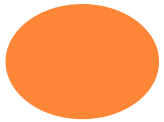
Lending Club Case Study

Submitted by:

- Aditya Kumar Jain
- Aditya Singh

Contents



Problem Statement

Problem:

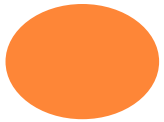
- You work for a consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Objective:

- Use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

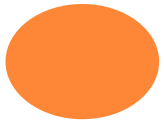
Constraints:

- 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 3 possible scenarios described below:
    - **Fully paid:** Applicant has fully paid the loan (the principal and the interest rate)
    - **Current:** Applicant is in the process of paying the installments, i.e. the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.
    - **Charged-off:** Applicant has not paid the installments in due time for a long period of time, i.e. he/she has 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 and so this data is not available with the company (and thus in this dataset)



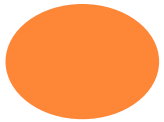
Data Summary

- Loan.csv file contains 39717 rows and 111 columns.
- There is no sub-headers or sub-Footer in the given data set.



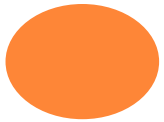
Data Cleaning

1. Loading data from loan CSV
2. Checking for null values in the dataset
3. Checking for unique values
4. Checking for duplicated rows in data
5. Dropping Records & Columns
6. Common Functions
7. Data Conversion
8. Outlier Treatment
9. Imputing values in Columns



Data Cleaning

1. **Loading data from loan CSV:** While loading the dataset, some of the variables had mixed datatypes so they have to be converted accordingly as per analysis.
2. **Checking for null values in the dataset:** There're many columns with null values. So they had to be dropped as they won't play a role in the analysis of the dataset. Roughly 48% of the columns were dropped.
3. **Checking for unique values:** If the column has only a single unique value, it does not make any sense to include it as part of our data analysis. We need to find out those columns and drop them from the dataset. 9 columns had such unique values and they were removed.
4. **Checking for duplicated rows in data:** No duplicate rows were found.
5. **Dropping Records and Columns:**
   - ✓ Dropped records where **loan_status="Current"** as the loan in progress cannot provide us insights as to whether the borrower is likely to default or not.
   - ✓ Dropping columns where missing data is **>=65%** as these columns will skew our data analysis and they need to be removed.
   - ✓ Dropping extra columns containing text like **collection_recovery_fee, delinq_2yrs, desc, earliest_cr_line, emp_title, id, inq_last_6mths, last_credit_pull_d, last_pymnt_amnt, last_pymnt_d, member_id, open_acc, out_prncp, out_prncp_inv, pub_rec, recoveries, revol_bal, revol_util, title, total_acc, total_pymnt, total_pymnt_inv, total_rec_int, total_rec_late_fee, total_rec_prncp, url, zip_code** as these will not contribute to loan pass or fail.



Data Cleaning

6. **Common Functions:** Common functions were created for repeating common operations like plotting bar graphs, box plots, histograms, countplots, binning etc.
7. **Data Conversion**: Converted columns like **debt to income (dti), funded amount (funded_amnt), funded amount investor (funded_amnt_inv) and loan amount (loan_amnt) to float** to match the data. Also converted **loan date (issue_d)** to **DateTime (format: yyyy-mm-dd).**
8. **Outlier Treatment:** Calculated the **Inter-Quartile Range (IQR)** and filtering out the outliers outside of lower and upper bound. During Outlier analysis the following observations were made
   - ✓ The annual income of most of the loan applicants is between 40K - 75K USD
   - ✓ The loan amount of most of the loan applicants is between 5K - 15K
   - ✓ The funded amount of most of the loan applicants is between 5K - 14K USD
   - ✓ The funded amount by investor for most of the loan applicants is between 5K - 14K USD
   - ✓ The interest rate on the loan is between 9% - 14%
   - ✓ The monthly installment amount on the loan is between 160 - 440
   - ✓ The debt to income ration is between 8 - 18

Data Cleaning

9. **Imputing values in Columns:**
   - ✓ **Replaced missing values of annual_inc with the corresponding mode value of annual_inc of the emp_length annual_inc field:** They Employment length has **1015** missing values, which means either they are **not employed or self-employed (business owners).** Considering they have a decent average annual income, we have assumed that these are business owners and we have added their employment duration with the mode value of **emp_length** which is **10+ years**.
   - ✓ Mapped employment length with the respective number of years in int.
   - ✓ Imputed **NONE** values as **OTHER** for **home_ownership.**
   - ✓ Replaced the '**Source Verified**' values as '**Verified**' since both values mean the same thing i.e. the loan applicant has some source of income which is verified.
   - ✓ There are **660 null values** for **pub_rec_bankruptcies**. Dropped those rows as they cannot be imputed.

Post Data cleaning and Pre-processing of dataset, we were left with **36094** rows × **18** columns.

Univariate Analysis

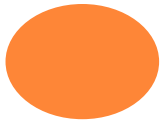
- ✓ **Univariate analysis** is a statistical method used to analyze and summarize data sets consisting of **one variable**. It deals with the analysis of a single variable, rather than multiple variables, to understand its distribution, central tendency and dispersion.
- ✓ It was carried out for both **Categorical** and **Quantitative** Variables

A. **Categorical Variables:**

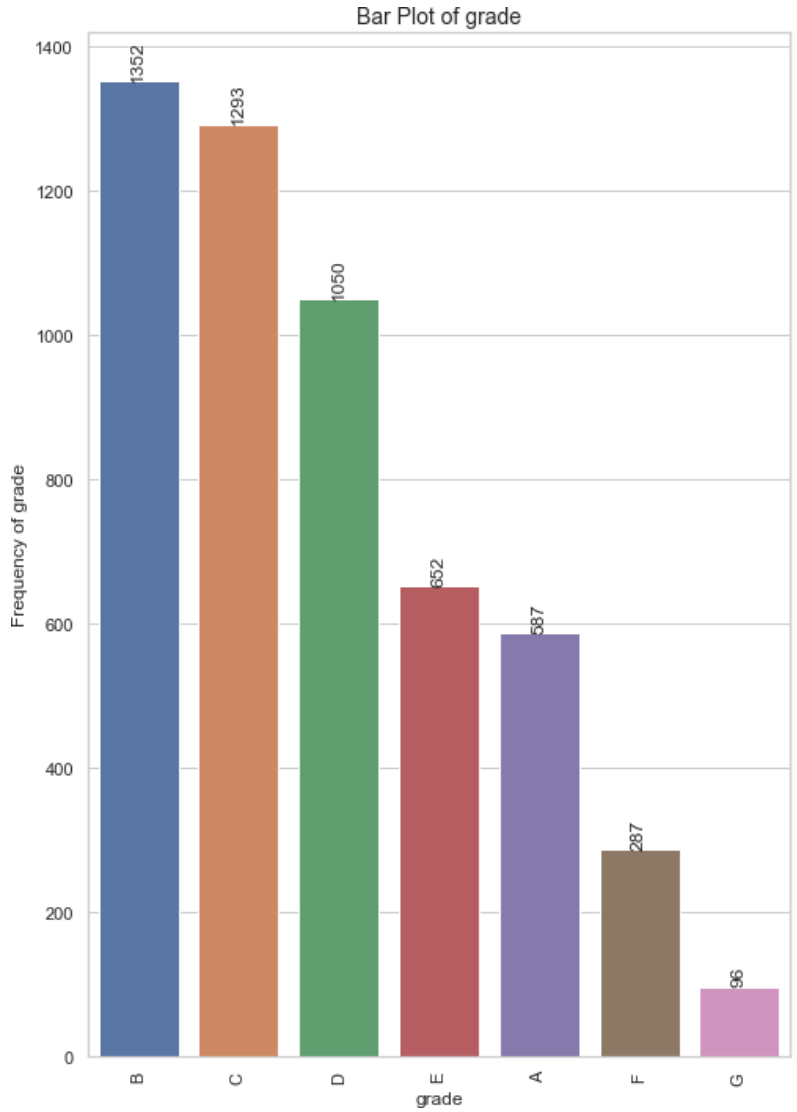
Ordered	Unordered
✓ Grade (grade) ✓ Sub grade (sub_grade) ✓ Term (36 / 60 months) (term) ✓ Employment length (emp_length) ✓ Issue year (issue_y) ✓ Issue month (issue_m) ✓ Issue quarter (issue_q)	✓ Address State (addr_state) ✓ Loan purpose (purpose) ✓ Home Ownership (home_ownership) ✓ Loan status (loan_status) ✓ Loan paid (loan_paid)

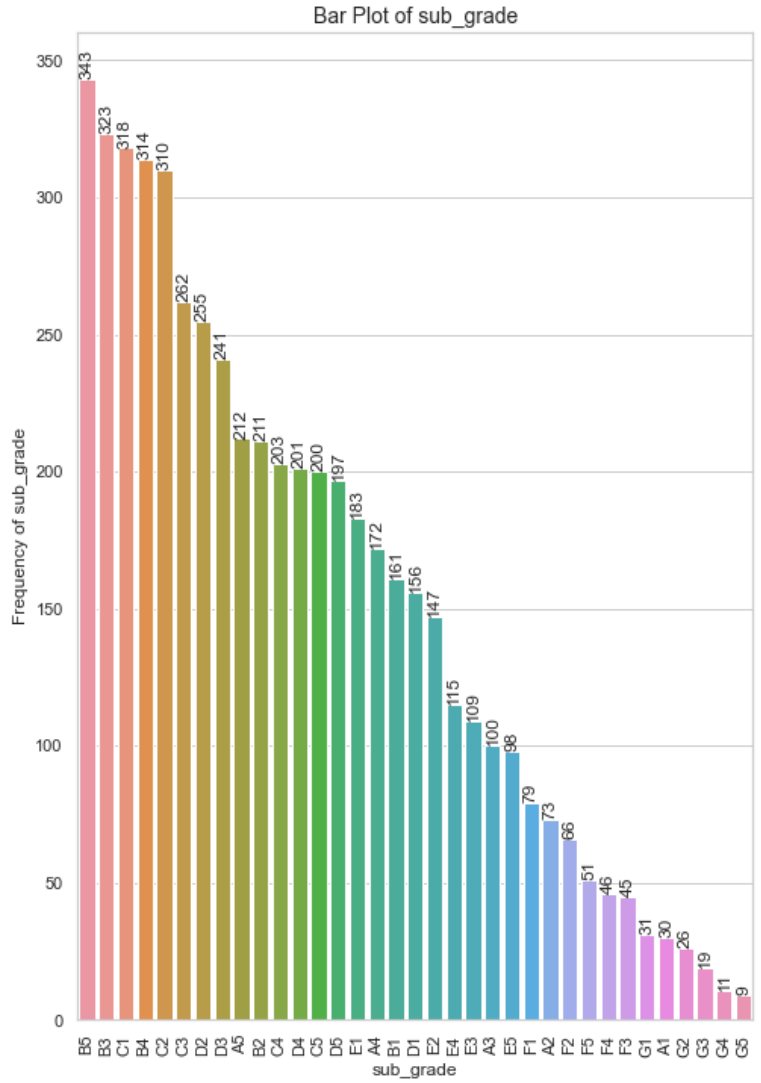
B. **Quantitative Variables:**

- ✓ **Interest rate bucket (int_rate_bucket)**
- ✓ **Annual income bucket (annual_inc_bucket)**
- ✓ **Loan amount bucket (loan_amnt_bucket)**
- ✓ **Funded amount bucket (funded_amnt_bucket)**
- ✓ **Debt to Income Ratio (DTI) bucket (dti_bucket)**
- ✓ **Monthly Installment (installment)**



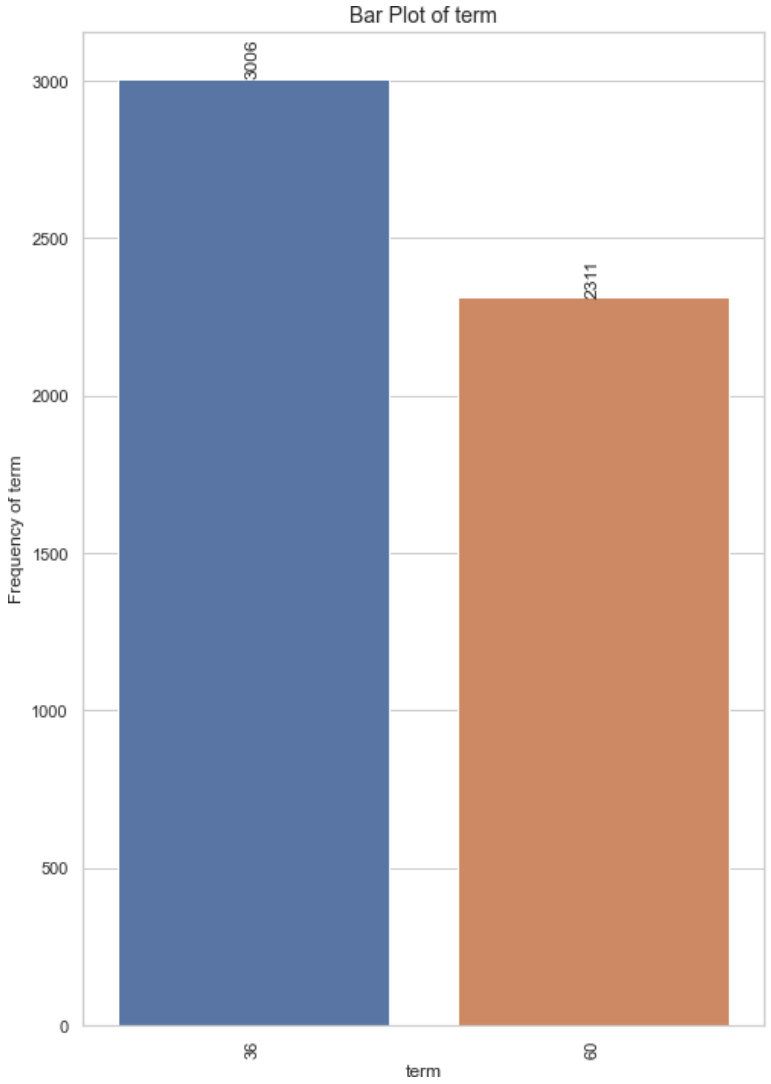
Univariate Analysis (Unordered Categorical)

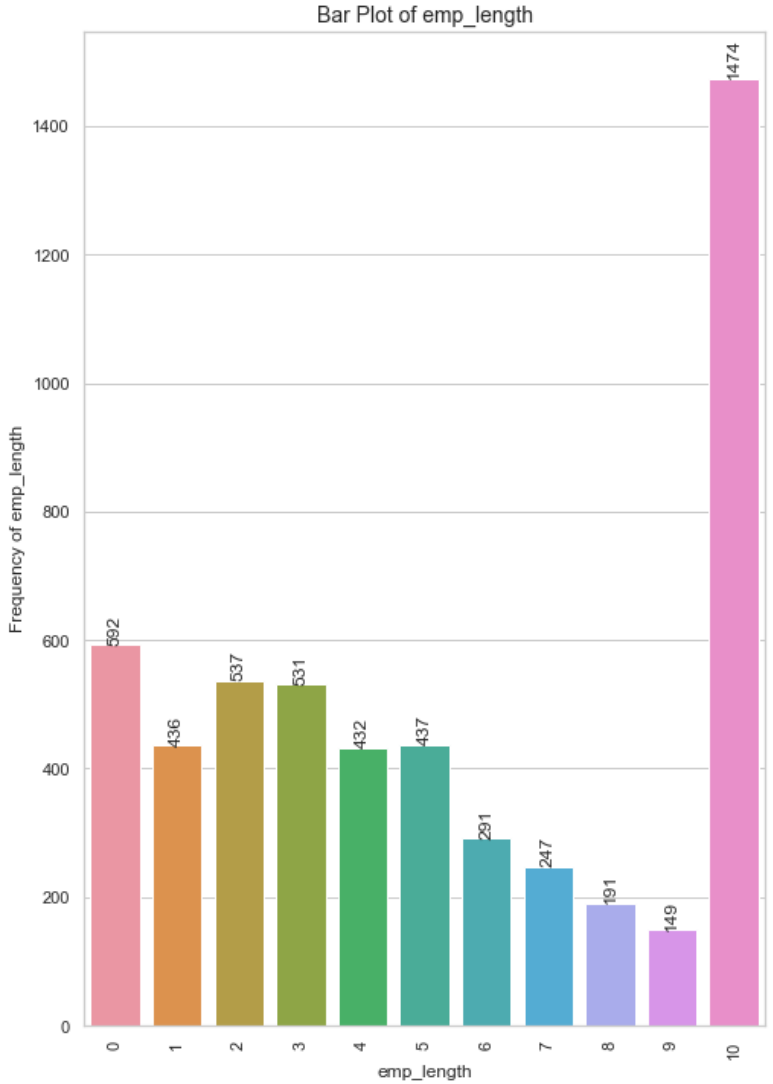




Grade and Sub-Grade

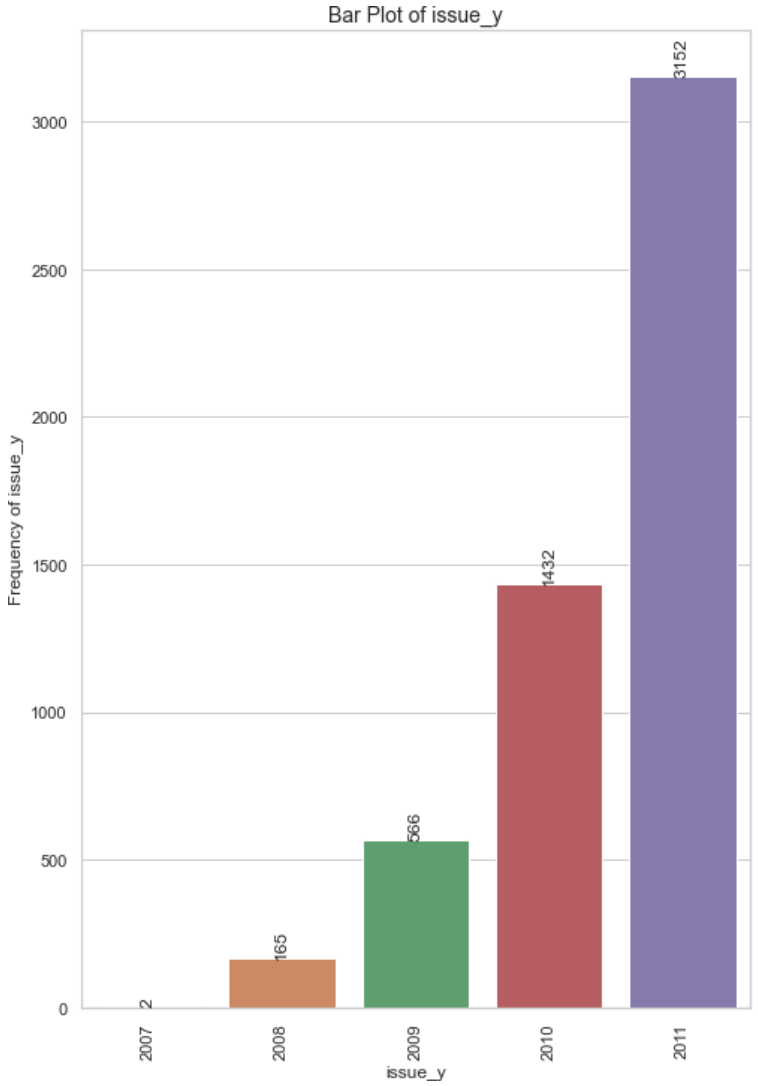
Univariate Analysis (Unordered Categorical)

