Bank Loan Case Study

Final Project-2

Project Description

The financial industry is constantly evolving and facing new challenges in the digital era. One of the major challenges is to assess the **creditworthiness** of customers who apply for various types of loans. Creditworthiness is the ability of a customer to repay a loan based on their income, assets, liabilities, and credit history. However, not all customers have a sufficient credit history to evaluate their creditworthiness. Some customers may take advantage of this and default on their loans, causing financial losses for the lenders.

To address this challenge, I performed **exploratory data analysis** (EDA) on a dataset of bank loan applications. EDA is a process of examining, summarizing, and visualizing data to gain insights and understanding. The dataset contains information about customers who applied for different types of loans. The dataset also includes information about their personal and financial attributes. The dataset also indicates whether the customers had payment difficulties or not.

The **objective** of this project was to analyze patterns in the data and identify factors that influence the likelihood of loan default. Loan default is the failure to repay a loan according to the agreed terms. The factors that influence loan default can be divided into two categories: **consumer attributes** and **loan attributes**. By understanding how these factors affect loan defaults, the company can improve its decision-making process and reduce its risks.

Approach

The approach I followed for this project consisted of the following steps:

- **Identify Missing Data and Deal with it Appropriately:** I used Excel functions like ***COUNT, ISBLANK,*** and ***IF*** to identify missing data in the dataset. Missing data can affect the accuracy and validity of the analysis, so it is essential to handle it effectively. I decided to impute the missing values using the median of each variable, as it is less sensitive to outliers than the mean. I created a bar chart to visualize the proportion of missing values for each variable.
- **Identify Outliers in the Dataset:** I used Excel functions like ***QUARTILE, IQR***, and conditional formatting to detect and identify outliers in the dataset. Outliers are extreme values that deviate significantly from the rest of the data. They can have a large impact on the analysis and distort the results, so it is important to identify them and investigate their causes. I focused on numerical variables such as income, loan amount, and interest rate. I created scatter plots to visualize the distribution of these variables and highlight the outliers. I applied thresholds to determine if the outliers were valid data points or required further investigation.

- **Analyze Data Imbalance:** I used Excel functions like ***COUNTIF*** and ***SUM*** to determine if there was data imbalance in the dataset. Data imbalance occurs when one class of the target variable has a much higher frequency than the other class. This can affect the accuracy of predictive models that aim to classify customers into default or non-default groups. I calculated the ratio of data imbalance by comparing the class frequencies of the target variable, which indicates whether a customer had payment difficulties or not. I created a bar chart to visualize the distribution of the target variable and highlight the class imbalance.
- **Perform Univariate, Segmented Univariate, and Bivariate Analysis:** I used Excel functions like ***COUNT, AVERAGE, MEDIAN***, and ***statistical functions*** to perform univariate analysis on individual variables. **Univariate** analysis helps to understand the distribution, central tendency, and variability of each variable. I used Excel features like ***filters, sorting,*** and ***pivot tables*** to perform segmented and bivariate analysis on different scenarios, such as customers with payment difficulties and all other cases. **Segmented** analysis helps to compare variable distributions for different groups or categories. **Bivariate** analysis helps to explore relationships between variables and the target variable using frequency distribution. I created bar charts, box plots, stacked bar charts, grouped bar charts, scatter plots to visualize the distributions and relationships of variables.
- **Identify Top Correlations for Different Scenarios:** I used Excel functions like ***CORREL*** to calculate correlation coefficients between variables and the target variable within each segment. Correlation coefficients measure the strength and direction of linear relationships between two variables. They range from ***-1*** to ***1***, where ***-1*** indicates a perfect negative relationship, ***0*** indicates no relationship, and ***1*** indicates a perfect positive relationship. I ranked the correlations to identify the top indicators of loan default for each scenario. I created correlation matrices to visualize the correlations between variables within each segment. I highlighted the top correlated variables for each scenario using different colors.

Tech-Stack Used

The software and its version that I used for this project was **Microsoft Excel 365**. Microsoft Excel is a spreadsheet software that allows users to perform various data analysis and visualization tasks using its built-in functions and features. The purpose of using Microsoft Excel for this project was to perform ***data cleaning, manipulation,*** and ***visualization*** using its built-in functions and features. Some of the functions and features that I used are:

- **COUNT, ISBLANK, IF:** These functions help to identify missing data in the dataset and count the number of non-empty cells.
- **AVERAGE, MEDIAN:** These functions help to calculate the mean and median of a range of cells, which can be used for imputation or descriptive analysis.
- **QUARTILE, IQR:** These functions help to calculate the quartiles and interquartile range (IQR) of a range of cells, which can be used for identifying outliers or creating box plots.
- **Conditional Formatting:** This feature helps to apply formatting rules to cells based on certain conditions, such as highlighting outliers or coloring cells by value.

- **COUNTIF, SUM:** These functions help to calculate the frequency or sum of a range of cells that meet a given criterion, which can be used for analyzing data imbalance.
- **CORREL:** This function helps to calculate the correlation coefficient between two ranges of cells, which can be used for bivariate analysis or identifying top correlations.
- **Filters, Sorting, Pivot Tables:** These features help to filter, sort, or summarize data based on different criteria or categories, which can be used for segmented or bivariate analysis.
- **Charts:** This feature helps to create various types of charts or graphs to visualize the data, such as histograms, bar charts, box plots, pie charts, scatter plots, heatmaps, etc.

Insights

Here are some of the insights and knowledge I gained while working on this project:

1. Identify Missing Data and Deal with it Appropriately:

One of the first steps in data analysis is to **identify and deal with missing data**. Missing data can affect the accuracy and validity of the analysis, as it can introduce bias, reduce statistical power, or distort the results. Therefore, it is essential to handle missing data effectively and appropriately.

To identify missing data in the bank loan application dataset, I used Excel functions like ***COUNT, ISBLANK,*** and ***IF***. These functions help to count the number of non-empty cells, check if a cell is blank, and return a value based on a condition. I applied these functions to each column of the dataset and calculated the number of missing values for each variable.

The table below summarizes the variables with missing data.

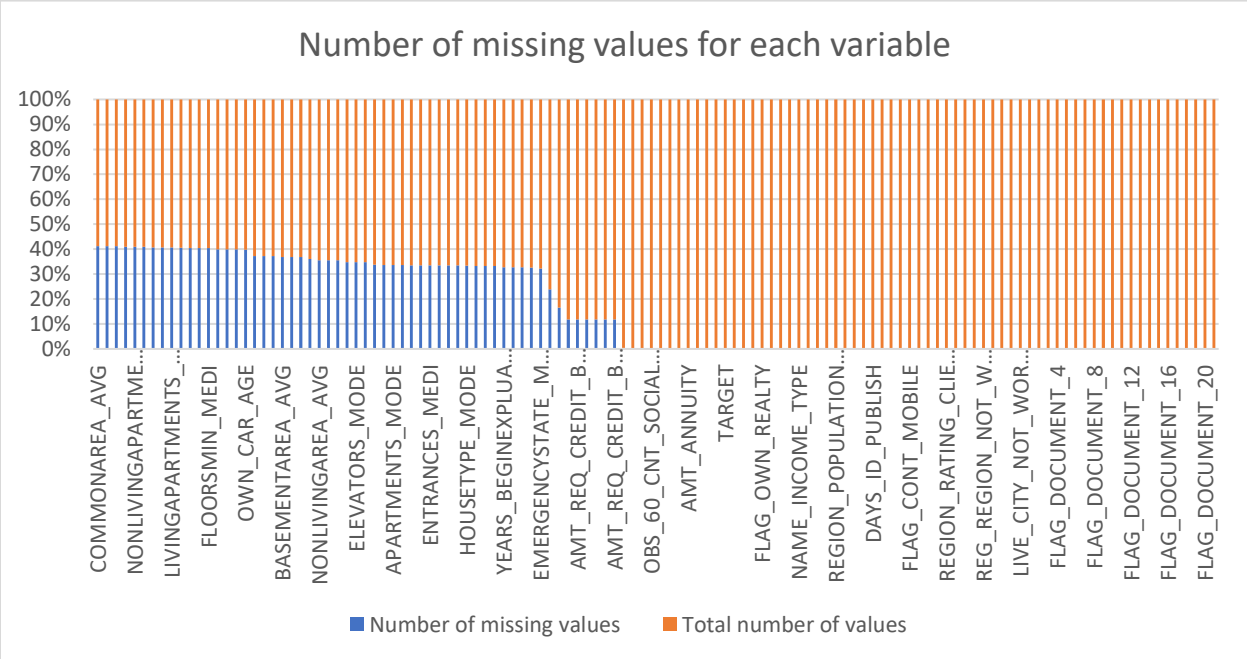
Variable	Number of missing values	Total number of values
COMMONAREA_AVG	34960	49999
COMMONAREA_MODE	34960	49999
COMMONAREA_MEDI	34960	49999
NONLIVINGAPARTMENTS_AVG	34714	49999
NONLIVINGAPARTMENTS_MODE	34714	49999
NONLIVINGAPARTMENTS_MEDI	34714	49999
LIVINGAPARTMENTS_AVG	34226	49999
LIVINGAPARTMENTS_MODE	34226	49999
LIVINGAPARTMENTS_MEDI	34226	49999
FONDKAPREMONT_MODE	34191	49999
FLOORSMIN_AVG	33894	49999
FLOORSMIN_MODE	33894	49999
FLOORSMIN_MEDI	33894	49999
YEARS_BUILD_AVG	33239	49999

YEARS_BUILD_MODE	33239	49999
YEARS_BUILD_MEDI	33239	49999
OWN_CAR_AGE	32950	49999
LANDAREA_AVG	29721	49999
LANDAREA_MODE	29721	49999
LANDAREA_MEDI	29721	49999
BASEMENTAREA_AVG	29199	49999
BASEMENTAREA_MODE	29199	49999
BASEMENTAREA_MEDI	29199	49999
EXT_SOURCE_1	28172	49999
NONLIVINGAREA_AVG	27572	49999
NONLIVINGAREA_MODE	27572	49999
NONLIVINGAREA_MEDI	27572	49999
ELEVATORS_AVG	26651	49999
ELEVATORS_MODE	26651	49999
ELEVATORS_MEDI	26651	49999
WALLSMATERIAL_MODE	25459	49999
APARTMENTS_AVG	25385	49999
APARTMENTS_MODE	25385	49999
APARTMENTS_MEDI	25385	49999
ENTRANCES_AVG	25195	49999
ENTRANCES_MODE	25195	49999
ENTRANCES_MEDI	25195	49999
LIVINGAREA_AVG	25137	49999
LIVINGAREA_MODE	25137	49999
LIVINGAREA_MEDI	25137	49999
HOUSETYPE_MODE	25075	49999
FLOORSMAX_AVG	24875	49999
FLOORSMAX_MODE	24875	49999
FLOORSMAX_MEDI	24875	49999
YEARS_BEGINEXPLUATATION_AVG	24394	49999
YEARS_BEGINEXPLUATATION_MODE	24394	49999
YEARS_BEGINEXPLUATATION_MEDI	24394	49999
TOTALAREA_MODE	24148	49999
EMERGENCYSTATE_MODE	23698	49999
OCCUPATION_TYPE	15654	49999
EXT_SOURCE_3	9944	49999
AMT_REQ_CREDIT_BUREAU_HOUR	6734	49999
AMT_REQ_CREDIT_BUREAU_DAY	6734	49999
AMT_REQ_CREDIT_BUREAU_WEEK	6734	49999
AMT_REQ_CREDIT_BUREAU_MON	6734	49999
AMT_REQ_CREDIT_BUREAU_QRT	6734	49999
AMT_REQ_CREDIT_BUREAU_YEAR	6734	49999

NAME_TYPE_SUITE	192	49999
OBS_30_CNT_SOCIAL_CIRCLE	168	49999
DEF_30_CNT_SOCIAL_CIRCLE	168	49999
OBS_60_CNT_SOCIAL_CIRCLE	168	49999
DEF_60_CNT_SOCIAL_CIRCLE	168	49999
EXT_SOURCE_2	126	49999
AMT_GOODS_PRICE	38	49999
AMT_ANNUITY	1	49999
CNT_FAM_MEMBERS	1	49999
DAYS_LAST_PHONE_CHANGE	1	49999
SK_ID_CURR	0	49999
TARGET	0	49999
NAME_CONTRACT_TYPE	0	49999
CODE_GENDER	0	49999
FLAG_OWN_CAR	0	49999
FLAG_OWN_REALTY	0	49999
CNT_CHILDREN	0	49999
AMT_INCOME_TOTAL	0	49999
AMT_CREDIT	0	49999
NAME_INCOME_TYPE	0	49999
NAME_EDUCATION_TYPE	0	49999
NAME_FAMILY_STATUS	0	49999
NAME_HOUSING_TYPE	0	49999
REGION_POPULATION_RELATIVE	0	49999
DAYS_BIRTH	0	49999
DAYS_EMPLOYED	0	49999
DAYS_REGISTRATION	0	49999
DAYS_ID_PUBLISH	0	49999
FLAG_MOBIL	0	49999
FLAG_EMP_PHONE	0	49999
FLAG_WORK_PHONE	0	49999
FLAG_CONT_MOBILE	0	49999
FLAG_PHONE	0	49999
FLAG_EMAIL	0	49999
REGION_RATING_CLIENT	0	49999
REGION_RATING_CLIENT_W_CITY	0	49999
WEEKDAY_APPR_PROCESS_START	0	49999
HOUR_APPR_PROCESS_START	0	49999
REG_REGION_NOT_LIVE_REGION	0	49999
REG_REGION_NOT_WORK_REGION	0	49999
LIVE_REGION_NOT_WORK_REGION	0	49999
REG_CITY_NOT_LIVE_CITY	0	49999
REG_CITY_NOT_WORK_CITY	0	49999

LIVE_CITY_NOT_WORK_CITY	0	49999
ORGANIZATION_TYPE	0	49999
FLAG_DOCUMENT_2	0	49999
FLAG_DOCUMENT_3	0	49999
FLAG_DOCUMENT_4	0	49999
FLAG_DOCUMENT_5	0	49999
FLAG_DOCUMENT_6	0	49999
FLAG_DOCUMENT_7	0	49999
FLAG_DOCUMENT_8	0	49999
FLAG_DOCUMENT_9	0	49999
FLAG_DOCUMENT_10	0	49999
FLAG_DOCUMENT_11	0	49999
FLAG_DOCUMENT_12	0	49999
FLAG_DOCUMENT_13	0	49999
FLAG_DOCUMENT_14	0	49999
FLAG_DOCUMENT_15	0	49999
FLAG_DOCUMENT_16	0	49999
FLAG_DOCUMENT_17	0	49999
FLAG_DOCUMENT_18	0	49999
FLAG_DOCUMENT_19	0	49999
FLAG_DOCUMENT_20	0	49999
FLAG_DOCUMENT_21	0	49999

To visualize the proportion of missing values for each variable, I created a ***column chart*** using Excel's chart feature. The column chart below shows the proportion of missing values for each variable before imputation.



To deal with missing data in the bank loan application dataset, I decided to impute the missing values using the **median** of each variable. The median is the middle value of a sorted list of values, which divides the list into two equal halves. The median is ***less sensitive*** to outliers than the mean, which is the average value of a list of values.

To impute the missing values using the median, I used Excel functions like ***AVERAGE*** and ***MEDIAN***. These functions help to calculate the mean and median of a range of cells, which can be used for imputation or descriptive analysis. I applied these functions to each column of the dataset and replaced the blank cells with the median value of each variable.

By identifying and dealing with missing data in the bank loan application dataset, I was able to prepare the data for further analysis and ensure its quality and reliability.

2. Identify Outliers in the Dataset:

Another important step in data analysis is to **identify outliers**. Outliers are extreme values that deviate significantly from the rest of the data. They can have a large impact on the analysis and distort the results, such as skewing the distribution, inflating the variance, or affecting the correlation. Therefore, it is important to identify outliers and investigate their causes.

To identify outliers in the bank loan application dataset, I used Excel functions like ***QUARTILE, IQR,*** and ***conditional formatting***. These functions help to calculate the ***quartiles*** and ***interquartile range*** (IQR) of a range of cells, which can be used for identifying outliers or creating box plots. Quartiles are values that divide a sorted list of values into four equal parts, such that each part contains 25% of the data. The ***first quartile*** (Q1) is the median of the lower half of the data, and the ***third quartile*** (Q3) is the median of the upper half of the data. The IQR is the difference between Q3 and Q1, which measures the spread of the middle 50% of the data. **Outliers** are usually defined as values that are more than 1.5 times the IQR above Q3 or below Q1.

I applied these functions to each column of the dataset and calculated the quartiles and IQR for each variable. I focused on numerical variables as they are more likely to have outliers than categorical variables. The table below summarizes the quartiles, IQR, and number of outliers for each of the numerical variables.

	Q1	Q3	IQR	Lower Threshold	Upper Threshold	Number of Outliers
LIVINGAREA_MEDI	0.075	0.075	0	0.075	0.075	24847
APARTMENTS_MEDI	0.0874	0.0874	0	0.0874	0.0874	24512
YEARS_BEGINEXPLUATATION_MEDI	0.9816	0.9821	0.0005	0.98085	0.98285	23115
NONLIVINGAREA_MEDI	0.0031	0.0031	0	0.0031	0.0031	22365
EXT_SOURCE_1	0.502257247	0.502257247	0	0.502257247	0.502257247	21827
BASEMENTAREA_MEDI	0.0757	0.0757	0	0.0757	0.0757	20781
LANDAREA_MEDI	0.0488	0.0488	0	0.0488	0.0488	20262
BASEMENTAREA_AVG	0.086	0.088945716	0.002945716	0.081581425	0.093364291	19121

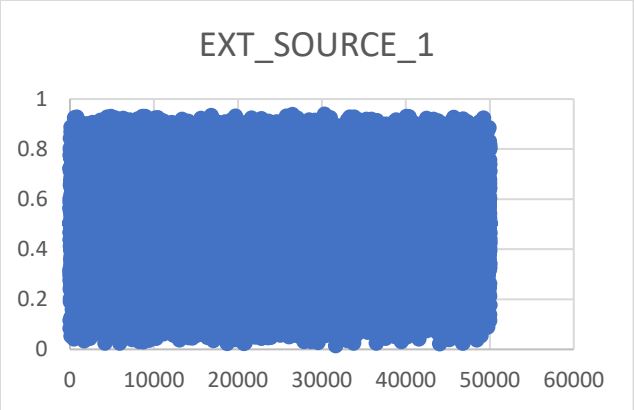
ENTRANCES_MEDI	0.1379	0.1379	0	0.1379	0.1379	19056
ENTRANCES_MODE	0.1379	0.1379	0	0.1379	0.1379	19017
ENTRANCES_AVG	0.1379	0.150550613	0.012650613	0.118924081	0.169526532	18640
LANDAREA_AVG	0.0631	0.066352022	0.003252022	0.058221967	0.071230055	18458
YEARS_BUILD_AVG	0.751638687	0.751638687	0	0.751638687	0.751638687	16760
YEARS_BUILD_MEDI	0.7585	0.7585	0	0.7585	0.7585	16371
FLOORSMIN_AVG	0.231649662	0.231649662	0	0.231649662	0.231649662	16105
LIVINGAPARTMENTS_AVG	0.100435573	0.100435573	0	0.100435573	0.100435573	15773
LIVINGAPARTMENTS_MEDI	0.0761	0.0761	0	0.0761	0.0761	15674
COMMONAREA_AVG	0.044796489	0.044796489	0	0.044796489	0.044796489	15039
COMMONAREA_MEDI	0.0207	0.0207	0	0.0207	0.0207	15020
LIVINGAPARTMENTS_MODE	0.0551	0.0551	0	0.0551	0.0551	14964
YEARS_BUILD_MODE	0.8236	0.8301	0.0065	0.81385	0.83985	14914
FLOORSMAX_MEDI	0.1667	0.1667	0	0.1667	0.1667	14692
FLOORSMAX_MODE	0.1667	0.1667	0	0.1667	0.1667	14353
NONLIVINGAPARTMENTS_AVG	0.0077	0.009096637	0.001396637	0.005605044	0.011191593	14060
REGION_RATING_CLIENT	2	2	0	2	2	13035
REGION_RATING_CLIENT_W_CITY	2	2	0	2	2	12658
APARTMENTS_MODE	0.063	0.084	0.021	0.0315	0.1155	12336
DAYS_EMPLOYED	-2786	-292	2494	-6527	3449	11712
REG_CITY_NOT_WORK_CITY	0	0	0	0	0	11608
NONLIVINGAREA_MODE	0	0	0	0	0	11590
FLOORSMAX_AVG	0.1667	0.225466801	0.058766801	0.078549799	0.313617002	11486
COMMONAREA_MODE	0	0.0028	0.0028	-0.0042	0.007	11282
APARTMENTS_AVG	0.09065	0.117771378	0.027121378	0.049967933	0.158453445	10922
FLOORSMIN_MEDI	0.2083	0.2083	0	0.2083	0.2083	10603
FLOORSMIN_MODE	0.2083	0.2083	0	0.2083	0.2083	10492
FLAG_WORK_PHONE	0	0	0	0	0	9963
ELEVATORS_MEDI	0	0	0	0	0	9110
LIVE_CITY_NOT_WORK_CITY	0	0	0	0	0	8985

