Exploratory Data Analysis(EDA)

ON

LOAN CREDIT DATA

Problem Statement:
Given the Data Set of Loan the task was to perform EDA on the Data Set involving the following steps

- Fixing rows and columns
- Handling missing values and Outliers
- Fixing the errors in terms of Format, Structure and types
- Univariate,Bivariate and Multivariate Analysis of the Data

Data: Previous_application.csv, application_data.csv, columns_description

Output: Identification off the patterns observed in the loan defaulters and top 10 highly correlated Variables

Handling Missing values:

- The application data contains where some of the columns have more 45% of the data is missing.Imputing these missing values with statistic variable like mean,median or mode will result in high risk of bias and inaccuracy and directly results in impacting the model’s performance. So, these columns can be dropped from the data set.
- This is similar for both application_data and previous_data set
- So the columns like COMMONAREA_MEDI, COMMONAREA_AVG, COMMONAREA_MODE NONLIVINGAPARTMENTS_MODE .NONLIVINGAPARTMENTS_AVG etc are dropped
- And the columns which are cannot effect the Target variable are also removed.Below are few examples of the those columns
- Examples: 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE','FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_3' etc
- The Null of the Column OCCUPATION_TYPE is replaced with “Unknown” and the Other Numerical Categorical variables are imputed with the mode values

Handling Outliers:

- Boxplots can be utilized to visualize the outliers in each of the columns in the dataframe
- The methods of binning and capping can be used for imputing the outliers
- For column ‘AMT_INCOME_TOTAL’ are binned into separate groups based on the occupation type and the outliers in these groups are replaced by the 99th percentile value in each of the groups
- And the other numerical column having the outliers are imputed by the capping method of using the 99th percentile value

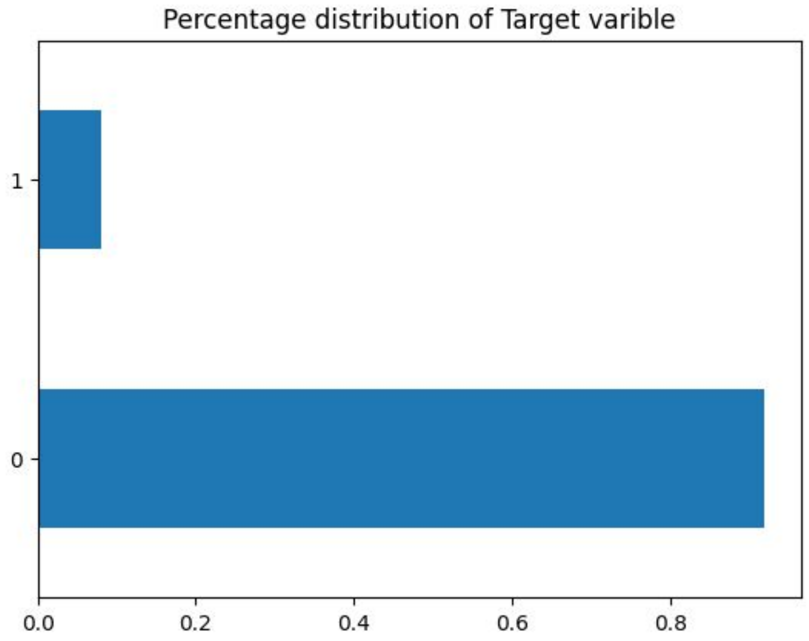
Fixing the errors in categorical columns:

- The values_counts method can be used for each of the column and looks if all the data following the correct structure and format. These can be fixed by replacing them with the expected values
- Categorical columns like 'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE','NAME_FAMILY_STATUS' etc are following the expected structure and format.

Univariate Analysis:

Aim: Understanding the Distribution of the Target Variable

Graph:

