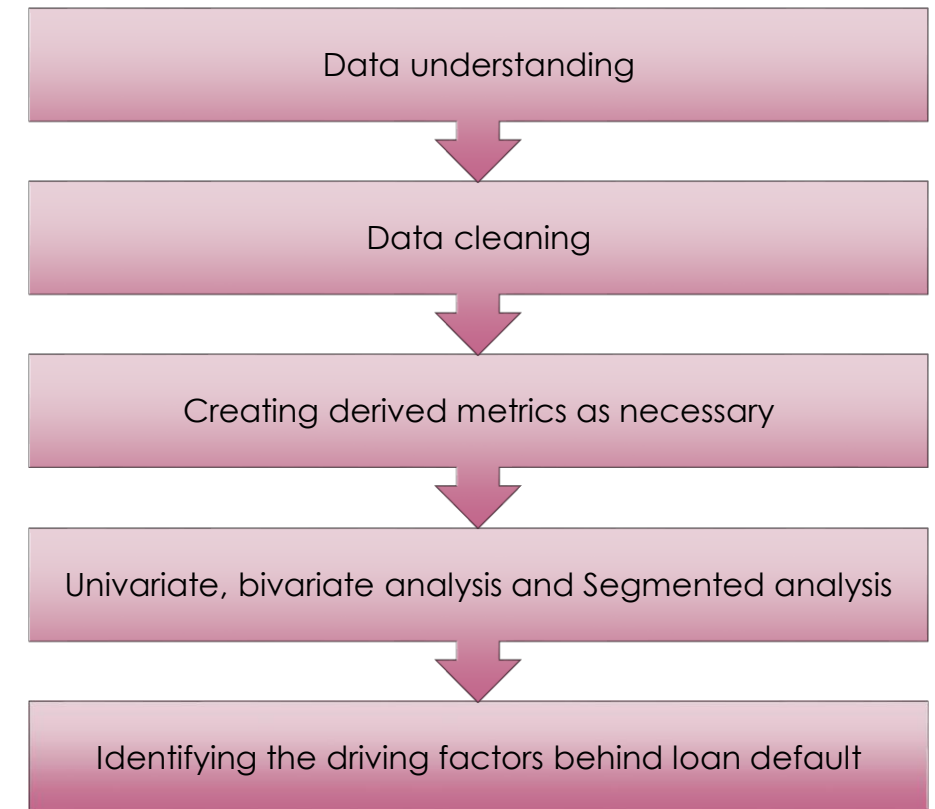# EDA Group Case study – Loan defaulters analysis

PRADNYA PAITHANKAR

DR. RAJENDRA WARKE

VAIBHAV PALKAR

SAMEER SINHA

INTRODUCTION

- Preamble: Consumer finance company faces challenge of
  - Lending loans to 'risky' applicants
  - Controlling financial loss (called credit loss) due to risky applicants
- Company wants to support portfolio risk assessment data based intelligence
- Case study objective
  - To identify the risky applicants
  - understand the driving factors (or driver variables) behind loan default
- Inputs available
  - Loans data with 39717 records and 111 data fields
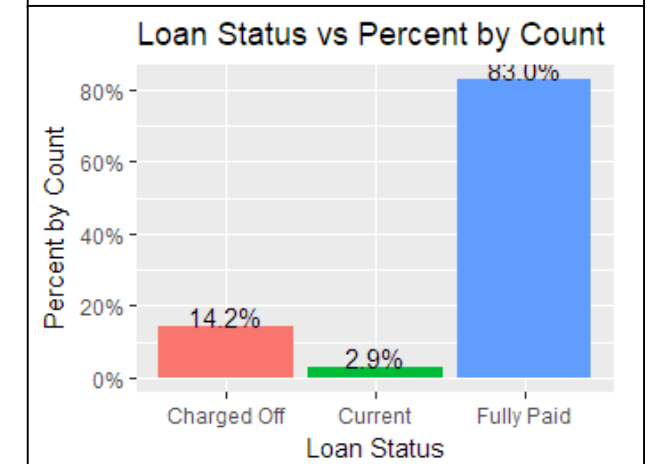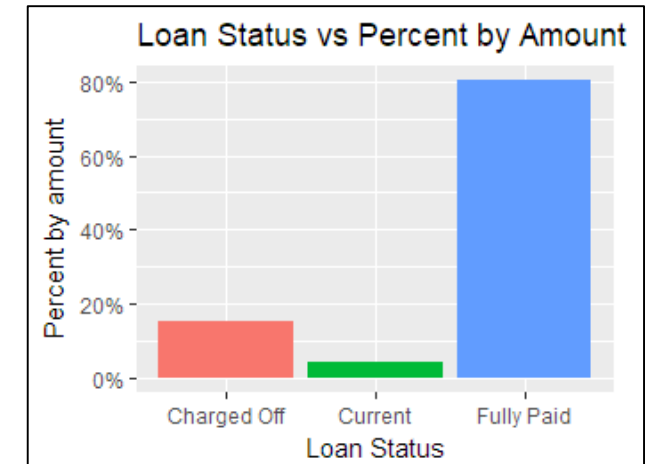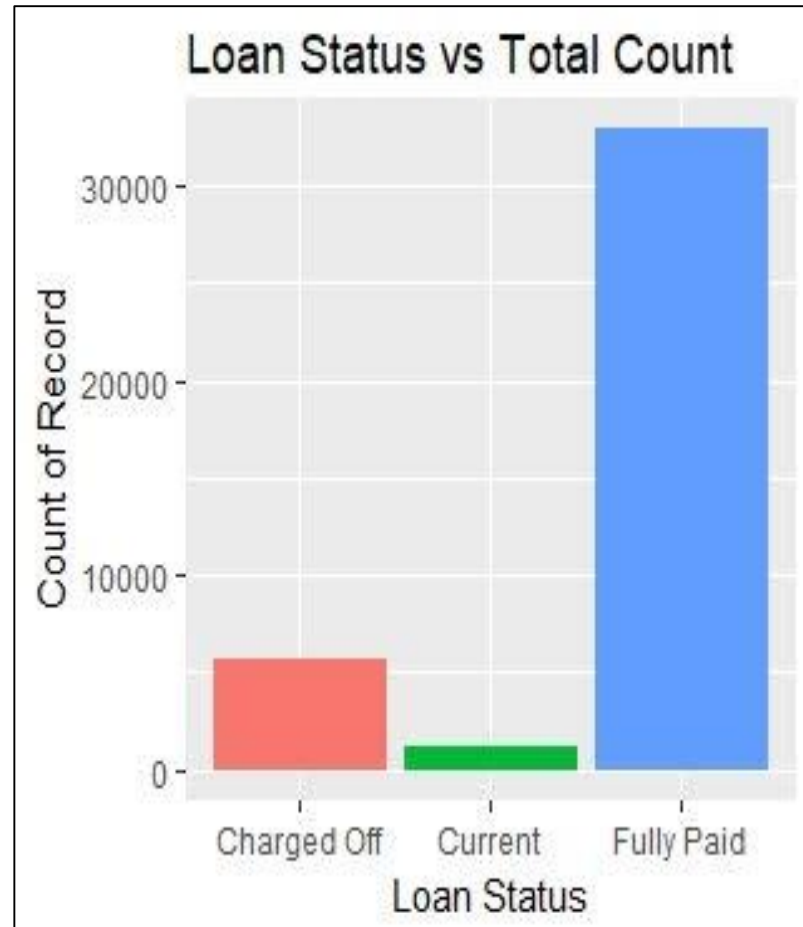
Methodology followed

