LENDING CLUB CASE STUDY

BY – SAGARIKA BHUYAN & ROHIT PANDEY

CONTENT

- Problem statement
- Objectives
- Dataset understanding
- Approach
- Segmented Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Result
- Recommendations
- Team

PROBLEM STATEMENT

ending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit ss is the amount of money lost by the lender when the borrower refuses to pay or runs away with e money owed. In other words, borrowers who default cause the largest amount of loss to the nders. In this case, the customers labeled as **'charged-off'** are the **'defaulters'**.

ompany wants to understand the driving factors (or driver variables) behind loan default, i.e. the riables which are strong indicators of default. The company can utilize this knowledge for its rtfolio and risk assessment.

he data given contains information about past loan applicants and whether they ‘defaulted’ or not. he aim is to identify patterns which indicate if a person is likely to default, which may be used for king actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at higher interest rate, etc.

OBJECTIVE

The aim of this case study is to identify the risky or default borrowers who are not paying the money owed to the lenders using EDA.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

We are using two datasets provided one is Loan Data set which contains the complete loan data for all loans issued through the time period 2007 to 2011. And data dictionary which describes the meaning of these variables

The main objective is to find variables in the dataset which influences for loan to be defaulted. Identify the pattern between multiple variables to understand whether a lender would become a defaulter at a later point in time or not.

Which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

DATASET UNDERSTANDING

The dataset given contains information about past loan applicants and whether they ‘defaulted’ or not.

When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

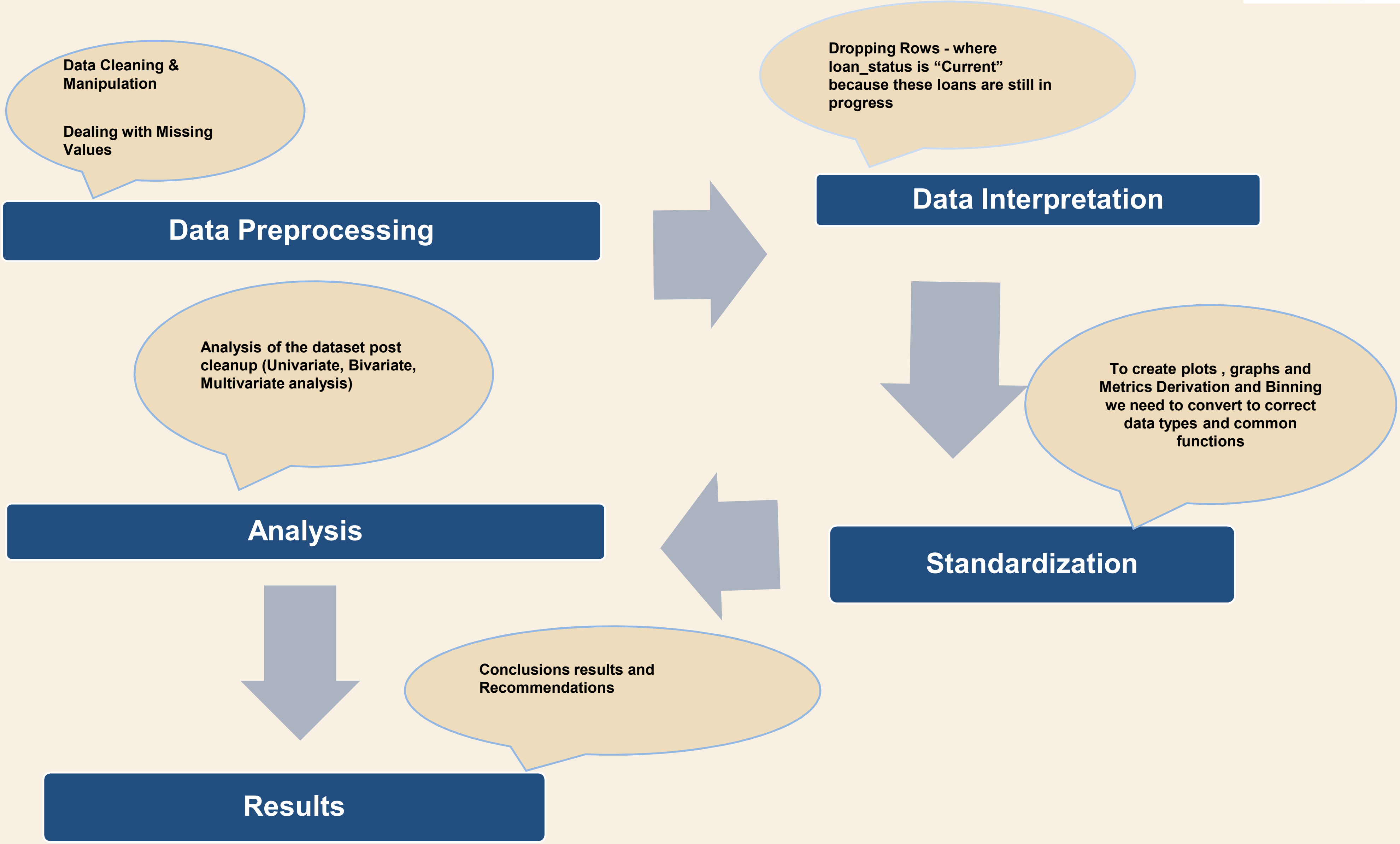
- **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
- **Current**: Applicant is in the process of paying the installments, i.e. the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.
- **Charged-off**: Applicant has not paid the installments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