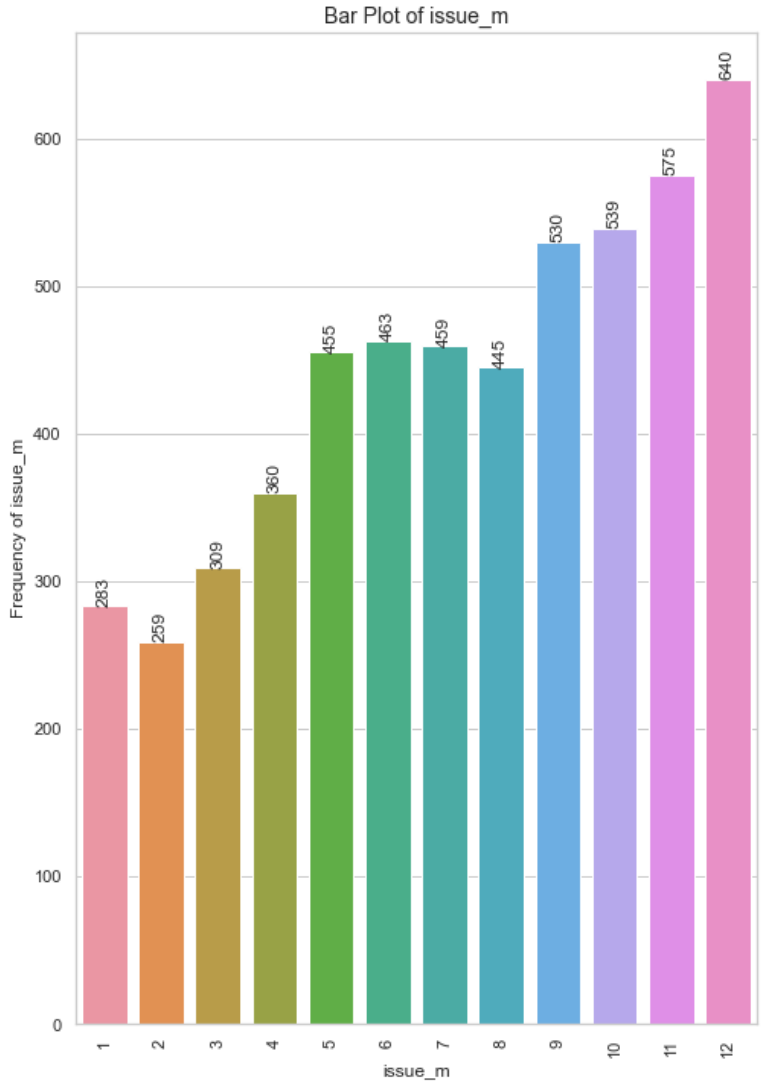




Term and Employment Length

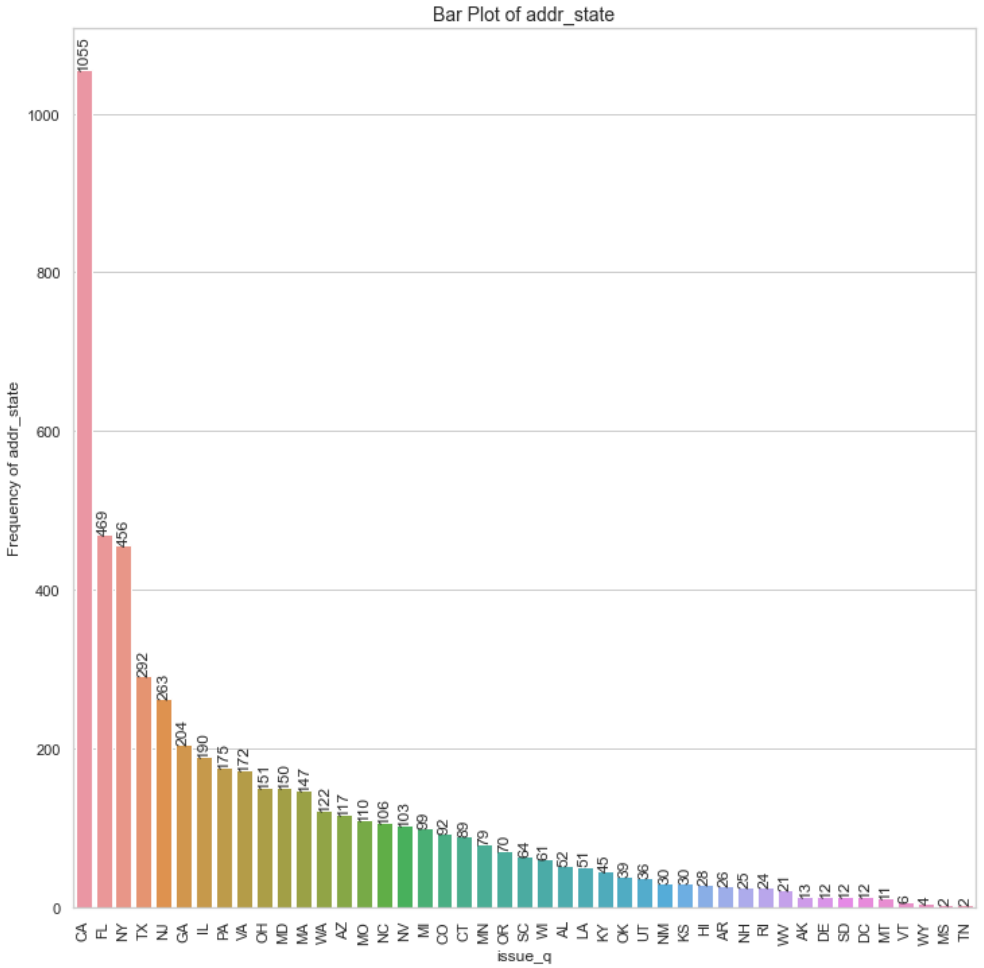
Univariate Analysis (Unordered Categorical)

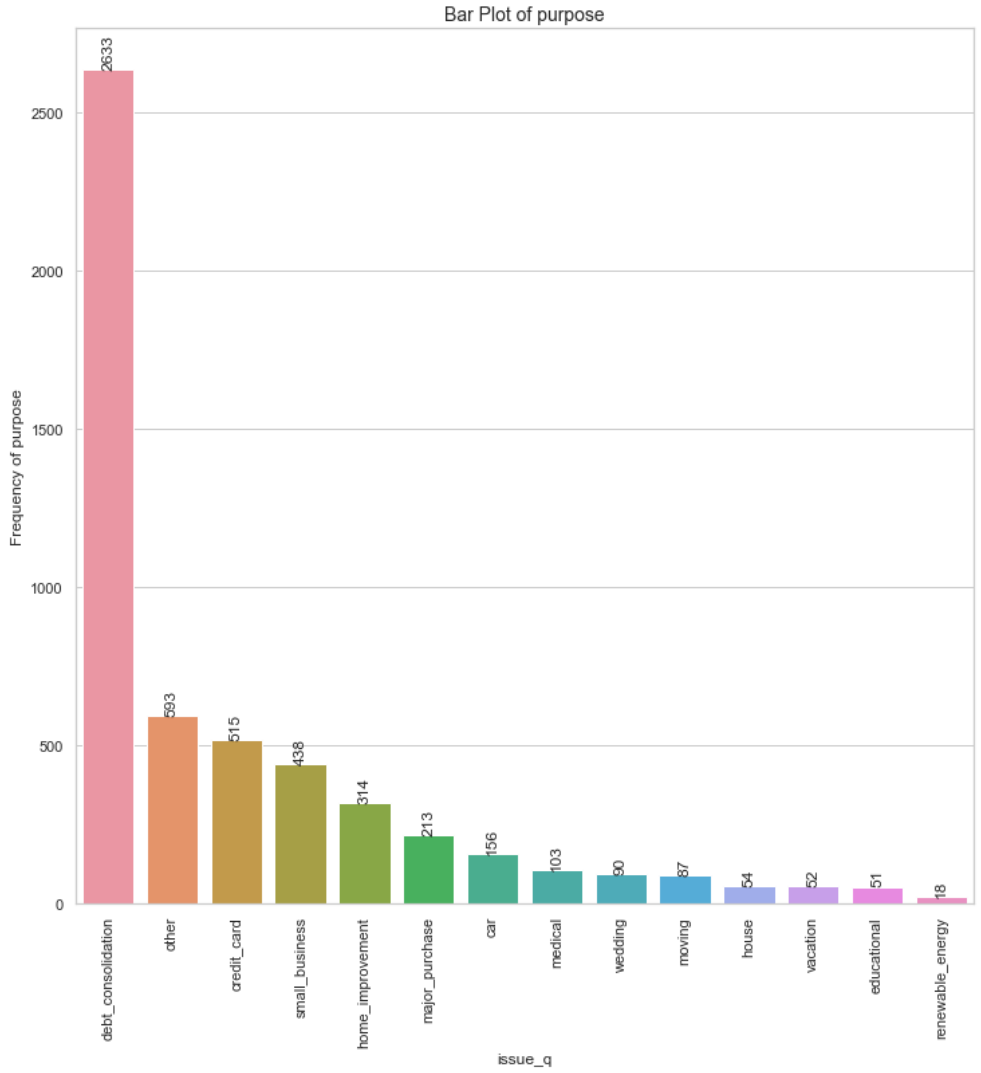




Term and Employment Length

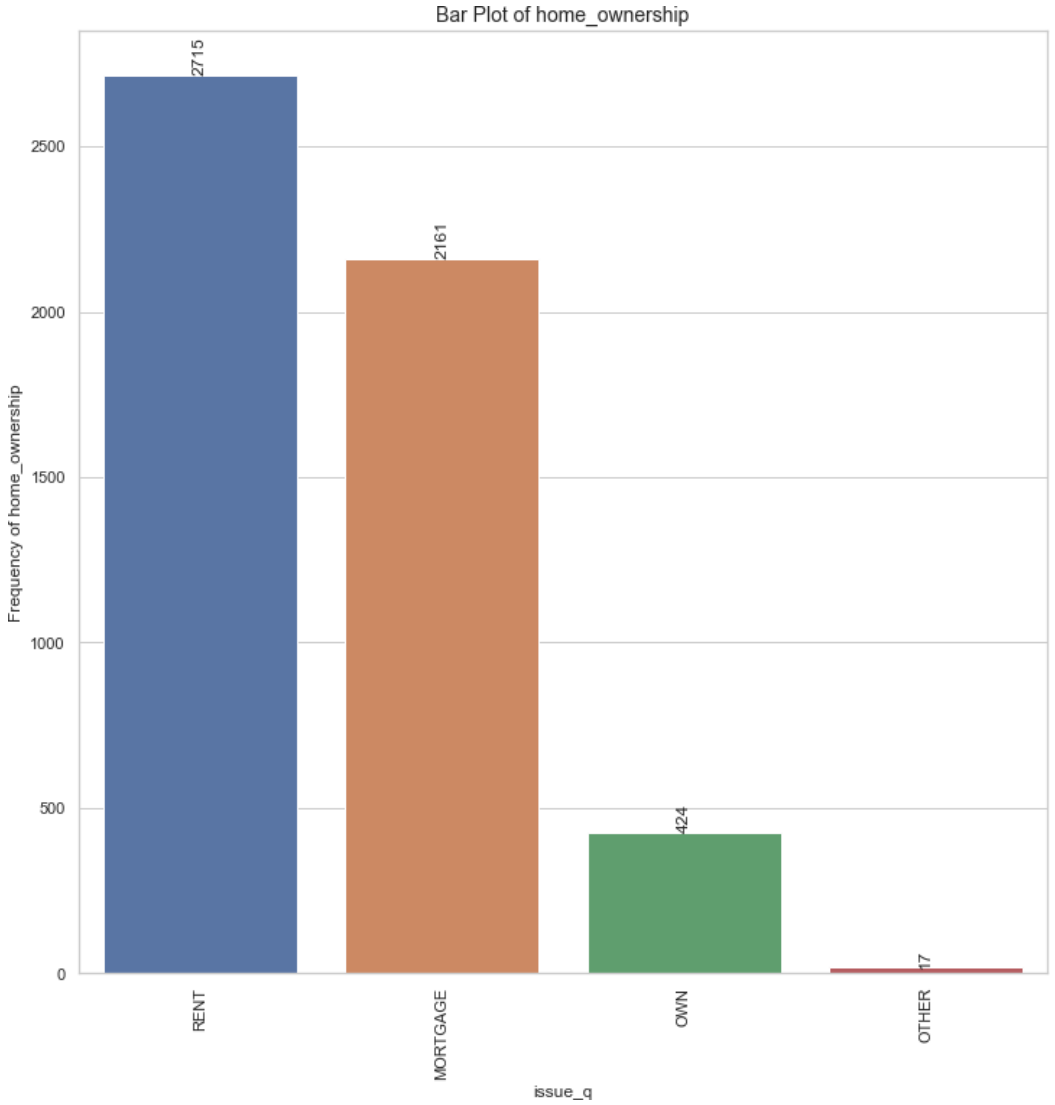
Univariate Analysis (Unordered Categorical)





Address State and Purpose of Loan

Univariate Analysis (Unordered Categorical)



Various types of Home Ownerships

Univariate Analysis (Categorical Variables)

Observations & Inferences:

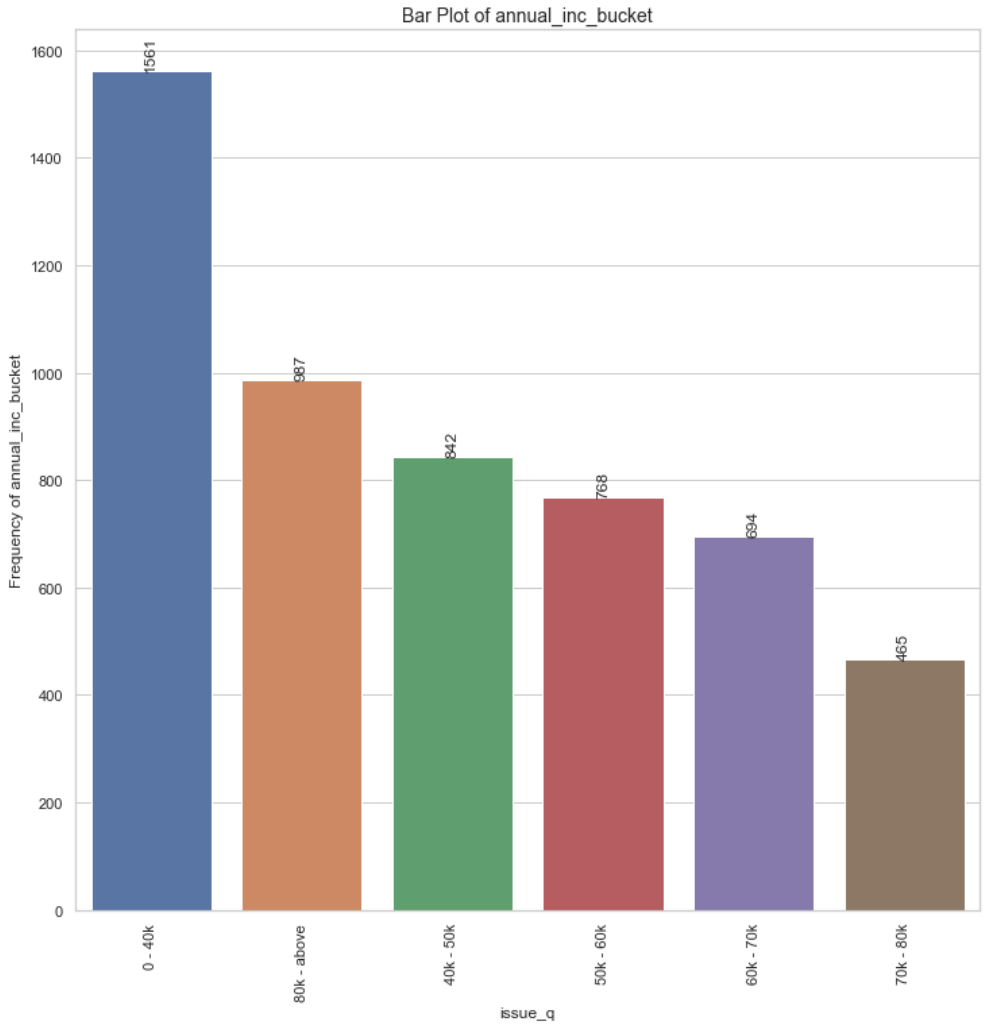
A. Ordered Categorical Variables:

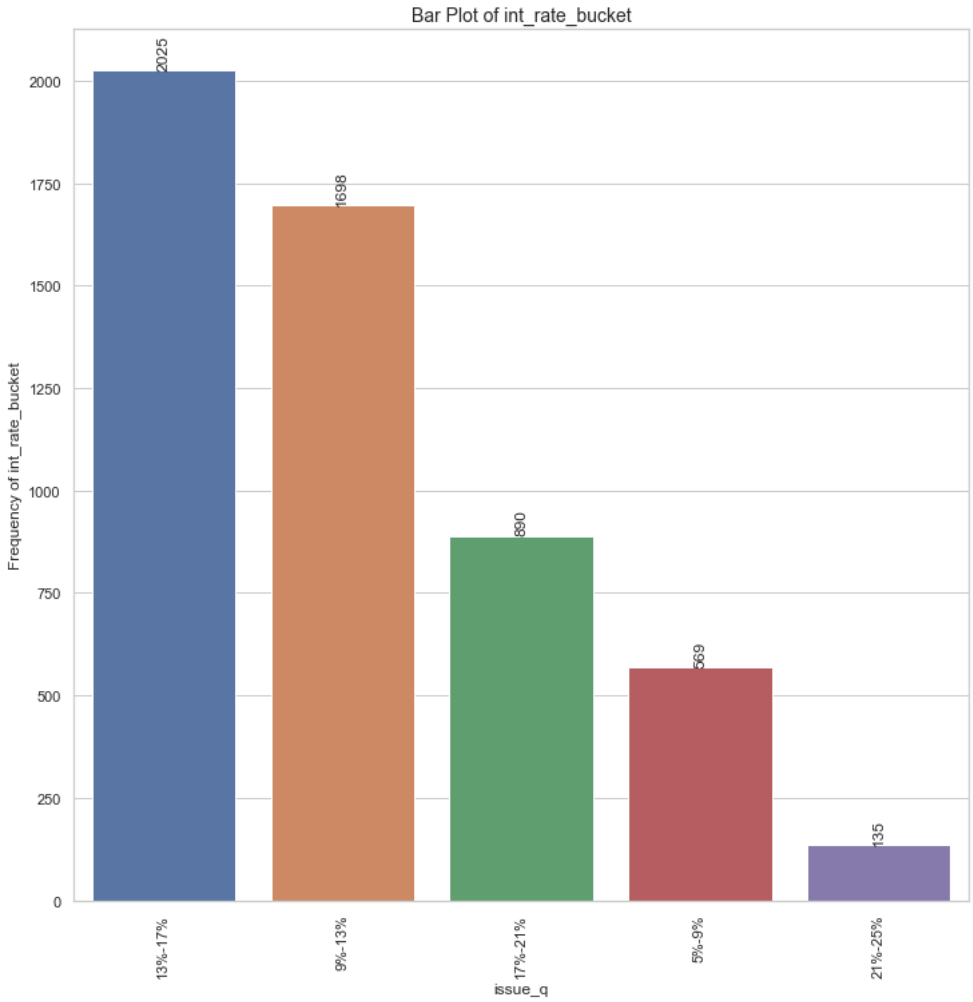
- ✓ Grade B had the highest number of "Charged off" loan applicants, with a total of 1,352 applicants, indicating that applicants with this credit grade faced challenges in repaying their loans.
- ✓ Short-term loans with a duration of 36 months were the most popular among "Charged off" applicants, with 3,006 applications. This suggests that a significant portion of applicants who experienced loan default chose shorter repayment terms.
- ✓ Applicants who had been employed for more than 10 years accounted for the highest number of "Charged off" loans, totaling 1,474. This indicates that long-term employment history did not necessarily guarantee successful loan repayment.
- ✓ The year 2011 recorded the highest number of "Charged off" loan applications, totaling 3,152, signaling a positive trend in the number of applicants facing loan defaults over the years. This could be indicative of economic or financial challenges during that year.
- ✓ "Charged off" loans were predominantly taken during the 4th quarter, with 2,284 applications, primarily in December. This peak in loan applications during the holiday season might suggest that financial pressures during the holidays contributed to loan defaults.

