Lending Club Case Study Through Exploratory Data Analysis

-Submitted By,

Vishnu Vardhan Gudla

Lending Club:

- Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.

- Borrowers can easily access lower interest rate loans through a fast online Interface.

Problem Statement:

- Use Exploratory Data Analysis to understand how consumer attributes and loan attributes influence the tendency of default.

Business Objectives:

- Identify Risky loan applicants to reduce credit loss using EDA

- Understand the driving factors behind loan default, i.e., the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Analysis Approach

1.Remove Null Values from rows and columns
2. Identifying and removing unnecessary columns
3. Standardizing text and numbers
4. Removing Outliers

Clean Data → Univariate Analysis → Segmented Univariate Analysis → Bivariate Analysis → Observations and Recommendations

Univariate Analysis

1. Analyzing Target column
2. Analyzing categorical and numerical columns
3. plotting

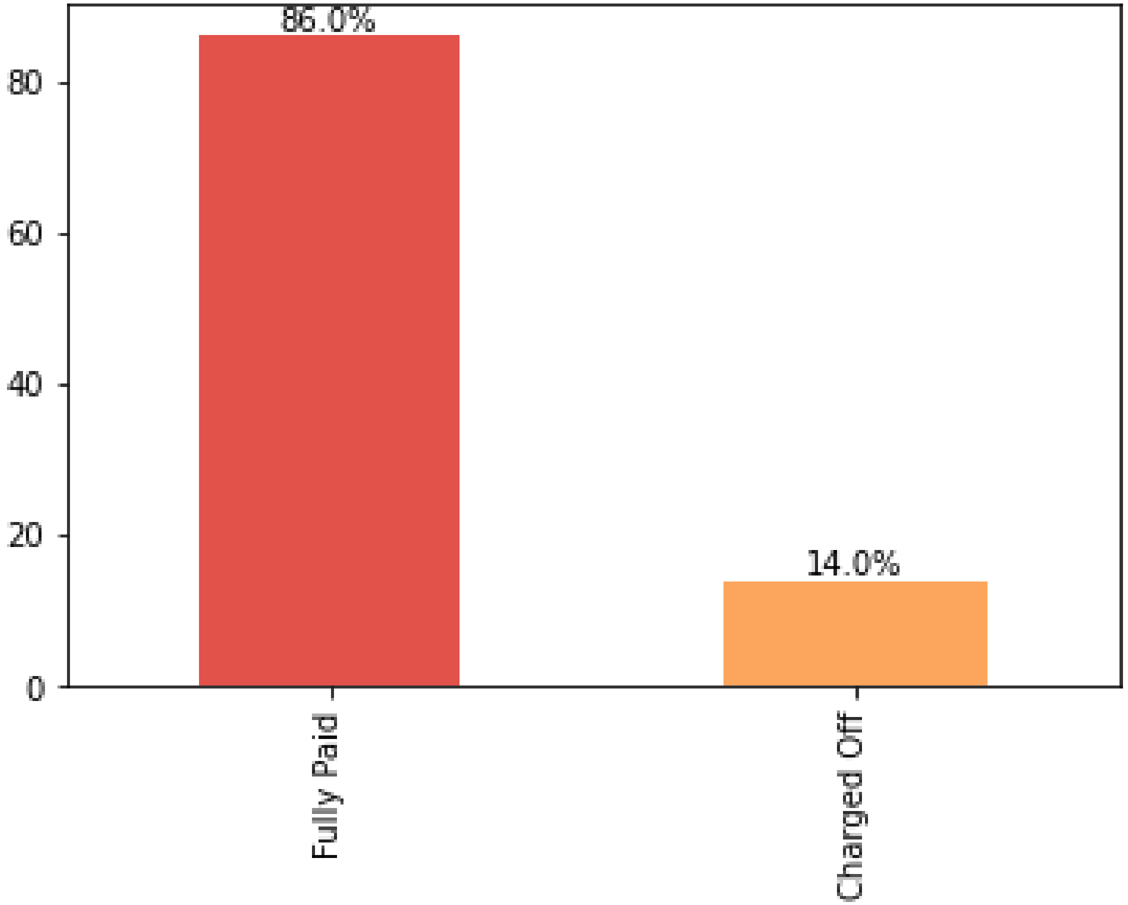
Segmented Univariate Analysis

1. Segment Numerical variables into categories
2. Derived Columns

Bivariate Analysis

1. Analyze two features combinedly and how they are effecting target variable
2. Perform correlation analysis for numerical columns
3. Perform countplot/barplot for categorical variable

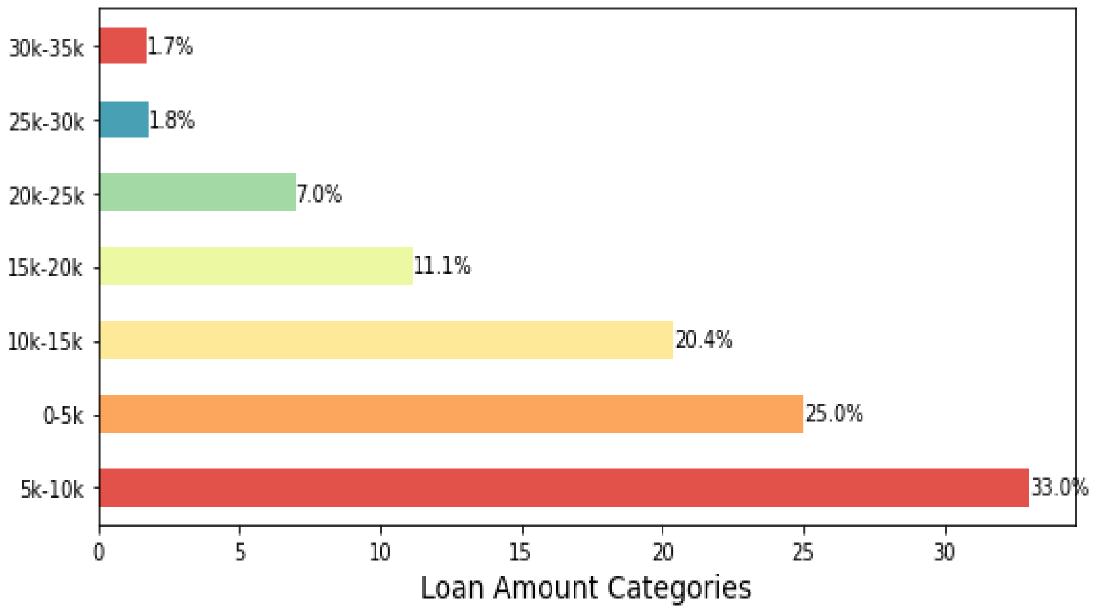
Analyzing Target feature Loan_Status

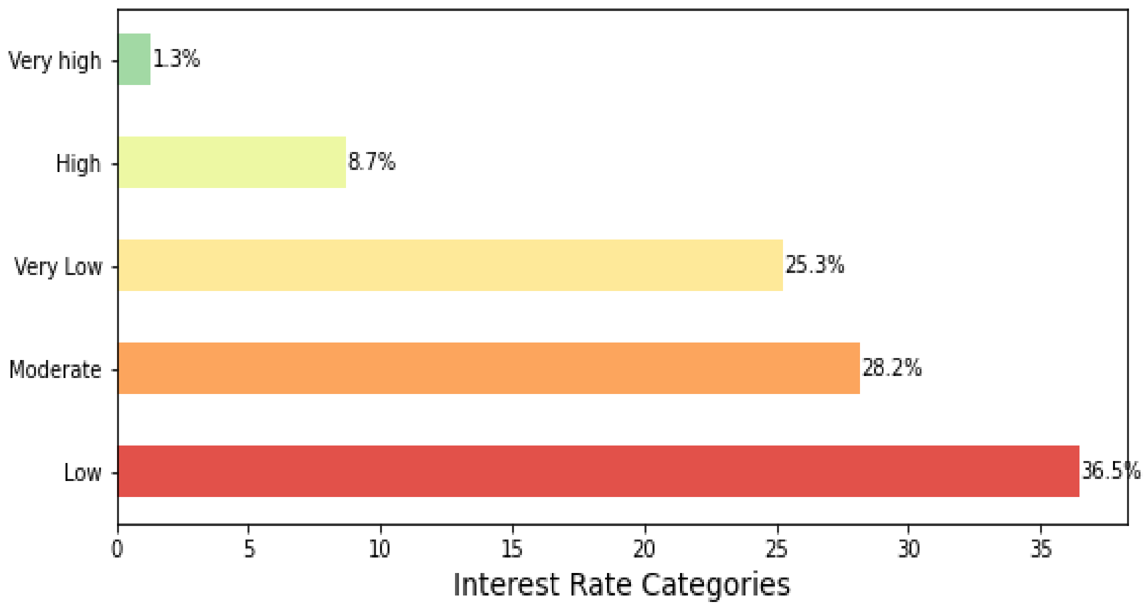


Observation:

- 14% of the Lending Club loans are defaulted
- lending club recovered 86% of loans successfully from year 2007 to 2011

Univariate Analysis

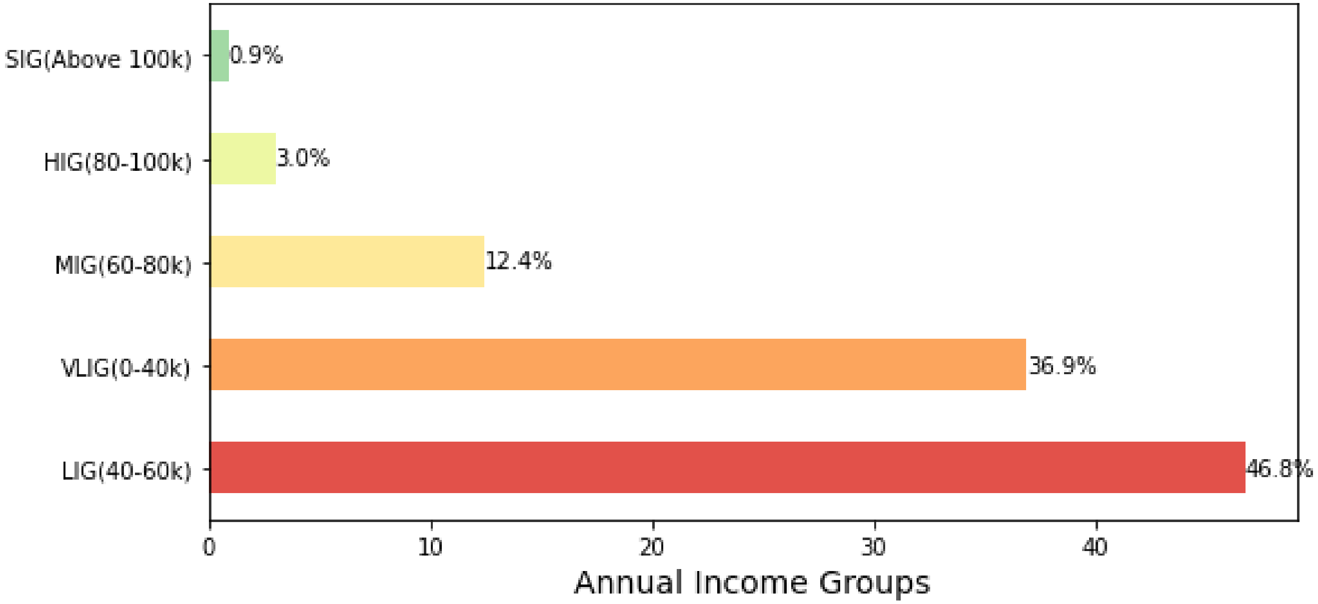




Observation:

- 33 % of the total loans fall under 0-5K loan amount category Interest rates of the most lending club loans are low.

