# Credit EDA Assignment

By- Debanik Kundu (Batch-C55)

# Problem Statement - I

- **Introduction**: This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.
- **Business Understanding**: The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyzed the patterns present in the data. This will ensure that the applicants 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 data given below 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):
- **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 at different stages of the process.
- In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

- **Business Objectives:** This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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 applicants using EDA is the aim of this case study. In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.  The company can utilize this knowledge for its portfolio and risk assessment. To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.
- **Data Understanding:** 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 on 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.

# Problem Statement -II

- **Results Expected by Learners**

Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.

Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

**Hint:** Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.

Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points. Identify if there is data imbalance in the data. Find the ratio of data imbalance.

**Hint:** How will you analyze the data in case of data imbalance? You can plot more than one type of plot to analyze the different aspects due to data imbalance. For example, you can choose your own scale for the graphs, i.e. one can plot in terms of percentage or absolute value. Do this analysis for the **'Target variable'** in the dataset ( **clients with payment difficulties** and **all other cases**). Use a mix of univariate and bivariate analysis etc.

**Hint:** Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights.

Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

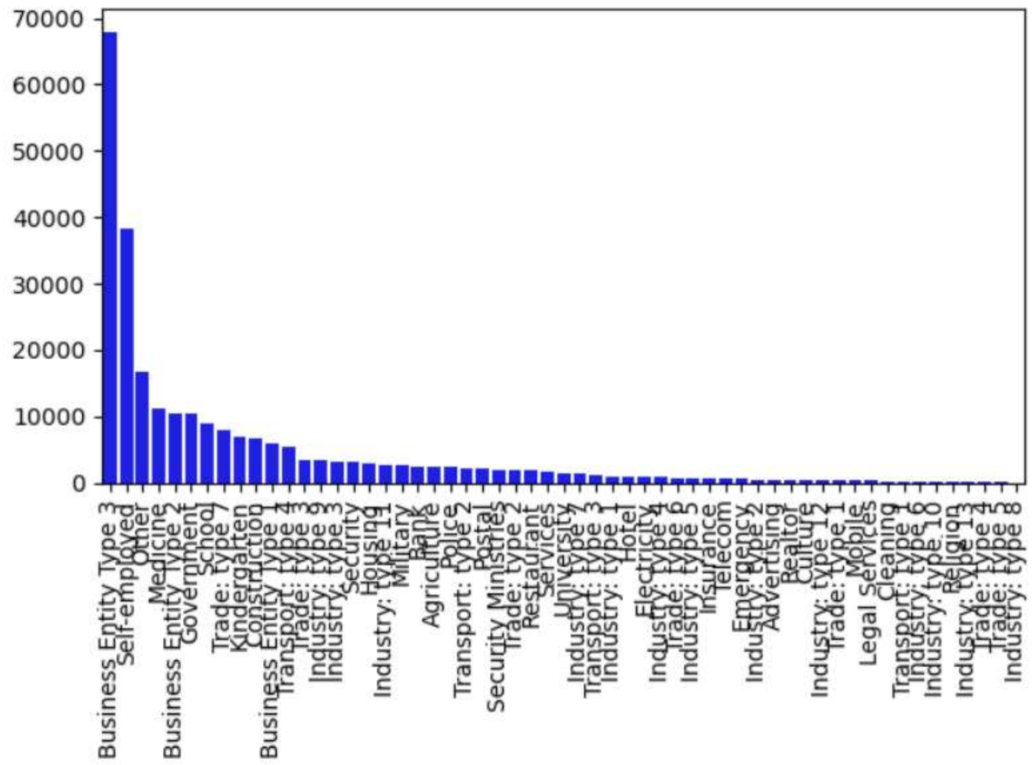
Find the top 10 correlation for the **Client with payment difficulties** and **all other cases** (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there.  Say, there are 5+1(target) variables in a dataset: **Var1, Var2, Var3, Var4, Var5, Target**. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.

Include visualizations and summaries the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the **clients with payment difficulties with all other cases.**

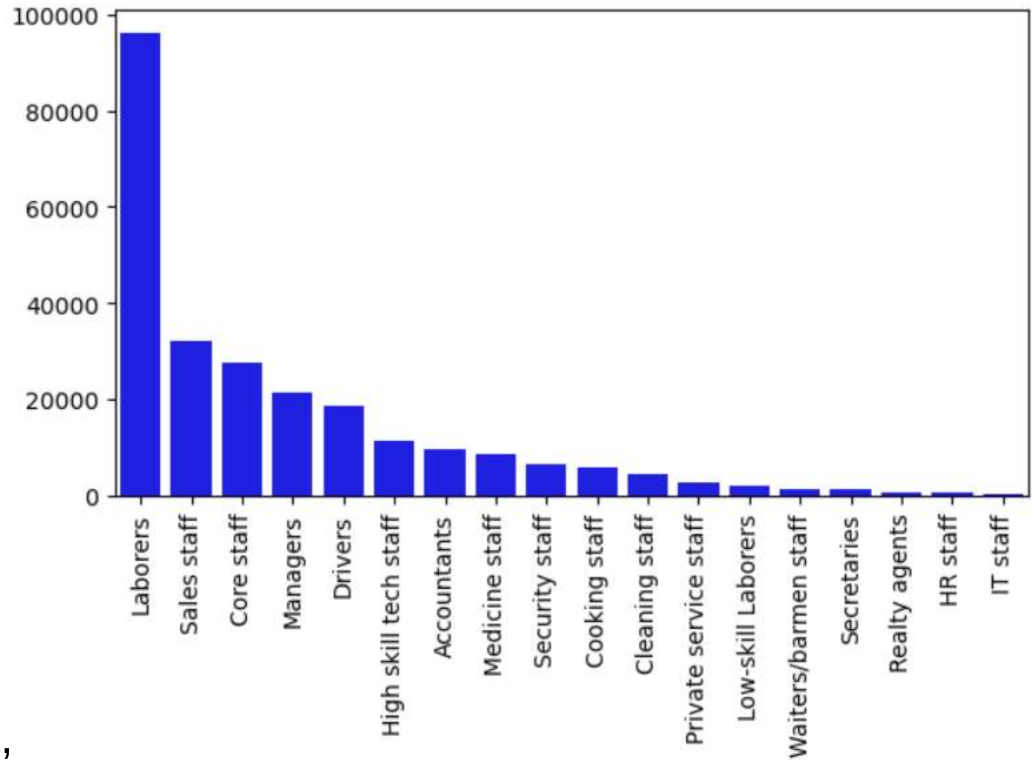
# Plot Explanation and steps taken to process over the data.

# Univariant Analysis

- “ORGANIZATION_TYPE” column is one of the categorical column which represents the type of organization the most number of loan requests come from.
- From the graph its completely understandable that employee “Business Entity Type 3” are reached out by banks for loan.
- Least reached out are from “Industry: type 8” organization.



- “OCCUPATION_TYPE” column is one of the categorical column which represents the type of occupation the loan requestee is from.
- From the graph its completely understandable that employee “Laboure's” are reached most for loan.
- Least reached for loan are from “IT Staff” occupation.
- Surely, banks reached out to people who does daily wage works.

