Exploratory Data Analysis- Lending Club Case Study

Group Members:
Saaigoutham Ashokkumar
Pankaj Mishra

Lending Club Case Study

BUSINESS CONTEXT: A consumer finance company which specializes in lending various types of loans to its customers has been experiencing the defaults by the borrowers. The company has assigned the task to identify the possible trend, patterns, factors and parameters, which could possibly help them in identifying and understanding the reasons for such default using Exploratory Data Analysis. (EDA) technique.

PROBLEM STATEMENT

To understand and analyze the given data using EDA techniques for identifying the major impact factors or trends for the reported loan defaults by its customers. Given Consumer attributes and loan attributes details should be utilized in the best possible manner to give recommendations. The primary objective of the Risk analysis to be carried out in the study is to minimize the risk for losing money by company in the lending duly taking care of the aspect that no potential loan applicant likely to repay the loan should get rejected, as it would result in financial loss to the company.

PROPOSED METHODOLOGY

1. Data Understanding

- Data quality, Data Interpretation, Understanding of the variables, size & shape of the dataset, checking of available parameters & attributes, checking of Missing or Null values, data formatting issues etc will be identified .

2. Data Cleaning & Handling

- Data quality issues will be handled with aim to optimize the dataset for better analysis with handling of Missing values, imputation of missing values possible and acceptable as per business context and target analysis, Data formatting corrections, Data type corrections, Sanity checks, Grouping of the data for derived metrics for future analysis etc will be carried out.

3. Data Visualization and Advanced data cleaning

- Using data visualization tools outlier handling and treatment will be carried out to prepare the dataset for Bivariate and Multivariate analysis.

4. Data Analysis

- Univariate, Bivariate and Multi Variate Analysis techniques will be used for carrying out the analysis to identify the impact of various variables in data on the tendency to default by a loan customer.

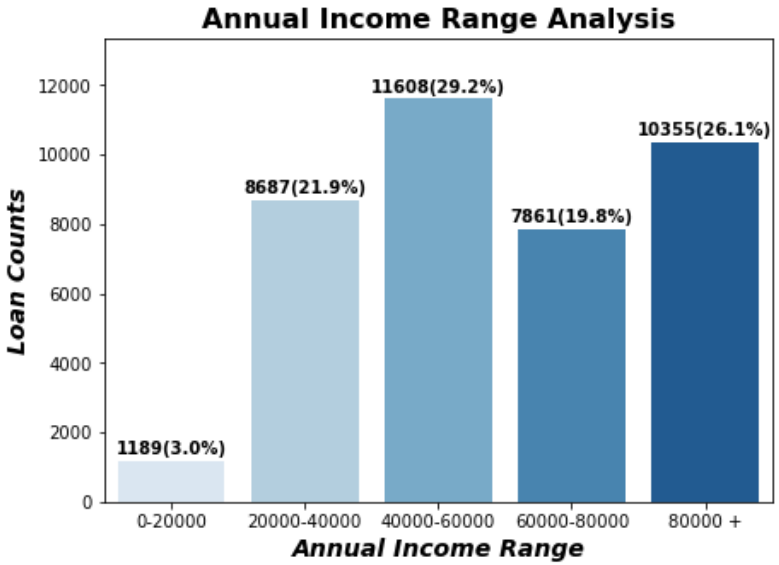
Data Cleaning

- Removal of NA values
  - Removed the columns having NA values for 90% and above
  - Removed the rows having NA values for entire columns (if any)
  - Removed duplicate records (if any)
- Derived multiple columns from the given data
  - Segregated date column into month and year
  - Created total amount column from term and installment
  - Created ranges for loan, interest, income, dti, total amount, fund approved
- Formatting the column
  - Removed unwanted spaces
  - Removed unwanted information
- Dropping the Column which not used for analysis
- Sanity Checks

Data Point Observation

- On analyzing the dataset, following points were observed and accordingly data analysis was carried out: Since the given data had details about the customers, approved loans and repayment details, it was obvious that the “Customer Behavior variables” essentially required for approving the loans will not be of any use in our study to identify focused on identification of the reasons and parameters for the default. Hence such variables were dropped for facilitating the better analysis.
- Details pertaining to the existing Customers in the given data were not used as they had no history of any default.

