Problem Statement

The aim of this study was to analyze various clients that the bank give loans such that it reduces their risk of losses.

If the customer is tend to default on a given loan then bank should avoid giving loan to such person.

If the customer will be able to pay the loan back, then not giving loans to such person will result in business loss to the bank.

So this study aims to find the risk associated with various customers and reduce the financial losses incurred to bank by only giving loans to suitable clients.

Assumptions

- There is only 1 assumptions taken i.e data which is XNA, XAP are considered as Unknown variables

Overall Approach

- The whole approach taken is as following:-

i.) Understanding the structure of the Data.

ii.) Data Cleaning and Missing Value Handling

iii.) Data Imputing

iv.) Outlier Detection

v.) Data Imbalance Analysis

vi.) Univariate Analysis

vii.) Bivariate Analysis

Overall Approach

- All the steps mentioned in the previous slide have been applied to three datasets:-

i.) Application data

ii.) Previous data

iii.) Merged data

Understanding the structure of the Data.

- Application data has 307511 rows and 122 columns
- There are 16 categorical and 106 numeric type data in application data
- Previous Application data has 1670214 rows and 37 columns
- There are 16 categorical and 21numeric type data in previous data

Data Cleaning and Missing Value Handling

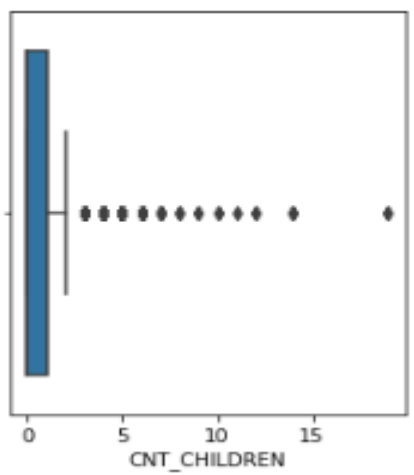
- There are 49 columns with more than 40% null values in application data.
- There are 11 columns in previous data with more than 40% null values
- The null values have been dropped from both datasets
- The Flag columns, Amount columns, and contact flag columns have been dropped from both datasets based on the correlation values .

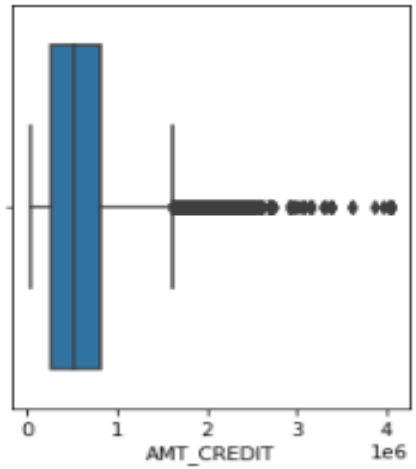
Data Imputing

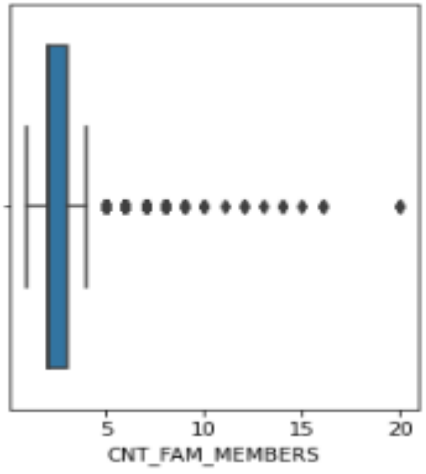
- For the categorical attributes a new category (‘Unknown’) has been created while for numeric attributes the missing data has been imputed with median values for both the datasets

Outlier Detection

- Many attributes seems to have outliers.
- These are detected by the boxplots. A few of them are shown below.
- For Application Data

