Credit Risk Analysis in Banking and Financial Services

EDA Credit Assignment

For Upgrad and IIIT Bangalore

Credit EDA Assignment for Upgrad & IIIT Bangalore

- Mahesh Kumar Arya

Problem Statement

- Financial institutions find it hard to lend loans for customers due to their insufficient or non-existent credit history.
- Identify and reject the loan for the customers who are not likely to repay the loan
- Identify and approve the loan for the customers who are likely to repay the loan and reduce the loss of business to company

Objective

The Objective of this case study is to:

- Identify pattern from the dataset and understand if the customer has any difficulty in clearing the loan amount.
- Identify which segments are more likely to default the loan and which segment of people are likely to pay their instalments.
- Understand and identify the factors that trigger the customer to default the loan.

Expected Outcome

1. Outline key data inferences and their desired outputs
2. Data analysis pictorially represented by using Graph and Charts generated using Matplotlab and Seaborn libraries.

The Data Analysis involves below steps:

1. Data Loading
2. Data Cleaning
3. Data Analysis
4. Data Interpretation
5. Data Visualization

Data Loading

Data loading technique:

- Using “read_csv” method from CSV file into data frame using Python.

Method:

- Load the data(CSV files) into pandas DataFrame.

Methods used to study the structure of the data:

- head()
- Describe()
- Shape()
- Info()

Data Cleaning

- Data cleaning technique adopted involves using Python to “standardize the data”
- The data cleaning technique involves cleaning up of data using below ways:

  Identifying outliers, checking for NULL values, unknown values, inconsistent spelling etc.

Data Cleaning – Null Values

- In the given data set it is found that there are 64 columns that contain NULL more than 30%.
- There are 3 approaches to handle NULL in dataset.
  - The NULL values can be replaced by mean/median of the column.
  - The NULL values rows/col can be deleted based on the relevance.
  - The NULL values can be ignored.

- Note:
  - In this EDA the NULL values are ignored as per the instructions

Data Cleaning - Outliers

- In the given dataset there are outliers present in the columns such as CNT_CHILDREN, AMT_INCOME_TOTALetc columns.
- In the normal scenario the outliers will be handled in one or more ways:
  - Imputation of outliers
  - Delete outliers
  - Binning the values
  - Cap the outliers
- In the given dataset it would have been better to delete the outliers in the CNT_CHILDREN column as the values of outliers were 16,17,12 and it is highly unlikely to have 16 children in 1 house.
- The other approach would be to replace these high values with 1,assuming that these families have at least 1 child and the other digit was a typing error.

Data Cleaning – Unknown Values

- The application data set contained an unknown value ‘XNA’ in few columns. Ex: CODE_GENDER column consists 4 rows with this value.
- Hence the XNA value is replaced by ‘F’ (Female). Because the column contains maximum ‘F’ as value.

```
inp1_appdata[inp1_appdata["CODE_GENDER"]=="XNA"].shape
```

```
(4, 122)
```

```
inp1_appdata["CODE_GENDER"].value_counts()
```

```
F      202448
M      105059
XNA         4
Name: CODE_GENDER, dtype: int64
```

Calculate Imbalance Ratio

- The imbalance ratio is calculated on the TARGET between defaulters and non-defaulter's values. The imbalance ratio is 11.39%.

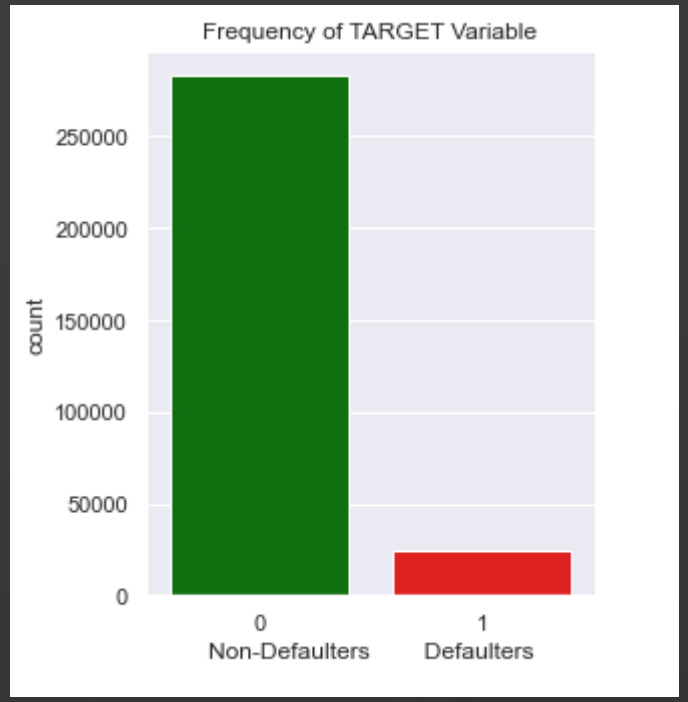
```
# Checking for Imbalance in TARGET Column.
cust_nondefaulter = inp1_appdata[inp1_appdata.TARGET==0]
cust_defaulter = inp1_appdata[inp1_appdata.TARGET==1]

data_imbalance = round(len(cust_nondefaulter)/len(cust_defaulter),2)
print(" Imbalance ratio is",data_imbalance)
```

```
 Imbalance ratio is 11.39
```

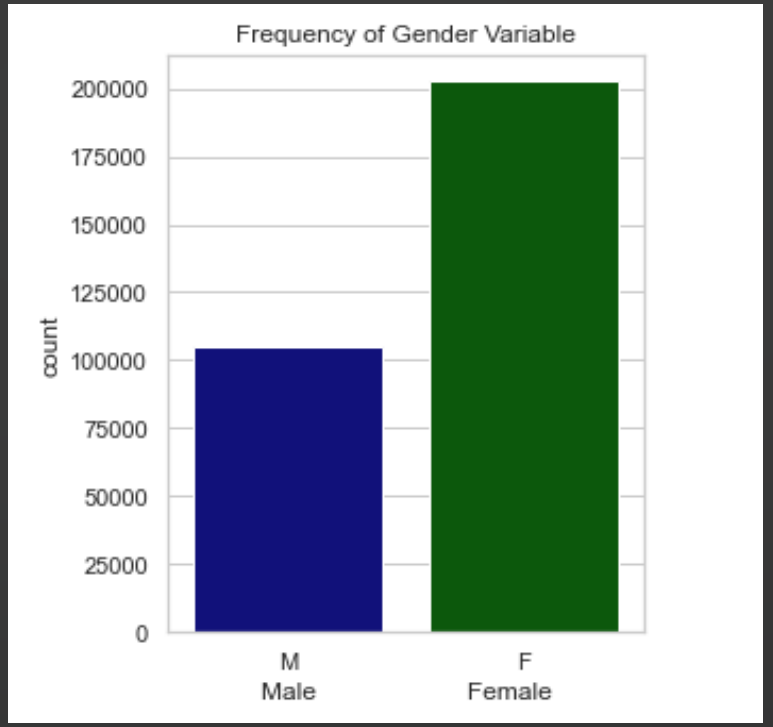
Data Analysis - Univariate Analysis of Target variable

- The graph shows that the count of non defaulters is Comparatively more than the defaulters.



