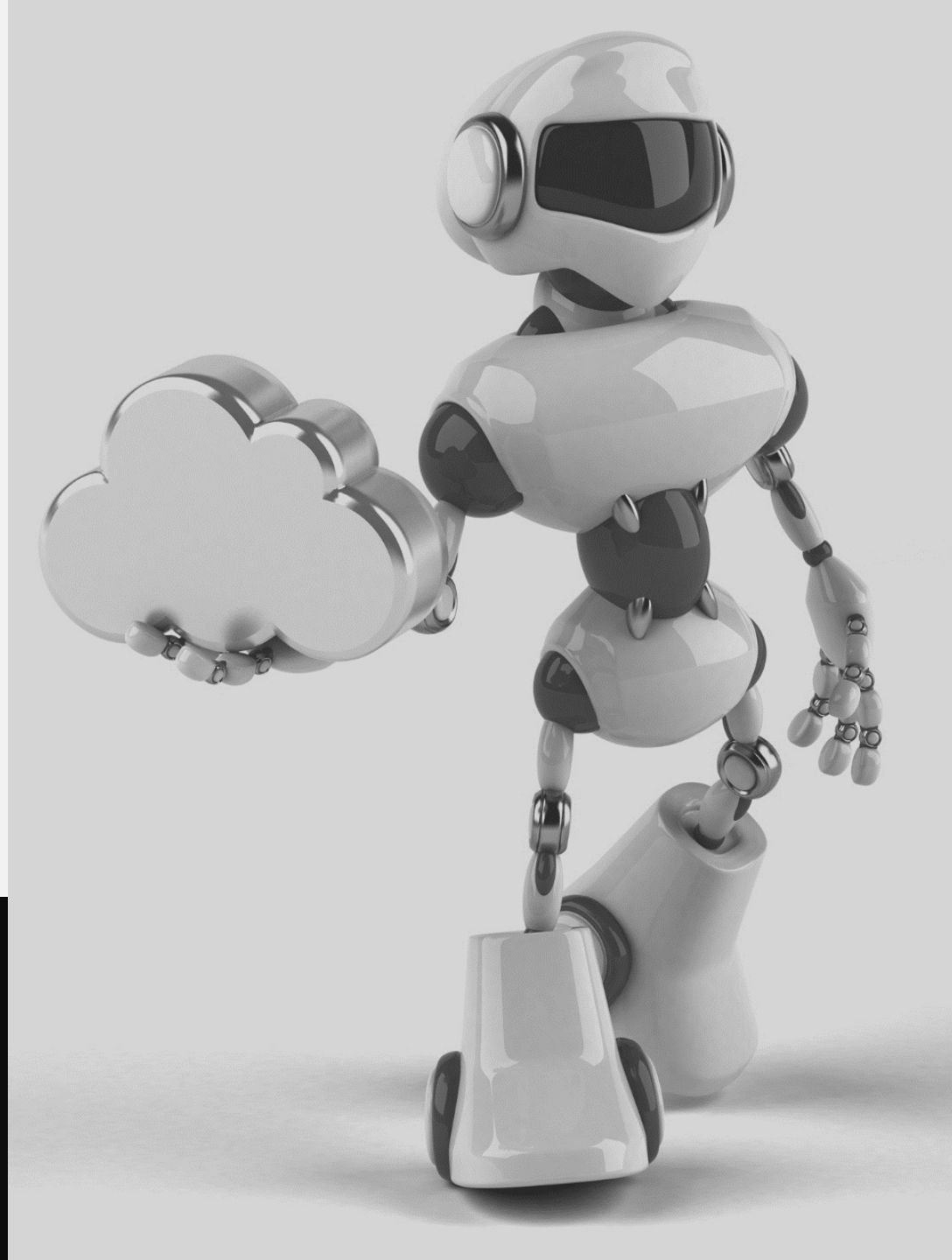
Lending Club Case Study

- Mittinpreet Singh Nayyar
- Melvin John

Table of Contents

Problem Statement

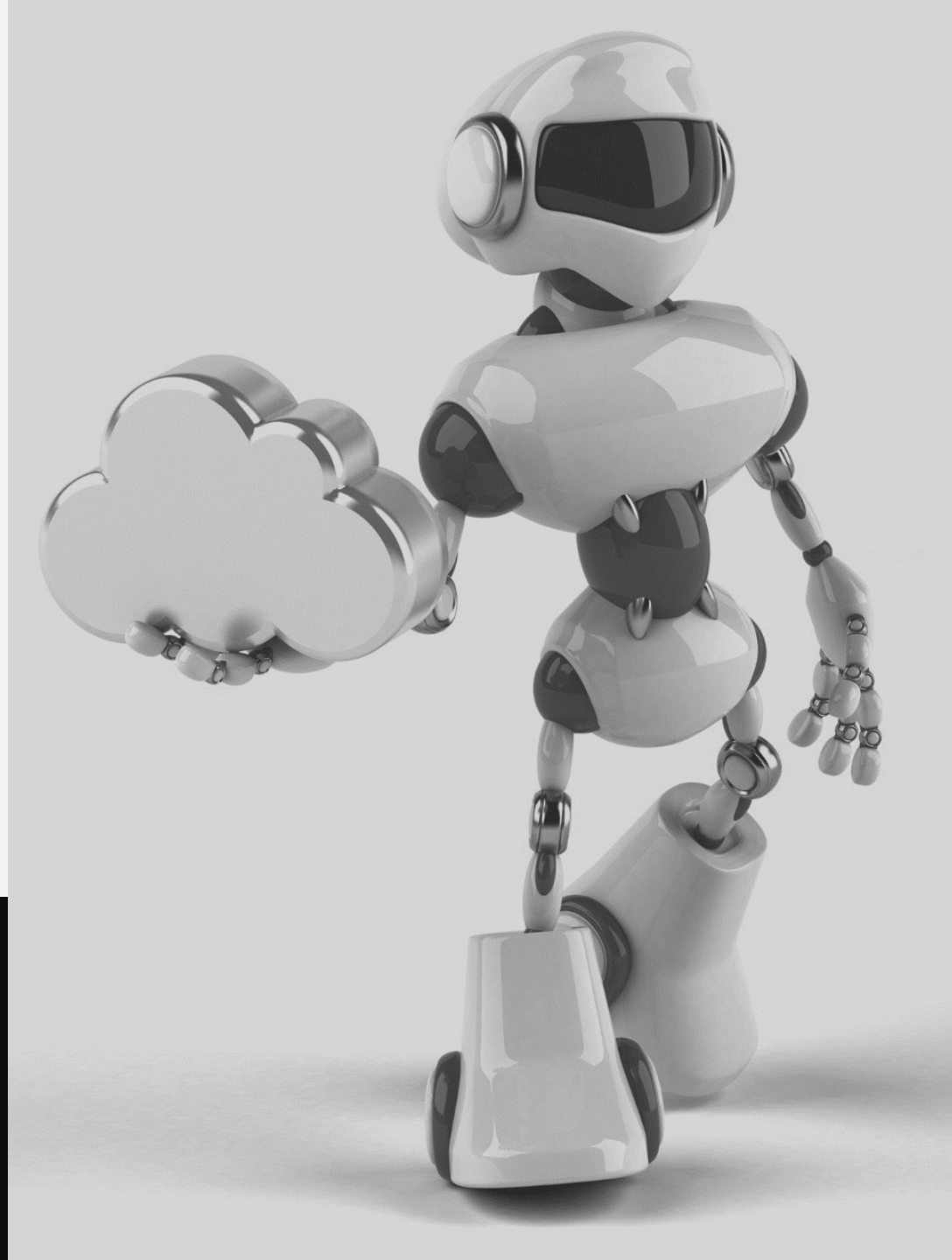


Problem Statement

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 scena
  - **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.**
  - **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.

