EDA CASE STUDY
LOAN RISK ANALYSIS

Prepared By- **SANCHIT KUMAR SHARMA**

Email Id- Sanchit.alig@gmail.com

THERE ARE MANY TYPES OF RISKS ASSOCIATED WITH ANY LOAN.

- There are different types of risks associated with loans and in this case study we have analysed the data and found different outcomes and parameters on which the loans should be lended to applicants.
- The Fintech companies and banks earning source is from the interests earned from the loans so it is essential to lend loans to customers.
- But it is also essential to lend loans to only customers who are likely to repay the loans.
- We will analyse the criteria on which the loans should be distributed among defaulters and Non defaulters.

IMPORTANT POINT

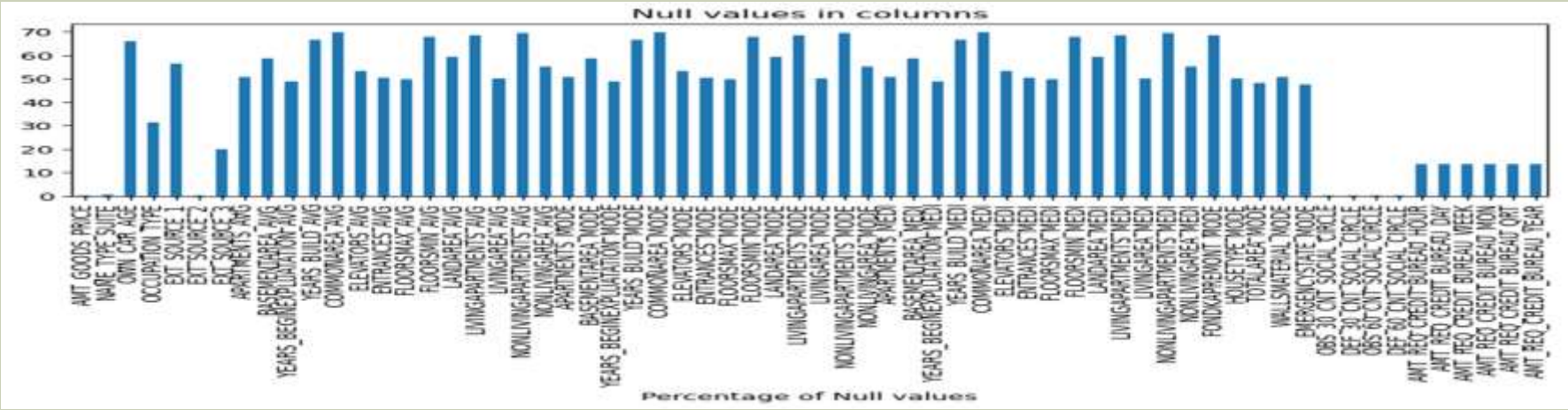
- The loans should be lended to customers who are able to repay the loans because if the loans are not lended to non defaulters than it will cause revenue losses for the banks and Loan lending companies.
- Also loans must not be lended to the customers who are not able to repay the loans because due to defaulters also heavy losses are suffered by banks and Loan lending companies.
- So, the non defaulters should get the loan and defaulters must not get the loans.
- Now , we will understand the analyses from next slide.

ANALYSIS DONE AND STEPS

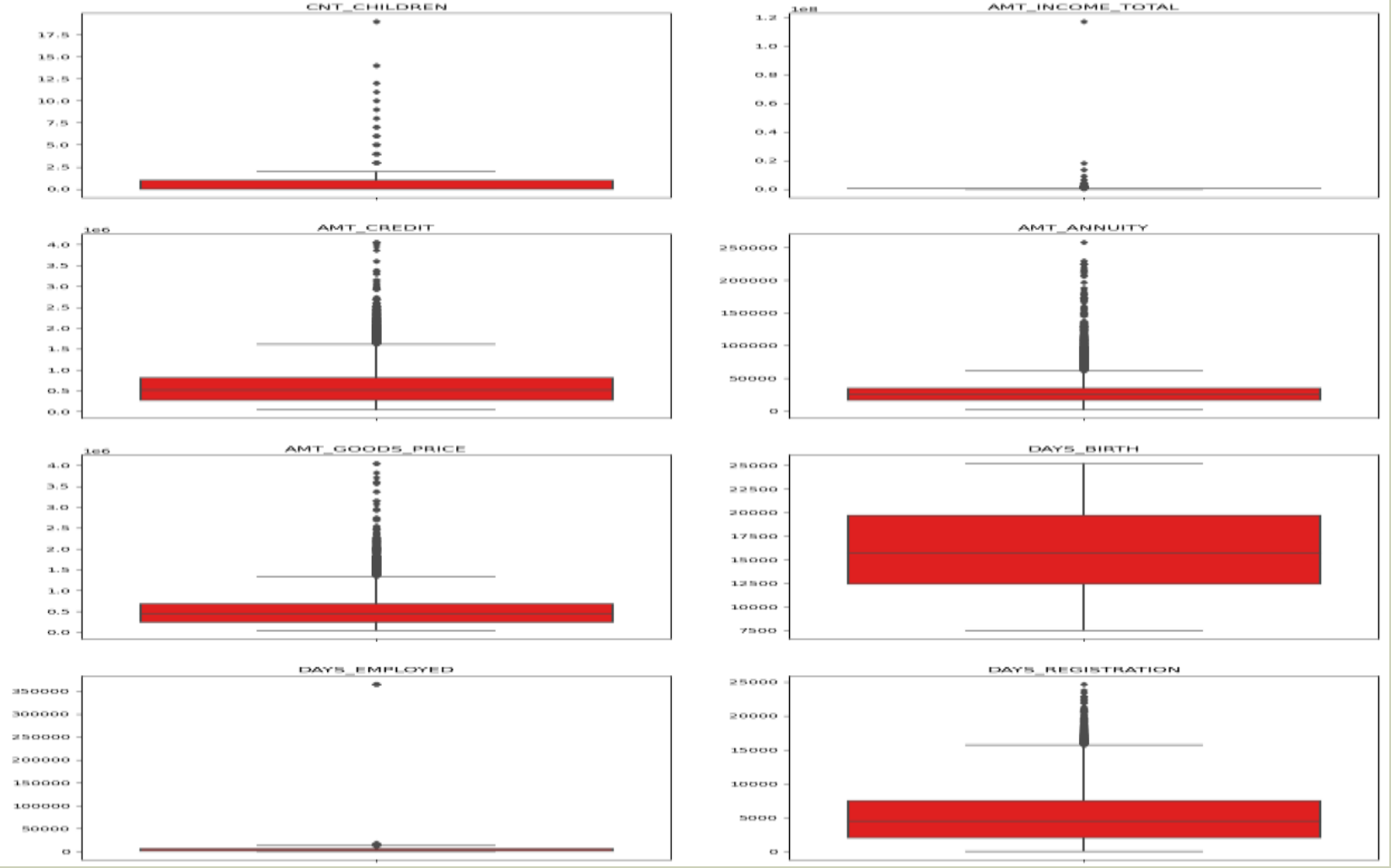
- Understanding the Data
- Data Cleaning and imputation necessary values
- Finding Outliers in the Data
- Data imbalance check between Non Defaulters and Defaulters
- Univariate and Bivariate Analysis with Respect to Non Defaulters and Defaulters.
- Finding the Correlation between the columns with respect to Non Defaulters and Defaulters and after that plotting correlation analyses on heat map.
- Merging the Data frames and understanding outcomes of the data

