FINAL PROJECT REPORT — Python Programming for Data Science

Data Processing – Individual Project

Beatrice Maria Infurna

Academic Year: 2025/2026

Table of Contents

Introduction

The goal of this project was to create a complete basetable using the Czech bank dataset provided, and the aim of this report is to describe the basetable that was built.
To do this, we were given eight different tables in asc format, which had to be imported and processed in a Jupyter notebook, and then combined to construct the final basetable.

The Jupyter notebook is organized into the following main steps:

1. **Importing data**
2. **Converting data column and Basic formatting**
3. **Building the Base Relational Table**
4. **Timeline Filtering**
5. **Feature Engineering (Independent variables)- 1996,** further divided into:
    - Account table features
    - Transaction based features
    - K-symbol features
    - Operation type features
    - Balanced-based features
    - RFM features
    - Permanent order features
    - Previous loan/card history before 1997
    - Card features
    - Extra features
6. **Feature Engineering (Dependent variables)- 1997**
7. **Cleaning and export**
8. **Visualizations**

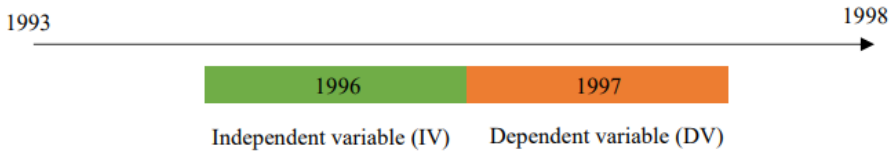
Basetable Construction

Basetable requirements

The basetable had the following criteria:

- **Granularity**: "*is client who is account owner. It was assumed that "all the accounts' activities are done only by the owner"*.
  Because of this, starting from section 3 of the Jupyter notebook, where the base relational table was built, the "OWNER" type from the disp table was selected.
  The disp table distinguishes between "DISPONENT" and "OWNER", and we only kept the owners. This resulted in the dataframe called "Owners", with about 4500 rows representing clients.
  It was not assumed that there was a one-to-one relationship between clients and accounts. To make sure that clients who own multiple accounts were still handled correctly, most variables were first calculated at the account level and later aggregated at the client_id level.
  This kept the granularity of the basetable at client-level while still considering all possible accounts a client may have.

- **Columns:** "*Independent and dependent variables following the timeline in Figure 1.*
  > The time window of the independent variables: 1996 (1 year).
  >The time window of the dependent variables: 1997 (1 year)."



  Because of this, in step 4 of the Jupyter notebook a Timeline Filtering section was created.
  Accounts used for calculating the independent variables were selected only if they were opened between 01/01/1993 and 31/12/1995.
  If an account was created in 1996, it was not considered, since we would not have enough data for the full independent variable window.
  All independent variables were then built using only data from 01/01/1996 to 31/12/1996.
  The dependent variables were created using only events that happened in the same way, but in 1997 instead of 1996.

It was specified that the independent variables should be *“calculated for clients who have sufficient information in the independent variables time window (i.e. 1996).”*.
To make sure of this, an additional filter was added to keep only accounts that had at least one transaction in 1996. If an account had no transactions, it would mean no incoming amounts, no outgoing amounts, no k-symbols, no operations, no balances, and no recency. In other words, there would be no observable behavior in the 1996 window.
After applying this activity filter, there were 3602 unique account IDs with at least one transaction in 1996.
When combining both filters, meaning the time filter and the activity filter, we ended up with **2239 active unique accounts** that were created before 1996 and that had observable behavior in 1996. These were the accounts used to build the independent variables.

Data Sources

The project uses eight tables from the Czech bank dataset.
These tables are client, account, disp, trans, loan, order, card, and district.
Each table contains different information about the bank’s customers and their financial activity.
After loading the data, all the tables were cleaned and then merged in order to build the final dataset that was used to create the basetable.

Data Cleaning and Transformations

The tables in the dataset had different formats, structures and types of information.
All tables were first loaded and formatted so that they could be used together .
Any column containing dates was converted into a proper datetime format, since this was necessary for timeline filtering and feature creation.

One of the most important transformations was the processing of the birth number in the client table.
The birth number encodes both the client’s birthdate and their gender.
To extract the real birthdate, the number had to be separated into year, month, and day, and the month needed to have 50 subtracted from it for female clients.
Once the true birthdate was reconstructed, the age of each client at the end of 1996 was calculated.

