Lending Club Case Study

Problem Statement:

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. the company wants 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.

Targeted Business Outcome:

The aim is to identify patterns which indicate if a person is likely to default, 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.

Given data and insights:

1. The data about past loan applicants and whether they 'defaulted' or not. The complete loan data for all loans issued through the time period 2007 to 2011.
2. The Loan accepted on 3 scenarios 1. Fully Paid, 2. Current, 3. Charged-off
3. The dictionary to understand the variables

Approach:

1. Understand the data set
2. Data preparation (cleaning, standardizing and treating outliers)
3. Data Analysis
4. Conclusion

1. Understanding the data set

1.1 There are no headers are footers in the given data set
1.2 There are 39717 rows and 111 columns in total
1.3 Out of 111 columns there are 54 columns are all null values
1.4 Out of 111 columns there are 9 columns which has same values throughout
➔ ***pymnt_plan, initial_list_status, collections_12_mths_ex_med, policy_code, application_type, acc_now_delinq, chargeoff_within_12_mths, delinq_amnt, tax_liens***

1.5 Out of 111 Columns there are 3 columns which has all values unique
➔ ***Id, member_id, url***

1.6 From observing the Dictionary, it's clear that we have customer behaviour variables, which were not available at the time of loan application
➔ ***pub_rec, delinq_2yrs, earliest_cr_line, inq_last_6months, open_acc,r evol_bal, revol_util, total_acc, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp,total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, last_credit_pull_d, application_type***

1.7 There are some irrelevant columns which will not contribute to our analysis
➔ ***emp_title, desc, title, zip_code, addr_state, chargeoff_within_12_mths, mths_since_last_record, mths_since_last_delinq'***

1.8 ***funded_amnt, funded_amnt_inv, loan_amnt*** has same distribution
1.9 We have ***grade and sub_grade*** , sub_grade is sub category of the variable grade

2. Data Preparation

2.1 Data Cleaning:

1. Dropped off **all the columns** which values are **all null**
2. Dropped off **all the columns** which values are **all unique**
3. Dropped off **all the columns** which values are **all same throughout**
4. Dropped off **all the columns** which are **not relevant** for our analysis
5. We will limit our analysis with variable "**grade**" hence dropped the column "**sub_grade**"
6. Dropped "**funded_amnt_inv**" and "**funded_amnt**" as we will consider "**loan_amnt"** which has similar distribution
7. Dropped the records where the ***loan_status ='Current'***, as this will not be useful for our analysis

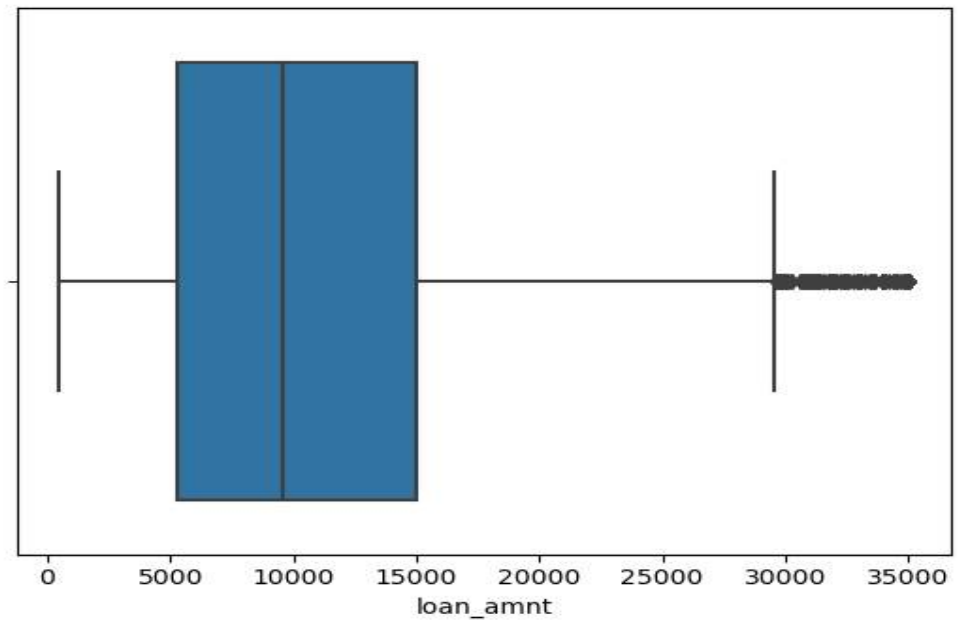
After cleaning the data, we have the relevant variables in our data set

➔ ***loan_amnt, term, int_rate, installment, grade, emp_length, home_ownership, annual_inc verification_status, issue_d, loan_status, purpose, dti***

2.2 Data Standardizing

1. Removed "%" from "***int_rate***" value and converting the values to float
2. Removed "years" and "year" from "***emp_length***" for better visualization
3. Derived two variable "***issue_month***", "***issue_year***" from "***issue_d"***
4. Imputing the value for "***home_ownership***" variable where value is "**NONE"** and assume that to "**OTHER"**
5. Assigning the Mode value to null values in "**emp_length"** as the **mode 10+years** value has far higher frequency than that of the next most frequent value. It is safe to assign the mode value.
6. As variable "***term***" has only two values 30 months and 60 months, we will keep them as it
7. we will keep "***emp_length***" as it's is and treat them as categorical variable, and 10+ years can be any practical number
8. Handling Outliers of numeric variable in our current data set

A. Loan Amount (loan_amnt) outliers

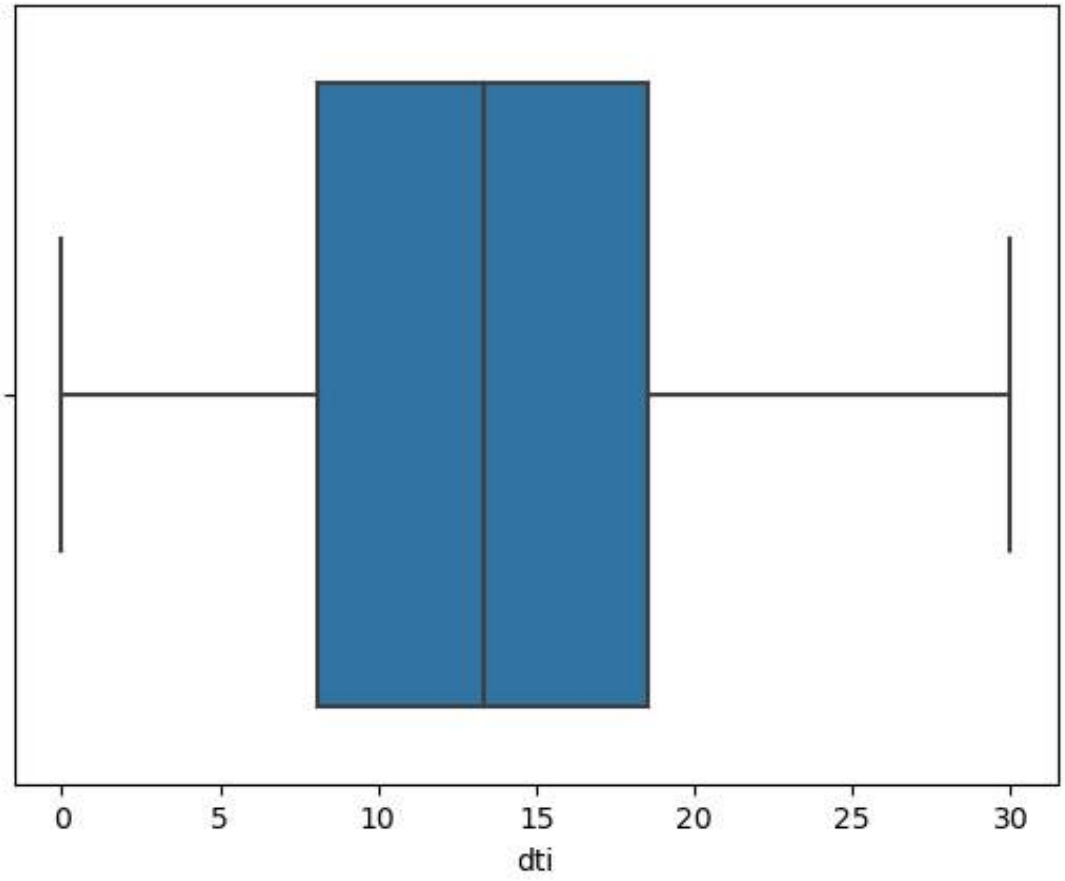


Quantile info - Loan Amount

0.25 5300.0
0.50 9600.0
0.75 15000.0
0.90 21600.0
0.95 25000.0
0.97 28000.0
0.98 30000.0
0.99 35000.0

There are clearly outliers, observing the quantile information of "***loan_amnt". There is a sudden spike at 0.99 quantile hence we will limit our data with 0.98 quantile***

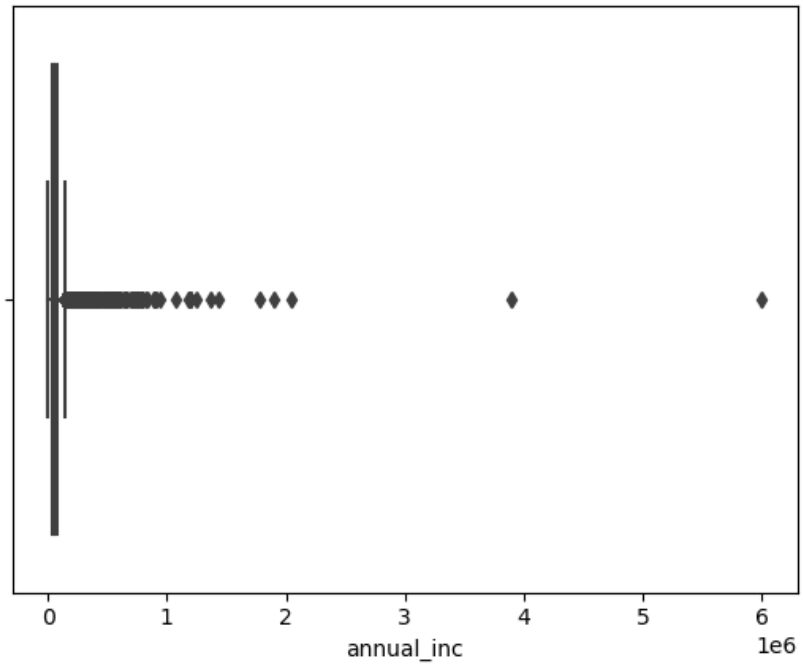
B. Treating Debt to Income ratio("dti") outliers



Quantile info - debt to income (dti)
0.50 13.3600
0.75 18.5600
0.90 22.2900
0.95 23.8000
0.97 24.5000
0.98 24.8400
0.99 26.5259

The variable "*dti*" (debt to income) does not have any outliers, we will keep the variable as it is.

C. Treating Annual Income Outliers (annual_inc)



Quantile info- Annual income

0.25 40000.00
0.50 57814.42
0.75 80000.00
0.90 113000.00
0.95 140000.00
0.97 160000.00
0.98 180000.00
0.99 225000.00

The value after 0.95 quantile seems to be disconnected from the distribution. Hence, we will restrict our data set to 0.95 quantile of Annual income to avoid the irrelevant outlier data

There are 36103 rows and, 15 columns as listed below, these are the most relevant data for our analysis

Important relevant variables are listed below

- ➔ ***loan_amnt,term, int_rate, installment, grade, emp_length, home_ownership, annual_inc, verification_status, issue_d, loan_status, purpose, dti, issue_month, issue_year.***

We will do Univariate and Bivariate analysis on our current data set

From Observing the current data set,

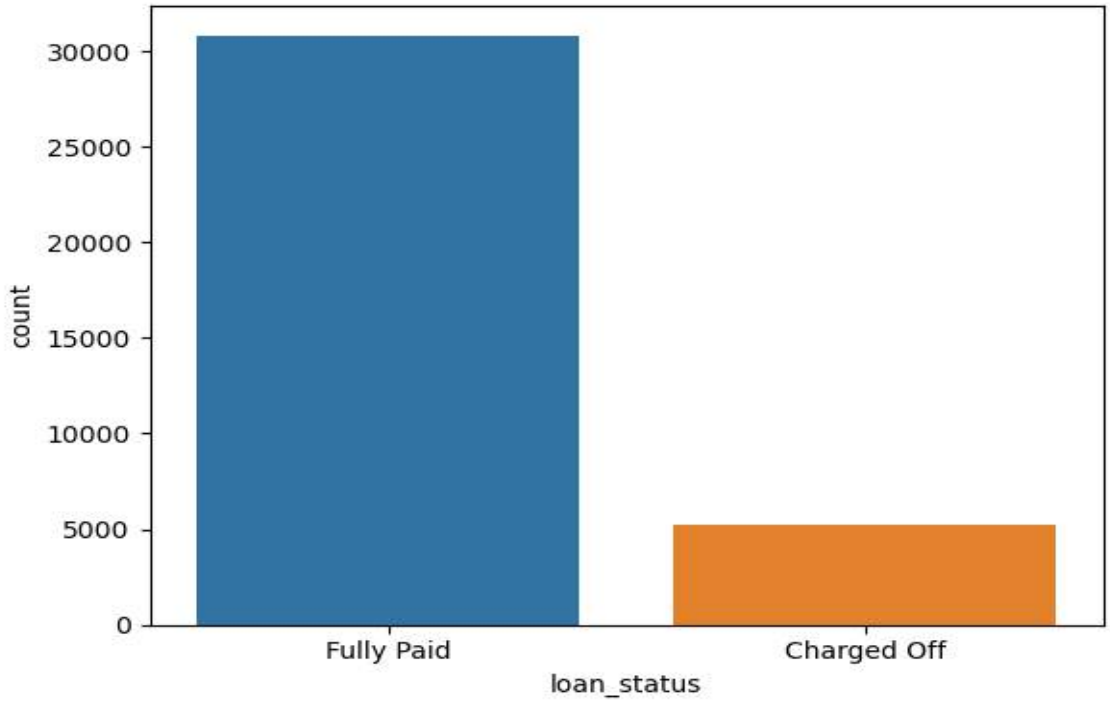
Loan status
Fully Paid 30841
Charged Off 5262

