Credit EDA Assignment

Problem Statement

The loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history Suppose you work for a consumer finance company specializing in lending various types of loans to urban customers. You have to use EDA to analyze 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 on 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 for 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.
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 the 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.

Presented by
Aseem Pasha

Overall Approach to the Eda:

- 1. **Understanding Domain Variables using Data Dictionary given:**

  'columns_description.csv' is data dictionary which describes the meaning of the variables.

- **2. Importing Necessary Libraries such as Numpy,Pandas,Matplotlib,seaborn**

  Numpy & Pandas:It helps Loading the Data & any scientific computation

  Seaborn & Matplotlib: It helps in visualizing univariate and bivariate data

- **3. Importing the Data file**
- **4. Checking the structure of Data**

  No of Columns & Rows present in the Dataframe

  Type of Data present i,e:int/float

  Checking Numerical variables i,e:Count,Max,Min,std etc

- **5. Data Cleaning & Manipulation**

  Identifying the percentage of Missing value of each column

  Deleting Columns having missing values above 40% which are not related to the outcome

- **6.Imputing values**

  If the Column is categorical Column we impute with Mode

  If the column is continuous Column we impute with Median

- **7.Segmentation**

  Separating important Columns into Categorical ,Continuous, Id Columns for ease in performing looping Condition to analyze all columns

- **8. Analysis**

  Univariate Analysis

  Histogram :It is used to see the bucket-wise frequency distribution of a continuous variable

  Box plot: Displays the minimum, first quartile, median, third quartile, and maximum.

Distribution plot:The distribution plot is suitable for comparing range and distribution for groups of numerical data.Pie chart:A pie chart is graph that represents the data in the circular graph.

- **Bi-Variate Analysis**

  Scatter plot:Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another.

  Bar Plot:A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

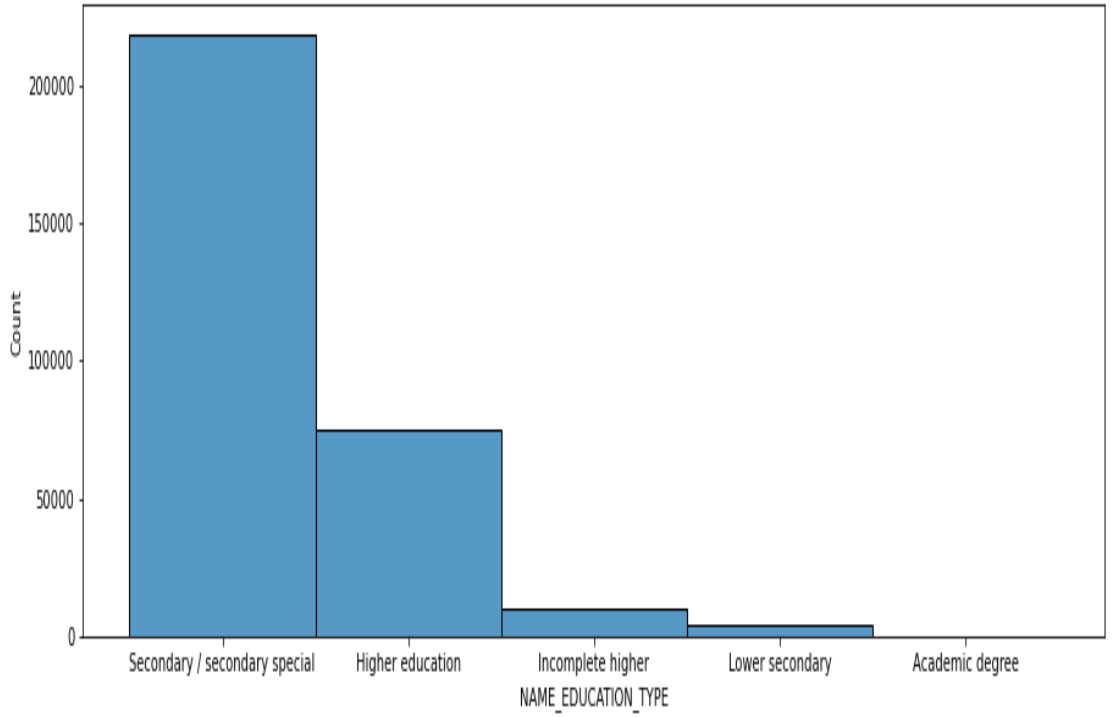
  Box plot

- **Multivariate Analysis**

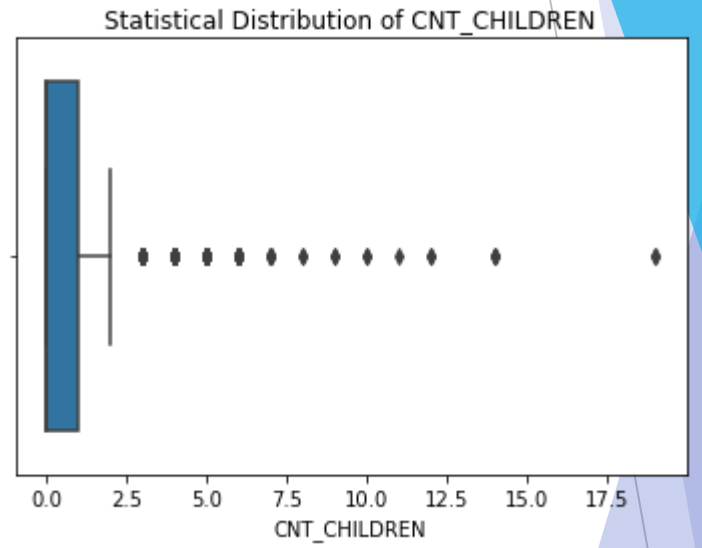
  Heat Map:A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colours

