Lending Club – Case Study

Saurav Suman

Megha Jain

Agenda

Introduction

Problem Statement

Data Analysis & Clean Up

Exploratory Data Analysis

Recommendation

Introduction

Consumer finance company which specializes in lending various types of loans to urban customers.

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

Two types of risks are associated with the bank's decision:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

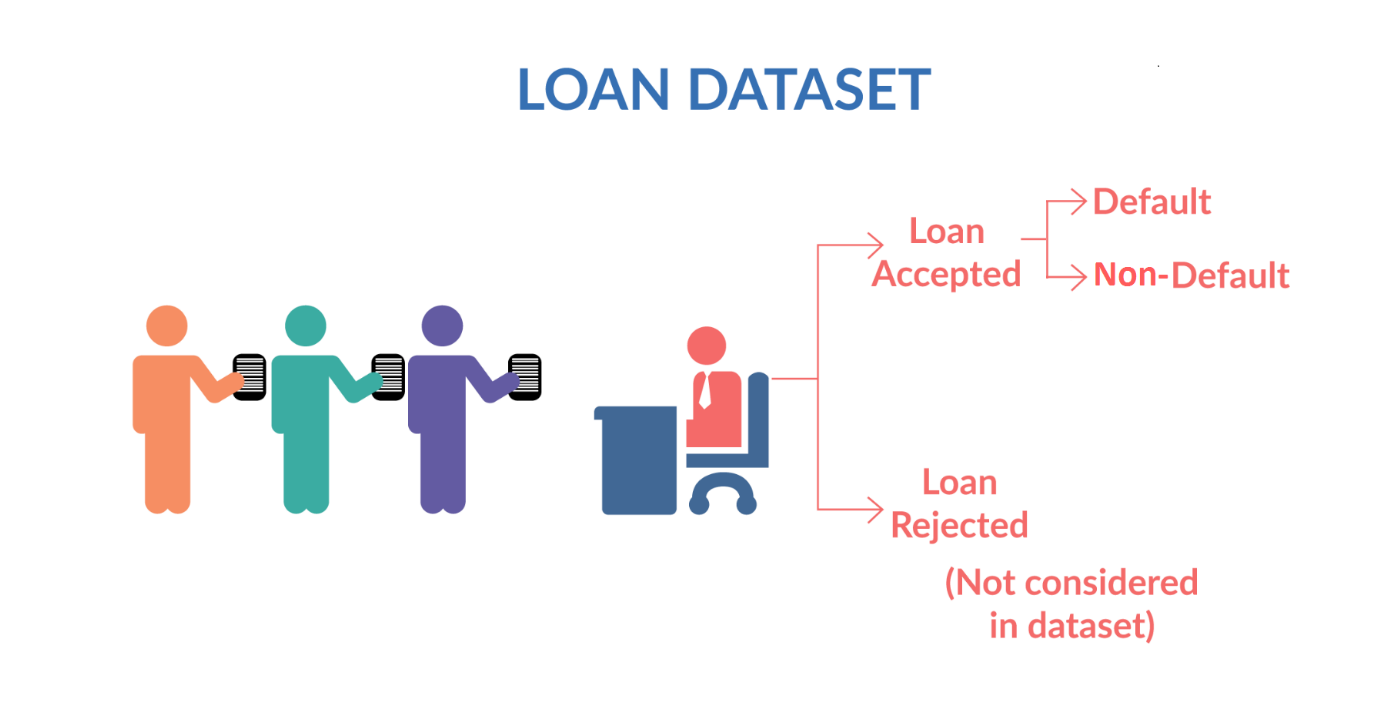
2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

Problem Statement

- Understanding driver variable behind loan default and listing down which variable can indicate strong defaults

- Steps to achieve it

  - ➢ Use EDA to understand how consumer attributes and loan attributes influence the tendency of default.
  - ➢ Identify the key driving factors causing Loan to default.
  - ➢ The dataset consists of Loan applications that has been accepted by Bank. The loan may be currently in one of the two states: Default or Non-Default.



Data Understanding

- Read the CSV and imported basic libraries for EDA
- The Loan Dataset provided by the Company has 111 columns and 39717 rows.
- Many columns data count is zero.
- Loan Amount ranges from 500-3500.
- The data will be analyzed and treated during various phases to make it ready for analysis.

Data Cleaning : Missing Values

- Found two types of missing value-
  I. Entire column value is missing
  II. Some of the column is missing in no particular pattern
- Both needs to be treated differently.
  - There are 55 Columns that have all null values – These can be dropped from the dataset.
  - There are two columns with majority of missing data

Column Name	Missing Value %
mths_since_last_delinq	64.6
mths_since_last_record	92.9

  - Also, there is a Loan Description column ('desc') contain 32% Null values. As it is a user entered text value, the data in the column is inconstant and it will not add value to our analysis.
  - These columns can be dropped.

Duplicate Values

- The duplicate values in the dataset skews the result. So, the duplicates can be dropped.
- In this dataset, there are no duplicate values.

Removal of Redundant or Constant Features

- The dataset has columns which has a constant value across rows.
- Dropped following column. (Reason- These columns have single value)

tax_liens	pymnt_plan
initial_list_status	application_type
out_prncp	acc_now_delinq
out_prncp_inv	chargeoff_within_12_mths
collections_12_mths_ex_med	delinq_amnt
policy_code	

Filtering Irrelevant Data

- After filtering the columns that are relevant to the analysis, the dataset rows needs to be filtered only for the relevant data.
- The scope of the analysis is to evaluate the Loan Applications that resulted in ‘Fully Paid’ or ‘Charged Off’ status.
- The Loans that are ongoing can be dropped from the dataset.
- Also, there are Features that are available only after the application has been accepted and Loan has been sanctioned to the borrowers. So, those features can be dropped.

