Credit EDA Assignment

Name : Upeksh Naik

Batch : Data Science Program - July 2024

Assignment : Credit EDA Assignment

Data Understanding.

In the provided data (csv files) , the data is about the application of the customers for the loans and the loans are divided into two parts .

- Current Application for loan.
- Previous Application for loan.

Data is containing the information about the customers which have age, count of family members, and many info.

There are more 40% columns have null values and having unwanted data.

There are many unwanted columns which doesn’t have importance in getting valid customers.

Also Both the data are merged with each other using left join and merge function.

And columns have data like credit loan, previous loan tenure for repayment, goods price, income type of customers, and many things.

Many columns have more than 10lakhs+ data are missing values.

The dataset have 1430155 columns and 157 rows

Data Cleaning.

Univariate Analysis

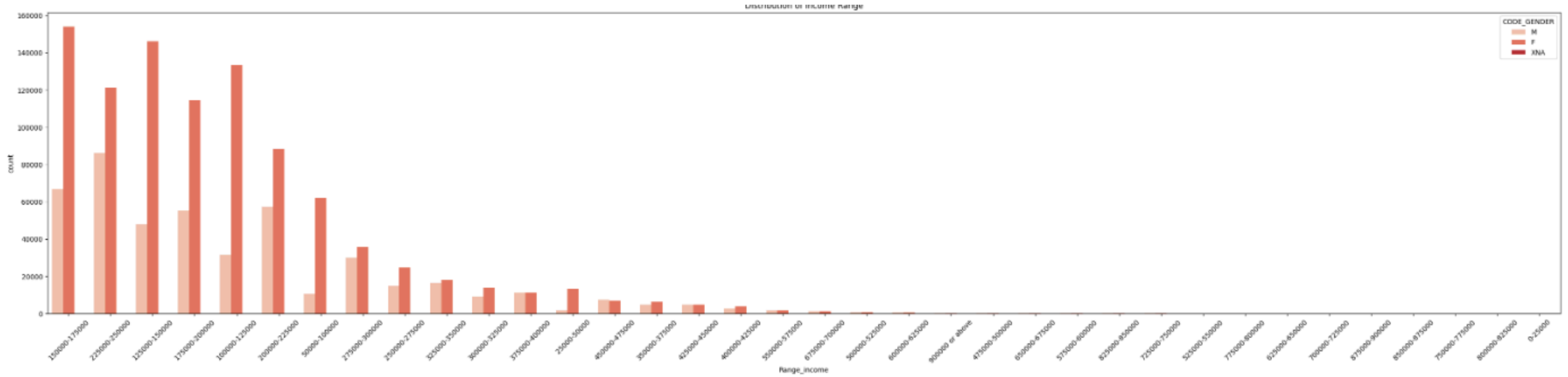
Data Cleaning.

- Converted the days columns into years by dividing it by -365 and changed their datatypes to int64.
- Deleted the days columns.
- Replaced the null values in AMT_ANNUITY of current and previous data by mean of the column.
- Replaced the null values in AMT_CREDIT of current and previous data by mean of the column.
- Replaced the null values in AMT_GOODS_PRICE of current and previous data by mean of the column.
- Replaced the null values in NAME_TYPE_SUITE of current and previous data by mode of the column.
- Replaced the null values in NAME_CONTRACT_TYPE of current and previous data by mode of the column.
- Replaced the null values in OWN_CAR_AGE of current and previous data by 0.
- Replaced the null values in OCCUPATION_TYPE of current and previous data by mode of the column.
- Replaced the null values in EXT_SOURCE_1,EXT_SOURCE_3,EXT_SOURCE_2 of current data by mean of the column.
- Dropped columns like SK_ID_PREV and unwanted data columns from the dataset.
- Created a function to analyze the defaulter using the observation columns in which it will say default to the customer who has more than 50% of negative responses by their social surrounding.
- Dropping the columns which has more than 50% null values .and missing values.
- Converting the columns which are matching with int63 and int32 by numerical.

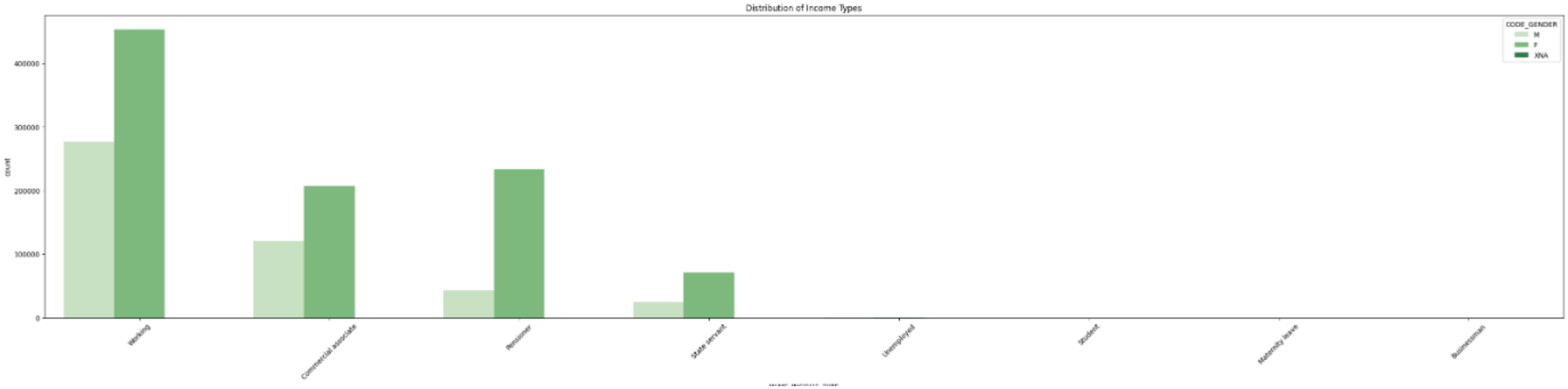
Bivariate and Multivariate Analysis

Data Cleaning.