DATA CLEANING AND MANIPULATION

- The graph shows percentage of Null values in columns of application_data dataset and we have cleaned the data by dropping columns with null values more than 40%.
- Now, we found that columns with null values more than 1% and replaced these null values with the median value.
- After that, we dropped flag columns that were irrelevant after analysing them.
- Occupation_type columns was important so we imputed “missing“ in place of null values in it.
- Now, after cleaning and necessary manipulation we move forward for the next analysis of the cleaned data.



IDENTIFYING OUTLIERS

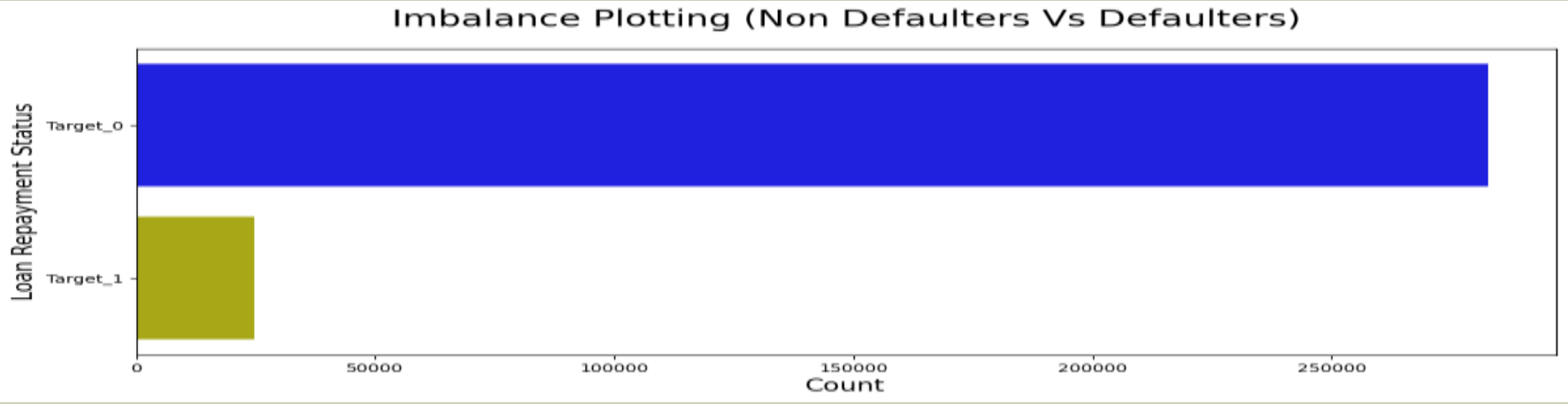


CONCLUSION OF OUTLIERS

- 1st quartile is missing for CNT_CHILDREN which means most of the data are present in the 1st quartile
- In AMT_INCOME_TOTAL only single high value data point is present as outlier.
- AMT_CREDIT has little bit more outliers.
- 1st quartiles and 3rd quartile for AMT_ANNUITY is moved towards first quartile.
- 1st quartiles and 3rd quartile for DAYS_EMPLOYED is stays towards first quartile. It can be seen that in current application data
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- DAYS_BIRTH has no outliers which means the data available is reliable.
- DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.