Metadata Understanding



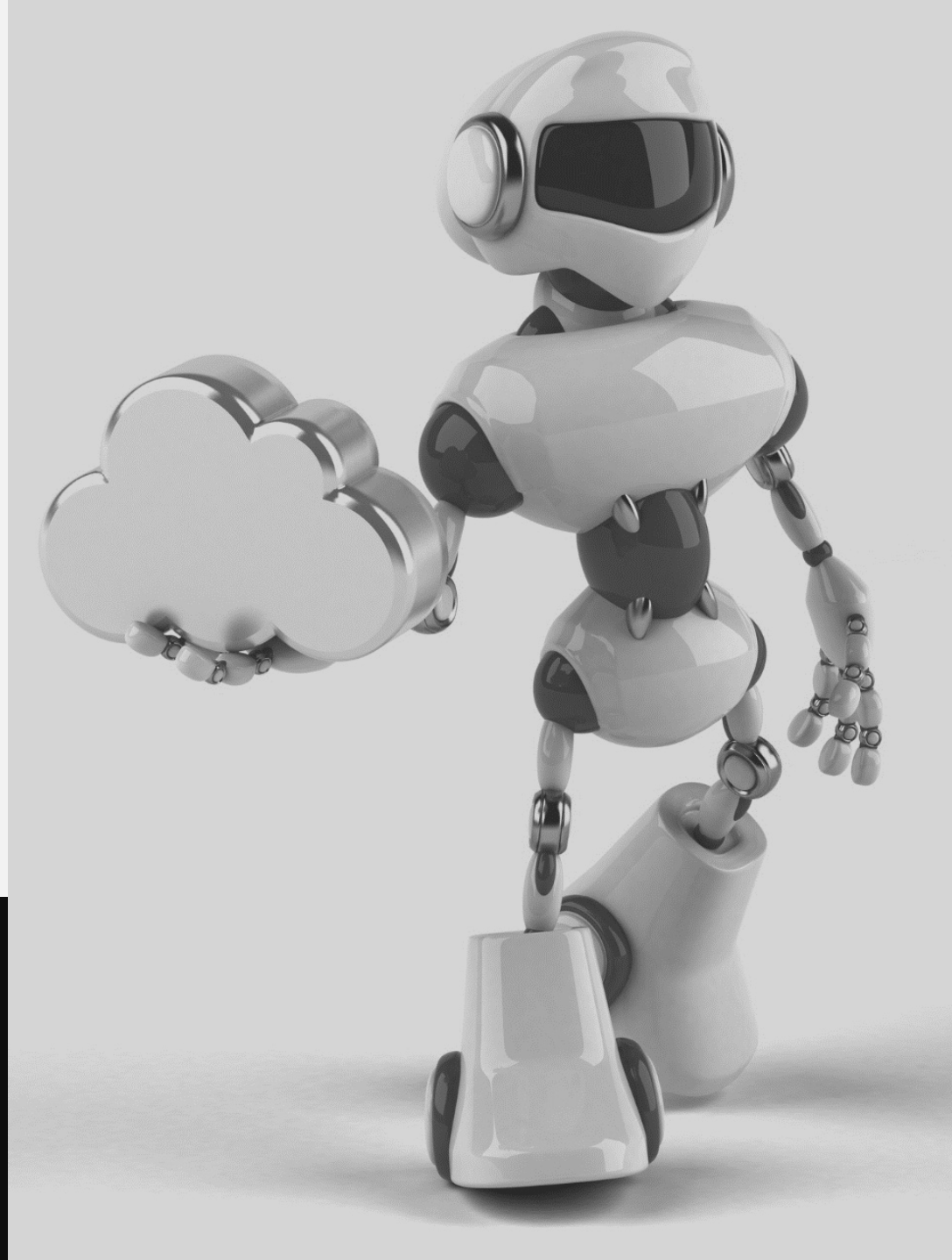
Meta Data Understanding

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-repo
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' combi
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.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
il_util	Ratio of total current balance to high credit/credit limit on all install acct
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_fi	Number of personal finance inquiries
inq_last_12m	Number of credit inquiries in past 12 months
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)

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



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.

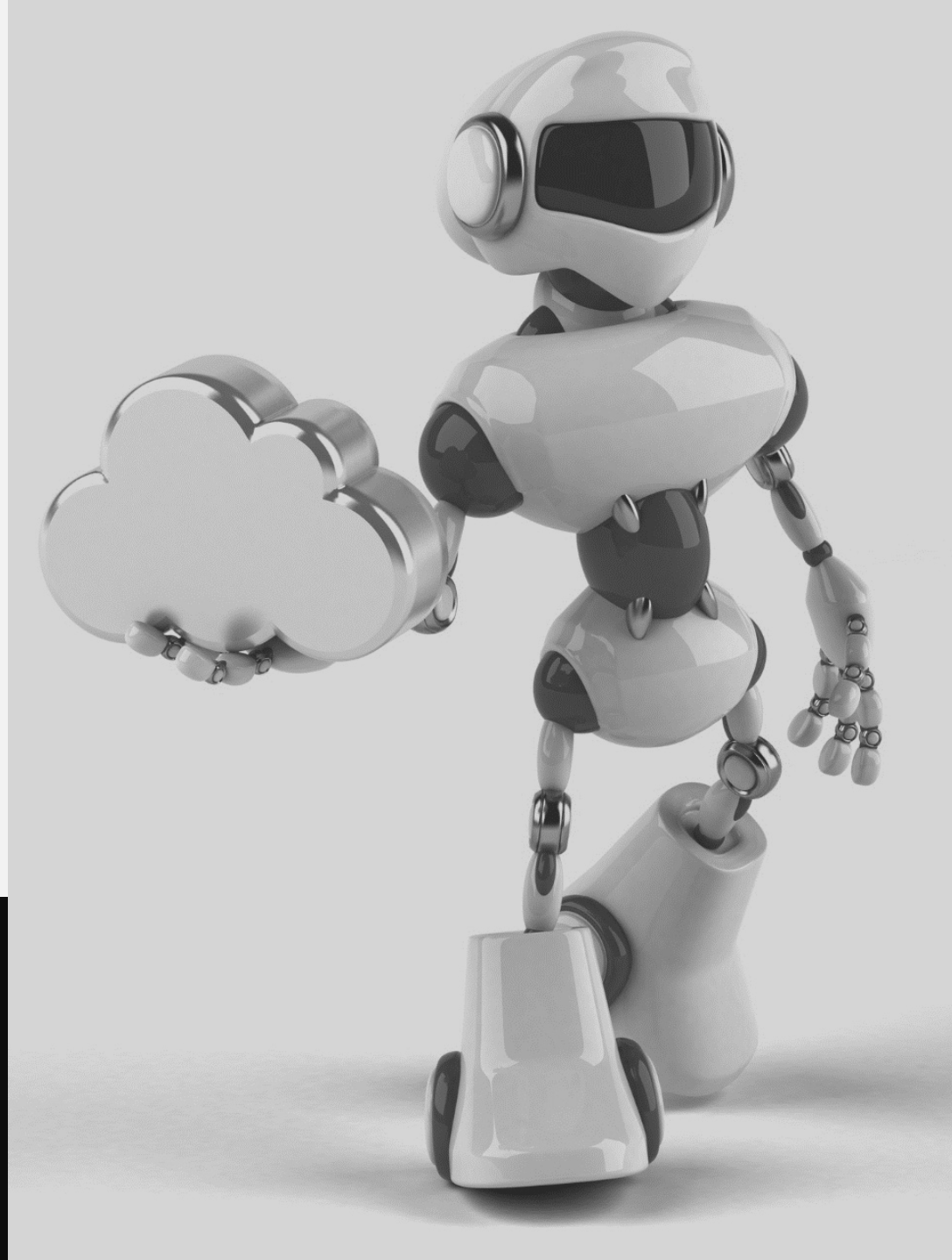
❖ **Customer Demographics:**

- **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 Cleaning and Manipulation

