



# Presentation for Credit EDA Assignment




## Problem Statement:

The aim of this case study is to identify patterns which indicate if a client has difficulty paying their installments 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.

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.



The dataset contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- All other cases:** All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

- 1.Approved:** The Company has approved loan Application
- 2.Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
- 3.Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- 4.Unused offer:** Loan has been cancelled by the client but on different stages of the process.



This dataset has 3 files as explained below:

1. *'application\_data.csv'* contains all the information of the client at the time of application.

The data is about whether a **client has payment difficulties**.

2. *'previous\_application.csv'* contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer**.

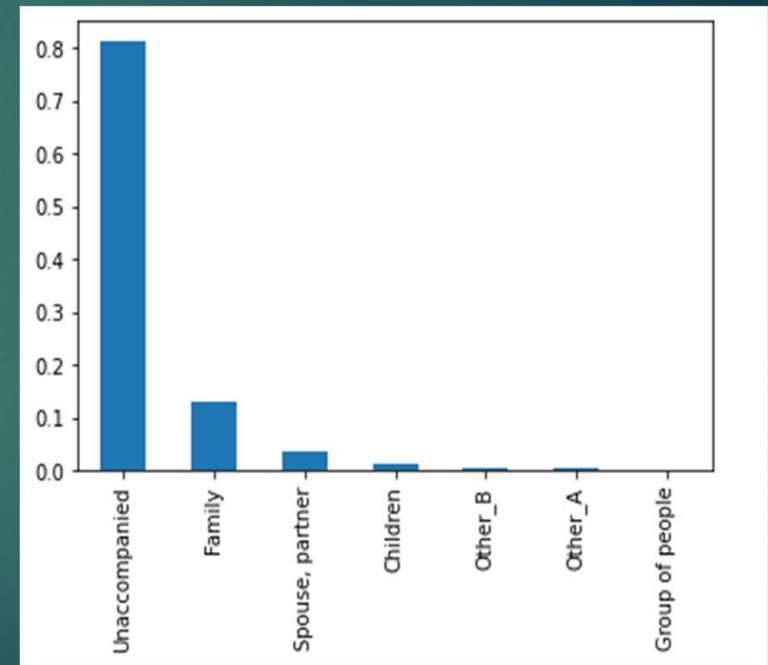
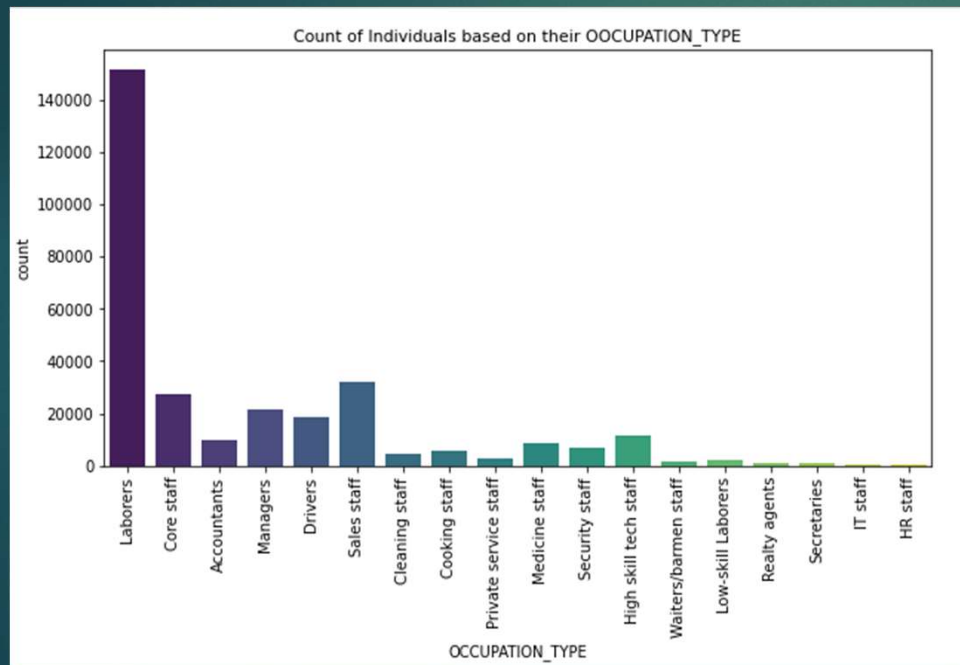
3. *'columns\_description.csv'* is data dictionary which describes the meaning of the variables.

## Steps to be performed for analysing the data

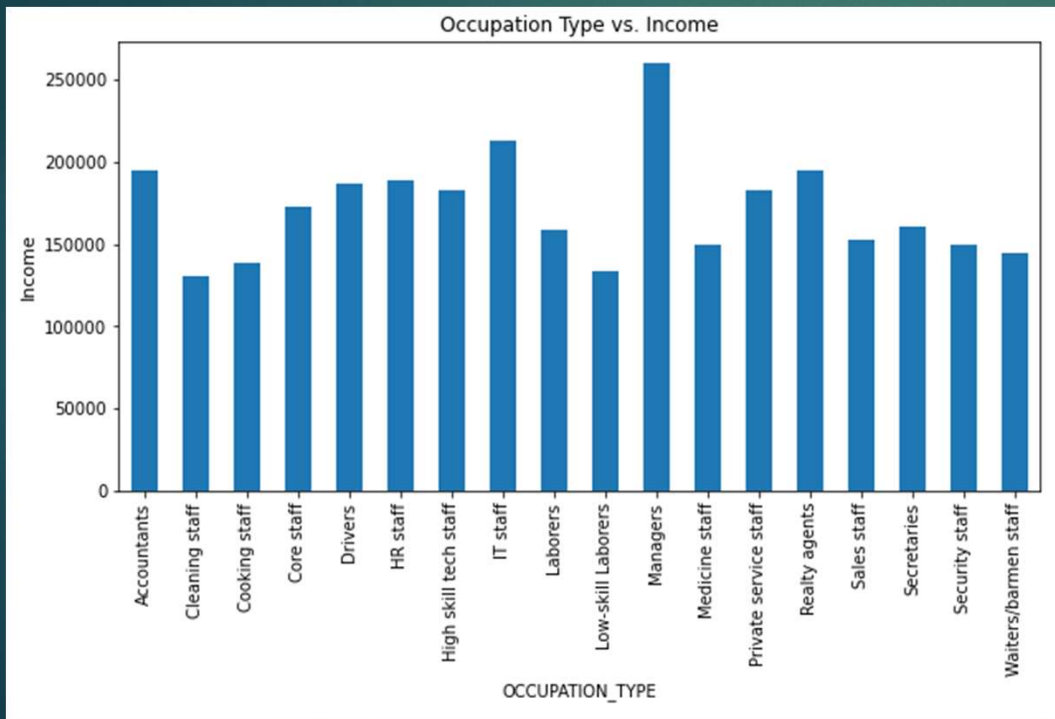
1. Import the libraries
2. Read the application\_data.csv file
3. Check the shape of the dataframe
4. Find the percentage of missing values for all columns
5. Remove the columns with high missing percentage value (>50%)
6. Impute the values based on their column type (categorical/numerical) with mean, median, mode.
7. Check Data type for all columns and change the data type if required.
8. Find the outliers and treat the outliers by providing appropriate reasoning
9. Check if you need to bin any variable in different categories
10. Check the Imbalance percentage
11. Divide the dataset in to two
12. perform univariate and bivariate analysis
13. Find the correlation for numerical columns and get better insight by using heatmap
14. Merge the two datasets (application\_data, previous\_application) and perform univariate and bivariate analysis to find some pattern

- Majority of the individuals applying for loans are laborers, Sales Staff, Core Staff, Managers and Drivers
- HR Staff, IT Staff, Realty agents, Secretaries and Waiters/Barmen Staff have least interest in taking loans

Most of the times, people come unaccompanied while applying for the loan



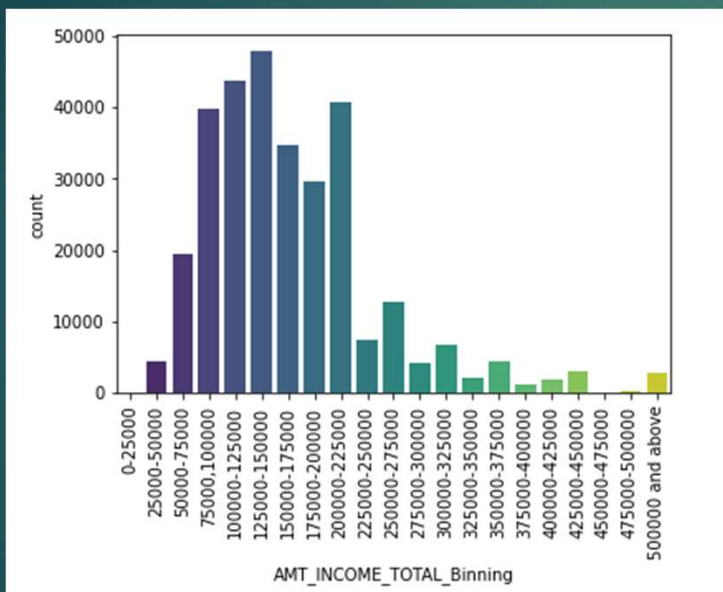
- we can say that Managers, IT staff, Realty agents, Accountants and HR staff are having Highest Average Income
- Cleaning Staff, Low-skill laborers and cooling staff are having Least Average Income



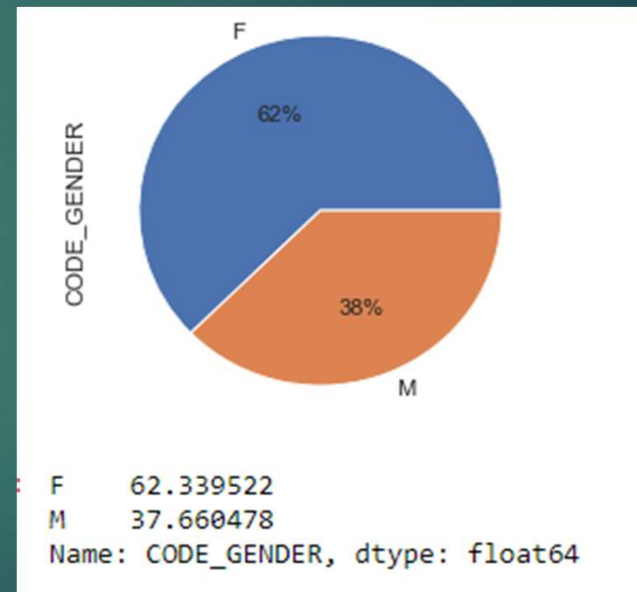
Based on the graph, we conclude that person who is earning between 25% - 50%(Low) are more likely to take the loan



Income range from 125000-150000, 100000-125000, 200000-225000, 75000-100000 has high chances of applying for loan

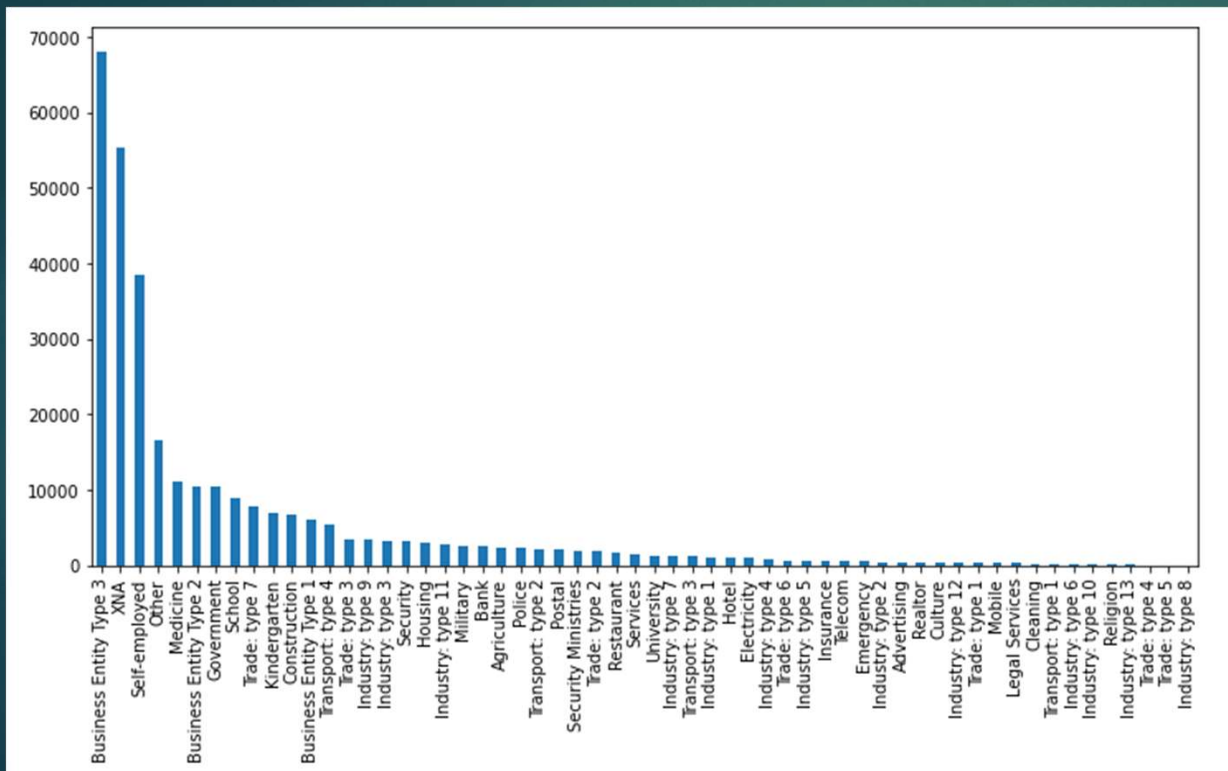


From the above Bar Graph and the percentage of values in CODE\_GENDER, 62% females are taking loans whereas 37% Males are taking the loans.

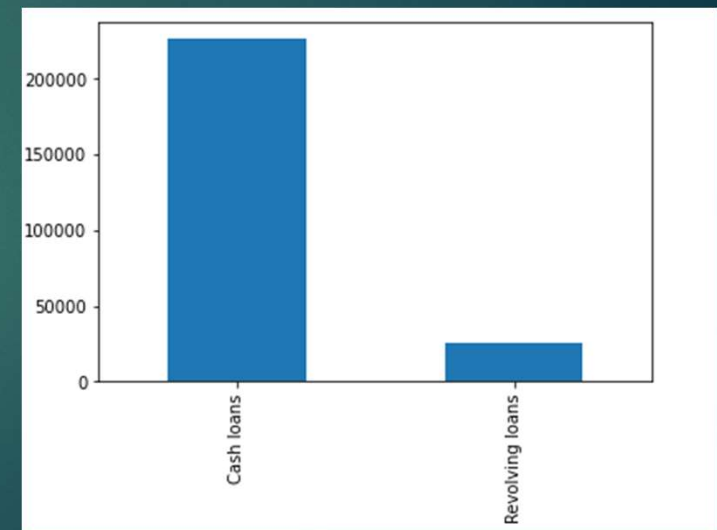




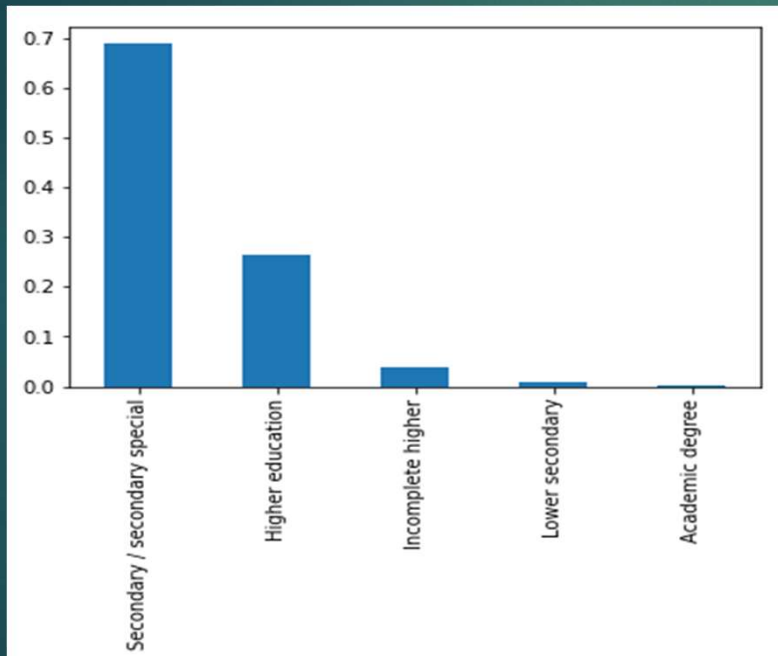
- XNA Value percentage is around 18% so we can drop the rows which are having the XNA Values. Business Entity Type 3, Self Employed are the types of organization which are more likely to take the loan



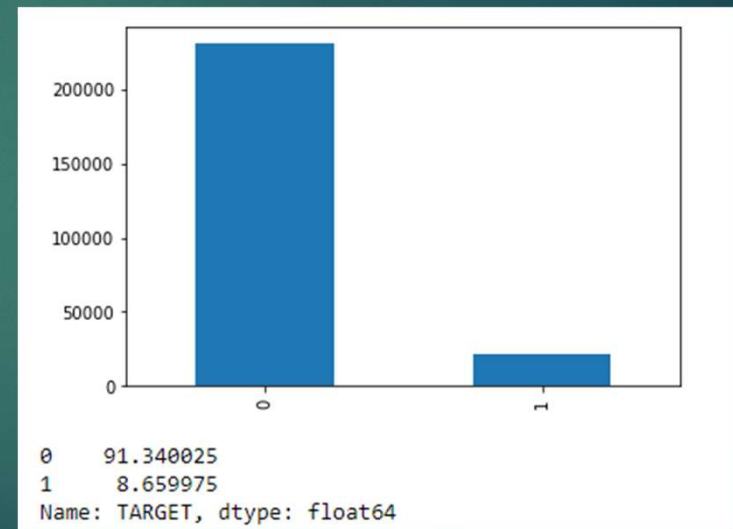
- From the Bar Graph for NAME\_CONTRACT\_TYPE, we can conclude that people are more likely to take cash loans rather than revolving loans

