Lending Club Case Study

Shashidhar Pattar

Shriyan Arcot

Introduction

This case study revolves around lending various types of loans to urban customers. The analysis is done using EDA and came up with different solutions for the Lending Club on what basis the loan can be approved for a customer. We have considered multiple methodologies.

EDA on the Dataset

The methodologies used are as follows

1. Data Cleaning
2. Univariate Analysis
3. Segmented Univariate
4. Bivariate Analysis
5. Multivariate Analysis
6. Derived Metrics

Data Cleaning

Data cleaning is an essential step in preparing data for analysis or modelling. It involves identifying and correcting errors, handling missing values, and transforming raw data into a usable format. Python provides powerful libraries and tools for data cleaning. Below is an overview of some common tasks and techniques for cleaning data using Python.

Main tasks need to performed on the data :

1. Handling Missing Values
2. Removing Duplicates
3. Handling Outliers
4. Transforming data in required format
   a. Standardizing / Normalizing Data
   b. Encoding Categorical Data
   c. Parsing dates

5. Cleaning Text Data (Using regex and other techniques)

Data Cleaning (Contd.)

- Clear all the columns with Null Values. We had 55 columns with null values which were removed in the subsequent steps.

```
Number of Columns with Null Values :  55
Columns to be removed are of length  :
['next_pymnt_d',
 'mths_since_last_major_derog',
 'annual_inc_joint',
 'dti_joint',
'percent_bc_gt_75',
 'tot_hi_cred_lim',
  …
 'total_bal_ex_mort',
 'total_bc_limit',
 'total_il_high_credit_limit']
```

- Cleared all columns with Unique values as well since they don't add any value to the analysis.
- We can clear off or drop columns with 0 or nan values along with some text fields like desc.
- Calculated the missing values in multiple columns and made sure the data is cleaned up further.

Data Cleaning using Data Dictionary file (cleanup unnecessary data):

A data dictionary file is also been provided along with the loan data. This file gives a brief describtion about all the columns of the loan transaction file. Let's use this file to get a better understanding of the 50 columns we are left with. This will help to further narrow down on the columns.

There are a few columns having a single unique value or all unique values. These may have nothing much to contribute to the analysis i.e. columns which are far from being the driving columns towards loan default. Some other columns are redundant.

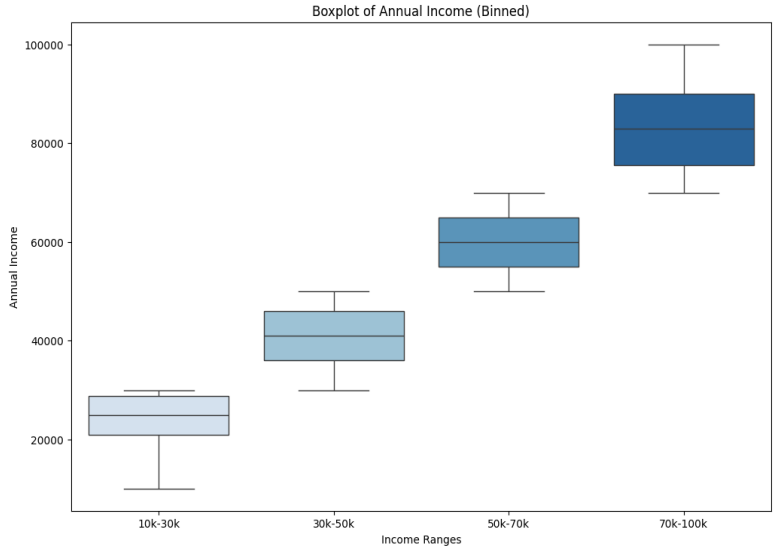
These columns are:

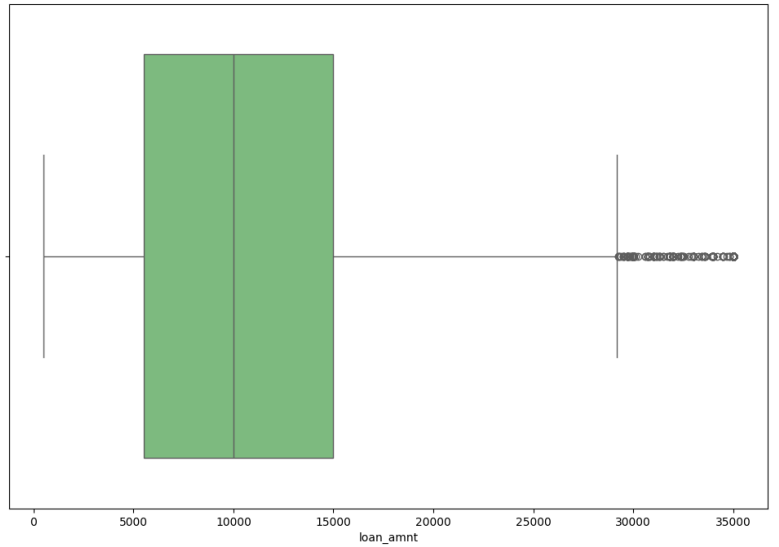
policy_code : value for entire dataset is '1', indicating all are publicly available, therefore nothing to contribute for analysis, to be removed.
application_type : value for entire dataset is 'INDIVIDUAL', indicating all are individual applications not joint, therefore nothing to contribute for analysis, to be removed.
acc_now_delinq : value for entire dataset is '0', therefore can be removed.
id : Since it cannot be used for any other purpose in the analysis we will be dropping it.
member_id : Member ID eventhough its unique and its per user it wont contribute much in the analysis
pymnt_plan : All the values for entire dataset is 'n', therefore can be removed.
url : This is URL is not useful as we cannot check the details by calling the URL as well.
zip_code : Zip code is not a complete one so its not valid in the analysis now.
initial_list_status : The data present in the Columns are f and its not useful for the analysis so removing it.
delinq_amnt : Since the values are 0 we are not considering it for analysis.

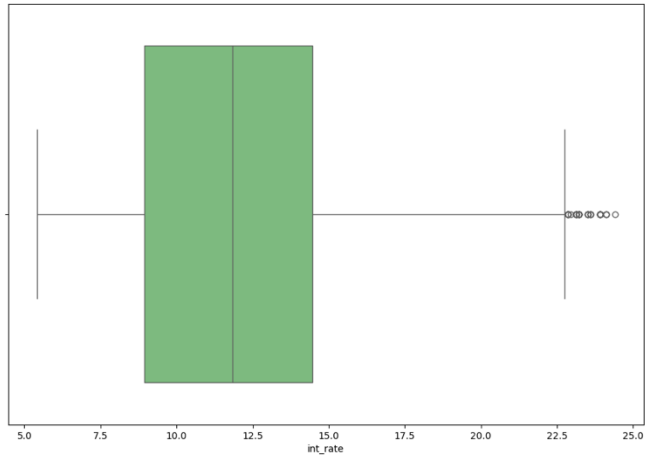
After cleaning up the data saved the final data set in final_loan_data.csv.

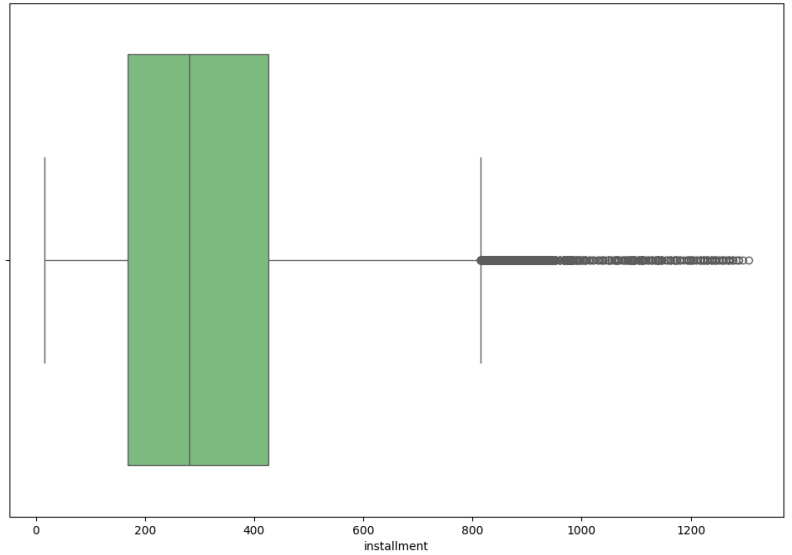
Univariate and Segmented Univariate Analysis

- Univariate Analysis was done one major items like loan amount, installment, annual income and others.
- We found that there are many outliers in the system.
- Few of the plots are as follows
- Annual Income , Interest Rates, Installment and Loan Amount etc.