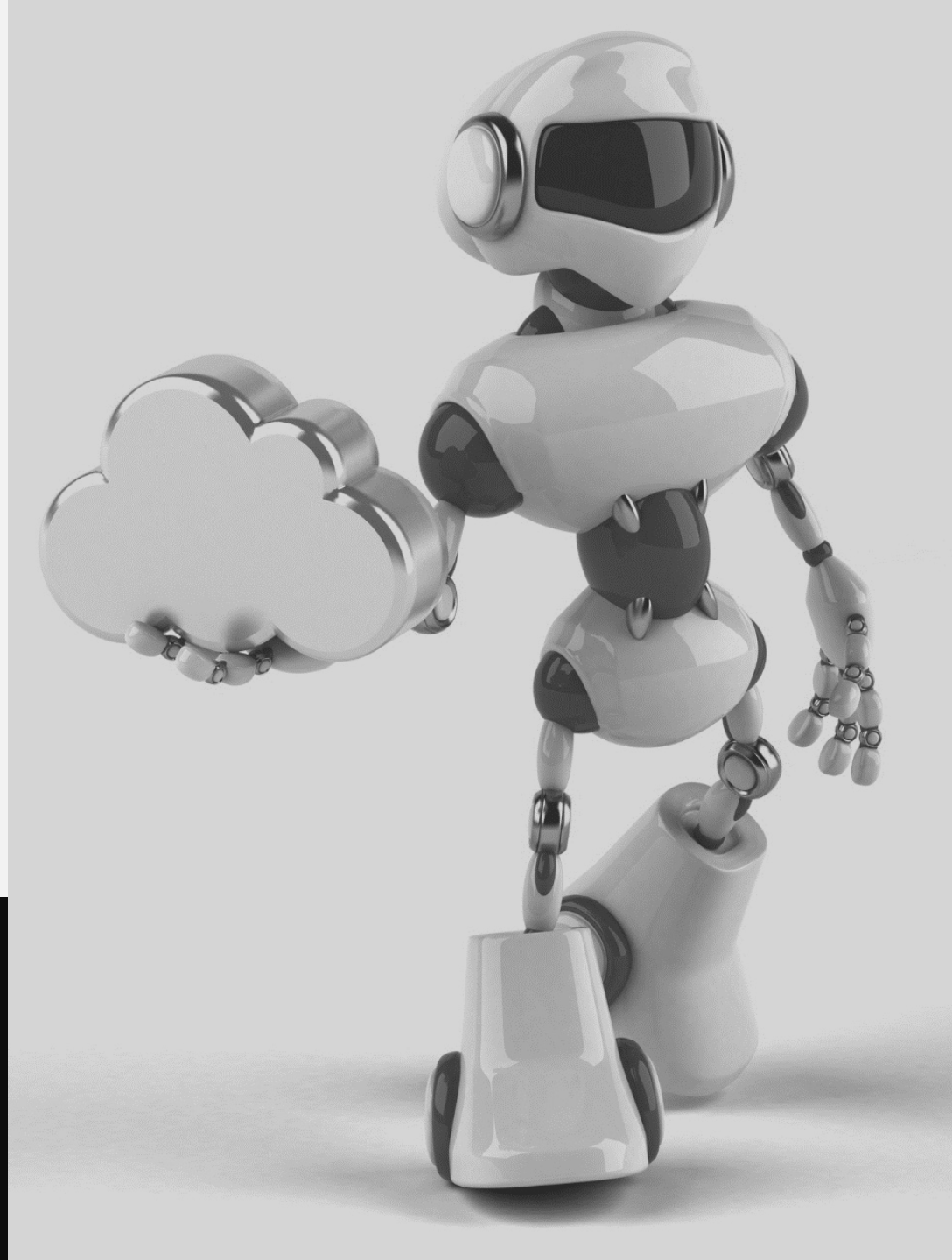


Data Cleaning and Manipulation

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

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

Univariate Analysis



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

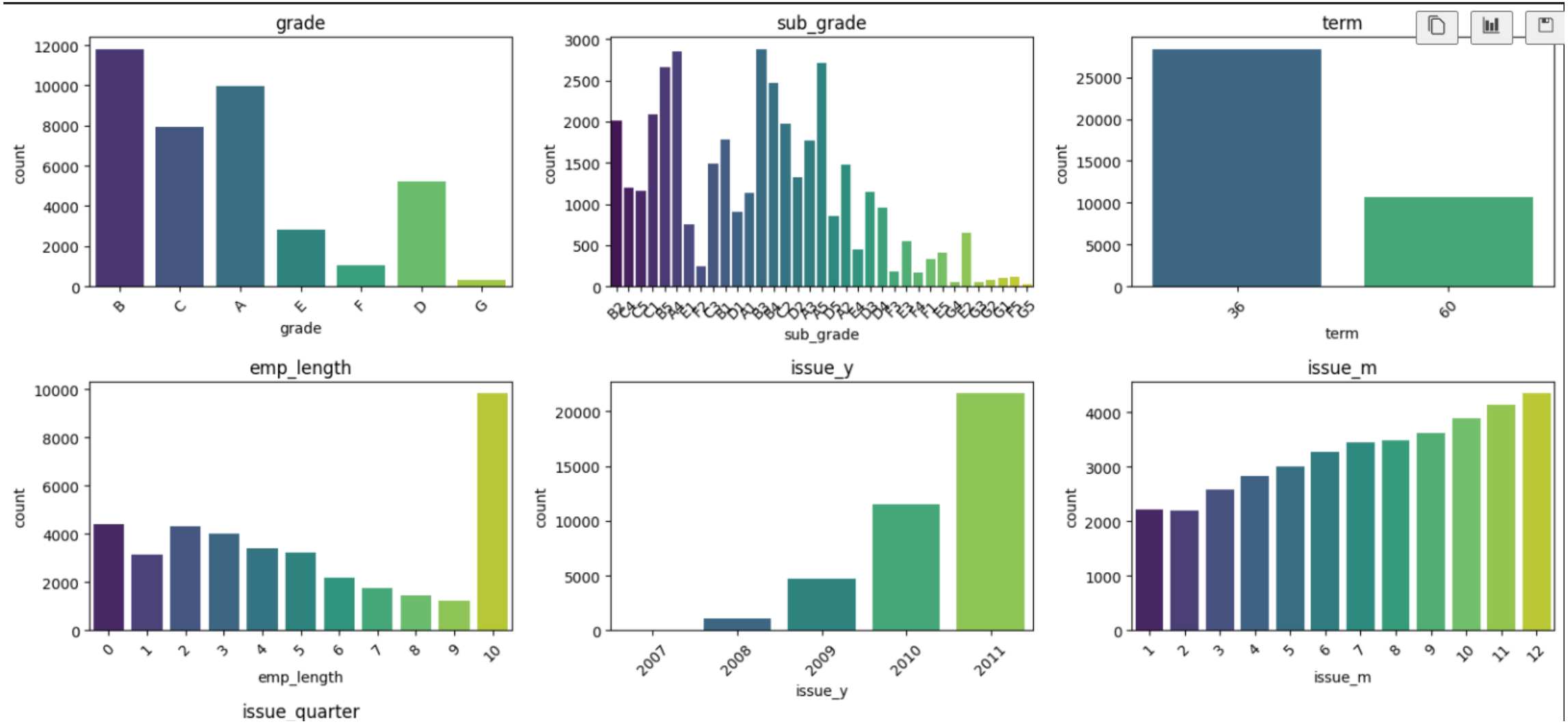
A. Categorical Variables:

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



Univariate Analysis – Categorical Variables

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.

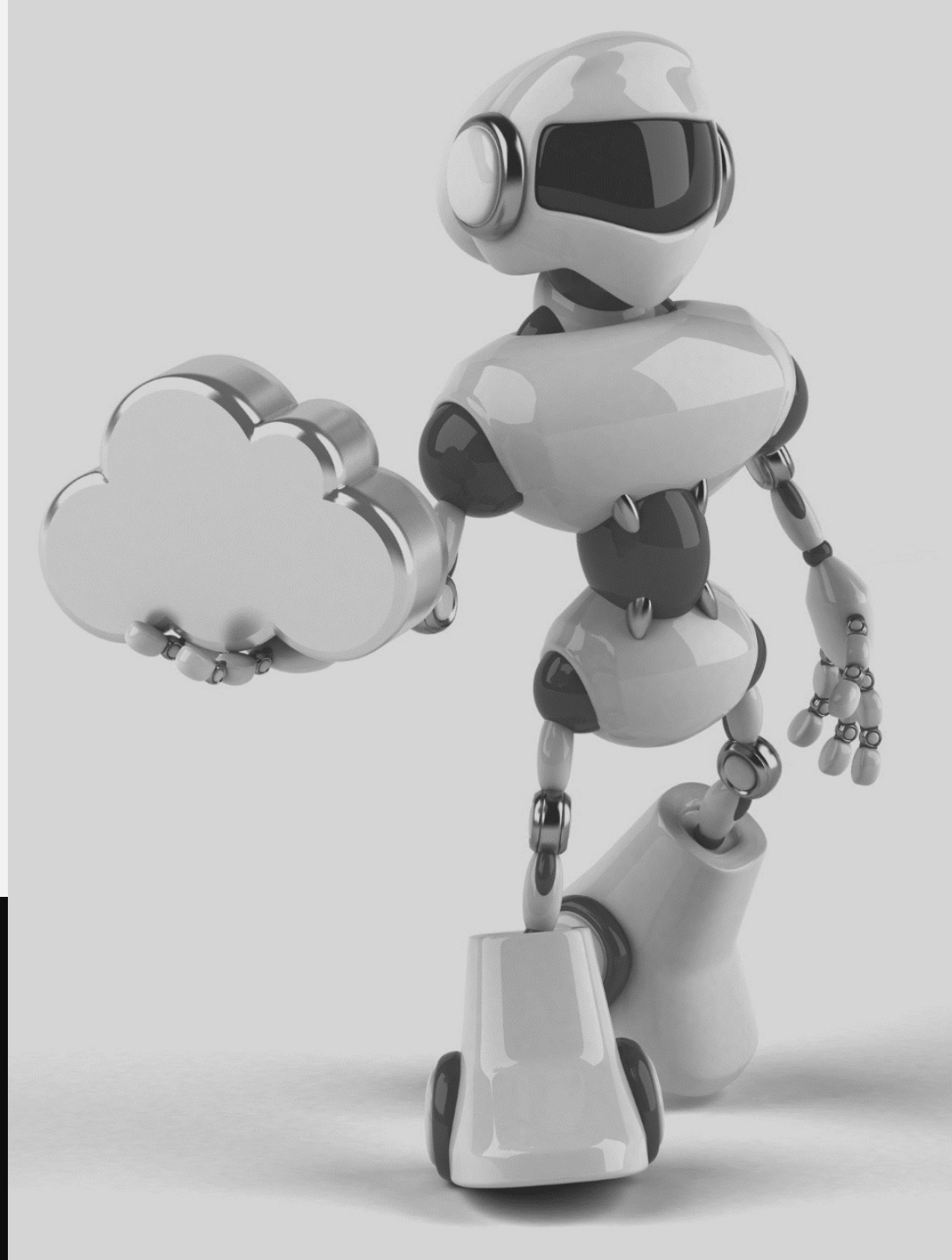
B. Unordered Categorical Variables:

- 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

- 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

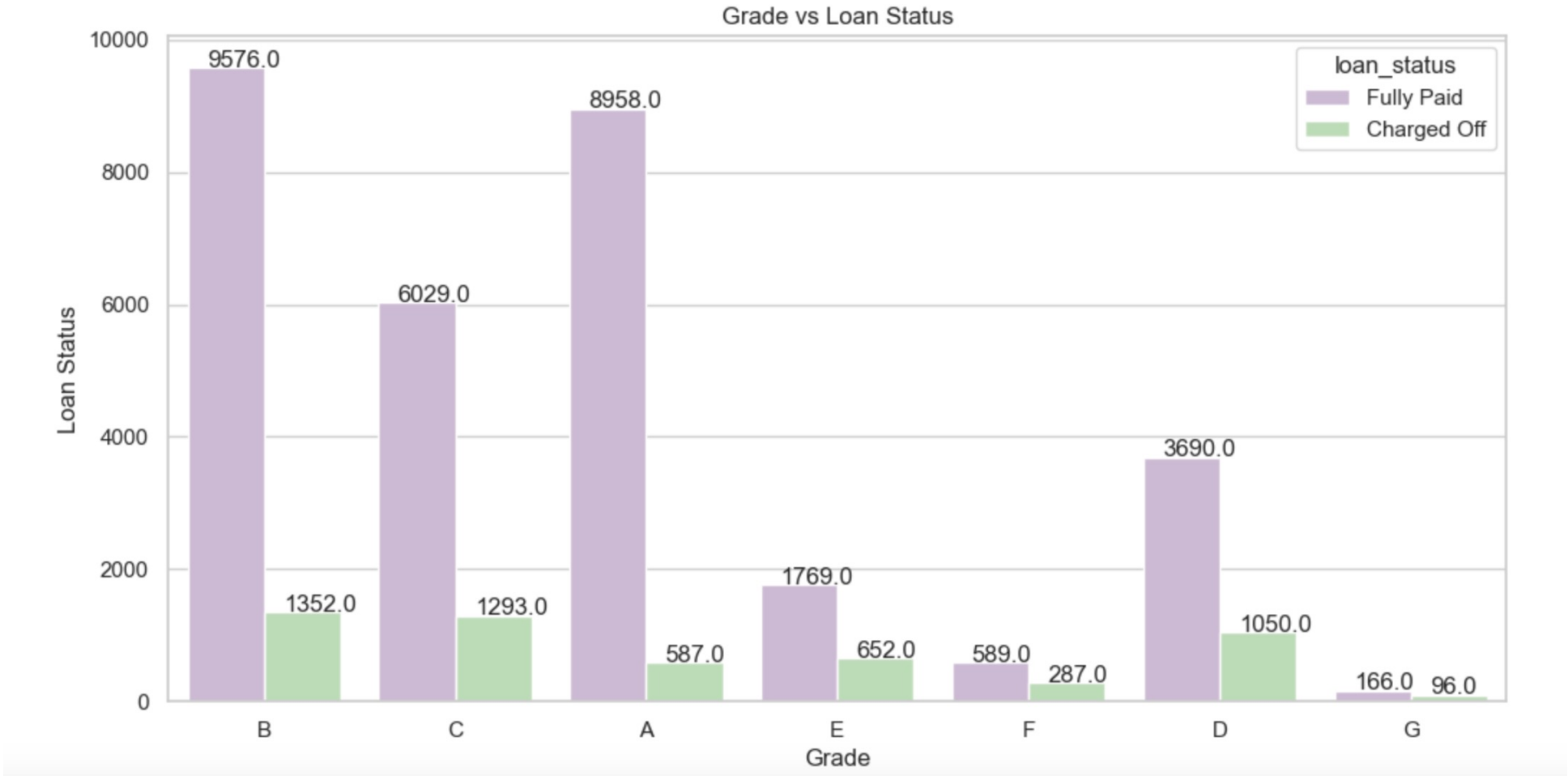
- Bivariate analysis is a statistical method that involves the simultaneous analysis of two variables (factors). It aims to determine the empirical relationship between them. The analysis can be used to test hypotheses, identify patterns, or explore relationships between the variables.
- It was carried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

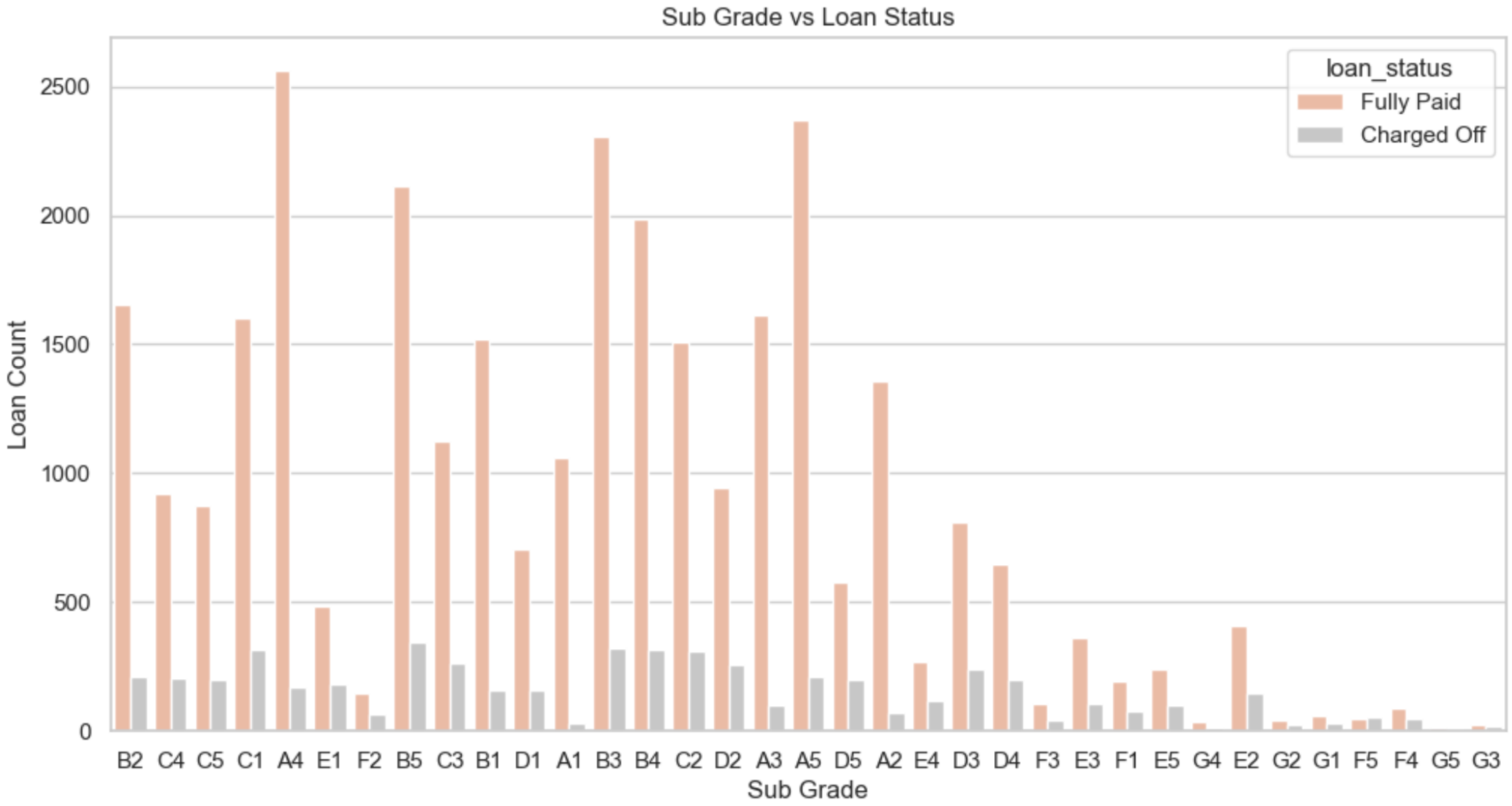
Quantitative Variables:

- 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)

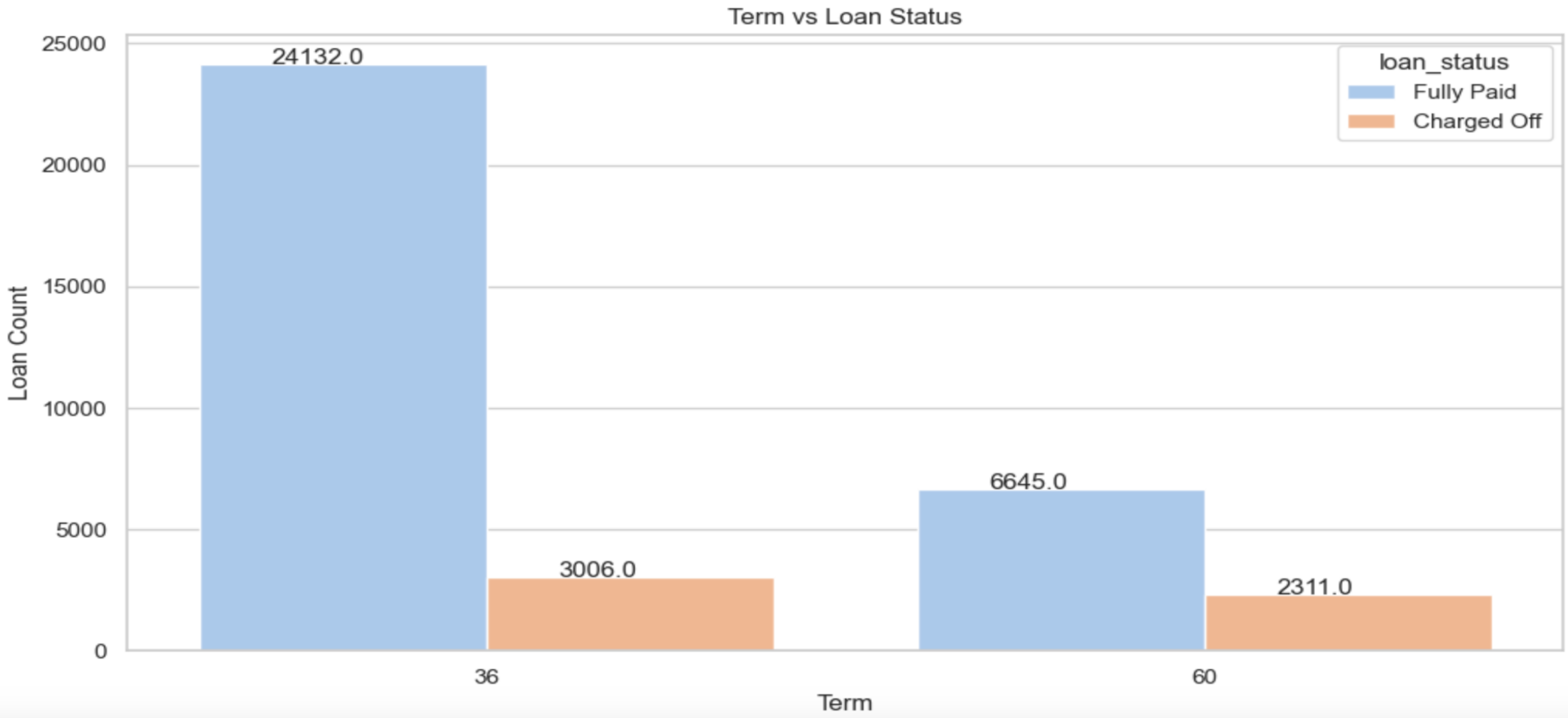
Bivariate Analysis – Grade vs. Loan Status



