Insights into Lending Risks

Optimizing Loan Approvals Through Identification Risky Applicant

Amaan Shaikh
Senior Analyst at Tiger Analytics

Table of content

Background

Company:

Largest online loan marketplace facilitating personal and business loans.

Challenge:

High financial losses due to loan defaults.

Focus:

Analyze historical loan data (2007–2011) to identify default risky applicants.

Business Problem

Problem:

Financial losses caused by loan defaults.

Solution:

Need for identifying risky loan applicants during the approval process.

Dataset Overview

loan.csv: Historical loan data (2007–2011).

Data_Directory.xlsx: Variable descriptions.

Key Features: Loan amount, interest rate, annual income, credit history, etc.

loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership	annual_inc	v
5000	5000	4975.00	36 months	10.65%	162.87	B	B2	NaN	10+ years	RENT	24000.00	
2500	2500	2500.00	60 months	15.27%	59.83	C	C4	Ryder	< 1 year	RENT	30000.00	
2400	2400	2400.00	36 months	15.96%	84.33	C	C5	NaN	10+ years	RENT	12252.00	
10000	10000	10000.00	36 months	13.49%	339.31	C	C1	AIR RESOURCES BOARD	10+ years	RENT	49200.00	
3000	3000	3000.00	60 months	12.69%	67.79	B	B5	University Medical Group	1 year	RENT	80000.00	

Objectives

Conduct exploratory data analysis (EDA) to:

- Identify strong indicators of loan default.
- Highlight potential risky applicants.
- Provide actionable recommendations.

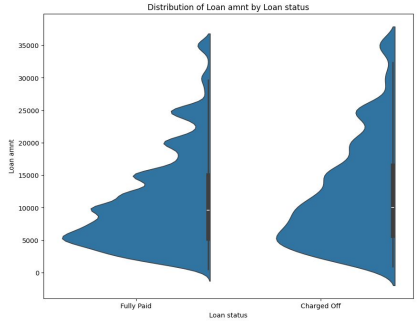
Data Cleaning & Preprocessing

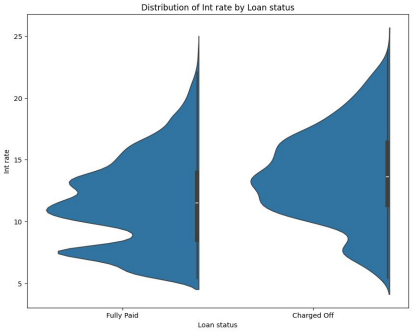
- Handled missing values and outliers.
- Applied log transformations to reduce skewness.
- Engineered features: `credit_history_length`, `time_since_last_payment`, etc.

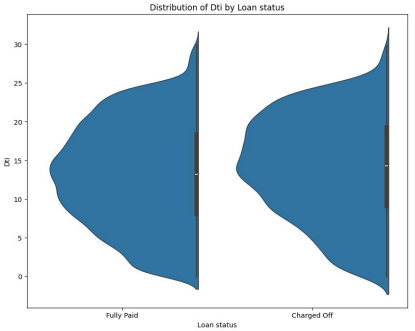
Key Insights – Numerical Variables (Part-1)

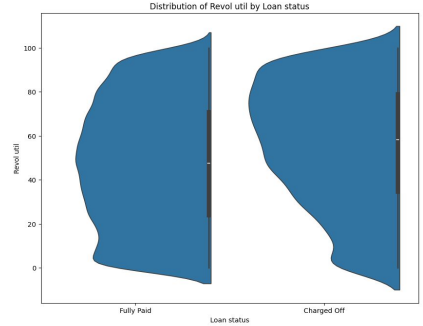
Risky Applicant Indicators:

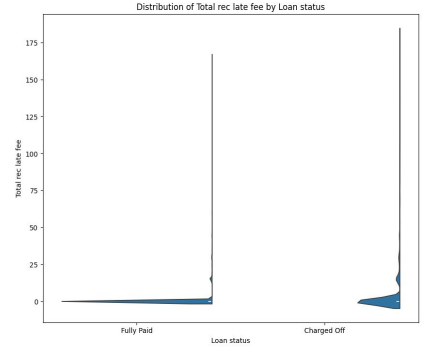
- 🔺 High loan amounts[Loan amnt]
- 🔺 High interest rates [int rate].
- 🔺 High debt-to-income ratio[dti]
- 🔺 High revolving utilization rate[Revol util].
- 🔺 High count of late fee received to date[Total rec late fee]
- 🔺 High day count since last payment[Time since last payment]
- 🔻 Low annual income [Annual inc]
- 🔻 Low short credit history [Credit history length]
- 🔻 Less day count since last credit pull [Time Since last credit pull]

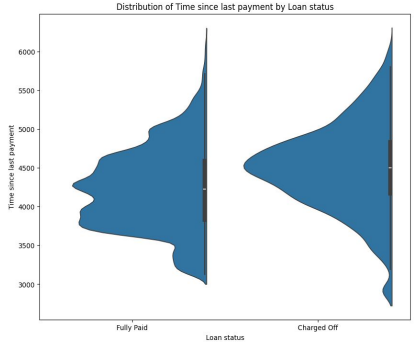


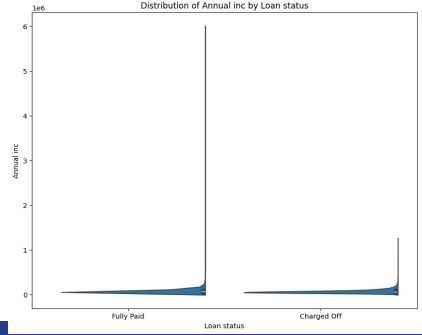


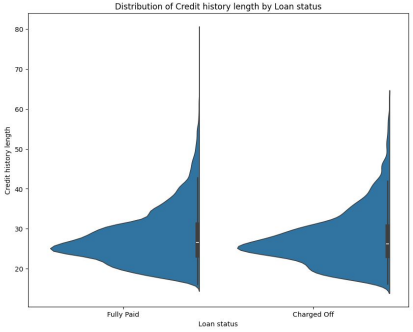


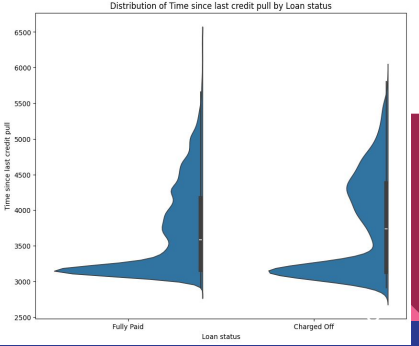








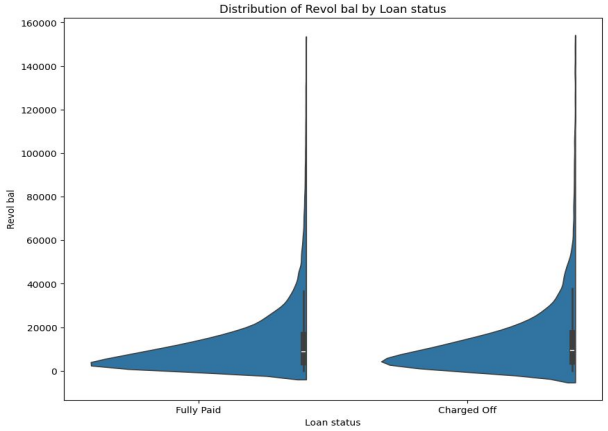


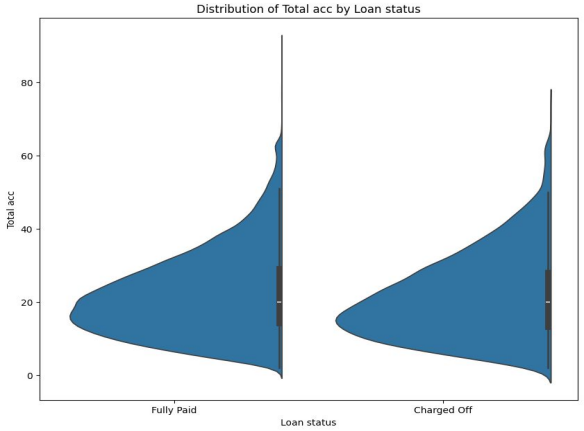


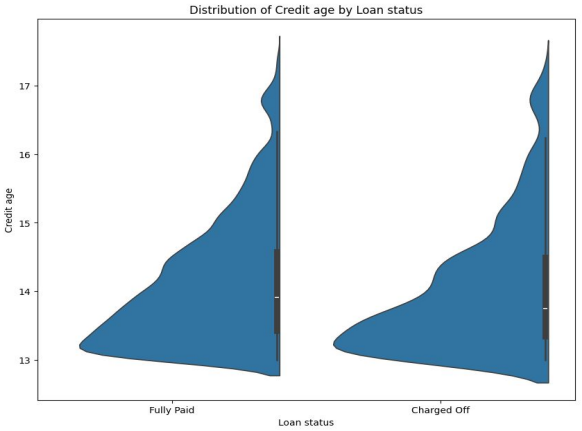
Key Insights – Numerical Variables (Part-2)

Factors with Minimal or No Impact visible on charts:

- ★ Total credit revolving balance [Revol bal]
- ★ Total number of credit lines currently in borrower's credit file[Total acc]
- ★ Total number of years borrower's credit history present [Credit age]



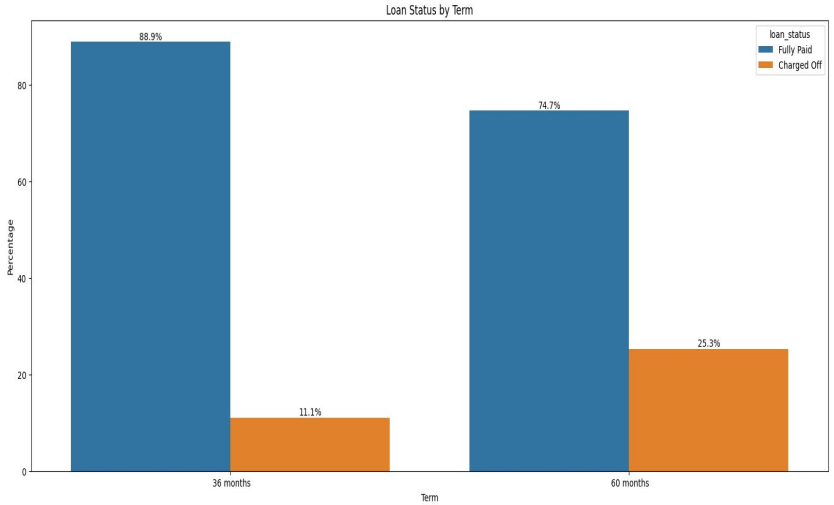


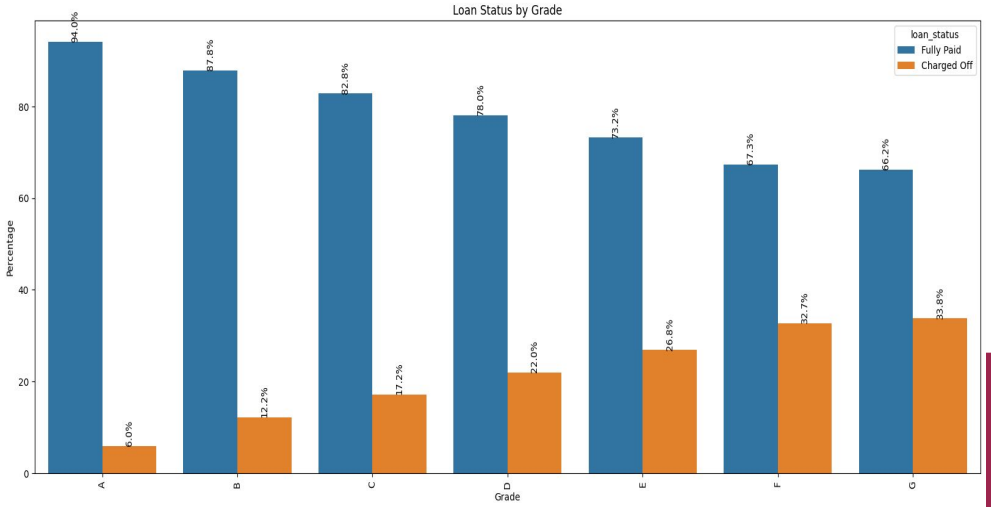


Key Insights – Categorical Variables (Part-1)

Risky Indicators:

- Loans with a 60-month term have a significantly higher default percentage. [Term]
- As grade value increasing, default percentages increase, indicating a strong correlation with risk. [Grade]

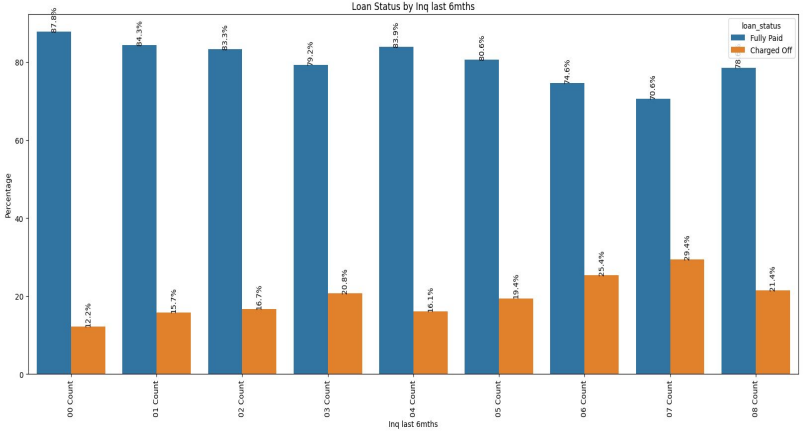


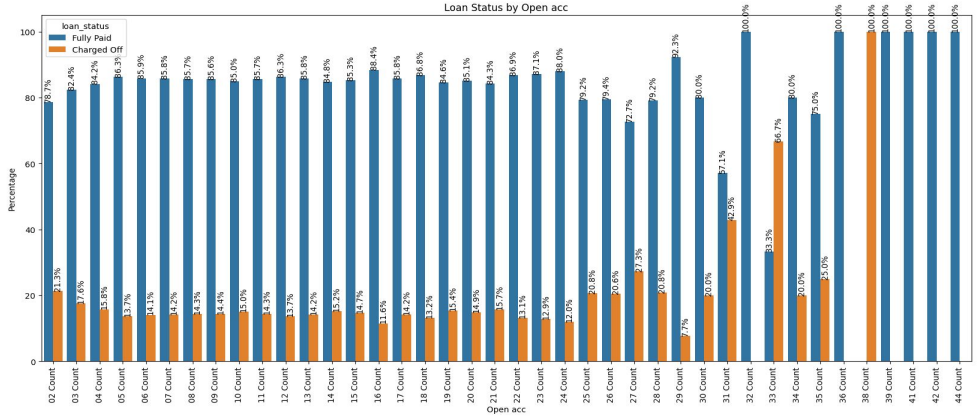


Key Insights – Categorical Variables (Part-2)

Risky Indicators:

- Higher inquiries in the last 6 months correlate with increased default percentages. [Inq last 6mths]
- Higher number of open accounts is associated with increased default risk. [Open acc]

