Lending Club Case Study

Problem Statement:
Lending Club, a platform that specializes in providing diverse loans to city dwellers, is grappling with the significant task of managing its loan approval process. The company needs to make judicious decisions when assessing loan applications to reduce financial losses, which primarily arise from loans given to applicants deemed “high-risk.”

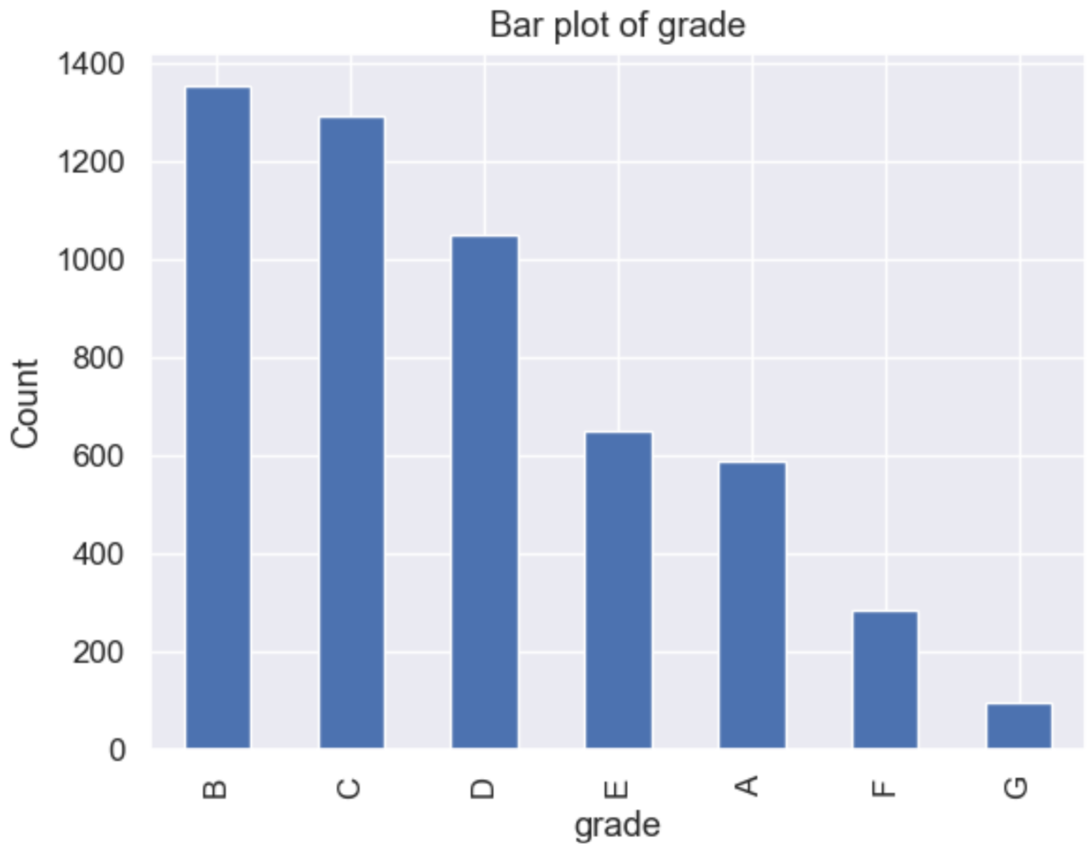
These financial losses, known as credit losses, happen when borrowers default or are unable to repay their loans. In layman’s terms, borrowers who are categorized as “charged-off” are the ones who cause the most substantial losses to the company.

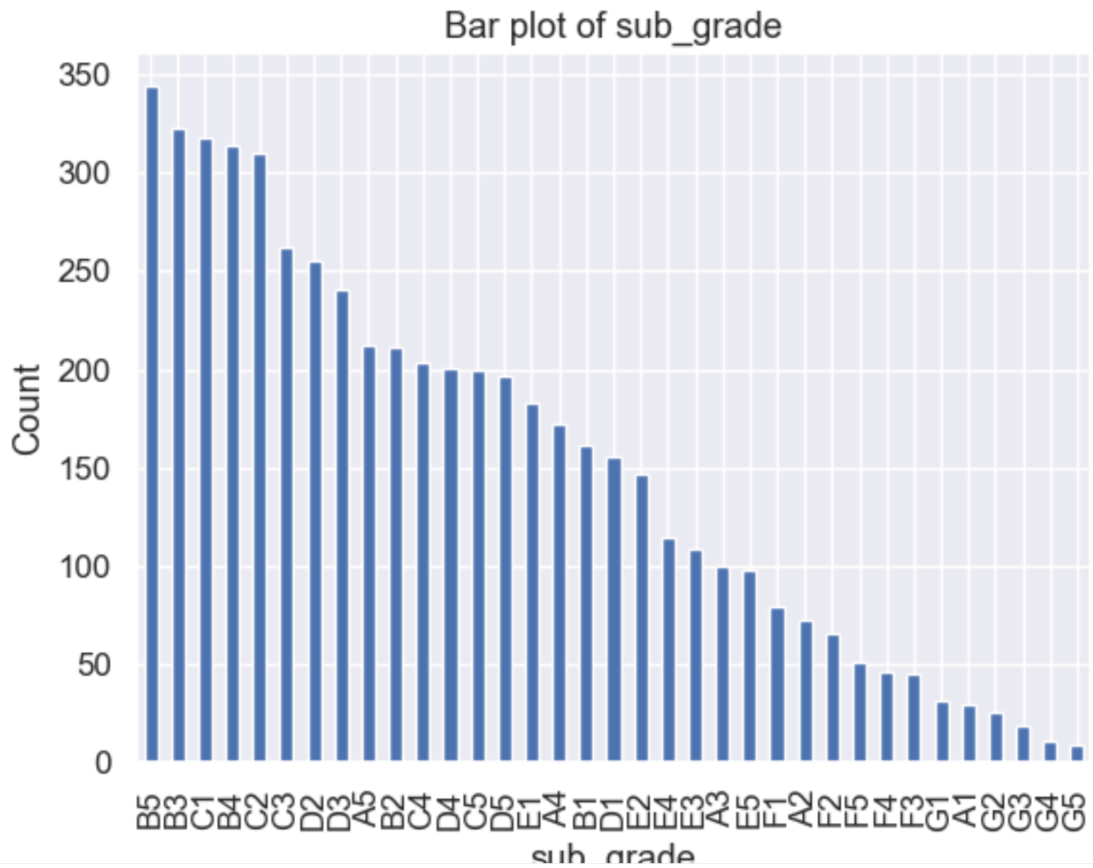
The main goal of this task is to help Lending Club reduce credit losses. This challenge stems from two possible situations:

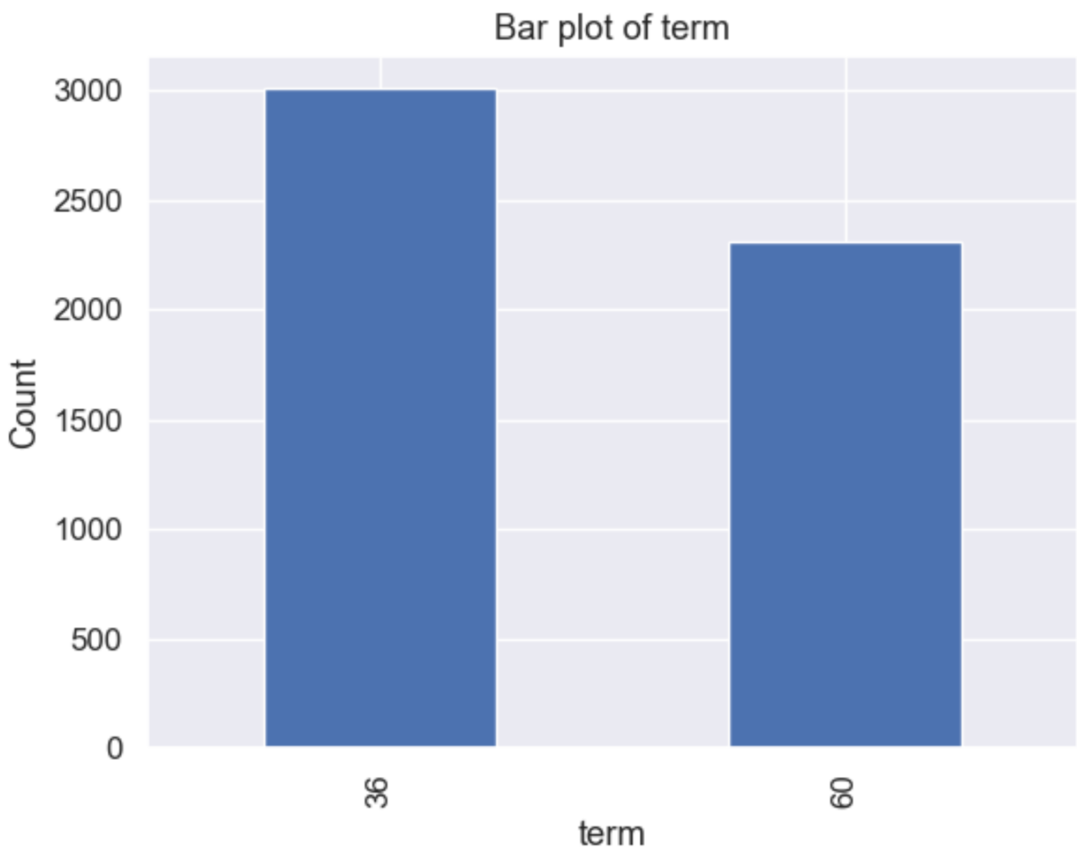
It’s vital to identify applicants who are likely to repay their loans as they can bring profits to the company through interest payments. Turning down such applicants would lead to a missed business opportunity. Conversely, approving loans for applicants who are unlikely to repay and are at risk of default can result in significant financial losses for the company.

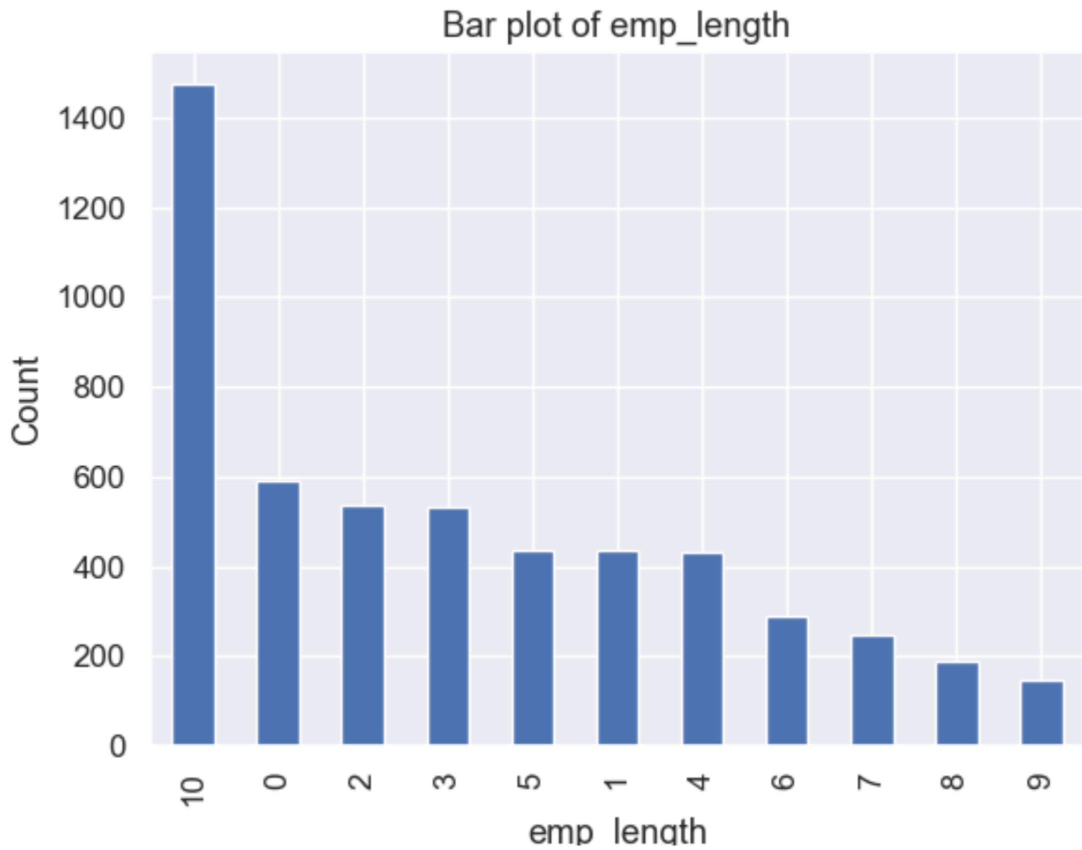
Approach:
Understand dataset and identify key driver variables to either approve, reject or provide loan with higher percentage for risky customers.
Steps:
Load dataset
Clean data by removing columns with null value, columns that have just one or two values which doesn’t help with analysis or those like id, member id or columns that are non-categorizable like name, title, url, desc.
Impute and remove rows wherever possible
Attend outliers
Perform univariate, bivariate and multivariate analysis