Data understanding

↓

Data cleaning

↓

Creating derived metrics as necessary

↓

Univariate, bivariate analysis and Segmented analysis

↓

Identifying the driving factors behind loan default

Input data – Preliminary understanding

The loans are having following status and distribution

- \# 32950 fully paid their loan
- \# 1140 current running loan
- \# 5627 candidates are defaulter

▶ % by count of loans

- Fully Paid - 83%
- **Charged off – 14%**
- Current – 3%

▶ % by sum of loan amounts

- Fully Paid - 80.5%
- **Charged off –15%**
- Current - 4.5%

IIITB Upgrad EDA Group case study

Loan Status vs Total Count

Loan Status vs Percent by Amount

Loan Status vs Percent by Count

# Data Cleaning - Basic

- Checked duplicated rows – No duplicate records found

- Removed all those columns where

  - All values are NA

  - All values are same and unique - This will not impact the analysis

  - Values are either zero or NA – This will also not impact the analysis

  - **This reduced the data set to 48 variables**

- Removed % sign from columns to make it continuous numeric variable – int_rate, revol_util

- Converted annual_income into integer

- Removed current loans - To analyze defaulters in comparison with fully paid loans, again reduced columns with all values same and unique

 - **This reduced data set to 38577 rows and 46 variables**

Data Analysis – Variables with impact

- **Grade and Interest rate**, as they are linked, have same impact on the defaulters

- While lending, lower credit risk **Grades** leads to higher interest rates, however it does not reduce defaulters

Grade vs Interest Rate

Grade vs Percent by Count

Interest Rate vs Percent by Count
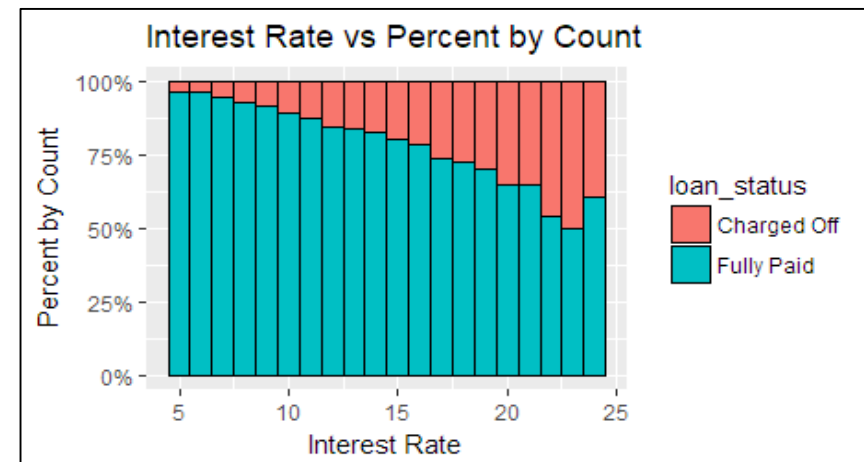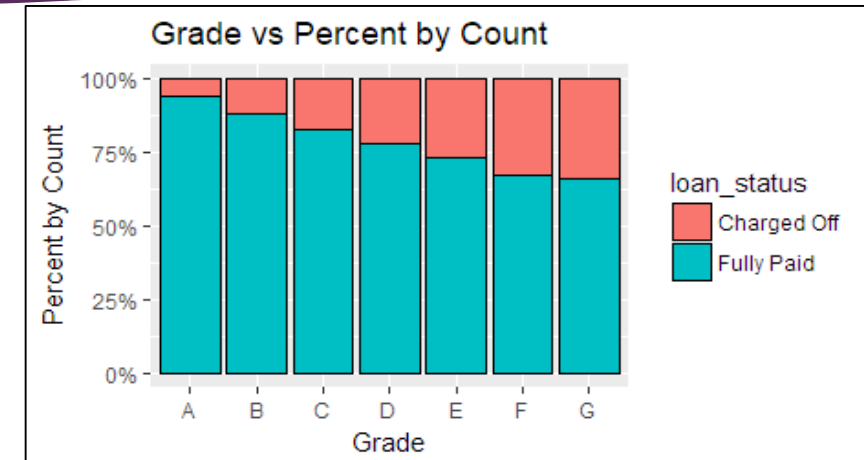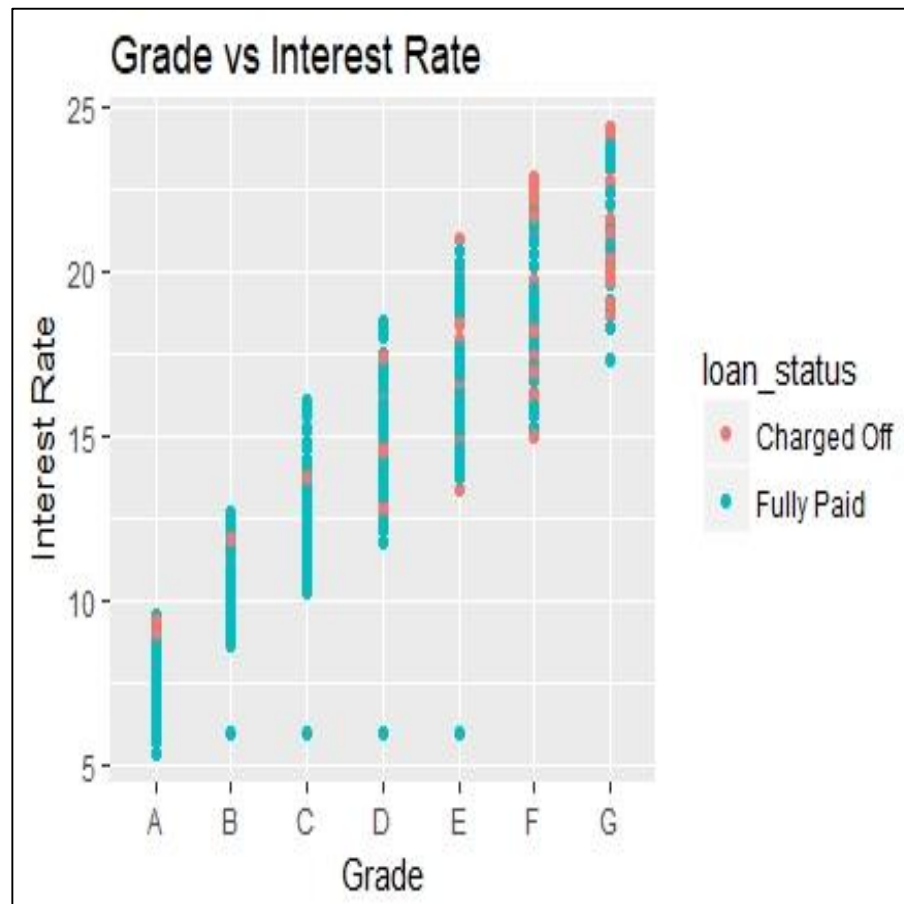
Data Analysis – Variables with impact