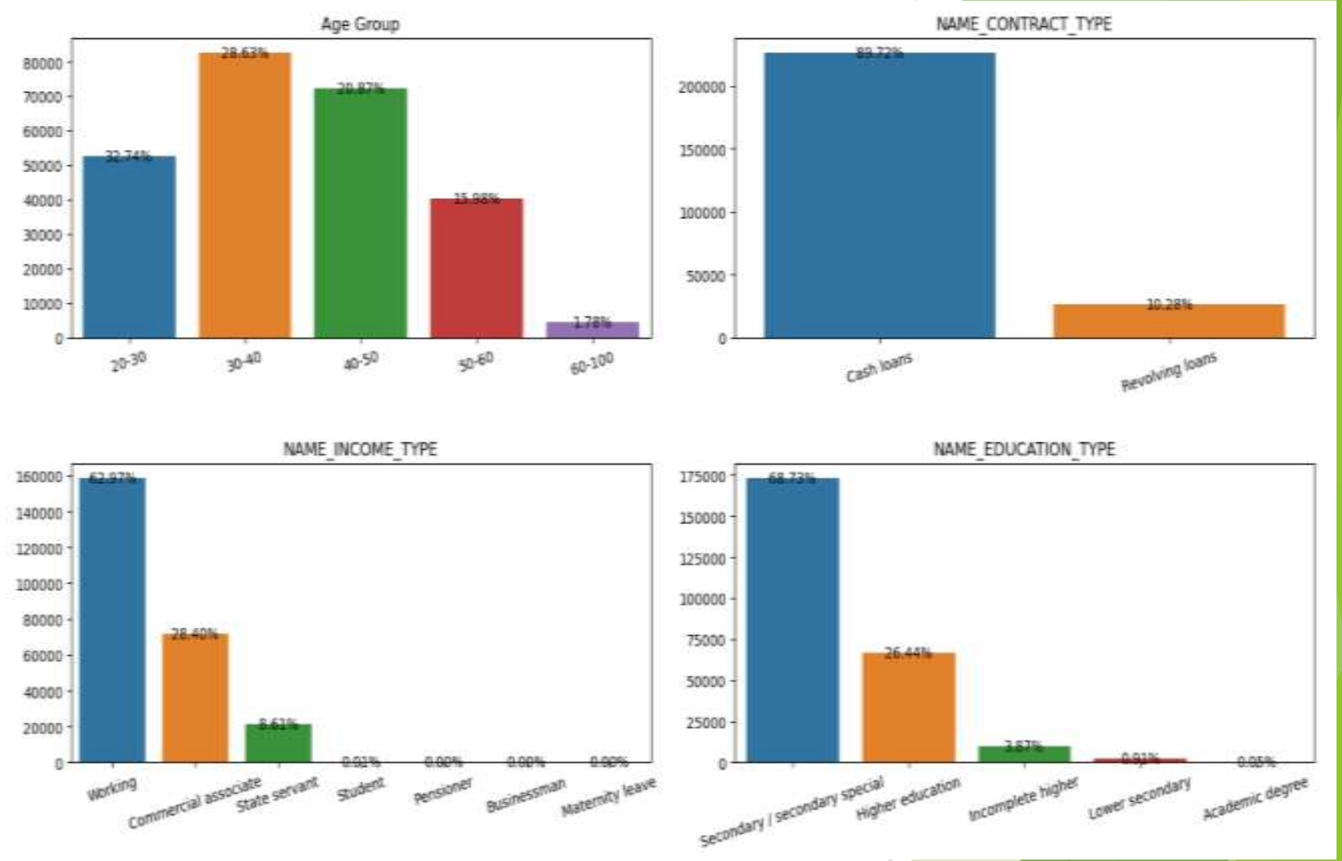


From this 2 graphs we can get an overview that,

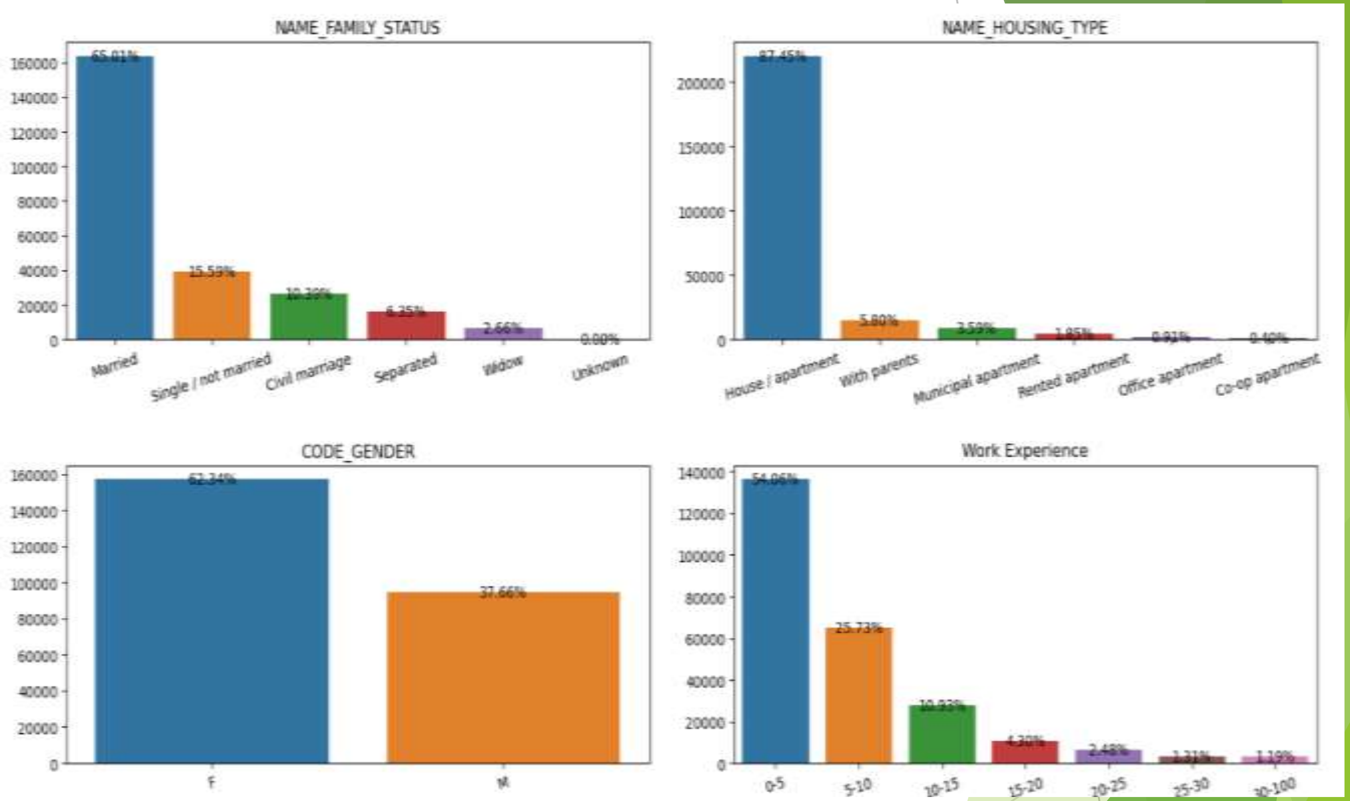
1. Most People who were reached  for Loan are Laborers and they can possibly from “Business Entity Type 3” organization.
2. Least People who reached out for Loan are IT Staffs and they can possibly from “Industry: type 8” organization.

- The following insights are,

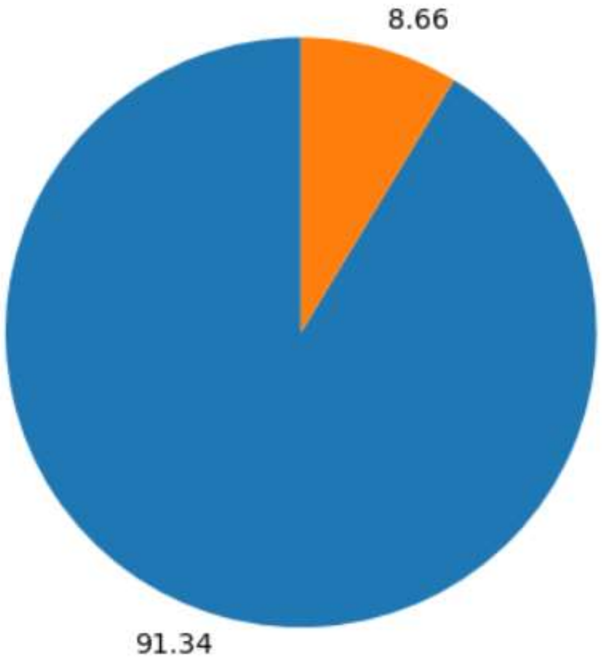
- Bank have reached majority for loan from 30-40(28.63%) & 40-50(20.87%) Age groups.
- 62.97% of clients who were reached out to opt for loan belong to Working Income Type.
- 68.73% clients with Secondary/Secondary Special education type have been reached out most.
- 89.72% loans are for Cash loans.



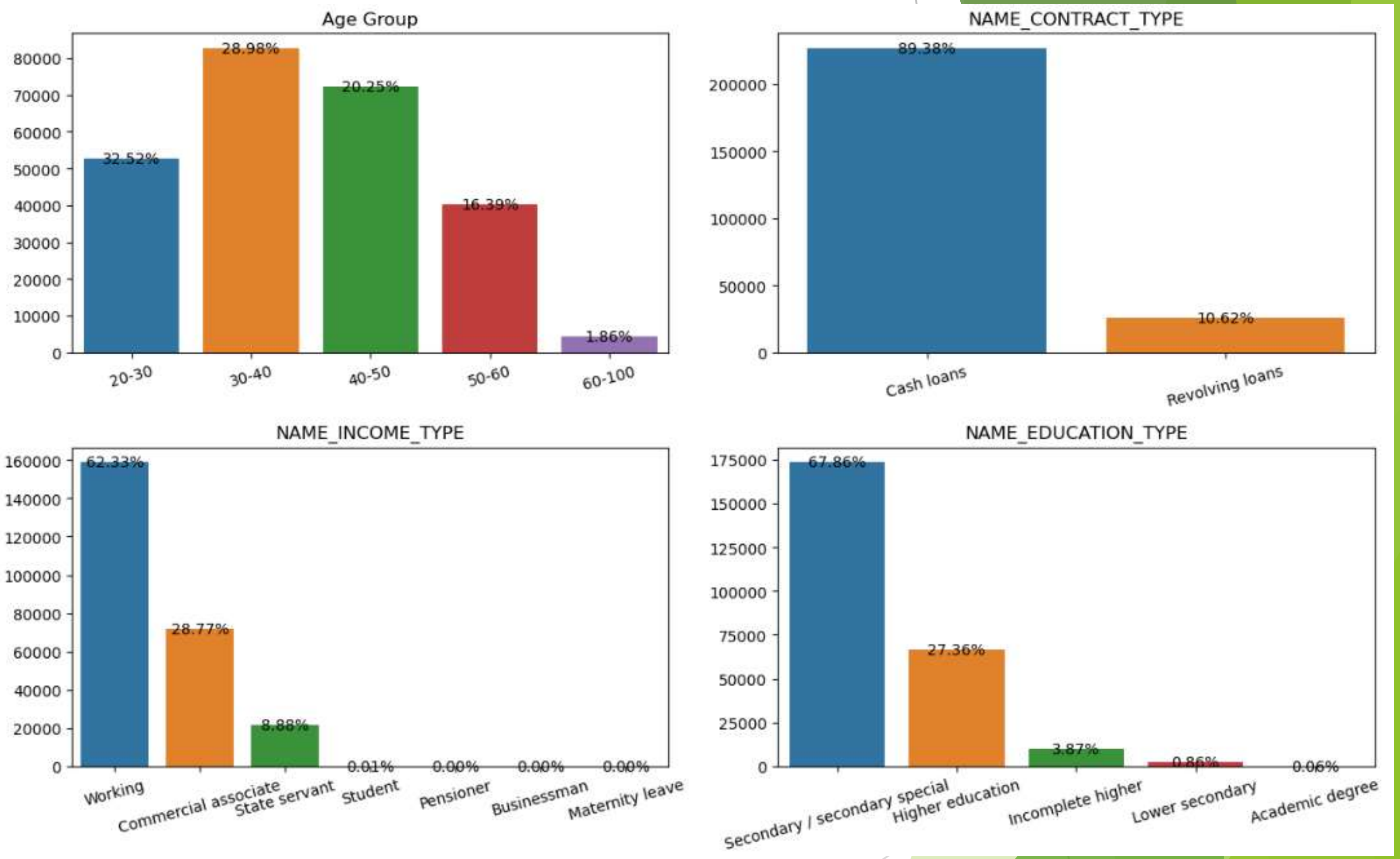
- Married people(65.01%) tend to apply more for loans.
- Majority of the Clients who have applied for the loan have their own house/apartment. Around 87.45% clients are owning either a house or an apartment.
- Female clients (62.34%) are more as compared to males. This may be because banks charge less rate of interest for -females.
- Clients with work experience between 0-5 years have applied most(54.06%) are reached out regarding loan.