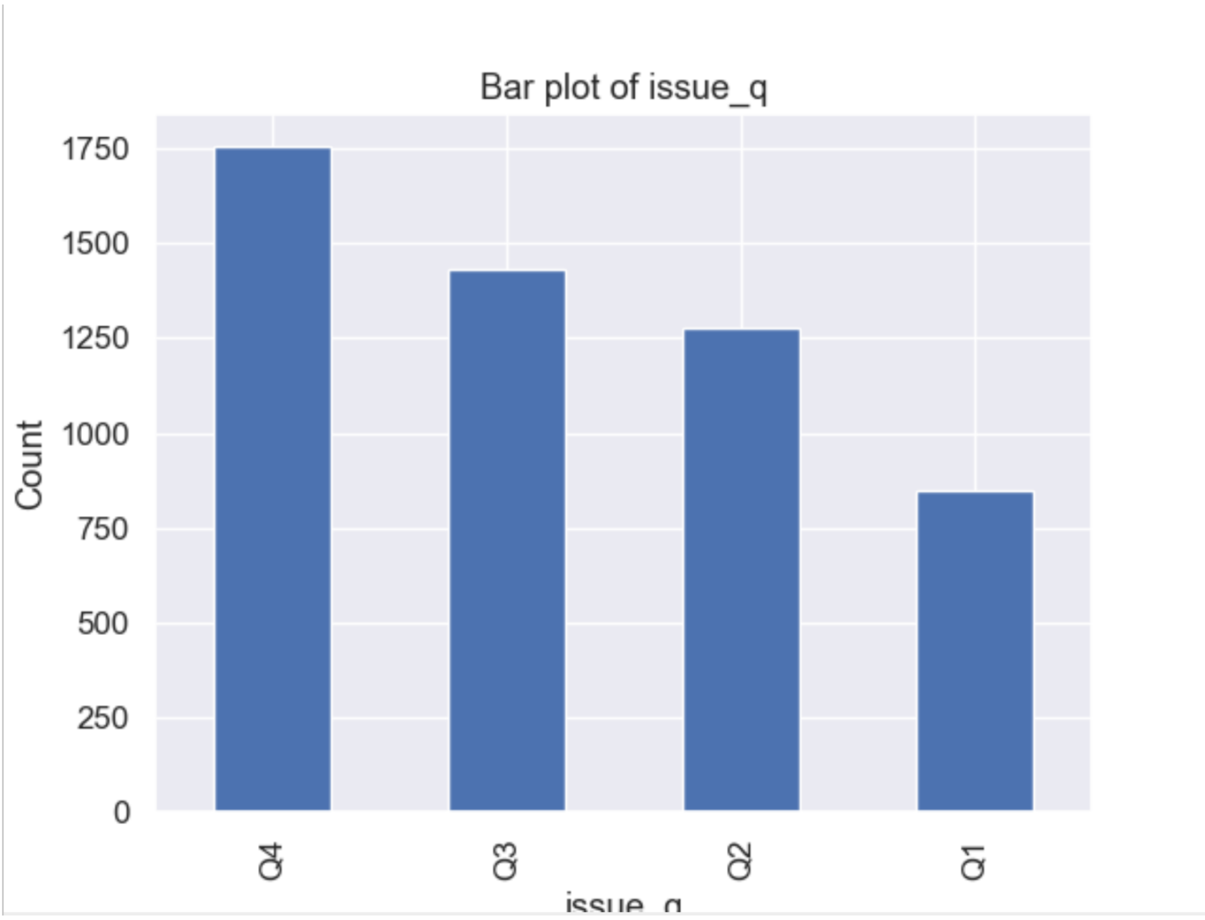
Univariate results – Ordered Categorical Variables:


Bar plot of grade
1400
1200
1000
800
600
400
200
0
Count
B
C
D
E
A
F
G
grade

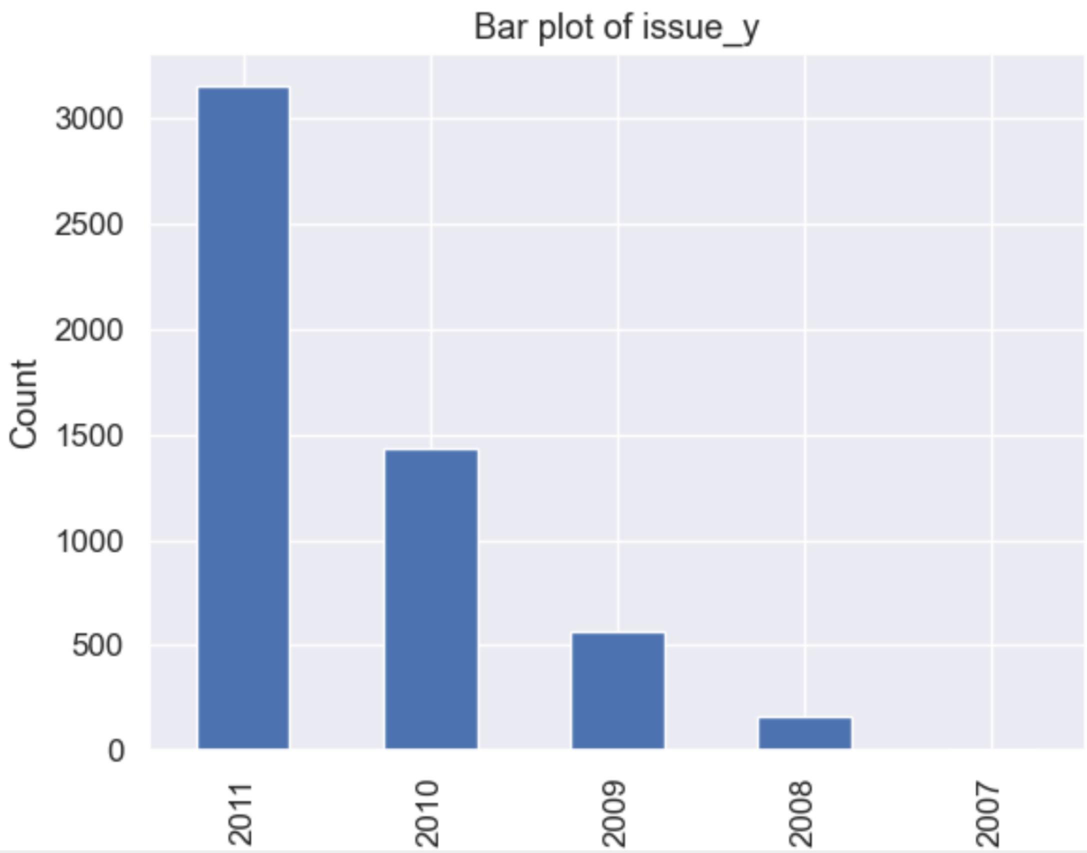

Bar plot of sub_grade
350
300
250
200
150
100
50
0
Count
B5 B3 C1 B4 C2 C3 D2 D3 A5 B2 C4 D4 C5 D5 E1 A4 B1 D1 E2 E4 E3 A3 E5 F1 A2 F2 F5 F4 F3 G1 A1 G2 G3 G4 G5
sub_grade


Bar plot of term
Count
3000
2500
2000
1500
1000
500
0
36
60
term


Bar plot of emp_length
Count
1400
1200
1000
800
600
400
200
0
10
0
2
3
5
1
4
6
7
8
9
emp_length


Bar plot of issue_q
Count
1750
1500
1250
1000
750
500
250
0
Q4
Q3
Q2
Q1


Bar plot of issue_m
Count
600
500
400
300
200
100
0
12
11
10
9
6
7
5
8
4
3
1
2
issue_m



Inference:
Grade B applicants topped the “Charged Off” list hinting at repayment difficulties for this group.

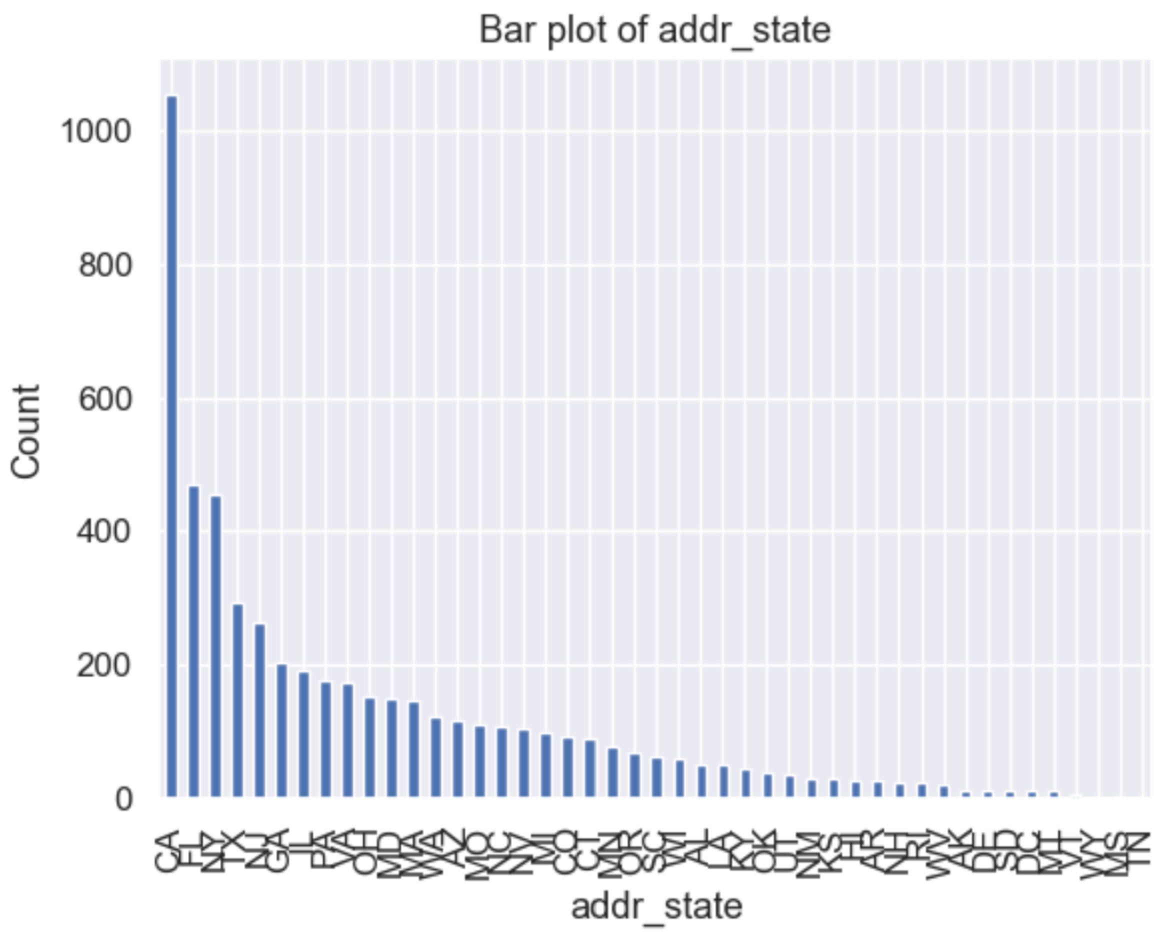
Short-term loans (36 months) were favoured by “Charged Off” applicants suggesting defaults were higher among those opting for shorter repayment periods.

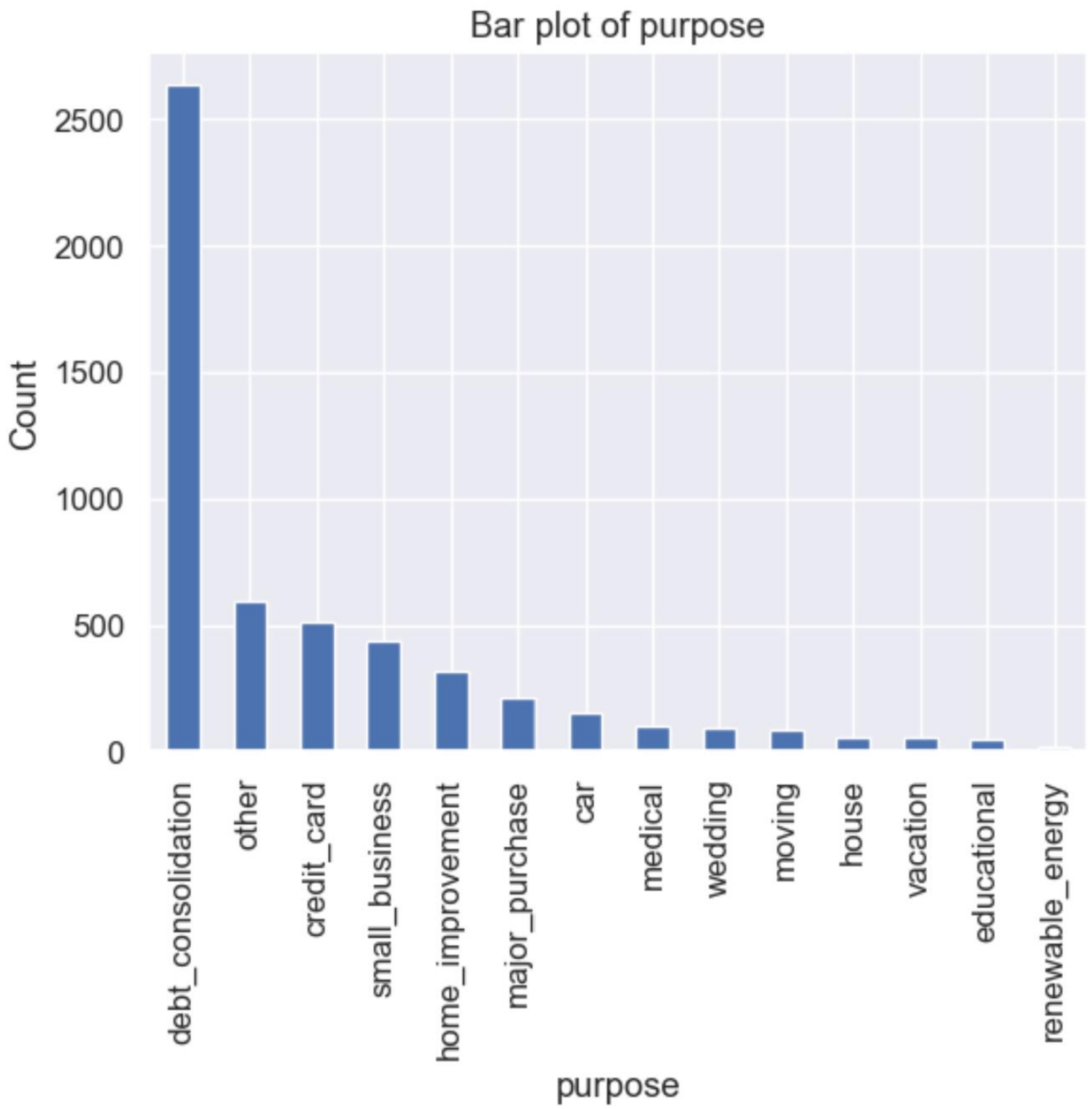
Applicants with over 10 years of employment had the most “Charged Off” applications indicating that a lengthy employment history didn’t necessarily equate to successful loan repayment.