In this case study, we are using EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

APPROACH

APPROACH



Data Cleaning & Manipulation

Dealing with Missing Values

Data Preprocessing

Dropping Rows - where loan_status is “Current” because these loans are still in progress

Data Interpretation

To create plots , graphs and Metrics Derivation and Binning we need to convert to correct data types and common functions

Standardization

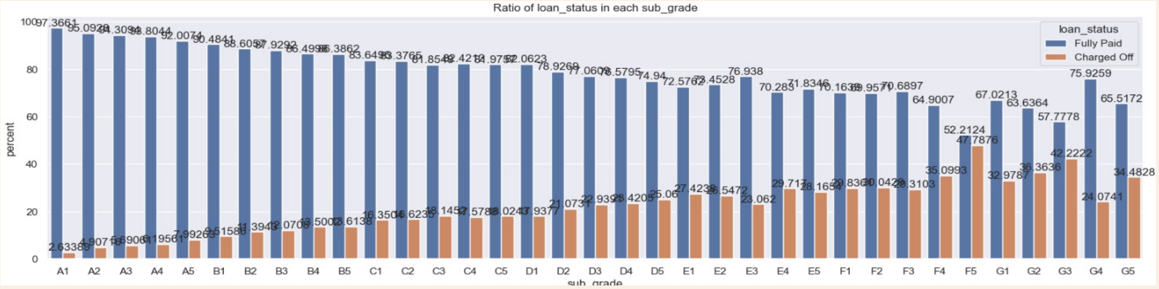
Analysis of the dataset post cleanup (Univariate, Bivariate, Multivariate analysis)