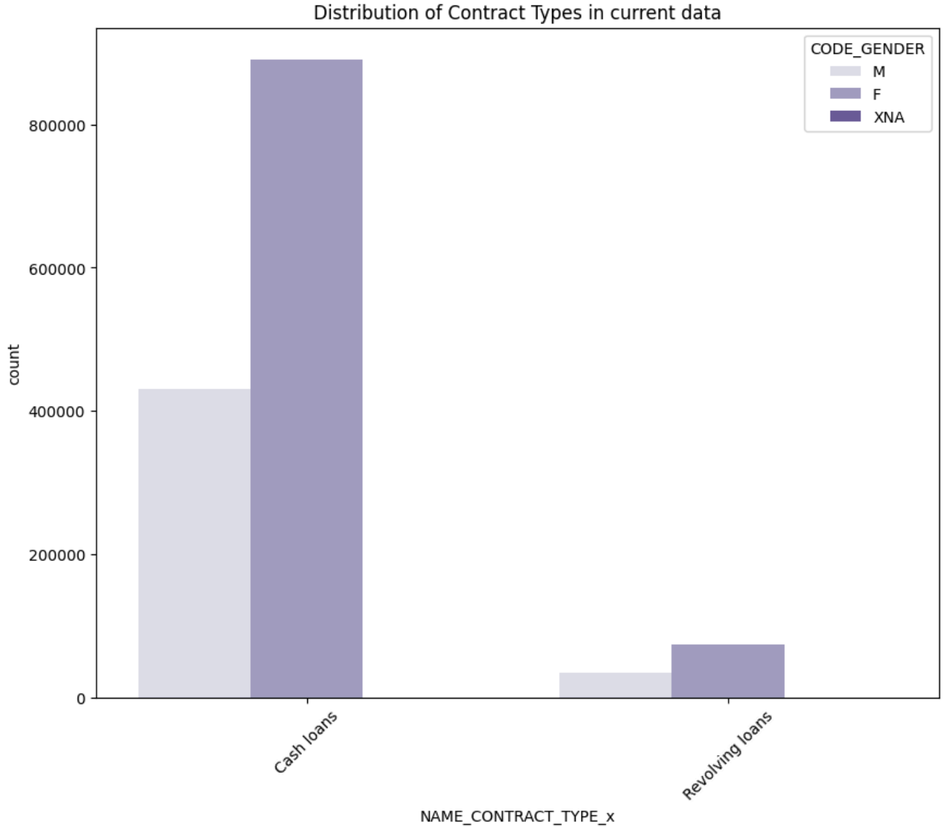
- Created and assigned the income columns to the ranges of incomes.
- Ranges are from 0 to 900000 and more.
- Data visualization using the boxplot and barplots.
- And Correlation between credit loan of previous and current application
- Also created heatmap correlation
- Created the purpose of loans for the customers and also granted loan by bank
- Outliers in the Amount Credit in previous and current application
- Removed the outliers



- The Above **Bar Plot** represents the Distribution of **Income Ranges** by dividing into Males and Females and also including the XNA.
- In the above plotted graph the **light orange bars represents the males group** and **orange bars represents the females group.**
- The **x-axis** is the count of the peoples and **y-axis** is the ranges from **0-900000 and above incomes ranges**.