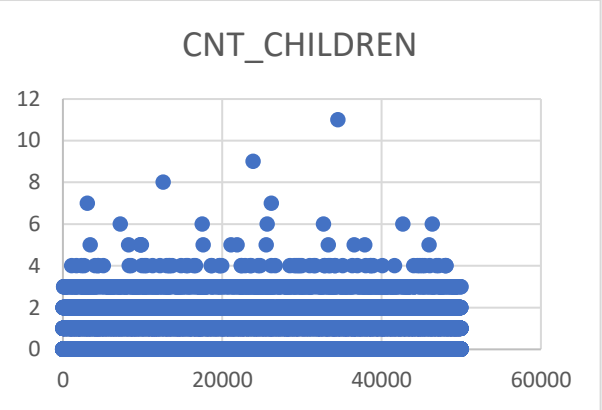
FLAG_EMP_PHONE	1	1	0	1	1	8926
LIVINGAREA_AVG	0.0751	0.107689534	0.032589534	0.026215699	0.156573836	8738
ELEVATORS_MODE	0	0	0	0	0	8690
AMT_REQ_CREDIT_BUREAU_QRT	0	0	0	0	0	8134
AMT_REQ_CREDIT_BUREAU_MON	0	0	0	0	0	7140
YEARS_BEGINEXPLUATATION_AVG	0.978035833	0.9821	0.004064167	0.971939582	0.988196251	6779
OWN_CAR_AGE	0	5	5	-7.5	12.5	6154
NONLIVINGAPARTMENTS_MEDI	0	0	0	0	0	6110
LANDAREA_MODE	0	0.0311	0.0311	-0.04665	0.07775	5754
DEF_30_CNT_SOCIAL_CIRCLE	0	0	0	0	0	5642
NONLIVINGAPARTMENTS_MODE	0	0	0	0	0	5599
FLAG_DOCUMENT_6	0	0	0	0	0	4335
TOTALAREA_MODE	0	0.0701	0.0701	-0.10515	0.17525	4329
DEF_60_CNT_SOCIAL_CIRCLE	0	0	0	0	0	4108
FLAG_DOCUMENT_8	0	0	0	0	0	4038
TARGET	0	0	0	0	0	4026
REG_CITY_NOT_LIVE_CITY	0	0	0	0	0	3998
ELEVATORS_AVG	0	0.078677814	0.078677814	-0.118016721	0.196694535	3871
LIVINGAREA_MODE	0	0.0728	0.0728	-0.1092	0.182	3830
YEARS_BEGINEXPLUATATION_MODE	0.9811	0.9871	0.006	0.9721	0.9961	3644
OBS_30_CNT_SOCIAL_CIRCLE	0	2	2	-3	5	3224
OBS_60_CNT_SOCIAL_CIRCLE	0	2	2	-3	5	3166
NONLIVINGAREA_AVG	0.0061	0.028293807	0.022193807	-0.027190711	0.061584516	3023
FLAG_EMAIL	0	0	0	0	0	2783
BASEMENTAREA_MODE	0	0.0623	0.0623	-0.09345	0.15575	2635
REG_REGION_NOT_WORK_REGION	0	0	0	0	0	2496
AMT_GOODS_PRICE	238500	679500	441000	-423000	1341000	2387
AMT_INCOME_TOTAL	112500	202500	90000	-22500	337500	2295

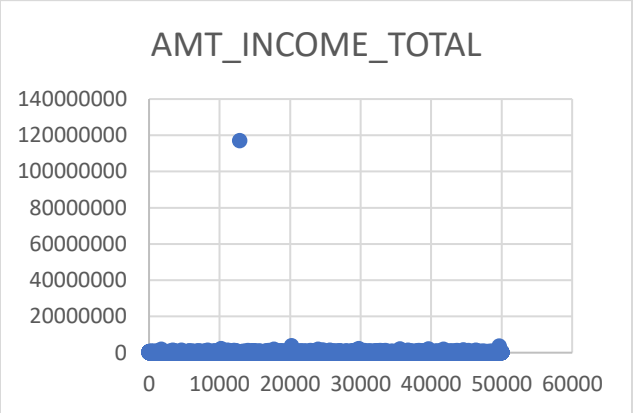
LIVE_REGION_NOT_WORK_REGION	0	0	0	0	0	1982
REGION_POPULATION_RELATIVE	0.010006	0.028663	0.018657	-0.0179795	0.0566485	1329
AMT_REQ_CREDIT_BUREAU_WEEK	0	0	0	0	0	1314
AMT_ANNUITY	16456.5	34596	18139.5	-10752.75	61805.25	1188
AMT_CREDIT	270000	808650	538650	-537975	1616625	1063
FLAG_DOCUMENT_5	0	0	0	0	0	785
REG_REGION_NOT_LIVE_REGION	0	0	0	0	0	750
CNT_CHILDREN	0	1	1	-1.5	2.5	723
CNT_FAM_MEMBERS	2	3	1	0.5	4.5	683
EXT_SOURCE_3	0.417099668	0.638043528	0.22094386	0.085683878	0.969459318	629
AMT_REQ_CREDIT_BUREAU_YEAR	0	3	3	-4.5	7.5	552
FLAG_DOCUMENT_16	0	0	0	0	0	501
FLAG_DOCUMENT_18	0	0	0	0	0	425
HOUR_APPR_PROCESS_START	10	14	4	4	20	353
AMT_REQ_CREDIT_BUREAU_HOUR	0	0	0	0	0	295
AMT_REQ_CREDIT_BUREAU_DAY	0	0	0	0	0	272
FLAG_DOCUMENT_11	0	0	0	0	0	213
FLAG_DOCUMENT_9	0	0	0	0	0	184
FLAG_DOCUMENT_13	0	0	0	0	0	161
FLAG_DOCUMENT_14	0	0	0	0	0	158
FLAG_CONT_MOBILE	1	1	0	1	1	101
DAYS_REGISTRATION	-7463.5	-1998	5465.5	-15661.75	6200.25	96
DAYS_LAST_PHONE_CHANGE	-1573	-270	1303	-3527.5	1684.5	63
FLAG_DOCUMENT_15	0	0	0	0	0	41
FLAG_DOCUMENT_19	0	0	0	0	0	35
FLAG_DOCUMENT_20	0	0	0	0	0	26
FLAG_DOCUMENT_21	0	0	0	0	0	19
FLAG_DOCUMENT_17	0	0	0	0	0	15
FLAG_DOCUMENT_7	0	0	0	0	0	11
FLAG_DOCUMENT_4	0	0	0	0	0	9
FLAG_DOCUMENT_2	0	0	0	0	0	2
FLAG_MOBIL	1	1	0	1	1	1
FLAG_DOCUMENT_10	0	0	0	0	0	1
DAYS_BIRTH	-19644	-12378.5	7265.5	-30542.25	-1480.25	0

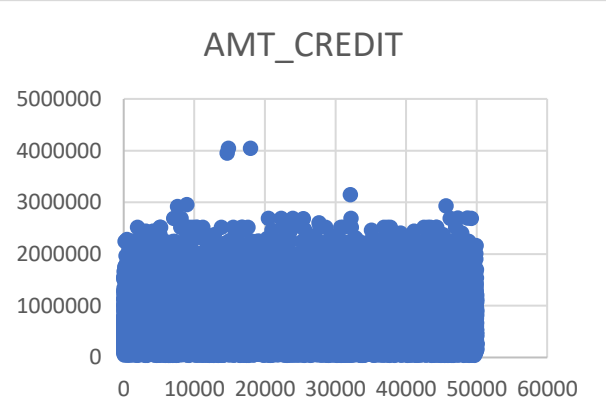
DAYS_ID_PUBLISH	-4297	-1722	2575	-8159.5	2140.5	0
FLAG_PHONE	0	1	1	-1.5	2.5	0
EXT_SOURCE_2	0.392220206	0.663160518	0.270940313	-0.014190263	1.069570987	0
FLAG_DOCUMENT_3	0	1	1	-1.5	2.5	0
FLAG_DOCUMENT_12	0	0	0	0	0	0

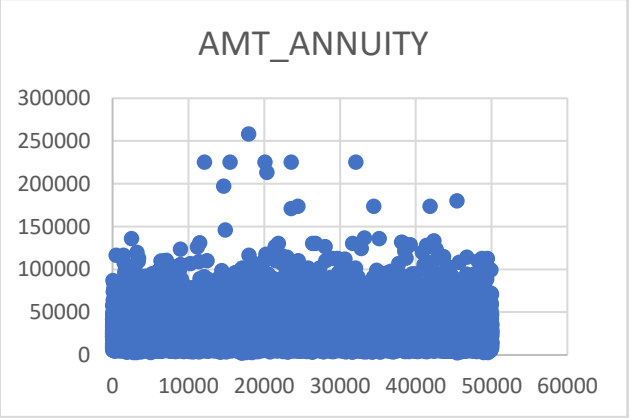
To visualize the outliers in the bank loan application dataset, I created ***box plots*** using Excel's chart feature. The box plots below show the distribution of each variable.

