Lending Club Case Study

Saket Lalpura

Sandip Joshi

Background Information

This 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.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters’.

Problem Statement

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 utilize this knowledge for its portfolio and risk assessment.

If one is able to 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.

Understanding of data

Types of variables

- Customer information
- Loan related information
- Customer related information

Customer information
Member Id
Customer Id
Employment tittle
Employment length
Grade
Home Ownership

Loan related
Id
Loan Amount
Funder Amount
Term
Interest Rate

Customer Related Information
Verification
Dit
Delinq_2yrs
Inq_last_6mths
Revol uti

Cleaning of Data

First we check for duplicates of id, there are none so we move to next step

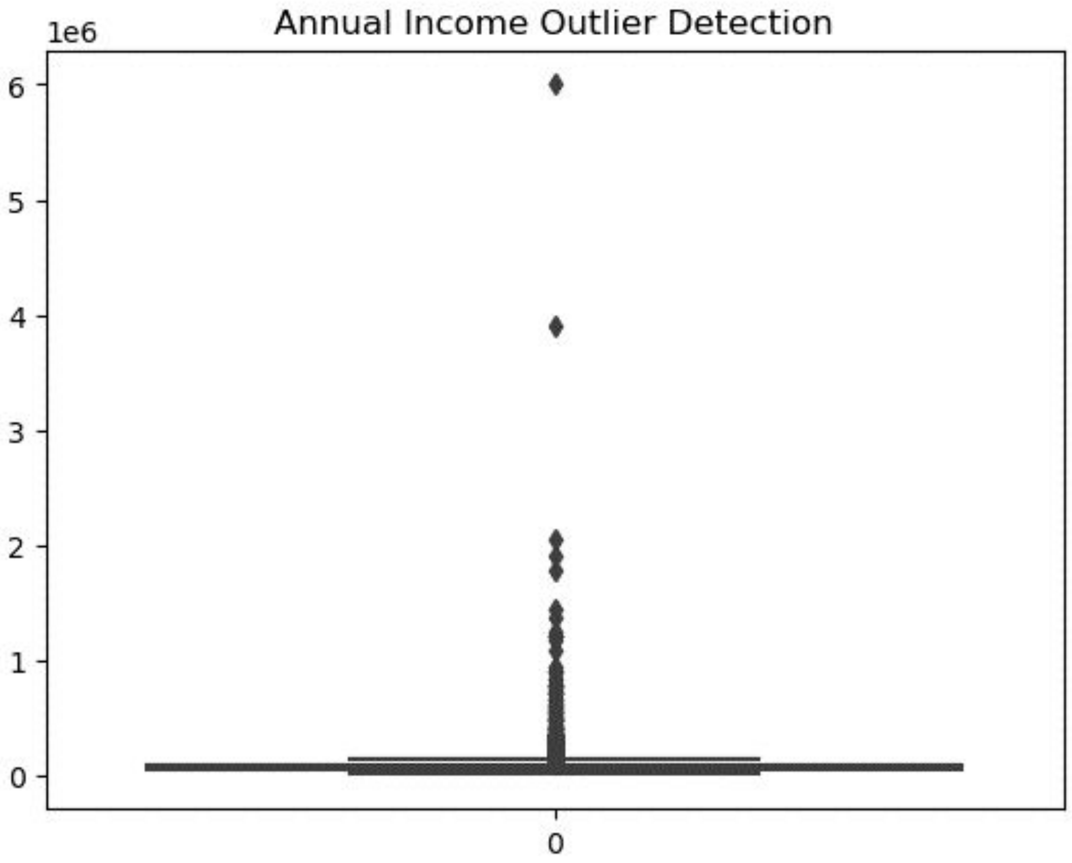
which is we need to remove the row which have a “Current” value for loan status, Since this loan is running at the time this data is incomplete as paid off or cleared off prospective

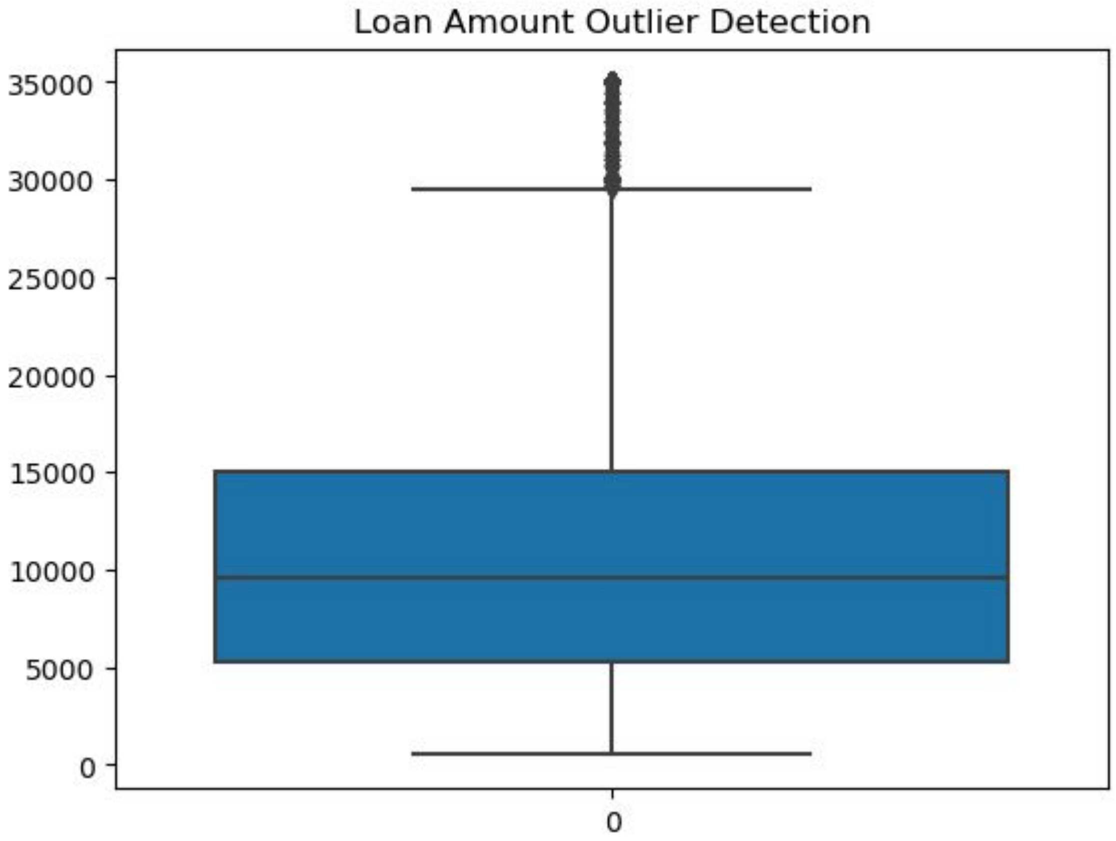
After that we need to remove the following columns which does not have any data or only 1 data or less data variation or data which is not related to or depending upon another column so those column are not going to make any impact over the analysis.

