Lending Club Case Study

This case study presents an analysis of loan data from Lending Club, focusing on key factors influencing loan default risk. The analysis aims to provide actionable insights for business decisions regarding loan approvals.

Objective:

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 data-set. Also the company wants to understand the driving factors behind loan default for detail analysis. The company can utilize this knowledge for its portfolio and risk assessment.

This can also be understood by:

-> Identifying applicants who are likely to repay their loans are important, as they can generate profits through interest payments. Rejecting such applicants would result in a loss of potential business.

-> On the other hand, approving loans for applicants not likely to repay and at risk of default can lead to financial losses for the company.

Data Description:

The data-set comprises 39,717 rows and 111 columns, capturing loan data from 2007 to 2011 which is sufficient to conduct analysis based on past data.The goal is to identify patterns indicating the likelihood of default, which can inform decisions such as loan denial, reducing loan amounts, or offering loans to risky applicants at higher interest rates.

Some important characters:
· Annual Income(annual_inc): Higher income boosts approval chances.
· Home Ownership(home_ownership): Provides collateral, increasing approval likelihood.
· Employment Length(emp_length): Longer tenure suggests financial stability.

· Debt to Income (dti): Lower DTI increases approval odds.
· State (addr_state): Useful for demographic analysis.
· Loan Amount (loan_amt): Amount requested by the borrower.
· Grade (grade): Creditworthiness rating.
· Term (term): Loan duration in months.
· Issue Date (issue_d): When the loan was approved.
· Purpose (purpose): Reason for the loan.
· Verification Status (verification_status): Whether borrower info is verified.
· Interest Rate (int_rate): Annual interest on the loan.
· Installment (installment):: Monthly repayment amount.
· Public Records (public_rec): Higher values reduce approval chances.
· Bankruptcy Records: More records lower approval odds.

Loan Status (loan_status):

- Fully-Paid: Customers who have fully repaid their loans.
- Charged-Off: Customers who have defaulted on their loans.
- Current: Loans still in progress, excluded from this analysis.

Data Cleaning & Handling:

1. Column Removal:

· Dropped Columns with All Null Values: Columns that contained only null values were removed from the data-set.

- Excluded Non-Informative Columns: Removed columns that did not provide meaningful information, such as : next_pymnt_d, mths_since_last_record, mths_since_last_delinq, desc, collections_12_mths_ex_med, delinq_amnt, chargeoff_within_12_mths, acc_now_delinq, and pymnt_plan.

2. Data Cleaning:

· Removed Percentage Symbols: The % symbol was stripped from values in the int_rate and revol_util columns.

3. Data Type Conversion:

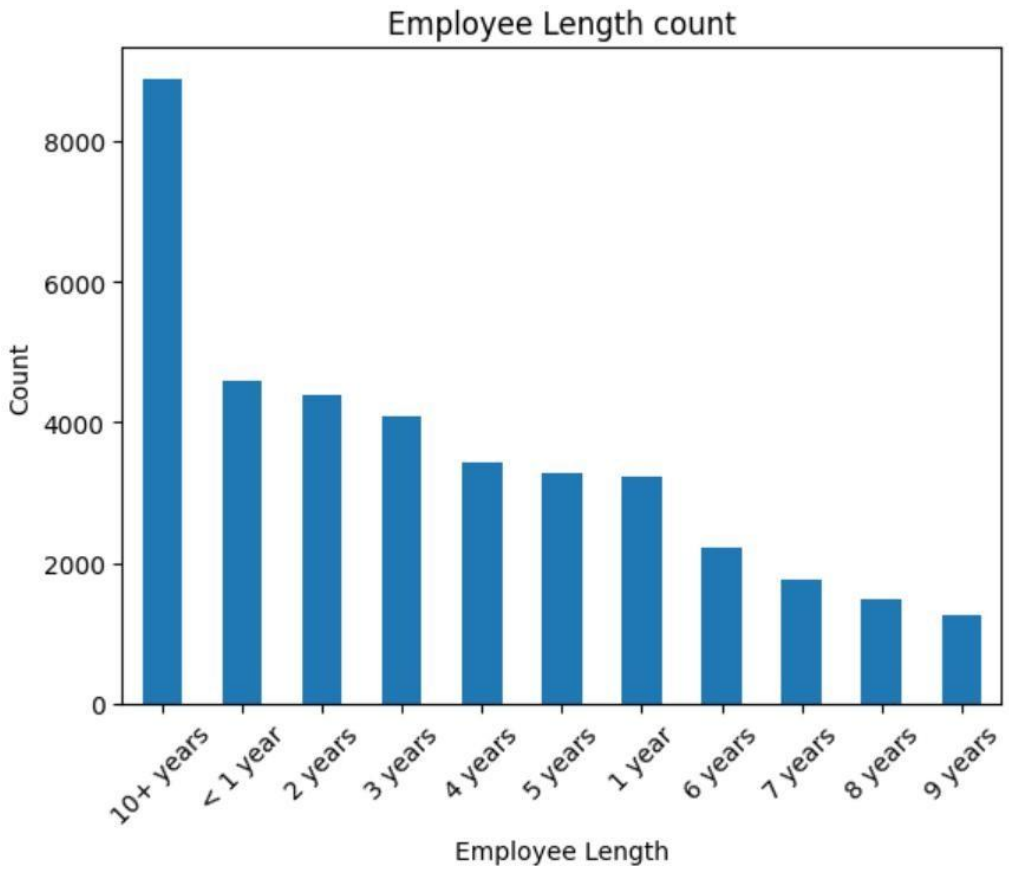
· Updated Data Types: Converted required columns to appropriate data types, such as int, float, or string.

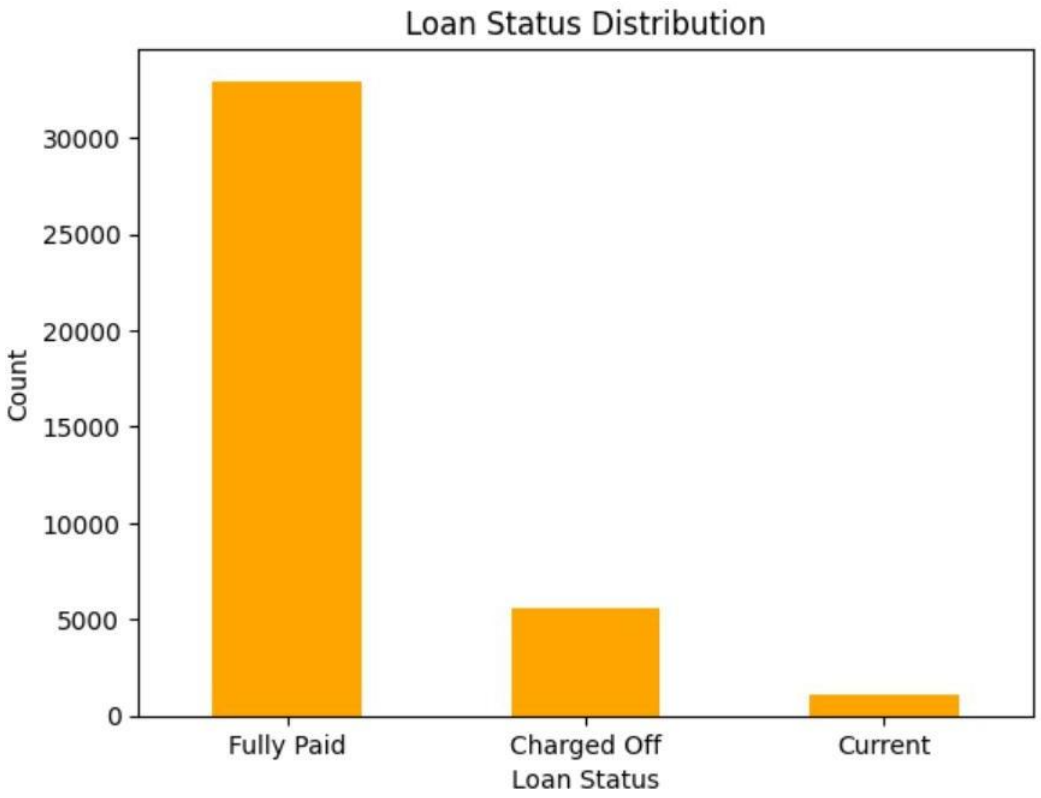
Derived Columns:

