Lending Club Case Study

SUBMITTED BY :

MOHAMMAD TASLEEM ARIF

MEENAKSHI GUPTA

Background

Background:

Lending company 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. The Customer applies loans to the company on basis of different Customer attribute & loan attribute the company decides whether to give loan or not.

When a person applies for a loan, there are two types of decisions that could be taken by the company:

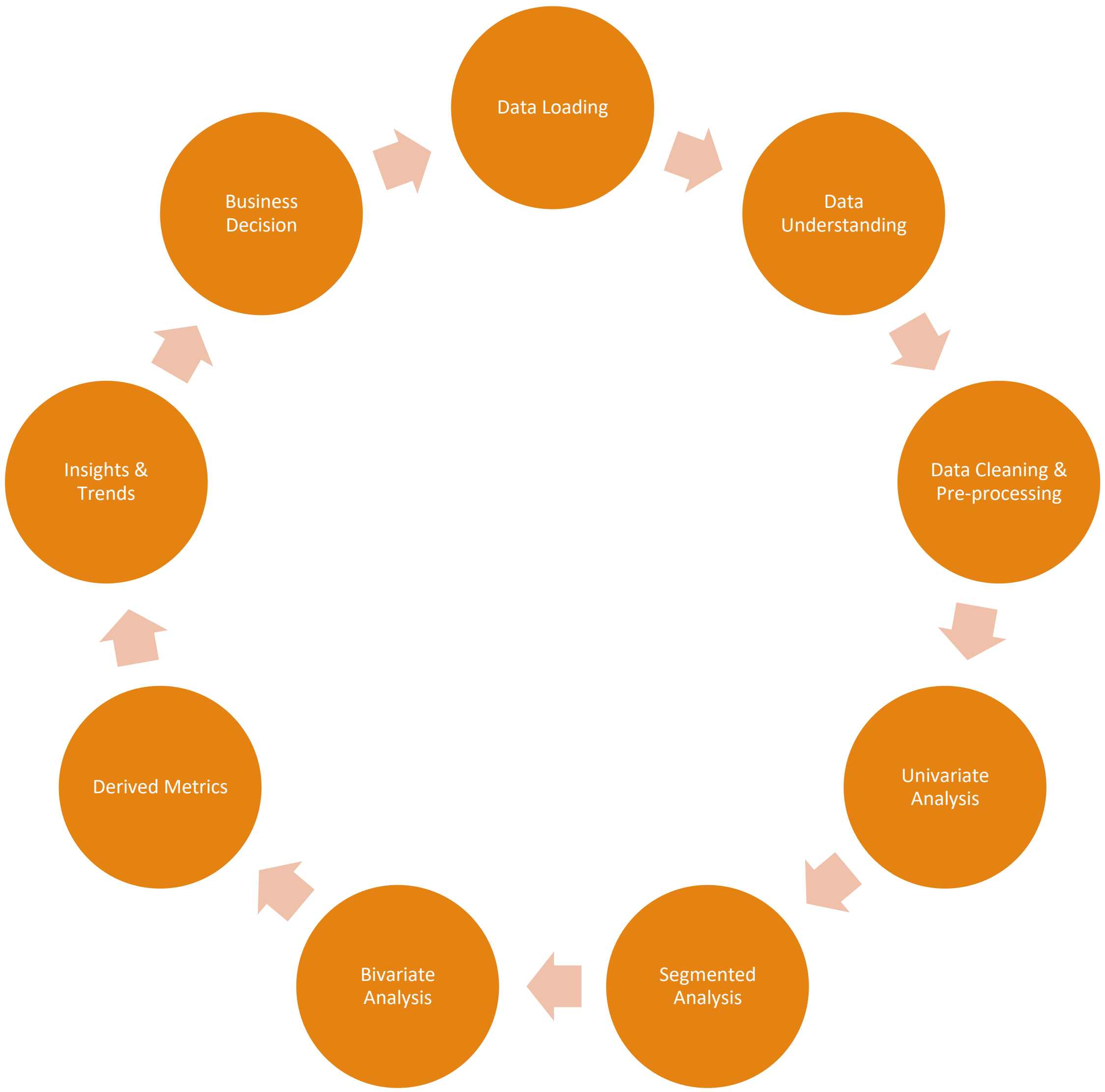
- Loan accepted: If the company approves the loan, there are 3 possible scenarios Fully Paid, Current & Charged Off
- Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.).

Business Objective:

This project aims to develop a comprehensive understanding of factors influencing loan default rates within a consumer finance company by employing exploratory data analysis (EDA) techniques, we will analyze historical loan data to identify key attributes that correlate with loan repayment behaviors. The insights derived from this analysis will enable the company to construct a robust credit risk model, aiding in the assessment of loan applicants and mitigating financial losses due to defaults. Ultimately, this project seeks to optimize lending decisions and enhance overall portfolio performance.

