CREDIT RISK ANALYTICS

Team Members:
Naga Sri Vyshnavi Kanuri
Shatrughan Singh Gusain
G Venkata Sai Charan
Arshak Roshan C
Nagesha S
Akanksha Mishra
Dharna Bhavani
Radha Barsaiya
Keerthana Senthil Raja (Co-Team Lead)
Mariya Saji (Team-Lead)

Problem Statement

- This project focuses on Exploratory Data Analysis (EDA) in credit risk analytics for a consumer finance company.
- The main goal is to identify patterns and key variables that predict loan default.
- The insights gained will inform better decision-making in loan approvals, risk reduction, and portfolio optimization.
- The project aims to provide concise and informative results to support informed choices in credit risk analytics.

Introduction

Understanding the Challenge

- Within the dynamic realm of consumer finance, lending institutions encounter a formidable challenge: the ability to discern between loan applicants capable of responsible repayment and those at risk of default.
- A significant complication arises from the absence of robust credit histories among certain applicants, necessitating precise risk assessment procedures.
- This challenge not only affects financial stability but also leads to missed business opportunities for lending companies.

❑ The Imperative of Risk Assessment

- Risk analytics, a pivotal facet of modern financial services, is instrumental in navigating the intricate landscape of lending.
- It facilitates the data-driven decision-making process essential for minimizing financial risks and optimizing loan portfolios.

Steps Included In The Project

1. Data Understanding
2. Missing Data Handling
3. Outlier Identification
4. Data Imbalance Analysis
5. Univariate Analysis
6. Bivariate Analysis
7. Top Correlations for Risk Assessment

1. Data Understanding

We have 2 datasets to work on:

1. Application_data
    - It contains all the information of the client at the time of application. The data is about whether a client has payment difficulties or not.
    - Data Structure: (307511 rows, 122 columns)
    - No. of Categorical Variables: 16
    - No. of Numerical Variables: 106
    - Target/Output Variable: TARGET

2. Previous_application

- It contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Canceled, Refused or Unused offer.
- Data Structure: (1048575 rows,37 columns)
- No. of Categorical Variables: 16
- No. of Numerical Variables: 21
- Target/Output Variable: NAME_CONTRACT_STATUS

2. Missing Data Handling

- **Merging the Data**
  - The first step involved identifying common columns between the two datasets, 'application_data' and 'previous_application_data.'
  - We calculated the number of unique clients in both datasets. This information helps us categorize clients as retained, new, or churn.
  - We merged the datasets using the "SK_ID_CURR" column via a left join. This process combines client information across both datasets.

❑ Handling the Missing data

- Columns with over 60% missing data are dropped.
- Numeric columns having above 0% to 60% missing data are imputed with the mean.
- Categorical columns are below 40% are imputed with the mode.
- We also created some special category in columns having missing values greater than mode.

Outcomes

```
common_col
```

```
['SK_ID_CURR',
 'NAME_CONTRACT_TYPE',
 'AMT_ANNUITY',
 'AMT_CREDIT',
 'AMT_GOODS_PRICE',
 'WEEKDAY_APPR_PROCESS_START',
 'HOUR_APPR_PROCESS_START',
 'NAME_TYPE_SUITE']
```

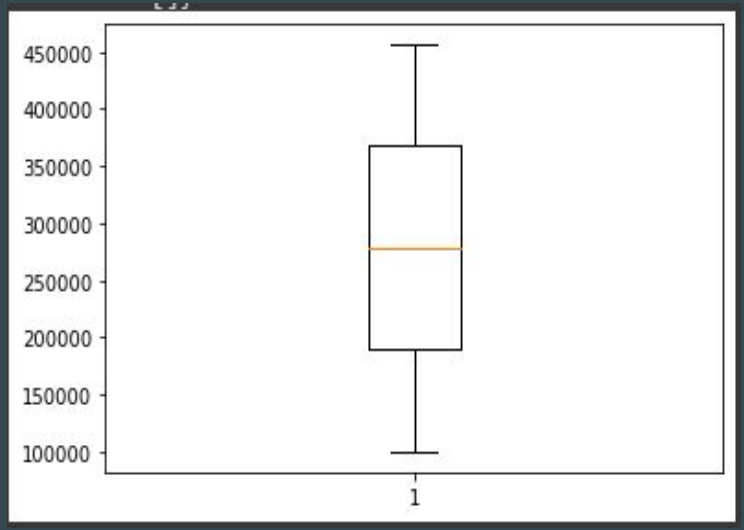
- Retain clients : 291057
- New clients : 16454
- Shape after merging : (1430155, 158)
- Shape after handling missing value : (1430155, 137)

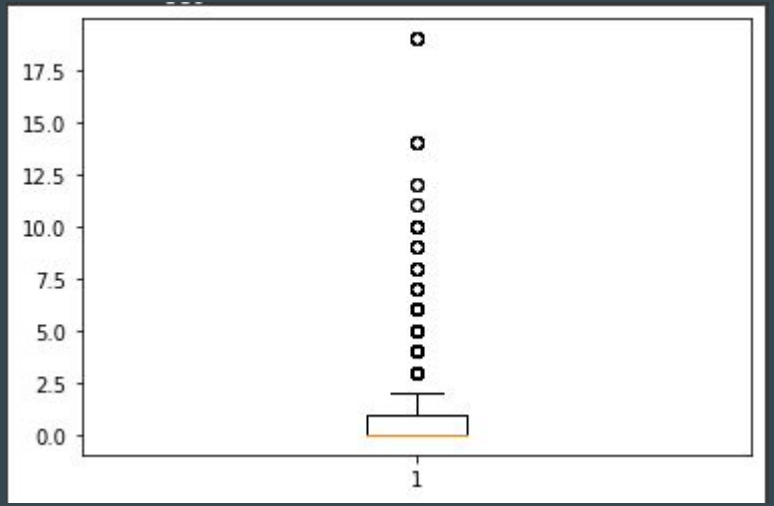
3. Outlier Handling

- **Steps Involved**
  - Calculated the correlation matrix.
  - For handling outliers, we used log transform for positive numeric columns.
  - If columns are having negative values we found the outliers, but didn't transformed it.
  - Analyzed boolean column to have binary values (0 or 1).
  - Investigated columns with numeric values exclusively between 0 and 1.
  - Examined categorical columns and checked the number of unique values in each.

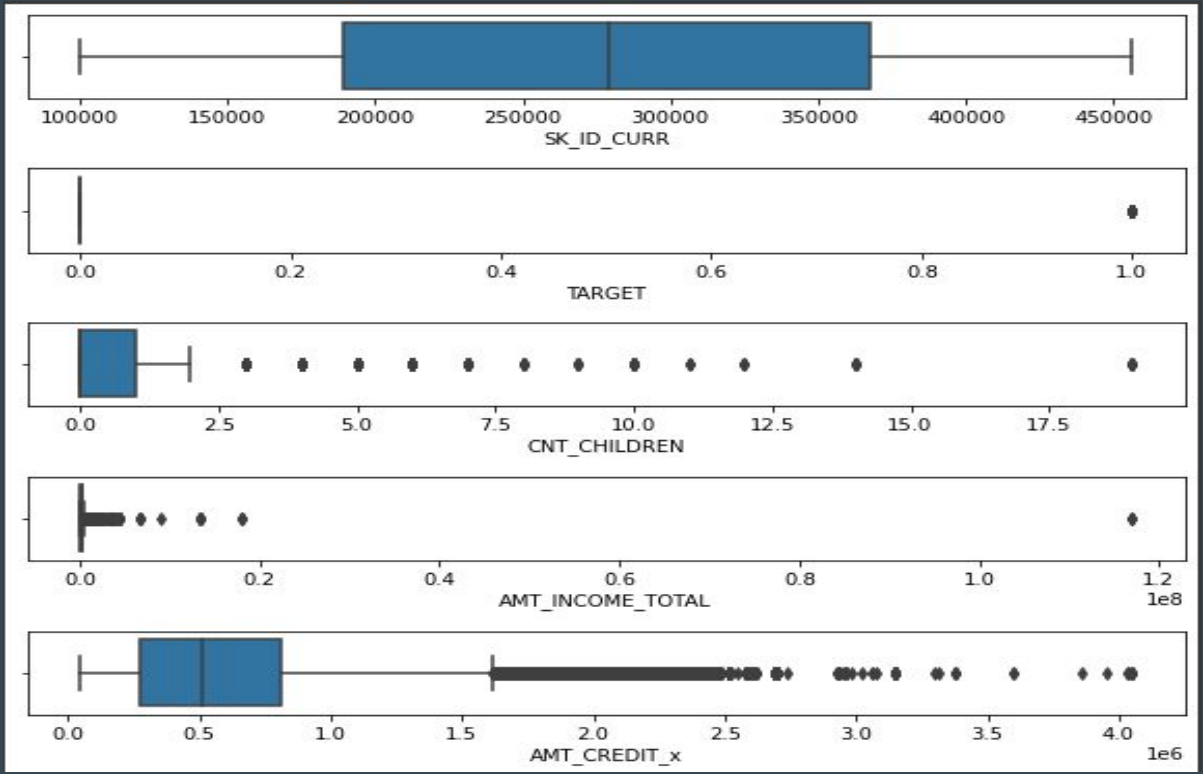
Outcomes

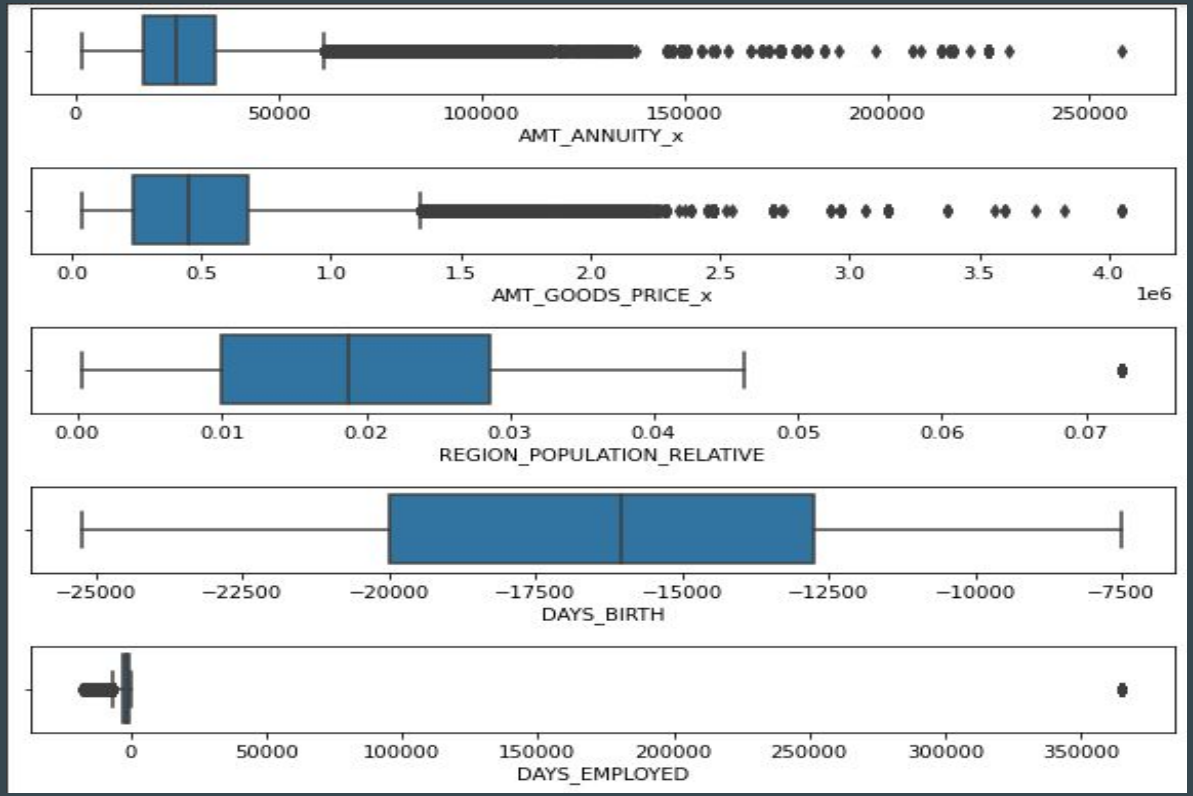
- SK_ID_CURR has no outliers and CNT_CHILDREN contains outliers.