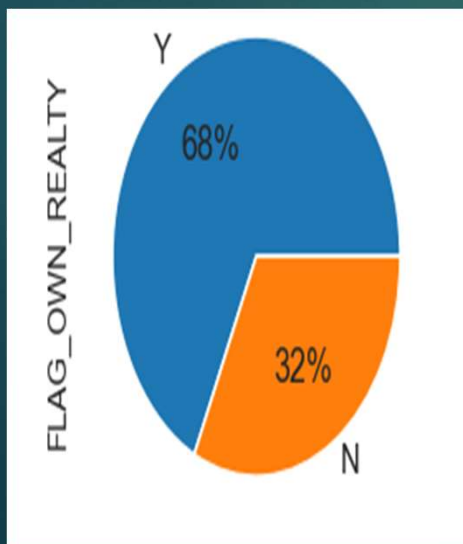
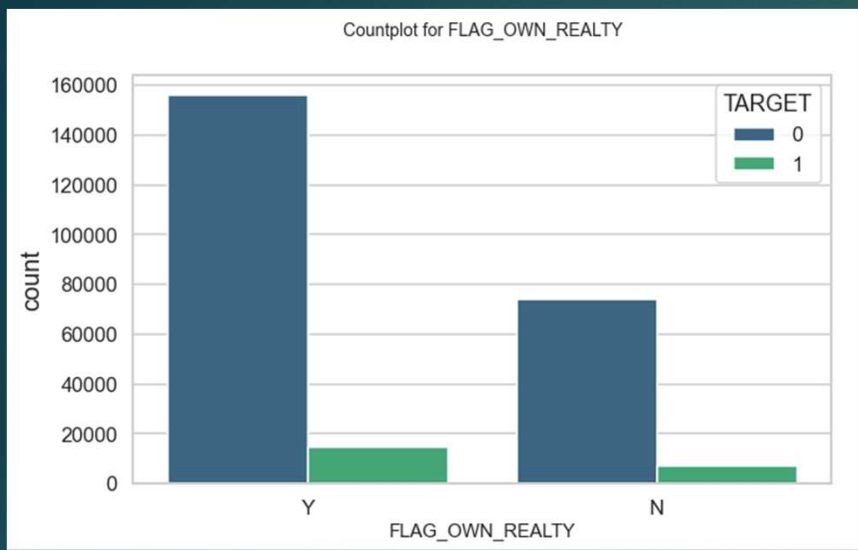


From the above Bar Graph, we can conclude that Secondary/secondary special are more likely to take loans from Bank



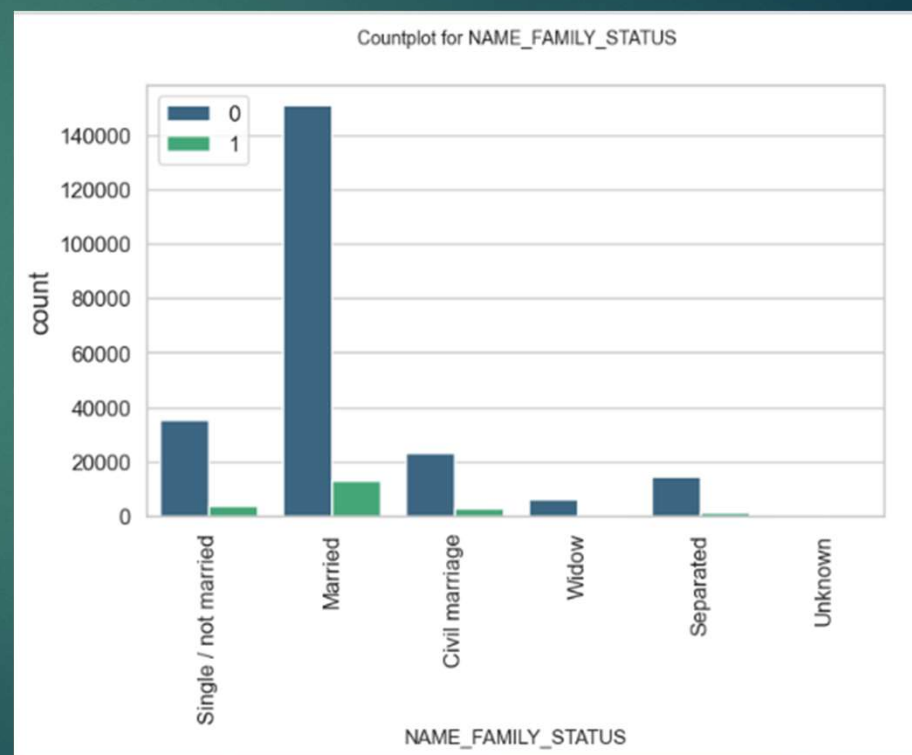
- From the Bar Graph, we can conclude that 90% of the individuals have paid their loan successfully and 8% of individuals are facing the difficulties while paying the loan
- Also we can clearly say that this is imbalanced dataset
- We need to analyze the dataset further by splitting the dataset into two parts based on 0 and 1.

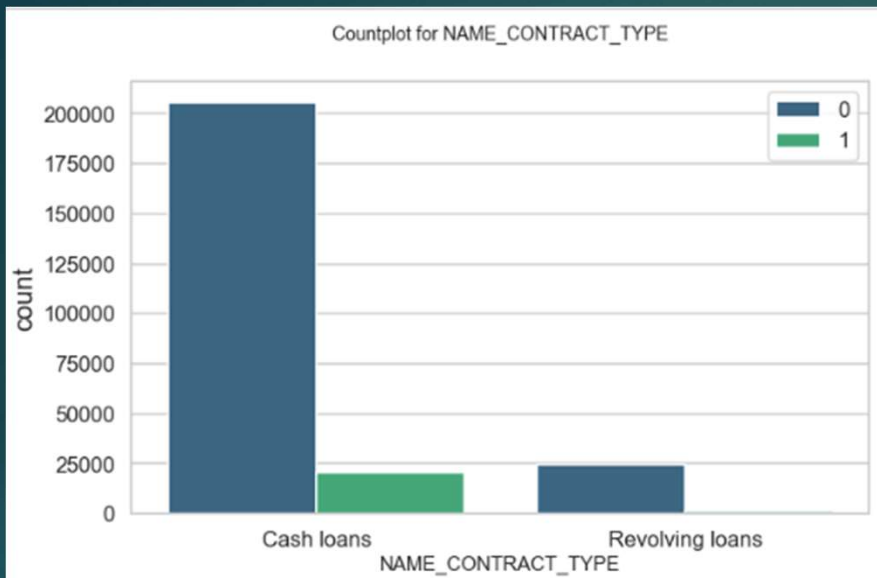




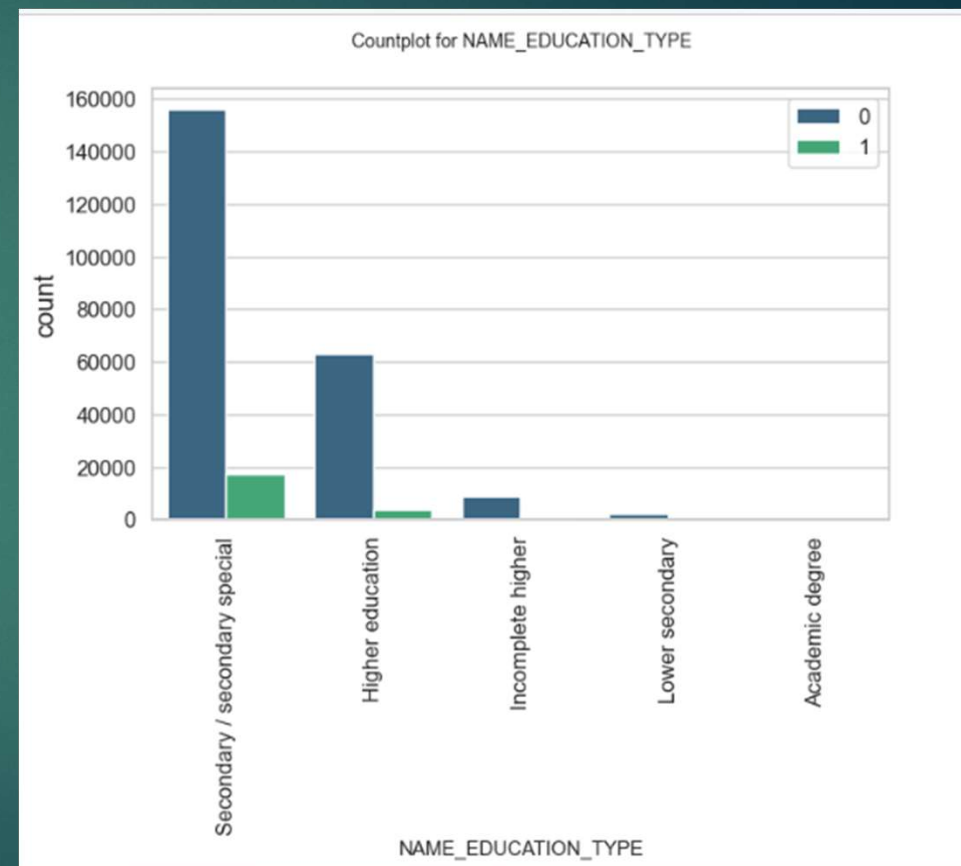
- Individuals having house/flat are more likely to take the loans
- People who are married are tend to take the loan followed by Single/Not Married as compared to other categories

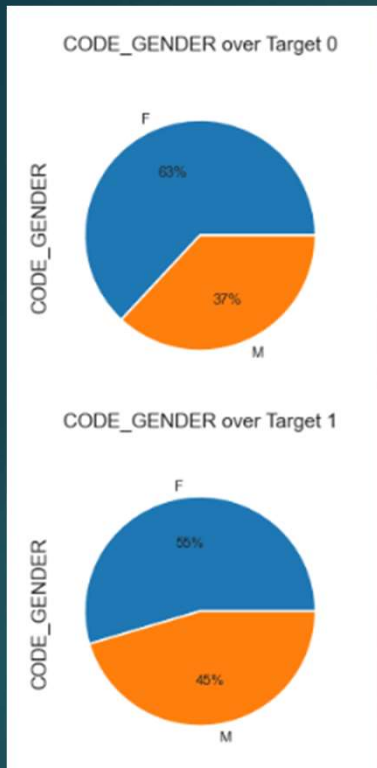
## Univariate Analysis on Categorical Columns



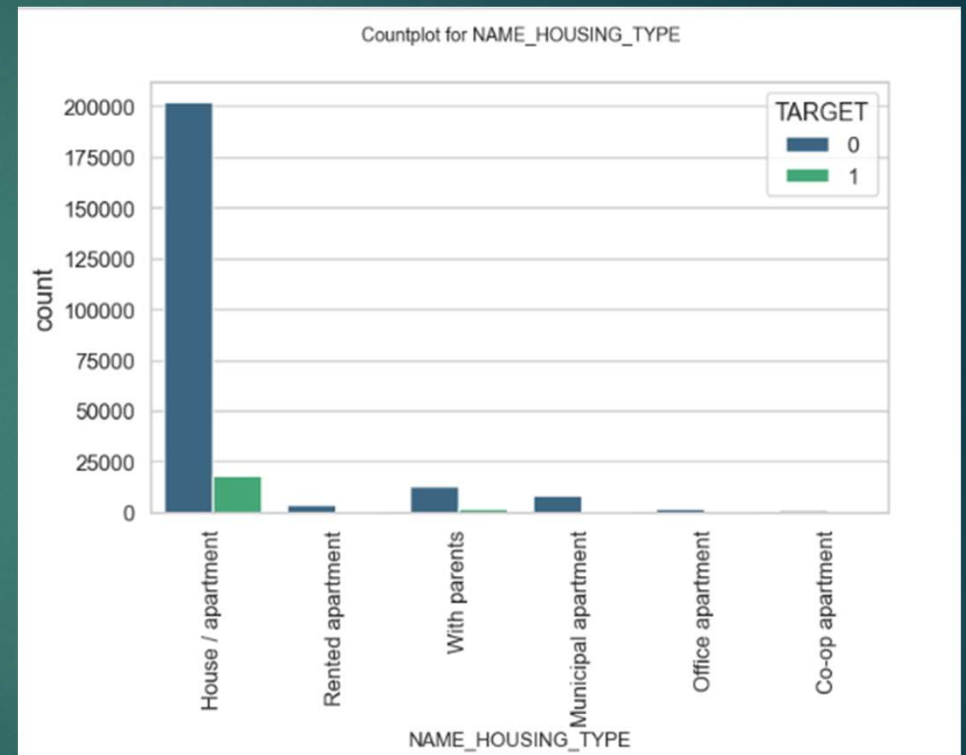


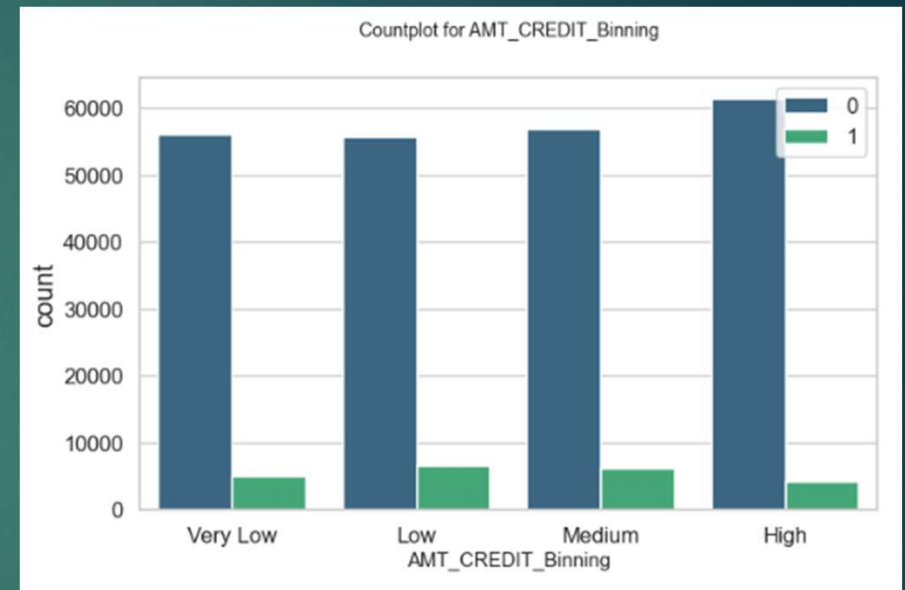
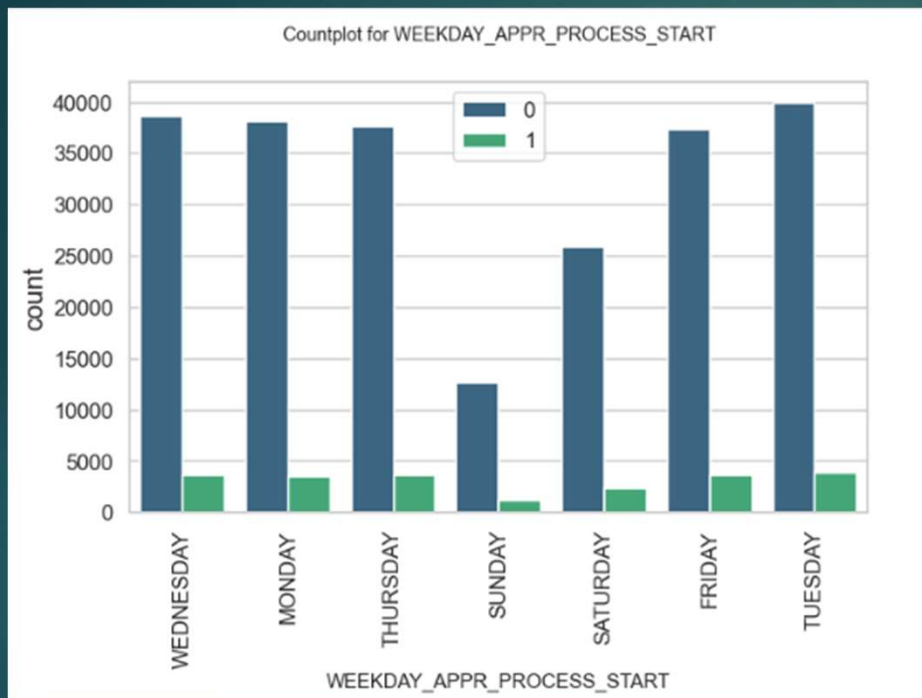
- People are more likely to take the cash loans
- We can conclude that Secondary/ Secondary Special are taking high number of loans



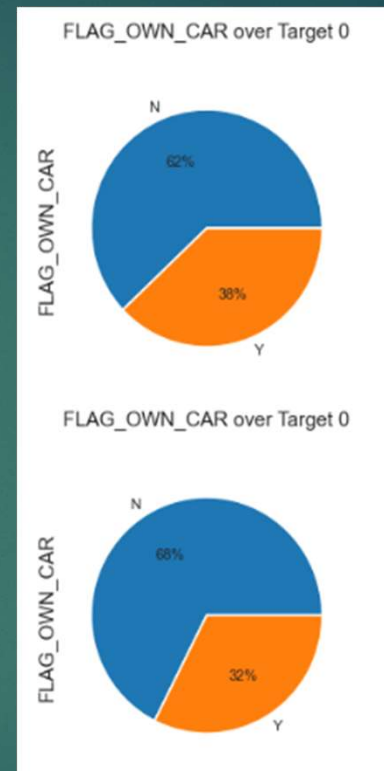
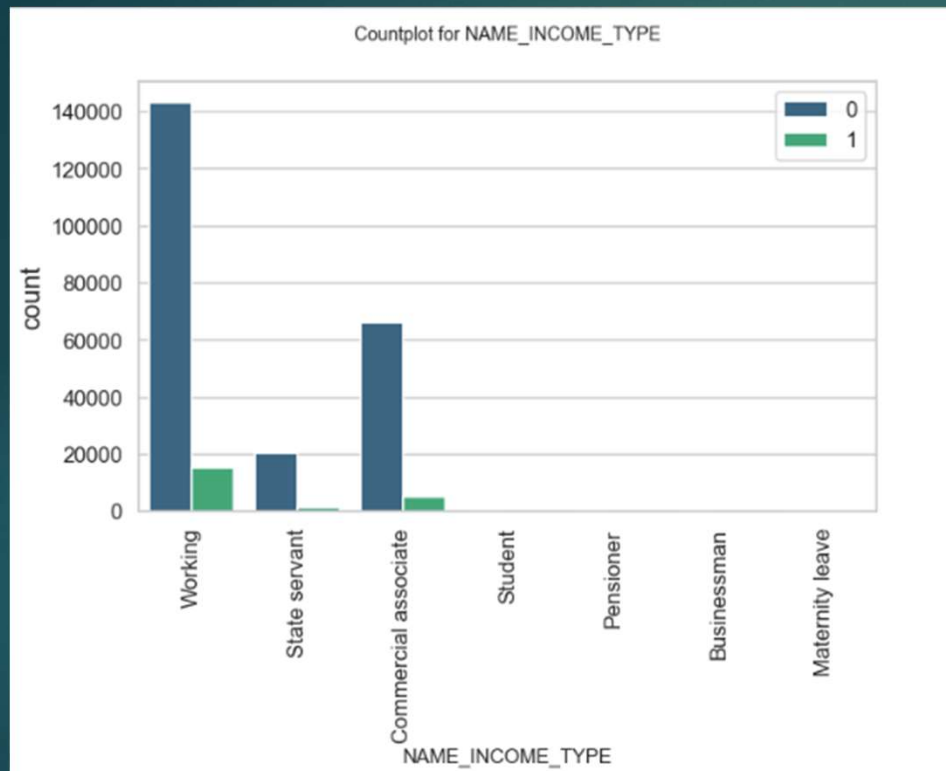


