Lending Club Case Study

Problem Statement:

To identify variables that may result in a Charged Off or full payment.

Understanding the drivers of loan default risk is essential for effective risk management and lending operations. By leveraging insights from data analysis, we can enhance our ability to identify and mitigate credit risk, ultimately contributing to sustainable lending practices and improved financial outcomes.

Exploratory Data Analysis

EDA is a crucial step in understanding the dataset and extracting insights from it. Here are some steps involved in EDA to identify driver variables in identifying whether custom will default or not:

Data Collection:

- Gather the dataset from relevant sources, ensuring it includes all necessary attributes for analysis.

Data Cleaning:

- **Handle missing values:** Identify missing values in the dataset and decide how to handle them (imputation, deletion, etc.).
- **Remove duplicates:** Check for and remove any duplicate rows in the dataset.
- **Data type conversion:** Convert data types if necessary (e.g., converting string values to numeric).

Data Exploration:

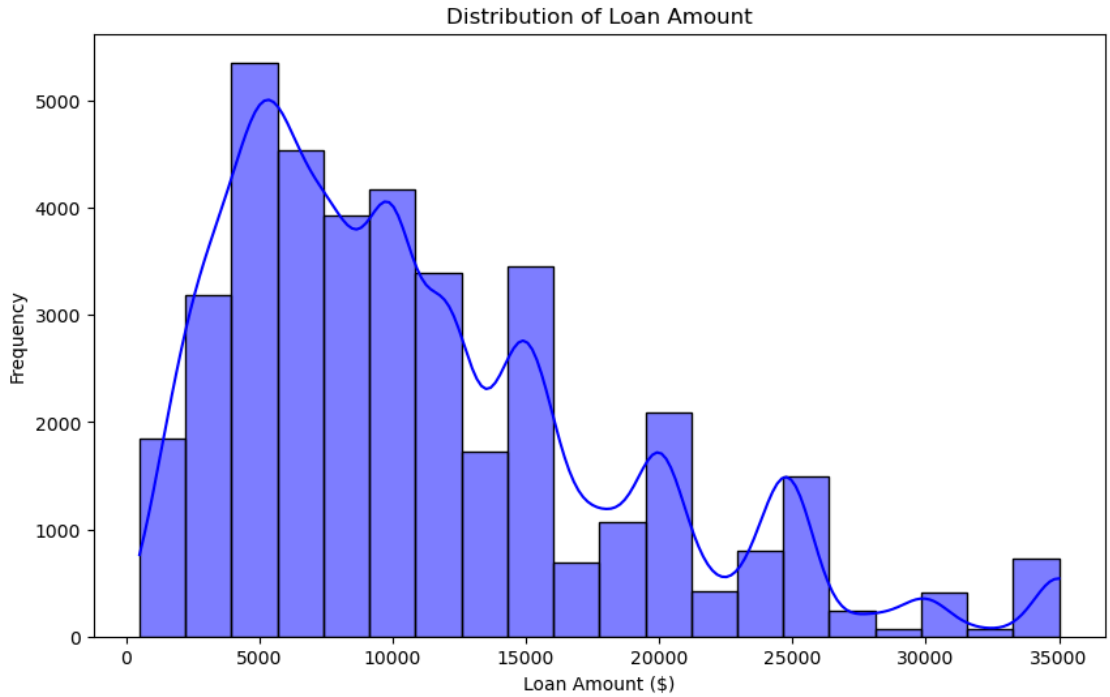
- **Summary statistics:** Compute descriptive statistics (mean, median, min, max, etc.) for numeric variables.
- **Univariate analysis:** Analyze individual variables to understand their distribution using histograms, box plots, etc.
- **Bivariate analysis:** Explore relationships between pairs of variables using scatter plots, correlation matrices, etc.

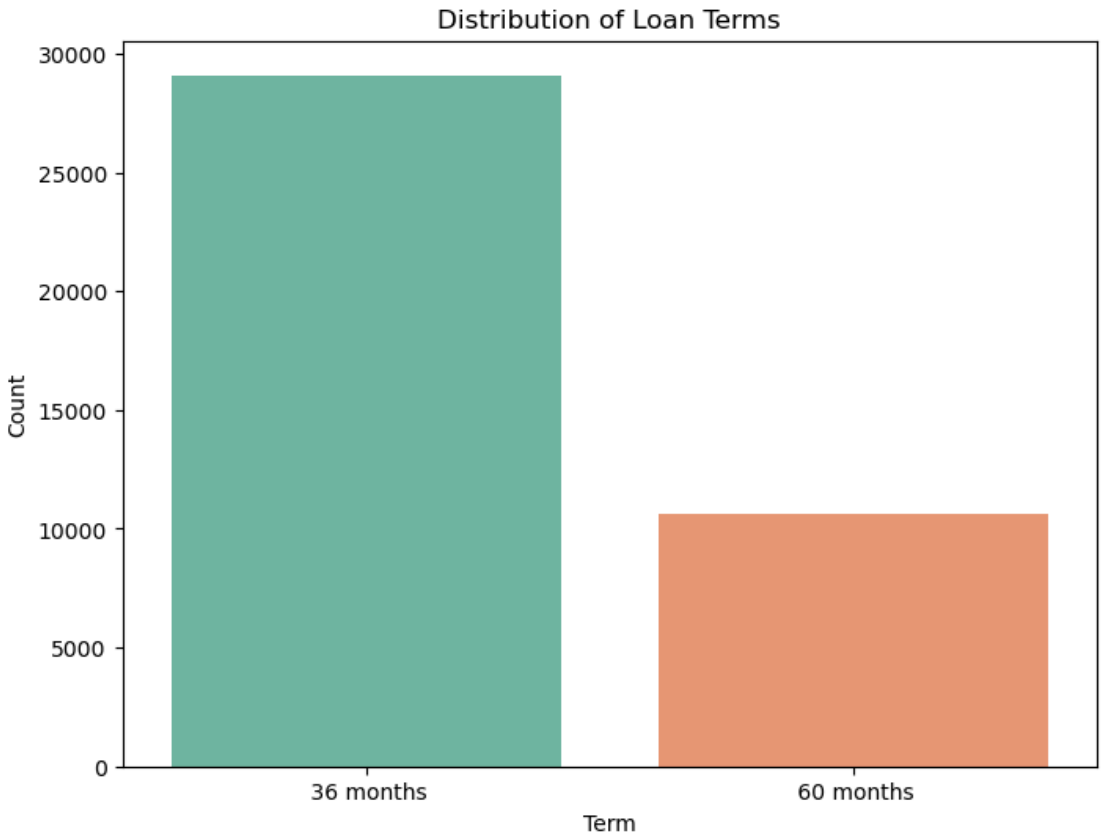
Data Visualization:

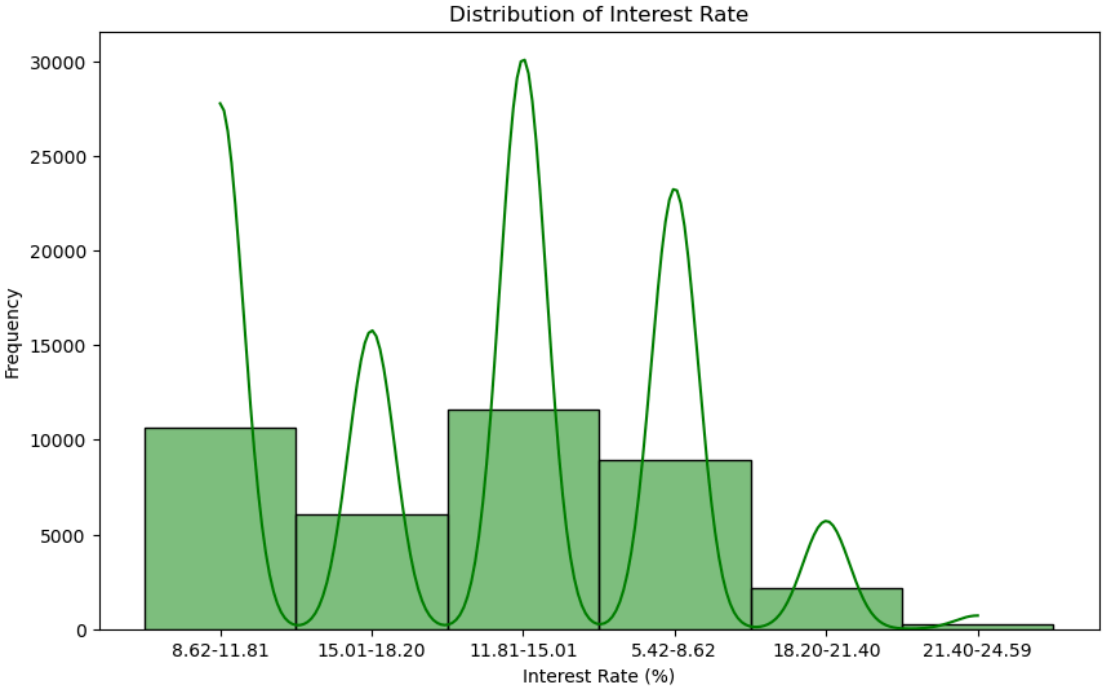
- Use various visualization techniques such as histograms, box plots, scatter plots, correlation matrices, etc., to visually represent the data and gain insights. This Markdown text will

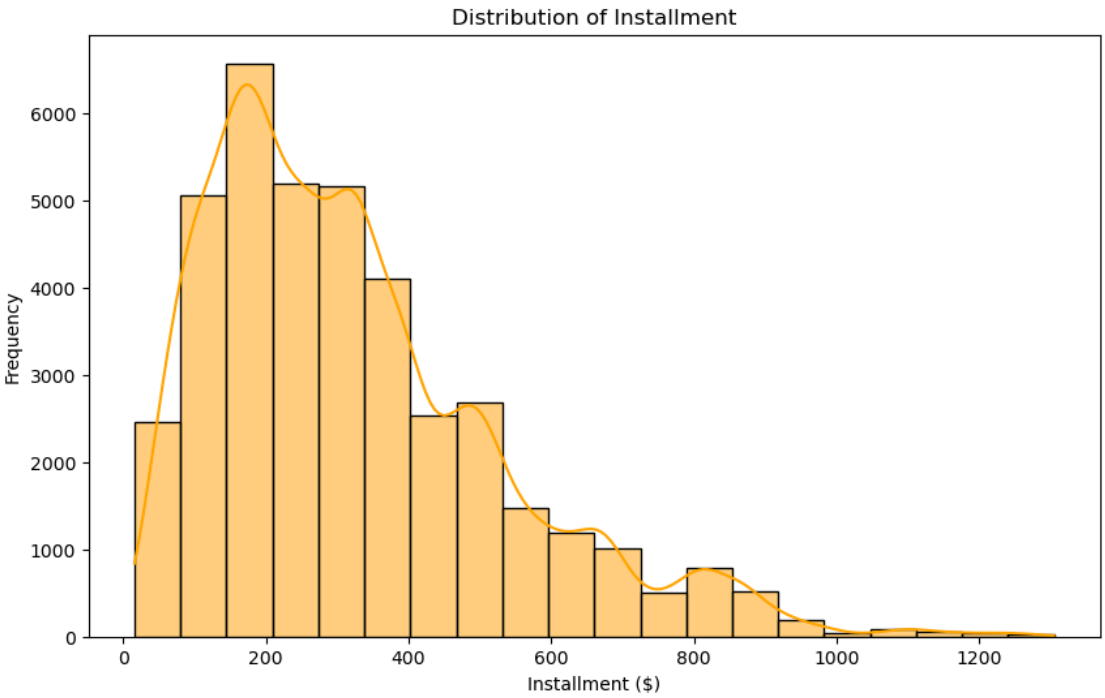
Insight of Univariant and Bivariant Analysis