Cleaning Text Data

- The Text columns can be very interesting. It may contain plain text or description, or actual values (numeric or categorical) but with a character or word like years, %, $, + symbols.
- The plain text or description column values must be examined, if it is irrelevant then it can be dropped.
    - Job Title
    - Borrower's Title
- The columns contain % or years needs to be stripped of the extra character and converted to their specific datatype.
    - Interest Rate
    - Revolution Utilization
    - Employee Tenure

Incorrect Data Type

- There are various columns that contain Integer or Float values, but the datatype is object or string.
- The Columns needs to typecasted to relevant datatype.

Feature Engineering

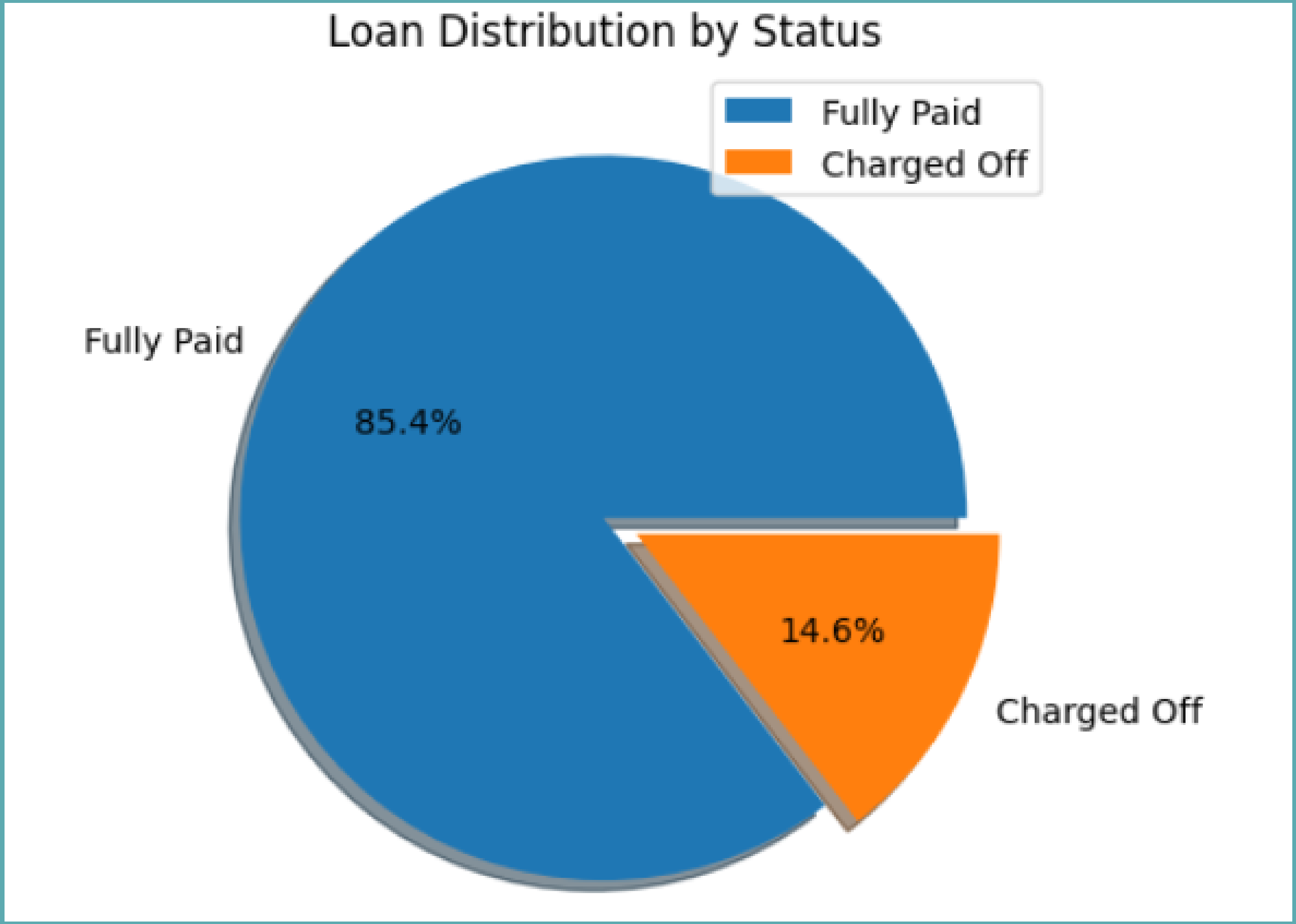
- Feature engineering is the process of creating new features or transforming existing ones to improve the quality of analysis.
- Derived columns can be created by extracting information from a column, like Year, Month, Day, Weekday or Weekend from a Date column.
- The numeric values can be converted into a segment or categorical column by defining the range.

Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is the process of analyzing and visualizing datasets to uncover patterns and correlation between the data.
- We can plot various type of charts on the dataset:
    - Univariate Analysis
    - Segmented Univariate Analysis
    - Bivariate Analysis
    - Multivariate Analysis
- After transformation we left with 38527 rows & 32 columns.