Name of Columns							
funded_amnt_inv	next_pymnt_d	tot_coll_amt	open_rv_12m	avg_cur_bal	mort_acc	num_bc_tl	num_tl_op_past_12m
pymnt_plan	collections_12_mths_ex_med	tot_cur_bal	open_rv_24m	bc_open_to_buy	mths_since_recent_bc	num_il_tl	pct_tl_nvr_dlq
url	mths_since_last_major_derog	open_acc_6m	max_bal_bc	bc_util	mths_since_recent_bc_dlq	num_op_rev_tl	percent_bc_gt_75
zip_code	policy_code	open_il_6m	all_util	chargeoff_within_12_mths	mths_since_recent_inq	num_rev_accts	tax_liens
addr_state	application_type	open_il_12m	total_rev_hi_lim	delinq_amnt	mths_since_recent_revol_delinq	num_rev_tl_bal_gt_0	tot_hi_cred_lim
initial_list_status	annual_inc_joint	open_il_24m	inq_fi	mo_sin_old_il_acct	num_accts_ever_120_pd	num_sats	total_bal_ex_mort
out_prncp	dti_joint	mths_since_rcnt_il	total_cu_tl	mo_sin_old_rev_tl_op	num_actv_bc_tl	num_tl_120dpd_2m	total_bc_limit
out_prncp_inv	verification_status_joint	total_bal_il	inq_last_12m	mo_sin_rcnt_rev_tl_op	num_actv_rev_tl	num_tl_30dpd	total_il_high_credit_limit
total_pymnt_inv	acc_now_delinq	il_util	acc_open_past_24mths	mo_sin_rcnt_tl	num_bc_sats	num_tl_90g_dpd_24m	

Outliers Detection

The next step of cleaning data is taking out the outliers from the data base





So we took 2 important parameters which is Annual income and Loan Amount take out the outliers

