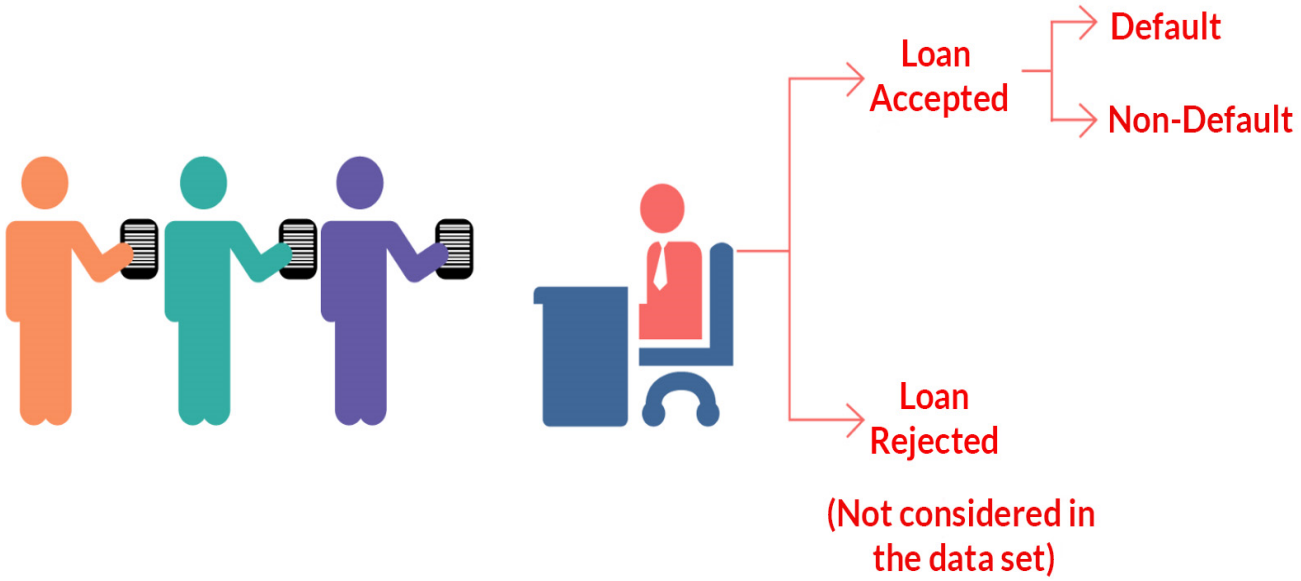
EDA Case Study

DONE BY:
IFNA ANSARI

Problem Statement



Loan Default Analysis: Understanding the factors contributing to loan defaulting.

Identifying Patterns: Discovering patterns related to applicant characteristics, loan attributes, and default rates.

The consumer finance company aims to mitigate the risks associated with lending by leveraging Exploratory Data Analysis (EDA) to identify patterns indicative of loan default, enabling informed decision-making to minimize both the loss of potential business from rejecting creditworthy applicants and the financial loss resulting from approving risky applicants.

Understanding the Problem

- The business problem revolves around the challenges faced by a consumer finance company in making informed decisions about loan applications. The two primary concerns involve

1. the risk of losing potential business if a good applicant is rejected and
2. the financial losses incurred if a risky applicant is approved and subsequently defaults on the loan.

- The objective is to use data analysis to identify patterns that indicate the likelihood of loan default, allowing the company to optimize its decision-making process and minimize credit loss.

When a person applies for a loan, there are two types of decisions that could be taken by the company:

Loan accepted: If the company approves the loan, there are three possible scenarios, as described below:

Fully paid: The applicant has fully paid the loan (the principal and the interest amount).

Current: The applicant is in the process of paying the instalments, i.e., the tenure of the loan is not yet completed. These candidates are not labelled as ‘defaulted’.

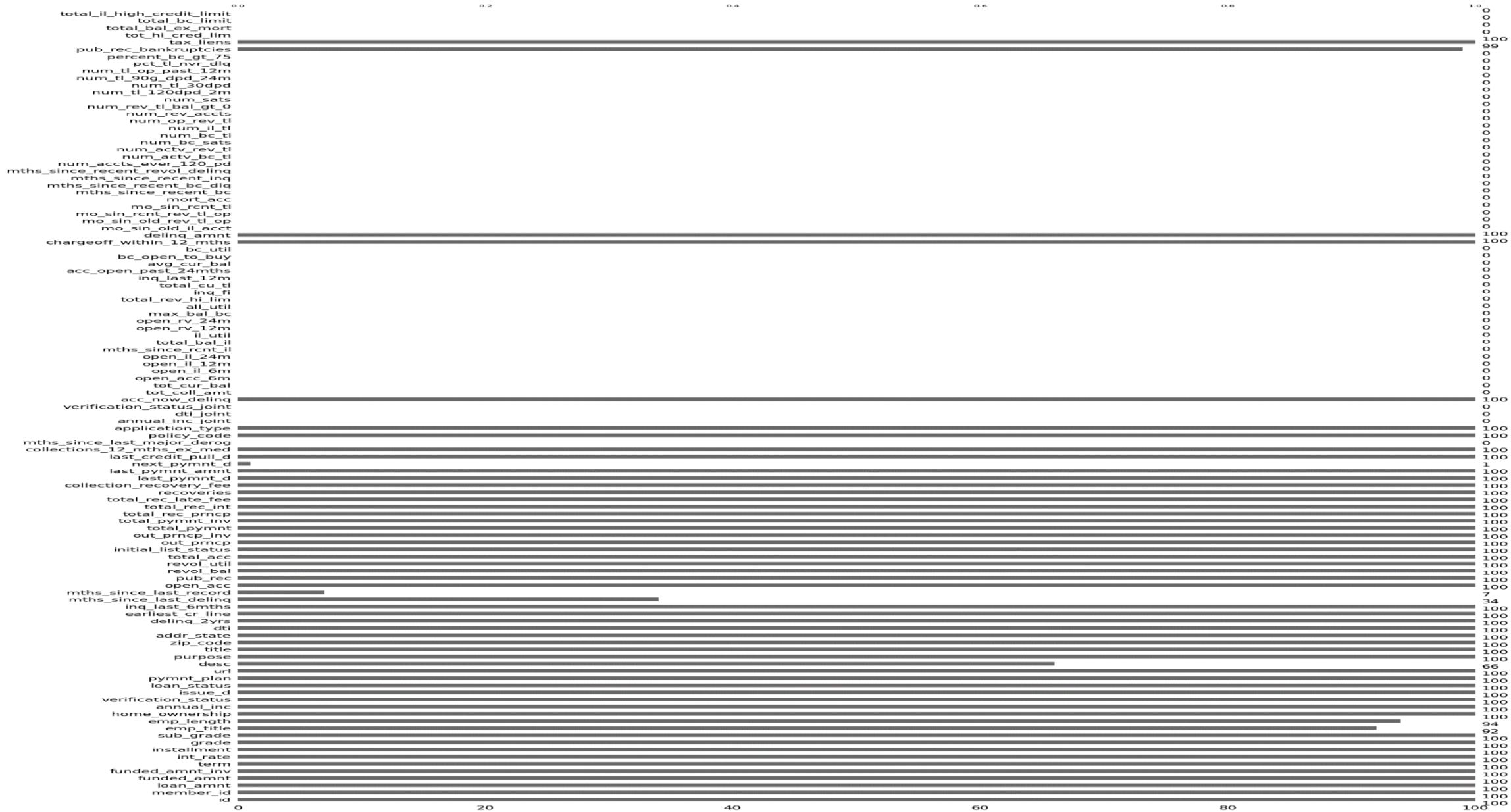
Charged-off: The applicant has not paid the instalments in due time for a long period, i.e., they have defaulted on the loan.

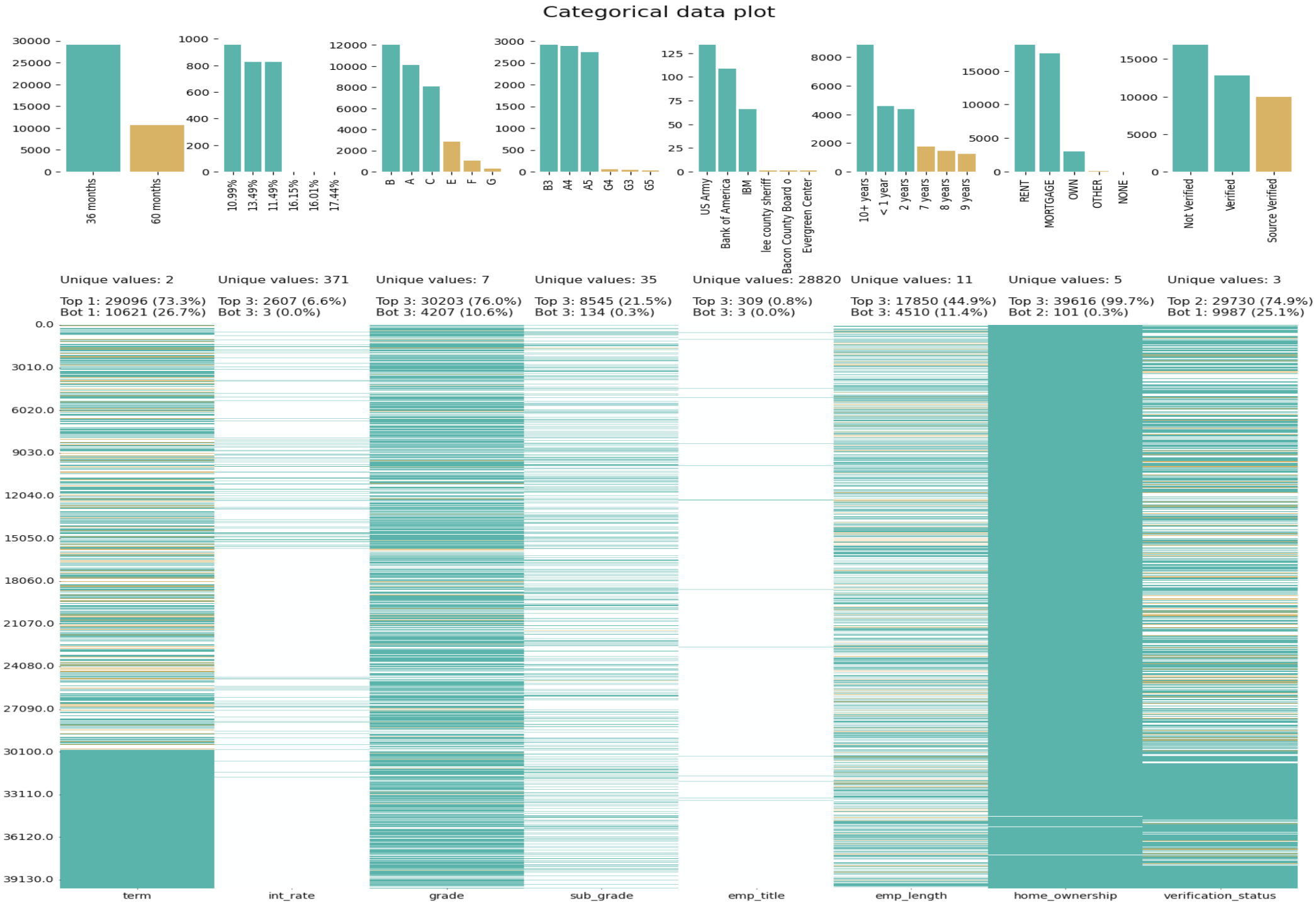
Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements, etc.). Since the loan was rejected, there is no transactional history of those applicants with the company; so, this data is not available with the company (and thus, in this data set).

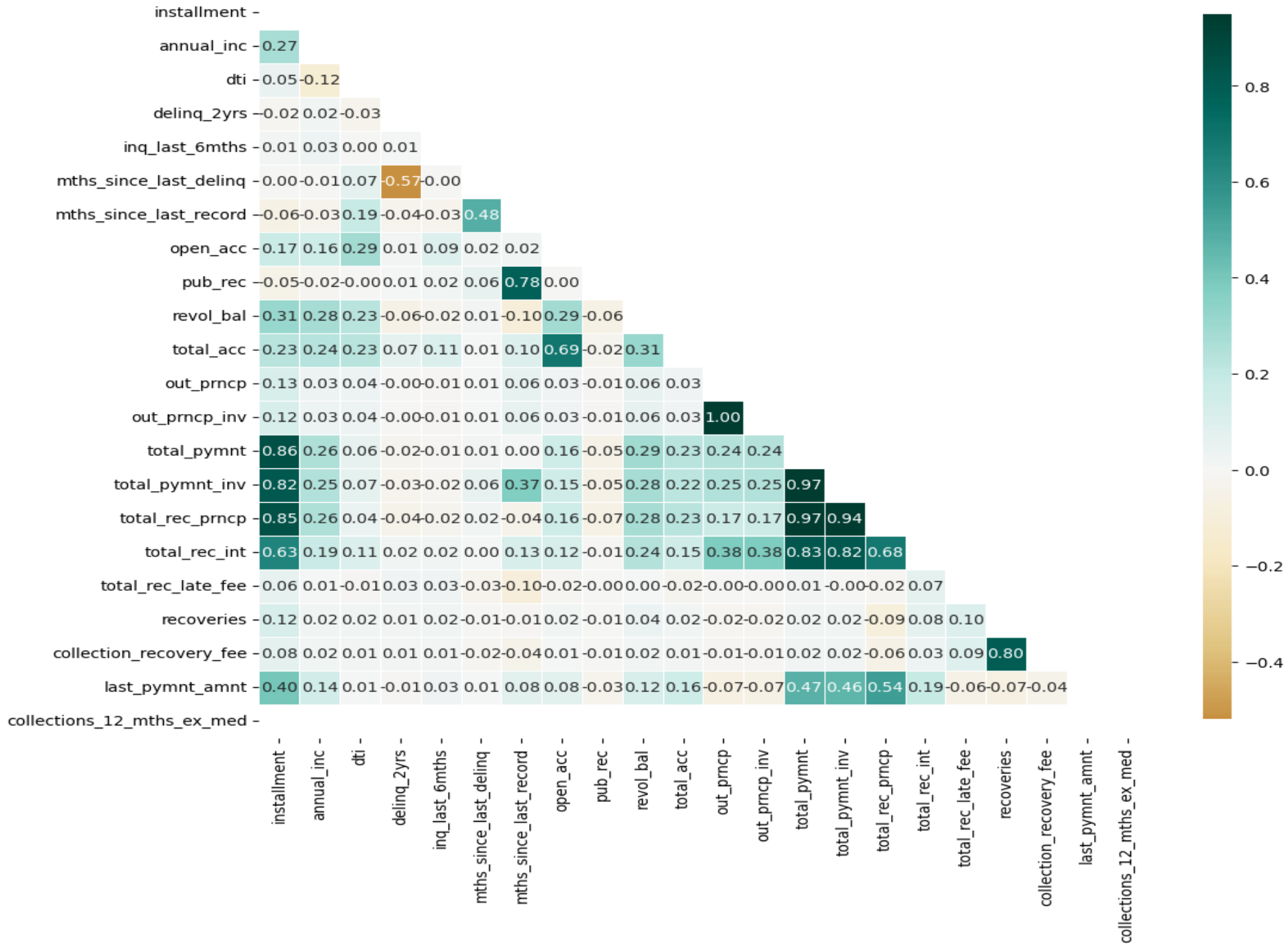
Tasks

1. Understanding the Data
2. Data Cleaning and Manipulation
3. Data Analysis
4. Insights and Recommendations

Understanding the Data

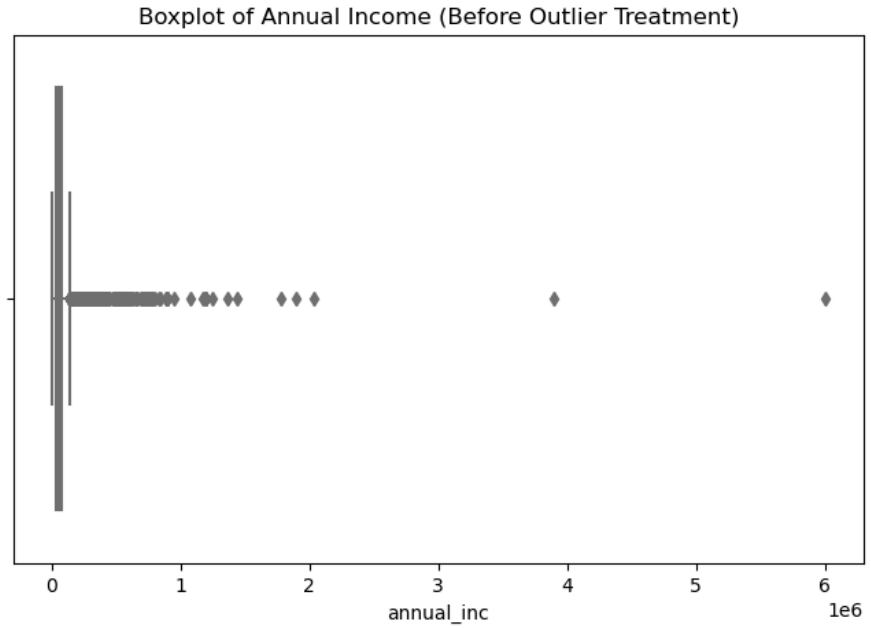


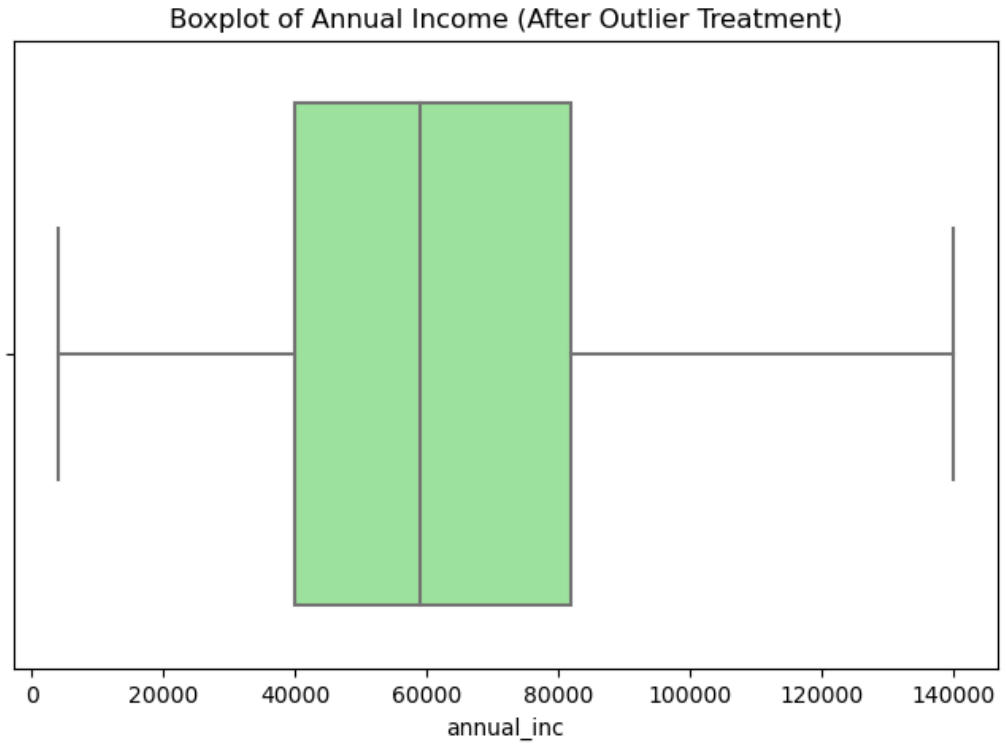

Categorical data plot
36 months
60 months
10.99%
13.49%
11.49%
16.15%
16.01%
17.44%
B
A
C
E
F
G
B3
A4
A5
G4
G3
G5
US Army
Bank of America
IBM
lee county sheriff
Bacon County Board o
Evergreen Center
10+ years
< 1 year
2 years
7 years
8 years
9 years
RENT
MORTGAGE
OWN
OTHER
NONE
Not Verified
Verified
Source Verified
Unique values: 2
Top 1: 29096 (73.3%)
Bot 1: 10621 (26.7%)
Unique values: 371
Top 3: 2607 (6.6%)
Bot 3: 3 (0.0%)
Unique values: 7
Top 3: 30203 (76.0%)
Bot 3: 4207 (10.6%)
Unique values: 35
Top 3: 8545 (21.5%)
Bot 3: 134 (0.3%)
Unique values: 28820
Top 3: 309 (0.8%)
Bot 3: 3 (0.0%)
Unique values: 11
Top 3: 17850 (44.9%)
Bot 3: 4510 (11.4%)
Unique values: 5
Top 3: 39616 (99.7%)
Bot 2: 101 (0.3%)
Unique values: 3
Top 2: 29730 (74.9%)
Bot 1: 9987 (25.1%)
0.0
3010.0
6020.0
9030.0
12040.0
15050.0
18060.0
21070.0
24080.0
27090.0
30100.0
33110.0
36120.0
39130.0
term
int_rate
grade
sub_grade
emp_title
emp_length
home_ownership
verification_status


