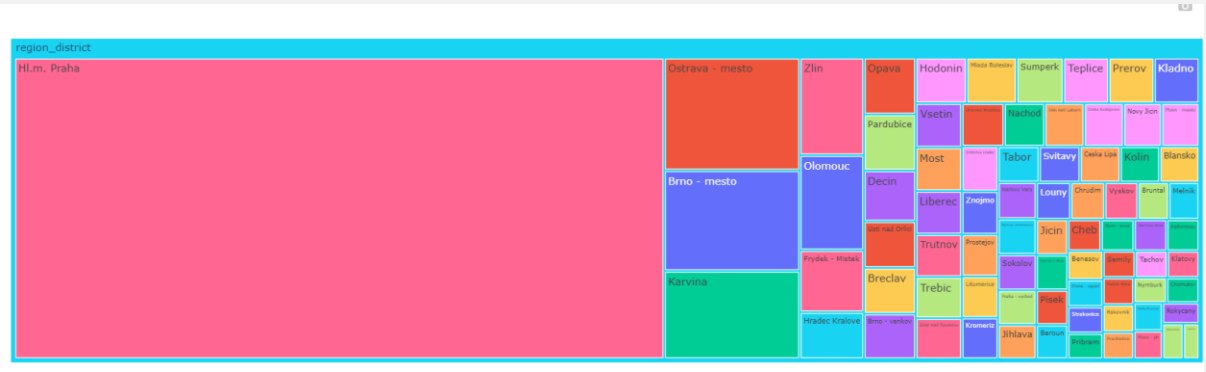
This graph gives us information on the correlation of the number of transfers with age in some regions. Thus, we can already draw a small conclusion related to the fact that the number of transfers exceeding a million occurs among clients who were born after about 1935. This conclusion may be related to the fact that older clients have fewer financial transfer needs, and this conclusion, in turn, can help us at some point in the analysis and calculation of certain companies.

Card issued in 97 by region and average salary:

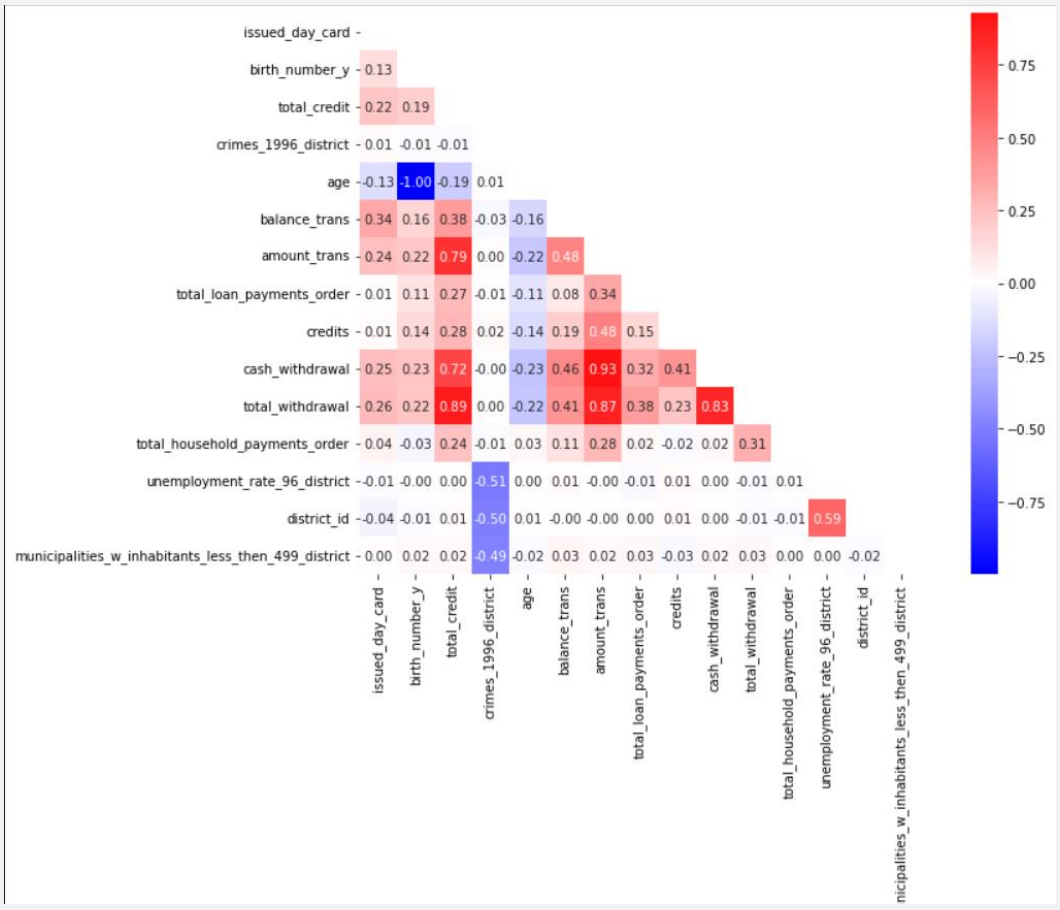


We can see that the population with high avg. salary is based Prague, however there are more cards issued in other regions, hence the bank should focus on these high earners and target them as potential clients.