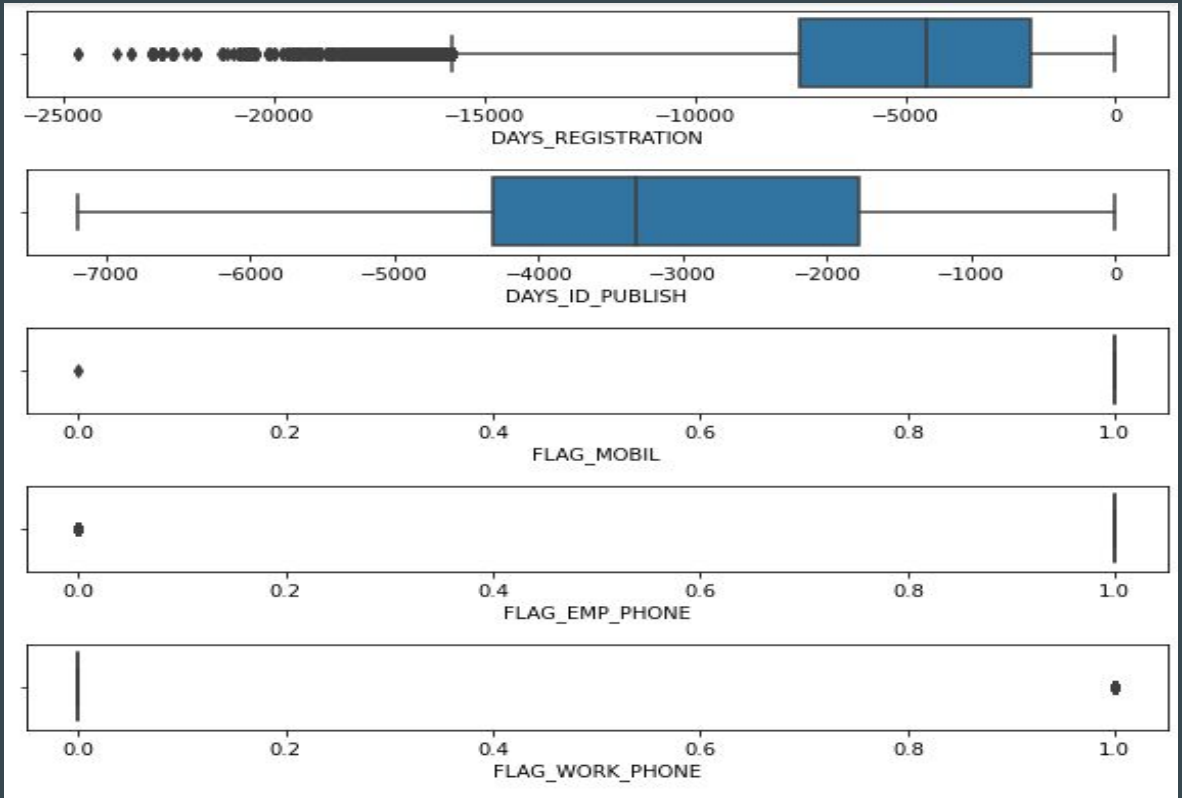


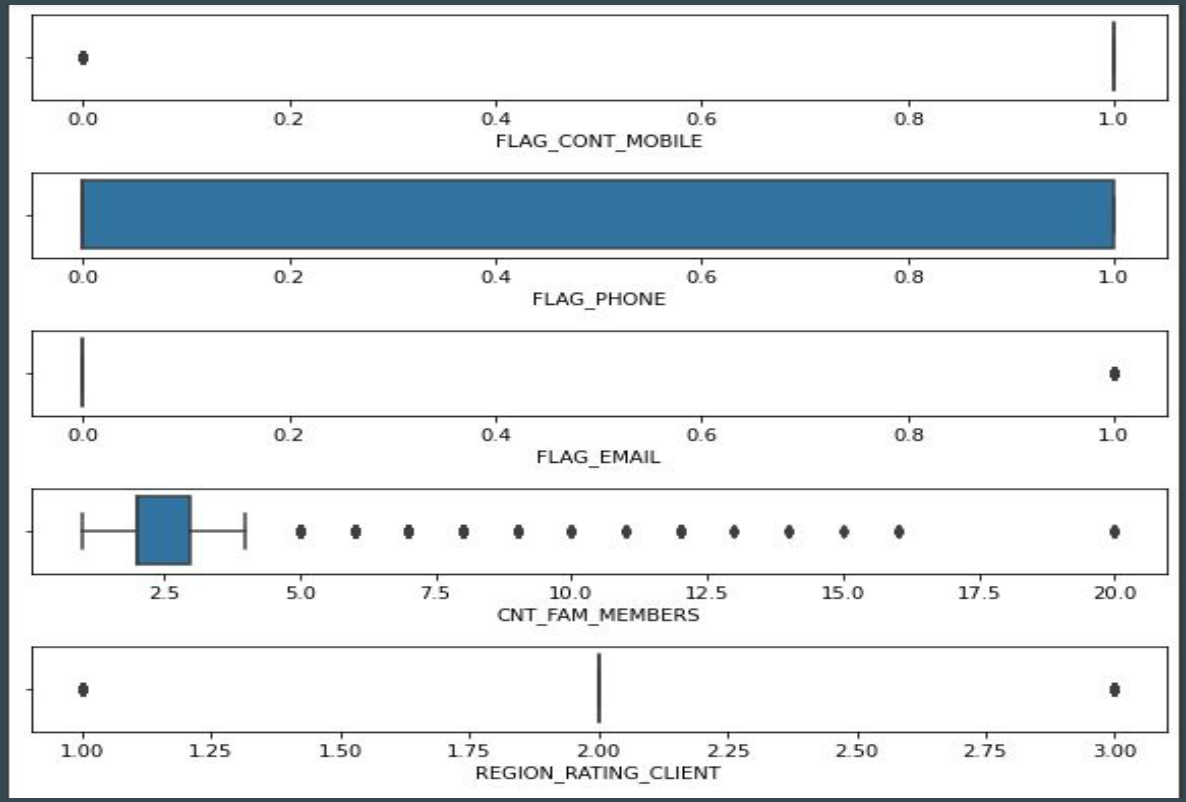


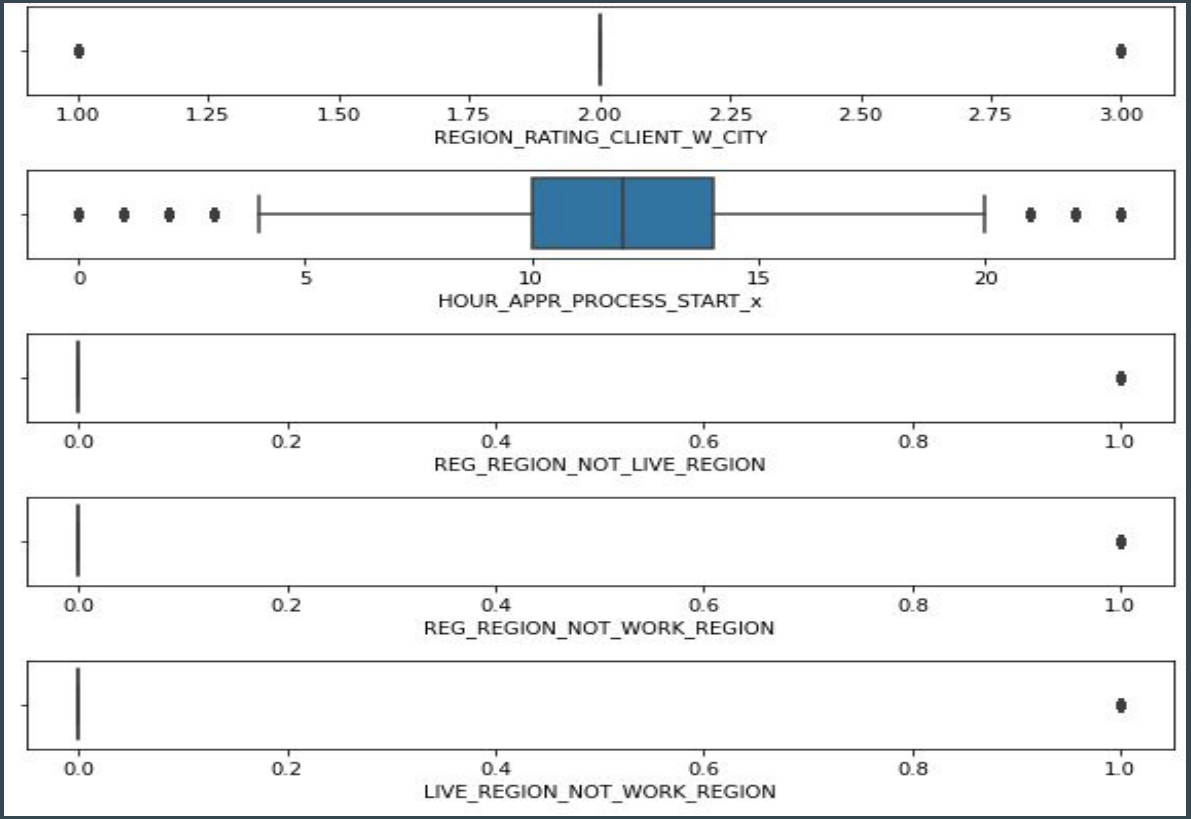
- Outlier detection in numerical columns

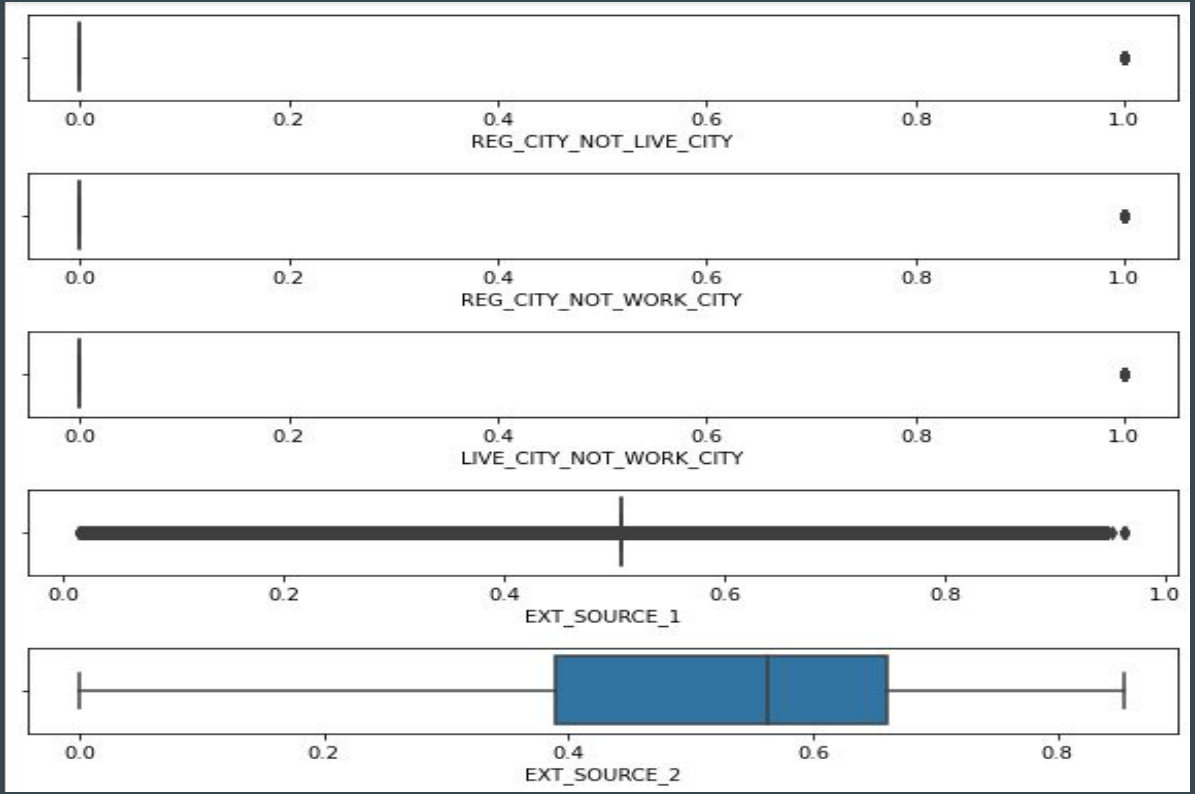



0
50000
100000
150000
200000
250000
AMT_ANNUITY_x
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6
AMT_GOODS_PRICE_x
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
REGION_POPULATION_RELATIVE
−25000
−22500
−20000
−17500
−15000
−12500
−10000
−7500
DAYS_BIRTH
0
50000
100000
150000
200000
250000
300000
350000
DAYS_EMPLOYED

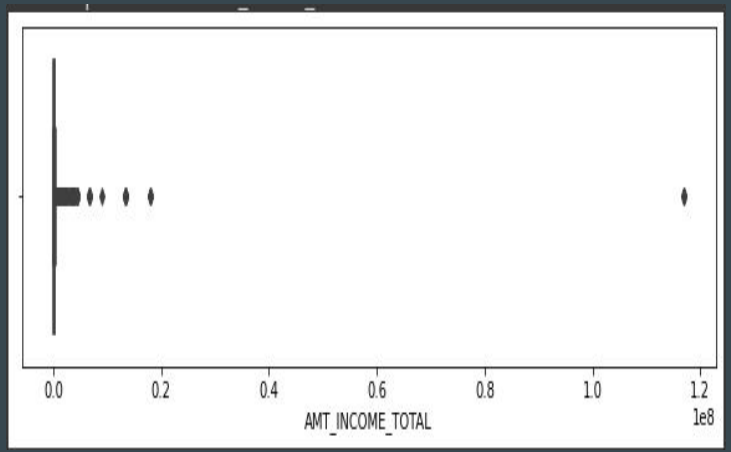

−25000
−20000
−15000
−10000
−5000
0
DAYS_REGISTRATION
−7000
−6000
−5000
−4000
−3000
−2000
−1000
0
DAYS_ID_PUBLISH
0.0
0.2
0.4
0.6
0.8
1.0
FLAG_MOBIL
0.0
0.2
0.4
0.6
0.8
1.0
FLAG_EMP_PHONE
0.0
0.2
0.4
0.6
0.8
1.0
FLAG_WORK_PHONE

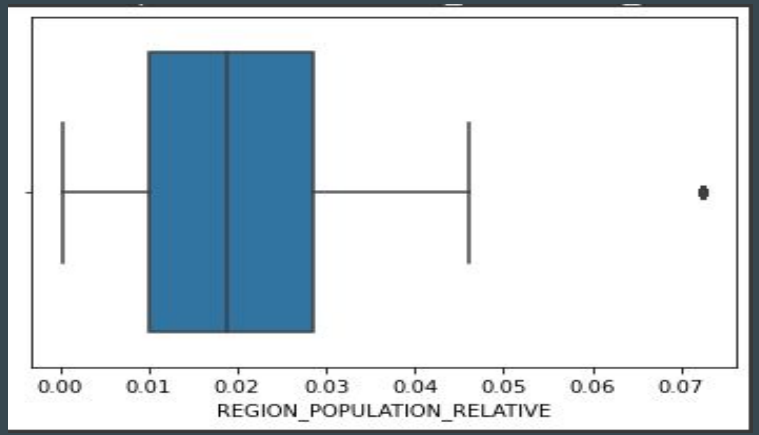

0.0
0.2
0.4
0.6
0.8
1.0
FLAG_CONT_MOBILE
0.0
0.2
0.4
0.6
0.8
1.0
FLAG_PHONE
0.0
0.2
0.4
0.6
0.8
1.0
FLAG_EMAIL
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
CNT_FAM_MEMBERS
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
REGION_RATING_CLIENT


1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
REGION_RATING_CLIENT_W_CITY
0
5
10
15
20
HOUR_APPR_PROCESS_START_x
0.0
0.2
0.4
0.6
0.8
1.0
REG_REGION_NOT_LIVE_REGION
0.0
0.2
0.4
0.6
0.8
1.0
REG_REGION_NOT_WORK_REGION
0.0
0.2
0.4
0.6
0.8
1.0
LIVE_REGION_NOT_WORK_REGION

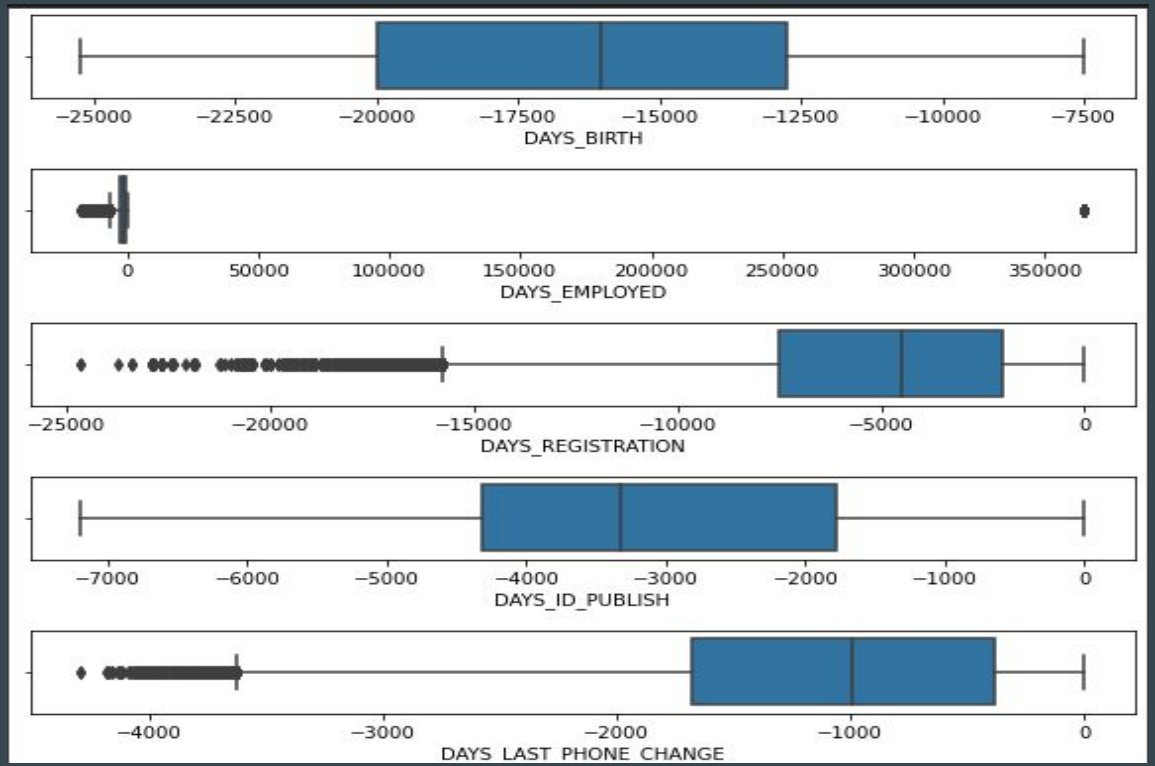

0.0
0.2
0.4
0.6
0.8
1.0
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
EXT_SOURCE_1
EXT_SOURCE_2

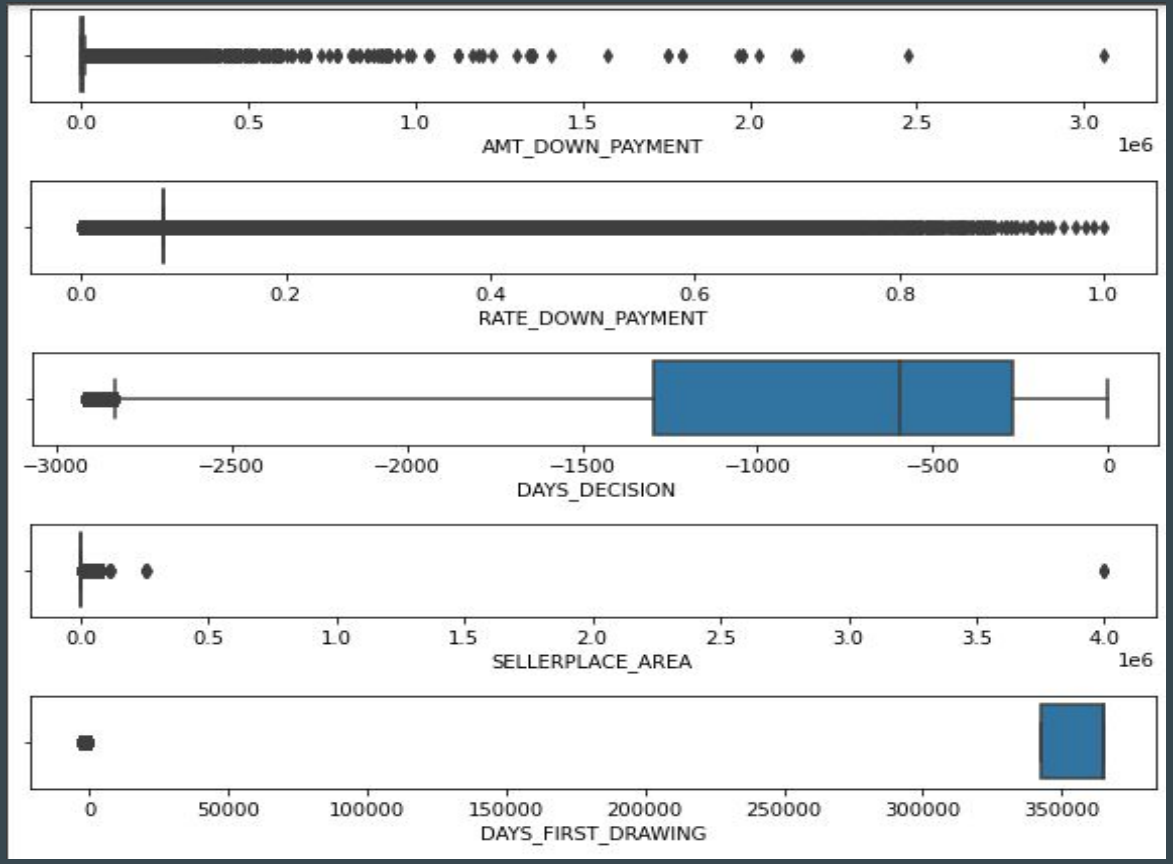
- There are 106 numerical columns so it's not possible to place all plots here.
- There are so many outliers in AMT_INCOME_TOTAL and not much in REGION_POPULATION_RELATIVE

