Lending Club Case Study

1. Introduction

1.1 Project Overview

The primary goal of this project is to analyze loan data and build a predictive model to identify loans that are likely to be charged off. This involves data cleaning, univariate and bivariate analysis, and the development of a logistic regression model to predict loan status.Develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

1.2 Dataset Description

The dataset used in this project is sourced from a financial institution and contains various features related to loans, such as loan amount, interest rate, term, grade, employment length, home ownership status, verification status, loan status, and loan purpose.

2. Data Preprocessing

2.1 Handling Missing Values

To ensure the quality of our data, we first address missing values. Columns with more than 50% missing values are dropped, followed by the removal of rows with any missing values.

2.2 Data Cleaning

The 'int_rate' column contains percentage signs, which are removed and the column is converted to float.

3. Exploratory Data Analysis

3.1 Univariate Analysis

3.1.1 Categorical Variables

The distribution of categorical variables such as 'term', 'grade', 'sub_grade', 'emp_length', 'home_ownership', 'verification_status', 'loan_status', and 'purpose' is visualized using count plots.

3.1.2 Numerical Variables

The distribution of numerical variables such as 'loan_amnt', 'int_rate', 'installment', 'annual_inc', and 'dti' is visualized using histograms with kernel density estimates.

3.2 Bivariate Analysis

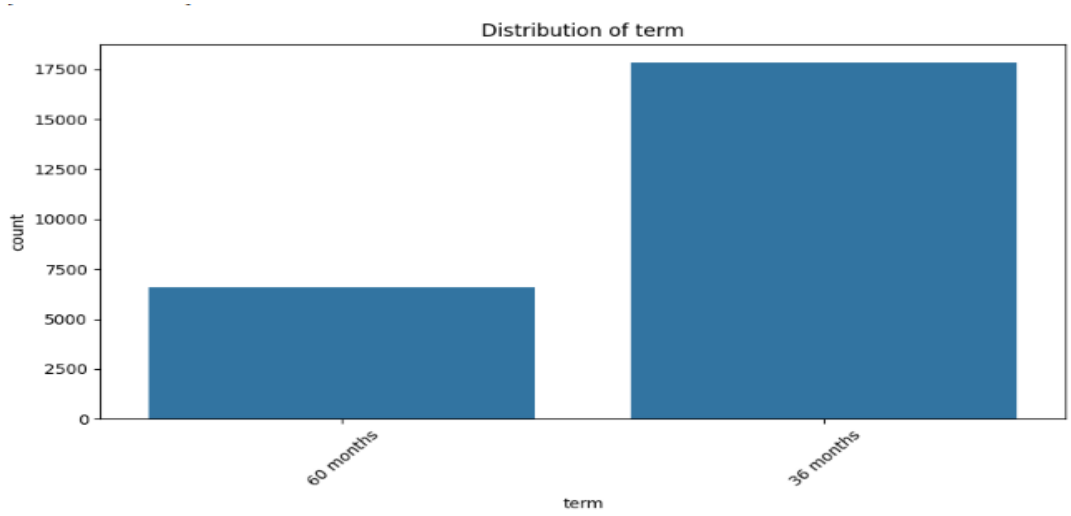
3.2.1 Categorical Variables vs Loan Status

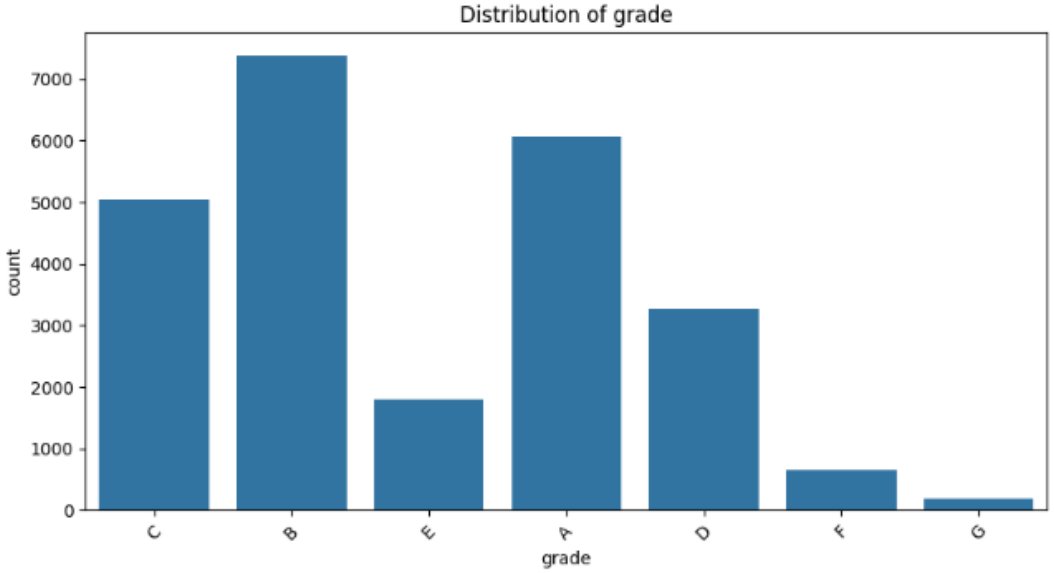
The relationship between categorical variables and loan status is analyzed using count plots with loan status as the hue.

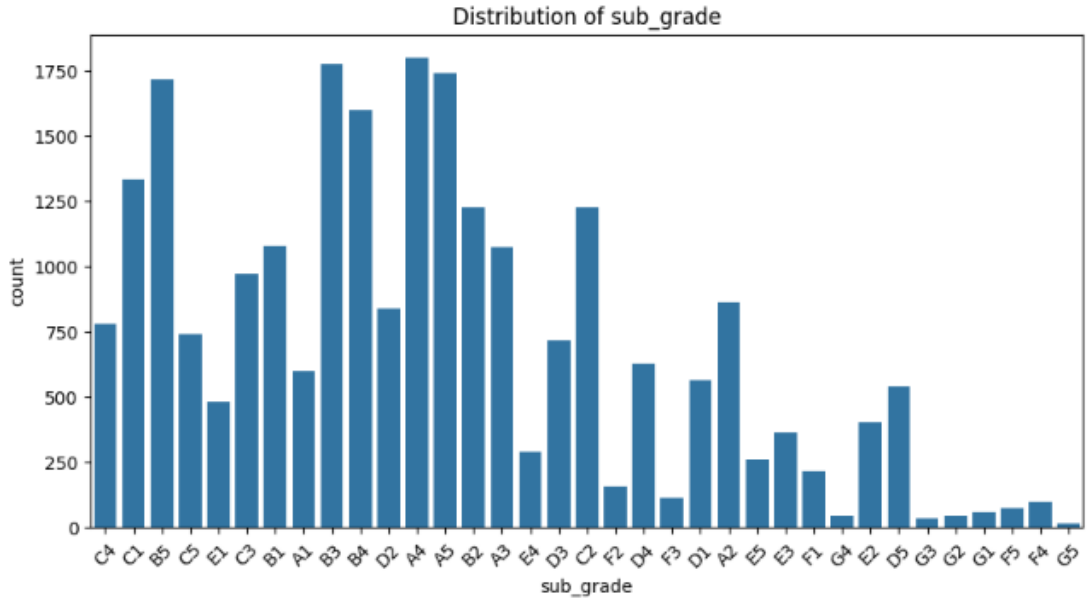
3.2.2 Numerical Variables vs Loan Status