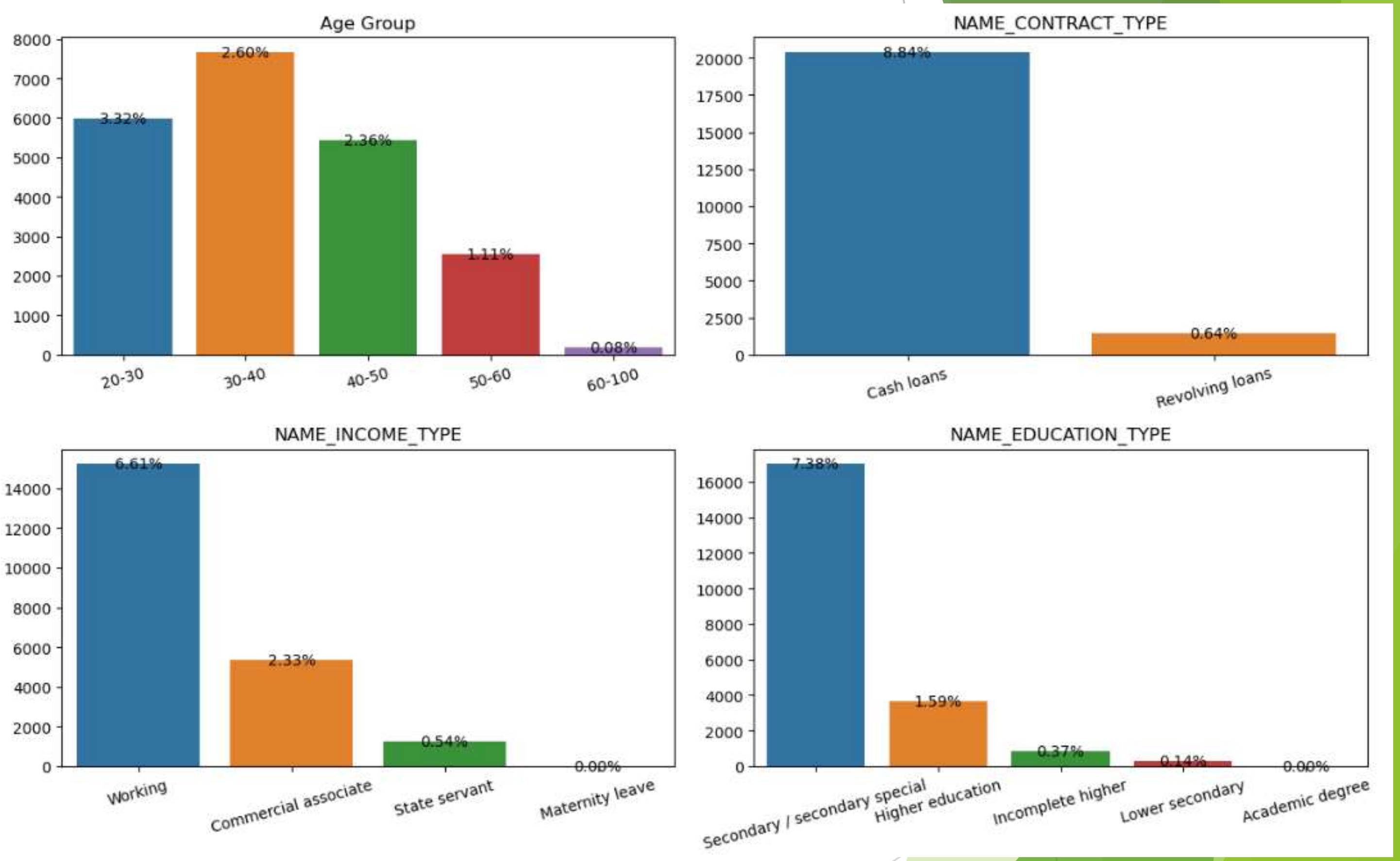
- This is a pie plot of the “Target” column which indicates that weather they have payment and non-payment issues.
- Clearly we can see that, “YES” or “1” response is very less (8.66%), compared to “NO” or “0” response which is very high (91.34%).
- Considering this we can say there are very less people who have payment related issue, so bank should consider looking for some scheme to attend those non-payment issue in there offers



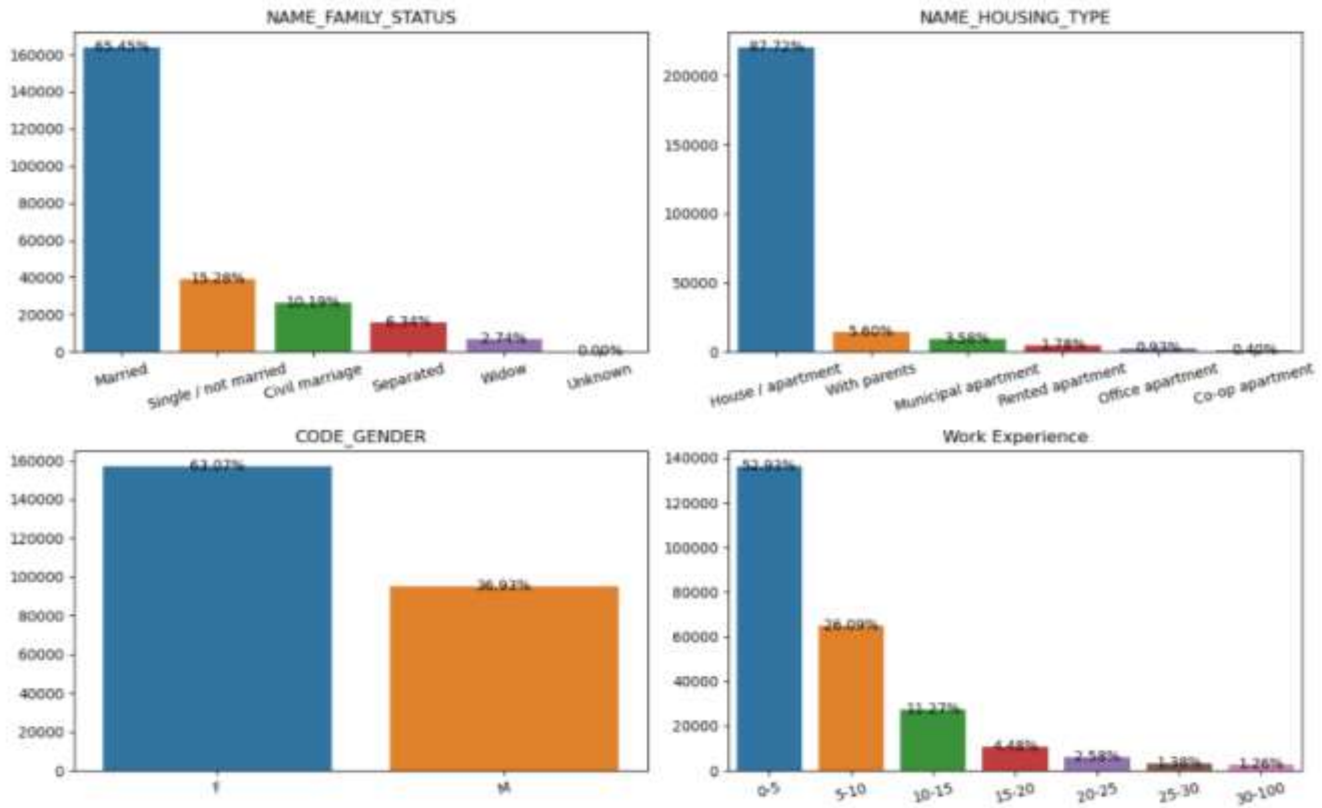
- In this plot we visualized the different attributes based on non-payment issue.
- The highest issue,
  - In terms of age group is “30-40”, which comprises of 28.98% of total average 91.34%.
  - In terms of contract it is for “Cash loans” which is about 89.38 % of total average.
  - For income types it is visible from “Working” class which is about 62.33%.
  - The people who have education till Secondary/Secondary special type of education i.e. about 67.86%.



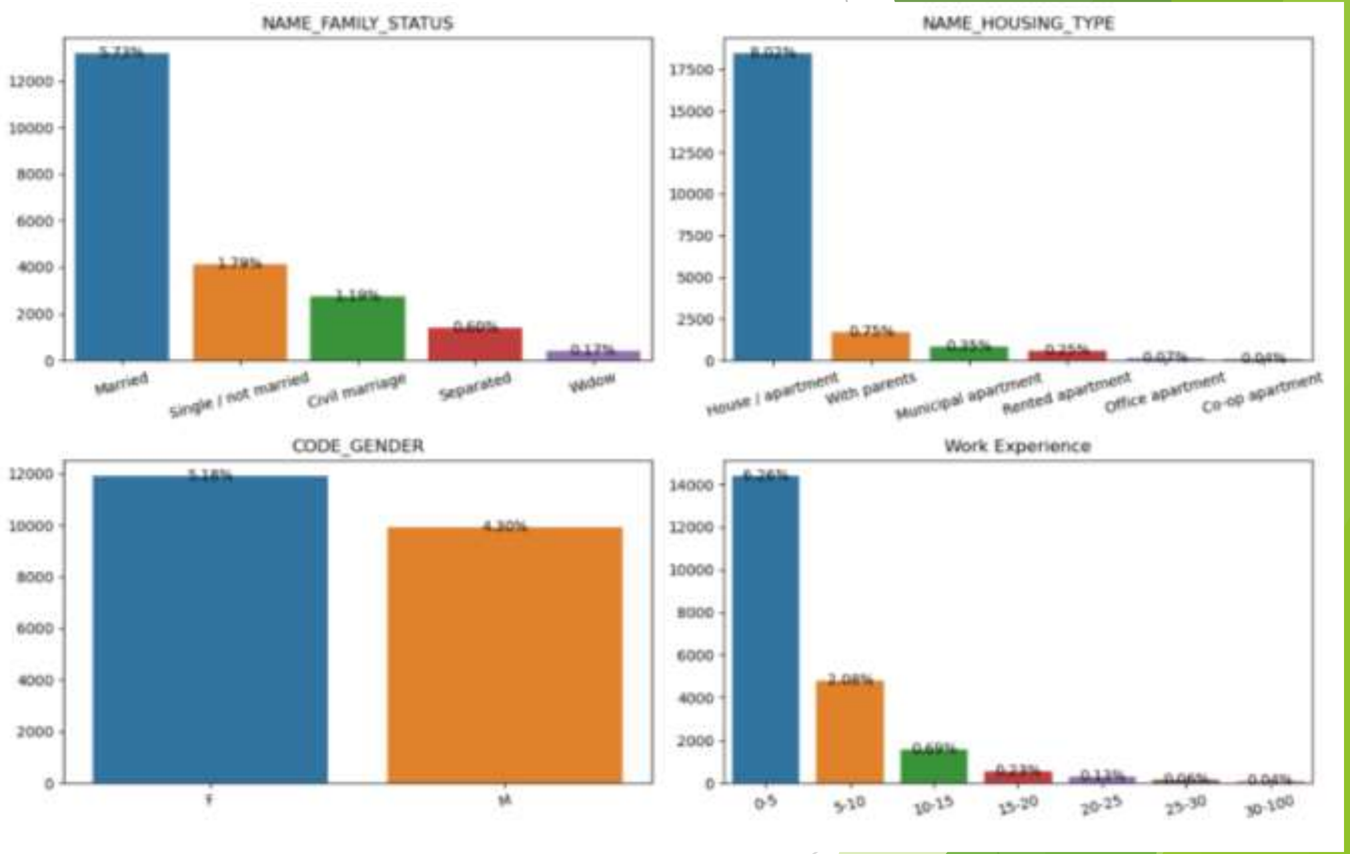
- In this plot we visualized the different attributes based on payment issue.
- The highest issue,
  - In terms of age group is “30-40”, which comprises of 2.60% of total average 8.66%.
  - In terms of contract it is for “Cash loans” which is about 8.84 % of total average.
  - For income types it is visible from “Working” class which is about 6.61%.
  - The people who have education till Secondary/Secondary special type of education i.e. about 7.38%.



