Lending Club Case Study

By:

Mahesh Kavatage
Prathvi Bhatti

Problem Statement

- A **consumer finance company** which specialises 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:

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

Types of Decisions

1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
    1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
    2. **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
    3. **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e., he/she has **defaulted** on the loan
2. **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)

Business Objective

- 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 assessment.
- 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.

Data Understanding and Analysis

- Understand the Data Dictionary
- Analyze the Loan Data set

- EDA process
  - Data Cleaning
  - Univariate analysis
  - Bivariate analysis

Data Cleanup

- Raw data set has 39717 Rows with 111 Columns
- Drop Columns with all NaN (Found 54 such columns)
- Drop Columns with Single/Constant values (Found 9 columns)
- Find Missing Values (df.isNull.sum()) (Found 3 columns)
  - 'next_pymnt_d','mths_since_last_record','mths_since_last_delinq’
- Drop irrelevant columns (Found 11 columns)
  - 'id','member_id','funded_amnt','url','title','desc','emp_title','zip_code','addr_state','last_credit_pull_d','revol_bal','delinq_2yrs’
- Ignore the loan status of type 'Current' as this cannot be used to derive the 'default' case.
- Remove the columns that are substituted by another column values

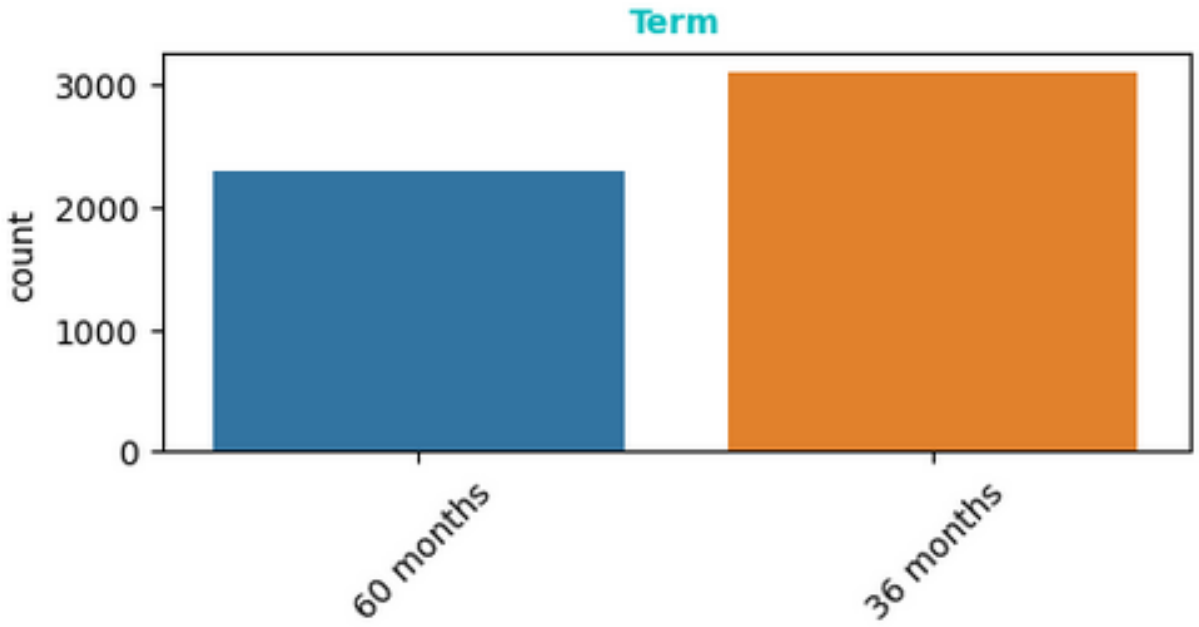
Data Cleanup Continued….

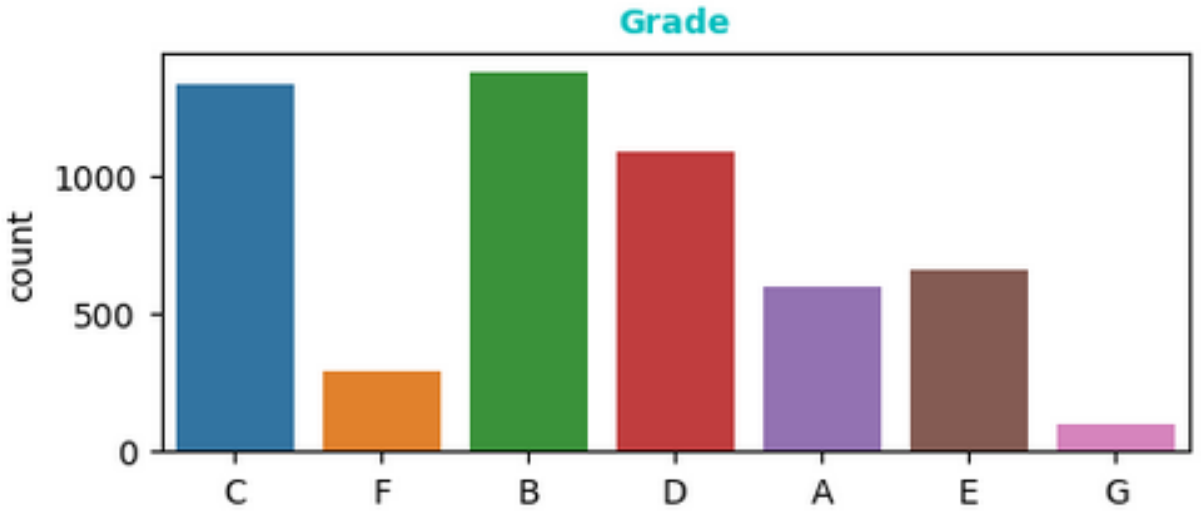
- Format the data
  - Emp_length
- Split the issue_d column into two separate variables
  - issue_month
  - issue_year
- Analyze numeric fields for the outliers
  - Found a field “annual_inc” that had invariably high values above 0.95 percentile and have been removed

Analysis

- Loan process approval considers the following
  - Fully paid
  - Current
  - Charged-off
- Customers labelled with ‘Charged-off’ are the deafulters.
- Since the driving factor is to understand the defaulters, the applicants labelled “Charged-off” are of interest.
- Hence all the analysis has been done considering the loan_status as “charged_off”