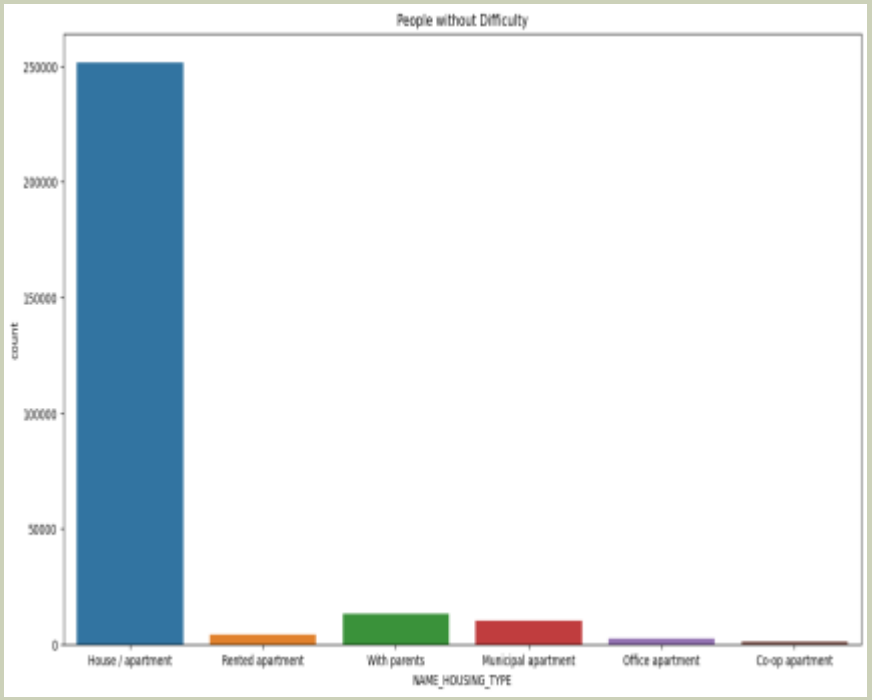
DATA IMBALANCE

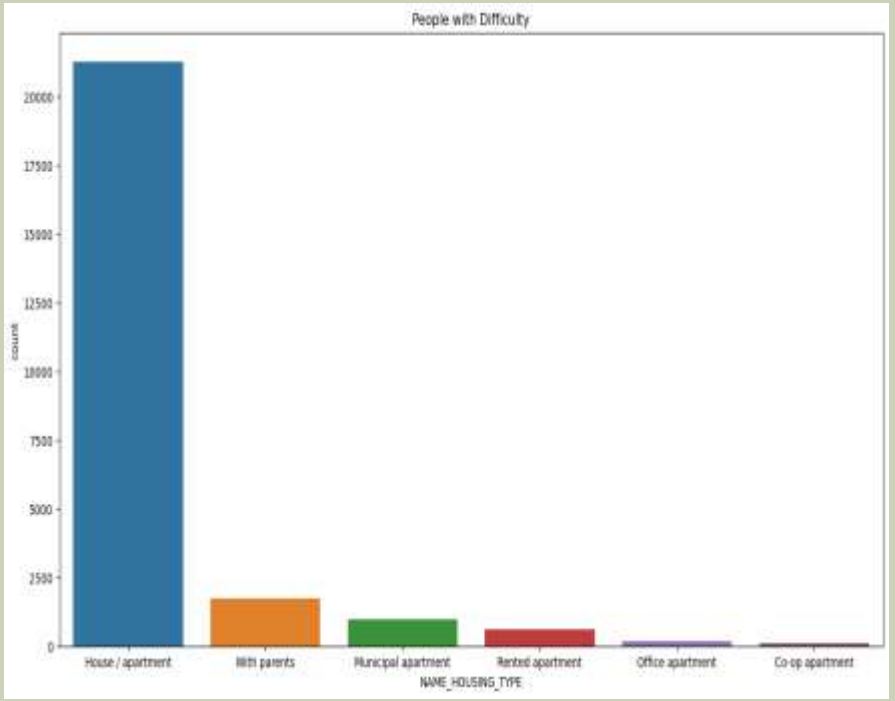
- Non Defaulters Percentage is 91.93%
- Defaulters Percentage is 8.07%
- Imbalance Ratio with respect to Non Defaulters and Defaulters is given: 11.39/1 (approx.)
- This shows that practically Non Defaulters are much more higher as compared to Defaulters. Which is in real scenario also true, because if the number Of Defaulters will be high than the bank will suffer heavy losses.



MULTIVARIATE AND BIVARIATE ANALYSIS

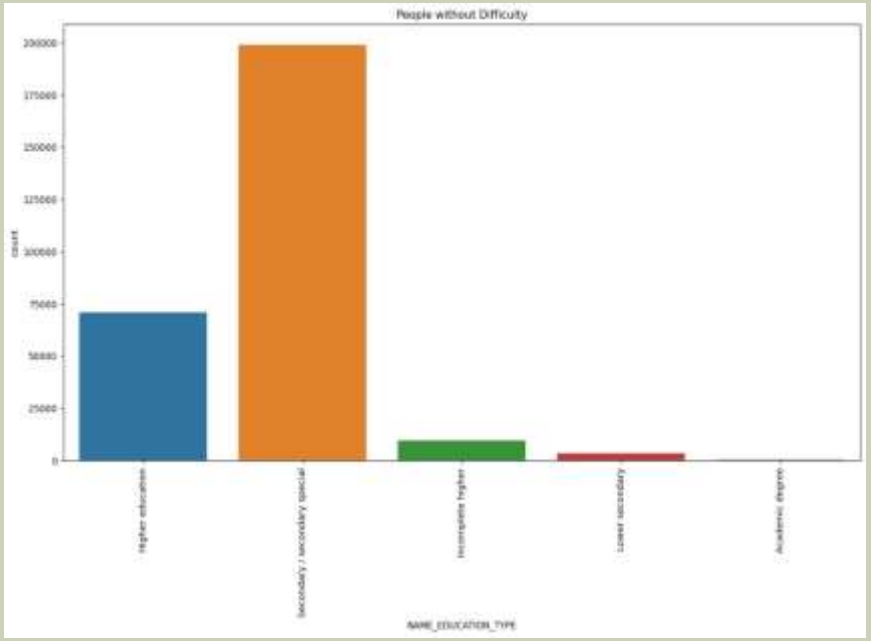
- Analysis of non Defaulters and Defaulters bases on Housing Type.
- Majority of people live in House/apartment.
- People living in office apartments have lowest default rate.