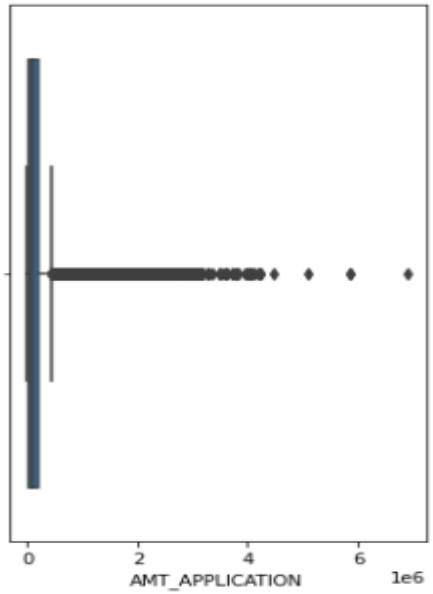


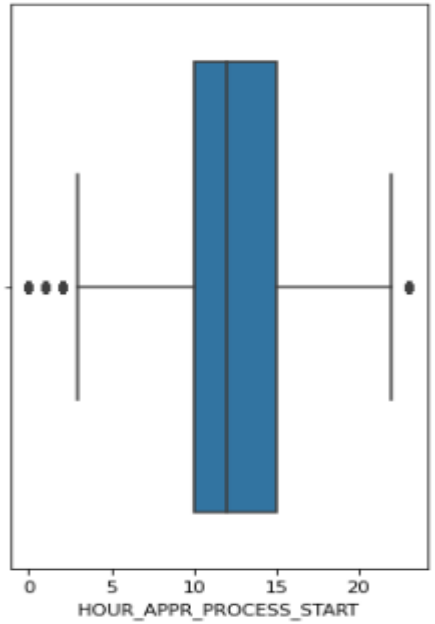


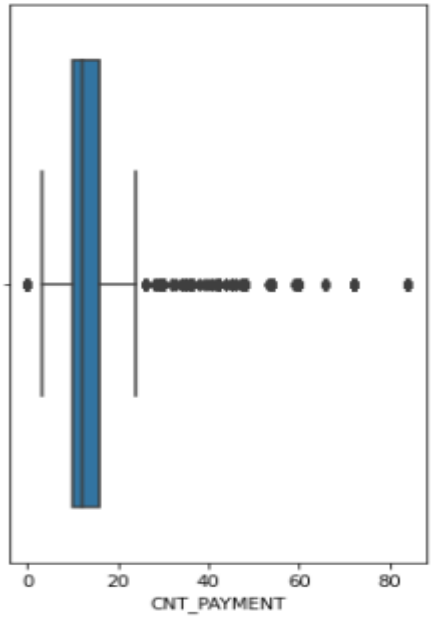


Outlier Detection

- For Previous Data

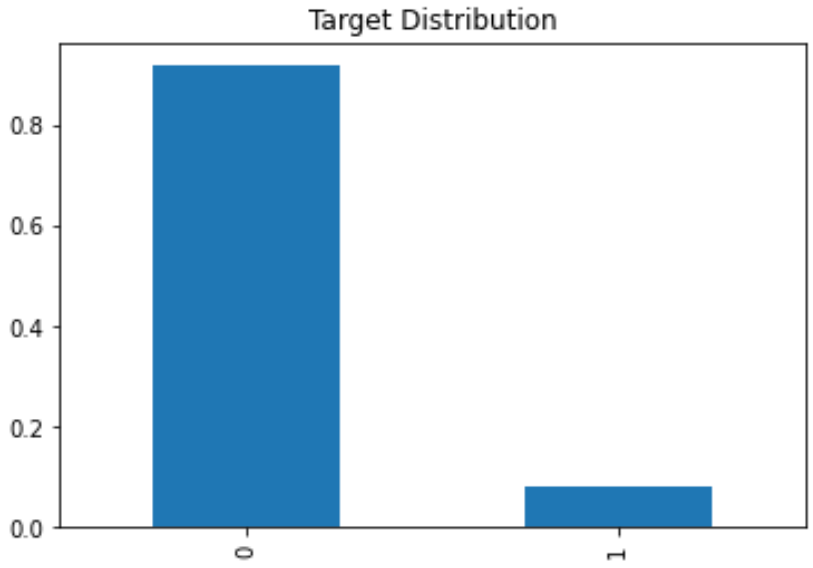






Data Imbalance Analysis

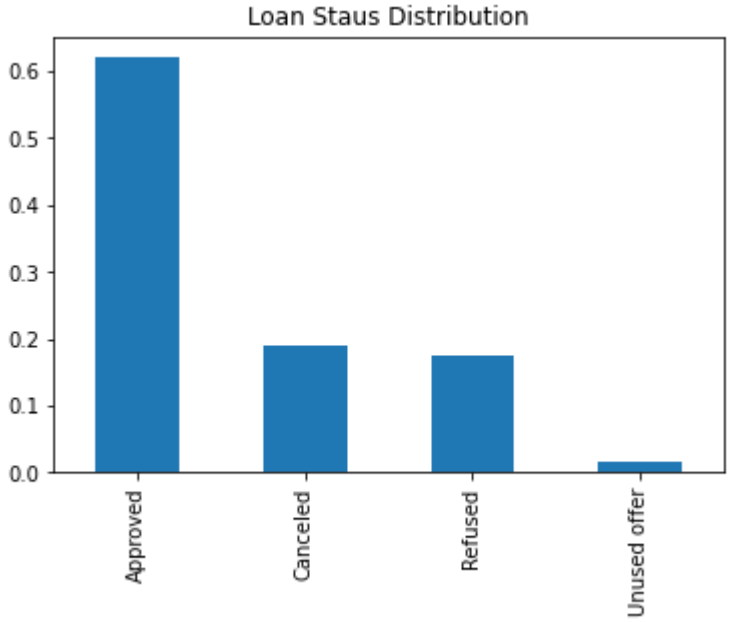
- Application Data



- There is 91.92711805431351 % of Re-payers and 8.072881945686495% of Defaulters

Data Imbalance Analysis

- Previous Data

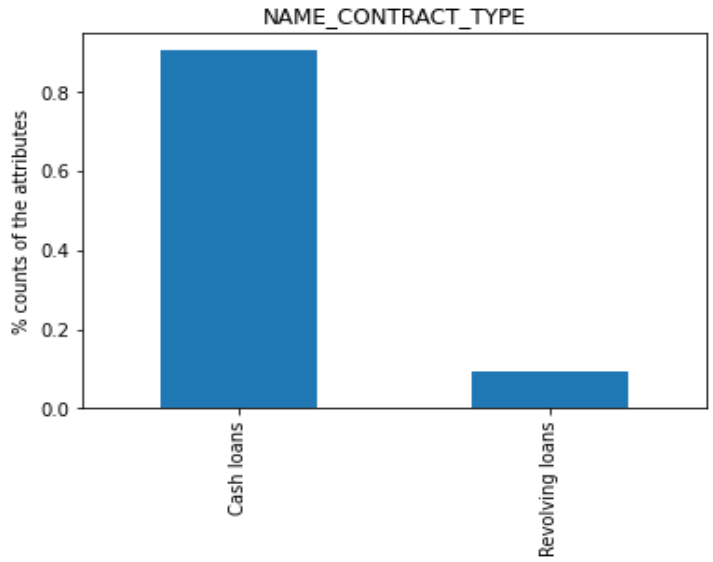


