EDA Credit Assignment

Problem Statement – I

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Problem Statement - II

Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.
Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
Identify if there is data imbalance in the data. Find the ratio of data imbalance.
Use a mix of univariate and bivariate analysis etc.
Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
Find the top 10 correlation for the **Client with payment difficulties** and **all other cases** (Target variable).
Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: **Var1, Var2, Var3, Var4, Var5, Target**. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.
Include visualisations and summarise the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the **clients with payment difficulties with all other cases.**

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment. To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Loan Application Data

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_
0	100002	1	Cash loans	M	N	Y	0	202500.0	
1	100003	0	Cash loans	F	N	N	0	270000.0	1
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	
3	100006	0	Cash loans	F	N	Y	0	135000.0	
4	100007	0	Cash loans	M	N	Y	0	121500.0	
...	...	...	...	...	...	...	...	...	
307506	456251	0	Cash loans	M	N	N	0	157500.0	
307507	456252	0	Cash loans	F	N	Y	0	72000.0	
307508	456253	0	Cash loans	F	N	Y	0	153000.0	
307509	456254	1	Cash loans	F	N	Y	0	171000.0	
307510	456255	0	Cash loans	F	N	N	0	157500.0	

307511 rows × 122 columns

Previous Loan Application Data

	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
...	...	...	...	...	...	...	...	...
1670209	2300464	352015	Consumer loans	14704.290	267295.5	311400.0	0.0	267295.5
1670210	2357031	334635	Consumer loans	6622.020	87750.0	64291.5	29250.0	87750.0
1670211	2659632	249544	Consumer loans	11520.855	105237.0	102523.5	10525.5	105237.0
1670212	2785582	400317	Cash loans	18821.520	180000.0	191880.0	NaN	180000.0
1670213	2418762	261212	Cash loans	16431.300	360000.0	360000.0	NaN	360000.0

1670214 rows × 37 columns

Assumptions:

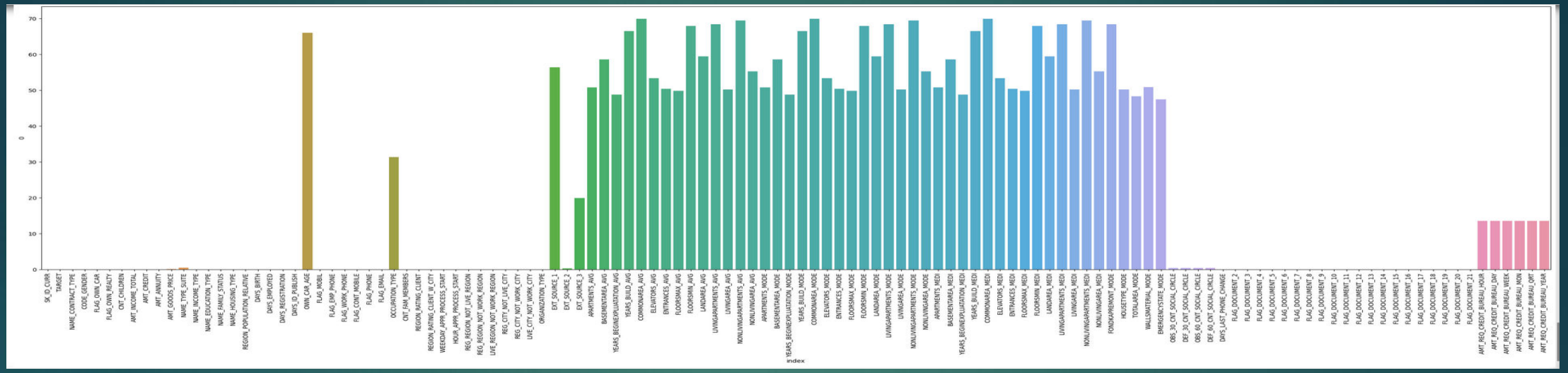
- There are so many missing values in columns so I dropped the columns which are having null values more than 40% in Application data.
- There are outliers in the numerical columns so I handle the outliers with imputation of mean, median and mode in Application data.
- There are some columns which needs Standardization like DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION so,its converted into years.

Approach & Methodology:

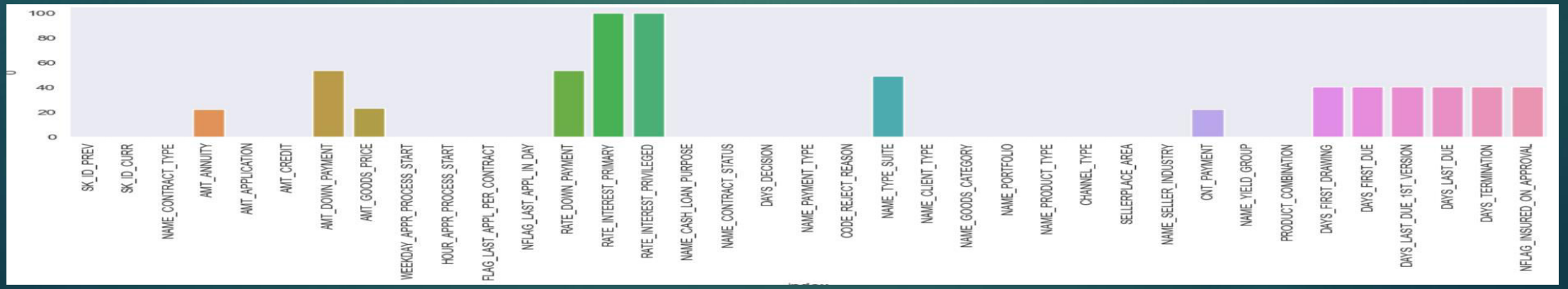
- Checking the missing values
- Handling outliers.
- Differentiates numerical columns and categorical columns.
- Univariate analysis, Bivariate analysis.
- Data Imbalanced Ratio.
- Merge the data sets.
- Univariate and Bivaraite analysis.
- Correlations.

