Loan Credit Risk EDA Case Study

By Bhavitha B

The case study

- The case study is an EDA (Exploratory Data Analysis) on a finance company that lends money to its clients/customers in the form of loans. We can help the business by determining measures to reduce the risk of losing money while lending to customers using this case study. Using the analyzed data, we can also assist the firm in focusing/targeting on clients who are capable of repaying the loan amount, hence assisting the business in growing.

EDA aim

- The objective of the EDA is to identify the driving factors which leads to loan defaulting. By analyzing their history using the previous data, the loan/credit amount can be amended for the risky applicant, interest rates can be hiked or the loan application can be rejected in case of a bad history.

Steps:

- Understanding the data
- Data cleaning(includes dealing with missing values, outliers and standardization)
- Check for Data Imbalance
- Performing univariate, segmented univariate & bivariate, multivariate analysis and derive the insights
- Understanding the previous application data
- Clean the data and merge with application data
- Perform bivariate analysis on the merged data and derive the insights

Problem statement

The application data consists of two Target categories. They are the risk associated with the applicant.

- 1- applicant had issues in the past while applying for a loan or is unlikely to pay the loan amount on time, thus being problematic for business
- 0: no issues for the applicant and the applicant is likely to repay the loan amount, thus applicant safe for business.

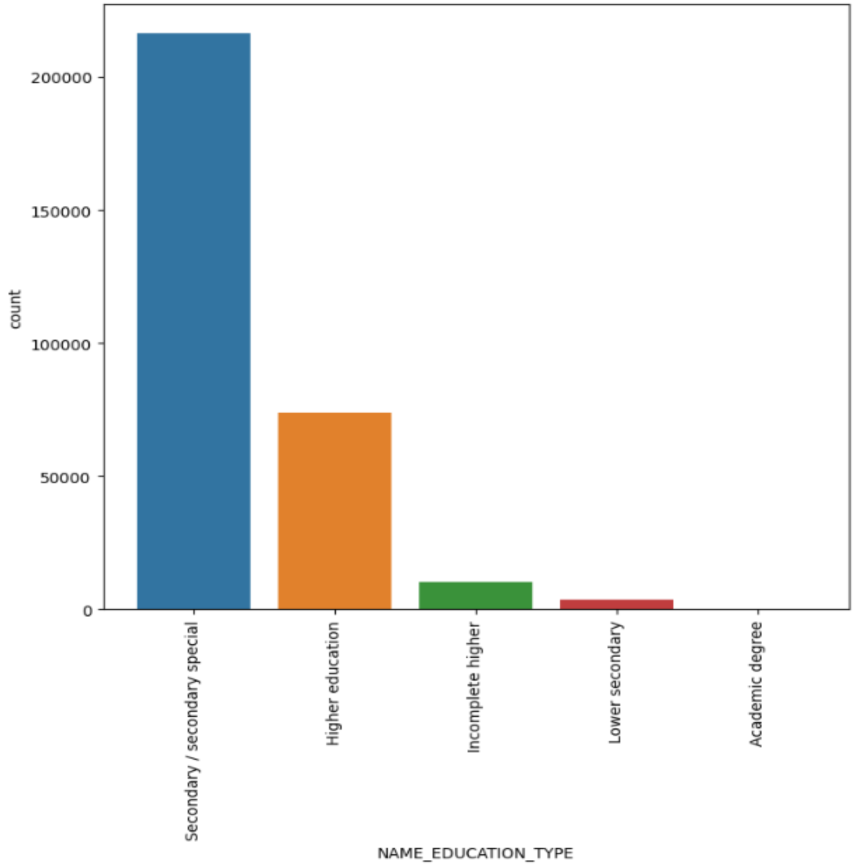
Understanding the given data

- This data is 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.

Data cleaning

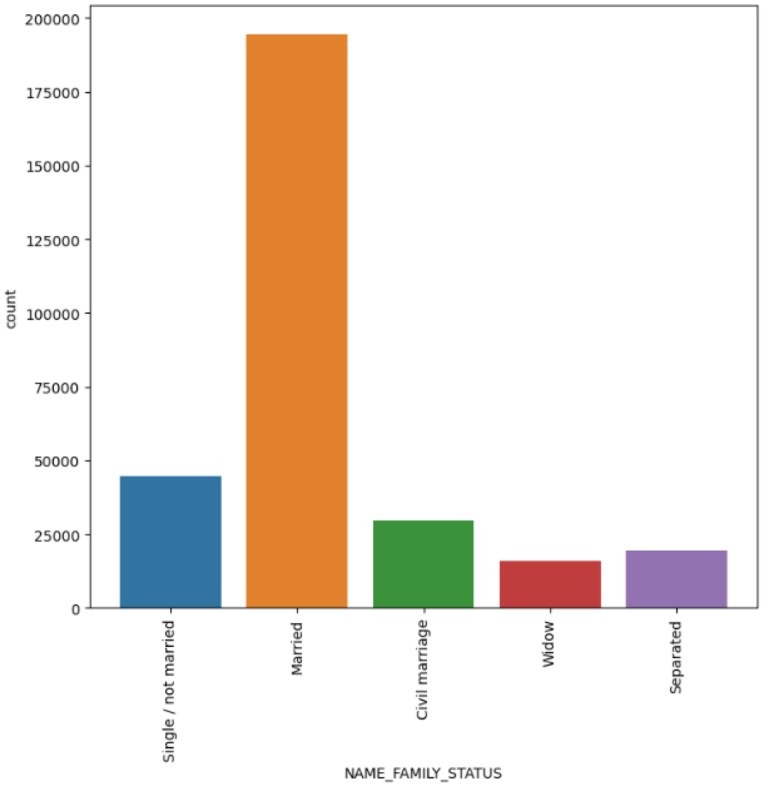
- When preparing data for analysis, it's important to detect and address any issues that may affect the accuracy and validity of the results. This includes identifying missing values, outliers, and abnormalities, and deciding whether to eliminate them from the dataset or impute missing values.
- To detect outliers and abnormalities, we can use statistical techniques such as box plots. Once identified, we can choose to remove or alter these values, depending on the nature and extent of the outlier.
- Standardizing or normalizing the data can help make it simpler to compare and analyze. This involves transforming the data to have a standard mean and standard deviation, or scaling the data to a standard range.
- If the data requires it, we may need to convert data types to ensure that the data is in the proper format for analysis. For example, converting a days field to a years field.
- Lastly, creating new variables from the existing data can provide additional insights or improve predictive accuracy. This can be done by combining variables or creating new derived variables based on the existing data.

Univariate analysis: Education type



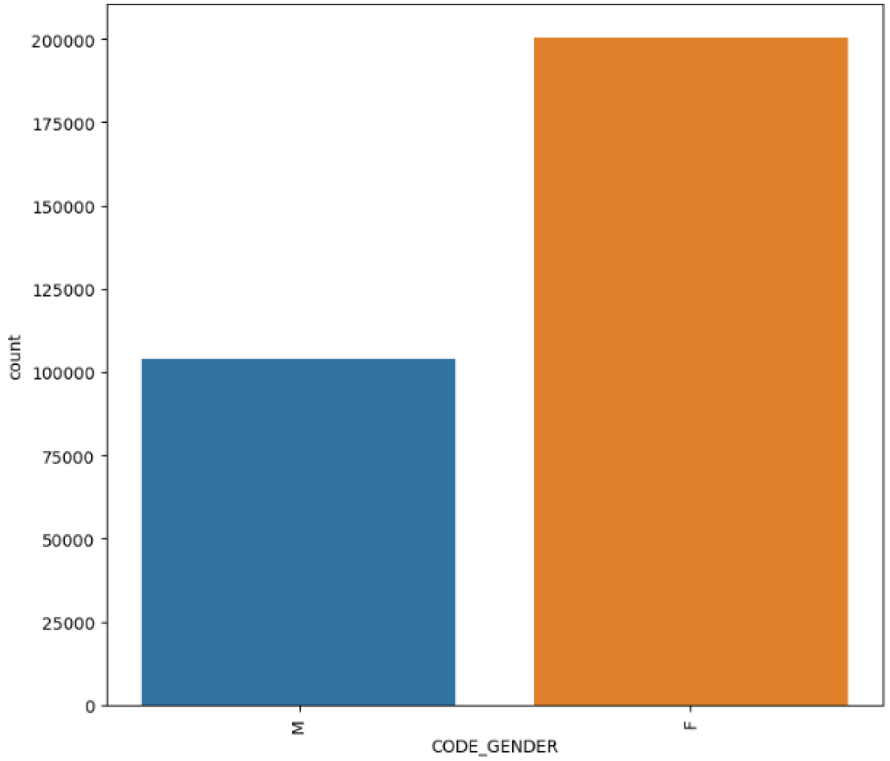
People with Secondary/Secondary Special education have mostly have applied for the loan.