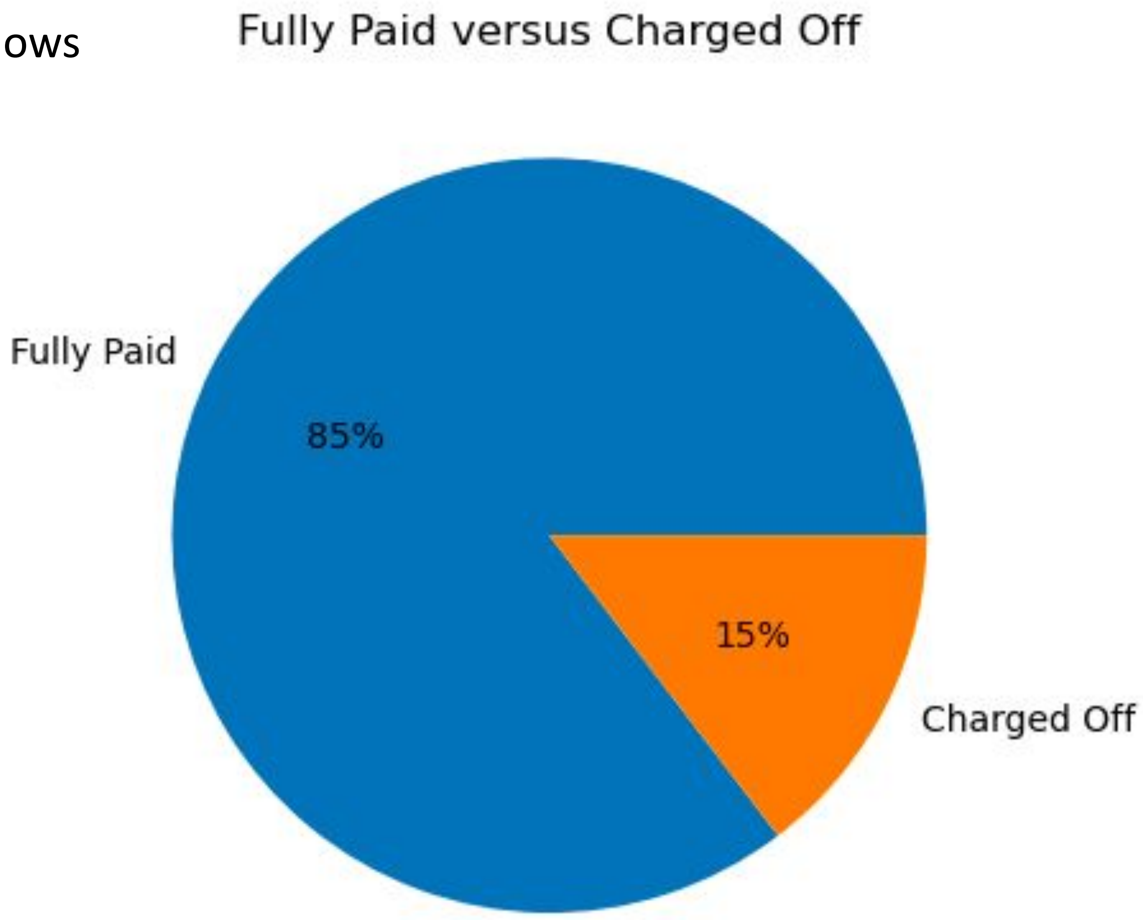
Basic Analysis

Now our data is ready for analysis

We have found the ratio of fully paid and charged off as follows

1. Fully paid customers
   85 %

2. Charged off customers
   15%



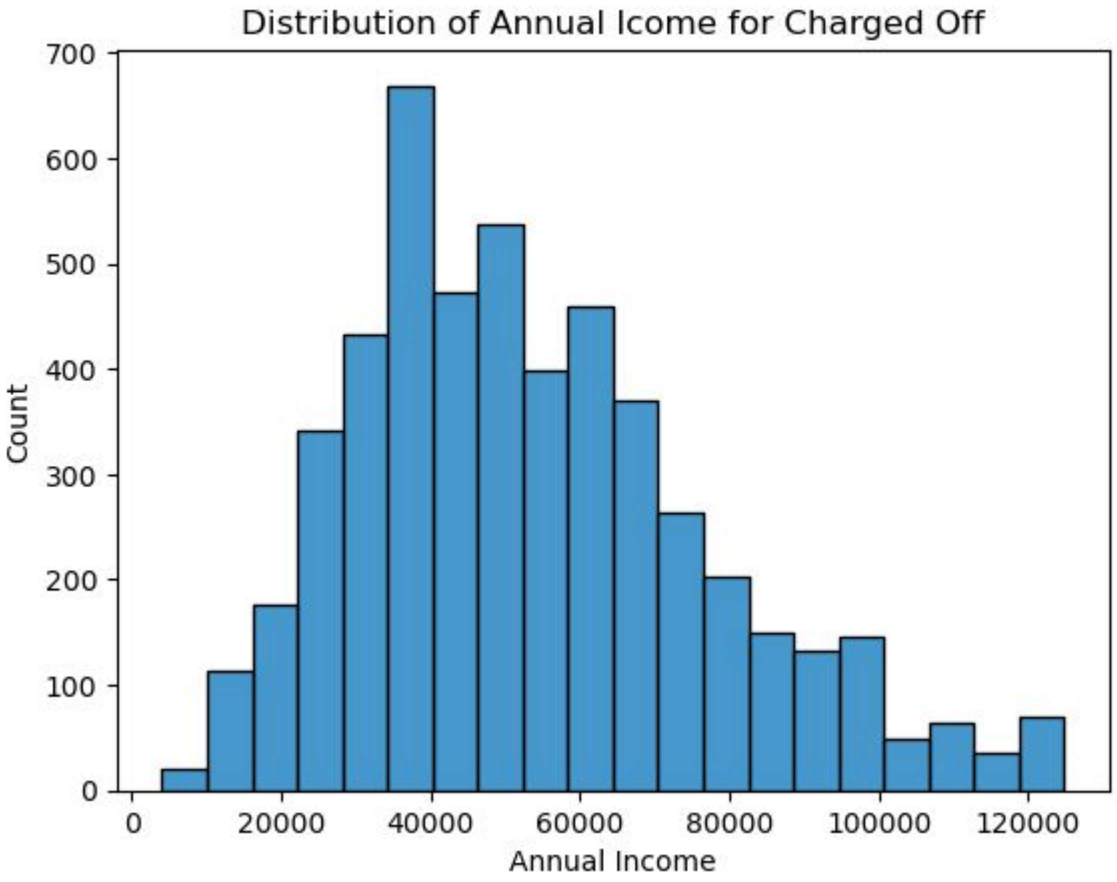
Univariate Analysis 1

Frequency of Defaulter based on their Annual Income

The histogram shows a clear insight where the highest bucket (near less than 40,000) shows a higher number of defaulters.