- There is 62.074740123121956 % of Loan approved by the bank

Univariate Analysis for Application Data

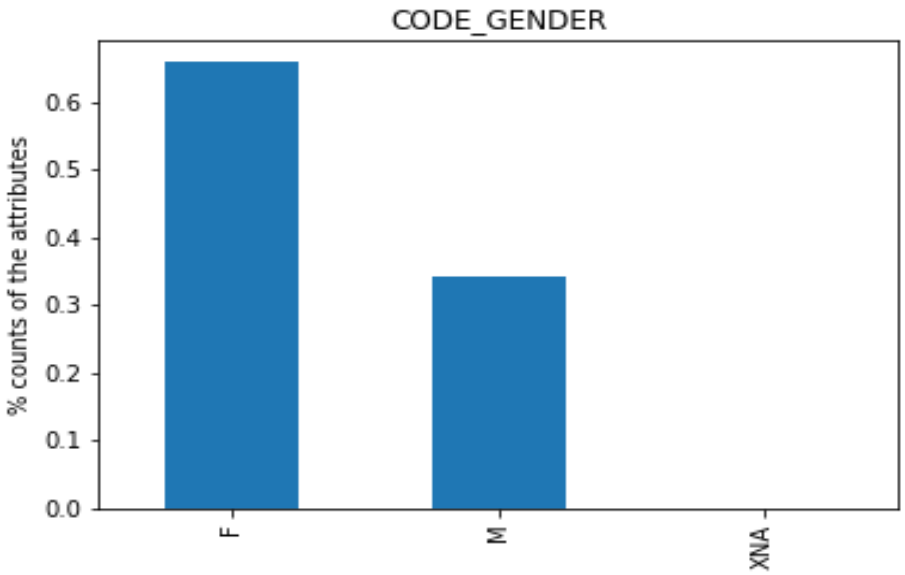
- The analysis of a few attributes are present in the subsequent slides

i.) NAME_CONTRACT_TYPE



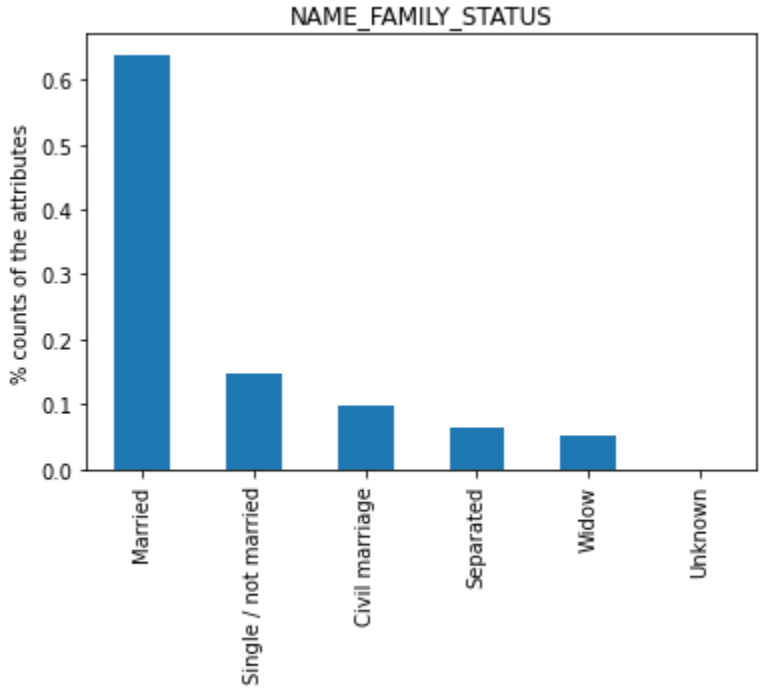
- Cash loans asked is much more as compared to Revolving Loans

ii.) CODE_GENDER



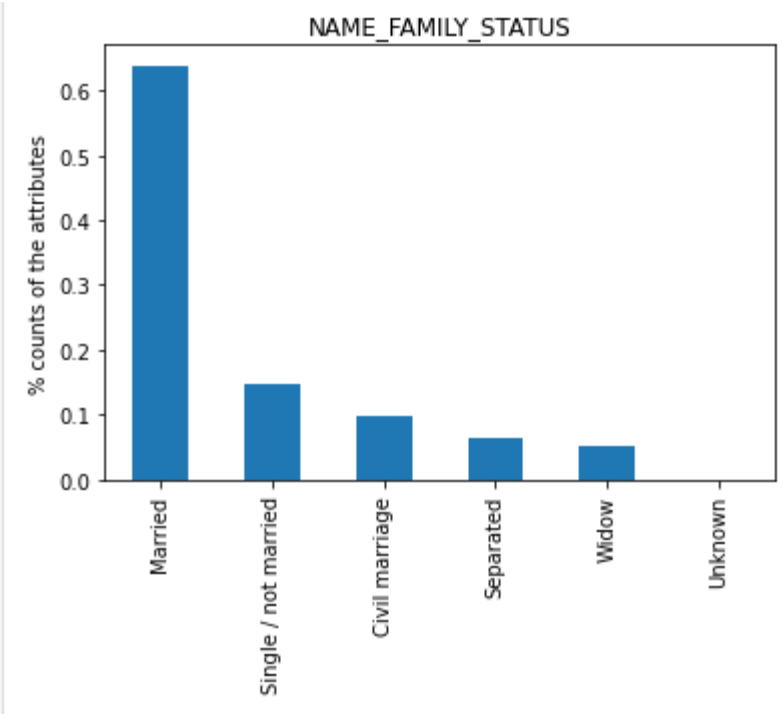
- More number of females have asked for loans as compared to males (More than 60% are females)