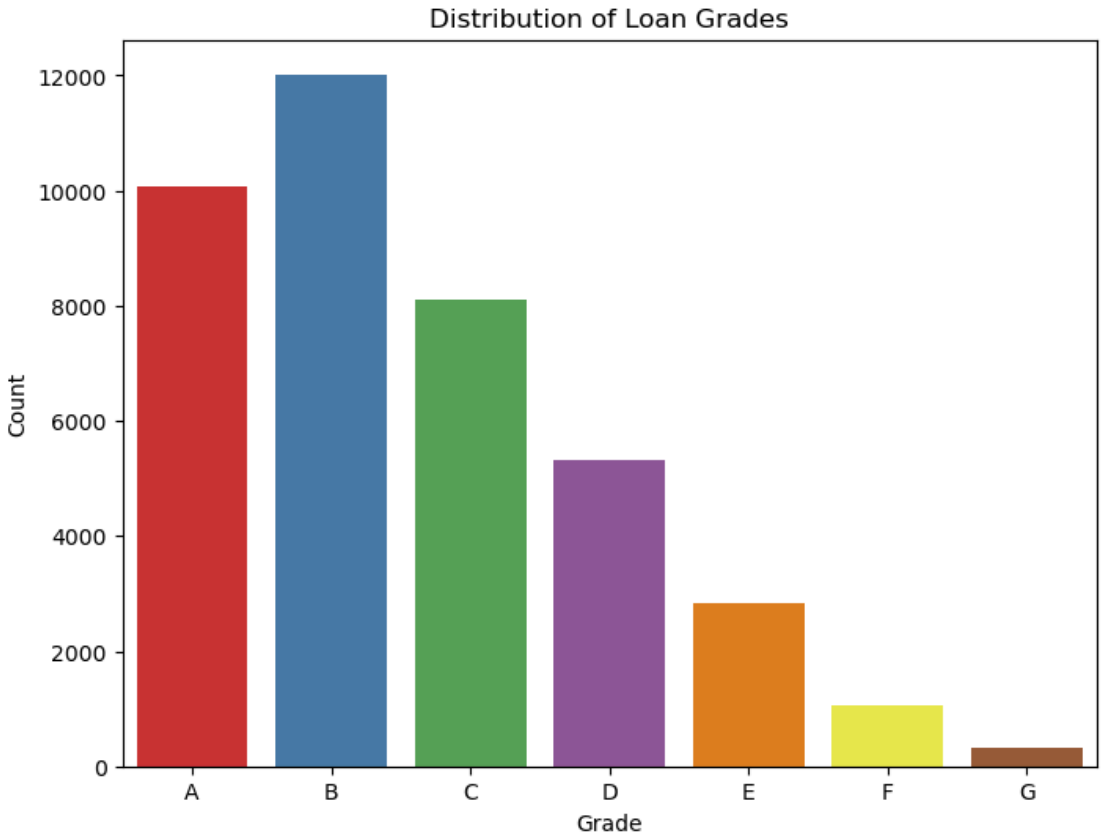
1) Univariant Analysis

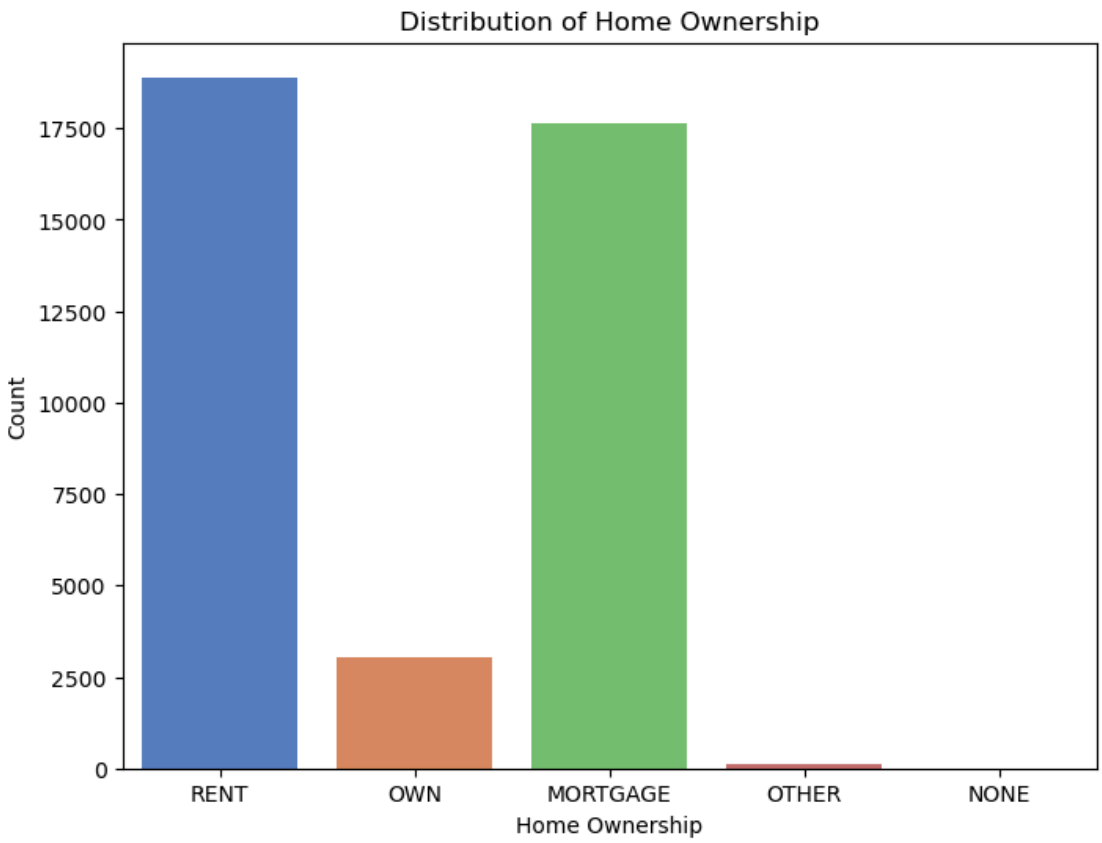


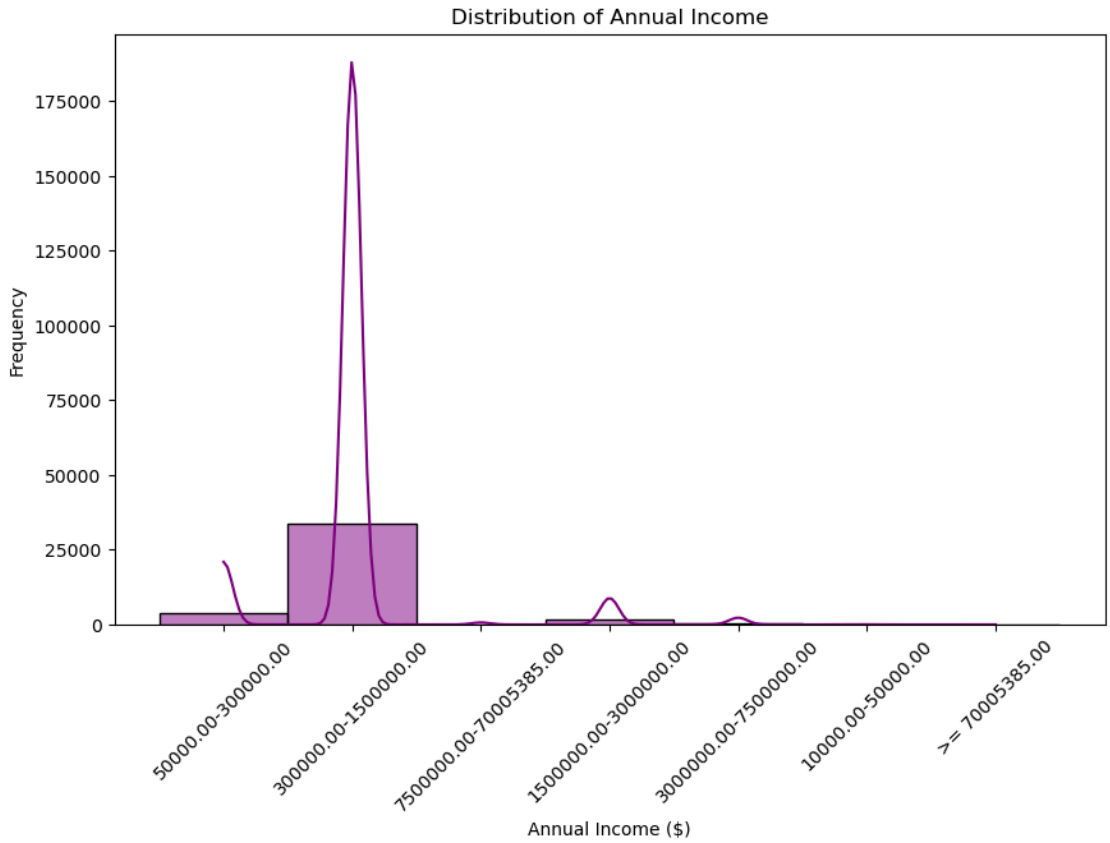

Distribution of Loan Terms
30000
25000
20000
15000
10000
5000
0
Count
36 months
60 months
Term

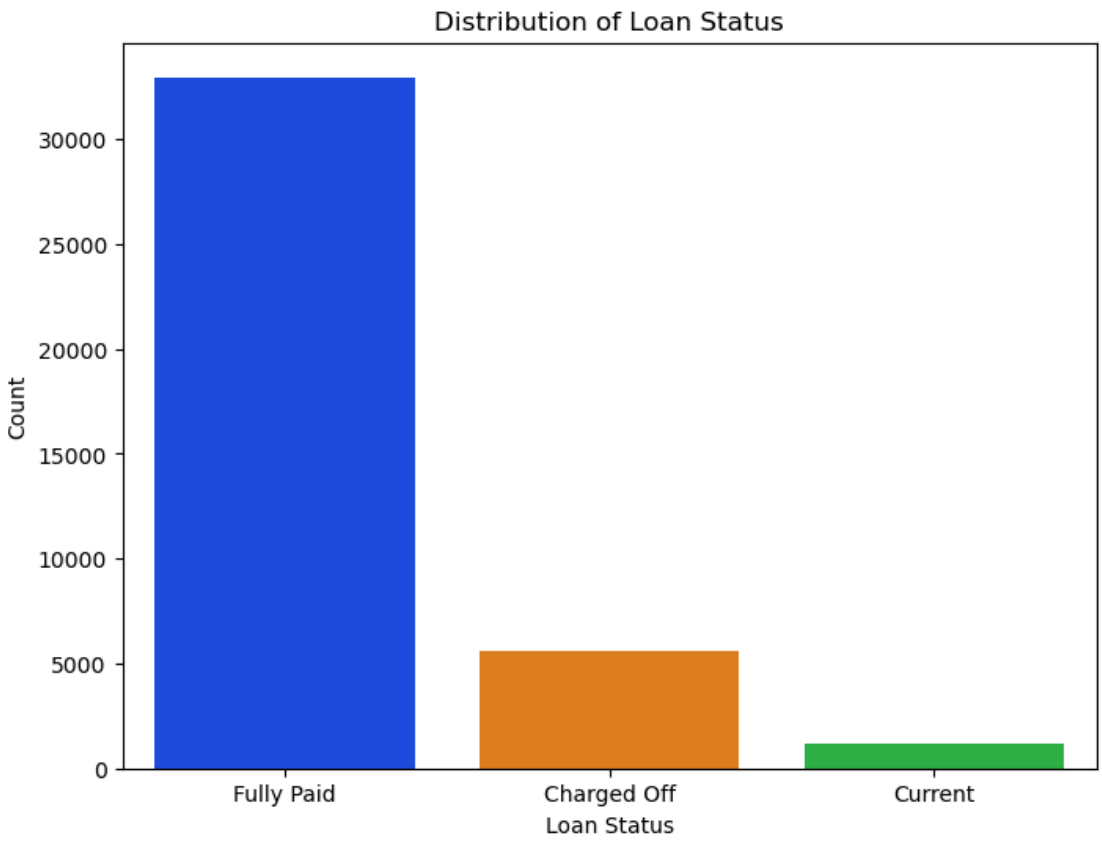

Distribution of Interest Rate
30000
25000
20000
15000
10000
5000
0
Frequency
8.62-11.81
15.01-18.20
11.81-15.01
5.42-8.62
18.20-21.40
21.40-24.59
Interest Rate (%)

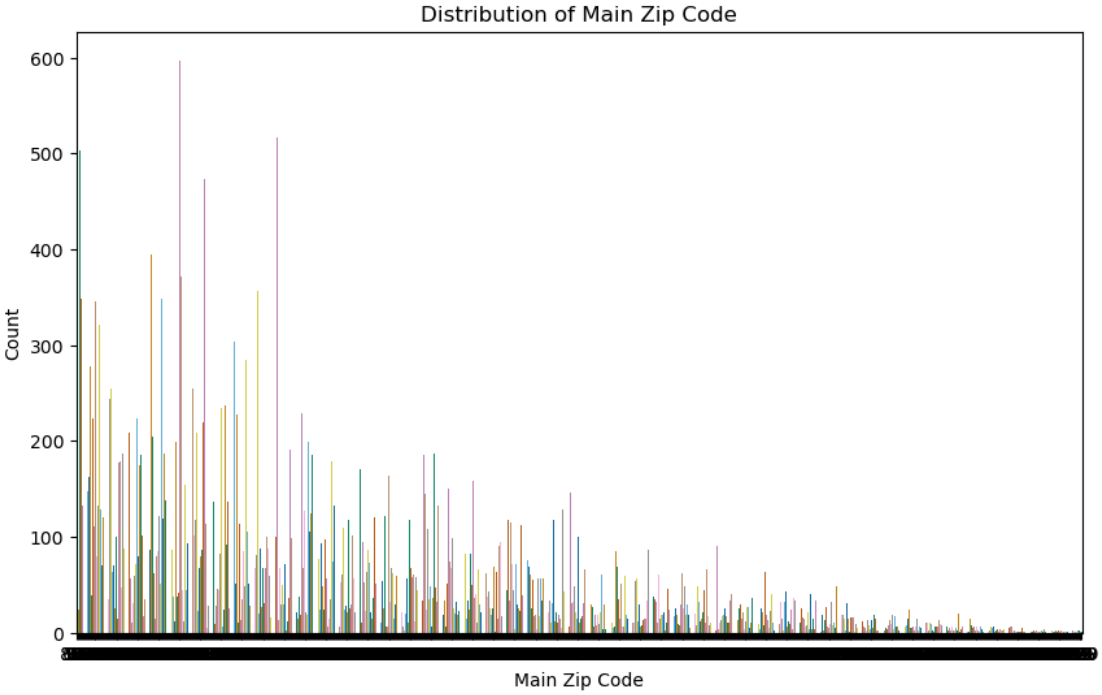

Distribution of Installment
Frequency
Installment ($)
0
1000
2000
3000
4000
5000
6000
0
200
400
600
800
1000
1200