- Female tends to take more loans
- People with House/Apartment are tend to take more loans

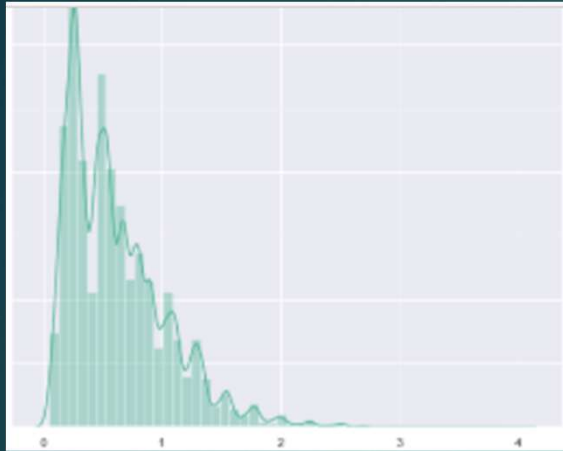




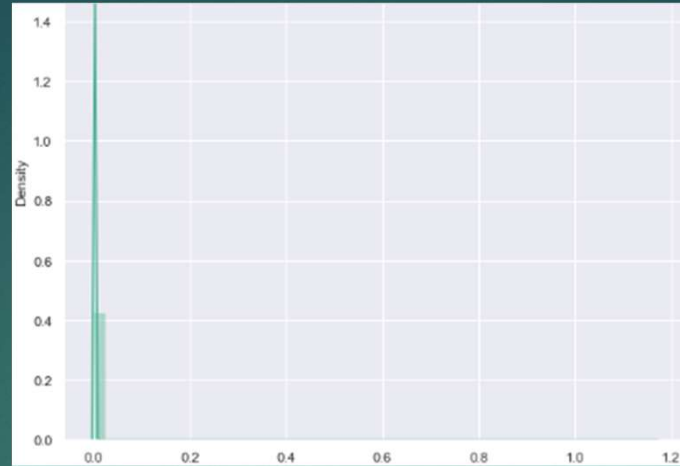
- People who start their application on Sunday are less likely to take loan whereas people who start their Application on Tuesday are more likely to take loan
- Income with 125000-150000 are more tend to take the loan



- People whose income type is working and commercial associate are more likely to take the loan
- people with no car tend to take more loans



AMT\_CREDIT



AMT\_INCOME\_TOTAL

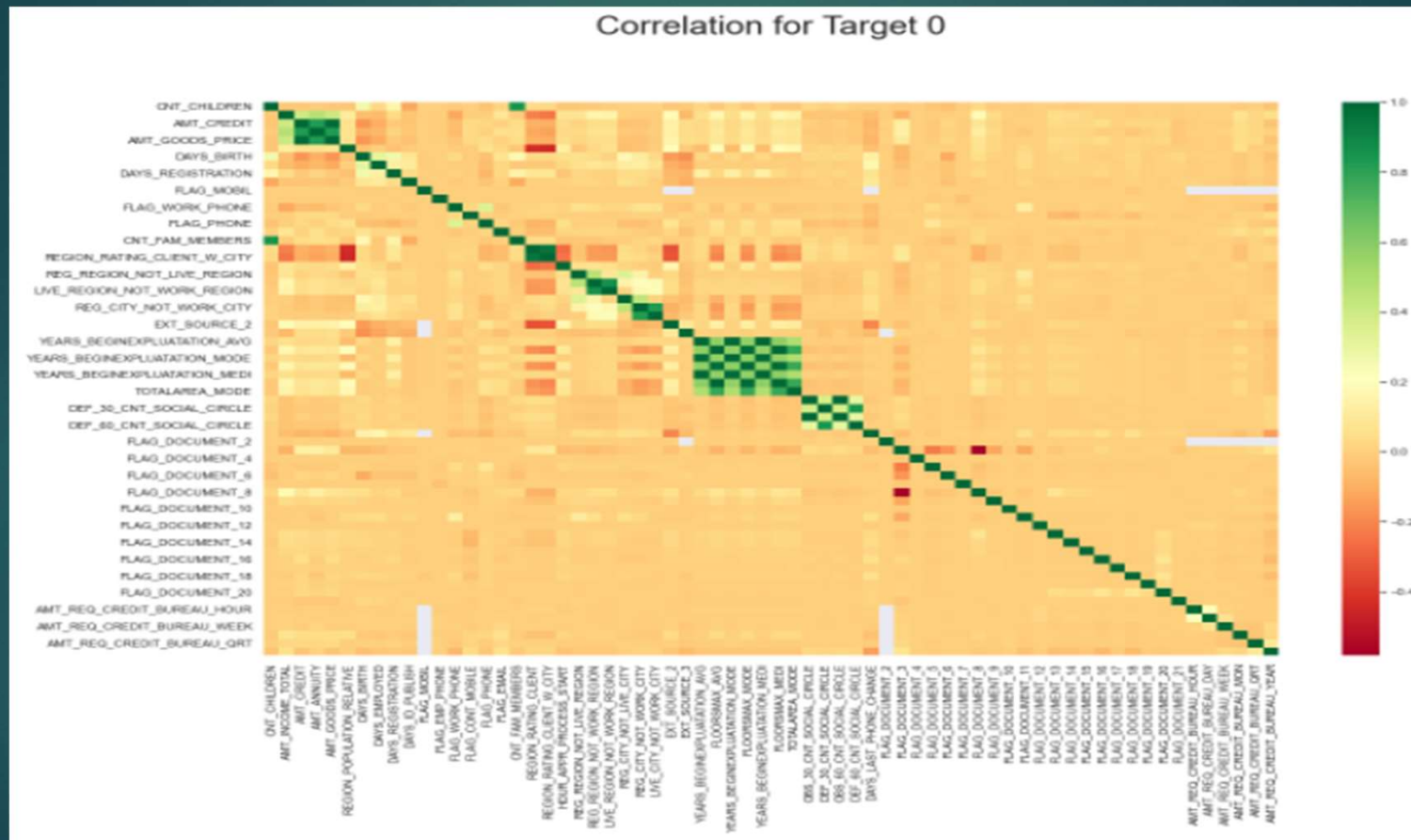
- For less goods amount people take loans
- Low amount annuity has high number of loans
- People whose ids got published between 4000 days and 5000 days ago tend to take more loans
- Nuclear family tends to take more loans

## Univariate Analysis on Numerical Columns of Application\_Data

- People with lower total income are less likely to take loans
- People who just got employed tends to take more loans
- People who retired tends to take more loans
- High number of applications are filed in between 10 AM to 2 PM
- People with age between 10000-15000 days tend to take more loans



Using Heatmap, we have shown the Correlation for Application\_Data which consists of Target=0



Using Heatmap, we have shown the Correlation for Application\_Data which consists of Target=1



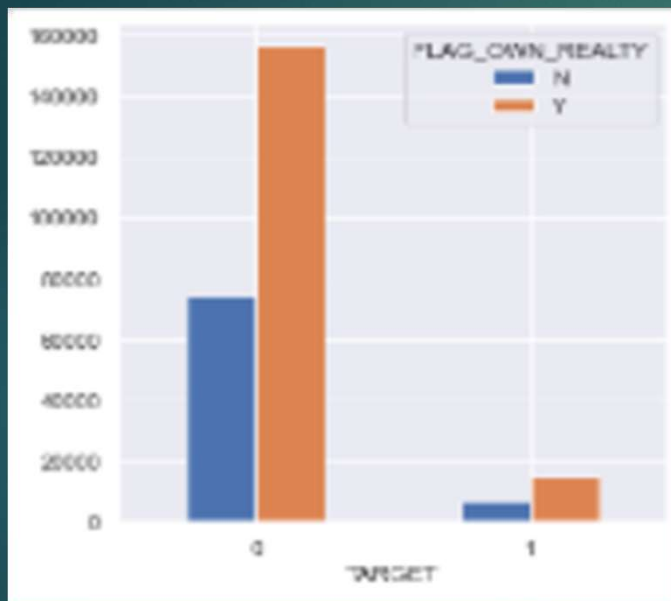
	Variable1	Variable2	Coorelation
2411	FLOORSMAX_MEDI	FLOORSMAX_AVG	1.00
2689	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
2342	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.99
412	AMT_GOODS_PRICE	AMT_CREDIT	0.99
2413	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.99
2275	FLOORSMAX_MODE	FLOORSMAX_AVG	0.99
2206	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.97
2344	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.96
1379	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.95
1226	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89

	Variable1	Variable2	Coorelation
2689	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
2411	FLOORSMAX_MEDI	FLOORSMAX_AVG	1.00
2342	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	1.00
2275	FLOORSMAX_MODE	FLOORSMAX_AVG	0.99
2413	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.99
2344	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.98
2206	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.98
412	AMT_GOODS_PRICE	AMT_CREDIT	0.98
1379	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96
1226	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89

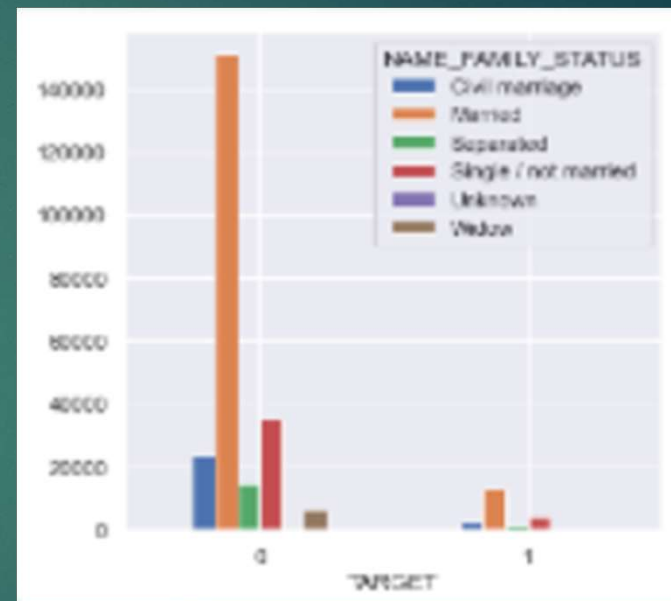
## Insights from Correlation

- When defaulted, major portion of decision is taken by- (YEARS\_BEGINEXPLUATATION\_MEDI AND YEARS\_BEGINEXPLUATATION\_AVG) (OBS\_60\_CNT\_SOCIAL\_CIRCLE AND OBS\_30\_CNT\_SOCIAL\_CIRCLE) (FLOORSMAX\_MEDI AND FLOORSMAX\_AVG)
- When not defaulted, major portion of decision is taken by- (FLOORSMAX\_MEDI AND FLOORSMAX\_AVG) (OBS\_60\_CNT\_SOCIAL\_CIRCLE AND OBS\_30\_CNT\_SOCIAL\_CIRCLE) (FLOORSMAX\_MEDI AND FLOORSMAX\_MODE)
- Values - YEARS\_BEGINEXPLUATATION\_MEDI and YEARS\_BEGINEXPLUATATION\_AVG are more correlated in default case than non-default

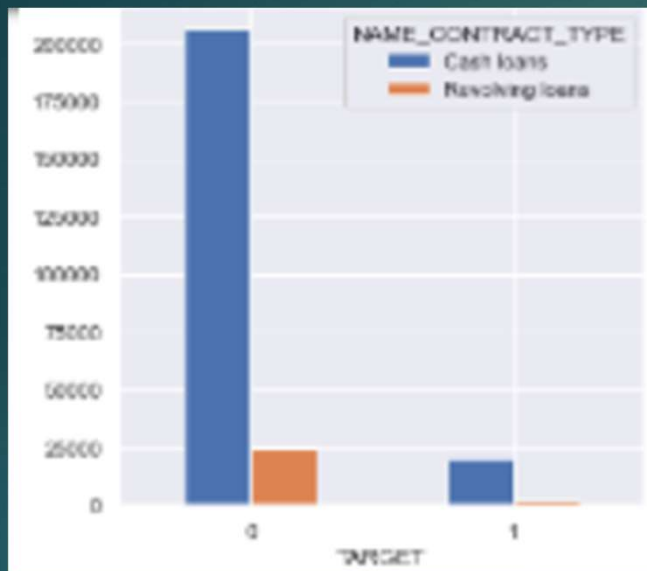
Perform the Bivariate Analysis for categorical columns in Application\_Data



