International Institute of Information Technology Bangalore

upGrad

Lending Club Case Study

upGrad and IIITB Machine Learning & AI Program - November 2024

Submitted by-
Tejasva Dhyani
Yatin Gupta

Contents

- ➔ Problem Statement
- ➔ Data Understanding
- ➔ Data Cleaning and Manipulation
- ➔ Univariate Analysis
- ➔ Bivariate Analysis
- ➔ Multivariate Analysis
- ➔ Recommendations
- ➔ Conclusion

Problem Statement

➔ **Lending Club** is one of the largest online marketplace that facilitates personal loans, business loans, and financing of medical procedures to it’s borrowers.

➔ The company faces a challenge in identifying loan applications that are considered as **“Risky”**

➔ The company wants to minimise the major source of **Credit Loss**. For this, it wants to understand the **driving factors** that may result into a **“Charged-Off”** loan in order to minimise the Credit loss.

➔ Our **objective** is to analyse the company’s **existing data** of the approved loans and provide **5 driving factors** that would eventually help in the company’s cause.

Data Understanding

- Lending Club provides us with a **Loan Dataset** that contains the details of all the loan applications which were approved by the company from 2007 to 2011.
- The company also provides us with a **Data Dictionary** that contains the meaning of all the columns from the Loan dataset.

Data Cleaning and Manipulation

Following steps were performed in order to clean data:

- **Fix Columns**

  1. Identified columns/variables that have only 1 unique value. These variables cannot help in gaining any useful insight. Thus, removed these columns. This check was also executed after all the cleaning.

  2. Identified columns/variables that have more 50% null values and those values cannot be imputed. Removed those columns as those columns can impact the insight and give the wrong result.

  3. Identified columns that are neither categorical nor quantitative. Such as desc, url. These columns cannot help in analysis. So removed them.

  4. Identified columns that have unique values equal/near to total rows count such as id, member_id, title, emp_title. These columns cannot help in deriving any trend in data so removed them to avoid complexity in analysis and remain focused on important variables.

  5. Identified columns that can neither be driving variables nor target variables such as delinq_2yrs, inq_last_6mths, revol_bal and removed them to avoid complexity in analysis and remain focused on important variables.

  6. Identified columns that are giving redundant insight or not worth insight and removed them such as total_acc. We kept instead open_acc.

  7. Identified columns that are showing inconsistent trend such as revol_util. Removed them to avoid complexity in analysis and remain focused on important variables.

Data Cleaning and Manipulation Cont..

- **Fix Missing Values / Rows**

1. Checked columns/variables non-null values count. Analyzed row/column data of each column which shows null values and analyzed if those null values can be imputed. For all the cases found that those values cannot be imputed. Thus removed rows that have null columns.

2. Checked for rows that have all columns same i.e duplicate rows. Either those rows were not there or they were cleared in step 1.

3. There were no unnecessary header, footer, total/subtotal, column indicator rows. Empty rows were either not there or might got cleared in step 1.

4. Remove rows where loan_status is current. Because data where loan_status is current cannot help us in understanding variables that can drive default behaviour of customer.

5. Removed rows where home_ownership is NONE. This record cannot provide us any insight.

6. Not exactly during row analysis but during Bivariate analysis, we identified outliers in quantitative variables such as annual_inc, int_rate, funded_amnt, installment, open_acc and removed them to perform proper analysis.

Data Cleaning and Manipulation Cont..

- **Fix Data Type**

  1. Removing unit is generally recommended in data analysis because it provides ease of parsing, standardization, improve storage efficiency and simplify data visualization. Removed units such as percentage, months from the columns and convert them to pure integer or float type. ‘term’ had formatting issues as well. This type of fix was done for columns such as term, int_rate.

  2. Update emp_length so that it’s values lies between 0-10.

  3. Convert columns which contains month-year in string format to datetime format such as earliest_cr_line, issue_d.

- **Fix Invalid Values**

  1. earliest_cr_line contains value greater than current year which is not possible. Removed those values.