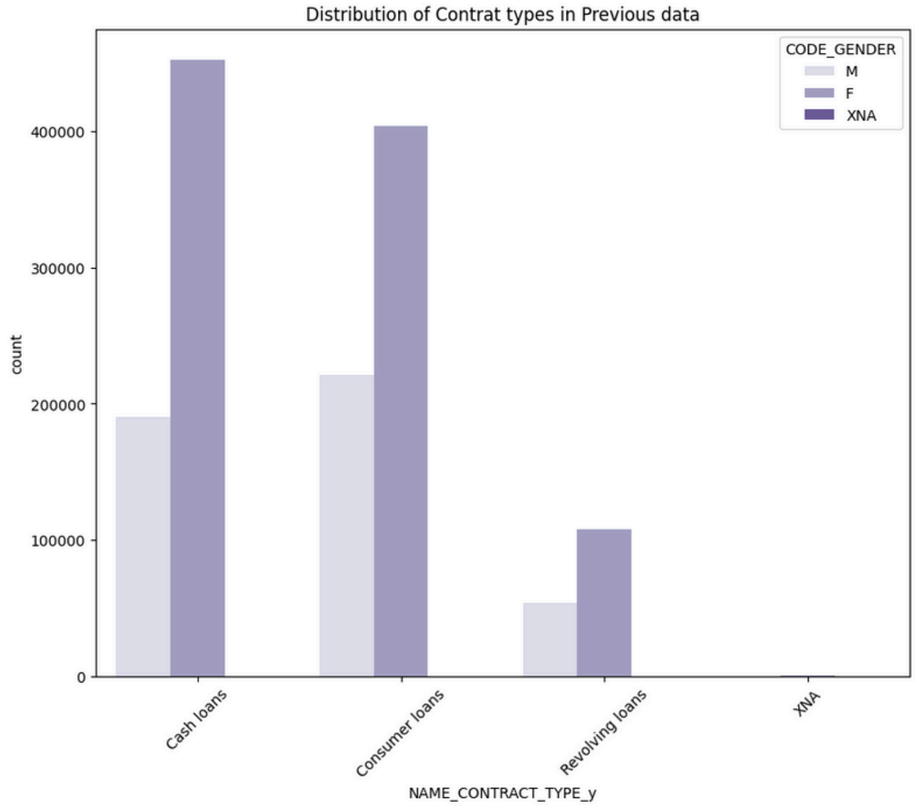


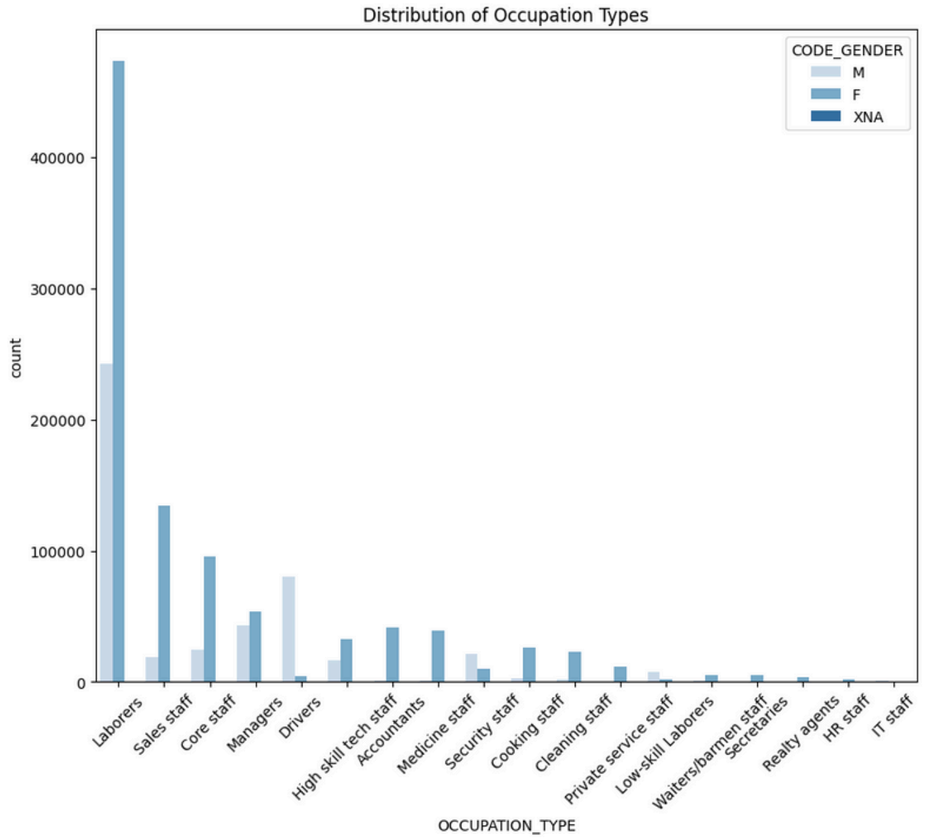
- The Above **Bar Plot** represents the Distribution of **Income Types** by dividing into Males and Females and also including the XNA.
- In the above plotted graph the **light green bars represents the males group** and **green bars represents the females group.**
- The **x-axis** is the count of the peoples and **y-axis** is the **income types.**



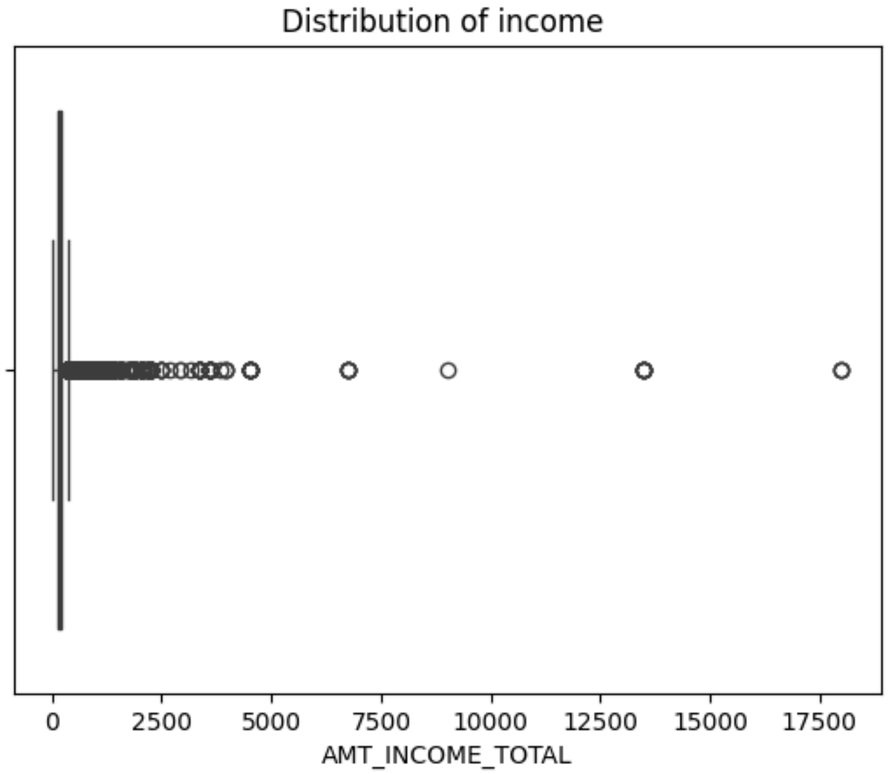
- The Above **Bar Plot** represents the Distribution of **Contract Types in current data** by dividing into Males and Females and also including the XNA.
- In the above plotted graph the **light purple bars represents the males group** and **purple bars represents the females group.**
- The **x-axis** is the count of the peoples and **y-axis** is the **contract types in current data**.