Univariate analysis: Family status



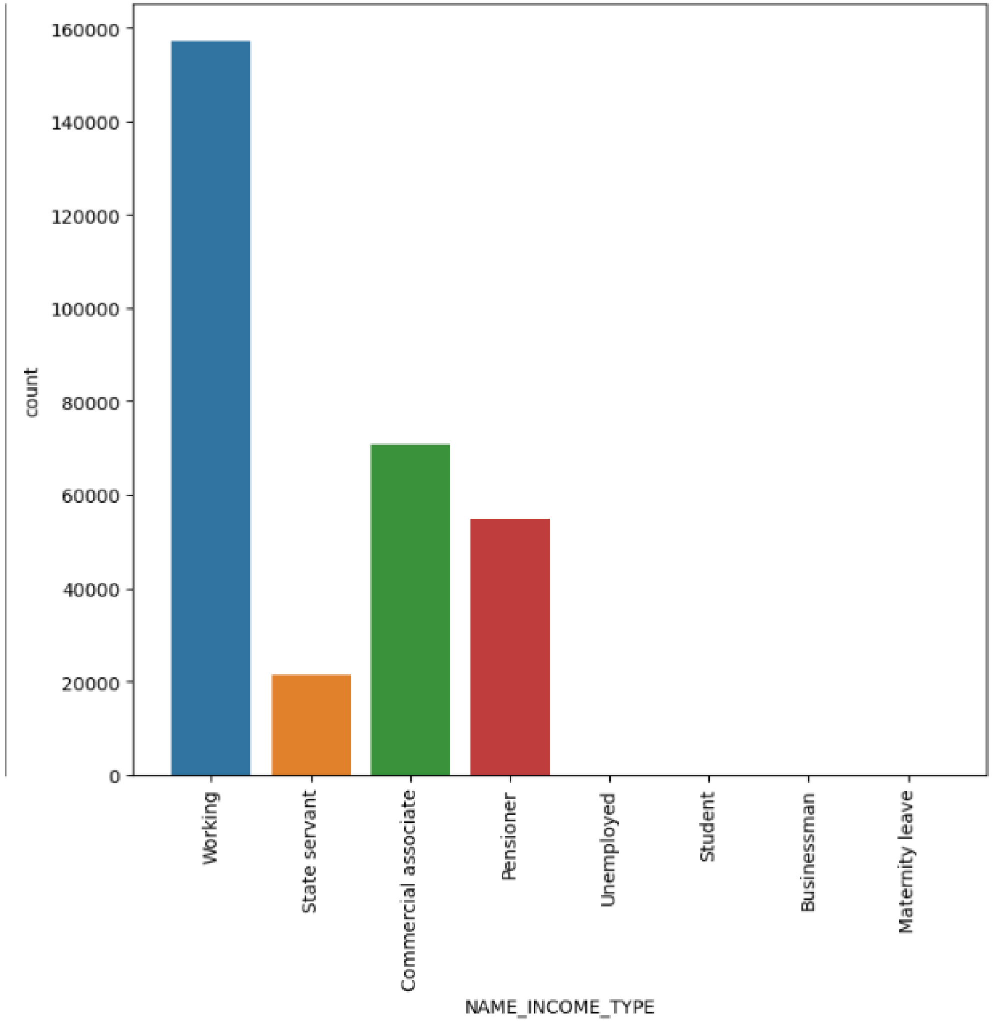
Mostly married people are tend to take more loans

Univariate analysis: Gender



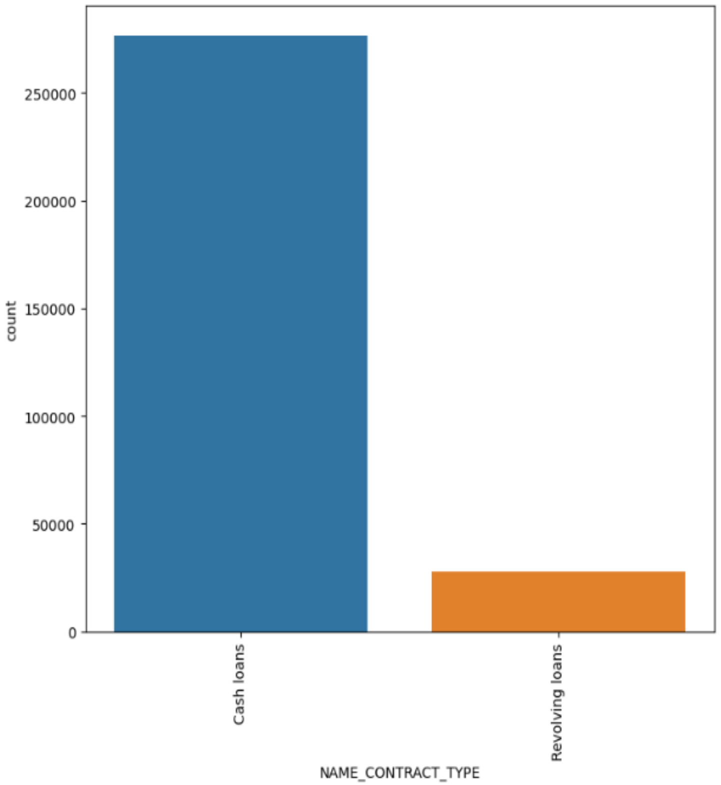
Mostly females
have taken
more loans

Univariate analysis: Income type



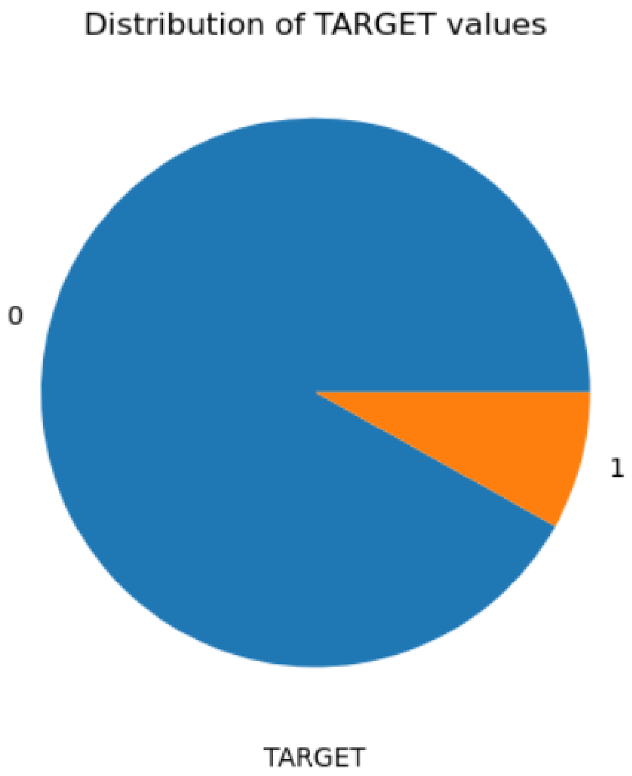
More than 50% of clients who have applied for loan belong to Working Income Type.

Univariate analysis: Contract type



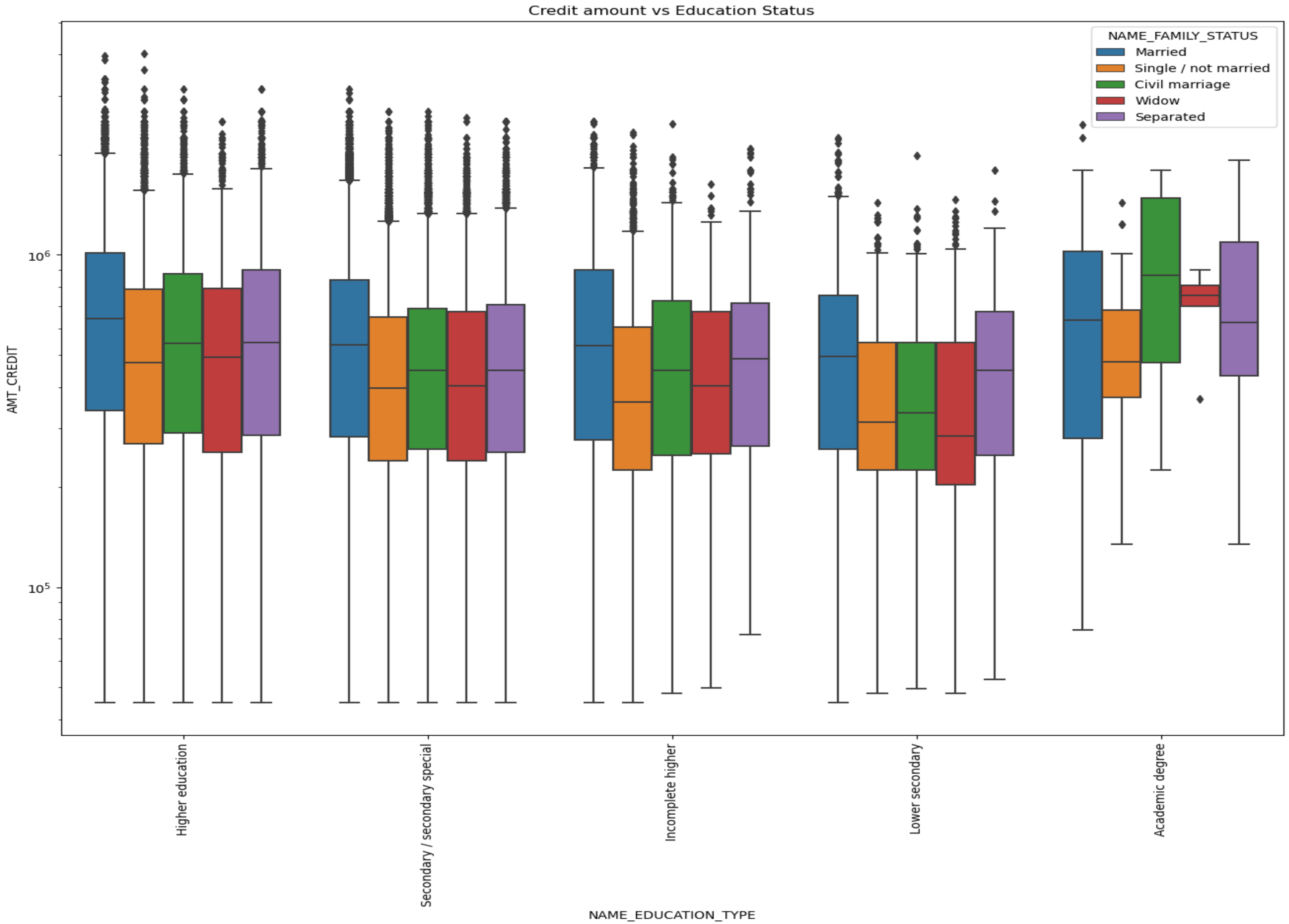
Majority of the applicants preferred cash loans