  Pair Plot:Plot pairwise relationships in a dataset

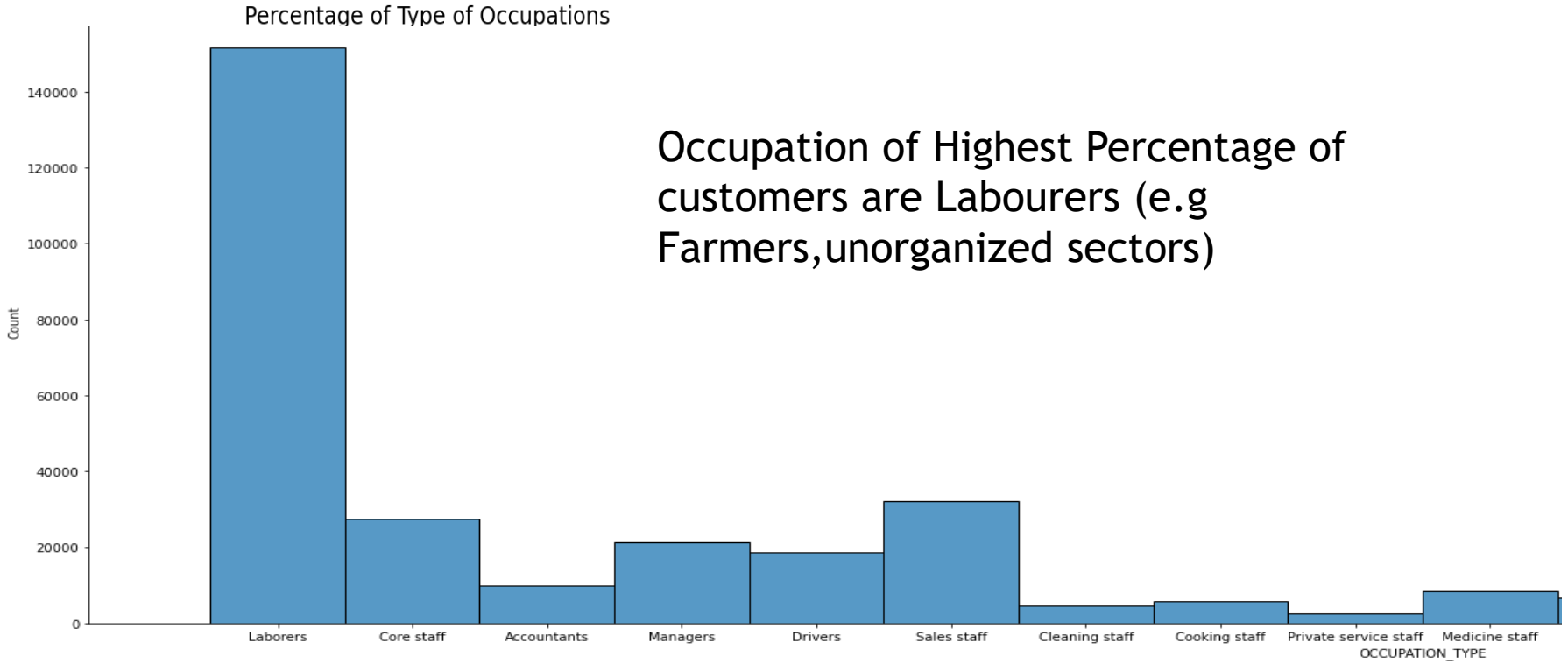
Univariate Analysis



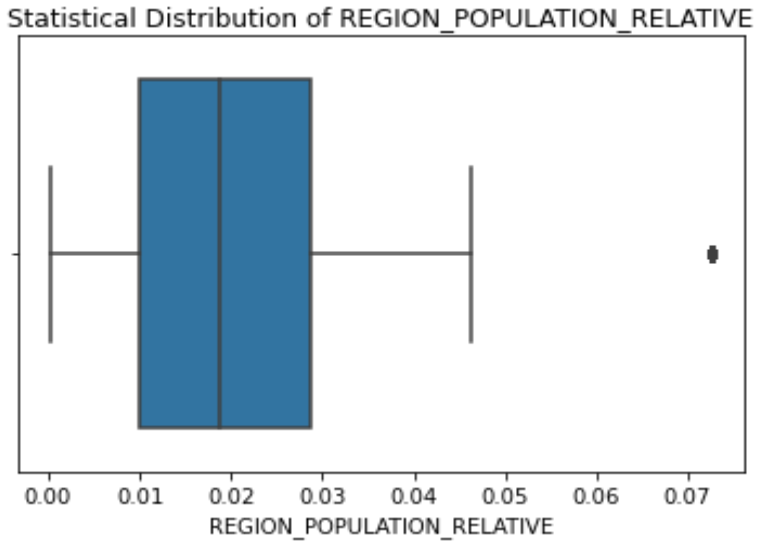
Most of the customers applying for Loan have completed their Secondary Education