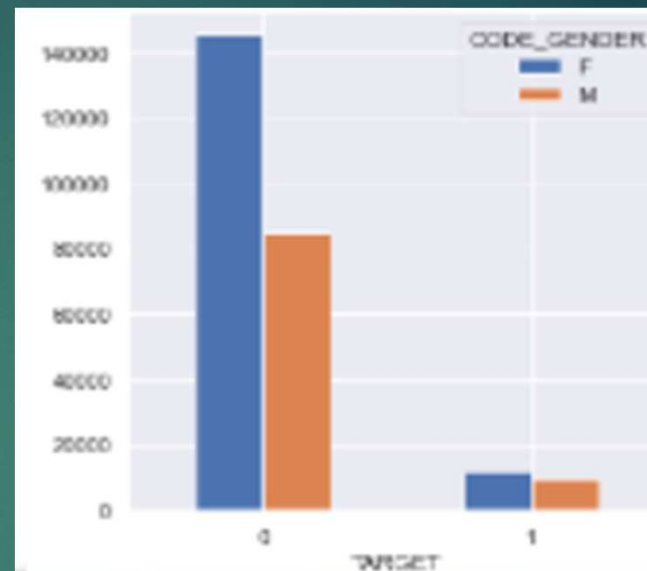
TARGET vs FLAG\_OWN\_REALTY



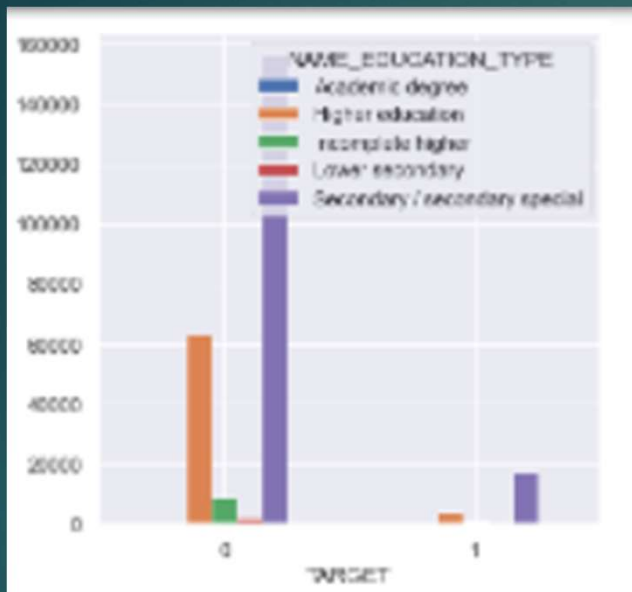
TARGET vs NAME\_FAMILY\_STATUS



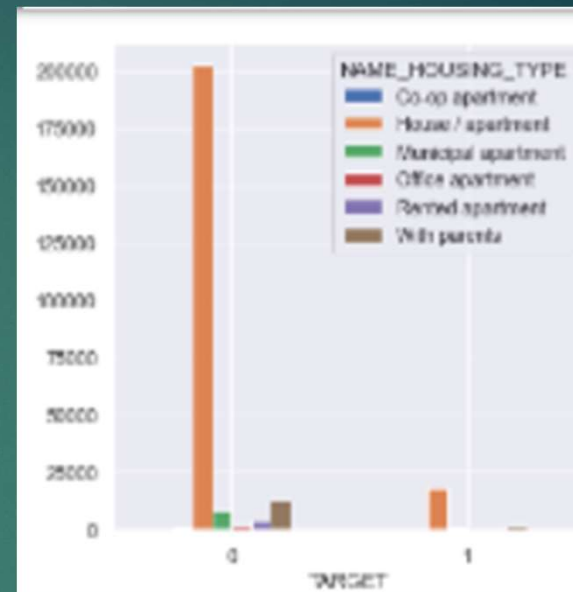
TARGET vs NAME\_CONTRACT\_TYPE



TARGET vs CODE\_GENDER



TARGET vs NAME\_EDUCATION\_TYPE



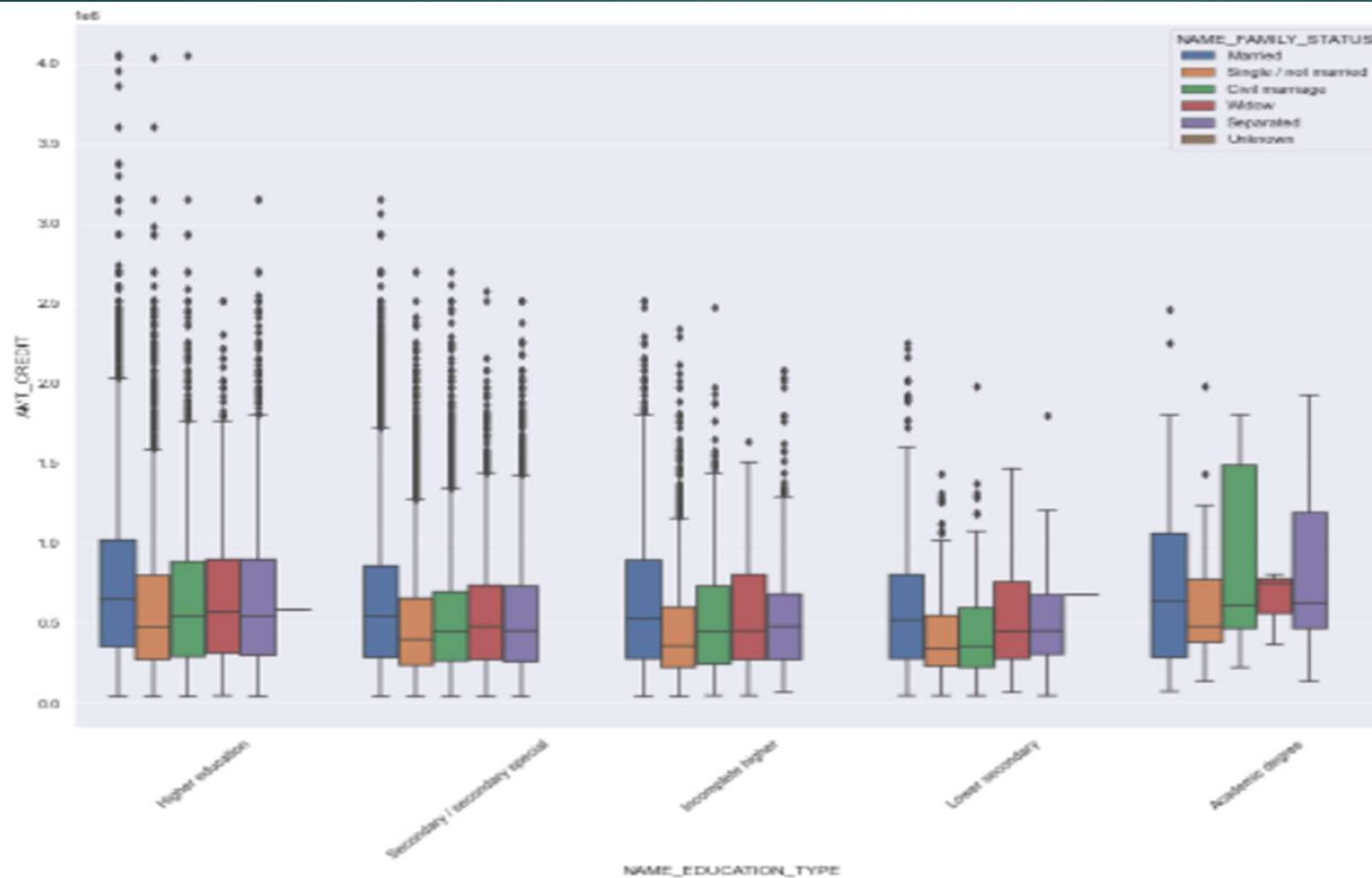
TARGET vs NAME\_HOUSING\_TYPE

## Insights from Bivariate Analysis of Categorical Columns

- People whose income type is working and commercial associate are more likely to take the loan
- people with no car tend to take more loans
- People who start their application on sunday are less likely to take loan where as people who start their Application on Tuesday are more likely to take loan
- Income with 125000-150000 are more tend to take the loan
- Female tends to take more loans
- People with House/Apartment are tend to take more loans
- People are more likely to take the cash loans
- We can conclude that Secondary/ Secondary Special are taking high number of loans
- Individuals having house/flat are more likely to take the loans
- People who are married are tend to take the loan followed by Single/Not Married as compared to other categories



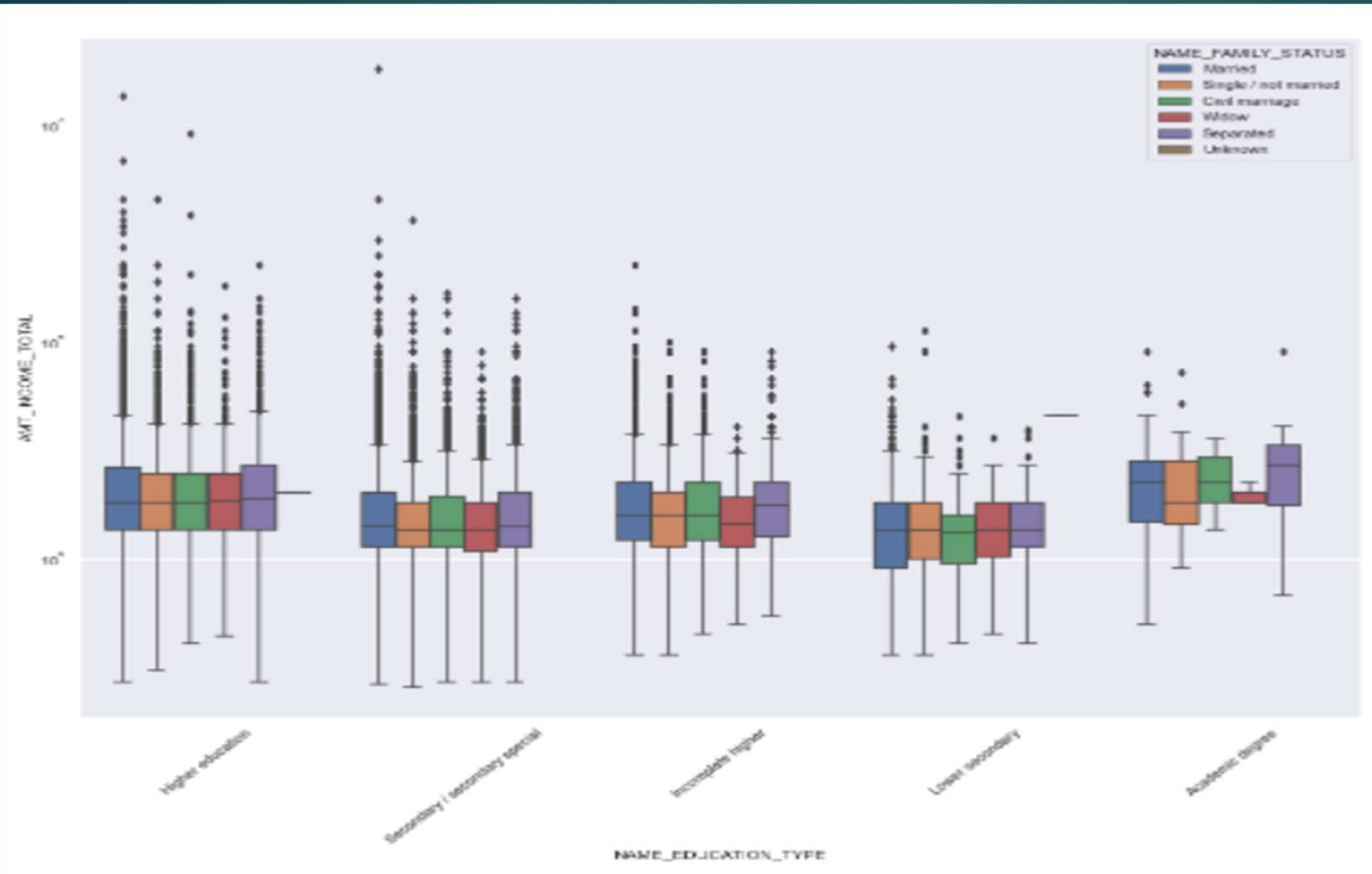
## NAME\_EDUCATION\_TYPE vs AMT\_CREDIT vs NAME\_FAMILY\_STATUS



- From the above box plot we can conclude that Family status of 'Civil Marriage', 'Marriage' and 'Separated' of Academic degree education are having higher number of credits than others. Also, higher education of family status of 'Marriage', 'Single' and 'Civil Marriage' are having more outliers. Civil marriage for Academic degree is having most of the credits in the third quartile.

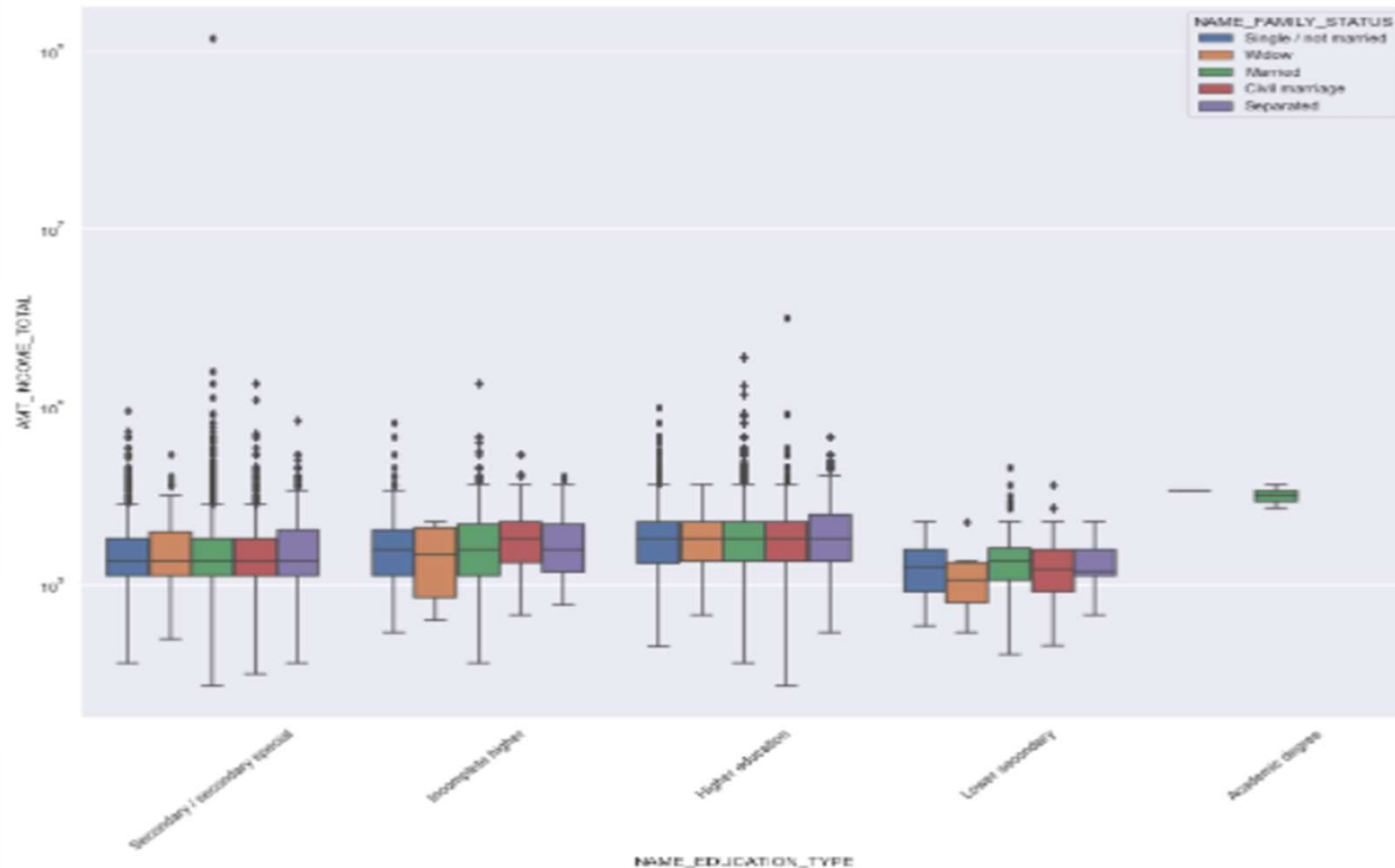


## NAME\_EDUCATION\_TYPE vs AMT\_INCOME\_TOTAL vs NAME\_FAMILY\_STATUS



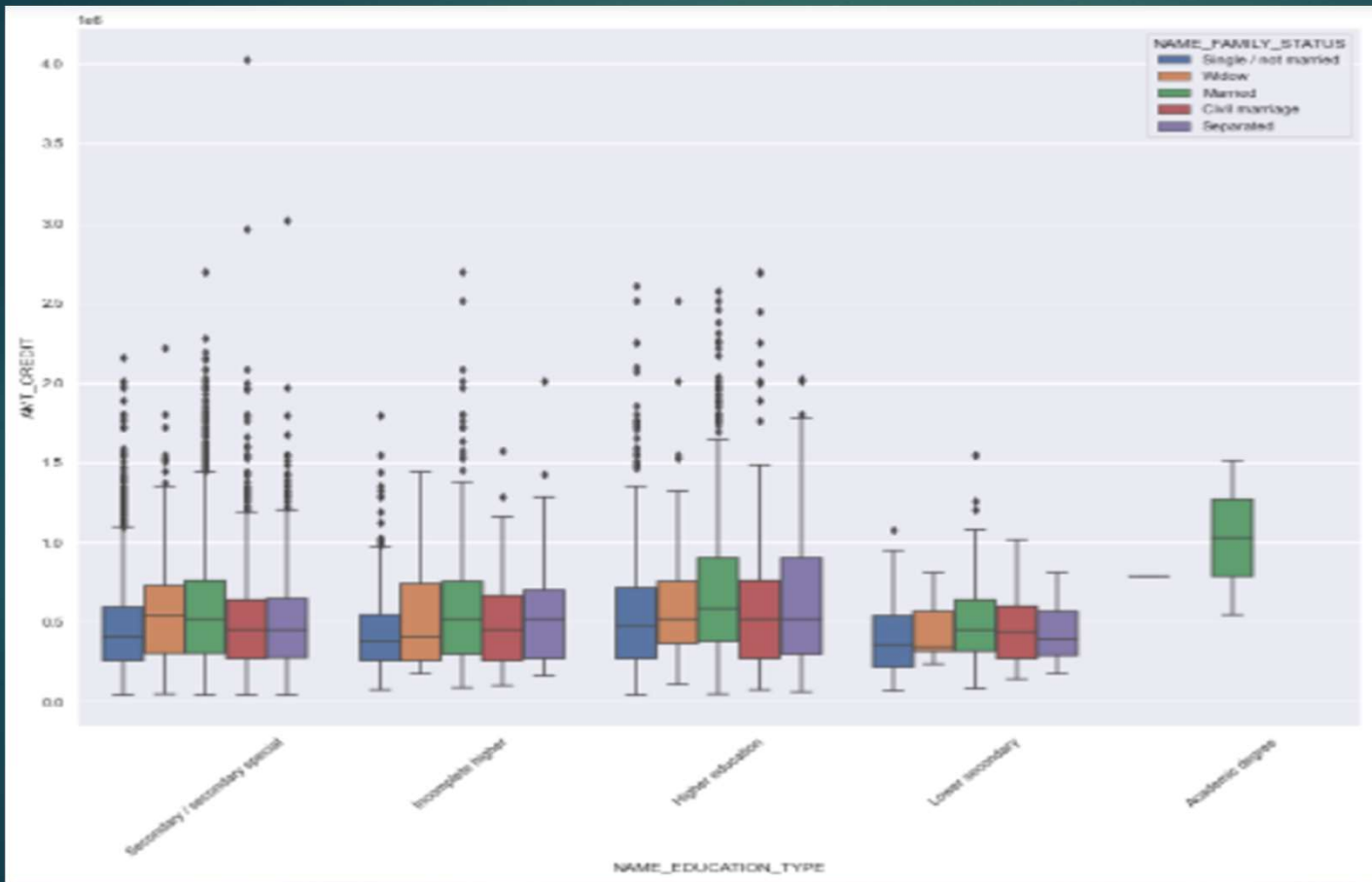
From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status. It does contain many outliers. Less outlier are having for Academic degree but there income amount is little higher than Higher education. Lower secondary of civil marriage family status are having less income amount than others.

## NAME\_EDUCATION\_TYPE vs AMT\_INCOME\_TOTAL vs NAME\_FAMILY\_STATUS



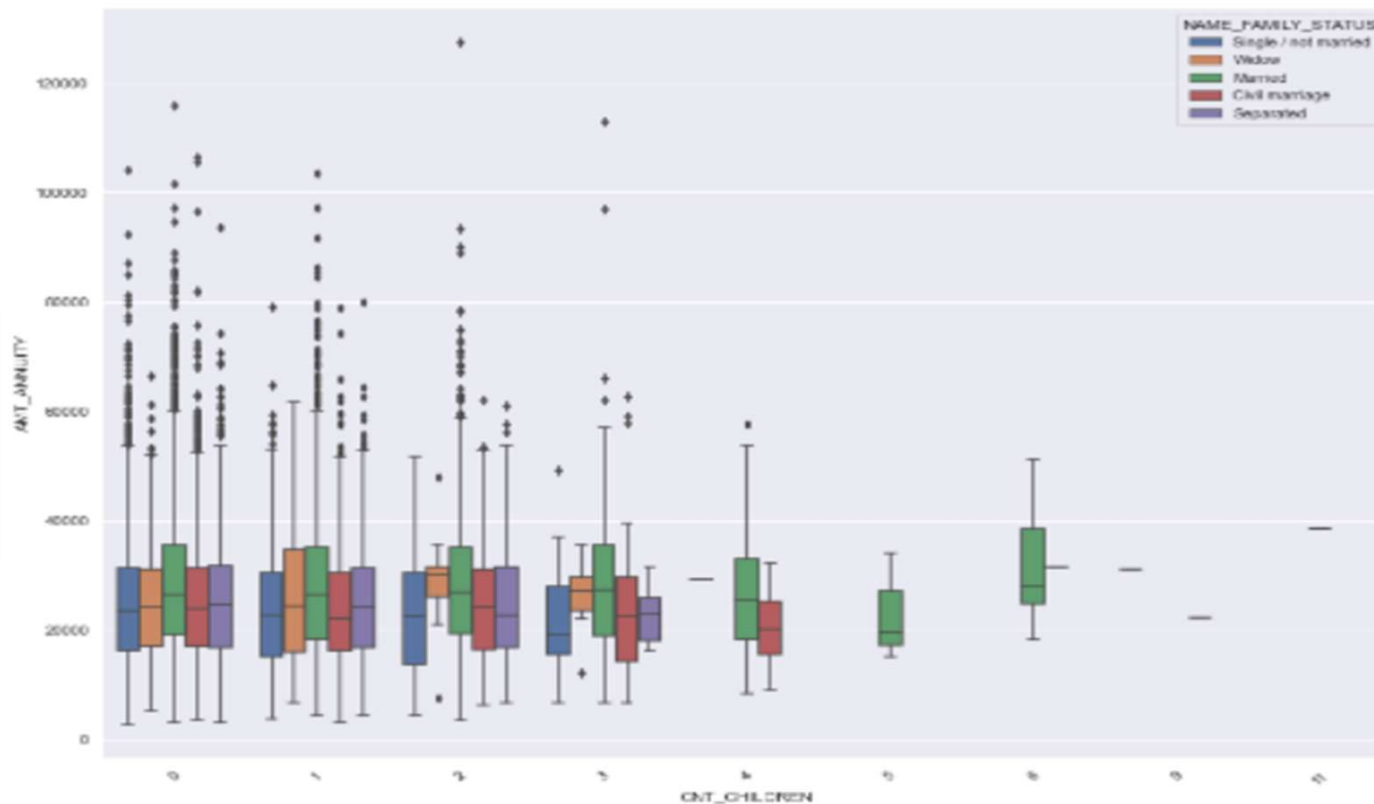
There is some similarity with Target 0, From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status. Less outlier are having for Academic degree but there income amount is little higher than Higher education. Lower secondary are have less income amount than others.

