Lending Club Case Study

SUBMISSION

Group Name:

1. Anupkumar Narayanakurup

Business Understanding

We work for a **consumer finance company** which is specialised in lending various types of loans to urban customers. Upon receiving a loan application, the company has to approve/reject the request based on the applicant's profile.

The following are two types of risks associated with the bank's decision,

Loss of business: Not approving the loan even if the applicant is likely to repay.

Financial Loss: Approving the loan when the applicant is likely to default.

Business Objective – Problem Statement

The aim of this case study is **to identify the traits of risky loan applicants** using Exploratory Data Analysis - **EDA**. In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

This knowledge helps the company during **Risk Assessment** process and approving loan for the right applicants thereby **avoiding potential loss of business** and also at the same time **minimising the financial/credit loss***.

**Credit loss is the amount of money lost by the lender when the borrower refuses to pay the money owed.*

Data Understanding

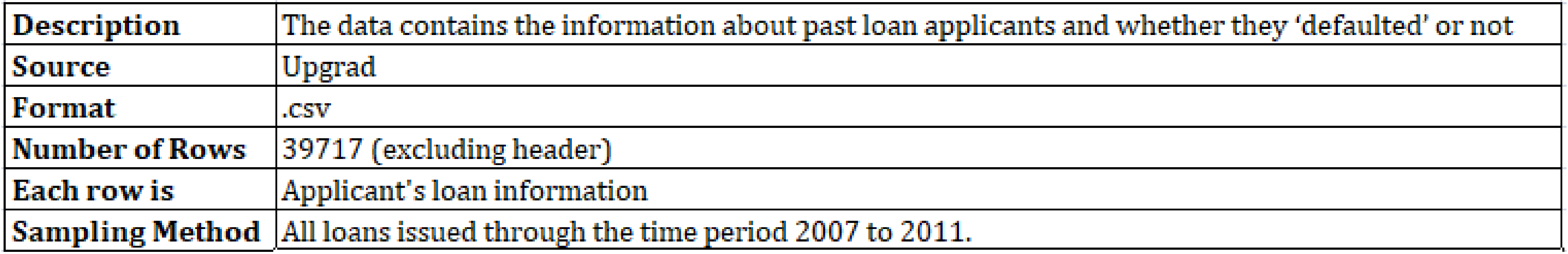
The dataset provided contains the information about past loan applicants (2007 to 2011) and their repayment status. We can term the dataset as **Private** because it has sensitive information and also customer transaction details.

The following are three types of repayment statuses for any approved loan,

- **Fully Paid:** Applicant has fully paid the principal and the interest rate.
- **Current:** Applicant is in the process of paying the instalments
- **Charged – off:** Applicant has defaulted on the loan

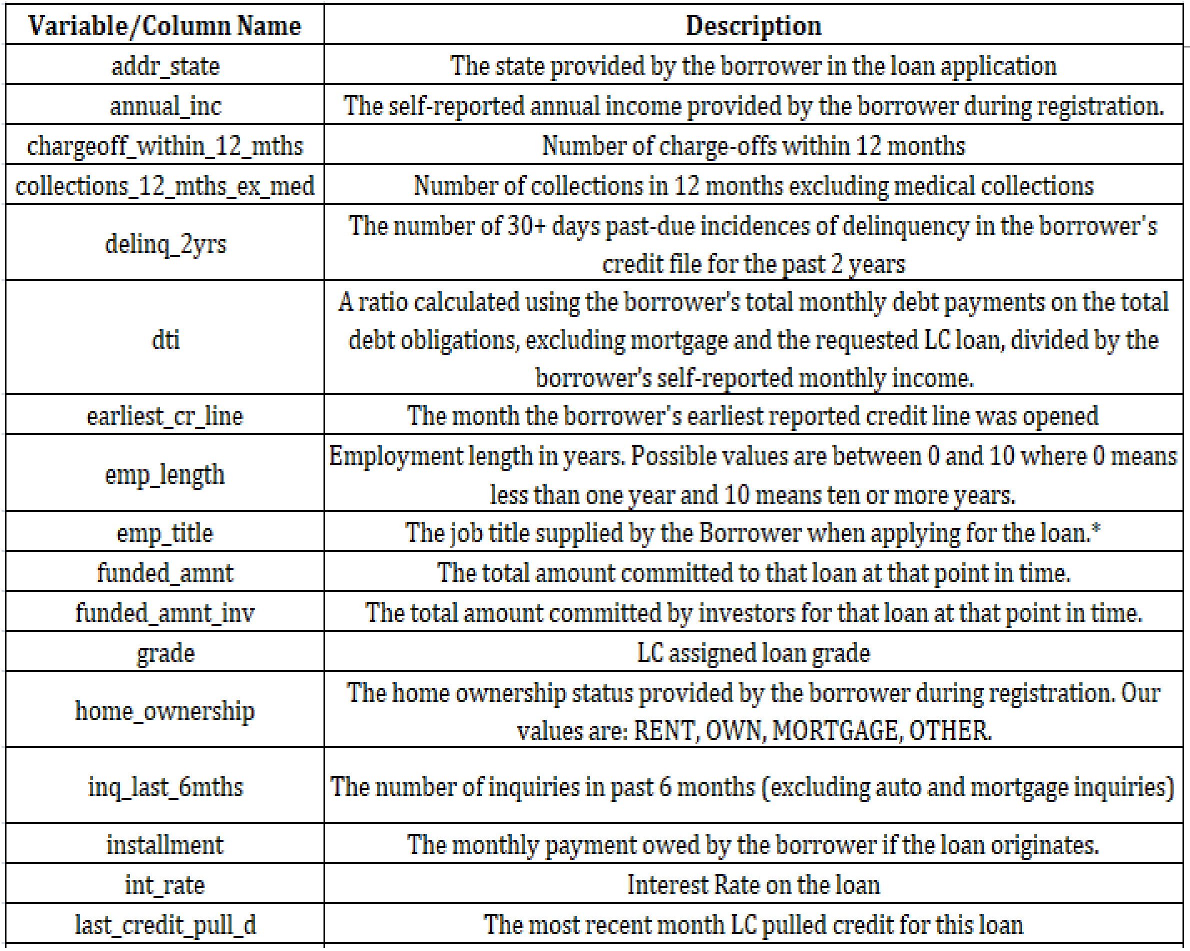
No details are available for rejected applications.