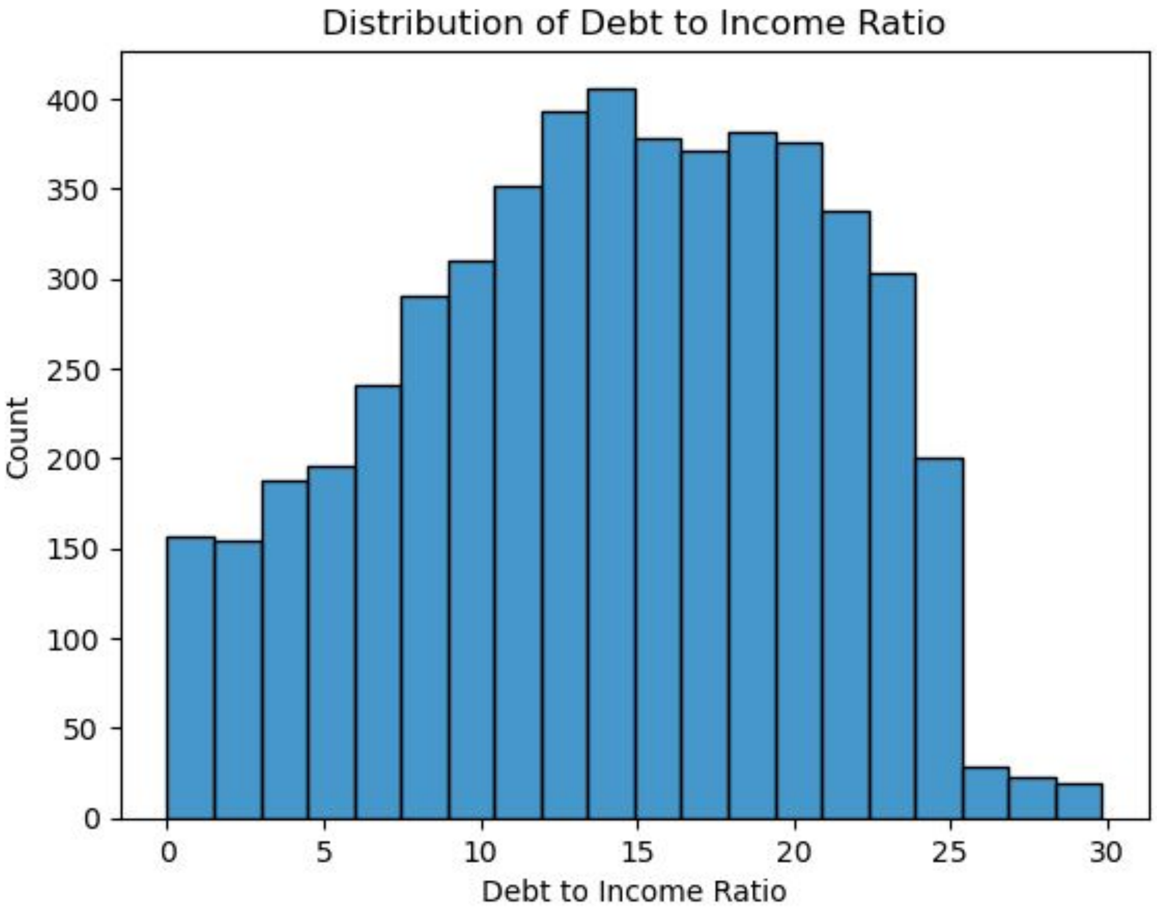
The frequency of defaulters sharply increases with income up to 40,000, after which it declines, suggesting that even higher income brackets may still experience defaults but less frequently.

This trend highlighting the complexity of financial behavior across different income levels.



Univariate Analysis 2

Frequency of Defaulter based on their Debt to Income Ratio



The histogram shows that the debt-to-income ratio peaks between 10 and 15, with counts from 300 to 400, suggesting this range has the highest concentration of defaulters.

A curve shape is observed as the ratio increases, with counts increase from 0 to 15 before sharply declining after 25, indicating that extremely huge ratios correlate with fewer defaulters.

This pattern show that a moderate debt-to-income ratio may heads to higher default rates, however extremely high ratios may not strongly indicate increased defaults.