The current ratio between Fully Paid and Charged off is approximately 6:1

Current data infers that for every 6 applicant 1 applicant is a defaulter

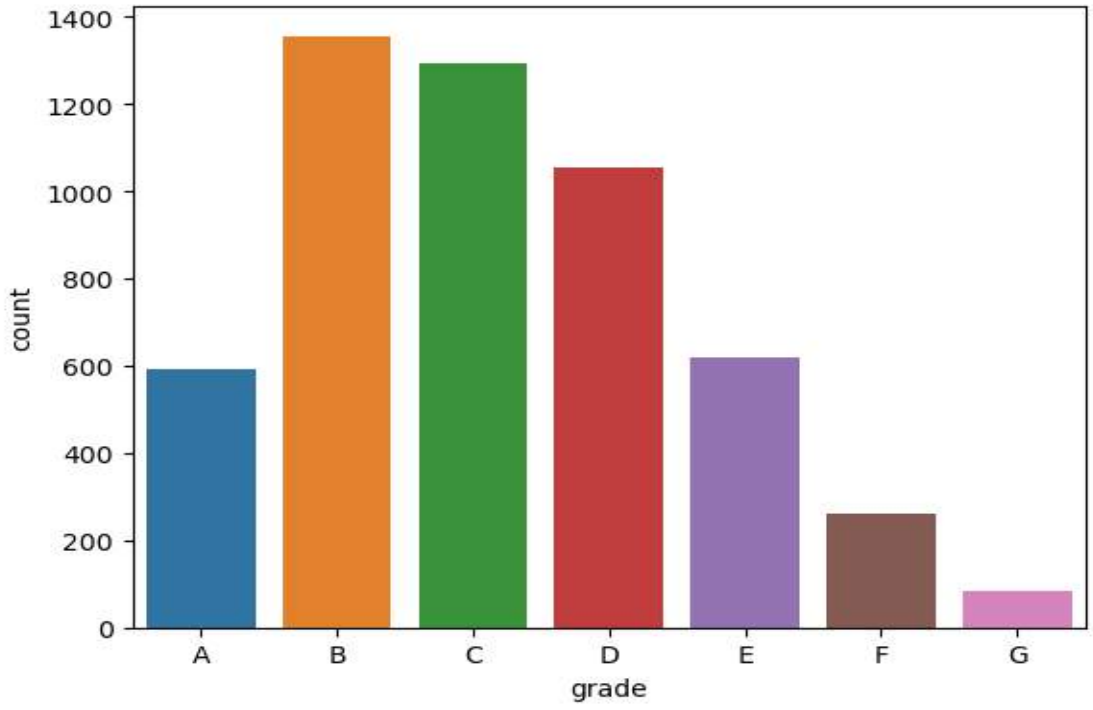
3. Data Analysis

We will first visualize the target variable "loan_status" from the current data

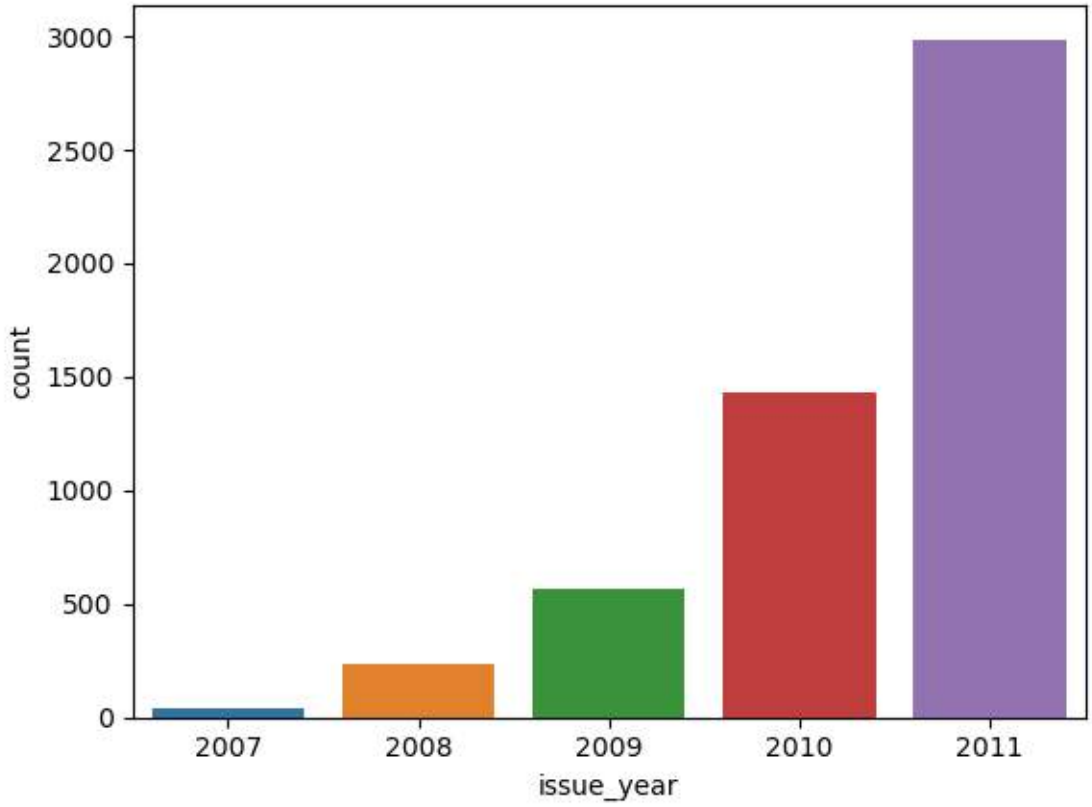


Univariate Analysis on the variables in the current data

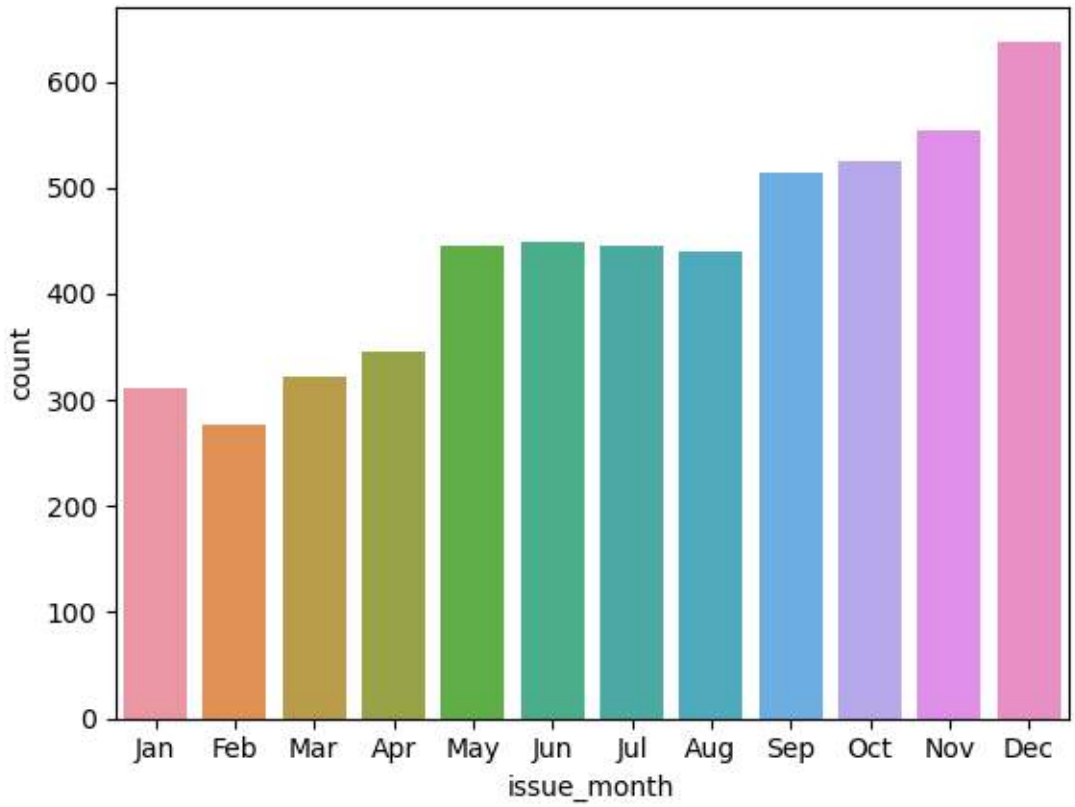
Analysing Grade



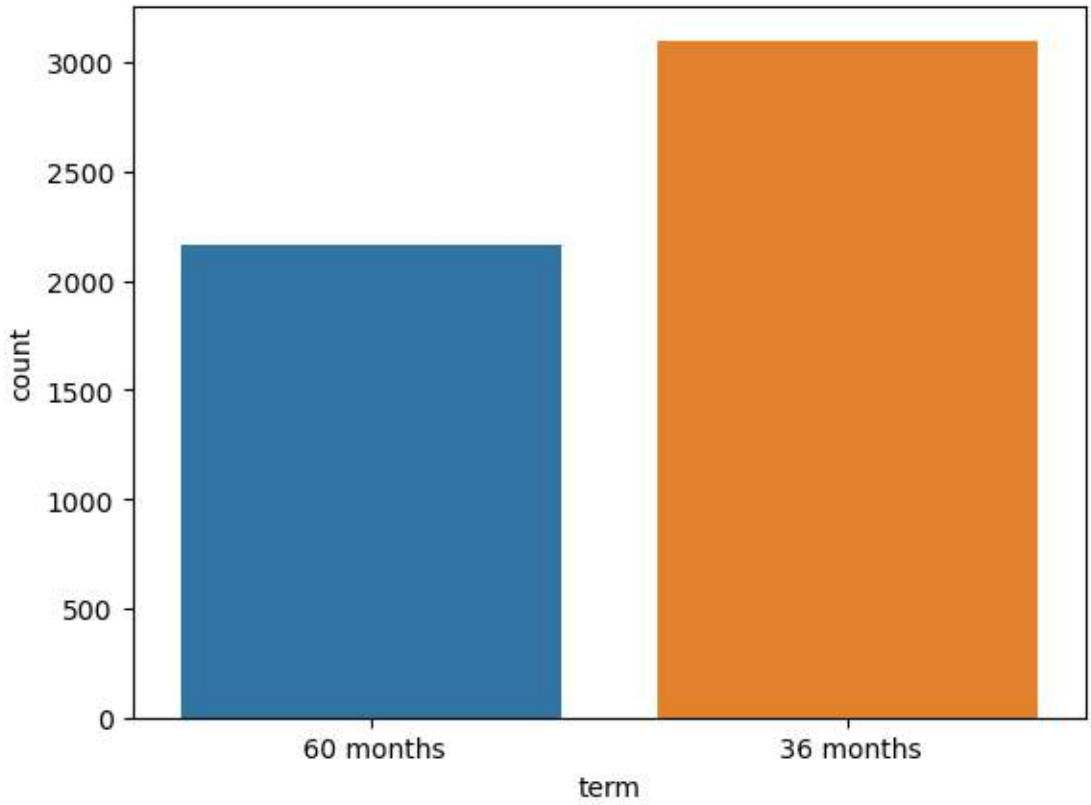
Analysing Issue Year



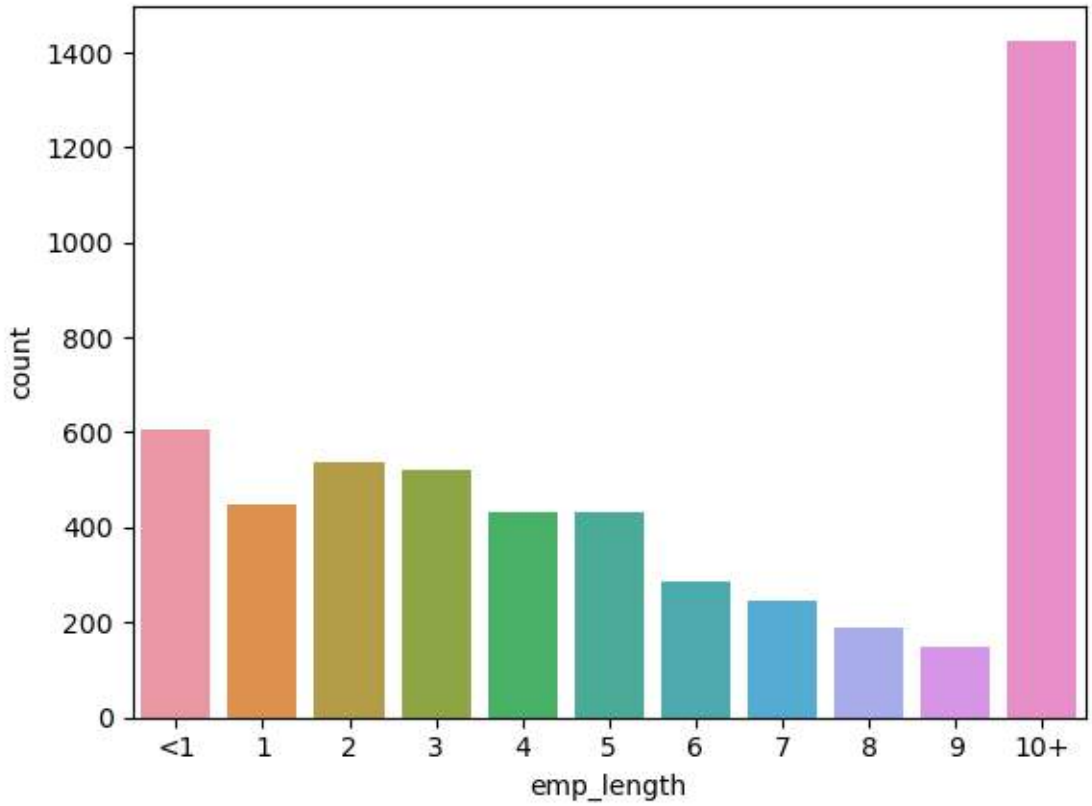
Analysing issue month



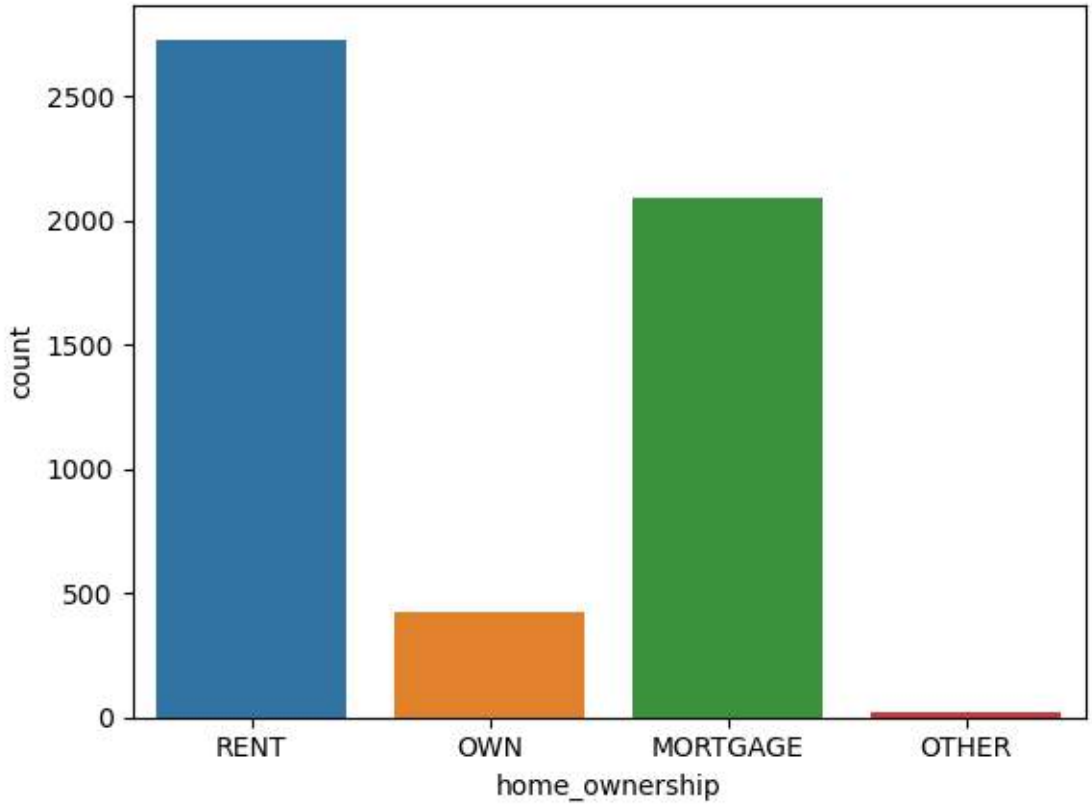