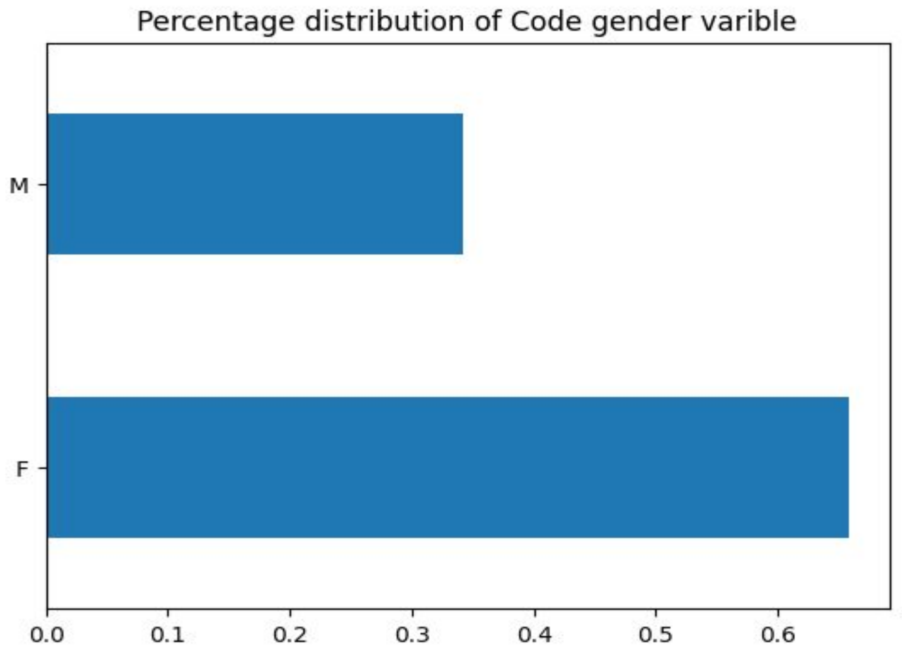


Conclusion: 92% of the Target variable consist of clients who pay on time and 8% of the clients

Univariate Analysis:

Aim: Understanding the Gender Distribution among the clients

Graph:

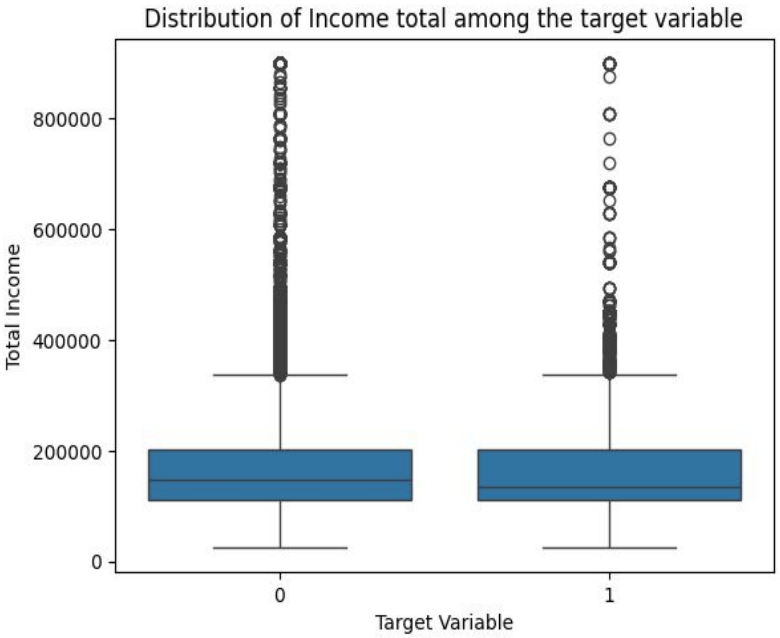


Conclusion: 66% of the Clients applying for loans are Females(F) and the 34% are Males(M)

Bivariate Analysis:

Aim: Understanding the Income Distribution with Target 0 and 1

Graph:

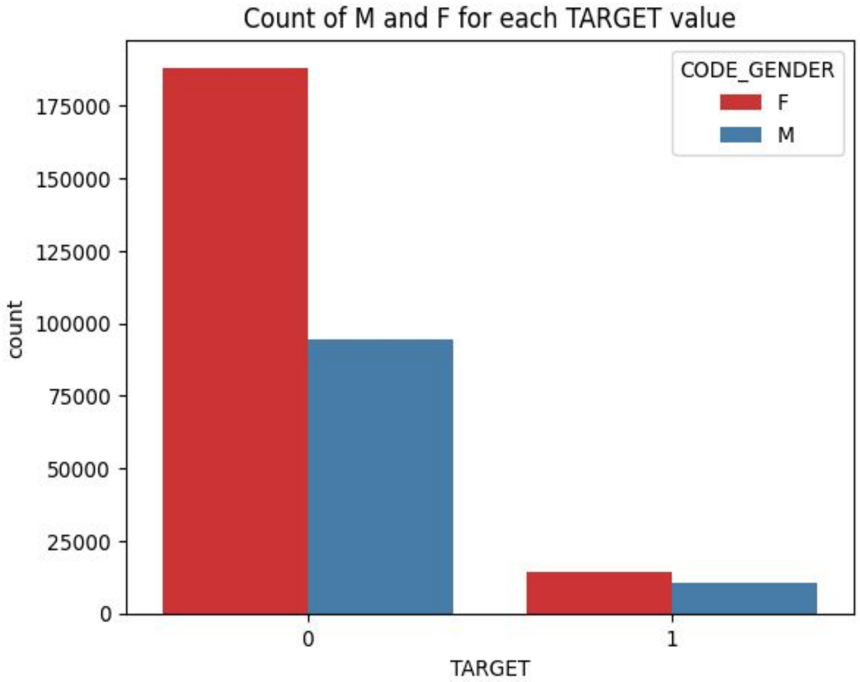


Conclusion: There is no major difference spotted between the max, median or the IQR between

Bivariate Analysis:

Aim: Understanding the count of M and F for the Target 0 and 1

Graph:

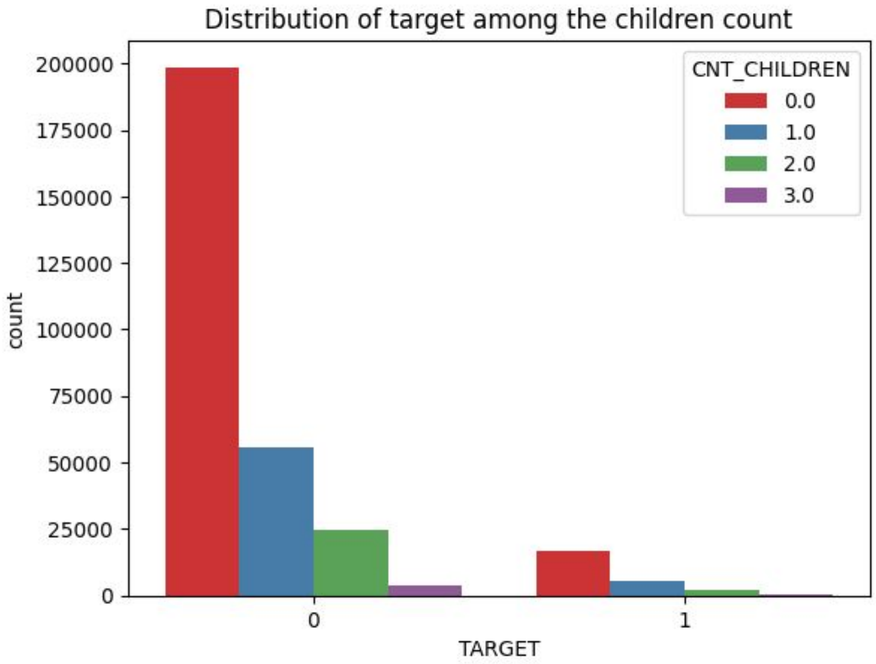


Conclusion: It is observed the Females have higher percentage in number of loans and relatively

Bivariate Analysis:

Aim: Understanding the distribution of the Target 0 and 1 with respect to the number of children

Graph:

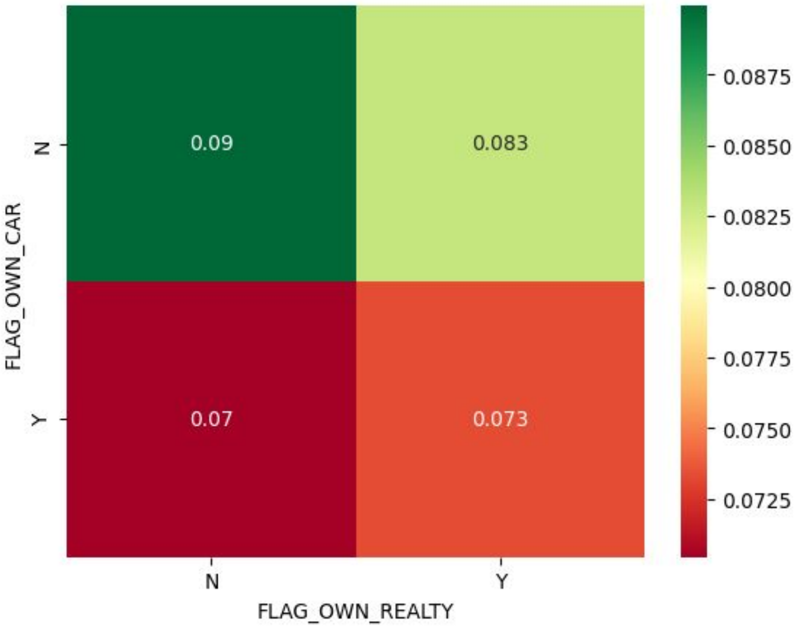


Conclusion: It is observed that the customers with zero children are more likely to opt for taking

Multivariate Analysis:

Aim: Understanding the distribution of the Target 0 and 1 with respect customers owning car and reality

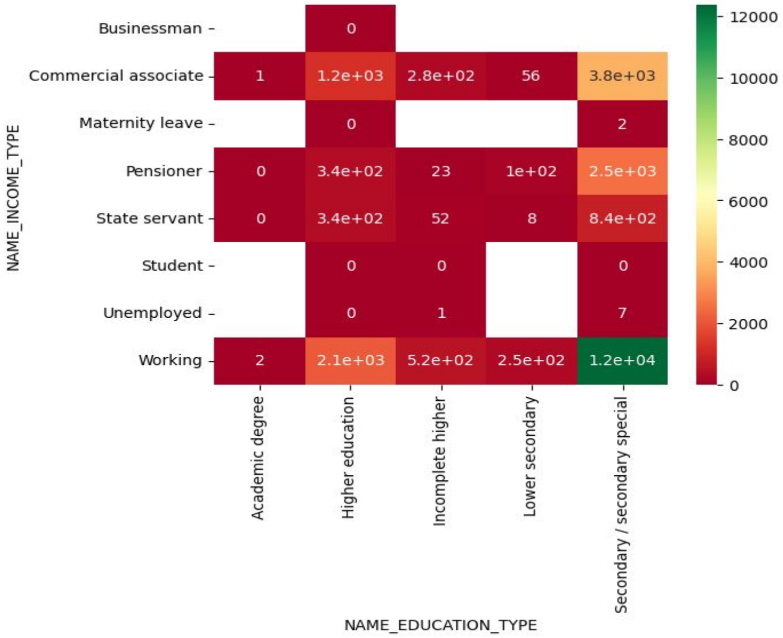
Graph:



Multivariate Analysis:

Aim: Understanding the distribution of the Target 0 and 1 with respect Income type and Education type

Graph:

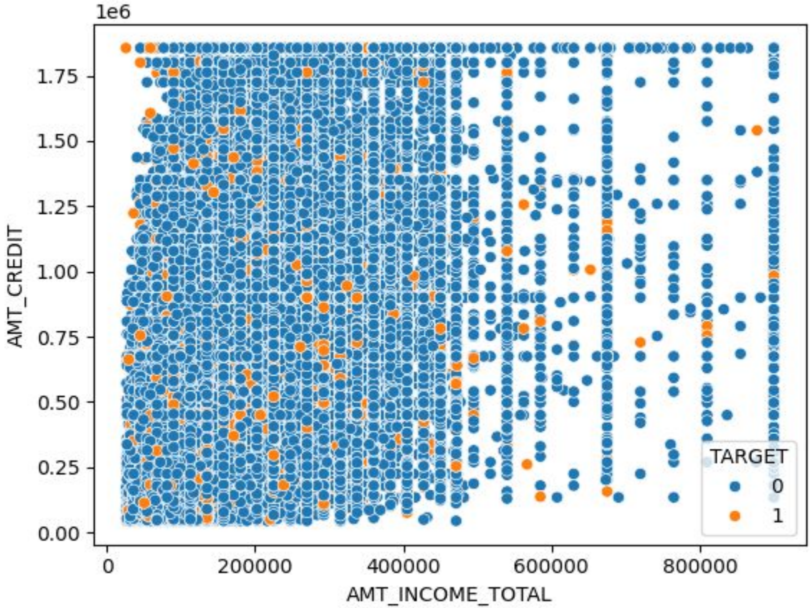


Conclusion: The clients who have 3 children and either single or civil marriage have higher percentage of the delayed payments

Multivariate Analysis:

Aim: Understanding the distribution of the Target 0 and 1 with respect customers owning car and reality

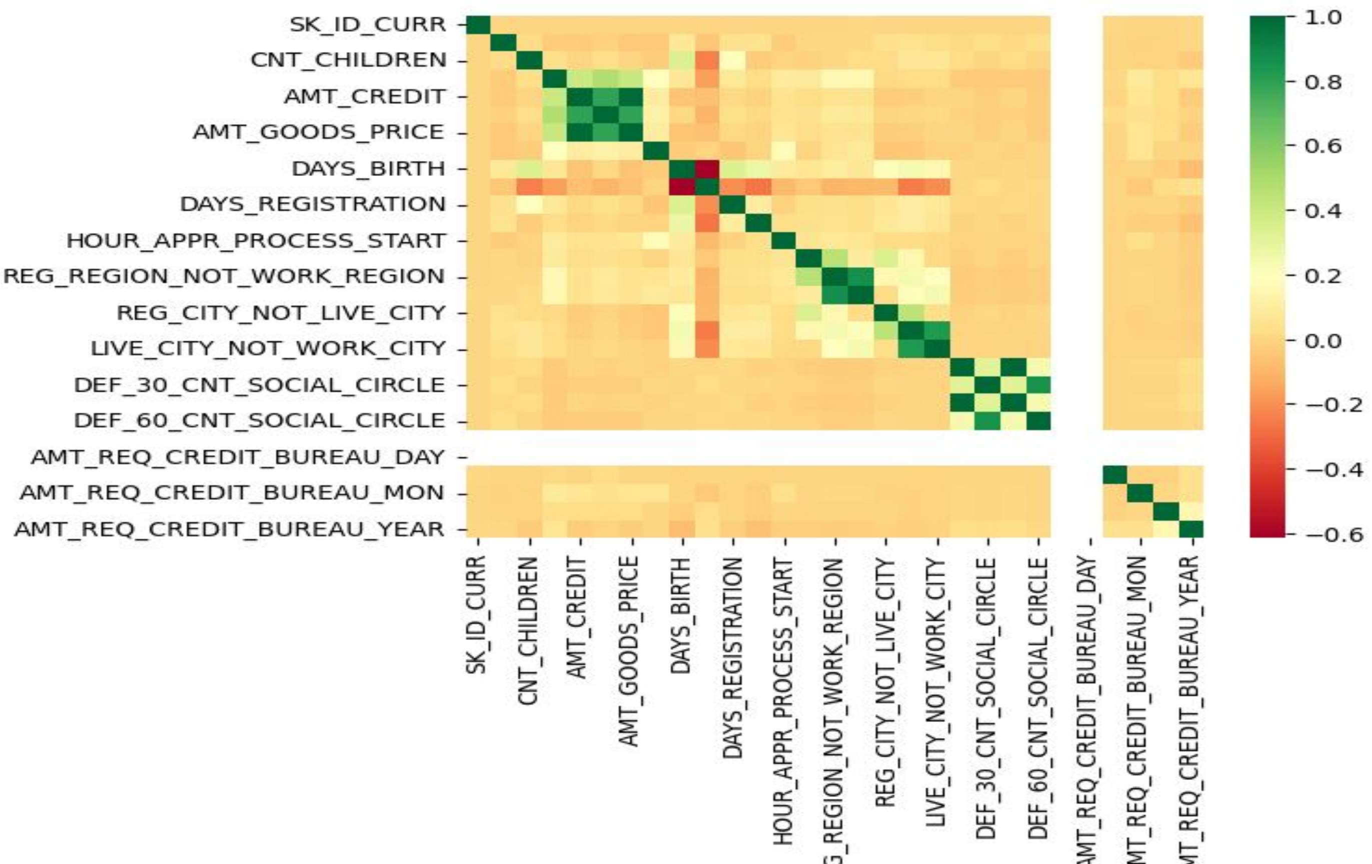
Graph:



Conclusion:

The clients who have lower income have taken more number of loans and also have higher percentage of delayed payments

Correlation between the variables:



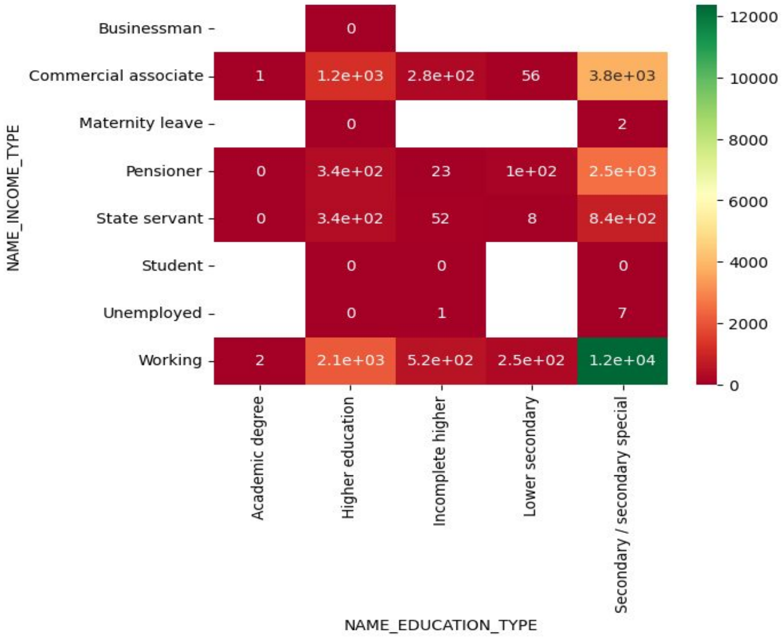
Top 10 correlated variables:

Variable-1	Variable - 2	correlation value
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998264
AMT_CREDIT	AMT_GOODS_PRICE	0.986447
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.860495
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.852307
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.825558
AMT_ANNUITY	AMT_GOODS_PRICE	0.790507
AMT_CREDIT	AMT_ANNUITY	0.787564
DAYS_EMPLOYED	DAYS_BIRTH	0.615908
AMT_ANNUITY	AMT_INCOME_TOTAL	0.475616
REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	0.450989

Multivariate Analysis:

Aim: Understanding the distribution of the Target 0 and 1 with respect Income type and Education type

Graph:



Conclusion: The clients who have 3 children and either single or civil marriage have higher percentage of the delayed payments

Merging and preprocessing the Data Frame:

- Merging both the data frames of previous_application and application_data on 'SK_CURRENT_ID' by inner join to have all the applications in the application_data and get the columns added for only rows which are present in the application_data
- Dropping the columns which are non-applicable/effect the target variable