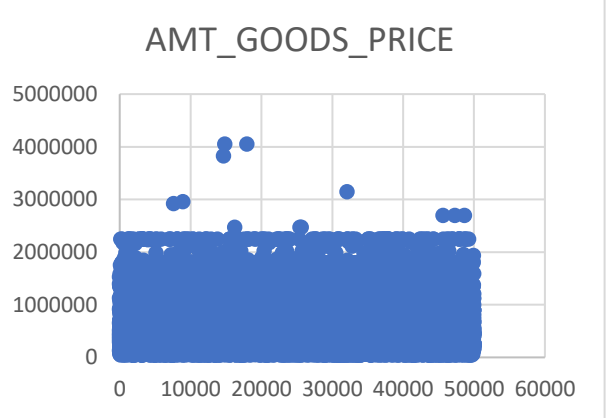


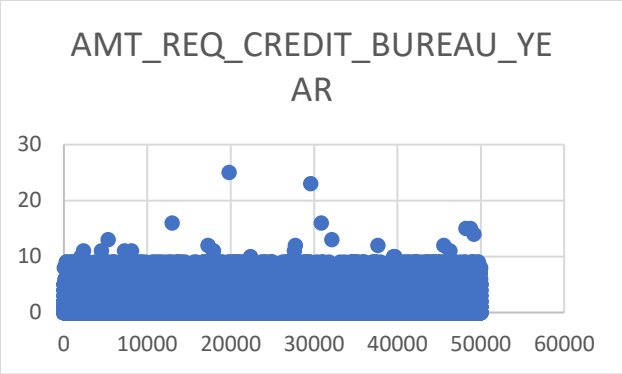


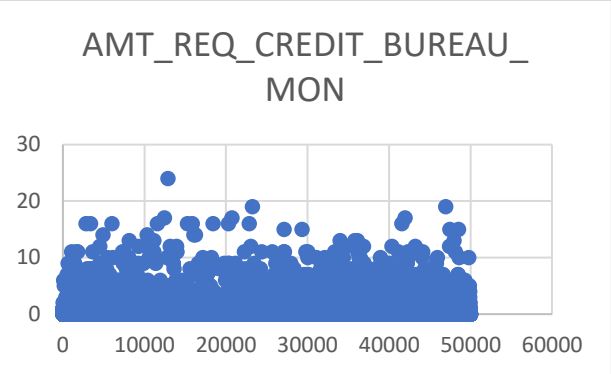


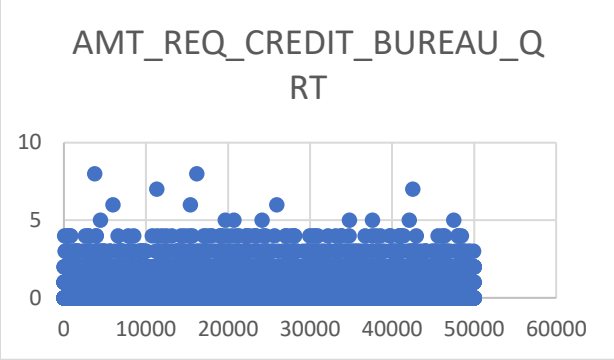


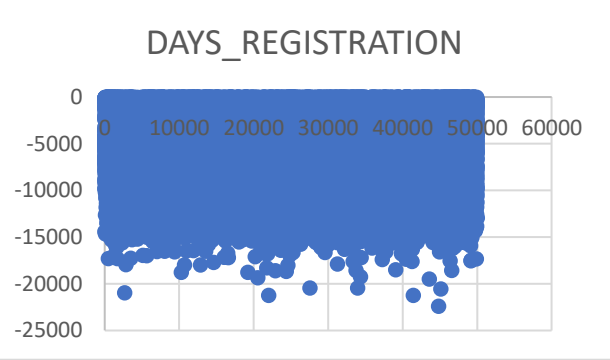


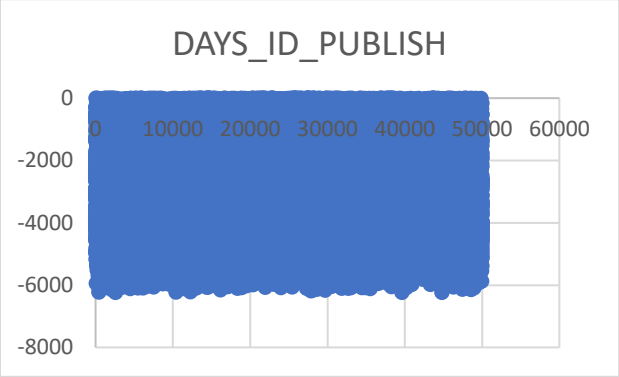


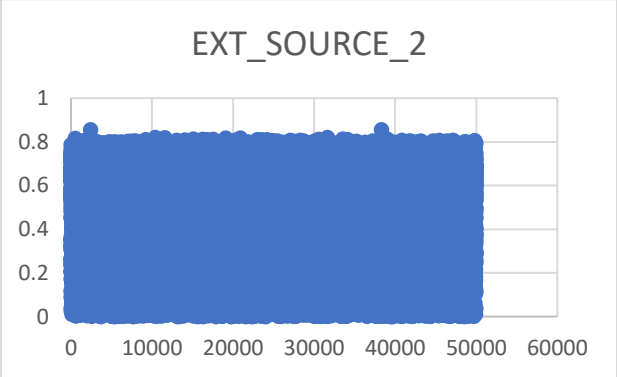

AMT_REQ_CREDIT_BUREAU_YEAR
30
20
10
0
0
10000
20000
30000
40000
50000
60000

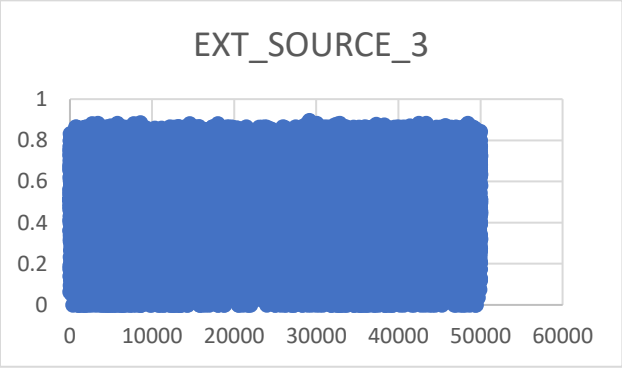

AMT_REQ_CREDIT_BUREAU_MON
30
20
10
0
0
10000
20000
30000
40000
50000
60000

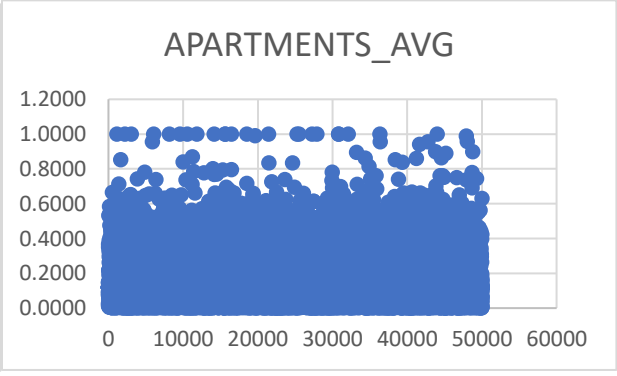

AMT_REQ_CREDIT_BUREAU_QRT
10
5
0
0
10000
20000
30000
40000
50000
60000

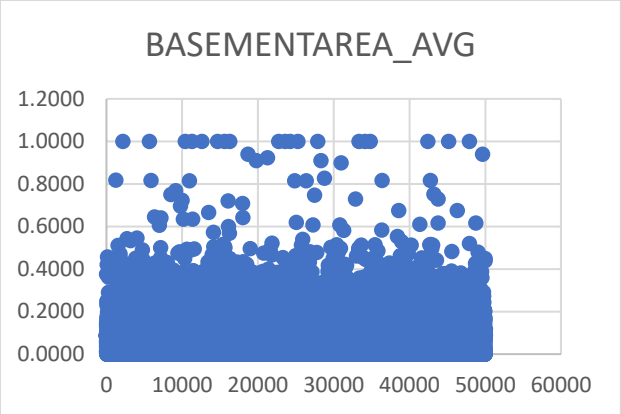

DAYS_REGISTRATION
0
-5000
-10000
-15000
-20000
-25000
0
10000
20000
30000
40000
50000
60000

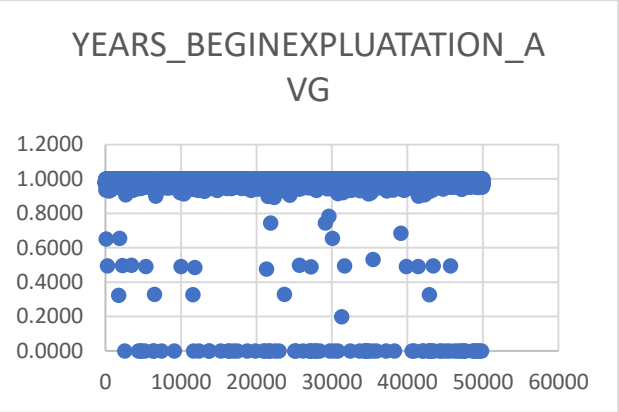

DAYS_ID_PUBLISH
0
-2000
-4000
-6000
-8000
0
10000
20000
30000
40000
50000
60000

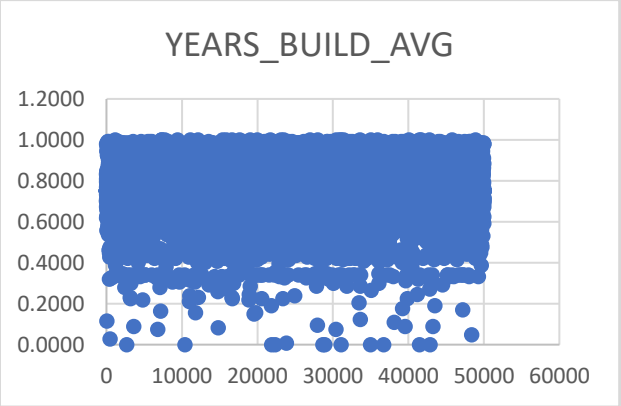

EXT_SOURCE_2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

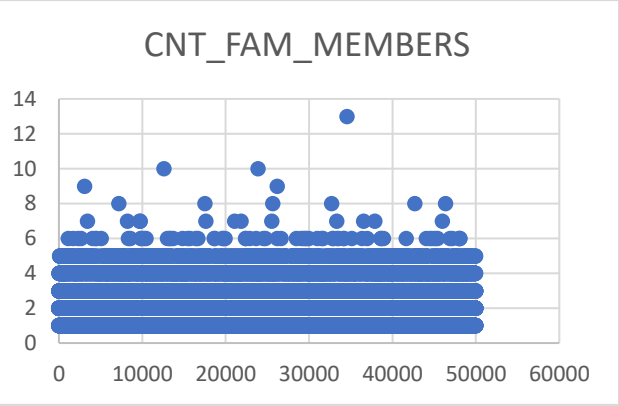

EXT_SOURCE_3
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

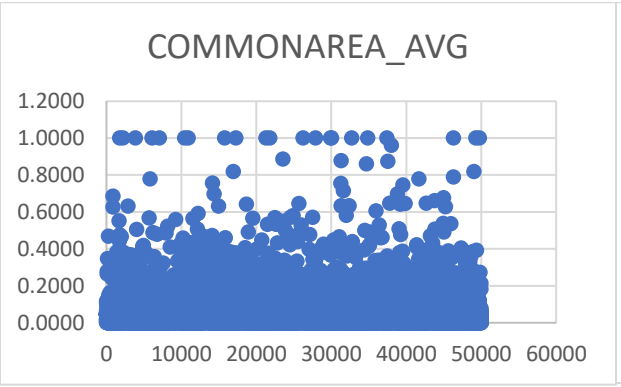

APARTMENTS_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

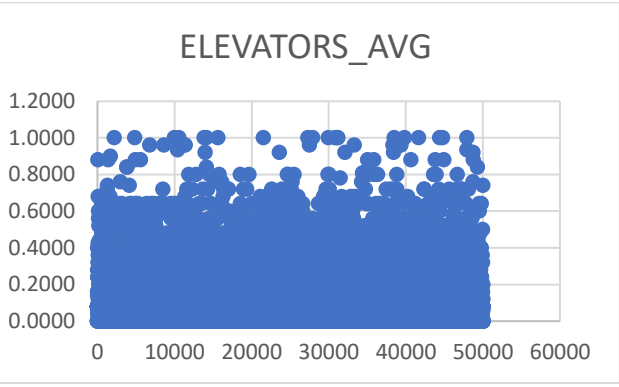

BASEMENTAREA_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

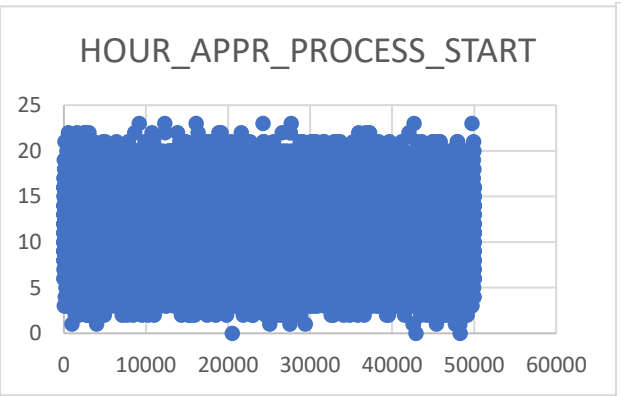

YEARS_BEGINEXPLUATATION_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

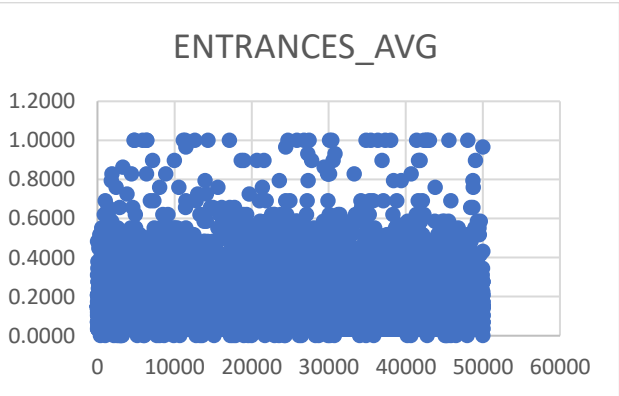

YEARS_BUILD_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

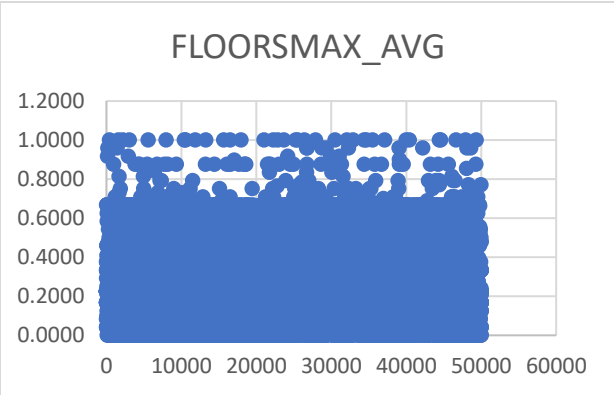

CNT_FAM_MEMBERS
14
12
10
8
6
4
2
0
0
10000
20000
30000
40000
50000
60000

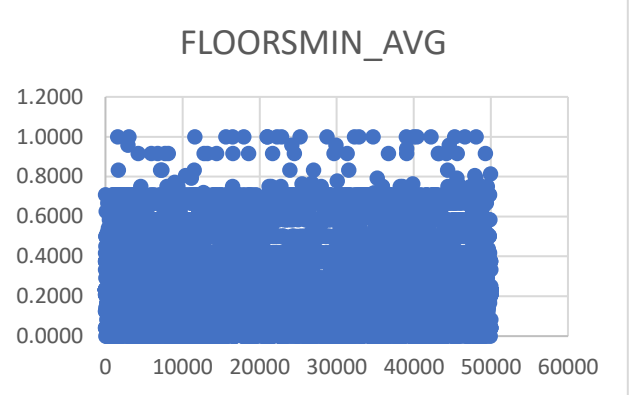

COMMONAREA_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

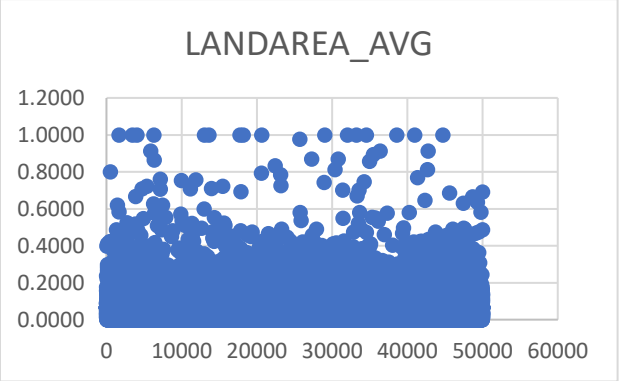

ELEVATORS_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

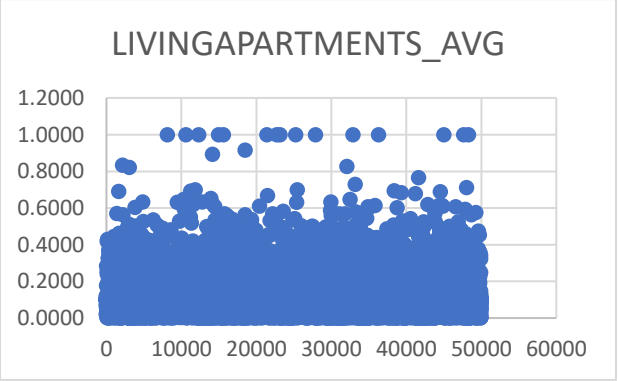

HOUR_APPR_PROCESS_START
25
20
15
10
5
0
0
10000
20000
30000
40000
50000
60000

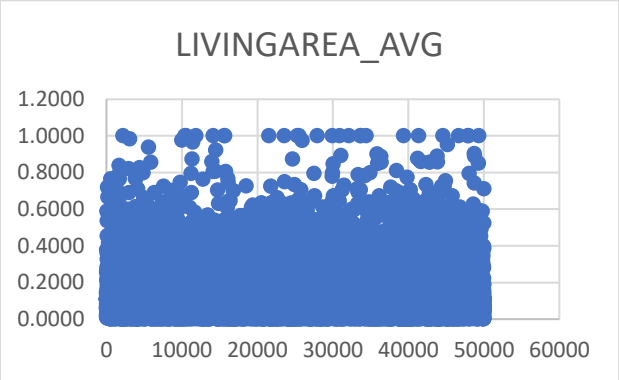

ENTRANCES_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

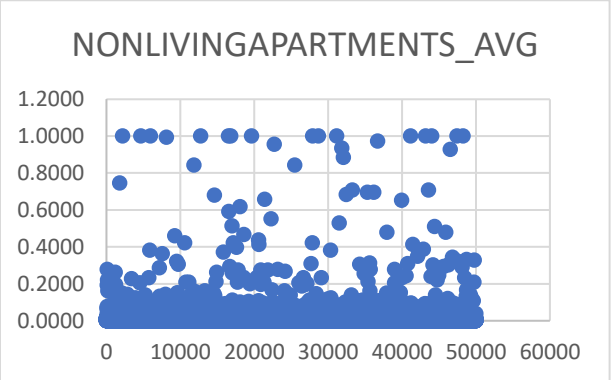

FLOORSMAX_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

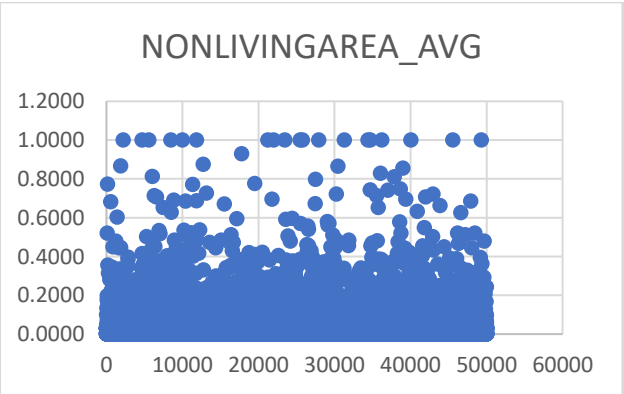

FLOORSMIN_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

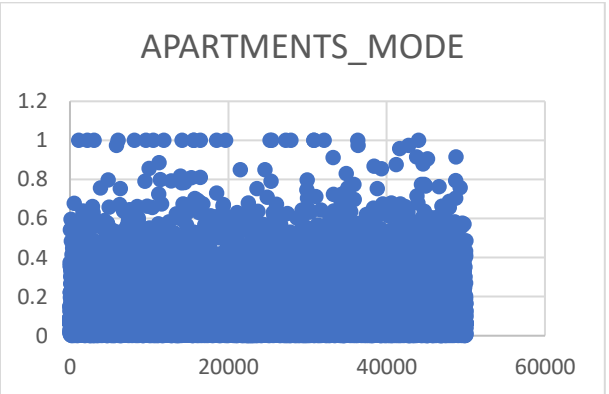

LANDAREA_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

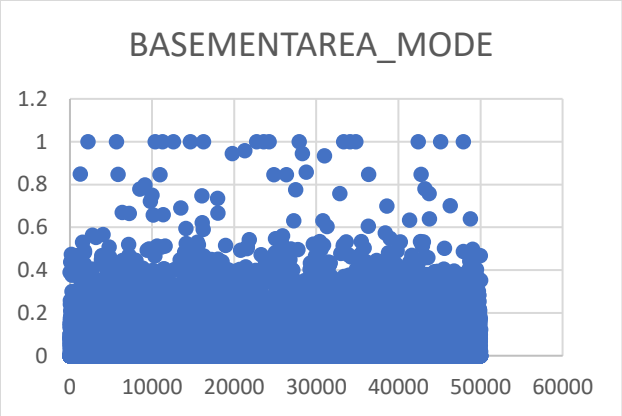

LIVINGAPARTMENTS_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

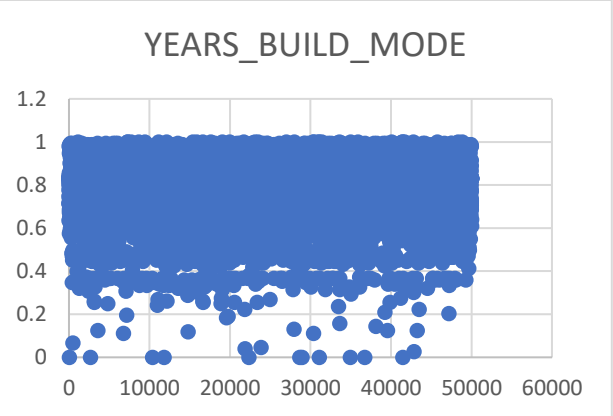

LIVINGAREA_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

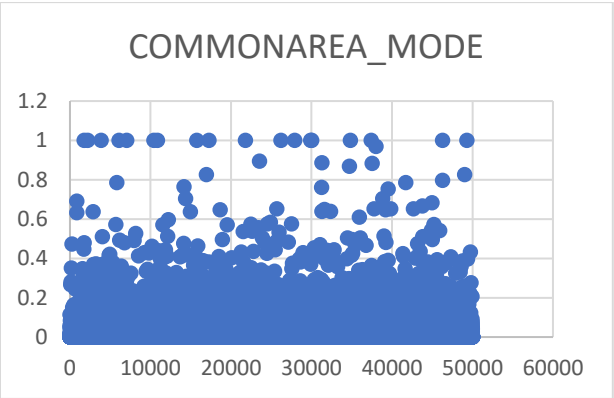

NONLIVINGAPARTMENTS_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

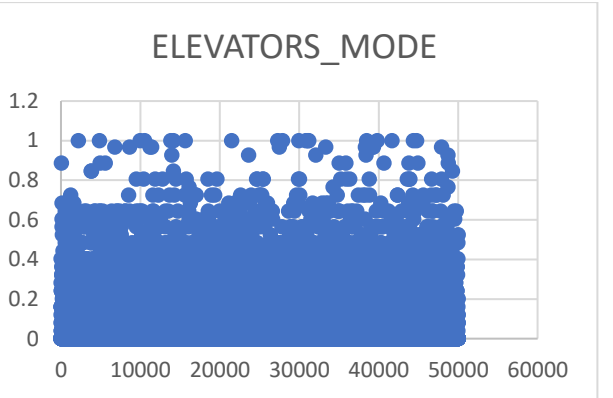

NONLIVINGAREA_AVG
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
10000
20000
30000
40000
50000
60000

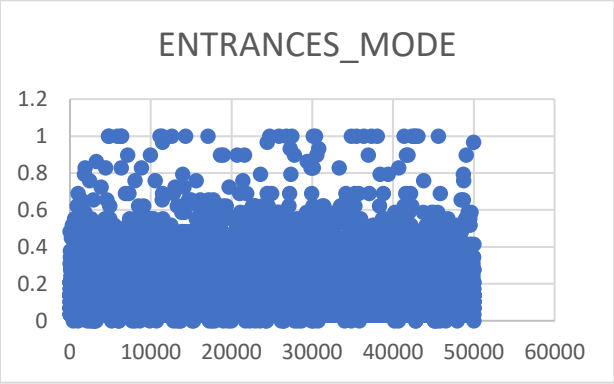

APARTMENTS_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
20000
40000
60000