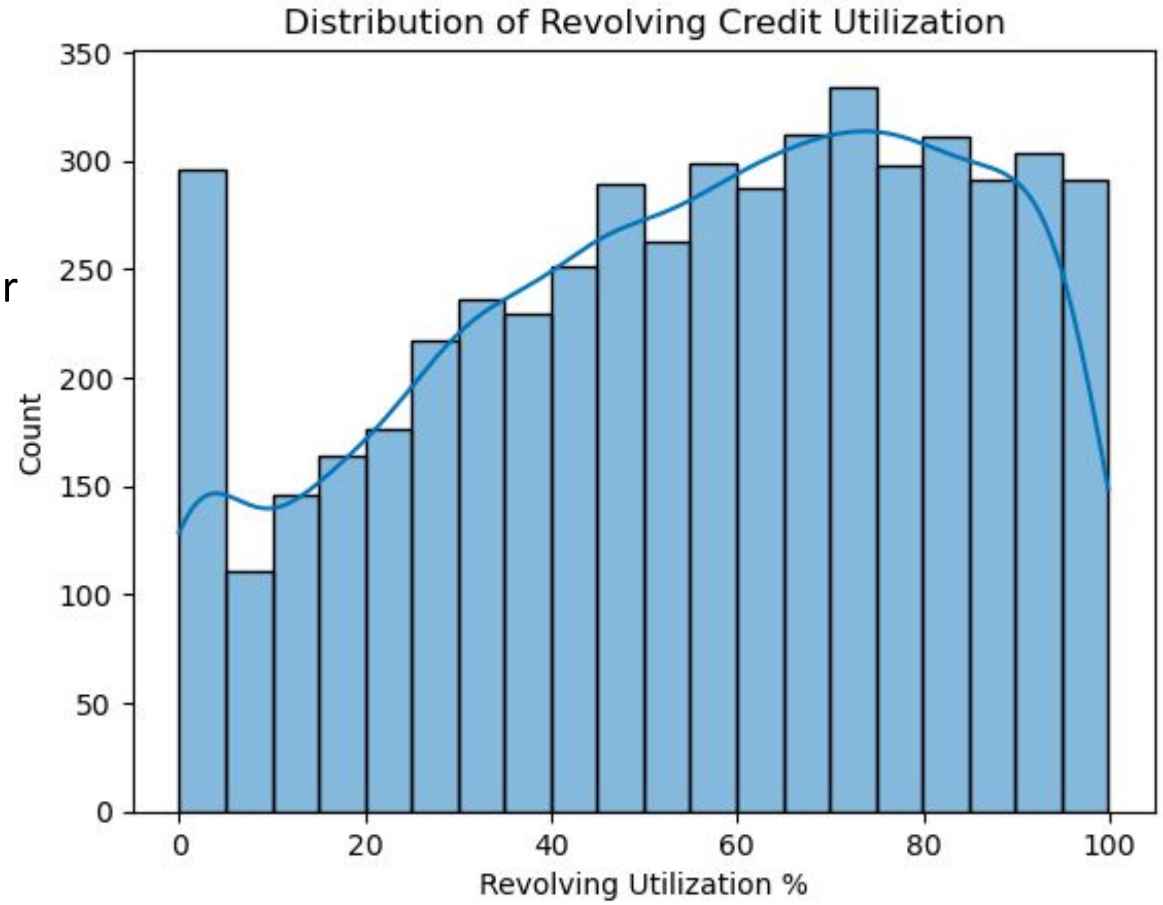
Univariate Analysis 3

Frequency of Defaulter based on their Revolving Utilization %

The histogram shows a higher concentration of accounts with 0% utilization (300 counts), followed by a sharp increase at 100% utilization, highlighting a key threshold for defaulter behavior.

The highest frequency occurs between 60% to 80% utilization (320 counts), suggesting this range may be a critical risk zone for defaults, while counts decline again at 80% and 100%.

This pattern indicates that moderate utilization levels (20% to 60%) correlate with increasing counts of defaulters.



Bivariate Analysis 2

Difference between fully paid and default customers based on Loan Interest Rate

Creating a derived integer variable column
for people who have defaulted or not

- Value 1 for Charged off

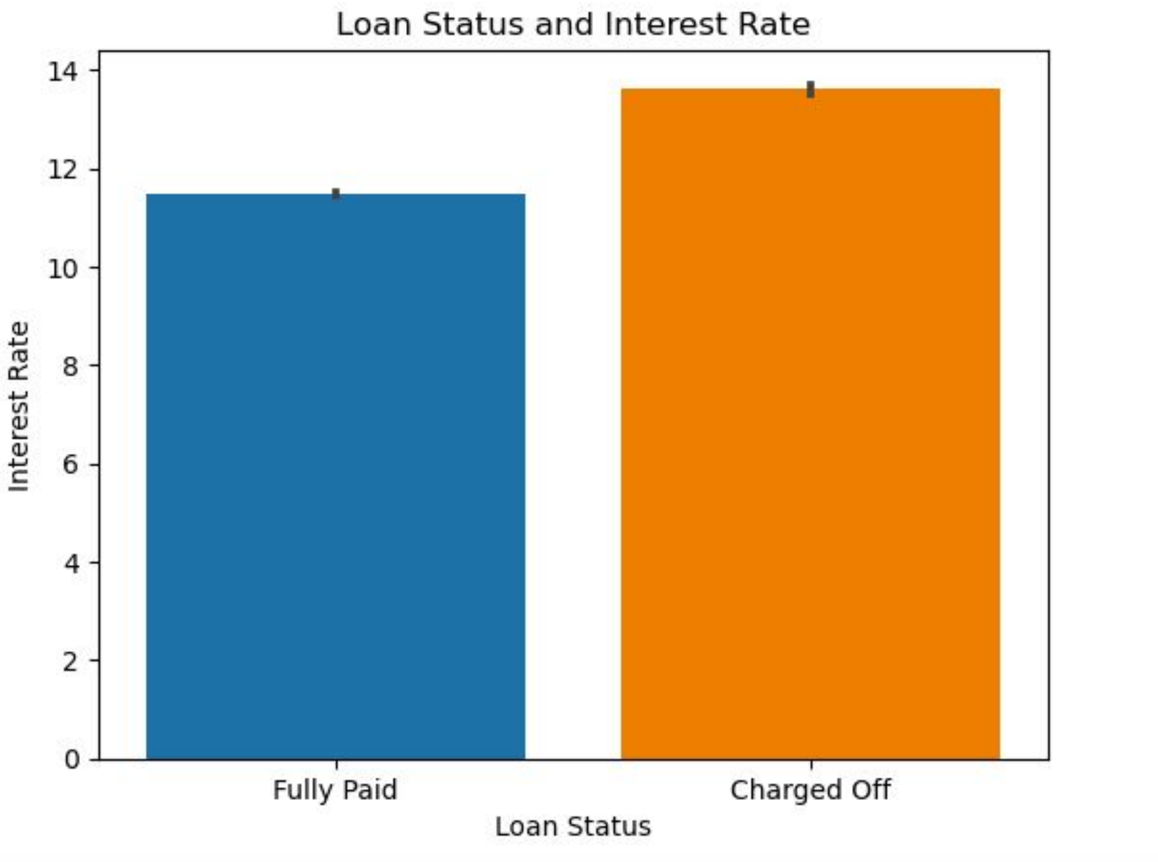
And

- Value 0 for Fully Paid

to make the Analysis easier
(one of the driving variables for loan default)

This Bivariate analysis show significant difference between customers of fully paid and defaulter on based on interest rate