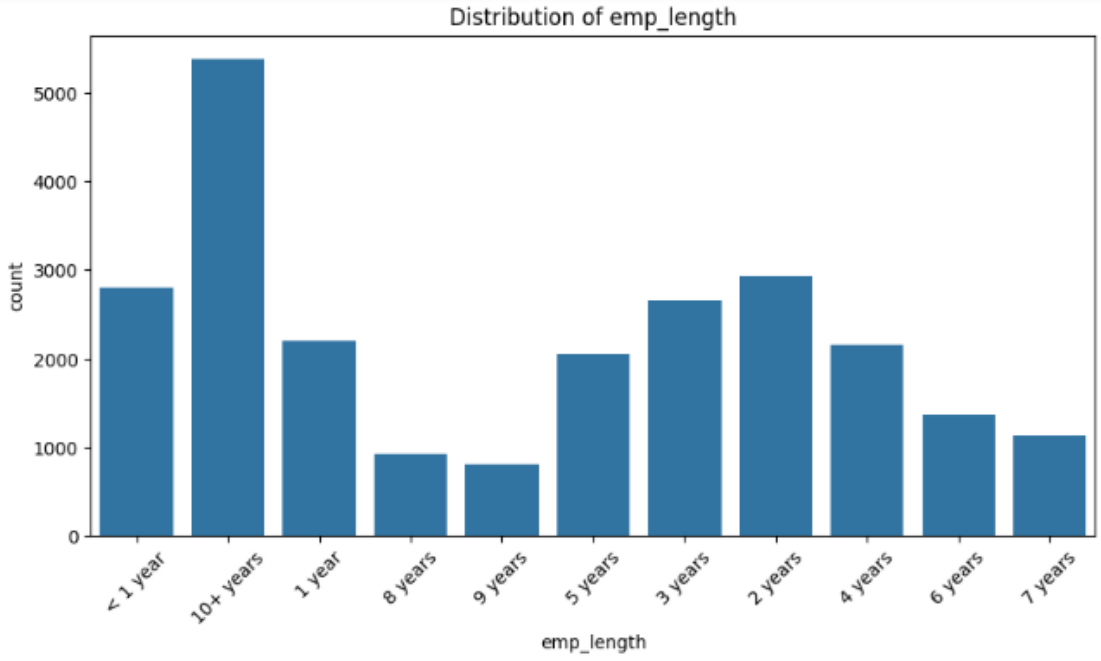
The relationship between numerical variables and loan status is analyzed using box plots.

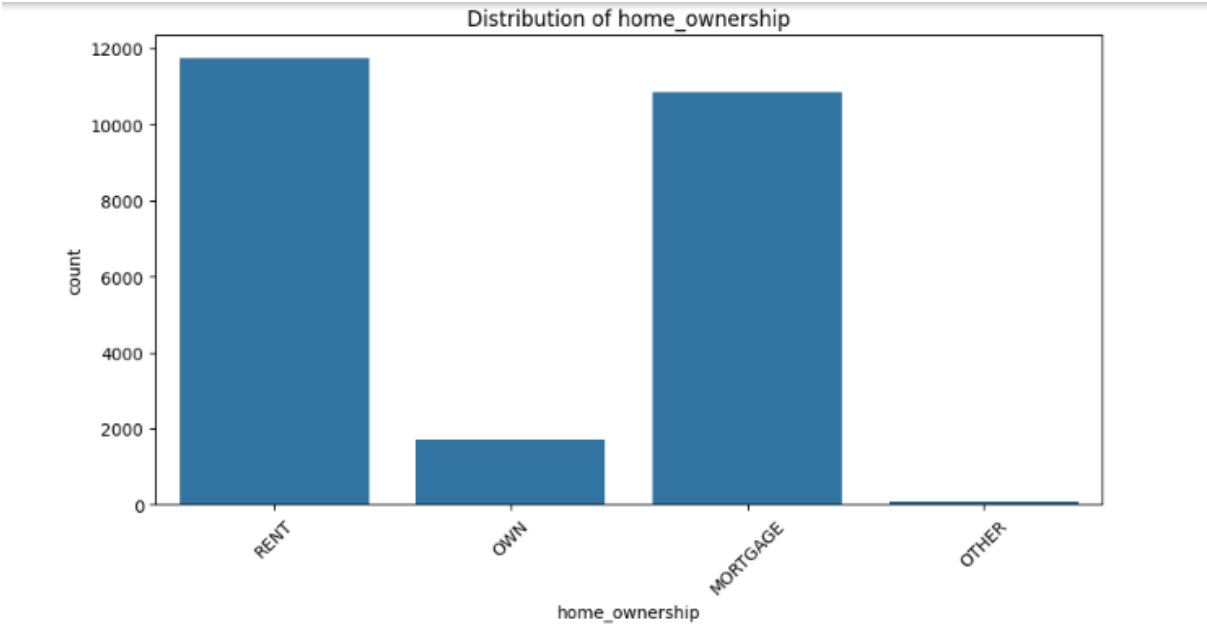
EDA Results:

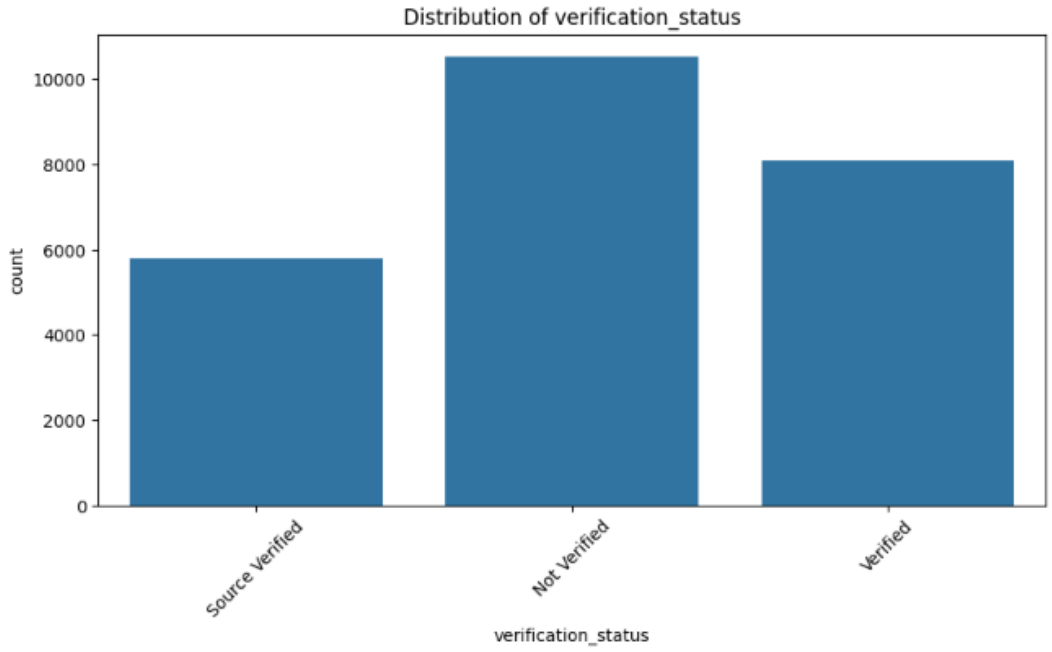


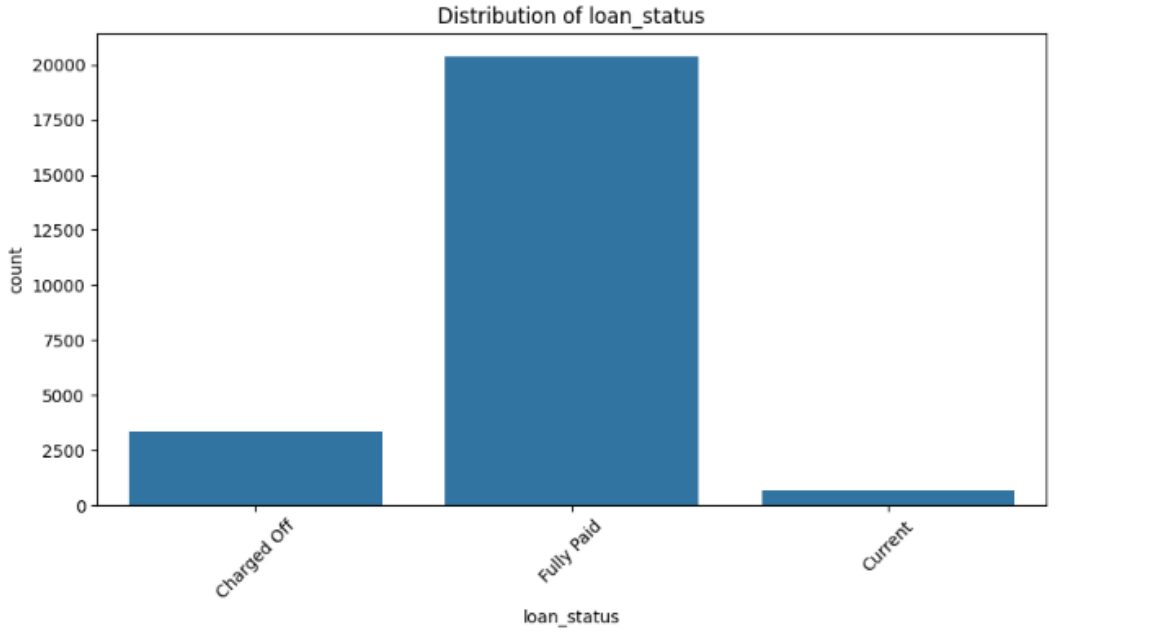


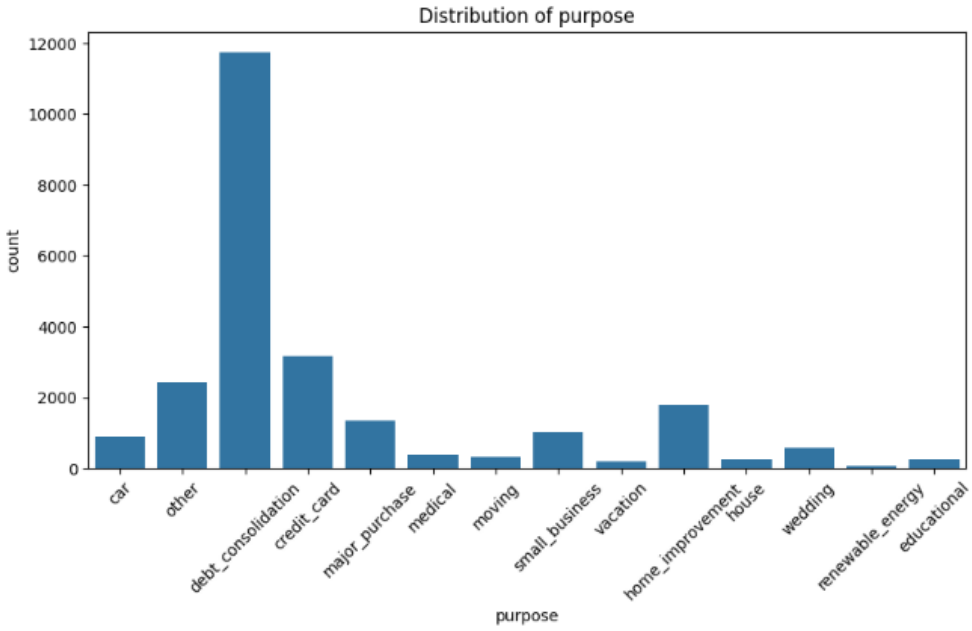

Distribution of sub_grade
count
sub_grade
0
250
500
750
1000
1250
1500
1750
C4 C1 B5 C5 E1 C3 B1 A1 B3 B4 D2 A4 A5 B2 A3 E4 D3 C2 F2 D4 F3 D1 A2 E5 E3 F1 G4 E2 D5 G3 G2 G1 F5 F4 G5

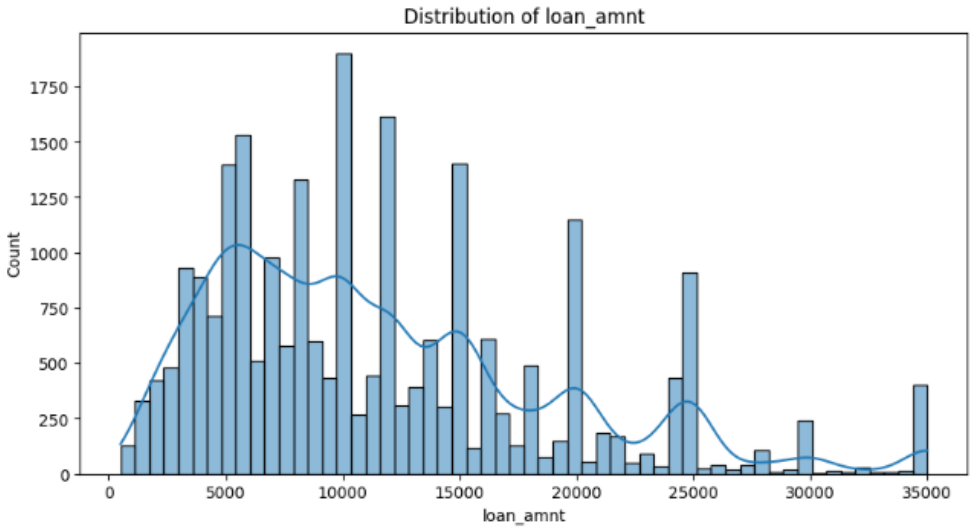

Distribution of emp_length
count
emp_length
0
1000
2000
3000
4000
5000
< 1 year
10+ years
1 year
8 years
9 years
5 years
3 years
2 years
4 years
6 years
7 years

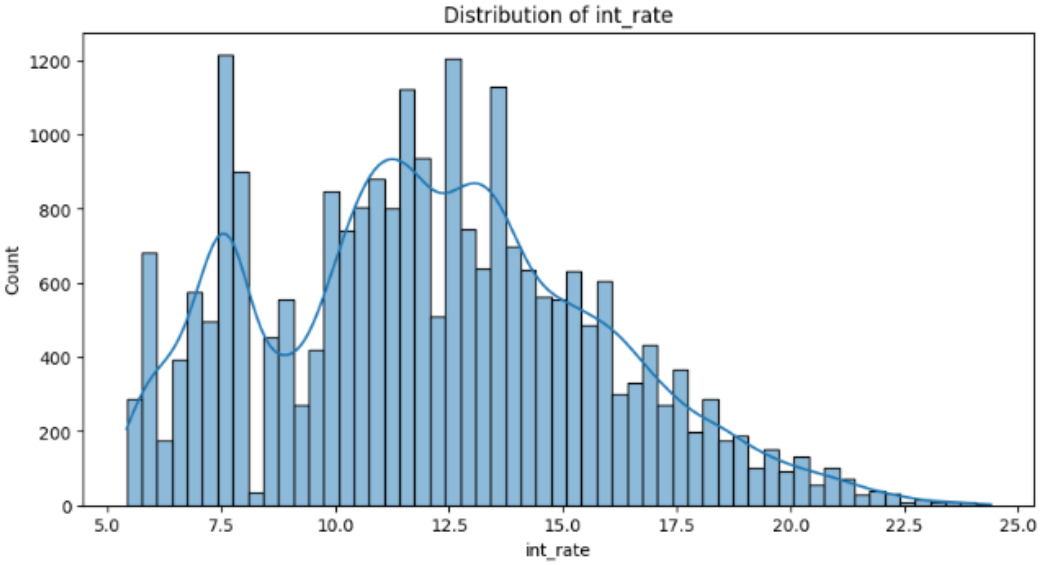

Distribution of home_ownership
12000
10000
8000
6000
4000
2000
0
count
RENT
OWN
MORTGAGE
OTHER
home_ownership