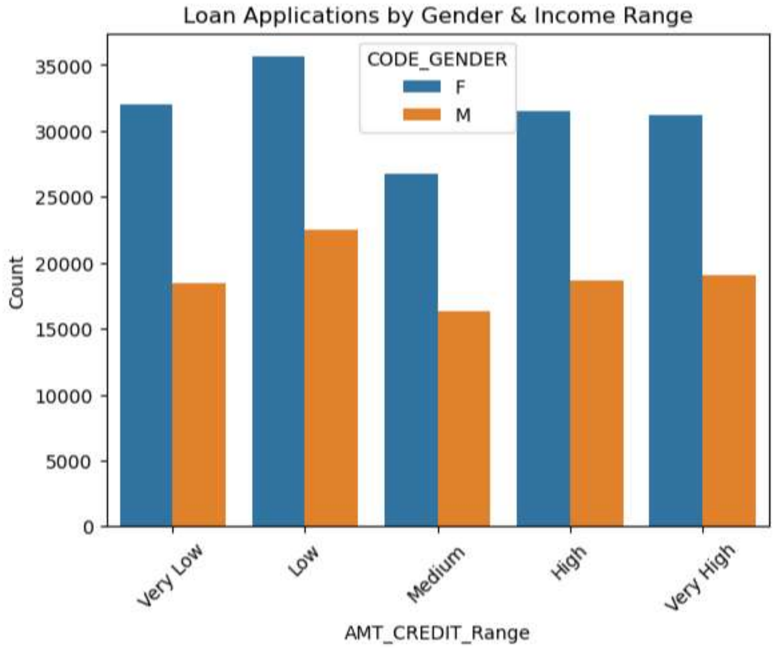
- In this plot we visualized the different attributes based on non-payment issue.
- The highest issue,
  - Is with the client with family status is “Married”, which is about 65.45% of total 91.34%.
  - The client housing type is of “House/Apartments”, which is about 87.72%.
  - Females are of more count than men, which is about 62.02%.
  - Client having work experience with 0-5 years are about 52.92%.



- In this plot we visualized the different attributes based on payment issue.
- The highest issue,
- Is with the client with family status is “Married”, which is about 5.75% of total 8.66%.
- The client housing type is of “House/Apartments”, which is about 8.02%.
- Females are of more count than men, which is about 5.18%.
- Client having work experience with 0-5 years are about 6.26%.

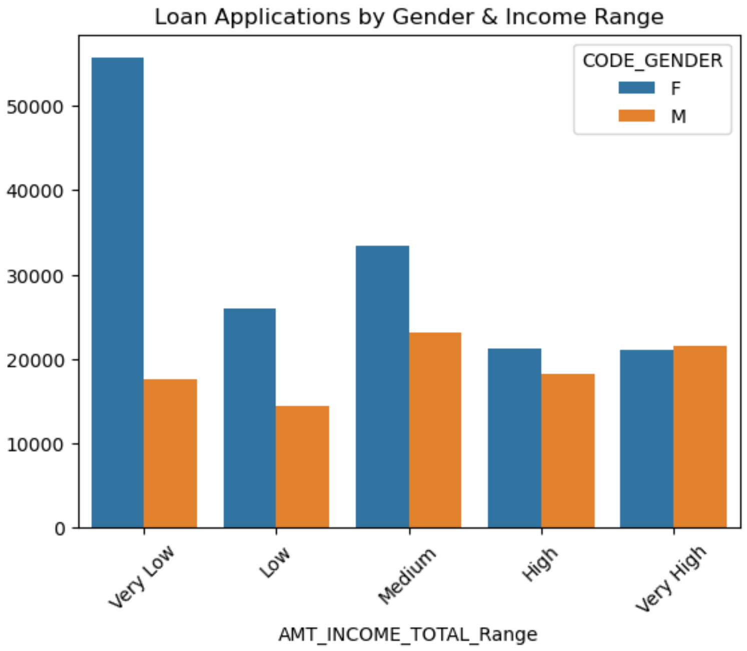


- This is the plot for count of loan application for different credit ranges based on gender(M or F).
- Females are mostly applying for Low credit loans.
- -Males are applying for low credit loans. But compared to women that is less even for high credit loan as they have more credit amount.
- Female applied the most for loan in every category. Since, they have good credit number compared to men.



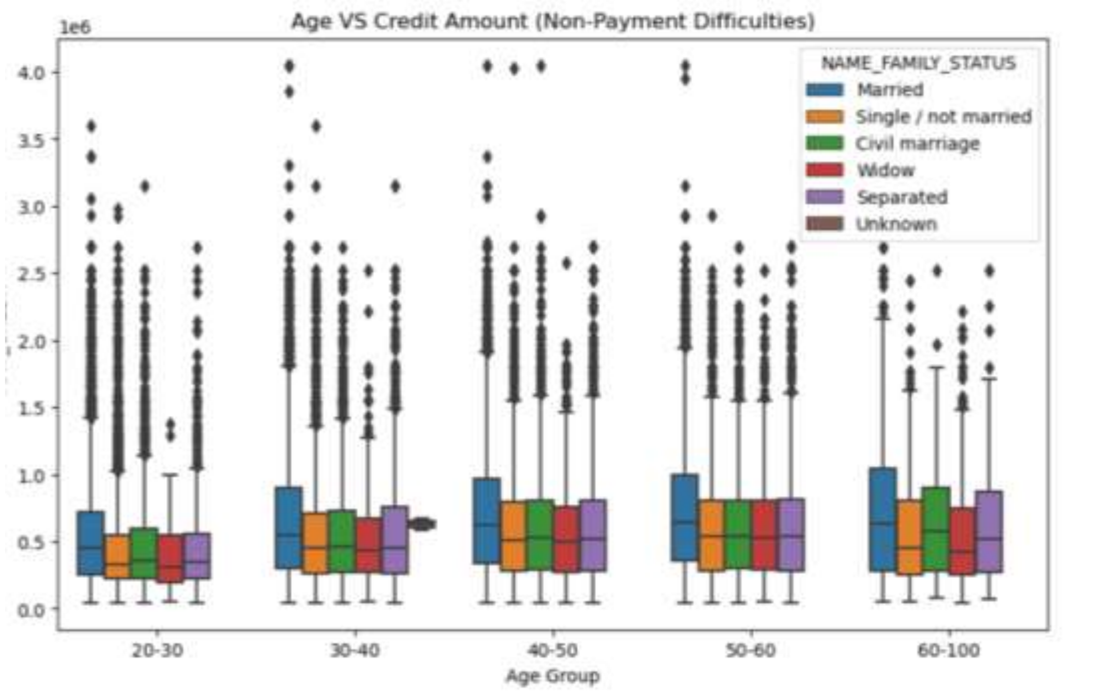
# Bivariant Analysis

- This is the plot for count of loan application for different income ranges based on gender(M or F).
- As its clearly visible that the loan application of women are far greater in number in all of income range other than “Very high”.
- We can say that as the credit amount of women are more than men, so bank would prefer to reach out to women client for proper recovery at the time.

