LENDING CLUB CASE STUDY (EDA)

Members

1. Vinay Kumar
2. Sharath kurup

Contents

Business Objectives

- Lending Club is a consumer finance marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return. Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss.

- The objective is to pinpoint 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.

- We need to identify patterns which indicate if a person is likely to default (charged off), which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. and With the help of loan, consumer attributes and loan attributes influence the tendency of default.

- In essence, lending club company want to clearly understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. Lending company can utilize this knowledge for its portfolio and risk assessment.

Data Description

- We are being provided with historical customer loan application data. It contain customer's credit history and lending club information.
- We have around 111 columns and 39k+ records.
- This provide us ample data to perform us to perform analysis and understand their relationship between each other and identify their effect on success or defaulting of a loan.
- We will be using only those column that will have a direct impact on status of loan. After identifying them we will create required dataset by performing data cleaning and conversion.

Problem solving methodology

The data analysis consists four main parts:

- Data understanding
- Data cleaning (cleaning missing values, removing duplicate columns and so)
- Data Analysis
- Recommendations

Data Understanding

Leading Attribute

- Loan Status - Key Attribute (*loan_status*). The column has three distinct values
  - Fully-Paid - The customer has paid the loan completely
  - Charged-Off - The customer is "Charged-Off" or has "Defaulted"
  - Current - These are in progress loan payments and cannot contribute to conclusive evidence if the customer will default of pay in future
    - For the given case study, "Current" status rows will be ignored

Important Columns

Following are leading columns for EDA

- **Customer Related Data**
  - Annual Income (annual_inc) - Annual Income of the Customer.
  - Home Ownership (home_ownership) - Wether the customer owns a home or not.
  - Employment Length (emp_length) - Employment tenure of a customer.
  - Debt to Income (dti) - The percentage of the salary which goes towards paying loan. Lower DTI is normally preferred while approval of loan
  - State (addr_state) - Location of the customer.

Data Understanding -II

- **Loan Attributes**
  - Loan Amount (loan_amt)
  - Grade (grade)
  - Term (term)
  - Loan Date (issue_date)
  - Purpose of Loan (purpose)
  - Verification Status (verification_status) - whether the customer details are verified or not.
  - Interest Rate (int_rate)
  - Installment (installment)
  - Public Records Bankruptcy (public_rec_bankruptcy) - Number of bankruptcy records publically available for the customer.

Data Understanding -III

- **Ignored Columns**
  - Customer Behavior Columns - Customer behavior variables generate post the approval of loan applications. Thus, these attributes will not be considered towards the loan approval/rejection process. So, these will be removed

```
'delinq_2yrs', 'earliest_cr_line', 'last_pymnt_amnt','inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc',
                    'total_pymnt','total_pymnt_inv', 'total_rec_prncp',
                   'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee',
                  'last_pymnt_d','last_credit_pull_d'snn
```

  - Granular Data - Granular column like sub grade will be removed.Columns which describe next level of details which may not be required for the analysis as grade column can be used.

Data Cleaning and Preprocessing

1. Load the CSV file in loan_dataset dataframe
2. Remove all the columns that have just NA columns
3. Remove the current loan records
4. Remove all the columns which has more than 60 % missing values.
5. Remove ignored columns(behavioral and granular)
6. Remove description or textual columns
7. Remove behavioral columns
8. Remove fund_amnt and fund_amnt_inv as they have high positive correlation with loan amnt
9. Fix the data discrepancy such as data type, blanks and duplicates.
10. Create bins/categories for loan characteristics.
11. Create derived columns from Issue_date to check year and month wise.
12. Concluded column required for analysis
13. Common function

Univariate Analysis

Univariate analysis is deals analysis of a single variable to understand its distribution, central tendency and dispersion.

➢ *Categorical variable*

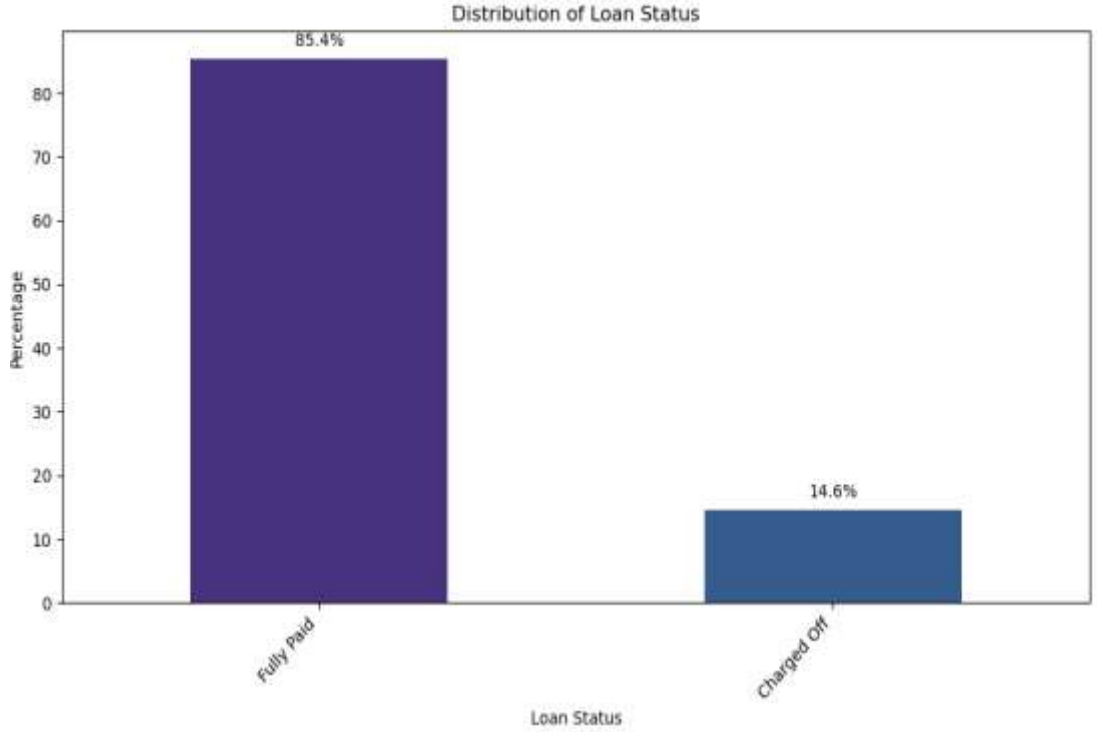
Ordered	UnOrdered
Grade(grade)	Loan Status(loan_status)
Employment tenure(emp_length)	Address State(addr_state)
Term(term)	Home Ownership(home_ownership)
Issue Year(issue_year)	Verification status(verification_status)
Issue Month(issue_month)	Purpose(purpose)

➢ *Quantitative variable*

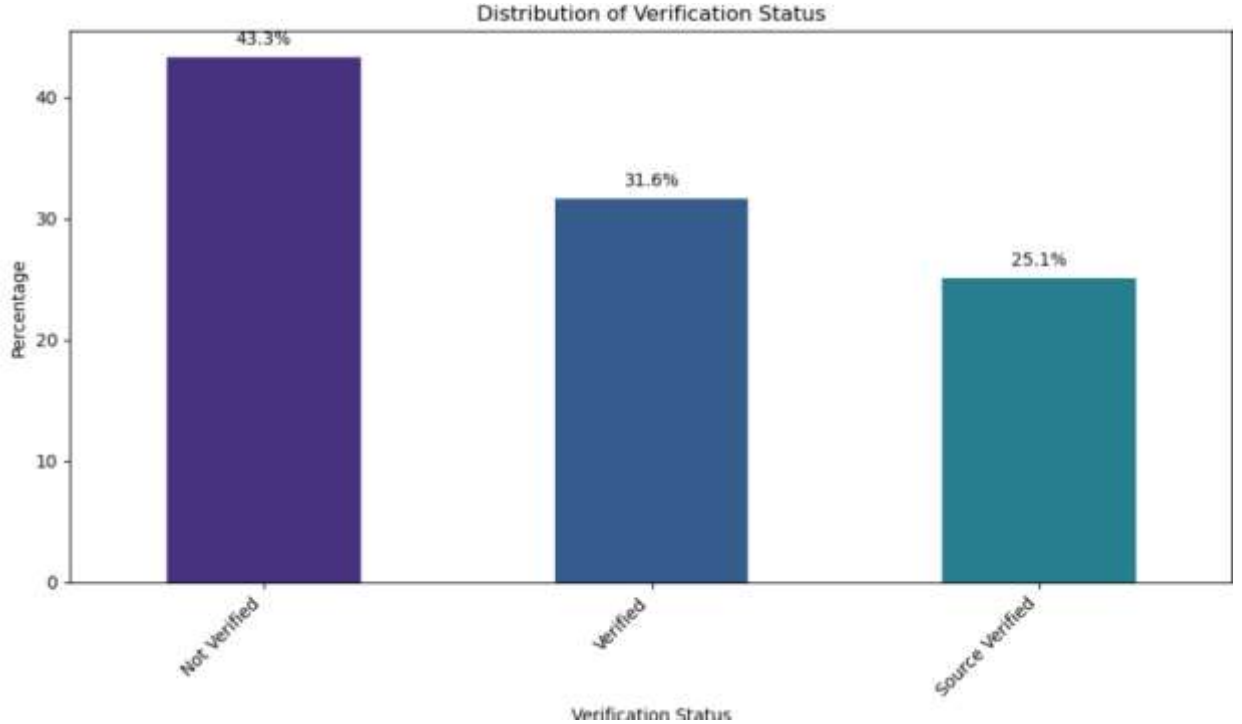
- Interest Rate (int_rate)
- *Debt to income ratio(dti)*
- *loan amount (loan_amnt)*
- *Annual income (annual_amnt)*
- *Monthly instalments(installment)*
- *Public record of bankruptcy(pub_rec_bankruptcies)*

Univariate Analysis(Unordered categorical)

Loan Status



Verification Status

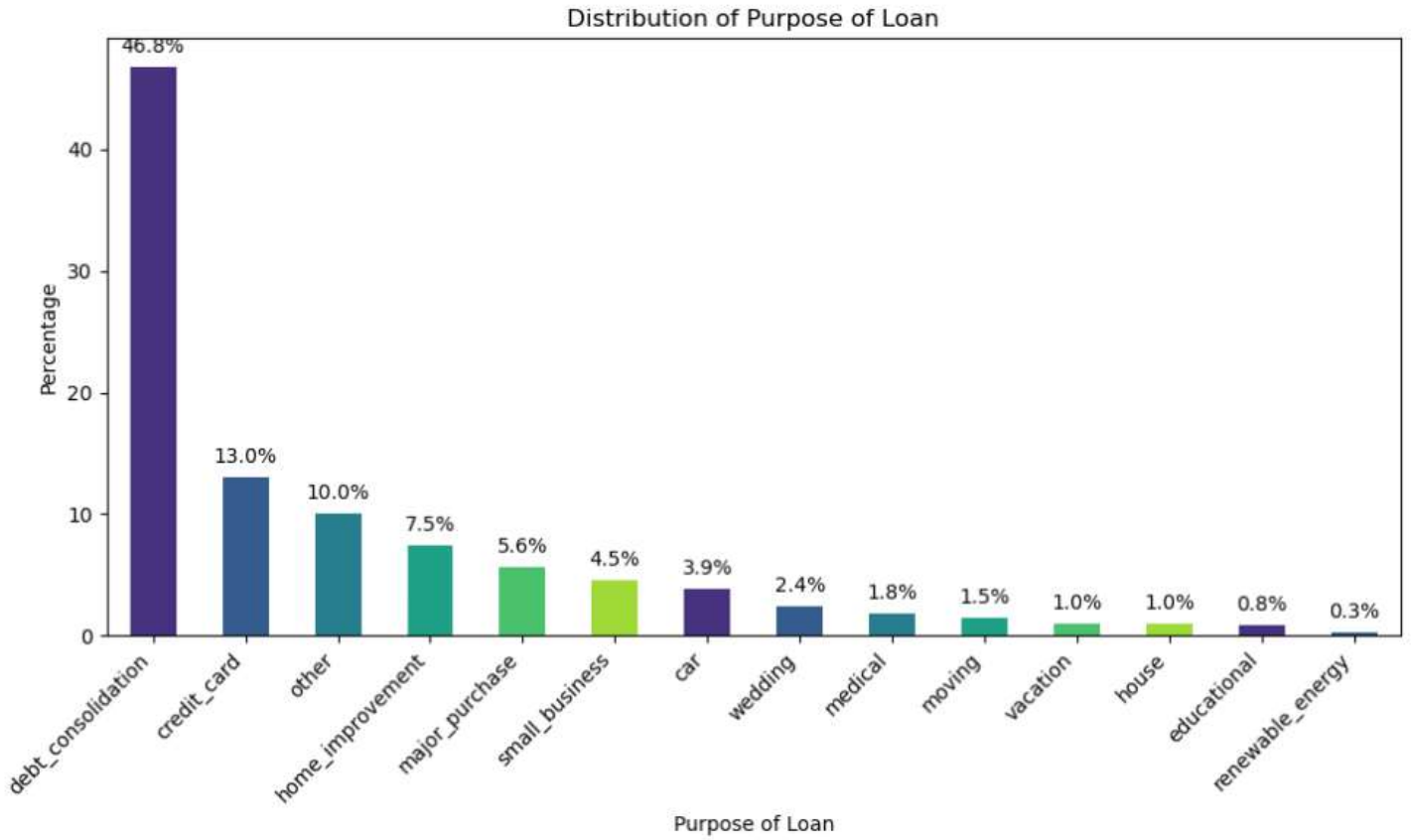


Observation: *Around 14.6% of the loan are charged off. Which is high by industry standard*

Observation: More than 40% of loan application are not verified. This is right behavior

Univariate Analysis(Unordered categorical)

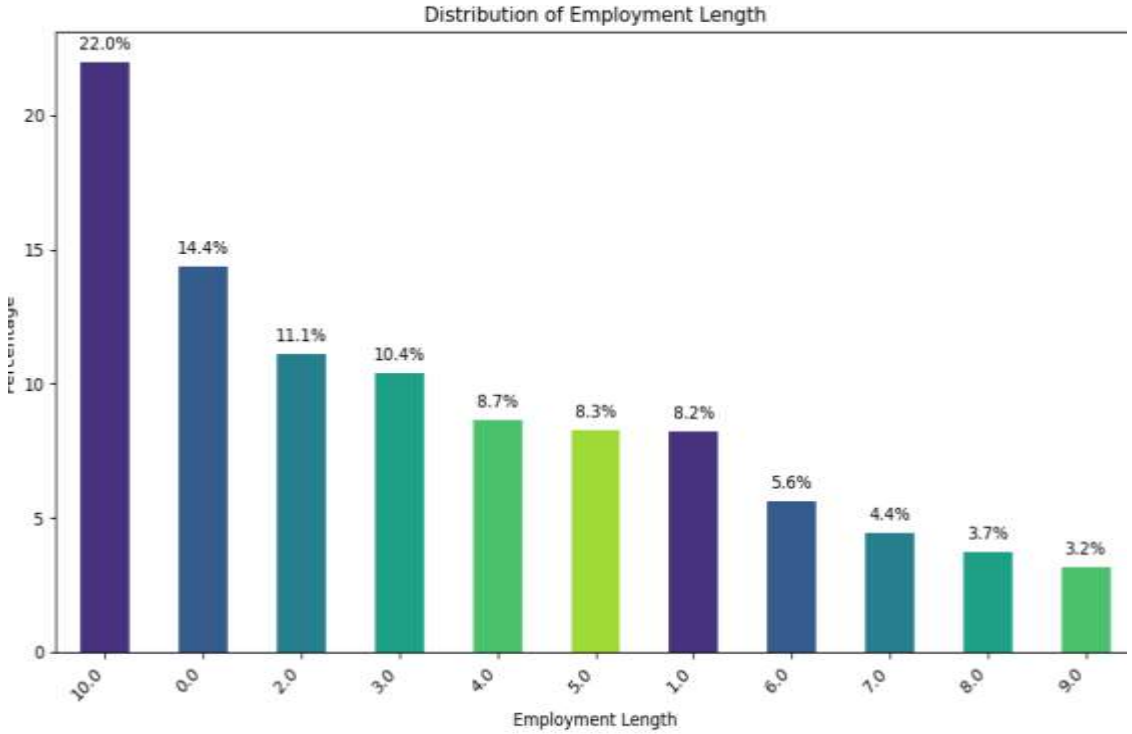
Loan Purpose



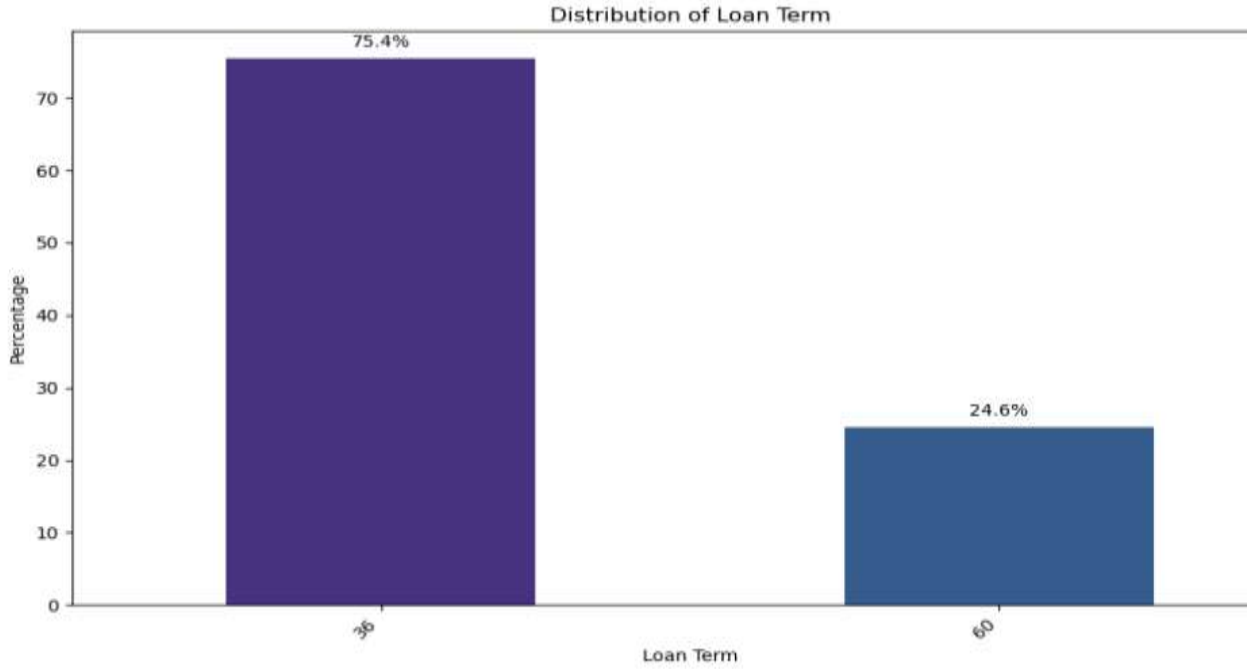
Observation – Majority of the loan is taken for debt consolidation and credit card payment

