Loantap Business Case Study
PDF Link :
https://drive.google.com/file/d/12Hht1lKV5uqBdKT_kNUWEzQ8YRCVNIP8/view?usp=sharing

LoanTap is an online platform committed to delivering customized loan products to millennials. They innovate in an otherwise dull loan segment, to deliver instant, flexible loans on consumer friendly terms to salaried professionals and businessmen. The data science team at LoanTap is building an underwriting layer to determine the creditworthiness of MSMEs as well as individuals. LoanTap deploys formal credit to salaried individuals and businesses 4 main financial instruments:

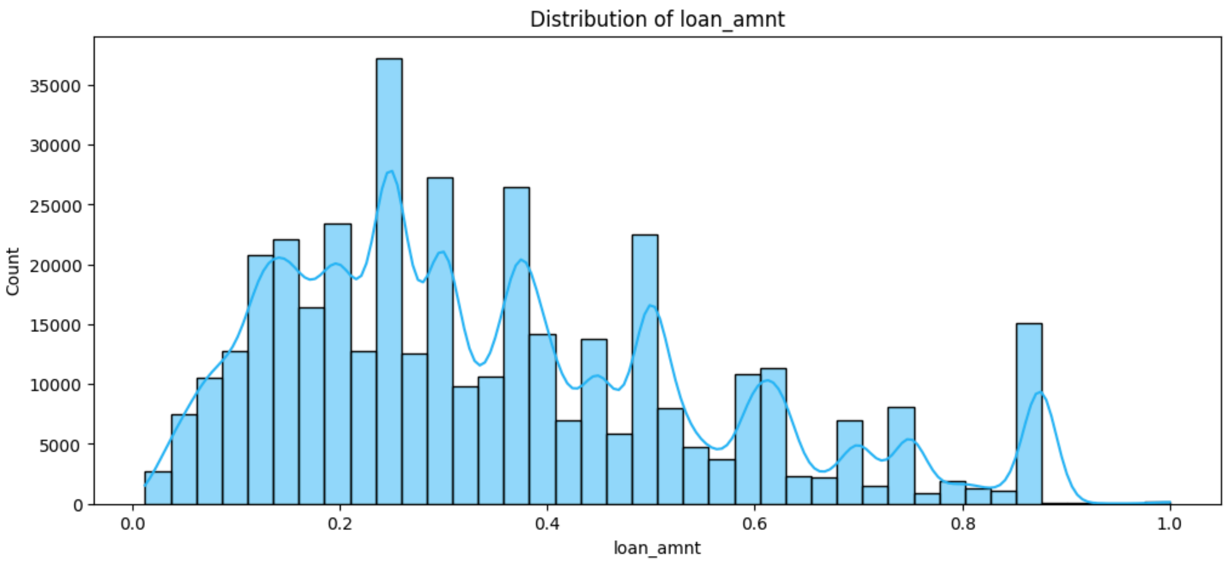
Exploratory Data Analysis

- There seems to be multiple instances of null values throughout the data
- These null values were handled in using the following methods
    1. Filling them with median value for numerical columns
    2. Filling them with the mode for the categorical columns
- There seem to be no duplicate values
- **Insights**
    - Many borrowers favor 36-month loan terms, indicating a preference for manageable monthly payments.
    - A significant number of applicants have mortgaged homes, which may signal financial stability or a reliance on property-backed loans.
    - The fact that most loans are fully repaid highlights borrowers' commitment to meeting their financial obligations and suggests strong lending criteria.
    - Differences in mean and median values for key metrics like loan amount and revolving balance hint at possible outliers.
    - The high number of individual applicants underscores the importance of personal loans as a key market segment.
    - Debt consolidation being the most common loan purpose suggests that borrowers frequently use loans to restructure or minimize high-interest debt.