B. Unordered Categorical Variables:

- ✓ California had the highest number of "Charged off" loan applicants, with 1,055 applicants. For such applicants, the lending company needs to implement stricter eligibility criteria or credit assessments due to a higher number of "Charged off" applicants from this state.
- ✓ Debt consolidation was the primary loan purpose for most "Charged off" loan applicants, with 2,633 applicants selecting this option. The lending company needs to exercise caution when approving loans for debt consolidation purposes, as it was the primary loan purpose for many "Charged off" applicants.
- ✓ The majority of "Charged off" loan participants, totaling 2,715 individuals, lived in rented houses. The lending company must assess the financial stability of applicants living in rented houses, as they may be more susceptible to economic fluctuations.
- ✓ A significant number of loan participants, specifically 5,317 individuals, were loan defaulters, unable to clear their loans. The lending company should enhance risk assessment practices, including stricter credit checks and lower loan-to-value ratios, for applicants with a history of loan defaults. They should offer financial education and support services to help borrowers manage their finances and improve loan repayment outcomes.

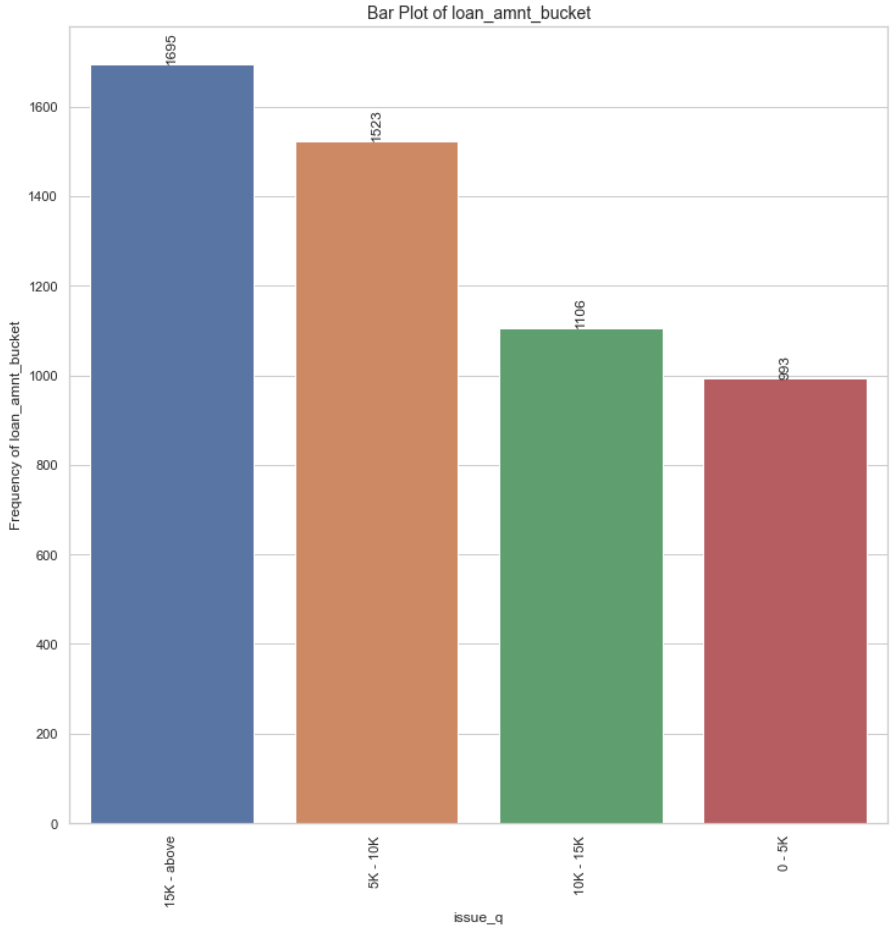
Univariate Analysis (Quantitative Variables)

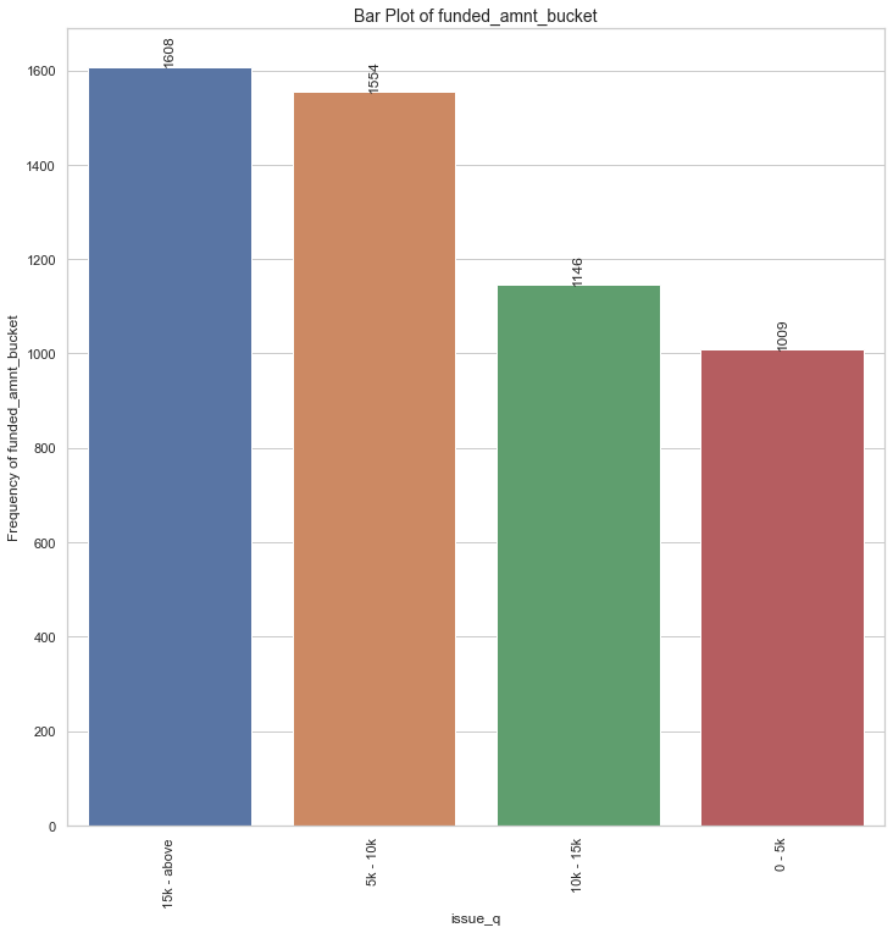




Buckets of Annual Income Status and Loan Interest Rates

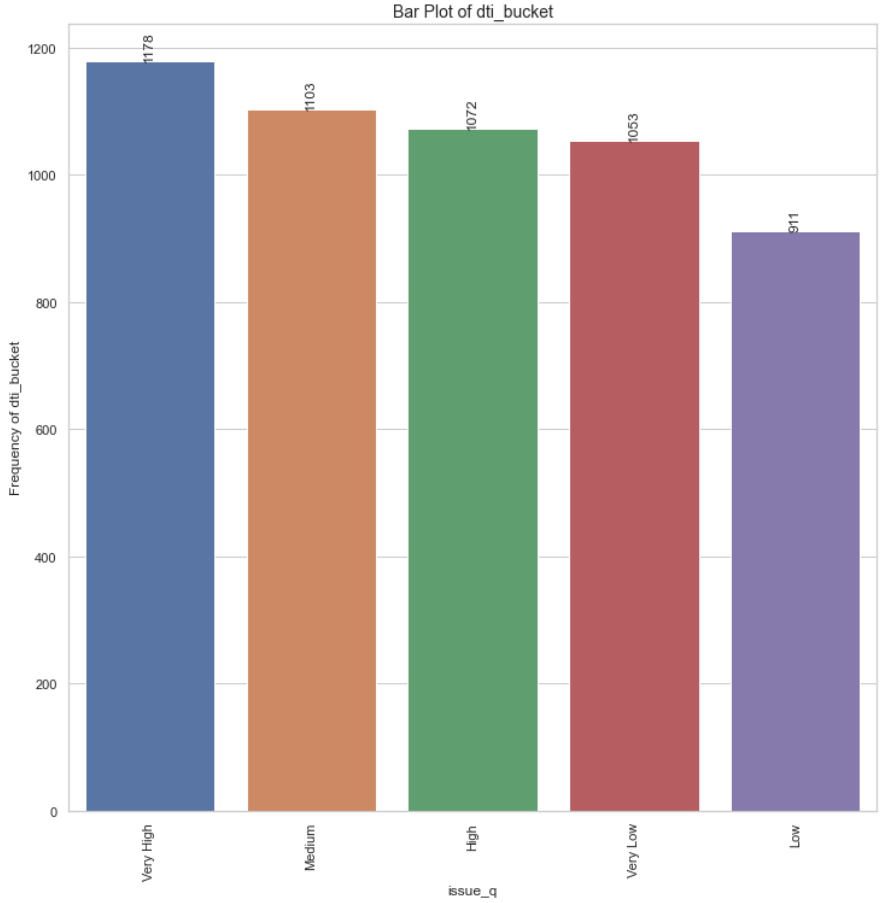
Univariate Analysis (Quantitative Variables)



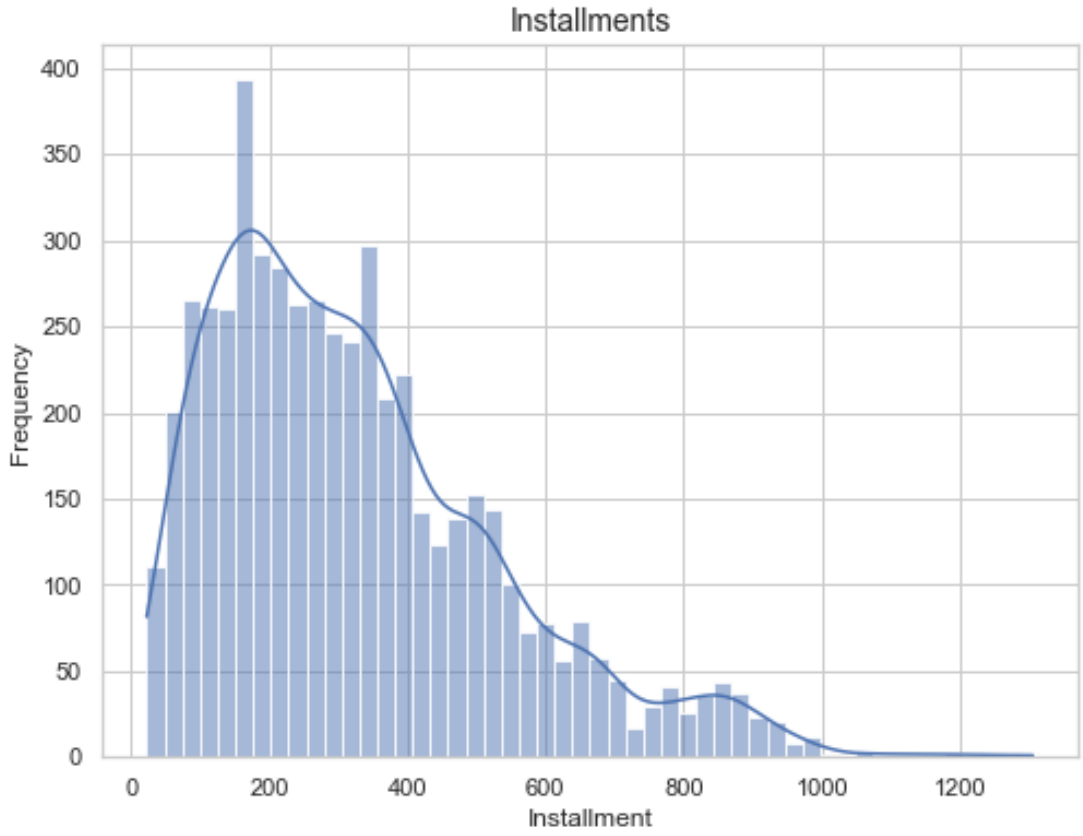


Buckets of Loan Amount and Funded Amount

Univariate Analysis (Quantitative Variables)



Bucket of Debt to Income Ratio (DTI)

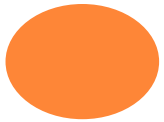


Histogram of Installment (For Defaulted Loans)

Univariate Analysis (Quantitative Variables)

Observations & Inferences:

- ✓ 1,561 loan applicants who charged off had annual salaries less than 40,000 USD. The lending company should exercise caution when lending to individuals with low annual salaries. They should implement rigorous income verification and assess repayment capacity more thoroughly for applicants in this income bracket.
- ✓ Among loan participants who charged off (2,025), a considerable portion belonged to the interest rate bucket of 13%-17%. To reduce the risk of default, the lending company should consider offering loans at lower interest rates when possible.
- ✓ 1,695 loan participants who charged off received loan amounts of 15,000 USD and above. The lending company should evaluate applicants seeking higher loan amounts carefully. They should ensure the applicants must have a strong credit history and repayment capability to handle larger loans.
- ✓ 1,608 loan participants who charged off received funded amounts of 15,000 USD and above. The lending company should ensure that the funded amounts align with the borrower's financial capacity. They should conduct thorough credit assessments for larger loan requests.
- ✓ Among loan participants who charged off, 1,178 loan applicants had very high debt-to-income ratios. The lending company should implement strict debt-to-income ratio requirements to prevent lending to individuals with unsustainable levels of debt relative to their income.
- ✓ Among loan participants who charged off, it's observed that the majority of them had monthly installment amounts falling within the range of 160-440 USD. The lending company should closely monitor and assess applicants with similar installment amounts to mitigate the risk of loan defaults.



Bivariate Analysis

- ✓ Bivariate analysis is a statistical method that involves the simultaneous analysis of two variables (factors). It aims to determine the empirical relationship between them. The analysis can be used to test hypotheses, identify patterns, or explore relationships between the variables.
- ✓ It was carried out for both Categorical and Quantitative Variables

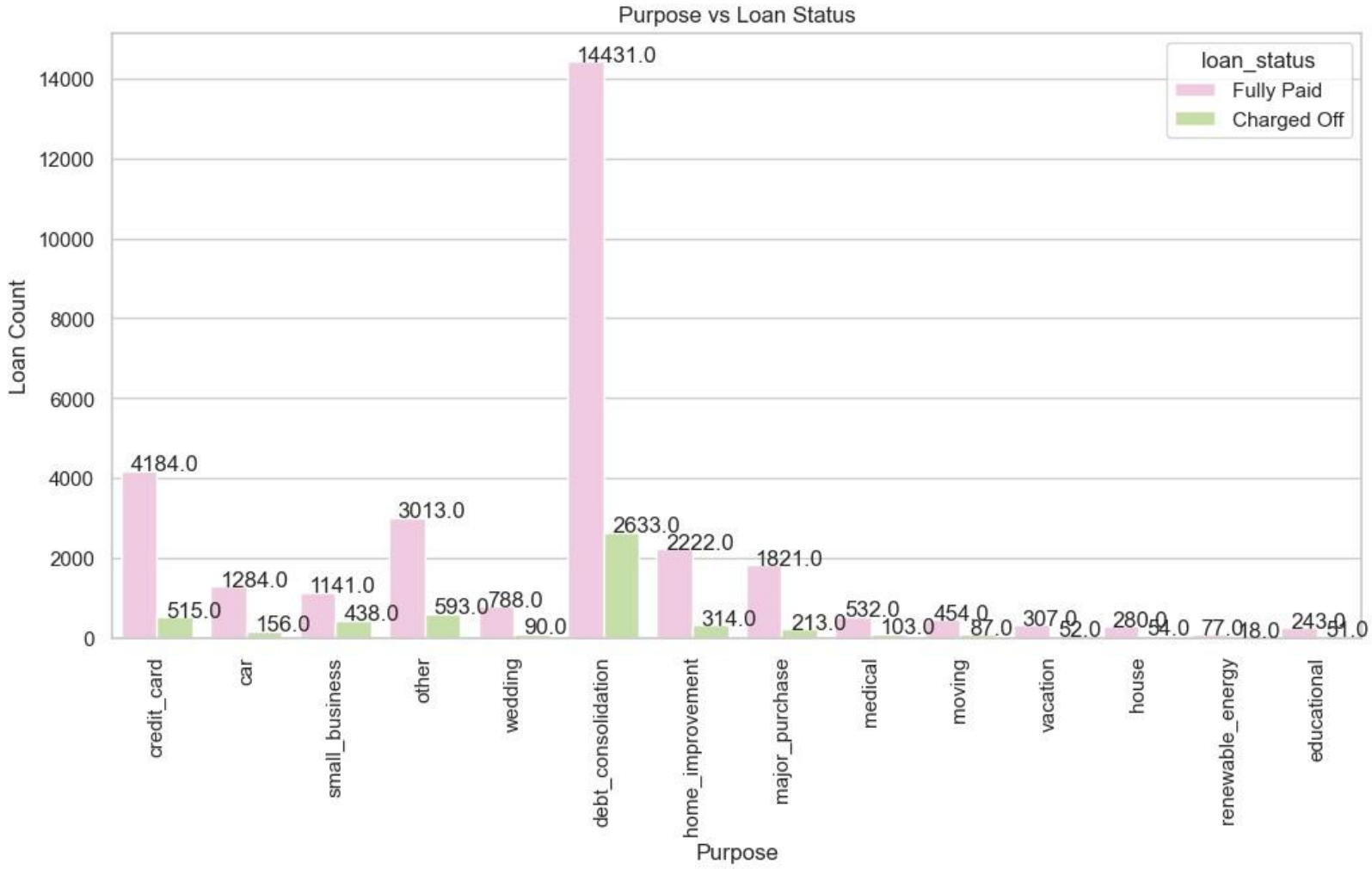
A. **Categorical Variables:**

Ordered	Unordered
✓ Grade (grade) ✓ Sub grade (sub_grade) ✓ Term (36 / 60 months) (term) ✓ Employment length (emp_length) ✓ Issue year (issue_y) ✓ Issue month (issue_m) ✓ Issue quarter (issue_q)	✓ Loan purpose (purpose) ✓ Home Ownership (home_ownership) ✓ Verification Status (verification_status) ✓ Address State (addr_state)