Meta Data - Dataset

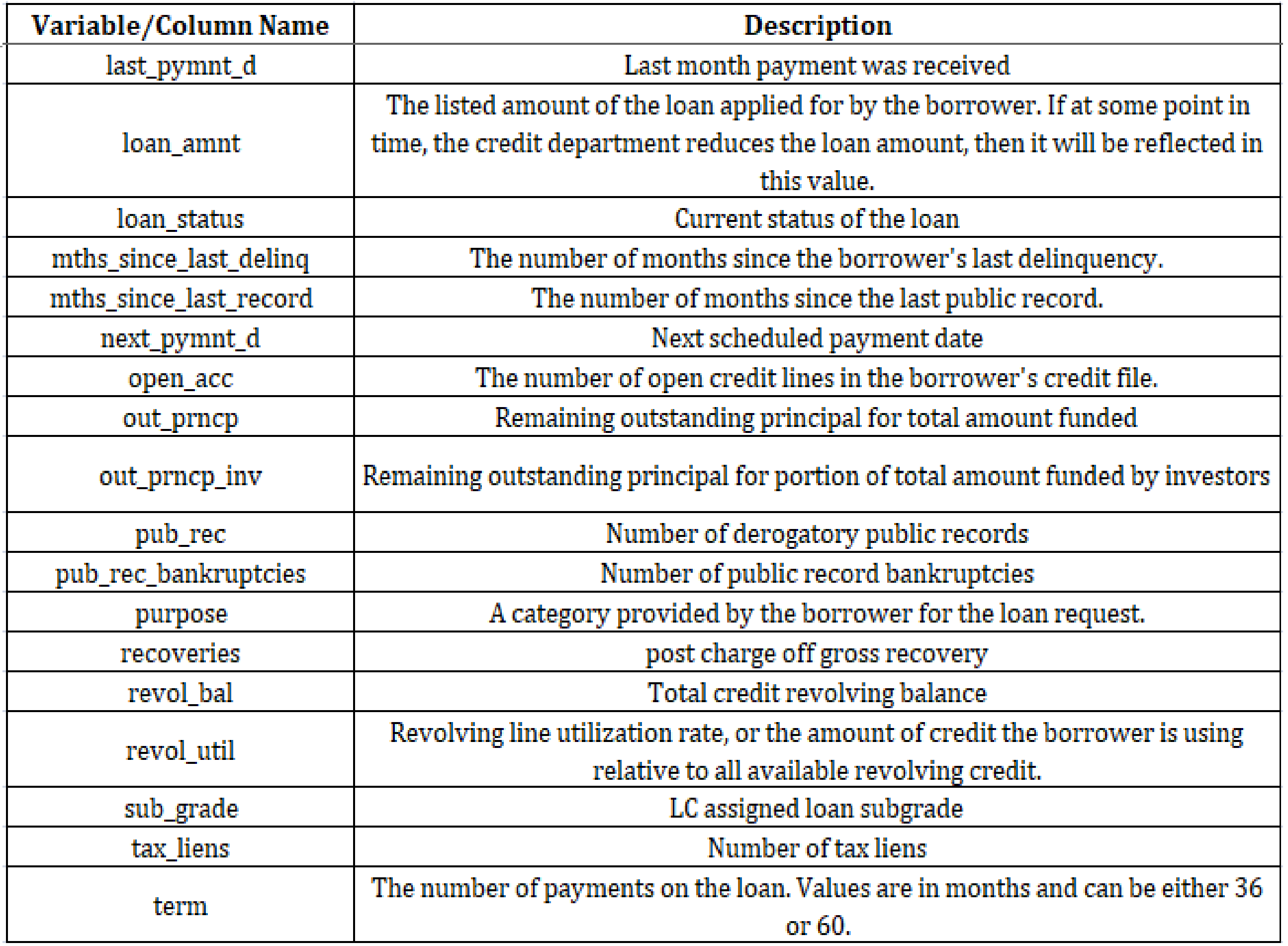


Description	The data contains the information about past loan applicants and whether they 'defaulted' or not
Source	Upgrad
Format	.csv
Number of Rows	39717 (excluding header)
Each row is	Applicant's loan information
Sampling Method	All loans issued through the time period 2007 to 2011.

Meta Data – column/variables

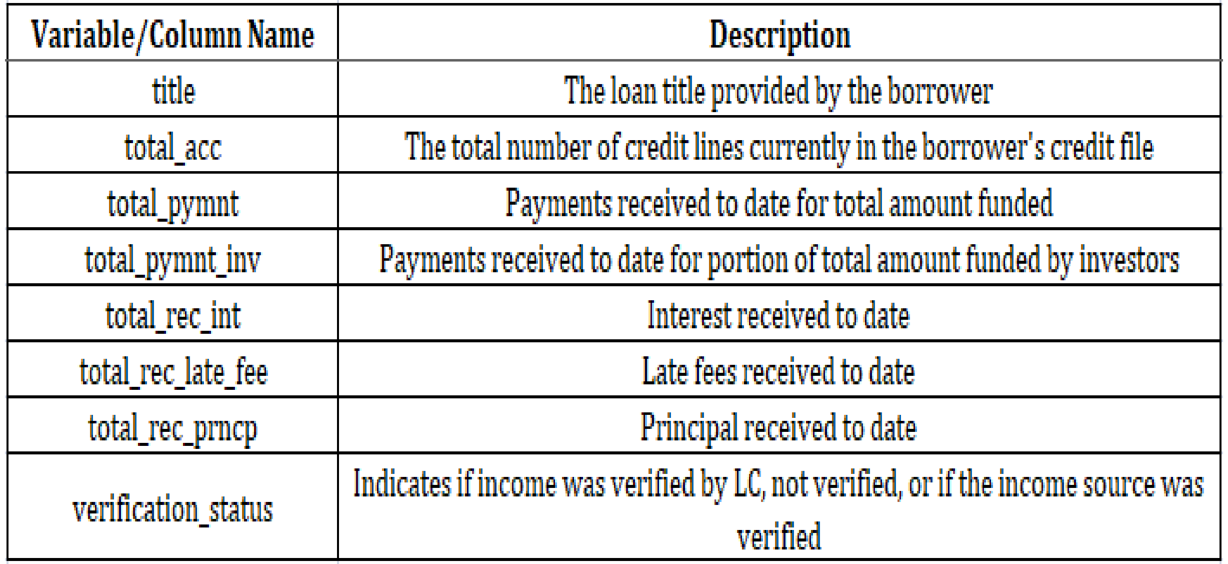


Variable/Column Name	Description
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
last_credit_pull_d	The most recent month LC pulled credit for this loan