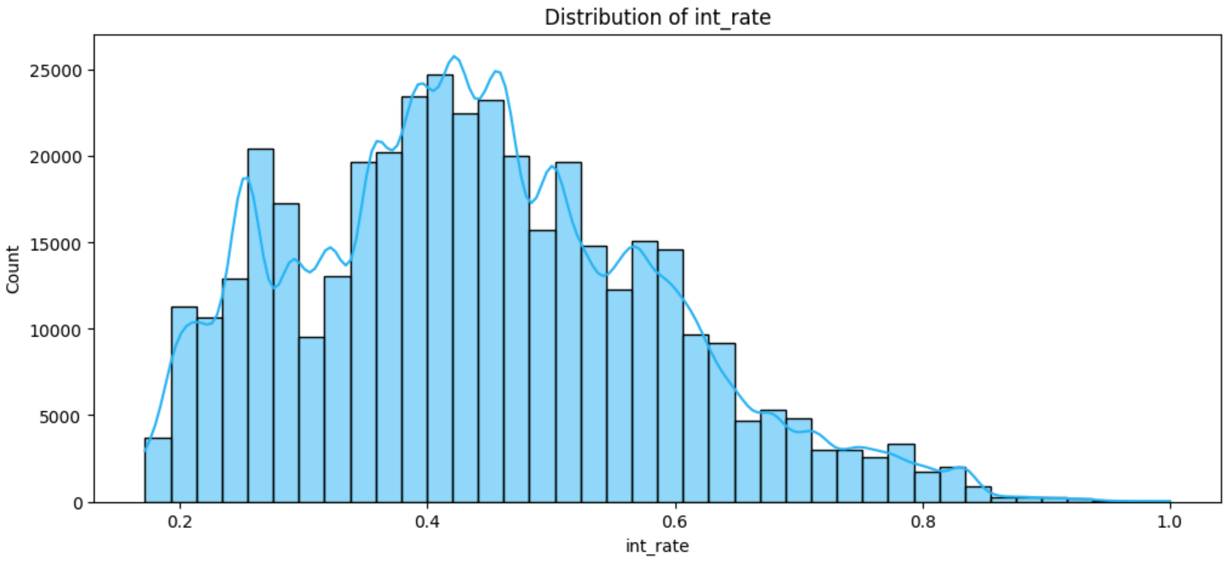
Univariate Analysis

Numerical data

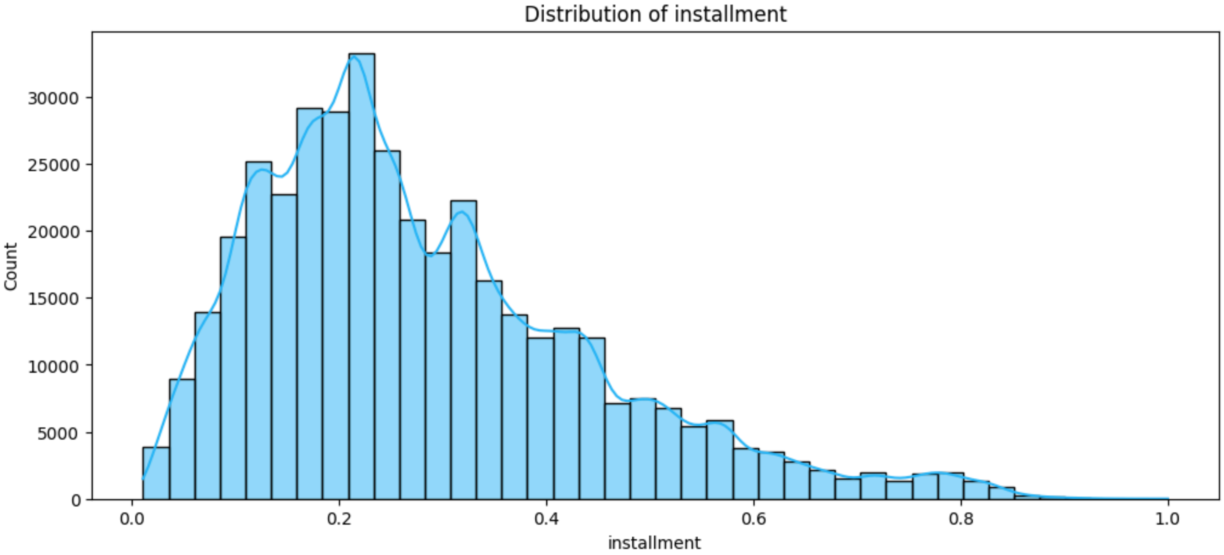
Loan amount



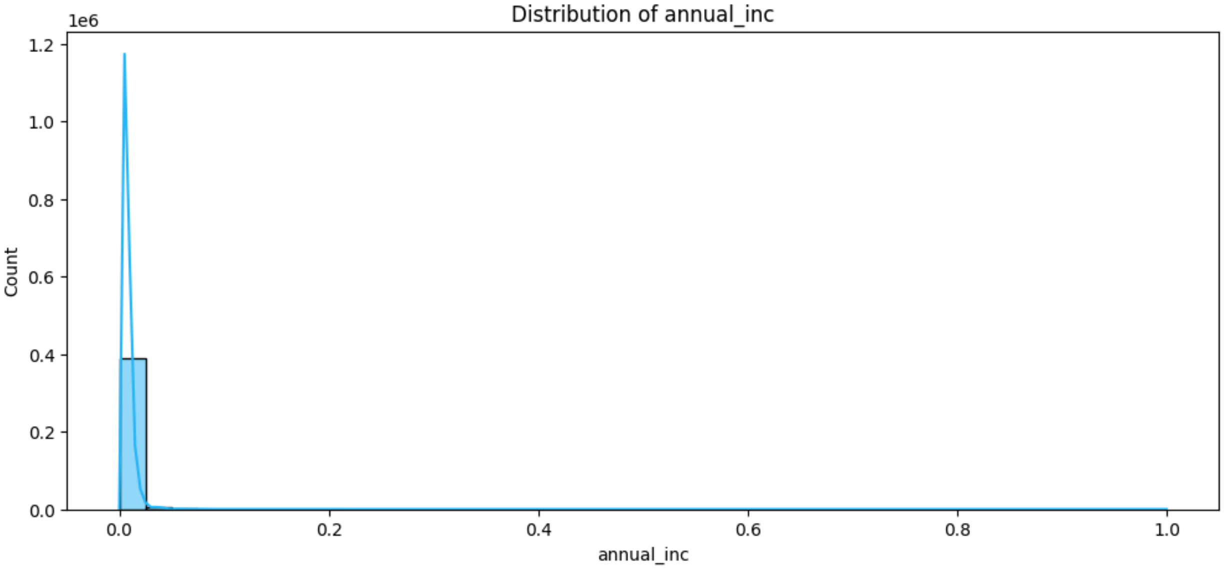
Interest rate



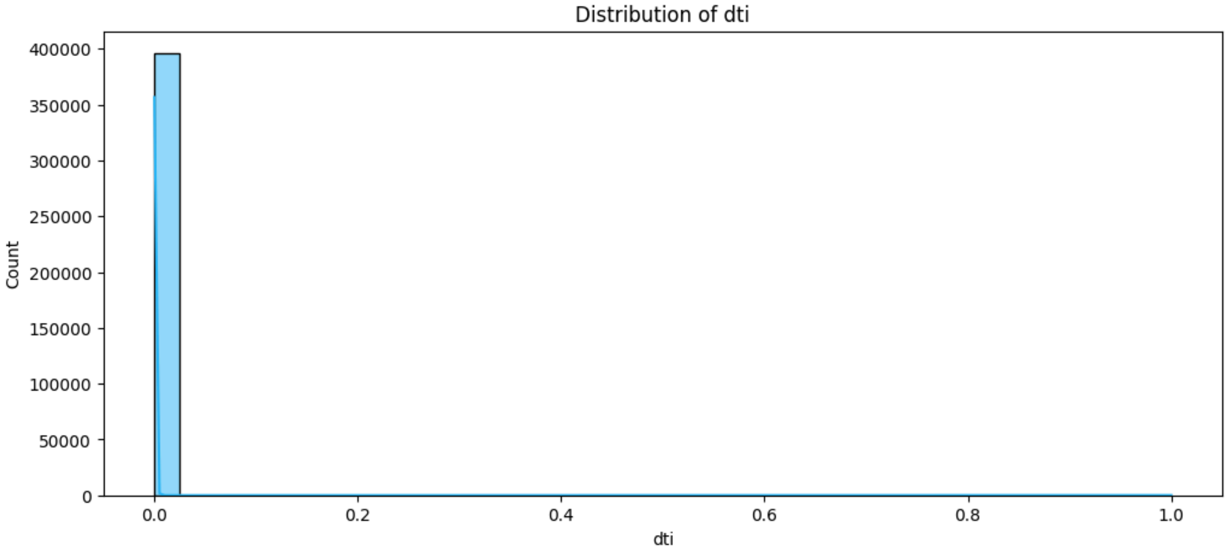
Installment



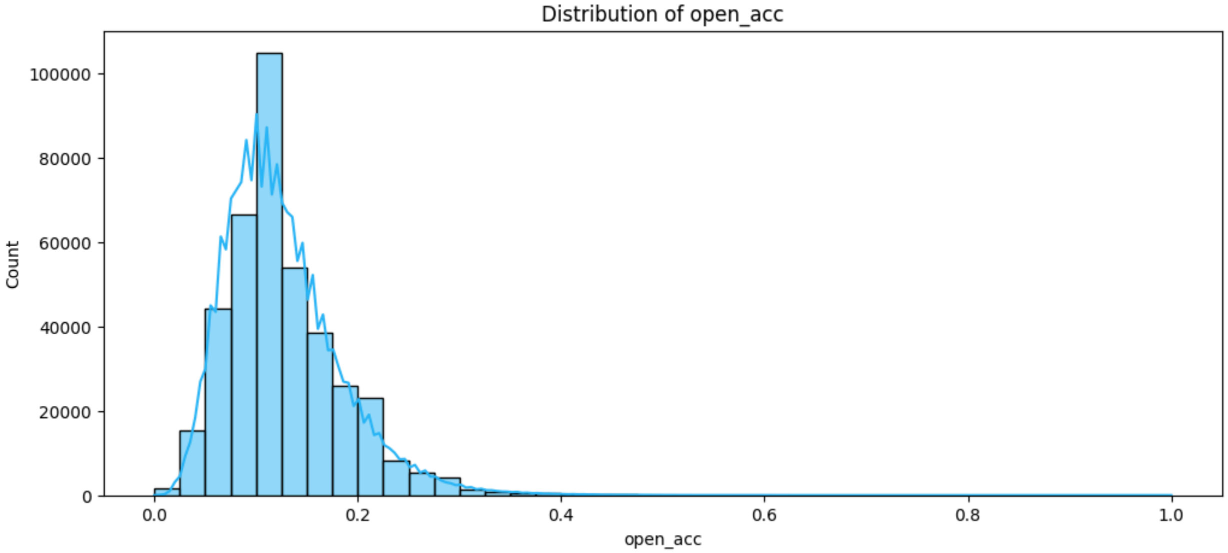
Annual Income



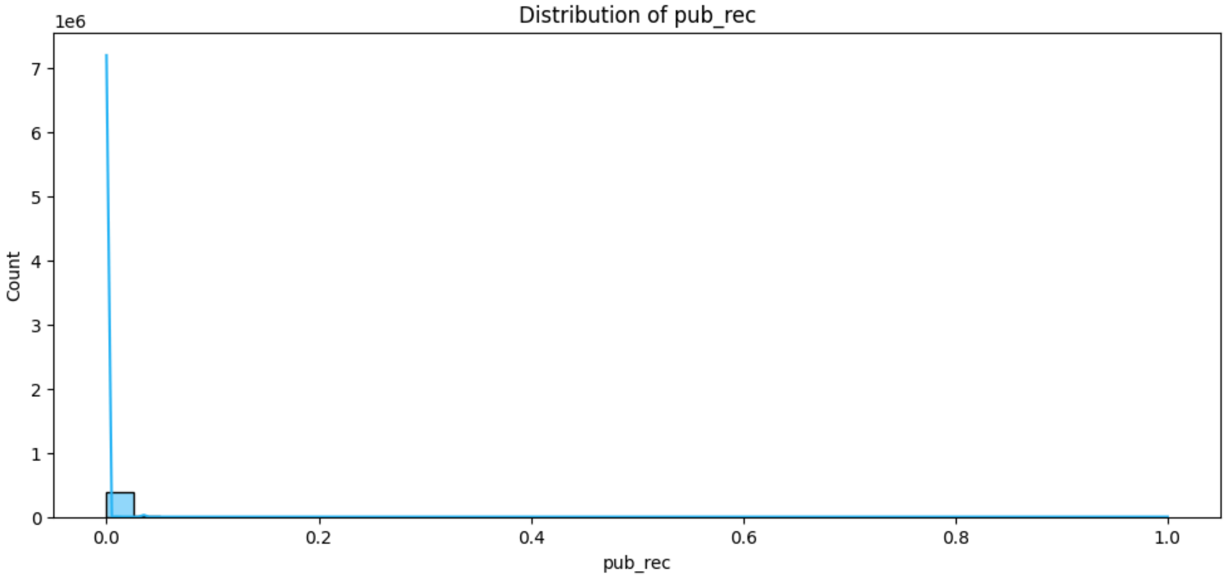
Debt to income ratio



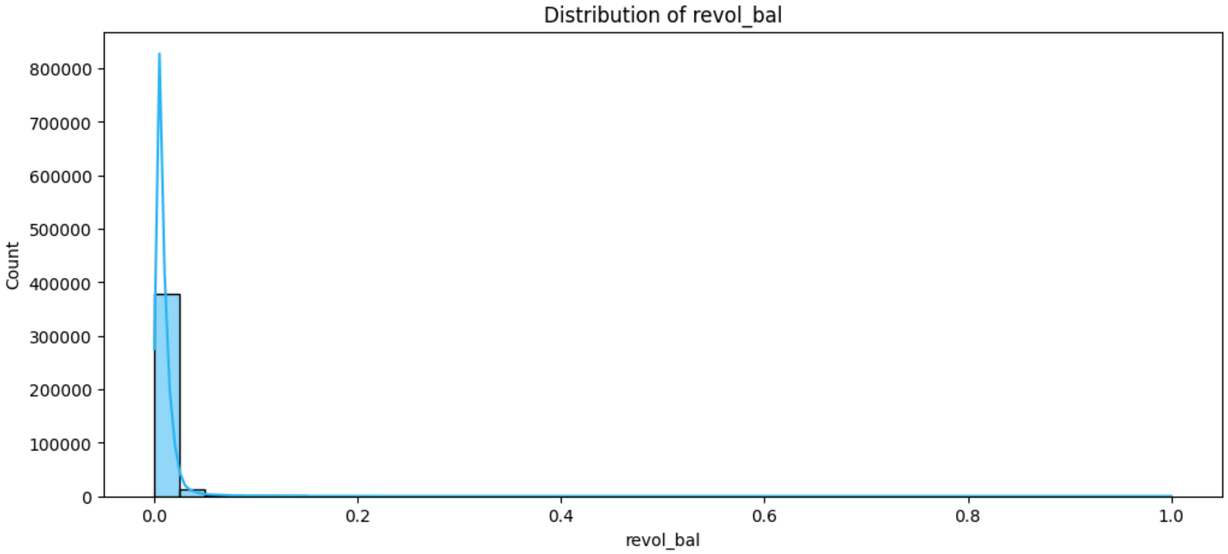
Open credit lines



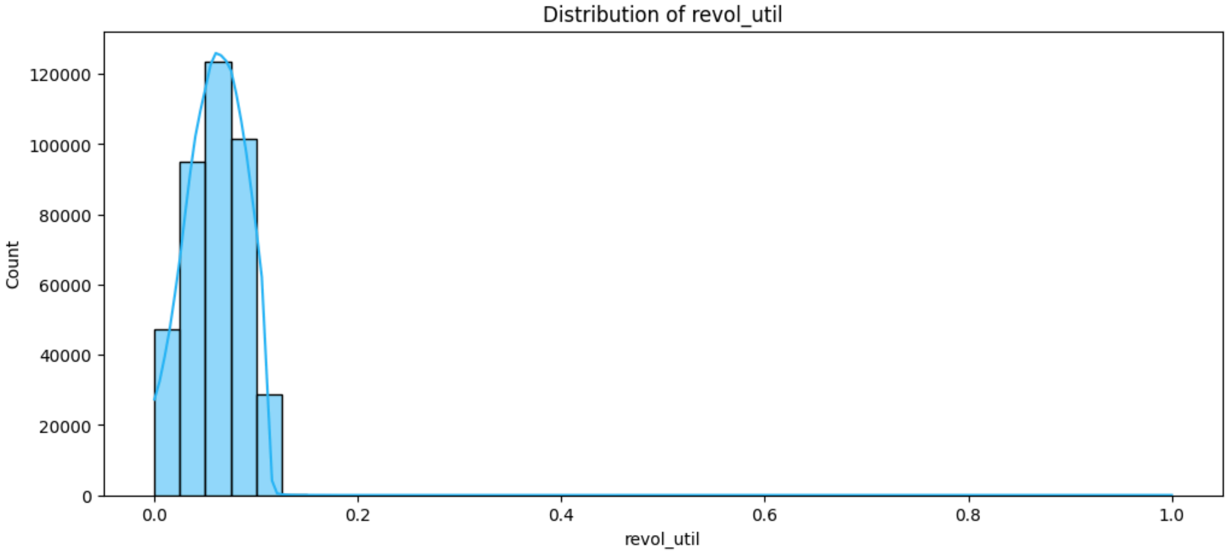
Number of derogatory public records



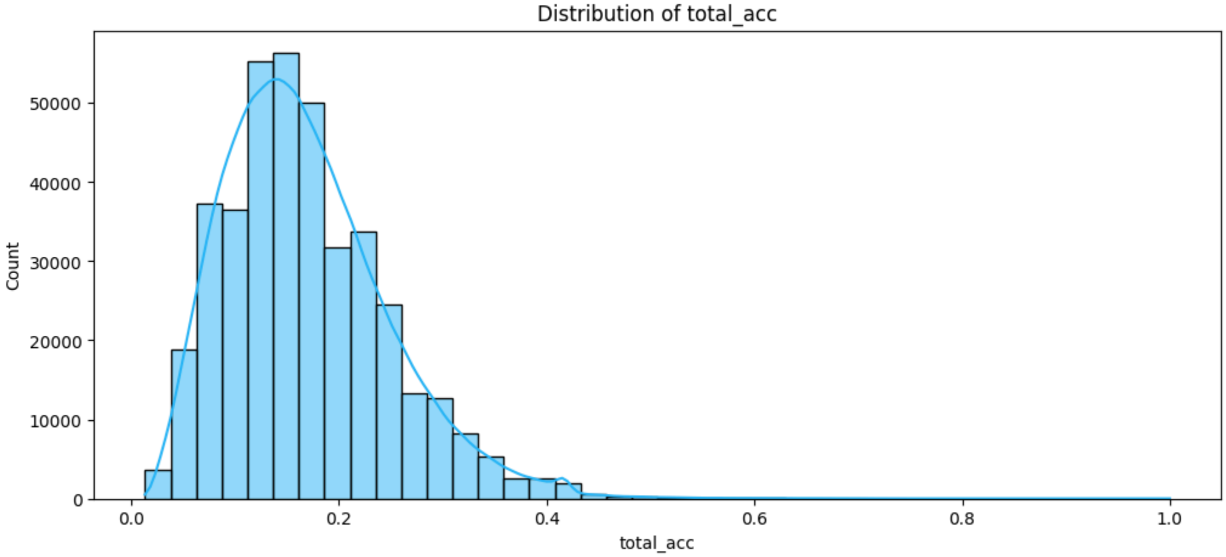
Credit Revolving balance



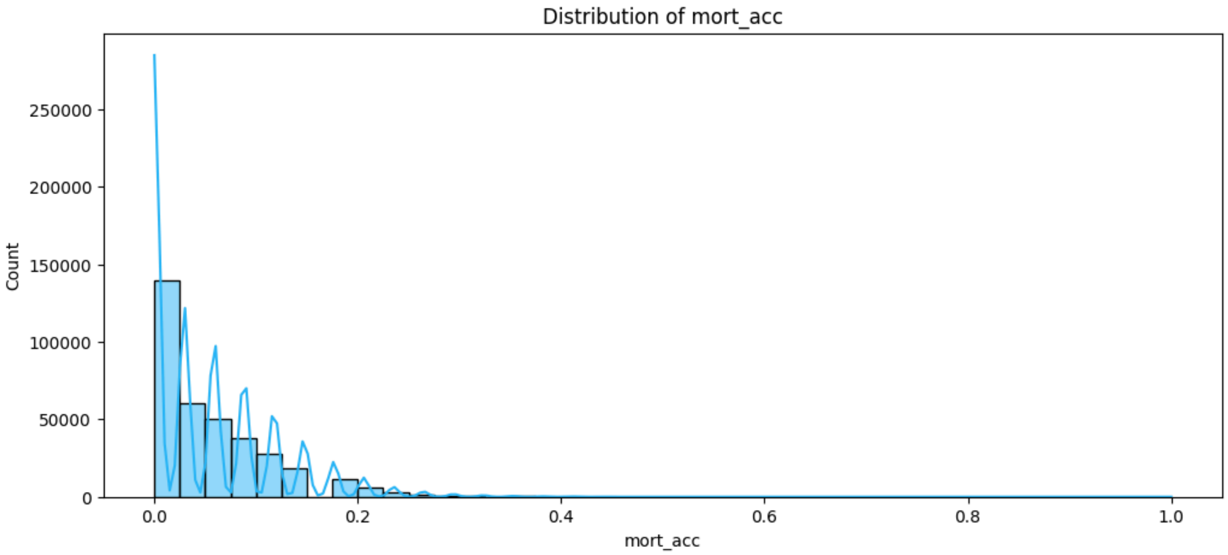
Revolving line utilization rate



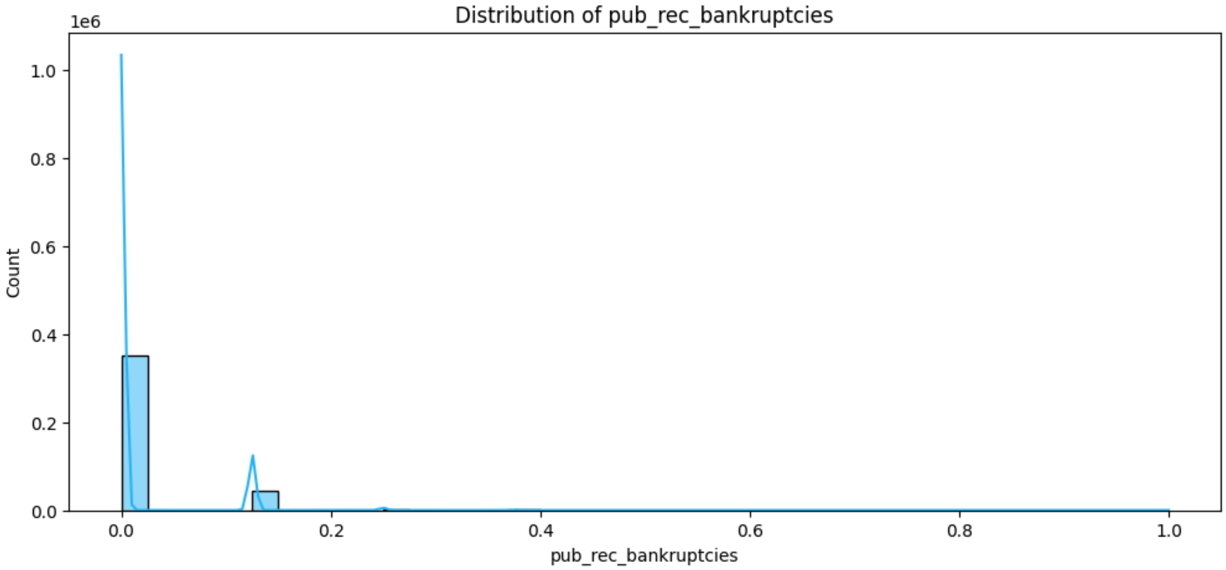
Total credit lines



Mortgage accounts

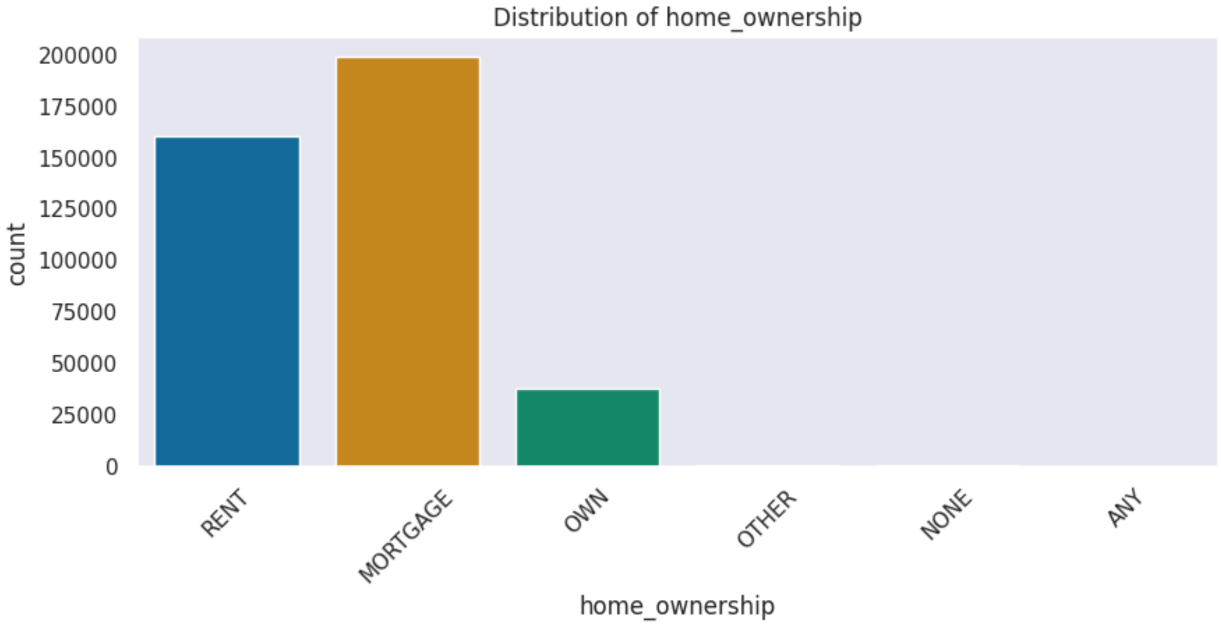


Public record bankruptcies

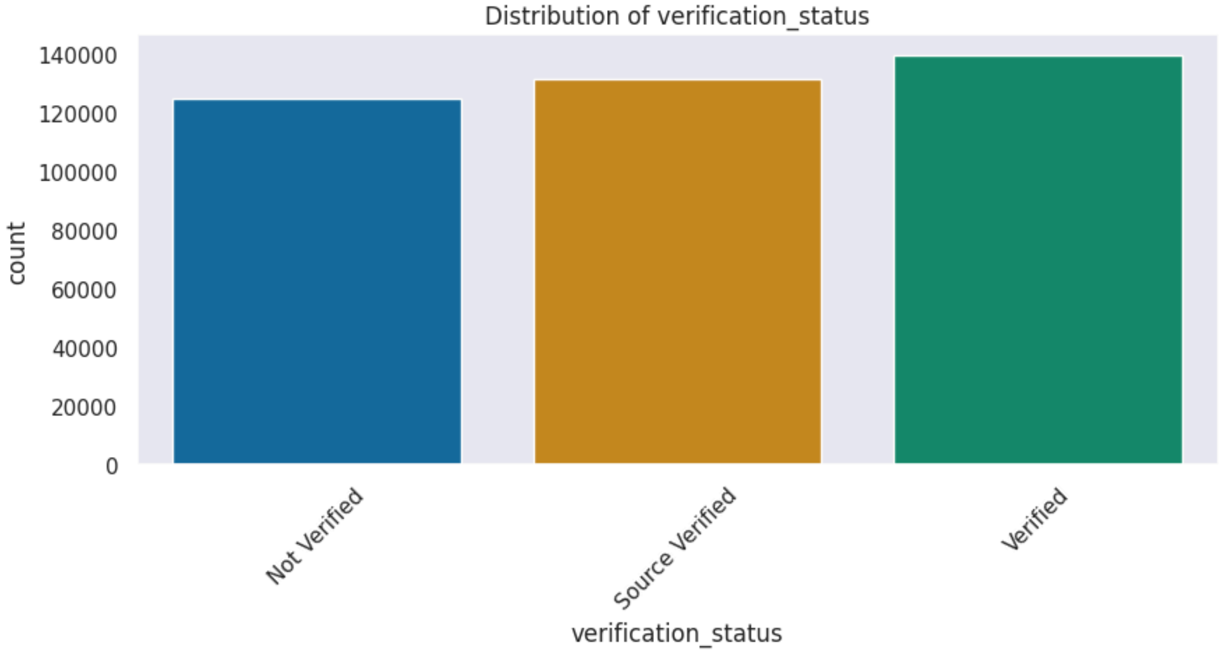


Categorical data

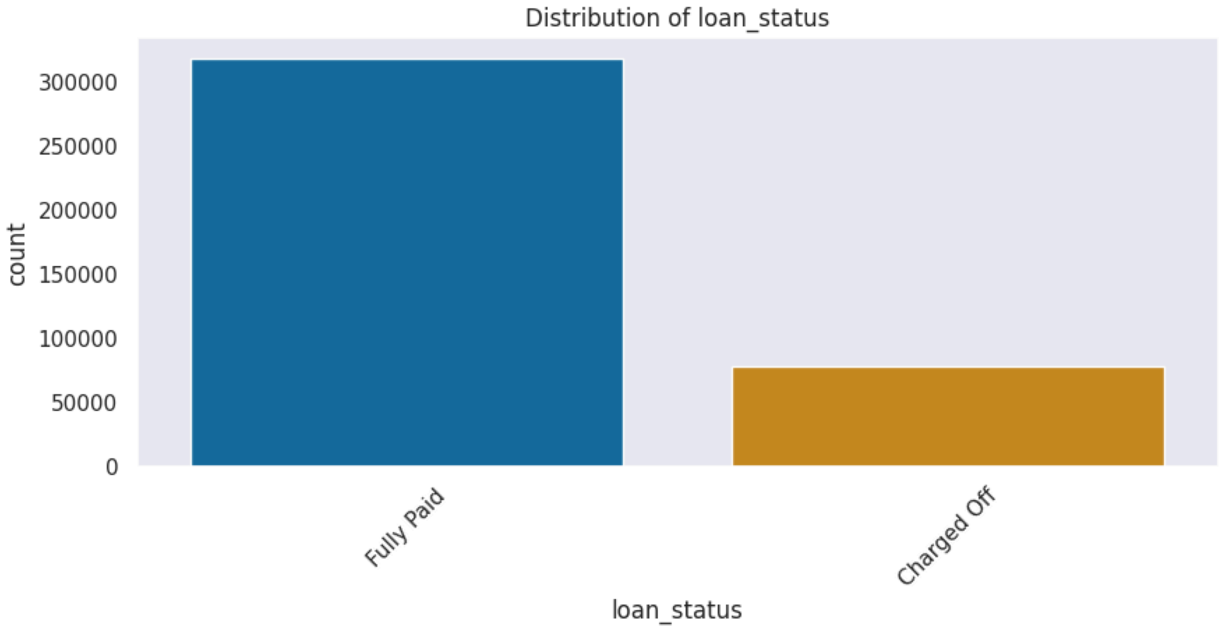
Home ownership



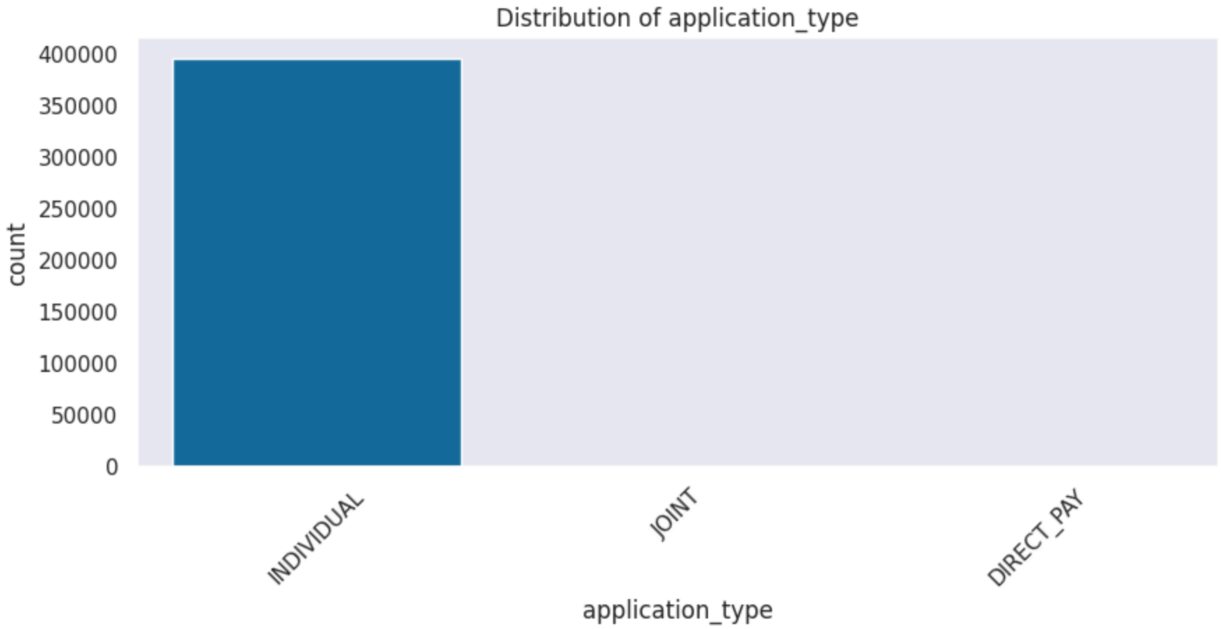
Verification Status



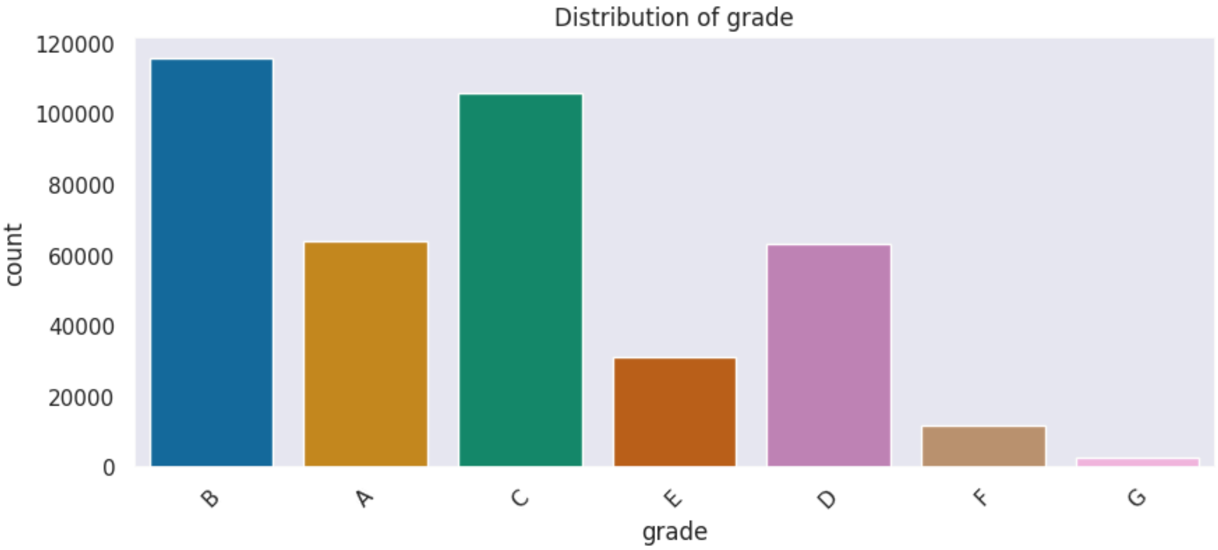
