Exploratory Data Analysis

Credit EDA Case Study

Prepared by- Arundhati Deshpande

Problem Statement

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. When the company receives a loan application, the company must 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.

- When a client applies for a loan, there are four types of decisions could be taken by the client/company:
  - Approved: The Company has approved loan Application
  - 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.
  - Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
  - Unused offer: Loan has been cancelled by the client but on different stages of the process.

Business Objective

- This case study aims 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. Identification of such applicant's using EDA is the aim of this case study.

Approach to EDA

The Steps followed in this analysis are as below:

1. Data Understanding / Data Sourcing
2. Data Cleaning
3. Data Analysis
    1. Univariate - Numerical and Categorical Variables
    2. Correlation
    3. Bivariate - Numeric to Numeric, Numeric- Categorical, Categorical-Categorical
    4. Multivariate
    5. Data Visualization
4. Final Observations / Insights

Data Cleaning

- First step - Imported all the useful libraries and loaded the datasets in the python notebook. Reviewed the rows and columns, shape of datasets, and got a broad idea of various data types.
- Our second and most important step in EDA is Data Cleaning.
- Steps followed in Data cleaning are:
    1. Identifying the data types
    2. Imputing/removing missing values
    3. Handling outliers
    4. Standardizing the values
    5. Binning

After closely observing the data and the null values, it was observed that a lot of columns had missing values greater than 40%. These columns didn’t seem to have much relevance in our analysis and so were removed from our Analysis

Data Cleaning

- For columns having missing values below 10% - corresponding rows were dropped as these wouldn’t affect our analysis.
- Columns having missing values >10% and <40% are treated separately.
  - For Categorical missing variables – imputed missing values with Mode values
  - Column like ‘OCCUPATION_TYPE’ which had 31% missing values – left these as ‘missing’
- Next crucial step in Data cleaning is looking for outliers. For outliers checking, we plot a boxplot and based on the data in hand and the type of outlier, we either impute them, drop them or cap them or create binning.
- In this case, Numerical variables have outliers that are continuous, so we did not drop them.
- Next step is to standardize the columns for better analysis followed by binning of some variables like Age and Income

Data Analysis

Analysis of the datasets is done one at a time.

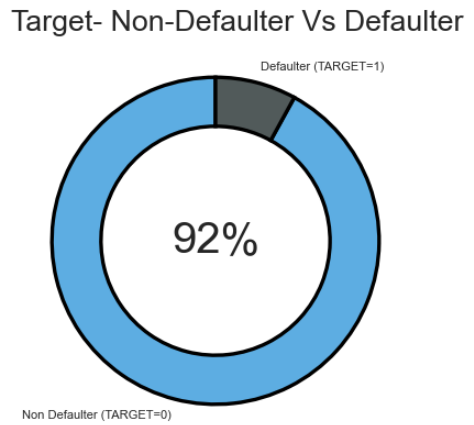
- First, we checked the imbalance % of this dataset. It is observed that about 92% of the applicants are Non-Defaulters and 8% are Defaulters.
- We divided application_data dataset into two datasets based on Target variable:
  - Target-0 : Non-Defaulters
  - Target-1: Defaulters
- Analysis started with Univariate analysis on numerical and categorical variables followed by Bivariate Analysis and Multivariate Analysis in relation to Target variables.
- Univariate concentrates on one variable at a time and Bivariate uses 2 variables. In our case we used one variable as Target variable and the other as we found relevant.
- The common tools used for this were value_counts function, groupby functions, mean, median, quantiles, pivot tables, barplots, countplots, box plots, pieplots, etc.
- Similar procedure is carried out in 2nd dataset (previous_applications)
- Finally, merged these two datasets and some made some observations.

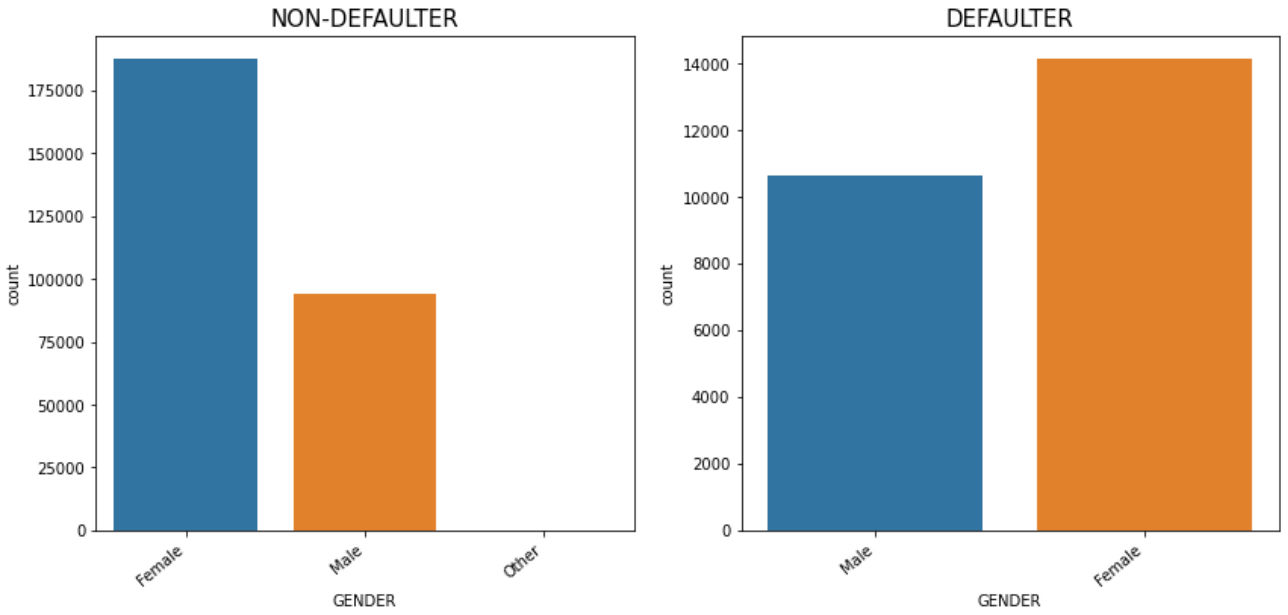
Data Analysis: ‘application_data’ dataset

- Data Imbalance – 92% Non Defaulters

Some General Observations from Univariate Analysis:

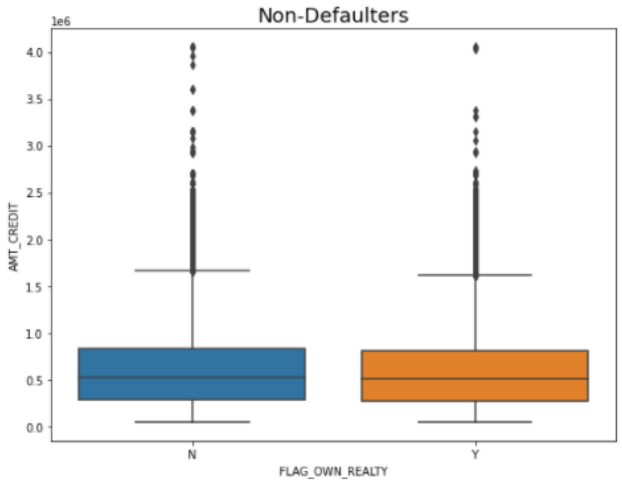
- Maximum number of applicants are Females.
- Applicants are mostly from Tier 2 regions
- Maximum number of applicants have a family size of 2-4 people.
- Credit Amount and Goods Price have Similar trend
- Most of the applicants do not own car.

