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

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

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

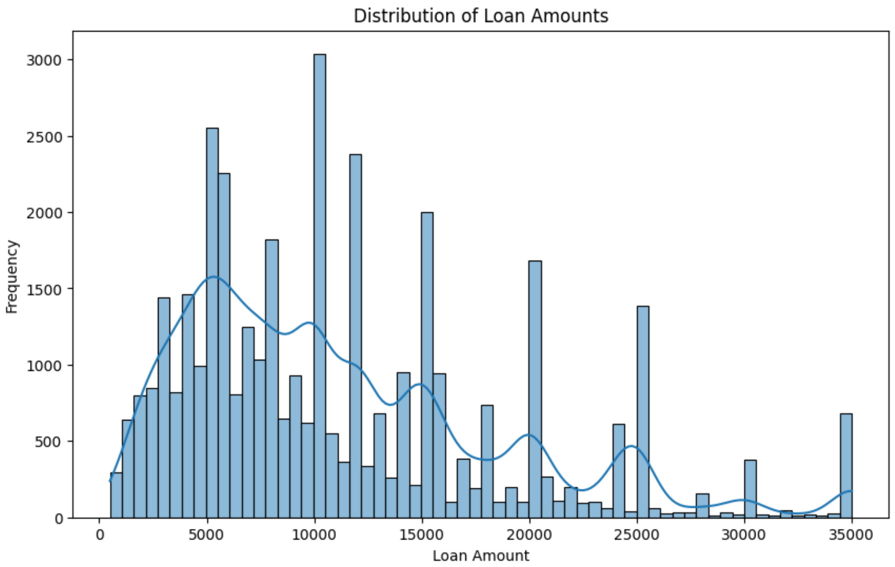
Rationale:

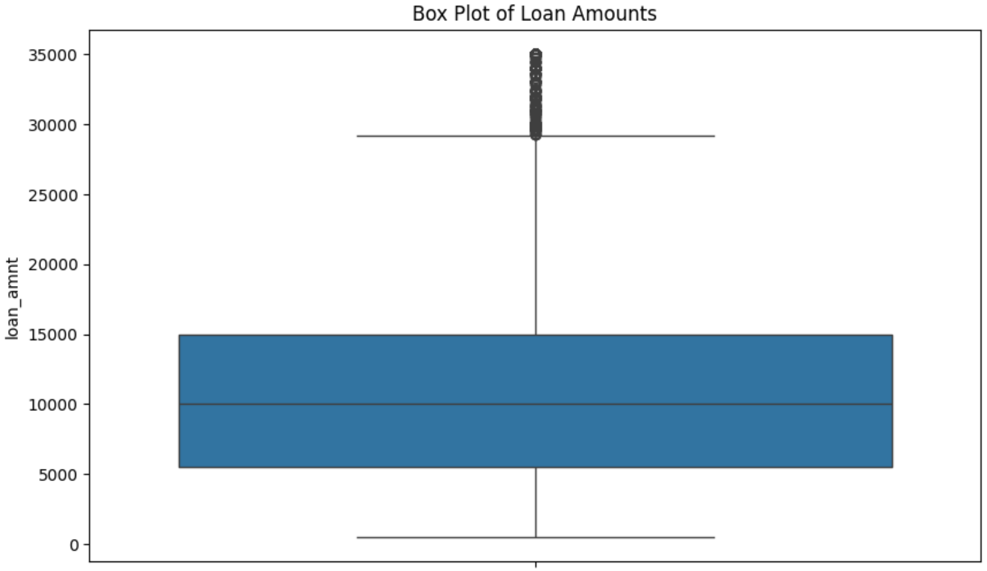
- These variables are dependent on the original loan terms and payment history, which have been analyzed through their primary variables.
- Any outliers or unusual distributions in these dependent variables would already be reflected in the primary variables.
- Skipping detailed univariate analysis for 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