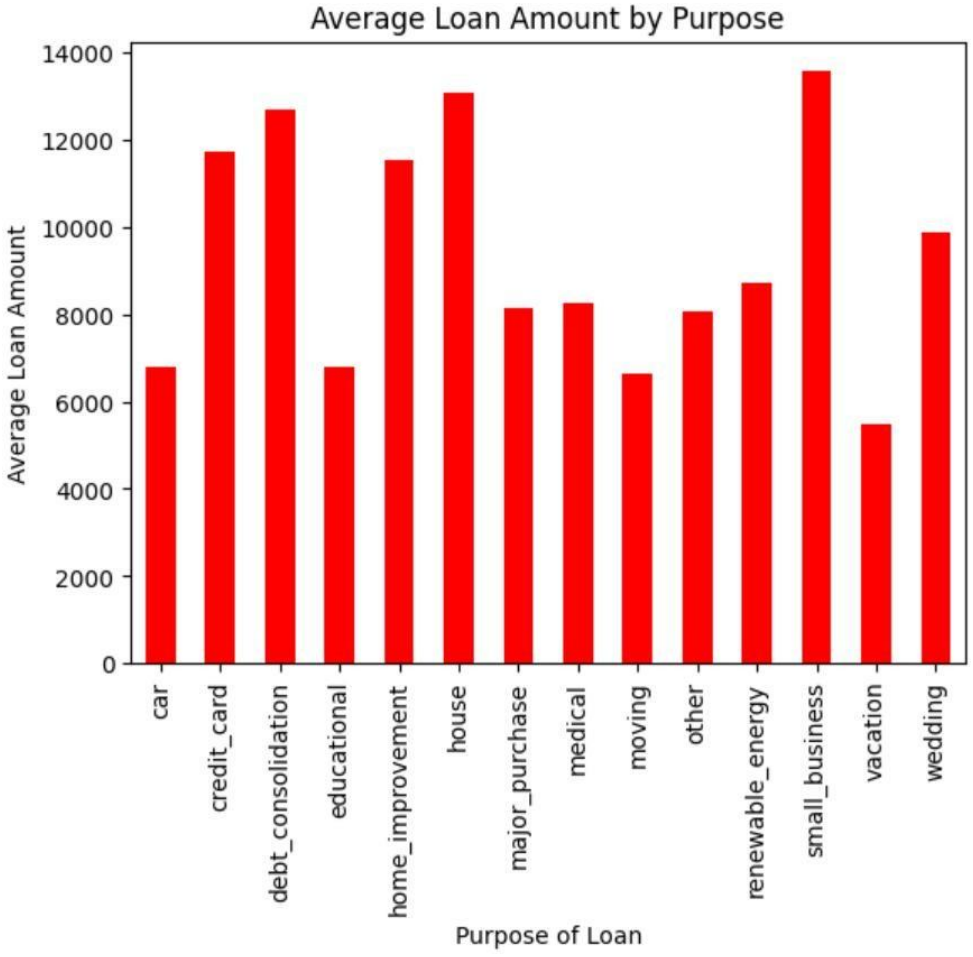
· Loan Status Numeric Conversion: Created a new column, loan_status_deriv, which assigns numeric values to different loan statuses.
· Paying Capacity Calculation: Added a new column, paying_capacity, calculated as the ratio of loan_amnt to annual_inc (loan_amnt/annual_inc).

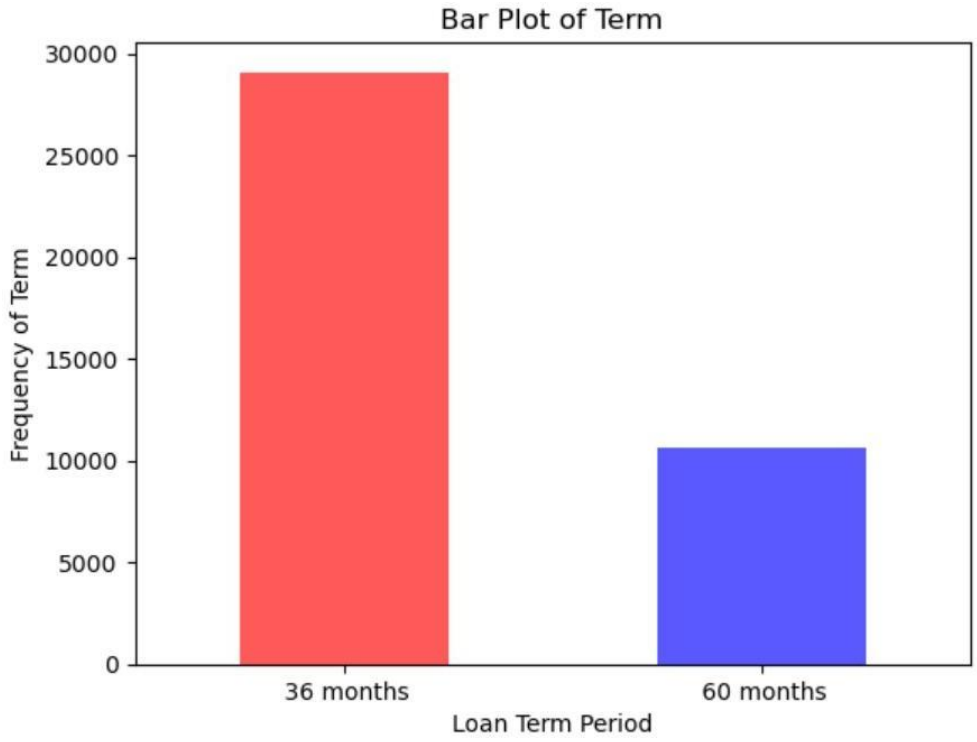
Univariate & Segmented Analysis:

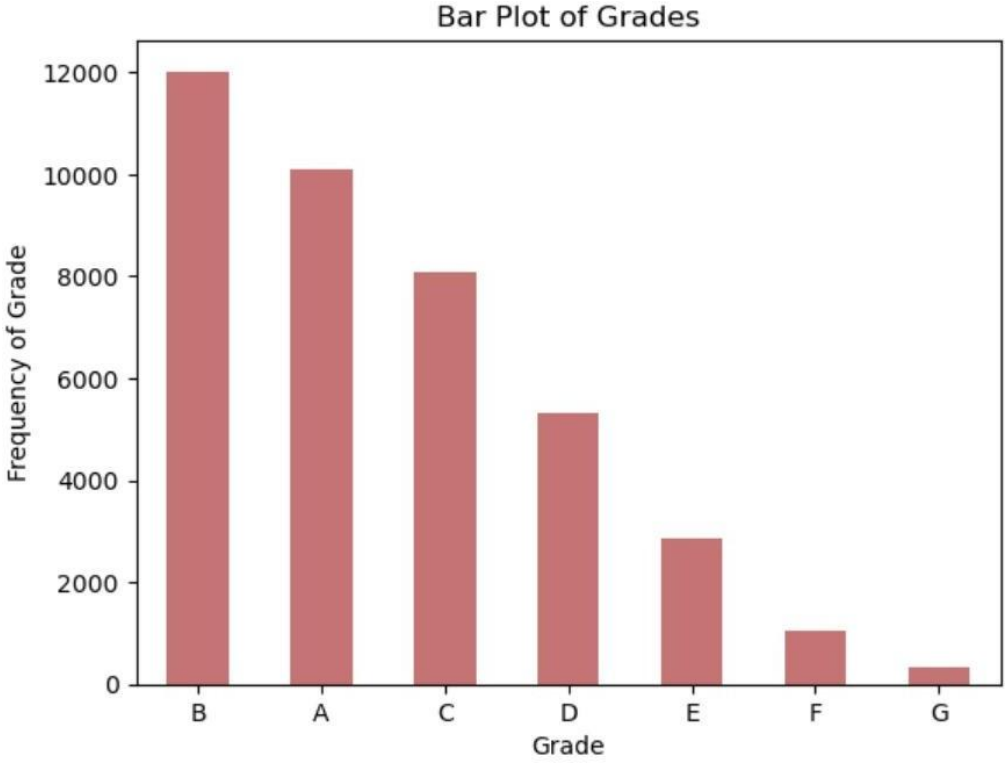
It's a statistical method used to analyze and summarize data-sets containing only 1 variable which can either be categorical or quantitative.
Lets see some of these analysis in form of graphs

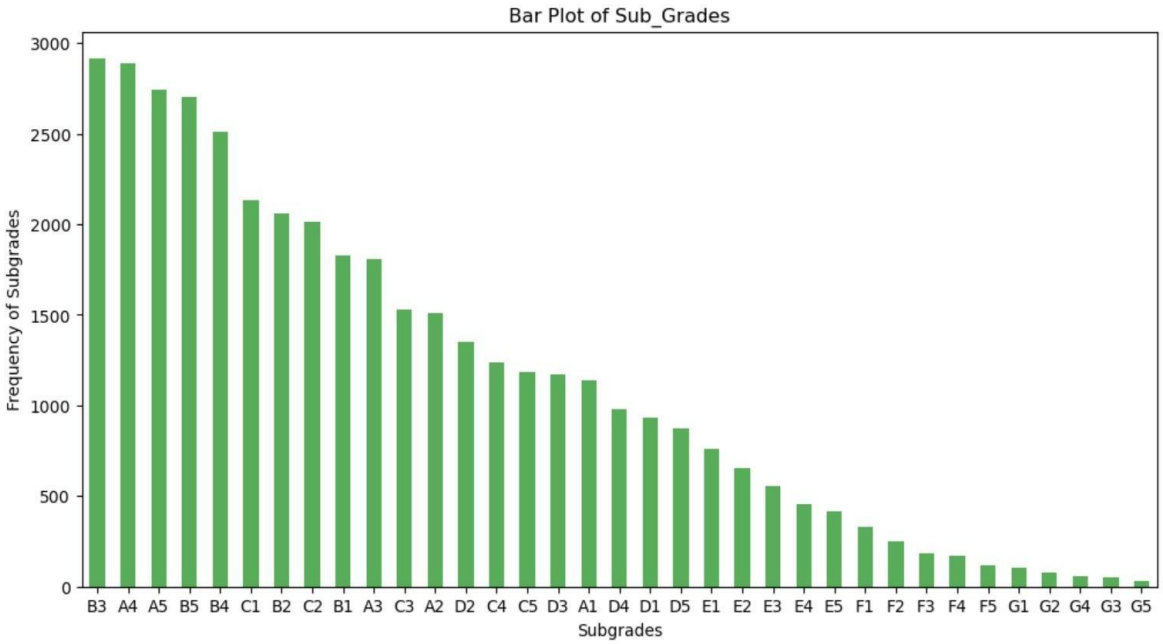


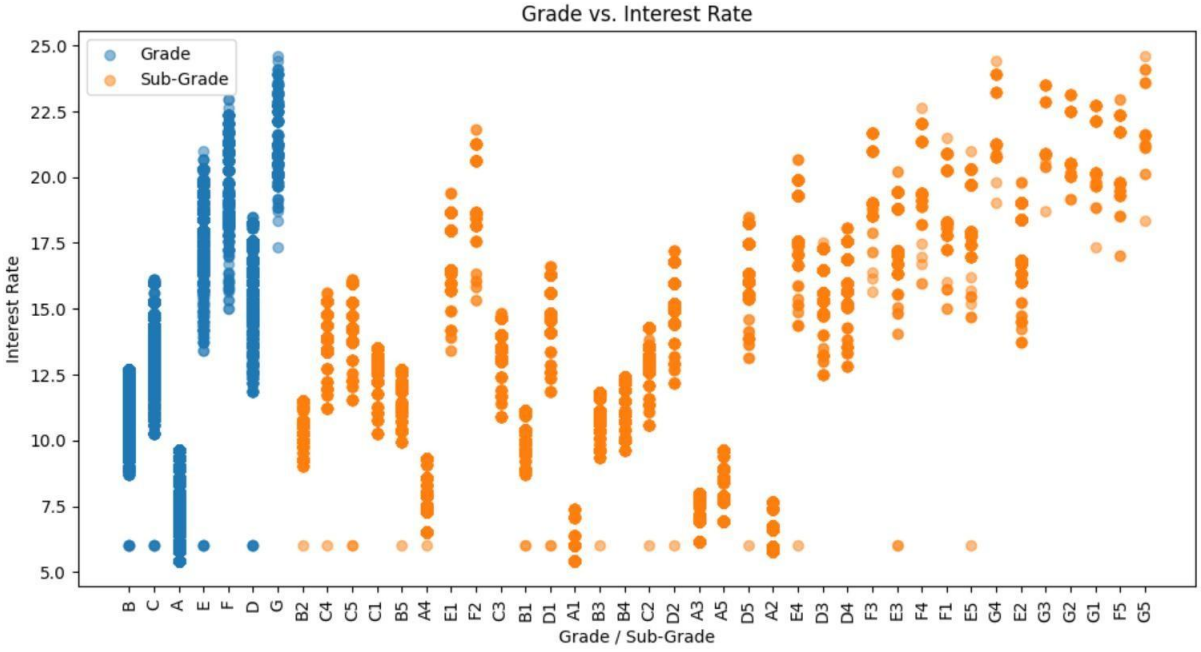

Loan Status Distribution
Count
30000
25000
20000
15000
10000
5000
0
Fully Paid
Charged Off
Current
Loan Status