Univariate Analysis

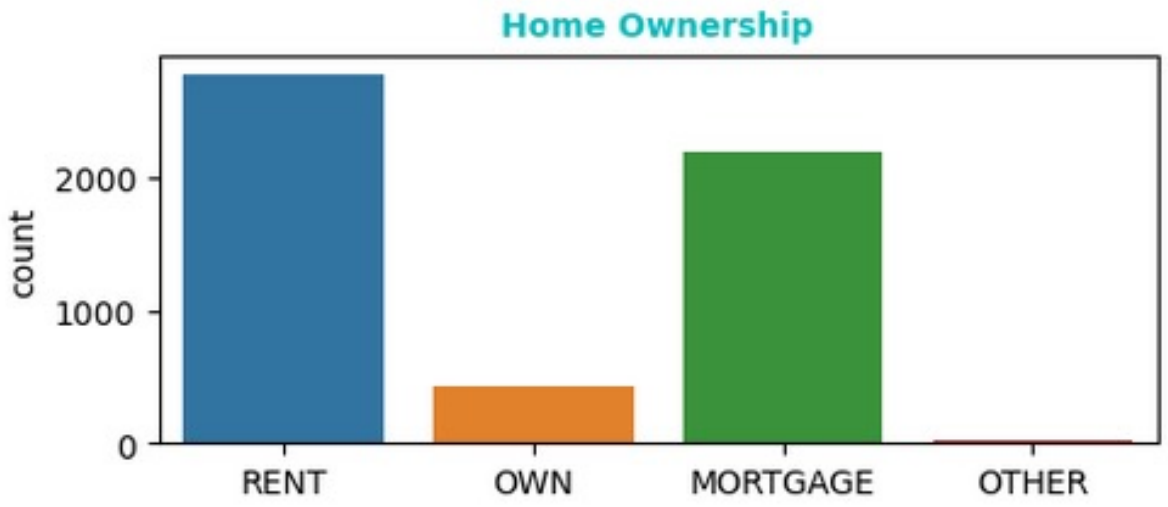




Chances of defaulting when:

- Applicants with 36 months tenure
- Applicants with grade ‘B’

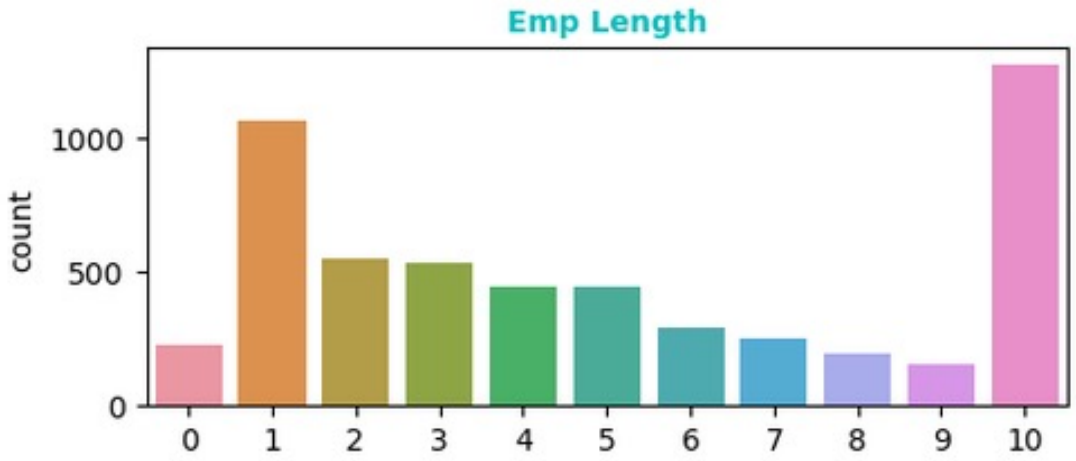
Univariate Analysis

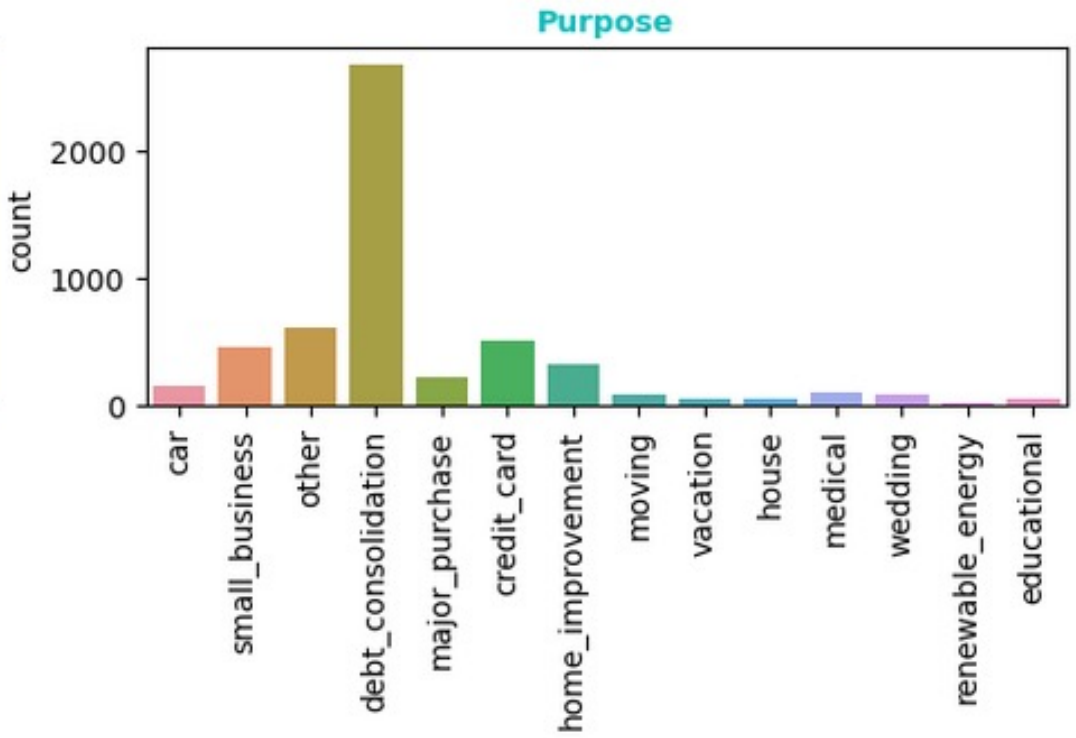


Chances of defaulting when:

- Applicants with Rented home
- Applicants with verification status set to ‘Not Verified’