Higher interest rate might have chance customer may tend to default

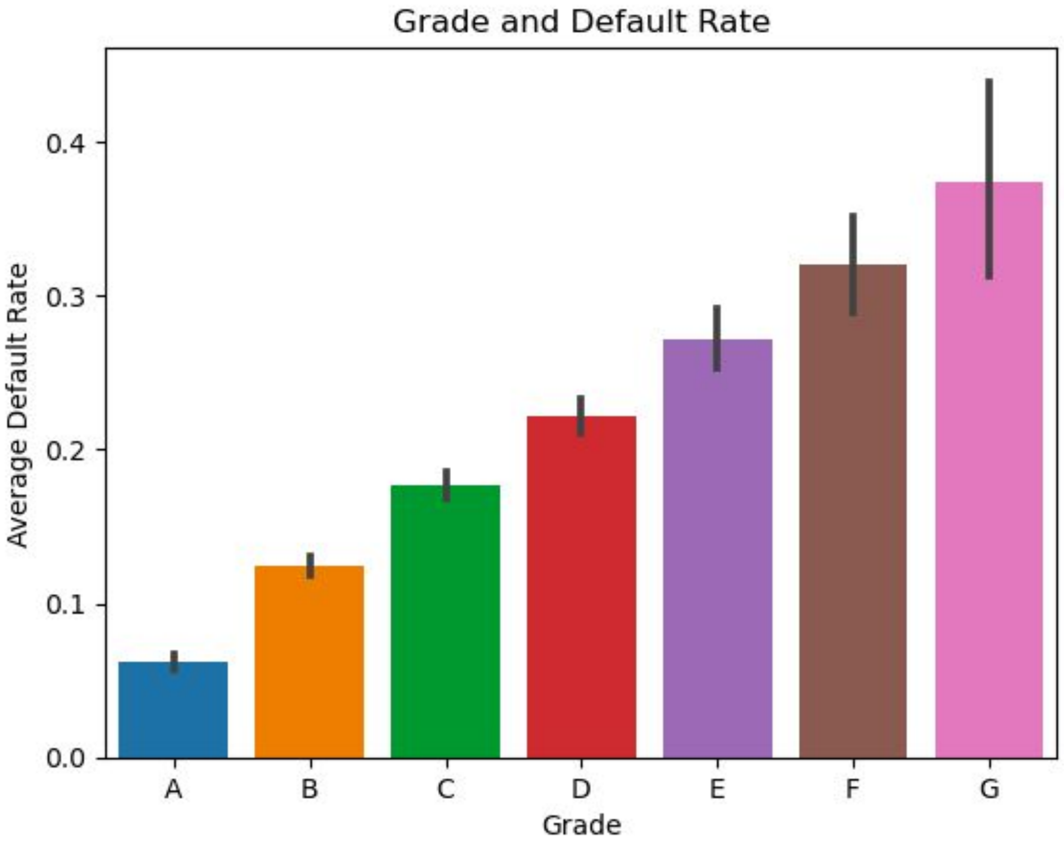


Bivariate Analysis 3

Comparison between “Grades” based on Average Default Rate

This bivariate analysis shows as lower the grade customer have high chance to make load default.

Average default rate gradually increase with grades shows the trend of default rate

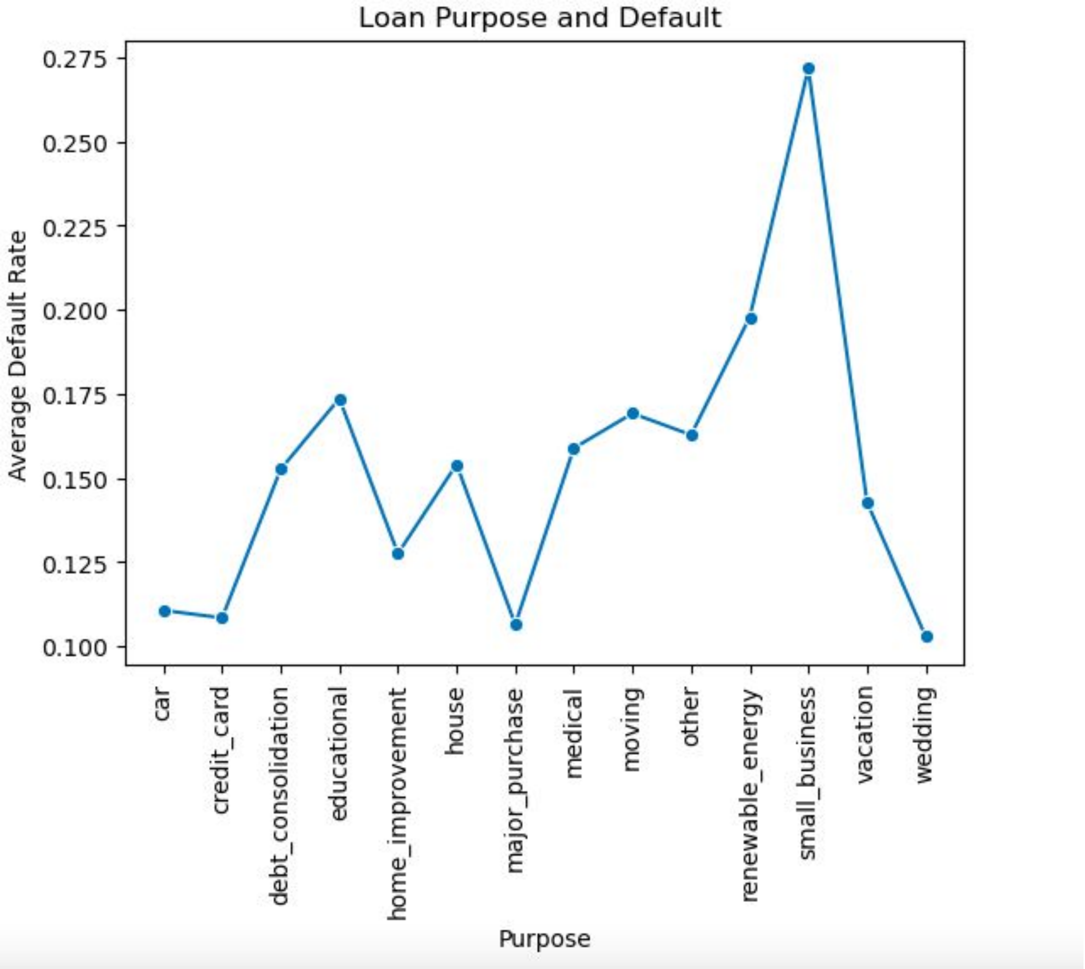


Bivariate Analysis 4

Comparison between “Purposes” based on Average Default Rate

This bivariate analysis shows customer with purpose of small business have higher rates of default the loan.

