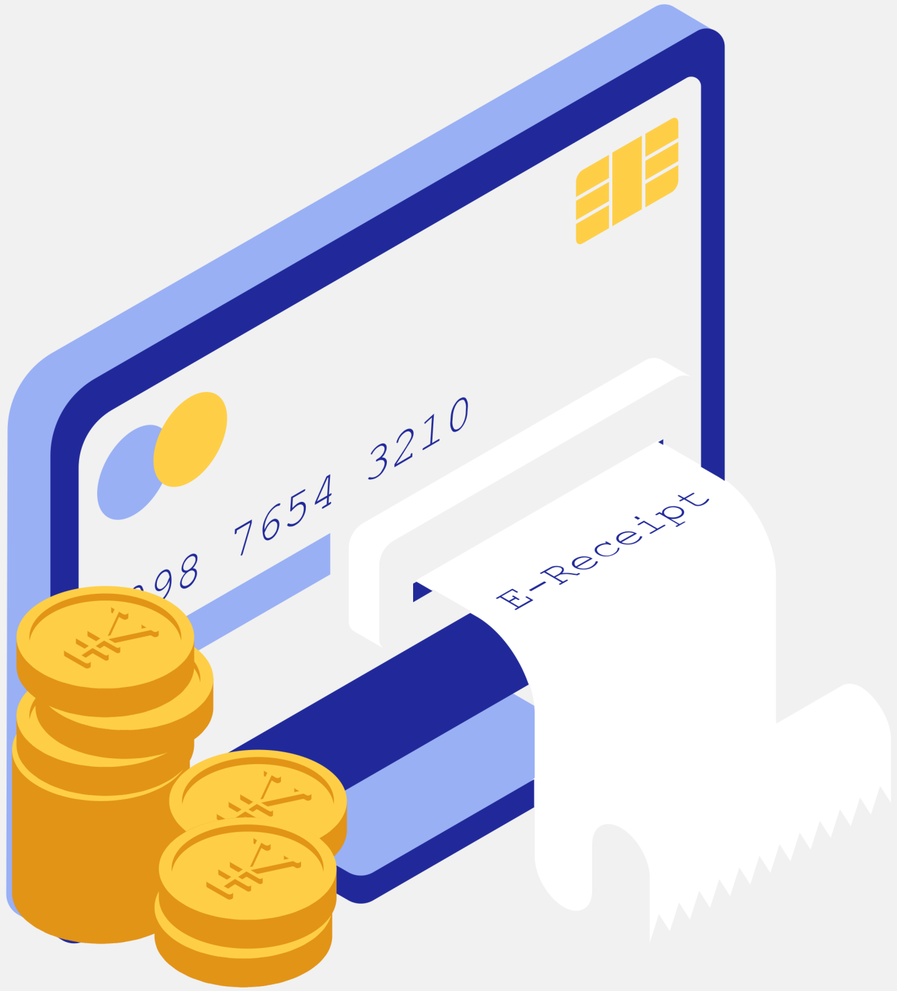


Exploratory Data Analysis

Bank Loan Risk Analytics

Name: Dang Thi Truc
upGrad & IIITB | Data Science Program (Global) - C11

Agenda

Exploratory Data Analysis

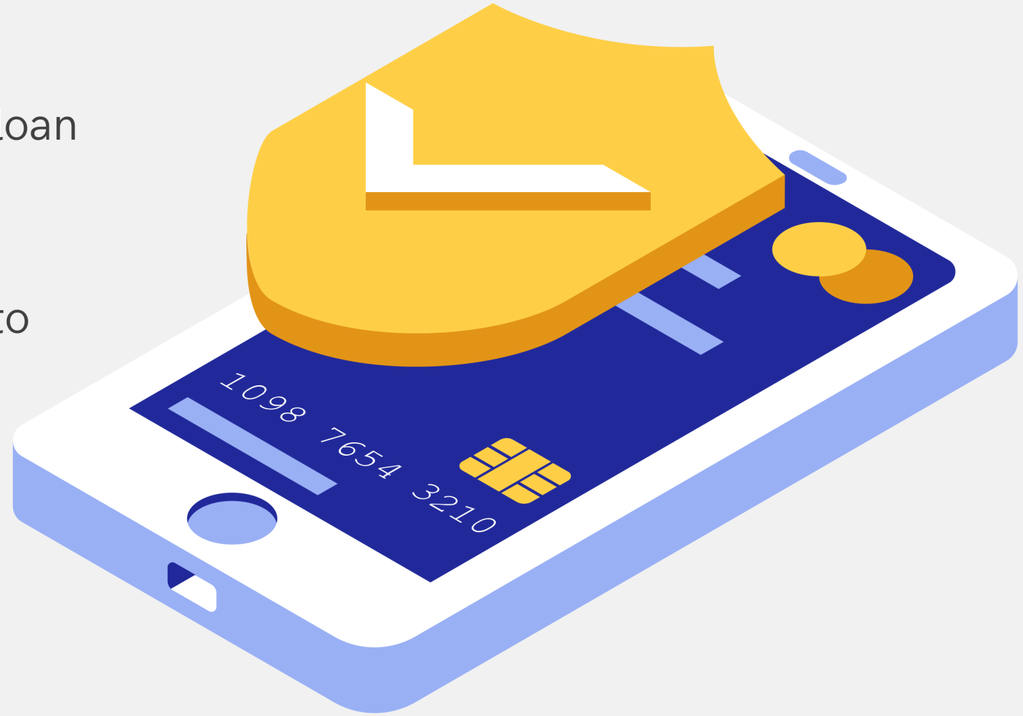
- Business Problems and Objectives
- Dataset Understanding
- Data Cleaning and Handling
- Exploratory Data Analysis
- Conclusion and Suggestion

Business Problem

When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.

- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.



Objective

EDA the dataset to analysis the patterns in the data, so that ensure the applicants capable of repaying the loan are not rejected.

Data Understanding

Application Dataset

- The data given below contains the information about the loan application at the time of applying for the loan.
- Shape: 307511 rows and 122 columns
- Target variable: TARGET column two types of scenarios: The client with payment difficulties and All other cases (Paid on time)

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0

Previous Dataset

- - Previous_application file contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- Shape: 1670214 rows and 37 columns
- Target variable: NAME_CONTRACT_STATUS contains four types of decisions that could be taken by the client/company): Approved, Cancelled, Refused, Unused offer.

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0

Data Cleanning and Handling

✓	Step 1	Drop columns containing many missing values
✓	Step 2	Drop unimportance columns
✓	Step 3	Data profiling before processing
✓	Step 4	Handling missing values
✓	Step 5	Identifying outliers
✓	Step 6	Standardlising values

Drop unnecessary columns

- Drop all columns that contain more than 40% missing values on the total record of application and previous application dataset.
- Define that some normalized columns have the same characteristics as the removed columns, so we can also drop them from the dataset. EXT_SOURCE_2, EXT_SOURCE_3,REGION_POPULATION_RELATIVE are dropped from the dataset.

OWN_CAR_AGE	65.990810
EXT_SOURCE_1	56.381073
APARTMENTS_AVG	50.749729
BASEMENTAREA_AVG	58.515956
YEARS_BEGINEXPLUATATION_AVG	48.781019
YEARS_BUILD_AVG	66.497784
COMMONAREA_AVG	69.872297
ELEVATORS_AVG	53.295980
ENTRANCES_AVG	50.348768
FLOORSMAX_AVG	49.760822
FLOORSMIN_AVG	67.848630
LANDAREA_AVG	59.376738
LIVINGAPARTMENTS_AVG	68.354953
LIVINGAREA_AVG	50.193326
NONLIVINGAPARTMENTS_AVG	69.432963
NONLIVINGAREA_AVG	55.179164
APARTMENTS_MODE	50.749729
BASEMENTAREA_MODE	58.515956
YEARS_BEGINEXPLUATATION_MODE	48.781019
YEARS_BUILD_MODE	66.497784
COMMONAREA_MODE	69.872297
ELEVATORS_MODE	53.295980

	0
AMT_DOWN_PAYMENT	53.636480
RATE_DOWN_PAYMENT	53.636480
RATE_INTEREST_PRIMARY	99.643698
RATE_INTEREST_PRIVILEGED	99.643698
NAME_TYPE_SUITE	49.119754
DAYS_FIRST_DRAWING	40.298129
DAYS_FIRST_DUE	40.298129
DAYS_LAST_DUE_1ST_VERSION	40.298129
DAYS_LAST_DUE	40.298129
DAYS_TERMINATION	40.298129
NFLAG_INSURED_ON_APPROVAL	40.298129

Data Profiling before processing

Using the Pandas profiling library to quickly look at the total dataset

Observations from the Application profiling

- ORGANIZATION_TYPE has a high cardinality: 58 distinct values
- NAME_CONTRACT_TYPE is highly imbalanced (54.6%)
- NAME_TYPE_SUITE is highly imbalanced (66.1%)
- NAME_EDUCATION_TYPE is highly imbalanced (52.8%)
- NAME_HOUSING_TYPE is highly imbalanced (72.1%)
- OCCUPATION TYPE has 96391 (31.3%) missing values
- AMT REQ CREDIT BUREAU_HOUR has 41519 (13.5%) missing values
- AMT REO CREDIT BUREAU DAY has 41519 (13.5%) missing values
- AMT REO CREDIT BUREAU WEEK has 41519 (13.5%) missing values
- AMT REQ CREDIT BUREAU MON has 41519 (13.5%) missing values
- AMT REO CREDIT BUREAU QRT has 41519 (13.5%) missing values
- AMT REO CREDIT BUREAU YEAR has 41519 (13.5%) missing values
- AMT_INCOME_TOTAL is highly skewed (y1 = 391.5596541)
- FLAG_MOBIL is highly skewed (y1 = -554.5367436)
- FLAG_CONT _MOBILE is highly skewed (y1 = -23.08117235)
- AMT_REQ_CREDIT_BUREAU_DAY is highly skewed (y1 = 27.04350471)
- AMT_REQ_CREDIT_BUREAU_QRT is highly skewed (y1 = 134.365776)

Data Profiling before processing

Using the Pandas profiling library to quickly look at the total dataset

Observations from the Previous Application profiling

- FLAG_LAST_APPL_PER_CONTRACT is highly imbalanced (95.4%)
- NAME_CASH_LOAN_PURPOSE is highly imbalanced (71.5%)
- CODE_REJECT_REASON is highly imbalanced (66.2%)
- NAME_GOODS_CATEGORY is highly imbalanced (52.9%)
- AMT ANNUITY has 372235 (22.3%) missing values
- AMT GOODS PRICE has 385515 (23.1%) missing values
- CNT_PAYMENT has 372230 (22.3%) missing values
- SELLERPLACE _AREA is highly skewed (y1 = 529.6202788)

Handling missing values

Based on the observations of Pandas profiling, we will handle missing values in Application dataset

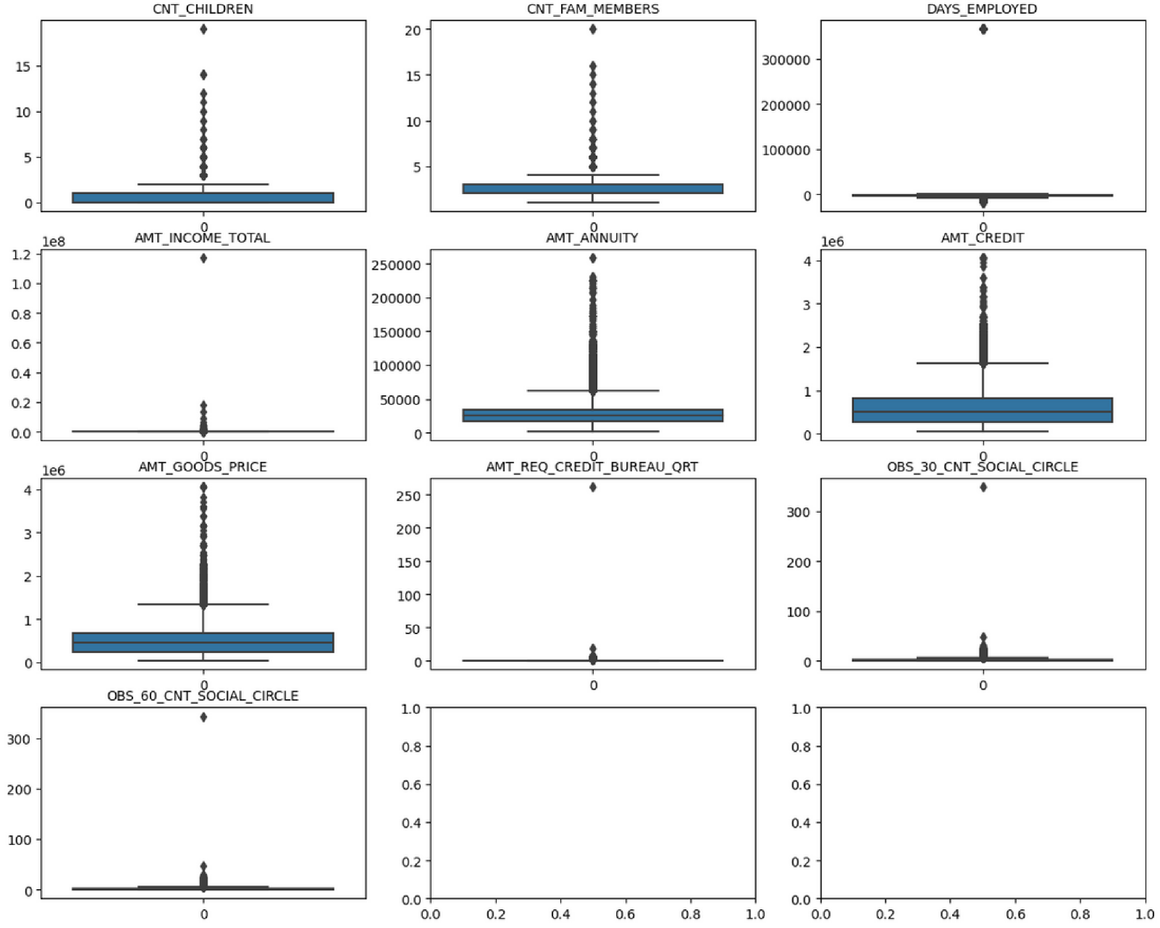
OCCUPATION_TYPE NAME_TYPE_SUITE	AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR	AMT_GOODS_PRICE OBS_30_CNT_SOCIAL_CIRCLE DEF_30_CNT_SOCIAL_CIRCLE OBS_60_CNT_SOCIAL_CIRCLE DEF_60_CNT_SOCIAL_CIRCLE DAYS_LAST_PHONE_CHANGE CNT_FAM_MEMBERS
Both columns have been completely missing at random. We will replace missing values with "unknown.".	Replacing missing values with the median instead of the mean to impute missing values because by using the median to impute missing values, we can keep the skewness of the distribution and avoid introducing biases that might arise from using the mean.	These columns also have a small number of missing values. So we are going to drop missing values as they don't affect the analysis.

Handling missing values

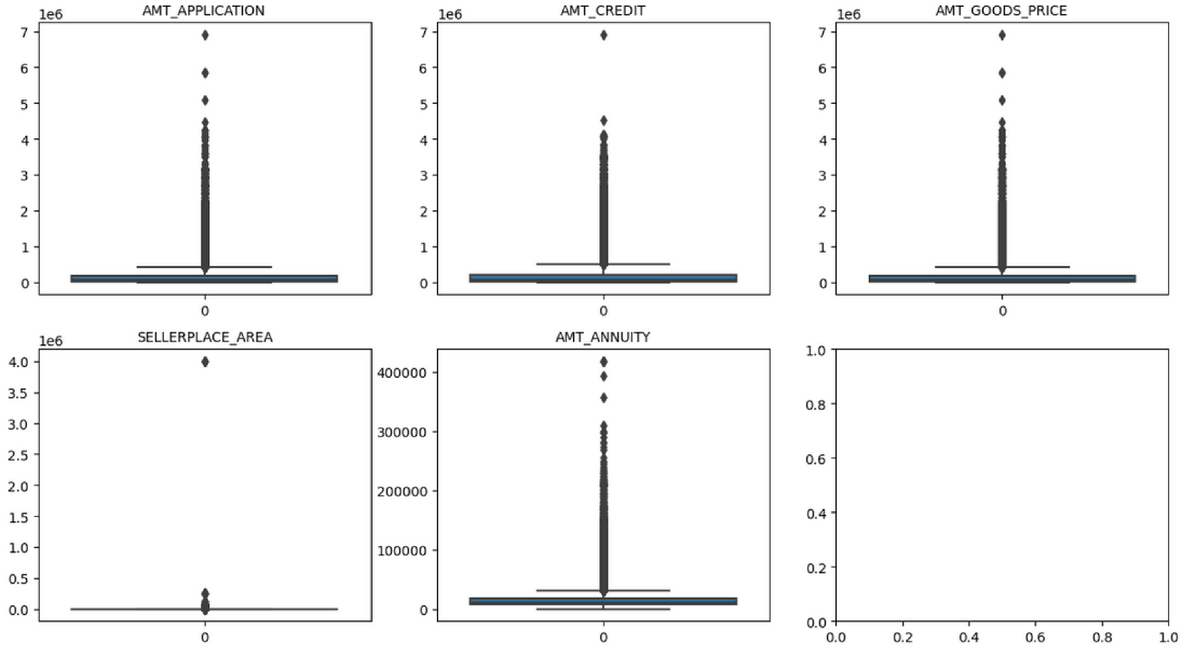
Based on the observations of Pandas profiling, we will handle missing values in Previous Application dataset.

AMT_ANNUITY AMT_GOODS_PRICE CNT_PAYMENT	AMT_CREDIT PRODUCT_COMBINATION
Replacing missing values with the median instead of the mean to impute missing values because by using the median to impute missing values, we can keep the skewness of the distribution and avoid introducing biases that might arise from using the mean.	We can also see that these columns also have a small number of missing values. So we are going to drop missing values as they don't affect the analysis.