Univariate Analysis(Ordered categorical)

Employment Length



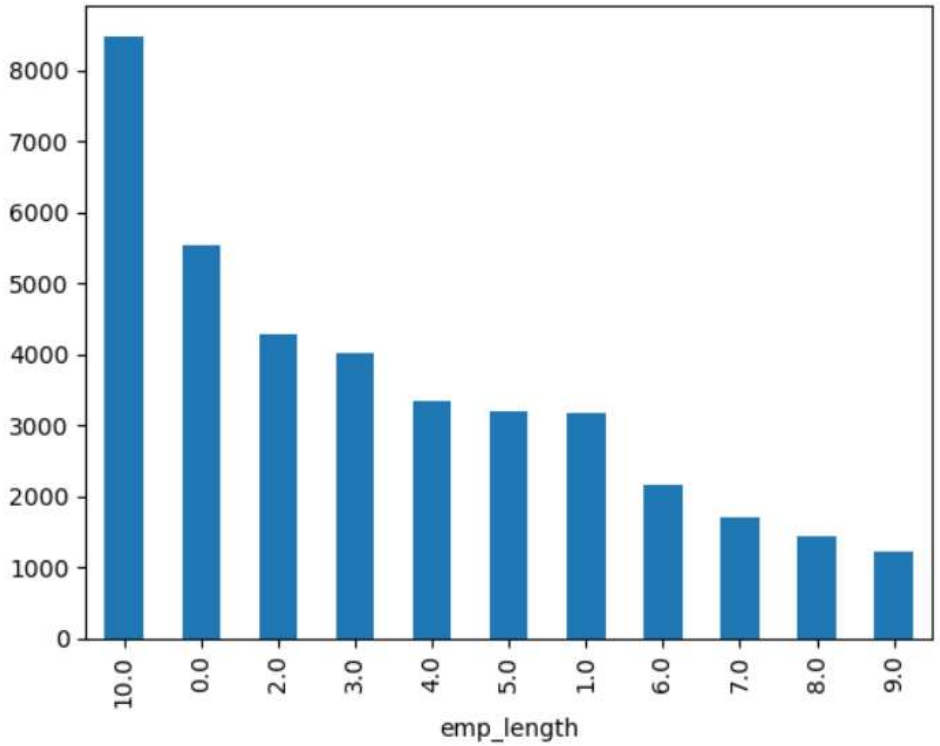
Loan Term



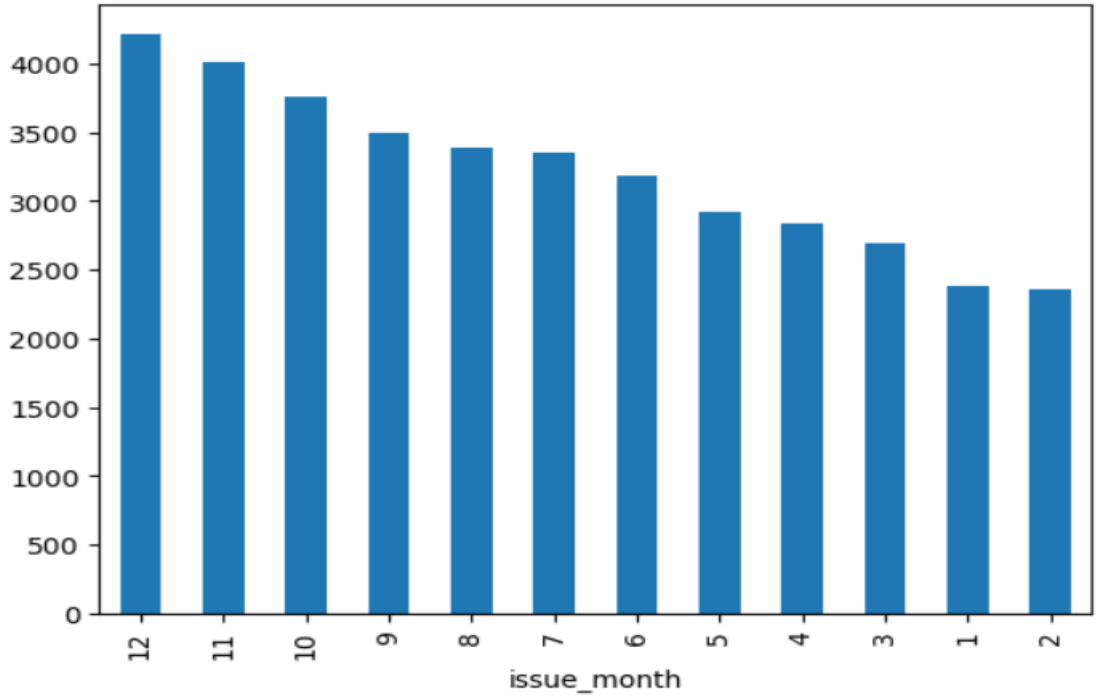
Observation: *Customer with 10+ employment tenure are more likely to take loan*
Observation: 3/4 of loan is for 36 month and 1/4 is for 60 month. Higher term usually result in default by industry standards

Univariate Analysis(Ordered categorical)

Loan Issue Year



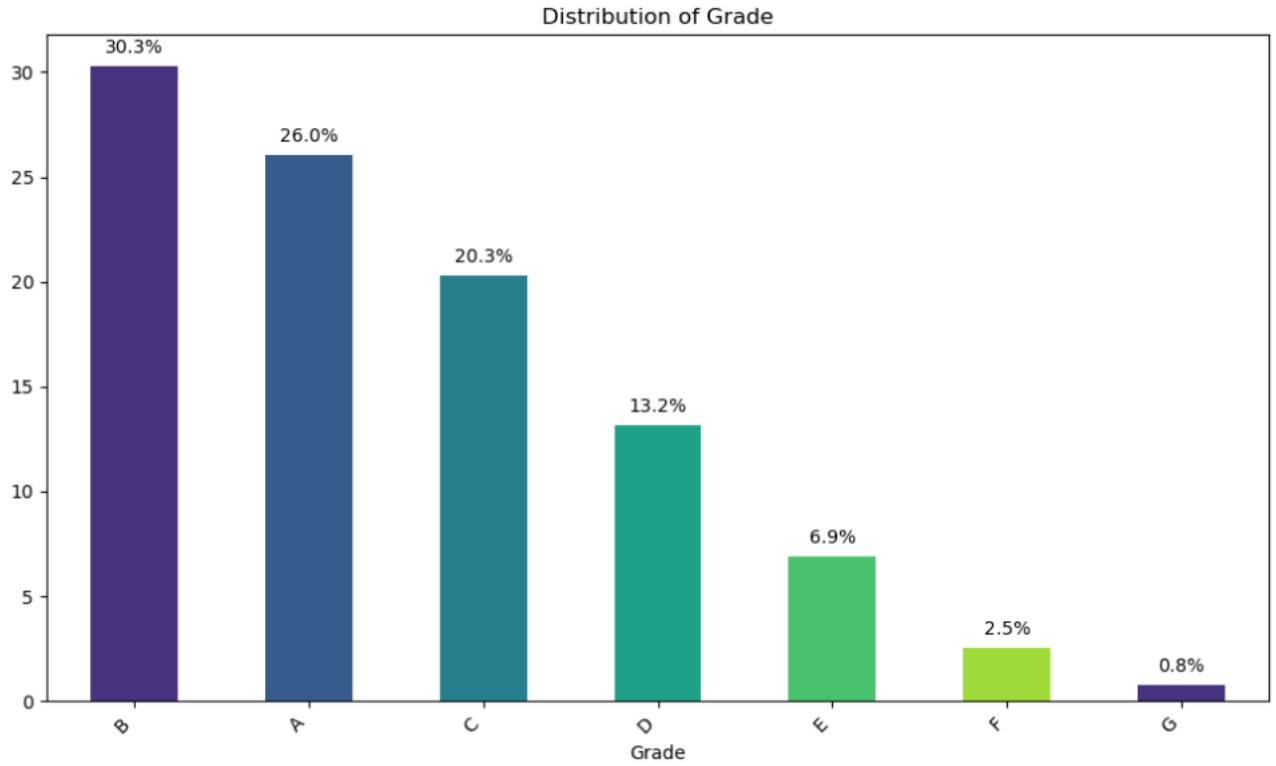
Loan Issue Month



Observation: *Number of loan approved are increasing per year*
Observation: Most of the loan is taken in last quarter.

Univariate Analysis(Ordered categorical)

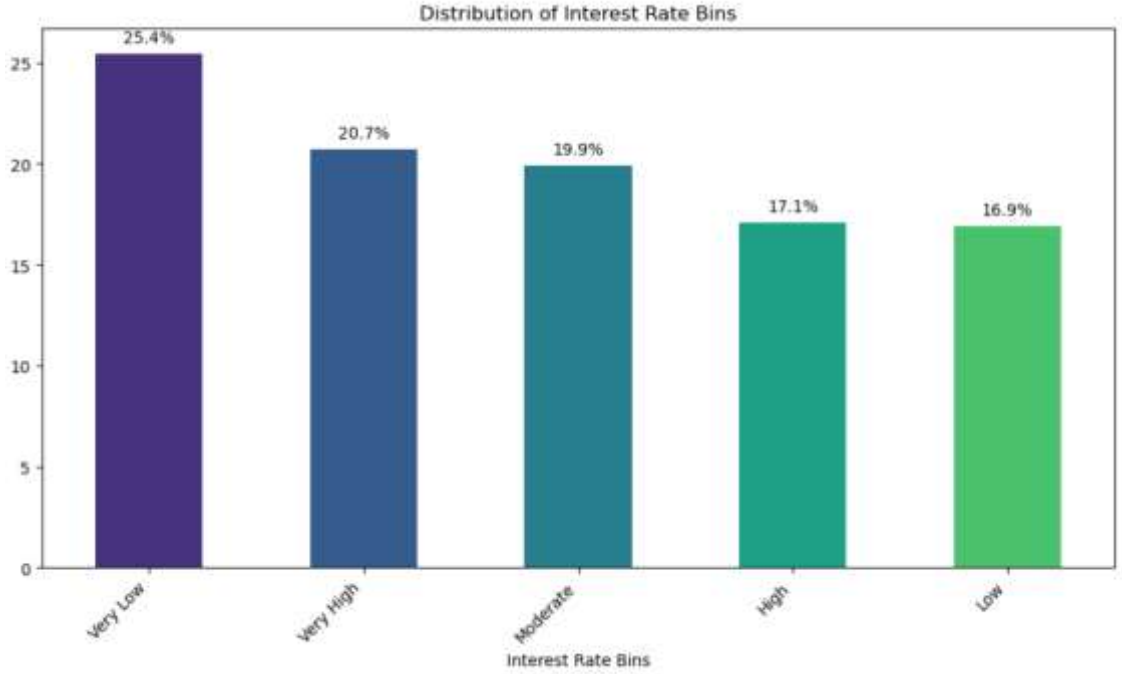
Grade

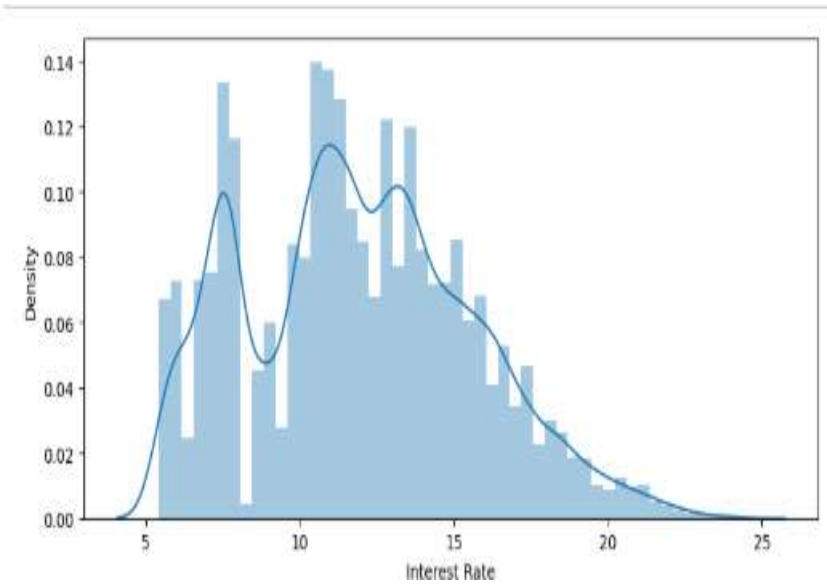


Observation: *Majority of loan have high grade*

Univariate Analysis(Quantitative)

Interest Rate





Observation: There appear to be two main peaks:

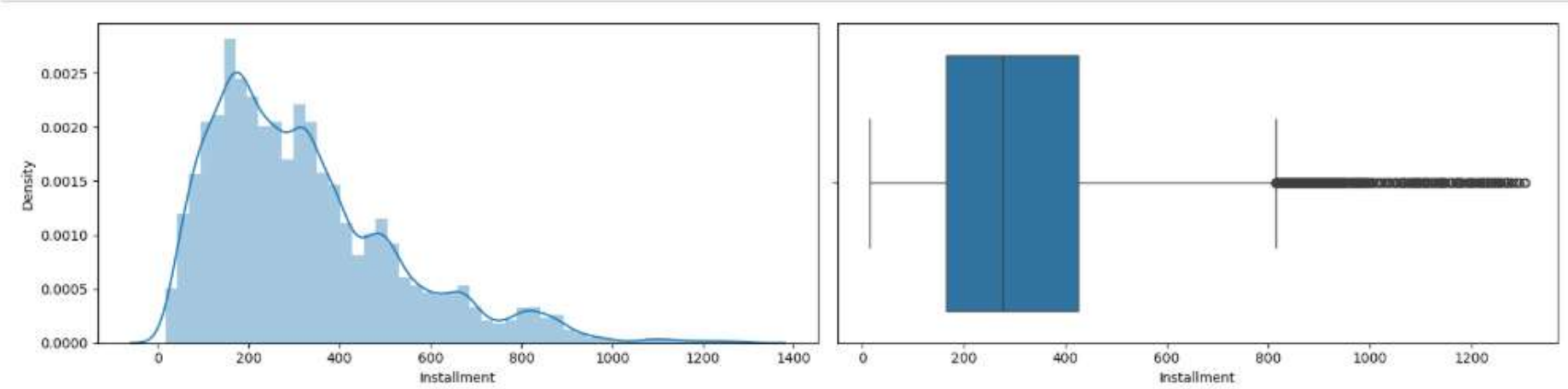
One around 7-8%

Another around 10-13%

We can say most of the loan interest rate are between 5 to 13

Univariate Analysis(Quantitative)

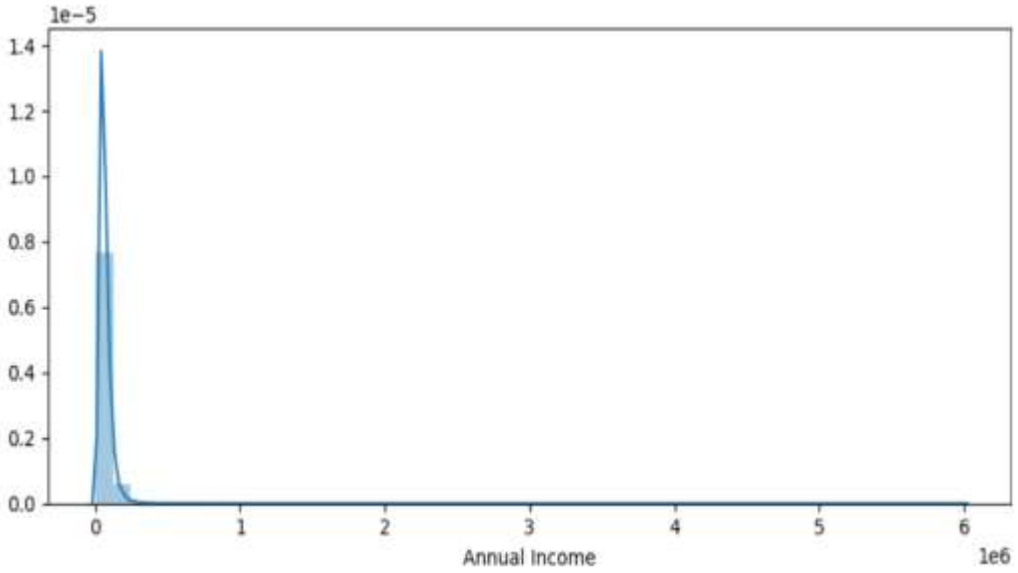
Installment

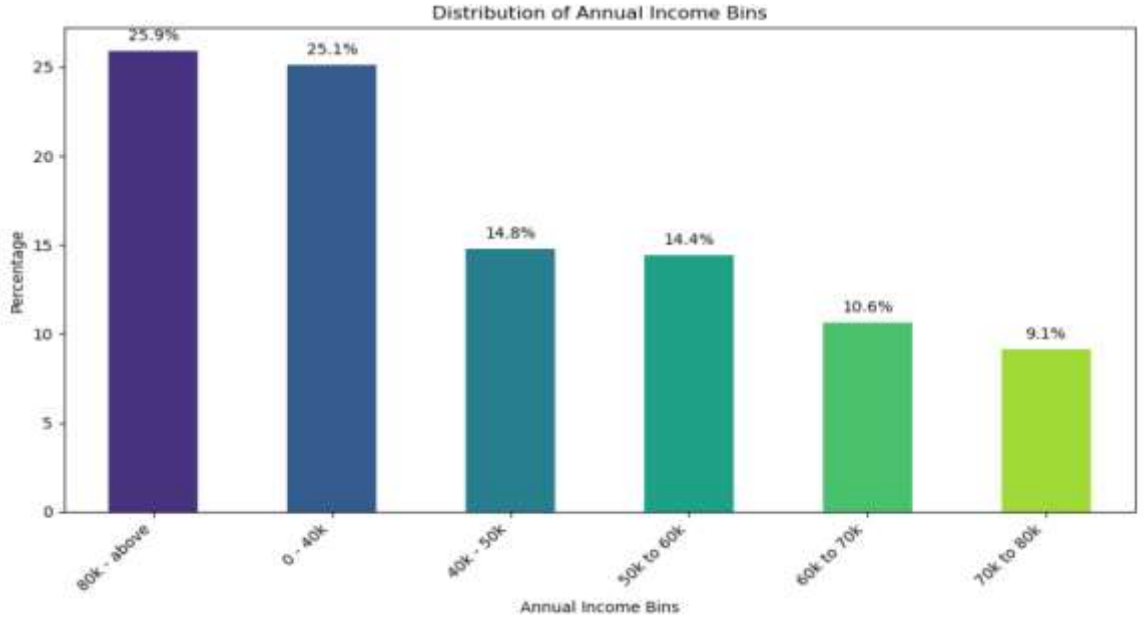


Observation: *Most of the loan seems to be having installment between 200 to 400.*
There are few high installments, this could be because of high loan amount with high interest