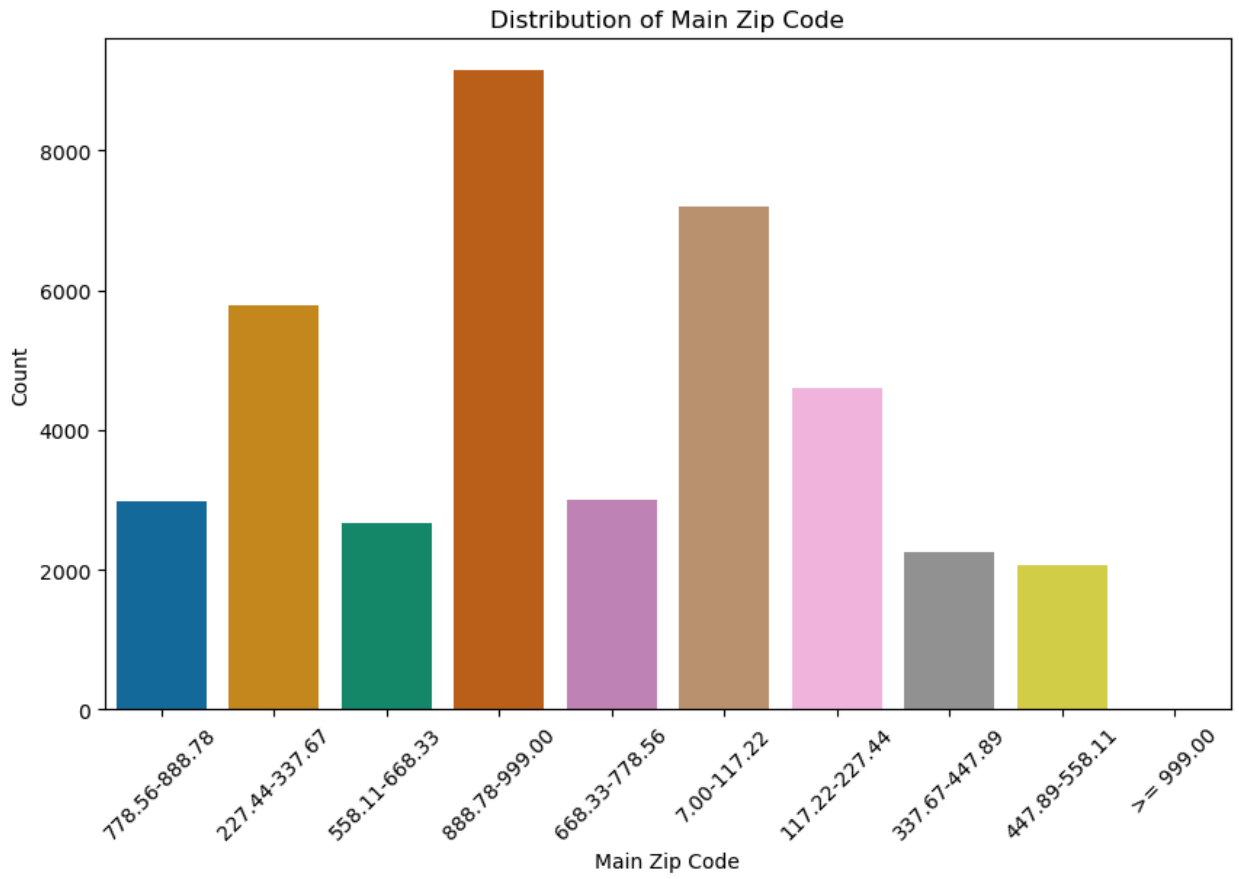

Distribution of Loan Grades
Count
Grade
0
2000
4000
6000
8000
10000
12000
A
B
C
D
E
F
G


Distribution of Home Ownership
Count
17500
15000
12500
10000
7500
5000
2500
0
RENT
OWN
MORTGAGE
OTHER
NONE
Home Ownership


Distribution of Annual Income
Frequency
175000
150000
125000
100000
75000
50000
25000
0
50000.00-300000.00
300000.00-1500000.00
7500000.00-70005385.00
1500000.00-3000000.00
3000000.00-7500000.00
10000.00-50000.00
>= 70005385.00
Annual Income ($)


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


Distribution of Main Zip Code
600
500
400
300
200
100
0
Count
Main Zip Code



Univariant Analysis Inferences:

Loan Amount (loan_amnt):

- This suggests that borrowers often request moderate to high loan amounts. The peak amount being just above 5000

Term (term):

- The majority of loans have a term of 36 months, indicating that borrowers prefer shorter repayment periods.
- However, a notable portion of loans also have a term of 60 months, indicating longer repayment commitments.

Interest Rate (int_rate):

- Interest rates vary widely, with a significant portion falling between 11% and 15%.
- Some loans have relatively high-interest rates exceeding 15%, suggesting higher risk associated with those loans.

Installment (installment):

- Monthly installments are distributed across a range of values,most between 0 to 400 with peak being close to 200
- This indicates variability in repayment obligations, with some borrowers facing higher monthly payment burdens.

Grade (grade):

- Loan grades are predominantly in the B and C categories, indicating moderate risk levels.
- However, there is also a presence of loans in lower-grade categories (e.g., D and E), which may represent higher risk.

Home Ownership (home_ownership):

- Renters constitute a significant portion of borrowers, followed by those with mortgages.
- This indicates a diverse mix of home ownership statuses among borrowers.

Annual Income (annual_inc):

- Annual incomes vary widely, with a range of 3,00,000 to 15,00,000 have a more possibility of applying for a loan
- However, there is a notable presence of lower-income borrowers, which may pose higher default risk.

Loan Status (loan_status):

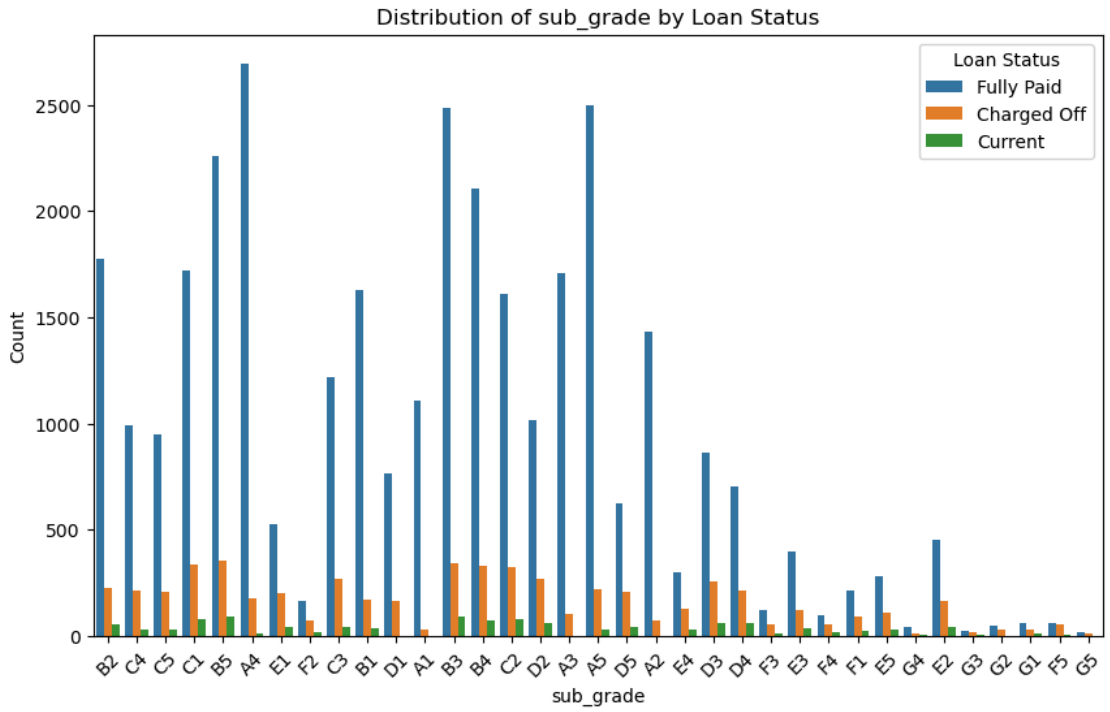
- The majority of loans are labeled as "Fully Paid," indicating successful repayment.
- However, there is also a portion of loans labeled as "Charged Off," indicating default or non-repayment.

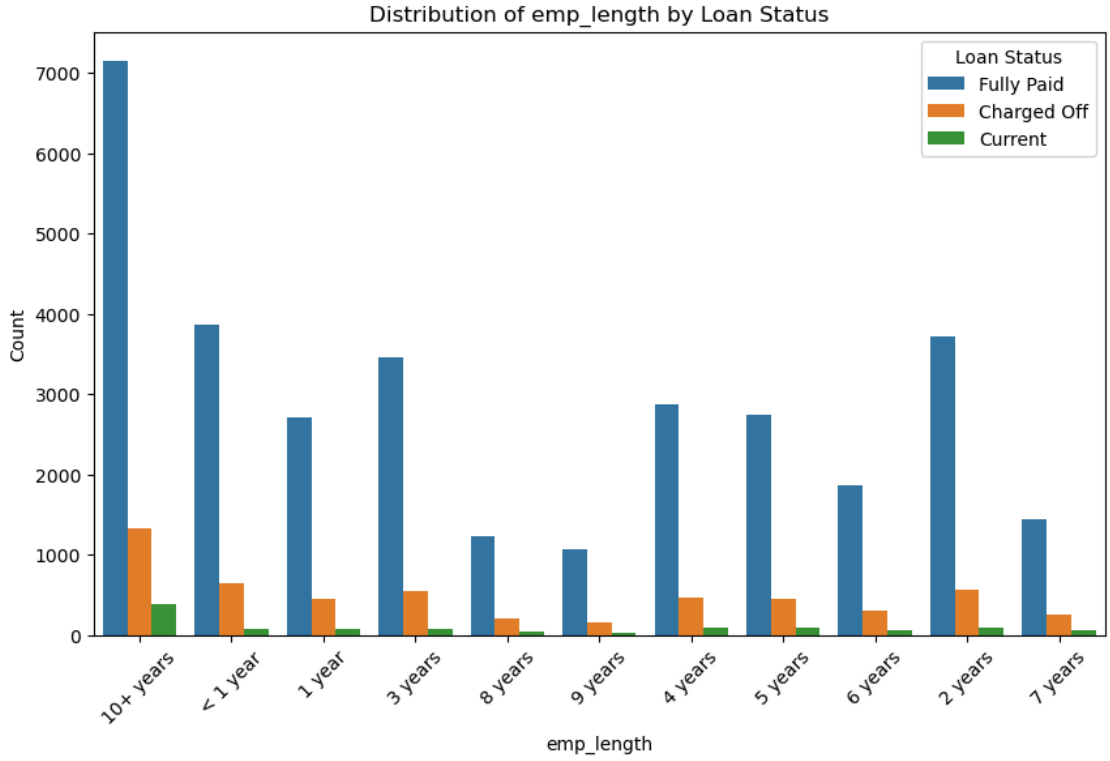
Main Zip Code (main_zip):

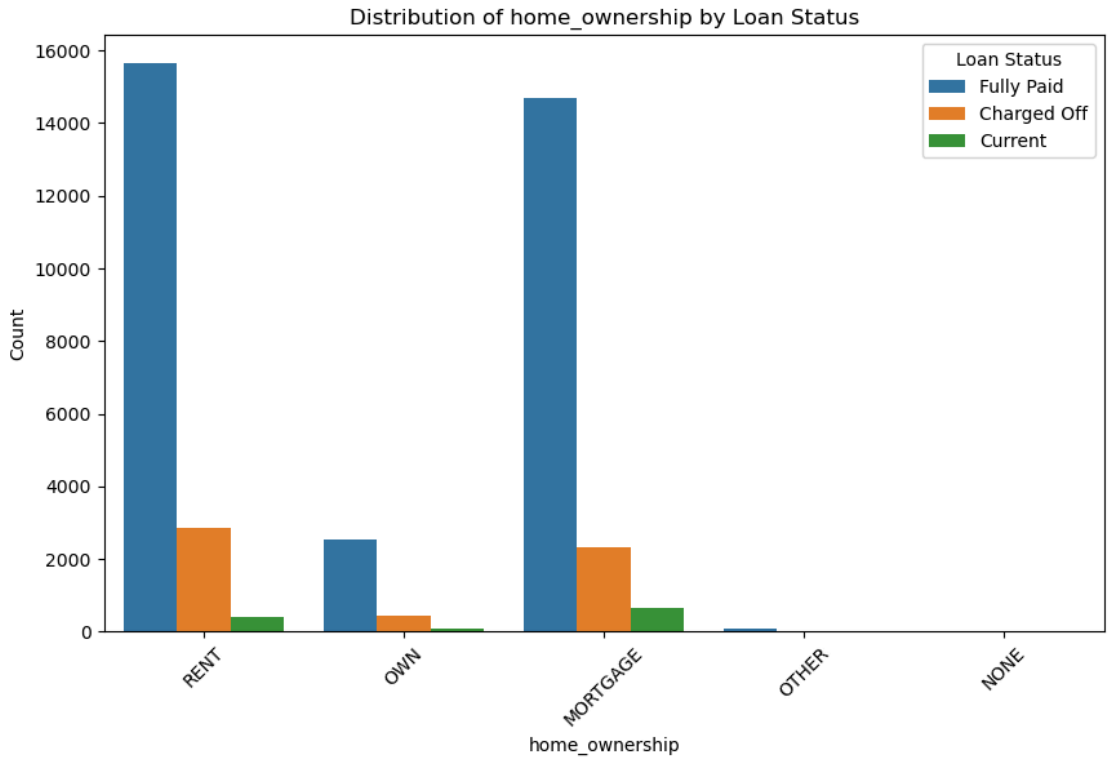
- The distribution of main zip codes shows variations in borrower locations.
- Certain zip codes may have higher concentrations of loan applicants, warranting further investigation into regional risk factors.
- Zip codes between 888 and 999 have taken more loans

