Contents

Problem Statement

- **Lending Club**, a Consumer Finance marketplace specializing in offering a variety of loans to urban customers, faces a critical challenge in managing its loan approval process. When evaluating loan applications, the company must make sound decisions to minimize financial losses, primarily stemming from loans extended to applicants who are considered **"Risky".**
- These financial losses, referred to as **Credit Losses**, occur when borrowers fail to repay their loans or default. In simpler terms, borrowers labeled as **"Charged-Off"** are the ones responsible for the most significant losses to the company.
- The primary objective of this exercise is to assist Lending Club in mitigating credit losses. This challenge arises from two potential scenarios:

    1. **Identifying applicants likely to repay their loans is crucial, as they can generate profits for the company through interest payments. Rejecting such applicants would result in a loss of potential business.**
    2. **On the other hand, approving loans for applicants not likely to repay and at risk of default can lead to substantial financial losses for the company.**

- The objective is to pinpoint applicants at risk of defaulting on loans, enabling a reduction in credit losses. This case study aims to achieve this goal through Exploratory Data Analysis (EDA) using the provided dataset.
- In essence, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Data Description

Lending Club provided us with customer's historical data. This dataset contained information pertaining to the borrower's past credit history and Lending Club loan information. The total dataset consisted of over 39717 records and 111 columns, which was sufficient for our team to conduct analysis. Variables present within the dataset provided an ample amount of information which we could use to identify relationships and gauge their effect upon the success or failure of a borrower fulfilling the terms of their loan agreement.

LoanStatNew	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
all_util	Balance to credit limit on all trades
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
·ded_amnt	The total amount committed to that loan at that point in time.
·l_amnt_inv	The total amount committed by investors for that loan at that point in time.
	LC assigned loan grade
·ship	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.

We required only the variables that had a direct or indirect response to a borrower's potential to default. To achieve this, we prepared the data by choosing select variables that would best fit this criteria.

Data Understanding

Dataset Attributes:

Primary Attribute

Loan Status: The Principal Attribute of Interest (loan_status). This column consists of three distinct values:

- ✓ **Fully-Paid:** Signifies customers who have successfully repaid their loans.
- ✓ **Charged-Off:** Indicates customers who have been labeled as "Charged-Off" or have defaulted on their loans.
- ✓ **Current:** Represents customers whose loans are presently in progress and, thus, cannot provide conclusive evidence regarding future defaults.

For the purposes of this case study, rows with a "Current" status will be excluded from the analysis.

Decision Matrix:

Loan Acceptance Outcome - There are three potential scenarios:

Fully Paid - This category represents applicants who have successfully repaid both the principal and the interest rate of the loan.

Current - Applicants in this group are actively in the process of making loan installments; hence, the loan tenure has not yet concluded. These individuals are not categorized as 'defaulted.'

Charged-off - This classification pertains to applicants who have failed to make timely installments for an extended period, resulting in a 'default' on the loan.

Loan Rejection - In cases where the company has declined the loan application (usually due to the candidate not meeting their requirements), there is no transactional history available for these applicants. Consequently, this data is unavailable to the company and is not included in this dataset.

Data Understanding

Key Columns of Significance:

The provided columns serve as pivotal attributes, often referred to as predictors. These attributes, available during the loan application process, significantly contribute to predicting whether a loan will be approved or rejected. It's important to note that some of these columns may be excluded due to missing data in the dataset.

- **Customer Demographics:**
  - **Annual Income (annual_inc):** Reflects the customer's annual income. Typically, a higher income enhances the likelihood of loan approval.
  - **Home Ownership (home_ownership):** Indicates whether the customer owns a home or rents. Home ownership provides collateral, thereby increasing the probability of loan approval.
  - **Employment Length (emp_length):** Represents the customer's overall employment tenure. Longer tenures signify greater financial stability, leading to higher chances of loan approval.
  - **Debt to Income (dti):** Measures how much of a person's monthly income is already being used to pay off their debts. A lower DTI translates to a higher chance of loan approval.
  - **State (addr_state):** Denotes the customer's location and can be utilized for creating a generalized demographic analysis. It may reveal demographic trends related to delinquency or default rates.
- **Loan Characteristics:**
  - **Loan Amount (loan_amt):** Represents the amount of money requested by the borrower as a loan.
  - **Grade (grade):** Represents a rating assigned to the borrower based on their creditworthiness, indicating the level of risk associated with the loan.
  - **Term (term):** Duration of the loan, typically expressed in months.
  - **Loan Date (issue_d):** Date when the loan was issued or approved by the lender.
  - **Purpose of Loan (purpose):** Indicates the reason for which the borrower is seeking the loan, such as debt consolidation, home improvement, or other purposes.
  - **Verification Status (verification_status):** Represents whether the borrower's income and other information have been verified by the lender.
  - **Interest Rate (int_rate):** Represents the annual rate at which the borrower will be charged interest on the loan amount.
  - **Installment (installment):** Represents the regular monthly payment the borrower needs to make to repay the loan, including both principal and interest.
  - **Public Records (public_rec):** Refers to derogatory public records, which contribute to loan risk. A higher value in this column reduces the likelihood of loan approval.
  - **Public Records Bankruptcy (public_rec_bankruptcy):** Indicates the number of locally available bankruptcy records for the customer. A higher value in this column is associated with a lower success rate for loan approval.

Data Understanding

Excluded Columns:

In our analysis, we will not consider certain types of columns. It's important to note that this is a general categorization of the columns we will exclude from our approach, and it does not represent an exhaustive list.

- **Customer Behavior Columns** - Columns that describe customer behavior will not be factored into our analysis. The current analysis focuses on the loan application stage, while customer behavior variables pertain to post-approval actions. Consequently, these attributes will not influence the loan approval/rejection process.
- **Granular Data** - Columns providing highly detailed information that may not be necessary for our analysis will be omitted. For example, while the "grade" column may have relevance in creating business outcomes and visualizations, the "sub grade" column is excessively granular and will not be utilized in our analysis.
- **54** columns contain **NA** values only, and these columns will be removed namely **acc_open_past_24mths, all_util, annual_inc_joint, avg_cur_bal, bc_open_to_buy, bc_util, dti_joint, il_util, inq_fi, inq_last_12m, max_bal_bc, mo_sin_old_il_acct, mo_sin_old_rev_tl_op, mo_sin_rcnt_rev_tl_op, mo_sin_rcnt_tl, mort_acc, mths_since_last_major_derog, mths_since_rcnt_il, mths_since_recent_bc, mths_since_recent_bc_dlq, mths_since_recent_inq, mths_since_recent_revol_delinq, num_accts_ever_120_pd, num_actv_bc_tl, num_actv_rev_tl, num_bc_sats, num_bc_tl, num_il_tl, num_op_rev_tl, num_rev_accts, num_rev_tl_bal_gt_0, num_sats, num_tl_120dpd_2m, num_tl_30dpd, num_tl_90g_dpd_24m, num_tl_op_past_12m, open_acc_6m, open_il_12m, open_il_24m, open_il_6m, open_rv_12m, open_rv_24m, pct_tl_nvr_dlq, percent_bc_gt_75, tot_coll_amt, tot_cur_bal, tot_hi_cred_lim, total_bal_ex_mort, total_bal_il, total_bc_limit, total_cu_tl, total_il_high_credit_limit, total_rev_hi_lim, verification_status_joint**
- Certain columns contain only **0** values, and these columns will also be dropped.
- **9** Columns with **single value** that do not contribute to the analysis will be removed.
- Columns with values that are single value but have other values as NA will be treated as constant and dropped.
- Columns with more than **65%** of data being empty (**mths_since_last_delinq, mths_since_last_record**) will be dropped.
- Columns (**id, member_id**) will be dropped as they are index variables with unique values and do not contribute to the analysis.
- Columns (**emp_title, desc, title**) will be dropped as they contain descriptive text (nouns) and do not contribute to the analysis.
- The redundant column (**url**) will be dropped. Further analysis reveals that the URL is a static path with the loan ID appended as a query, making it redundant compared to the (**id**) column.
- **660** records for **pub_rec_bankruptcies** are dropped due to **missing values**
- These columns capture customer behavior recorded after loan approval and are not available at the time of loan approval. Thus, these variables will not be included in the analysis.
- Columns to be dropped: (**delinq_2yrs, earliest_cr_line, inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, 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_d, last_pymnt_amnt, last_credit_pull_d, application_type**)

Data Cleaning & Pre-processing

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 & Pre-processing

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 & Pre-processing

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 & Pre-processing

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

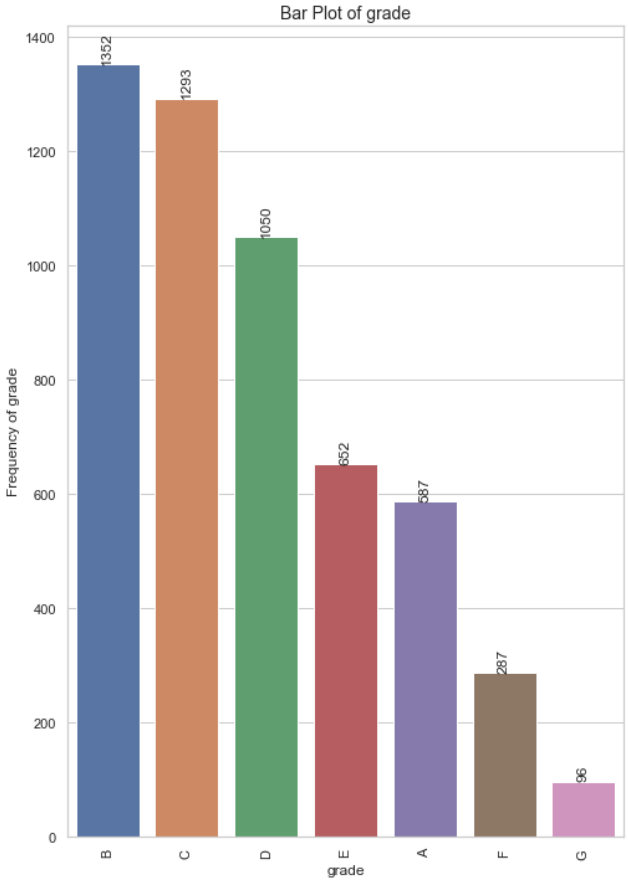
A. **Categorical Variables:**

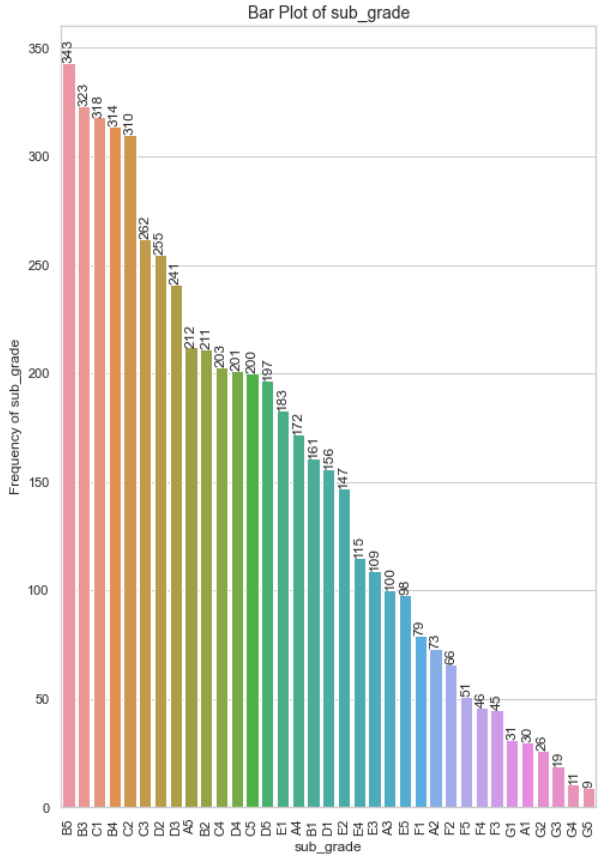
Ordered	Unordered
✓ Grade (grade) ✓ Sub grade (sub_grade) ✓ Term (36 / 60 months) (term) ✓ Employment length (emp_length) ✓ Issue year (issue_y) ✓ Issue month (issue_m) ✓ Issue quarter (issue_q)	✓ Address State (addr_state) ✓ Loan purpose (purpose) ✓ Home Ownership (home_ownership) ✓ Loan status (loan_status)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

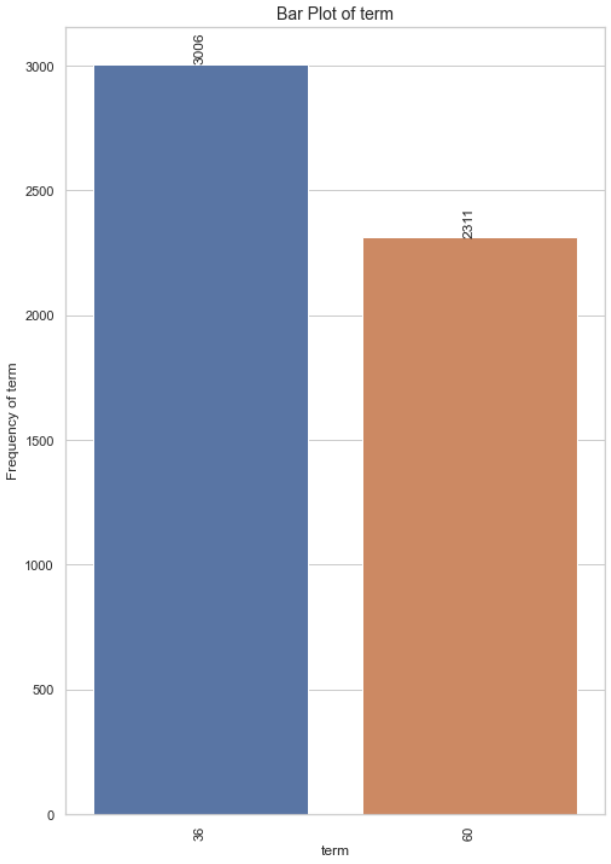
Univariate Analysis (Unordered Categorical)

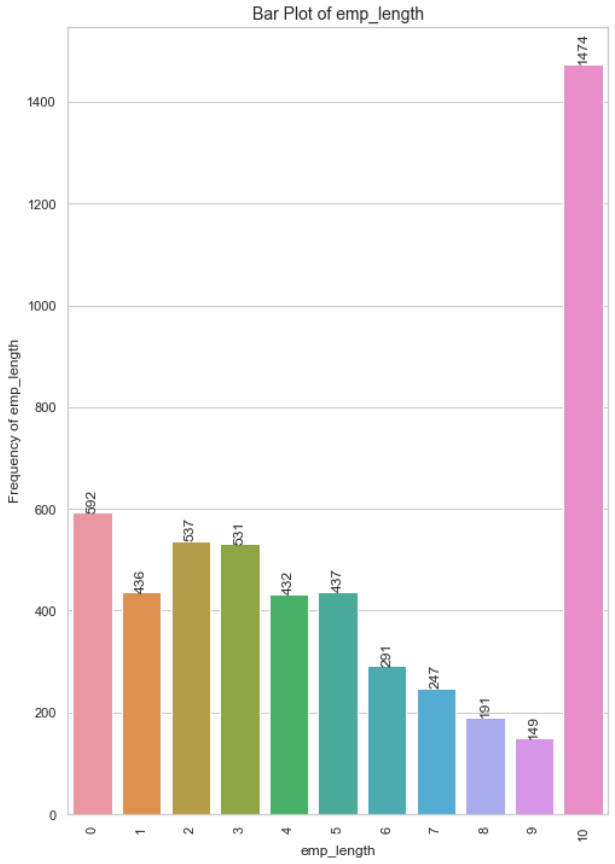




Grade and Sub-Grade

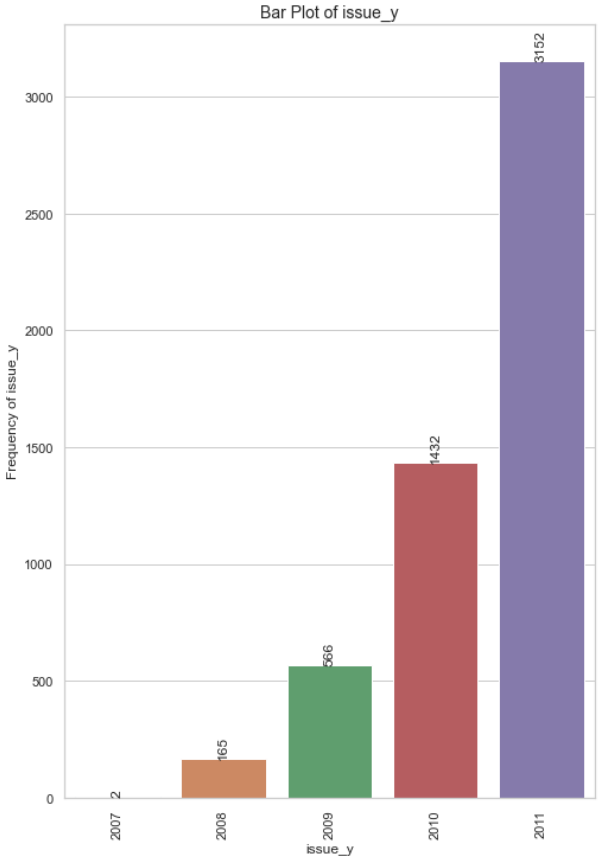
Univariate Analysis (Unordered Categorical)