Observation during Analysis

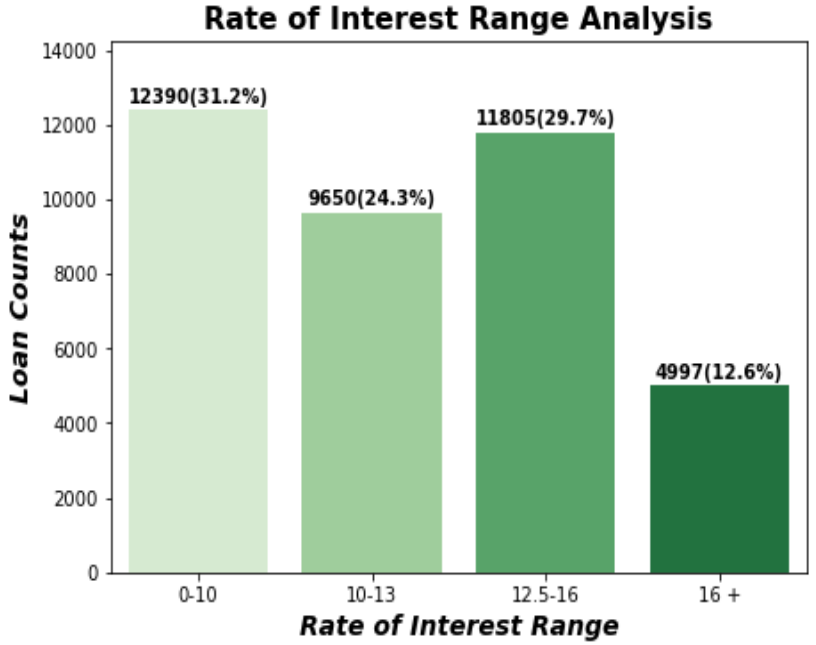


Annual Income Analysis

More number of person - 11608(29.2%) applied for loan having Annual Income Range between 40K USD and 60K USD

Rate of Interest Analysis

Maximum number of Loans were approved - 12390(31.2%) for Interest Rate Range between 0% and 10%

