Credit EDA Case Study

Team Members:

Kushagra Misra

Minakshi Maurya

Outline

- Problem Statement
- Data Understanding
- Data Cleaning
- Data Analysis
- Conclusion statement

Problem Statement

- Based on data provided we have to figure out potential customers which will return the loan for the loan providing company.
- We have to segregate applicants on Good, Low Risk and High Risk, where
  - Good applicant will return the loan on time and will not have any issue with payment
  - Low Risk applicant has less changes of having issues with payment
  - High Risk applicant should be rejected
- Given data is having 2 dataset Application data and Previous Application data
  - Application Data is having 307511 records and 122variables(column)
  - Previous application is having 1670214 records 37 variables(columns)

Data Understanding

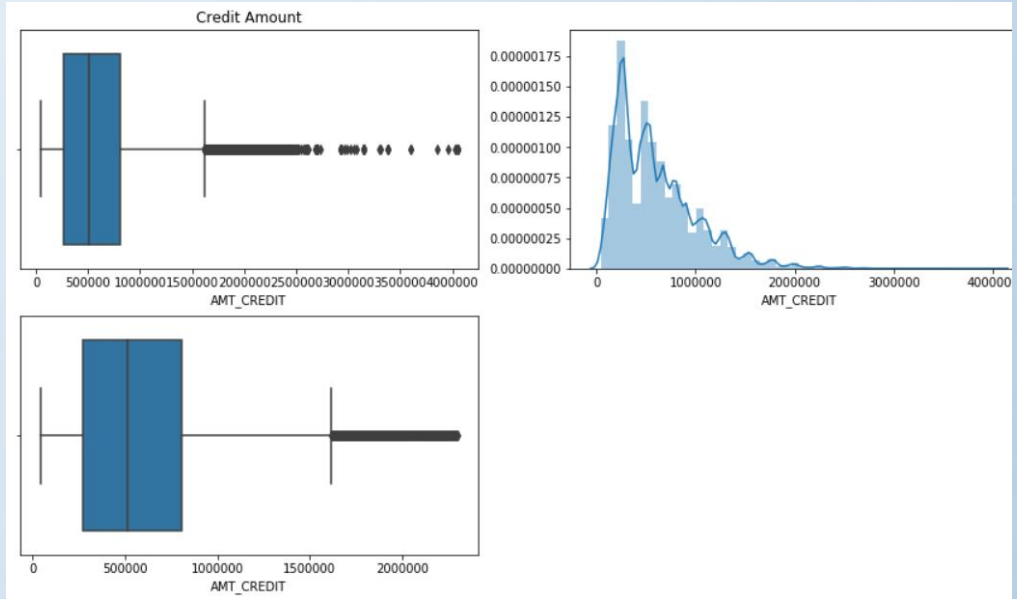
Missing Values in Dataset

- In Application data columns (variables) are having lot of missing values . Around 41 columns are with 50% or more than 50% null values. In order to do the EDA these variables will not have significant insight due high percentage of missing values, hence these columns have been dropped.
- In Previous Application also columns(variables) are having lot of missing data. Around 14 columns are with 205 or more than 20% columns. In order to do the EDA these variables will not have significant insight due high percentage of missing values, hence these columns have been dropped.

Outliers in Dataset

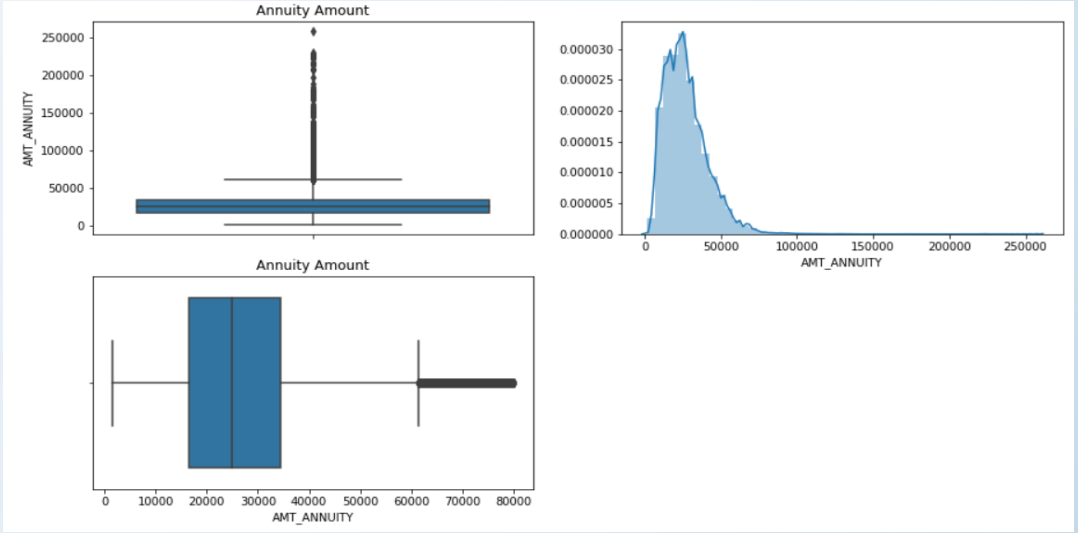
- In Application data, Outlier analysis is performed on Income, Credit amount, Annuity amount, Good's Price and birth days Variables.
- In Previous Application , Outliers analysis is performed on applied Loan amount and approved Loan amount.

✓ **Total Income**: For AMT_CREDIT variable as per box plot credit amount greater than 1500000 are outliers but from dist plot its visible that amount greater than 150000 are having some frequency, hence as per plot credit amount greater than 2300000 are rare and lies under outliers.