Data Analysis - Univariate Analysis on Gender Variable

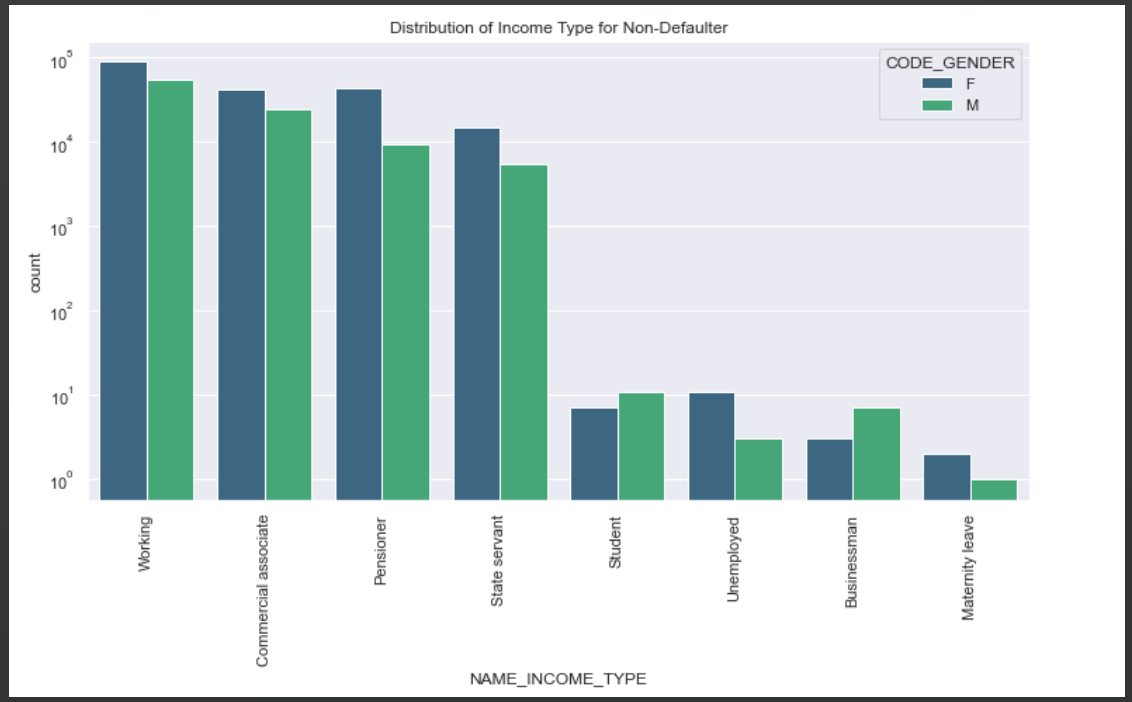
- The graph shows that there are a greater number of females who are involved in applying for loans than men.



Data Analysis - Distribution of Income type for Non defaulter

Inferences from the graph

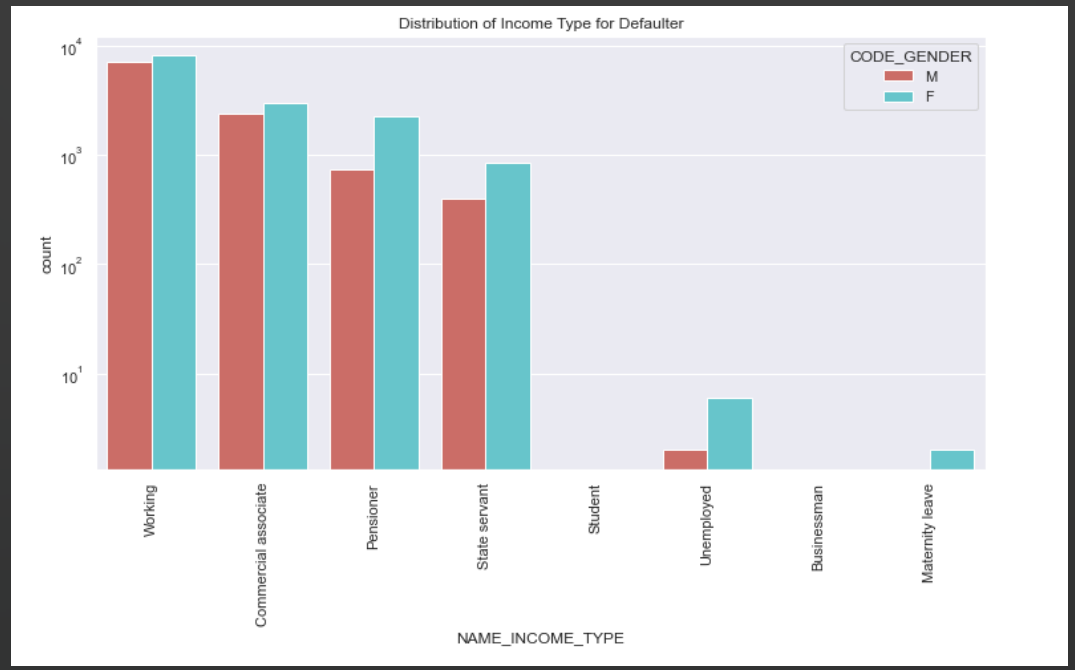
- For the income type working, commercial associate, state servant the number of credits are higher than student, businessman, unemployed.
- Women are having higher credit rate than men.