Univariate Analysis(Quantitative)

Annual Income

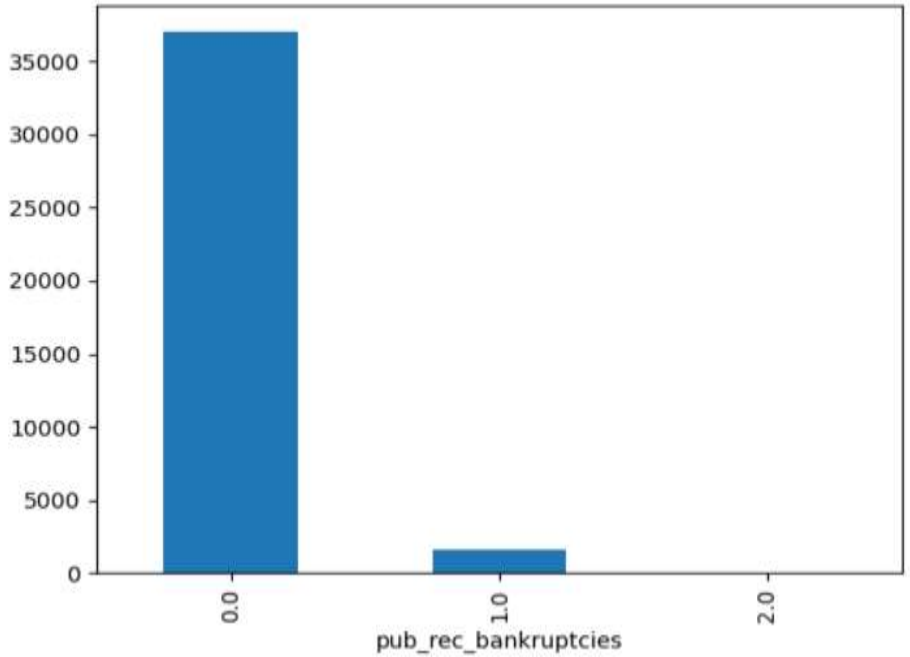




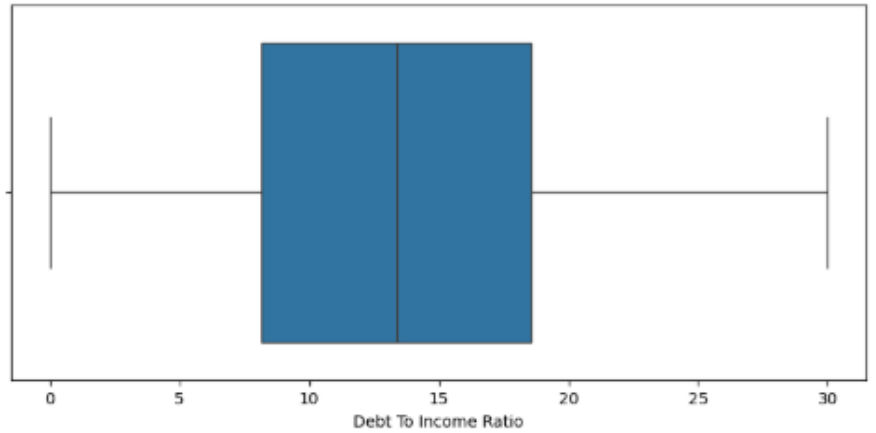
Observation: **There are applicant with very high income.**

Univariate Analysis(Quantitative)

Public record of bankruptcy

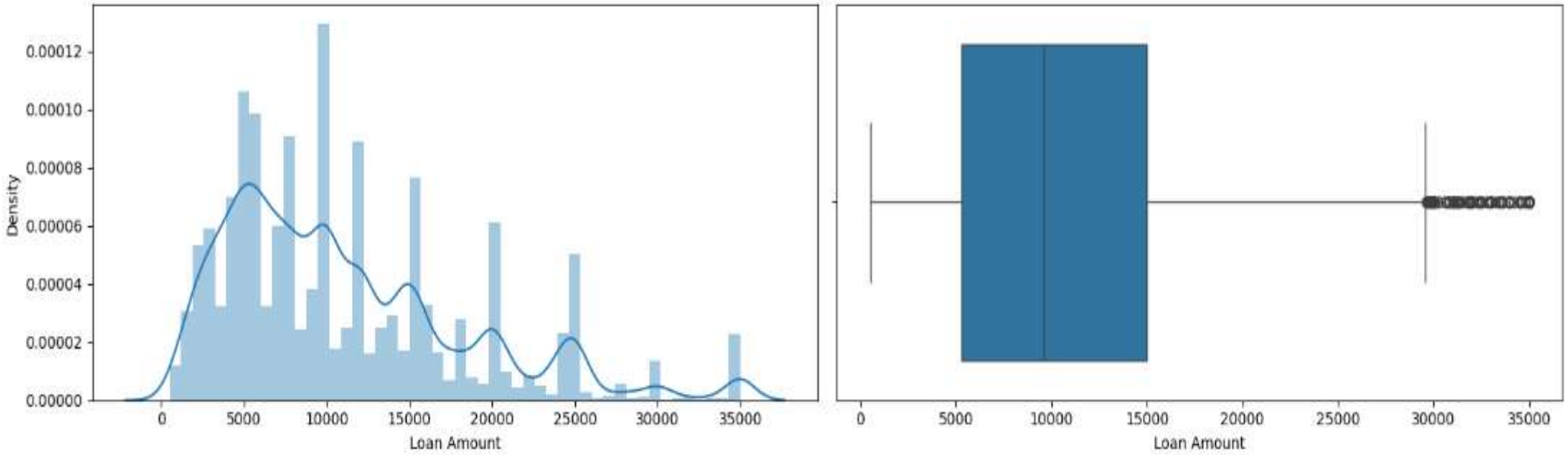


Debt to Income Ratio



Observation: Majority of the loan applicant does not have bankruptcy record**.** Most of the loan has DTI between 8 to 18.

Univariate Analysis(Quantitative)



Observation : Based on both the histogram and the box plot, the majority of loan amounts are concentrated betwee $5,000 and $15,000,with a particularly high concentration between $5,000 and $10,000.

Loan amounts above $20,000 are less frequent.

Univariate Analysis

Observation & Inference

- Around 14.6% of the loan are charged off. Which is high by industry standard.
- Majority of application comes from state CA.
- Majority of applicant does not have own house. They are rented or mortgaged.
  This could be risky as house act as a good collateral.
- Majority of application are not verified. This is not the best practice.
- 3/4 of loan is for 36 month and 1/4 is for 60 month. Higher term usually result in default by industry standards.
- majority of loan amounts are concentrated between 5,000 and 15,000
- Majority of interest rate is between 5 to 13.
- Some borrowers are paying significantly higher interest rates. This might be due to lower credit scores, higher risk loans, or specific loan characteristics.
- Debit consolidation and credit card is the most common reason to take loan
- Most of the loan are taken in the last quarter of the year. This fits with the loan purpose. As people usually try to clear their debt by EOY

Univariate Analysis

Observation & Inference

- There is gradual increase in number of loans per year
- Majority of the loan applicant has 10+ year of experience. This is because complacency is shown for such candidate
- Majority of loan has B grade. Overall, most of them have high grade.
- Max number of loan have 200- 400 installment amount
- Majority of the applicant either have very high income or have very low income.

Bivariate Analysis

Bivariate analysis aims to determine the relationship 2 variables(factor). The analysis can be used to test hypotheses, identify patterns, or explore relationships between the variables.

It was carried out for both Categorical and Quantitative Variables

- **Categorical Variables**:

Ordered	UnOrdered
Grade(grade)	Loan Status(loan_status)
Employment tenure(emp_length)	Address State(addr_state)
Term(term)	Home Ownership(home_ownership)
Issue Year(issue_year)	Verification status(verification_status)
Issue Month(issue_month)	Purpose(purpose)

- **Quantitative variable**

* Interest Rate Bucket(int_rate_bucket)

* Debt to income ratio Bucket(dti_bucket)

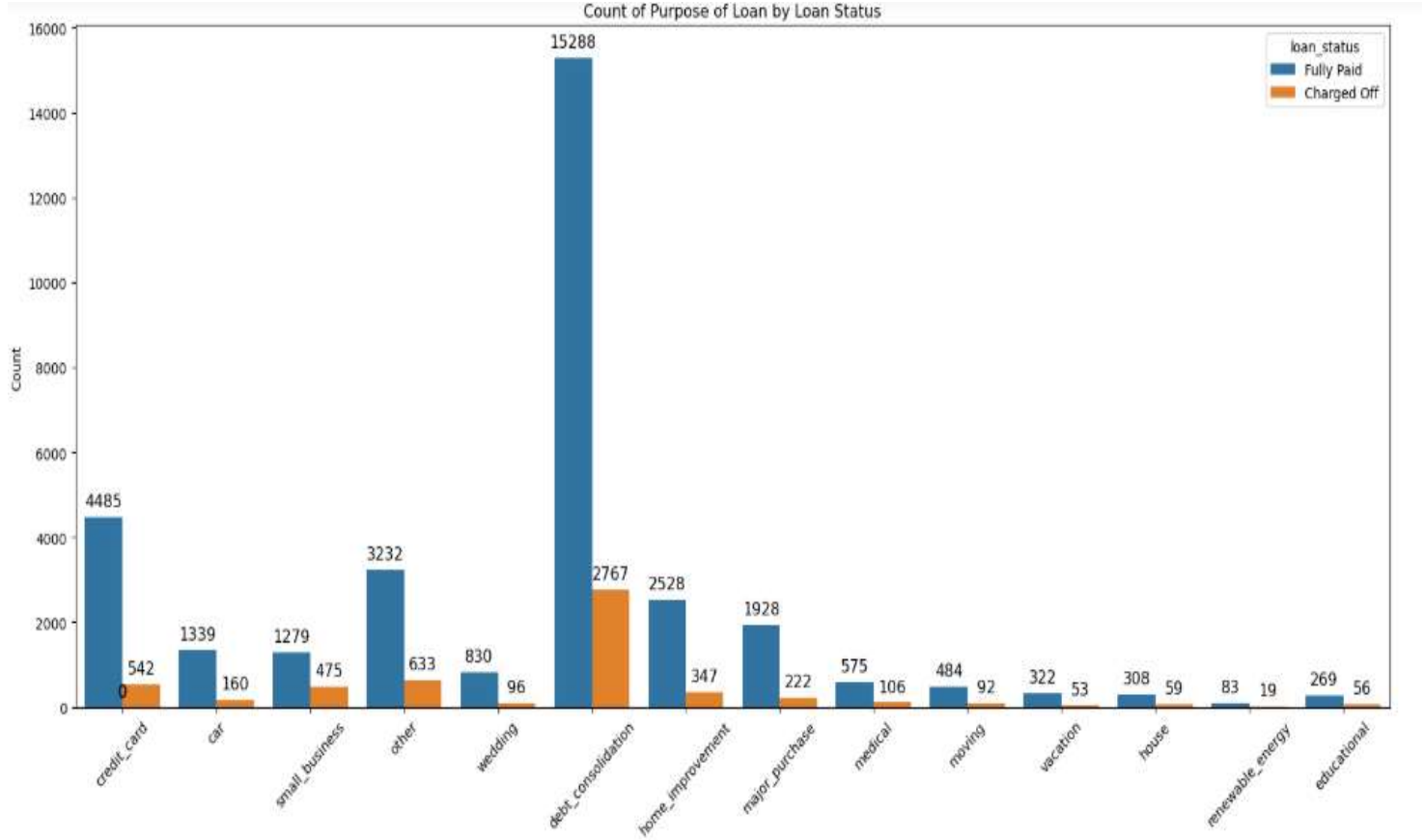
* loan amount Bucket(loan_amnt_bucket)

* Annual income Bucket (annual_amnt_bucket)

* Monthly instalments Bucket(installment_bucket)

Bivariate Analysis (Unordered Categorical)

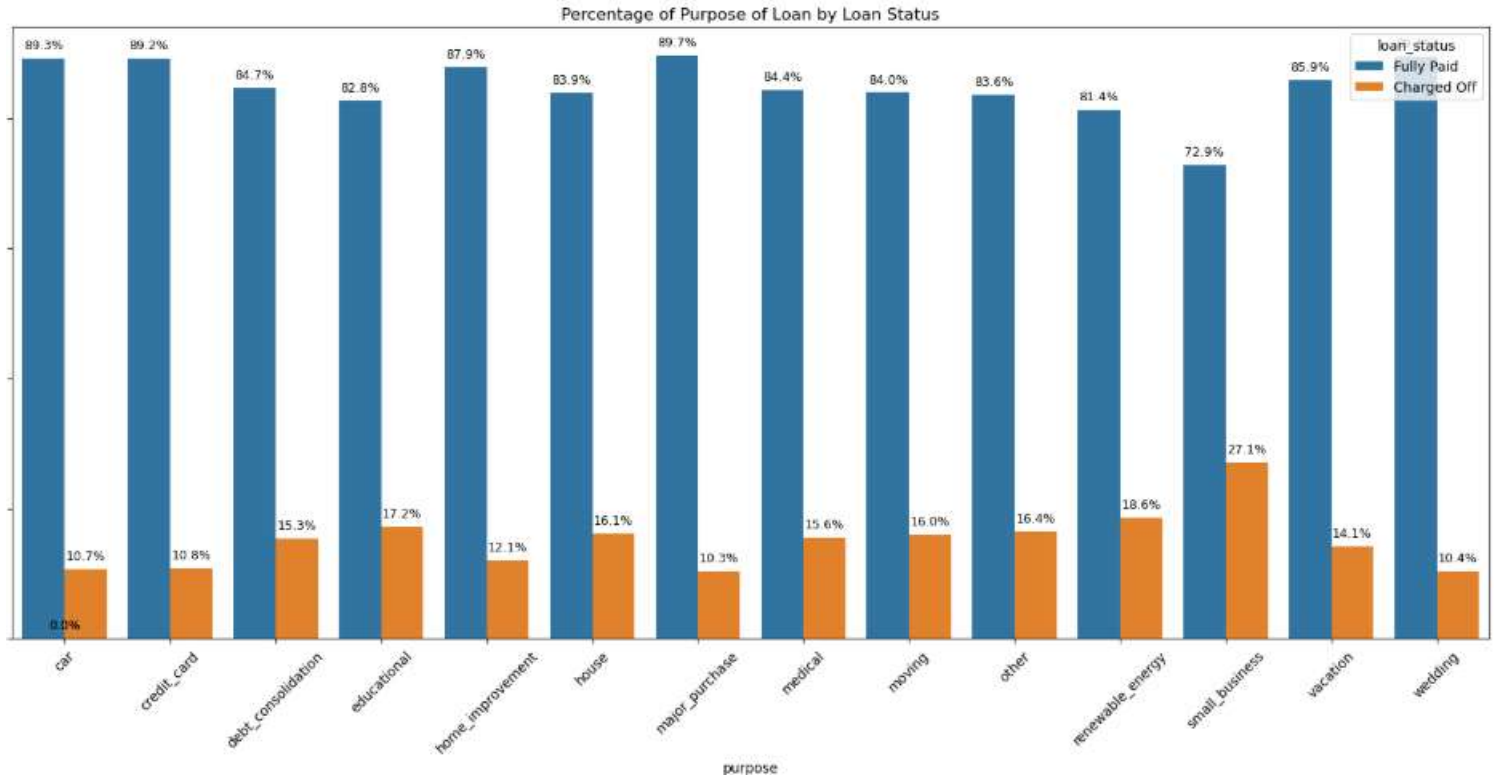
Purpose vs Loan status



Observation – debt_consolidation, credit card payment has high charge off count. Also small business has a higher percentage of charged off loan (Can be seen in below plot)

Bivariate Analysis (Unordered Categorical)