Data Imbalance



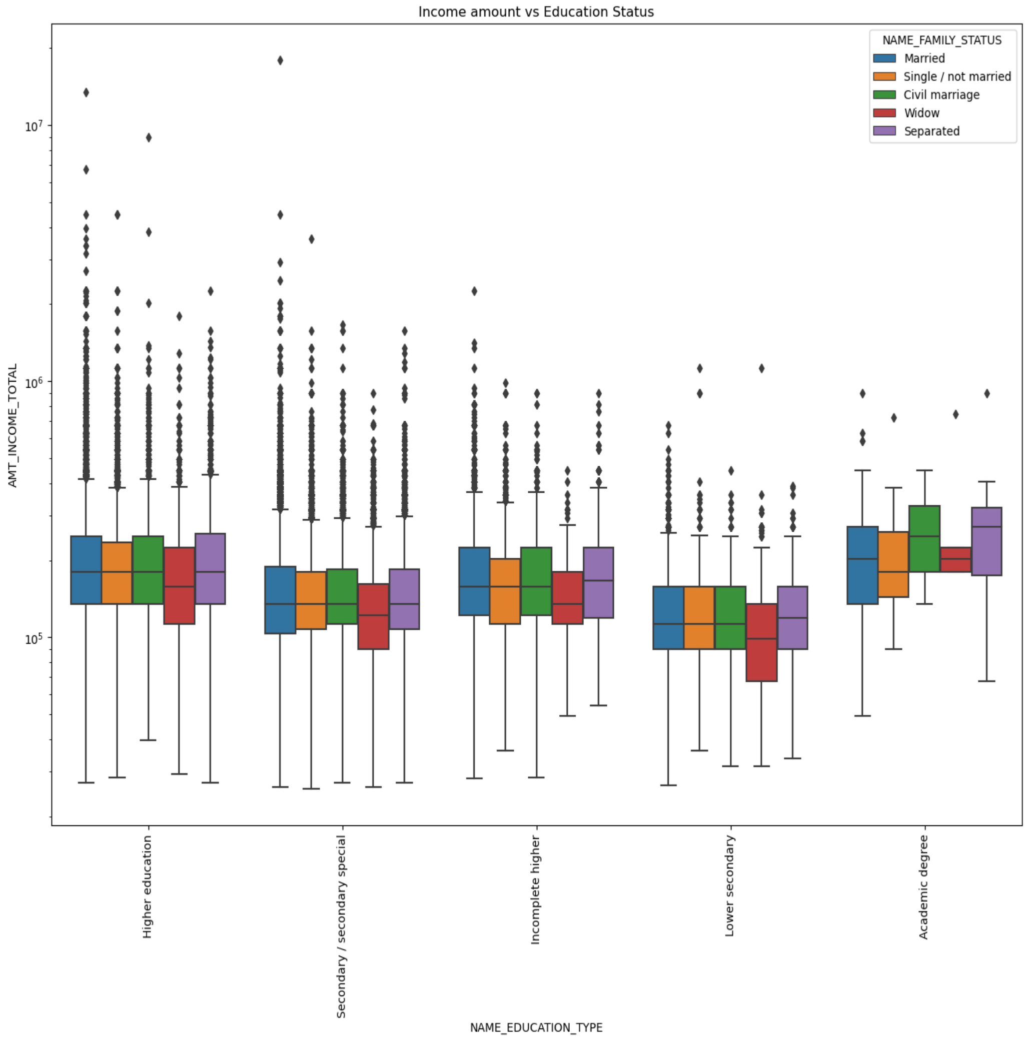
92% of the applicants are non-defaulters.
8% of the applicants have issues with loan repayment

Bivariate analysis: Credit amount vs Education status with non difficulty in payment



Married individuals are interested in taking loans.
Education level affects interest in taking loans.
The box plot suggests that:
Civil marriage, marriage, and separated family statuses with academic degrees have higher credit numbers.
Higher education levels for family statuses of marriage, single, and civil marriage have more outliers.
Civil marriage with an academic degree has most of the credits in the third quartile.

Bivariate analysis: Income amount vs Education status with difficulty in payment



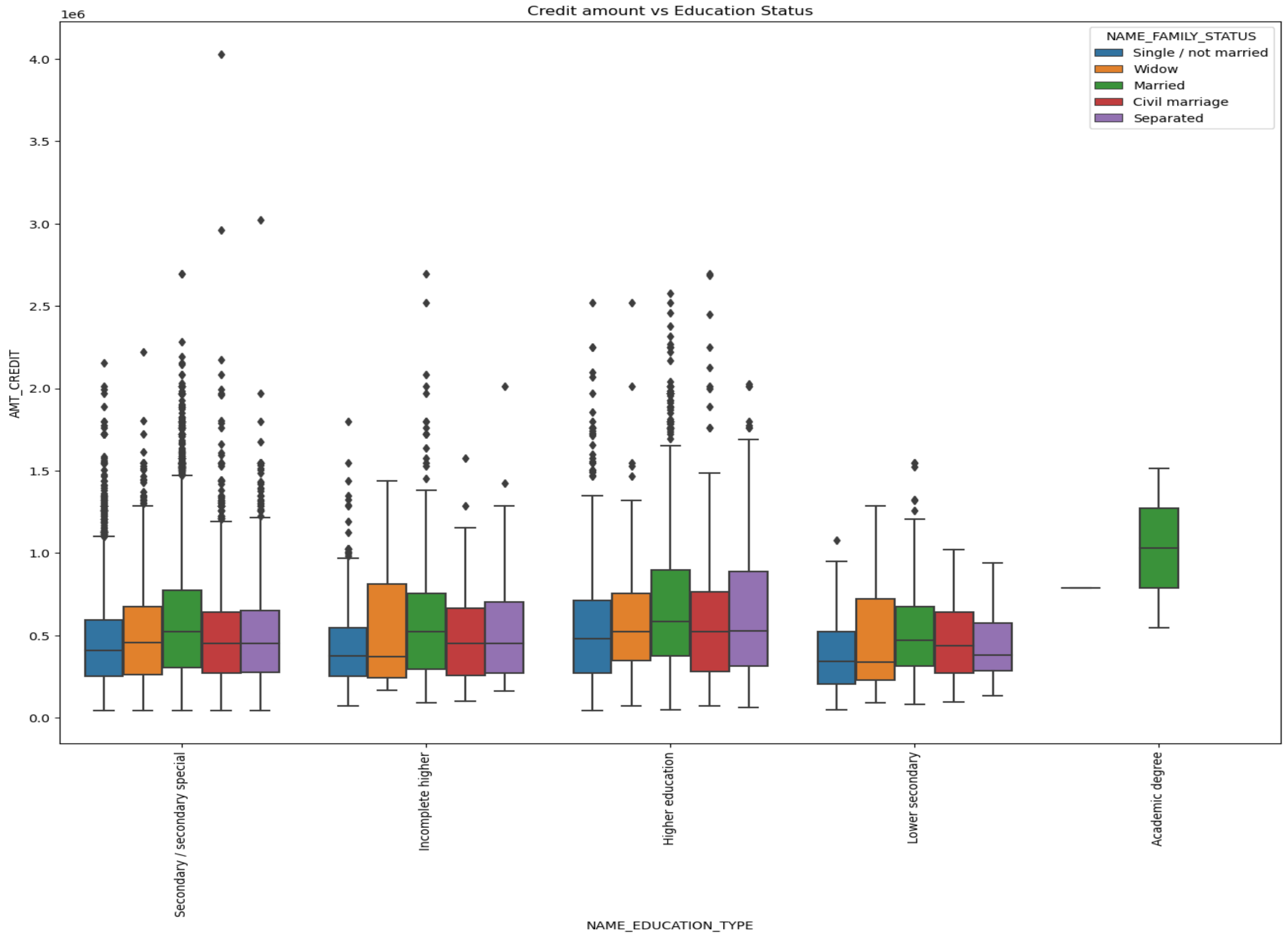
•For the "Higher education" education type, the income amount is mostly equal to the family status.

•There are many outliers for this education type.

•For "Academic degree" education type, there are fewer outliers, but the income amount is slightly higher than for "Higher education."

•For the "Lower secondary" education type and civil marriage family status, the income amount is less than for others.