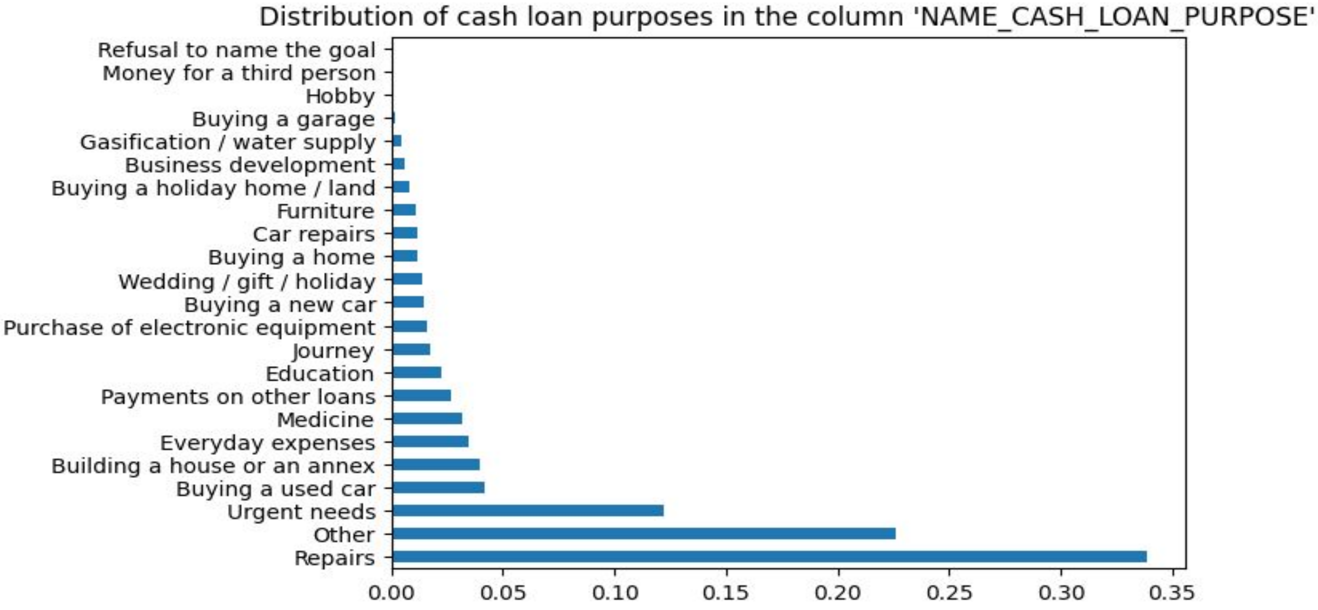
Below are columns dropped:

'WEEKDAY_APPR_PROCESS_START_CURR','HOUR_APPR_PROCESS_START_CURR','WEEKDAY_APPR_PROCESS_START_PREV','HOUR_APPR_PROCESS_START_PREV', 'REG_REGION_NOT_LIVE_REGION', 'REG_REGION_NOT_WORK_REGION', 'LIVE_REGION_NOT_WORK_REGION', 'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CITY', 'FLAG_LAST_APPL_PER_CONTRACT','NFLAG_LAST_APPL_IN_DAY'

Univariate Analysis:

Aim: Understanding the distribution of different purposes in the Loans data

Graph:

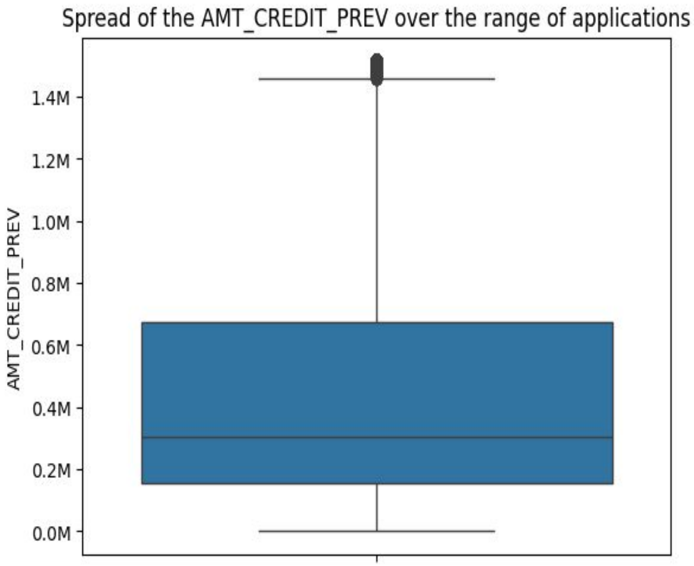


Conclusion: The clients with the purpose of repairs have the higher percentage in number of

Univariate Analysis:

Aim: Spread of credit amount over the range of previous loans applications

Graph:

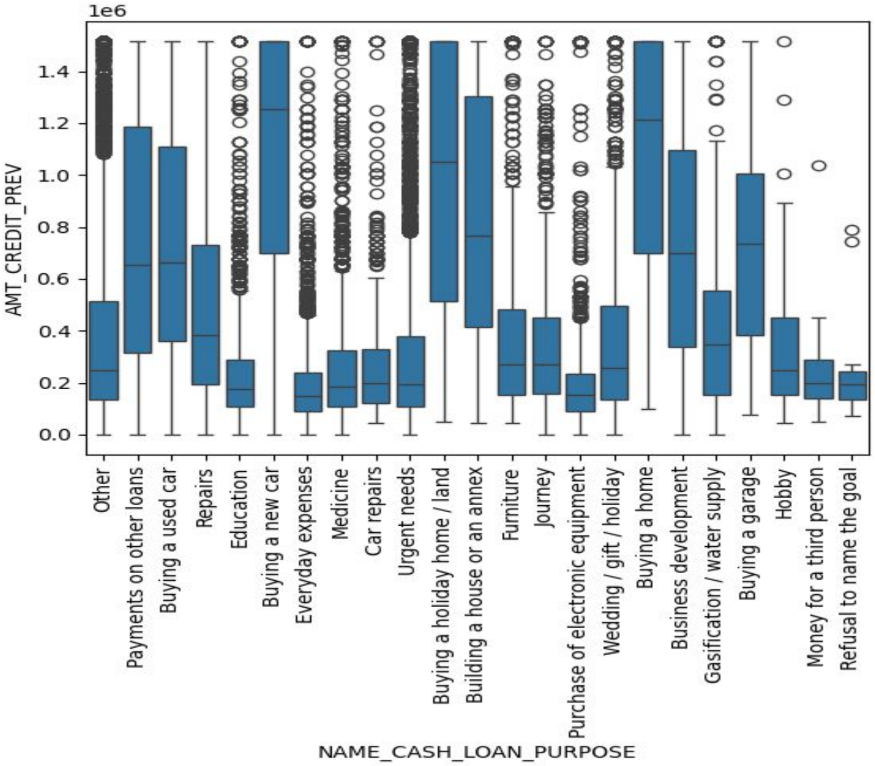


Conclusion: The clients have applied for the loan amount between 0.2 Millions to 0.62 Millions

Bivariate Analysis:

Aim: Spread of credit amount over the range of loans purposes

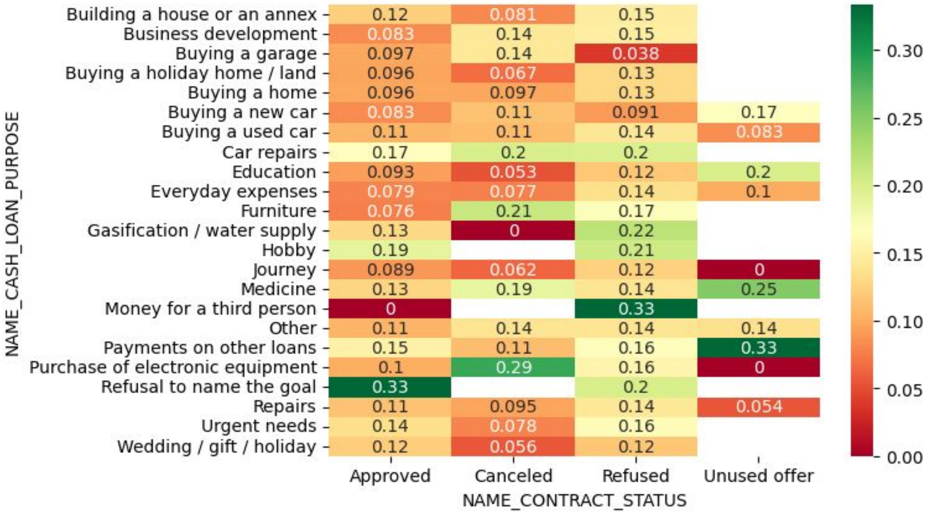
Graph:



Multivariate Analysis:

Aim: Spread of target variable between the Name cash loan purpose and the Name contract status over the range of previous loans applications

Graph:



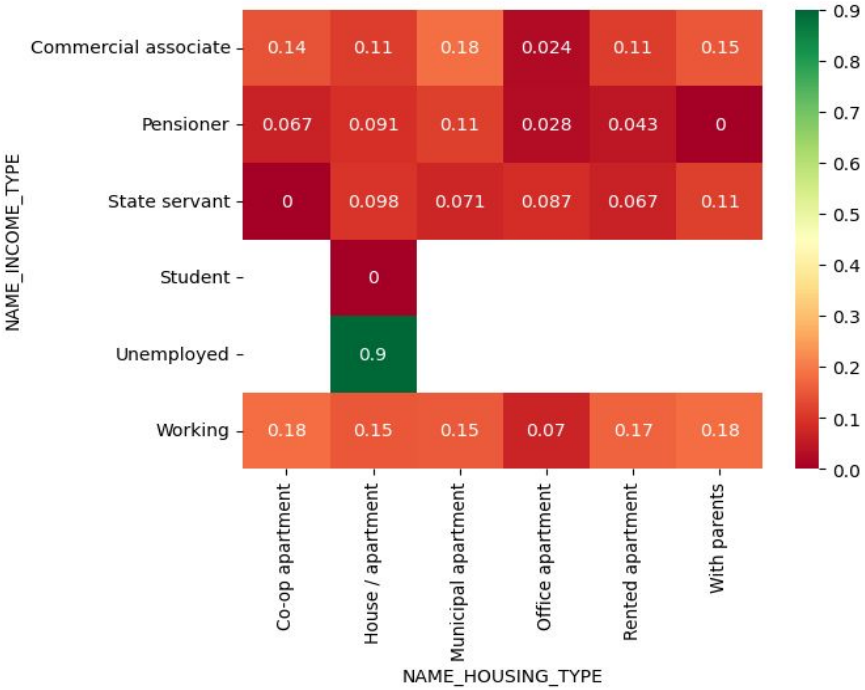
Conclusion:

```
1. The clients who refused to name the loan purpose are having high chances of becoming the
defaulters
```

