Lending Club – Risk Analysis Case Study

Context

A consumer finance company specializing in lending different types of loans to urban customers has historical data about the loans disbursed. This historical data in CSV format along with a data dictionary describing the definitions of columns is provided.

Objective

The objective is to perform risk analysis to identify patterns which indicate the likelihood of a person taking loan to default. This insights will help the consumer finance company for taking informed decisions about loan disbursals such as denying the loan, reducing the loan amount, lending (to risky applicants) at a higher interest rate or a combination like reduced amount at a higher interest rate

Outcome

Identified patterns and recommendations based on analysis of the historical loan data given

Business Understanding

Loan approvals are done based on the applicant's profile. Loan approval has 2 types of risks:

1. If loan is approved and applicant is not likely to pay leads to financial loss
2. If loan is not approved and applicant is likely to pay leads to business loss

Hence it is imperative that the finance company does some deep analysis on the applicant's profile by applying the patterns that lead to defaulting, to reduce the financial and business loss risks

Approach

A 3 stage approach was taken to solve the given business problem, as detailed below:

Stage 1 - Data Preparation: In this stage the given dataset was examined in the light of the data dictionary and a relevant dataset for the next stage of data analysis was prepared using the data preparation techniques like irrelevant column removal, removals of non-impacting columns with same values, fixing invalid/NAN values with appropriate default values, performing data validations, datatype fixing and identifying the columns that are required for analysis.

Stage 2 - Data Analysis: In this stage, the dataset prepared in the previous stage of data preparation was analyzed. It started with classifying the parameters into categorial (ordered/un-ordered) and continuous variables. Then the parameters were further classified as independent and dependent parameters. Subsequently, univariate analysis was performed to get basic understanding of the parameters and to find and treat the outliers. Then bivariate analysis was performed to identify the behavioral correlations between parameters. Finally, a multi-variate analysis was performed by creating a derived metric called credit score.

Stage 3 - Conclusion: In this stage, the observations from the analysis were studied to identify the patterns leading to defaulting loans. The identified patterns were summarized, and suitable recommendations were made to augment the finance company to take informed decision about loan approval which may include denying the loan or reducing the loaning amount or increasing the interest or reducing the tenure or a combination of these.

Stage 1 – Data Preparation

Following actions were taken on the given dataset:

1. Dropped columns containing values as Null/NAN in all rows
2. Dropped columns which have same values for all rows
3. Dropped columns which have values as 0 or NAN in all rows
4. Dropped irrelevant and non-impacting columns
5. Fixed data types where values of mixed data types are present using a conversion map
6. Validated data of categorical columns to ensure no rows with invalid data are present
7. Fixed missing values
8. Performed deduplication of data
9. Excluded rows which have null values for certain columns, as null is a valid value and hence these rows can be ignored in the analysis

Stage 1 – Data Preparation on

10. Basic validations were performed on the numeric data to ensure the data integrity
    1. Loan amount consistency
    2. Loan term consistency : 36/60 months
    3. last_payment_date <= next_payment_date
    4. earliest_cr_line <= issue_d
    5. open_acc <= total_acc
    6. total_rec_pricpal <= loan_amnt
    7. pub_rec > pub_rec_bankruptcies
    8. (total_pymnt + total_pymnt_inv) should be approximately equal to (total_rec_prncp + total_rec_int + total_rec_late_fee + recoveries)

Resultant dataset had 47 columns, including the id and member_id, which are the ID columns.

Stage 2 – Data Analysis

Univariate Analysis of Numerical Variables

Univariate Analysis

Out of the numeric variables, dependent variables were skipped in univariate analysis

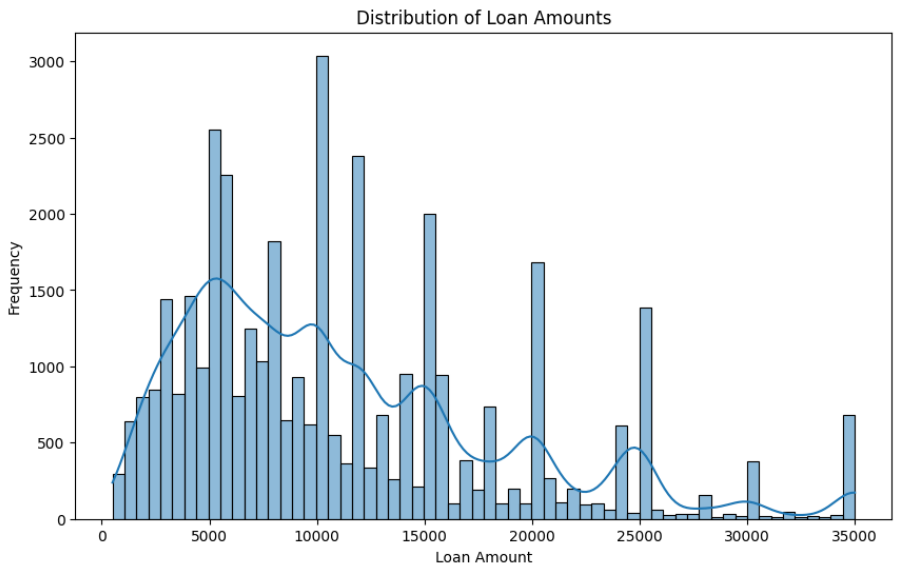
Rationale:

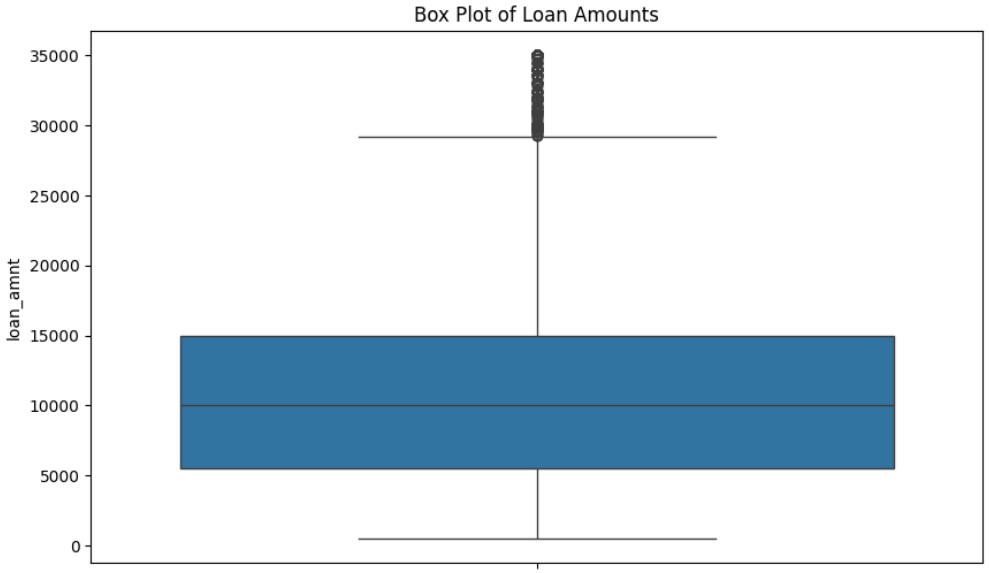
- These variables are dependent on the original loan terms and payment history, which have been analyzed through their primary variables.
- Any outliers or unusual distributions in these dependent variables would already be reflected in the primary variables.
- Skipping detailed univariate analysis for these variables allows for a more efficient focus on other critical aspects of the dataset.

Conclusion: The primary variables related to loan amounts, interest rates, and installments will be thoroughly analyzed. As a result, the dependent variables listed above are considered aligned with the primary data and do not require separate detailed analysis.

Stage 2 – Data Analysis

Loan Amount





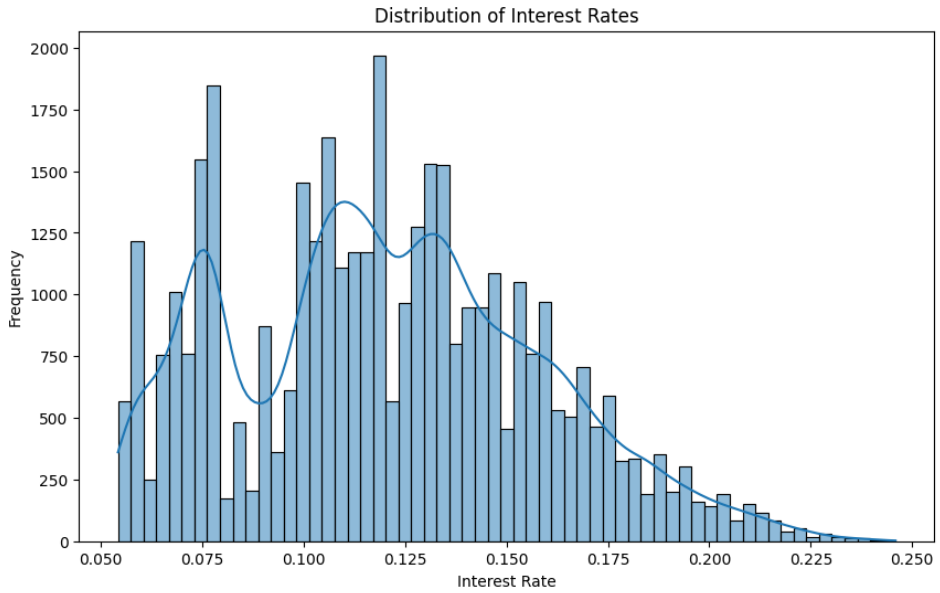
Results of Loan_amnt univerate analysis

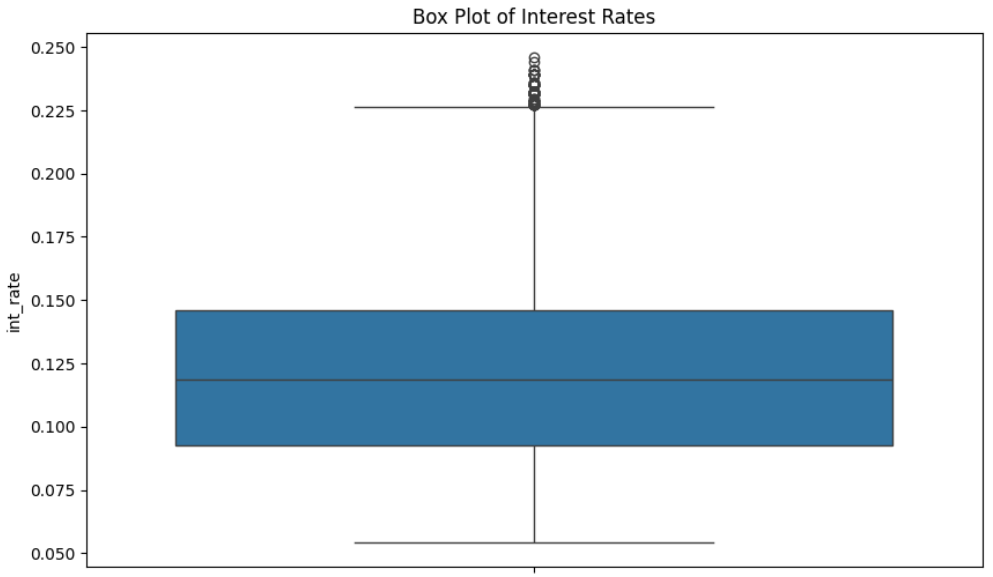
- The loan amounts range from 500 to 35,000, with a mean of approximately 11,248.
- The median loan amount is 10,000, indicating a balanced distribution around this value.
- The standard deviation of 7,470 suggests moderate variability in loan amounts.
- The interquartile range (IQR) is 9,500, showing a reasonable spread between the 25th and 75th percentiles.
- Both the minimum and maximum values are within the expected range for personal loans, with no extreme outliers observed.

Conclusion: All data should be included in the analysis as the distribution appears normal and reflects the typical range of loan amounts.

Stage 2 – Data Analysis

Interest Rate





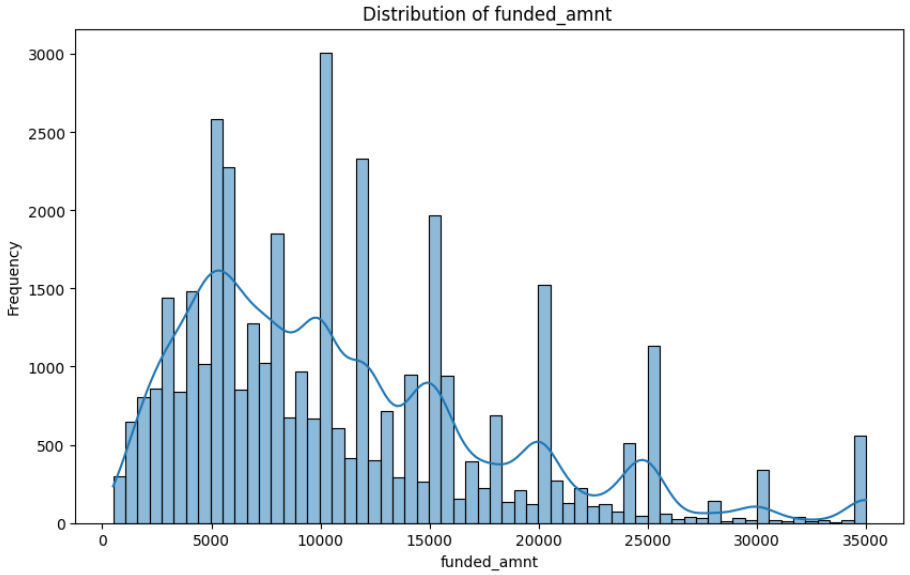
Results of int_rate univerate analysis

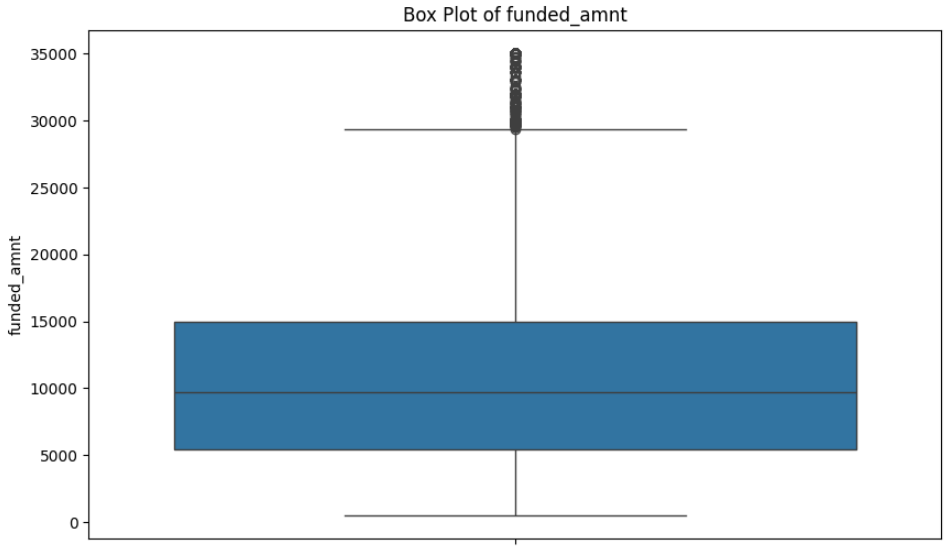
- The interest rates range from 5.42% to 24.59%, with a mean of approximately 12.04%.
- The median interest rate is 11.86%, indicating a slightly lower concentration of interest rates around this value.
- The standard deviation of 3.74% suggests a moderate variability in interest rates.
- The interquartile range (IQR) is 5.36%, with rates ranging from 9.25% (25th percentile) to 14.61% (75th percentile), showing a reasonable spread for loan interest rates.
- Both the minimum and maximum values are within the expected range for personal loans, with no extreme outliers observed.

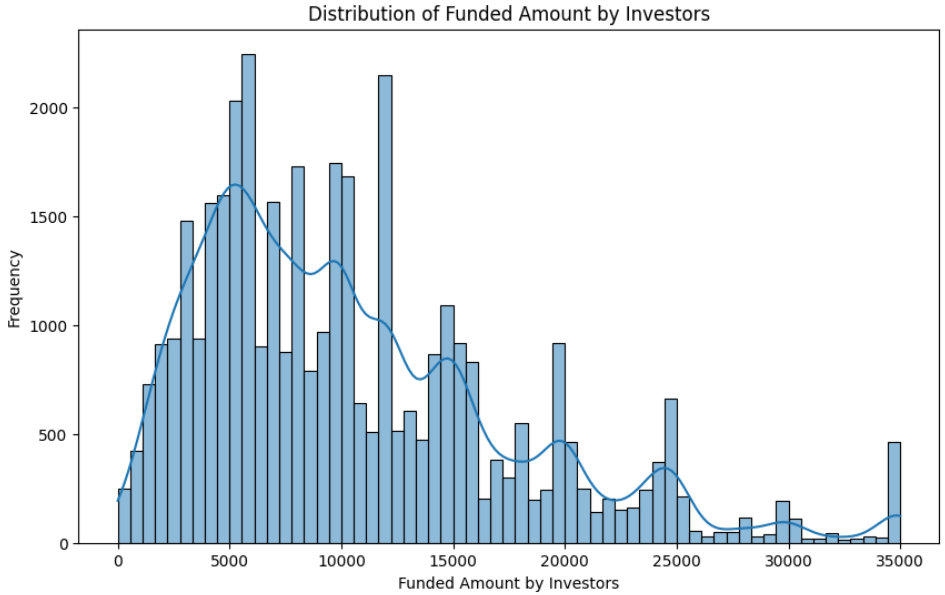
Conclusion: All data should be included in the analysis as the distribution appears normal and reflects the typical range of interest rates offered to borrowers.

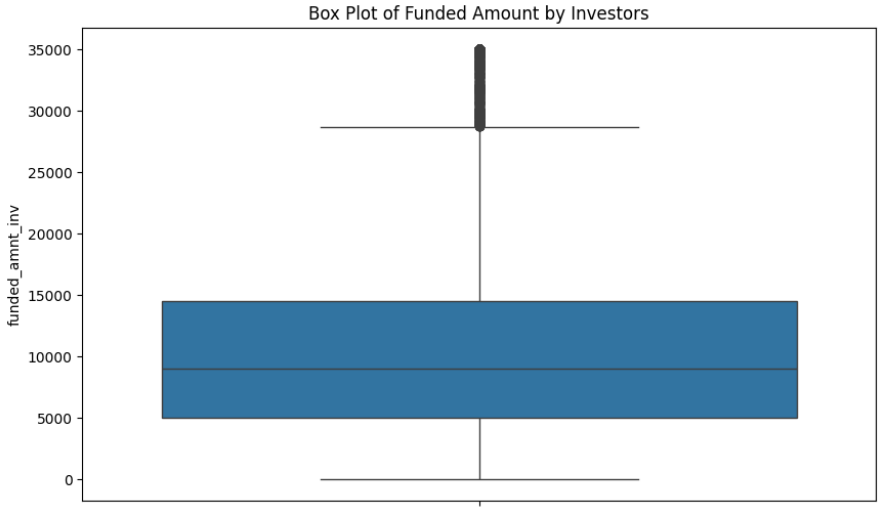
Stage 2 – Data Analysis

Funded Amount & Funded Amount By Investors





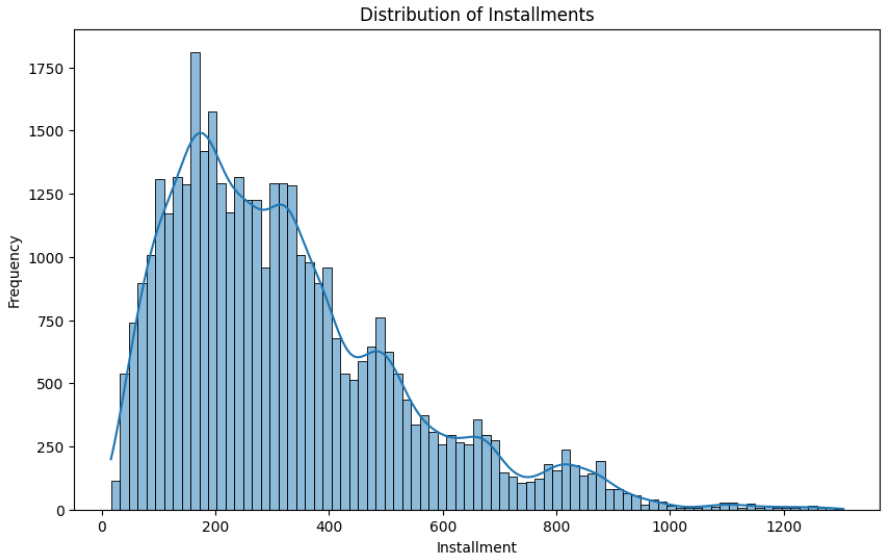


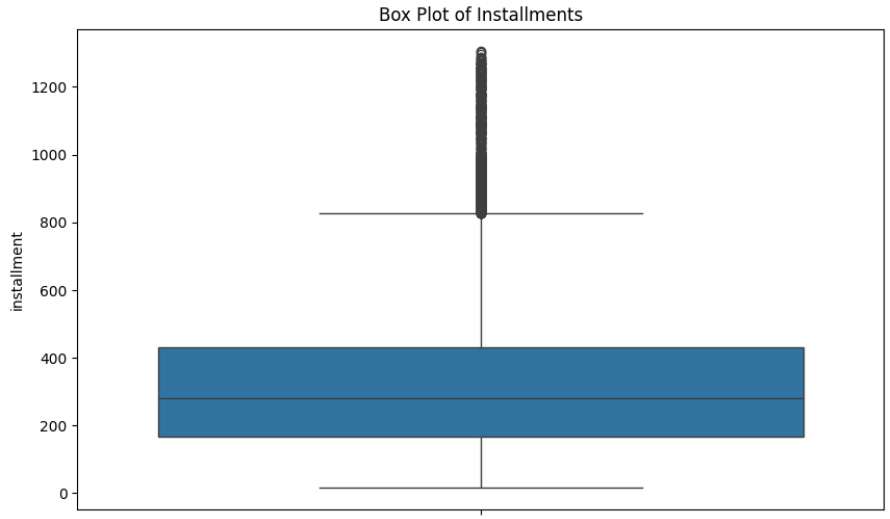


Conclusion: The funded amount and funded amount by investors exhibit the same behaviour as loan amount