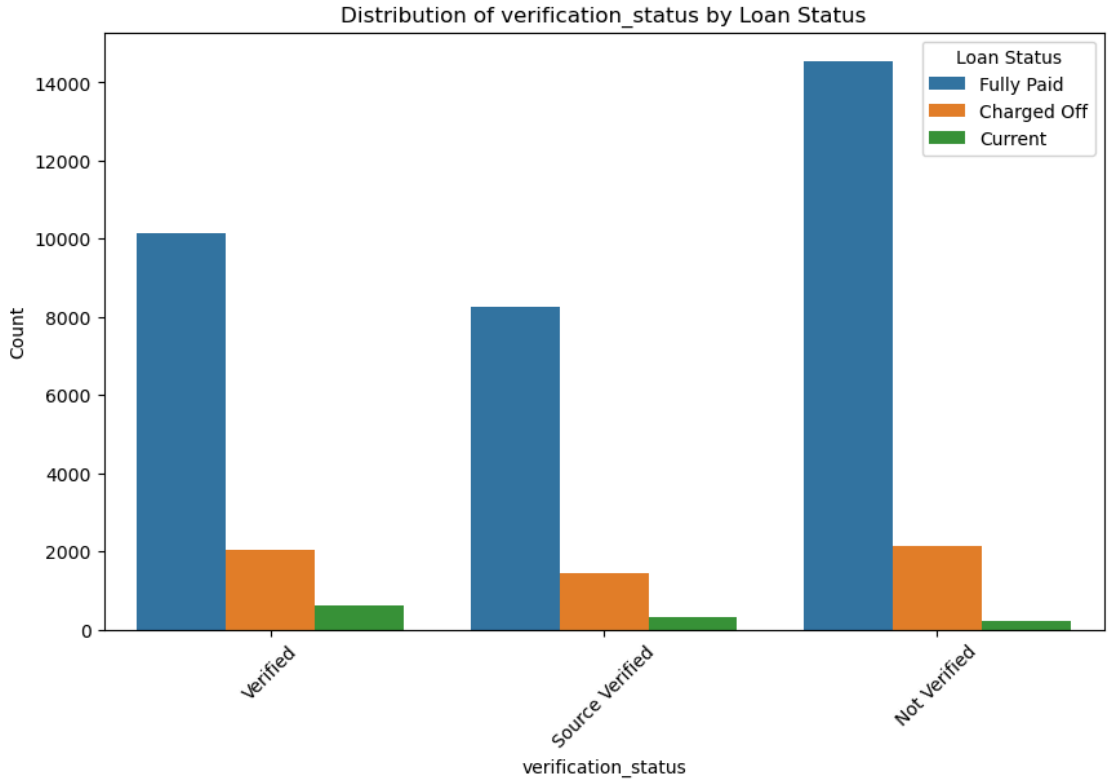
2) Bivariant Analysis

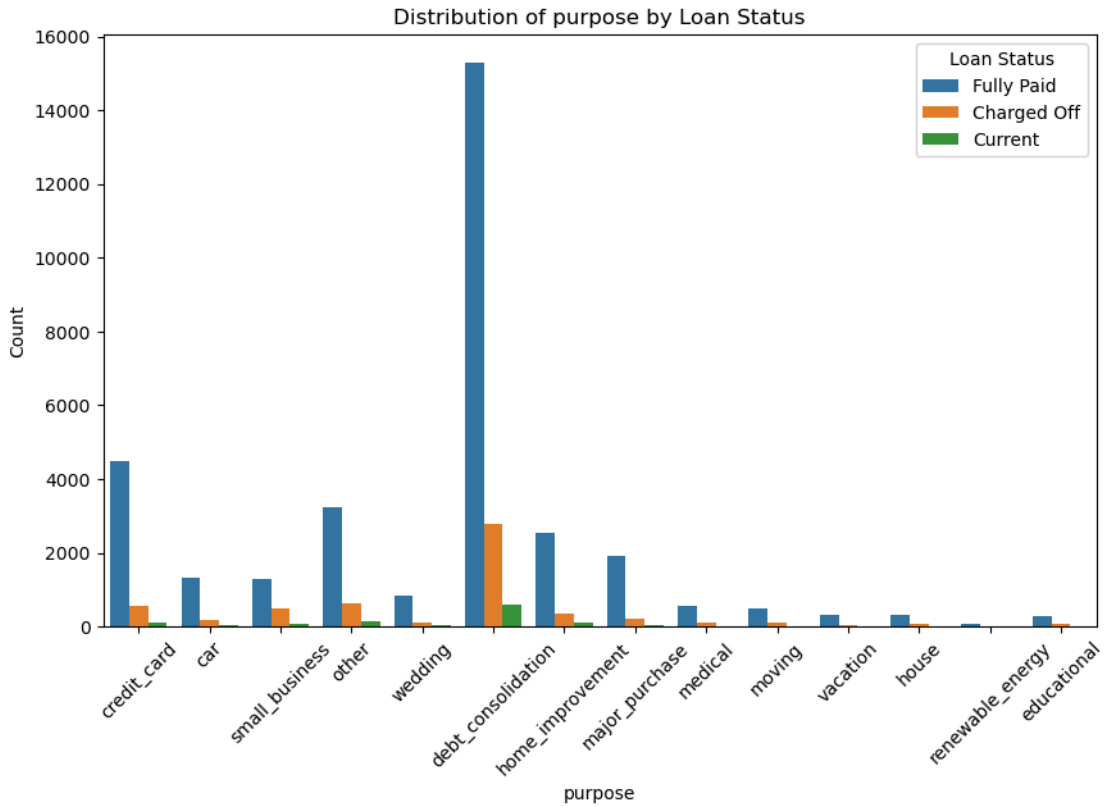
Bivariant Categorical Analysis

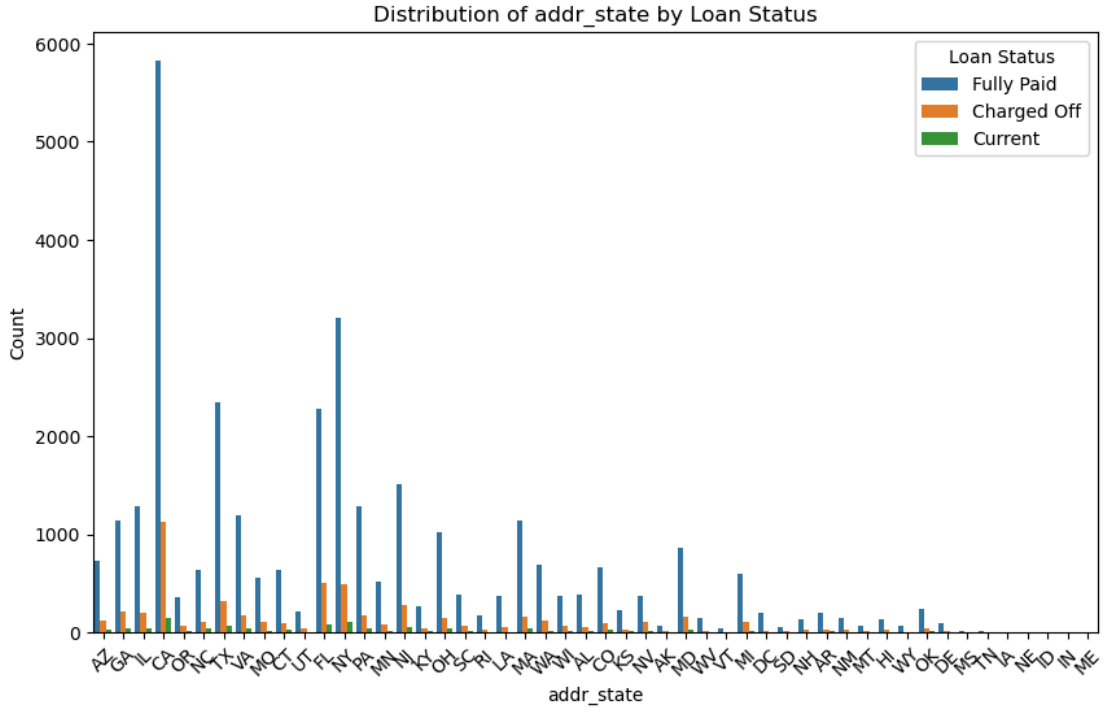

Distribution of sub_grade by Loan Status
Loan Status
Fully Paid
Charged Off
Current
Count
0
500
1000
1500
2000
2500
B2 C4 C5 C1 B5 A4 E1 F2 C3 B1 D1 A1 B3 B4 C2 D2 A3 A5 D5 A2 E4 D3 D4 F3 E3 F4 F1 E5 G4 E2 G3 G2 G1 F5 G5
sub_grade

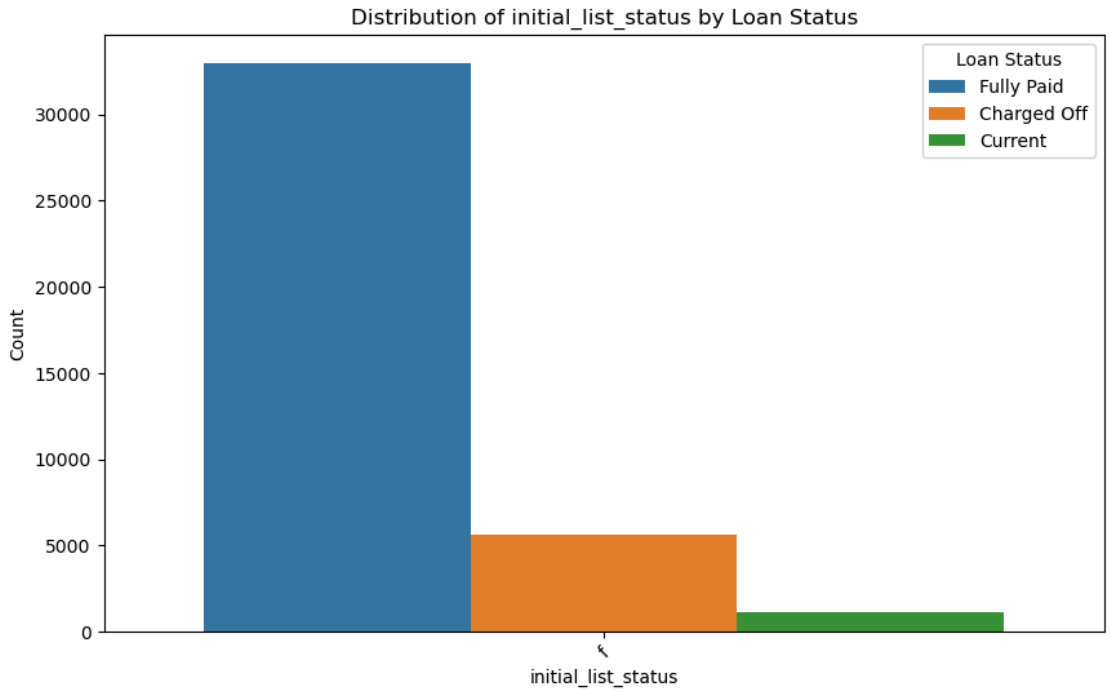

Distribution of emp_length by Loan Status
Loan Status
Fully Paid
Charged Off
Current
Count
0
1000
2000
3000
4000
5000
6000
7000
10+ years
< 1 year
1 year
3 years
8 years
9 years
4 years
5 years
6 years
2 years
7 years
emp_length

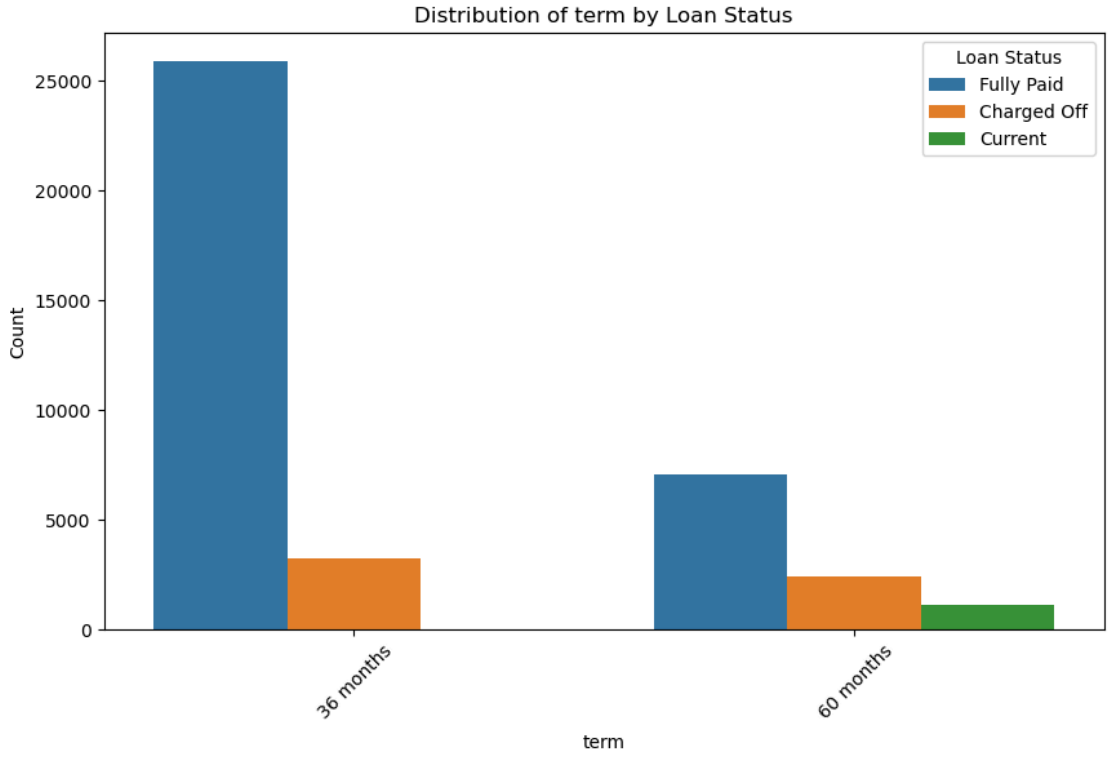

Distribution of home_ownership by Loan Status
Loan Status
Fully Paid
Charged Off
Current
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership