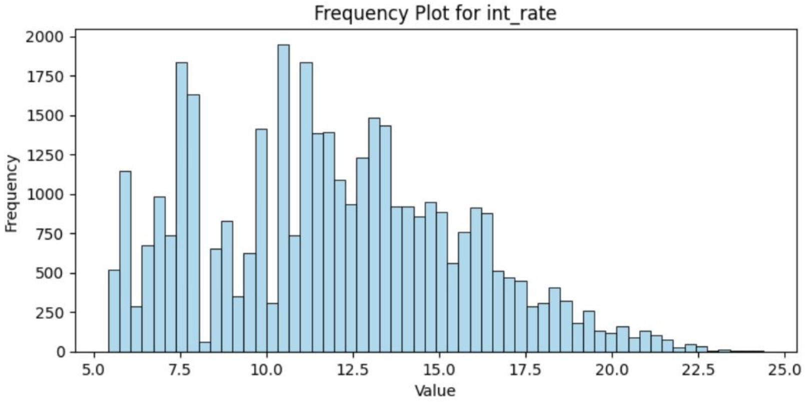
  2. Analyzed if columns follows expected format such as zip_code.

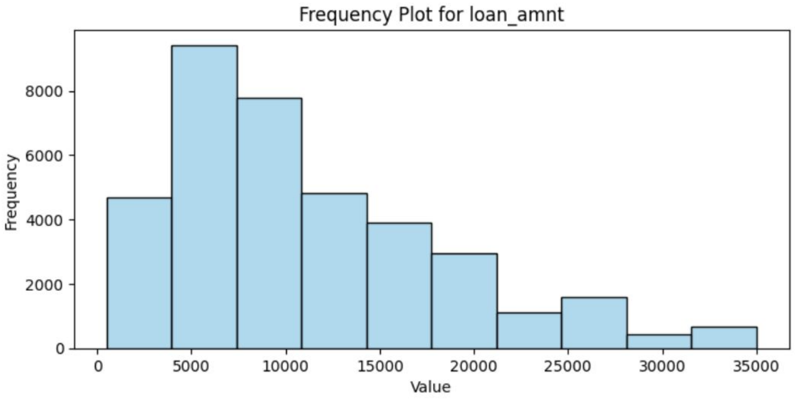
  3. Check if loan amount is smaller than funded amount or funded amount is smaller than funded invested amount.

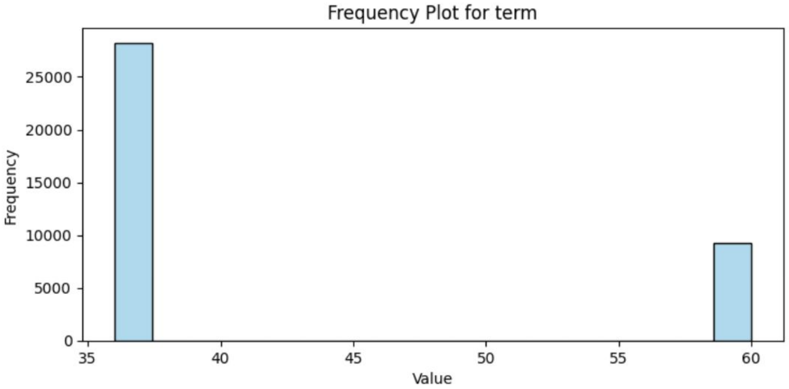
In the end, after all cleaning we remained with 21 columns and 37428 rows with 0 non-null columns.