Data Analysis - Distribution of Income type for Defaulter

Inferences from the graph

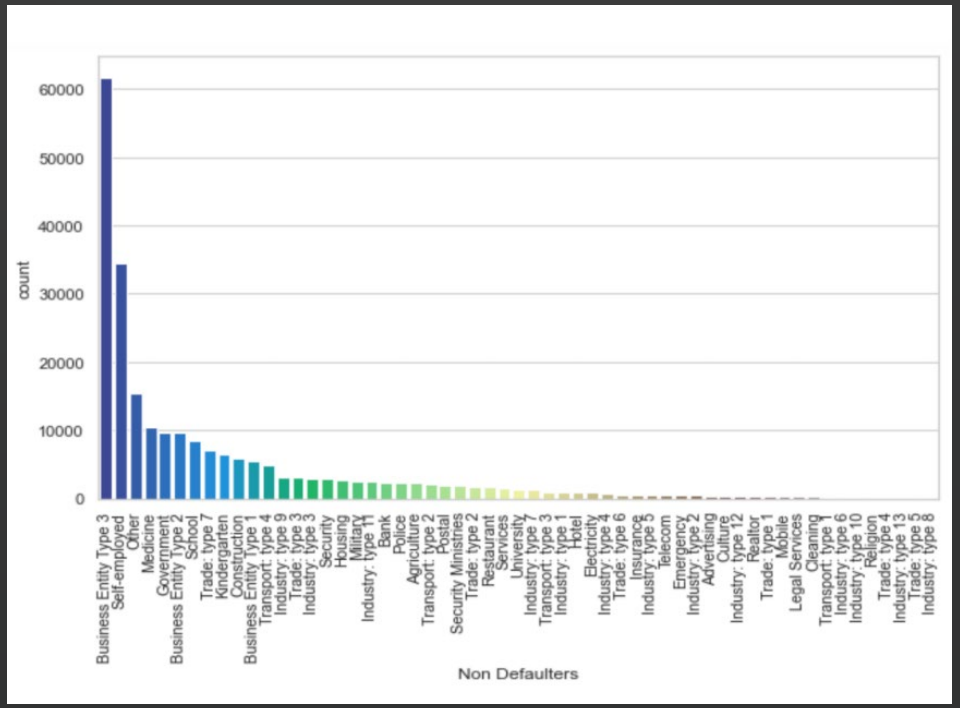
- The women are having more loan than men in Defaulters.
- Student and businessmen are almost Nil, which means they are the safe customers, and they pay the instalments on time.



Data Analysis - Distribution of Organization type for Non-Defaulter

Inferences from the graph

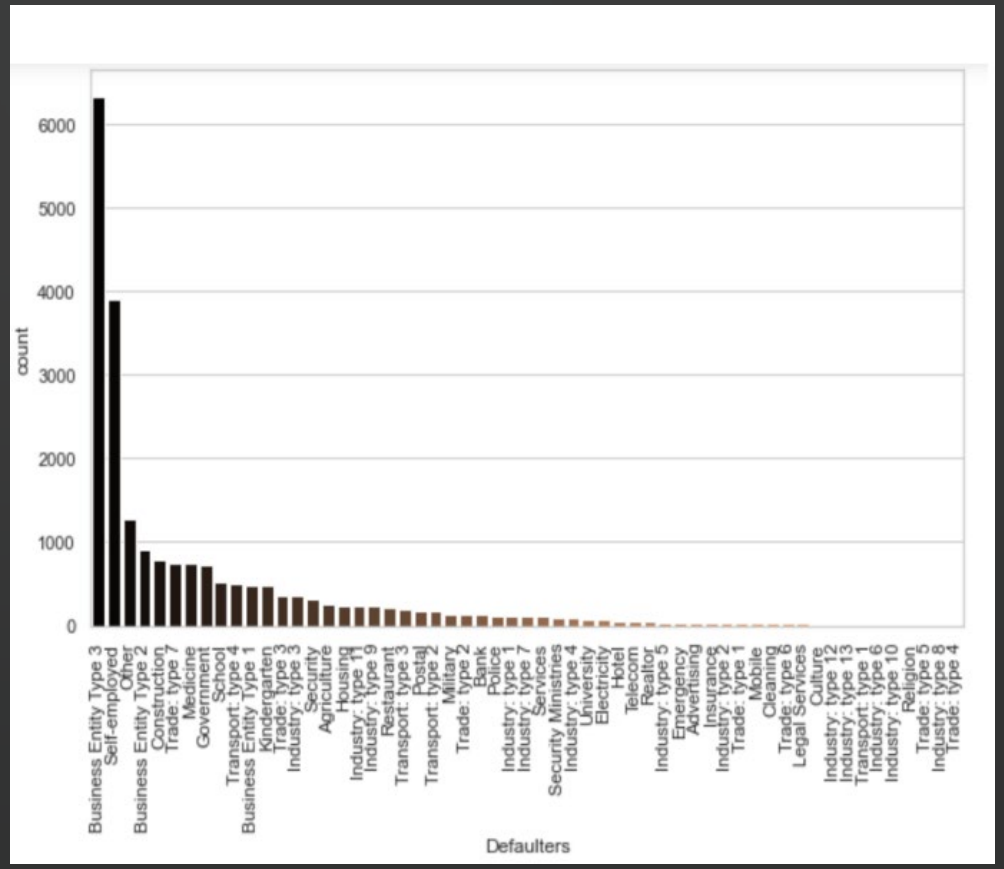
- Customers who have applied for loans are from business, government, people who are in medicinal Field.
- Customers are almost Nil from industry type sector, transport etc.
- There is a steep fall from business customers and other sector people which means businesspeople will apply for huge and heavy loans.



Data Analysis - Distribution of Organization type for Defaulter

Inferences from the graph:

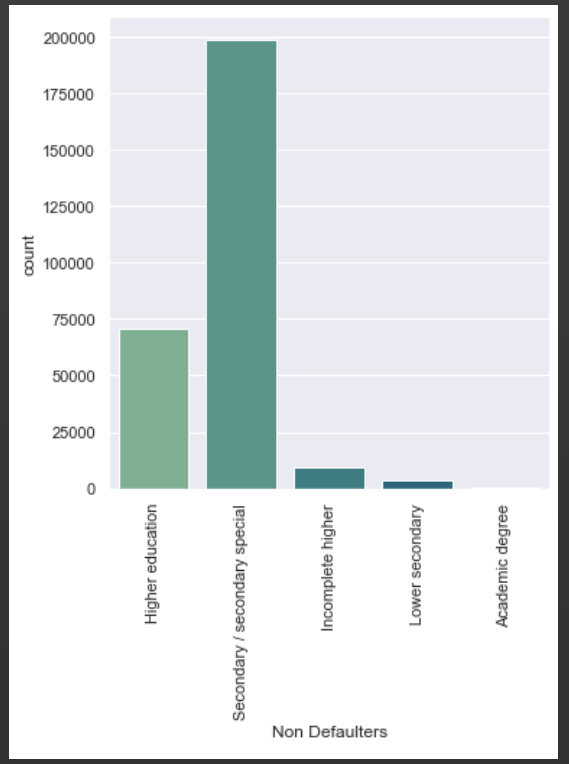
- The customers who have applied for loans are from business and self-employed sector.
- Customers with a smaller number of loans are from services, culture, industry type 8,10,12…