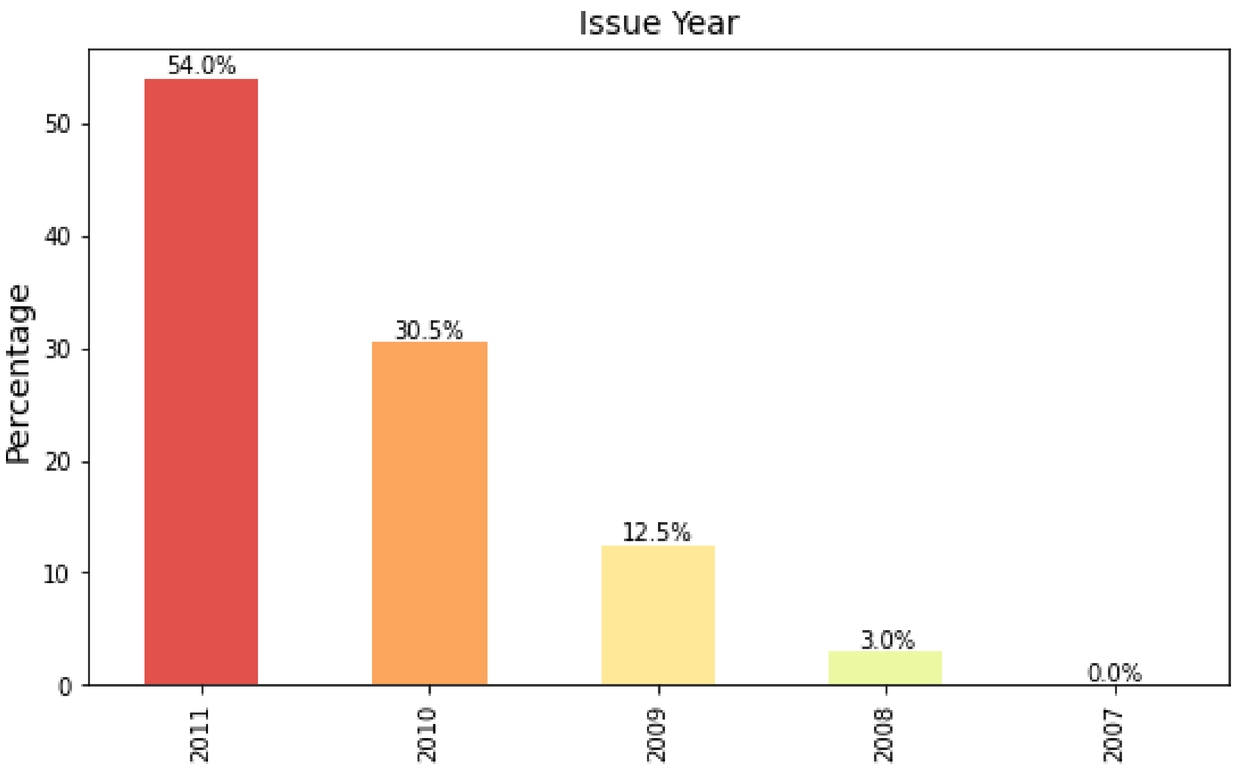
Univariate Analysis

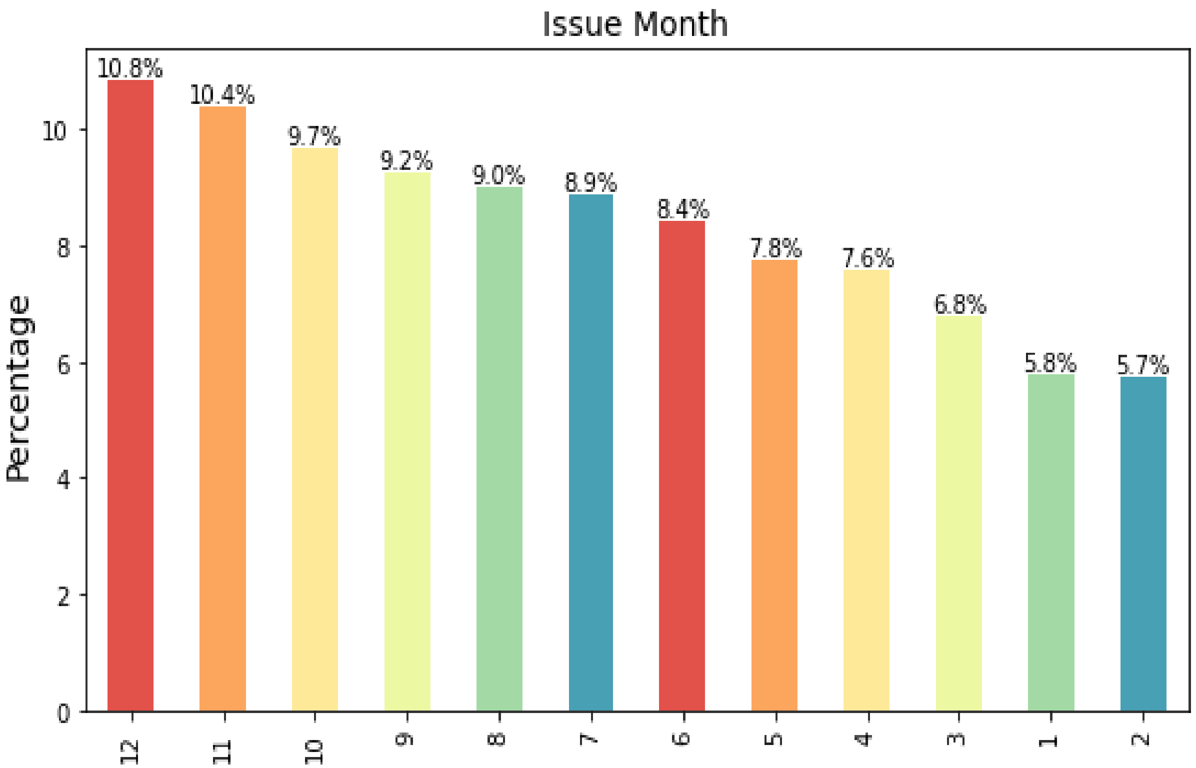


Observation:

- People under Low Income Group (LIG) applied for maximum (47%) of loans

Univariate Analysis





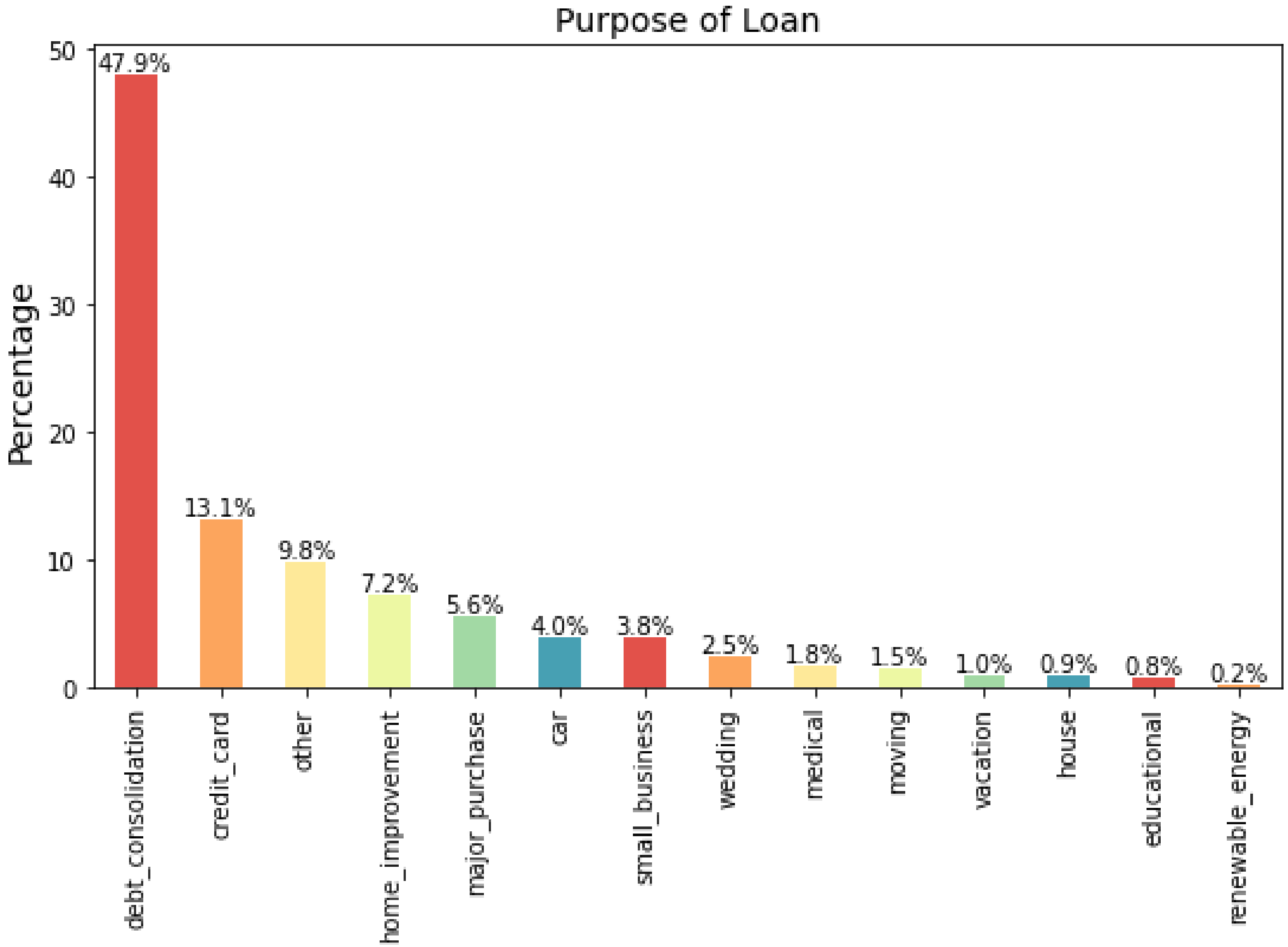
Observation:

- Maximum number (54%) of the loans are issued in the year 2011.