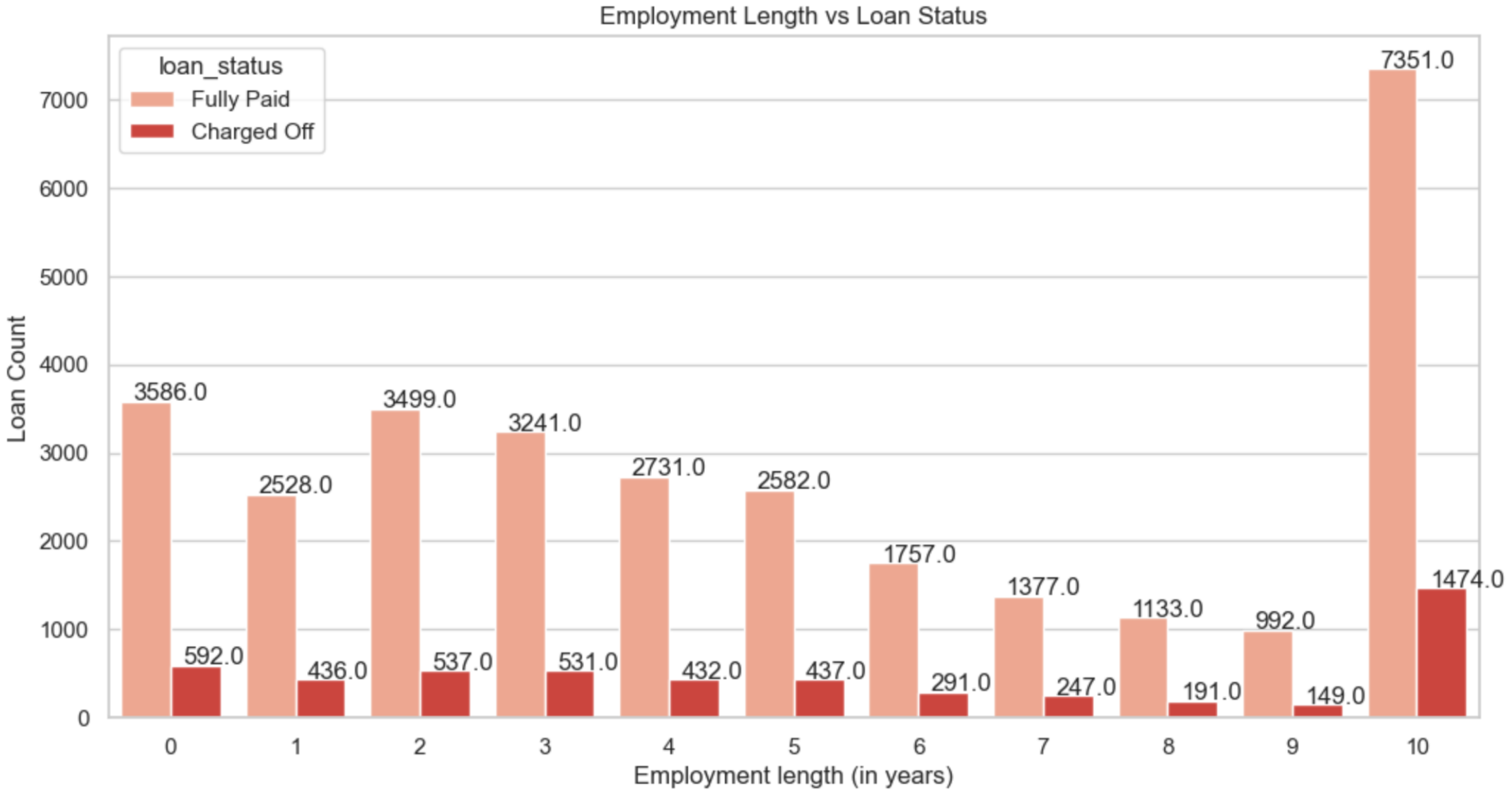
Bivariate Analysis – Sub Grade vs Loan Status



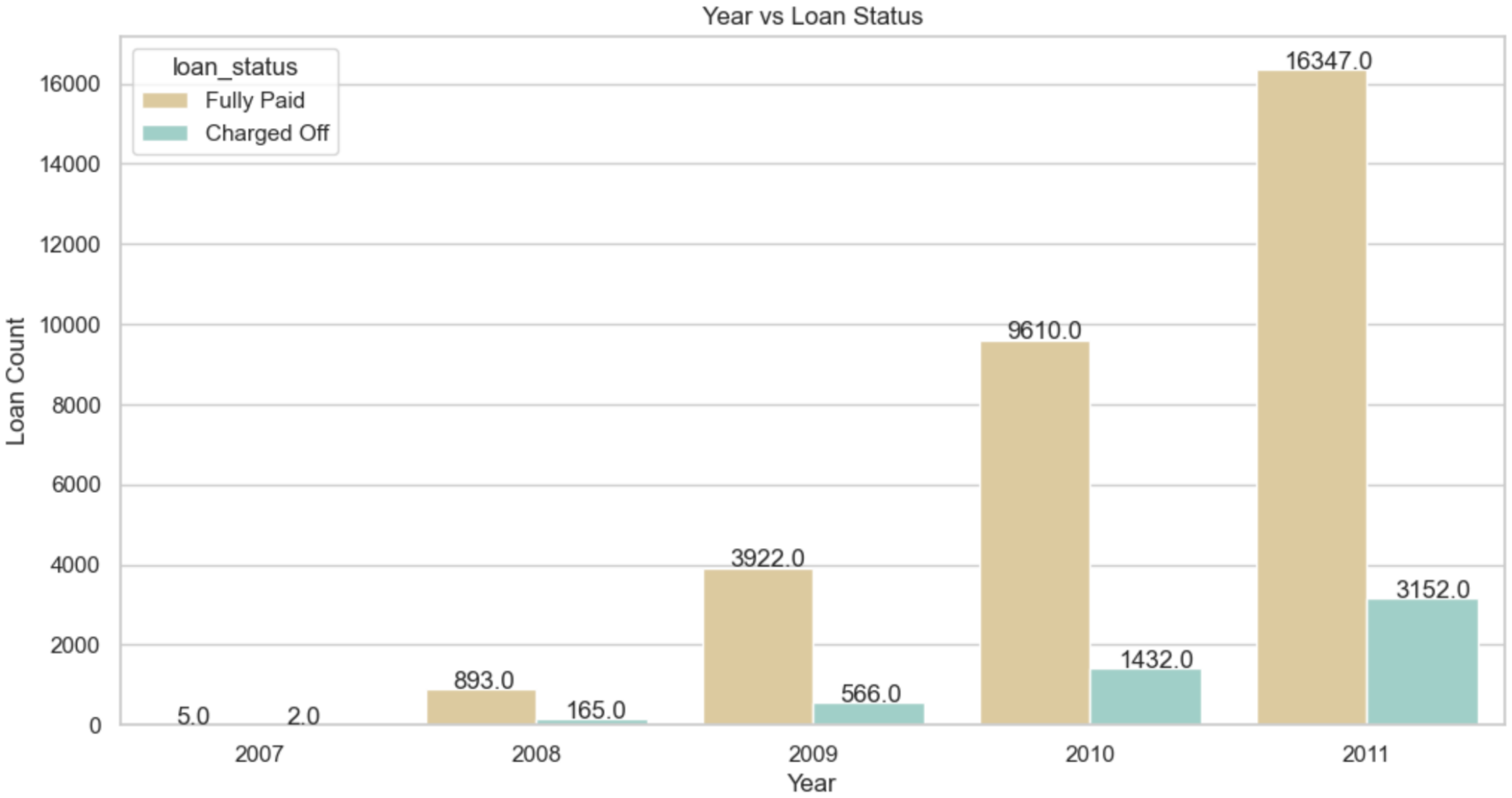
Bivariate Analysis – Term vs Loan Status