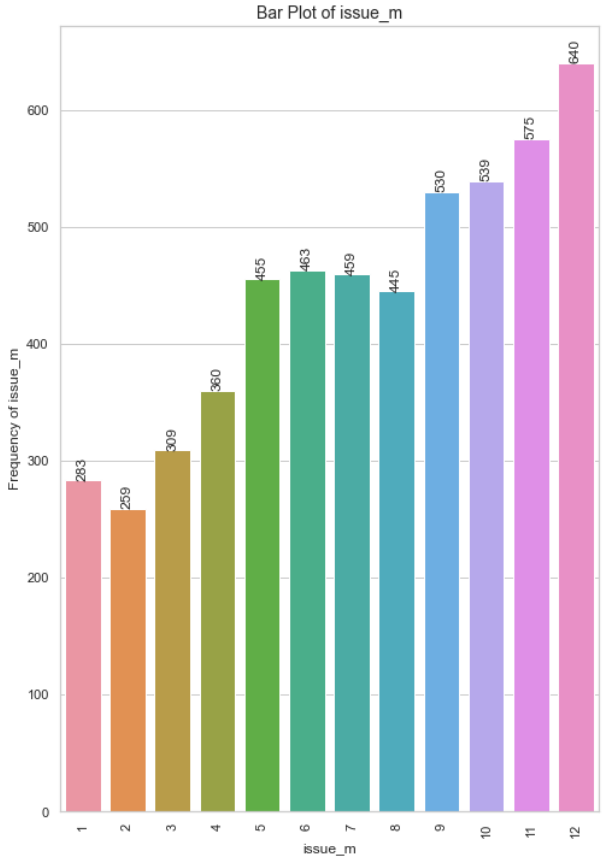




Term and Employment Length

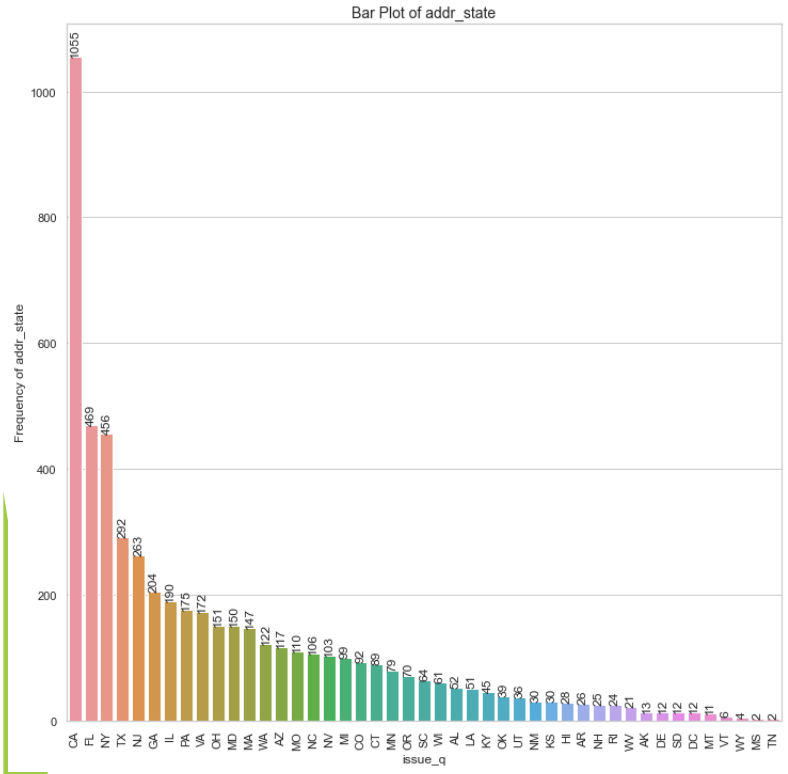
Univariate Analysis (Unordered Categorical)

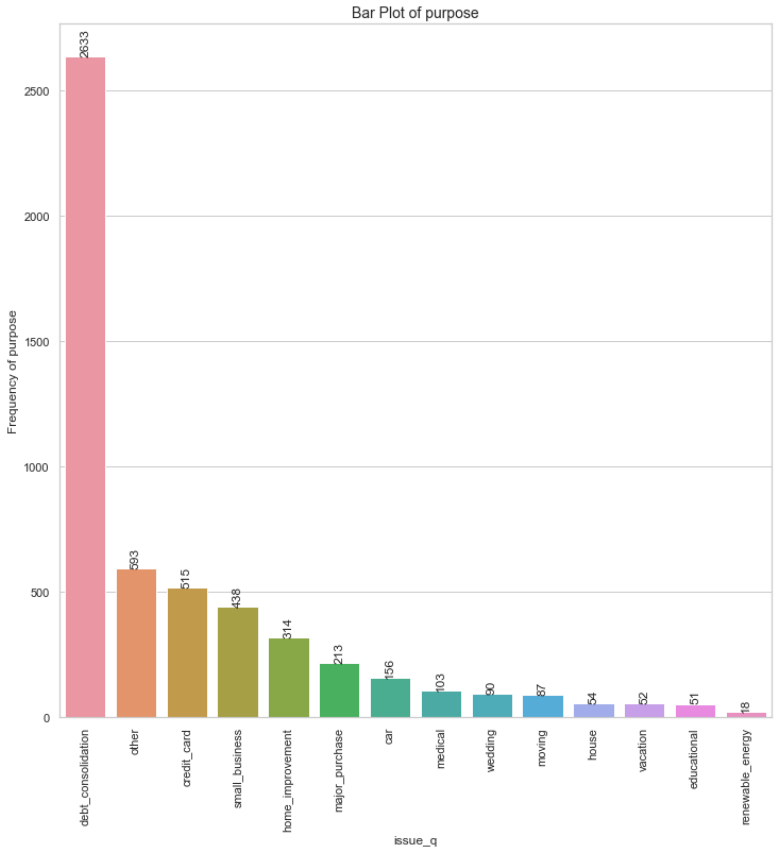




Term and Employment Length

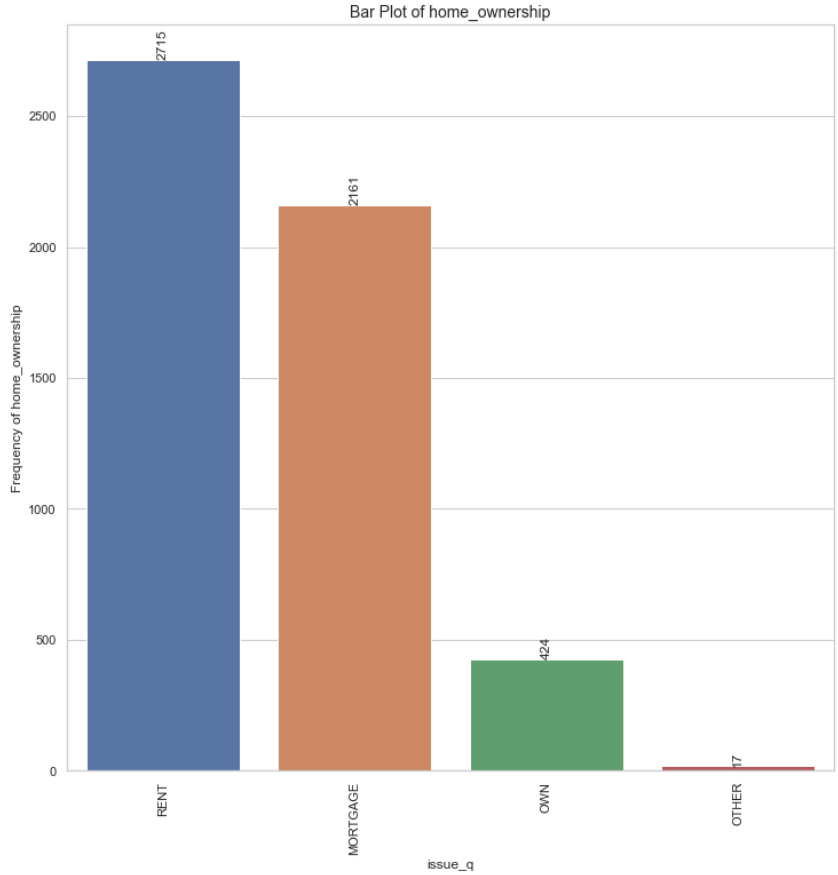
Univariate Analysis (Unordered Categorical)





Address State and Purpose of Loan

Univariate Analysis (Unordered Categorical)



Various types of Home Ownerships

Univariate Analysis (Categorical Variables)

Observations & Inferences:

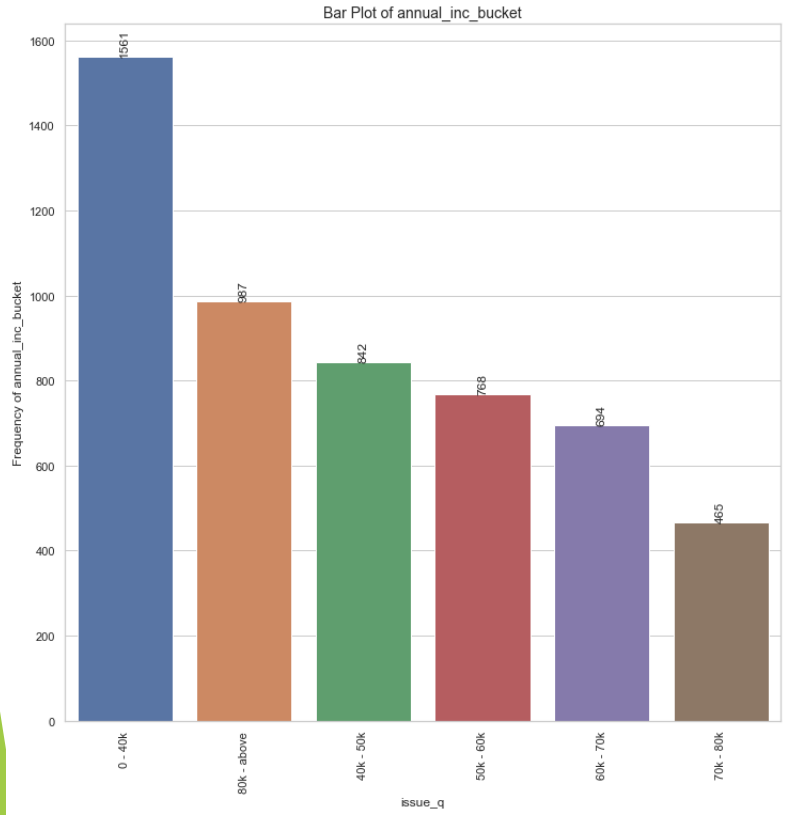
A. **Ordered Categorical Variables:**

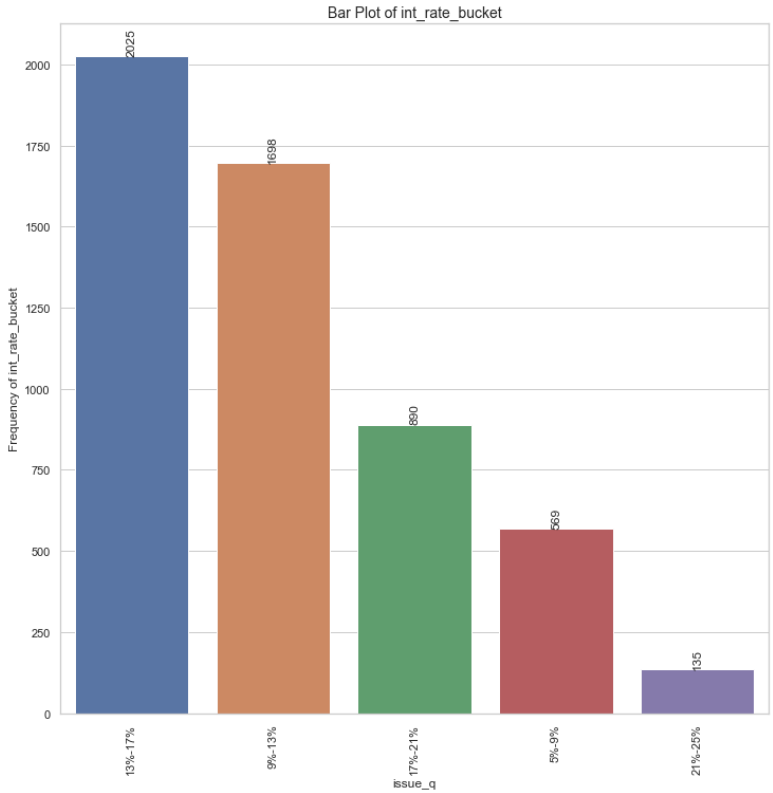
- ✓ Grade B had the highest number of "Charged off" loan applicants, with a total of 1,352 applicants, indicating that applicants with this credit grade faced challenges in repaying their loans.
- ✓ Short-term loans with a duration of 36 months were the most popular among "Charged off" applicants, with 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.

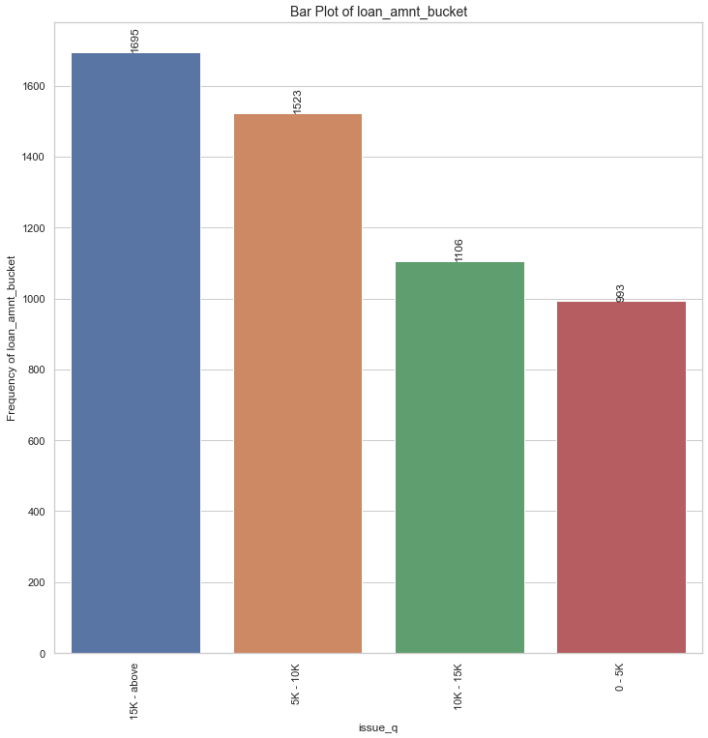
Univariate Analysis (Quantitative Variables)

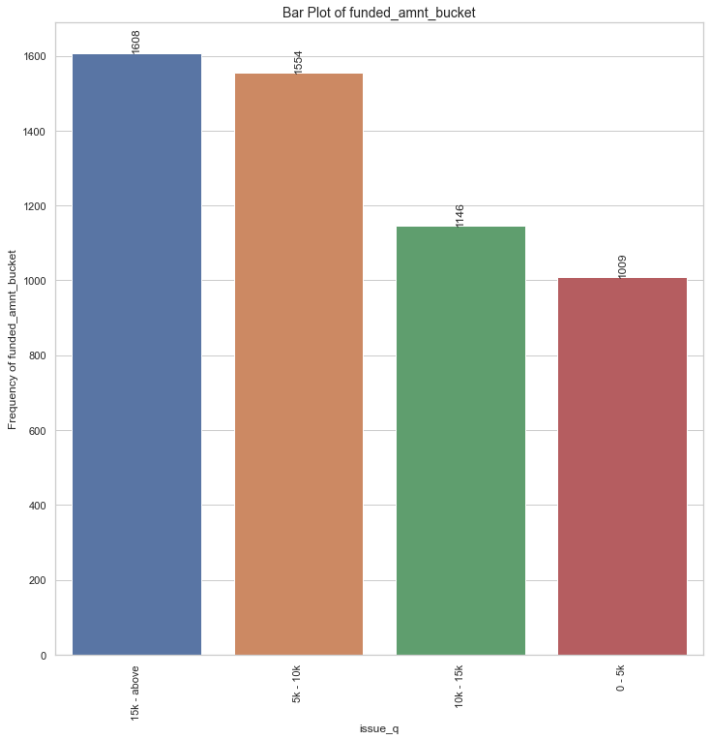




Buckets of Annual Income Status and Loan Interest Rates

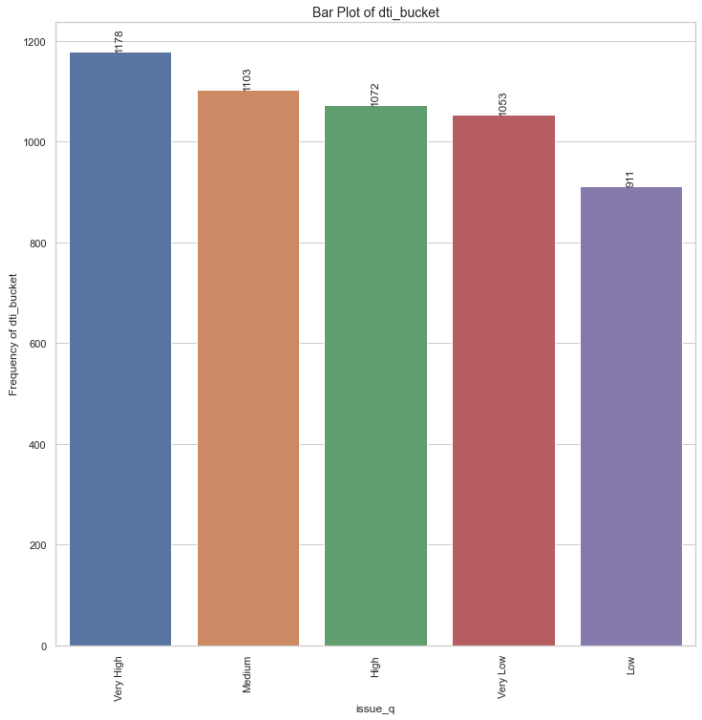
Univariate Analysis (Quantitative Variables)