The second highest rate of customer have purpose of renewable energy followed by education loan.

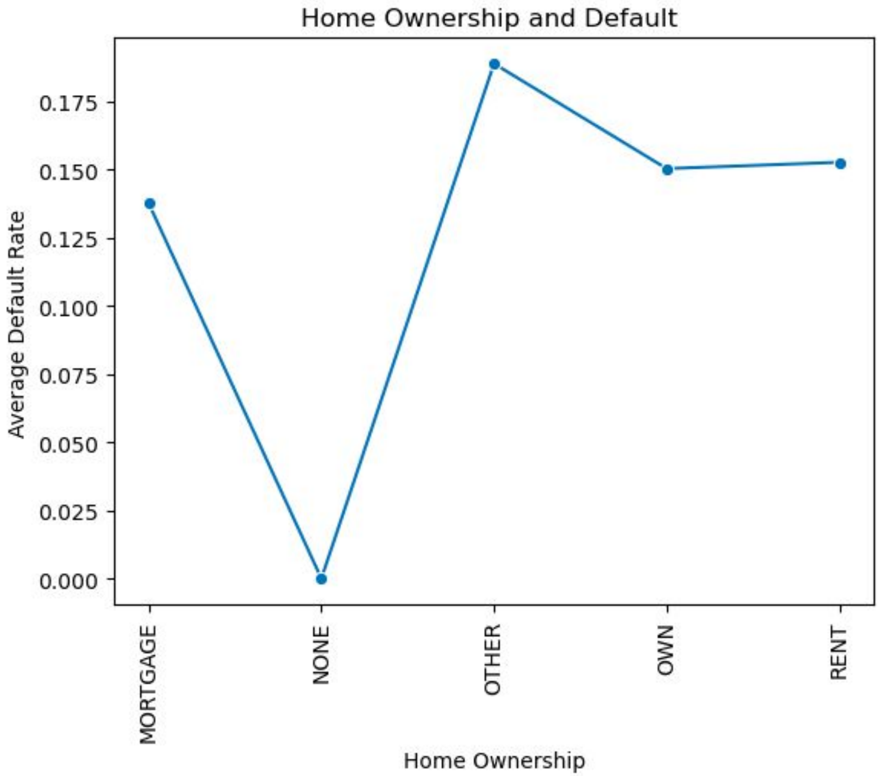


Bivariate Analysis 5

Comparison between “Home Ownership” based on Average Default Rate

This bivariate analysis shows the impact of home ownership does not have any significant variation in terms of average default rate.

Surprisingly None home ownership have zero default rate.



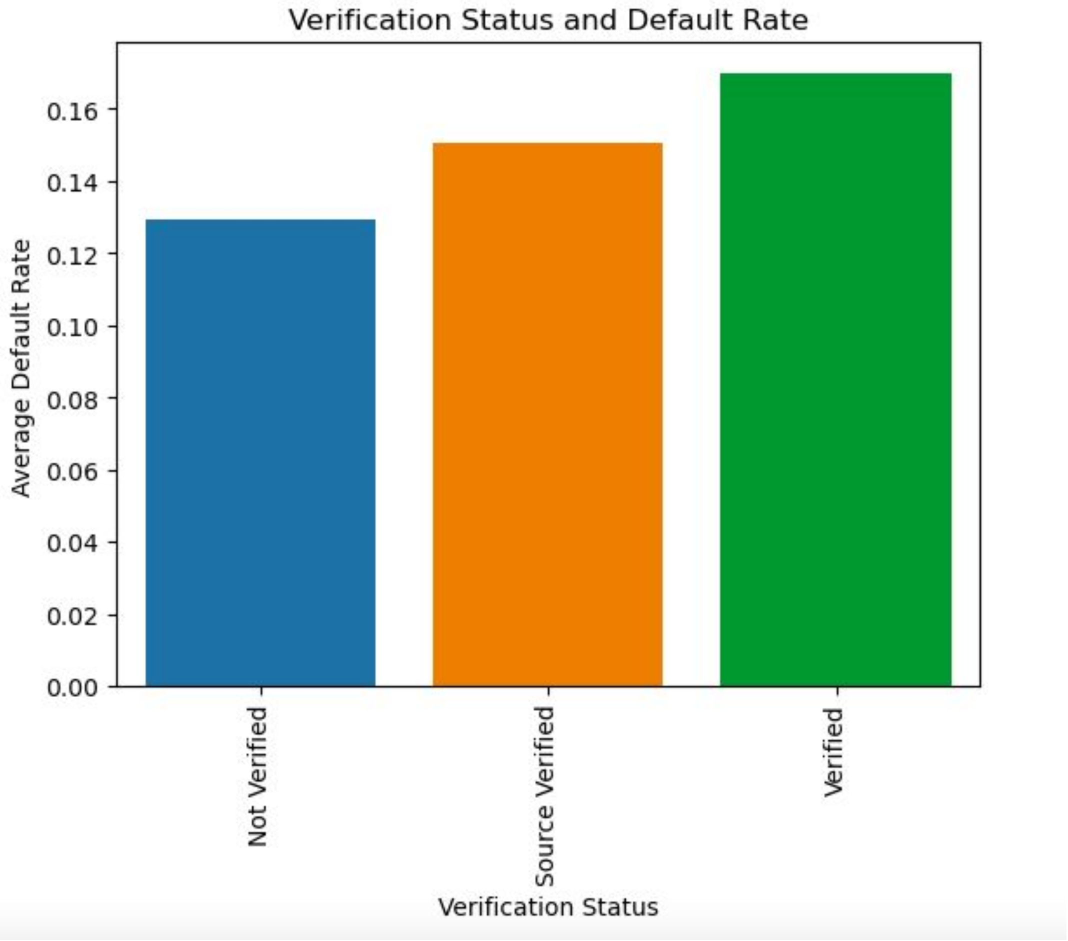
Bivariate Analysis 5

Comparison between "Verification Status" based on Average Default Rate.