Feature-correlation (pearson)
installment
annual_inc
dti
delinq_2yrs
inq_last_6mths
mths_since_last_delinq
mths_since_last_record
open_acc
pub_rec
revol_bal
total_acc
out_prncp
out_prncp_inv
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_amnt
collections_12_mths_ex_med
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Data Cleaning

- Handle the missing values
- Outlier detection and treatment

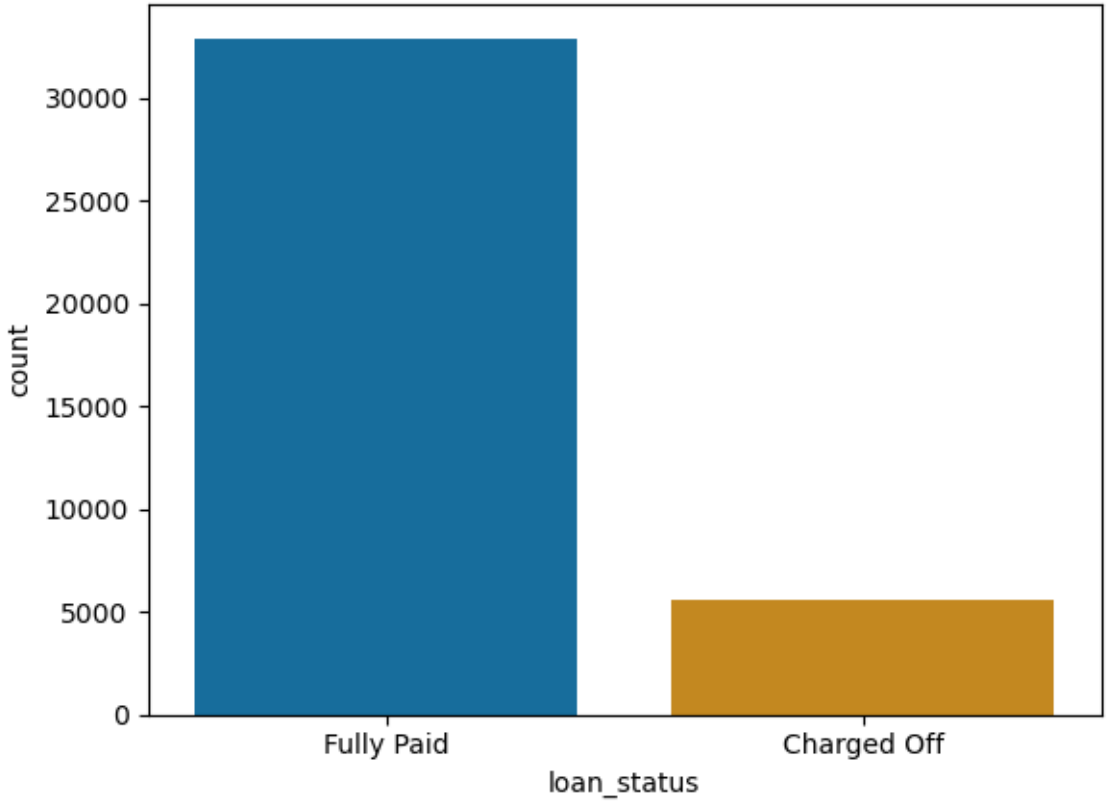


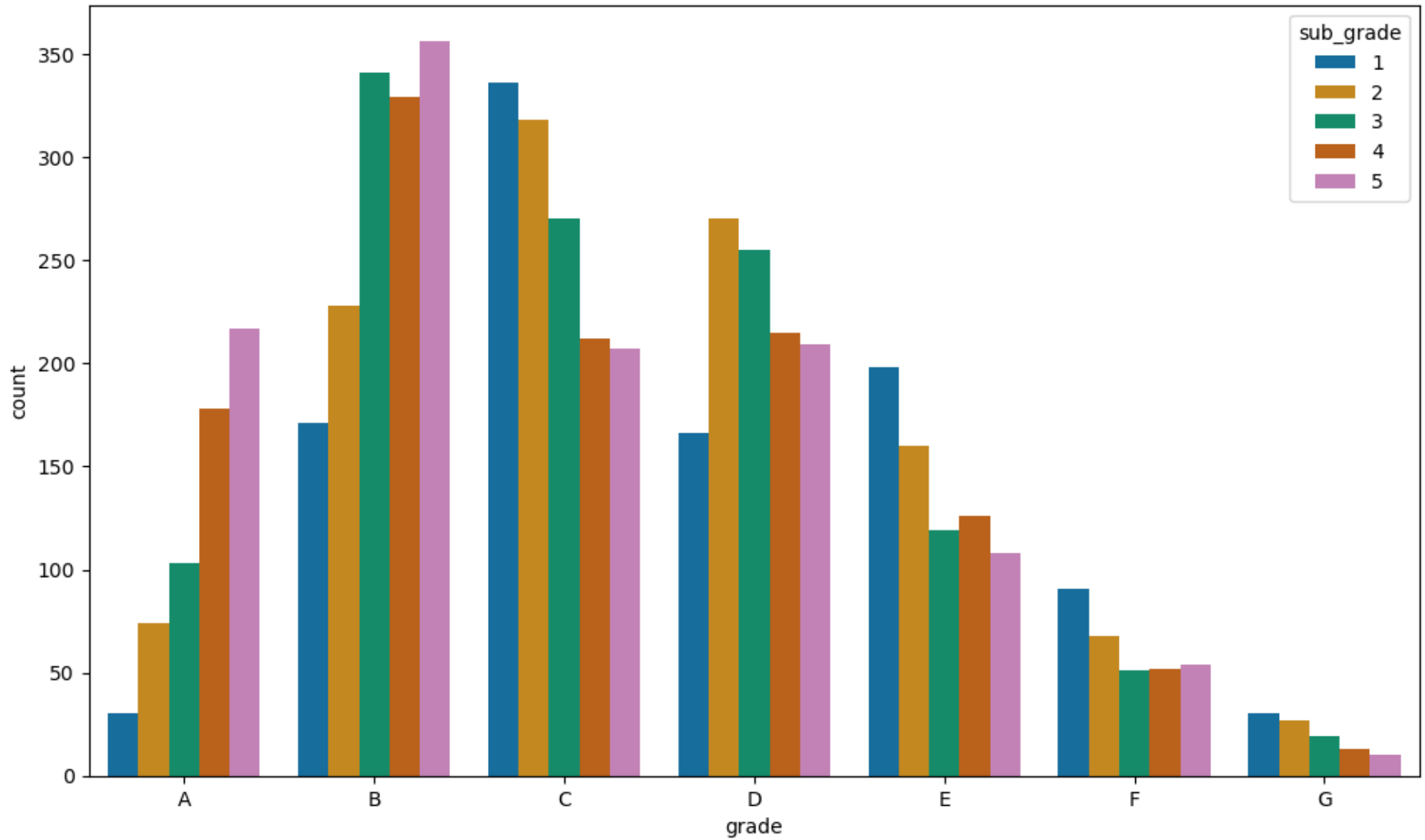


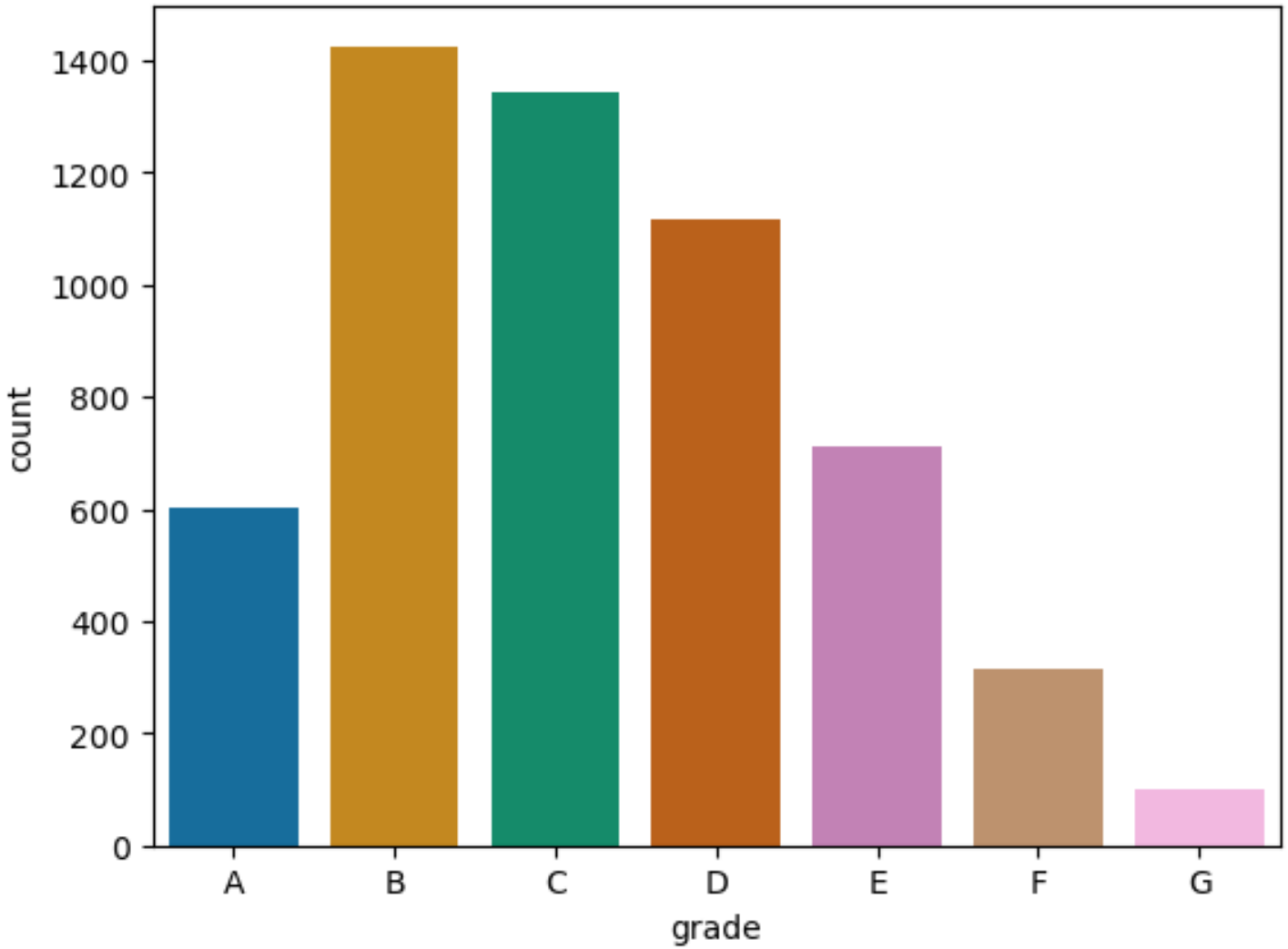
DATA ANALYSIS

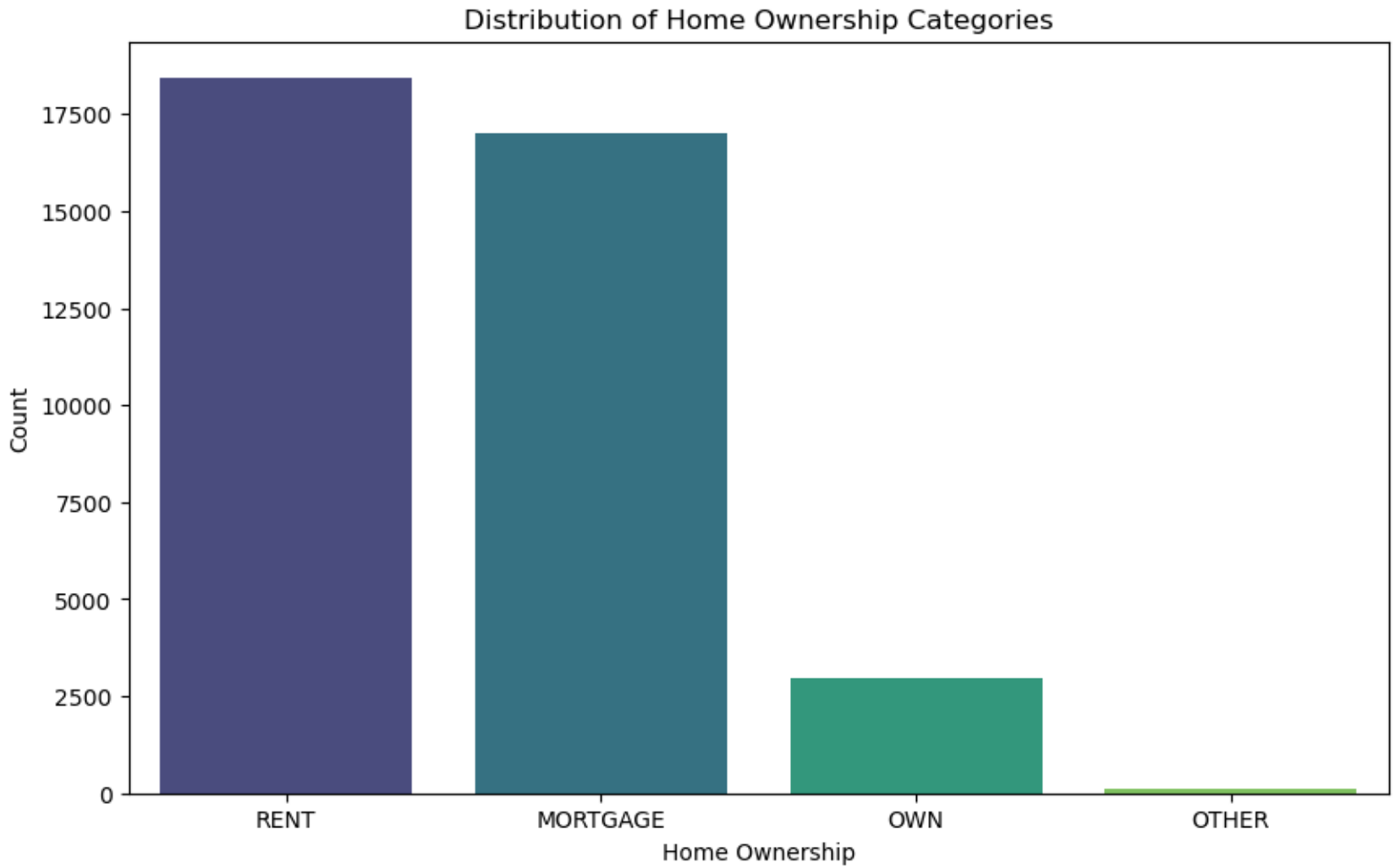
- ***UNIVARIATE ANALYSIS***
- ***BIVARIATE ANALYSIS***
- ***MULTIVARIATE ANALYSIS***