Univariate Analysis

Frequency Plots





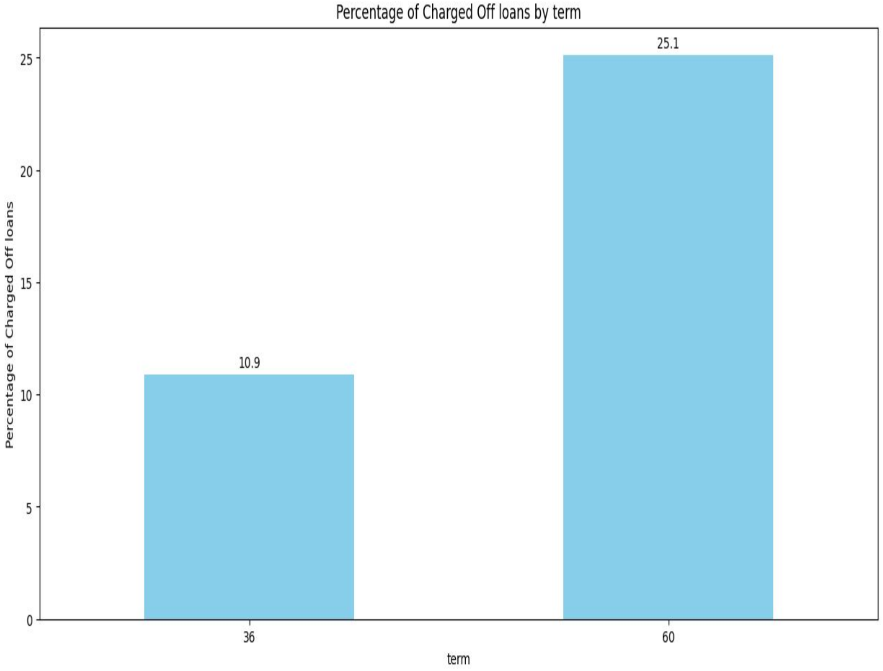


- ➔ Duration for all the loans is either 3 years or 5 years.
- ➔ Majority of loans are of the amount between $4000 and $15000 and at an interest rate between 9% and 15%

Bivariate Analysis

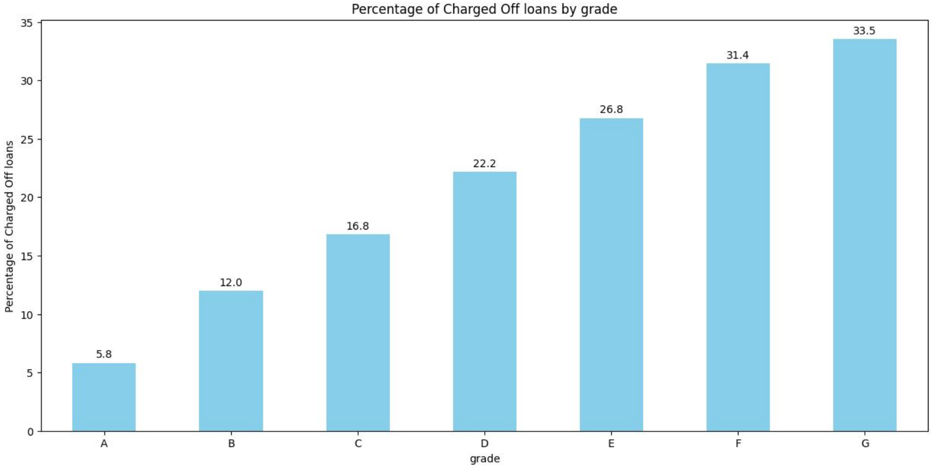
Percentage of Charged off loans segmented by Term

- ➔ Loans issued for longer duration i.e. 5 years are 2.5 times more likely get defaulted as compared to loans issued for 3 years.



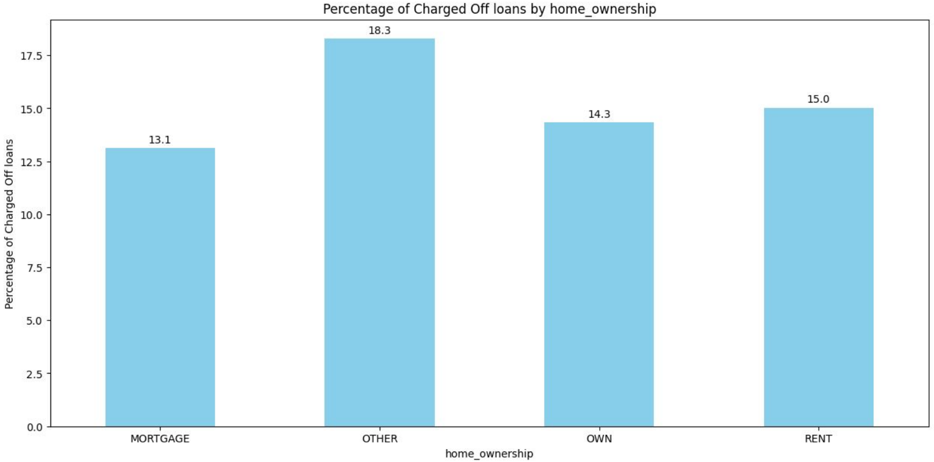
Percentage of Charged off loans segmented by Grade