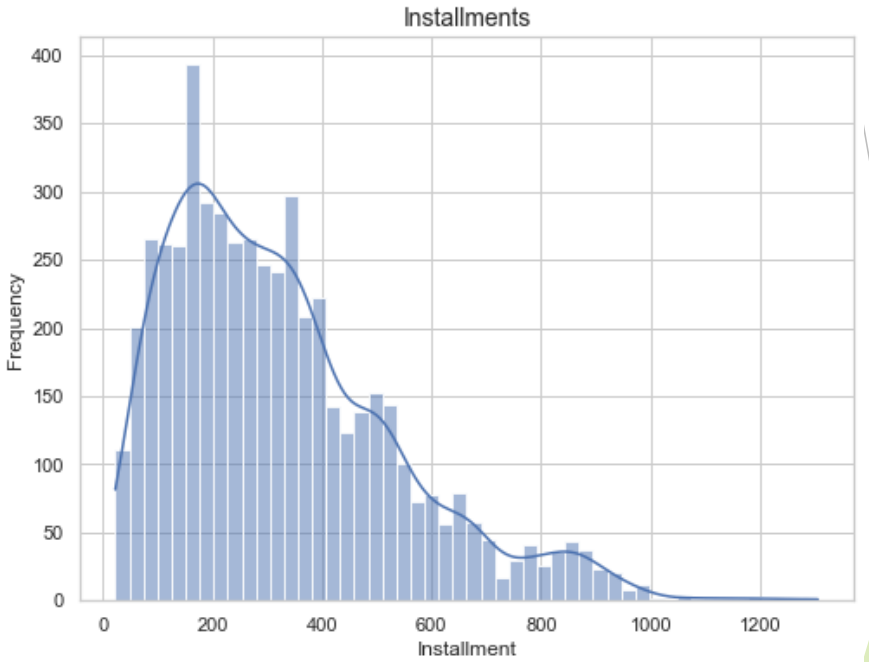


Buckets of Loan Amount and Funded Amount

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

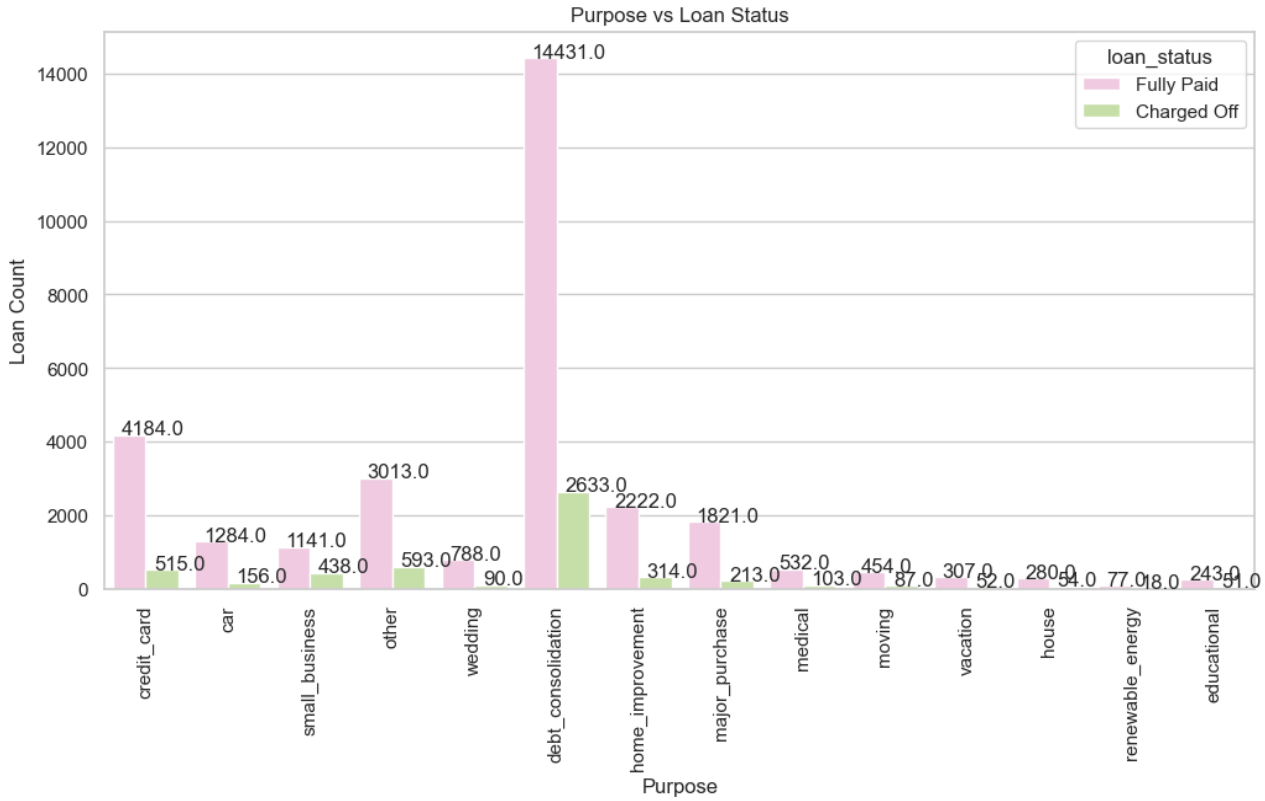
A. **Categorical Variables:**

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:**

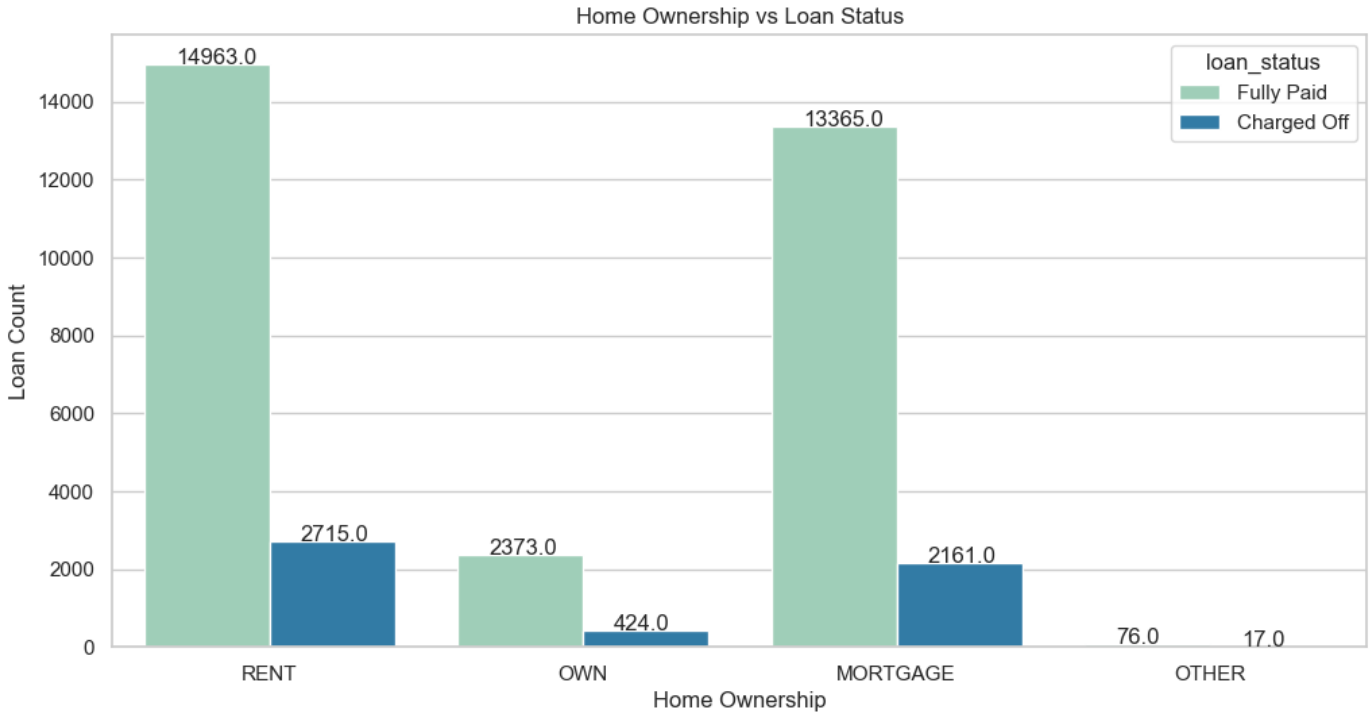
- ✓ **Int Rate Bucket (int_rate_bucket)**
- ✓ **Debt to Income Bucket (dti_bucket)**
- ✓ **Annual Income Bucket (annual_inc_bucket)**
- ✓ **Funded Amount Bucket (funded_amnt_bucket)**
- ✓ **Loan Amount Bucket (loan_amnt_bucket)**

Bivariate Analysis (Unordered Categorical)



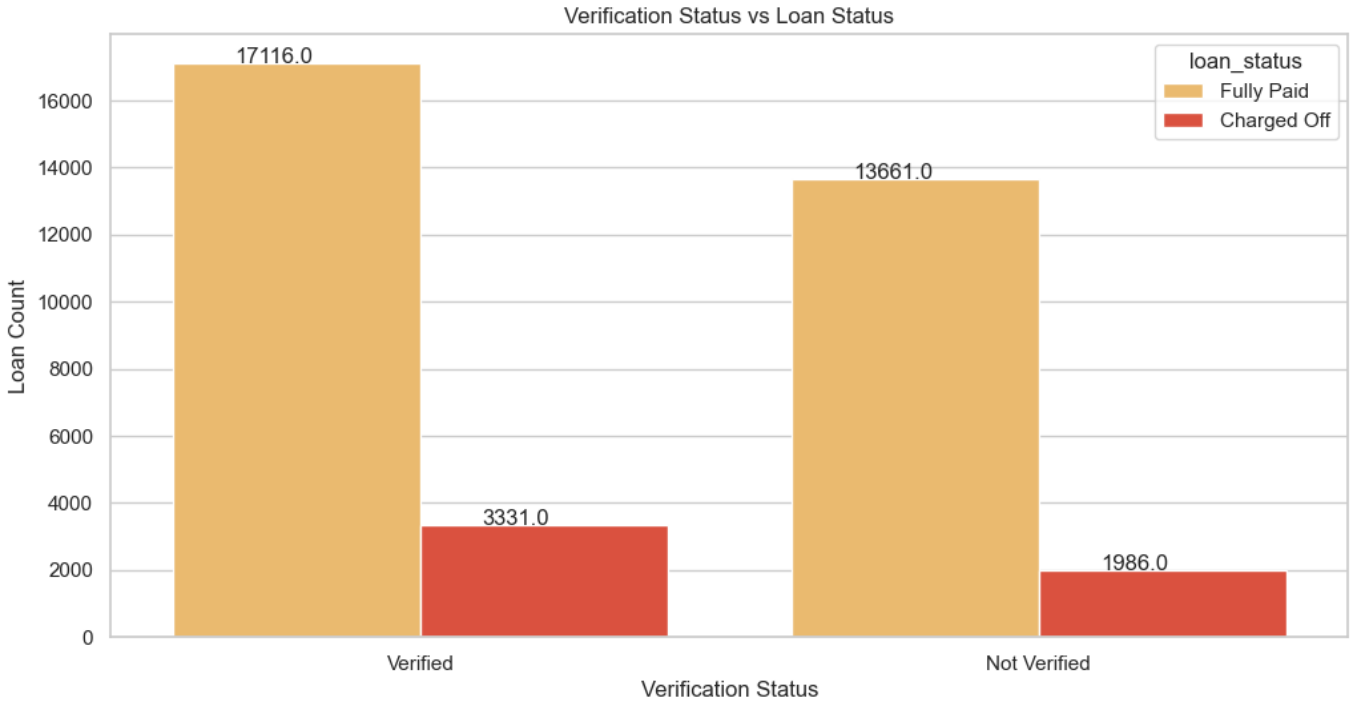
Purpose of Loan v/s Status of Loan

Bivariate Analysis (Unordered Categorical)



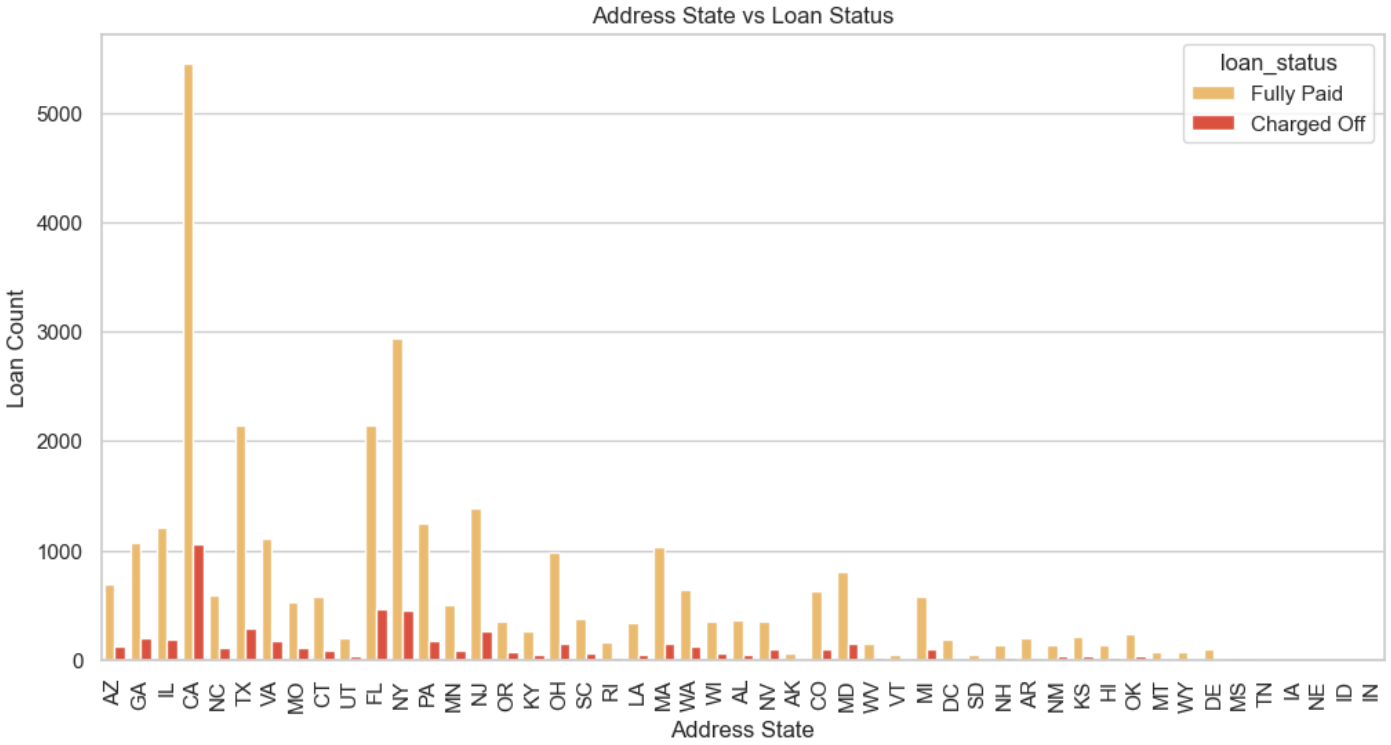
Home Ownership v/s Status of Loan

Bivariate Analysis (Unordered Categorical)