- The Above **Bar Plot** represents the Distribution of **Contract Types in previous data** by dividing into Males and Females and also including the XNA.
- In the above plotted graph the **light purple bars represents the males group** and **purple bars represents the females group.**
- The **x-axis** is the count of the peoples and **y-axis** is the **contract types in previous data**.

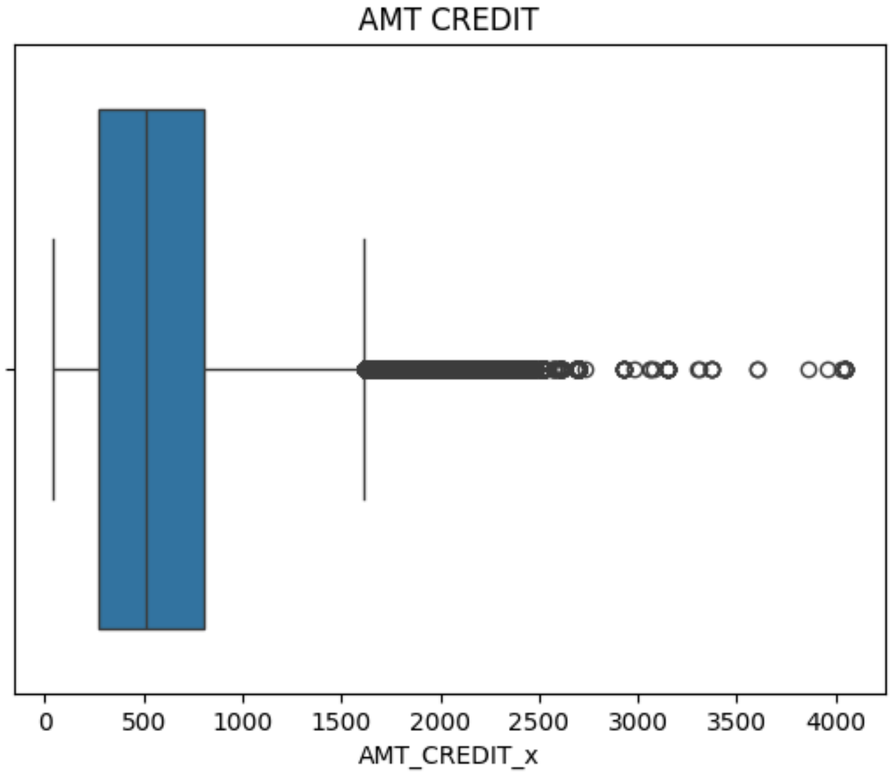




- The Above Bar Plot represents the Distribution of **occupation types** by dividing into Males and Females and also including the XNA.
- In the above plotted graph the **light blue bars represents the males group** and **blue bars represents the females group.**
- The **x-axis** is the count of the peoples and **y-axis** is **occupation types**.