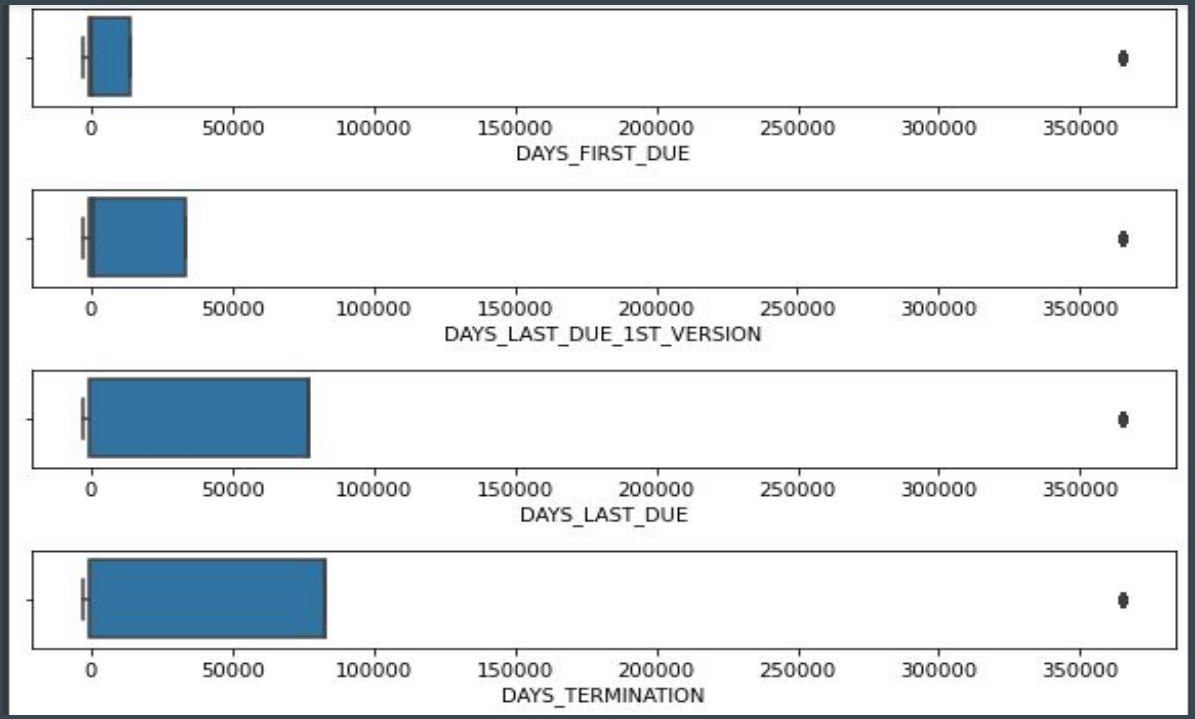




- Outliers in negative columns

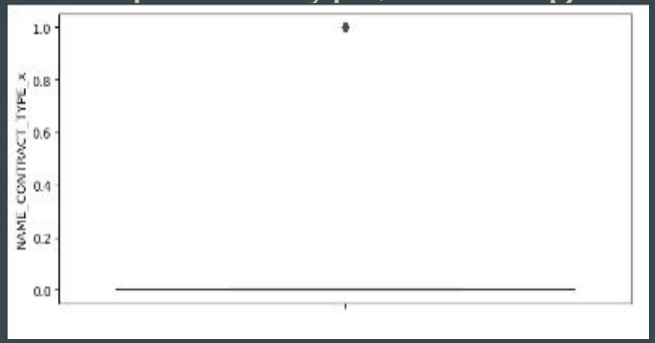


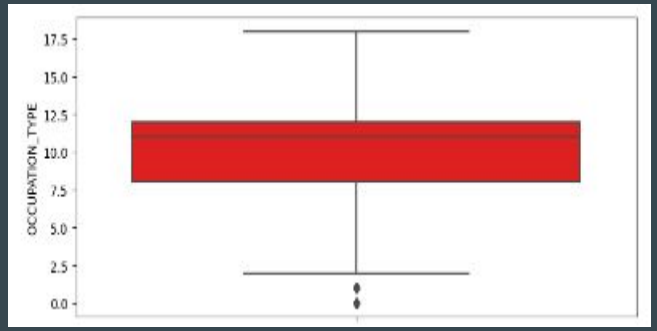

0.0
0.5
1.0
1.5
2.0
2.5
3.0
AMT_DOWN_PAYMENT
1e6
0.0
0.2
0.4
0.6
0.8
1.0
RATE_DOWN_PAYMENT
−3000
−2500
−2000
−1500
−1000
−500
0
DAYS_DECISION
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
SELLERPLACE_AREA
1e6
0
50000
100000
150000
200000
250000
300000
350000
DAYS_FIRST_DRAWING

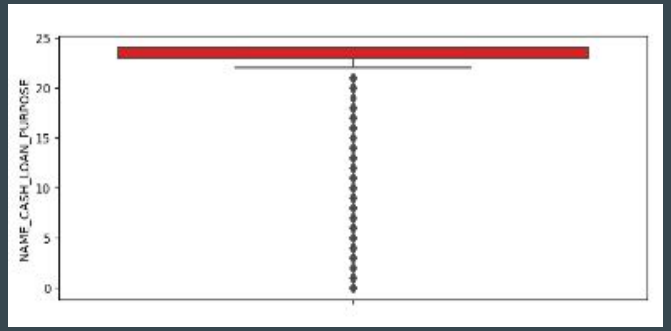


- Not inserted all plots of outliers

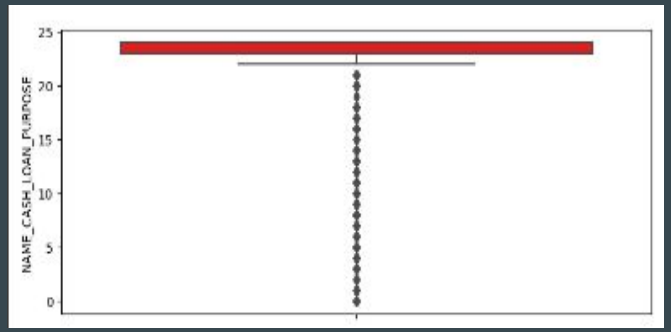
- Outliers in categorical columns, Name_Contract_Type_x, Name_Type Suit_x, Occupation_Type, and Flag_Last_Appl_Per_Contract.

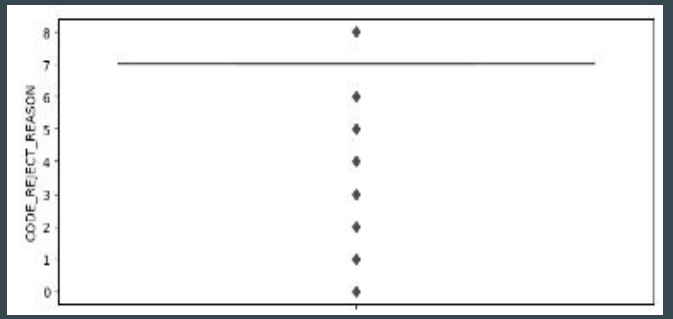


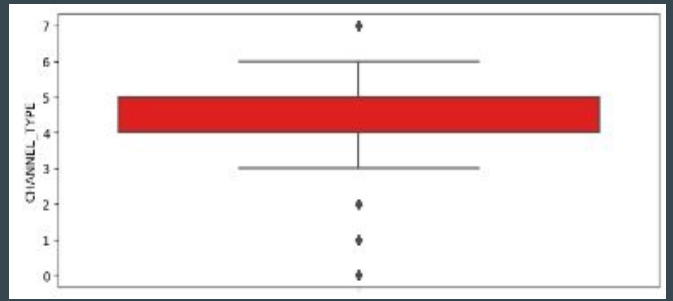




- Name_cash_Loan_Purpose, Code_Reject_Reason, and Channel_Type.

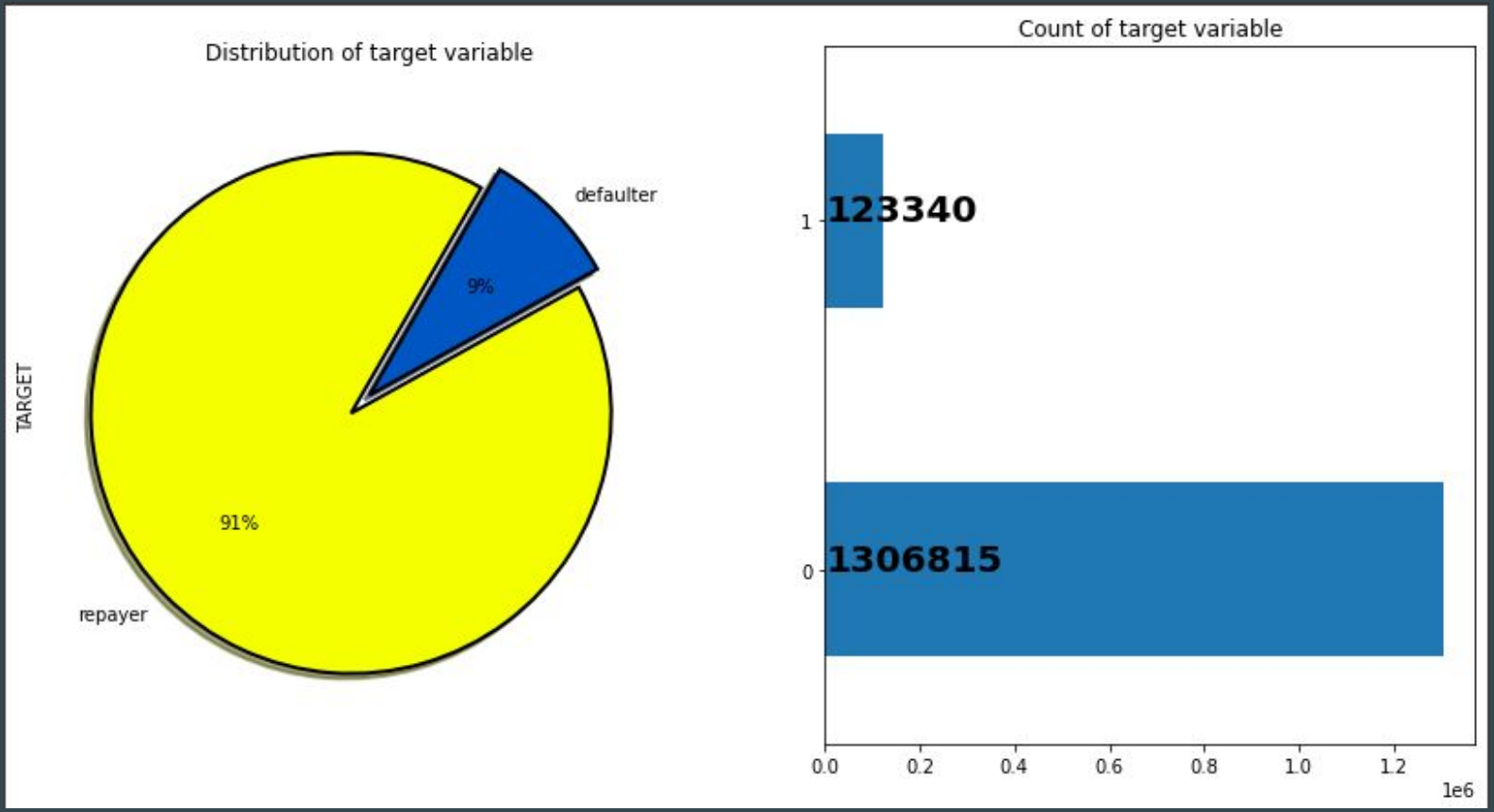






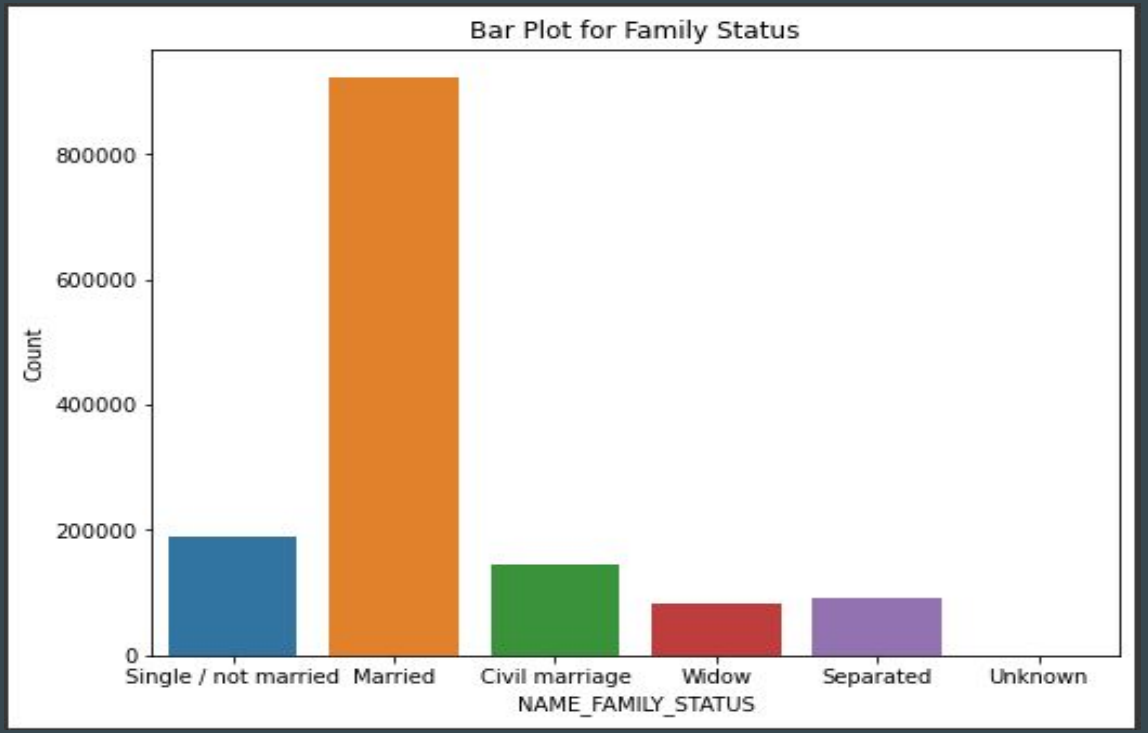
4. Data Imbalance Ratio

- Separated the dataset into two subsets: "Target0" for records with a "TARGET" value of 0 (repayers) and "Target1" for records with a "TARGET" value of 1 (defaulters).
- Calculated the imbalance ratio by dividing the number of records in "Target0" by the number of records in "Target1".
- Plotted the distribution of the "TARGET" variable
- The below visualizations indicate that approximately 9% of the records are defaulters (class 1), while 91% are repayers (class 0), illustrating the class imbalance in the dataset.