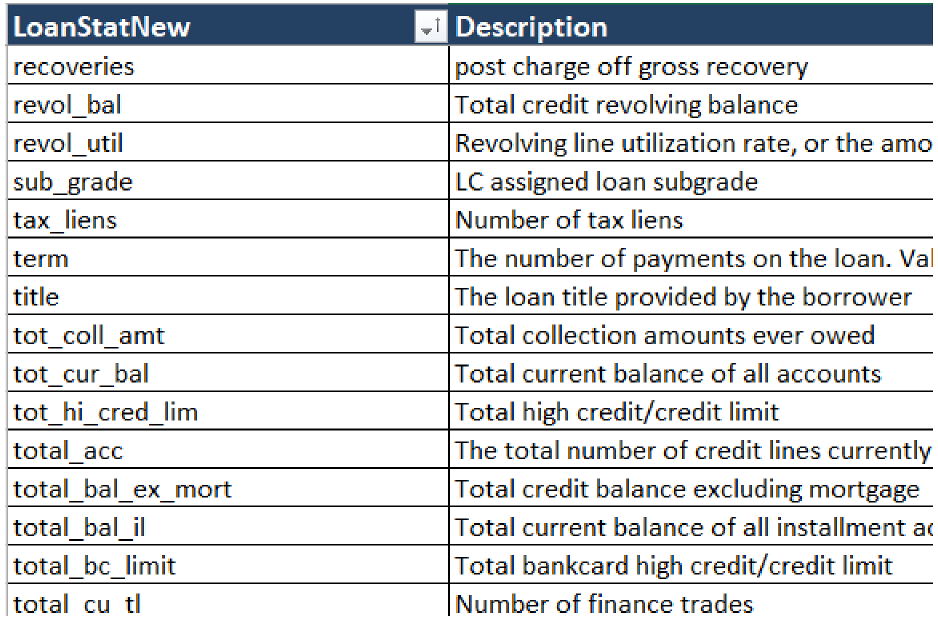
Step for Analysis

The primary goal is to uncover underlying patterns & find the insights & trends on basis of data provided.



Data Loading

We will be loading data for analyzing, Lending Club provided us with historical data on their customers. This dataset includes details about the borrower's past credit history & information about their loans from Lending Club. The dataset is extensive, containing over 39717 records & 111 columns, giving large volume & variety of information to analyze. This data allowed us to identify key factors & relationships that could influence a borrower's for fully paying or defaulting the payment of loan in their agreed tenure.



LoanStatNew	Description
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amo
sub_grade	LC assigned loan subgrade
tax_liens	Number of tax liens
term	The number of payments on the loan. Val
title	The loan title provided by the borrower
tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all accounts
tot_hi_cred_lim	Total high credit/credit limit
total_acc	The total number of credit lines currently
total_bal_ex_mort	Total credit balance excluding mortgage
total_bal_il	Total current balance of all installment ac
total_bc_limit	Total bankcard high credit/credit limit
total_cu_tl	Number of finance trades

To effectively predict a borrower's potential default, we focused on columns which will directly or indirectly impact the outcome of analysis.

Data Understanding

Key Variables :

We focused on a subset of critical variables directly related to loan performance & borrower reliability. These include:

- **Customer Demographic:** Variables such as Home Ownership (home_ownership),state of residence (addr_state), employment length (emp_length), job title (emp_title) help in understanding borrower profiles.
- **Financial Metrics:** Variables like annual income (annual_inc), debt-to-income ratio (dti), and credit utilization (revol_util)offer insights into the financial health of the borrowers.
- **Credit History:** Metrics such as the delinq_2yrs) ,earliest_cr_line , fico_range_high & fico_range_low provide a snapshot of the borrower's creditworthiness.
- **Loan Characteristics**: Information on loan amounts (funded_amnt, funded_amnt_inv), loan grade (grade), Sub Grade(sub_grade), Term(term), Loan Date (issue_d), Purpose of Loan (purpose), Verification Status (verification_status), Interest Rate (int_rate), Installment (installment), Public Records Bankruptcy (public_rec_bankruptcy) and loan descriptions (desc) helps to assess the loan's risk profile.

Data Cleaning & Pre-processing

Data Quality: To ensure robust analysis, the data was subjected to thorough cleaning. This included handling missing values, correcting inconsistencies, and ensuring uniformity across all variables. The step by step process for it as follows:

- Checking for null values in the dataset: There're many columns with null values. So they had to be dropped as they won't play a role in the analysis of the dataset. Roughly 48% of the columns were dropped.
  - **54 columns contain** NA values only, and these columns will be removed namely acc_open_past_24mths, all_util, annual_inc_joint, avg_cur_bal, bc_open_to_buy, bc_util, dti_joint, il_util, inq_fi, inq_last_12m, max_bal_bc, mo_sin_old_il_acct, mo_sin_old_rev_tl_op, mo_sin_rcnt_rev_tl_op, mo_sin_rcnt_tl, mort_acc, mths_since_last_major_derog, mths_since_rcnt_il, mths_since_recent_bc, mths_since_recent_bc_dlq, mths_since_recent_inq, mths_since_recent_revol_delinq, num_accts_ever_120_pd, num_actv_bc_tl, num_actv_rev_tl, num_bc_sats, num_bc_tl, num_il_tl, num_op_rev_tl, num_rev_accts, num_rev_tl_bal_gt_0, num_sats, num_tl_120dpd_2m, num_tl_30dpd, num_tl_90g_dpd_24m, num_tl_op_past_12m, open_acc_6m, open_il_12m, open_il_24m, open_il_6m, open_rv_12m, open_rv_24m, pct_tl_nvr_dlq, percent_bc_gt_75, tot_coll_amt, tot_cur_bal, tot_hi_cred_lim, total_bal_ex_mort, total_bal_il, total_bc_limit, total_cu_tl, total_il_high_credit_limit, total_rev_hi_lim, verification_status_joint

- Checking for columns where missing data is >=50% & removing them. 3 columns were dropped.