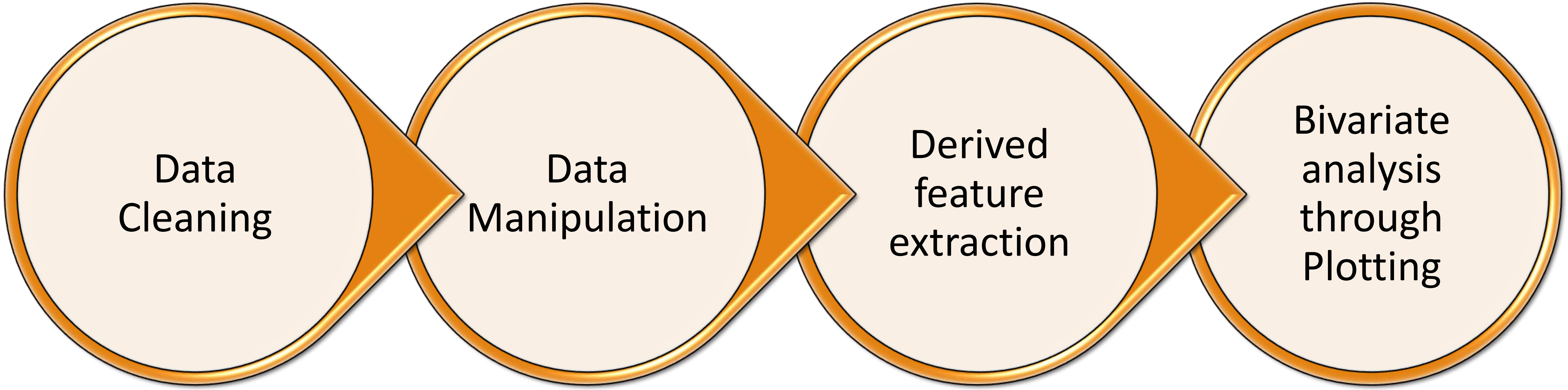
Variable/Column Name	Description
last_pymnt_d	Last month payment was received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_record	The number of months since the last public record.
next_pymnt_d	Next scheduled payment date
open_acc	The number of open credit lines in the borrower's credit file.
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
tax_liens	Number of tax liens
term	The number of payments on the loan. Values are in months and can be either 36 or 60.

Meta Data – column/variables



Variable/Column Name	Description
title	The loan title provided by the borrower
total_acc	The total number of credit lines currently in the borrower's credit file
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified

Overall Process followed:



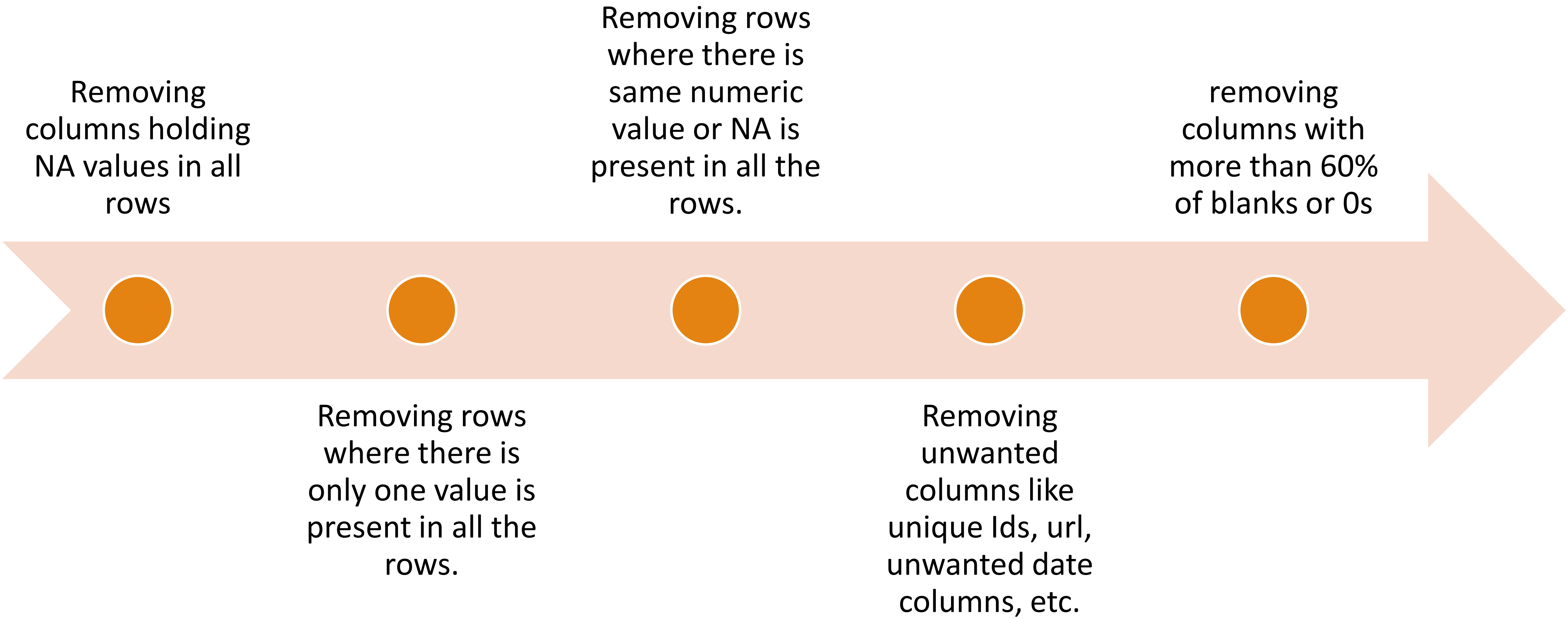
Data Cleaning

Data Manipulation

Derived feature extraction

Bivariate analysis through Plotting

Data Quality Issues – Cleaning



Removing columns holding NA values in all rows

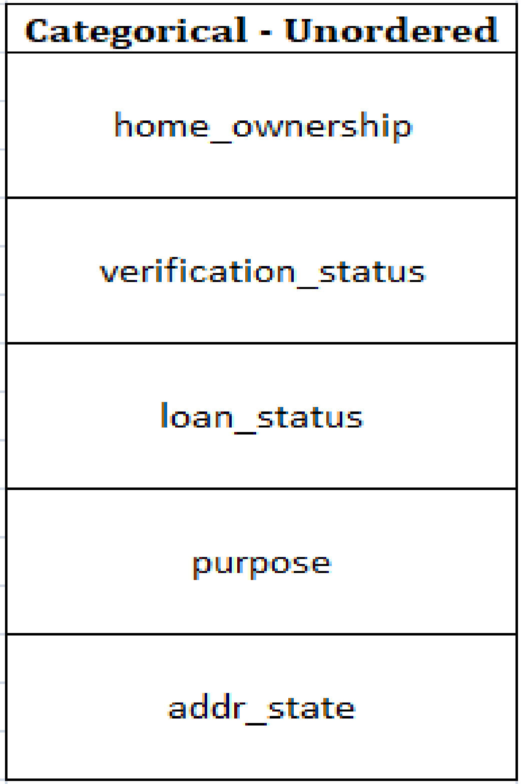
Removing rows where there is only one value is present in all the rows.