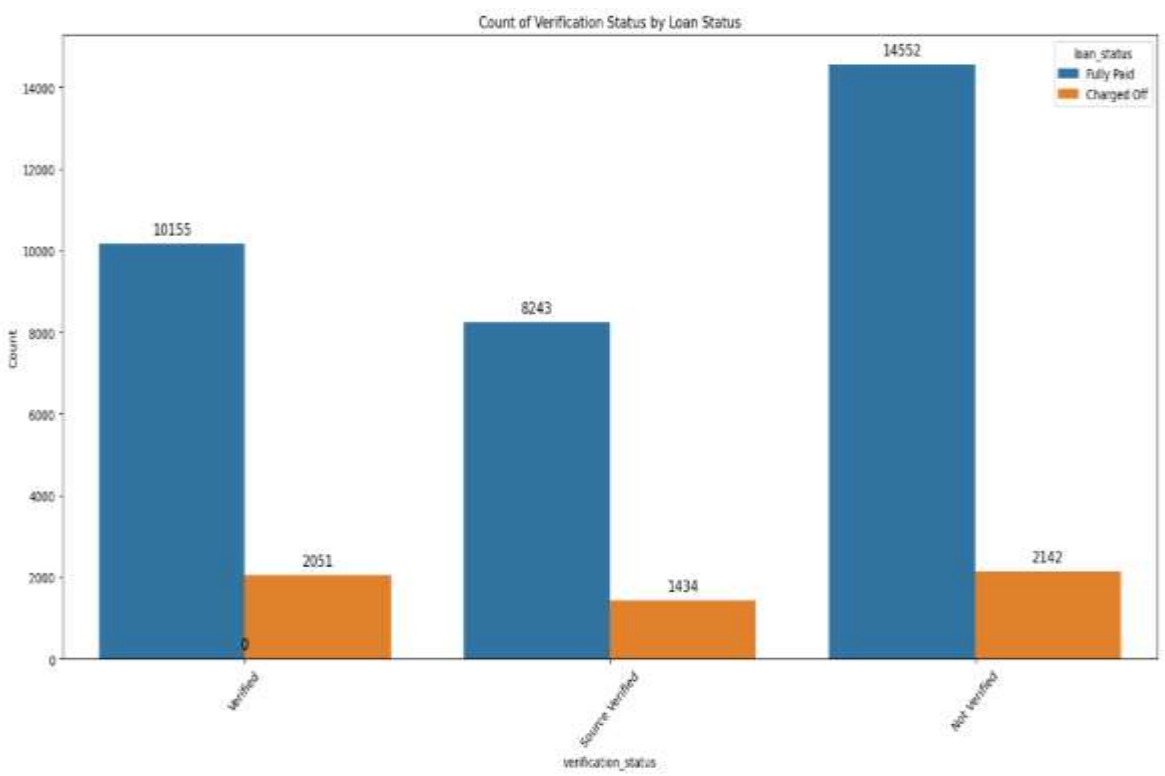
Purpose vs Loan status II

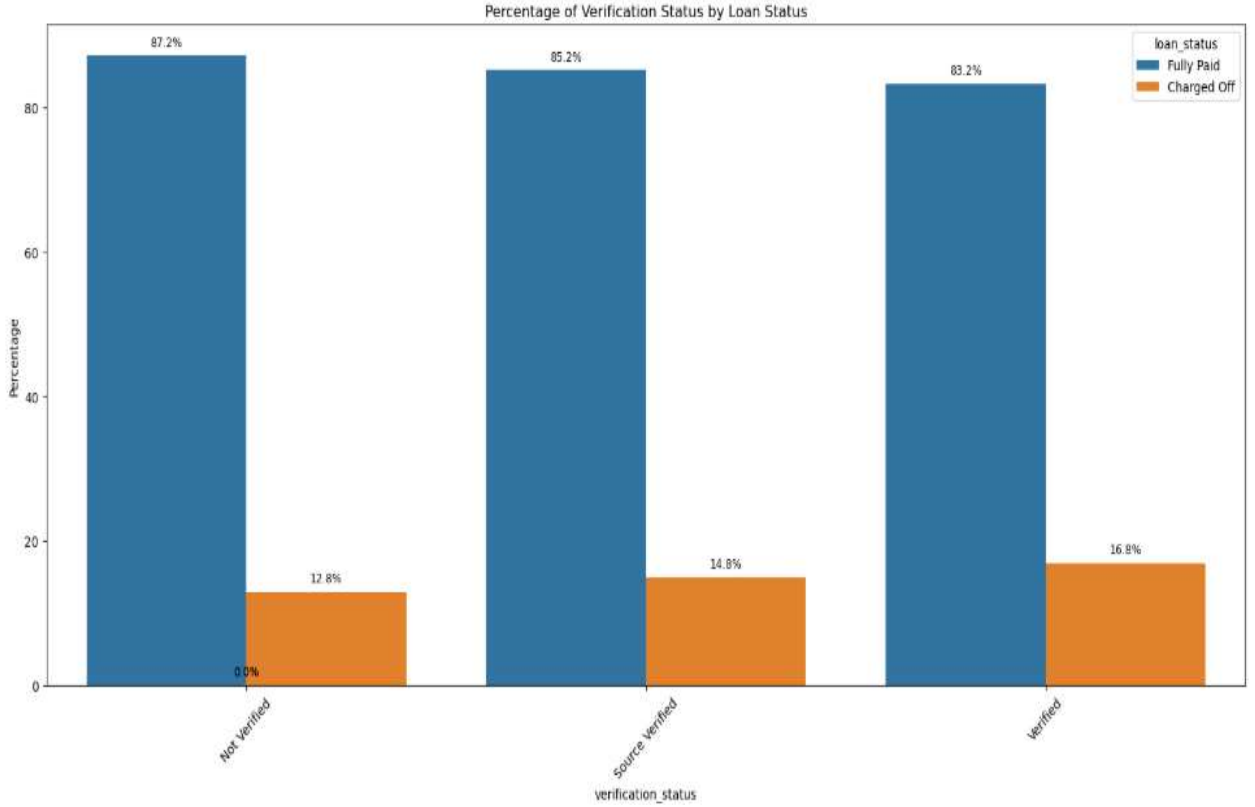


Observation – As mentioned above small business related loan has higher percentage of charged off ratio

Bivariate Analysis (Unordered Categorical)

Verification status vs Loan status

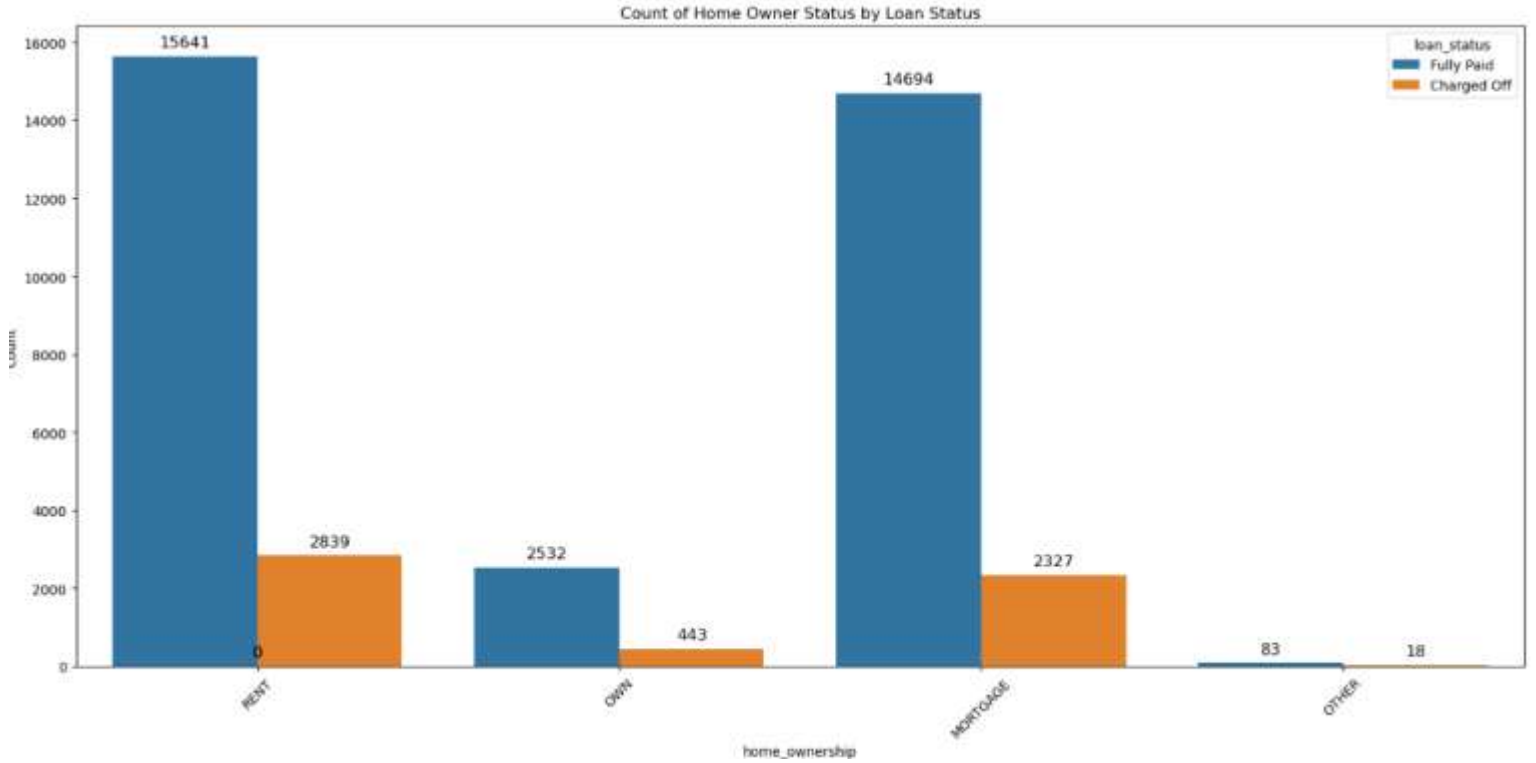




Observation – Verified loan has higher charged off count and percentage. This point to either corruption or incompetency.

Bivariate Analysis (Unordered Categorical)

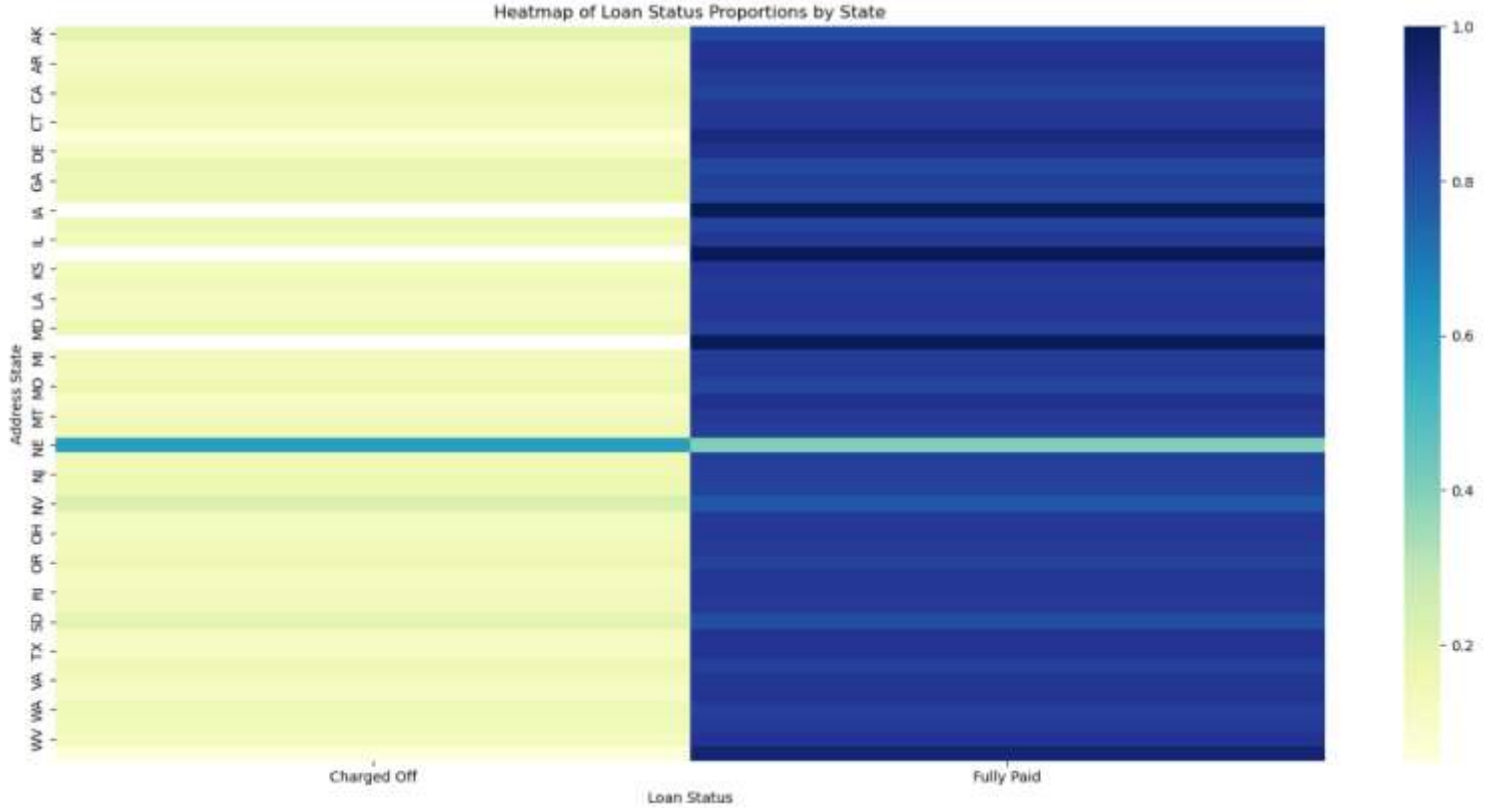
Home ownership vs Loan status



Observation – Verified loan has higher charged off count and percentage. This point to either corruption or incompetency.

Bivariate Analysis (Unordered Categorical)

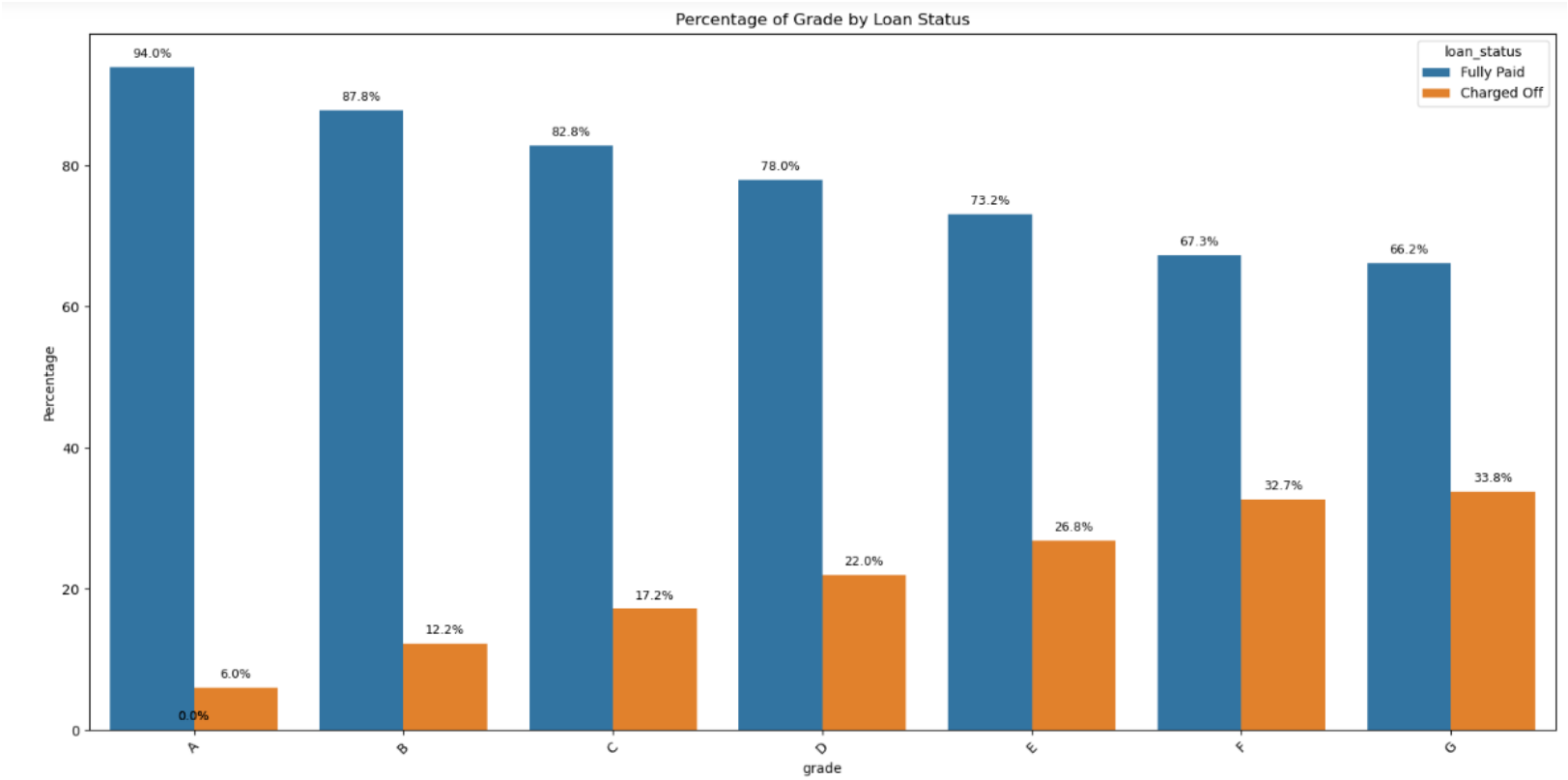
State vs Loan status



Observation - State of applicant and loan being charged off is not related, except for state NE. But it has a very low count, so can be ignored as not much loan data is available for it to better understand the trend.

Bivariate Analysis (Ordered Categorical)

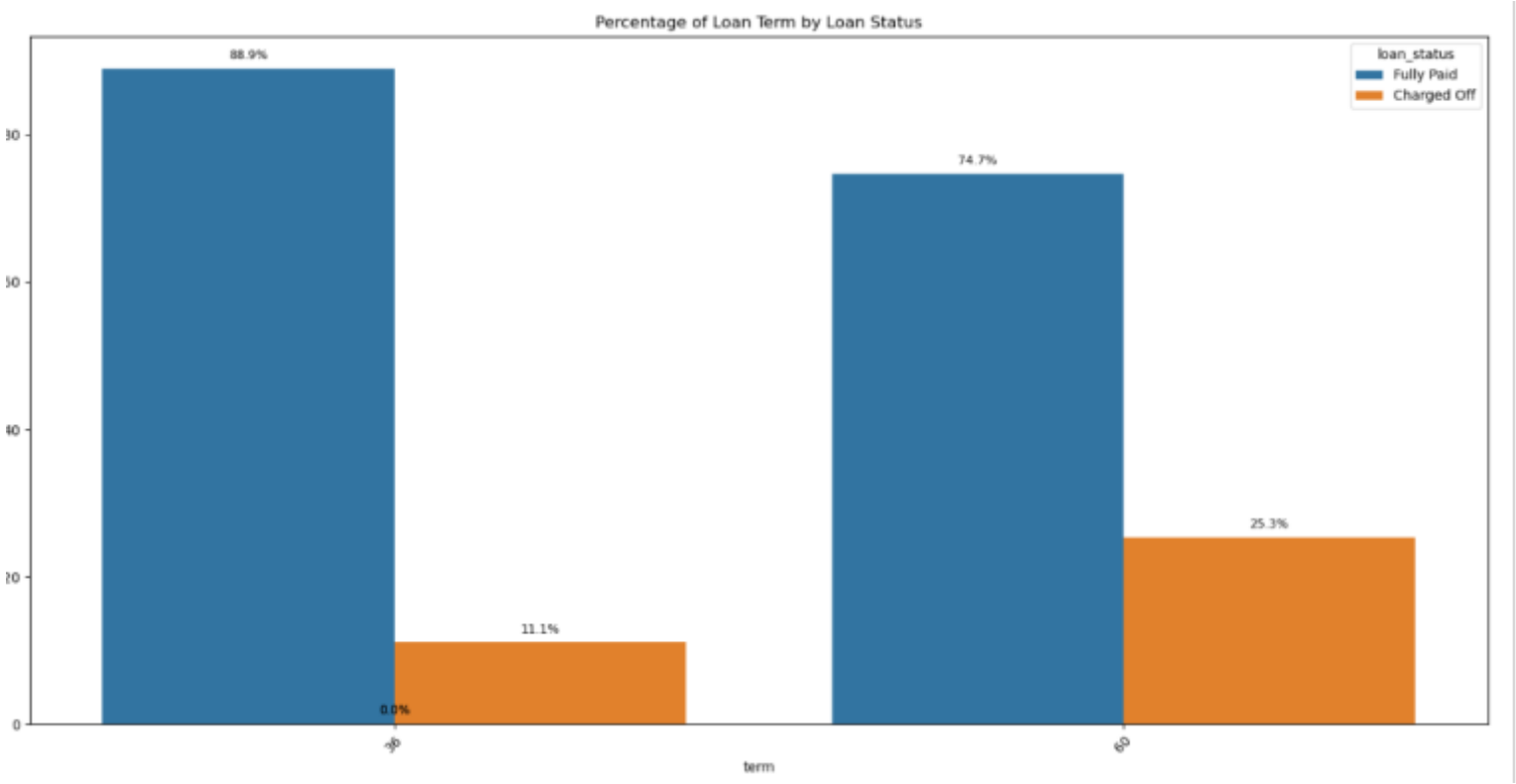
Grade vs Loan status



- Observation – **Charge off possibility increases as grade is decreasing**

Bivariate Analysis (Ordered Categorical)