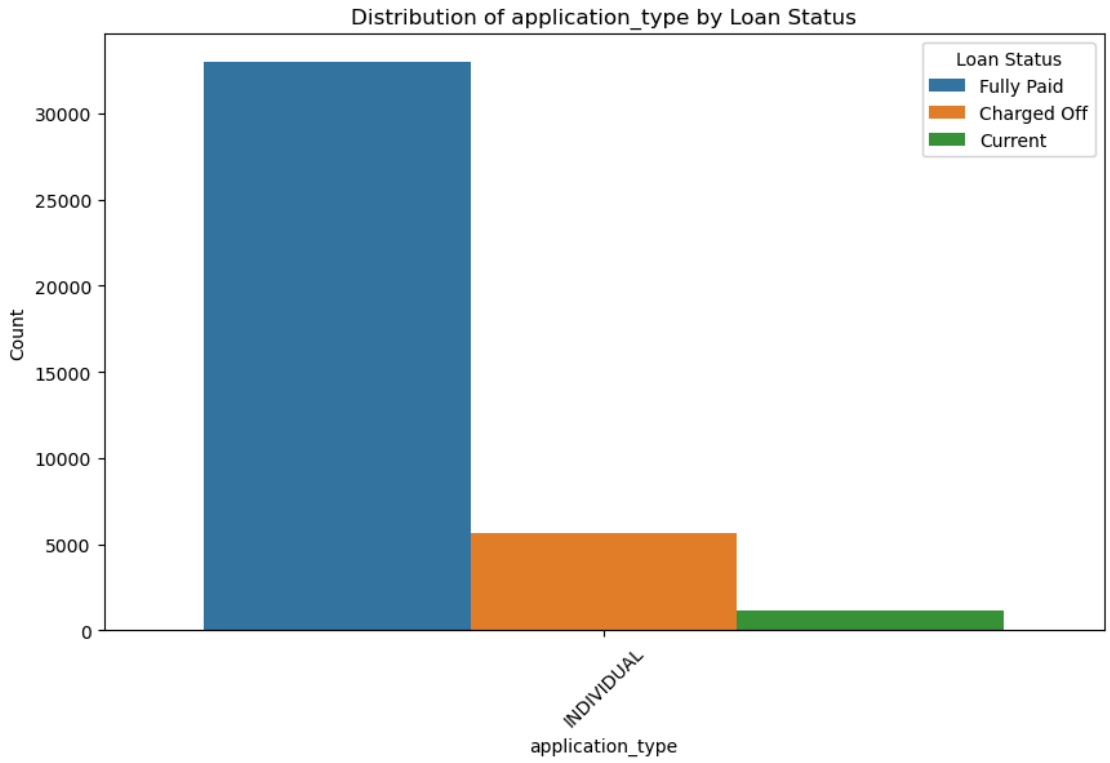

Distribution of verification_status by Loan Status
Loan Status
Fully Paid
Charged Off
Current
Count
14000
12000
10000
8000
6000
4000
2000
0
Verified
Source Verified
Not Verified
verification_status

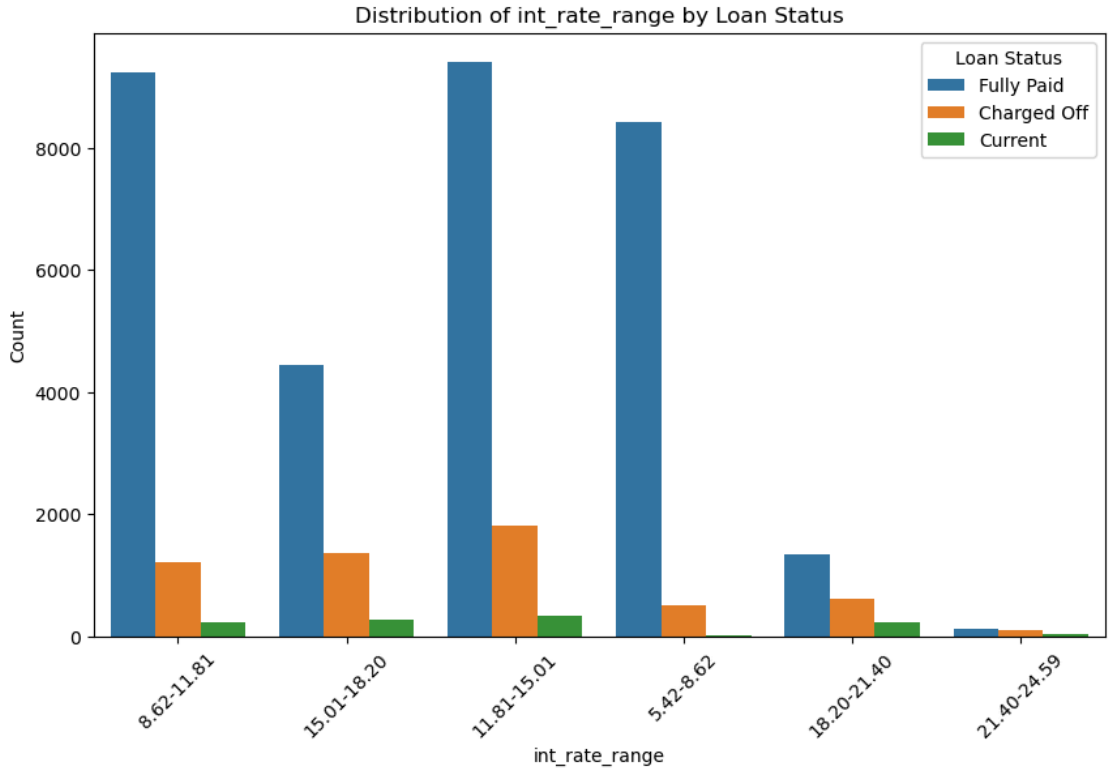

Distribution of purpose by Loan Status
Loan Status
Fully Paid
Charged Off
Current
Count
purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational

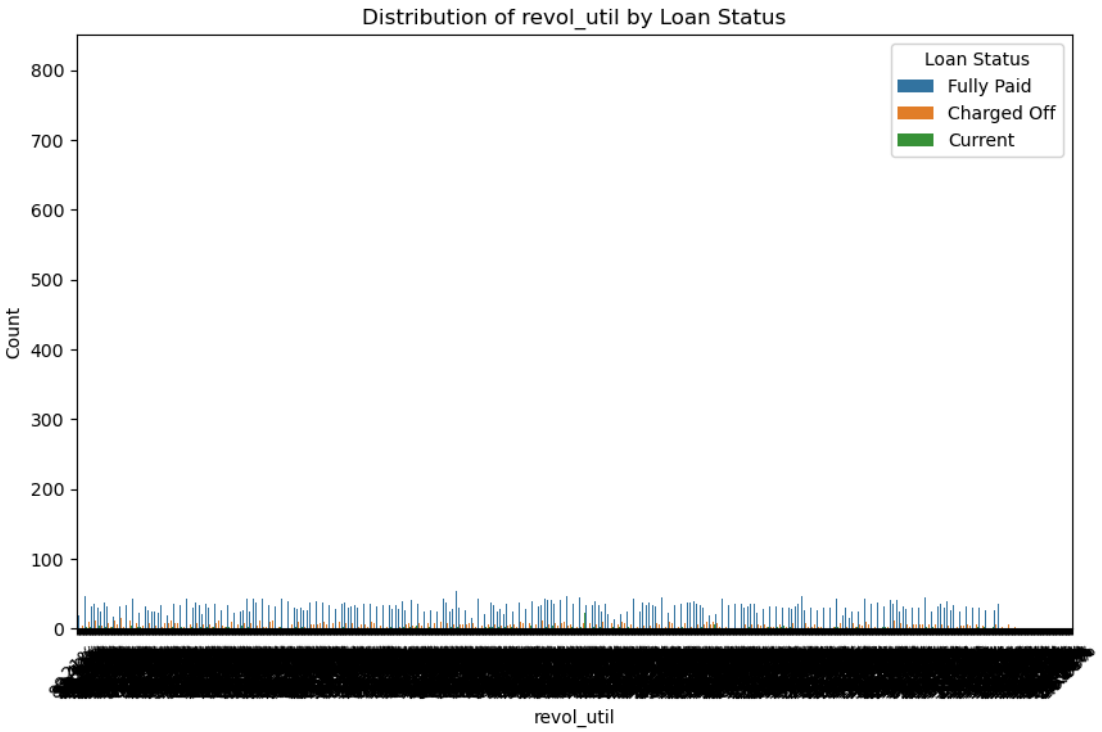

Distribution of addr_state by Loan Status
Loan Status
Fully Paid
Charged Off
Current
Count
addr_state

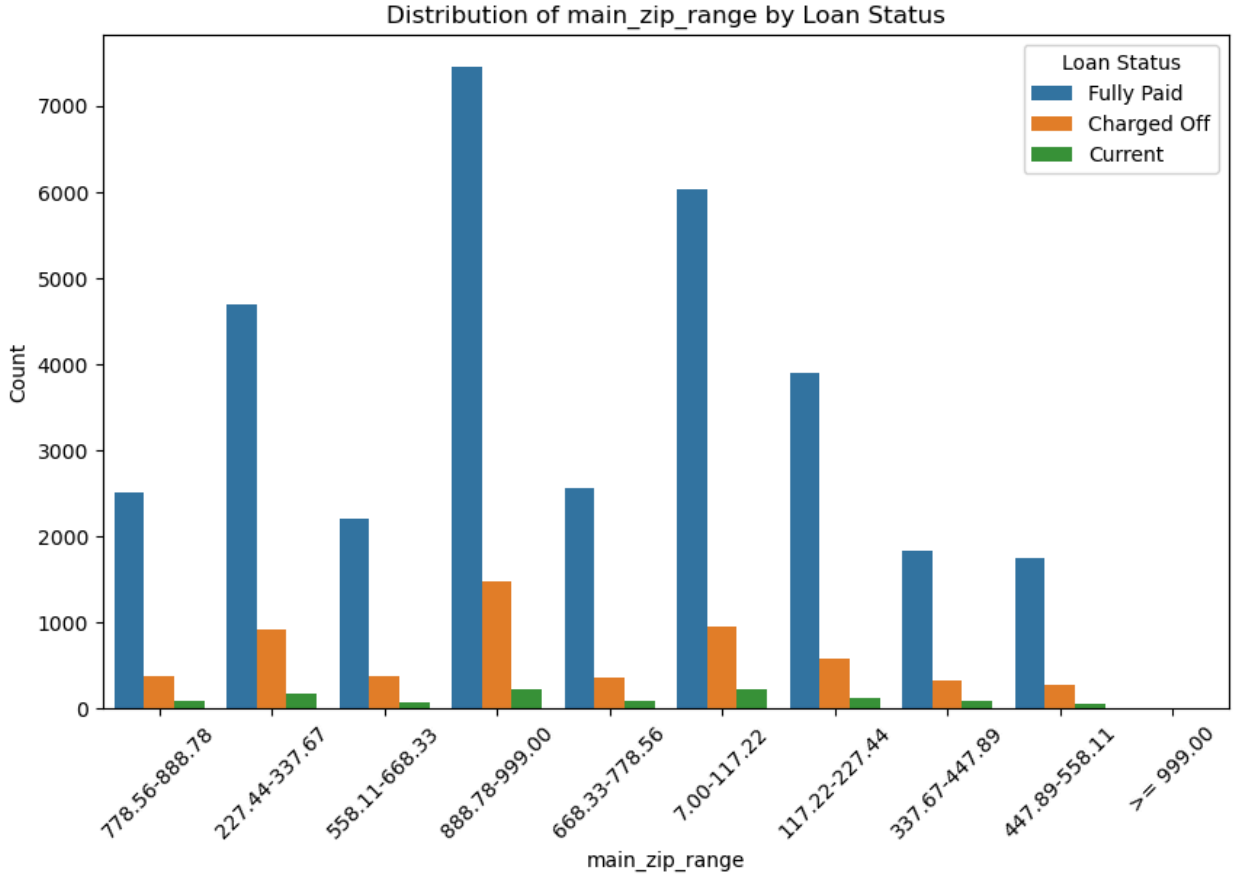

Distribution of initial_list_status by Loan Status
Loan Status
Fully Paid
Charged Off
Current
30000
25000
20000
15000
10000
5000
0
Count
initial_list_status


Distribution of term by Loan Status
Loan Status
Fully Paid
Charged Off
Current
25000
20000
15000
10000
5000
0
Count
36 months
60 months
term


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


Distribution of int_rate_range by Loan Status
Loan Status
Fully Paid
Charged Off
Current
Count
8000
6000
4000
2000
0
8.62-11.81
15.01-18.20
11.81-15.01
5.42-8.62
18.20-21.40
21.40-24.59
int_rate_range