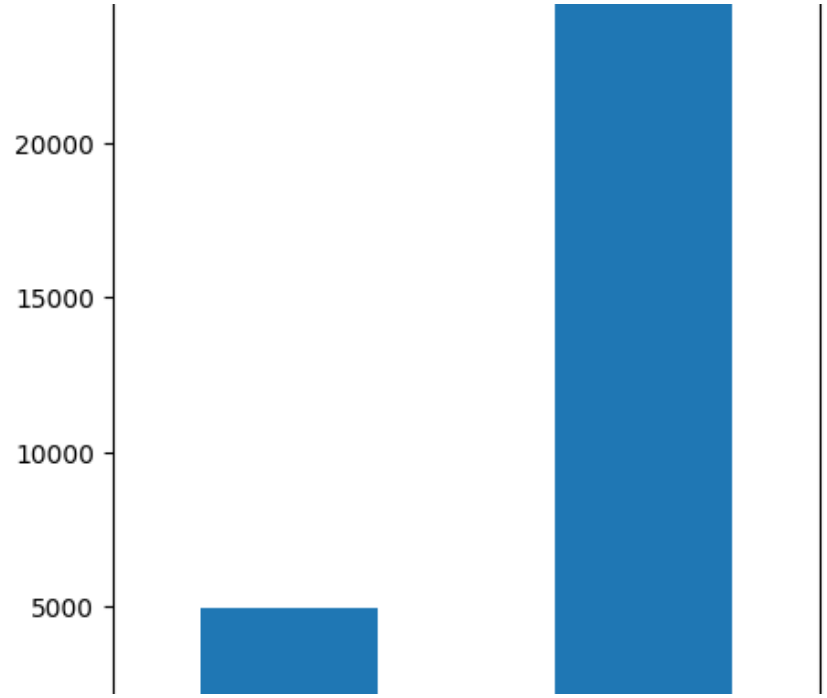


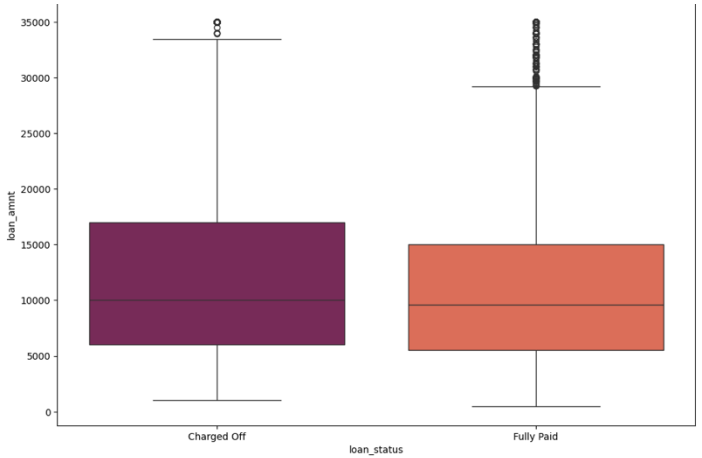




Bivariate Analysis

- Bivariate is the most important part of the analysis as we can compare multiple factors and come to a conclusion.
- First and fore most thing would be to check loan amount vs the loan status. Which will give us an idea of how many are fully paid and how many are charged off.
- **Observation** : Approximately 14% of loans in the dataset are defaulted. Any variable that increases percentage of default to higher than 16.5% should be considered a business risk. (16.5 is 18% higher than 13.98 - a large enough increase)



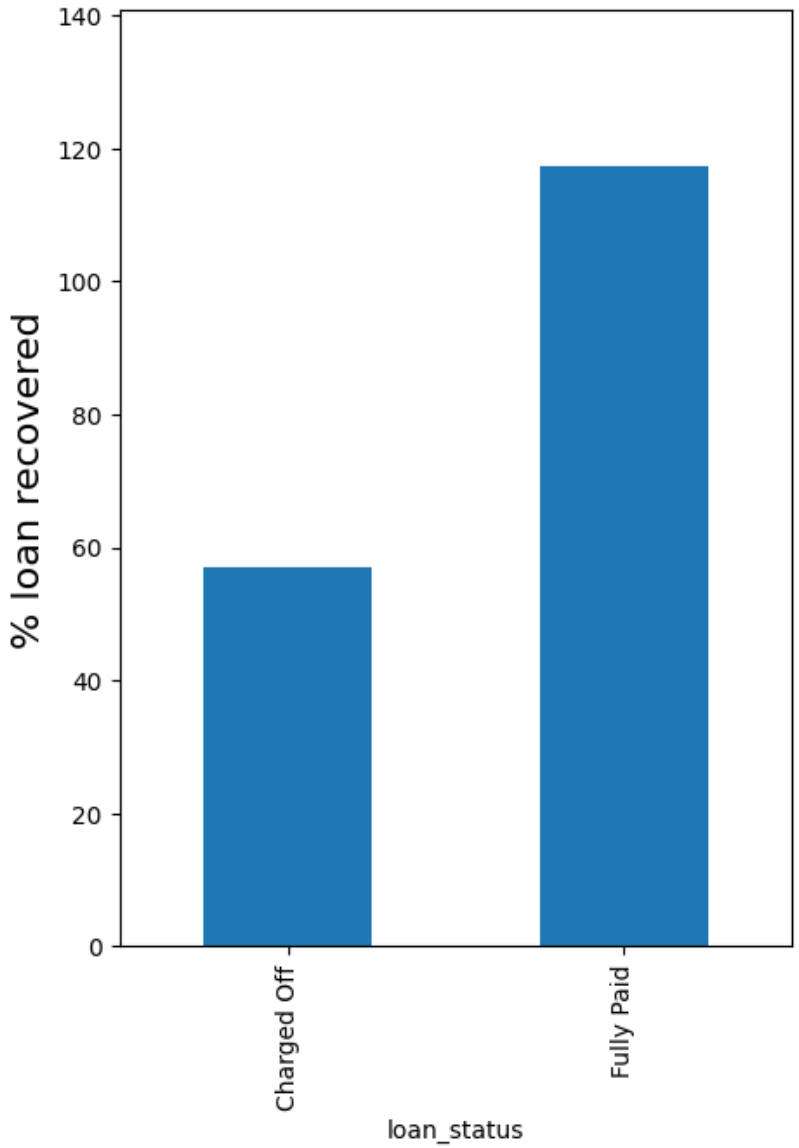


Bivariate Analysis

Percentage Loan Recovered :

Observation :

Lending Club only recovers 57% of the loan amount when loans are defaulted. On fully paid up loans, the company makes 17% profit.

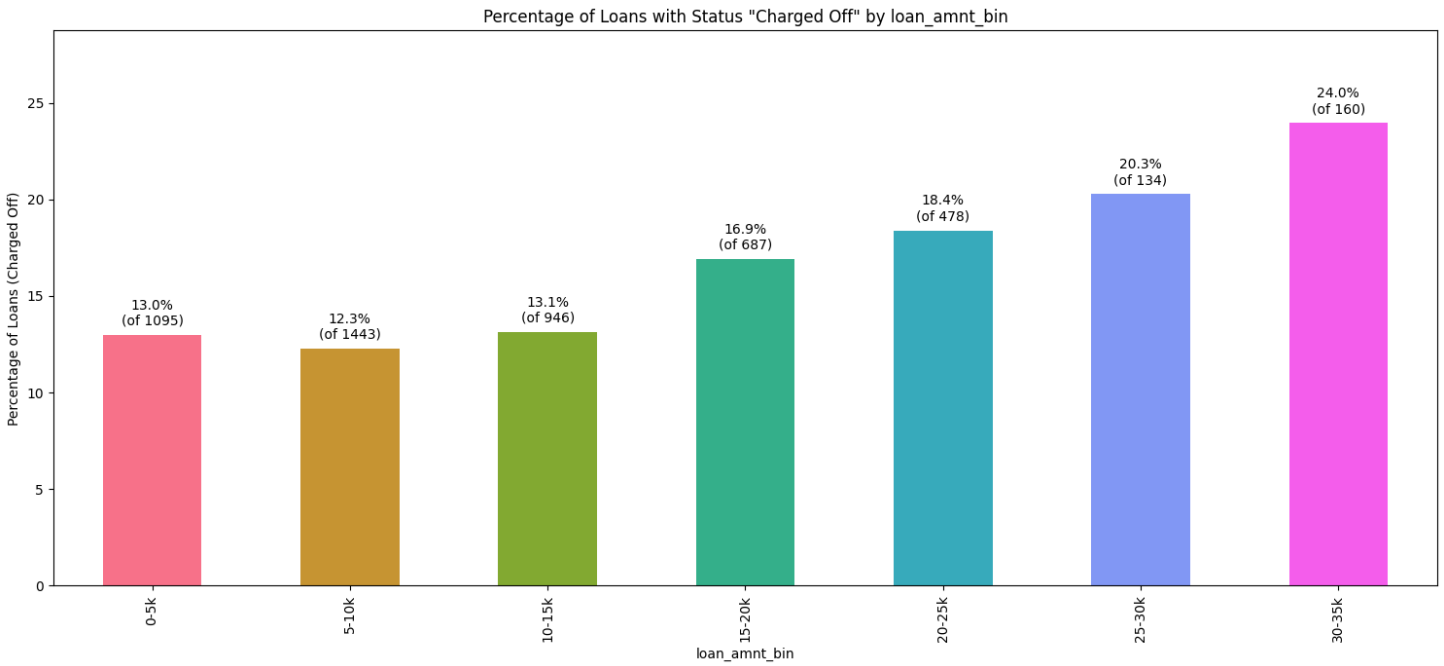


Bivariate Analysis - Binning

Let’s have a percentage distribution of data on the basis of loan amount bin and percentage of loans charged off.