Univariate Analysis

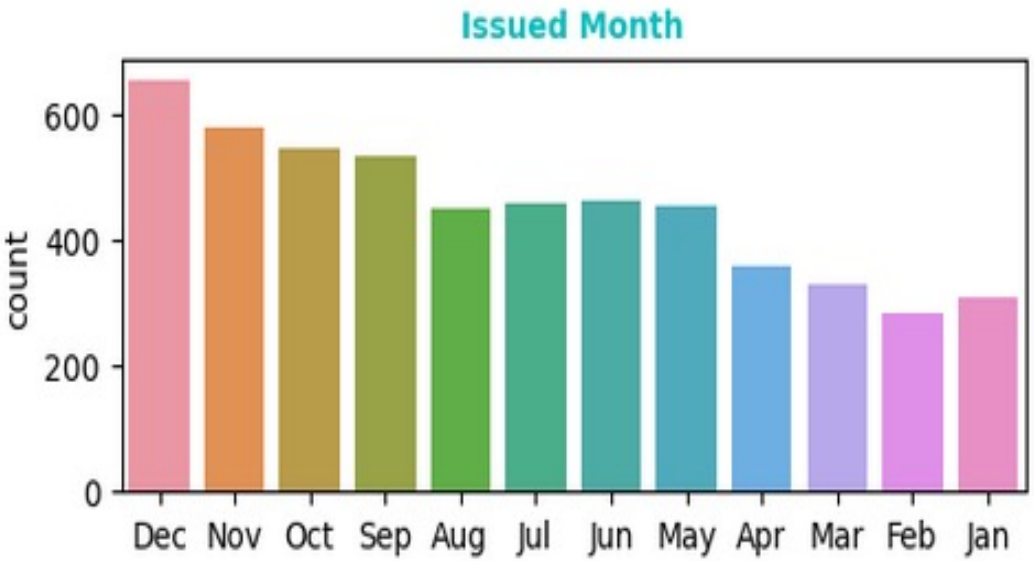


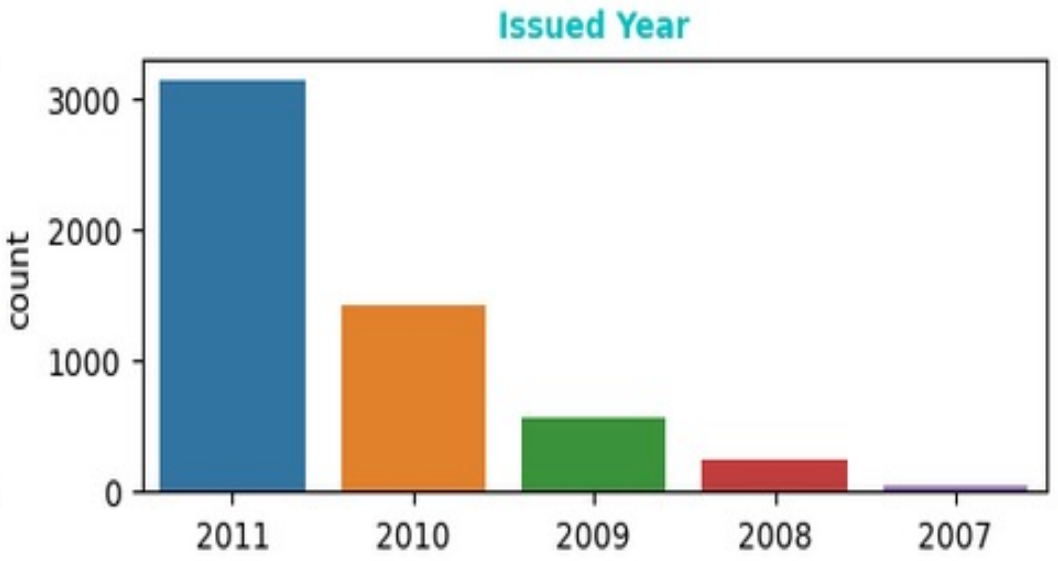


Chances of defaulting when:

- Applicants with employment length or experience of 10+ years
- Applicants with loan purpose of Debt Consolidation

