Group Project

EDA Case Study Findings

SUBMISSION

Group Name: Data Analytics – Chennai Champs

1. Prabhakar Kalaiselvan
2. Rohit Sipani
3. Raghu Teja
4. Natarajan Ganapathi

Abstract

Background	❖ The customer is an online lending market place. ❖ Provides personal loans, business loans, and financing of medical procedures.
Objective	❖ Company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk.

Data For Analysis

Historical data of loan accounts data with fields containing following group of information:

- ❖ Customer Personal Information
- ❖ Customer Demographic Data
- ❖ Spend & Repayment behaviour data
- ❖ Past credit risk/score data
- ❖ Collection, Charge off and Recovery Data

Data to Insights: Approach

- ➢ Data preparation, Cleansing and various exploratory data analysis techniques in R
- ➢ Generation of insights through Data Visualizations & EDA in R and Tableau.
- ➢ Identify suspected list of variables contributing the "Loan Default"
- ➢ Apply the domain understanding to provide context to the analysis and to make meaningful observations and conclusions.

Solution Methodology

UpGrad

- Load data to R.
- Review structure, format and visual inspection of data
- Identify fields needing cleansing, type, format conversions

2

- Plot histograms, boxplots for cont variables & counts / freq for categorical variables.
- Plots for bivariate & segmented bivariate to understand dependencies.
- Create separate datasets/filters/ aggregates as needed

4

- Consolidate observations from each of the analysis and plots.
- Present observations, context and explanations to include/exclude variables.

6

- Review and Understand metadata dictionary.
- Review Datatypes and formats
- Review literature on Credit Risk scorecard models and potential data fields used in analysis.

1

- Delete fields containing all/ majority NAs/Blanks/0 etc.
- Delete fields that are not relevant to analysis based data distribution
- Delete fields that are related to repayment behaviour, collections & recoveries etc.

3

- Identify initial suspect list of variables out of plots and analysis
- Understand the business context of the analysis and apply accordingly to scrutinize the list
- Identify the final list of variables

5

- Summarize final list of variables.
- Summarize Insights and Conclusions

7

Business Understanding & Metrics used for Analysis

KPIs / Measures Used

- Number of Defaults : Count of loan records where status is Charge off
- Default Ratio: Number of Defaults / Number of Loan records

1. Loan Characteristics

loan_amnt
term
int_rate
installment
purpose
verification_status
issue_d

2. Customer Info & Demographics

emp_title
emp_length
home_ownership
annual_inc
zip_code
addr_state
title
grade & sub_grade
dti

3. Customer Payment Behaviour

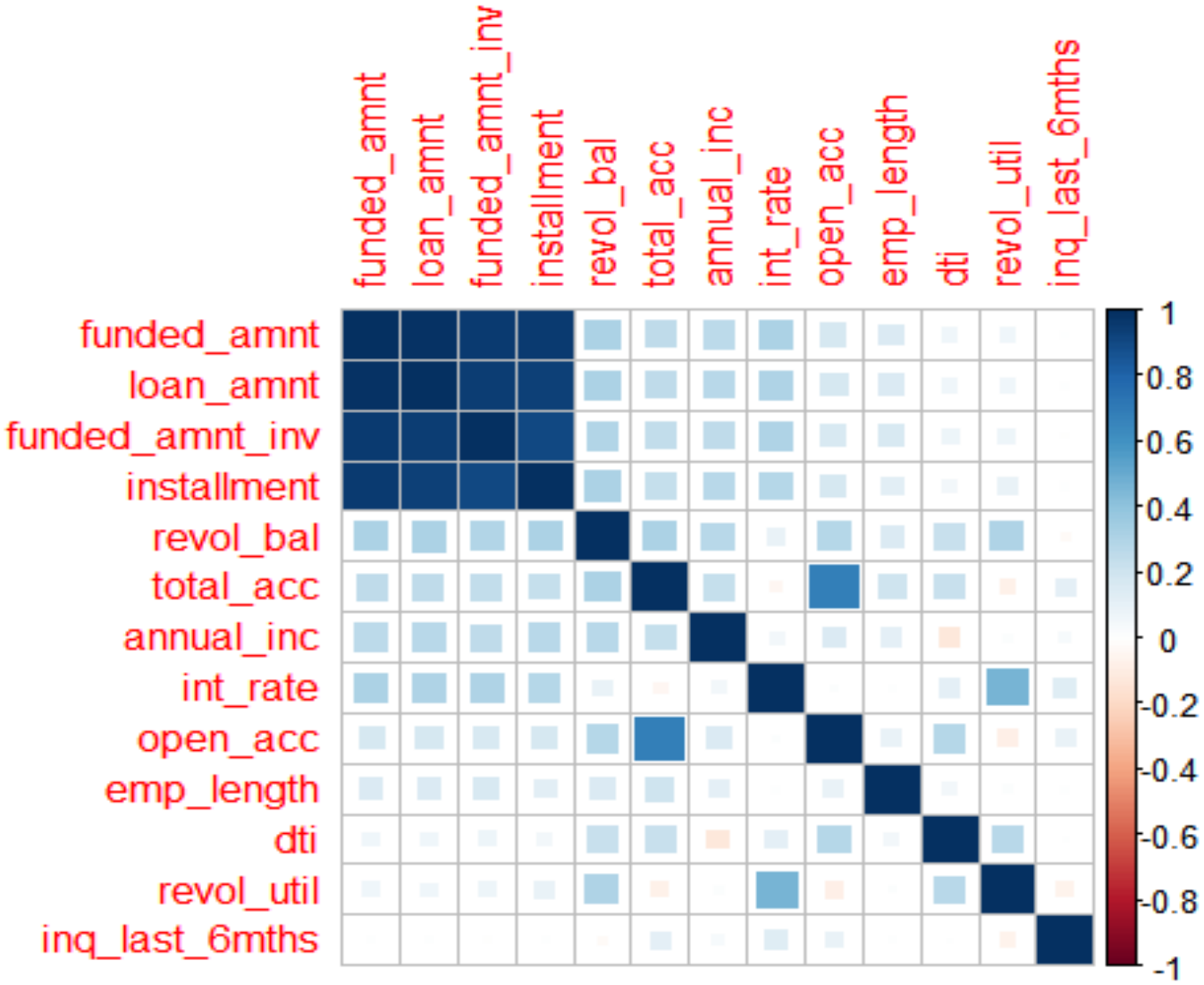
out_prncp
total_pymnt
recoveries
last_pymnt_d
last_pymnt_amnt
next_pymnt_d
delinq_2yrs
revol_bal
revol_util
total_acc
chargeoff_within_12_mths