Bivariate analysis: Credit amount vs Education status with non difficulty in payment

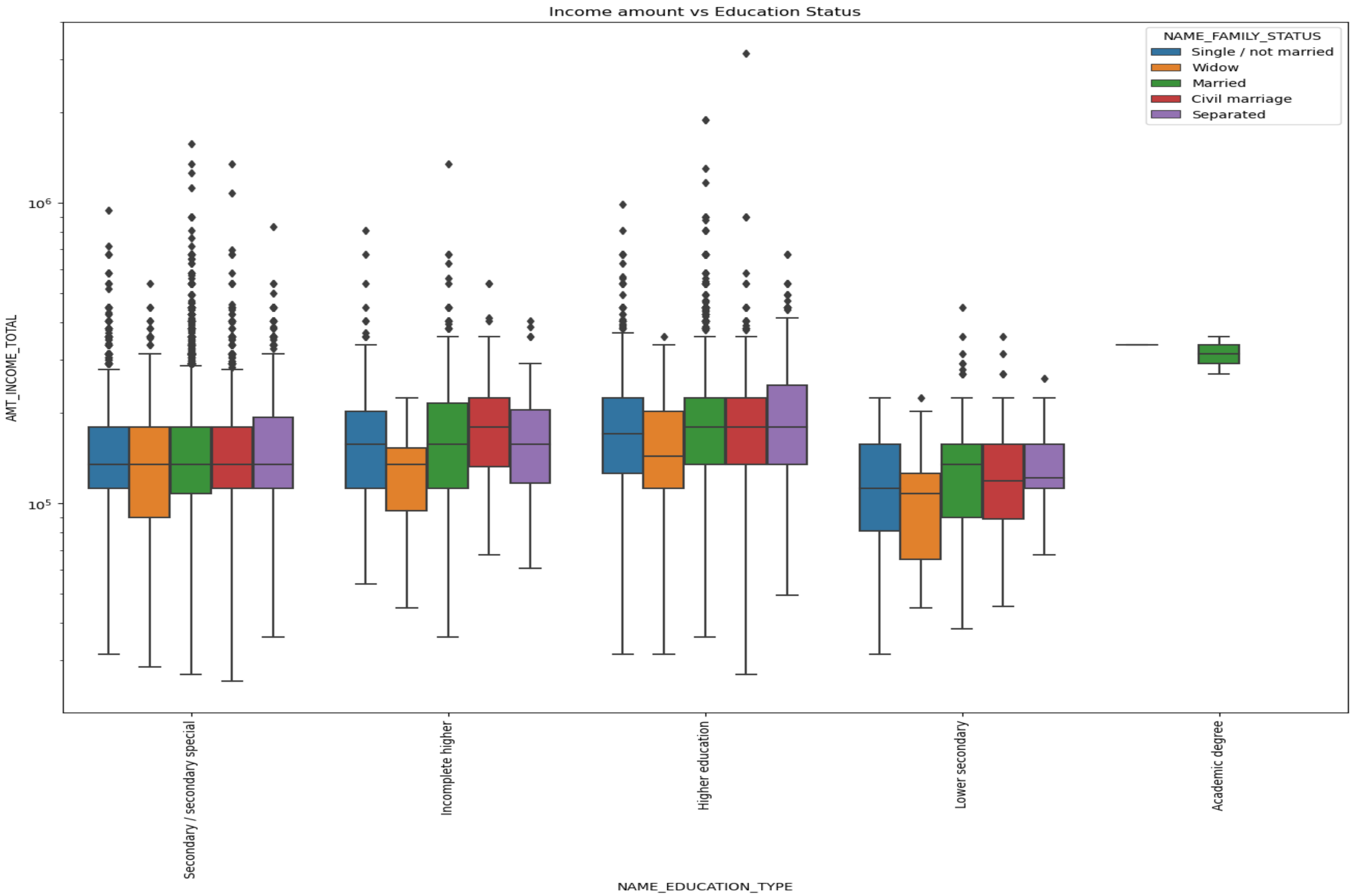


•Similar to Target 0, the box plot shows that the family statuses of "civil marriage," "marriage," and "separated" with an academic degree education have higher numbers of credits than others.

•Most of the outliers are from the "Higher education" and "Secondary" education types.

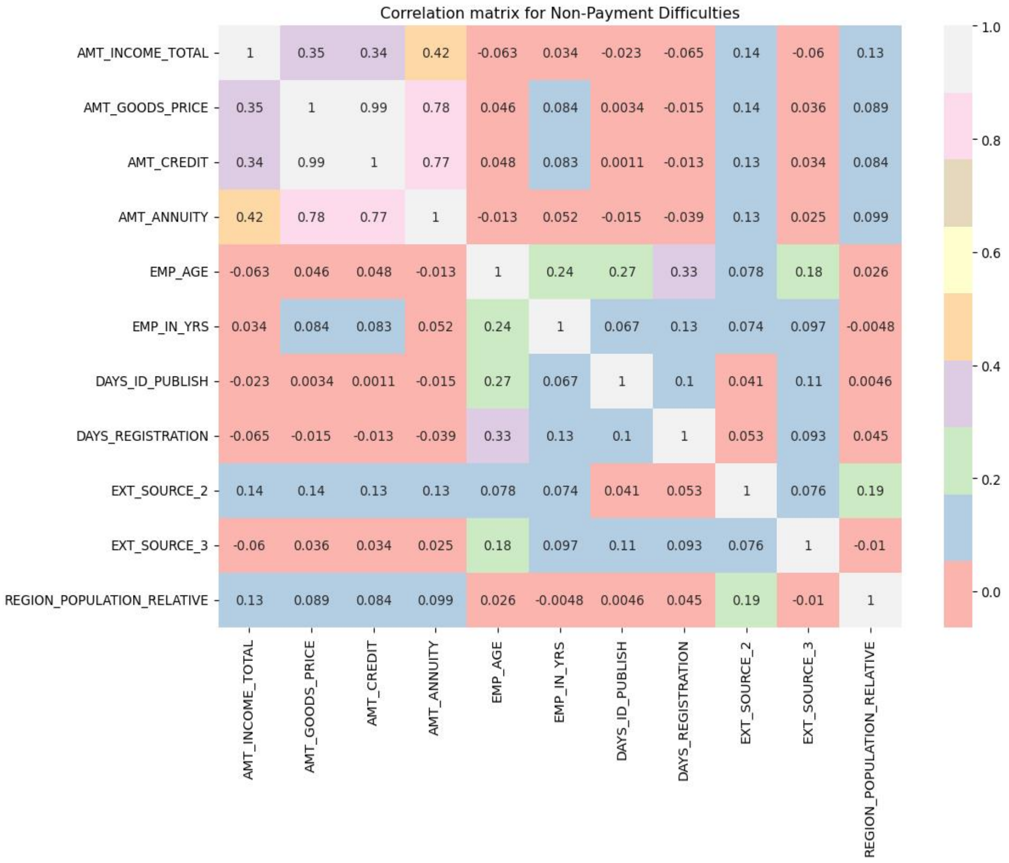
•For the "Academic degree" education type and civil marriage family status, most of the credits are in the third quartile.

Bivariate analysis: Income amount vs Education status with difficulty in payment



- Similar to Target0, for the "Higher education" education type, the income amount is mostly equal to the family status.
- There are fewer outliers for the "Academic degree" education type, but their income amount is slightly higher than for "Higher education."
- For the "Lower secondary" education type, the income amount is less than for others.