Analysing Term



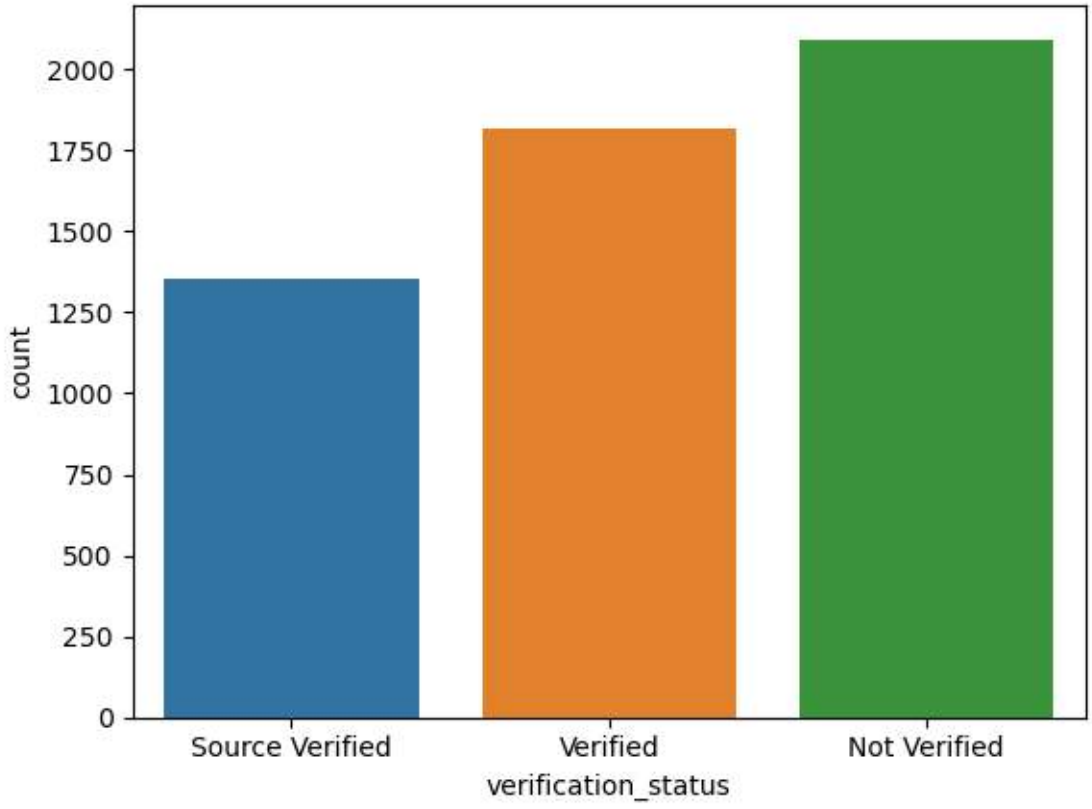
Analysing Employee Length



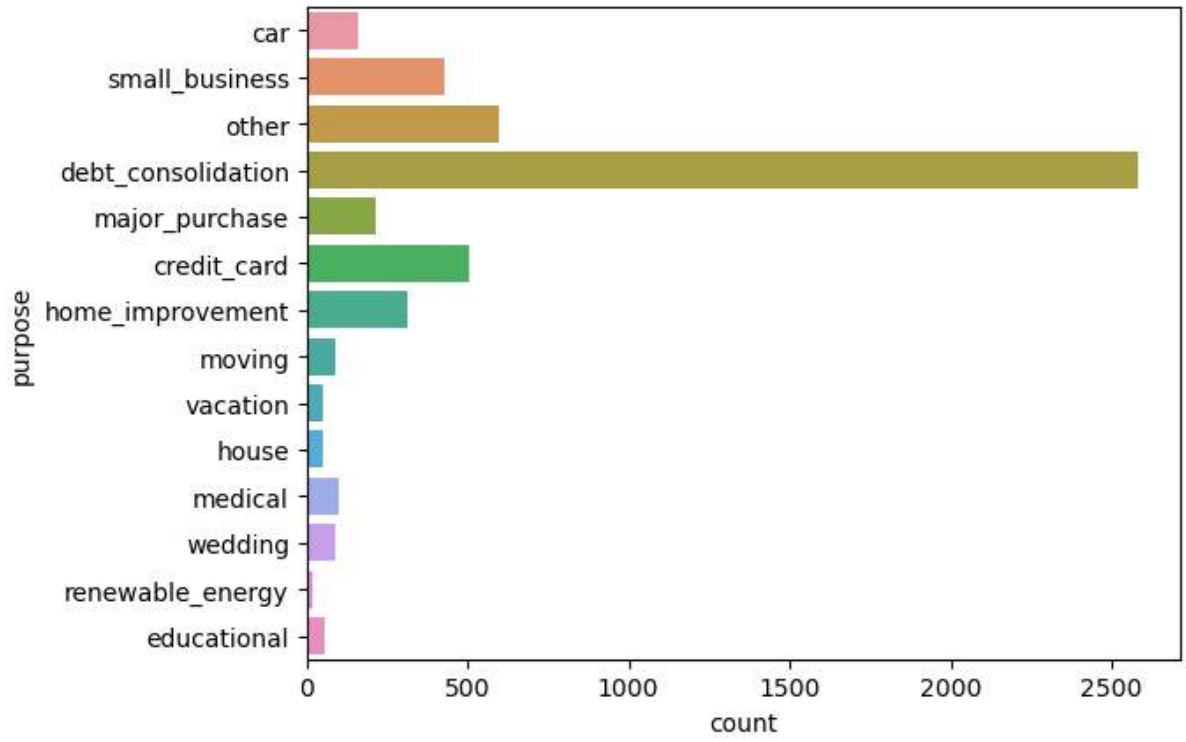
Analysing Home Ownership



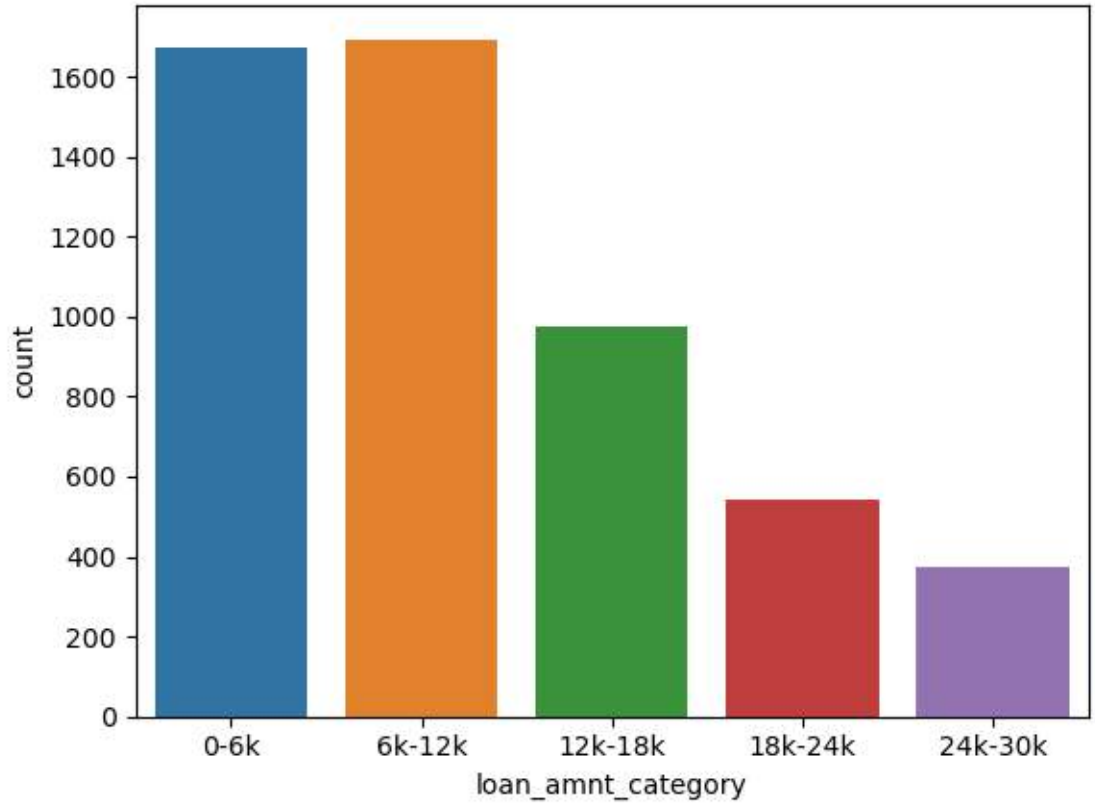
Analysing Verification Status



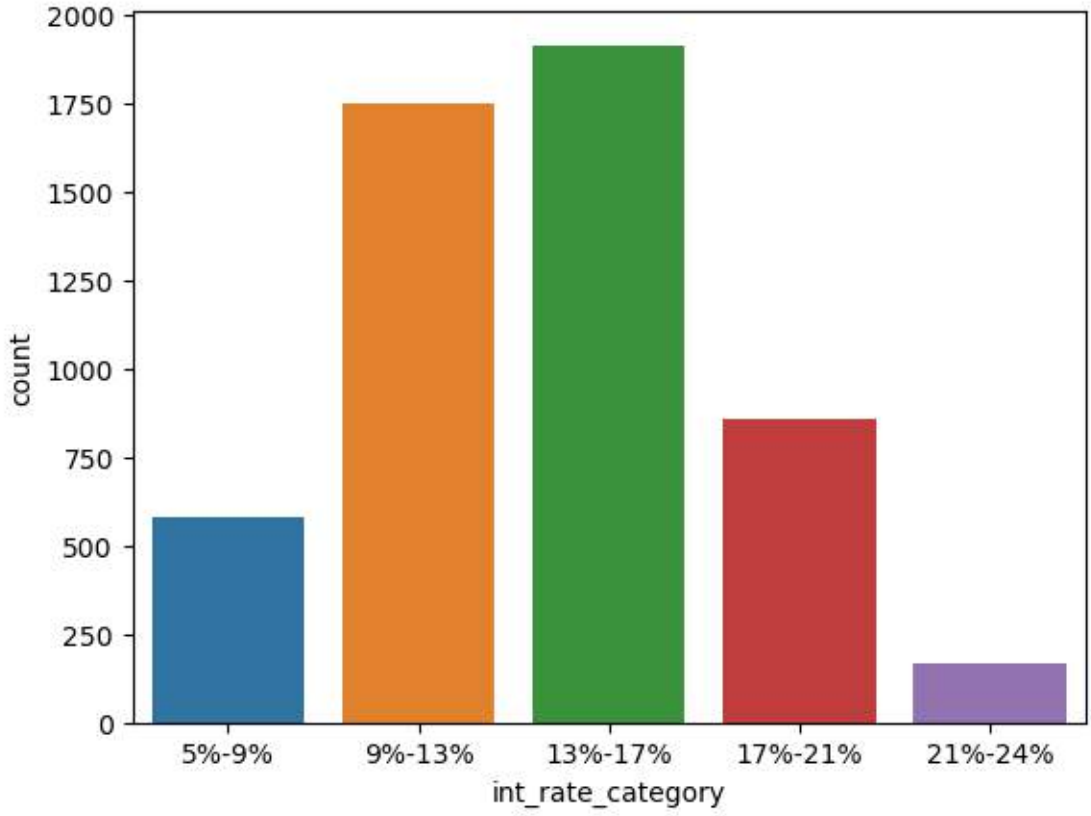
Analysing Purpose



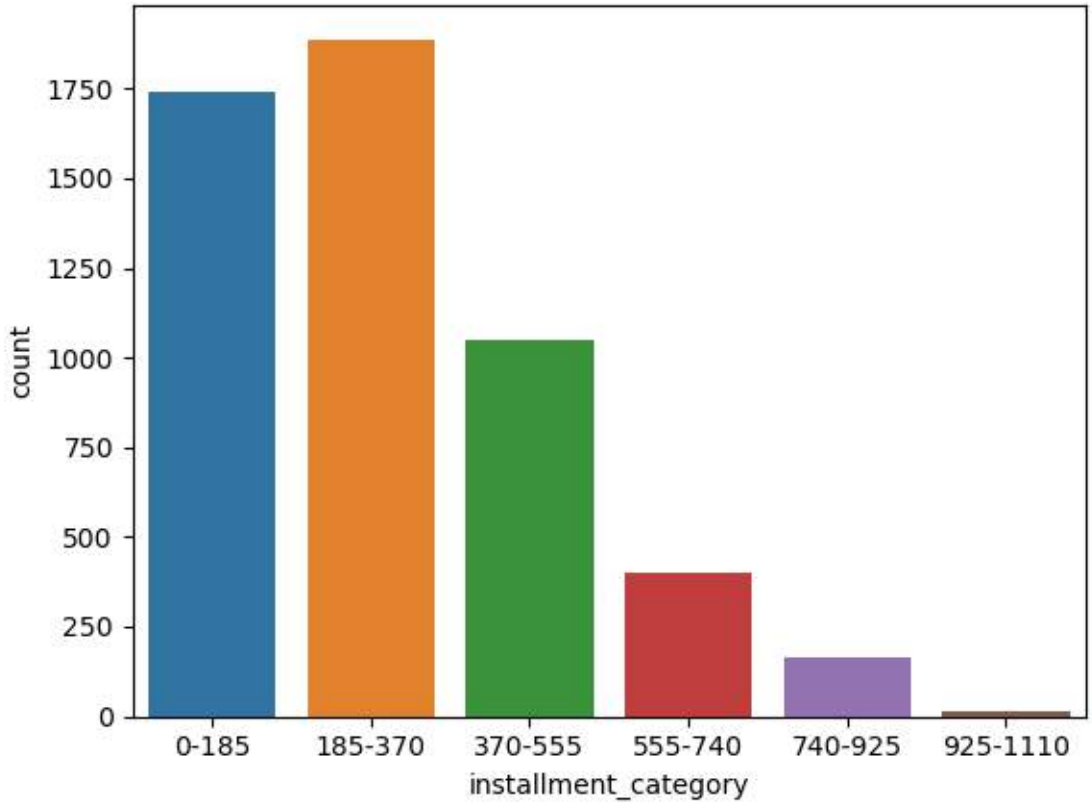
Analysing Loan Amount (Visualization loan_amnt as categorical variable)



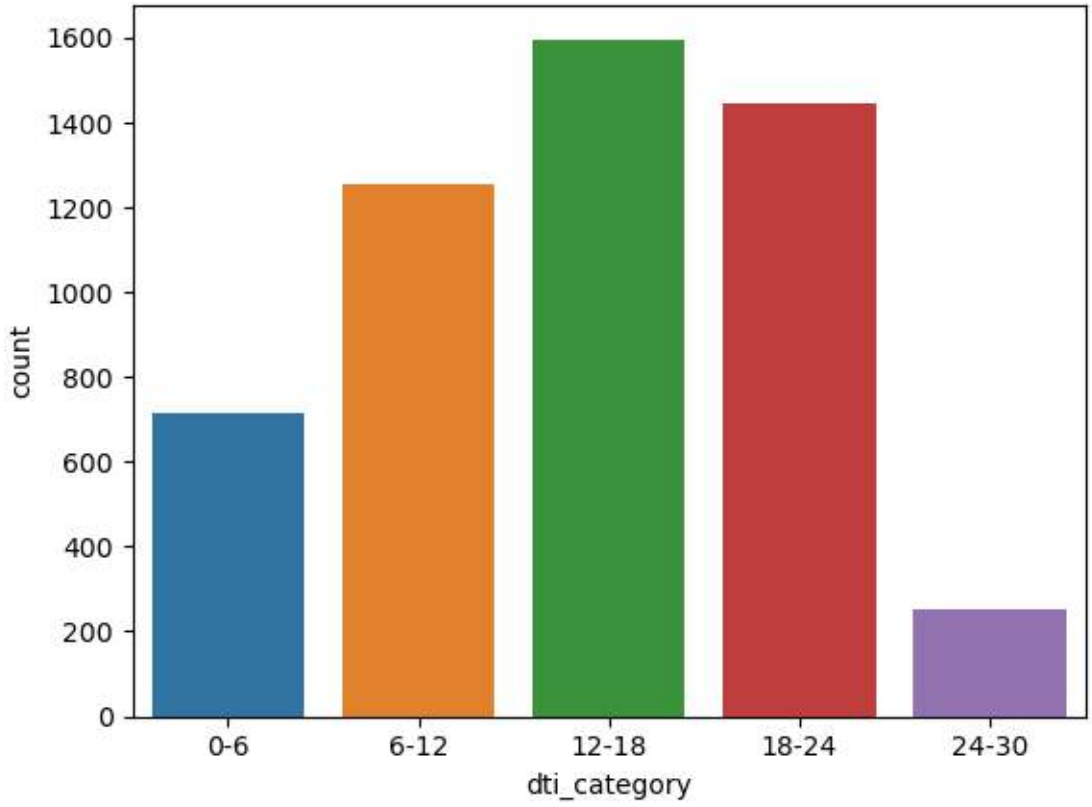
Analysing Interest Rate (Visualizing int_rate as categorical variable)



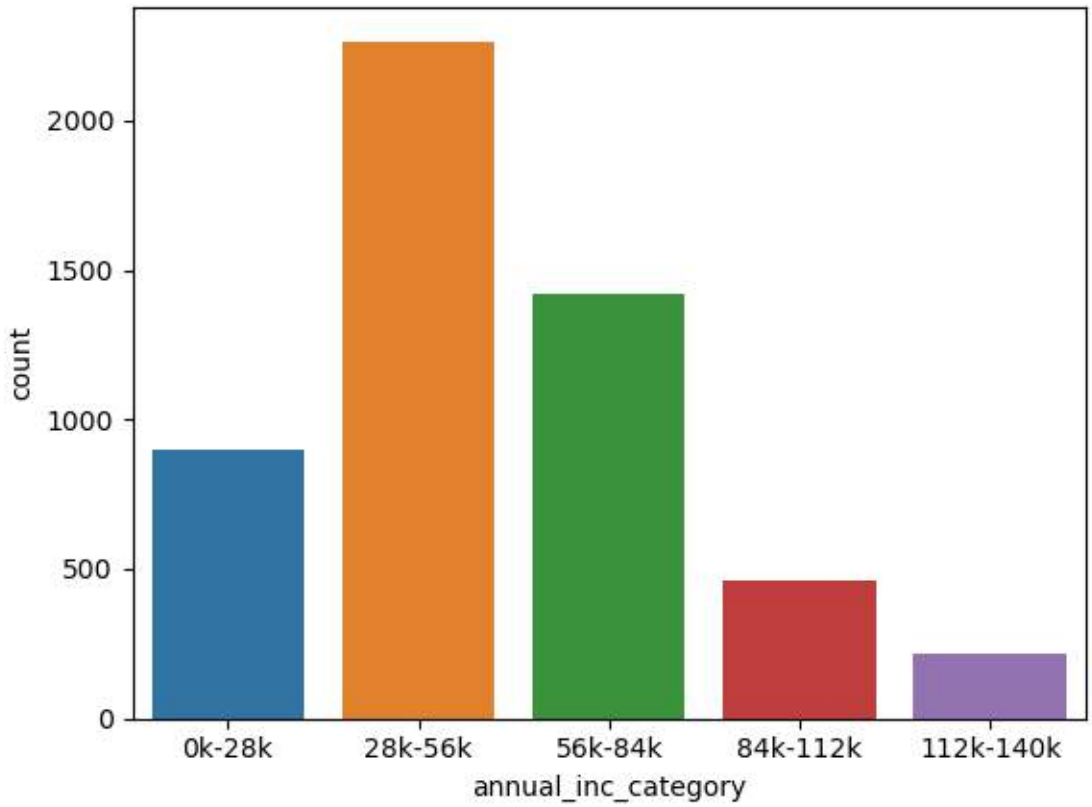
Analysing Instalment (Visualizing instalment as categorical variable)