Average Loan Amount by Purpose
Average Loan Amount
14000
12000
10000
8000
6000
4000
2000
0
car
credit_card
debt_consolidation
educational
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding
Purpose of Loan

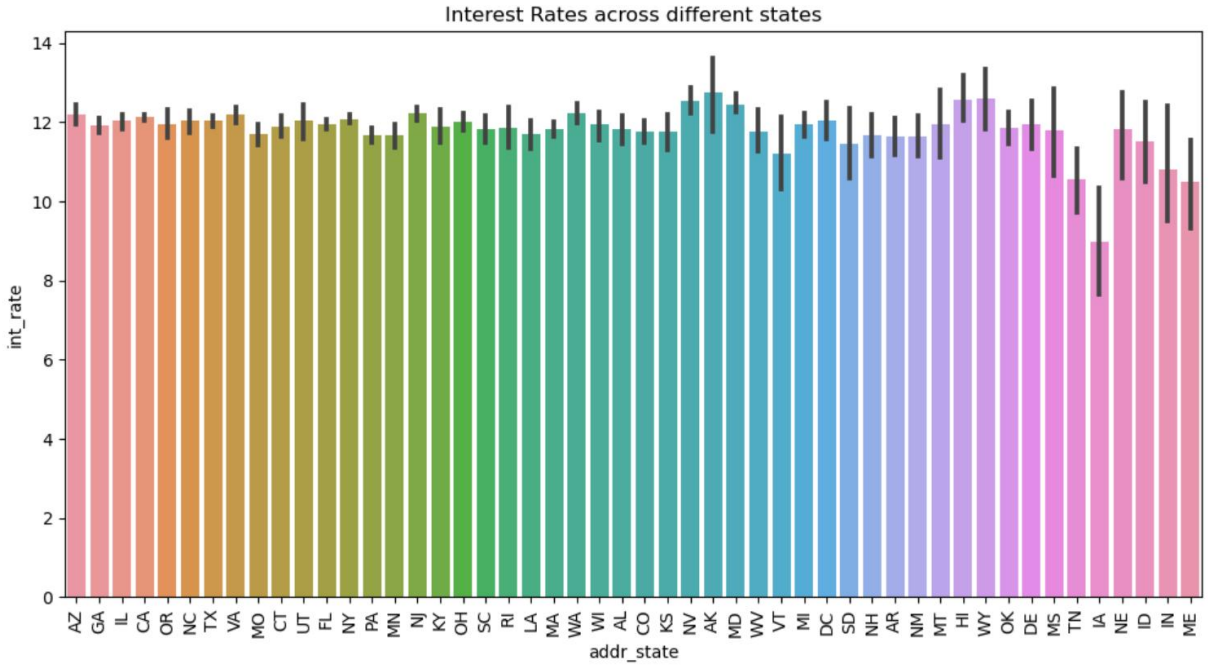

Bar Plot of Term
30000
25000
20000
15000
10000
5000
0
Frequency of Term
36 months
60 months
Loan Term Period

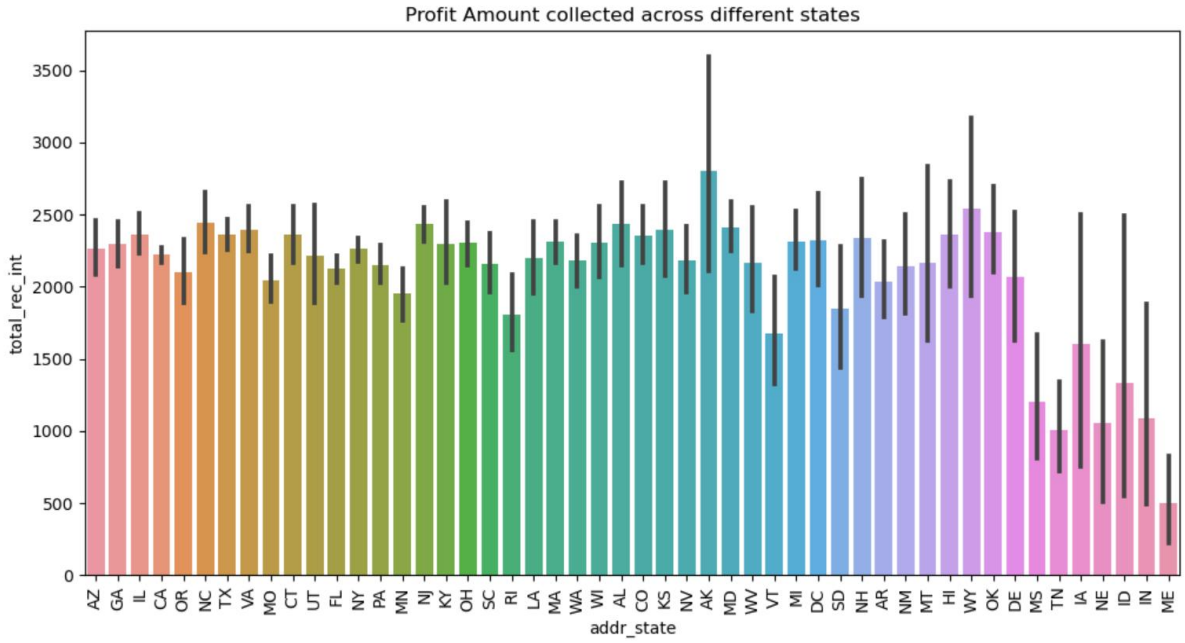

Bar Plot of Grades
12000
10000
8000
6000
4000
2000
0
Frequency of Grade
B
A
C
D
E
F
G
Grade