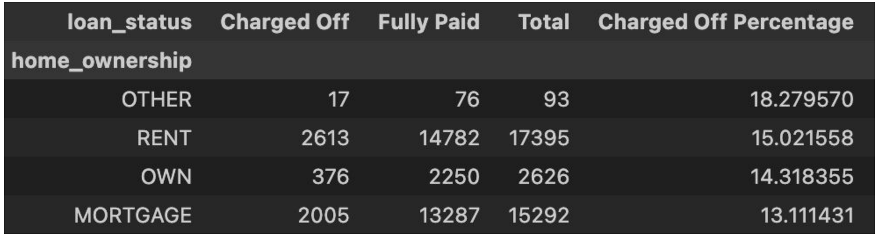
- ➔ Customers who are marked as “G” grade has highest percentage of defaulted applications.
- ➔ We can conclude that as grade is shifting from A to G, chances of default are also increasing



Percentage of Charged off loans segmented by Home ownership

- ➔ Although “Others” have the highest chance of getting charged off, we are ignoring this value since the number of applications are comparatively very low.
- ➔ Customers who live in a “Rented” home are more likely to get defaulted as compared to other values.

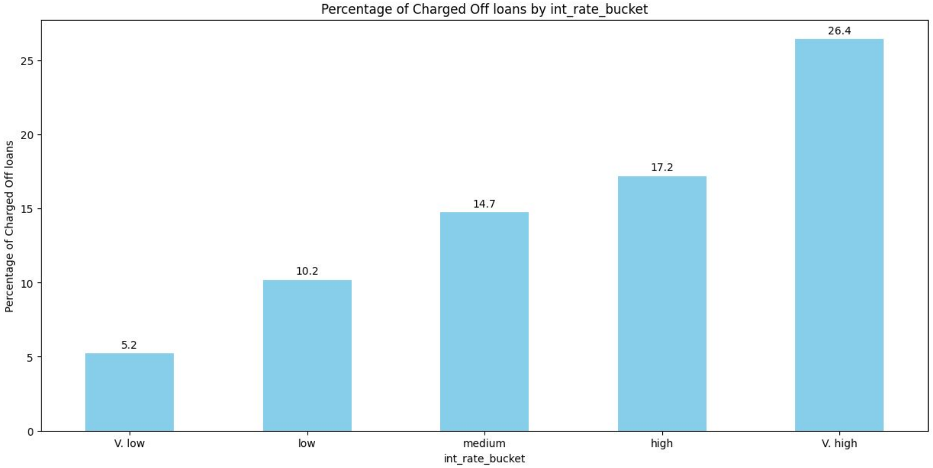




loan_status	Charged Off	Fully Paid	Total	Charged Off Percentage
home_ownership				
OTHER	17	76	93	18.279570
RENT	2613	14782	17395	15.021558
OWN	376	2250	2626	14.318355
MORTGAGE	2005	13287	15292	13.111431

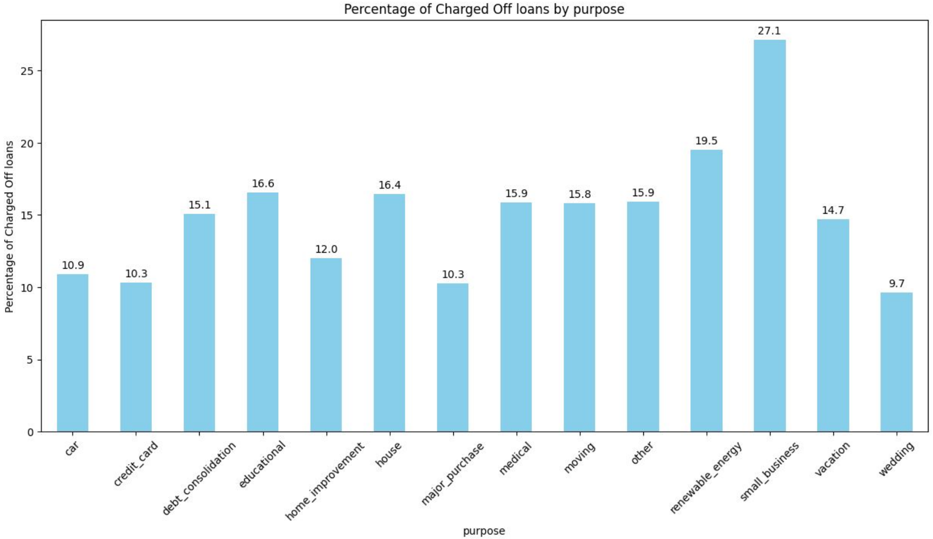
Percentage of Charged off loans segmented by Interest Rate

- ➔ We removed the outliers in the interest rate column and then created 5 equal quartile buckets on the interest rate values
- ➔ We can see as the interest rate increases, the chances of the loan application getting charged off also increases