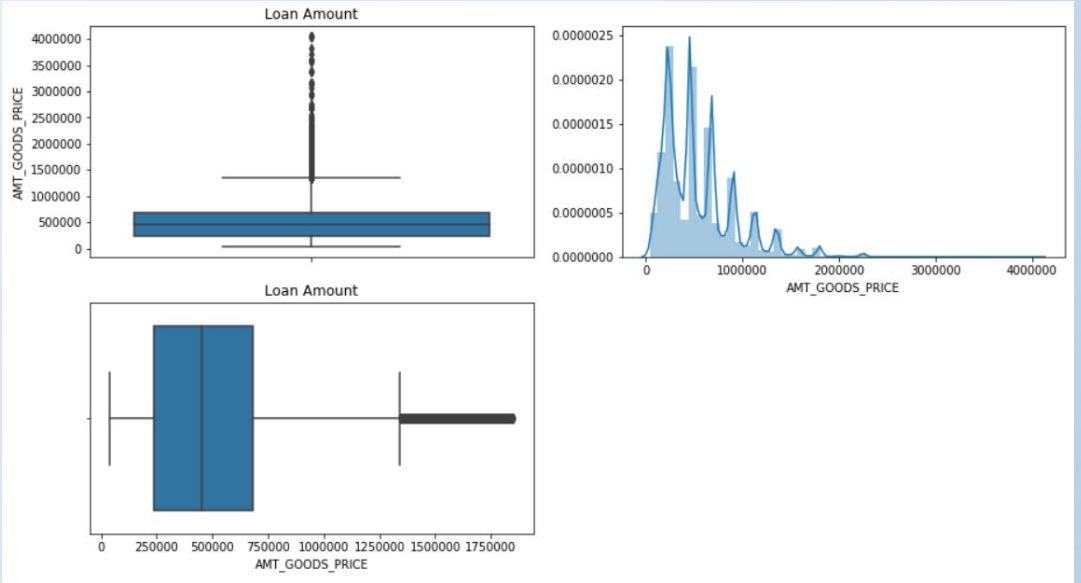


Data Understanding continue…..

- ✓ **Loan Annuity:** For AMT_ANNUITY variable as per box plot annuity amount greater than 80000 and via dist plot also same outliers are visible. Annuity amount greater than 80000 is rare and lies under outliers.

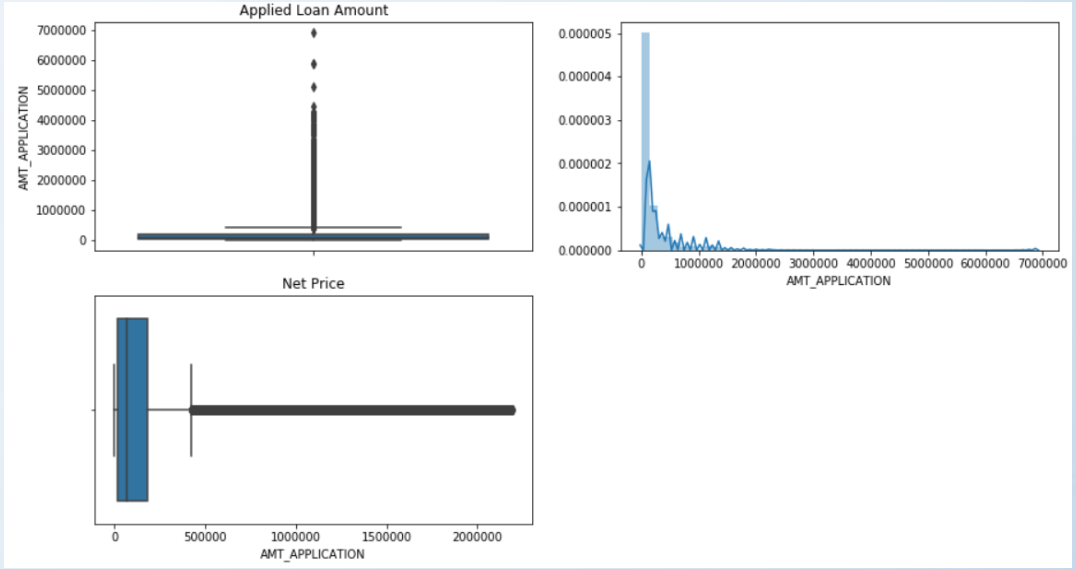


- ✓ **Loan Amount:** For AMT_GOODS_PRICE variable as per box plot credit amount greater than 1500000 are outliers but from dist plot its visible that amount greater than 150000 are having some frequency, hence as per plot credit amount greater than 1850000 are rare and lies under outliers.

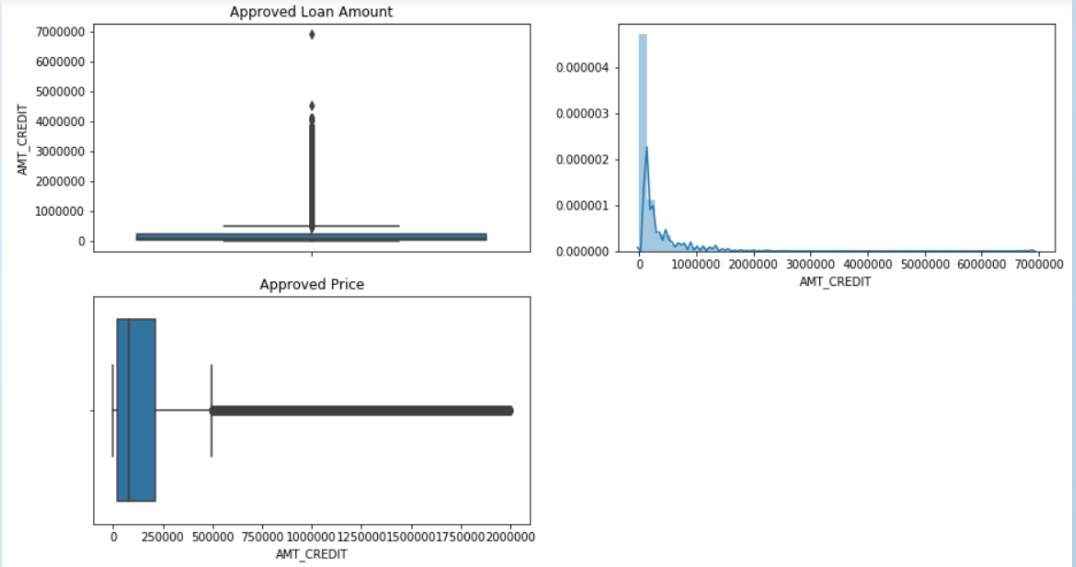


Data Understanding continue…..