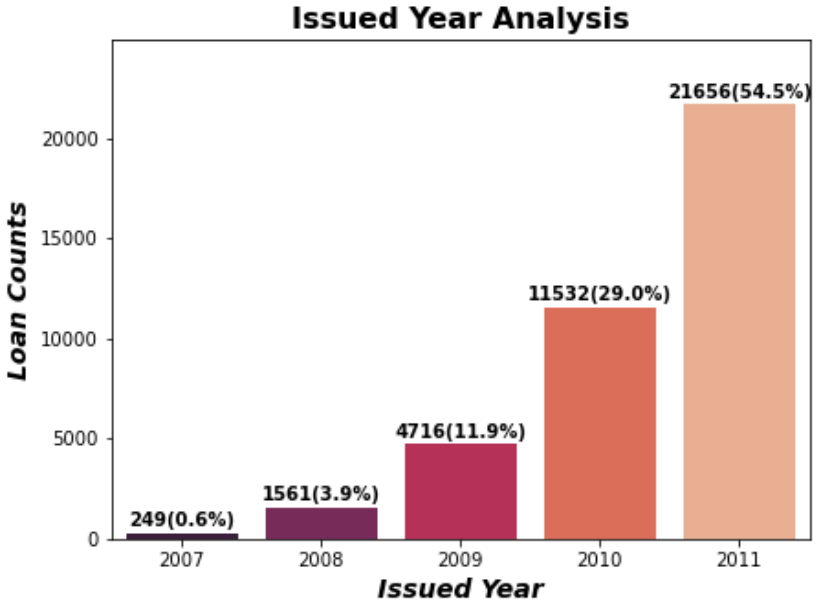


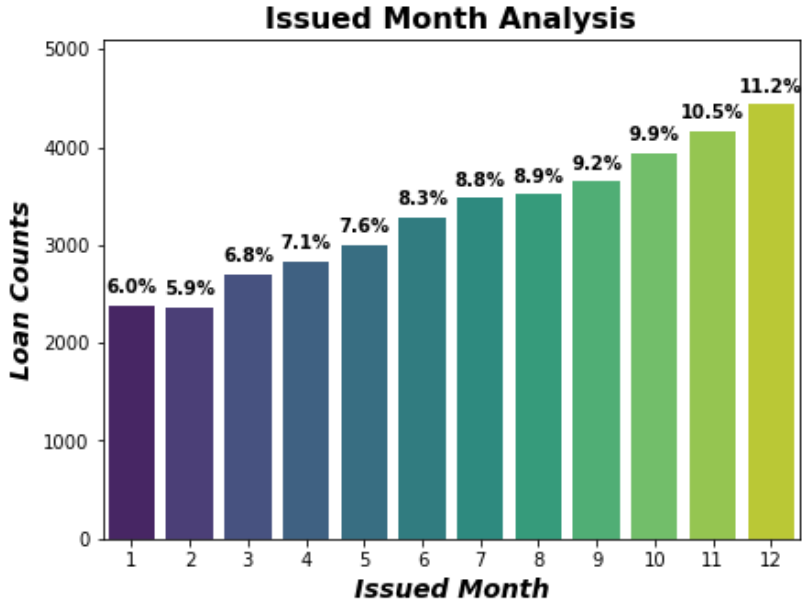
Observation during Analysis

Issued Year and Issued Month Analysis

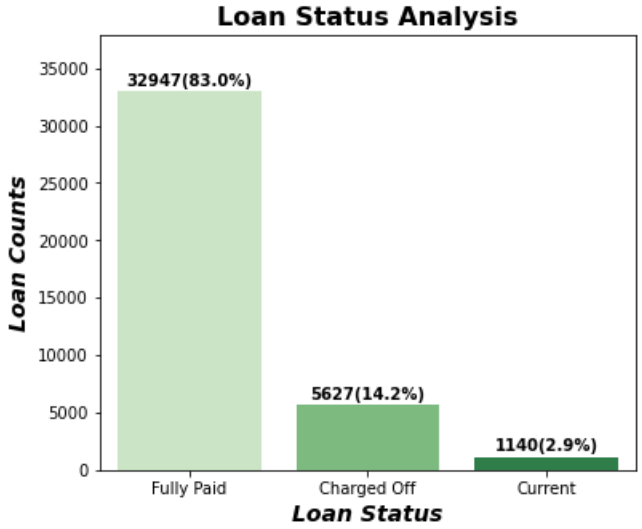
1) Maximum number of Loans - 21656(54.5%) issued in the year 2011

2) The rate of Issuance of Loan were continuously increasing from January to December