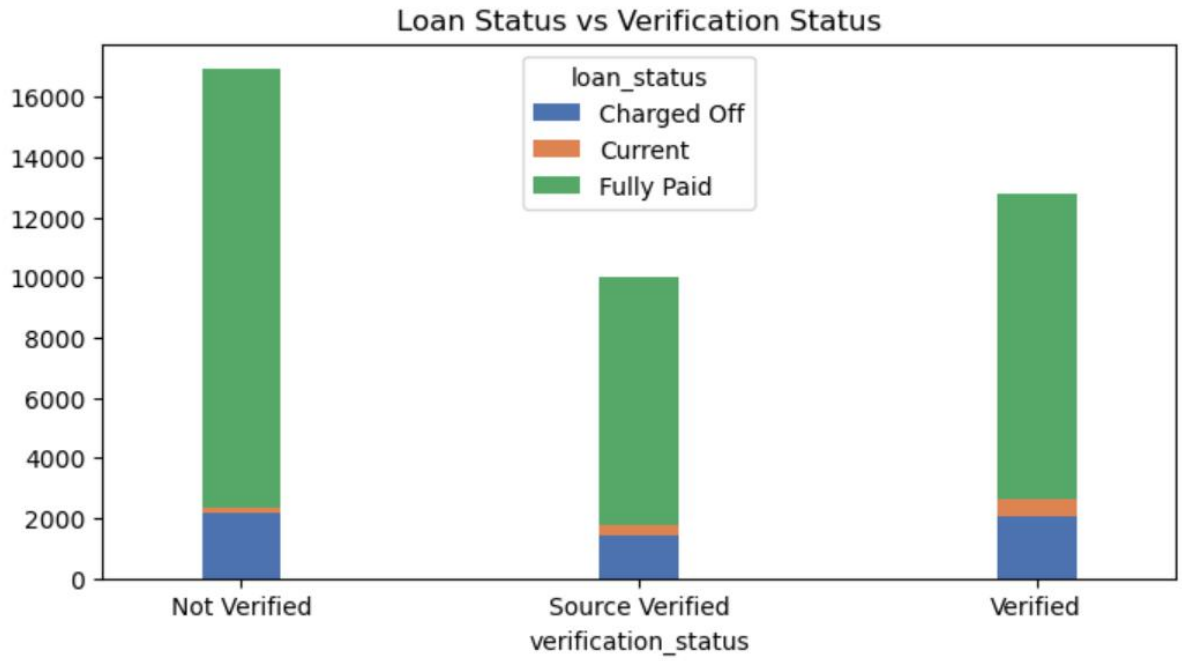
Bivariate Analysis:

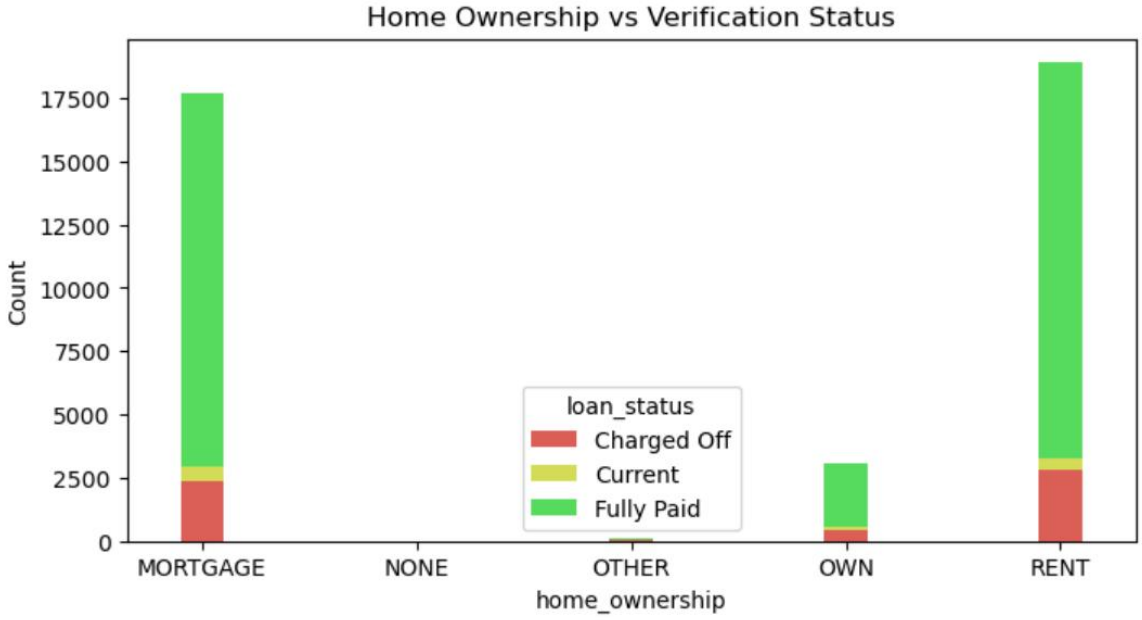


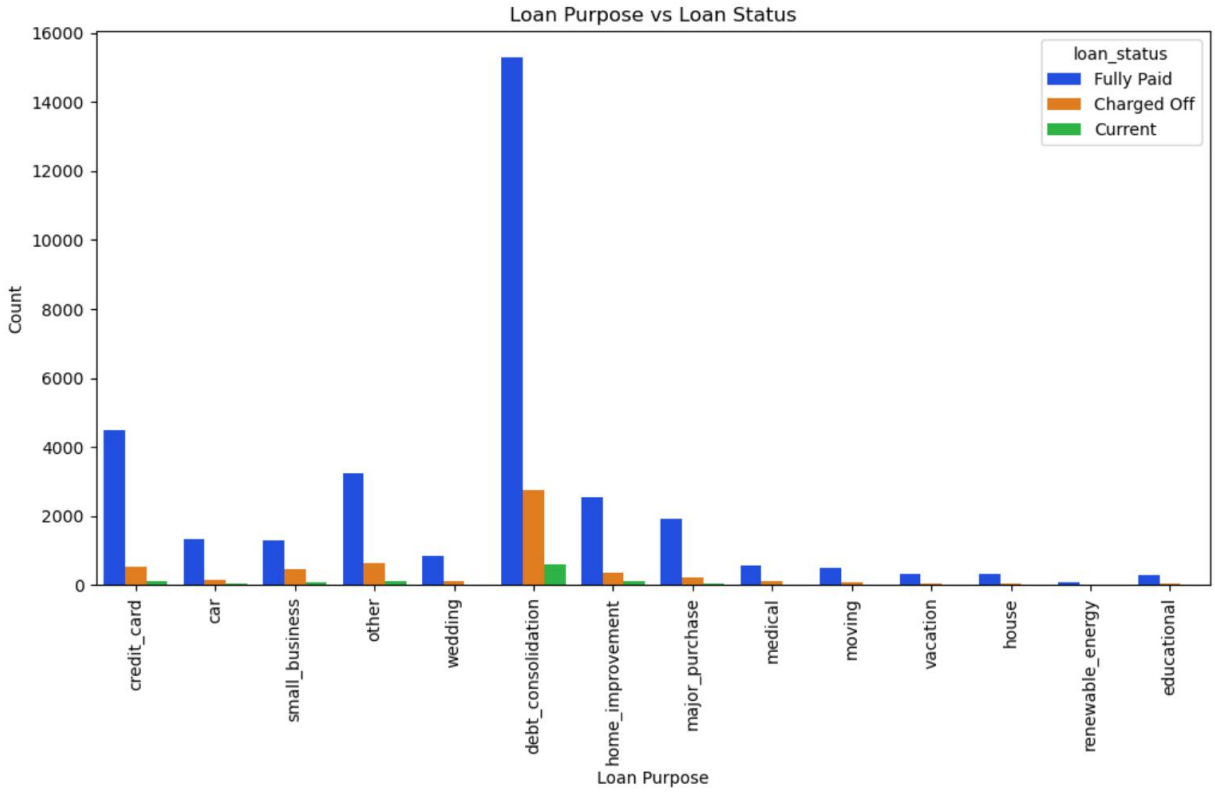


These 2 graphs show the interest rates and corresponding amounts collected from different states.