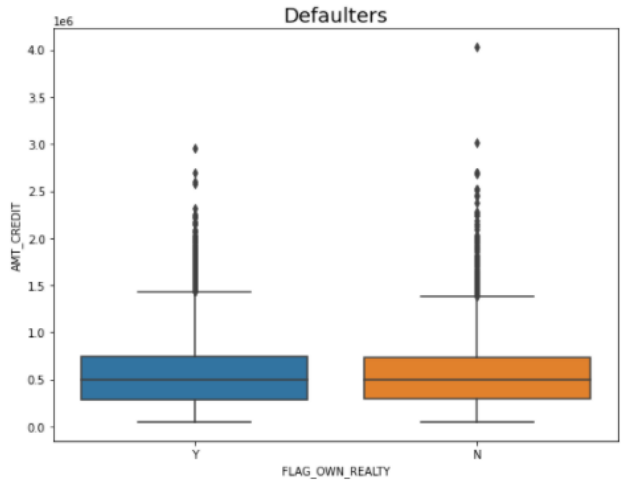


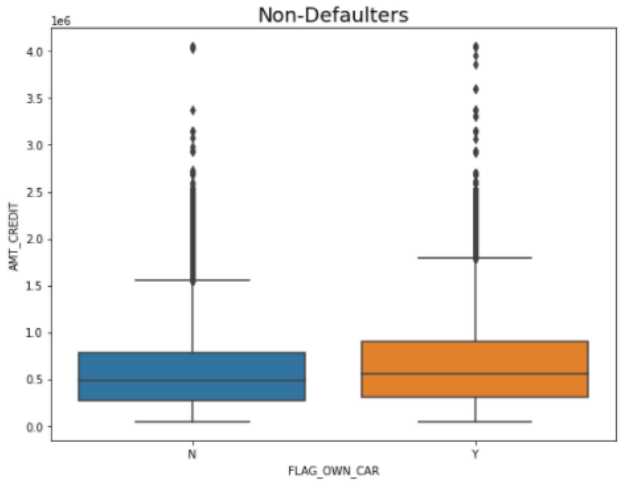


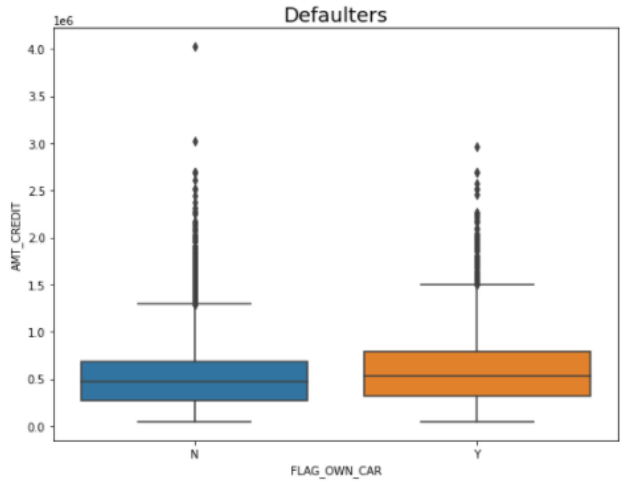
Data Analysis: ‘application_data’ dataset