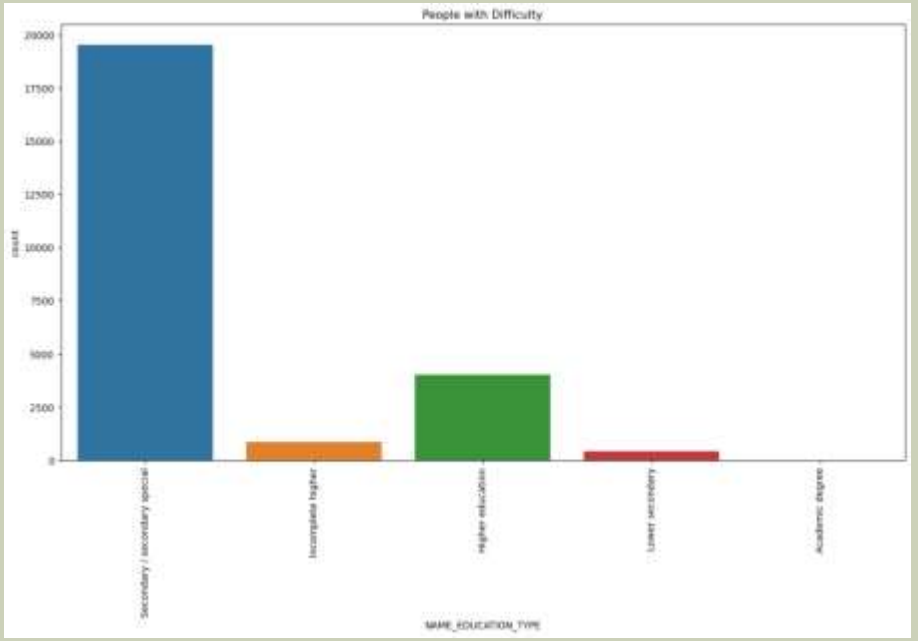




UNIVARIATE ANALYSIS BASED ON EDUCATION TYPE

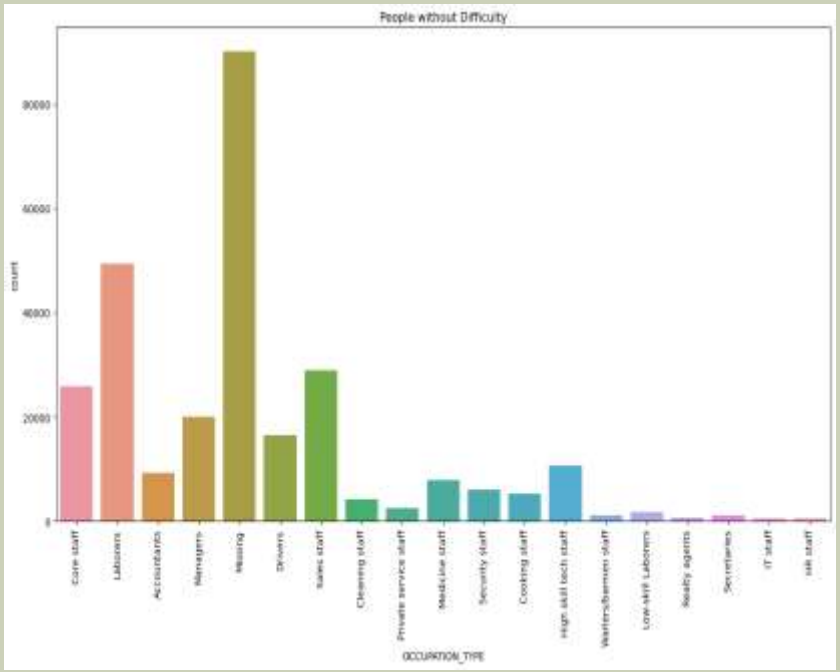
- People with Academic degree are least likely to default.
- Majority of applicants have Secondary/secondary special education, followed by clients with Higher education.
- Very few applicants have an academic degree.