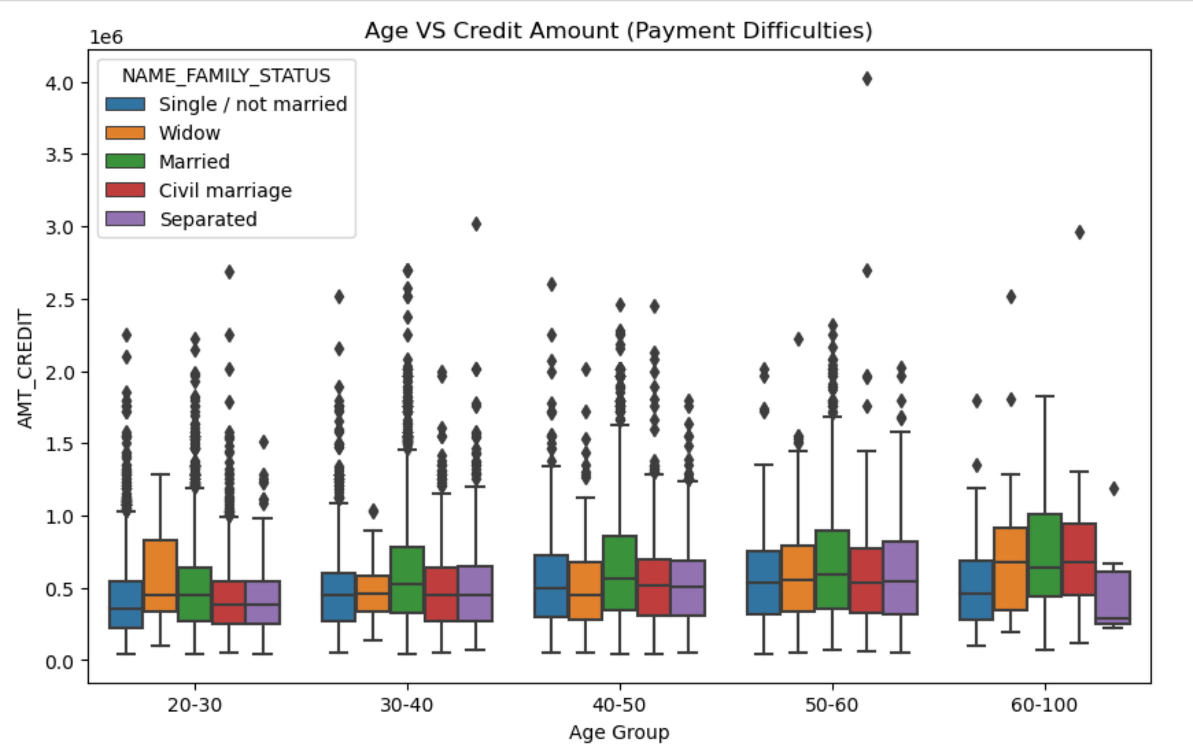


# Multivariant Analysis

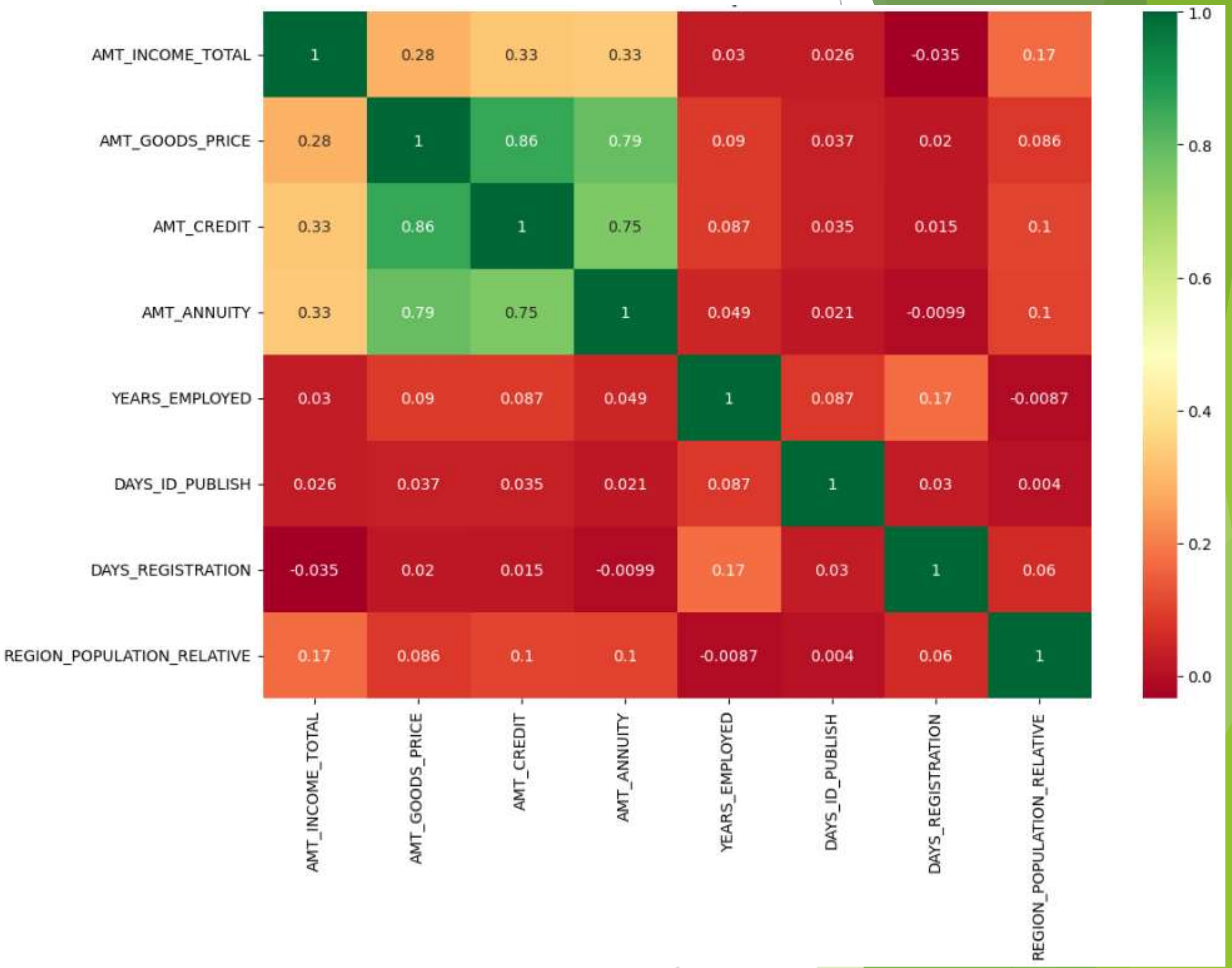
- The plot here is for age group against the credit amount based on the family status of the client, who have non-payment difficulties.
- There are many outliers for each age group, we can say there can be some exception in each age group who can have a very high credit amount compared to rest of the group.
- Also, in every section “Married” family status is having a higher median value for all.
- This shows married clients have some problem which is not related to payment.



- The plot here is for age group against the credit amount based on the family status of the client, who have payment difficulties.
- “Married” family seems to have payment difficulties, but also applied for most number of loans.

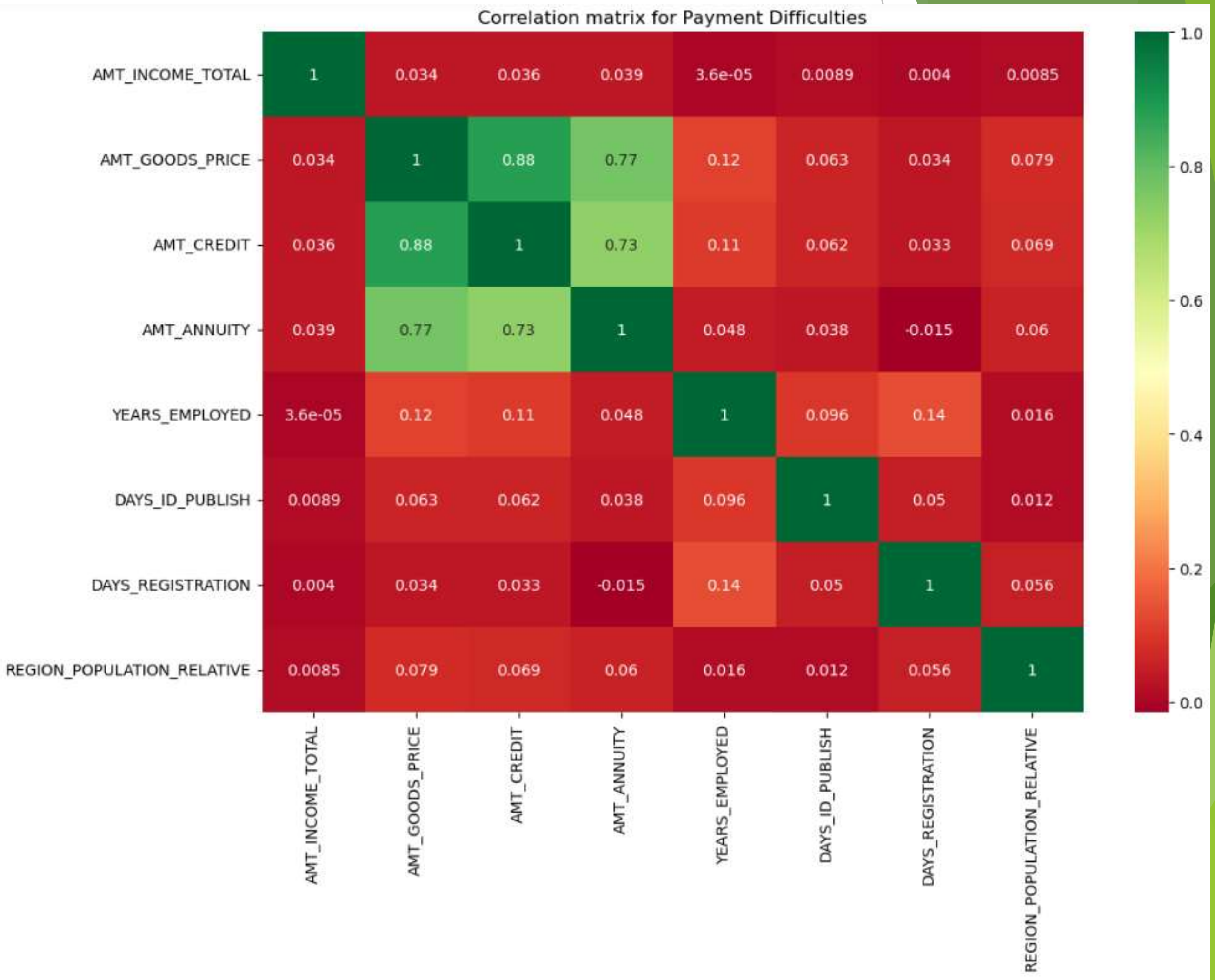


- The heatmap plot of all the important numerical attributes for non-payment issue.
- In terms of it we can see “AMT_CREDIT” and “AMT_GOODS_PRICE” are having a positive correlation among them.
- Also, “AMT_CREDIT” and “DAYS_REGISTRATION” are having the least or negative correlation among them.



| | AMT_INCOME_TOTAL | AMT_GOODS_PRICE | AMT_CREDIT | AMT_ANNUITY | YEARS_EMPLOYED | DAYS_ID_PUBLISH | DAYS_REGISTRATION | REGION_POPULATION_RELATIVE |
|---|---|---|---|---|---|---|---|---|
| AMT_INCOME_TOTAL | 1 | 0.28 | 0.33 | 0.33 | 0.03 | 0.026 | -0.035 | 0.17 |
| AMT_GOODS_PRICE | 0.28 | 1 | 0.86 | 0.79 | 0.09 | 0.037 | 0.02 | 0.086 |
| AMT_CREDIT | 0.33 | 0.86 | 1 | 0.75 | 0.087 | 0.035 | 0.015 | 0.1 |
| AMT_ANNUITY | 0.33 | 0.79 | 0.75 | 1 | 0.049 | 0.021 | -0.0099 | 0.1 |
| YEARS_EMPLOYED | 0.03 | 0.09 | 0.087 | 0.049 | 1 | 0.087 | 0.17 | -0.0087 |
| DAYS_ID_PUBLISH | 0.026 | 0.037 | 0.035 | 0.021 | 0.087 | 1 | 0.03 | 0.004 |
| DAYS_REGISTRATION | -0.035 | 0.02 | 0.015 | -0.0099 | 0.17 | 0.03 | 1 | 0.06 |
| REGION_POPULATION_RELATIVE | 0.17 | 0.086 | 0.1 | 0.1 | -0.0087 | 0.004 | 0.06 | 1 |