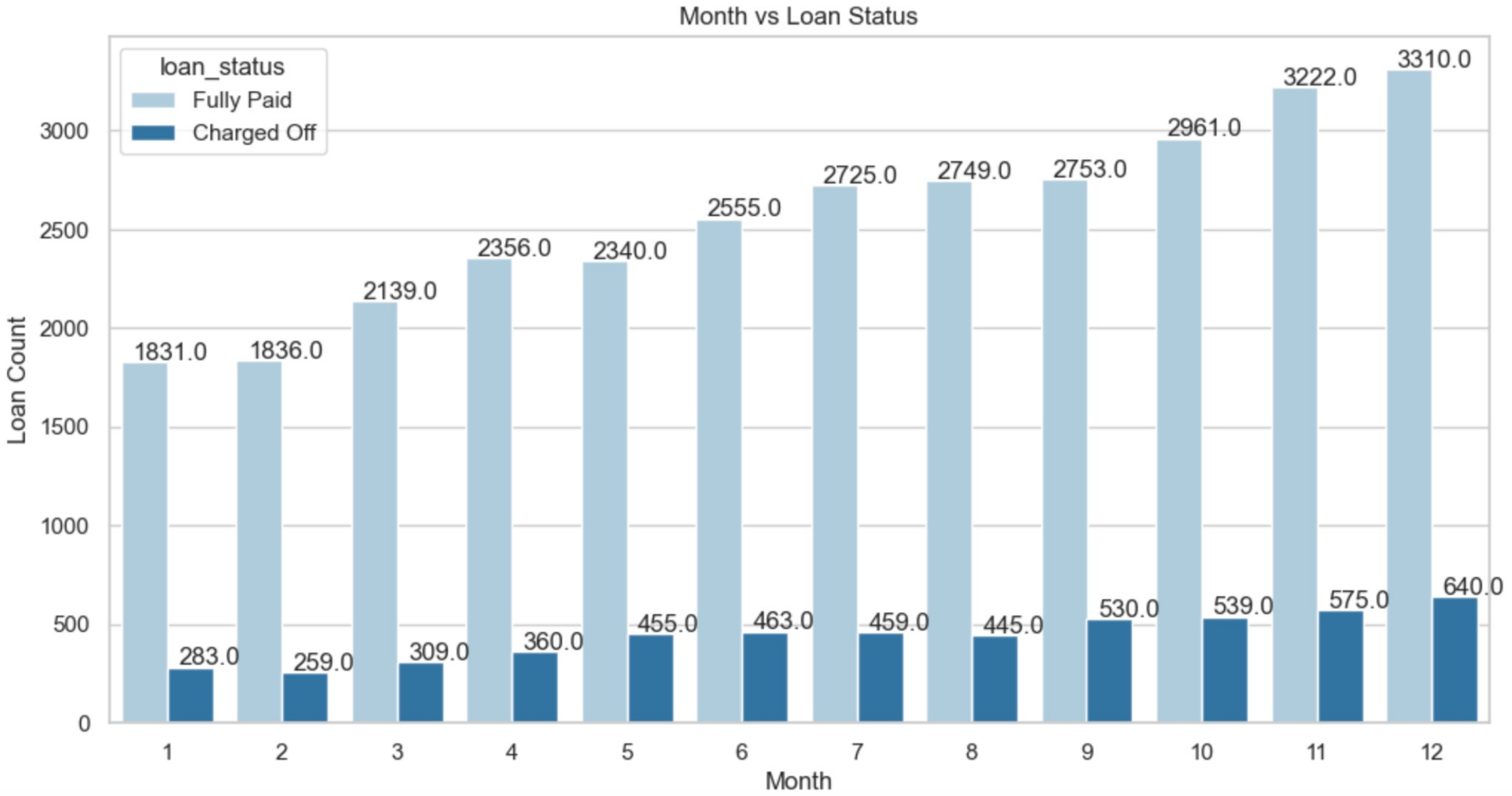
Bivariate Analysis – Employment Length vs Loan Status