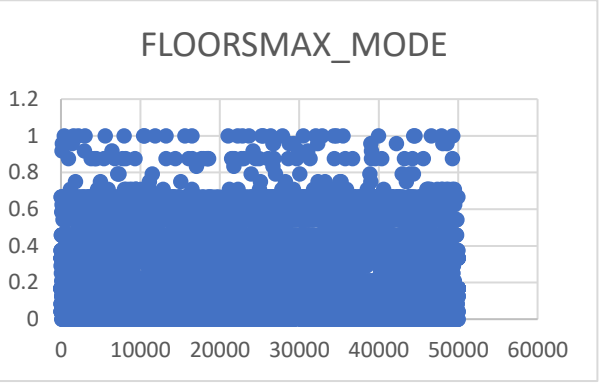

BASEMENTAREA_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

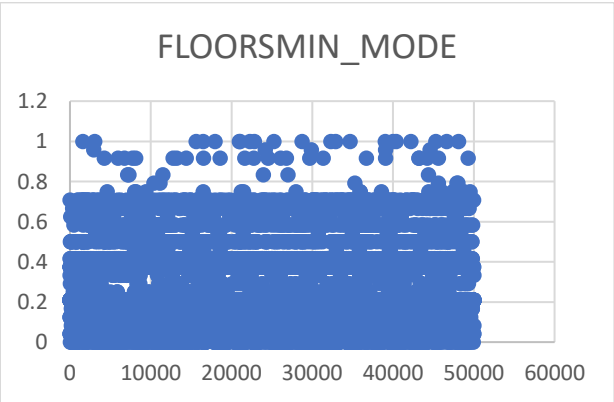

YEARS_BUILD_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

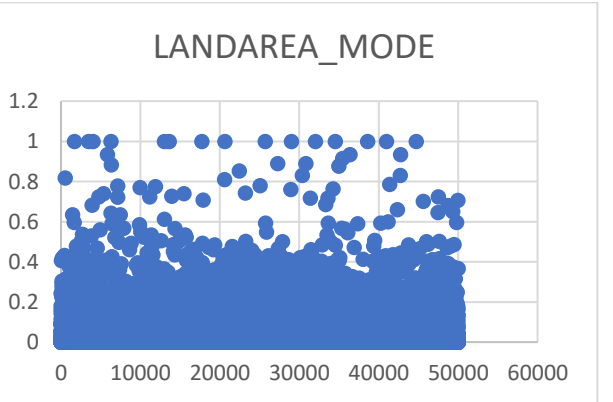

COMMONAREA_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

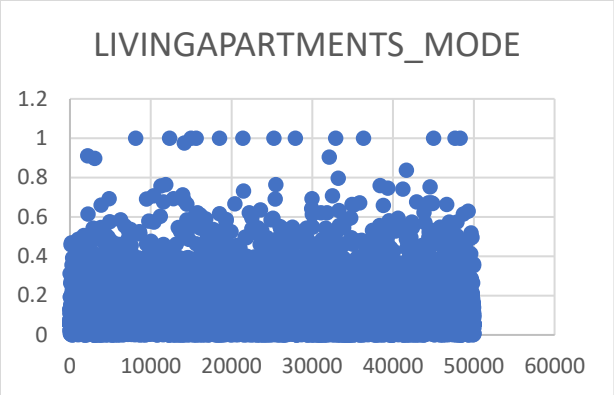

ELEVATORS_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

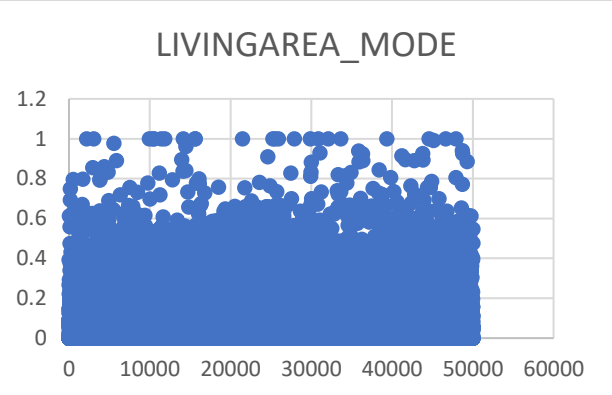

ENTRANCES_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

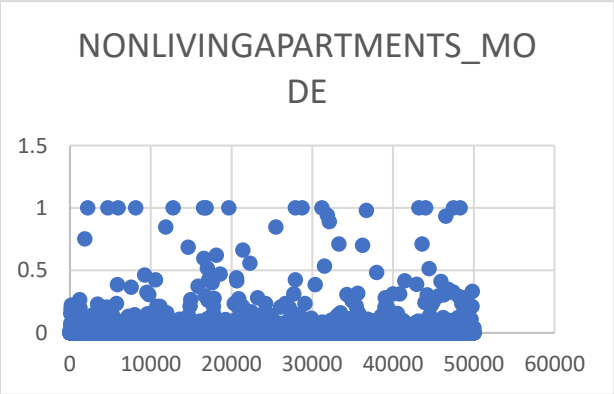

FLOORSMAX_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

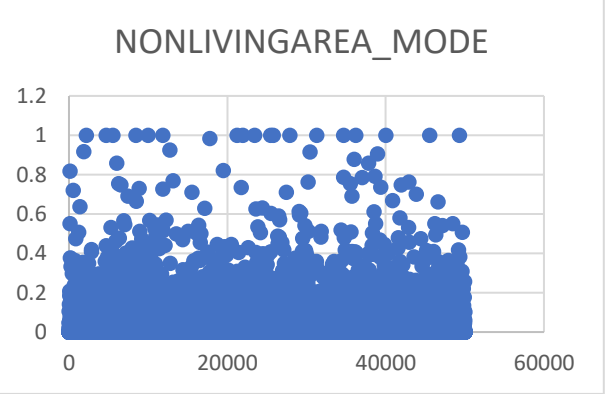

FLOORSMIN_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

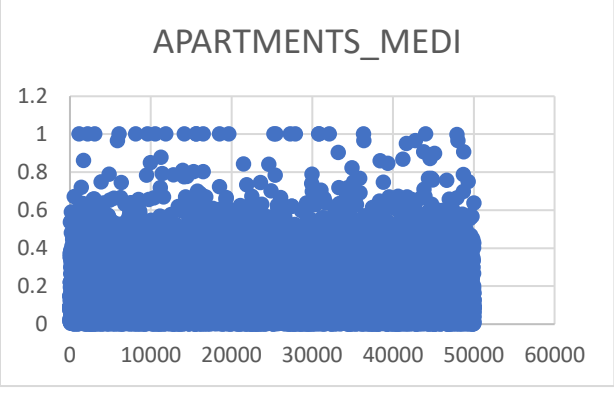

LANDAREA_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

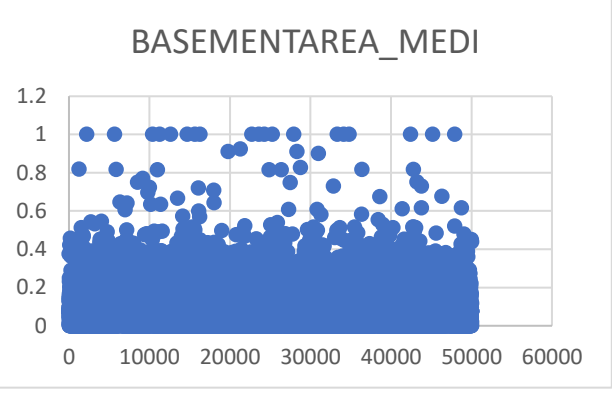

LIVINGAPARTMENTS_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

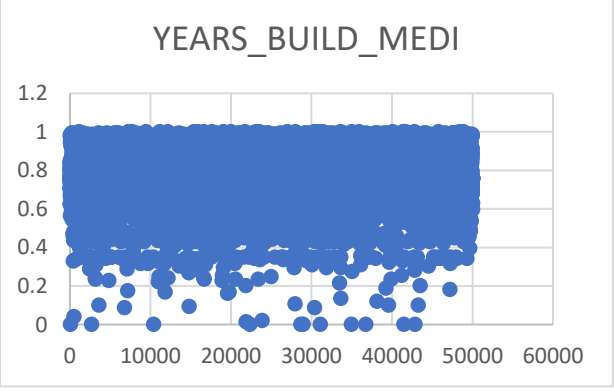

LIVINGAREA_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

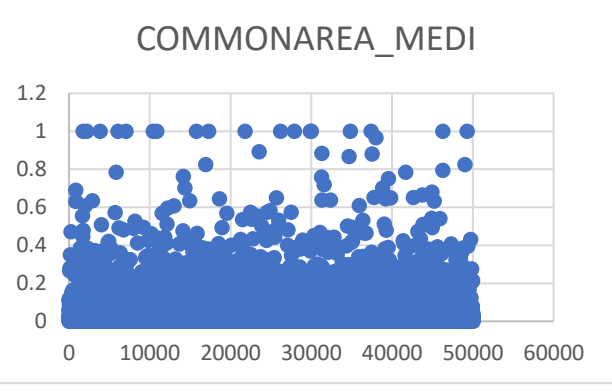

NONLIVINGAPARTMENTS_MODE
1.5
1
0.5
0
0
10000
20000
30000
40000
50000
60000

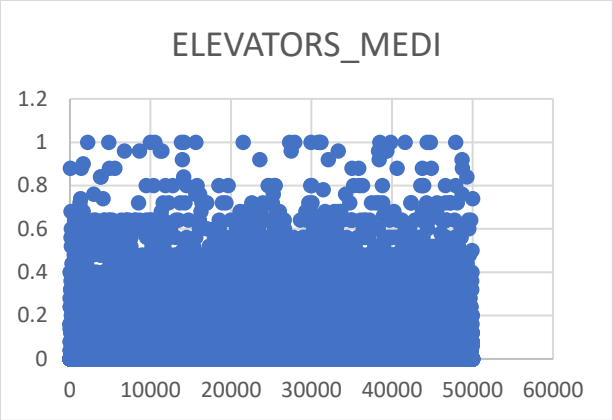

NONLIVINGAREA_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
20000
40000
60000

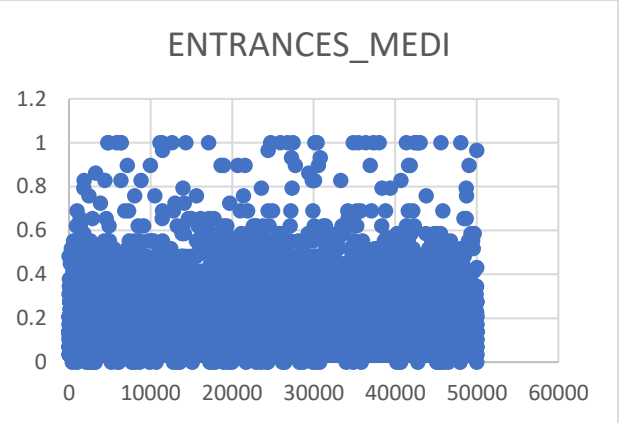

APARTMENTS_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

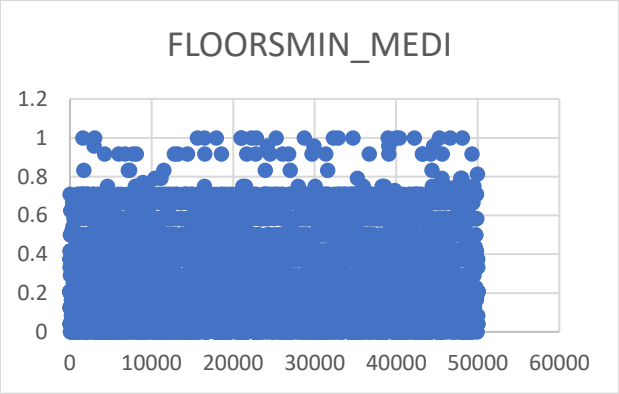

BASEMENTAREA_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

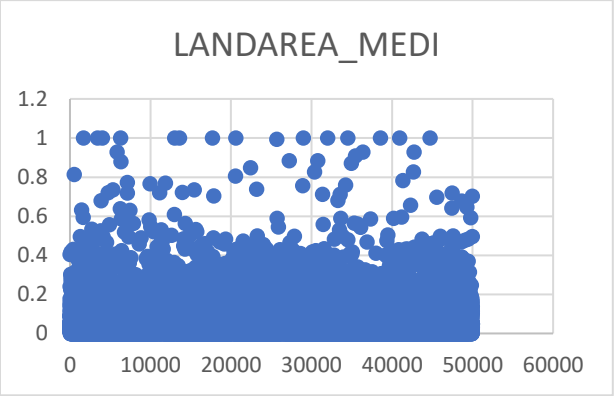

YEARS_BUILD_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

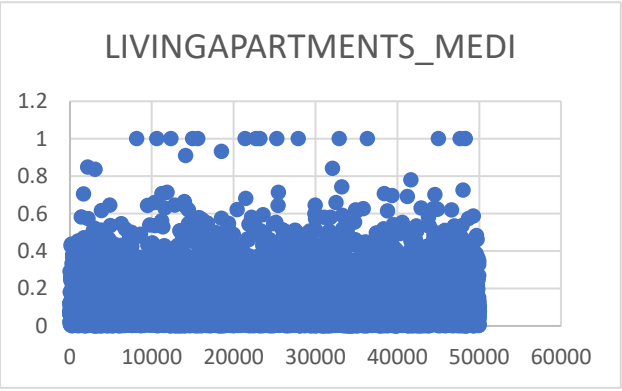

COMMONAREA_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

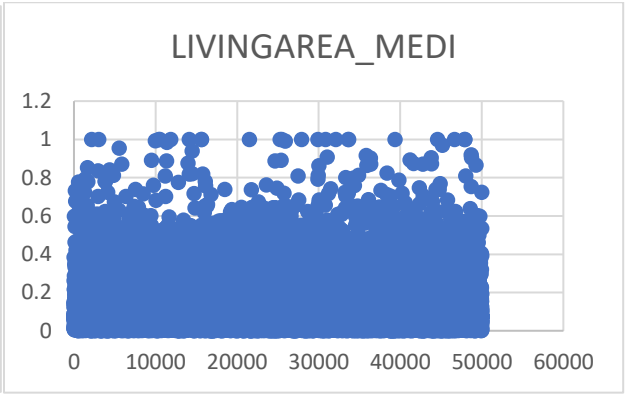

ELEVATORS_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

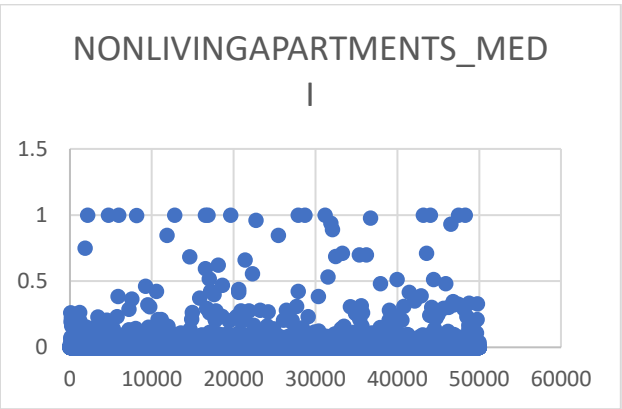

ENTRANCES_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

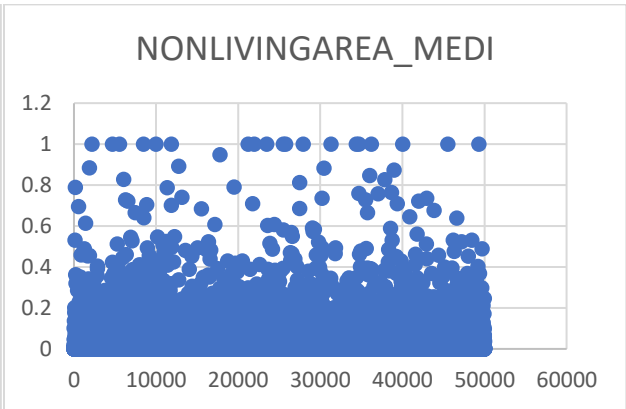

FLOORSMIN_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

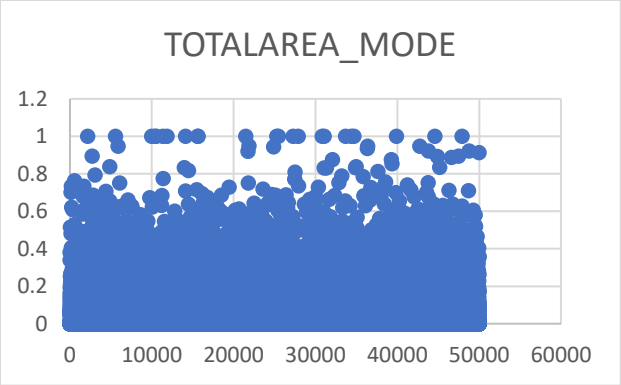

LANDAREA_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

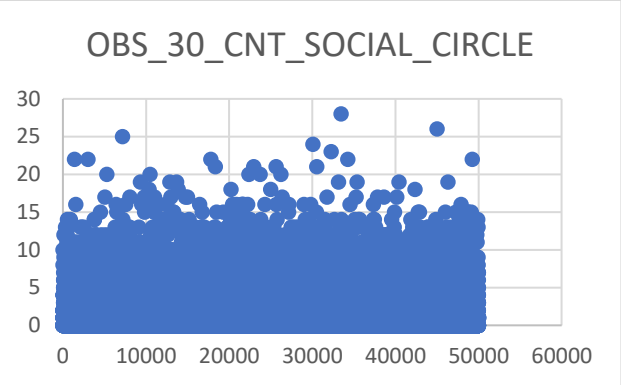

LIVINGAPARTMENTS_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

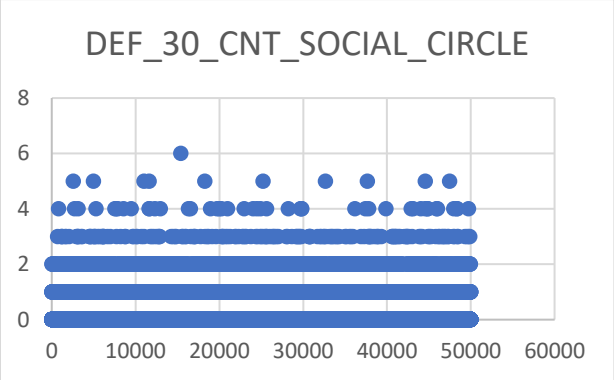

LIVINGAREA_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

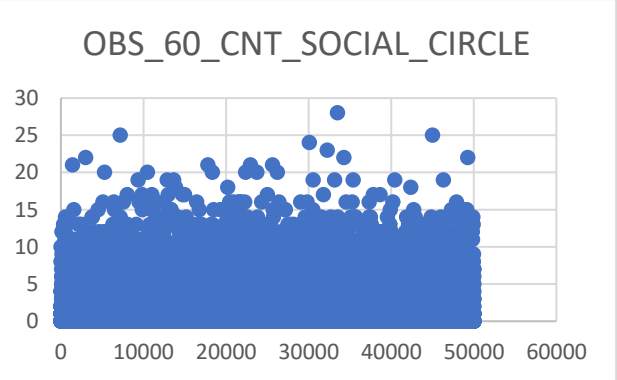

NONLIVINGAPARTMENTS_MEDI
1.5
1
0.5
0
0
10000
20000
30000
40000
50000
60000

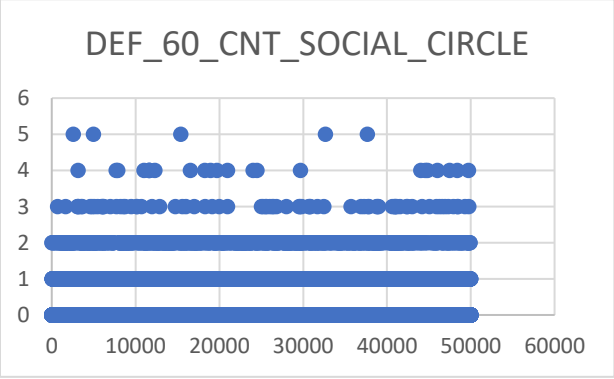

NONLIVINGAREA_MEDI
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

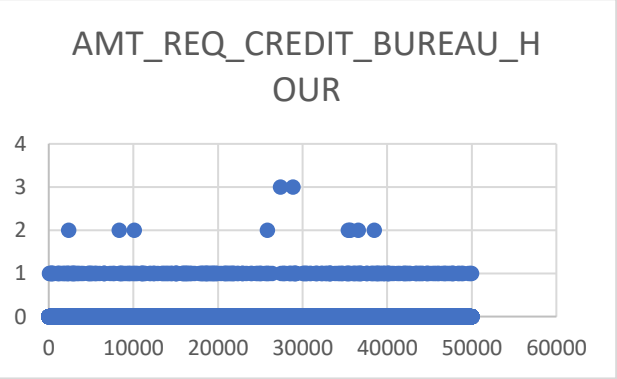

TOTALAREA_MODE
1.2
1
0.8
0.6
0.4
0.2
0
0
10000
20000
30000
40000
50000
60000

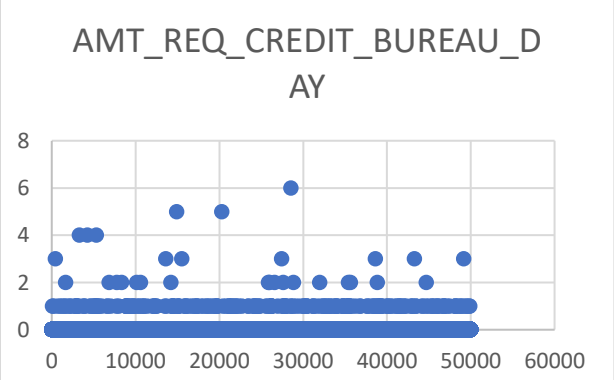

OBS_30_CNT_SOCIAL_CIRCLE
30
25
20
15
10
5
0
0
10000
20000
30000
40000
50000
60000

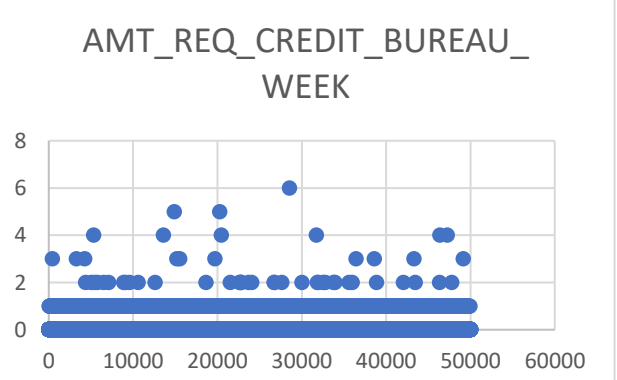

DEF_30_CNT_SOCIAL_CIRCLE
8
6
4
2
0
0
10000
20000
30000
40000
50000
60000


OBS_60_CNT_SOCIAL_CIRCLE
30
25
20
15
10
5
0
0
10000
20000
30000
40000
50000
60000


DEF_60_CNT_SOCIAL_CIRCLE
6
5
4
3
2
1
0
0
10000
20000
30000
40000
50000
60000


AMT_REQ_CREDIT_BUREAU_HOUR
4
3
2
1
0
0
10000
20000
30000
40000
50000
60000


AMT_REQ_CREDIT_BUREAU_DAY
8
6
4
2
0
0
10000
20000
30000
40000
50000
60000


AMT_REQ_CREDIT_BUREAU_WEEK
8
6
4
2
0
0
10000
20000
30000
40000
50000
60000

3. Analyze Data Imbalance:

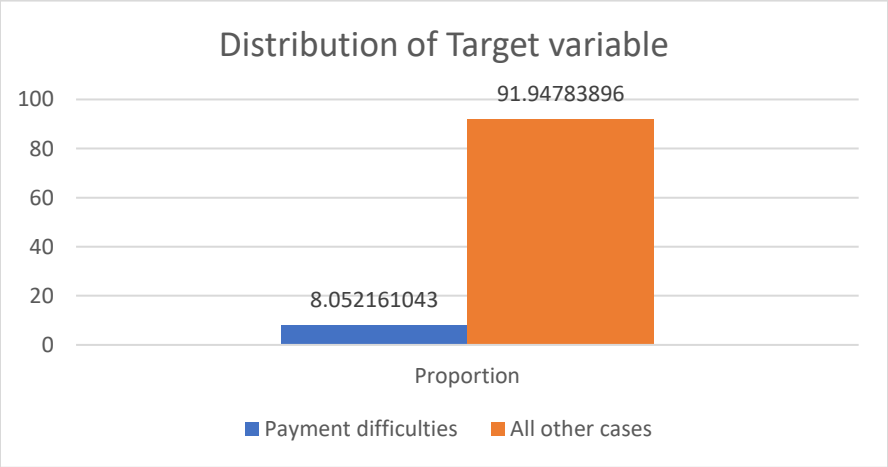
Another important step in data analysis is to **analyze data imbalance**. Data imbalance occurs when one class of the target variable has a much higher frequency than the other class. This can affect the accuracy of predictive models that aim to classify customers into default or non-default groups, as they may be biased towards the majority class and ignore the minority class. Therefore, it is important to analyze data imbalance and apply appropriate techniques to balance the data.

To analyze data imbalance in the bank loan application dataset, I used Excel functions like ***COUNTIF*** and ***SUM***. These functions help to calculate the frequency or sum of a range of cells that meet a given criterion, which can be used for analyzing data imbalance or segmented analysis. I applied these functions to the target variable, which indicates whether a customer had payment difficulties or not.

The results showed that there was data imbalance in the bank loan application dataset. The table below summarizes the class frequencies and proportions of the target variable.

Target variable	**Frequency**	**Proportion**
Payment difficulties	4026	8.052161
All other cases	45973	91.947839
Total	49999	100

To visualize data imbalance in the bank loan application dataset, I created a ***bar chart*** using Excel's chart feature. The bar chart below shows the proportion of each class in the target variable.



By analyzing data imbalance in the bank loan application dataset, I was able to understand the distribution of the target variable and its implications for predictive modeling. Data imbalance is a common problem in many real-world datasets, especially for binary classification problems.

4. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

One of the main steps in data analysis is to **perform univariate, segmented univariate, and bivariate analysis**. These types of analysis help to gain insights into the driving factors of loan default by exploring the characteristics and relationships of consumer and loan attributes.

Univariate analysis is the analysis of a single variable to understand its distribution, central tendency, and variability. It helps to describe the basic features of a variable, such as its range, mean, median, mode, standard deviation, skewness, kurtosis, etc. It also helps to identify outliers or anomalies in the data.