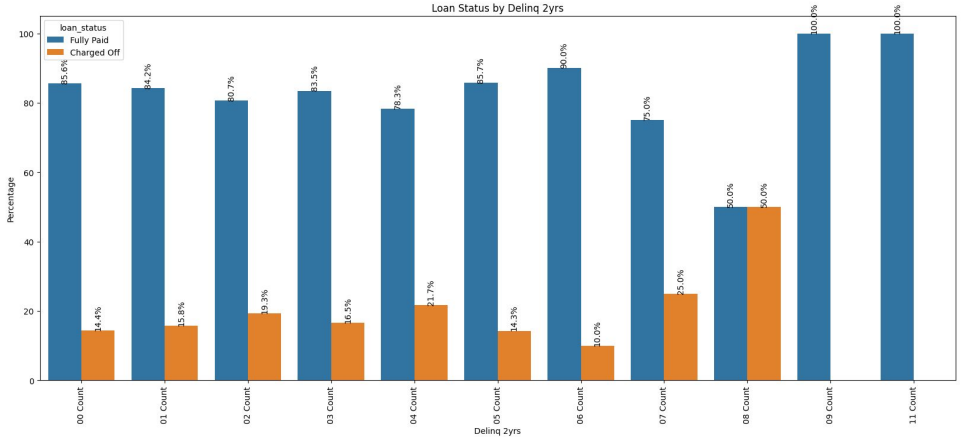


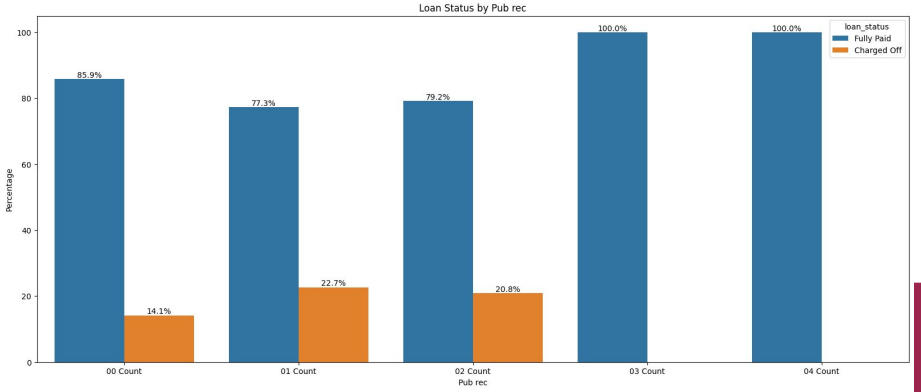


Key Insights – Categorical Variables (Part-3)

Factors with Minimal Impact visible on charts:

- Generally insignificant, but higher 30+days past due incidences of delinquency counts may indicate potential risk. [Delinq 2yrs]
- Public records have minimal impact on loan status. [Pub rec]

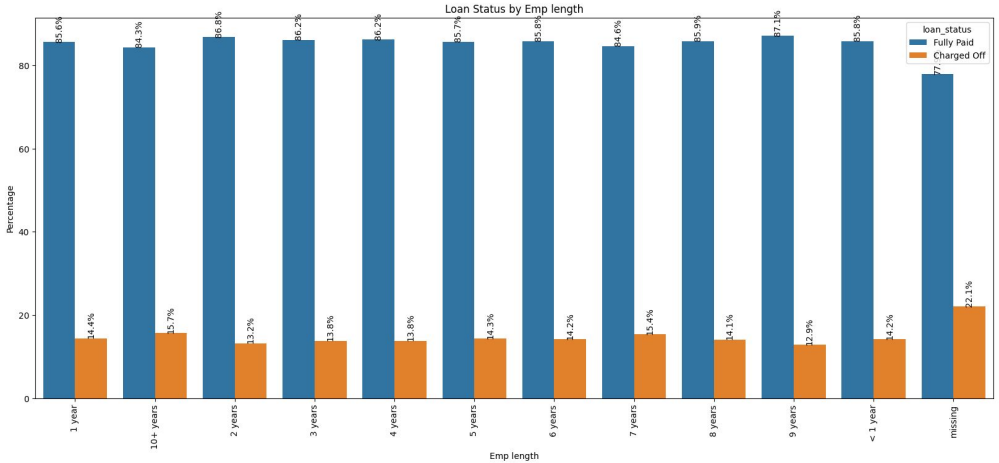


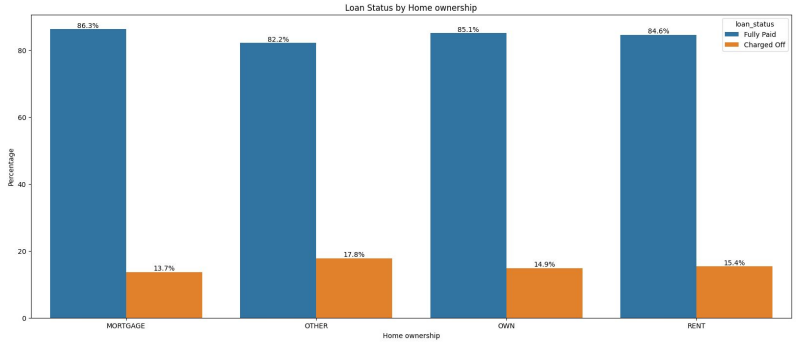


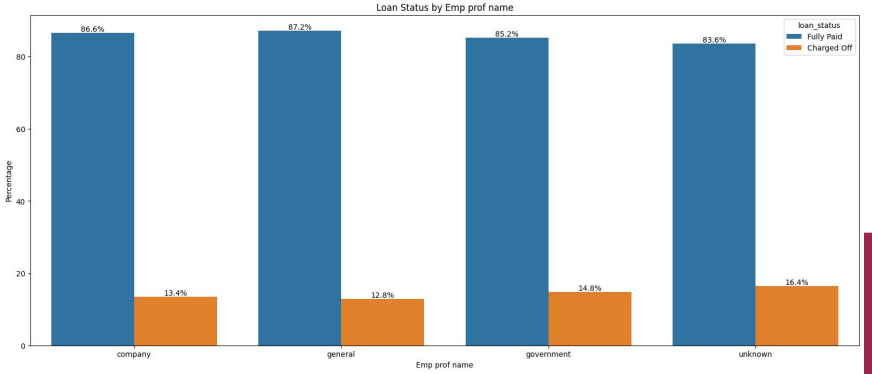
Key Insights – Categorical Variables (Part-4)

Factors with No Impact visible on charts:

- No significant impact of Employment length, though missing data shows a slight increase in default percentage. Also, missing data has highest default percentage. [Emp length]
- No notable differences in default percentages based on home ownership.
- Profession names show no strong relationship with default risk. [Emp prof name]



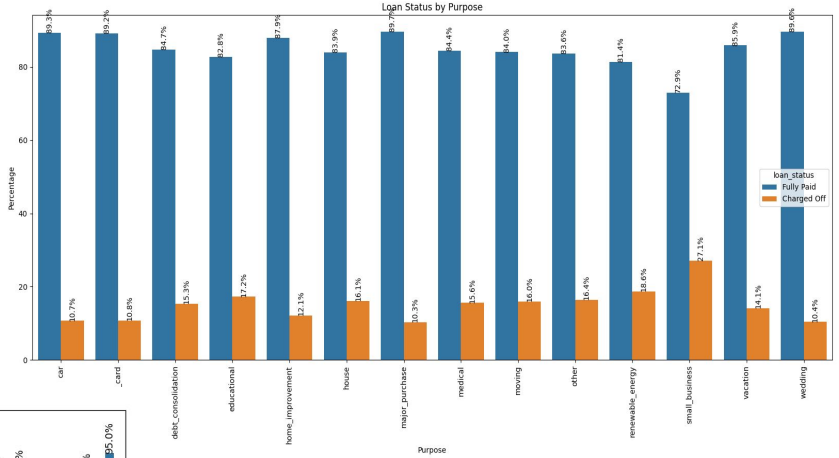


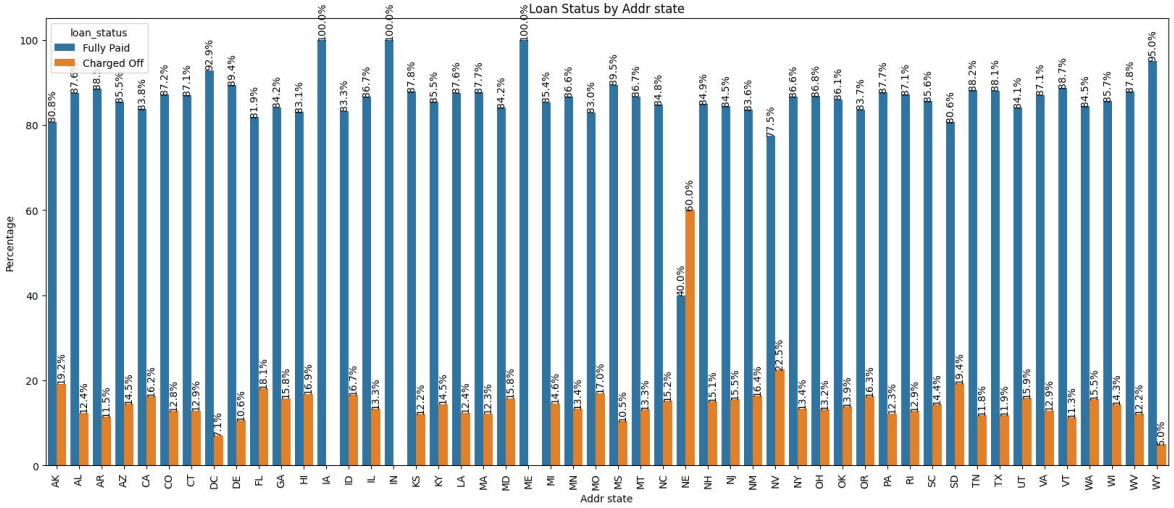


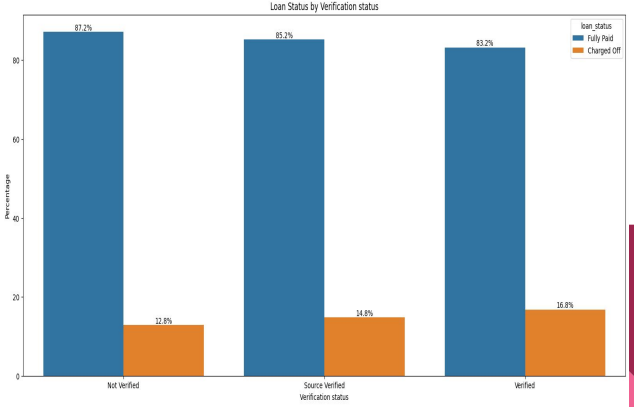
Indirect Impact Factors

Factors with No Direct Impact but Possible Indications:

- Loans with a 60-month term have significantly higher default percentages, indicating a potential need to assess risk more rigorously for longer-term loans.
- Verified applicants show a higher default percentage, suggesting potential flaws in the verification process.
- Loans for small businesses have the highest default rates, indicating a need for stricter background checks for certain purposes.
- Nebraska (NE) shows a 60% default rate, suggesting validation processes vary by state and may require improvement.







Recommendations

- Revise credit approval policies for:
  - **High** loan amounts or interest rates or DTI ratio.
  - **Higher number** of open accounts or inquiries
  - **Short** credit history
  - Longer-term loans
- Strengthen verification processes.
- Implement stricter checks for:
  - Small business loans.
  - High-risk states like Nebraska.
  - Precise data gathering such as Employee length (Employment experience year)

Challenges & Assumptions

🍕Challenges:

- Skewed numerical variables persist post-transformation.
- High granularity in categorical variables (e.g., addr_state, purpose).
- Balancing outliers vs noise.
- Interpreting categorical variables effectively.

💡Assumptions:

- loan_status reliably represents borrower risk.
- Economic conditions during the dataset period are stable.
- Data accurately represents past trends.
- Missing data is random and unbiased.
- Retained outliers to capture risk-related anomalies.

Conclusion

- ➔ Identified key drivers of loan default.
- ➔ Provided actionable insights to minimize credit risk.
- ➔ Enhanced decision-making for loan approvals.