Identifying the outliers



Application dataset



Previous Application dataset

Standardlise values

We can see that some columns contain error values, such as the DAYS columns. We will standardize the values of these columns.

```
#convert negative values to positive values
def convert_values(x):
    if x < 0:
        return abs(x)
    else:
        return x
```

✓ 0.0s

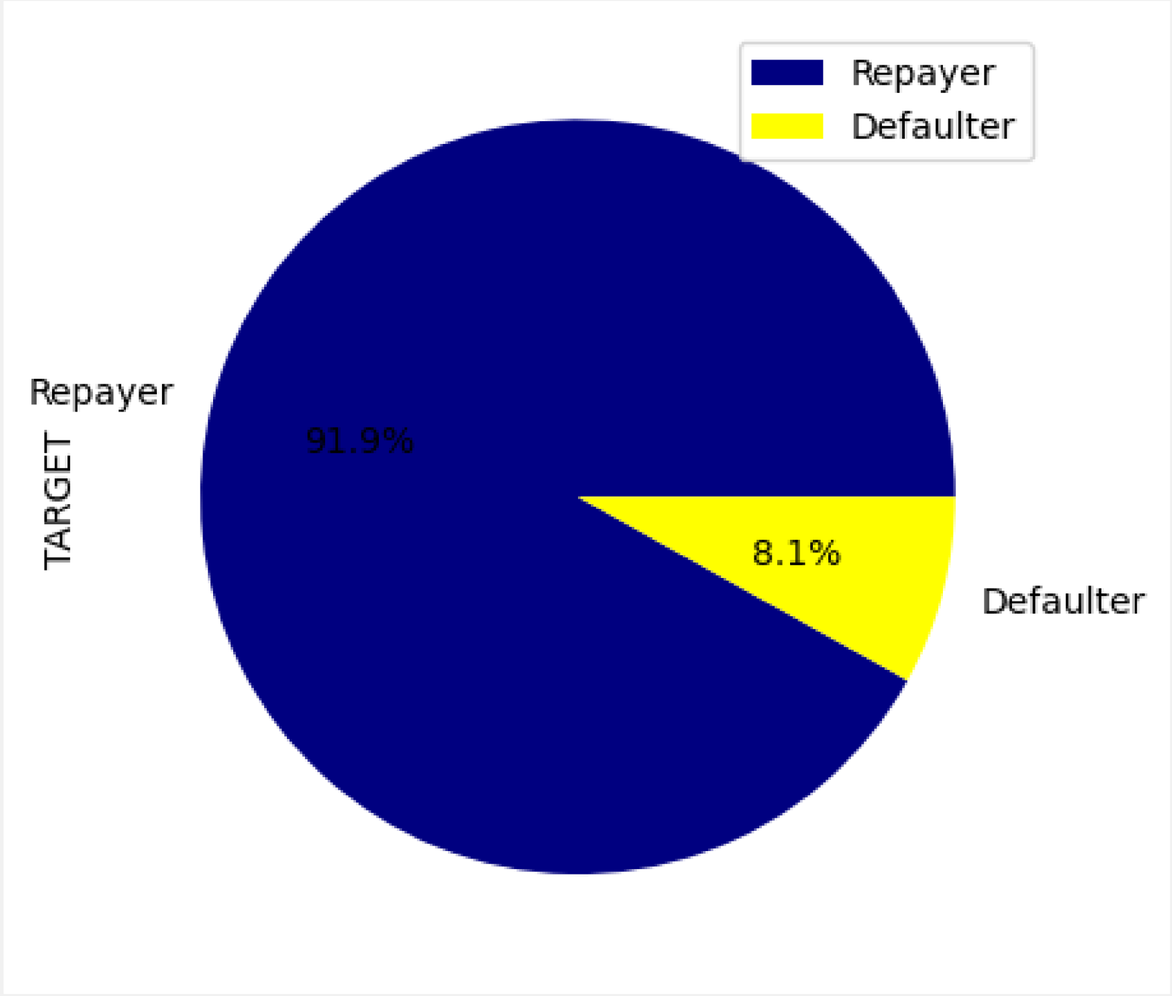
```
application['DAYS_EMPLOYED'] = application['DAYS_EMPLOYED'].apply(convert_values)
application['DAYS_BIRTH']= application['DAYS_BIRTH'].apply(convert_values)
application['DAYS_REGISTRATION'] = application['DAYS_REGISTRATION'].apply(convert_values)
application['DAYS_ID_PUBLISH'] = application['DAYS_ID_PUBLISH'].apply(convert_values)
application['DAYS_LAST_PHONE_CHANGE'] = application['DAYS_LAST_PHONE_CHANGE'].apply(convert_values)
```

✓ 0.2s

Exploratory Data Analysis

- Univariate and bivariate analysis
  - Categorical Variables
  - Numerical Variables
  - Merged Dataset
- Correlation and Multivariate analysis
  - Categorical Variables
  - Numerical Variables

Univariate and Bivariate analysis



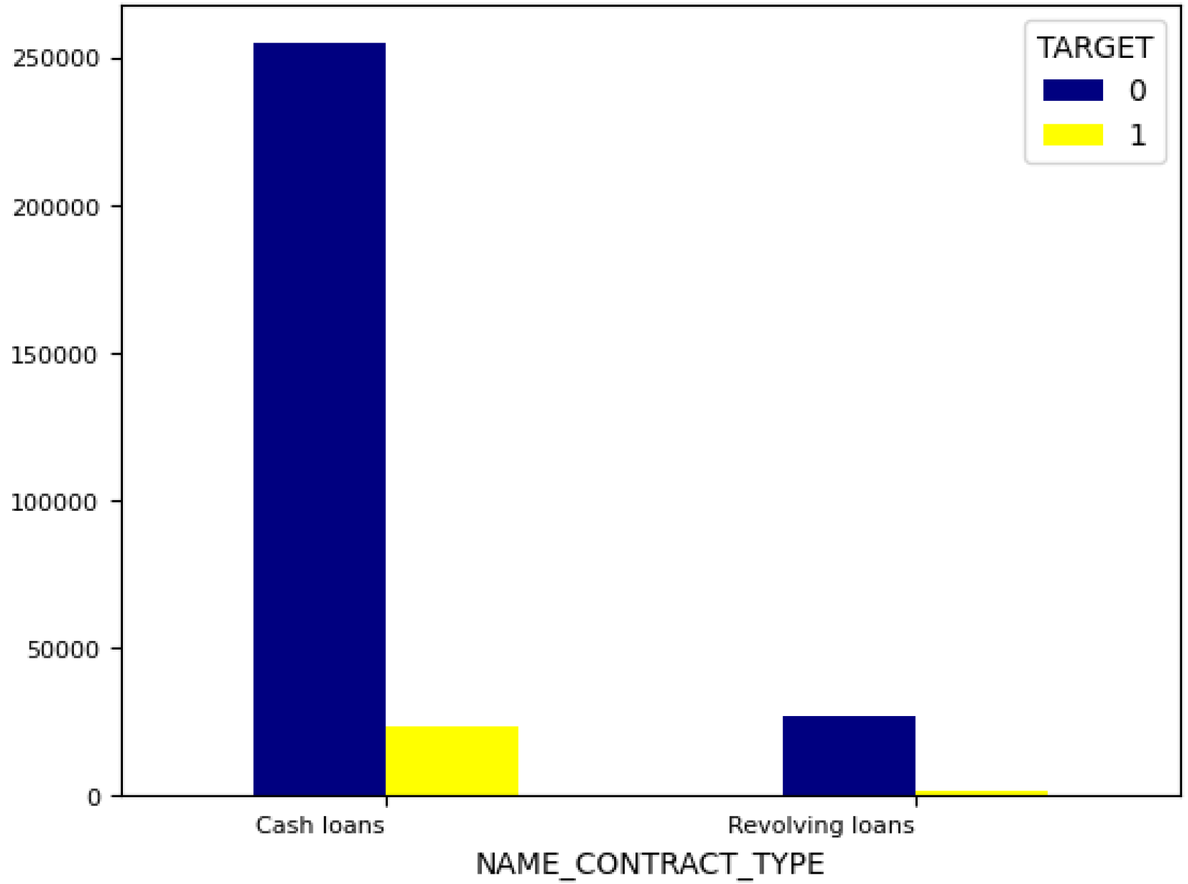
Target variable

- The target column is imbalanced with 91% repairs and 8.1% defaulters.
- The bank should identify the characteristics of each type to avoid loan default.

Univariate and Bivariate analysis

Contract Type

- The majority of applications are for cash loans. The variable is imbalance, so we have to apply a new scale to compare these two types of contracts.
- According to the ratio, clients who apply for cash loans are more likely to default than clients who apply for revolving loans.

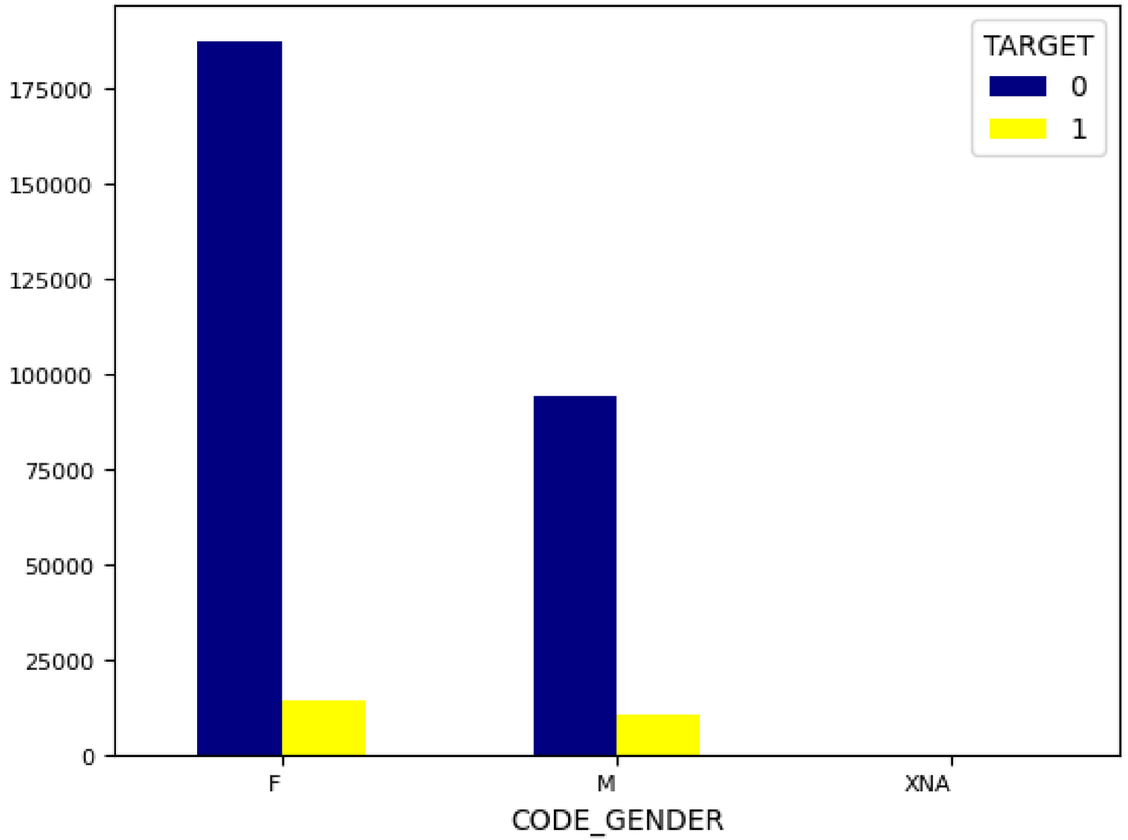


- The ratio of repayer to defaulter for cash loans is 10.981.
- The ratio of repayer to defaulter of revolving loans:17.087

Univariate and Bivariate analysis

Gender

- The number of female customers is higher than the number of male customers.
- Based on the ratio, female clients have a higher ability to pay on time than male clients.

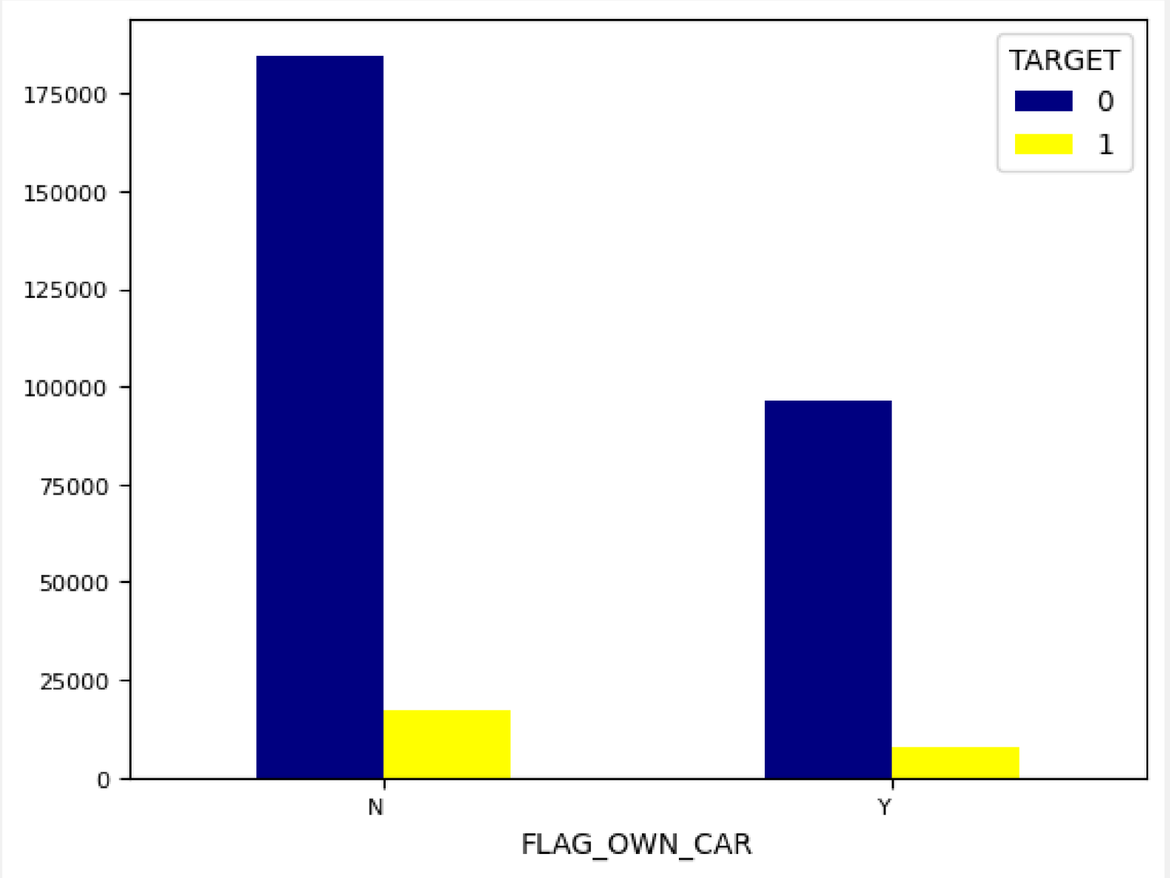


- The ratio of repayer to defaulter for females is 13.267.
- The ratio of repayer to defaulter for males is 8.832

Univariate and Bivariate analysis

Car owners

- Most of the clients don't own cars.
- Based on the ratio, car owners have a higher ability to pay on time than non-car owners. It can explain the fact that car owners have higher financial ability than non-car owners. Hence, they are less likely to default.

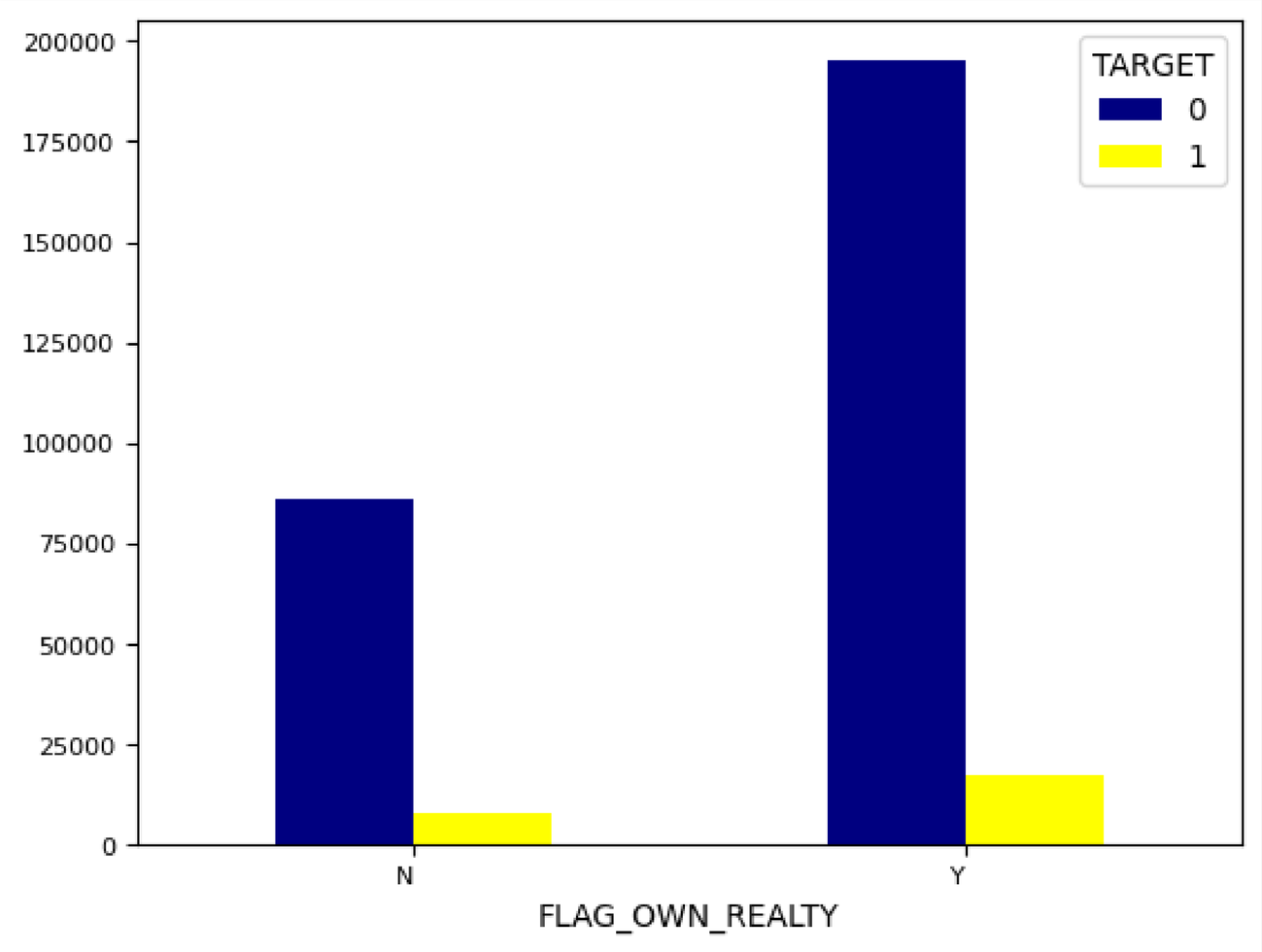


- The ratio of repayer to defaulter of Car owners: 12.785
- The ratio of repayer to defaulter of Non-car owners: 10.739

Realty Owners

- The majority of the clients own realty.
- The realty owners have a higher ability to pay on time than non-realty owners. However, we can see that gap between two types isn't transparent because the realty is a high value asset. Clients with average finances are difficult to have the realty, but that does not mean they will have a high probability of defaulting. Therefore, in this case, we can consider that the correlation between realty ownership and default is nonexistent.