Moreover, it is important to notice that the card data table required a supplementary step because of the format “YYMMDD hh:mm:ss”., so we firstly divided the string into 2 part,

then dropped the “hh:mm:ss”, so that it could be treated like the other tables following the format “YYMMDD”.

As we mentioned already, the disp table was used to connect clients to their accounts.
Only clients marked as “OWNER” were kept, since the basetable must be built at the client level under the assumption that all account activity belongs to the owner.

The account table includes the account creation date, which was essential for timeline filtering.
Only accounts created on or before 31 December 1995 were kept, as was explained previously.

The transaction table was used to extract all transaction behavior during the year 1996.
Values such as the number of transactions, total incoming and outgoing amounts, and the minimum, maximum, and average transaction values were computed.
The ending balance and the recency of the last transaction were also created from this table.

Permanent orders, loans, cards and district tables were all cleaned as well and merged through client and account IDs.
Permanent order information was used to generate features related to repeated payments.
The loan and card tables were used mainly to create the dependent variables for 1997 and also to build the client’s history before 1997.
The district table provides regional characteristics, but in this project it did not significantly affect the feature set.

Toward the end of the notebook, a final cleaning step was also performed.
This included removing missing values, dropping duplicate or unnecessary columns, and renaming columns when needed to keep the final basetable clean and consistent.

Variables of the Basetable

General Overview

The final basetable contains one row per client and includes all the independent and dependent variables that were created throughout the notebook.

Below is a description of the different types of variables included in the basetable.

Demographics

The basetable starts with the client identifier, which is the column "**client_id".** This column is used only for identification and is not meant to be used as a feature for modeling.

The demographic variables include **"gender"** and **"age_1996".** These were obtained by transforming the birth number, which allowed us to calculate the client's age at the end of 1996 and to identify whether the client is male or female.

We encoded gender to make sure it was numeric, with female=1 and male= 2

Account-related Features

Account-related variables include **"account_age_days",** which shows how many days the client's account was open for by the end of 1996. There are also the columns "**freq_POPLATEK PO OBRATU"** and **"freq_POPLATEK TYDNE",** which indicate the type of account fee frequency associated with the client's account.

Transaction Features

Multiple independent variables come from the transaction table, all based on data from 1996.
These include:

- **"txn_count_1996",** that is the total number of transactions in that year,
- "**incoming_total_1996"** and "**outgoing_total_1996",** which show total incoming and outgoing amounts.
- **"avg_txn_1996",** "**min_txn_1996"** and "**max_txn_1996"** which describe the characteristics of the client's transactions.
- **"ending_balance_1996"** , that represents the last recorded account balance in 1996,
- **"recency_1996"** that shows how many days passed between the end of 1996 and the client's most recent transaction.

K-symbol Features

There is another group of variables related to the K-symbol information.
These include:

- **“k_pojistne_count_1996”,**
- **“k_sipo_count_1996”,**
- **“k_sluzby_count_1996”,**
- **“k_urok_count_1996”,**
- **“k_sankc_urok_count_1996”,**
- **“k_duchod_count_1996”,**
- **“k_uver_count_1996**”.
  Each of these represents the number of transactions with a specific category of purpose.
- **“k_symbol_diversity_1996**” measures how many different K-symbol types the client used during 1996.

Operation Type Features

The operation type variables describe the types of banking actions performed by the client.
These include:

- **“op_vyber_kartou_count_1996”,**
- **“op_vklad_count_1996”,**
- **“op_prevod_z_uctu_count_1996”,**
- **“op_vyber_count_1996”**
- **“op_prevod_na_ucet_count_1996”.**

All mainly describe the count for each type of operation.

The variable **“op_type_diversity_1996**” represents how many different operation types the client performed in that year.

Permanent Order Features

Permanent order variables include:

- “**order_count”,** which represents how many permanent orders the client had in 1996**,**
- **“order_amount_total”,** which is the total amount involved in these orders,
- Several variables such as “**po_pojistne_count”** and **“po_sipo_count”** that show the number of permanent orders of each type.
- The variable **“po_type_diversity_1996**” summarizes how diverse these orders were.