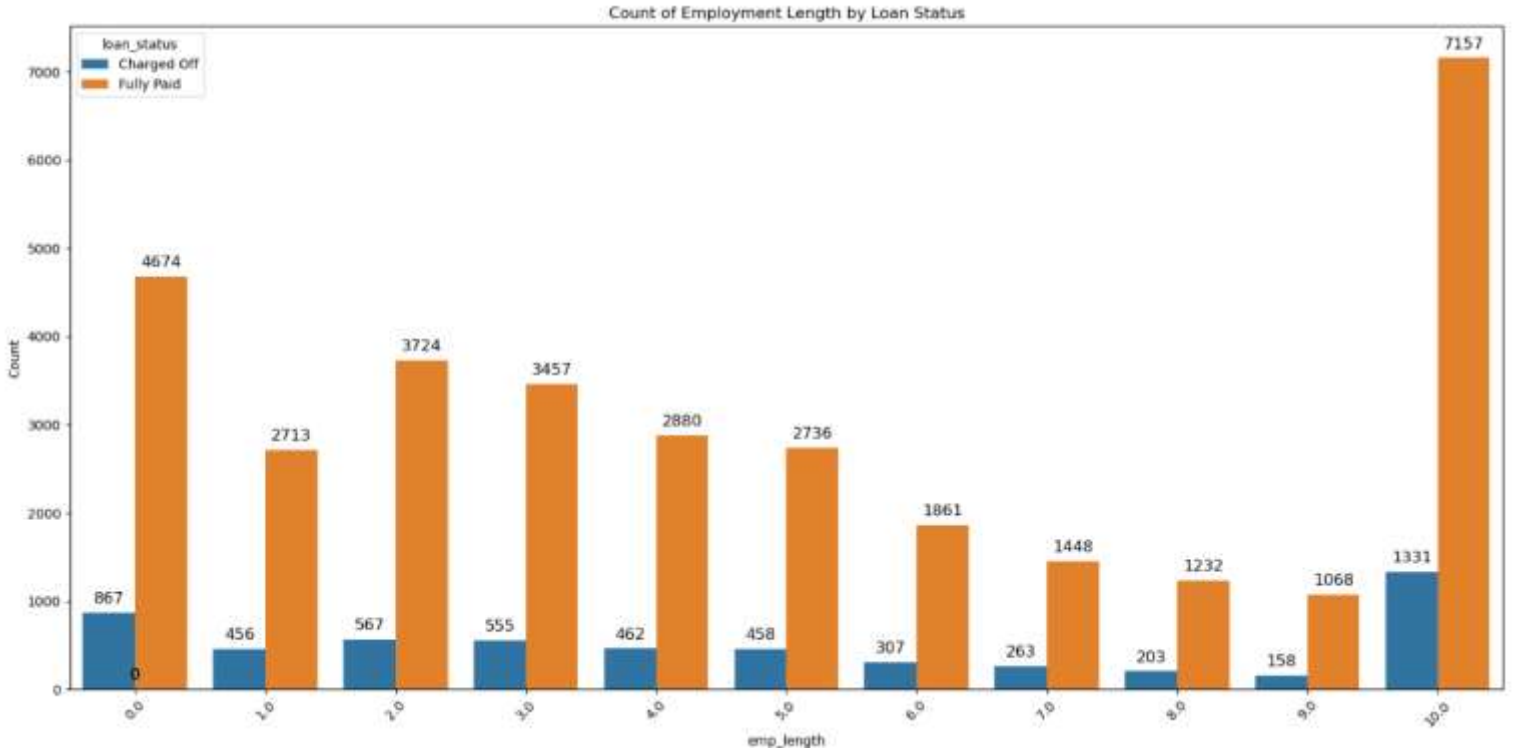
Term vs Loan status



- Observation – Longer tenure results in higher charge off

Bivariate Analysis (Ordered Categorical)

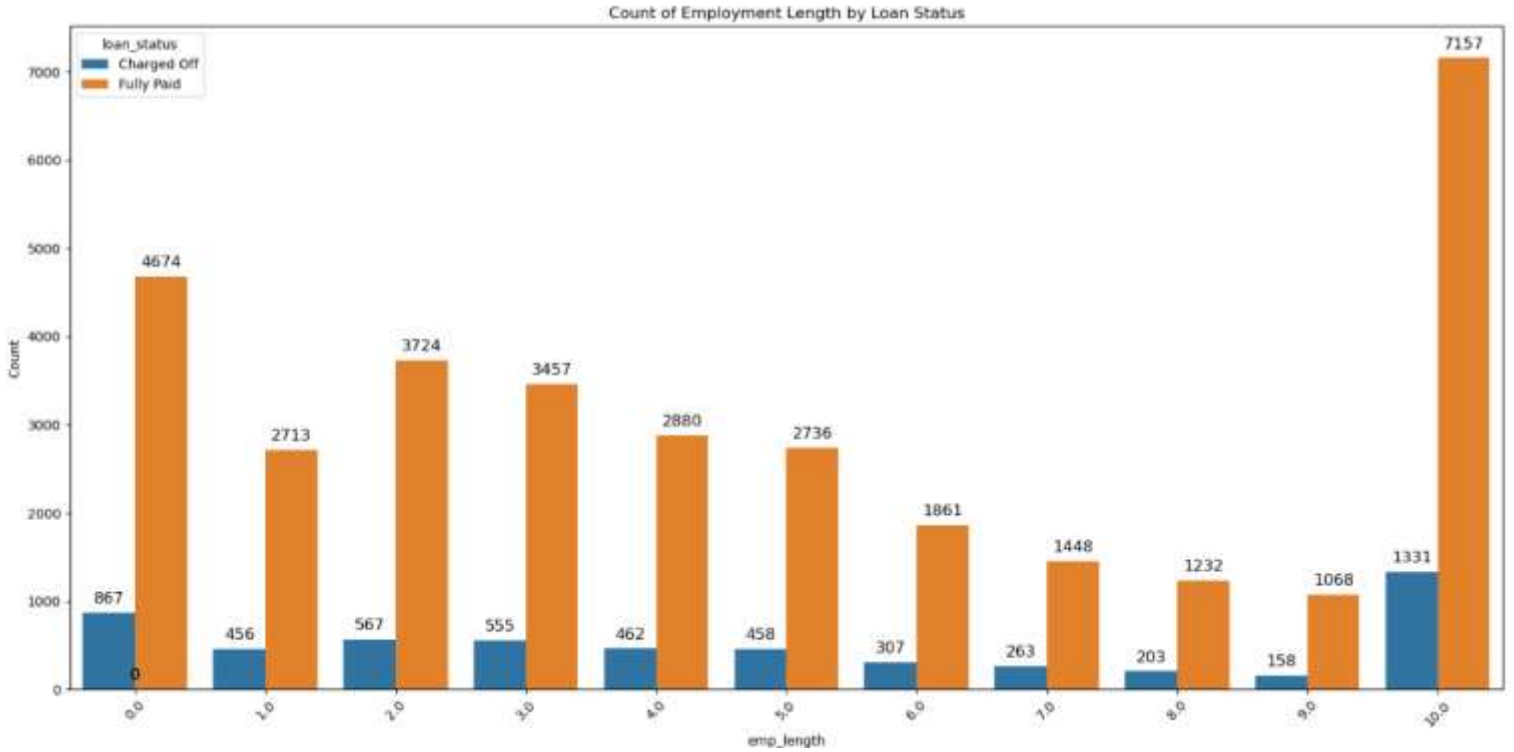
Employee length vs Loan status



- Observation - High employment tenure customer are defaulting higher.

Bivariate Analysis (Quantitative variable)

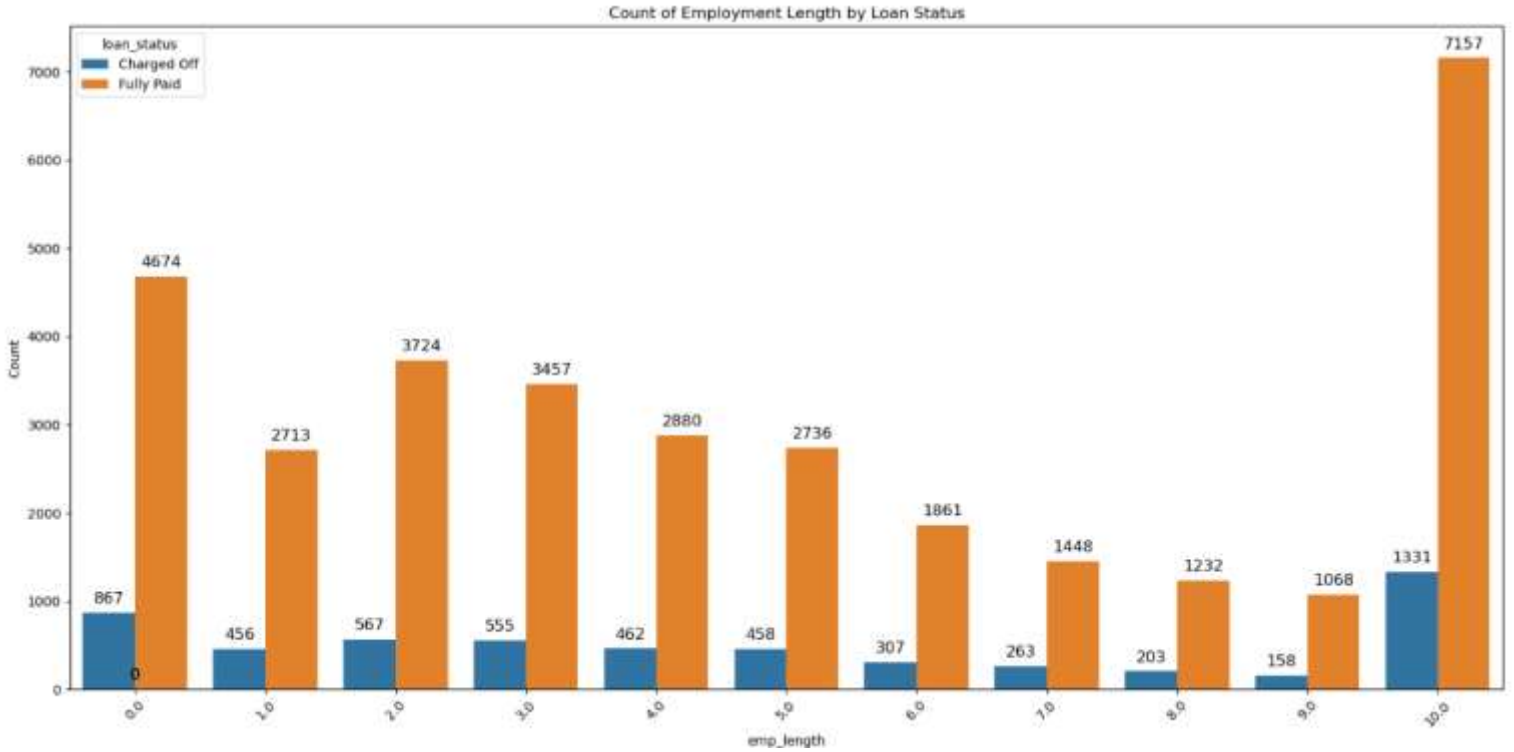
Loan Amount bin vs Loan status



- Observation - High employment tenure customer are defaulting higher.

Bivariate Analysis (Quantitative variable)

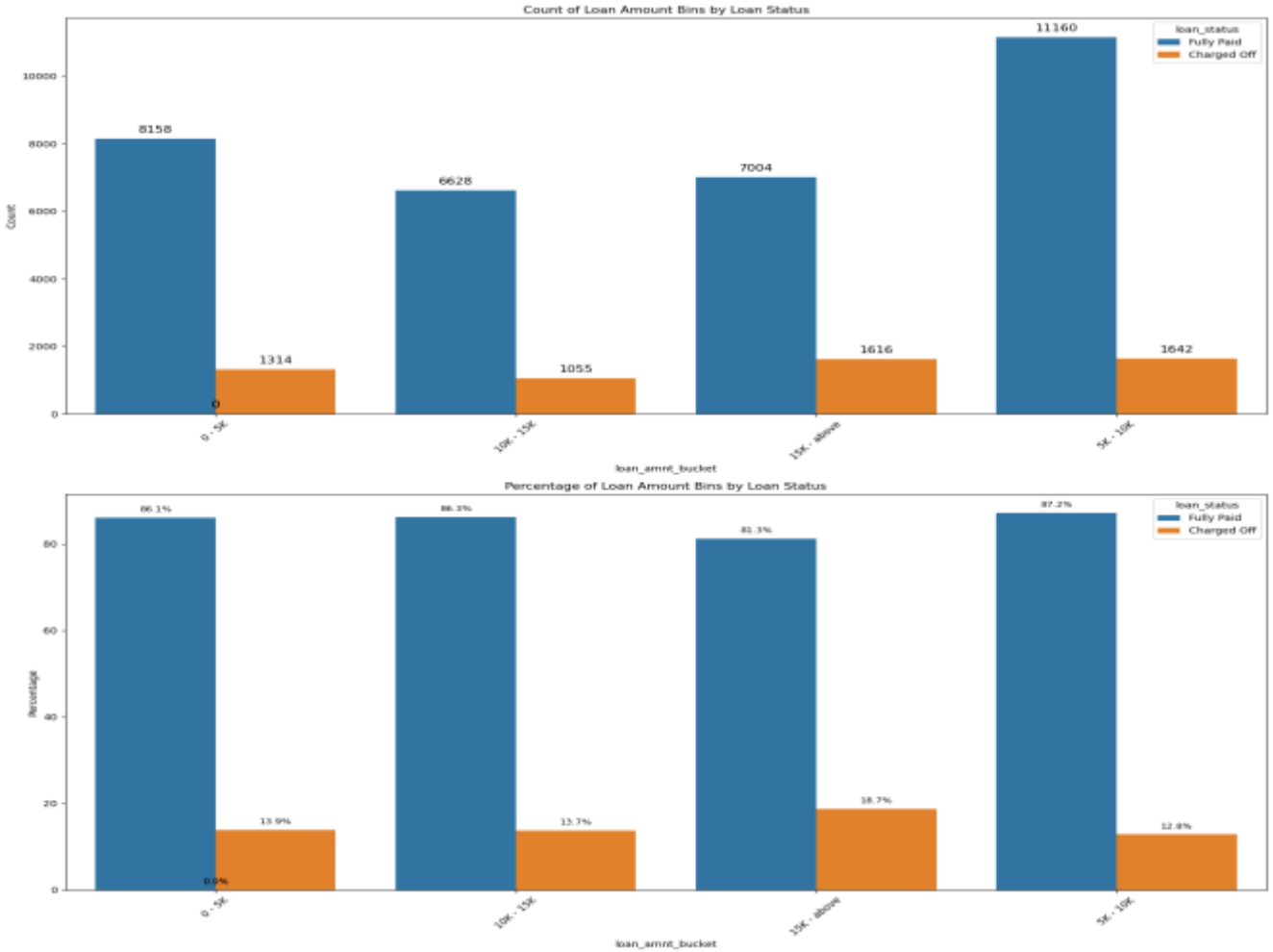
Loan Amount bin vs Loan status

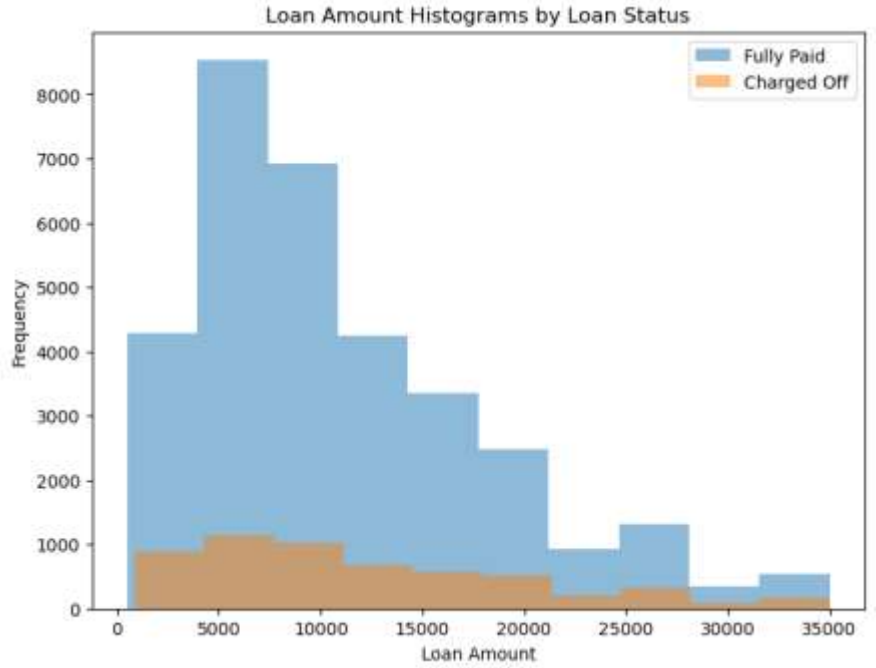


- Observation - High employment tenure customer are defaulting higher.