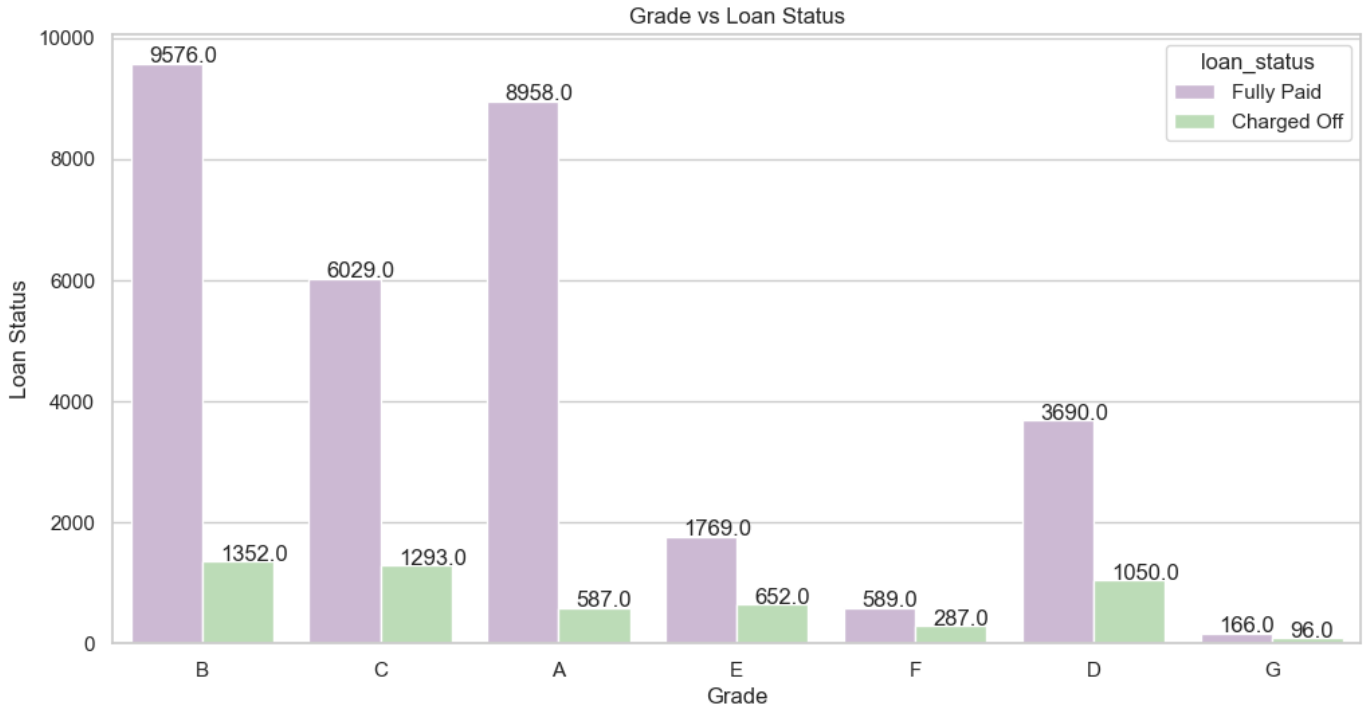
Verification Status of Loan v/s Status of Loan

Bivariate Analysis (Unordered Categorical)



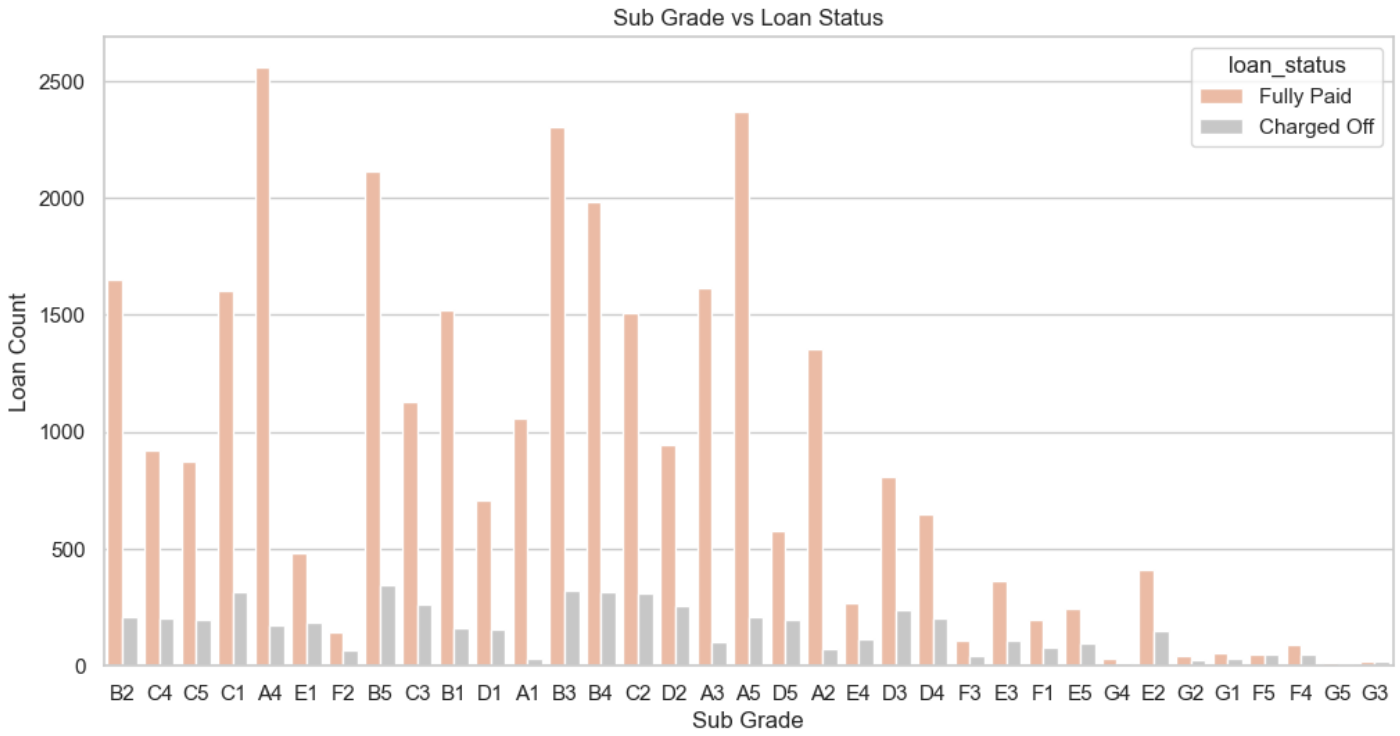
Address State v/s Status of Loan

Bivariate Analysis (Ordered Categorical)



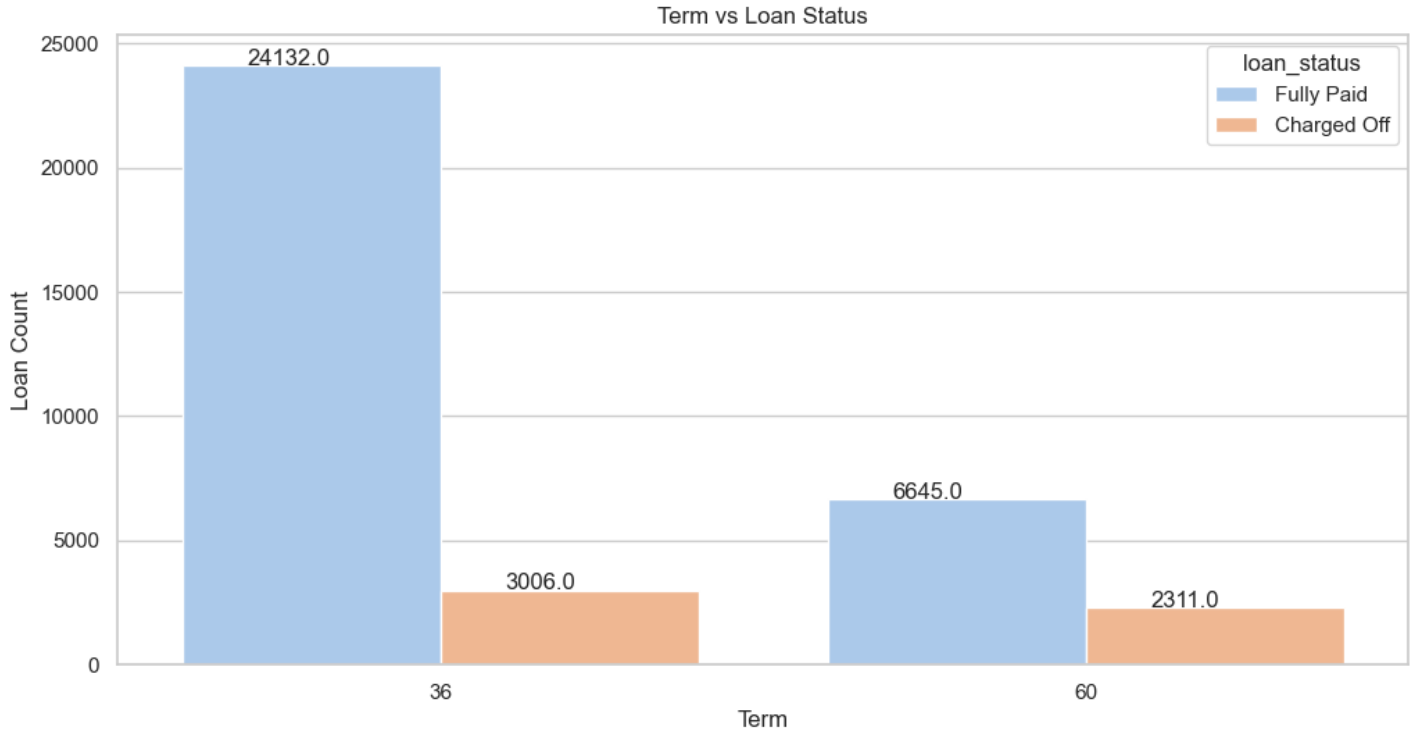
Loan Grade v/s Status of Loan

Bivariate Analysis (Ordered Categorical)



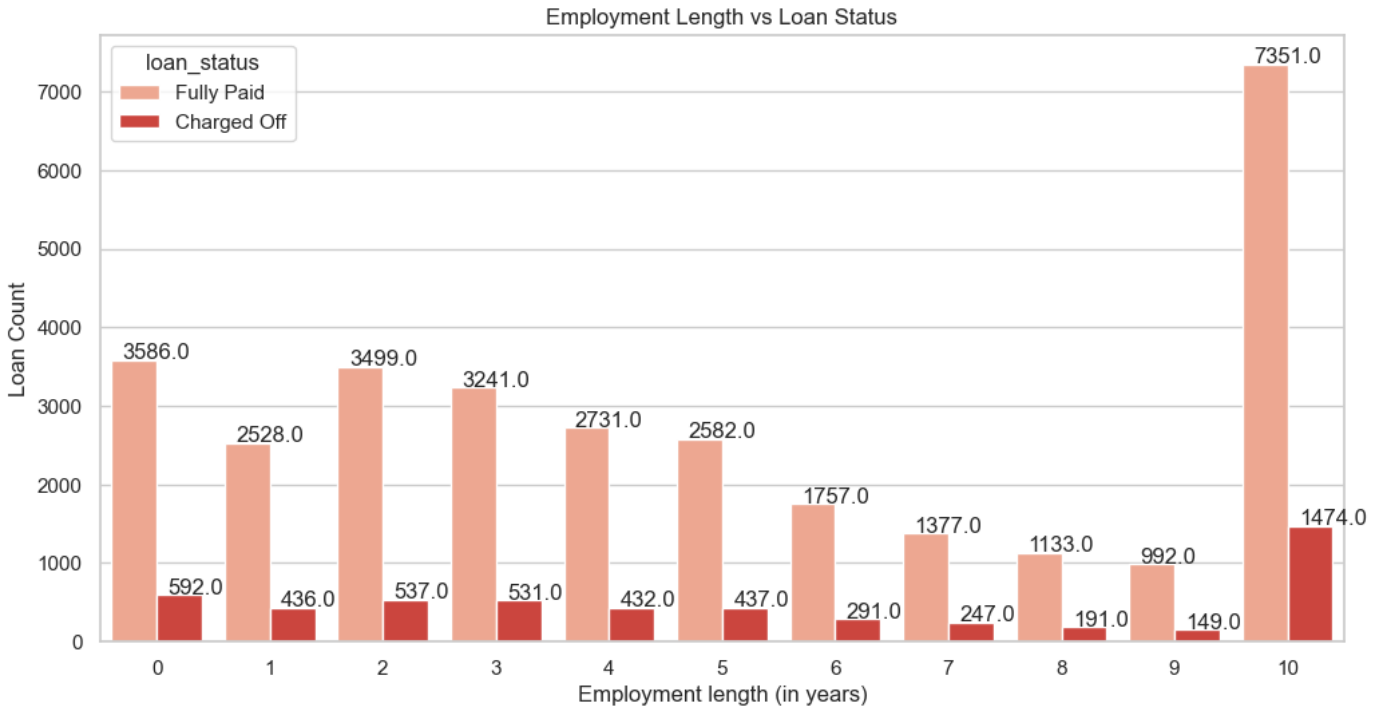
Loan Sub-Grade v/s Status of Loan

Bivariate Analysis (Ordered Categorical)



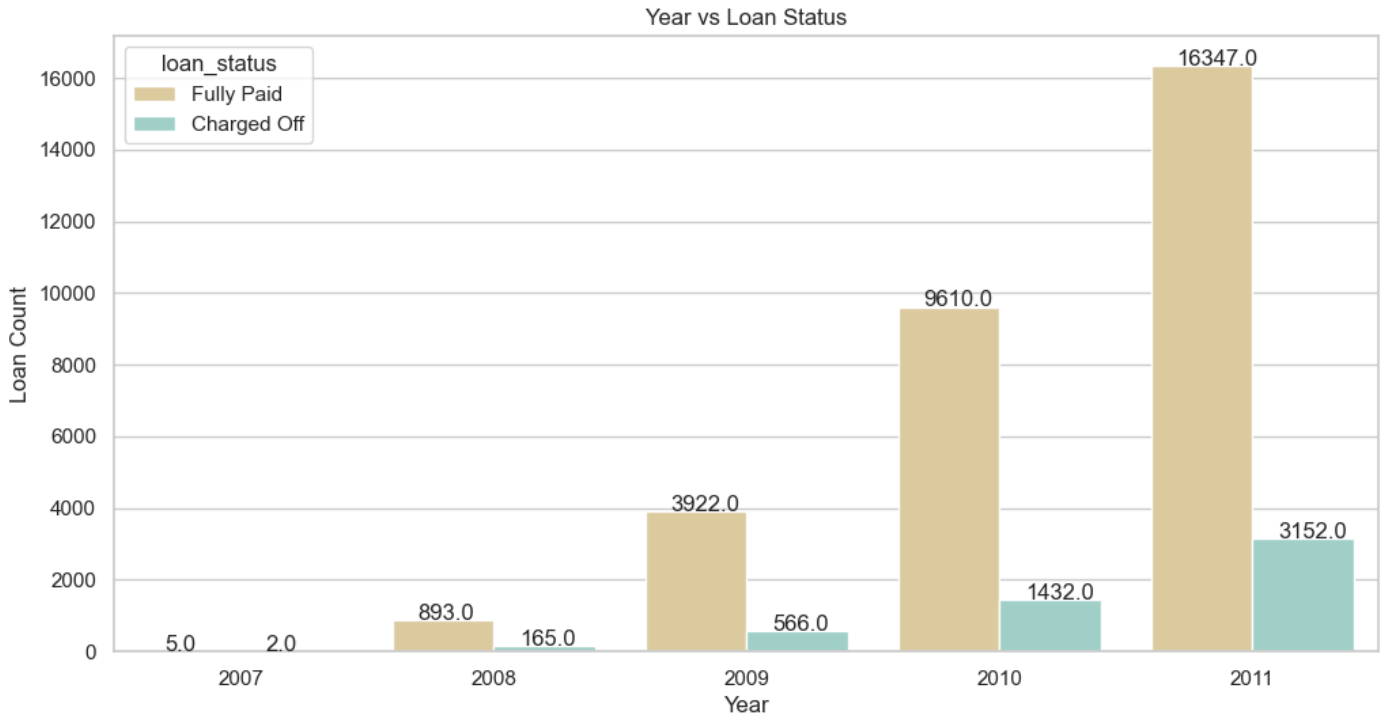
Term of Loan v/s Status of Loan

Bivariate Analysis (Ordered Categorical)