- Owning car and house does not affect anything. Most of them do not own them



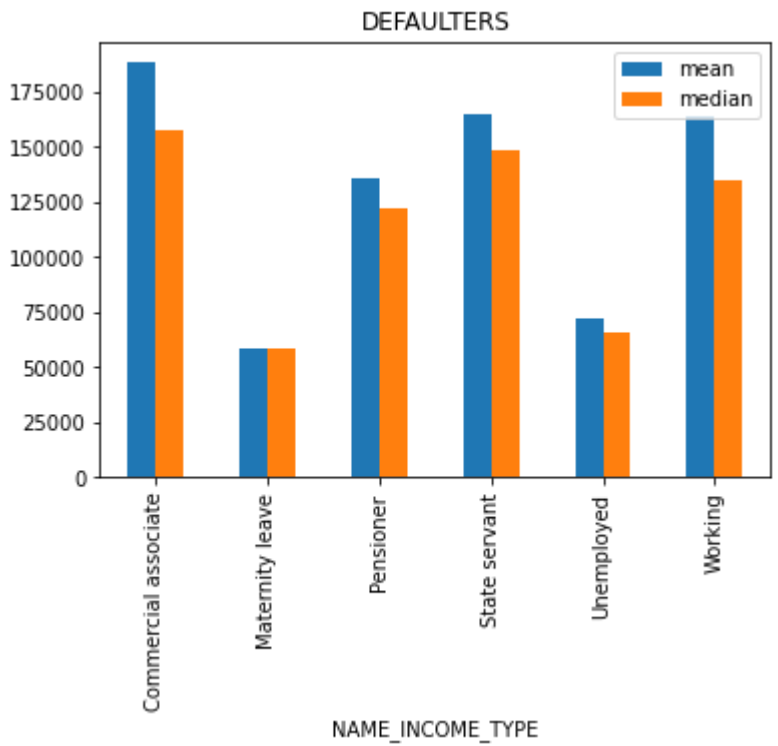


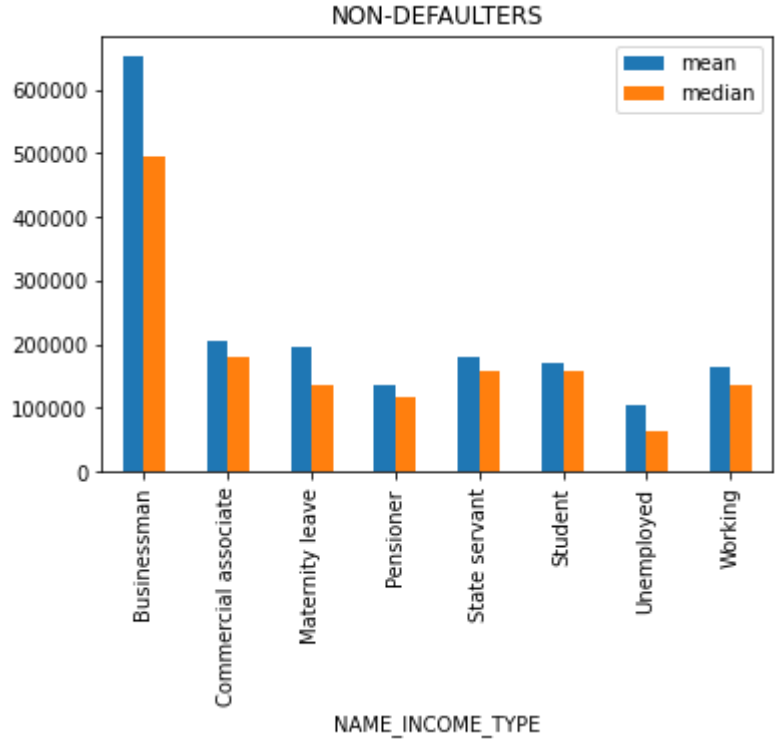




Data Analysis: ‘application_data’ dataset

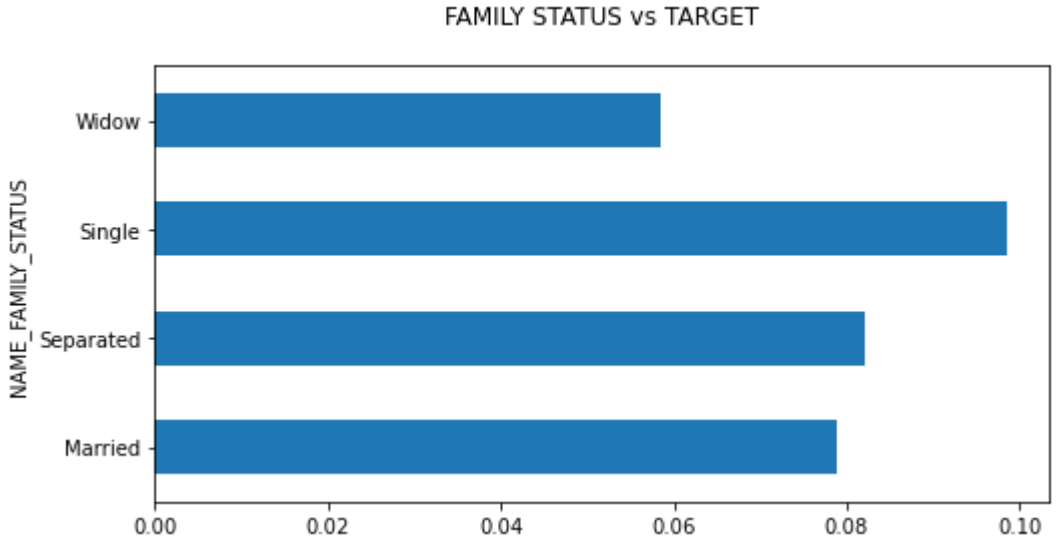
- Based on Income types we can say that Commercial associates are more likely to be defaulters and businessman are more likely to give business as these are less defaulters

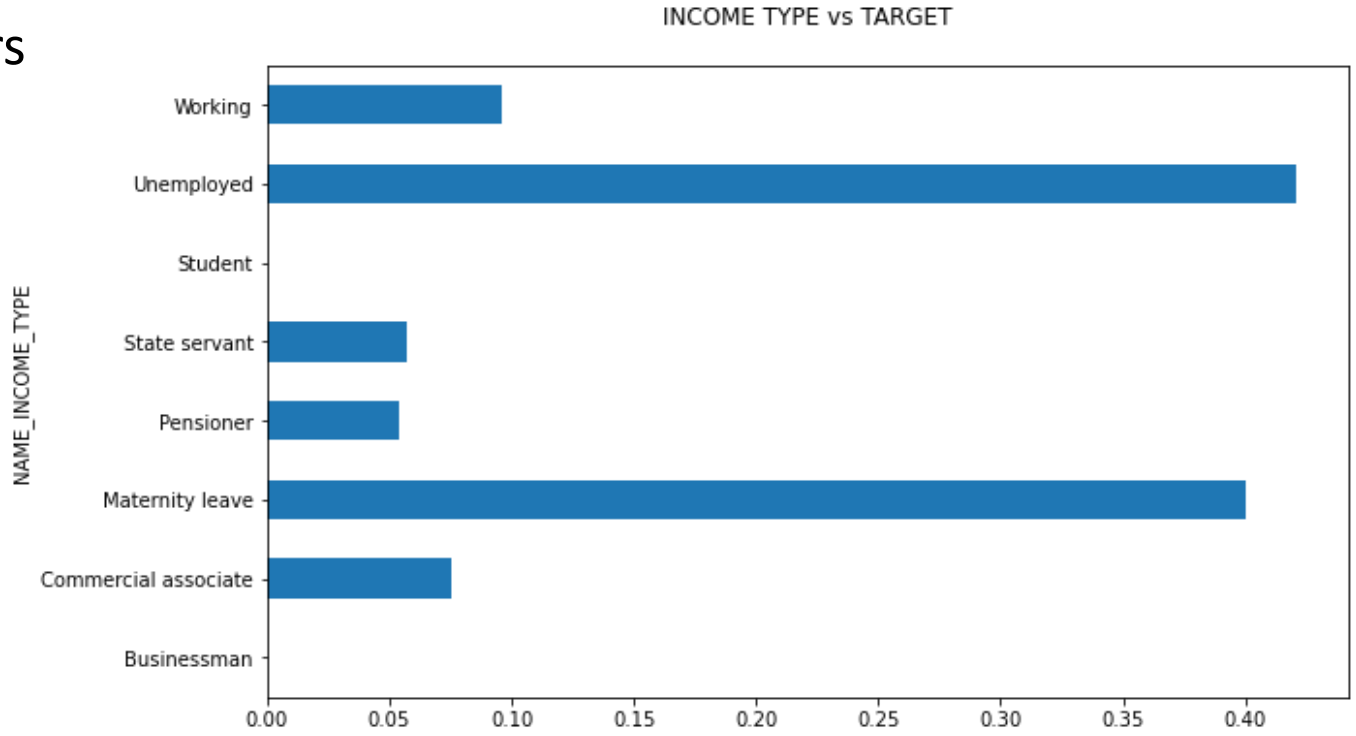




Data Analysis: ‘application_data’ dataset

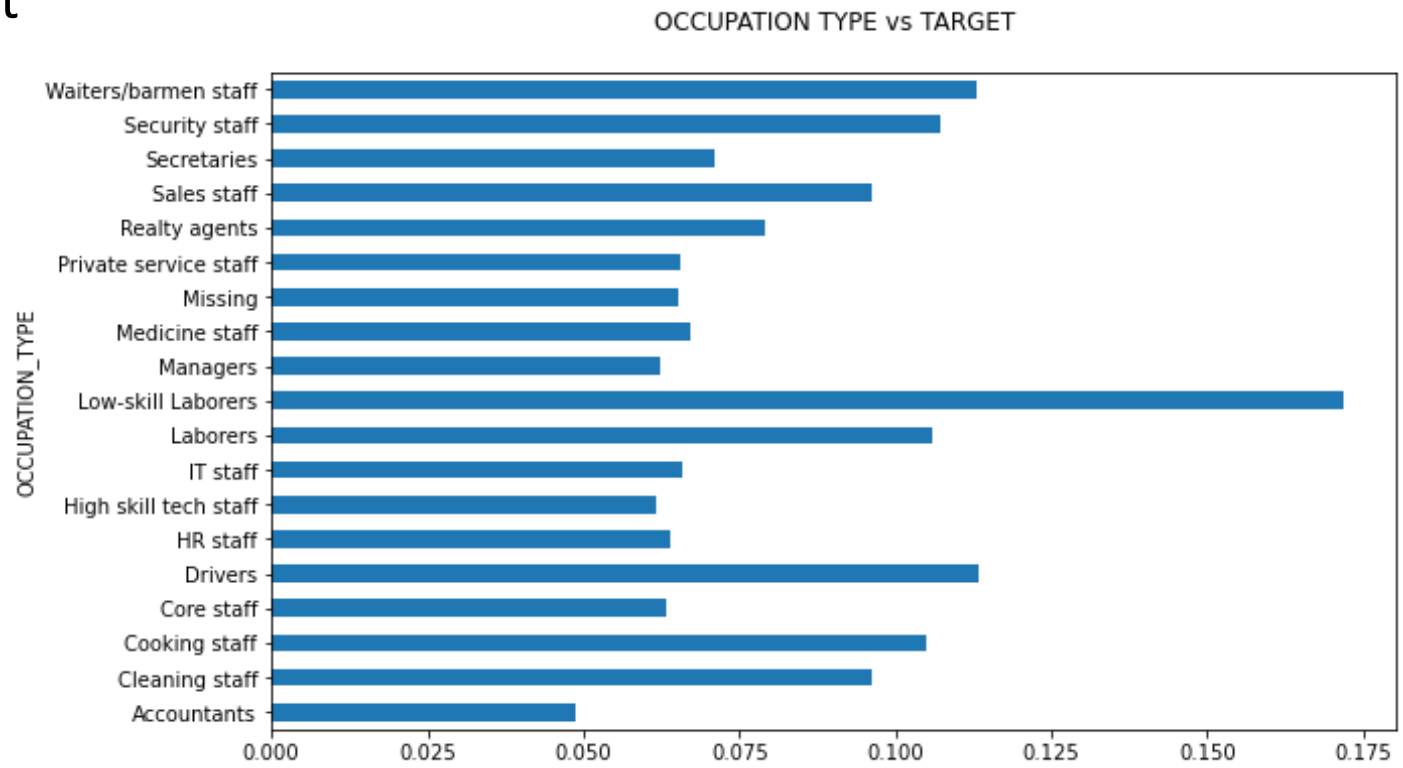
- Applicants who are unemployed and are on maternity leave are more likely to be defaulters. These are high risk applicants
- Applicants who are single are high risk as they are more likely to be defaulters
- Separated applicants also are defaulters