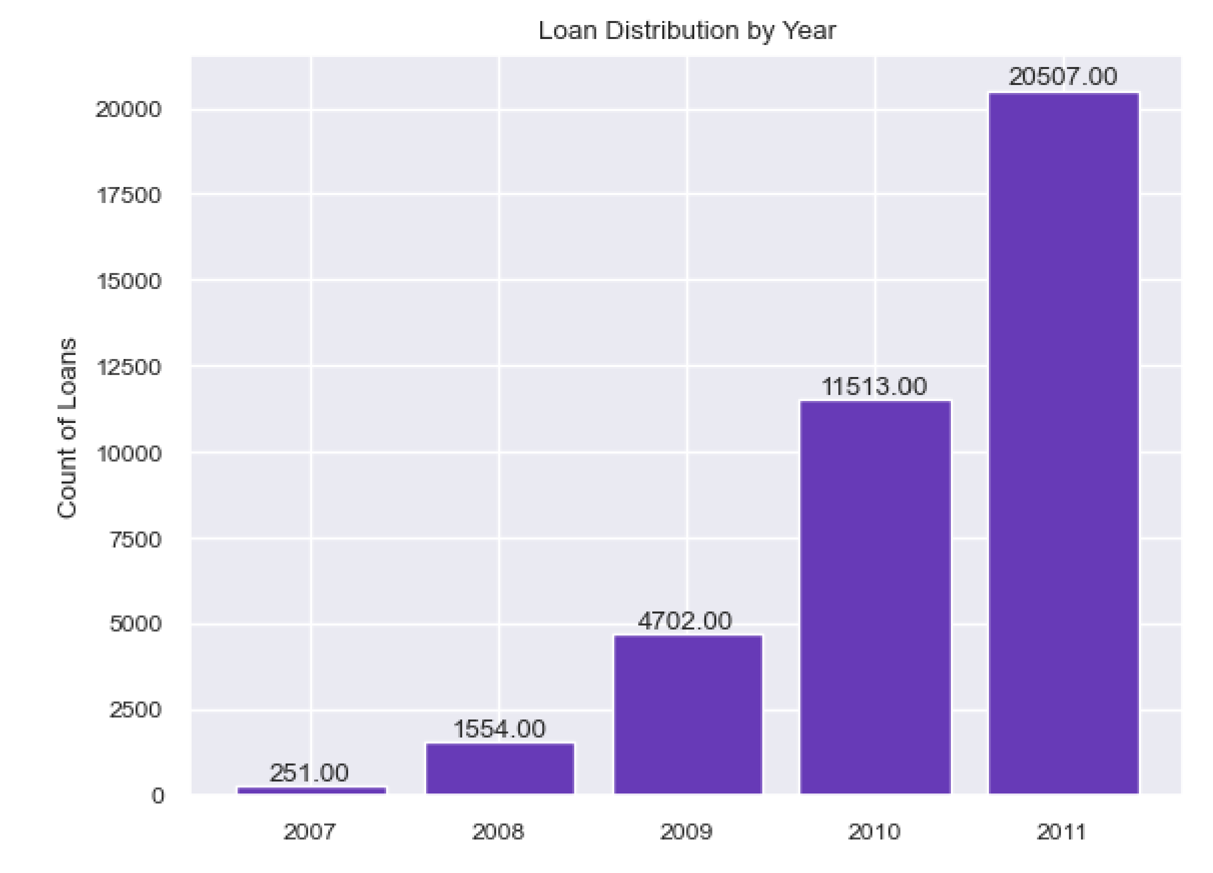
Univariate Analysis : Loan Status

- 85% of the loan is fully paid.