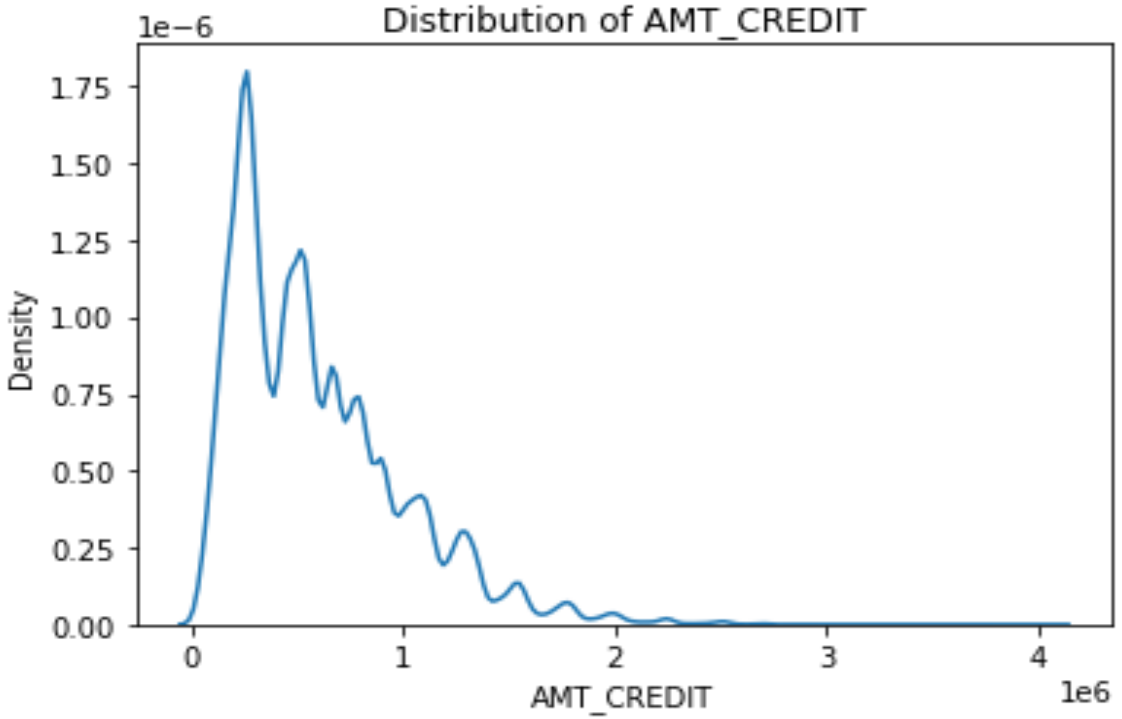
The customers applying for loan have mostly 1 children but there are outliers as well



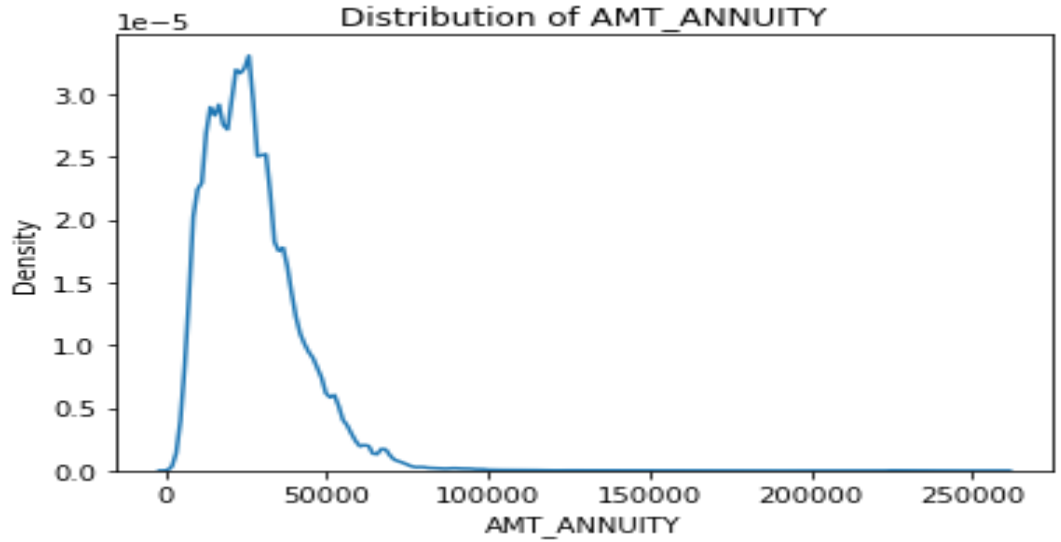
Occupation of Highest Percentage of customers are Labourers (e.g Farmers,unorganized sectors)