B. **Quantitative Variables:**

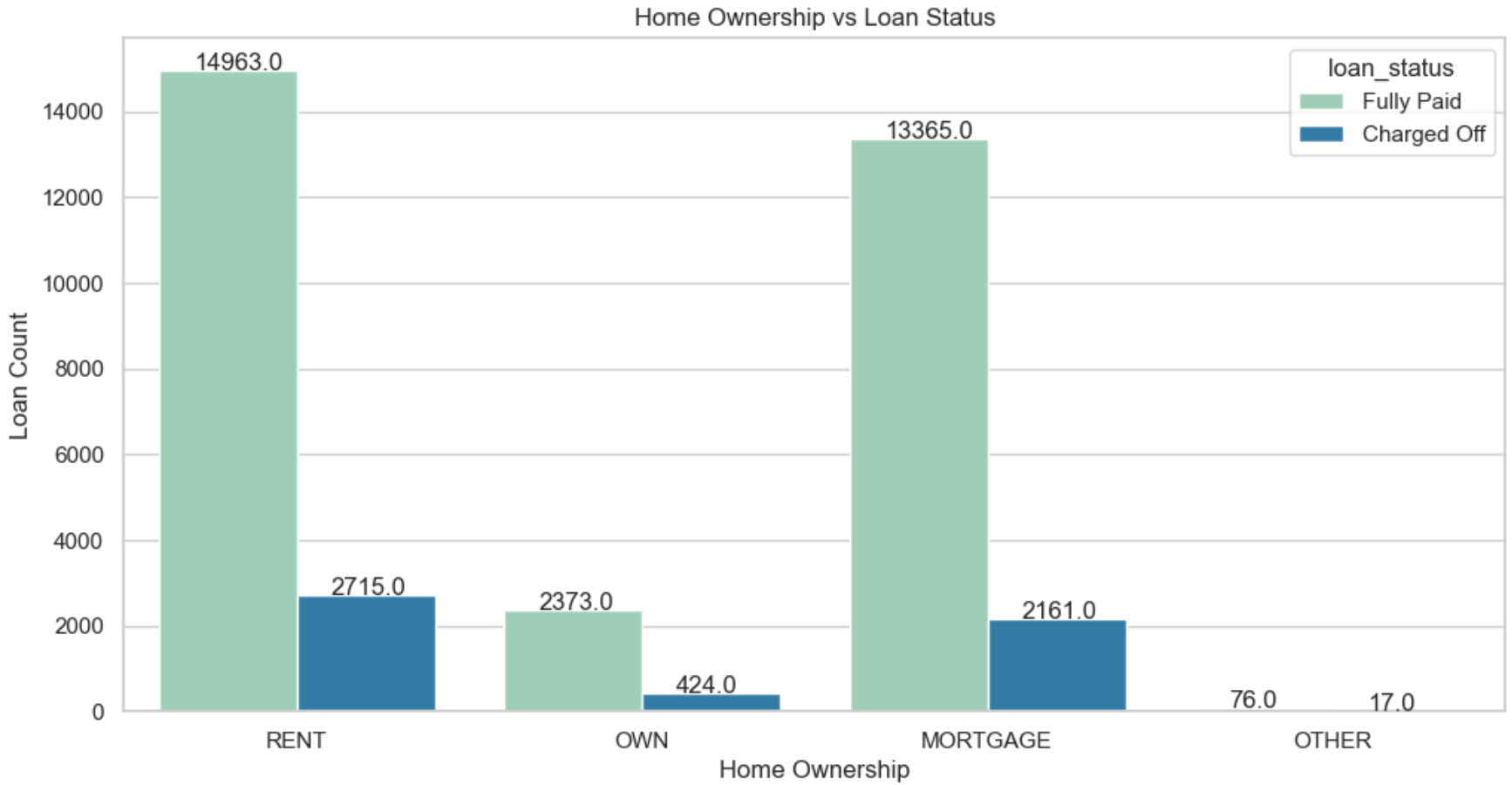
- ✓ **Int Rate Bucket (int_rate_bucket)**
- ✓ **Debt to Income Bucket (dti_bucket)**
- ✓ **Annual Income Bucket (annual_inc_bucket)**
- ✓ **Funded Amount Bucket (funded_amnt_bucket)**
- ✓ **Loan Amount Bucket (loan_amnt_bucket)**

Bivariate Analysis (Unordered Categorical)



Purpose of Loan v/s Status of Loan

Bivariate Analysis (Unordered Categorical)

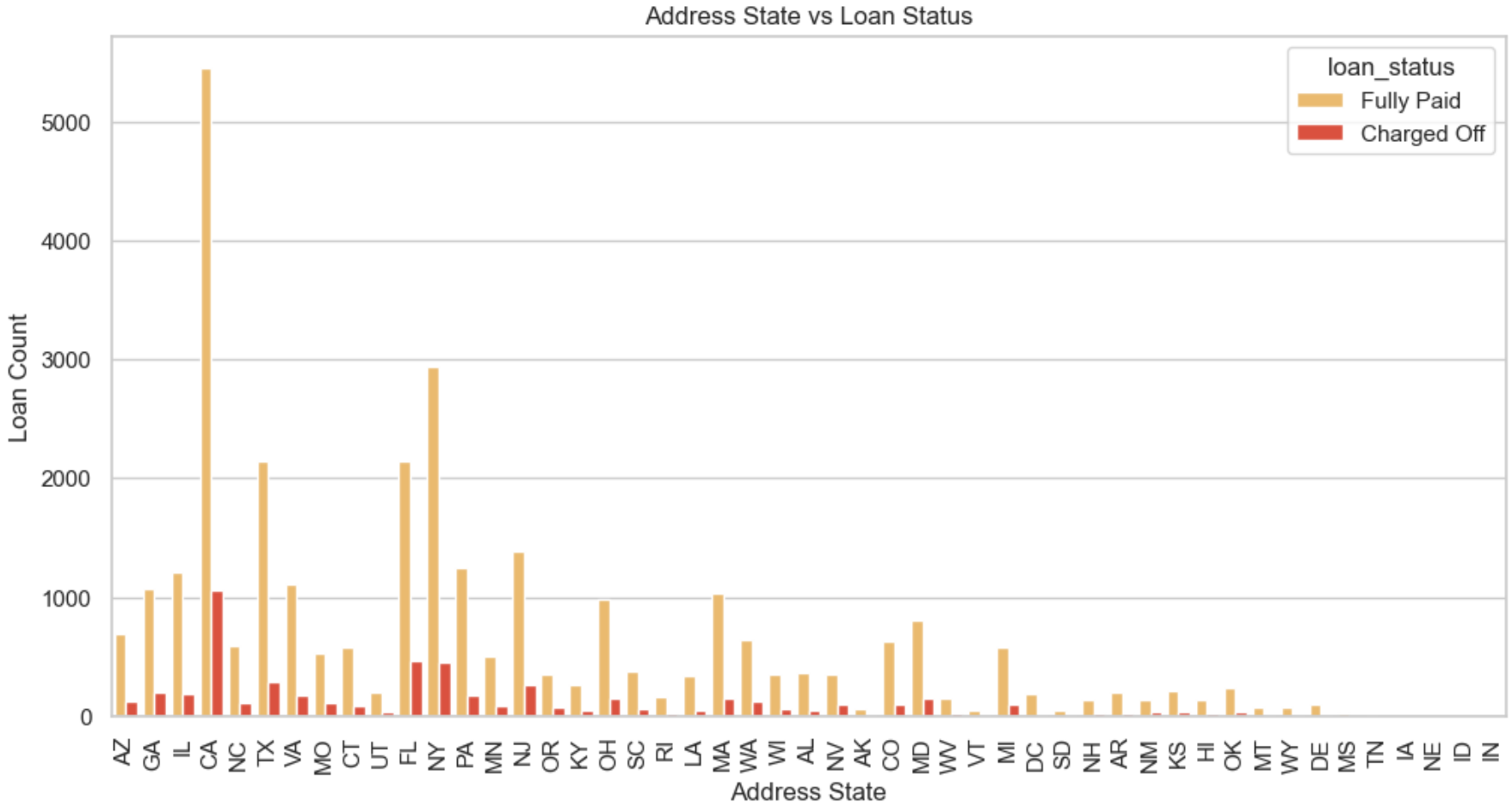


Home Ownership v/s Status of Loan

Bivariate Analysis (Unordered Categorical)

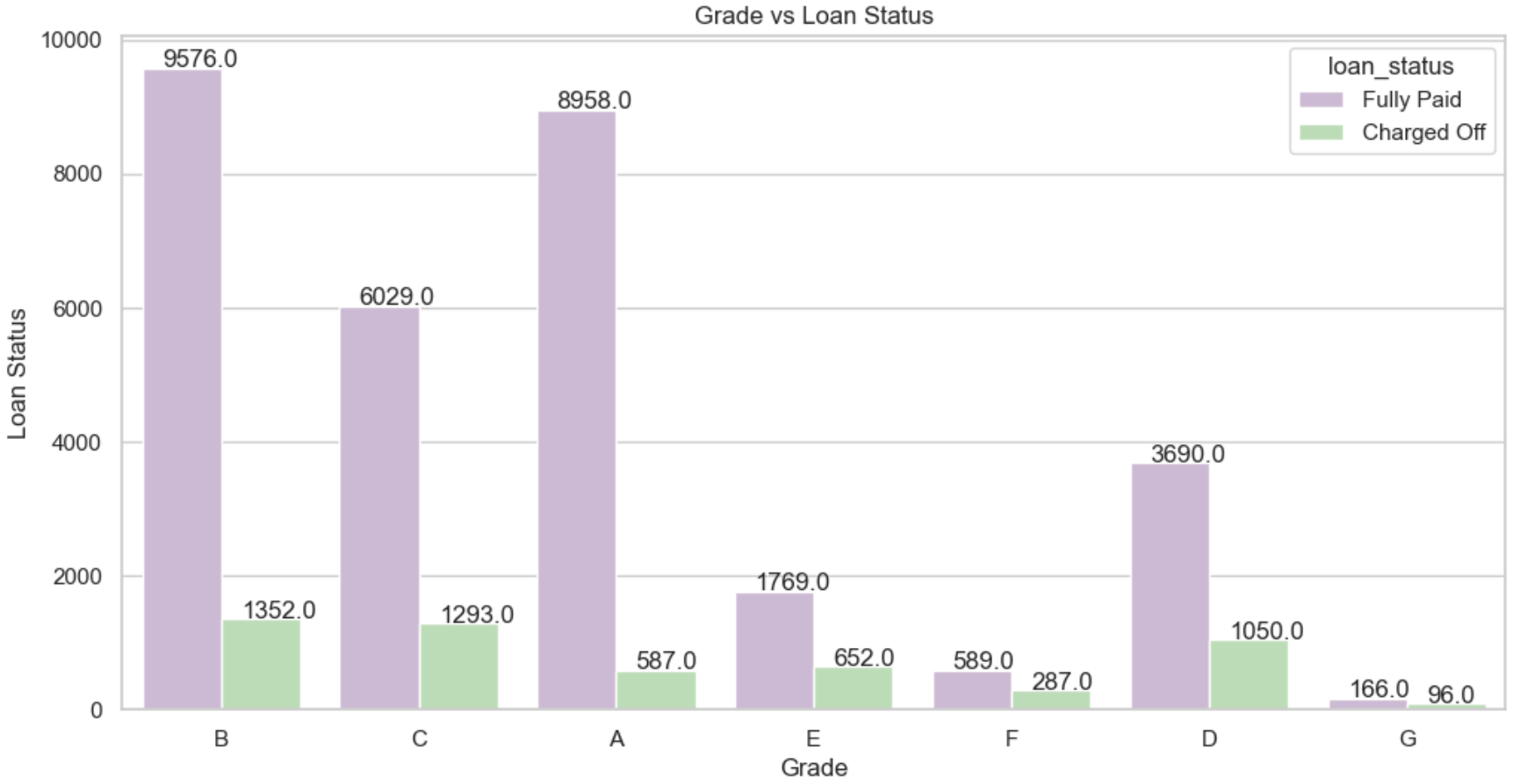
Verification Status of Loan v/s Status of Loan

Bivariate Analysis (Unordered Categorical)



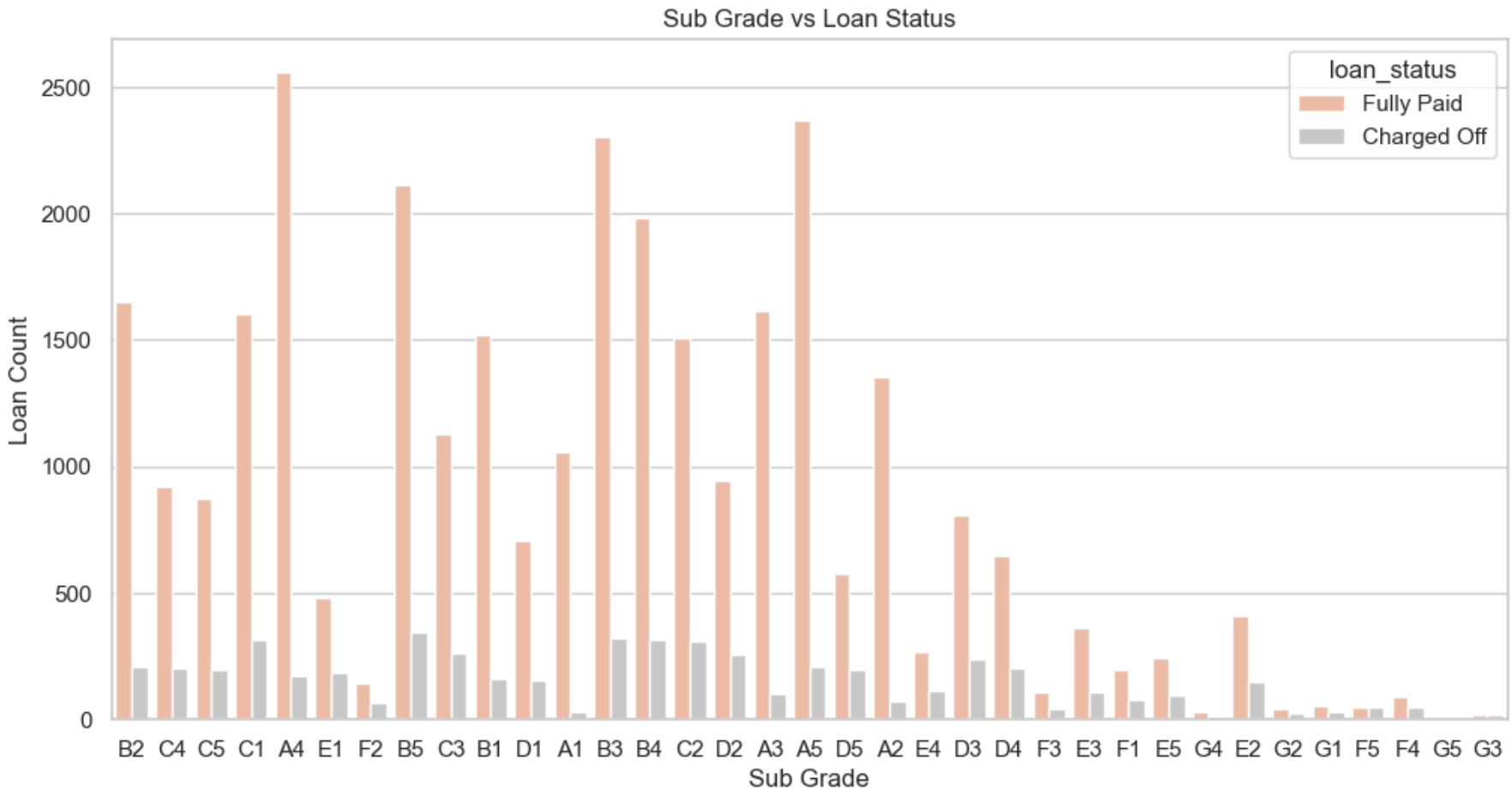
Address State v/s Status of Loan

Bivariate Analysis (Ordered Categorical)



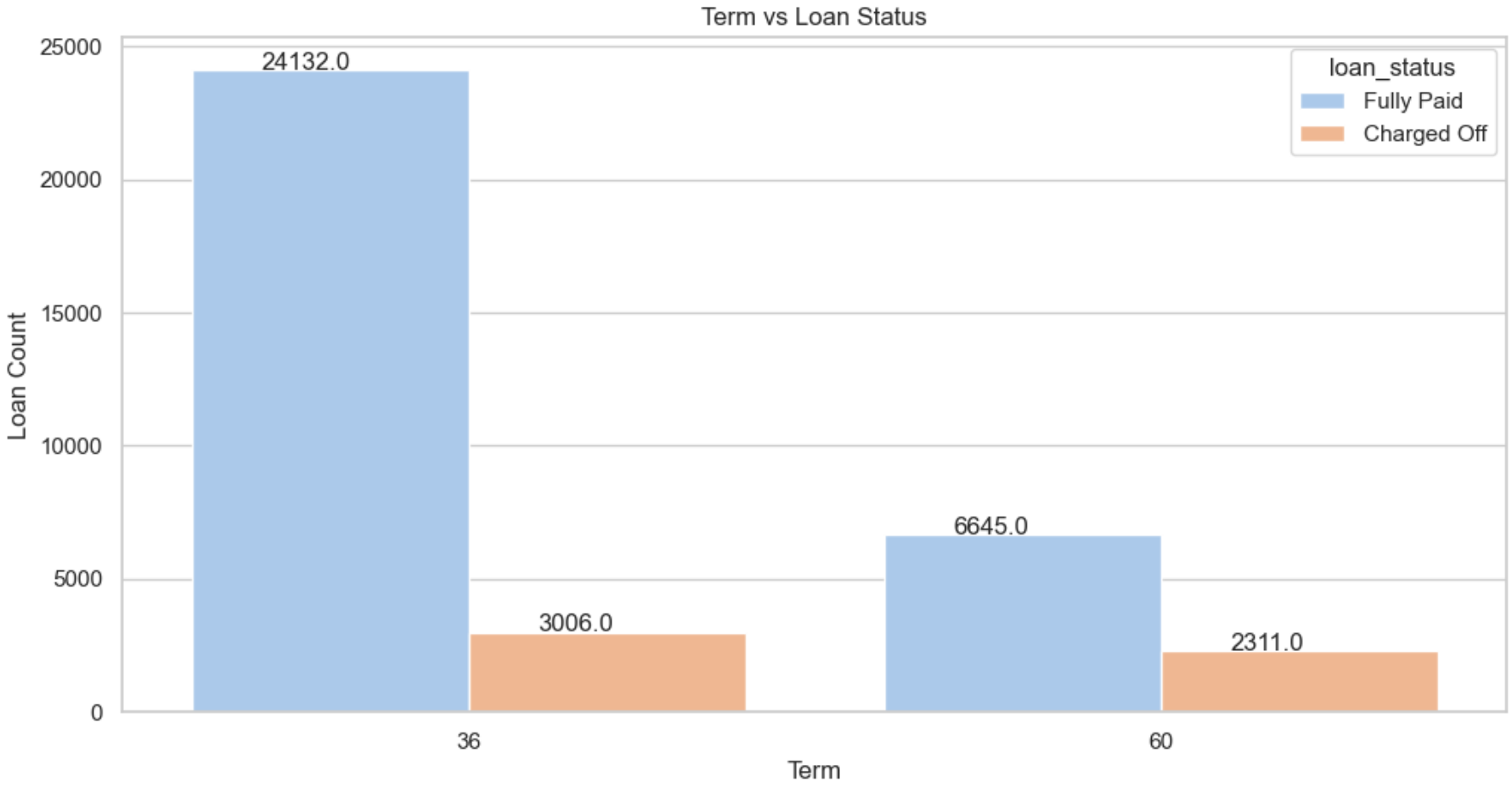
Loan Grade v/s Status of Loan

Bivariate Analysis (Ordered Categorical)