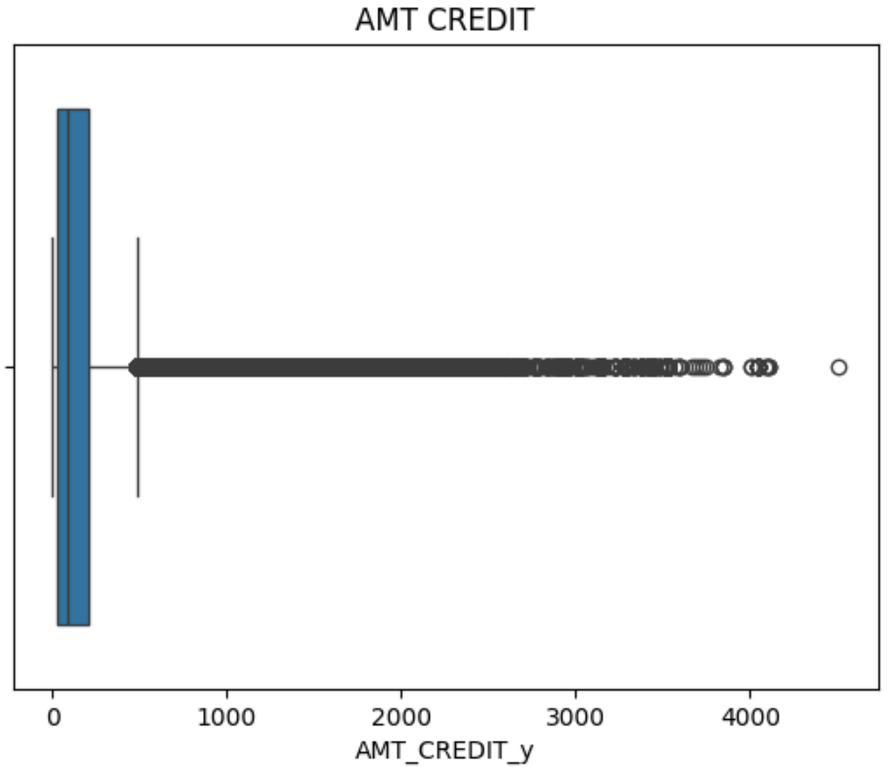


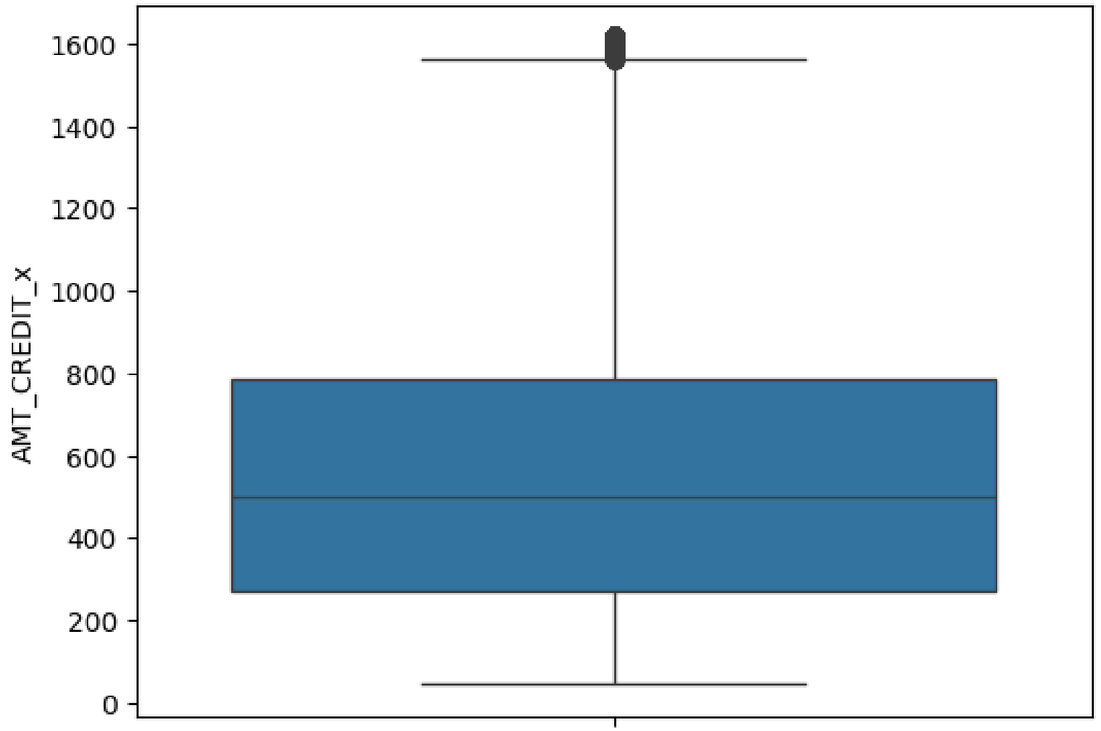
- The Above **boxplot** represents the Distribution of **Income.**
- It is showing the numbers of outliers in total incomes of applied customers for loan.
- **Removing the outliers from the AMT_INCOME_TOTAL.**
- Distribution is based on Total income of the customers.



- The Above **boxplot** represents the Distribution of **loan credit.**
- It is showing the numbers of outliers in loan credit of applied customers for loan in current data.
- **Removing the outliers from the AMT_CREDIT_x.**
- Distribution is based on Total credit loan wanted by customers.