Observation:

- Maximum number (11%) of the loans are issued in the month of December.

Univariate Analysis



Observation:

- Maximum number (54%) of the lending club loans are applied for the purpose of debt consolidation.

Univariate Analysis for Numerical Variables

	loan_amnt	int_rate	dti	installment	revol_bal	revol_util	total_acc	open_acc
count	34328.000000	34328.000000	34328.000000	34328.000000	34328.000000	34328.000000	34328.000000	34328.000000
mean	10974.302319	11.956669	13.412565	319.483303	12182.173095	49.103013	21.651859	9.195642
std	7202.198570	3.685531	6.627309	203.142133	12254.557962	28.212408	10.692519	4.241137
min	500.000000	5.420000	0.000000	16.080000	0.000000	0.000000	2.000000	2.000000
25%	5450.000000	8.940000	8.330000	166.800000	3688.000000	25.900000	13.000000	6.000000
50%	9600.000000	11.830000	13.510000	277.350000	8708.000000	49.700000	20.000000	9.000000
75%	15000.000000	14.460000	18.640000	419.847500	16389.250000	72.500000	28.000000	12.000000
max	35000.000000	24.400000	29.990000	1305.190000	78935.000000	99.900000	55.000000	39.000000

Observation:

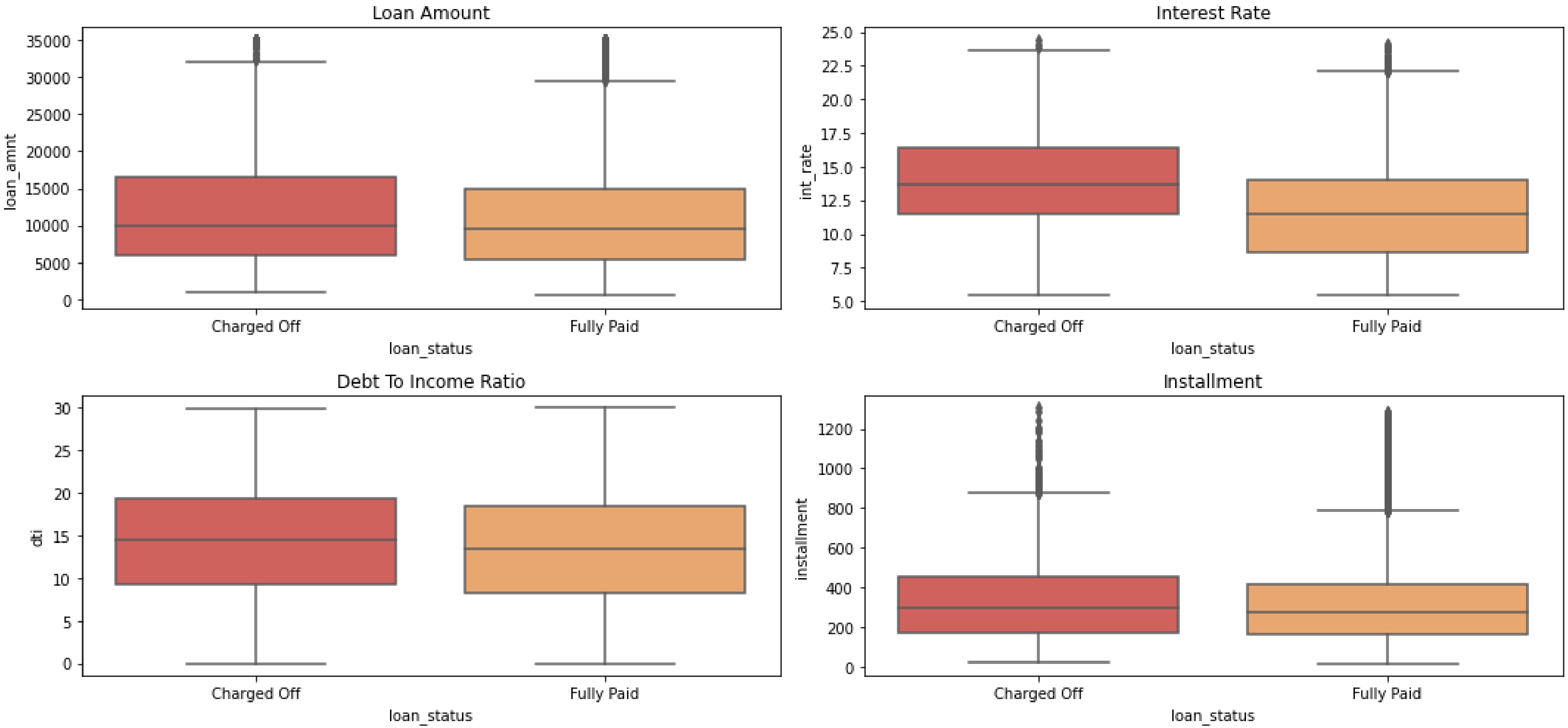
Loan Amount:
75% of the loans have an loan amount below 15000 dollars.
mean is less the median, which indicates that the loan amount data is negatively skewed.

Interest Rate:
Majority of the loans have an interest rate between 9-12%
Mean and Median are almost similar, which indicates zero skew and data following perfect normal distribution

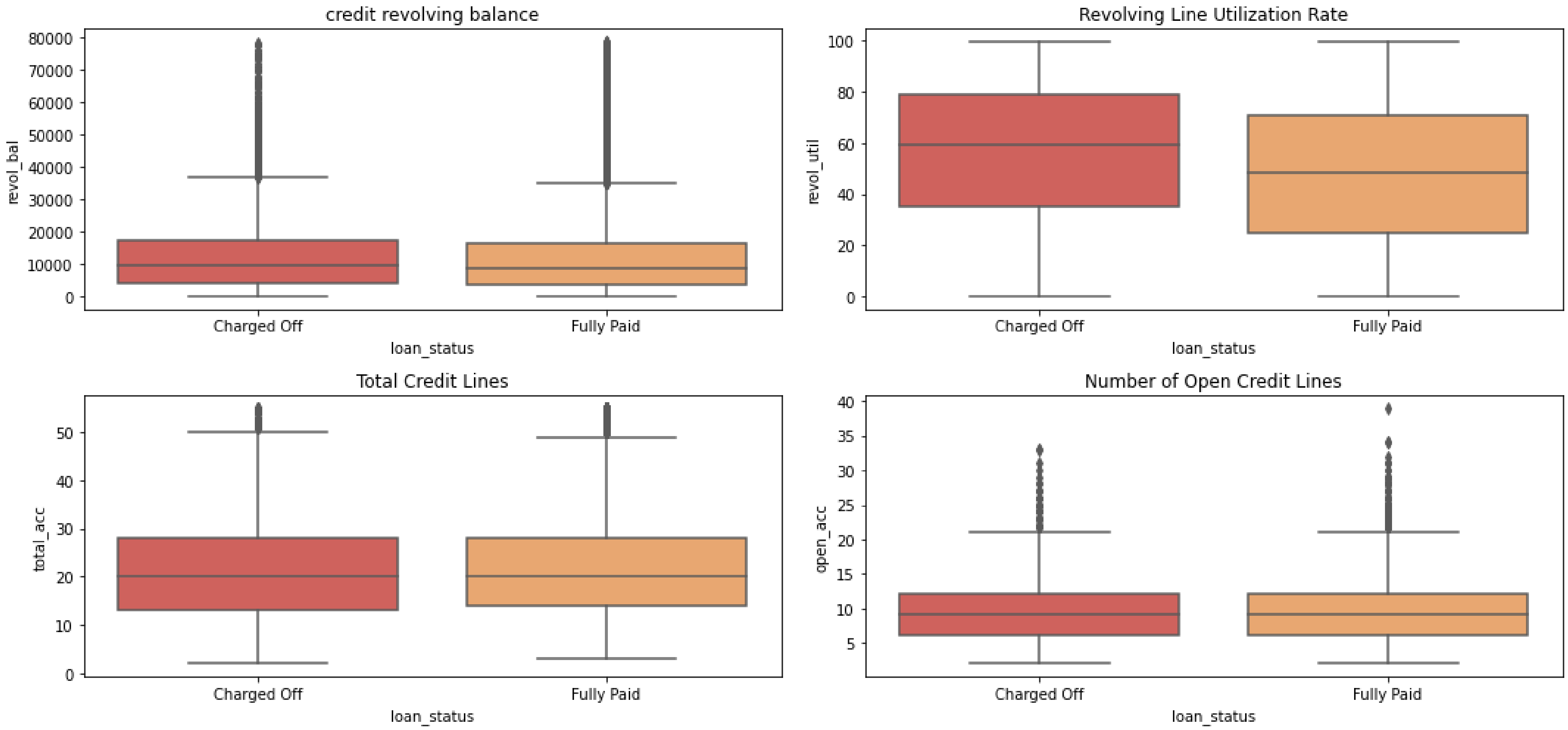
Debt-to-Income Ratio: ranges between 0 and 30 with inter quartile range of 10

Installment: The IQR is 253 and mean greater than median, indicating the data is positively skewed.

Bivariate Analysis for Numerical Variables



Bivariate Analysis for Numerical Variables



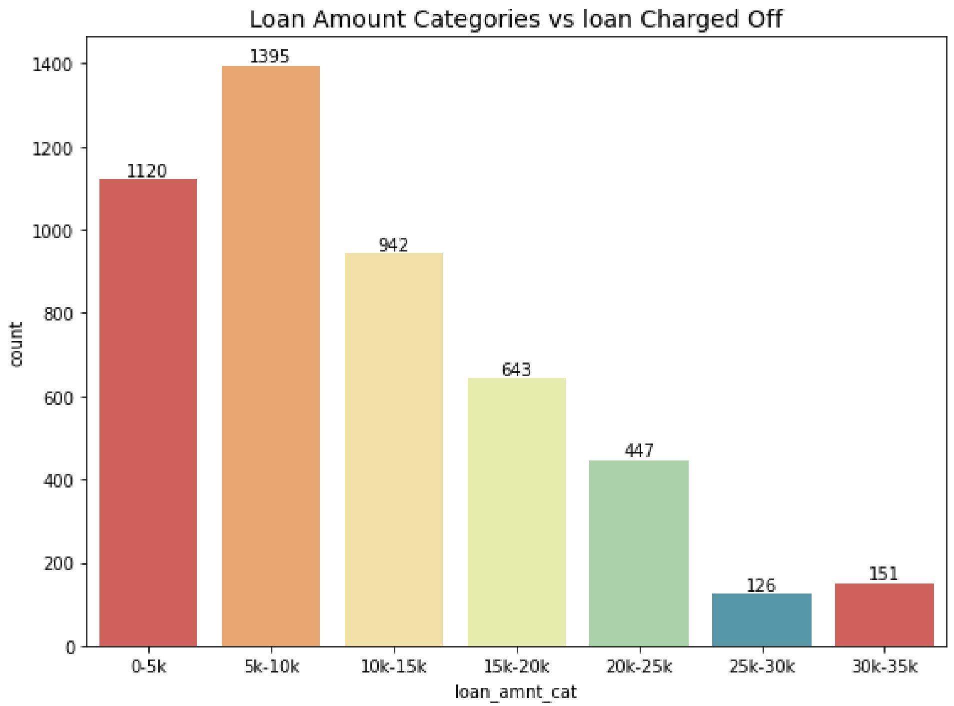
Univariate Analysis for Numerical Variables