GitHub Repository

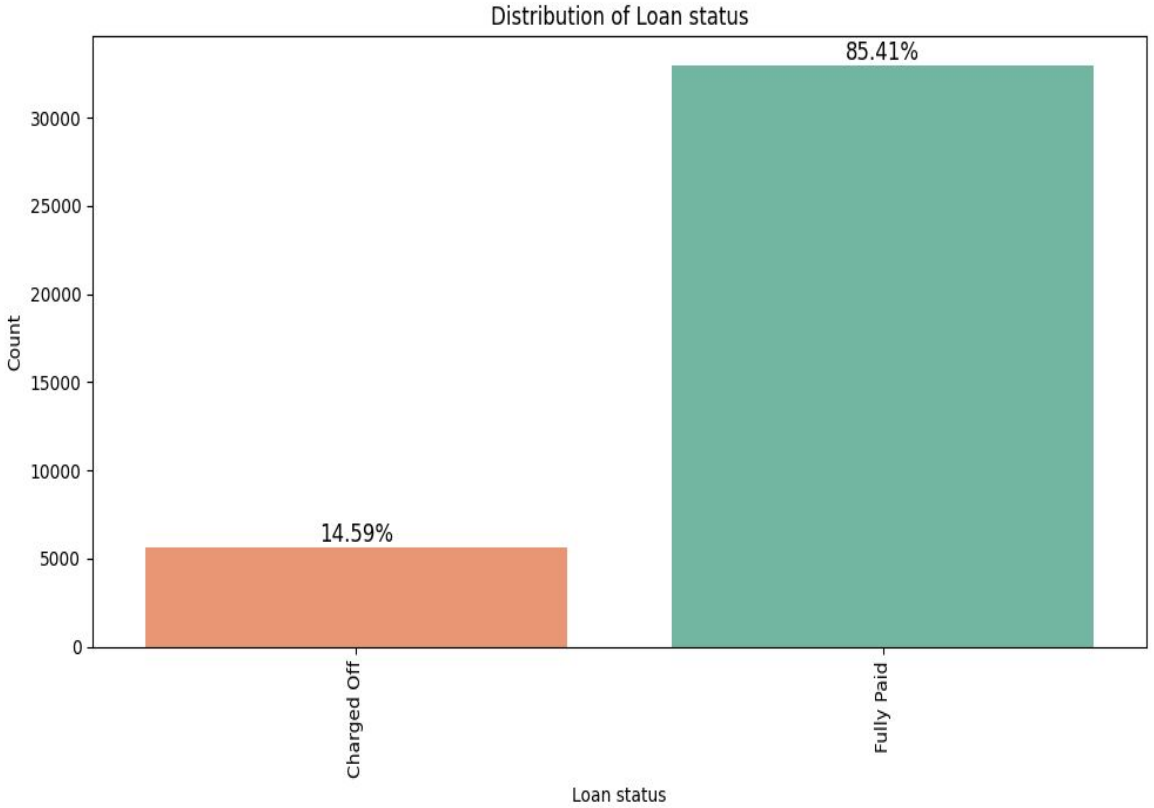
Link: insights-into-lending-risks

Reference: https://github.com/amaan2398/insights-into-lending-risks

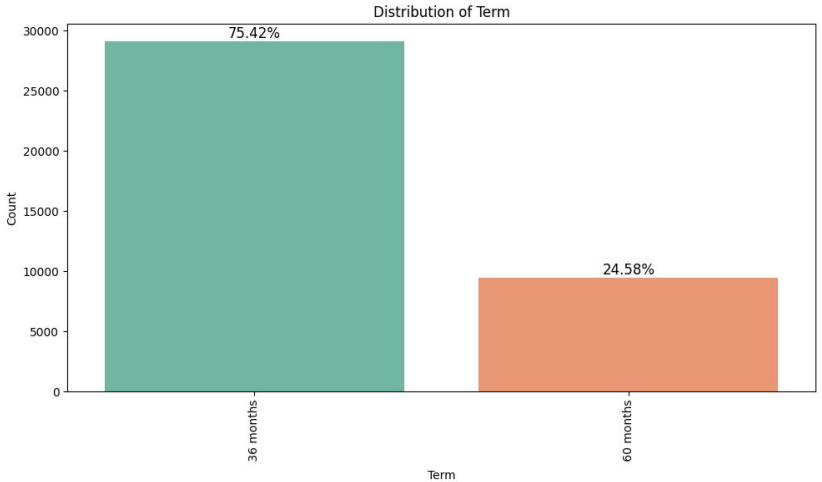
Thank You

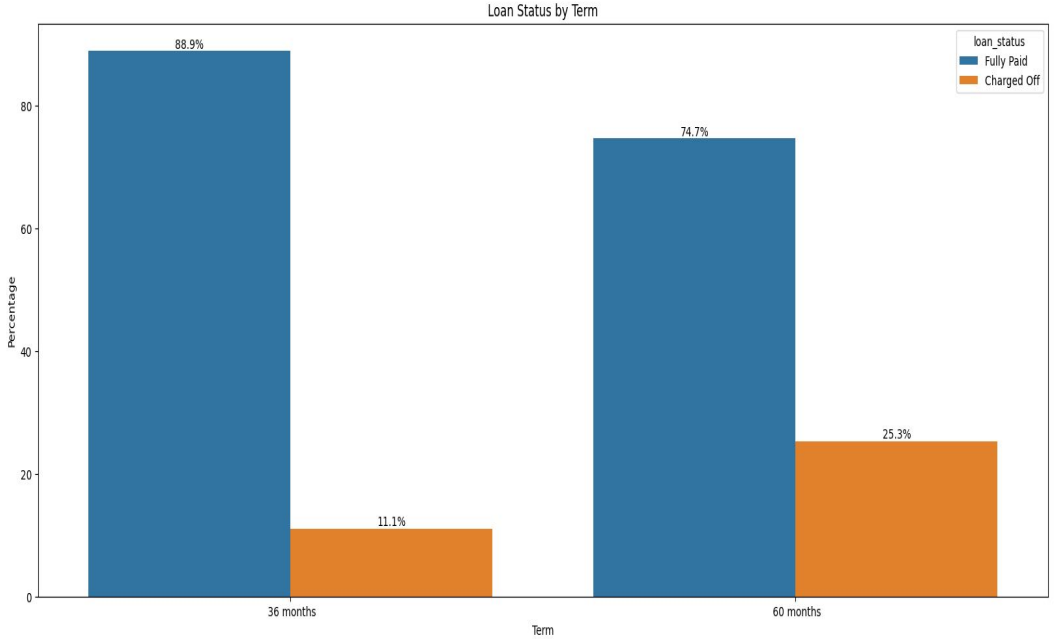
Visualizations Overview

Loan Status

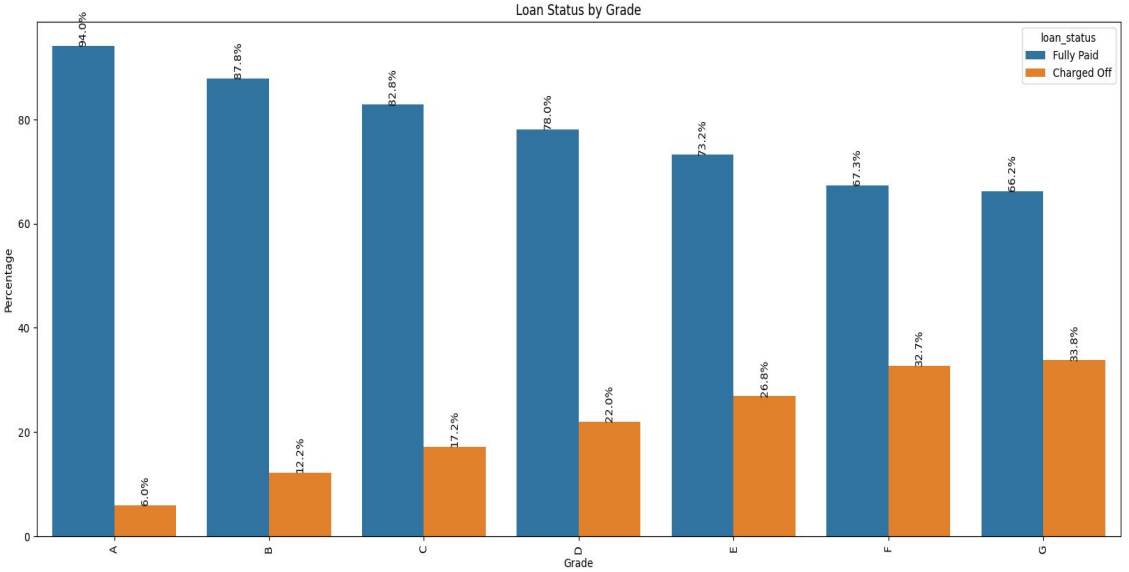

Distribution of Loan status
85.41%
14.59%
30000
25000
20000
15000
10000
5000
0
Count
Charged Off
Fully Paid
Loan status

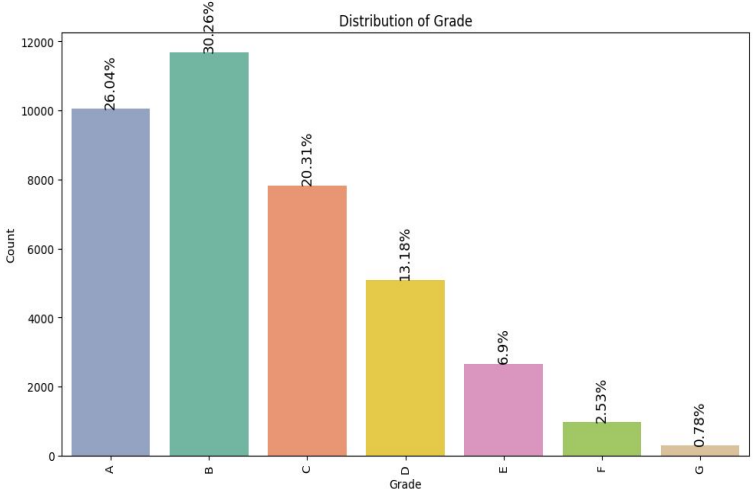
Term


Distribution of Term
75.42%
24.58%
Count
Term
36 months
60 months

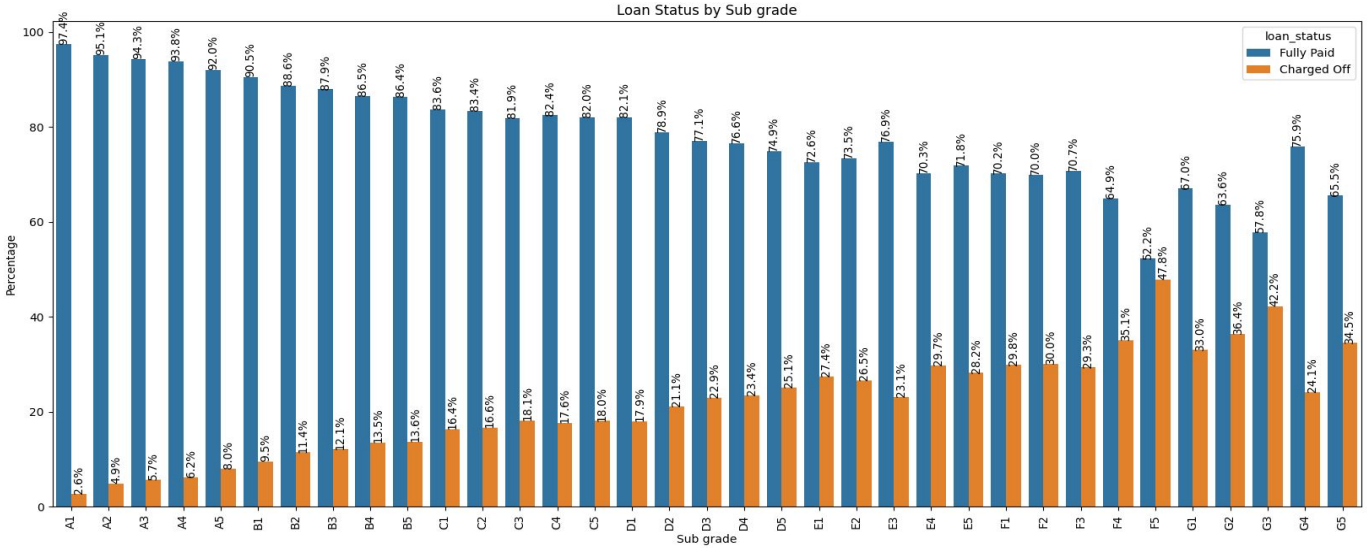

Loan Status by Term
loan_status
Fully Paid
Charged Off
88.9%
11.1%
74.7%
25.3%
Percentage
Term
36 months
60 months

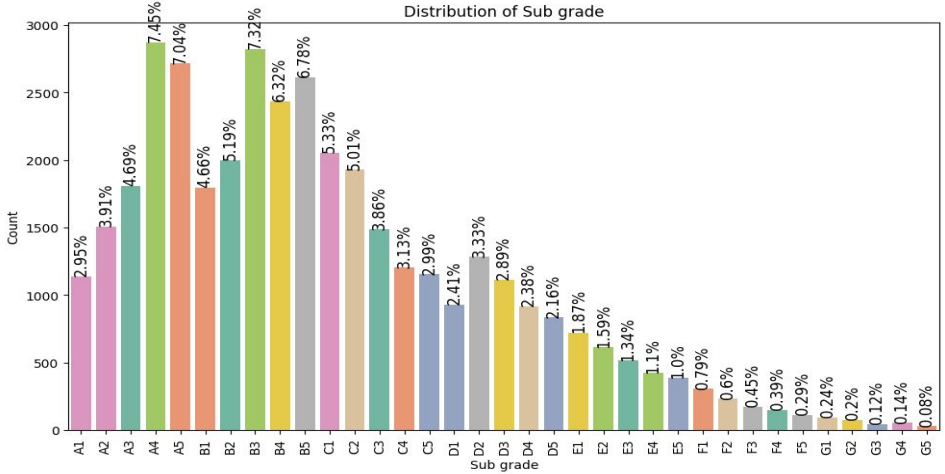
Grade


Loan Status by Grade
loan_status
Fully Paid
Charged Off
Percentage
Grade
A
B
C
D
E
F
G
94.0%
6.0%
87.8%
12.2%
82.8%
17.2%
78.0%
22.0%
73.2%
26.8%
67.3%
32.7%
66.2%
33.8%

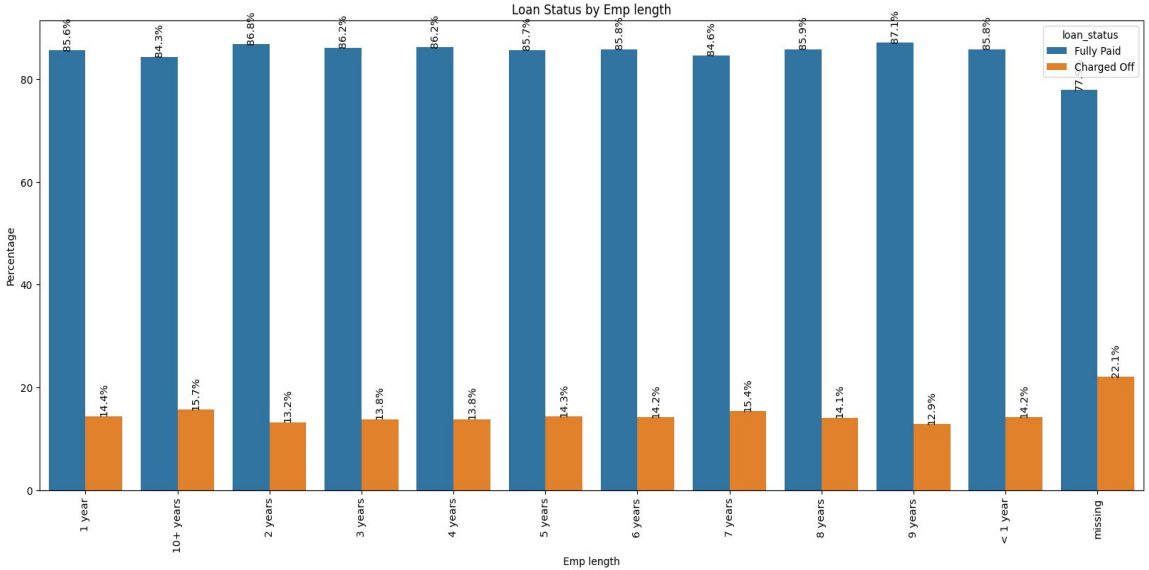

Distribution of Grade
Count
Grade
A
B
C
D
E
F
G
26.04%
30.26%
20.31%
13.18%
6.9%
2.53%
0.78%

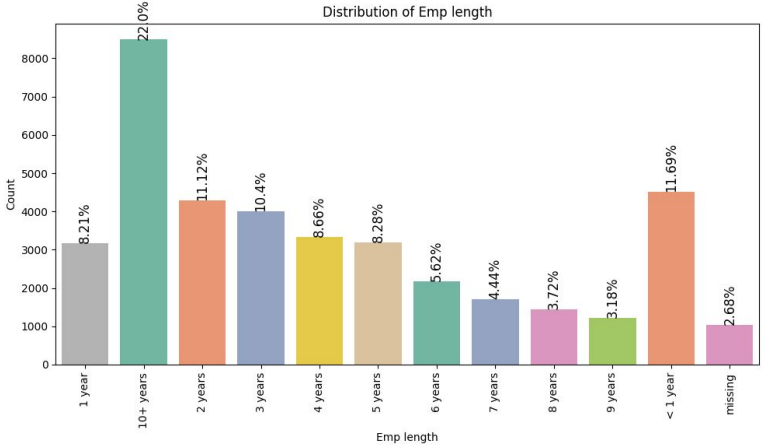
Sub Grade


Loan Status by Sub grade
loan_status
Fully Paid
Charged Off
Percentage
Sub grade
A1 97.4% 2.6%
A2 95.1% 4.9%
A3 94.3% 5.7%
A4 93.8% 6.2%
A5 92.0% 8.0%
B1 90.5% 9.5%
B2 88.6% 11.4%
B3 87.9% 12.1%
B4 86.5% 13.5%
B5 86.4% 13.6%
C1 83.6% 16.4%
C2 83.4% 16.6%
C3 81.9% 18.1%
C4 82.4% 17.6%
C5 82.0% 18.0%
D1 82.1% 17.9%
D2 78.9% 21.1%
D3 77.1% 22.9%
D4 76.6% 23.4%
D5 74.9% 25.1%
E1 72.6% 27.4%
E2 73.5% 26.5%
E3 76.9% 23.1%
E4 70.3% 29.7%
E5 71.8% 28.2%
F1 70.2% 29.8%
F2 70.0% 30.0%
F3 70.7% 29.3%
F4 64.9% 35.1%
F5 52.2% 47.8%
G1 67.0% 33.0%
G2 63.6% 36.4%
G3 57.8% 42.2%
G4 75.9% 24.1%
G5 65.5% 34.5%