▶ Maximum number of loans are availed by people with **annual income** <300000 (Hence graph is zoomed for the same limit)

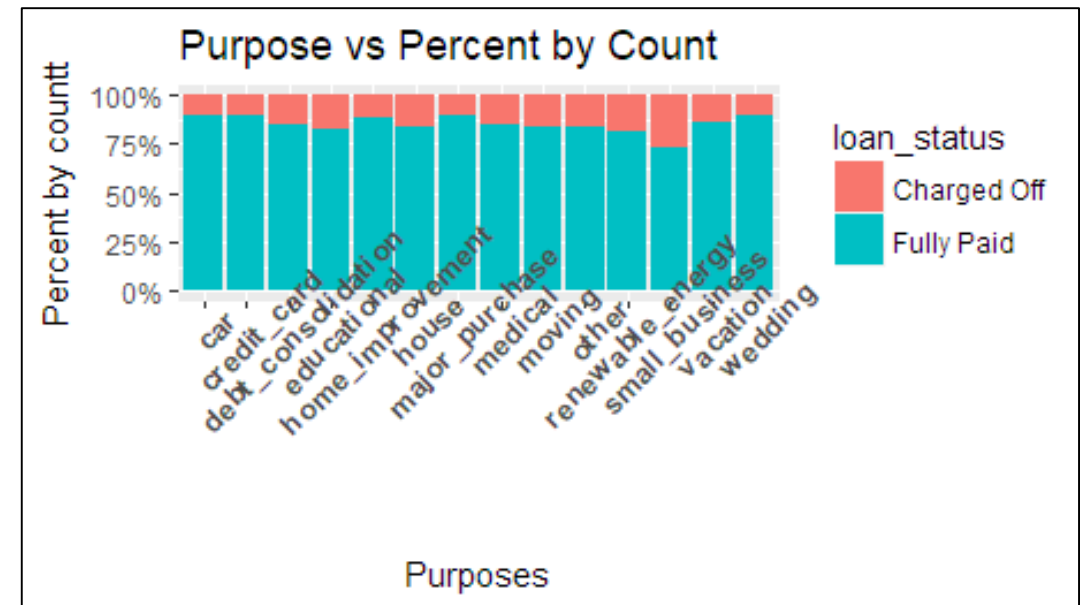▶ % of loans getting defaulted are more for the annual income <225000
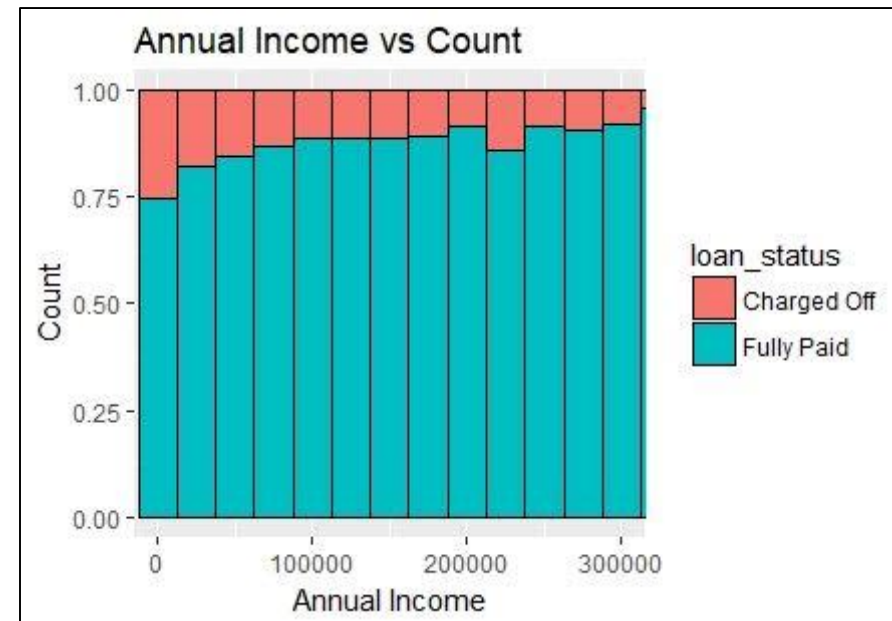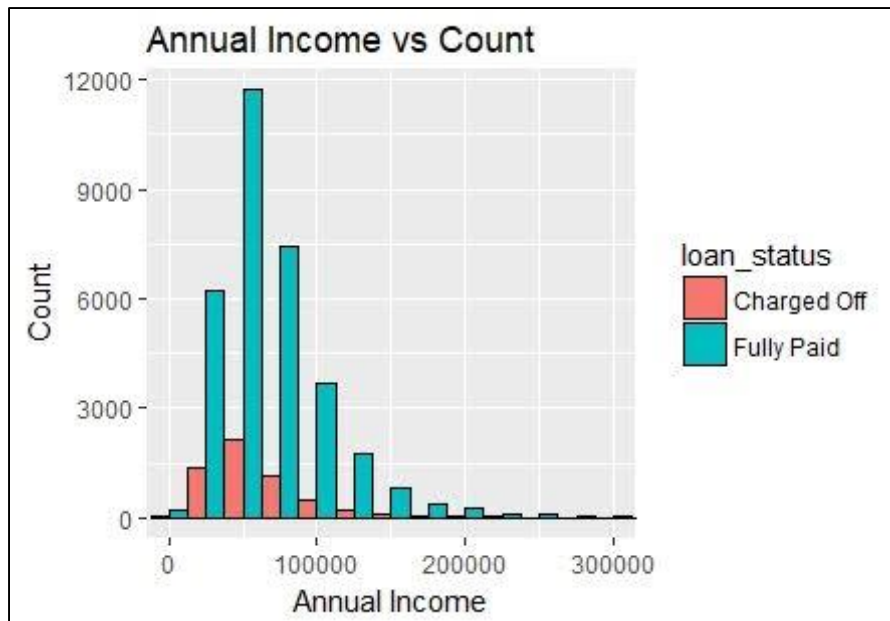
Data Analysis – Variables with impact