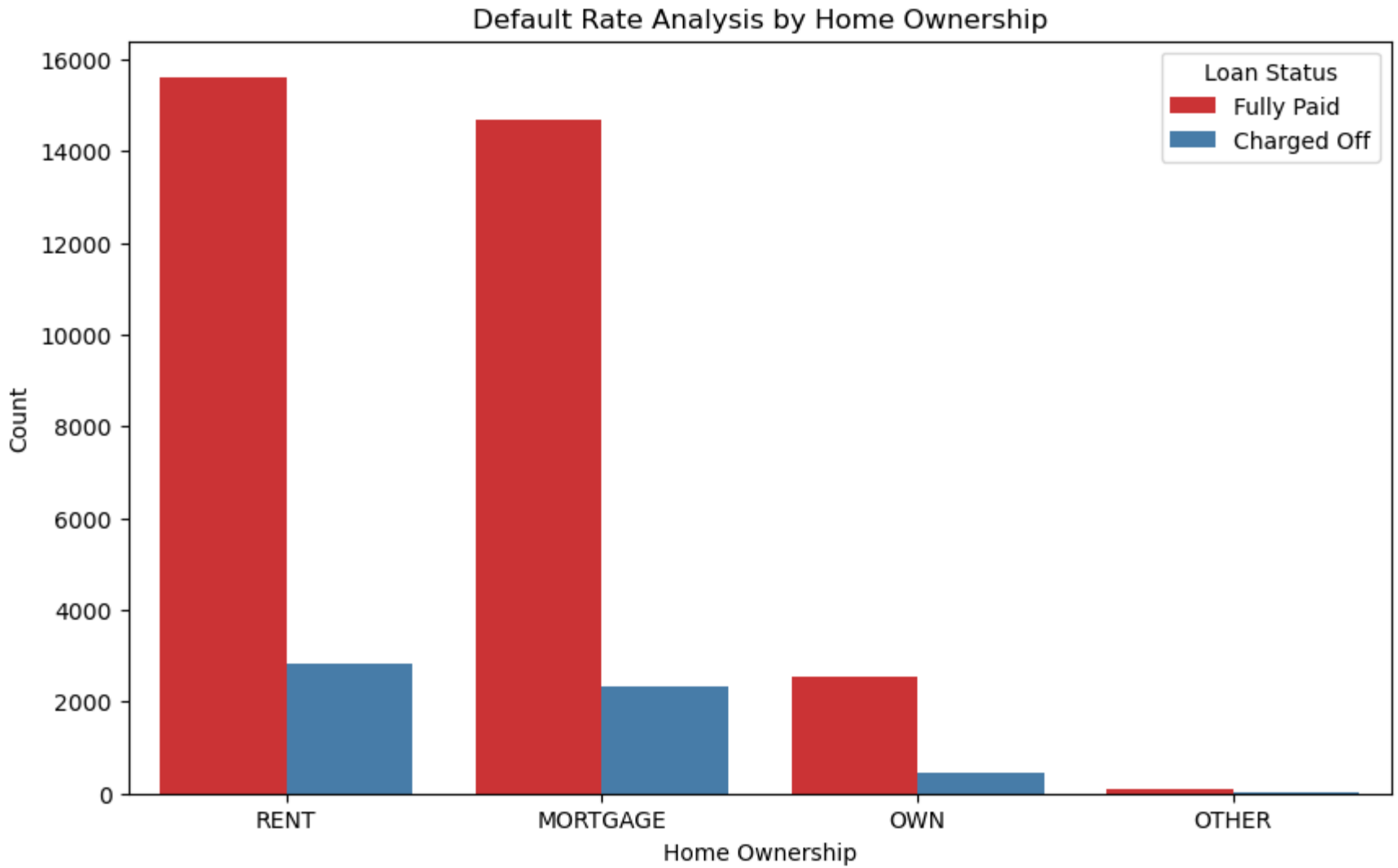
Visualizing Categorical Data

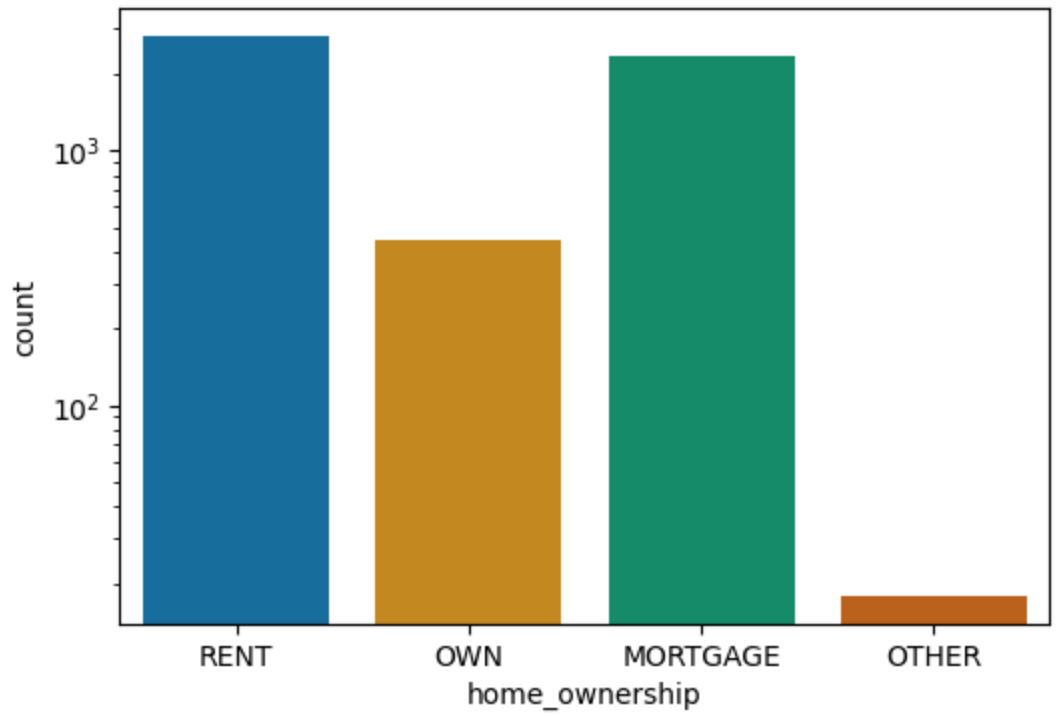


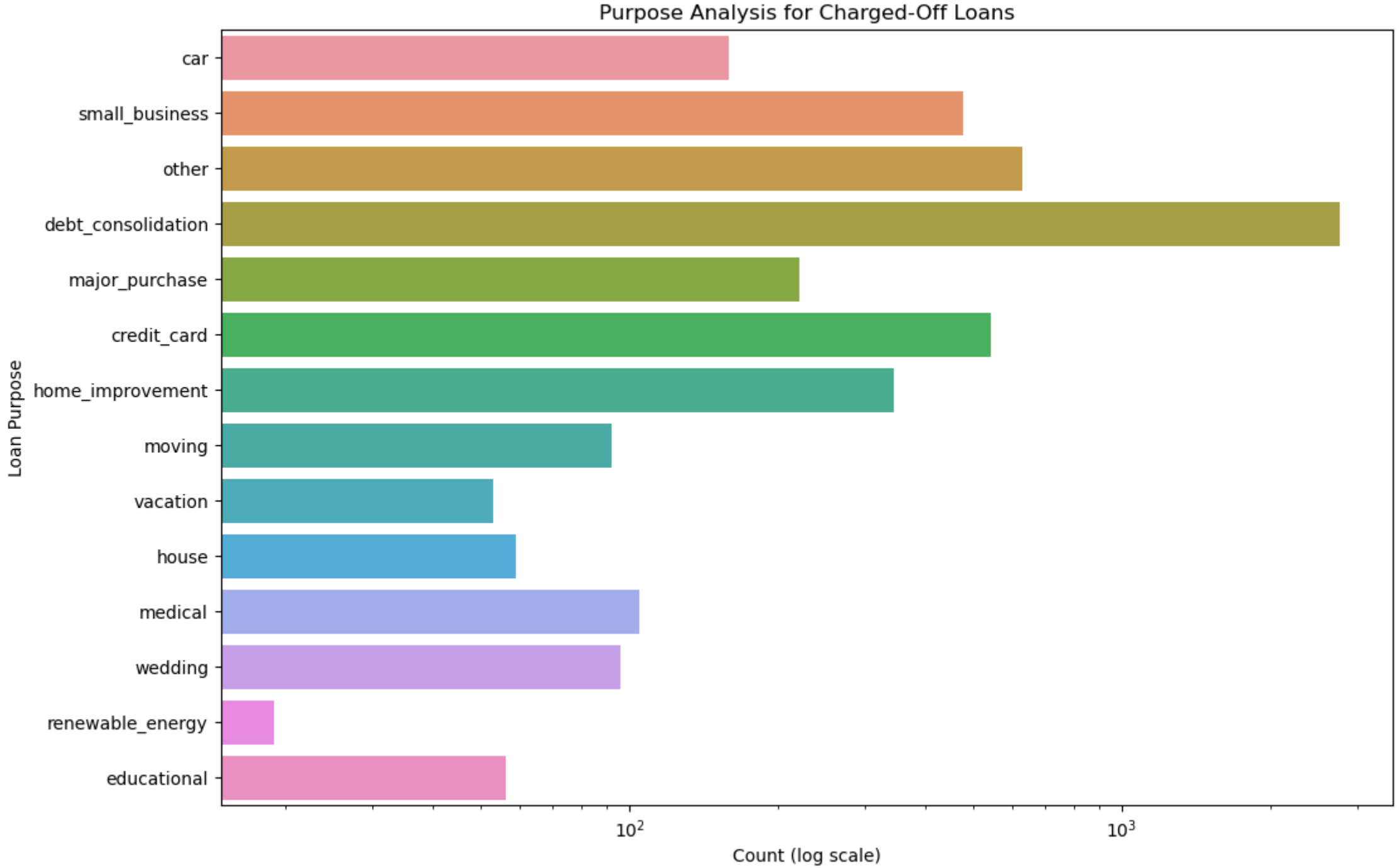

sub_grade
1
2
3
4
5
count
0
50
100
150
200
250
300
350
A
B
C
D
E
F
G
grade

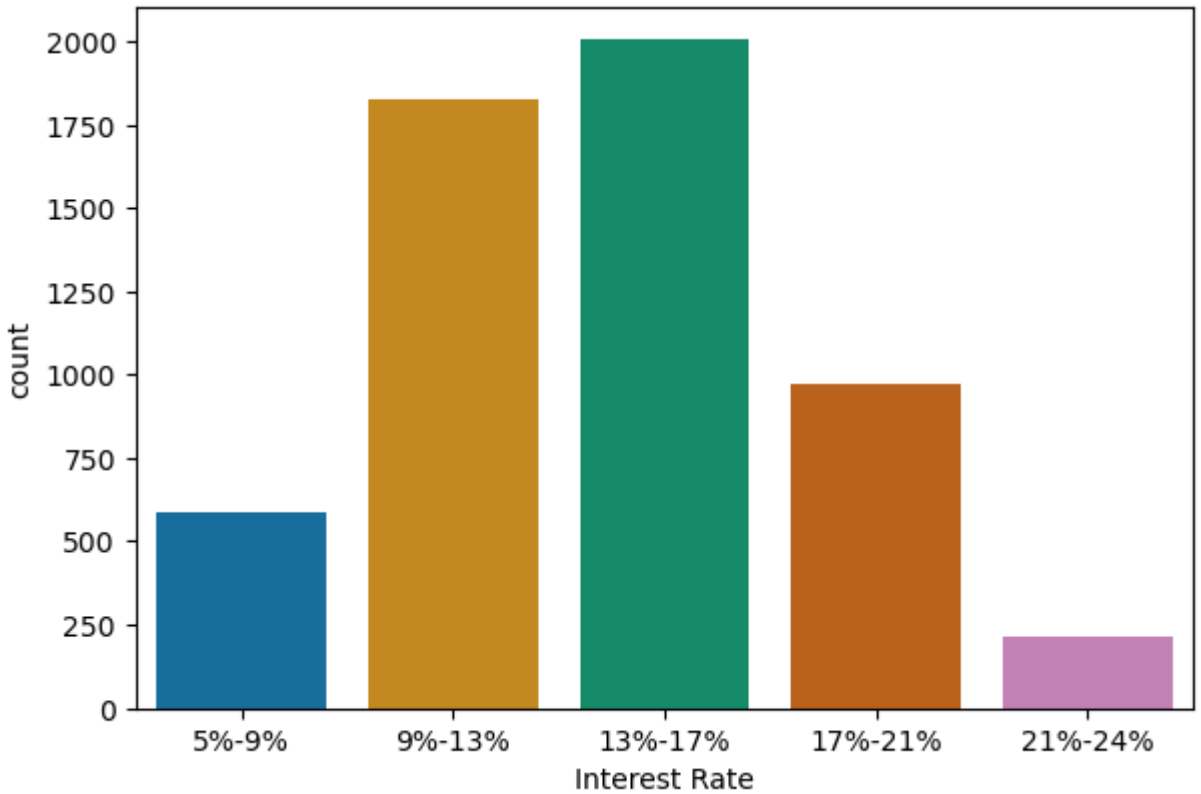

1400
1200
1000
800
600
400
200
0
count
A
B
C
D
E
F
G
grade

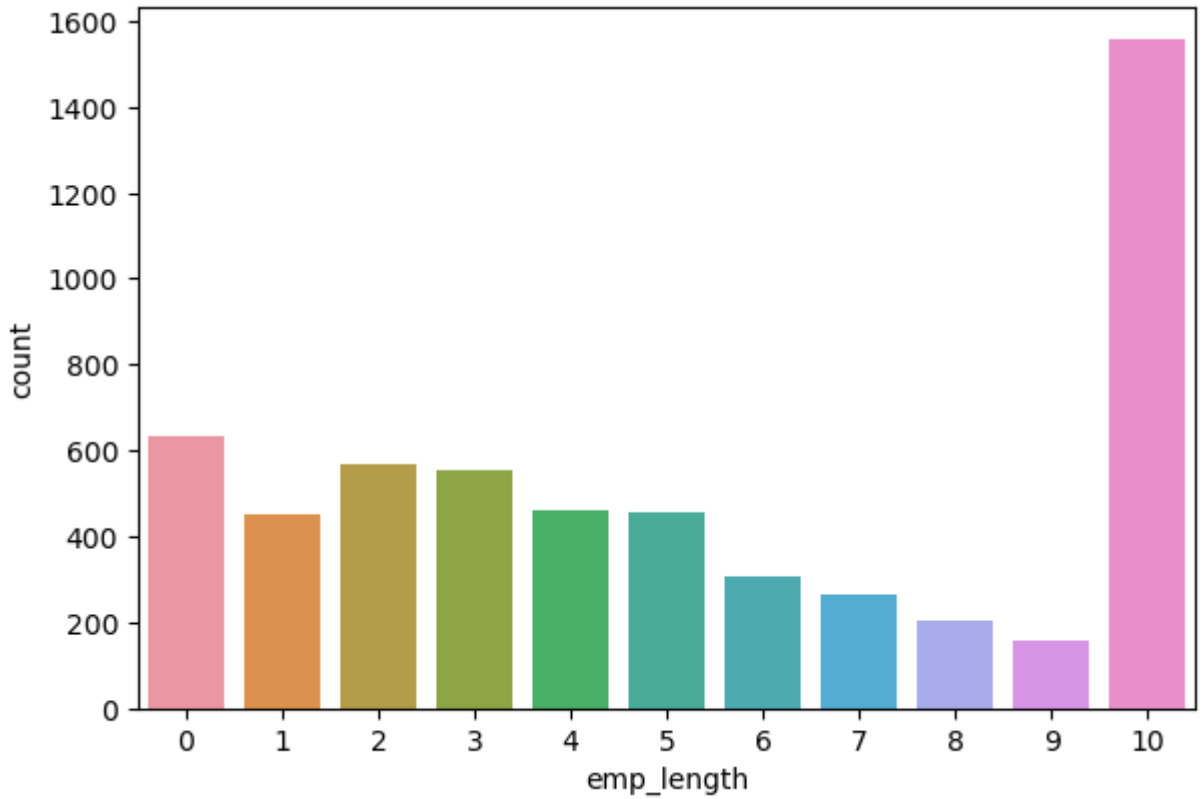

Distribution of Home Ownership Categories
Count
17500
15000
12500
10000
7500
5000
2500
0
RENT
MORTGAGE
OWN
OTHER
Home Ownership

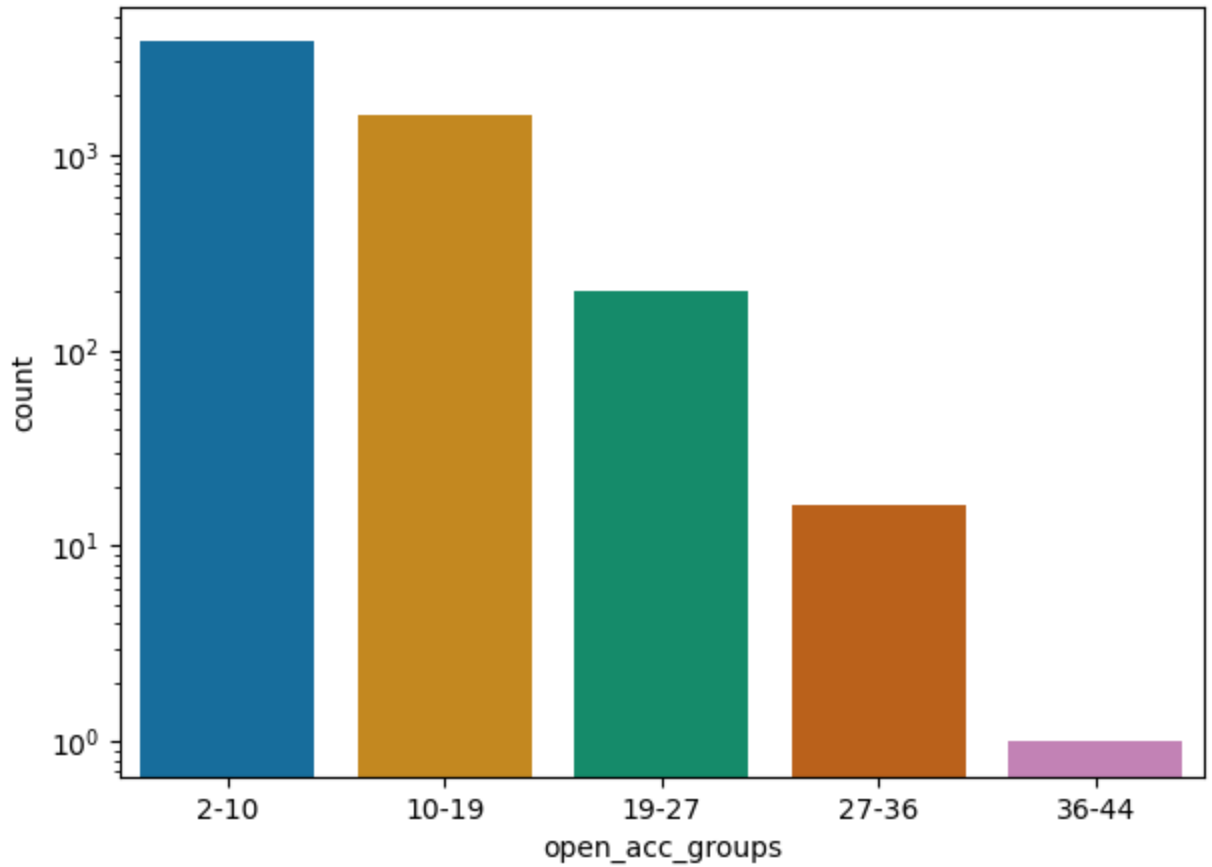

Default Rate Analysis by Home Ownership
Loan Status
Fully Paid
Charged Off
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
RENT
MORTGAGE
OWN
OTHER
Home Ownership