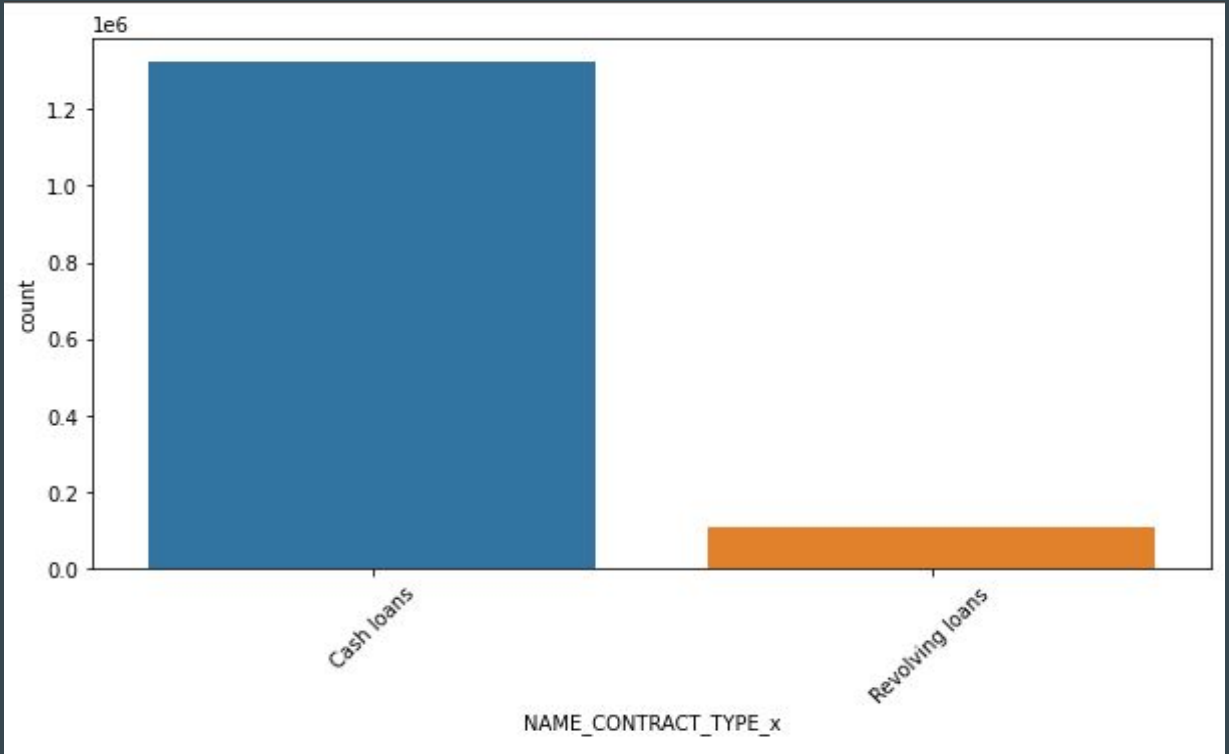

Distribution of target variable
TARGET
defaulter
9%
91%
repayer
Count of target variable
1
123340
0
1306815
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e6

5. Univariate Analysis

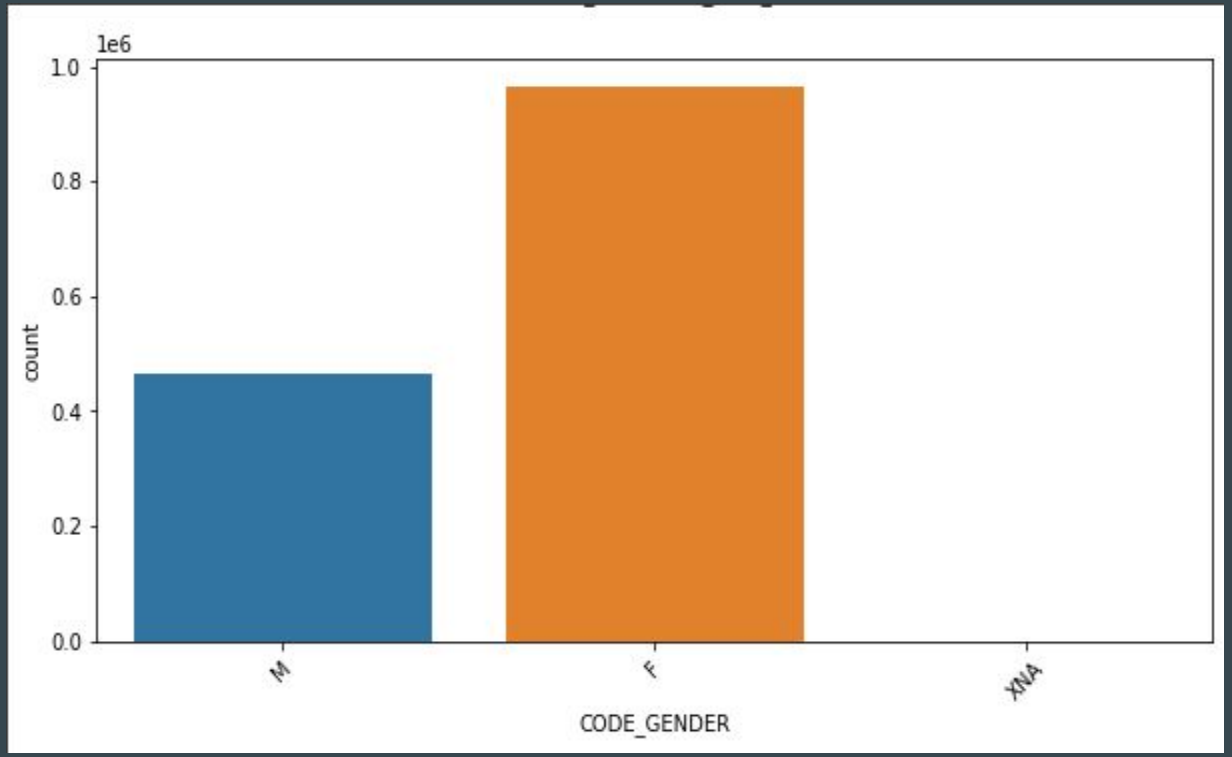
- Univariate analysis is a technique in data analysis that focuses on exploring and understanding a single variable in isolation.
- Analyzed each categorical variable using bar plot.
- Analyzed some numerical variables using boxplot and histogram.
- Found some insights from the analysis.



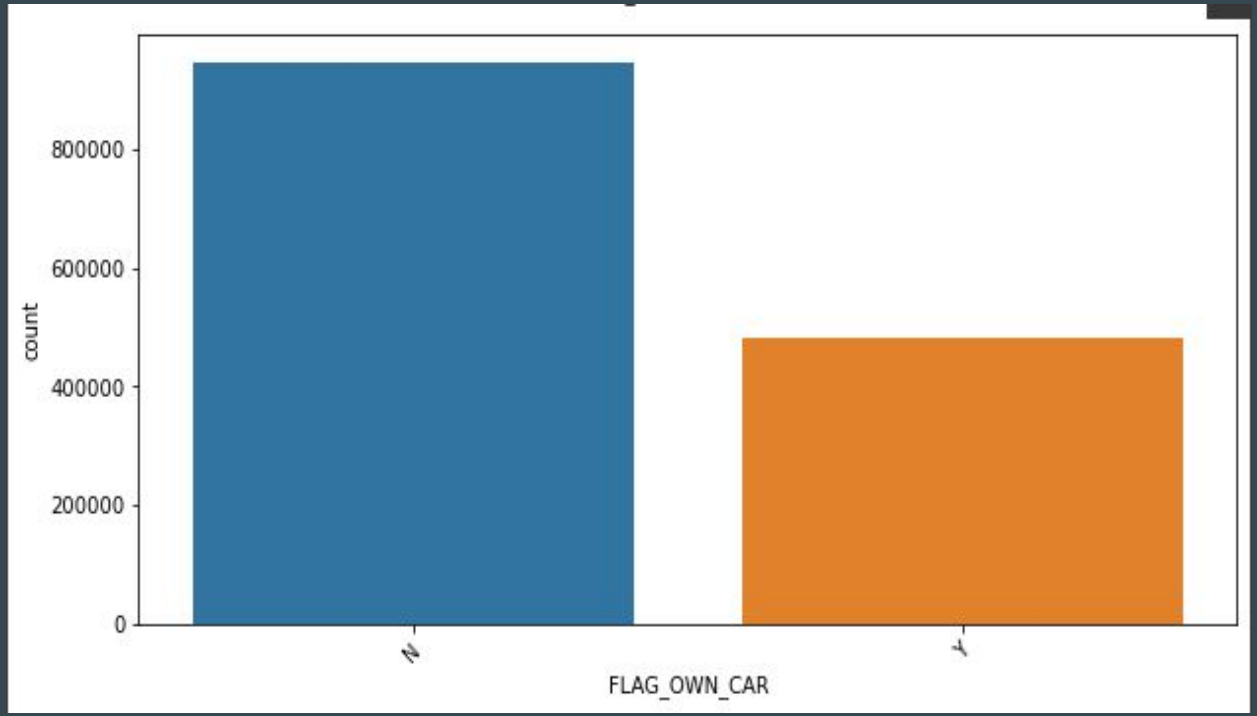
- The majority of applicants have family status as "married".
- Applicants having family status as "Widow", "Separated" are comparatively less.



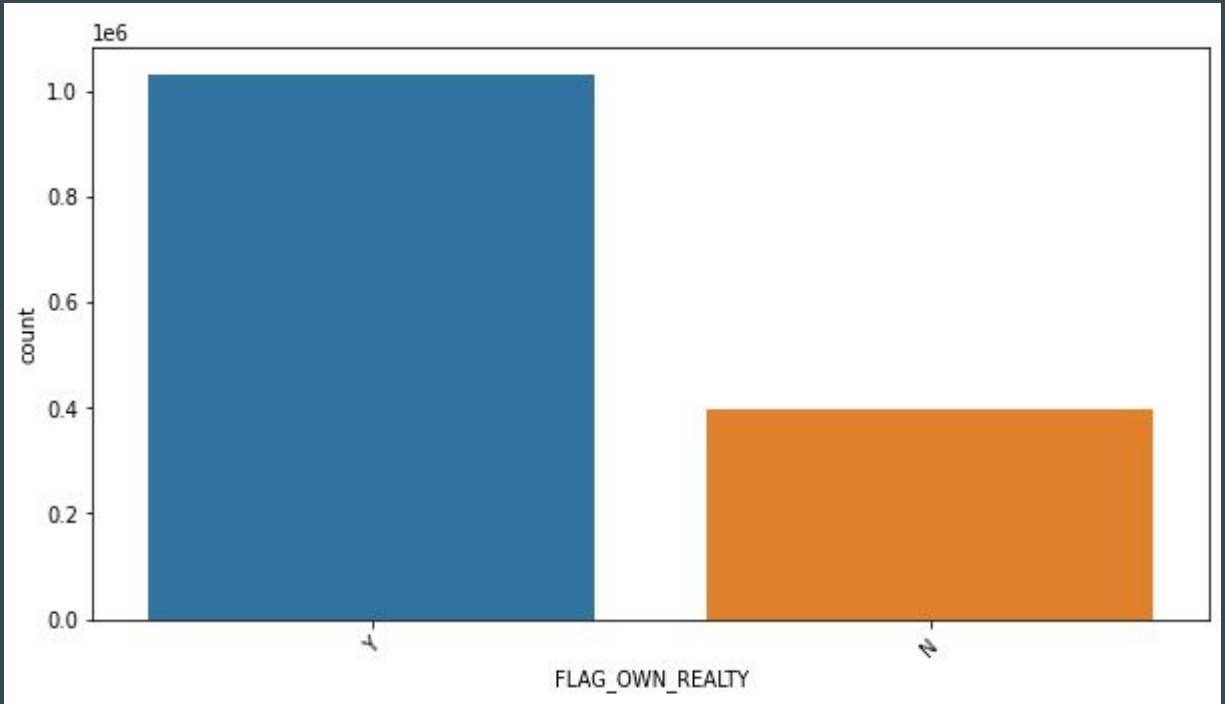
- The majority of applicants are applied for cash loans.



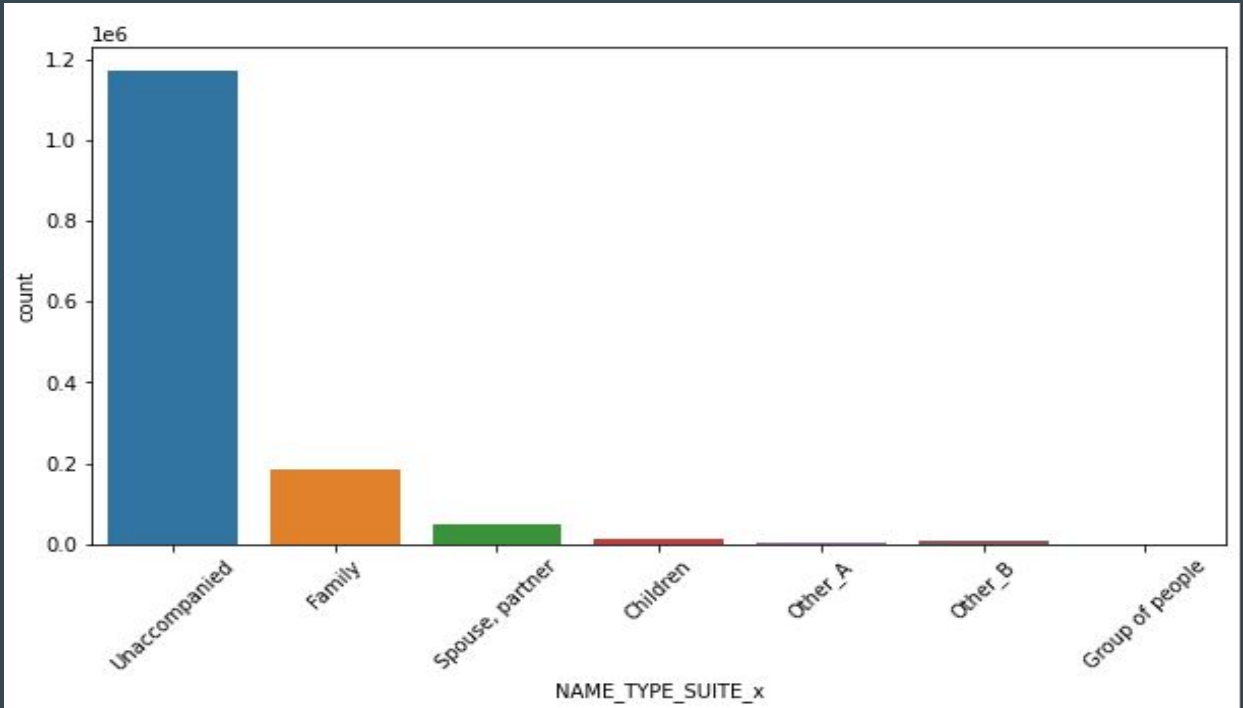
- The majority of applicants are females



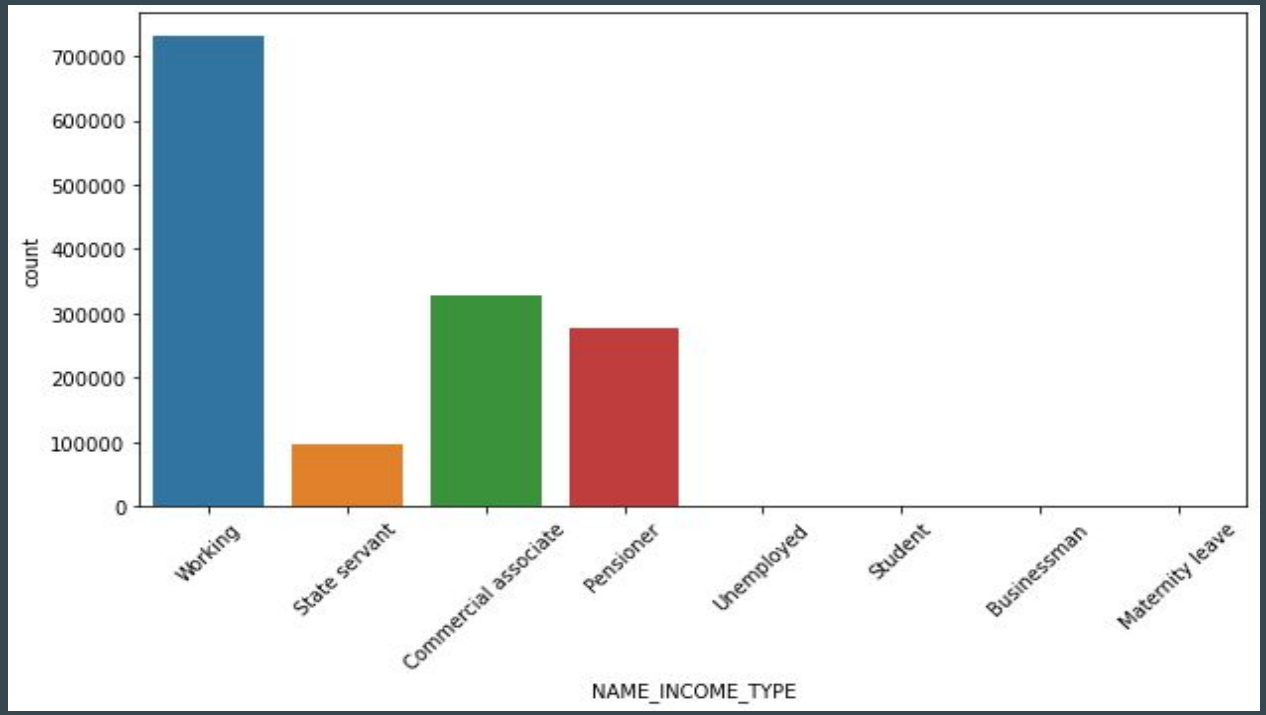
- Majority of applicants doesn’t own car, so there is a chance to getting loan.



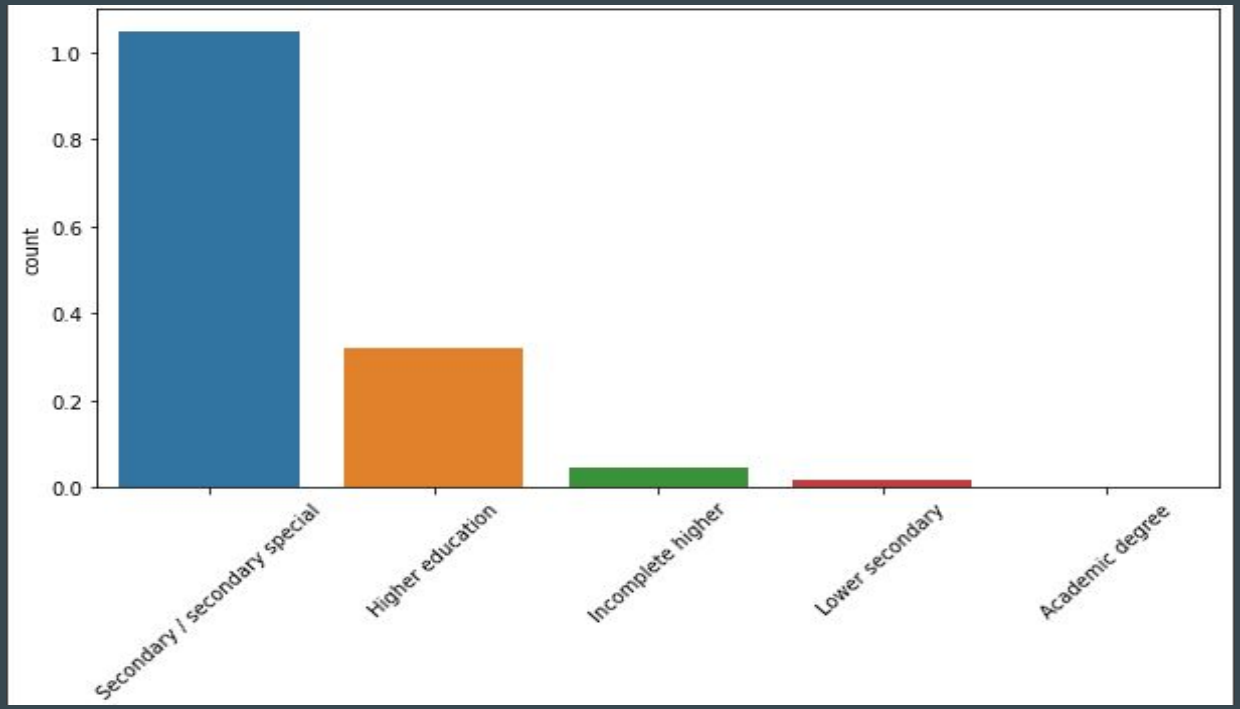
- Majority of the applicants got flag because they own realty.