Stage 2 – Data Analysis

Installment



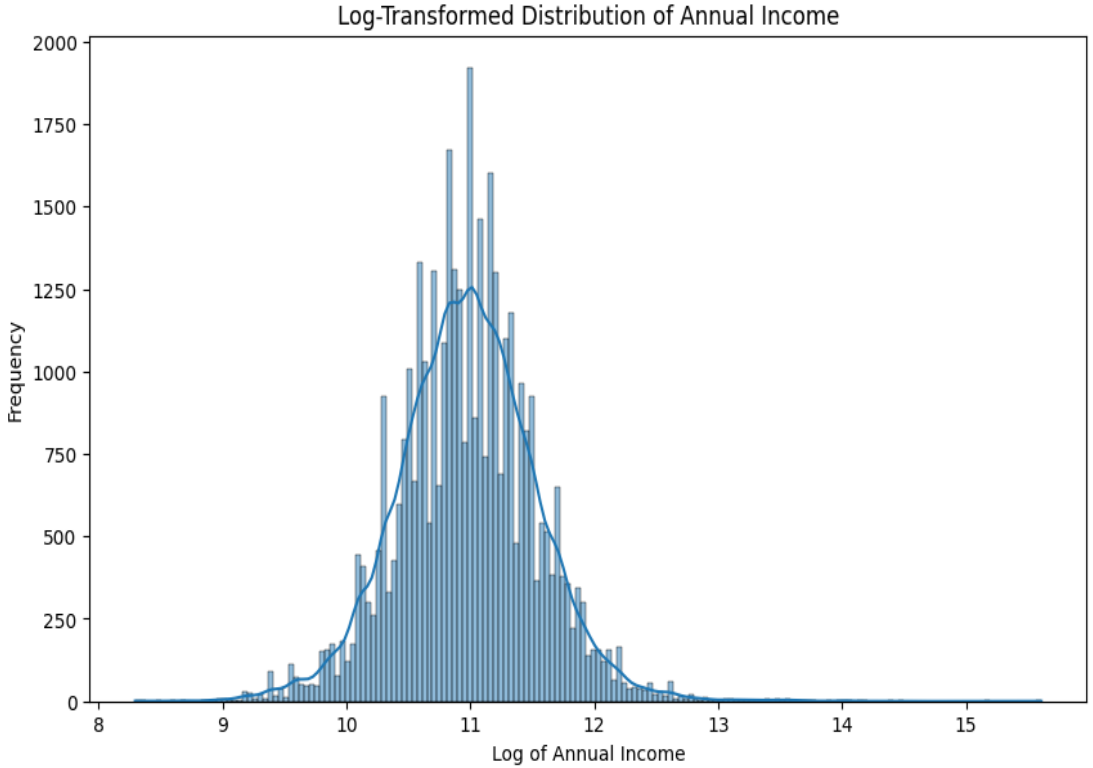


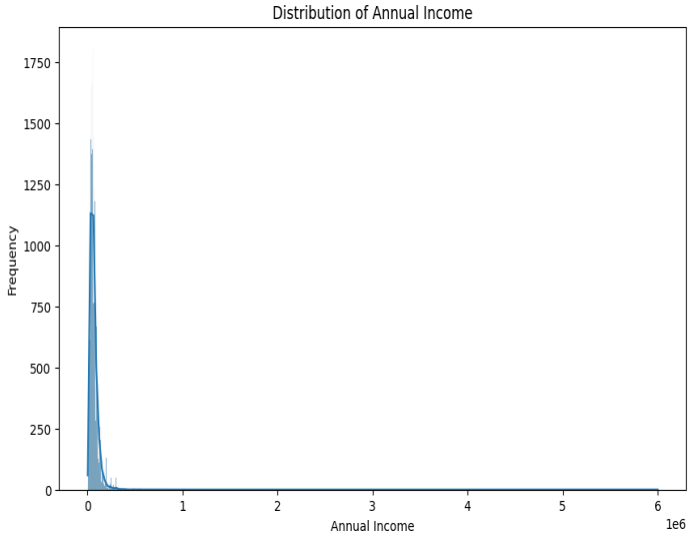
Results of int_rate univerate analysis

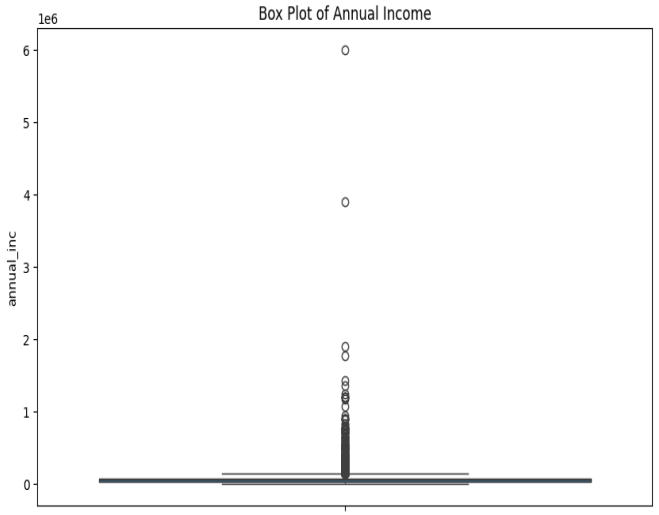
- The installment is a derived metric based on the loan_amnt and int_rate, calculated using the loan's principal, interest rate, and term. Given that we've already determined that there are no significant outliers in loan_amnt and int_rate, and that these variables are within expected ranges, the same reasoning applies to installment.
- We can also see this with the above histo and box plot.

Stage 2 – Data Analysis

Annual Income







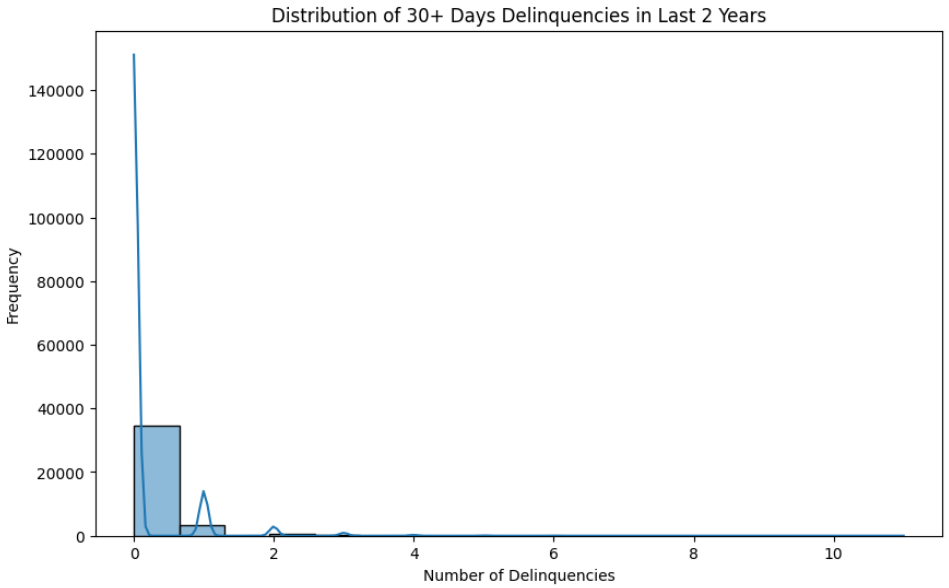
Results of annual_inc univerate analysis

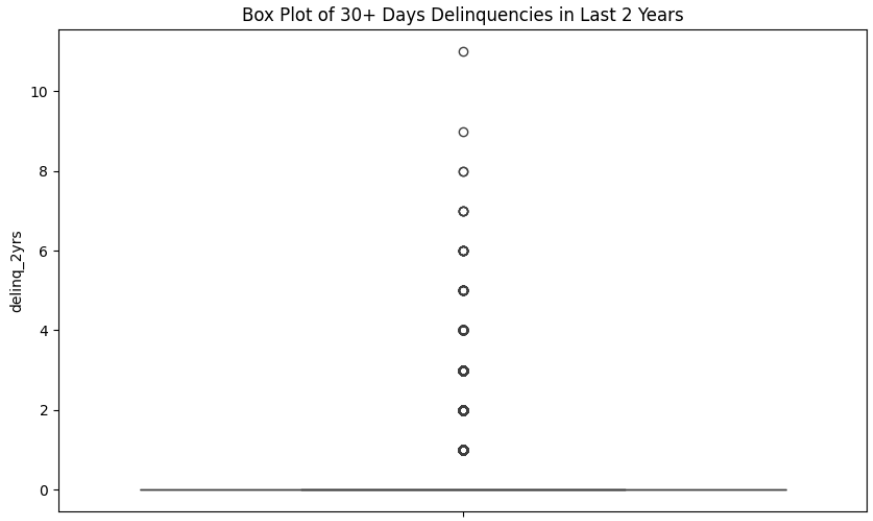
- The annual incomes range from 4,000to6,000,000, with a mean of approximately $68,970.
- The median income is $59,020, indicating a slight skew toward higher incomes.
- The standard deviation of $63,165 suggests significant variability in income levels.
- The distribution is right-skewed, with high-income outliers notably impacting the mean.

Conclusion: While outliers exist, all data should be included in the analysis. Special handling, such as log transformation, may be needed in bivariate or multivariate analyses involving annual_inc.

Stage 2 – Data Analysis

Delinq 2 years



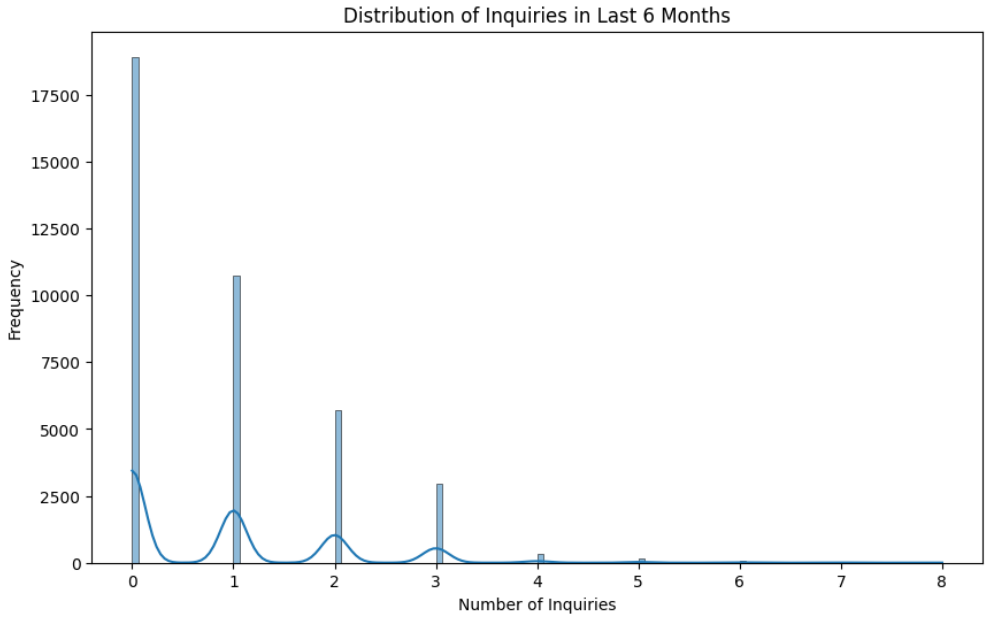


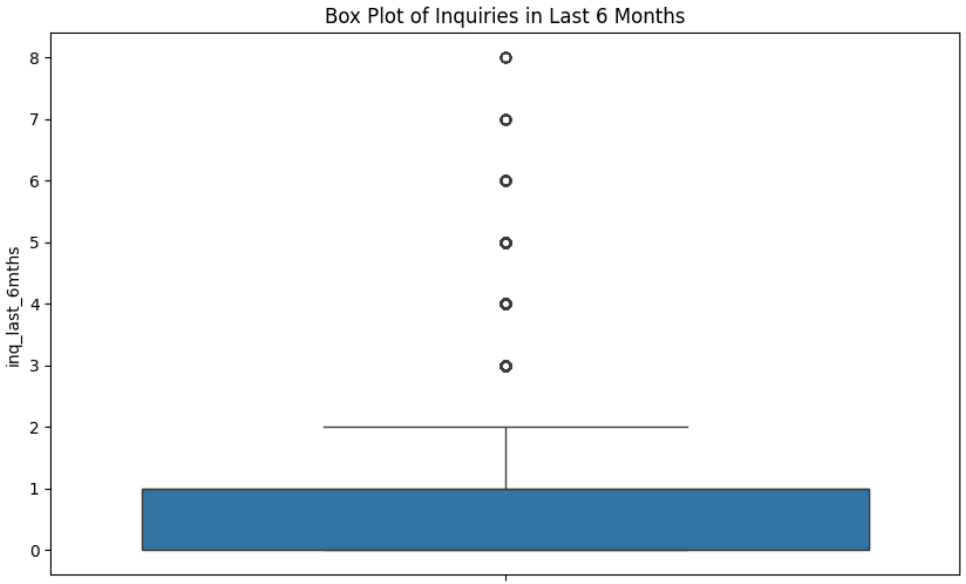
Results of delinq_2yrs Univerate analysis

- The summary statistics for delinq_2yrs indicate that the majority of borrowers have had no 30+ days delinquencies in the past 2 years.
- Given that the data is skewed but expected (many borrowers with no delinquencies), we should include all data in analysis when we move to bivariate analysis, particularly examining the relationship between delinq_2yrs and loan performance, we will consider the outliers more carefully.

Stage 2 – Data Analysis

Inquiries in last 6 months



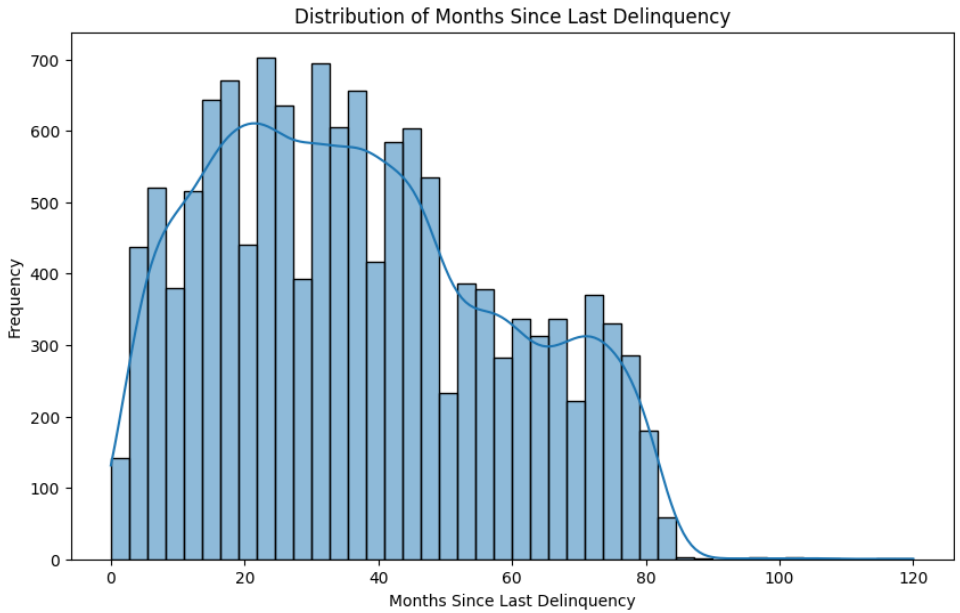


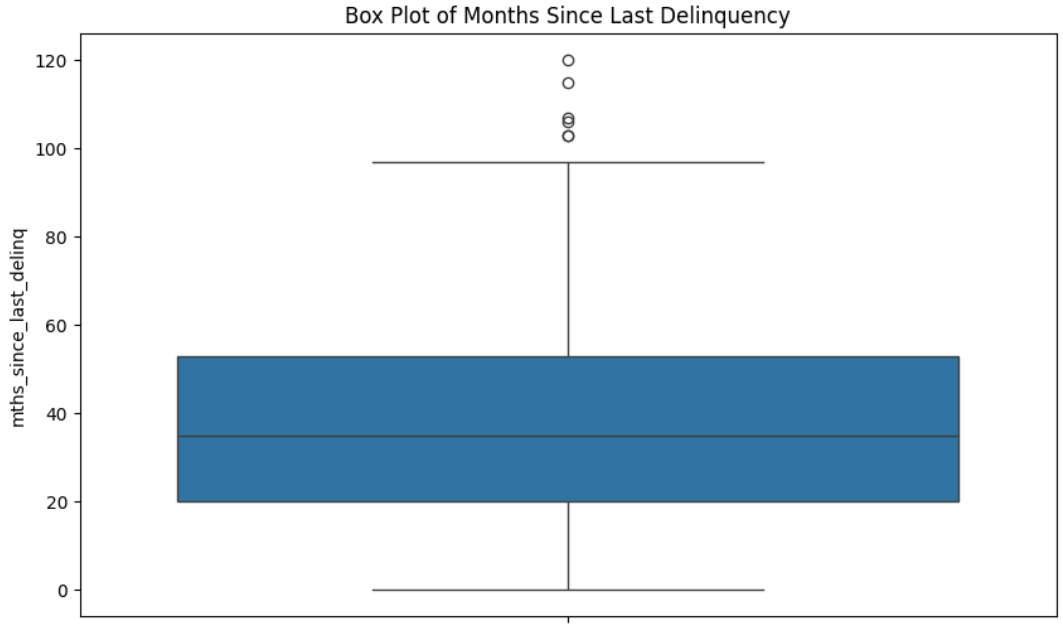
Results : inq_last_6mths Univerate analysis

- The number of inquiries in the last 6 months ranges from 0 to 8, with a mean of approximately 0.87.
- The median number of inquiries is 1, indicating that most borrowers had one or fewer inquiries in the last 6 months.
- The standard deviation of 1.07 suggests that the number of inquiries varies, but most values are close to the mean.
- The 25th and 75th percentiles are both 0 and 1, respectively, showing that the majority of borrowers had either 0 or 1 inquiry

Conclusion: The distribution is right-skewed, with no extreme outliers. All data should be included in the analysis as it represents typical borrower behavior.