Note: Only sample fields are mentioned here. Detailed list is part of the R code in comments.

Data Quality Issues & Cleansing

Data Quality Issues	Action Taken
More than half of the fields just contained NAs	Fields were removed.
There were fields that contained same data value or default values for most records	Fields were excluded from Analysis & Removed.
There were fields that contain most of the values as Zero	Fields were excluded from Analysis & Removed
There were numerical fields which were not rounded off	Rounded off those numerical fields
Interest Rate and Revolving Utilizing Rate were present as character format with having “%” symbol	“%” symbol was removed and was converted to numeric
There were fields which was free text field and had many unique values	Fields were excluded from Analysis & Removed
Fields with month and year information was not stored in dataframe as Date	Field was converted in date format and additional fields for year and month was derived

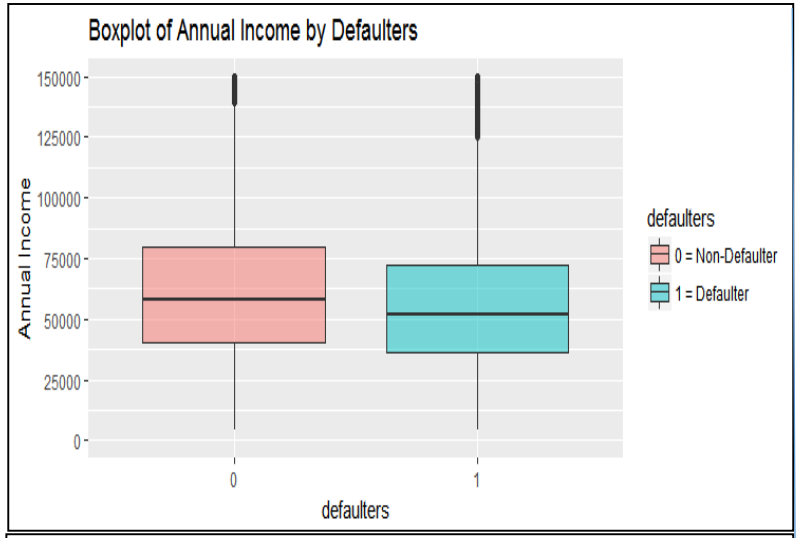
Correlation Matrix for Numerical Variables