Analysing Debt to Income (Visualizing dti as categorical variable)



Analysing Annual Income (Visualizing annual_inc as categorical variable)



Observation from univariate Analysis:

The Charge off is higher in the following scenarios:

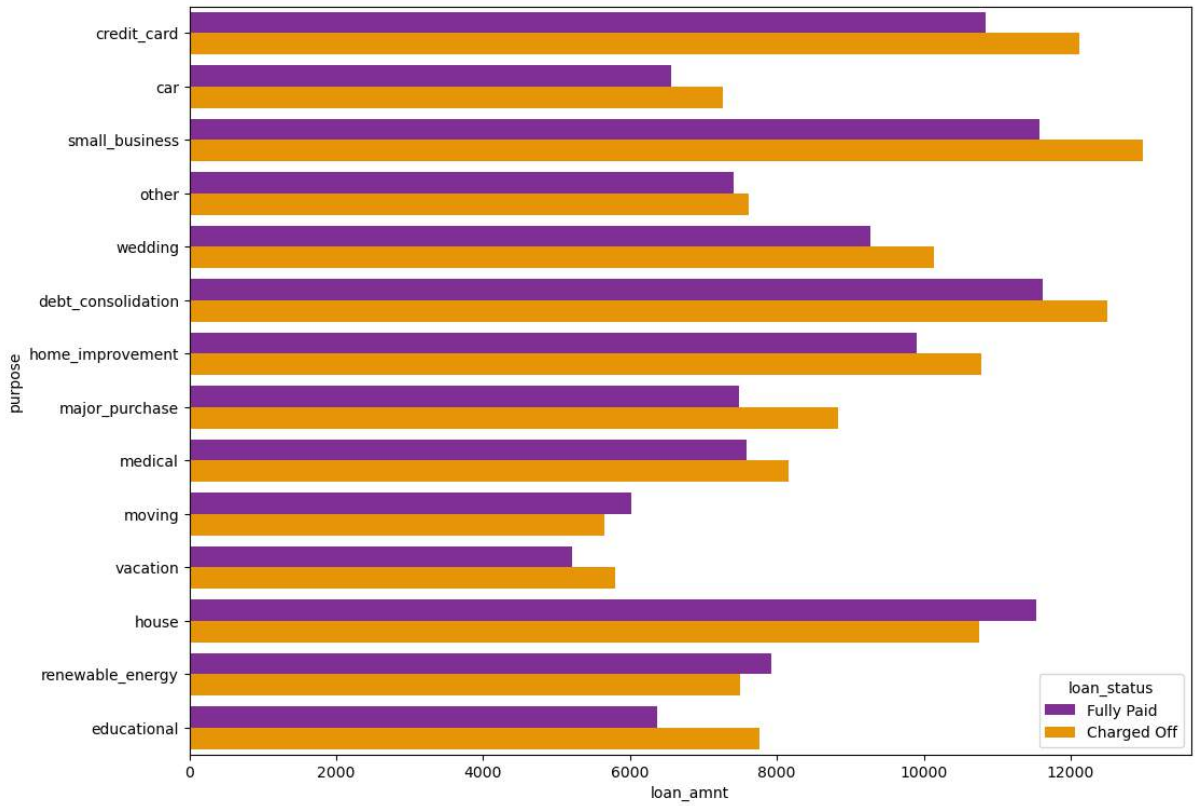
1. **Applicants with grade B**
2. **More charged off occurred in the year 2011**
3. **More charged off occurred in the month of December**
4. **Term with 36 months**
5. **Employment length more than 10 years charged off in high numbers, this could be a fact that all the applicants above 10 years are in the category 10+ years**
6. **Applicant who is residing in rented home**
7. **Applicants whose source are not verified**
8. **Applicants used loan for debt consolidation**
9. **Applicants with interest rate between 13% and 17%**
10. **Applicants with instalments between 185 and 370**
11. **Applicants with debt-to-income ratio between 12 and 18**
12. **Applicants with annual income between 28k and 56K are charged off in high numbers**

Further analysing the data combination

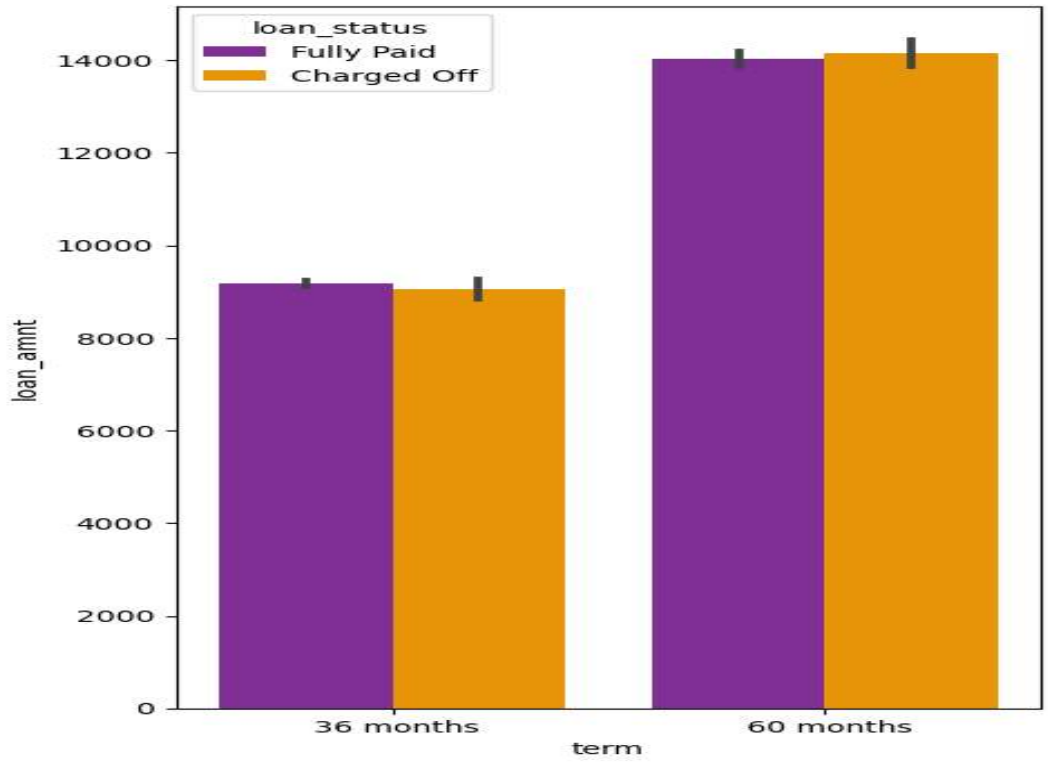
Bivariate Analysis:

Analysing loan amount with other variables against the target variable loan status