- Checking for unique values: If the column has only a single unique value, it does not make any sense to include it as part of our data analysis. We need to find out those columns and drop them from the dataset. 9 columns had such unique values and they were removed

- Checking for duplicated rows in data: No duplicate rows were found.

- Checking for columns with text value & unique ID
  - Columns (member_id,id,url) to be dropped as it unique LC assigned Id for the borrower member, will not help in analysis
  - Column zip_code is a masked data and cannot be used as input for the analysis.
  - Columns (emp_title , title, desc) will be dropped as they contain descriptive text & do not help in analysis.

Data Cleaning & Pre-processing

  - Column pub_rec_bankruptcies has mostly null values, so we are dropping it due to missing value.
  - After Column dropping we are left with 33 columns.

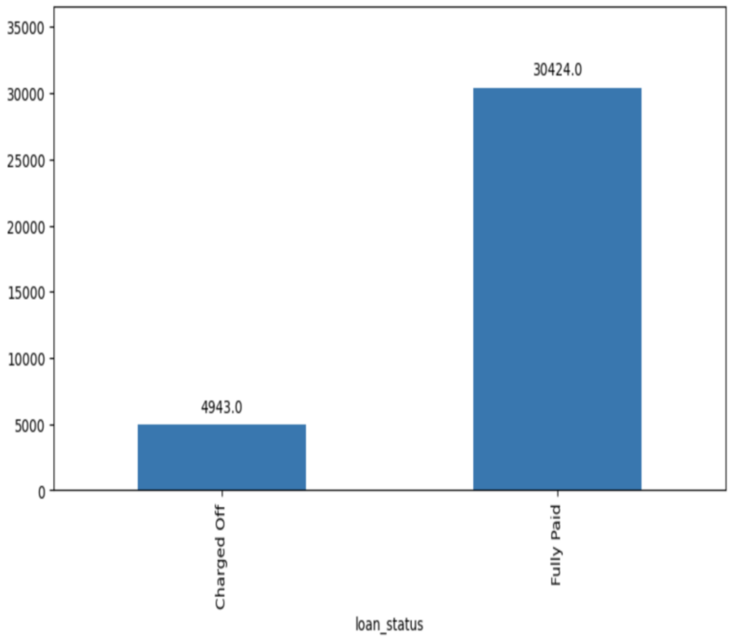
- Filling null values :
  - Column emp_length has null values, replacing it with mode value i.e 10+ years.
  - Column revol_util has null values, replacing it with median i.e 49.3

- Data Conversion:
  - Column term removing months & converting to int.
  - Columns int_rate & revol_util removing % & converting it to float.
  - Column emp_length if <1 year is assumed as 0 & 10+ years assumed as 10 & converting it to int.

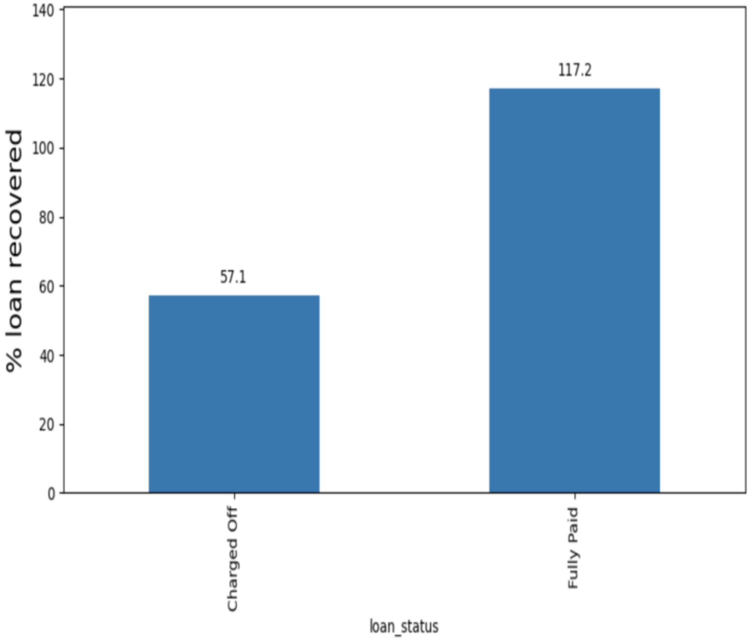
- Derived Columns
  - Column issue_d , getting issue_month & issue_year columns.
  - Column earliest_cr_line getting days, from difference of today’s date & earliest_cr_line.

- Outlier Removal : Calculated the Inter-Quartile Range (IQR) and filtering out the outliers outside of lower and upper bound
  - Funded_amt_inv , upper fence turns around 28k whereas max is 35k, so no need to remove outliers**.**
  - Annual_inc , we need to remove outliers as there is vast difference using 99 percentile we will remove it.