Removing rows where there is same numeric value or NA is present in all the rows.

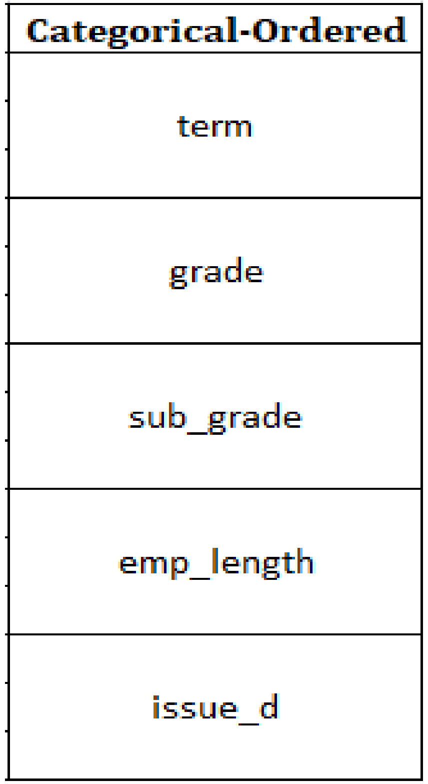
Removing unwanted columns like unique Ids, url, unwanted date columns, etc.

removing columns with more than 60% of blanks or 0s

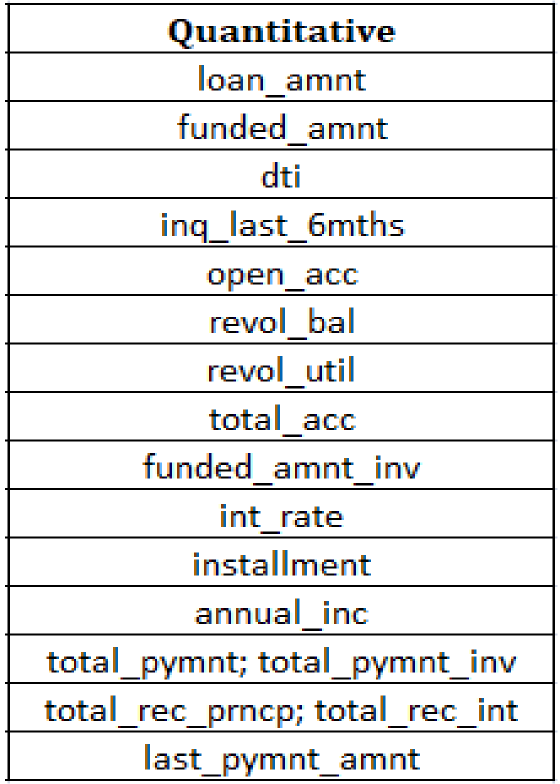
Final Variables applicable for the data analysis



Categorical - Unordered
home_ownership
verification_status
loan_status
purpose
addr_state

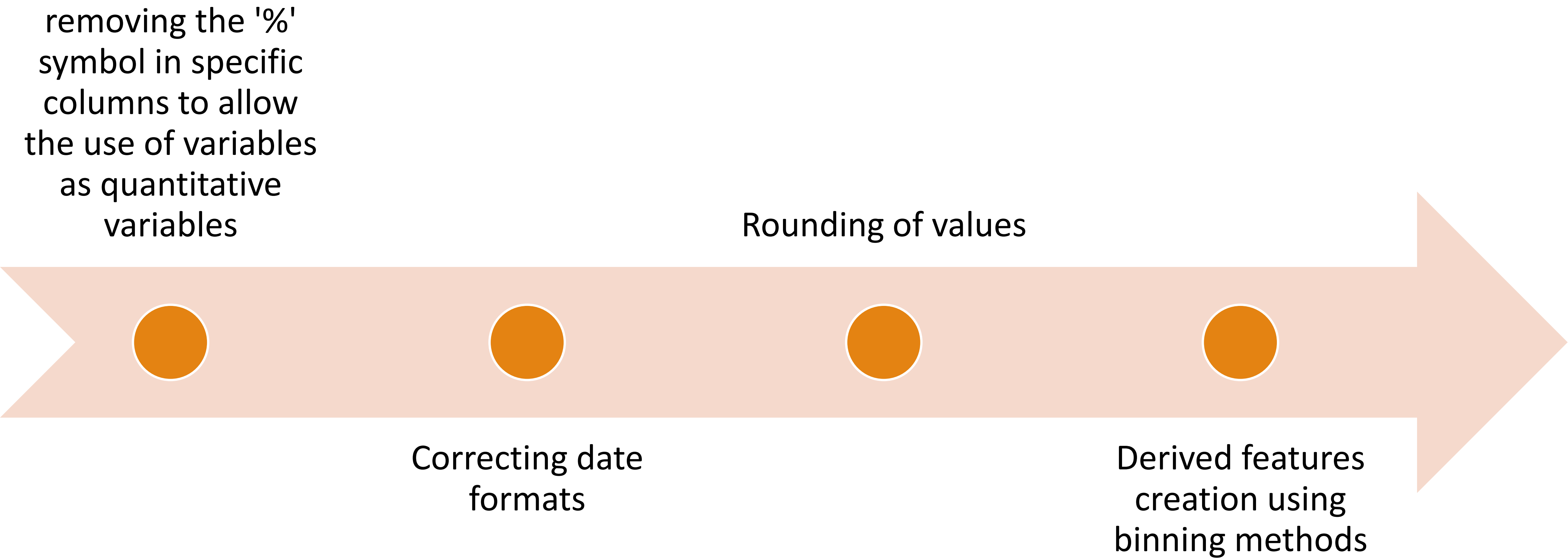


Categorical-Ordered
term
grade
sub_grade
emp_length
issue_d



Quantitative
loan_amnt
funded_amnt
dti
inq_last_6mths
open_acc
revol_bal
revol_util
total_acc
funded_amnt_inv
int_rate
installment
annual_inc
total_pymnt; total_pymnt_inv
total_rec_prncp; total_rec_int
last_pymnt_amnt

