Lending Club Case Study

Submitted by:-
Piyush Kumar Roy
Prathamesh Kulkarni

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:
  - **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.

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

Data Understanding

Primary Attribute

Loan Status: The target field (*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 this case study, rows with a "Current" status will be excluded from the analysis.

Data Understanding

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 are actively in the process of making loan installments, hence, the loan tenure has not yet concluded.
- **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.

Data Understanding

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, rents, or mortgages. Homeownership 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.
- **Revolving Utilities (revol_utils):** Measures how much pending debt the applicant is yet to pay. Lowering the value increases the probability of loan approval.
- **State (addr_state):** Denotes the customer's location and can be utilized to create a generalized demographic analysis. It may reveal demographic trends related to delinquency or default rates.

Data Understanding

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.
- **Term (term):** Duration of the loan in months.
- **Loan Date (issue_d):** Date when the loan was issued.
- **Purpose of Loan (purpose):** Indicates the reason for which the borrower is seeking the loan, such as debt consolidation, home improvement, vacation, or other purposes.
- **Verification Status (verification_status):** Represents whether the details provided by the borrower have been verified by the lender or not.
- **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.

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.

- There are columns that represent the detailed information or behavior of borrowers which won't be utilized for the analysis, such as desc, id, url, zip_code, and emp_title.

- There are columns which represent post-default data, such as collection_recovery_fee, recoveries, and total_rec_late_fee.

Data Cleaning & Pre-processing

- **Loading data from loan CSV**
- **Checking for null values in the dataset:**
  1. There are 54 empty columns in the data. Dropping those columns from the dataset and then checking the dataset shape
  2. There are 3 columns in the data where missing data is more than equal to 60%. Dropping those columns from the dataset and then checking the dataset shape
- **Checking for unique values:**

  There are 9 columns in the data which have single value. Dropping those columns from the dataset and then checking the dataset shape
- **Dropping records:**

  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 etc.

Data Cleaning & Pre-processing

- **Data Conversion and Standardization:**

  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).

- **Checking for duplicated rows in data:**

  There are no duplicate rows in the data.

- **Common Functions:**

  Made common functions to plot boxplot, barplot, lineplot, countplot and heatmaps

- **Outlier Treatment:**

  1. Made boxplots for annual income, loan amount, interest rates, funded amount, installments, debt to income ratio
  2. Dropped few outliers wherever required

Data Cleaning & Pre-processing

Observations and Inferences from Boxplots

- The **annual income** of most loan applicants is between **$40,000 - $80,000**.
- The **loan amount** for most applicants is between **$5,000 - $15,000**.
- The **debt-to-income ratio** is between **8 - 18**.
- The **interest rate** on loans is between **8% - 14%.**
- The **installments** on the loans range from **$160 - $400**.
- The **funded amount by investors** for most applicants is between **$5,000 - $14,000**.

Univariate Analysis

1. Categorical Variables

Ordered Categorical Data:

1. Grade (grade)
2. Subgrade (sub_grade)
3. Term (36 / 60 months) (term)
4. Employment Length (emp_length)
5. Issue Year (issue_y)
6. Issue Quarter (issue_q)
7. Public Derogatory Records (pub_rec)
8. Number of Enquiries (inq_last_6mths)

Univariate Analysis

Unordered Categorical Data:

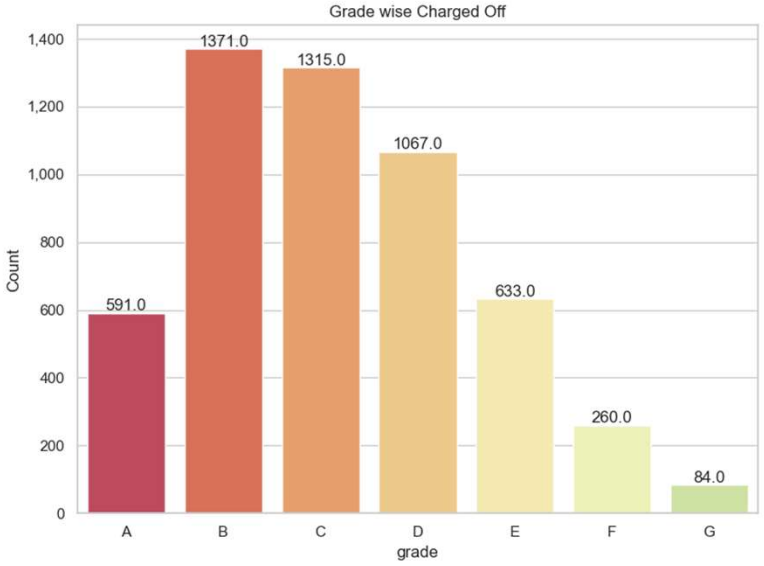
- Address State (addr_state)
- Loan Purpose (purpose)
- Home Ownership (home_ownership)
- Loan Status (loan_status)
- Verification Status (verification_status)