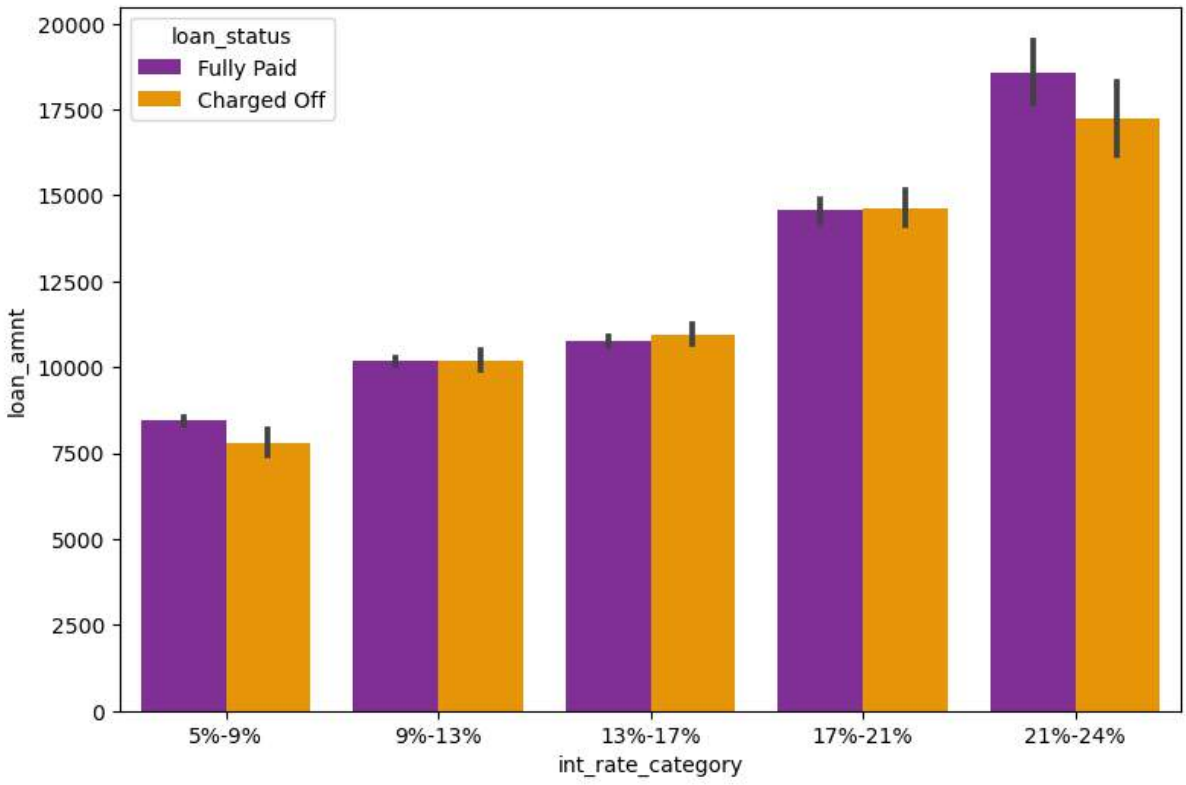
Loan Amount and Purpose



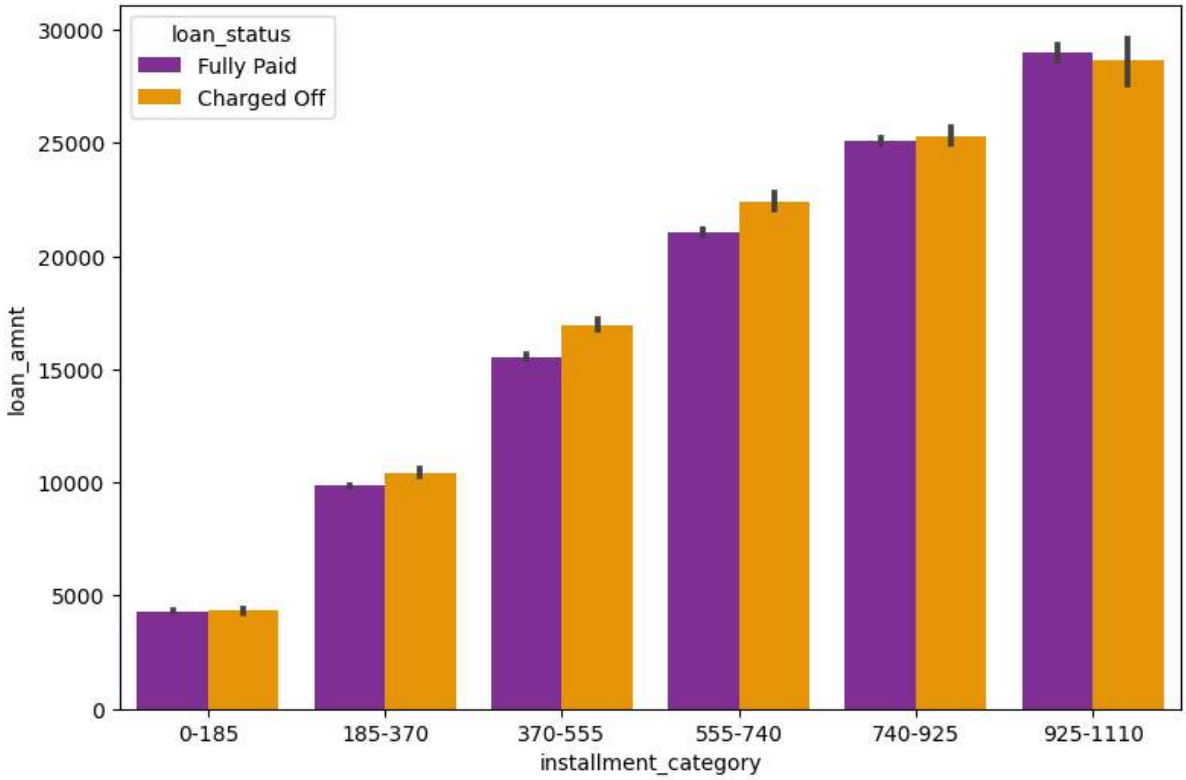
Loan Amount and Term



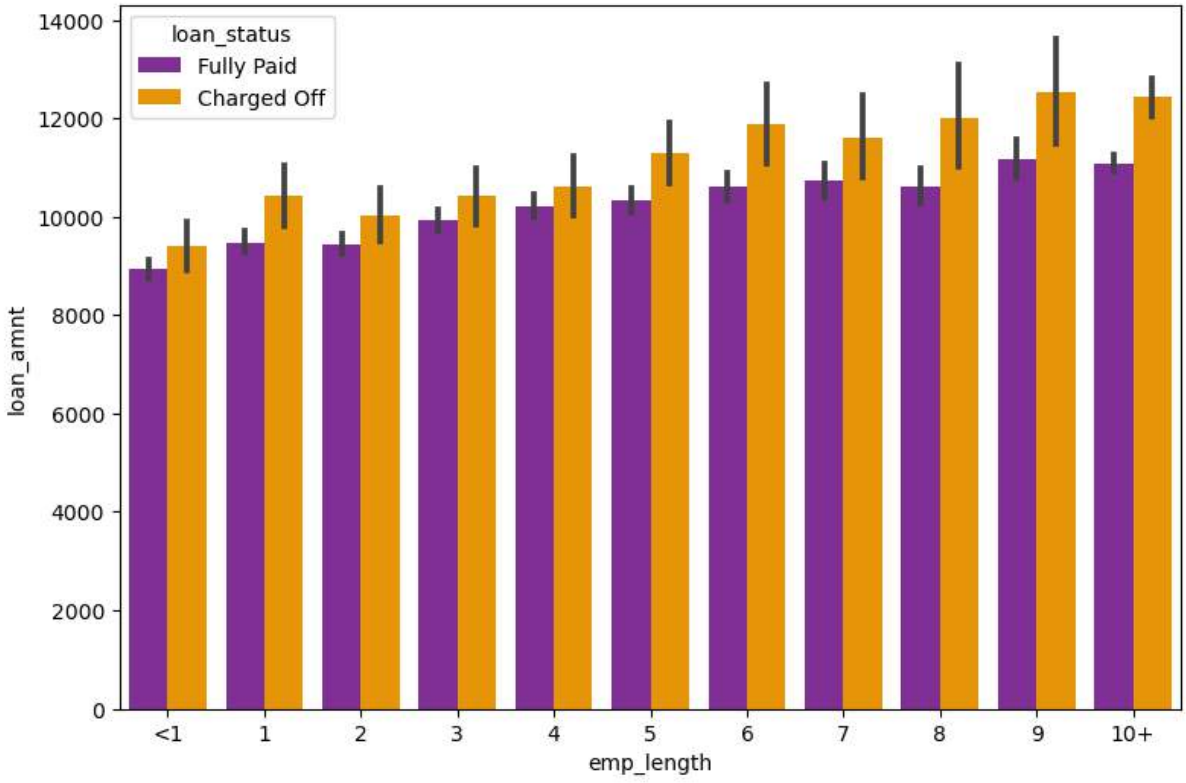
Loan Amount and Interest Rate



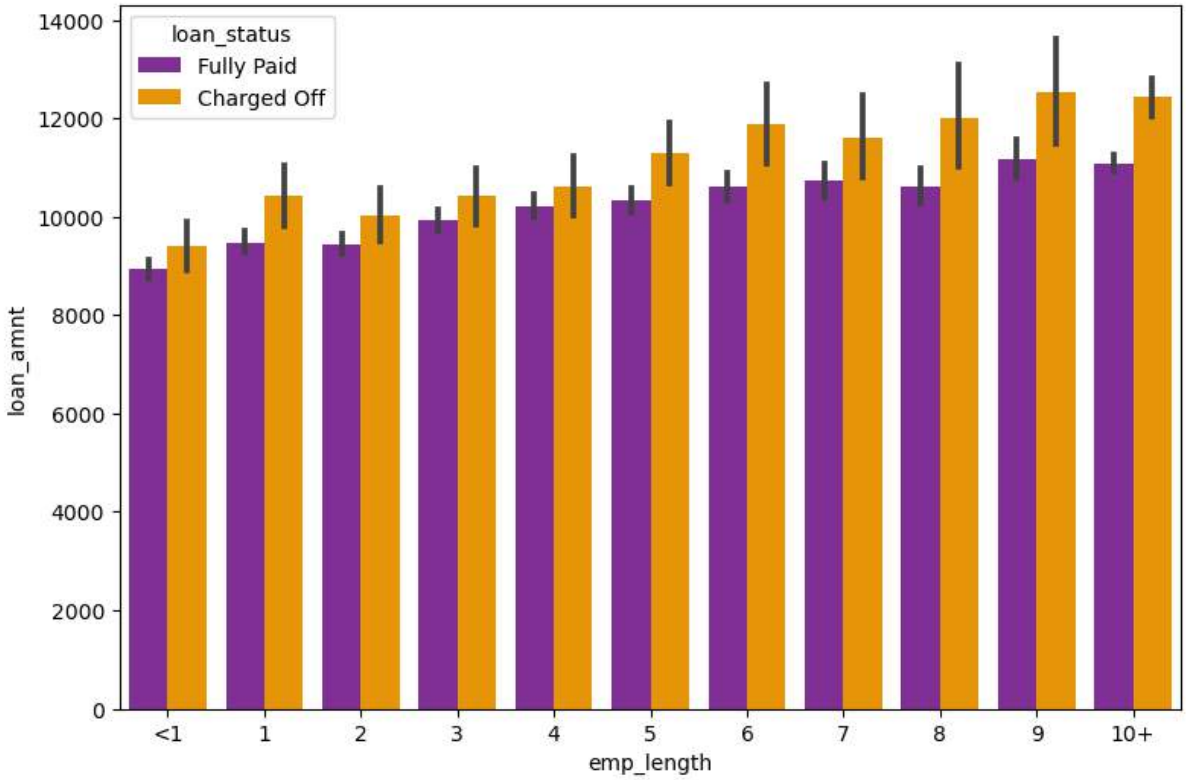
Loan Amount and Instalment



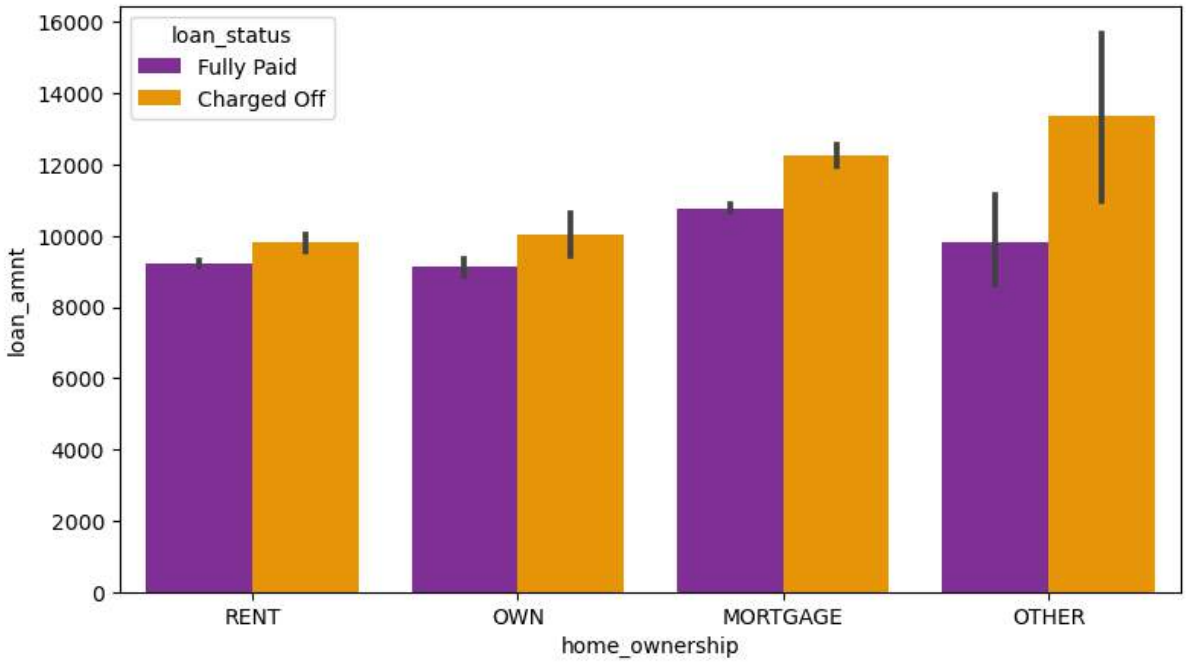
Loan Amount and Grade



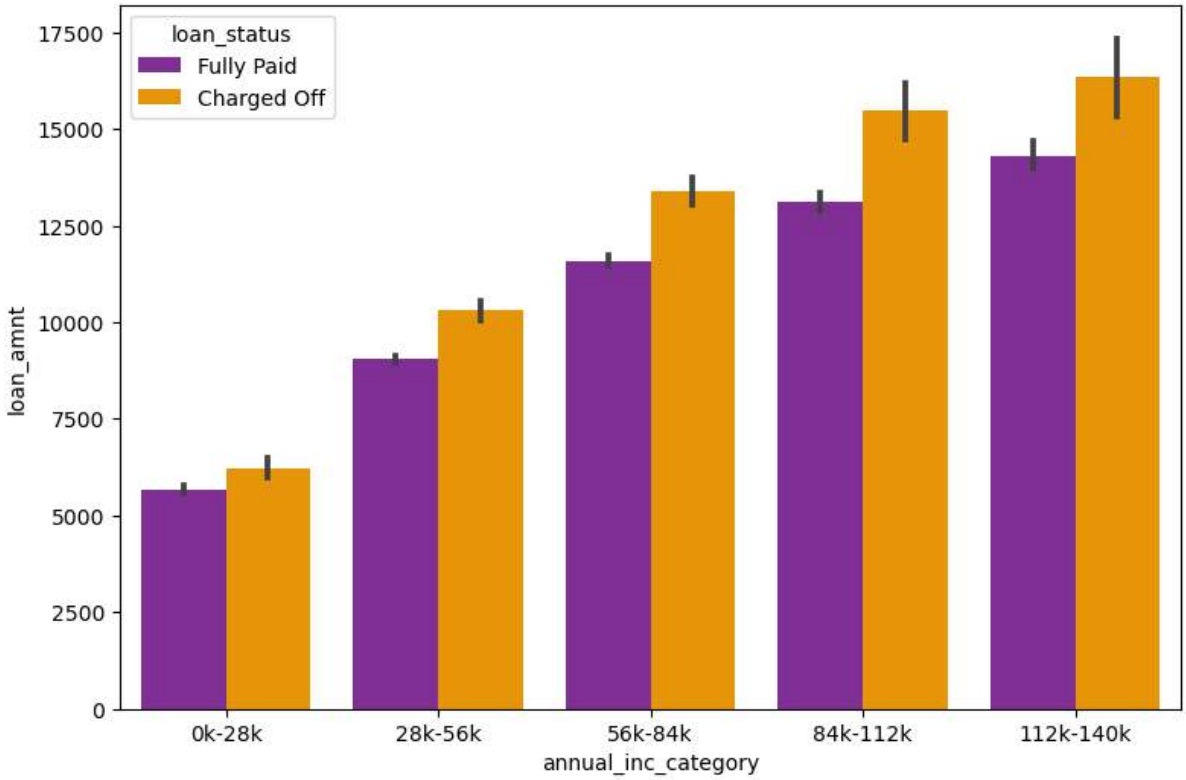
Loan Amount and Employee Length