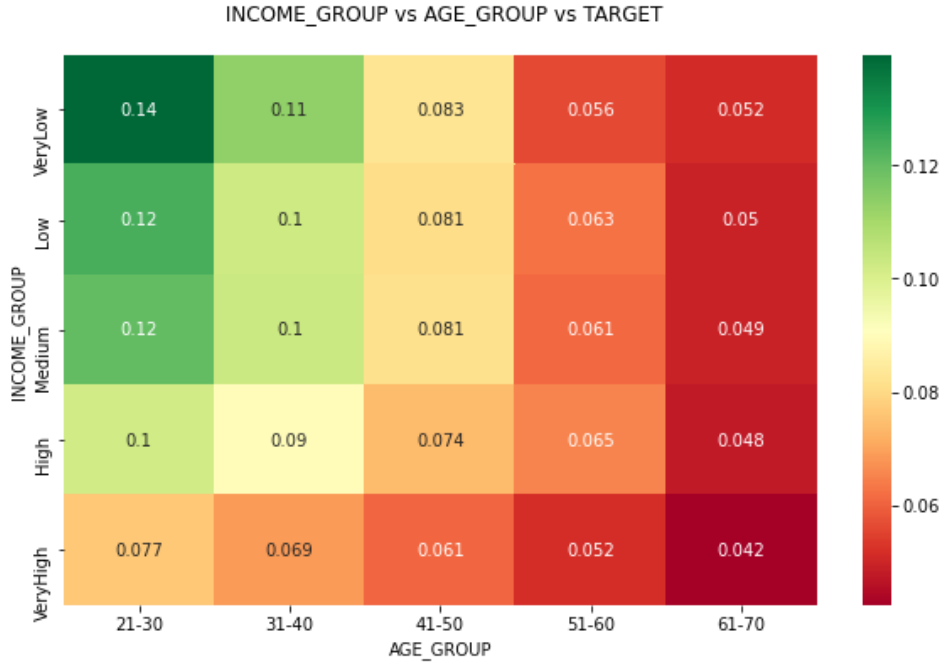
Data Analysis: ‘application_data’ dataset

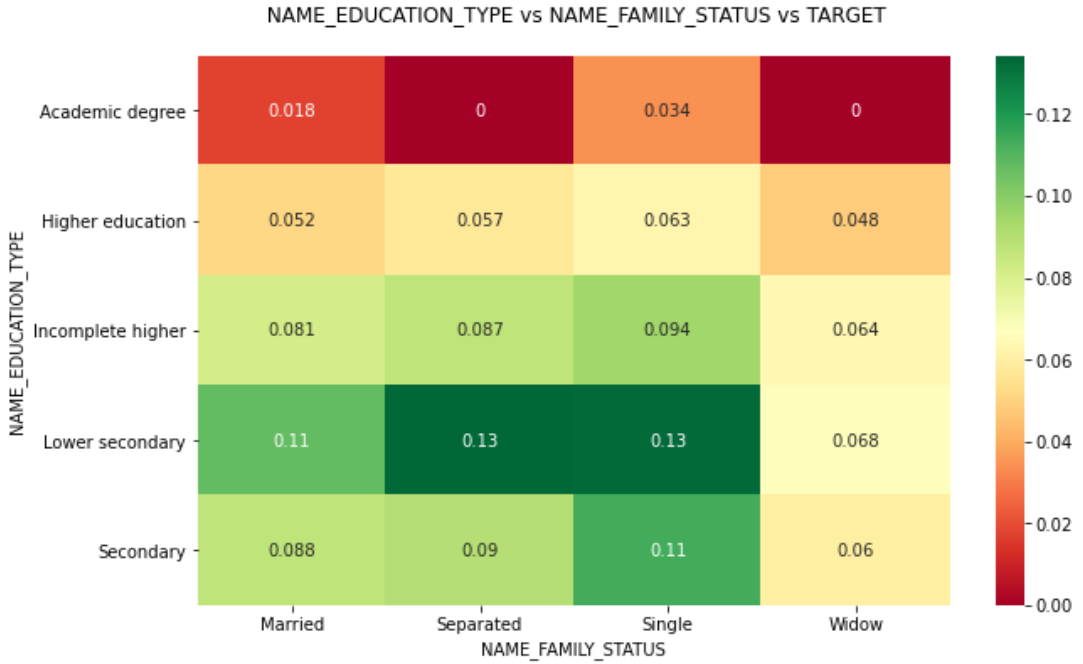
- Low skilled labourers are more likely to be defaulters. These are high risk applicants.
- Drivers, Laborers, cleaning staff, cooking staff, waiters, security staff - These are the categories who have very less income and that might be the reason that they are more likely to default



Data Analysis: ‘application_data’ dataset

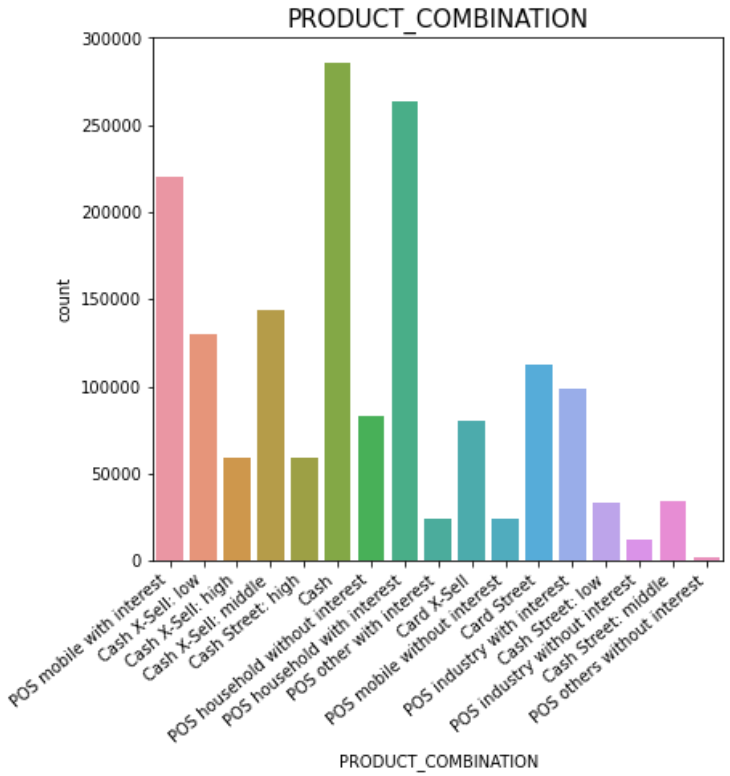
- Applicants with lower secondary education, secondary education and who are single and/or separated are more likely to be defaulters.
- Applicants having Academic degree are having very less chance of defaulting
- Young Applicants (age between 21-30) having low income are more likely to be defaulters.
- Applicants having age between 31-40 with very low income are also most likely to be defaulters