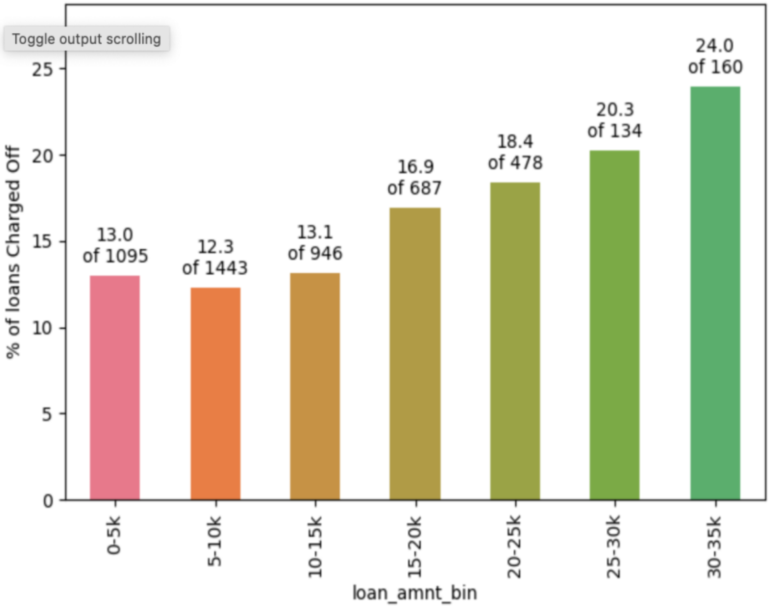
Overall status of loans allotted



Approx 14% of loans in the datasets are defaulted.



57% of the amount is recovered from charged off loan, while 17% profit is made on Fully Paid loans.



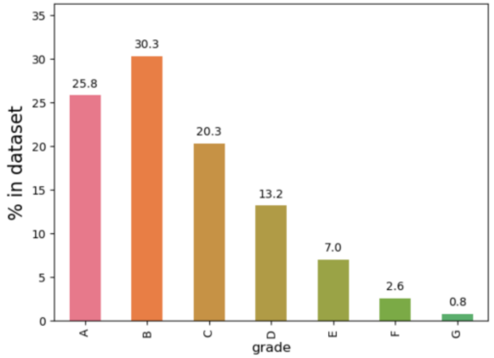
As the loan amount increases , the % of charged off loan also increases. Hence, higher the loans, the risk of defaulters increases.

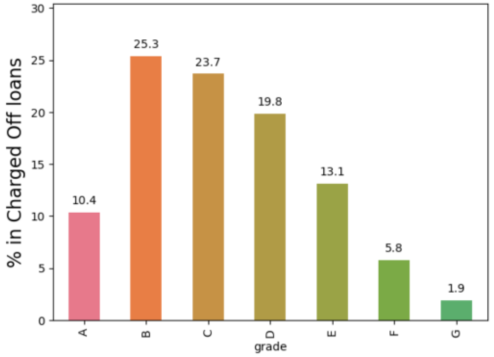
Univariate Analysis – Loan Grade Analysis

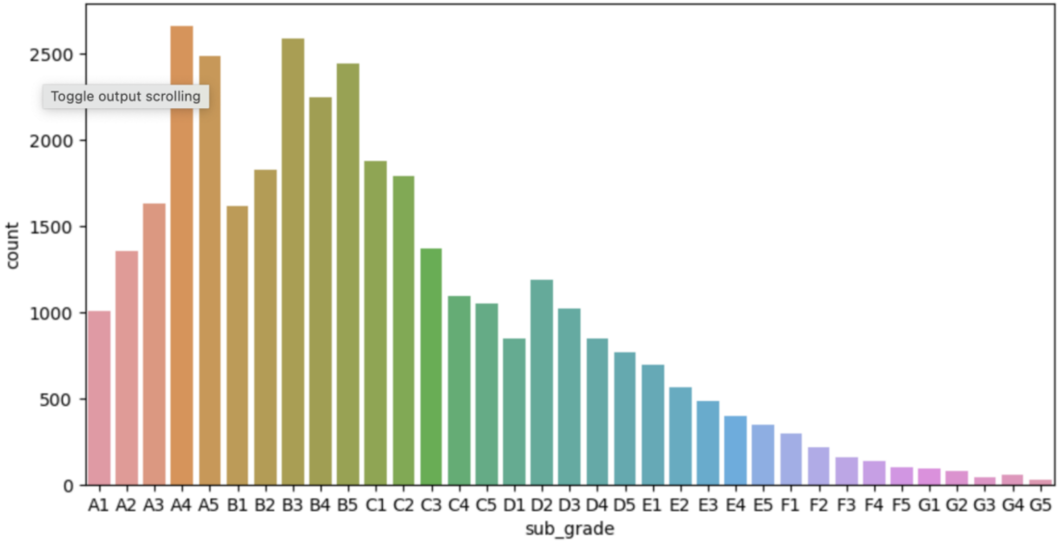
- Grade A and B loans are safe. The percentages in full dataset are much higher than percentages in Charged Off loans.

- Grade D, E, F, G loans are less safe. We should plot grade by percentage Charged Off by category

