₹ Lending Club Case Study

Case study group:

Priyanka Gulati
Pankaj Kumar Agrawal

Contents

- Objective
- Data Understanding
- Data Cleaning and Pre-Processing
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Recommendations

Objective

The Objective of this case study is to help a consumer finance company understand the consumer and loan attributes that influence the tendency of default by using Exploratory Data Analysis (EDA) techniques.

The Lending company wants to analyse the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Learnings from the Case Study

- Identification of risky loan applicants will help the lending company in risk assessment and portfolio management thereby cutting down the amount of credit loss.
- EDA is a powerful tool to get a basic understanding of the data trends.
- Understanding of Risk analytics domain to identify logical variables that are indicators of default.

Data Understanding

Lending Company's historical data is provided with customer and loan attributes. The data consists of 39717 rows and 111 columns. This is a rich dataset which is used to analyse the driving factors of a loan default.

Dataset Attributes

Primary Attribute-Loan Status

The primary field on interest is the loan_status which indicates where it's a defaulted loan. It has 3 distinct values:

- **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
- **Current**: Applicant is in the process of paying the installments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
- **Charged-off**: Applicant has not paid the installments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

For the purpose of case study, Current loan status is removed from the analysis and this is an inconclusive category and hence not useable in identify default risk drivers.

Data Understanding

Some of the key risk drivers that represent customer & loan attributes that could drive default risk based on our domain understanding are summarized below:

Customer Attributes

- Annual Income (annual_inc): Self-reported annual income provided by the borrower.
- Home Ownership (home_ownership): The home ownership status provided by the borrower during registration. Values are: RENT, OWN, MORTGAGE, OTHER.
- Employment Length (emp_length): Employment length in years.
- State (addr_state): The state provided by the borrower in the loan application

Loan Characteristics

- Loan Amount(loan_amnt): The listed amount of the loan applied for by the borrower.
- Grade(grade): LC assigned loan grade
- Term(term): The number of payments on the loan.
- Purpose of Loan(purpose): Purpose of loan provided by the borrower on loan request
- Verification Status(verification_status): Indicates if income was verified by LC, not verified, or if the income source was verified
- Interest Rate(int_rate): Interest Rate on the loan
- Installments(installment): Monthly payment owed by the borrower if the loan originates
- Public Records of Bankruptcy(pub_rec_bankruptcies): Number of public record bankruptcies

Some of the columns have been excluded from the analysis and the reasons are summarized below:

- Un-useful Post loan features (like collection_recovery_fee etc.)
- Payment Amounts which are more appropriate for Loss Estimation rather than serving as default risk drivers (like total_pymnt, total_rec_late_fee etc.)
- User IDs (like id, member_id)
- Text Columns (like desc, emp_title, title etc.)
- Other unituitive variables (like url, zip_code etc.)

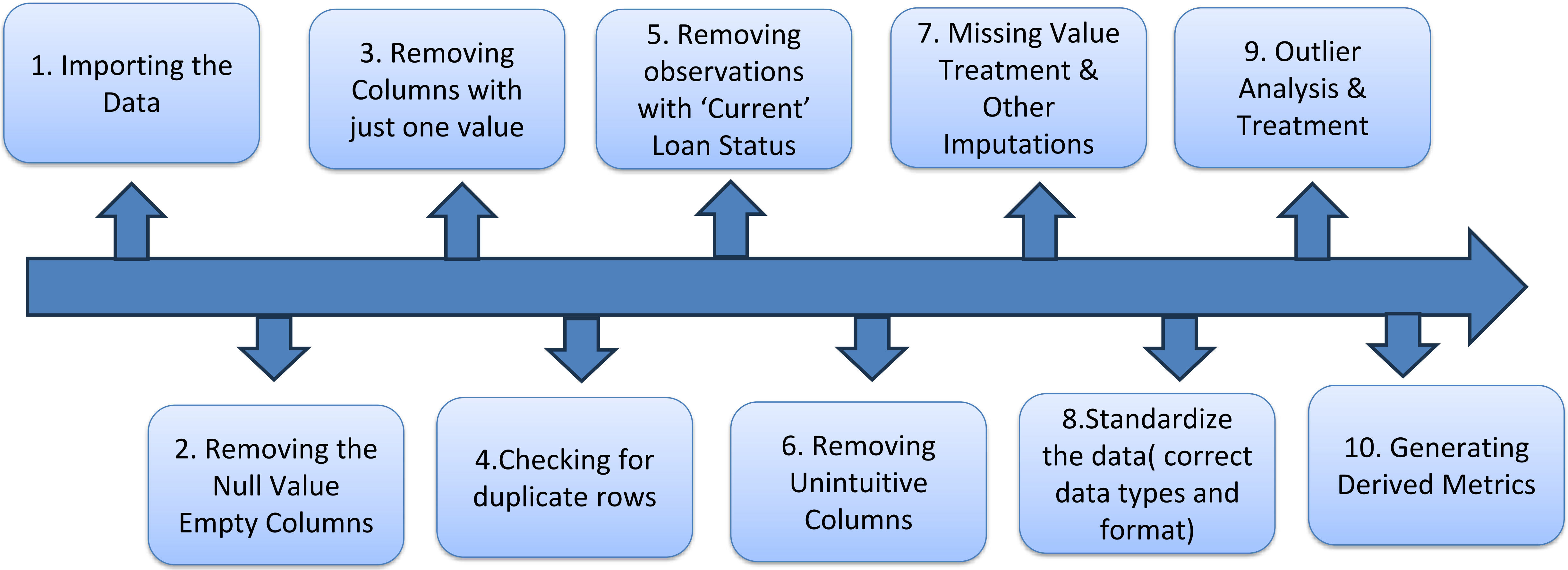
Data Cleaning & Pre-Processing

Data Cleaning & Pre-processing steps have been summarized the diagram below.
Key highlights from each step are as follows :

1. Importing the Data- Importing the csv file and conducting basic checks like shape, descriptions, column types , data types.
2. Removing the Null Value Empty Columns- A lot of columns had only null values. Hence 49% of the columns had to be removed.

3. Removing Columns with just one value-Columns with just one unique value would not contribute inidentifying default risk drivers. Hence, 9 such columns are removed from the data

4. Checking for Duplicate Rows- Duplicate rows were checked but not identified.


1. Importing the Data
2. Removing the Null Value Empty Columns
3. Removing Columns with just one value
4.Checking for duplicate rows
5. Removing observations with 'Current' Loan Status
6. Removing Unintuitive Columns
7. Missing Value Treatment & Other Imputations
8.Standardize the data(correct data types and format)
9. Outlier Analysis & Treatment
10. Generating Derived Metrics

Data Cleaning & Pre-Processing

5. Removing observations with ‘Current’ Loan Status - As these observations cannot provide any insights in identifying the default risk drivers they are removed from the data.
6. Removing Unintuitive Columns- Excluding the unintuitive columns containing un-useful post loan features, payment amounts which are more appropriate for loss estimation rather than serving as default risk drivers, user IDs, text columns, other un-intuitive variables.
7. Missing Value Treatment & Other Imputations- Replacing pub_rec_bankruptcies with median value, emp_length with mode value and revol_util with mean value.Replacing NONE with OTHER in home_ownership.
8. Standardize the data(correct data types and format)- converting term to float post-removal of ‘months’ string, int_rate to float post-removal of % sign, revol_util to float post removal of % sign,issue_d to datetime
9. Outlier Analysis & Treatment- Outliers were detected for several quantitative variables but removed only based on Annual Income as it had extreme outliers(above Q3+1.5IQR).
10. Generating Derived Metrics – Generated year, month and quarter variable from issue date and binned quantitative variables