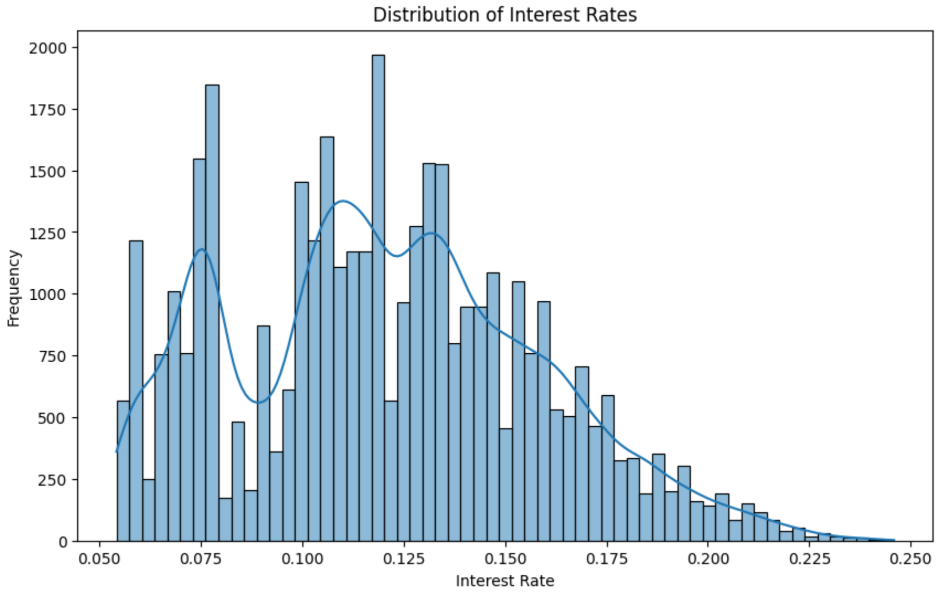
Results of Loan_amnt univerate analysis

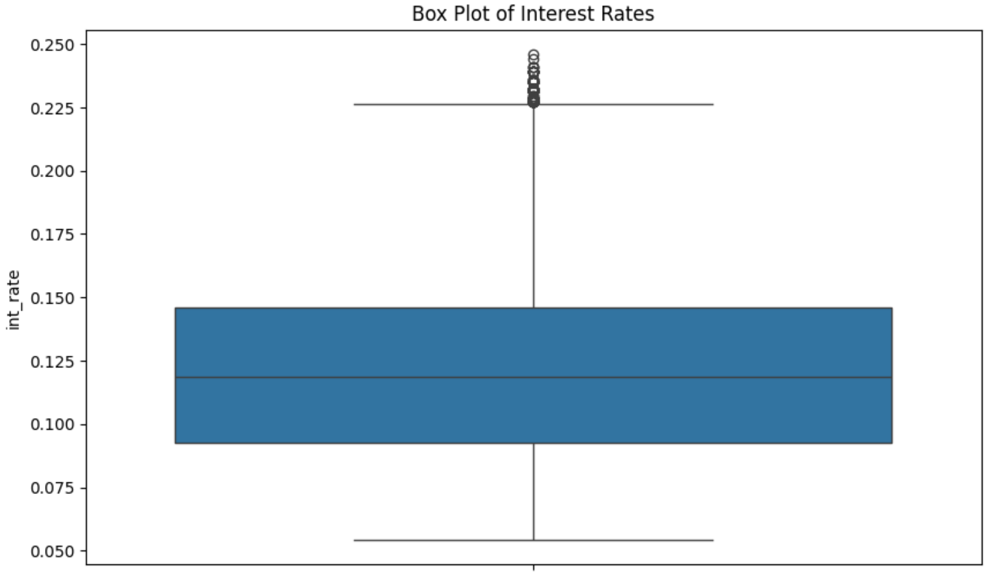
- The loan amounts range from 500 to 35,000, with a mean of approximately 11,248.
- The median loan amount is 10,000, indicating a balanced distribution around this value.
- The standard deviation of 7,470 suggests moderate variability in loan amounts.
- The interquartile range (IQR) is 9,500, showing 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