▶ **60 months payment term** leads to more defaulters percentage
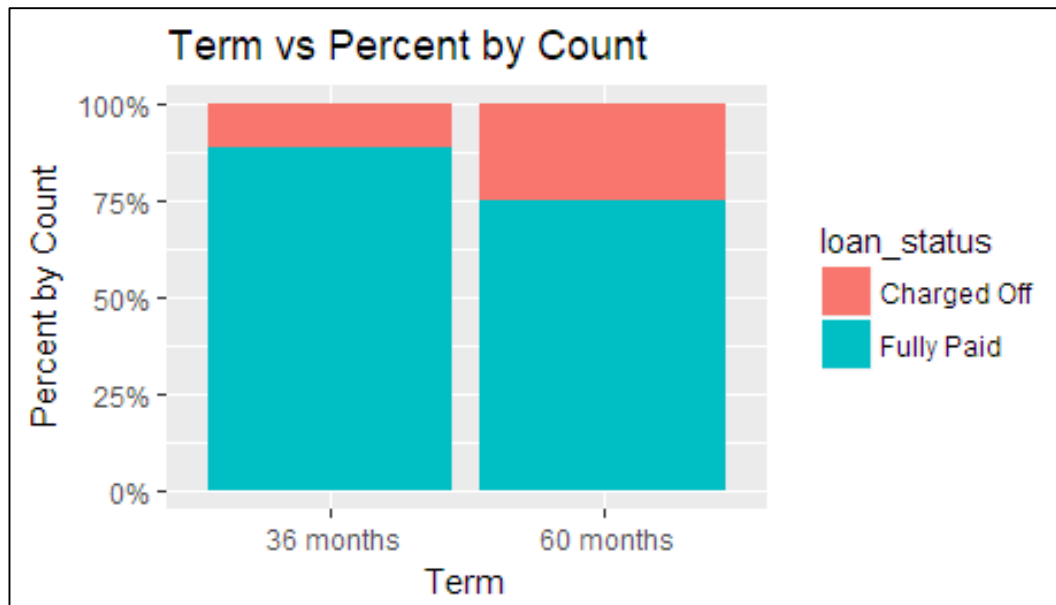
▶ **Purpose** of Small business, renewable energy, educational, debt consolidation have significant impact on defaulters %
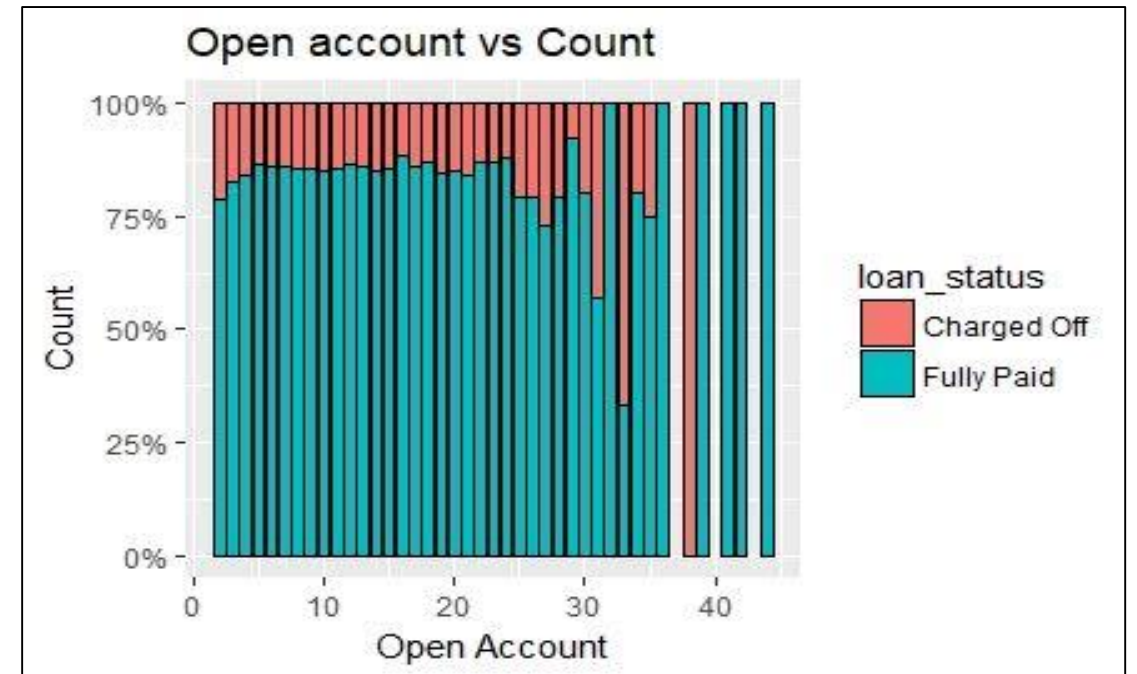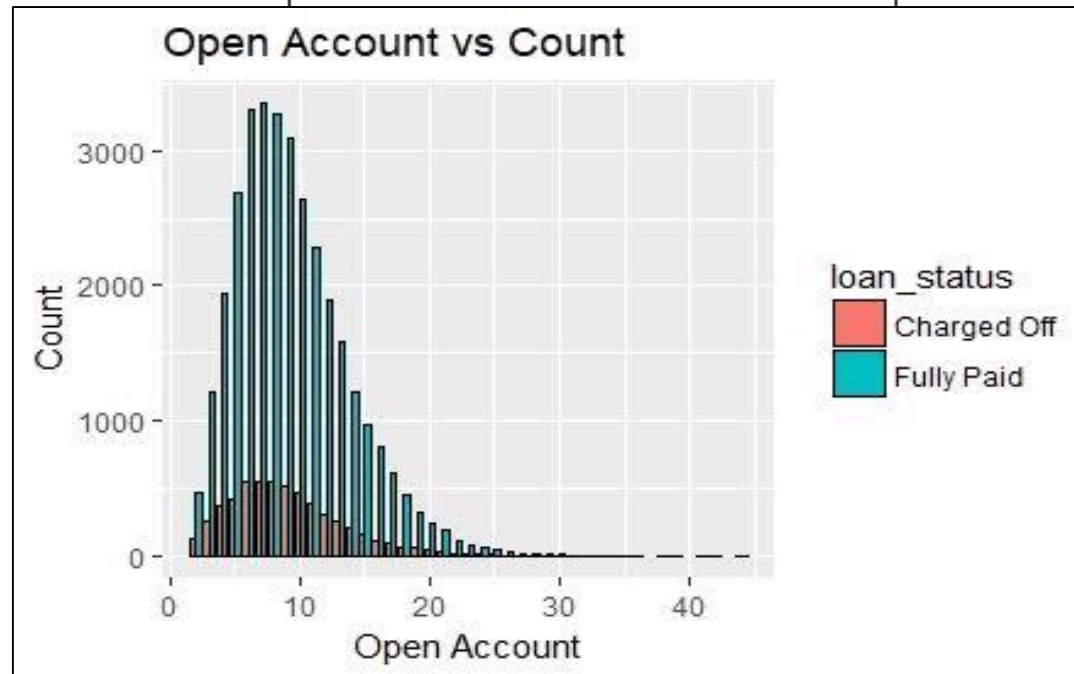
Data Analysis – Variables with impact