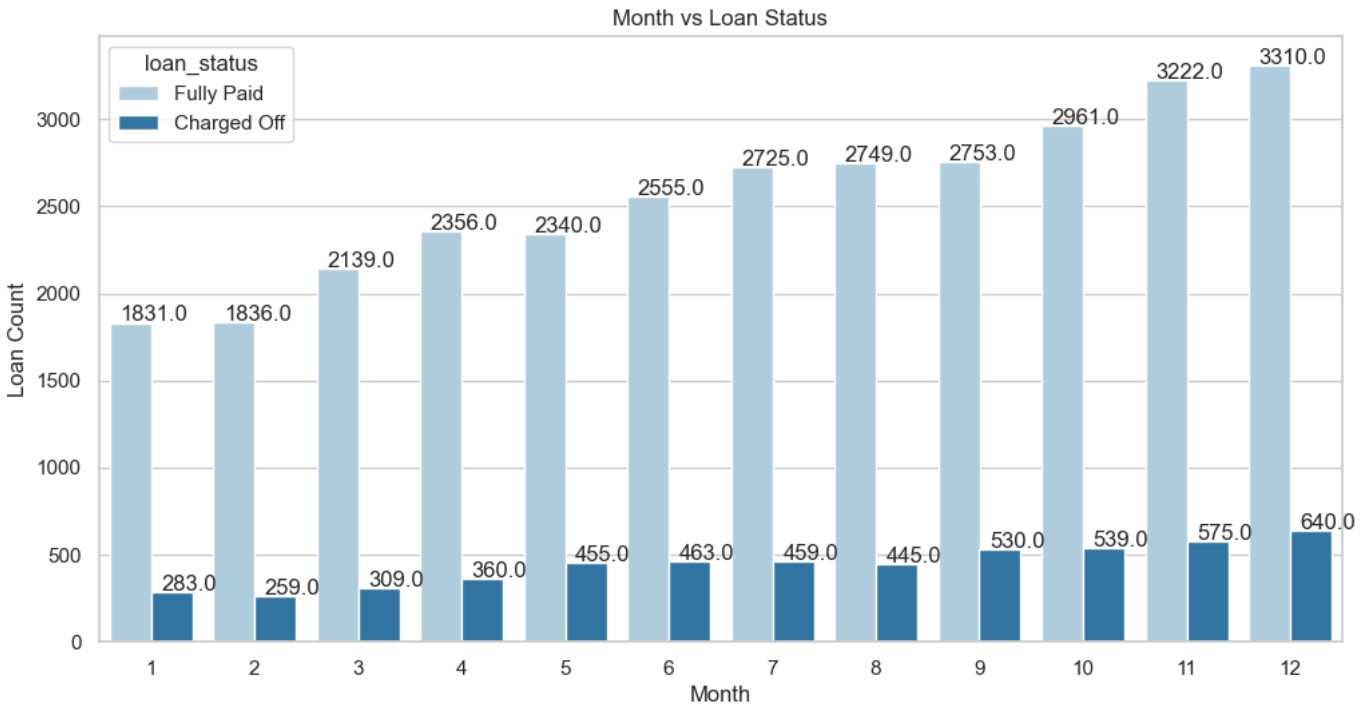
Employment Length of Customer v/s Status of Loan

Bivariate Analysis (Ordered Categorical)



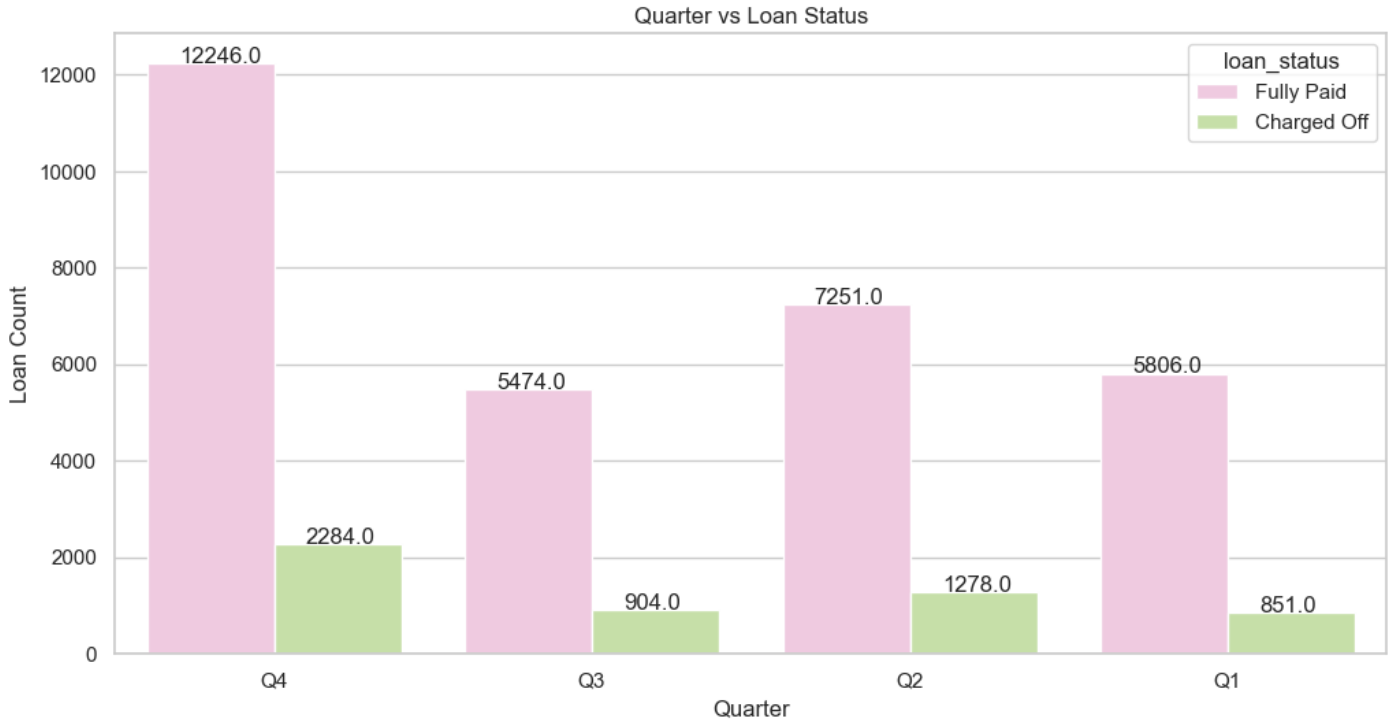
Year the Loan was given to Customer v/s Status of Loan

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.

B. **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:

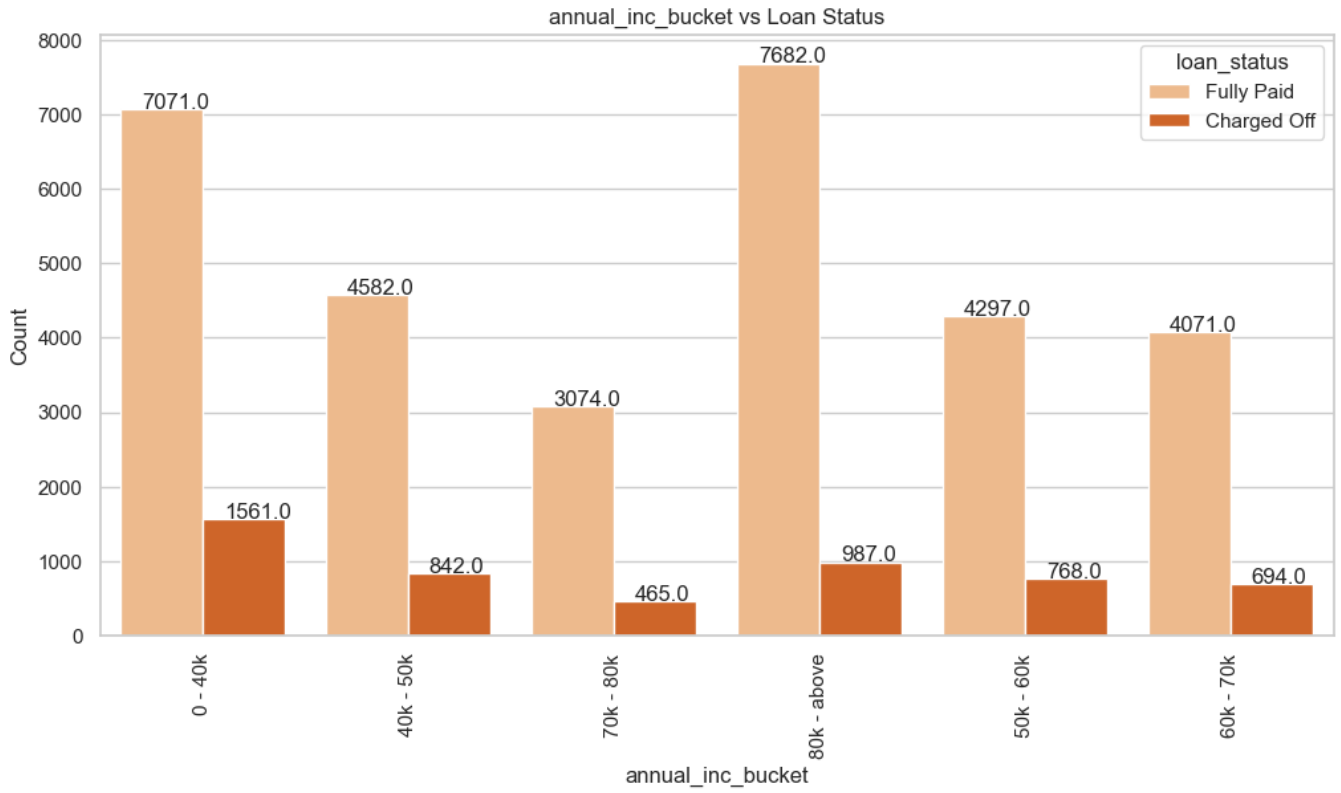
A. **Ordered Categorical Variables:**

- ✓ **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:**

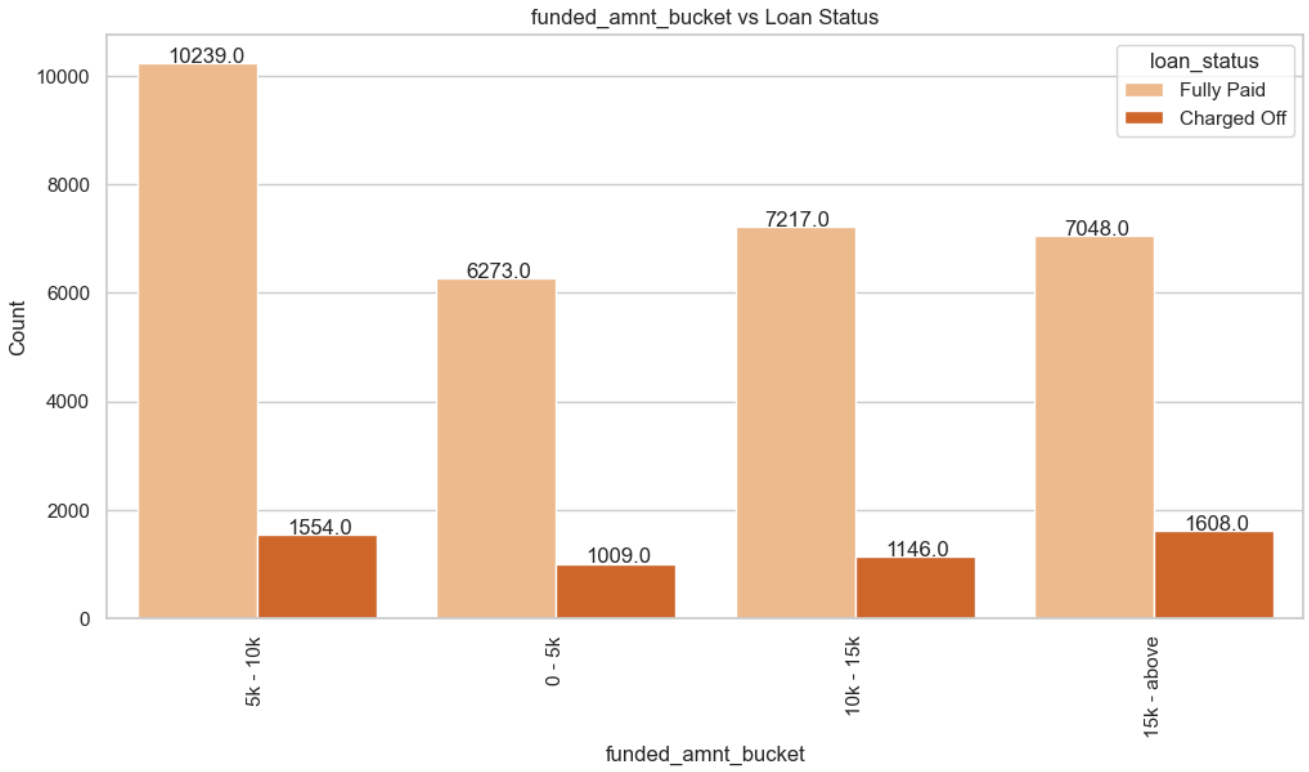
- ✓ **Debt Consolidation Risk**: Since debt consolidation is the category with the maximum number of loans and high default rates, the company should carefully evaluate applicants seeking debt consolidation loans and potentially adjust interest rates or offer financial counseling services.
- ✓ **Housing 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)



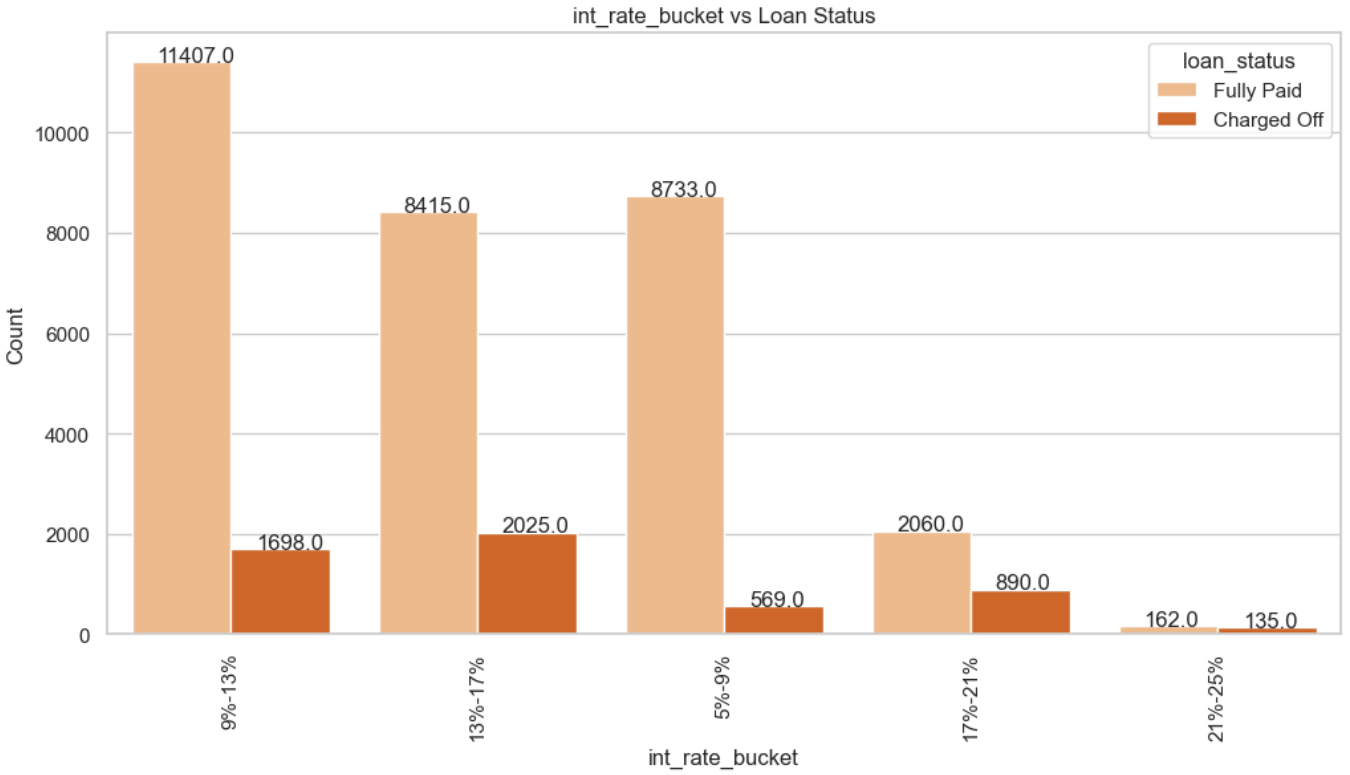
Bucket of Annual Income v/s Status of Loan

Bivariate Analysis (Quantitative Variables)