Bivariate Analysis – Year vs Loan Status

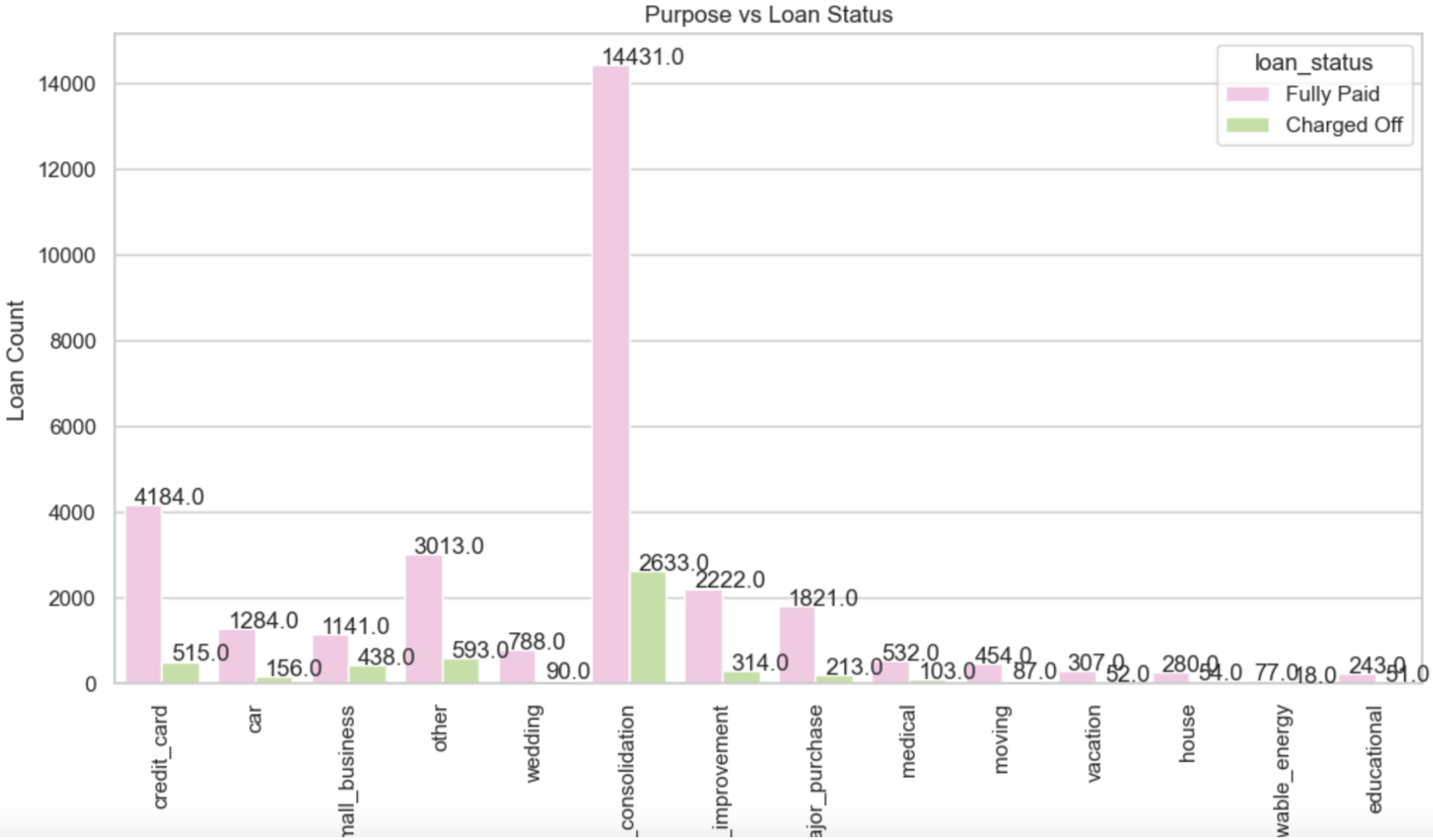


Bivariate Analysis – Month vs Loan Status

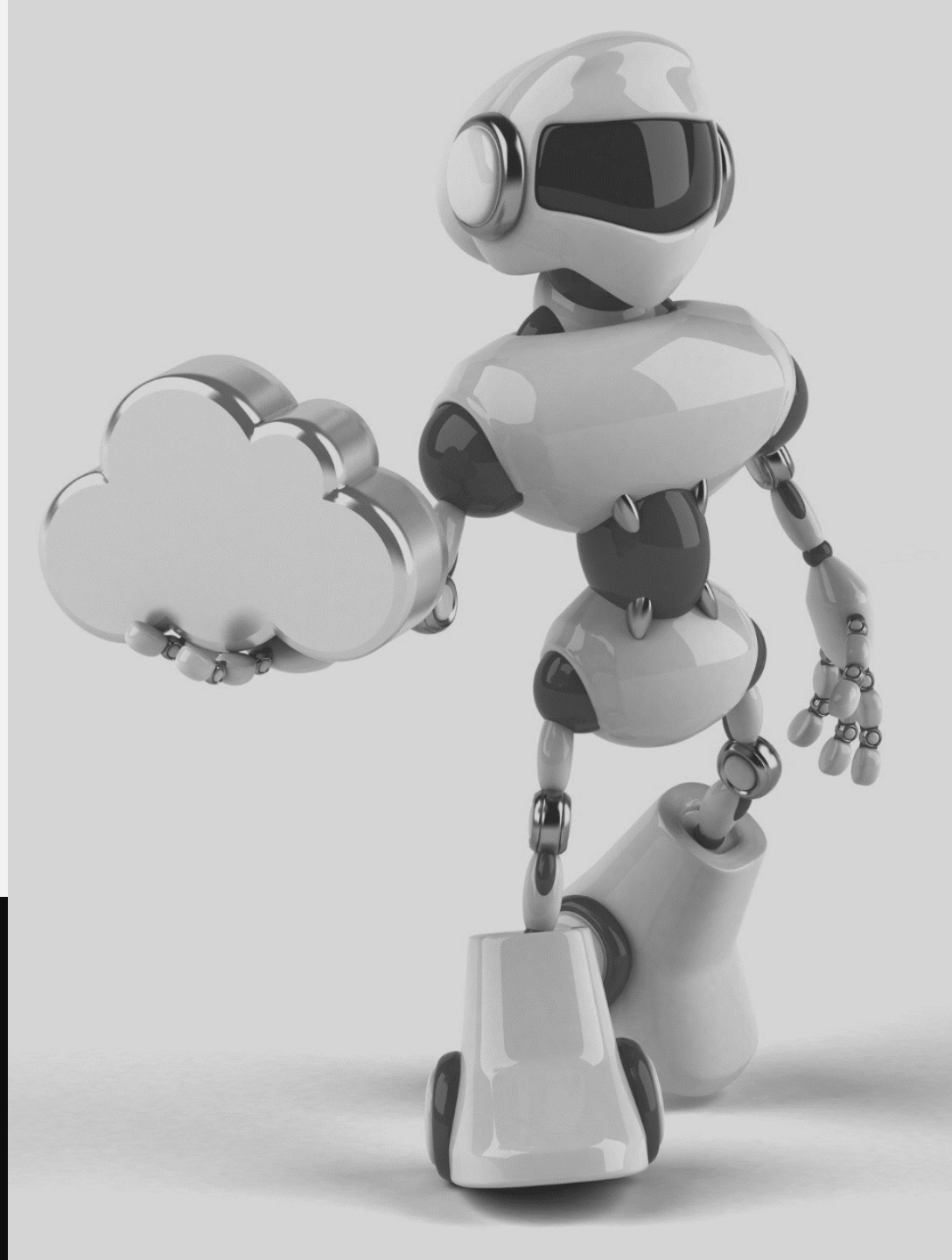


Bivariate Analysis – Quarter vs Loan Status

Bivariate Analysis – Purpose vs Loan Status



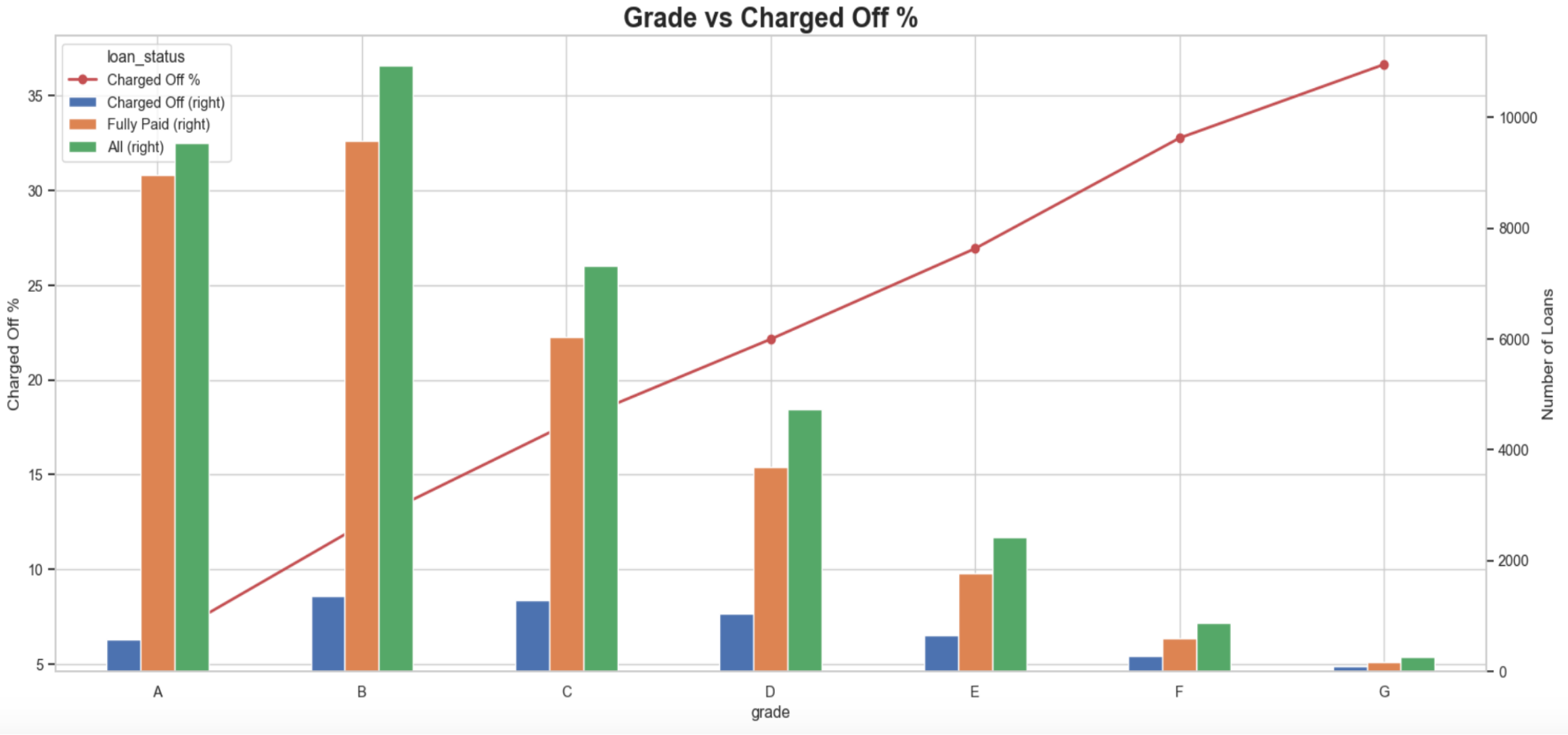
Multivariate Analysis



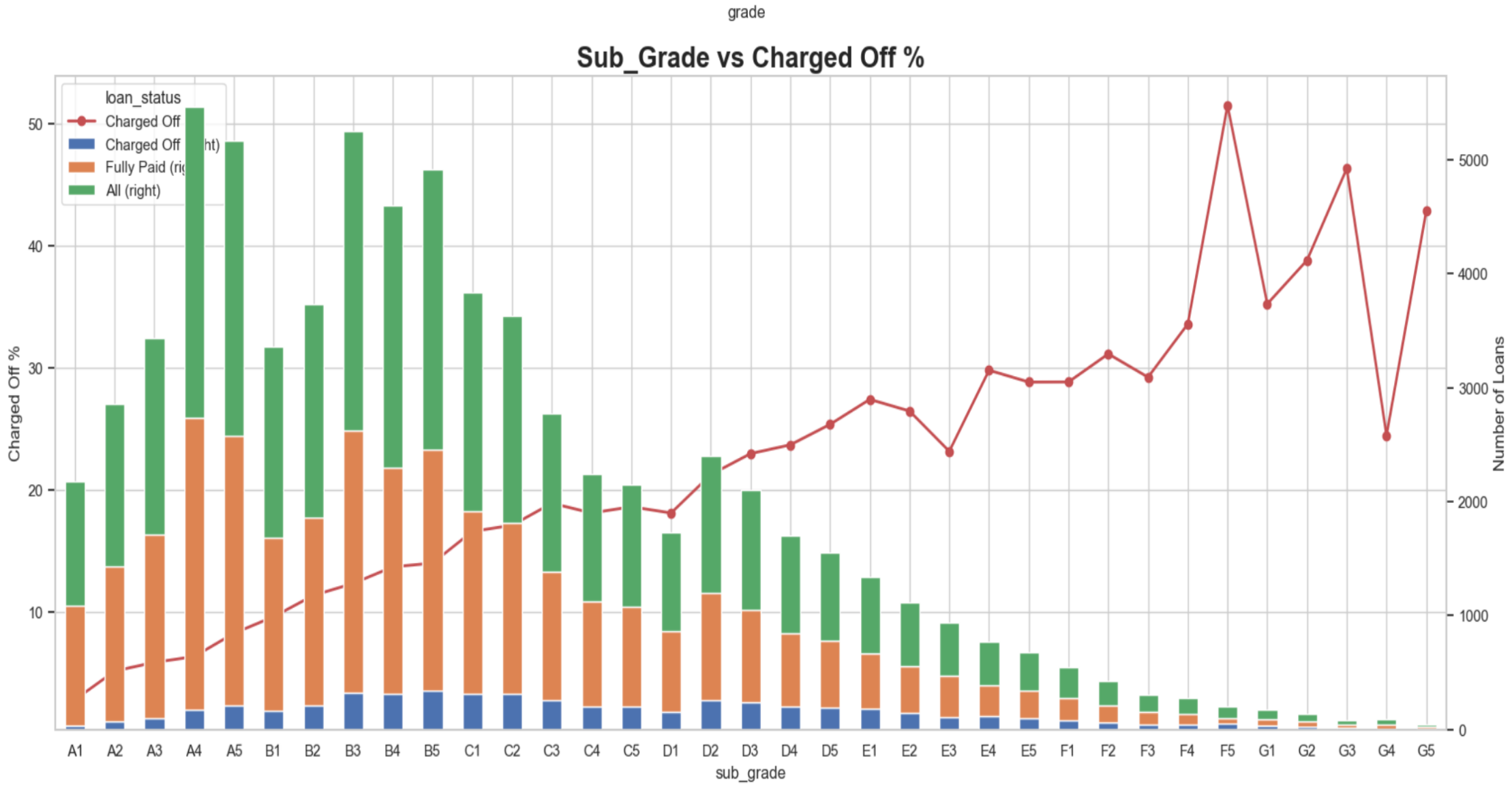
Multivariate Analysis

- **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

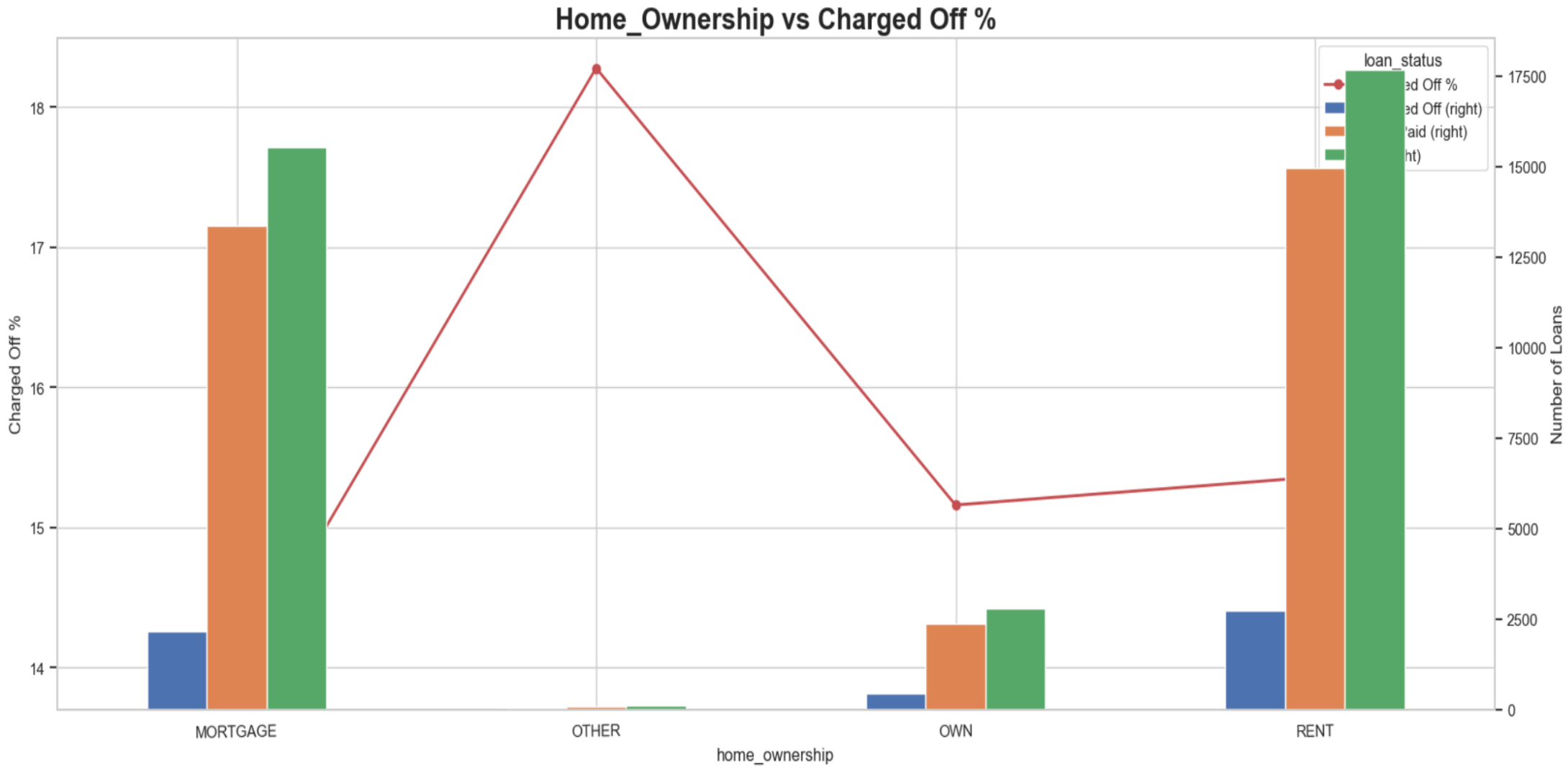