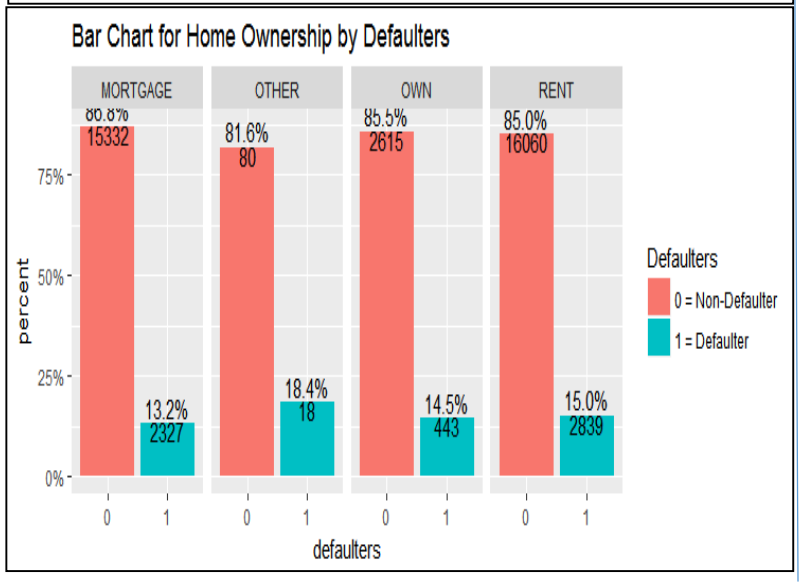


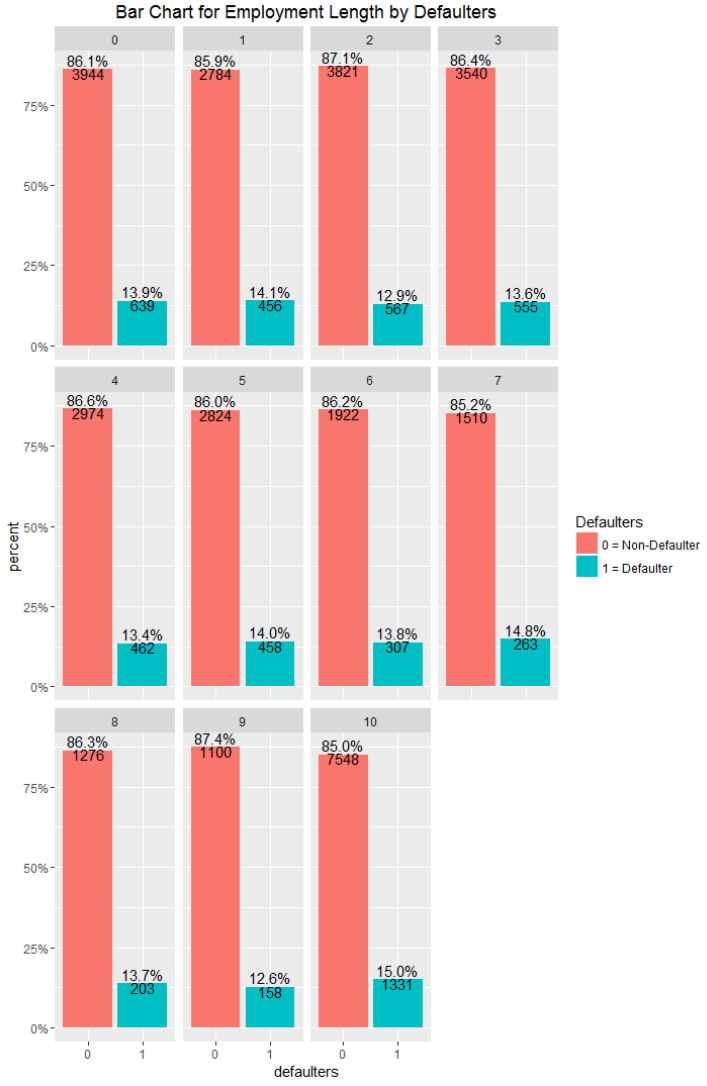
Observations

1. Loan amount, funded amount, funded amount by investor and instalment are highly correlated to each other and would provide the same insights. Hence, only one variable could be kept for EDA analysis and other variables could be removed.