Distribution of Interest Rates


Box Plot of Interest Rates

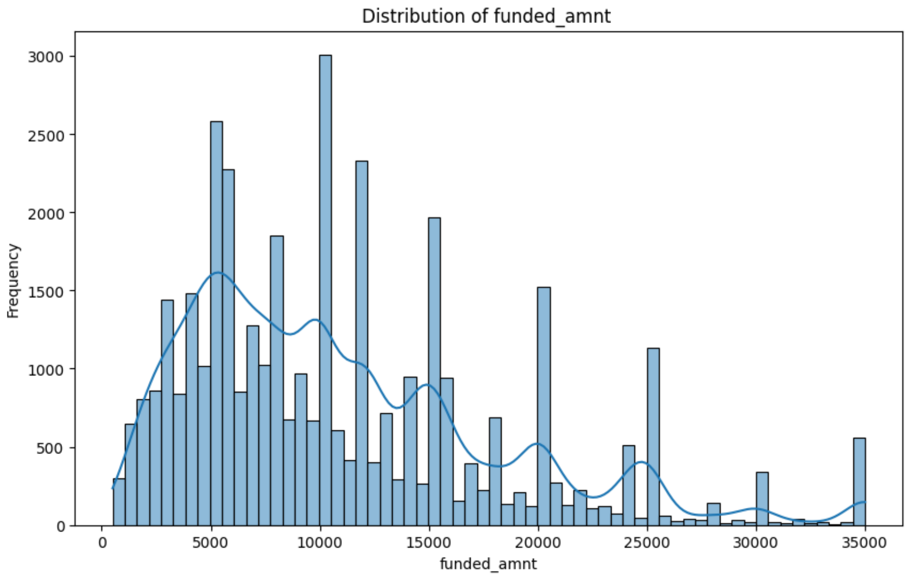
Results of int_rate univerate analysis

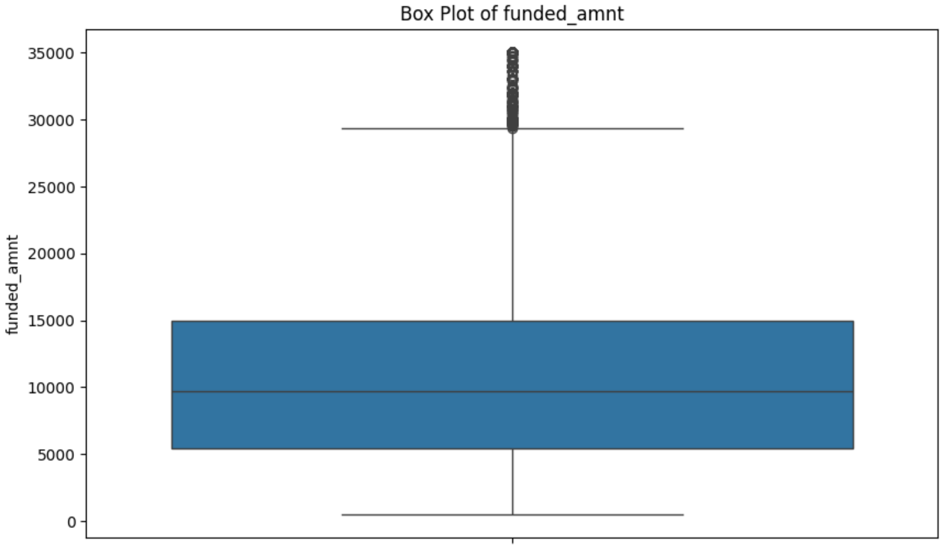
- The interest rates range from 5.42% to 24.59%, with a mean of approximately 12.04%.
- The median interest rate is 11.86%, indicating a slightly lower concentration of interest rates around this value.
- The standard deviation of 3.74% suggests a moderate variability in interest rates.
- The interquartile 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.