Data Analysis - Distribution of education for Non-Defaulter

Inferences from the graph

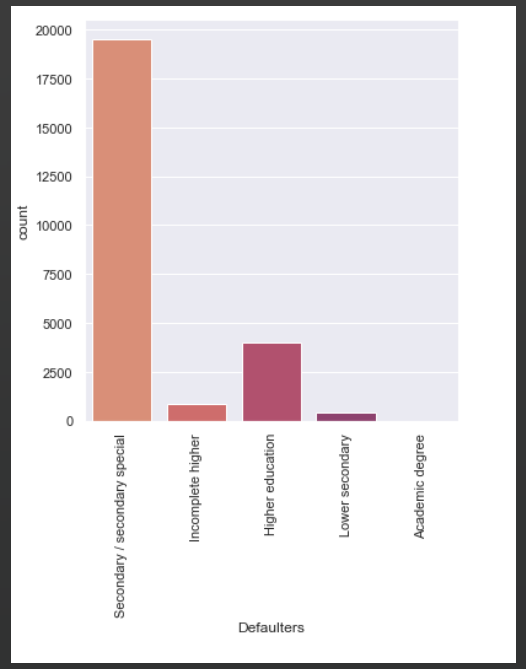
- The highest loan is applied by customers with secondary education and the lowest loans are from the academic degree people.
- There is a steep difference between the secondary education and other degrees.



Data Analysis - Distribution of education for Defaulter

Inferences from the graph

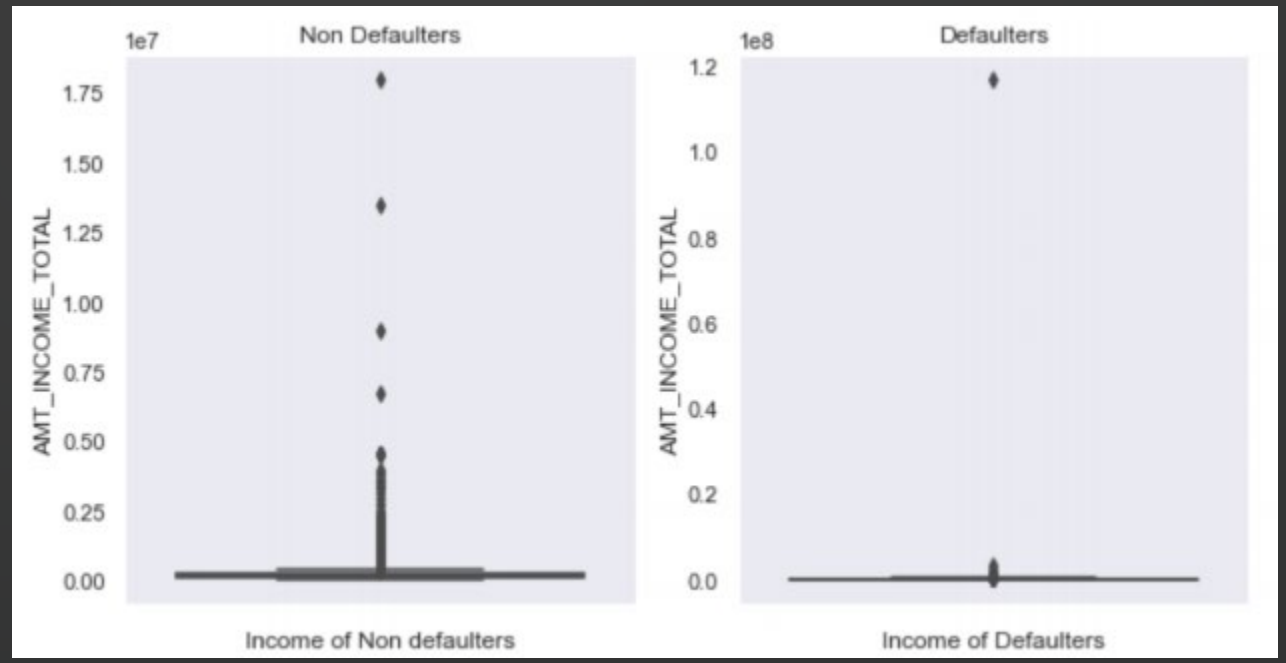
- The highest loan appliers are from secondary education background.
- The academic degree customers almost pay the instalments on time, and they do not possess any risk to the bank.
- Banks should be more careful with the customers with secondary special education background.



Data Analysis - Outlier in income for Non defaulter and defaulter

Inferences from the graph

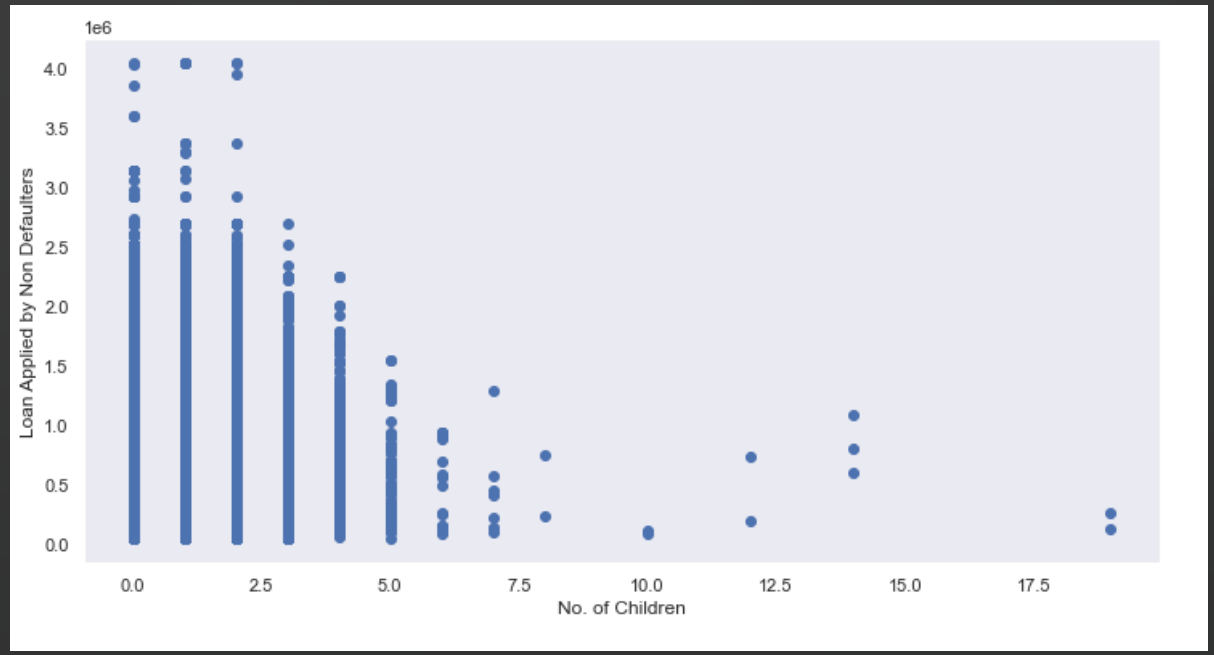
- There is a prominent outlier in both non defaulter and defaulter income category.
- The quartiles are very slim for the defaulters compared to non-defaulters