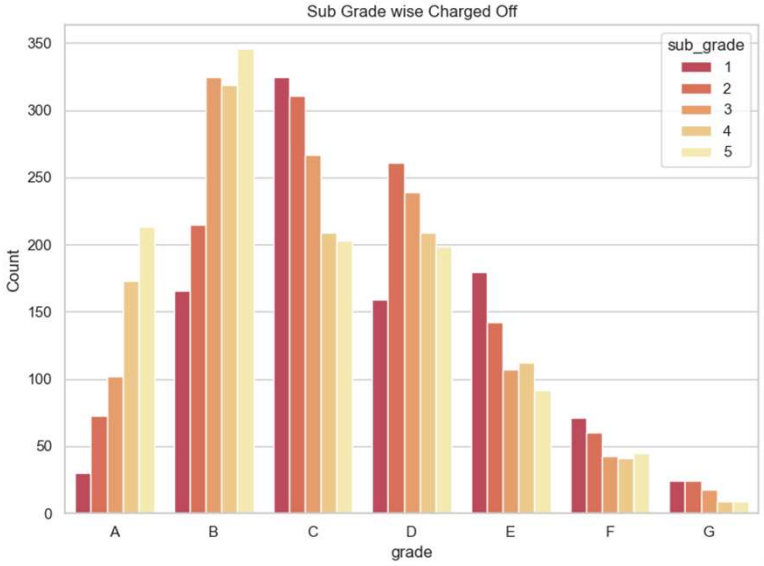
Univariate Analysis

Quantitative Variables

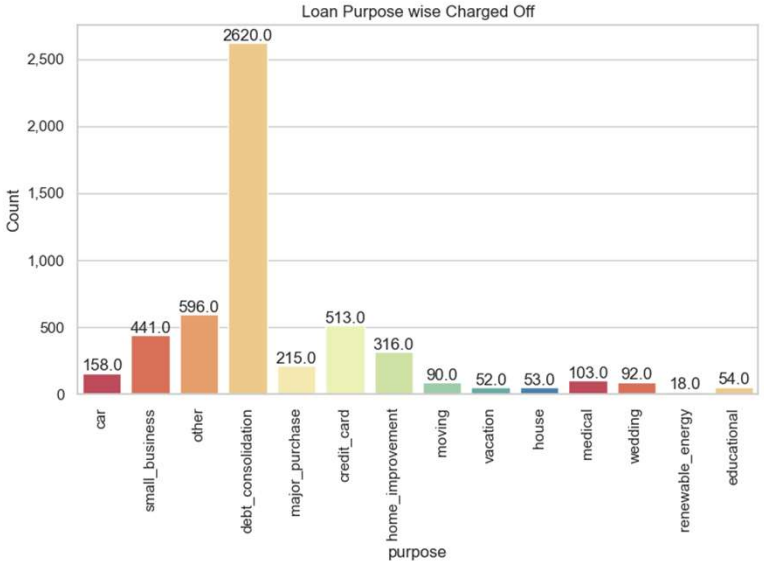
- Interest Rate Bucket (int_rate_range)
- Annual Income Bucket (annual_inc_range)
- Loan Amount Bucket (loan_amnt_range)
- Debt to Income Ratio (DTI) Bucket (dti_range)
- Revolving Utilities (revol_util_range)
- Funded Amount Invested Bucket (funded_amnt_inv_range)

