Lending Club Case Study

Submitted by:

Group Facilitator – Neha Khajuria

Group Member - Vrushali Ranjalkar

Table of Content

Problem Statement

- Lending Club, a Consumer Finance marketplace specializing in offering a variety of loans to urban customers, faces a critical challenge in managing its loan approval process. When evaluating loan applications, the company must make sound decisions to minimize financial losses, primarily stemming from loans extended to applicants who are considered "**Risky**".
- These financial losses, referred to as Credit Losses, occur when borrowers fail to repay their loans or default. In simpler terms, borrowers labeled as "Charged-Off" are the ones responsible for the most significant losses to the company.
- The primary objective of this exercise is to assist Lending Club in mitigating credit losses. This challenge arises from two potential scenarios.
  - Identifying applicants likely to repay their loans is crucial, as they can generate profits for the company through interest payments. Rejecting such applicants would result in a loss of potential business.
  - On the other hand, approving loans for applicants not likely to repay and at risk of default can lead to substantial financial losses for the company.
- The objective is to find the applicants at risk of defaulting on loans, enabling a reduction in credit losses. This case study aims to achieve this goal through Exploratory Data Analysis (EDA) using the provided dataset
- In essence, 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.

Data Understanding

Dataset Attributes

Primary Attribute:

Loan Status: The Principal Attribute of Interest (loan_status). This column consists of three distinct values:

- ✓ **Fully-Paid:** Signifies customers who have successfully repaid their loans.
- ✓ **Charged-Off:** Indicates customers who have been labeled as "Charged-Off" or have defaulted on their loans.
- ✓ **Current:** Represents customers whose loans are presently in progress and, thus, cannot provide conclusive evidence regarding future defaults.

Decision Matrix:

Loan Acceptance Outcome- There are three potential scenarios:

- ✓ **Fully Paid-** This category represents applicants who have successfully repaid both the principal and the interest rate of the loan.
- ✓ **Current-** Applicants in this group are actively in the process of making loan installments; hence, the loan tenure has not yet concluded. These individuals are not categorized as 'defaulted.'
- ✓ **Charged-off-** This classification pertains to applicants who have failed to make timely installments for an extended period, resulting in a 'default' on the loan.

Loan Rejection- In cases where the company has declined the loan application (usually due to the candidate not meeting their requirements), there is no transactional history available for these applicants. Consequently, this data is unavailable to the company and is not included in this dataset.

****Note: For the purposes of this case study, rows with a "**Current**" and "**Fully-Paid**" status will be excluded from the analysis as this will not help in driving any facts.*

Data Understanding Cont.

Key Columns of Significance:
The provided columns serve as pivotal attributes, often referred to as predictors. These attributes, available during the loan application process, significantly contribute to predicting whether a loan will be approved or rejected. It's important to note that some of these columns may be excluded due to missing data in the dataset

- **Customer Demographics:**
  - ✓ **Annual Income** (annual_inc): Reflects the customer's annual income. Typically, a higher income enhances the likelihood of loan approval.
  - ✓ **Home Ownership** (home_ownership): Indicates whether the customer owns a home or rents. Home ownership provides collateral, thereby increasing the probability of loan approval.
  - ✓ **Employment Length** (emp_length): Represents the customer's overall employment tenure. Longer tenures signify greater financial stability, leading to higher chances of loan approval.
  - ✓ **Debt to Income** (dti): Measures how much of a person's monthly income is already being used to pay off their debts. A lower DTI translates to a higher chance of loan approval.
  - ✓ **State** (addr_state): Denotes the customer's location and can be utilized for creating a generalized demographic analysis. It may reveal demographic trends related to delinquency or default rates.

Data Understanding Cont.

Key Columns of Significance:

- **Loan Characteristics:**
  - ✓ **Loan Amount** (loan_amt): Represents the amount of money requested by the borrower as a loan.
  - ✓ **Grade** (grade): Represents a rating assigned to the borrower based on their creditworthiness, indicating the level of risk associated with the loan.
  - ✓ **Term** (term): Duration of the loan, typically expressed in months.
  - ✓ **Loan Date** (issue_d): Date when the loan was issued or approved by the lender.
  - ✓ **Purpose of Loan** (purpose): Indicates the reason for which the borrower is seeking the loan, such as debt consolidation, home improvement, or other purposes.
  - ✓ **Verification Status** (verification_status): Represents whether the borrower's income and other information have been verified by the lender.
  - ✓ **Interest Rate** (int_rate): Represents the annual rate at which the borrower will be charged interest on the loan amount.
  - ✓ **Installment** (installment): Represents the regular monthly payment the borrower needs to make to repay the loan, including both principal and interest.
  - ✓ **Public Records** (public_rec): Refers to derogatory public records, which contribute to loan risk. A higher value in this column reduces the likelihood of loan approval.
  - ✓ **Public Records Bankruptcy** (public_rec_bankruptcy): Indicates the number of locally available bankruptcy records for the customer. A higher value in this column is associated with a lower success rate for loan approval.

Data Understanding Cont.

Excluded Columns:

In our analysis, we will not consider certain types of columns. It's important to note that this is a general categorization of the columns we will exclude from our approach, and it does not represent an exhaustive list.

- **Customer Behaviour Columns**- Columns that describe customer behaviour will not be factored into our analysis. The current analysis focuses on the loan application stage, while customer behaviour variables pertain to post-approval actions. Consequently, these attributes will not influence the loan approval/rejection process.
- **Granular Data** - Columns providing highly detailed information that may not be necessary for our analysis will be omitted. For example, while the "grade" column may have relevance in creating business outcomes and visualizations, the "sub grade" column is excessively granular and will not be utilized in our analysis.
- **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
- Certain columns contain only 0 values, and these columns will also be dropped.
- **9** Columns with single value that do not contribute to the analysis will be removed.

Data Understanding Cont.

Excluded Columns:

- Columns with values that are single value but have other values as NA will be treated as constant and dropped.
- **9** Columns with **single value** that do not contribute to the analysis will be removed
- Columns with more than 65% of data being empty (**mths_since_last_delinq, mths_since_last_record**) will be dropped.

Columns (**id, member_id**) will be dropped as they are index variables with unique values and do not contribute to the analysis.

- Columns (**emp_title, desc, title**) will be dropped as they contain descriptive text (nouns) and do not contribute to the analysis.
- The redundant column (**url**) will be dropped. Further analysis reveals that the URL is a static path with the loan ID appended as a query, making it redundant compared to the (id) column.
- **660** records for **pub_rec_bankruptcies** are dropped due to missing values
- These columns capture customer behavior recorded after loan approval and are not available at the time of loan approval. Thus, these variables will not be included in the analysis.
- Columns to be dropped:

  (delinq_2yrs, earliest_cr_line, inq_last_6mths, open_acc, pub_rec, revol_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)

Data Cleaning & Pre-processing

For data cleaning and pre-processing, following steps have been followed:

1. Loading data from loan CSV
2. Checking for null values in the dataset 3.
3. Checking for unique values 4
4. Checking for duplicated rows in data
5. Dropping Records & Columns
6. Common Functions
7. Data Conversion
8. Outlier Treatment
9. Imputing values in Columns

Data Cleaning & Pre-processing

For data cleaning and pre-processing, following steps have been followed:

1. **Loading data from loan CSV** – Data have been loaded using the python pandas and as a first step we did the analysed the headers, footers in csv. Along with this, we check the shape of the rows and columns we have initially in loan_csv. [39717 * 111]
2. **Checking for null values in the dataset -** Overall there were 48.65% columns were having the NA Values, so we dropped all these columns which were 54 in number. After this drop we left with 39717 rows and 54 columns.
3. **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.
4. **Remove columns with high percentage of missing values** - Following columns having the null percentage more than 60%, and if these columns imputed, this will skew the analysis.
    1. next_pymnt_d **0.97**
    2. mths_since_last_record **0.93**
    3. mths_since_last_delinq **0.65**
5. **Checking for duplicated rows in data** - No duplicate rows were found.

Data Cleaning & Pre-processing

For data cleaning and pre-processing, following steps have been followed:

6. **Data Conversion/Dervived Columns**: Converted columns like debt to income (dti), funded amount (funded_amnt), funded amount investor (funded_amnt_inv) and loan amount (loan_amnt) to float to match the data. Also converted loan date (issue_d) to DateTime (format: yyyy-mm-dd).
7. **Outlier Treatment**: Calculated the Inter-Quartile Range (IQR) and filtering out the outliers outside of lower and upper bound. During Outlier analysis the following observations were made :
   - ✓ The annual income of most of the loan applicants is between 40K - 75K USD
   - ✓ The loan amount of most of the loan applicants is between 5K - 15K
   - ✓ The funded amount of most of the loan applicants is between 5K - 14K USD
   - ✓ The funded amount by investor for most of the loan applicants is between 5K - 14K USD
   - ✓ The interest rate on the loan is between 9% - 14%
   - ✓ The monthly installment amount on the loan is between 160 - 440
   - ✓ The debt to income ration is between 8 - 18
8. **Imputing values in Columns:**
   - ✓ Replaced the 'Source Verified' values as 'Verified' since both values mean the same thing i.e. the loan applicant has some source of income which is verified
   - ✓ There are 660 null values for pub_rec_bankruptcies. Dropped those rows as they cannot be imputed
   - ✓ The Employment length has 1015 missing values, which means either they are not employed or self employed (business owners). Considering they have a decent average annual income, we have assumed that these are business owners and we have added their employment duration with the mode value of emp_length which is 10+ years

Data Analysis

- As a part of data analysis, following analysis to be involved:

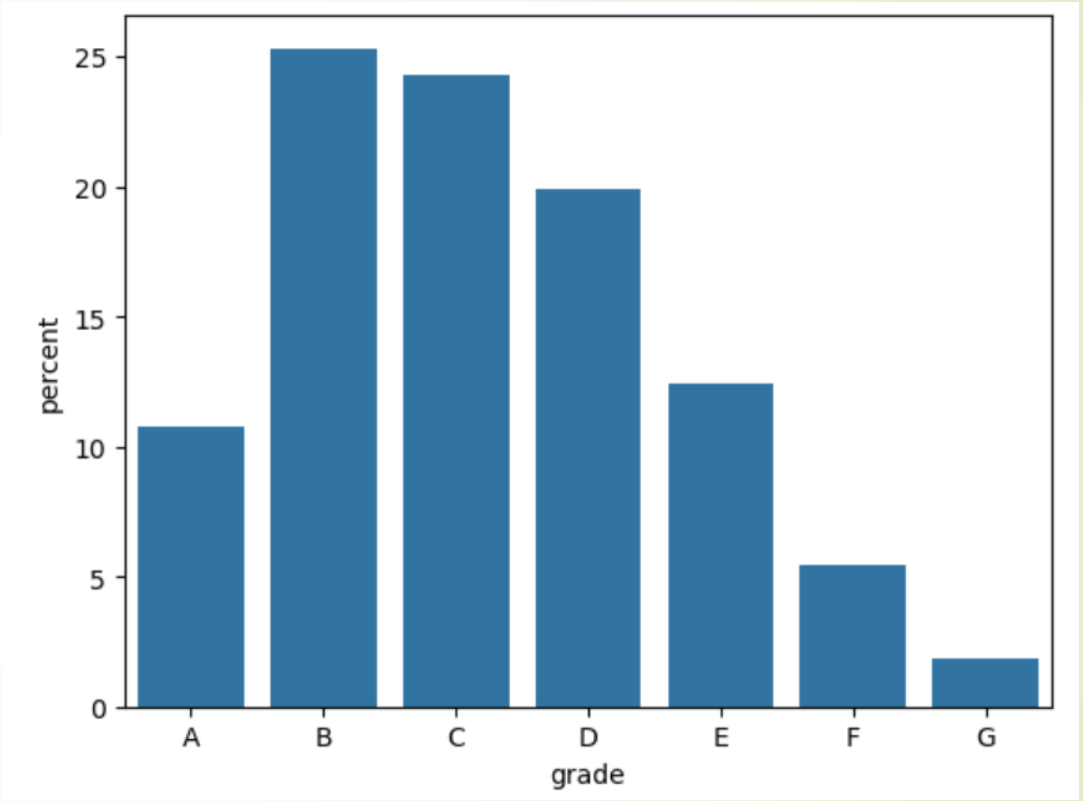
1. Univariant Analysis
2. Bivariant Analysis
3. Correlation Analysis

Univariant Analysis for Grades

Insight:

Most of the charged off loans are from category –

- B with 25.31% defaults
- C with 24.30% defaults-
- D with 19.94% default



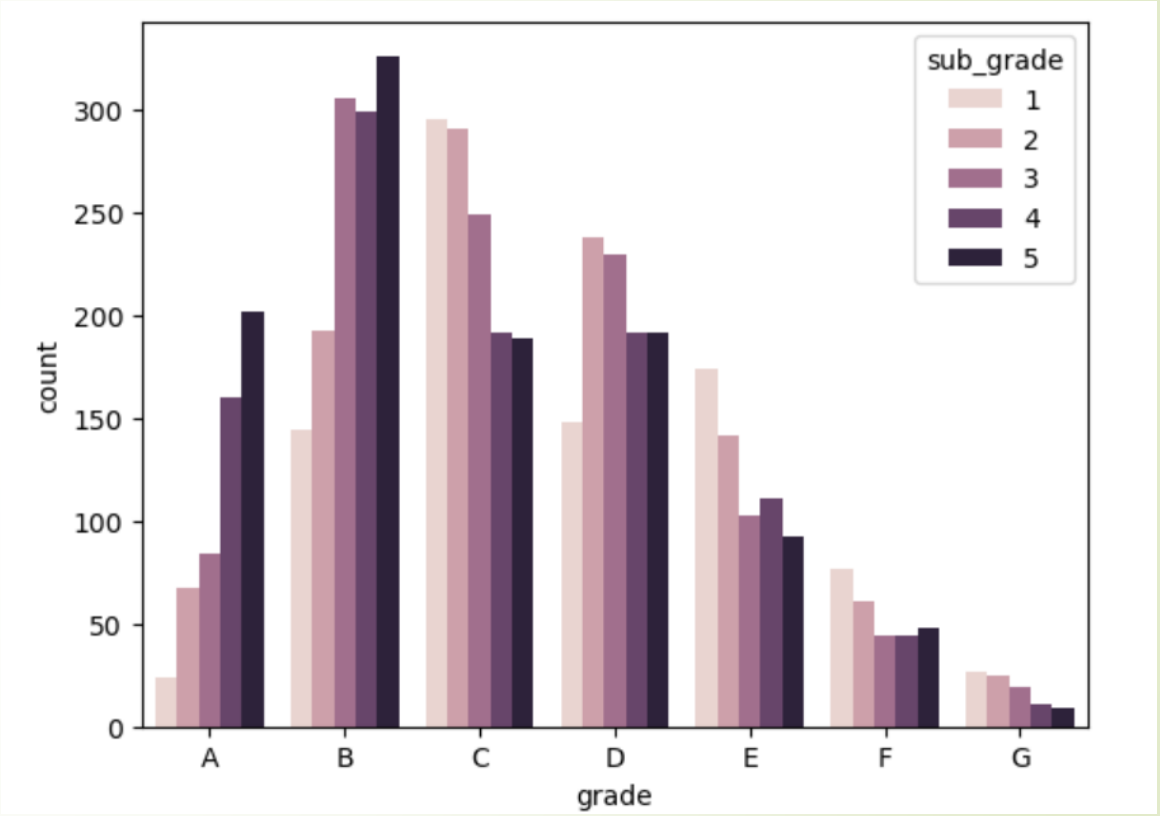
****** Category B and C together makes up for 49.61% of charged off loans***

Univariant Analysis for Sub-Grades

Insight:

There isn't a clear pattern in charged off loans for sub-categories.-

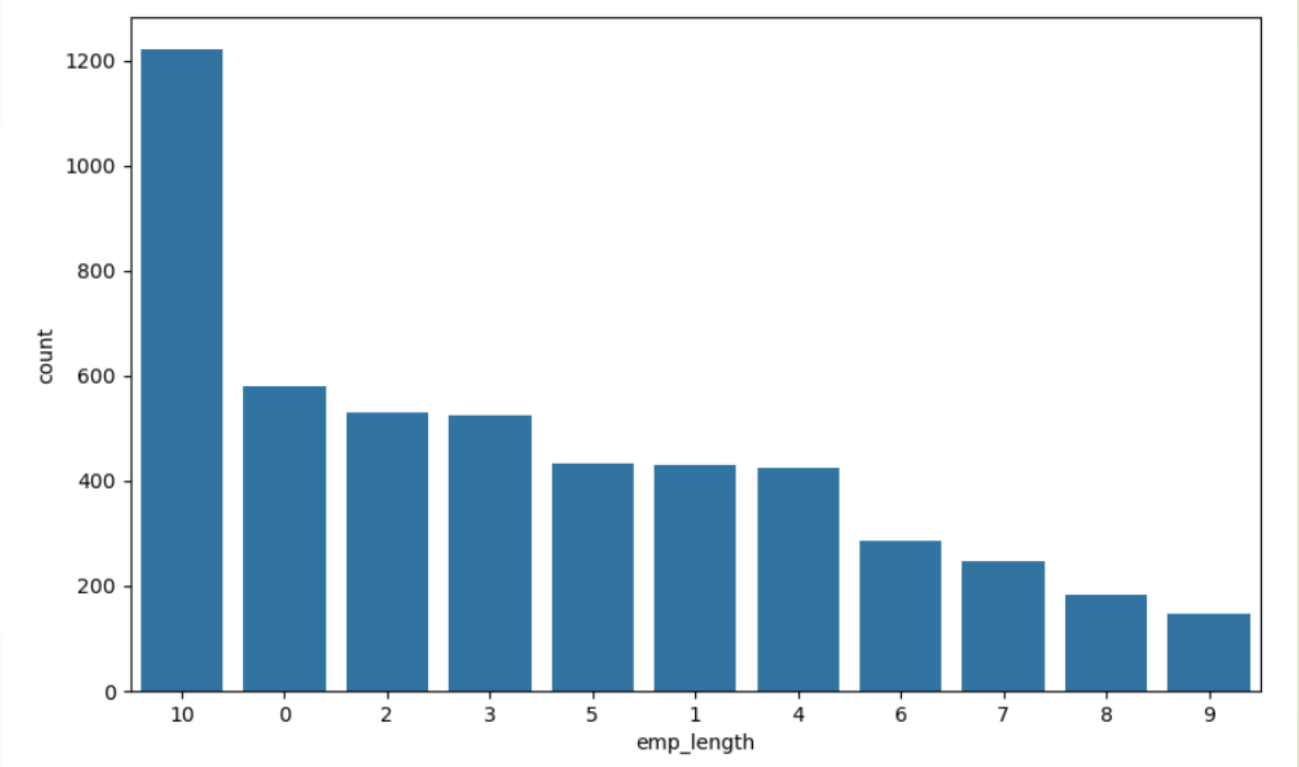
- Sub-Grade 5 is dominating in category B followed by 3 and 4
- Sub-Grade 1 has the most frequency in cat C followed by 2 and 3
- Sub-Grade 2 is winning in cat D followed by 3 and 4.



Univariant Analysis for Employee Length

Insight:

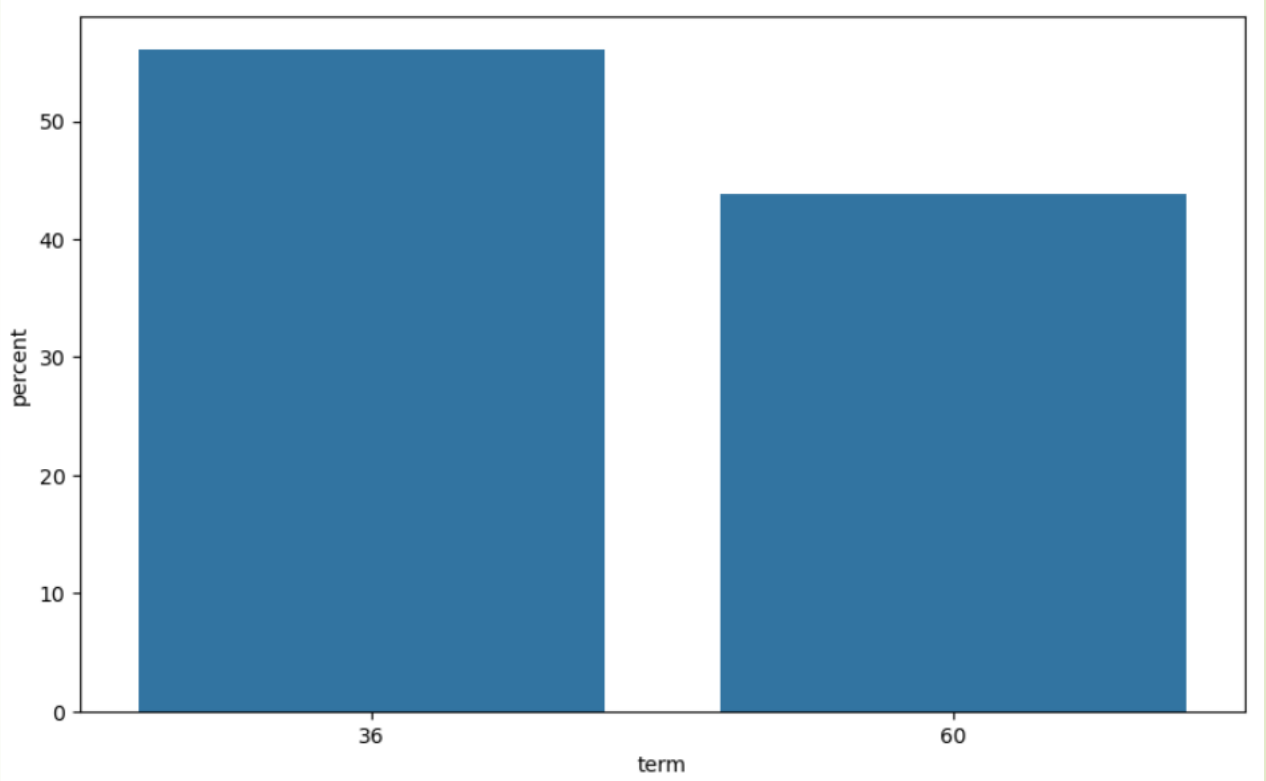
- Loan applications for applicants employed for more than 10years tends to be the most defaulted loans followed by < 1 and 2 Year



Univariant Analysis for Loan Term

Insight:

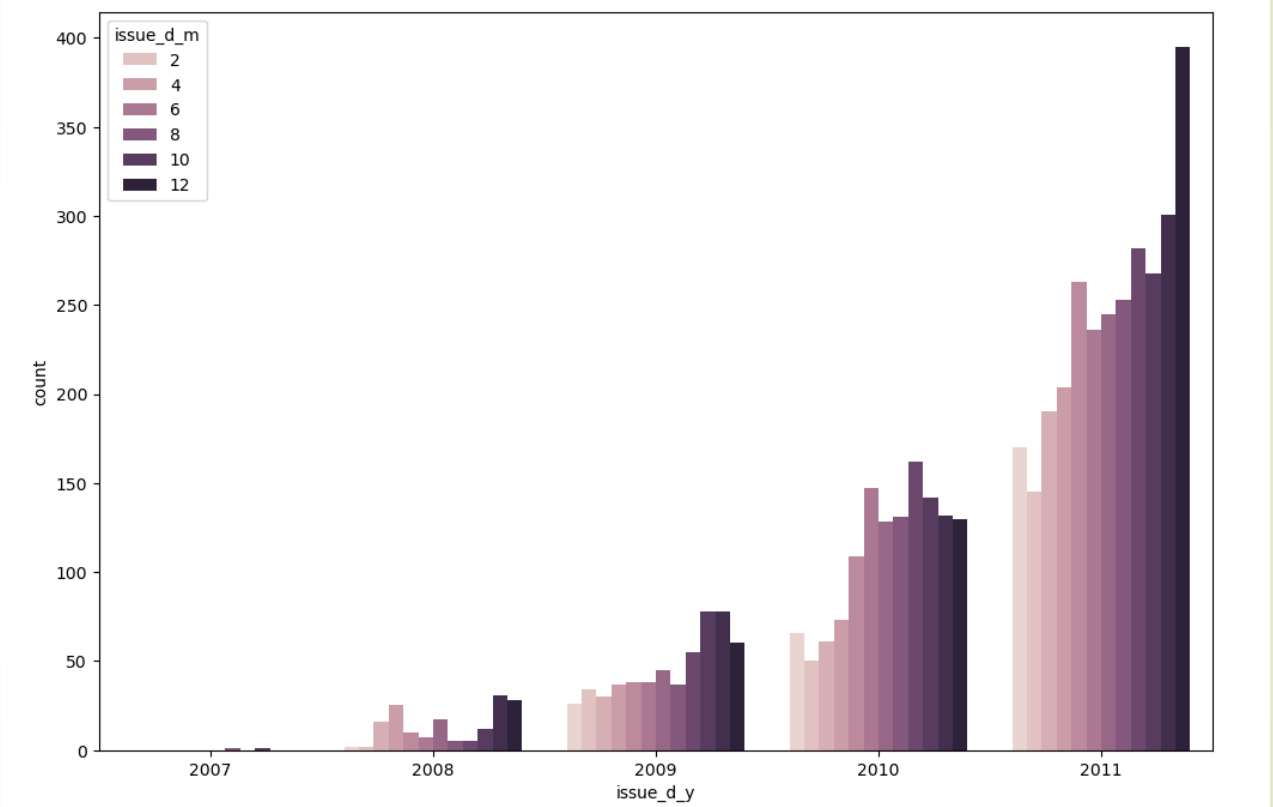
- More than 50% of defaulted loans are taken for lower term.