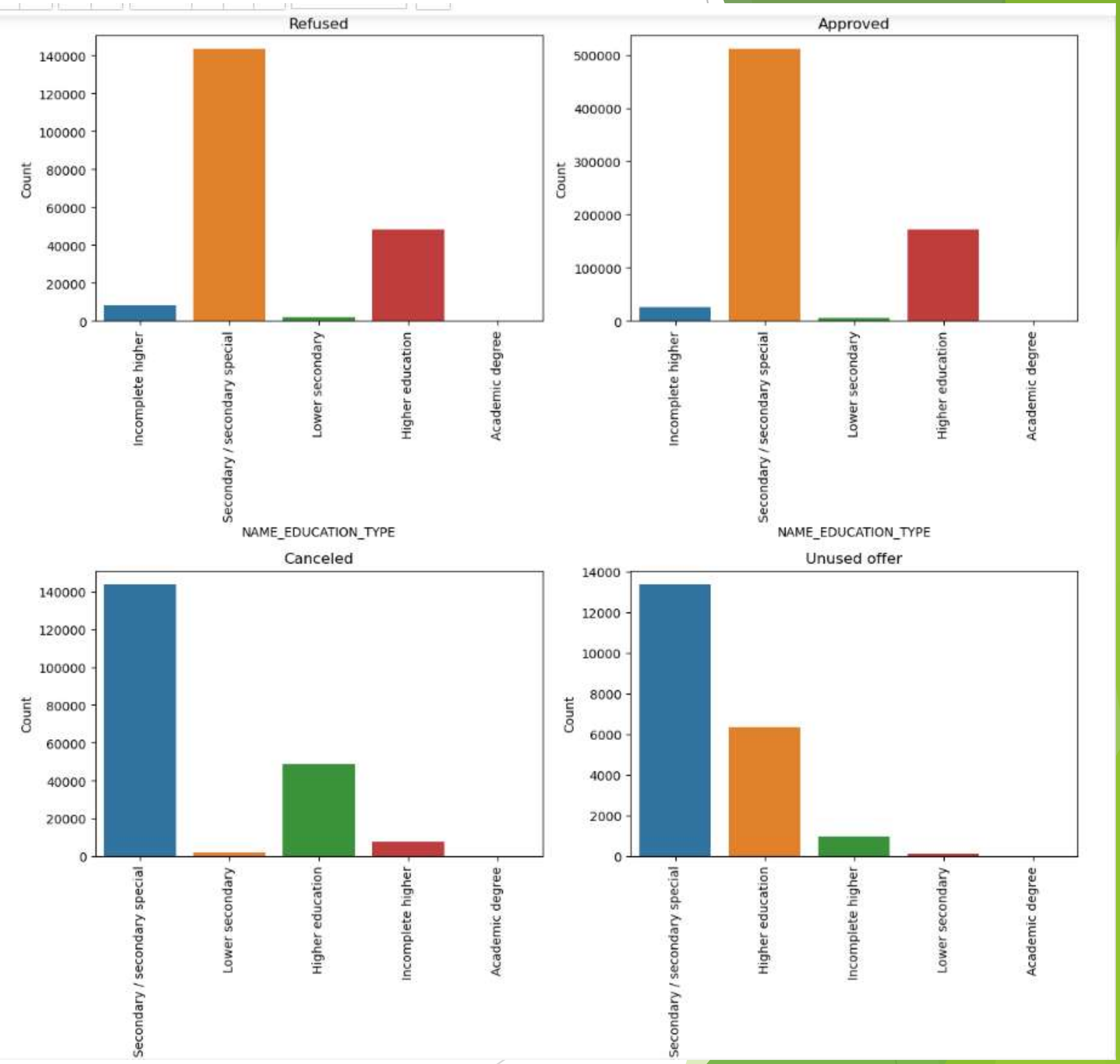
- The heatmap plot of all the important numerical attributes for payment issue.
- In terms of it we can see “AMT_CREDIT” and “AMT_GOODS_PRICE” are having a positive correlation among them.
- Also, “AMT_INCOME_TOTAL” and “YEARS_EMPLOYEED” are having the least or negative correlation among them.


Correlation matrix for Payment Difficulties

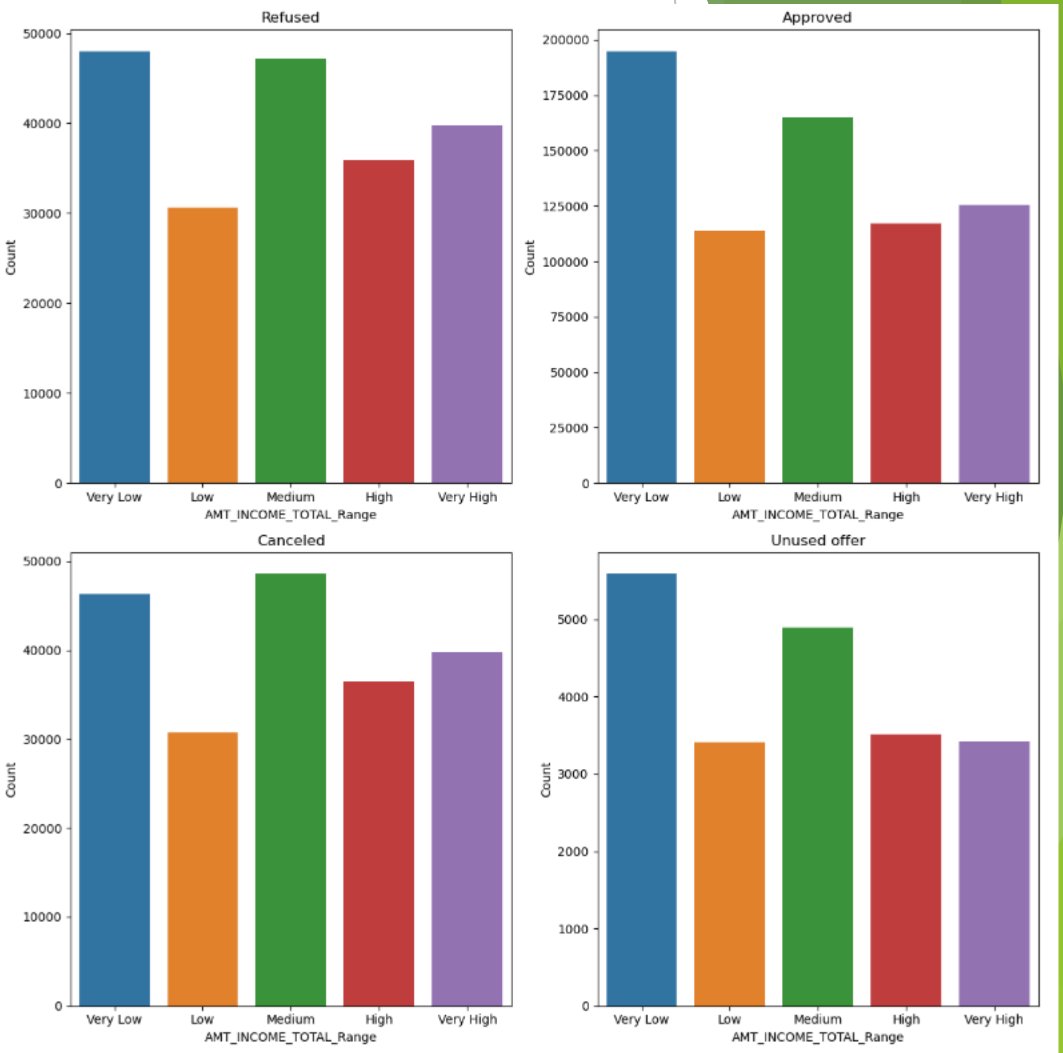
| | AMT_INCOME_TOTAL | AMT_GOODS_PRICE | AMT_CREDIT | AMT_ANNUITY | YEARS_EMPLOYED | DAYS_ID_PUBLISH | DAYS_REGISTRATION | REGION_POPULATION_RELATIVE |
|---|---|---|---|---|---|---|---|---|
| AMT_INCOME_TOTAL | 1 | 0.034 | 0.036 | 0.039 | 3.6e-05 | 0.0089 | 0.004 | 0.0085 |
| AMT_GOODS_PRICE | 0.034 | 1 | 0.88 | 0.77 | 0.12 | 0.063 | 0.034 | 0.079 |
| AMT_CREDIT | 0.036 | 0.88 | 1 | 0.73 | 0.11 | 0.062 | 0.033 | 0.069 |
| AMT_ANNUITY | 0.039 | 0.77 | 0.73 | 1 | 0.048 | 0.038 | -0.015 | 0.06 |
| YEARS_EMPLOYED | 3.6e-05 | 0.12 | 0.11 | 0.048 | 1 | 0.096 | 0.14 | 0.016 |
| DAYS_ID_PUBLISH | 0.0089 | 0.063 | 0.062 | 0.038 | 0.096 | 1 | 0.05 | 0.012 |
| DAYS_REGISTRATION | 0.004 | 0.034 | 0.033 | -0.015 | 0.14 | 0.05 | 1 | 0.056 |
| REGION_POPULATION_RELATIVE | 0.0085 | 0.079 | 0.069 | 0.06 | 0.016 | 0.012 | 0.056 | 1 |

# Analysis of previous_application.csv

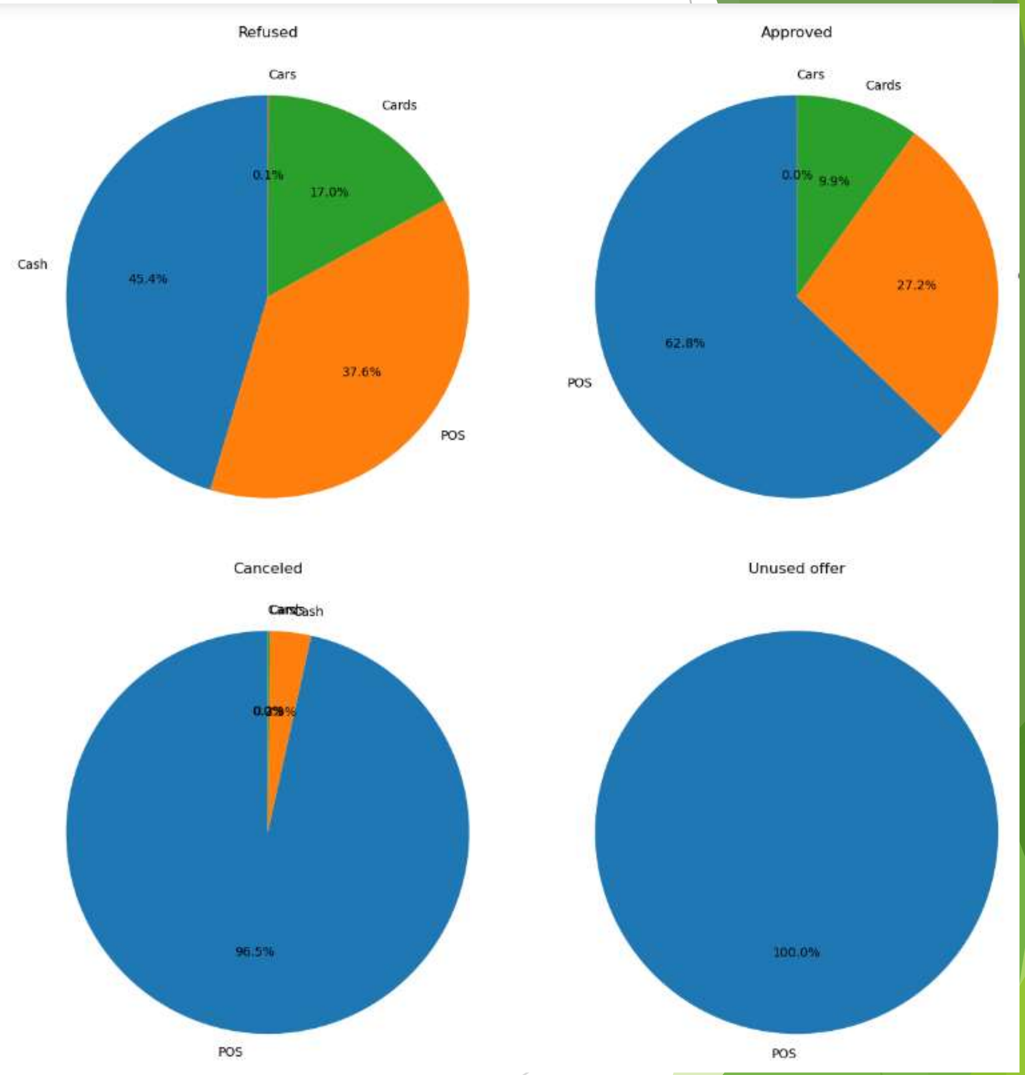
- It is the plot of previous clients and plot on there request status for loan based on education.
- We can see that most of the approved and canceled loans belong to applicants with Secondary / Secondary Special education type. Also, they only belong under most unused offer section for Education.