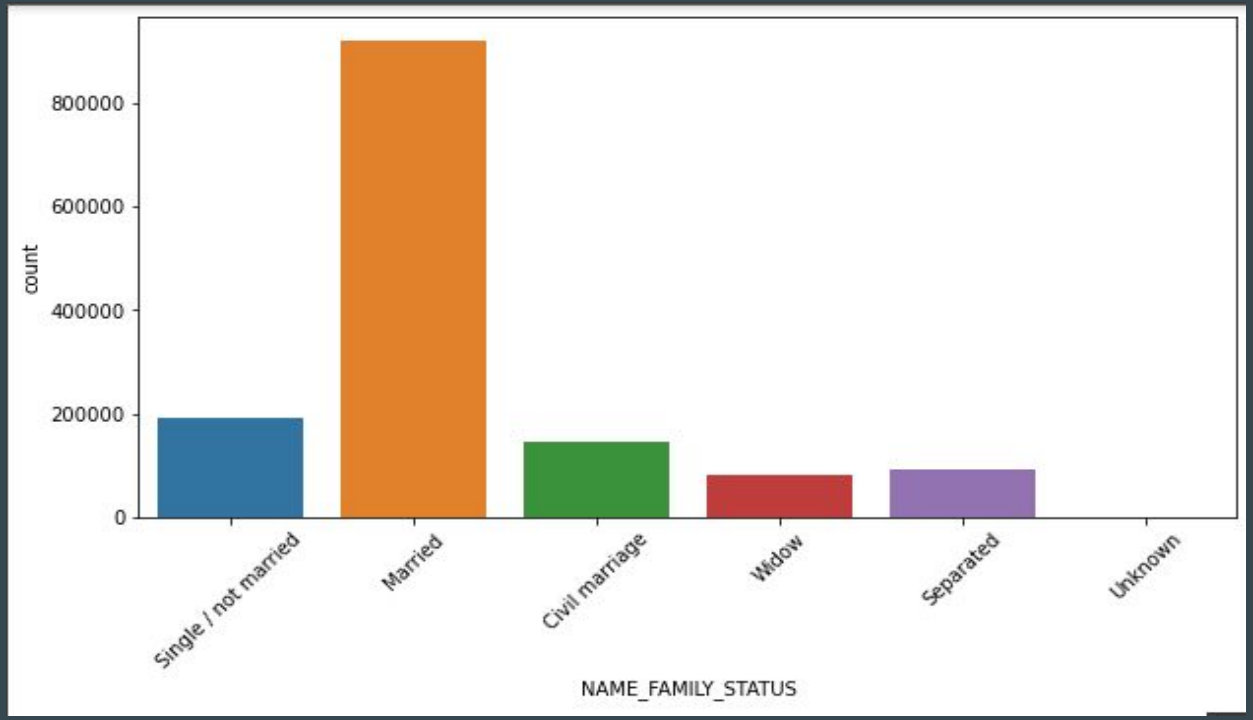
- Majority of the applicants are unaccompanied while they applying for loan.



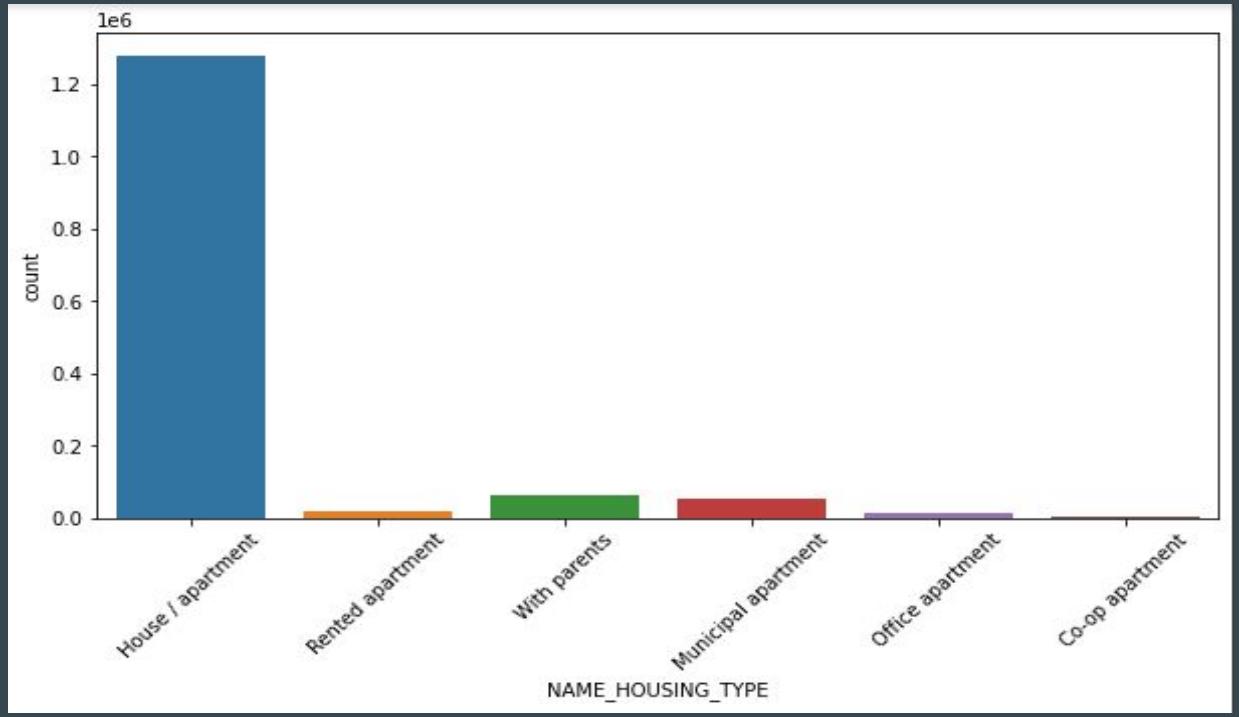
- Majority of the applicant's income type is "working".



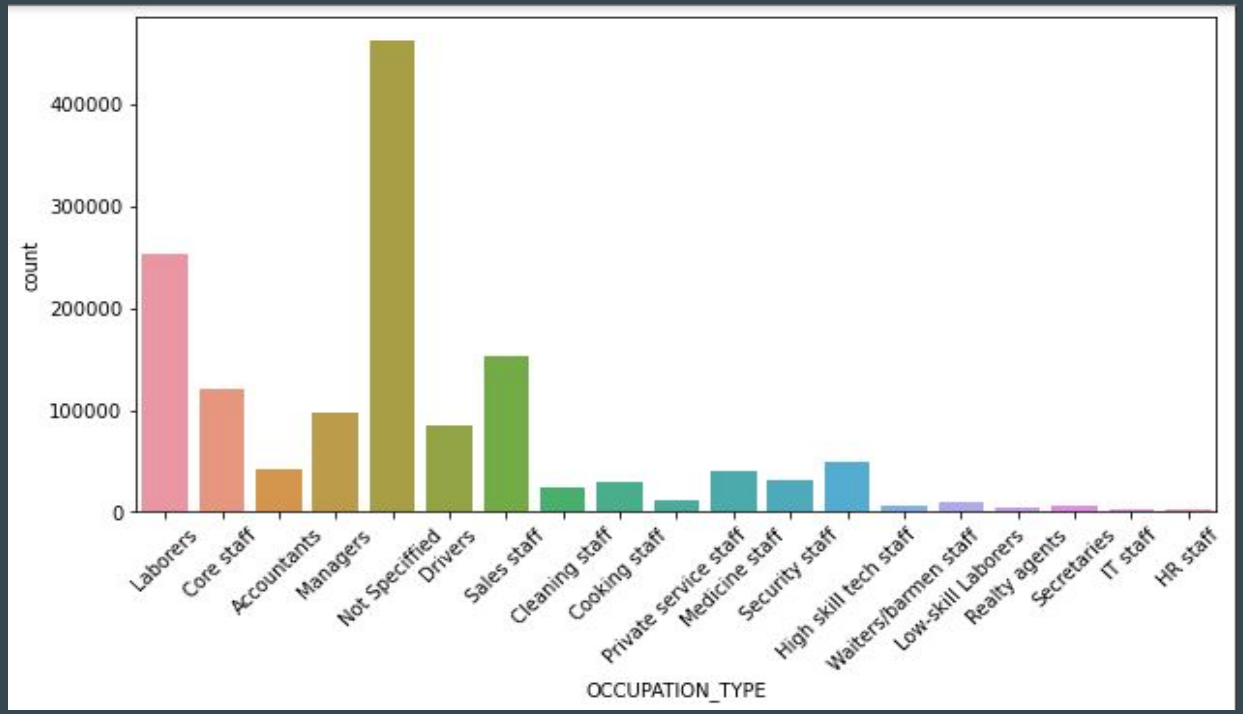
- The majority of applicants have “secondary/secondary special” education.



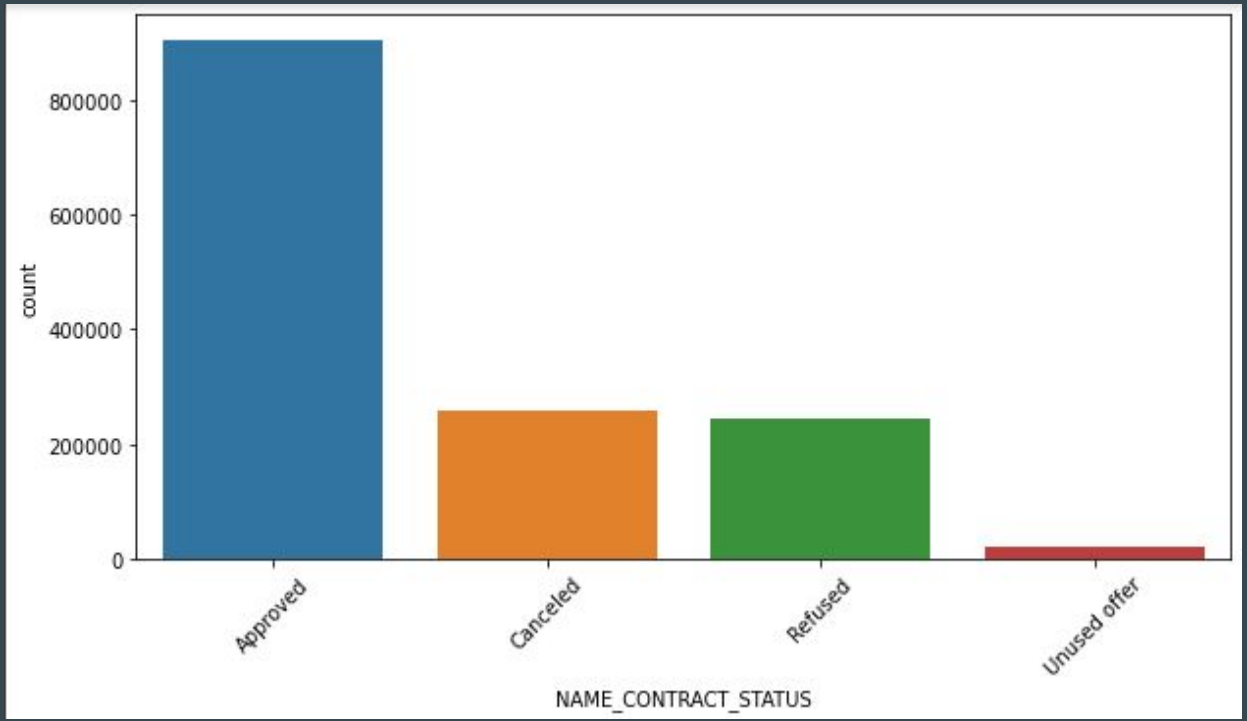
- Majority of the applicants are married.



- The majority of applicants have their own house/apartment.



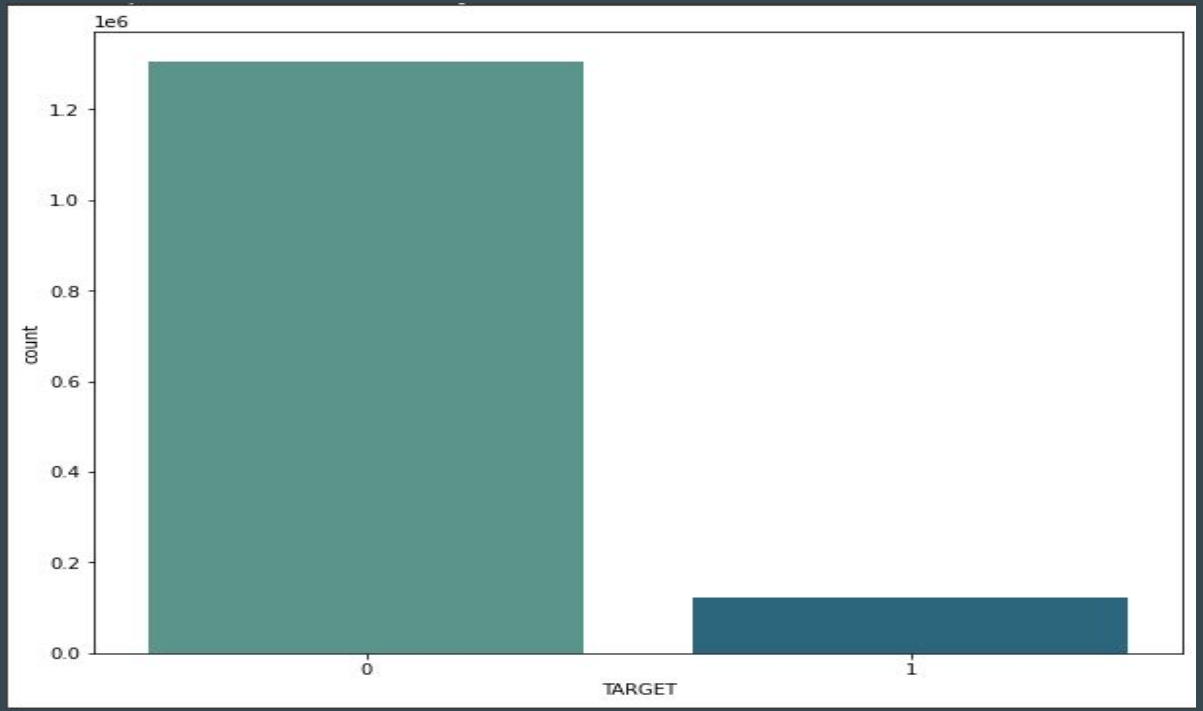
- The majority of applicants didn’t specified their occupation.



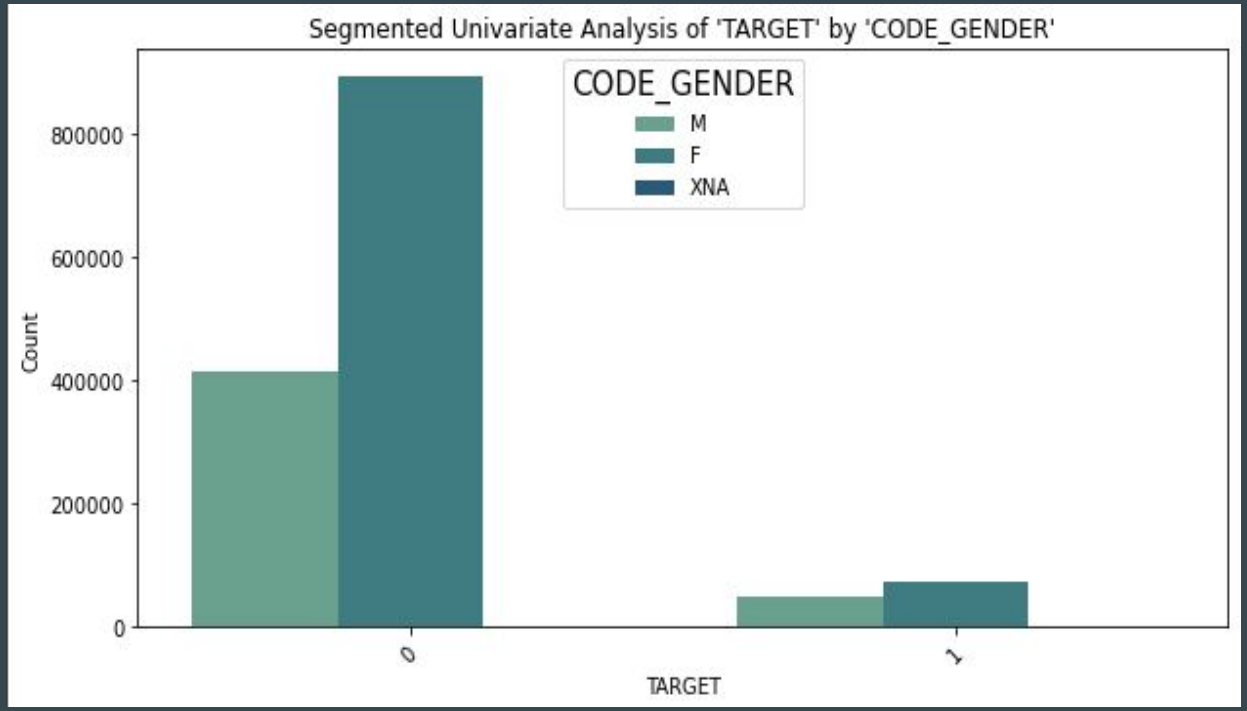
- The majority of previous applications have approved.