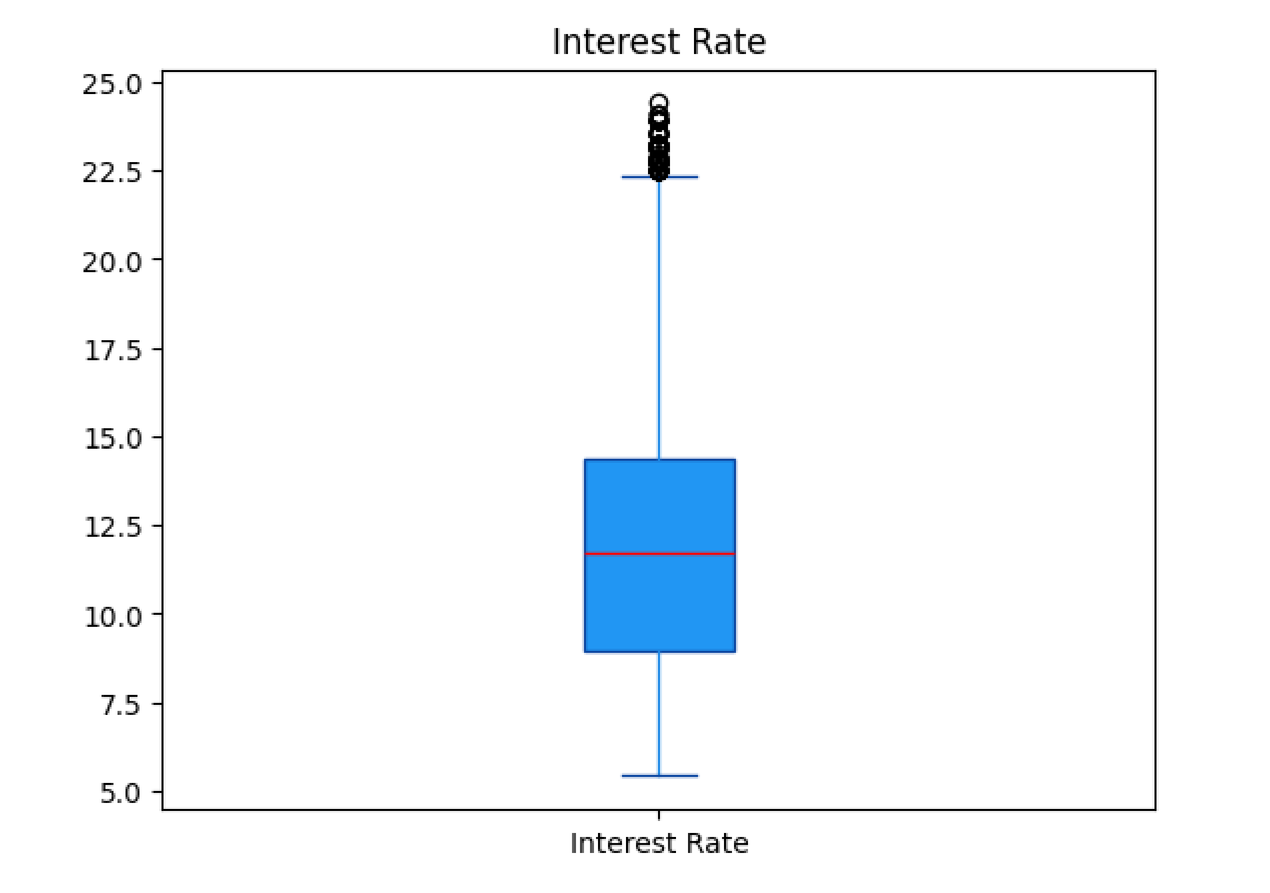
Univariate Yearly Loan Distribution

- Loan Funding is increasing YoY



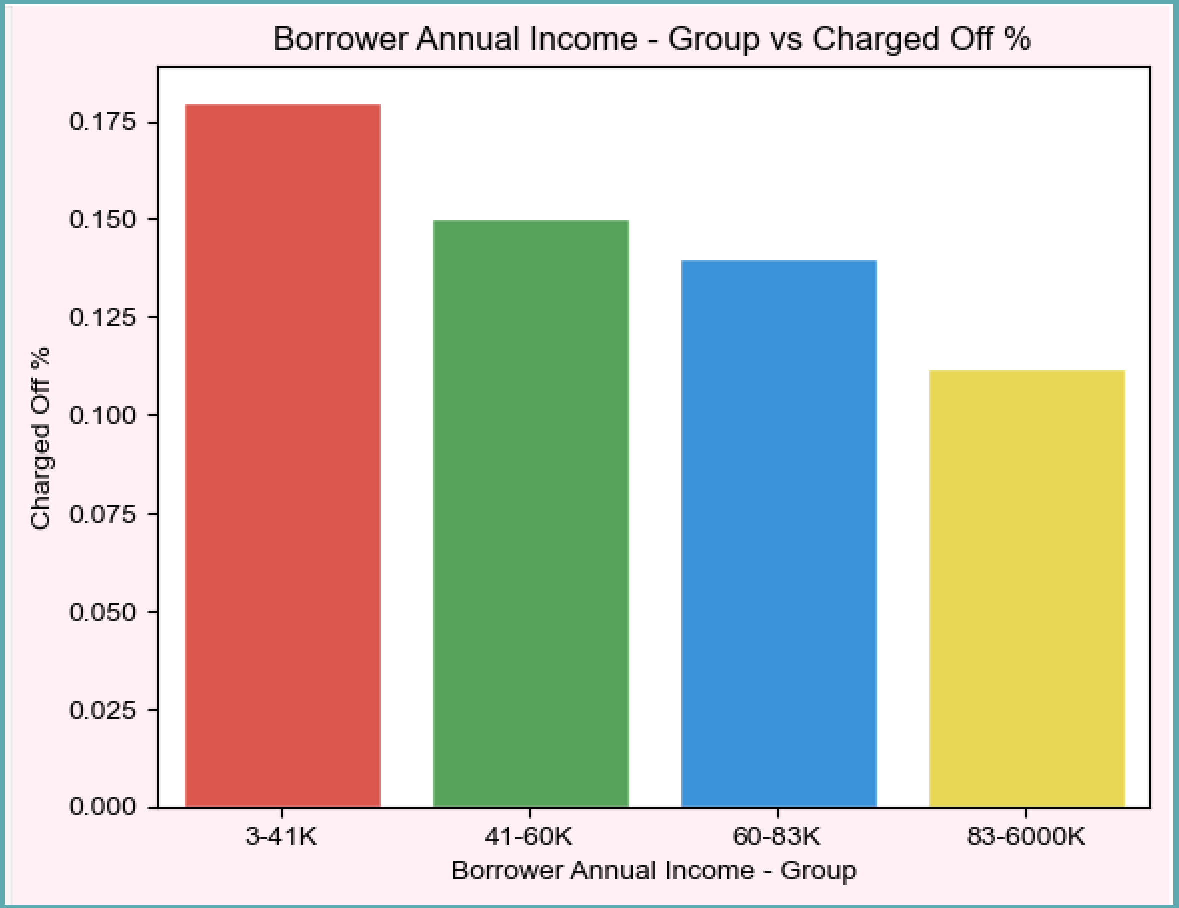
Univariate Analysis : Interest Rate

- 8-14 % of interest constitute majority of Loans.
- 65 instances of outlier constitute from 22.5% to 25%