Loan & Card History

There are also variables related to the client's loan and card history before 1997.
These include:

- **"has_previous_loan",**
- **"has_card_pre_1997**",
- **"loan_granted_1996",**
- **"card_issued_1996"**
- **"previous_loan_amount_total",**
- **"previous_loan_problem",**
- **"card_count_pre_1997".**

These variables show whether the client had previous loans or cards, how many they had, and whether there were any issues with the loan.

The first four were created using dummy variables, where 1 is yes, 0 is no.

Aggregated Ratios and extra

In addition to these, the basetable includes ratio-based features such as:

- **"ratio_credit_to_debit_1996",**
- **"ratio_incoming_total_1996",**
- **"ratio_outgoing_total_1996",**
- **"avg_txn_norm_1996".**

These help describe the client's financial behaviour more proportionally rather than using only absolute values.

Finally, the variable "**n_accounts"** indicates how many accounts the client owned.

Dependent Variables

Two dependent variables were created for the basetable: "**dv_loan_1997"** and **"dv_card_1997".**

Both variables use information from the year 1997, and the first dependent variable shows whether the client received a loan in 1997. If the client had any loan issued during that year, the value is 1, otherwise it is 0.

The second dependent variable does the same, but is related to weather the client received a card in 1997

Both dependent variables are binary and follow the time rules required by the project.

Basetable Summary

The final basetable is the result of all the cleaning, merging and feature engineering steps. It contains one row per client, and all the independent variables are based on data from 1996 only. The dependent variables come from 1997.

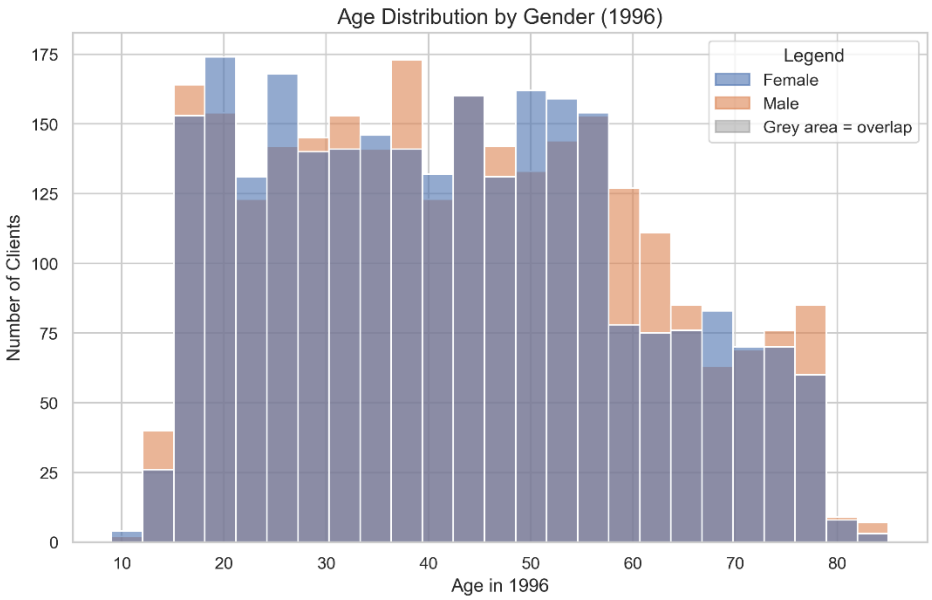
At the end of the merging process, the final basetable contains 2239 clients and 60 columns. This includes the client ID, all independent variables and both dependent variables.

After removing the raw columns that were only used during preparation, such as the birth number and the separate date components, dropping duplicates and renaming, the number went down to 53, for a total of 52 active features (we do not include client id), divided into 50 independent and 2 dependent variables.

All values in the basetable are numeric, and there are no missing values. The dataset is clean, consistent and ready to be used for modeling.

Descriptive Analysis

Age Distribution by Gender

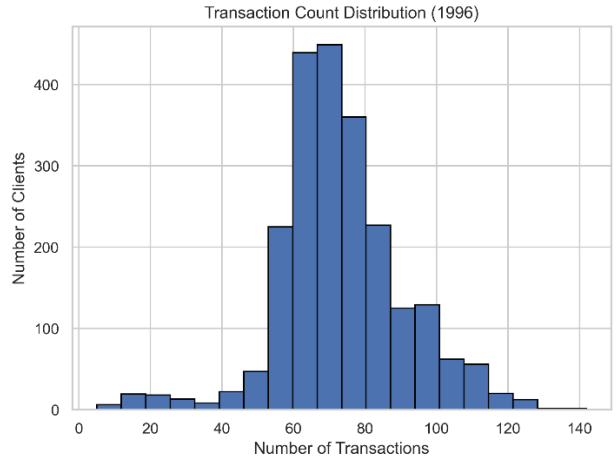


The age distribution of the clients in 1996 shows most clients are between 20 and 60 years old.