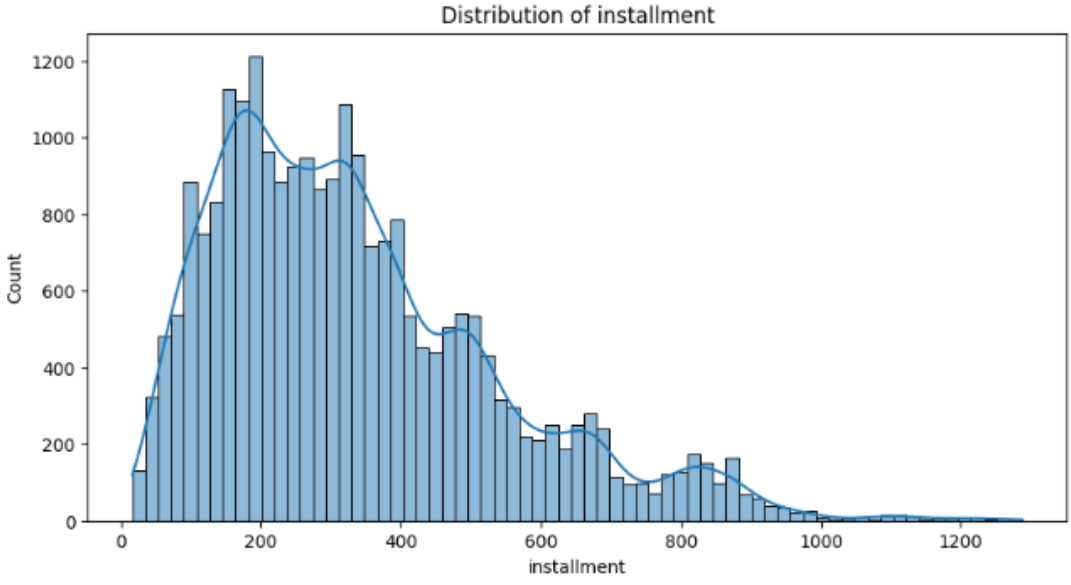

Distribution of verification_status
10000
8000
6000
4000
2000
0
count
Source Verified
Not Verified
Verified
verification_status

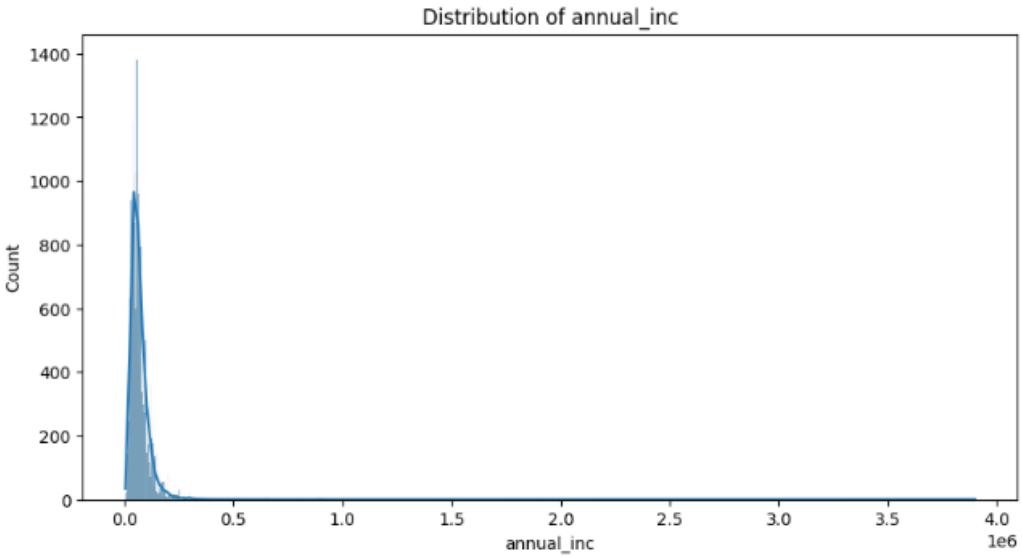

Distribution of loan_status
count
20000
17500
15000
12500
10000
7500
5000
2500
0
Charged Off
Fully Paid
Current
loan_status

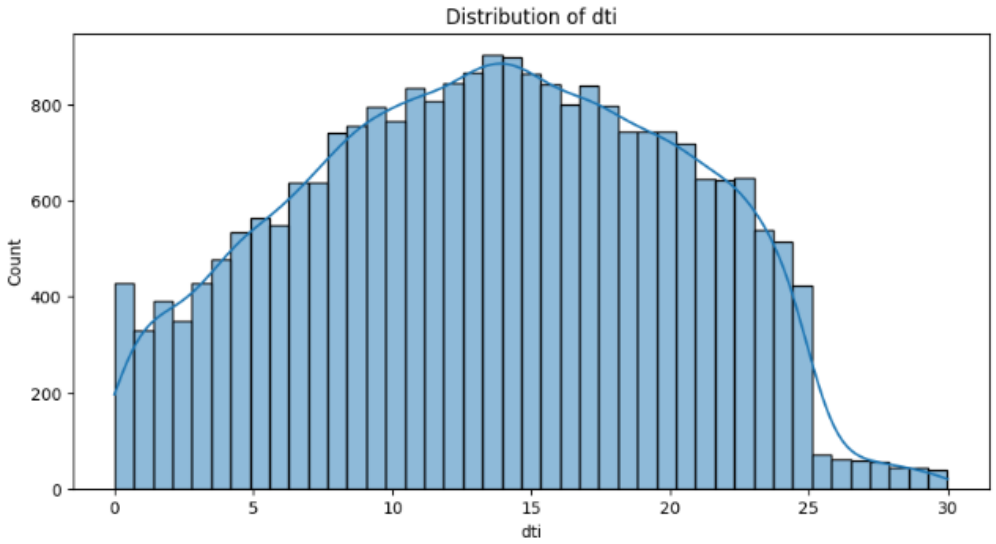

Distribution of purpose
count
12000
10000
8000
6000
4000
2000
0
car
other
debt_consolidation
credit_card
major_purchase
medical
moving
small_business
vacation
home_improvement
house
wedding
renewable_energy
educational
purpose