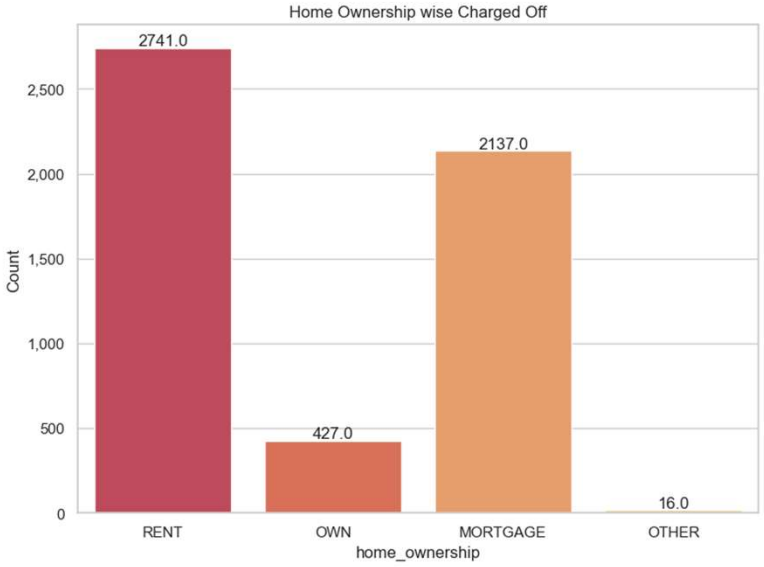
Univariate Analysis



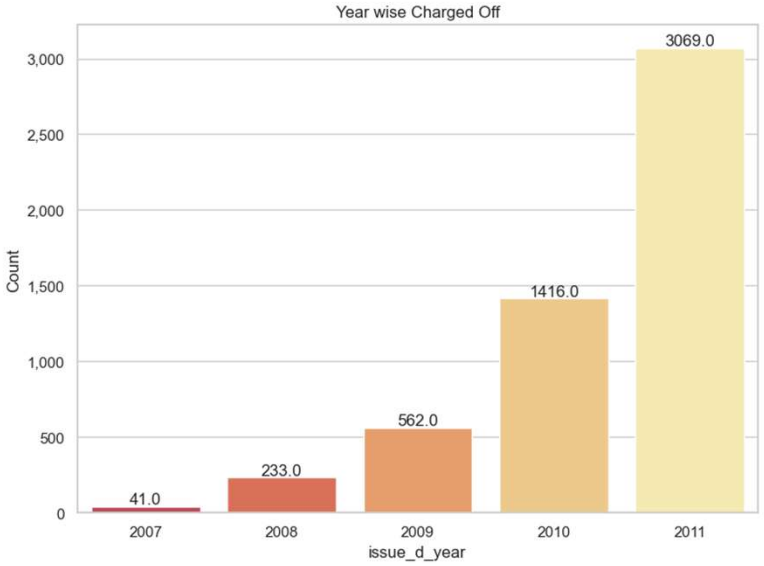


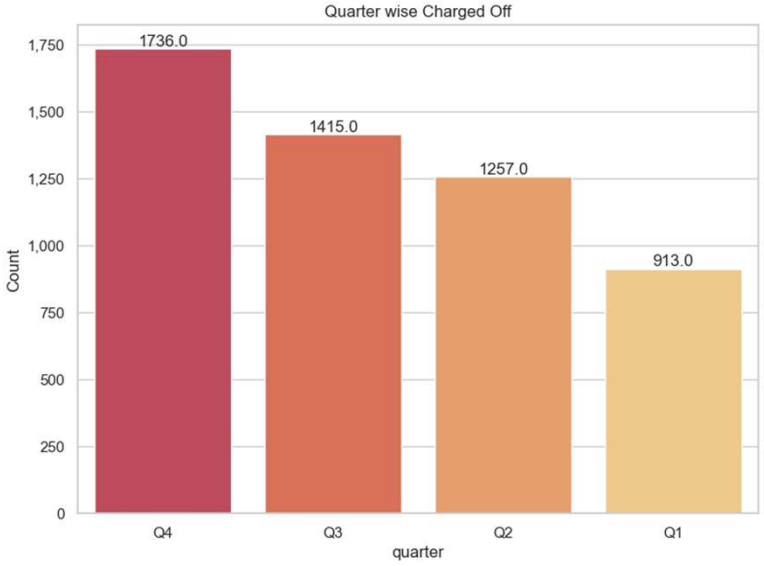
Univariate Analysis



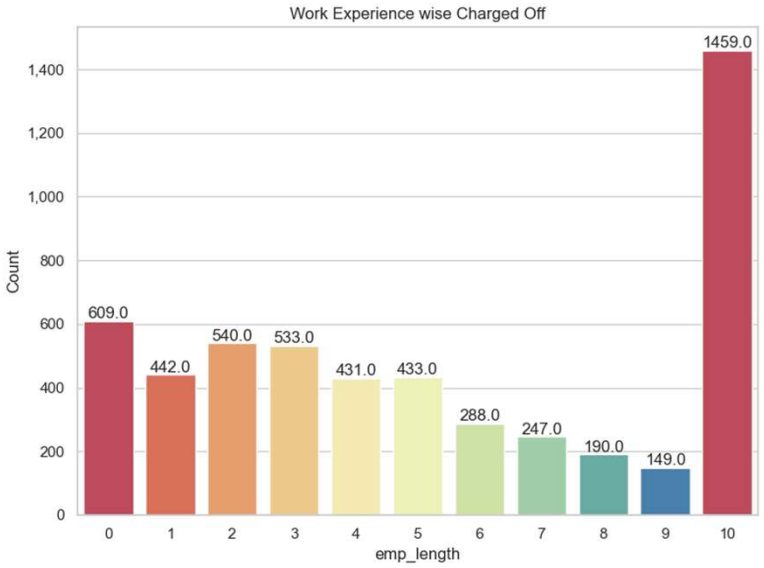


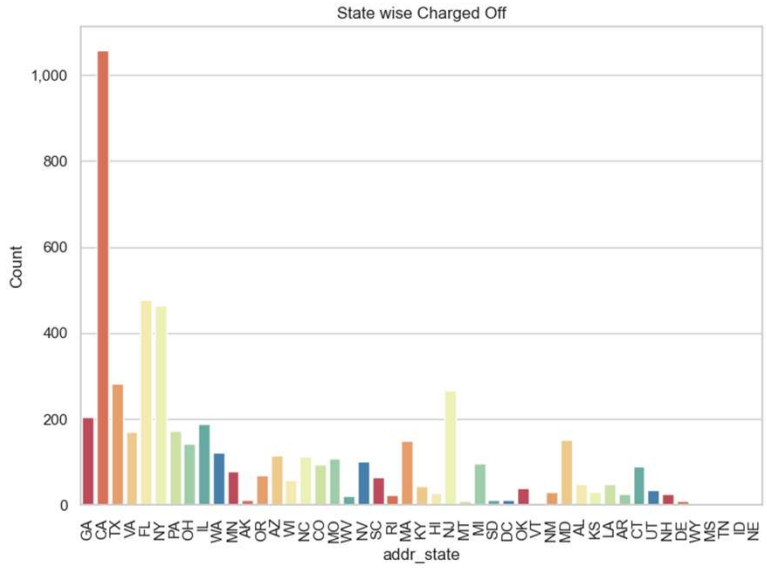
Univariate Analysis



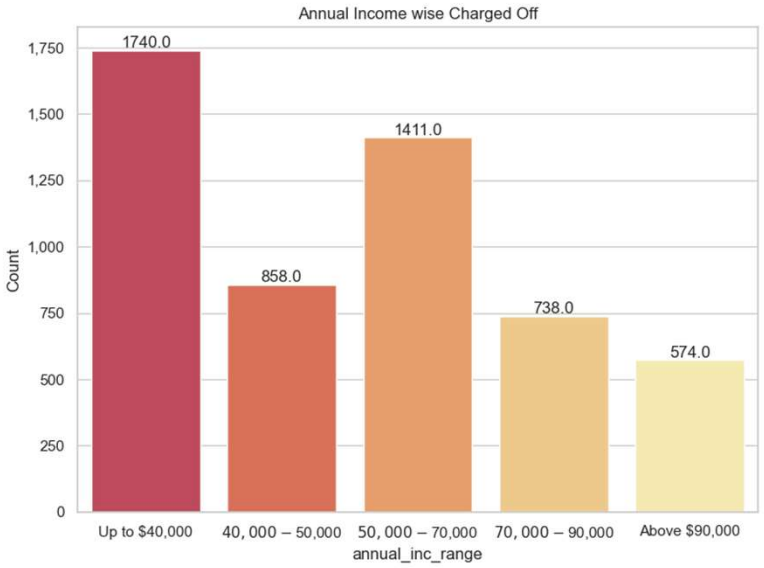


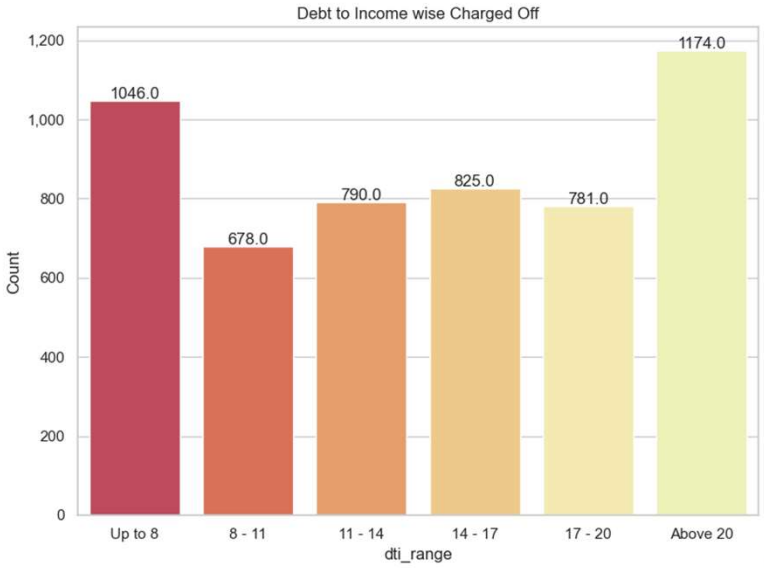
Univariate Analysis



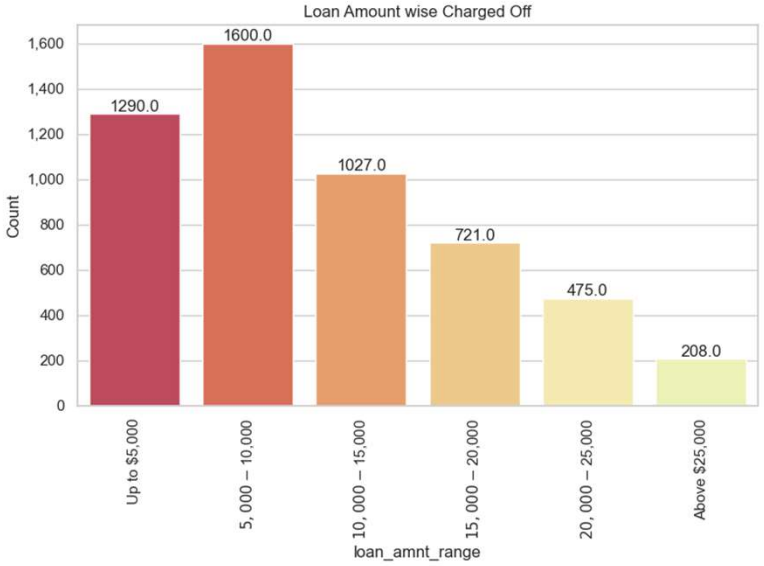


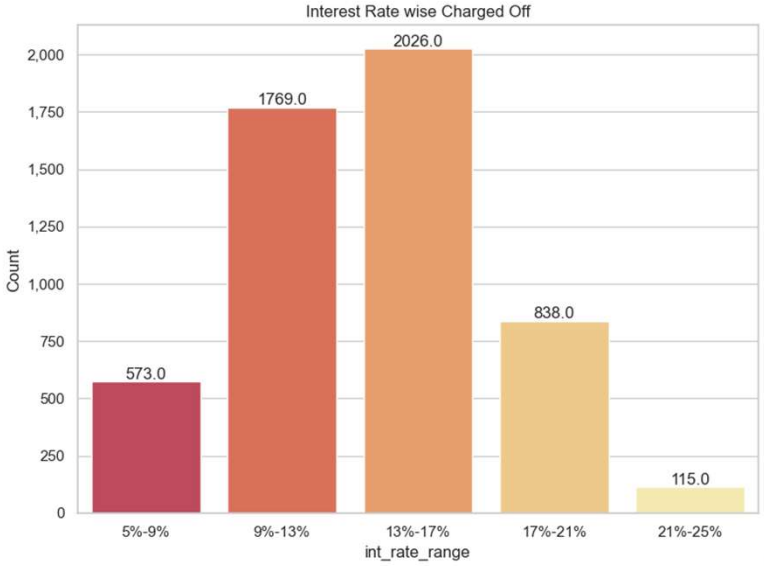
Univariate Analysis





Univariate Analysis





Univariate Analysis

Ordered Categorical Fields

- **Grade B and Grade C** had the highest number of "Charged off" loan applicants, with 1,371 and 1,321 applicants respectively, indicating that applicants with these credit grades faced challenges in repaying their loans.
- Among the two grades, **Sub-grade 5 in Grade B** and **Sub-grade 1 in Grade C** had the highest number of "Charged Off" applicants, showing that applicants with these specific sub-grades had difficulty repaying their loans.
- **Short-term loans of 36 months** had the highest number of "Charged off" applicants, with 3,063 defaults, indicating that many applicants who chose shorter repayment terms struggled to repay their loans.
- **Applicants employed for more than 10 years** had the highest number of "Charged off" loans, totaling 1,459, showing that long-term employment does not necessarily guarantee successful loan repayment.
- The year **2011** recorded the highest number of "Charged off" loan applications, totaling 3,069, possibly due to the **2011 U.S. Debt Ceiling Crisis**.
- **Fourth quarter loans** had the most "Charged off" applicants (1,736), suggesting that financial pressures during the holiday season may have contributed to defaults.