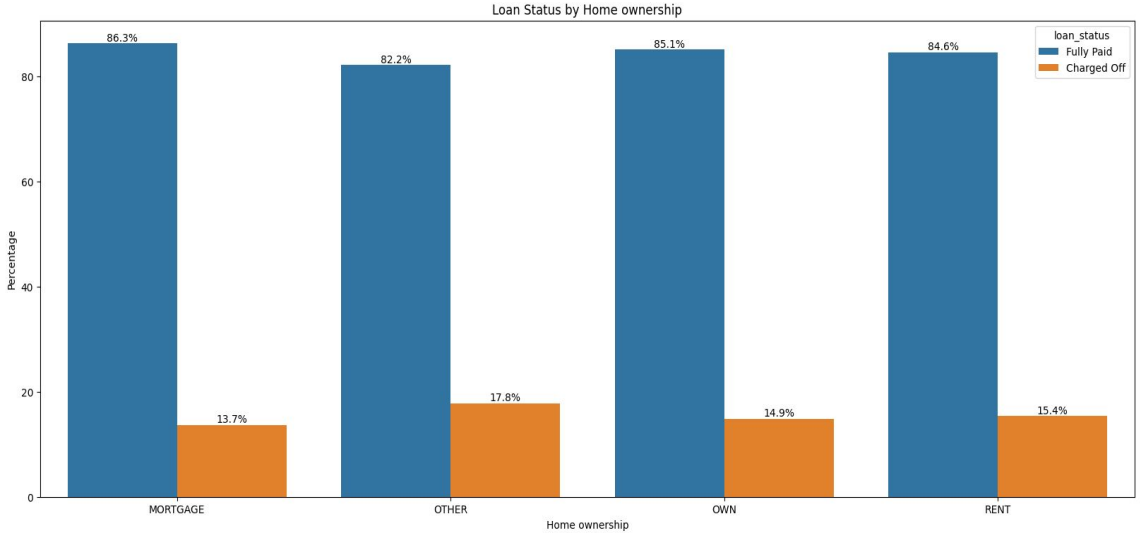

Distribution of Sub grade
Count
Sub grade
A1 2.95%
A2 3.91%
A3 4.69%
A4 7.45%
A5 7.04%
B1 4.66%
B2 5.19%
B3 7.32%
B4 6.32%
B5 6.78%
C1 5.33%
C2 5.01%
C3 3.86%
C4 3.13%
C5 2.99%
D1 2.41%
D2 3.33%
D3 2.89%
D4 2.38%
D5 2.16%
E1 1.87%
E2 1.59%
E3 1.34%
E4 1.1%
E5 1.0%
F1 0.79%
F2 0.6%
F3 0.45%
F4 0.39%
F5 0.29%
G1 0.24%
G2 0.2%
G3 0.12%
G4 0.14%
G5 0.08%

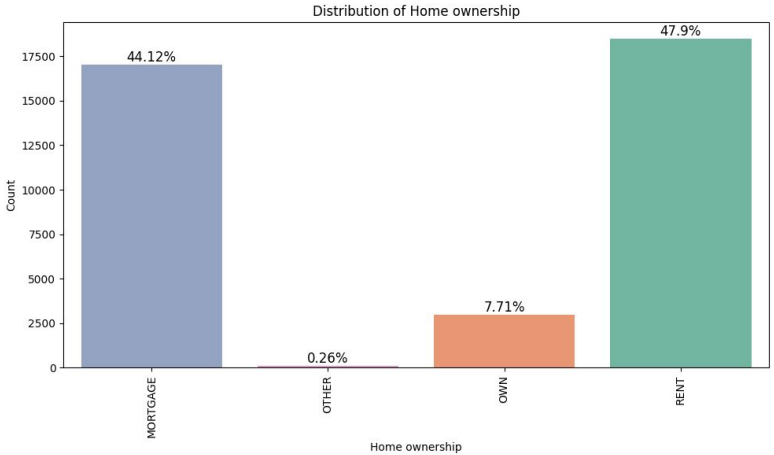
Employee Length



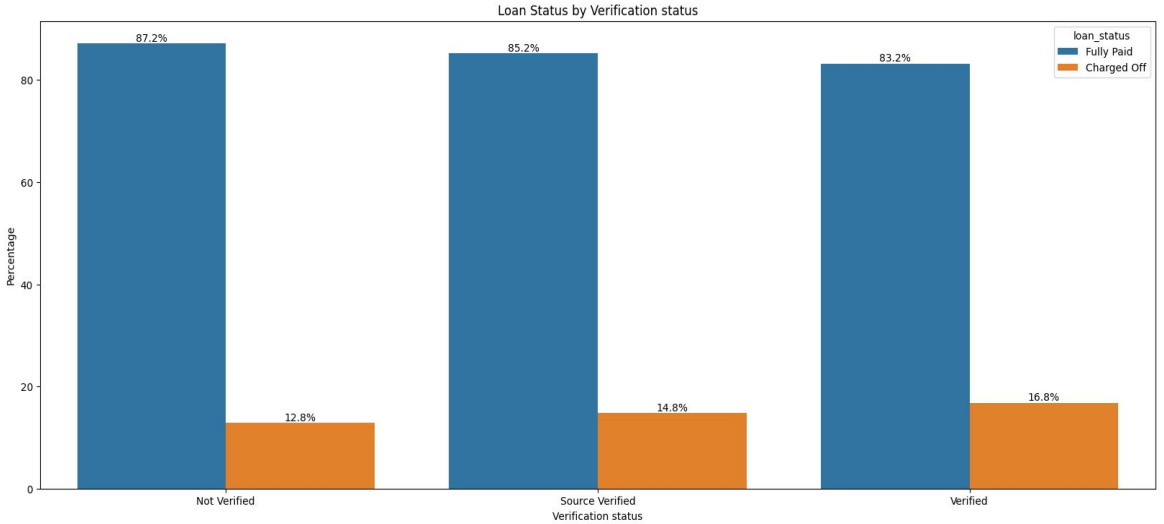


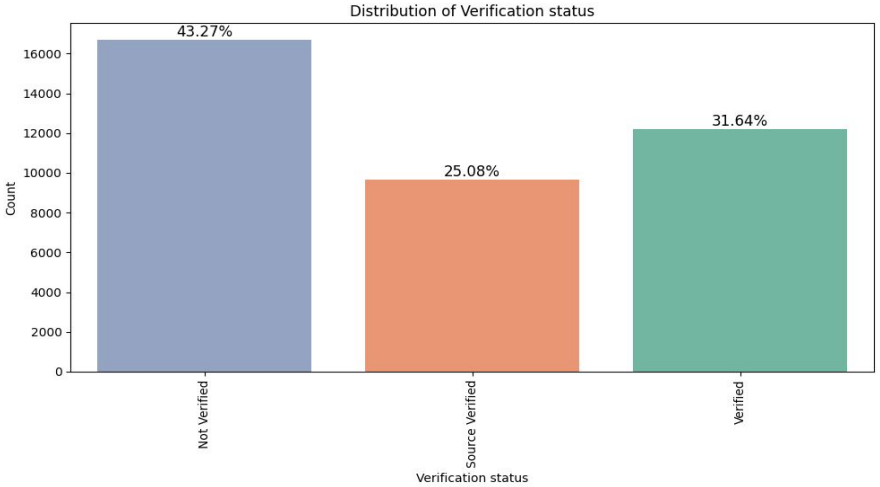
Home Ownership



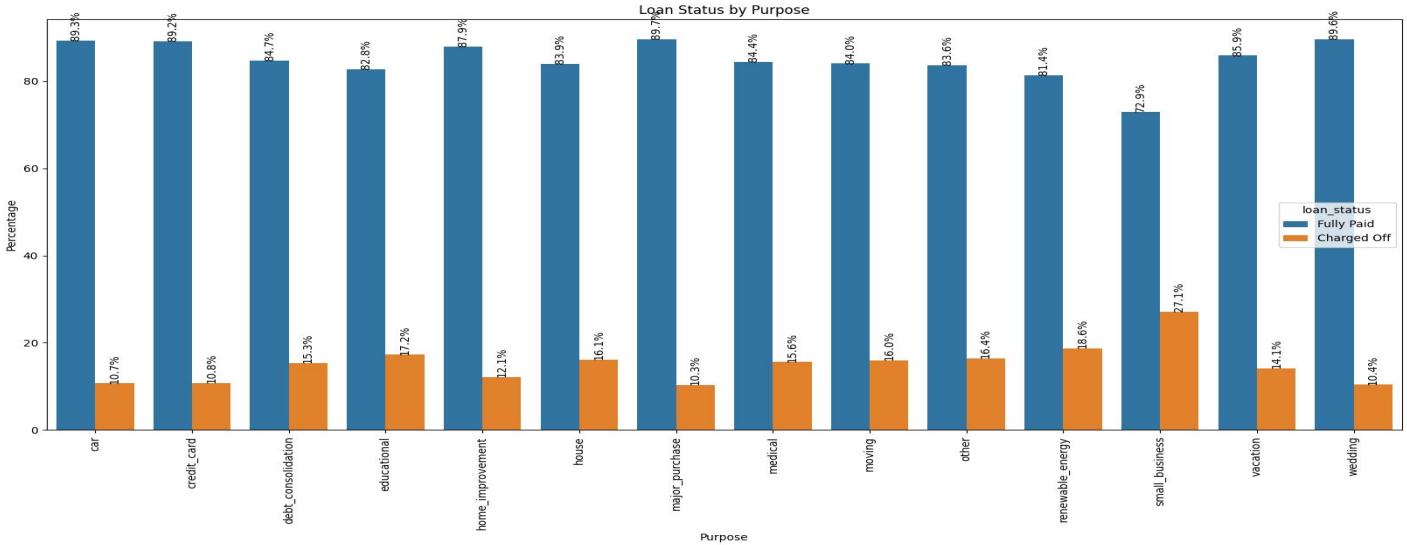


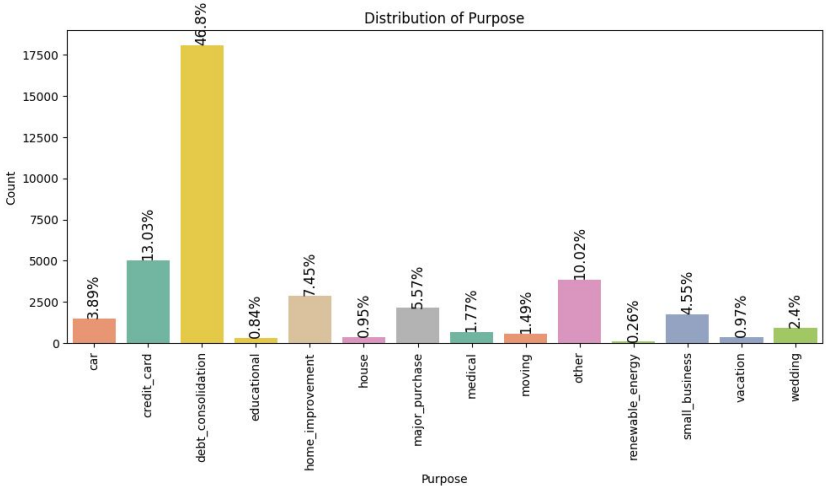
Verification Status



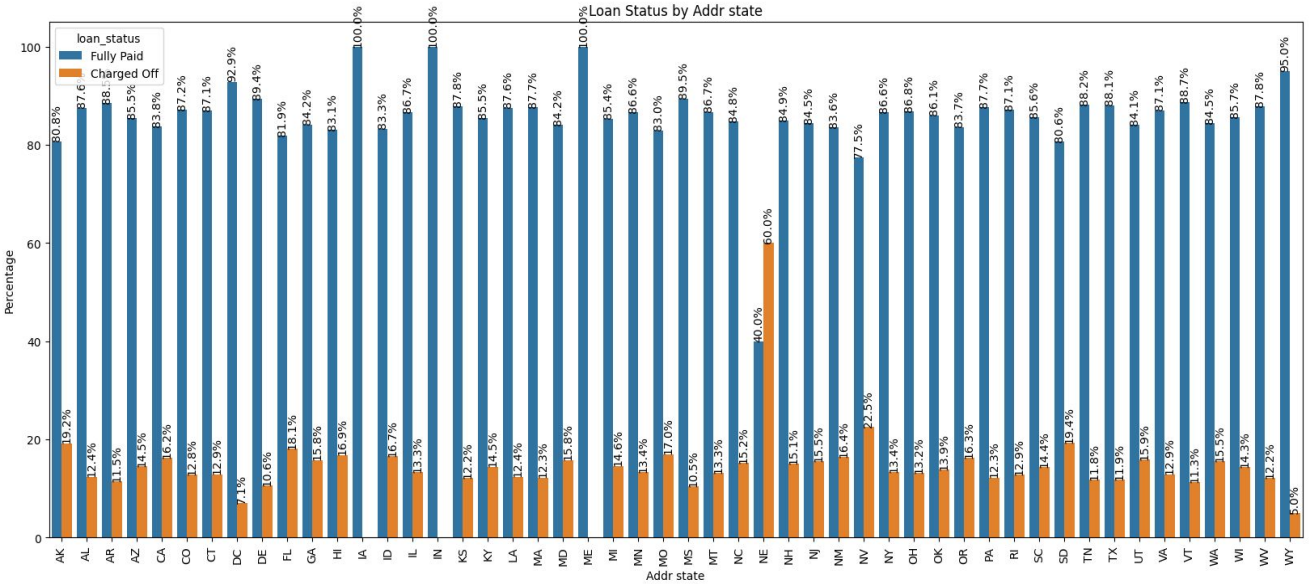


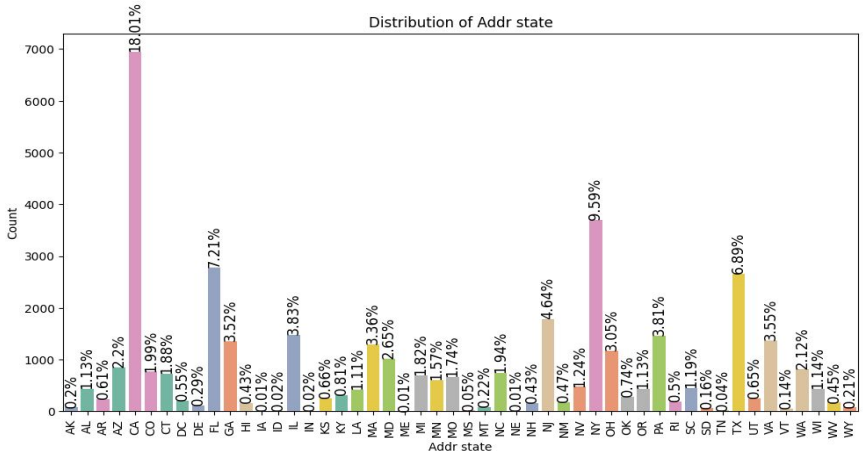
Purpose


Loan Status by Purpose
Percentage
Purpose
loan_status
Fully Paid
Charged Off
car
credit_card
debt_consolidation
educational
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding

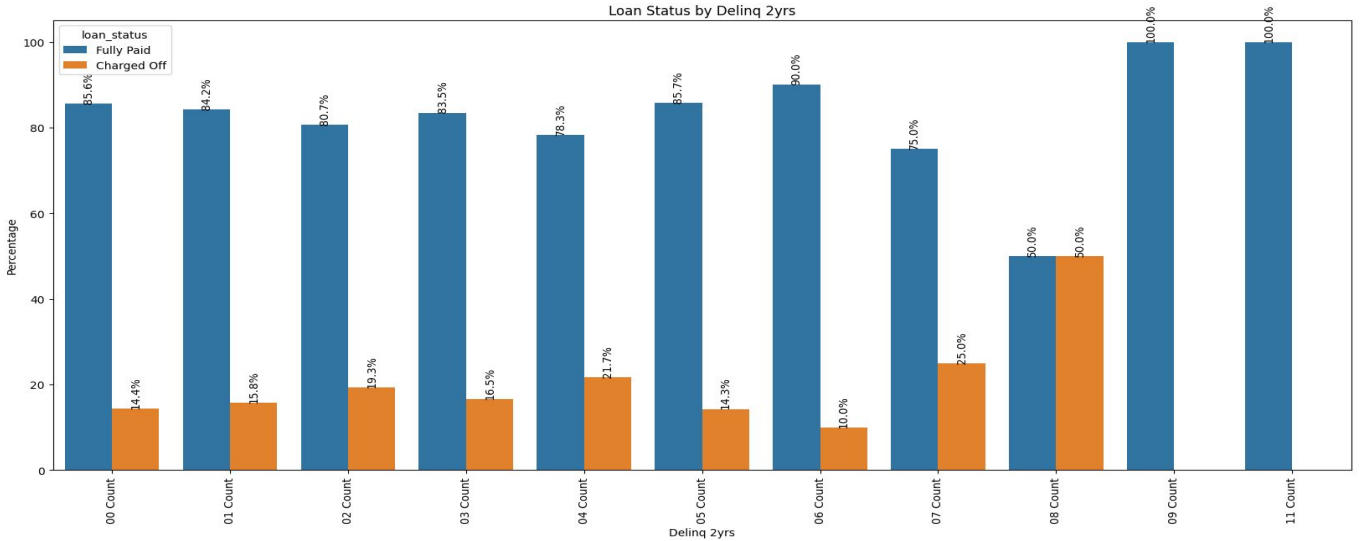

Distribution of Purpose
Count
Purpose
3.89%
13.03%
46.8%
0.84%
7.45%
0.95%
5.57%
1.77%
1.49%
10.02%
0.26%
4.55%
0.97%
2.4%