## NAME\_EDUCATION\_TYPE vs AMT\_CREDIT vs NAME\_FAMILY\_STATUS



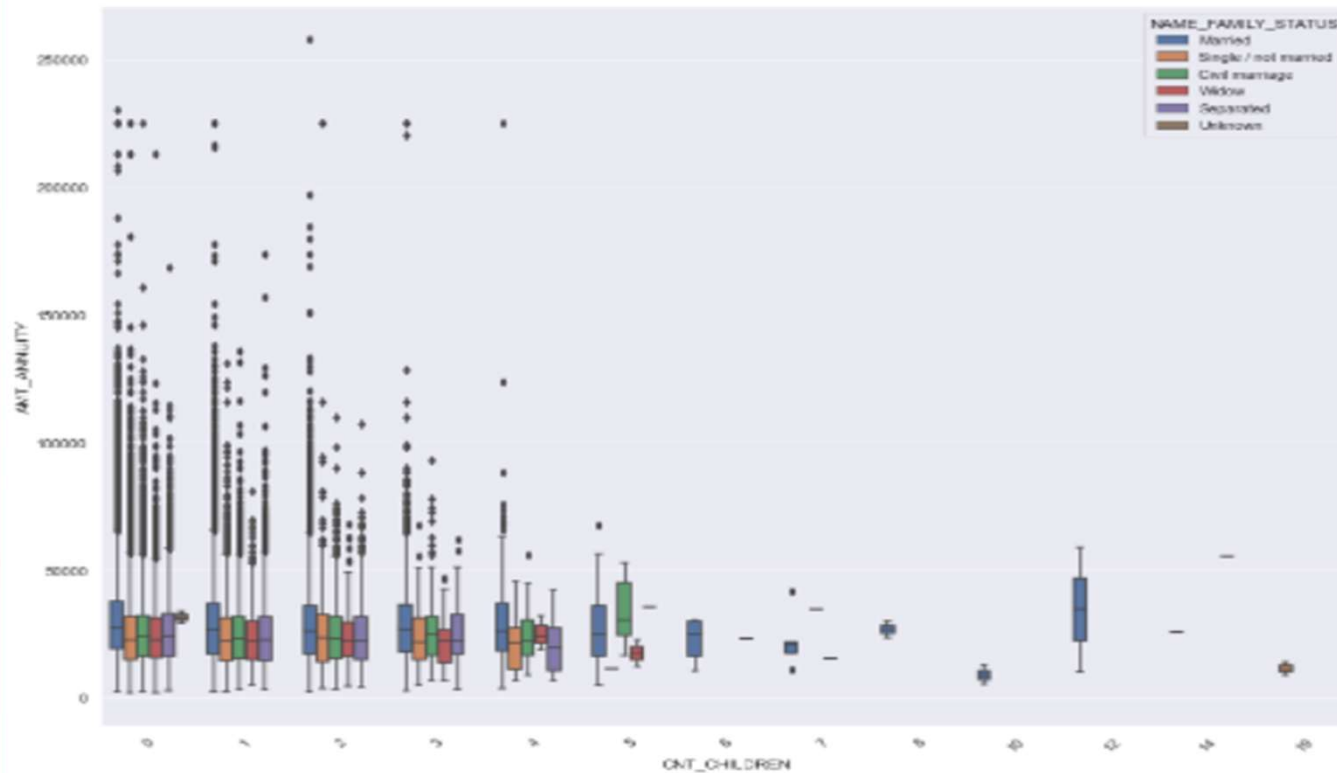
It is similar with Target 0  
From the above box plot we can say that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others. Most of the outliers are from Education type 'Higher education' and 'Secondary'. Civil marriage for Academic degree is having most of the credits in the third quartile.

## CNT\_CHILDREN vs AMT\_ANNUITY vs NAME\_FAMILY\_STATUS



People who are married and having six children are having high annual income and they tend to take loan

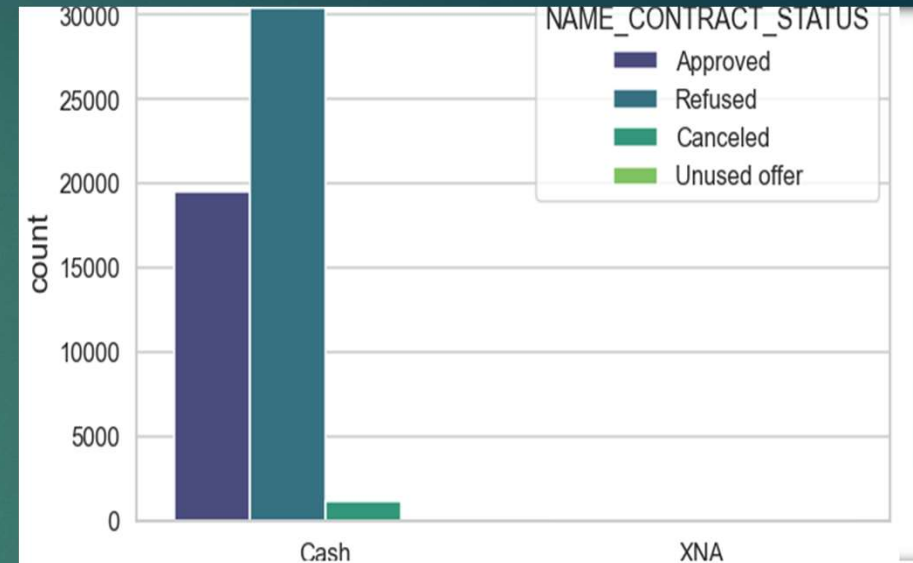
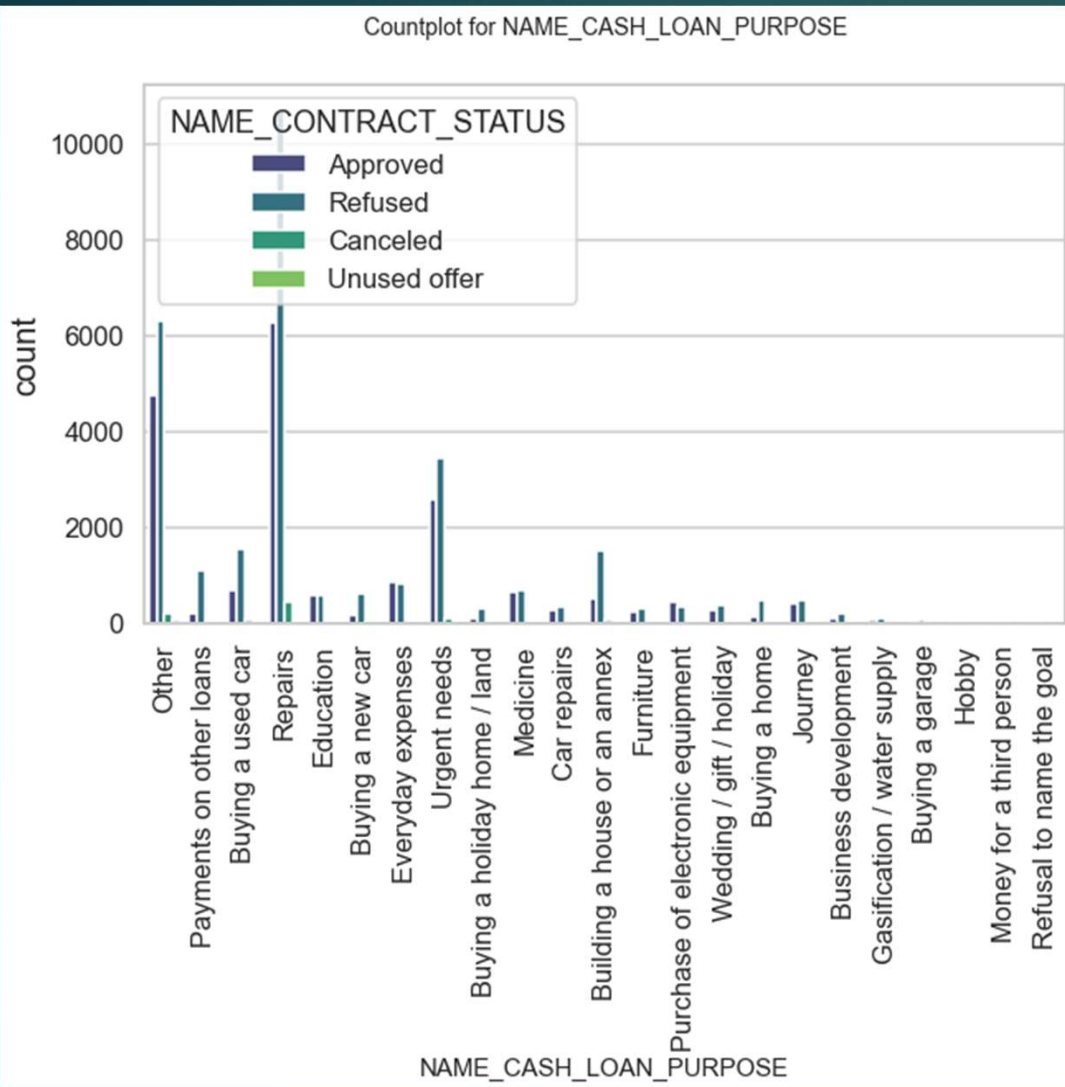
## CNT\_CHILDREN vs AMT\_ANNUITY vs NAME\_FAMILY\_STATUS



From the above graph, people who are married and having 12 children are likely to take the loan

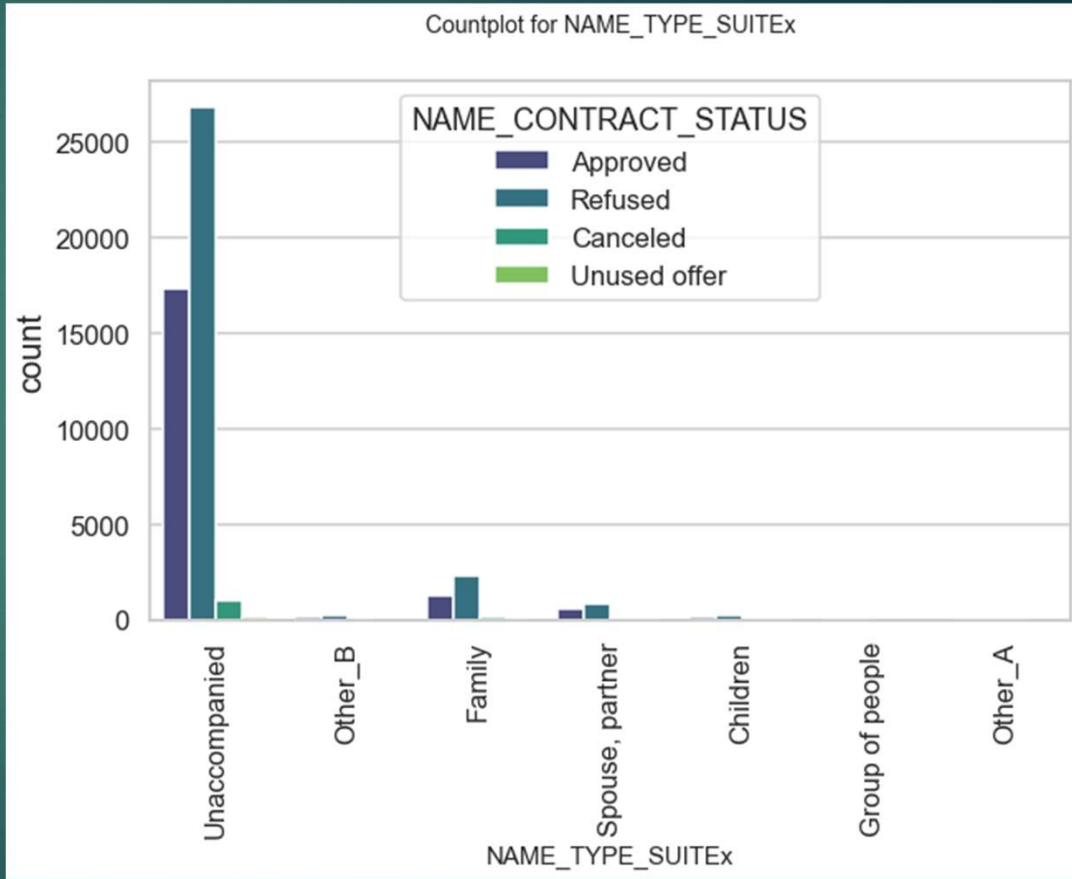
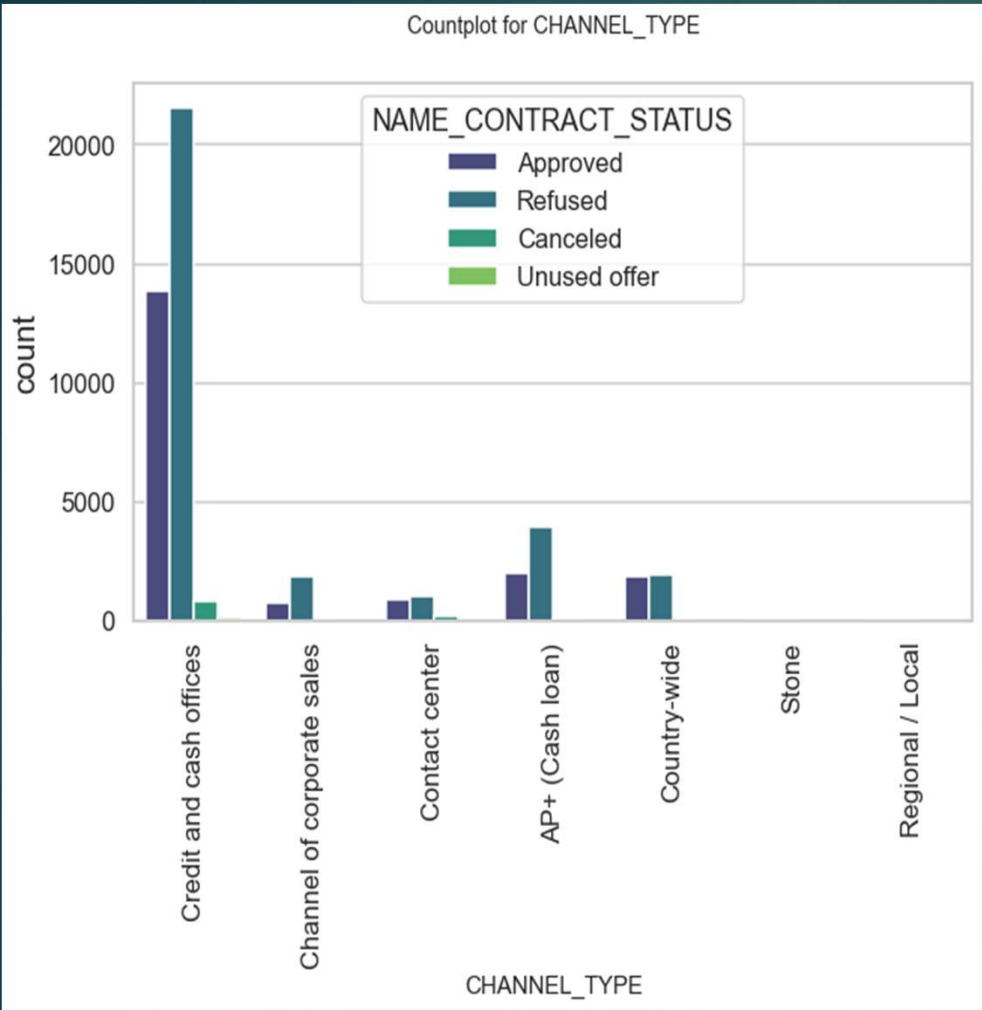
## Insights from Bi variate Analysis on Numerical column in Application\_Data

- There exists more clients who changed their registration details after 4000 days of approval of loan.
- For few not default clients, time taken to publish id's are higher than default clients.
- The application process start hours taken for default and not default cases are similar.
- In non default cases, people keep their phone numbers for greater time.
- People with greater number of days born count are less likely to default.
- In non default case AMT\_GOODS PRICE contains more outliers than default case.
- In default case, most of the clients amount annuity tends to be greater than 25000(median value).
- Whose credit amount is greater than 50000 tends to be less default than compared to default cases and vice versa.
- people with higher no of employment days are less likely to default.
- Majority of defaulting people are having less total income.