▶ Loans with income **Verified status** has slightly more % of defaulted loans. This needs to be further analyzed

▶ Loans with **employment length not provided** have more % of defaulters
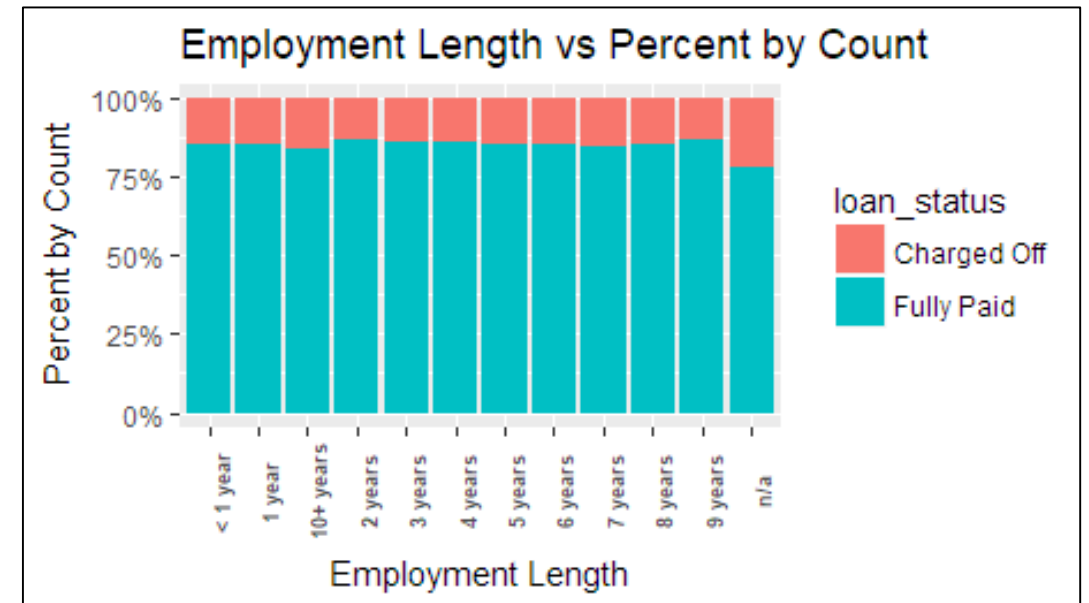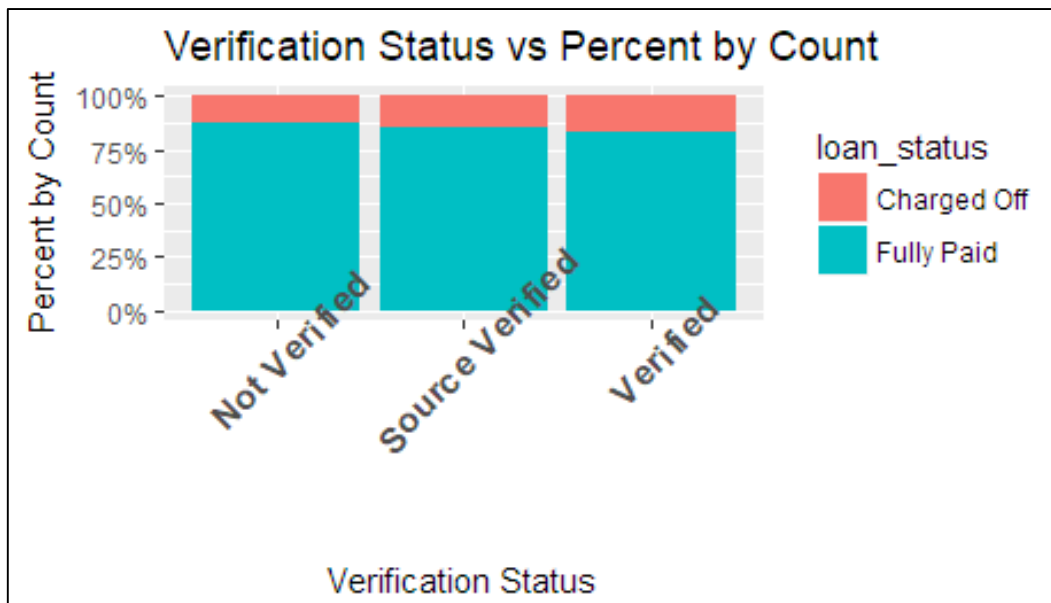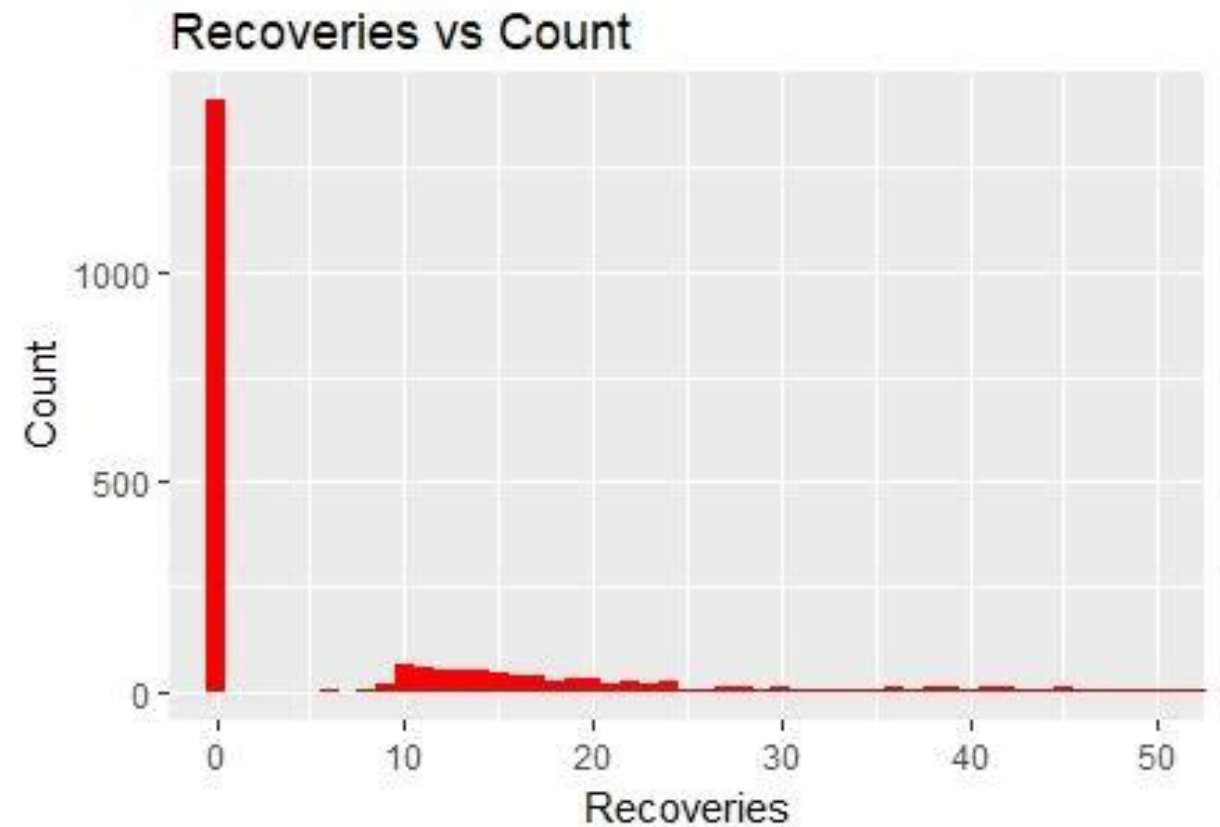
Data Analysis – Variables with impact