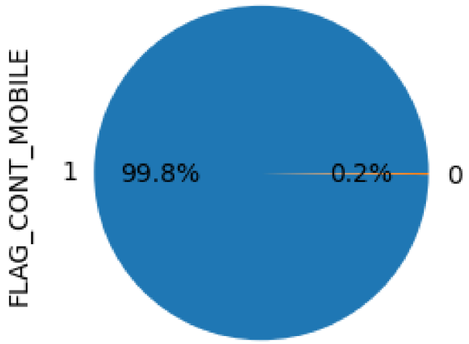
Univariate and Bivariate analysis

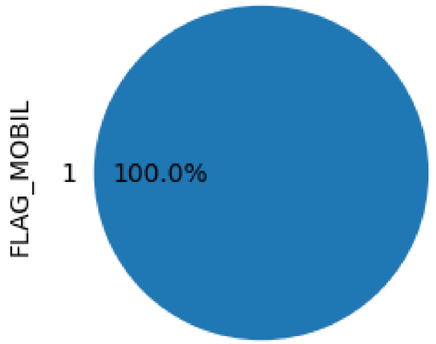


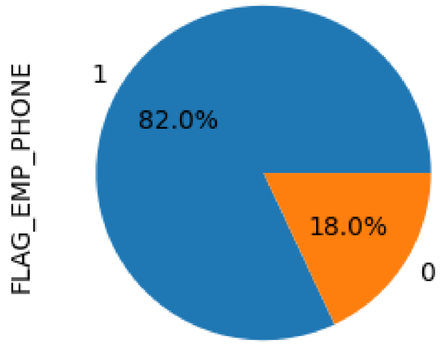
- The ratio of repayer to defaulter of Realty owners: 11.537
- The ratio of repayer to defaulter of Non-realty owners: 10.985

Univariate and Bivariate analysis

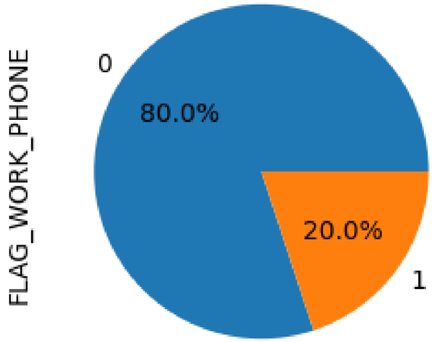
Provided Contact Information

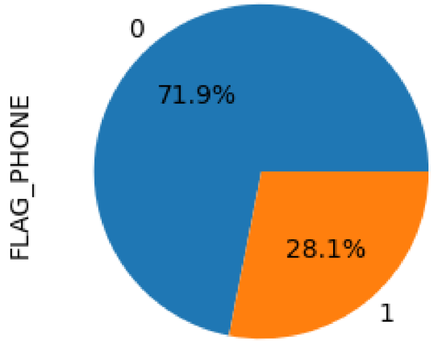


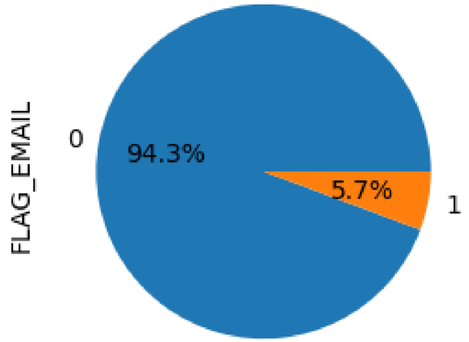




Following the charts, these columns are data imbalances. The clients just provided their mobile phone and don't want to share their email, work phone, or home phone. Also, 99.8% of mobile phones are reachable. In conclusion, the missing information about contact can be affected by other reasons, not because the clients are likely to default because they don't provide information.

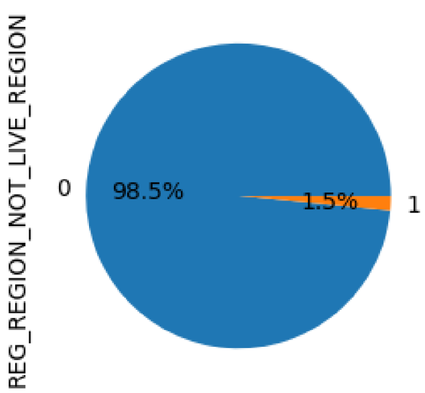


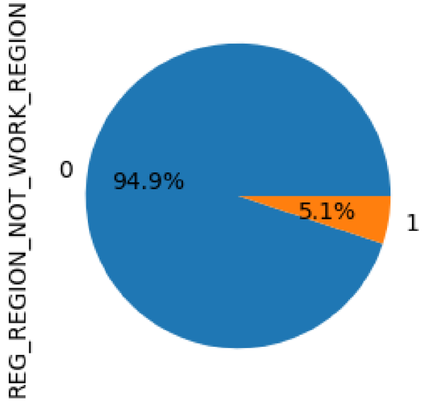


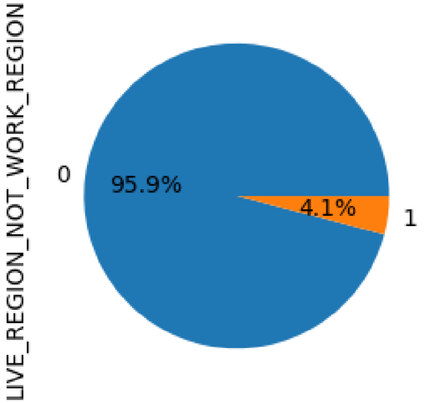


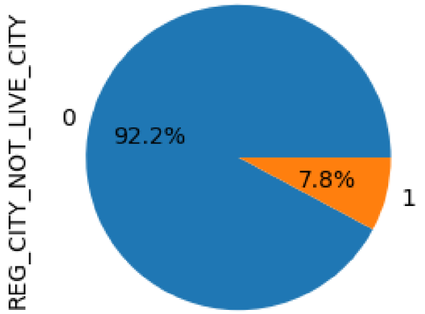
Address Information

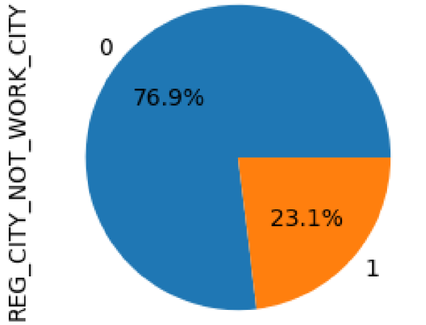
The permanent address, contract address, and work address of the majority of customers are matched together at the regional level. The difference between living in a city and working in a city is high, but in fact, it is also normal.

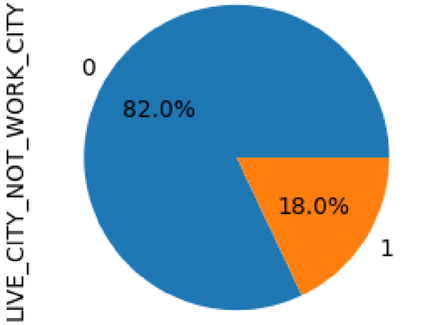






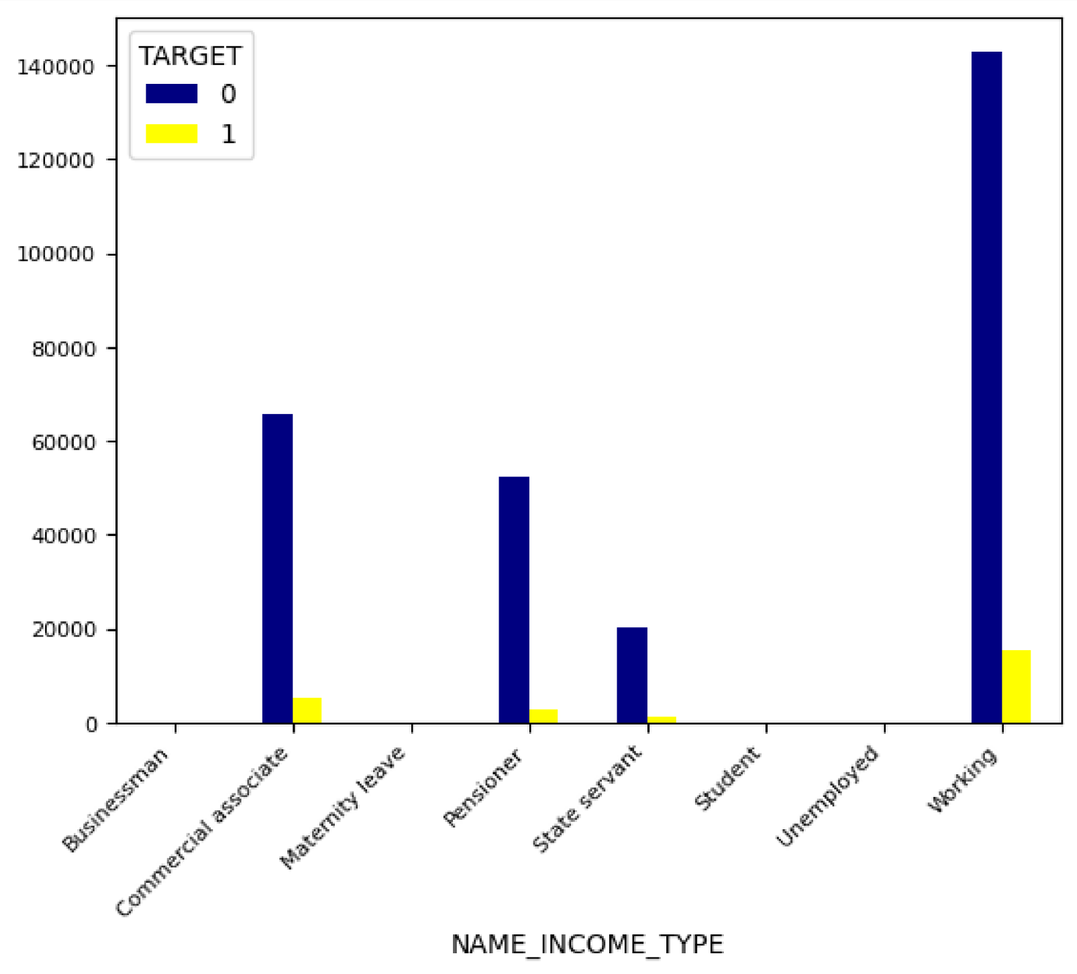


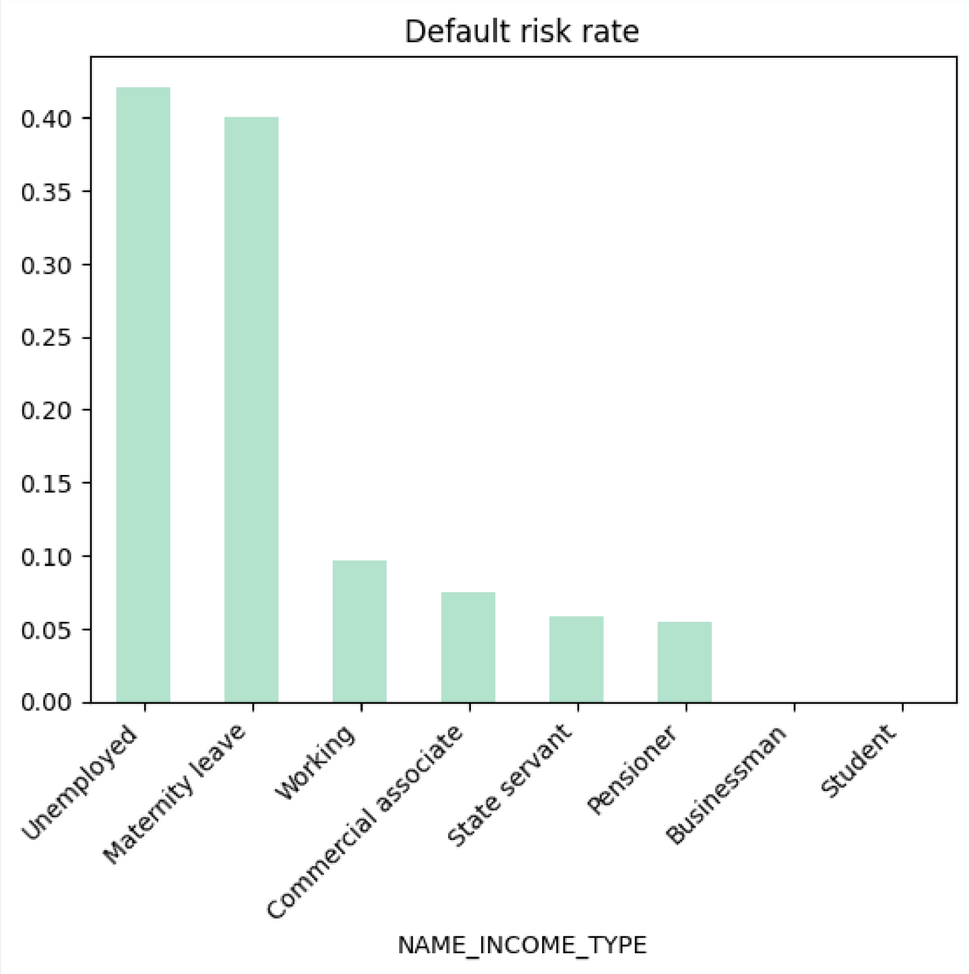




Univariate and Bivariate analysis

Income Type

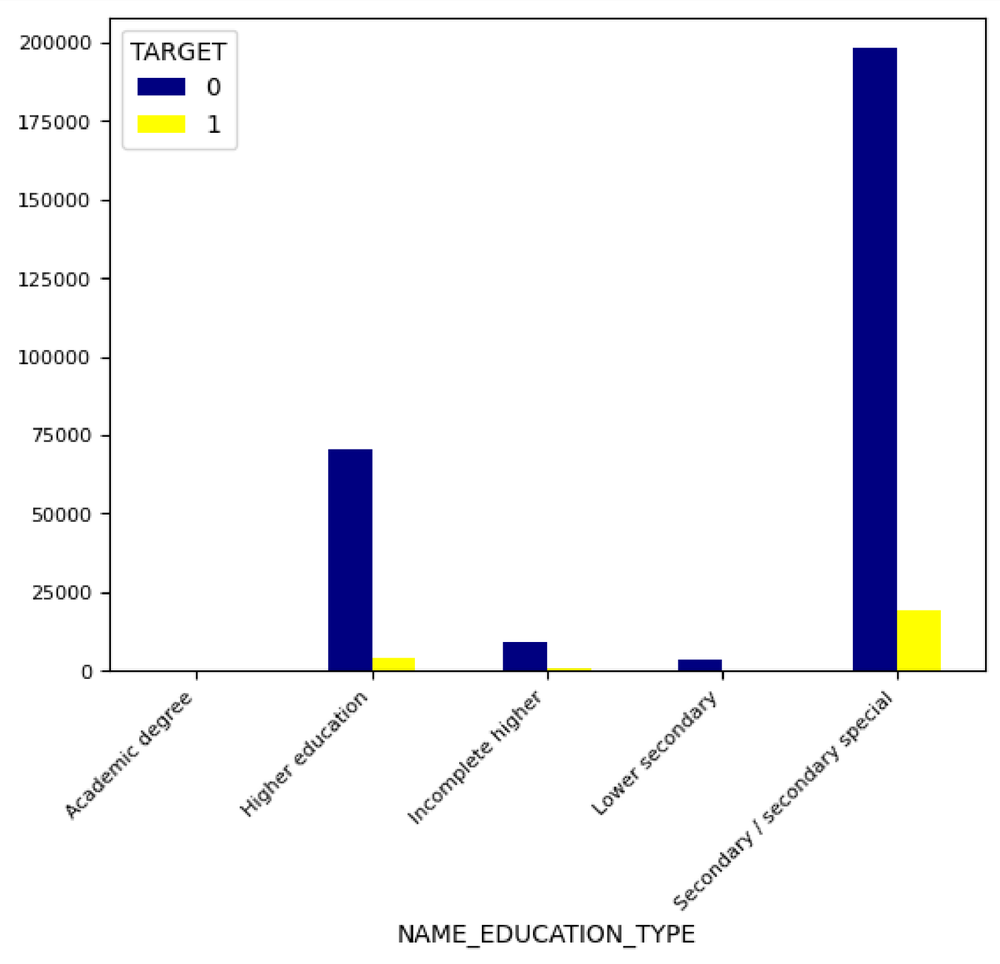


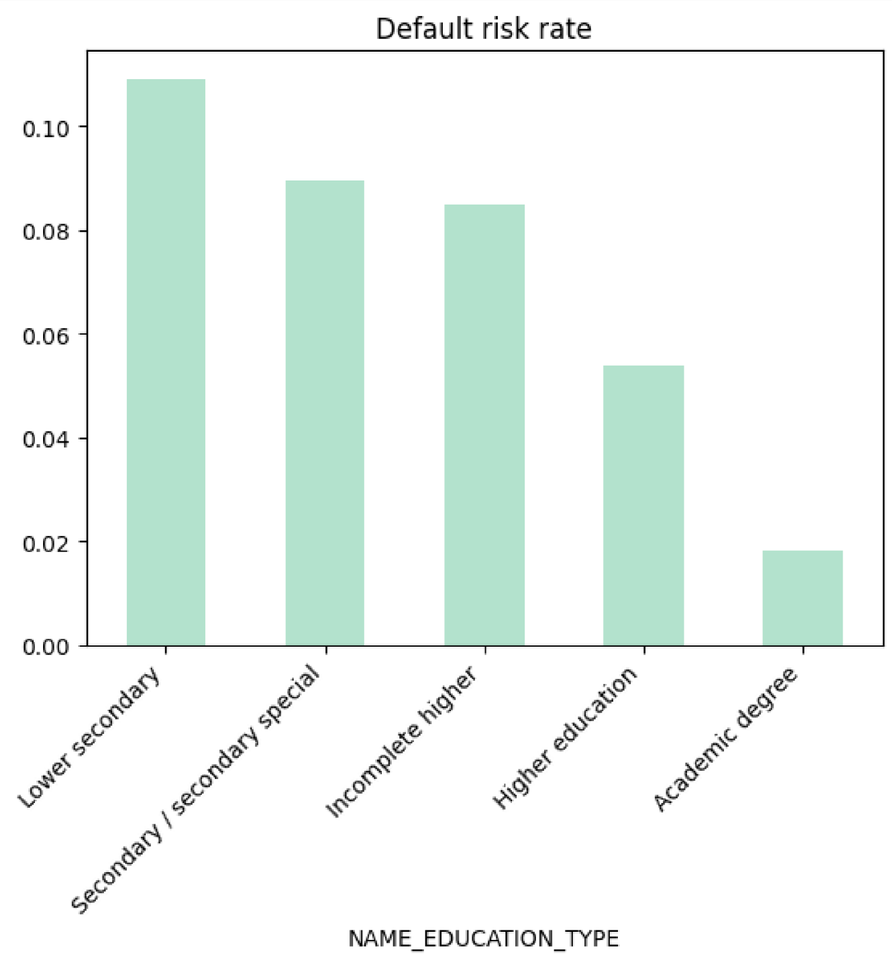


The working type has the most repayers, and the default risk rate is much lower than that of the unemployed and maternity leave. It can explain the fact that clients with income from subsidies find it difficult to pay loans on time.