Multivariate Analysis



Multivariate Analysis



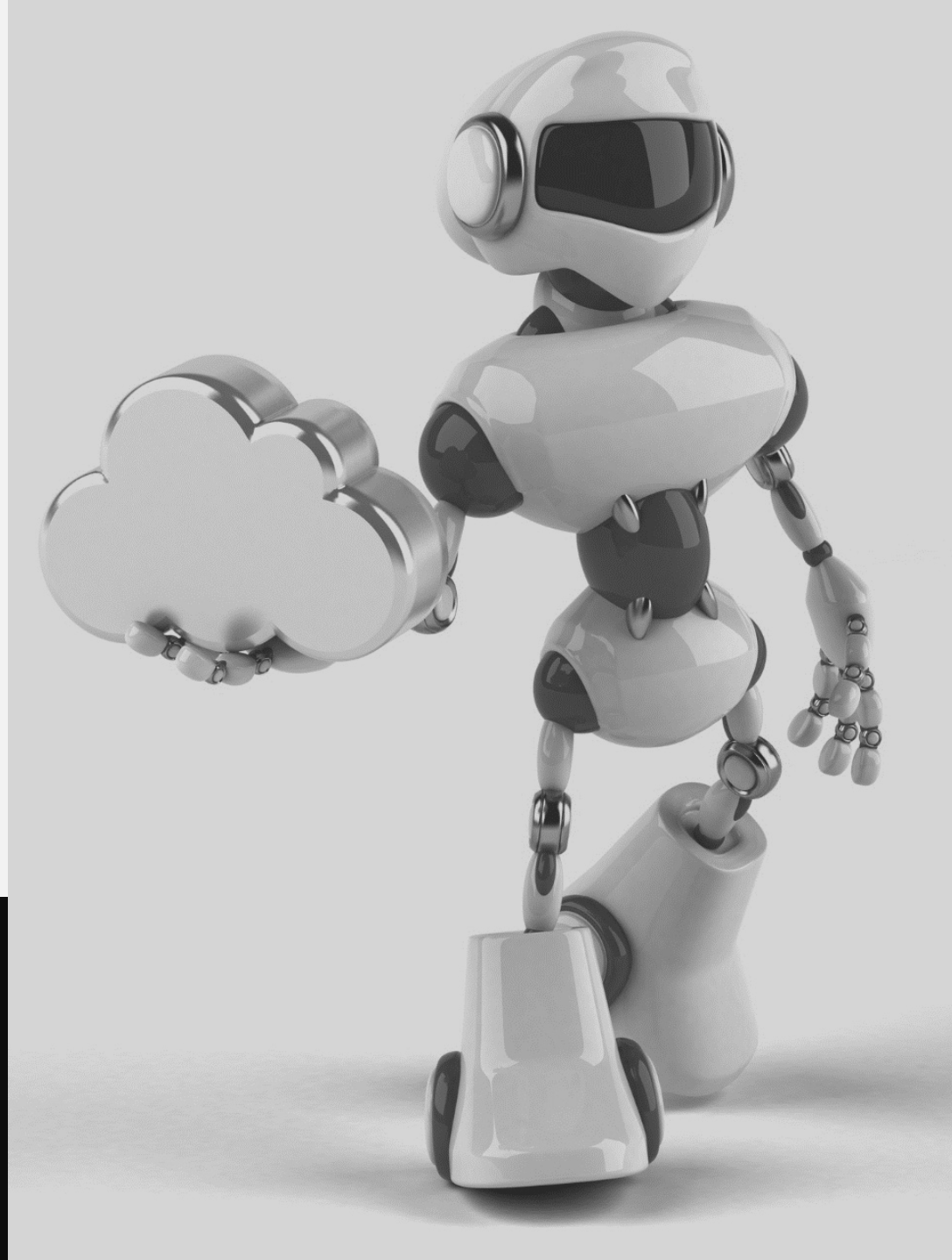
Multivariate Analysis



Summary of Multivariate Analysis

- 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 tendancy to default the loan.
- Borrowers from states CA, FL, NJ have maximum tendency to deafult 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

Suggestions



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.


Thank You
Happy to Connect