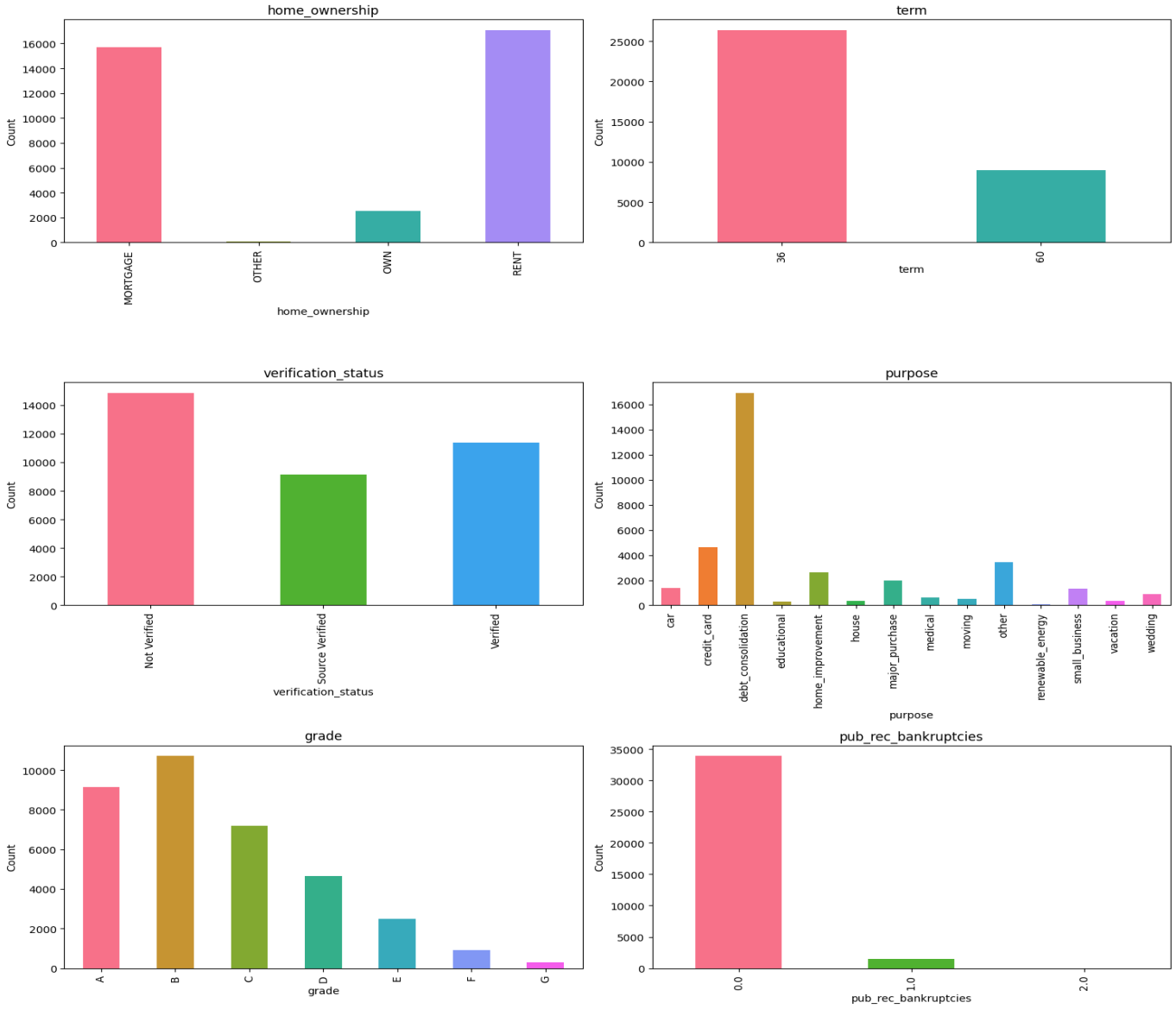
Observation :

The percentage of charged off loans increases substantially as we go up the loan amount buckets. Most loans are below 20000 amount. The higher loans, though lesser in number, carry a substantially higher risk of default.



Categorical Variable Analysis

- We can see that the categorical variables like grade, verification status, term etc are being plotted on the basis of loan amount.

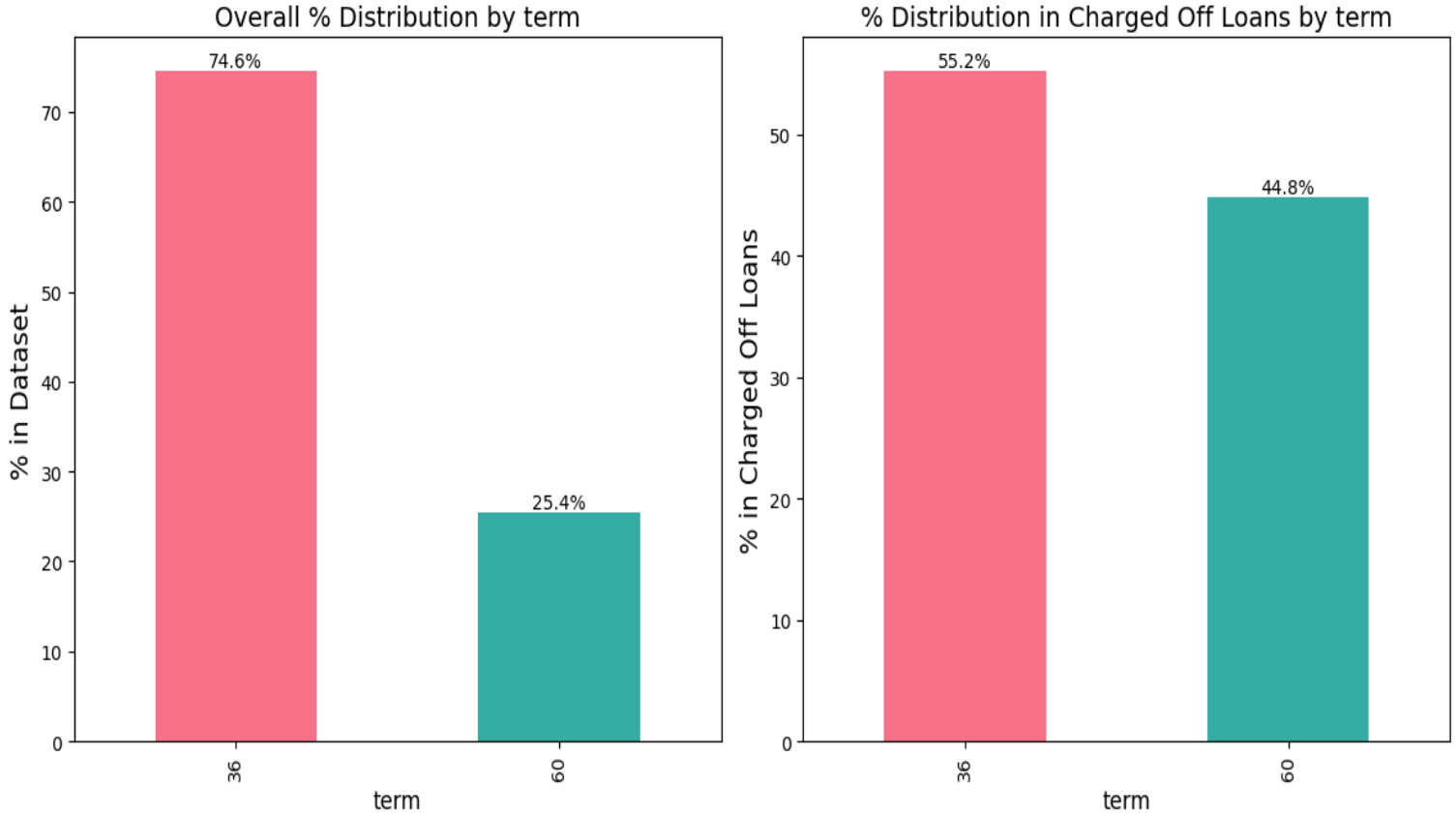


Loan Status vs Term

Observation :

Around 75% of the total loans are given for duration of 3 years. while just 25% of the loans are those given for 5 years.