Univariate and Bivariate analysis

Education Type

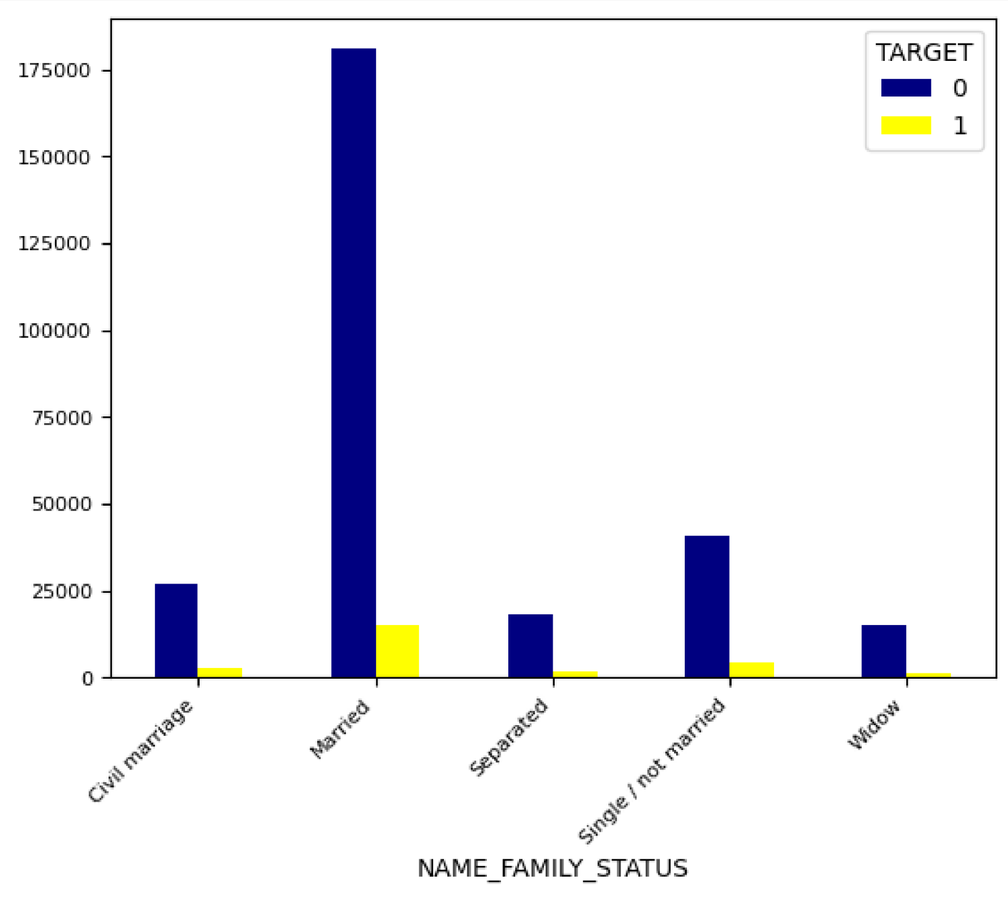


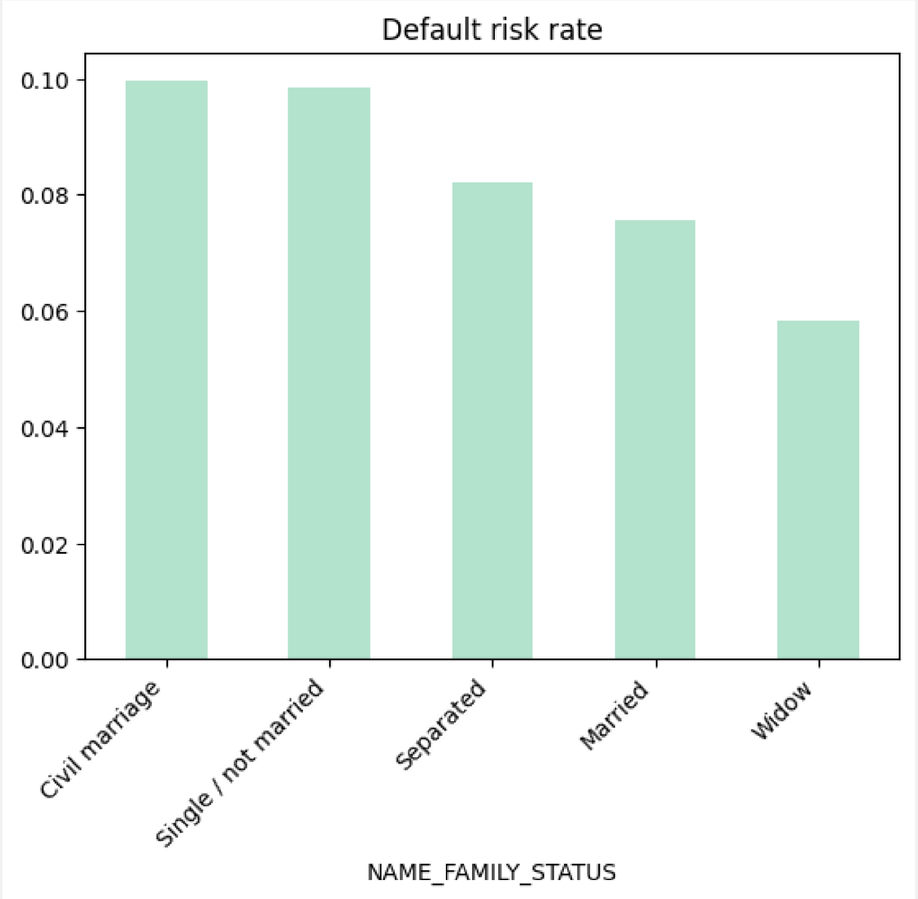


- The number of clients who have secondary or secondary special types is very high, but this type also has a high default rate.
- The default risk decreased with education levels. It can explain the fact that clients with high education levels have more awareness of the need to pay the loans on time.

Univariate and Bivariate analysis

Family Status

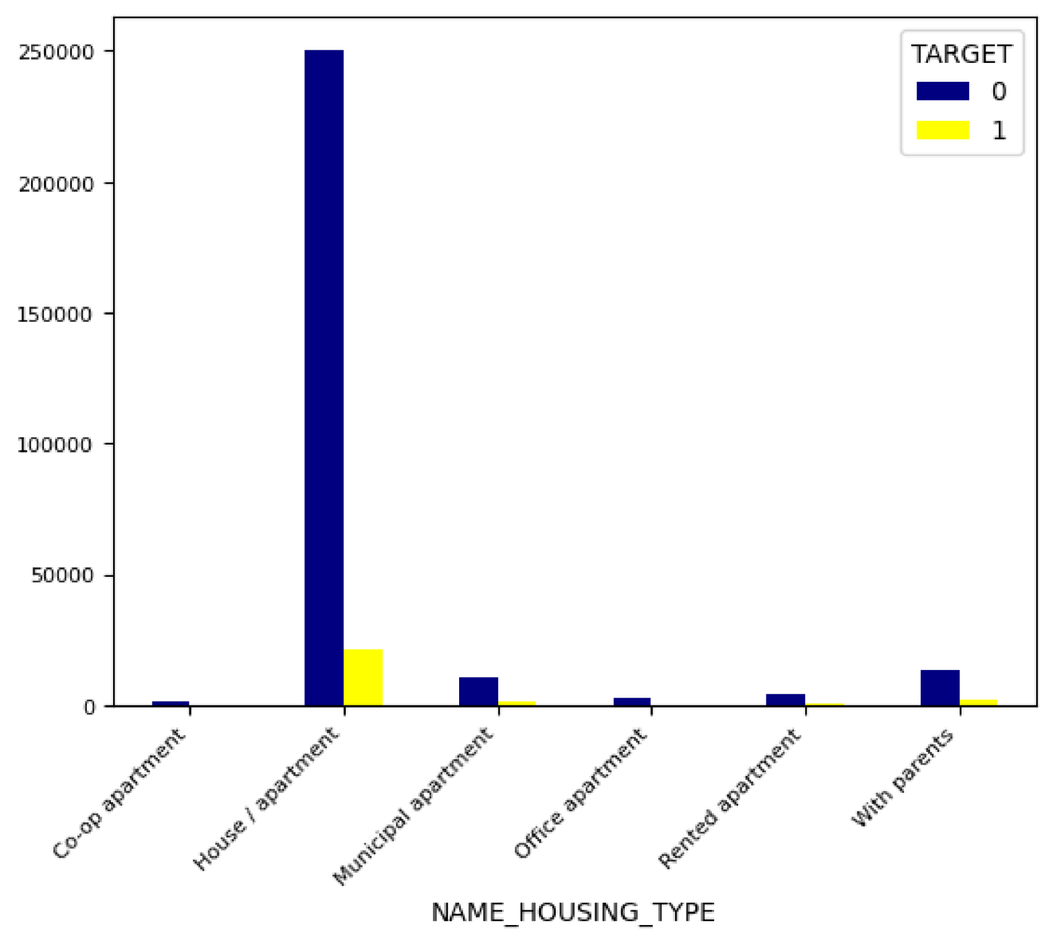


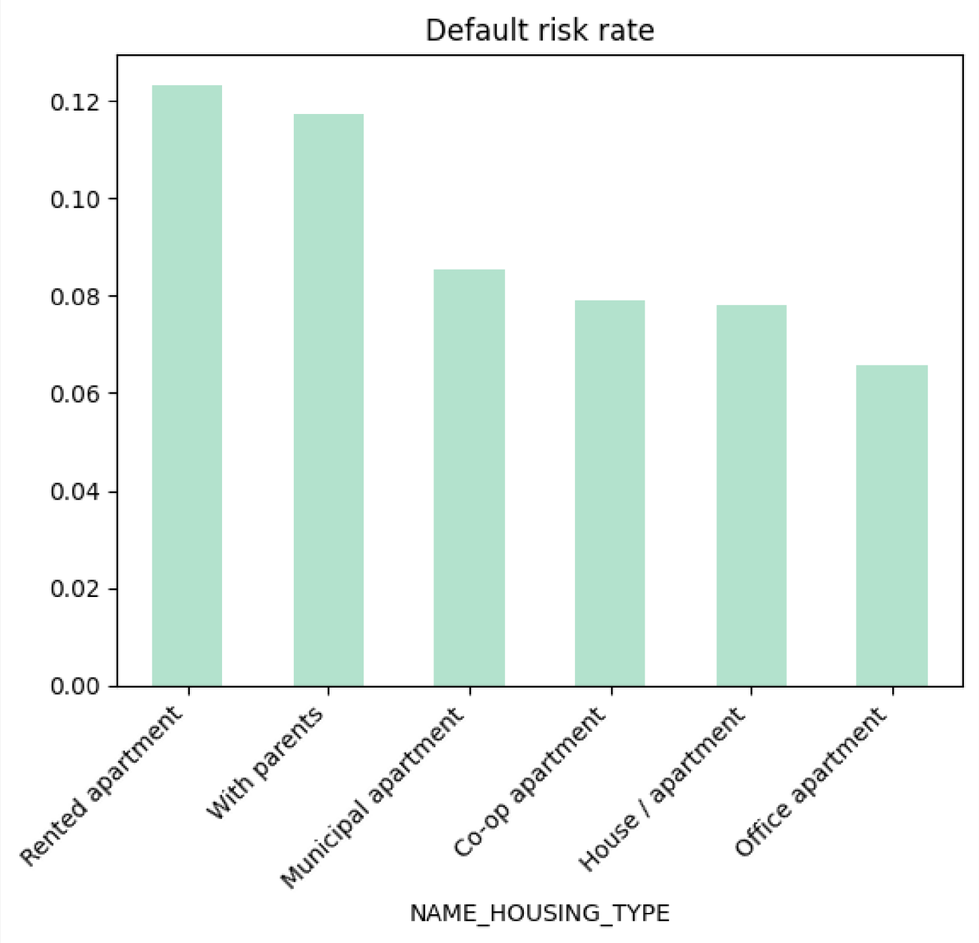


- Married clients are more likely to pay on time than other groups because the quantity is very high and the default risk is lower than for others.
- Single/unmarried is the second most common, followed by married, but it has a high default risk. The bank should be more careful with this type.

Univariate and Bivariate analysis

Housing Type

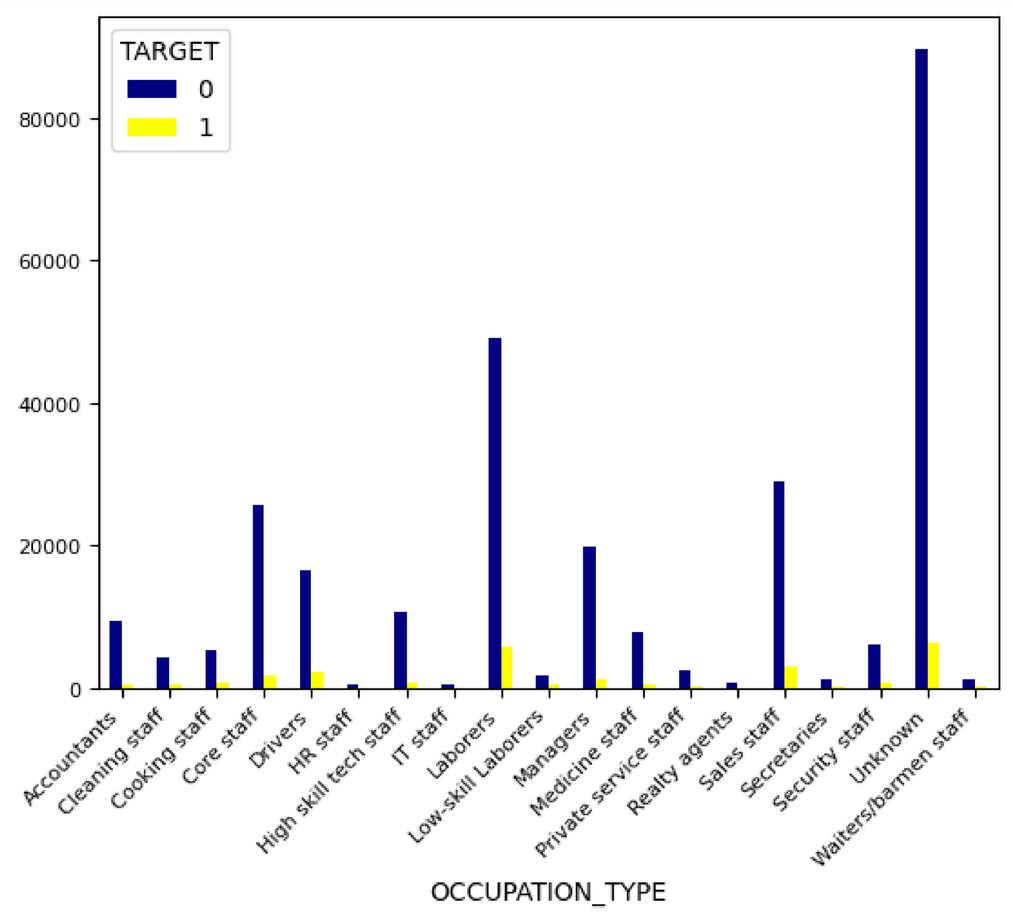


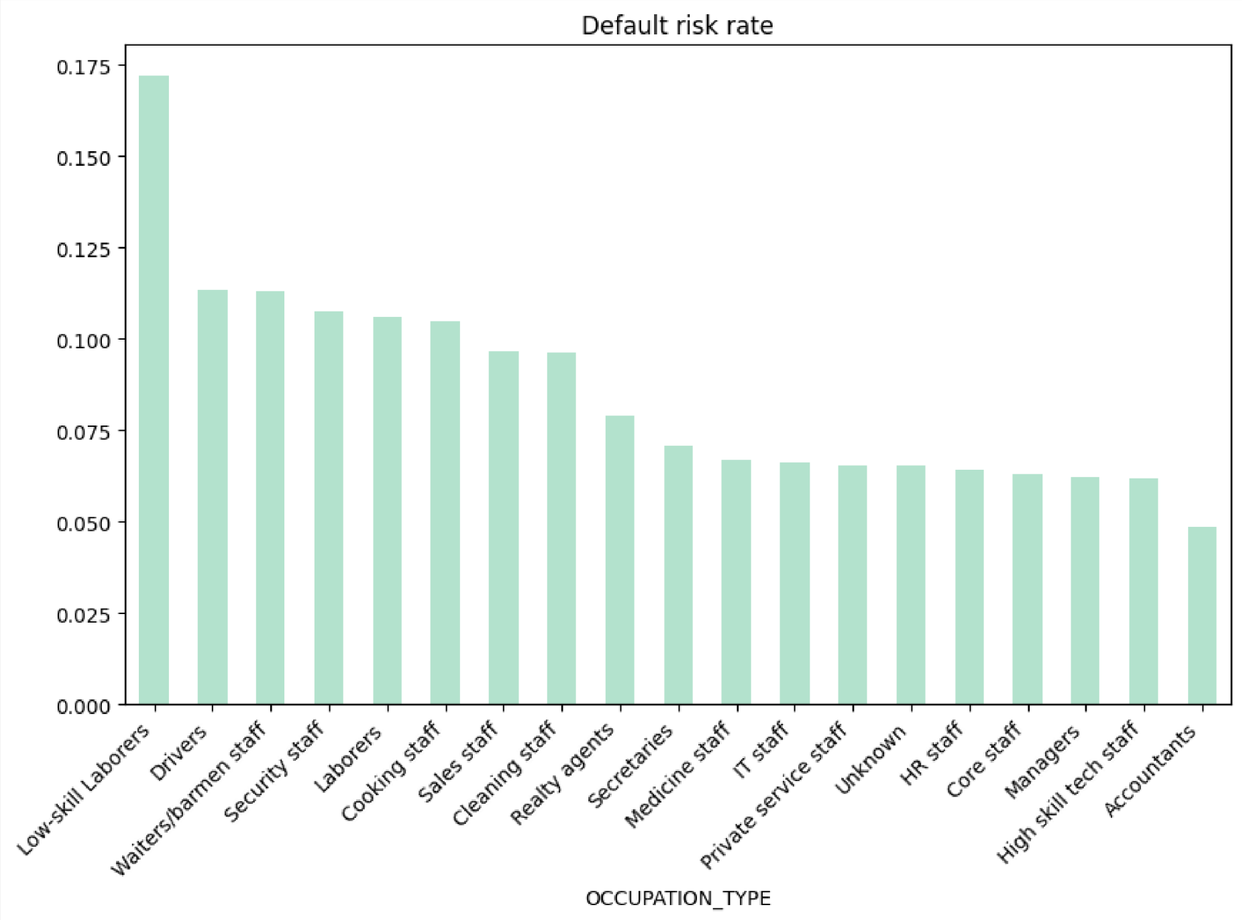


- The clients who have a house or apartment are more likely to pay on time than other groups because the quantity is very high and the default risk is lower than for others.
- The clients who have to rent department or live with parents are likely to default.