Stage 2 – Data Analysis



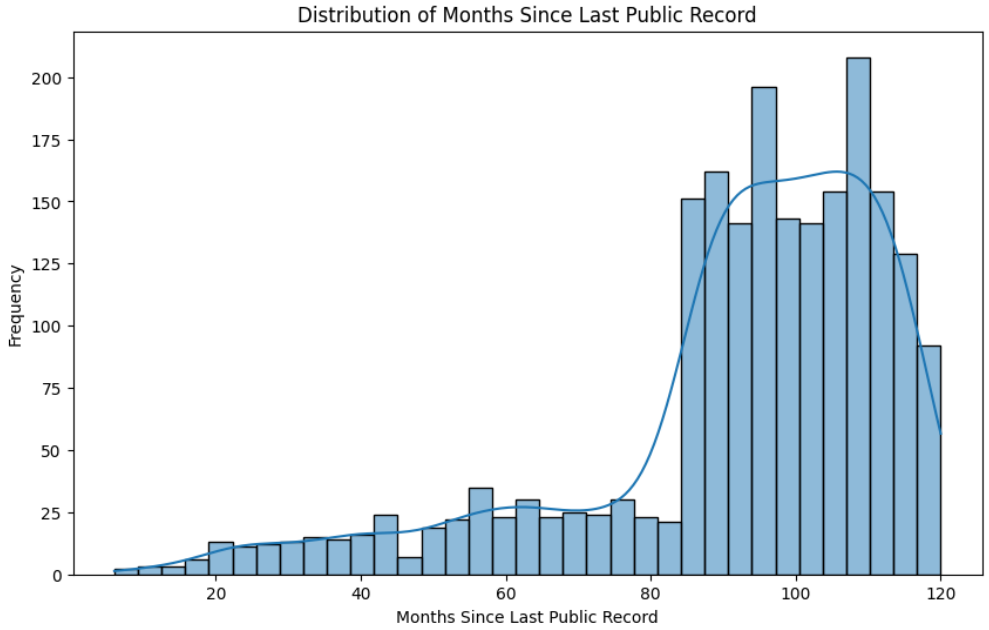


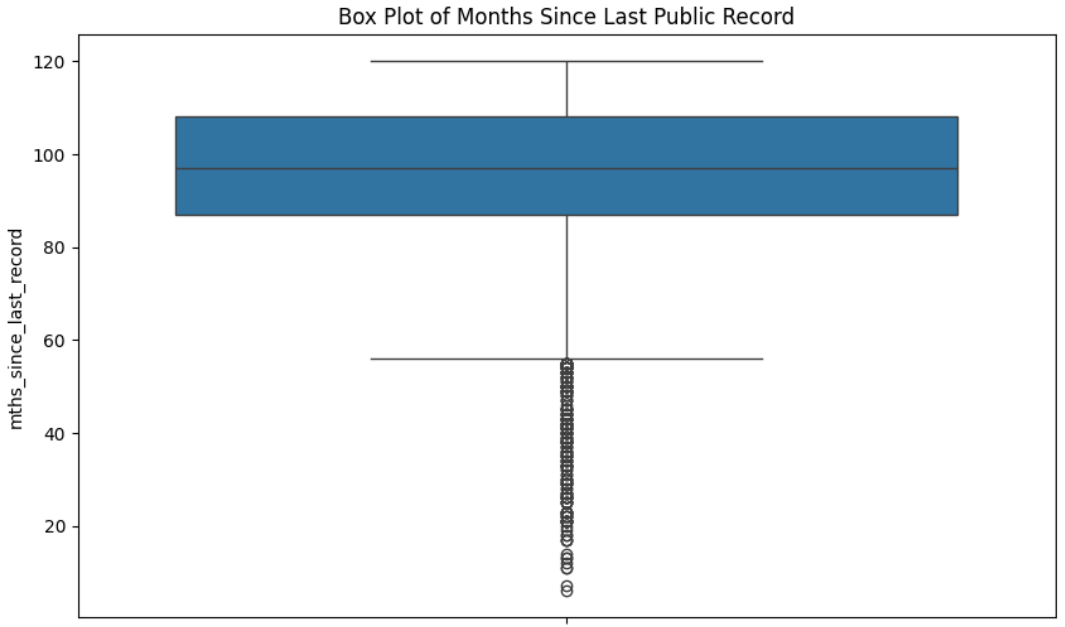
Months since last Delinquency

Results: mths_since_last_delinq Univerate analysis

- The number of months since the last delinquency has a mean of approximately 37 months.
- The standard deviation is fine at 21 months, indicating significant variability in the data.
- The minimum value is 0 months, representing recent delinquencies.
- The 25th percentile is at 20 months, while the 50th and 75th percentiles are at 35 and 53 months.
- This data is only for members who have delinq in previous loans. We should do biverate analysis to check if it has any relation to defaulters.

Stage 2 – Data Analysis



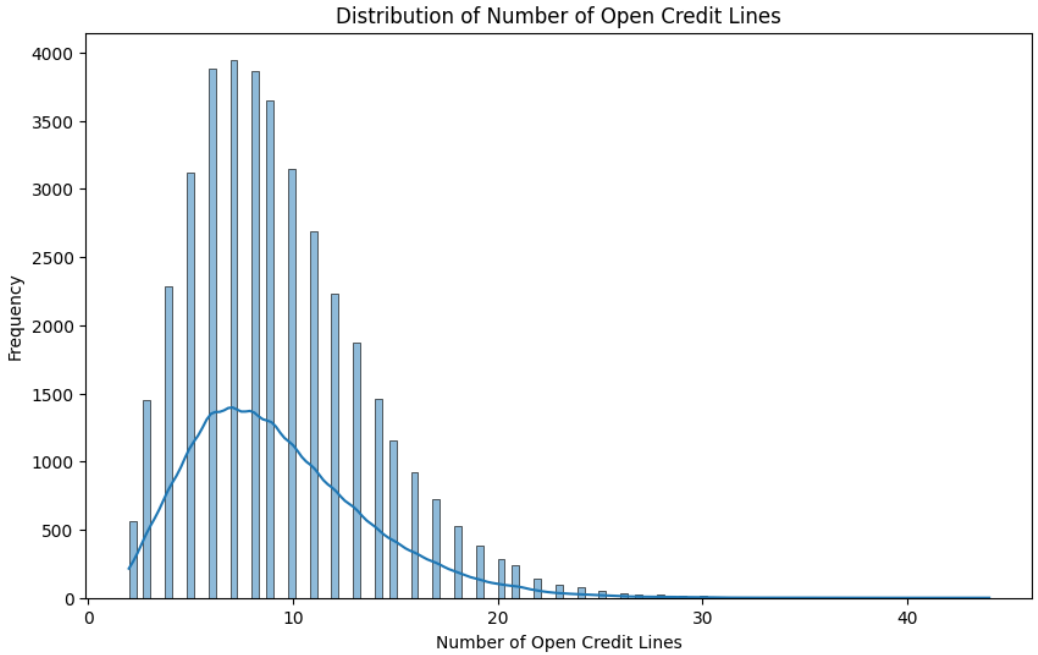


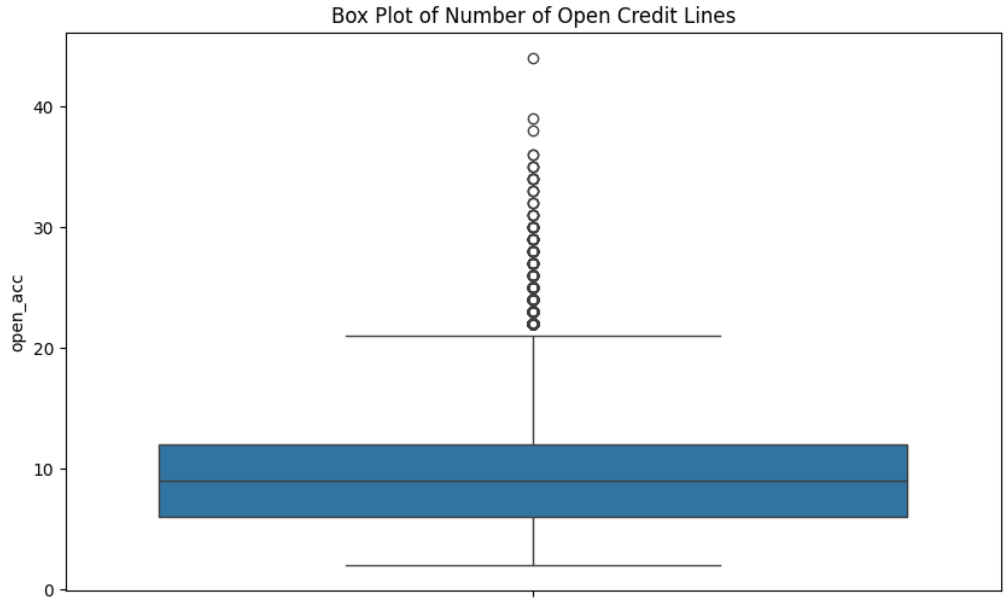
Months since last Public Record

Results mths_since_last_record Univerate analysis

- The average time since the last public record is approximately 92 months, indicating that most records occurred over seven years ago.
- The standard deviation is around 22 months, reflecting some variability in the timing of public records.
- The most recent public record in the dataset occurred 6 months ago.
- The distribution of data shows that 25% of the records are from 87 months ago or less, the median is 97 months, and 75% are from 108 months ago or less.
- The maximum value is 120 months, suggesting that the data primarily includes public records from within the past 10 years.
- This data is only for members who have public record from previous loans. We should do biverate analysis to check if it has any relation to defaulters.

Stage 2 – Data Analysis





Open Credit Lines

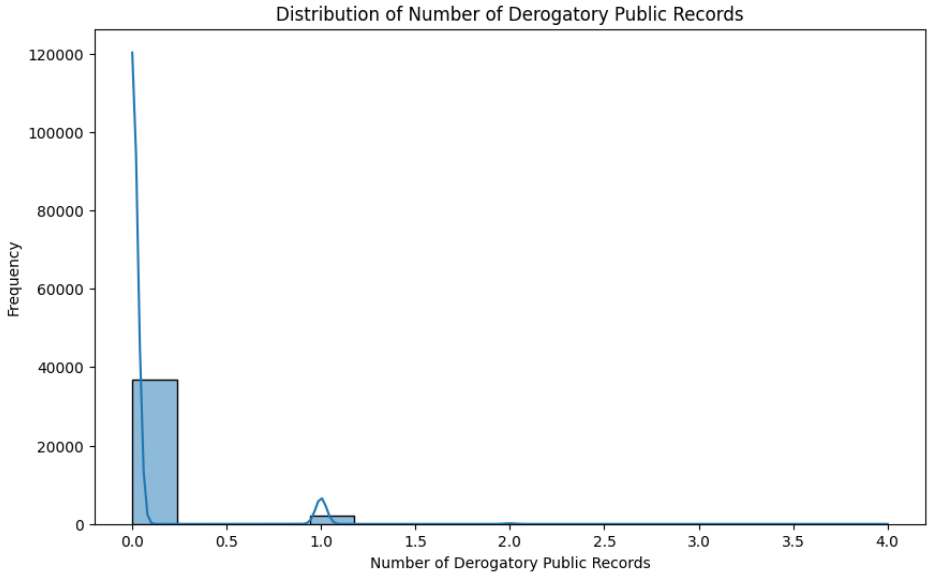
Results: open_acc Univerate analysis

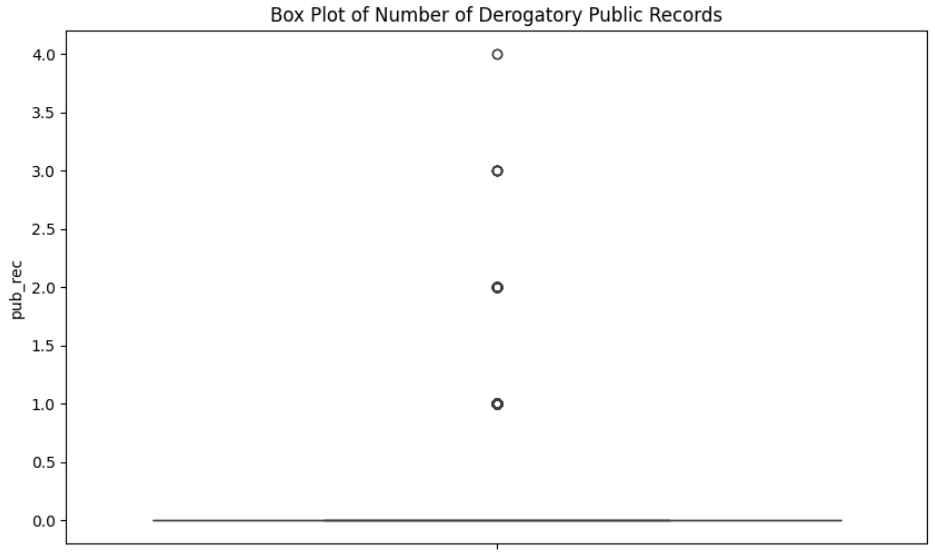
- The number of open credit lines ranges from 2 to 44, with a mean of approximately 9.29.
- The median number of open accounts is 9, indicating a balanced distribution around this value.
- The standard deviation of 4.38 suggests moderate variability in the number of open credit lines among borrowers.
- The 25th percentile is at 6 open accounts, while the 75th percentile is at 12, showing a reasonable spread in the data.

Conclusion: The distribution appears normal, with no extreme outliers. All data should be included in the analysis as it represents typical borrower credit profiles.

Stage 2 – Data Analysis

Number of Derogatory Public Records

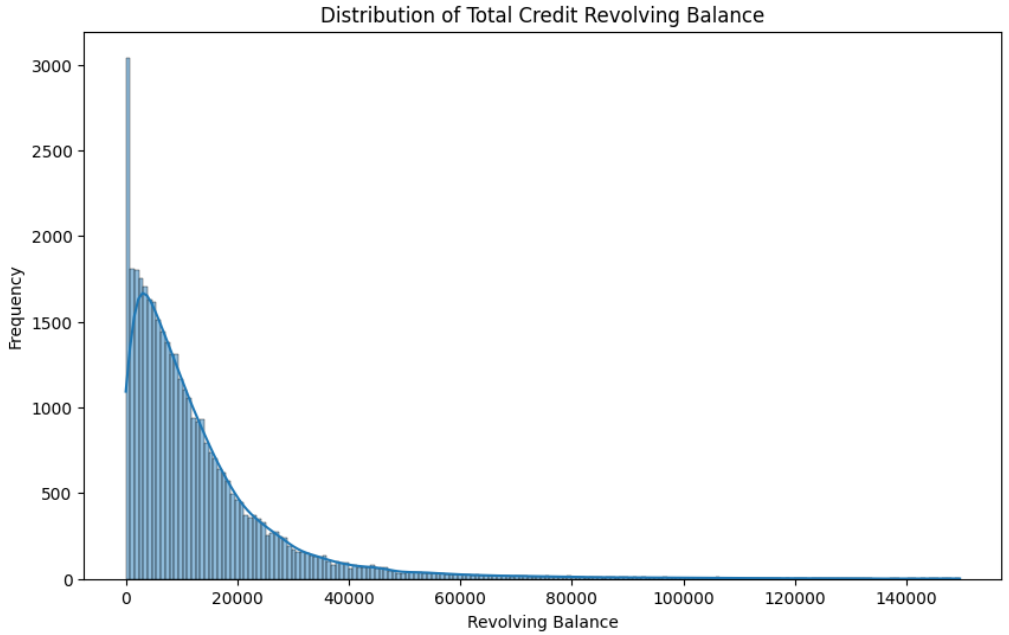


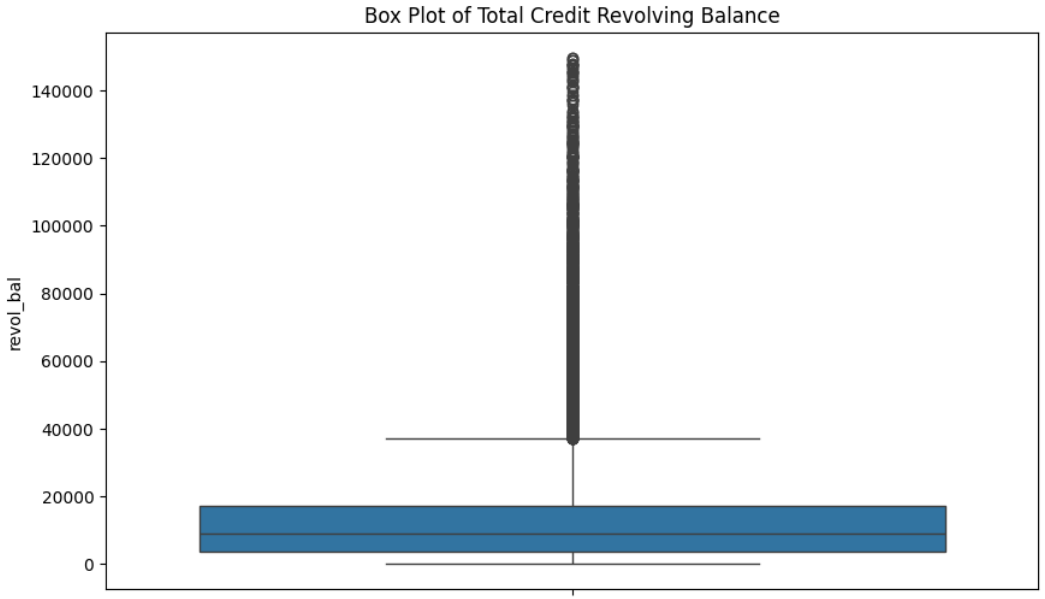


Results of pub_rec Univerate analysis

- The number of derogatory public records ranges from 0 to 4, with a mean of approximately 0.055.
- The median value is 0, indicating that the majority of borrowers have no derogatory public records.
- The standard deviation of 0.238 suggests very low variability, with most values close to 0.
- The 25th, 50th, and 75th percentiles are all 0, showing that derogatory public records are rare in this dataset.
- Conclusion: The distribution is heavily skewed towards 0, with no extreme outliers. All data should be included in the analysis, as it accurately reflects the rarity of derogatory public records among borrowers.
- This data is only for members who have public record from previous loans. We should do biverate analysis to check if it has any relation to defaulters.

Stage 2 – Data Analysis





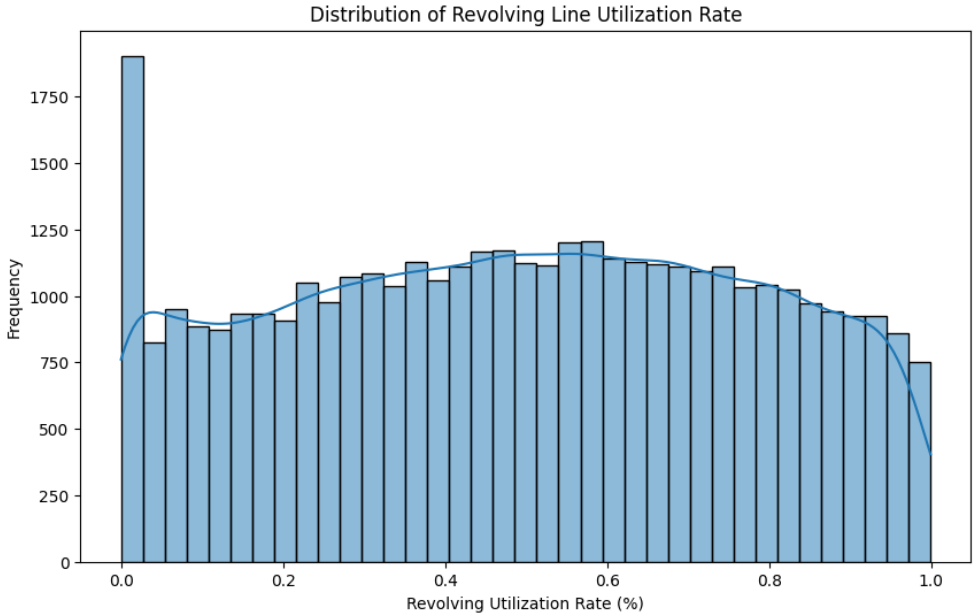
Total Credit Revolving Balance

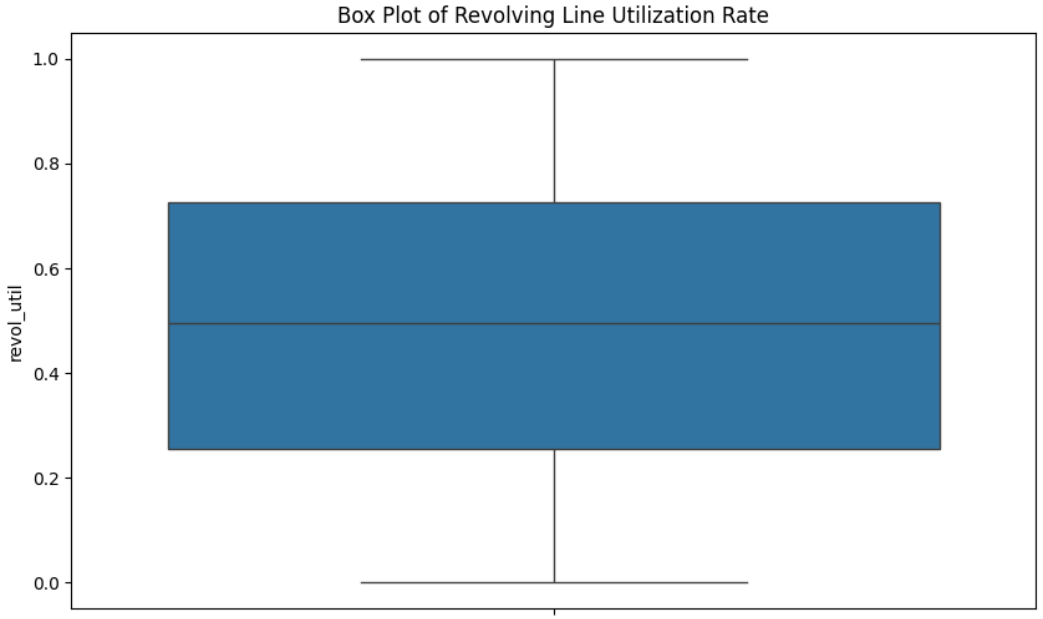
Results of revol_bal Univerate analysis

- The total credit revolving balance ranges from 0to149,588, with a mean of approximately $13,381.
- The median revolving balance is $8,868, indicating that half of the borrowers have a balance below this amount.
- The standard deviation of $15,829 suggests significant variability in revolving balances among borrowers.
- The 25th percentile is 3,734,whilethe75thpercentileis17,063, showing a wide range in the distribution of revolving balances

Conclusion: The distribution shows a reasonable spread of revolving balances with no extreme outliers. All data should be included in the analysis as it reflects the typical distribution of credit revolving balances.