Data Quality Issues – Manipulation



removing the '%' symbol in specific columns to allow the use of variables as quantitative variables

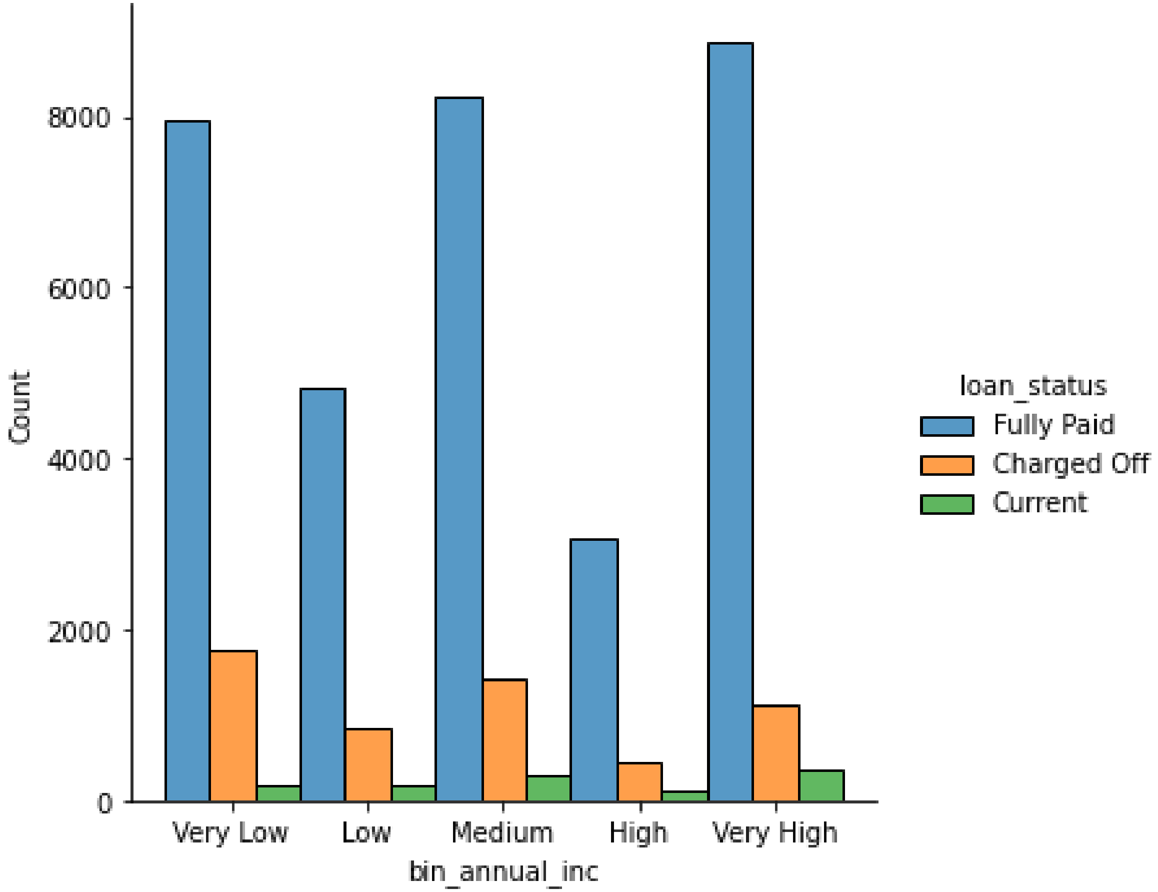
Correcting date formats

Rounding of values

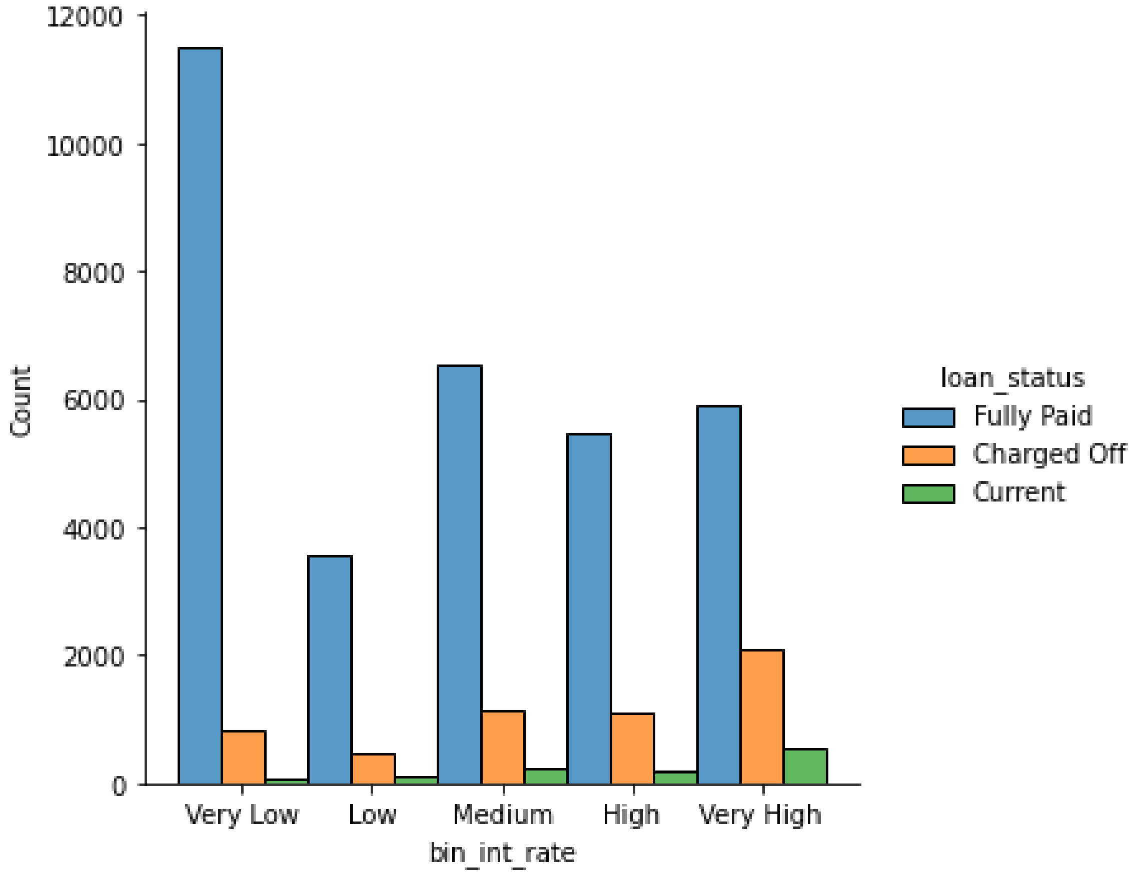
Derived features creation using binning methods

Bivariate Analysis method is preferred for further analysis, as the intention of the business case is to find the dependencies between the different attributes and its dependencies on a loan application becoming default or charged-off.