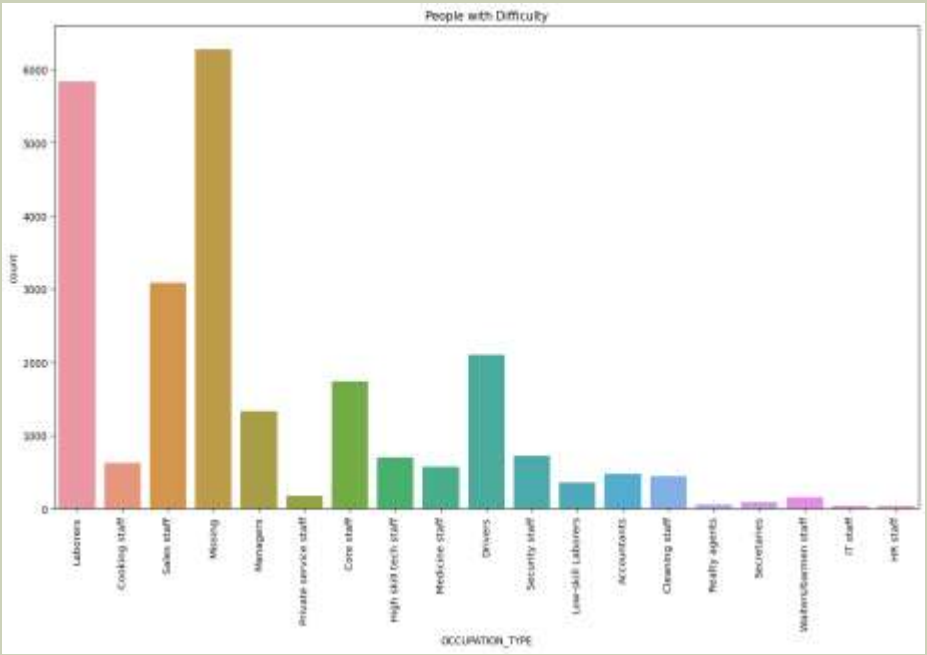




UNIVARIATE ANALYSIS BASED ON OCCUPATION

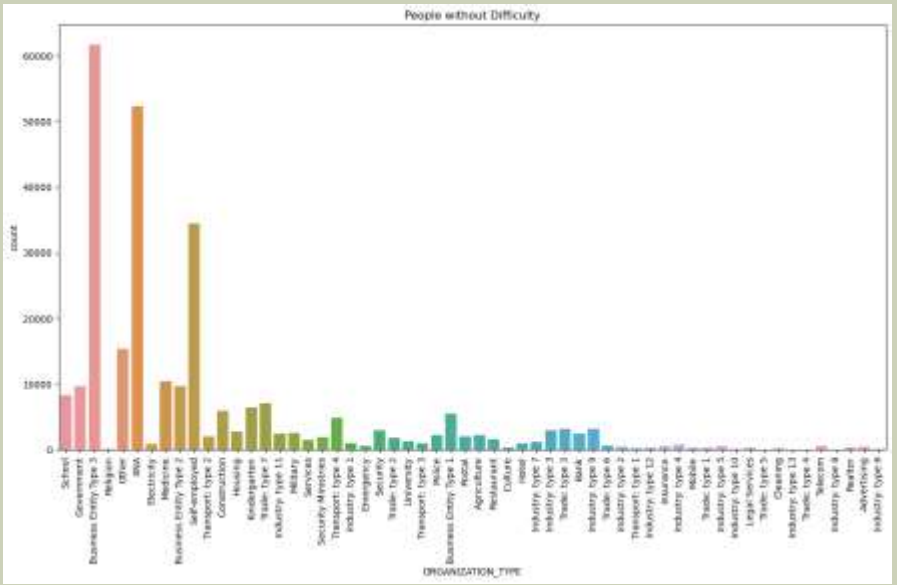
- Most of the loans are taken by Laborers, followed by Sales staff.
- IT staff and HR staff are less likely to apply for Loan.

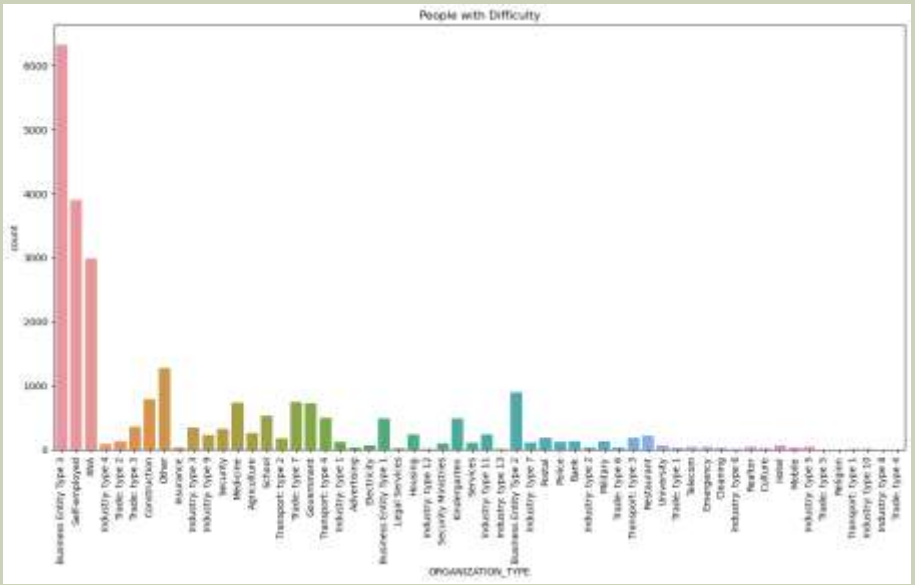




UNIVARIATE ANALYSIS BASED ON ORGANIZATION TYPE

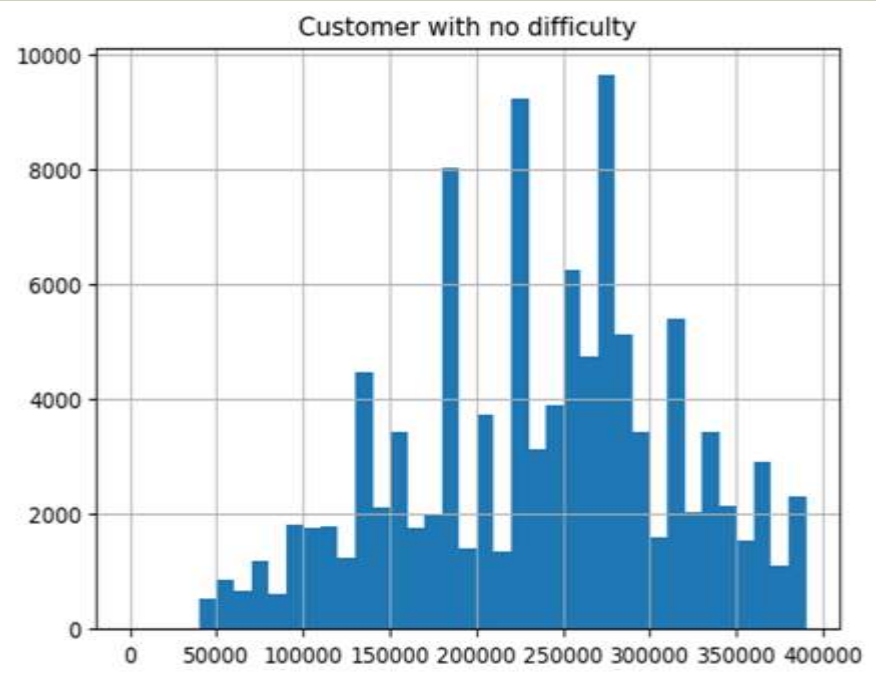
- Self employed people have relative high default rate, to be safer side loan disbursement should be avoided.
- Most of the people application for loan are from Business Entity Type 3
- For a very high number of applications, Organization type information is unavailable(XNA)