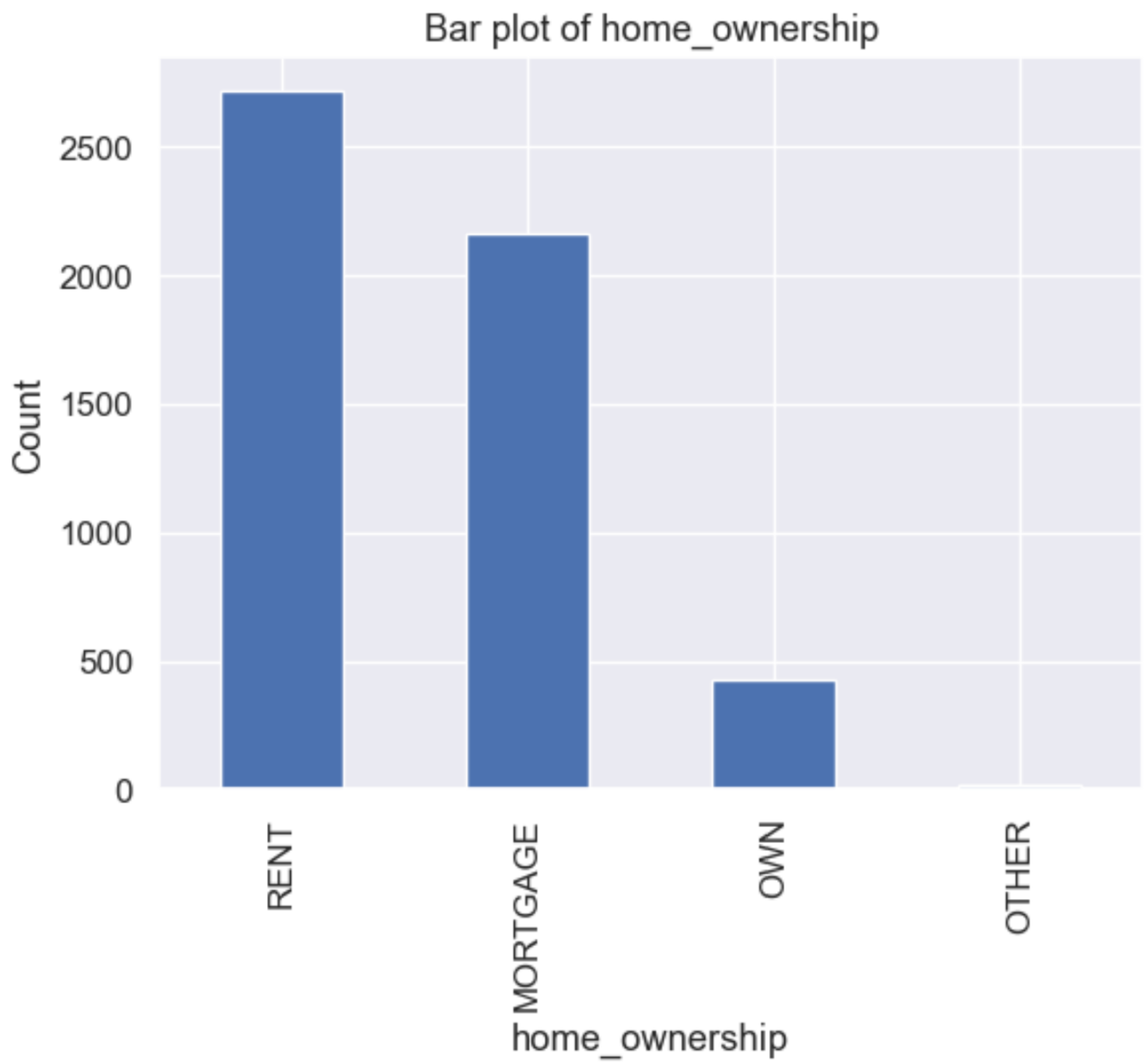
The year 2011 saw the highest number of “Charged Off” applications possibly reflecting economic or financial hardships during that period.

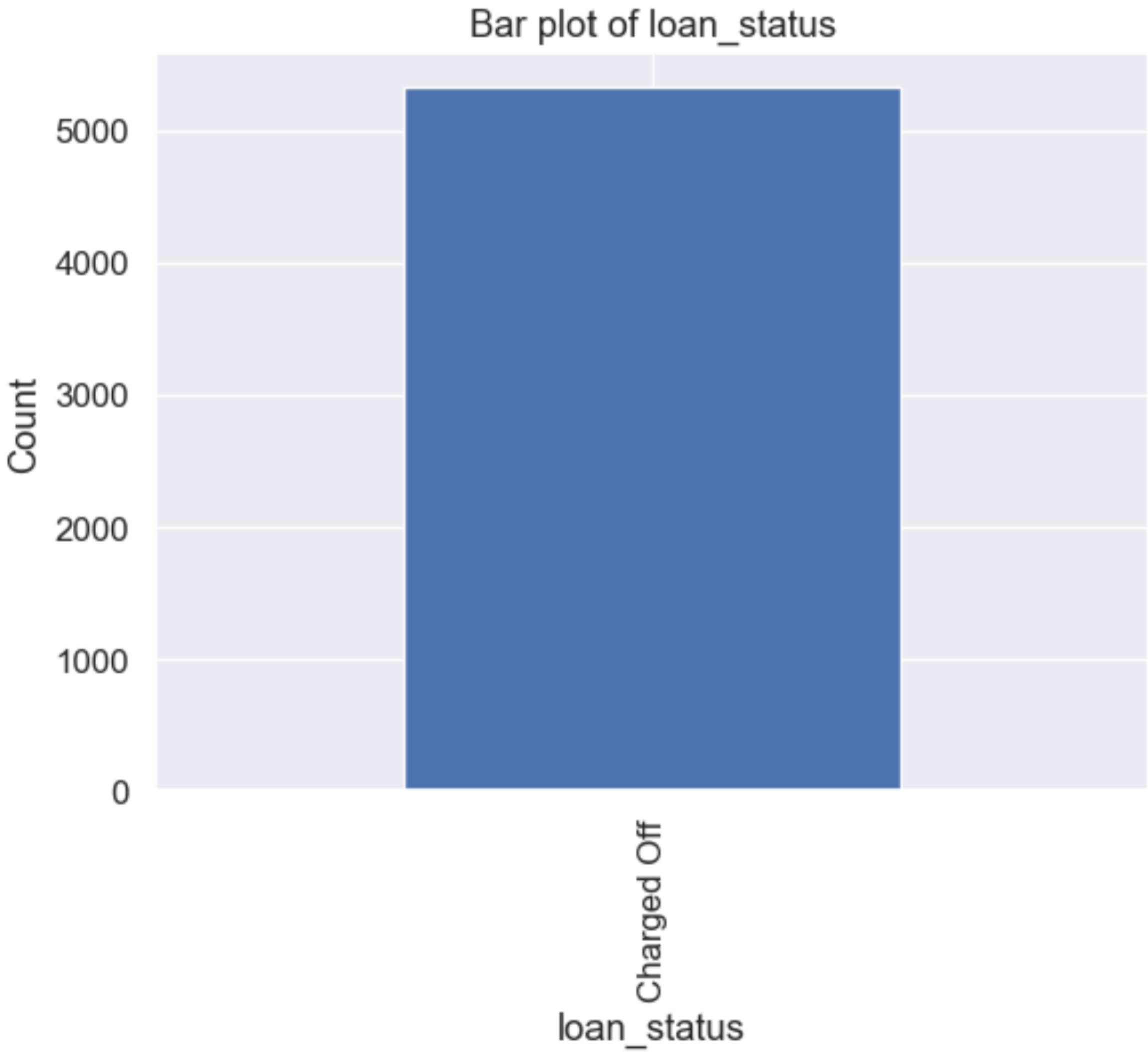
The 4th quarter, particularly December, witnessed a surge in “Charged Off” loans implying that holiday season financial stress could be a factor in loan defaults.

Univariate results – UnOrdered Categorical Variables:


Bar plot of addr_state
Count
1000
800
600
400
200
0
addr_state


Bar plot of purpose
2500
2000
1500
1000
500
0
Count
debt_consolidation
other
credit_card
small_business
home_improvement
major_purchase
car
medical
wedding
moving
house
vacation
educational
renewable_energy
purpose


Bar plot of home_ownership
Count
2500
2000
1500
1000
500
0
RENT
MORTGAGE
OWN
OTHER
home_ownership



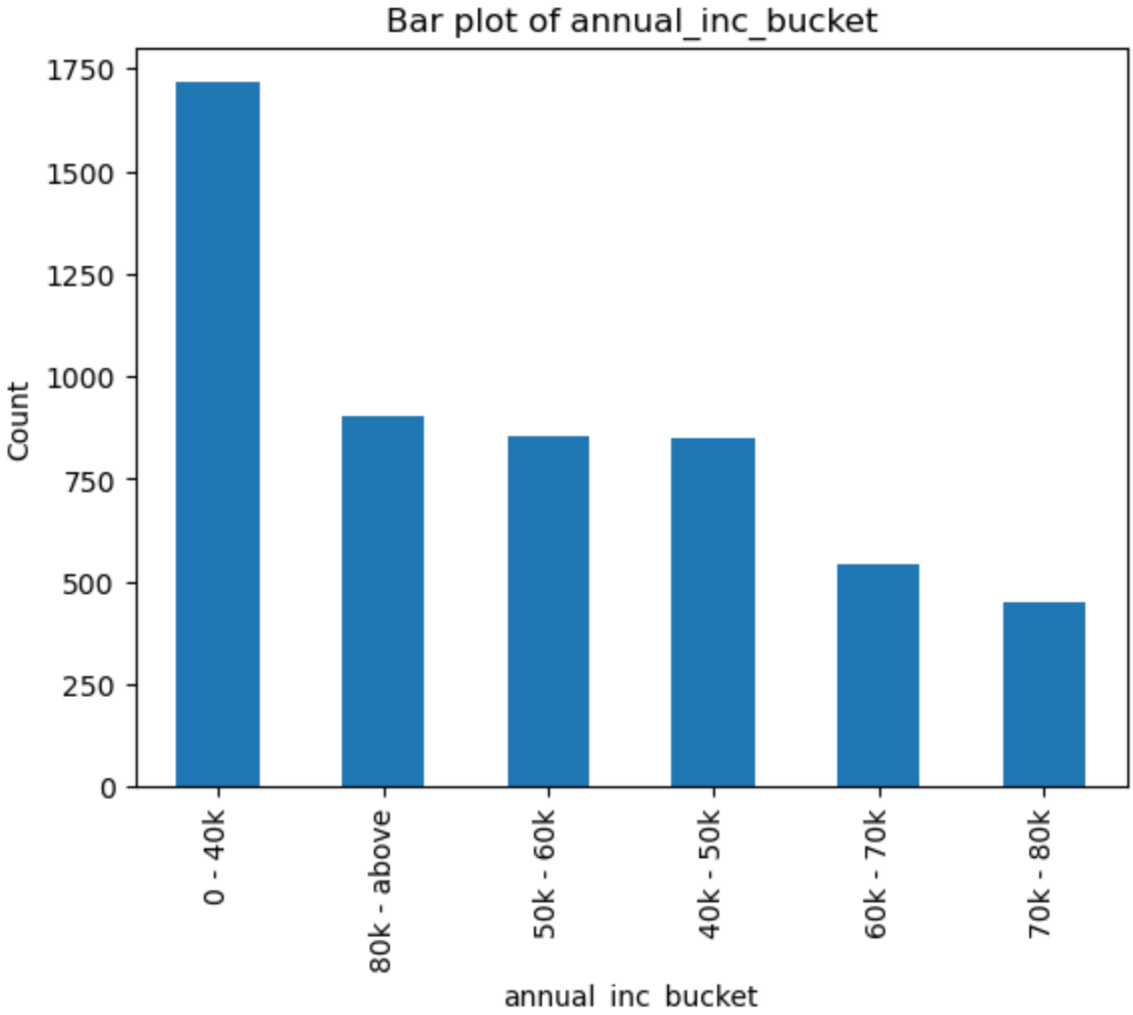
Inference:
California had the most "Charged Off" applicants indicating a need for stricter credit assessments.

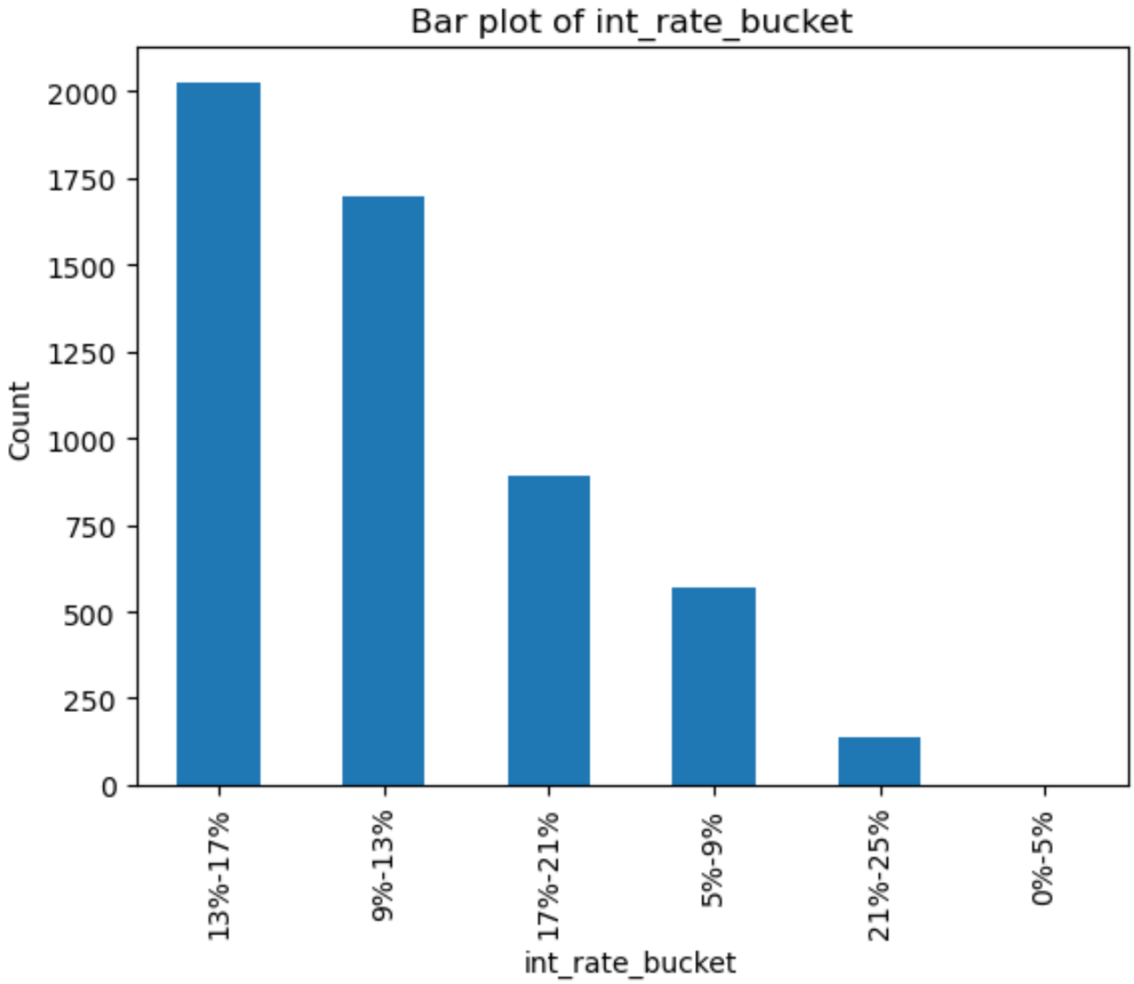
Debt consolidation was the main reason for most "Charged Off" loans suggesting caution for such loans.