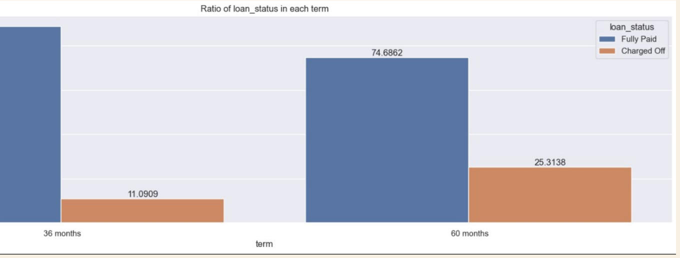
Analysis

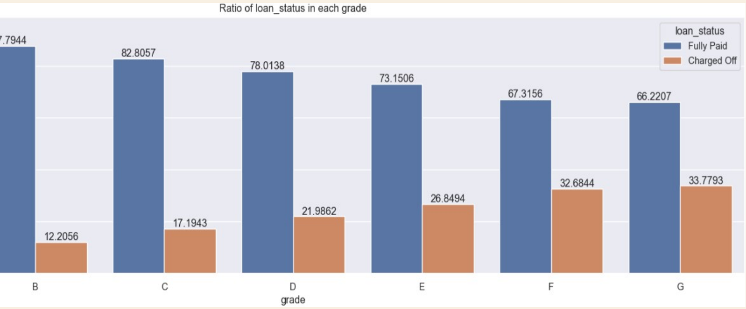
Conclusions results and Recommendations

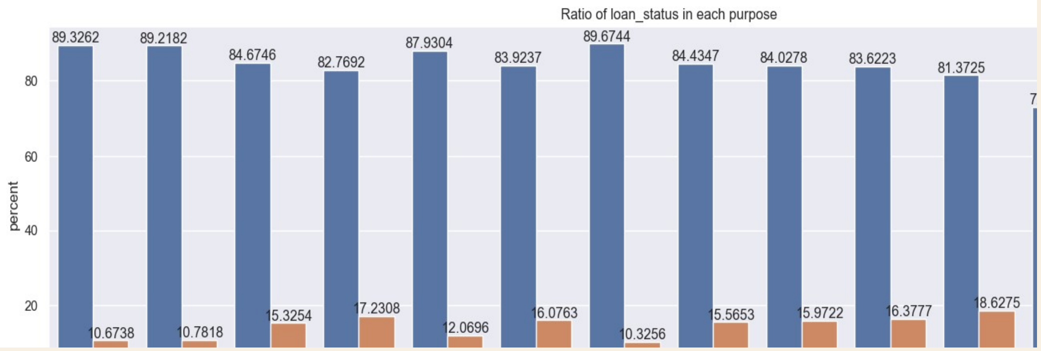
Results

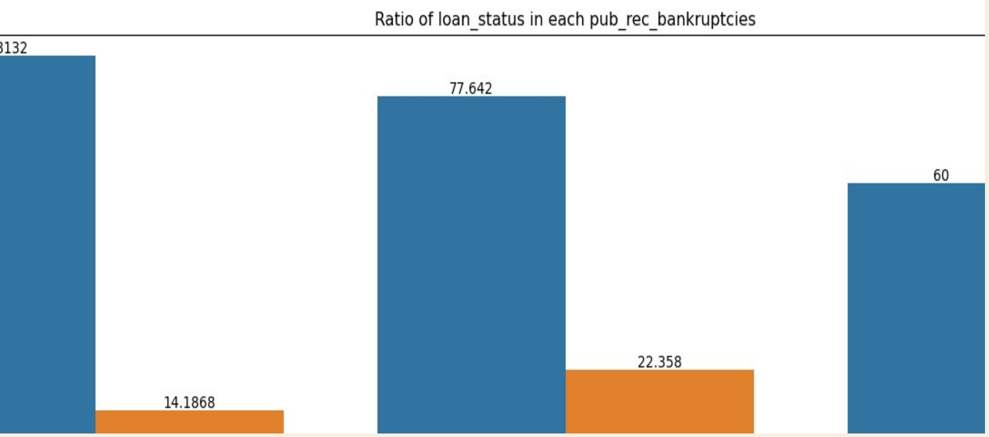
Segmented Univariate Analysis

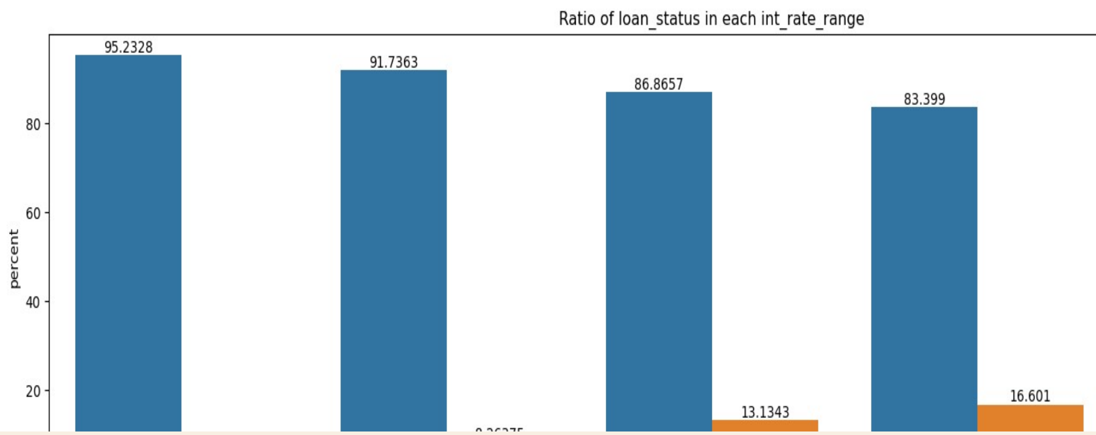




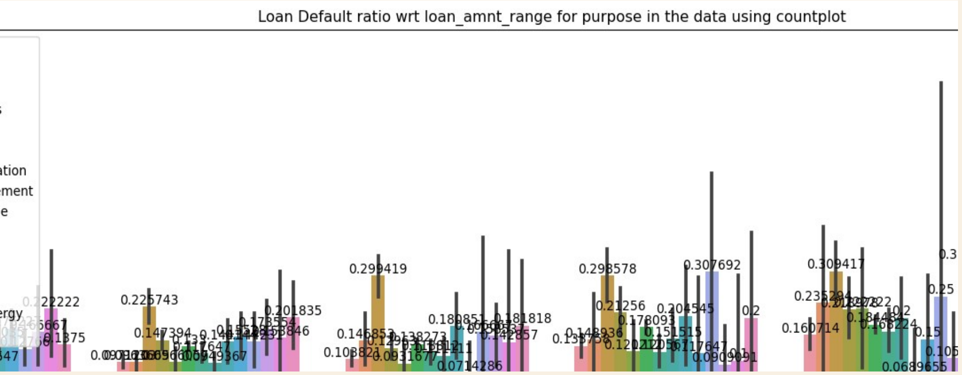


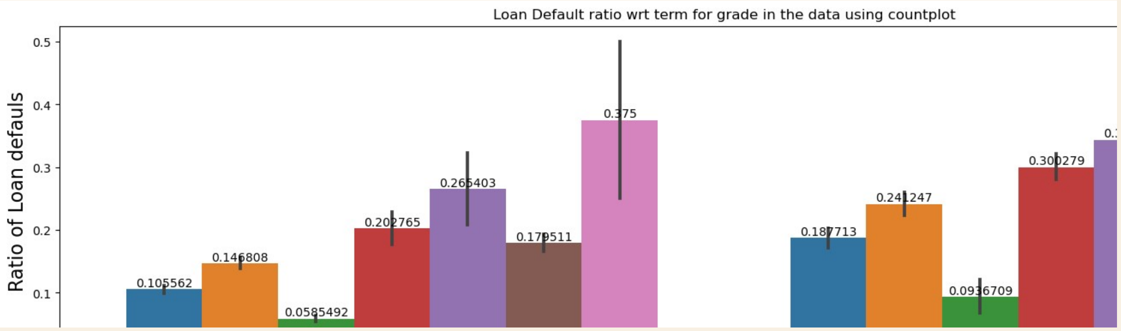


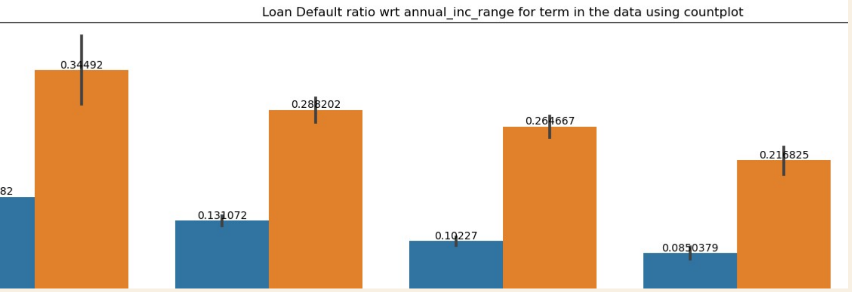


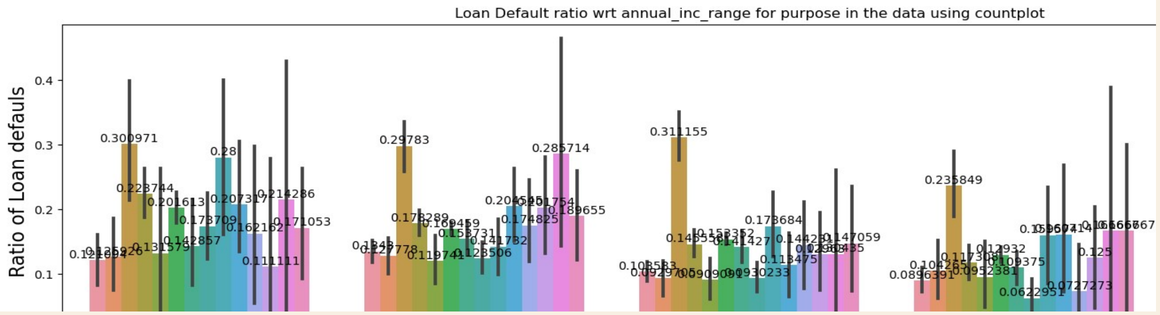