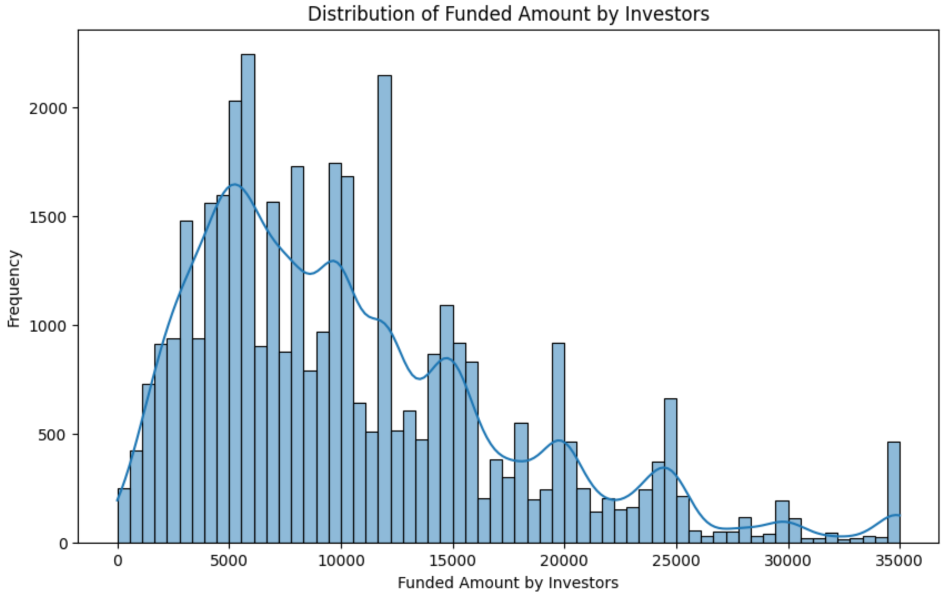
Conclusion: All data should be included in the analysis as the distribution appears normal and reflects the typical range of interest rates offered to borrowers.