Observation during Analysis

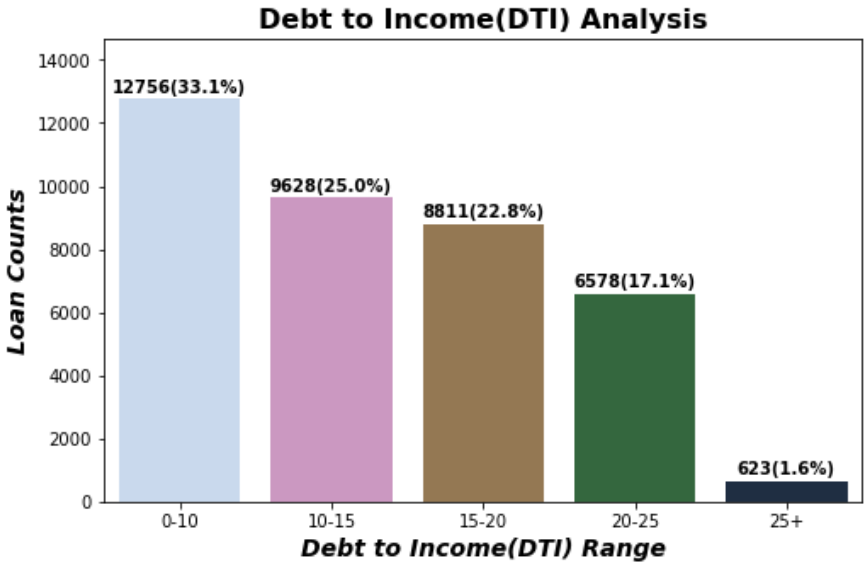


Loan Status Analysis

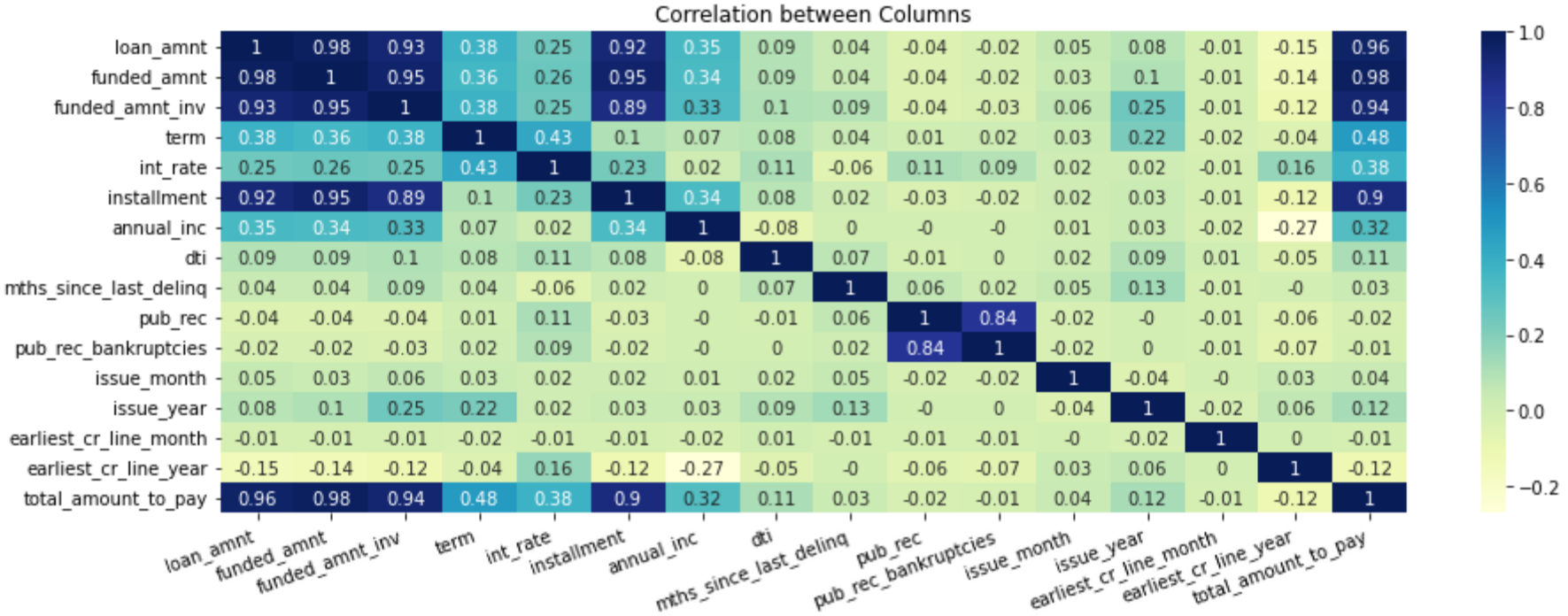
More number of Loans – 32947(83%) were fully paid and 5627(14.2%) number of loan were defaulted

Debt to Income (DTI) Analysis

Maximum number of Loans - 12756(33.1%) were issued for Debt to Income(DTI) in Range between 0 and 10