Bivariate Analysis (Quantitative variable)

Loan Amount bin vs Loan status

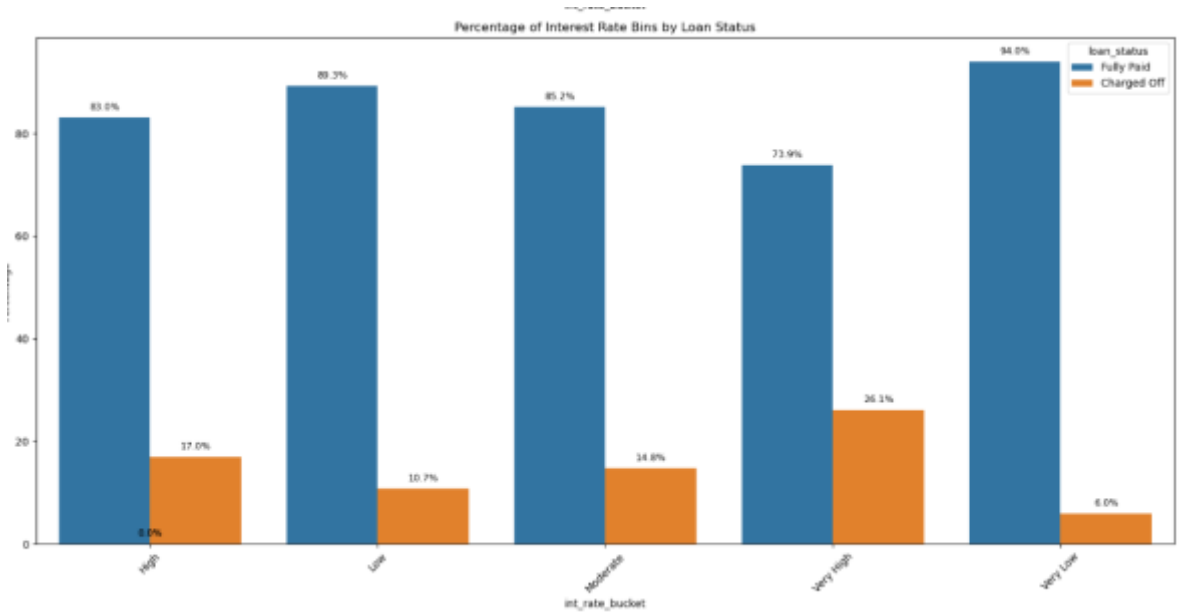


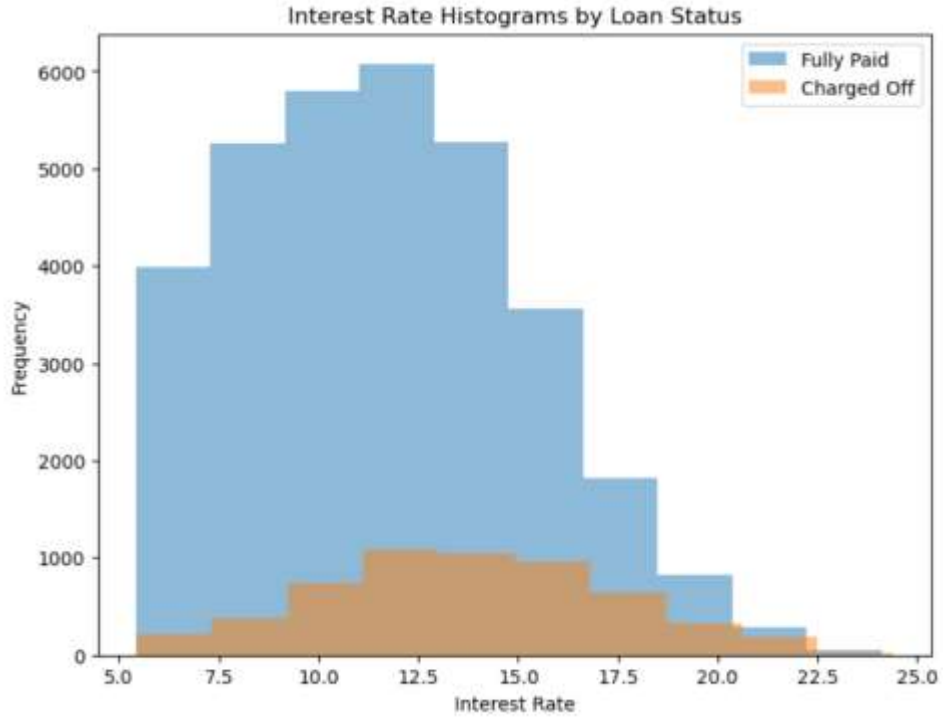


Observation- most of the charge off happening in 5600 to 16500 loan amount. Even in it 5k to 10k is very high. Also for loan amount greater than 25k, seeing higher charge off ratio

Bivariate Analysis (Quantitative variable)

- Interest Rate bucket vs Loan Status

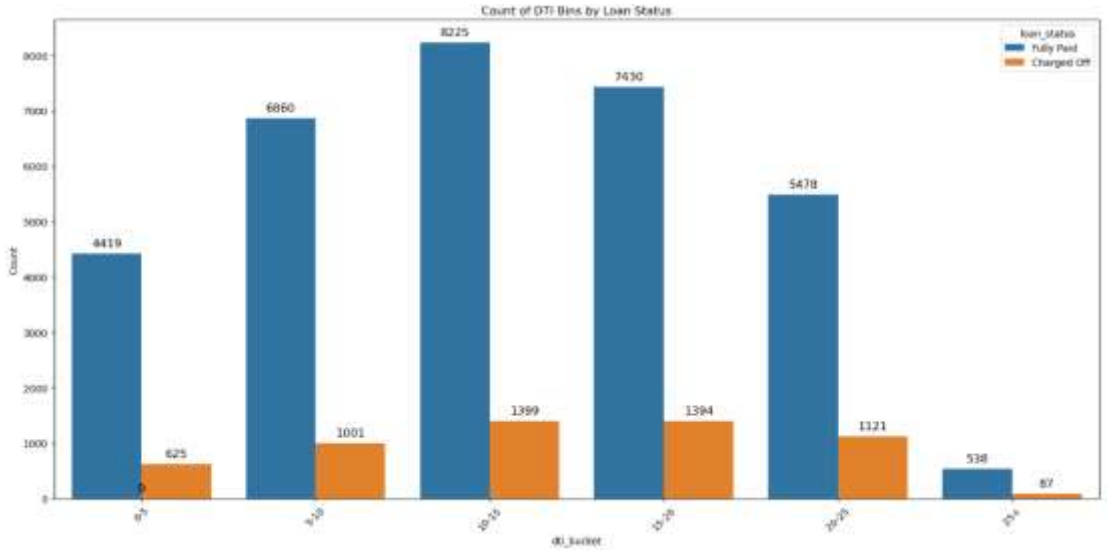


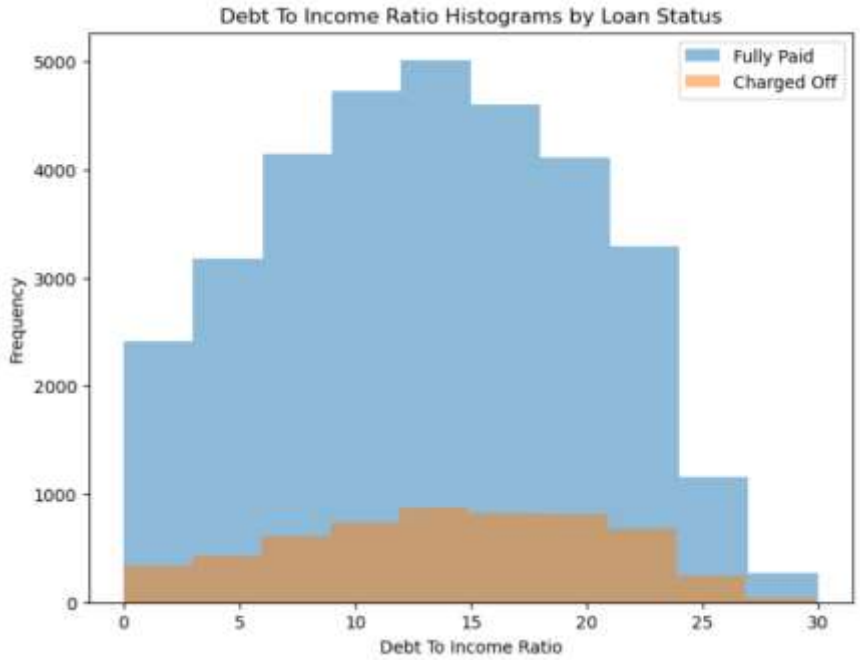


Observation- We can see high charge off for High and very high interest rate.After breaking down using histogram we can see that it is high between 13 to 17 and for interest more than 20%

Bivariate Analysis (Quantitative variable)

- DTI bucket vs Loan Status

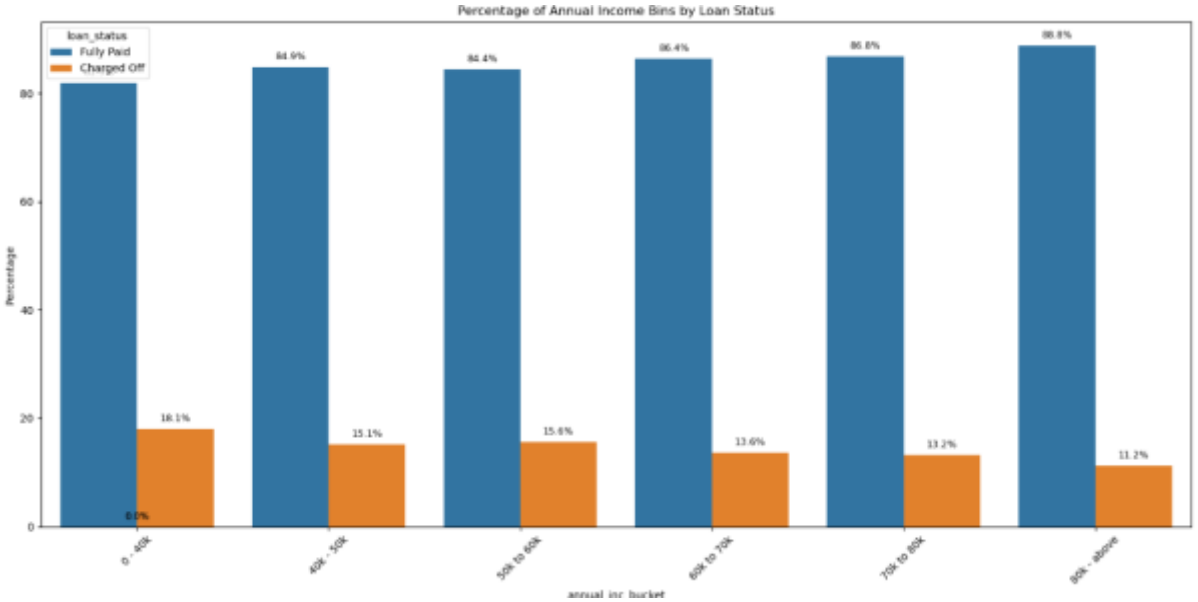


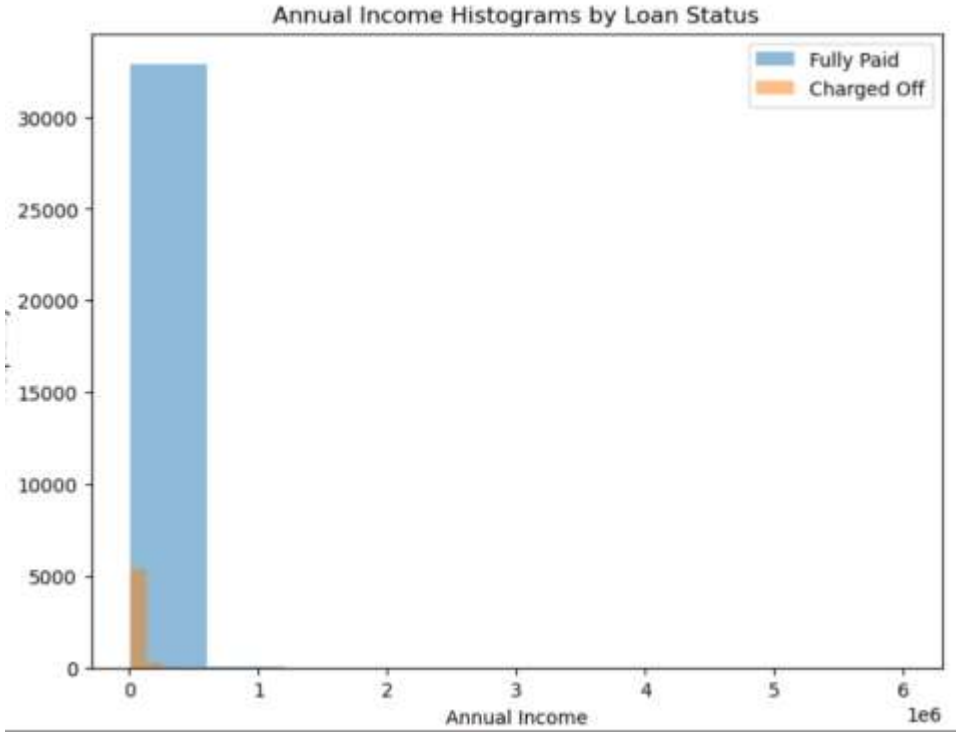


Observation- When DTI is between 10 to 25 we are seeing high charge off of loan. So we can say that higher the DTI value , higher will be the chance of loan being default

Bivariate Analysis (Quantitative variable)

- Annual Income Bucket vs Loan Status

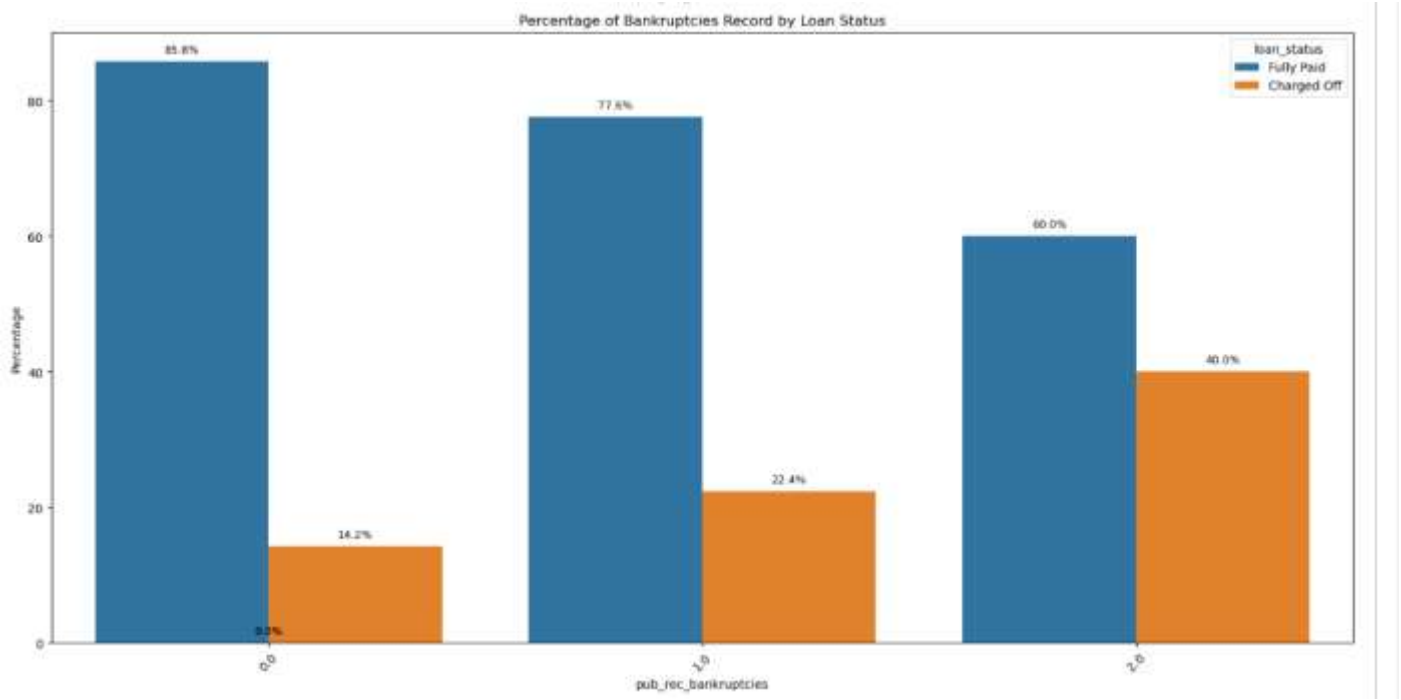




Observation-- Charge off loan are higher when the annual income is low. In Other word, there is a negative correlation

Bivariate Analysis (Quantitative variable)

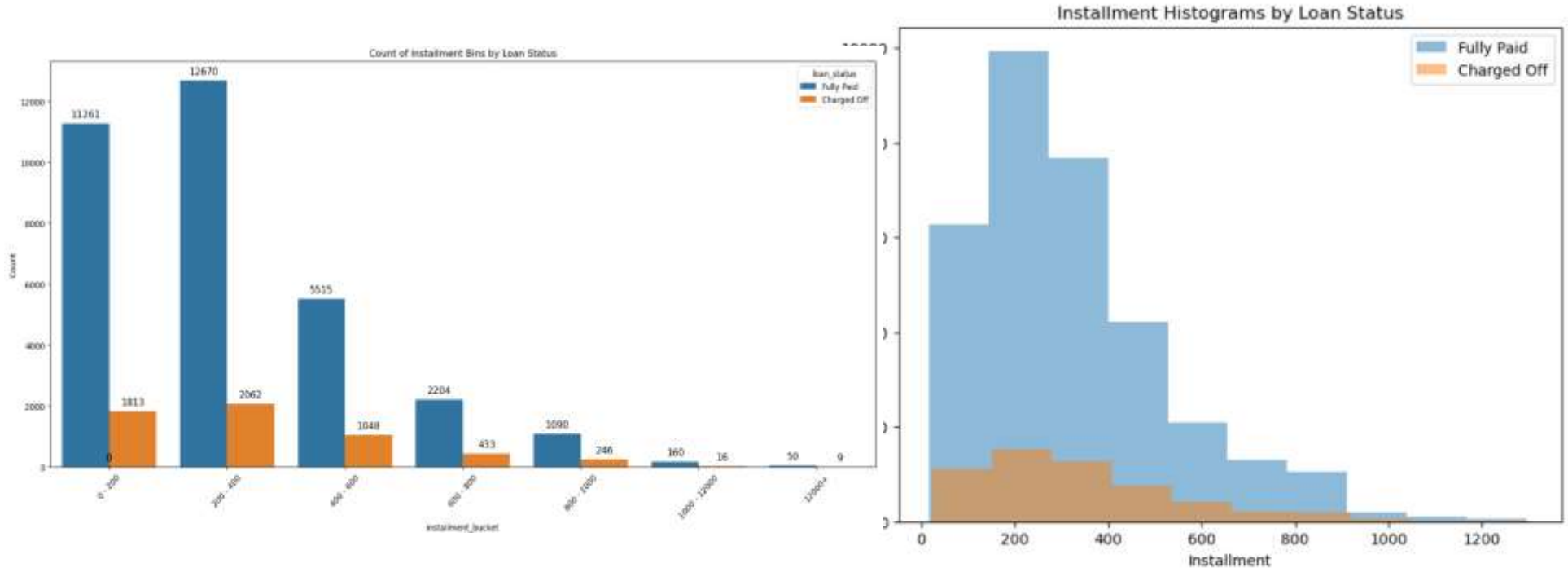
- Public record of bankruptcy vs Loan Status



Observation- Higher change of loan defaulting is there, when we have public record of bankruptcy.