Missing Value Ratio:

Application Data

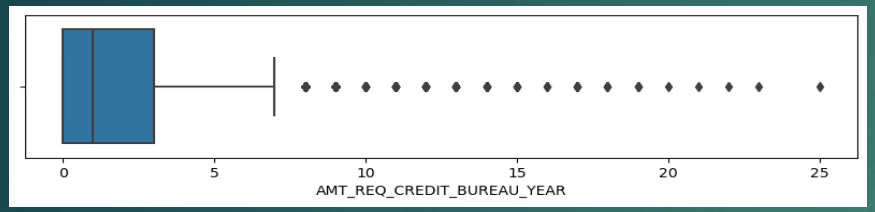


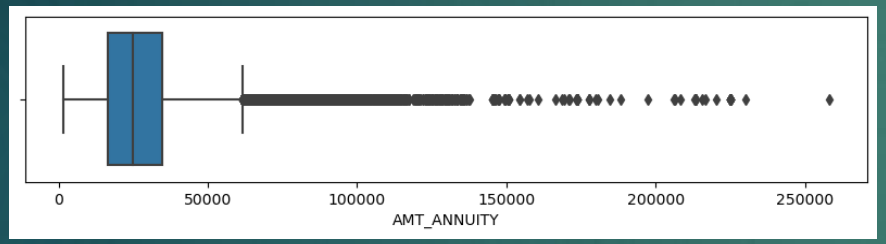
Previous Loan Application Data

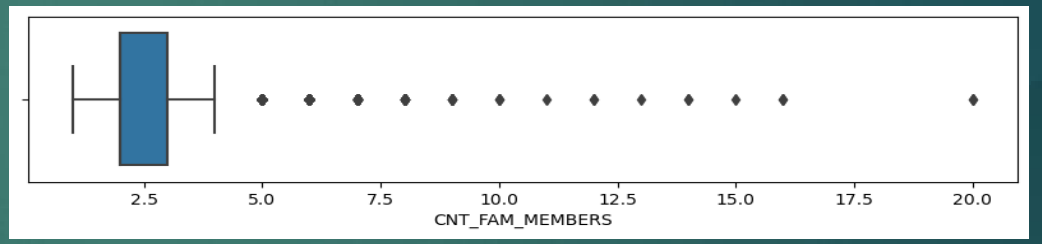


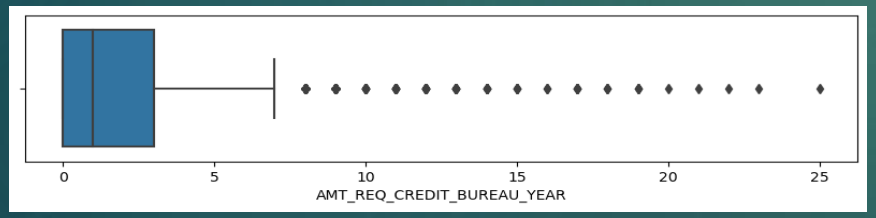
Outliers in Application Data:

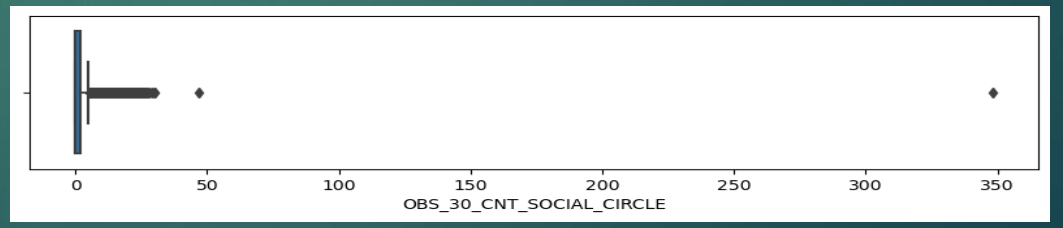
There so many columns with outliers here are some insights with box plots.











Diffrentiate Between Categorical columns and Numerial Columns¶

Categorical columns

NAME_CONTRACT_TYPE
CODE_GENDER
FLAG_OWN_CAR
FLAG_OWN_REALTY
AMT_ANNUITY
AMT_GOODS_PRICE
NAME_TYPE_SUITE
NAME_INCOME_TYPE
NAME_EDUCATION_TYPE
NAME_FAMILY_STATUS
NAME_HOUSING_TYPE
OCCUPATION_TYPE
WEEKDAY_APPR_PROCESS_START
ORGANIZATION_TYPE