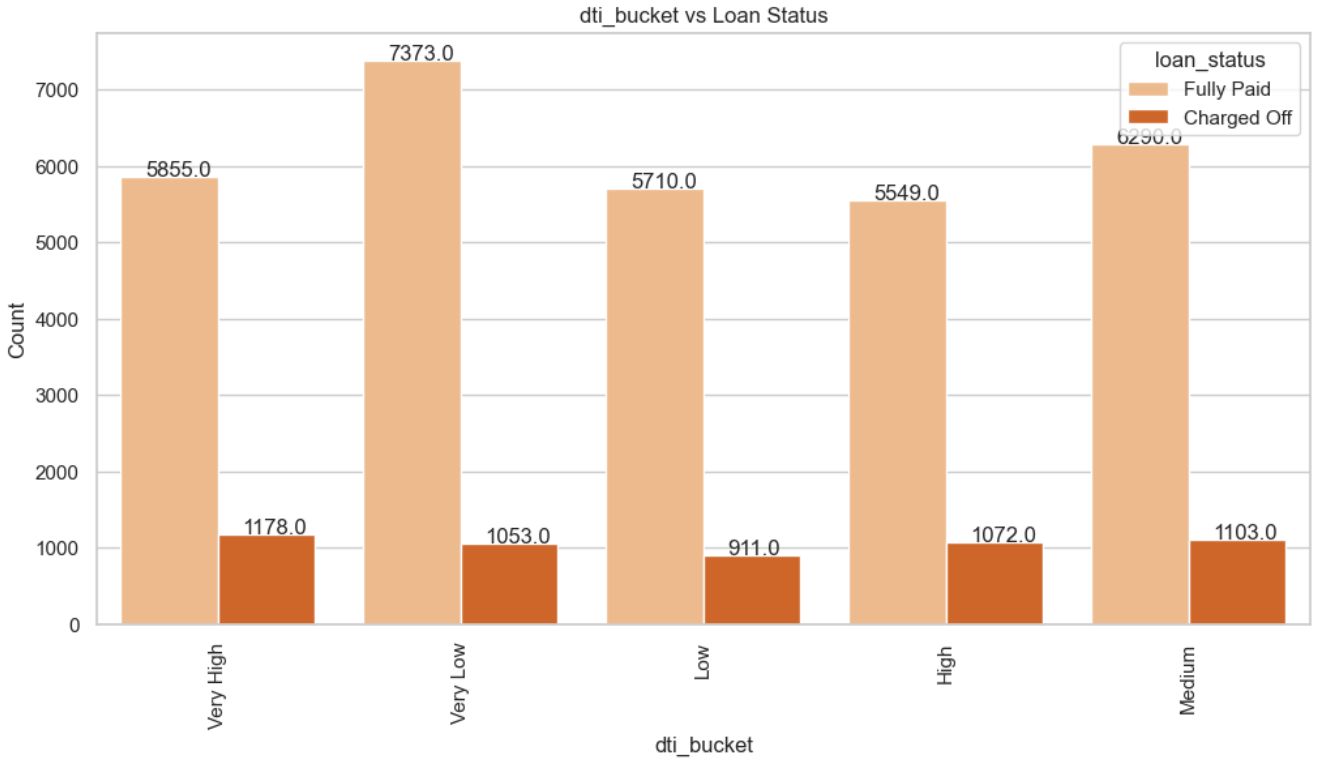
Bucket of Amount which was Funded v/s Status of Loan

Bivariate Analysis (Quantitative Variables)



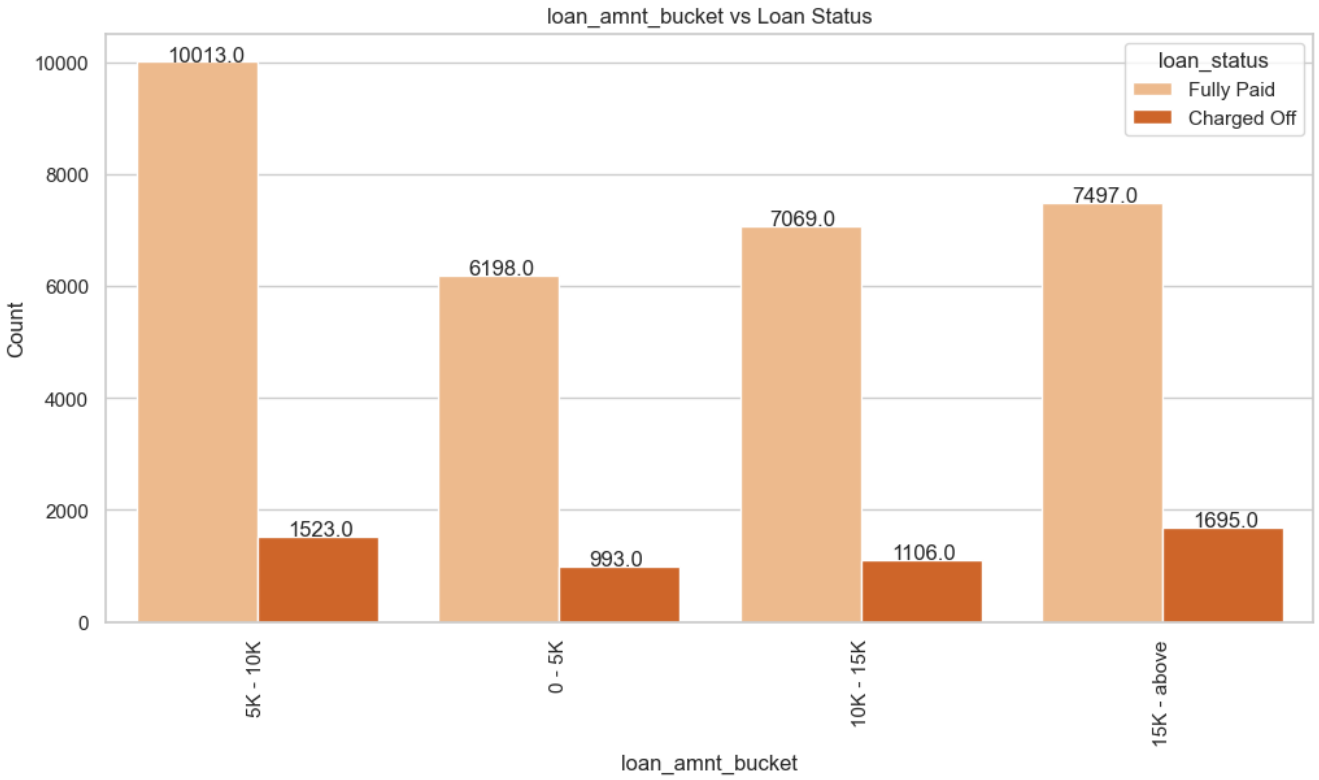
Bucket of Interest Rate of the loan v/s Status of Loan

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:

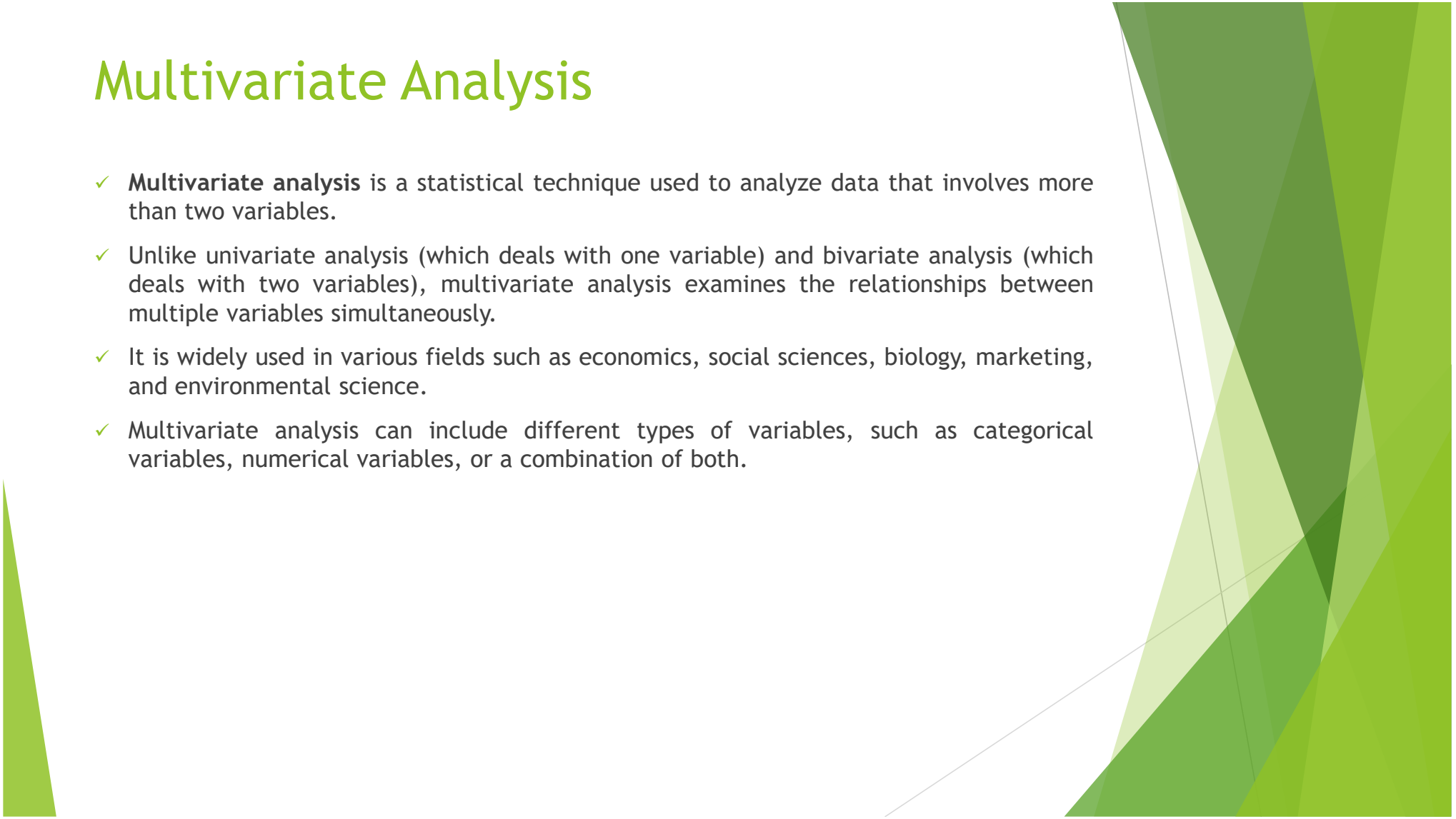
- ✓ A majority of the loan applicants who defaulted received loan amounts of $15,000 or higher.
- ✓ The majority of loan applicants who charged off had significantly high Debt-to-Income (DTI) ratios.
- ✓ A significant portion of loan applicants who defaulted received loans with interest rates falling 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:

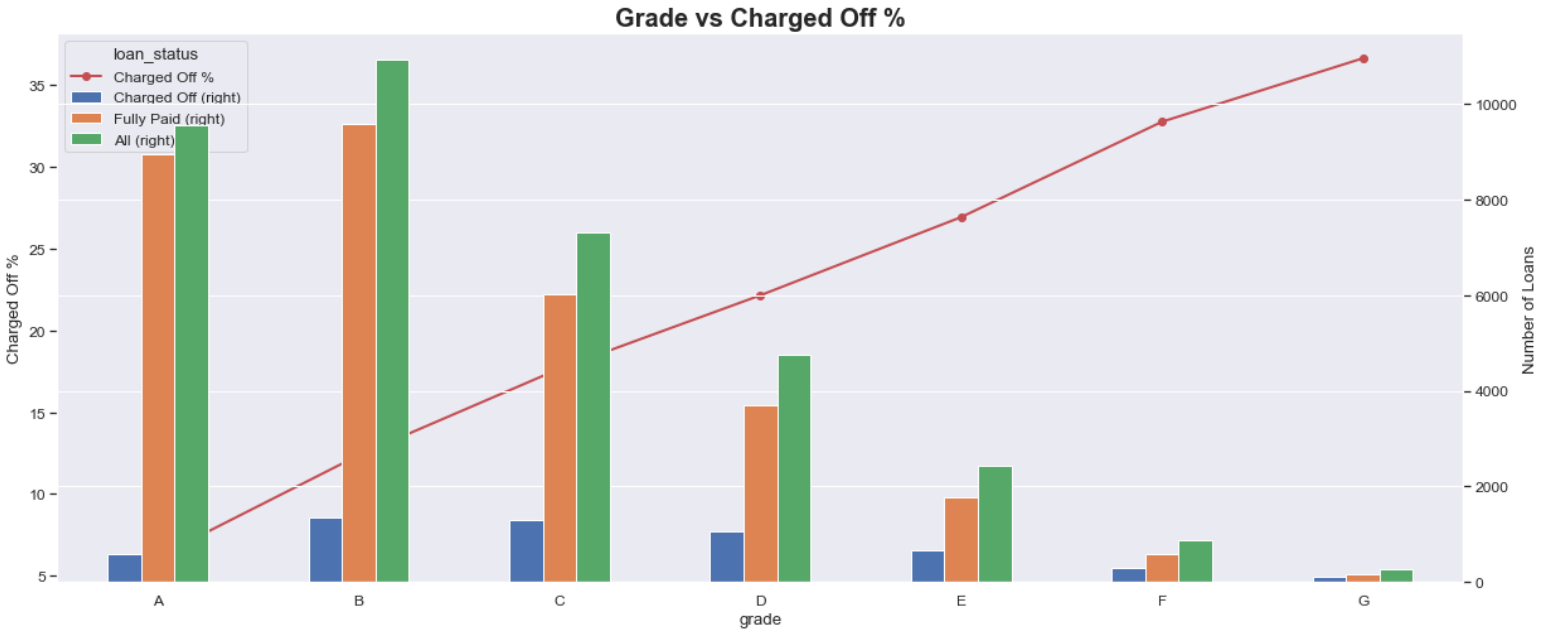
- ✓ **High Loan Amounts**: Applicants receiving loan amounts of $15,000 or higher are more likely to default. The company can mitigate this risk by conducting more thorough assessments for larger loan requests and potentially capping loan amounts for higher-risk applicants.
- ✓ **DTI and Interest Rates**: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