Univariant Analysis for Loan Year

Insight:

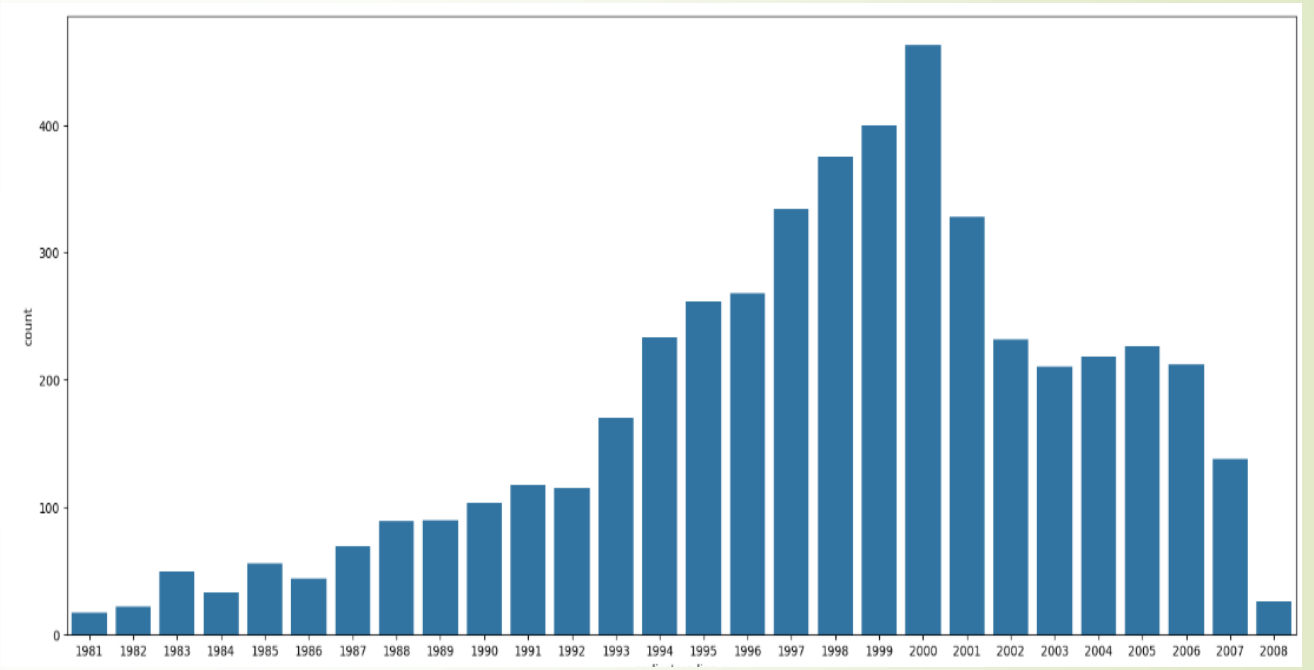
- We could see an upward trend in number of defaulted loans over the years.
- Most of the defaulted loans tends to be approved around end of year, this coincides with the holiday seasons.



Univariant Analysis for Earliest Credit Line

Insight:

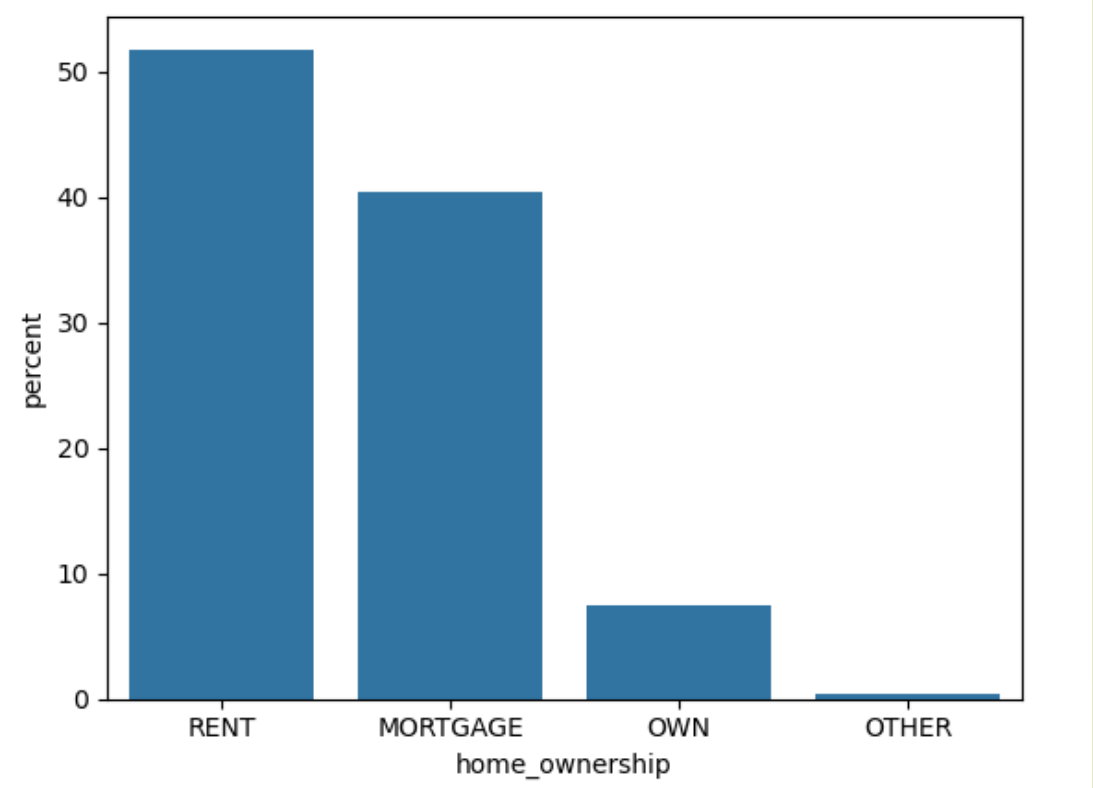
- We could infer that a long credit history doesn't necessarily means the ability for repayment.
- Also charged off loans peaked for customers who started their credit in 2000 and is on a downward trend ever since.



Univariant Analysis for Home Ownership

Insight:

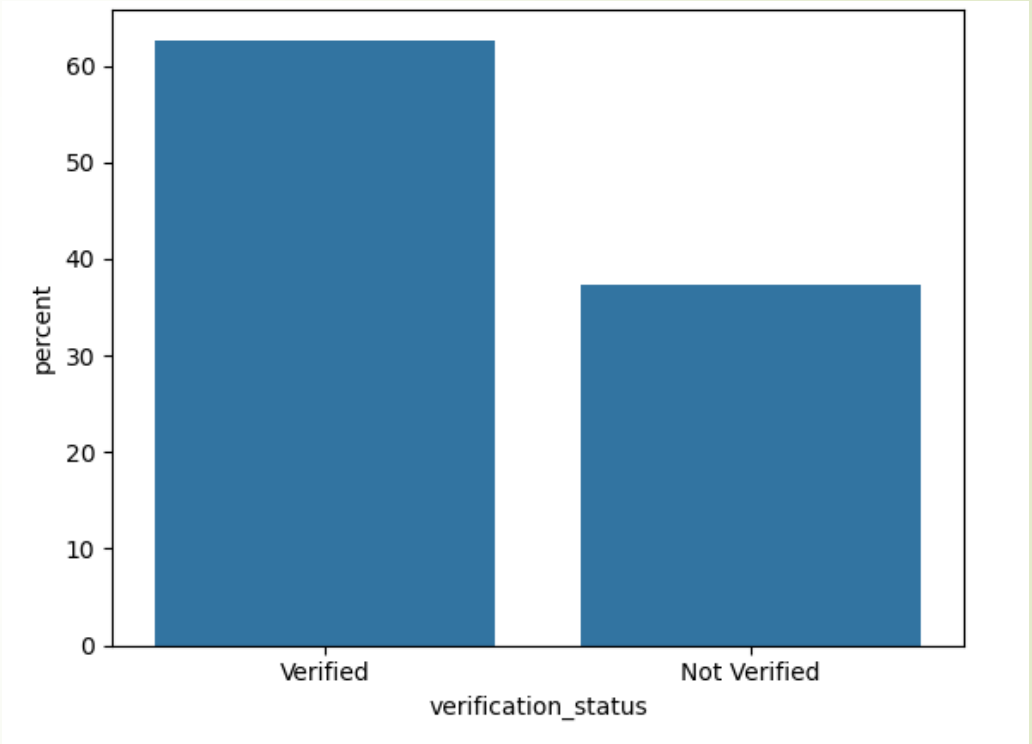
- Customers with Rented and Mortgaged homes make up for the majority of charged off loans.
- This might be due to the additional financial commitments for rent and mortgage payments.
- The LC should take extra precautions when considering borrowers ability to pay the installments if they have other fixed financial commitments.



Univariant Analysis for Verification Status

Insight:

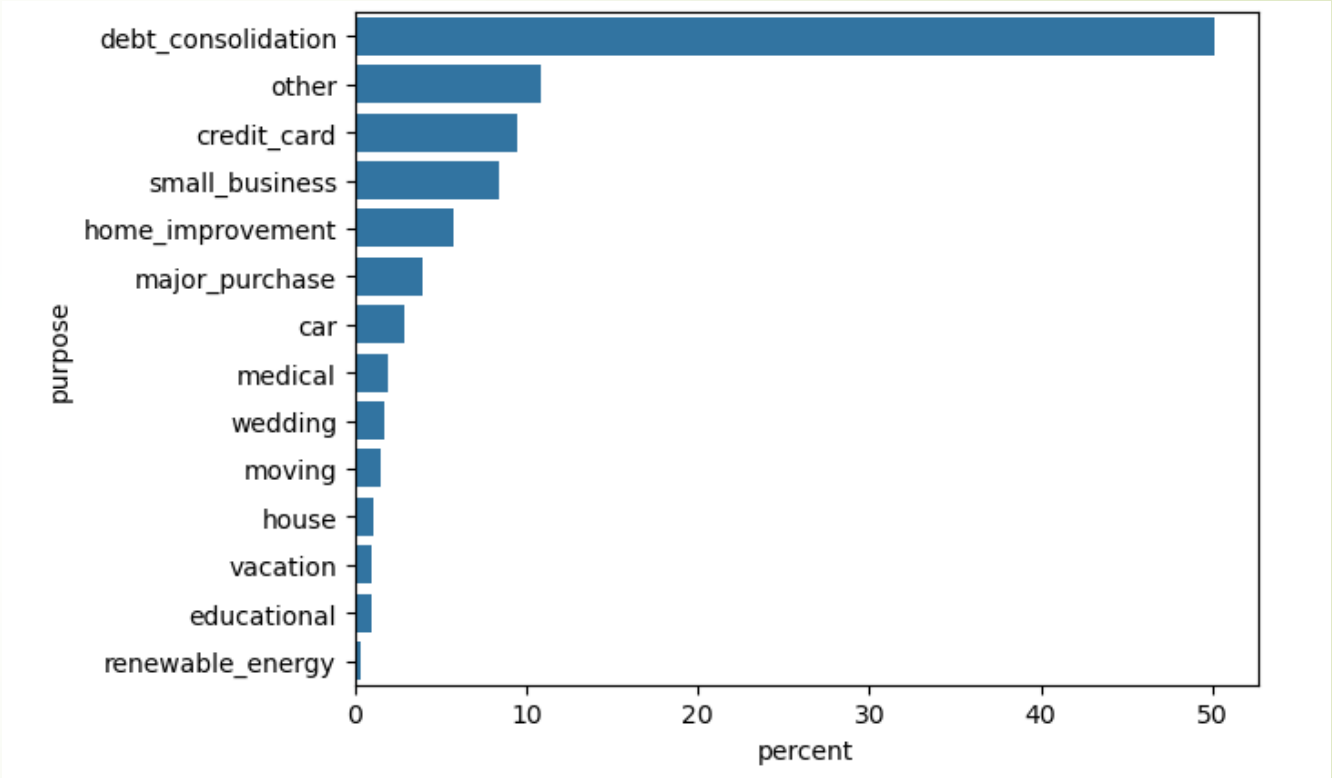
- Verified income is not a strong indicator of loan repayment capacity as most of the charged off loans are from verified category.
- LC needs to apply stricter methods for loan repayment capacity decisions.