- ✓ **Applied Loan Amount:** For AMT_APPLICATION variable as per box plot application amount greater than 500000 but from dist plot its visible that amount greater than 500000 are having some frequencies, hence as per insight applied loan amount greater than 2200000 are rare and lies under outliers .



- ✓ **Approved Loan Amount:** For AMT_CREDIT variable as per box plot credit amount greater than 500000 are outliers but after plotting dist plot its visible that amount greater than 500000 are having some frequency, hence as per insight credit amount greater than 2000000 are rare and lies under outliers.

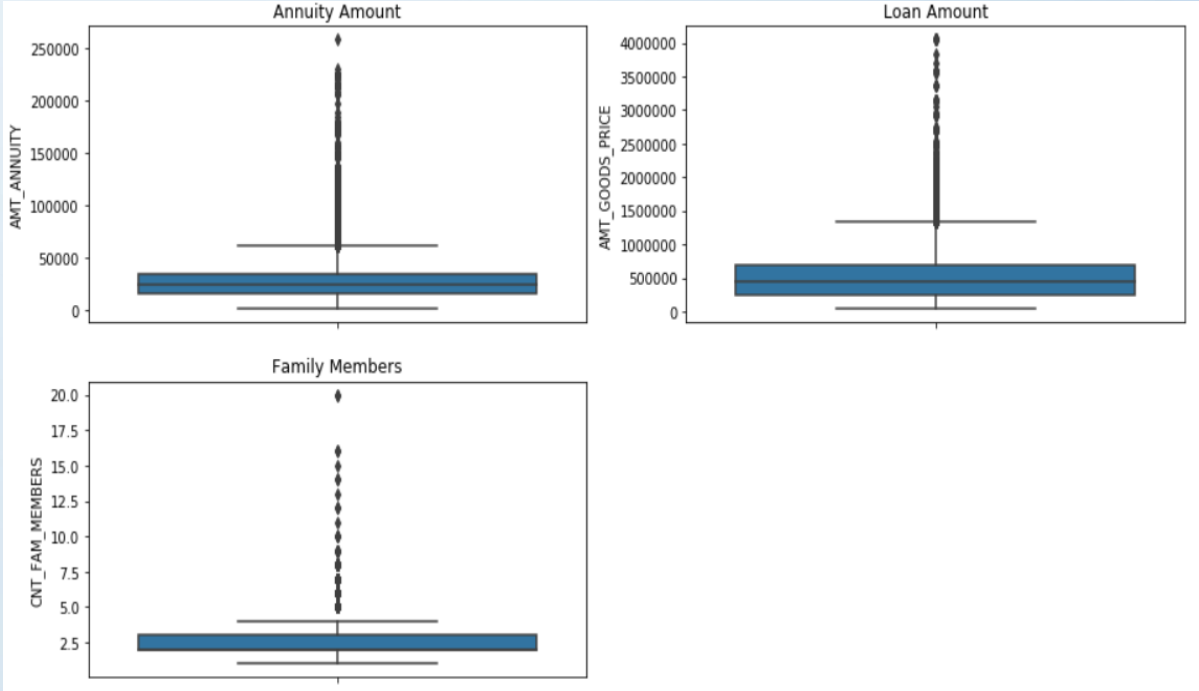


Data Cleaning and manipulation

Missing value Imputation:

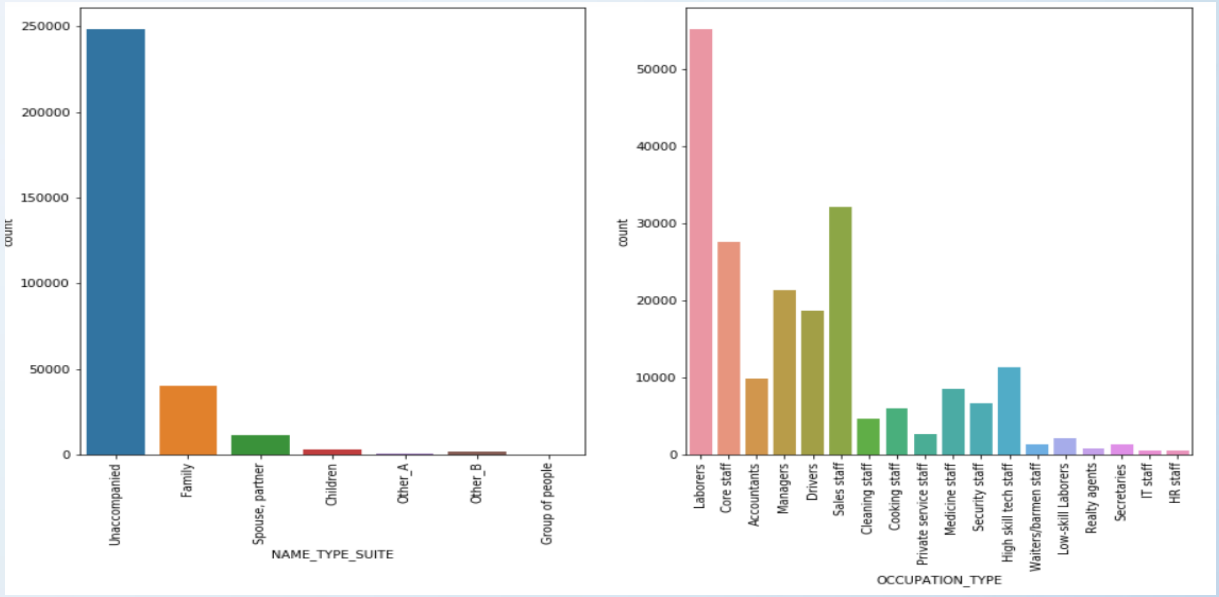
- Missing value imputation has been done for the variables having missing values less than 13%
- In Application data missing values analysis has been on Annuity amount, Loan amount, Family member, Occupation type and accompanying person.
- In previous Application data missing values analysis has been performed on Approved loan amount and Product combination