Univariate Analysis

Univariate Analysis is a statistical method to analyze each variable in the dataset. **Univariate analysis is used to study overall distribution and default distribution and using Segmented Univariate analysis to study default rate trends(in the case study).**

The categorical and quantitative variables used for univariate analysis are identified as below:

Categorical Variables

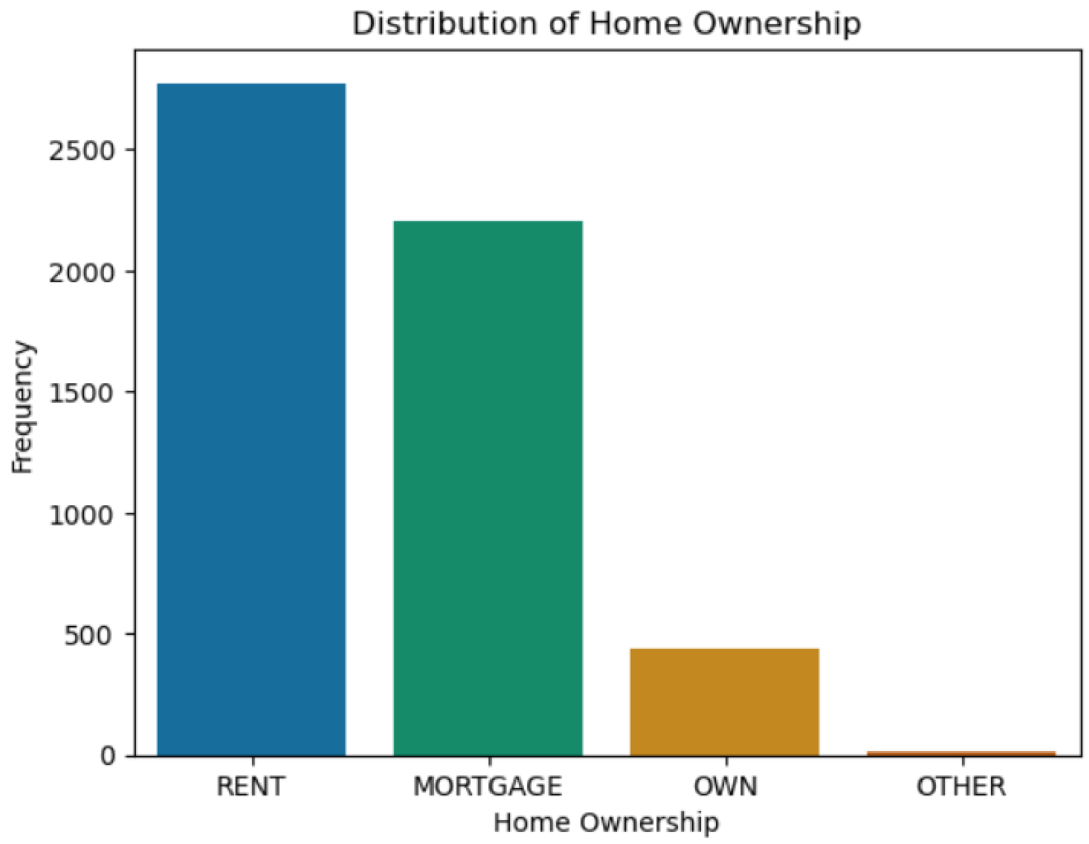
Ordered	Unordered
• Term (term) • Grade (grade) • Sub Grade (sub_grade) • Employment Length (emp_length) • Loan Issue Year (issue_y) • Loan Issue Month (issue_m) • Loan Issue Quarter (issue_q)	• Loan Status (loan_status) • Home Ownership (home_ownership) • Verification Status (verification_status) • Purpose (purpose) • State (addr_state)

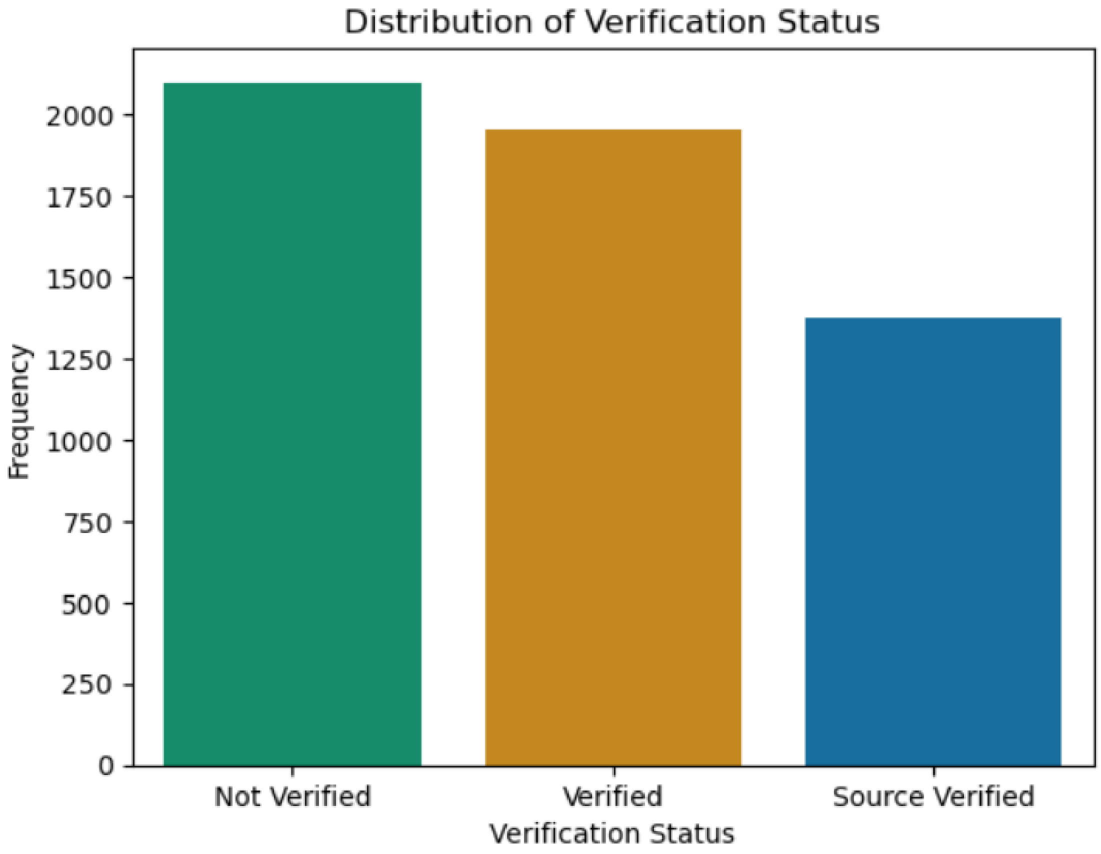
Quantitative Variables

- Loan Amount (loan_amount_binned)
- Funded Amount (funded_amnt_binned)
- Funded Amount commited by investors (Funded_amnt_inv_binned)
- Interest Rate (int_rate_binned)
- Installment (installment_binned)
- Annual Income (annual_inc_binned)
- Debt to Income Ratio (dti_binned)
- Credit Revolving Balance (revol_bal_binned)
- Revolving Line Utilisation Rate (revol_util_binned)
- Total Credit Lines (total_acc_binned)
- Open Credit Lines (open_acc_binned)
- Public Record Bankruptcies (pub_rec_bankruptcies