Customer Demographics
Income, Home Ownership, Employment





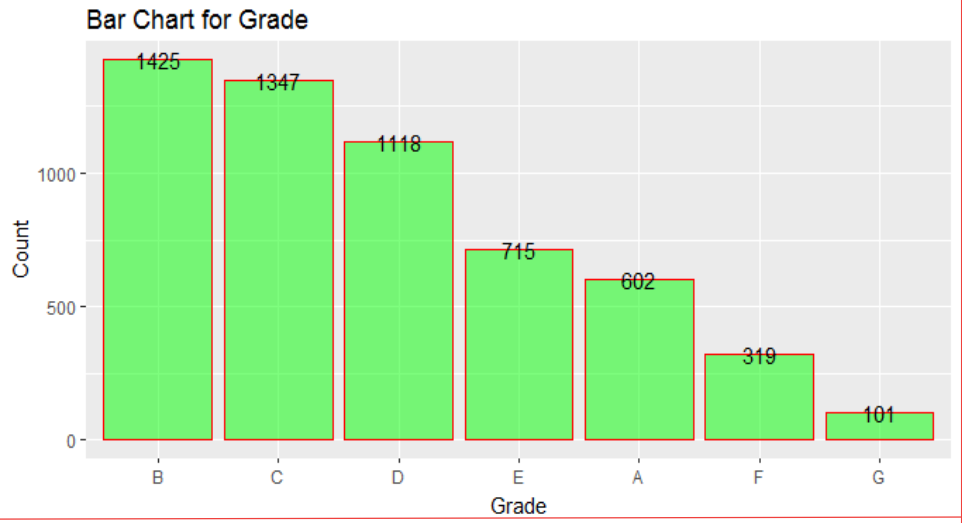


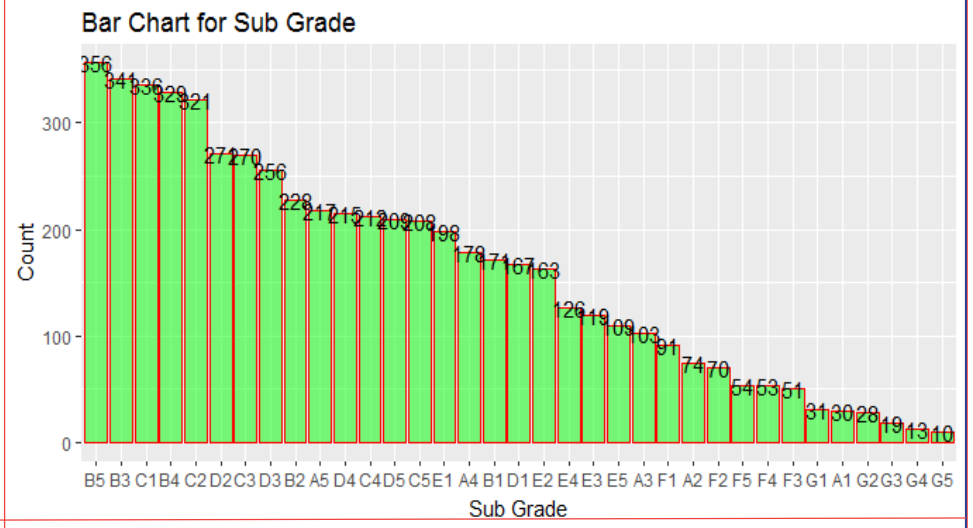
Observations

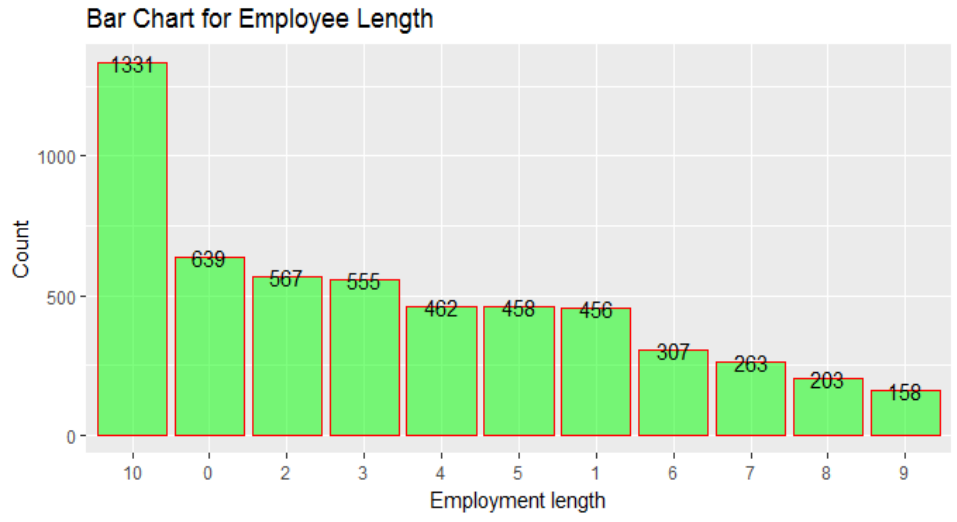
1. Loan given to borrowers having lower annual income are most likely to get defaulted
2. The loan are more likely to default for the borrowers having home ownership as “other” category
3. The loan are more likely to default for the borrowers having employment length of 10 years or more

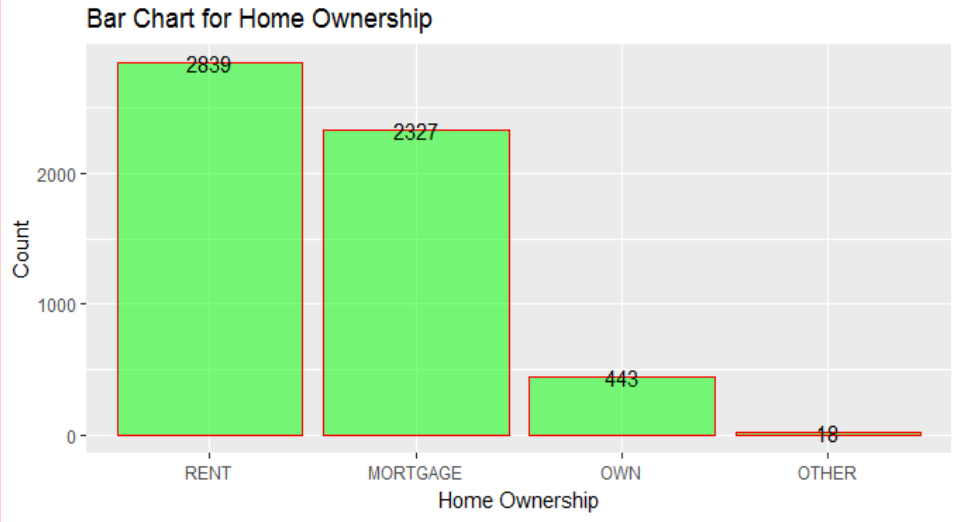
Note: Only Sample visuals included in the deck. Rest of the visuals are in R code submitted.