Loan Sub-Grade v/s Status of Loan

Bivariate Analysis (Ordered Categorical)

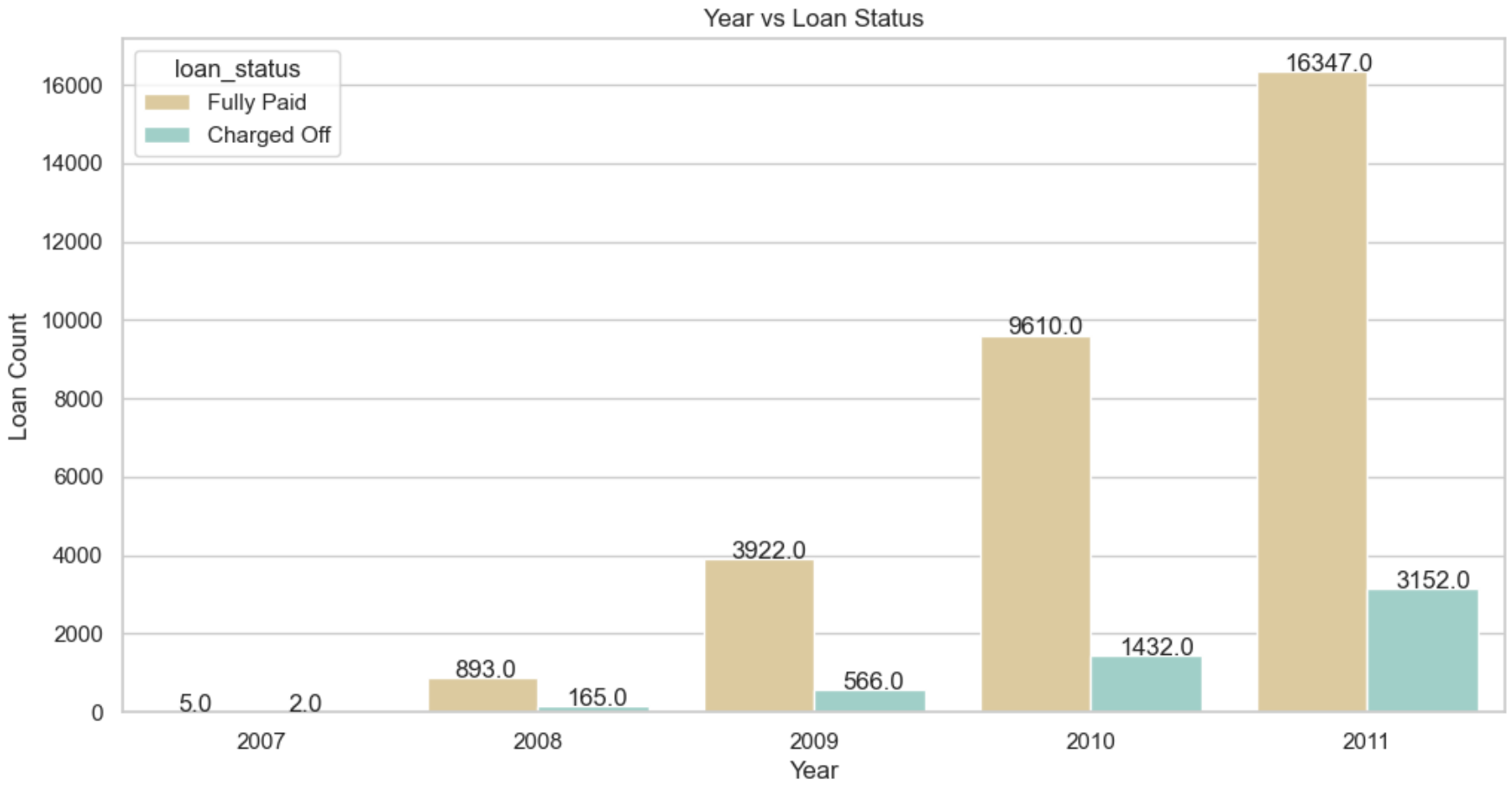


Term of Loan v/s Status of Loan

Bivariate Analysis (Ordered Categorical)

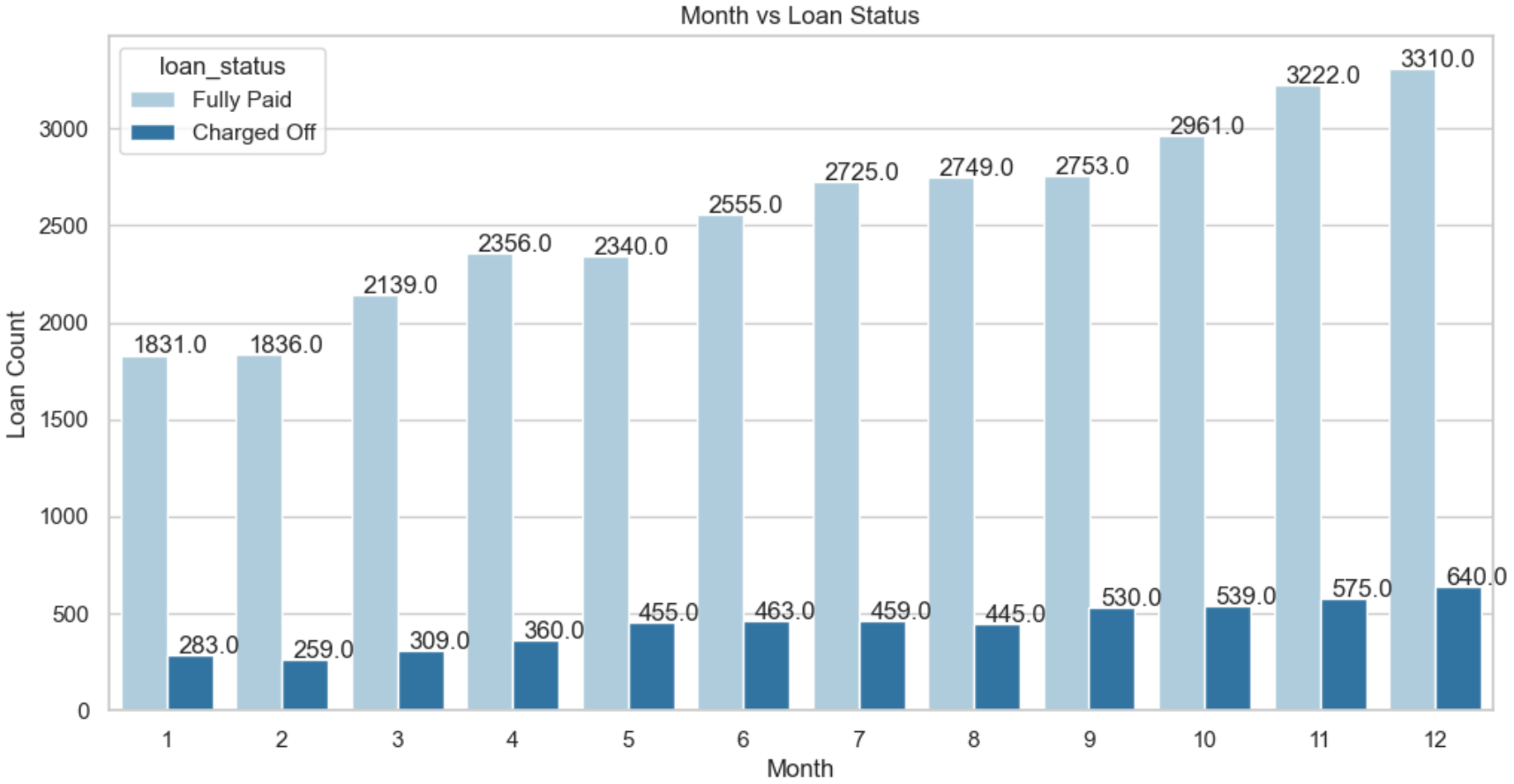
Employment Length of Customer v/s Status of Loan

Bivariate Analysis (Ordered Categorical)



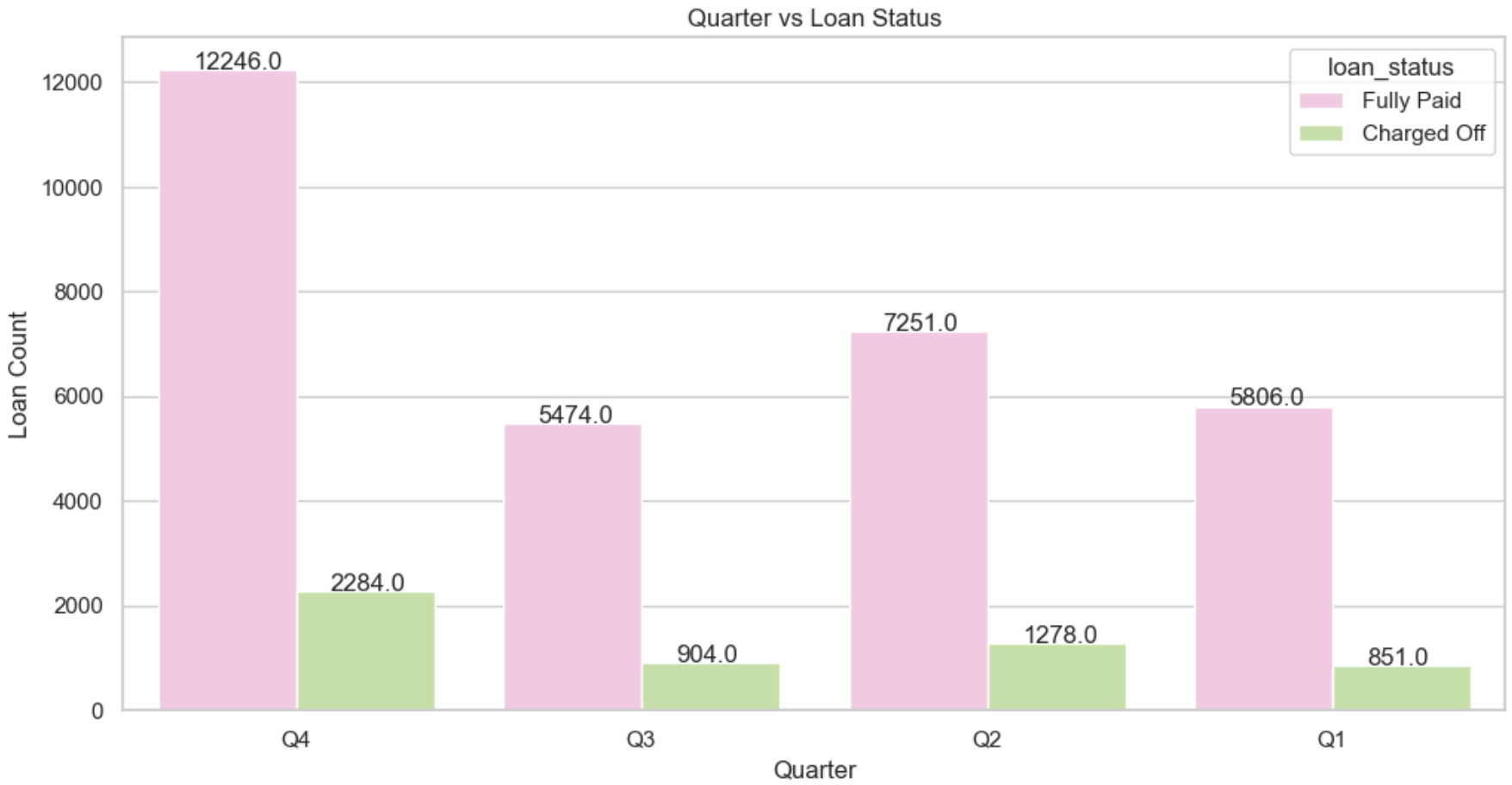
Year the Loan was given to Customer v/s Status of Loan

Bivariate Analysis (Ordered Categorical)



Month during which the Loan was given to Customer v/s Status of Loan

Bivariate Analysis (Ordered Categorical)



Quarter during which the Loan was given to Customer v/s Status of Loan

Bivariate Analysis (Categorical Variables)

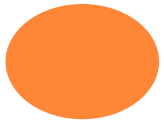
Observations:

A. **Ordered Categorical Variables:**

- ✓ The loan applicants belonging to Grades B, C, and D contribute to most of the "Charged Off" loans.
- ✓ Loan applicants belonging to Sub Grades B3, B4, and B5 are more likely to charge off.
- ✓ Loan applicants applying for loans with a 60-month term are more likely to default than those taking loans for 36 months.
- ✓ Most loan applicants have ten or more years of experience, and they are also the most likely to default.
- ✓ The number of loan applicants has steadily increased from 2007 to 2011, indicating a positive trend in the upcoming years.
- ✓ December is the most preferred month for taking loans, possibly due to the holiday season.
- ✓ The fourth quarter (Q4) is the most preferred quarter for taking loans, primarily because of the upcoming holiday season.

A. **Unordered Categorical Variables:**

- ✓ Debt consolidation is the category where the maximum number of loans are issued, and people have defaulted the most in the same category.
- ✓ Loan applicants who live in rented or mortgaged houses are more likely to default.
- ✓ Verified loan applicants are defaulting more than those who are not verified.
- ✓ Loan applicants from the states of California (CA), Florida (FL), and New York (NY) are most likely to default.



Bivariate Analysis (Categorical Variables)

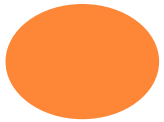
Inferences:

A. Ordered Categorical Variables:

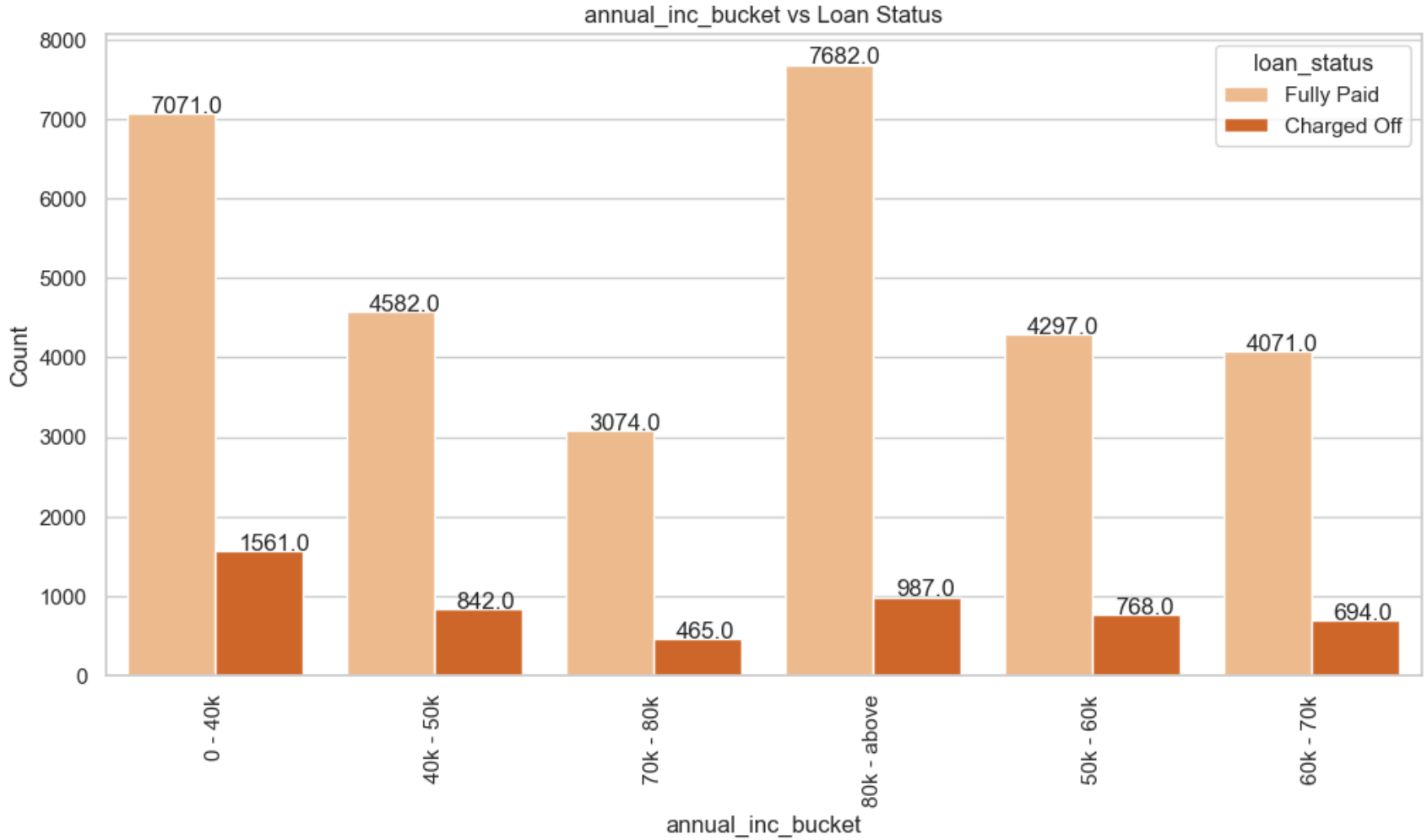
- ✓ **Risk Assessment for Grades B, C, and D:** Since loan applicants from Grades B, C, and D contribute to most of the "Charged Off" loans, the company should consider implementing stricter risk assessment and underwriting criteria for applicants falling into these grades.
- ✓ **Subgrades B3, B4, and B5:** Pay special attention to applicants with Subgrades B3, B4, and B5, as they are more likely to charge off. Implementing additional risk mitigation measures or offering them lower loan amounts could be considered.
- ✓ **Term Length**: Given that applicants opting for 60-month loans are more likely to default, the company should consider evaluating the risk associated with longer-term loans and potentially limiting the maximum term or adjusting interest rates accordingly.
- ✓ **Experience and Default Probability:** Loan applicants with ten or more years of experience are more likely to default. This suggests that experience alone may not be a reliable indicator of creditworthiness. The company should use a more comprehensive credit scoring system that factors in other risk-related attributes.
- ✓ **Positive Growth Trend:** The steady increase in the number of loan applicants from 2007 to 2011 indicates growth in the market. The company can capitalize on this trend by maintaining a competitive edge in the industry while keeping risk management practices robust.
- ✓ **Seasonal Trends:** December and Q4 are peak periods for loan applications, likely due to the holiday season. The company should anticipate increased demand during these periods and ensure efficient processing to meet customer needs

B. Unordered Categorical Variables:

- ✓ **Debt Consolidation Risk**: Since debt consolidation is the category with the maximum number of loans and high default rates, the company should carefully evaluate applicants seeking debt consolidation loans and potentially adjust interest rates or offer financial counseling services.
- ✓ **Housing Status and Default Risk**: Applicants living in rented or mortgaged houses are more likely to default. This information can be considered in the underwriting process to assess housing stability and its impact on repayment ability.
- ✓ **Verification Process**: Verified loan applicants are defaulting more than those who are not verified. The company should review its verification process to ensure it effectively assesses applicant creditworthiness and consider improvements or adjustments.
- ✓ **Geographic Risk**: Loan applicants from states like California (CA), Florida (FL), and New York (NY) are more likely to default. The company should monitor regional risk trends and adjust lending strategies or rates accordingly in these areas.