iii.) NAME_EDUCATION_TYPE



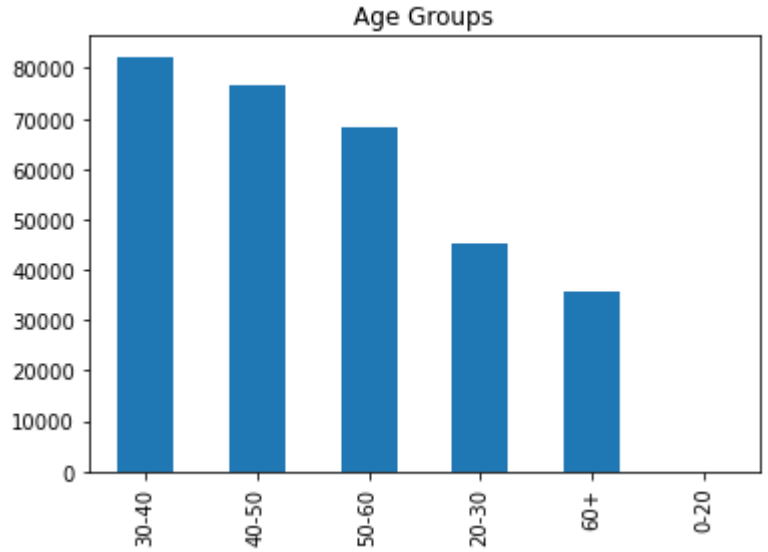
- Almost all the customers are educated at least at Secondary level

iv.) NAME_FAMILY_STATUS



- More than 60% of customers are married

- We have binned a few numerical columns to get some better insights on the data.



- Most of the applicants belongs to age group between 30-40

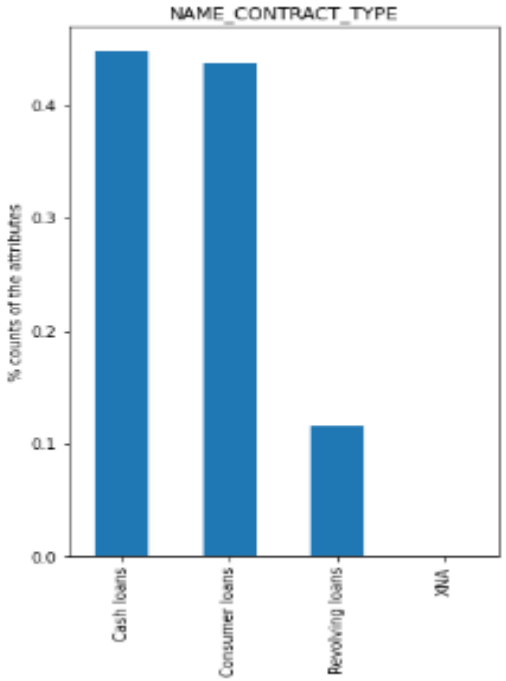
A few insights on numerical columns

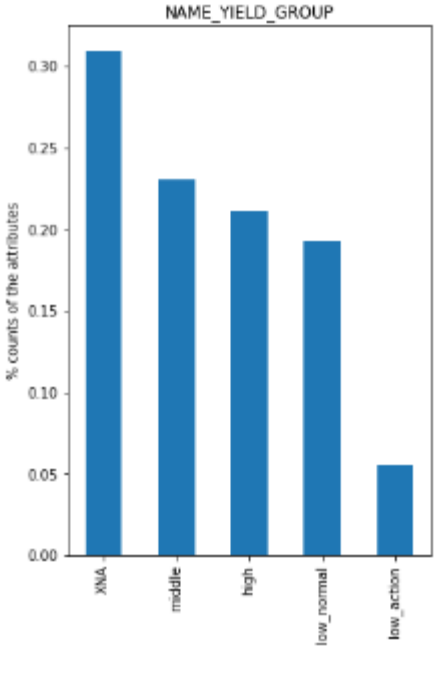
i.) A lot of customers doesn't have any children. It might be because most have them are young.

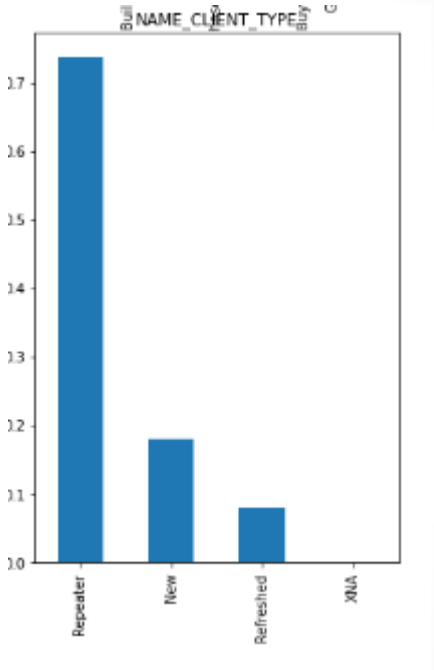
ii.) The average amount of income is Rs. 1,68,797.9 with highest being Rs. 11,70,00,000.

iii.) The average amount of credit asked is Rs. 5,99,026 with highest being Rs. 40,50,000.