Some other bivariate analysis are:


Loan Status vs Verification Status
loan_status
Charged Off
Current
Fully Paid
16000
14000
12000
10000
8000
6000
4000
2000
0
Not Verified
Source Verified
Verified
verification_status


Home Ownership vs Verification Status
Count
17500
15000
12500
10000
7500
5000
2500
0
loan_status
Charged Off
Current
Fully Paid
MORTGAGE
NONE
OTHER
OWN
RENT
home_ownership



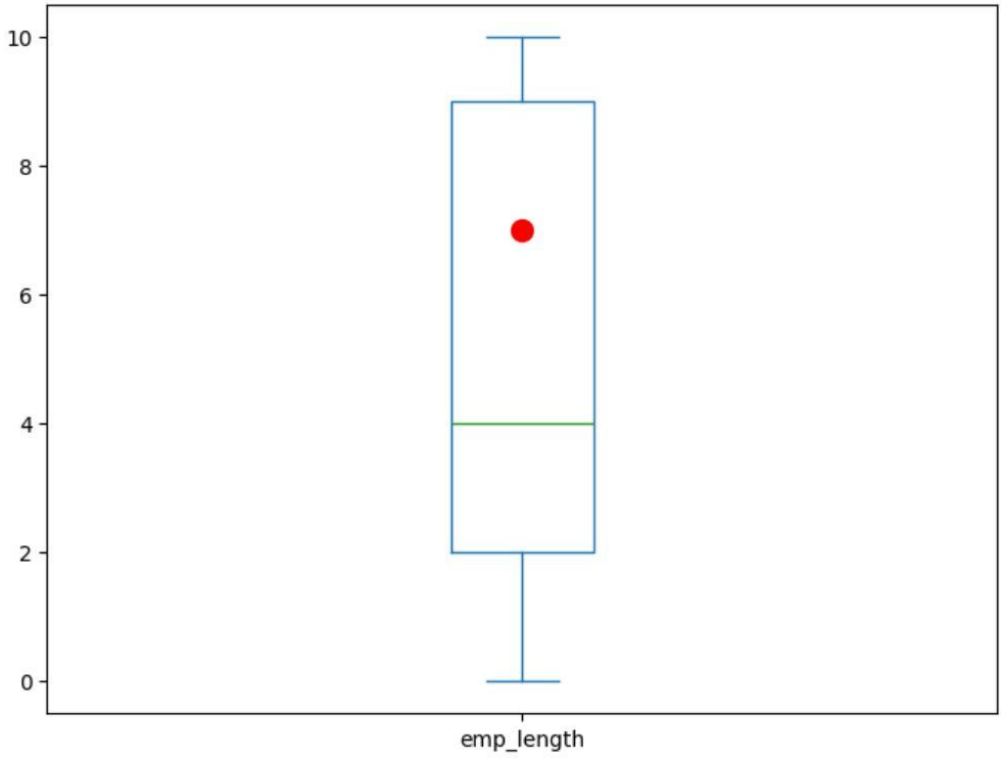
Driving Factors:

Let’s take a loan candidate. Let's say we are evaluating the application of a candidate with id 1069102.

Driving Factors which will be looked:

1. Employment Length

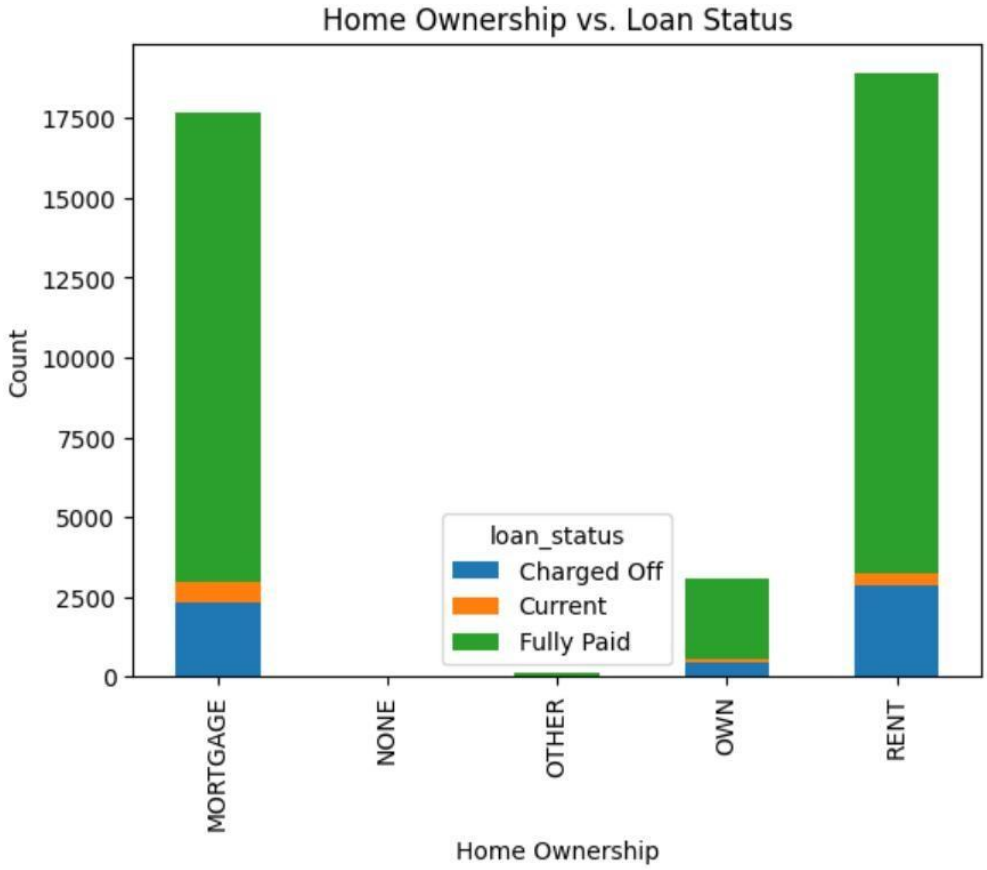
Below box plot shows that, among all the applicants, the given loan applicant has good employment length. Therefore, he is likely to have a stable income source and has lower credit risk.



2. Home Ownership

The below plot illustrates the relationship between home ownership and loan status. It indicates that applicants who own their homes are less likely to experience loan charge-offs.