Customer Demographics
Grade & Sub-Grade (Credit Rating), Income, Employment







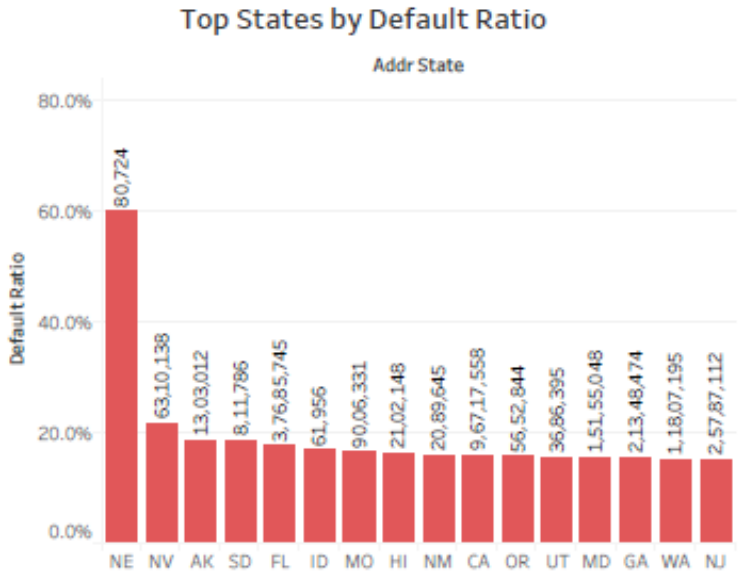


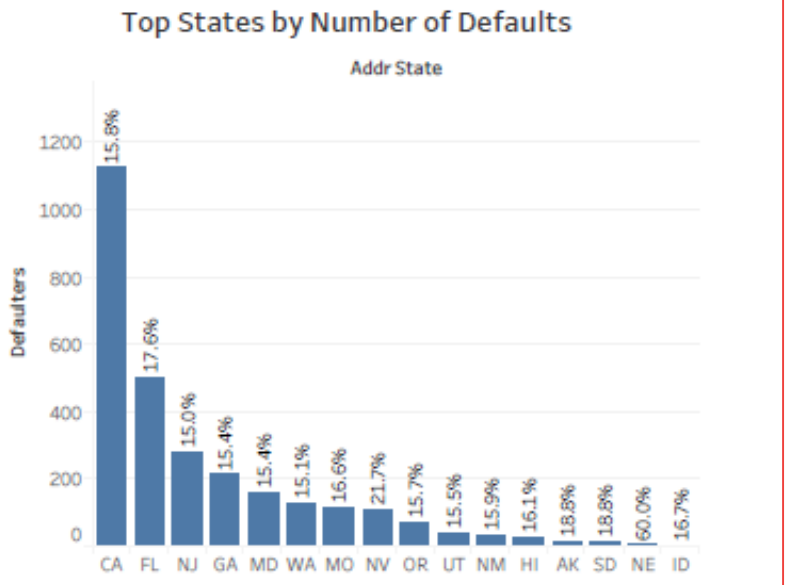
Observations

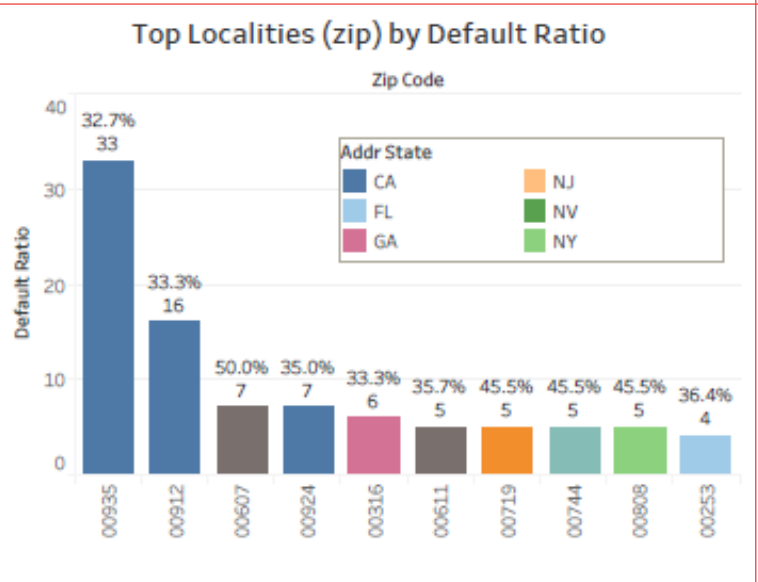
1. Customers with Loan Grades B, C, & D shows high number of defaults.
2. Similar pattern is shown in the Sub Grades also with B, C & D sub grades are contributing to high number of defaults.
3. Number of loans defaulted was higher when borrowers was having employment length of 10 Years or more.
4. People staying in rented home or mortgaged their property seems to be defaulting more than people living in their own homes.

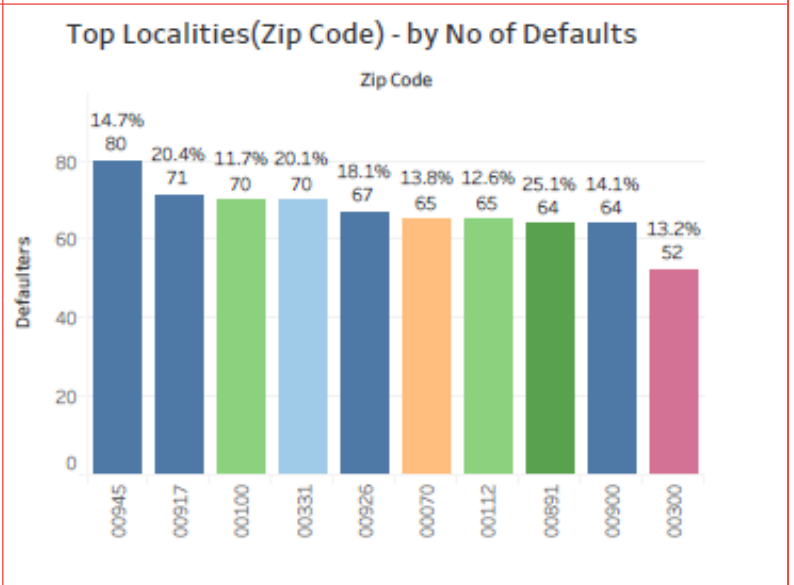
Note: Only Sample visuals included in the deck. Rest of the visuals are in R code submitted.