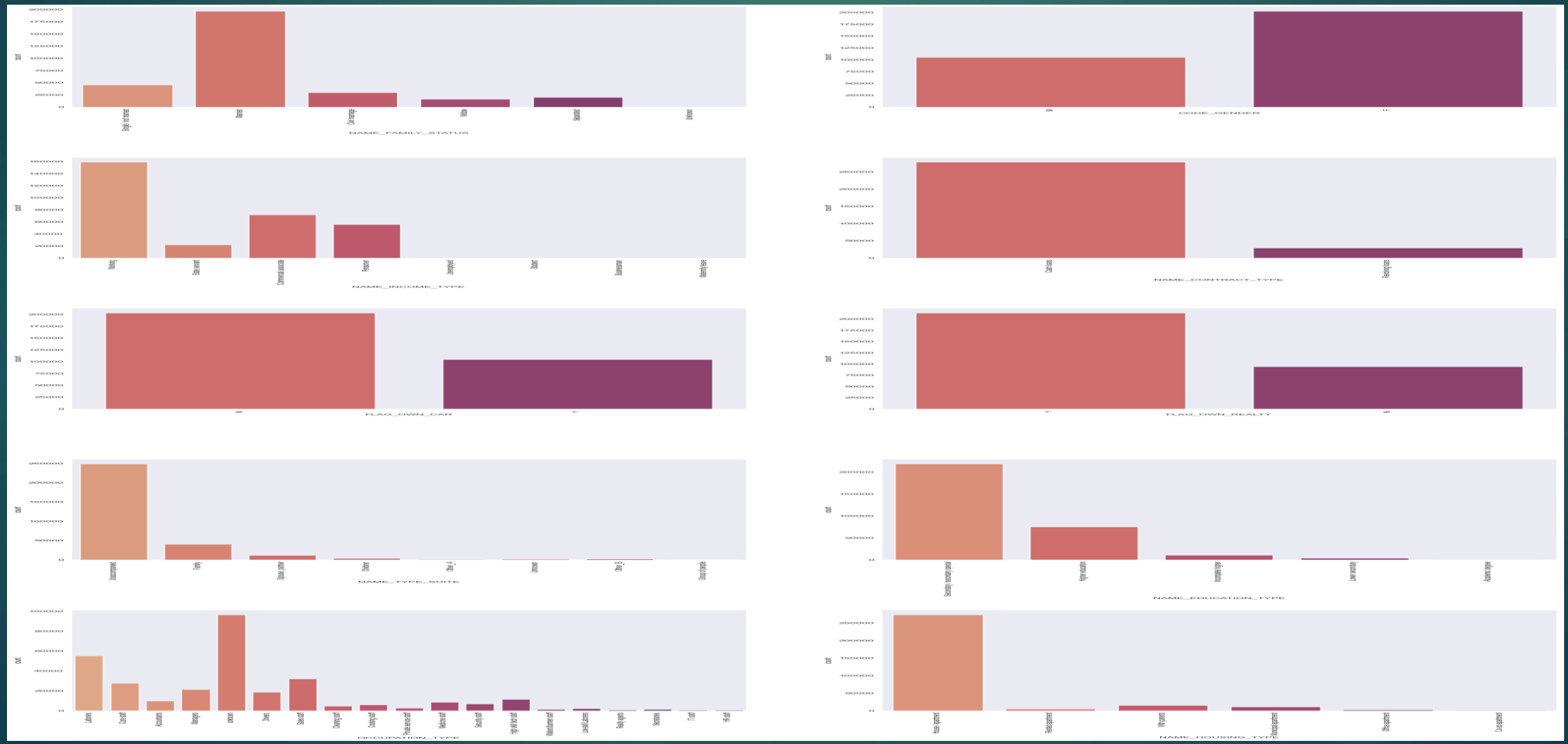
Numerial Columns

SK_ID_CURR
TARGET
CNT_CHILDREN AMT_INCOME_TOTAL
AMT_CREDIT
REGION_POPULATION_RELATIVE
FLAG_MOBIL FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
ETC.

UNIVARIATE ANALYSIS:

Some univariate analysis was performed on the below columns:

NAME_FAMILY_STATUS, CODE_GENDER,NAME_INCOME_TYPE, NAME_CONTRACT_TYPE, FLAG_OWN_CAR, FLAG_OWN_REALTY, NAME_TYPE_SUITE, NAME_EDUCATION_TYPE, OCCUPATION_TYPE, NAME_HOUSING_TYPE



BIVARIATE ANALYSIS:

Some bivariate analysis was performed on the below columns:

NAME_FAMILY_STATUS, CODE_GENDER,NAME_INCOME_TYPE, NAME_CONTRACT_TYPE, FLAG_OWN_CAR, FLAG_OWN_REALTY, NAME_TYPE_SUITE, NAME_EDUCATION_TYPE, OCCUPATION_TYPE, NAME_HOUSING_TYPE

OBSERVATIONS:

Target1 :
- The client with payment difficulties, Target0 : The all other type.(repayor).

CODE_GENDER :
- The % of defaulters are more in Female than Male.

NAME_INCOME_TYPE :
- Student and businessman are higher in percentage of loan repayment.
- Working, State servant and Commercial associates are higher in default percentage.
- Maternity category is significantly higher problem in repayment.

NAME_CONTRACT_TYPE :
- For contract type ‘Cash loans’ are high in number of credits than ‘Revolving loans’ contract type.
- By above graph 'Revolving loans' is small amount compared to ‘Cash loans’.

OCCUPATION_TYPE
- HR staff, Secretaries, Realty agents and IT staff are very less in both.
- Laborers, Drivers and Low skill Laborers are higher in percentage of loan repayment.

NAME_HOUSING_TYPE
- people living in House apartment are more in percentage of loan repayment.

NAME_EDUCATION_TYPE
- People are having secondary education and higher education are significantly higher in loan repayment.

NAME_FAMILY_STATUS
- Married are higher with non difficulties with payment.

FLAG_OWN_REALTY and FLAG_OWN_CAR
- The people who having realty and not having car are having significantly higher in payment with difficulties.