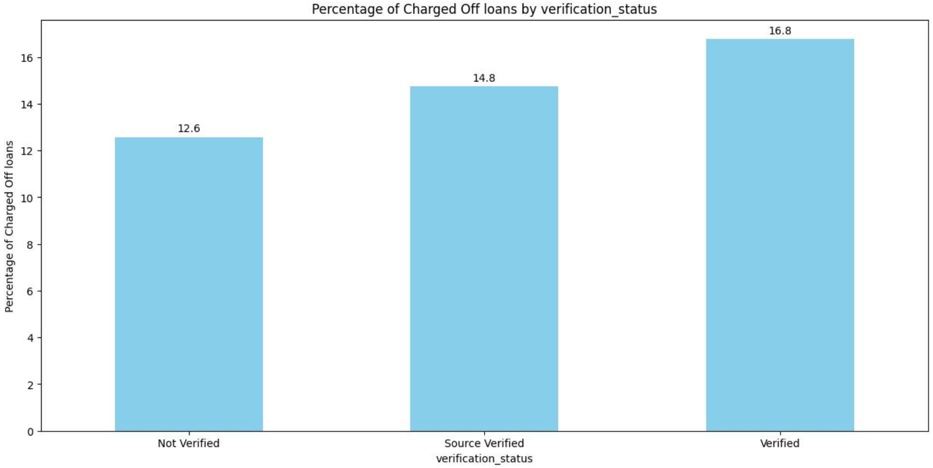
Percentage of Charged off loans segmented by Purpose

- ➔ Loans that are given for the purpose of small business are the riskiest with a chance of 27% to get defaulted
- ➔ Wedding loans or credit card loans are among the least risky loans.



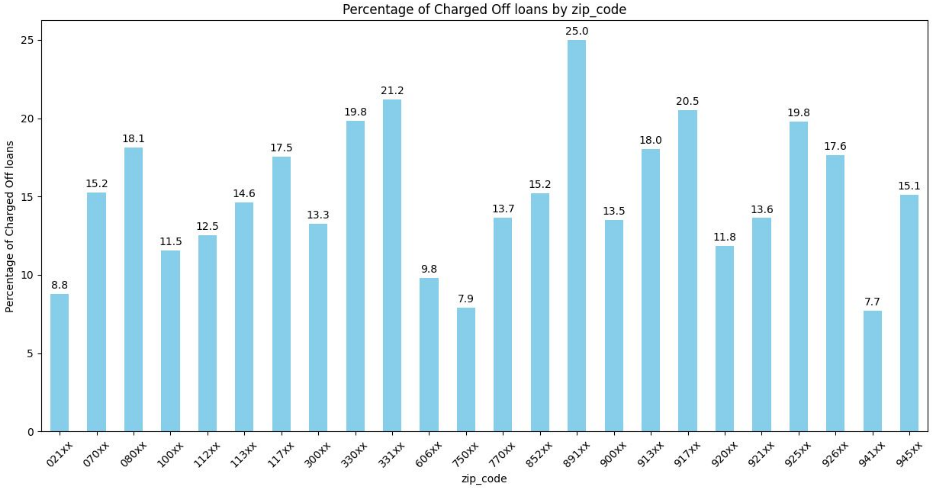
Percentage of Charged off loans segmented by Verification Status

- ➔ Verified loan applications are more riskier than the ones that are not verified.



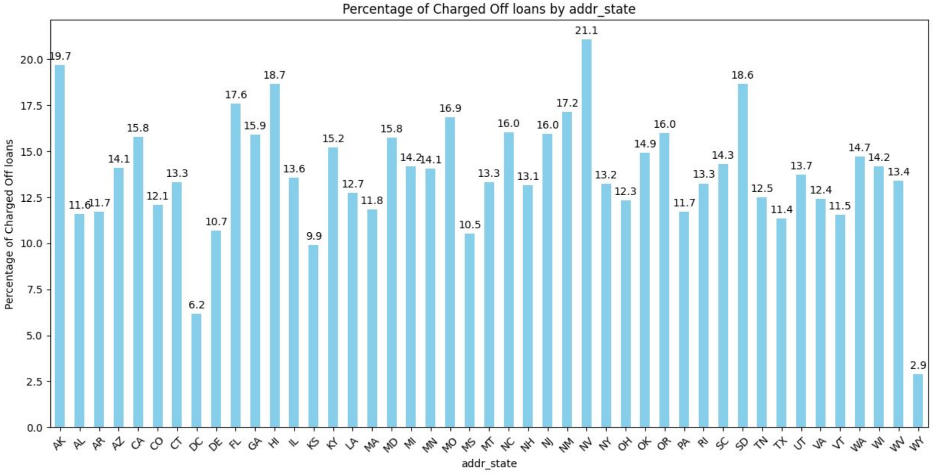
Percentage of Charged off loans segmented by Zip Code