Among Charged Off loans, percentage of term 60 months rises to 45%. The higher term loans have a higher chance of default.



Loan Status vs Purpose

Observation :

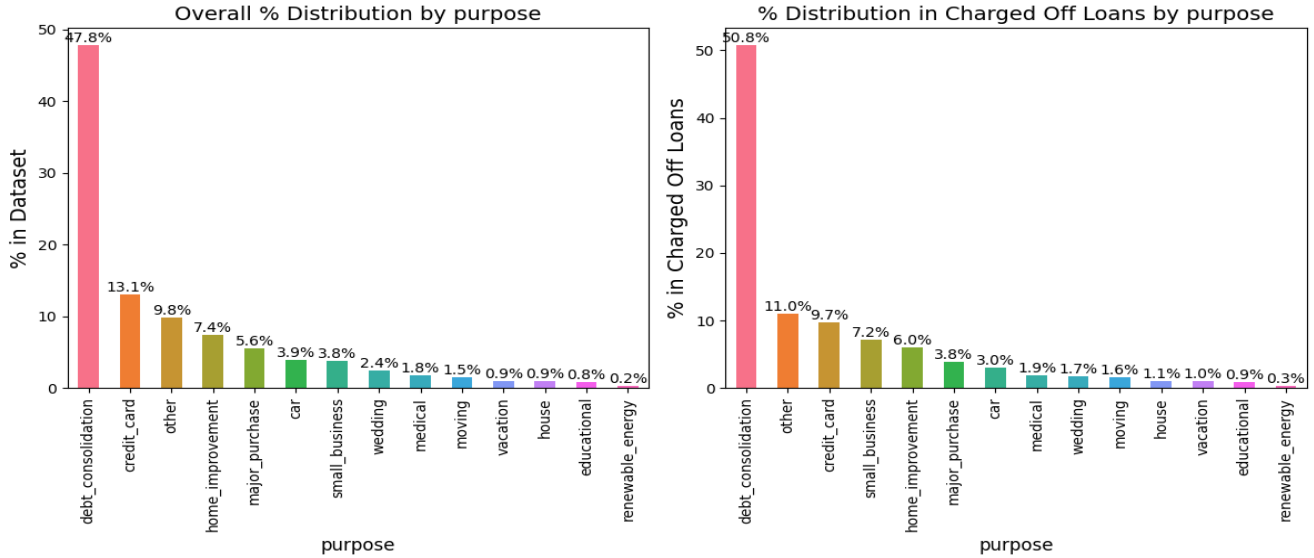
The category small_business percentage doubles from 3.8 to 7.2 for Charged Off loans.

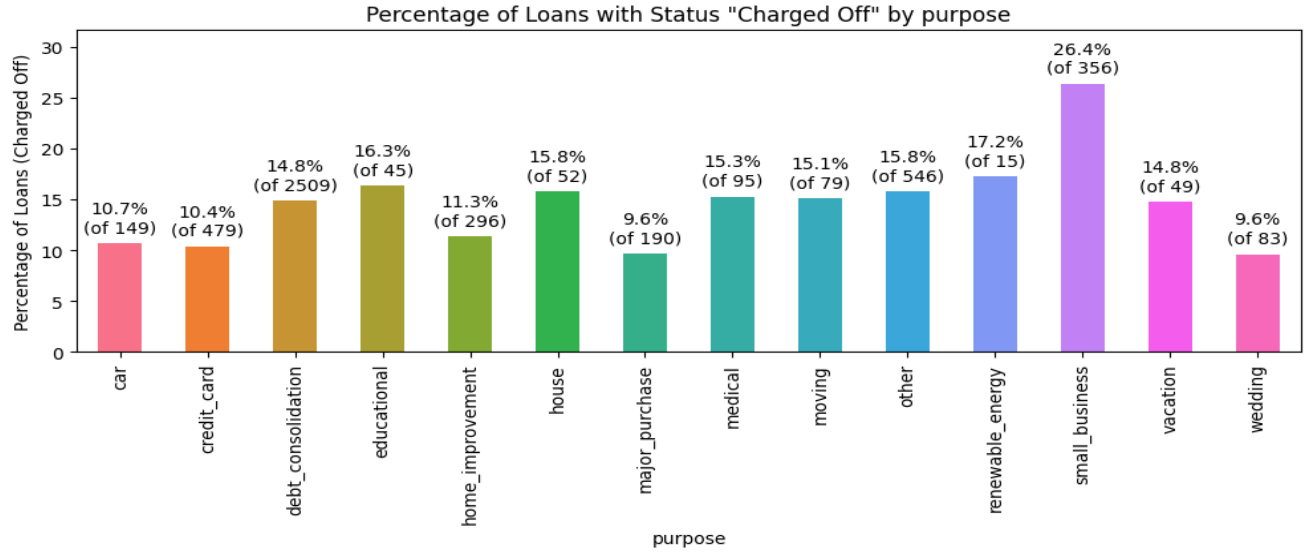
Let's see how the categories of this variable behave.

26.4 % of loans for small business are Charged Off. Making them the most risky purpose.

Approximately 50% of the loans are issued for the purpose of dept consolidation.

17% of the loans for renewable_enrgy are charged Off, but the number is too less to be of significance.