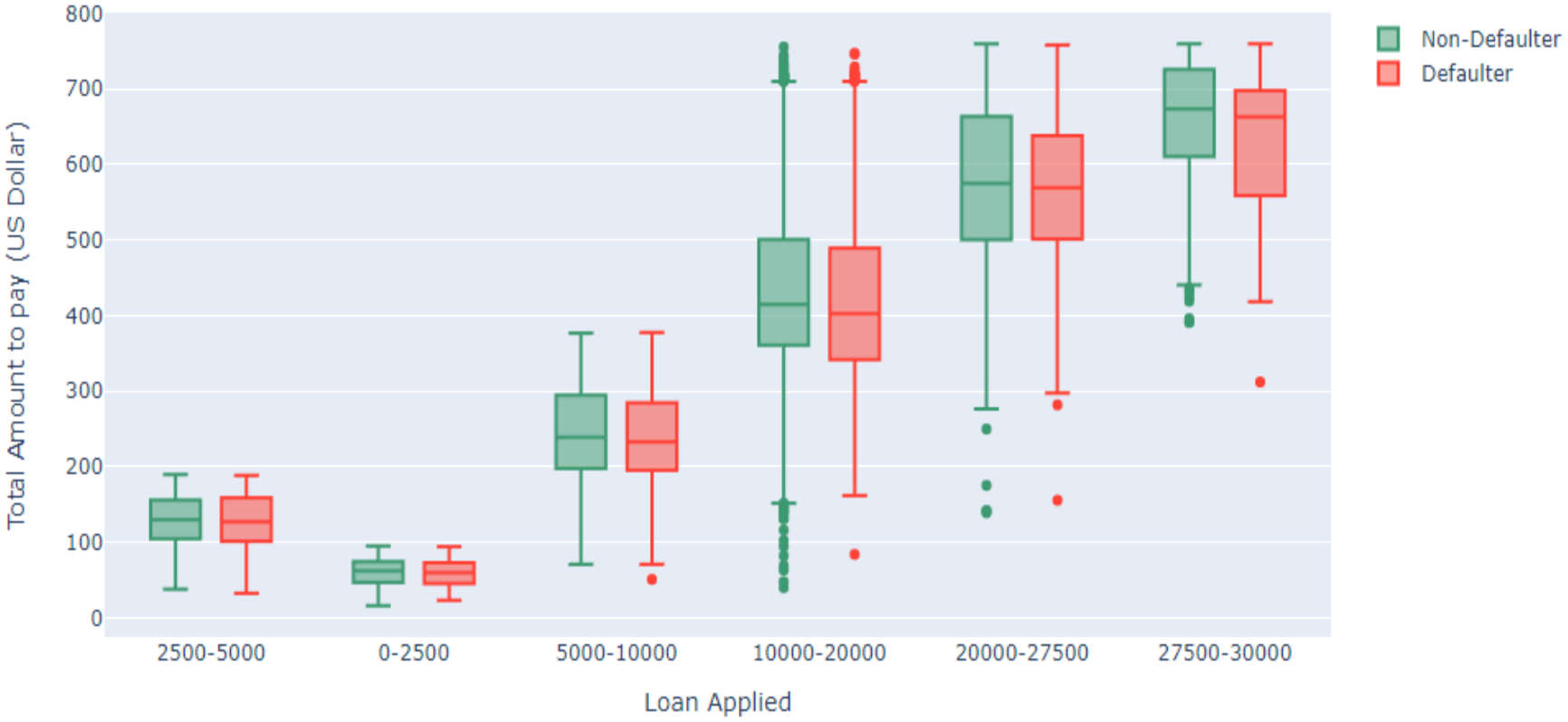


Relationship between variables



- Strong Positive Correlation (> 0.75 or higher only is considered) is seen between loan_amnt,funded_amnt and funded_amnt_inv, which is due to direct logical and mathematical relationship between them.
- Strong Positive Correlation (> 0.75 or higher only is considered) is also seen between loan_amnt/funded_amnt/funded_amnt_inv and installment/total_amount_to_pay (calculated using term and Installment) and would be further analysed for impact on target variable.
- Some Negative (> -0.1 or higher only is considered) Correlation is also seen between between loan_amnt (funded_amnt & funded_amnt_inv and earliest_cr_line and would be further analyzed for impact on target variable.
- Strong Positive Correlation (> 0.75 or higher only is considered) is also seen between pub_rec and pub_rec_bankruptcies, which is logical as delinquency instances are invariable part of public record instances.
- Some Negative (> -0.1 or higher only is considered) Correlation is also seen between between annual_inc and earliest_cr_line and would be further analysed for impact on target variable.
- Some Negative (-0.081) Correlation is also seen between between annual_inc and dti and would be further analysed for impact on target variable.