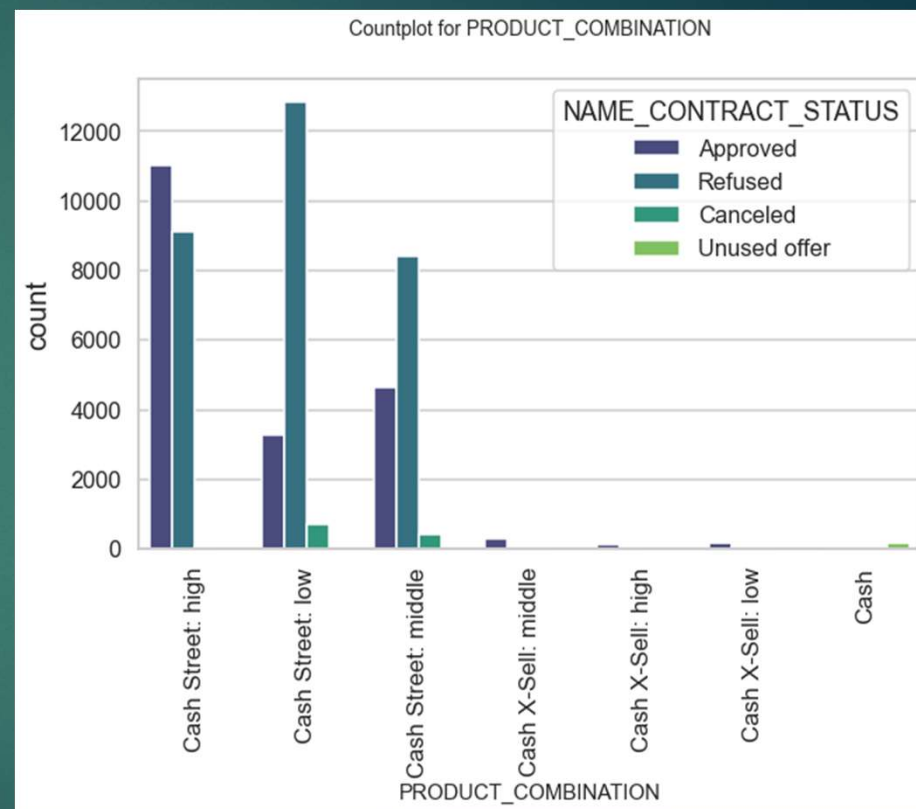
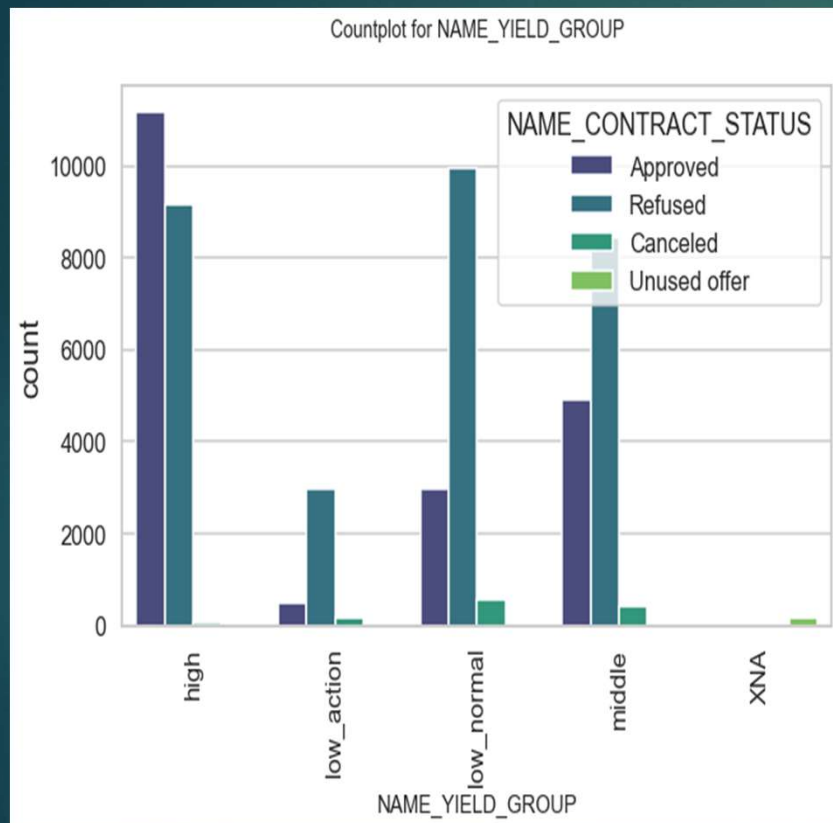


Univariate Analysis of Categorical Columns  
after merging the data

1. Repairs category from NAME\_CASH\_LOAN\_PURPOSE has high chances of rejecting the loan
2. NAME\_PORTFOLIO refuses cash

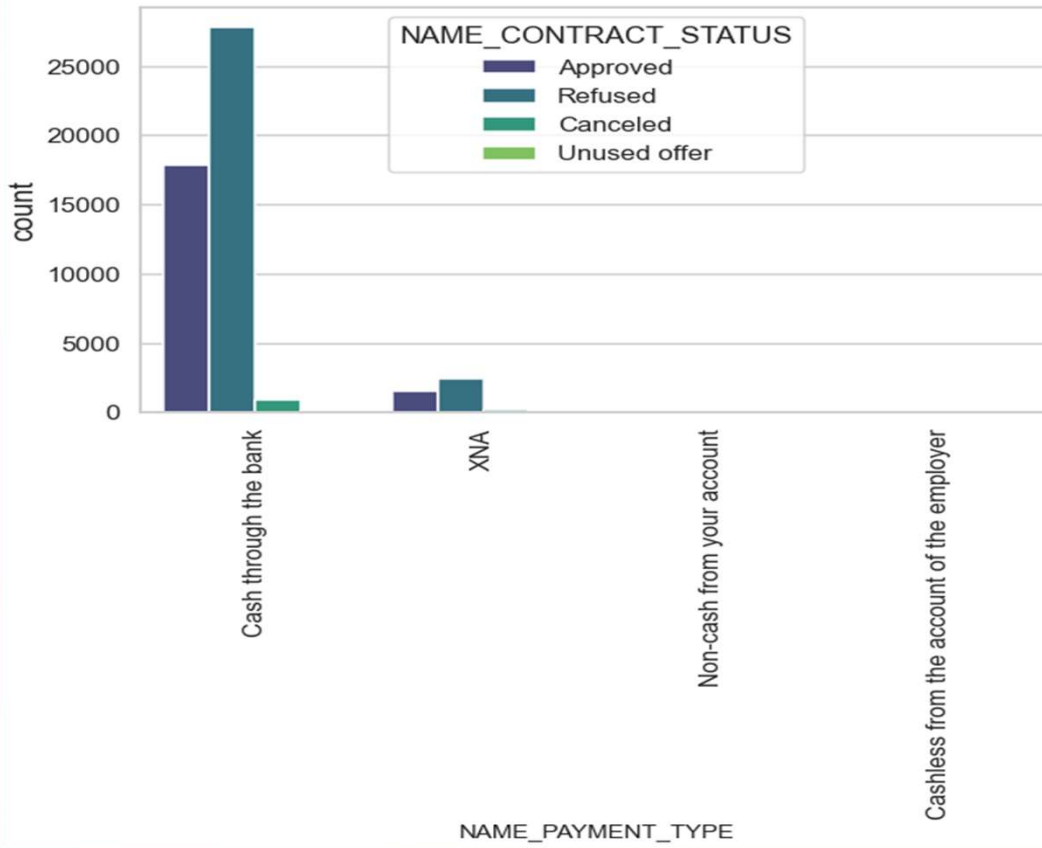




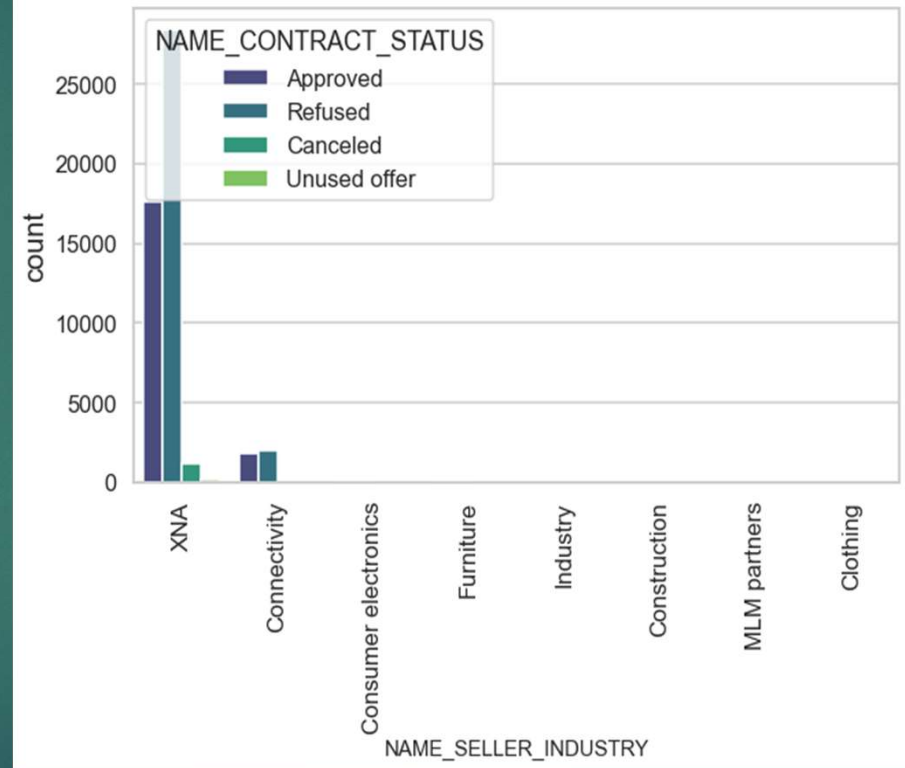


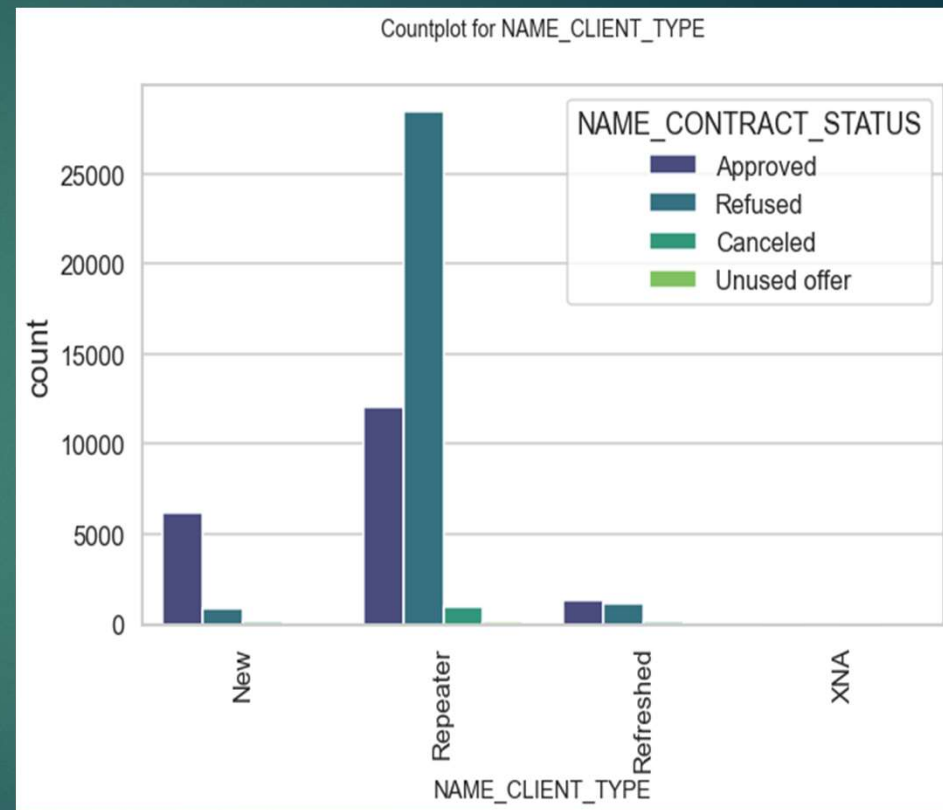
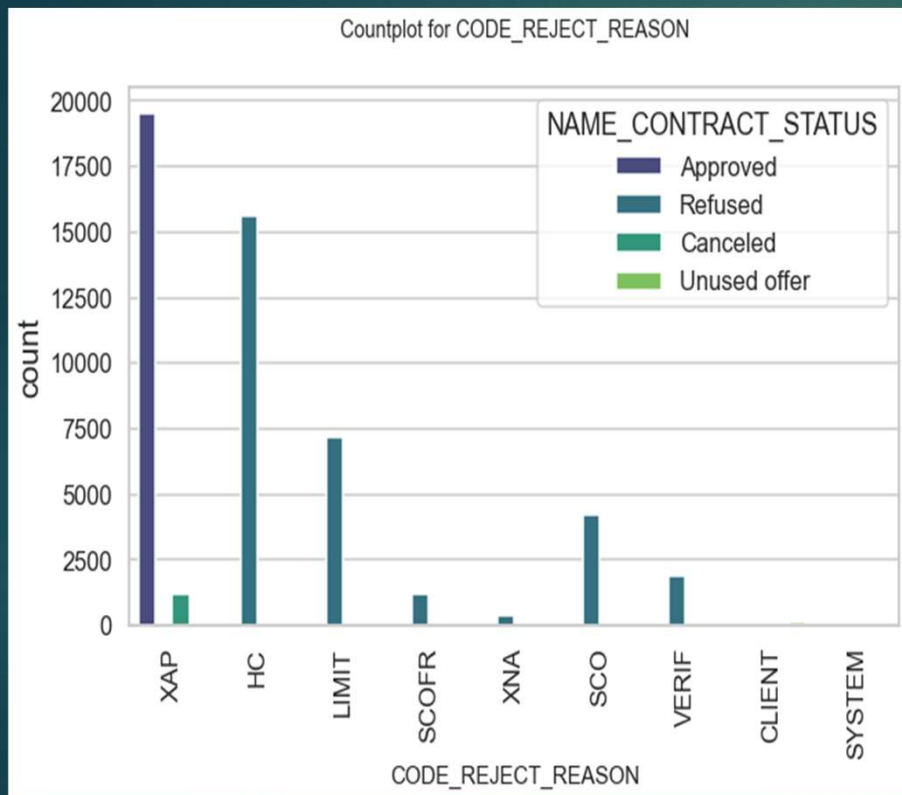


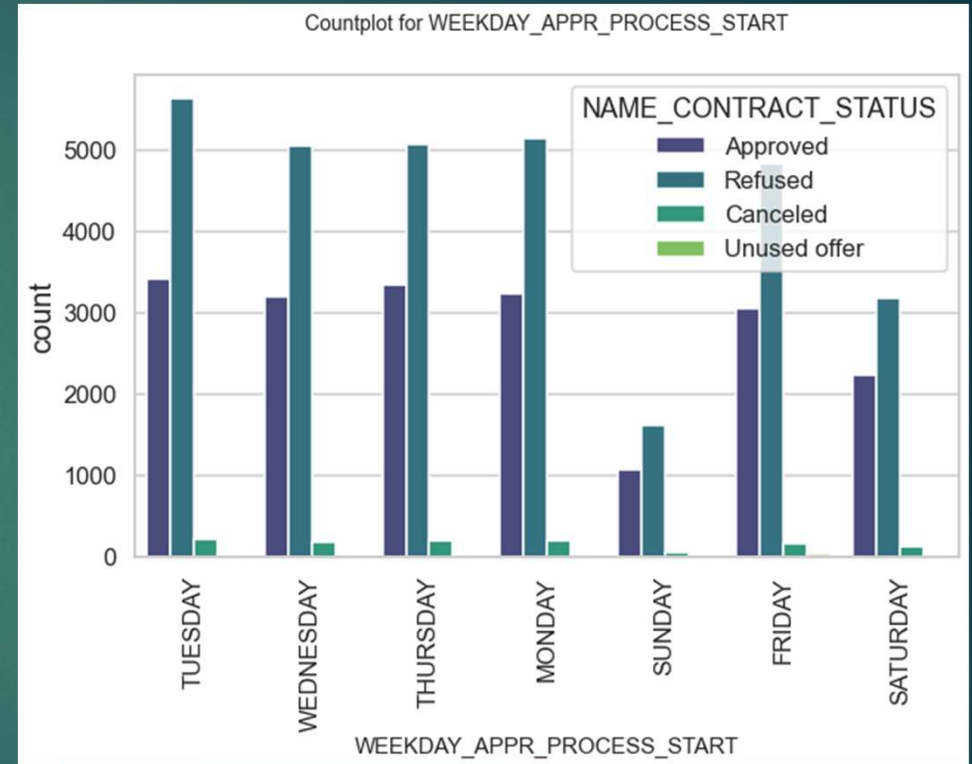
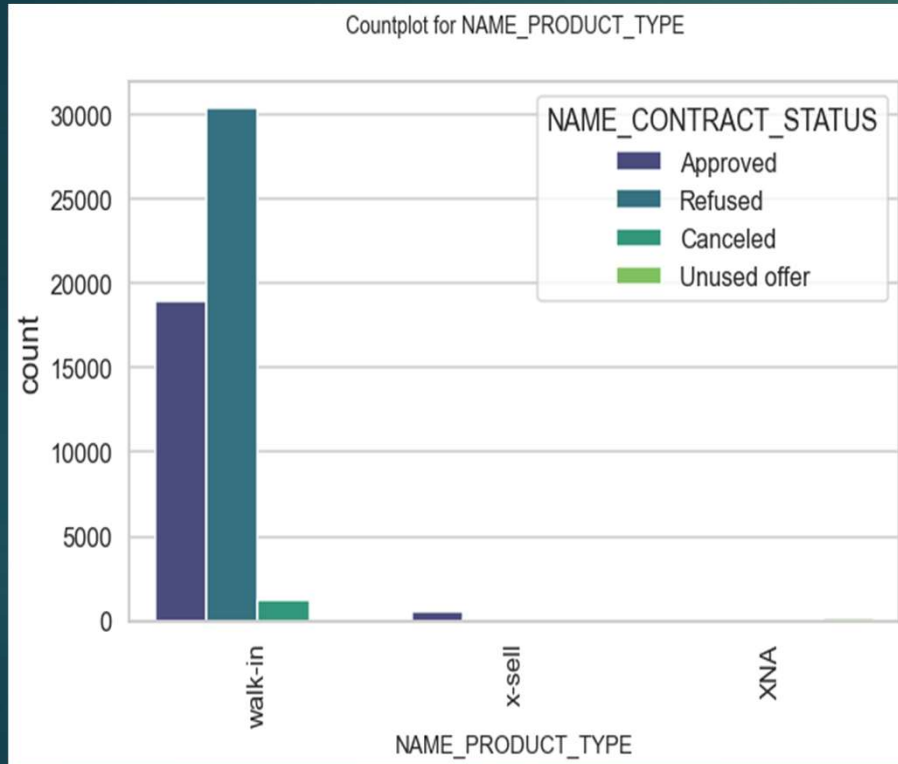
Countplot for NAME\_PAYMENT\_TYPE