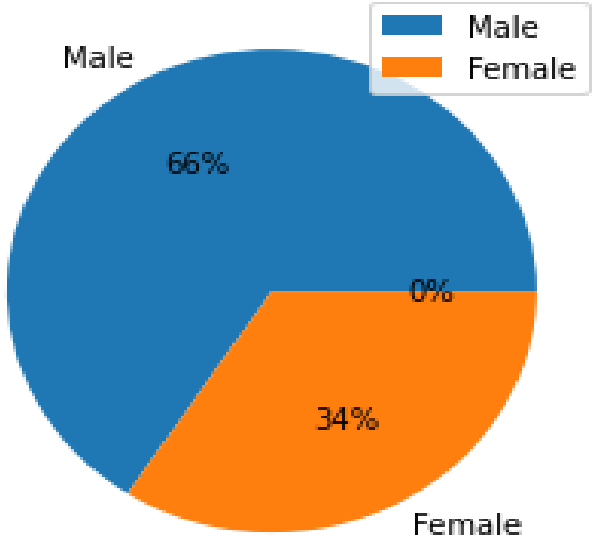
Customers applying for loan are not from highly populated regions



The Credit Amout of Loan Mostly ranges from 25,000 to 1 Lakh rupees



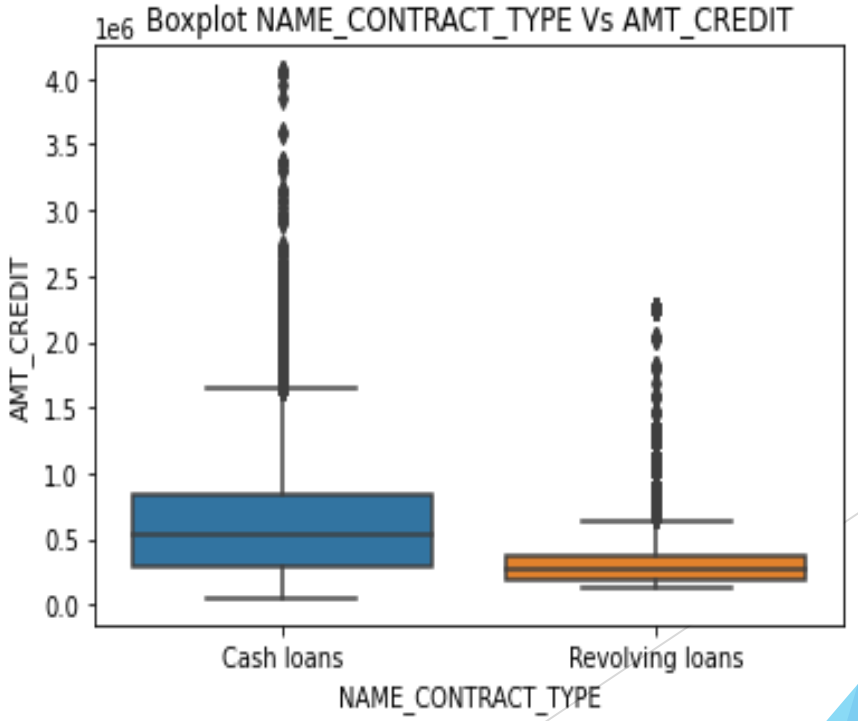
Annuity loans are a type of loan that is repaid in monthly installments over the course of many years and the range of Loan Amount ranges upto 50,000/-
Eg:Houshold,Daily usage products on Emi etc