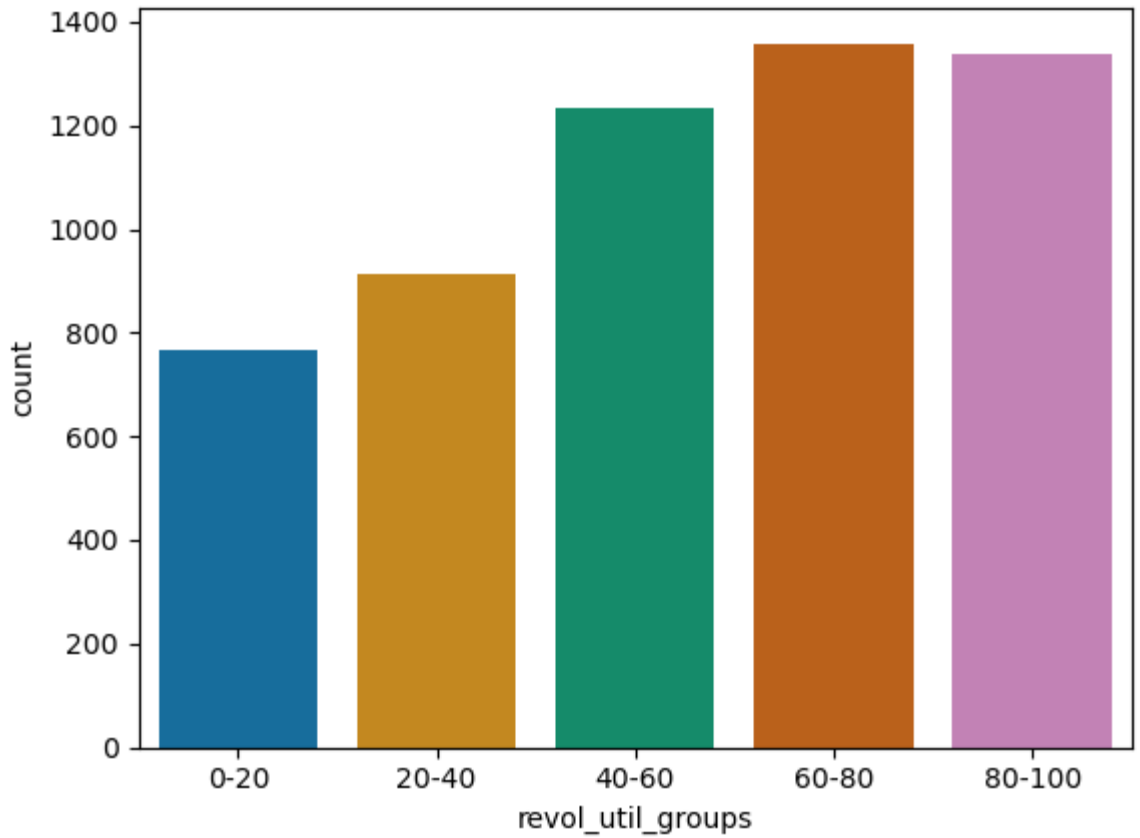

count
10^3
10^2
RENT
OWN
MORTGAGE
OTHER
home_ownership

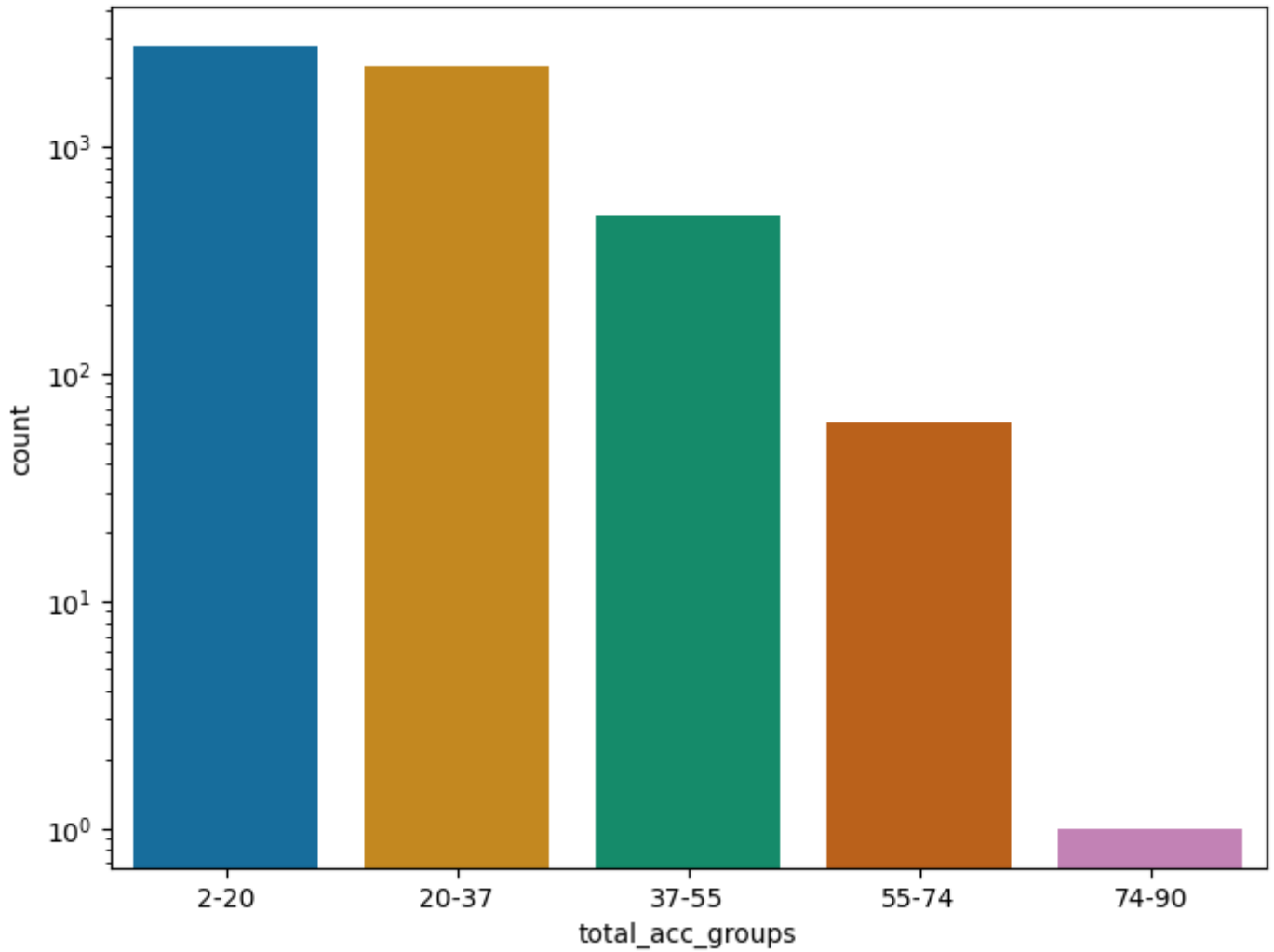

Purpose Analysis for Charged-Off Loans
car
small_business
other
debt_consolidation
major_purchase
credit_card
home_improvement
moving
vacation
house
medical
wedding
renewable_energy
educational
Loan Purpose
10^2
10^3
Count (log scale)

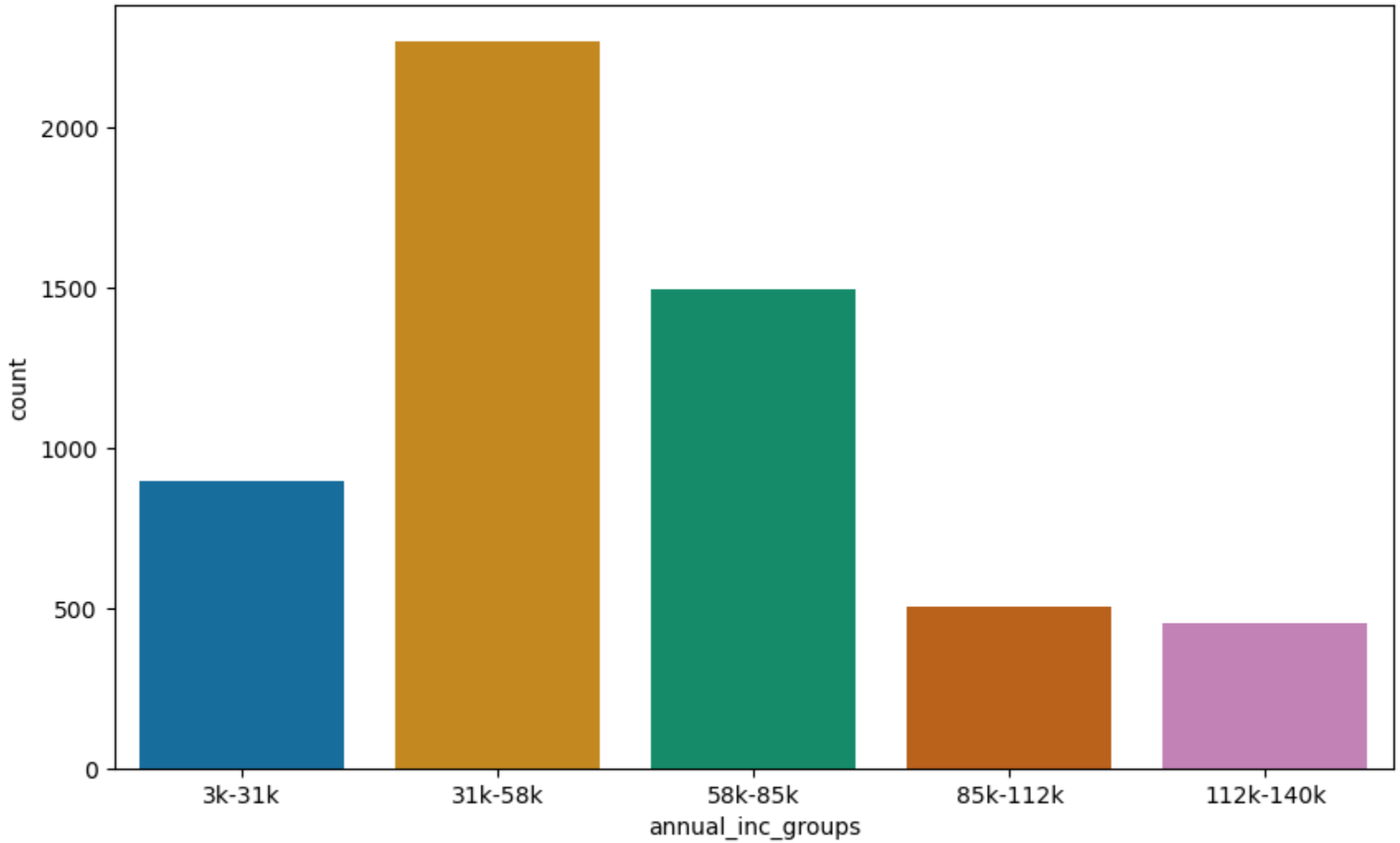

2000
1750
1500
1250
1000
750
500
250
0
count
5%-9%
9%-13%
13%-17%
17%-21%
21%-24%
Interest Rate

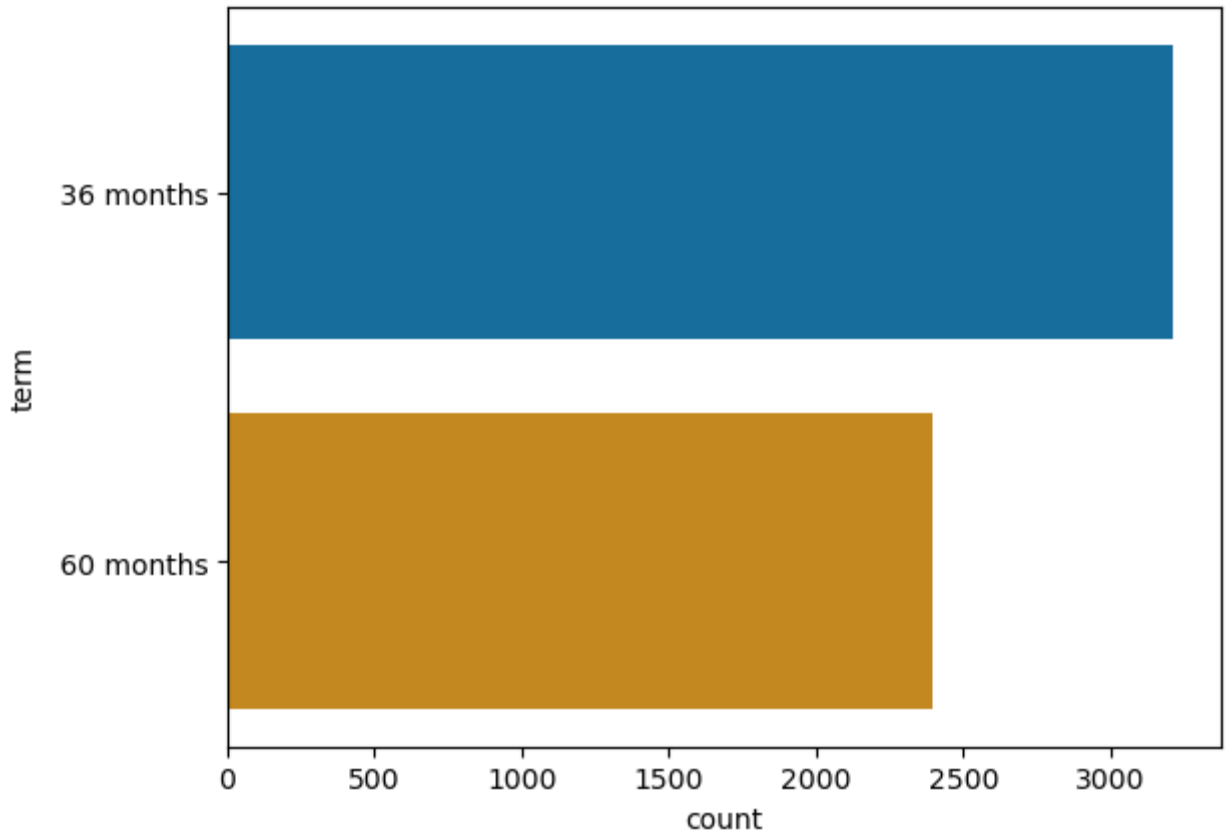

1600
1400
1200
1000
800
600
400
200
0
count
0
1
2
3
4
5
6
7
8
9
10
emp_length

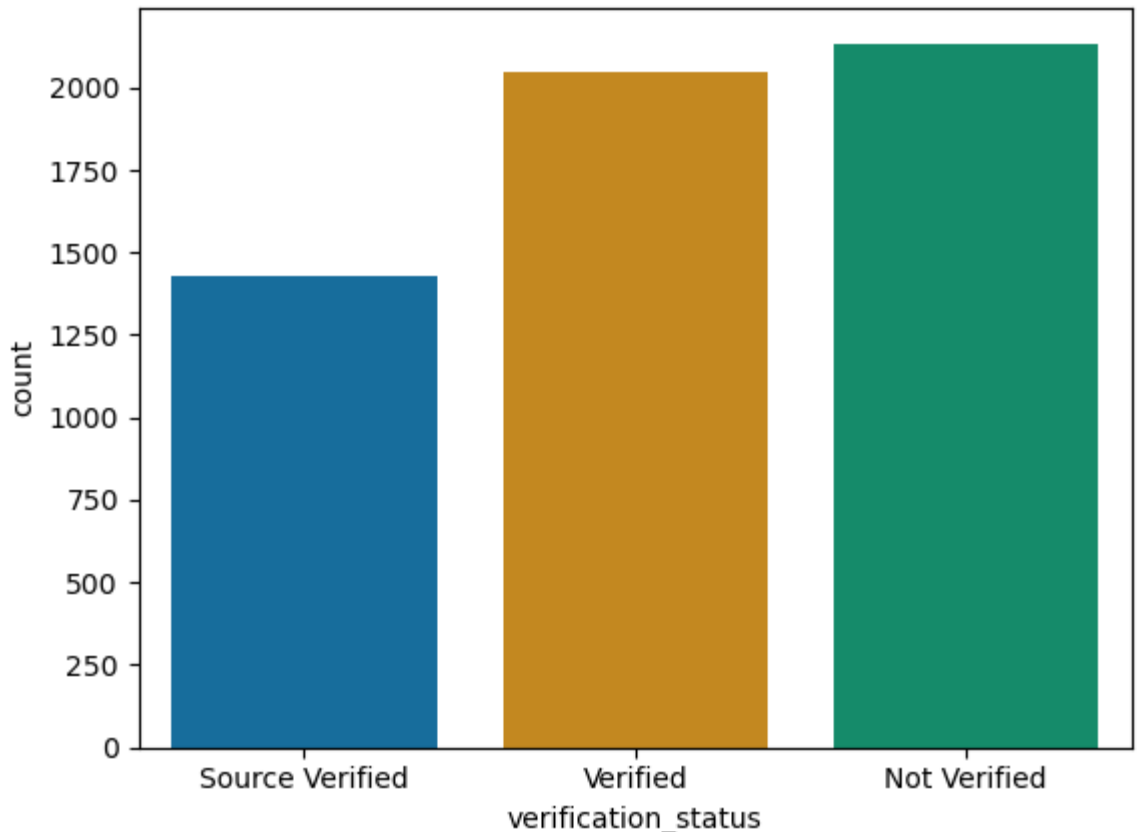

count
10^3
10^2
10^1
10^0
2-10
10-19
19-27
27-36
36-44
open_acc_groups

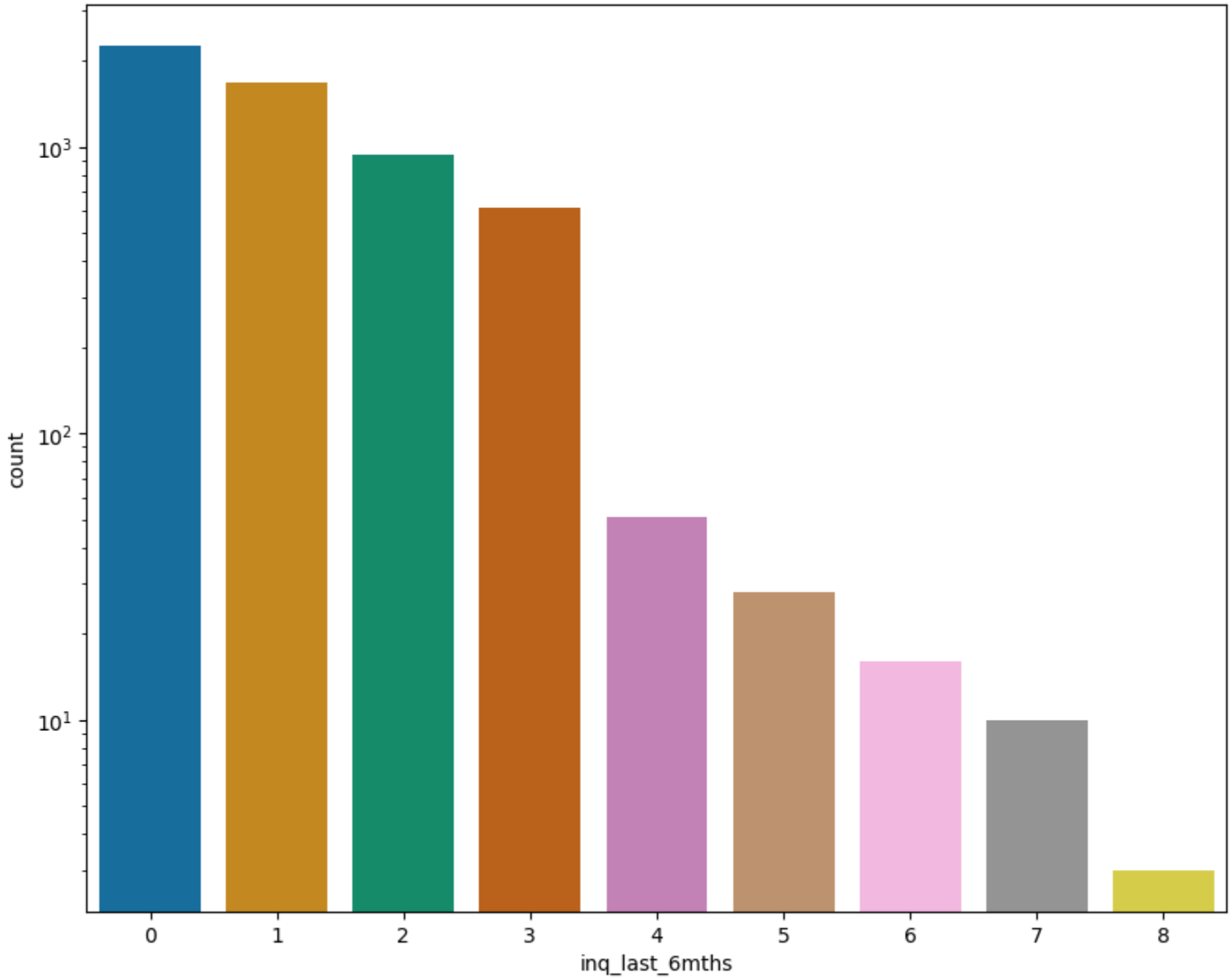

1400
1200
1000
800
600
400
200
0
count
0-20
20-40
40-60
60-80
80-100
revol_util_groups