Demographic Analysis – By Geography





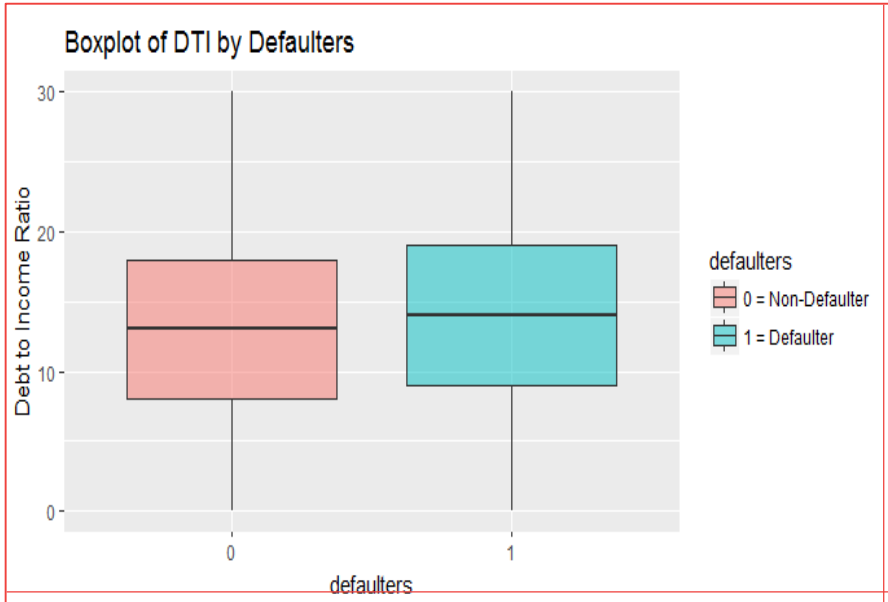


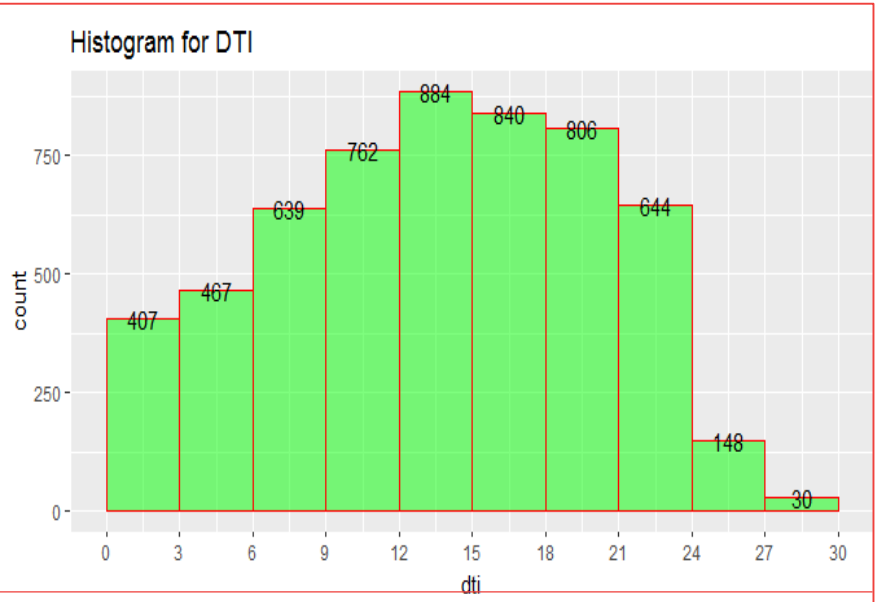


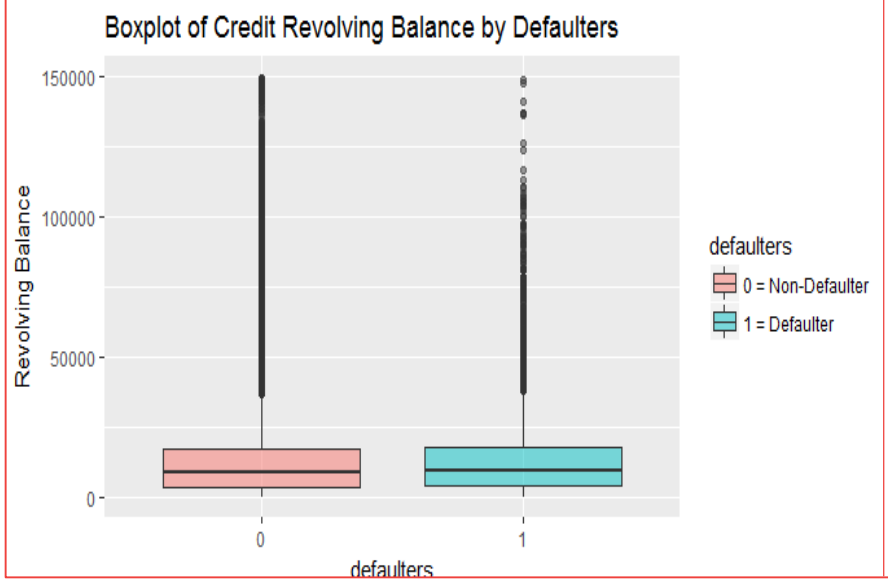
Observations

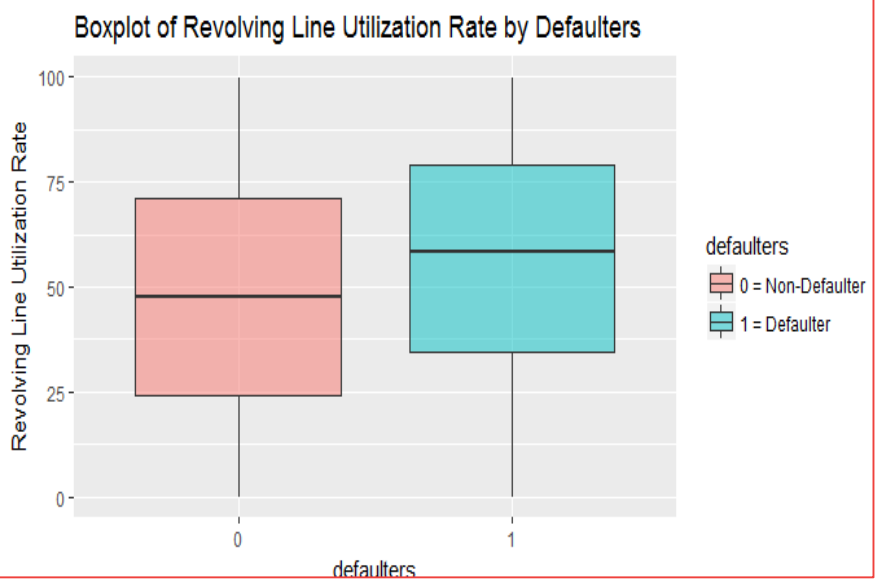
1. California(CA), Florida(FL), NewYork (NY), Texas (TX) & New Jersey(NJ) were the top 5 states by number of defaults hence maximum losses.
2. Going by the default ratio, NE seems to be the state with maximum default ratio. Given number of loans in NE overall, difficult to conclude. Followed by NV, SD, AK, & FL states in-terms of Default Ratio.
3. Similar analysis is also done to identify the top zip codes where the defaulters are located.
4. More investigation required about the top localities as pointed out in the visuals.

Behavioural Variables (Spend behaviour)
Credit Limit Usage, Account Balance & DTI