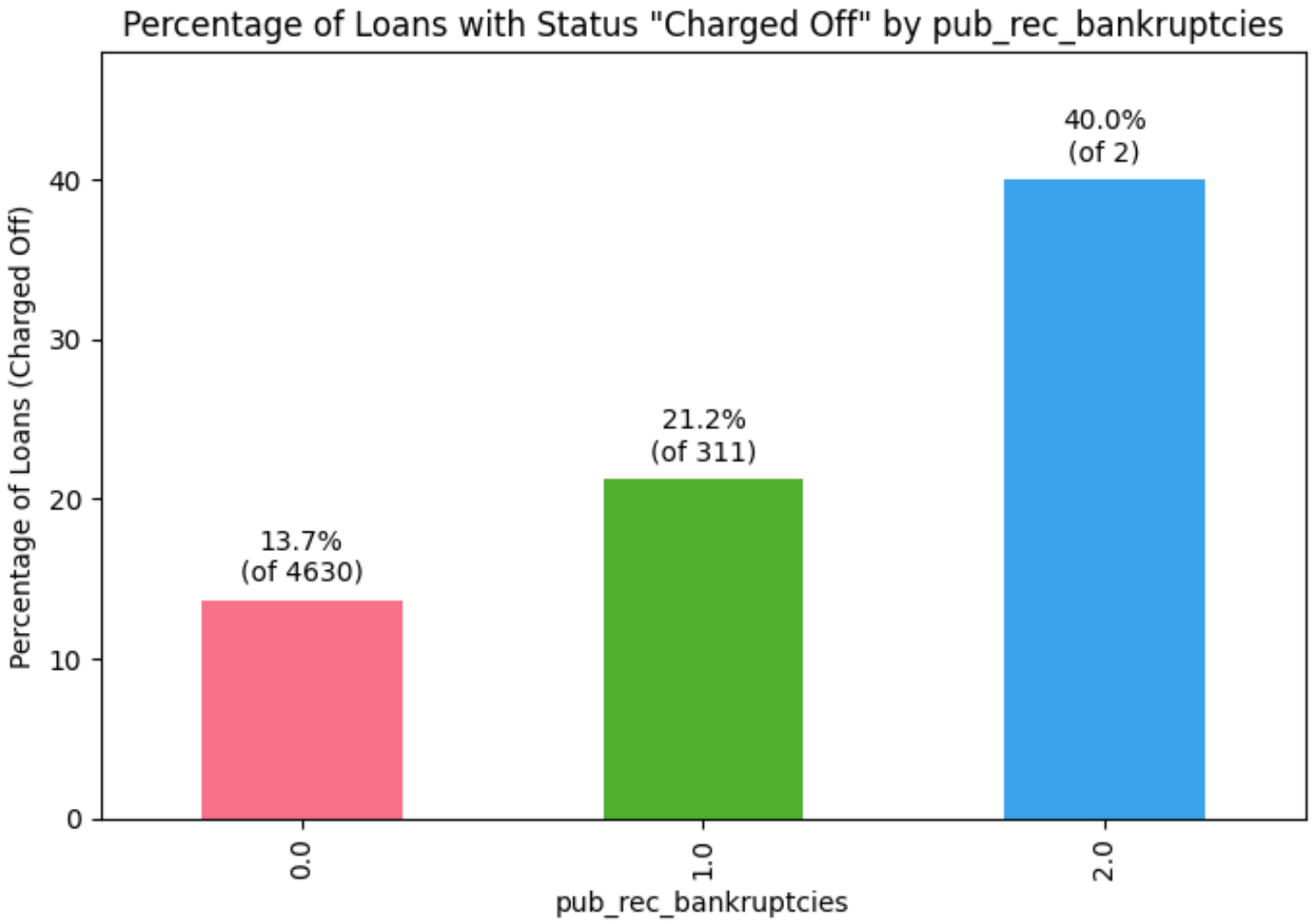




Loan Status vs Public Record of Bankruptcies

Observation :

The percentage of Charged Off loans is markedly higher when the borrower has a prior record of bankruptcy.

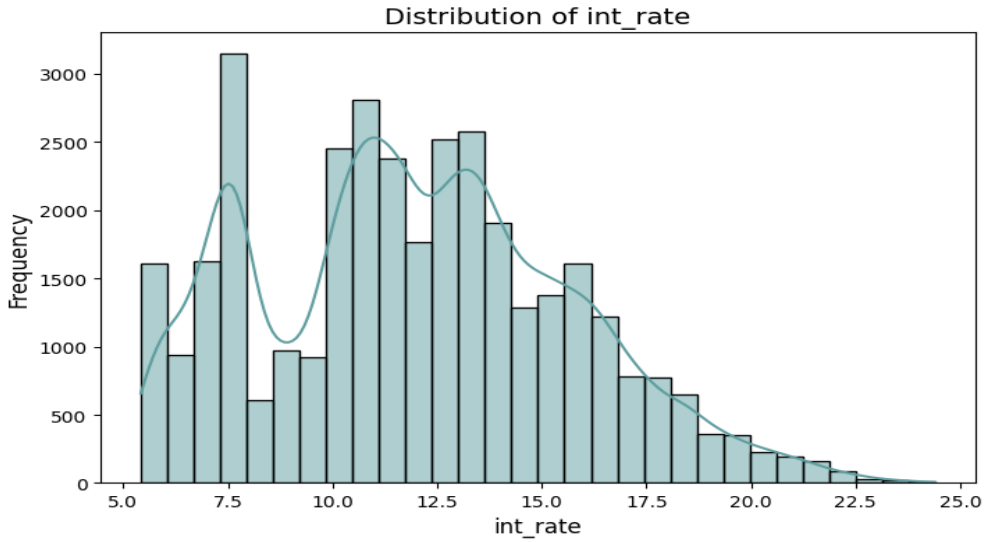


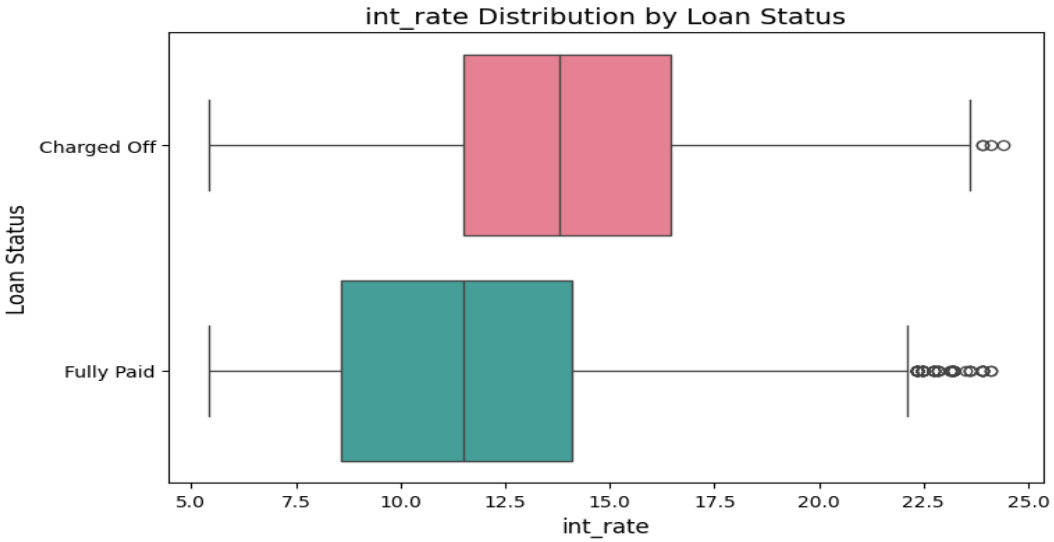
Interest Rate vs Frequency and Loan Status

Overall, the interest rate varies from 5.42% to 24.4% with average interest rate of 11.66%.

The interest rate for Charged Off loans appear to be higher than for Fully paid. This is naturally expected. As, the risk increases the rate of interest imposed on the loan also increases.

	count	mean	std	min	25%	50%	75%	max
loan_status								
Charged Off	4943.0	13.929828	3.647619	5.42	11.49	13.79	16.45	24.40
Fully Paid	30424.0	11.667243	3.613734	5.42	8.59	11.49	14.09	24.11





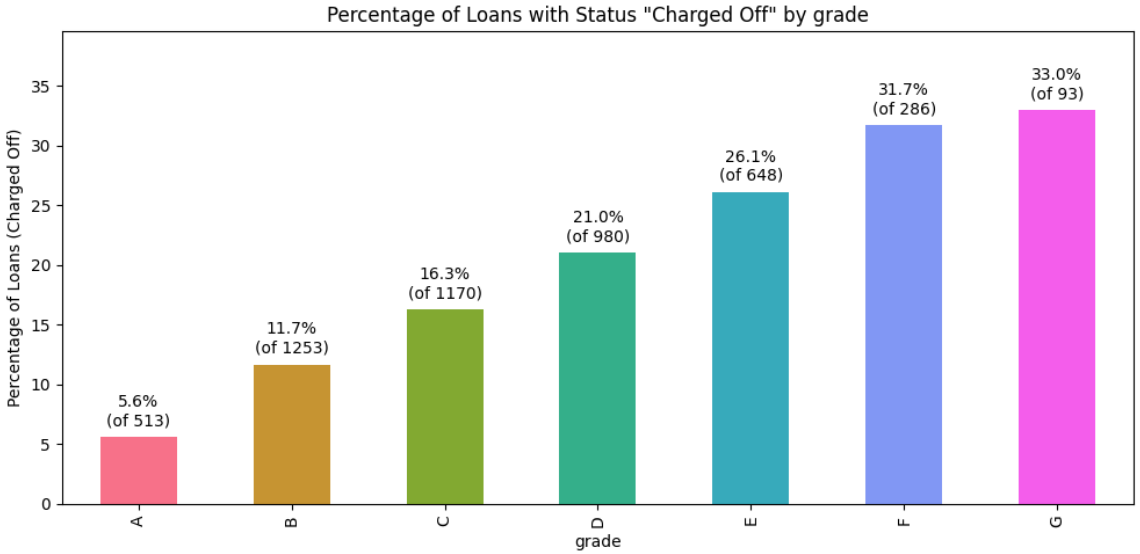
Grade vs Loan Status

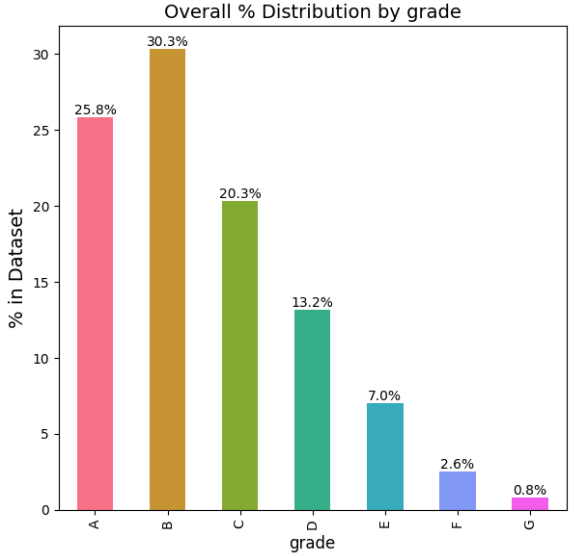
Observations :