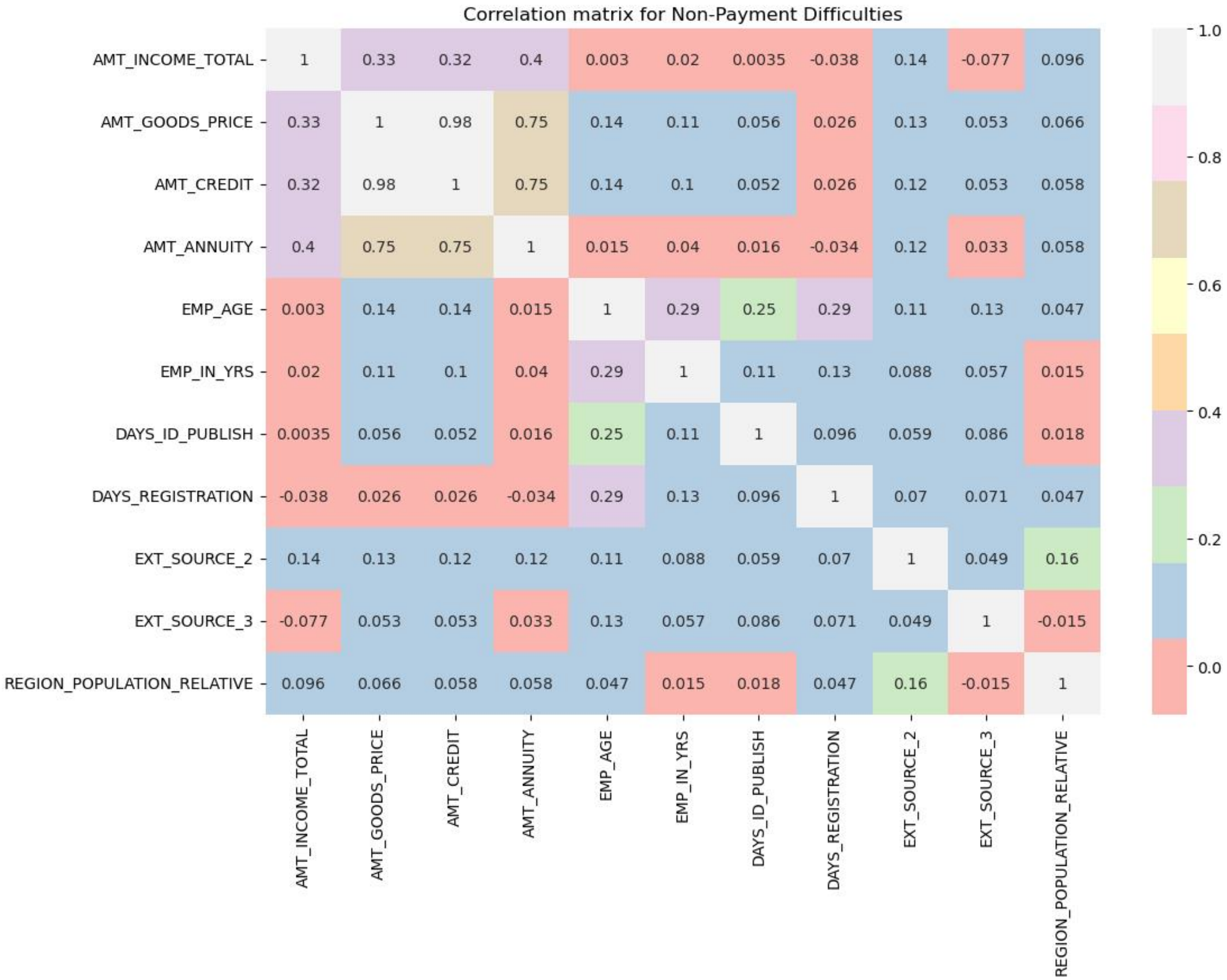
Multivariate analysis: for target 0


Correlation matrix for Non-Payment Difficulties

	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	AMT_CREDIT	AMT_ANNUITY	EMP_AGE	EMP_IN_YRS	DAYS_ID_PUBLISH	DAYS_REGISTRATION	EXT_SOURCE_2	EXT_SOURCE_3	REGION_POPULATION_RELATIVE
AMT_INCOME_TOTAL	1	0.35	0.34	0.42	-0.063	0.034	-0.023	-0.065	0.14	-0.06	0.13
AMT_GOODS_PRICE	0.35	1	0.99	0.78	0.046	0.084	0.0034	-0.015	0.14	0.036	0.089
AMT_CREDIT	0.34	0.99	1	0.77	0.048	0.083	0.0011	-0.013	0.13	0.034	0.084
AMT_ANNUITY	0.42	0.78	0.77	1	-0.013	0.052	-0.015	-0.039	0.13	0.025	0.099
EMP_AGE	-0.063	0.046	0.048	-0.013	1	0.24	0.27	0.33	0.078	0.18	0.026
EMP_IN_YRS	0.034	0.084	0.083	0.052	0.24	1	0.067	0.13	0.074	0.097	-0.0048
DAYS_ID_PUBLISH	-0.023	0.0034	0.0011	-0.015	0.27	0.067	1	0.1	0.041	0.11	0.0046
DAYS_REGISTRATION	-0.065	-0.015	-0.013	-0.039	0.33	0.13	0.1	1	0.053	0.093	0.045
EXT_SOURCE_2	0.14	0.14	0.13	0.13	0.078	0.074	0.041	0.053	1	0.076	0.19
EXT_SOURCE_3	-0.06	0.036	0.034	0.025	0.18	0.097	0.11	0.093	0.076	1	-0.01
REGION_POPULATION_RELATIVE	0.13	0.089	0.084	0.099	0.026	-0.0048	0.0046	0.045	0.19	-0.01	1

As we can see from above correlation heatmap,
Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa. Credit amount is higher to densely populated area. The income is also higher in densely populated area.

Multivariate Analysis: for target 1

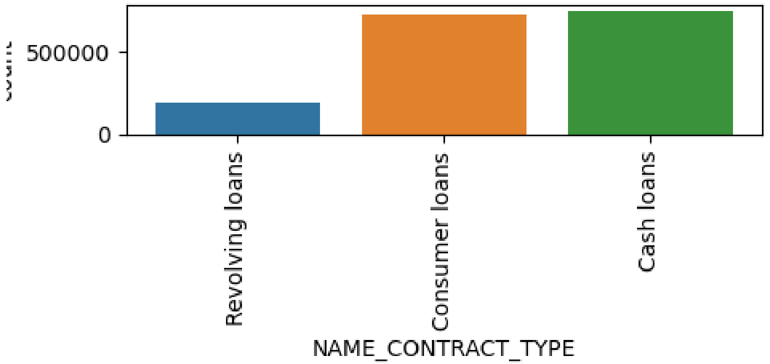


Correlation matrix for Non-Payment Difficulties

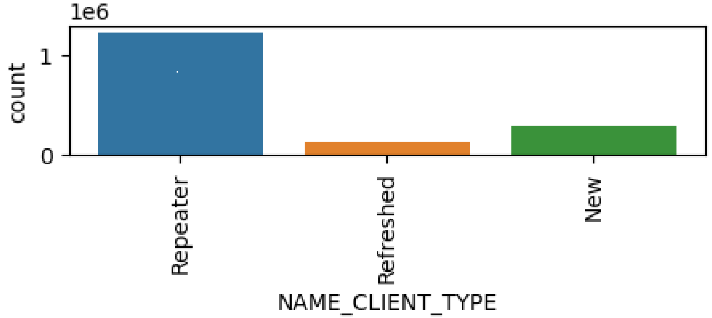
	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	AMT_CREDIT	AMT_ANNUITY	EMP_AGE	EMP_IN_YRS	DAYS_ID_PUBLISH	DAYS_REGISTRATION	EXT_SOURCE_2	EXT_SOURCE_3	REGION_POPULATION_RELATIVE
AMT_INCOME_TOTAL	1	0.33	0.32	0.4	0.003	0.02	0.0035	-0.038	0.14	-0.077	0.096
AMT_GOODS_PRICE	0.33	1	0.98	0.75	0.14	0.11	0.056	0.026	0.13	0.053	0.066
AMT_CREDIT	0.32	0.98	1	0.75	0.14	0.1	0.052	0.026	0.12	0.053	0.058
AMT_ANNUITY	0.4	0.75	0.75	1	0.015	0.04	0.016	-0.034	0.12	0.033	0.058
EMP_AGE	0.003	0.14	0.14	0.015	1	0.29	0.25	0.29	0.11	0.13	0.047
EMP_IN_YRS	0.02	0.11	0.1	0.04	0.29	1	0.11	0.13	0.088	0.057	0.015
DAYS_ID_PUBLISH	0.0035	0.056	0.052	0.016	0.25	0.11	1	0.096	0.059	0.086	0.018
DAYS_REGISTRATION	-0.038	0.026	0.026	-0.034	0.29	0.13	0.096	1	0.07	0.071	0.047
EXT_SOURCE_2	0.14	0.13	0.12	0.12	0.11	0.088	0.059	0.07	1	0.049	0.16
EXT_SOURCE_3	-0.077	0.053	0.053	0.033	0.13	0.057	0.086	0.071	0.049	1	-0.015
REGION_POPULATION_RELATIVE	0.096	0.066	0.058	0.058	0.047	0.015	0.018	0.047	0.16	-0.015	1

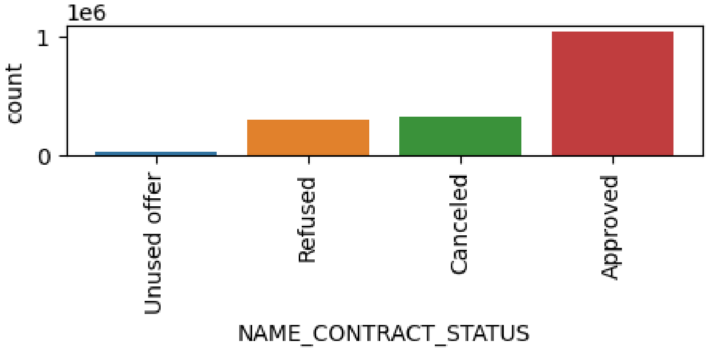
Age, income and employment years have negative correlation

Univariate analysis on previous applications



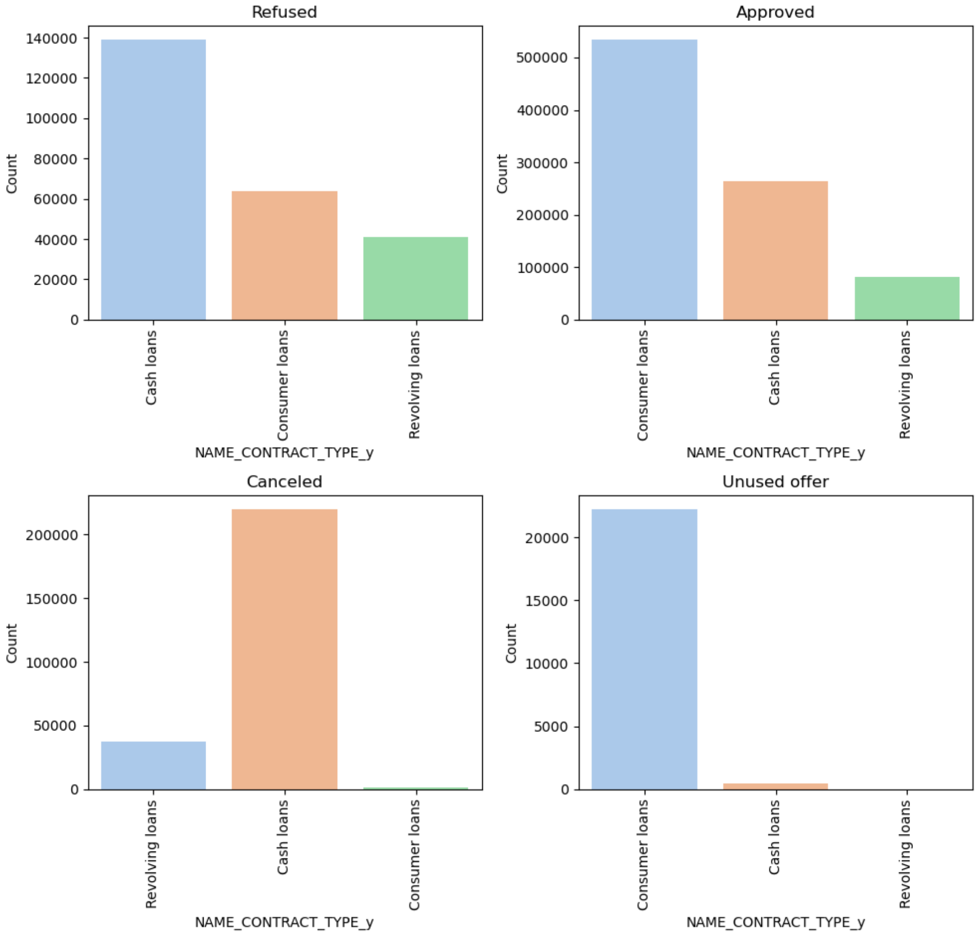
A large proportion of the loan applicants are repeat customers and the most commonly requested type of loan is 'Cash Loans'. Additionally, a majority of the loan applications are approved while only a small percentage of them are left unused.