- ✓ **Annuity Amount:** As per insight for Annuity amount there are lot of Outliers, hence missing values can be impute with median value (24903) of the variable.
- ✓ **Loan Amount:** As per insight for Loan amount there are lot of Outliers, hence missing values can be impute with median value (450000) of the variable
- ✓ **Family Members:** As per insight for Family Members there are not much Outliers, hence missing values can be impute with mean value (2) of the variable

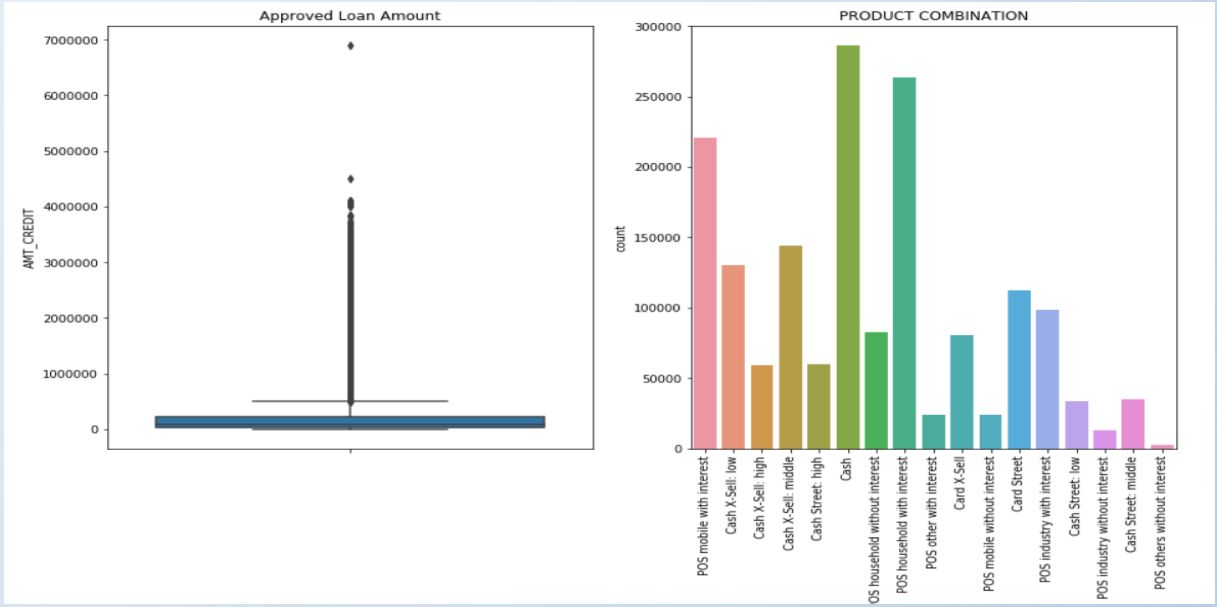


Data Cleaning and manipulation Continue..

- ✓ **Accompanying person :** As per insight for accompanying person frequency of Unaccompanied is more and that is the mode of variable, hence missing values can be impute with “Unaccompanied”.
- ✓ **Occupation type:** As per insight for Occupation Type frequency of Laborers is more and that is the mode of variable, hence missing values can be impute with “Laborers”.



- ✓ **Approved Loan Amount in Previous Application:** As per insight for Approved Loan amount there are lot of Outliers, hence missing values can be impute with median value (80541) of the variable.
- ✓ **Product Combination:** As per insight for Product Combination frequency of Cash is more and that is the mode of variable, hence missing values can be impute with “Cash”.



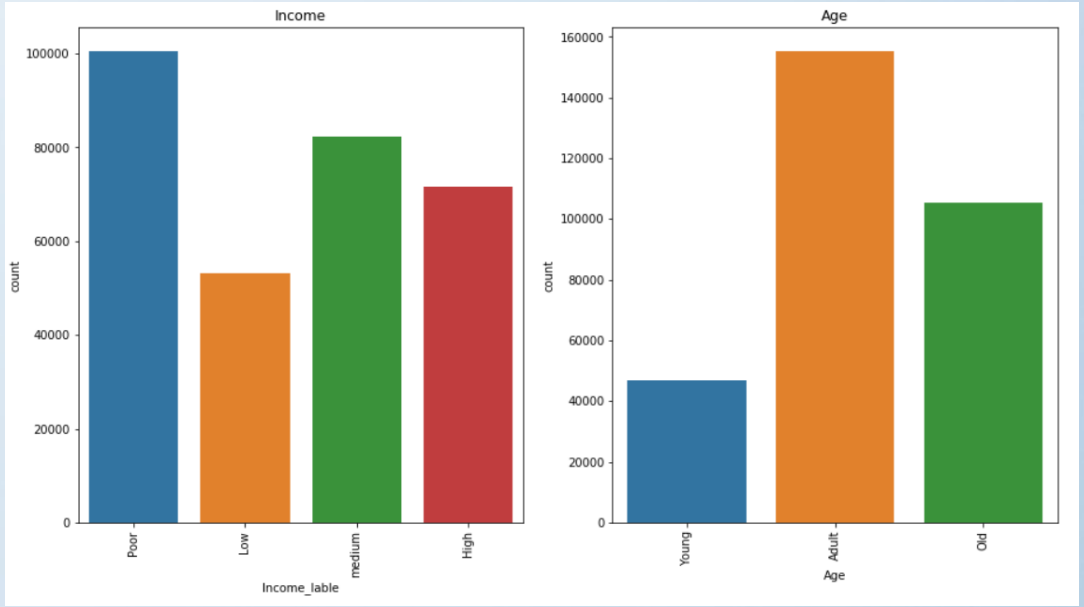
Data Cleaning and manipulation Continue..

Data Manipulation:

- In Application data DAYS_REGISTRATION, CNT_FAM_MEMBERS, DAYS_LAST_PHONE_CHANGE variables are having data type Float64 as these variables should be integer, hence converted datatype for the variables from Float64 to INT64
- As CNT_FAM_MEMBERS variable is having null values ,hence while changing data type, it was throwing error for null values. To resolve the error first have imputed missing values with “Unaccompanied” and then converted data type to INT64.
- As DAYS_LAST_PHONE_CHANGE variable is having null values ,hence while changing data type, it was throwing error for null values. To resolve the error, as this is having very less null values, have dropped rows and then converted data type to INT64.
- In Application DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE variables are having negative values which have been converted to positive.