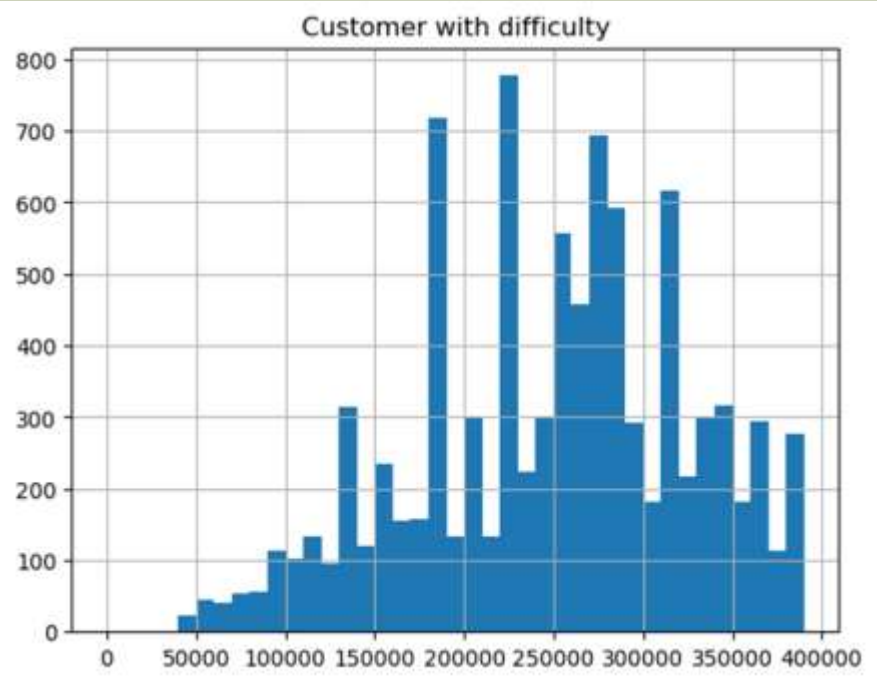




UNIVARIATE ANALYSIS BASED ON AMOUNT CREDIT

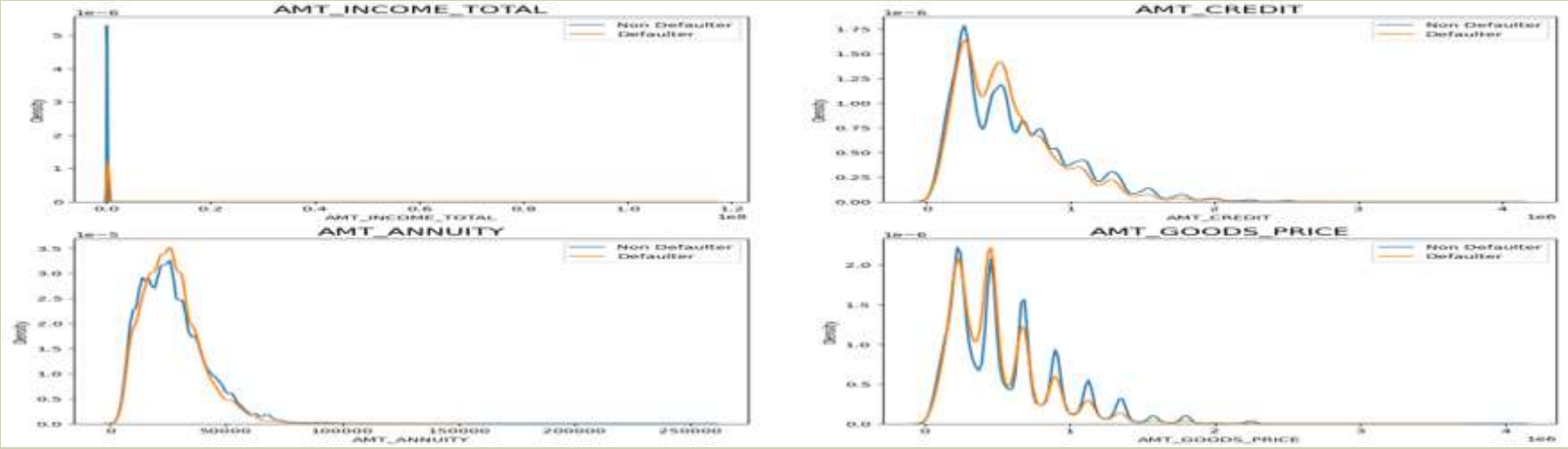
- **Most of the loan defaulters are the customers whose income ranges in between 260000 to 290000.**