- ➔ To analyse the zip code data, we considered only those zip codes that have more than 200 loans
- ➔ Loans issued in the zip code 891xx are considered as the riskiest
- ➔ Loans issued in zip codes 941xx and 750 can be categorized as least riskiest.


Percentage of Charged Off loans by zip_code

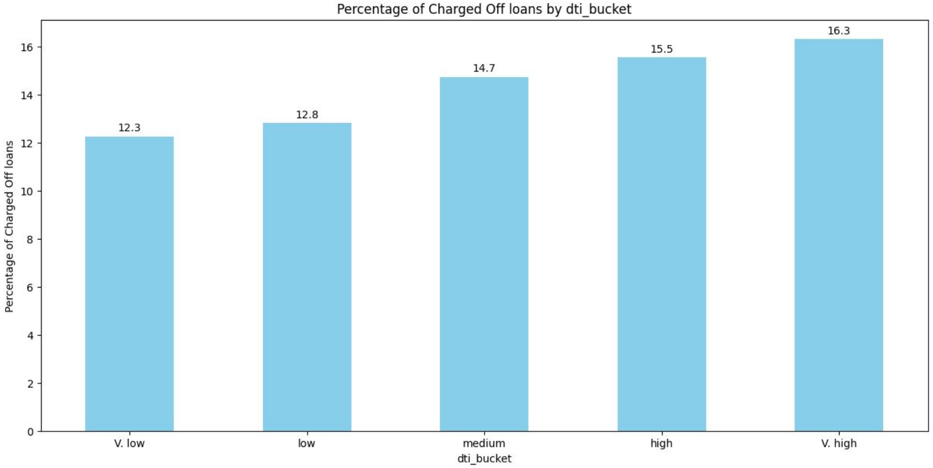
Percentage of Charged off loans segmented by State

- ➔ To analyse the State data, we considered only those states that have more than 10 loans
- ➔ Loans issued in the states of NV, AK and SD are considered as the riskiest
- ➔ Loans issued in states of WY and DC can be categorized as least riskiest.



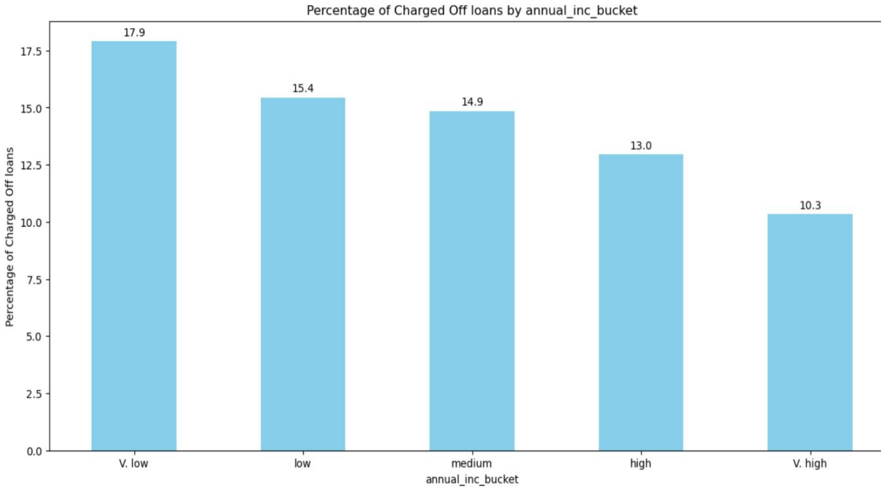
Percentage of Charged off loans segmented by Debt to Income ratio

- ➔ We removed the outliers in the dti column and then created 5 equal quartile buckets on the dti values
- ➔ We can see as the dti increases, the chances of the loan application getting charged off also increases