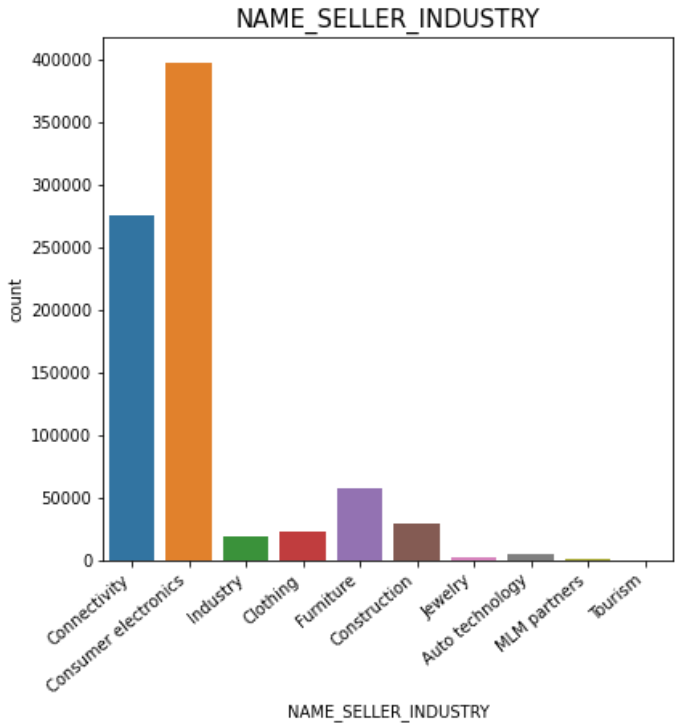




Data Analysis: ‘previous_application’ dataset

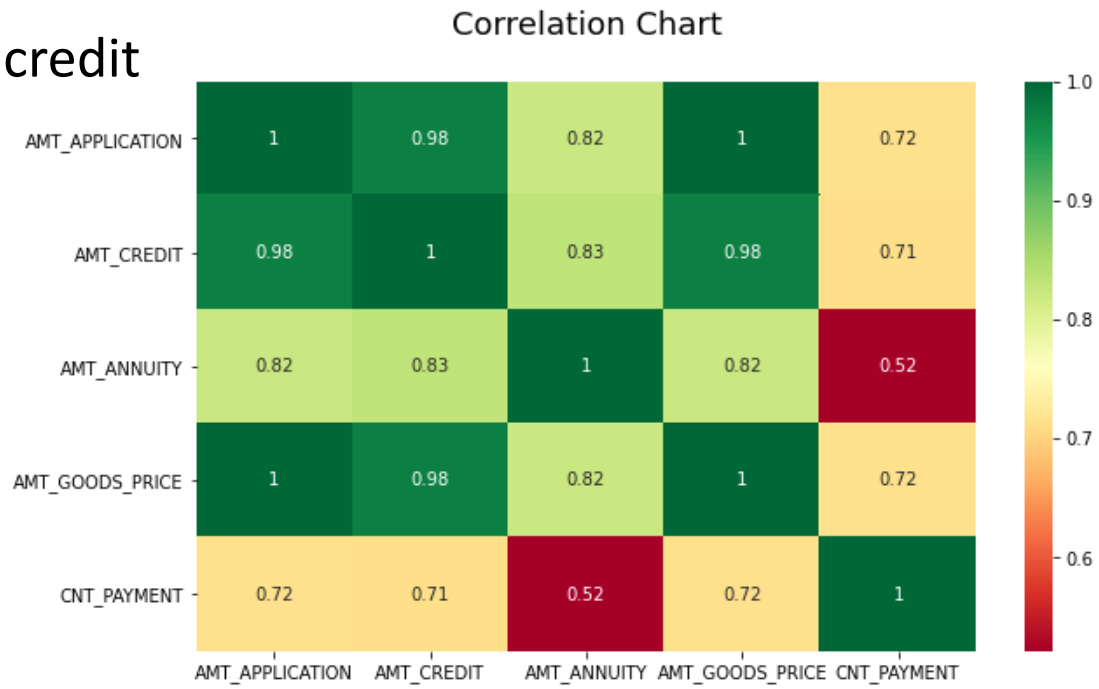
- Cash loans are more mostly preferred.
- Loans were asked mostly for consumer electronics, computers, audio/video
- A lot of them are also for Furnitures etc.,





Data Analysis: ‘previous_application’ dataset

- We can see the strong correlation between all of these parameters. All these variables are interlinked.
- Most of the clients received the credit amount as per they requested in the application
- Credit amount depends on the goods price
- Credit amount also depends on Term of previous credit