Loan status



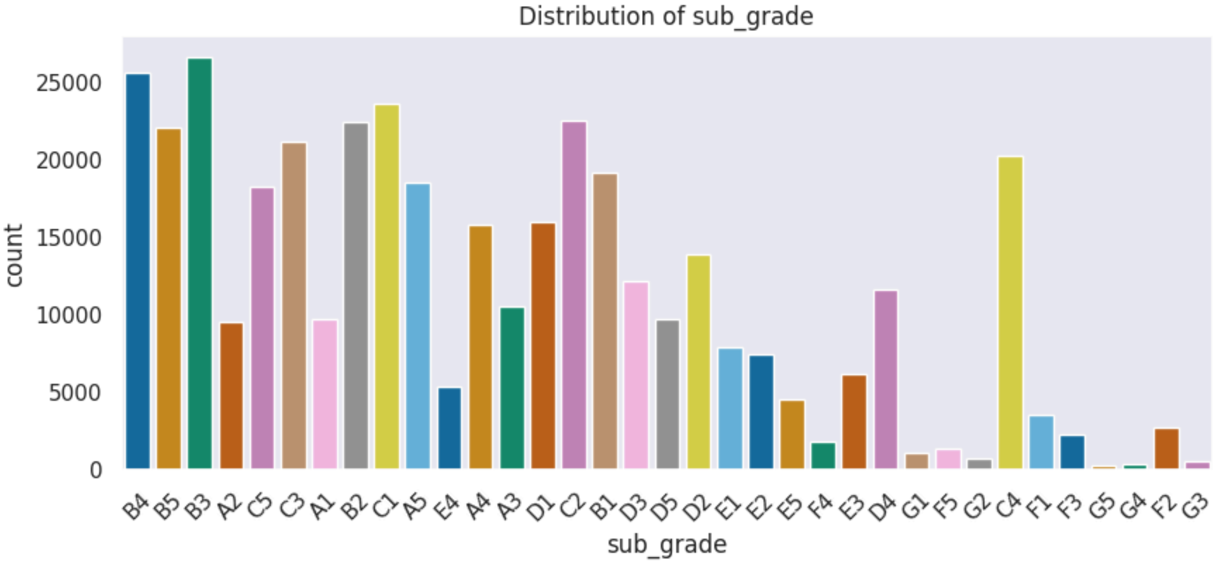
Application type



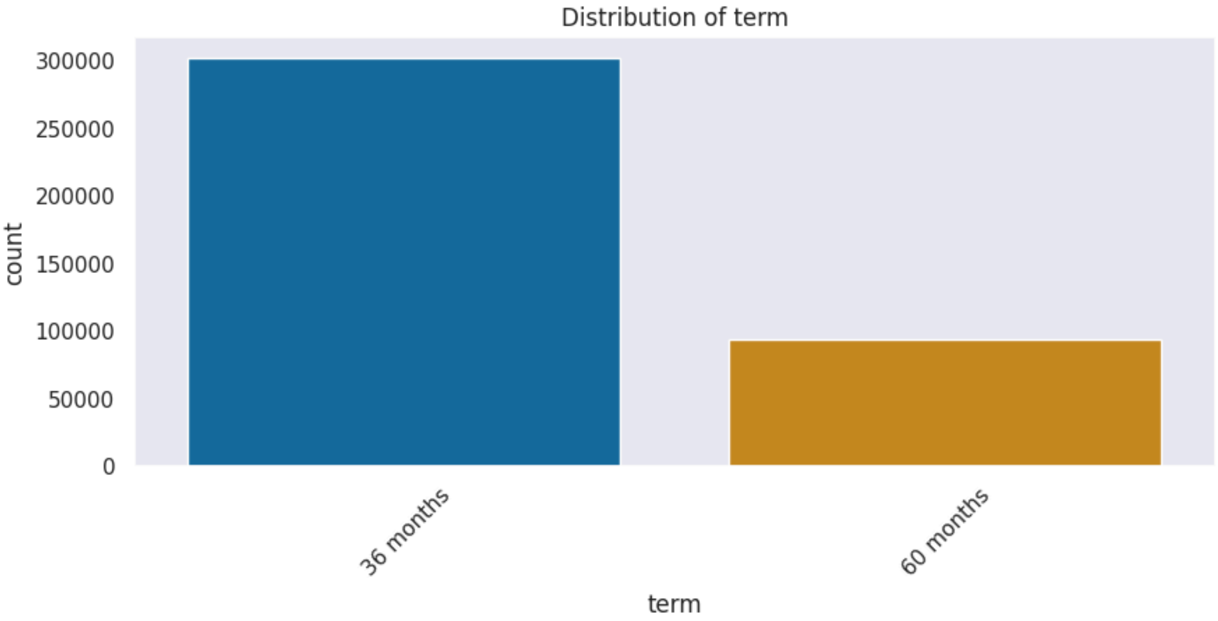
Grade



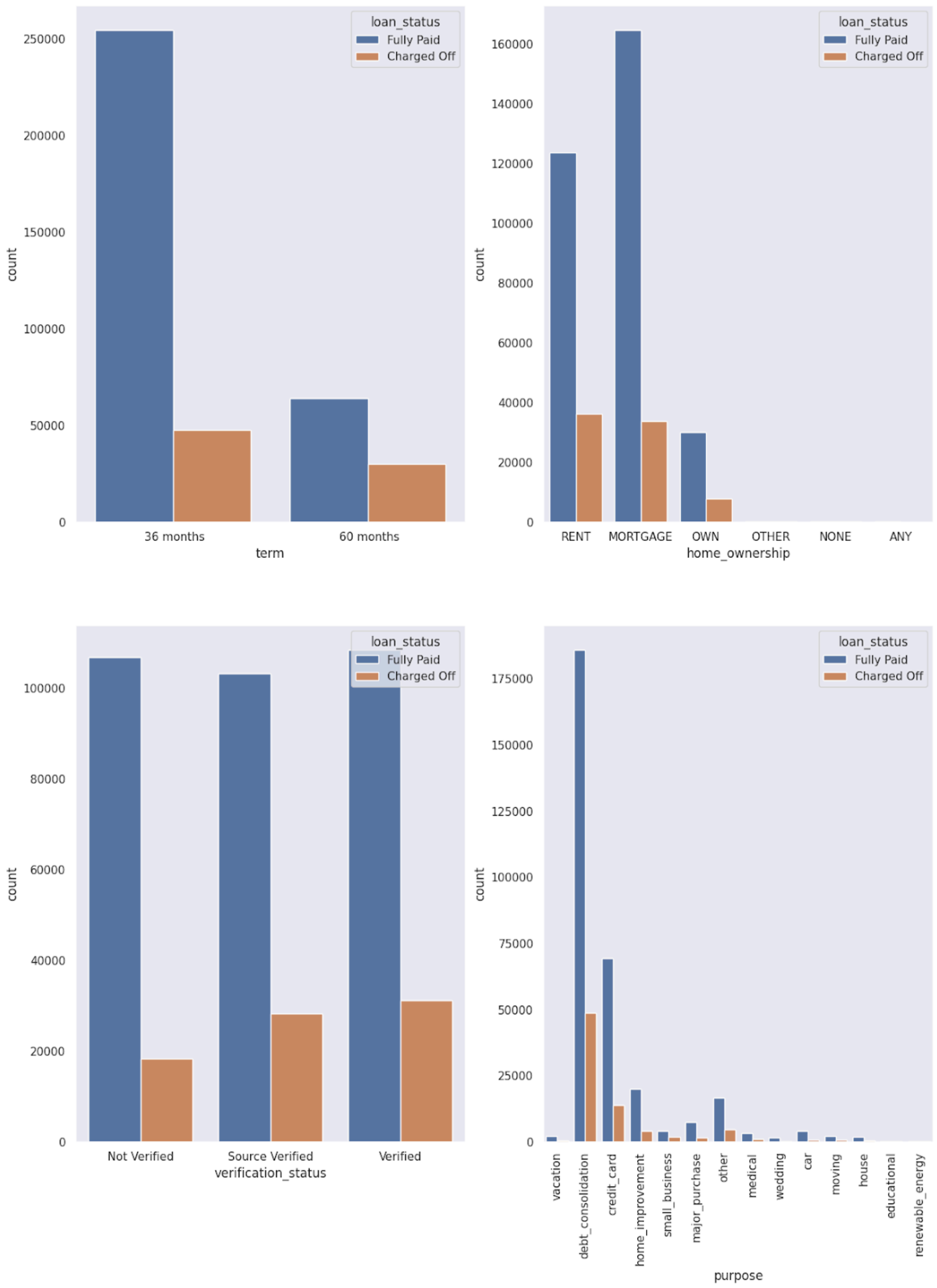
Sub grade

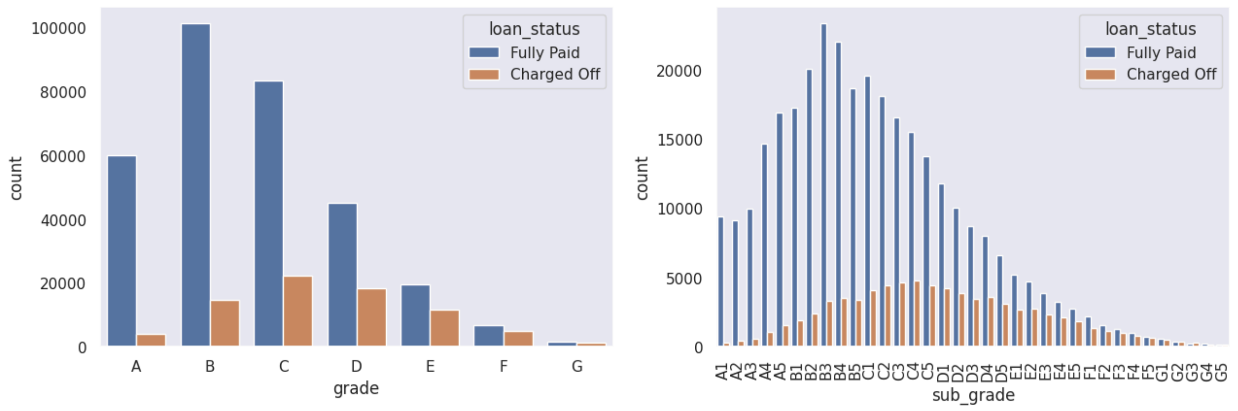


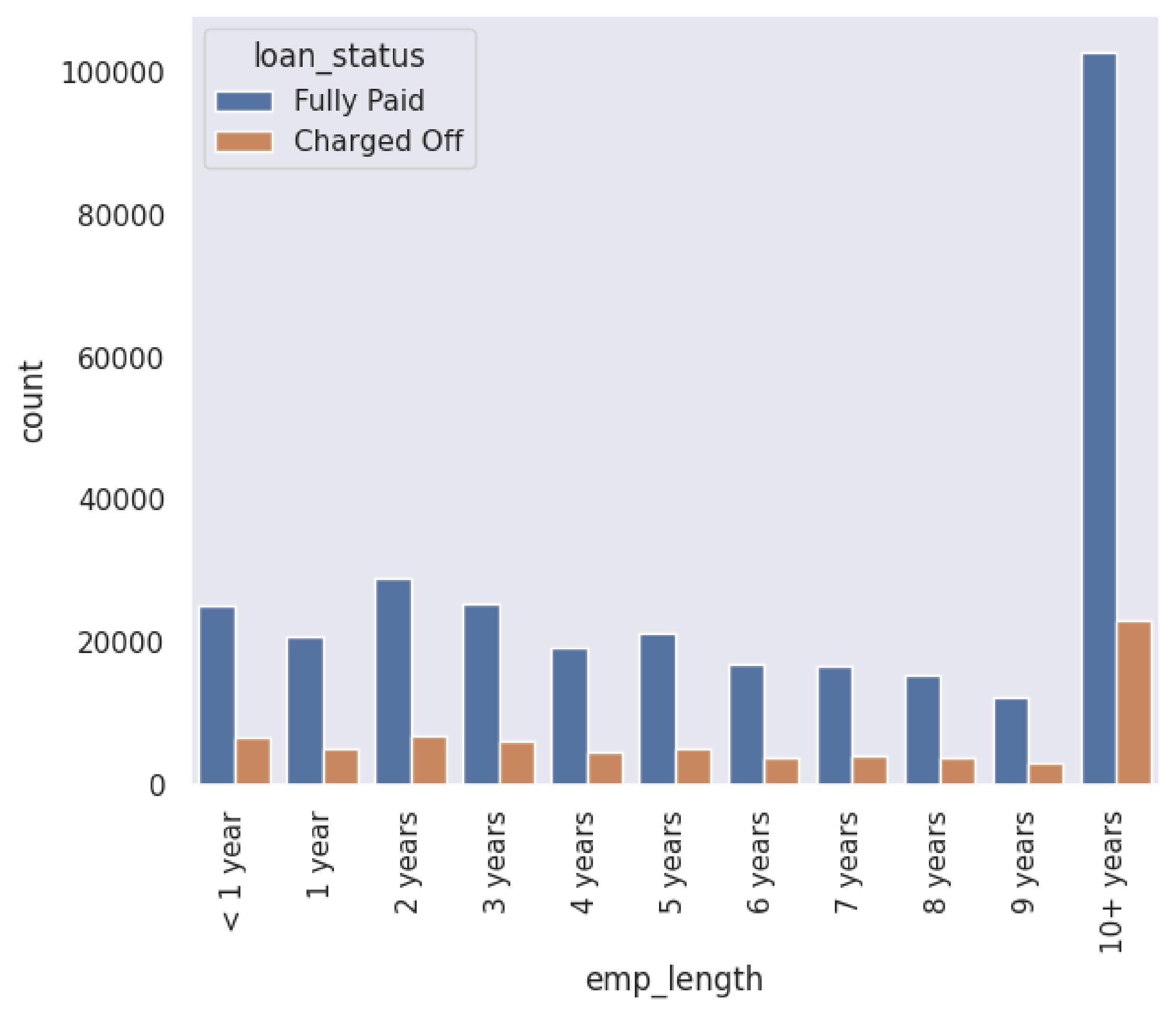
Term



Bivariate Analysis


loan_status
Fully Paid
Charged Off
count
250000
200000
150000
100000
50000
0
36 months
60 months
term
loan_status
Fully Paid
Charged Off
count
160000
140000
120000
100000
80000
60000
40000
20000
0
RENT
MORTGAGE
OWN
OTHER
NONE
ANY
home_ownership
loan_status
Fully Paid
Charged Off
count
100000
80000
60000
40000
20000
0
Not Verified
Source Verified
Verified
verification_status
loan_status
Fully Paid
Charged Off
count
175000
150000
125000
100000
75000
50000
25000
0
vacation
debt_consolidation
credit_card
home_improvement
small_business
major_purchase
other
medical
wedding
car
moving
house
educational
renewable_energy
purpose

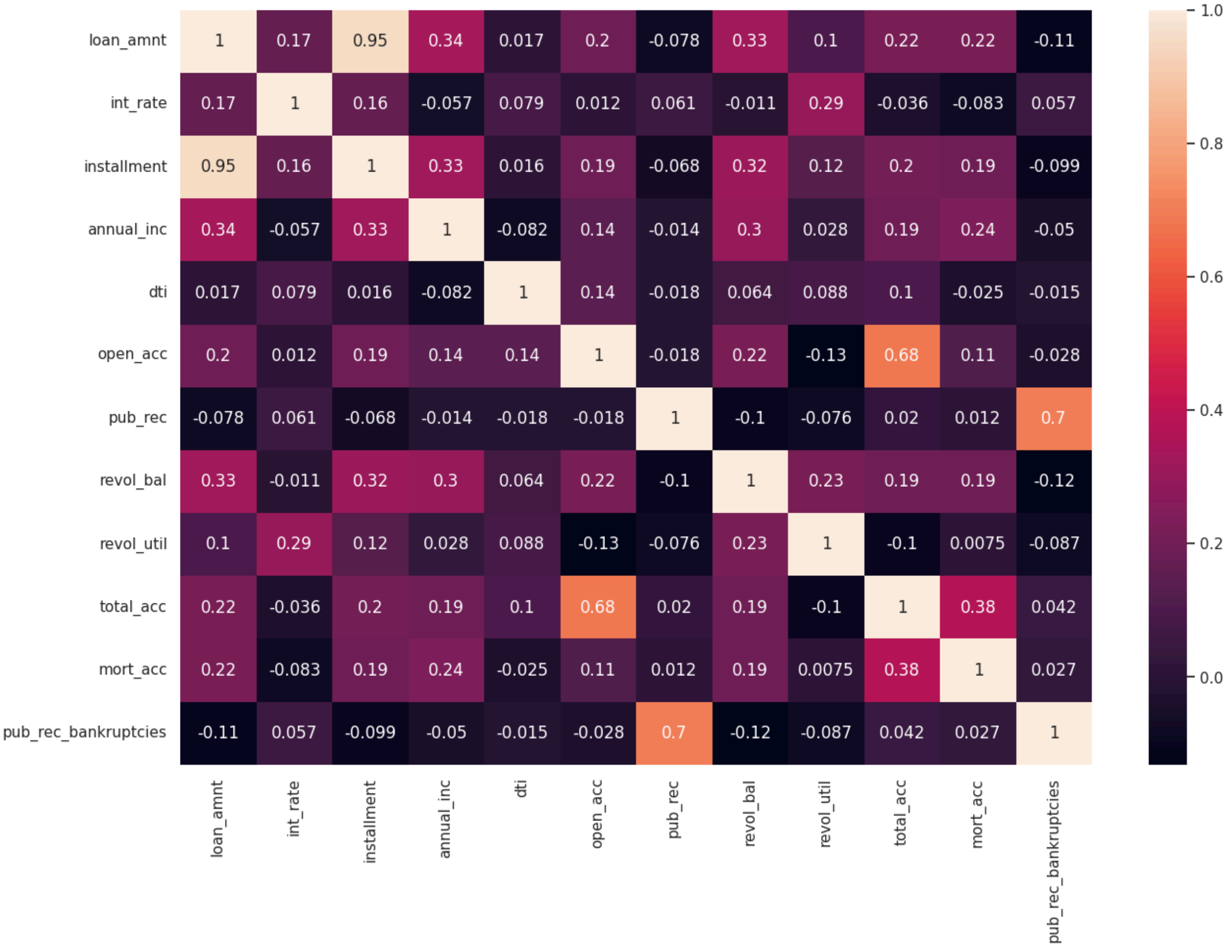

loan_status
Fully Paid
Charged Off
count
100000
80000
60000
40000
20000
0
A
B
C
D
E
F
G
grade
loan_status
Fully Paid
Charged Off
count
20000
15000
10000
5000
0
A1 A2 A3 A4 A5 B1 B2 B3 B4 B5 C1 C2 C3 C4 C5 D1 D2 D3 D4 D5 E1 E2 E3 E4 E5 F1 F2 F3 F4 F5 G1 G2 G3 G4 G5
sub_grade


loan_status
Fully Paid
Charged Off
count
100000
80000
60000
40000
20000
0
< 1 year
1 year
2 years
3 years
4 years
5 years
6 years
7 years
8 years
9 years
10+ years
emp_length

Insights

- **Loan Duration:** The 36-month loan term is the most favored, showing a strong completion rate among borrowers.
- **Types of Loans:** Mortgages and rental loans are the most common, while debt consolidation loans remain a popular choice.
- **Credit Profiles:** Borrowers with a "B" credit grade and a "B3" subgrade tend to have the best repayment history.
- **Professions & Approval:** Managers and teachers see the highest loan approval rates, indicating lender confidence in these occupations.
- **Repayment Behavior:** Those with over a decade of employment display a solid history of timely loan repayments.