Stage 2 – Data Analysis





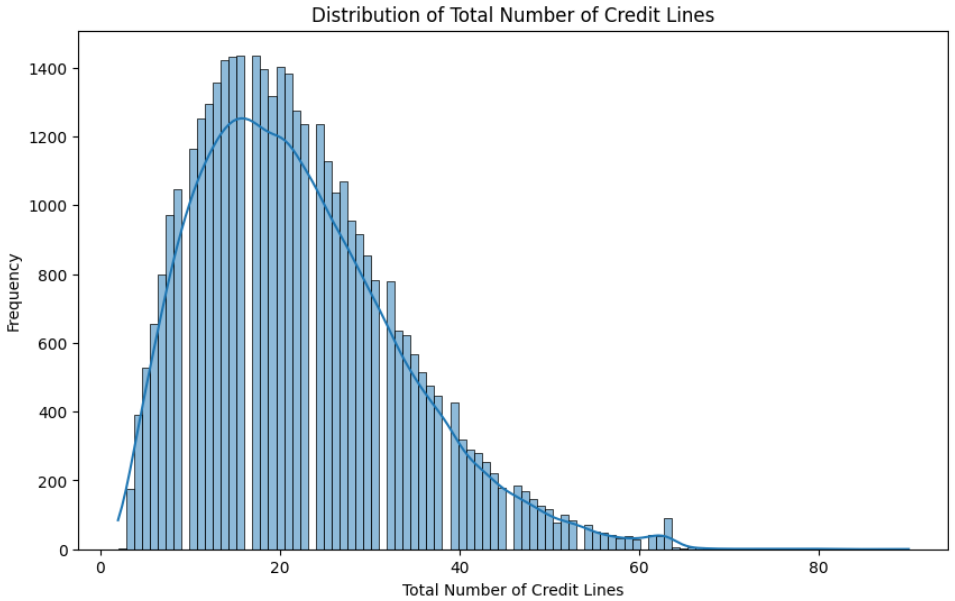
Revolving Line Utilization Rate

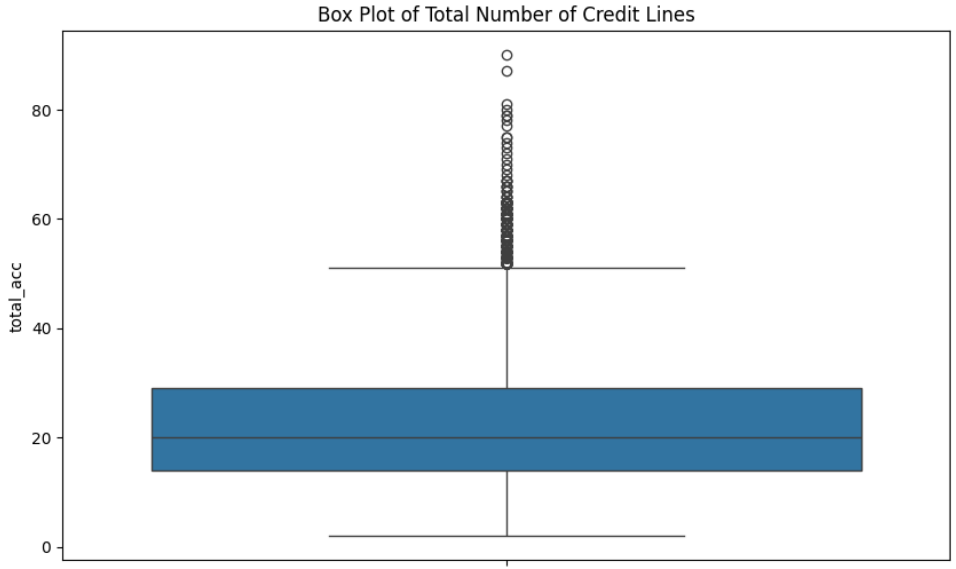
Results revol_util Univerate analysis

- The revolving line utilization rate ranges from 0% to 99.9%, with a mean of approximately 48.98%.
- The median utilization rate is 49.5%, indicating that half of the borrowers are using nearly half of their available revolving credit.
- The standard deviation of 28.30% suggests a wide variability in utilization rates among borrowers.
- The 25th percentile is at 25.6%, and the 75th percentile is at 72.5%, showing a significant spread in how borrowers use their revolving credit.

Conclusion: The distribution shows a reasonable spread with no extreme outliers. All data should be included in the analysis as it reflects the typical usage of revolving credit by borrowers.

Stage 2 – Data Analysis





Total Credit Lines

Results: total_acc Univerate analysis

- The total number of credit lines ranges from 2 to 90, with a mean of approximately 22.14.
- The median value is 20 credit lines, indicating a balanced distribution around this value.
- The standard deviation of 11.39 suggests significant variability in the total number of credit lines among borrowers.
- The 25th percentile is at 14 credit lines, and the 75th percentile is at 29 credit lines, showing a broad range in the data.

Conclusion: The distribution shows a wide spread with no extreme outliers. All data should be included in the analysis as it reflects the typical number of credit lines in borrowers' credit files.

Stage 2 – Data Analysis

Univariate Analysis of Ordered Categorical Variables

These categorical variables have a natural order or ranking.

term: Number of payments on the loan (e.g., 36 months, 60 months).

Encoded as an integer type to preserve the order in the analysis.

grade: Loan grade (e.g., A, B, C, D, E, F, G).

Will be encoded as an ordered categorical variable.

sub_grade: Loan sub-grade (e.g., A1, A2, B1, B2, etc.).

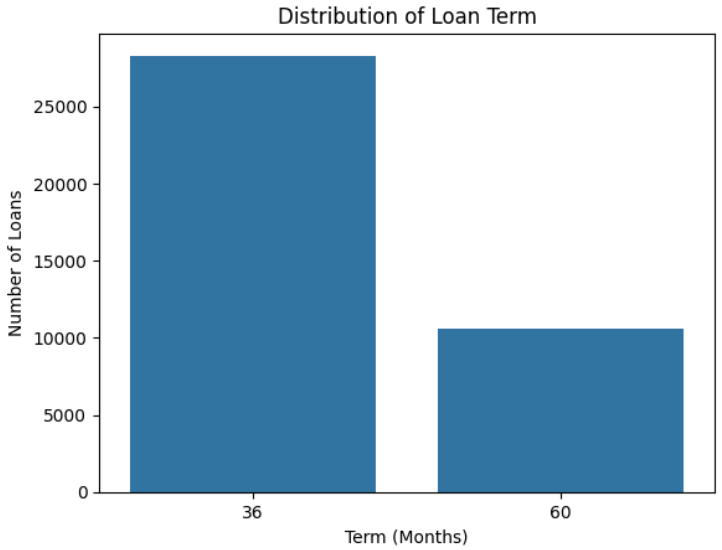
Will be encoded as an ordered categorical variable.

emp_length: Length of employment (e.g., <1 year, 1-2 years, 10+ years).

Will be encoded as an ordered categorical variable.

Stage 2 – Data Analysis

Term

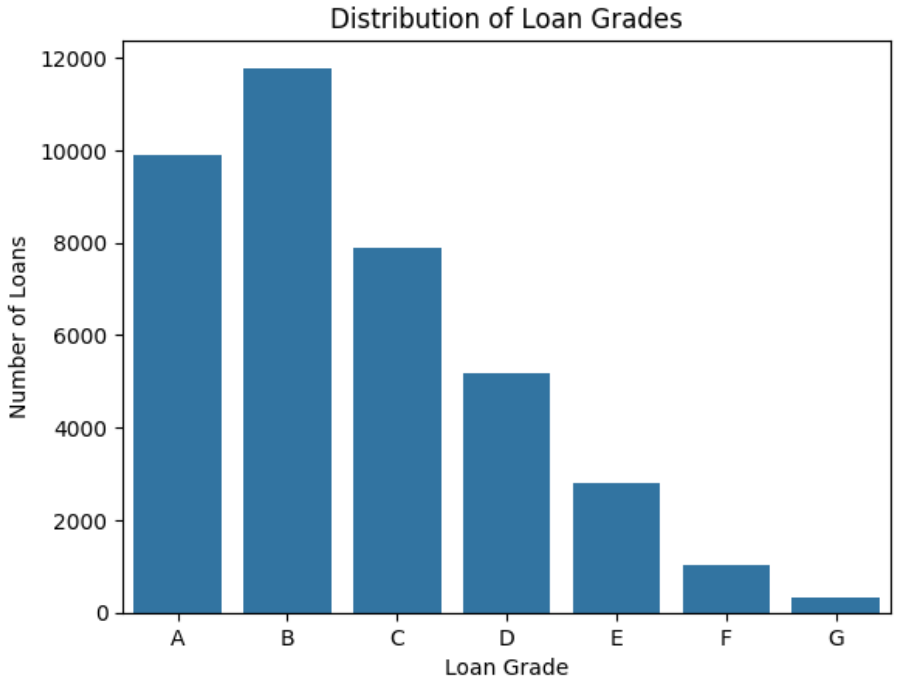


Results: term Univerate analysis

- Dominance of 36-Month Term: The majority of loans (28,289) have a 36-month term, indicating that this is the most common loan duration in your dataset.
- 60-Month Term: A significant number of loans (10,592) have a 60-month term, but it is less common compared to the 36-month term.

Stage 2 – Data Analysis

Grade

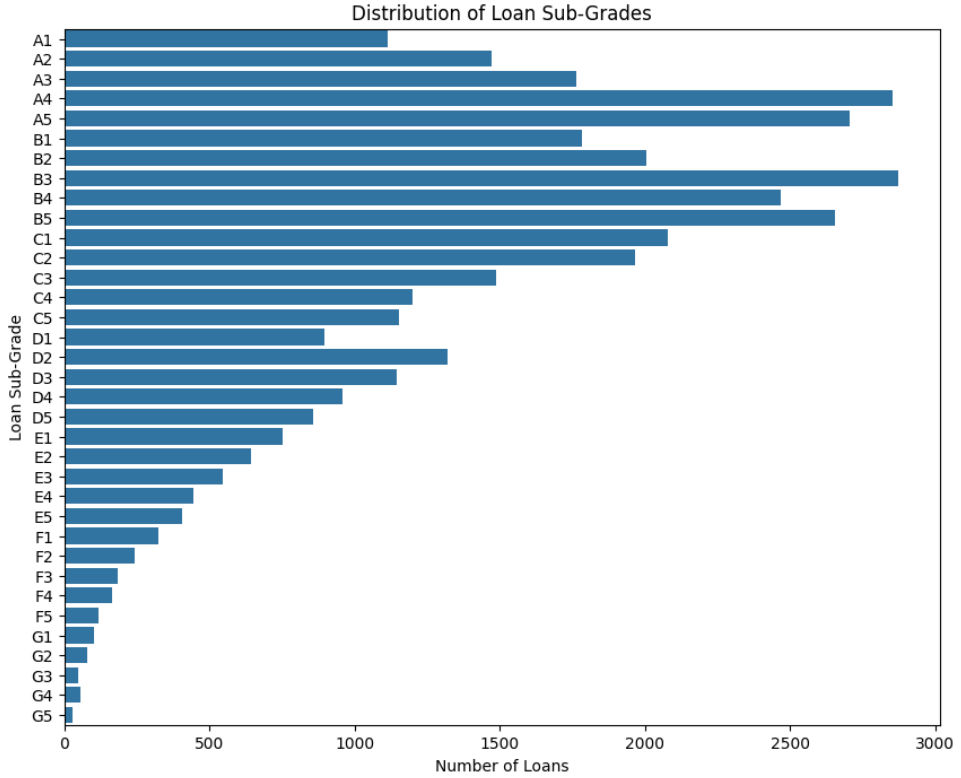


Results: grade Univerate analysis

- **Observation**: The distribution shows a skew towards the higher grades (A, B), with fewer loans in the lower grades (E, F, G).
- **Interpretation**: This skew reflects a typical risk management strategy, where lenders prefer to offer more loans to borrowers with better credit profiles.

Stage 2 – Data Analysis

Sub Grade

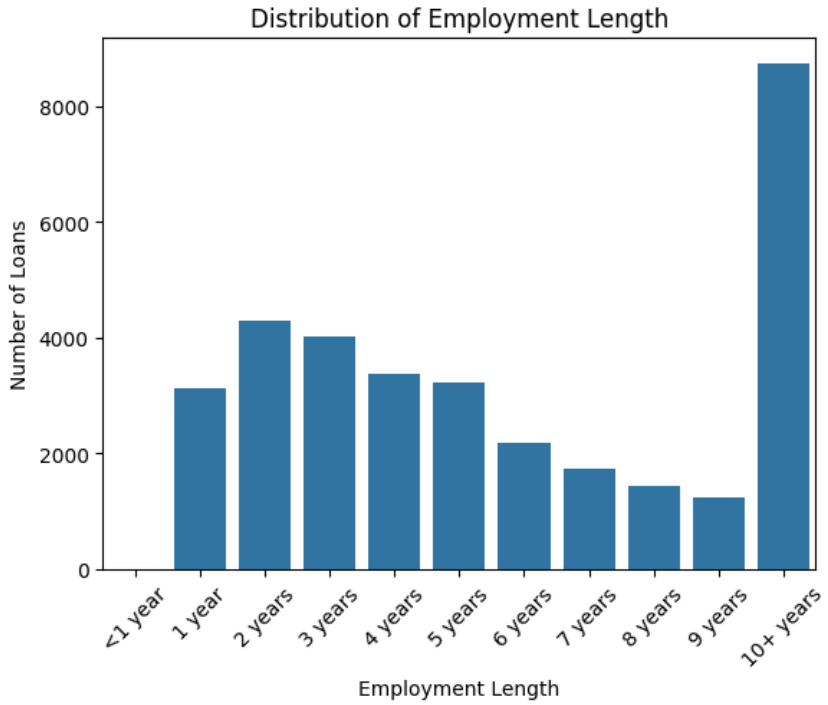


Results: sub_grade Univerate analysis

- **Observation**: The distribution shows a skew towards the higher grades (A, B), with fewer loans in the lower grades (E, F, G).
- **Interpretation**: This skew reflects a typical risk management strategy, where lenders prefer to offer more loans to borrowers with better credit profiles.

Stage 2 – Data Analysis

Length of Employment



Results: sub_grade Univerate analysis

- The data shows that most borrowers have been employed for 10+ years, which might correlate with higher creditworthiness.
- Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

Stage 2 – Data Analysis

Univariate Analysis of Ordered Categorical Variables

These categorical variables do not have a natural order or ranking.

id: Loan ID : Converted to string data type.

member_id: Member ID : Converted to string data type.

emp_title: Job title of the borrower. : Converted to string data type.

home_ownership: Home ownership status (e.g., Rent, Own, Mortgage) : Converted to enum data type.

verification_status: Income verification status (e.g., Verified, Not Verified) : Converted to enum data type.

loan_status: Current status of the loan (e.g., Fully Paid, Charged Off, Current) : Converted to enum data type.

desc: Loan description : Converted to string data type.

purpose: Purpose of the loan (e.g., Debt consolidation, Credit card) : Converted to string data type.

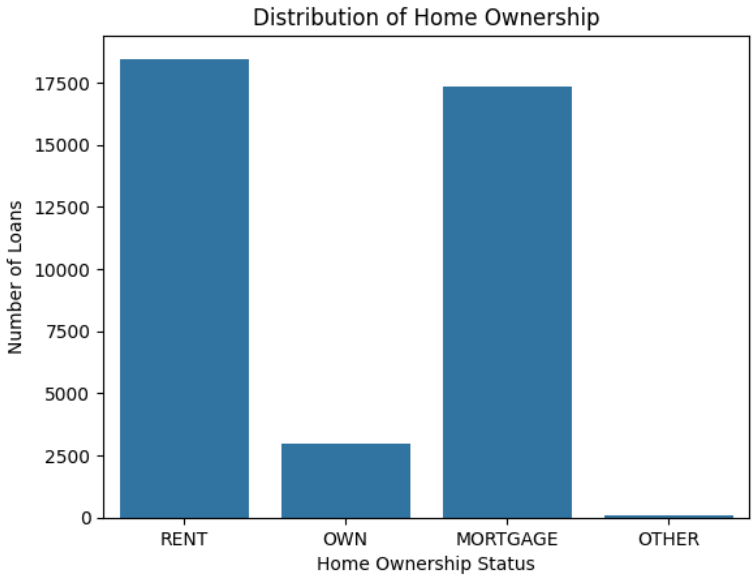
title: Loan title : Converted to string data type.

zip_code: First 3 digits of the borrower's zip code : Converted to string data type.

addr_state: State of the address provided by the borrower : Converted to string data type.

Stage 2 – Data Analysis

Home Ownership Status

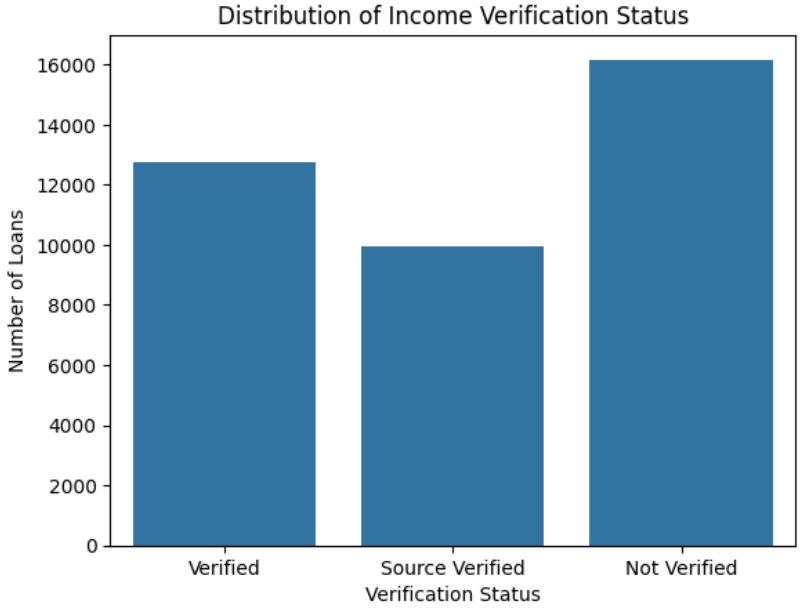


Results: home_ownership Univerate analysis

- The distribution of home ownership status shows a heavy skew towards renters and mortgage holders, with fewer borrowers owning their homes outright.
- Negligible OTHER Category: The OTHER category will have a very small bar, showing that this is a rare home ownership status among borrowers.

Stage 2 – Data Analysis

Income Verification Status

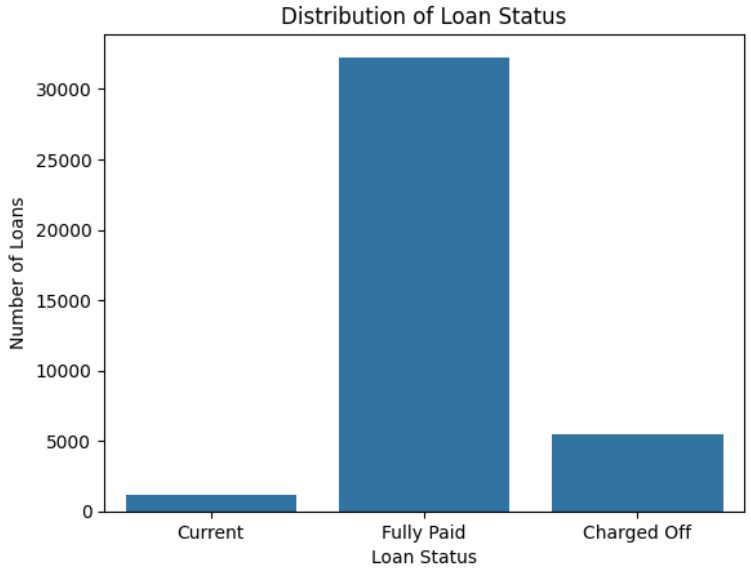


Results: verification_status Univerate analysis

- The high number of Not Verified loans could suggest a higher risk profile for the loan portfolio. Loans that have not undergone full verification may have a higher likelihood of default, which is important to consider in risk assessments.

Stage 2 – Data Analysis

Loan Status

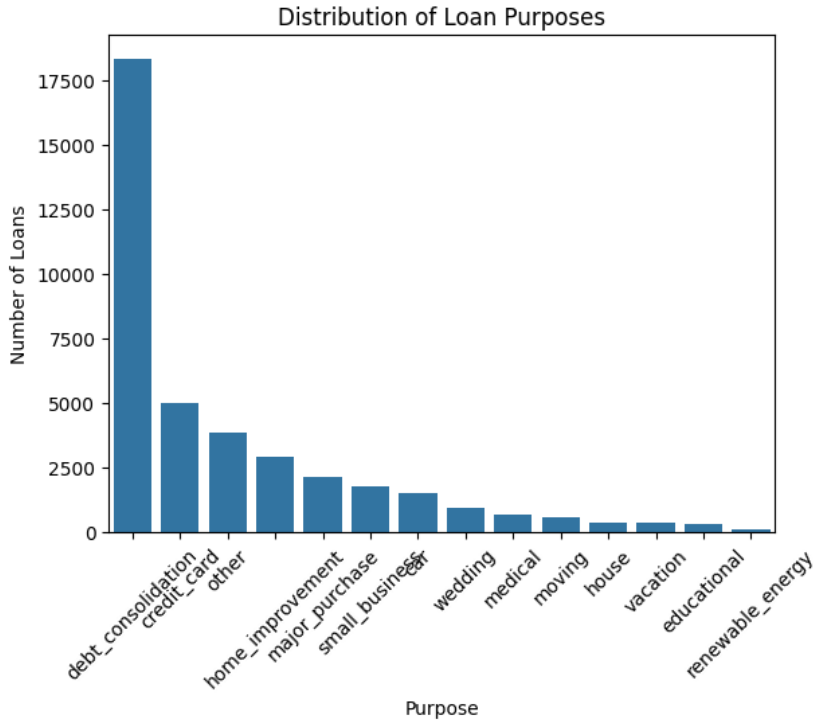


Results: loan_status Univerate analysis

- **Majority of Loans Fully Paid Off**: The analysis shows that the majority of loans in the dataset have been fully paid off.
- **Charged Off Loans**: There are 5,476 loans in the dataset that have been charged off, representing cases where borrowers have defaulted on their obligations.