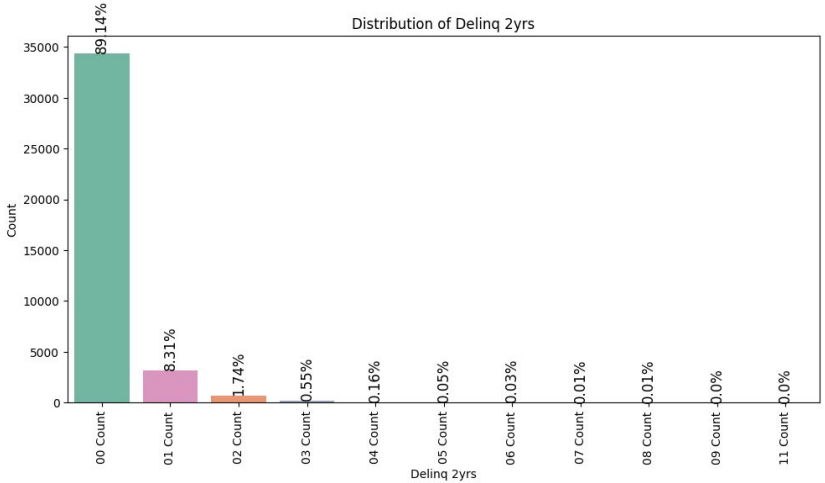
Address State



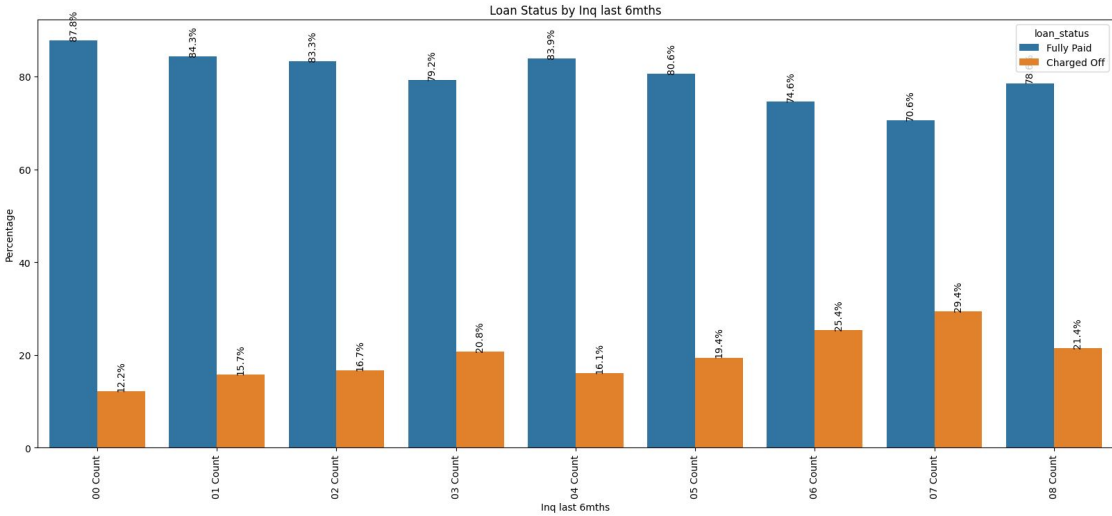


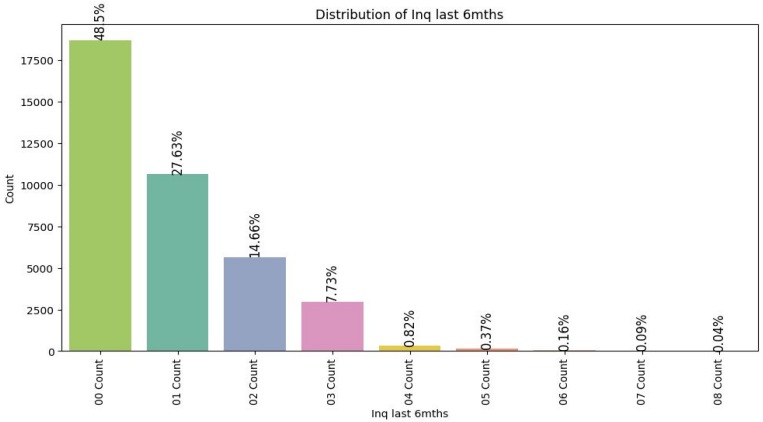
Delinquency



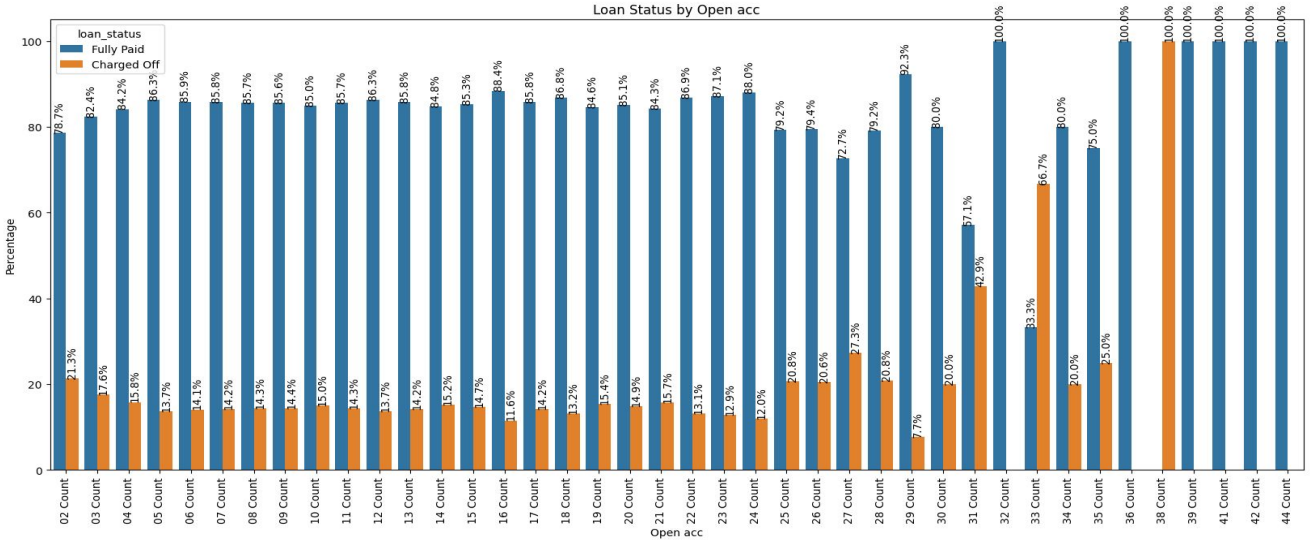


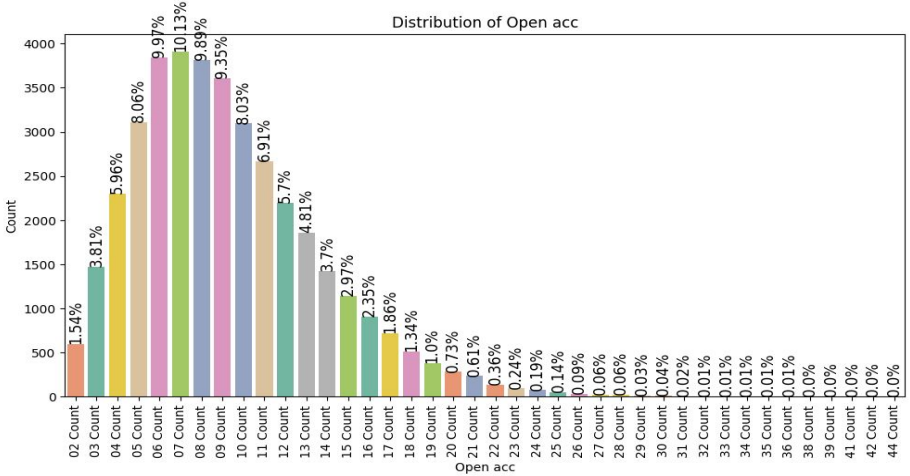
Inquiry Last 6 months



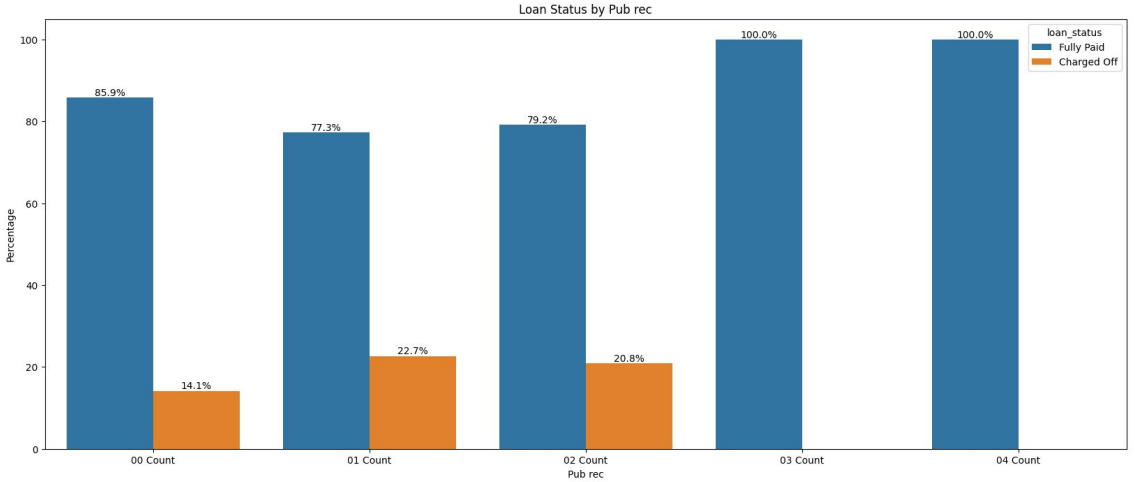


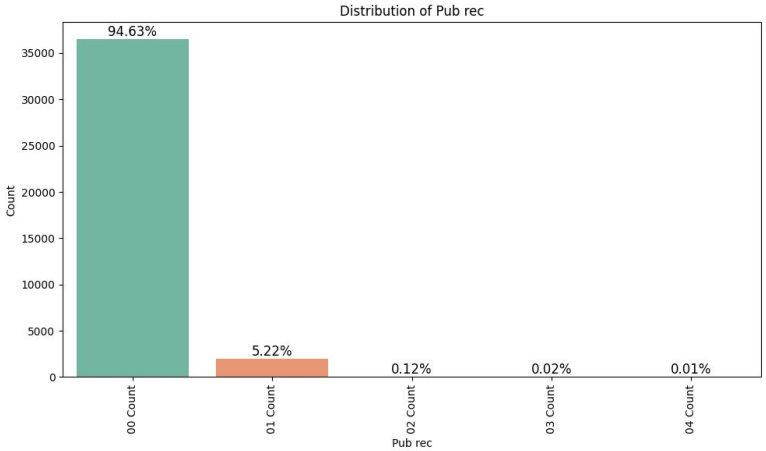
Open Account



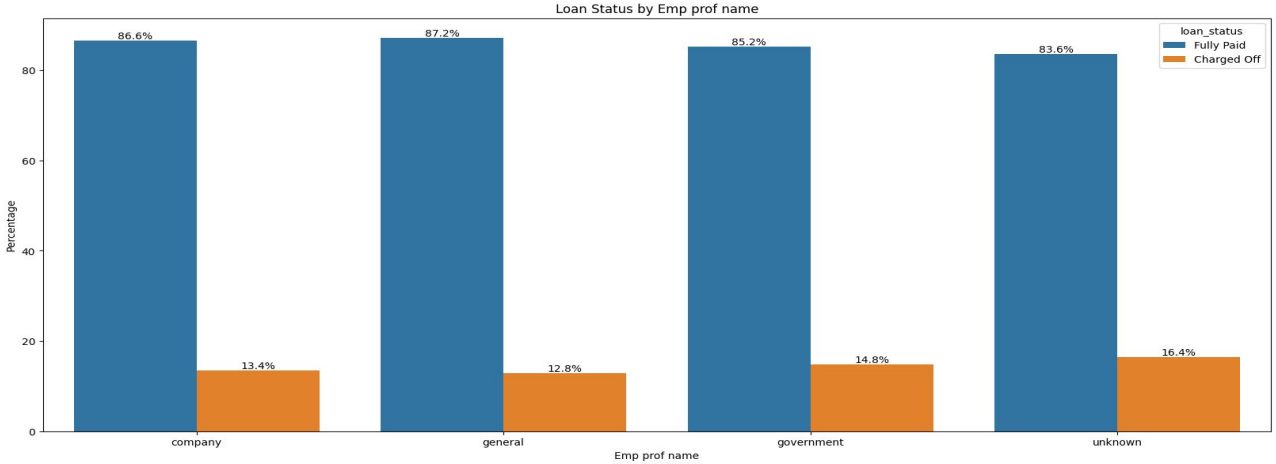


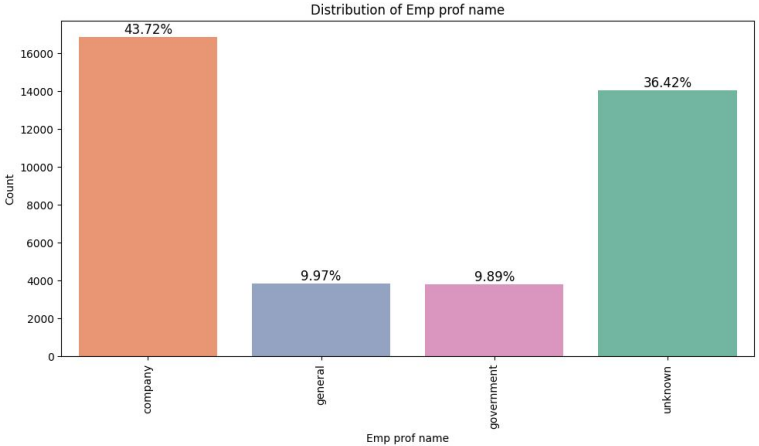
Number of derogatory public records



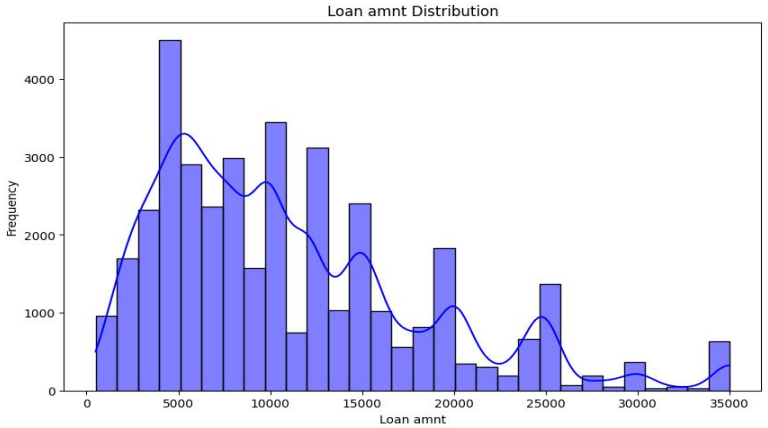


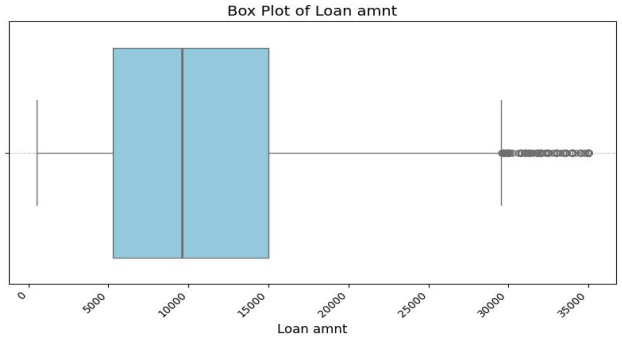
Employee profession

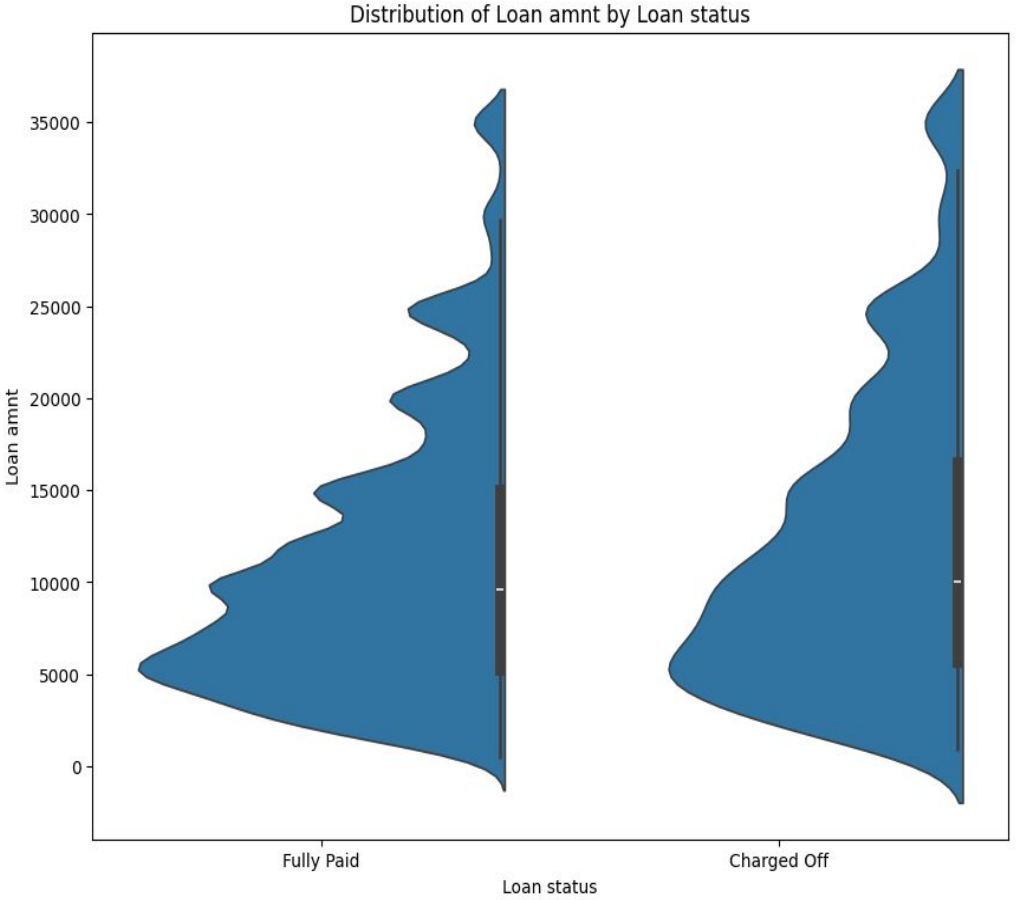




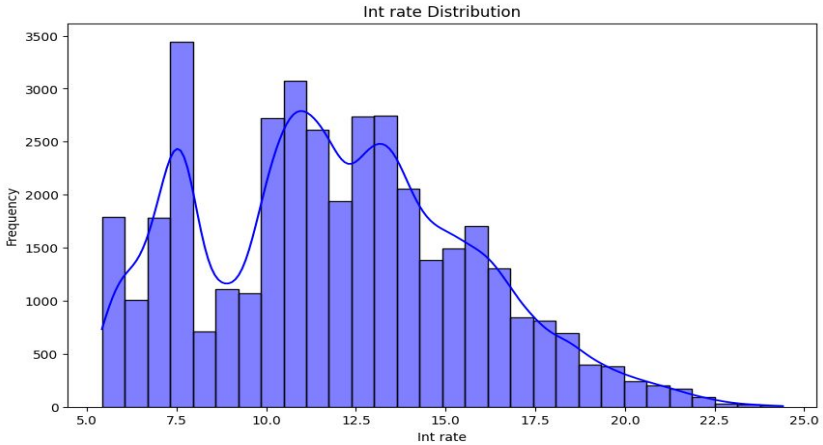
Loan Amount

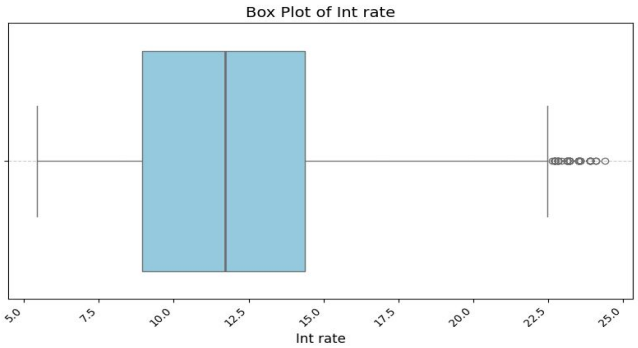

Loan amnt Distribution
Frequency
Loan amnt
0
1000
2000
3000
4000
0
5000
10000
15000
20000
25000
30000
35000


Box Plot of Loan amnt
Loan amnt
0
5000
10000
15000
20000
25000
30000
35000

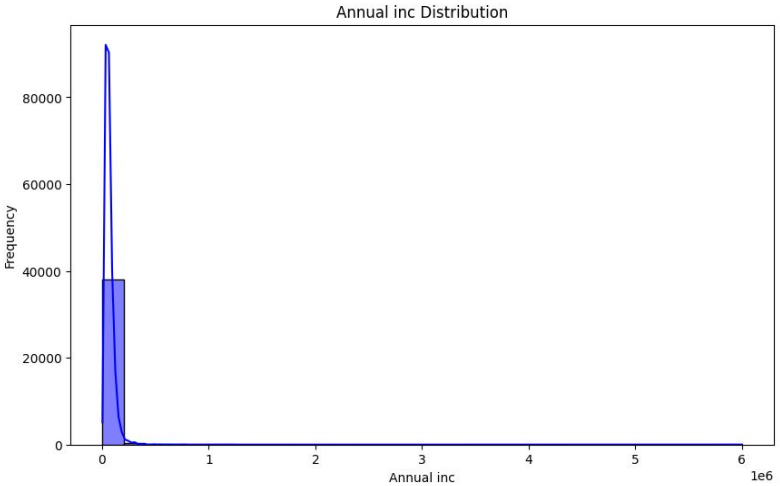

Distribution of Loan amnt by Loan status
Loan amnt
0
5000
10000
15000
20000
25000
30000
35000
Fully Paid
Charged Off
Loan status