Univariate Analysis

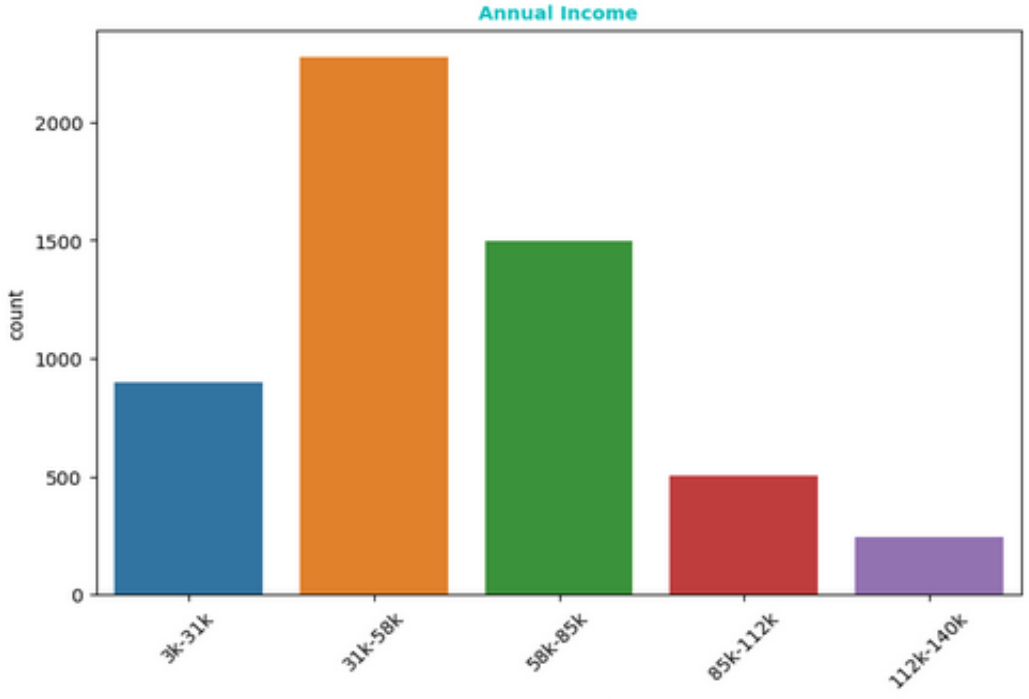


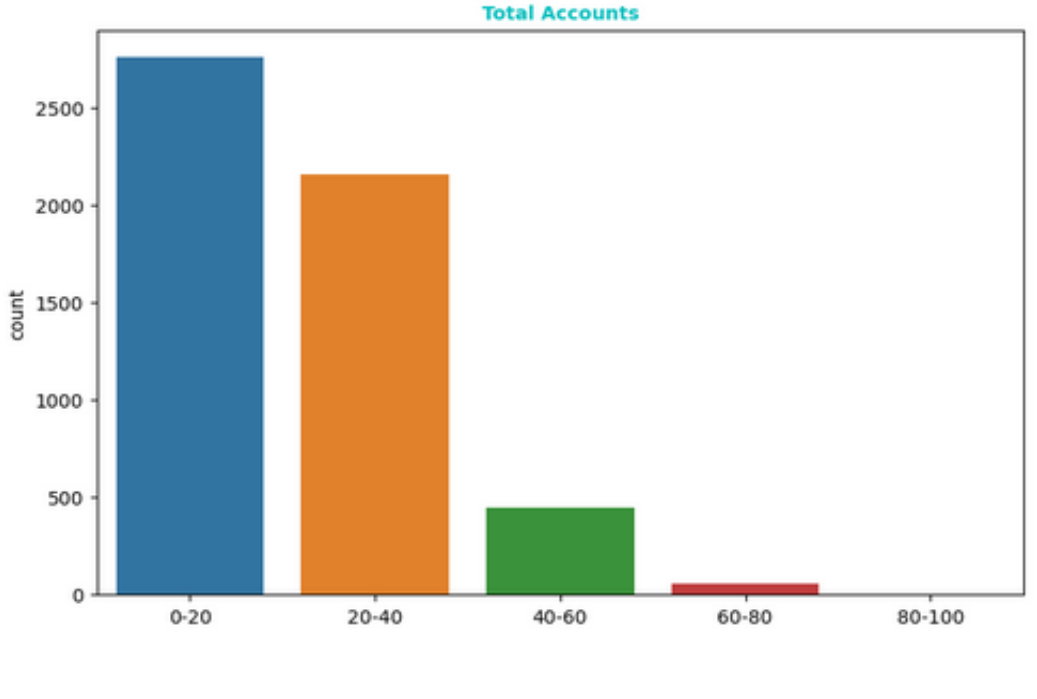


Chances of defaulting when:

- Applicants were issued loan in the month of Dec and in the year 2011

Segmented Univariate Analysis

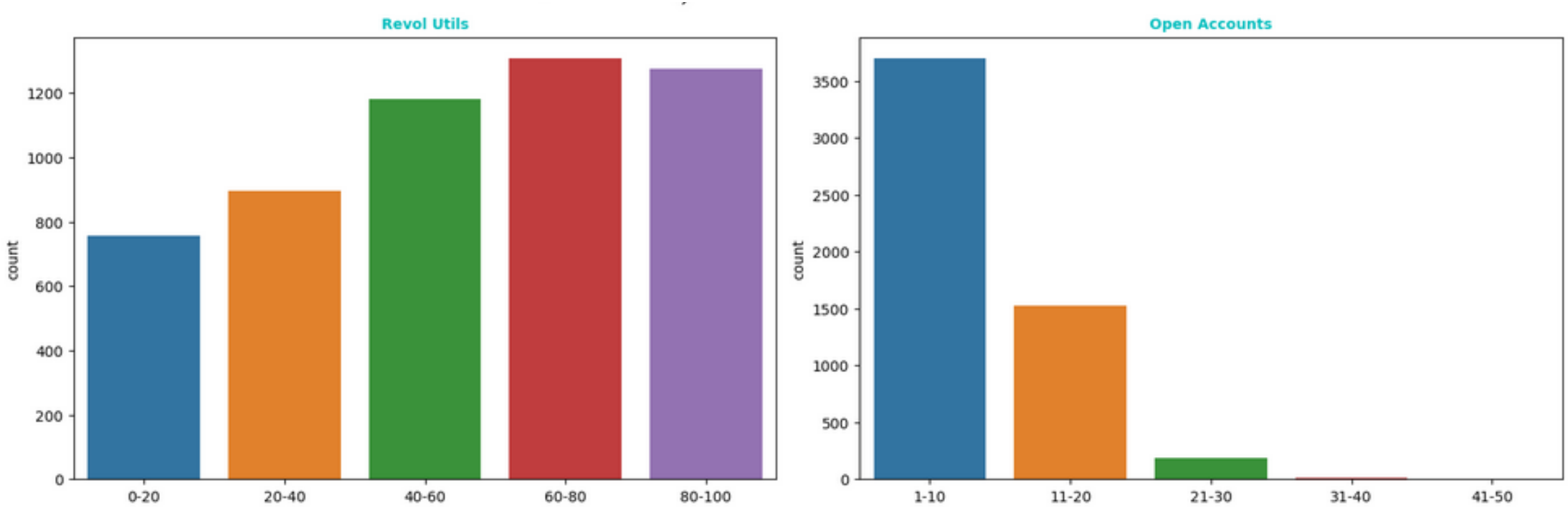




Chances of defaulting when:

- Applicants having annual income between 31-35K
- Applicants having total accounts around 20