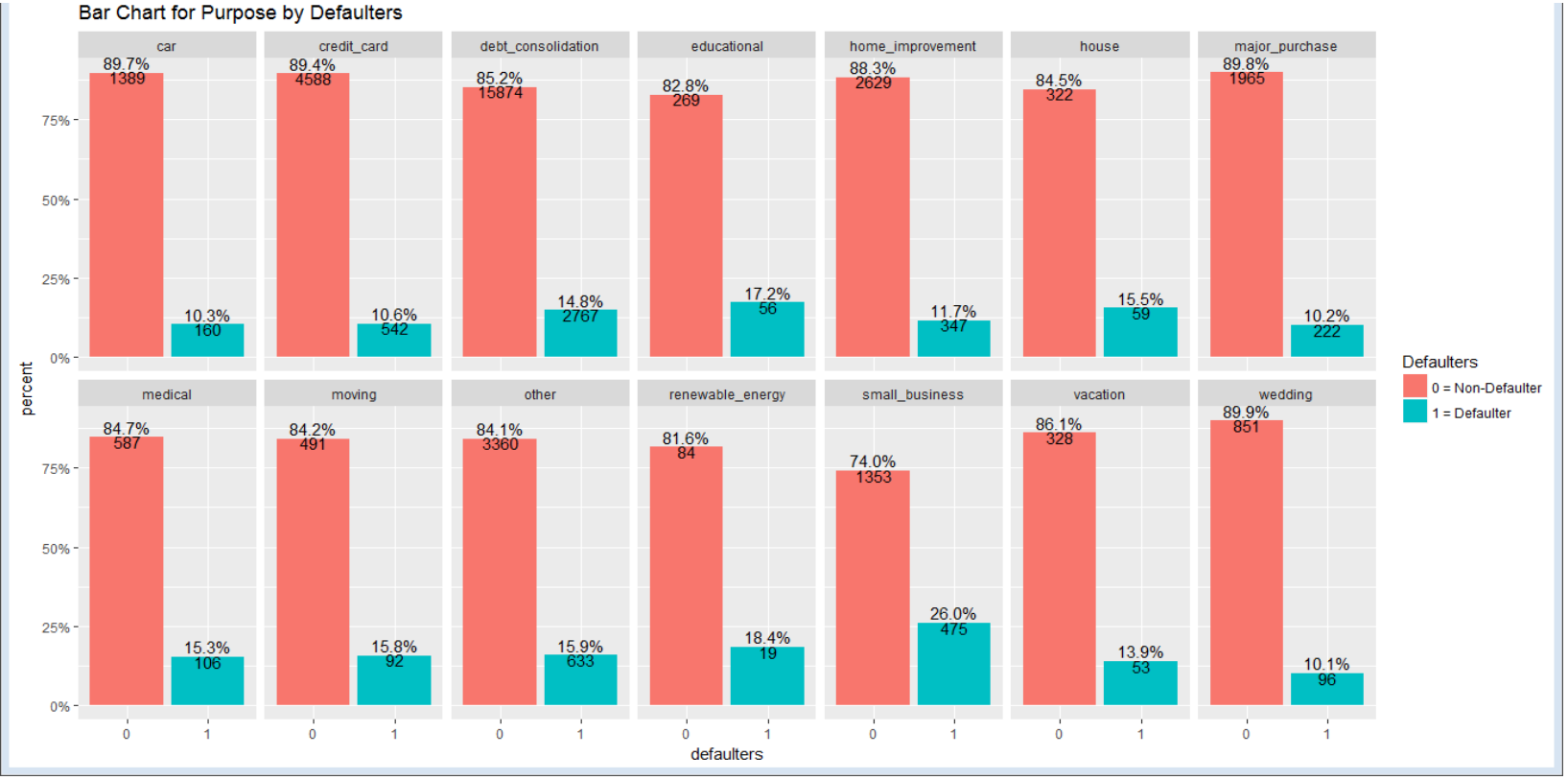


Observations

1. Number of loans defaulted was higher for borrower having debt to income ratio is between 8% to 25%.
2. Given that dataset did not contain any data points at DTI > 30%, we can safely state that higher the DTI, higher is the chance of default.
3. Higher the account balance (rev balance), higher the chances of default.
4. Similarly, people who utilize the credit limit more seems to exhibit higher rate of default.

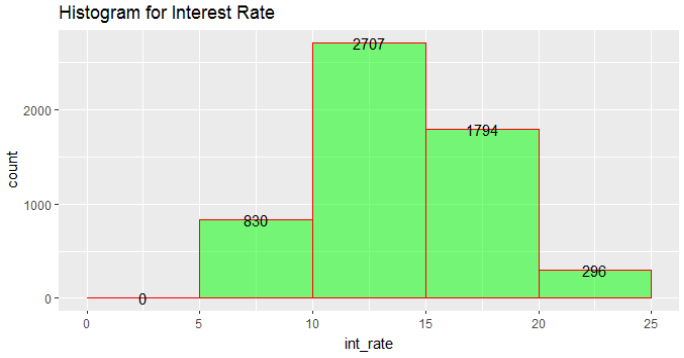
Assumption: If DTI > 30%, we assume the loans are not approved at all, as it directly impact the repayment capability of the borrowers.

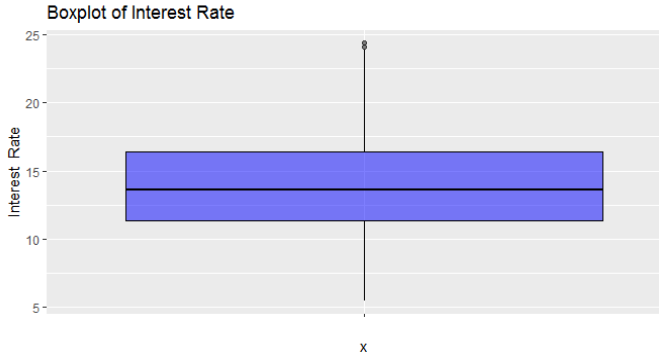
Application Variables (Purpose & Interest Rate)



Observations

1. When the purpose of the loan is **debt consolidation**, then it has higher chances of default, followed by other, **home improvement**.
2. Looking at the ratio, the top two purposes with higher chance of default are **small business** & **education**
3. Higher Interest Rate shows higher default likelihood as it directly impacts the borrower repayment capability.





Conclusions - Summary of Problems

Variable	Explanation
❖ **Term of Loan**	Long term loans have higher chance of default
❖ **Interest rate**	Higher the Interest Rate, Higher the borrower playout, increasing the burden on Borrowers. Hence high Interest is likely to cause higher default. Even banks charge higher interest rate on high risk loans.
❖ **Grade & SubGrade**	Grade & Sub-grade themselves are some sort of credit ratings assigned by rating agencies or lenders to the borrowers. Hence these can impact the possibility of default. Poorer the score, higher the chances of default.
❖ **Emp Length**	People working for more than 10 years are likely to default more. Possible explanation can be that this will be mid-career segment, the segment that is likely in significant job/employment stress with higher expenses.
❖ **Home Ownership**	People staying in rented home OR mortgaged their property likely to default more. Higher rent payments and mortgage charges directly affect their repayment capability.
❖ **Annual Income**	Lower the Income, higher the chances of default.
❖ **Purpose of Loan**	Debt consolidation has shown highest number of defaults and small business shown highest rate of default. People doing debt consolidation are already under the loan burden and small business have higher risk in terms of business success and cashflow explaning the higher defaults.
❖ **Address State and Zipcode**	Customers from certain states and localities shown higher rate of default. This needs geo specific analysis.
❖ **DTI by Defaulters**	Higher the DTI, higher the chances of default. It directly affects the repayment capability of the customers as they are already repaying loans from their income.
❖ **Rev line utilization rate**	High utilization indicate high spending, hence higher risk of default.