Univariate Analysis – Unordered Categorical

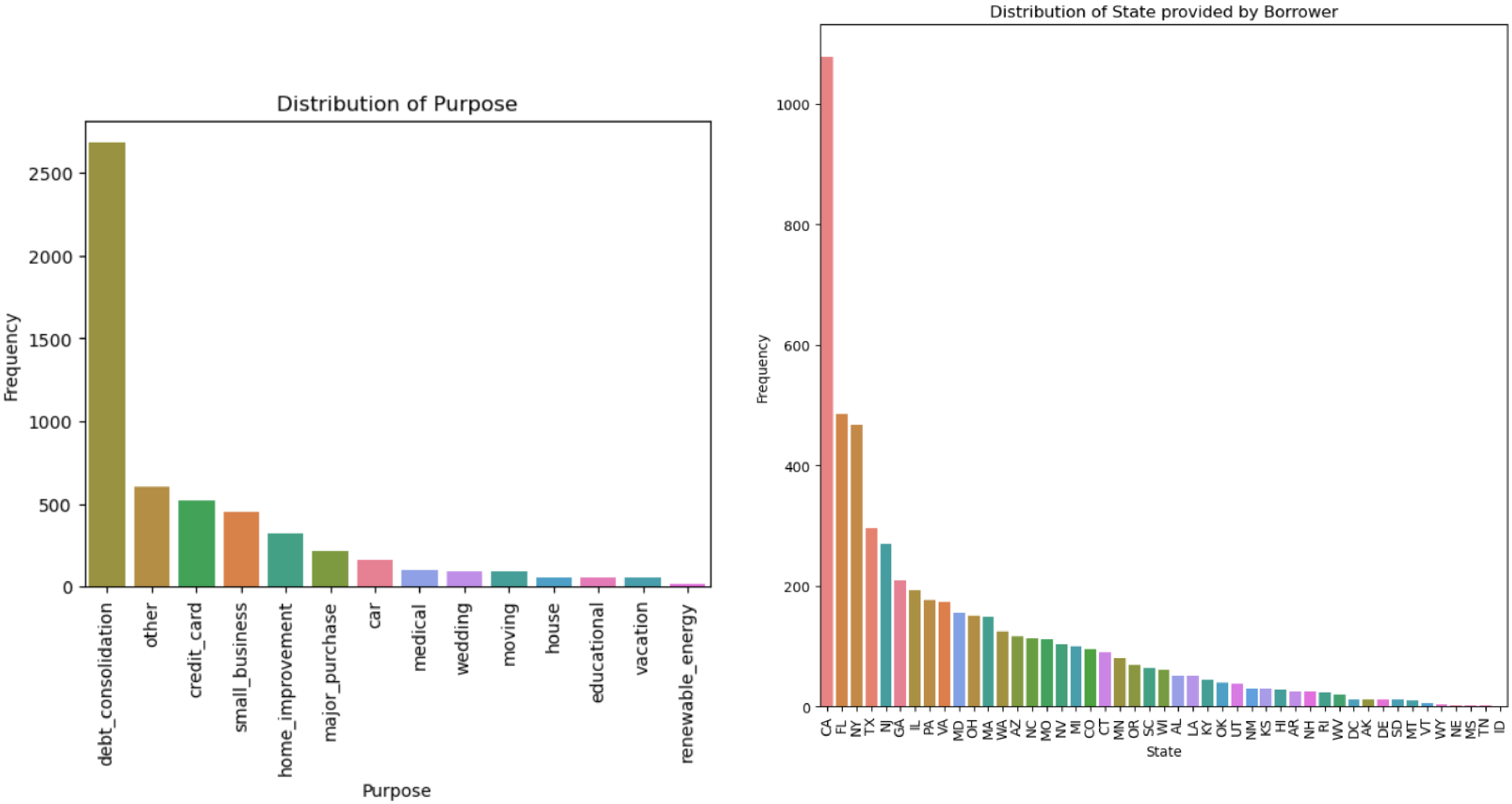
Analysing Distribution of variables on entire clean dataset (Defaults+Non-Defaults) as well as defaulted population.