▶ Significant count of records lie for open accounts number <25

▶ For **open accounts** values 2 to 4, % of defaulters and count of defaulters is higher as compared to other values of open accounts
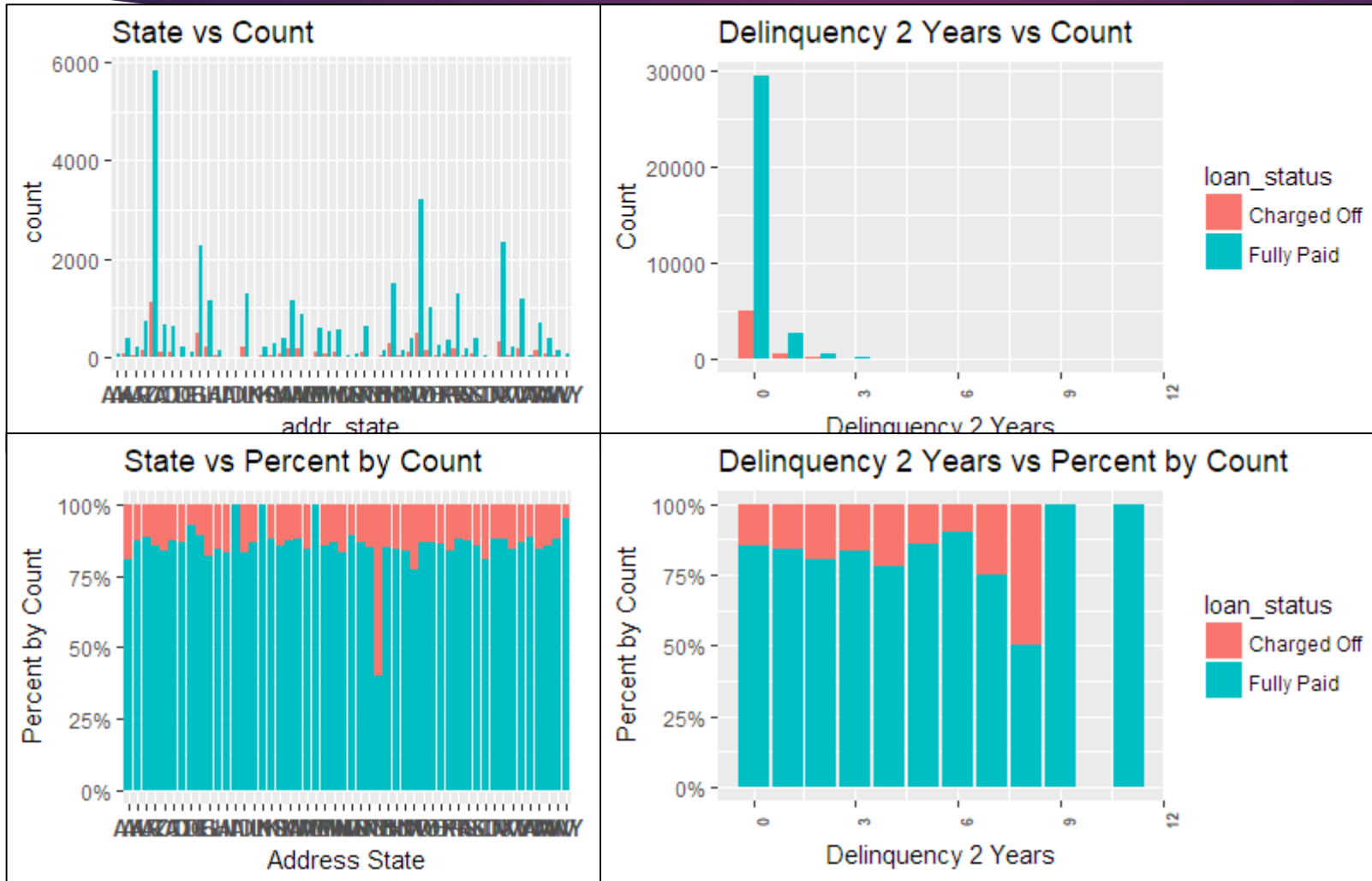
Data Analysis – Variables with impact

▶ Huge number of loans (1750+) could not have any recovery post charged-off.