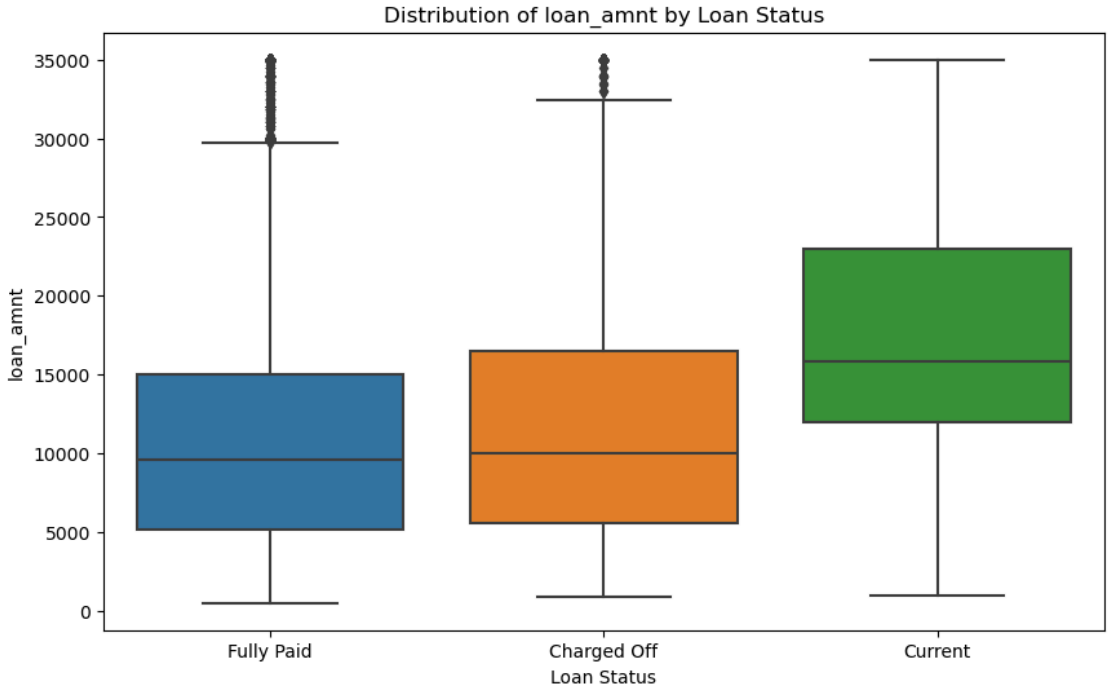

Distribution of revol_util by Loan Status
Loan Status
Fully Paid
Charged Off
Current
Count
800
700
600
500
400
300
200
100
0
revol_util


Distribution of main_zip_range by Loan Status
Loan Status
Fully Paid
Charged Off
Current
Count
7000
6000
5000
4000
3000
2000
1000
0
778.56-888.78
227.44-337.67
558.11-668.33
888.78-999.00
668.33-778.56
7.00-117.22
117.22-227.44
337.67-447.89
447.89-558.11
>= 999.00
main_zip_range

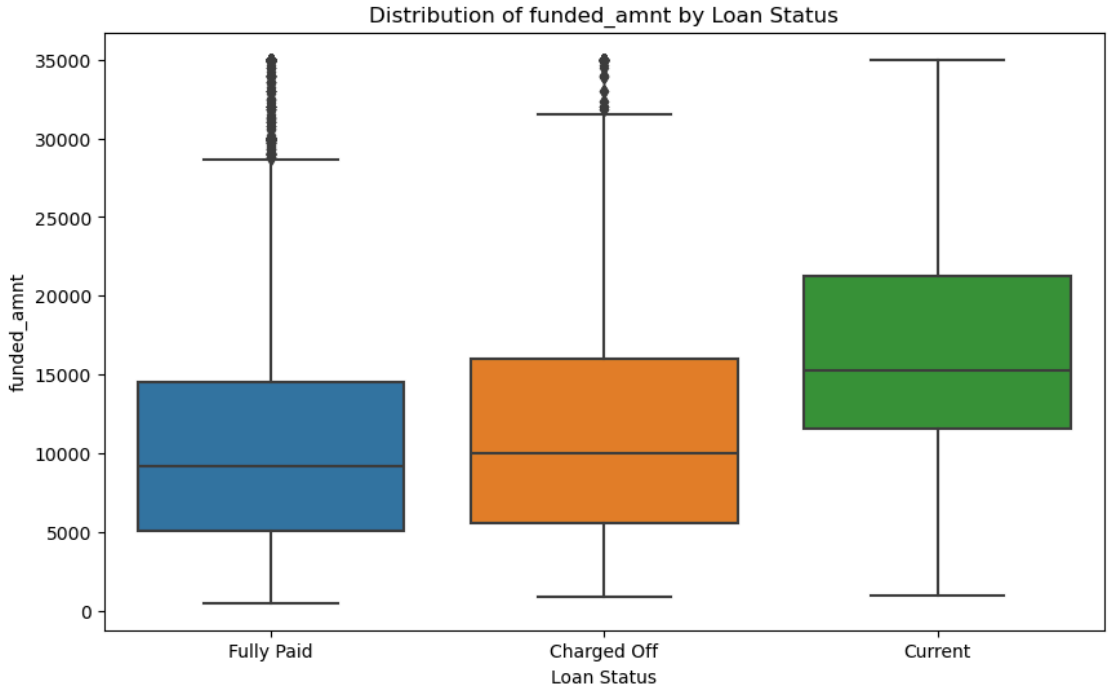
Bivariant Categorical Analysis Inferences

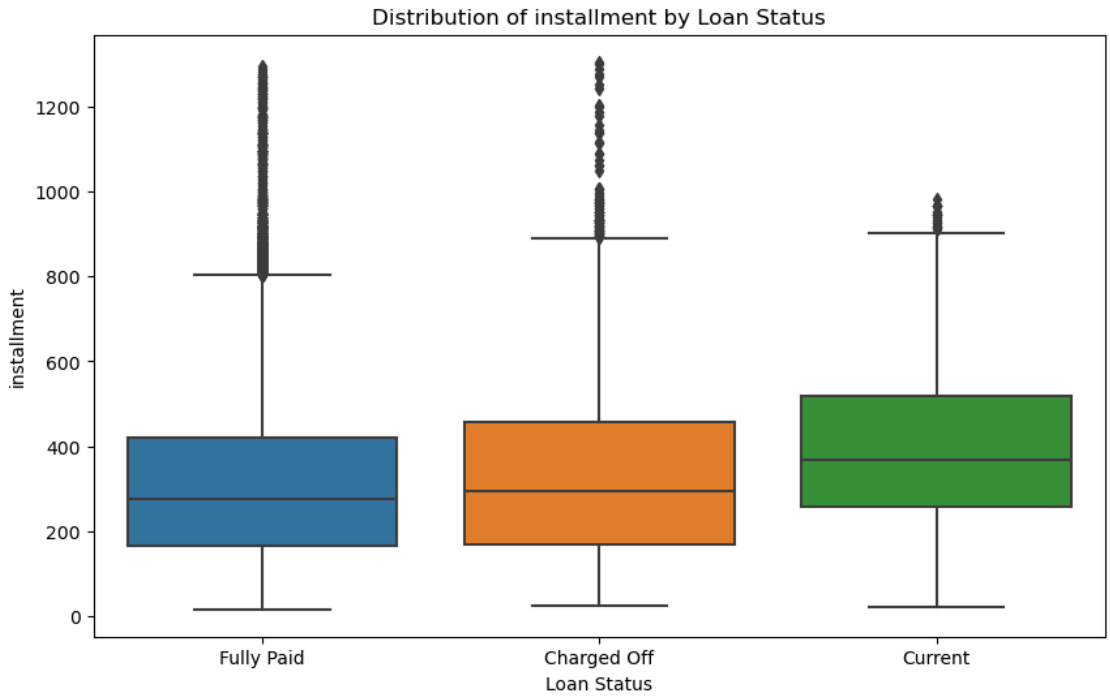
-
- **Grade and Sub-grade:**
  - Higher grade loans (A, B) have a lower proportion of charged-off loans compared to lower grade loans (C, D).
  - Similarly, within each grade, lower sub-grades tend to have higher proportions of charged-off loans.
- **Employment Length:**
  - Borrowers with longer employment lengths (e.g., 10+ years) tend to have a lower proportion of charged-off loans compared to those with shorter employment lengths.
- **Home Ownership:**
  - Borrowers who rent their homes have a slightly higher proportion of charged-off loans compared to those who own homes or have a mortgage.
- **Verification Status:**
  - Loans with verified income tend to have a slightly lower proportion of charged-off loans compared to those with income not verified.
- **Purpose:**
  - Debt consolidation is the most common purpose for loans, and it also has a higher proportion of charged-off loans compared to other purposes.
- **Address State:**
  - The distribution of loan status across states varies, CA and NY seems to have higher charged off.
- **Initial List Status:**
  - Loans with an initial list status of 'f' (whole) have a slightly higher proportion of charged-off loans compared to those with a status of 'w' (fraction).
- **Term:**
  - Loans with a term of 60 months have a slightly higher proportion of charged-off loans compared to those with a term of 36 months.
- **Application Type:**
  - Individual applications have a lower proportion of charged-off loans compared to joint applications.
- **Interest Rate (int_rate):**
  - Loans with higher interest rates tend to have a higher proportion of charged-off loans.
- **Revolving Line Utilization Rate (revol_util):**
  - Loans with higher revolving line utilization rates tend to have a slightly higher proportion of charged-off loans.
  - **Zip Code Range(main_zip_range):**
  - Zip code with range of 227 and 337 have higher range of paying loans, so borrowers applying in this range can be given loans
  -