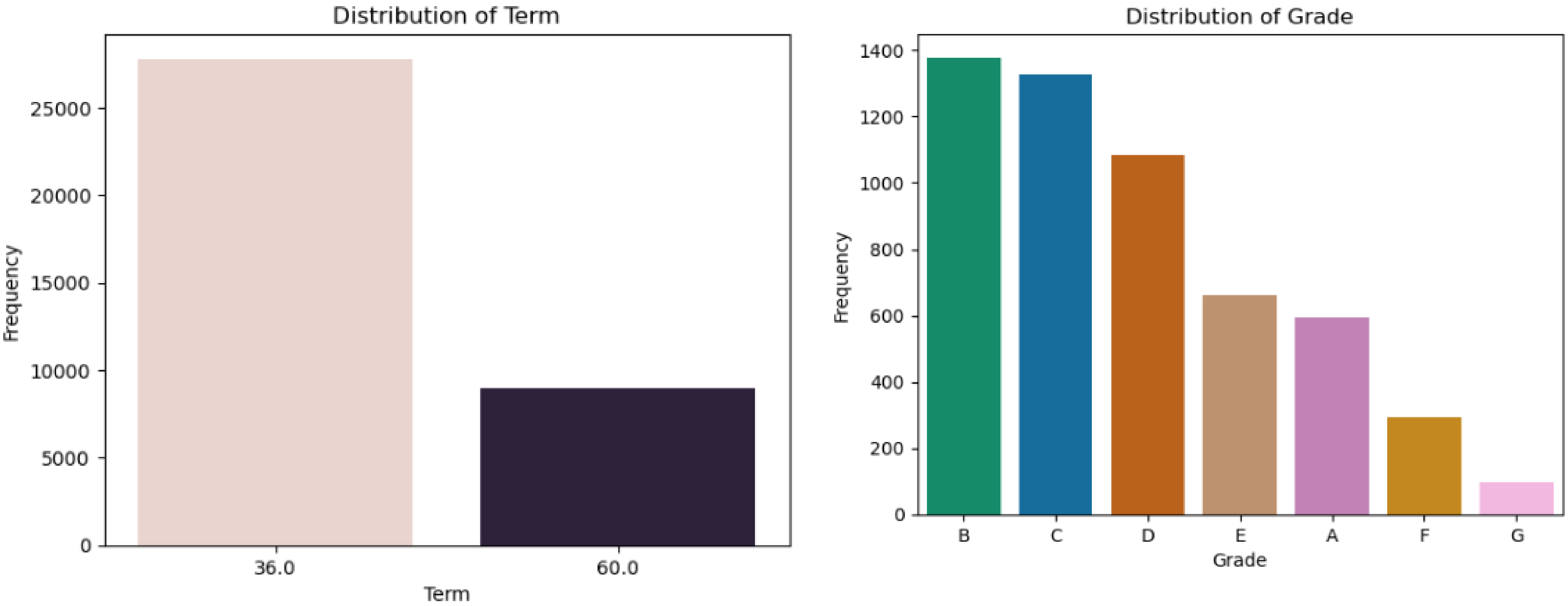
Home Ownership & Verification Status

Univariate Analysis – Unordered Categorical



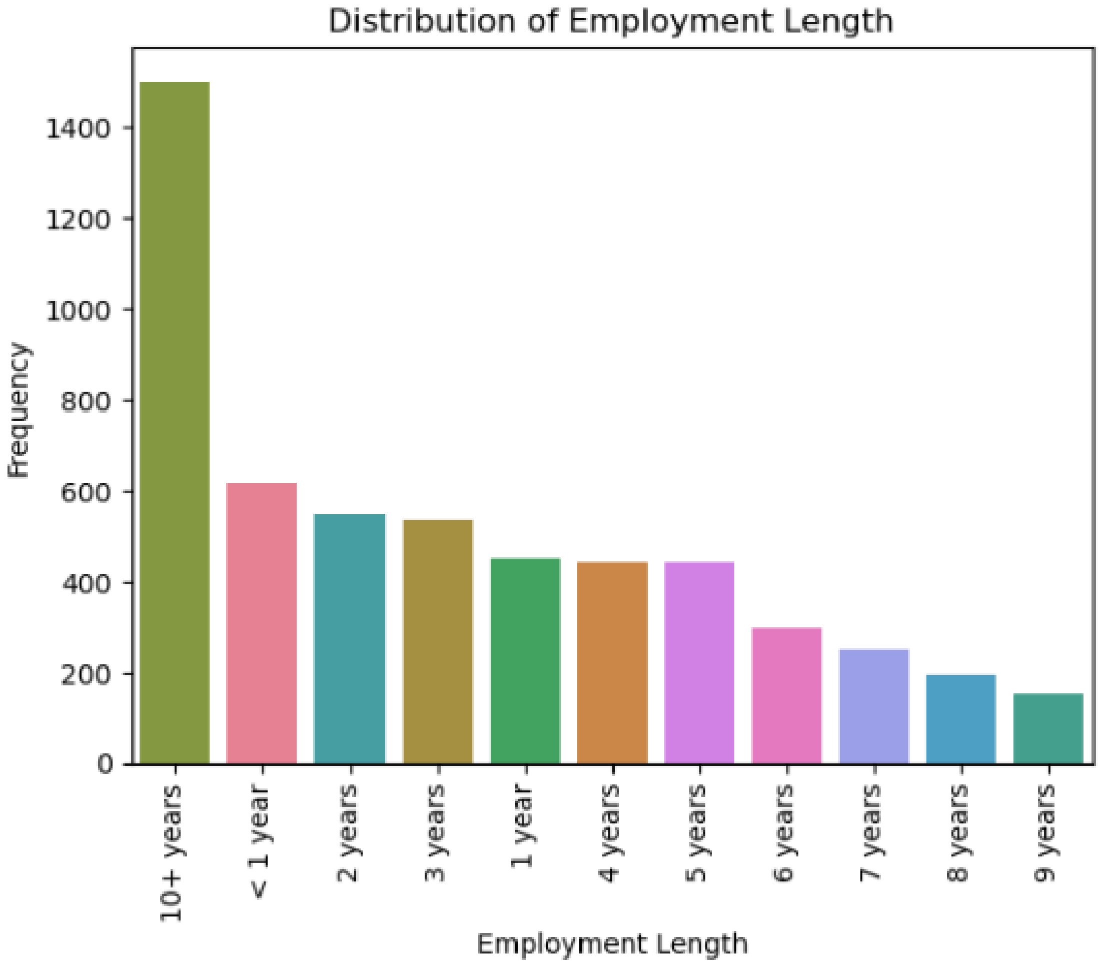
Purpose and State

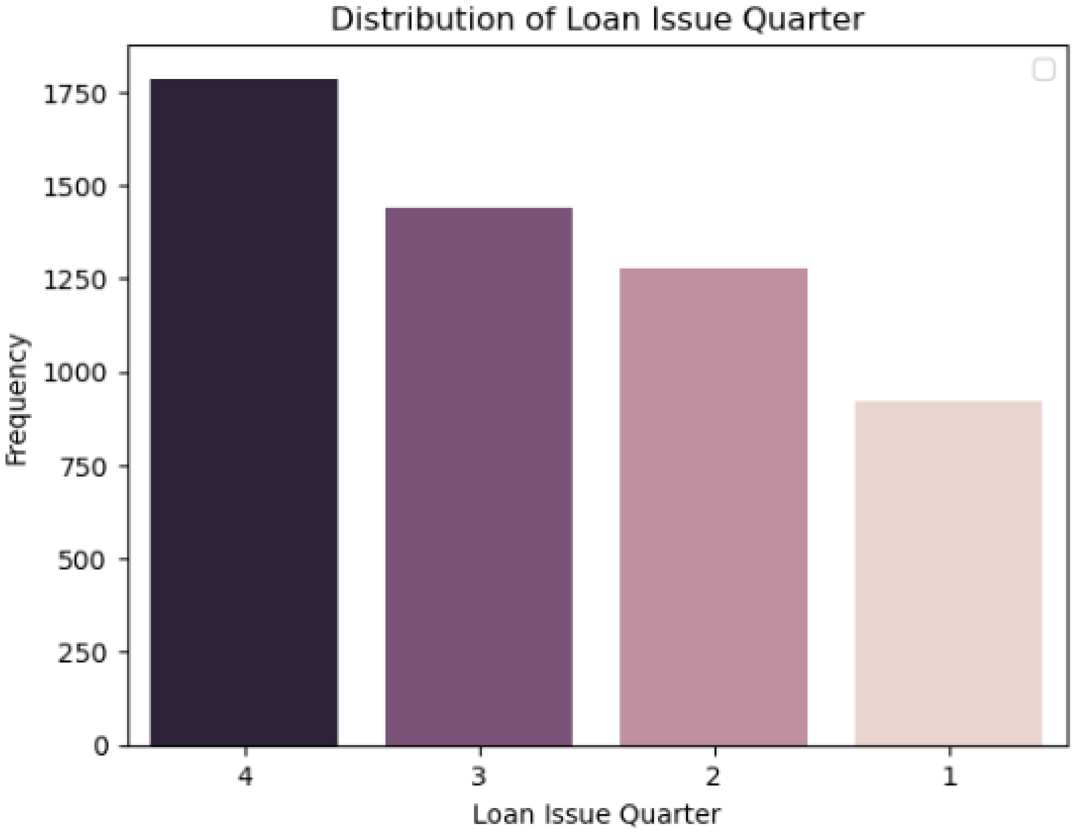
Univariate Analysis – Ordered Categorical