Data Analysis - Distribution of number of children for Non-defaulter and Defaulter

Inferences from the graph

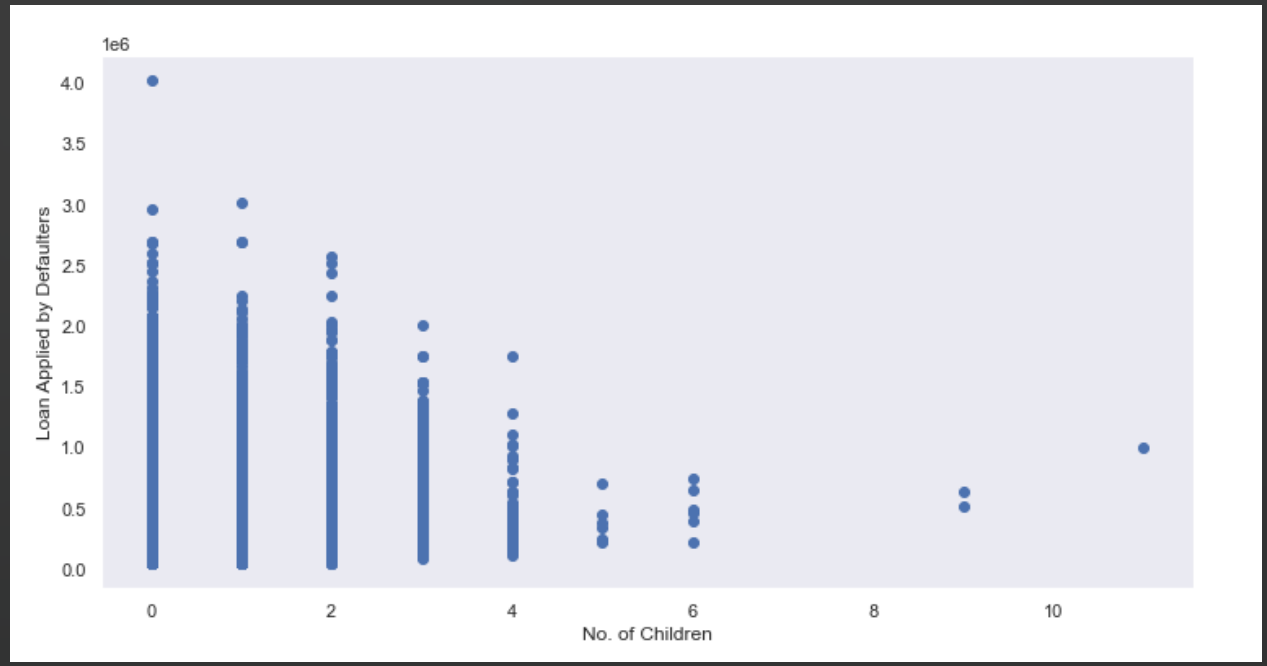
- The loan applied by non defaulters are more with 0 children in the family.
- The loan applied gradually has decreased with the increase in number of children.
- We can see outliers with respect to number of children here, the graph shows number of children as 10,12,15 which is highly unlikely.



Data Analysis

Inferences from the graph

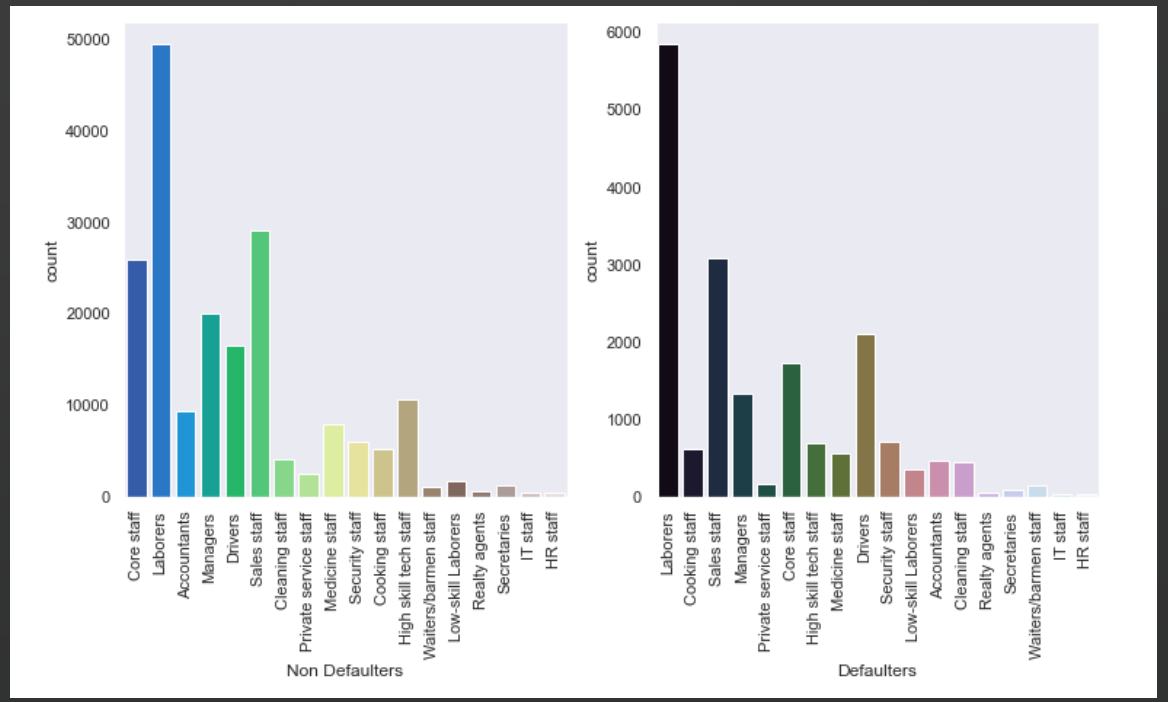
- The loan applied by customers with o children is very high.
- The number of loans gradually decreases as the number of children increases.
- There are outliers in the data with respect to number of children, and it can be ignored as per the instructions.