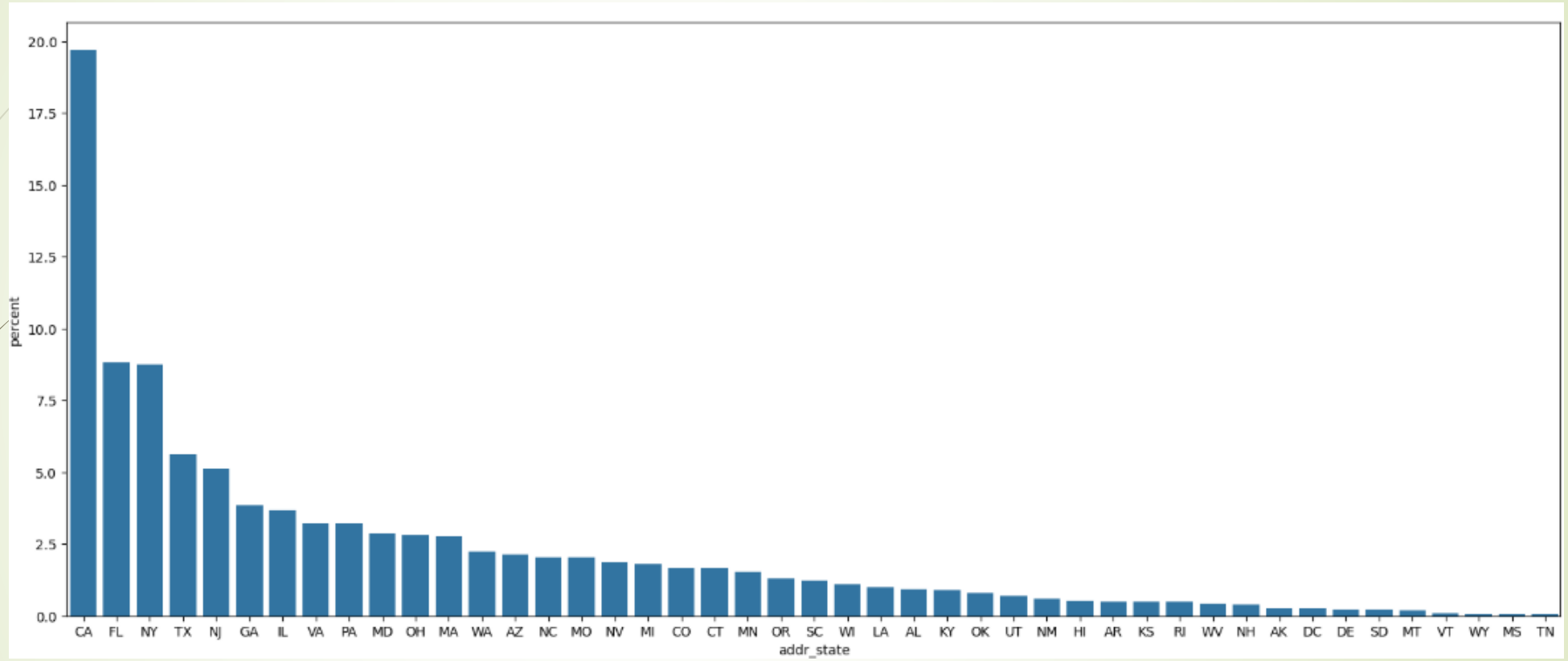
Univariant Analysis for Purpose

Insight:

- Customer who take loans for Debt Consolidation are at a higher risk of defaulting as they are already under financial pressure and might not meet the commitment



Univariant Analysis for Address State



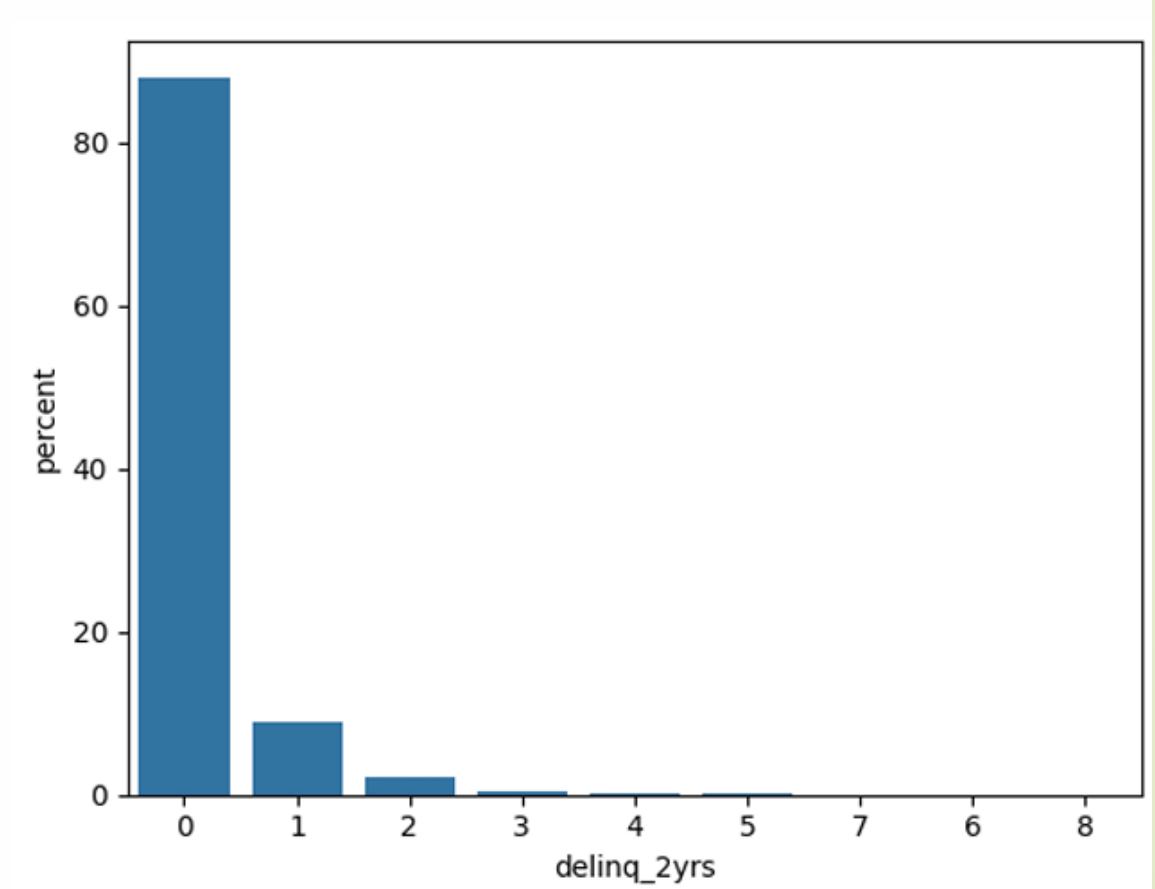
Insight:

- Most of the defaulted borrowers come from high GDP states such as California, Florida, New York, Texas and New Jersey.

Univariant Analysis for Delinquency 2yrs

Insight:

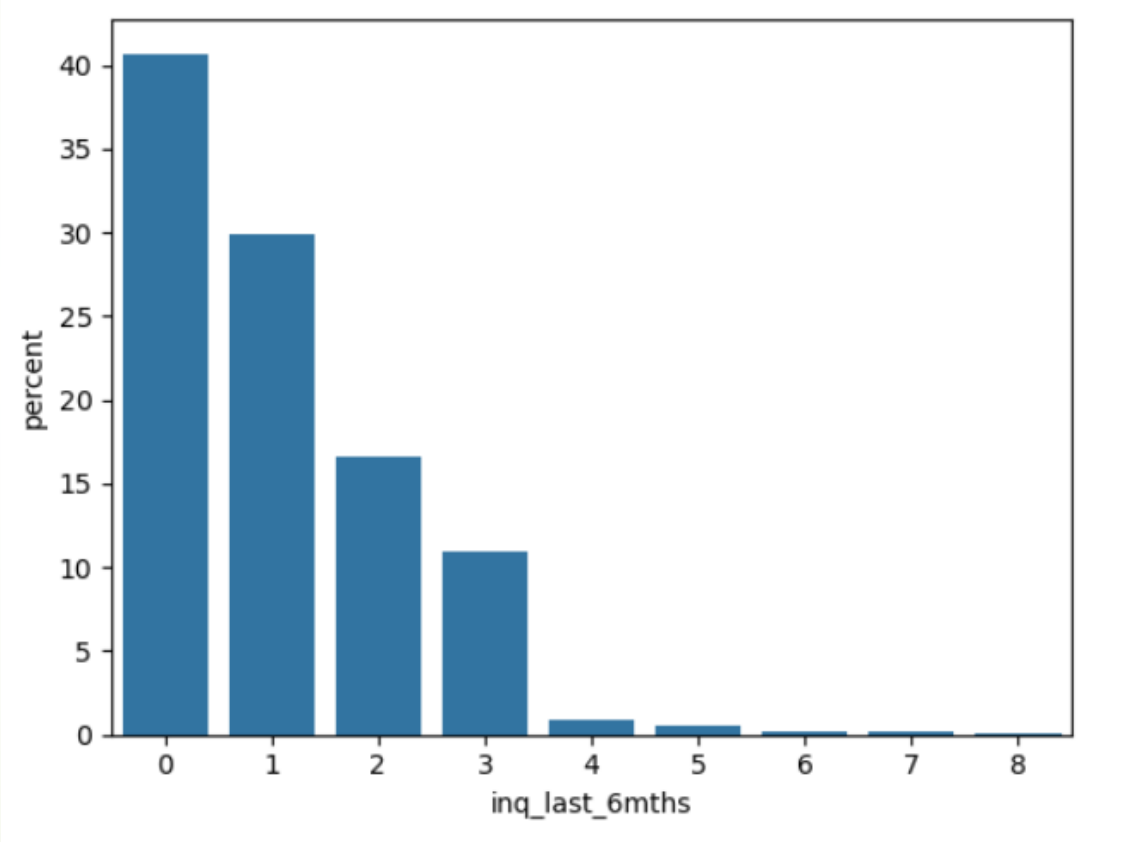
- Customers with no past delinquency in 2 years have higher risk of defaulting



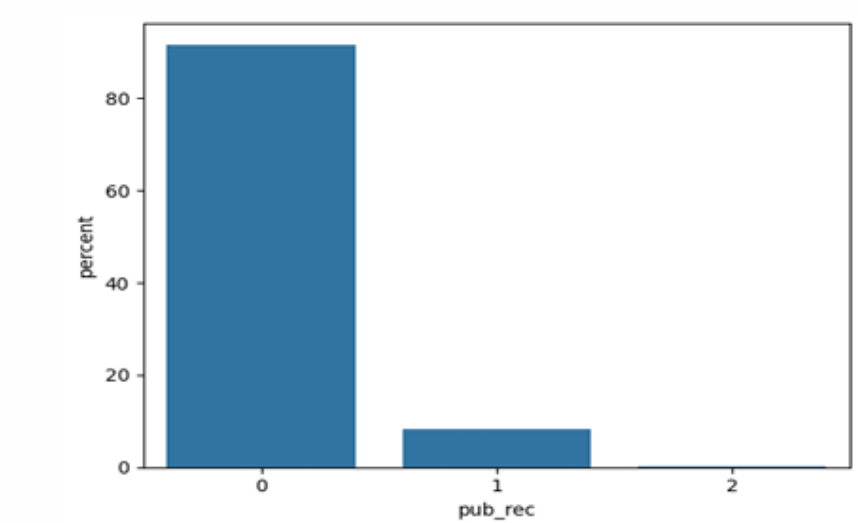
Univariant Analysis for Credit Inquiry in Last 6mths

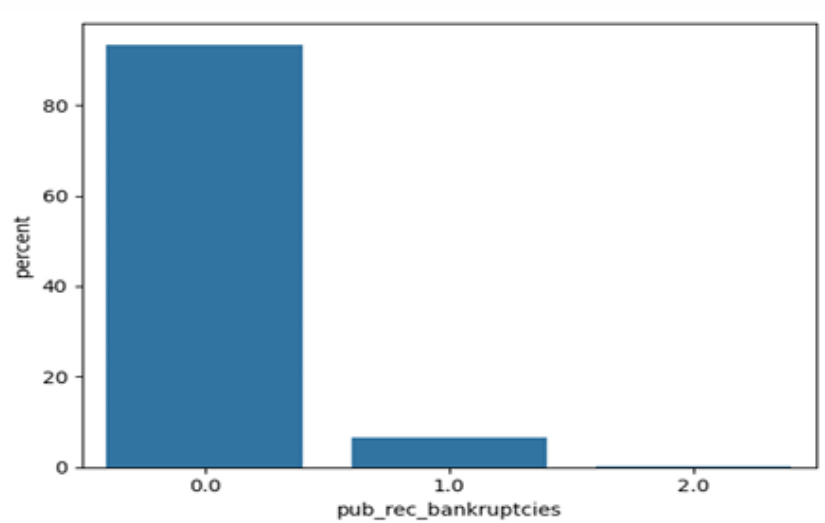
Insight:

- Customers with high number of credit inquiries are less likely to default.