Binning:

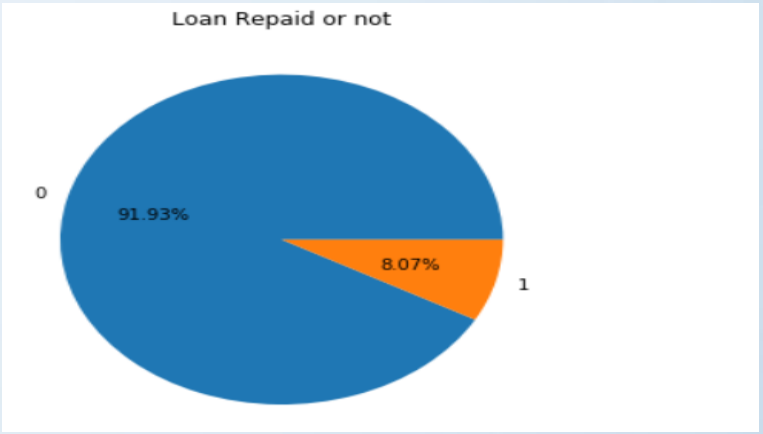
- Binning has been performed on In Application data for AMT_INCOME_TOTAL(Total Income) and DAYS_BIRTH
  - ✓ **AMT_INCOME_TOTAL :** Based on Quantile have binned AMT_INCOME_TOTAL column.
  - ✓ **DAYS_BIRTH :** After converting days to age by dividing the column by 365 , have binned in 3 categories**.**



Data Analysis

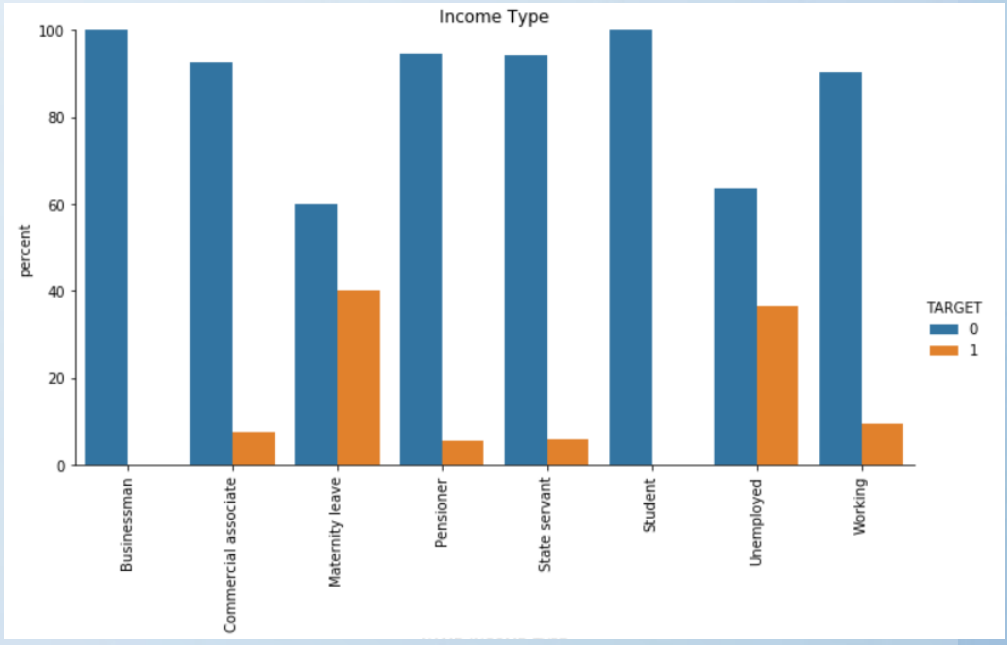
Imbalance Percentage :

- In Application data Imbalance percentage has been checked with respect to TARGET (Default or Non Default)
- As per the pie plot its highly imbalance because for defaulter its only 8.07% and for non defaulter its 91.93%



Analysis based on Applicants Income Type

- 100% Businessmen and students are paying loan amount on time or we can say do not have any payment difficulties`
- 40% of Maternity Leave and 36.36% of Unemployed are Defaulters or have payment difficulties`
- Most of the people who are working as Commercial associate or pensioner or state servant or belong to working class are paying their installment on time



Data Analysis continue…..

Correlation between Numeric variables with respect to TARGET

- Non Defaulter (TARGET = 0) :

	FEATURE_1	FEATURE_2	CORRELATION
211	DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999756
94	AMT_CREDIT	AMT_GOODS_PRICE	0.987022
438	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950149
62	CNT_CHILDREN	CNT_FAM_MEMBERS	0.878572
116	AMT_ANNUITY	AMT_GOODS_PRICE	0.776423
93	AMT_CREDIT	AMT_ANNUITY	0.771298
185	DAYS_BIRTH	DAYS_EMPLOYED	0.625928
189	DAYS_BIRTH	FLAG_EMP_PHONE	0.621888
173	REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	0.539006
174	REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT_W_CITY	0.537302

- Defaulter (TARGET = 1) :