Univariate and Bivariate analysis

Occupation Type



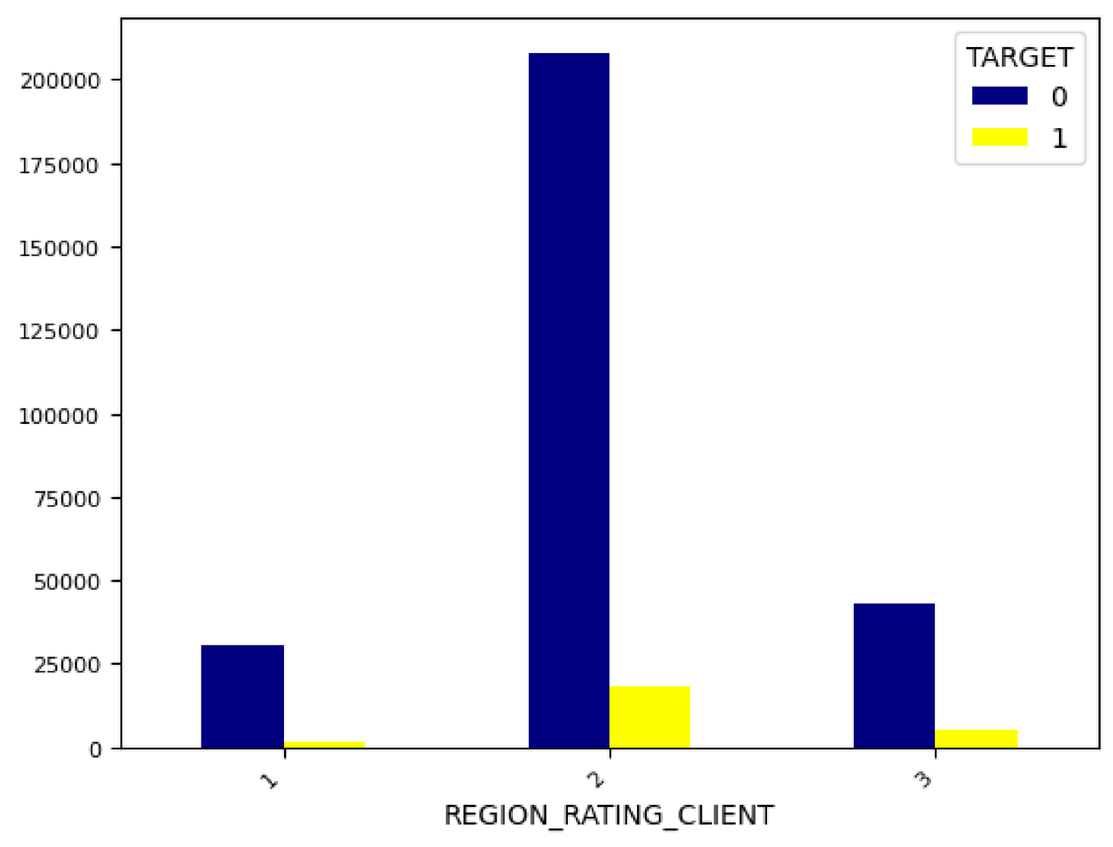


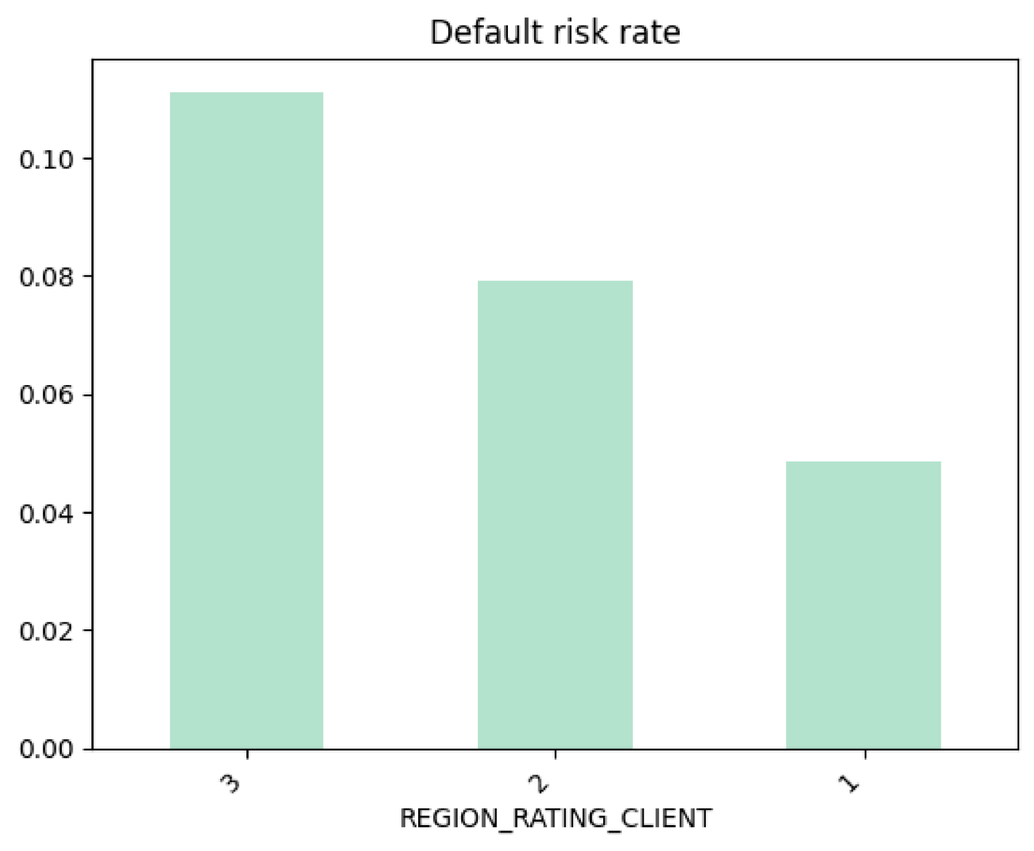
- Laborers and sales staff have the highest number of repayers, but the default risk is also high.
- Managers and core staff are also common and look like they provide more safety with low default risk.
- Drivers have a significant number of defaulters, but there is a very high default risk.

Categorical variables

Univariate and Bivariate analysis

Region Rating

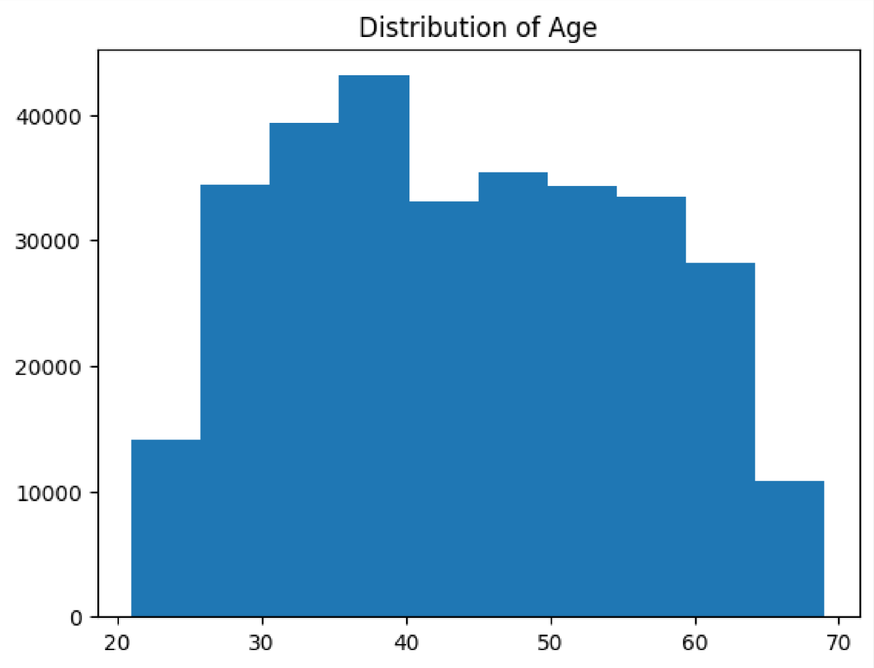


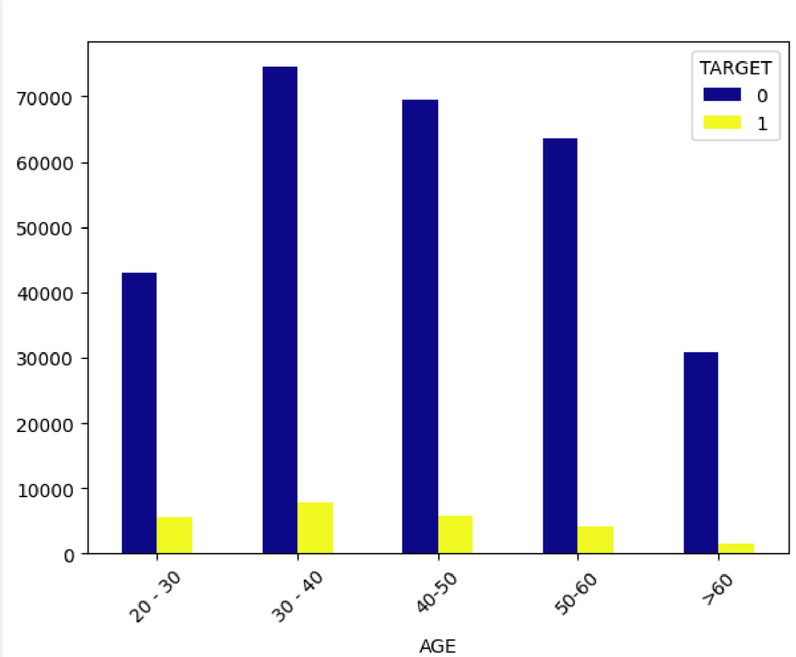


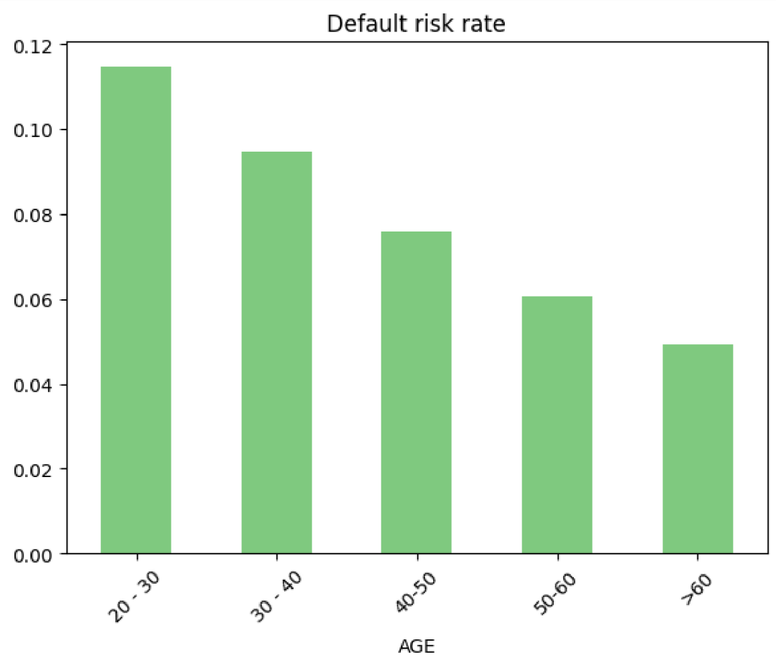
- The majority of clients living in the region have a rating 2, but the region with a rating of 3 has the highest default risk.

Univariate and Bivariate analysis

Age Group



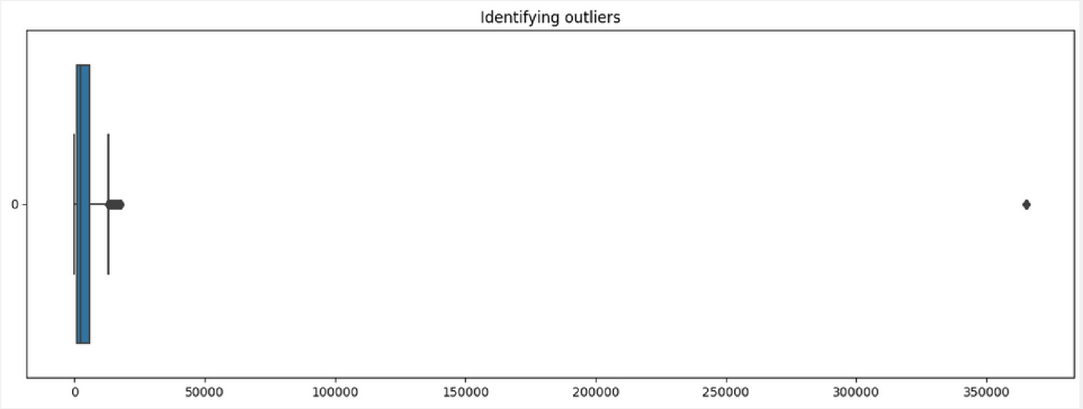




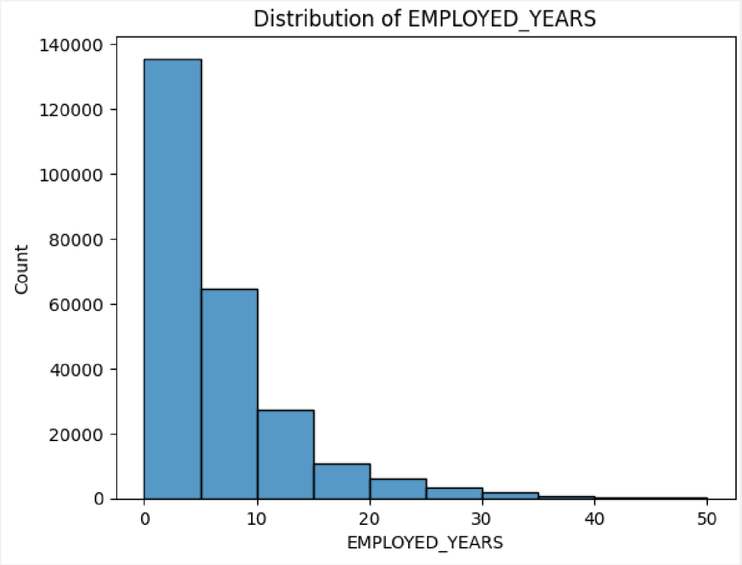
- The age group from 30 to 50 has the highest number of repayers; however, the default risk decreases with increasing age.
- The age group of 40-50 is more safe than 30 - 40 age group.

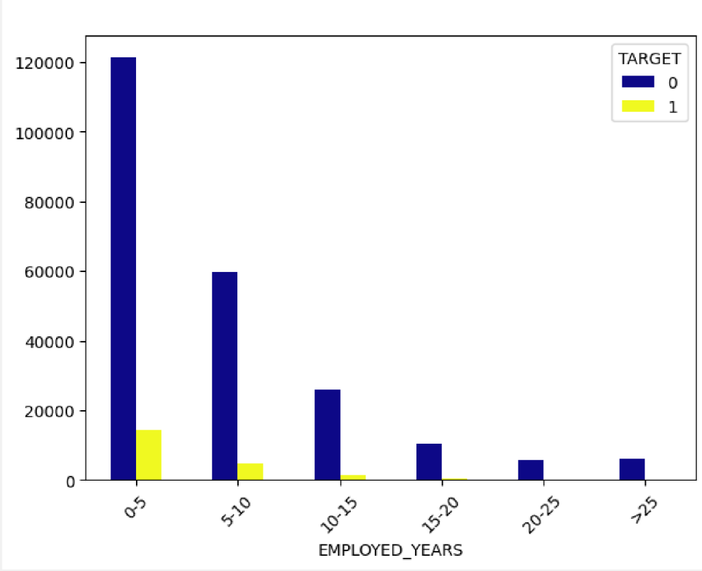
Univariate and Bivariate analysis

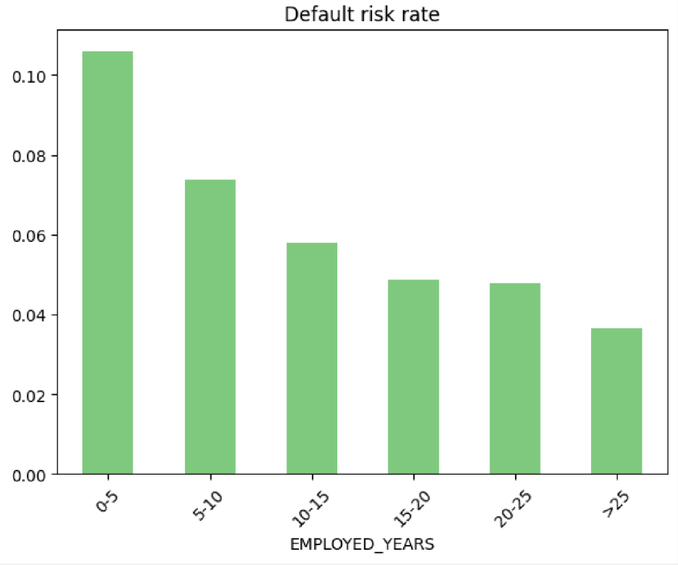
Days Employed



As the variable has outliers, we will remove them before analyzing the patterns.