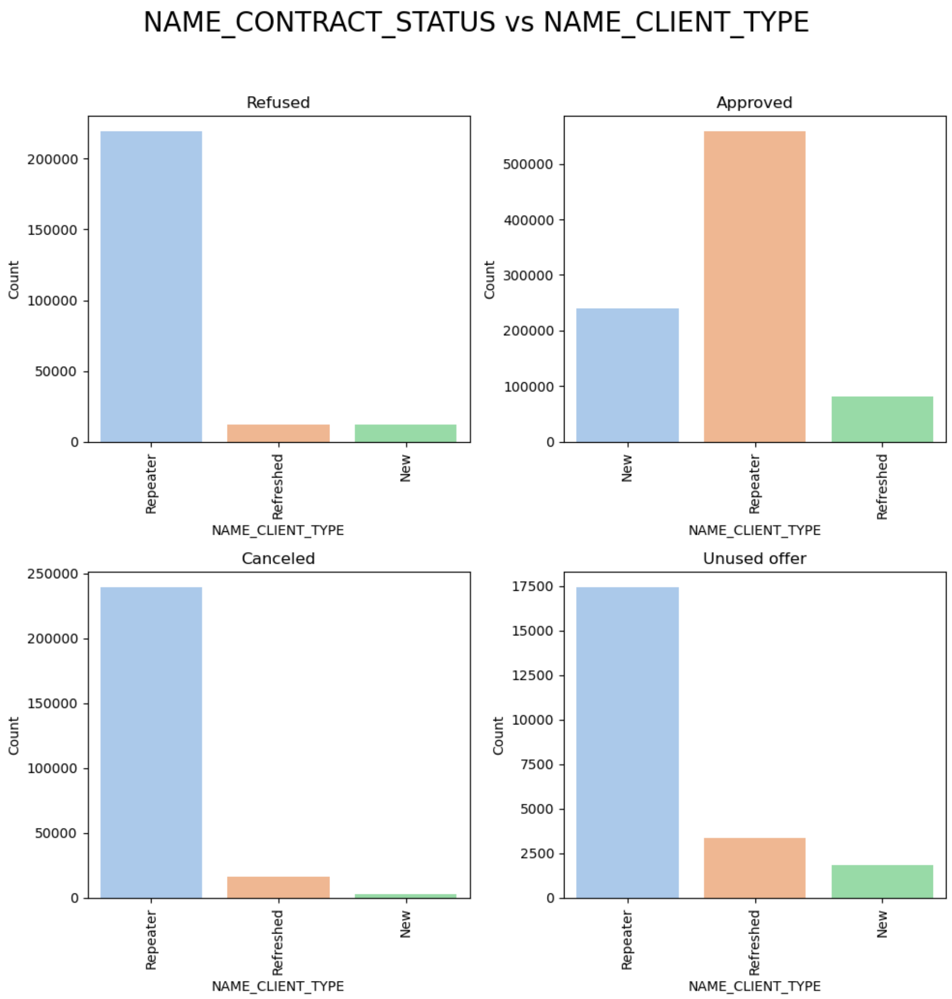
Analysis on Merged data

NAME_CONTRACT_STATUS vs NAME_CONTRACT_TYPE_y



- Banks mostly approve Consumer Loans.
- Most of the Refused_ & Cancelled loans are cash loans.

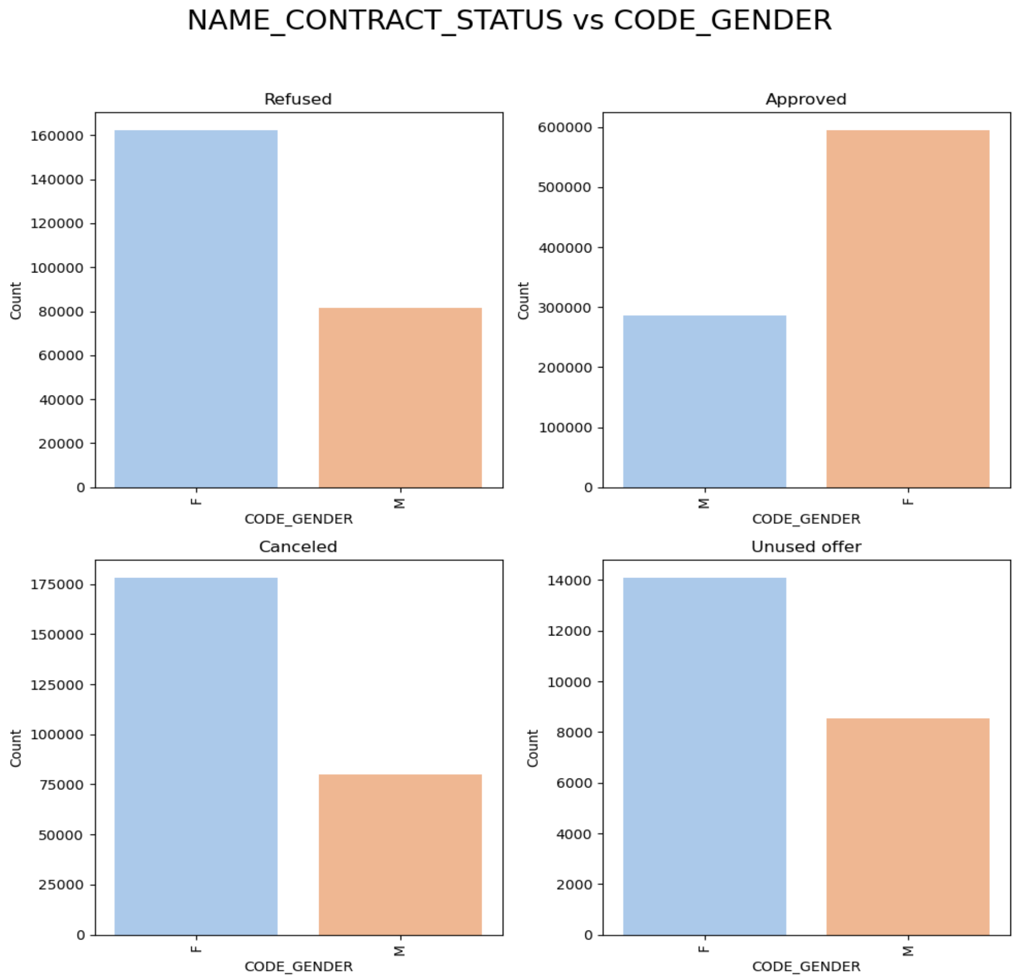
Analysis on Merged data



•The majority of approved, refused, and cancelled loans belong to old clients.

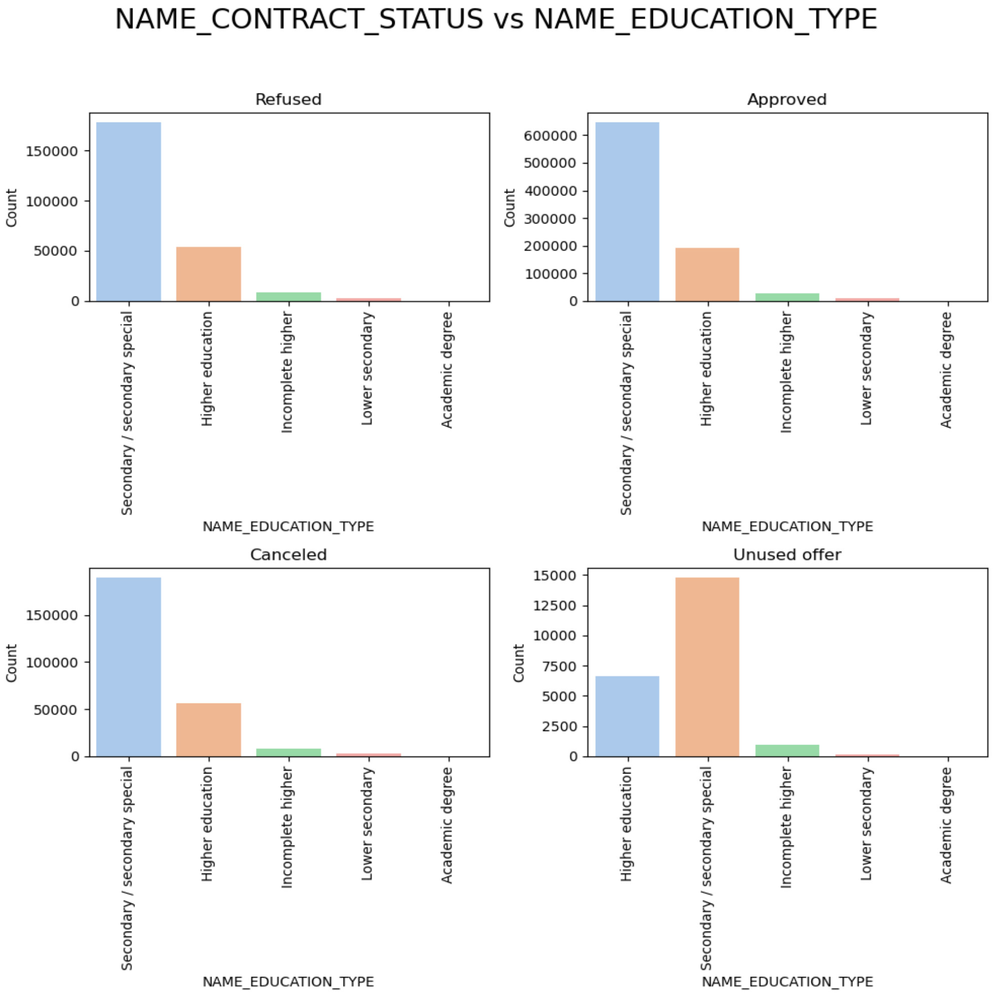
•Approximately 27.4% of loans were provided to new customers.