Countplot for NAME\_SELLER\_INDUSTRY



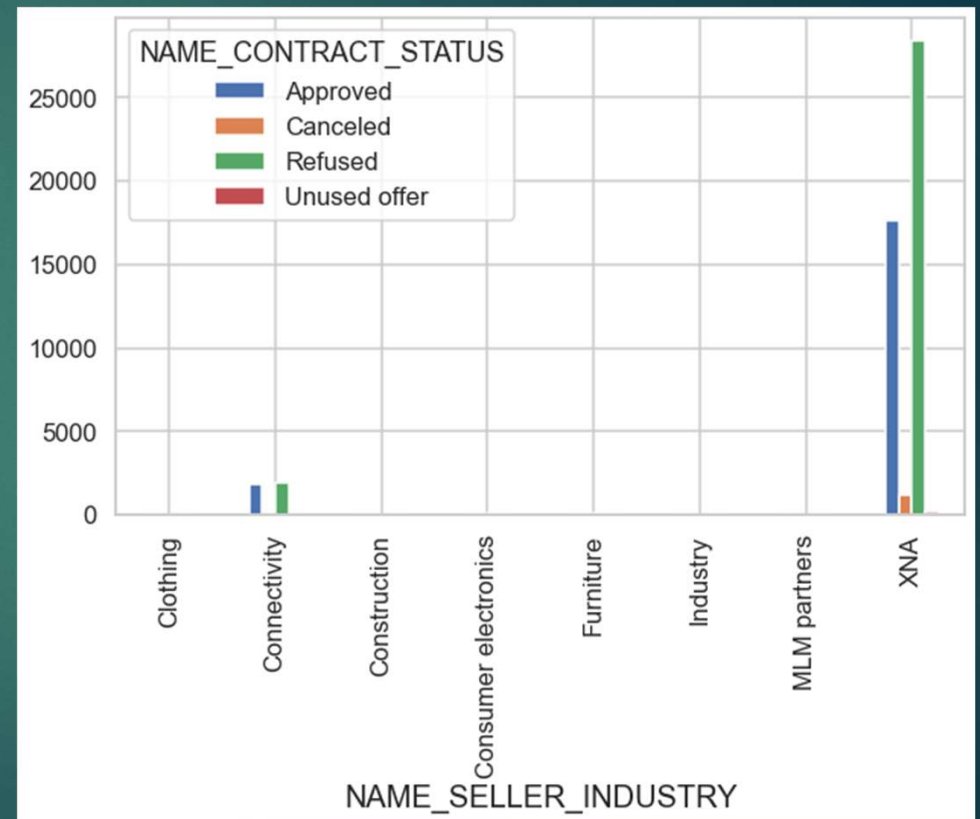
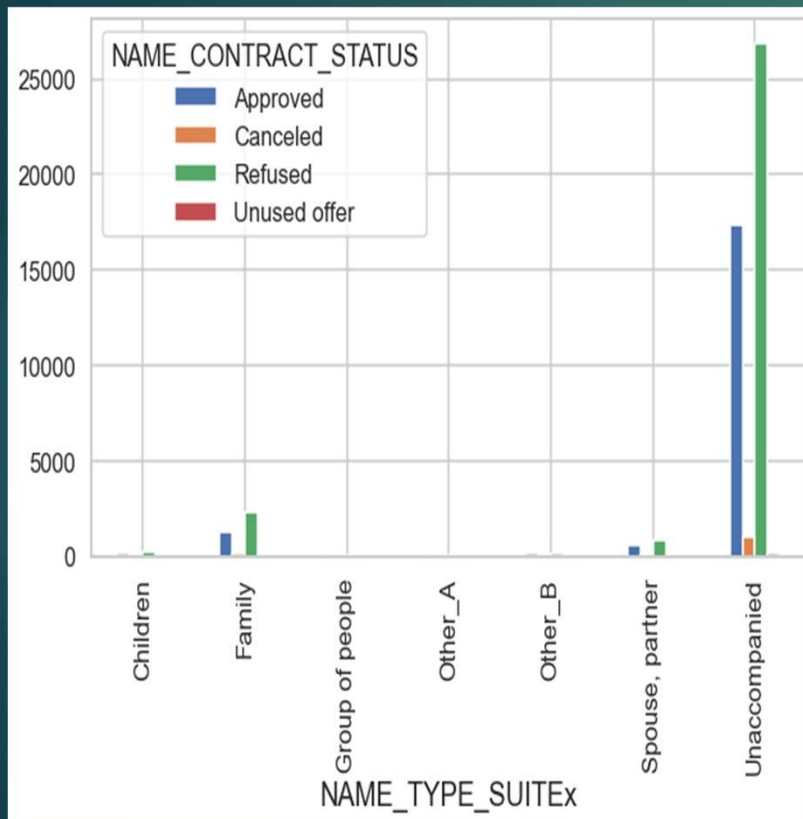


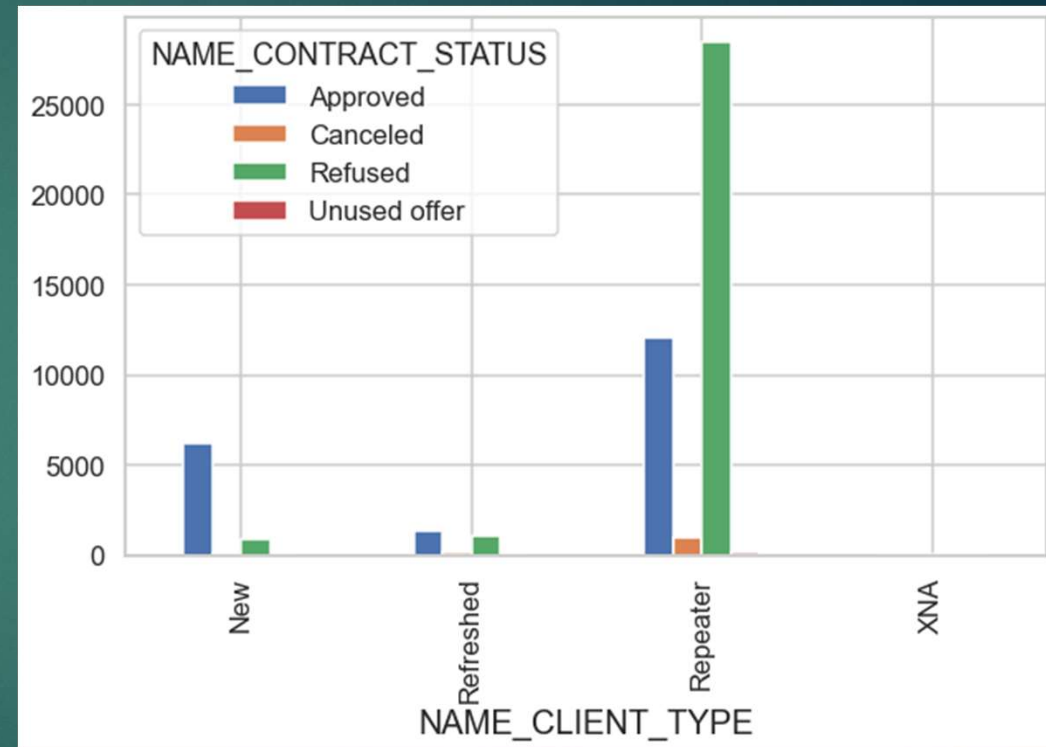
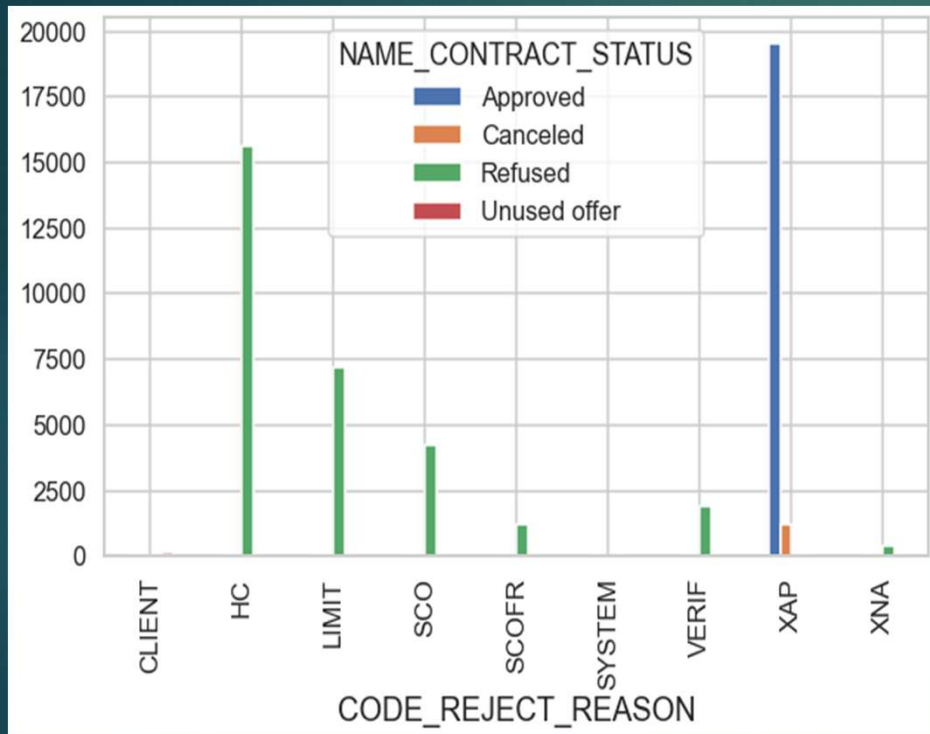


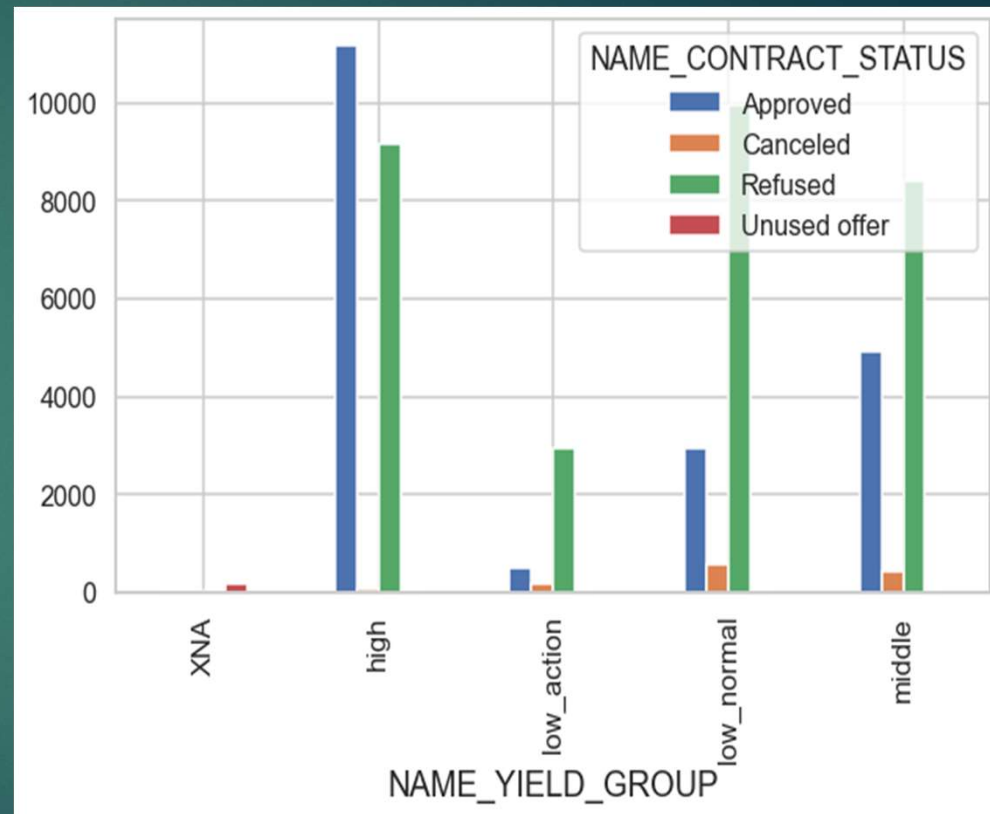
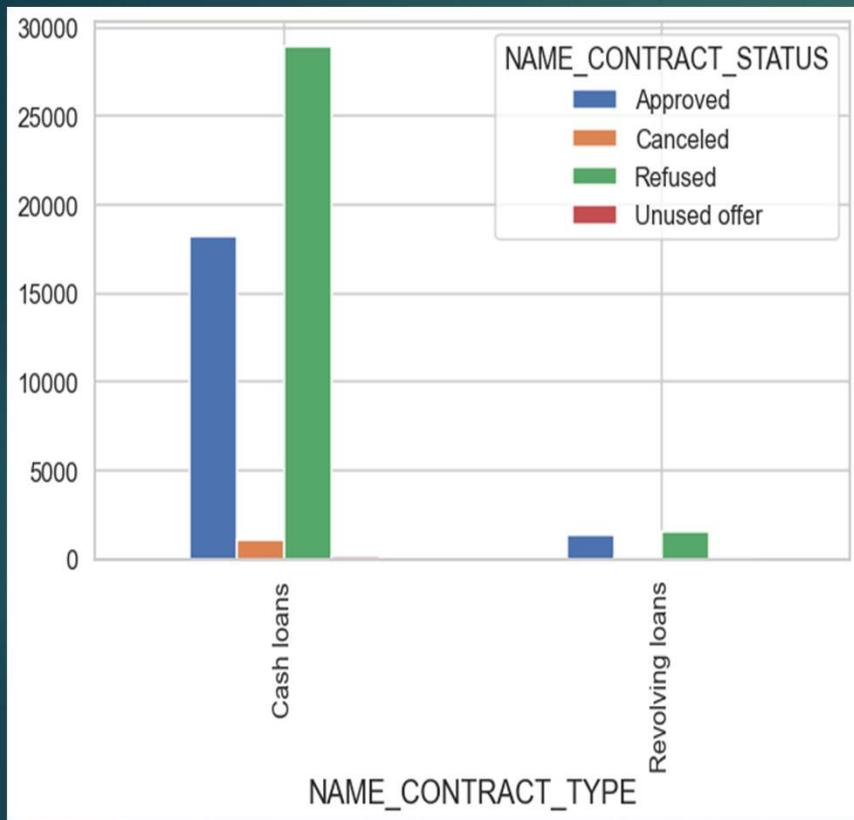
## Insights from Univariate Analysis of Categorical columns after merging the data

- Repeater has highest number of approved loans.
- Middle NAME\_YIELD\_GROUP has highest approval.
- Value of AMT\_CREDIT\_Binning does not affect loan approvals.
- Medium AMT\_INCOME\_TOTAL\_Binning the approval is highest .
- In previous application saturday has the highest approval rate but in current application it is Tuesday.
- Both in NAME\_CONTRACT\_TYPEx and NAME\_CONTRACT\_TYPE unaccompanied has the highest number.
- Currently bank is only giving two types of loans -Cash and Revolving Loans.
- Previously bank was providing Cash, Revolving and Consumer loans.
- Number of consumer loans were highest previously and now highest number is Cash loans

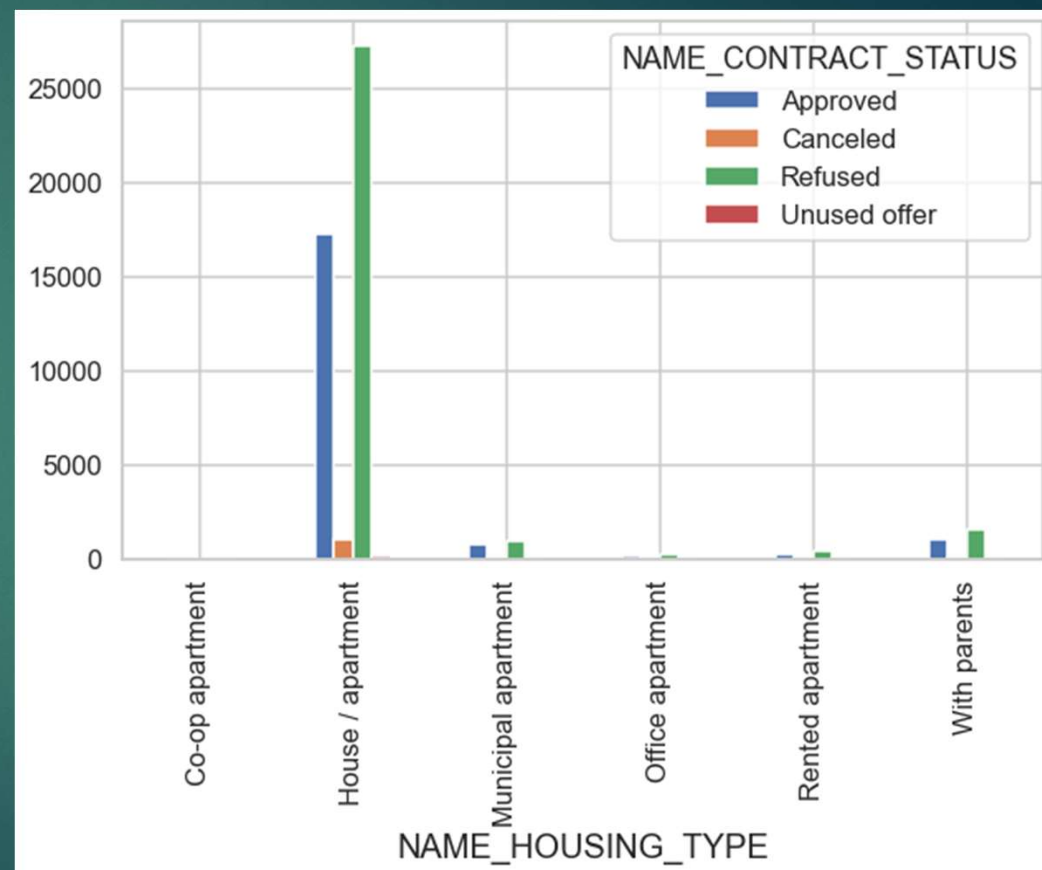
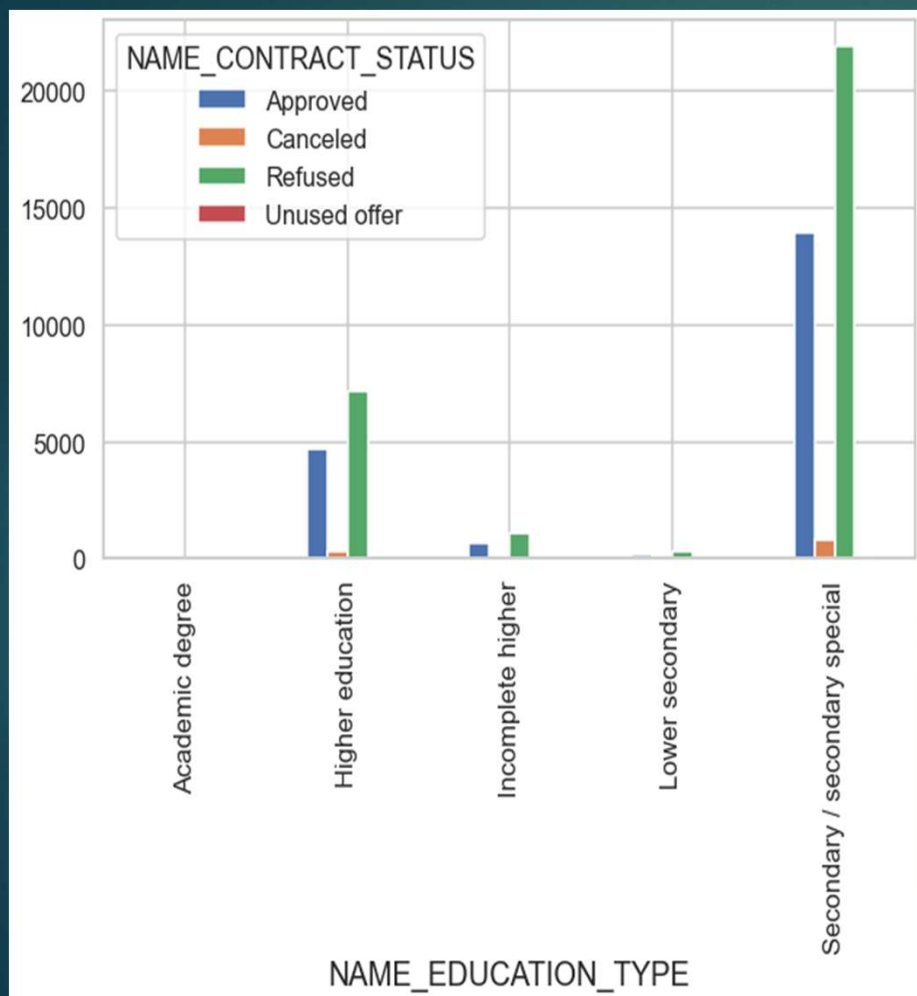
## Bivariate Analysis of Categorical columns after merging the data