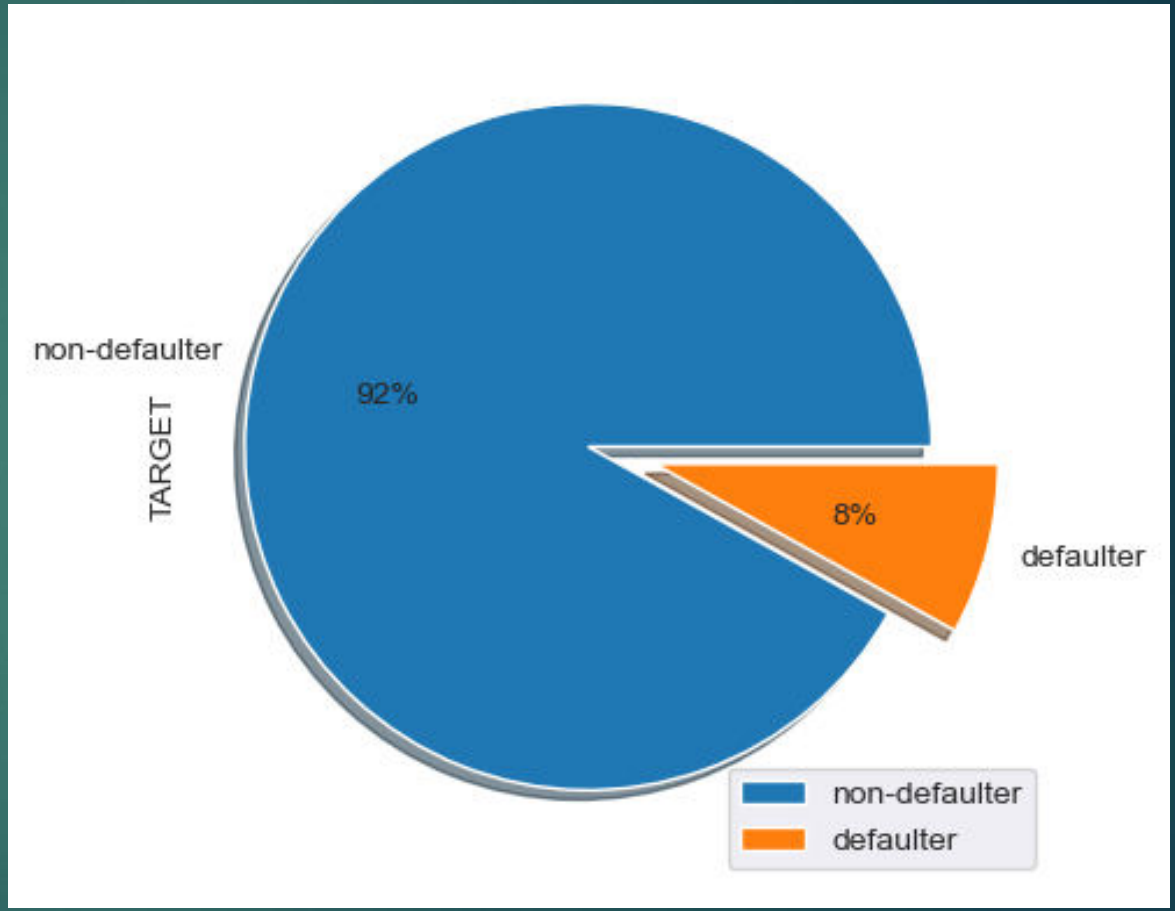
NAME_TYPE_SUITE
- The people unaccompanied are more in percentage of loan repayment.

RATIO OF DATA IMBALANCED:

To check the data is balance or not we will count the values in the target column in the application data.

Data Imbalanced Ratio : 11.39

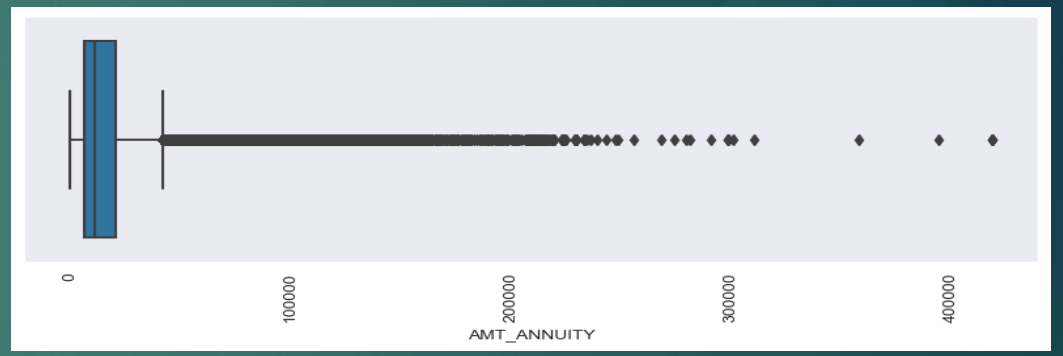
Data is highly imbalanced

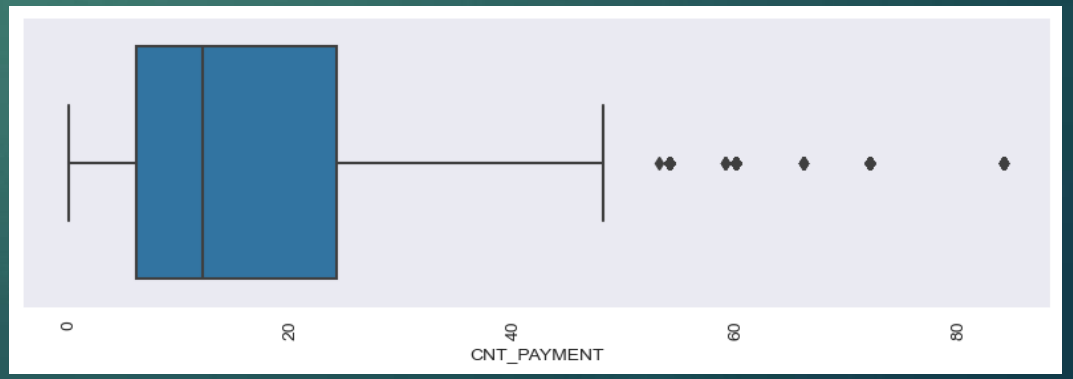


OUTLIERS IN PREVIOUS LOAN DATA :

There are some insights of outliers with box plots.