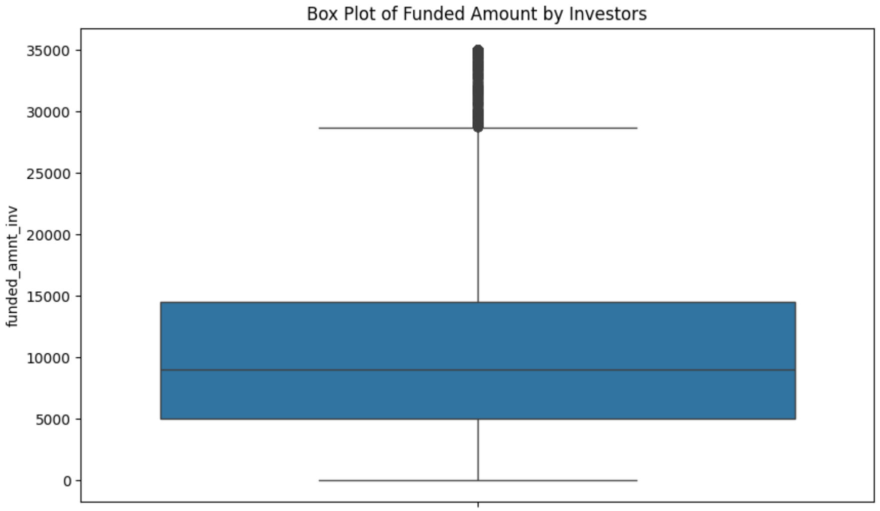
Stage 2 – Data Analysis

Funded Amount & Funded Amount By Investors


Distribution of funded_amnt


Box Plot of funded_amnt

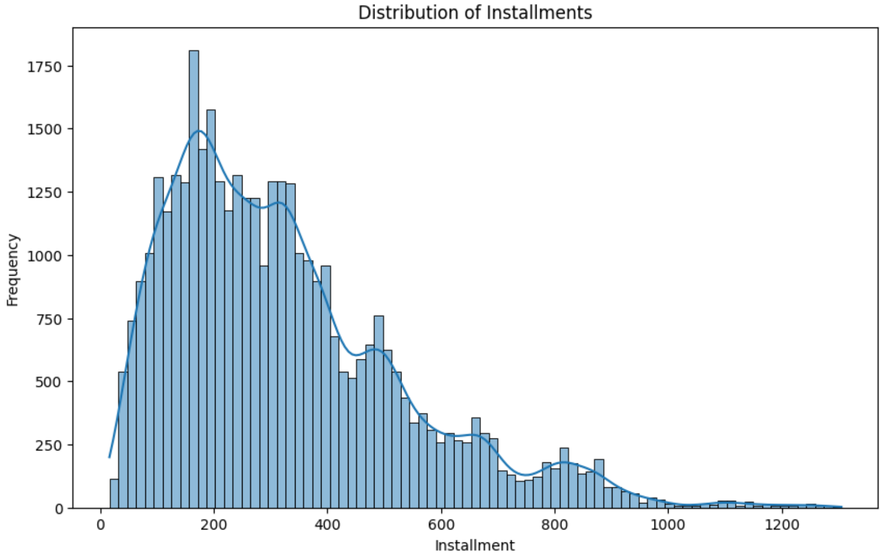

Distribution of Funded Amount by Investors

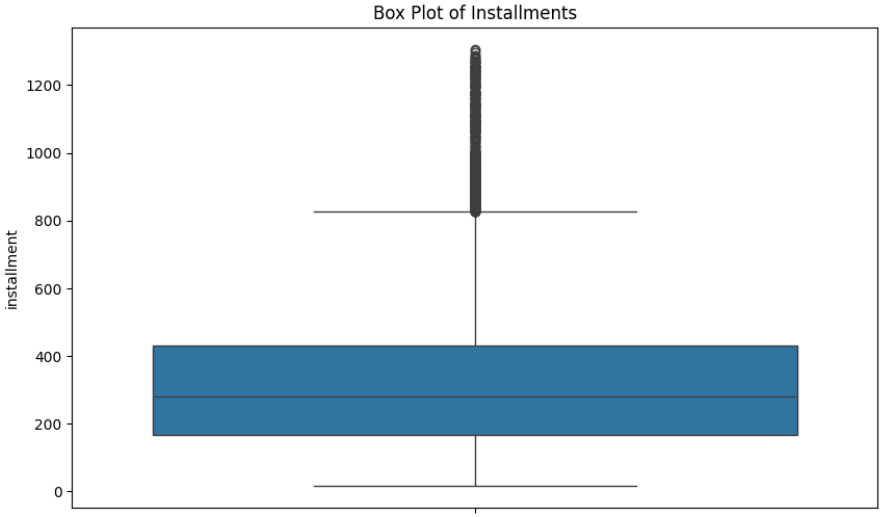

Box Plot of Funded Amount by Investors

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

Stage 2 – Data Analysis

Installment



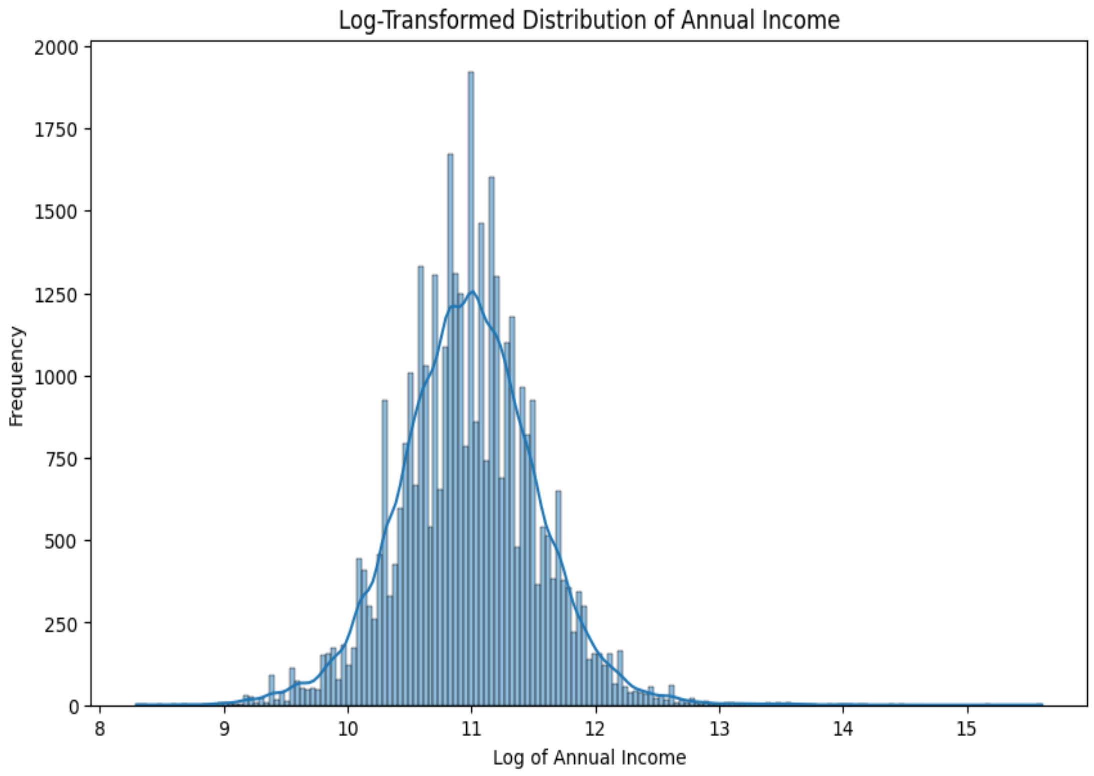


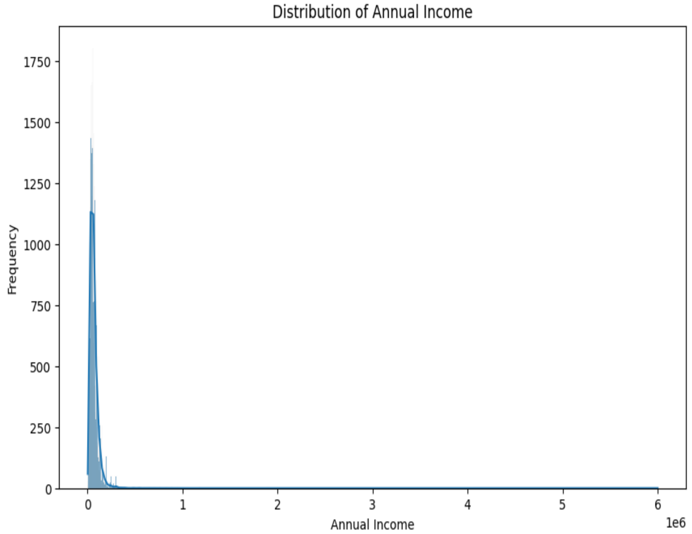
Results of int_rate univerate analysis