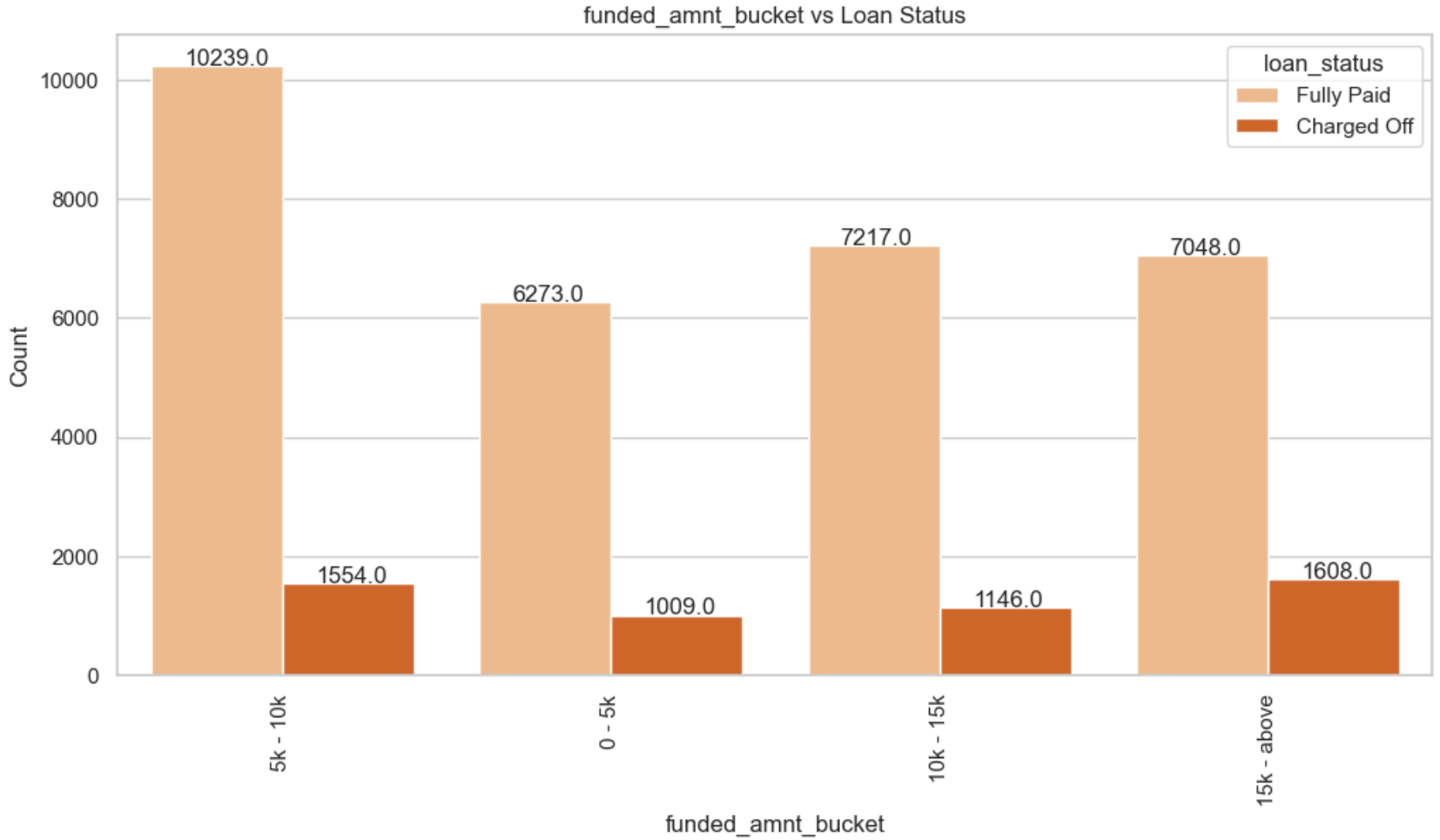


Bivariate Analysis (Quantitative Variables)



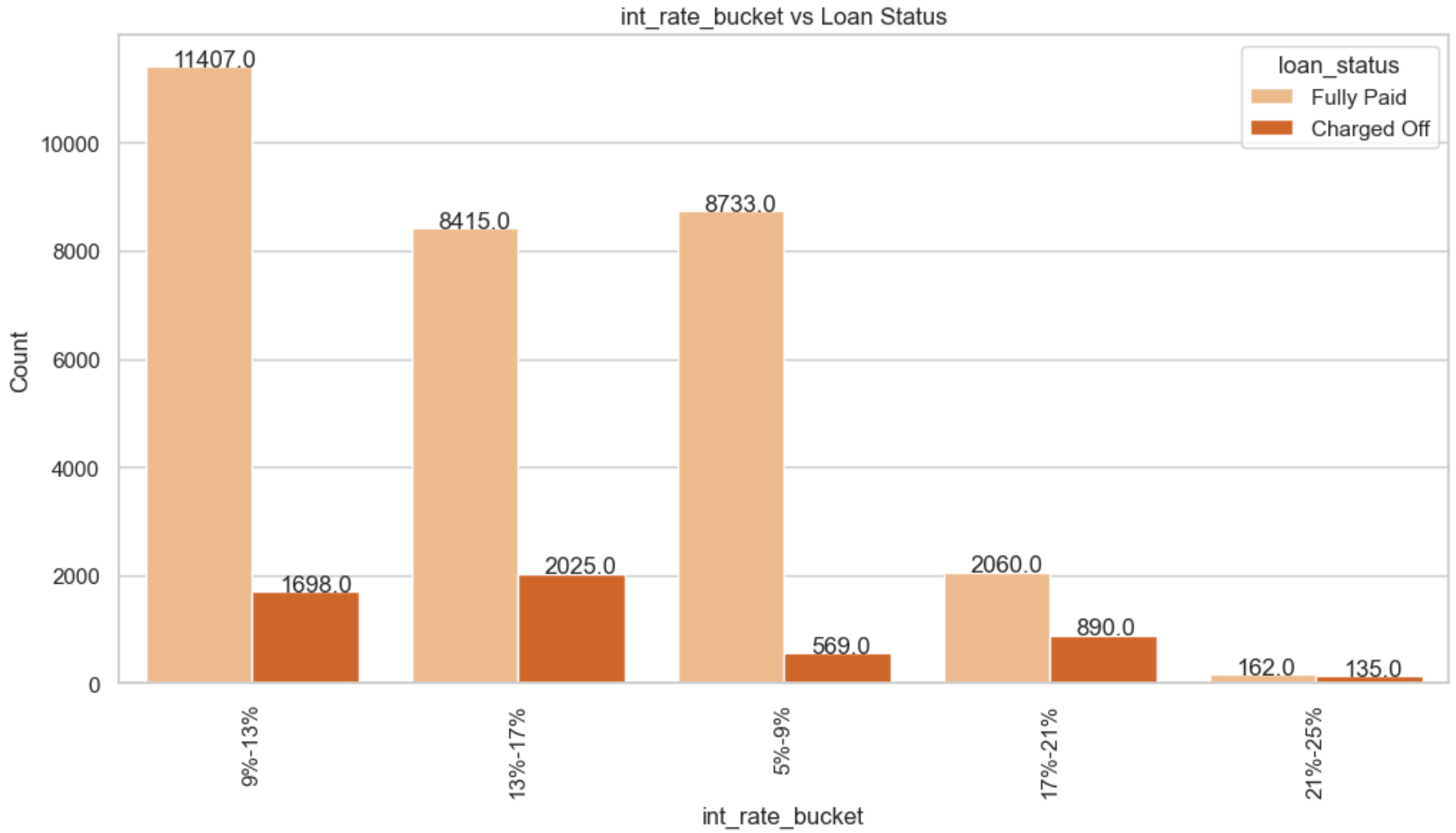
Bucket of Annual Income v/s Status of Loan

Bivariate Analysis (Quantitative Variables)



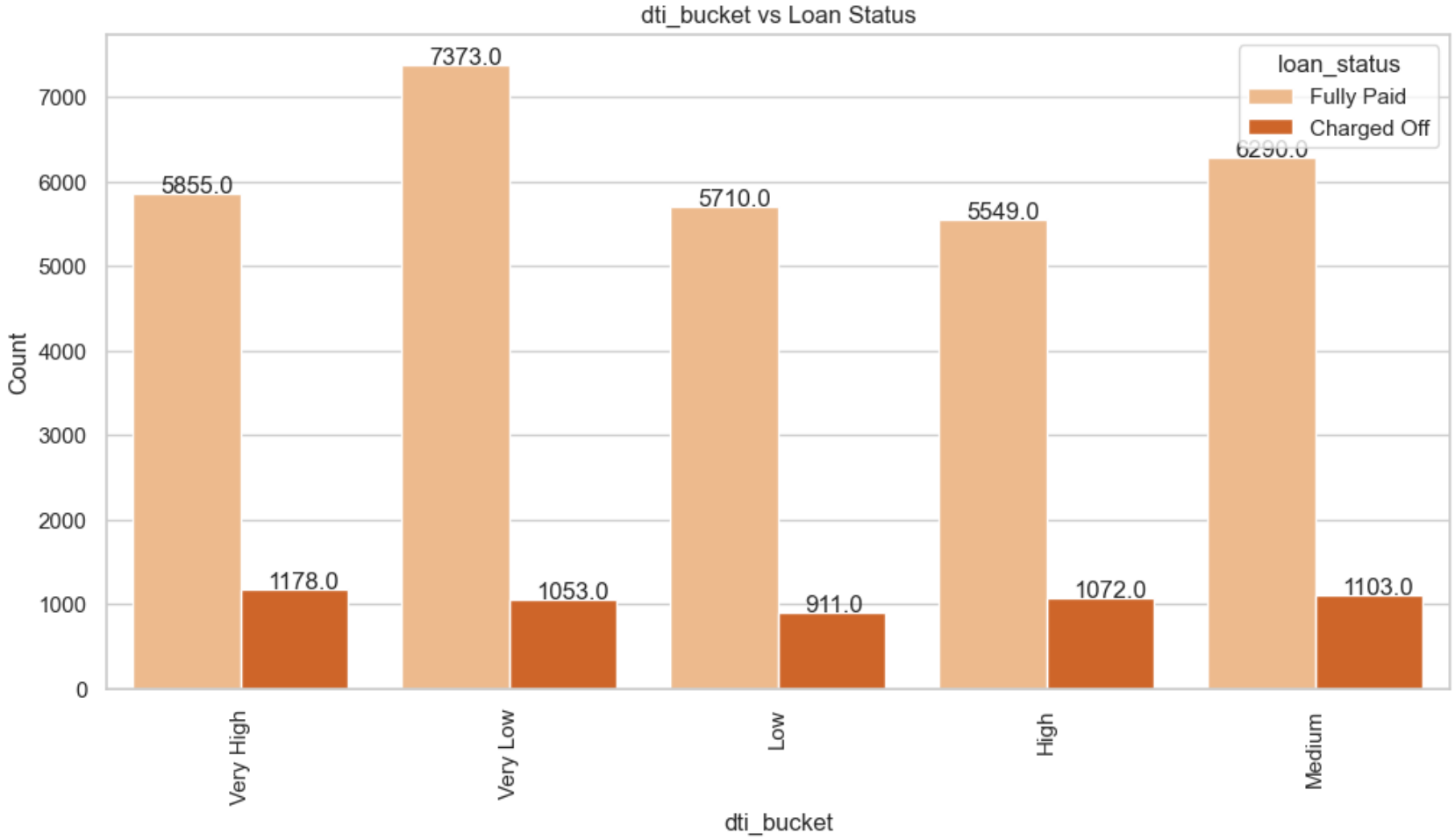
Bucket of Amount which was Funded v/s Status of Loan

Bivariate Analysis (Quantitative Variables)



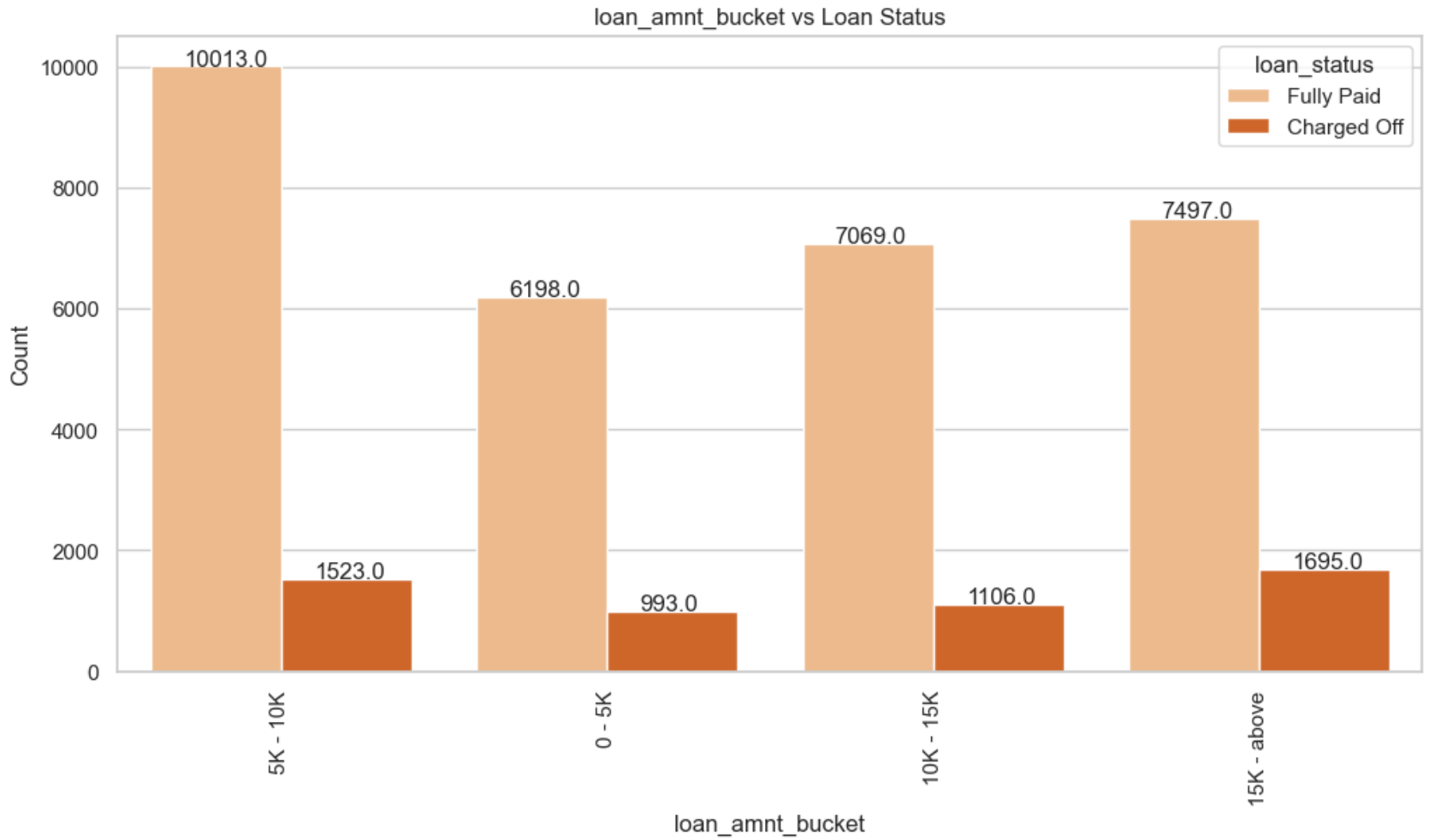
Bucket of Interest Rate of the loan v/s Status of Loan

Bivariate Analysis (Quantitative Variables)



Bucket of Debt to Income Ratio of the Customer v/s Status of Loan

Bivariate Analysis (Quantitative Variables)



Bucket of Loan Amount associated with the customer v/s Status of Loan

Bivariate Analysis (Quantitative Variables)

Observations:

- ✓ A majority of the loan applicants who defaulted received loan amounts of $15,000 or higher.
- ✓ The majority of loan applicants who charged off had significantly high Debt-to-Income (DTI) ratios.
- ✓ A significant portion of loan applicants who defaulted received loans with interest rates falling within the range of 13% to 17%.
- ✓ A majority of the loan applicants who charged off reported an annual income of less than $40,000.

Inferences:

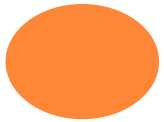
- ✓ **High Loan Amounts**: Applicants receiving loan amounts of $15,000 or higher are more likely to default. The company can mitigate this risk by conducting more thorough assessments for larger loan requests and potentially capping loan amounts for higher-risk applicants.

- ✓ **DTI and Interest Rates**: High Debt-to-Income (DTI) ratios and interest rates in the 13%-17% range are associated with defaults. The company should review its interest rate determination process and consider adjusting rates based on DTI ratios to better align with the borrower's ability to repay.