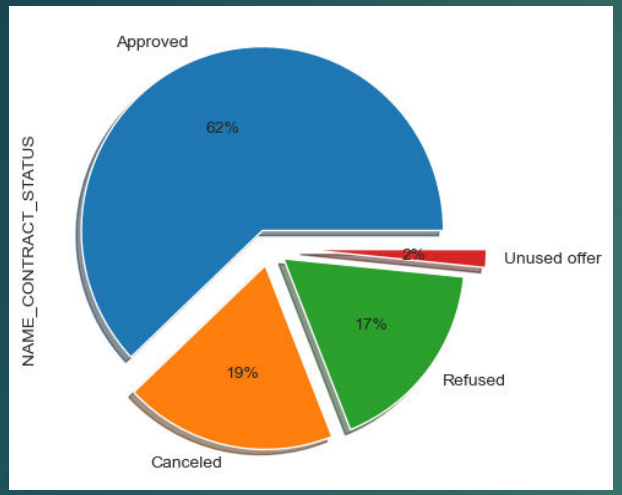
Handle these outliers by creating bins.



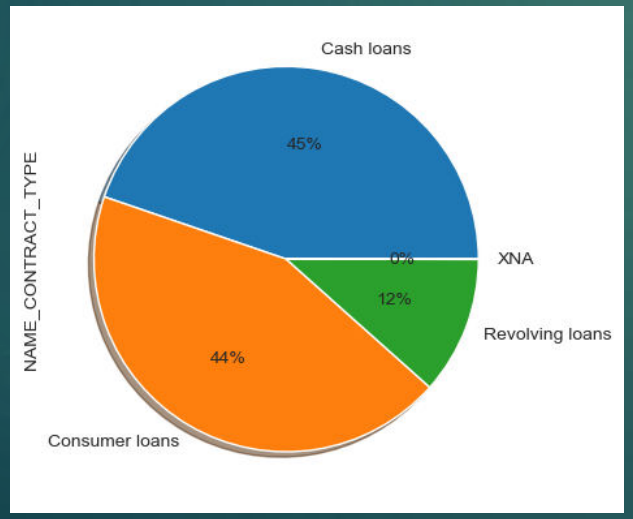


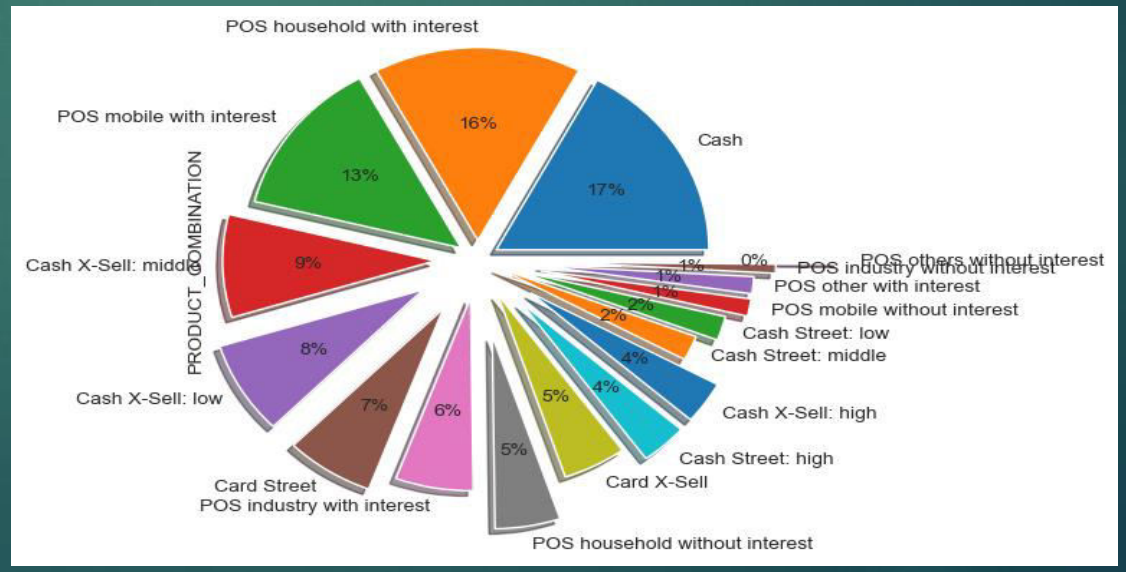
Univariate Analysis :

Some univariate analysis was performed on the below columns:



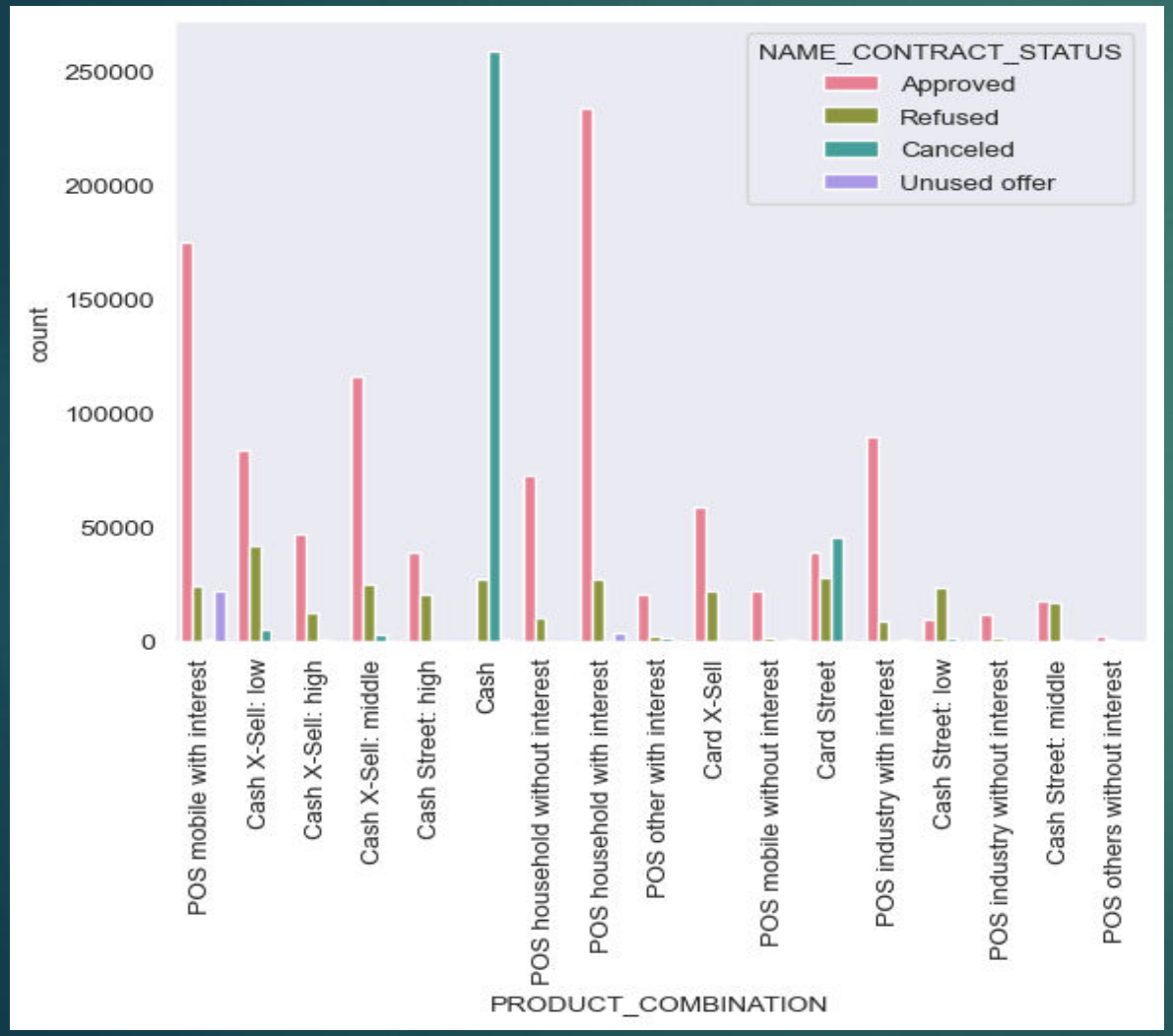
- CONTRACT TYPE
- CONTRACT STATUS
- PRODUCT COMBINATIONS

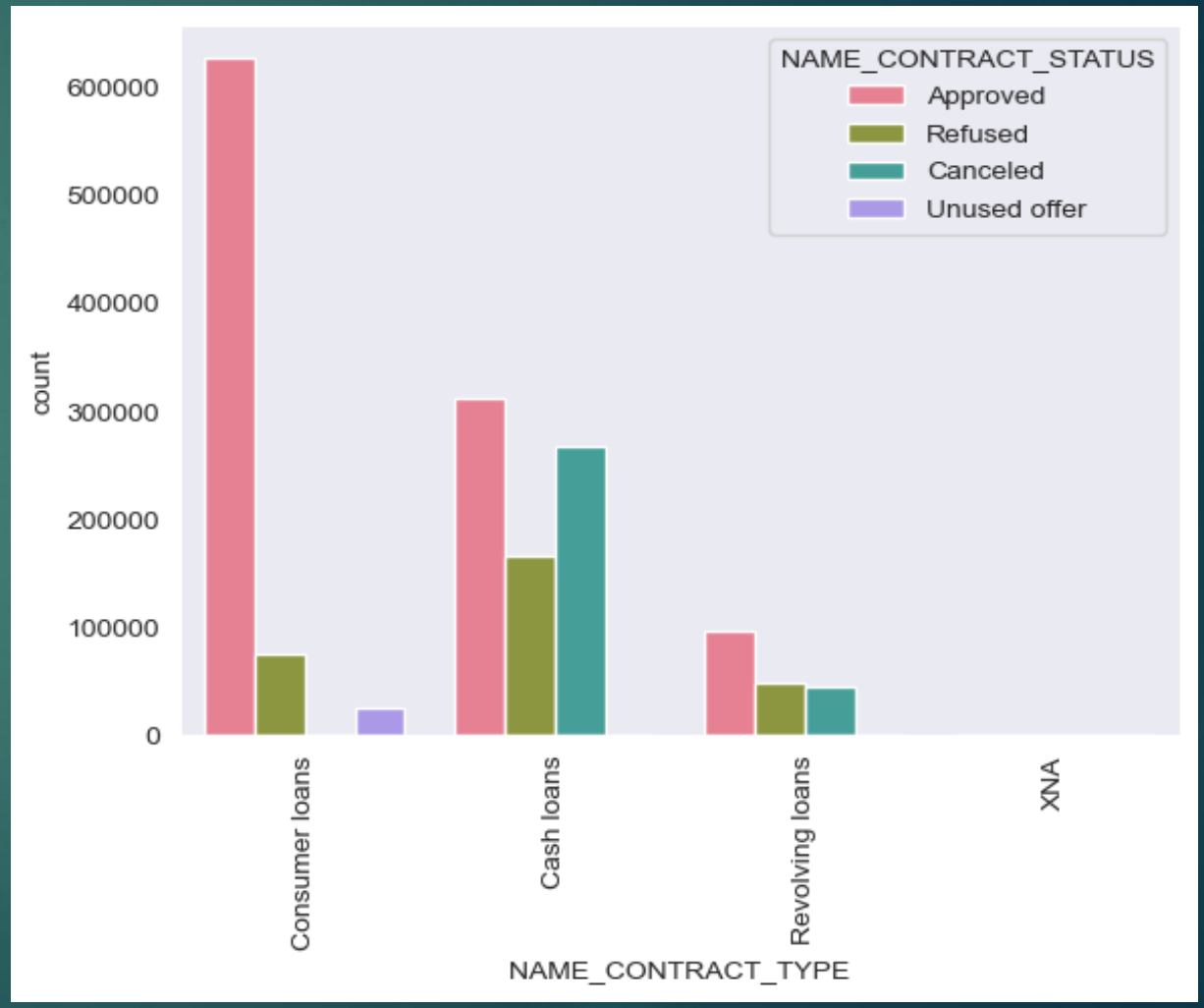




BIVARIATE ANALYSIS:

Some bivariate analysis was performed on the below columns:

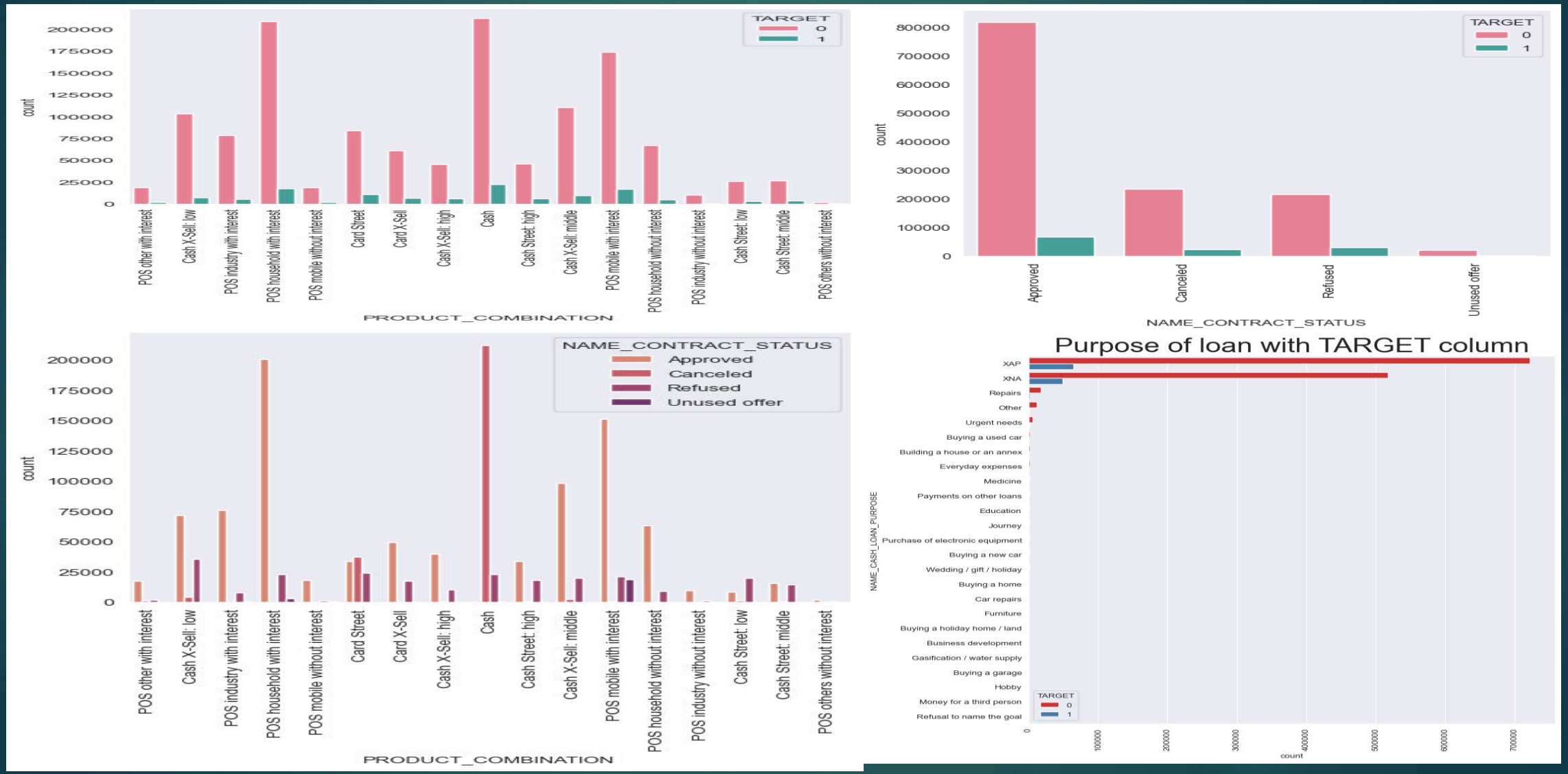




OBSERVATIONS:

- The product combination with cash is having significantly higher in cancelled ratio of loan.
- Consumer loans are high in approved loan status.
- Cash loan type are significantly high in cancelation of loan.
- POS mobile with interest are high in unused offers.
- Consumer loans are not having any calcelations.

BIVARIATE ANALYSIS AFTER MERGING APPLICATION DATA WITH THE PREVIOUS LOAN APPLICATIONS:



OBSERVATIONS:

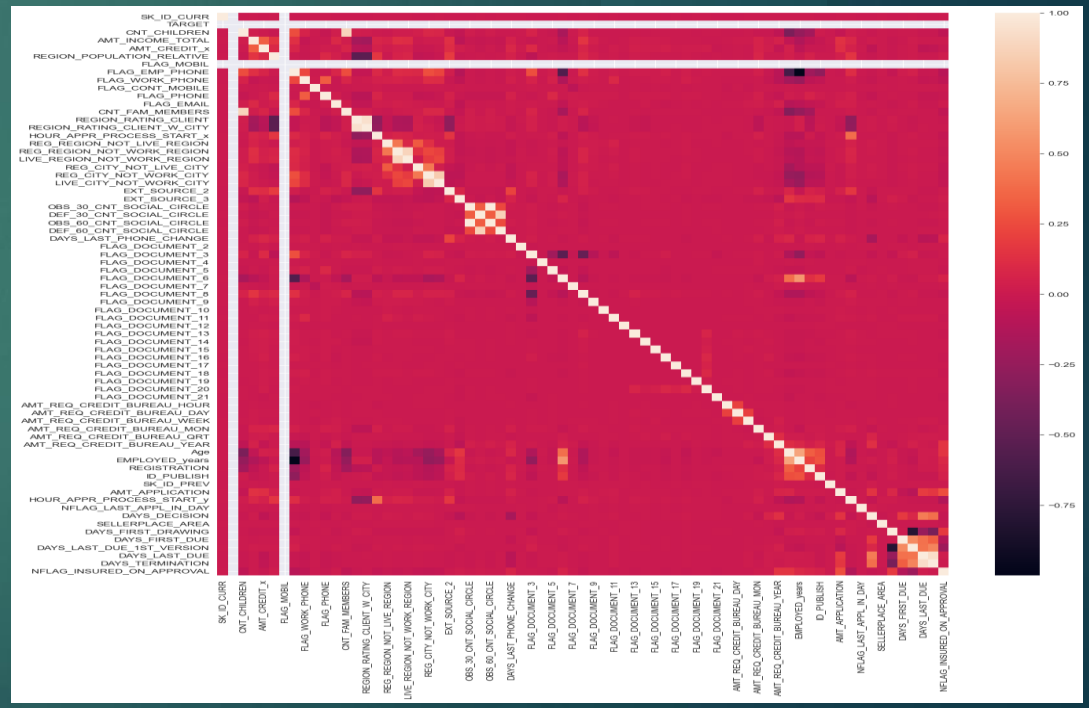
- The product combination with cash is having significantly higher in cancelled ratio of loan.
- Consumer loans are high in approved loan status.
- Most of loan rejection was from 'Repairs'
- Most of the Repairs are with payment difficulties.

CORRELATIONS:

The highest corelation (1.0) is between OBS_60_CNT_SOCIAL_CIRCLE WITH OBS_30_CNT_SOCIAL_CIRCLE.

Correlation between people, payment with all others

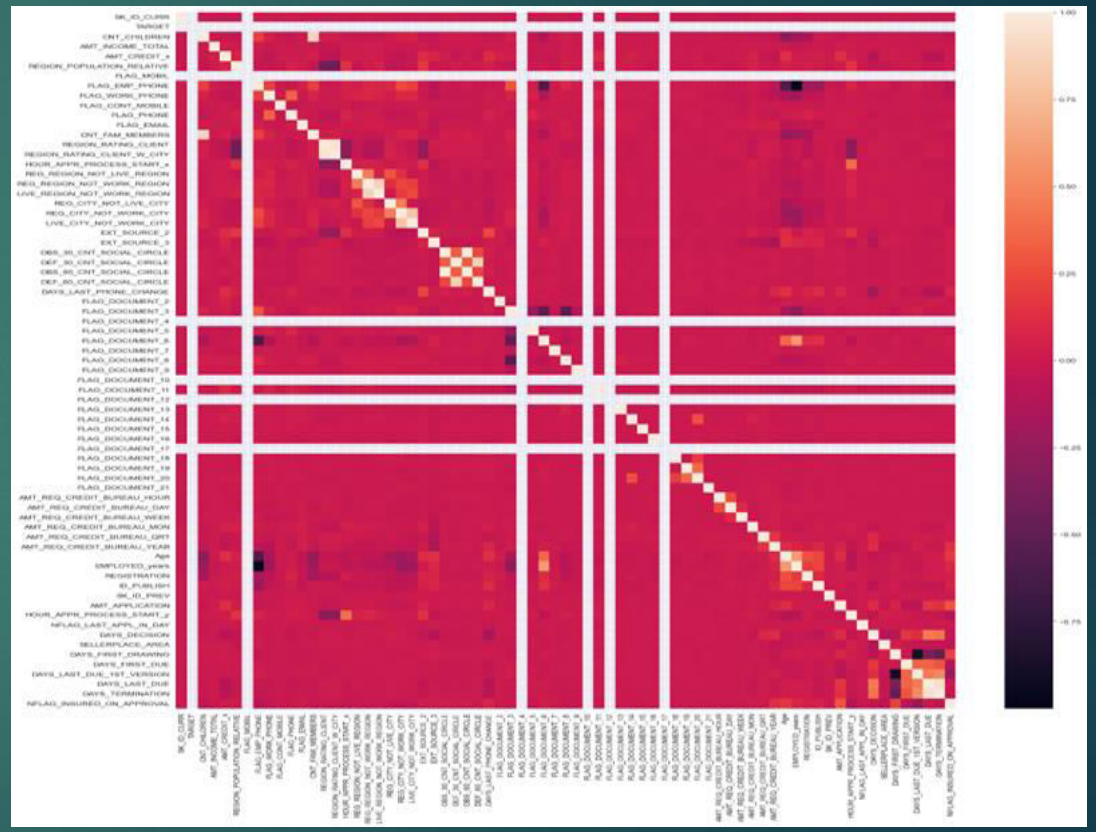
	Var1	Var2	Correlation
1870	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
3983	EMPLOYED_years	FLAG_EMP_PHONE	1.00
1007	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.94
4967	DAYS_TERMINATION	DAYS_LAST_DUE	0.93
854	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
1295	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.88
1942	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
1511	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.84
4822	DAYS_LAST_DUE_1ST_VERSION	DAYS_FIRST_DRAWING	0.80
3912	Age	FLAG_EMP_PHONE	0.63



Correlation between people, payment with difficulties

The highest corelation (1.0) is between (EMPLOYED_years with FLAG_EMP_PHONE)

	Var1	Var2	Correlation
3983	EMPLOYED_years	FLAG_EMP_PHONE	1.00
1870	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
1007	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96
4967	DAYS_TERMINATION	DAYS_LAST_DUE	0.94
854	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
4822	DAYS_LAST_DUE_1ST_VERSION	DAYS_FIRST_DRAWING	0.89
1295	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.87
1942	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
1511	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.79
4031	EMPLOYED_years	Age	0.59



THANK YOU