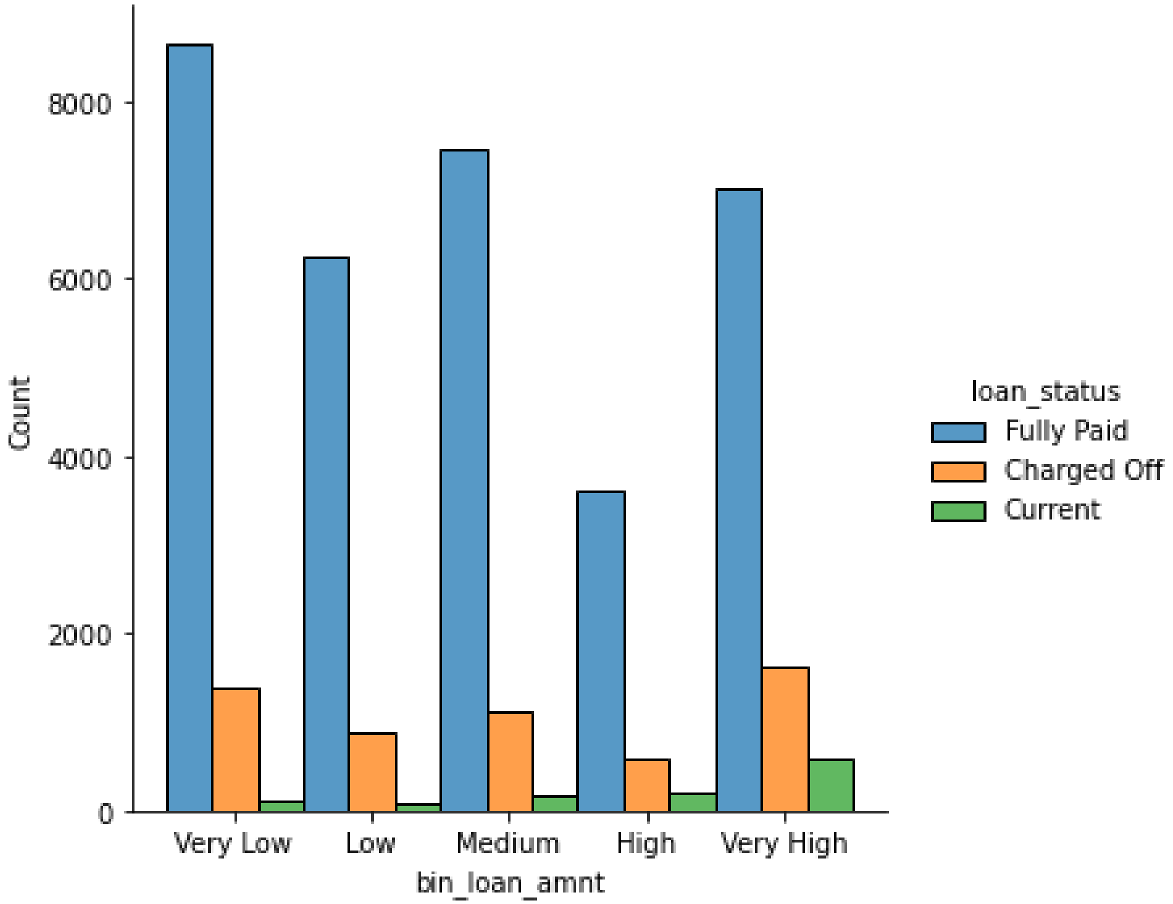
Bivariate Analysis : Annual Income vs. Loan Payment Status



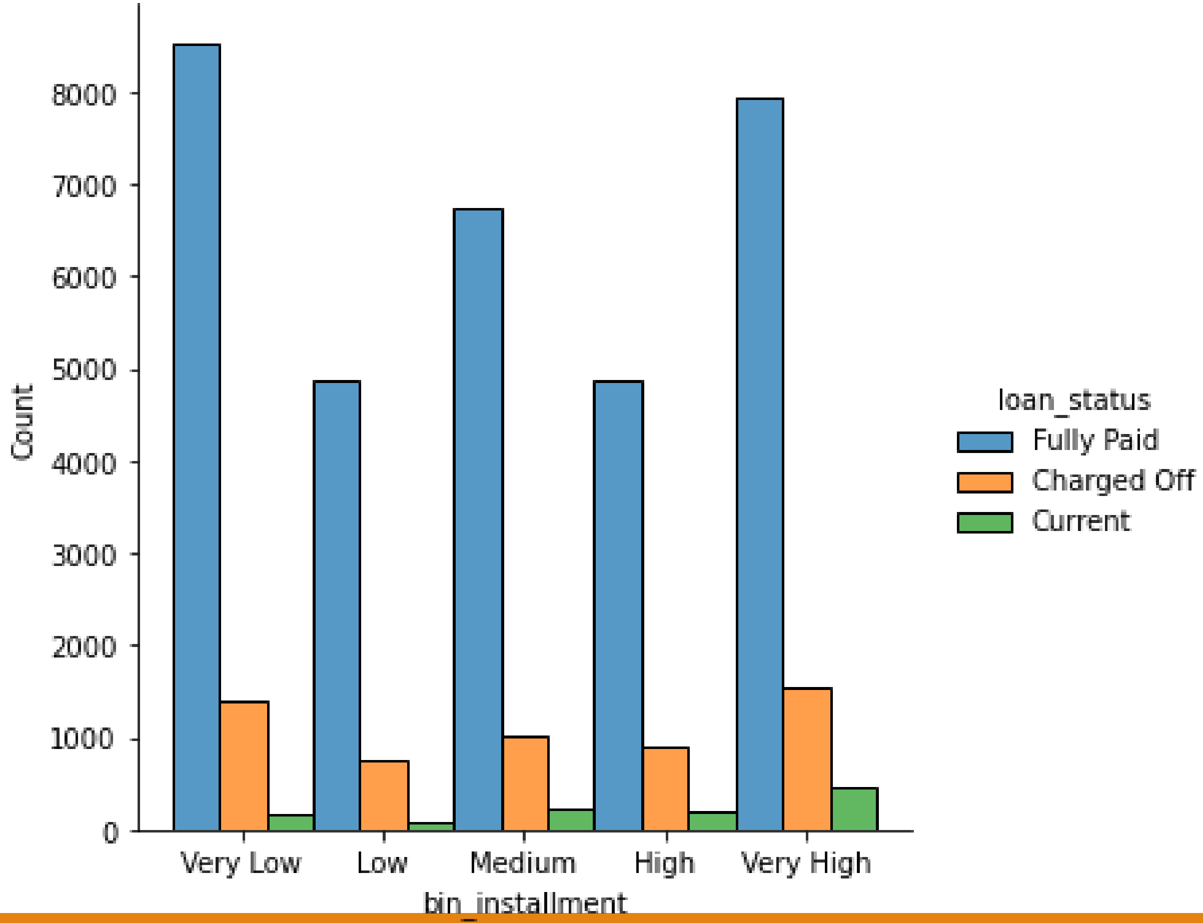
Bivariate Analysis : Loan Status vs. Interest Rate