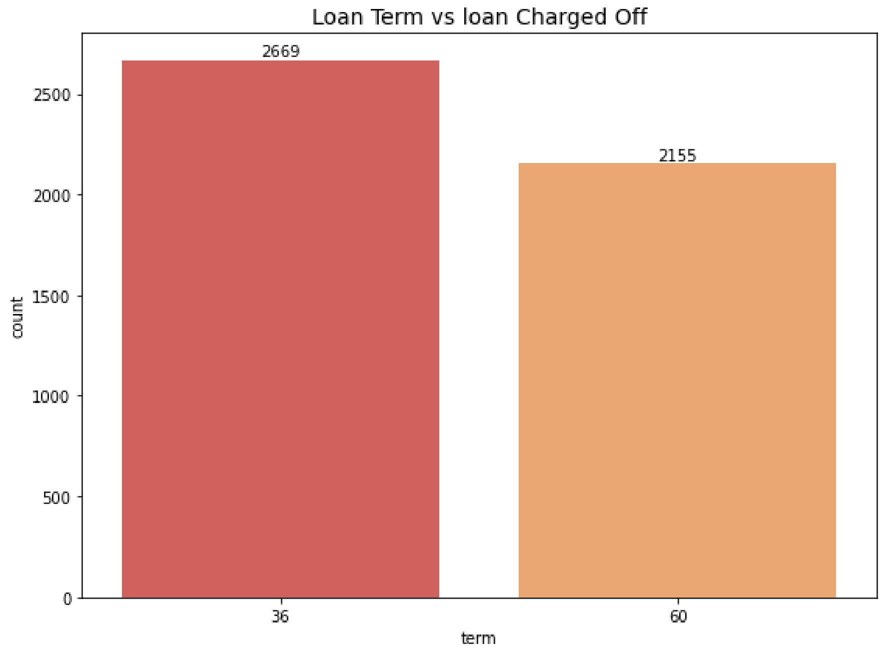
Observation:

After analyzing the above numerical columns with loan status, the major reasons for loan Charged Off/default are:

- Higher Loan amount
- Interest Rate
- Higher Debt-to-Income Ratio
- Higher Revolving line Utilization rate

Bivariate Analysis



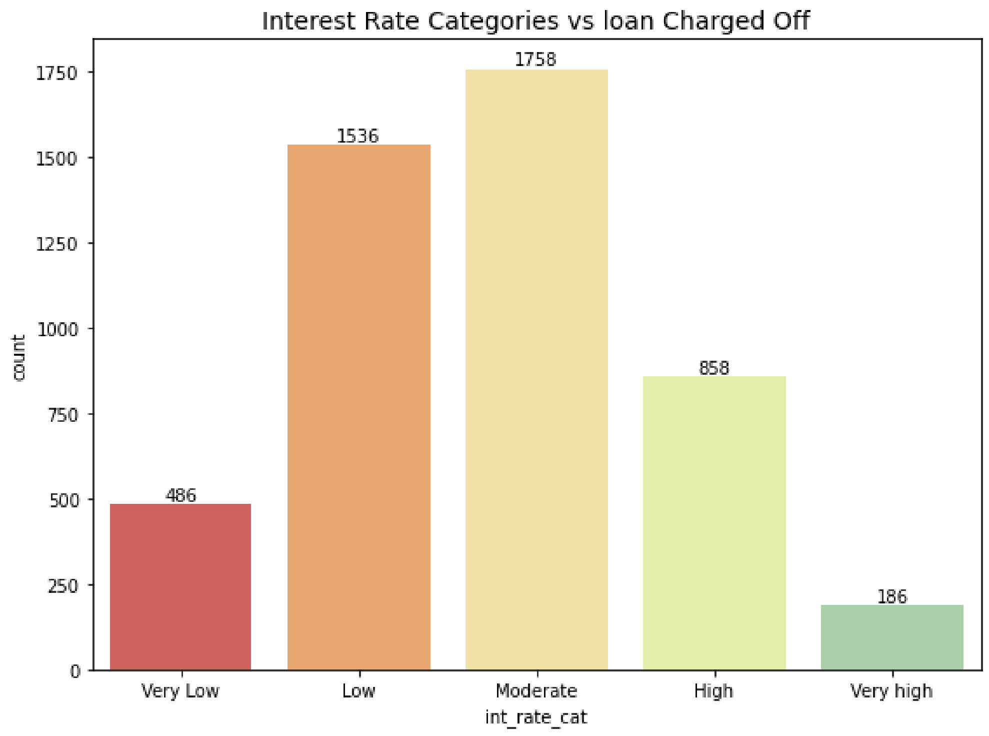


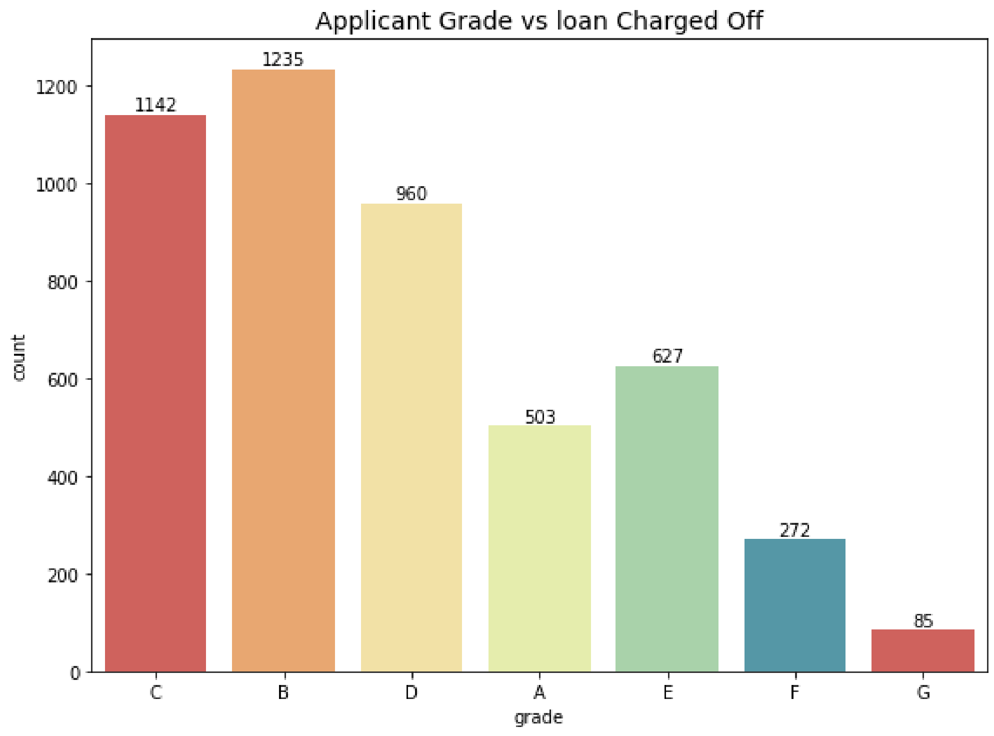
Observation:

The loan Charged Off/default happens when:

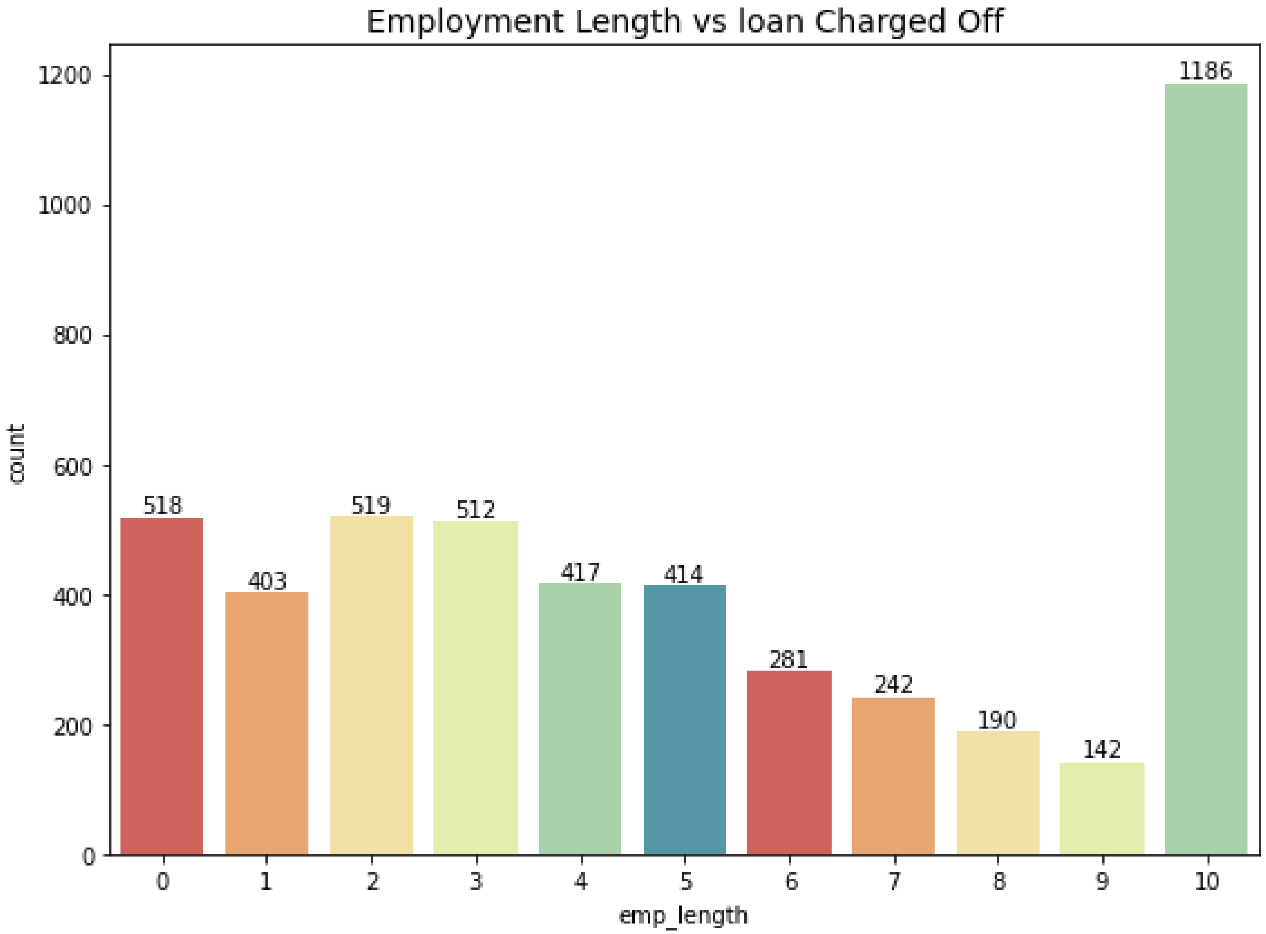
- loan amount in range 5-10K
- loan term is less (3 years)