Furthermore, based on the following chart, the largest customer base seems to be in Praha district, which also seems to have our largest customer base, but the population to customer ratio could be better here if we compare this district to other districts and their population.



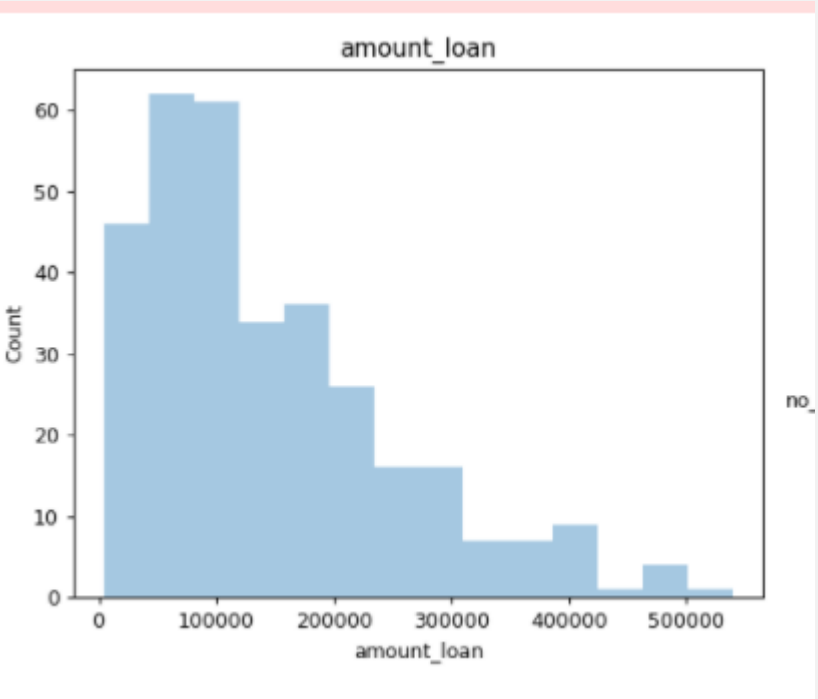
Below, we can check the corelation between top features as per random forest, XGBoost and Gradient Boosting models (the model was running on from randomly selected features)

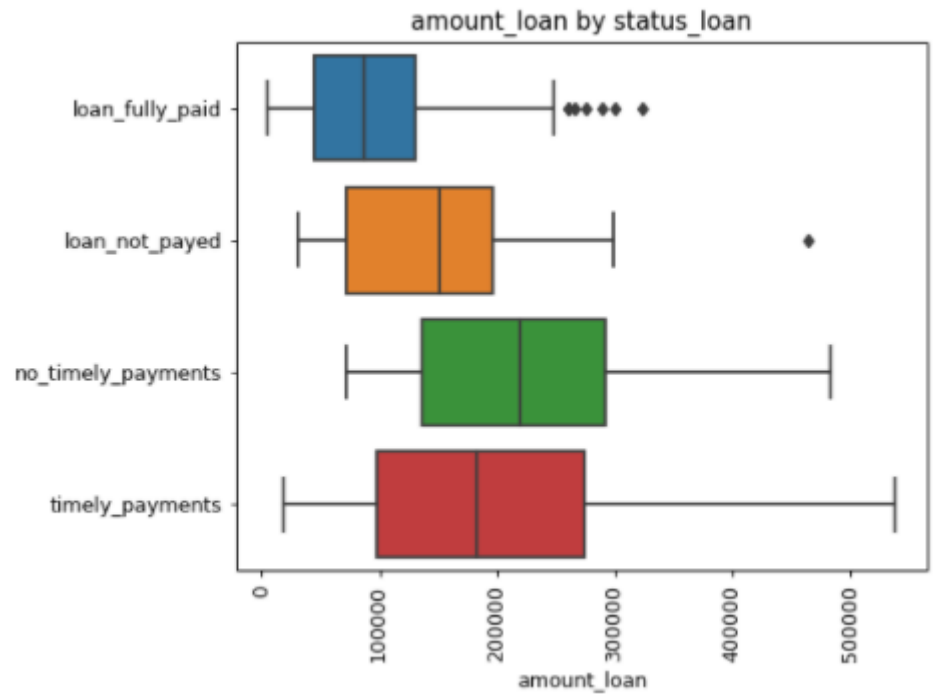


We got the following model results based on the 1st iteration of predictions; it may include some bias as the feature selection was not completely done as it was out of scope for this project.