Data Analysis - Analysis of occupation type non-defaulter and defaulter

Inferences from the graph

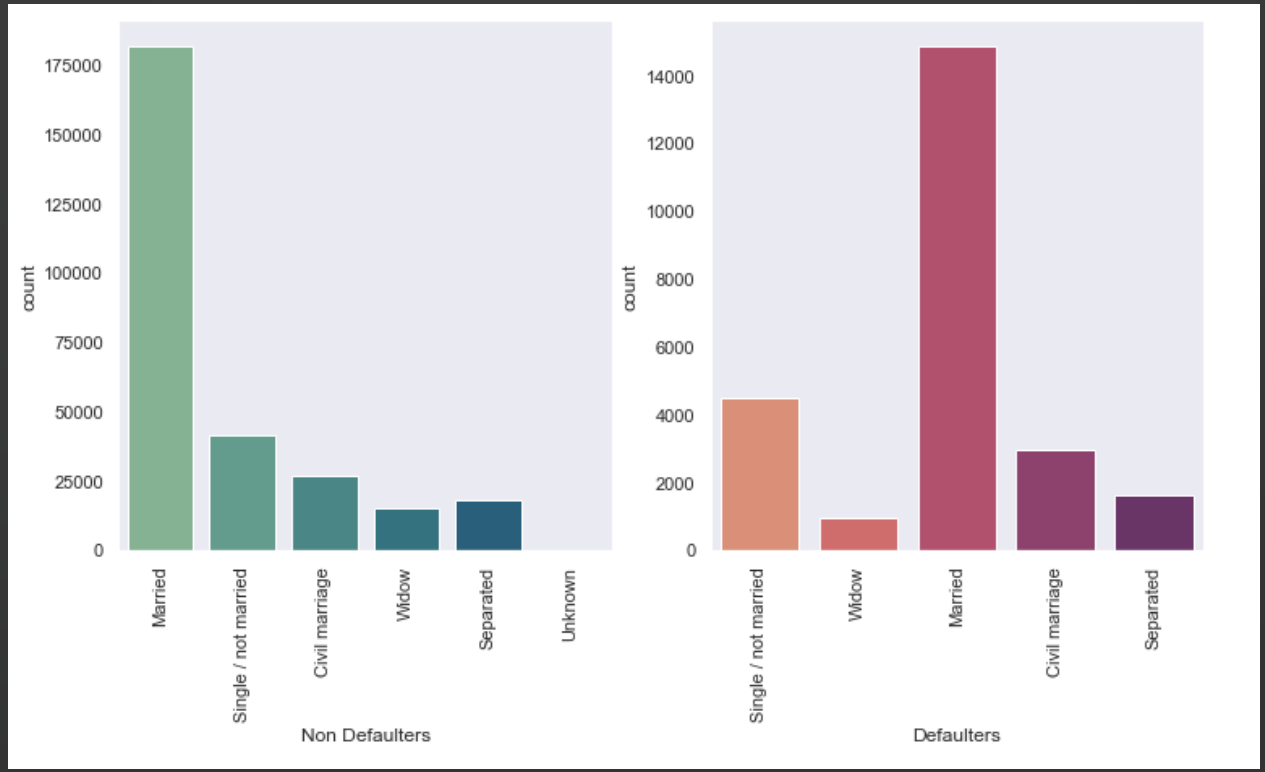
- For the non-defaulters the core staff, laborers and sales staff are high in number who apply for loans.
- For the defaulters the laborer's, sales staff, drivers core Staff are more likely to have payment difficulty.
- Banks should be careful while dealing with these customers with above occupation type



Data Analysis - Analysis of family status of non defaulter and defaulter

Inferences from the graph

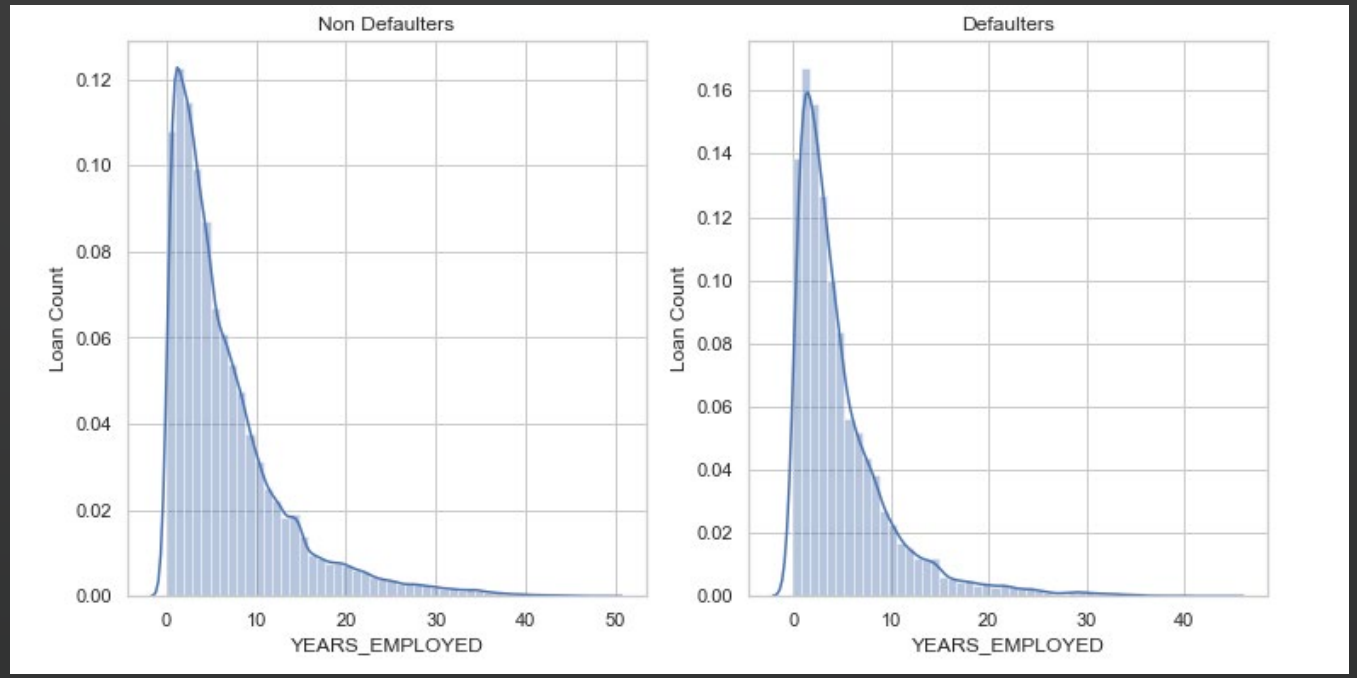
- Married people tend to have more difficulty in paying off the loans in the defaulter's data.
- Widow, separated, single status people are disciplined when it comes to paying the Instalments



Data Analysis - Analysis of number of years of employment for non defaulter and defaulter

Inferences from the graph

- The graph shows that the greater number of customers fall into o years which means, most of the people are not working/don't work.
- These set of people possess threat to the bank as they are not employed and don't have a permanent income and hence cannot pay the loan in time.