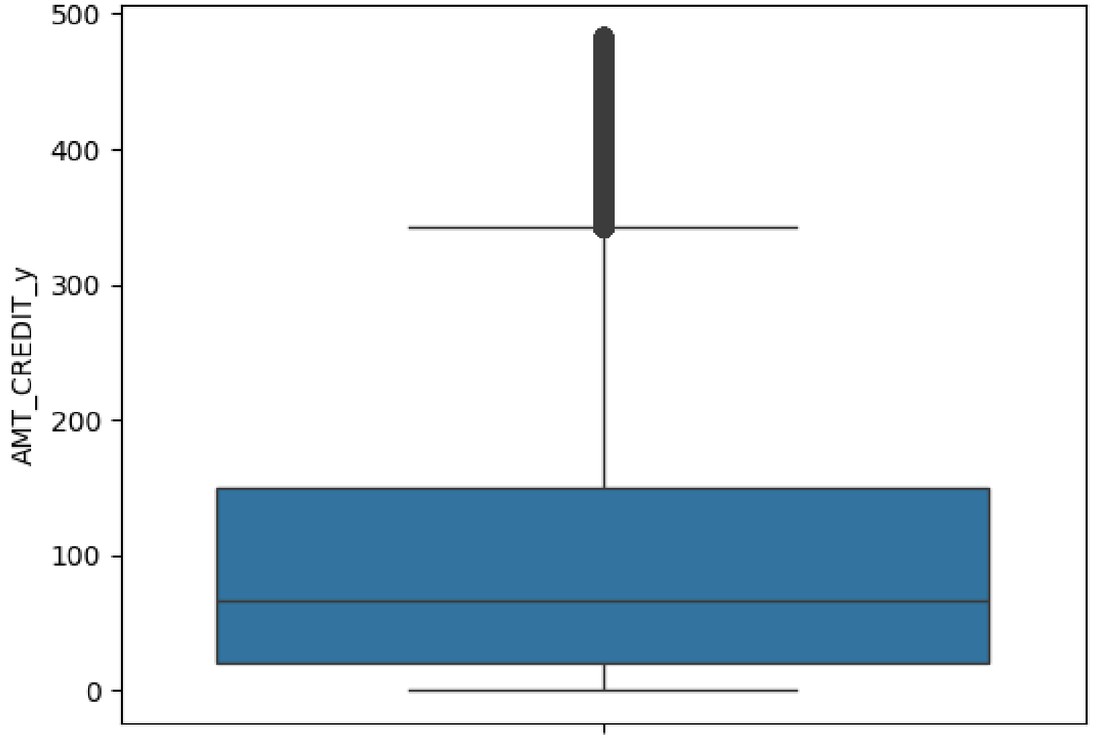
- The Above **boxplot** represents the Distribution of **loan credit.**
- It is showing the numbers of outliers in loan credit of applied customers for loan in previous data.
- **Removing the outliers from the AMT_CREDIT_y.**
- Distribution is based on Total credit loan wanted by customers.

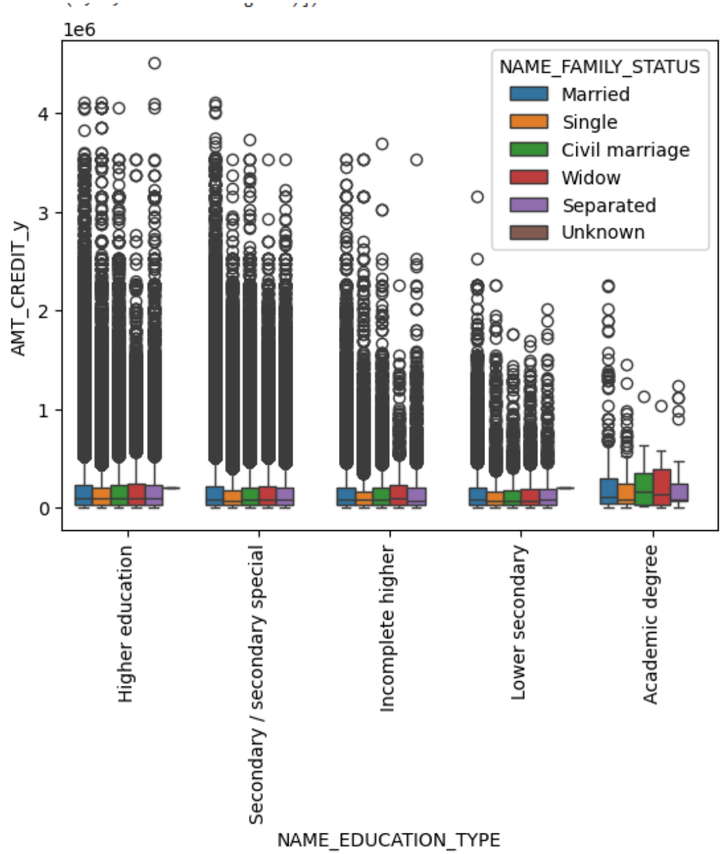




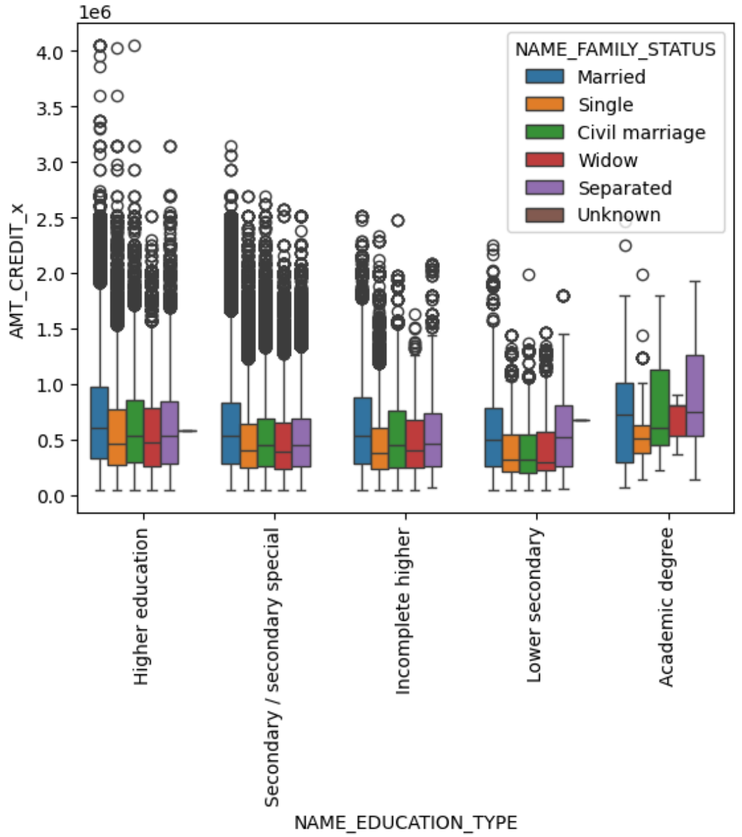
- The outliers have been removed from the **AMT_CREDIT_x.**