6. Segmented Univariate Analysis

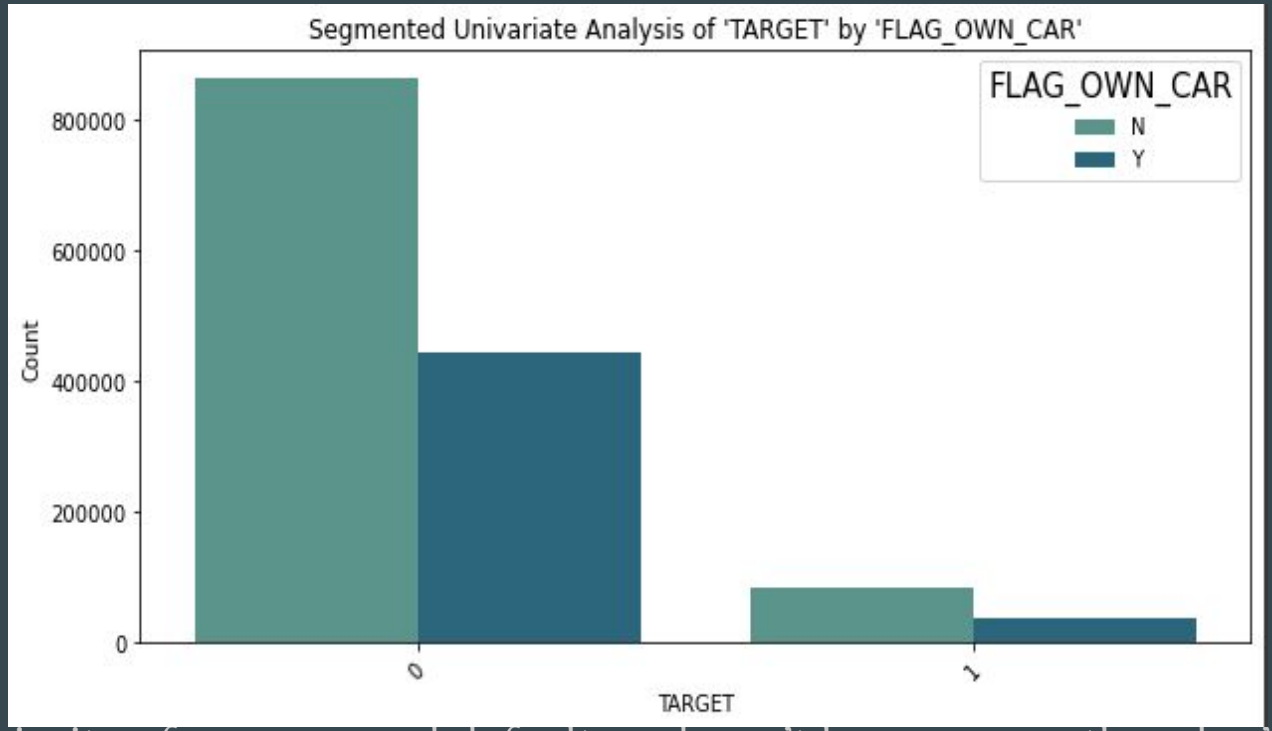
- Segmented univariate analysis is a powerful technique for exploring the relationship between a target variable and different segments or categories within a dataset.
- It helps identify patterns and variations in how the target variable behaves with respect to different segments.
- Here, we conducted segmented univariate analysis for several variables based on the "TARGET" variable, which represents whether applicants are repayers or defaulters.
- Count plot is used in this analysis.



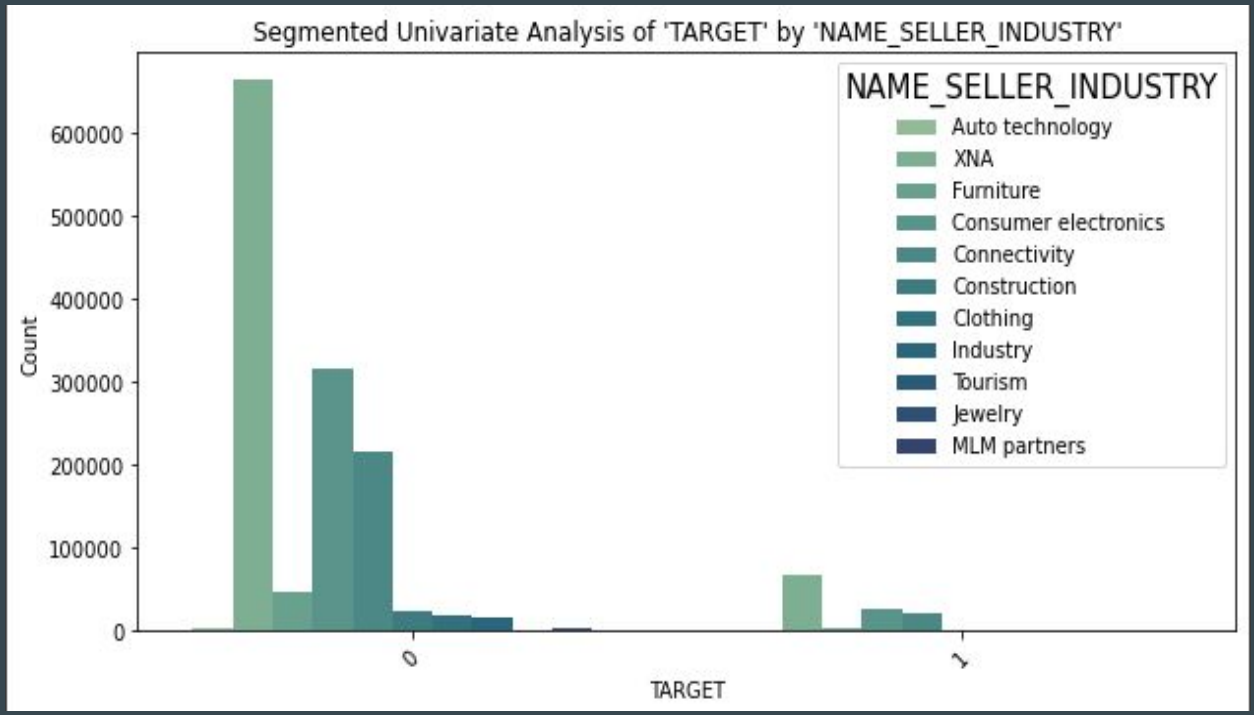
- Majority of applicants can repay the loan. It is represented as 0.



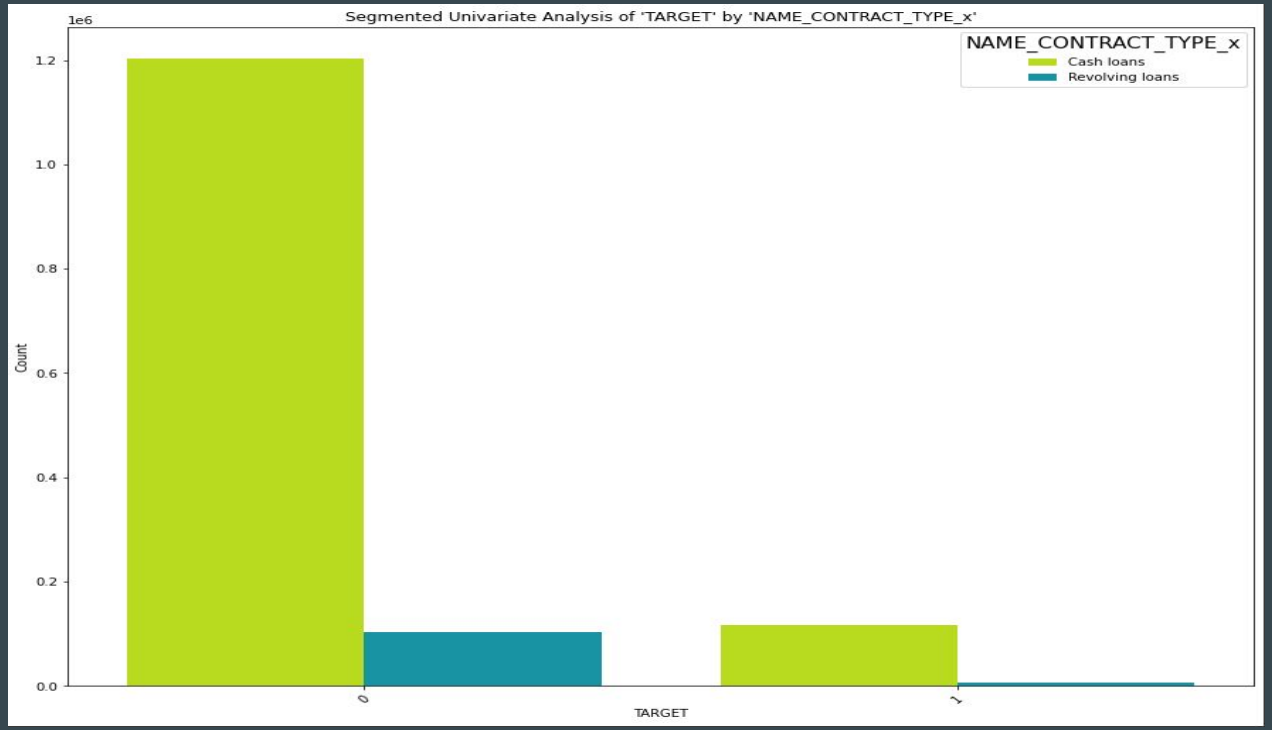
- Female applicants are able to repay the loans rather than male applicants. Defaulters are less in both gender.



- Majority of repayers and defaulters doesn't have cars, so they don't get flag for the application.



- The applicants not specified their industry have more chance to repay the loan. Applicants having consumer electronics is in the second position in the order.



- The clients applied for cash loans are more likely to repay the loan.

7. Bivariate Analysis

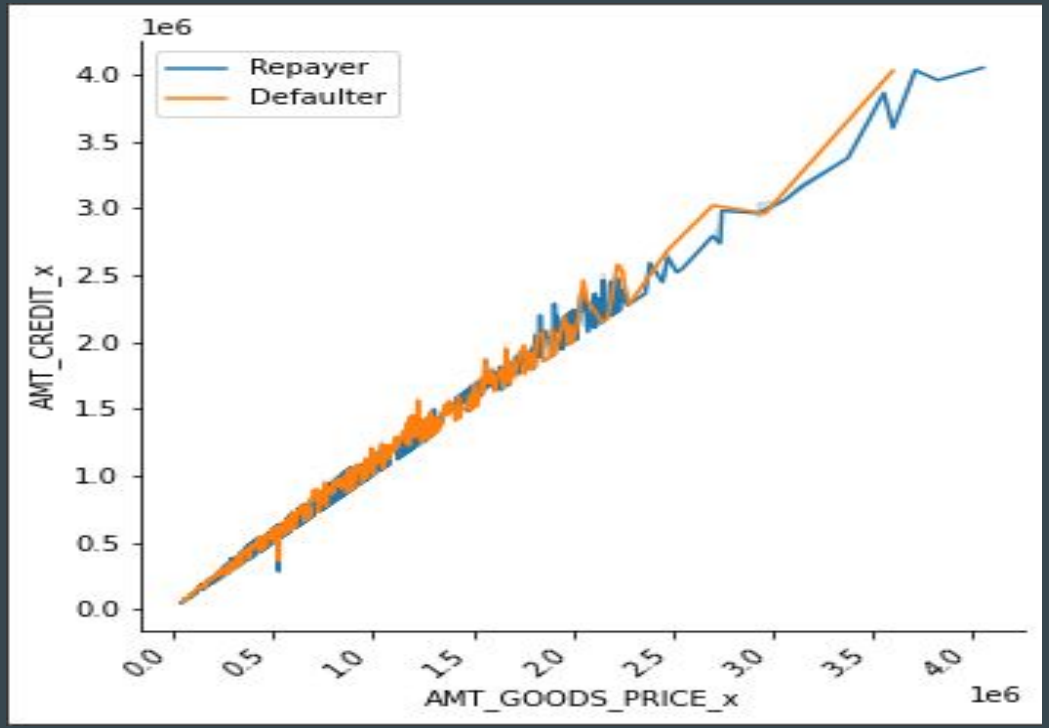
- Bivariate analysis is a powerful tool for exploring the relationships between two variables.
- In our analysis, we've considered the relationship between various numerical attributes and the target variable which represents whether an applicant is a repayer or defaulter.
- We got some key findings from the analysis.
- We used heatmap, relplot, scatterplot for the analysis.
- We found the best correlated variables for repayers and defaulters separately.

- Top 10 correlated variables of repayers

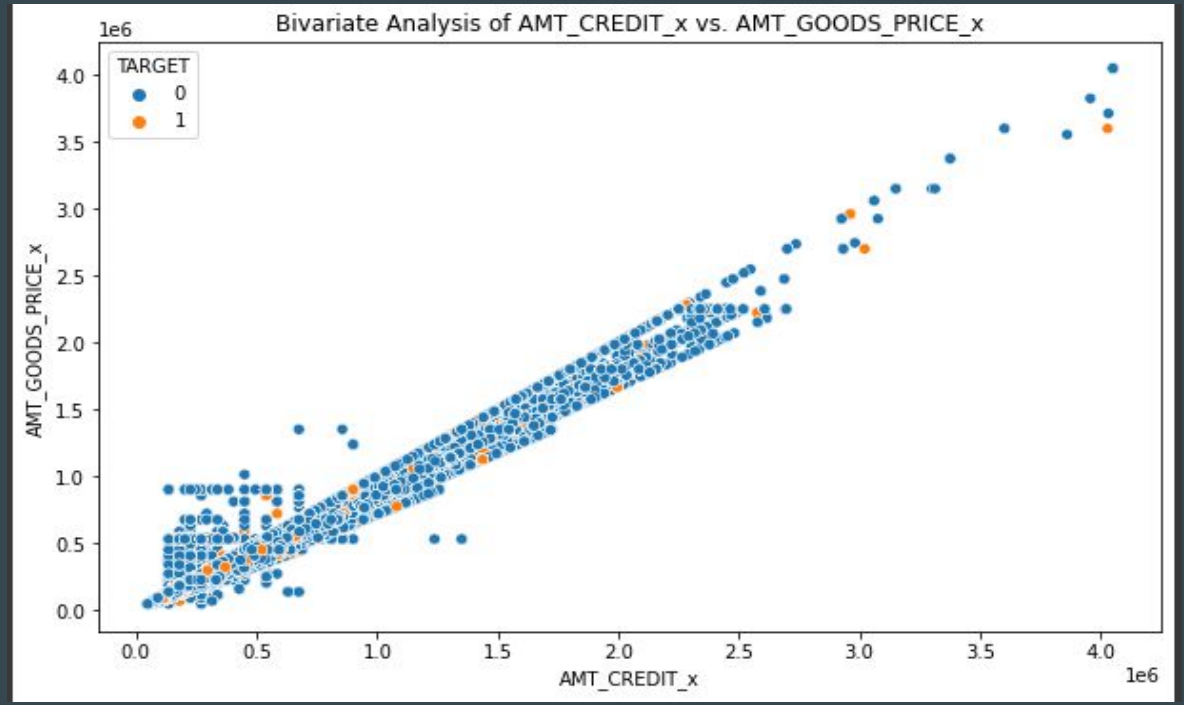
	VAR1	VAR2	Correlation
242	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.944635
182	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878168
323	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.874188
404	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.835286
107	DAYS_EMPLOYED	DAYS_BIRTH	0.627513
210	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.525737
236	REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	0.524537
377	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.431449
296	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.424962
78	DAYS_BIRTH	CNT_CHILDREN	0.367343