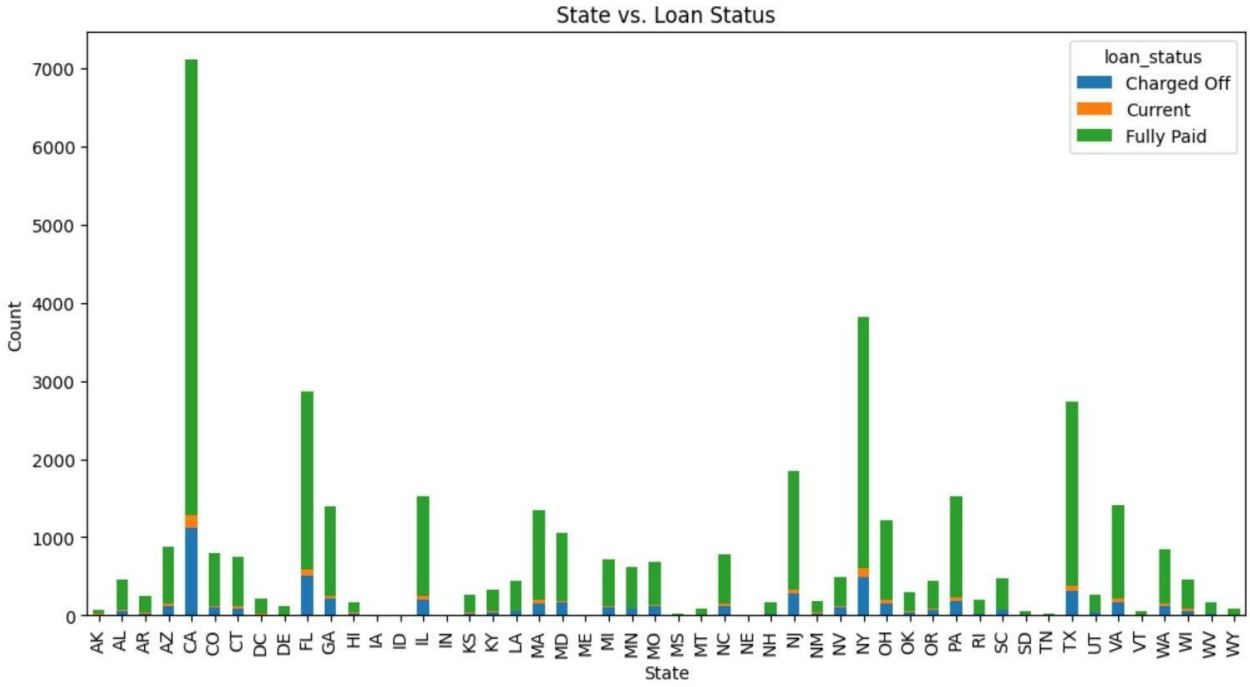
For the candidate in question, who has a home ownership status of "MORTGAGE," this suggests a lower credit risk based on this parameter. This implies that, from the home ownership perspective, the candidate is at a reduced risk of loan default.



3. State Wise Analysis

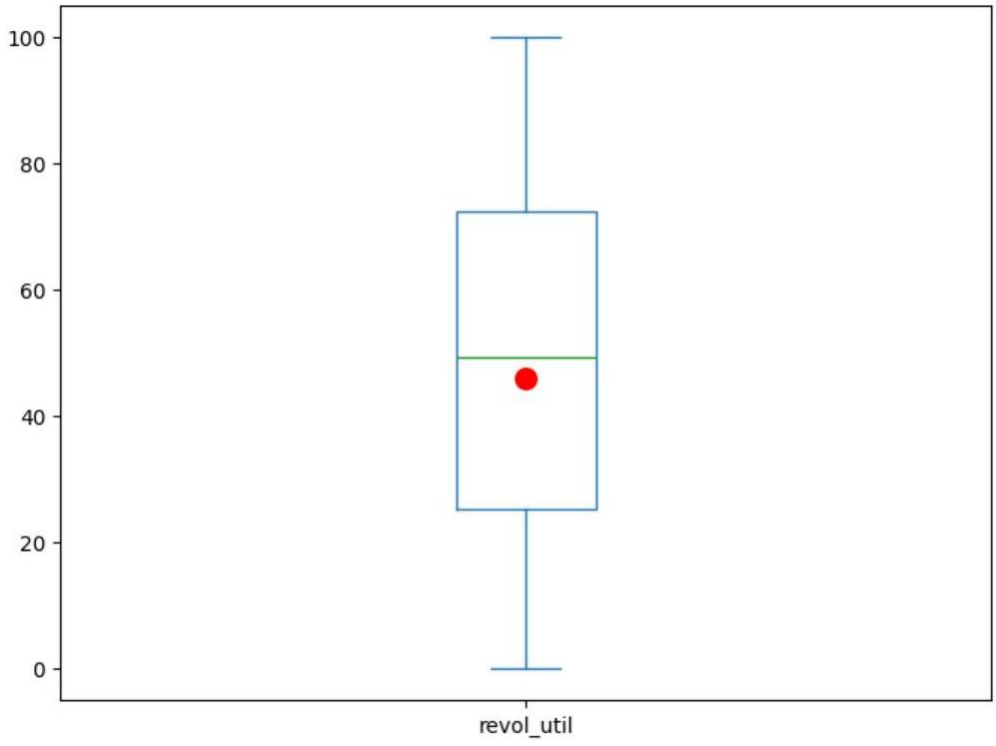
The bar plot reveals variations in loan status across different states. States like GA, FL, and CA exhibit higher charged-off rates, while NY and IL demonstrate a predominance of fully paid loans. These findings suggest regional differences in borrower behavior or economic conditions that may influence loan performance.

The given candidate is from OH state.



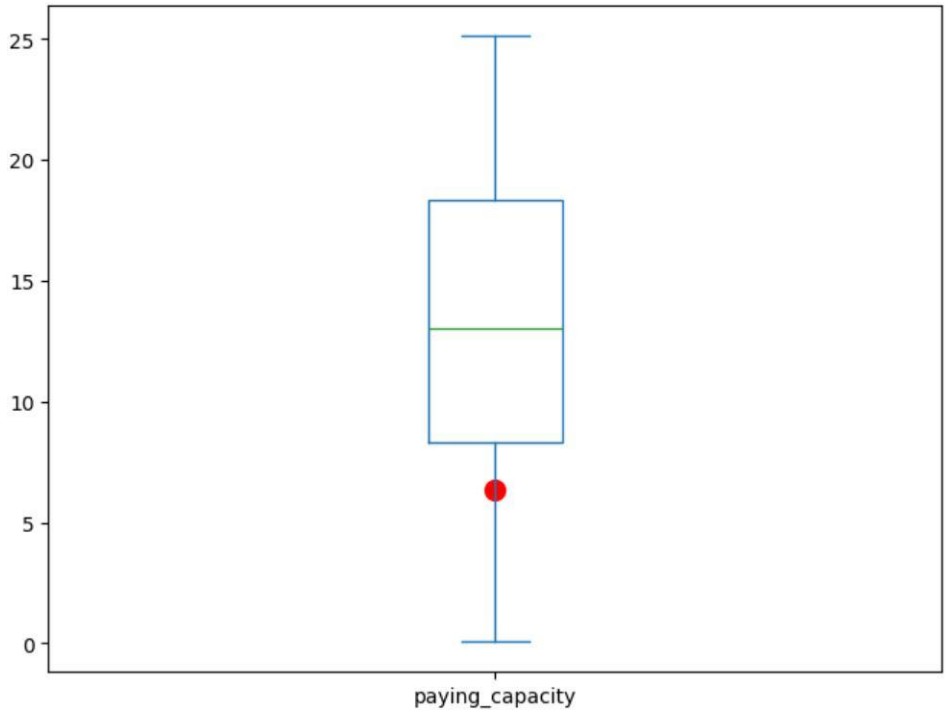
4. Revolving Line Utilization

Lower revolving line utilization indicates lower credit risk of the candidate. As seen in the below plot, candidate has lower credit risk then median with respect to revolving line utilization