Term & Grade

Univariate Analysis – Ordered Categorical





Employment Length & Loan Issue Quarter

Univariate Analysis – Categorical

Observations and Inferences

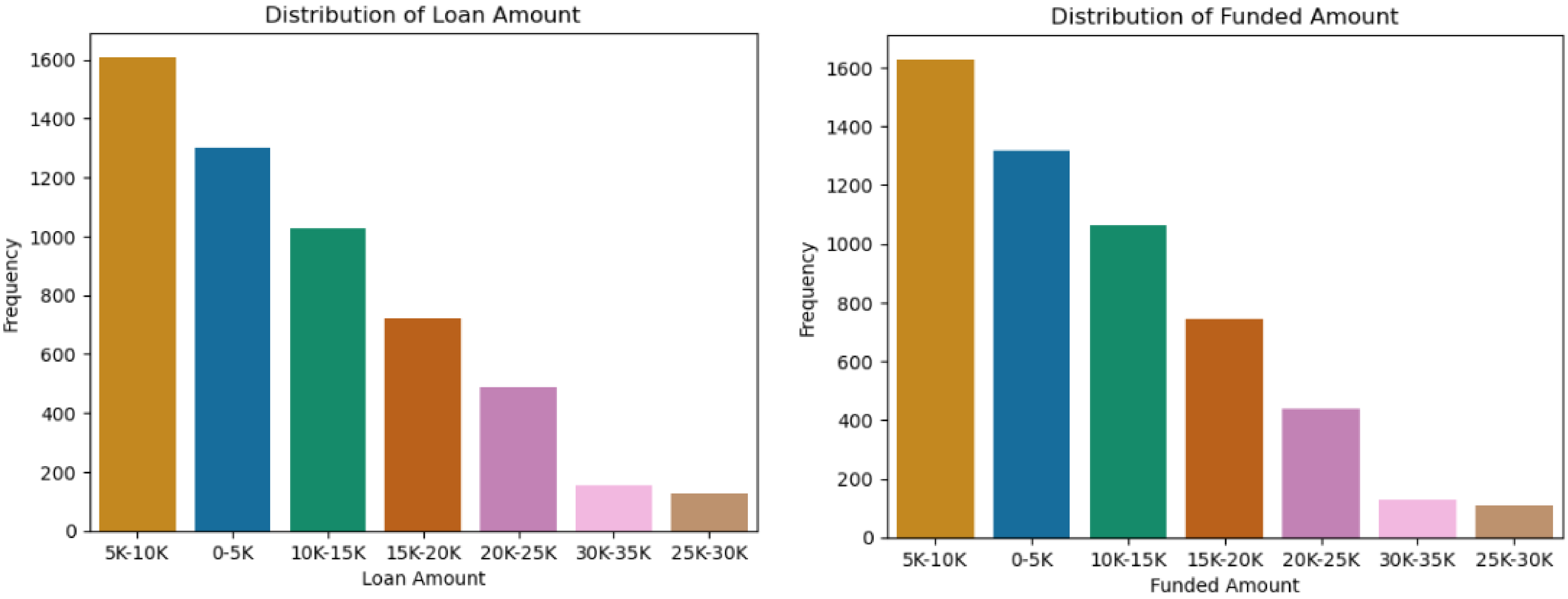
Unordered Categorical

- Rented Households had the highest number of observations and charge-off cases. The lending company should evaluate the loan applicants carefully before lending to rented households as they are more susceptible to default.
- Loan Applicants whose income source are not verified have higher number of observations and charge-offs when compared with verified, followed by verified from source. But the difference in not verified and verified count is not very significant.
- Loans taken for Debt consolidation purpose had the highest number of observations and charge offs. The lending company should be cautious when approving loans for debt consolidation purpose.
- California had the highest number of observations and charge-offs. The lending company can implement stricter credit policies in California.

Ordered Categorical

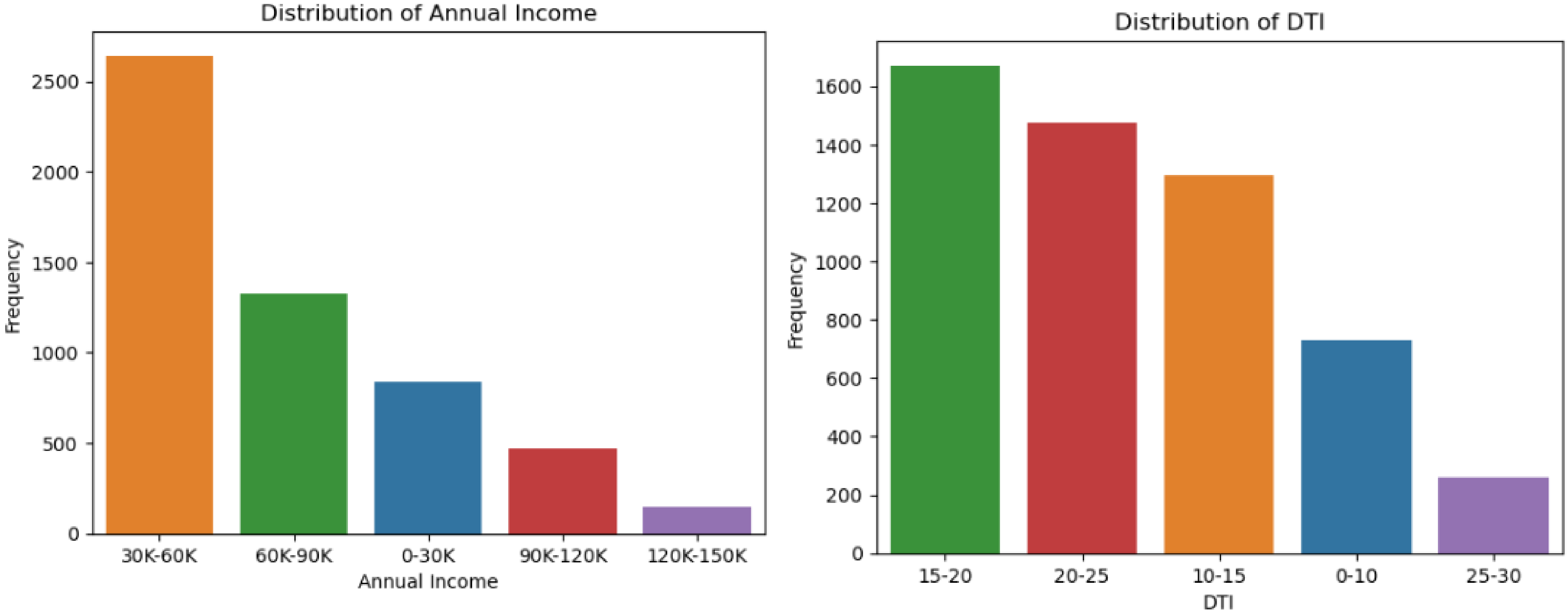
- Short term loans (term of 36 months) had higher observations and charge-offs. The lending company should evaluate the loans carefully for shorter terms.
- Loans with B grade had highest number of observations and charge-offs indicating these customers faced difficulty in repaying the loans.
- Loans with 10 + years of experience had higher number of observations and charge-offs. Although, the total number of employees with 10+ years of experience is significantly higher than in other categories.
- Loans issued in 2011 had the highest number of observations and charge-offs. One potential reason could be economic difficulties in the specific year.
- Loans issued in fourth quarter had the highest number of observations and charge-offs. One potential reason for this can be that loan underwriters are providing loans with less due diligence to complete their targets.