The distribution seems balanced between men and women.
The grey area in the graph represents the overlap between the two genders, meaning both male and female clients fall into those age bins in similar numbers.

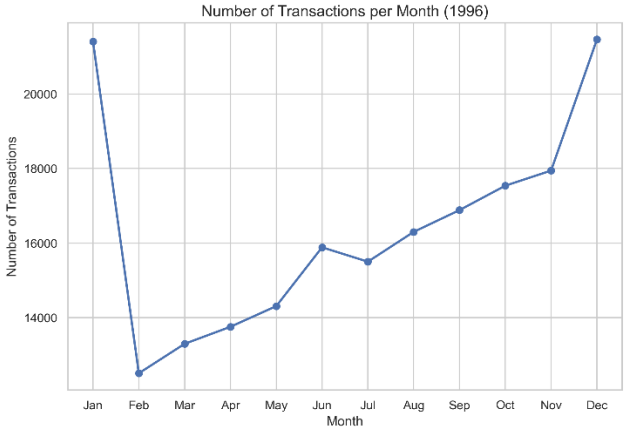
There is a slight decrease in the number of older clients, especially after the age of 60. The youngest clients in the dataset are around 10 years old, while the oldest ones are above 80, but most clients fall within typical working-age groups.
Overall, the age distribution does not show any extreme skewness, and both genders follow a very similar pattern.

Transaction Behavior



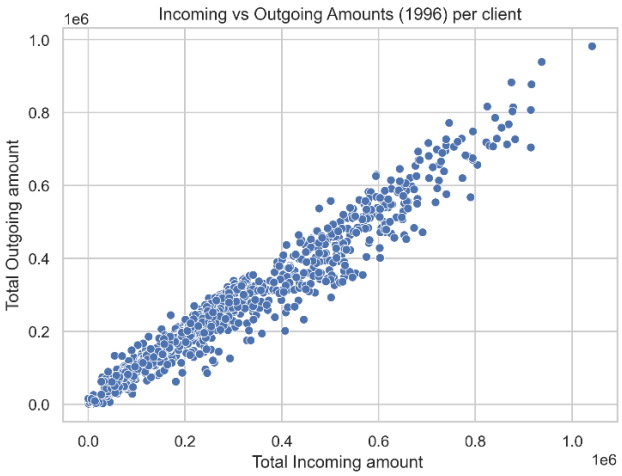
Transaction count distribution

Most clients have between 40 and 100 transactions during 1996. The distribution is spread out, showing that clients have very different levels of activity.
The majority of clients fall in the middle range, while only a small number show extremely low or extremely high transaction counts.
There is no sign of abnormal spikes or gaps, meaning the transaction count variable behaves as expected.



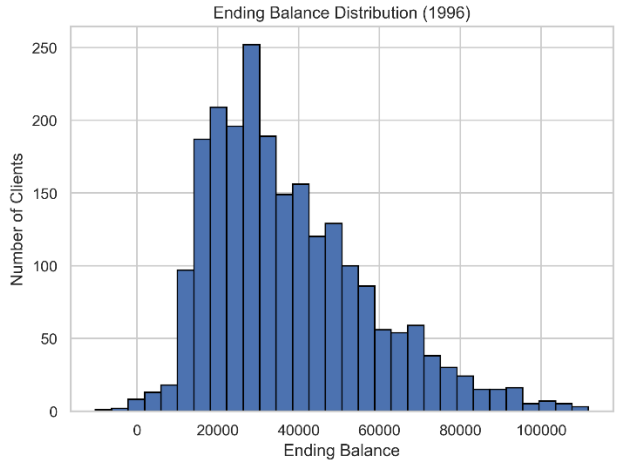
Monthly Transaction Activity (1996)

The line chart shows how many transactions were performed each month in 1996. The pattern seems stable throughout the year, with slightly higher activity at the beginning and at the end of the year, and a small dip during the summer months. This is a common seasonal trend in banking data, since many payments and transfers tend to be made around the start and the end of the year.