Correlation Matrix



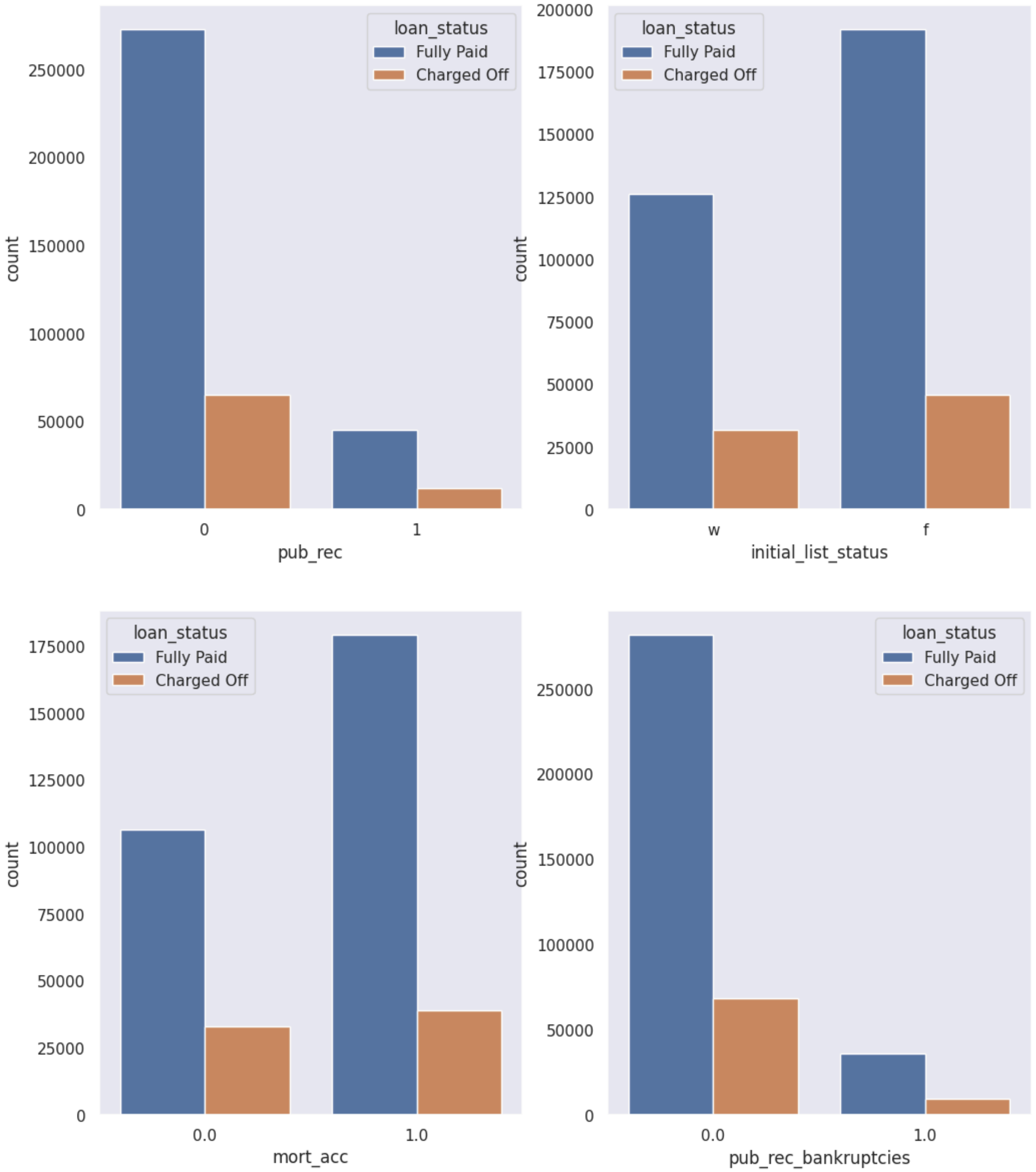
	loan_amnt	int_rate	installment	annual_inc	dti	open_acc	pub_rec	revol_bal	revol_util	total_acc	mort_acc	pub_rec_bankruptcies
loan_amnt	1	0.17	0.95	0.34	0.017	0.2	-0.078	0.33	0.1	0.22	0.22	-0.11
int_rate	0.17	1	0.16	-0.057	0.079	0.012	0.061	-0.011	0.29	-0.036	-0.083	0.057
installment	0.95	0.16	1	0.33	0.016	0.19	-0.068	0.32	0.12	0.2	0.19	-0.099
annual_inc	0.34	-0.057	0.33	1	-0.082	0.14	-0.014	0.3	0.028	0.19	0.24	-0.05
dti	0.017	0.079	0.016	-0.082	1	0.14	-0.018	0.064	0.088	0.1	-0.025	-0.015
open_acc	0.2	0.012	0.19	0.14	0.14	1	-0.018	0.22	-0.13	0.68	0.11	-0.028
pub_rec	-0.078	0.061	-0.068	-0.014	-0.018	-0.018	1	-0.1	-0.076	0.02	0.012	0.7
revol_bal	0.33	-0.011	0.32	0.3	0.064	0.22	-0.1	1	0.23	0.19	0.19	-0.12
revol_util	0.1	0.29	0.12	0.028	0.088	-0.13	-0.076	0.23	1	-0.1	0.0075	-0.087
total_acc	0.22	-0.036	0.2	0.19	0.1	0.68	0.02	0.19	-0.1	1	0.38	0.042
mort_acc	0.22	-0.083	0.19	0.24	-0.025	0.11	0.012	0.19	0.0075	0.38	1	0.027
pub_rec_bankruptcies	-0.11	0.057	-0.099	-0.05	-0.015	-0.028	0.7	-0.12	-0.087	0.042	0.027	1

Insights

- Higher annual income (annual_inc) is linked to bigger loan amounts, as people with more earnings can afford larger loans.
- Installment amount (installment) has a slight positive link, since bigger loans usually come with higher installments.
- Total accounts (total_acc) and mortgage accounts (mort_acc) show a weak positive link—those with more credit history might qualify for larger loans.
- Annual income (annual_inc) has a slight negative link—borrowers with higher income often get lower interest rates.
- People with higher incomes may also have more credit accounts, leading to a weak positive connection between total accounts (total_acc) and mortgage accounts (mort_acc).
- Revolving balance (revol_bal) and credit utilization (revol_util) have a positive link—those with higher credit balances tend to use more of their available credit.
- Number of open accounts (open_acc) and total accounts (total_acc) have a weak positive relationship, as those with more open accounts usually have more total accounts.

Data preprocessing

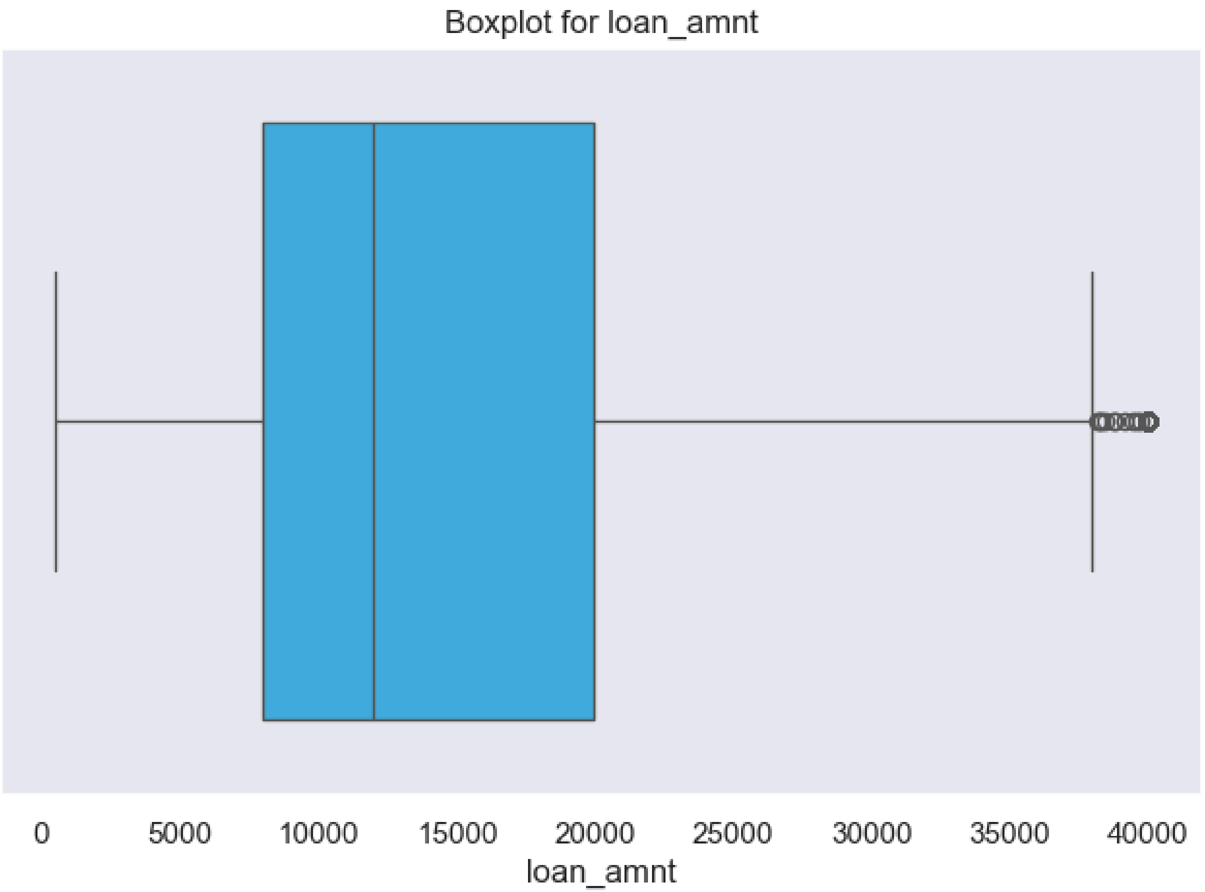
The columns pub_rec, mort_acc and pub_rec_bankruptcies were encoded