Data Analysis: ‘previous_application’ dataset

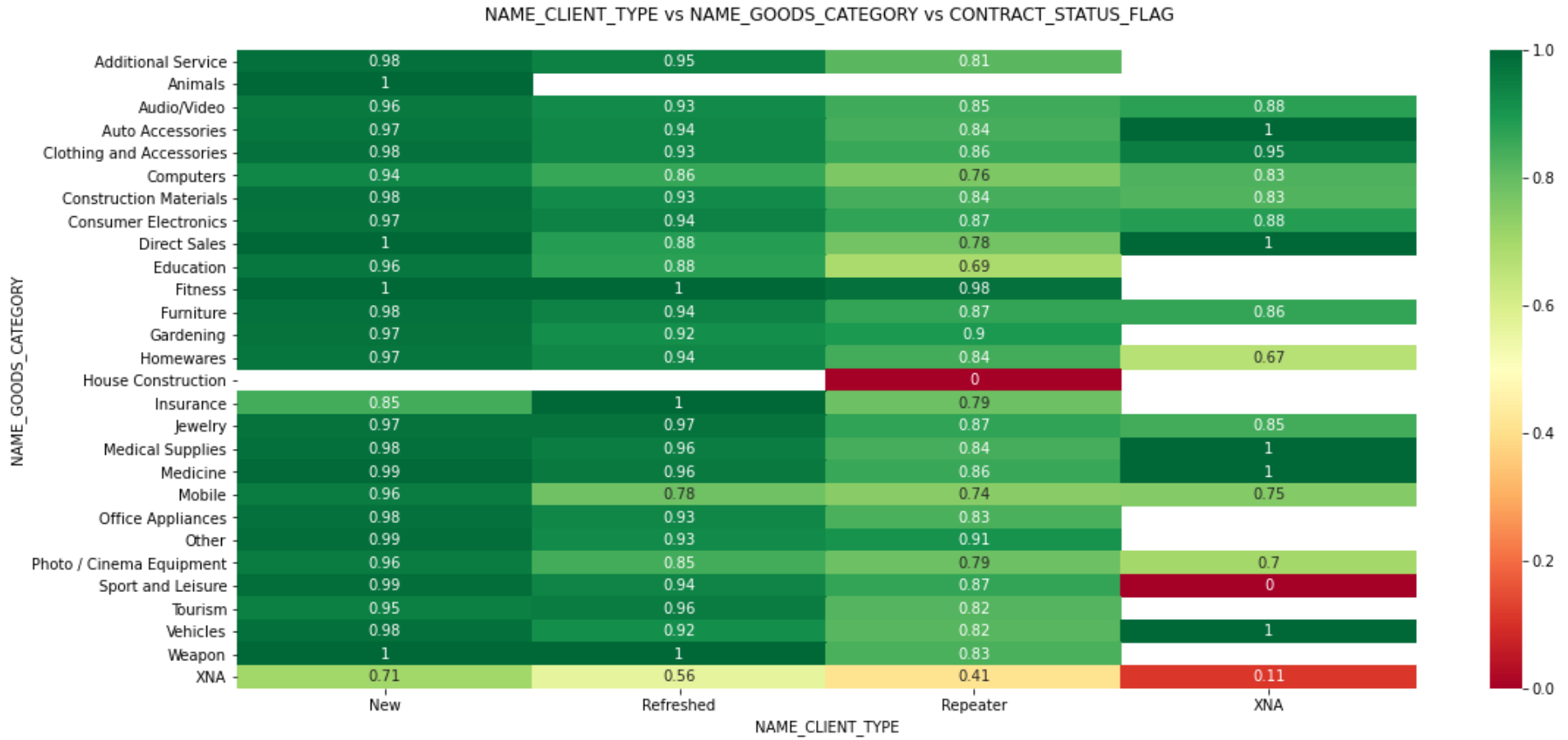
- When we look at the average amounts, we can say that
- Credit Amount is more for Tourism and MLM Partners
- Next inline are Sellers from clothing, Industry, Construction and Furniture etc
- Repeaters are asking for higher credit amount than new applicants

Data Analysis: ‘previous_application’ dataset

- Most of the loans approved were for new applicants and most of them were for Animals and Fitness. This heatmap gives the correlation between them and we can say most of the new loans are for Animals and Fitness
- We can see loan for fitness is also high by repeaters and refreshed clients
- This also concludes that approval of loans are given is fairly given in all Goods categories.
- Applicants who are repeaters have less likely chance to get the loan approved. This is the case with almost every category on CASH_LOAN_PURPOSE. This would be a loss on business
- New applications are getting preference on cash loans

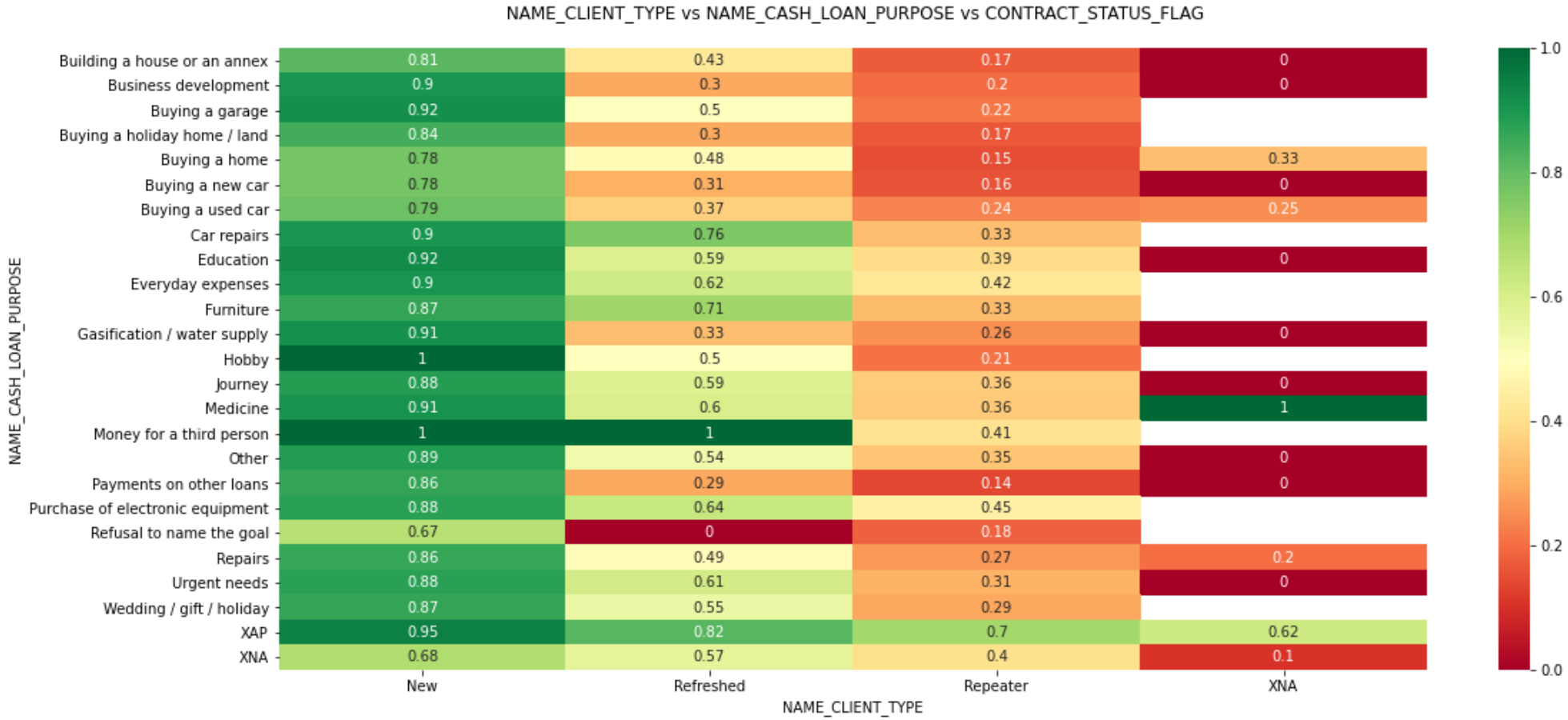
Data Analysis: ‘previous_application’ dataset