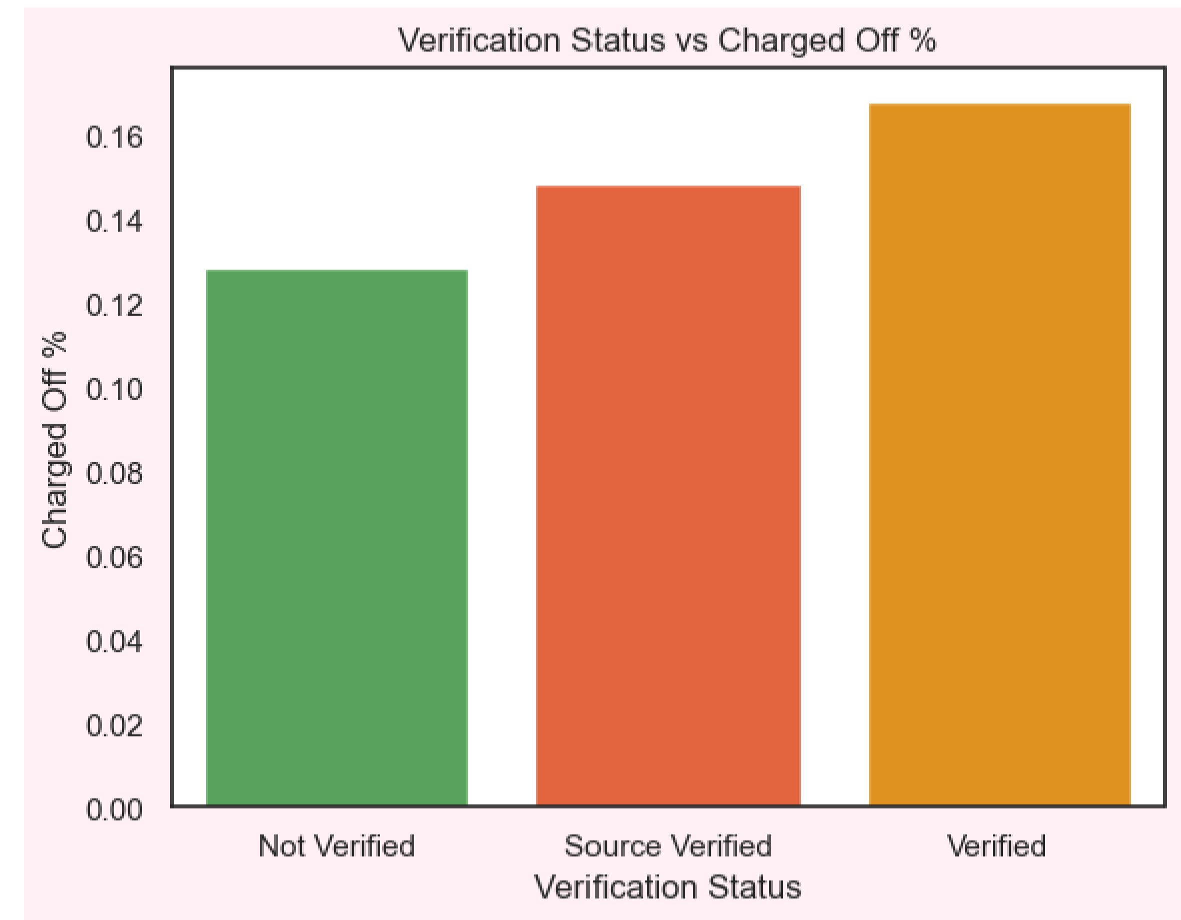
Income Vs. Defaulters

- Income group found directly correlated to defaulters life.
- Lower the income – Higher the chance of charged off and vice versa.



Bivariate Analysis : Verification Status

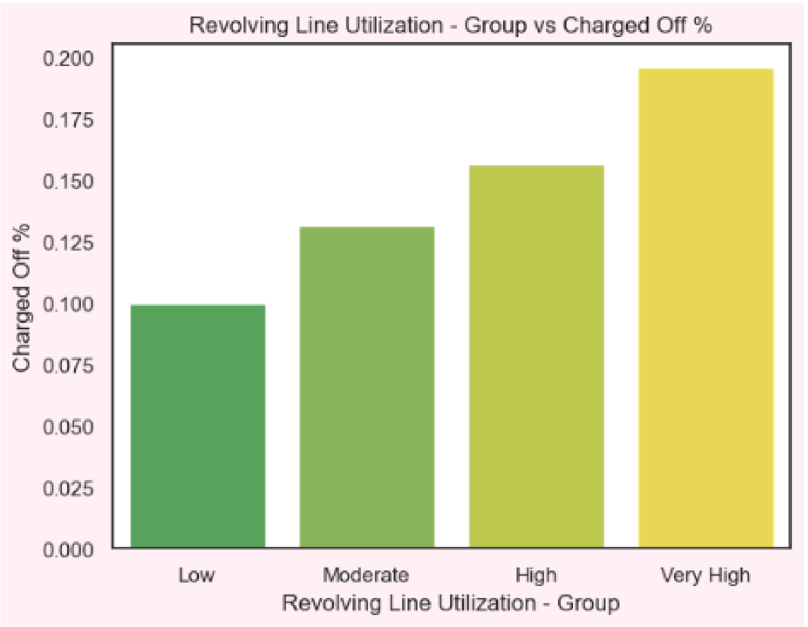
- Found customers who are not verified are less likely to default .
- Most of the customers who have defaulted are verified either by bank or by source.
- This clearly signifies bank's faulty verification system.