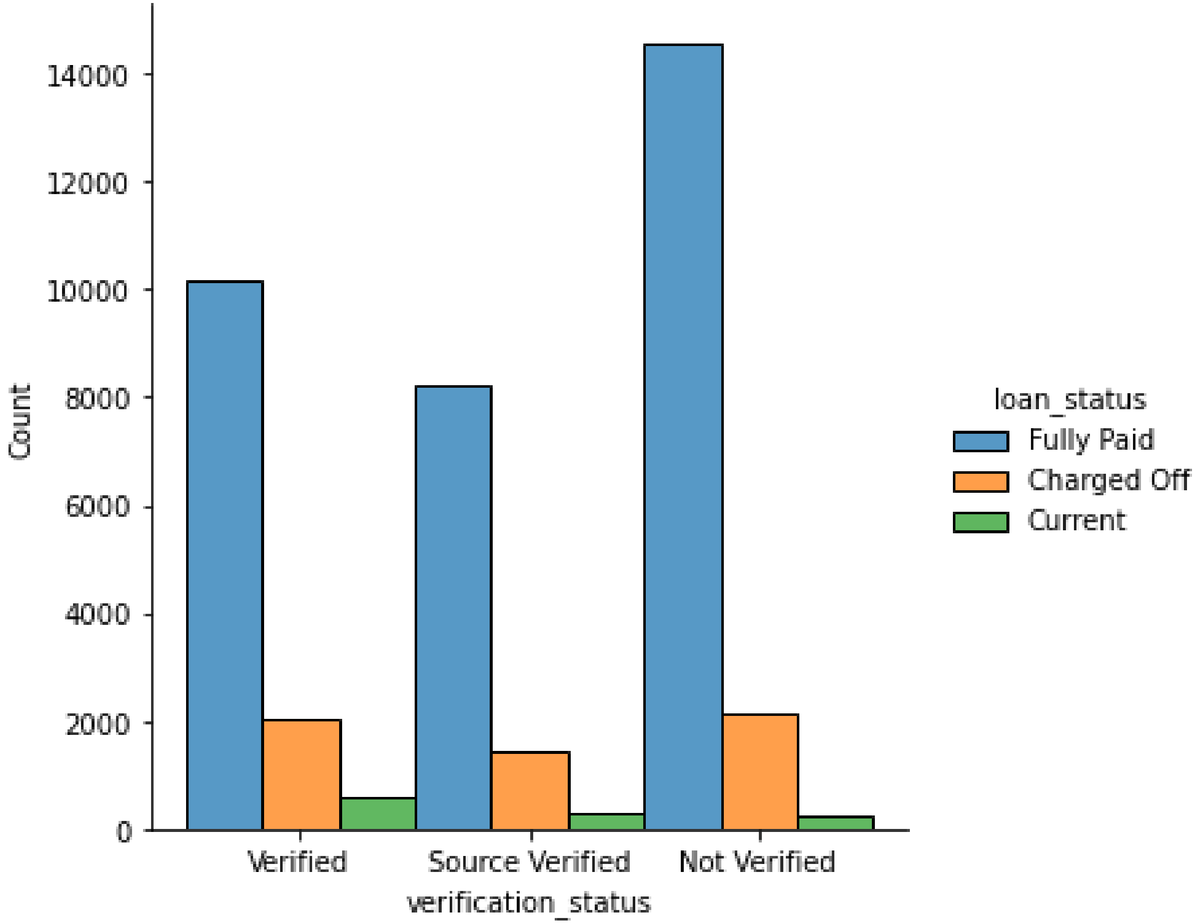
Bivariate Analysis : Loan Status vs. Loan Amount