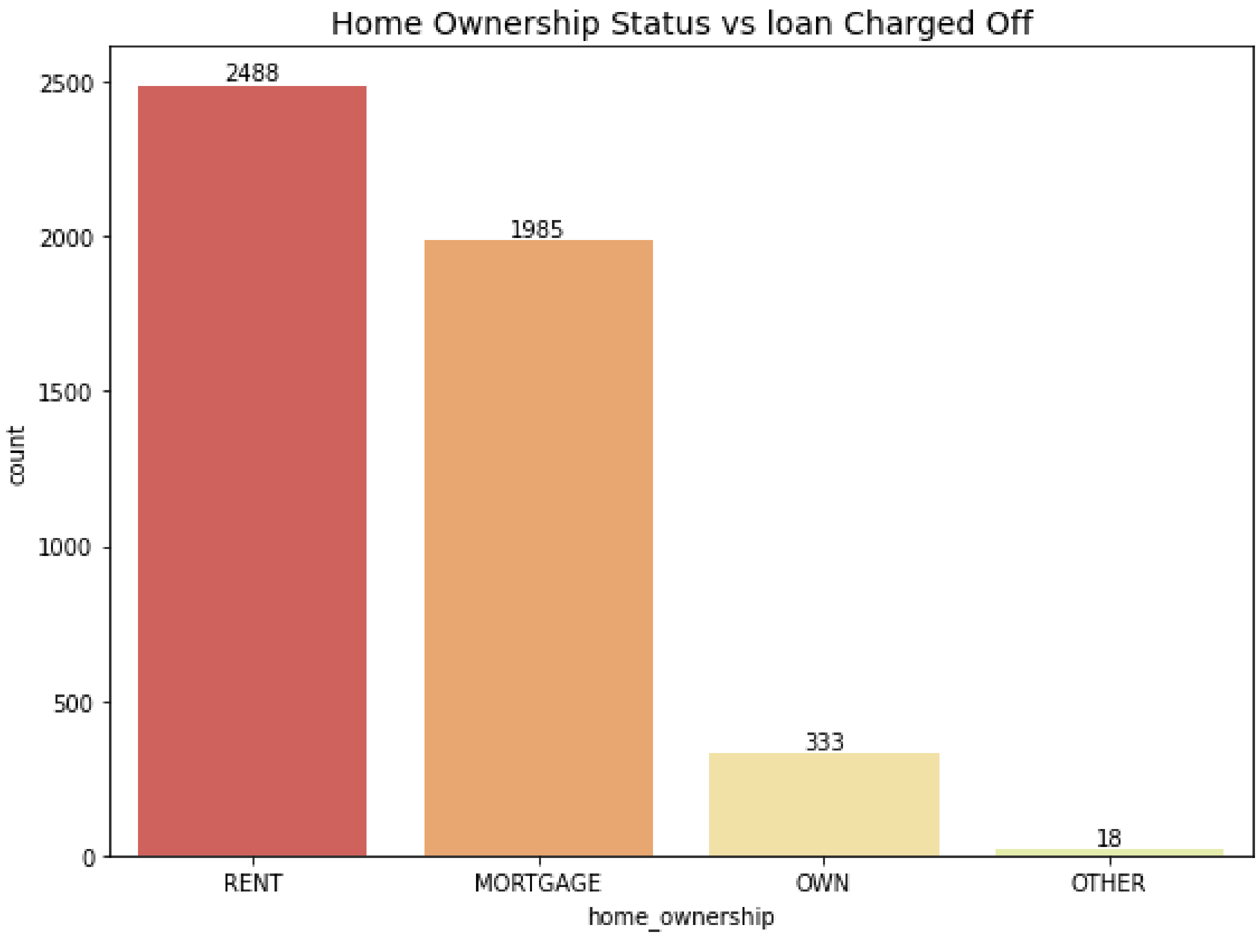
Bivariate Analysis



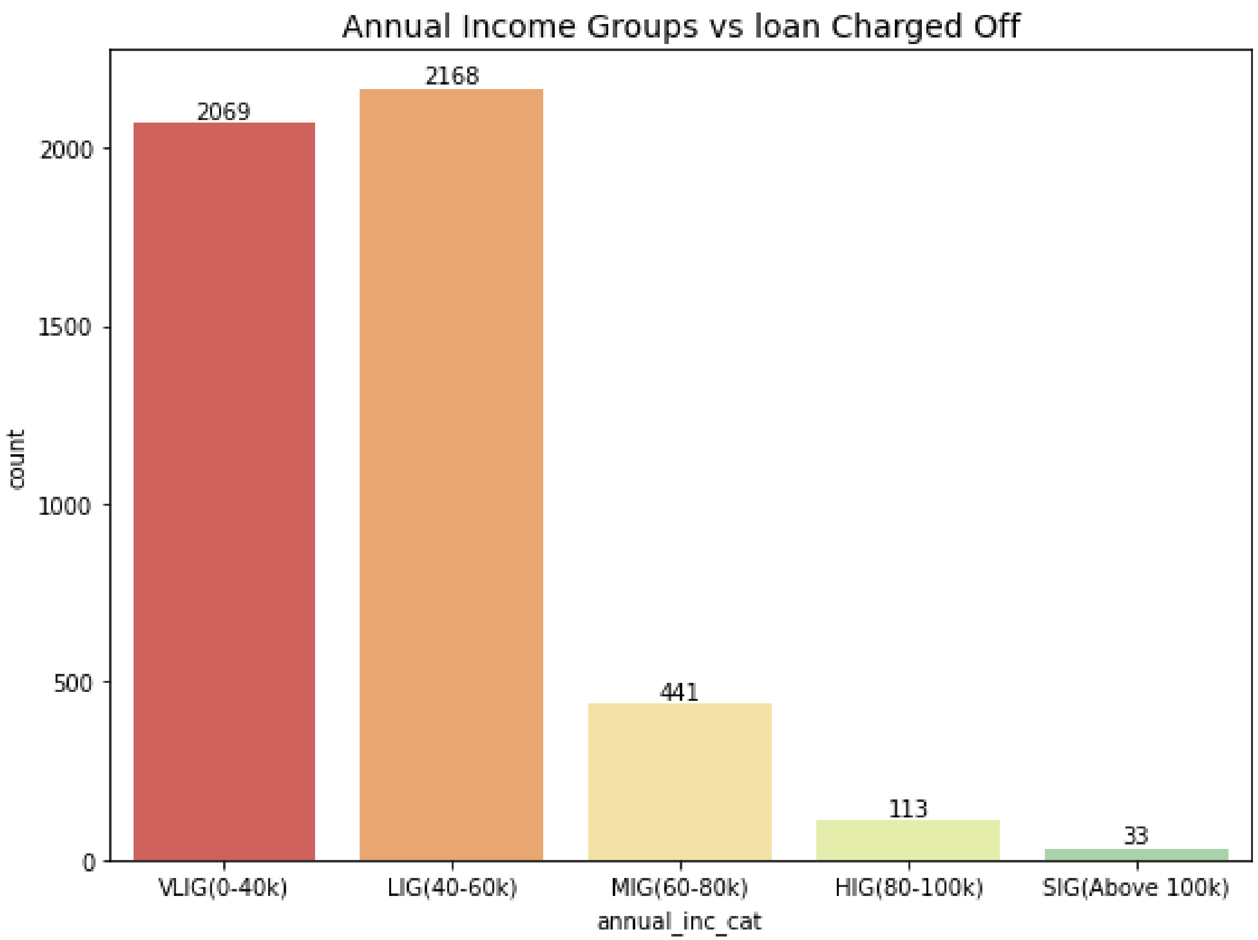


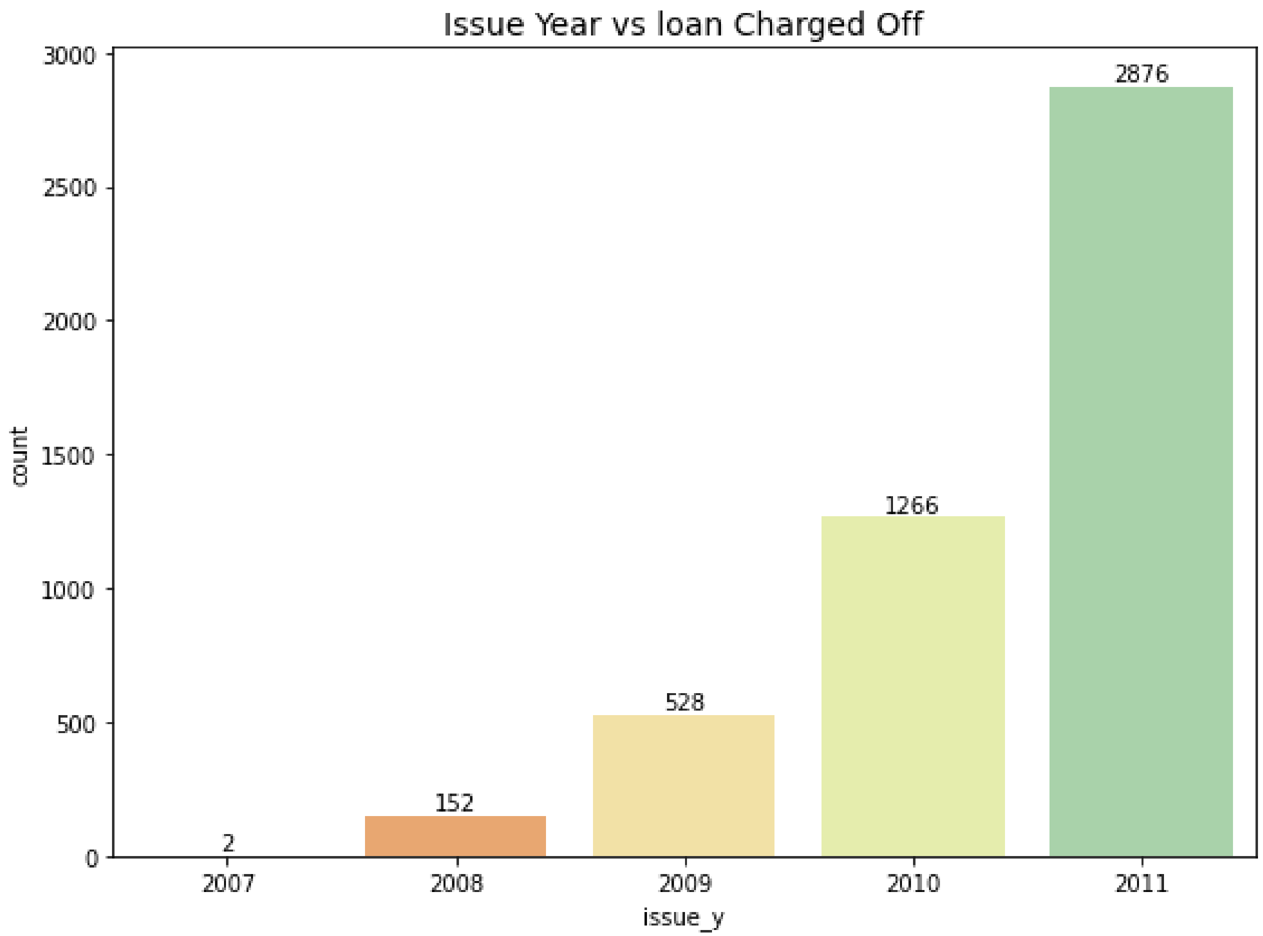
Bivariate Analysis



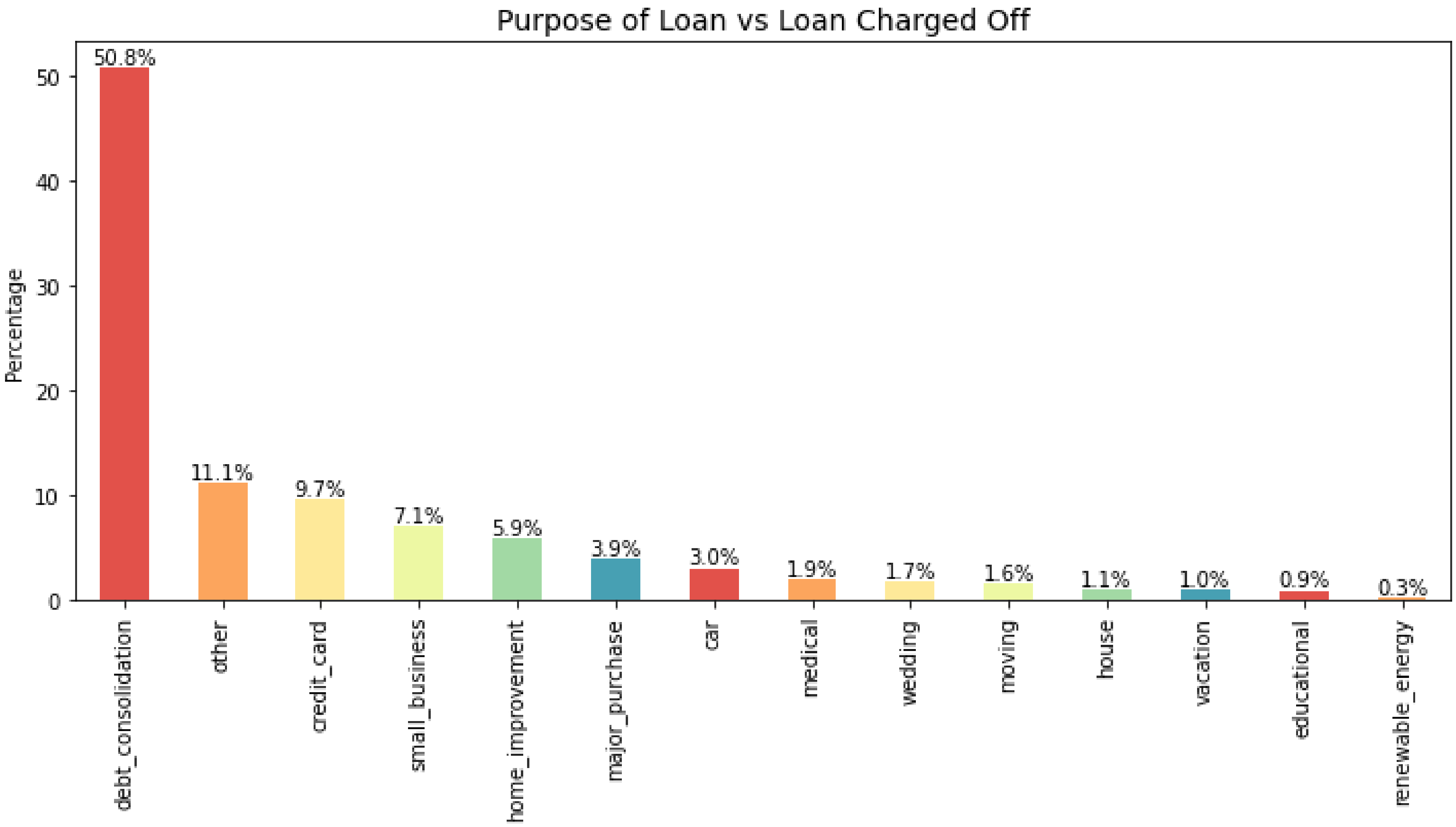


Bivariate Analysis