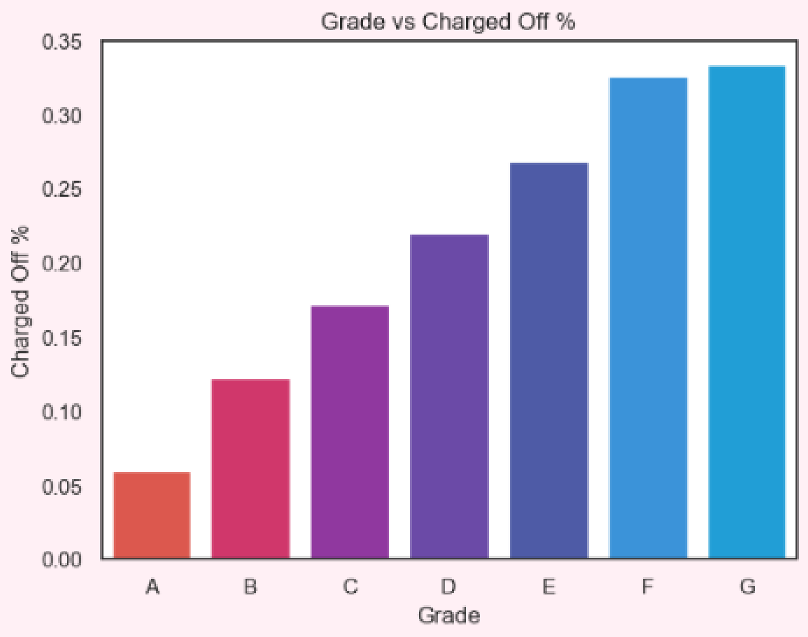


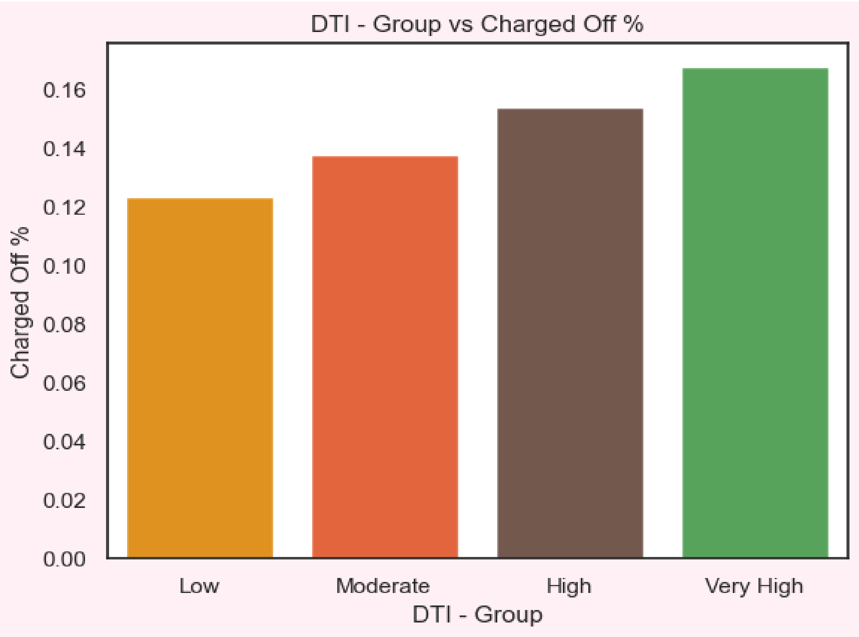
Assumption : Verified Status signifies bank internal verification system checks all the customer filled details are True

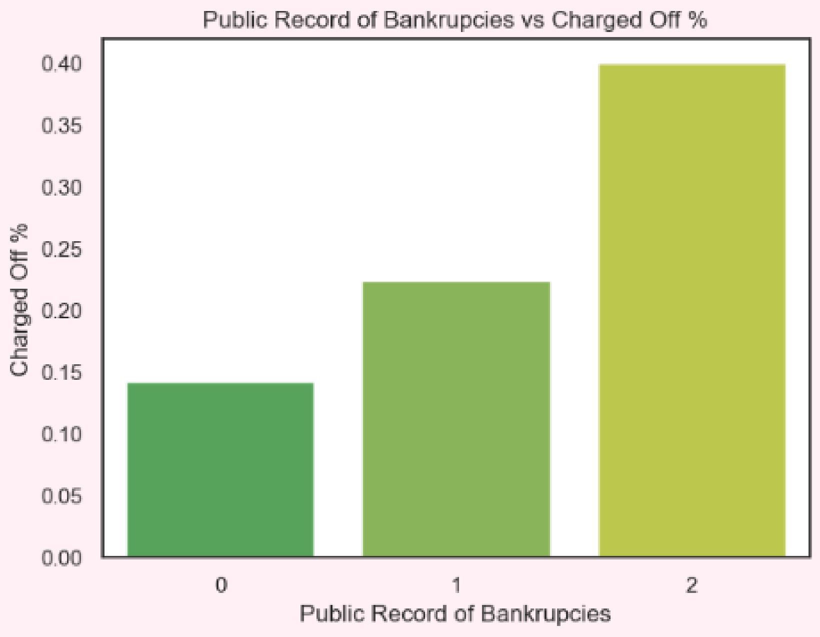
Bivariate Analysis DTI

- Following shows positive correlation with defaulters
  1. Grade
  2. Higher DTI ratio
  3. Public Bankruptcies record
  4. Revolving Line Utilization