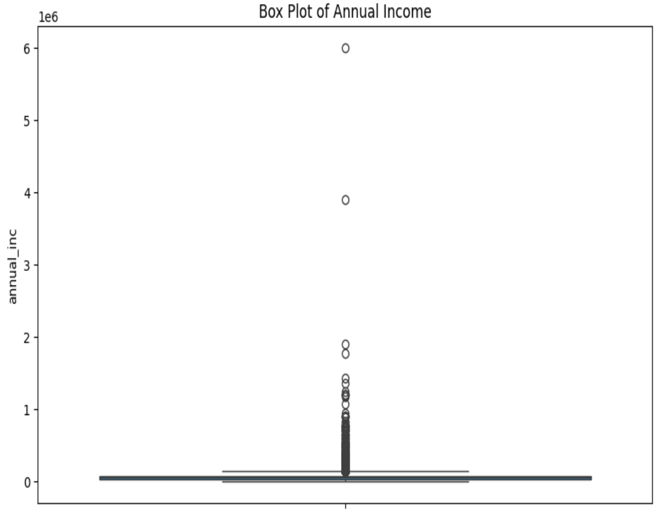
- The installment is a derived metric based on the loan_amnt and int_rate, calculated using the loan's principal, interest rate, and term. Given that we've already determined that there are no significant outliers in loan_amnt and int_rate, and that these variables are within expected ranges, the same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

Stage 2 – Data Analysis

Annual Income







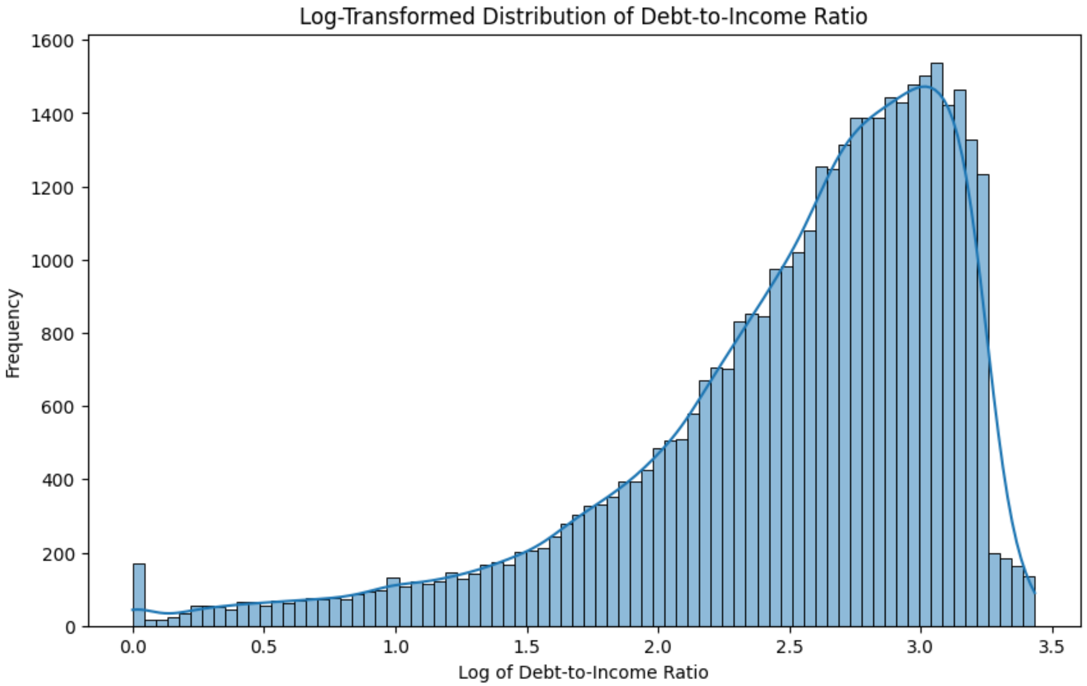
Results of annual_inc univerate analysis