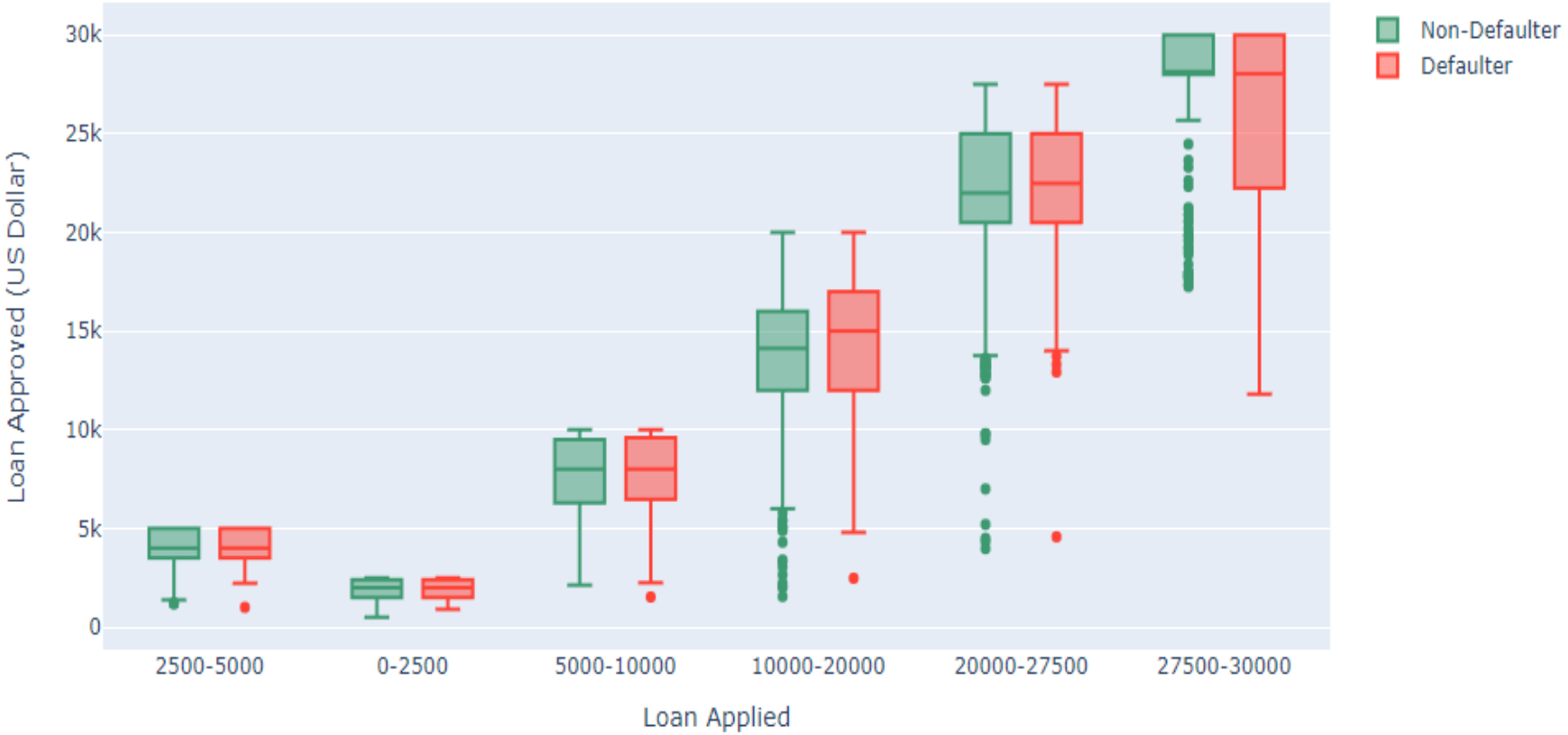
Effect of Installment variable on tendency to default



OBSERVATION:

No significant trend is seen except the trend where default is starting at a lower Installment value of approx 520 USD i.e. much lower than the minimum installment value for Non-Defaulters in the loan range of 27,500-30,000.