Segmented univariate analysis is the analysis of a single variable for different groups or categories. It helps to compare the distribution of a variable across different segments, such as customers with payment difficulties and all other cases. It also helps to identify significant differences or similarities between segments.

Bivariate analysis is the analysis of two variables to explore their relationship. It helps to measure the strength and direction of the association between two variables, such as correlation or contingency. It also helps to identify patterns or trends in the data.

To perform **univariate analysis** on the bank loan application dataset, I used Excel functions like ***COUNT, AVERAGE, MEDIAN***, and statistical functions. These functions help to calculate the frequency, mean, median, and other descriptive statistics of a range of cells, which can be used for univariate analysis. I applied these functions to each column of the dataset and calculated the descriptive statistics for each variable.

The table below summarizes the descriptive statistics for each variable.

Variable	Frequency		Mean	Median	Mode	Standard Deviation
TARGET	N/A		0.08052161	0	0	0.27210175
NAME_CONTRACT_TYPE	Cash loans	45276	N/A	N/A	Cash loans	N/A
	Revolving loans	4723				
CODE_GENDER	F	32823	N/A	N/A	F	N/A
	M	17174				
	XNA	2				
FLAG_OWN_CAR	N	32949	N/A	N/A	N	N/A
	Y	17050				
FLAG_OWN_REALTY	Y	34691	N/A	N/A	Y	N/A
	N	15308				

CNT_CHILDREN	N/A		0.419848397	0	0	0.72403855
AMT_INCOME_TOTAL	N/A		170767.5905	145800	135000	531819.095
AMT_CREDIT	N/A		599700.5815	514777.5	450000	402415.434
AMT_ANNUITY	N/A		27107.33399	24939	9000	14562.802
AMT_GOODS_PRICE	N/A		538992.3491	450000	450000	369720.822
NAME_TYPE_SUITE	Unaccompanied	40627	N/A	N/A	Unaccompanied	N/A
	Family	6549				
	Spouse, partner	1849				
	Children	542				
	Other_B	259				
	Other_A	137				
	Group of people	36				
NAME_INCOME_TYPE	Working	26010	N/A	N/A	Working	N/A
	Commercial associate	11543				
	Pensioner	8920				
	State servant	3512				
	Unemployed	6				
	Student	5				
	Businessman	2				
	Maternity leave	1				
NAME_EDUCATION_TYPE	Secondary / secondary special	35572	N/A	N/A	Secondary / secondary special	N/A
	Higher education	12167				
	Incomplete higher	1620				
	Lower secondary	620				
	Academic degree	20				
NAME_FAMILY_STATUS	Married	32094	N/A	N/A	Married	N/A

	Single / not married	7306				
	Civil marriage	4859				
	Separated	3142				
	Widow	2597				
	Unknown	1				
NAME_HOUSING_TYPE	House / apartment	44368	N/A	N/A	House / apartment	N/A
	With parents	2399				
	Municipal apartment	1845				
	Rented apartment	769				
	Office apartment	427				
	Co-op apartment	191				
REGION_POPULATION_RELATIVE	N/A		0.020798283	0.01885	0.03579	0.01376058
DAYS_BIRTH	N/A		-16022.0421	-15731	-11039	4361.40027
DAYS_EMPLOYED	N/A		63219.42449	-1221	365243	140794.606
DAYS_REGISTRATION	N/A		-4977.28267	-4490	-3	3525.54831
DAYS_ID_PUBLISH	N/A		-2996.79718	-3261	-4360	1509.23541
OWN_CAR_AGE	N/A		4.100622012	0	0	8.98235174
FLAG_MOBIL	N/A		0.99998	1	1	0.00447218
FLAG_EMP_PHONE	N/A		0.82147643	1	1	0.38295671
FLAG_WORK_PHONE	N/A		0.199263985	0	0	0.39945092
FLAG_CONT_MOBILE	N/A		0.99797996	1	1	0.04489989
FLAG_PHONE	N/A		0.277725555	0	0	0.44788177
FLAG_EMAIL	N/A		0.055661113	0	0	0.22926841

OCCUPATION_TYPE	Laborers	24606	N/A	N/A	Laborers	N/A
	Sales staff	5160				
	Core staff	4434				
	Managers	3489				
	Drivers	3044				
	High skill tech staff	1852				
	Accountants	1621				
	Medicine staff	1403				
	Security staff	1140				
	Cooking staff	963				
	Cleaning staff	739				
	Private service staff	447				
	Low-skill Laborers	357				
	Waiters/barmen staff	228				
	Secretaries	212				
	Realty agents	123				
	HR staff	101				
	IT staff	80				
CNT_FAM_MEMBERS	N/A		2.158943179	2	2	0.91132365
REGION_RATING_CLIENT	N/A		2.051661033	2	2	0.50797787
REGION_RATING_CLIENT_W_CITY	N/A		2.030720614	2	2	0.50222142
WEEKDAY_APPR_PROCESS_START	TUESDAY	8741	N/A	N/A	TUESDAY	N/A
	MONDAY	8385				
	WEDNESDAY	8355				
	FRIDAY	8286				
	THURSDAY	8149				
	SATURDAY	5467				
	SUNDAY	2616				
HOUR_APPR_PROCESS_START	N/A		12.05256105	12	10	3.25258368
REG_REGION_NOT_LIVE_REGION	N/A		0.0150003	0	0	0.12155487

REG_REGION_NOT_WORK_REGION	N/A		0.049920998	0	0	0.21778393
LIVE_REGION_NOT_WORK_REGION	N/A		0.039640793	0	0	0.19511577
REG_CITY_NOT_LIVE_CITY	N/A		0.079961599	0	0	0.27123645
REG_CITY_NOT_WORK_CITY	N/A		0.232164643	0	0	0.4222177
LIVE_CITY_NOT_WORK_CITY	N/A		0.179703594	0	0	0.38394422
ORGANIZATION_TYPE	Business Entity Type 3	11101	N/A	N/A	Business Entity Type 3	N/A
	XNA	8924				
	Self-employed	6240				
	Other	2717				
	Medicine	1817				
	Government	1716				
	Business Entity Type 2	1704				
	School	1450				
	Trade: type 7	1210				
	Kindergarten	1090				
	Construction	1066				
	Business Entity Type 1	953				
	Transport: type 4	837				
	Trade: type 3	550				
	Security	550				
	Industry: type 3	542				
	Industry: type 9	537				
	Housing	489				
	Industry: type 11	489				
	Military	458				
	Bank	435				

	Transport: type 2	392				
	Agriculture	392				
	Postal	370				
	Police	366				
	Security Ministries	331				
	Trade: type 2	307				
	Restaurant	289				
	Services	284				
	University	222				
	Industry: type 7	209				
	Transport: type 3	191				
	Hotel	182				
	Industry: type 1	159				
	Electricity	147				
	Industry: type 4	140				
	Trade: type 6	108				
	Telecom	106				
	Industry: type 5	103				
	Emergency	93				
	Insurance	89				
	Industry: type 2	78				
	Advertising	68				
	Trade: type 1	66				
	Culture	64				
	Realtor	61				
	Mobile	56				
	Industry: type 12	53				
	Legal Services	44				
	Cleaning	40				
	Transport: type 1	28				

	Industry: type 10	21				
	Industry: type 13	15				
	Religion	14				
	Industry: type 6	12				
	Industry: type 8	8				
	Trade: type 5	8				
	Trade: type 4	8				
EXT_SOURCE_1	N/A		0.502257247	0.502257	0.50226	0.13942126
EXT_SOURCE_2	N/A		0.513823595	0.564892	0.51382	0.190924
EXT_SOURCE_3	N/A		0.511881408	0.511881	0.51188	0.17426614
APARTMENTS_AVG	N/A		0.117771378	0.117771	0.11777	0.07610399
BASEMENTAREA_AVG	N/A		0.088945716	0.088946	0.08895	0.05330931
YEARS_BEGINEXPLUATATION_AVG	N/A		0.978035833	0.978036	0.97804	0.04042235
YEARS_BUILD_AVG	N/A		0.751638687	0.751639	0.75164	0.06545755
COMMONAREA_AVG	N/A		0.044796489	0.044796	0.0448	0.04332342
ELEVATORS_AVG	N/A		0.078677814	0.078678	0.07868	0.09183826
ENTRANCES_AVG	N/A		0.150550613	0.150551	0.15055	0.07103618
FLOORSMAX_AVG	N/A		0.225466801	0.225467	0.22547	0.10290264
FLOORSMIN_AVG	N/A		0.231649662	0.23165	0.23165	0.09168226
LANDAREA_AVG	N/A		0.066352022	0.066352	0.06635	0.05072318
LIVINGAPARTMENTS_AVG	N/A		0.100435573	0.100436	0.10044	0.05211357
LIVINGAREA_AVG	N/A		0.107689534	0.10769	0.10769	0.07814067
NONLIVINGAPARTMENTS_AVG	N/A		0.009096637	0.009097	0.0091	0.0272195
NONLIVINGAREA_AVG	N/A		0.028293807	0.028294	0.02829	0.04605914

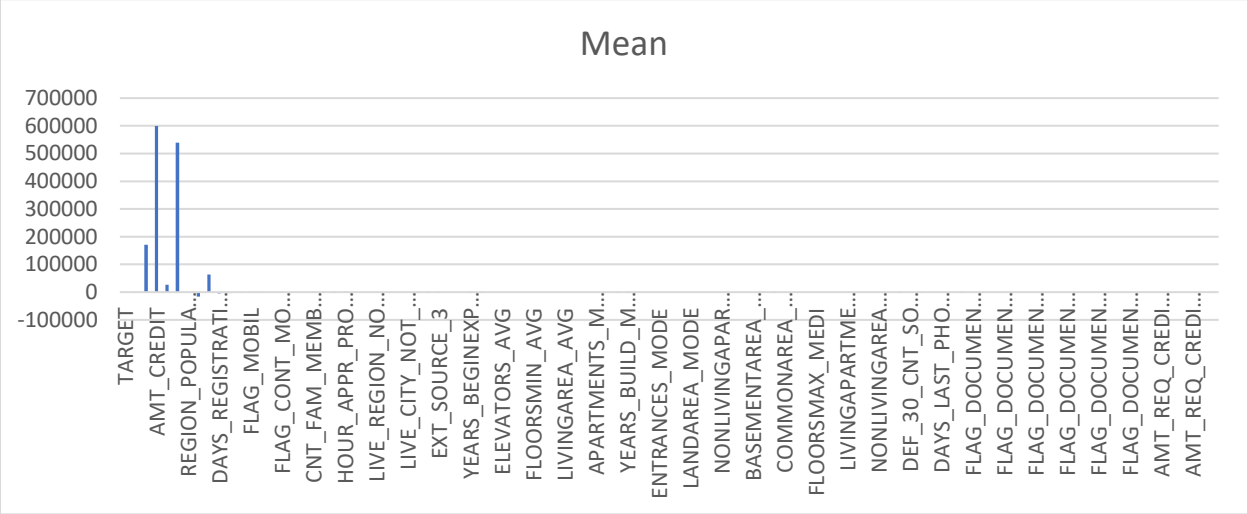
APARTMENTS_MODE	N/A	0.088325193	0.063	0.063	0.08002937
BASEMENTAREA_MODE	N/A	0.036613646	0	0	0.06948641
YEARS_BEGINEXPLUATATION_MODE	N/A	0.982134435	0.9871	0.9871	0.04438761
YEARS_BUILD_MODE	N/A	0.806255675	0.8301	0.8301	0.07189258
COMMONAREA_MODE	N/A	0.012832437	0	0	0.04642595
ELEVATORS_MODE	N/A	0.034629145	0	0	0.0975161
ENTRANCES_MODE	N/A	0.141917624	0.1379	0.1379	0.07169509
FLOORSMAX_MODE	N/A	0.194230753	0.1667	0.1667	0.10588539
FLOORSMIN_MODE	N/A	0.214729313	0.2083	0.2083	0.09220685
LANDAREA_MODE	N/A	0.026298378	0	0	0.06028225
LIVINGAPARTMENTS_MODE	N/A	0.070943803	0.0551	0.0551	0.05993952
LIVINGAREA_MODE	N/A	0.052763393	0	0	0.09509357
NONLIVINGAPARTMENTS_MODE	N/A	0.002513974	0	0	0.02602805
NONLIVINGAREA_MODE	N/A	0.012161901	0	0	0.04885051
APARTMENTS_MEDI	N/A	0.102563807	0.0874	0.0874	0.0782386
BASEMENTAREA_MEDI	N/A	0.081035579	0.0757	0.0757	0.05343355
YEARS_BEGINEXPLUATATION_MEDI	N/A	0.979772085	0.9816	0.9816	0.04108815
YEARS_BUILD_MEDI	N/A	0.757310536	0.7585	0.7585	0.06476116
COMMONAREA_MEDI	N/A	0.02792131	0.0207	0.0207	0.04432964
ELEVATORS_MEDI	N/A	0.036347927	0	0	0.09962652
ENTRANCES_MEDI	N/A	0.143921136	0.1379	0.1379	0.07152279
FLOORSMAX_MEDI	N/A	0.196035665	0.1667	0.1667	0.10724018
FLOORSMIN_MEDI	N/A	0.215791878	0.2083	0.2083	0.09261477

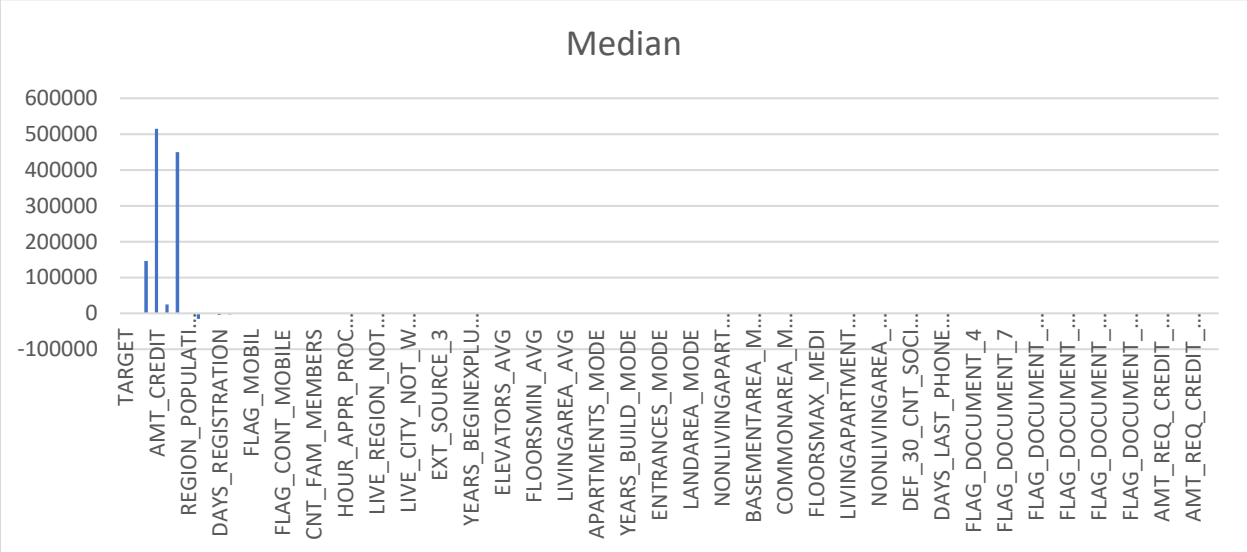
LANDAREA_MEDI	N/A		0.056224766	0.0488	0.0488	0.05192749
LIVINGAPARTMENTS_MEDI	N/A		0.08418763	0.0761	0.0761	0.0542552
LIVINGAREA_MEDI	N/A		0.091821816	0.075	0.075	0.08095229
NONLIVINGAPARTMENTS_MEDI	N/A		0.002732713	0	0	0.02716768
NONLIVINGAREA_MEDI	N/A		0.014403616	0.0031	0.0031	0.04831972
FONDKAPREMONT_MODE	reg oper account	46236	N/A	N/A	reg oper account	N/A
	reg oper spec account	1967				
	not specified	898				
	org spec account	898				
HOUSETYPE_MODE	block of flats	49528	N/A	N/A	block of flats	N/A
	specific housing	255				
	terraced house	216				
TOTALAREA_MODE	N/A		0.05309386	0.0049	0	0.09305016
WALLSMATERIAL_MODE	Panel	36107	N/A	N/A	Panel	N/A
	Stone, brick	10571				
	Block	1505				
	Wooden	896				
	Mixed	347				
	Monolithic	302				
	Others	271				
EMERGENCYSTATE_MODE	No	49642	N/A	N/A	No	N/A
	Yes	357				
OBS_30_CNT_SOCIAL_CIRCLE	N/A		1.41600832	0	0	2.29968521
DEF_30_CNT_SOCIAL_CIRCLE	N/A		0.141342827	0	0	0.43987537
OBS_60_CNT_SOCIAL_CIRCLE	N/A		1.398947979	0	0	2.2793928
DEF_60_CNT_SOCIAL_CIRCLE	N/A		0.09800196	0	0	0.35670842

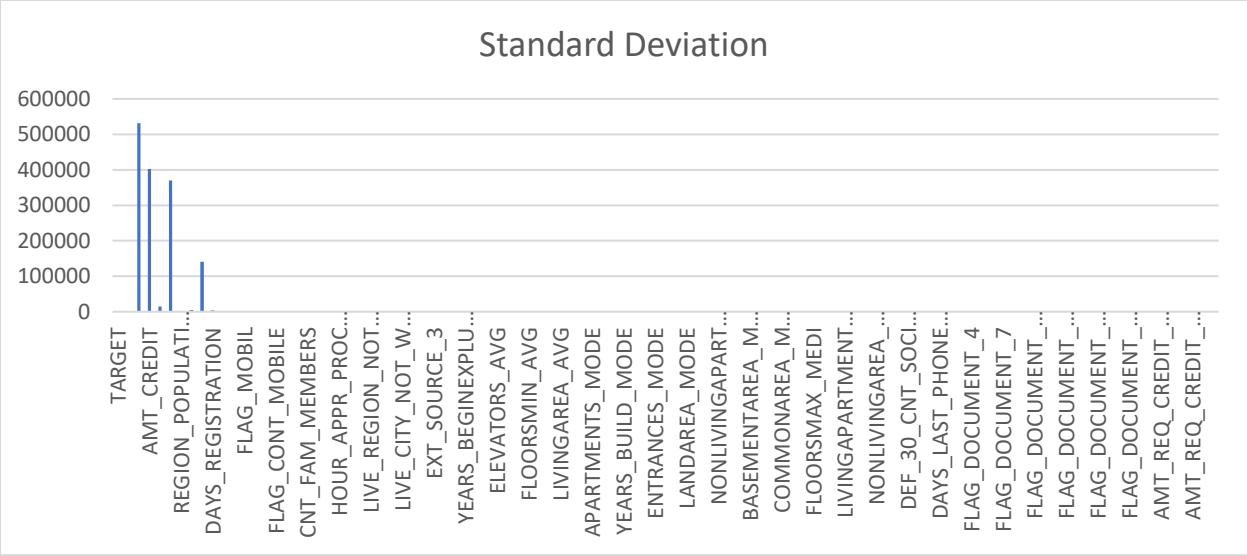
DAYS_LAST_PHONE_CHANGE	N/A	-964.276886	-755	0	829.488489
FLAG_DOCUMENT_2	N/A	4.00008E-05	0	0	0.00632456
FLAG_DOCUMENT_3	N/A	0.712254245	1	1	0.45271651
FLAG_DOCUMENT_4	N/A	0.000180004	0	0	0.01341547
FLAG_DOCUMENT_5	N/A	0.015700314	0	0	0.12431461
FLAG_DOCUMENT_6	N/A	0.086701734	0	0	0.2814003
FLAG_DOCUMENT_7	N/A	0.000220004	0	0	0.01483106
FLAG_DOCUMENT_8	N/A	0.080761615	0	0	0.2724714
FLAG_DOCUMENT_9	N/A	0.003680074	0	0	0.06055249
FLAG_DOCUMENT_10	N/A	2.00004E-05	0	0	0.00447218
FLAG_DOCUMENT_11	N/A	0.004260085	0	0	0.0651308
FLAG_DOCUMENT_12	N/A	0	0	0	0
FLAG_DOCUMENT_13	N/A	0.003220064	0	0	0.05665474
FLAG_DOCUMENT_14	N/A	0.003160063	0	0	0.05612611
FLAG_DOCUMENT_15	N/A	0.000820016	0	0	0.02862447
FLAG_DOCUMENT_16	N/A	0.0100202	0	0	0.09959917
FLAG_DOCUMENT_17	N/A	0.000300006	0	0	0.01731826
FLAG_DOCUMENT_18	N/A	0.008500 17	0	0	0.09180461
FLAG_DOCUMENT_19	N/A	0.000700014	0	0	0.02644878
FLAG_DOCUMENT_20	N/A	0.000520 01	0	0	0.02279803
FLAG_DOCUMENT_21	N/A	0.000380008	0	0	0.01949027
AMT_REQ_CREDIT_BUREAU_HOUR	N/A	0.006140123	0	0	0.08162454
AMT_REQ_CREDIT_BUREAU_DAY	N/A	0.006500 13	0	0	0.10048957

AMT_REQ_CREDIT_BUREAU_WEEK	N/A	0.02802056	0	0	0.18087612
AMT_REQ_CREDIT_BUREAU_MON	N/A	0.233884678	0	0	0.86868247
AMT_REQ_CREDIT_BUREAU_QRT	N/A	0.225824516	0	0	0.57162696
AMT_REQ_CREDIT_BUREAU_YEAR	N/A	1.627692554	1	0	1.84994779

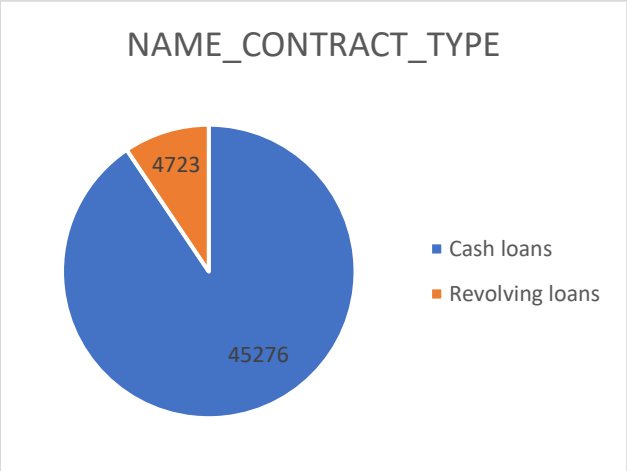
To visualize the **univariate analysis** of the bank loan application dataset, I created ***column charts*** and ***pie charts*** using Excel's chart feature. The column charts below show the mean, median, and standard deviation for each of the numerical variables.