Incoming vs Outgoing Totals

The scatterplot compares the total incoming amount with the total outgoing amount for each client in 1996. Each point on the chart represents one client.
The axes show the total amount of money received and spent during the year.
Because financial amounts vary a lot from client to client, a log scale was used in the graph to make the pattern easier to see.
The points form a clear upward trend, showing a normal pattern in banking data, which means that clients who receive more money also tend to spend more. The plot also shows that most clients have moderate financial activity, while only a few have extremely high income or spending.

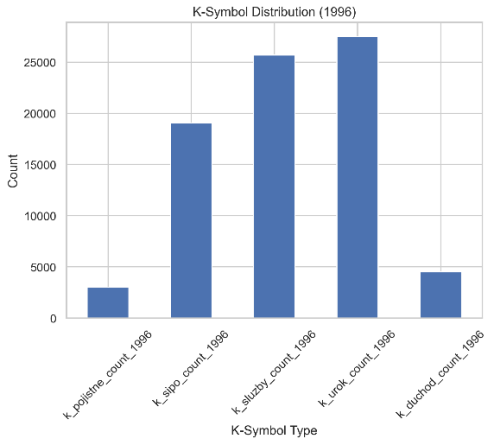


Ending Balance Distribution (1996)

The histogram shows how clients' ending balances were distributed at the end of 1996. Most clients have positive balances, with a large group around the lower and middle

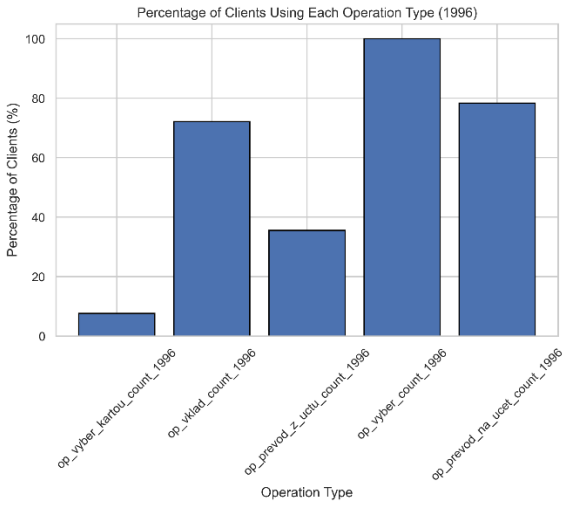
ranges. Only a few clients have very high ending balances.
There are also some clients with negative balances, but this is not unusual in banking datasets. Overall, the distribution looks reasonable and does not show any unusual patterns.

Other relevant distribution



K-symbol Distribution

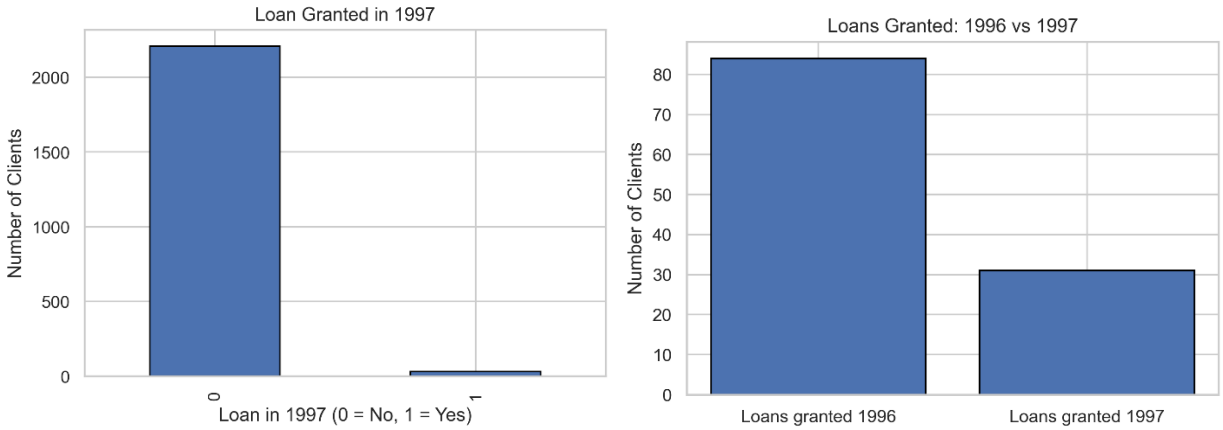
The bar chart shows the total number of K-symbol operations performed by all clients during 1996.
Some K-symbol types are much more common than others. For example, household payments (SIPO) and insurance payments appear very frequently, while categories such as interest or pension transfers are less common. This is predictable because most clients make everyday payment such as household and insurance payments, while only a portion receive pensions or interest payments. The distribution confirms that the K-symbol variables behave in a normal and realistic way for banking data.