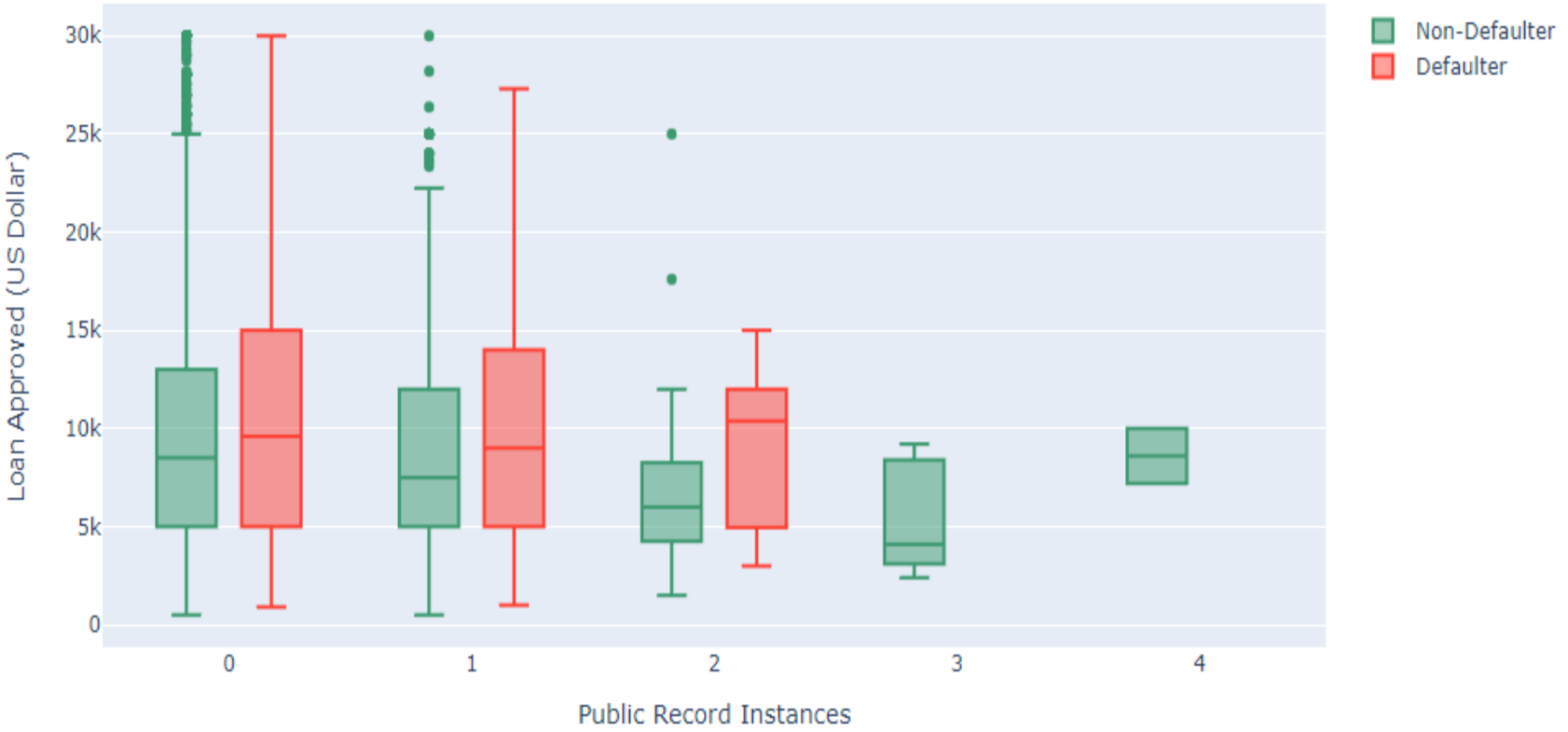
EFFECT OF LOAN AMOUNT VARIABLE ON TENDENCY TO DEFAULT



OBSERVATION:

A significant trend is for two applied loan ranges i.e. in 10,000-20,000 and 27,500-30,000 i.e. wherever approved loan amount by company is greater than approx 15,000-18,000 tendency to default is higher and lower range of value from which default has started has been considerably lower for loan approved by company amounts of approx 24,000 or higher.