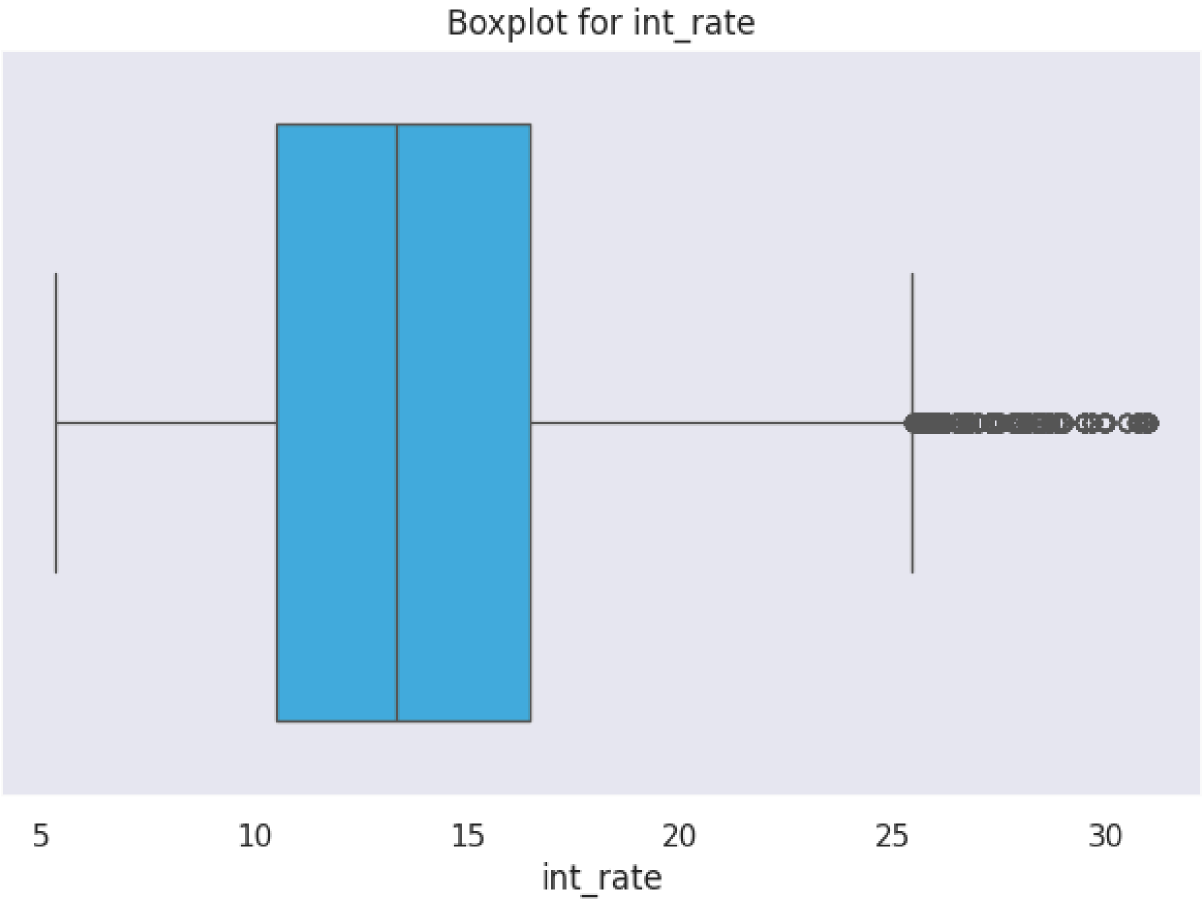


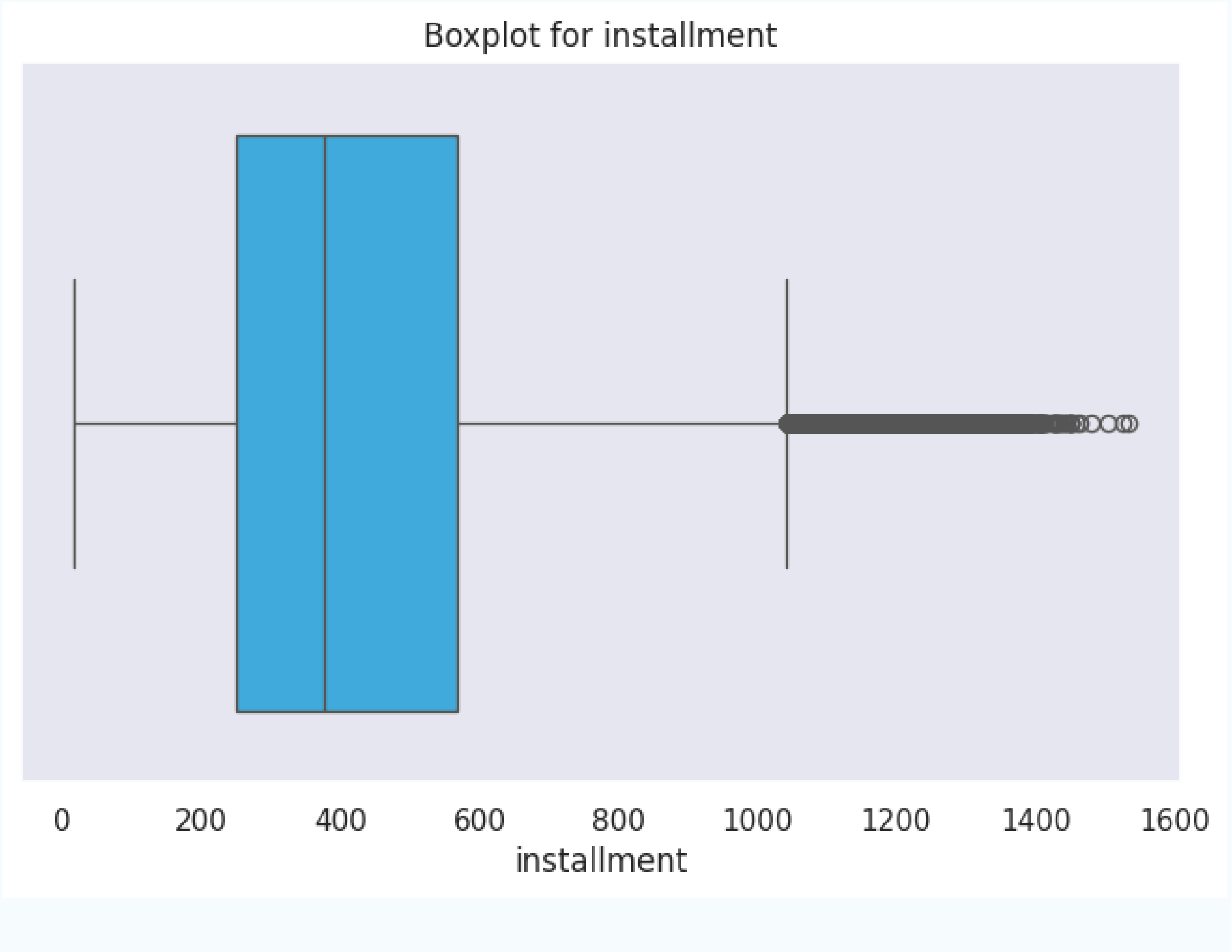
]Null values

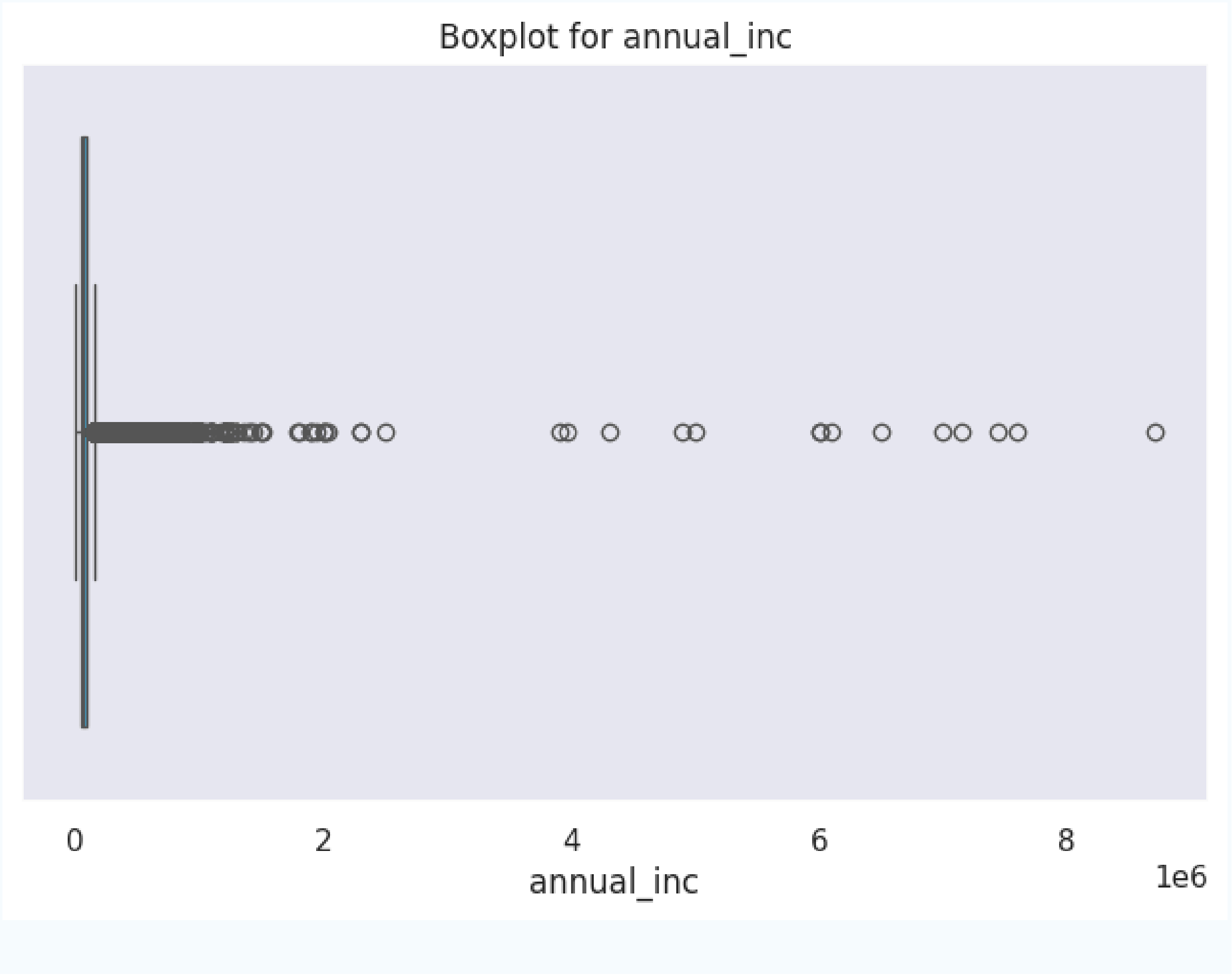
- For **numeric** columns, the null values were filled with **median**
- For **categorical** columns, the null values were filled with **mode**