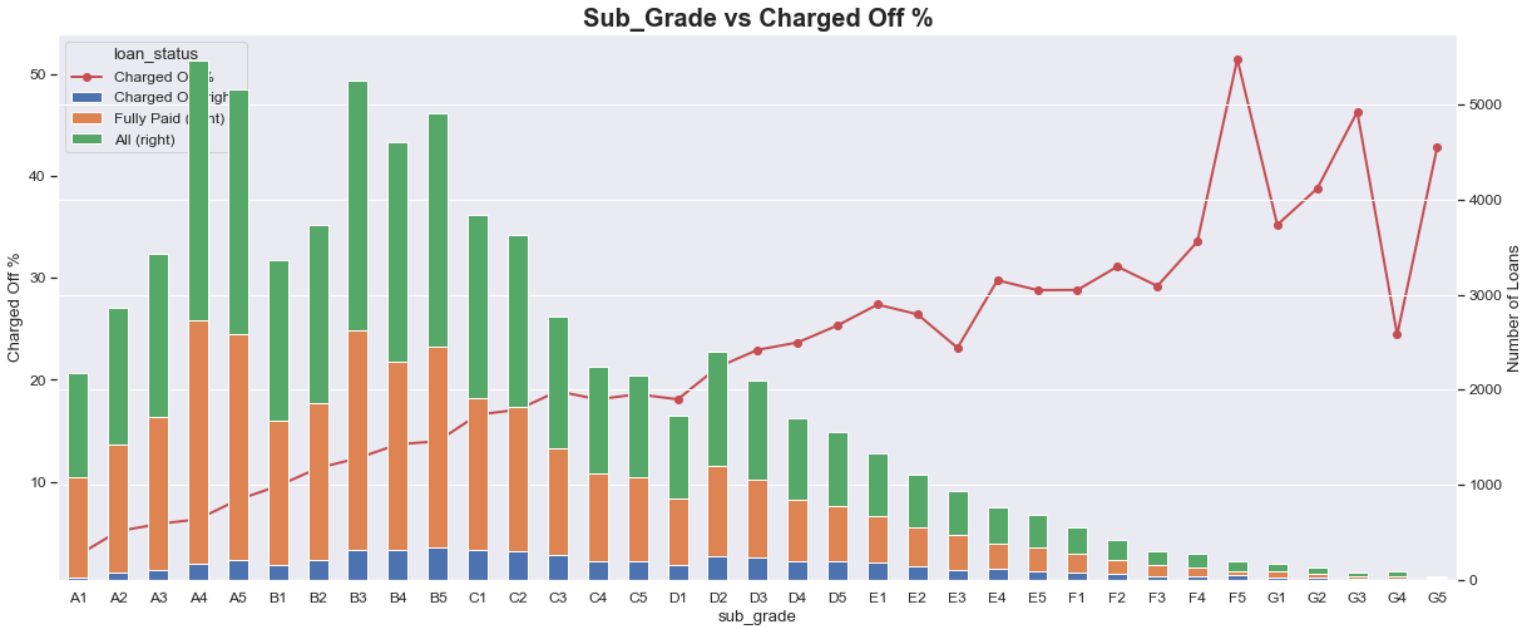


Multivariate Analysis



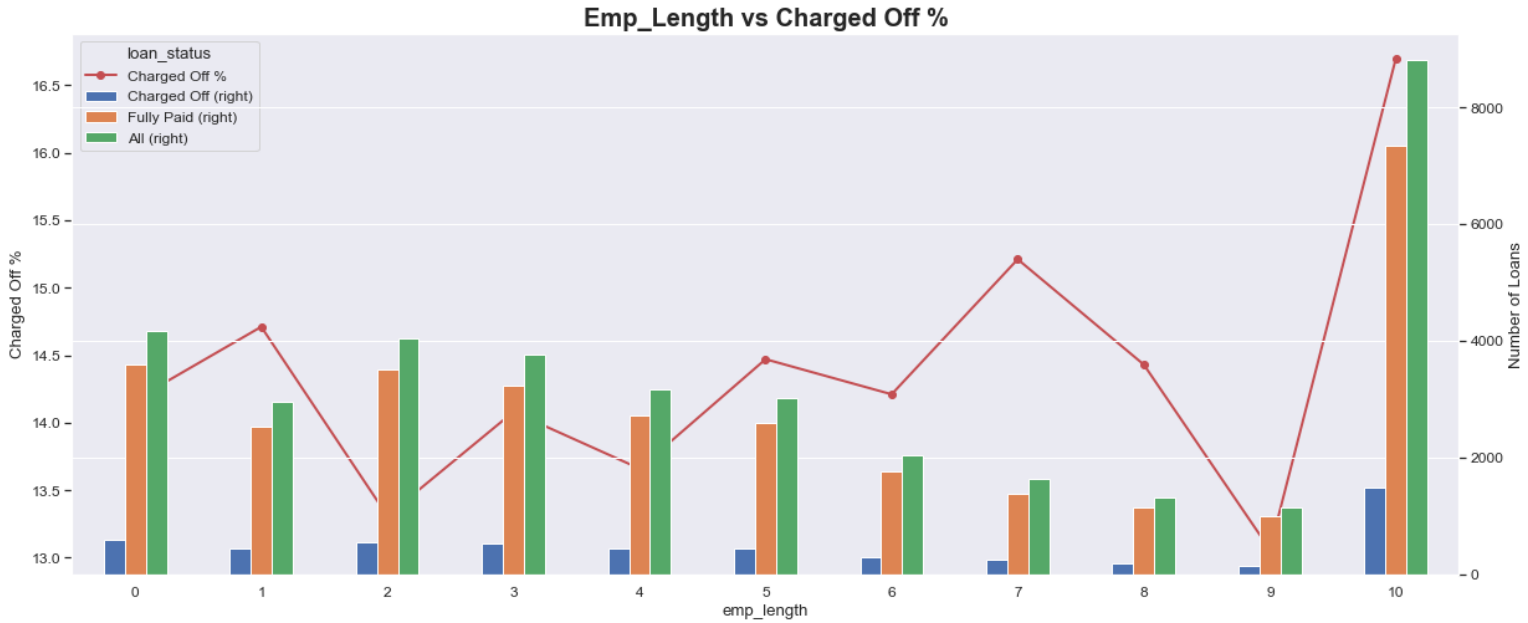
Grade v/s Percentage of Charged-off Loans

Multivariate Analysis



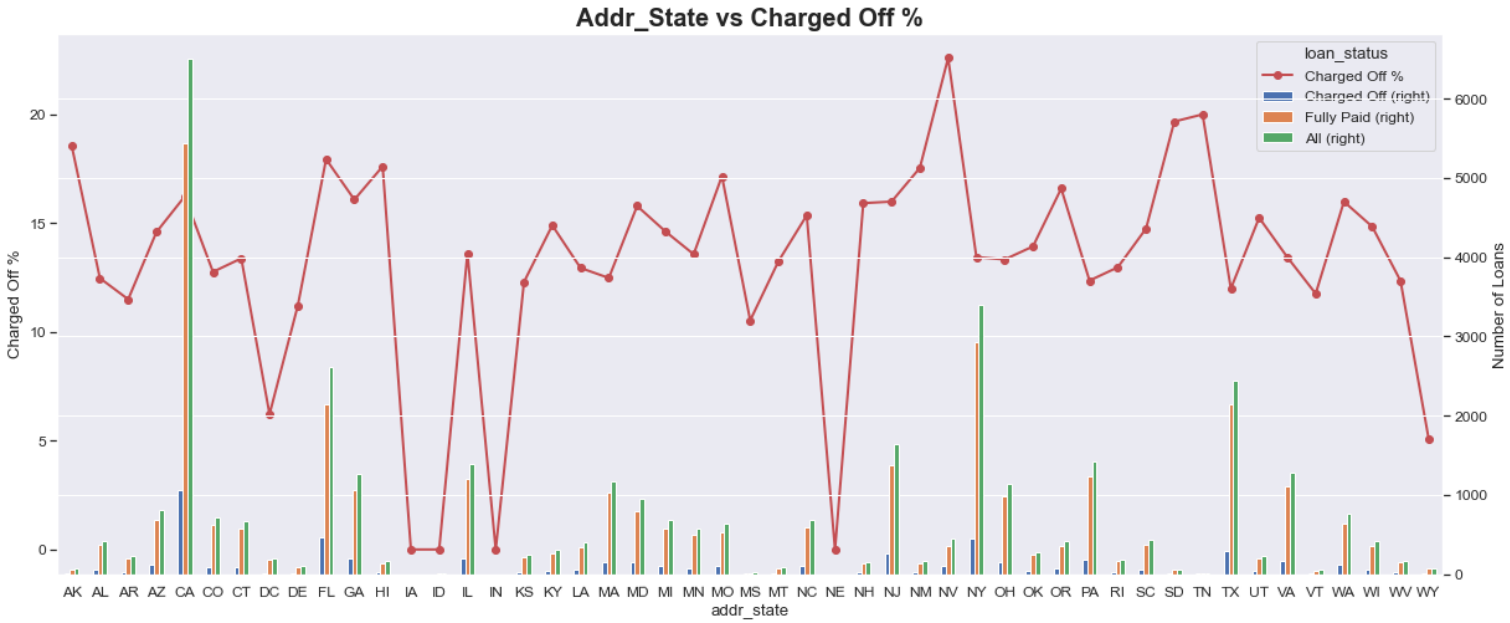
Sub-Grade v/s Percentage of Charged-off Loans

Multivariate Analysis



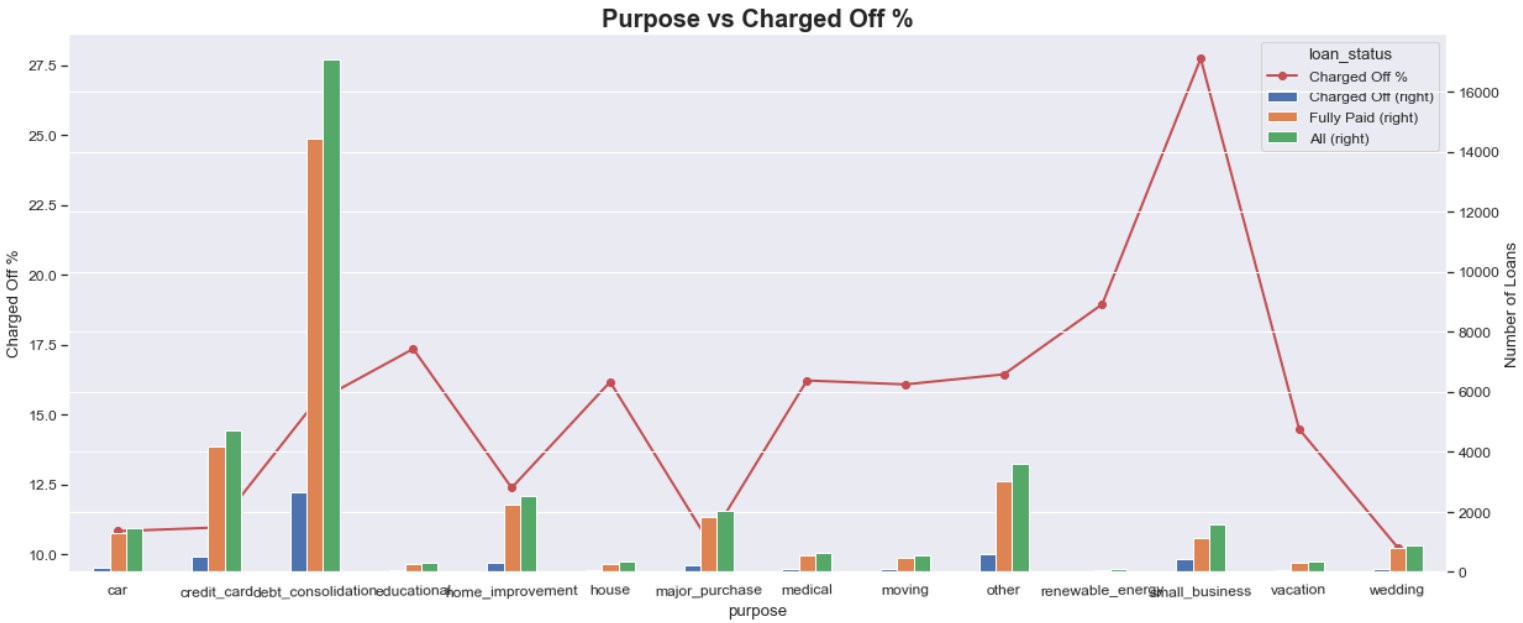
Employment Length (In Years) v/s
Percentage of Charged-off Loans

Multivariate Analysis



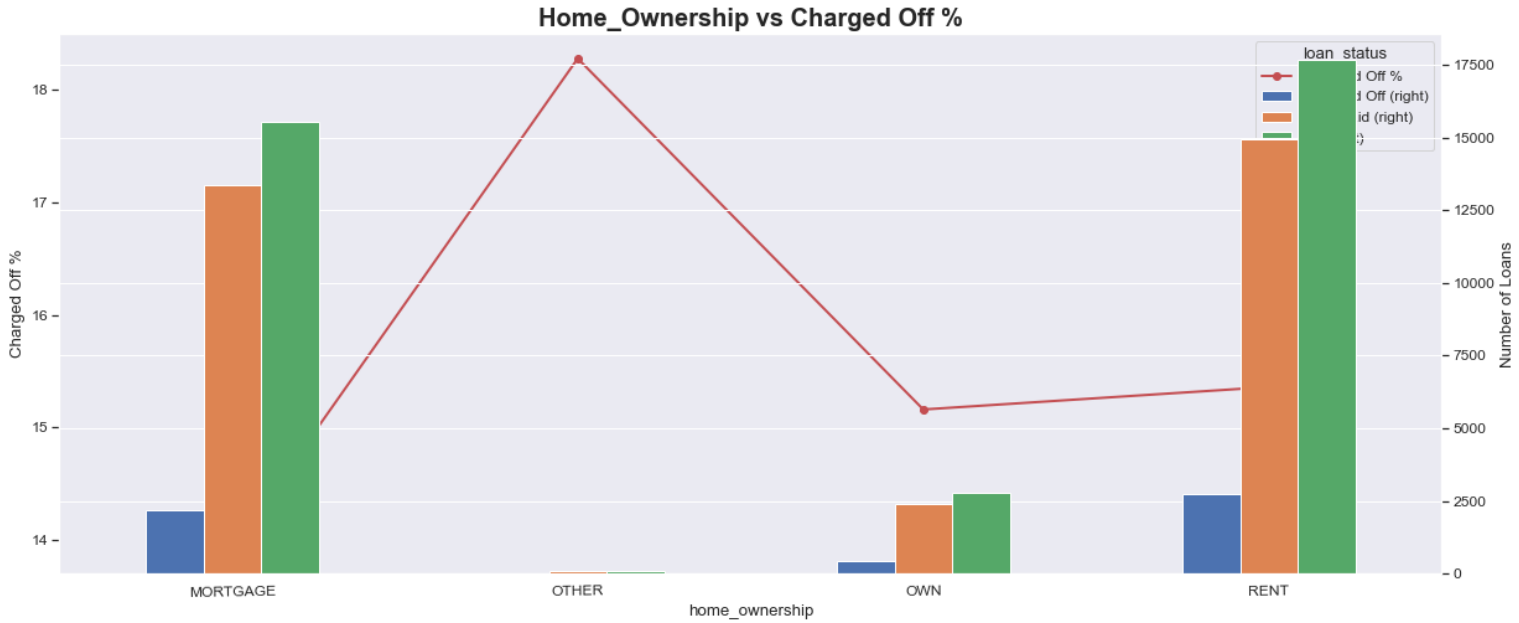
Address State v/s Percentage of Charged-off Loans

Multivariate Analysis



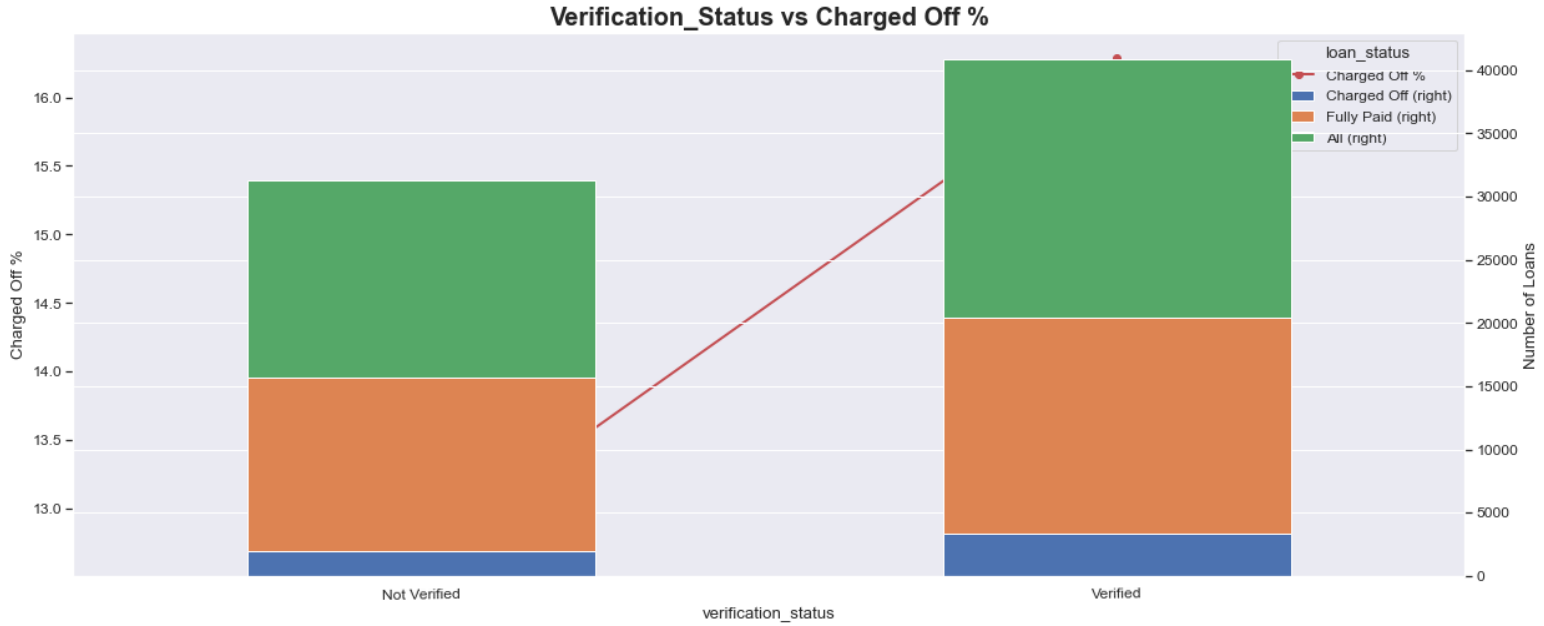
Purpose of Loan v/s Percentage of Charged-off Loans