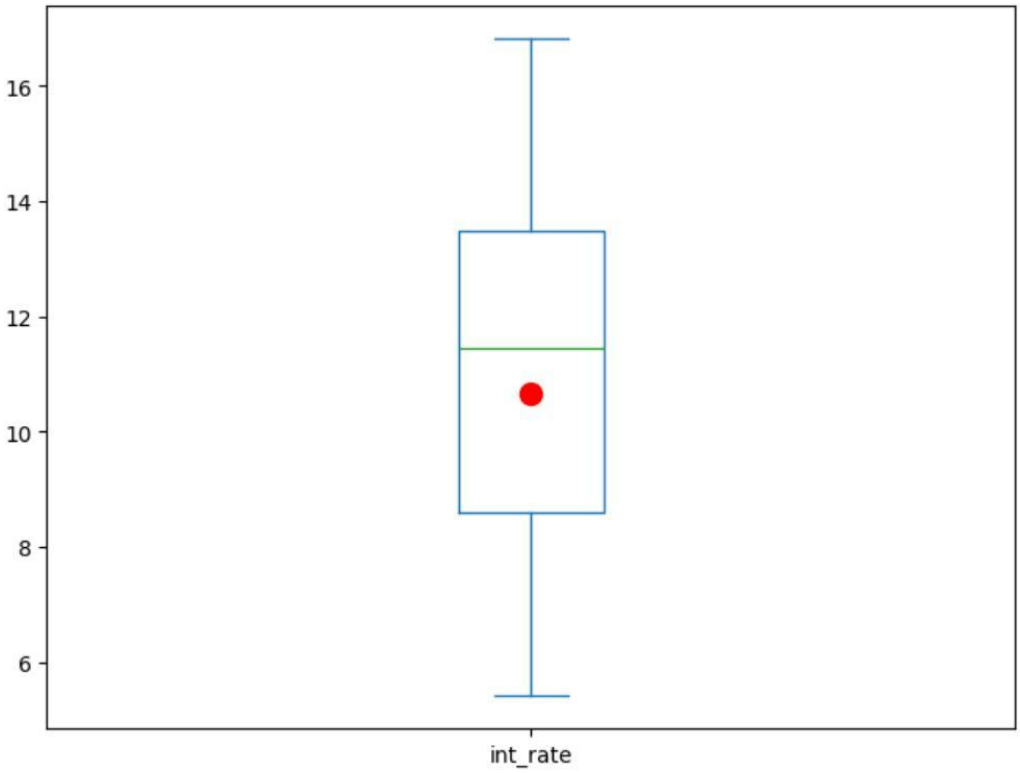
5. Paying Capacity(Derived column)

The box plot shows the distribution of paying capacity for a filtered data-set, highlighting the median, quartiles, and potential outliers. The scatter plot indicates that the individual loan candidate's paying capacity falls within the inter-quartile range of the overall distribution. And therefore the candidate has good paying capacity for the loan given his/her annual income



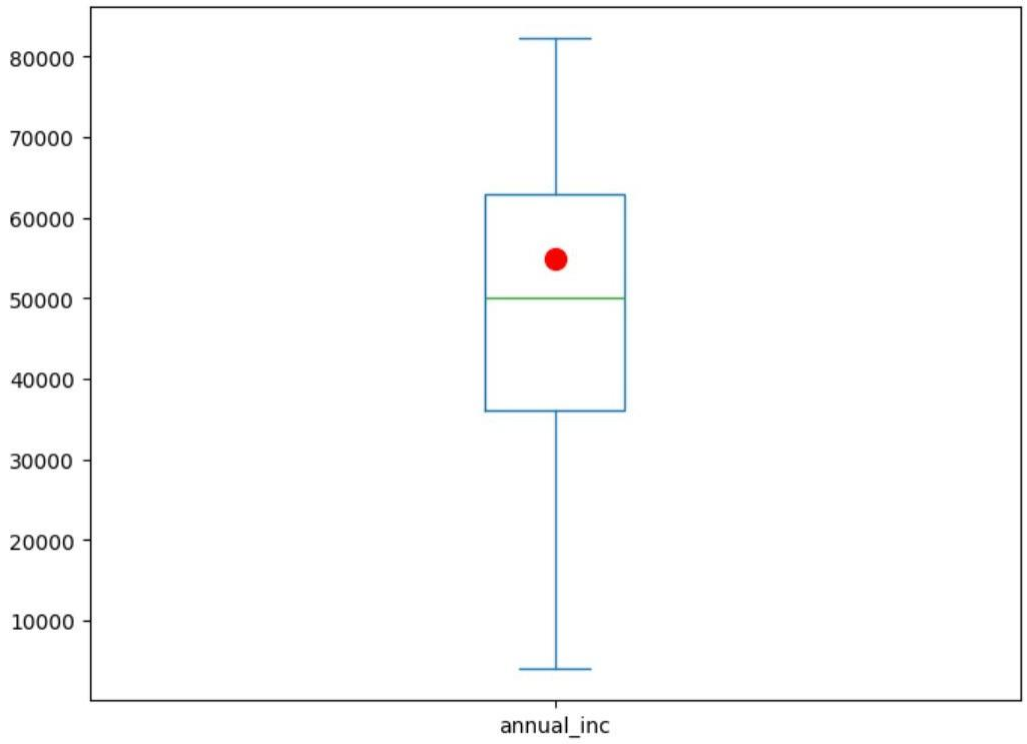
6. Interest Rate

Candidates already having loans with higher interest rates have higher credit default risk. The scatter plot indicates that the individual loan candidate's interest rate falls below the 75th percentile of the overall distribution, suggesting a lower interest rate compared to most others in the data-set.



7. Annual Income

The box plot illustrates the distribution of annual incomes for the filtered data-set, highlighting the median, quartiles, and potential outliers. The scatter plot indicates that the individual loan candidate's annual income falls above the 75th percentile of the overall distribution, suggesting a higher income level compared to most others in the data-set. This aligns with the established relationship between higher income and lower credit default risk.



Conclusions:

The analysis concludes that the given loan applicant does not have a high credit risk based on the examined factors. These findings provide valuable insights for the business in making data-driven loan approval decisions.