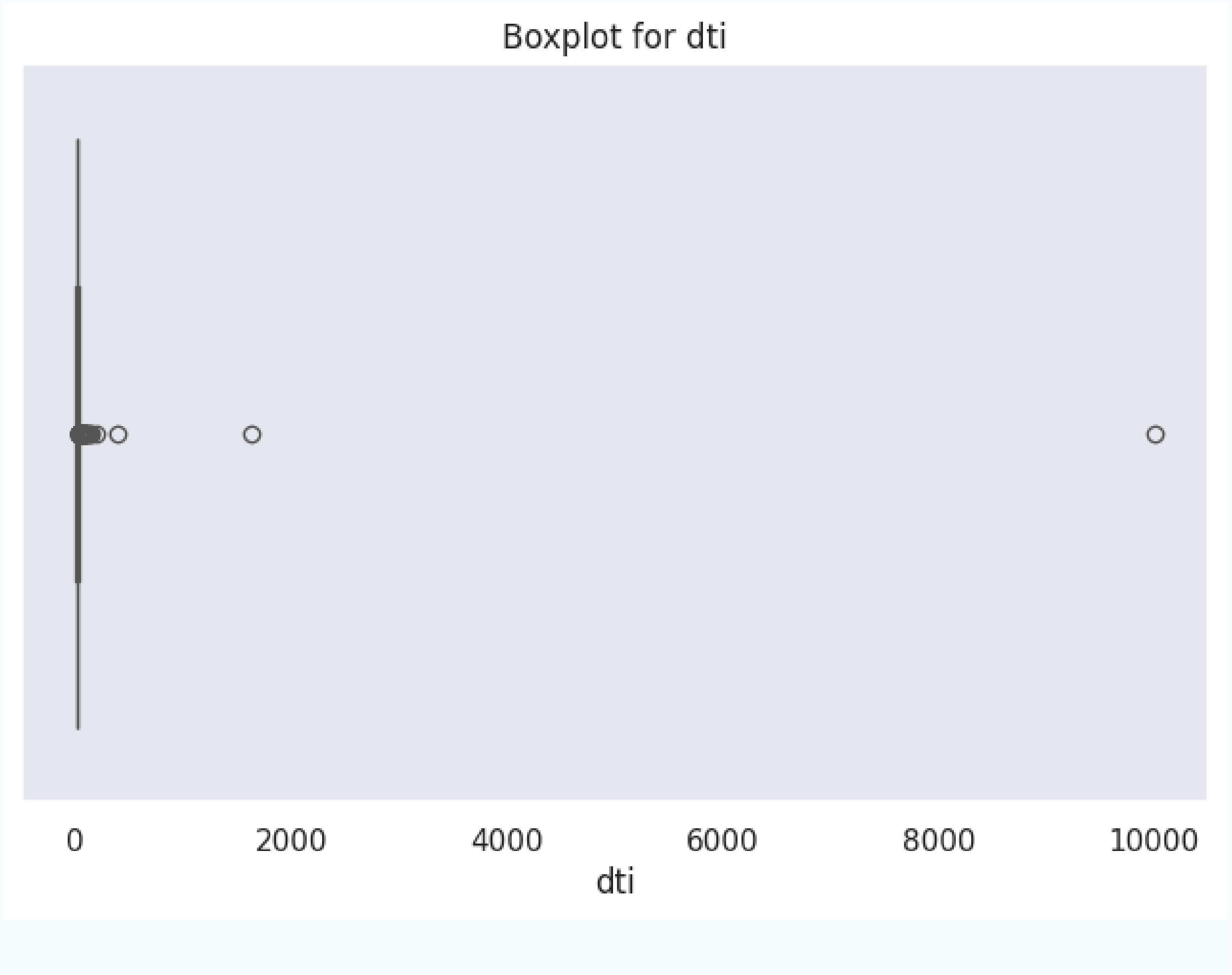
Outlier analysis



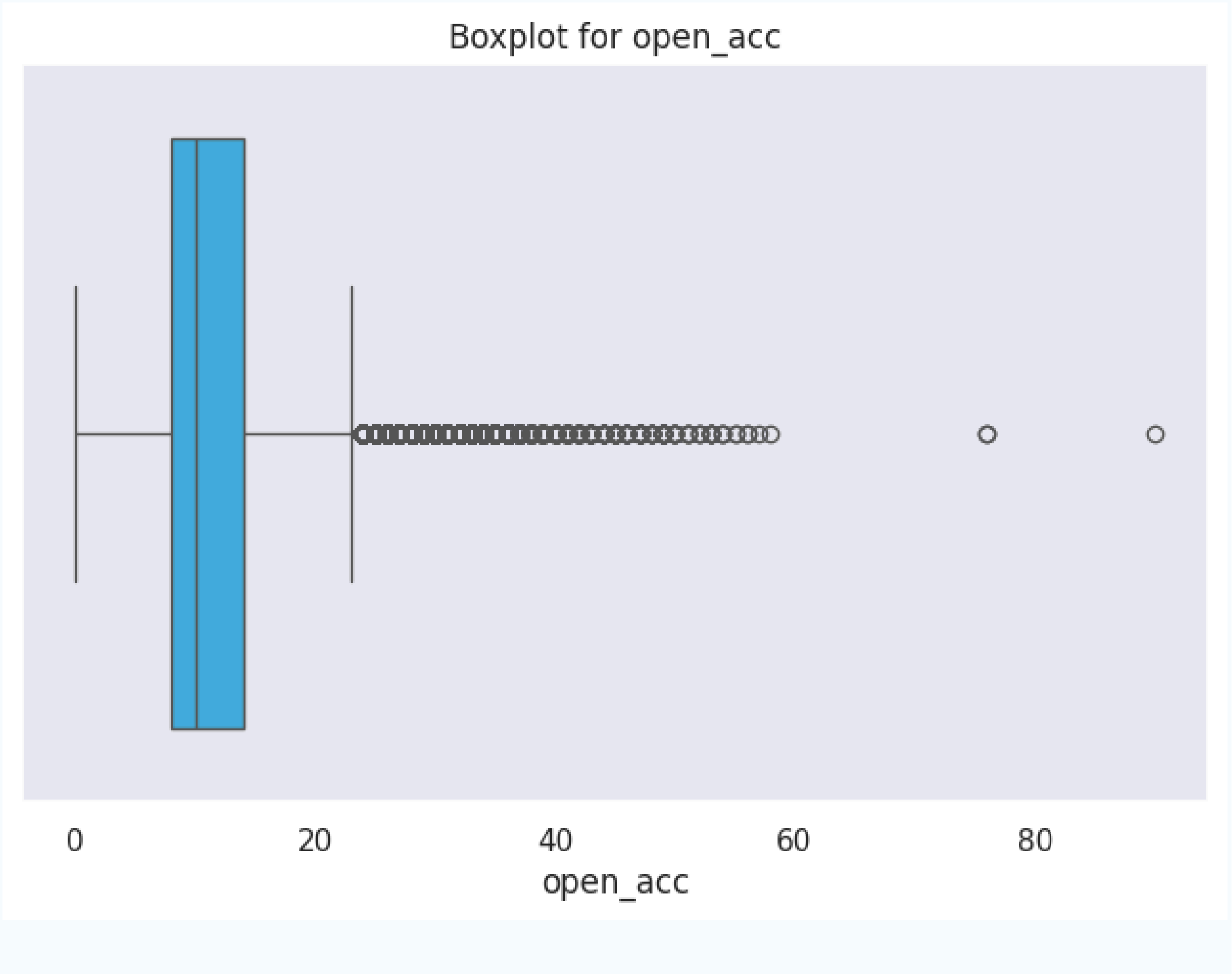

Boxplot for int_rate
5
10
15
20
25
30
int_rate

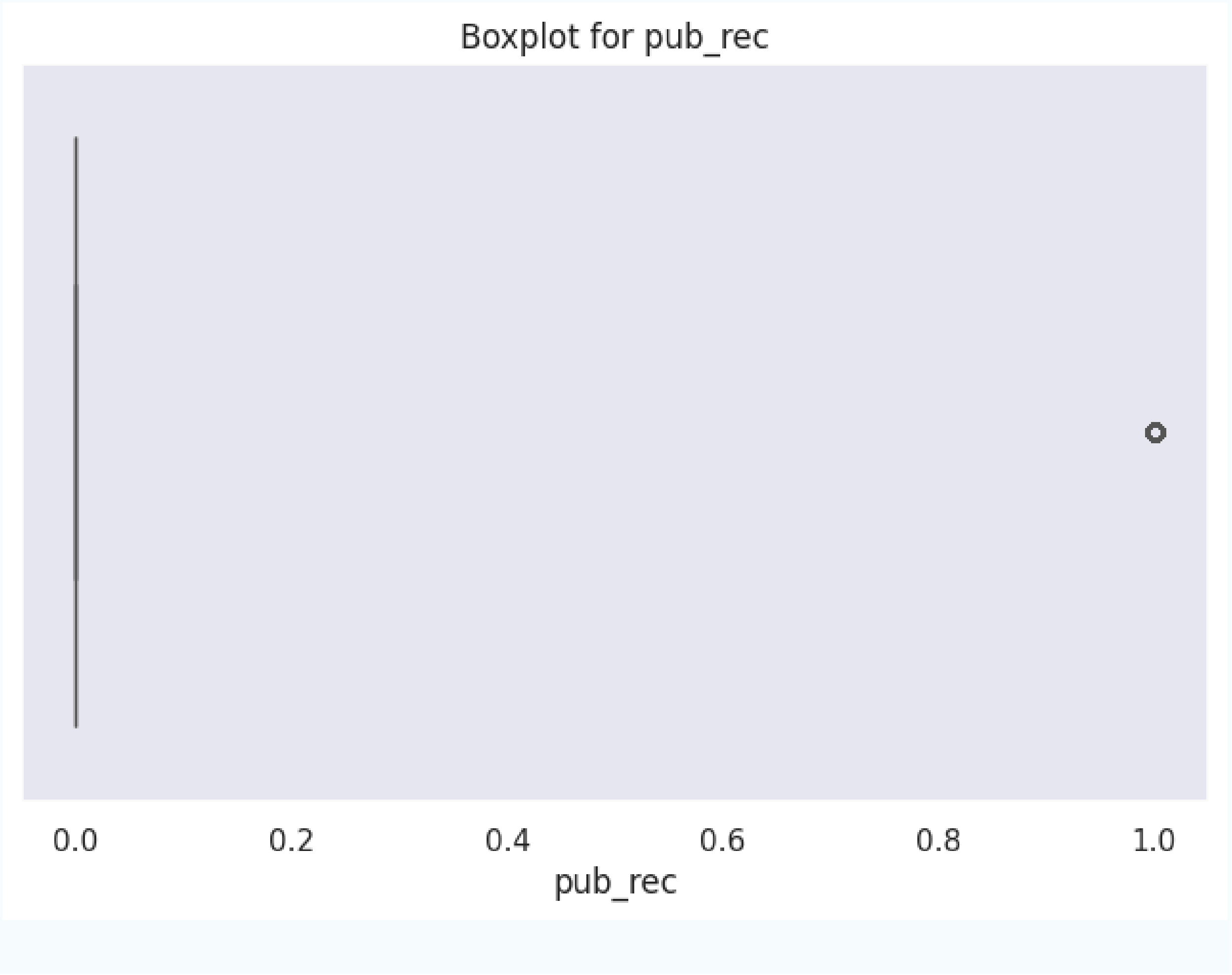

Boxplot for installment
0
200
400
600
800
1000
1200
1400
1600
installment

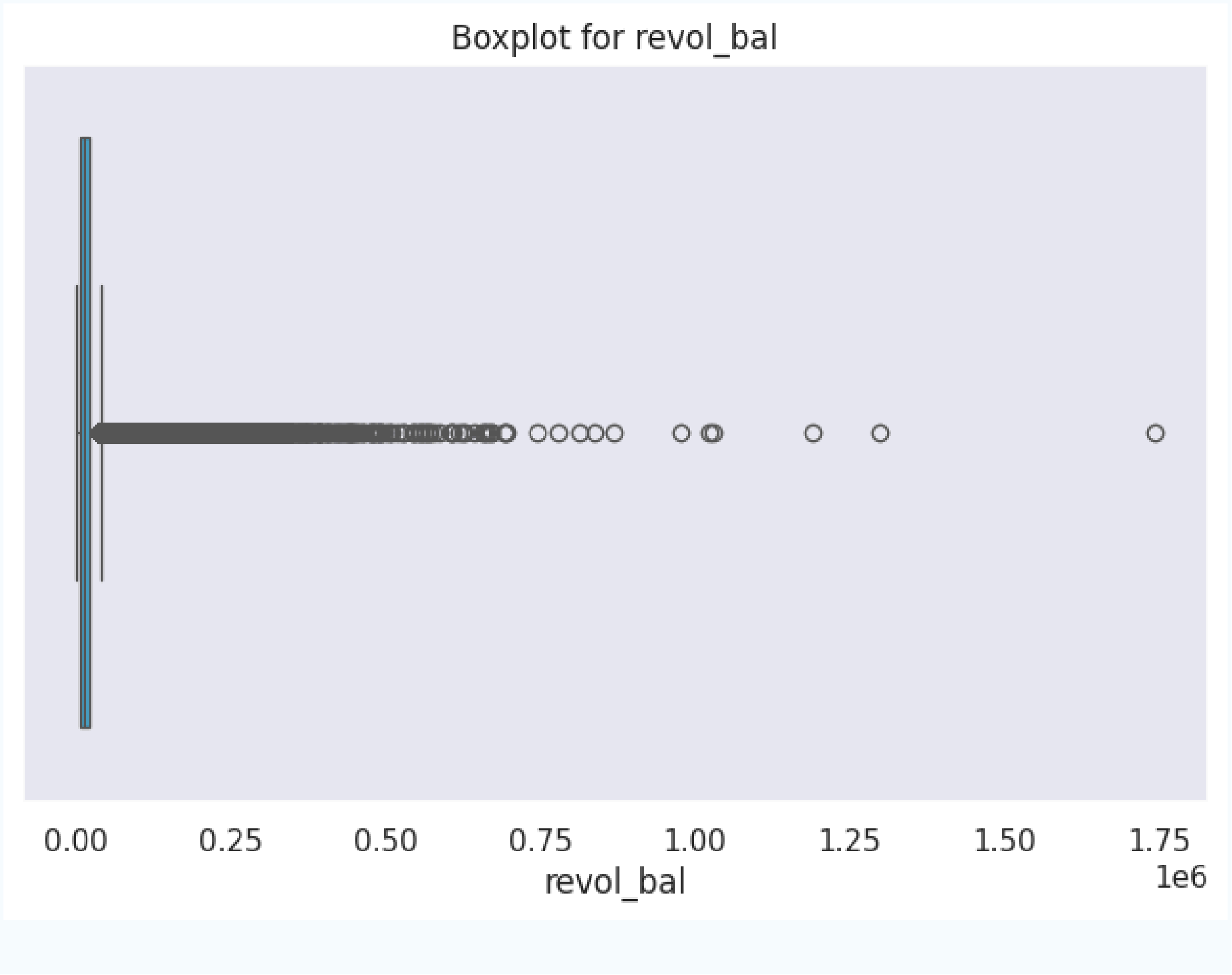

Boxplot for annual_inc
0
2
4
6
8
1e6
annual_inc

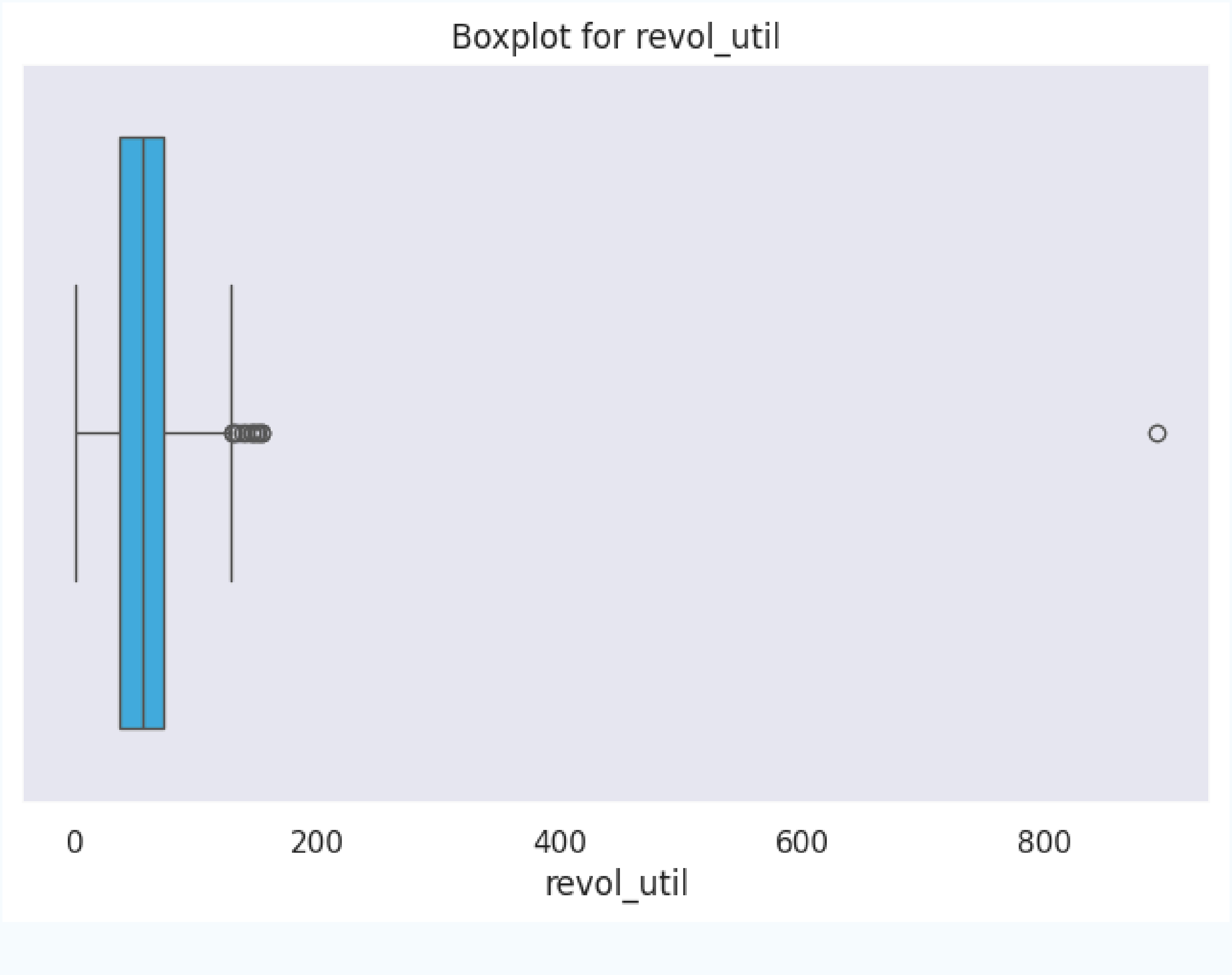


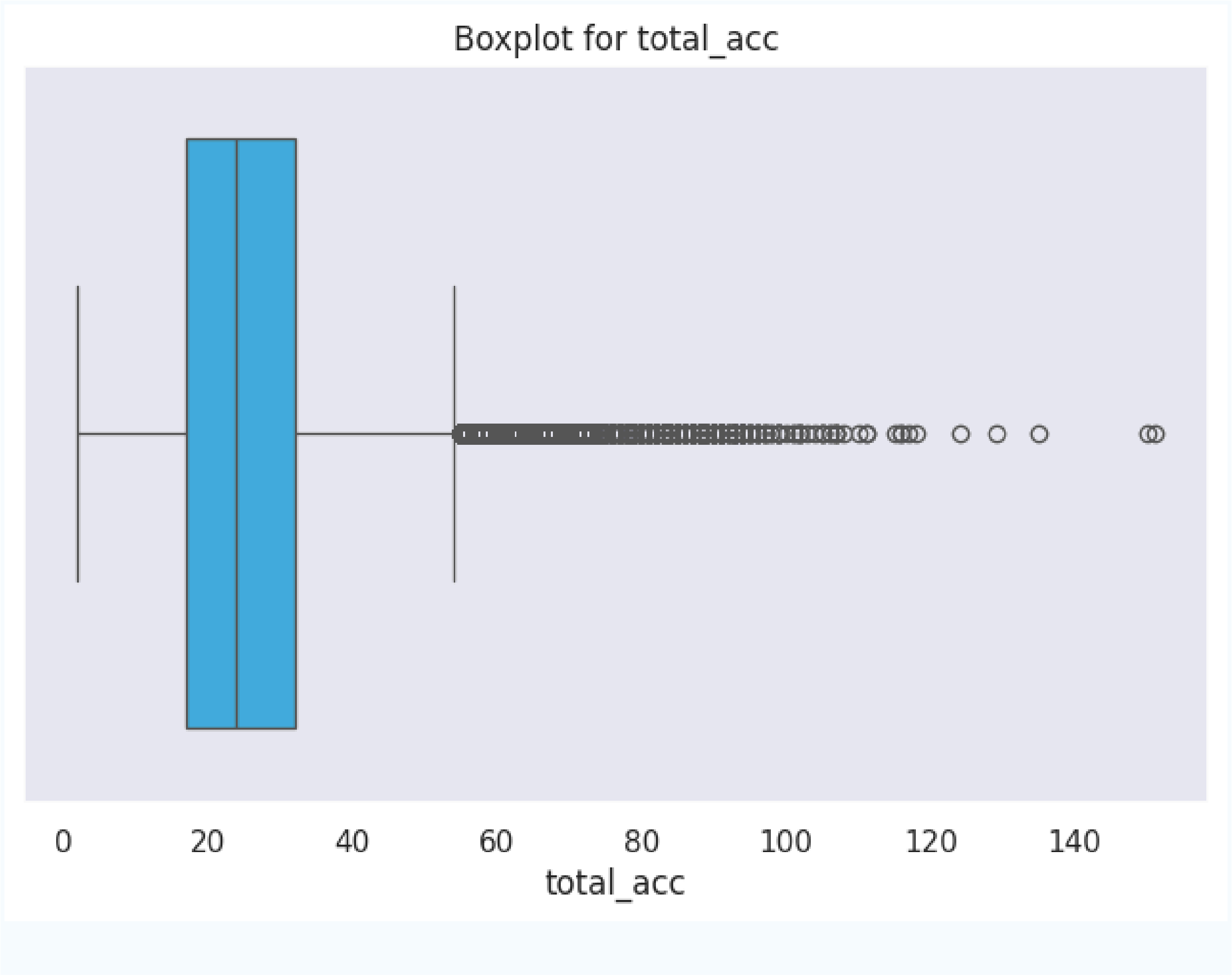
Boxplot for dti
0
2000
4000
6000
8000
10000
dti