Univariate Analysis – Quantitative



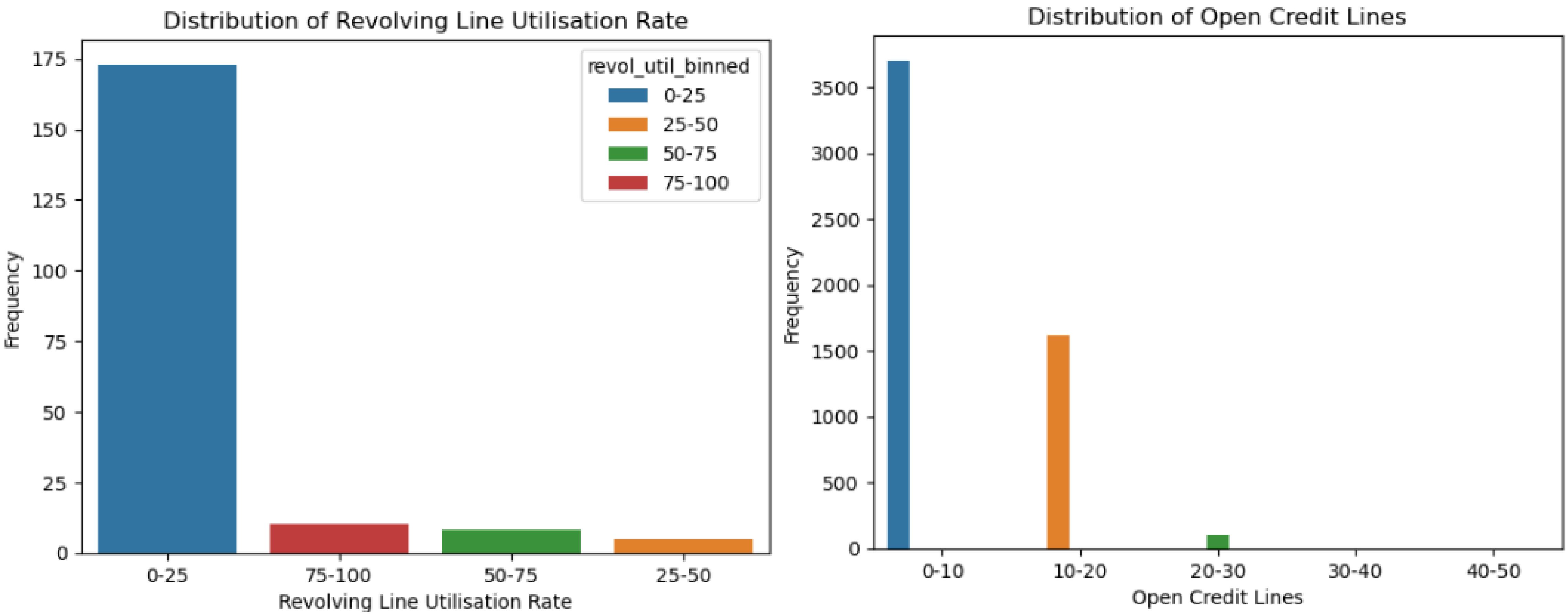
Loan Amount and Funded Amount

Univariate Analysis – Quantitative



Annual Income and DTI

Univariate Analysis – Quantitative



Revolving Line Utilisation Rate and Open Credit Lines

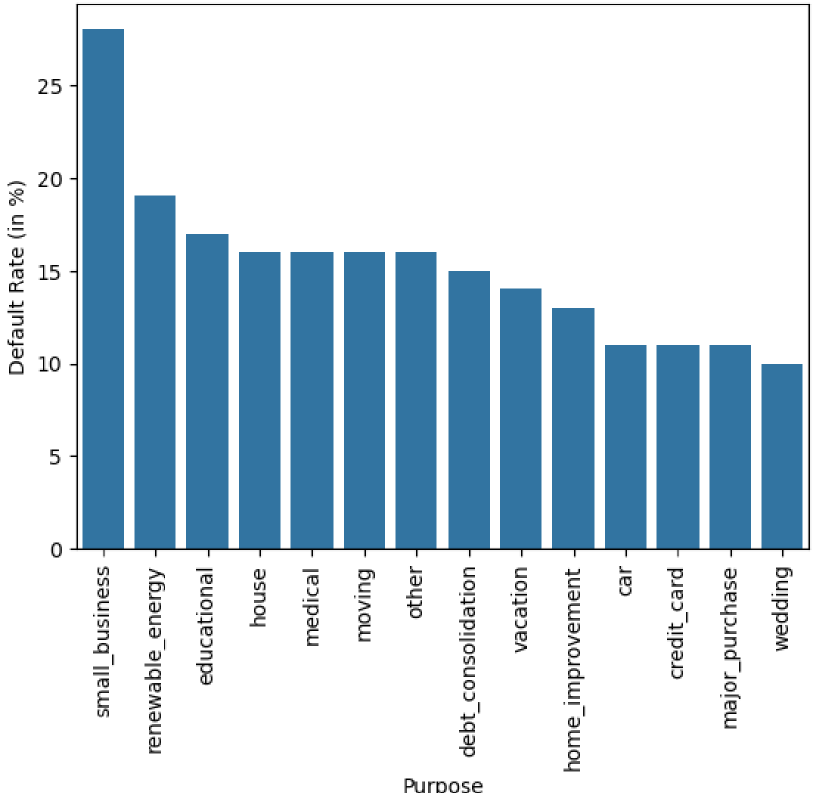
Univariate Analysis – Quantitative

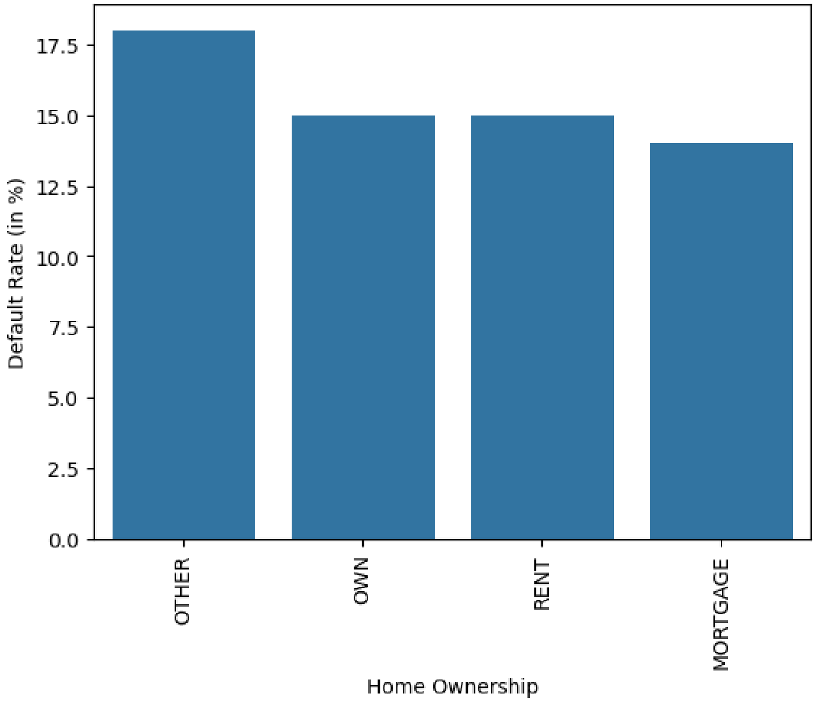
Observations and Inferences

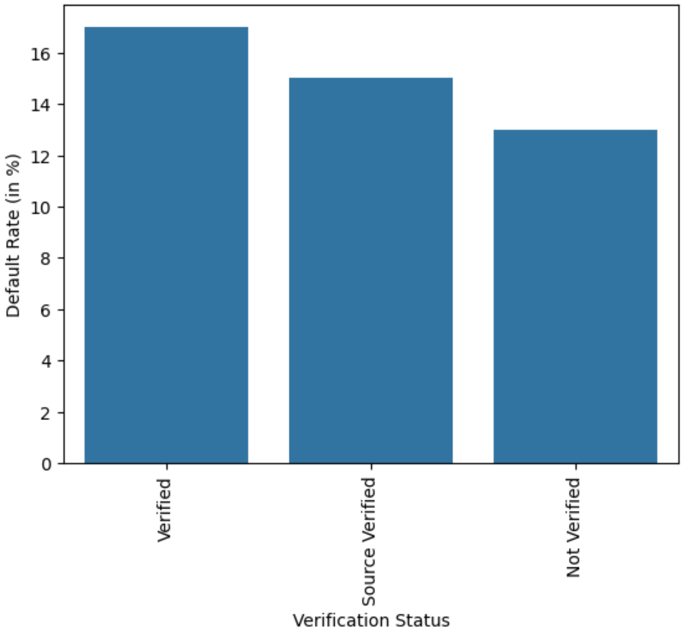
- Applied Loan Amount between 5K-10K has the highest number of observations and charge-offs. Lending company should be careful when sanctioning loans.
- Funded Loan Amount & Funded Loan Amount committed by investors between 5K-10K has the highest number of observations and charge-offs. Lending company should fund the loans only after diligent credit worthiness analysis.
- Applicants who are charged 10-15 % interest rate have the highest observations and charge-offs.
- Installments between 0-250 have the highest observations and charge-offs. Lending company should assess the risk of the applicants with similar installments.
- Applicants with income of 30K-60K have the highest observations and charge-offs. Lending company should thoroughly analyse the repaying capacity of low-middle income group.
- Applicants with Debt-to-Income ratio in 15-20 range have the highest observations and charge-offs. Lending company should be careful when lending to the highly indebted customers with very high debt/income ratio.
- Applicant with revolving credit balance in 0-25 K range and utilisation 0-25 range have high observations and charge-offs. Lending company should be cautious when lending to customers with less credit history or utilisation rates.
- Applicants with 0-10 open credit lines and 0-25 total credit lines have high observations and charge-offs. Lending company should be cautious when lending to customers with less credit history.
- Applicants with 0 public bankruptcy record have higher observations and charge-offs.