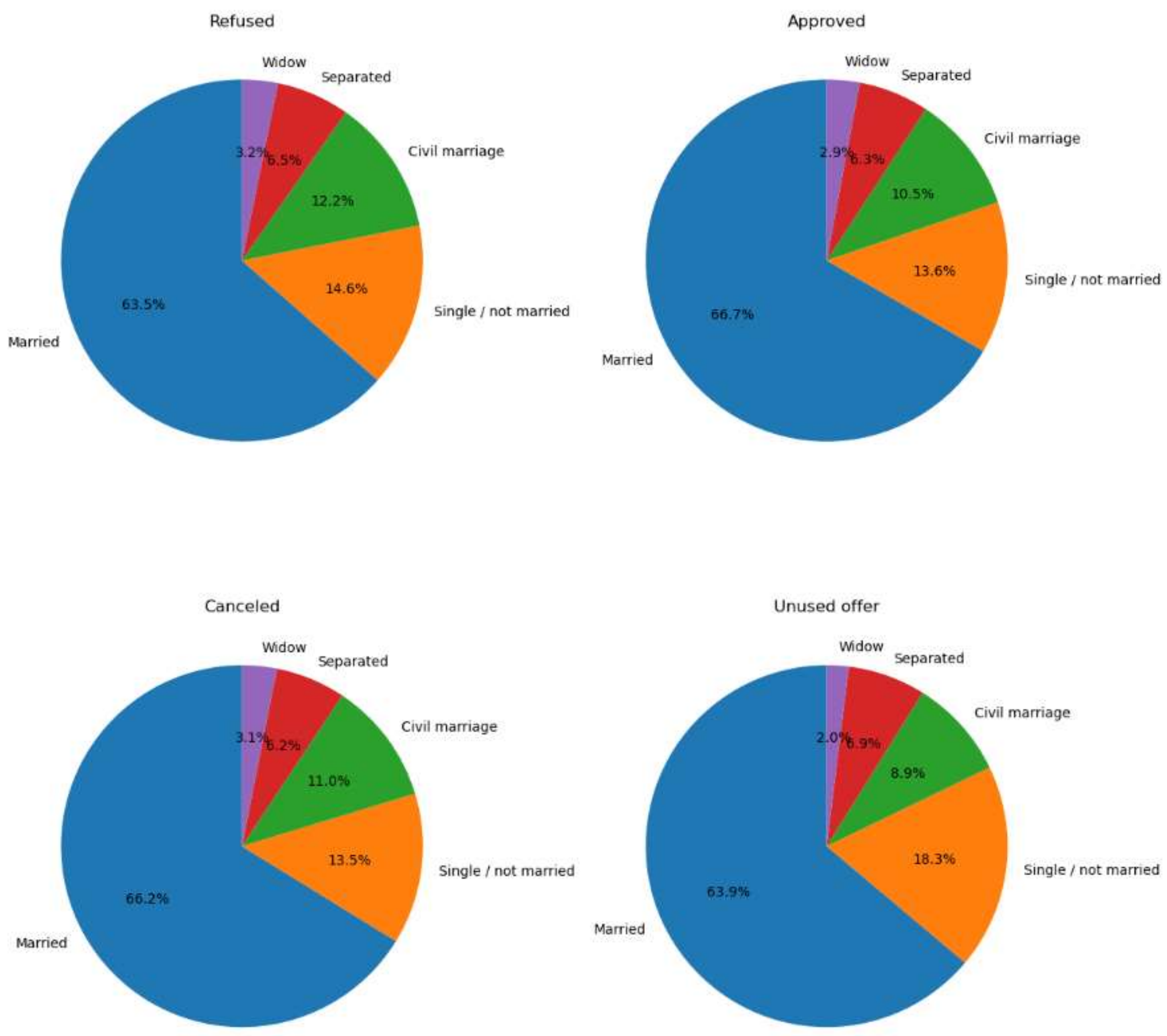
- It is the plot of previous clients and plot on there request status for loan based on income range.
- Most of the loans are getting approved for Applicants with very Low Income range it is possible they are opting for low credit loans. A large number of loan are rejected or canceled even though applicant belong to HIGH Income range. May be they have requested for quite HIGH credit range.



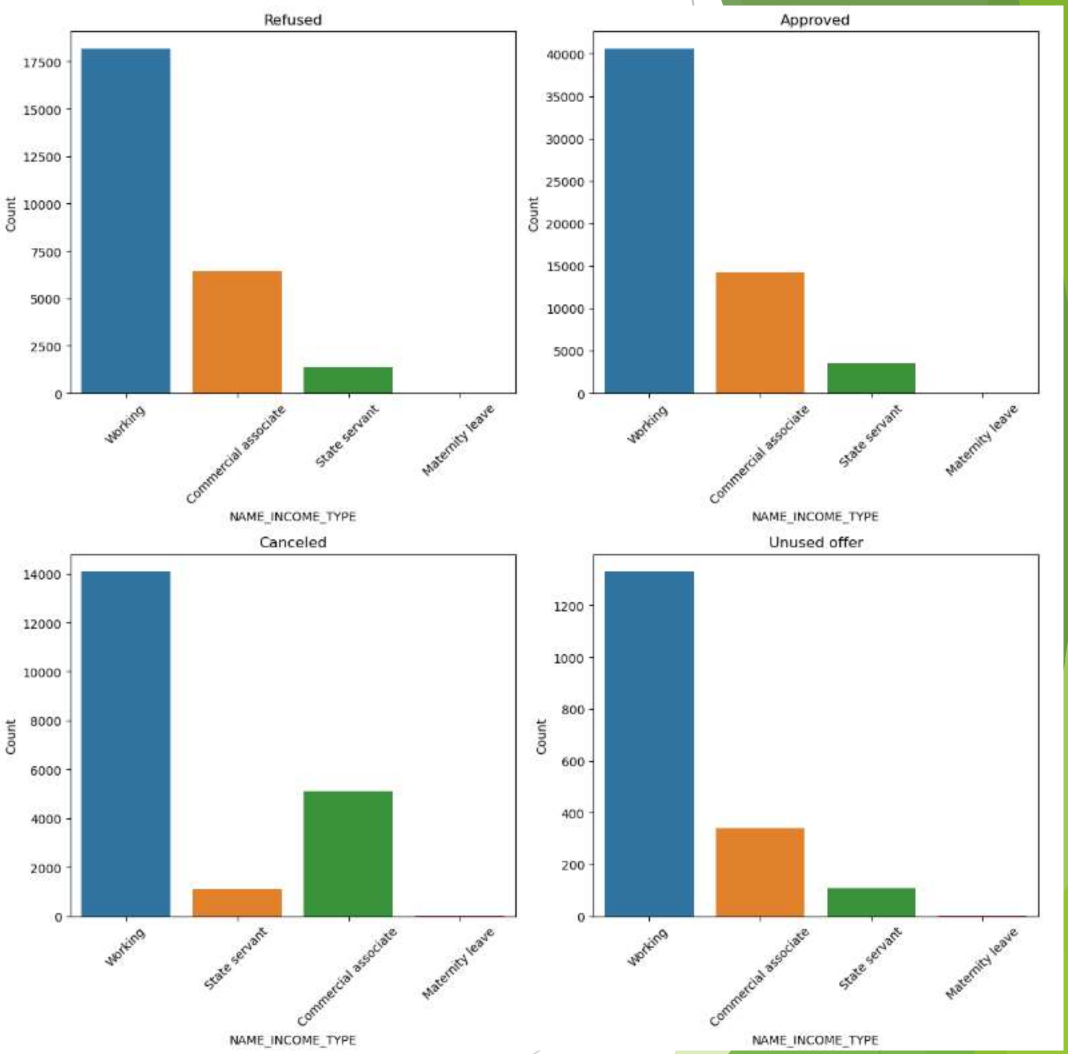
- This is the pie-plot of previous clients and plot on there request status for loan based on different portfolio of loans.
- 62.6% previous approved loans belong to POS name portfolio.
- Majority of the loans refused were cash loans.
- 96.5% loans that belong to POS were canceled.



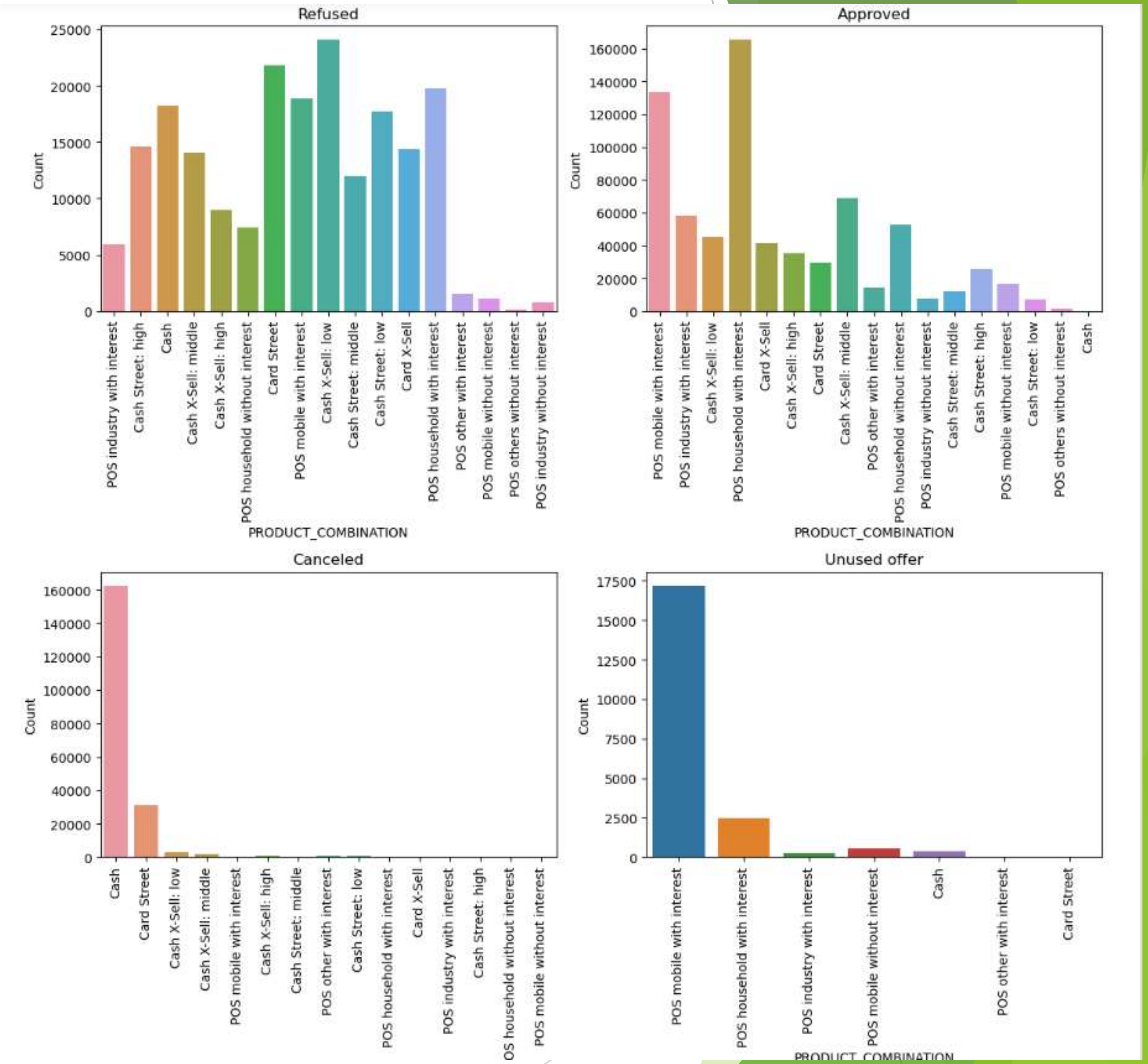
- This is the pie-plot of previous clients and plot on there request status for loan based on family status.
- Approved percentage of loans for married applicants is higher than the rest of the contract status (refused , canceled etc.).
- All the category have “Married” family status as common.