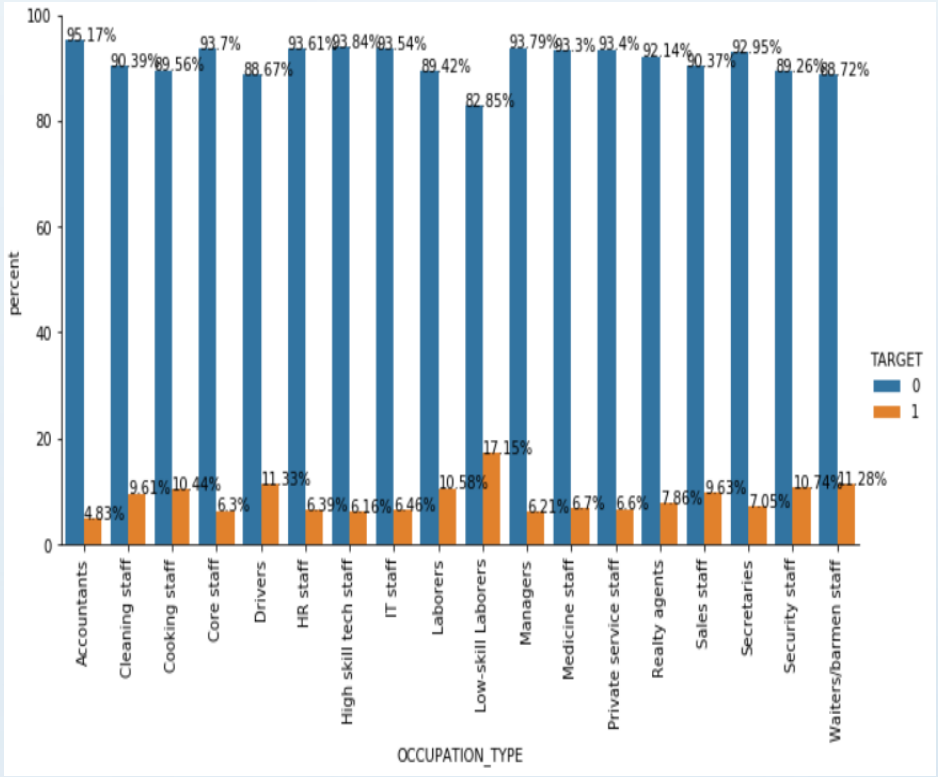
	FEATURE_1	FEATURE_2	CORRELATION
211	DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999705
94	AMT_CREDIT	AMT_GOODS_PRICE	0.982783
438	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
62	CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
116	AMT_ANNUITY	AMT_GOODS_PRICE	0.752295
93	AMT_CREDIT	AMT_ANNUITY	0.752195
185	DAYS_BIRTH	DAYS_EMPLOYED	0.581849
189	DAYS_BIRTH	FLAG_EMP_PHONE	0.578184
174	REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT_W_CITY	0.446977
173	REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	0.443236

Numeric Variables with high correlation are same for both Defaulter and non Defaulter

Data Analysis continue…..

Applicant analysis based on occupation (Considering 10% as the threshold **)**

- Applicant in occupation where they are facing issues in payment are Cooking staff, Drivers, Laborers, Low-skill Labors, Security staff, waiters/barmen staff
- People working as Accountant are more likely not to have any issue with the installments



Potential Customers beyond 10% threshold

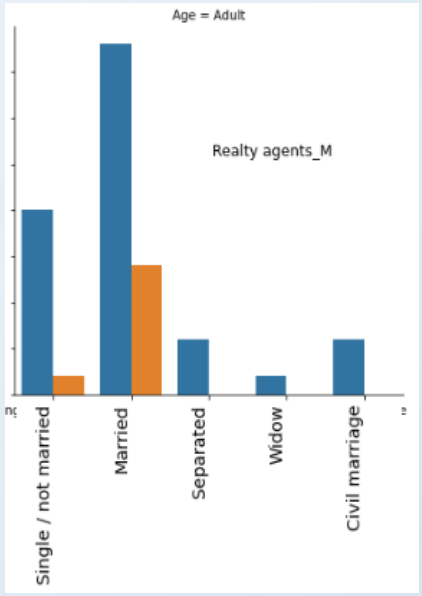
- There is good number of people who are not having issues in payment in Cooking staff, Drivers, Laborers, Low-skill Labors, Security staff, waiters/barmen staff
- We can perform Multi-Variant analysis on occupation and get more filtered data to segregate applicants who are potential customers

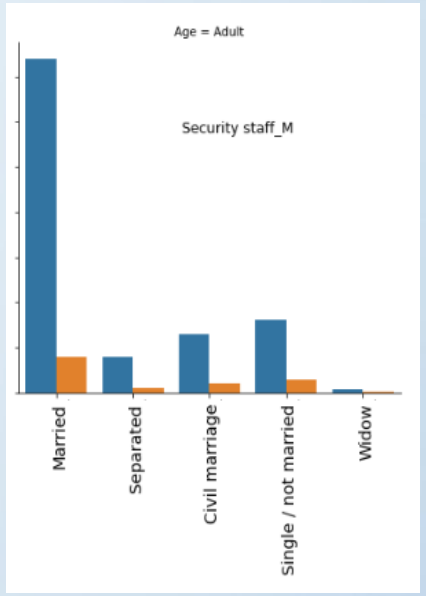
Data Analysis Multi-Variant

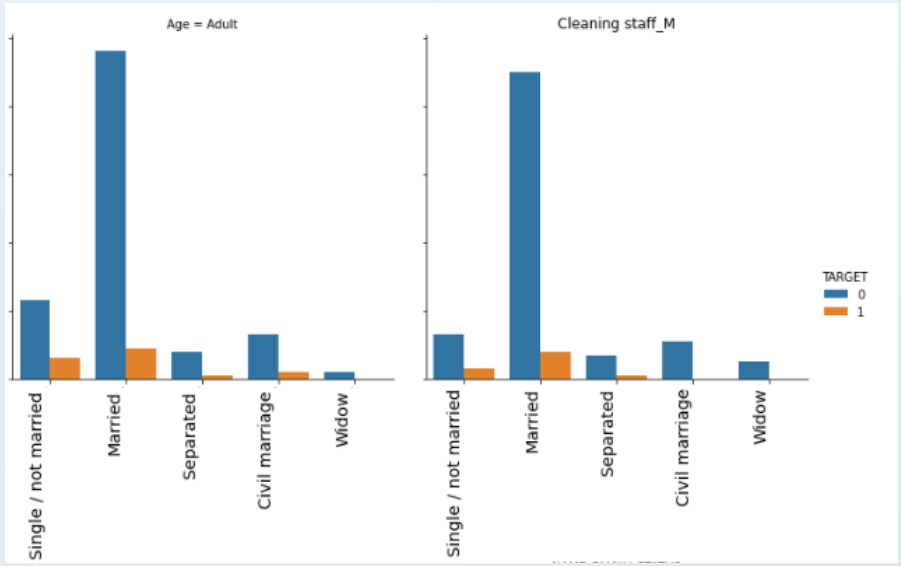
Multi-Variant Analysis on Columns
(Age, Family Status, Job Type and *Gender as* ***Male*****)**