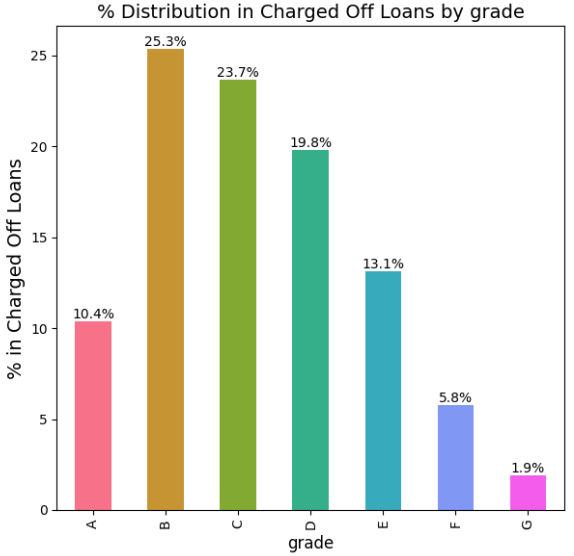
Nearly 30% of all loans in Grades F and G see a default.

Grade E onwards are risky, and less numerous. Lending Club should either refuse loans or charge high interest rates.

The grades A, B and C are safer than D, E and F. Lets now calculate the percentage of Loans charged off per grade.





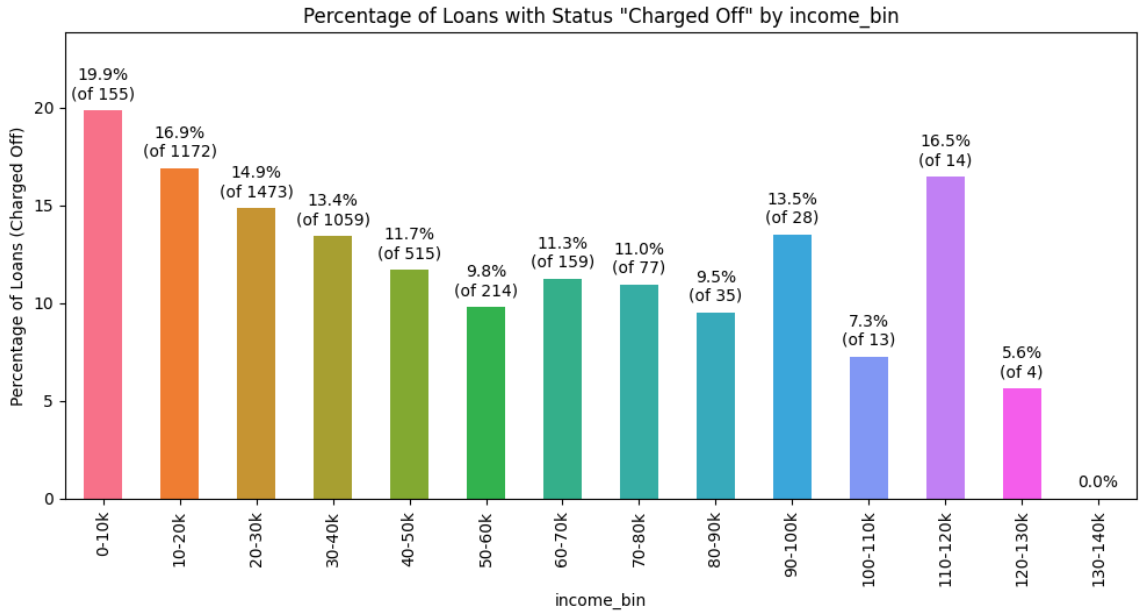


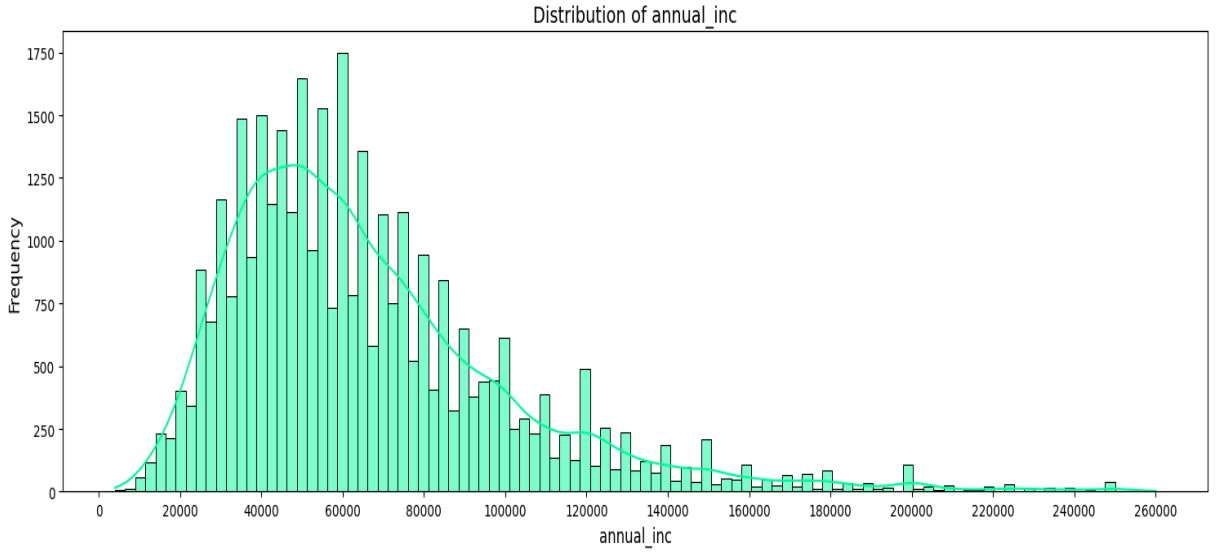
Annual Income and Loan Status

We can see that the bin size 10000 (income) looks good since the data is distributed properly.

Loan defaults are higher for lower income, and progressively reduce as incomes go up.

It will be interesting to see a bivariate analysis of defaults by income buckets and loan amounts later.



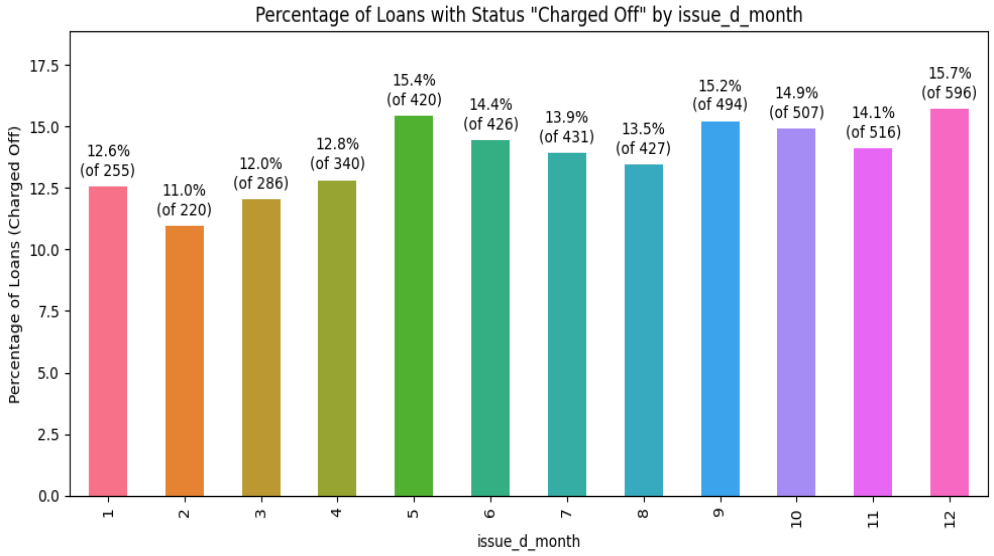


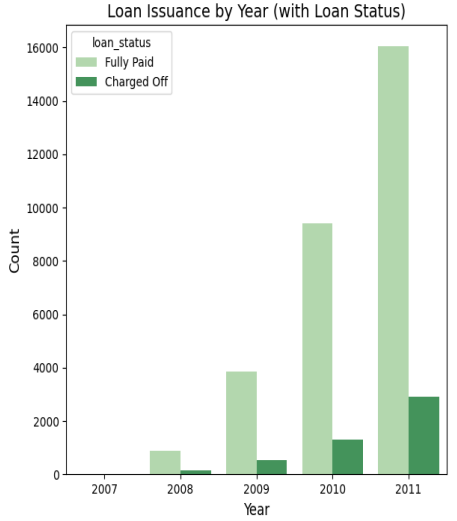
Loan Issued Month and Status

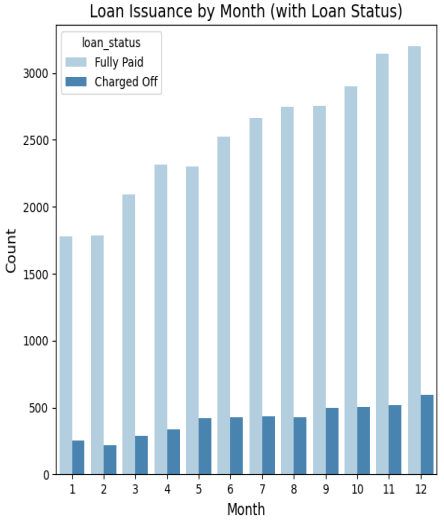
If we consider the above analysis we can see that from Aug,2007 to 2011 the loans issued have risen significantly.