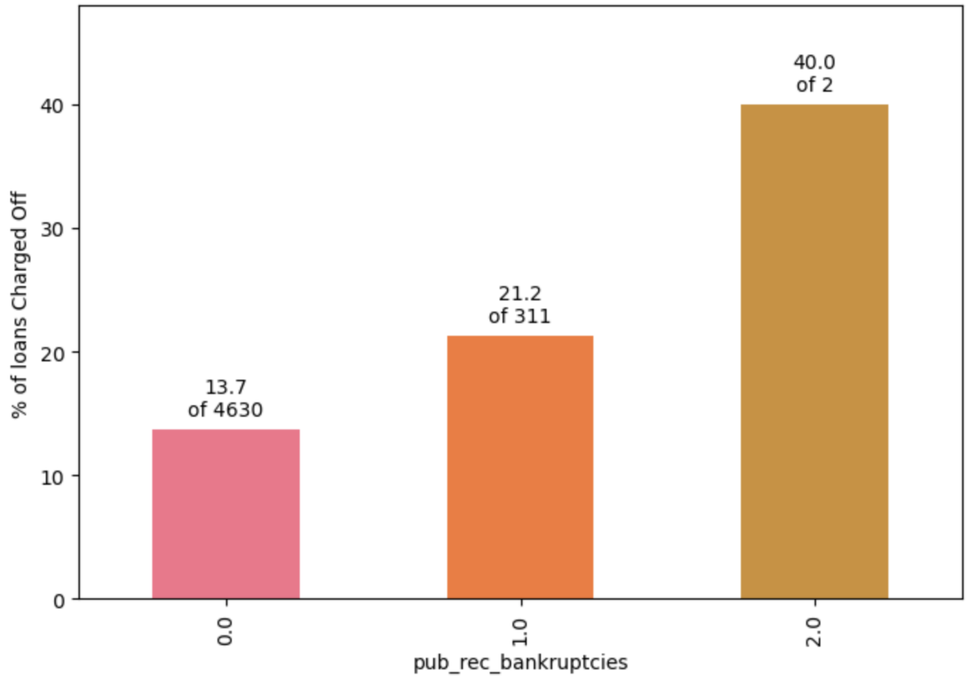
- Grading system of LC works properly.