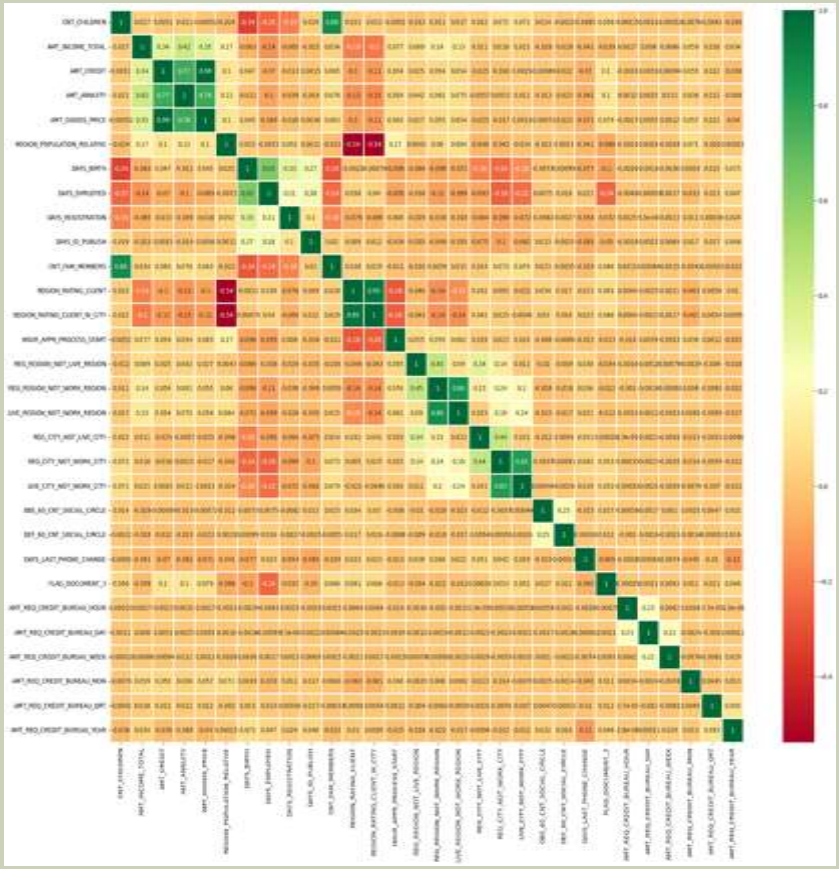


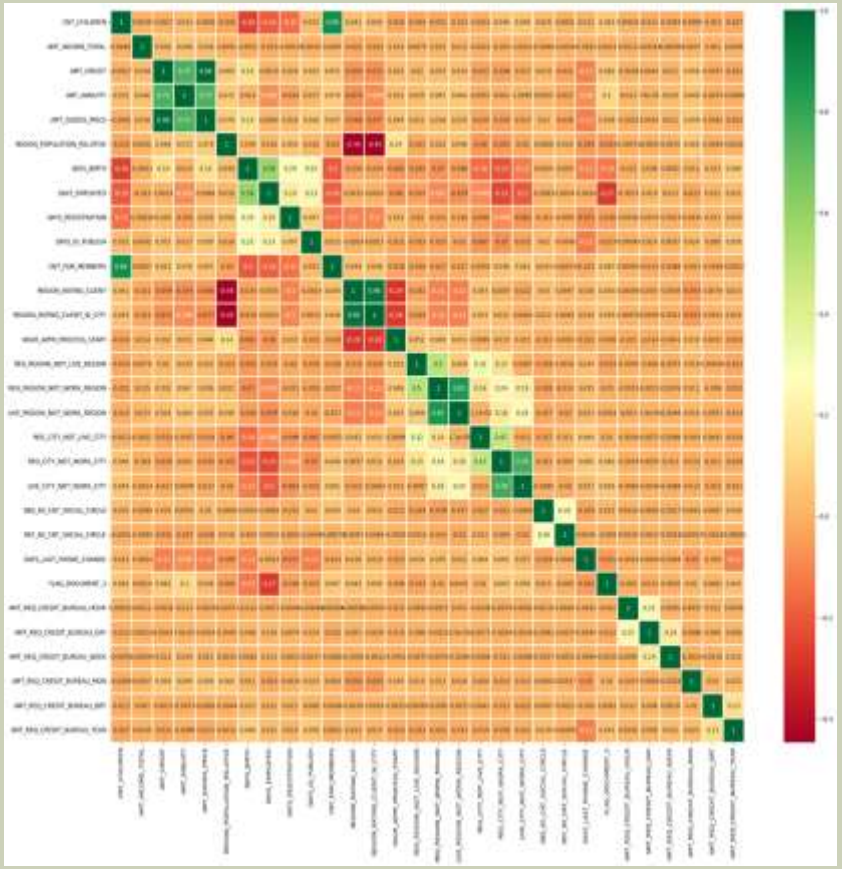
CONCLUSION OF ANALYSIS OF AMOUNT CREDIT, ANNUITY, INCOME, GOODS PRICE

- Most no of loans are given for goods price below 10 lakhs.
- Most people pay annuity below 50 thousand for the credit loan.
- Credit amount of the loan is mostly less then 10 lakhs.
- The non defaulters and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision.