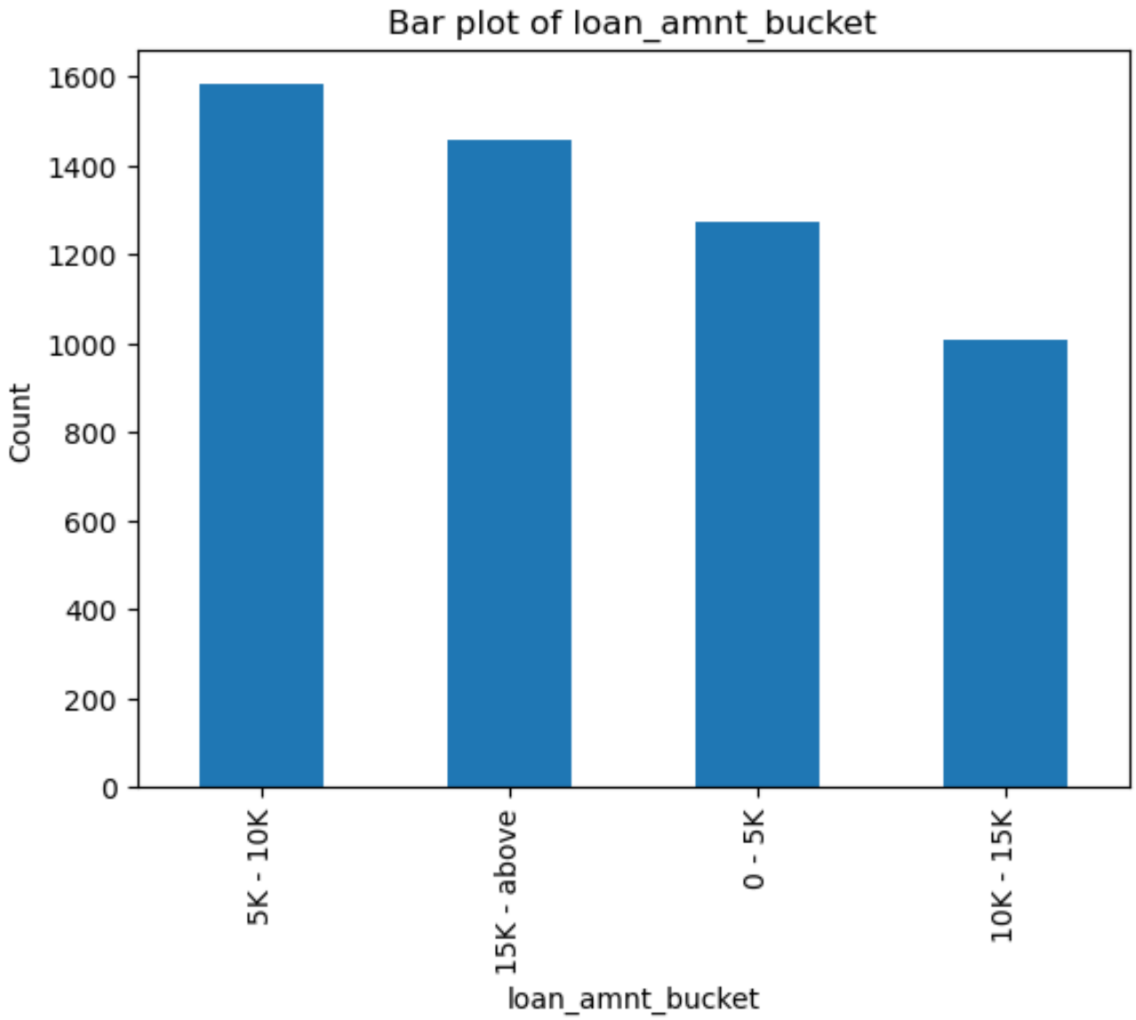
Most "Charged Off" applicants lived in rented, mortgagaed houses implying financial instability.

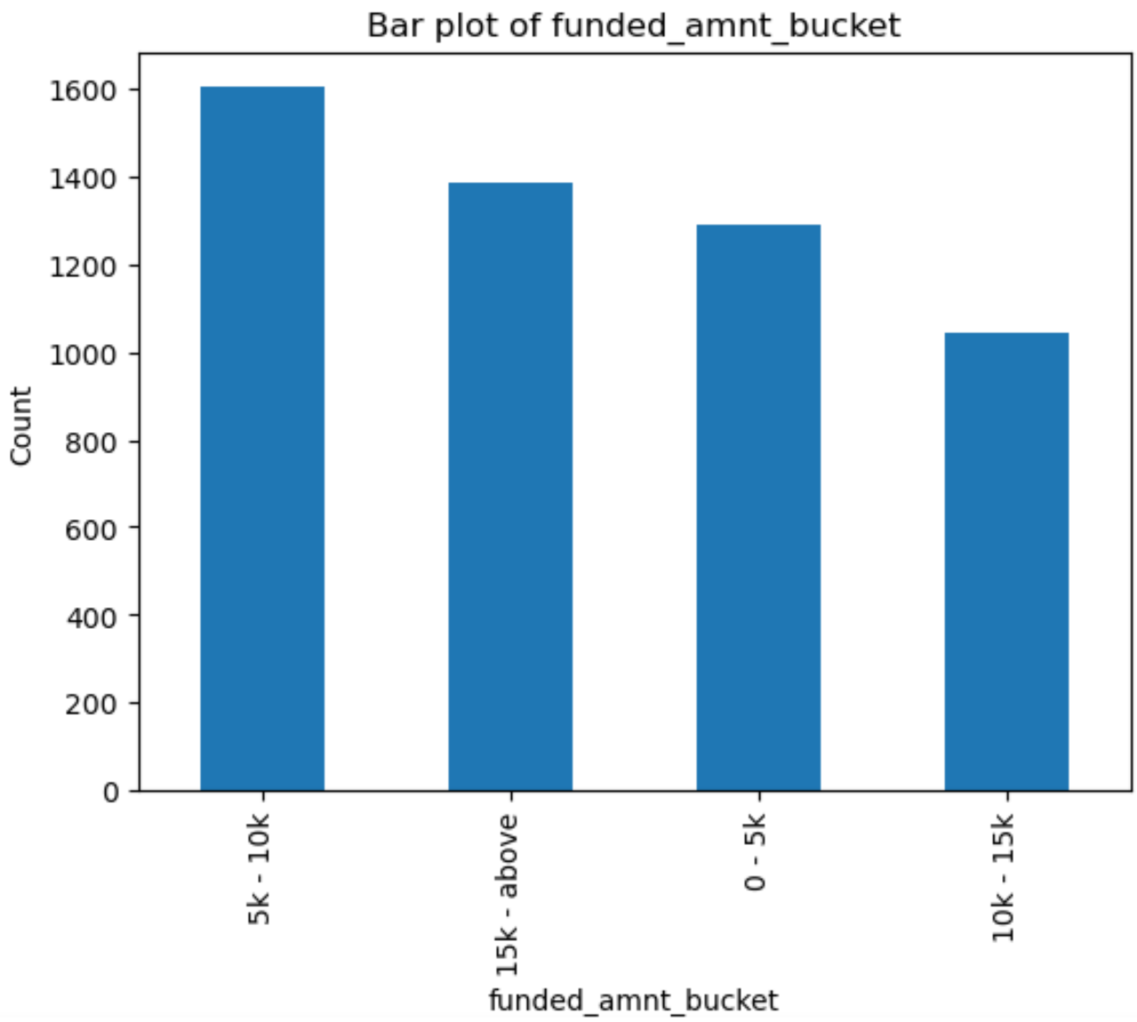
Many applicants were loan defaulters highlighting the need for enhanced risk assessment and financial education.

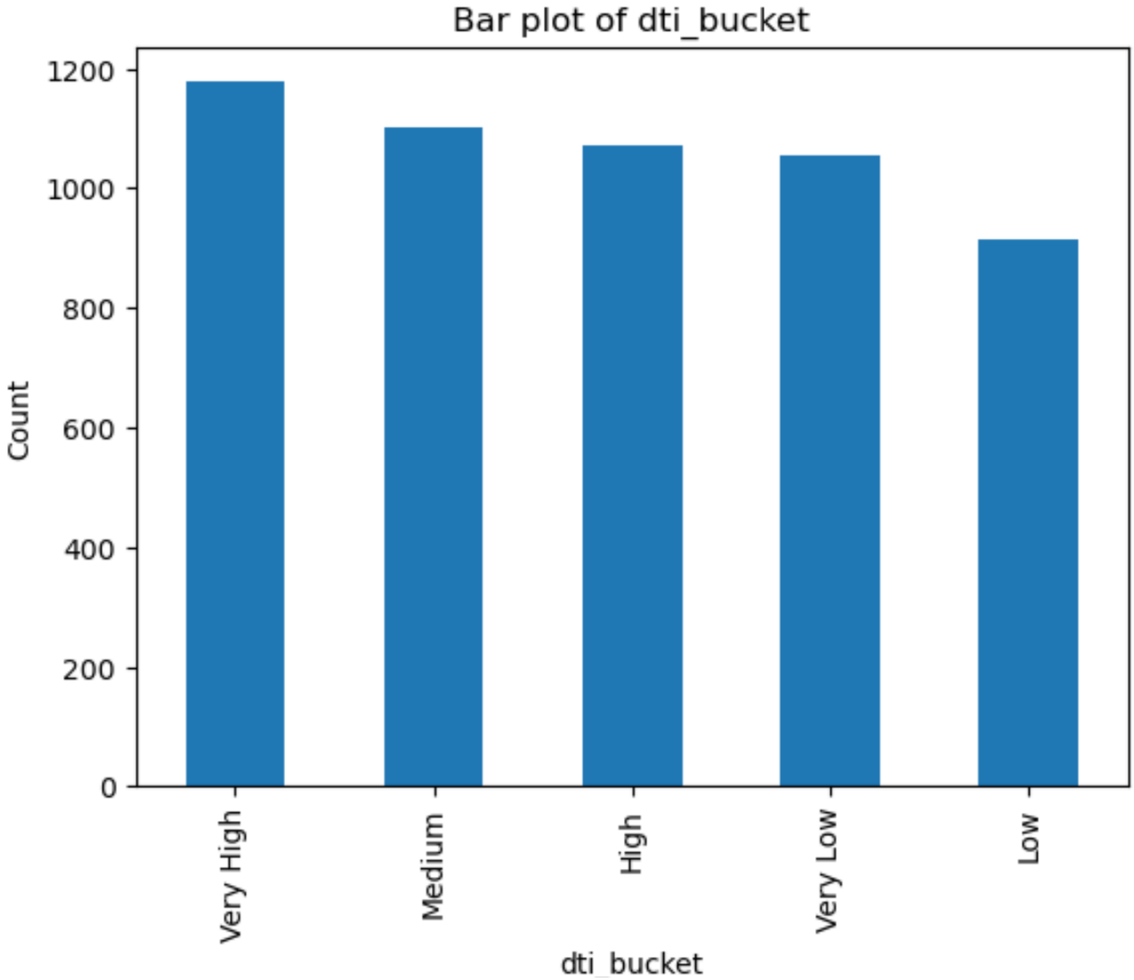
Univariate Analysis - Quantitative Variables




Bar plot of int_rate_bucket
Count
2000
1750
1500
1250
1000
750
500
250
0
13%-17%
9%-13%
17%-21%
5%-9%
21%-25%
0%-5%
int_rate_bucket


Bar plot of loan_amnt_bucket
Count
1600
1400
1200
1000
800
600
400
200
0
5K - 10K
15K - above
0 - 5K
10K - 15K
loan_amnt_bucket


Bar plot of funded_amnt_bucket
Count
1600
1400
1200
1000
800
600
400
200
0
5k - 10k
15k - above
0 - 5k
10k - 15k
funded_amnt_bucket



Inference:

"Charged Off" applicants earned less than **$40,000** annually. Lenders should enforce stringent income checks and repayment assessments for this group.

A significant portion of "Charged Off" applicants fell into the **13%-17%** interest rate bracket. Lenders could consider lower interest rates to reduce default risk.