Multivariate Analysis



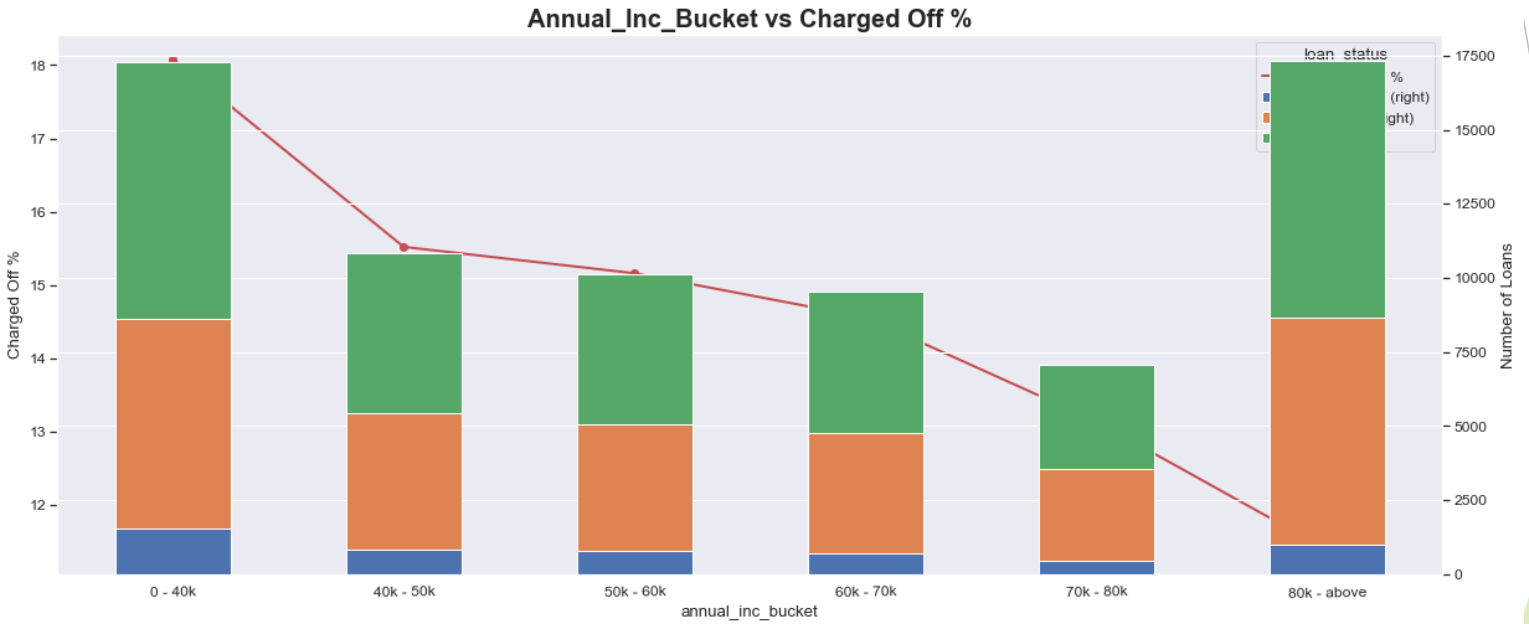
Home Ownership v/s Percentage of Charged-off Loans

Multivariate Analysis



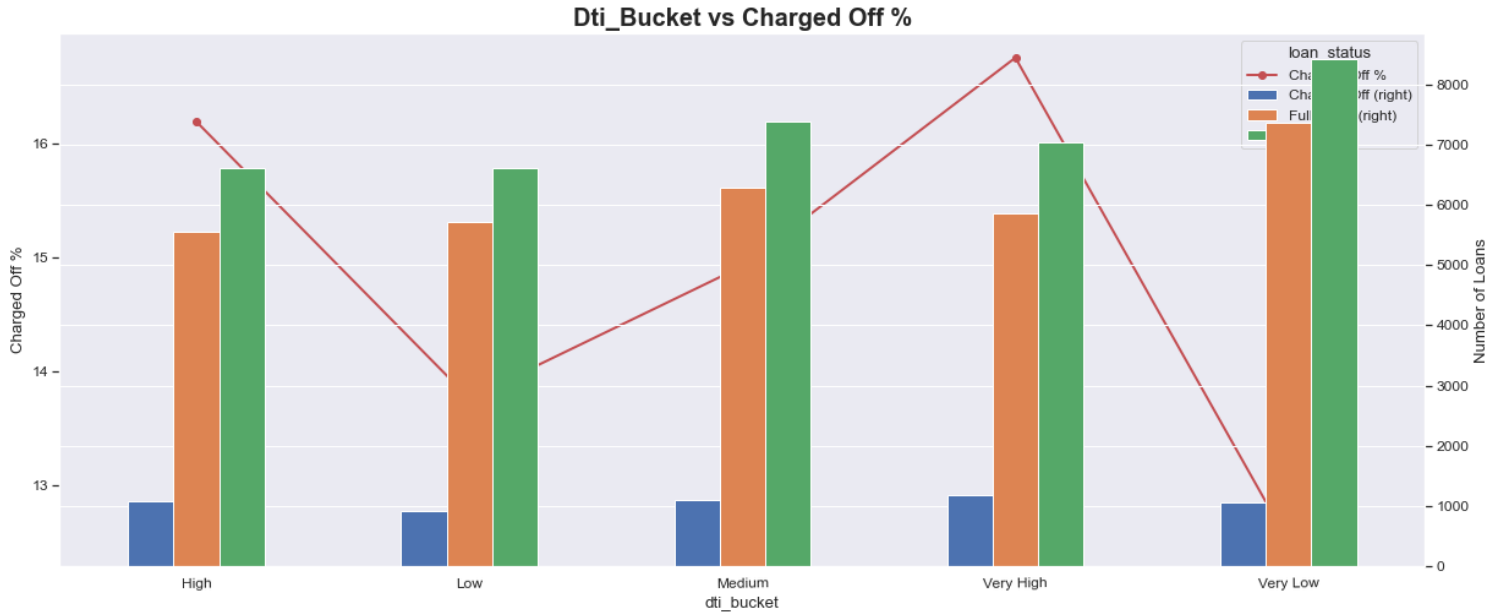
Verification Status of Loan v/s
Percentage of Charged-off Loans

Multivariate Analysis



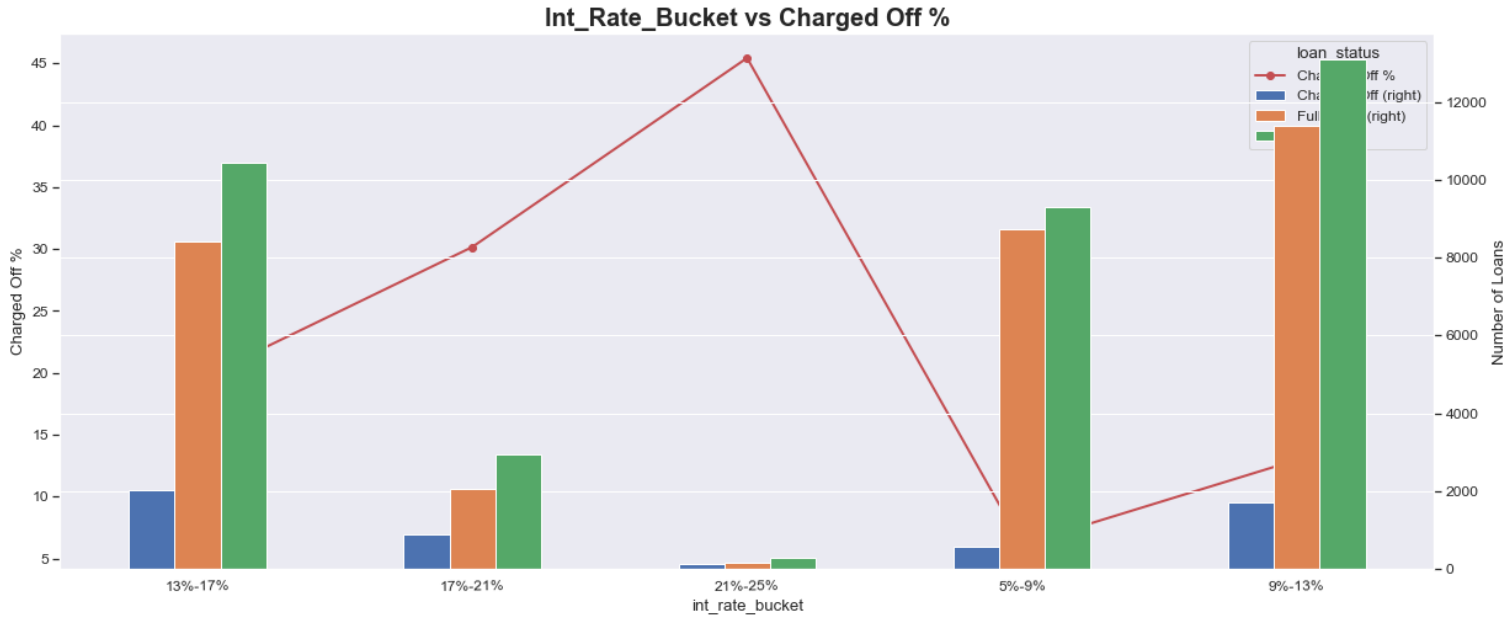
Buckets of Annual Income v/s
Percentage of Charged-off Loans

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

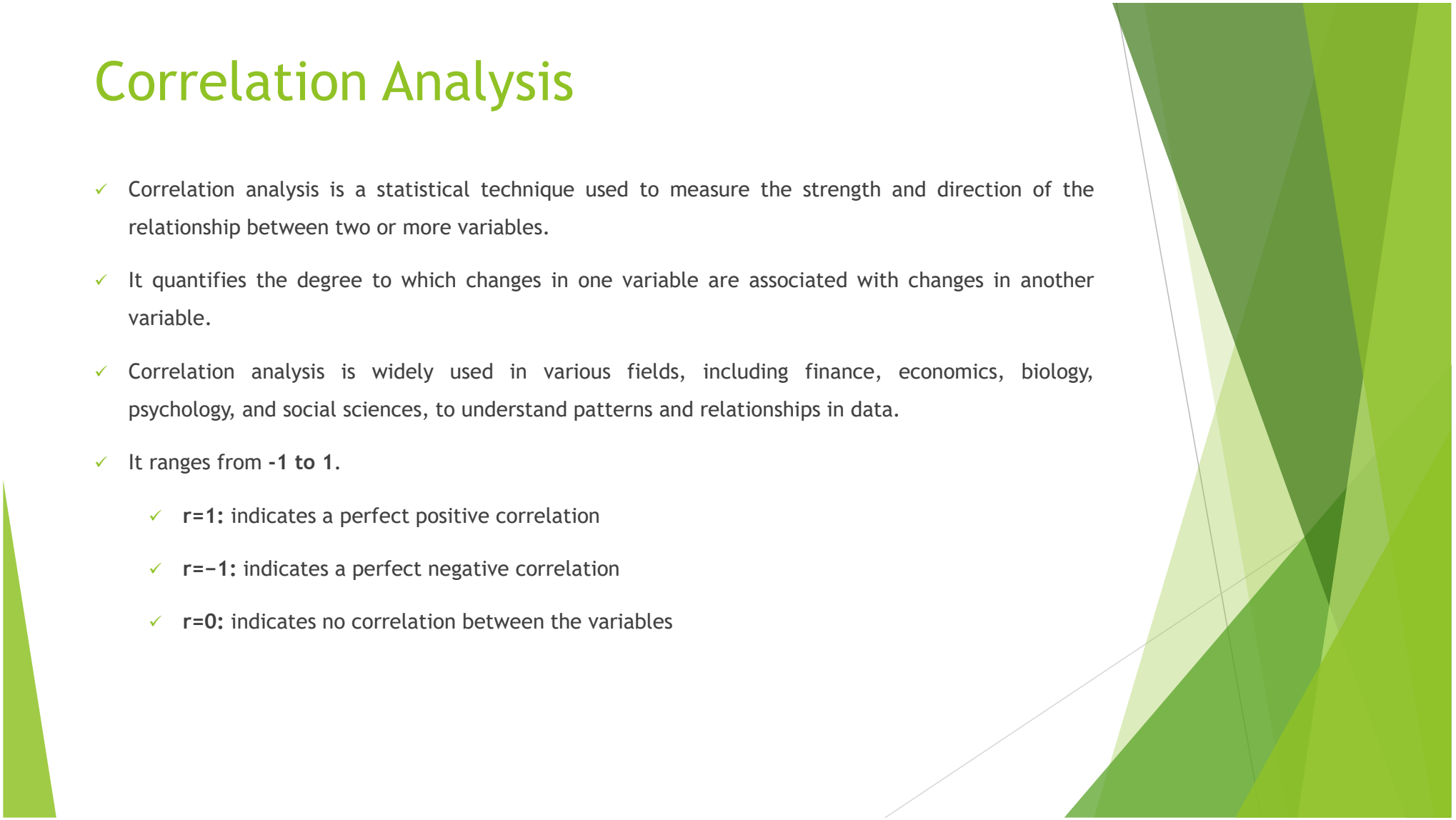
Multivariate Analysis

Observations & Inferences:

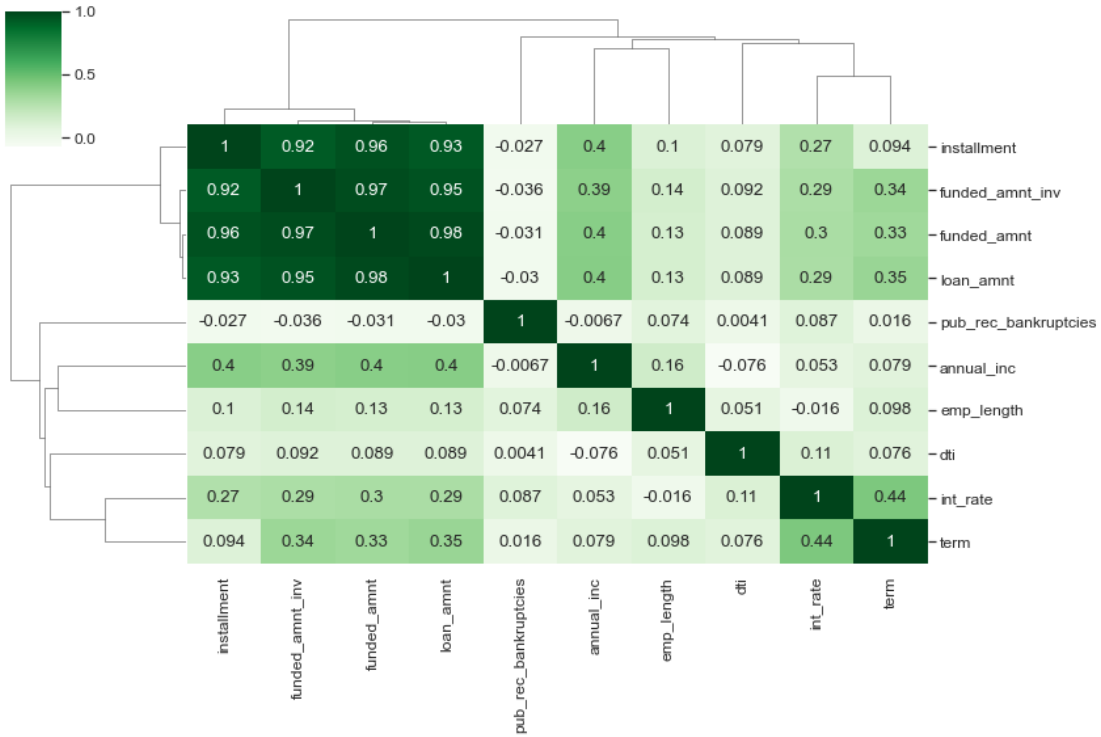
- ✓ Tendency to default the loan is likely with loan applicants belonging to B, C, D grades.
- ✓ Borrowers from sub grade B3, B4 and B5 have maximum tendency to default.
- ✓ Loan applicants with 10 years of experience has maximum tendency to default the loan.
- ✓ Borrowers from states CA, FL, NJ have 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**

Suggestions

- ✓ **Implement Stricter Criteria for Grades B, C, and D:** Consider implementing stricter risk assessment and underwriting criteria for applicants falling into Grades B, C, and D to minimize default risks.
- ✓ **Focus on Subgrades B3, B4, and B5:** Pay special attention to applicants with Subgrades B3, B4, and B5. Consider additional risk mitigation measures or offering lower loan amounts for these subgrades to reduce default rates.
- ✓ **Evaluate and Limit 60-Month Loans:** Evaluate the risk associated with 60-month loans. Consider limiting the maximum term or adjusting interest rates for longer-term loans to decrease the likelihood of defaults.
- ✓ **Comprehensive Credit Scoring System:** Develop a comprehensive credit scoring system that incorporates various risk-related attributes, as experience alone might not be sufficient to gauge creditworthiness.
- ✓ **Capitalizing on Market Growth:** Capitalize on the market's growth trend observed from 2007 to 2011 by maintaining a competitive edge in the industry while ensuring robust risk management practices.
- ✓ **Anticipate Peak Periods:** Anticipate increased loan applications during peak periods such as December and Q4. Ensure efficient processing to meet customer demands during these busy seasons.

Suggestions

- **Careful Evaluation for Debt Consolidation Loans:** Carefully evaluate applicants seeking debt consolidation loans, considering potential interest rate adjustments or offering financial counseling services to manage the associated risks.
- **Consider Housing Stability:** Take housing status into account during the underwriting process to assess housing stability and its impact on the applicant's ability to repay the loan.
- **Review Verification Process:** Review the verification process to ensure effective assessment of applicant creditworthiness. Consider improvements or adjustments based on the review findings.
- **Monitor & Adjust for Regional Risk Trends:** Monitor regional risk trends, especially in states like **California, Florida, and New York.** Adjust lending strategies or rates accordingly in high-risk regions.
- **Thorough Assessment for High Loan Amounts:** Conduct more thorough assessments for loan amounts of $15,000 or higher. Consider capping loan amounts for higher-risk applicants to mitigate potential defaults.
- **Adjust Interest Rates Based on DTI Ratios:** Review the interest rate determination process and consider adjusting rates based on Debt-to-Income (DTI) ratios to align with the borrower's ability to repay.
- **Consider Income Levels for Affordability:** Consider offering financial education resources and set maximum loan amounts based on annual incomes below $40,000 to ensure loan affordability for borrowers.