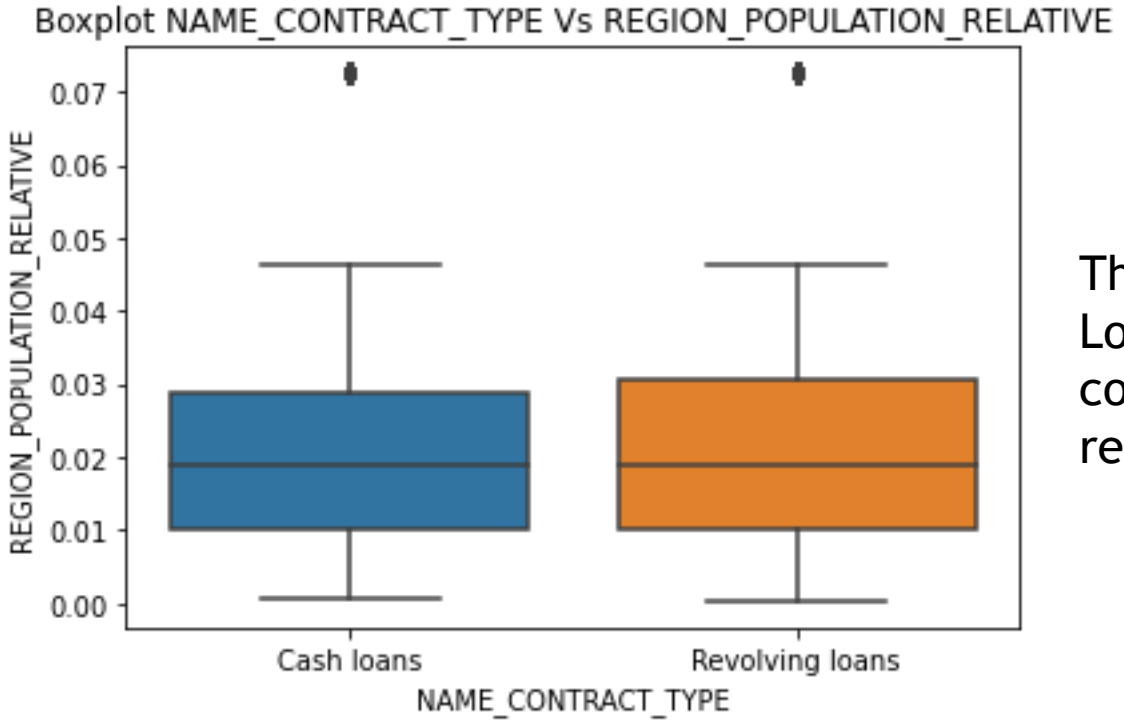


66 percent of Customers applying for Loan are Male

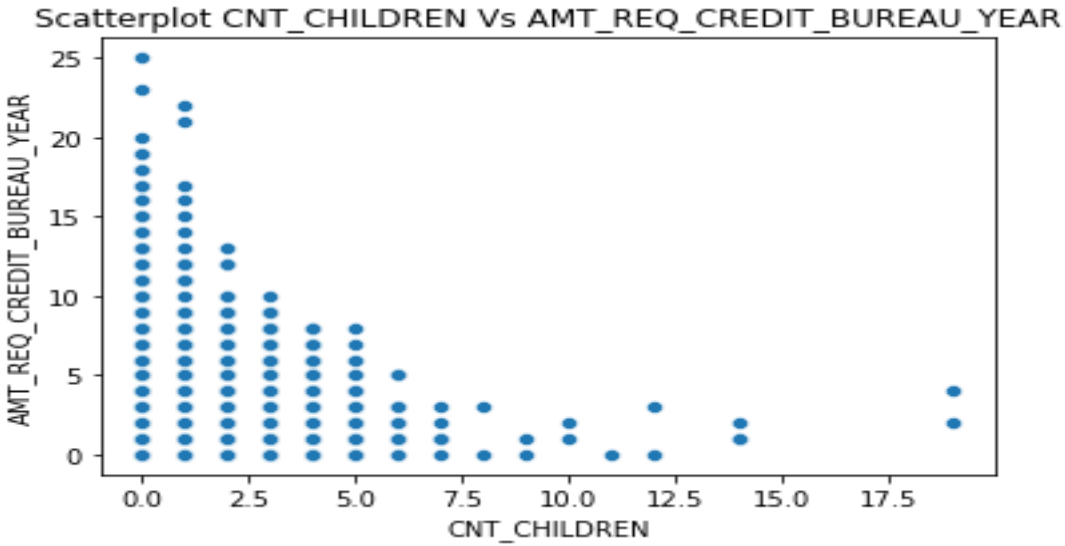
Bivariate Analysis

The Loan Amount of cash Loans is more than the revolving loans
Daily used credit cards is a good Example of revolving loans