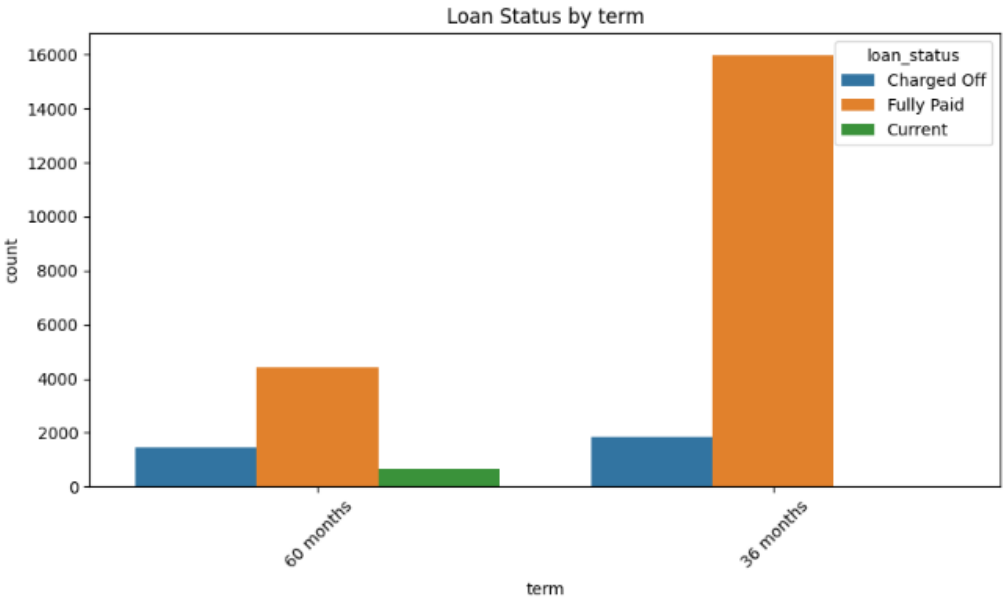

Distribution of loan_amnt
Count
loan_amnt
0
250
500
750
1000
1250
1500
1750
0
5000
10000
15000
20000
25000
30000
35000

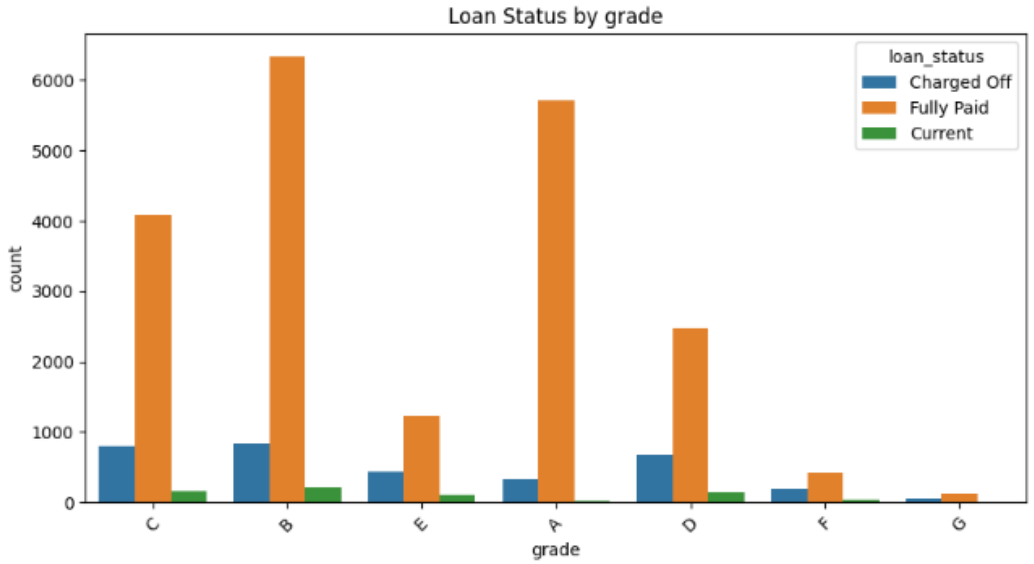

Distribution of int_rate
Count
int_rate
0
200
400
600
800
1000
1200
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0

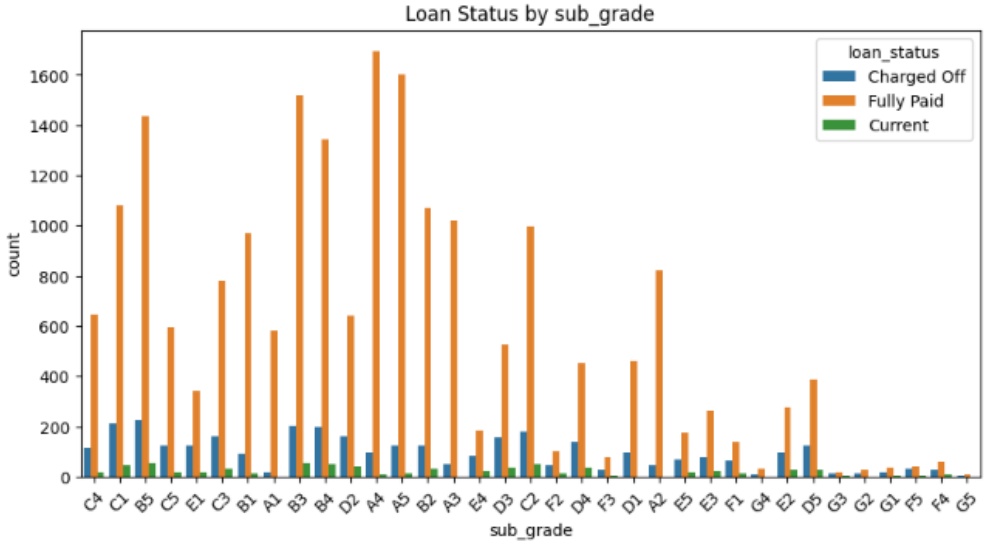

Distribution of installment
Count
installment
0
200
400
600
800
1000
1200
0
200
400
600
800
1000
1200