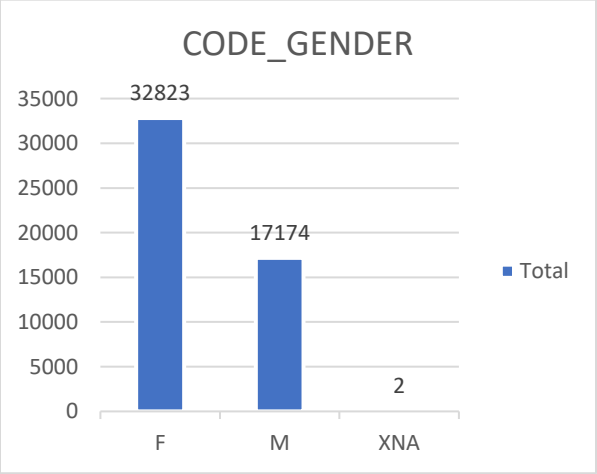


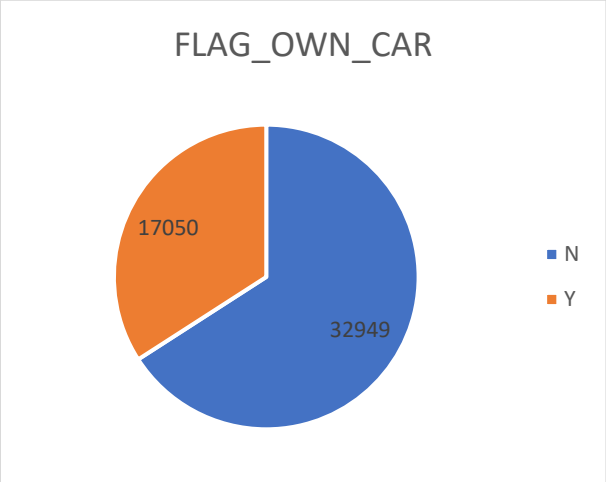


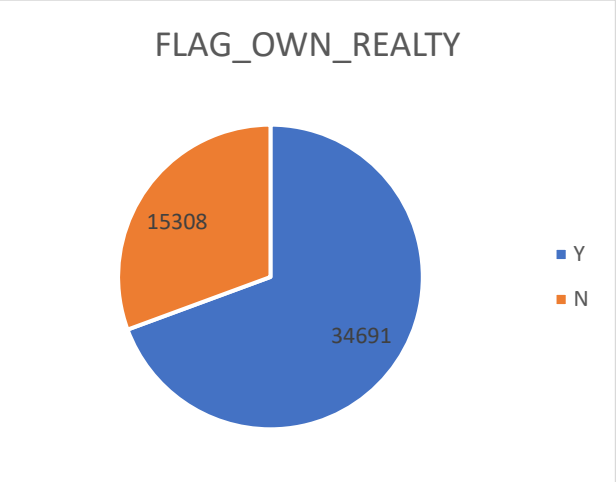


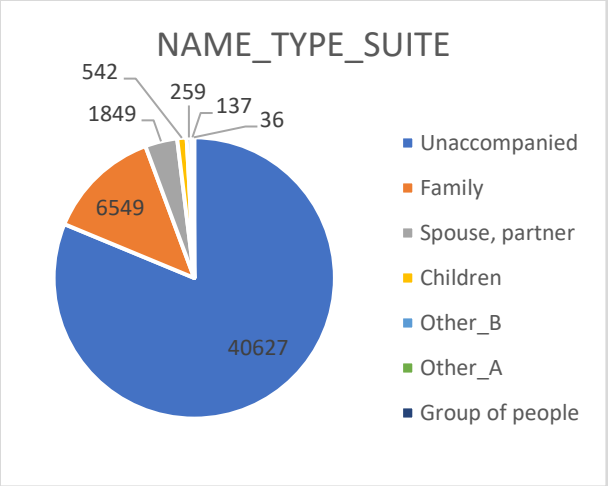
The charts below show the frequency or proportion of each category for each of the categorical variables.

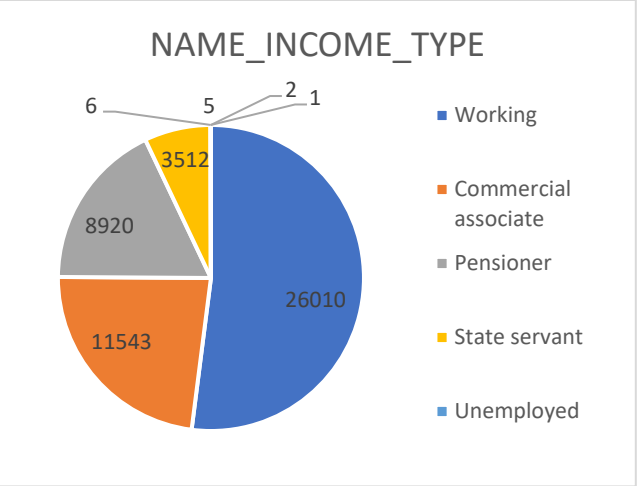


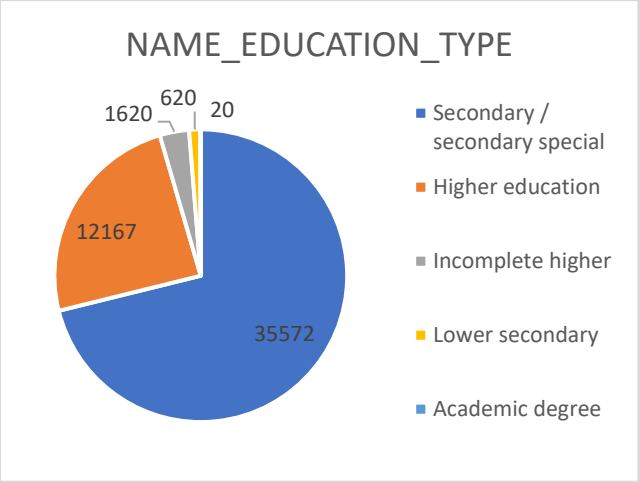


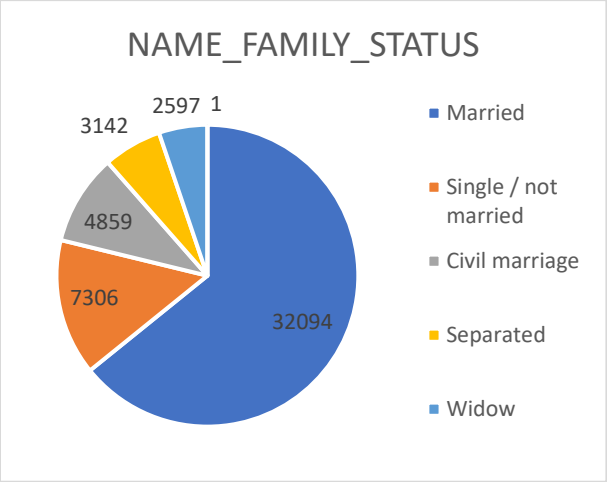


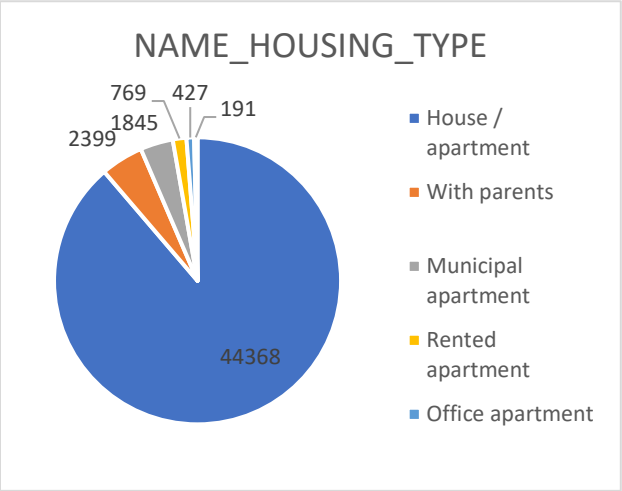


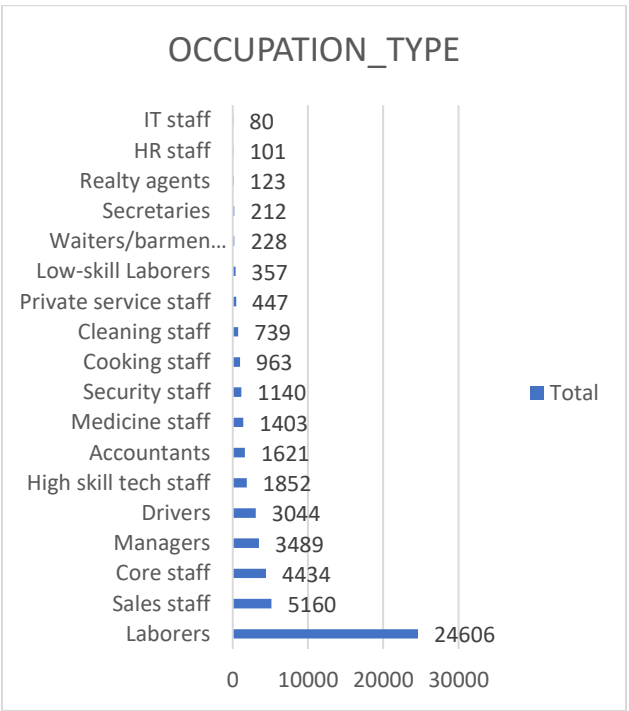

NAME_TYPE_SUITE
542
259
137
1849
36
6549
40627
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people

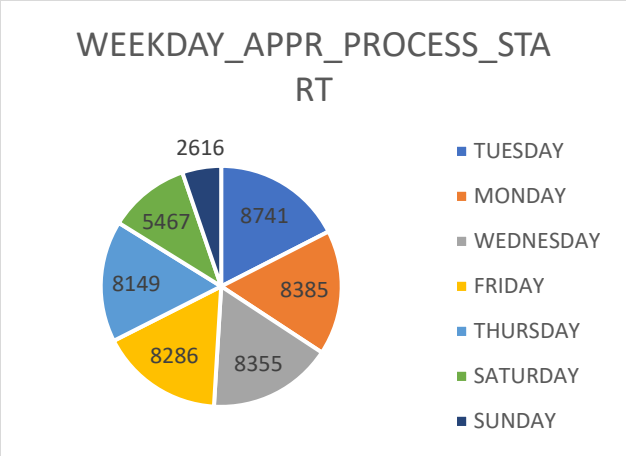

NAME_INCOME_TYPE
6
5
2
1
3512
8920
26010
11543
Working
Commercial associate
Pensioner
State servant
Unemployed

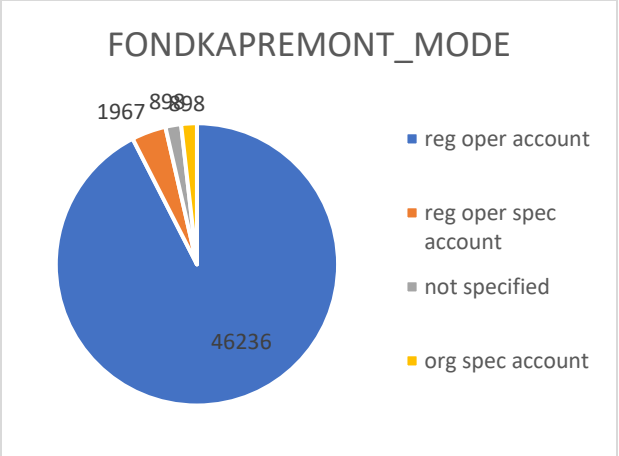

NAME_EDUCATION_TYPE
1620
620
20
12167
35572
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree

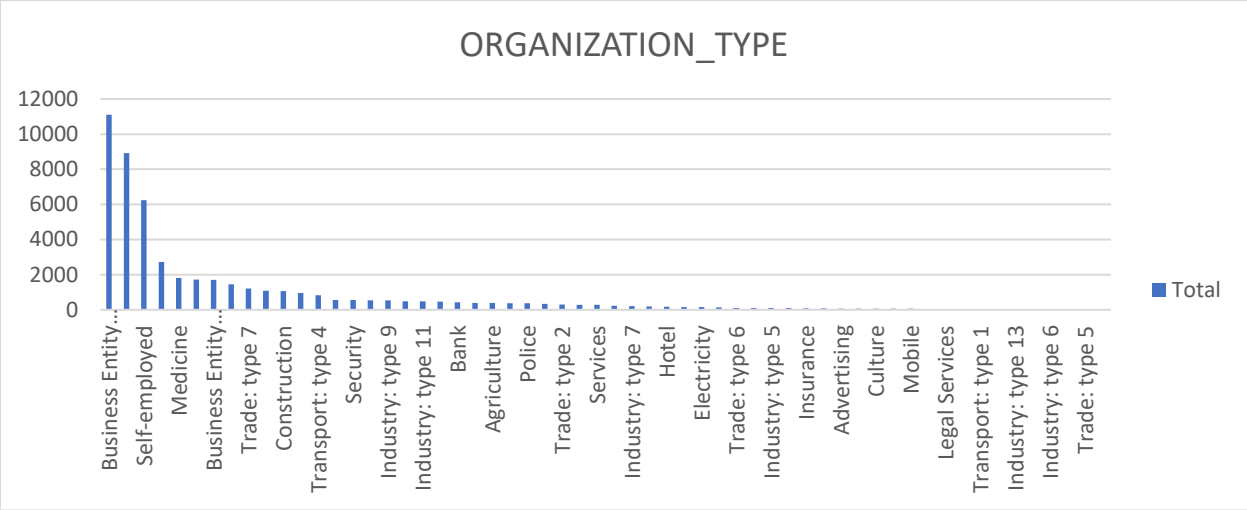

NAME_FAMILY_STATUS
2597
1
3142
4859
7306
32094
Married
Single / not married
Civil marriage
Separated
Widow

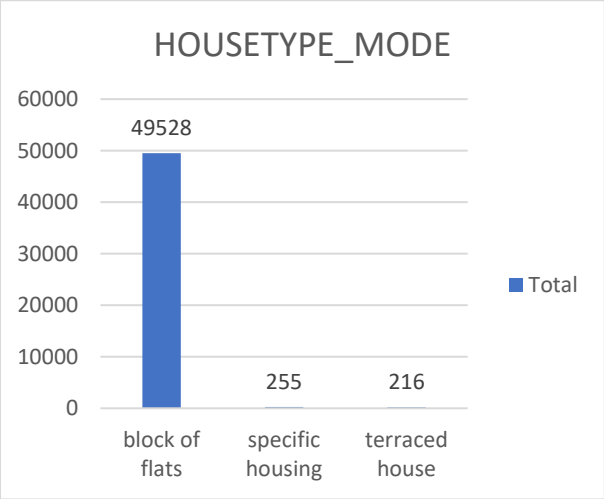

NAME_HOUSING_TYPE
769
427
191
1845
2399
44368
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment

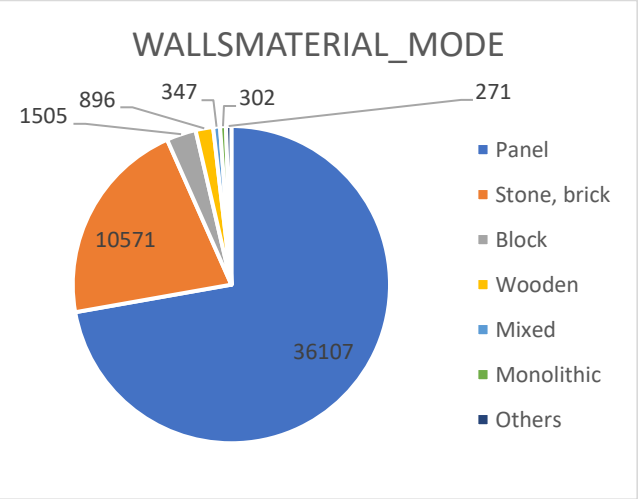

OCCUPATION_TYPE
IT staff 80
HR staff 101
Realty agents 123
Secretaries 212
Waiters/barmen... 228
Low-skill Laborers 357
Private service staff 447
Cleaning staff 739
Cooking staff 963
Security staff 1140
Medicine staff 1403
Accountants 1621
High skill tech staff 1852
Drivers 3044
Managers 3489
Core staff 4434
Sales staff 5160
Laborers 24606
0
10000
20000
30000
Total

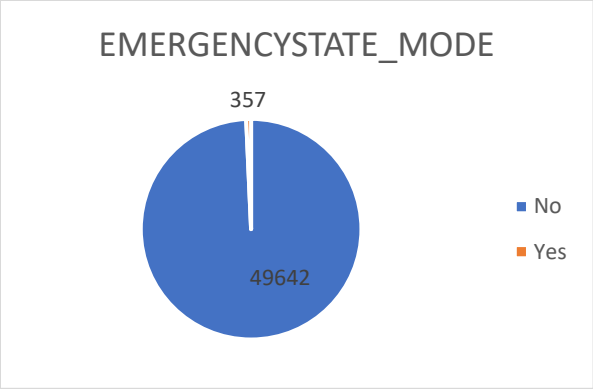

WEEKDAY_APPR_PROCESS_START
2616
8741
8385
8355
8286
8149
5467
TUESDAY
MONDAY
WEDNESDAY
FRIDAY
THURSDAY
SATURDAY
SUNDAY


FONDKAPREMONT_MODE
1967
46236
reg oper account
reg oper spec account
not specified
org spec account


ORGANIZATION_TYPE
12000
10000
8000
6000
4000
2000
0
Business Entity...
Self-employed
Medicine
Business Entity...
Trade: type 7
Construction
Transport: type 4
Security
Industry: type 9
Industry: type 11
Bank
Agriculture
Police
Trade: type 2
Services
Industry: type 7
Hotel
Electricity
Trade: type 6
Industry: type 5
Insurance
Advertising
Culture
Mobile
Legal Services
Transport: type 1
Industry: type 13
Industry: type 6
Trade: type 5
Total


HOUSETYPE_MODE
60000
50000
40000
30000
20000
10000
0
49528
255
216
block of flats
specific housing
terraced house
Total


WALLSMATERIAL_MODE
1505
896
347
302
271
10571
36107
Panel
Stone, brick
Block
Wooden
Mixed
Monolithic
Others



To perform **segmented univariate analysis** on the bank loan application dataset, I used Excel features like filters, sorting, and pivot tables. These features help to filter, sort, or summarize data based on different criteria or categories, which can be used for segmented or bivariate analysis. I applied these features to each column of the dataset and calculated the descriptive statistics for each variable for different scenarios, such as customers with payment difficulties and all other cases.