Operation Type Distribution

The chart shows the percentage of clients who used each operation type at least once in 1996. Incoming and outgoing transfers are the most common operations, which means most clients regularly move money between accounts.
Cash withdrawals and card withdrawals are also widely used, while deposits are much less common. Looking at the percentage of clients instead of total counts gives a clearer idea of which operations are used by most people in the dataset

Dependent variables

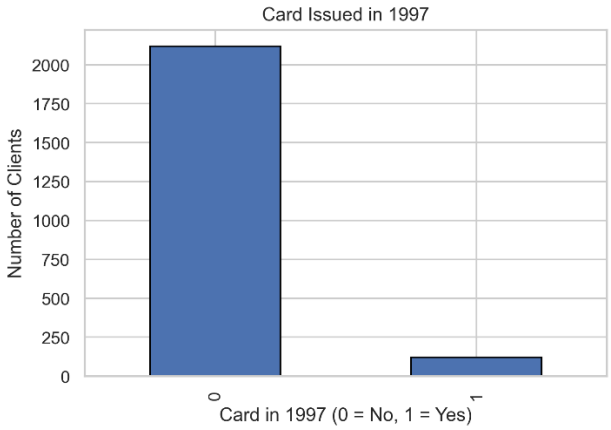


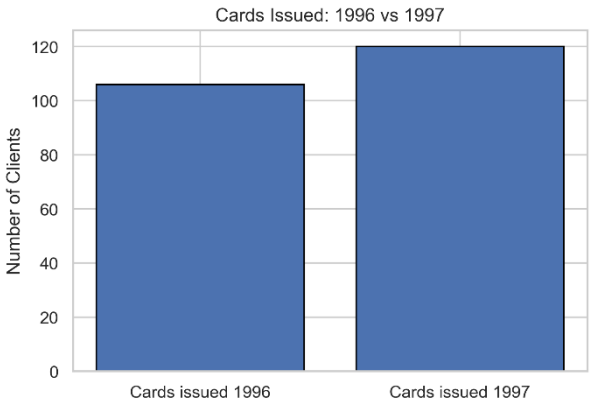
Loan Dependent Variable Distribution

The bar chart shows that only a very small percentage of clients received a loan in 1997. Almost all clients fall into the “0” category, which means no loan was granted, while only a few clients fall into the “1” category. This confirms that the loan dependent variable is

highly imbalanced. This type of imbalance is normal in banking data, since only a small fraction of clients apply for and receive a loan each year.

Looking at the comparison from the previous year, we can see that there was a drop of about 62,5% from 1996 to 1997, and I think this could be partially explained by our database design, which does not includes clients who opened their accounts during 1996 , which reduces the total number of observed loan events in 1997.