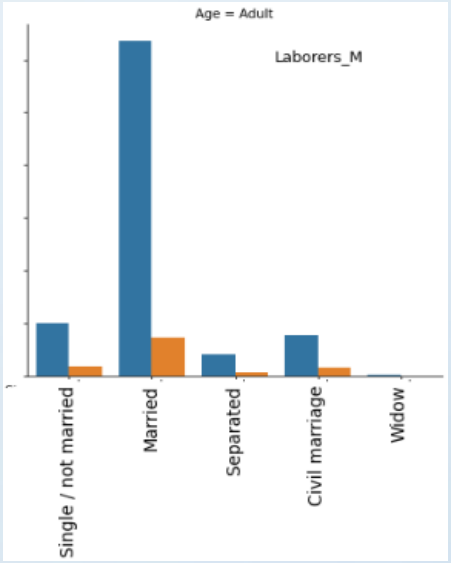
- ***Married Male Applicants in the age group of 30-50(Adults)*** in following sectors Realty agents, Security Staff, Cleaning Staff, Labours, Drivers, Low Skill Labours should be Rejected

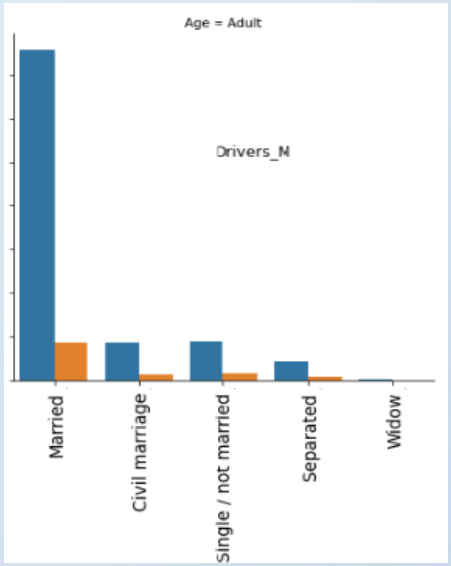
- We can observe from ***previous slide Reality agents*** comes under 10% threshold but from Multi-variant we can observer ***most of them are Married Male Applicants in the age group of 30-50(Adults)***.