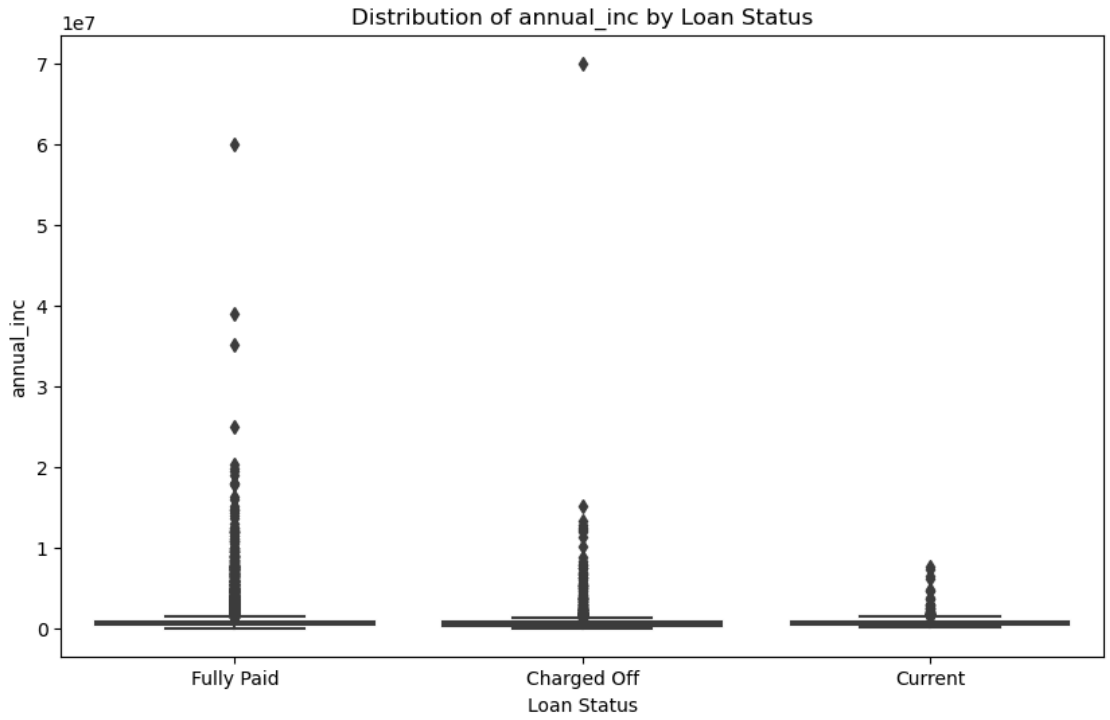
Bivariant Numerical Analysis

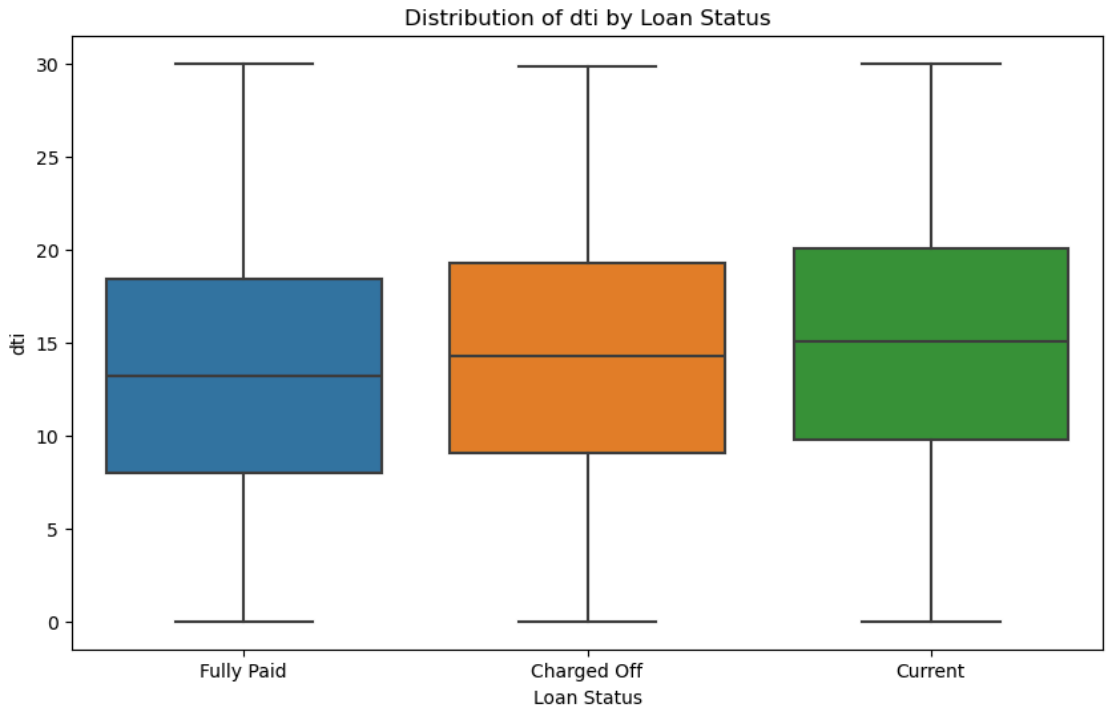


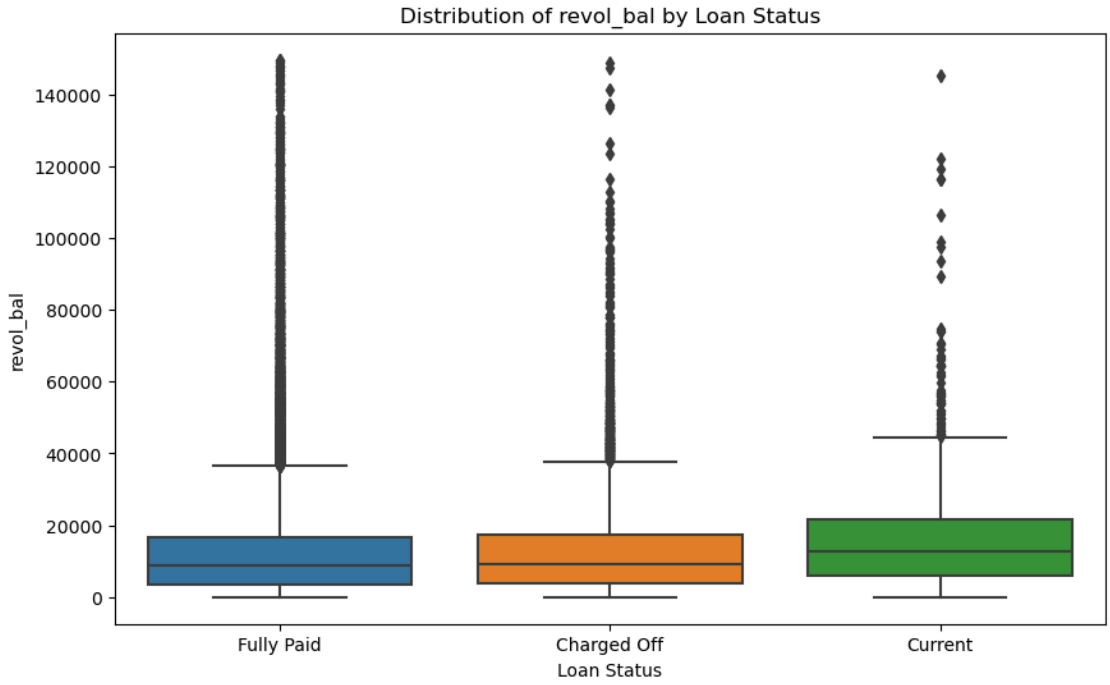
#

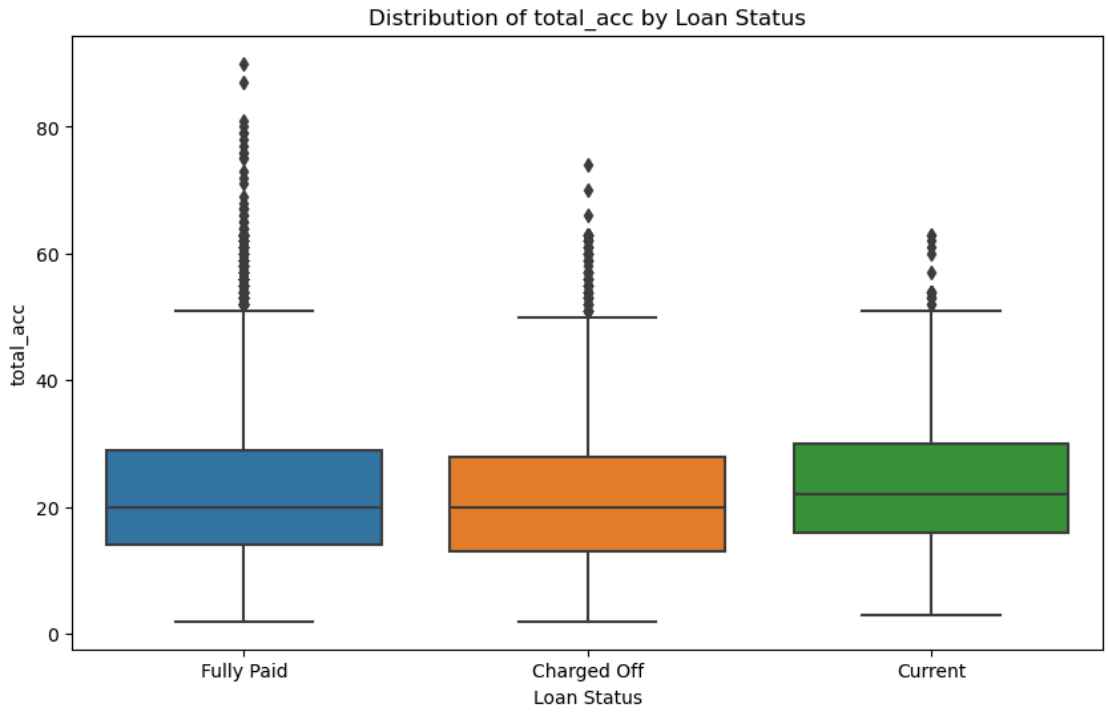


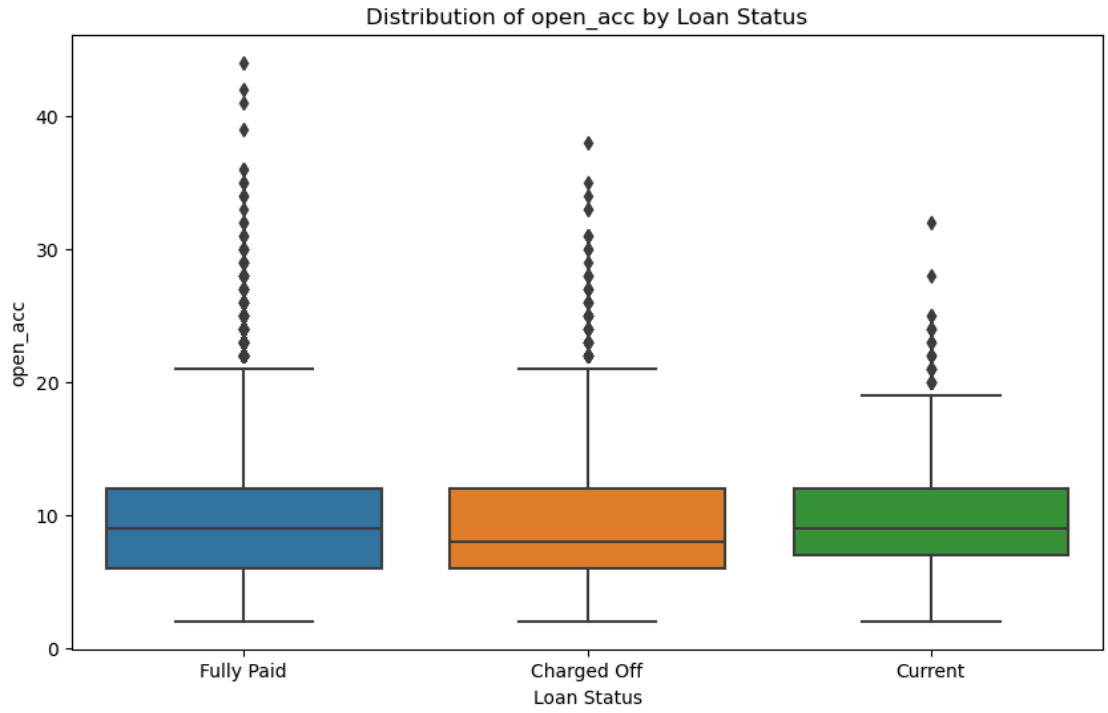

Distribution of installment by Loan Status
installment
1200
1000
800
600
400
200
0
Fully Paid
Charged Off
Current
Loan Status


Distribution of annual_inc by Loan Status
1e7
7
6
5
4
3
2
1
0
annual_inc
Fully Paid
Charged Off
Current
Loan Status


Distribution of dti by Loan Status
dti
30
25
20
15
10
5
0
Fully Paid
Charged Off
Current
Loan Status

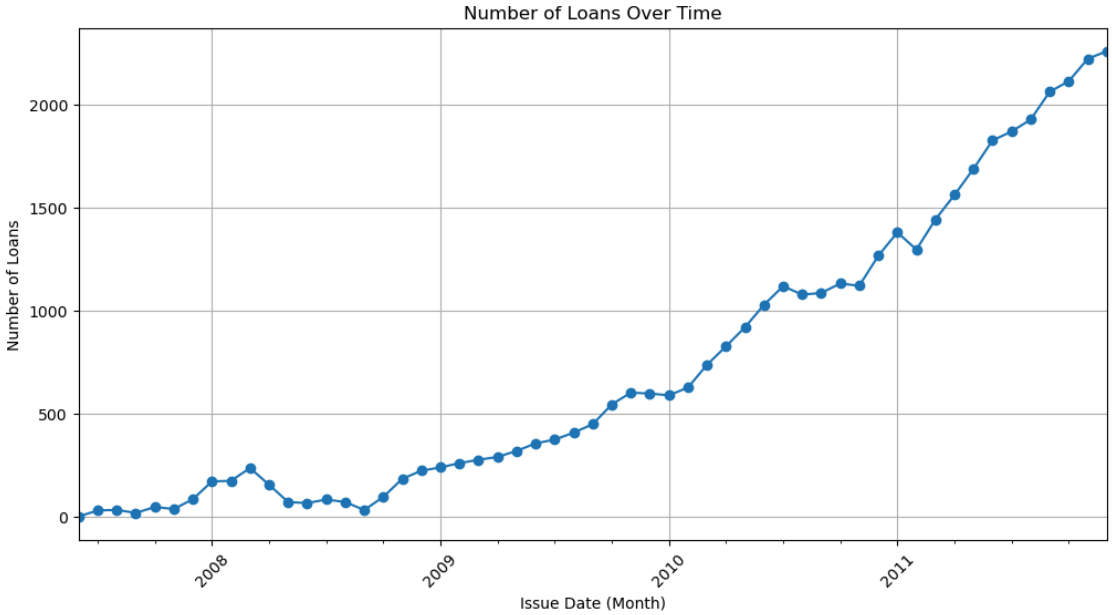

Distribution of revol_bal by Loan Status
revol_bal
140000
120000
100000
80000
60000
40000
20000
0
Fully Paid
Charged Off
Current
Loan Status

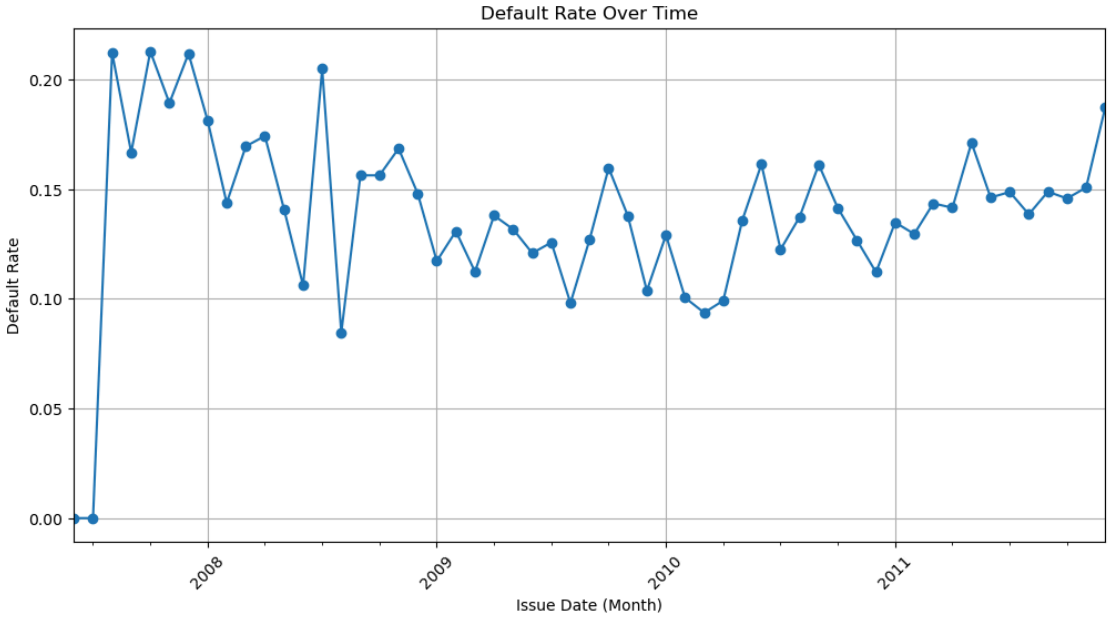

Distribution of total_acc by Loan Status
total_acc
0
20
40
60
80
Fully Paid
Charged Off
Current
Loan Status