Stage 2 – Data Analysis

Purpose

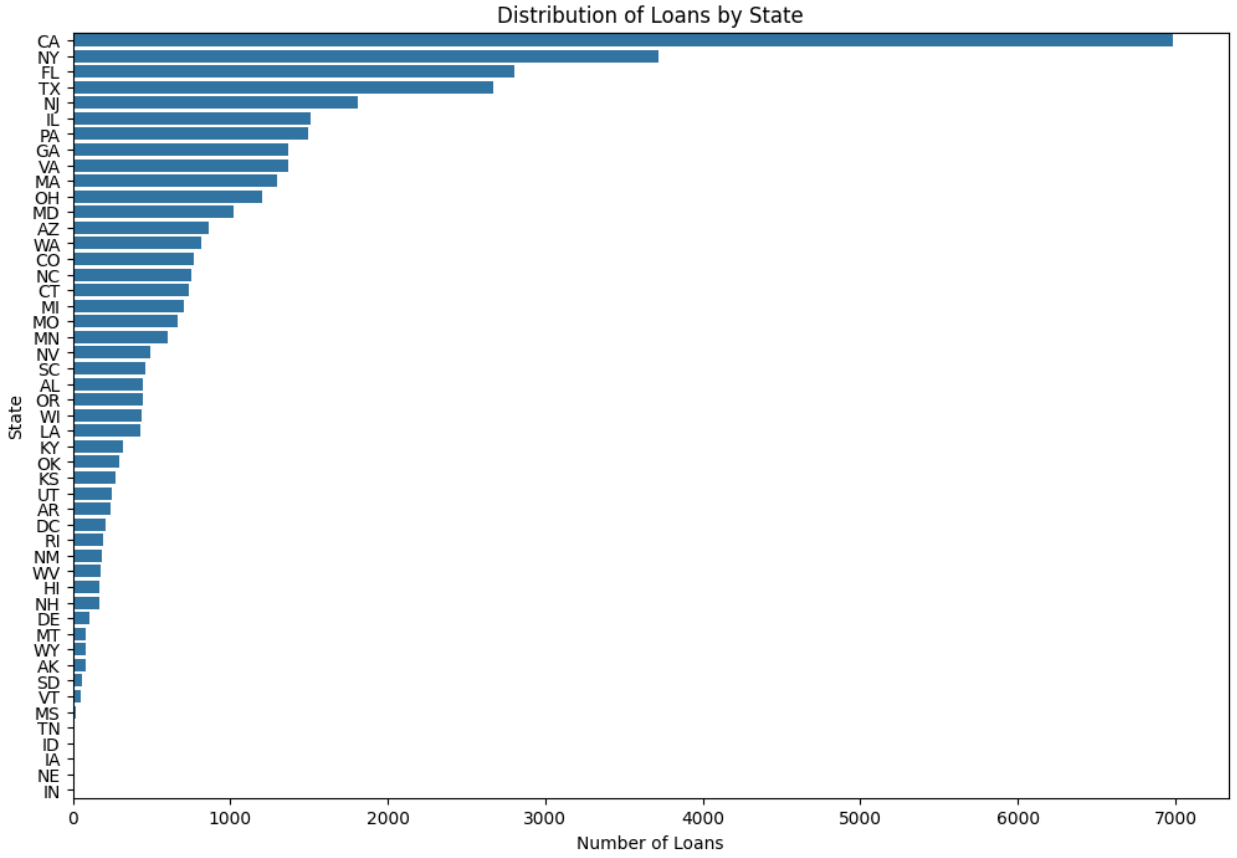


Results: purpose univerate analysis

- The dominance of Debt Consolidation and Credit Card loan purposes suggests that many borrowers are focused on managing and consolidating existing debt.
- The range of other loan purposes, from Home Improvement to Small Business, reflects the diverse financial needs that personal loans help fulfill.
- Niche Categories: Although less common, loan purposes like Educational, Renewable Energy, and Vacation show that some borrowers seek loans for more specific, and sometimes non-essential, purposes.

Stage 2 – Data Analysis

Loans by State



Results: addr_state univerate analysis

- High Concentration in California (CA)
- New York (3,717), Florida (2,804), and Texas (2,671) also have a large number of loans.
- States like New Jersey (1,811), Illinois (1,510), and Pennsylvania (1,496) have moderate numbers of loans.
- Some states like Iowa (1), Nebraska (1), and Indiana (1) have almost negligible loan counts.

Stage 2 – Data Analysis

Bivariate Analysis

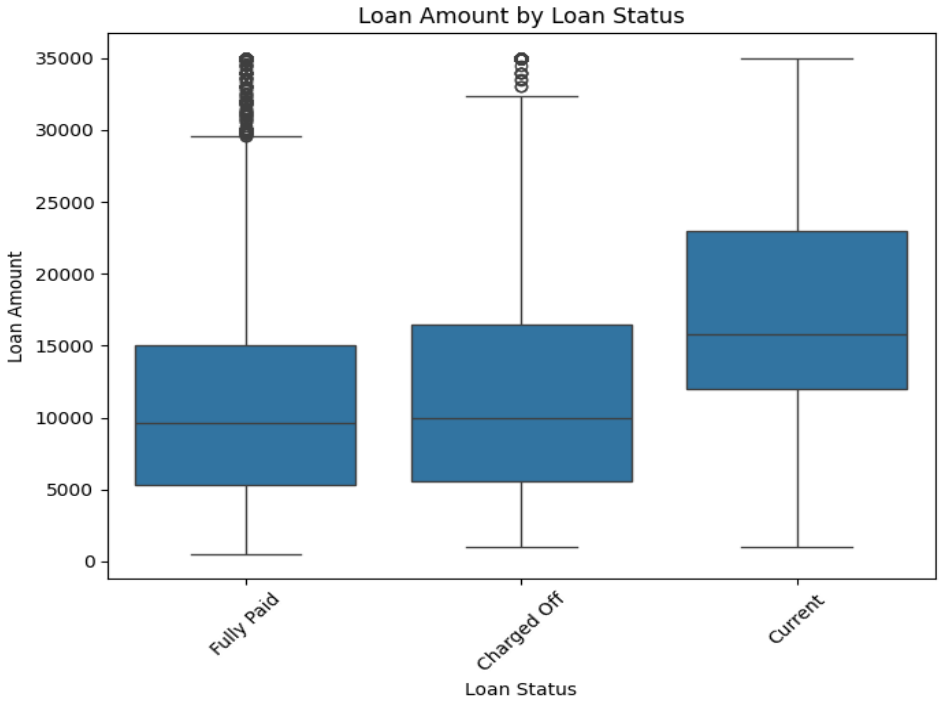
The bivariate analysis is performed using 2 variables at a time to check the impact of one variable on the other variable.

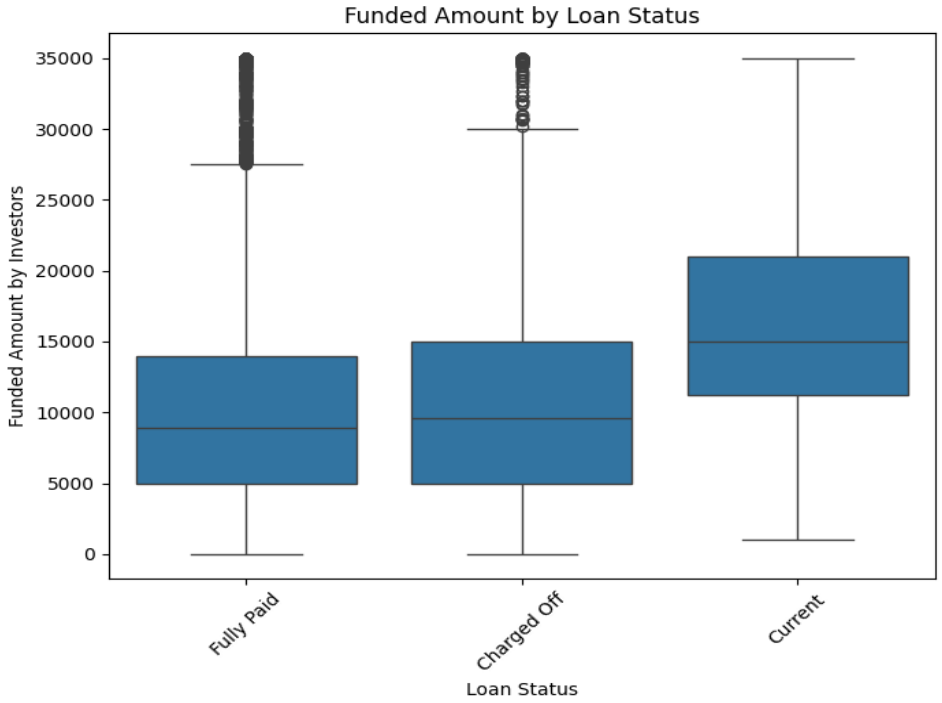
Bivariate analysis on the following pairs of variables was done:

- Loan Amount and Funded Amount by Loan Status
- Grade and Sub-Grade by Loan Status
- Home Ownership by Loan Status.
- Annual Income by Loan Status
- Debt To Income Ratio (DTI) by Loan Status
- Inquiries in last 6 months by Loan Status
- Public Records/Bankruptcyby Loan Status
- Employment Length by Loan Status
- Delinquency by Loan Status
- income by loan status (Segmeneted)
- Earliest Credit Line Age by Loan Status
- employment length by loan status
- Revolving Utilization Rate by Loan Status
- Loan Term by Loan Status
- Months since last delinq vs. Loan Status
- Open/Total Accounts by Loan Status
- verification status by Loan Status
- purpose by Loan Status
- State by Loan Status

Stage 2 – Data Analysis

Loan Amount/Funded Amount by Loan Status



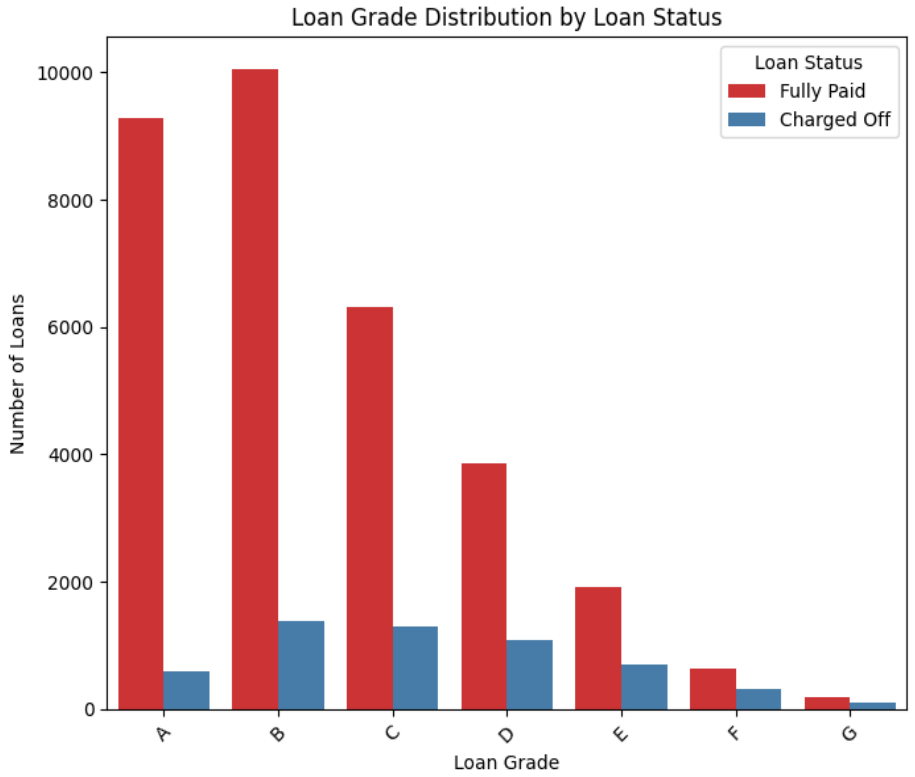


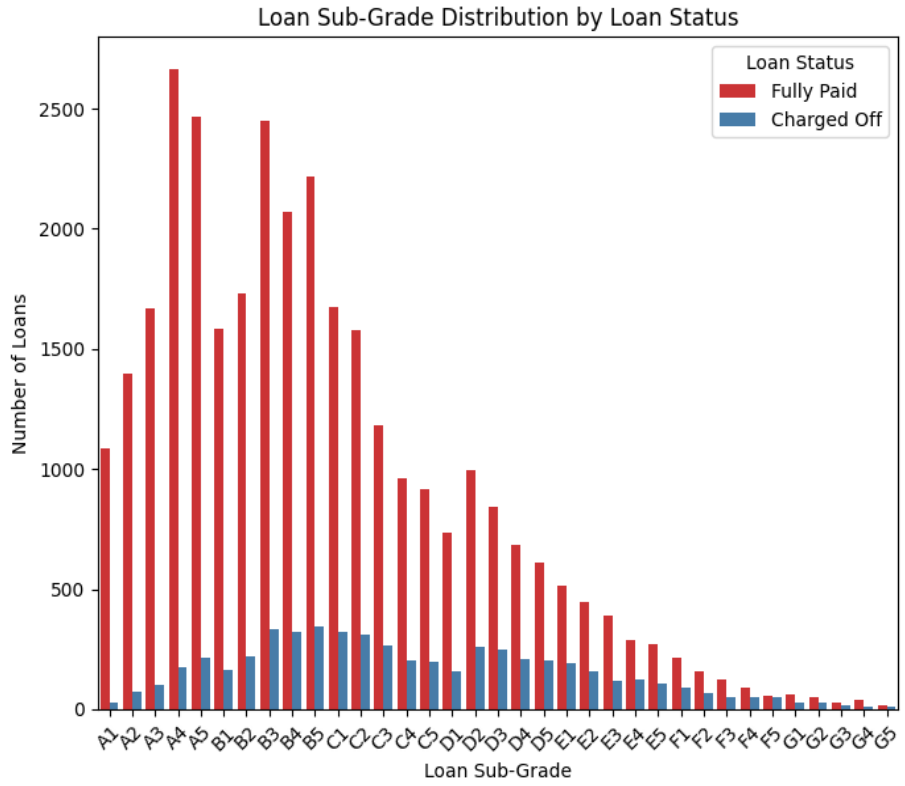
Results of Loan Amount and Funded Amount by Loan Status

- Graphs show similar distributions across all loan statuses (Fully Paid, Charged Off, Current, etc.), it suggests that the size of the loan or the amount funded by investors is not a key driver of default risk in this dataset.

Stage 2 – Data Analysis

Grade/Sub-Grade by Loan Status





- **Results of Grade and Sub-Grade by Loan Status**
- The plots indicate that lower loan grades (D, E, F, and G) have a significantly higher proportion of "Charged Off" loans compared to higher grades (A, B, C), which are mostly "Fully Paid." Similarly, within each grade, lower sub-grades (e.g., C5, D5) show a higher likelihood of default. This suggests that both lower grades and sub-grades are strong indicators of increased default risk, highlighting the importance of these factors in assessing loan risk.
- For risk management, loans in the lower grades (D and below) and lower sub-grades (e.g., C5, D5, etc.) should be treated with greater caution, possibly requiring higher interest rates or stricter approval criteria to mitigate the increased risk of default.

Grade distribution

A 0.060069 B 0.121329 C 0.171324

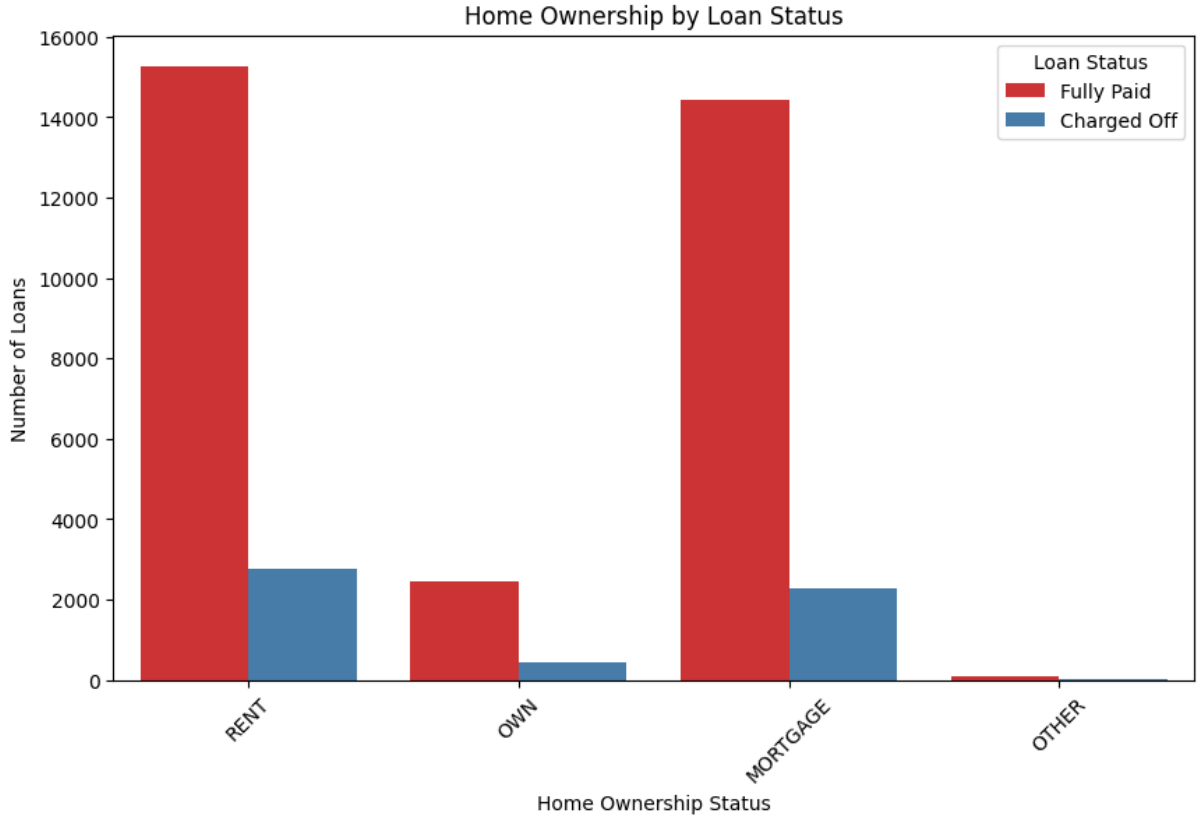
D 0.218182 E 0.267994 F 0.324948

G 0.335593

Name: Charged Off, dtype: float64

Stage 2 – Data Analysis

Home Ownership by Loan Status



home_ownership distribution

MORTGAGE 0.135643

OTHER 0.189474

OWN 0.146662

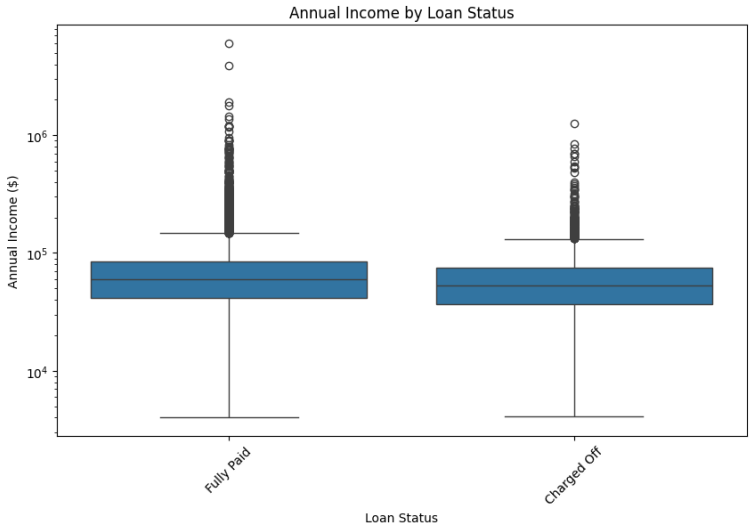
RENT 0.153322

Name: Charged Off, dtype: float64

- **Results of Grade and Sub-Grade by Loan Status**
- The plots indicate that lower loan grades (D, E, F, and G) have a significantly higher proportion of "Charged Off" loans compared to higher grades (A, B, C), which are mostly "Fully Paid." Similarly, within each grade, lower sub-grades (e.g., C5, D5) show a higher likelihood of default. This suggests that both lower grades and sub-grades are strong indicators of increased default risk, highlighting the importance of these factors in assessing loan risk.
- For risk management, loans in the lower grades (D and below) and lower sub-grades (e.g., C5, D5, etc.) should be treated with greater caution, possibly requiring higher interest rates or stricter approval criteria to mitigate the increased risk of default.