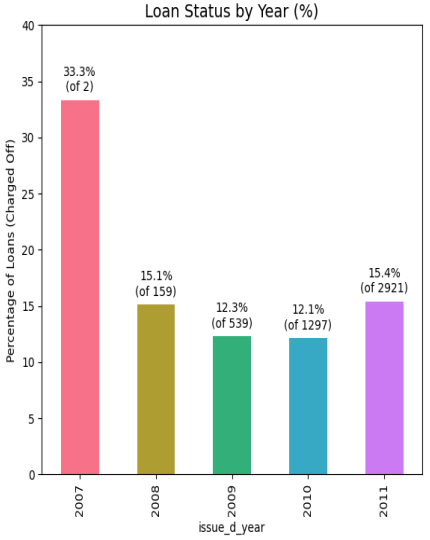
Within a year, the number of loan issued rises over the month from jan to December.

Month of loan is spread out and has no significant effect.





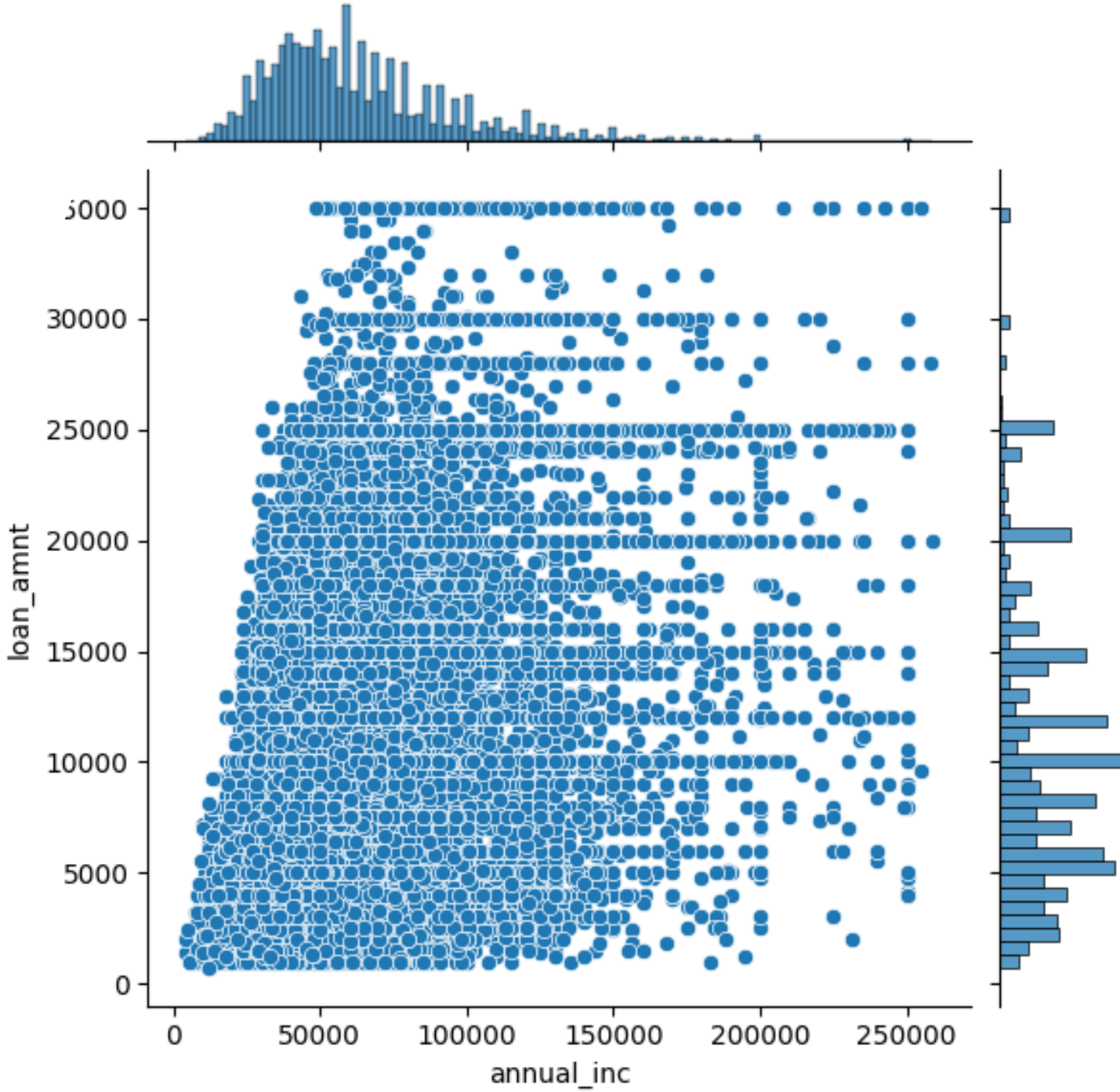




Loan Amount vs Annual Income

Observation :

There are people with average income lower than 50000 taking loans of 25000 or higher. These would be risky loans