Univariant Analysis for public delinquency and bankruptcy records





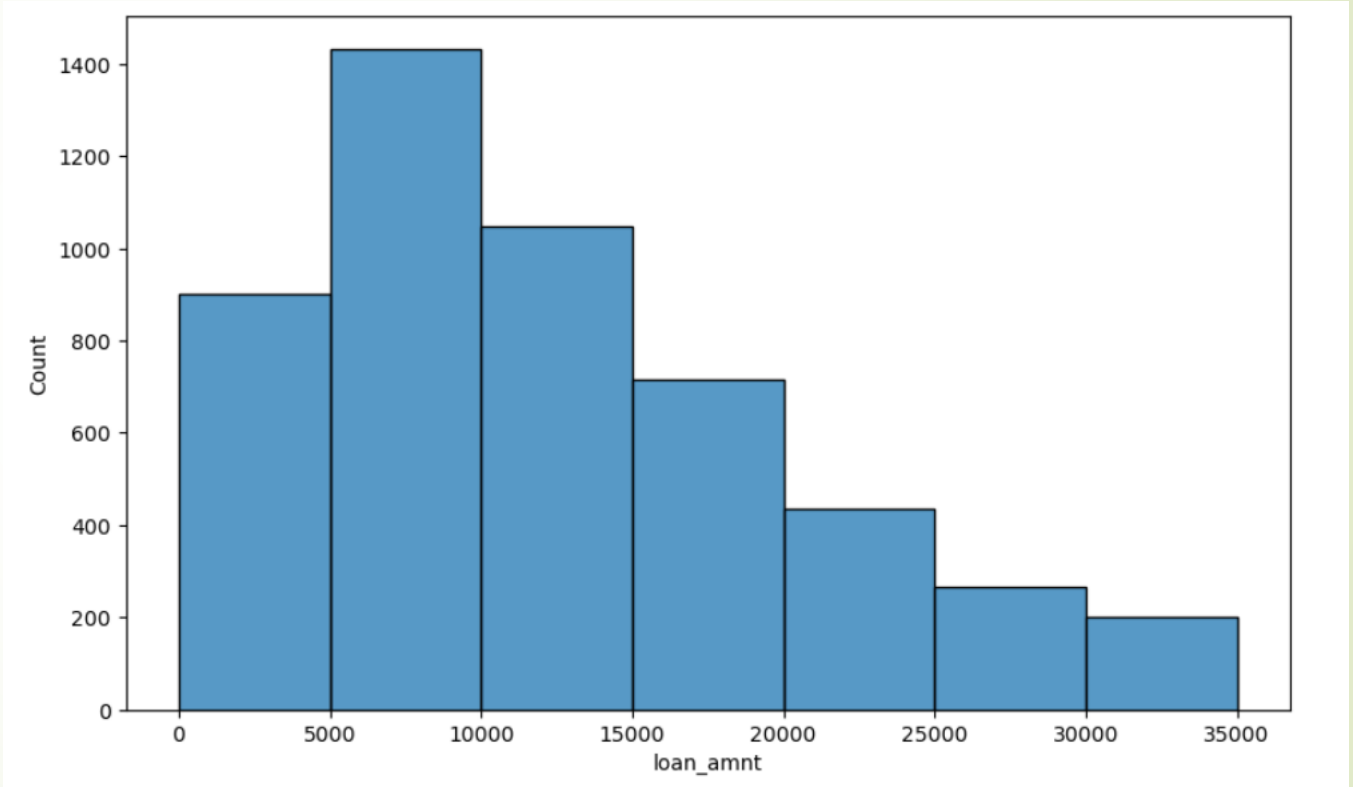
Insight:

- Customer with past public deliquency record or bankruptcy records are less likely to default on loans than the customer with no past public record.
- This might be due to the fact the customers who have public record tends to be more diligent in paying back what they owe as they are aware of the negative impact of such loans.

Univariant Analysis for Quantative Variable – Loan Amount

Insight:

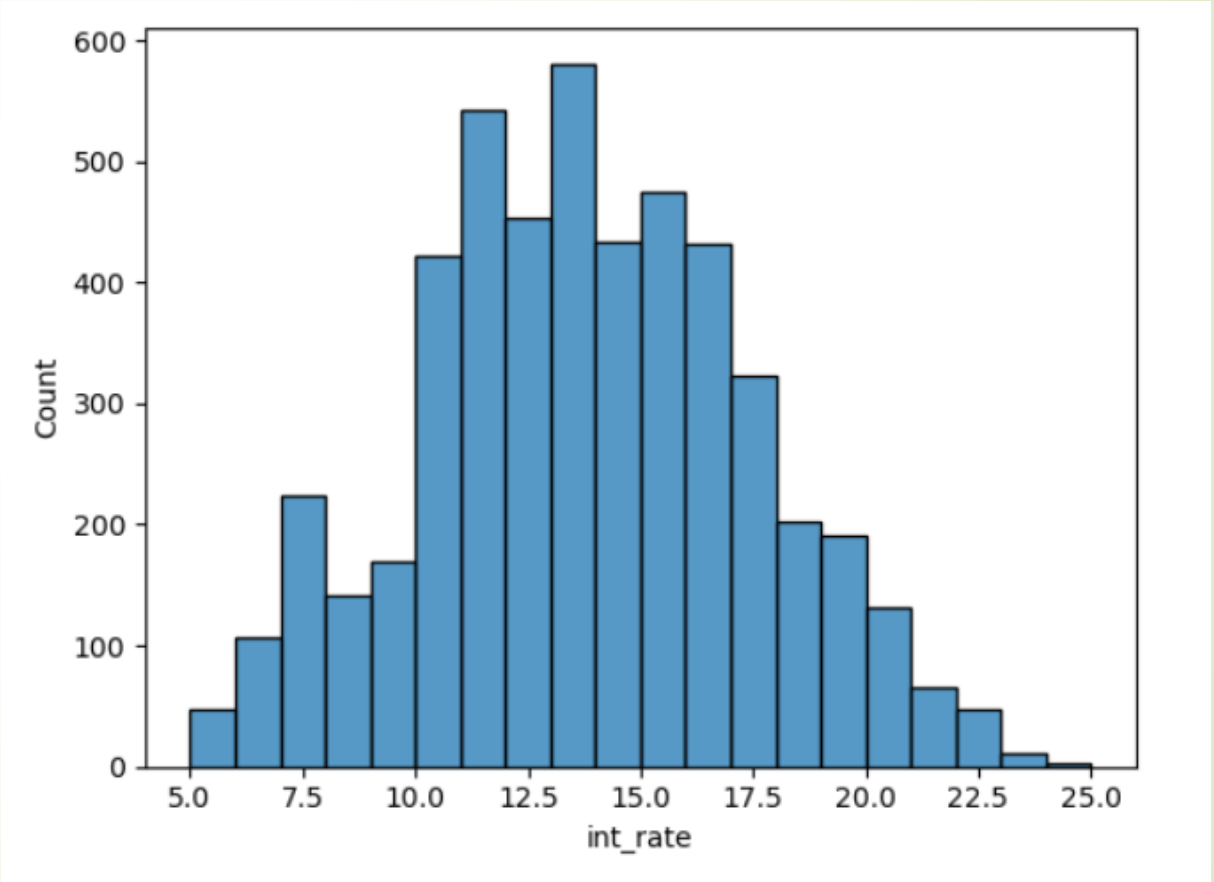
- Loans in the range of 5-15K are at higher risk of defaulting.
- The loan amount for charged off loans is left skewed, meaning that borrowers who borrow smaller amounts then to default more.



Univariant Analysis for Quantative Variable – Interest Rate

Insight:

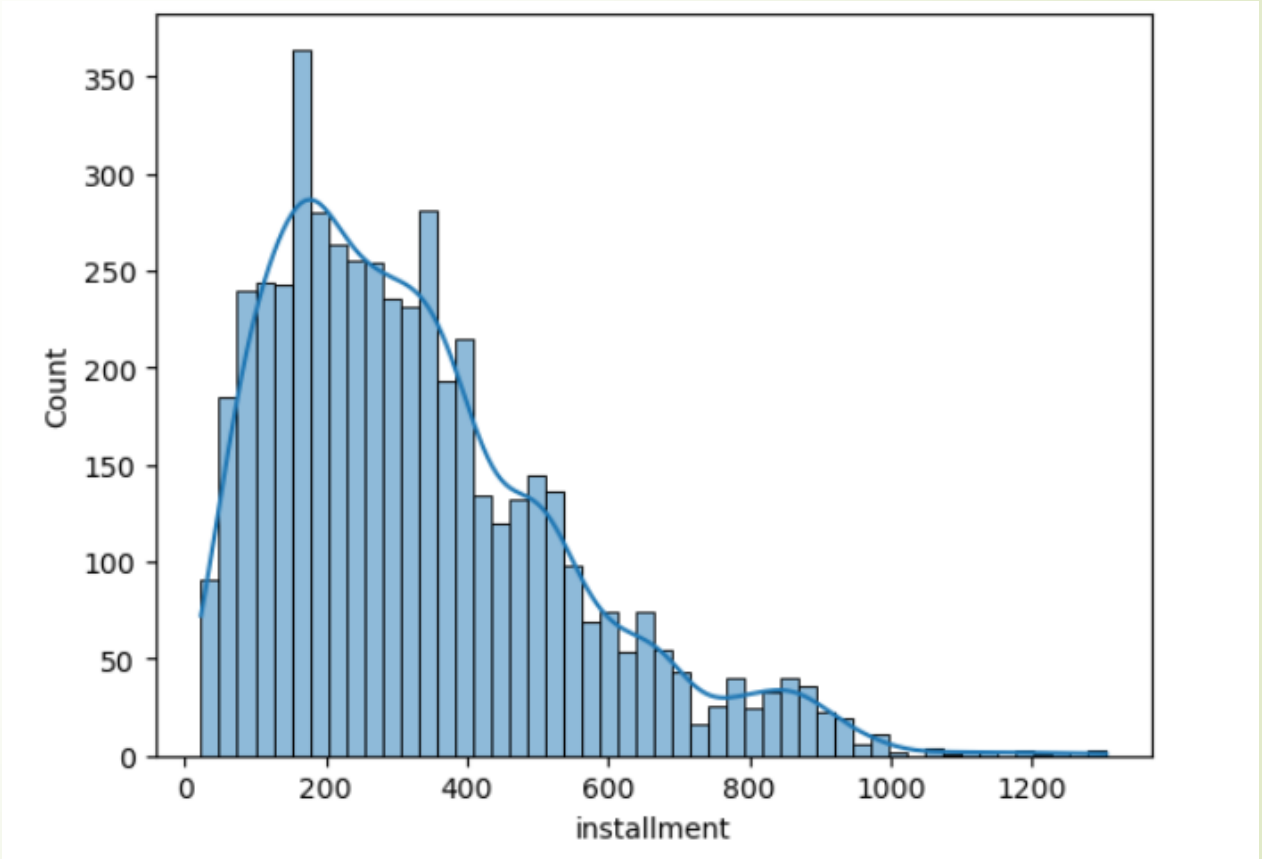
- We see a sharp uptick in defaulted loans between interest rate 10 & 17 after which the trend seems to slow down as int rate increases.