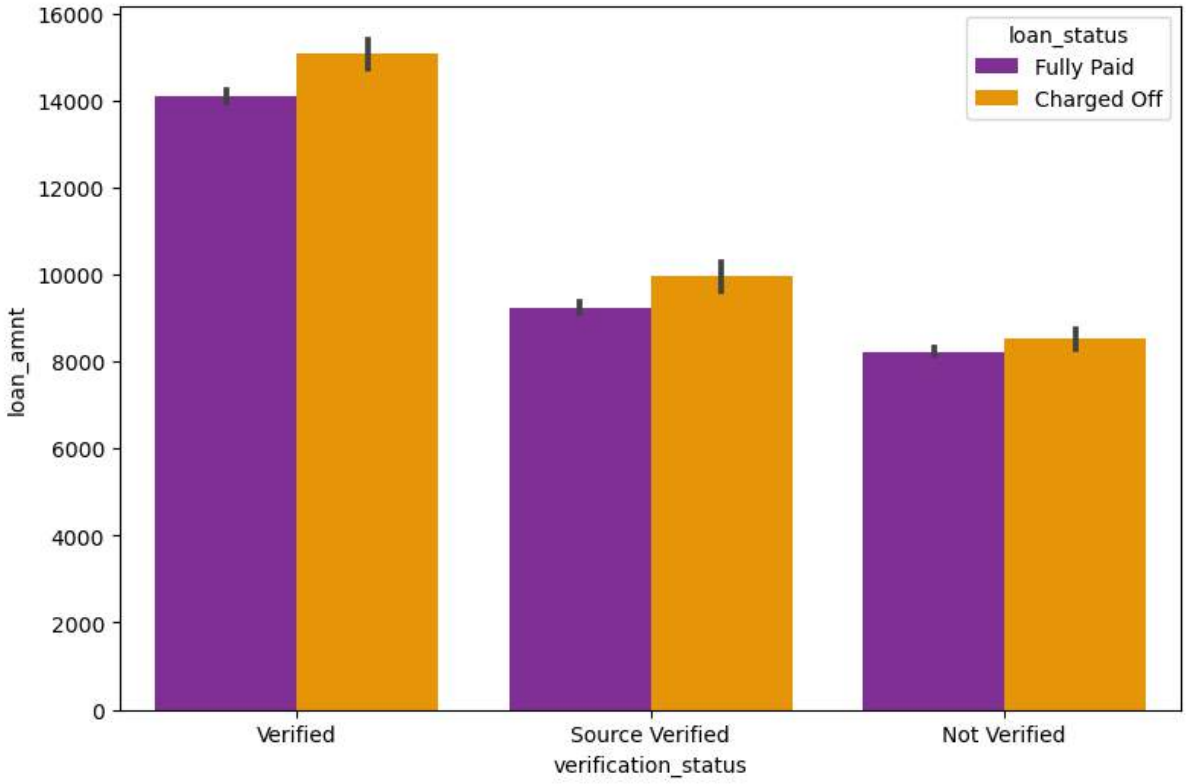
Loan Amount and Home Ownership



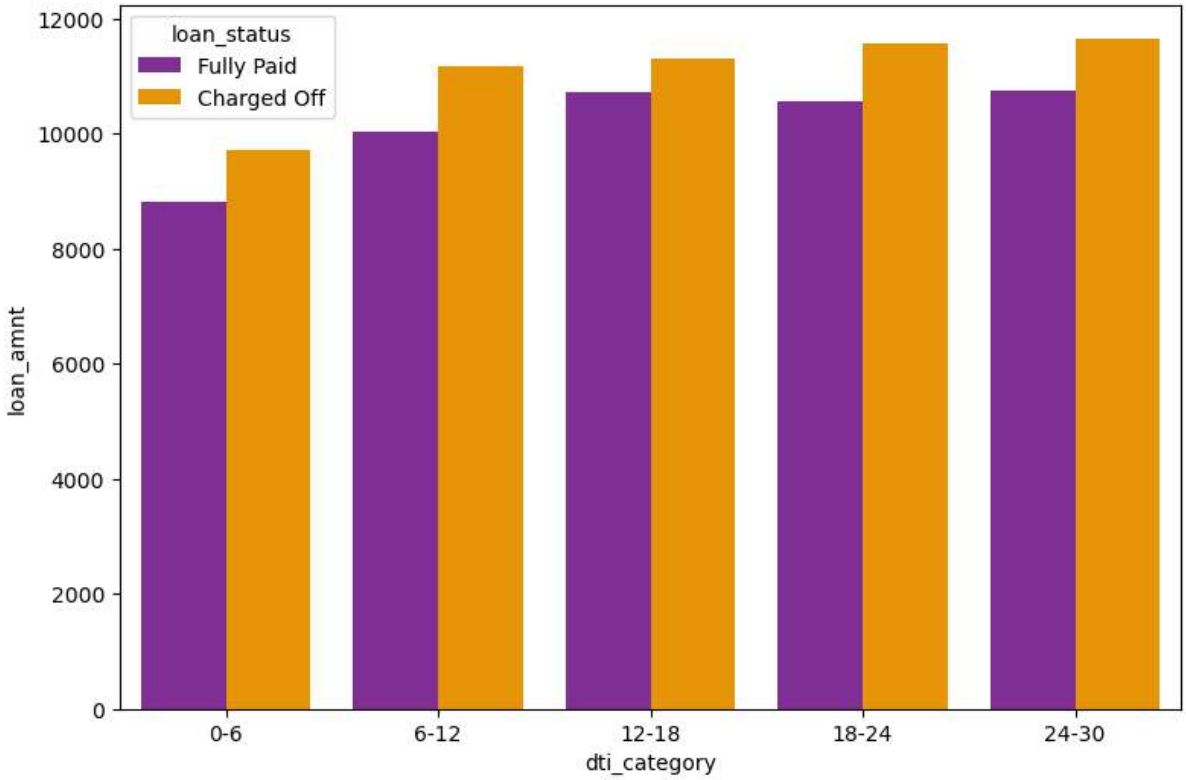
Loan Amount and Annual Income



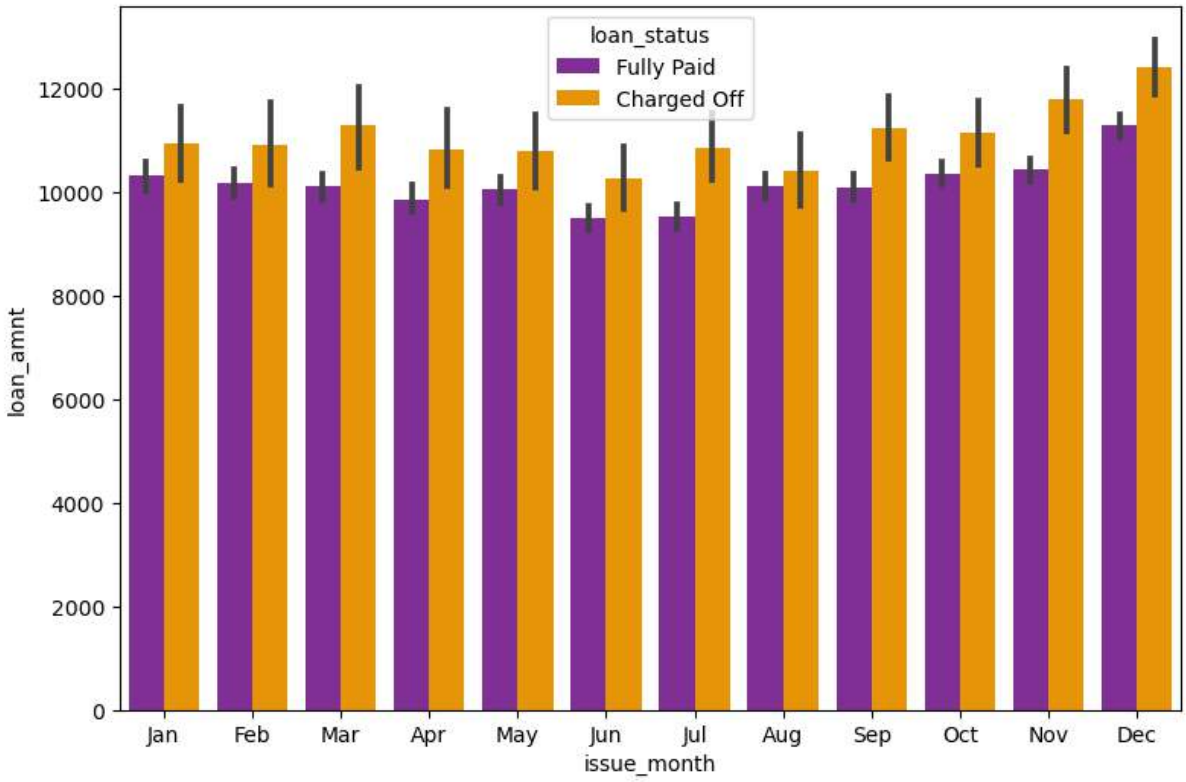
Loan Amount and Verification Status



Loan Amount and Debt-Income

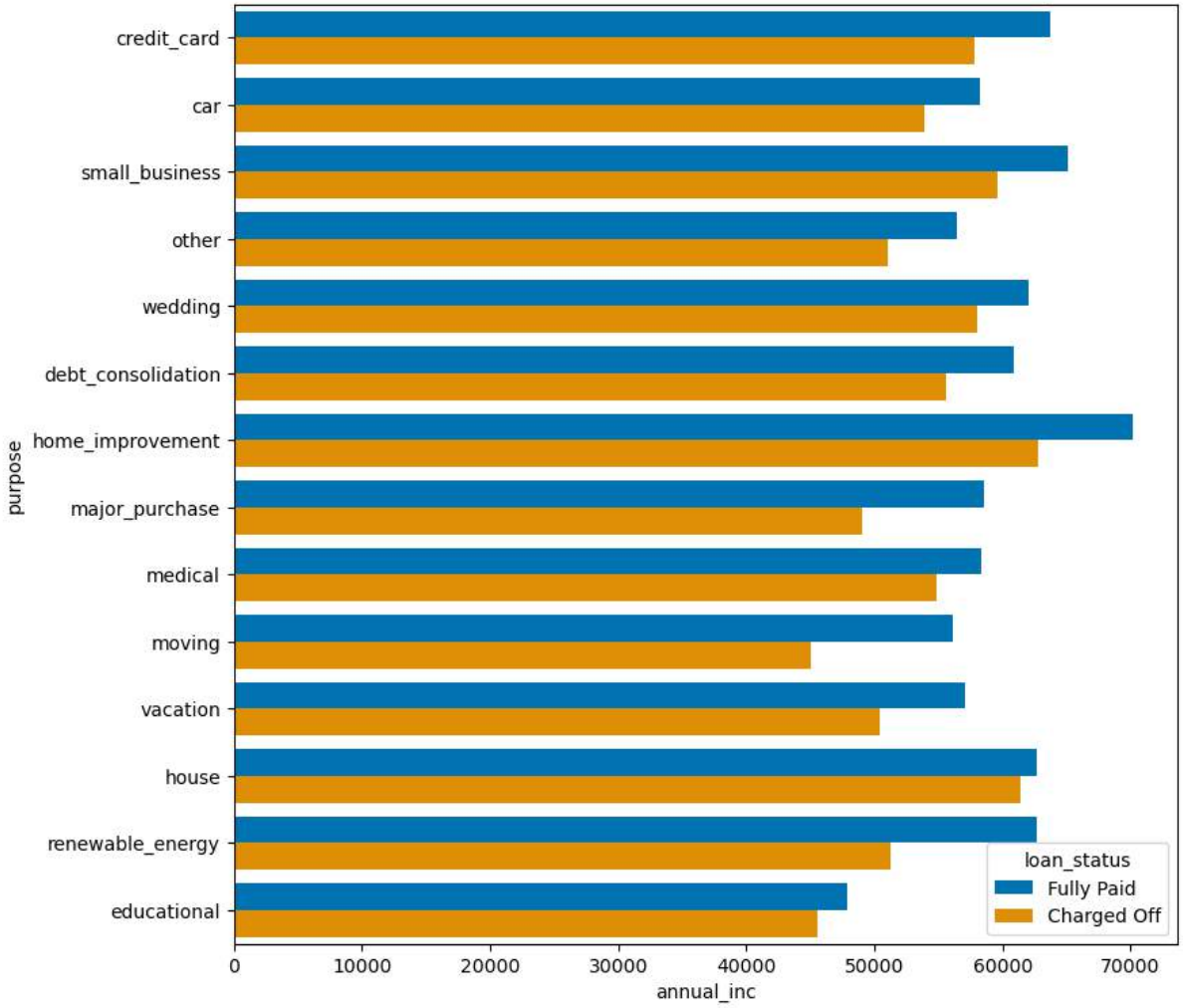


Loan Amount and Issue month

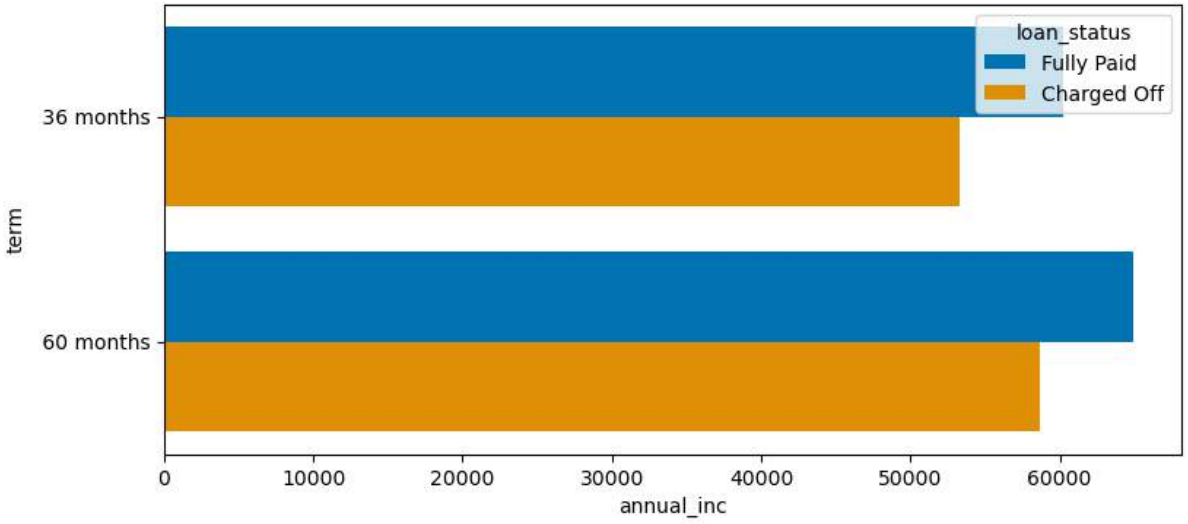


Analysing Annual Income with other variables against the target variable loan status