▶ These individuals / entities should be scrutinized thoroughly in future while providing loan as they could be willful defaulters

** Graph is zoomed only for significant values of recoveries

Recoveries vs Count

Data Analysis – Variables with limited impact

State vs Count

Delinquency 2 Years vs Count

State vs Percent by Count

Delinquency 2 Years vs Percent by Count

- State does not have significant impact but for state NE, though count of defaulters is less, the % of defaulters is significantly high

- Delinquency in 2 years does not have significant impact but still for no delinquency, almost 25% of the loans are defaulters. This needs to be further analyzed

Data Analysis – Variables with no visible impact

▶ Both the loan types (charged off and Fully Paid) have similar distribution of loan amount
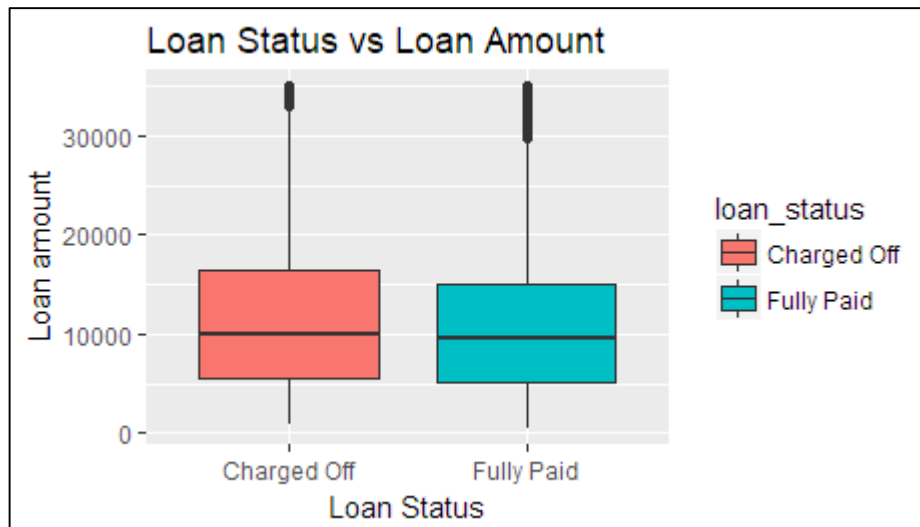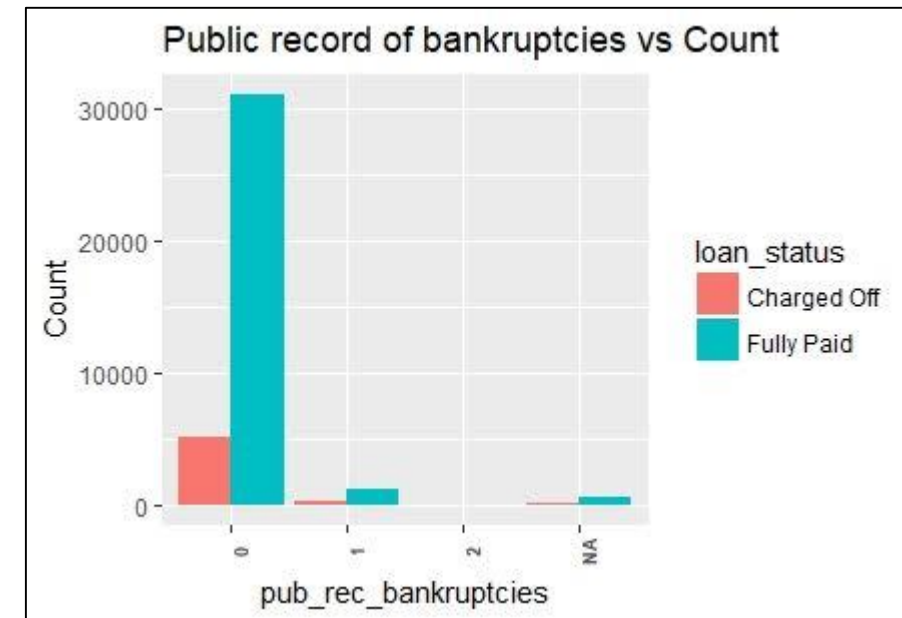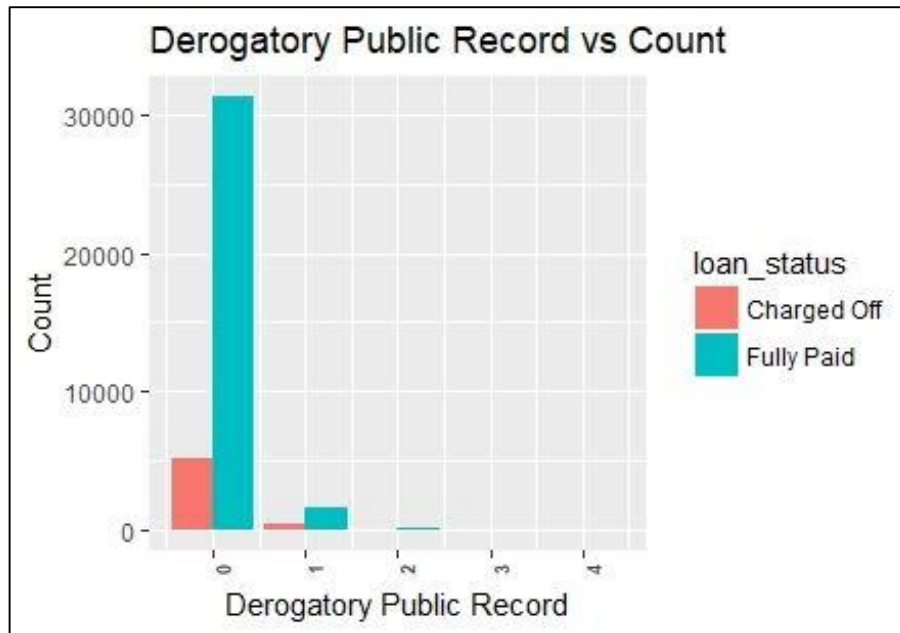
▶ Home ownership type does not impact the defaulter's list