Univariant Analysis for Quantative Variable – Instalment

Insight:

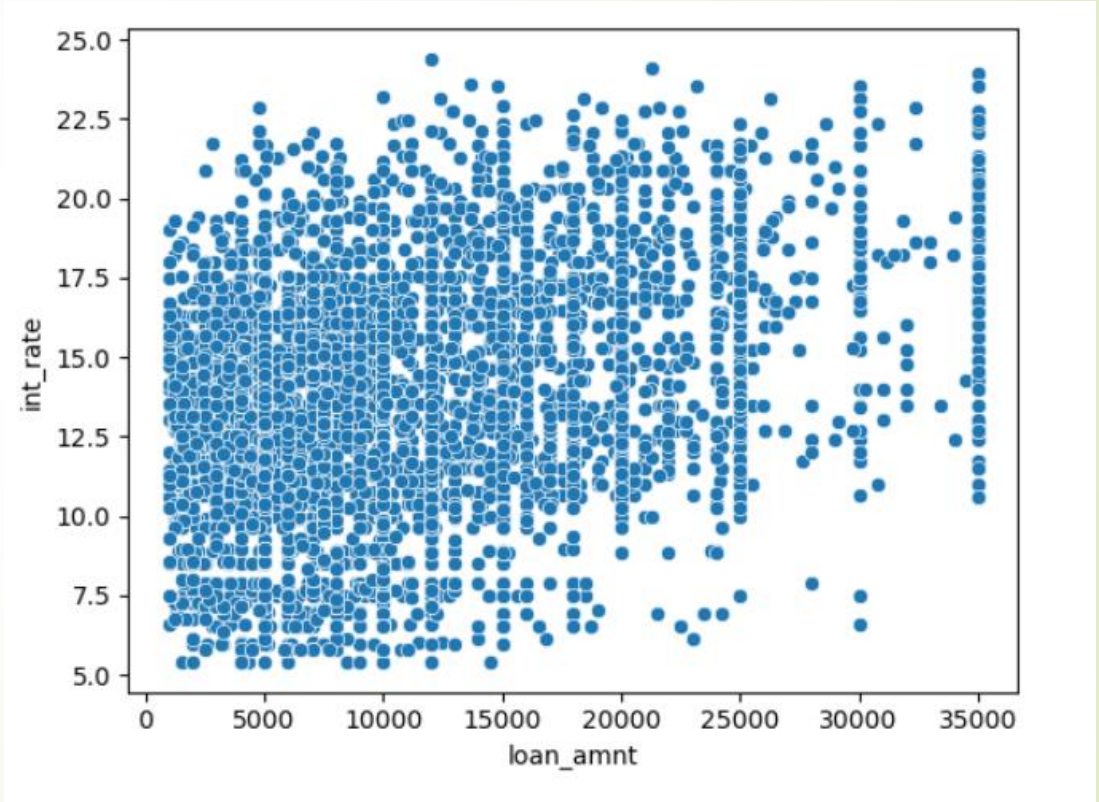
- From this we see that loans with installments between 75-400 faced issues with repayment.



Bi-Variant Analysis for Loan Amount and Interest Rate

Insight:

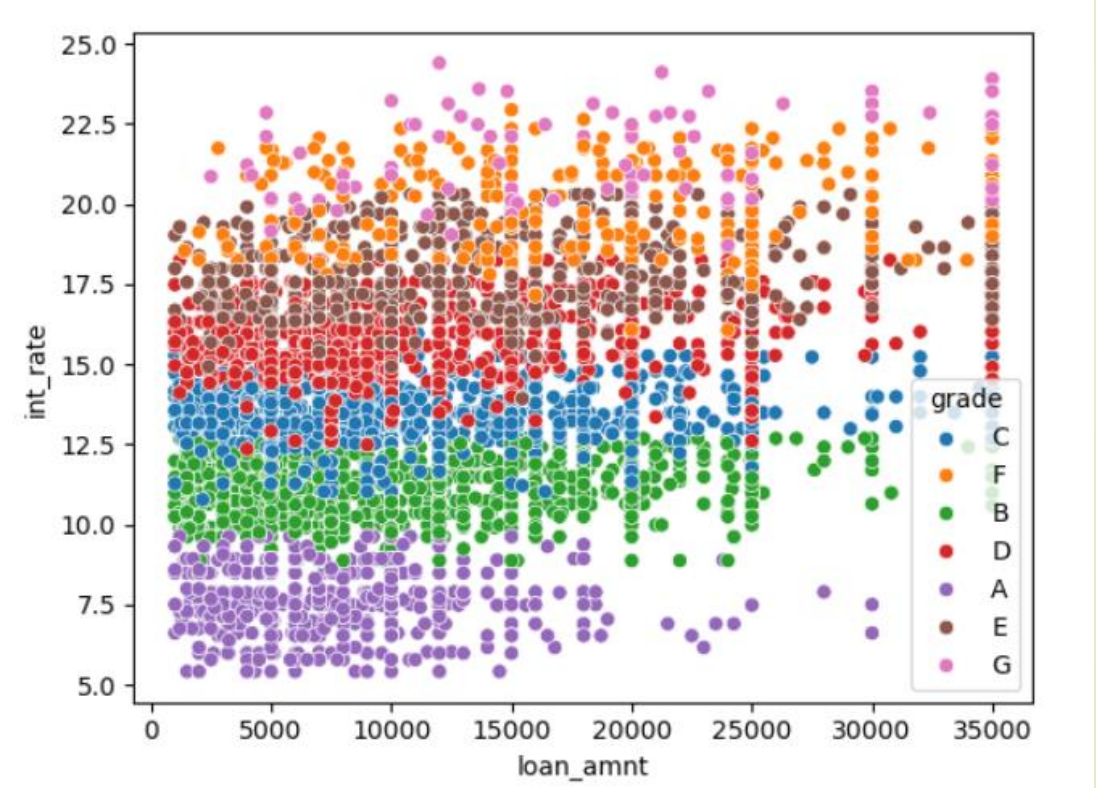
- From the above scatter plot, loan amount and int_rate does not tell us anything.
- Only thing that we can make from this is the loan is more concentrated on left side



Bi-variant Analysis for Loan Amount and Interest Rate Cont.

Insight:

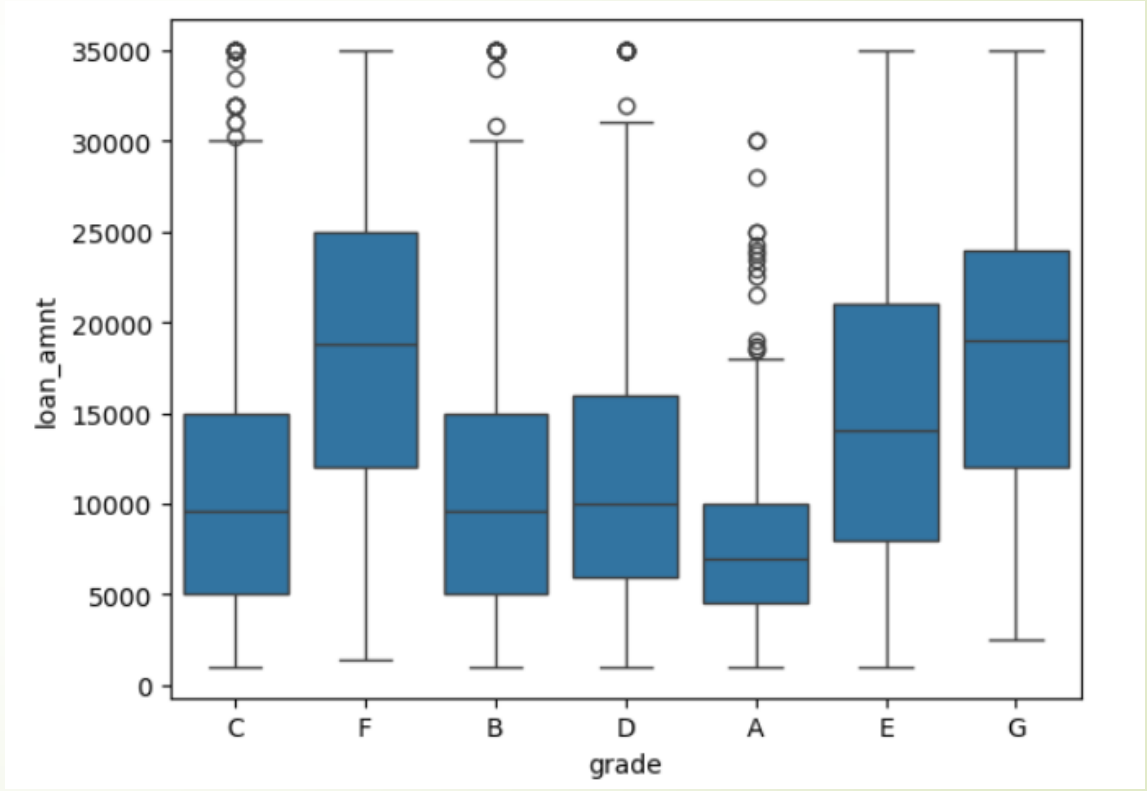
- High grade loans in A,B,C tend to have lower interest rate.
- Low grade loans which carry higher risk are offset with higher interest rate.



Bi-variant Analysis for Grade & Loan Amount

Insight:

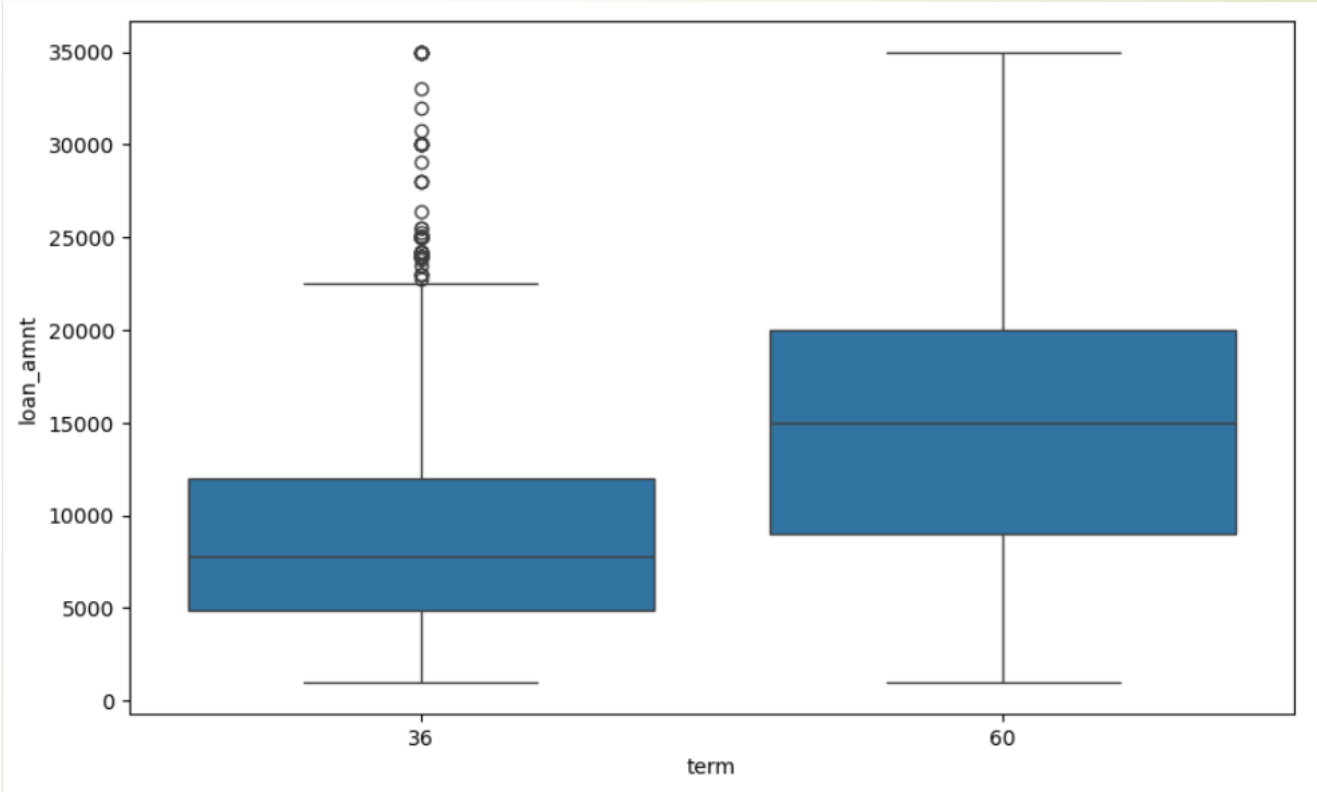
- Lower grade loans F,D,E and G have higher max and median values than higher grade loans A,B and C.
- Higher loan amounts were sanctioned for lower grade loans at a higher interest rate as inferred from last plot. Increasing the risk for LC.