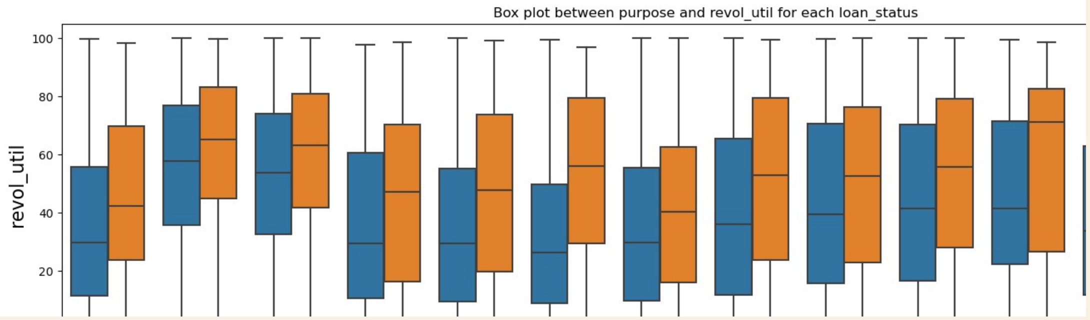
Bivariate Analysis


Loan Default ratio wrt loan_amnt_range for purpose in the data using countplot

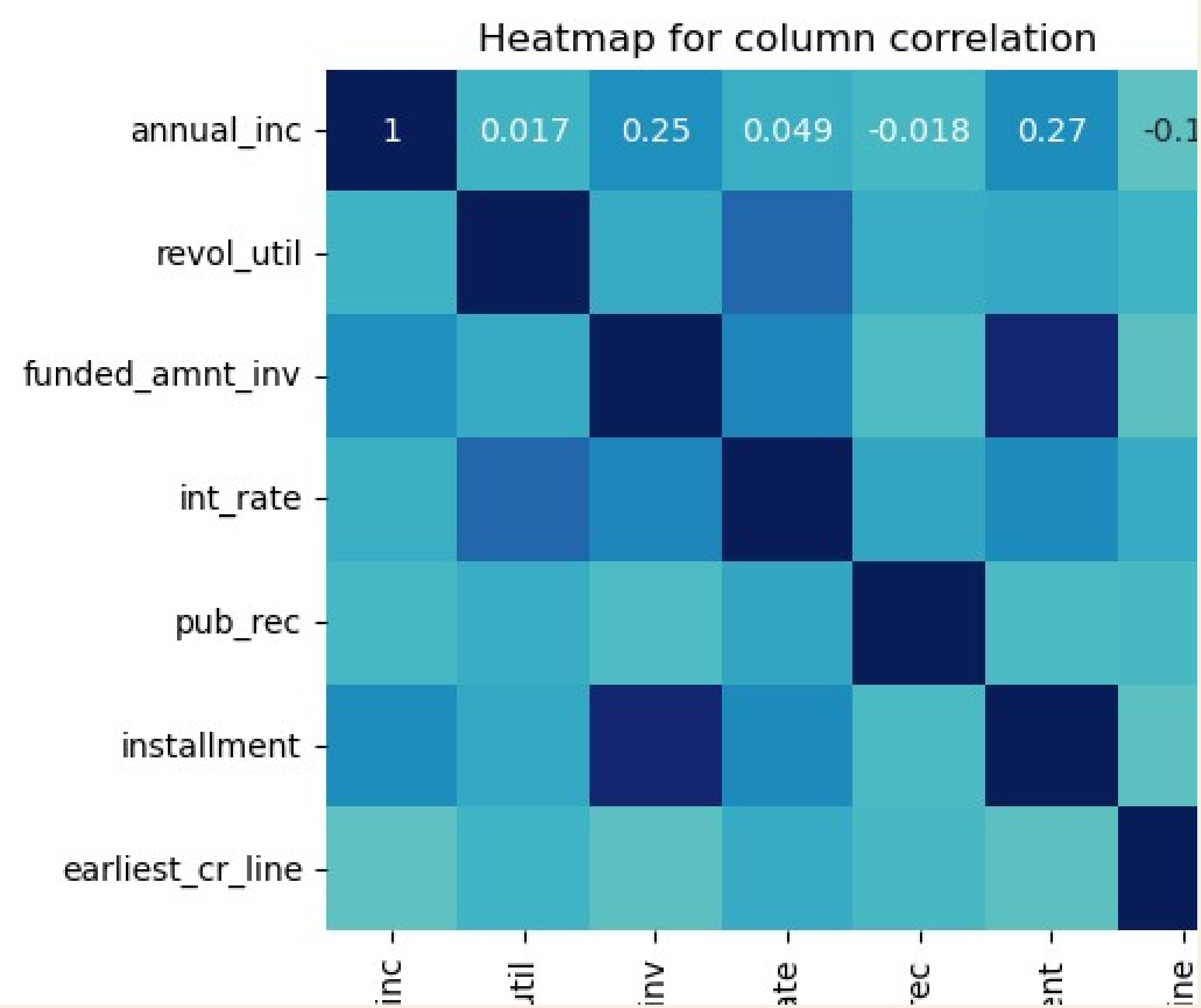

Loan Default ratio wrt term for grade in the data using countplot


Loan Default ratio wrt annual_inc_range for term in the data using countplot


Loan Default ratio wrt annual_inc_range for purpose in the data using countplot


Box plot between purpose and revol_util for each loan_status

Multivariate Analysis



Result

After dataset analysis, for loan approval process we can interpret below variables can be driving factors.

- Term
- Purpose
- Public Record for bankruptcies
- Annual income
- Interest rate
- Grade
- Sub Grade

We observed below % chance for loan to be default:

- 50% with purpose = medical and more than 25K loan amount
- 40% with Educational purpose and with 60 months tenure
- 40% with Small bus in E, G and 50% with F
- 37.5% for grade G with 36 months
- 34% with annual income < 25000 and 60 months tenure

Recommendations

Per analysis, below are recommendations for loan approval

- Implement Stricter Criteria for Grades E, F, and G
- Reject loans for medical purpose and loan amount more than 25K
- Reject loans for educational purpose and with 60 months tenure
- Reject loans for small business purpose with F grade

TEAM

Sagarika Bhuyan & Rohit Pandey