Loan Status vs Loan Amount

Home Ownership vs Count
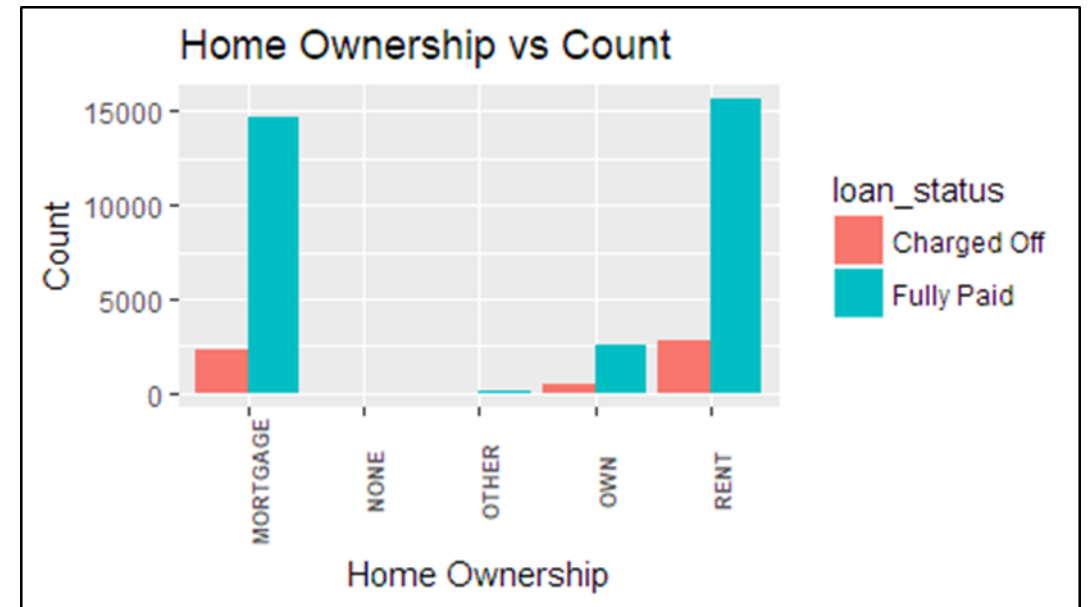
Data Analysis – Variables with no visible impact

▶ Derogatory public records and public record of bankruptcies do not have visible impact as even for nil records, defaulters are found

Data Analysis – Bivariate analysis

- Default_flag column has 2 values:
  - 1 : Defaulter
  - 0 : Not a defaulter
- Moderate positive correlation found between
  - Interest rate and Defaulters
- Inverse moderate correlation found between
  - Total payment, total principal, interest received Vs defaulters : As the total payment or total principal payment or total interest paid increases, chances of defaulting are reduced
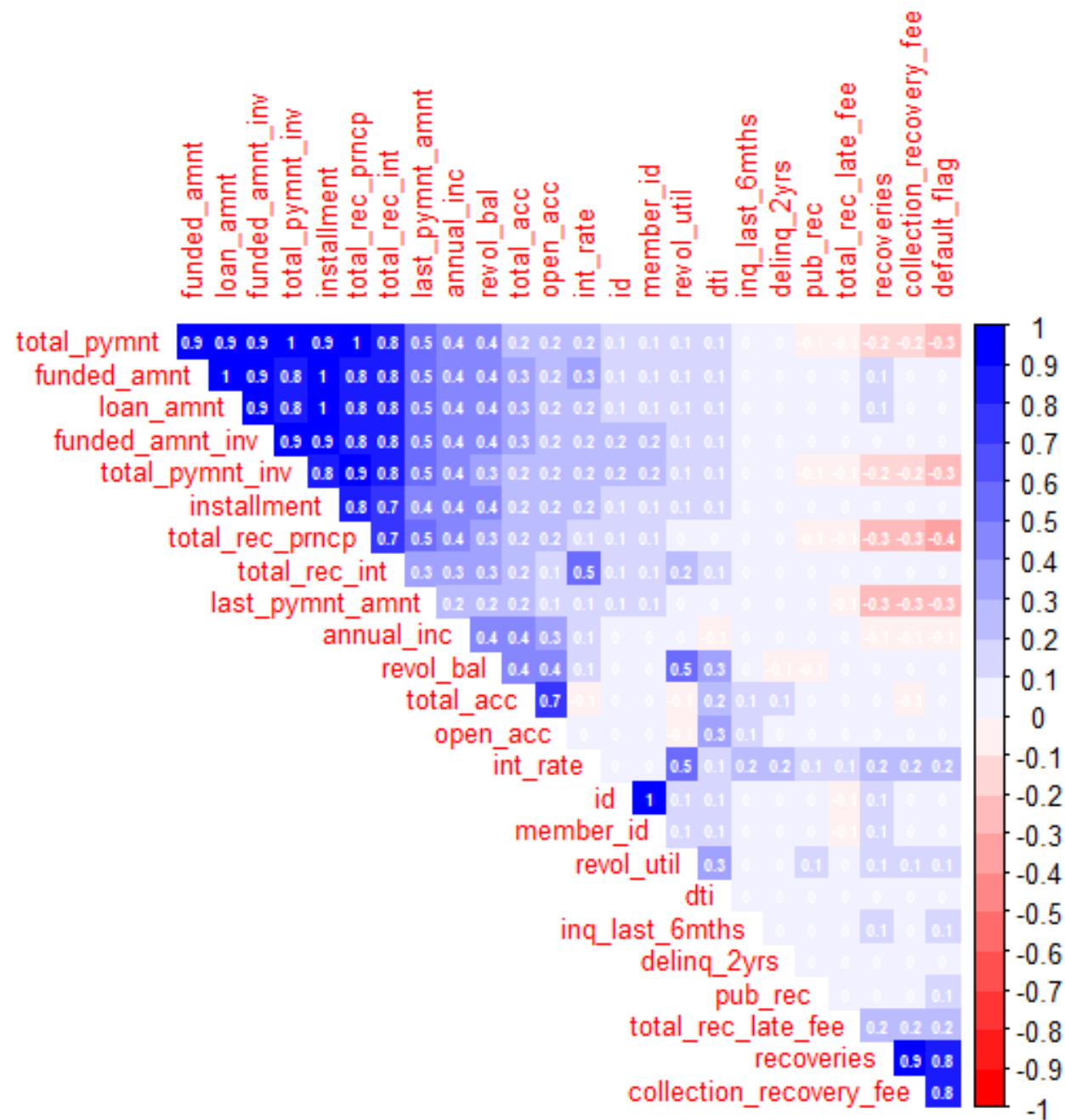
Summary of observations

▶ Following factors can have impact on risk of loans being defaulted

- ▶ LC grade (and hence interest rate) – Lower grades
- ▶ Annual income – Less than 225000
- ▶ Payment term – 60 month
- ▶ Purpose – small business, debt consolidation, educational
- ▶ Employment length – Not mentioned
- ▶ Verification status – Verified
- ▶ Open accounts – 2, 3 or 4
- ▶ Members from whom no recovery could be done post charged off

▶ Following factors need to be watched further for the impact

- ▶ State, Delinquency in last 2 years

▶ Following factors do not have visible impact

- ▶ Loan amount, Home ownership