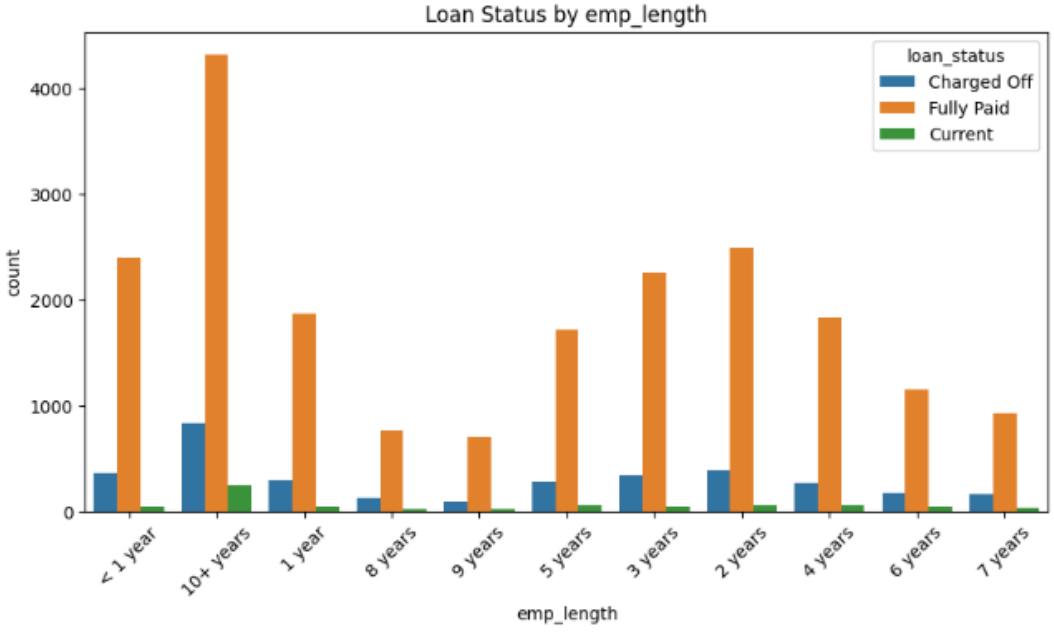

Distribution of annual_inc
Count
annual_inc
1e6

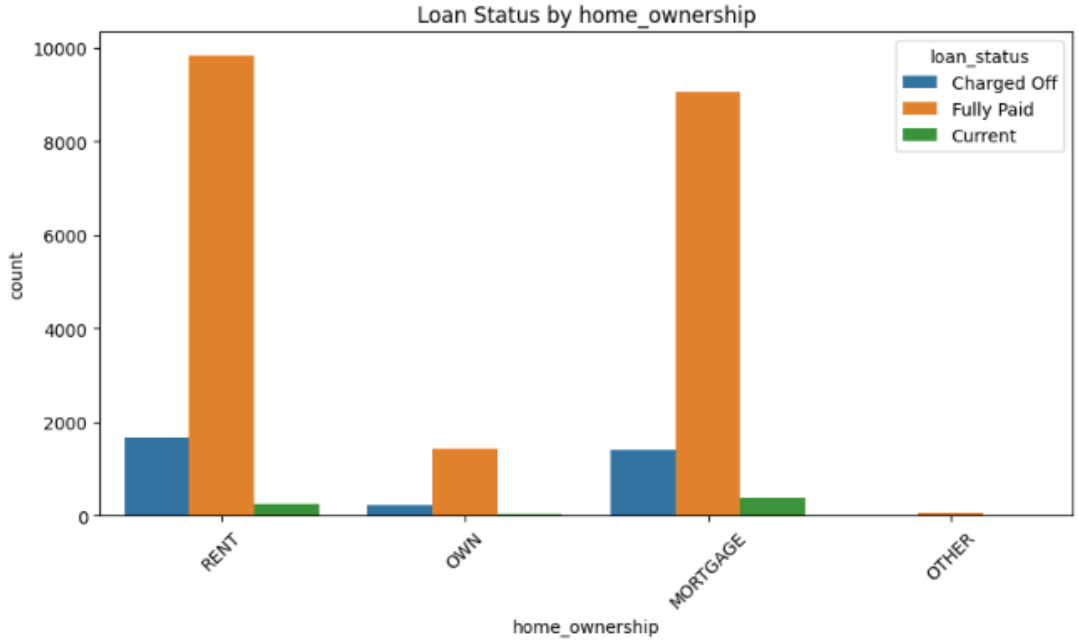

Distribution of dti
Count
dti

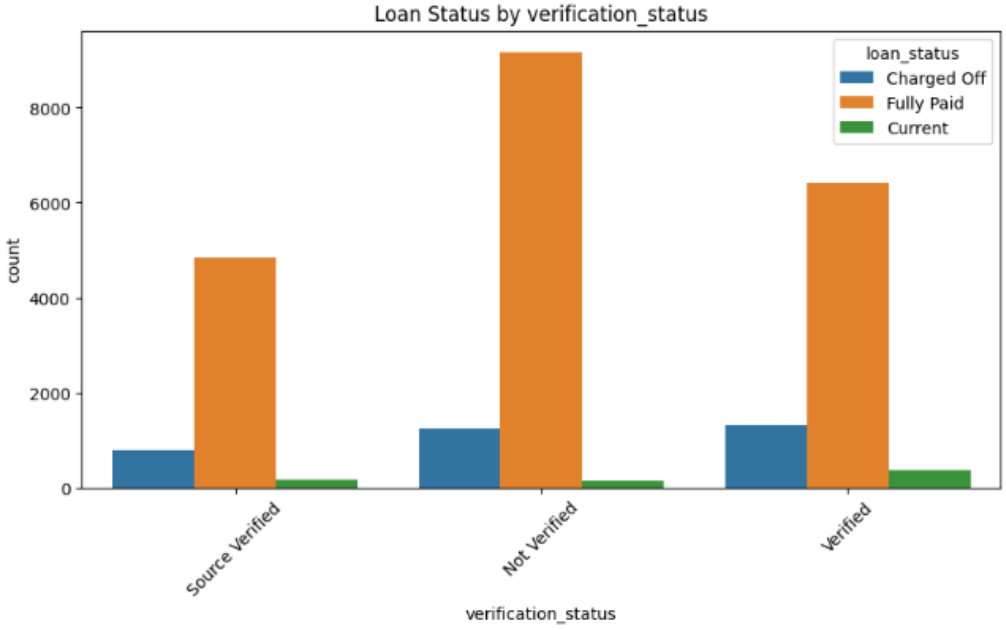

Loan Status by term
loan_status
Charged Off
Fully Paid
Current
count
60 months
36 months
term

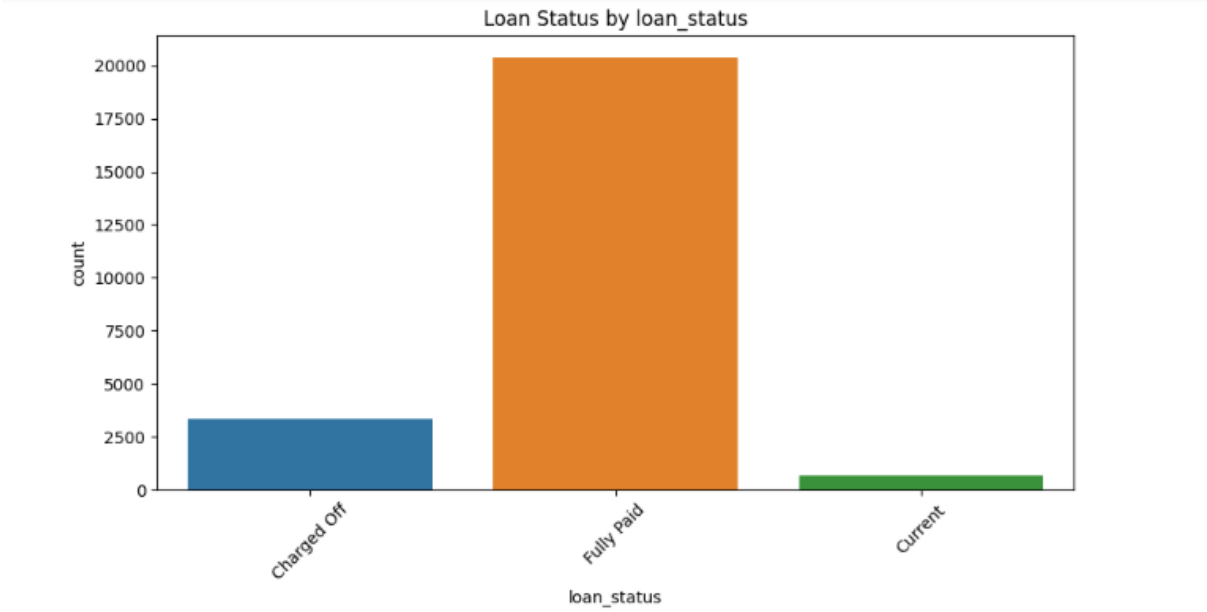

Loan Status by grade
loan_status
Charged Off
Fully Paid
Current
count
grade
C
B
E
A
D
F
G