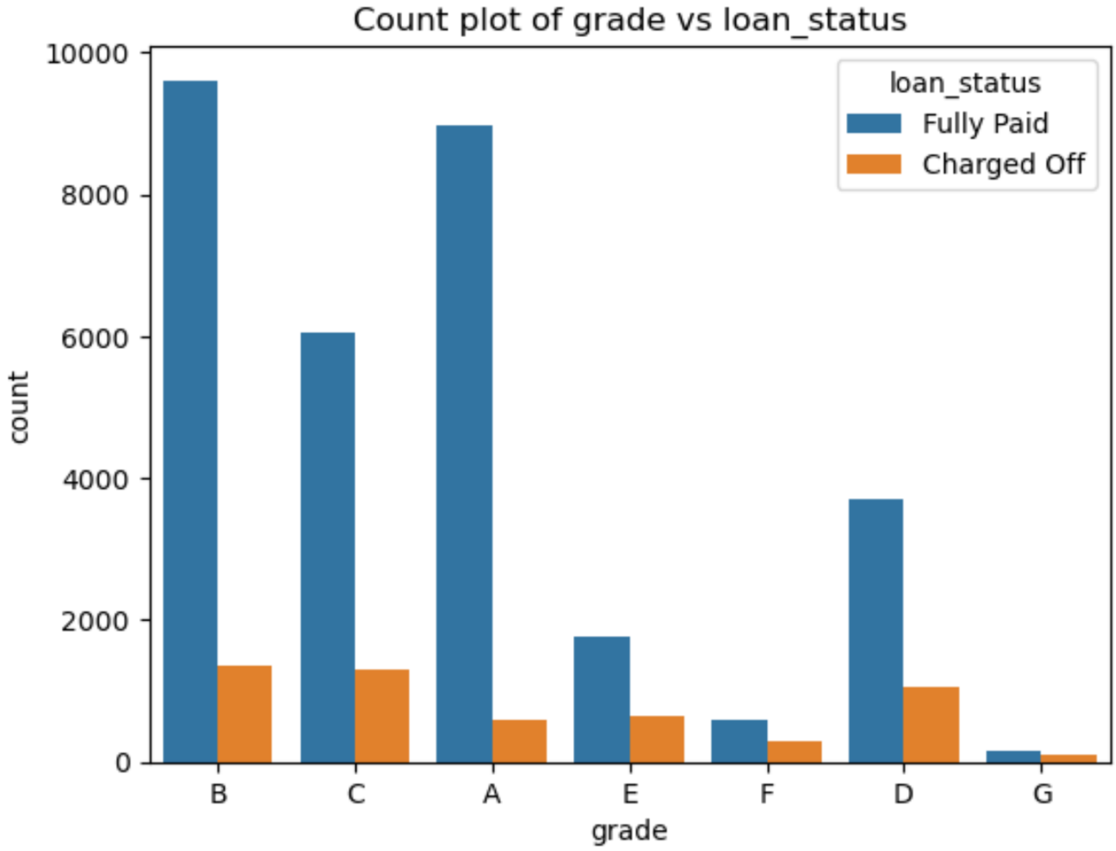
"Charged Off" applicants received loans of **$15,000** or more. Lenders should carefully evaluate applicants seeking larger loans, ensuring strong credit history and repayment capability.

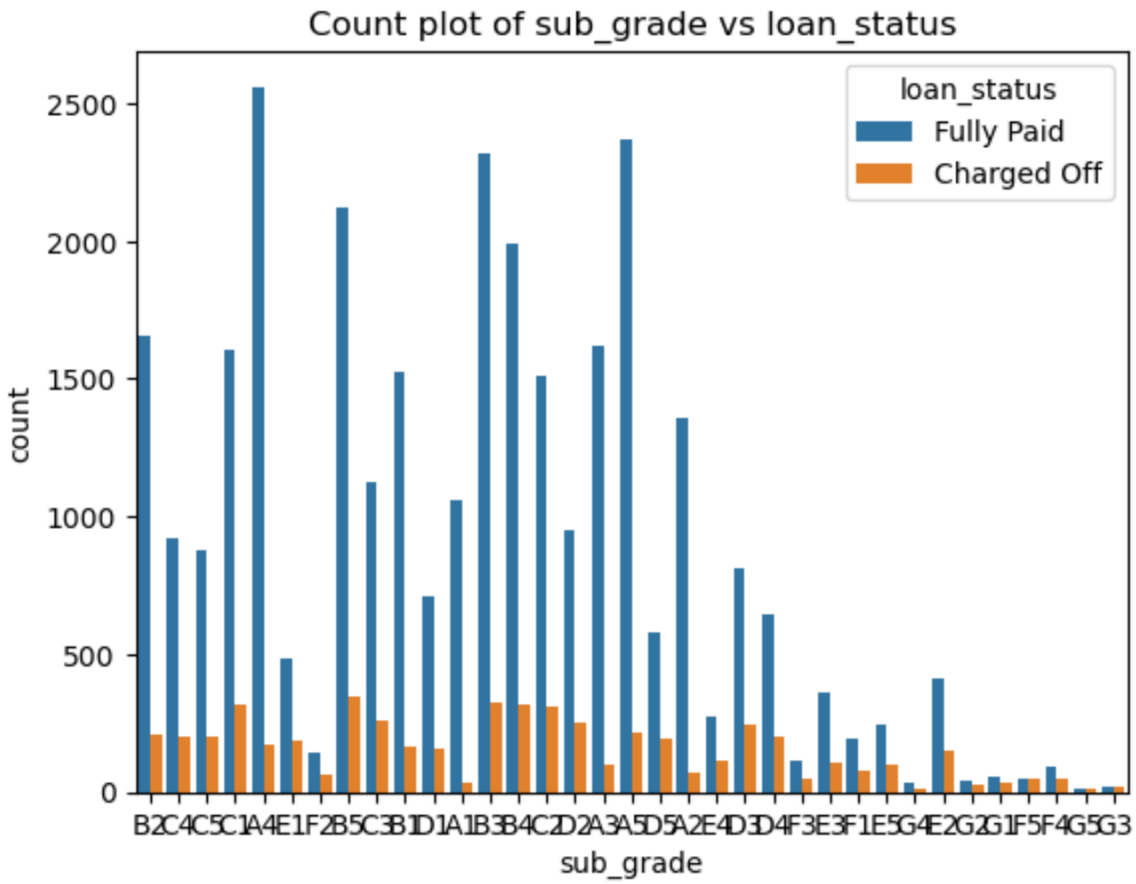
"Charged Off" applicants received funded amounts of **$15,000** or more. Lenders should align funded amounts with borrowers' financial capacity and conduct thorough credit assessments for larger loan requests.

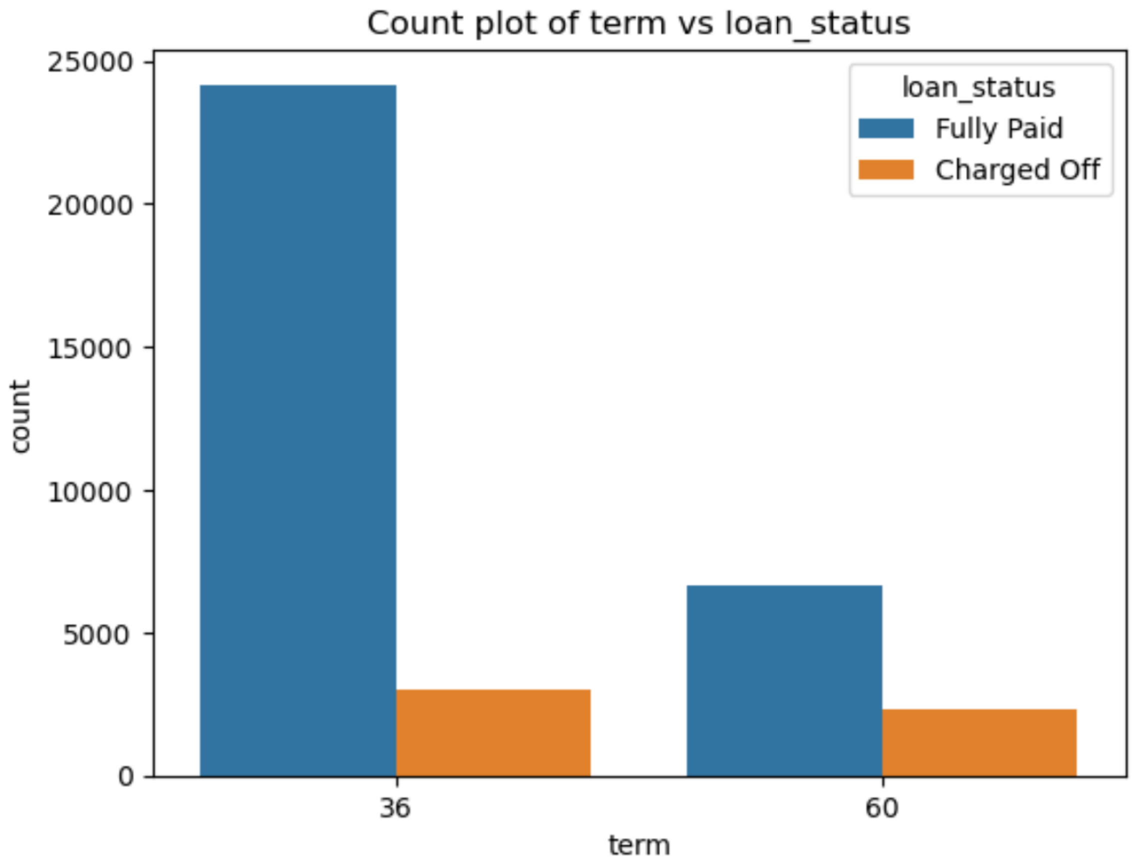
"Charged Off" applicants had high debt-to-income ratios. Lenders should enforce strict debt-to-income ratio requirements to avoid lending to individuals with unsustainable debt levels.

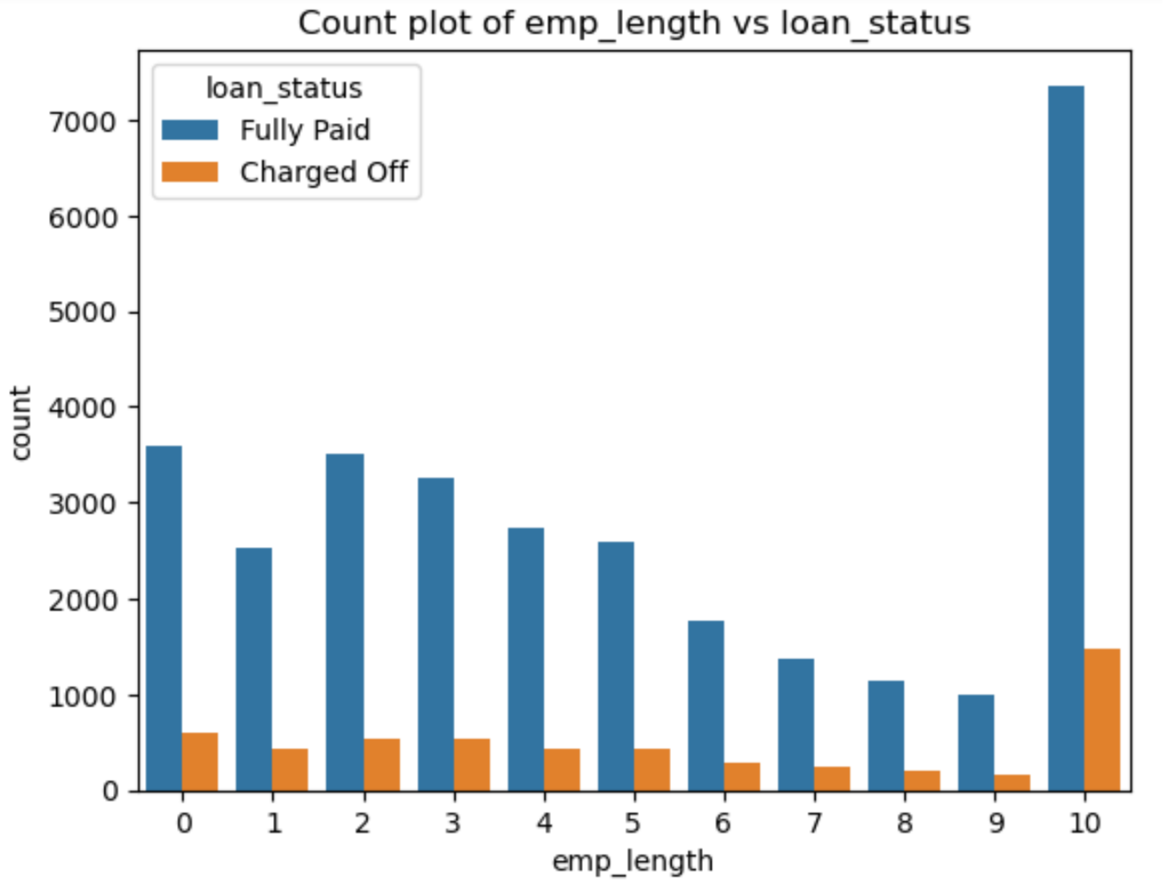
Most "Charged Off" applicants had monthly installments between **$160-$440**. Lenders should closely monitor and assess applicants with similar installment amounts to mitigate loan default risk.