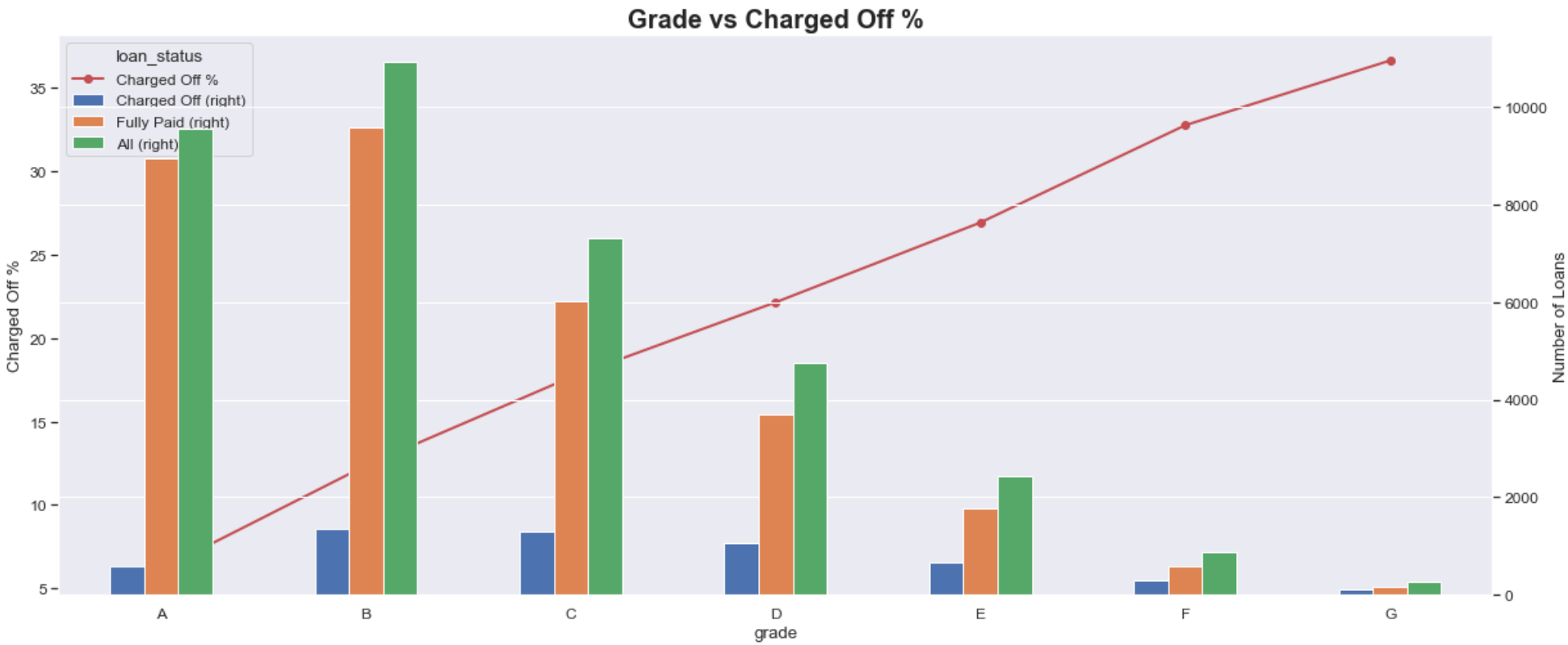
- ✓ **Low Annual Income**: Applicants with annual incomes less than $40,000 have a higher likelihood of defaulting. The company should consider offering financial education resources or setting maximum loan amounts based on income levels to ensure affordability for borrowers.

Multivariate Analysis

- **Multivariate analysis** is a statistical technique used to analyze data that involves more than two variables.
- Unlike univariate analysis (which deals with one variable) and bivariate analysis (which deals with two variables), multivariate analysis examines the relationships between multiple variables simultaneously.
- It is widely used in various fields such as economics, social sciences, biology, marketing, and environmental science.
- Multivariate analysis can include different types of variables, such as categorical variables, numerical variables, or a combination of both.

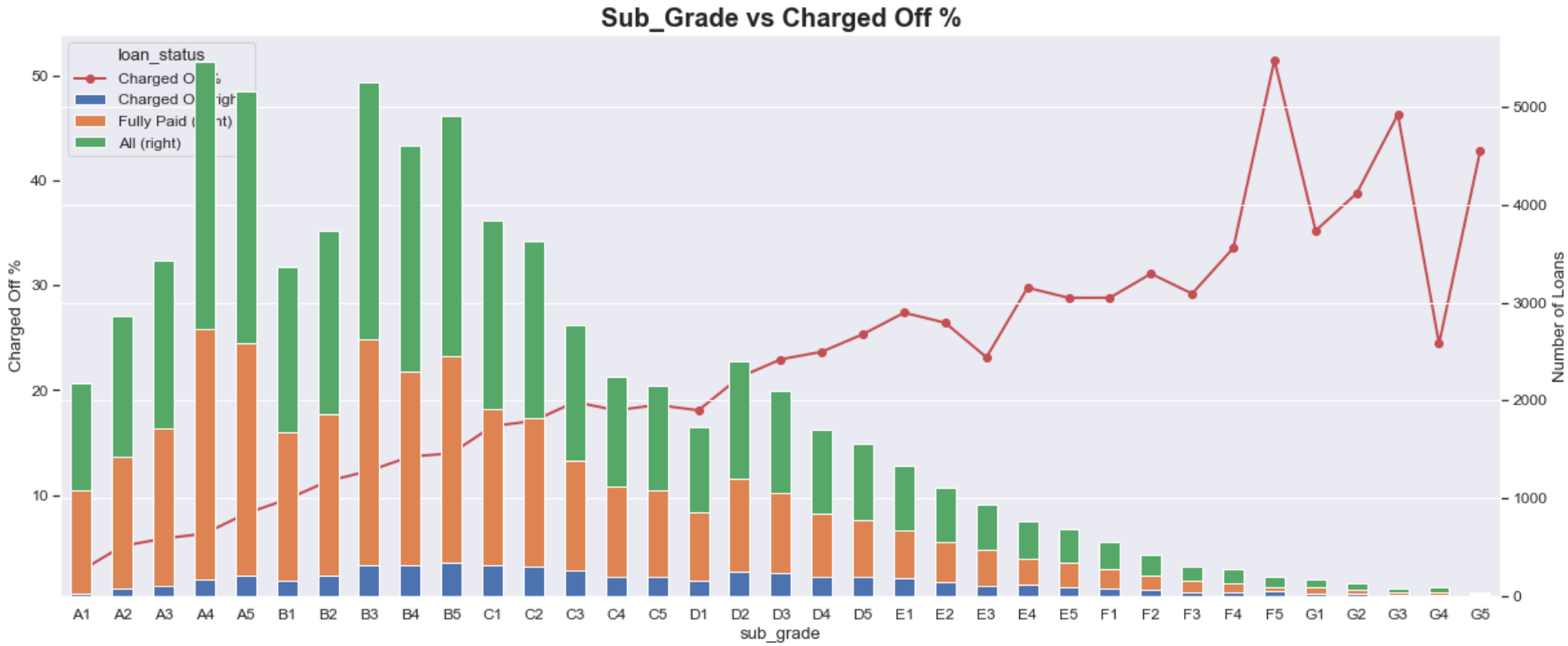


Multivariate Analysis



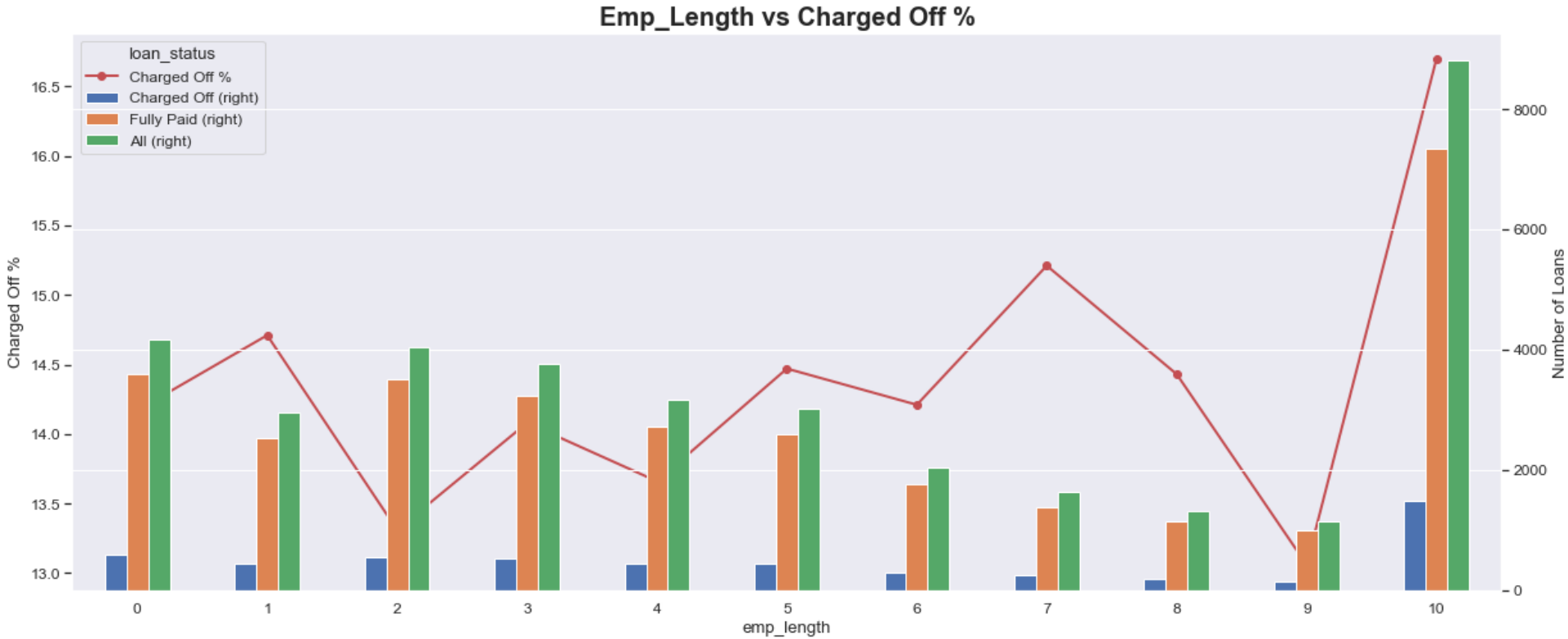
Grade v/s Percentage of Charged-off Loans

Multivariate Analysis



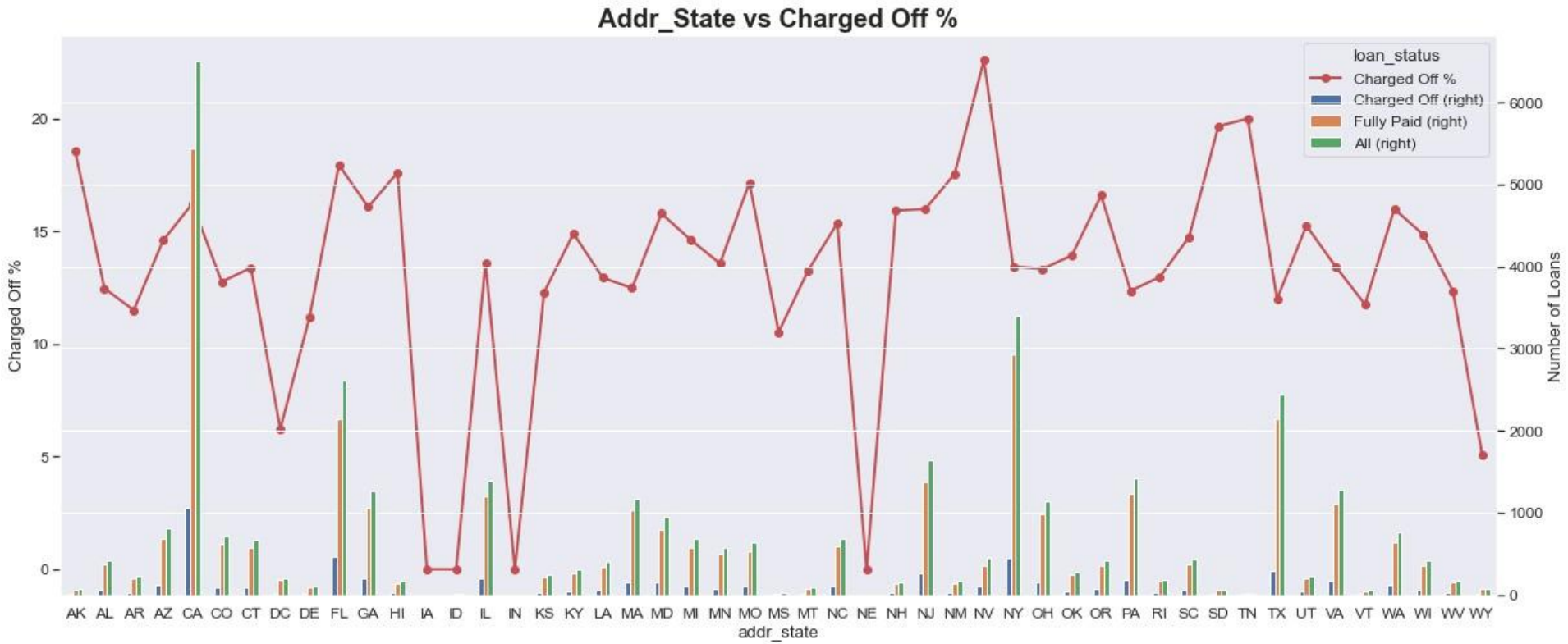
Sub-Grade v/s Percentage of Charged-off Loans

Multivariate Analysis



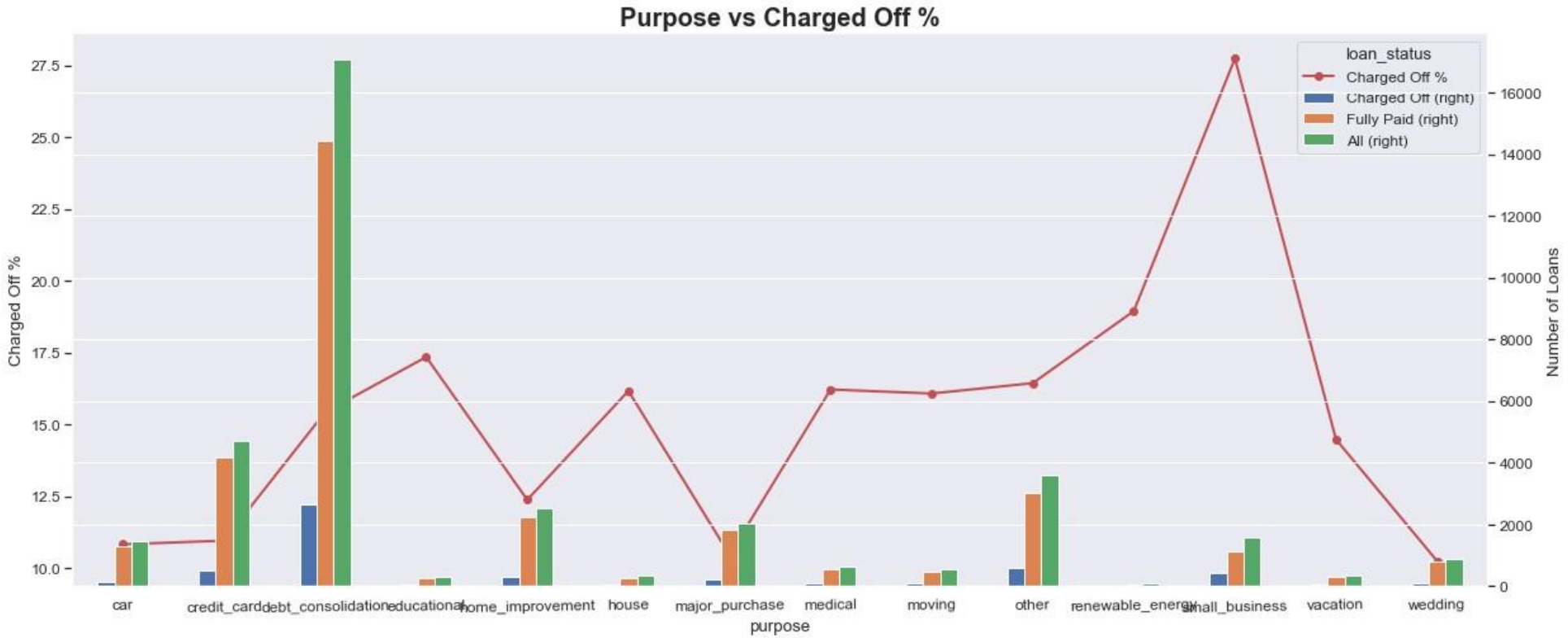
Employment Length (In Years) v/s
Percentage of Charged-off Loans

Multivariate Analysis



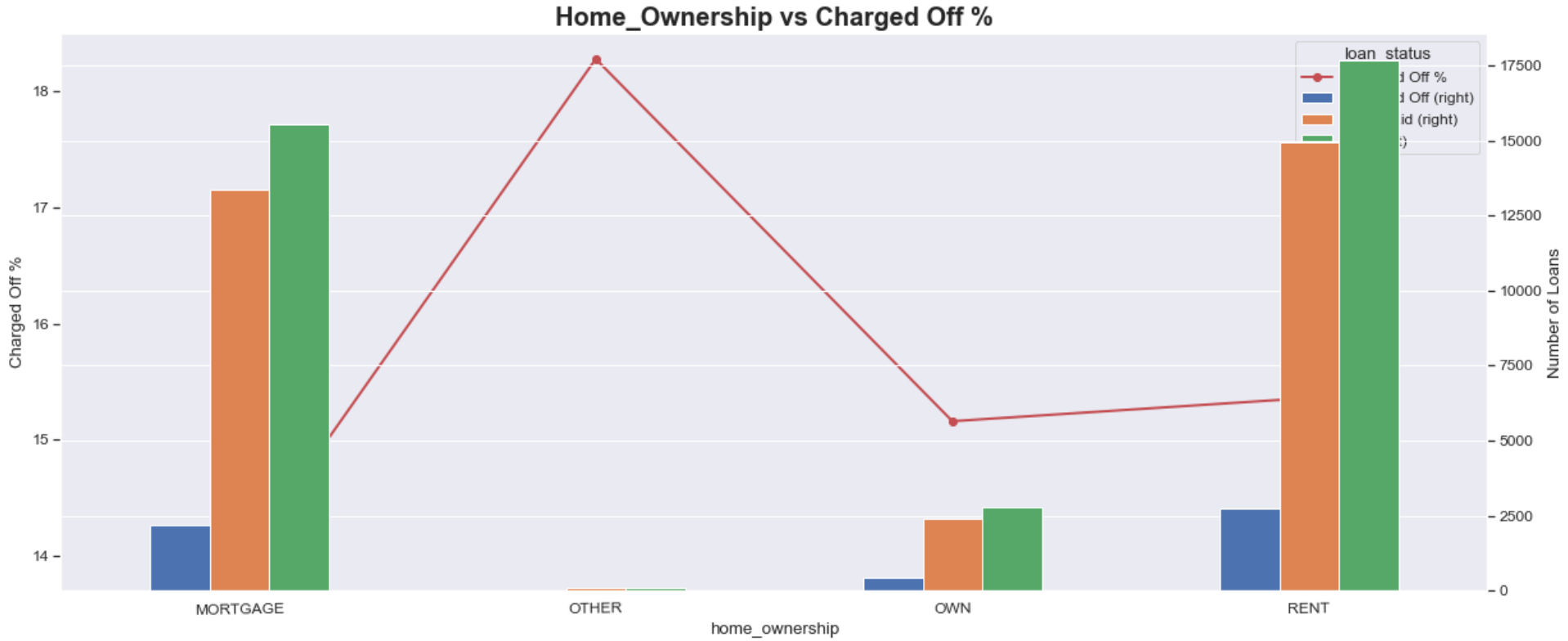
Address State v/s Percentage of Charged-off Loans