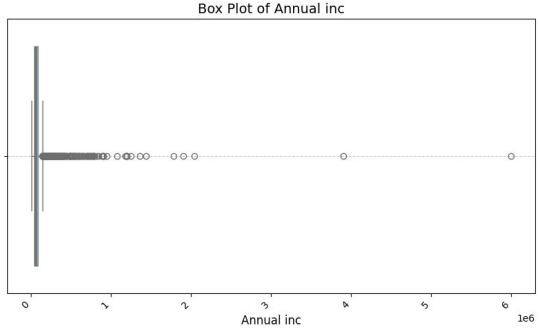
Interest rate

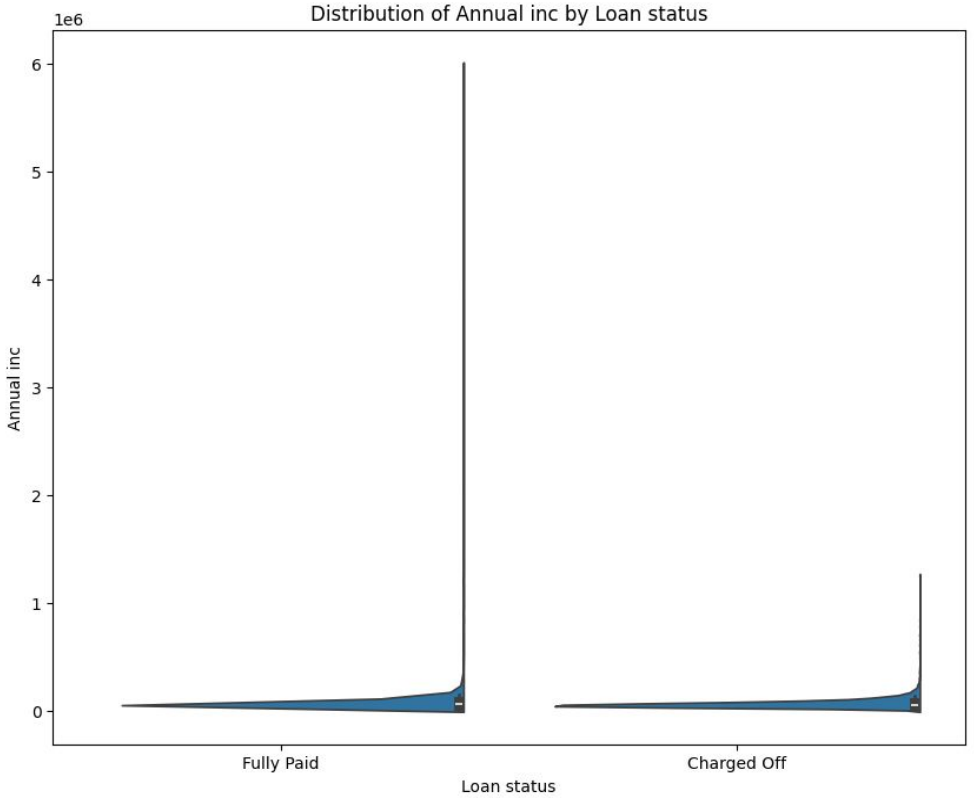




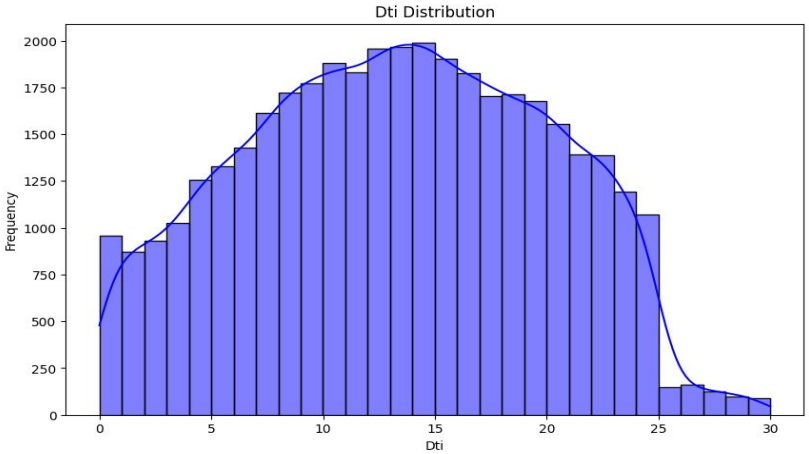
Annual Income

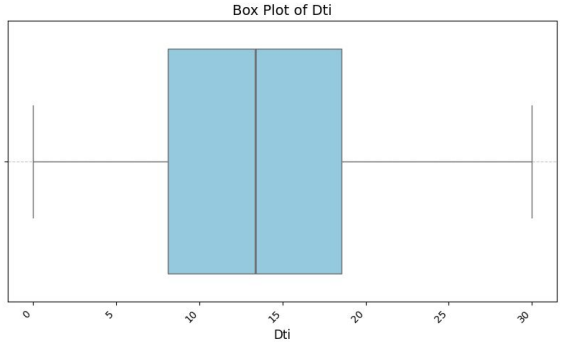


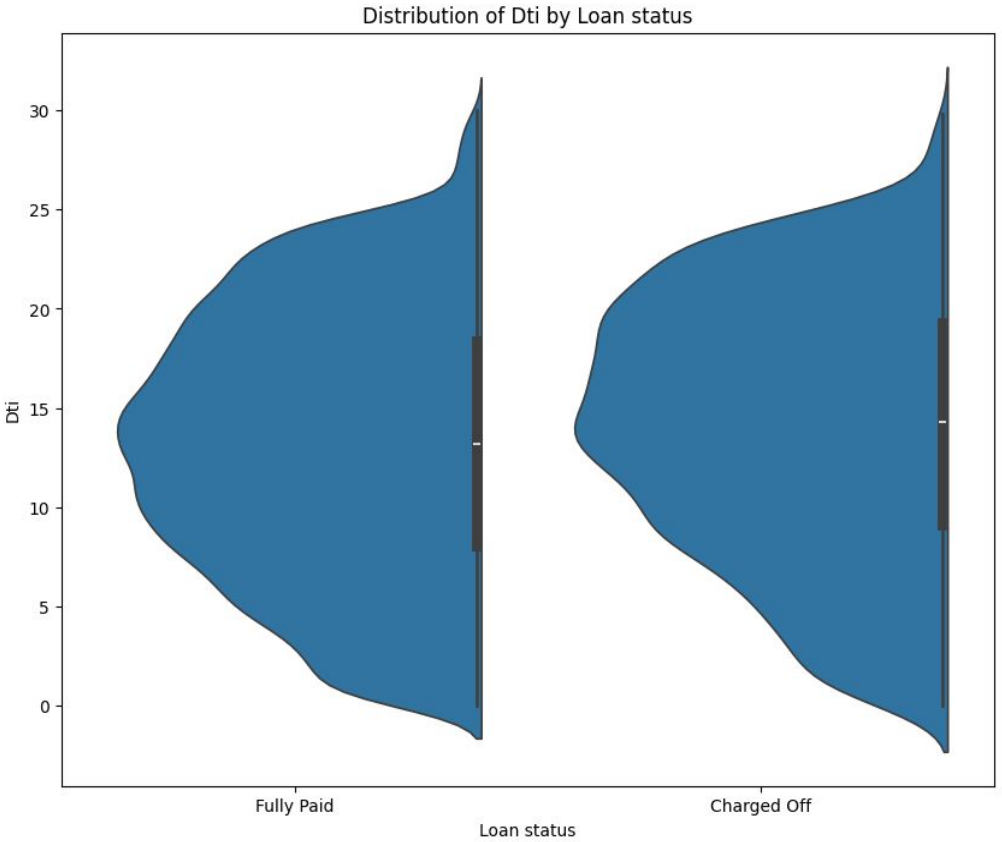




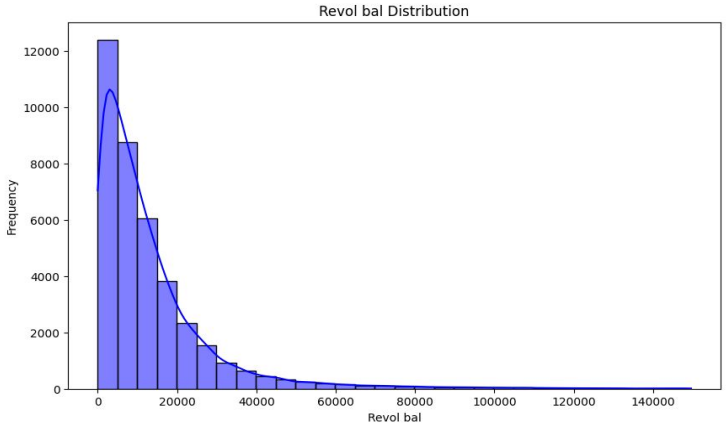
Debt to income ratio

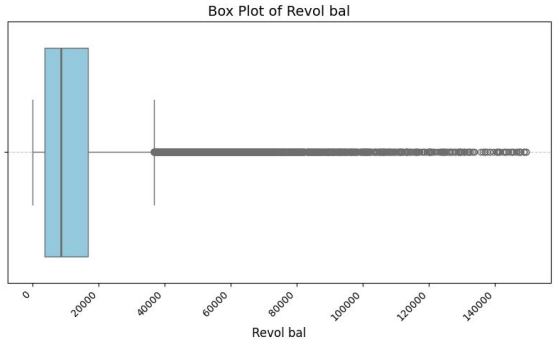


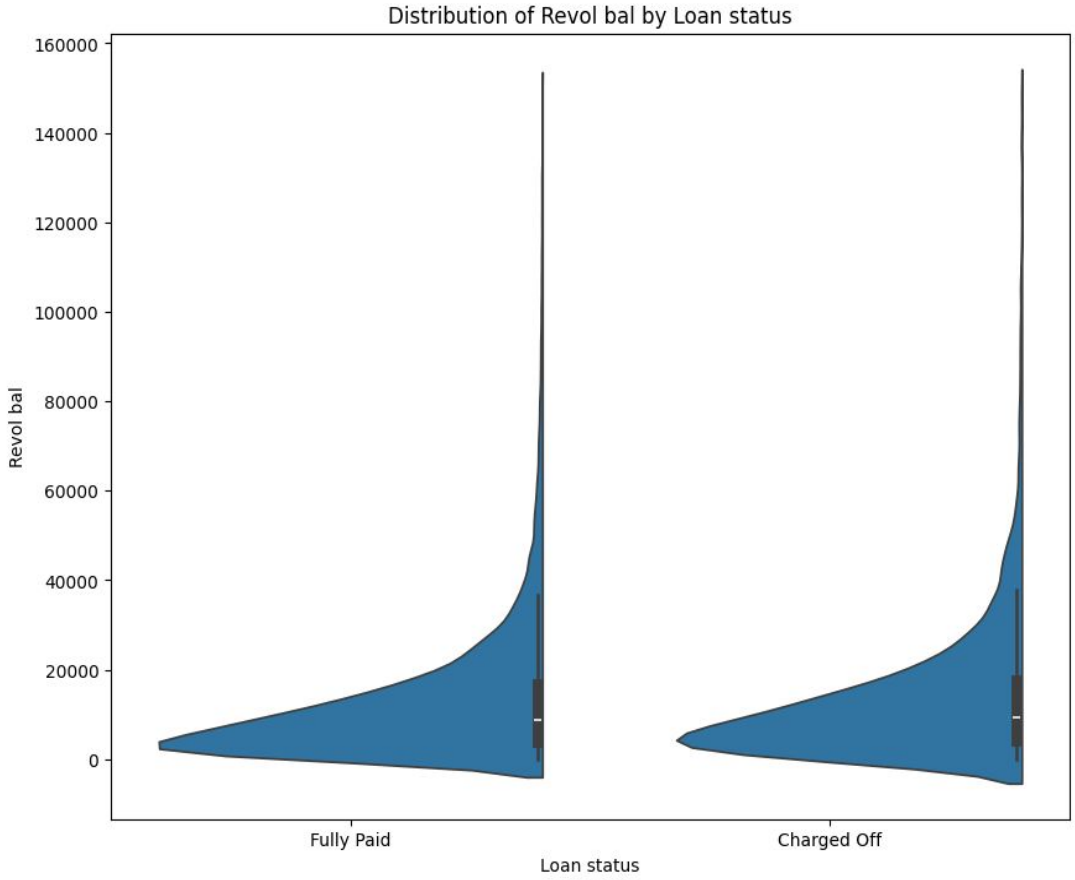




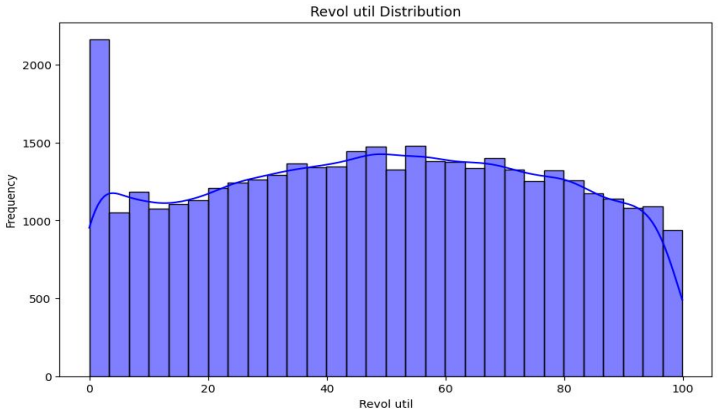
Revolving balance

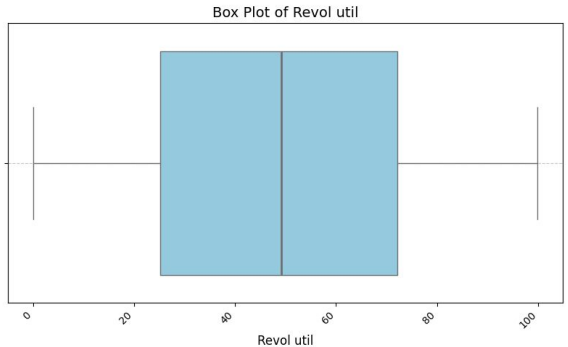


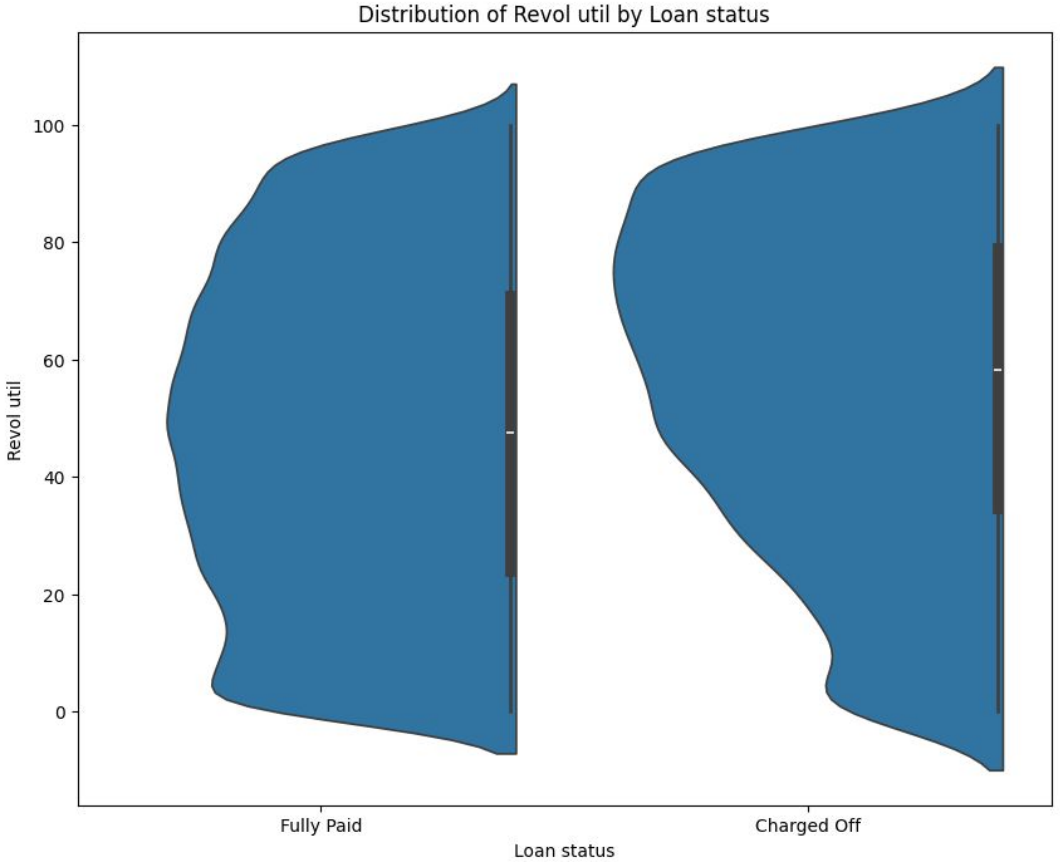




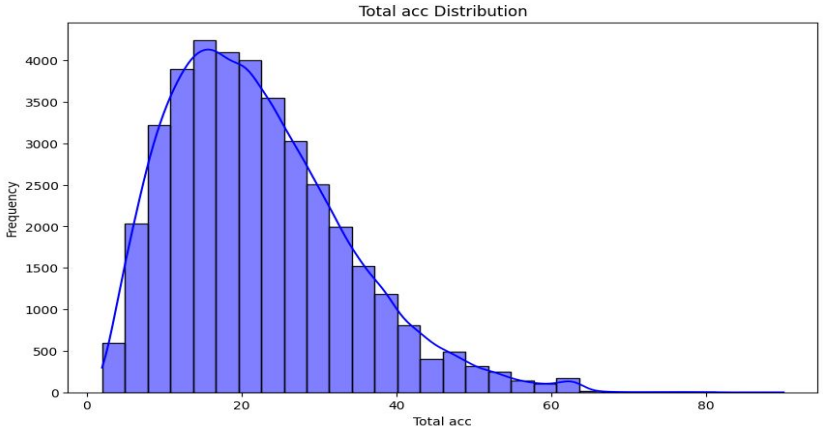
Revolving line utilization rate

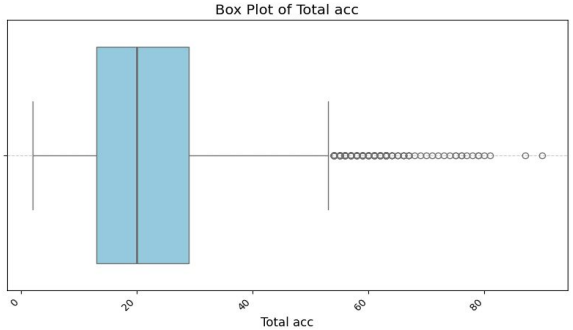


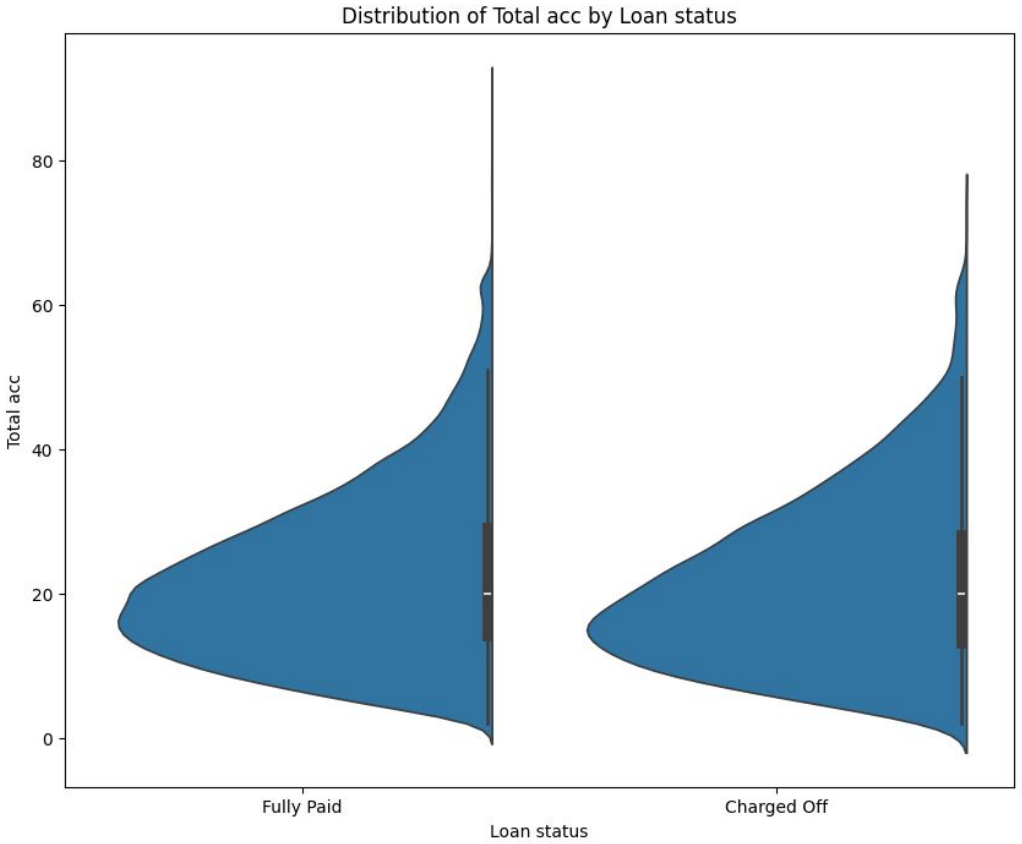




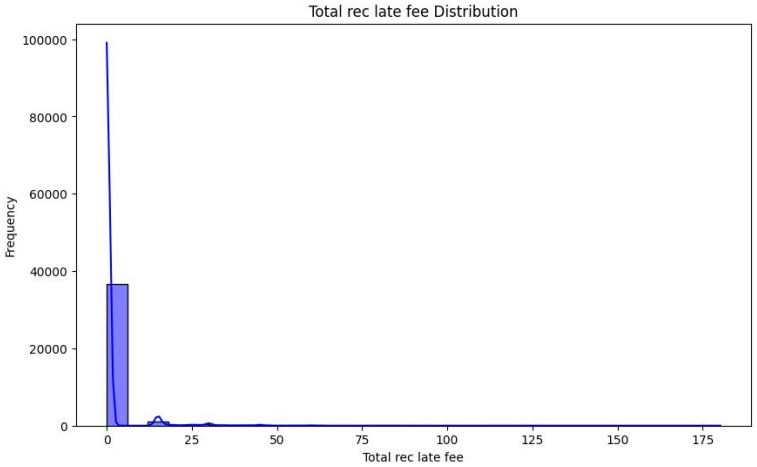
Total account

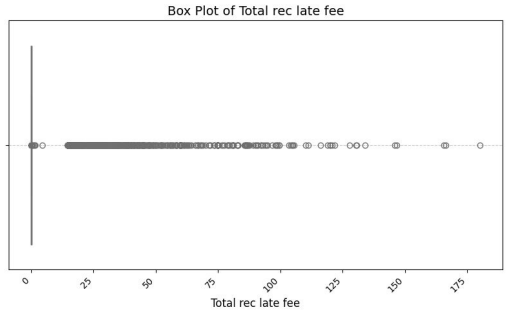