Percentage of Charged off loans segmented by Annual Income ratio

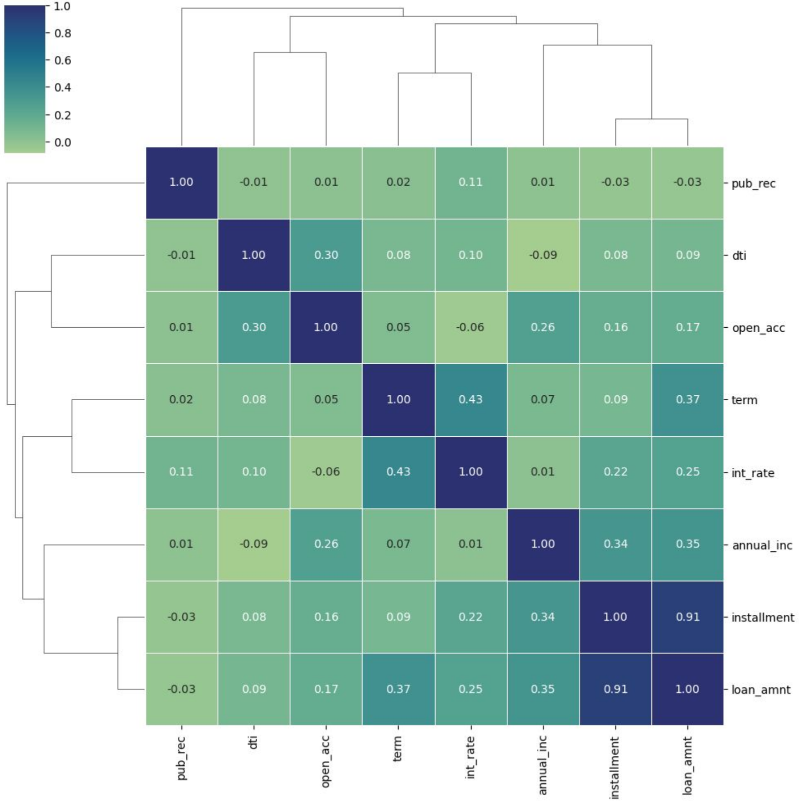
- ➔ We removed the outliers in the annual income column and then created 5 equal quartile buckets on the annual income values
- ➔ We can see as the annual income increase, charged-off percentage decreases.



Multivariate Analysis

Clustermap of quantitative variables

- ➔ There is a significant positive correlation between loan amount and installment.
- ➔ There is a comparable positive correlation between loan amount & annual income, term & interest rate, debt to income ratio & Number of open accounts
- ➔ Number of public records does not have a correlation between any of the parameters.



	pub_rec	dti	open_acc	term	int_rate	annual_inc	installment	loan_amnt
pub_rec	1.00	-0.01	0.01	0.02	0.11	0.01	-0.03	-0.03
dti	-0.01	1.00	0.30	0.08	0.10	-0.09	0.08	0.09
open_acc	0.01	0.30	1.00	0.05	-0.06	0.26	0.16	0.17
term	0.02	0.08	0.05	1.00	0.43	0.07	0.09	0.37
int_rate	0.11	0.10	-0.06	0.43	1.00	0.01	0.22	0.25
annual_inc	0.01	-0.09	0.26	0.07	0.01	1.00	0.34	0.35
installment	-0.03	0.08	0.16	0.09	0.22	0.34	1.00	0.91
loan_amnt	-0.03	0.09	0.17	0.37	0.25	0.35	0.91	1.00

Recommendations

Loan Acceptance

- ➔ Wedding loans or credit card loans
- ➔ Customers who live in a self owned or mortgaged house
- ➔ Loans issued at interest rates lower than 7.5%
- ➔ Customers who have low debt to income ratio
- ➔ A & B grade customers
- ➔ Customers living in states of NV, AK and SD

Loan Risk Factors

- ➔ Loans for the purpose of small businesses
- ➔ Customer having low income less than 56K.
- ➔ Customers who live in a rented house
- ➔ Loans issued at interest rates higher than 17%
- ➔ Customers who have high debt to income ratio
- ➔ F & G grade customers
- ➔ Customers living in states of WY and DC

Conclusion

As per our analysis, we found that the possibility of a loan getting default depends on the following driving factors

➔ Purpose of the loan
➔ Home ownership
➔ Interest rate
➔ Grade
➔ Debt to Income(dti) ratio
➔ Annual Income
➔ Term
➔ Address State

Other than above, the company must re-evaluate it’s verification process since we have seen that the verified applications are getting charged off more often than the ones that are not verified.