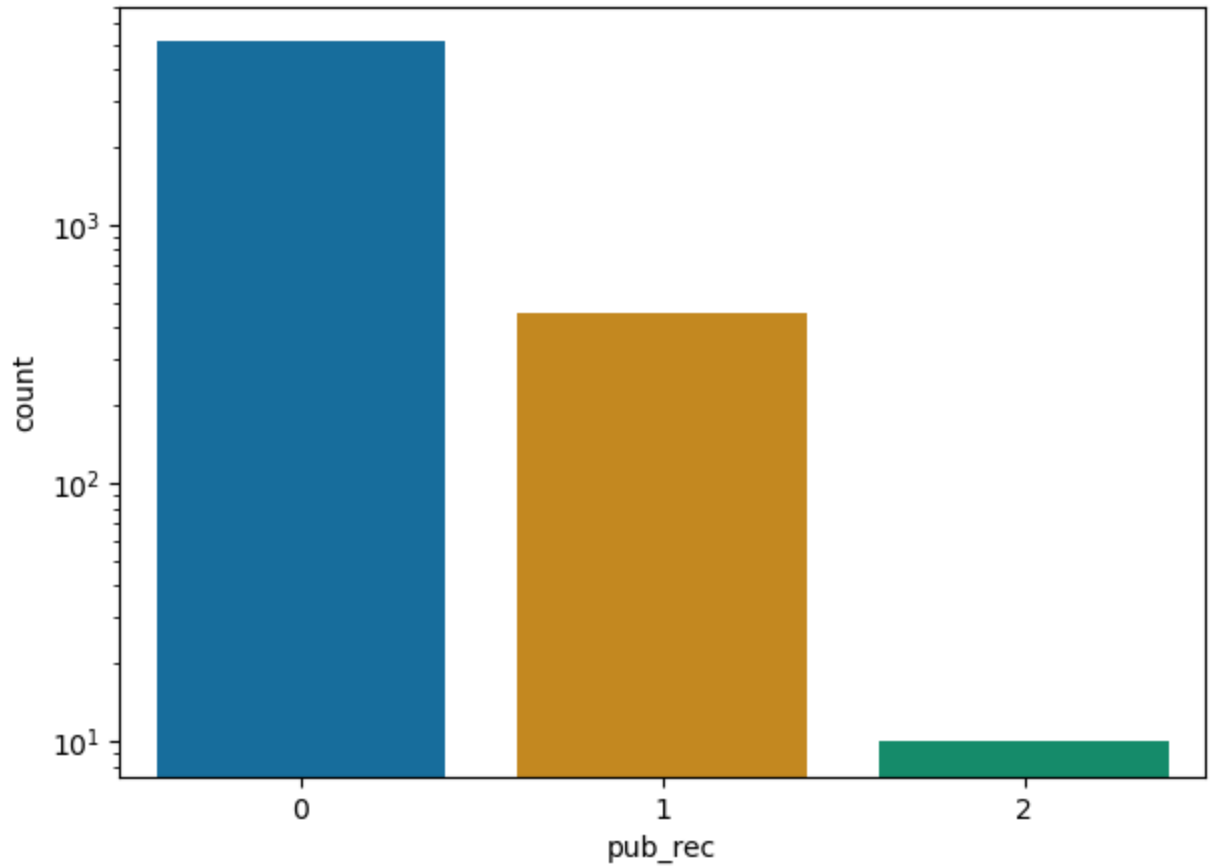

count
10^3
10^2
10^1
10^0
2-20
20-37
37-55
55-74
74-90
total_acc_groups

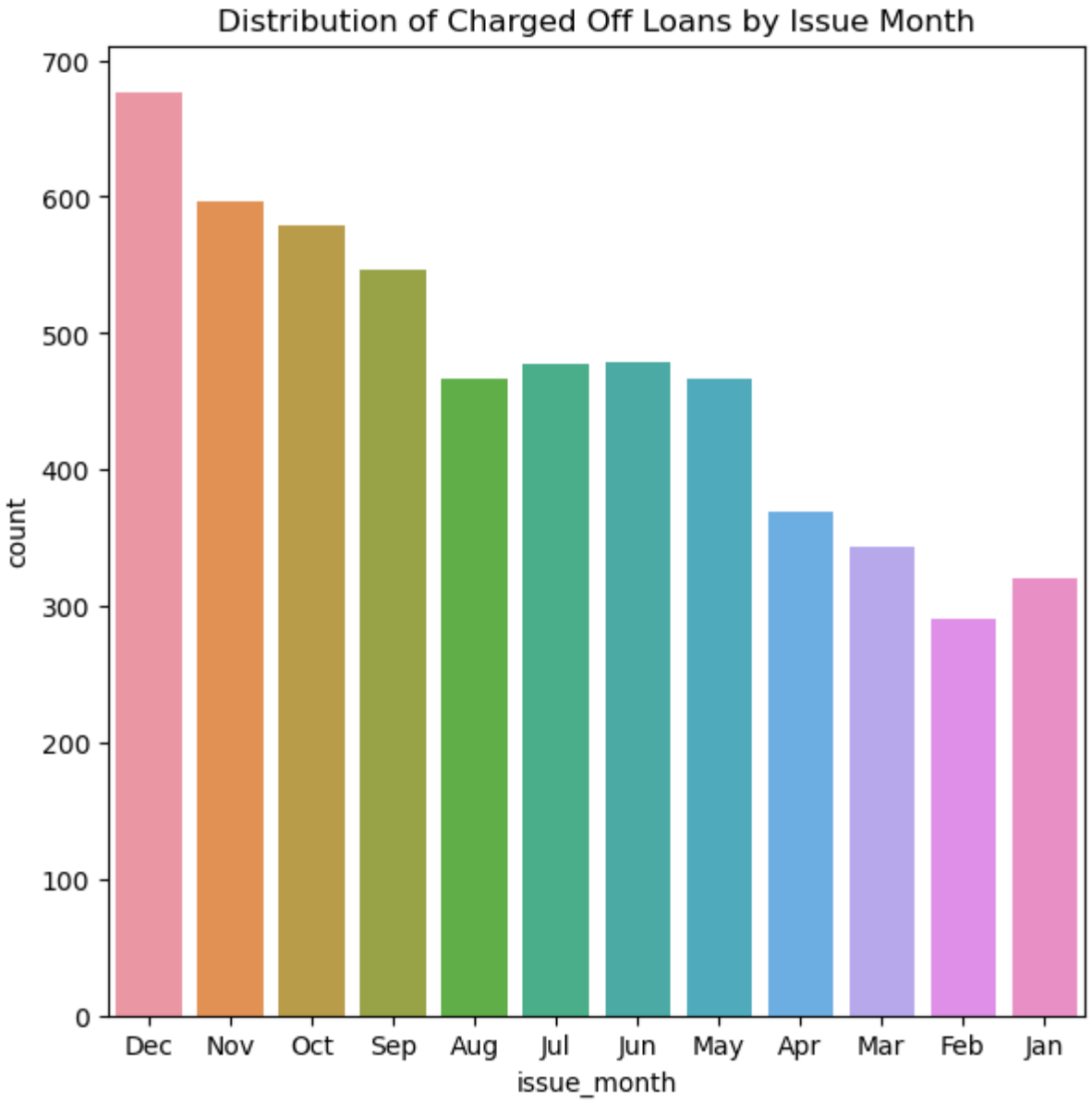

2000
1500
1000
500
0
count
3k-31k
31k-58k
58k-85k
85k-112k
112k-140k
annual_inc_groups

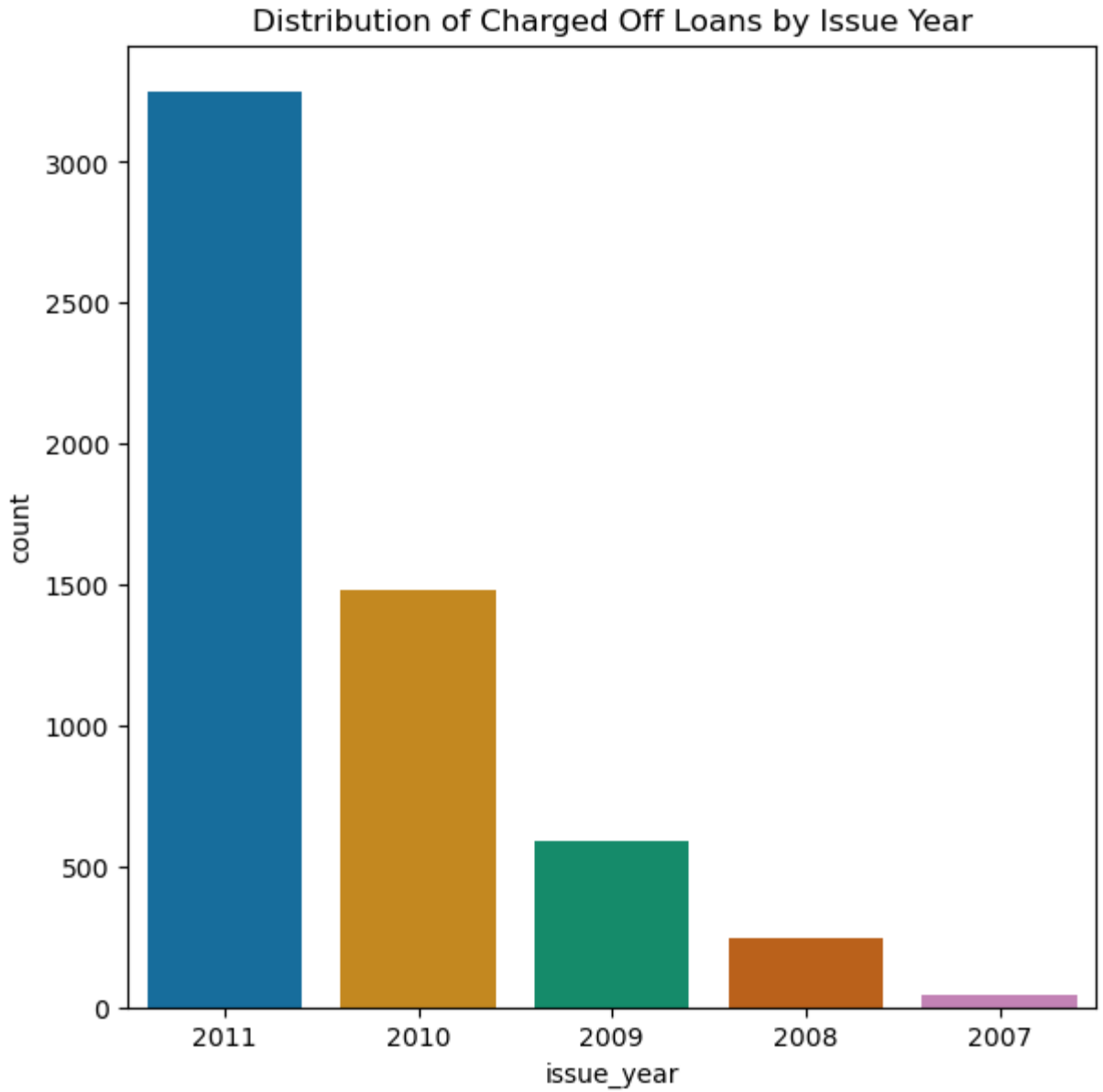

36 months
60 months
term
0
500
1000
1500
2000
2500
3000
count


2000
1750
1500
1250
1000
750
500
250
0
count
Source Verified
Verified
Not Verified
verification_status


count
10^3
10^2
10^1
0
1
2
3
4
5
6
7
8
inq_last_6mths


count
10^3
10^2
10^1
0
1
2
pub_rec


Distribution of Charged Off Loans by Issue Month
700
600
500
400
300
200
100
0
count
Dec
Nov
Oct
Sep
Aug
Jul
Jun
May
Apr
Mar
Feb
Jan
issue_month


Distribution of Charged Off Loans by Issue Year
3000
2500
2000
1500
1000
500
0
count
2011
2010
2009
2008
2007
issue_year

Analysis of Loan Issuance Patterns for Charged Off Loans

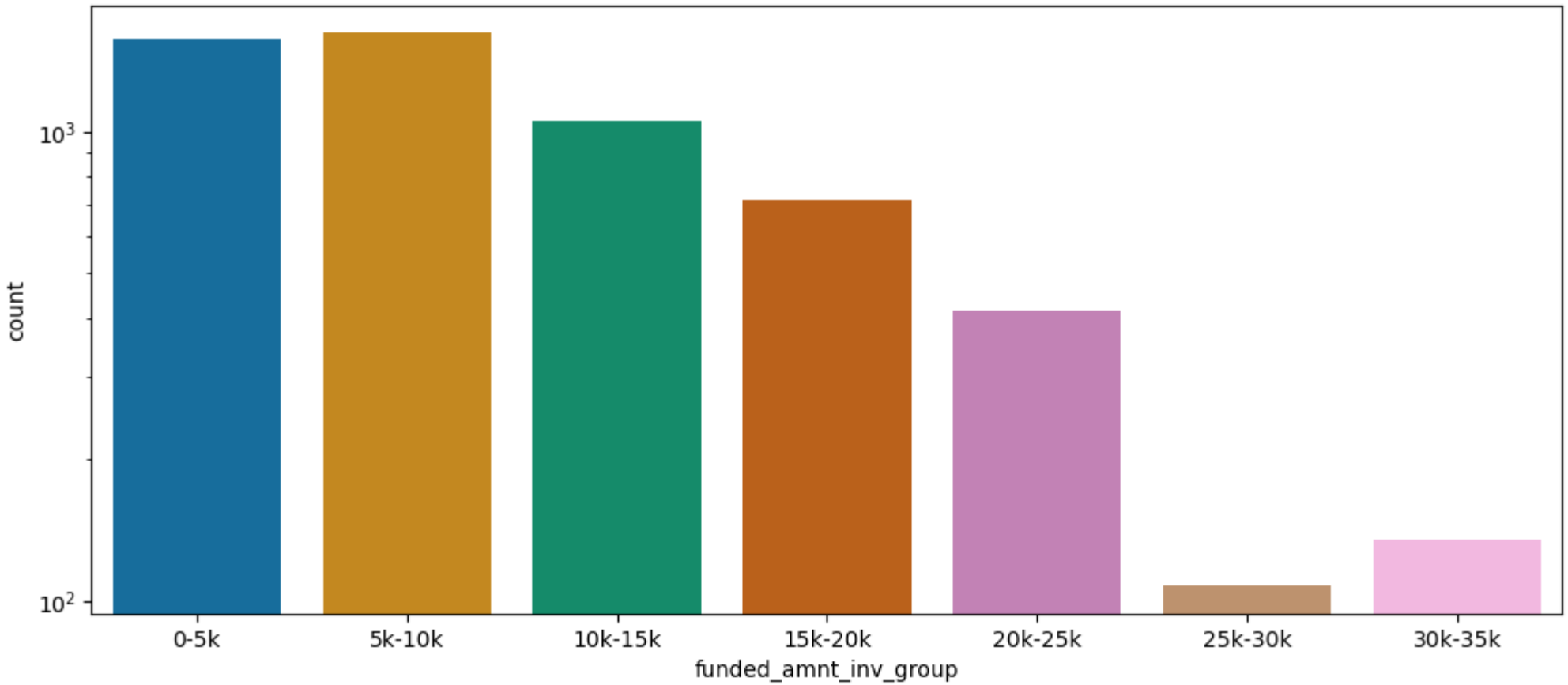
• Issue Month:

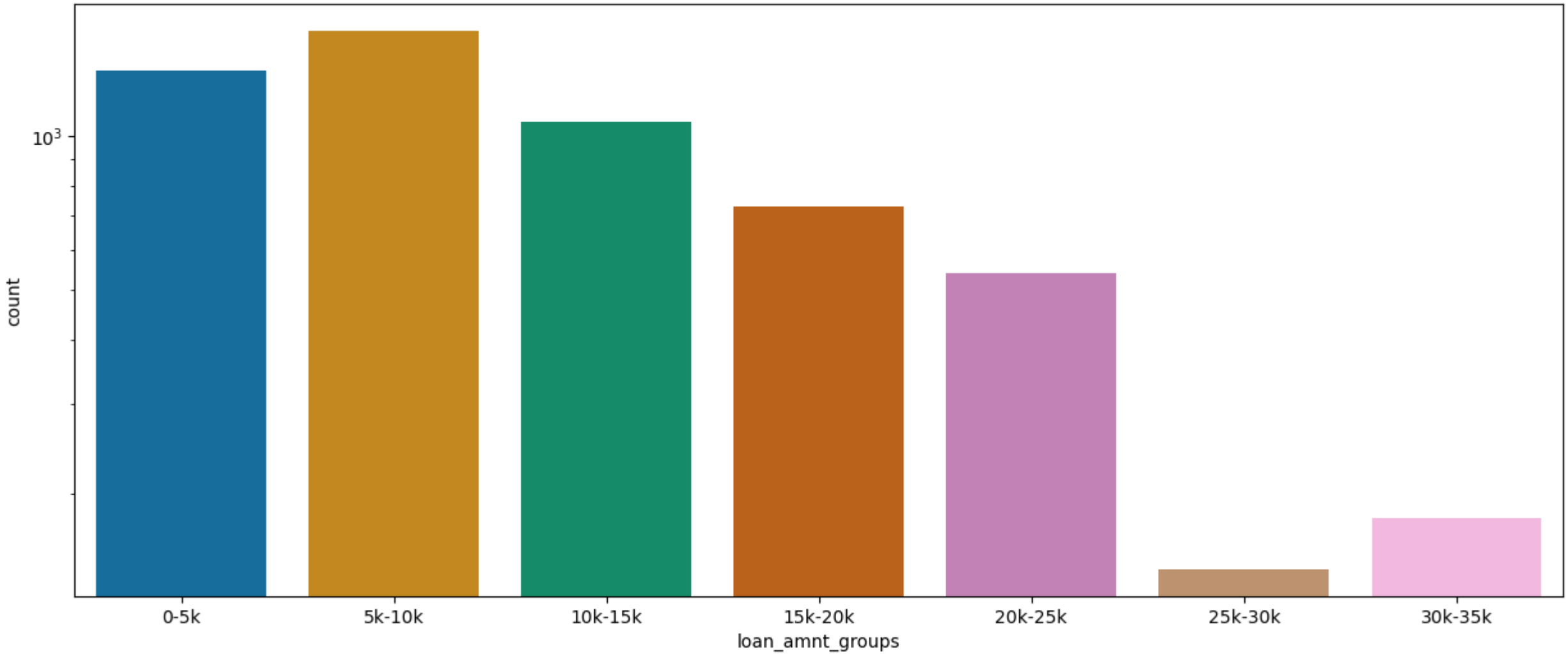
The analysis of 'Charged Off' loans based on the month of issuance reveals interesting patterns. The maximum number of defaults occurred when the loans were sanctioned in December. This observation could be attributed to various factors, such as year-end financial pressures or holiday-related expenses.