Analysis on Merged data



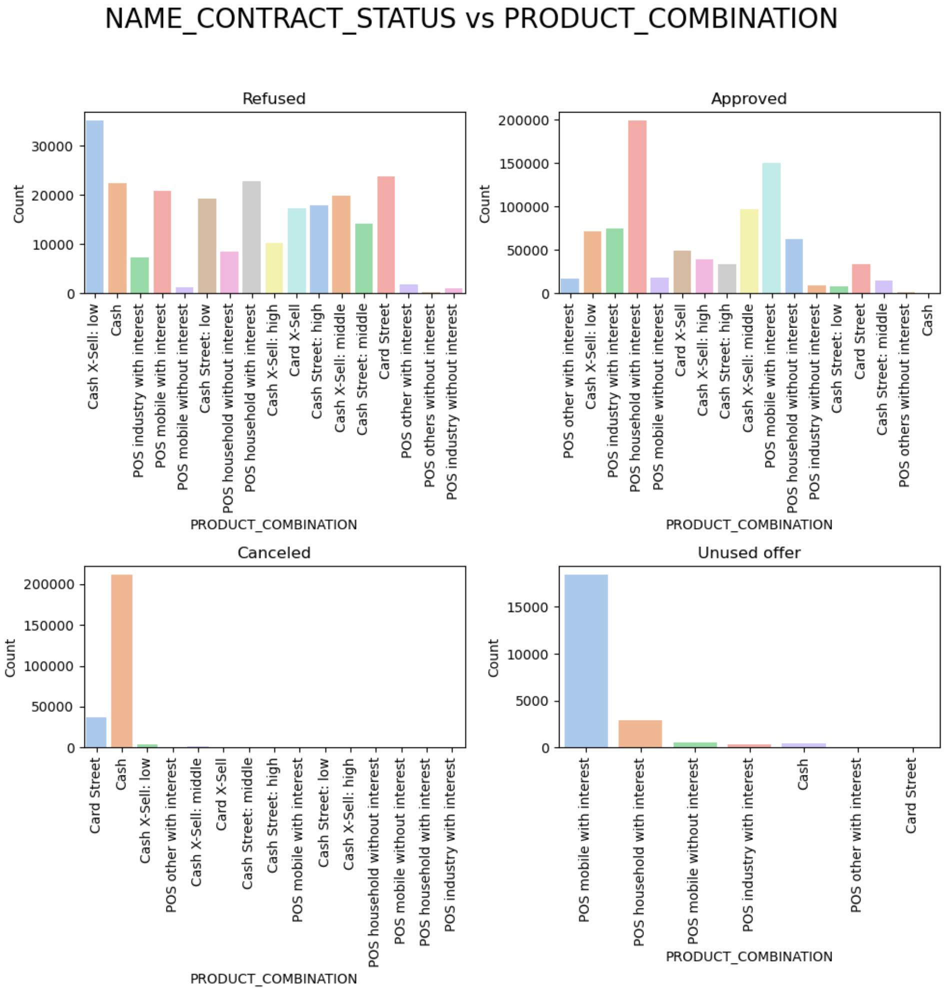
- The percentage of loans approved for females is higher compared to the percentage of loans refused for females.

Analysis on Merged data



The majority of approved loans belong to applicants with a "Secondary / Secondary Special" education type.

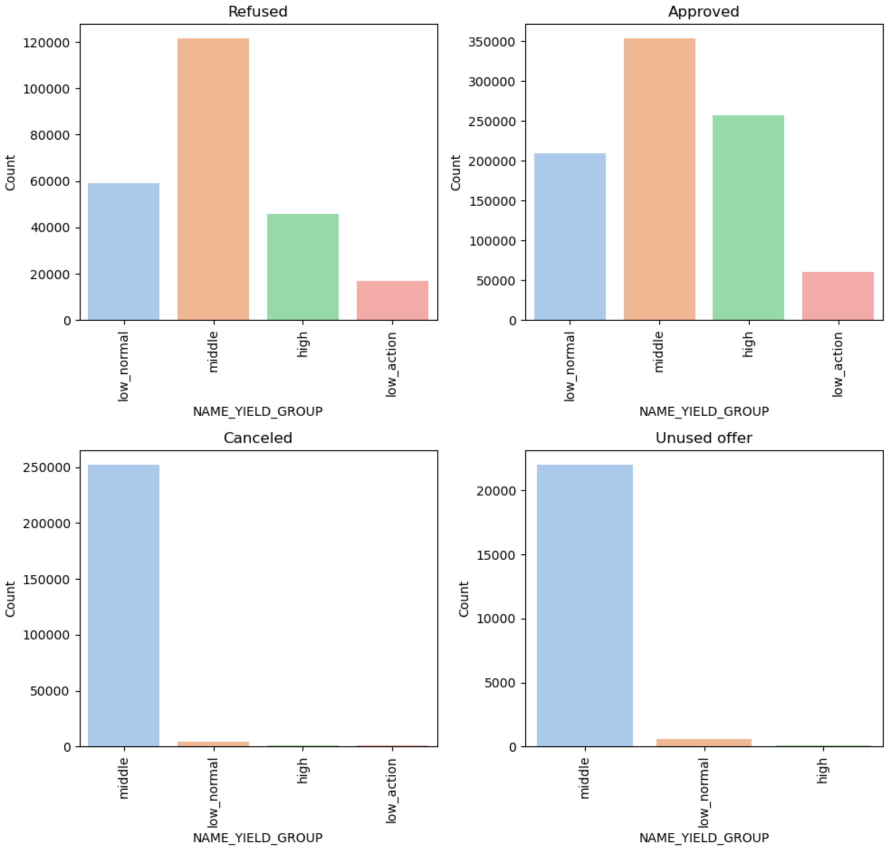
Analysis on Merged data



- The majority of approved loans belong to the "POS household with interest" and "POS mobile with interest" product combinations.
- 15% of refused loans belong to the "Cash X-Sell: low" product combination.
- The majority of canceled loans belong to the "Cash" category.
- 81.3% of unused offer loans belong to the "POS mobile with interest" product combination.

Analysis on Merged data

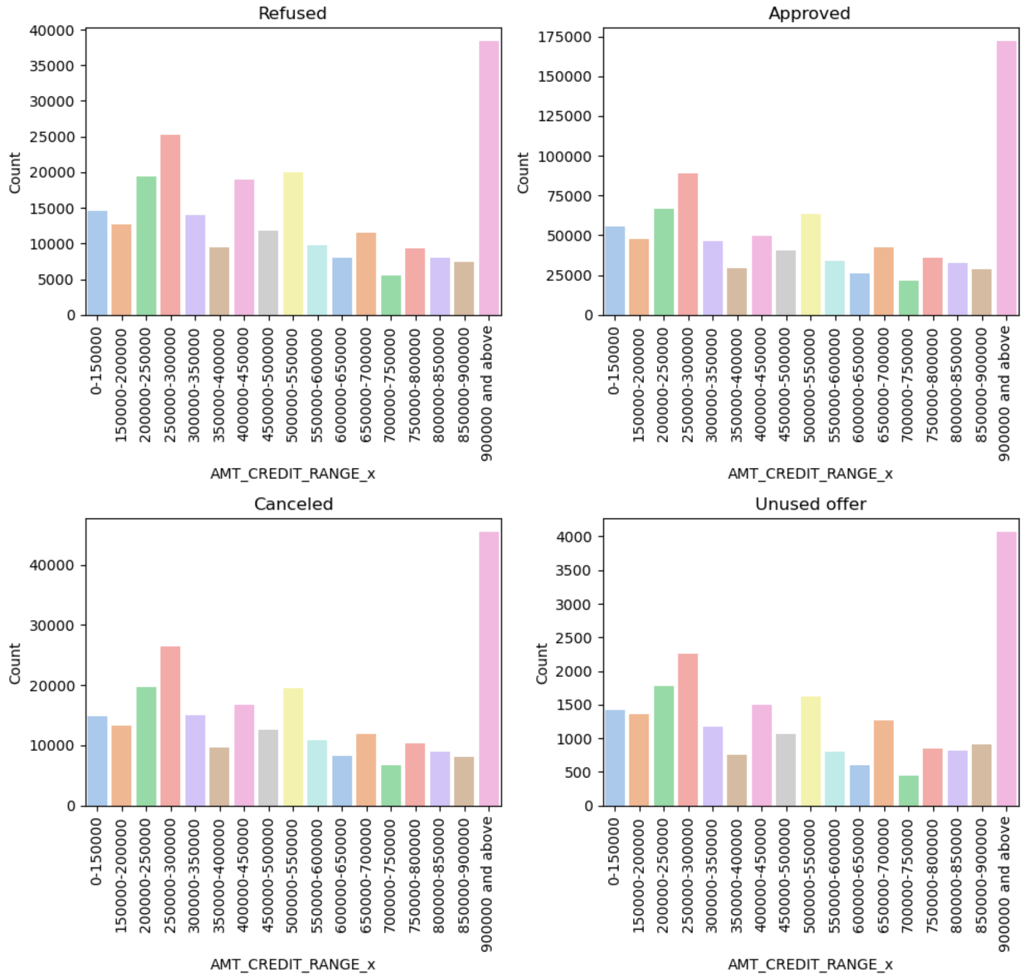
NAME_CONTRACT_STATUS vs NAME_YIELD_GROUP



- The majority of approved loans have a medium grouped interest rate.
- Loans with low or normal interest rates are getting refused or cancelled the most.