Data Analysis - Multivariate analysis of the application data – non-defaulters

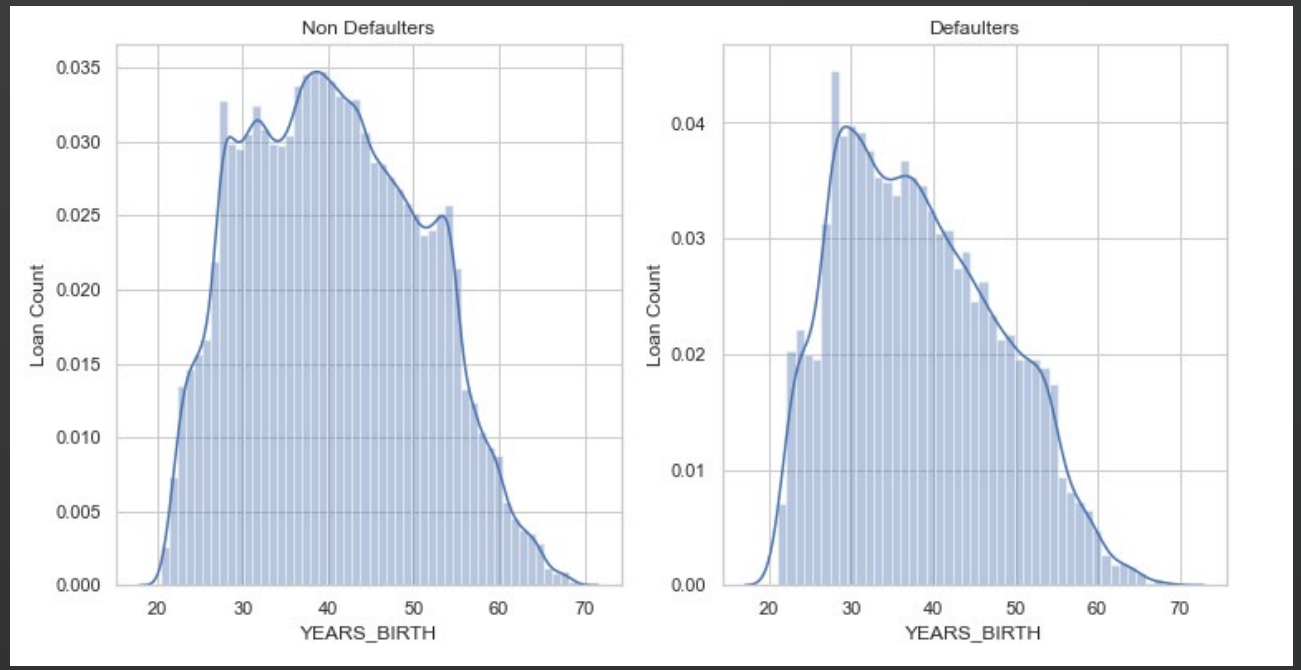
Inferences

- The customers permanent address does not match with current live city.
- Credit amount is higher for young customers.

Data Analysis - Analysis of number of birth years for non defaulter and defaulter

Inferences from the graph

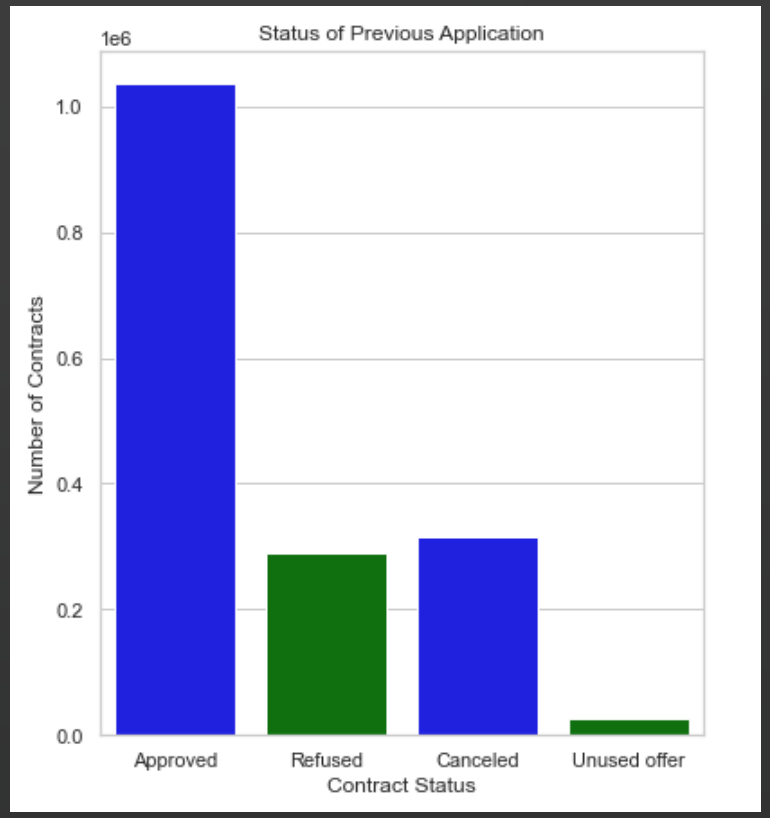
- For the non-defaulter's category, the graph is almost steady between 25-50, there is no steep slope.
- The greatest number of people in defaulters are from the age group of 25-50.
- There is a decline in the slope of the graph from 35 onwards for defaulters.
- Maximum people failing to repay the loan is from 20-25.