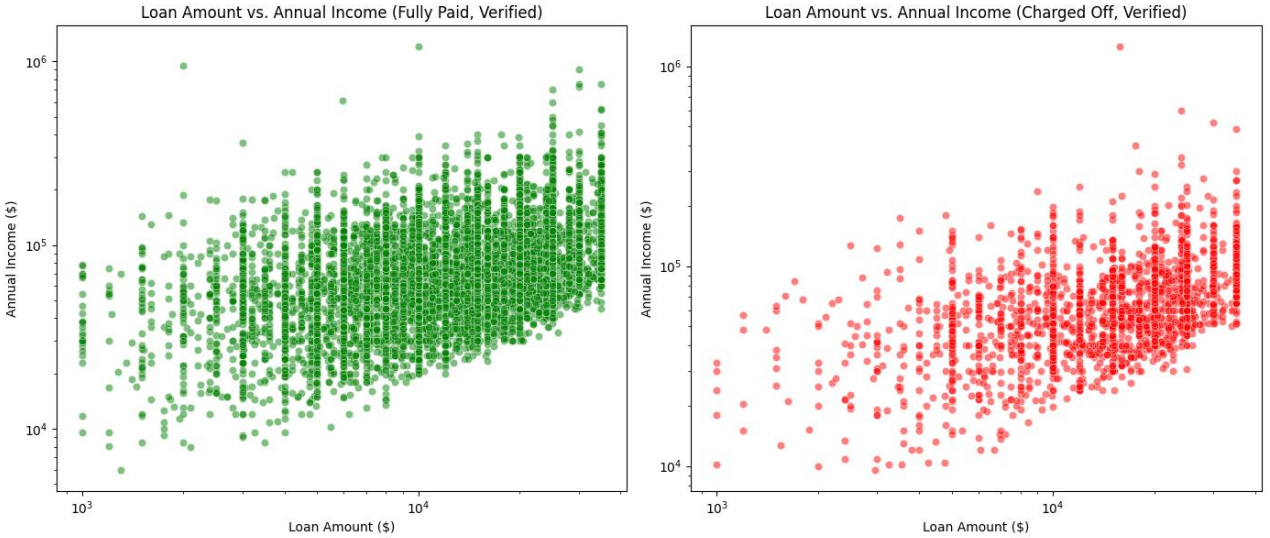
Stage 2 – Data Analysis

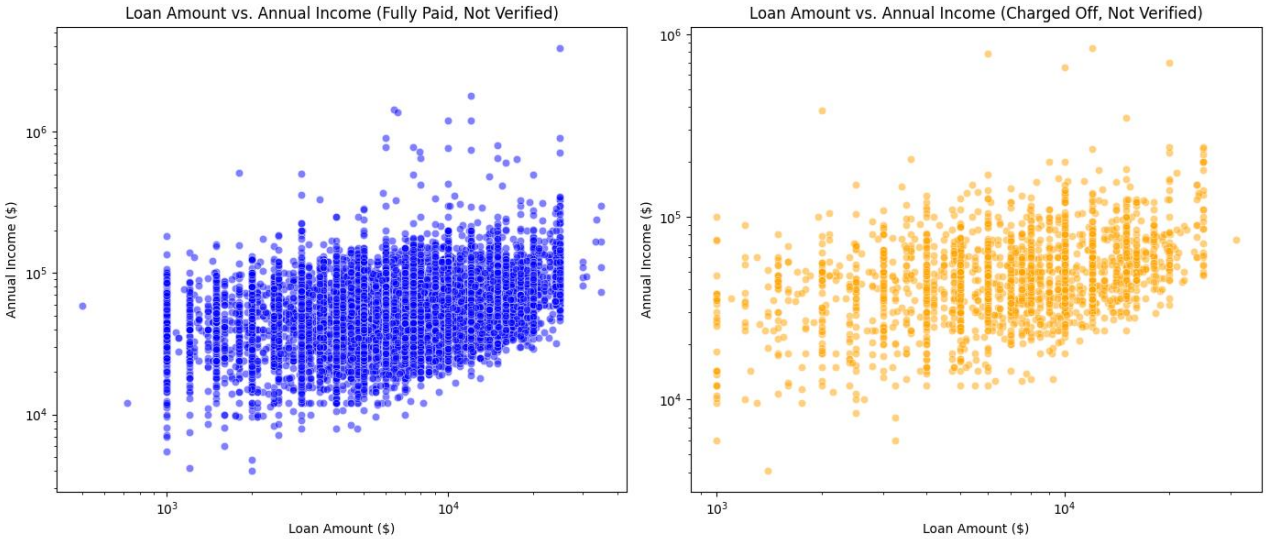
Annual Income by Loan Status



Results of Annual income, loan amount by Loan Status.

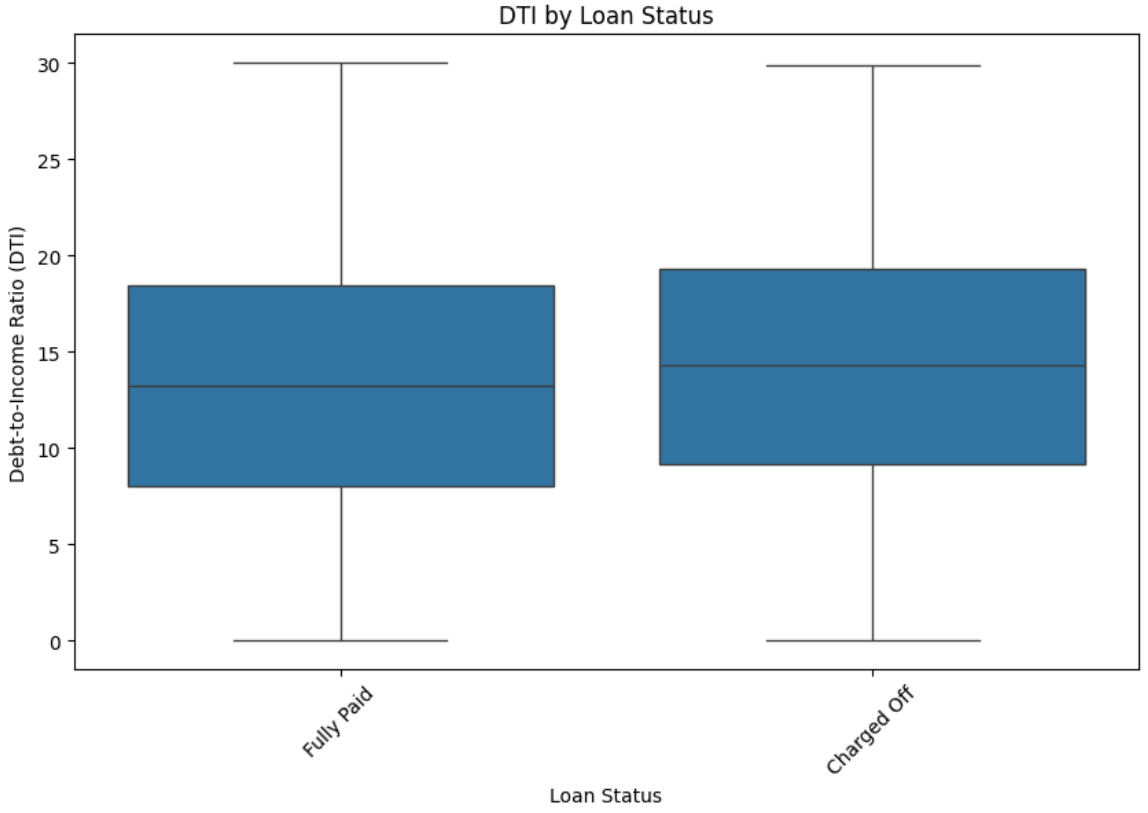
- The plot shows a dense concentration of data points across a wide range of loan amounts and annual incomes, with both "Fully Paid" and "Charged Off" loans spread throughout.
- This suggests that neither loan amount nor annual income alone is a strong predictor of loan default when visualized together like this





Stage 2 – Data Analysis

Debt To Income Ratio (DTI) by Loan Status

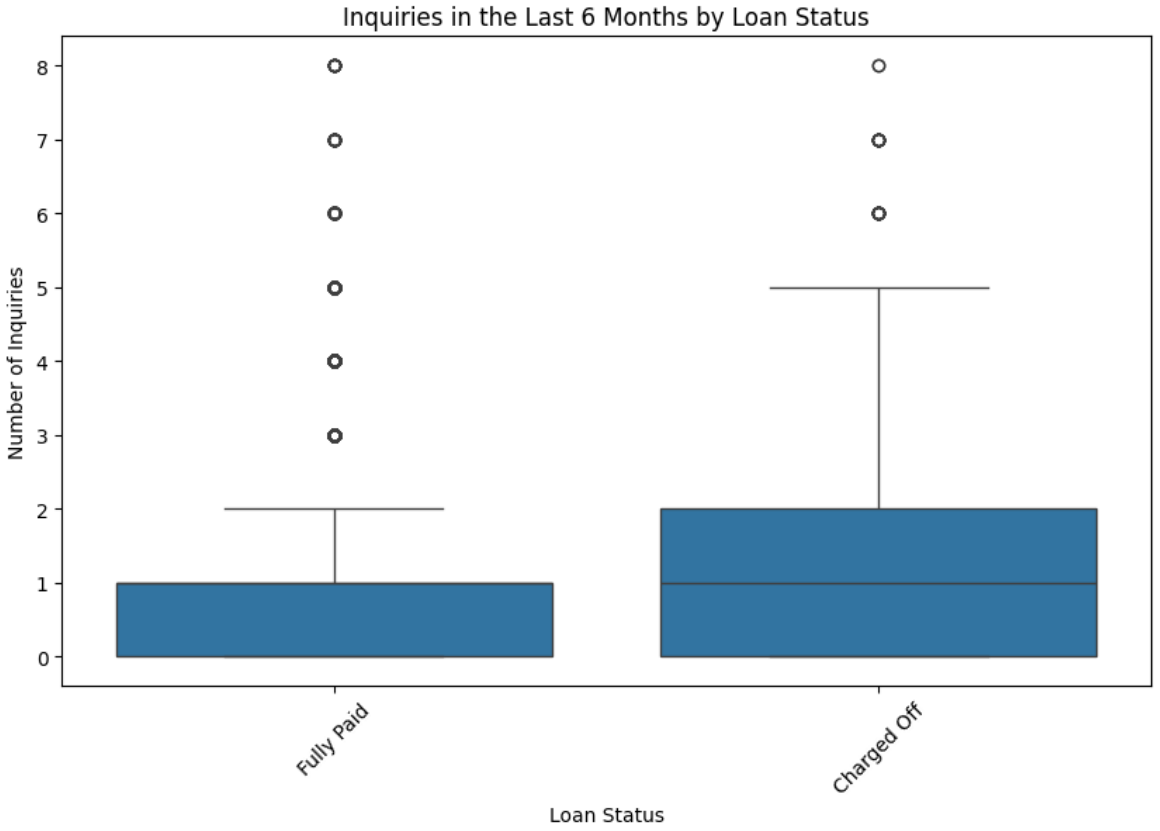


Results of Debt-to-Income Ratio (DTI) by Loan Status.

- The box plots show that the Debt-to-Income (DTI) ratios for both "Fully Paid" and "Charged Off" loans are quite similar.
- Both distributions have a similar median and interquartile range (IQR), with no significant differences observed between the two groups.

Stage 2 – Data Analysis

Inquiries in last 6 months by Loan Status

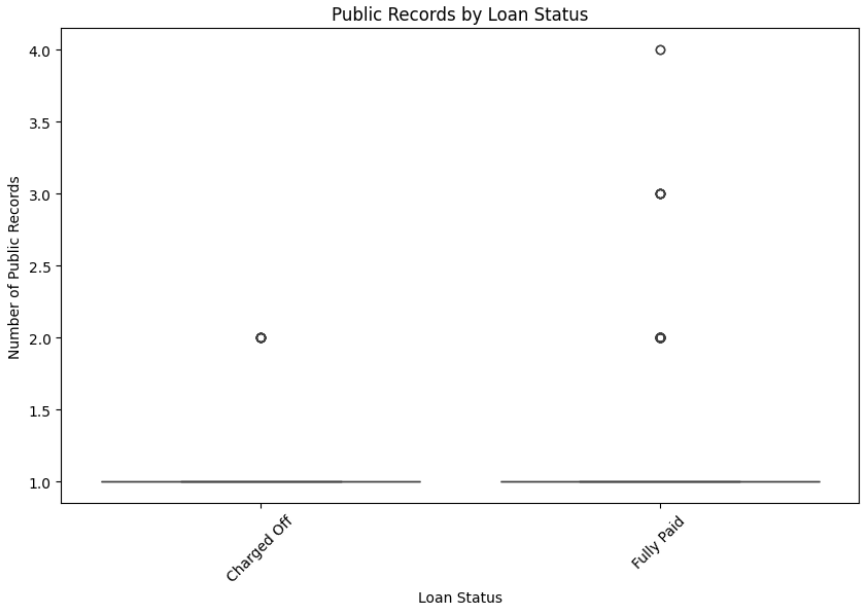


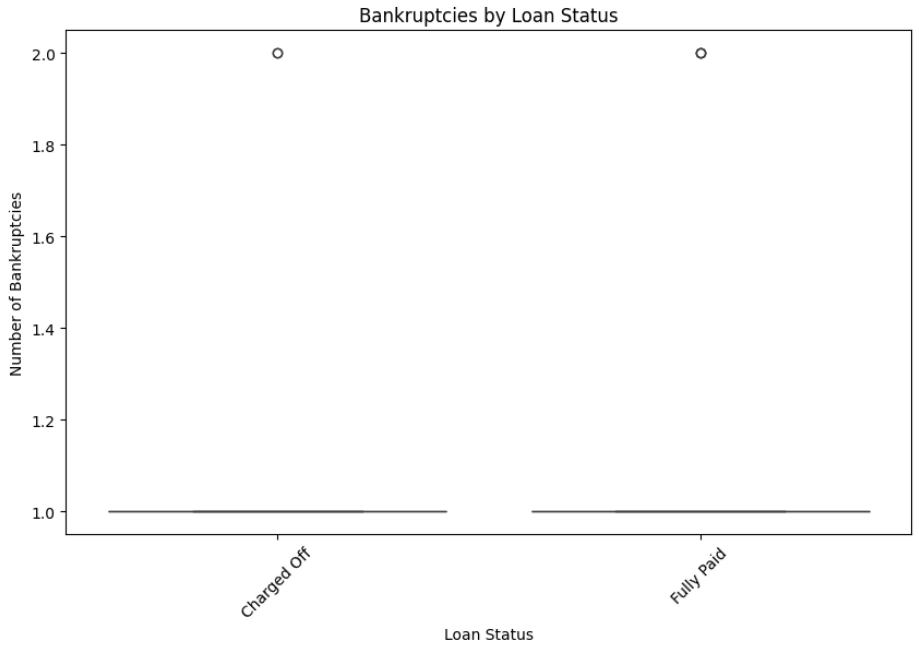
Results of Number of Inquiries in the Last 6 Months by Loan Status.

- The box plot shows that the number of credit inquiries in the last 6 months tends to be slightly higher for borrowers who "Charged Off" compared to those who "Fully Paid" their loans.
- Let's combine this with other variables to do more detailed analysis in multi-variate analysis to figure out the comprehensive risk profile.

Stage 2 – Data Analysis

Public Records/ Bankruptcy by Loan Status



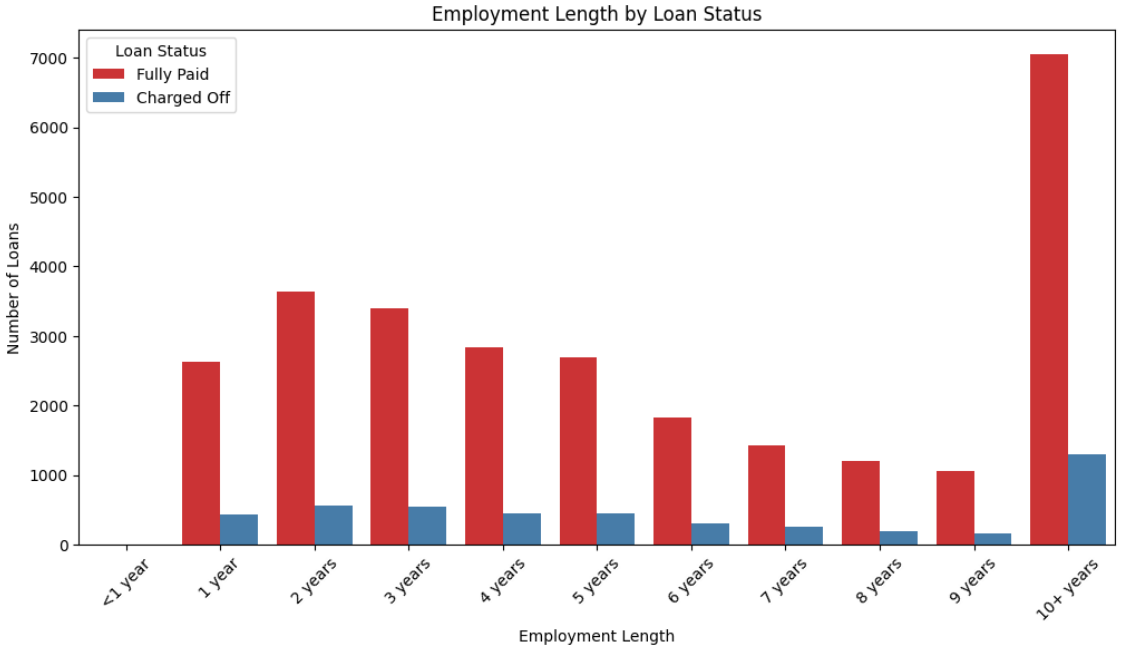


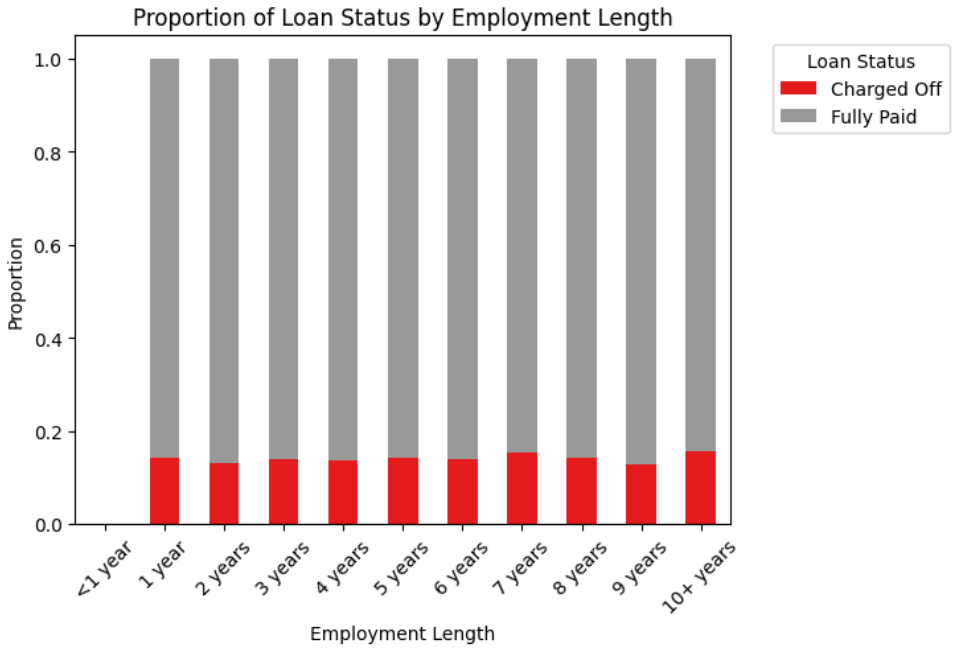
Results of Public Records and Bankruptcies by Loan Status.

- The fact that both "Fully Paid" and "Charged Off" loans are represented equally among those with multiple bankruptcies suggests that having more than one bankruptcy might not be a strong predictor of loan default in this dataset.