Analysis on Merged data

NAME_CONTRACT_STATUS vs AMT_CREDIT_RANGE_x

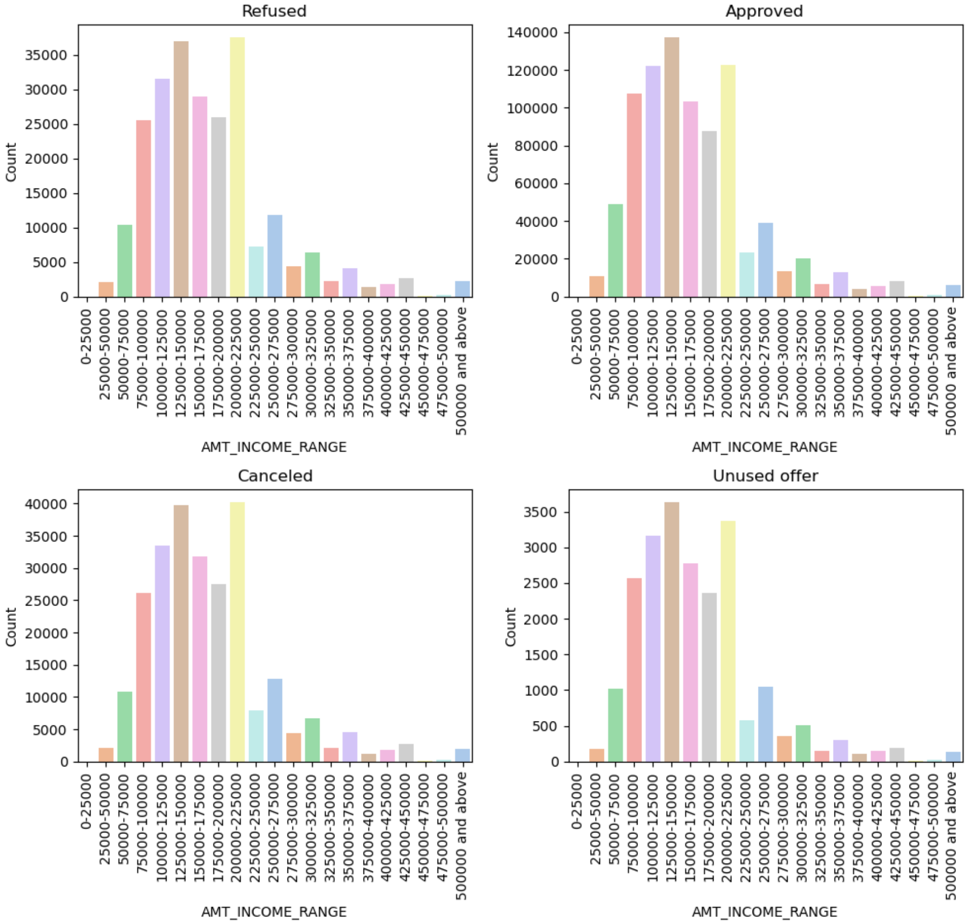


•The majority of approved loans belong to the "Very Low" and "High" credit range.

•Medium and Very Low credit range loans are getting cancelled and rejected the most.

Analysis on Merged data

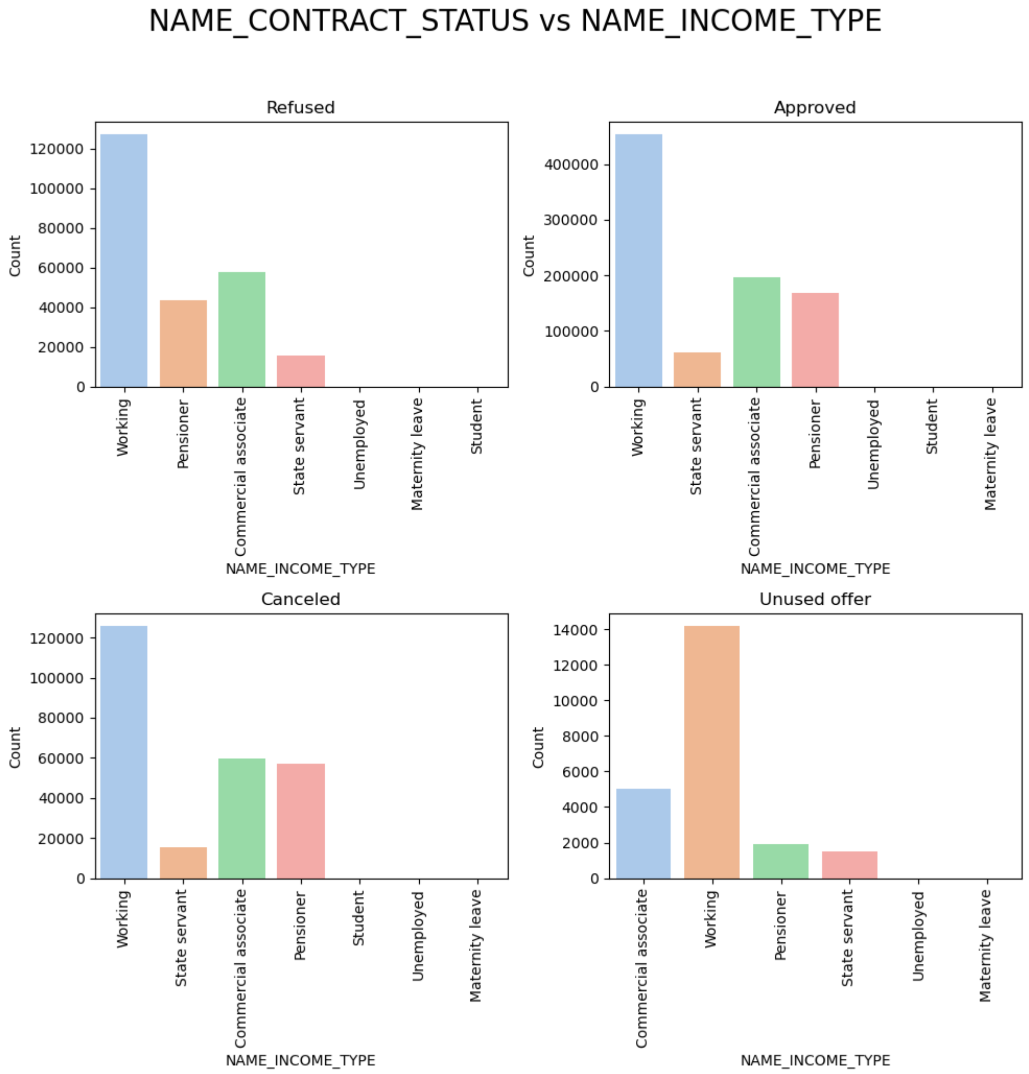
NAME_CONTRACT_STATUS vs AMT_INCOME_RANGE



•The majority of loans are getting approved for applicants with a Low Income range. It could be because they are opting for low credit loans.

•Almost 28% of loan applications are getting rejected or cancelled even though the applicant belongs to a High Income range. It could be because they have requested for a high credit range that is not within their capacity to repay.

Analysis on Merged data



Across all Contract Status (Approved, Refused, Canceled, Unused Offer), people with a Working income type are leading. This suggests that the majority of loans are coming from this income type class.

Insights

The analysis conducted provides several key insights into loan approval patterns at the bank. Consumer loans are the most commonly approved loan type, while cash loans make up the majority of refused and cancelled loans. The bank's customer base is largely made up of existing clients, with only a small percentage of loans provided to new customers. Additionally, a higher percentage of loans are approved for female applicants compared to those who are refused. Loans are most commonly approved for applicants with a Secondary/Secondary Special education type, and for those who are married. Loans previously approved tend to belong to the POS name portfolio. Refused and cancelled loans are most commonly associated with the Credit and cash offices channel type. Furthermore, approved loans tend to have a medium grouped interest rate, with most belonging to the Very Low and High Credit range. The bank also tends to approve loans for applicants with a low income range. Lastly, the Working income type appears to be the most common across all contract statuses. It is important to note that these findings are based on the available data and may not

Unsafe categories

Customers with lower levels of education beyond secondary school tend to default on their loans, particularly when their previous loan applications have been terminated or rejected. Additionally, men who are in a civil union fall into a group that has a higher likelihood of loan rejection. This group includes those who have had their loan applications rejected in the past.