Univariate Analysis for Previous Data







Univariate Analysis for Previous Data

- **A few Insights from Previous Data**:

i.) The number of Cash loans and Consumer loans are almost same. Together combined, they add up to more than 80% of the total loans.

ii.) Less number of application are processed on Sunday as compared to any other day of the week.

iii.) Most loan purposes are unknown, but apart from that most of the loan are taken for Repairs.

iv.) Above 60% of the loan applications are approved. Almost 20% applications are refused or cancelled each.

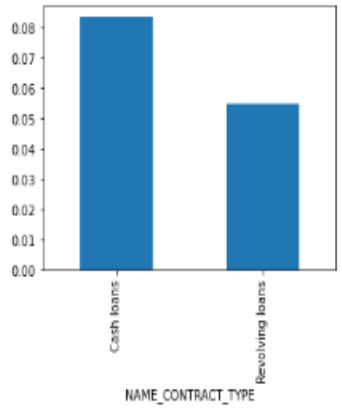
v.) More than 70% of the clients are repeater while less than 20% are New.

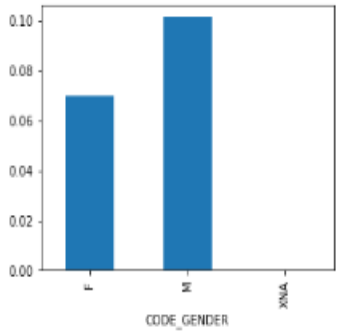
vi.) Most of the loans are taken for mobiles. It's only after the unknown category

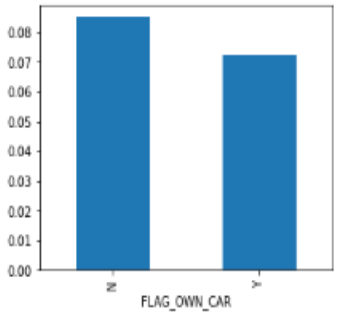
vii.) Most of the previous application is for POS.

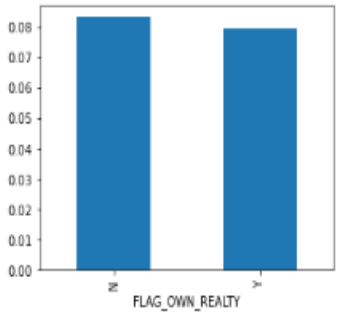
viii.) The number 1 seller industry is Consumer Electronics only after Unknown category

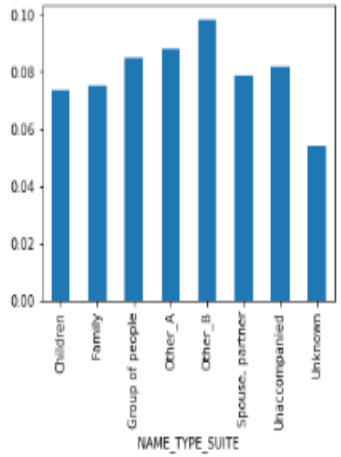
Bivariate Analysis on Application data

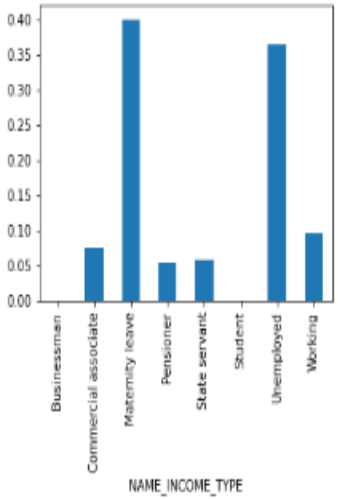


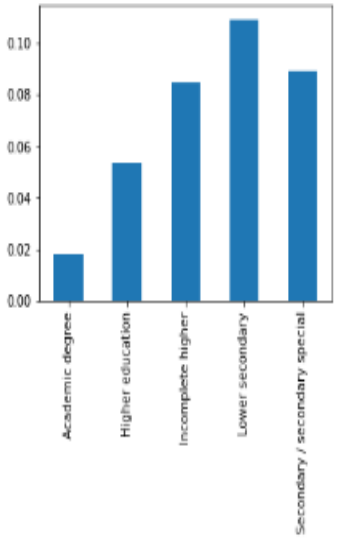


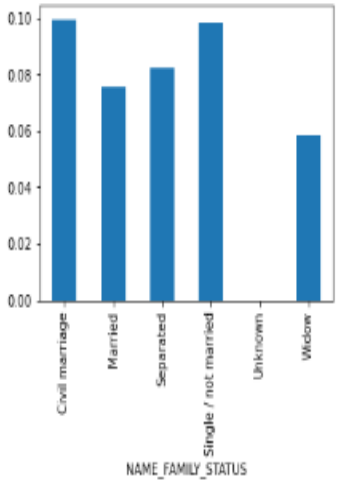












- **A few insights after bivariate analysis on Application data:**

Note: All the data is in %

i.) The number of defaulters in the cash loan is more as compared to Resolving loans

ii.) There are more number of male defaulters (Almost 10%)

iii.) There is a slight no. of more defaulters who don't own a car or realty as compared to who own these.

iv.) The people who are on Maternity Leave and Unemployed are among the highest no. of defaulters. Above 35% of people who come under these categories are defaulters

v.) The people with lower education i.e Lower Secondary are the highest defaulters in the education class.