- Chart supporting for inferences made in previous slides



Data Analysis: ‘previous_application’ dataset

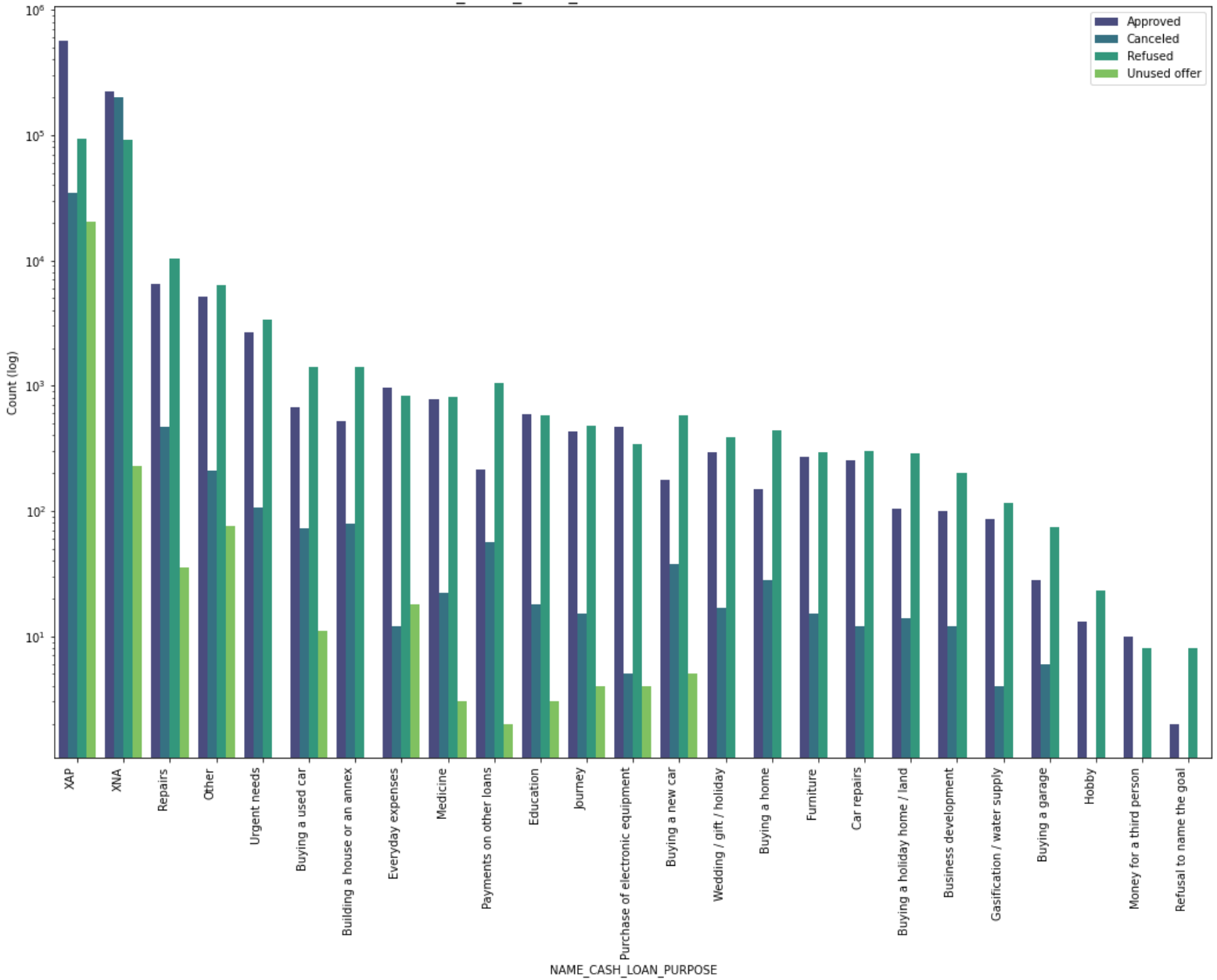
- Chart supporting for inferences made in previous slides

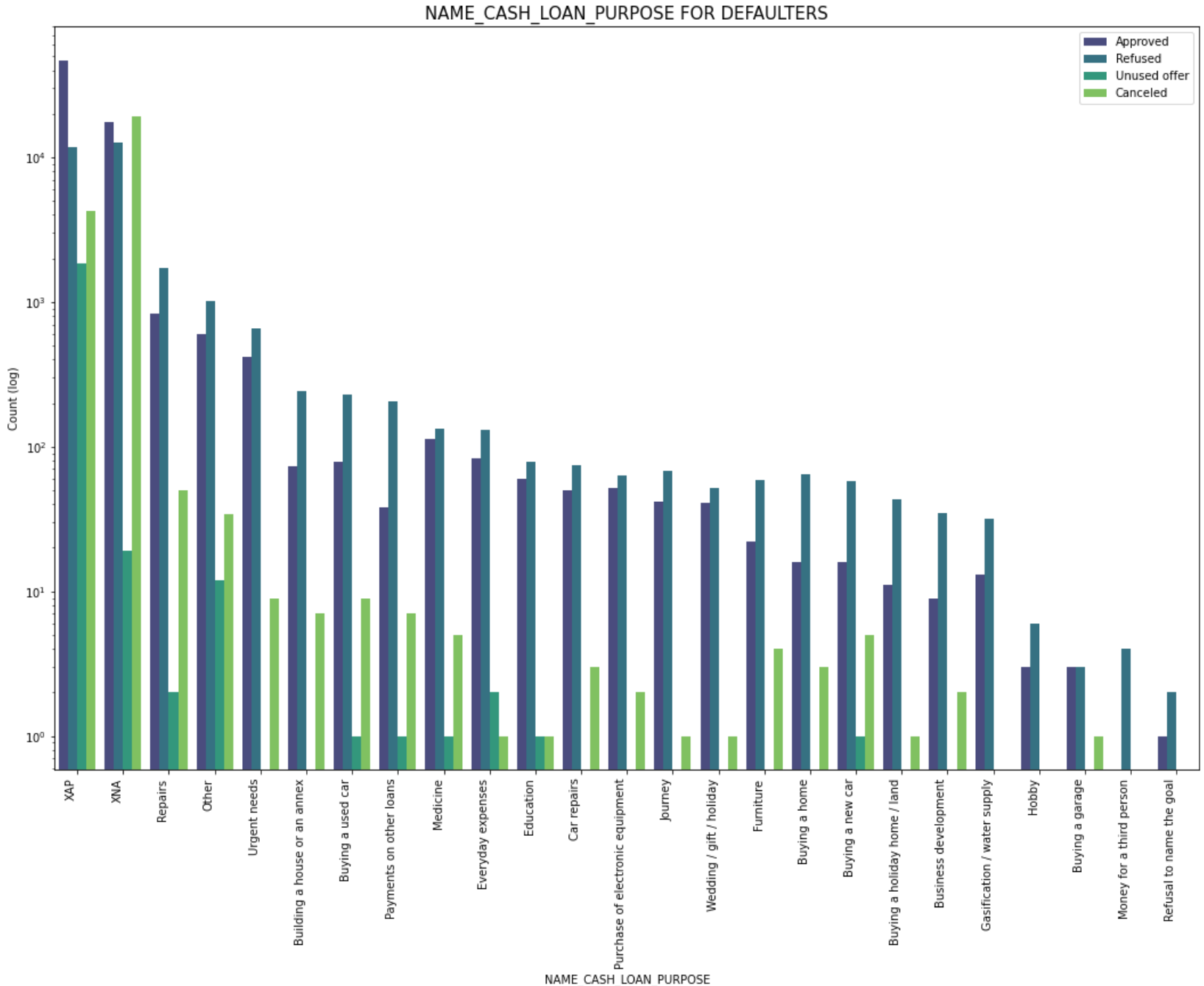


Data Analysis: After Merging the datasets

So far, we have analyzed the data from both datasets separately. Now see what more insights we can draw after merging them.