Stage 2 – Data Analysis

Employment Length by Loan Status



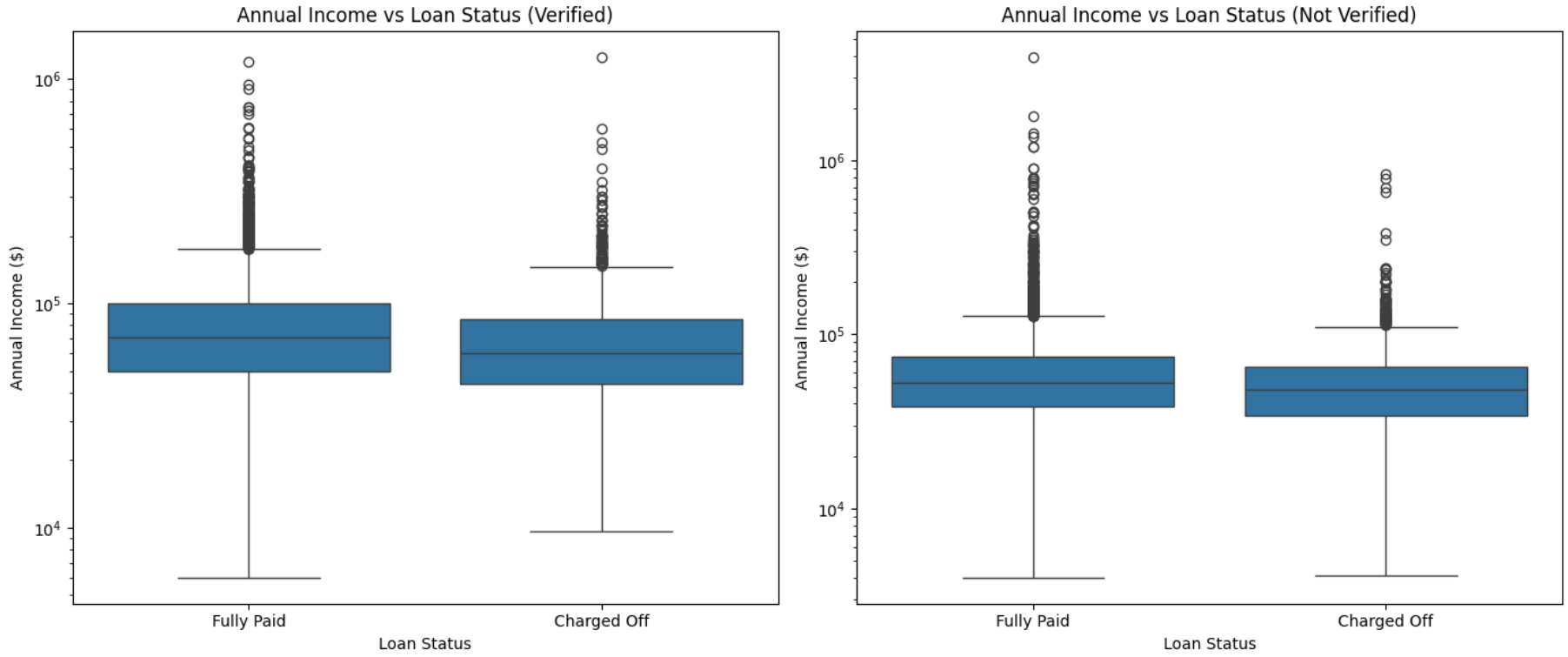


Results of Employment Length by Loan Status.

- Employment length does not appear to be a strong predictor of loan default. The similar proportions across categories indicate that other factors may be more influential in determining whether a borrower will default. In other words, borrowers with different lengths of employment history seem to have similar probabilities of defaulting on their loans

Stage 2 – Data Analysis

Income by Loan Status - Segmented Analysis

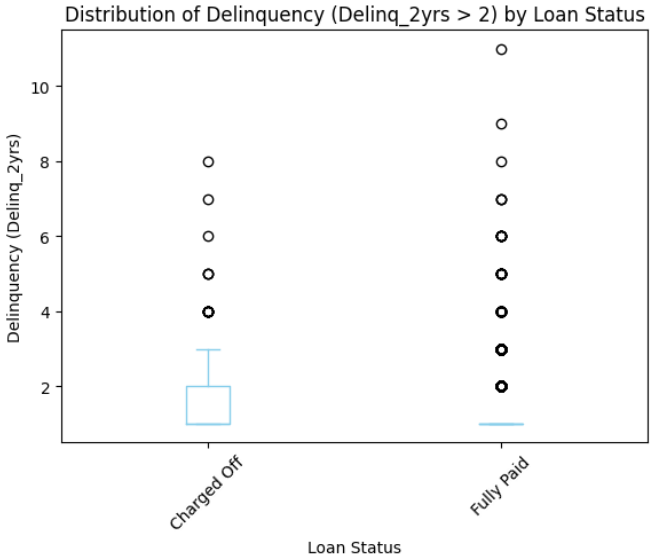


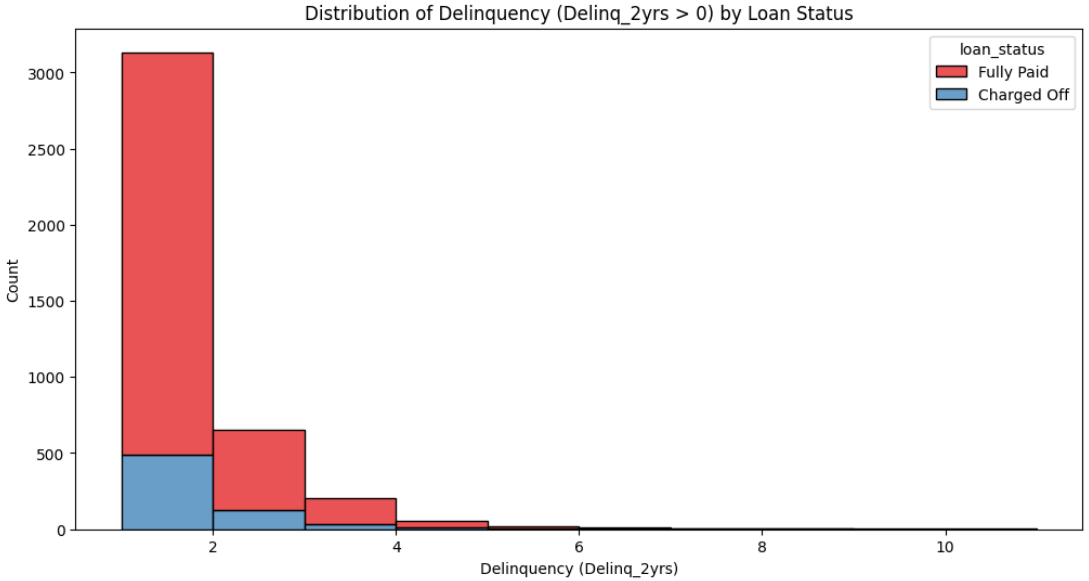
Results of Income by Loan Status – segmented analysis

- Both verified and non-verified income segments show similar patterns

Stage 2 – Data Analysis

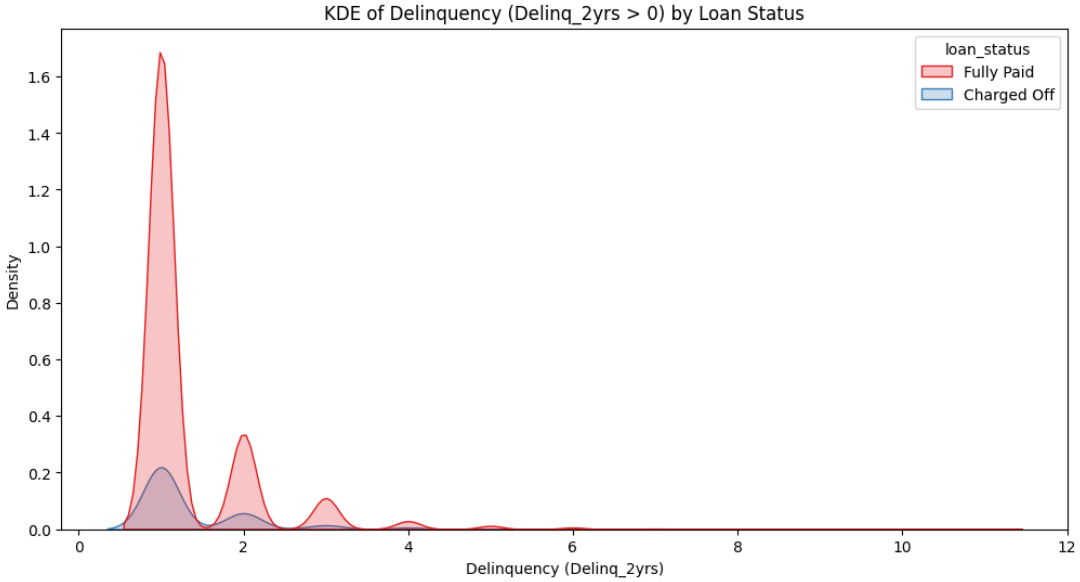
Delinquency by Loan Status





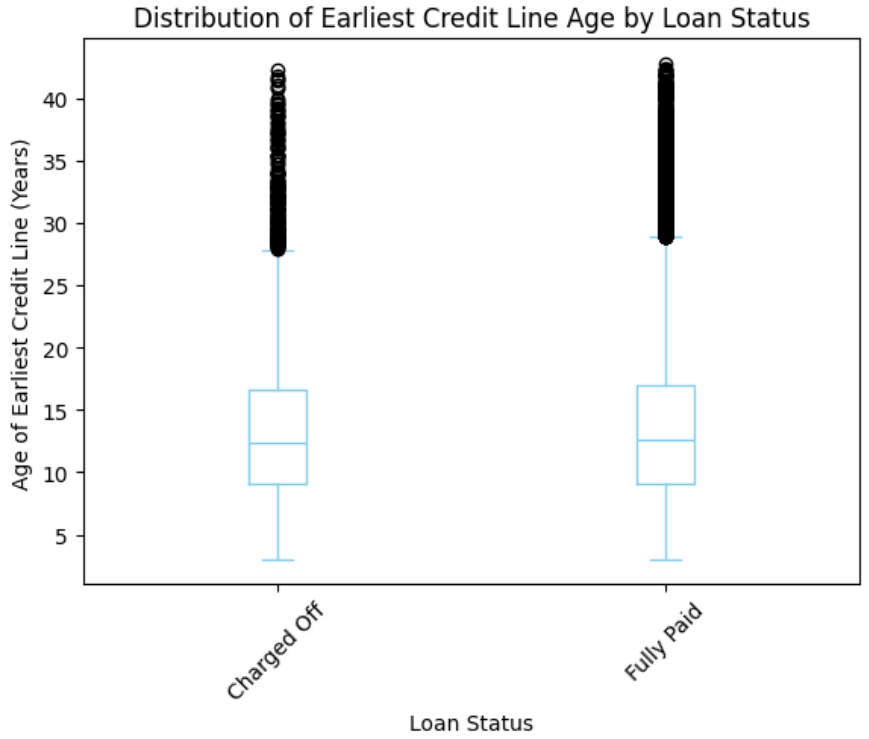
Results Delinquency (delinq_2yrs) by Loan Status with delinquency > 0

- This view focuses on loans that have experienced some level of delinquency, providing a clearer comparison between delinquent loans in different statuses.
- **Higher Delinquency in Charged-Off Loans**: 75th Percentile of Charged-Off Loans: A value of 2 means that 75% of charged-off loans have 2 or fewer delinquencies, but a significant portion may have more.
- **Lower Delinquency in Fully Paid Loans:** 75th Percentile of Fully Paid Loans: A value of 1 means that 75% of fully paid loans have 1 or fewer delinquencies.



Stage 2 – Data Analysis

Earliest Credit Line Age by Loan Status

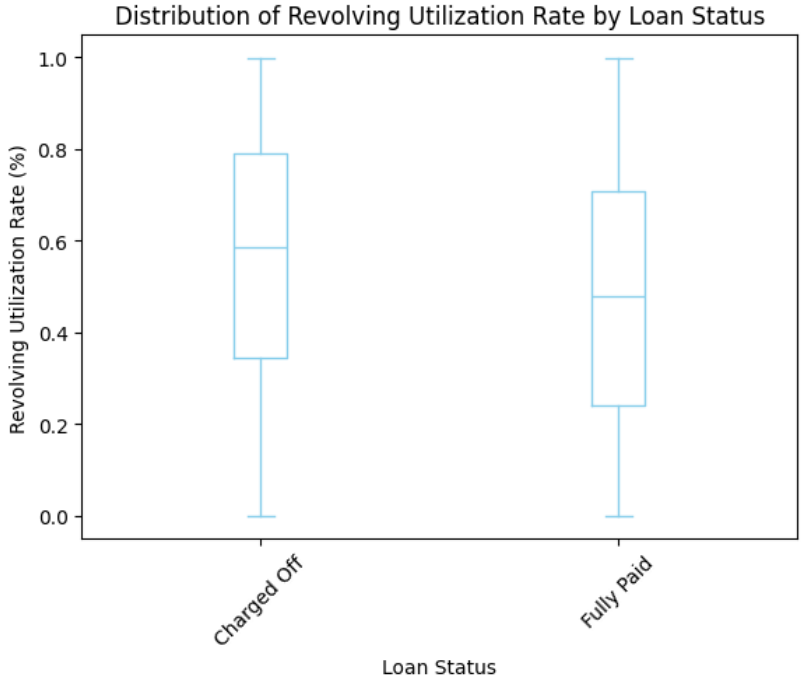


Results of earliest credit line by Loan Status

- Analysis shows that the age of the earliest credit line is similar across different loan statuses and does not reveal any significant patterns

Stage 2 – Data Analysis

Revolving Utilization Rate by Loan Status

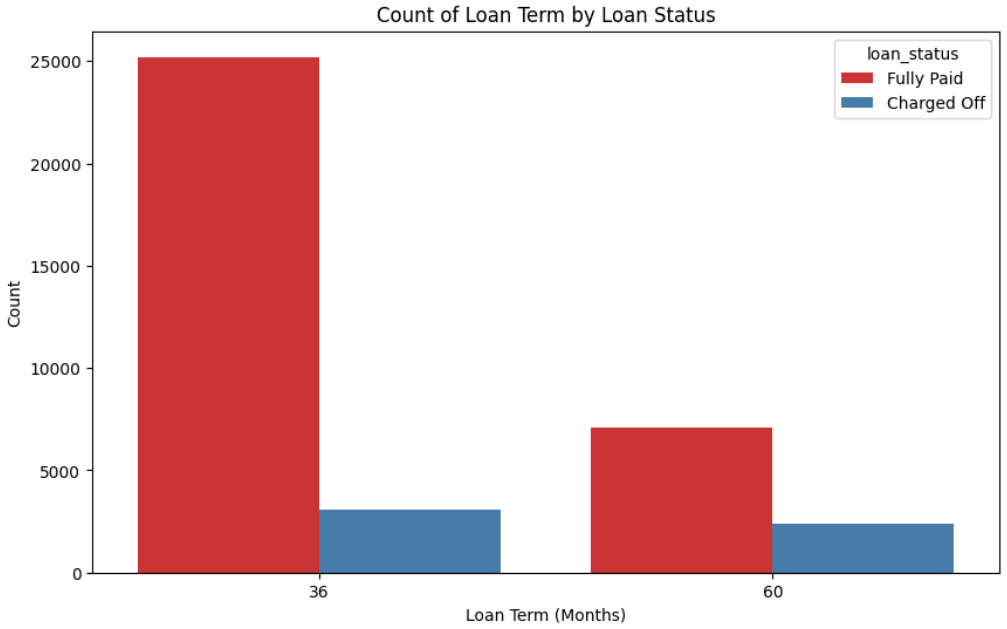


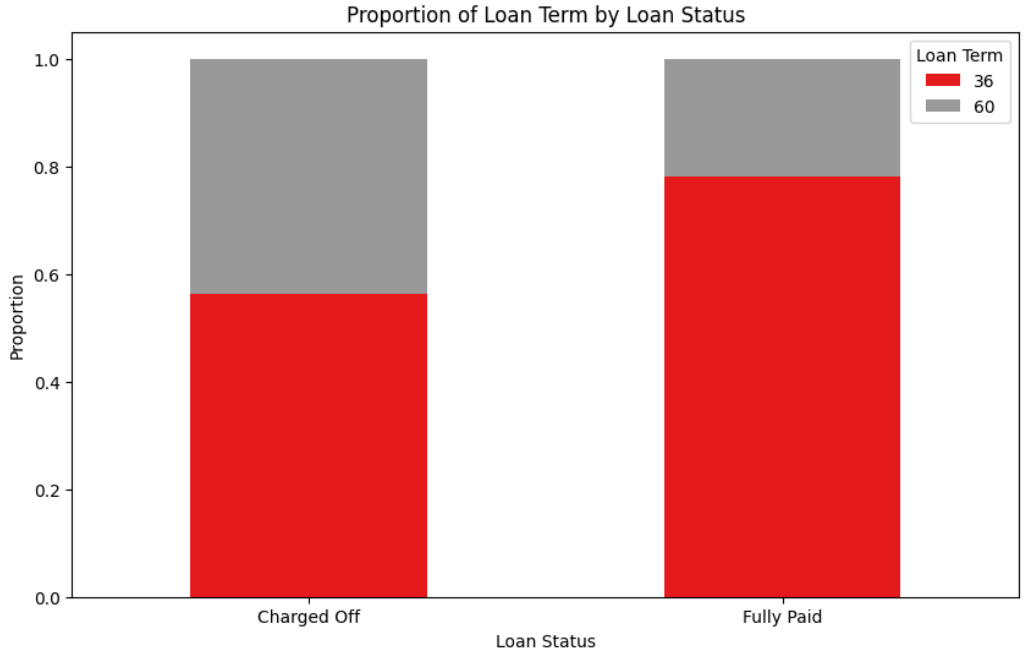
Results of Revolving Utilization Rate by Loan Status

- Charged-off loans have a higher average revolving utilization rate (0.557) compared to fully paid loans (0.477). This suggests that borrowers with higher utilization rates are more likely to have their loans charged off.
- Percentiles: 25th Percentile: Charged-off loans have a higher 25th percentile (0.345) compared to fully paid loans (0.241), indicating that even at lower utilization rates, charged-off loans tend to have higher values. 50th Percentile (Median): Charged-off loans have a median of 0.587 compared to 0.478 for fully paid loans, reinforcing that higher utilization rates are associated with charged-off loans. 75th Percentile: Charged-off loans have a higher 75th percentile (0.791) compared to fully paid loans (0.709), further indicating that high utilization rates are more common in charged-off loans.
- The analysis suggests that borrowers with higher revolving utilization rates are more likely to have their loans charged off. This insight can be useful for credit risk assessment and management

Stage 2 – Data Analysis

Loan Term by Loan Status



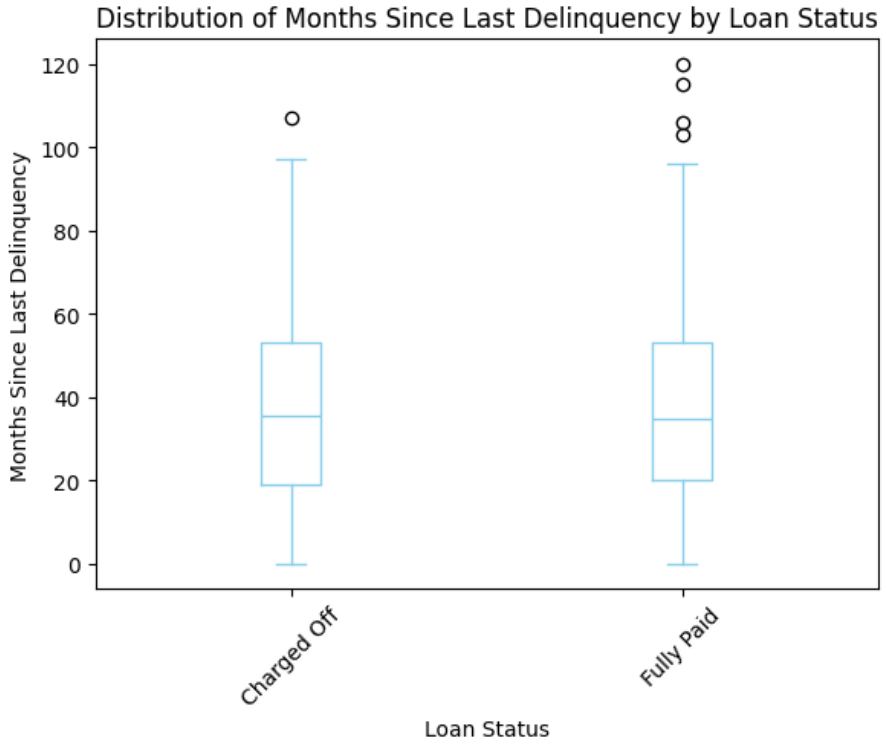


Result of Loan Term by Loan Status

- The analysis reveals that fully paid loans are more likely to have a 36-month term compared to charged-off loans, which are more evenly distributed between the two terms.
- This might suggest that shorter loan terms are less risky or more manageable, leading to a higher rate of full repayment.