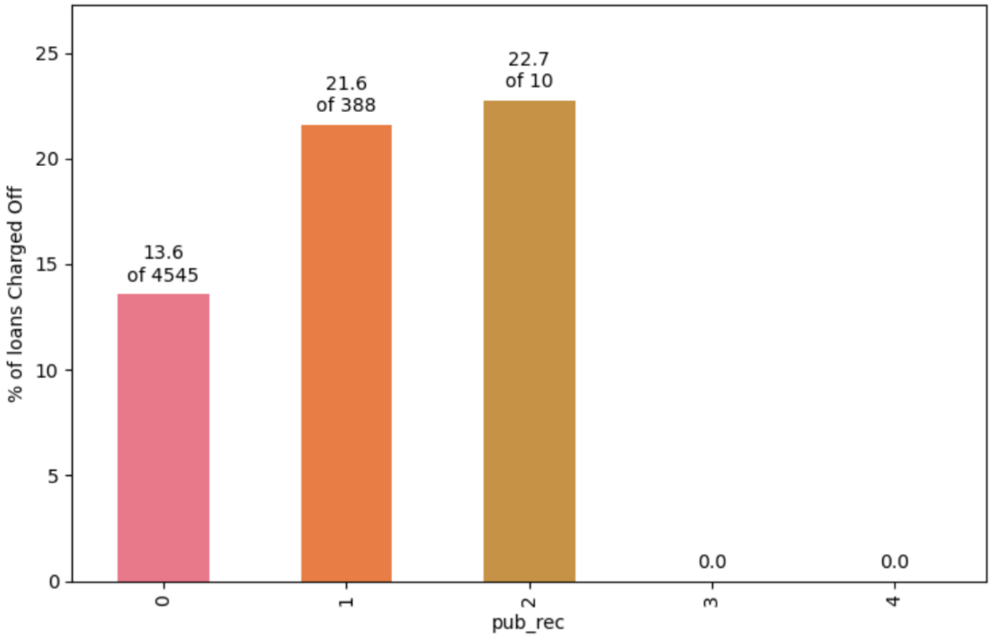




Univariate Analysis on prior Bad Records

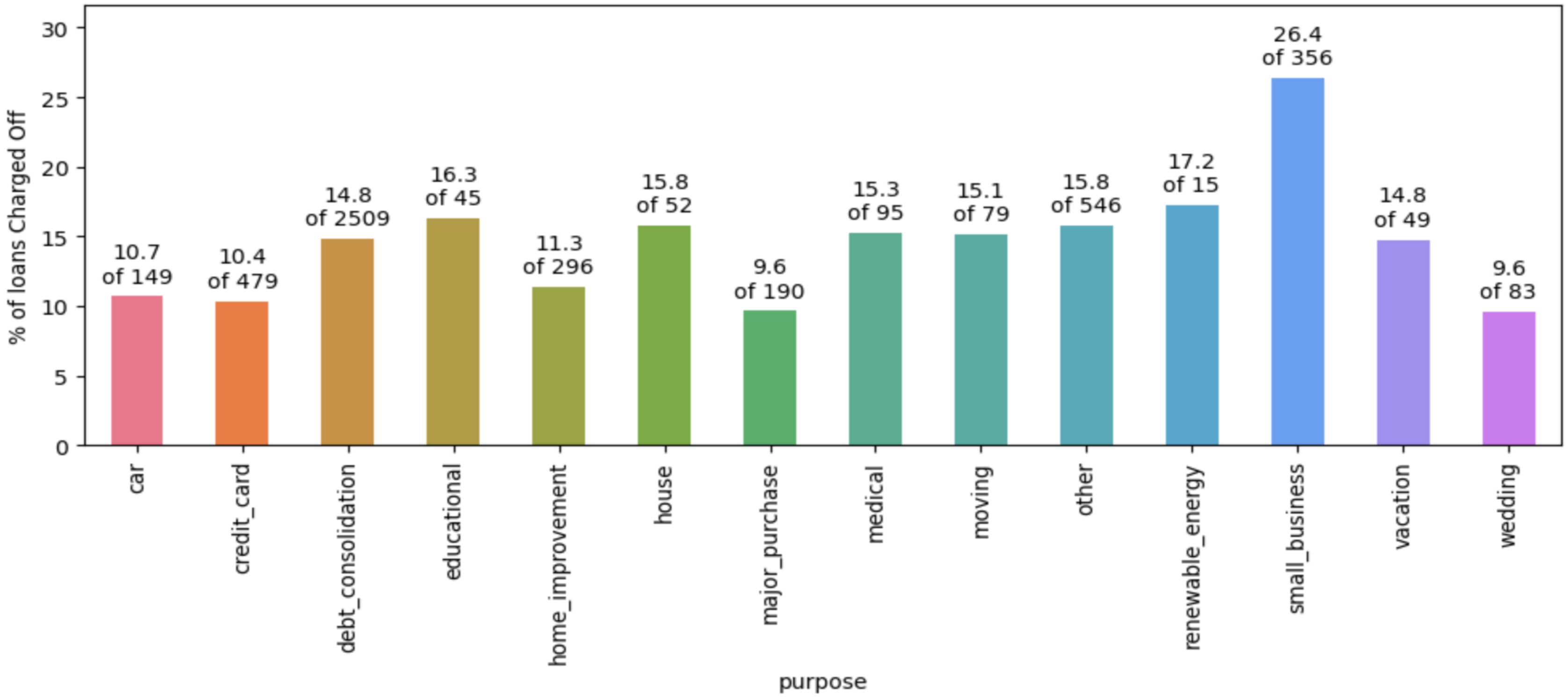


The percentage of Charged Off loans is markedly higher when the borrower has a prior record of bankruptcy.



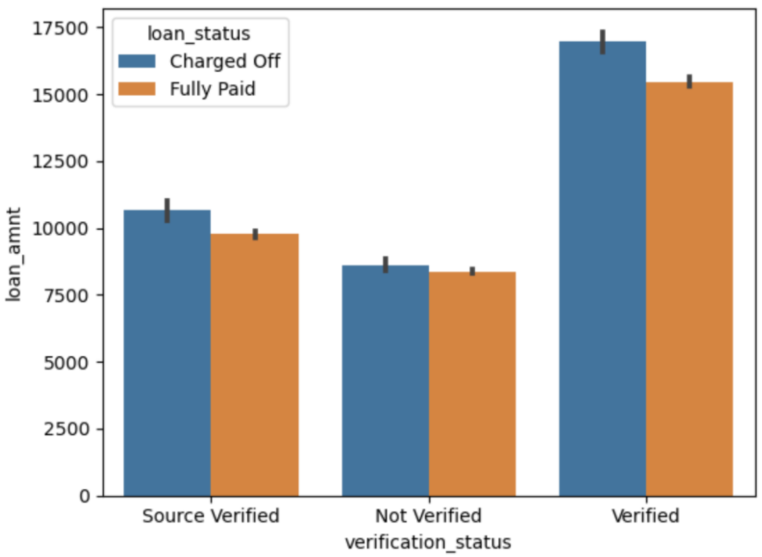
- 94% have no Public derogatory records. 5% have 1 derogatory record.
- Having even 1 derogatory record increases the chances of Charge Off significantly.

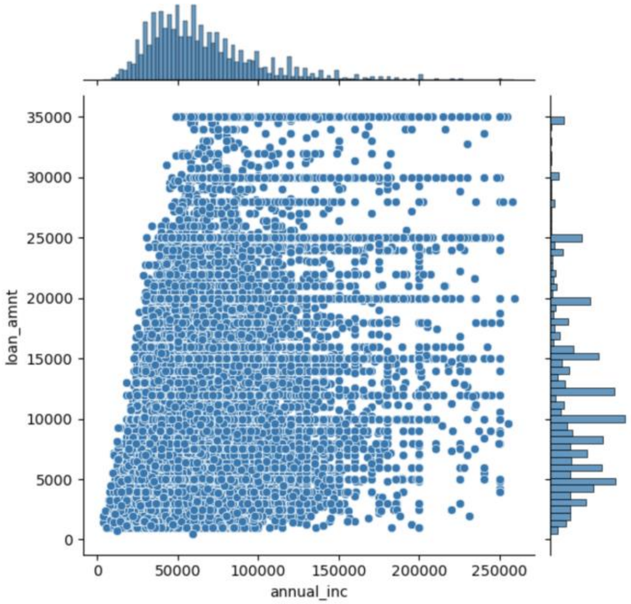
Segmented Analysis



26% of loans for small business are Charged Off. Making them the riskiest purpose.