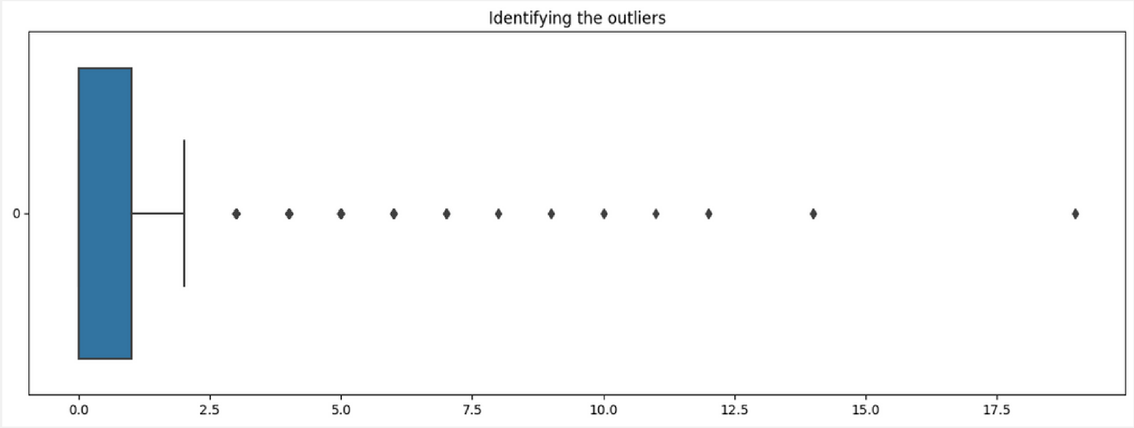




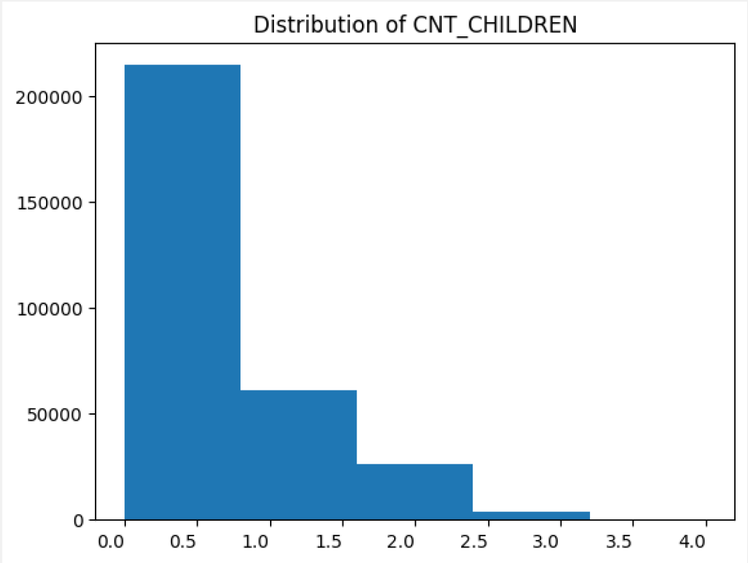
The majority of clients have worked for 0 to 5 years. The probability of default declines with more years of work. It proves that long-term employees who have saved a particular amount of money have more solid financial abilities than clients who have less experience.

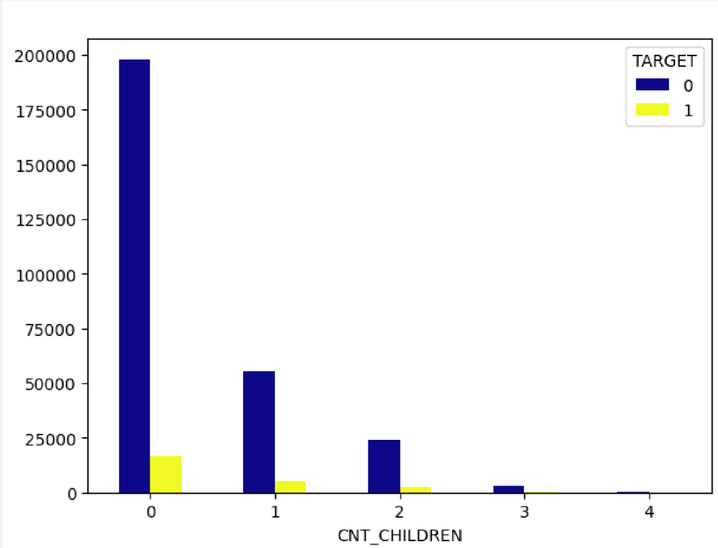
Univariate and Bivariate analysis

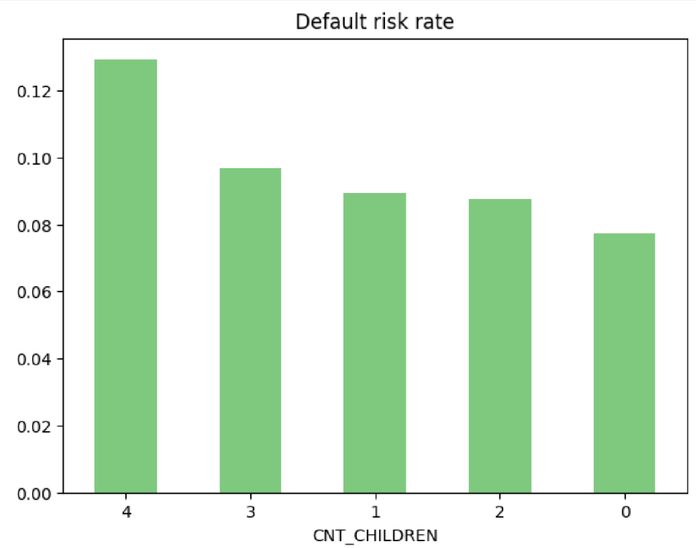
Children



As the variable has outliers, we will remove them before analyzing the patterns.



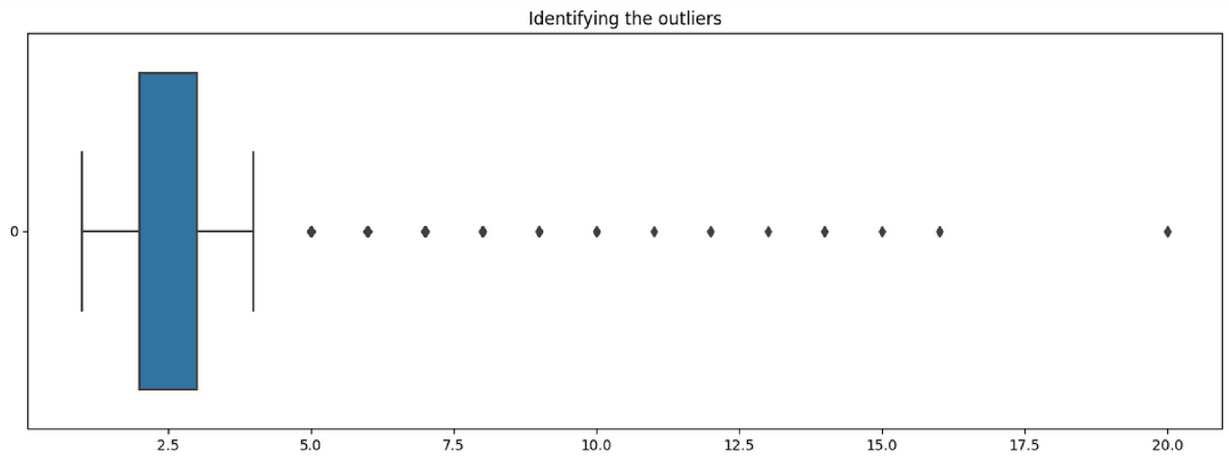




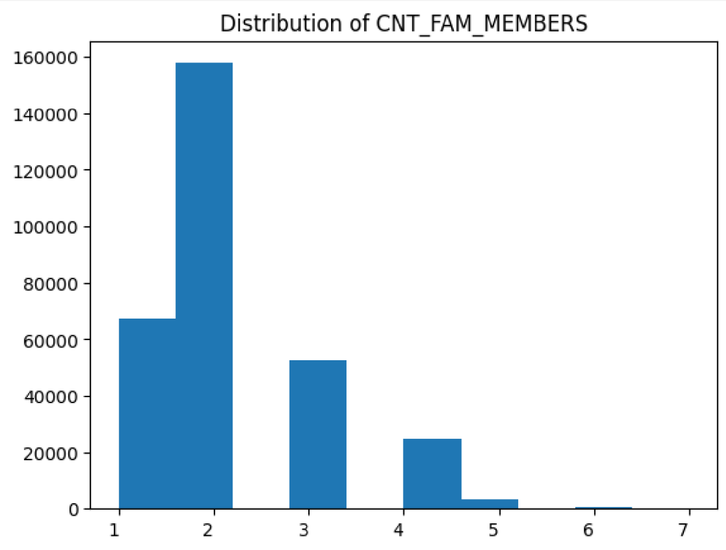
Clients who have several kids are more likely to default. Clients without kids will be the most secure. It explains the fact that clients have to pay a lot of money to raise their children, so it will be more difficult to pay on time.

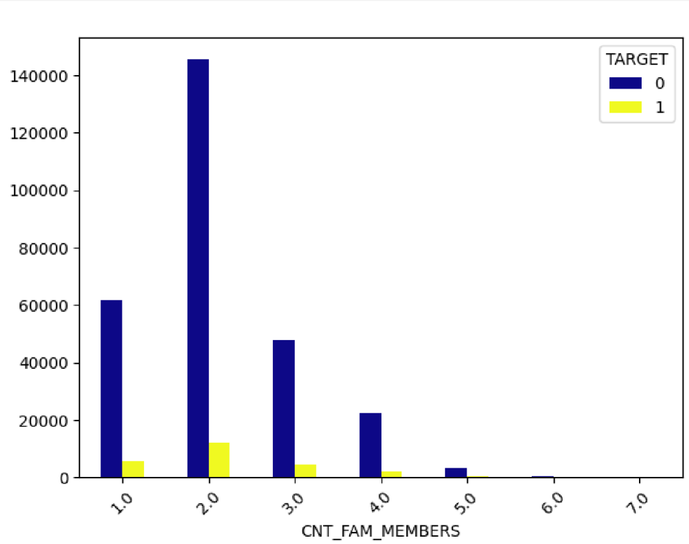
Univariate and Bivariate analysis

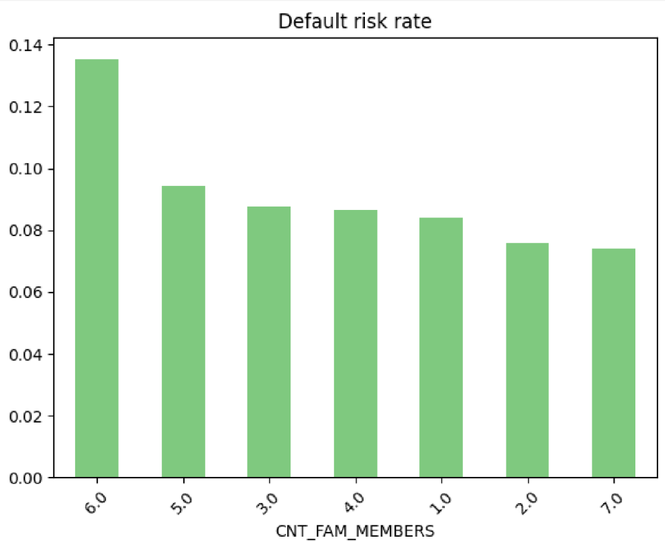
Family Members



As the variable has outliers, we will remove them before analyzing the patterns.



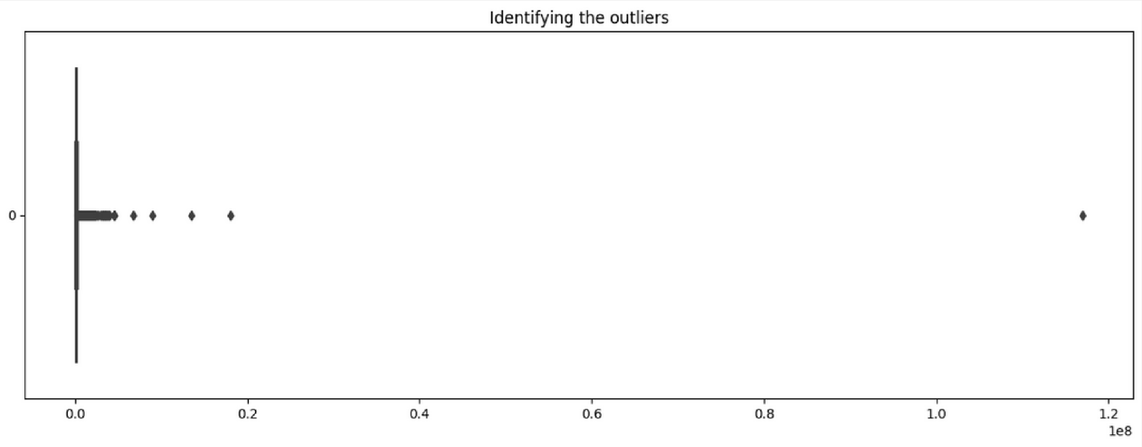




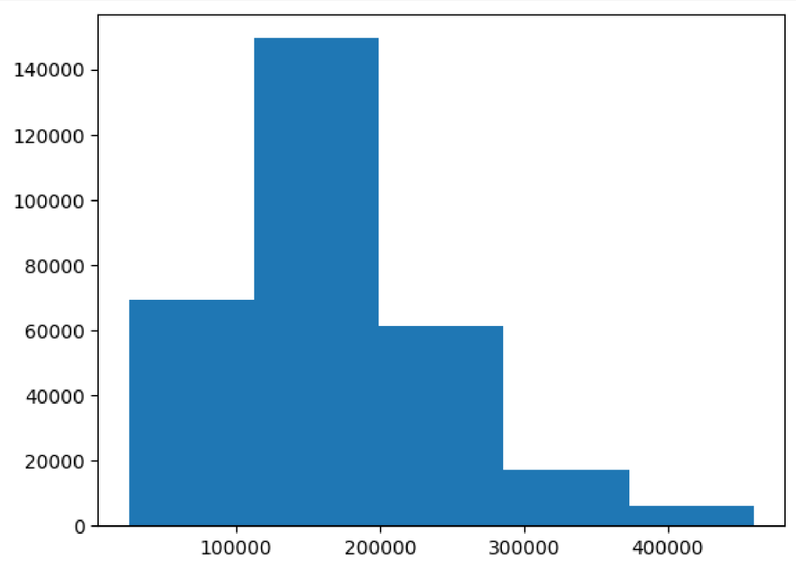
Clients with two members in their family are the safest group because they have the highest number of repayers and the lowest default risk.

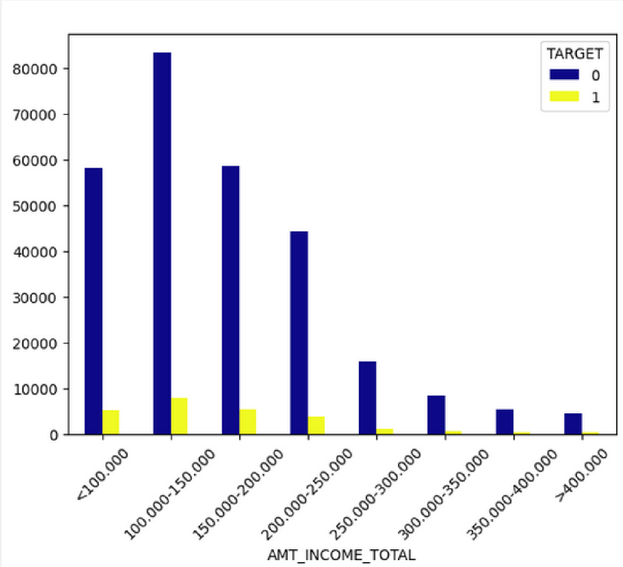
Univariate and Bivariate analysis

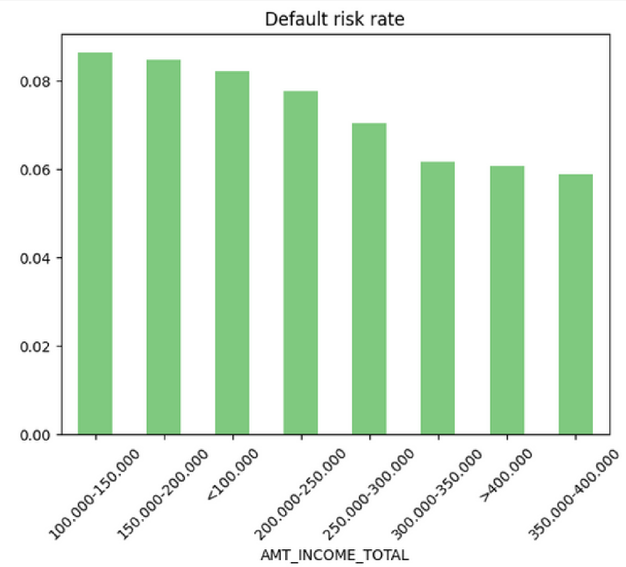
Income Amount



As the variable has outliers, we will remove them before analyzing the patterns.





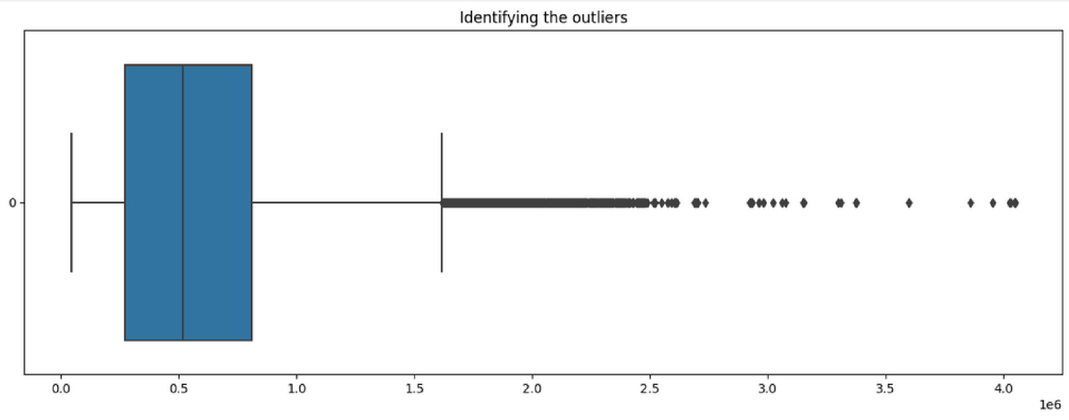


- The majority of clients earn between 100,000 and 150,000. However, this group also has a high default risk rate.
- The default risk rate decreases following the increasing income of clients.