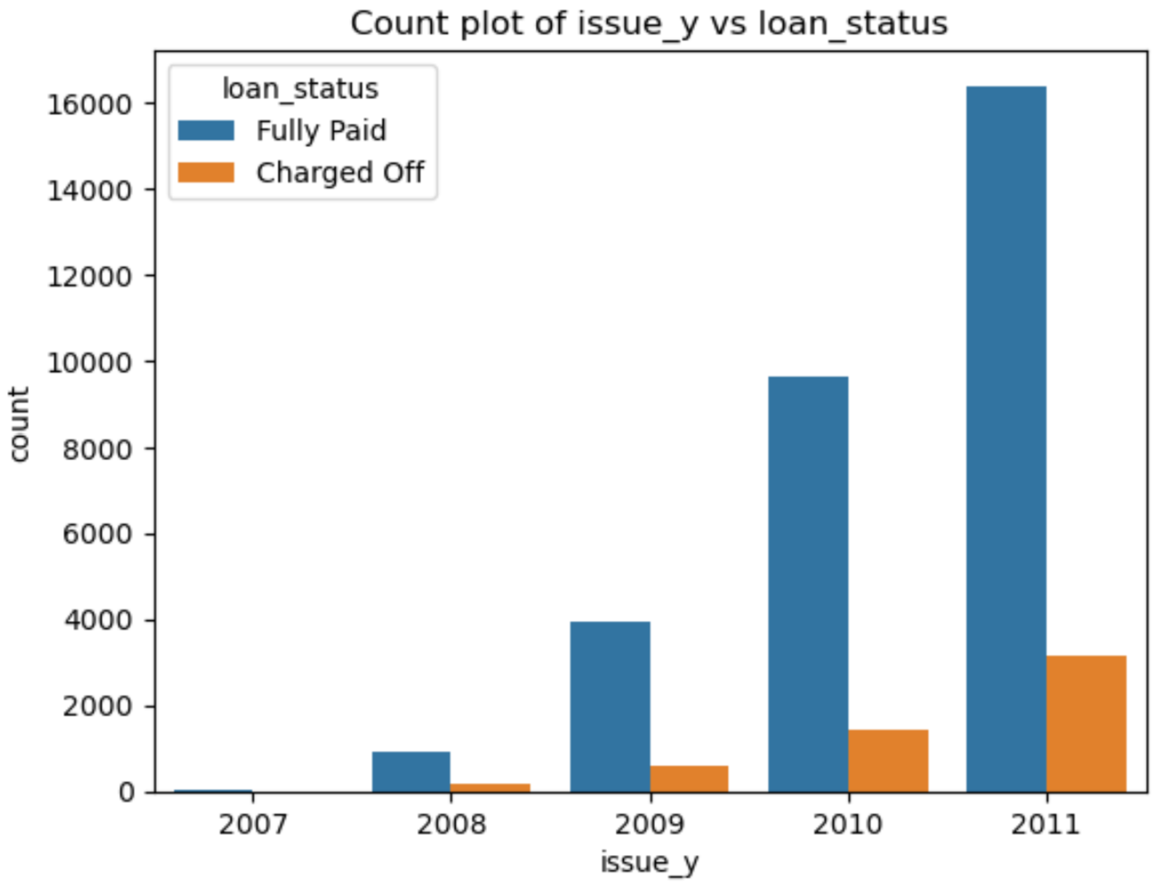
Bivariate Ordered categorical data

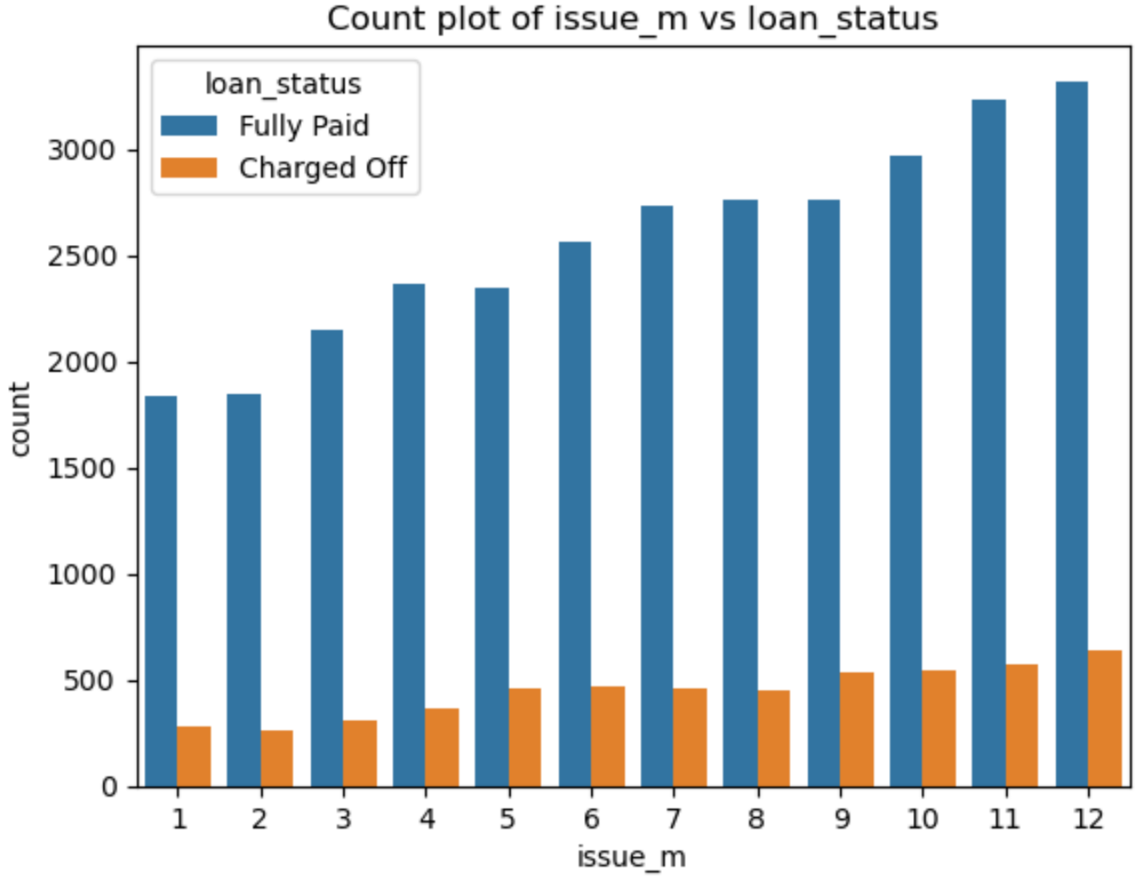


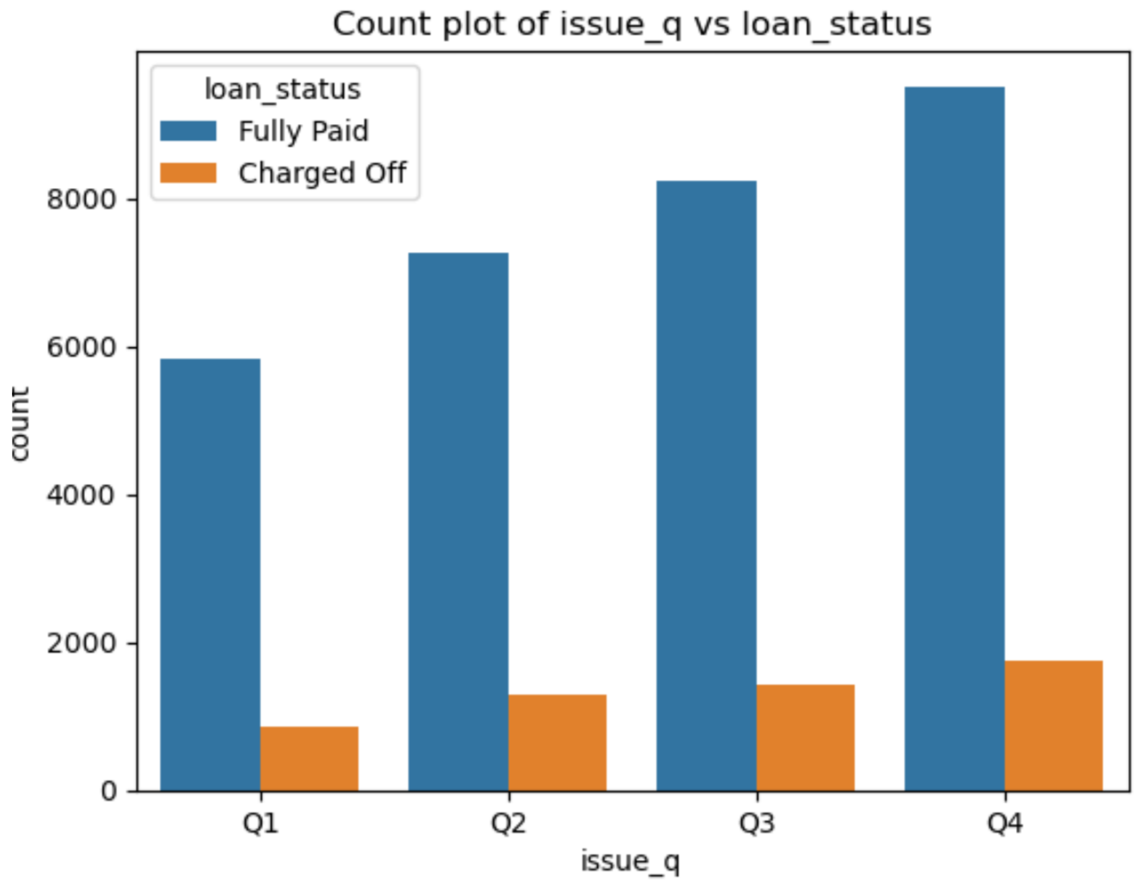

Count plot of sub_grade vs loan_status
loan_status
Fully Paid
Charged Off
count
0
500
1000
1500
2000
2500
B2 C4 C5 C1 A4 E1 F2 B5 C3 B1 D1 A1 B3 B4 C2 D2 A3 A5 D5 A2 E4 D3 D4 F3 E3 F1 E5 G4 E2 G2 G1 F5 F4 G5 G3
sub_grade


Count plot of term vs loan_status
loan_status
Fully Paid
Charged Off
count
0
5000
10000
15000
20000
25000
36
60
term

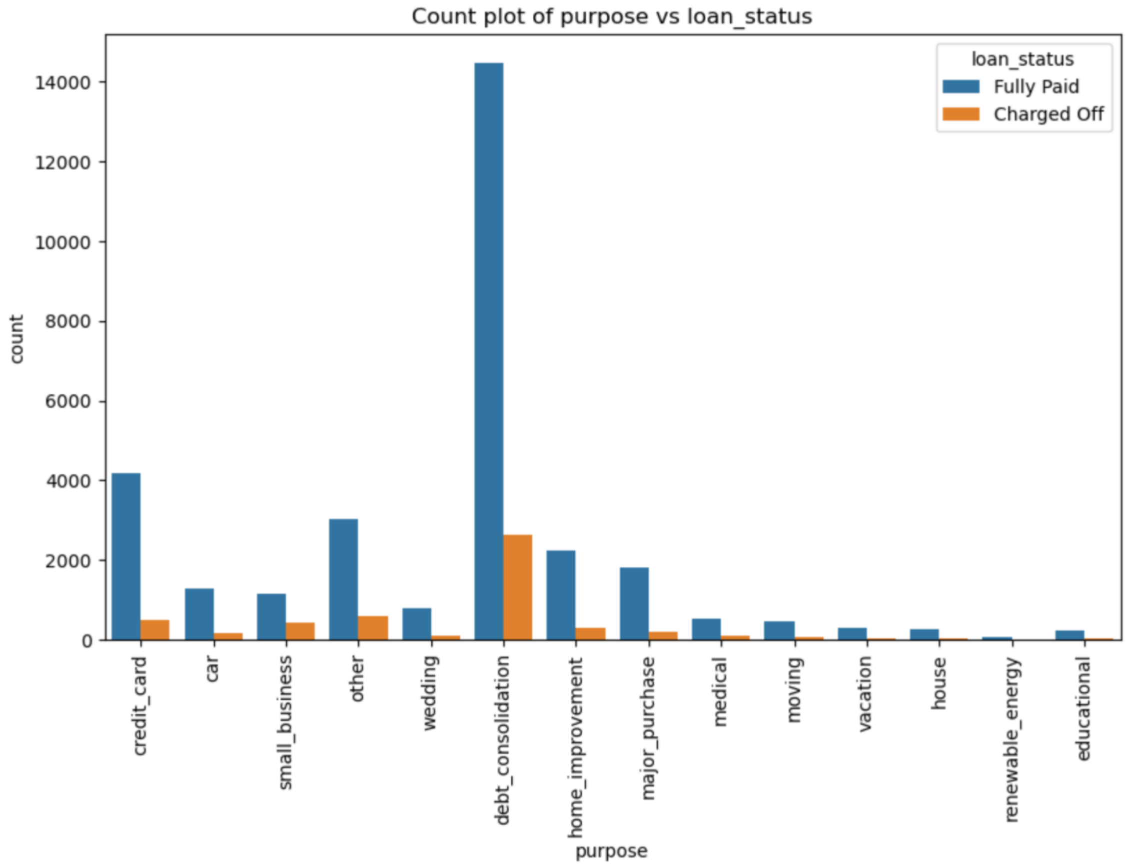

Count plot of emp_length vs loan_status
loan_status
Fully Paid
Charged Off
count
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
9
10
emp_length

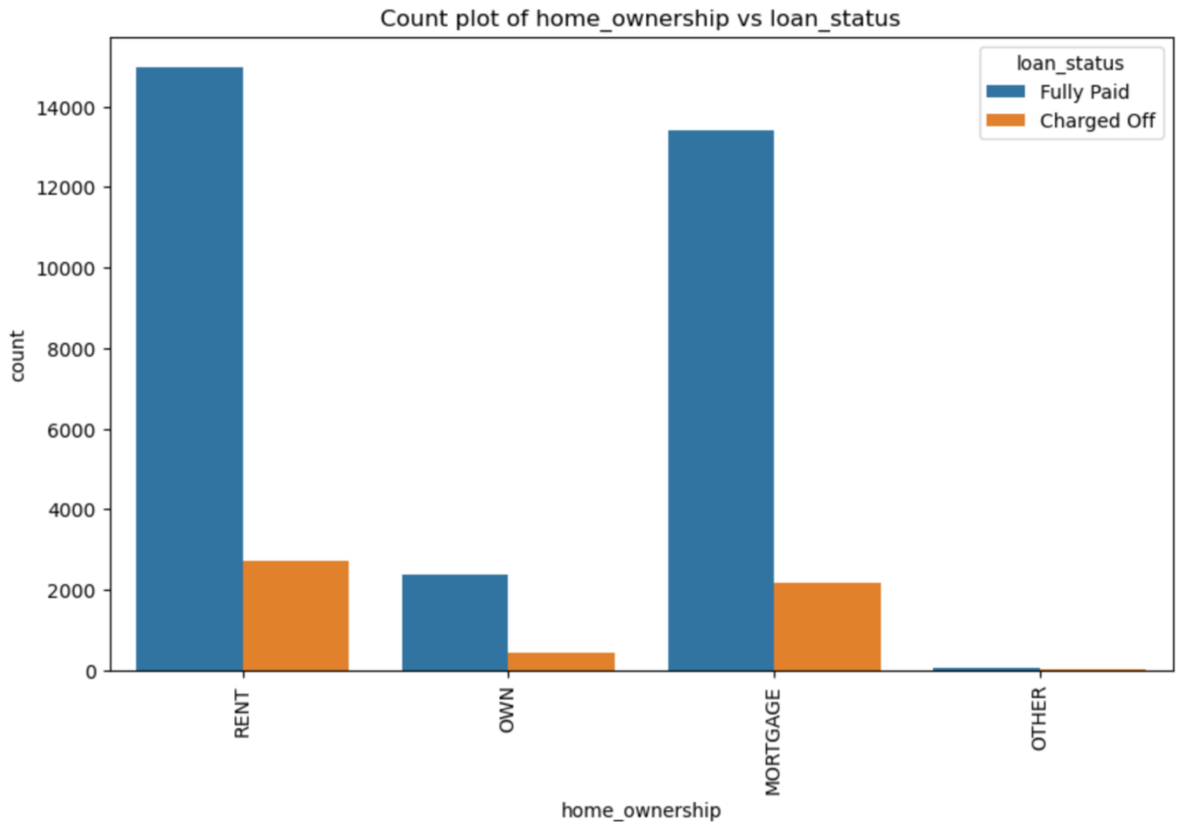

Count plot of issue_y vs loan_status
loan_status
Fully Paid
Charged Off
count
16000
14000
12000
10000
8000
6000
4000
2000
0
2007
2008
2009
2010
2011
issue_y