The results showed that there were insignificant differences in variable distributions for customers with payment difficulties and all other cases. The table below summarizes the descriptive statistics for each variable for different scenarios.

	Mean/Most Common for Payment difficulties	**Mean/Most Common for All other cases**
NAME_CONTRACT_TYPE	Cash loans	Cash loans
CODE_GENDER	F	F
FLAG_OWN_CAR	N	N
FLAG_OWN_REALTY	Y	Y
CNT_CHILDREN	0.419849161	0.419856794
AMT_INCOME_TOTAL	170774.5566	170766.9558
AMT_CREDIT	599686.8653	599704.4437
AMT_ANNUITY	27107.68645	27107.38213
AMT_GOODS_PRICE	538969.9615	538996.1091
NAME_TYPE_SUITE	Unaccompanied	Unaccompanied
NAME_INCOME_TYPE	Working	Working
NAME_EDUCATION_TYPE	Secondary / secondary special	Secondary / secondary special
NAME_FAMILY_STATUS	Married	Married
NAME_HOUSING_TYPE	House / apartment	House / apartment
REGION_POPULATION_RELATIVE	0.020797471	0.020798323
DAYS_BIRTH	-16022.06854	-16022.17331
DAYS_EMPLOYED	63220.40873	63220.70167
DAYS_REGISTRATION	-4977.15196	-4977.309252
DAYS_ID_PUBLISH	-2996.73025	-2996.814713

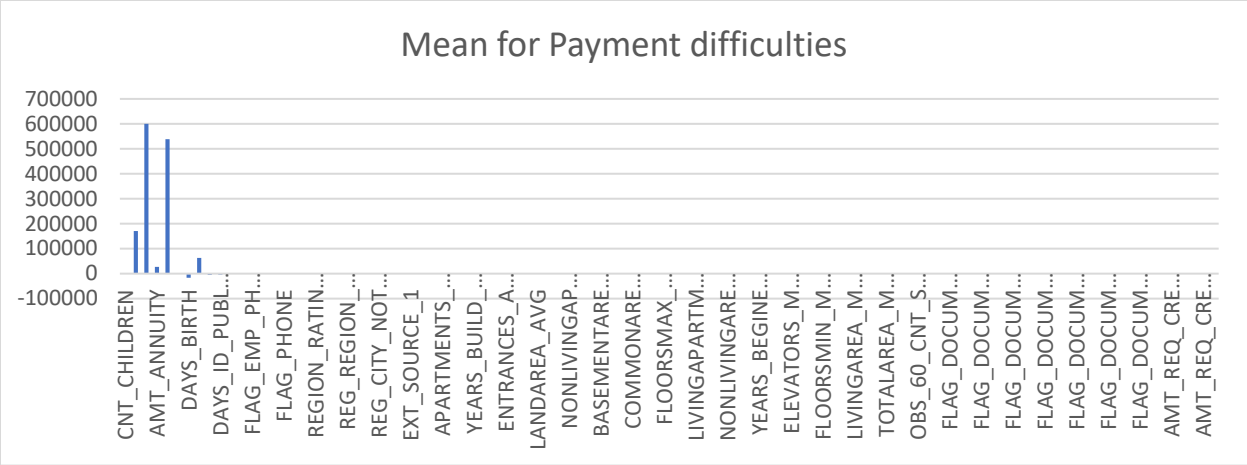
OWN_CAR_AGE	4.099725929	4.100704028
FLAG_MOBIL	0.999979995	0.999979999
FLAG_EMP_PHONE	0.821473583	0.821472859
FLAG_WORK_PHONE	0.199271811	0.199267971
FLAG_CONT_MOBILE	0.997979475	0.997979919
FLAG_PHONE	0.277772221	0.277711108
FLAG_EMAIL	0.05565447	0.055662226
OCCUPATION_TYPE	Laborers	Laborers
CNT_FAM_MEMBERS	2.15896133	2.158966359
REGION_RATING_CLIENT	2.051673435	2.051662066
REGION_RATING_CLIENT_W_CITY	2.030727989	2.030721229
WEEKDAY_APPR_PROCESS_START	TUESDAY	TUESDAY
HOUR_APPR_PROCESS_START	12.05275372	12.0526021
REG_REGION_NOT_LIVE_REGION	0.015003901	0.0150006
REG_REGION_NOT_WORK_REGION	0.049932983	0.049921997
LIVE_REGION_NOT_WORK_REGION	0.039650309	0.039641586
REG_CITY_NOT_LIVE_CITY	0.079980795	0.079963199
REG_CITY_NOT_WORK_CITY	0.232180367	0.232169287
LIVE_CITY_NOT_WORK_CITY	0.179706724	0.179707188
ORGANIZATION_TYPE	Business Entity Type 3	Business Entity Type 3
EXT_SOURCE_1	0.502264458	0.502265631
EXT_SOURCE_2	0.513823314	0.513828612
EXT_SOURCE_3	0.511887608	0.511888858
APARTMENTS_AVG	0.117766987	0.11777324
BASEMENTAREA_AVG	0.088942114	0.088946757
YEARS_BEGINEXPLUATATION_AVG	0.97803533	0.97803595
YEARS_BUILD_AVG	0.75163894	0.751641336
COMMONAREA_AVG	0.044800134	0.044797099
ELEVATORS_AVG	0.078670427	0.078679388
ENTRANCES_AVG	0.150547114	0.150552244
FLOORSMAX_AVG	0.225471625	0.225469644
FLOORSMIN_AVG	0.231654963	0.231651795
LANDAREA_AVG	0.066345885	0.066352611
LIVINGAPARTMENTS_AVG	0.100440638	0.100437178
LIVINGAREA_AVG	0.107674885	0.107691308
NONLIVINGAPARTMENTS_AVG	0.009096957	0.009096819

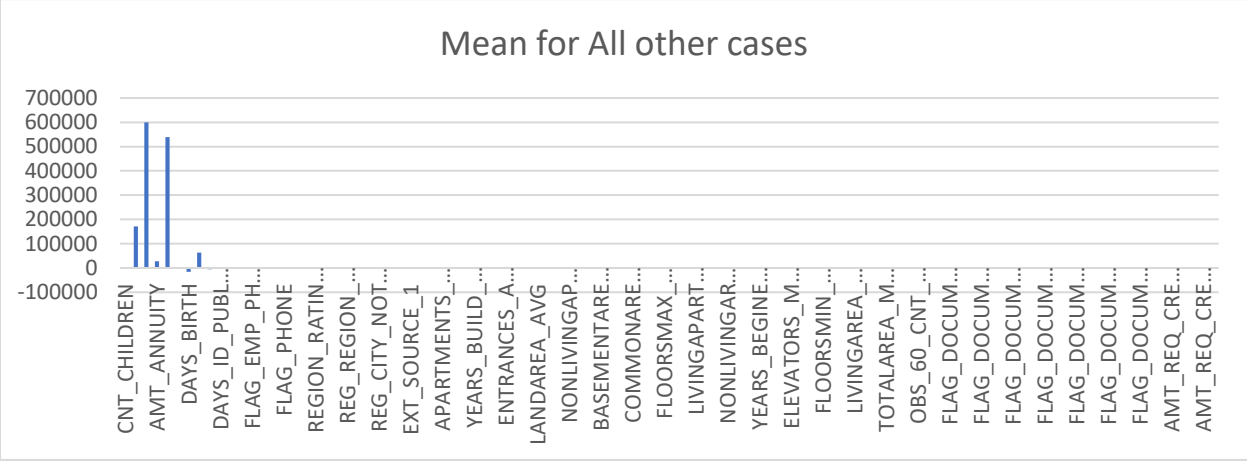
NONLIVINGAREA_AVG	0.028295248	0.028294372
APARTMENTS_MODE	0.088323452	0.088326455
BASEMENTAREA_MODE	0.036608086	0.036613613
YEARS_BEGINEXPLUATATIO N_MODE	0.982134189	0.982134633
YEARS_BUILD_MODE	0.806258445	0.806259118
COMMONAREA_MODE	0.012834005	0.012832405
ELEVATORS_MODE	0.034624566	0.034629837
ENTRANCES_MODE	0.141915134	0.141919083
FLOORSMAX_MODE	0.194236531	0.194232971
FLOORSMIN_MODE	0.214734189	0.214731107
LANDAREA_MODE	0.026294307	0.02629815
LIVINGAPARTMENTS_MODE	0.070948985	0.070944782
LIVINGAREA_MODE	0.052750061	0.052764053
NONLIVINGAPARTMENTS_ MODE	0.0025138	0.002514025
NONLIVINGAREA_MODE	0.01216255	0.012162144
APARTMENTS_MEDI	0.102558727	0.102565359
BASEMENTAREA_MEDI	0.081031456	0.081036461
YEARS_BEGINEXPLUATATIO N_MEDI	0.979771715	0.979772237
YEARS_BUILD_MEDI	0.757310925	0.757313197
COMMONAREA_MEDI	0.02792383	0.027921581
ELEVATORS_MEDI	0.036339448	0.036348654
ENTRANCES_MEDI	0.143917058	0.143922635
FLOORSMAX_MEDI	0.196038958	0.19603792
FLOORSMIN_MEDI	0.215797009	0.215793694
LANDAREA_MEDI	0.056218197	0.056225141
LIVINGAPARTMENTS_MEDI	0.084192222	0.084188904
LIVINGAREA_MEDI	0.091805557	0.091823267
NONLIVINGAPARTMENTS_ MEDI	0.002732592	0.002732767
NONLIVINGAREA_MEDI	0.014403753	0.014403904
FONDKAPREMONT_MODE	reg oper account	reg oper account
HOUSETYPE_MODE	block of flats	block of flats
TOTALAREA_MODE	0.053074325	0.053094624
WALLSMATERIAL_MODE	Panel	Panel
EMERGENCYSTATE_MODE	No	No
OBS_30_CNT_SOCIAL_CIRCL E	1.416148199	1.41599664
DEF_30_CNT_SOCIAL_CIRCL E	0.141376758	0.141305652
OBS_60_CNT_SOCIAL_CIRCL E	1.399083762	1.398935957

DEF_60_CNT_SOCIAL_CIRCLE	0.098025487	0.097963919
DAYS_LAST_PHONE_CHANGE	-964.2804929	-964.2734909
FLAG_DOCUMENT_2	4.00104E-05	4.00016E-05
FLAG_DOCUMENT_3	0.712285194	0.71224849
FLAG_DOCUMENT_4	0.000180047	0.000180007
FLAG_DOCUMENT_5	0.015704083	0.015700628
FLAG_DOCUMENT_6	0.086722548	0.086703468
FLAG_DOCUMENT_7	0.000220057	0.000220009
FLAG_DOCUMENT_8	0.080760998	0.080763231
FLAG_DOCUMENT_9	0.003680957	0.003680147
FLAG_DOCUMENT_10	2.00052E-05	2.00008E-05
FLAG_DOCUMENT_11	0.004261108	0.00426017
FLAG_DOCUMENT_12	0	0
FLAG_DOCUMENT_13	0.003220837	0.003220129
FLAG_DOCUMENT_14	0.003160822	0.003160126
FLAG_DOCUMENT_15	0.000820213	0.000820033
FLAG_DOCUMENT_16	0.010022606	0.010020401
FLAG_DOCUMENT_17	0.000300078	0.000300012
FLAG_DOCUMENT_18	0.008482205	0.00850034
FLAG_DOCUMENT_19	0.000700182	0.000700028
FLAG_DOCUMENT_20	0.000520135	0.000520021
FLAG_DOCUMENT_21	0.000380099	0.000380015
AMT_REQ_CREDIT_BUREAU_HOUR	0.006141597	0.006140246
AMT_REQ_CREDIT_BUREAU_DAY	0.00650169	0.00650026
AMT_REQ_CREDIT_BUREAU_WEEK	0.028027287	0.028021121
AMT_REQ_CREDIT_BUREAU_MON	0.233940825	0.233889356
AMT_REQ_CREDIT_BUREAU_QRT	0.225758697	0.225829033
AMT_REQ_CREDIT_BUREAU_YEAR	1.627723208	1.627705108

To visualize the segmented univariate analysis of the bank loan application dataset, I created ***stacked column charts*** using Excel's chart feature. Stacked column charts are graphical representations of the frequency or proportion of each category in a categorical variable for different segments using stacked bars of different colors.

The stacked column charts below show the mean distribution of each of the variables for different scenarios.

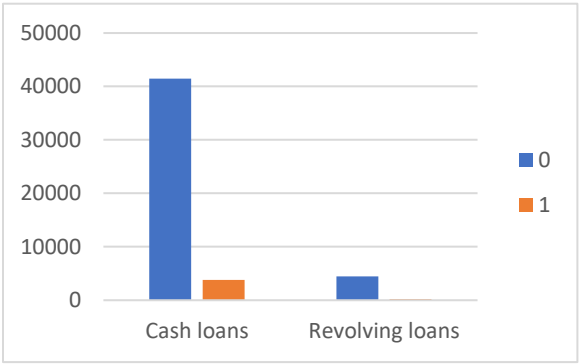




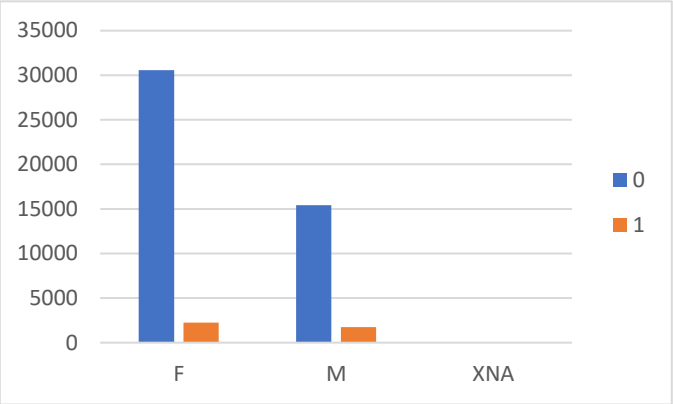
To perform **bivariate analysis** on the bank loan application dataset, I used Excel features like ***filters, sorting,*** and ***pivot tables***. These features help to filter, sort, or summarize data based on different criteria or categories, which can be used for bivariate analysis. I applied these features to each column of the dataset and calculated the frequency distribution of each variable for different scenarios, such as customers with payment difficulties and all other cases.

The table below summarizes the frequency distribution for each variable for different scenarios.

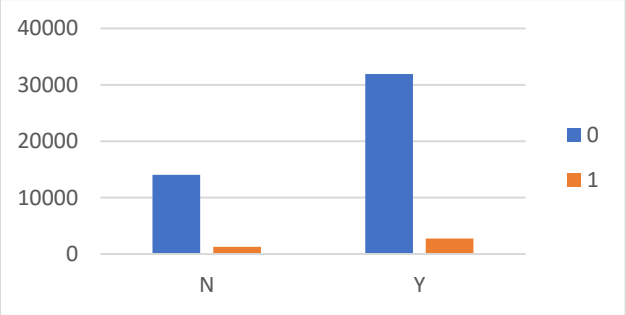
Count of NAME_CONTRACT_TYPE	**Column Labels**		
Row Labels	**0**	**1**	**Grand Total**
Cash loans	41484	3792	45276
Revolving loans	4489	234	4723
Grand Total	**45973**	**4026**	**49999**



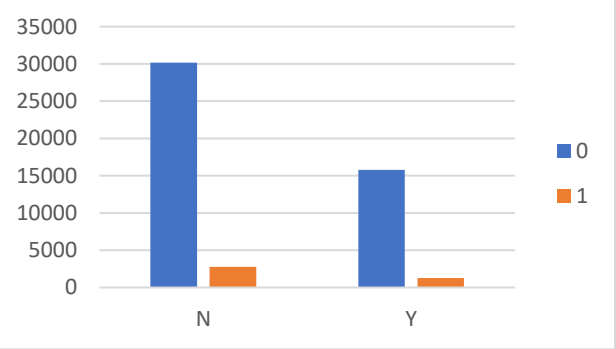
Count of CODE_GENDER	Column Labels		
Row Labels	0	1	Grand Total
F	30559	2264	32823
M	15412	1762	17174
XNA	2		2
Grand Total	**45973**	**4026**	**49999**



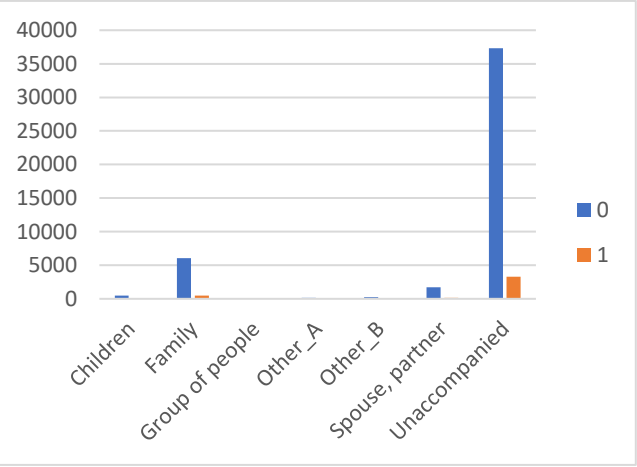
Count of FLAG_OWN_CAR	Column Labels		
Row Labels	0	1	Grand Total
N	30176	2773	32949
Y	15797	1253	17050
Grand Total	**45973**	**4026**	**49999**



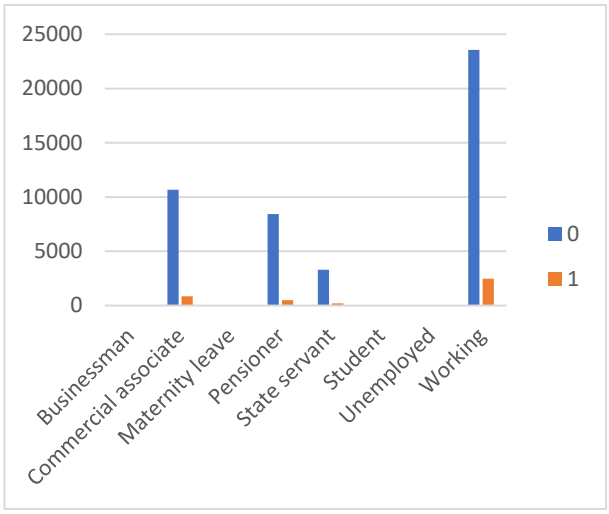
Count of FLAG_OWN_REALTY	Column Labels		
Row Labels	0	1	Grand Total
N	14034	1274	15308
Y	31939	2752	34691
Grand Total	**45973**	**4026**	**49999**



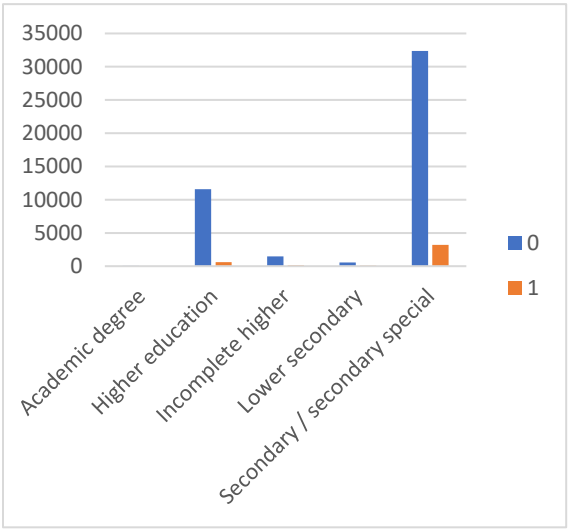
Count of NAME_TYPE_SUITE	Column Labels		
Row Labels	0	1	Grand Total
Children	495	47	542
Family	6050	499	6549
Group of people	35	1	36
Other_A	127	10	137
Other_B	231	28	259
Spouse, partner	1705	144	1849
Unaccompanied	37330	3297	40627
Grand Total	**45973**	**4026**	**49999**