Bivariate Analysis (Quantitative variable)

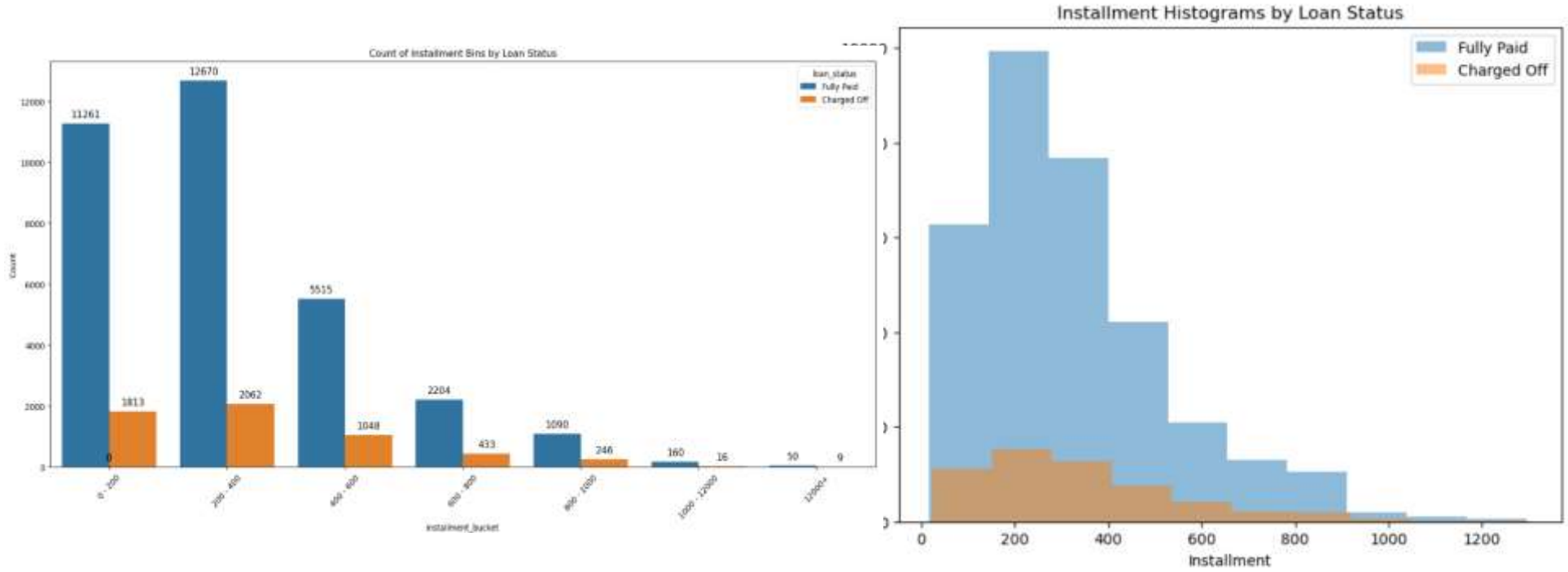
- Installment Bucket vs Loan Status



Observation- We are seeing high number of defaulted loan for instalment amount less than or equal to 1k. Some other variable is at play here. We need to perform multivariant analysis

Bivariate Analysis (Quantitative variable)

- Installment Bucket vs Loan Status



Observation- We are seeing high number of defaulted loan for instalment amount less than or equal to 1k. Some other variable is at play here. We need to perform multivariant analysis

Bivariate Analysis

Inference

- **Debt Consolidation and Credit Card purpose loans have higher charge-off rates:** - This indicate that loan taking to close other financial burden have a high chance of defaulting. We should have better verification of such loans. Also, we should check whether existing financial installment are paid on time, to see the repayment behavior.
- **Small Business loans also have a higher percentage of charge-offs:** This suggests that small business loans are inherently riskier, possibly due to higher uncertainty and lower collateral.
- **Verified loans have a higher charge-off count and percentage:** This could point to potential issues with the verification process or might indicate that borrowers with verified income might have different risk profiles. We should check for incompetency or corruption in the verification process.
- **Charge-off possibility increases as loan grade decreases:** This is expected, as lower grades correspond to higher risk borrowers.
- **Highest number of defaults occurs with applicants having 10 years of employment experience:** This might seem counterintuitive, but further investigation is needed to understand the underlying reasons.

Bivariate Analysis

Inference

- **Longer loan terms result in a higher possibility of charge-off:** This is expected as longer loan terms increase the overall risk and exposure for lenders.We should try to either reduce the tenure or increase interest rate to have better recovery loan sanctioned
- **Loan Amount and Charge-Offs:** A significant portion of charge-offs occurs for loan amounts between $5,600 and $16,500, with a particularly high concentration between $5,000 and $10,000. This suggests that loans within this range might represent a higher risk segment.
- **Interaction of Interest Rate and Loan Amount**: Loans with very high interest rates and large loan amounts exhibit a higher propensity for charge-off. This highlights the combined risk associated with high borrowing costs and substantial debt.
- **Interest Rate and Charge-Offs:** High charge-off rates are observed for loans with high and very high interest rates. Further analysis using a histogram reveals a particularly high concentration of charge-offs for interest rates between 13% and 17%, and also for rates exceeding 20%. This identifies specific interest rate ranges that warrant closer scrutiny.
- **Debt-to-Income Ratio (DTI) and Charge-Offs:** Loans with a DTI between 10 and 25 show elevated charge-off rates. This supports the general trend that higher DTI values are associated with an increased likelihood of loan default.

Bivariate Analysis

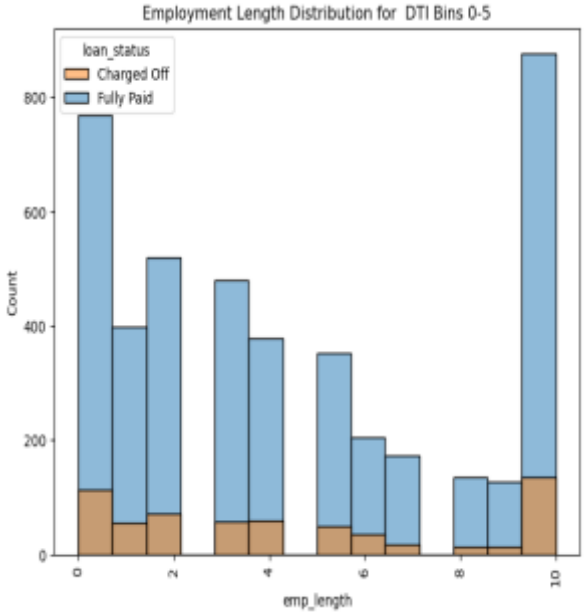
Inference

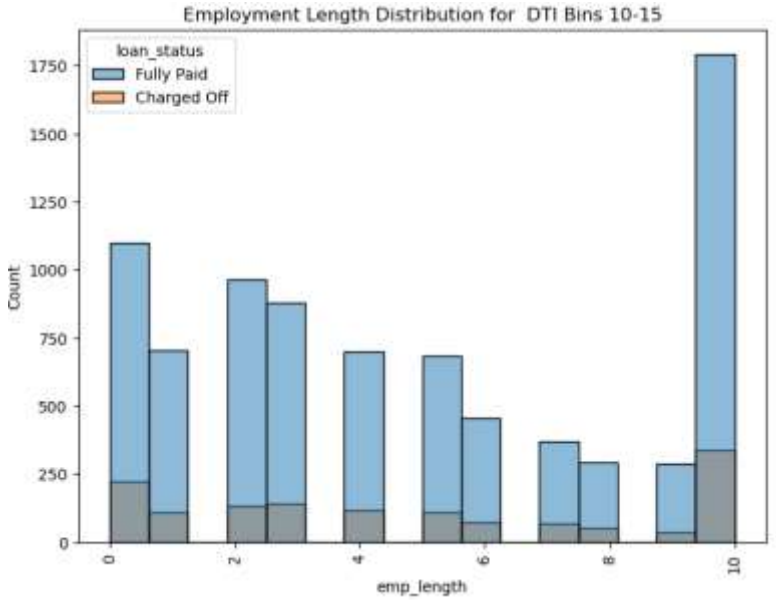
- **Annual Income and Charge-Offs:** Charge-off rates are higher for borrowers with lower annual incomes, indicating a negative correlation between income and loan repayment.
- **Installment Amount and Defaulted Loans:** A high number of defaulted loans are observed for installment amounts less than or equal to $1,000. This seemingly counterintuitive finding suggests that other variables are likely influencing default in this segment and necessitates multivariate analysis.
- **Bankruptcy Count and Default:** A higher count of bankruptcies in a borrower's history is associated with a higher chance of loan default. This is consistent with expectations, as prior bankruptcies indicate a higher risk profile.

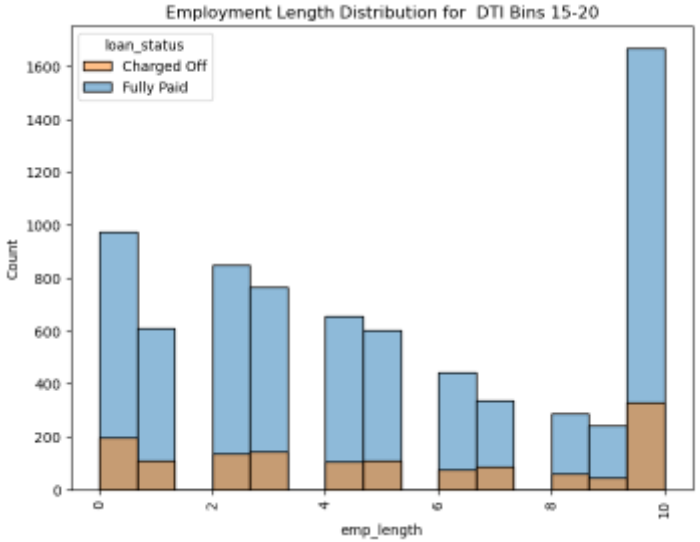
Multi Variant Analysis

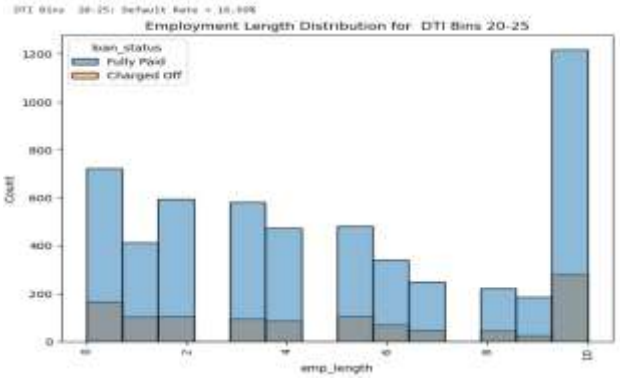
- Loan Amount vs Interest Rate vs Loan Status

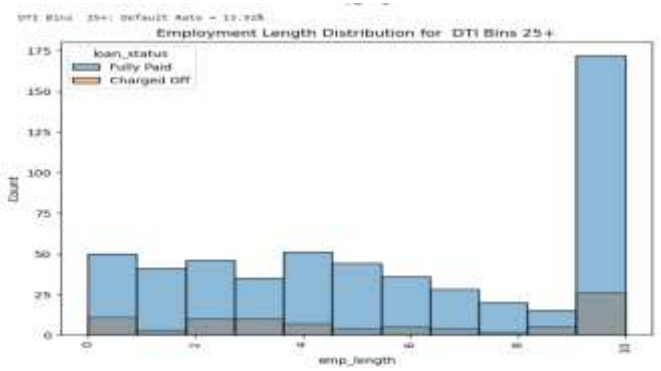
Multi Variant Analysis(Segmented)

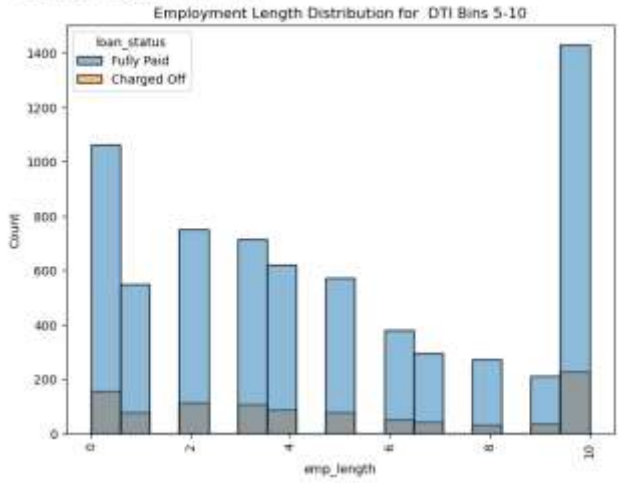






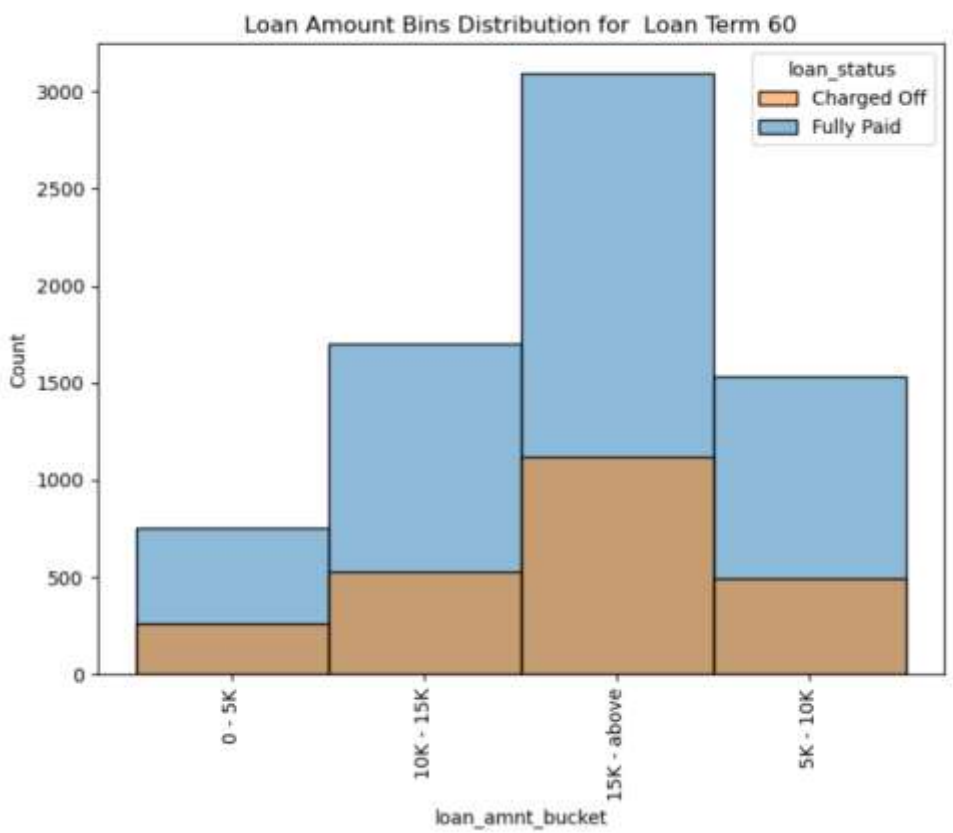




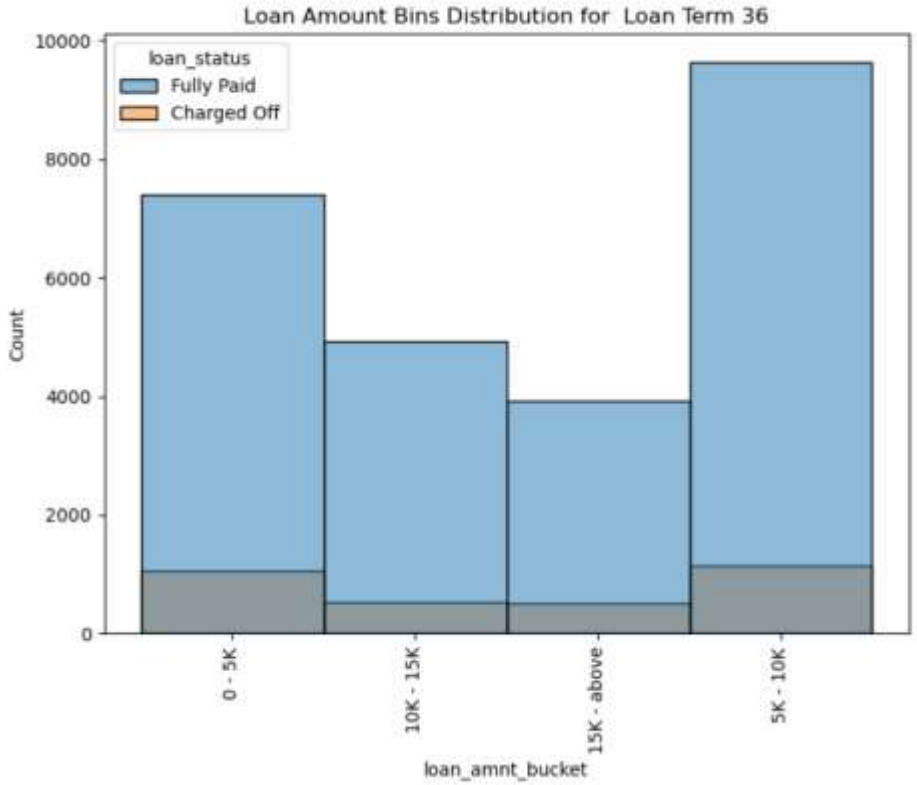


Multi Variant Analysis(Segmented)