vi.) There are more than 10% defaulters among people who are civil married and single.

vii.) The people who either live in Rented Apartment or with parents tend to default more with Around 12% default rate.

viii.) Low skilled Labourers have the highest default rate with approximately 17.5%

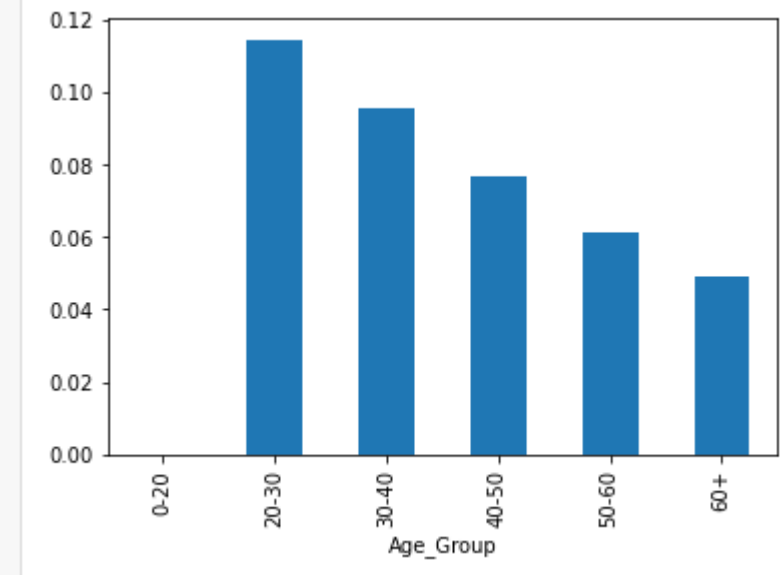
ix.) People in organization Transport: type 3 have highest rate of defaulters.

x.) The people with 9 or more number of people are usually defaulters.

xi.) The loans which are in range 40L-50L are defaulted more

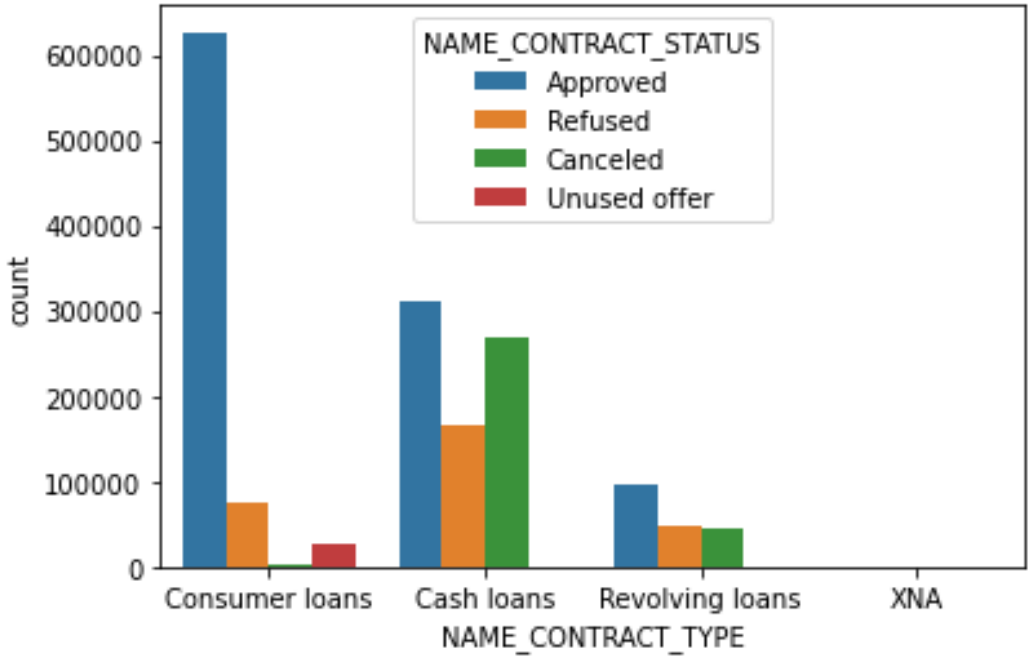
xii.) Surprisingly the people who earn more than 50L+ have the highest default rate. This may be because of the large loan amount