- The outliers have been removed from the **AMT_CREDIT_y.**

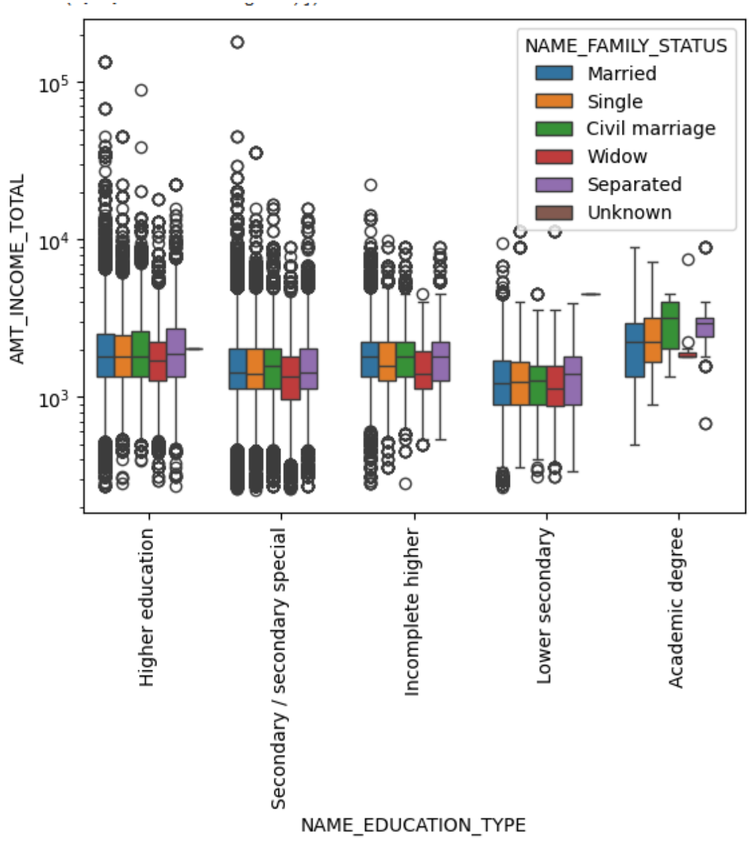




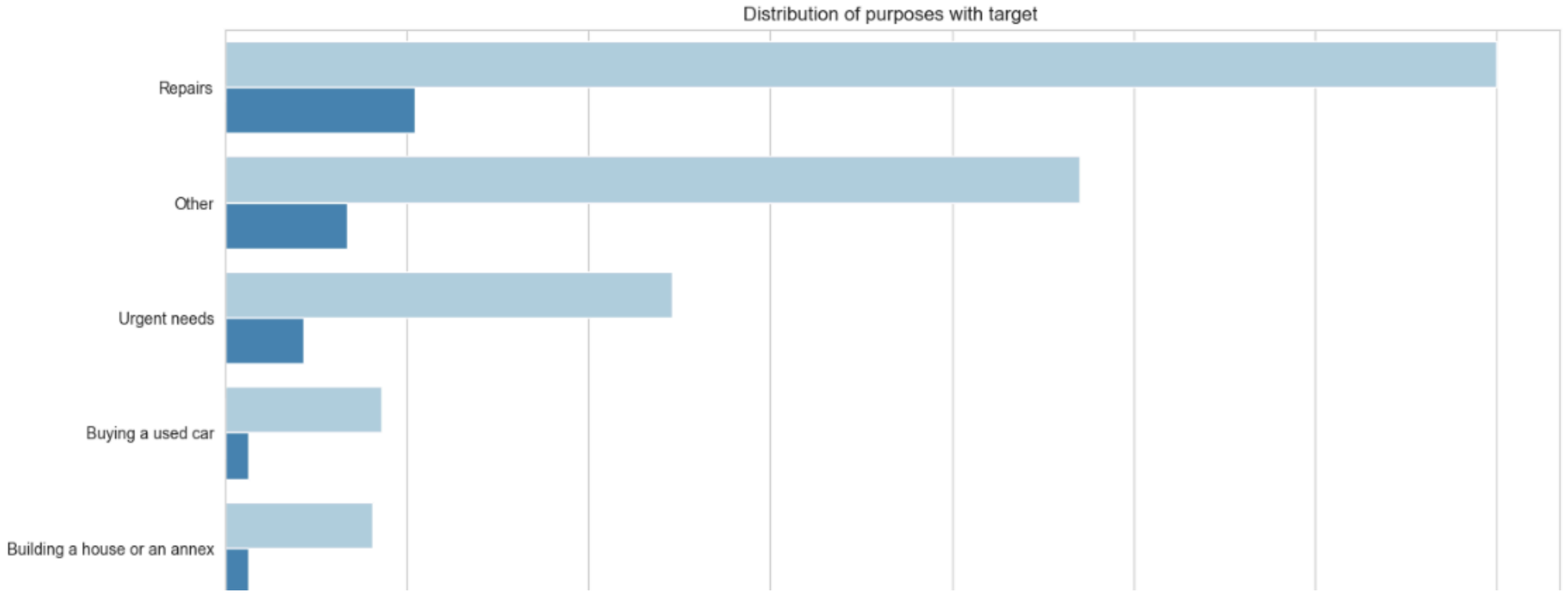
- The Above Bar Plot represents the Distribution of Education types of customers by dividing into family status (Married, Single, civil marriage, widow, separated, unknown) using amount credit loan in previous data.