Loan Term 60: Default Rate = 25.31%

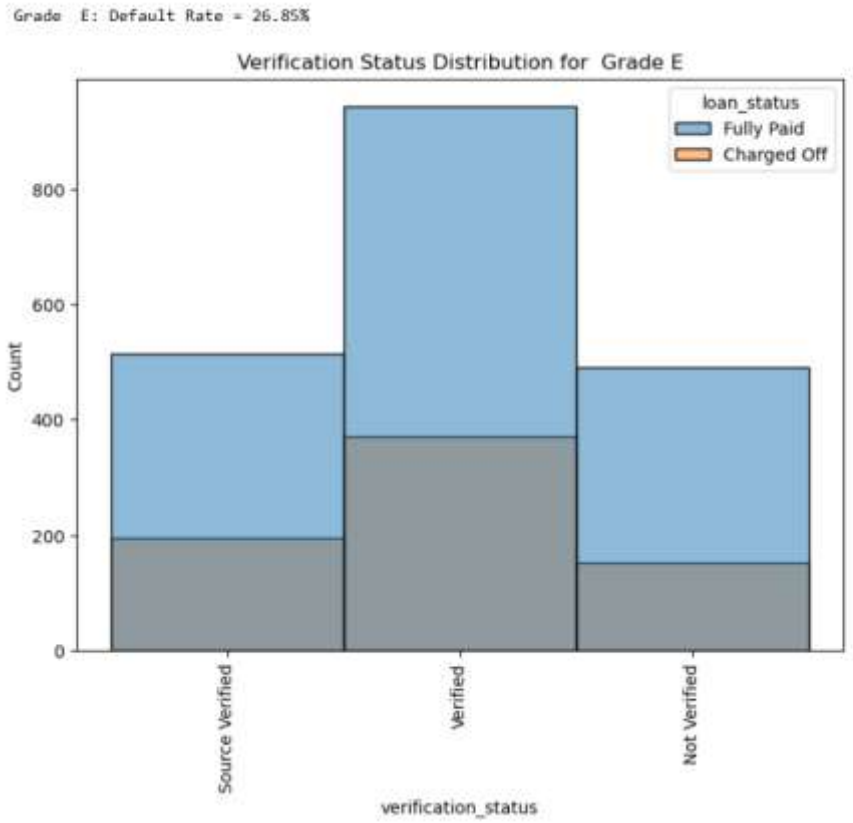


Loan Term 36: Default Rate = 11.09%



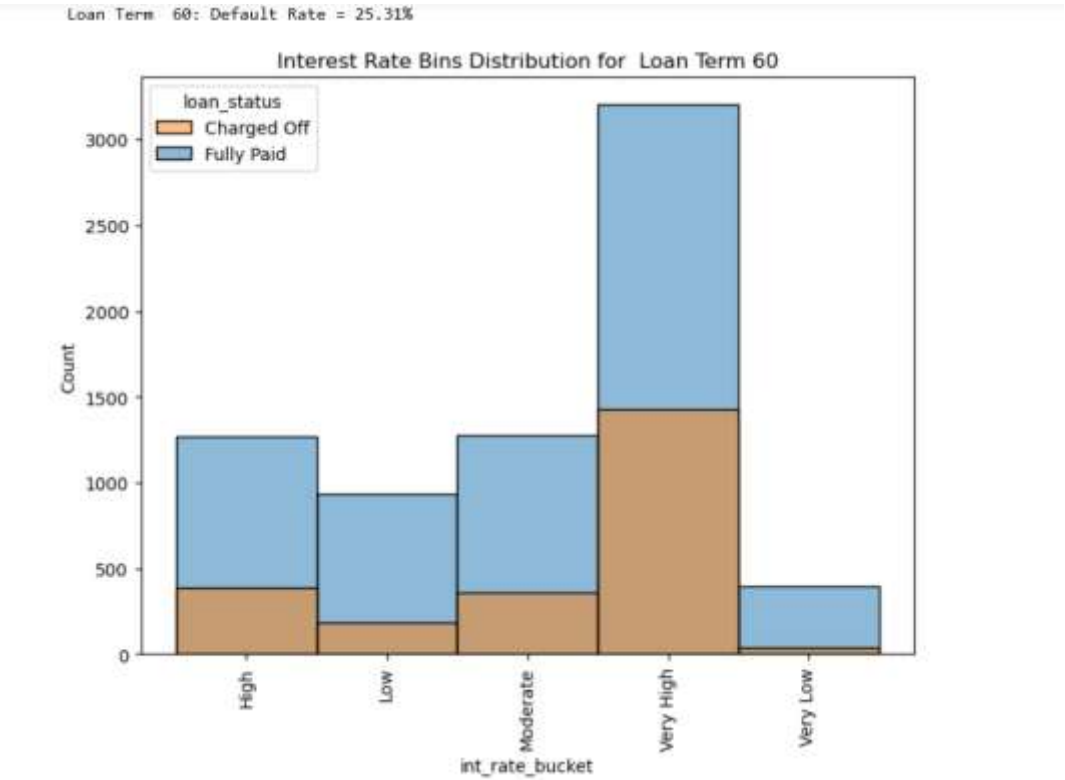
Observation - High term will with high loan amount will more likely lead to default of loan

Multi Variant Analysis(Segmented)



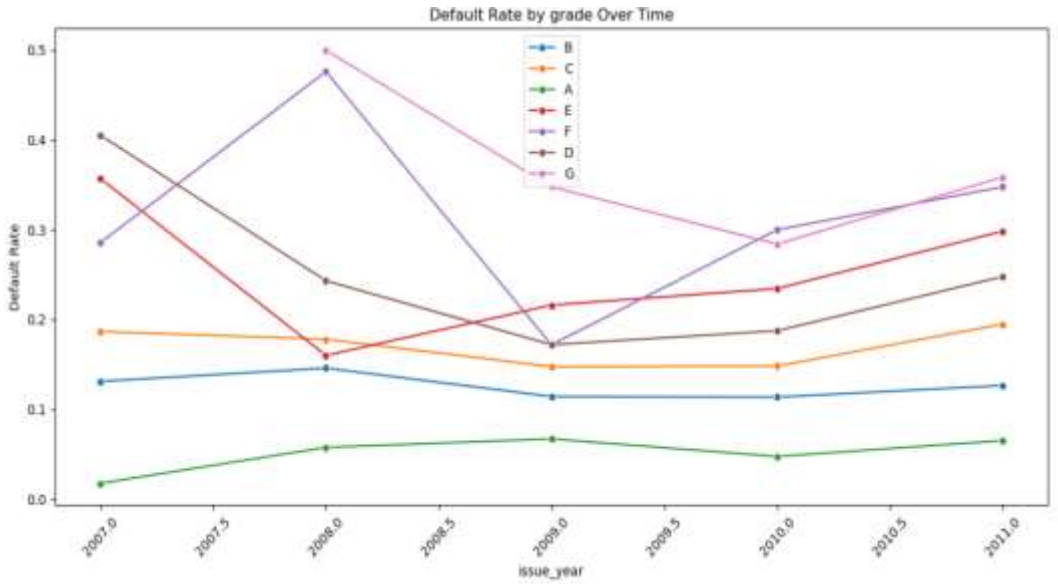
Observation - Poorly graded loan have high verification number. Seems issue in the process of verification

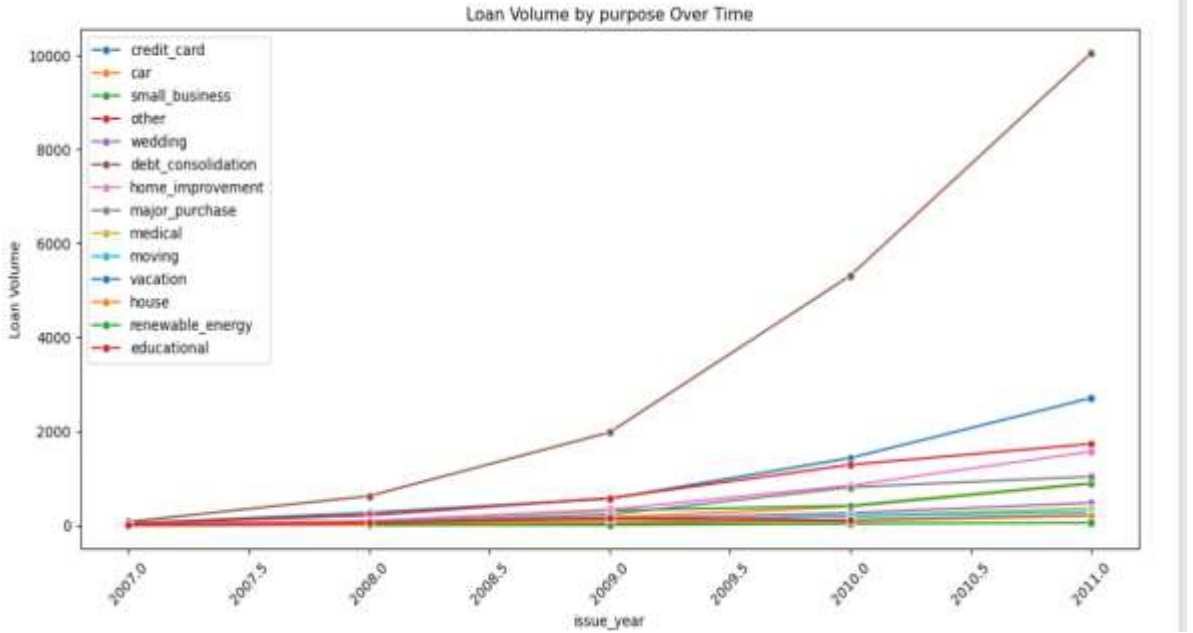
Multi Variant Analysis(Segmented)

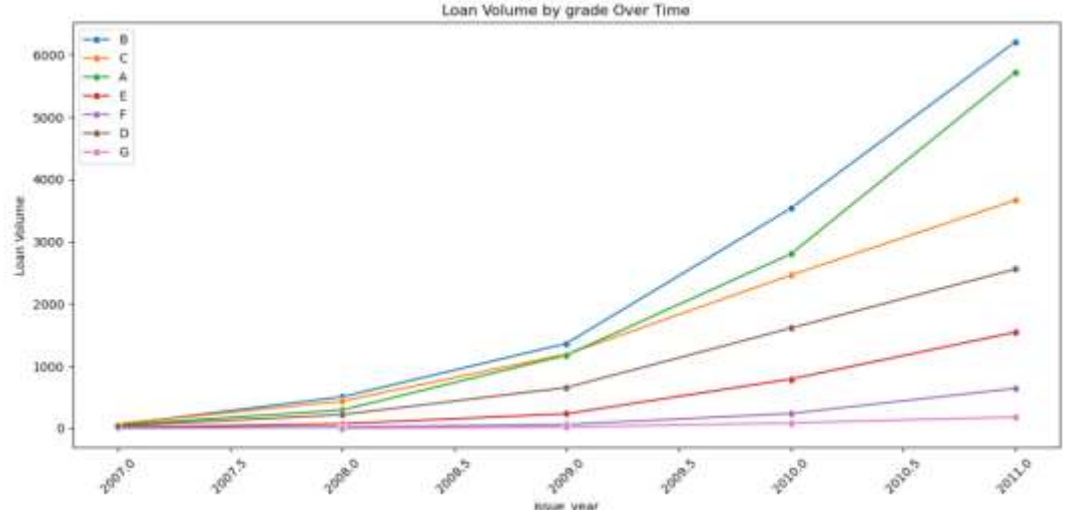


Observation - High term with high interest rate will result in charge off

Multi Variant Analysis(Line plots)



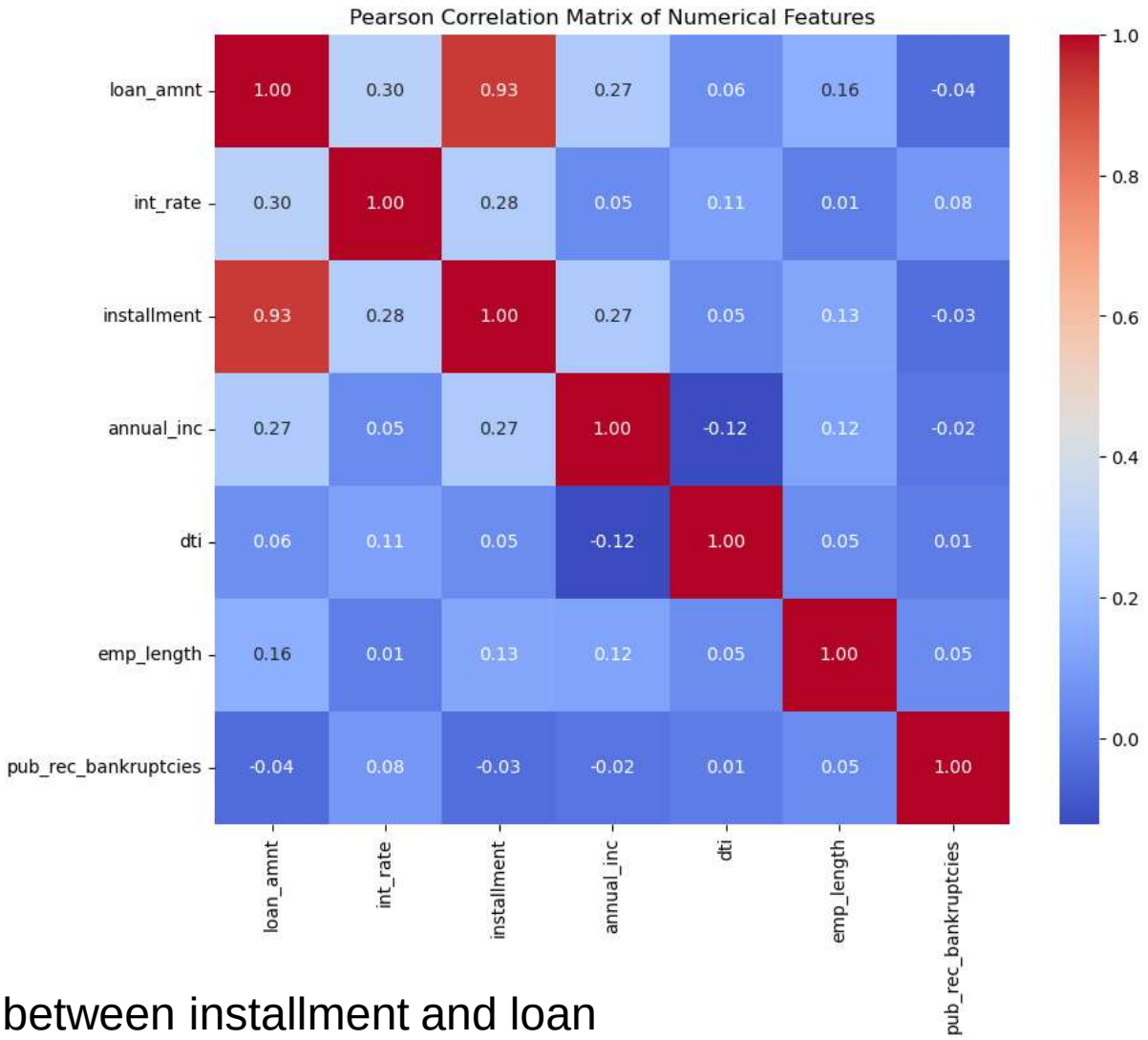




Key Observations for line plot:

- Higher Grades Have Lower Default Rates: This is the most consistent and expected trend. Grades A and B consistently have the lo followed by C, and then the lower grades (D, E, F, G) have progressively higher default rates. This reflects the risk assessment pro better creditworthiness (higher grades) are less likely to default.
- Impact of the 2008 Financial Crisis: There's a noticeable spike in default rates across all grades around 2008-2009. This correspon financial crisis, which significantly impacted loan performance across the board. The crisis likely led to job losses, economic hardsh difficulty for borrowers to repay their loans.
- Recovery After the Crisis: After the peak in 2008-2009, default rates generally declined, although some grades (especially the lowe gradual recovery.
- Volatility in Lower Grades: The lower grades (D, E, F, G) exhibit more volatility in their default rates compared to the higher grades. these borrowers are more sensitive to economic fluctuations.
- Grade F and G: Grades F and G have very high default rates, especially around the financial crisis. This reinforces the idea that th borrowers.

Multi Variant Analysis(Correlation***)***



High correlation between installment and loan amount.

Negative correlation between dti and annual income

Final Inference and Suggestions

Inference-I

- **Verification Status**: Verified loan are showing higher charged off rate.
- **High-Risk Loan Purposes**: Debt consolidation and credit card loans exhibit a higher percentage of defaults, indicating that borrowers seeking these loans might be in a financially vulnerable position.
- **Loan Amount and Interest Rate:** Most defaults occur for loan amounts between $5,600 and $16,500, particularly in the $5,000 to $10,000 range. These loan amounts often carry high to very high-interest rates, increasing the likelihood of default.
- **High Interest Rate**: High interest rate generally have high defaulting rates.
- **Grade and Homeownership**: Loan grade is negatively correlated with defaulting, implying that lower-grade loans are riskier. Additionally, loan applicants who do not own a house have a higher defaulting rate.

1) Low grade loan had a much higher rate of defaulting during 2008 financial crisis. So, keep watch during when market is in instability

Final Inference and Suggestions

Inference - II

- **Employment Tenure**: Unexpectedly, customers with longer employment tenures have a higher defaulting rate. This further investigated using pair plot and details mentioned in below point.

  1) Loan Amount and Employment Tenure Interaction: As per pair plot, customer with higher amount of loan.Higher amount is linked to higher interest. All these factor result in higher defaulting

  2) Confounding Factors for employee tenure: It's possible that other confounding factors are at play, such as:

  3) Age: Longer-tenured employees are likely older and might have different financial priorities or life events that impact their ability to repay (e.g., children's education, medical expenses).

  4) Complacency: Lenders might be more lenient with underwriting standards for longer-tenured employees, assuming lower risk, which could lead to approving some riskier loans.

- **Bankruptcies and DTI**: A higher count of bankruptcies increases the chance of defaulting. High DTI is also a cause of defaulting, and it is negatively correlated with annual income, suggesting that high DTI often implies lower annual income.
- **Installment and Loan Term**: Installment and loan amount have a high positive correlation. Large-term loans with high-interest rates increase the likelihood of defaulting.

Final Inference and Suggestions

Suggestion - I

- Lending club need to re-examine its application verification process. A large portion of verified application are getting defaulted, it is either incompetency at the verification team end or some sort of corruption is involved.
- If loan amount is less, then we should try to cap the loan term to 36 month.
- If loan amount is high, we should have a combination higher tenure with medium interest rate. In long term, it will help in recovery of larger amount for investor.
- Lenders need to be prepared for increased defaults during economic downturns. Lender could choose to lend less to lower-grade borrowers, especially during uncertain economic times, or charge significantly higher interest rates to compensate for the increased risk.
- Implement stricter criteria for debt consolidation and credit card refinancing loans, including lower loan amounts, higher credit score requirements, and stricter DTI limits.
- Verify the payoff of existing debts for debt consolidation loans to ensure responsible use of funds.

Final Inference and Suggestions

Suggestion - I

- Implement risk-based pricing, adjusting interest rates based on borrower risk profiles.
- Continuously monitor the relationship between interest rates and defaults to optimize pricing strategies.
- Perform thorough background checks for bankruptcies and consider stricter criteria for applicants with a history of bankruptcies.
- Enforce stricter DTI limits, especially for higher loan amounts or longer loan terms.
- Encourage shorter loan terms, especially for higher-risk borrowers, to reduce overall interest paid and the risk of default.
- For customer with long employment tenure, following factor need to taken into consideration

  - we need to check DTI value (to see whether existing financial burden is there).

  - Make sure that proper verification and risk analysis is done and no Complacency is done due to tenure

  - Even for longer-tenured employees, apply stricter underwriting standards for larger loan amounts. Don't assume that employment tenure alone guarantees low risk.

  - Use a combination capped loan amount (based on DTI, income and other factor) with high interest- short tenure, or low/medium interest with high tenure.

Links

Git link - GitHub - vinay1793/Lending_Club_Case_Study