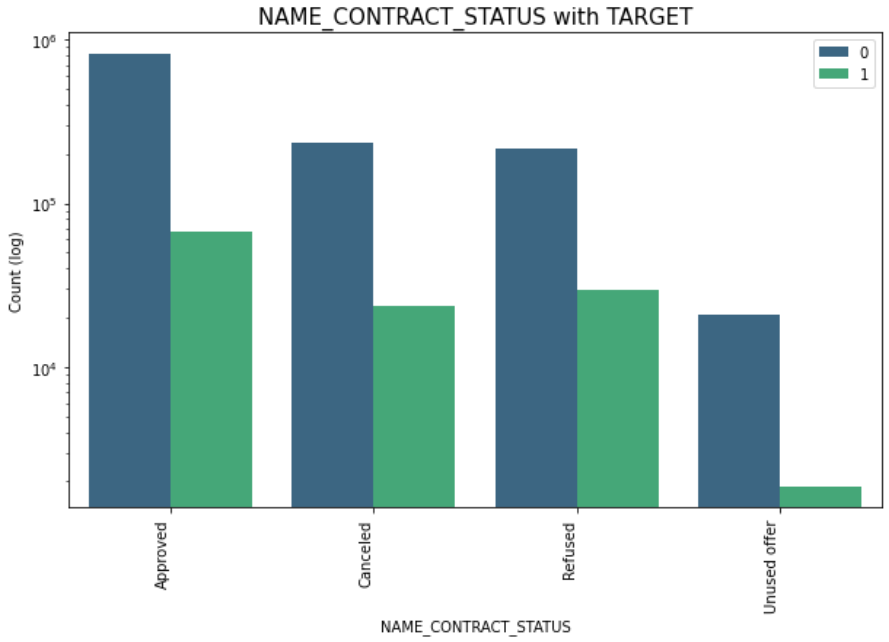
- We have plotted graphs showing Trend about the Non-defaulters and Defaulters separately against their previous applications.
- It seems that majority of the loans are requested for repairs and urgent needs
- The people who are Non-defaulters and pay the loans on time, have been refused to give loans by bank. This may affect the business.
- The dataset contains lot of missing information. Most of them are XAP and XNA. Bank should collect more relevant information about these clients

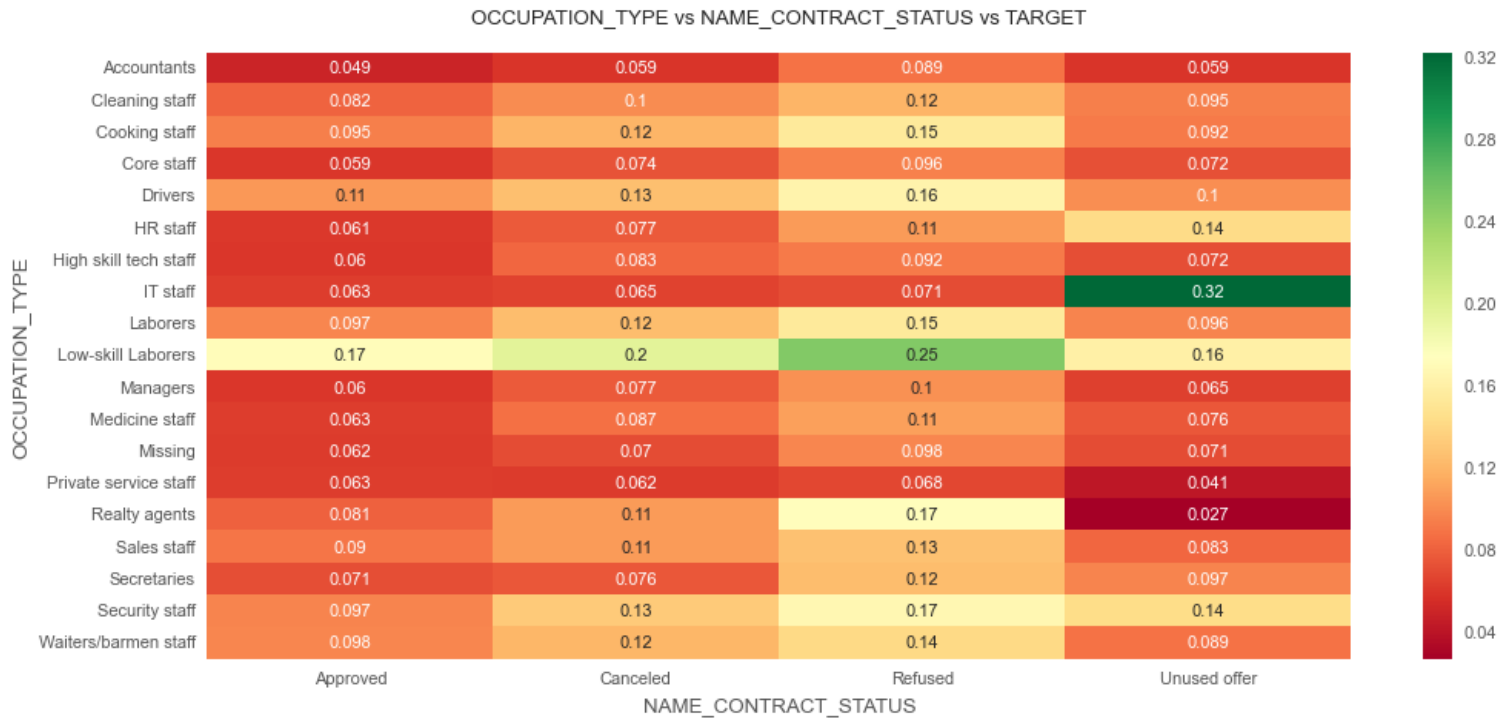

NAME_CASH_LOAN_PURPOSE FOR NON-DEFAULTERS
Approved
Canceled
Refused
Unused offer
Count (log)
10^6
10^5
10^4
10^3
10^2
10^1
XAP
XNA
Repairs
Other
Urgent needs
Buying a used car
Building a house or an annex
Everyday expenses
Medicine
Payments on other loans
Education
Journey
Purchase of electronic equipment
Buying a new car
Wedding / gift / holiday
Buying a home
Furniture
Car repairs
Buying a holiday home / land
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
NAME_CASH_LOAN_PURPOSE


NAME_CASH_LOAN_PURPOSE FOR DEFAULTERS
Approved
Refused
Unused offer
Canceled
Count (log)
10^4
10^3
10^2
10^1
10^0
XAP
XNA
Repairs
Other
Urgent needs
Building a house or an annex
Buying a used car
Payments on other loans
Medicine
Everyday expenses
Education
Car repairs
Purchase of electronic equipment
Journey
Wedding / gift / holiday
Furniture
Buying a home
Buying a new car
Buying a holiday home / land
Business development
Gasification / water supply
Hobby
Buying a garage
Money for a third person
Refusal to name the goal
NAME_CASH_LOAN_PURPOSE

Data Analysis: After Merging the datasets

- People who were approved of loans earlier defaulted less.
- People who were refused of loans defaulted more
- loan refused by bank are mostly for low skilled laborers
- IT staff have more unused offers





Final Observations / Insights

Inferences and recommendations from the complete Analysis:

- The bank should categorize every new client into 2 categories: high risk or low risk based on whether that person's probability of default.
- There were a lot of missing information in the data and bank should collect all relevant information to infer more accurately
- The people who are Non-defaulters and pay the loans on time, have been refused to give loans by bank. This may affect the business.
- Applicants who are unemployed and are on maternity leave are more likely to be defaulters. These are high risk applicants
- Applicants who are single are high risk as they are more likely to be defaulters