- **Applicants with no public derogatory history** accounted for 4,941 "Charged Off" loans, showing that having no prior derogatory records does not necessarily mean lower default risk.
- **Applicants with no inquiries** made up 2,191 "Charged Off" loans, indicating that the number of inquiries had little impact on default risk.

Univariate Analysis

Unordered Categorical Fields

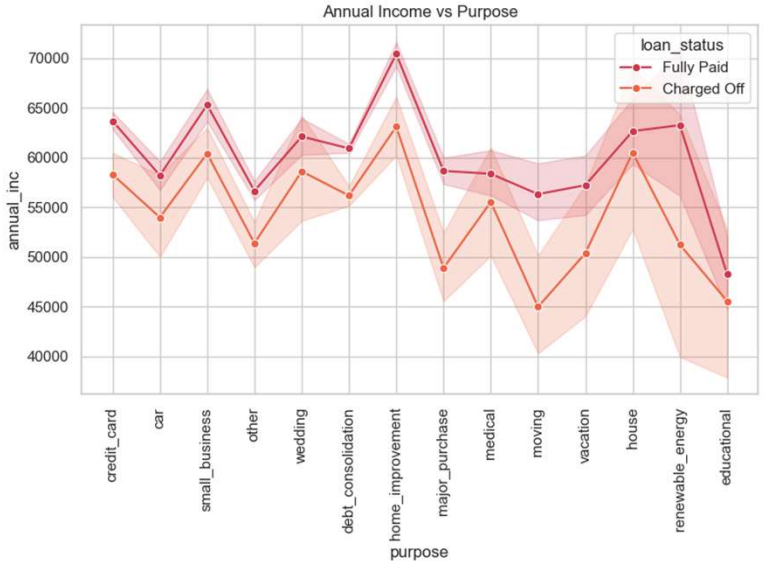
- **California, Florida, and New York** had the highest number of "Charged off" loan applicants, with 1,059, 478, and 463 defaults respectively. Stricter eligibility criteria may be needed in these states due to the higher number of defaults.
- One reason for the high number of defaults could be that **California, Florida, and New York** contribute more to the U.S. GDP, leading to a higher number of loan applications.
- **Debt consolidation** was the primary loan purpose for 2,620 "Charged off" applicants. The lending company should be cautious when approving loans for this purpose.
- **Renters** accounted for 2,741 "Charged off" loans, suggesting that applicants living in rented homes may be more susceptible to default due to economic fluctuations.
- A significant number of applicants (5,321) were **loan defaulters**. The company should improve risk assessment by enhancing credit checks and loan-to-value ratio requirements, as well as providing financial education to help borrowers manage their loans better.
- **Verified applicants** had more "Charged off" loans, with 3,236 defaults, indicating the need for stricter verification processes to reduce defaults.