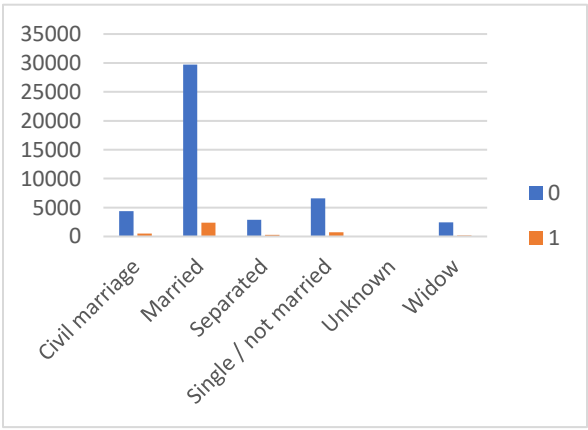
Count of NAME_INCOME_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Businessman	2		2
Commercial associate	10679	864	11543
Maternity leave	1		1
Pensioner	8419	501	8920
State servant	3314	198	3512
Student	5		5
Unemployed	4	2	6
Working	23549	2461	26010
Grand Total	**45973**	**4026**	**49999**



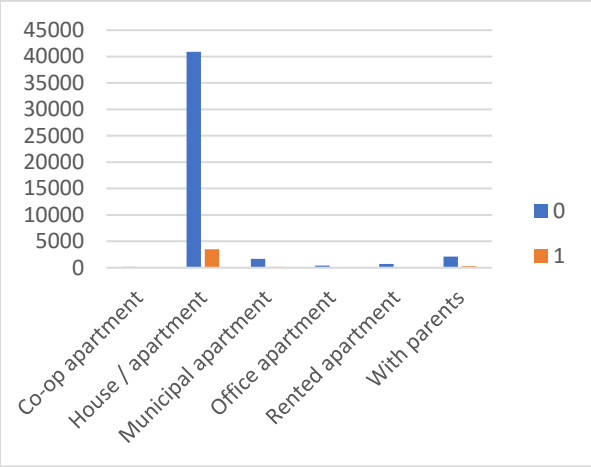
Count of NAME_EDUCATION_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Academic degree	20		20
Higher education	11561	606	12167
Incomplete higher	1482	138	1620
Lower secondary	547	73	620
Secondary / secondary special	32363	3209	35572
Grand Total	**45973**	**4026**	**49999**



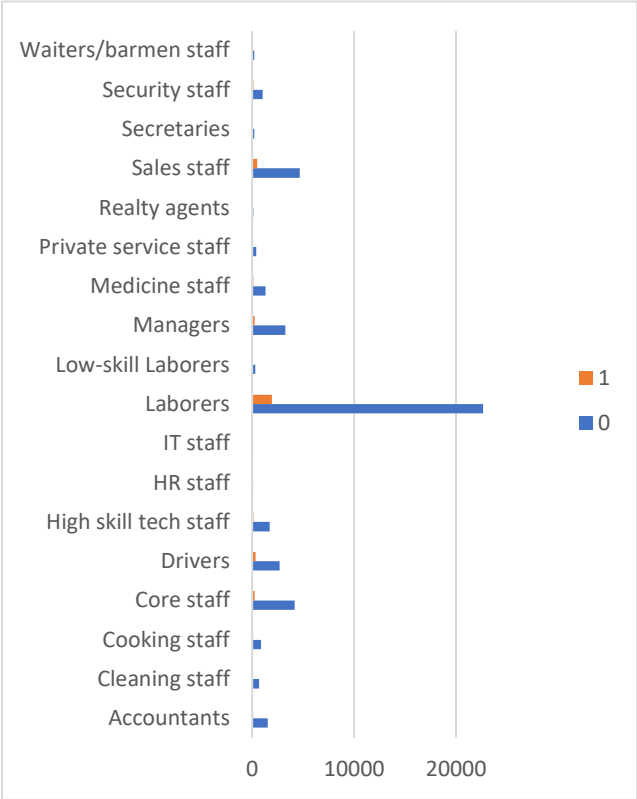
Count of NAME_FAMILY_STATUS	Column Labels		
Row Labels	0	1	Grand Total
Civil marriage	4377	482	4859
Married	29699	2395	32094
Separated	2870	272	3142
Single / not married	6577	729	7306
Unknown	1		1
Widow	2449	148	2597
Grand Total	**45973**	**4026**	**49999**



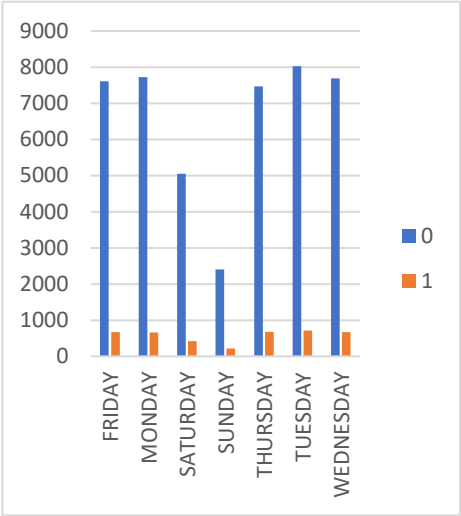
Count of NAME_HOUSING_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Co-op apartment	176	15	191
House / apartment	40895	3473	44368
Municipal apartment	1700	145	1845
Office apartment	398	29	427
Rented apartment	682	87	769
With parents	2122	277	2399
Grand Total	**45973**	**4026**	**49999**



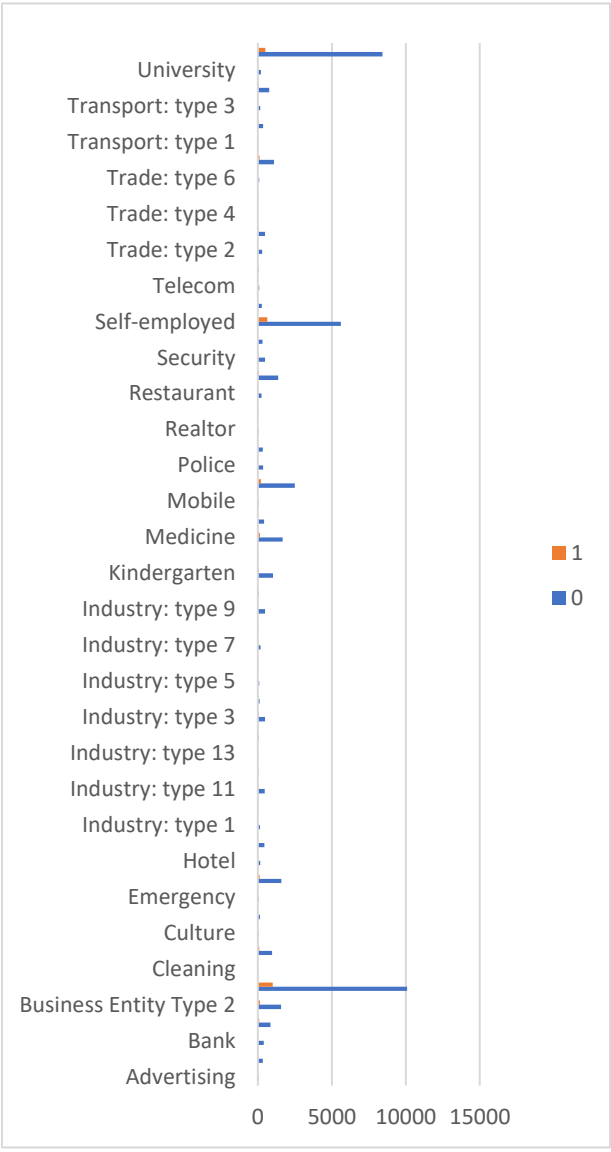
Count of OCCUPATION_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Accountants	1540	81	1621
Cleaning staff	671	68	739
Cooking staff	862	101	963
Core staff	4184	250	4434
Drivers	2706	338	3044
High skill tech staff	1734	118	1852
HR staff	92	9	101
IT staff	76	4	80
Laborers	22660	1946	24606
Low-skill Laborers	296	61	357
Managers	3246	243	3489
Medicine staff	1297	106	1403
Private service staff	410	37	447
Realty agents	110	13	123
Sales staff	4668	492	5160
Secretaries	203	9	212
Security staff	1015	125	1140
Waiters/barmen staff	203	25	228
Grand Total	**45973**	**4026**	**49999**



Count of WEEKDAY_APPR_PROCESS_START	Column Labels		
Row Labels	0	1	Grand Total
FRIDAY	7615	671	8286
MONDAY	7728	657	8385
SATURDAY	5048	419	5467
SUNDAY	2401	215	2616
THURSDAY	7471	678	8149
TUESDAY	8024	717	8741
WEDNESDAY	7686	669	8355
Grand Total	**45973**	**4026**	**49999**

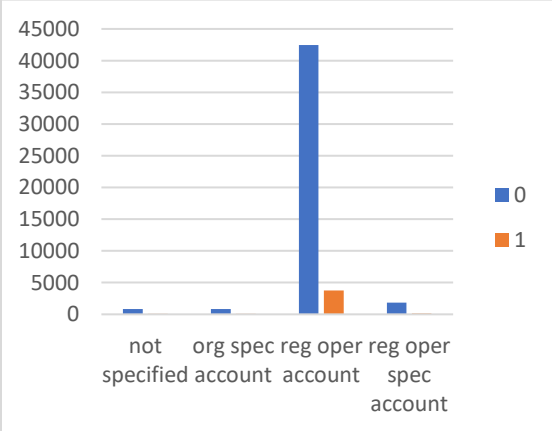


Count of ORGANIZATION_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Advertising	61	7	68
Agriculture	341	51	392
Bank	408	27	435
Business Entity Type 1	865	88	953
Business Entity Type 2	1571	133	1704
Business Entity Type 3	10087	1014	11101
Cleaning	37	3	40
Construction	958	108	1066
Culture	62	2	64
Electricity	134	13	147
Emergency	86	7	93
Government	1592	124	1716
Hotel	169	13	182
Housing	447	42	489
Industry: type 1	140	19	159
Industry: type 10	20	1	21
Industry: type 11	461	28	489
Industry: type 12	50	3	53
Industry: type 13	11	4	15
Industry: type 2	68	10	78
Industry: type 3	491	51	542
Industry: type 4	125	15	140
Industry: type 5	96	7	103
Industry: type 6	11	1	12
Industry: type 7	190	19	209

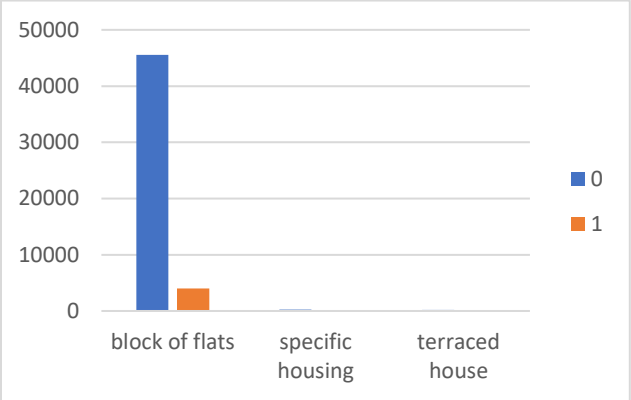


Industry: type 8	8		8
Industry: type 9	496	41	537
Insurance	82	7	89
Kindergarten	1024	66	1090
Legal Services	40	4	44
Medicine	1687	130	1817
Military	432	26	458
Mobile	52	4	56
Other	2509	208	2717
Police	348	18	366
Postal	343	27	370
Realtor	54	7	61
Religion	13	1	14
Restaurant	257	32	289
School	1372	78	1450
Security	488	62	550
Security Ministries	315	16	331
Self-employed	5612	628	6240
Services	260	24	284
Telecom	98	8	106
Trade: type 1	61	5	66
Trade: type 2	286	21	307
Trade: type 3	490	60	550
Trade: type 4	8		8
Trade: type 5	7	1	8
Trade: type 6	105	3	108
Trade: type 7	1090	120	1210
Transport: type 1	26	2	28
Transport: type 2	359	33	392
Transport: type 3	166	25	191
Transport: type 4	770	67	837
University	213	9	222
XNA	8421	503	8924
Grand Total	**45973**	**4026**	**49999**

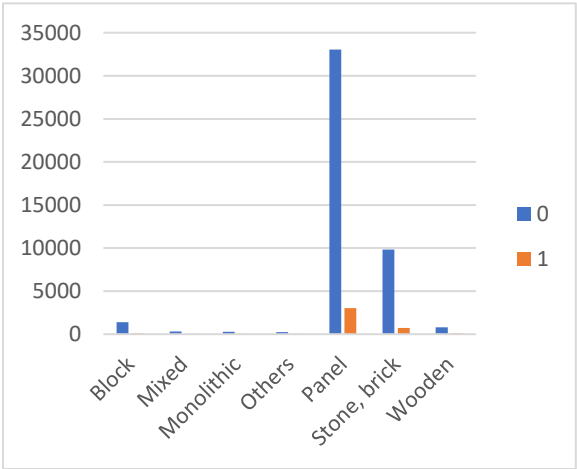
Count of FONDKAPREMONT_MODE	**Column Labels**		
Row Labels	**0**	**1**	**Grand Total**
not specified	839	59	898
org spec account	840	58	898
reg oper account	42454	3782	46236
reg oper spec account	1840	127	1967



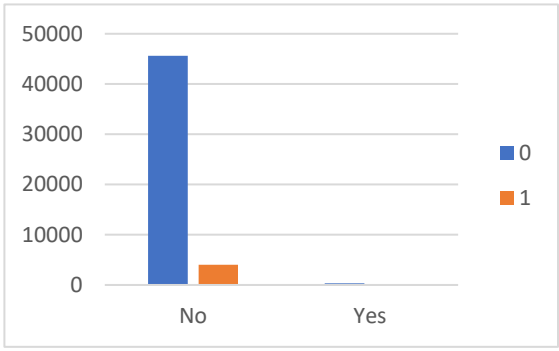
Count of HOUSETYPE_MODE	Column Labels		
Row Labels	0	1	Grand Total
block of flats	45552	3976	49528
specific housing	224	31	255
terraced house	197	19	216
Grand Total	**45973**	**4026**	**49999**



Count of WALLSMATERIAL_MODE	Column Labels		
Row Labels	0	1	Grand Total
Block	1406	99	1505
Mixed	323	24	347
Monolithic	291	11	302
Others	244	27	271
Panel	33062	3045	36107
Stone, brick	9832	739	10571
Wooden	815	81	896
Grand Total	**45973**	**4026**	**49999**



Count of EMERGENCYSTATE_MODE	Column Labels		
Row Labels	0	1	Grand Total
No	45645	3997	49642
Yes	328	29	357
Grand Total	**45973**	**4026**	**49999**



5. Identify Top Correlations for Different Scenarios:

One of the final steps in data analysis is to **identify top correlations for different scenarios**. Top correlations are the strongest and most significant associations between variables and the target variable within each segment. They help to identify the most influential factors of loan default for different groups of customers, such as customers with payment difficulties and all other cases.

To identify top correlations for different scenarios in the bank loan application dataset, I used Excel functions like ***CORREL*** and ***pivot tables***. These functions help to calculate the correlation coefficient

between two ranges of cells, which can be used for bivariate analysis or identifying top correlations. Pivot tables help to segment the data based on different scenarios, such as customers with payment difficulties and all other cases.

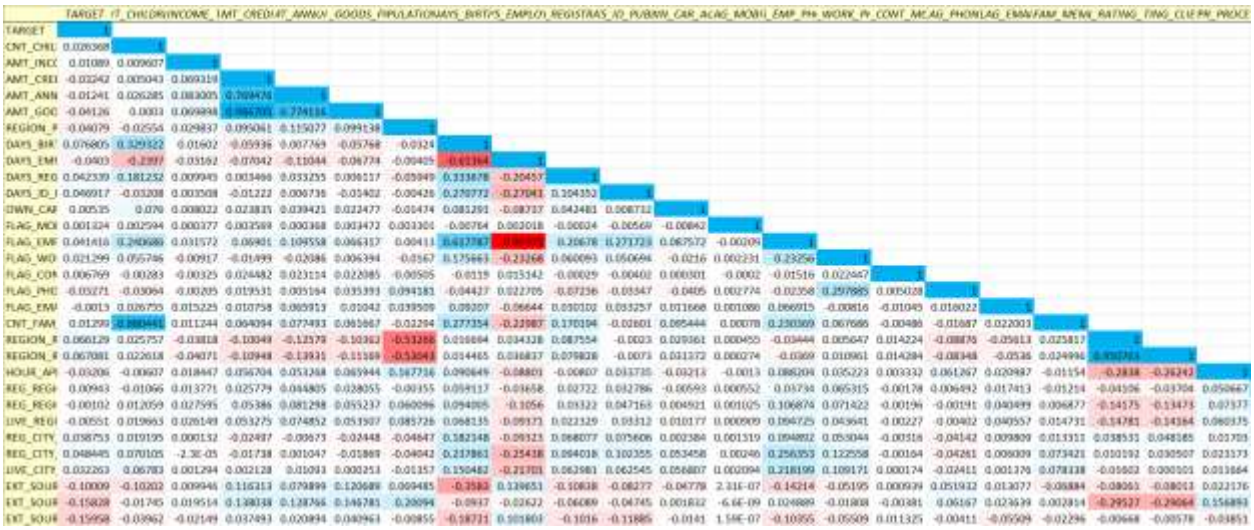
I applied these functions to each column of the dataset and calculated the correlation coefficients between variables and the target variable within each segment. I ranked the correlations to identify the top indicators of loan default for each scenario. The table below summarizes the top 10 correlations for each scenario.

	Customers with payment difficulties
DAYS_BIRTH	0.07681
REGION_RATING_CLIENT_W_CITY	0.06708
REGION_RATING_CLIENT	0.06613
DAYS_LAST_PHONE_CHANGE	0.05614
REG_CITY_NOT_WORK_CITY	0.04844
DAYS_ID_PUBLISH	0.04692
FLAG_DOCUMENT_3	0.04504
DEF_60_CNT_SOCIAL_CIRCLE	0.04438
DAYS_REGISTRATION	0.04234
DEF_30_CNT_SOCIAL_CIRCLE	0.04175

	All other cases
DAYS_BIRTH	0.07679
REGION_RATING_CLIENT_W_CITY	0.06708
REGION_RATING_CLIENT	0.06613
DAYS_LAST_PHONE_CHANGE	0.05613
REG_CITY_NOT_WORK_CITY	0.04845
DAYS_ID_PUBLISH	0.04693
FLAG_DOCUMENT_3	0.04505
DEF_60_CNT_SOCIAL_CIRCLE	0.04440
DAYS_REGISTRATION	0.04234
DEF_30_CNT_SOCIAL_CIRCLE	0.04177

To visualize the top correlations for different scenarios in the bank loan application dataset, I created correlation matrices and heatmaps using Excel's chart feature. Correlation matrices are graphical representations of the correlation coefficients between variables using a table format. Heatmaps are graphical representations of the correlation coefficients between variables using colors to indicate the intensity or magnitude of the association.

The correlation matrix below shows the correlation coefficients between variables for ***customers with payment difficulties***.



The correlation matrix below shows the correlation coefficients between variables for ***all other cases***.

By identifying top correlations for different scenarios in the bank loan application dataset, I was able to discover the most important factors of loan default for different groups of customers. Top correlations help to highlight the key differences or similarities between segments and provide insights into their behavior and preferences.

Result

The result of this project was a comprehensive **EDA** of bank loan applications using Microsoft Excel 365. The EDA helped me to gain insights into the factors of loan default and provide recommendations to the company. The EDA also showcased my skills and knowledge in using Microsoft Excel 365 as a versatile tool for data analysis and visualization.

The main findings and achievements of this project are:

- I identified and dealt with ***missing data*** and ***outliers*** using Excel functions and features. I imputed the missing values using the median and applied thresholds to identify the outliers. I created bar charts, box plots, and scatter plots to visualize the data quality.
- I analyzed ***data imbalance*** using Excel functions. I calculated the ratio of data imbalance by comparing the class frequencies of the target variable. I found that there was a significant data imbalance in the dataset. I created a pie chart and a bar chart to visualize the data distribution.
- I performed ***univariate, segmented univariate, and bivariate analysis*** using Excel functions and features. I calculated the descriptive statistics and correlation coefficients for each variable and scenario. I also found that there were significant differences and relationships between variables and scenarios. I created histograms, bar charts, box plots, stacked bar charts, grouped bar charts, and scatter plots, to visualize the data characteristics and associations.
- I identified ***top correlations*** for different scenarios using Excel functions. I segmented the data based on different scenarios and calculated the correlation coefficients within each segment. I ranked the correlations to identify the top indicators of loan default for each scenario. I created correlation matrices and heatmaps to visualize the top correlations.

By following these steps, I was able to perform a comprehensive EDA of bank loan applications using Microsoft Excel 365 as a powerful tool for data analysis and visualization. The EDA helped me to understand the Bank Loan Case Study and provide insights to the company.

Excel Sheet File