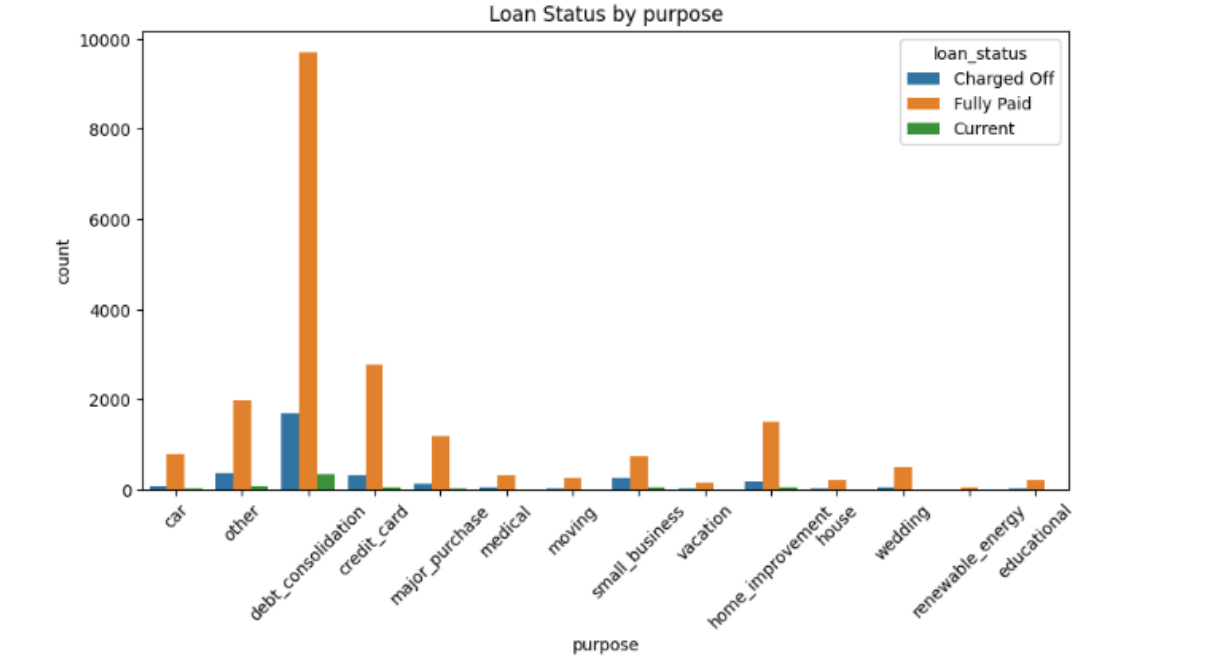

Loan Status by sub_grade
loan_status
Charged Off
Fully Paid
Current
count
sub_grade

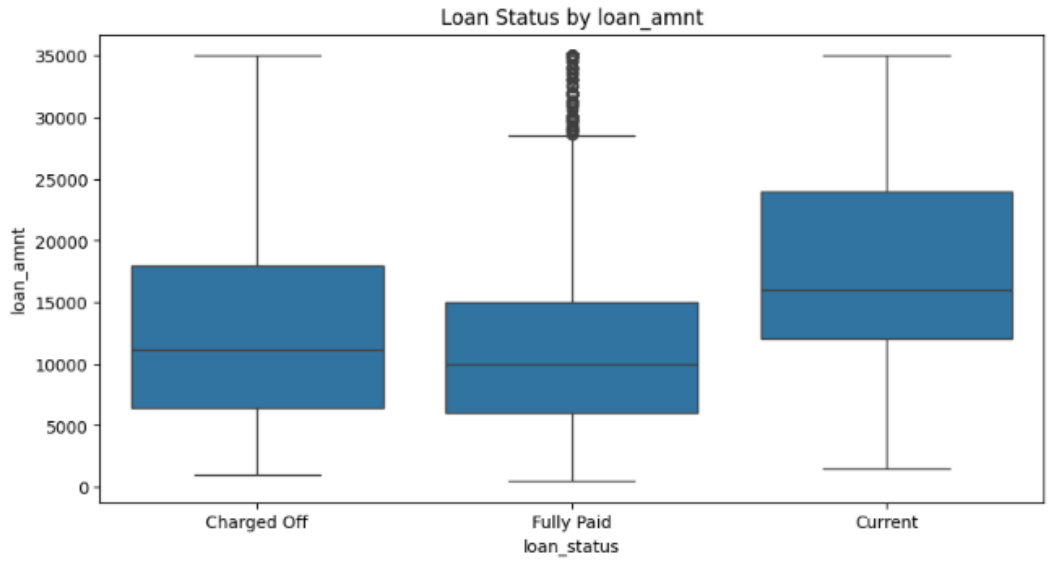

Loan Status by emp_length
loan_status
Charged Off
Fully Paid
Current
count
< 1 year
10+ years
1 year
8 years
9 years
5 years
3 years
2 years
4 years
6 years
7 years
emp_length

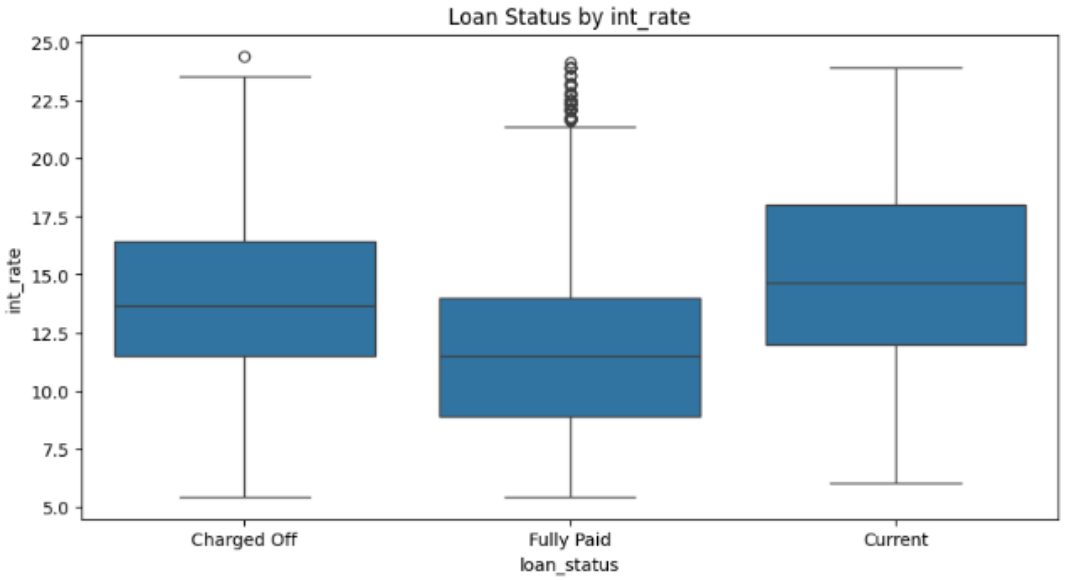

Loan Status by home_ownership
loan_status
Charged Off
Fully Paid
Current
count
10000
8000
6000
4000
2000
0
RENT
OWN
MORTGAGE
OTHER
home_ownership

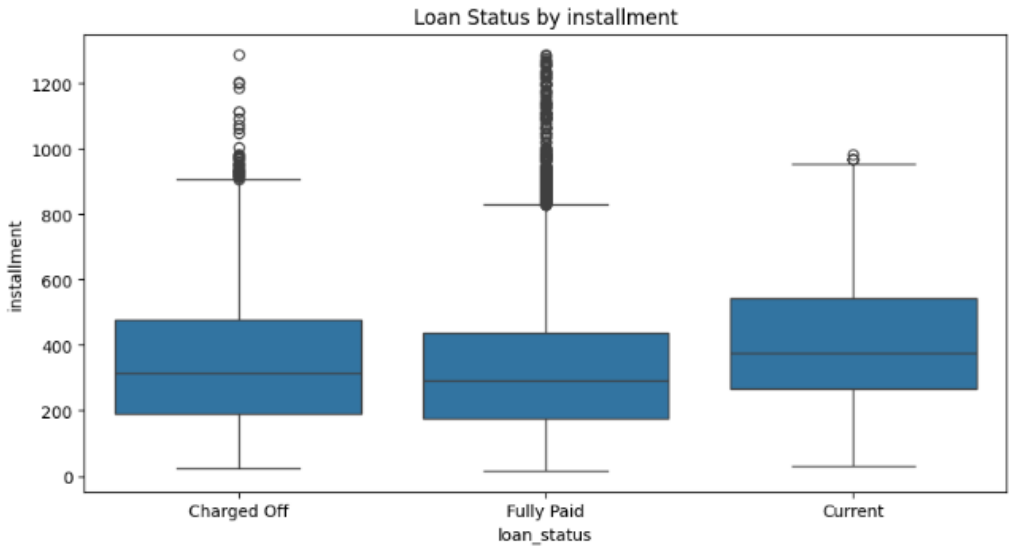

Loan Status by verification_status
loan_status
Charged Off
Fully Paid
Current
count
8000
6000
4000
2000
0
Source Verified
Not Verified
Verified
verification_status