	randomForest	boostedTree	XG
Accuracy	0.986607	0.977679	0.988839
AUC	0.996437	0.986540	0.991115

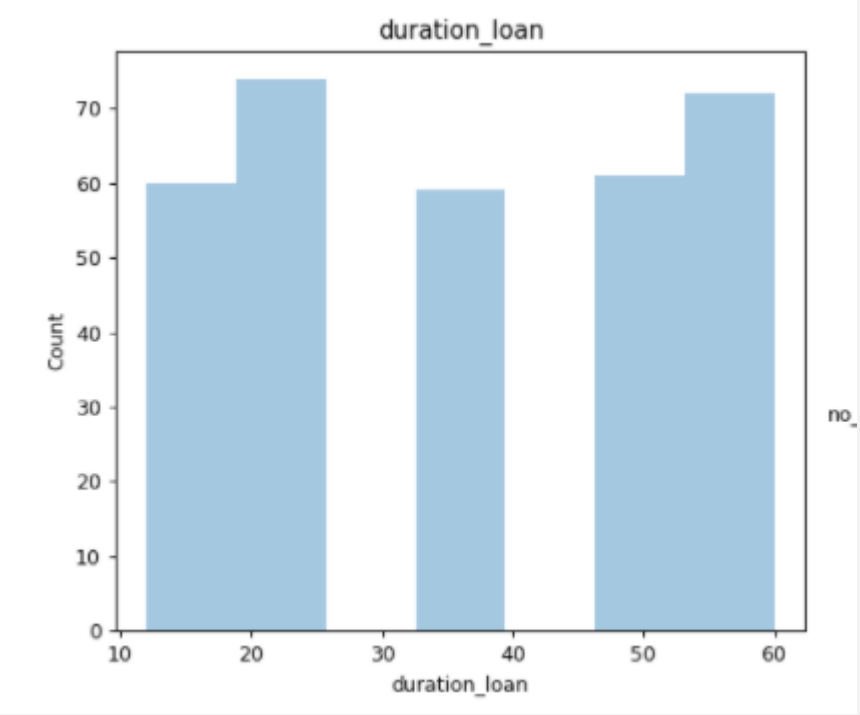
Loan Amount By Loan status:

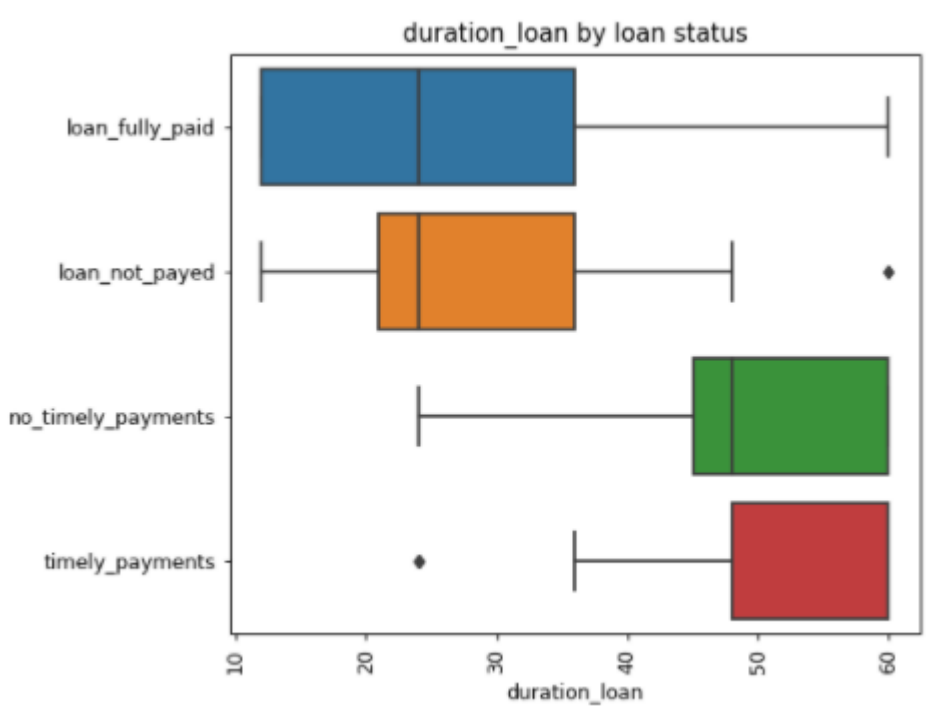




It looks like all loans are not unique. Loan amount of 100000 is the highest and appearing several times. And no timely payments of loan is higher with amount above 200000.