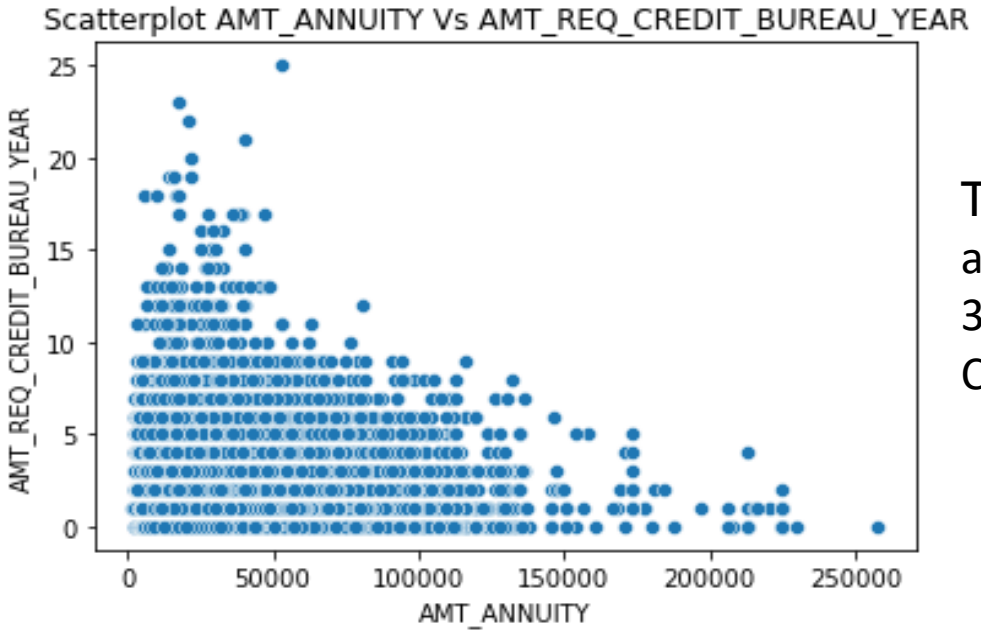




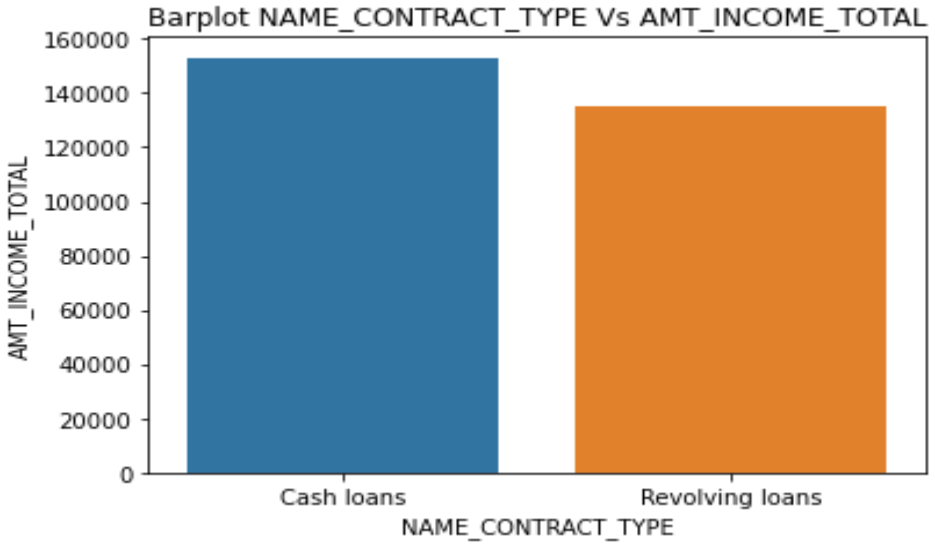
The people using revolving Loans are slightly more when compared to the populated region



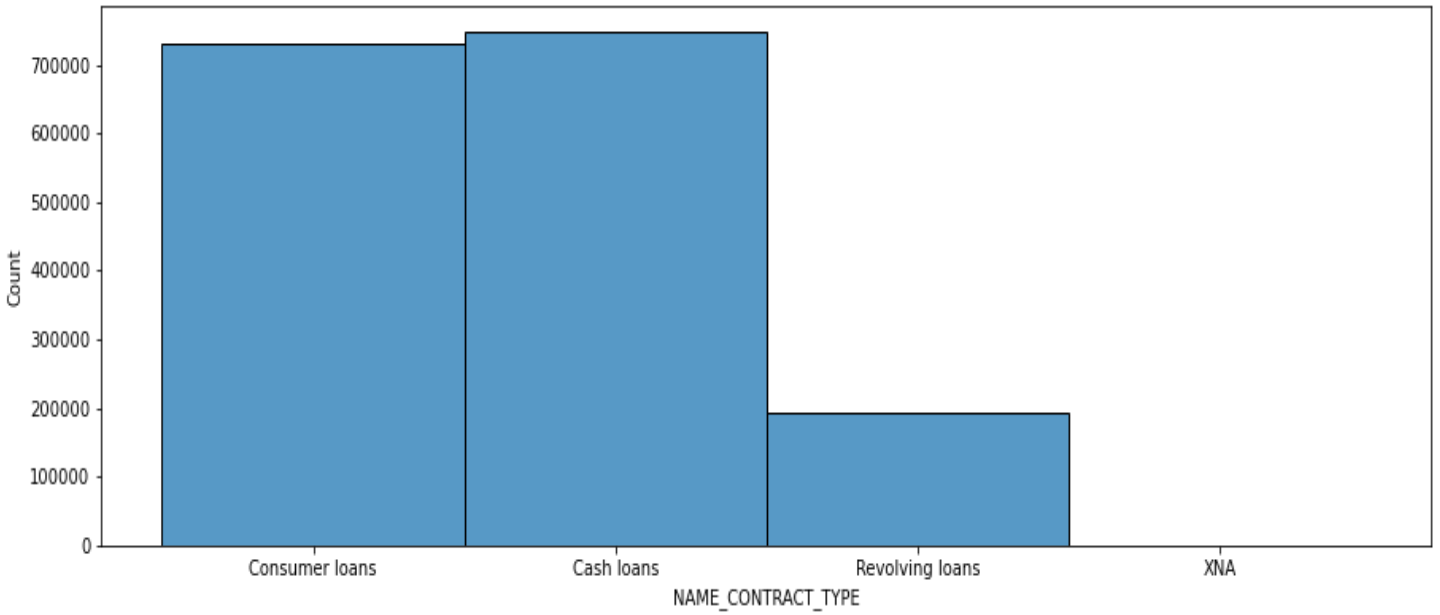
Number of enquiries to Credit Bureau about the client reduces with higher the count of children's