Bi-variant Analysis for Loan Amount & Term

Insight:

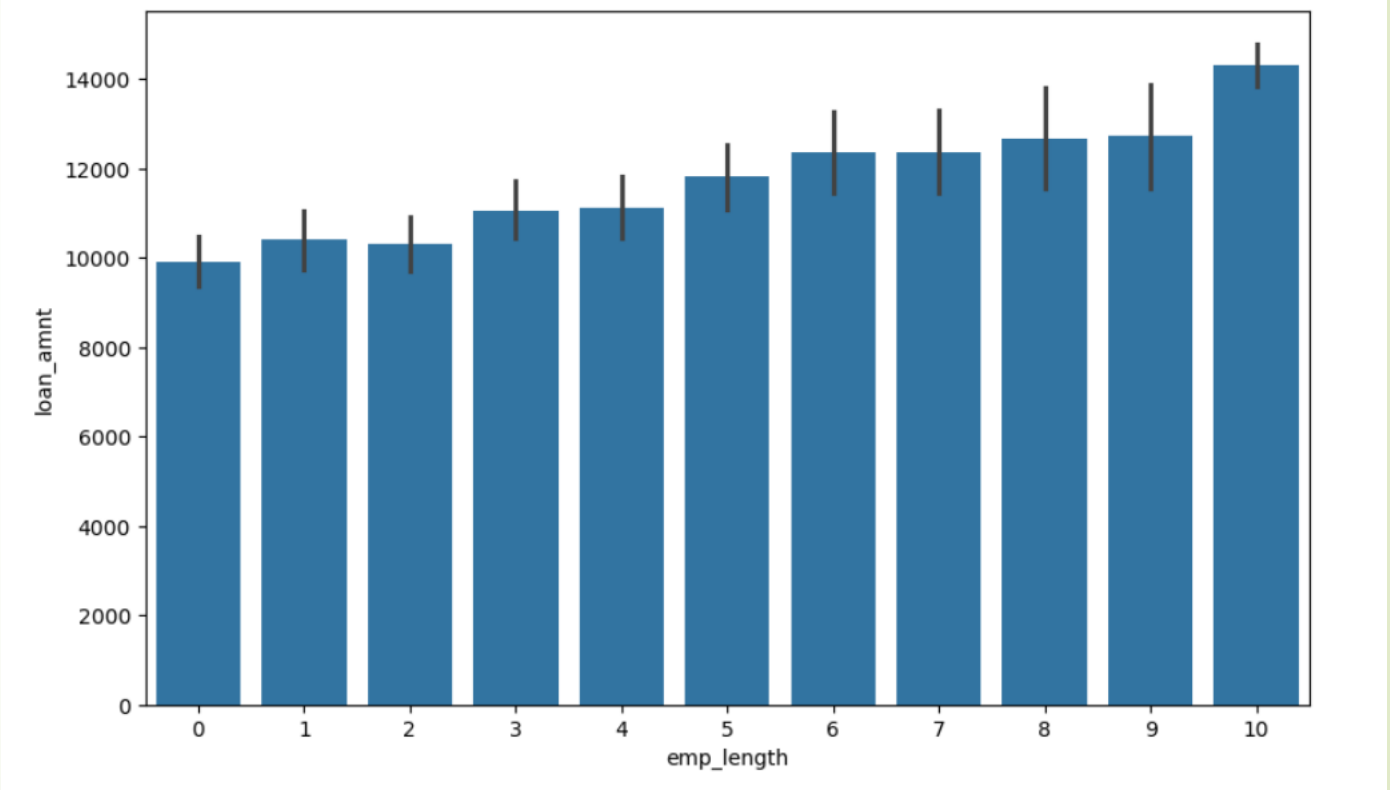
1. Lower term loans tend are disbursed for smaller amounts while higher terms loans have higher loan amounts.



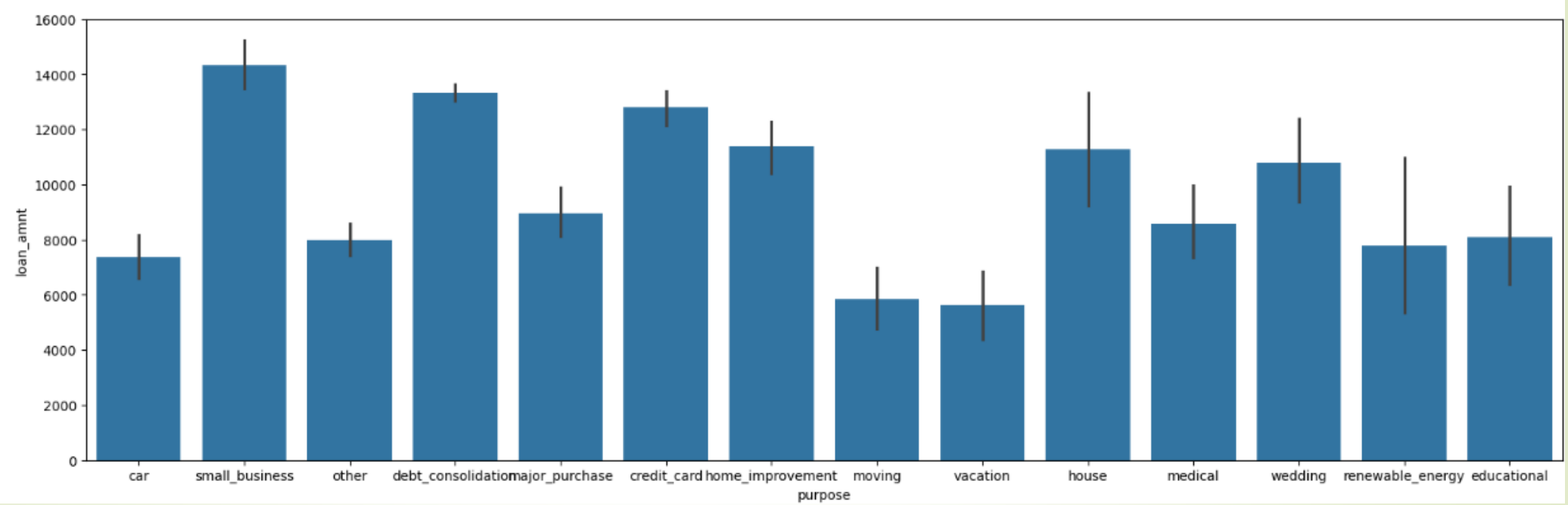
Bi-variant Analysis for Employee Length & Loan Amount

Insight:

- We see an upward trend in loan amount as the borrower employment length grows.
- This could be due to increasing financial stability and employees starting to settle in their personal lives.



Bi-variant Analysis for Purpose & Loan Amount



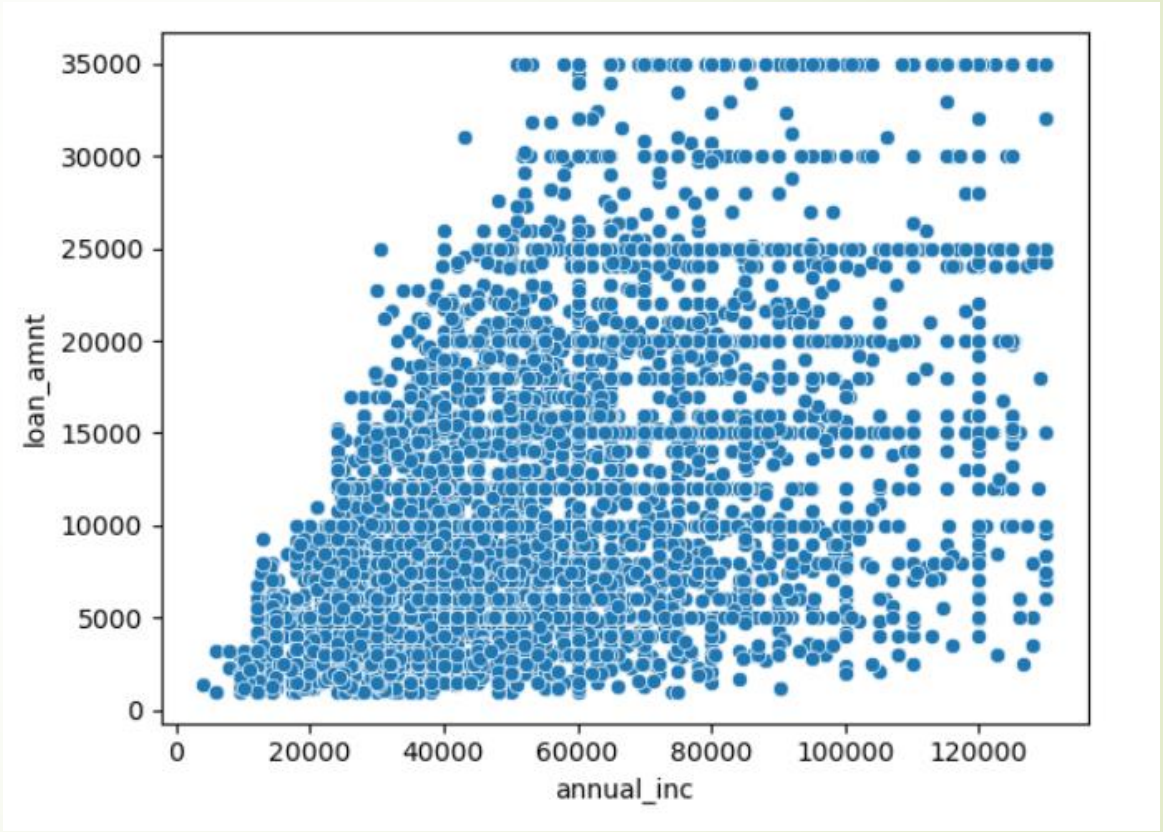
Insight:

1. Most high value loans are sanctioned for Small Business, followed by debt consolidation, house, home_improvement and wedding

Bi-variant Analysis for Annual Income & Loan Amount

Insight:

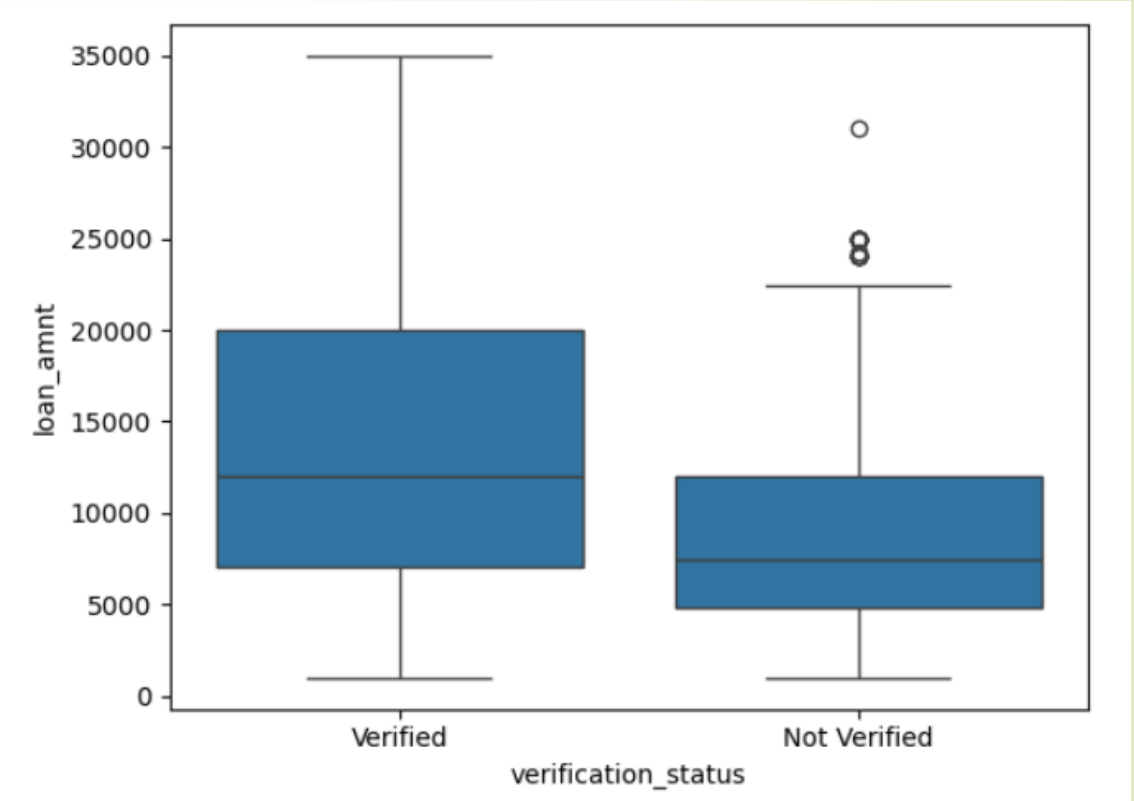
- Most of the loans are concentrated in the lower left and gradually scatter to up-right.
- Most defaulted loans are taken for smaller amounts and by people with lower income.