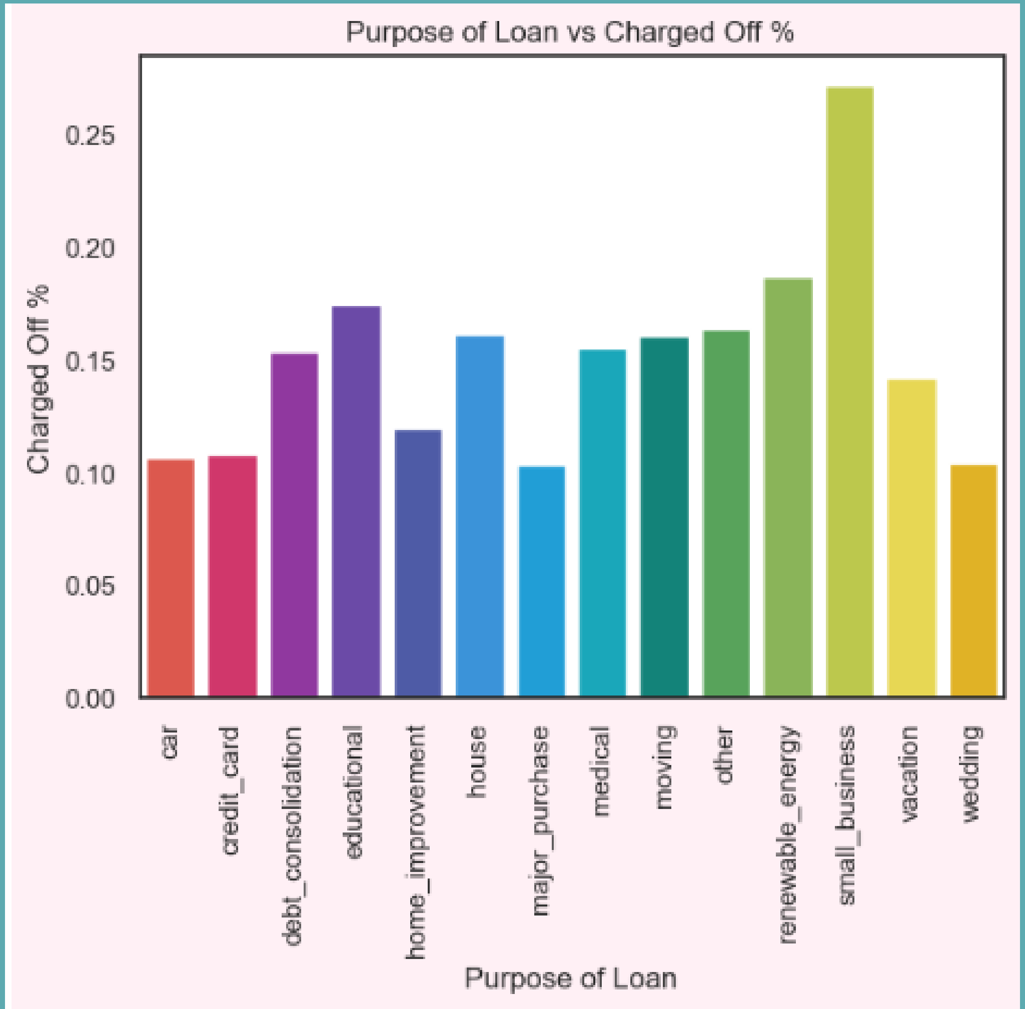






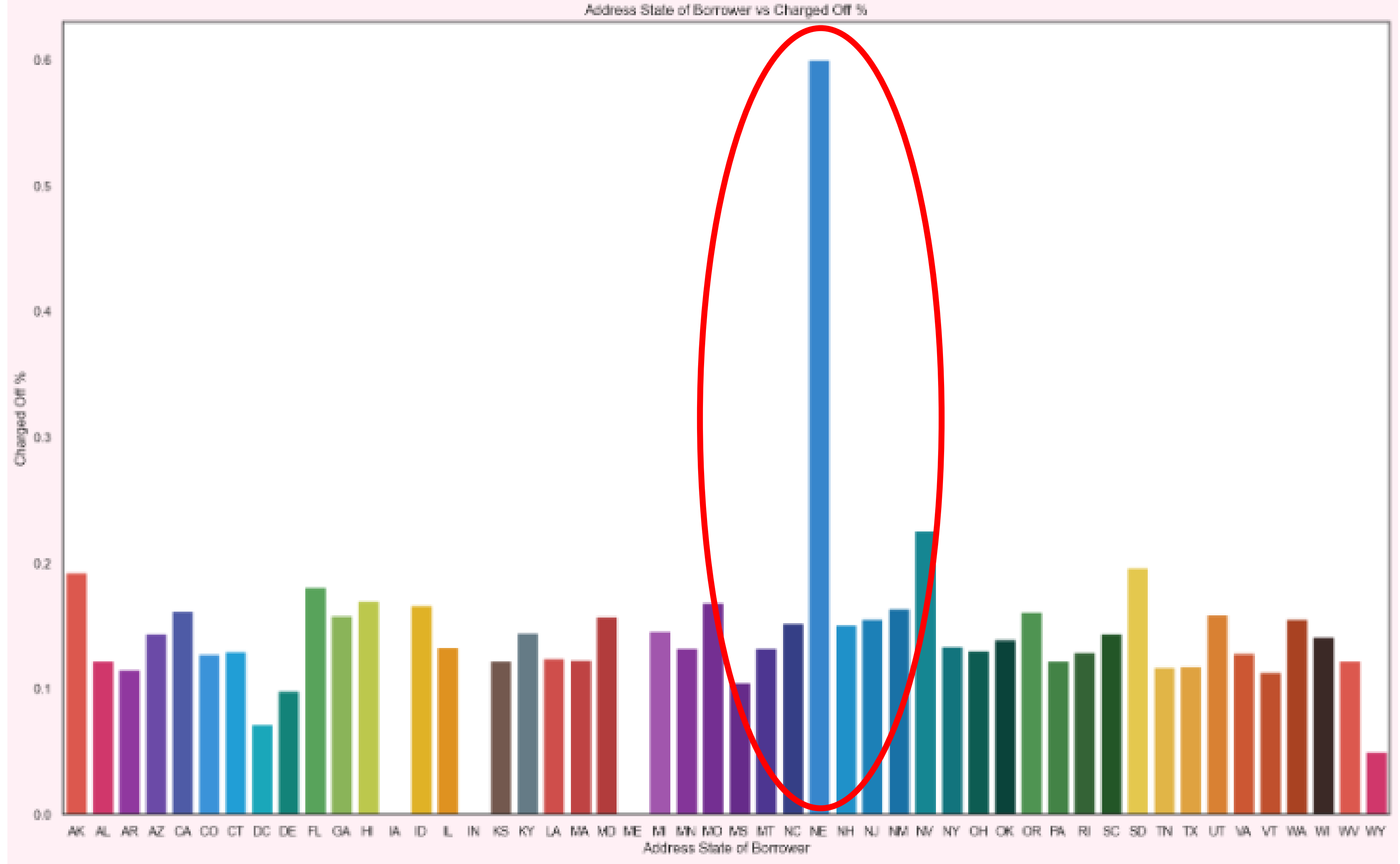
Bivariate Analysis : Purpose of Loan

- Customer taken loan stating reason as small business is likely to default the most



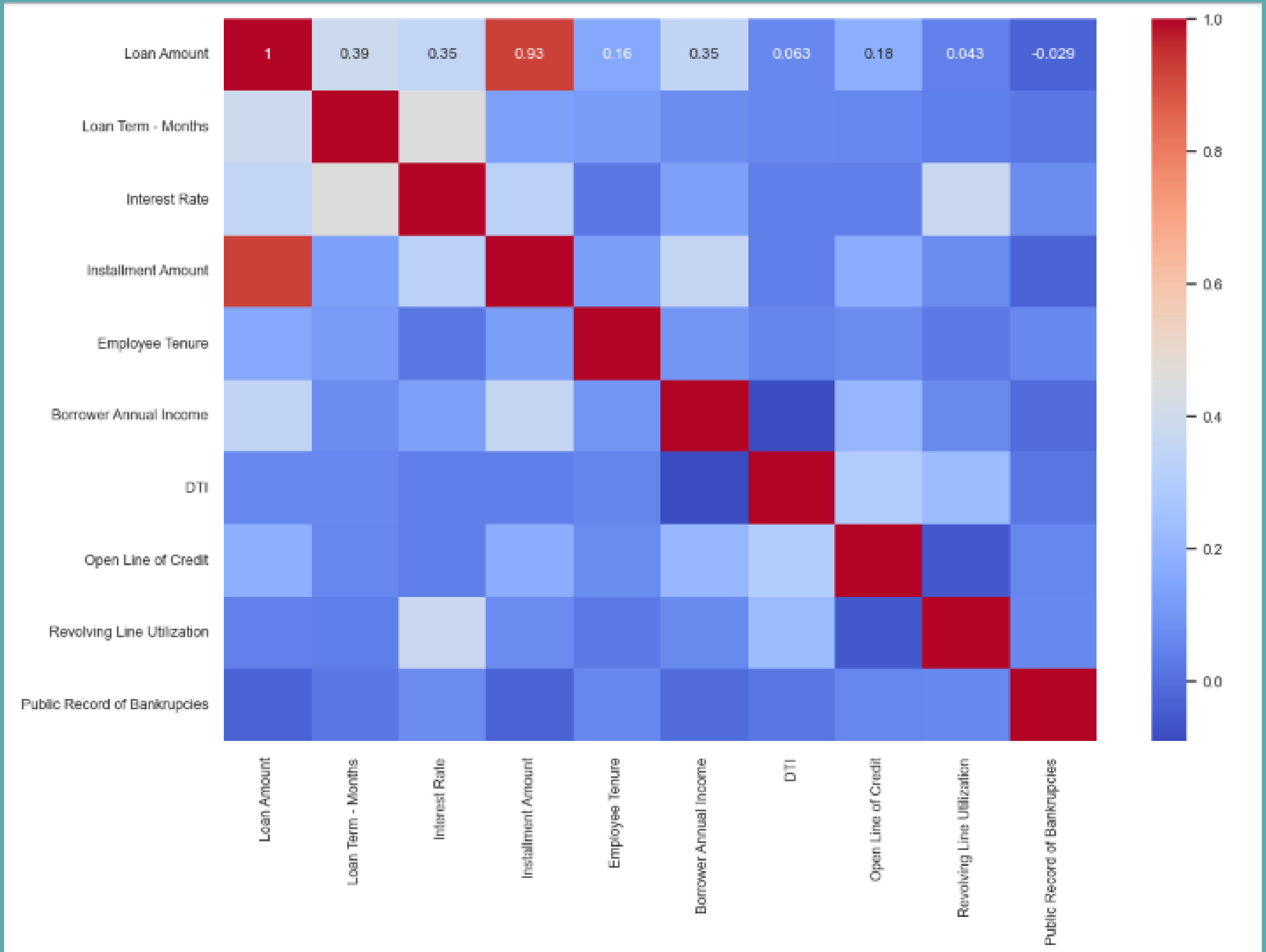
Bivariate Analysis : States

- More than 50% defaulters where resident/ addressed to city Nebraska



Multivariate Analysis

- Found highest positive correlation between instalment amount and Loan Amount.
- Also, no/negative correlation among
  - Public Bankruptcies and Loan amount
  - DTI and Borrower Annual Income.
  - Revolving line and open line of credit



Recommendations

- Company should work on the improvement of their verification system.
- State Nebraska should be put under caution, whenever resident apply for loan it will highlighted.
- Customer with bankruptcies history should have strong negative correlation with loan amount/Loan acceptance.
- While new loan sanction company should always consider following variables:
    1. Grade
    2. Higher DTI ratio
    3. Public Bankruptcies record
    4. Revolving Line Utilization
- Eg: Customer with A grade , low DTI ratio (less than 8), No bankruptcy record and low revolving line utilization can ideal loan candidate