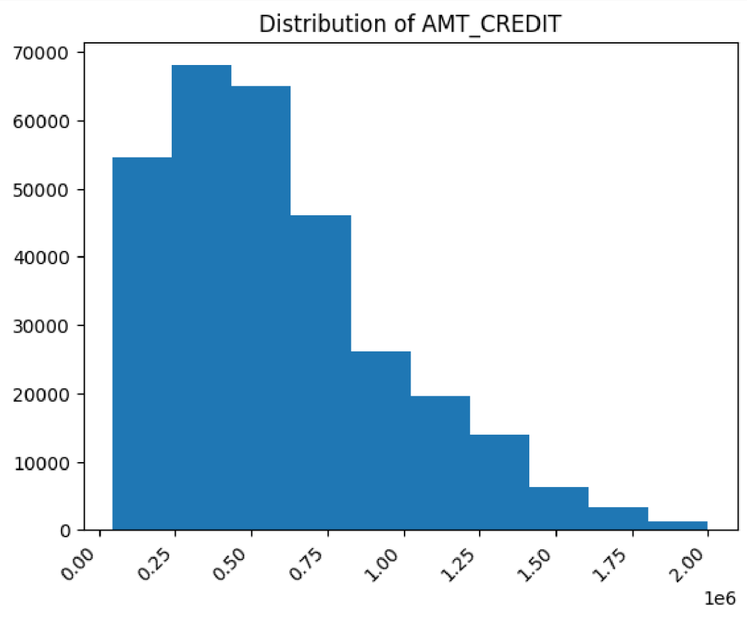
Numerical variables

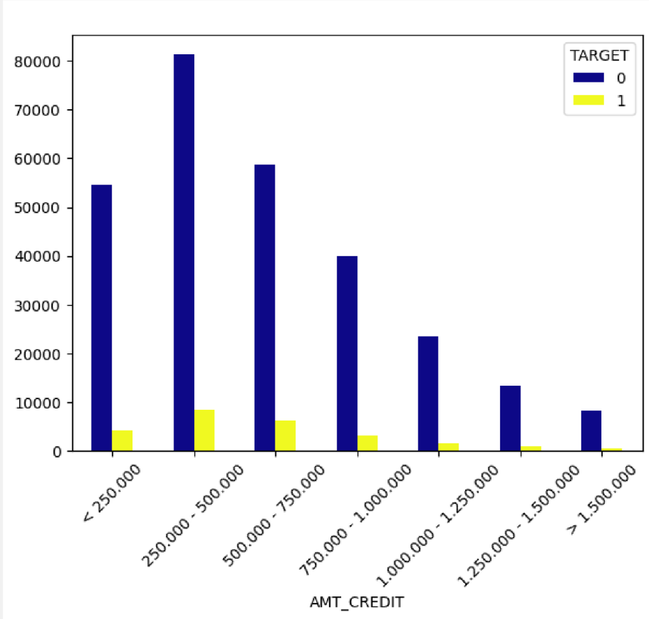
Univariate and Bivariate analysis

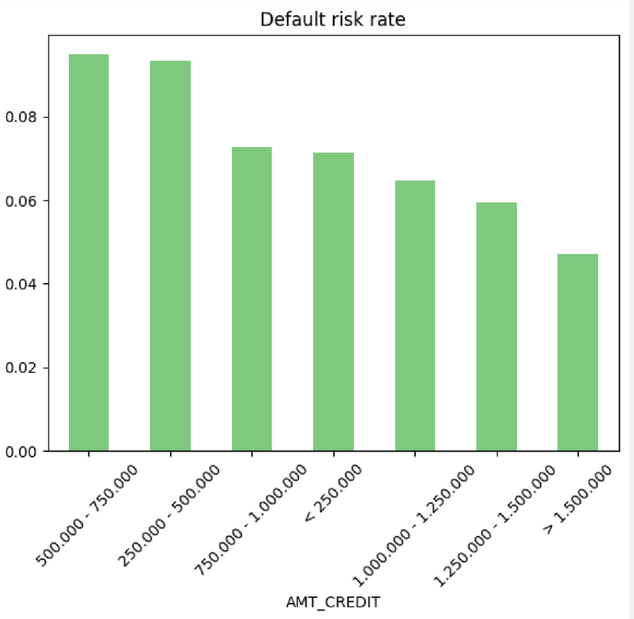
Credit Amount



As the variable has outliers, we will remove them before analyzing the patterns.



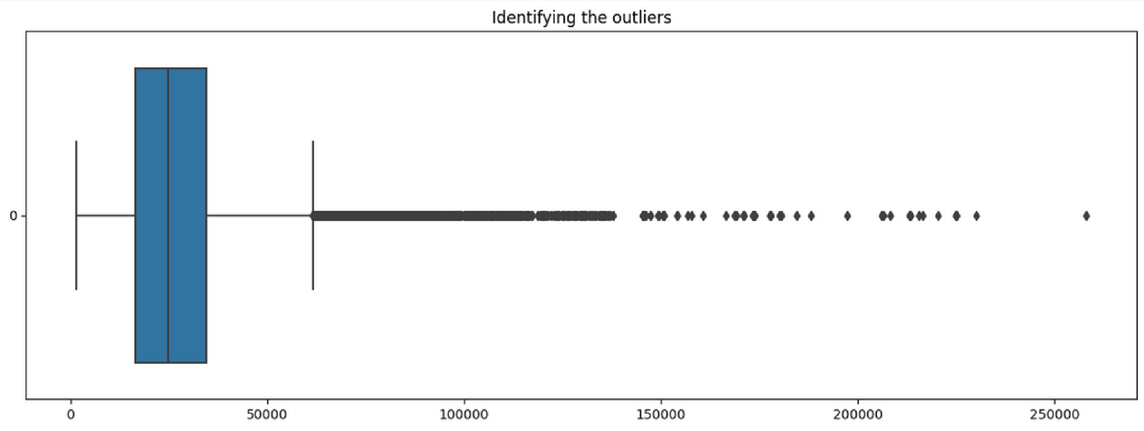




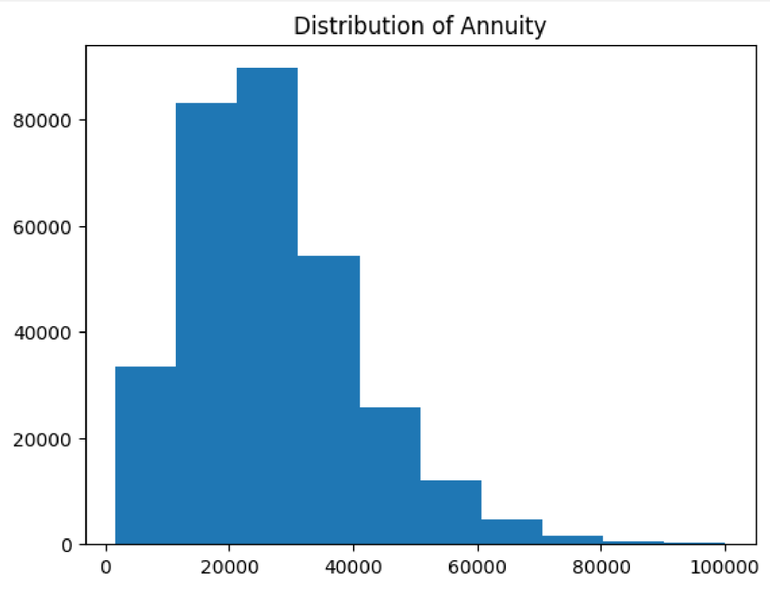
- The majority of clients applys for credit amount of loans between 250,000 and 500,000. However, this group also has a high default risk rate.
- The default risk rate decreases following the amount of credit.

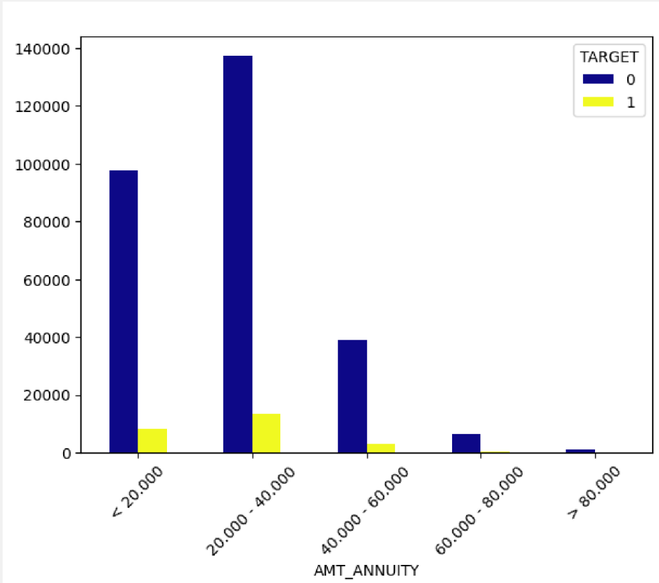
Univariate and Bivariate analysis

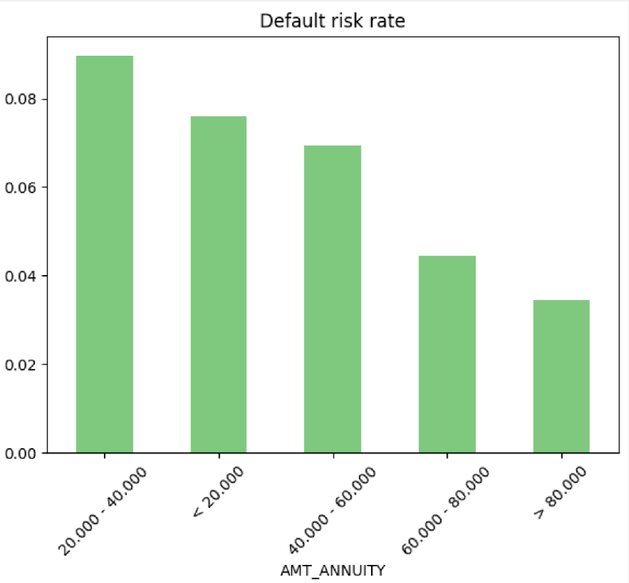
Annuity Amount



As the variable has outliers, we will remove them before analyzing the patterns.

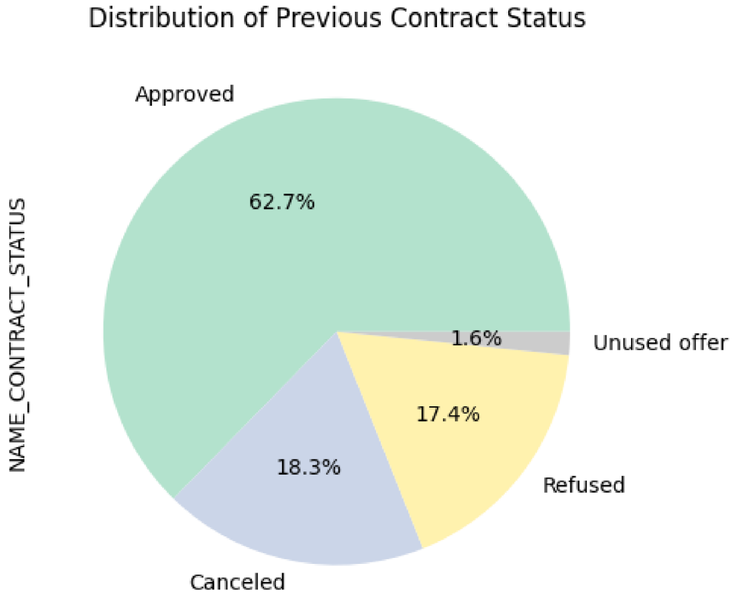


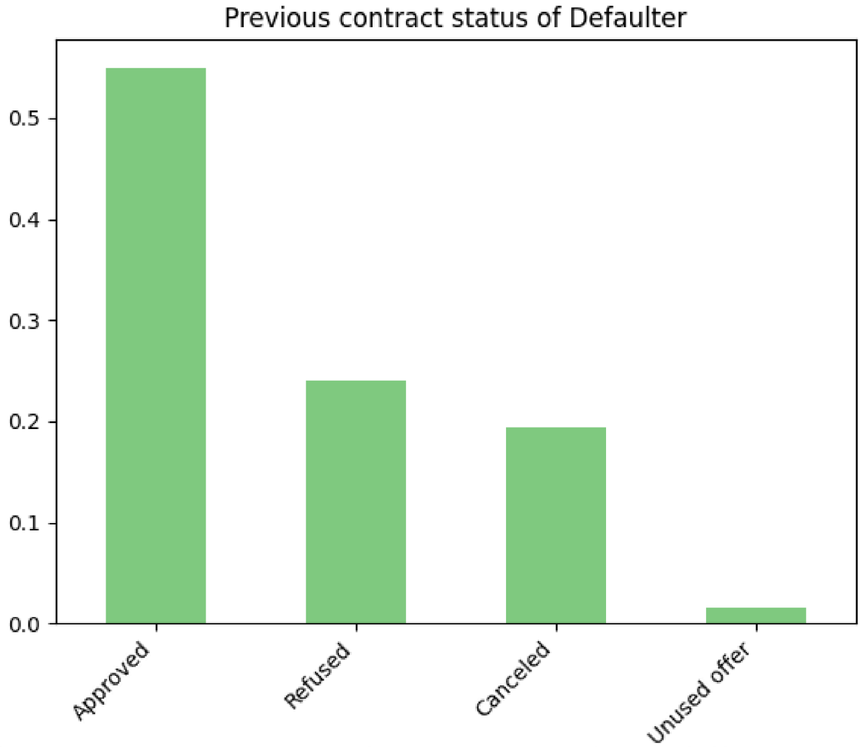




Although the number of applications under 20,000 is smaller than those between 20,000 and 40.000, this group is also safer as the default risk is reasonable. The default risk decreases with increasing annuity amounts.

Contract Status

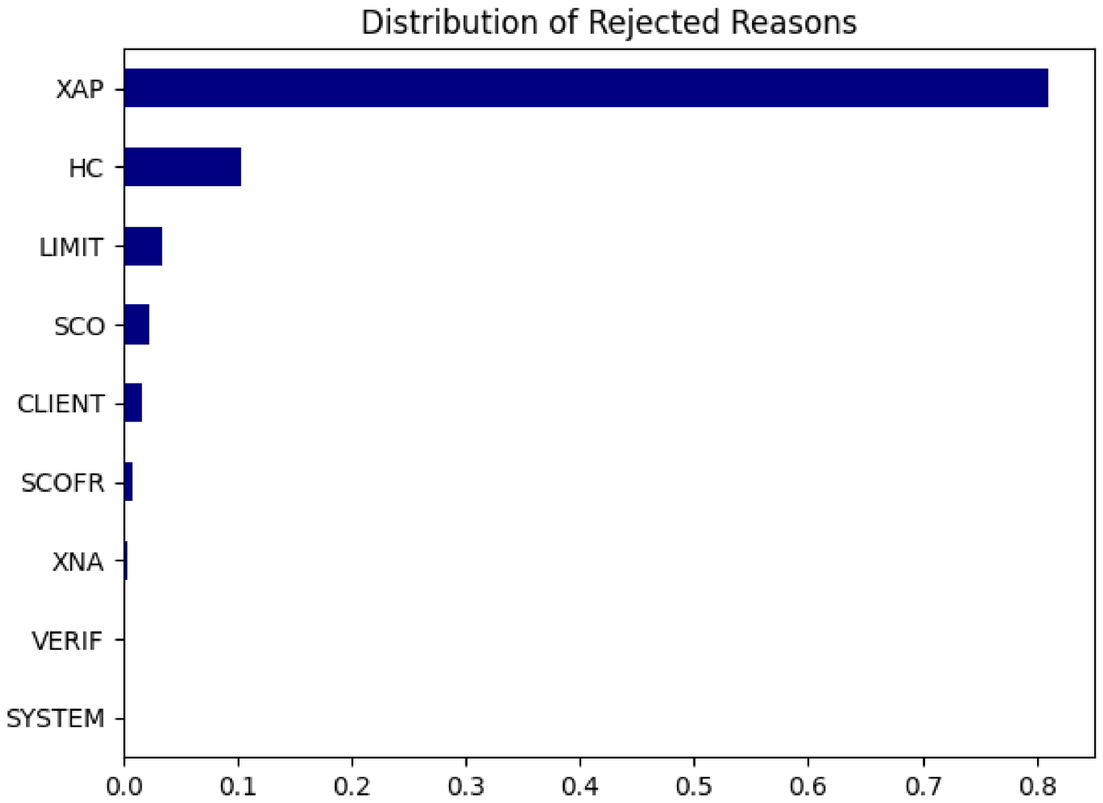


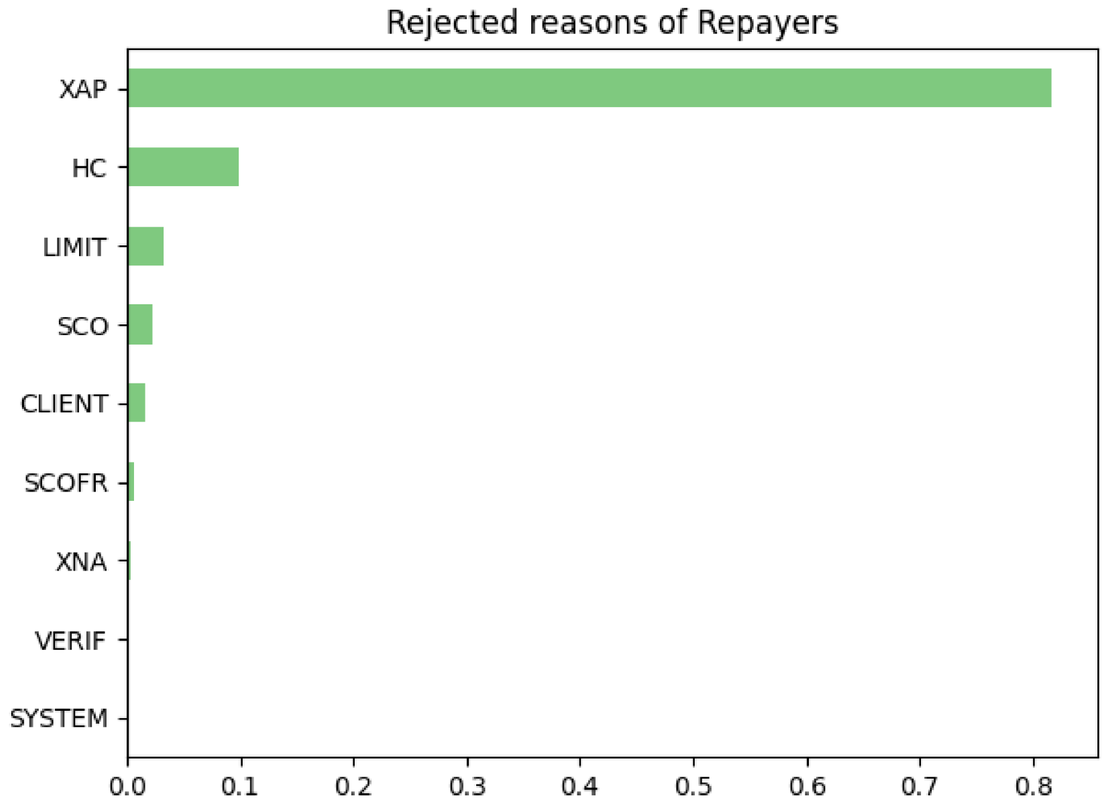


- 62.7% of clients who have been approved before return to continue to apply, while more than 30% of returned clients were rejected application previously.
- More than 50% of clients who have been approved for previous applications still face payment difficulties.