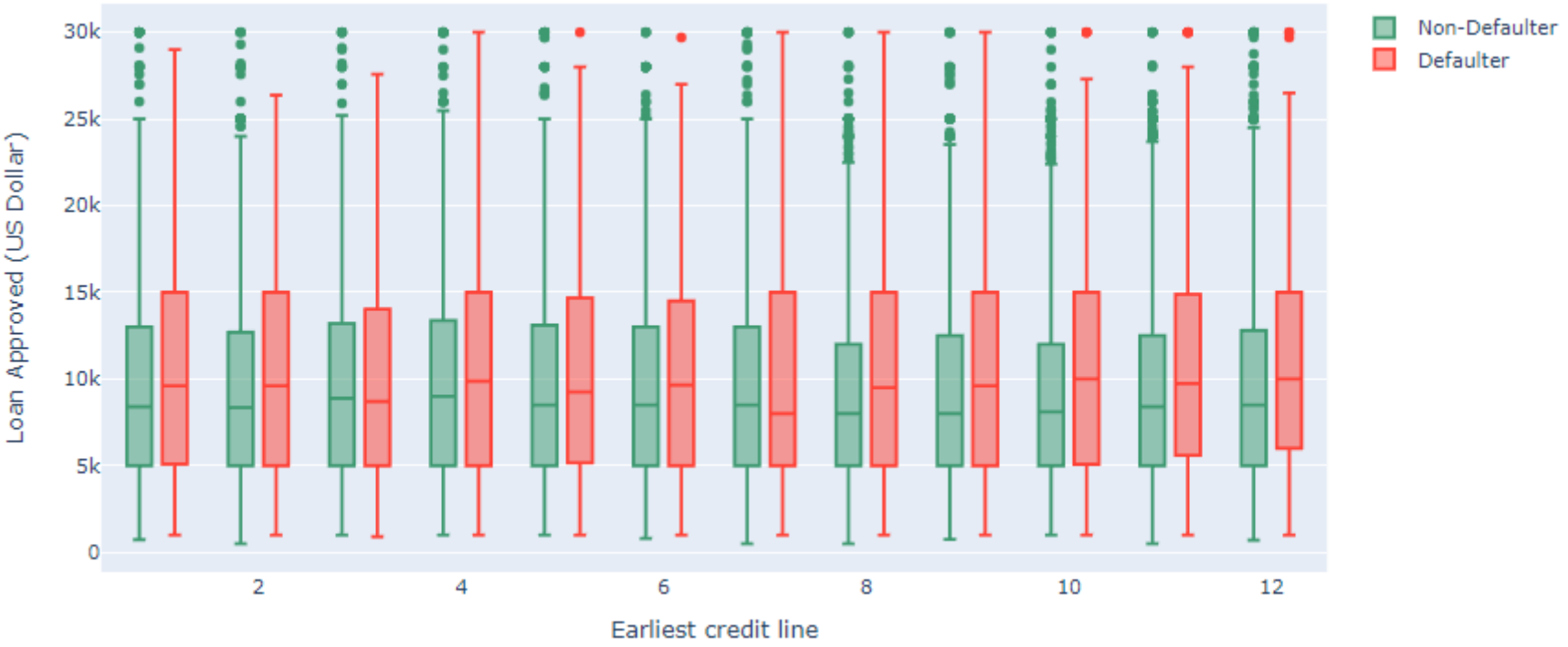
EFFECT OF PUBLIC RECORD VARIABLE ON TENDENCY TO DEFAULT



OBSERVATION:

An interesting trend is observed here, where no defaulters are there for cases with 3 or higher instances. Probably that means loan approval process must have been very efficient in such instances, however when cases with 1,2 & 3 instances are observed number of defaulters, approved loan range of default and median value for 50th percentile is also on higher side.
Particularly, in cases with 2 instances numbers of defaulters are very high in numbers.

Effect of Earliest credit line variable on tendency to default



OBSERVATION:

The number of defaulters for month August to December are increasing relatively in comparison to the trend visible for the months prior to August

RECOMMENDATIONS

- Loan Applications received for amounts ranging from 10K USD to 30K USD with LC assigned grading of '**G**' category needs to be scrutinized and checked as the tendency to default in such case is High
- Customers applying for the loan amounts in the range of 27,500 USD or higher should be checked for the past history including but not limited to the public record instances during the loan approval process as there were significant default customers in this loan range in cases company approved loan amounts were lower than the applied loan amounts. It seems some amount of risk was absorbed in such cases with approval of lesser loan amount but resulted in relatively high number of defaulters
- Small businesses loans have more defaulters, hence business loans dispersal should be strengthen with better checks for reducing the defaults
- Credit hungry applicants i.e. Customers with past history of >5 enquiry in last 6 months should be verified carefully with Due Credentials about Good Credit History
- People with significance past history of public records and bankruptcy must be handled with more precautions particularly for higher amount of loans. Few public record instances in Credit History shouldn't be taken lightly during loan application evaluation as this lapse was evident in the given data for more number of defaulters in such cases
- Loans approved for amounts higher than 10,000 USD and longer term of 60 months have defaulted double in numbers so loan approval process in such cases needs to be strengthened with more checks