Multivariate Analysis



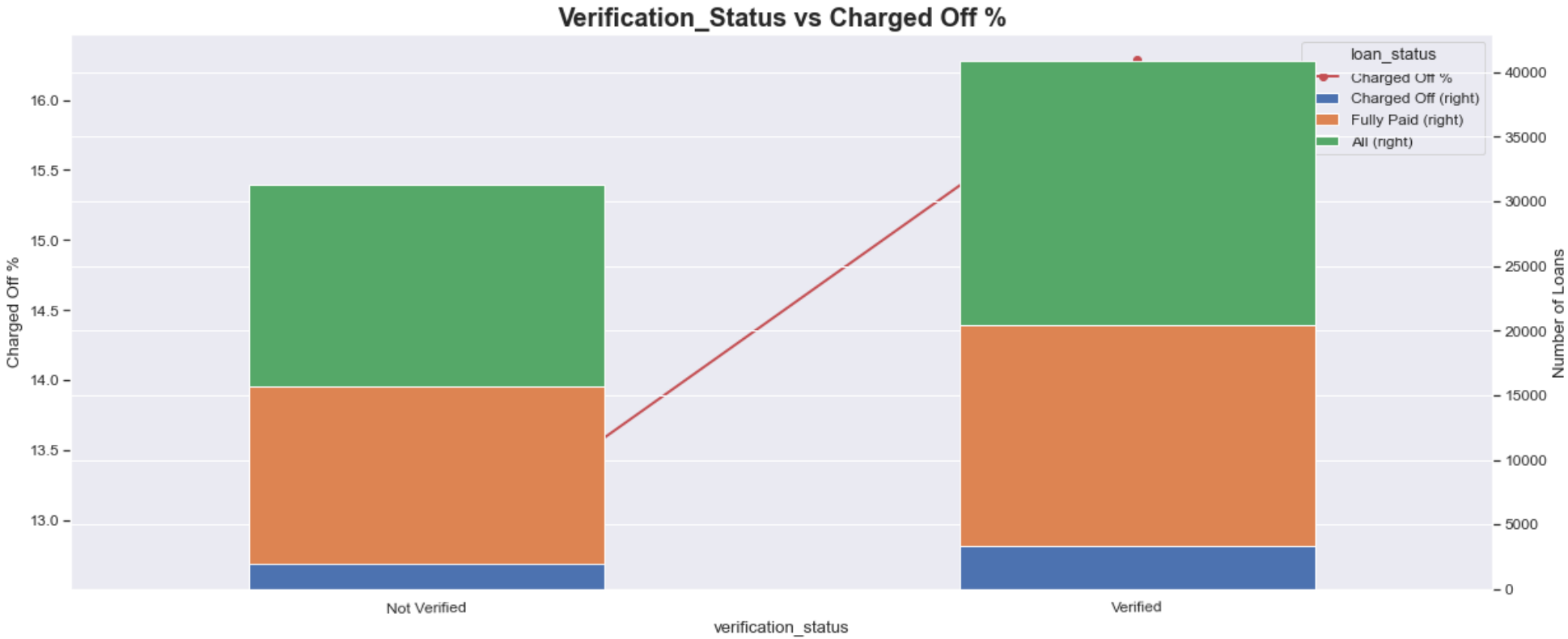
Purpose of Loan v/s Percentage of Charged-off Loans

Multivariate Analysis



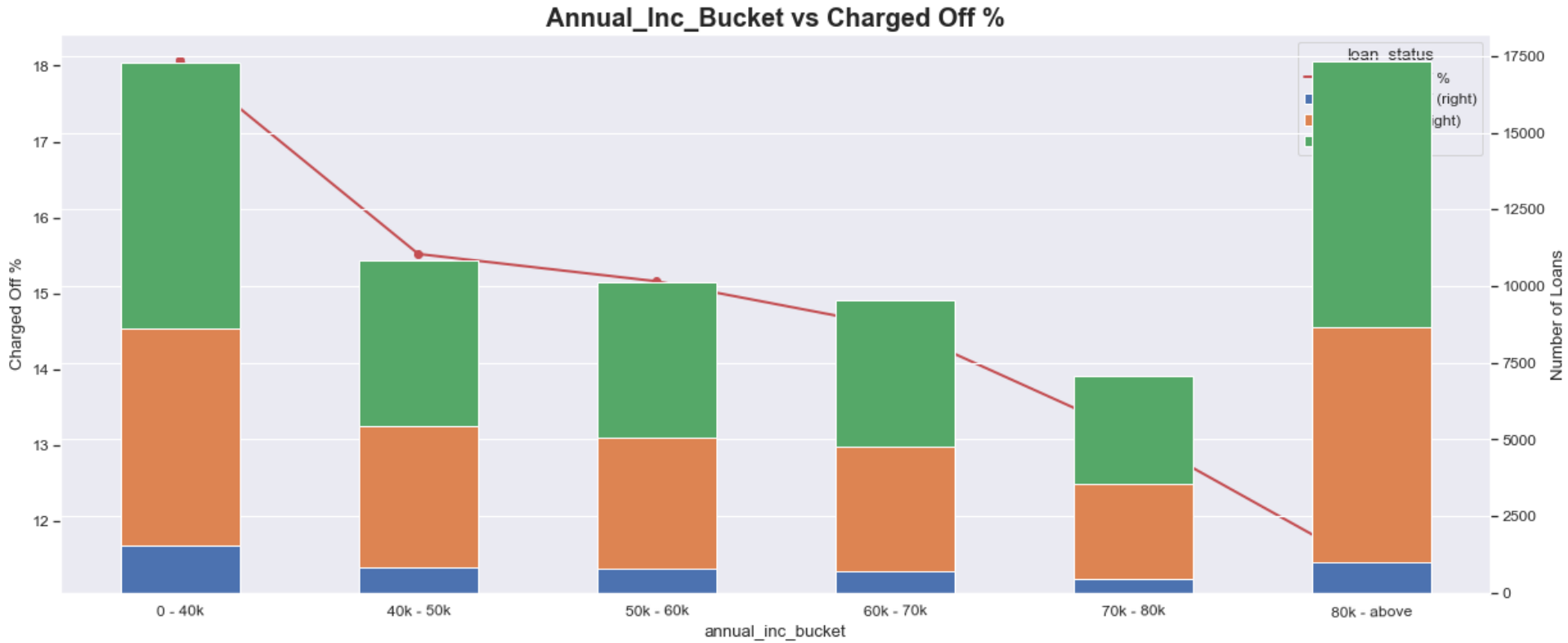
Home Ownership v/s Percentage of Charged-off Loans

Multivariate Analysis



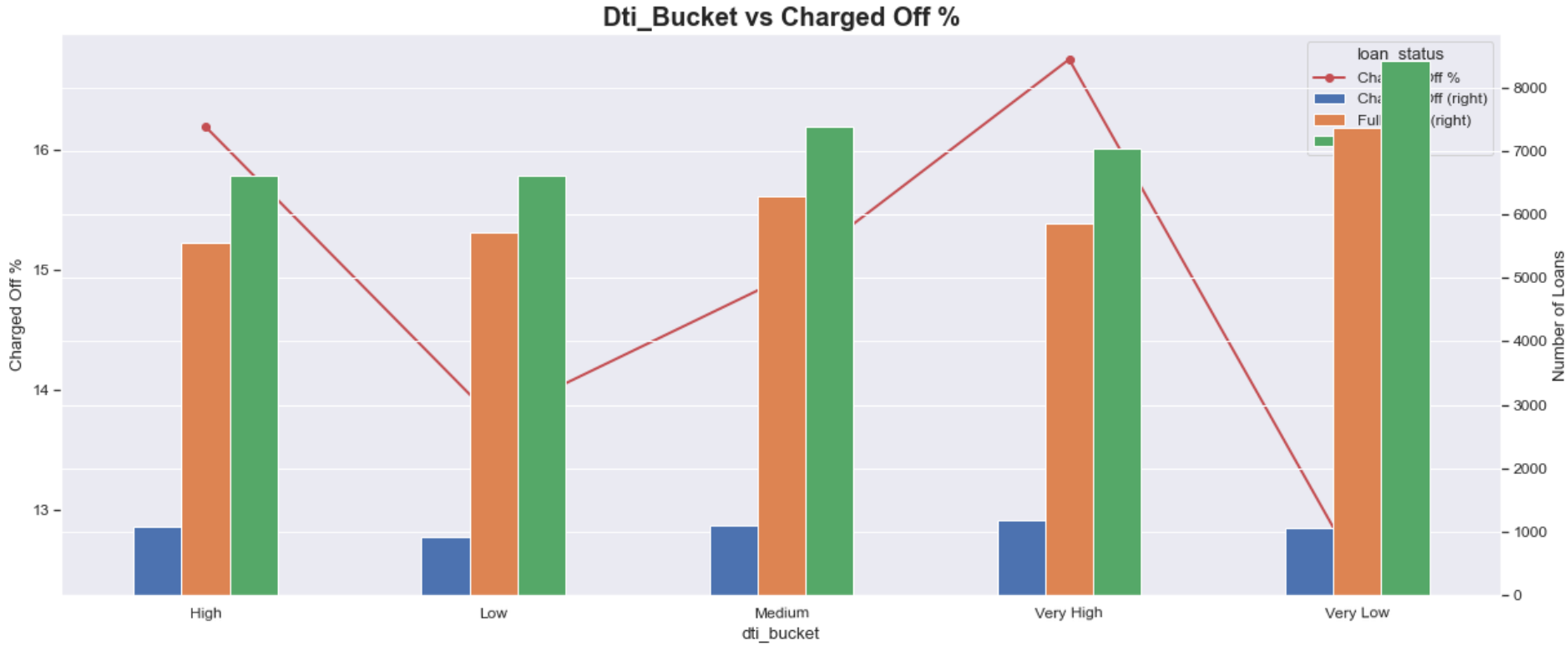
Verification Status of Loan v/s
Percentage of Charged-off Loans

Multivariate Analysis



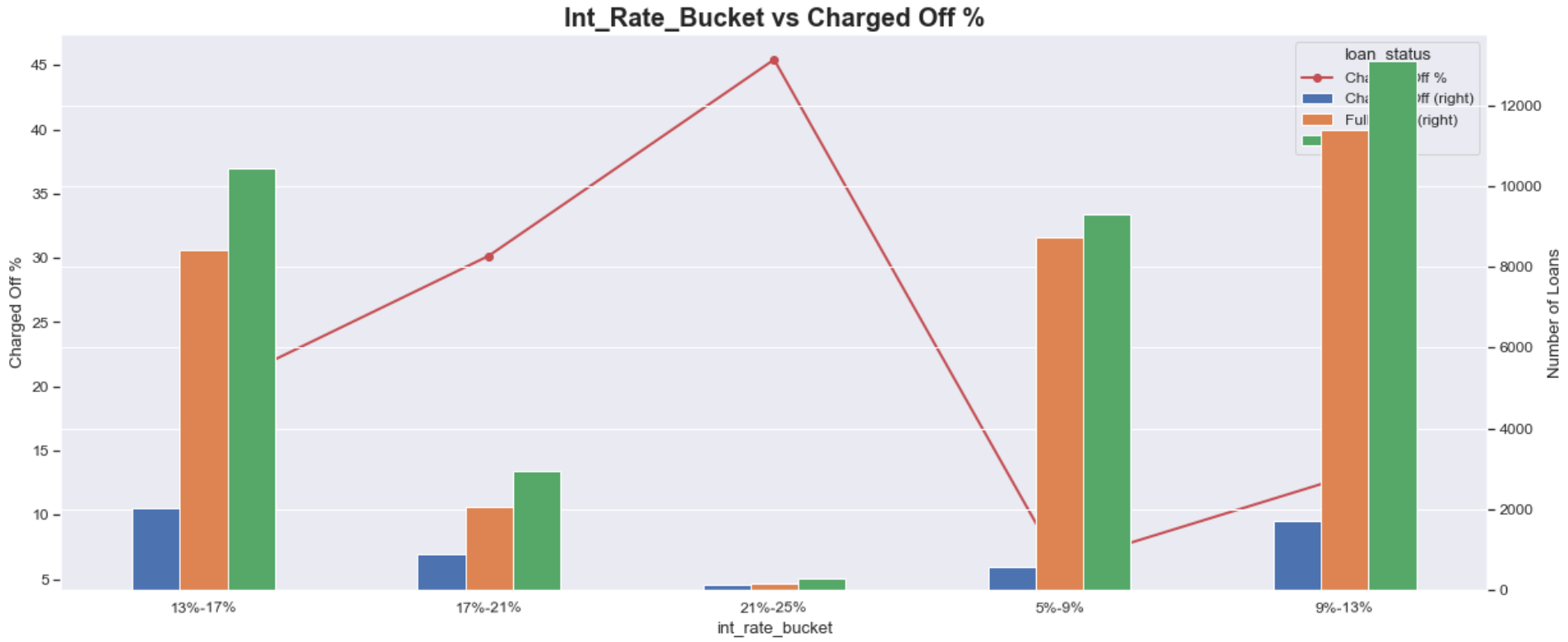
Buckets of Annual Income v/s
Percentage of Charged-off Loans

Multivariate Analysis



Buckets of Debt to Income Ratio (DTI)
v/s Percentage of Charged-off Loans

Multivariate Analysis

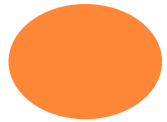


Buckets of Interest Rate v/s Percentage of Charged-off Loans

Multivariate Analysis

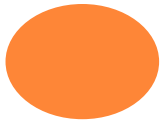
Observations & Inferences:

- ✓ Tendency to default the loan is likely with loan applicants belonging to B, C, D grades.
- ✓ Borrowers from sub grade B3, B4 and B5 have maximum tendency to default.
- ✓ Loan applicants with 10 years of experience has maximum tendency to default the loan.
- ✓ Borrowers from states CA, FL, NJ have maximum tendency to default the loan.
- ✓ Borrowers from Rented House Ownership have highest tendency to default the loan.
- ✓ The borrowers who are in lower income groups have maximum tendency to default the loan and it generally decreases with the increase in the annual income.
- ✓ The tendency to default the loan is increasing with increase in the interest rate.

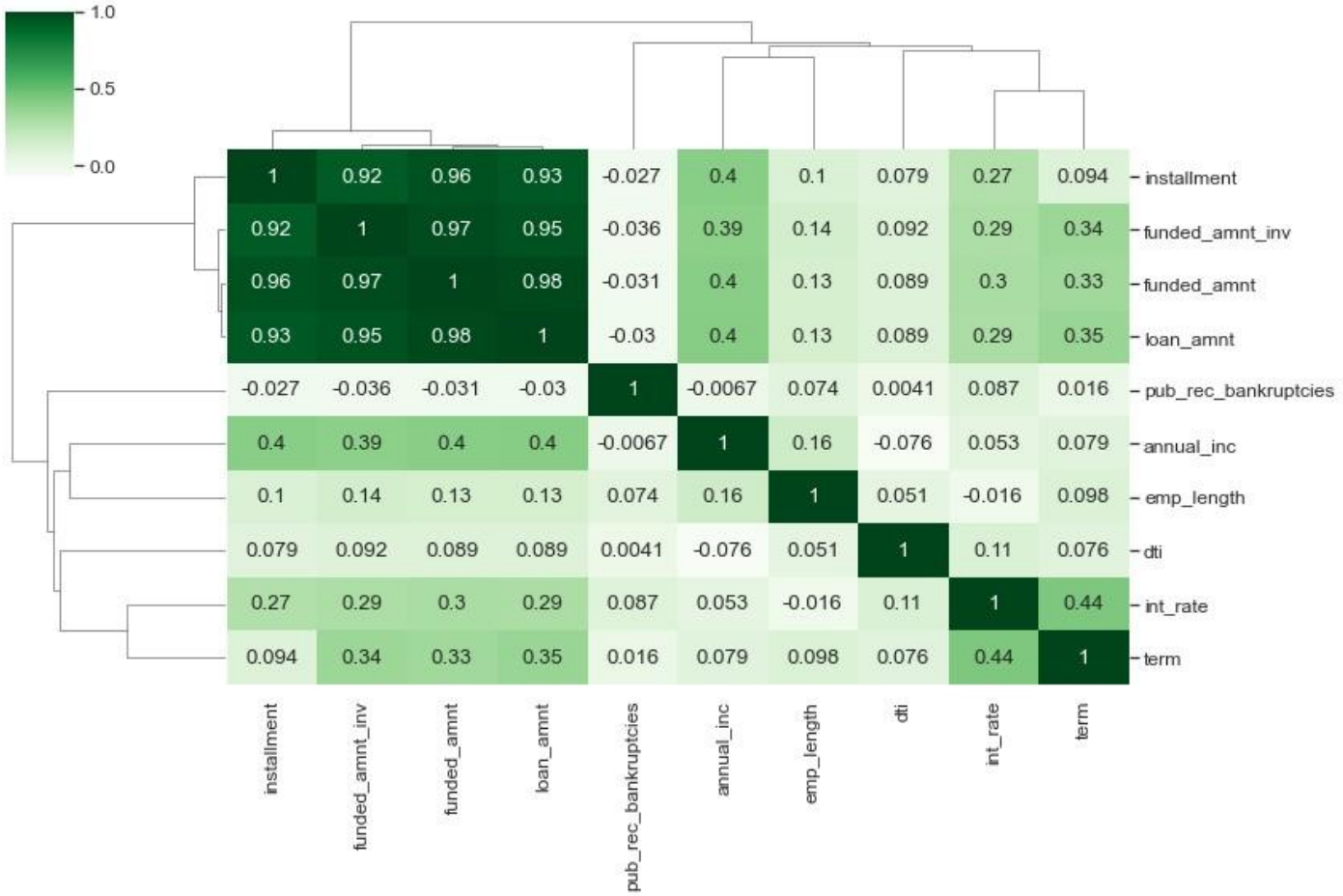


Correlation Analysis

- Correlation analysis is a statistical technique used to measure the strength and direction of the relationship between two or more variables.
- It quantifies the degree to which changes in one variable are associated with changes in another variable.
- Correlation analysis is widely used in various fields, including finance, economics, biology, psychology, and social sciences, to understand patterns and relationships in data.
- It ranges from **-1 to 1**.
  - **r=1:** indicates a perfect positive correlation
  - **r=−1:** indicates a perfect negative correlation
  - **r=0:** indicates no correlation between the variables



Correlation Analysis



Correlation Matrix among variables namely **installment, funded_amnt_inv, funded_amnt, loan_amnt, pub_rec_bankrupticies, annual_inc, emp_length, dti, int_rate , term**

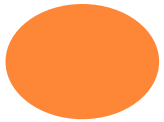
Correlation Analysis

Observations & Inferences:

- **Strong Correlation:**
  - ✓ **installment** has a strong correlation with **funded_amnt, loan_amnt, and funded_amnt_inv**
  - ✓ **term** has a strong correlation with **interest rate**
  - ✓ **annual_inc** has a strong correlation with **loan_amount**

- **Weak Correlation:**
  - ✓ **dti** has weak correlation with most of the fields
  - ✓ **emp_length** has weak correlation with most of the fields

- **Negative Correlation:**
  - ✓ **pub_rec_bankrupticies** has a negative correlation with almost every field
  - ✓ **annual_inc** has a negative correlation with **dti**



References & Useful Links

GitHub Repository Link:

https://github.com/Adityaajain/LendingClubCaseStudy

Thank You!