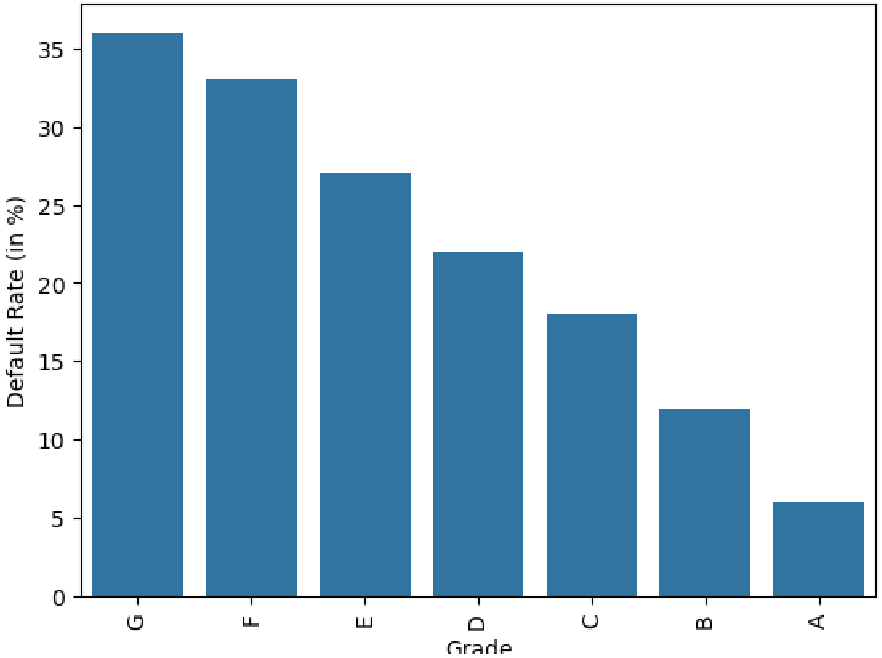
Segmented Univariate

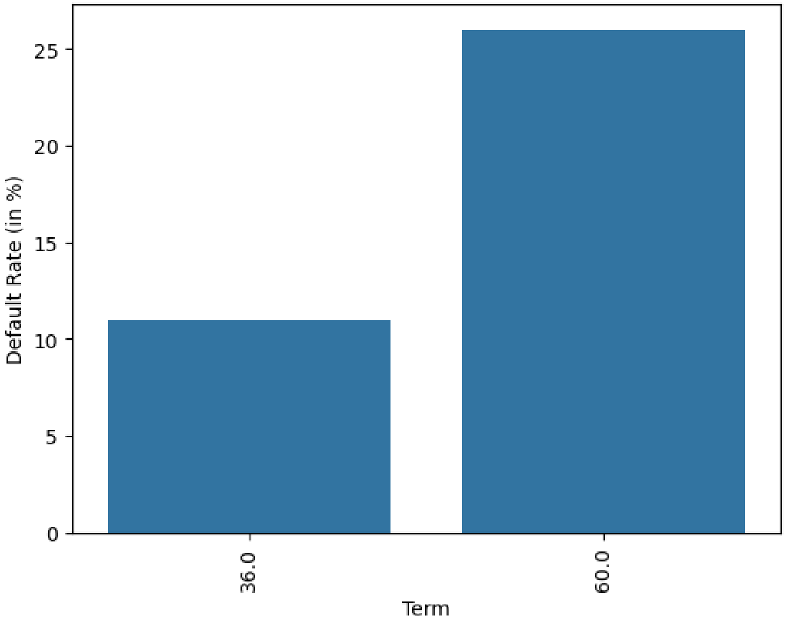
Analysing Default Rate Trends

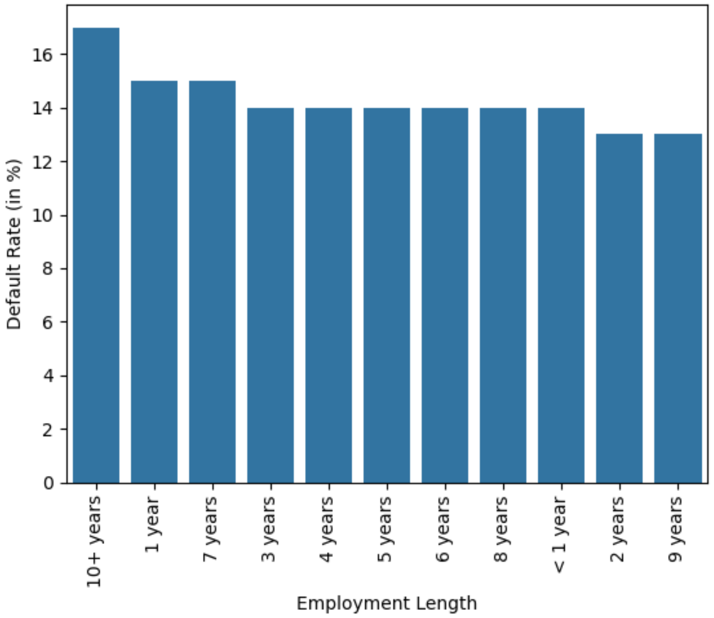






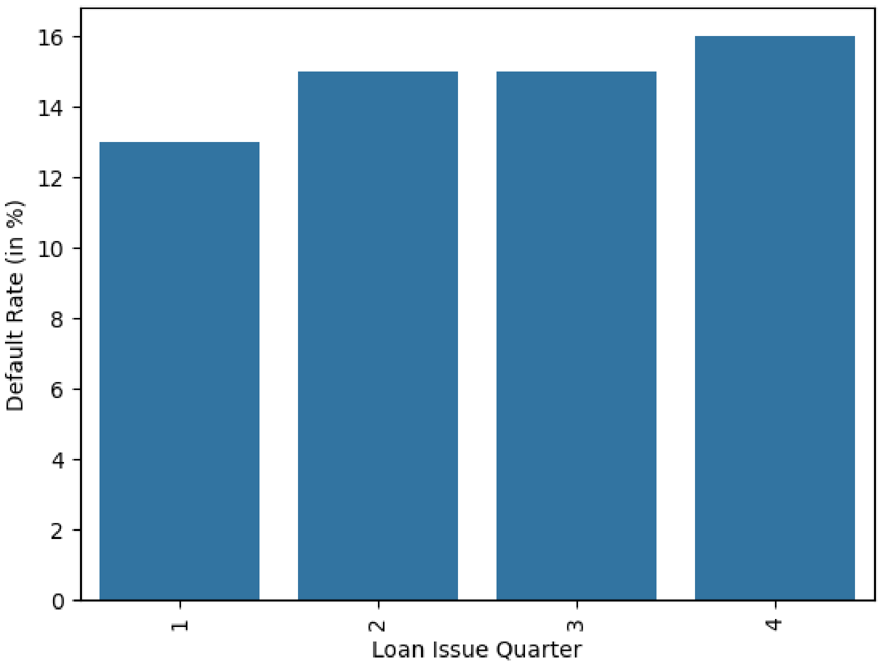


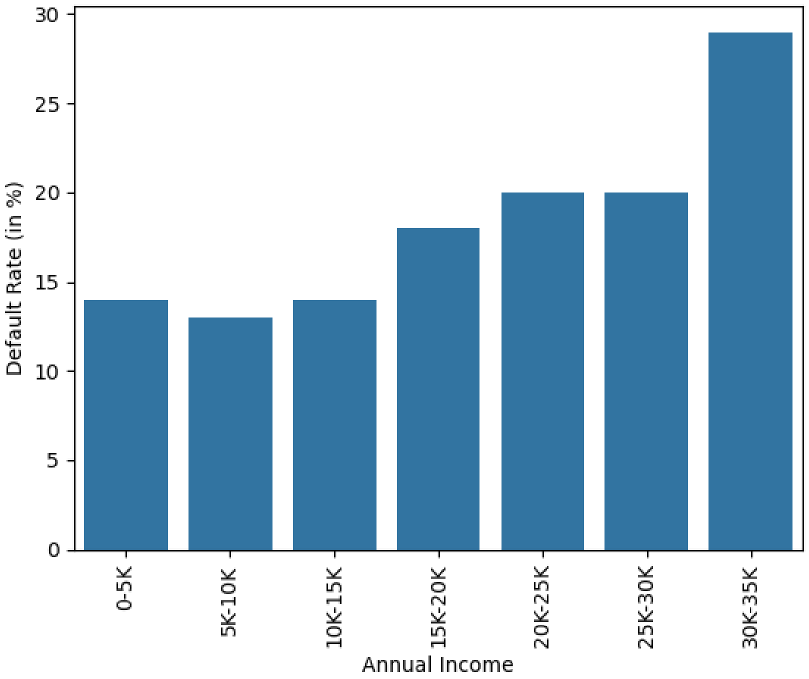


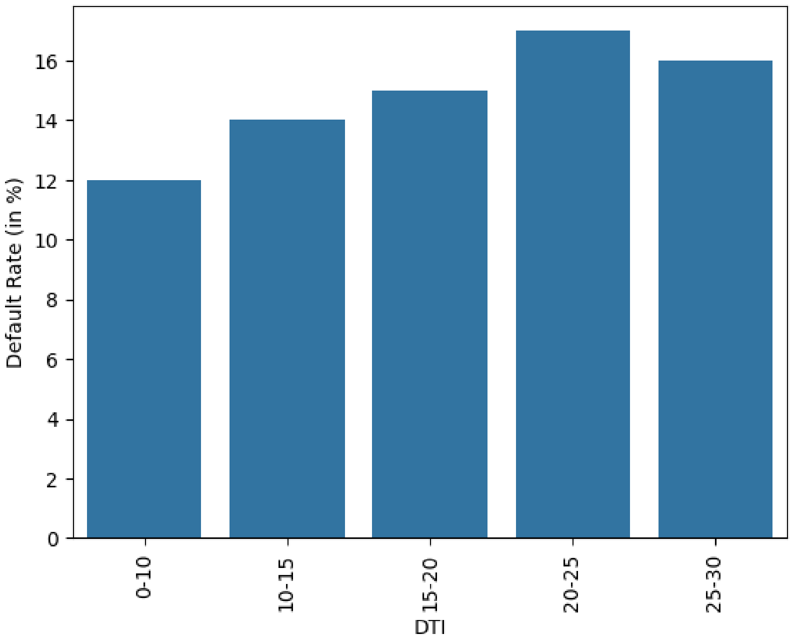


Segmented Univariate

Analysing Default Rate Trends







Observations and Inferences

•Loan applicants with Small Business have higher default rates.

•Other home ownership category has higher default rates.