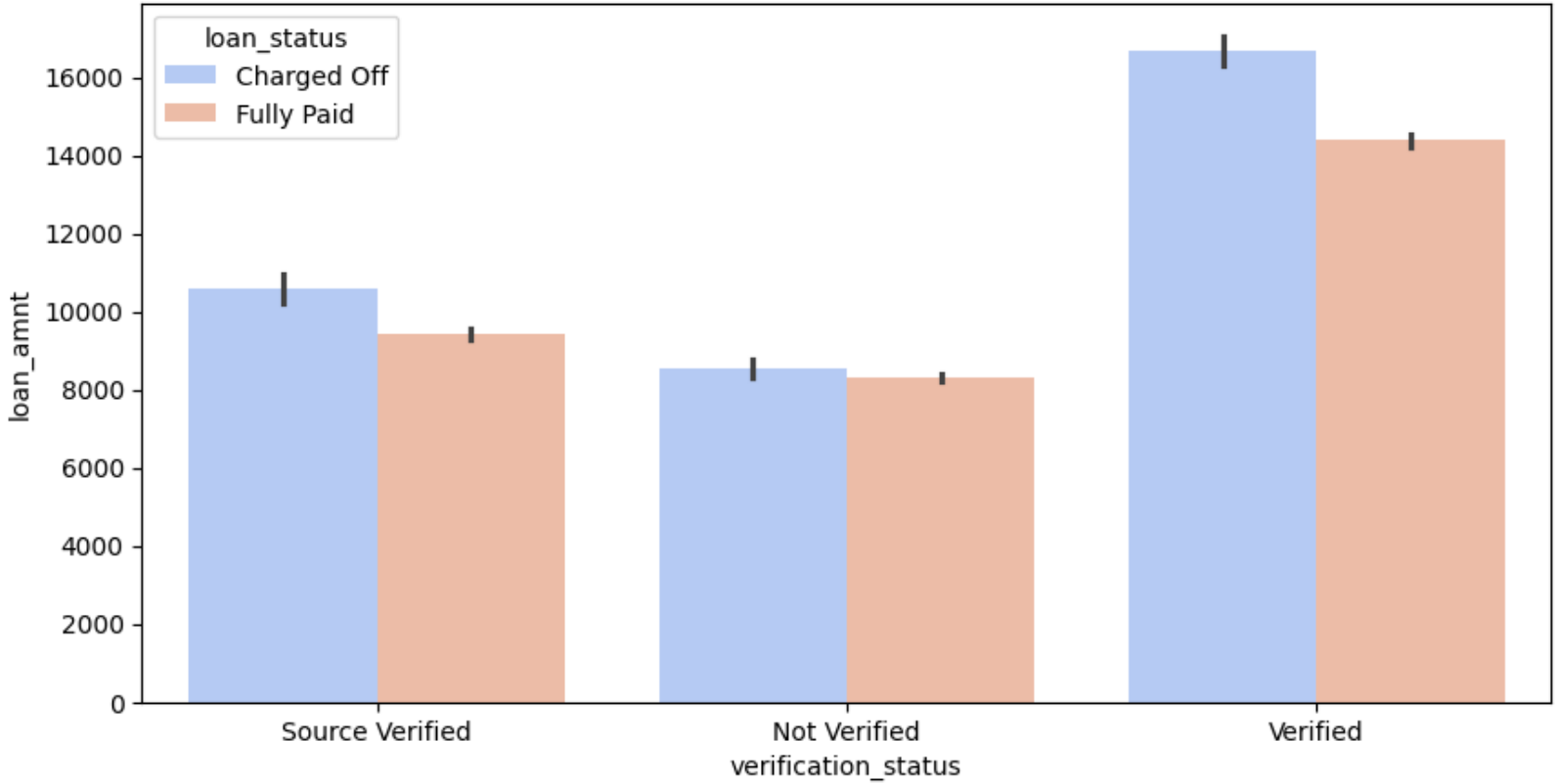
Multivariate Analysis

Loan Amount vs Verification Status vs Loan Status

Observation :

Higher loan amounts are Verified more often.

And others have lesser verification status.

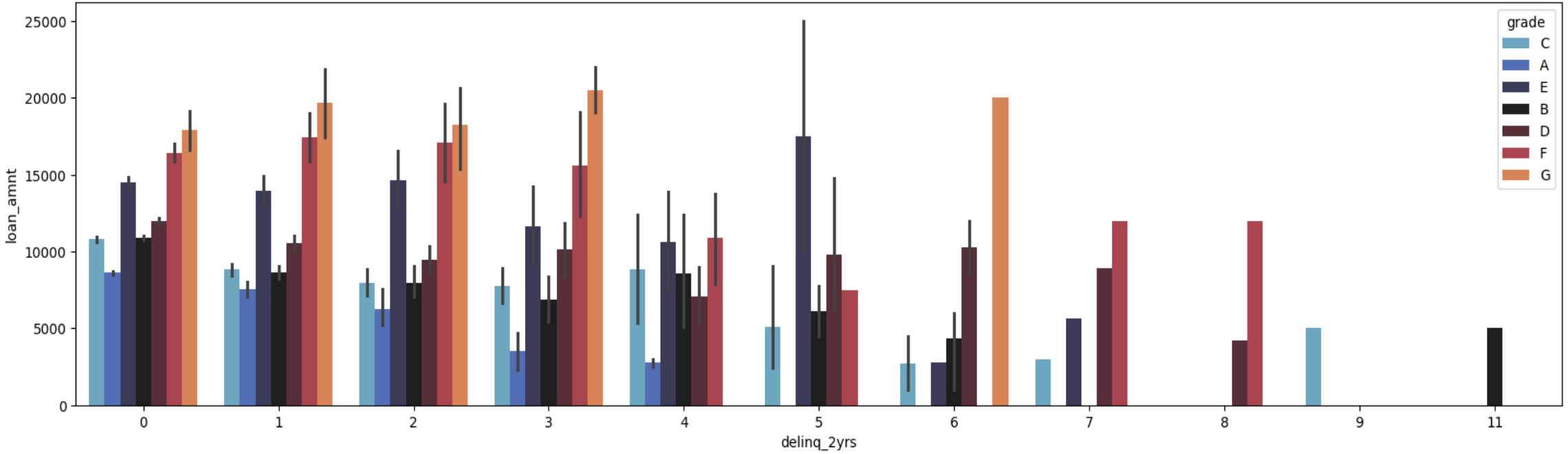


delinq_2yr VS loan amount VS grade

Observation :

Not many loans receive investment with higher number of delinquencies(>3). Despite the low loan amount request, these loans are considered risky and are not invested much in.

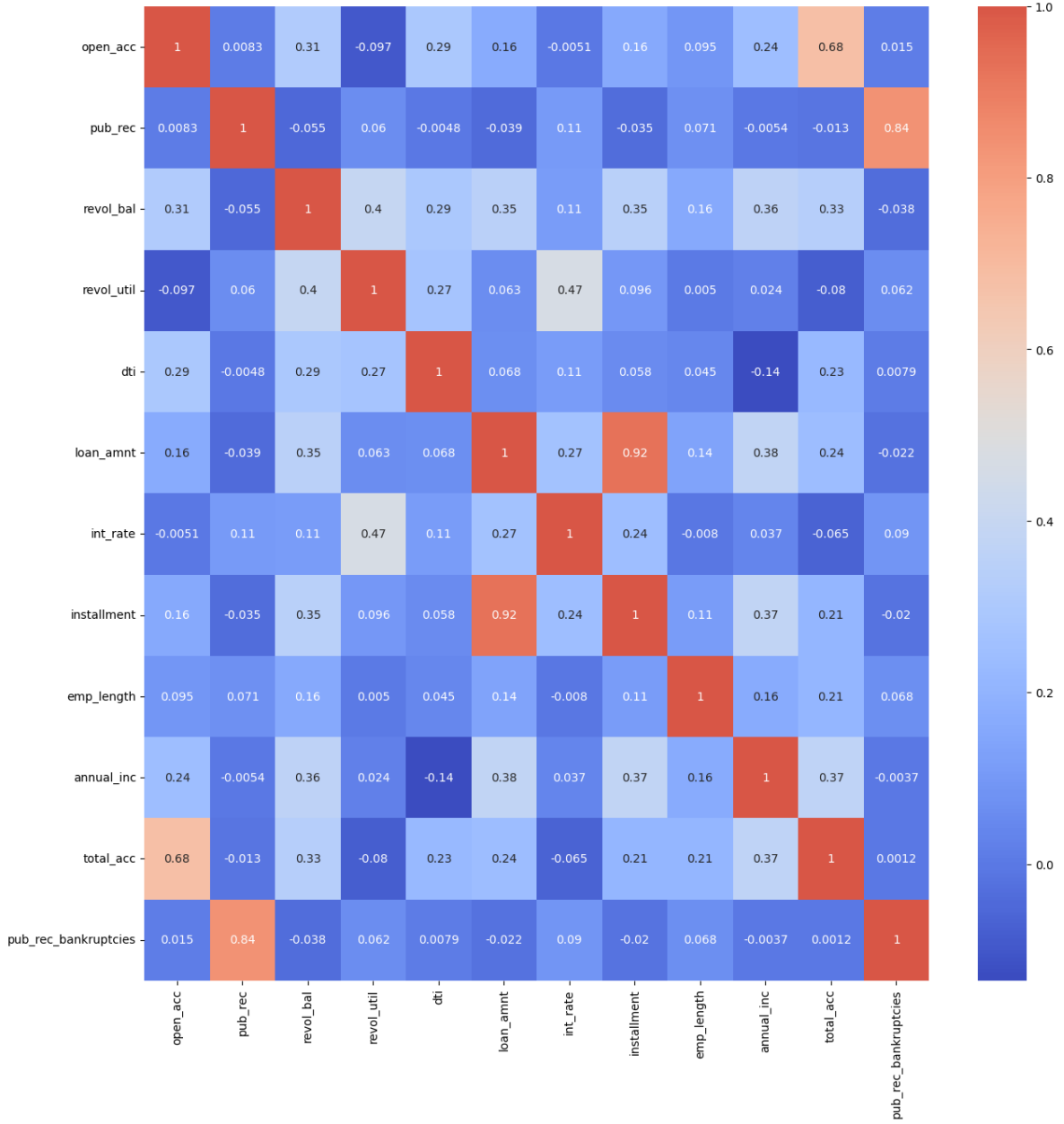
Lending club should further restrict their investment. We see loan amounts of >15000 on average for people having >=2 delinquencies.



Correlation

Observation :

- Strong Positive Correlations :
  - open_acc - total_acc -> 0.68
  - pub_rec - pub_rec_bankruptcies -> 0.84
  - revol_util - int_rate -> 0.47
  - loan_amnt - installment -> 0.92
- Weak Correlations :
  - total_acc - pub_rec_bankruptcies -> 0.0012
  - emp_length - revol_util -> 0.005
- Negative Correlations :
  - dti - annual_inc -> -0.14
  - open_acc - revol_util -> -0.097
  - pub_rec - loan_amnt -> -0.039



Thank you