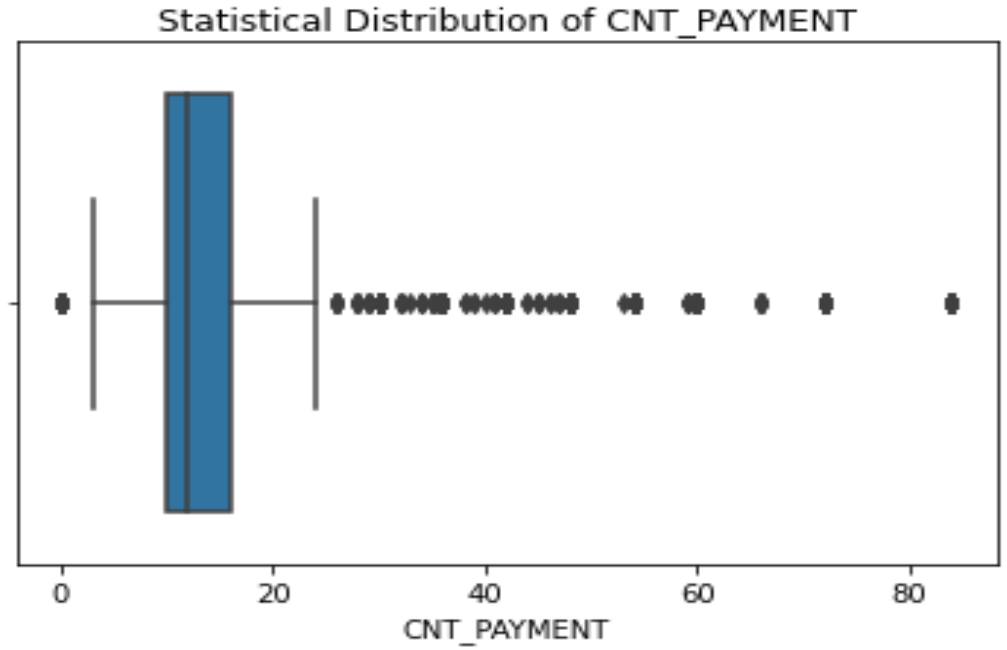
The Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application) decreases for Clients as the Loan Amount increases



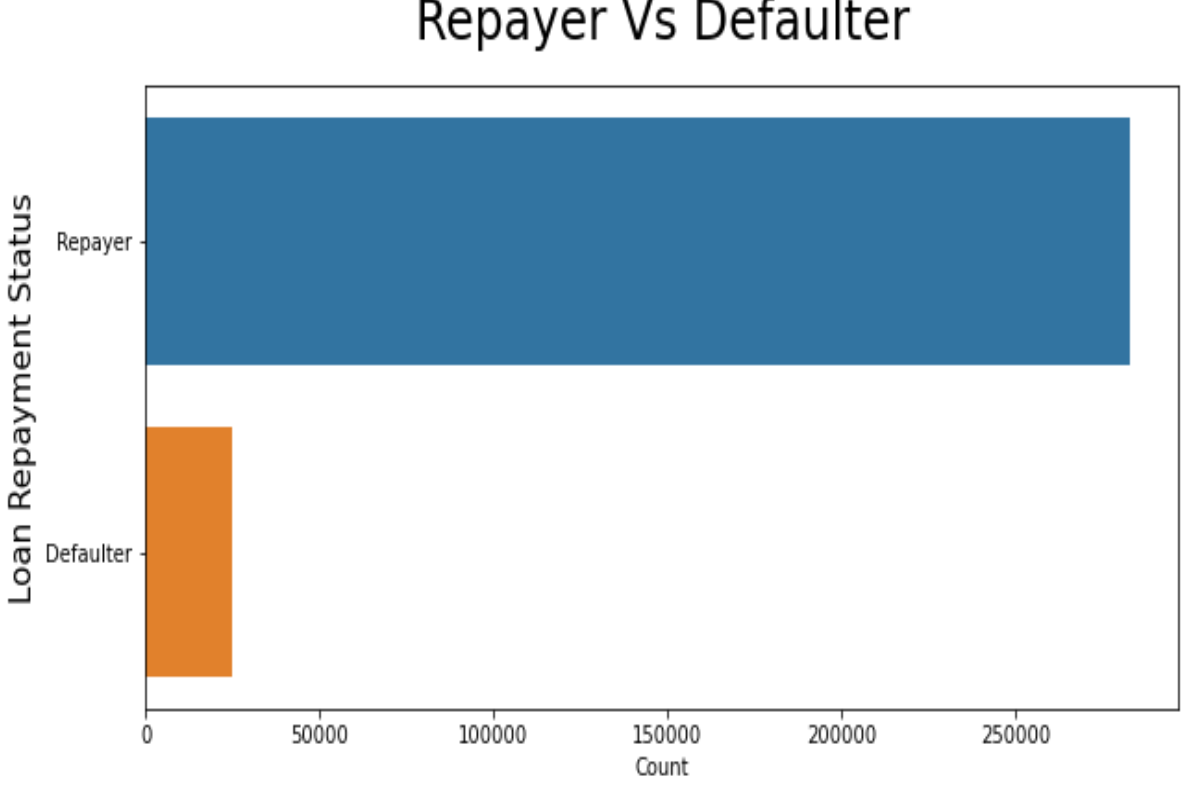
The Income of clients taking cash loans is more than revolving loans