Bivariate Analysis

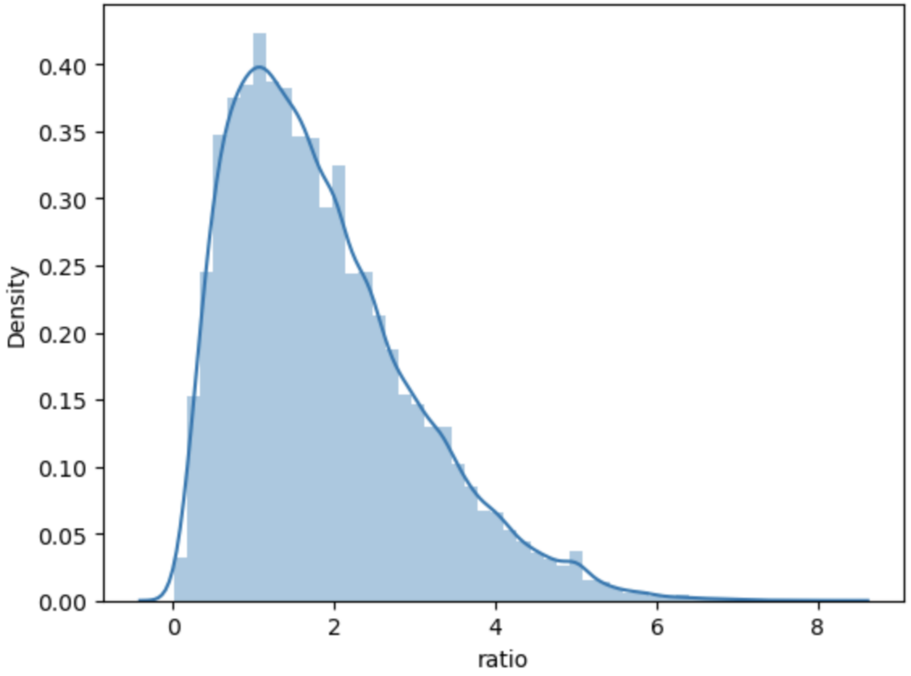


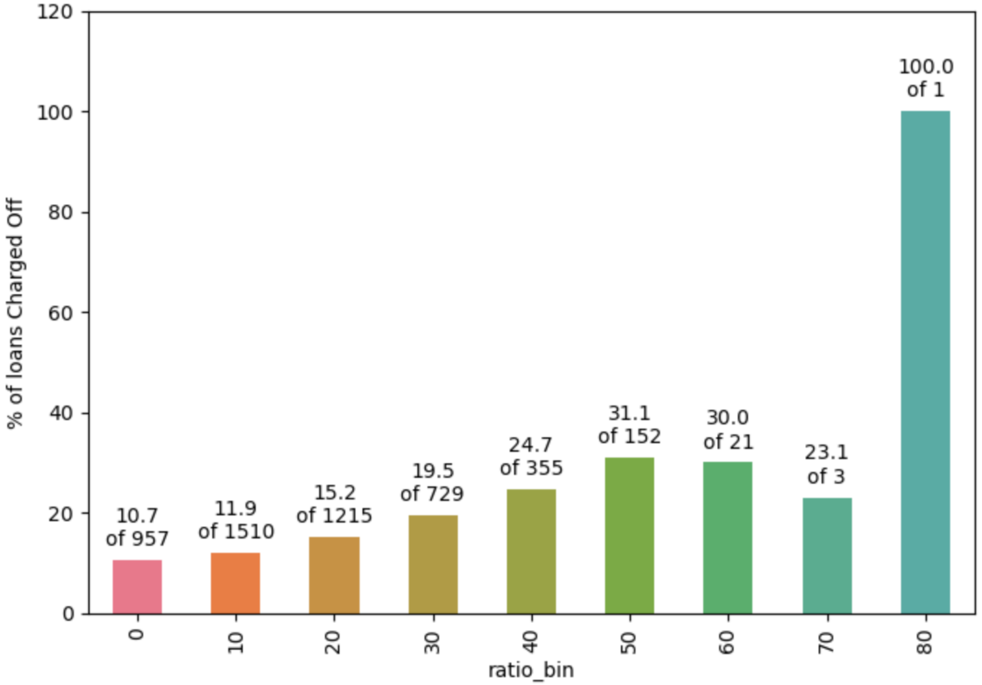


- Higher loan amounts are Verified more often.
- We already know that larger loans are less in number but see a higher charge off rate.
- This, combined with previous observation, explains why verified loans see a higher rate of default. It's not the verified status per se, it's the fact that higher loan amounts are riskier and are also verified more often by Lending Club.