This bivariate analysis shows the not-verified loan have less default rate.

Verified loan have high default rate.

And Source verified loan have default rates have between above two.

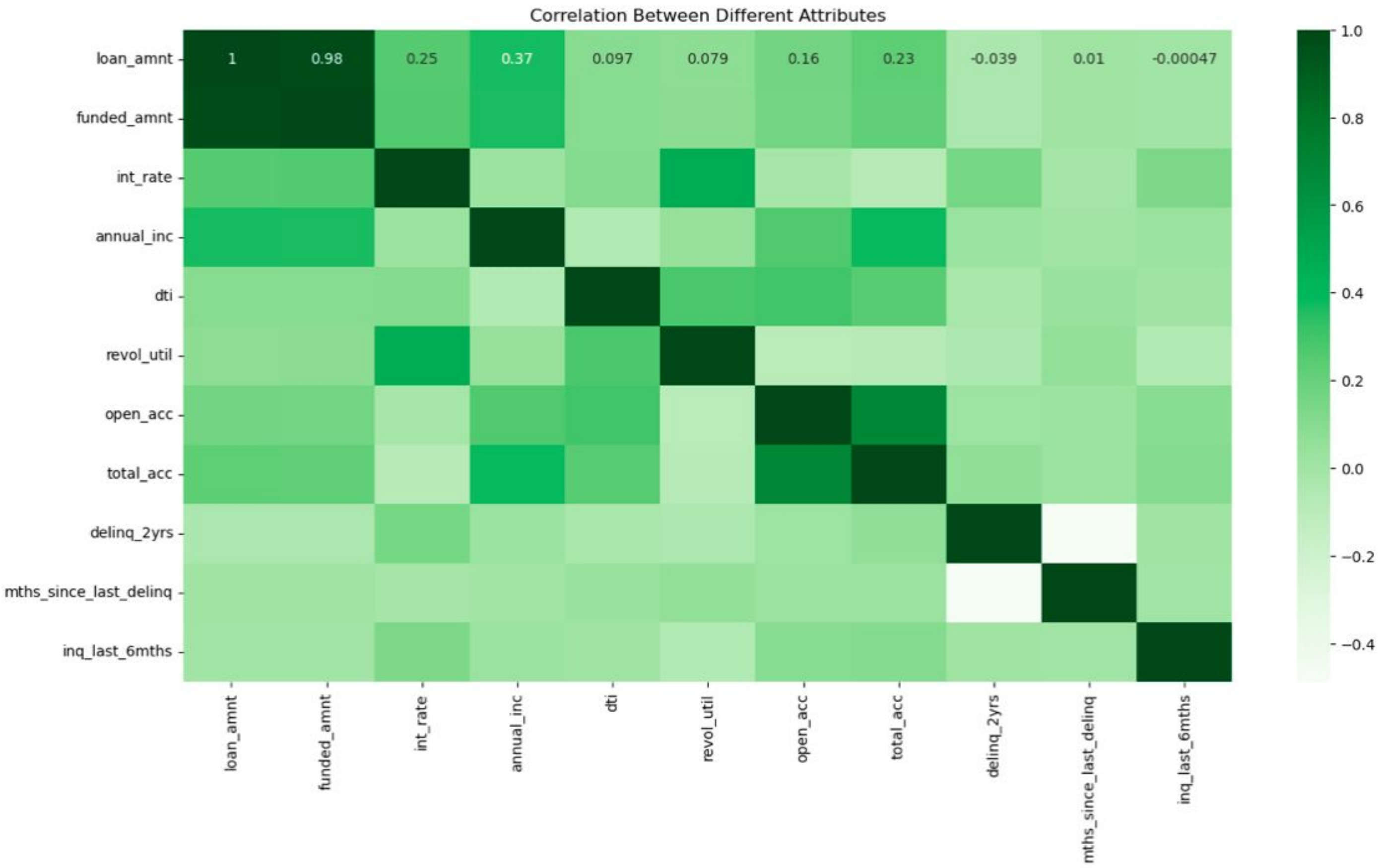


Multivariate Analysis

Correlation between numerical attributes

The moderate negative correlation (-0.48) between delinq_2yrs and mths_since_last_delinq suggests that borrowers who have gone longer without a delinquency are less likely to have had many delinquencies recently, which could indicate lower credit risk.

The moderate positive correlation (0.46) between int_rate and revol_util suggests that borrowers who use more of their available credit are typically charged higher interest rates. This relationship highlights the use of revol_util as a risk factor in Lending Club, where borrowers with higher revolving credit utilization are considered riskier and therefore face higher costs for borrowing.



Conclusion

For analysis of year 2007 to 2011 we found following insight

1. 85% loans are fully paid and 15% loans are got defaulted
2. Mid level earner have high chances of loan default, high and low income also experience default but less frequently.
3. Similarly high and low debt to income ratio have less frequency of default comparative middle DTI
4. For Revolving utilization default is max at 0 and then sharply less at next bucket and then gradually increase.
5. High interest rate customer may tends to default the loan.
6. Higher grade (A) may also experience default but less than other and this rate increase gradually according to grade.
7. Customer with purpose of small business may lean to default loan, followed by renewable energy and education.
8. Home ownership does not have so much impact on loan default.
9. Verified loan have max average rate of get default