- Top 10 correlated variables of defaulters

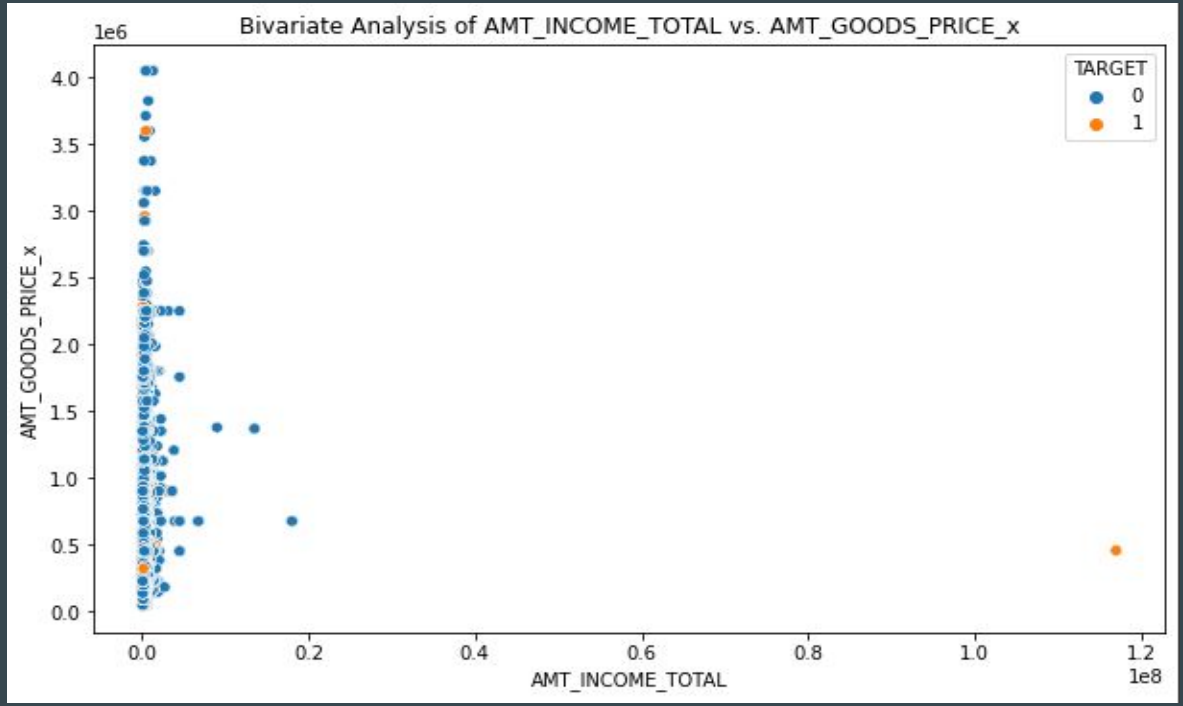
	VAR1	VAR2	Correlation
242	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956321
182	CNT_FAM_MEMBERS	CNT_CHILDREN	0.886005
323	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.871302
404	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.792598
107	DAYS_EMPLOYED	DAYS_BIRTH	0.580581
377	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.465897
296	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.460320
236	REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	0.431318
210	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.427898
348	REG_CITY_NOT_LIVE_CITY	REG_REGION_NOT_LIVE_REGION	0.298886



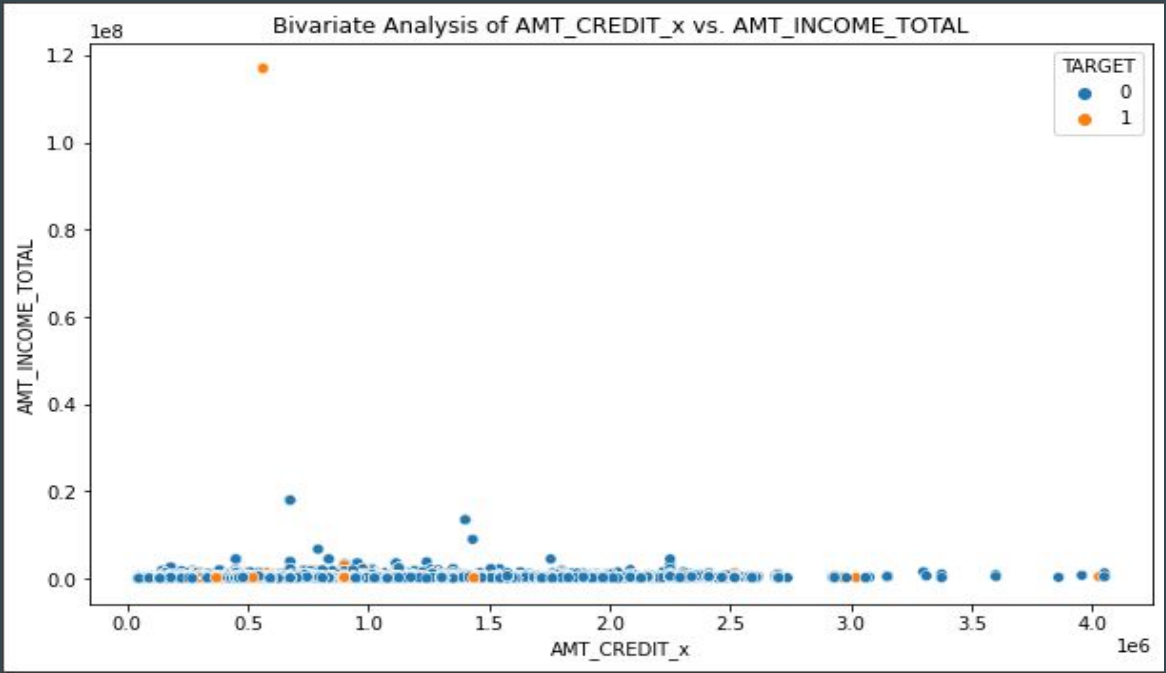
- When the credit amount goes beyond 25 Lakhs, there is an increase in defaulters.



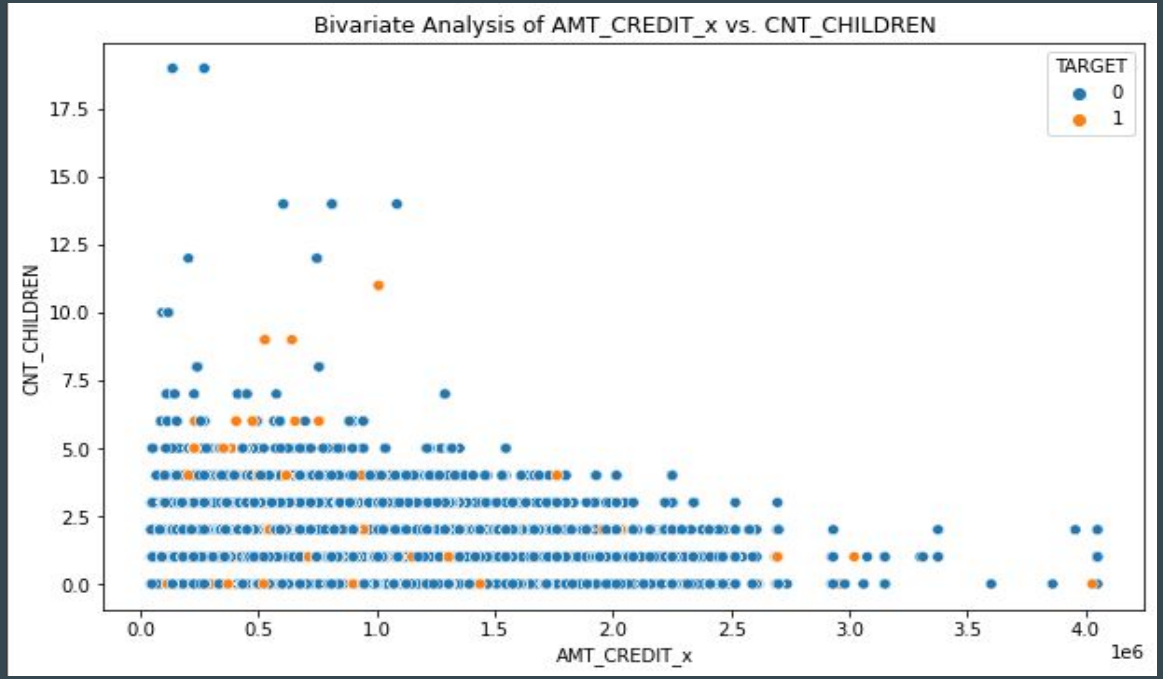
- These are linearly correlated and when credit amount increases defaulters are seen to be decreased.



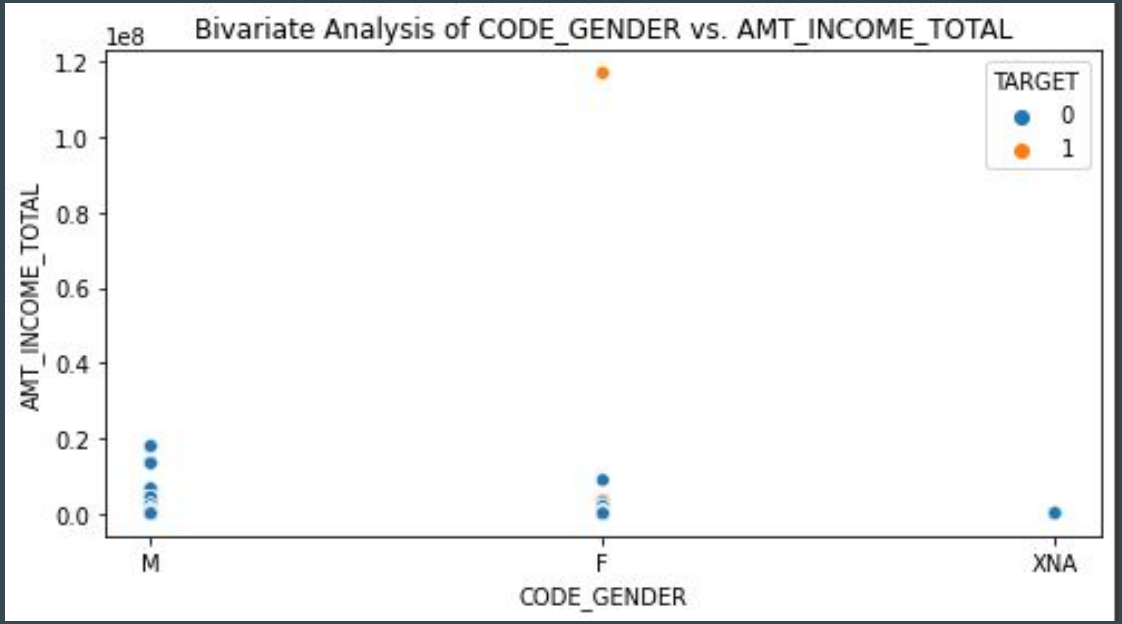
- According to the plot non-defaulters are in between the range of above 1.5 to below 2.5, so it is safe to give loans in this range only.



- People having more income don't tend to take loans but people having less income could turn to become defaulters. Low Income- High Defaulters, High Income- Low or Non-Defaulters.



- People having children (3-5) tend to be defaulters and people having fewer or no children tend to be non-defaulters.According to the plot, more than 1.5 are safe to give loans.



- Working Females are tend to repay loan than others,so providing loan for them reduces the risk.

Other Inferences:

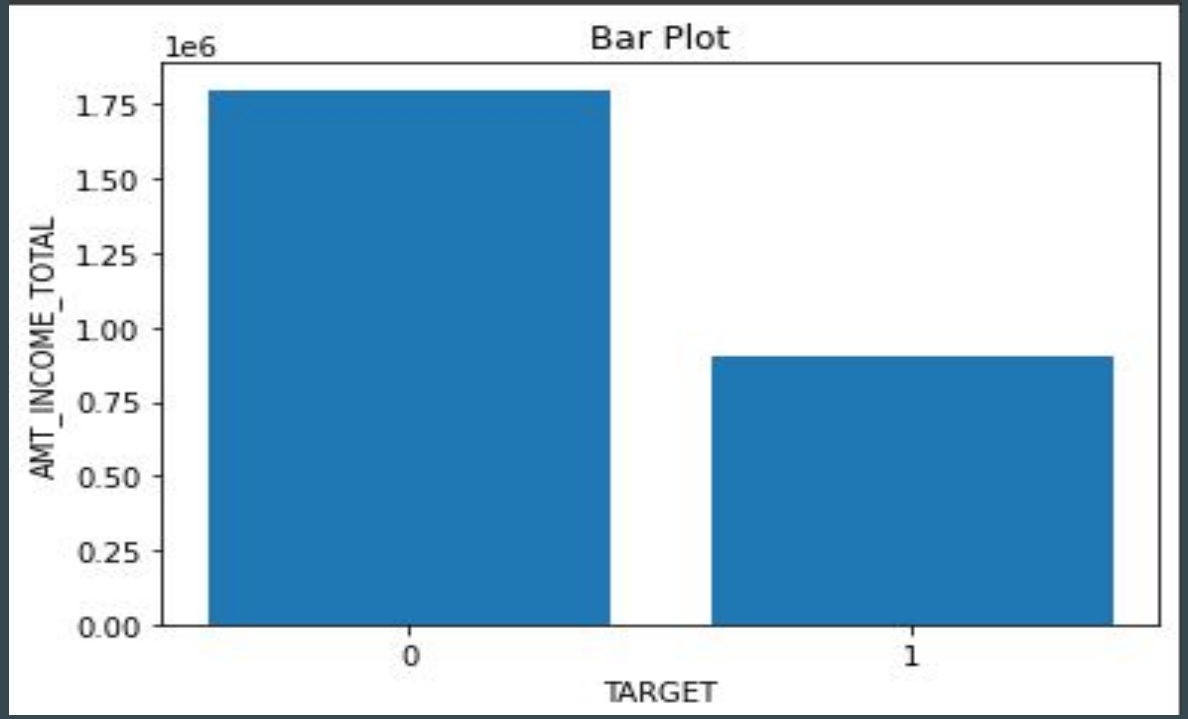
- When Annuity Amount > 15K and Good Price Amount > 20 Lakhs, there is a lesser chance of defaulters.
- Loan Amount(AMT_CREDIT) and Goods price(AMT_GOODS_PRICE) are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line.
- There are very less defaulters for AMT_CREDIT >20 Lakhs.

8. Correlation Analysis

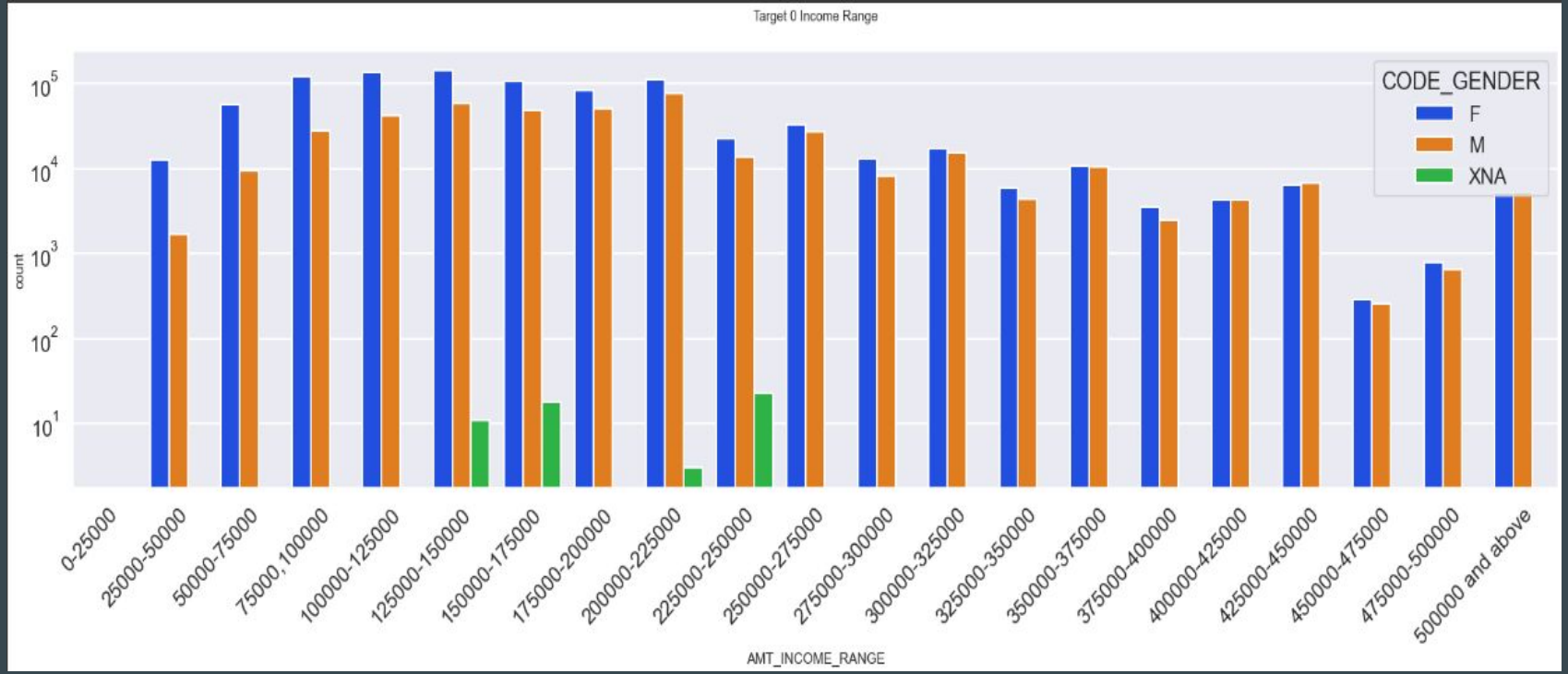
- Correlation analysis is a statistical technique used to evaluate the strength and direction of the linear relationship between two or more variables.
- We calculated the correlation of all attributes with the 'TARGET' variable. This helps in identifying which attributes are most strongly correlated with loan repayment status.
- We created bins for income and credit amount ranges, allowing us to see how different income and credit levels affect loan repayment.
- Analyzed using barplot, countplot.
- Found top 10 correlations.