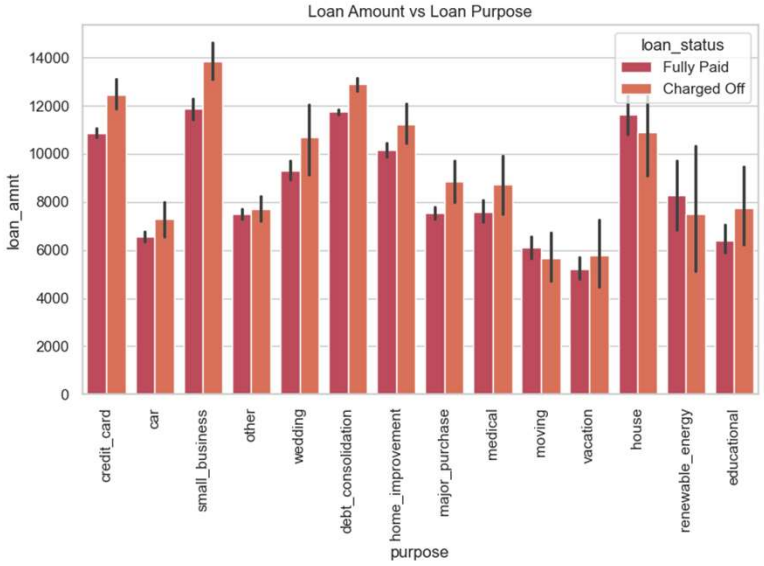
Univariate Analysis

Quantitative Fields

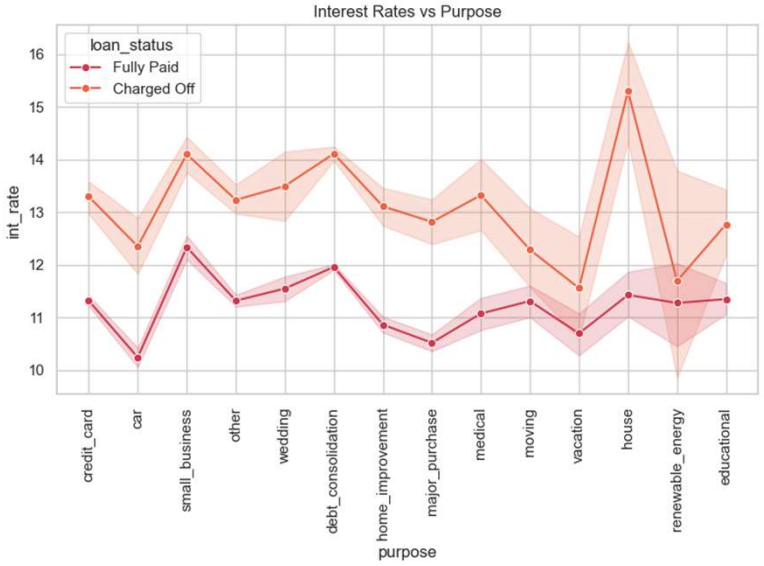
- **1,740 loan applicants** who charged off had **annual incomes below 40,000 USD**. More stringent income verification and repayment assessments are needed for applicants in this income range.
- **1,600 applicants** who charged off received loans between **5,000 USD and 15,000 USD**. These applicants should be closely evaluated for their creditworthiness and ability to repay.
- **2,026 applicants** who charged off fell into the **13% - 17% interest rate bucket**. Offering lower interest rates may reduce default risks.
- **1,174 applicants** who charged off had a **debt-to-income ratio above 20%**, indicating the need for stricter DTI ratio requirements.
- **1,544 applicants** who charged off had **75% of their debts still unpaid**, suggesting the need for stricter lending criteria for applicants with high levels of unpaid debt.
- **1,584 applicants** who charged off received **funded amounts between 5,000 USD and 15,000 USD**. The company should ensure that loan amounts align with the borrower's financial capacity, conducting thorough credit assessments accordingly.