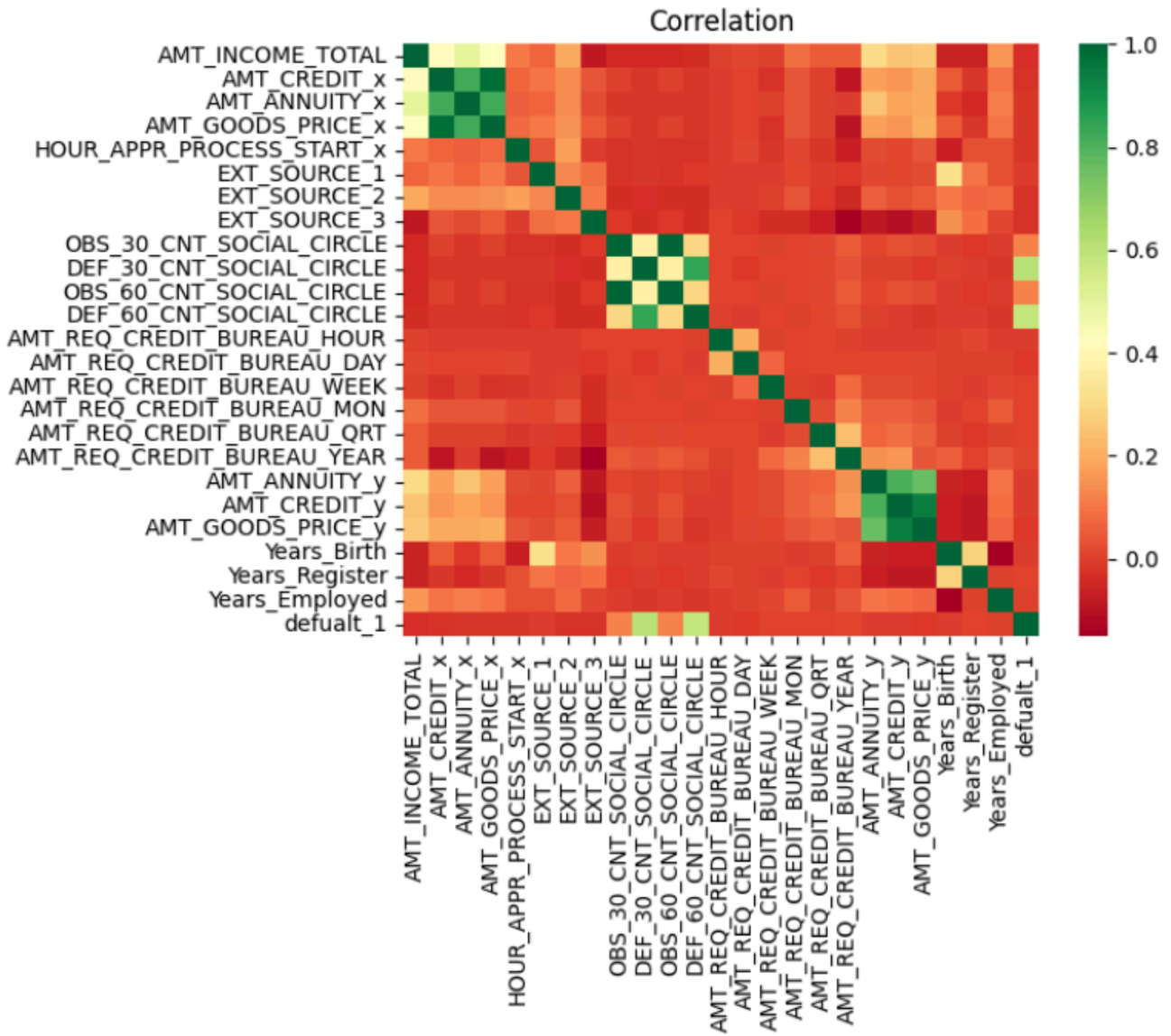
- The Above Bar Plot represents the Distribution of Education types of customers by dividing into family status (Married, Single, civil marriage, widow, separated, unknown) using amount credit loan in current data.



- The Above Bar Plot represents the Distribution of Education types of customers by dividing into family status (Married, Single, civil marriage, widow, separated, unknown) using amount of total income of the applied customers for loan in previous data.



- This is distribution of the purpose of customers for getting loan credit and bank getting the target as well .
- This shows that the customers and bank is lending the money as per the purposes of the customers.
- This help to get to whom does bank lends money or approved the loans.
- Maximum rejection loan by bank is for repairs.



- Heatmap which is represents the correlation between the current and previous application.
- This graph is showing which columns are correlated with each others
- Here we can see that only diagonal part is correlated and some outer diagonals are correlated.

Conclusion

- Now the conclusion is that bank should focus on the purpose of the customers .
- And also observe on the Observation of social of customers and their reviews also.
- Working profession also have less loan approves in Previous Application.
- Repairs purpose is also having least loan approvals in Previous Application.

Thank You