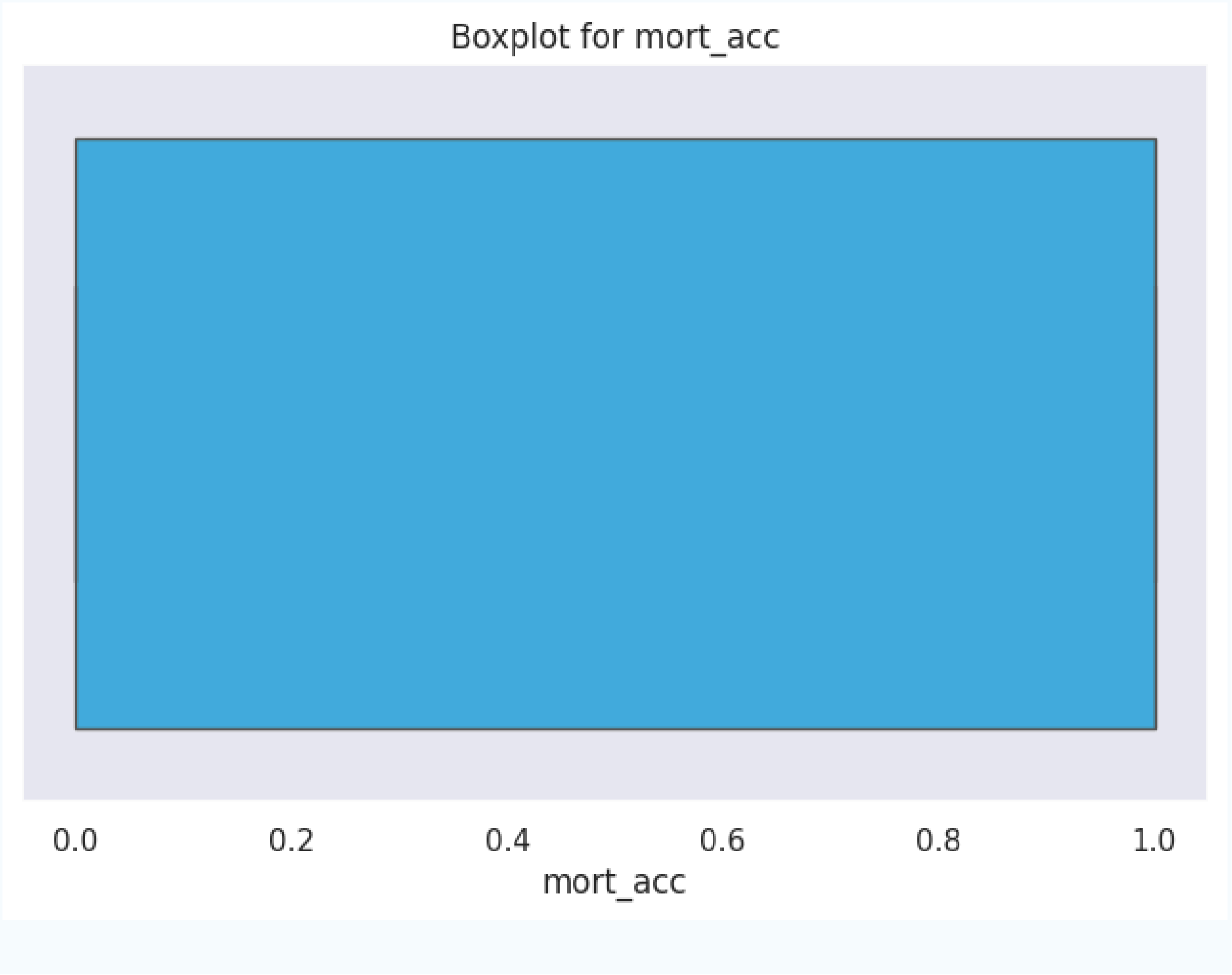

Boxplot for open_acc
0
20
40
60
80
open_acc

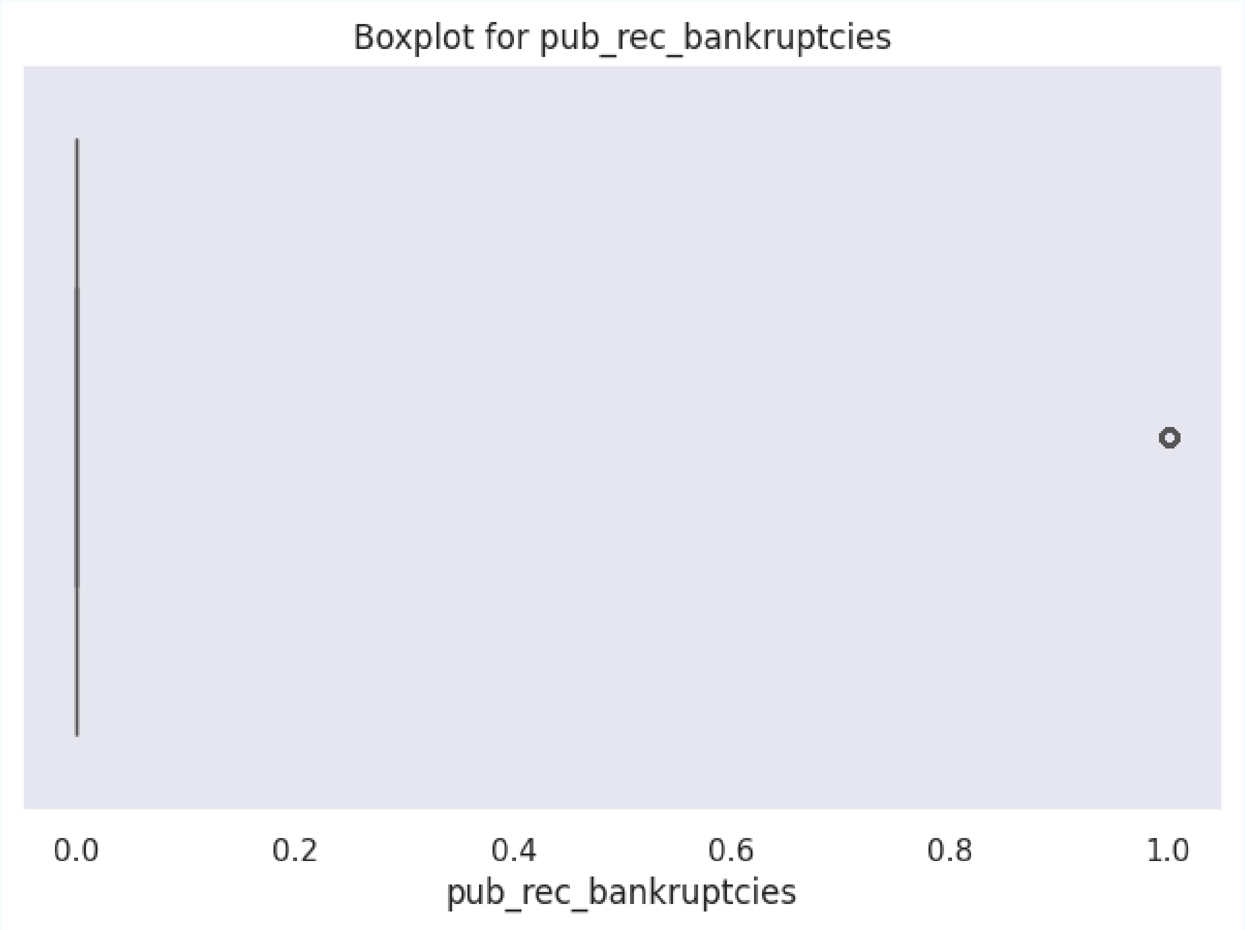

Boxplot for pub_rec
0.0
0.2
0.4
0.6
0.8
1.0
pub_rec


Boxplot for revol_bal
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e6
revol_bal


Boxplot for revol_util
0
200
400
600
800
revol_util


Boxplot for total_acc
0
20
40
60
80
100
120
140
total_acc


Boxplot for mort_acc
0.0
0.2
0.4
0.6
0.8
1.0
mort_acc



Outliers were treated

The columns term values and loan status were encoded

Zip code was extracted from the address and divided into 2 parts

1. Regional area : The first 3 digits
2. Delivery zone : The last 2 digits

This information was extracted through research

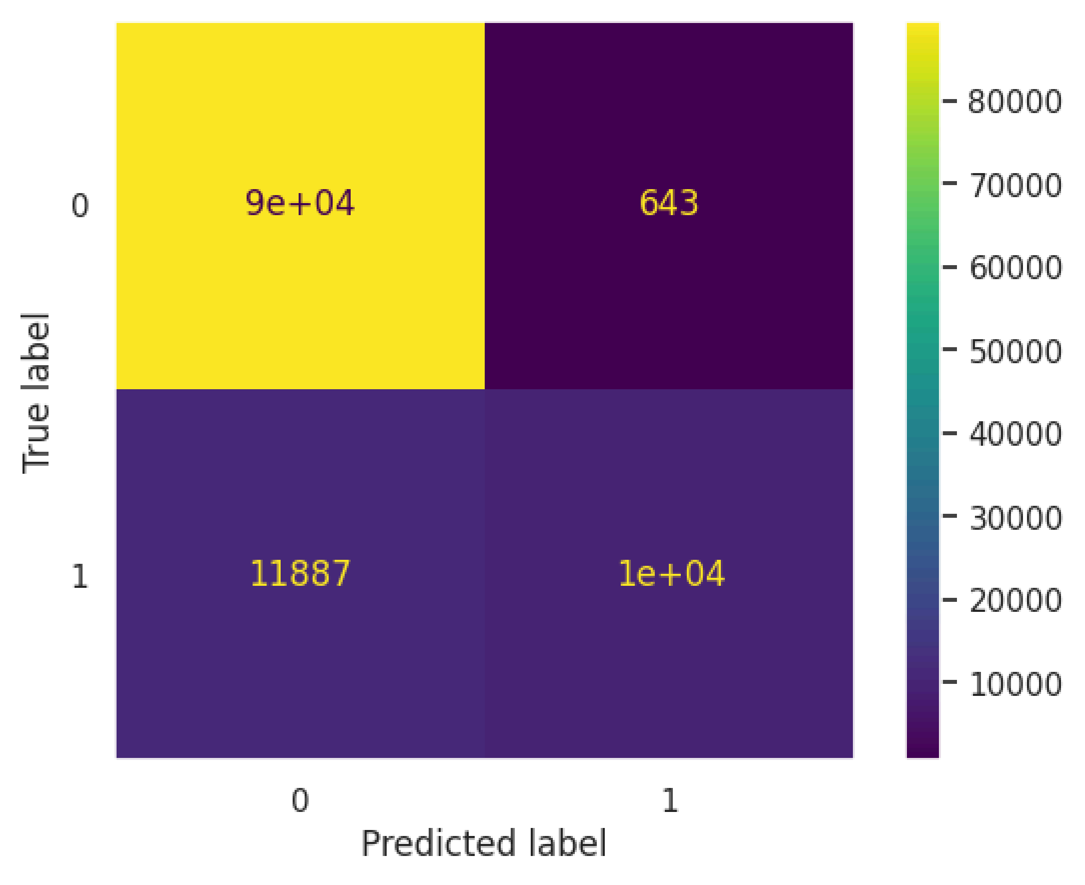
Columns issue_d, emp_title, title, sub_grade,address, earliest_cr_line, emp_length were dropped

One hot encoding was done on the categorical data

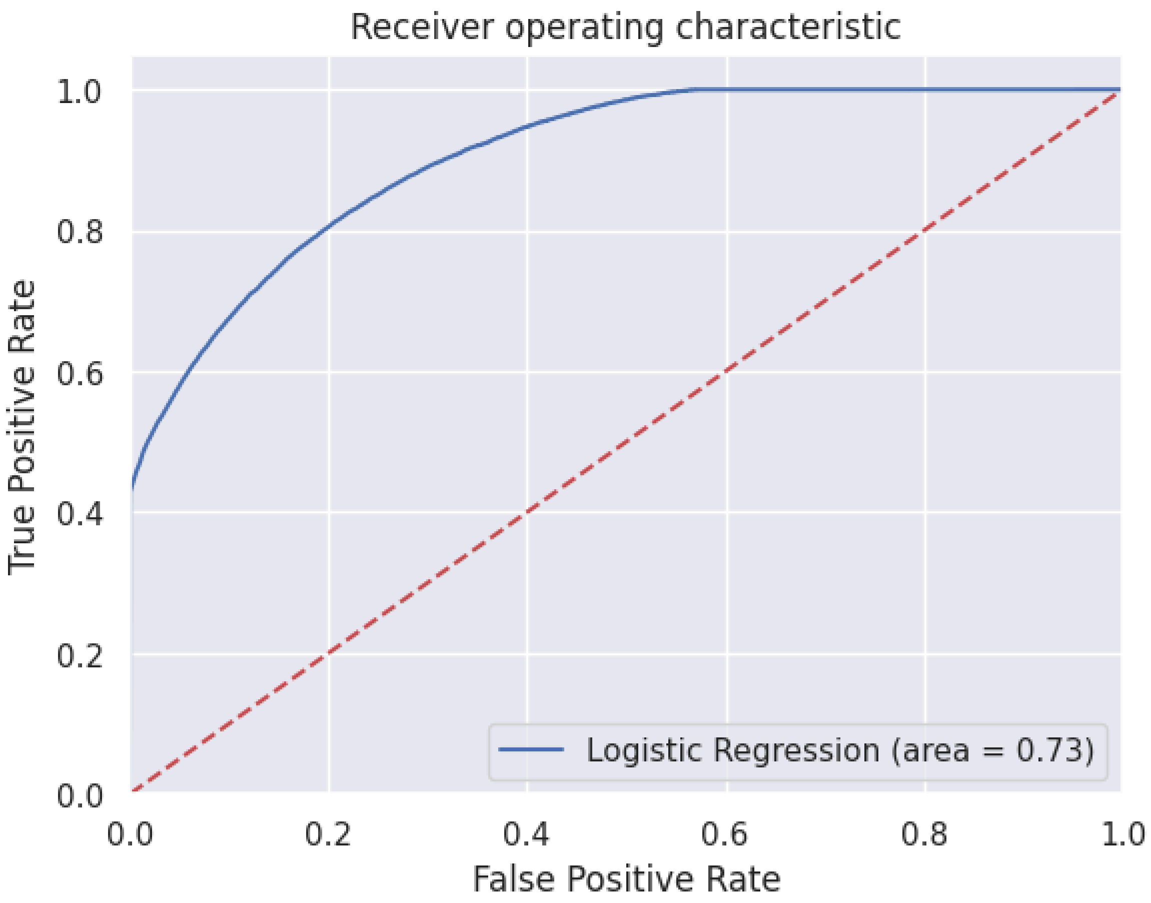
Model building

Logistic Regression model was trained which resulted in an accuracy of 88%

Confusion matrix

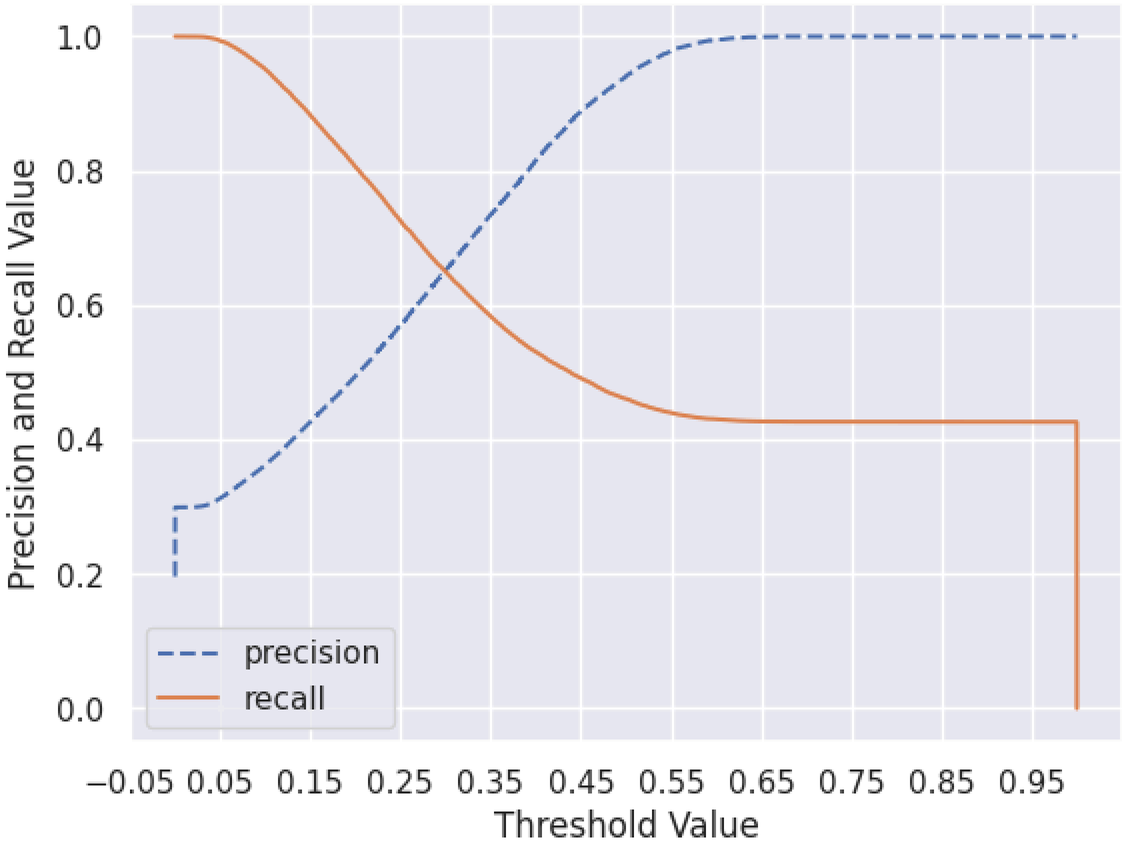


ROC Curve



- **The ROC-AUC curve reaches around 0.73, showing that the model is performing well.**
- **Better performance can be achieved by gathering more data, using a more advanced model, or fine-tuning hyperparameters.**

Precision recall curve



Insights

- Precision is highest at a 0.55 threshold—A high precision score means the model correctly identifies charged-off loans, helping businesses make more stable decisions.
- Recall is higher at lower thresholds but remains constant after 0.55, indicating the model consistently classifies actual predictions correctly.

Key Findings & Recommendations

- 80% of customers fully repay their loans, while 20% default.
- The model can help predict who is likely to repay or default, aiding decision-making.
- Based on the correlation heatmap, there's a very strong positive correlation between loan_amnt and installment (around 0.95). This makes sense because installment amounts are directly calculated from the loan amount, interest rate, and term.
- The majority of people have home ownership as mortgage
- People with grades ‘A’ are more likely to fully pay their loan, as seen in the countplot
- The top 2 afforded job titles are Teacher and Manager with almost equal count of 4389 and 4250 respectively.

- Precision would be the most important metric for a bank. The precision of 0.94 for the Charged off class means that when the model predicts a loan will default, its correct 94% of the time. This helps the bank avoid risky loans and minimize the loss
- The model has high precision (0.94) but low recall (0.46) for predicting defaults (class 1). This means the model is very accurate when it predicts a default, but it misses many actual defaults. For a bank, this translates to approving some loans that will eventually default, leading to potential financial losses, despite being confident about the loans it does reject.
- The features that heavily affected the outcome are
-
- Grade/sub-grade (strong relationship with default rates)
- Term (60-month loans default more often than 36-month loans)
- Interest rate (higher rates correlate with higher default probability)
- Home ownership status
- Verification status
- The results will be affected by geographical location, on a test wherein a model was trained without the information about regional area and delivery zone (zip), the accuracy dropped to 80% from 88%
- F1-score:
  - Negative class (Fully Paid): 94%
  - Positive class (Charged Off): 62%
- Cross-validation and test accuracy are similar, showing the model generalizes well to unseen data.
- ROC-AUC score is 0.73, meaning the model correctly classifies about 73% of cases—a good result, but improvement is possible.
- The precision-recall curve shows the trade-off—higher thresholds improve precision but reduce recall, and vice versa. The best threshold depends on business needs.
- Balancing the dataset significantly affects precision and recall for both classes.
- Logistic Regression test accuracy: 0.884, indicating reliable performance.