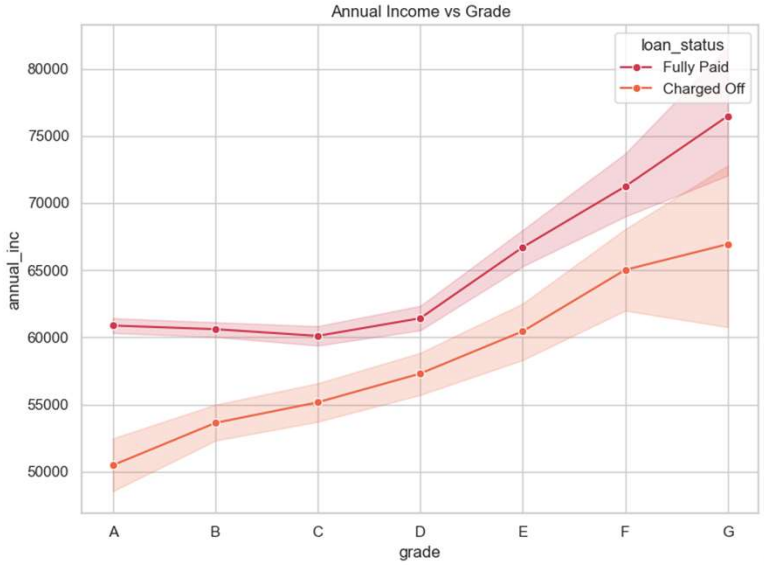
Bivariate Analysis



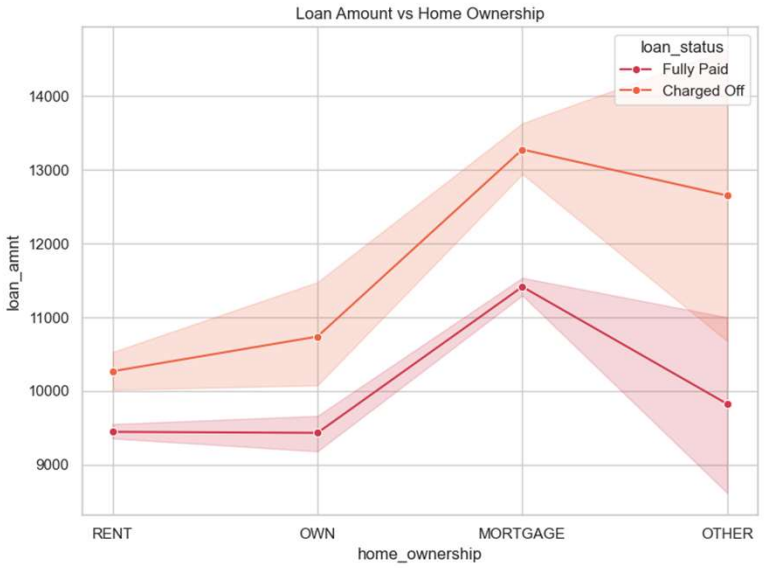


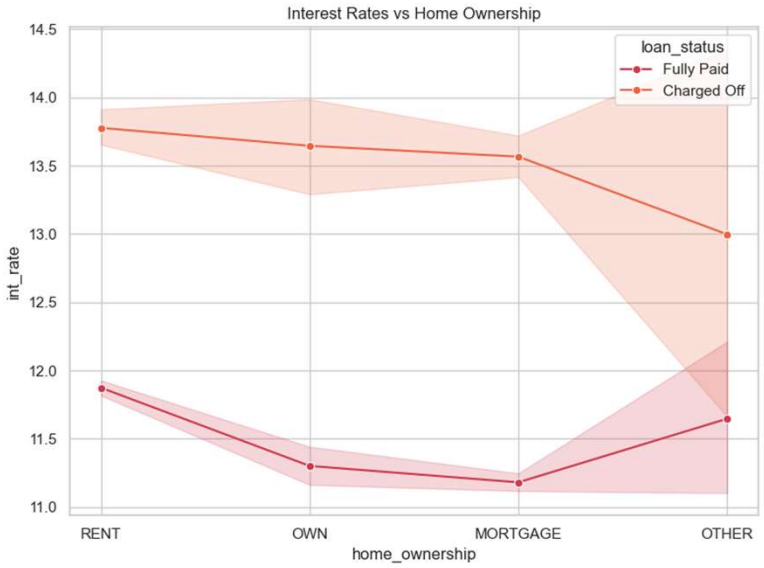
Bivariate Analysis



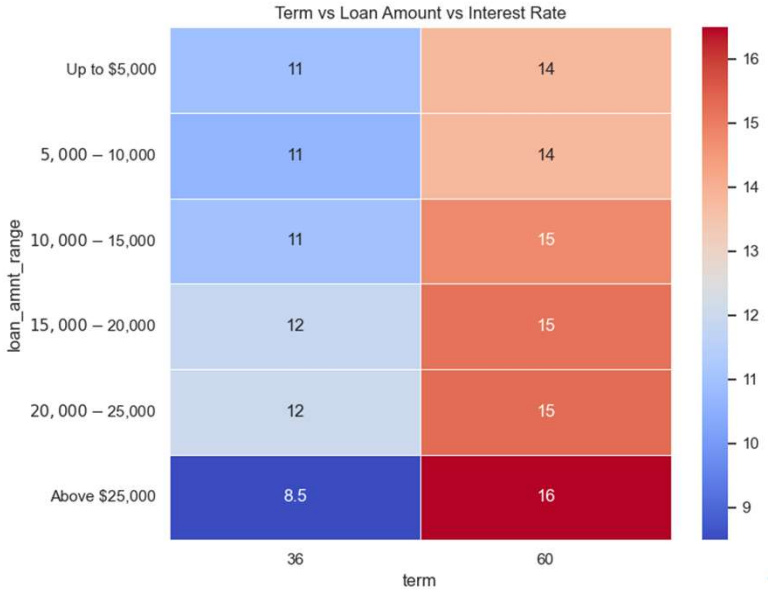


Bivariate Analysis

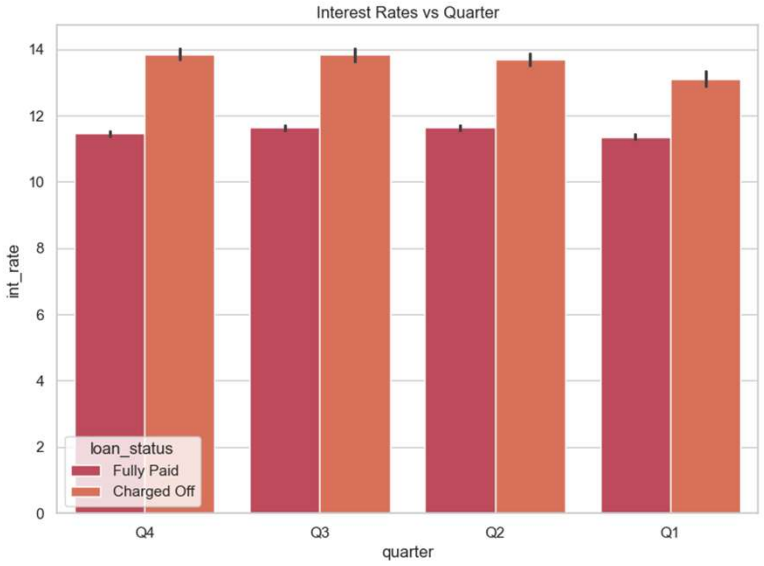


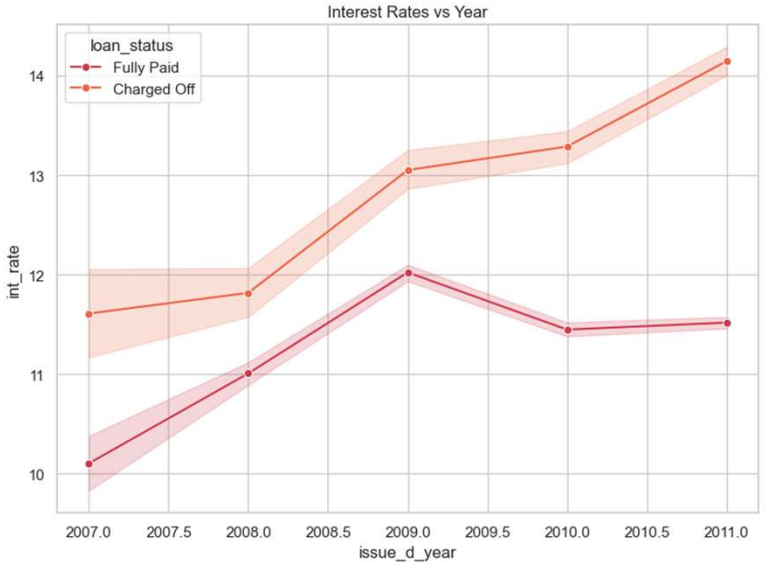


Bivariate Analysis

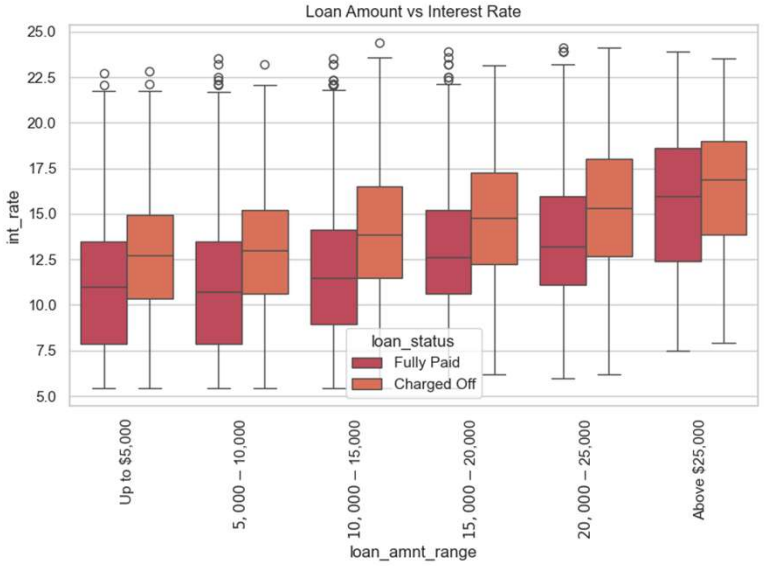


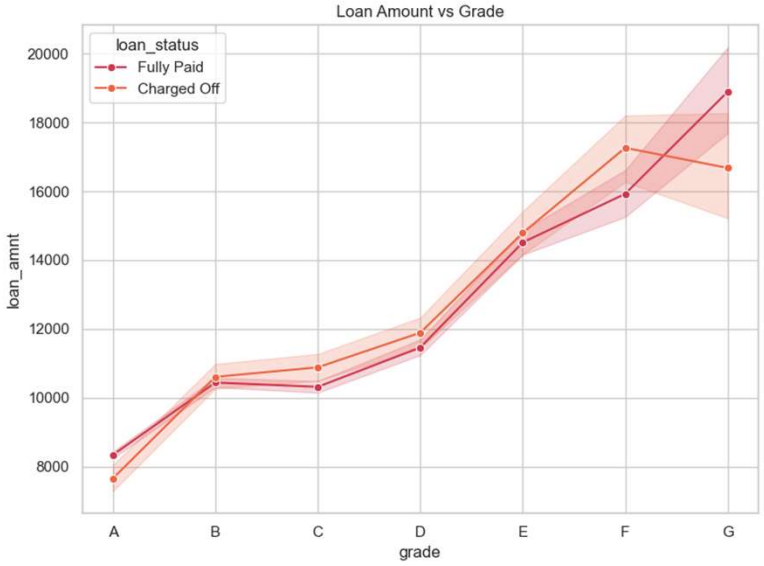
Bivariate Analysis



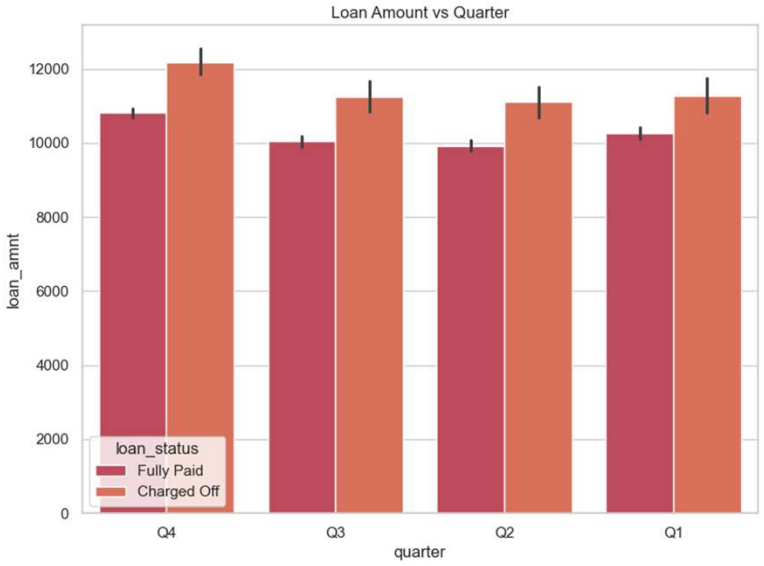


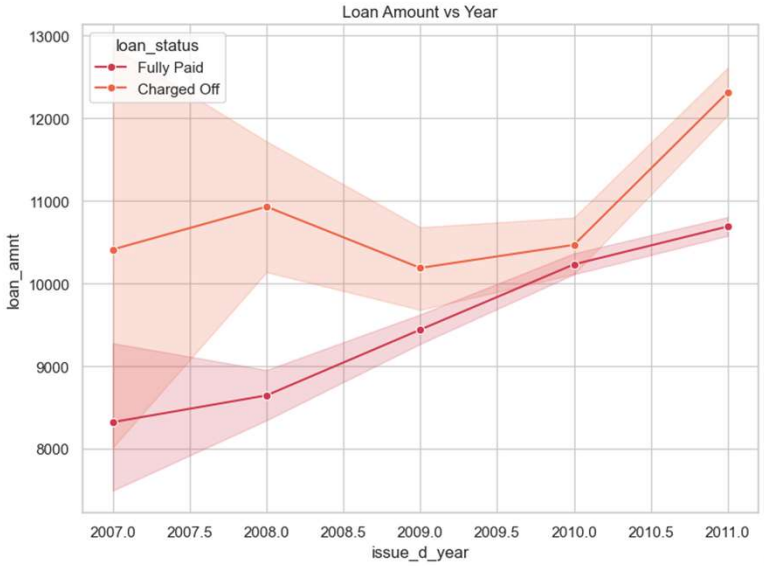
Bivariate Analysis



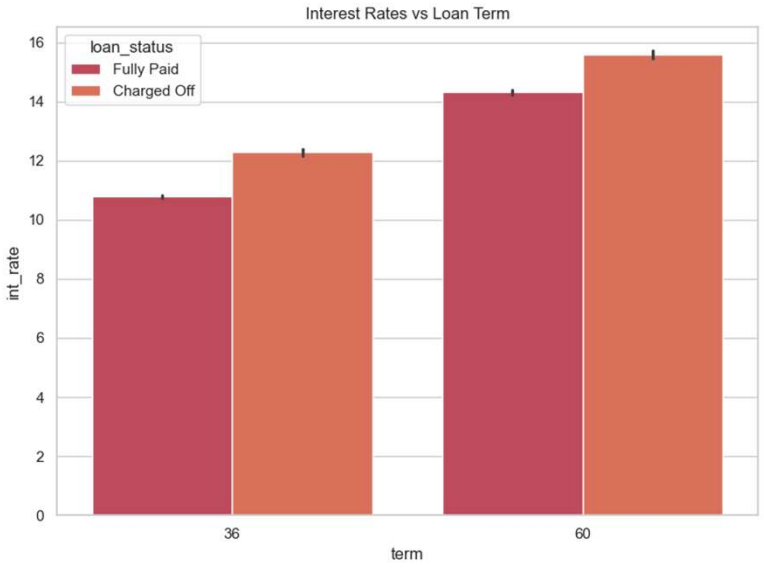


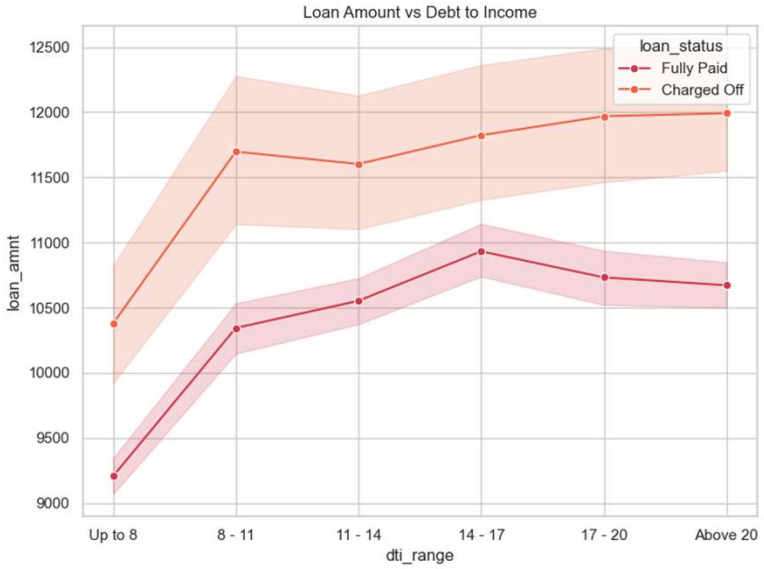
Bivariate Analysis





Bivariate Analysis





Bivariate Analysis

- Applicants with incomes above $60,000 are more likely to be "Charged off" for loan purposes such as small businesses, home improvements, and renewable energy. In contrast, lower-income applicants face higher risk for Education, Moving, and Major purchases.

- Applicants with higher incomes above $60,000 have a higher chance of being "Charged off" for Grades F and G, whereas lower-income applicants are at greater risk for Grades A, B, and C.