Bi-variant Analysis for Verification Status & Loan Amount

Insight:

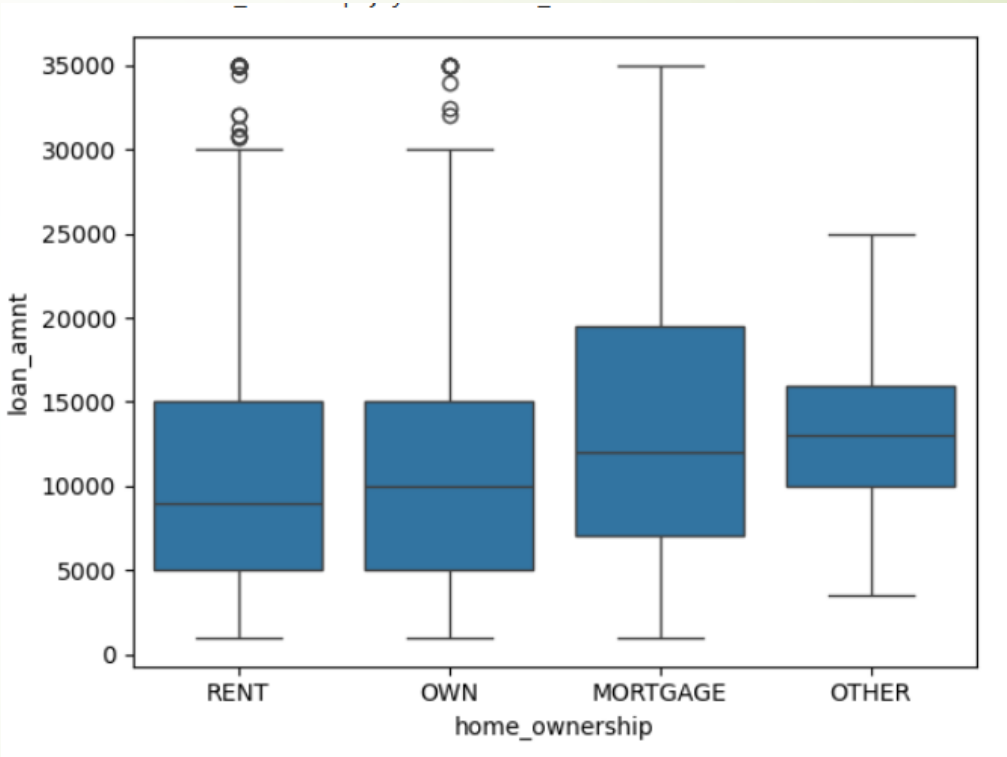
- Higher loan amounts are sanctioned for verified income status.
- There are some outlier in the not verified but looks like smaller loan amount are either not verified or rejected for which data is not available.



Bi-variant Analysis for Home Ownership & Loan Amount

Insight:

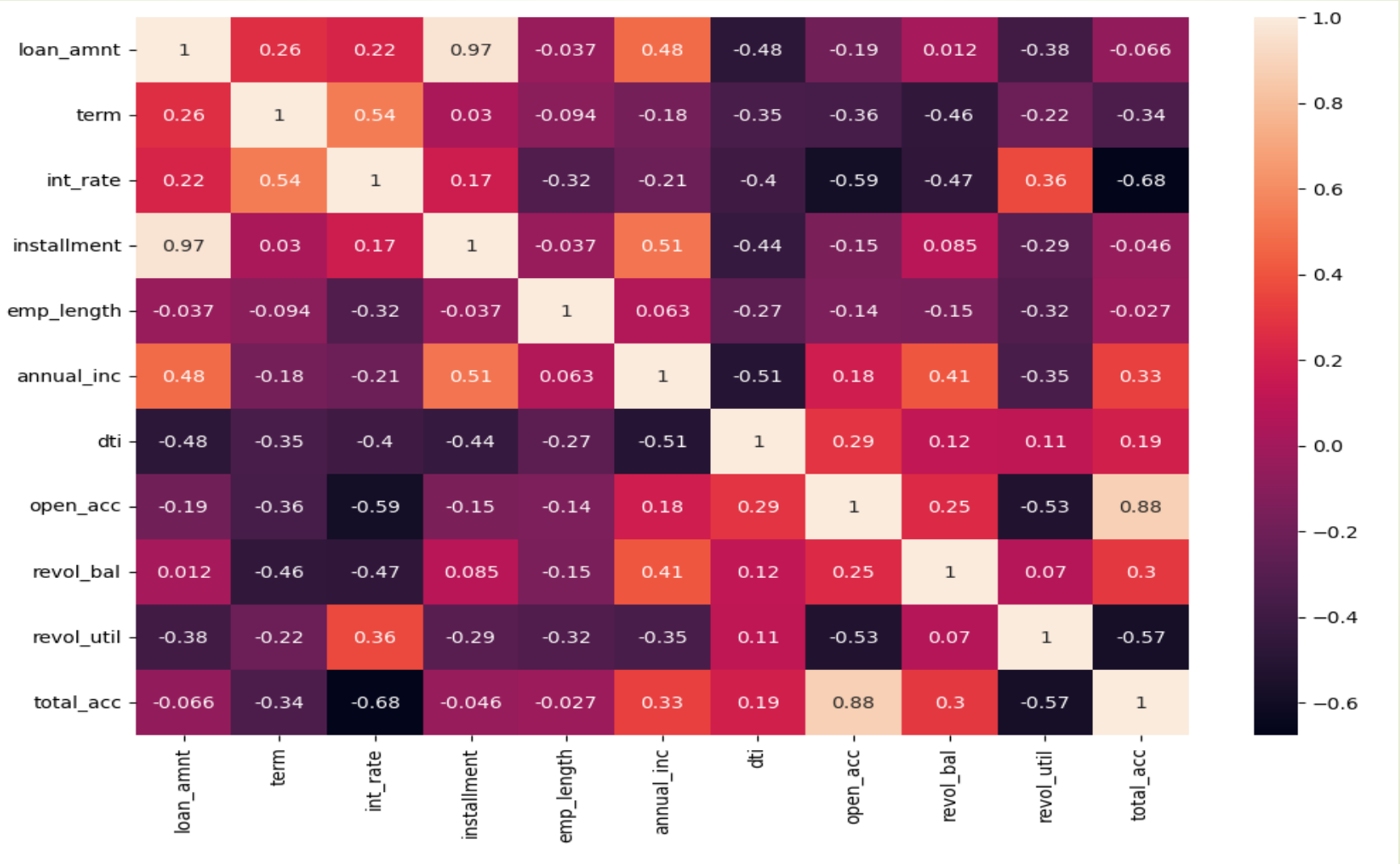
- People who are on rent or have own houses have similar loan requirement.
- People who have mortgages tend to go for higher loan amounts which could be explained by their higher financial responsibilities.



Correlation Analysis

- Correlation analysis is a statistical technique used to measure the strength and direction of the relationship between two or more variables.
- It quantifies the degree to which changes in one variable are associated with changes in another variable.
- Correlation analysis is widely used in various fields, including finance, economics, biology, psychology, and social sciences, to understand patterns and relationships in data. It ranges from-1 to 1.
  - r=1: indicates a perfect positive correlation
  - r=−1: indicates a perfect negative correlation
  - r=0: indicates no correlation between the variable

Correlation Analysis cont.



Correlation Analysis Cont.

Very Strong Correlations (.8 and above)	Strong Correlations (.6-.79)	Moderate Correlations (0.4-0.59)	Weak Correlations (0.2 -0.39)	Very Weak Correlations (0-0.19)
• loan amount with installment	• open accounts with total accounts	• term with loan_amnt • term with annual_inc • term with int_rate • installment with annual_inc • annual_incwith revol_balance	• loan_amnt with emp_length_mapping, int_rate and total_acc • int_ratewith installment and revol_util • installment with revol_baland total_acc • emp_lengthwith annual_inc • annual_incwith open_accand total_acc • dtiwith open_acc, revol_bal, revol_utiland total_acc • open_accwith revol_bal • revol_balwith revol_utiland total_acc	• int_rateand total_acc int_ratewith installment and revol_util • installment with revol_baland total_acc • emp_lengthwith annual_inc • annual_incwith open_accand total_acc • dtiwith open_acc, revol_bal, revol_utiland total_acc • open_accwith revol_bal • revol_balwith revol_utiland total_acc • loan_amntwith dti, revol_util and open_acc • term with installment, emp_length, annual_inc, dti,open_acc, revol_bal,revol_util& total_acc • int_ratewith emp_length, annual_inc,dti, open_acc, revol_bal, total_acc • installments with emp_length, dti, open_acc, revol_util • emp_lengthwith dti, open_acc, revol_bal, revol_util, total_acc • annual_incwith dti, revol_util • open_accwith revol_util

Analysis Summary & Recommendation

- **Loan Grade**
  - Grade B, C and D have the largest contribution in defaulted loans. LC needs better guidelines and assessments for grading the loans.
  - There isn't a clear pattern for sub-grades. LC needs a better framework to categories the loans.
- **Employment Experience**
  - Loans given to borrowers with 10+ Years of experiencers at a higher risk of being charged off.
  - There is a big jump (3%-24%) in loan defaults from 9 to 10+ Years, suggesting that granular details are needed for 10+ years experience category. Currently 10+ holds everyone with more than 9 years of experience.
- **Loan Term**
  - More than 50% of defaulted loans are taken for lower term, LC should consider the term decisions as small term loans tend to have higher installments which can impact repayment capacity

Suggestions

- **Implement Stricter Criteria for Grades B, C, and D:** Consider implementing stricter risk assessment and underwriting criteria for applicants falling into Grades B, C, and D to minimize default risks.
- **Focus on Subgrades B3, B4, and B5:** Pay special attention to applicants with Subgrades B3, B4, and B5. Consider additional risk mitigation measures or offering lower loan amounts for these subgrades to reduce default rates.
- **Evaluate and Limit 60-Month Loans:** Evaluate the risk associated with 60-month loans. Consider limiting the maximum term or adjusting interest rates for longer-term loans to decrease the likelihood of defaults.
- **Comprehensive Credit Scoring System:** Develop a comprehensive credit scoring system that incorporates various risk-related attributes, as experience alone might not be sufficient to gauge creditworthiness.
- **Capitalizing on Market Growth:** Capitalize on the market's growth trend observed from 2007 to 2011 by maintaining a competitive edge in the industry while ensuring robust risk management practices.
- **Anticipate Peak Periods:** Anticipate increased loan applications during peak periods such as December and Q4. Ensure efficient processing to meet customer demands during these busy seasons.
- **Review Verification Process:** Review the verification process to ensure effective assessment of applicant creditworthiness. Consider improvements or adjustments based on the review findings.
- **Careful Evaluation for Debt Consolidation Loans:** Carefully evaluate applicants seeking debt consolidation loans, considering potential interest rate adjustments or offering financial counseling services to manage the associated risks

Thank You!!