- Top 10 correlations

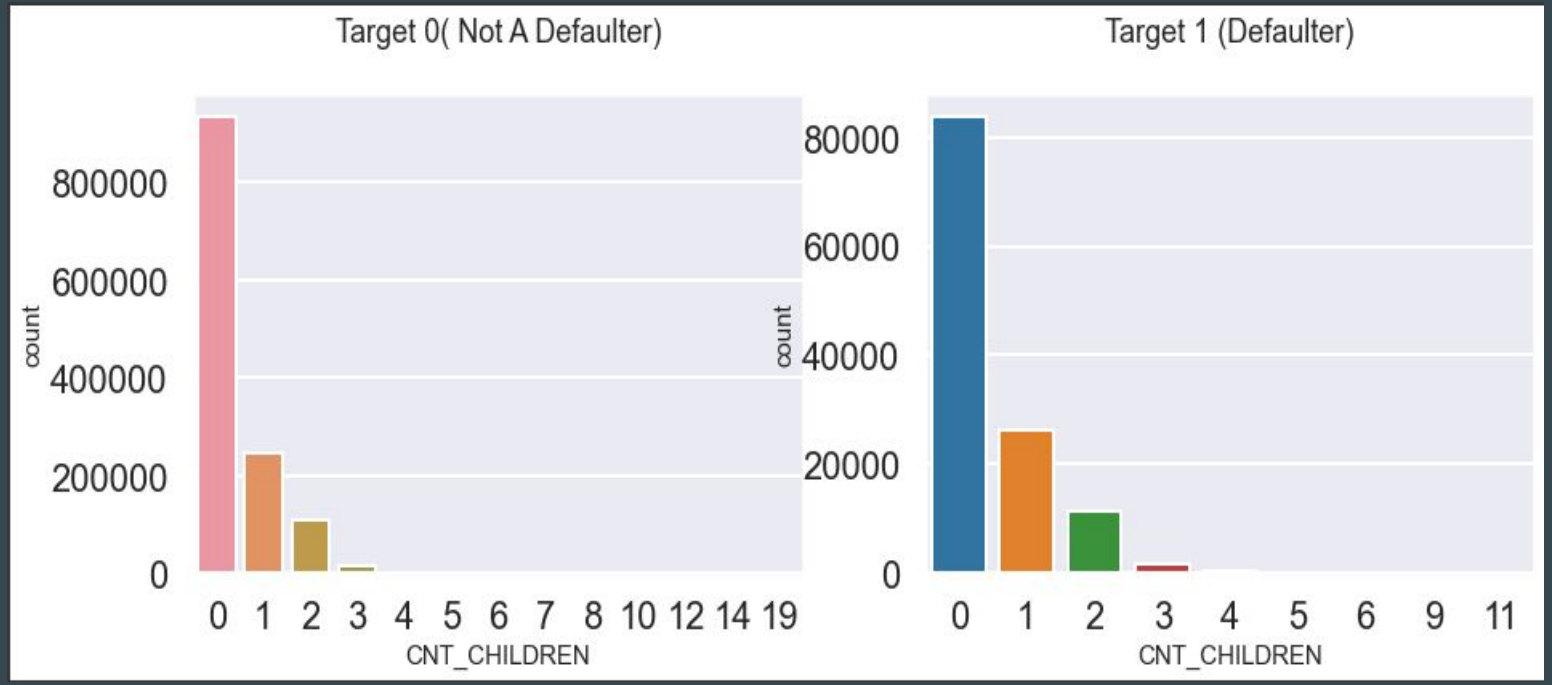
EXT_SOURCE_3	0.169590
EXT_SOURCE_2	0.155104
EXT_SOURCE_1	0.098944
DAYS_BIRTH	0.074314
REGION_RATING_CLIENT_W_CITY	0.059832
DAYS_LAST_PHONE_CHANGE	0.058110
REGION_RATING_CLIENT	0.057135
DAYS_ID_PUBLISH	0.050833
REG_CITY_NOT_WORK_CITY	0.049383
FLAG_EMP_PHONE	0.048353



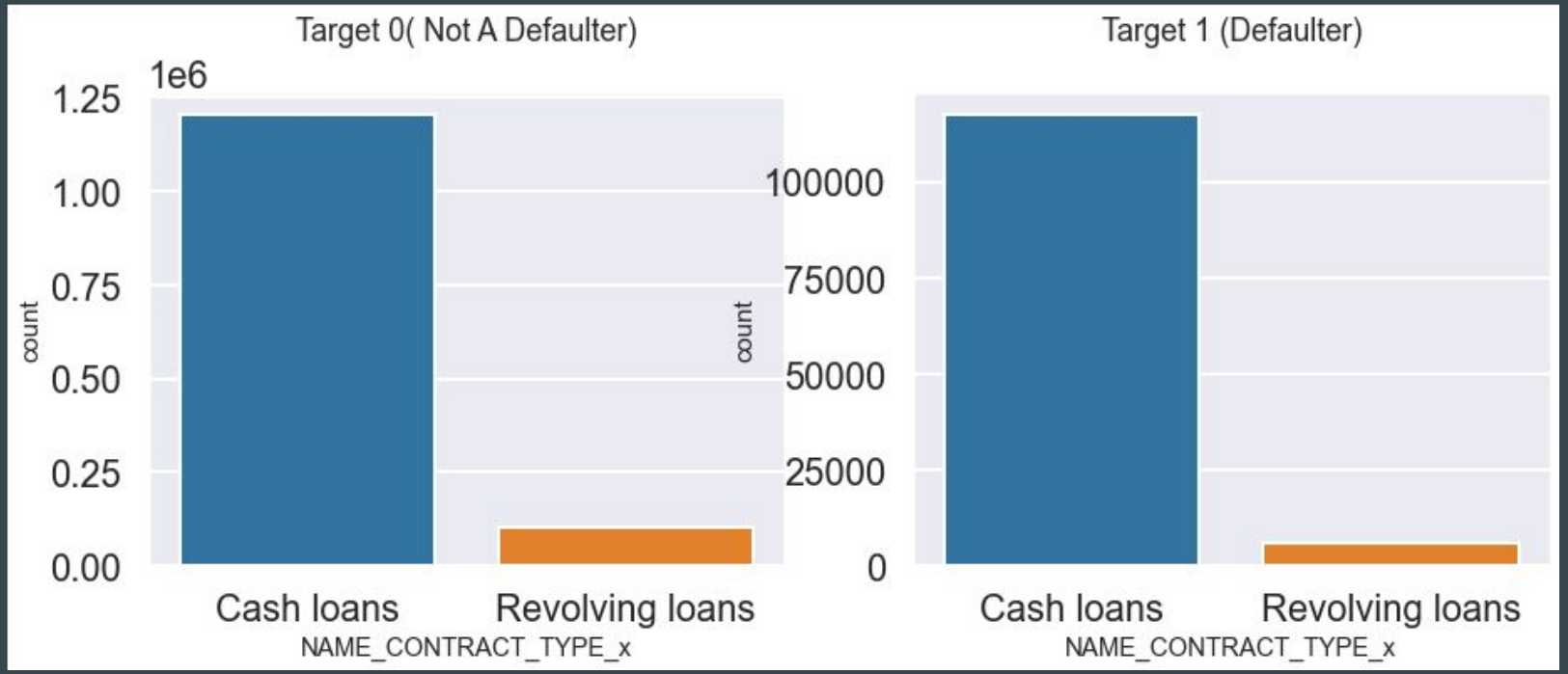
- If client has low income, they will be difficulty in repaying the loan.



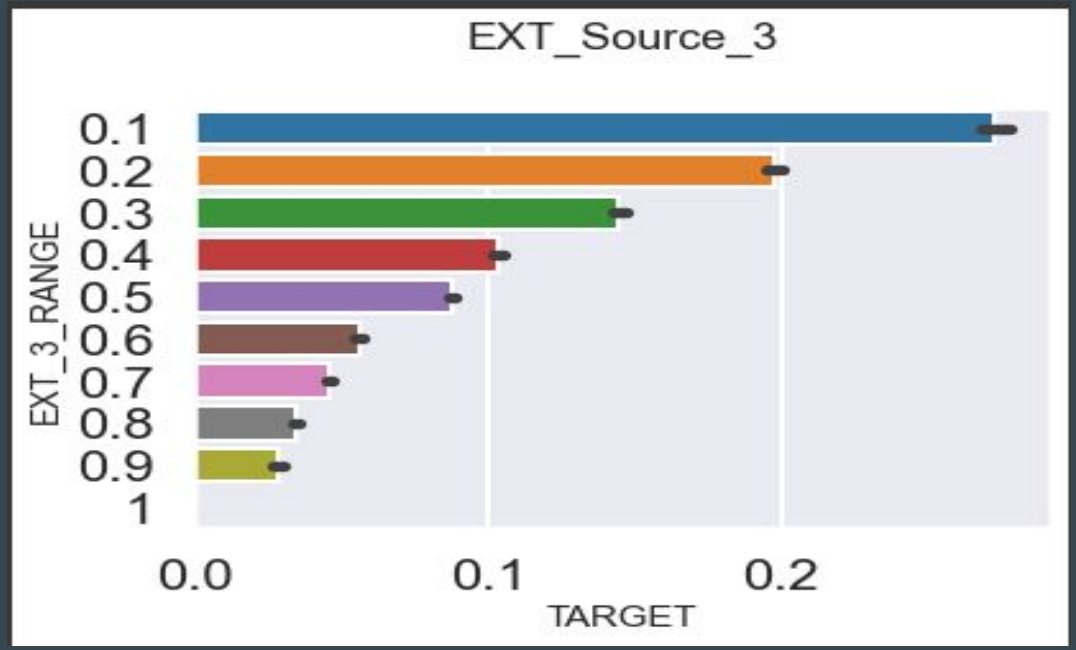
- Females are having high income, so they can repay the loan easily. Priority is given to them.



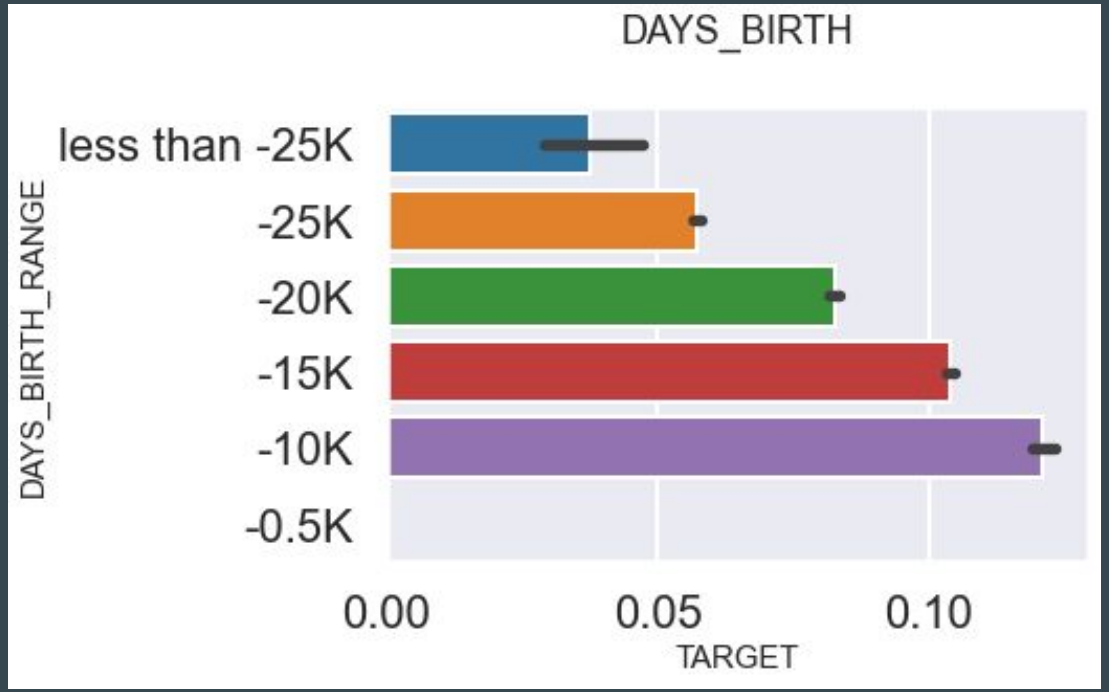
- Applicants having 0 children are likely to be non-defaulters.



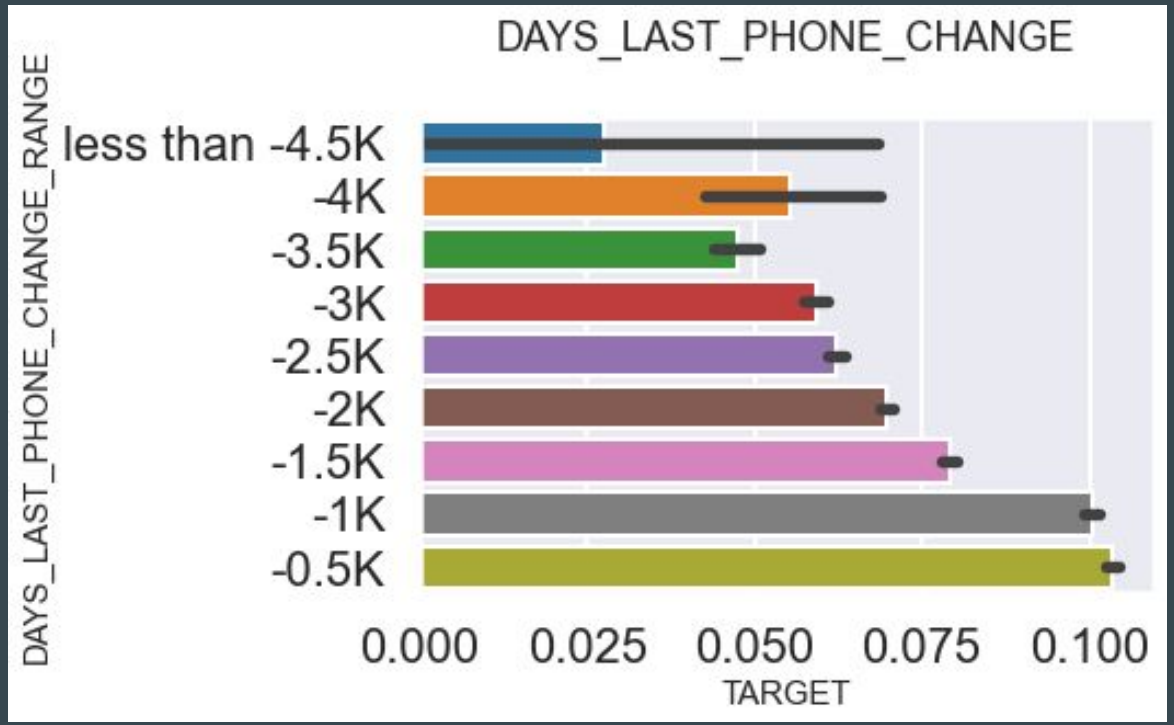
- The clients applied for cash loans are likely to be non-defaulters.



- Client having less external source faces more difficulty to repay the loan.



- New clients faces more difficulty to repay the loan compared to old clients.



- Client recently changed his/her phone number have less chance to get the loan.

- Client in region 1 are more likely to repay the loan, but client in region 3 faces difficulty.

Key Findings

- 9% of total applicants are defaulters and 91% are non-defaulters.
- Female applicants are more and they are more likely to repay the loan.
- The majority of applicants are applied for cash loans.
- Majority of repayers and defaulters doesn’t have cars, so they don’t get flag for the application.
- When the credit amount goes beyond 25 Lakhs, there is an increase in defaulters.
- when credit amount increases defaulters are seen to be decreased.

- People having more income don't tend to take loans but people having less income could turn to become defaulters.
- There are very less defaulters for AMT_CREDIT >20 Lakhs.
- Applicants having 0 children are likely to be non-defaulters.
- Client having less external source faces more difficulty to repay the loan.
- New clients faces more difficulty to repay the loan compared to old clients.
- Client recently changed his/her phone number have less chance to get the loan.

Conclusion and Future Work

- Cleaned and prepared the dataset for analysis, including handling missing values and handling outliers.
- Calculated Imbalance Ratio with 91% repayers and 9% defaulters in the dataset.
- Univariate Analysis: Explored the distribution of data, revealing insights into key features in the dataset.
- Segmented Univariate Analysis: Examined the impact of attributes like gender, car ownership, and income type on loan repayment.

- Bivariate Analysis: Investigated relationships between variables, particularly with AMT_GOODS_PRICE, AMT_CREDIT, AMT_TOTAL_INCOME, CNT_CHILDREN etc .
- Correlation Analysis: Identified attributes significantly correlated with loan repayment, providing insights for risk assessment.
- Future Work: Discussed potential next steps, including building predictive models and enhancing risk assessment techniques.

Thank You

Any Questions?