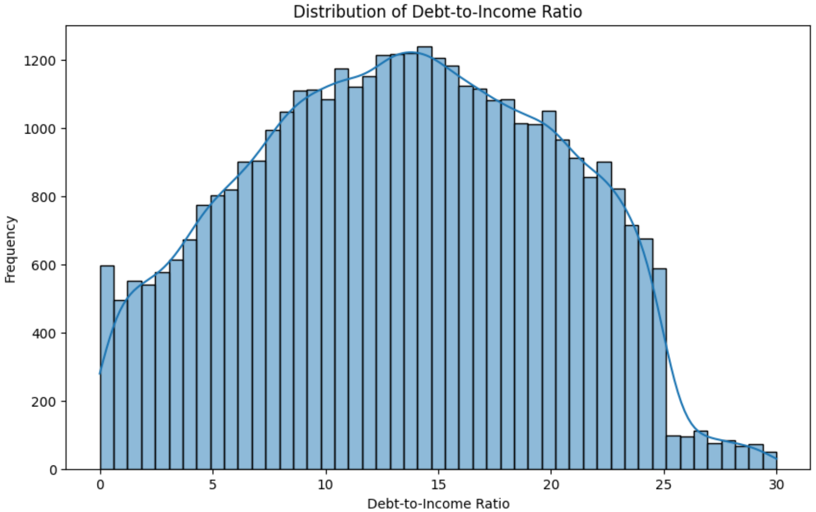
- The annual incomes range from 4,000to6,000,000, with a mean of approximately $68,970.
- The median income is $59,020, indicating a slight skew toward higher incomes.
- The standard deviation of $63,165 suggests significant variability in income levels.
- The distribution is right-skewed, with high-income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

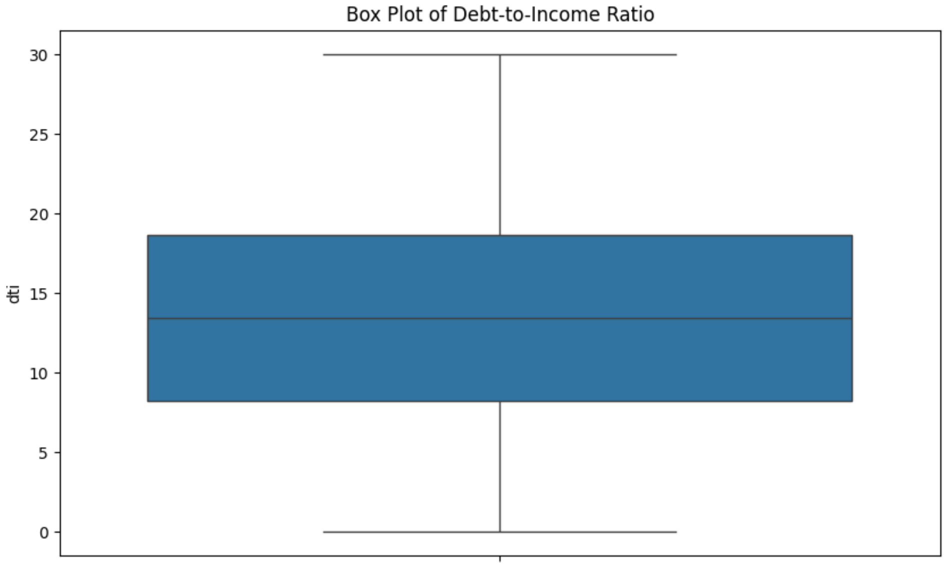
Conclusion: While outliers exist, all data should be included in the analysis. Special handling, such as log transformation, may be needed in bivariate or multivariate analyses involving annual_inc.