Multivariate Analysis:

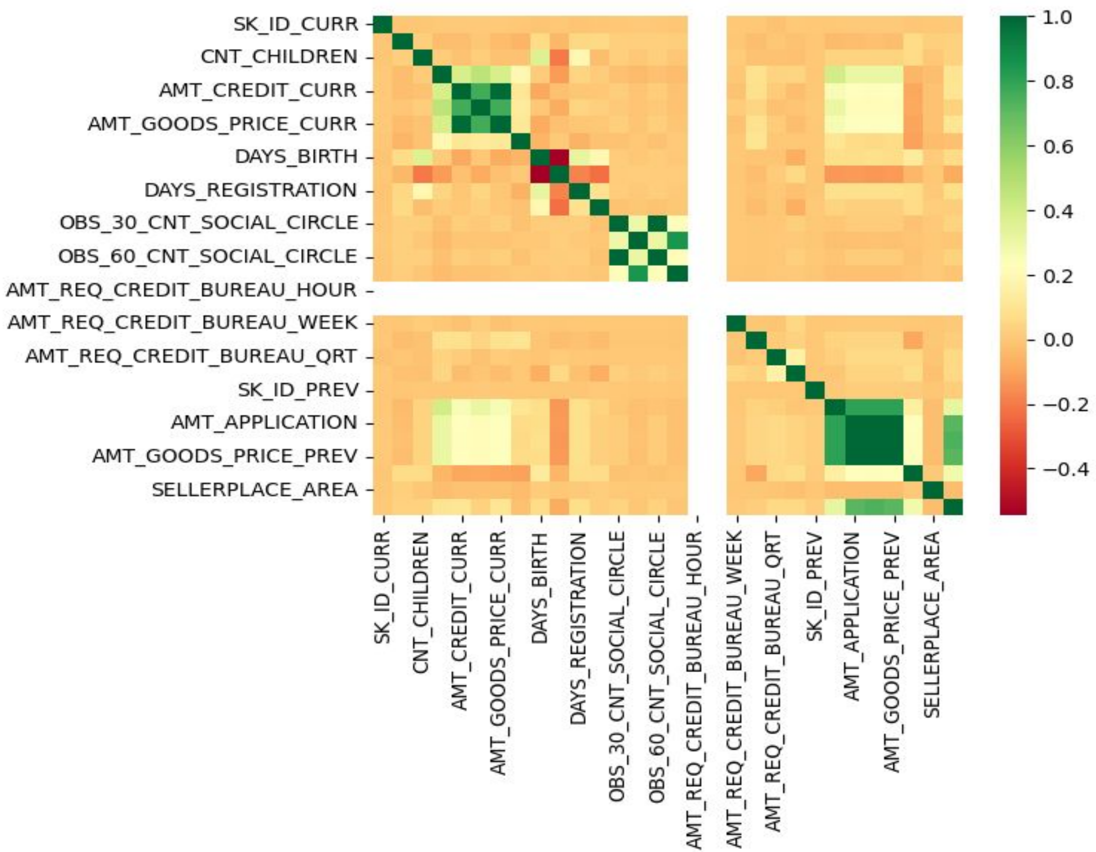
Aim: Spread of Target variable among Income type and Housing type

Graph:



Conclusion: The clients whose housing type is House and student,co-op Apartment and State

Correlation in Merged data:



Top 10 Correlations in the Merged Data:

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998552
AMT_CREDIT_PREV	AMT_APPLICATION	0.994726
AMT_GOODS_PRICE_PREV	AMT_CREDIT_PREV	0.994726
AMT_GOODS_PRICE_CURR	AMT_CREDIT_CURR	0.985216
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.848932
AMT_ANNUITY_PREV	AMT_GOODS_PRICE_PREV	0.805626
AMT_GOODS_PRICE_PREV	AMT_ANNUITY_PREV	0.805626
AMT_ANNUITY_PREV	AMT_CREDIT_PREV	0.802055
AMT_GOODS_PRICE_CURR	AMT_ANNUITY_CURR	0.760251
AMT_CREDIT_CURR	AMT_ANNUITY_CURR	0.759877

Conclusion:

The exploratory data analysis reveals clear patterns in loan repayment behavior:

- **Women** are more likely to repay loans on time than men.
- Clients with **no children** show better repayment behavior, while those with **three children** and either **single** or in **civil marriages** are more likely to default.
- Clients who **do not own both a car and real estate** are at higher risk of default, while those who own either or both have better repayment records.
- Borrowers with undisclosed **loan purposes** are more likely to default, whereas those requesting loans for a **new car** (even if rejected) tend to be reliable repayers.
- Clients living in **student housing, co-op apartments, or with parents**, and those who are **state servants or pensioners**, repay on time. In contrast, **unemployed** clients and those living in a **house** have higher default risk.