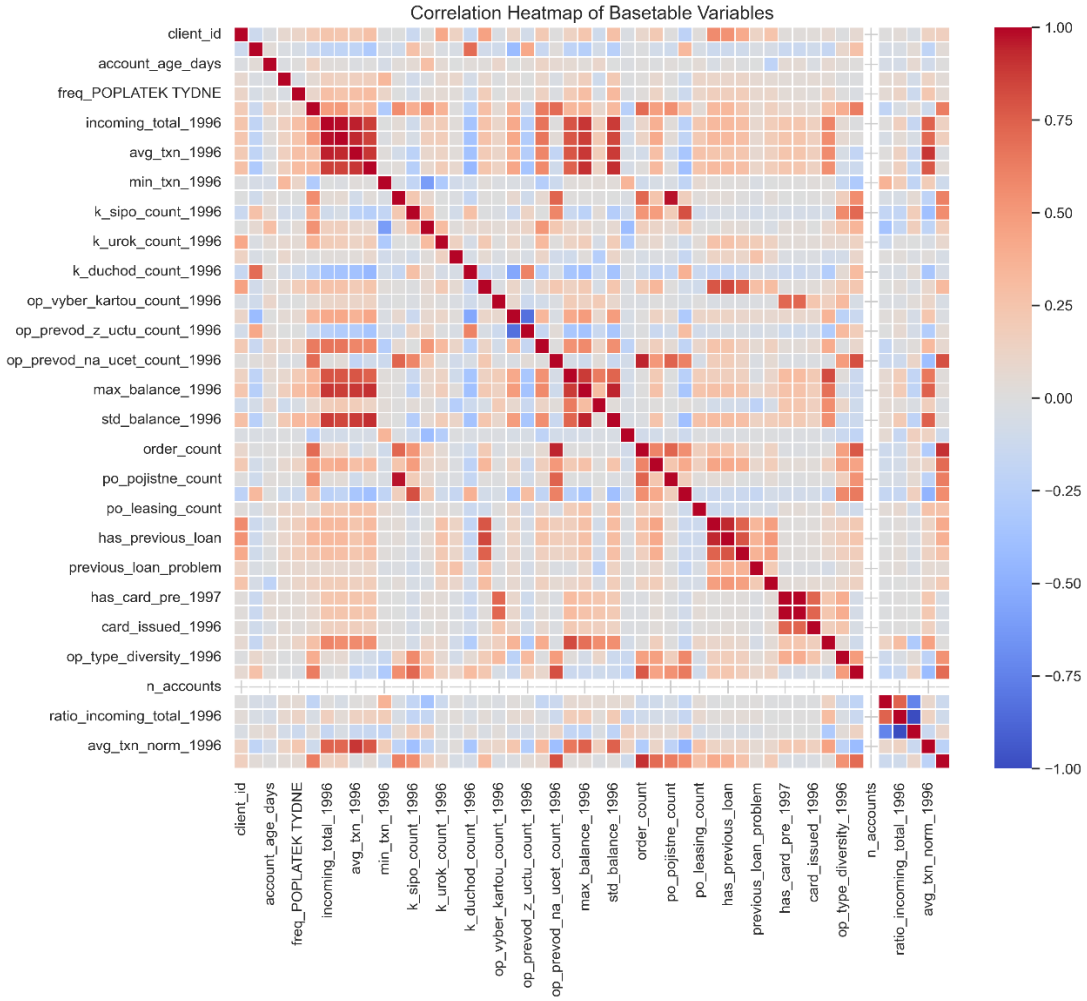


Card Dependent Variable Distribution

After correctly formatting the card issuance dates, the dependent variable dv_card_1997 shows that only a small number of clients received a card in 1997. The majority of clients belong to the “0” category, meaning no new card was issued, while a smaller group belongs to the “1” category. This creates an imbalanced dependent variable, which is normal for this dataset because only a minority of clients typically receive a new card in a given year.

We also include a comparison with the independent variable card_issued_1996, which shows that the number of cards issued increased slightly in 1997. This difference is small and seems reasonable given the size of the client base and the fact that card issuance is a relatively rare event.

Correlations Heatmap



The heatmap shows the correlations between all numeric variables in the basetable. Most variables have low pairwise correlations, which is expected because the dataset contains many different types of information such as demographics, transactions, K-symbols, balances and operations. Some related features, such as the different transaction statistics, show moderate positive correlations with each other. Age does not appear to be strongly correlated with most behavioral features. The dependent variables show only weak correlations with the independent variables, which is normal for financial datasets and confirms that predicting loans or card issuance is not a trivial task.

Conclusion

The project successfully produced a complete basetable for the predictive modeling task using the Czech bank dataset. All tables were loaded, cleaned and merged, and the correct timeline filters were applied so that the independent variables were based only on the full year of 1996, while the dependent variables were taken from 1997. The basetable contains one row per client and includes all the engineered features across demographics, transactions, K-symbols, operations, balances and permanent orders.

During the construction of the basetable, additional processing was required to correctly format several date fields. After all transformations, the final basetable contains 2239 clients and 50 active features, along with two dependent variables.

The descriptive analysis shows that most clients have moderate financial activity, and only a small portion received a loan or a new card in 1997, which creates imbalanced dependent variables. The correlations between features are mostly low, which is expected because the variables capture different types of information.