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

Derived Metrics



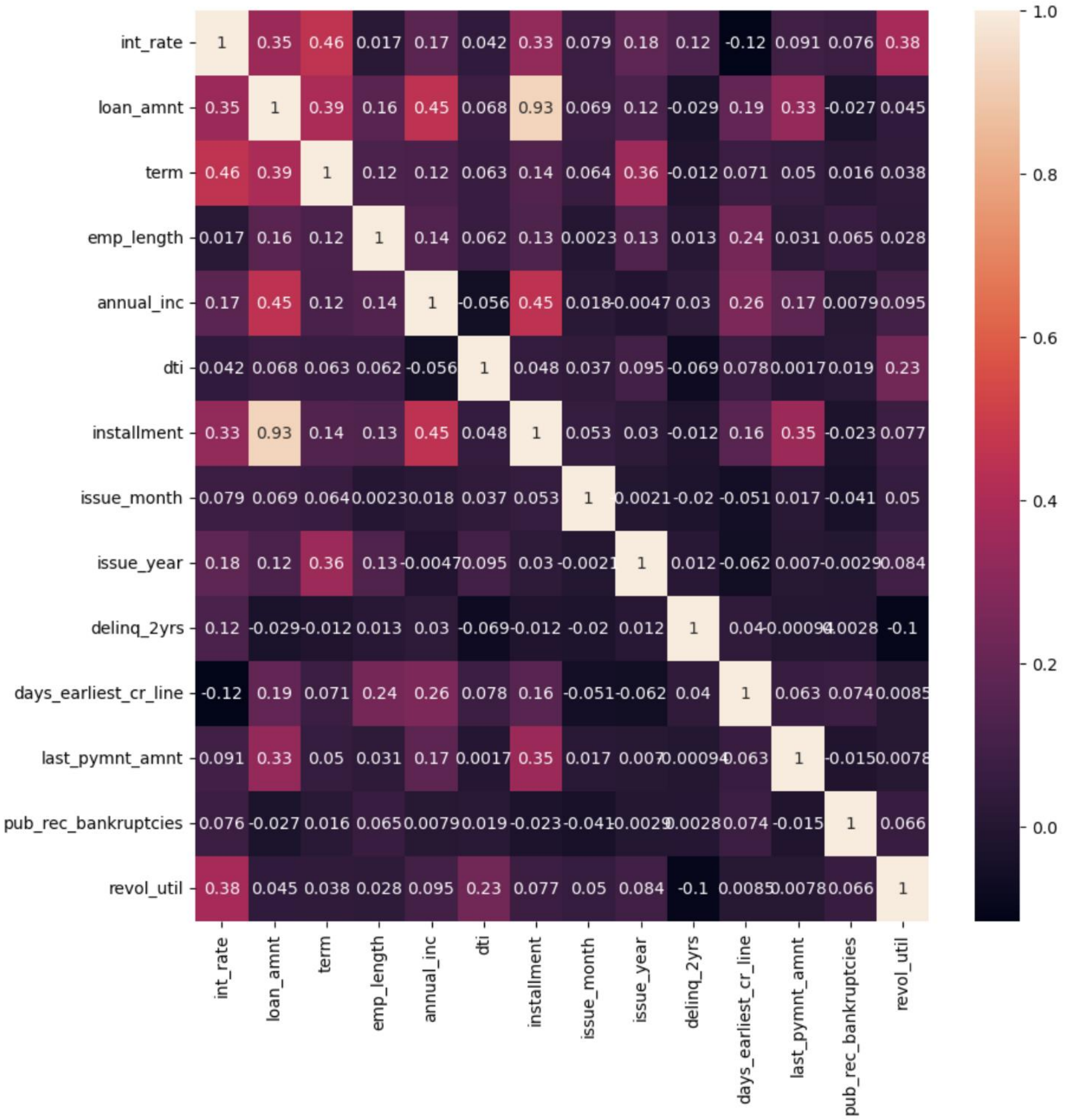


Observation:

- As long as loan amount is less than 20% of annual income, defaults are low.
- Loan amounts of 30% of annual income or higher see a high rate of default.

Correlation Analysis

- installment has strong correlation with funded_amnt, loan_amount
- term has strong correlation with int_rate.
- annual_inc has strong correlation with loan_amnt.
- dti and emp_length has weak correlation with most of the fields.
- dti has strong negative correlation with annual_inc



	int_rate	loan_amnt	term	emp_length	annual_inc	dti	installment	issue_month	issue_year	delinq_2yrs	days_earliest_cr_line	last_pymnt_amnt	pub_rec_bankruptcies	revol_util
int_rate	1	0.35	0.46	0.017	0.17	0.042	0.33	0.079	0.18	0.12	-0.12	0.091	0.076	0.38
loan_amnt	0.35	1	0.39	0.16	0.45	0.068	0.93	0.069	0.12	-0.029	0.19	0.33	-0.027	0.045
term	0.46	0.39	1	0.12	0.12	0.063	0.14	0.064	0.36	-0.012	0.071	0.05	0.016	0.038
emp_length	0.017	0.16	0.12	1	0.14	0.062	0.13	0.0023	0.13	0.013	0.24	0.031	0.065	0.028
annual_inc	0.17	0.45	0.12	0.14	1	-0.056	0.45	0.018	-0.0047	0.03	0.26	0.17	0.0079	0.095
dti	0.042	0.068	0.063	0.062	-0.056	1	0.048	0.037	0.095	-0.069	0.078	0.0017	0.019	0.23
installment	0.33	0.93	0.14	0.13	0.45	0.048	1	0.053	0.03	-0.012	0.16	0.35	-0.023	0.077
issue_month	0.079	0.069	0.064	0.0023	0.018	0.037	0.053	1	-0.0021	-0.02	-0.051	0.017	-0.041	0.05
issue_year	0.18	0.12	0.36	0.13	-0.0047	0.095	0.03	-0.0021	1	0.012	-0.062	0.007	-0.0029	0.084
delinq_2yrs	0.12	-0.029	-0.012	0.013	0.03	-0.069	-0.012	-0.02	0.012	1	0.04	-0.00094	-0.0028	-0.1
days_earliest_cr_line	-0.12	0.19	0.071	0.24	0.26	0.078	0.16	-0.051	-0.062	0.04	1	0.063	0.074	0.0085
last_pymnt_amnt	0.091	0.33	0.05	0.031	0.17	0.0017	0.35	0.017	0.007	-0.00094	0.063	1	-0.015	0.0078
pub_rec_bankruptcies	0.076	-0.027	0.016	0.065	0.0079	0.019	-0.023	-0.041	-0.0029	-0.0028	0.074	-0.015	1	0.066
revol_util	0.38	0.045	0.038	0.028	0.095	0.23	0.077	0.05	0.084	-0.1	0.0085	0.0078	0.066	1

Insights & Trends

- Loans allotted to Grade A and Grade B are safer.
- Loans allotted to people with prior bad records are riskier.
- Loans allotted to small business are of riskiest purpose.
- Loans background verification should be stricter, as percentage of charged off loan increases for non verified customer.
- As long as loan amount is less than 20% of annual income, default rate is low.