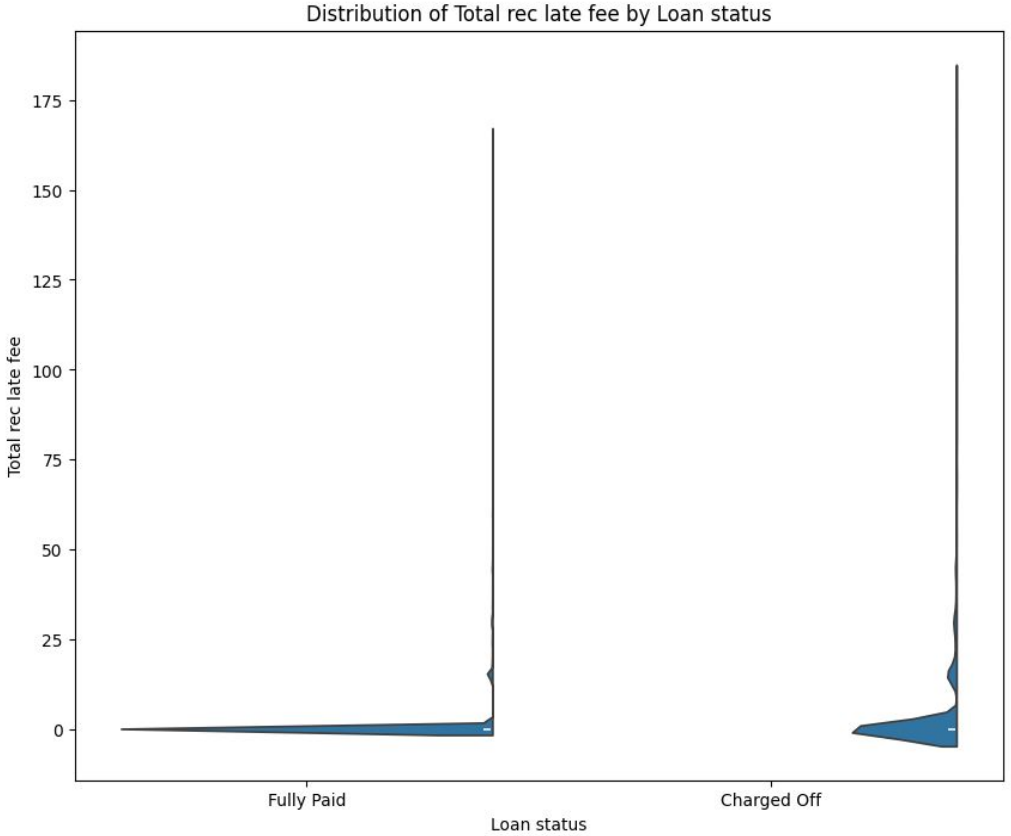




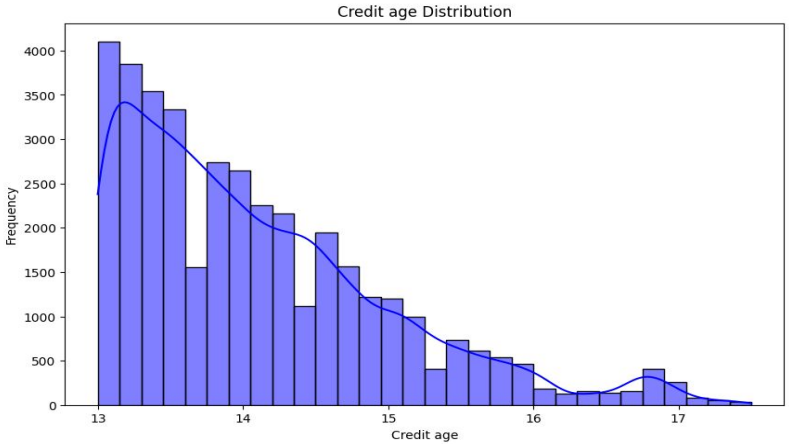
Total recovery late fee

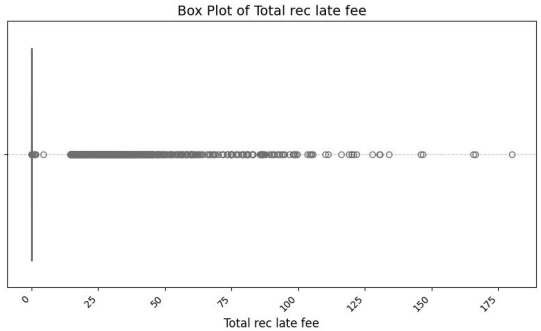


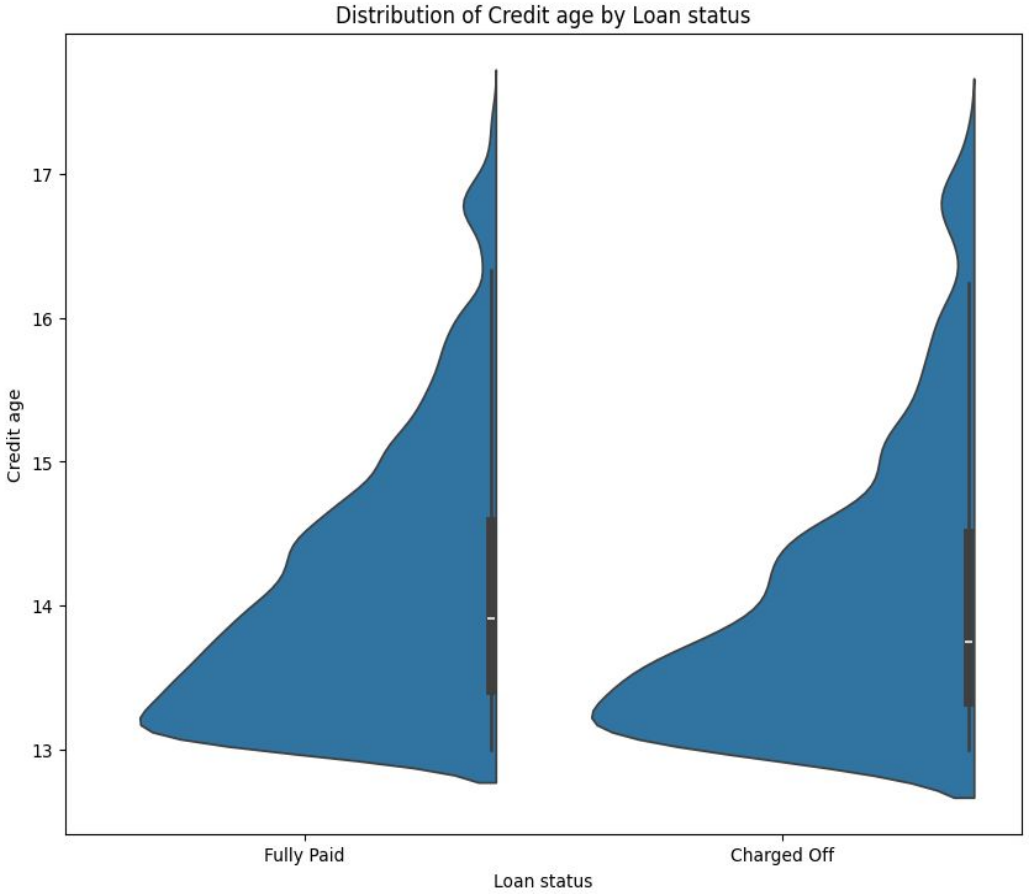




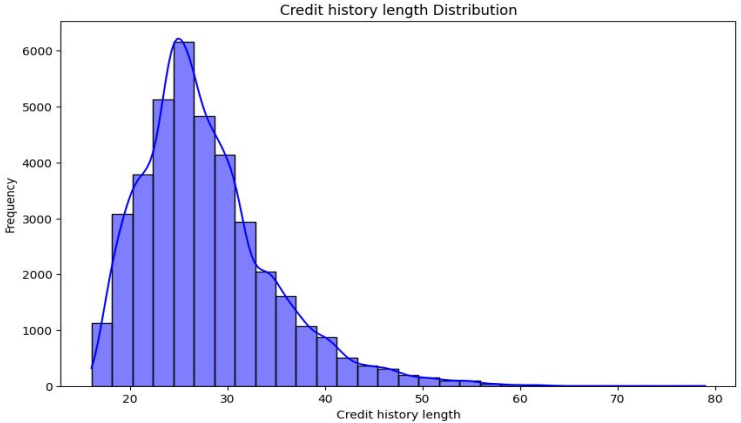
Credit Age

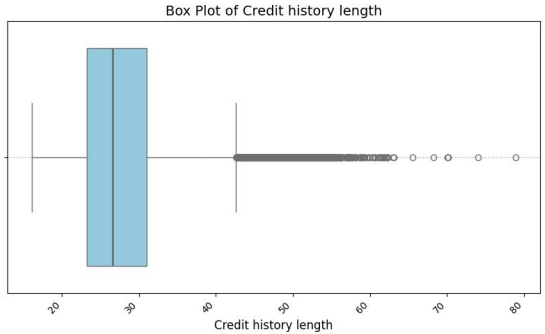

Credit age Distribution
Frequency
Credit age
0
500
1000
1500
2000
2500
3000
3500
4000
13
14
15
16
17

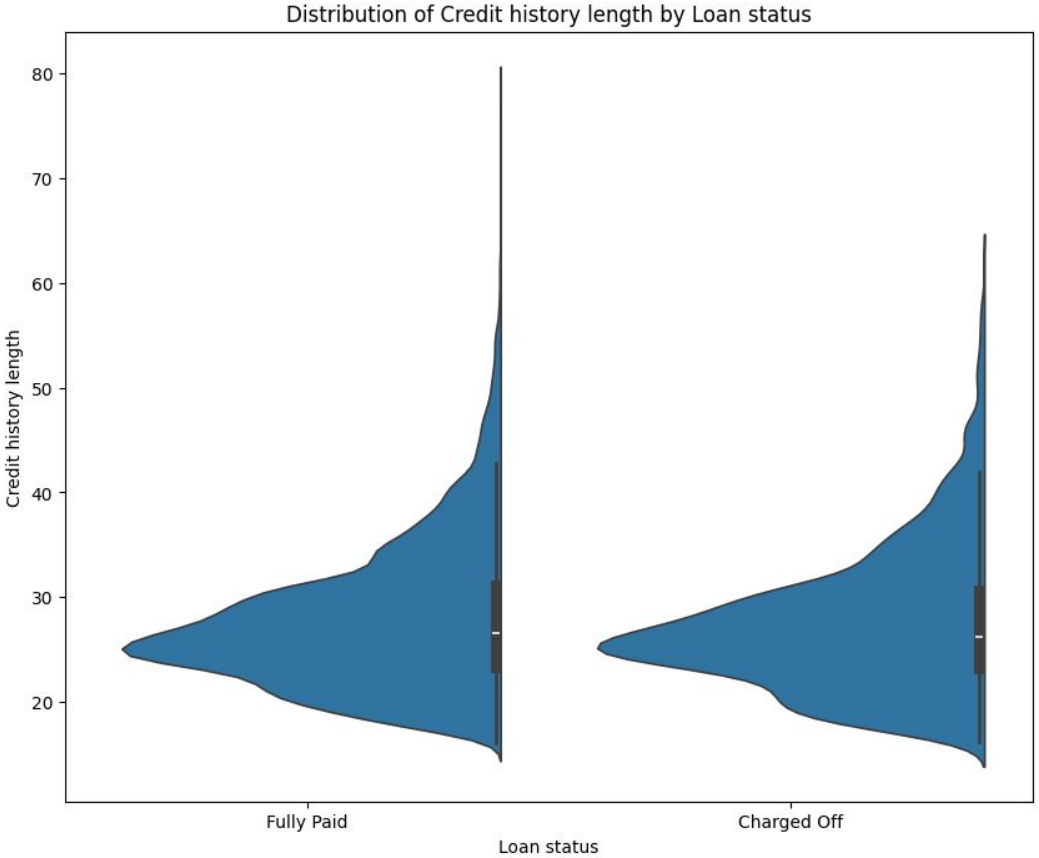

Box Plot of Total rec late fee
Total rec late fee
0
25
50
75
100
125
150
175


Distribution of Credit age by Loan status
Credit age
13
14
15
16
17
Fully Paid
Charged Off
Loan status

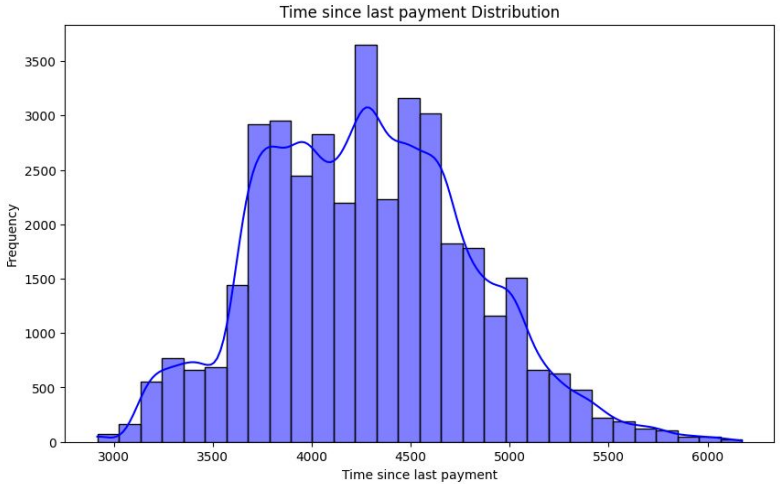
Credit History Length

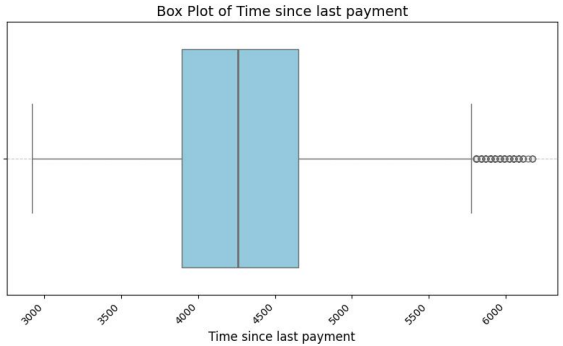






Time since last payment





Time since last credit pull