Data Analysis - Analysis of the previous application data

Inferences from the graph

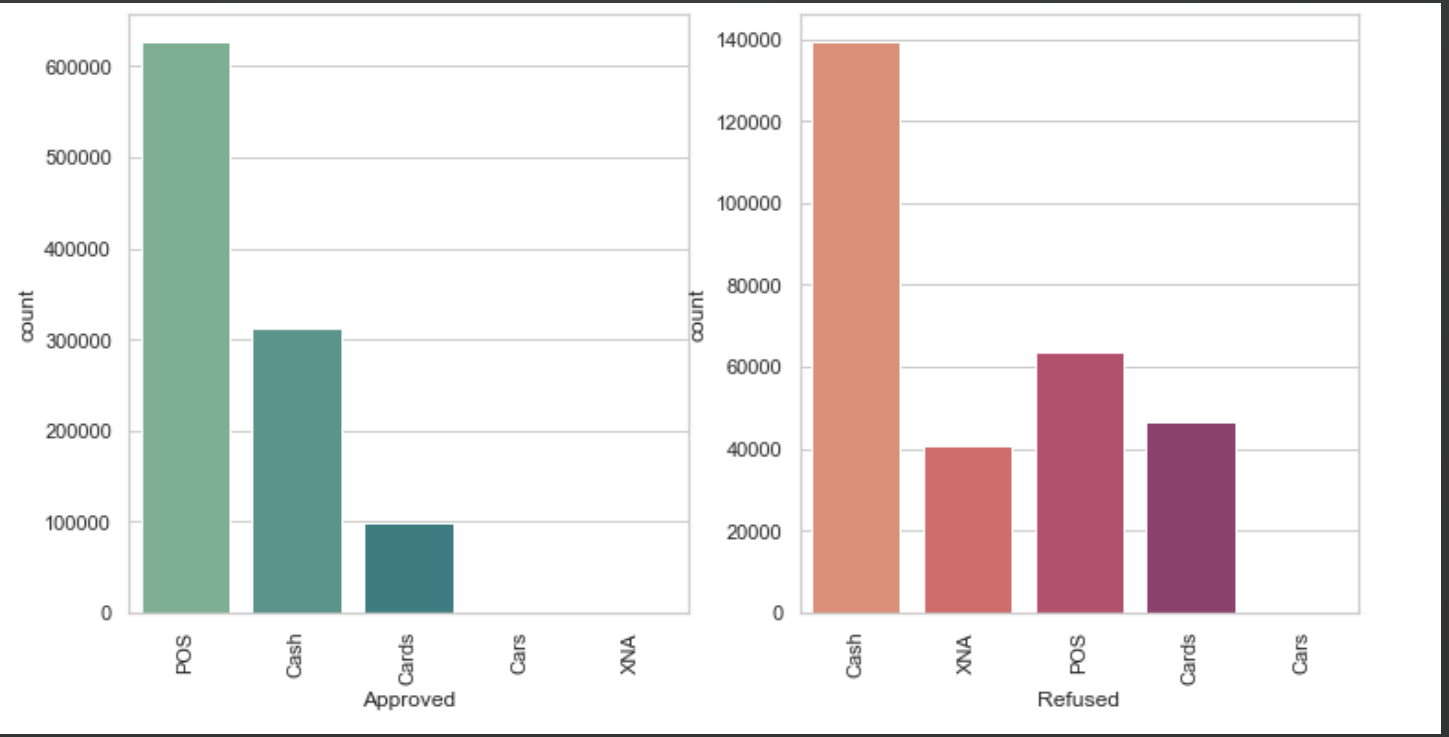
- In the previous application the data shows there are more
  - Approved status loan than refused, cancelled and unused.
  - The number of refused and cancelled loans are almost equal



Data Analysis - Analysis of previous loan status

Inferences from the graph

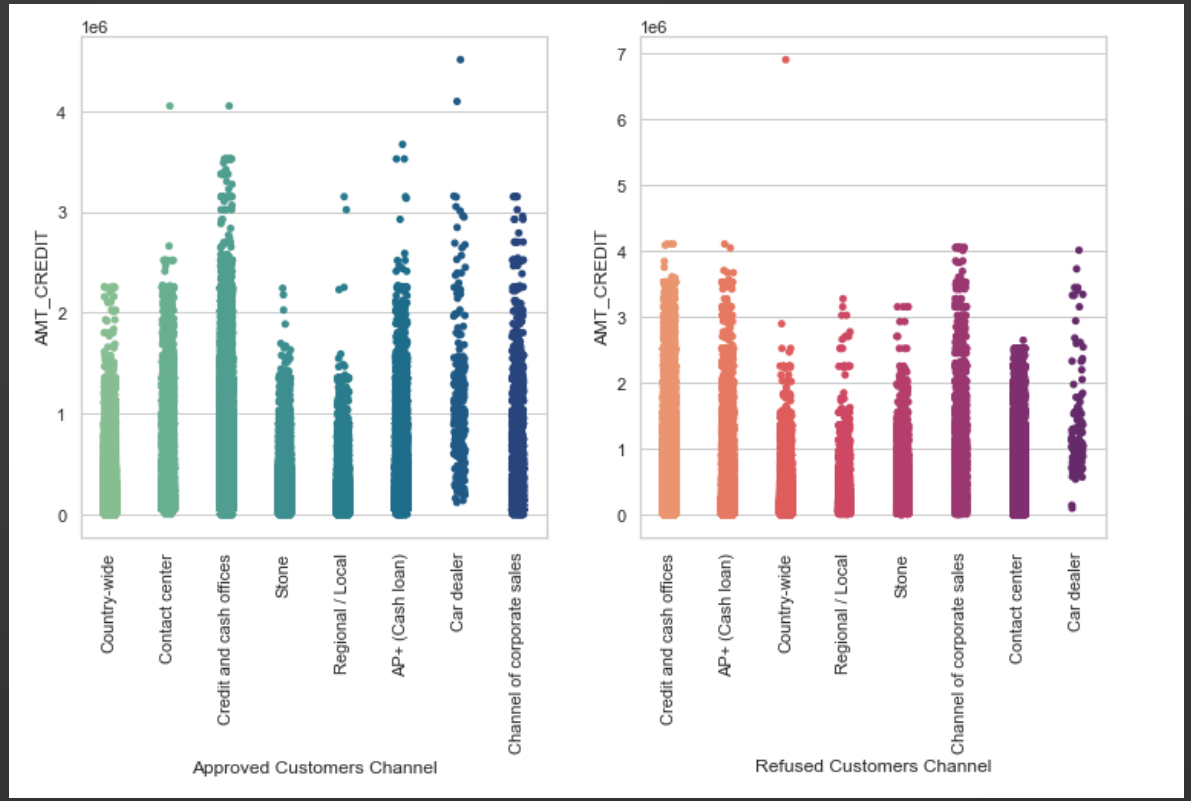
- It is evident from the graph that the greatest number of approved loans are for POS Section.
- Most number of refused loans are for the Category Cash



Data Analysis - Analysis of previous loan status

Inferences from the graph

- The graph depicts that most approved loans channel type are having outliers.
- Most of the approved loans are from channel of Corporate, credit and cash office.
- Most of the refused loans are from cash loan.



Data Analysis - Multivariate analysis of the application data – Defaulters

Inferences

- Less children customers will live in highly populated area.
- Income is higher in densely populated areas as opportunities will be more.