Stage 2 – Data Analysis

Months Since Last Delinq by Loan Status

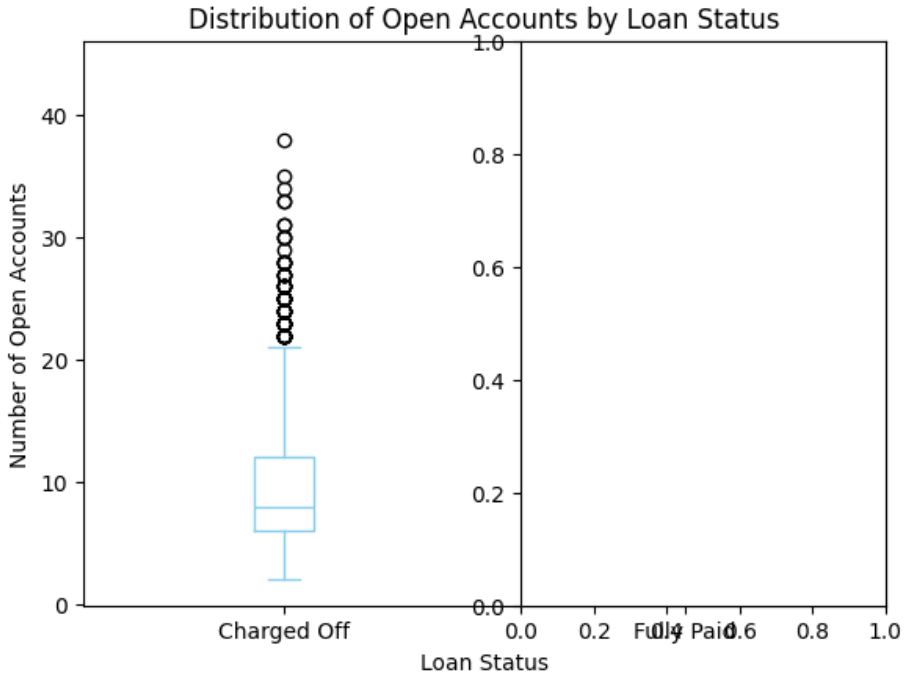


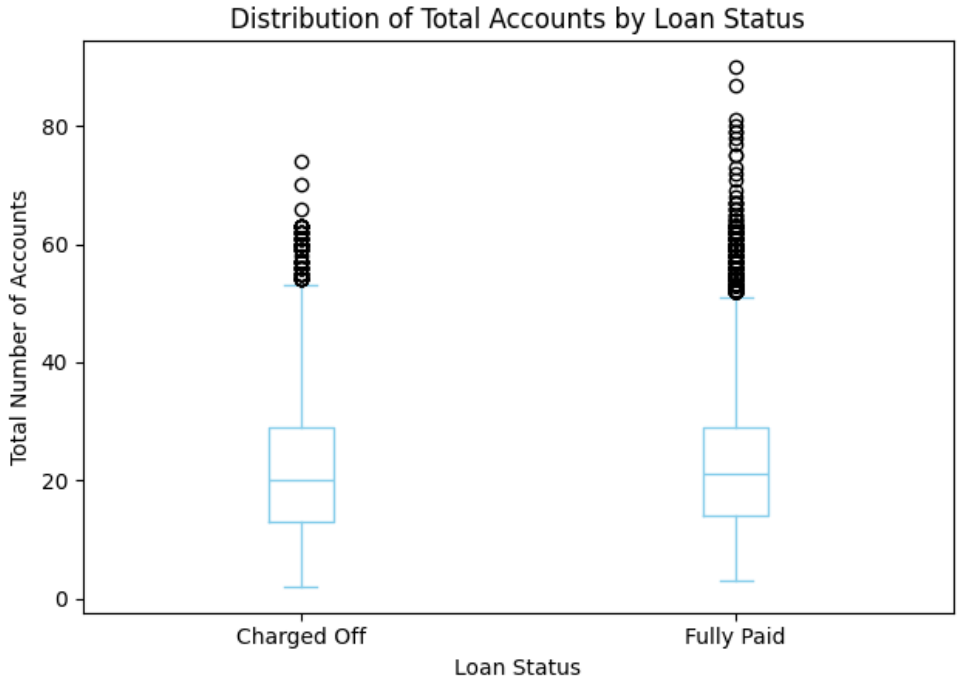
Result of mths_since_last_delinq by Loan Status

- The similar statistical profiles suggest that the timing of the last delinquency does not significantly differentiate between charged-off and fully paid loans.
- This means that, based on this feature alone, there isn't a strong distinction between the two loan statuses.

Stage 2 – Data Analysis

Open / Total Accounts by Loan Status



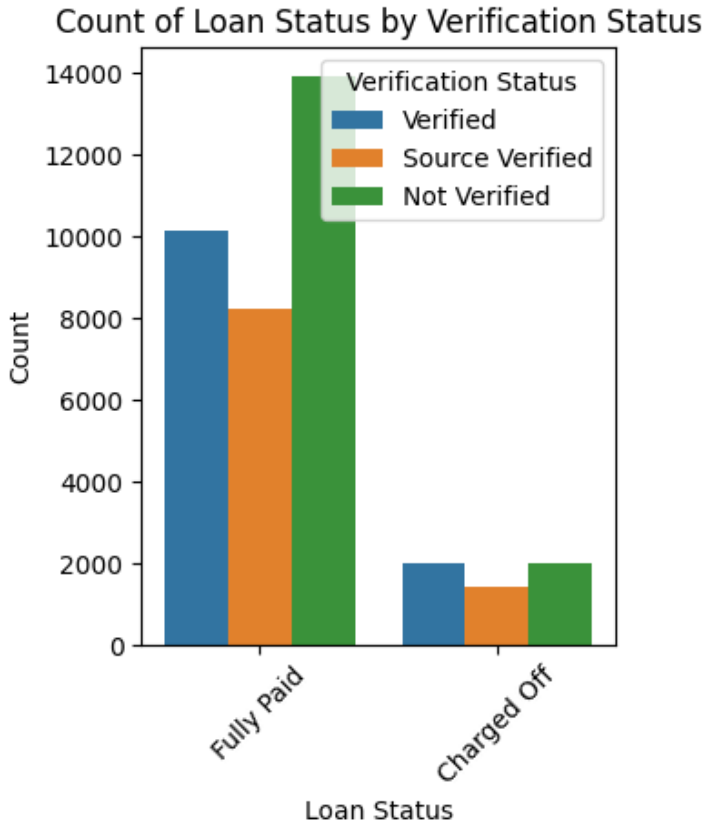


Results of Open Accounts and Total Accounts by Loan Status

- Overall, the statistics for both open_acc and total_acc are comparable between charged-off and fully paid loans.
- The slight differences in mean values and maximum ranges do not suggest a significant distinction in terms of the number of accounts between the two loan statuses.

Stage 2 – Data Analysis

Verification Status by Loan Status

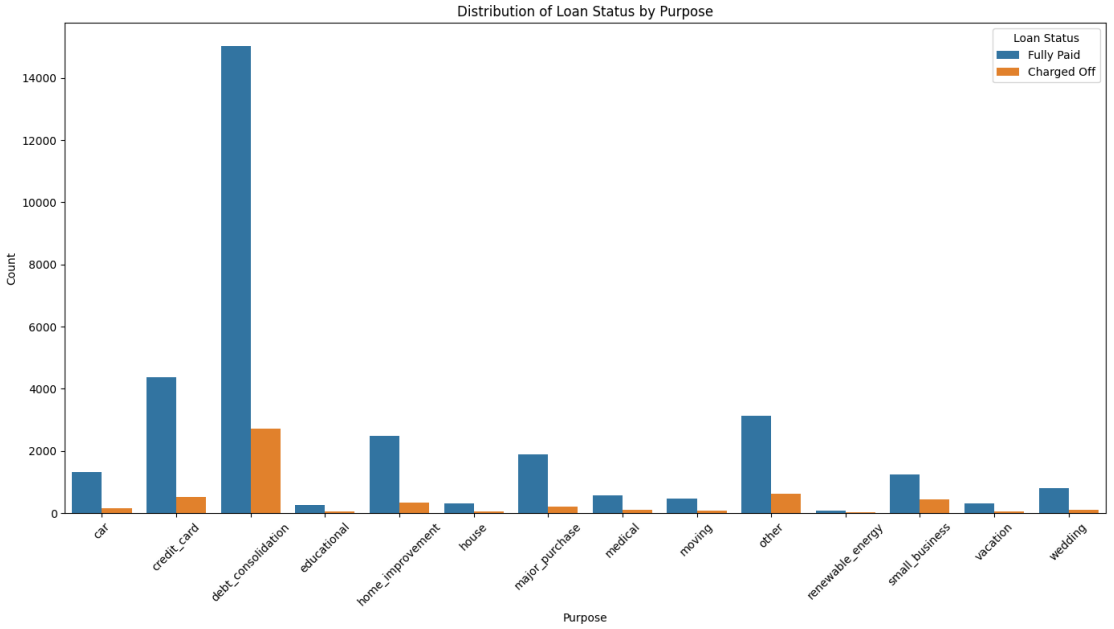


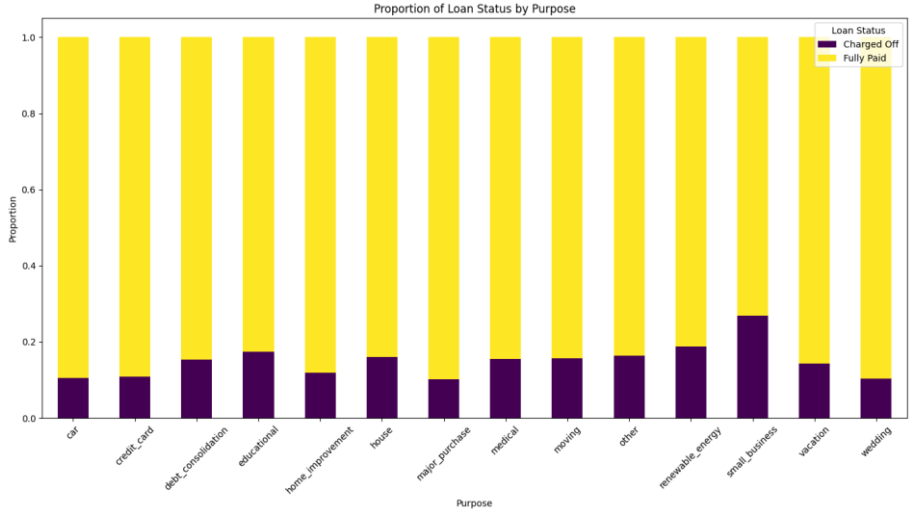
Results: Verification status biverate analysis with loan status

- Across all verification_status categories, "Fully Paid" loans dominate, indicating that verification status does not drastically alter the likelihood of a loan being fully paid compared to other loan statuses.

Stage 2 – Data Analysis

Purpose by Loan Status



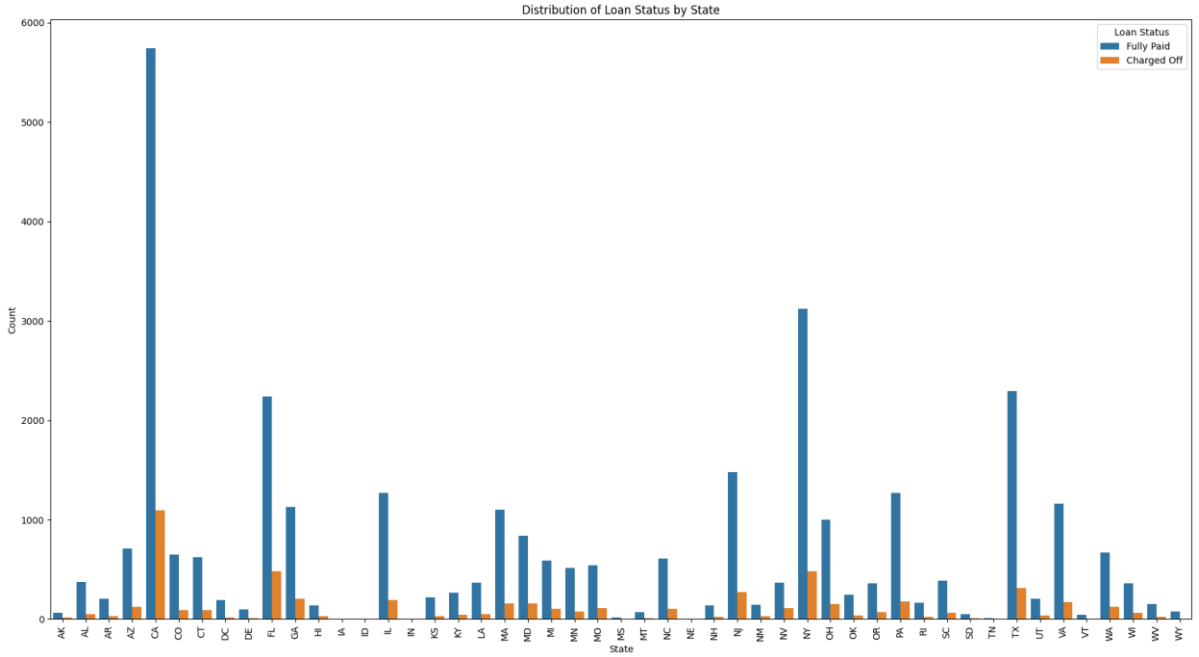


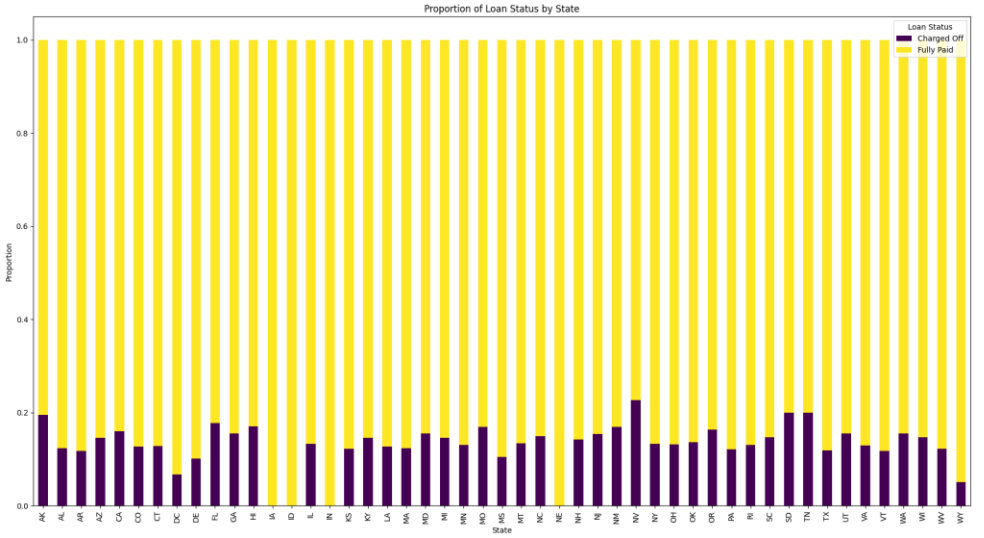
Results: Purpose biverate analysis with loan status

- The plot highlights that while most loans are paid in full regardless of purpose, certain categories like "Small Business" and "Renewable Energy" stand out as higher-risk investments.

Stage 2 – Data Analysis

State by Loan Status





- **Results: addr_state biverate analysis with loan status**
- California (CA), Florida (FL), and New York (NY) have higher counts of charged-off loans, but due to their large populations and economies, they also have high counts of fully paid loans.
- Nevada (NV) and Georgia (GA) show higher proportions of charged-off loans, making them higher-risk states relative to others.
- North Dakota (ND), South Dakota (SD), and Vermont (VT) have very low proportions of charged-off loans, suggesting a safer lending environment.

Stage 2 – Data Analysis

Summary – Bivariate Analysis

The bivariate analysis results can be summarized as below:

- The bivariate analysis reveals key risk factors associated with loan defaults.
- Lower loan grades (D, E, F, and G) and sub-grades (e.g., C5, D5) significantly correlate with higher default rates, highlighting the importance of these factors in assessing loan risk.
- Borrowers with more credit inquiries in the last 6 months and higher revolving utilization rates (average 0.557 for charged-off loans vs. 0.477 for fully paid loans) are more likely to default, underscoring the need to monitor these metrics closely.
- Delinquency history also plays a critical role, with charged-off loans showing higher delinquency counts.
- Shorter loan terms (36 months) are generally safer, as longer terms are more evenly distributed between fully paid and charged-off loans.
- Purpose-based analysis identifies "Small Business" and "Renewable Energy" loans as higher-risk categories, while geographical analysis points to certain states (e.g., NV, GA) as having higher proportions of defaults, indicating a higher risk lending environment in these regions.

Stage 2 – Data Analysis

Derived Metrics Analysis

Based on the results of the bivariate analysis, a derived metric called “Credit Score” was created including the key variables impacting the loan status.

Each variable has been standardized based on the patterns found in previous analysis.

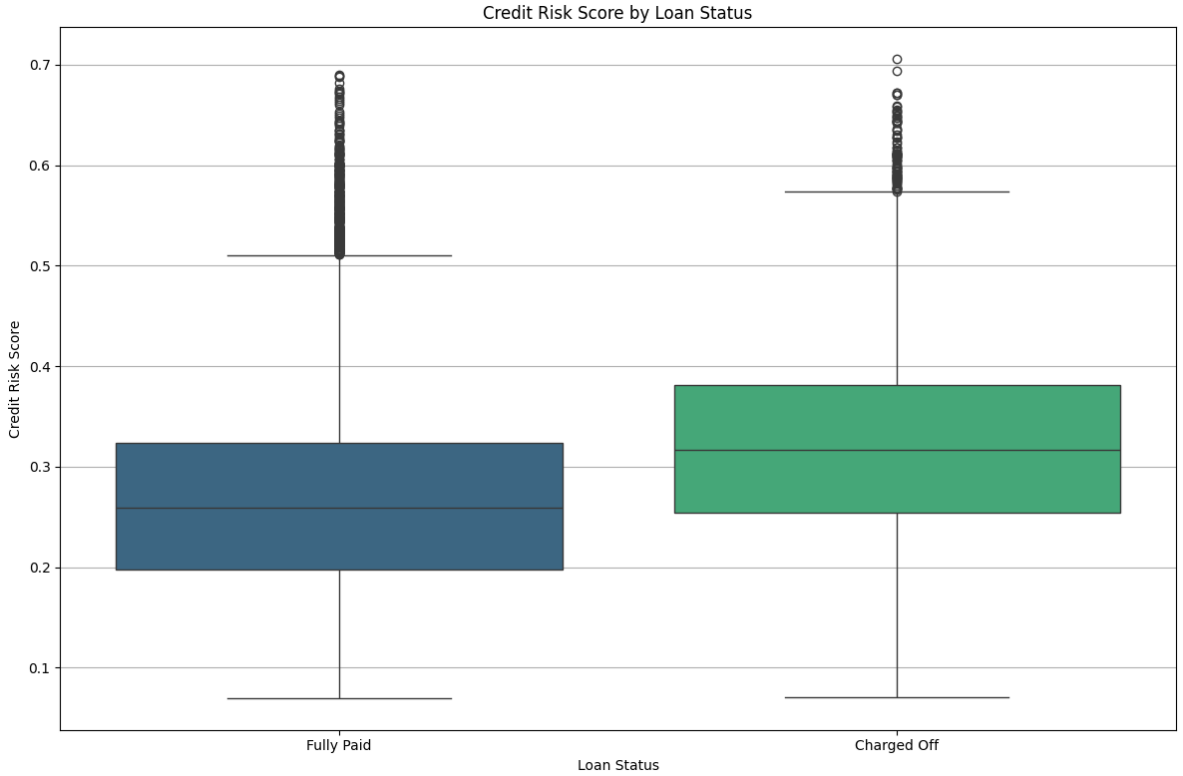
Each standardized variable was given a certain weightage based on the level of impact of that variable on the loan status.

The standardized variables along with the weightages used to create credit score are summarized below:

- 'Standardized_Grade': 0.10, # Slightly reduced to balance with other factors
- 'Standardized_Inquiries': 0.10, # Further reduced to emphasize stronger indicators
- 'Standardized_Delinquencies': 0.20, # Increased to reflect the importance of delinquency history
- 'Standardized_Utilization': 0.20, # Maintained as it correlates with financial stress
- 'Standardized_Loan_Term': 0.05, # Kept consistent as it has some influence
- 'Standardized_Purpose_Risk': 0.30, # Increased further to emphasize purpose
- 'Standardized_State_Risk': 0.05 # Reduced to minimal to focus on stronger predictors

Stage 2 – Data Analysis

Credit Score by Loan Status



Results of Credit Score by Loan Status multivariate analysis

1. The median credit risk score for "Charged Off" loans remains distinctly higher than that of "Fully Paid" loans, reinforcing the score's effectiveness in identifying higher-risk borrowers.
2. The interquartile range (IQR) for "Charged Off" loans still sits above that for "Fully Paid" loans, which is a positive indicator that the score captures the risk gradient effectively.
3. These results indicate that the further fine-tuning of weights has improved the performance of the Credit Risk Score. The adjustments have made the score more effective at distinguishing between borrowers who are likely to pay off their loans versus those who are at higher risk of default.

Stage 3 – Conclusion

What We Did:

We used various data points—like loan grades, borrower behavior, and loan purposes—to figure out what makes someone more likely to default on their loan. We built a Credit Risk Score that combines these factors and tells us how risky each loan is.

Key Findings:

- Loan Grades Matter: Loans with lower grades (D, E, F, G) had a much higher chance of defaulting compared to higher grades (A, B, C).
- High Utilization Rates Are Risky: Borrowers using a lot of their available credit were more likely to default.
- Loan Purpose is Key: Loans for things like Small Business and Renewable Energy showed higher default rates, so they're riskier.
- Past Behavior Predicts Future Risk: Borrowers with past delinquencies were more likely to default again, which makes sense—past behavior is a strong indicator of future actions.
- Location Matters: Some states like Nevada and Florida showed higher default rates, meaning geographical location plays a role in risk.

Credit Risk Score

We combined all these factors into a Credit Risk Score, fine-tuning the weights of each factor to reflect their impact on default risk.

The score effectively separated risky loans from safer ones, helping us identify which loans need stricter approval criteria or higher interest rates

Recommendations:

- Use the Score in Decision-Making: The Credit Risk Score can help the company decide which loans to approve, adjust loan terms, or set interest rates based on risk levels.
- Focus on High-Risk Areas: Be extra cautious with loans for risky purposes or in high-default states.
- Keep Updating the Score: Regularly check and adjust the Credit Risk Score to keep it accurate as market conditions change.