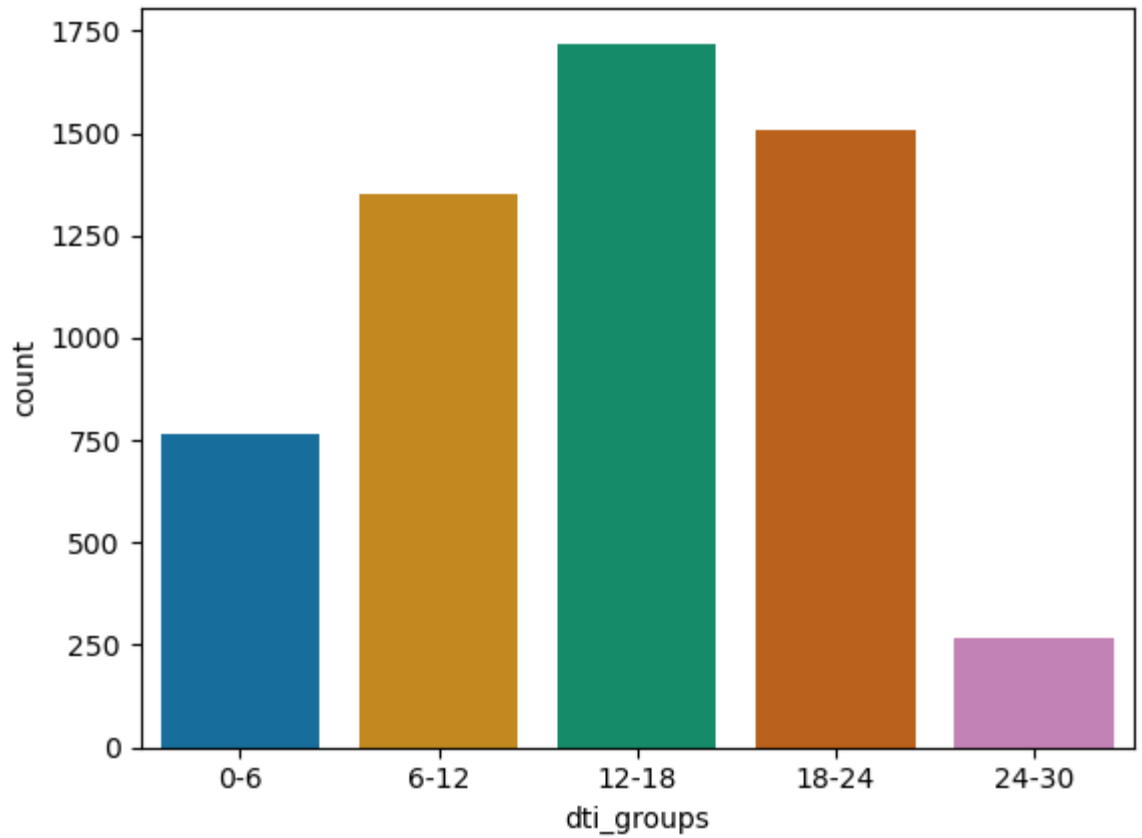
• Issue Year:

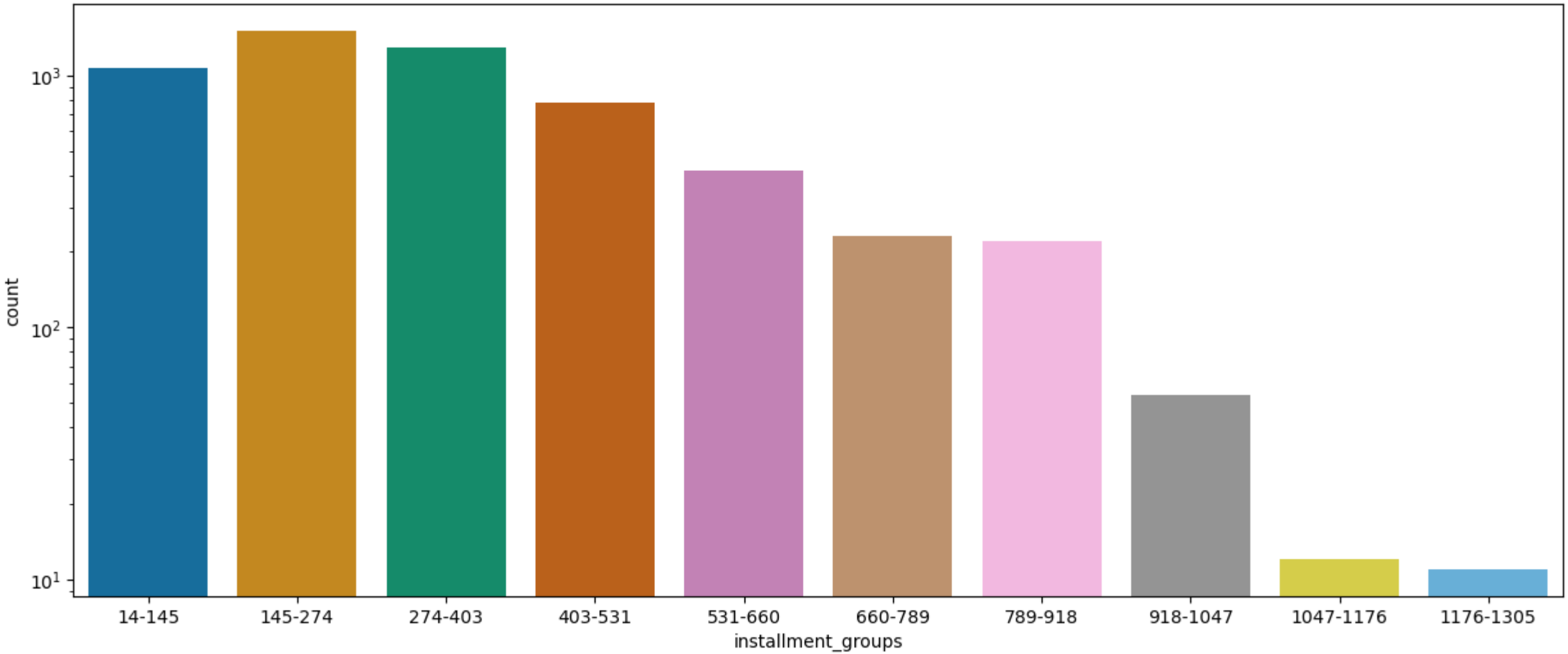
When examining the distribution by issue year, it is noteworthy that loans issued in the year 2011 show a higher occurrence of defaults compared to other years. This finding might indicate specific economic conditions or lending practices during that period.

These insights provide valuable information for understanding the relationship between loan issuance patterns and the likelihood of default, contributing to our broader analysis of loan defaulting behavior.


count
10^3
10^2
0-5k
5k-10k
10k-15k
15k-20k
20k-25k
25k-30k
30k-35k
funded_amnt_inv_group


10^3
count
0-5k
5k-10k
10k-15k
15k-20k
20k-25k
25k-30k
30k-35k
loan_amnt_groups


1750
1500
1250
1000
750
500
250
0
count
0-6
6-12
12-18
18-24
24-30
dti_groups


count
10^3
10^2
10^1
14-145
145-274
274-403
403-531
531-660
660-789
789-918
918-1047
1047-1176
1176-1305
installment_groups

These observations are derived from the analysis done so far and highlighted the potential factors that may be associated with a higher probability of defaulting in charged-off loans.

Here's a breakdown of the key findings:

House Ownership: Applicants with house ownership listed as 'RENT' are more likely to default.

Loan Purpose: Loans intended to clear other debts have a higher likelihood of default.

Interest Rate: Loans with interest rates in the range of 13-17% are associated with a higher probability of default.

Annual Income: Applicants with an annual income in the range of 31,201 - 58,402 are more likely to default.

Open Accounts: Applicants with 20-37 open accounts have a higher probability of defaulting.

Employment Length: Applicants with an employment length of 10 years are more likely to default.

Funded Amount by Investor: Loans with funded amounts between 5,000 - 10,000 have a higher likelihood of default.

Loan Amount: Loans in the amount range of 5,429 - 10,357 are associated with a higher probability of default.

DTI (Debt-to-Income Ratio): Loans with DTI in the range of 12-18 have a higher likelihood of default.

Monthly Installments: Loans with monthly installments between 145-274 are more likely to default.

Loan Term: Loans with a term of 36 months have a higher probability of default.

Verification Status: Loans with a status of 'Not Verified' are associated with a higher likelihood of default.

Inquiries in Last 6 Months: Loans with no inquiries in the last 6 months are more likely to default.

Derogatory Public Records: Loans with no derogatory public records have a higher likelihood of default.

Loan Purpose ('debt_consolidation'): Loans with the purpose of debt consolidation are more likely to default.

Grade and Subgrade: Loans with a grade of 'B' and a subgrade of 'B5' have a higher probability of default.

Date Issued: Loans issued in the later months of the year, particularly in 2011, show a higher occurrence of defaults, potentially related to economic conditions.

Analyzing annual income with other columns

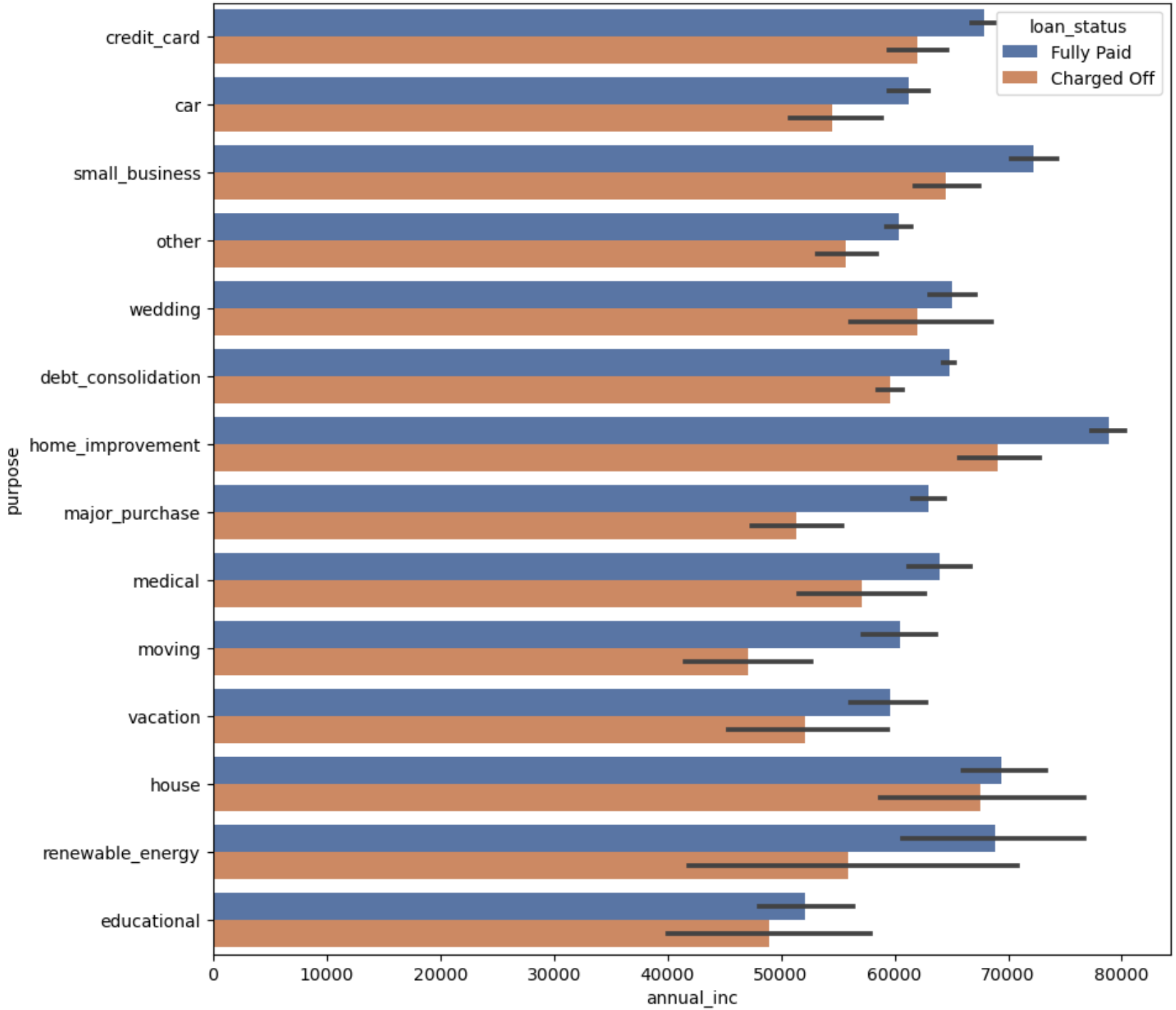
1.Annual income vs loan purpose

Insights from Annual Income vs. Loan Purpose Analysis

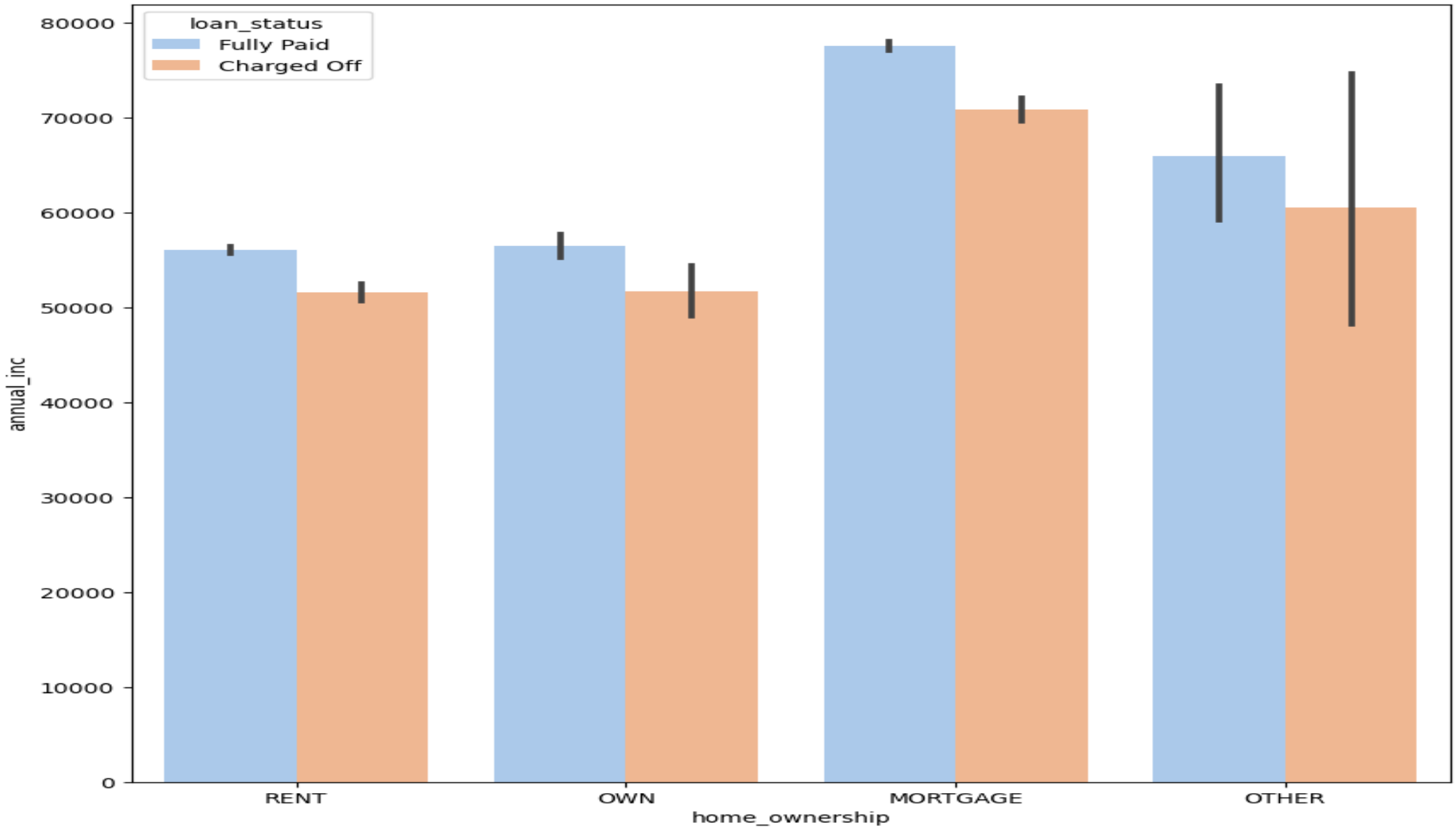
The bar plot illustrates the relationship between annual income and loan purpose, distinguishing between Charged Off and Fully Paid loan statuses. Different income levels are associated with varying loan purposes, and the color differentiation provides a clear view of the defaulting patterns across different purposes.