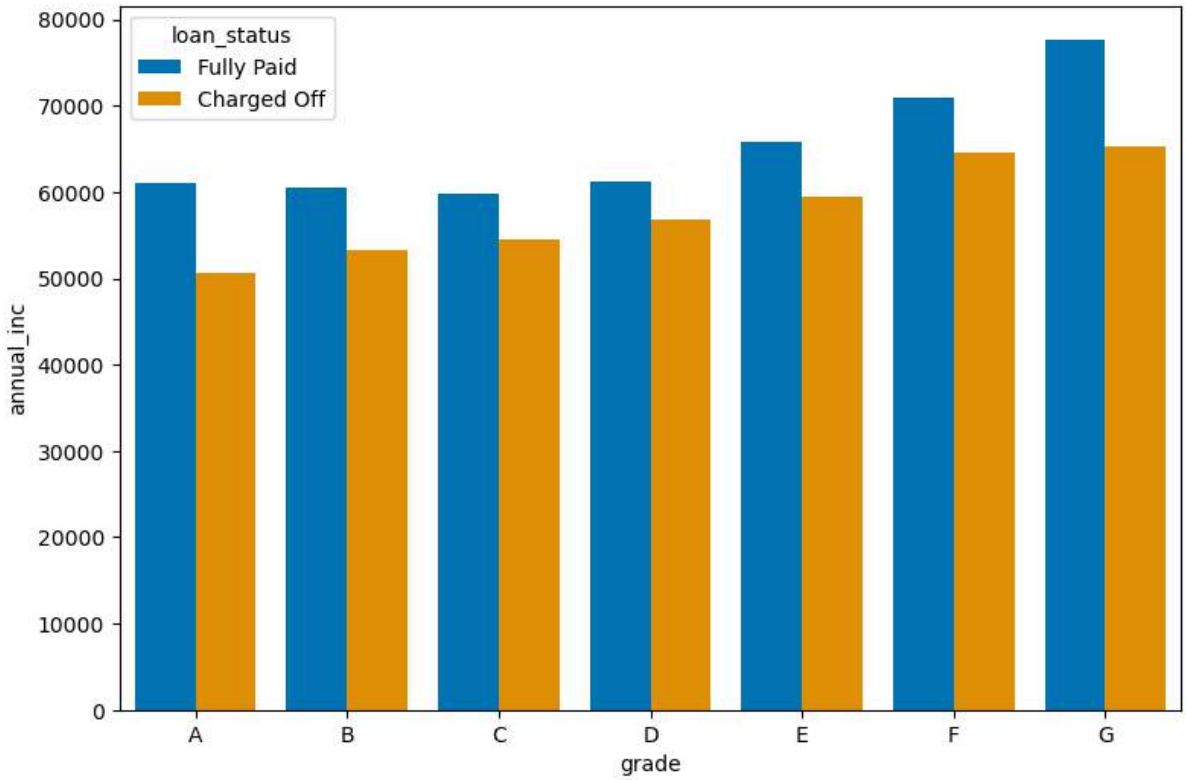
Annual Income and Purpose



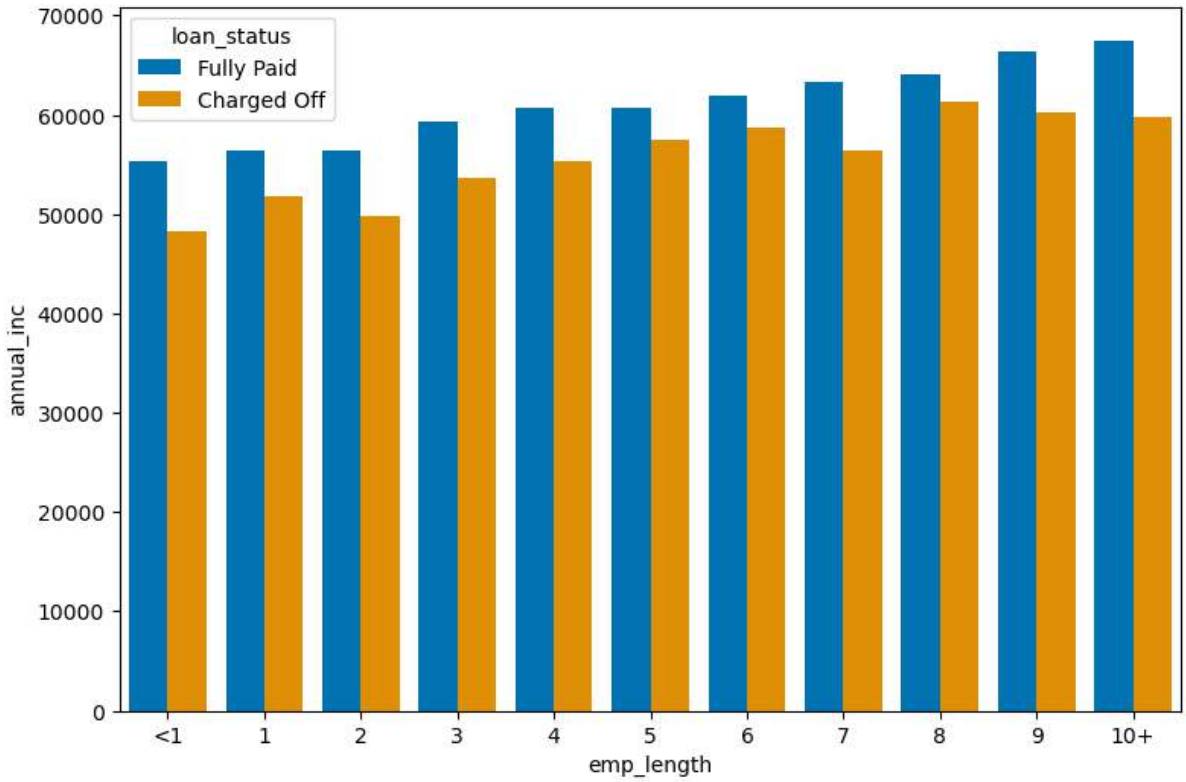
Annual Income and terms



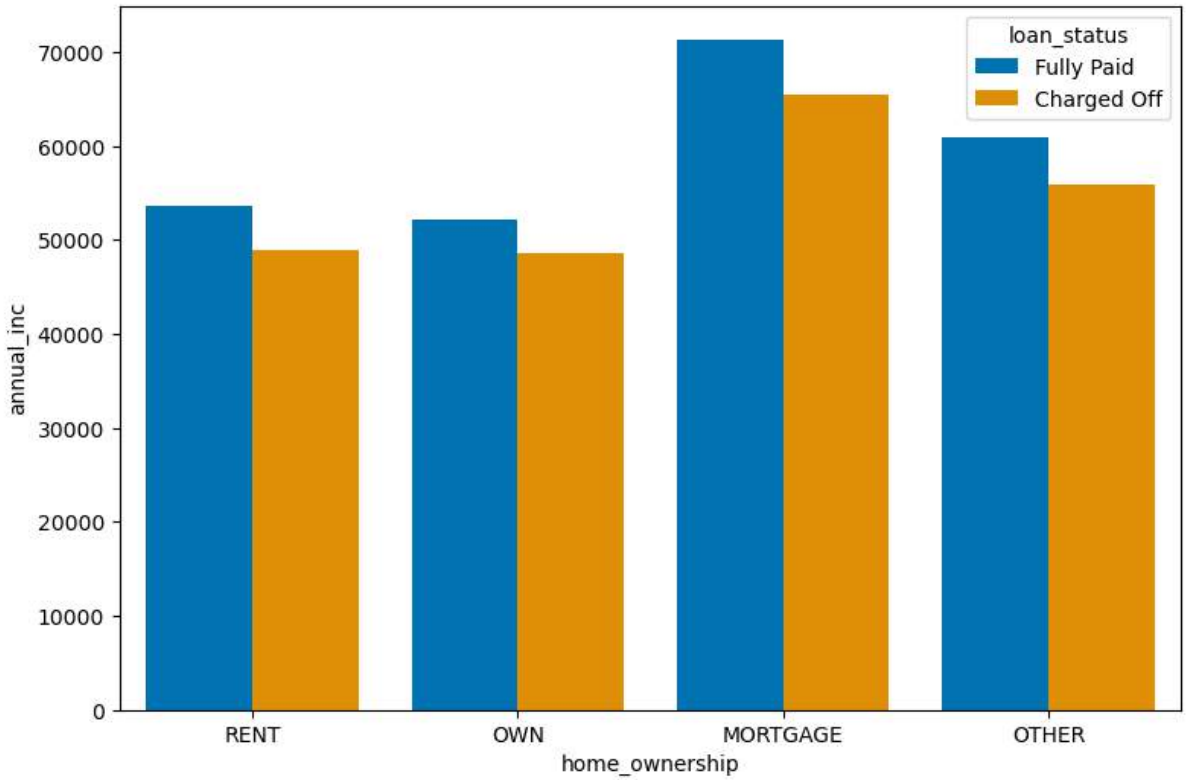
Annual Income and Grade



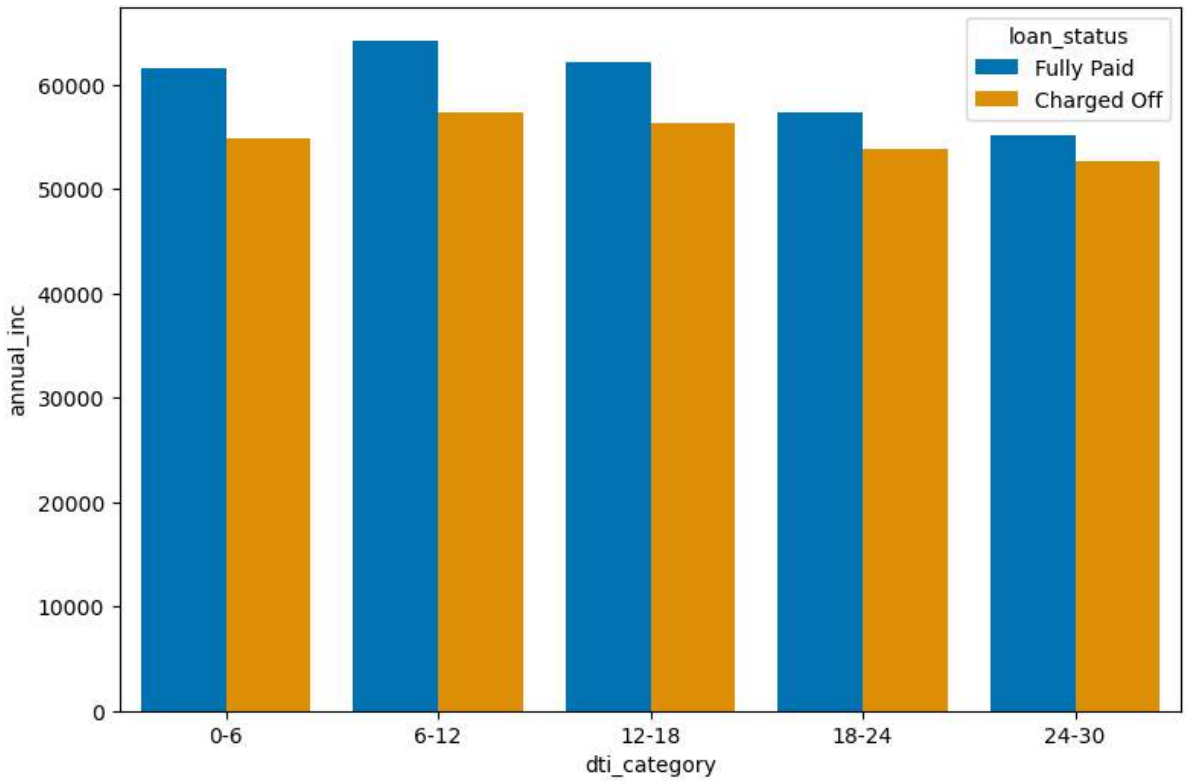
Annual Income and Employee Length



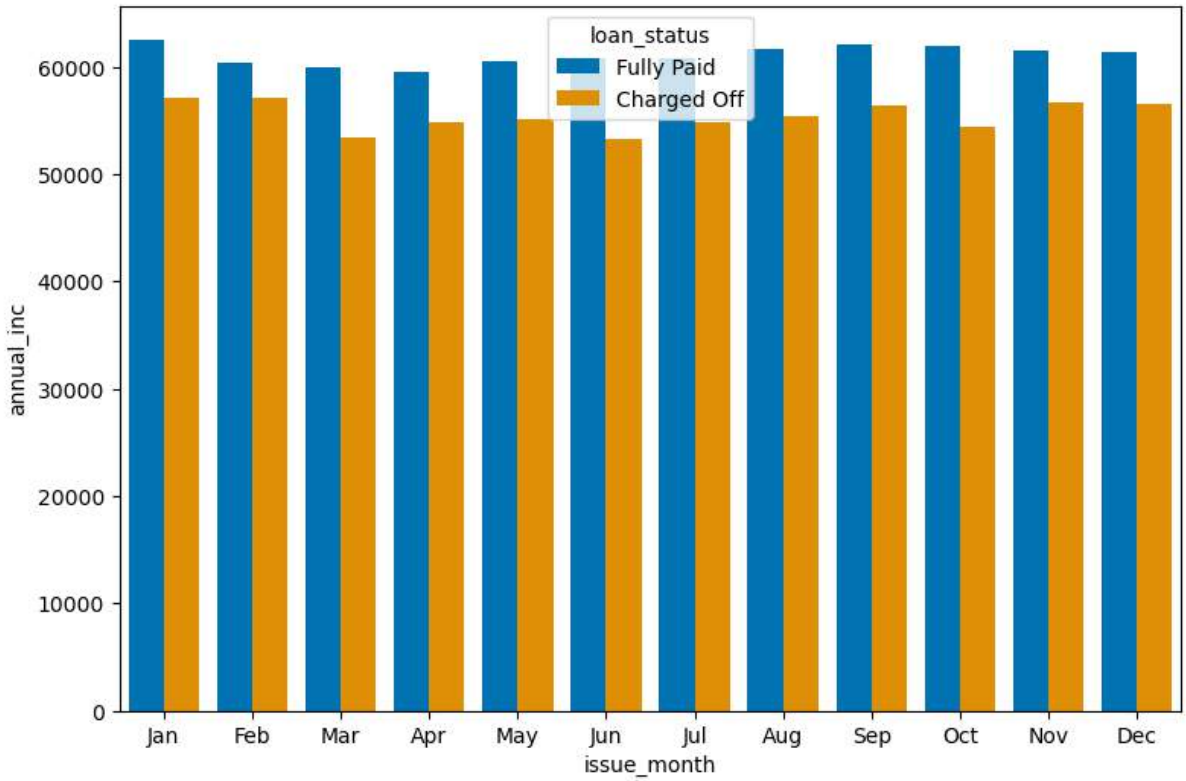
Annual Income and Home Ownership



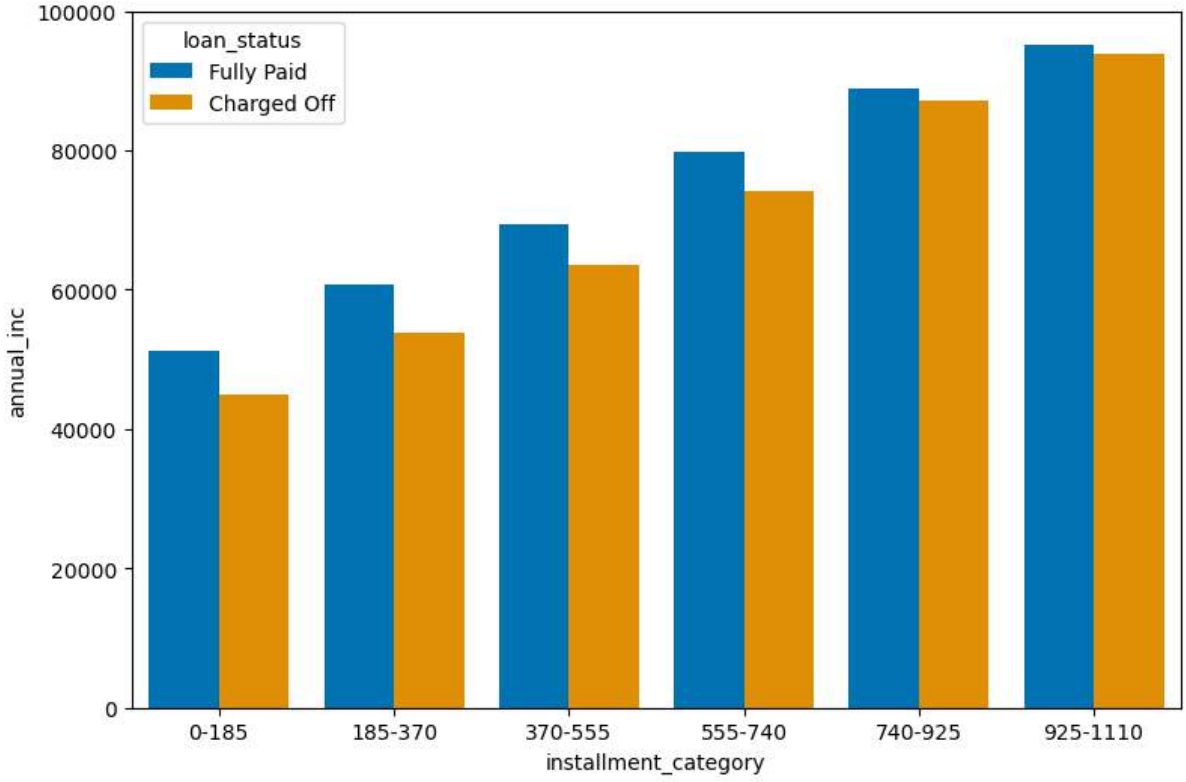
Annual Income and Debt-Income



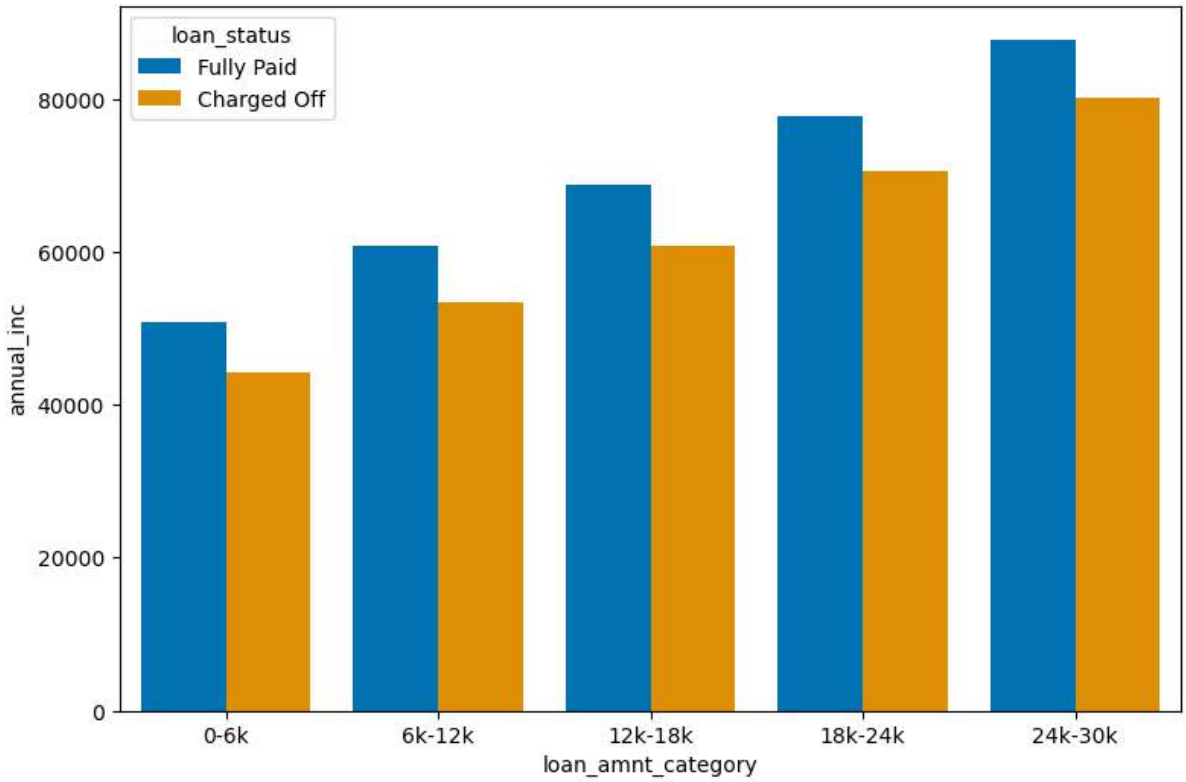
Annual Income and Issue Month



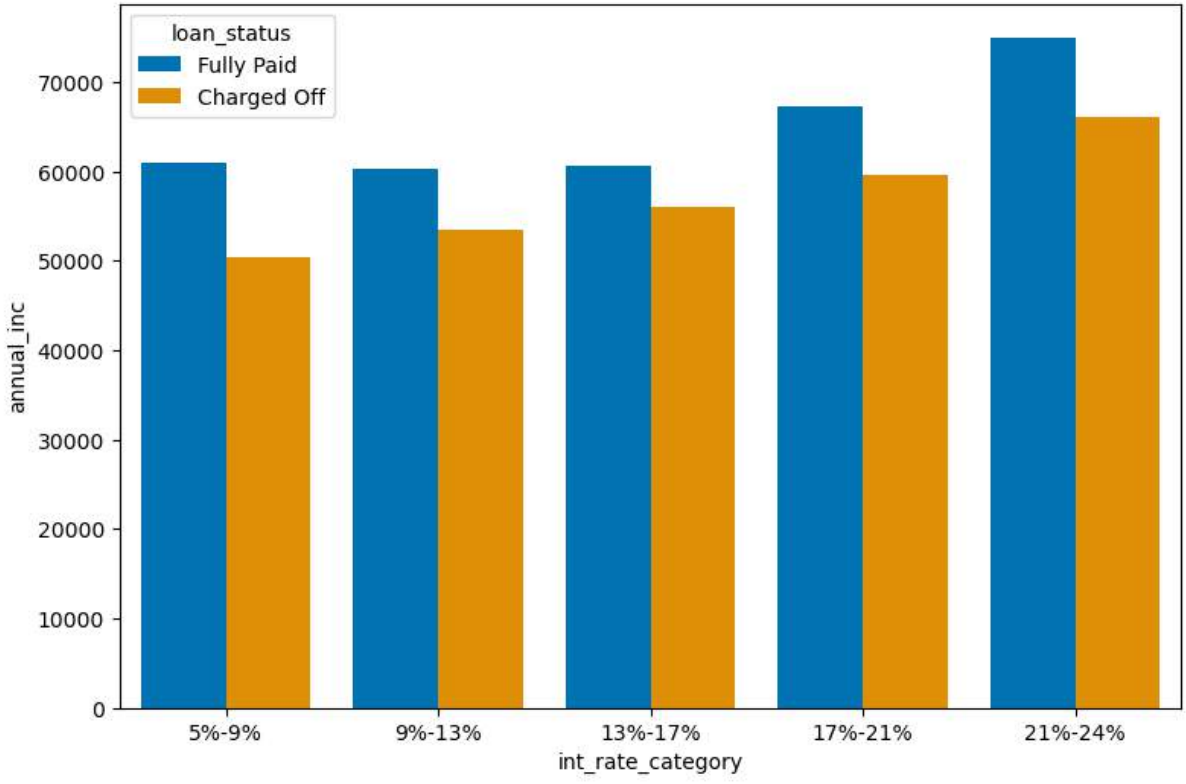
Annual Income and Instalment



Annual Income and Loan Amount

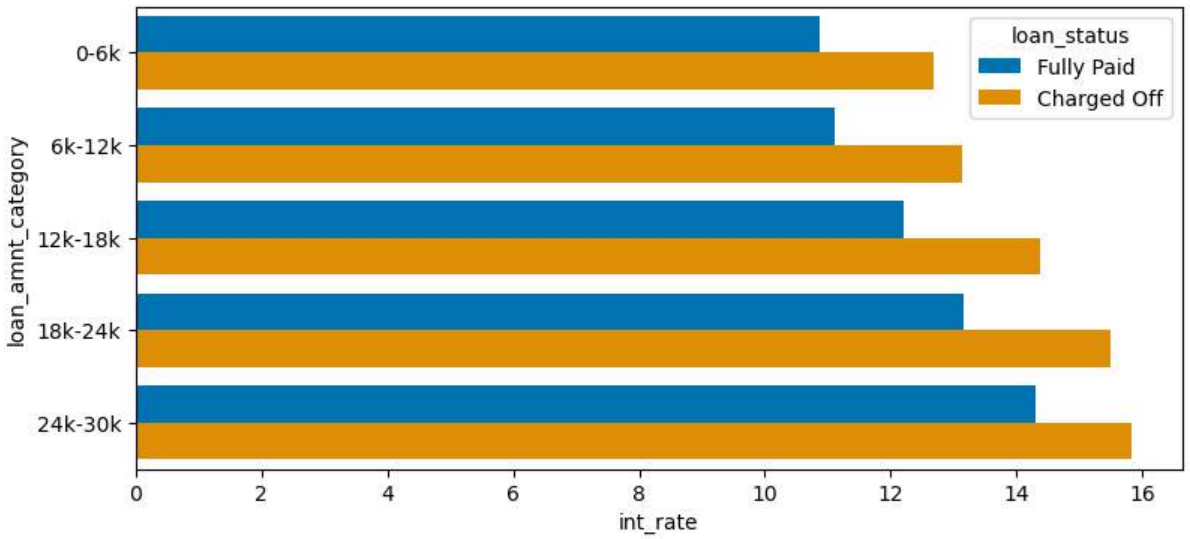


Annual Income and Interest Rate

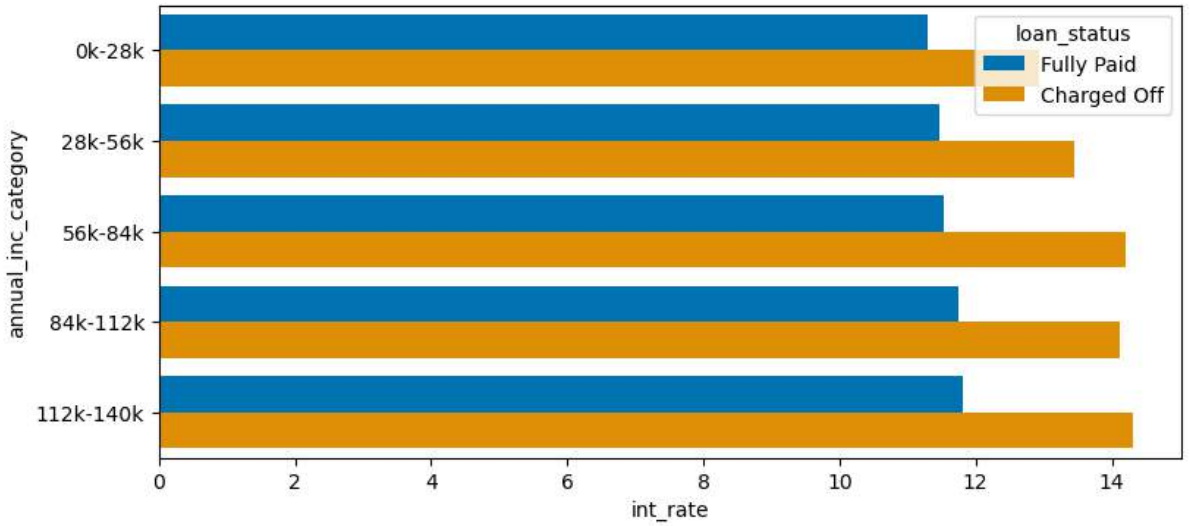


Analysing Interest Rate with other variables against the target variable loan status