Bivariate Analysis

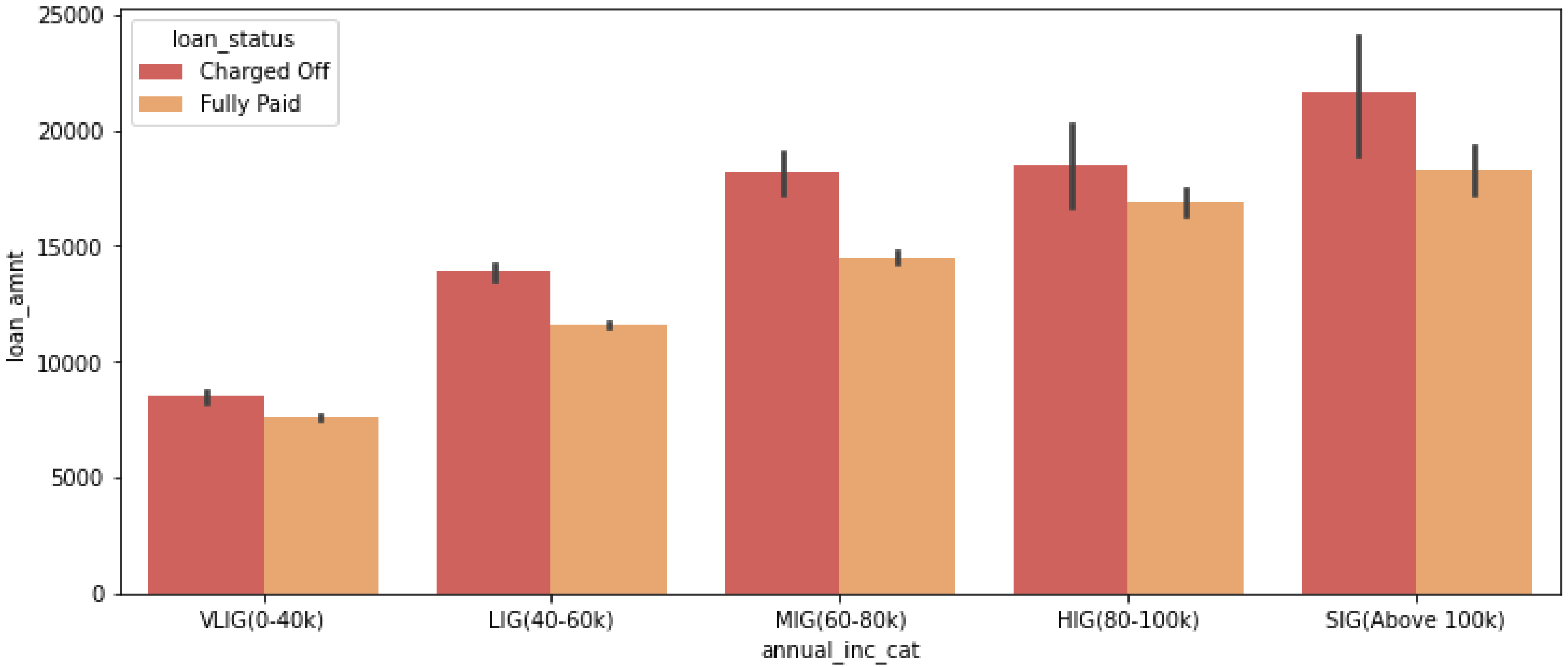


Bivariate Analysis

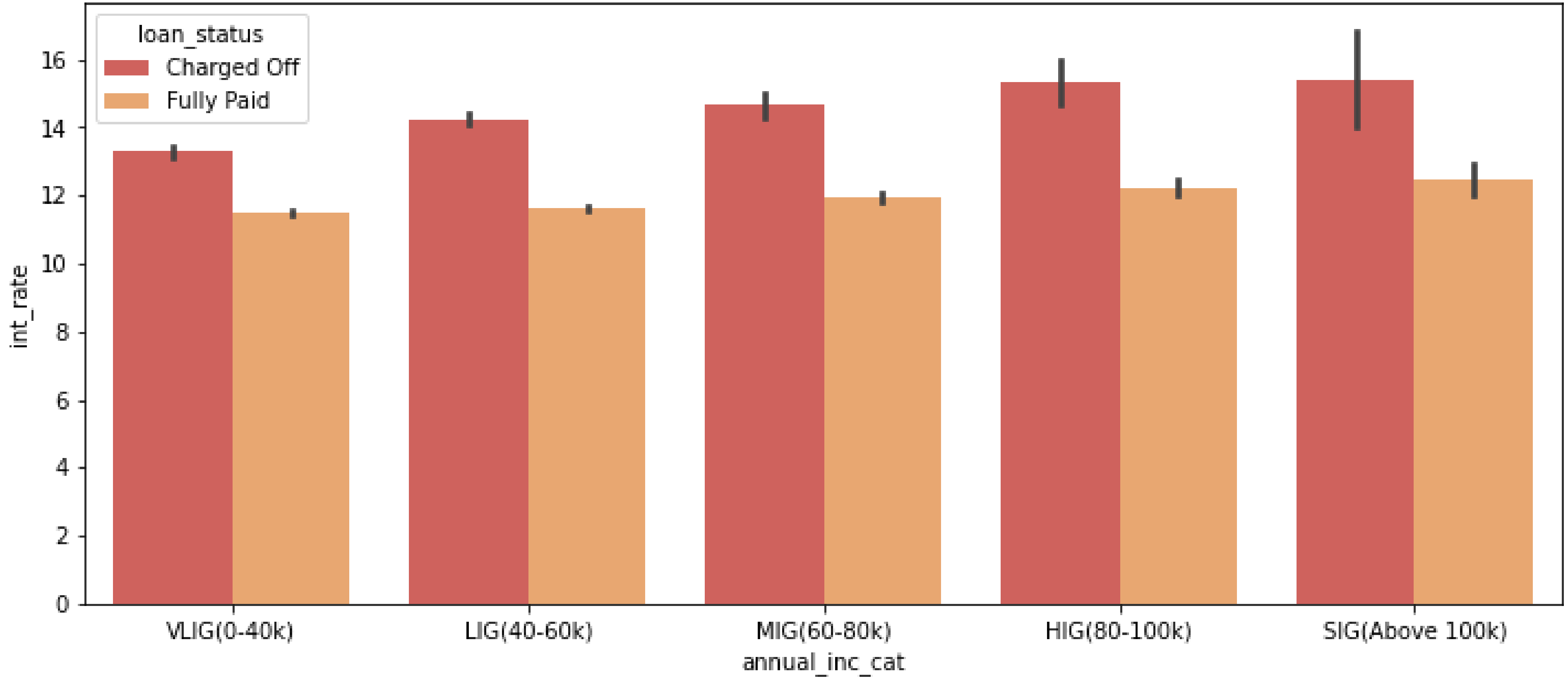
Observations: After analyzing the above categorical columns with loan status, the **loan Charged Off/default** happens when:

1. loan amount in range 5-10K
2. loan term is less (3 years)
3. Interest Rate is Moderate/low (More loans are aprroved with low/moderate interest rates)
4. Applicants belong to Grades B and C
5. Applicants have more than 10 years of employment experience (More loans are approved for higher cadre employee).
6. Applicants living in rented houses and have house/property mortgage.
7. Applicants belong to Low and very Low income groups.
8. Loans which are not properly verified are defaulted more.
9. Most of the loans are defaulted in the year 2011 and in the month of december.
10. loans of the applicants who have done 1 or 2 inquiries in the last 6 months.
11. loans applied by employess from US Army are defaulted more when compared to other organizations (More loans are approved for US Army employees).
12. Loans applied for debt consolidation purpose defaulted more.

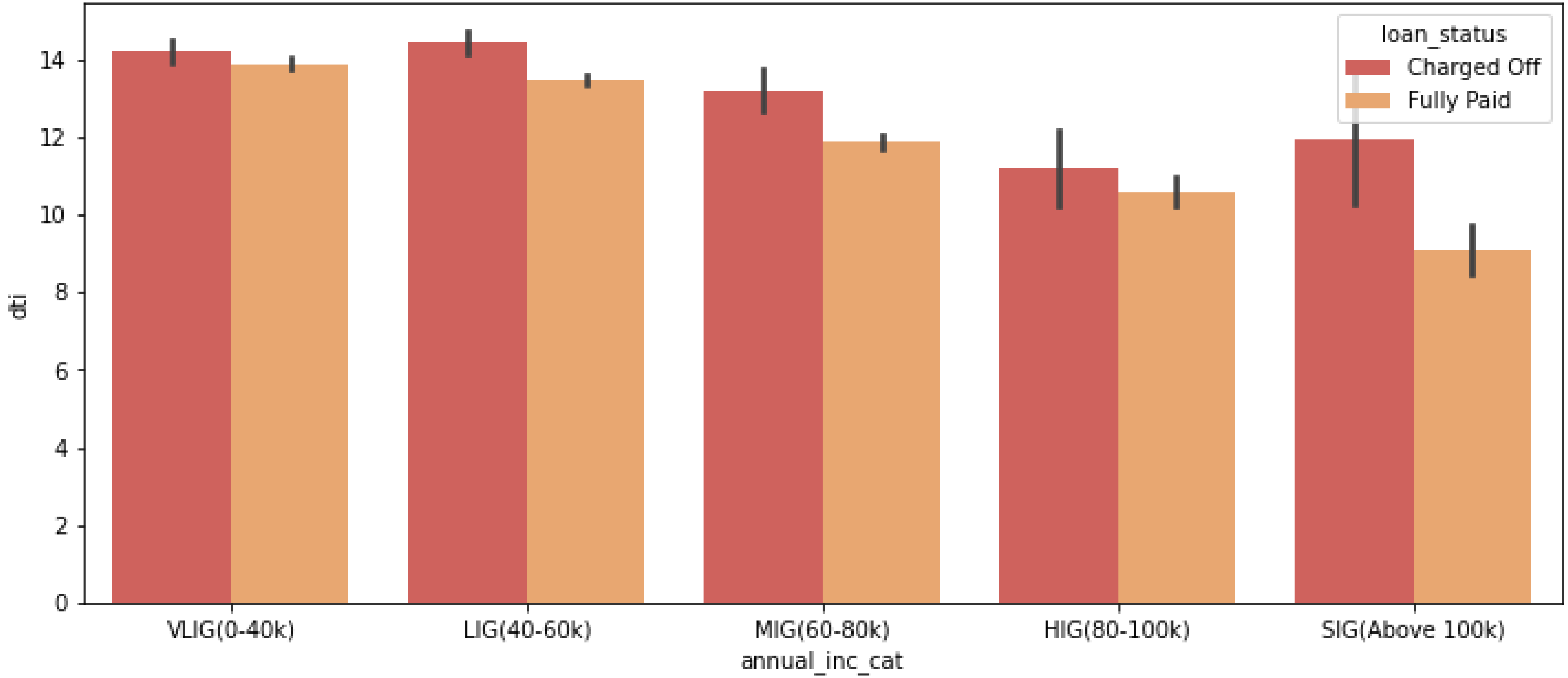
Bivariate Analysis



Bivariate Analysis



Bivariate Analysis

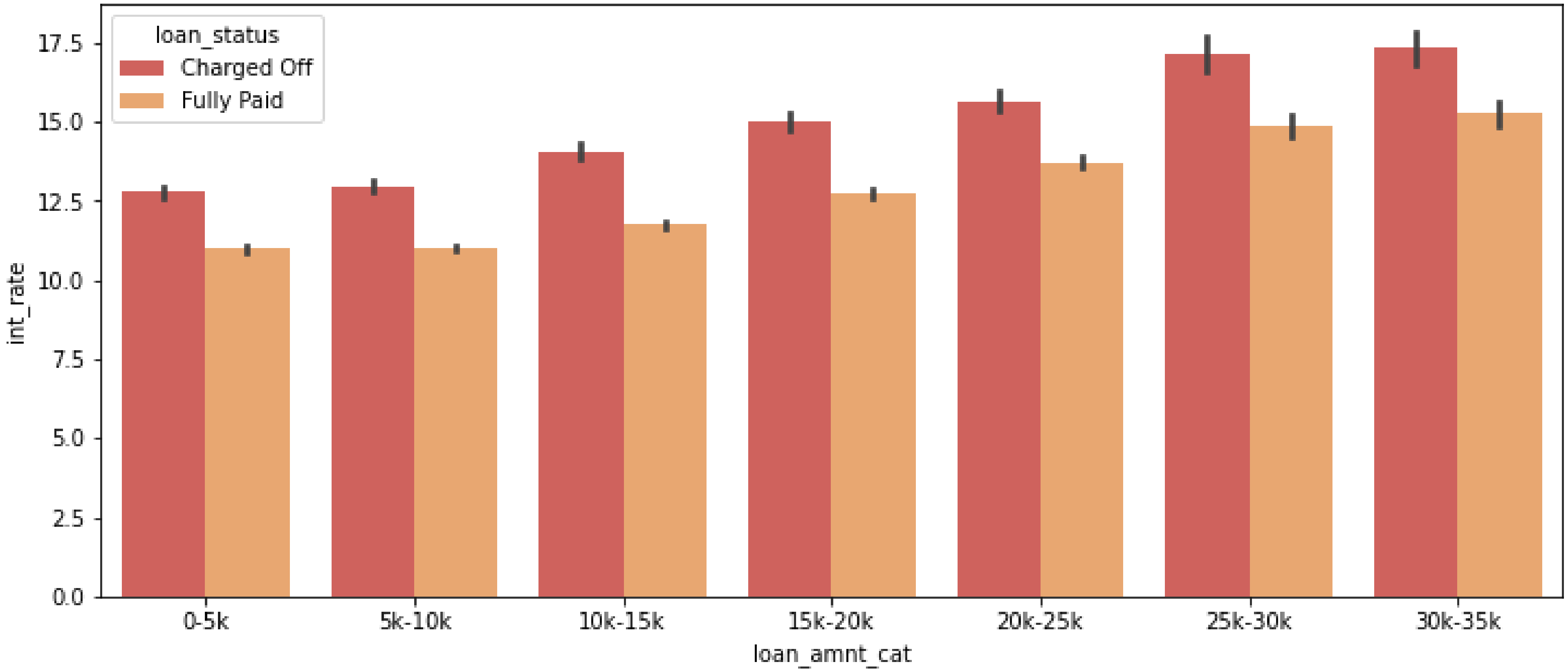


Bivariate Analysis

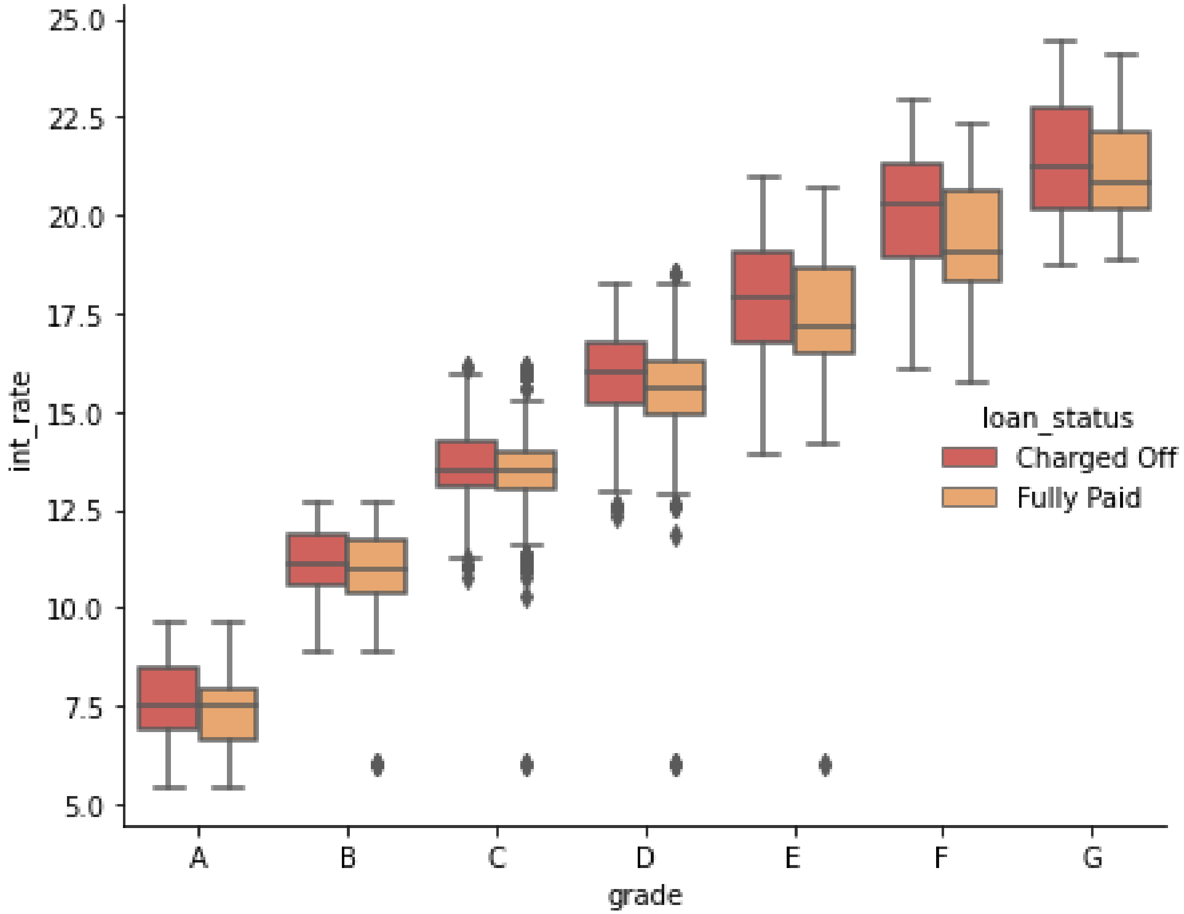
Observations: After analyzing the above categorical columns with loan status, the **loan Charged Off/default** happens when:

1. Applicants belonging to higher income groups (MIG-SIG) taking large amounts of loans.
2. As Applicants belonging to Superior Income geoups take large amounts of loans, their monthly installment is also high.
3. As Loans corresponding to VLIG and LIG have higher DTI ratio.
4. Applicants belonging to higher income groups with more lines of credit defaulted more.
5. Applicants belonging to MIG having more than 10 open lines of credit defaulted more.

Bivariate Analysis



Bivariate Analysis

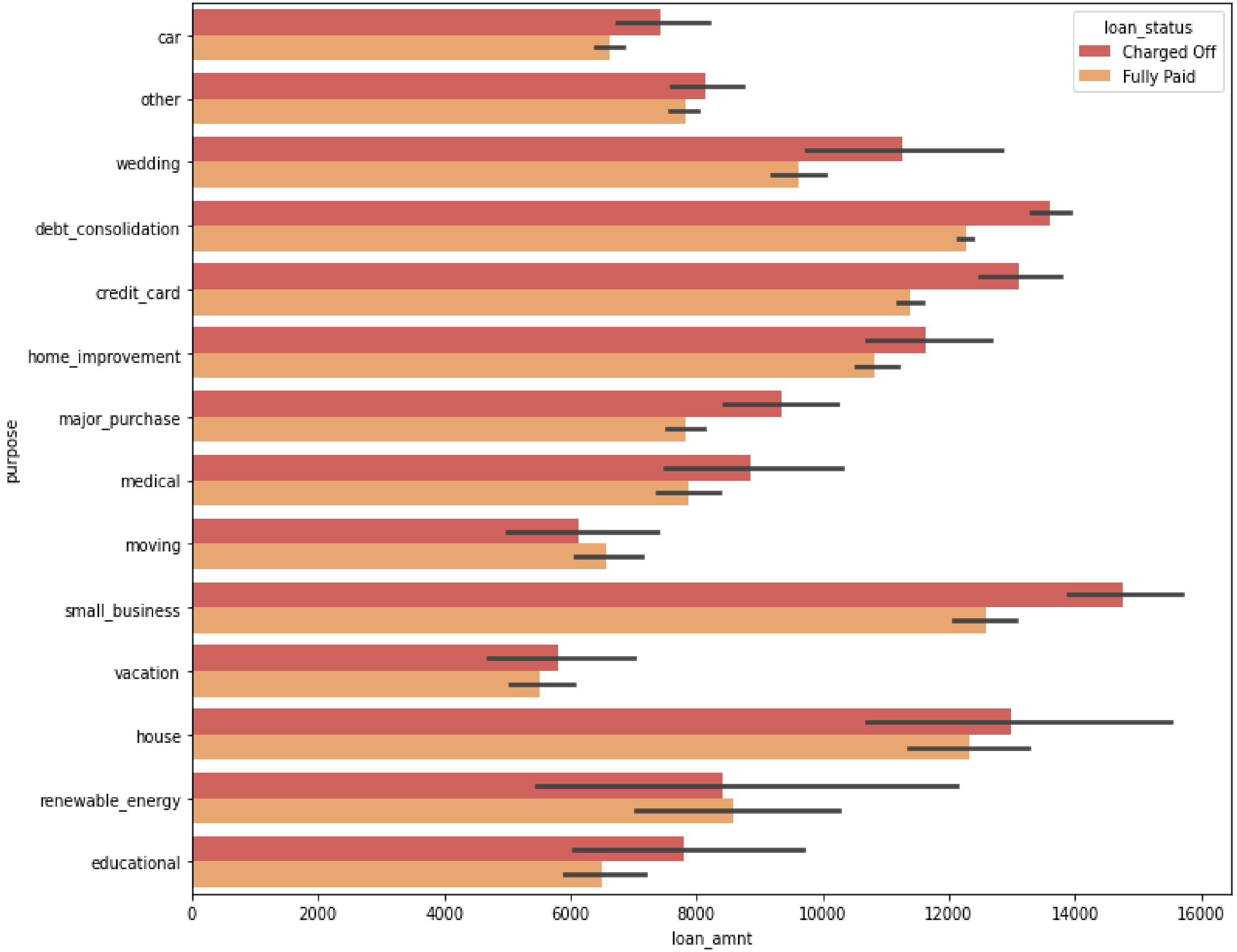


Observation:

The loan Charged Off/default happens when:

- Applicants belonging to lower grade (F and G) with higher loan amount.
- Interest Rates are higher for applicants belonging to lower grade (F and G)

Bivariate Analysis

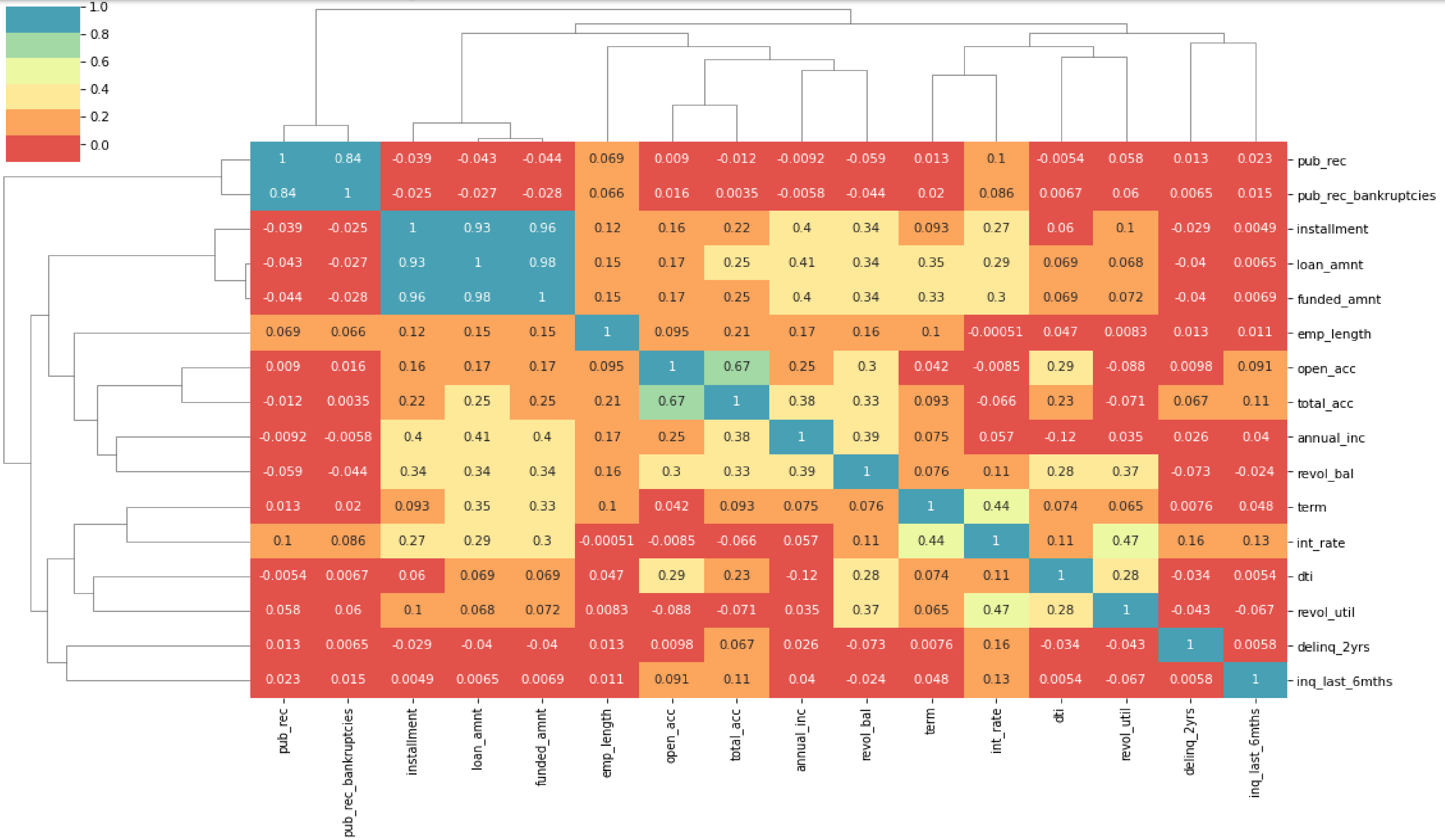


Observation:

The loan Charged Off/default happens when:

- Small Business and Debt Consolidation with higher average loan amount

Correlation Analysis



Correlation Analysis

Observation:

- Correlation between Interest rate and revolving line utilization rate is 0.47, which indicates that Lending club is charging higher interest rates for riskier loan applicant

- Correlation between loan amount and credit Revolving Balance is 0.35, which indicates that lending club is approving higher amount loans for applicants with higher credit revolving balance (Risk)

Recommendations

Recommendations:

- Stop – approving loans where Loan amount/income is higher
- Reduce – number of approvals where purpose is Debt Consolidation and small business or increase interest rates
- Stop – approving loans approving higher amount loans for applicants with higher credit revolving balance (Risk)
- Reduce-approving loans to people with employment experience greater than 10 years
- Reduce-approving loans to US government based organizations
- Stop- approving loans to people with higher number of inquires (low credit score) for higher loan amounts