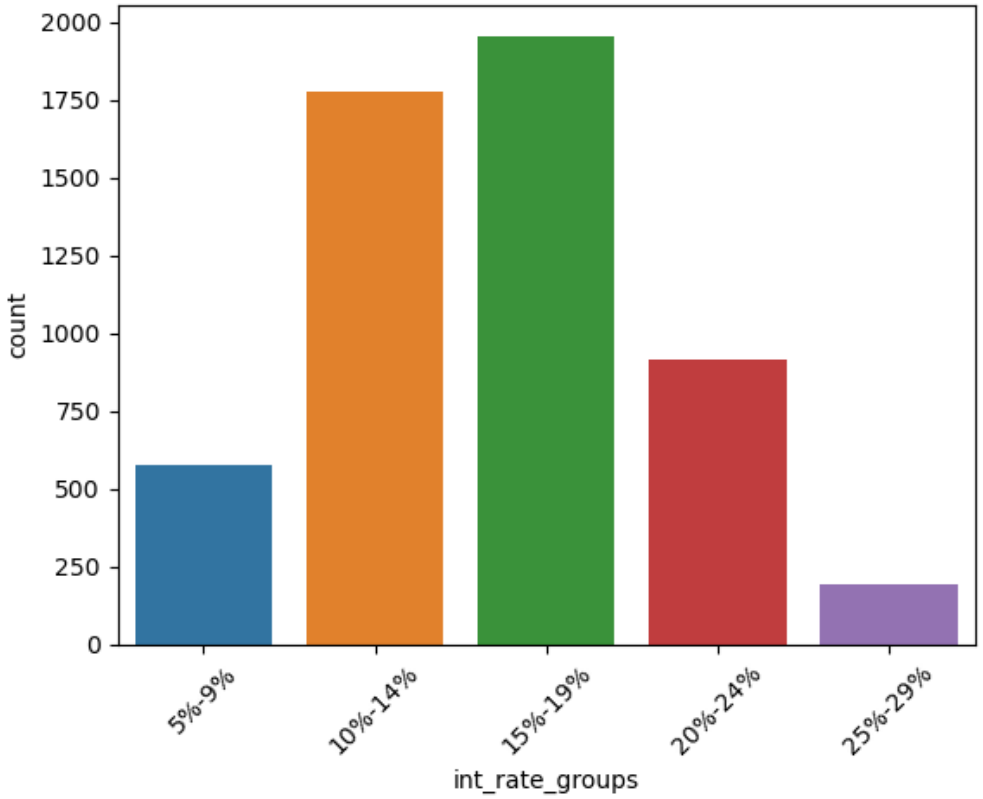
Segmented Univariate Analysis



Chances of defaulting when:

- Applicants with revolving line utilization rate between 60-80
- Applicants having open account around 10

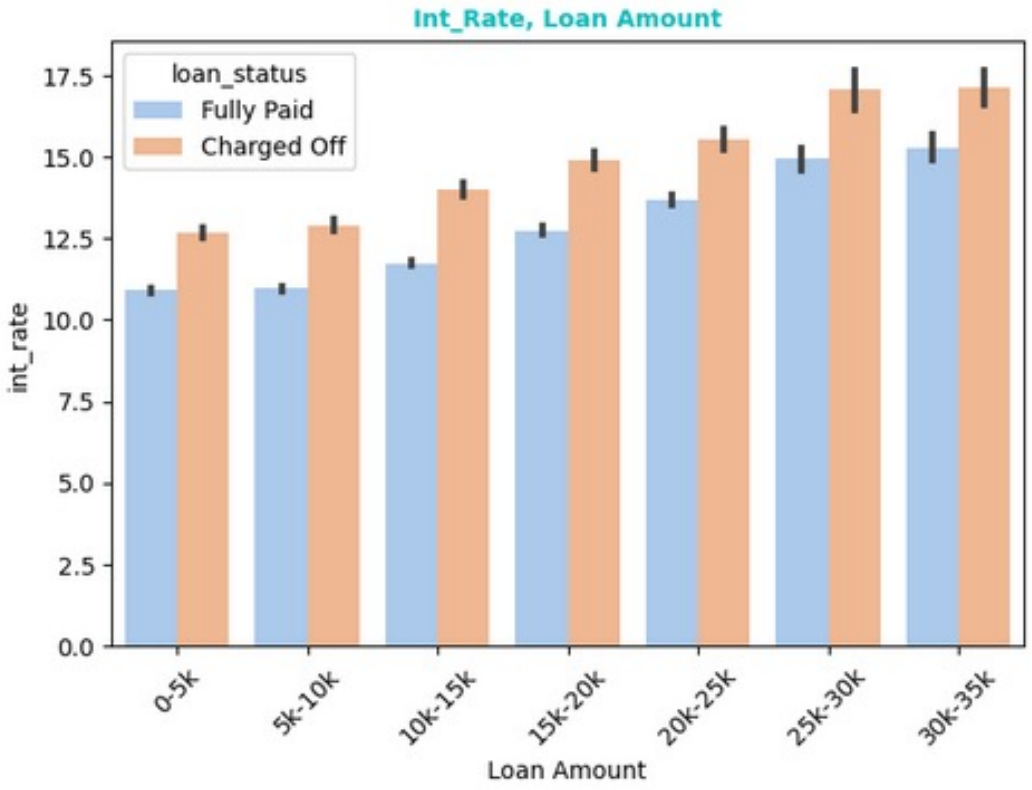
Segmented Univariate Analysis

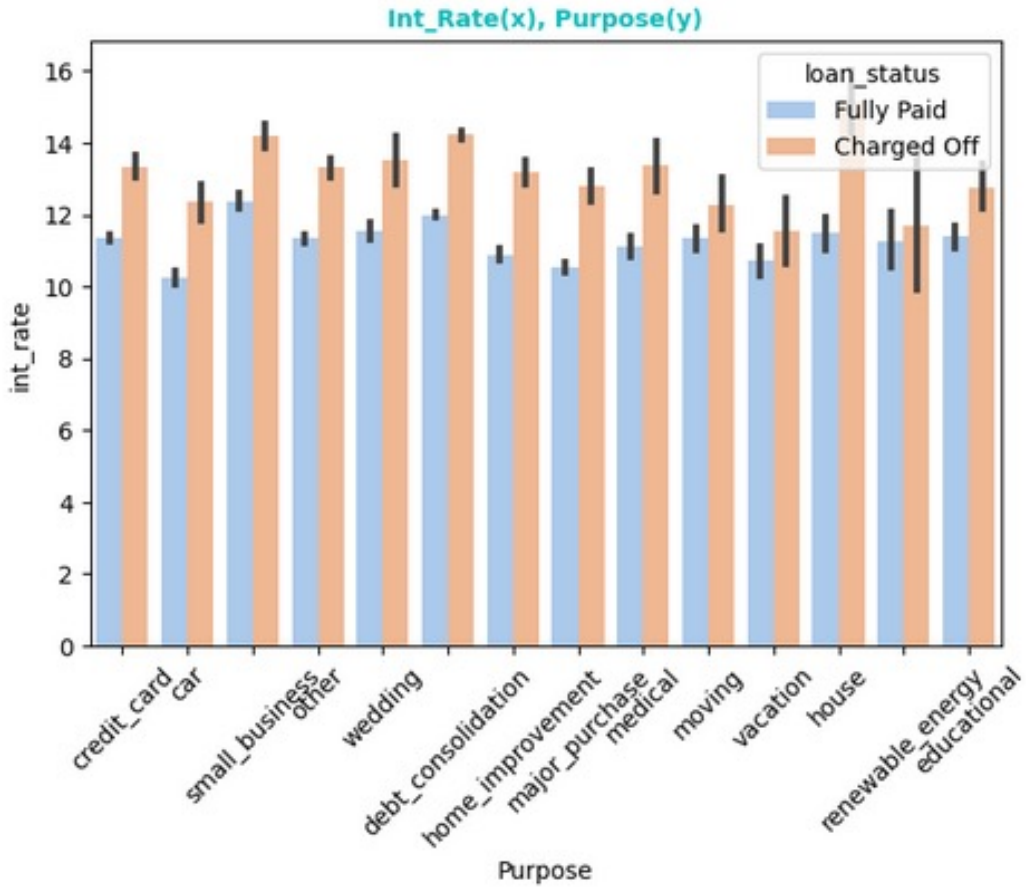


Chances of defaulting when:

- Applicants with interest rates between 15-19%

Bivariate Analysis

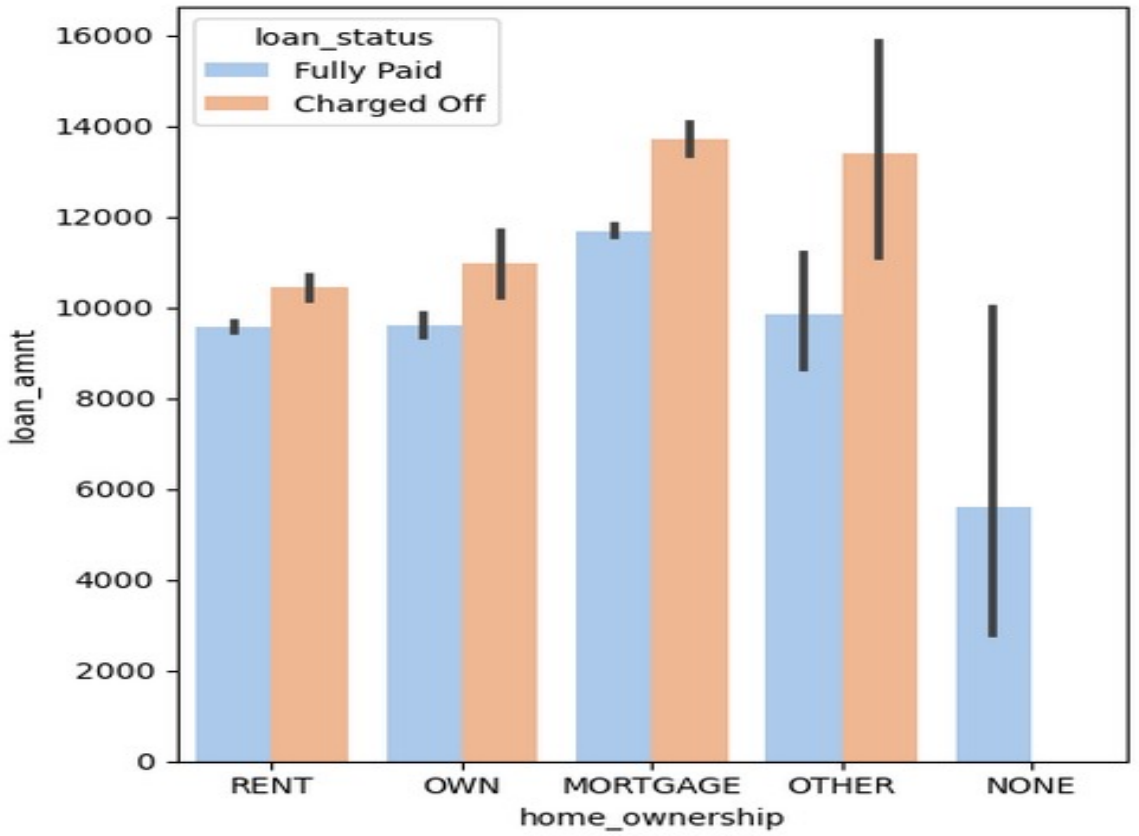


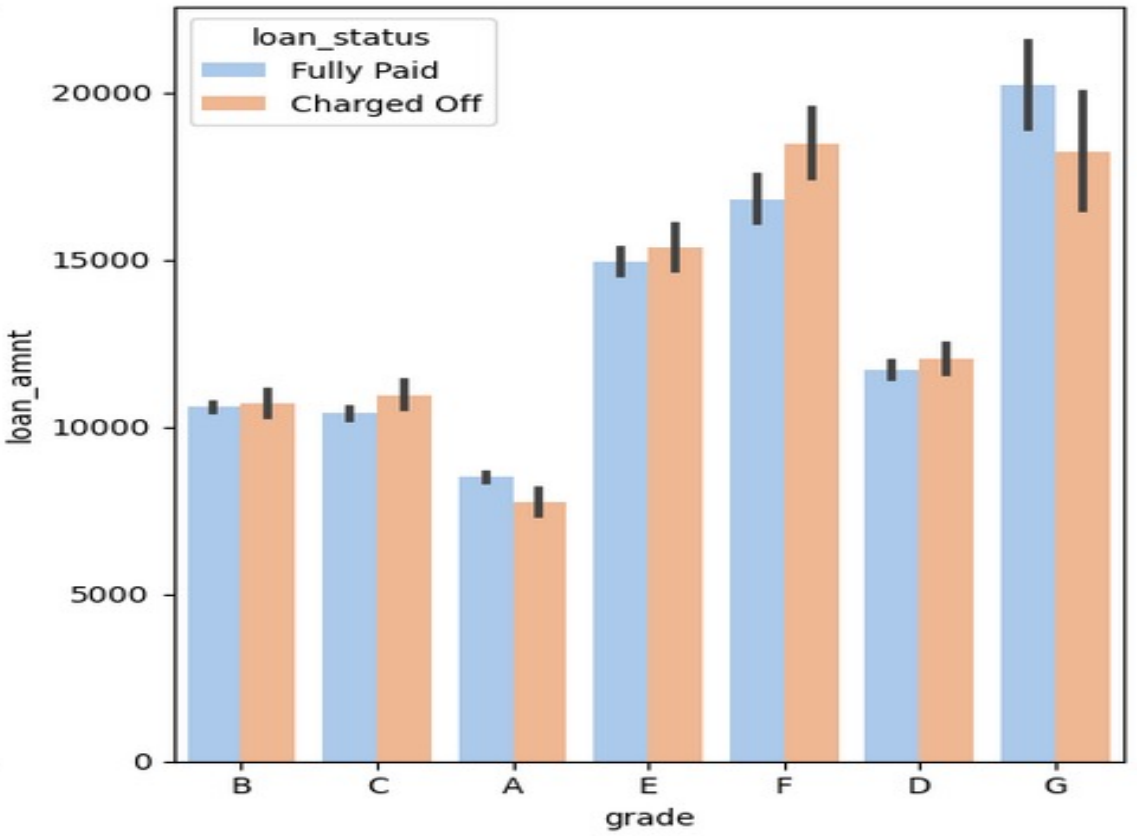


Chances of defaulting when:

- Applicants with interest rates between 15-19% with loan amounts 30-35K
- Applicants with purpose of buying house with interest rate of 12-15%

Bivariate Analysis

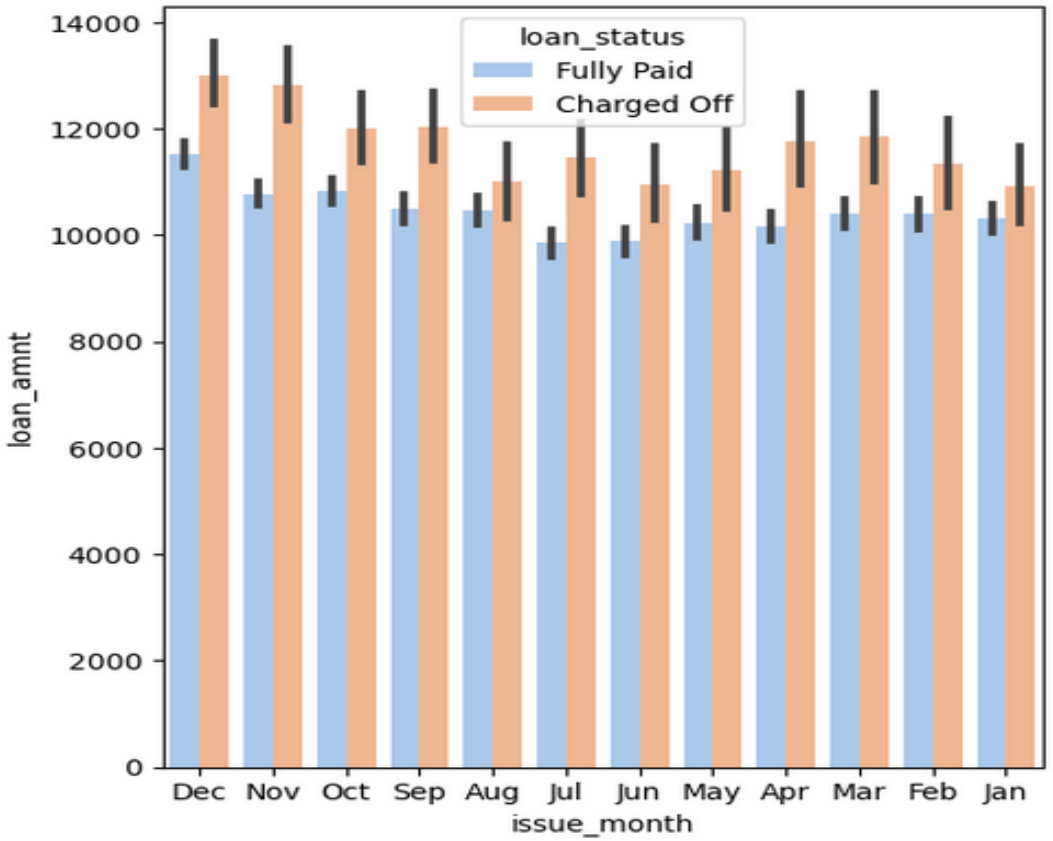


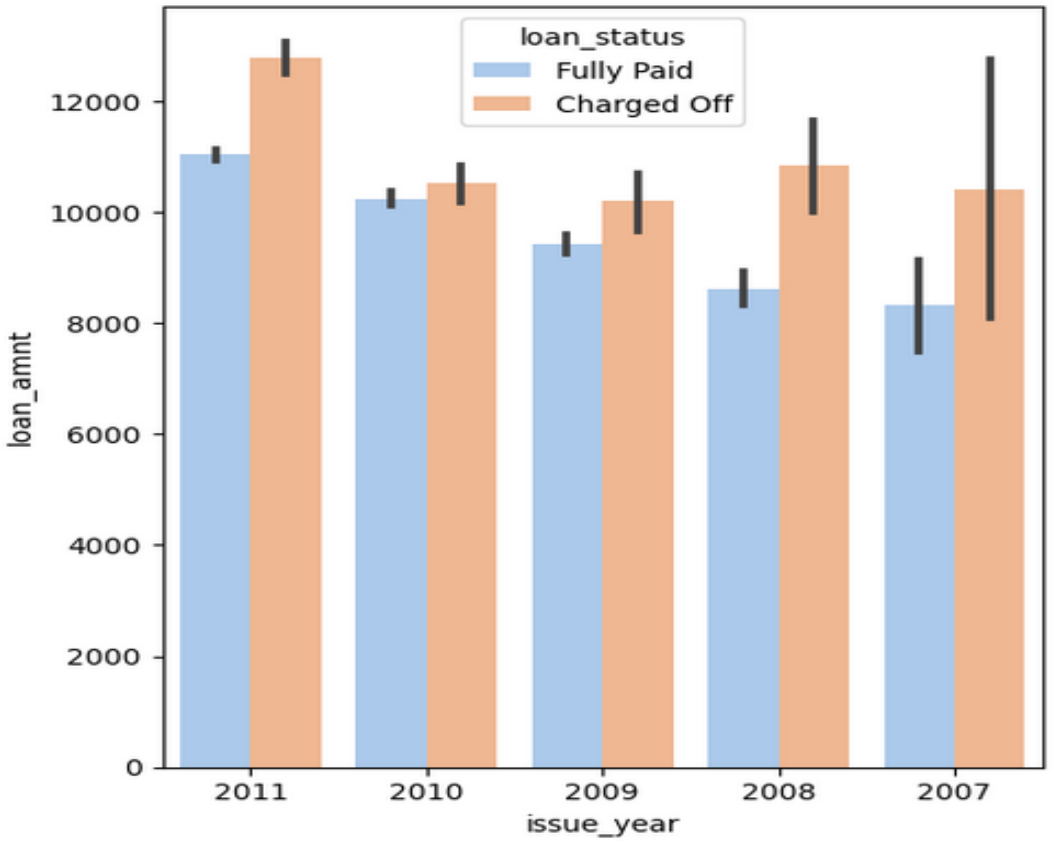


Chances of defaulting when:

- Applicants with MORTGAGE having applied for the loan amounts of 12-14K
- Applicants with loan grade “F” with loan amount between of 15-20K

Bivariate Analysis

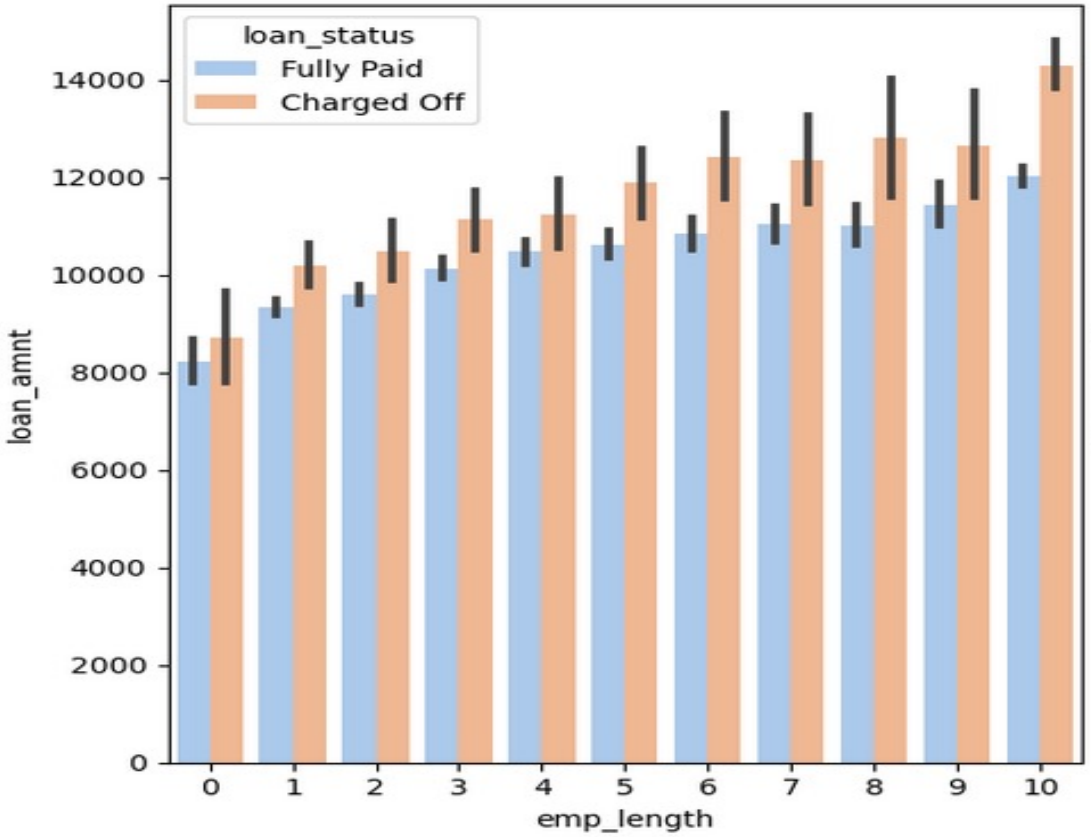


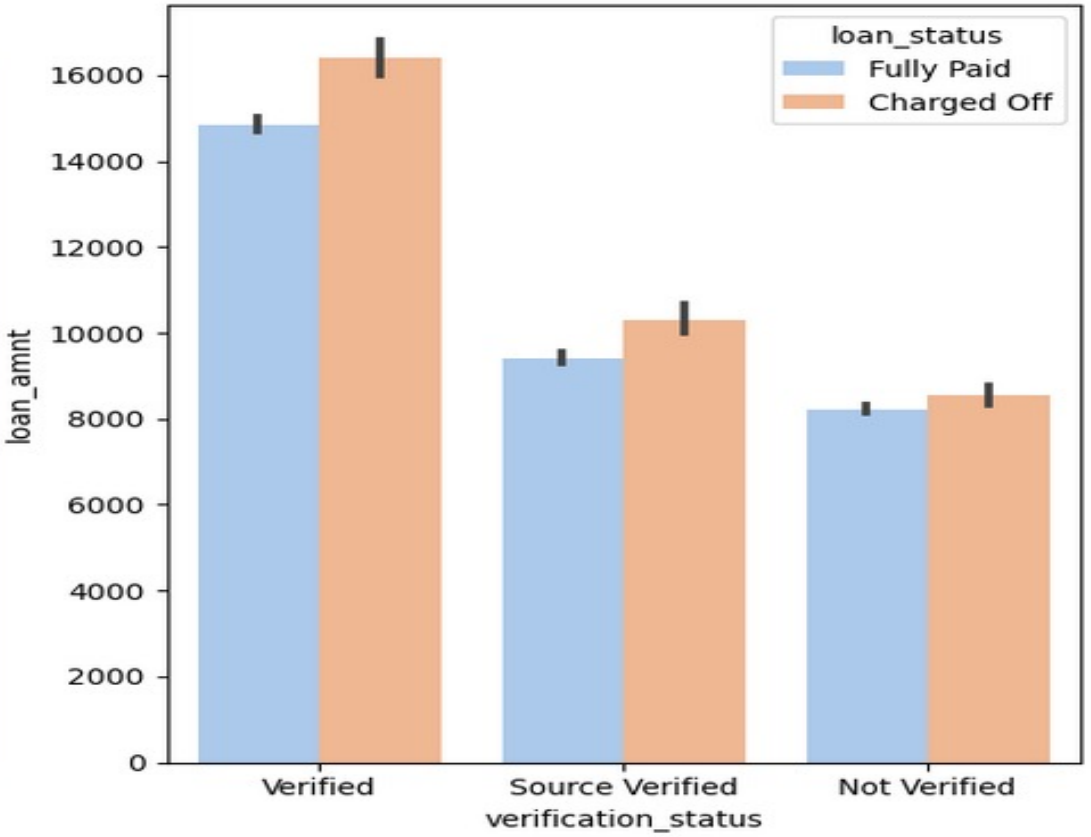


Chances of defaulting when:

- Applicants applied for the loan amounts between 12-14K in December month
- Applicants applied for the loan amounts between 11-13K in the year 2011

Bivariate Analysis





Chances of defaulting when:

- Applicants applied for the loan amounts between 12-14K with employment experience of 10+ years
- Applicants applied for the loan amounts between 15-17k and are verified.