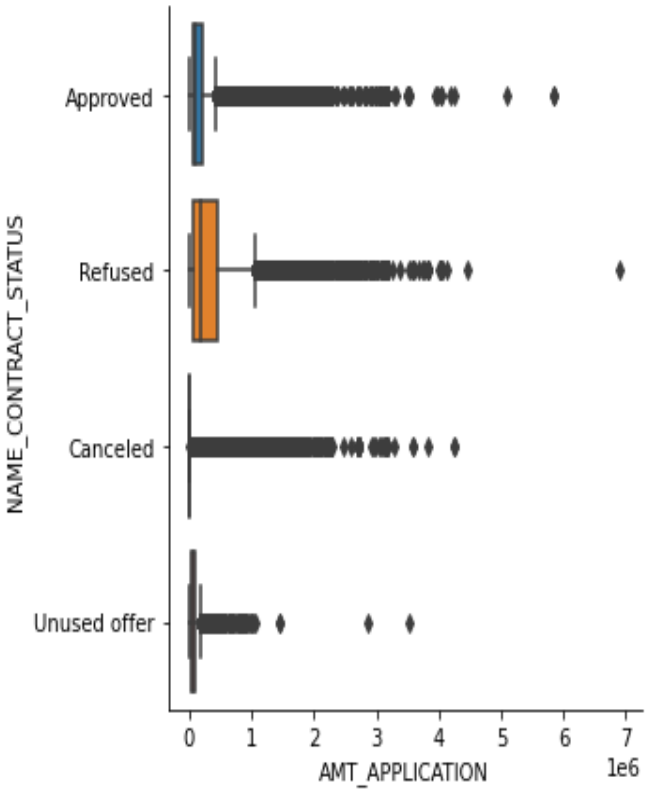
- Age Group Defaulter Distribution

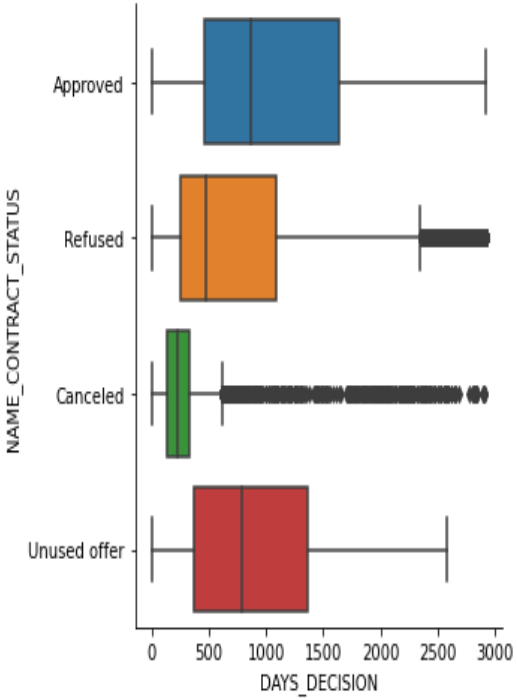


- The people in age group with 20-30 have highest default rate

Bivariate Analysis on Previous data




Approved
Refused
Canceled
Unused offer
NAME_CONTRACT_STATUS
0
1
2
3
4
5
6
7
AMT_APPLICATION
1e6


Approved
Refused
Canceled
Unused offer
NAME_CONTRACT_STATUS
0
500
1000
1500
2000
2500
3000
DAYS_DECISION

- **A few insights after bivariate analysis on Previous data:**

i.) Most of the Consumer loans are approved while there are varieties in Cash loans and Revolving loans

ii.) Most of the loan purpose is unknown but still the loans are approved

ii.) Most of the loans with Payment type as Cash through the bank are approved

iii.) Most of the clients are repeater. And their loans are approved more as compared to others

iv.) Many loans are rejected where channel type is Credit and Cash officers

v.) Many loan applications are rejected if the interest rates are unknown. Sounds odd

vi.) Most % of POS applications are Approved

vii.) The application where amount is small is small are usually approved. The median amount of rejected application id slightly higher

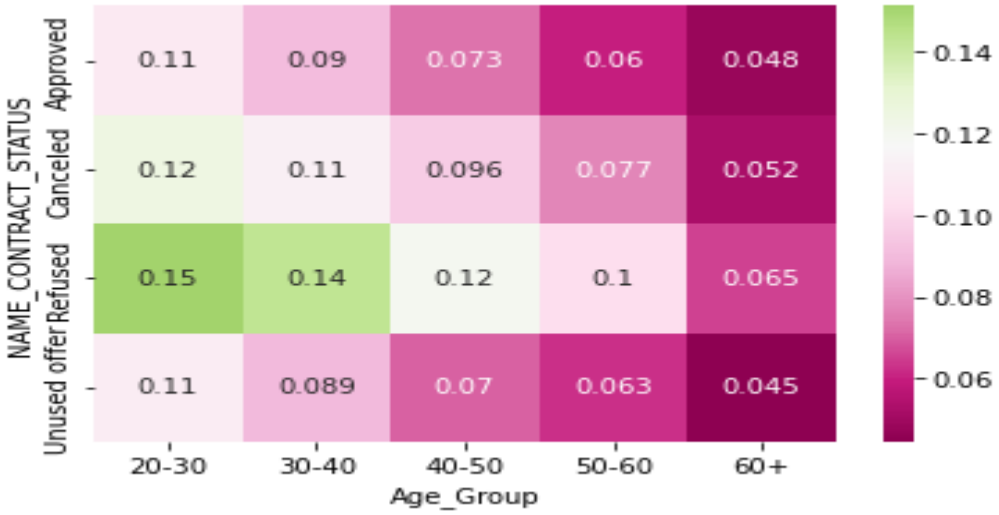
viii.) Usually approval process takes more number of days than refused application.

ix.) The terms with higher terms are rejected often

Multivariate Analysis using Merged data

- **A few information about the merged data**
- Merged data has 1413701 rows and 60 columns.
- The data type distribution of merged data is as follows:- category(3), float64(17), int64(14), object(26)