CORRELATION HEATMAP FOR NON DEFAULTERS AND DEFAULTERS

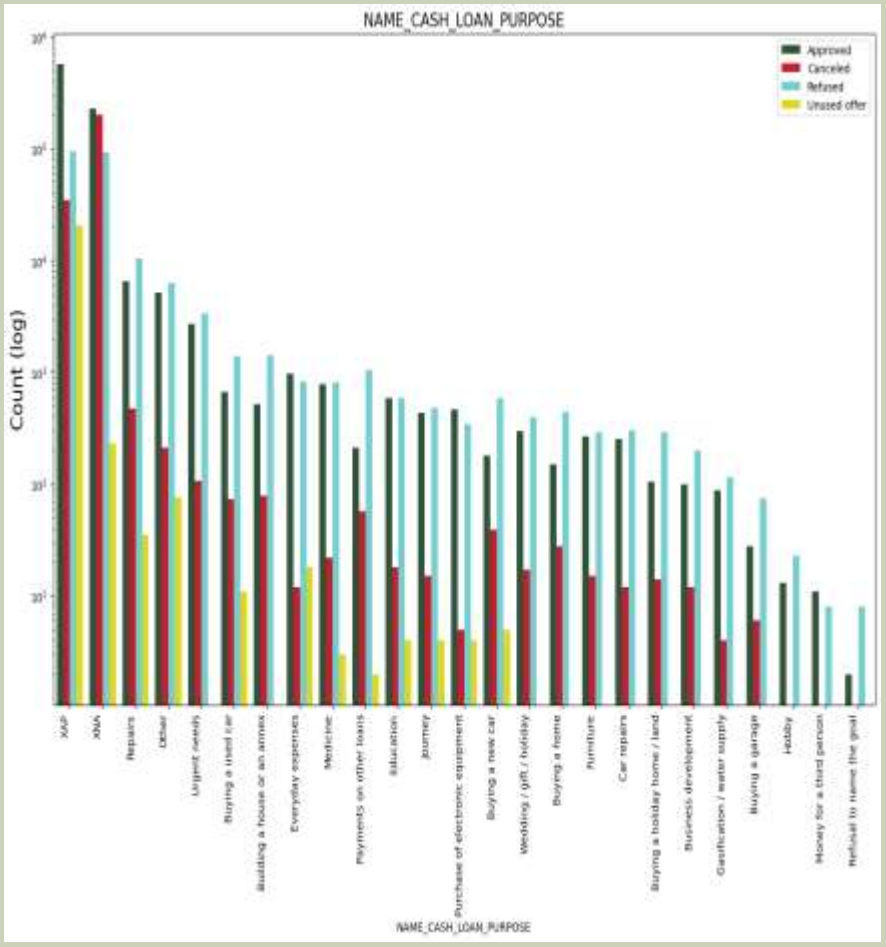


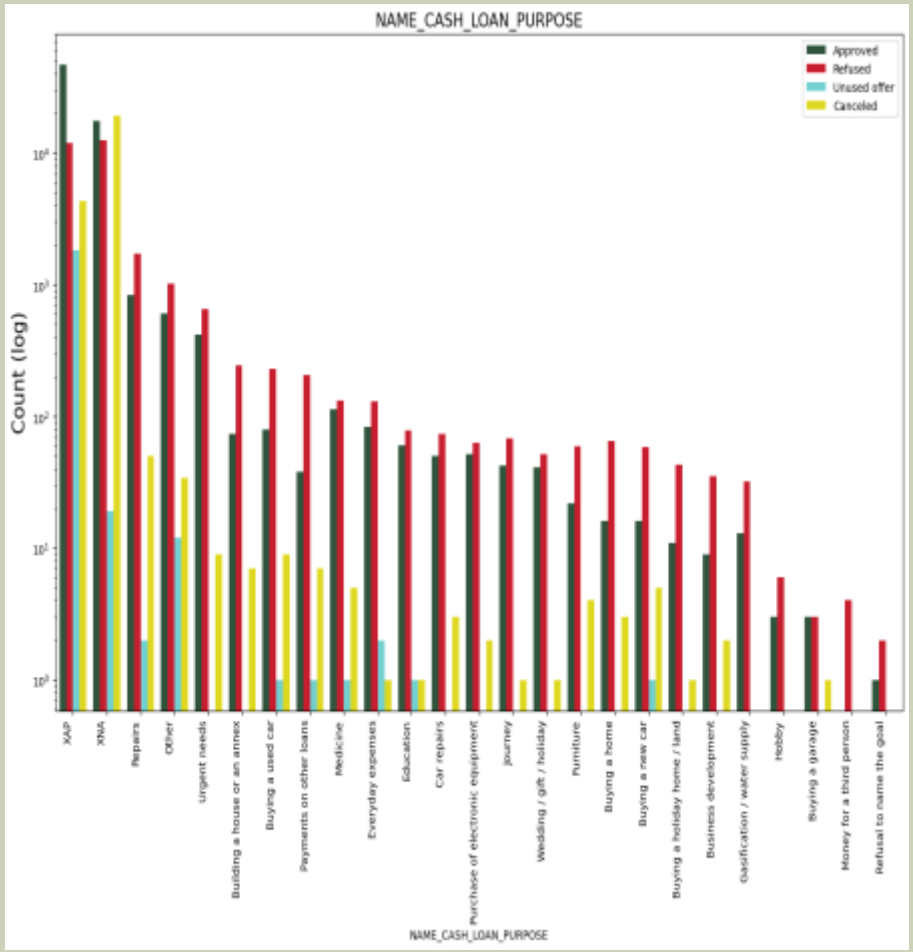


CONCLUSION OF CORRELATION AMONG DEFAULTERS AND DEFAULTERS

- Correlating factors amongst Non Defaulters and Defaulters
- **Credit amount is highly correlated with: Goods Price Amount, Loan Annuity, Total Income. We can also see that Non Defaulters have high correlation in number of days employed.**
- **Credit amount is highly correlated with good price amount which is same as Non Defaulters.**
- **Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to Non Defaulters(0.77)**
- **We can also see that Non Defaulters have high correlation in number of days employed(0.62) when compared to defaulters(0.58).**
- **There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among Non Defaulters.**
- **Days birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in Non Defaulters.**
- **There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to Non Defaulters(0.254)**

MERGED DATA FRAME ANALYSIS GRAPH





CONCLUSION OF MERGED DATA FRAME GRAPH

- Analysis of graph from merged data
- Loan purpose has high number of unknown values (XAP, XNA)
- Loan taken for the purpose of Repairs looks to have highest default rate
- Numerous applications for repairs or other requests have been turned down by banks or clients. This indicates that the bank views maintenance as high risk. Additionally, either they are turned down or the bank gives them a loan at a high interest rate that they cannot afford, so they decline the loan.

CONCLUSIONS FROM THE EDA CASE STUDY FOR CREDIT RISK ANALYSIS

- Non Defaulters Percentage is 91.93% and defaulters is 8.07%.
- Females are less likely to default to loans.
- People living in office apartments have lowest default rate.
- People with Academic degree are least likely to default.
- Most of the loans are taken by Laborers, followed by Sales staff.
- Applicants with rating 1 are less likely to default.
- The customers who owns real estate are double in numbers from those who don't known and it owning a real estate does not affect the defaulting.
- Applicants living in region rating 1 are less likely to default.
- Most of the loan defaults are for customers whose income ranges between 1100000-1500000.

CONCLUSIONS FROM THE EDA CASE STUDY FOR CREDIT RISK ANALYSIS

- Most of the loans are taken by Laborers, followed by Sales staff.
- Self employed people have relative high defaulting rate, to be safer side loan disbursement should be avoided or provide loan with higher interest rate to mitigate the risk of defaulting.
- Most of the people application for loan are from Business Entity Type 3.
- Loan taken for the purpose of Repairs looks to have highest default rate
- Huge number application have been rejected by bank or refused by client which are applied for Repair or Other. from this we can infer that repair is considered high risk by bank. Also, either they are rejected or bank offers loan on high interest rate which is not feasible by the clients and they refuse the loan.

CONCLUSIONS FROM THE EDA CASE STUDY FOR CREDIT RISK ANALYSIS

- From the EDA CASE STUDY Analysis we have drawn useful insights about Non Defaulters and defaulters which will be useful lending of loans.
- We can say that any applicant able to repay the loan should be lended the loan and the applicants not able to repay must be avoided the loan . Both, the criteria's are important and should be taken care of for better functioning and revenue of bank.