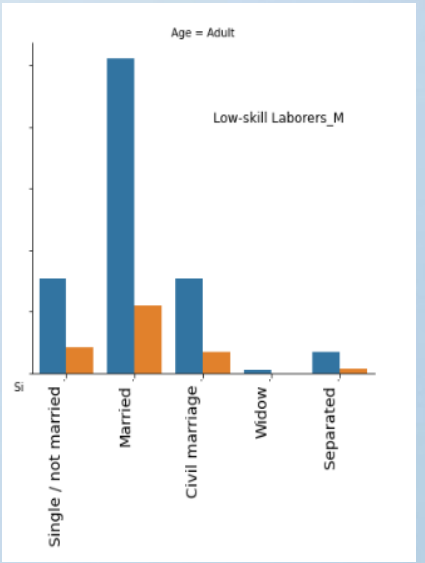








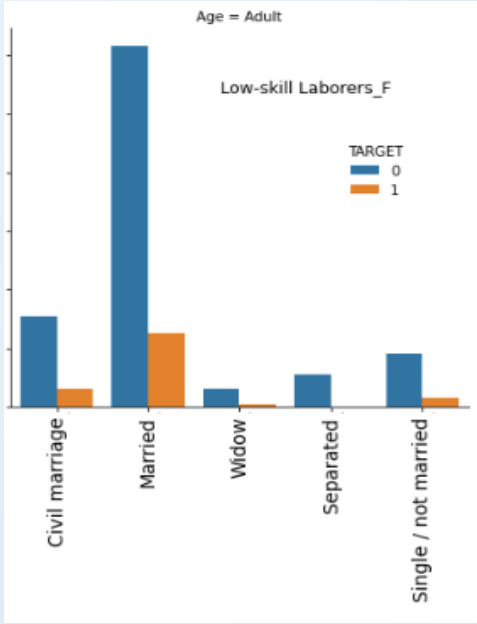


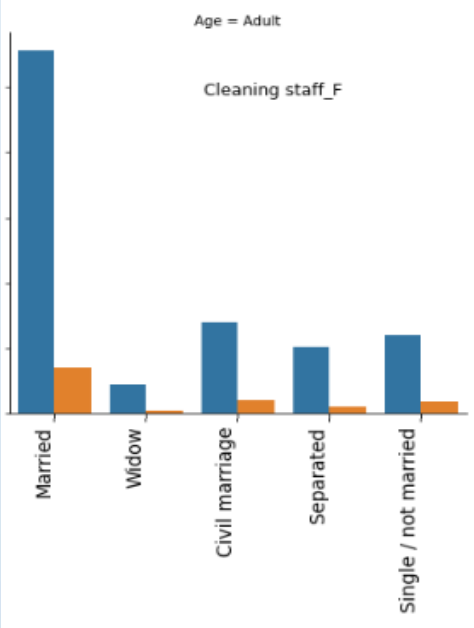


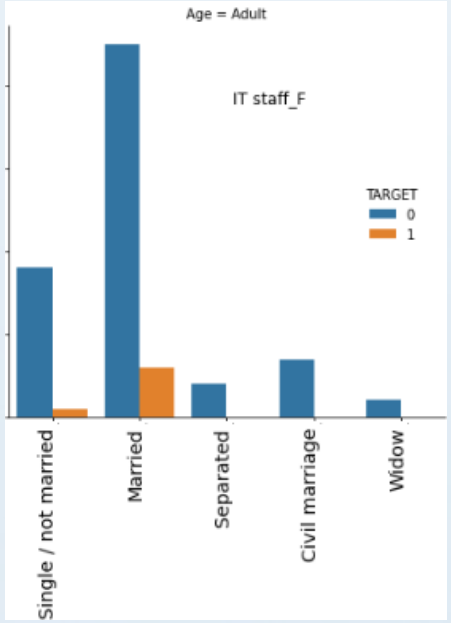
Data Analysis Multi-Variant

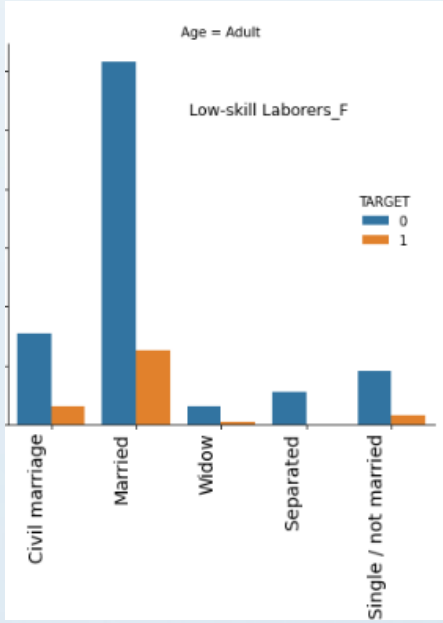
Multi-Variant Analysis on Columns
(Age, Family Status, Job Type and *Gender as* ***Female*****)**

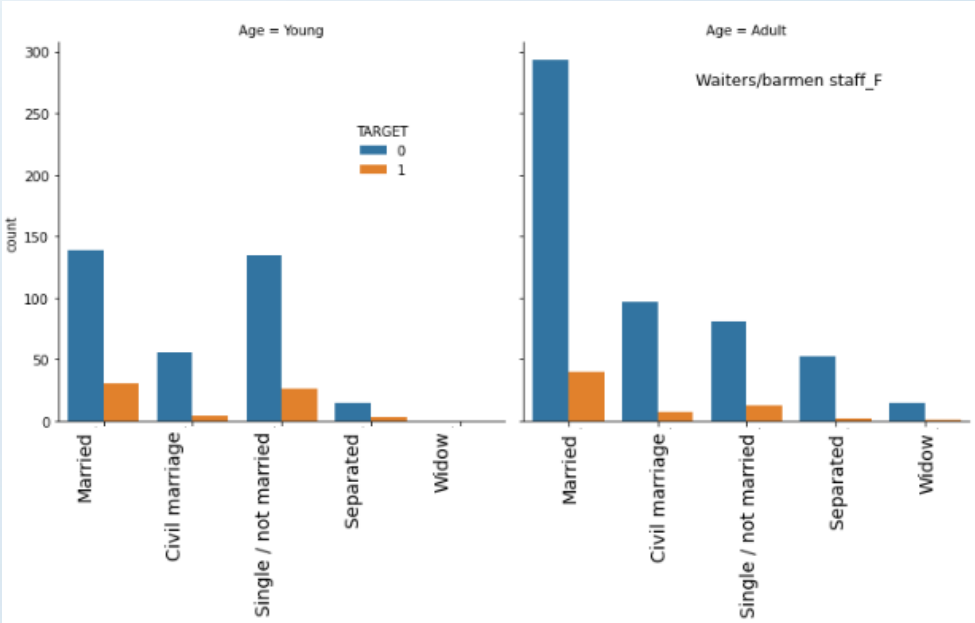
- ***Married Female Applicants in the age group of 30-50(Adults)*** in following sectors Realty agents, Security Staff, Cleaning Staff, Waiter/barmen staff, Labours, IT staff, Low Skill Labours should be Rejected
- We Observe Married Females applicants working as IT staff are having issues with their payment and should be reject or added extra interest to cover loss. *This data was not available in bi-variant analysis*
- Young Married female applicants working as waiter or barmen staff are also facing issues with their payments

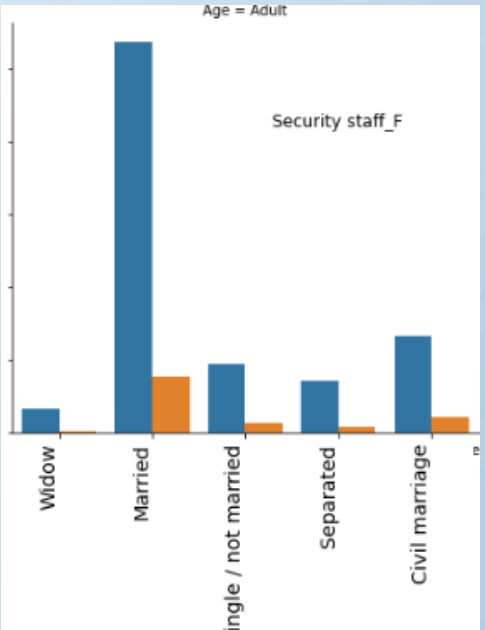












Conclusion

Based on all analysis on provided data we conclude there are 3 types of applicants available

Good Applicant (Normal interest Rate)	Risk (High interest Rate)	Reject
• Previously had no issues in payment • Lives in or from Region 1 • Male applicants he should be part of core staff, Accountants or manager, and age is more than 50 • For old age Females applicants with age more than 50 should be **labor, Core staff, Accountants, Managers, Drivers, Sales Staff, Private Service staff, Medical Staff, High skill labor, reality agent, Secretaries, IT staff or HR staff** and **Not Married** • **Male in age from 30-50 working as HR** • Applicants who have not defined their gender	• Previously had no issues in payment • Can live in any region • Married male in age group of 30-50 **Not working as Low Skill Labor, Reality Agent, Security Staff, Labors, Accountants, Drivers** • Married female applicants in age between 30-50 and **Not working in IT, Low skill labor, Waiter/barmen , Security Staff, cleaning or sales**	• Previously had payment issue • Applicant is Married and Maternity leave and in age group of 30-50. • Applicant is **Married** in **age group if 20-50** and **Unemployed.** • Applicant is **unemployed** and in **age group of more than 50**

Thank You