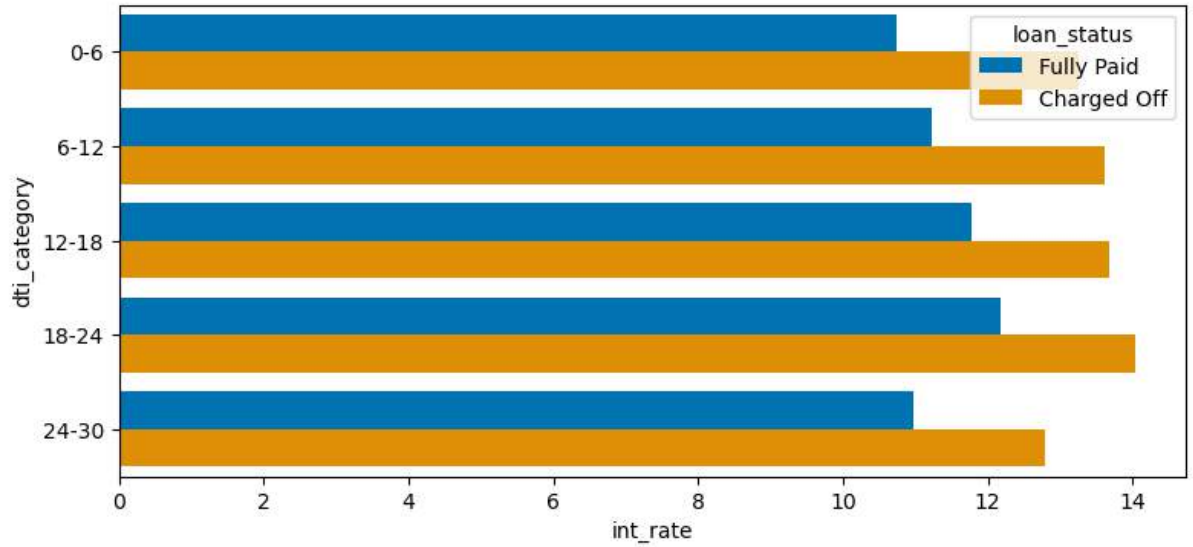
Analysing Interest Rate and Loan Amount



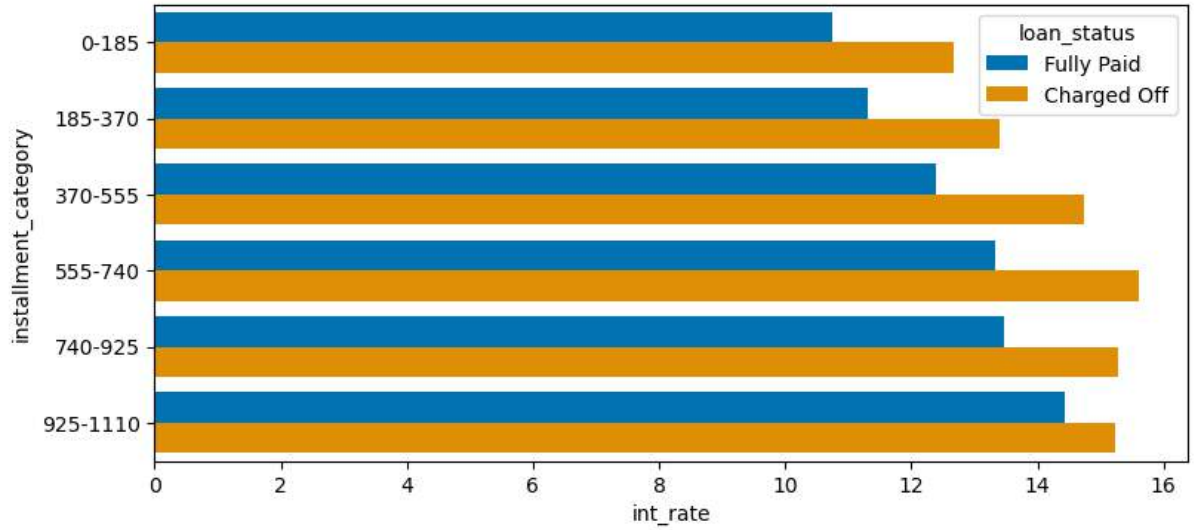
Analysing Interest Rate and Annual Income



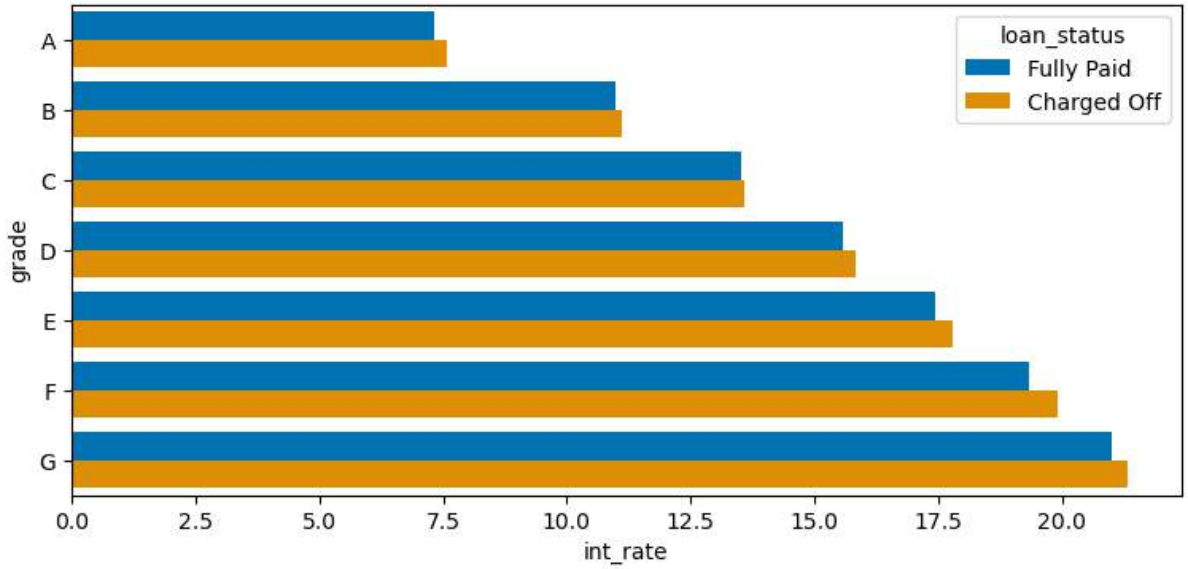
Analysing Interest Rate and Debt to Income



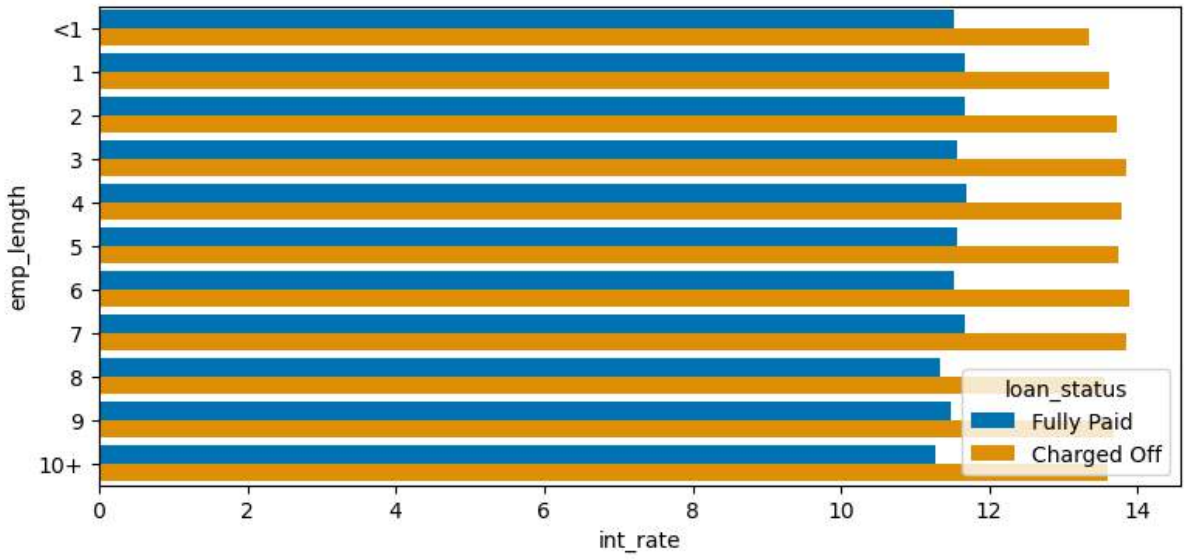
Analysing Interest Rate and Instalment



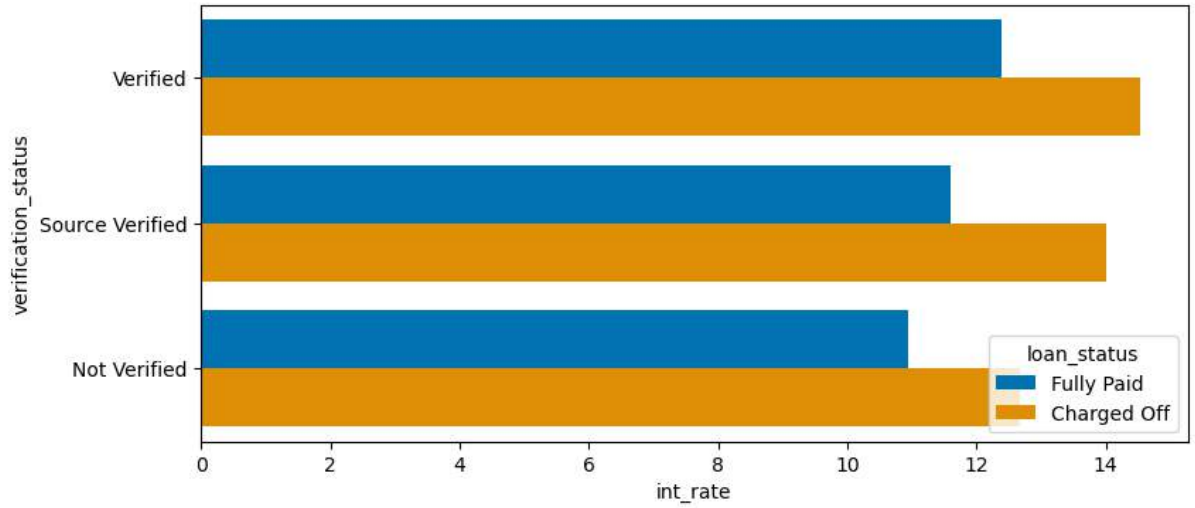
Analysing Interest Rate and Grade



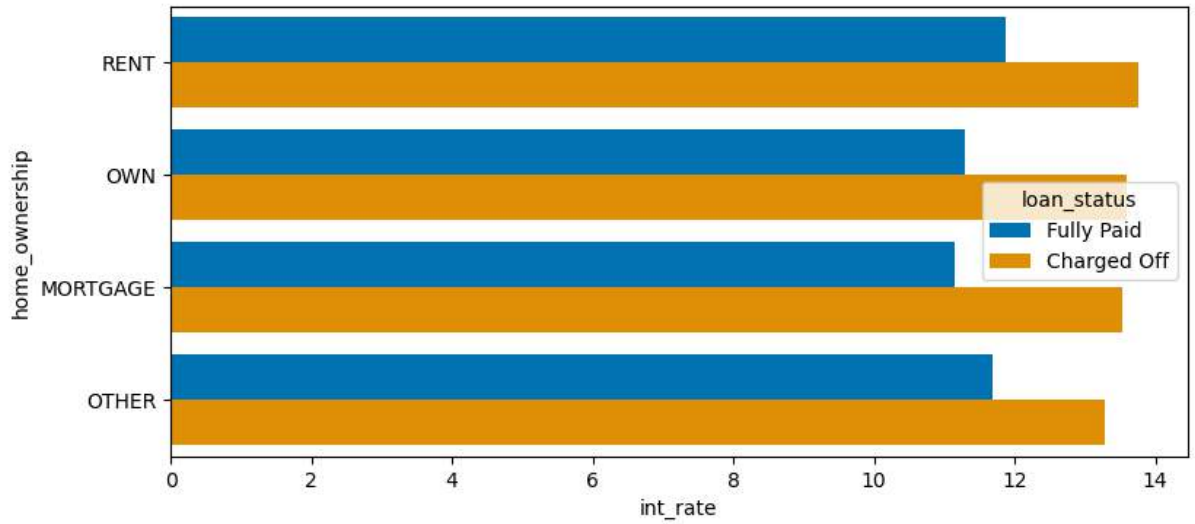
Analysing Interest Rate and Employee Length



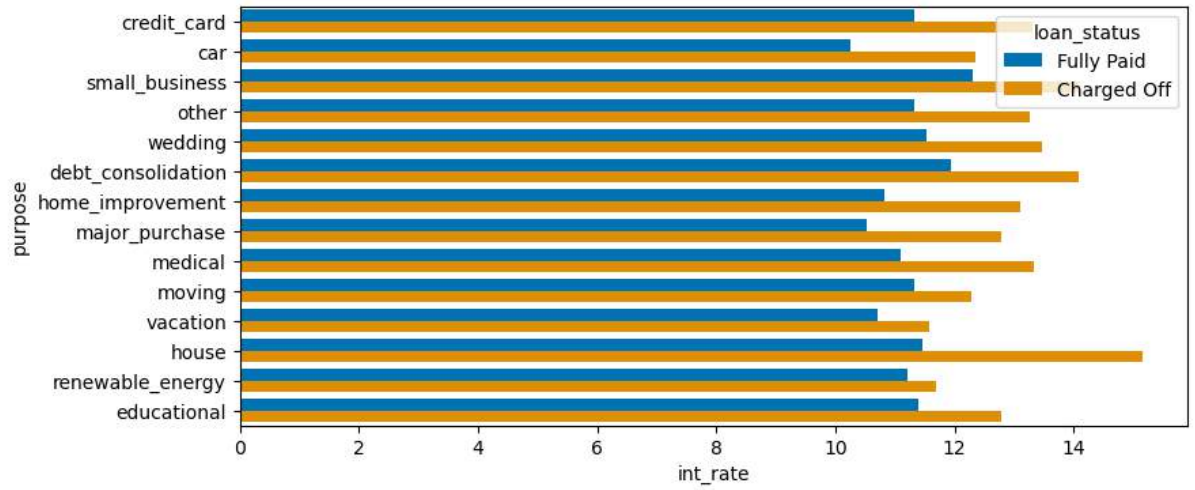
Analysing Interest Rate and Verification Status



Analysing Interest Rate and Home Ownership



Analysing Interest Rate and Home Ownership



Key Observation from Bivariate Analysis

The Charge off is higher in the following scenarios: (1K =1000)

1. If Loan Amount is higher than 10K for the purpose small business, debt consolidation, credit card
2. If Loan amount more than 12K, and term is 60 months
3. Loan amount higher than15K and interest rate is higher than 20%
4. Loan amount higher than 25K and instalment is higher than 925
5. Loan amount higher than 15K and Grade of the applicants are F and G
6. Applicants whose home is in Mortgage
7. Loan amount is higher and even if the verification is done
8. Applicant's Annual income is higher than 60k with purpose of loan is for high home improvement, small business
9. Applicant's Annual income is higher than 60k and falls in lower Grade as F and G
10. Interest Rate is higher than 14% and purpose of loan is small business

Recommendation:

1. Loan amount can be reduced those who fall under Grade F and G
2. Loan Interest Rate can be reduced if purpose of loan is small business, home improvement, credit card and debt consolidation (Loan amount higher than 15k and interest rate higher than 20% is likely to default)
3. Loan amount can be reduced if the home is already in mortgage

4. Conclusion

In general, from the given data the applicants with higher loan amount, higher interest rate and higher instalment are more likely to default. From our analysis we have identified strong indicators that contributes for the loan charge off, by implementing the recommendation we can reduce the risk of default.

Authors:

1. Antony John Sundar Aruldos
2. Subhrabindu Khuntia

Reference Material

1. UpGrad Material – Live session and recordings
2. Seaborn Libraries

   https://seaborn.pydata.org/tutorial/categorical.html

   https://seaborn.pydata.org/tutorial/distributions.html
3. Lending Club website

   https://www.lendingclub.com/personal-loan/rates-fees

   https://www.lendingclub.com/resource-center