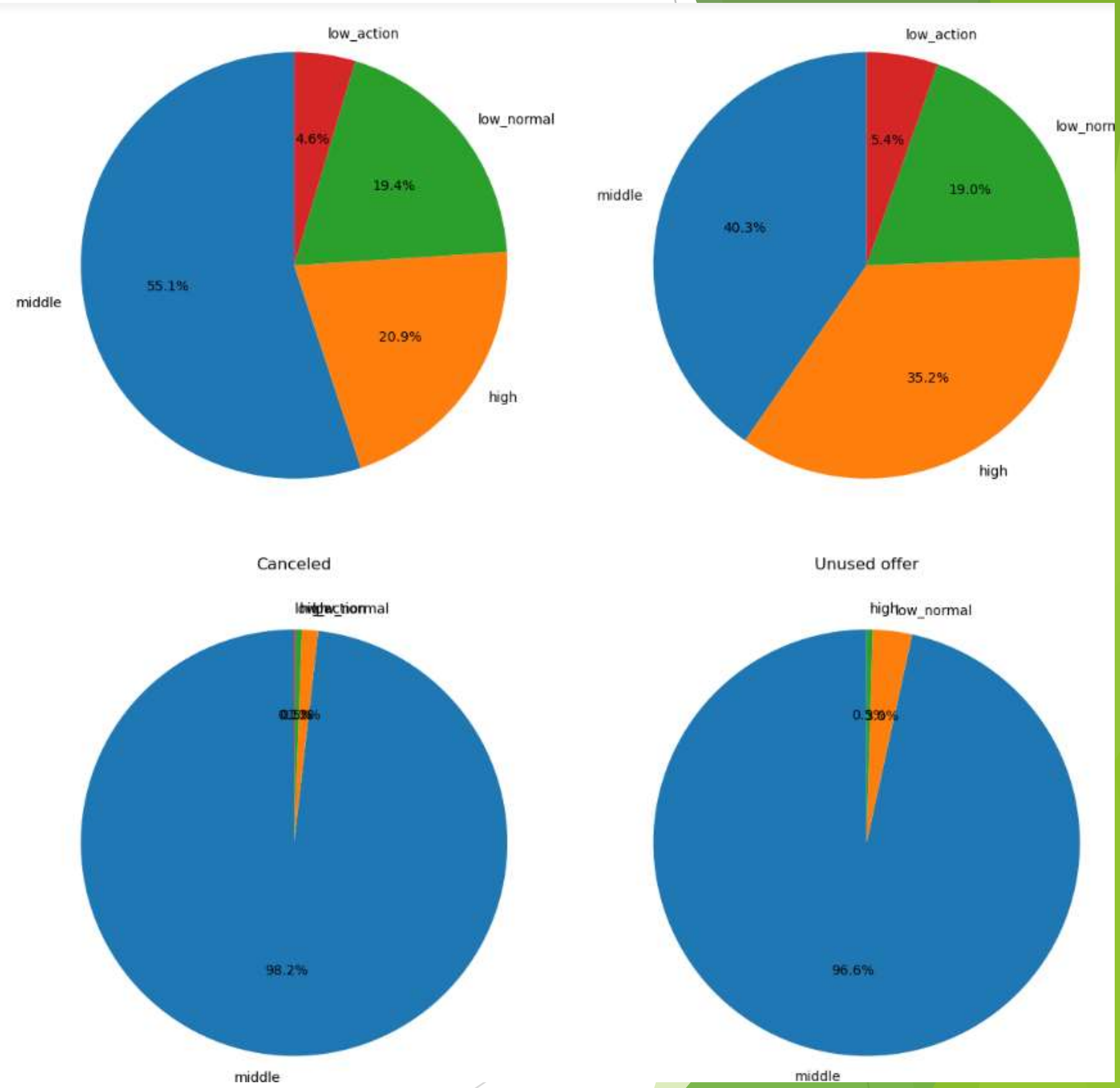
- It is the plot of previous clients after splitting the data frame based on “TARGET” and plot on there request status for loan based on income type.
- Working class income type people are mostly approved loan, but they are facing to payback the loan.
- Also the same Working class has been refused or canceled the loan.



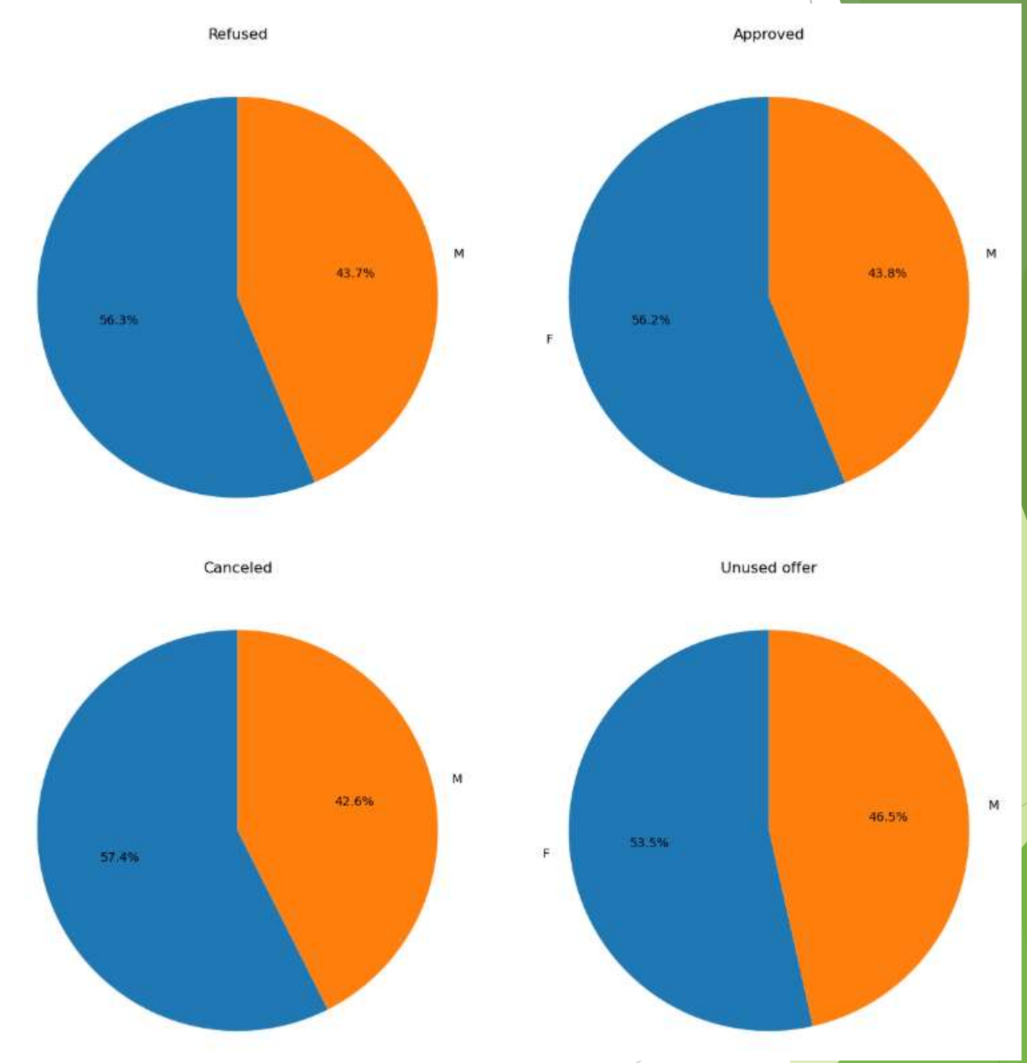
- It is the plot of previous clients after splitting the data frame based on “TARGET” and plot on there request status for loan based different product.
- Most of the approved loans belong to POS hosehold with interest & POS mobile with interest product combination.
- Most rejected loan is from "Cash X-Sell: low"
- Most of the canceled loans are for Cash category.
- Unused loan offer are from "POS mobile with interest".



- This is the pie-plot plot of previous clients after splitting the data frame based on “TARGET” and plot on there request status for Grouped interest rate into small medium and high.
- The defaulted loans which were approved are under "Medium" yield group loans, where clients had defaulted.
- Most of the Refused or Cancelled loans are also under "Medium" yield group loans.



- This is the pie-plot plot of previous clients after splitting the data frame based on “TARGET” and plot on there request status based on gender.
- Female are most approved of the loan, as they have high credit score, we saw they have huge default rate.
- Compared to loan provided to men the default rate is quite high among them.



# FINAL INSIGHTS:

- Based on the analysis, here are some of the insights we gathered:
- Most of Refused and Cancelled loans were cash loans.
- Percentage of loans approved for females is higher than the percentage refused.
- Most approved loans belong to Very Low and High Credit range.
- Large number of loans are approved for applicants with a low income range.
- Most approved loans belong to applicants with Secondary/Secondary Special education type.
- The percentage of loans approved for married applicants is higher than for other contract status categories (refused, canceled, etc.).
- Most of the approved loans have a medium grouped interest rate.
- Across all contract statuses (Approved, Refused, Canceled, Unused Offer), people with the Working income type are leading.
- Most of the loans that were previously approved belong to the POS name portfolio.
- Credit and cash channel type has the highest number of refused and canceled loans.
- Though we tried to simplify the most of the data frame and bring out insight, we can execute future analysis on the data frame to bring out more details out of it.