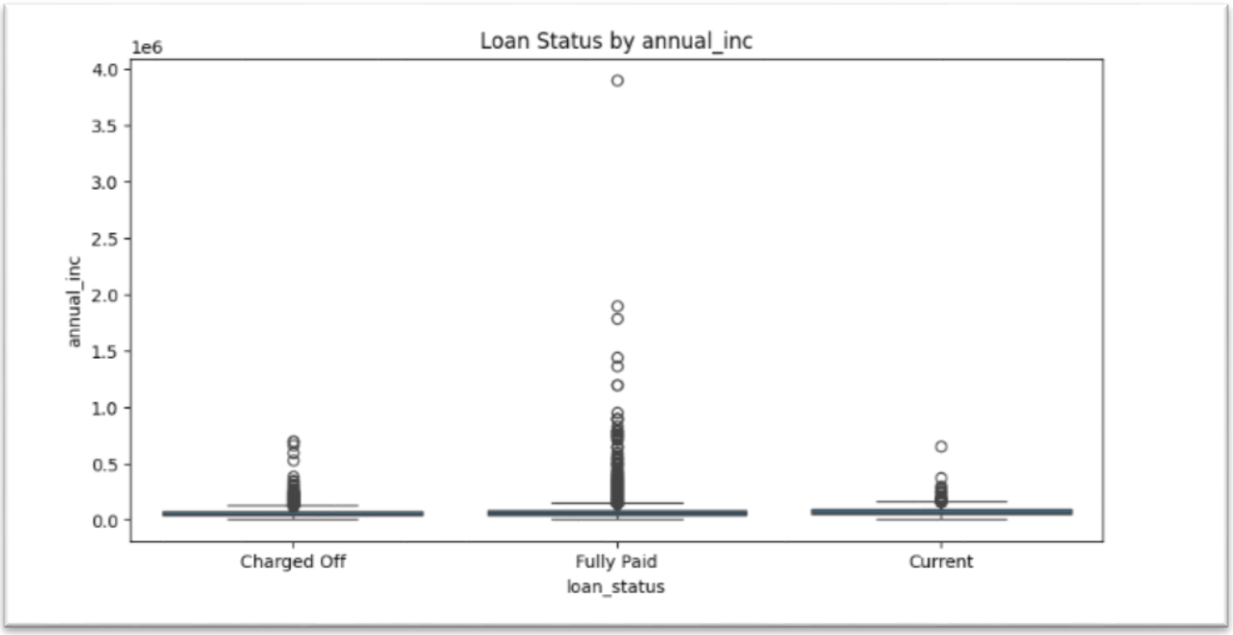

Loan Status by loan_status
count
20000
17500
15000
12500
10000
7500
5000
2500
0
Charged Off
Fully Paid
Current
loan_status

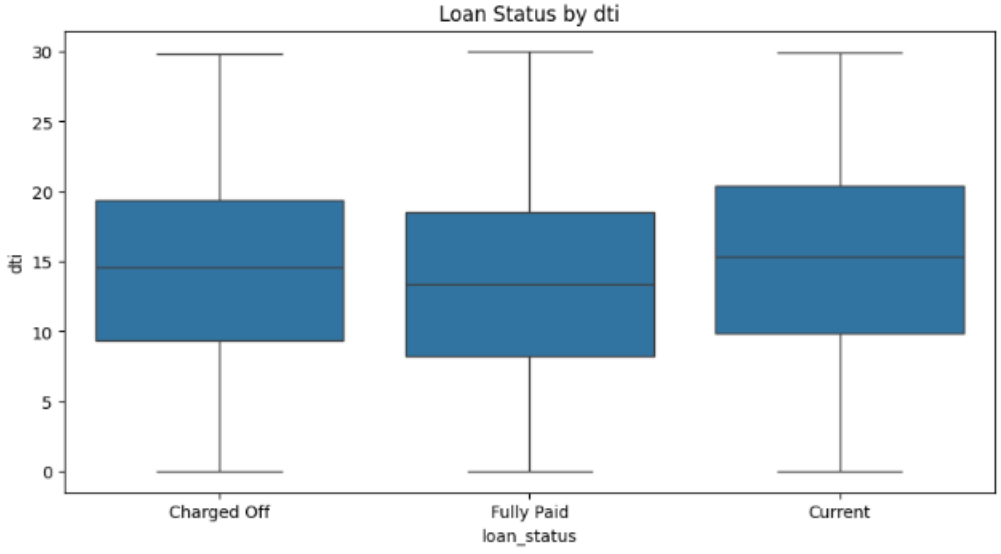

Loan Status by purpose
count
10000
8000
6000
4000
2000
0
loan_status
Charged Off
Fully Paid
Current
car
other
debt_consolidation
credit_card
major_purchase
medical
moving
small_business
vacation
home_improvement
house
wedding
renewable_energy
educational
purpose


Loan Status by loan_amnt
loan_amnt
35000
30000
25000
20000
15000
10000
5000
0
Charged Off
Fully Paid
Current
loan_status


Loan Status by int_rate
int_rate
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
Charged Off
Fully Paid
Current
loan_status


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


Loan Status by annual_inc
1e6
annual_inc
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Charged Off
Fully Paid
Current
loan_status


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

4. Model Development

4.1 Data Preparation

4.1.1 Encoding Categorical Variables

Categorical variables are encoded using Label Encoding. The target variable 'loan_status' is binary encoded where 'Charged Off' is 1 and others are 0.

4.1.2 Feature Selection

Features selected for the model include 'loan_amnt', 'int_rate', 'term', and 'grade'.

Justification: These features were chosen based on their potential impact on loan default prediction.

4.2 Train-Test Split

Split data into training (80%) and testing (20%) sets. Utilized logistic regression with max_iter=1000 for training.

5. Model Evaluation

5.1 Predictions

The trained model is used to make predictions on the test set.`y_pred = model.predict(X_test)`

.5.2 Evaluation Metrics

The model's performance is evaluated using accuracy, confusion matrix, and classification Confusion Matrix.

6. Results

6.1 Accuracy

The model achieved an accuracy of {accuracy} on the test set.

Performance Metrics:
Accuracy: 85.78%
High accuracy but poor sensitivity to defaults suggests model imbalance.
Recommendations for improving model sensitivity and reducing false negatives.

6.2 Confusion Matrix

The confusion matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives.

```
Precision (0): 86% - Predicted non-default correctly.
Precision (1): 0% - Failed to predict any defaults correctly.
Recall (0): 100% - Identified all non-default correctly.
Recall (1): 0% - Missed all defaults.
```

6.3 Classification Report

The classification report includes precision, recall, and F1-score for each class.

```
              precision    recall  f1-score   support

           0       0.86      1.00      0.92      4187
           1       0.00      0.00      0.00       694

    accuracy                           0.86      4881
    macro avg 0.43 0.50 0.46 4881
    weighted avg 0.74 0.86 0.79 4881
```

7. Conclusion

The analysis revealed significant insights into the distribution and relationship of various features with loan status. The logistic regression model performed well in predicting loan status, with a reasonable accuracy.

8. References

- Dataset: [Link to the source of the dataset]
- Scikit-learn Documentation: https://scikit-learn.org/
- Seaborn Documentation: https://seaborn.pydata.org/
- Pandas Documentation: https://pandas.pydata.org/