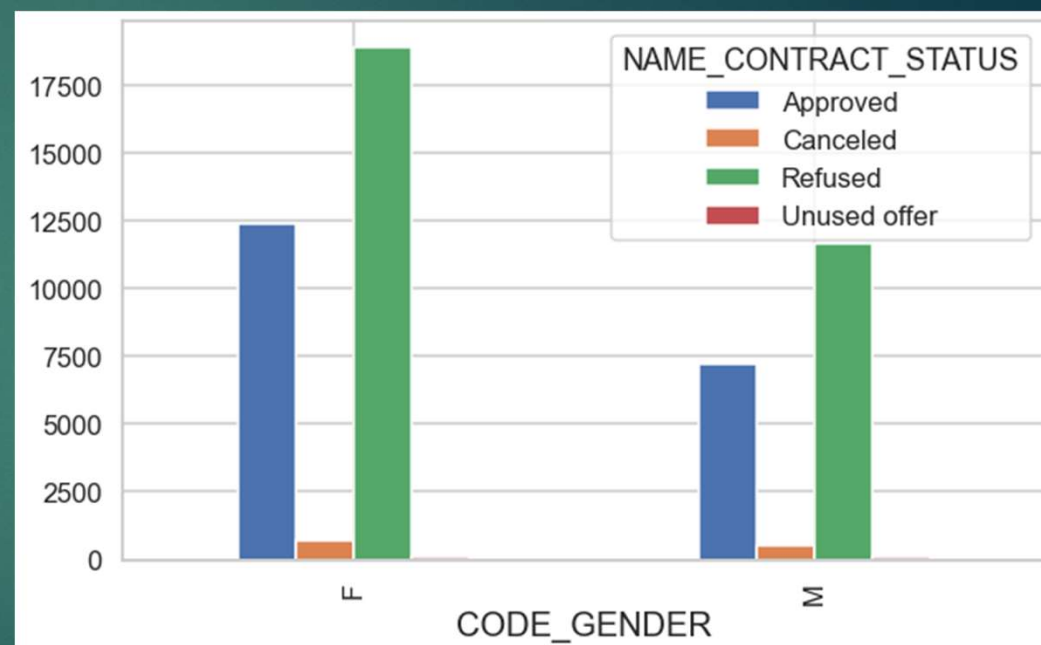
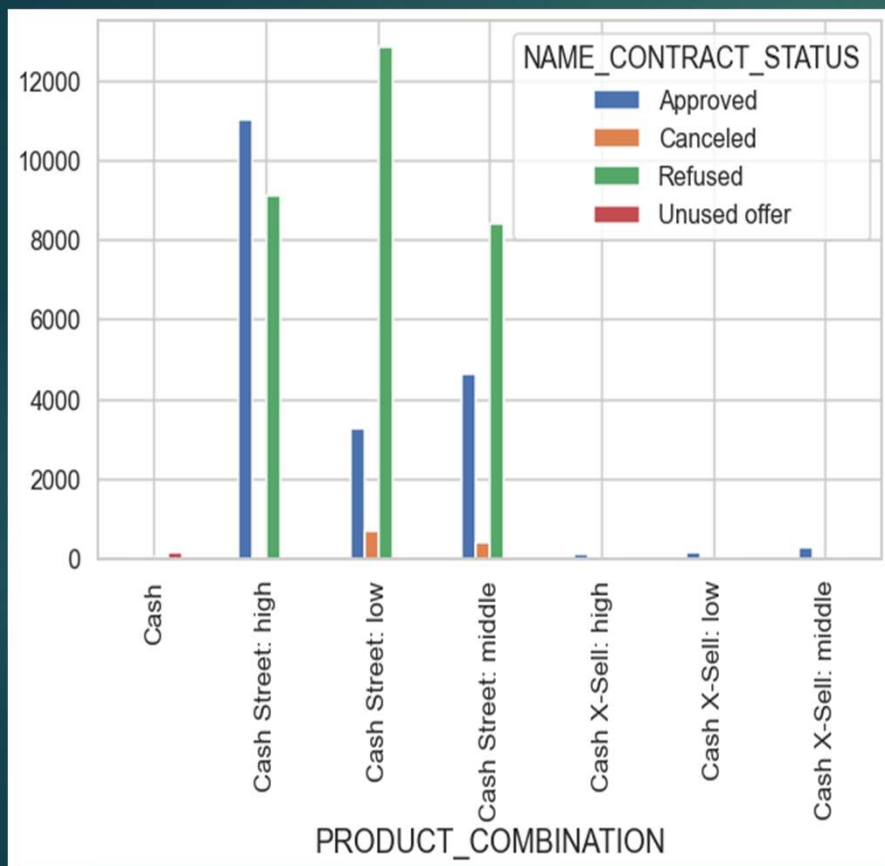


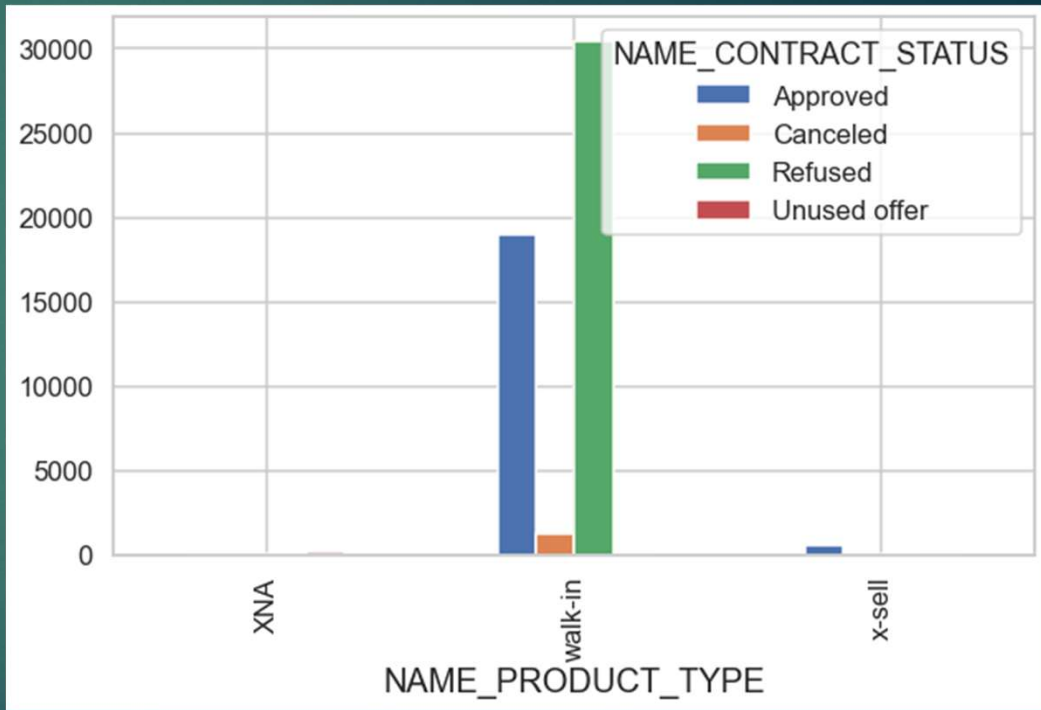
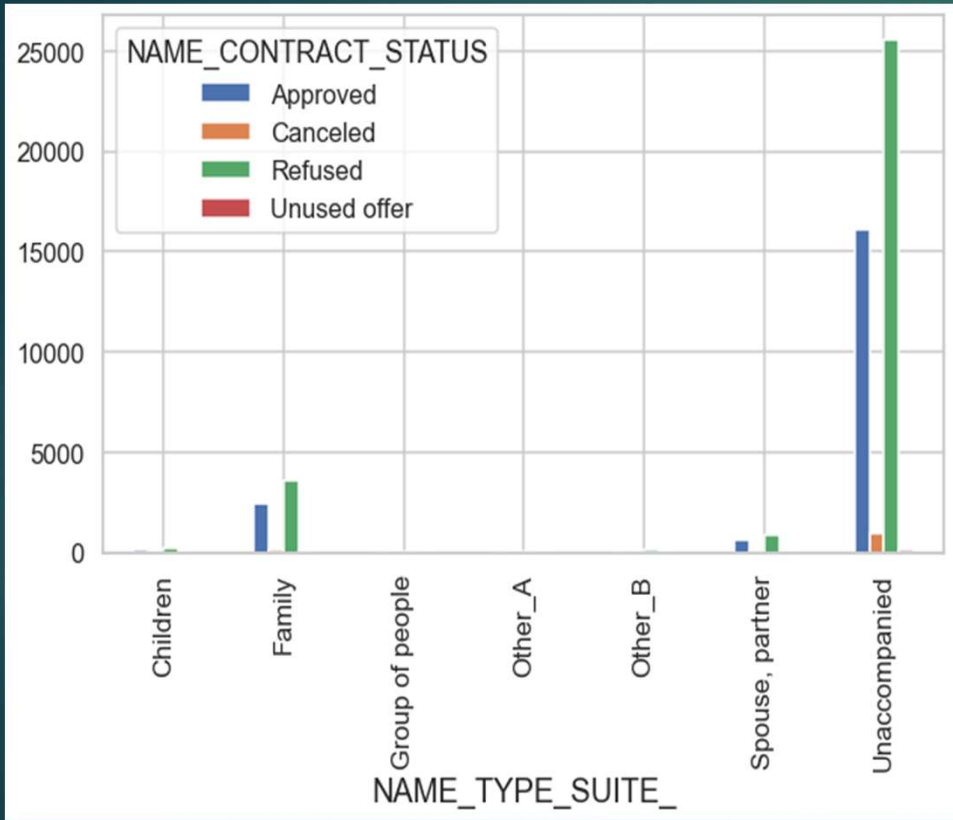








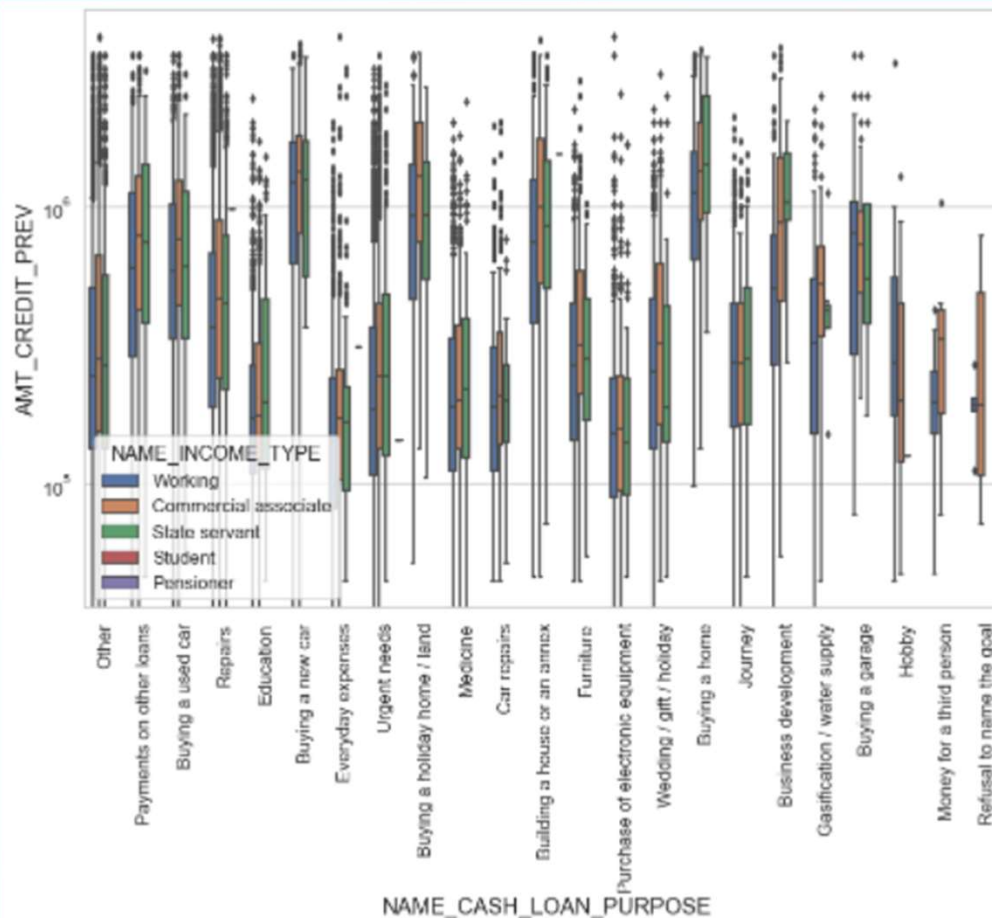




## Insights from Bivariate Analysis on categorical columns after merging the columns

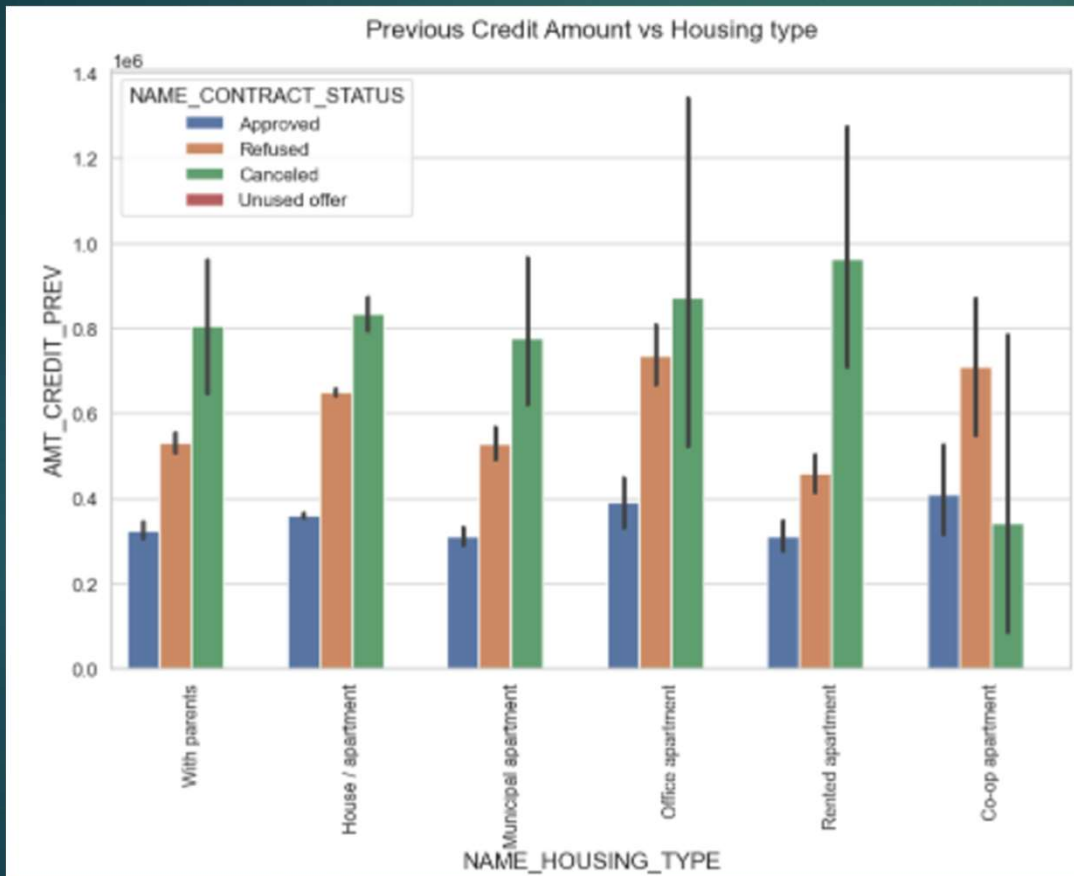
- people who are unaccompanied while applying for loan are having high chances of refuse
- HC has the high chances of refusing the loan
- Repeater Client has high chances of approving as well as refusing compared to other categories
- Cash Loans are having high chances of approving the loans
- High Yield group has the maximum chances of approval
- Secondary/secondary special has high chances of taking the loan
- People who are having house/apartment tends to take loans
- Cash Street: high product has higher chances of loan approval
- Female tends to take more loans
- Walk in product has high chances of taking the loans.
- Loan purposes with 'Repairs' are facing more difficulties in payment on time.
- There are few places where loan payment is significantly higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education' Hence we can focus on these purposes for which the client is having minimal payment difficulties.

## Bivariate analysis of numerical columns after merging the data



- From the above we can conclude some points- The credit amount of Loan purposes like 'Buying a home', 'Buying a land', 'Buying a new car' and 'Building a house' is higher. Income type of state servants have a significant amount of credit applied Money for third person or a Hobby is having less credits applied for.

## NAME\_HOUSING-TYPE vs AMT\_CREDIT\_PREV vs NAME\_CONTRACT\_STATUS



From the above graph, we can conclude that Rented apartment has higher chances of cancellation of loan

## CONCLUSION

1. Banks should focus more on contract type 'Student' , 'pensioner' and 'Businessman' with housing 'type other than 'Co-op apartment' for successful payments.
2. Banks should focus less on income type 'Working' as they are having most number of unsuccessful payments.
3. Also with loan purpose 'Repair' is having higher number of unsuccessful payments on time.
4. Get as much as clients from housing type 'With parents' as they are having least number of unsuccessful payments.