Age Group Wise Distribution



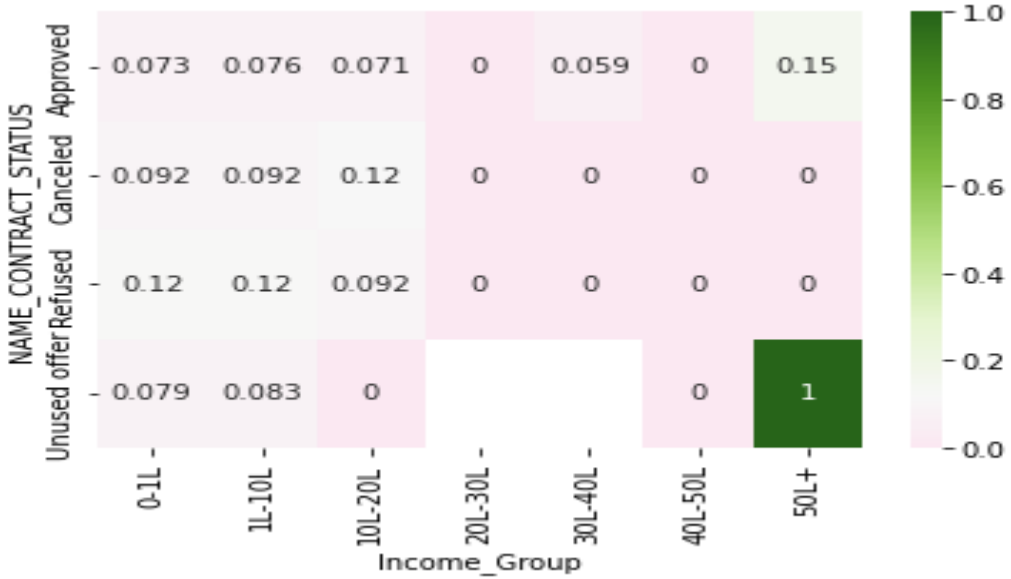
Insights:

i.) 15% application refused by the bank of age group 20-30 were defaulters hence avoiding the loss.

ii.) 11% defaulters were in age group 20-30 for which bank approved the loans resulting in business loss.

iii.) The people with age group above 60+ tend to default less and it's a good business for the bank

Income Group Distribution

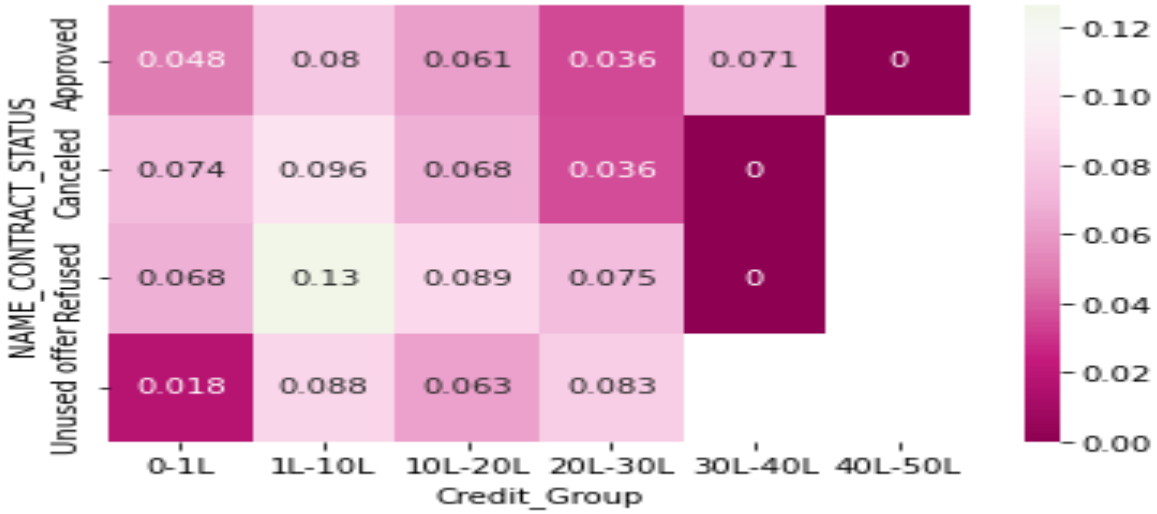


Insights:

i.) The bank has refused potential defaulters with income less than 10L who had 12% default rate

ii.) Most of the individuals earning between 20L -50L have repaid their loans. Hence bank has good business with these people

Credit Group Distribution



Insights:

i.) The bank has faced the most business loss(% wise) for the loans in range 30L-40L for which 7.1% people were the defaulters

ii.) There are no defaulters who have taken loans more than 50L+

iii.) The bank has rejected 13% application for loans between 1L-10L

Gender Wise Distribution

Insights:

i.) There is more business loss on loans given to males as compared to females.

ii.) The bank has refused more number of male applicant defaulters (% wise) as compared to females

Housing Type Distributio

Insights:

i.) The most loss incurred by the bank is from the people who either live in rented apartments or live with parents with default rates as 12% and 11% respectively.

ii.) The people with office apartment tend to default less as compared to others

Conclusions & Recommendations

- **Decisive Factor whether an applicant will be Repayer:**
- NAME_EDUCATION_TYPE: Academic degree has less defaults.
- NAME_INCOME_TYPE: Student and Businessmen have no defaults.
- DAYS_BIRTH: People above age of 50 have low probability of defaulting
- AMT_INCOME_TOTAL: Applicant with Income more income are less likely to default
- CNT_CHILDREN: People with zero to two children tend to repay the loans.

Conclusions & Recommendations

- **Decisive Factor whether an applicant will be Defaulter:**
- CODE_GENDER: Men are at relatively higher default rate
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
- NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
- OCCUPATION_TYPE: Avoid Low-skill Laborers as the default rate is huge.
- ORGANIZATION_TYPE: Organizations with highest % of loans that are not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).
  DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
- CNT_CHILDREN : Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
- AMT_CREDIT: When the credit amount goes beyond 3M, there is an increase in defaulters.