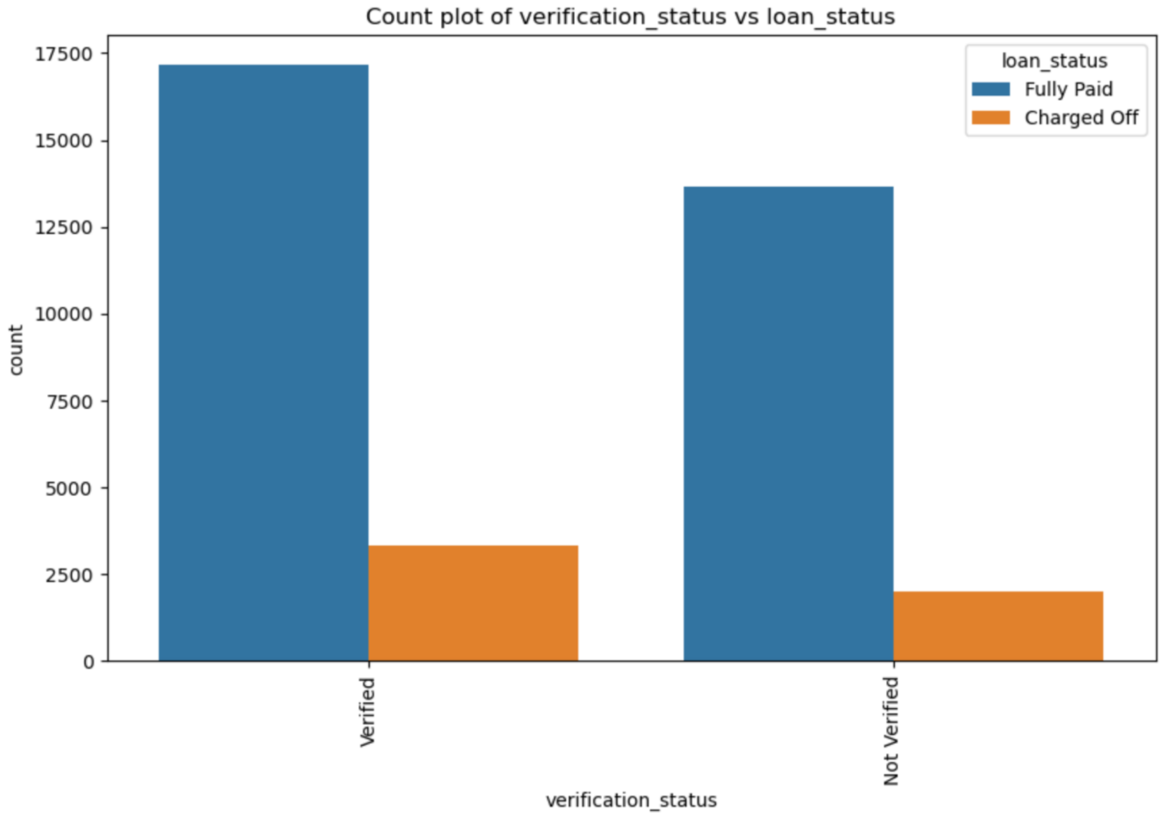

Count plot of issue_m vs loan_status
loan_status
Fully Paid
Charged Off
count
0
500
1000
1500
2000
2500
3000
1
2
3
4
5
6
7
8
9
10
11
12
issue_m

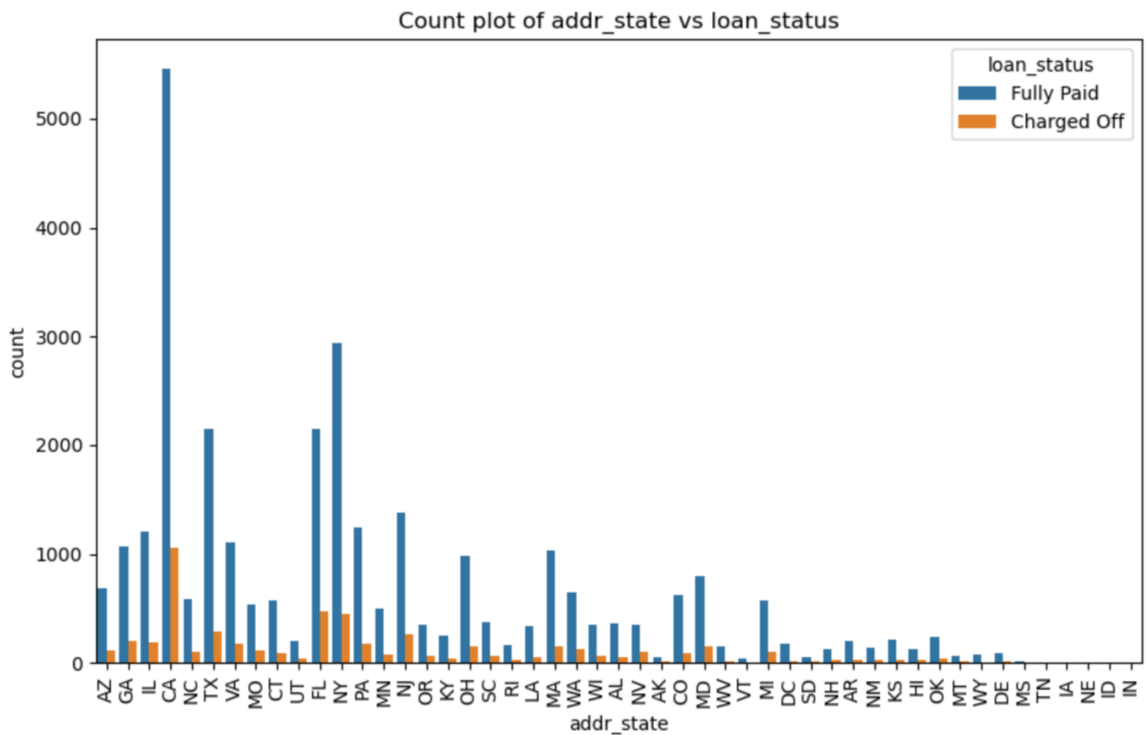

Count plot of issue_q vs loan_status
loan_status
Fully Paid
Charged Off
count
0
2000
4000
6000
8000
Q1
Q2
Q3
Q4
issue_q

Bivariate UnOrdered categorical data









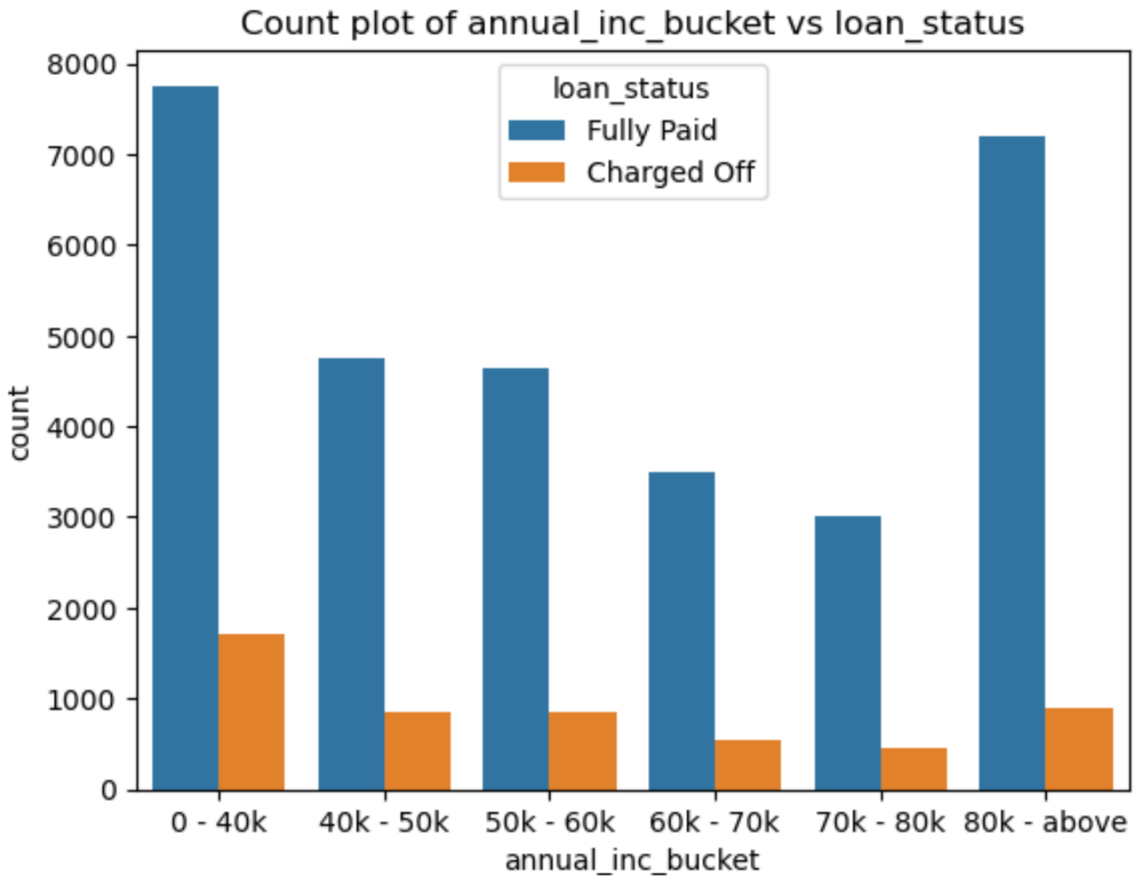
Inference:

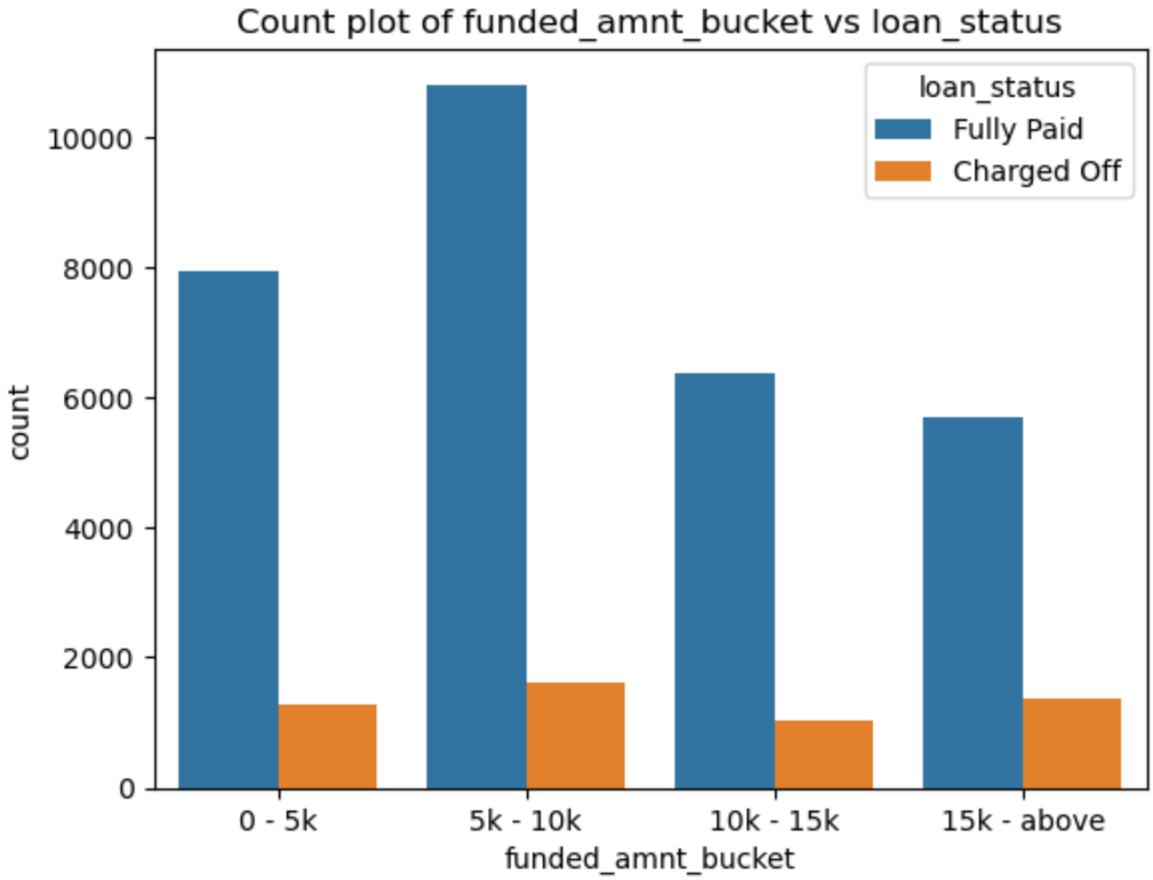
Risk Assessment: Stricter criteria should be applied to loan applicants from Grades B, C, and D due to their high charge-off rates.