While "debt_consolidation" has the highest number of loan applications and defaults, it's interesting to note that applicants with higher annual incomes tend to apply for loans related to "home_improvement," "house," "renewable_energy," and "small_businesses.“

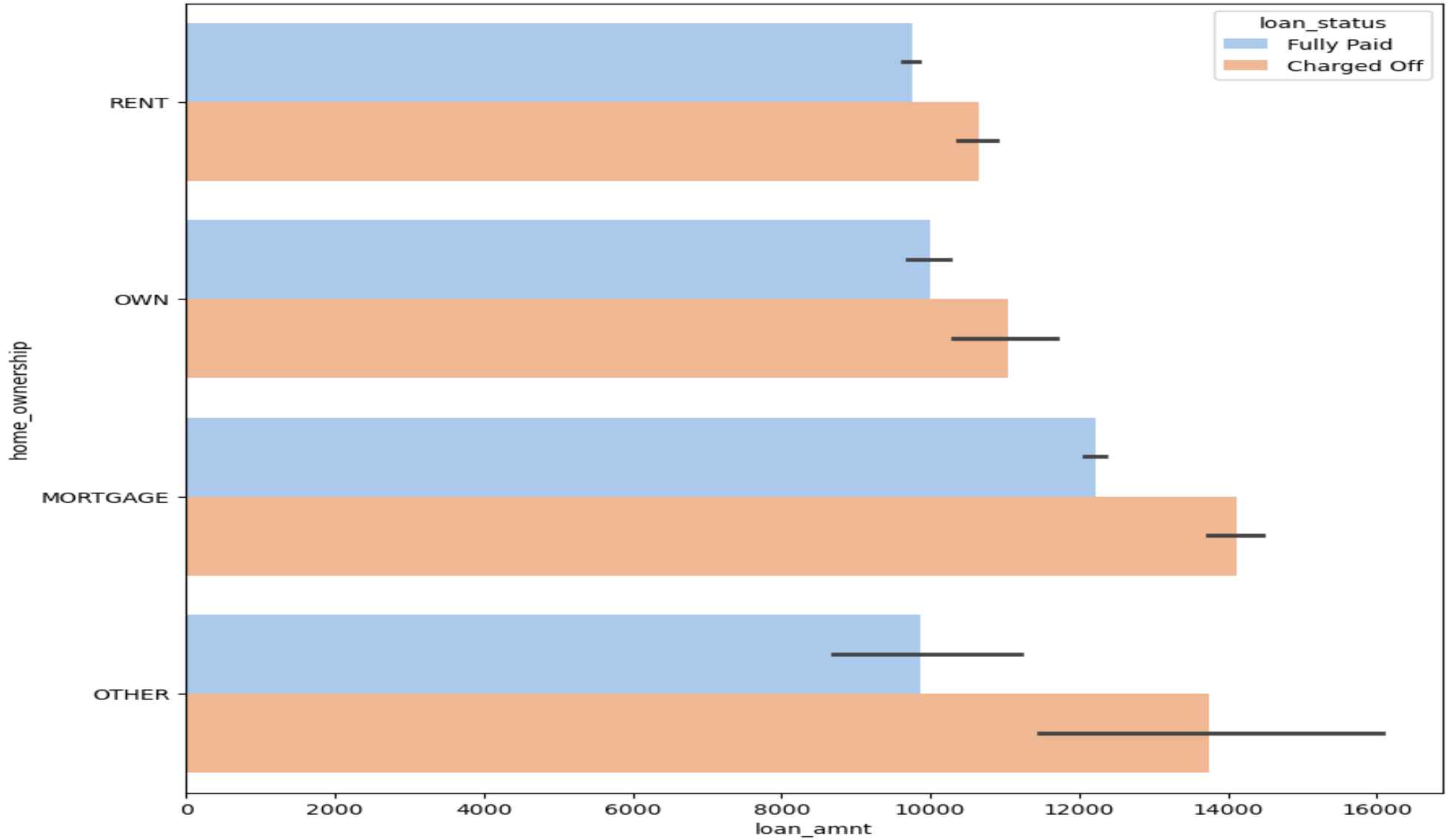
The purpose of the loan seems to vary based on the applicant's income level, indicating diverse financial goals and needs among borrowers.


loan_status
Fully Paid
Charged Off
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational
purpose
0
10000
20000
30000
40000
50000
60000
70000
80000
annual_inc

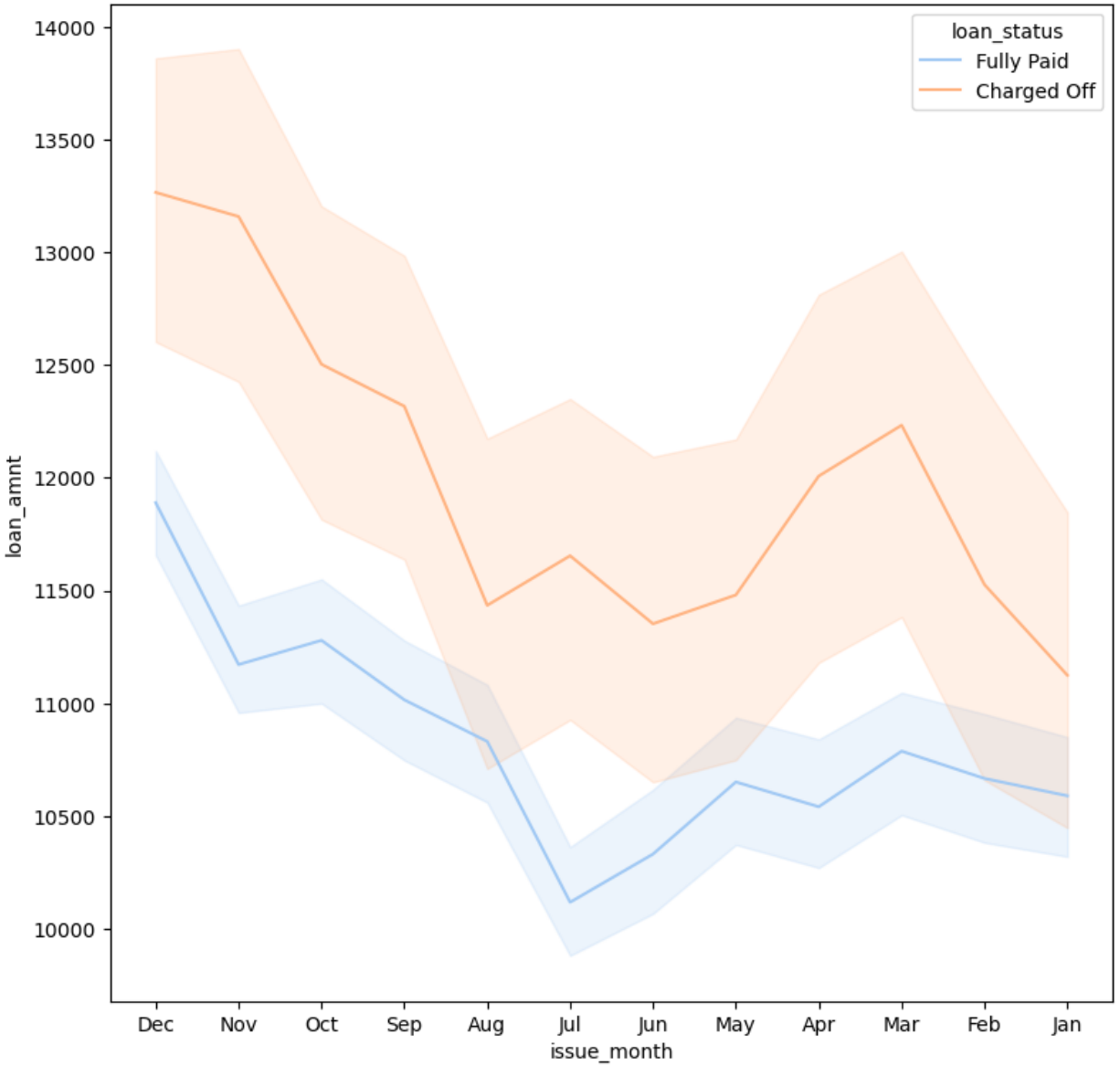
2.Annual income vs home ownership

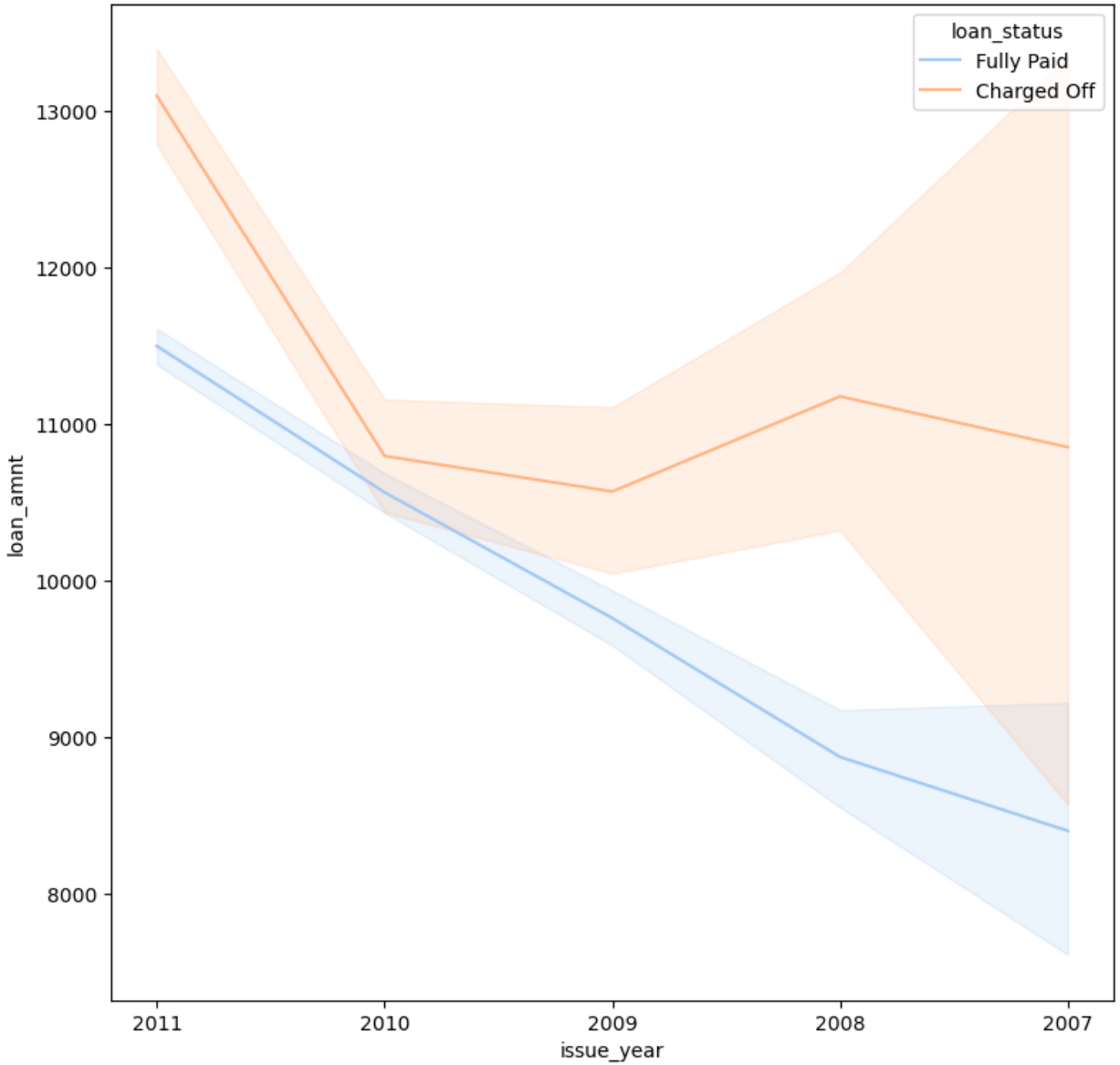


3.Loan vs House Ownership

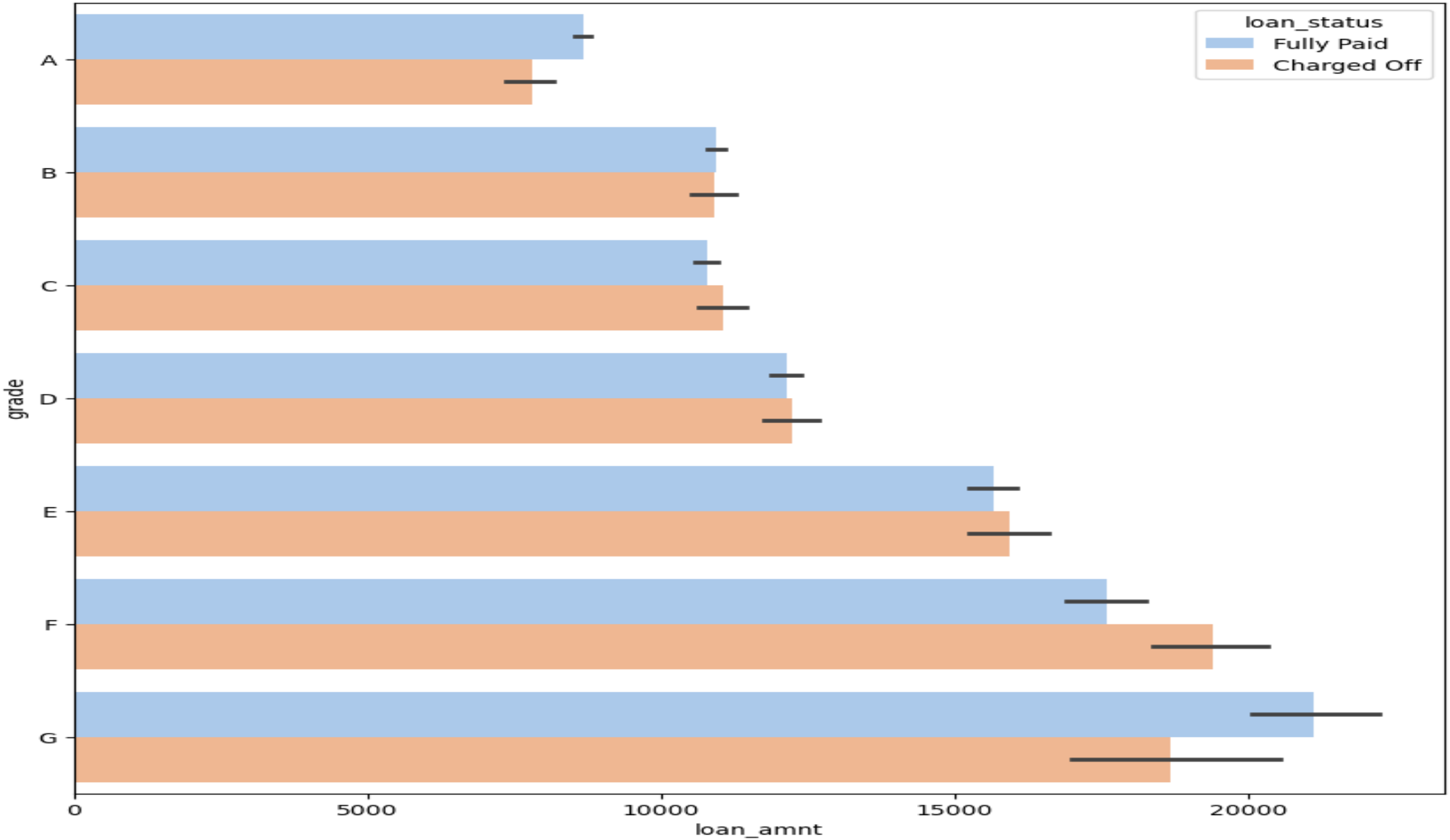


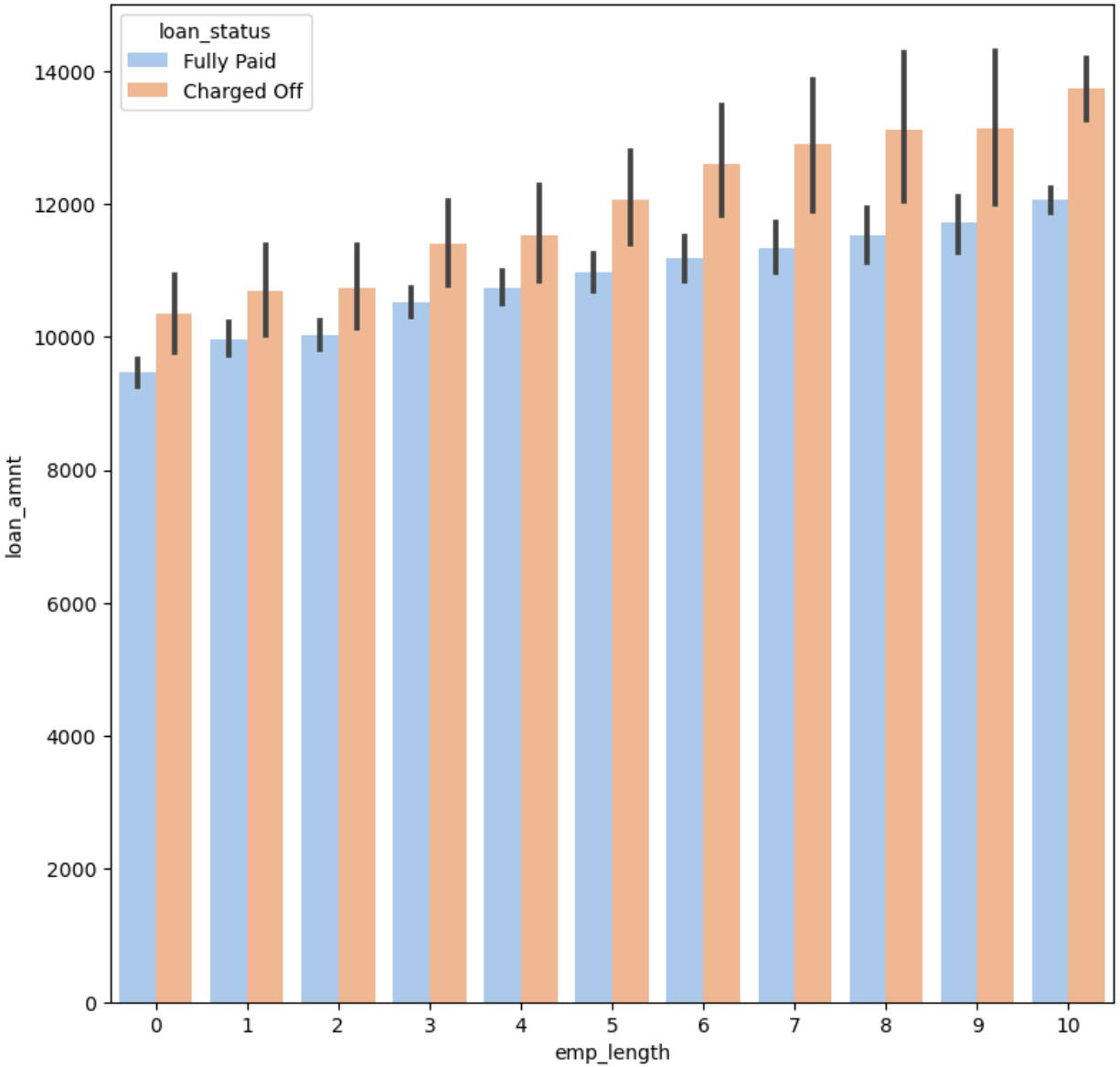
4.Loan amount vs month issued and year issued