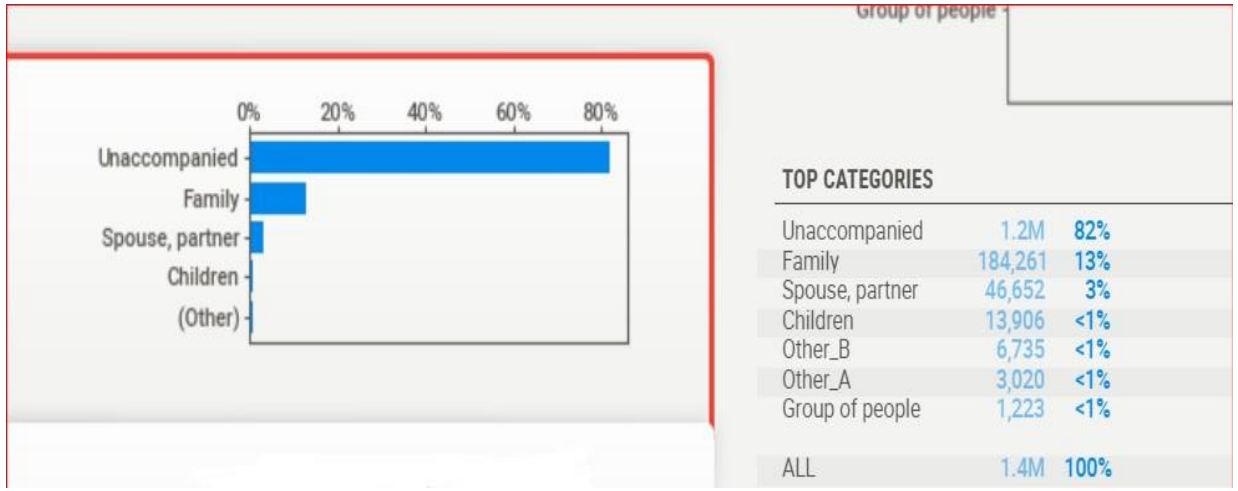
According to the previous cutomer data the consumer & cash Loans are Dominating



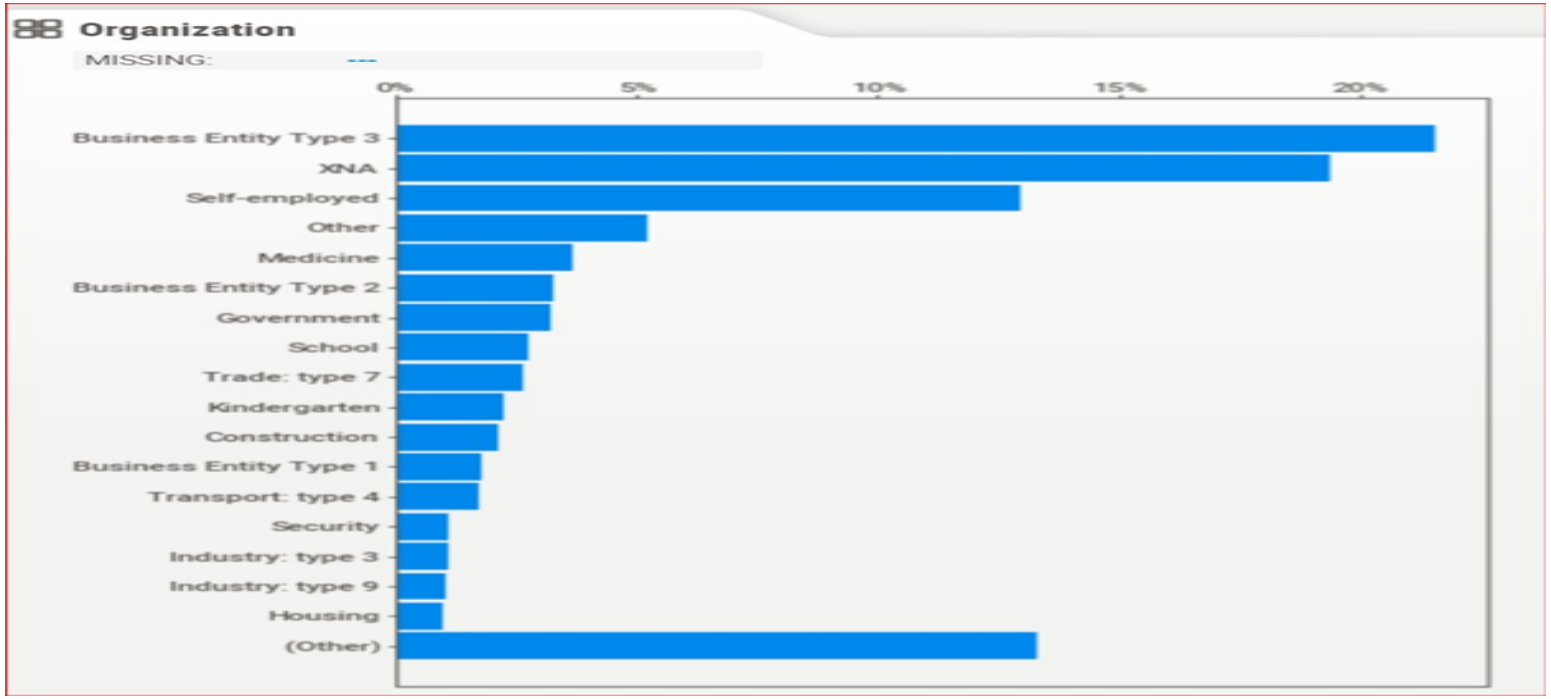
The terms for previous applications is same about 15-20
There are outliers for many Applications



There is Data Imbalance as Timely Loan repayers are Dominating



80 percent of clients are Unaccompanied or Bachelors



Majority of Clients work in business Entity Organizatons

Conclusions:

1. The clients applying for loan majorly are Male
2. Most of the customers have completed their Higher education and request for cash loans ranging from 25,000 to 1,00,000/-
3. As the income increases the Loan approval also Increases
4. There is Data Imbalance in Repayers & Defaulters
   Repayers are greater than the Defaulters
5. Cleints applying for Loan are not from highly populated regions

Thank You