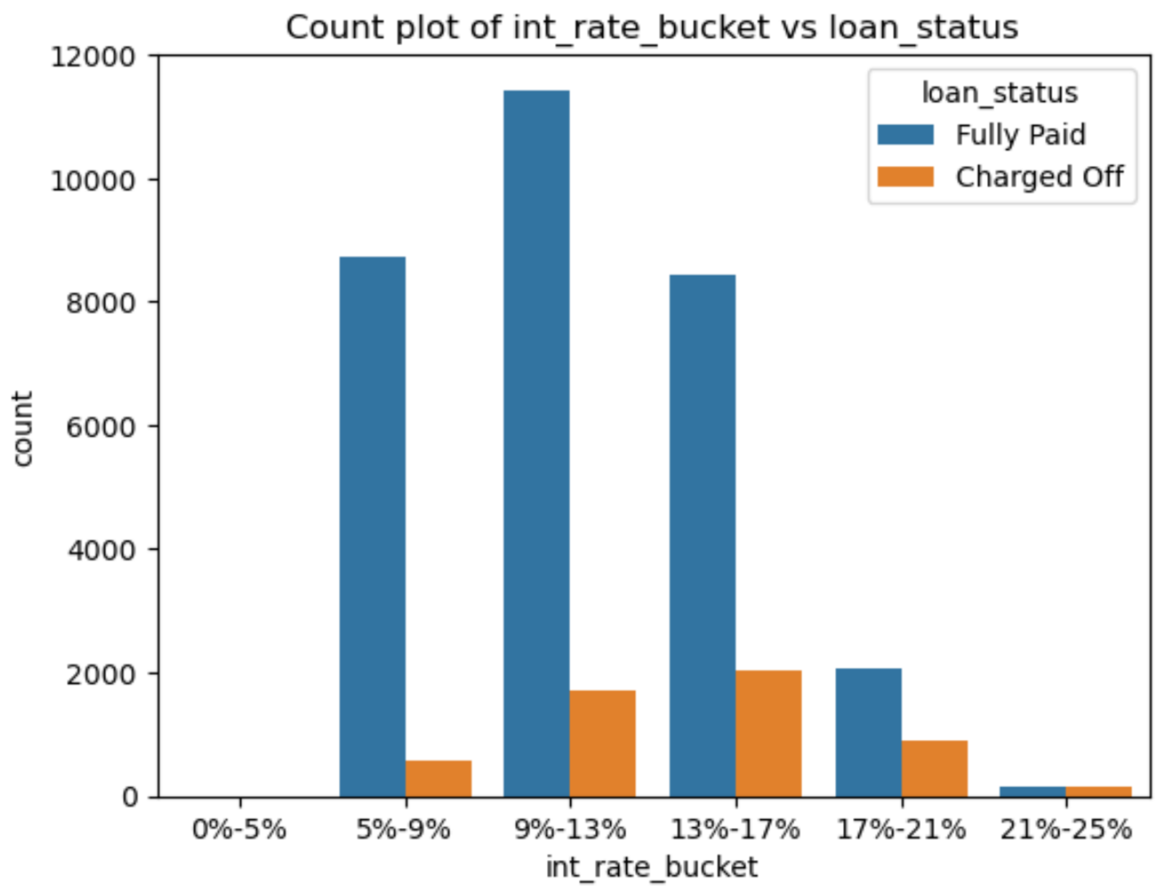
- **Subgrades B3, B4, and B5**: These applicants are more likely to charge off. Consider additional risk mitigation measures or lower loan amounts.

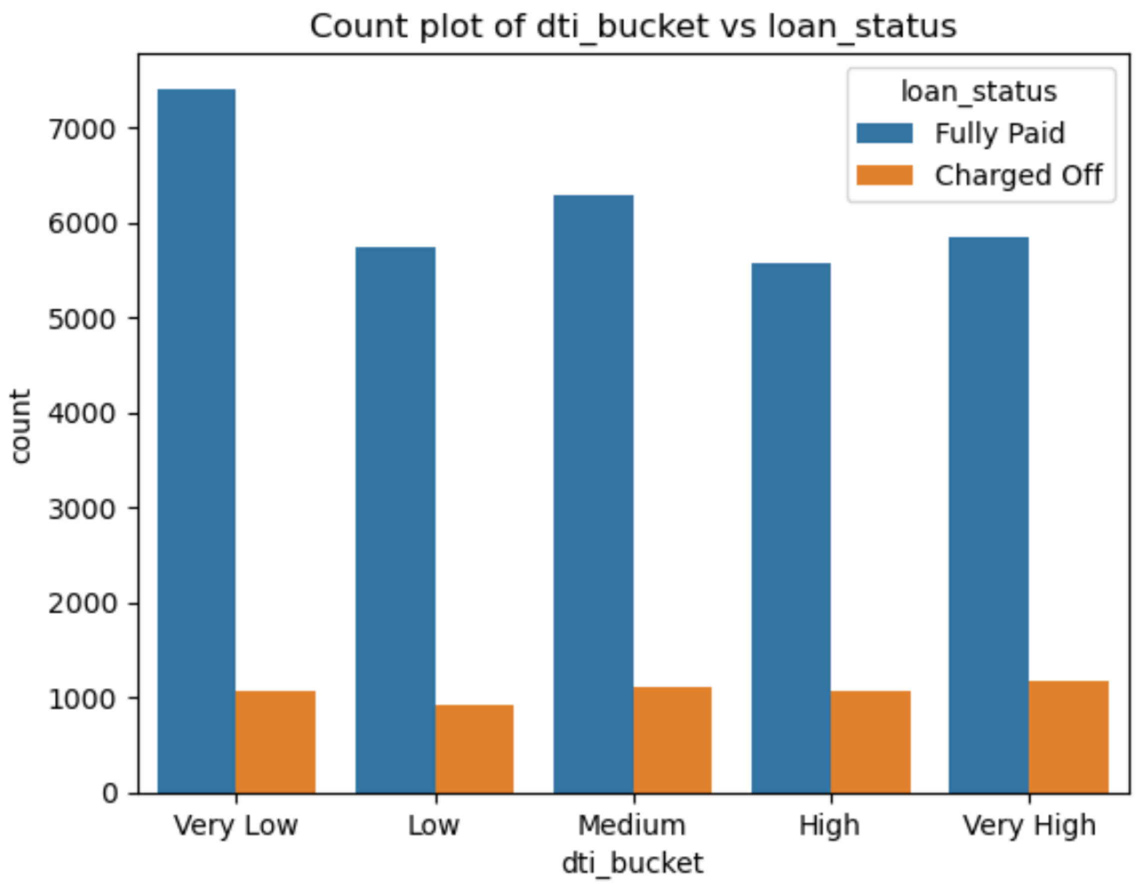
- **Term Length**: The risk of default is higher for 60-month loans. Consider limiting the maximum term or adjusting interest rates.
- **Experience**: Applicants with over ten years of experience have a higher default probability. A comprehensive credit scoring system should be used.
- **Growth Trend**: The increasing number of loan applicants from 2007 to 2011 indicates market growth. Maintain competitive edge and robust risk management.
- **Seasonal Trends**: December and Q4 see peak loan applications. Ensure efficient processing to meet increased demand.
- **Debt Consolidation Risk**: This category has the most loans and high default rates. Carefully evaluate these applicants and consider adjusting interest rates or offering counseling.
- **Housing Status**: Default risk is higher for applicants in rented or mortgaged houses. Consider this in the underwriting process.
- **Verification Process**: Verified applicants are defaulting more. Review the verification process for potential improvements.
- **Geographic Risk**: Default risk is higher in CA, FL, and NY. Monitor regional trends and adjust lending strategies accordingly.
- **High Loan Amounts**: Default risk is higher for loans of $15,000 or more. Consider thorough assessments or capping loan amounts for these applicants.
- **DTI and Interest Rates**: High DTI ratios and interest rates of 13%-17% are linked to defaults. Review the interest rate determination process and consider adjustments based on DTI.
- **Low Annual Income**: Default likelihood is higher for those earning less than $40,000 annually. Consider offering financial education or setting loan amounts based on income.

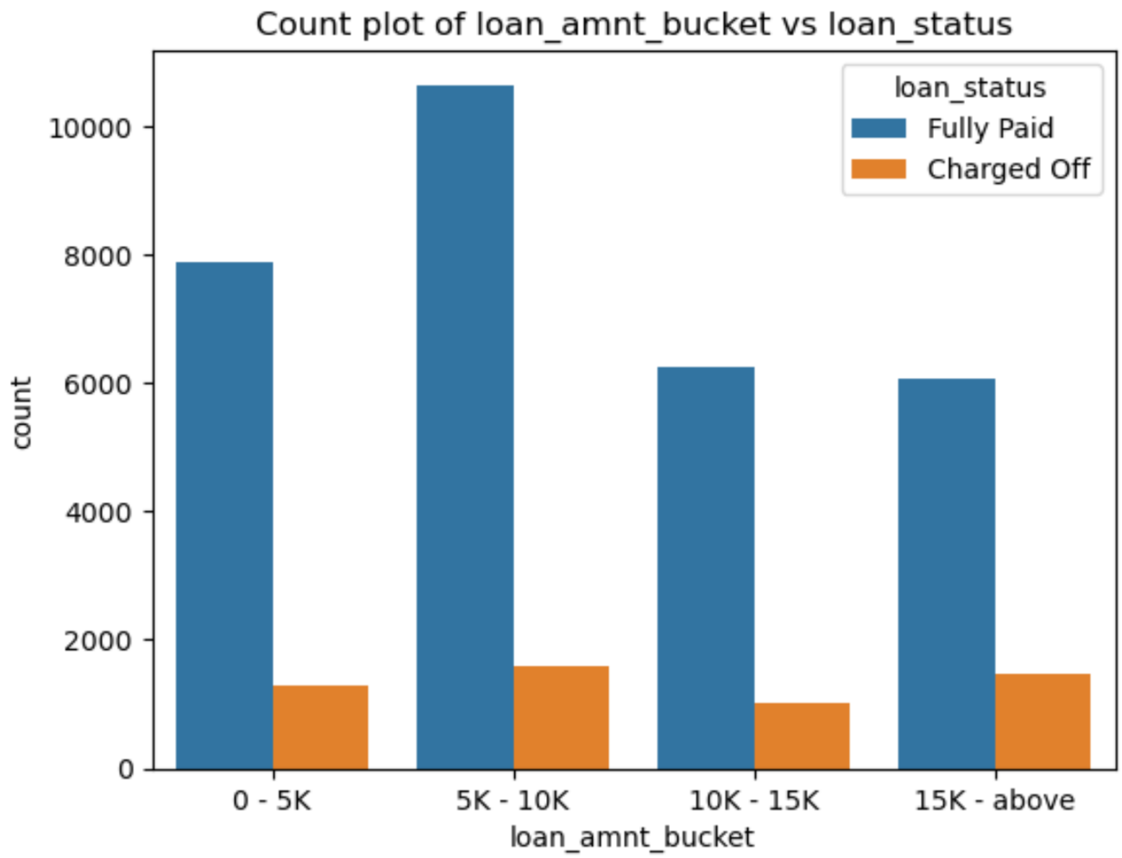
Bivariate quantitative variable






Count plot of int_rate_bucket vs loan_status
loan_status
Fully Paid
Charged Off
count
0
2000
4000
6000
8000
10000
12000
0%-5%
5%-9%
9%-13%
13%-17%
17%-21%
21%-25%
int_rate_bucket

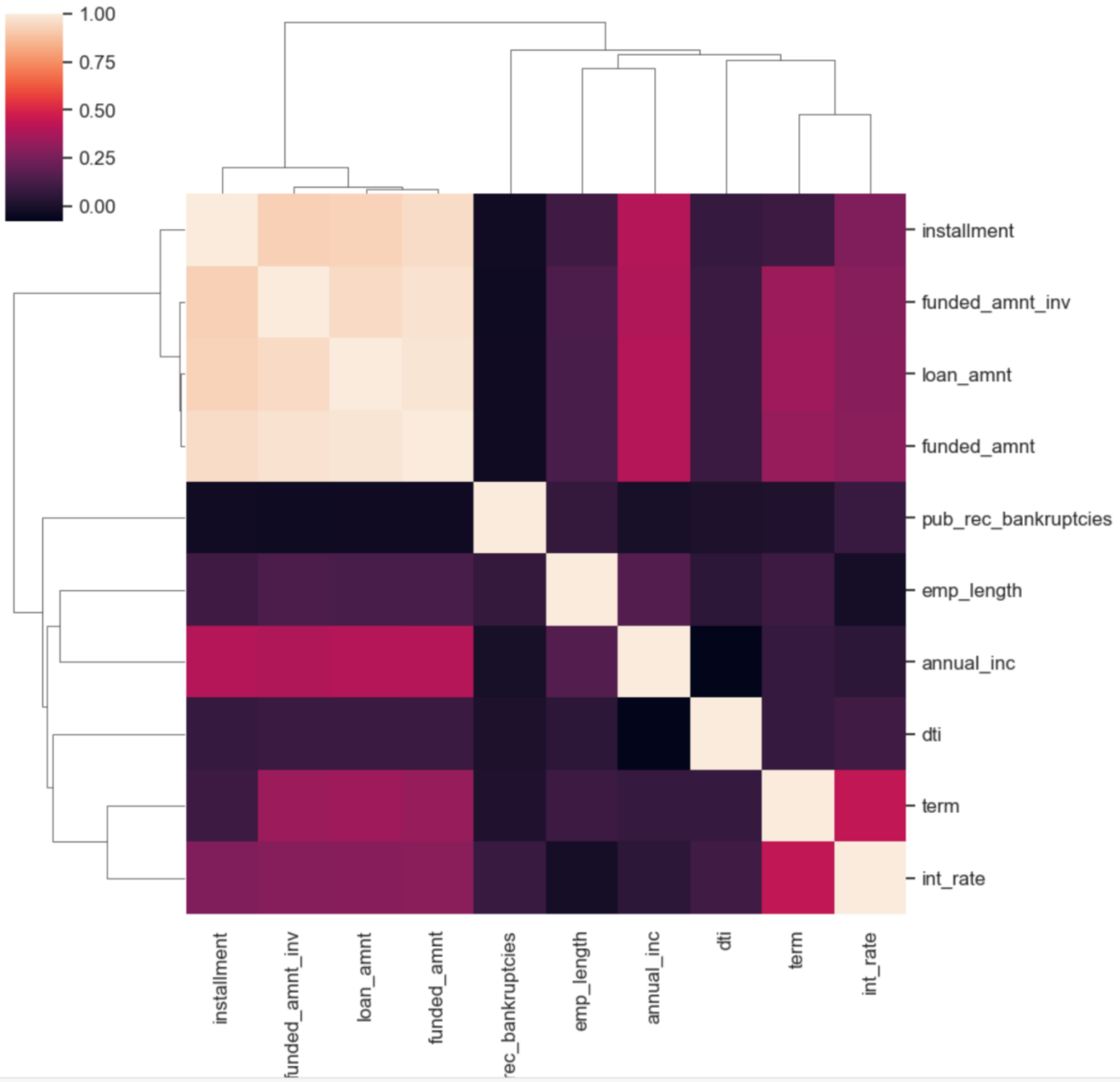

Count plot of dti_bucket vs loan_status
loan_status
Fully Paid
Charged Off
count
0
1000
2000
3000
4000
5000
6000
7000
Very Low
Low
Medium
High
Very High
dti_bucket



Inference:

- Most defaulters had loans of **$15,000 or more**.
- High Debt-to-Income (DTI) ratios were common among those who charged off.
- Defaulters often had interest rates between **13% and 17%**.
- Those who charged off typically reported an annual income **below $40,000**.

Multivariate:



Strong Correlation

- installment is strongly correlated with funded_amnt, loan_amnt, and funded_amnt_inv.
- term has a strong correlation with interest rate.
- annual_inc is strongly correlated with loan_amount.

Weak Correlation

- dti has weak correlations with most fields.
- emp_length also shows weak correlations with most fields.

Negative Correlation

- pub_rec_bankrupticies negatively correlates with almost all fields.
- annual_inc has a negative correlation with dti.