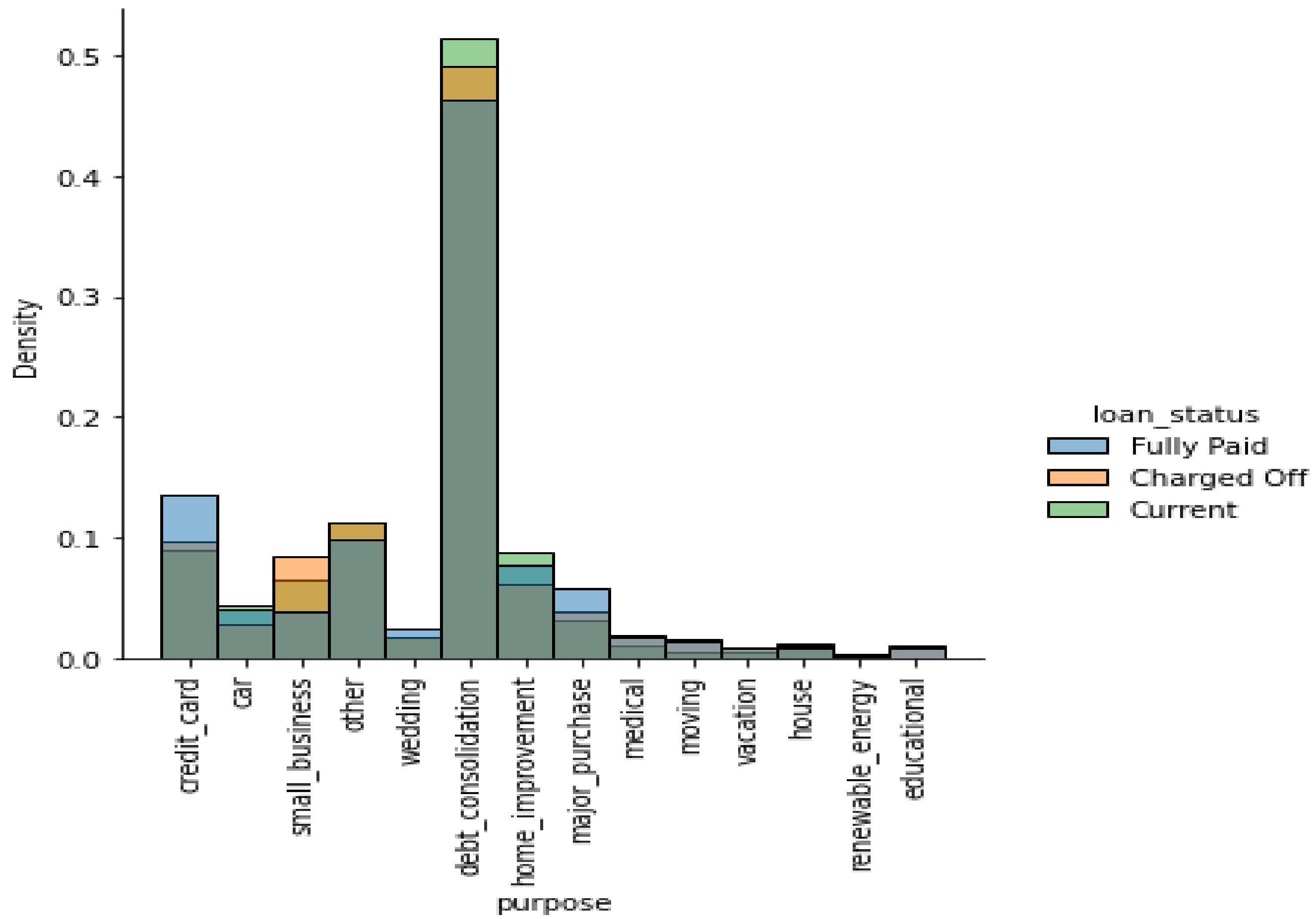
Bivariate Analysis : Loan Status vs. Instalment Amount



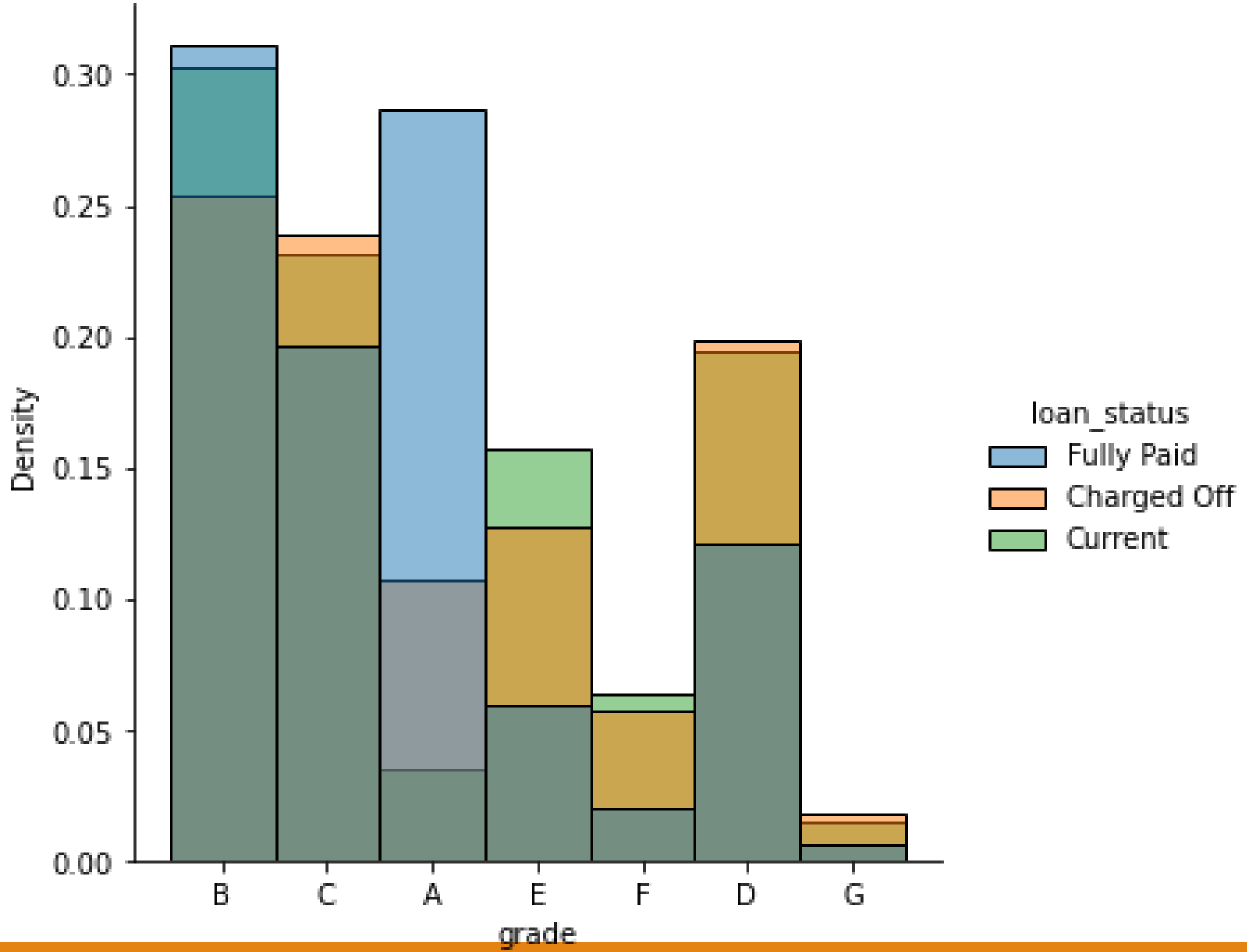
Bivariate Analysis : Loan Status vs. Verification Status



Bivariate Analysis : Loan Status vs. Purpose



Bivariate Analysis : Loan Status vs. Grade



Inferences

Based on the analysis, we can infer that the following variables are significant to determine the likelihood of charged-off/default.

1. Customers with **annual income** < 40000 are most likely to default

2. Loans with very high (>15%) of **interest rate** are most likely to be charged-off

3. Loan Amounts of > 15000 are most likely to be defaulted

4. Loans taken with the **purpose** of Small Business loans are most likely to be defaulted

5. Grades in the order of G, F, E and D have high chances of default