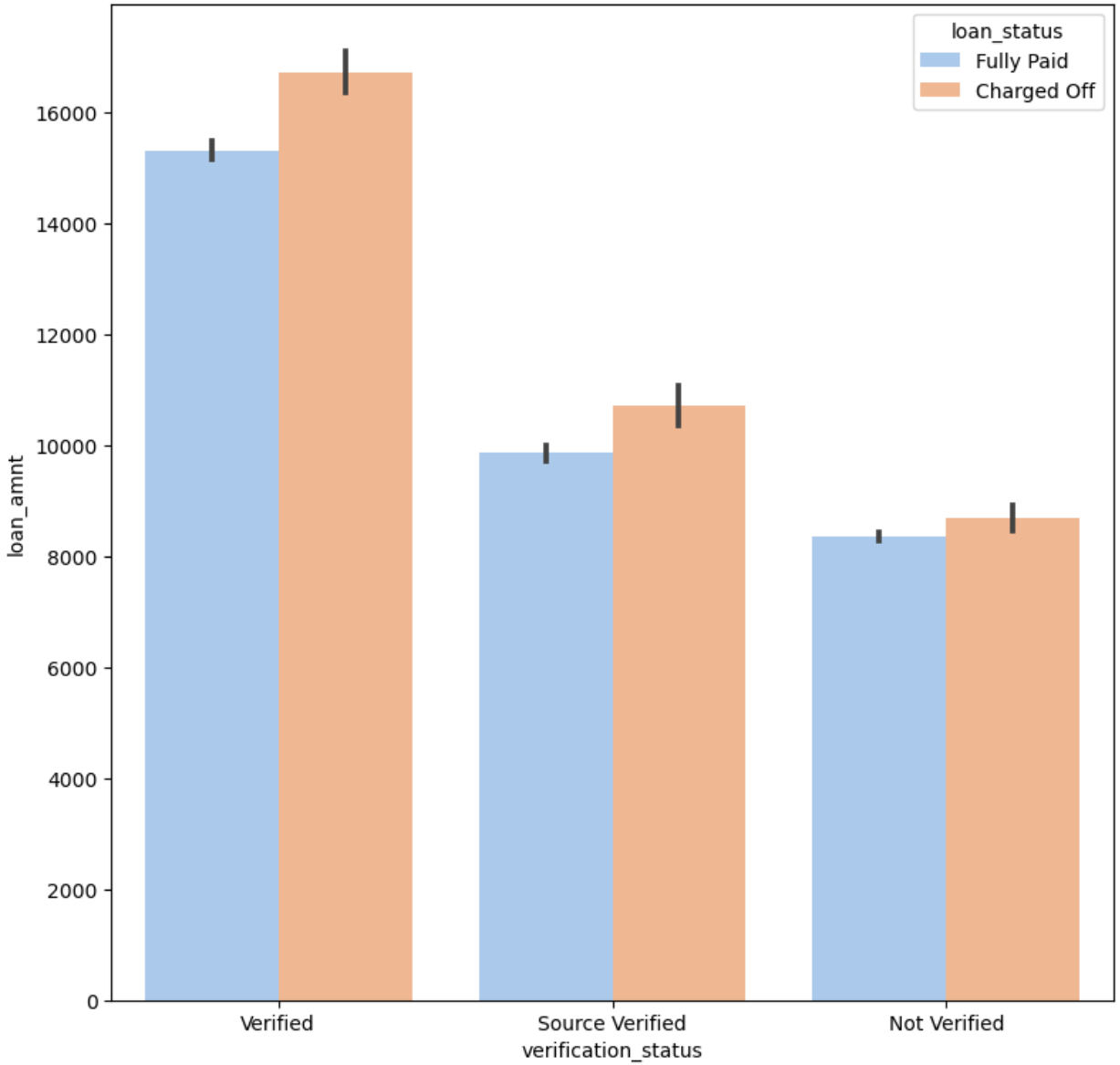




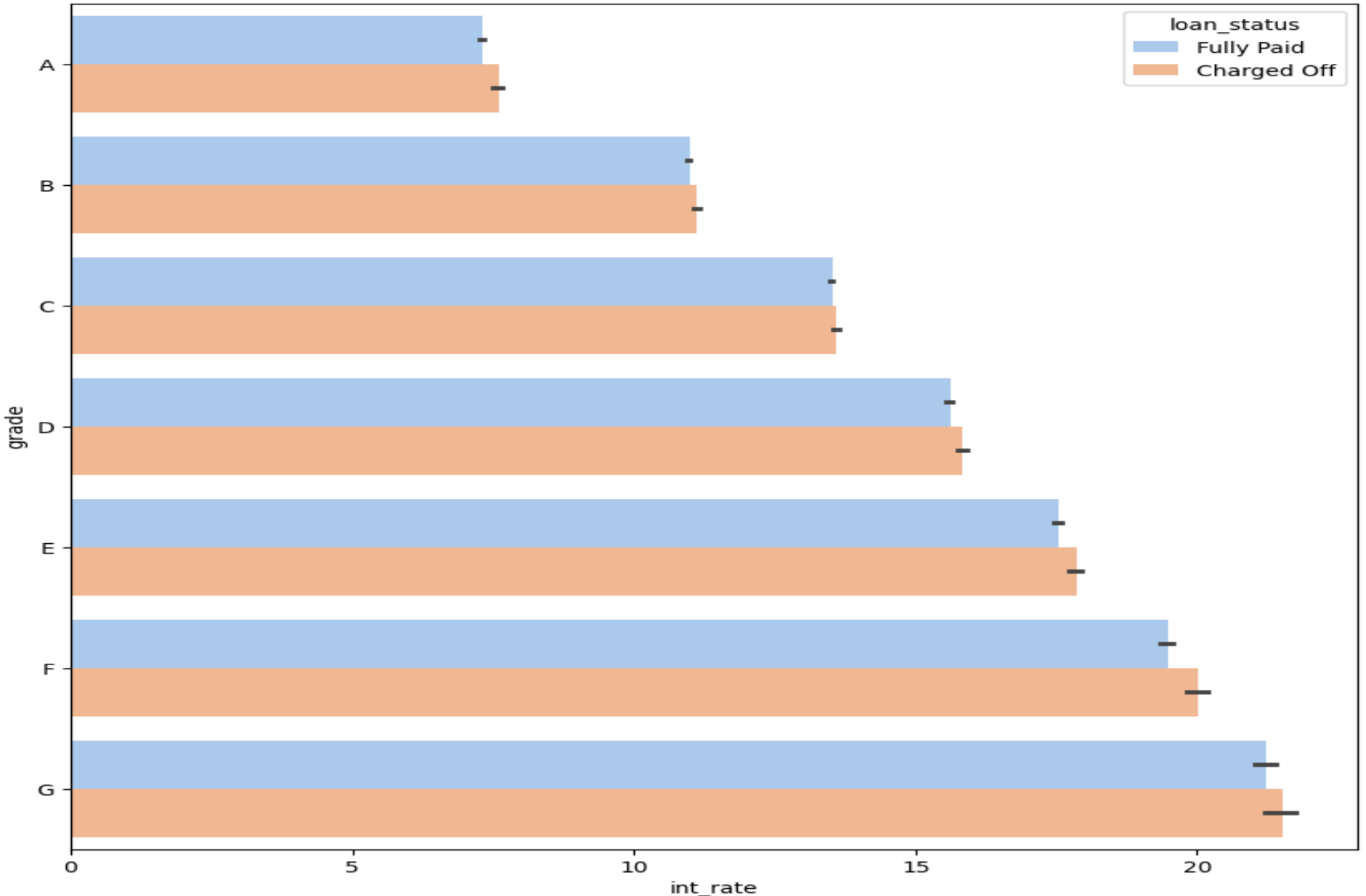
5.Loan amount vs Grade

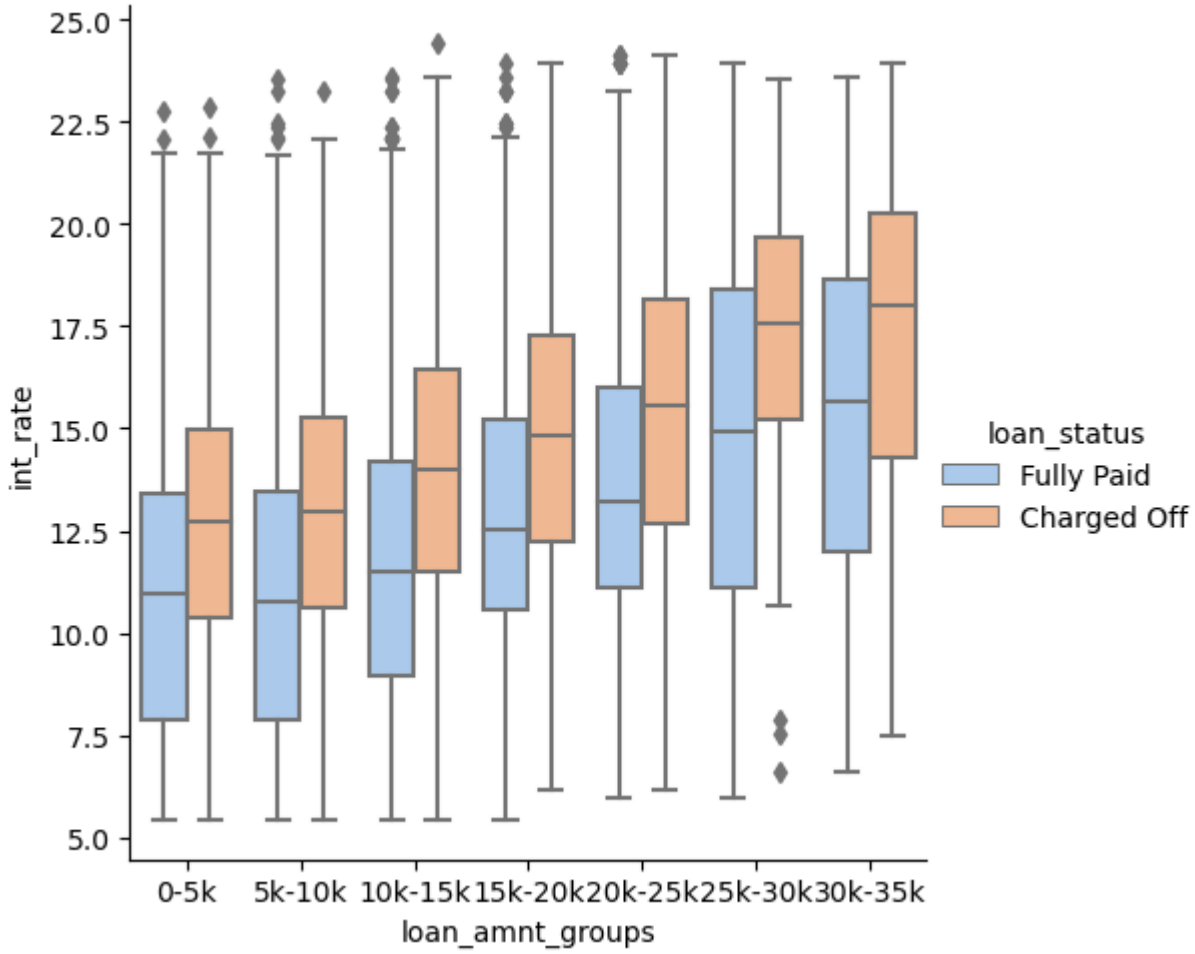



loan_status
Fully Paid
Charged Off
loan_amnt
0
2000
4000
6000
8000
10000
12000
14000
0
1
2
3
4
5
6
7
8
9
10
emp_length


loan_status
Fully Paid
Charged Off
loan_amnt
0
2000
4000
6000
8000
10000
12000
14000
16000
Verified
Source Verified
Not Verified
verification_status

grade vs interest rate



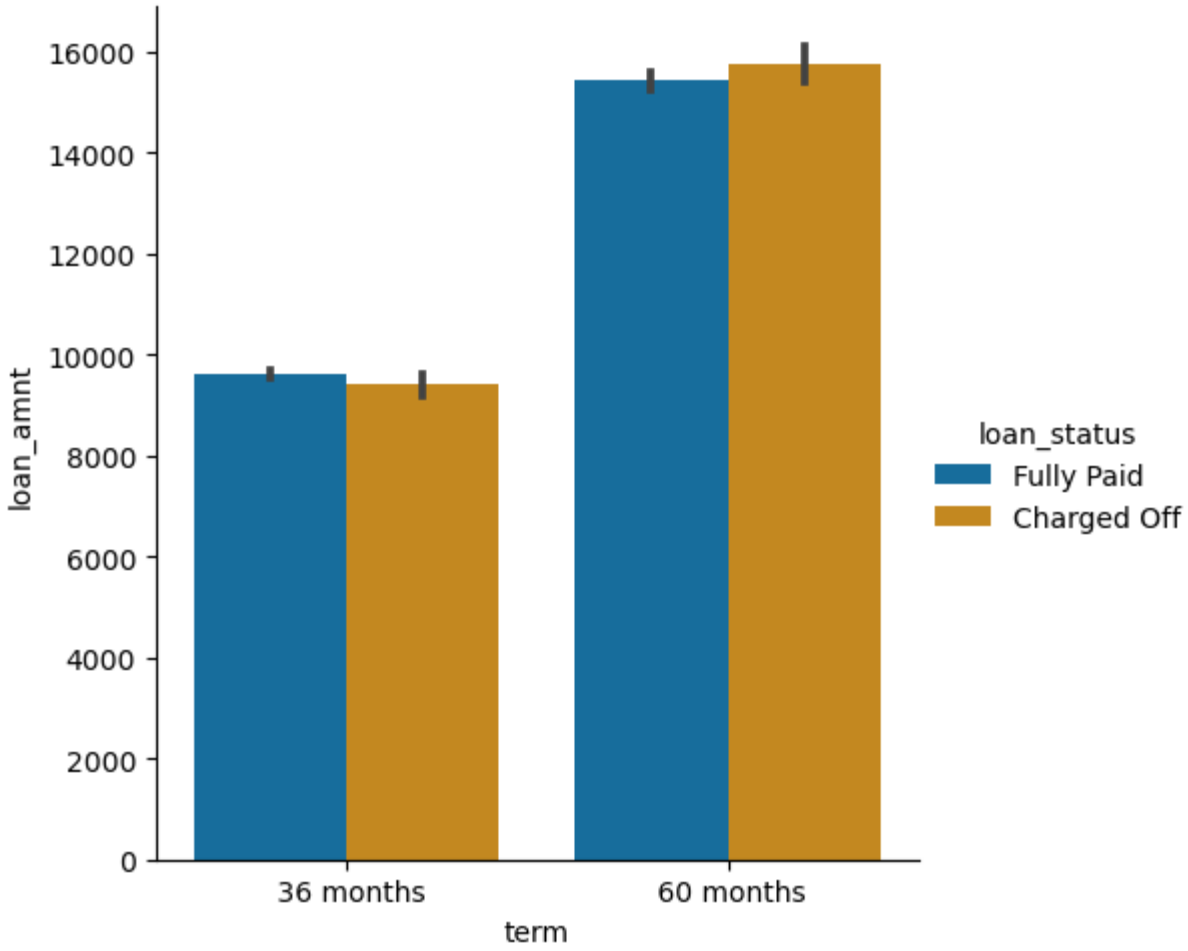


Interest Rate and Loan Defaulting

The analysis indicates that the interest rate for charged-off loans is consistently higher across all loan amount groups compared to fully paid loans.

This observation suggests a strong correlation between higher interest rates and the likelihood of loan defaulting.

The interest rate appears to be a significant driving factor influencing the default patterns in the dataset.



Loan Amount and Loan Term Analysis

There is no significant difference in loan amounts between applicants who applied and defaulted.

The observation suggests that applicants applying for long-term loans tend to request higher loan amounts.

The relationship between loan amount and loan term is an essential factor to consider in understanding borrowing patterns and potential defaulting behavior.

Insights from Charged Off Loan Analysis

The analysis of charged-off loans reveals the following insights:

Loan Purpose and Income:

Applicants taking loans for 'home improvement' with an income range of 60k - 70k are more likely to default.

Home Ownership and Income:

Applicants with 'MORTGAGE' home ownership and an income of 60k - 70k have a higher probability of default.

Interest Rate and Income:

Applicants charged an interest rate of 21-24% with an income of 70k-80k are more prone to default.

Loan Amount and Interest Rate:

Default probability is higher when the loan amount is in the range 30k - 35k, and the interest rate is between 15-17.5%.

Loan Purpose and Amount:

Small business loans with amounts greater than 14k have a higher likelihood of default.

Home Ownership and Loan Amount:

'MORTGAGE' home ownership with a loan amount in the range of 14k - 16k is associated with a higher default rate.

Grade and Loan Amount:

Loans with grade 'F' and amounts between 15k-20k have a higher chance of default.

Employment Length and Loan Amount:

Applicants with employment length of 10 years and loan amounts between 12k-14k are more likely to default.

Verification Status and Loan Amount:

Verified loans with amounts above 16k are associated with a higher probability of default.

Grade G and High Interest Rate:

Loans with grade 'G' and interest rates above 20% have an increased likelihood of default.

These insights provide a detailed understanding of the factors contributing to loan defaulting behavior.

RECOMMENDATIONS

1.Refine Loan Approval Criteria:
Refine criteria for applicants with 'home improvement' purposes and an annual income of 60k - 70k.

2.Mitigate Risk for 'MORTGAGE' Homeowners:
Implement risk mitigation for 'MORTGAGE' homeowners with an income of 60k - 70k to reduce default risks.

3.Adjust Interest Rates:
Consider adjusting interest rates for applicants earning 70k-80k, especially for those with rates of 21-24%.

4.Analyze Loan Amounts and Interest Rates:
Further analyze the relationship between loan amounts (30k - 35k) and interest rates (15-17.5%) for optimal lending practices.

5.Evaluate Small Business Loans:
Scrutinize small business loans exceeding 14k for better assessment and management of default risks.

6.Assess 'MORTGAGE' Homeownership Risk:
Evaluate the risk associated with 'MORTGAGE' homeownership and loan amounts (14k - 16k) for optimized lending decisions.

7.Manage Grade 'F' Loans:
Enhance monitoring of grade 'F' loans with amounts between 15k-20k to minimize defaults.

8.Consider Employment Length:
Consider employment length in loan approval decisions, especially for applicants with 10 years of history and amounts between 12k-14k.

9.Review Verification Status:
Review and assess default risk associated with verified loans, particularly those above 16k.

10.Evaluate Grade 'G' Loans:
Carefully evaluate and monitor grade 'G' loans with interest rates above 20% to mitigate potential defaults.

These recommendations aim to refine the company's loan approval processes, reduce default risks, and optimize lending strategies based on identified patterns and insights from the analysis.