- Applicants with a high Debt to Income Ratio (above 15), annual income less than $60,000, and a loan amount over $10,000 are more likely to be "Charged Off". Approving larger loans for applicants with low income and a high DTI ratio increases the risk. Keeping an income to loan ratio of 1:8 could reduce this risk.

- For lower loan amounts, interest rates above 12% increase the chances of being charged off, and for higher loan amounts, interest rates above 15% increase the risk. The company should maintain lower rates for loans below $18,000.

- For higher loan amounts, the chance of being "Charged off" is higher for Grades F and G. The company could lend larger amounts to higher-income applicants in these grades.

Bivariate Analysis

- Over the years, the risk of being charged off for higher loan amounts has decreased, except for 2011. The U.S. Debt Ceiling Crisis in 2011 may have contributed to the spike in defaulters.
- Loan amounts over $12,000 have a high risk of being charged off regardless of the quarter, although other factors also contribute.
- Applicants with higher loan amounts are more likely to be "Charged off" for loan purposes like small businesses and credits. For lower-income applicants, the risk is higher for Education, Vacation, and Medical loans.
- Applicants with higher loan amounts and home ownership as a mortgage face a higher risk of being charged off, whereas renters are at greater risk with lower loan amounts.
- Applicants with a debt-to-income ratio above 17 and loan amounts over $12,000 have a higher chance of being charged off.

Bivariate Analysis

- For long-term loans, interest rates above 14% increase the risk of being charged off, while for short-term loans, interest rates above 11% have a similar effect.
- Short-term loans with high loan amounts at lower interest rates reduce the risk of being charged off.
- Higher-income applicants applying for high loan amounts with a debt-to-income ratio around 15 reduce their risk of being charged off if the loan is granted at lower interest rates.

THANK YOU