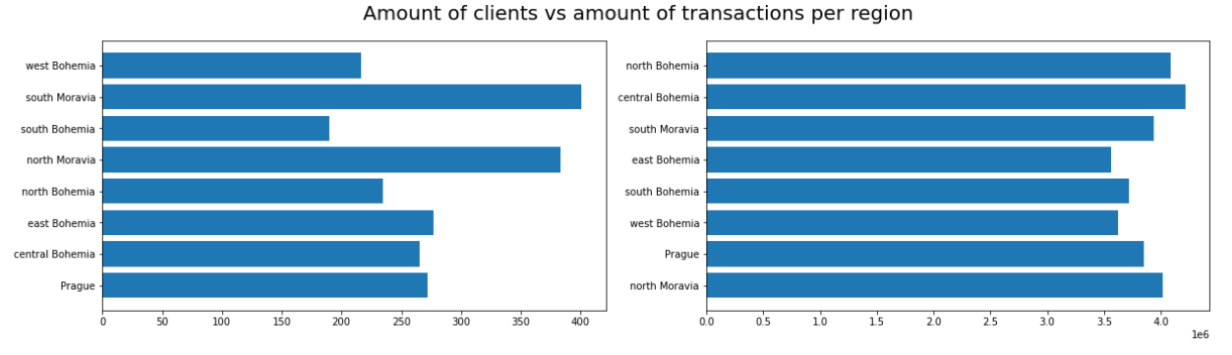
Duration of Loan-by-loan status:





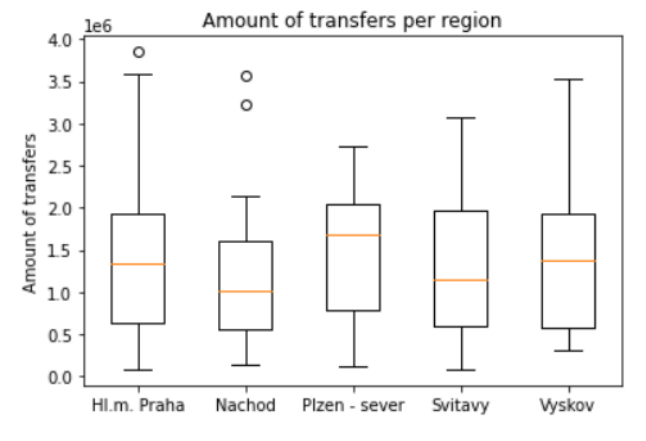
The count of number loan is higher for the duration of 15 to 25 months and 45 to 60 months. With the boxplot we can conclude that no timely payments are higher for the duration of 45 to 60 months whereas we have timely payments from 50 to 60 month.

Amount of client's vs number of transactions per region:



From this graph, you can see how in the same region, but the number of transactions does not always correlate with a different number of clients. This observation is important to keep in mind when working with these regions.

Boxplot of transactions per region:



This graph shows well how the net transfers are distributed in different regions.

References:

https://stackoverflow.com/questions/60869243/how-can-i-filter-dataframe-based-on-null-not-null-using-a-column-name-as-a-varia
https://stackoverflow.com/questions/17210646/python-subplot-within-a-loop-first-panel-appears-in-wrong-position/17211410

https://stackoverflow.com/questions/16096627/selecting-a-row-of-pandas-series-dataframe-by-integer-index

https://stackoverflow.com/questions/38541636/keep-other-variables-when-executing-get-dummies-in-pandas

https://numpy.org/doc/stable/user/whatisnumpy.html

https://pandas.pydata.org/docs/user_guide/index.html

https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

https://plotly.com/python/

https://plotly.com/python/line-and-scatter/

https://plotly.com/python/bar-charts/

https://plotly.com/python/line-and-scatter/

https://www.talend.com/resources/what-is-data-mart/

https://www.altexsoft.com/blog/what-is-data-mart/

https://www.pythoncheatsheet.org/

https://perso.limsi.fr/pointal/_media/python:cours:mementopython3-english.pdf

(https://iesegnet-my.sharepoint.com/personal/ayesha_noor_ieseg_fr/Documents/Microsoft

#Multiple YouTube videos related to manipulating data with pandas.