Univariate and Bivariate analysis

Rejected Reasons

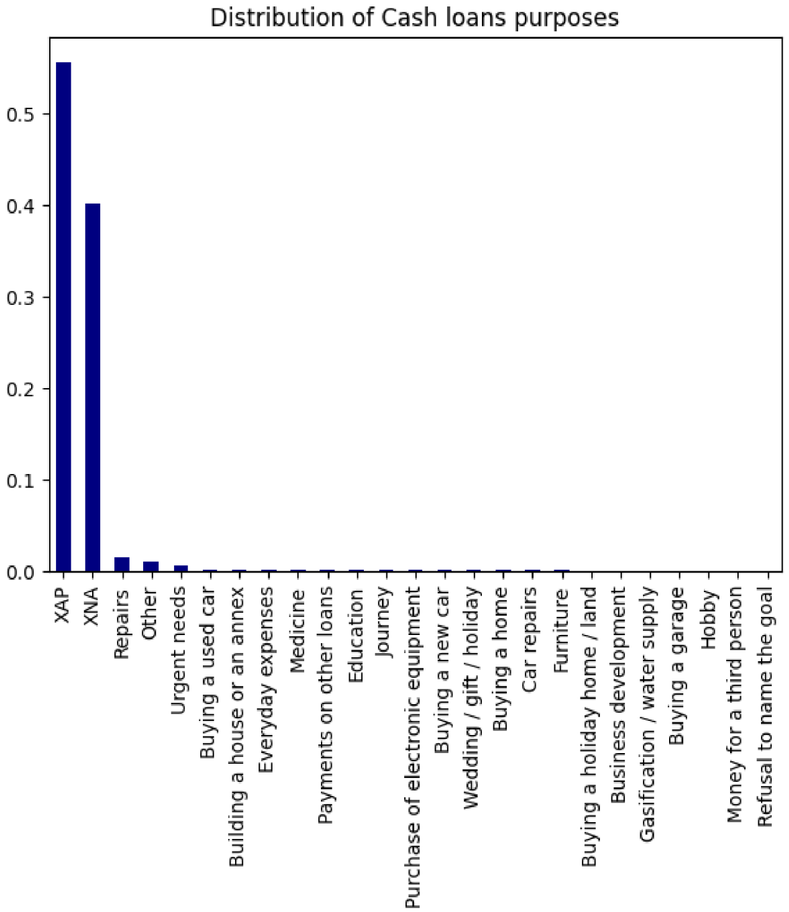


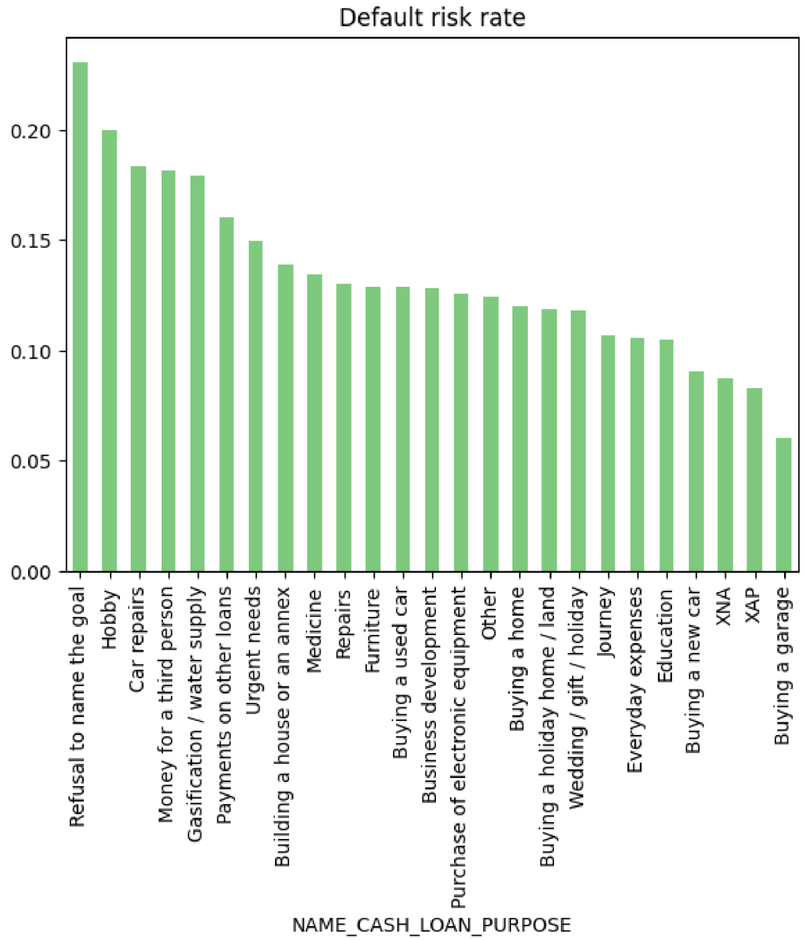


- Among the rejected applications, XAP is the most common reason.
- In the top rejected reasons of repayers, more than 80% of applications are rejected for XAP reasons. This may also indicate that, although repayers have a good payment history, they may still have difficulty meeting the requirements of the applications and may therefore be rejected.

Univariate and Bivariate analysis

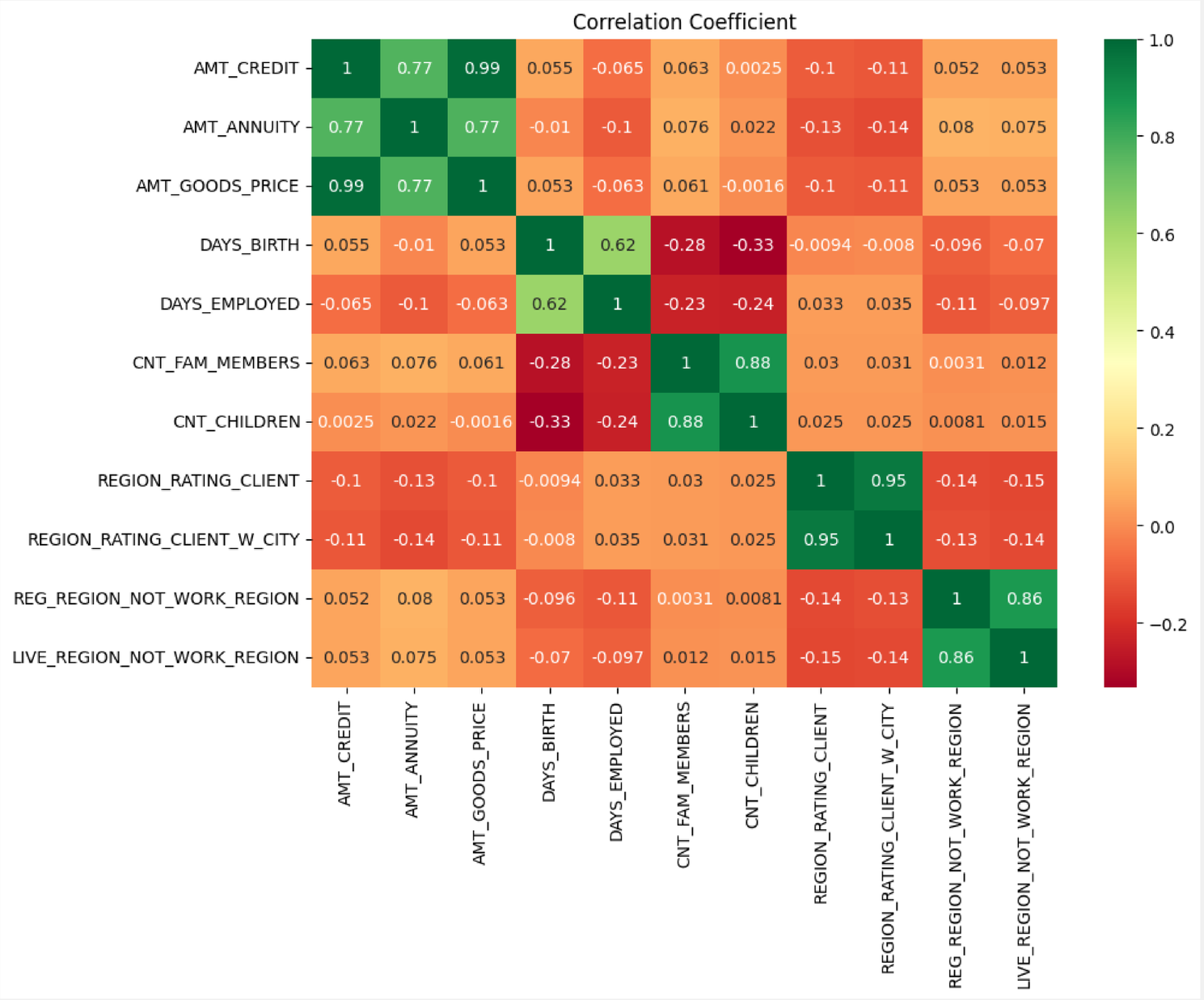
Cash Loan Purposes





- XAP and XNA are the most common purposes of applications.
- Refusal to name the goal and hobby purpose have a high default risk. It indicates that clients who don't have a specific cash loan purpose or are borrowing just for their hobby are likely to default because of the unclear information.

Correlation and Multivariate analysis



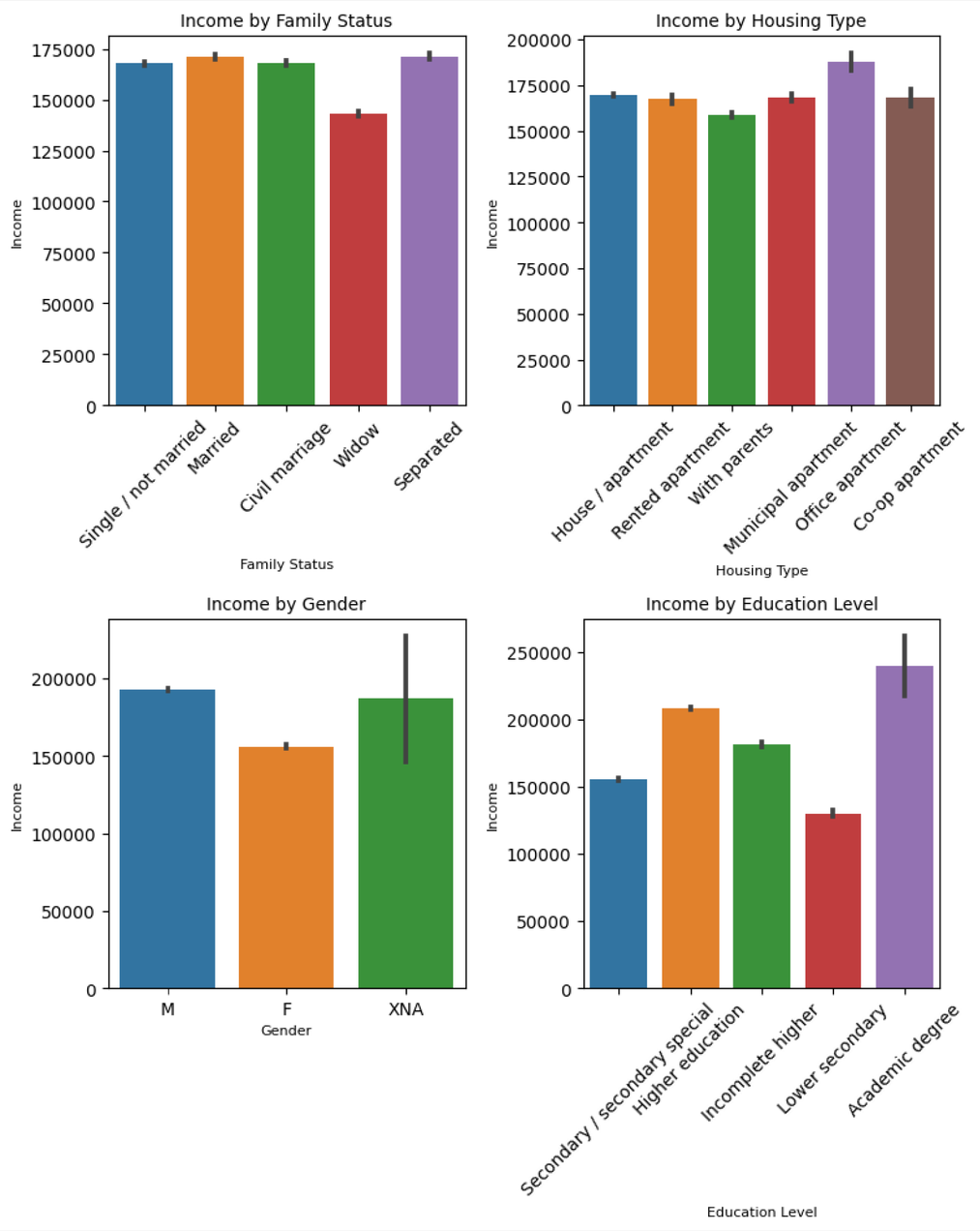
Numerical variables

Correlation and Multivariate analysis

Correlation observations	Connect with Target variable
AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE are high correlation together. It explains that clients who apply for larger amounts of credit also need to pay larger annuity amounts and are purchasing more expensive goods.	**It is secure for the bank because according to our analysis, the applicants for the high credit limit and annuity amount have a minimal likelihood of defaulting.**
DAYS_BIRTH and DAYS_EMPLOYED are high correlation. It explains that younger clients are more likely to be recently employed, while older clients may have a longer employment history.	**It is secure for the bank because according to our analysis, clients with old age and long employment periods have a minimal likelihood of default.**
CNT_FAM_MEMBERS and CNT_CHILDREN are hight correlation. This relationship may indicate that clients with larger families are more likely to have more children.	**The bank is at danger because according to our analysis, clients with a large number of children and family members are more likely to default on their loans.**
REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_CITY are hight correlation. It means clients live in high rating region and high rating city as well.	**The bank is at danger because according to our analysis, clients living in the higher rating region have a higher default risk.**
REG_REGION_NOT_WORK_REGION and LIVE_REGION_NOT_WORK_REGION are high correlation.	**It doesn't pose detrimental effects on the bank as we analysed.**

Numerical variables

Correlation and Multivariate analysis



- Family Status: Widows have the lowest income total, while other types are the same.
- Housing type: rented apartment and with parents who have low income. It proves that the financial ability of these clients isn't high, and that makes them have a high default risk.
- Gender: A male has a higher income than a female. Also, they have a higher default risk than women. It may be because females can manage their income better than males.
- Education level: A high education level can help people earn more money.

Numerical variables

Conclusion and Suggestion

Summary

According to the analysis, we define the characteristics of paid-on-time clients and likely-to-default clients as:

Characteristic	Paid on time	Likely to default
Contract Type	Resolving loans	Cashloans
Gender	Female	Male
Age	30 - 50 years old	Others
Car Owners	Yes	No
Realty Owners	Same probability	
Income Type	Others	Unemployed and maternity leave
Education Type	Others	Secondary or Secondary special / Low secondary
Housing Type	Others	Rented aparments / With parents
Family Status	Married/ Widow	Others
Occupation Type	Manager, Core staff	Laborers, Low-skill laborers, sales staff, drivers

Characteristic	Paid on time	Likely to default
Num of children	Low num of children	High num of children
Num of Family members	Low num of fam members (ideally = 2.0)	High num of fam members
Employed Years	High years of experience	Low years of experience
Region Rating	Low rating	High rating
Income Amount	High income	Low income
Credit Amount	Others	250,000 - 750,000
Annuity Amount	Others	20,000 - 40,000
Historical application	• Approved previously application. • Purpose: Have a specific purpose	• Cancelled/Rejected previous application. • Purpose: Refusal to name the goal and hobby purpose

Conclusion

Based on the summary, we have a big picture of the client's profile to verify the application. Also, banks should be more careful with some paradoxes.

01 Clients with low financial capacity are at high risk of default, as are those with low education and job levels.

02 Clients have many dependent persons are at high risk of default as they have more family expenses.

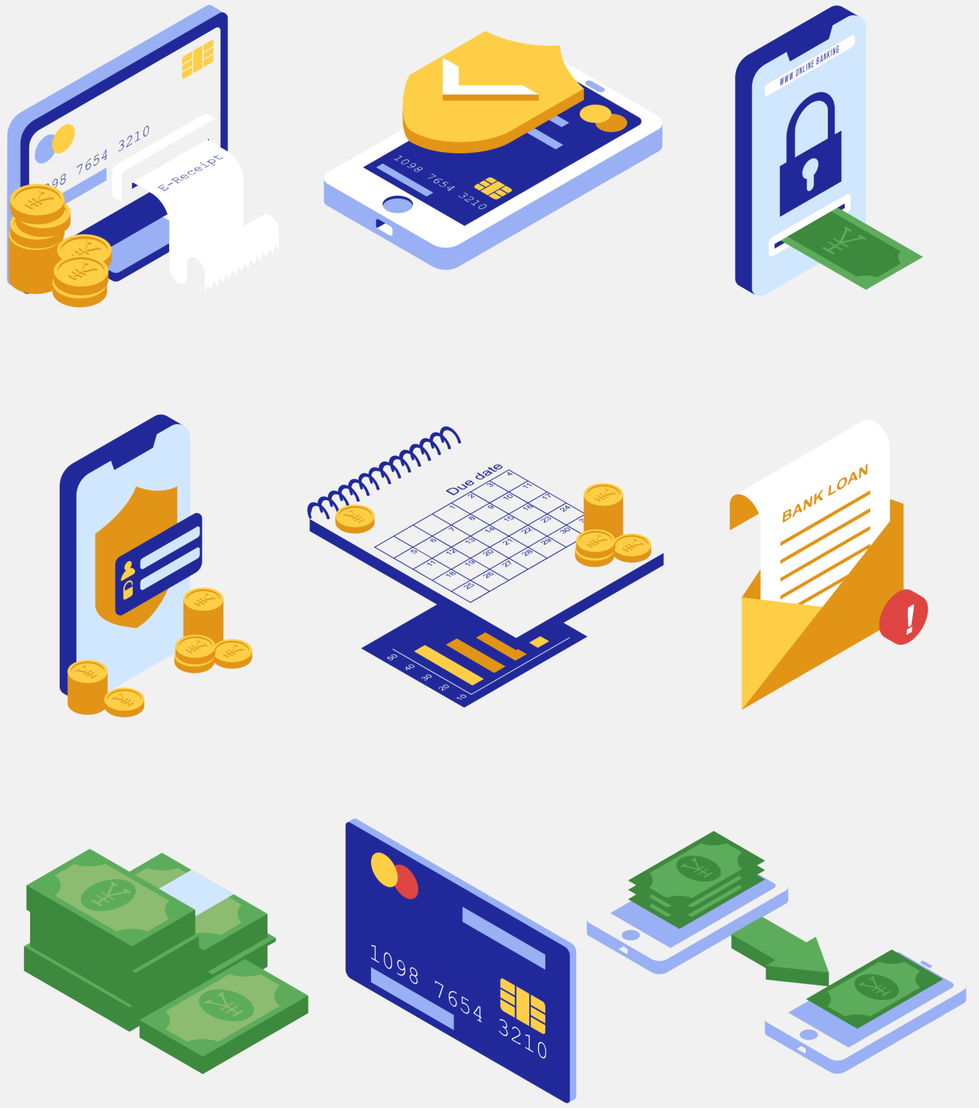
03 Clients living in urban areas may have a high standard of living, but that doesn't mean they can afford the loans.

04 More than 50% of clients who have been approved for previous applications still face payment difficulties. This is a risk to the bank.

05 More than 80% of clients have rejected the previous application, but they can still be potential clients because they can pay on time. The bank can't miss these clients.

06 The bank attracted more clients who were interested in cash loans, but they should check the application more carefully as it may contain more clients likely to default.

THANK YOU!


E-Receipt
1098 7654 3210
WWW.ONLINE BANKING
Due date
BANK LOAN
1098 7654 3210