Stage 2 – Data Analysis

Delinq 2 years







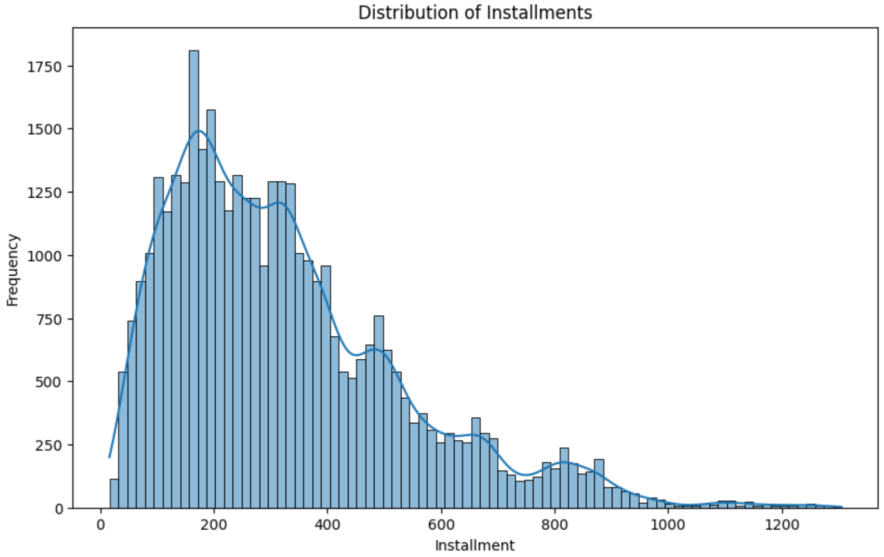
Results of dti univerate analysis

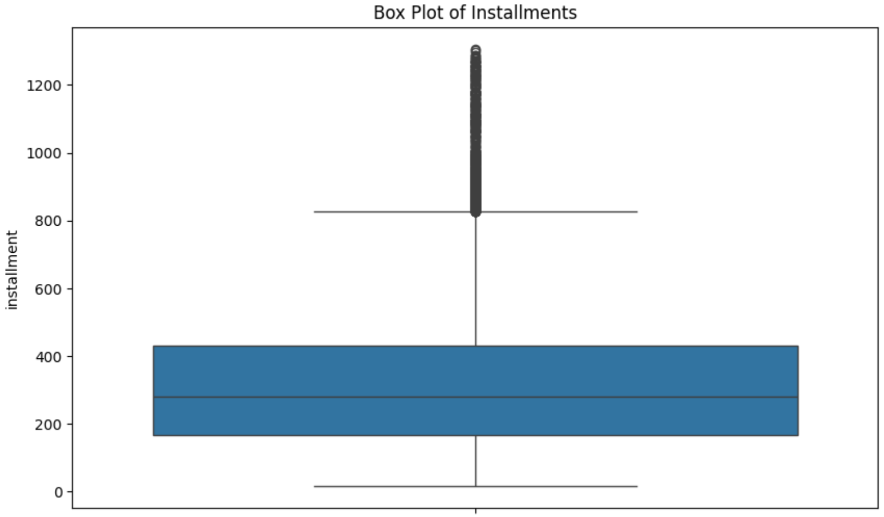
- The debt-to-income ratios range from 0% to 29.99%, with a mean of approximately 13.36%.
- The median DTI is 13.45%, indicating a balanced distribution around this value.
- The standard deviation of 6.67% suggests moderate variability in DTI among borrowers.
- The distribution is normal, with no extreme outliers, as the maximum value is within expected

Conclusion: All data should be included in the analysis as the distribution of DTI values is within the expected range and shows no significant outliers.

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 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.