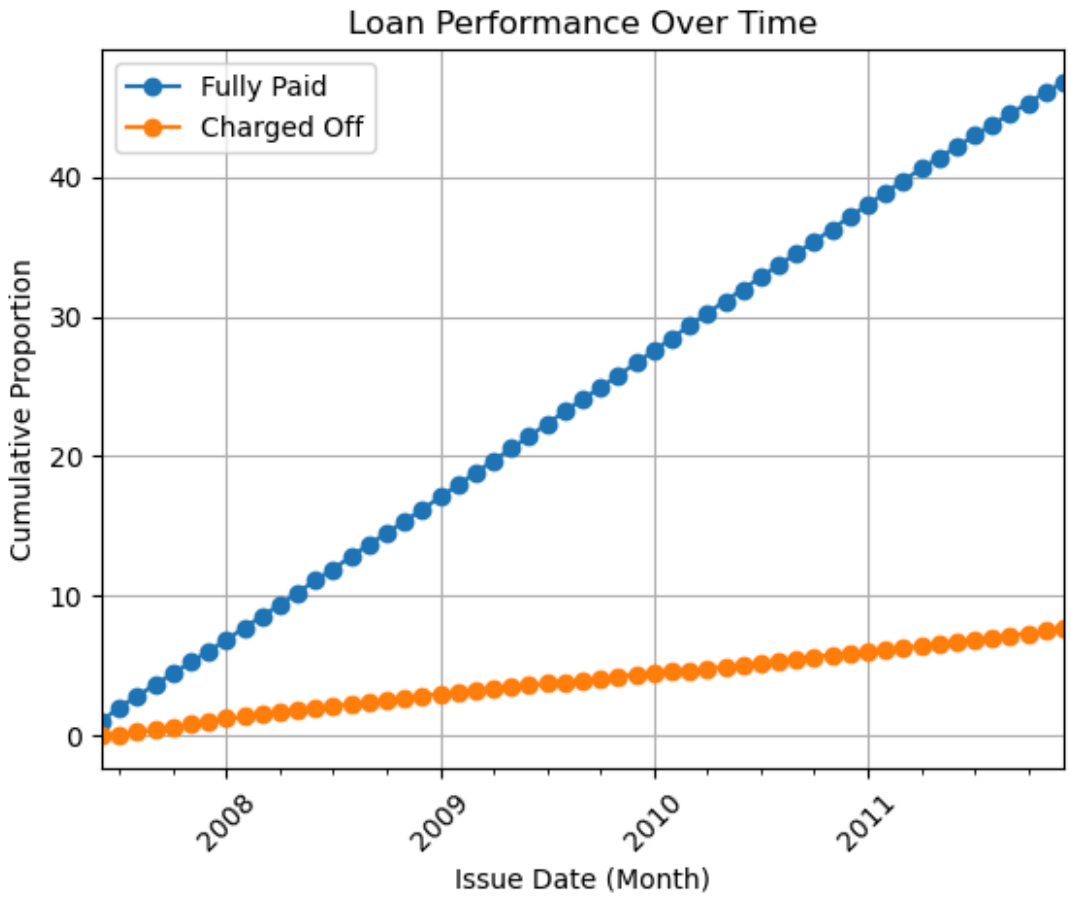

Distribution of open_acc by Loan Status
open_acc
0
10
20
30
40
Fully Paid
Charged Off
Current
Loan Status

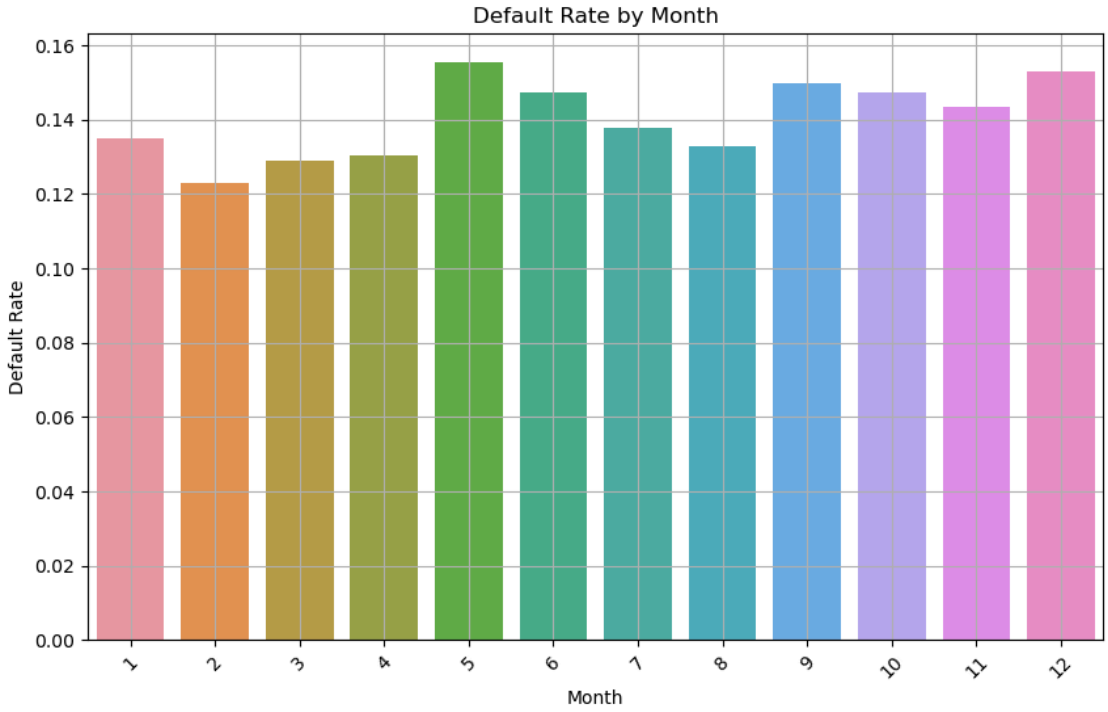
- Bivariant Numerical Inference
-
- **Loan Amount (loan_amnt):**
  - For both charged-off and fully paid loans, there is a wide range of loan amounts.
  - The median loan amount for charged-off loans appears slightly higher compared to fully paid loans, indicating that borrowers with larger loan amounts might have a higher likelihood of default.
- **Funded Amount (funded_amnt):**
  - Similar to loan amount, there is variability in funded amounts for both charged-off and fully paid loans.
  - The median funded amount for charged-off loans is also slightly higher compared to fully paid loans.
- **Monthly Installment (installment):**
  - The distribution of monthly installments is wider for charged-off loans compared to fully paid loans.
  - There are outliers with higher monthly installments in both categories, which might indicate loans with higher interest rates or longer terms.
- **Annual Income (annual_inc):**
  - Charged-off loans seem to have slightly lower median annual incomes compared to fully paid loans.
  - There is a wider spread of annual incomes for charged-off loans, suggesting that borrowers with varying income levels may default.
- **Debt-to-Income Ratio (dti):**
  - The debt-to-income ratio distributions for both loan statuses are similar, with some outliers indicating high DTI ratios.
  - There doesn't seem to be a significant difference in median DTI ratios between charged-off and fully paid loans.
- **Number of Open Credit Lines (open_acc):**
  - Charged-off loans tend to have a slightly higher median number of open credit lines compared to fully paid loans.
  - However, both distributions have a wide spread, indicating variability in the number of open credit lines.

Bivariant Time Based Analysis






Loan Performance Over Time
Fully Paid
Charged Off
Cumulative Proportion
0
10
20
30
40
2008
2009
2010
2011
Issue Date (Month)


Default Rate by Month
Default Rate
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
1
2
3
4
5
6
7
8
9
10
11
12
Month

Bivariant Time Based Inferenc

Trend Analysis - Number of Loans Over Time:

- The plot of the number of loans over time reveals an upward trend in loan volume, indicating consistent growth in lending activities throughout the analyzed period.

Default Rate Over Time:

- The default rate over time exhibits fluctuations, with periods of both higher and lower default rates. This variability suggests that external factors or shifts in lending practices may influence default rates over time.

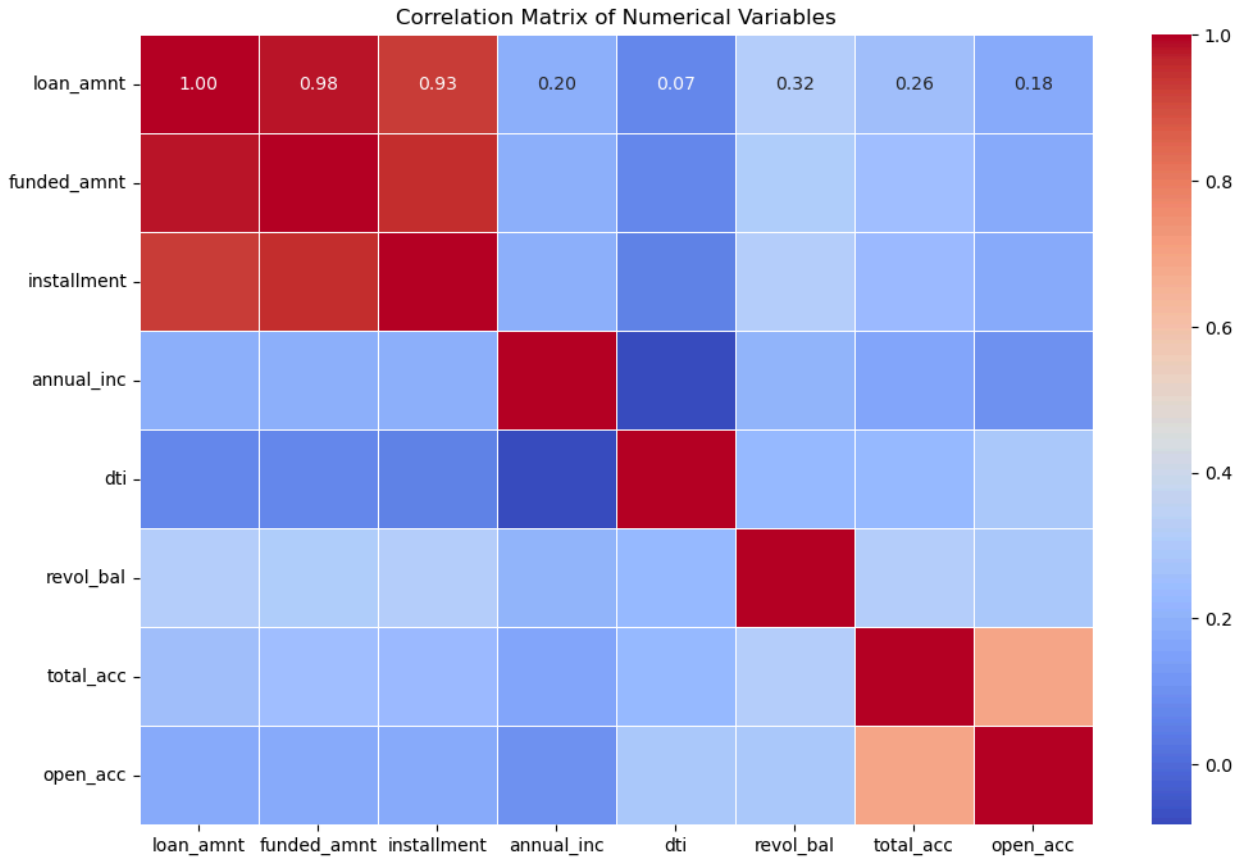
Loan Performance by Issuance Date:

- The plot illustrating loan performance over time showcases the cumulative proportion of fully paid and charged-off loans since issuance. It underscores the evolving performance of the loan portfolio, characterized by a gradual increase in fully paid loans and occasional spikes in charged-off loans.

Seasonal Analysis - Default Rate by Month:

- The analysis of default rates by month unveils potential seasonal patterns. Notably, default rates may surge during specific months of the year, possibly influenced by seasonal factors.

Bivariant Coreleation Analysis



Bivariant Coreleation Inference:

- **Loan Amount and Funded Amount:**
  - There is a strong positive correlation between loan amount and funded
    ## Bivariant Coreleation Analysis
  - amount, which is expected since funded amount is typically equal to or slightly lower than the requested loan amount.
- **Loan Amount and Installment:**
  - There is a moderate positive correlation between loan amount and installment, indicating that larger loan amounts are associated with higher installment payments.
- **Loan Amount and Annual Income:**
  - There is a weak positive correlation between loan amount and annual income, suggesting that borrowers with higher incomes may tend to apply for larger loan amounts.
- **Loan Amount and Total Balance Excluding Mortgage:**
  - There is a weak positive correlation between loan amount and total balance excluding mortgage, indicating that borrowers with higher existing balances may apply for larger loans.
  -

- **Revolving Balance and Total Credit Limit:**
  - There is a strong positive correlation between revolving balance and total credit limit, suggesting that borrowers with higher total credit limits tend to have higher revolving balances.

  Total Balance Excluding Mortgage and Annual Income:
  - There is a weak positive correlation between total balance excluding mortgage and annual income, indicating that borrowers with higher incomes may tend to have higher total balances excluding mortgage.
  -
-

- 3) Insights

- **Risk Factors Identification:** Certain borrower attributes (e.g., lower credit scores, higher debt-to-income ratios) and loan characteristics (e.g., longer terms, larger loan amounts) are associated with increased default risk.

- **Targeted Risk Mitigation:** By focusing on high-risk borrower segments and adjusting lending criteria accordingly, we can mitigate credit risk and optimize portfolio performance.

- **Enhanced Decision-Making:** Leveraging predictive models incorporating key risk factors can improve loan underwriting processes and inform strategic decisions in risk management.

- 3) Contributed By:

- Ajay Ramesh

-