•Verified Income Source have higher default rates.

•Nebraska has high default rates.

•Longer term (60 months) have higher default rates.

•Worse Grades (G) have high default rates.

•Applicants with longer employment (10 + years) have high default rates.

•Loans issued in Q4 have high default rates.

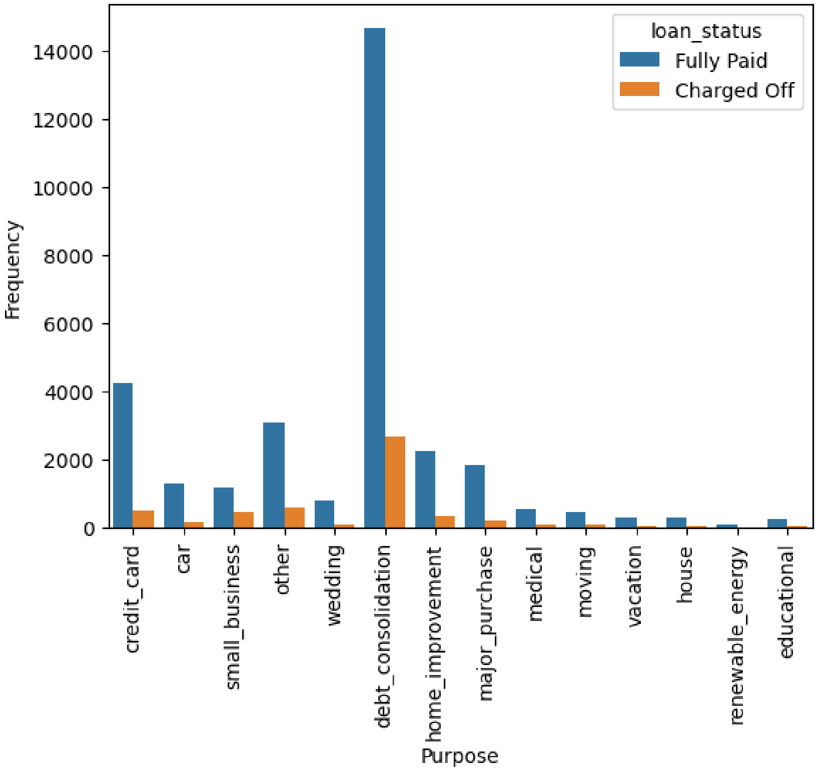
•Annual Income 0-30 K have high default rates.

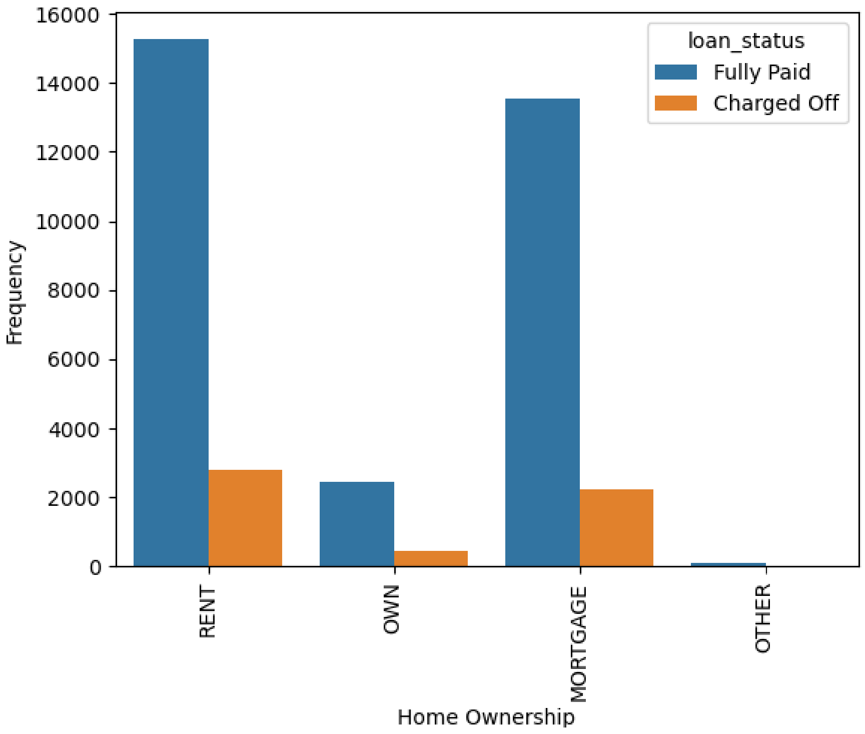
•High funded amount loans in the range 30-35K have high default rates.

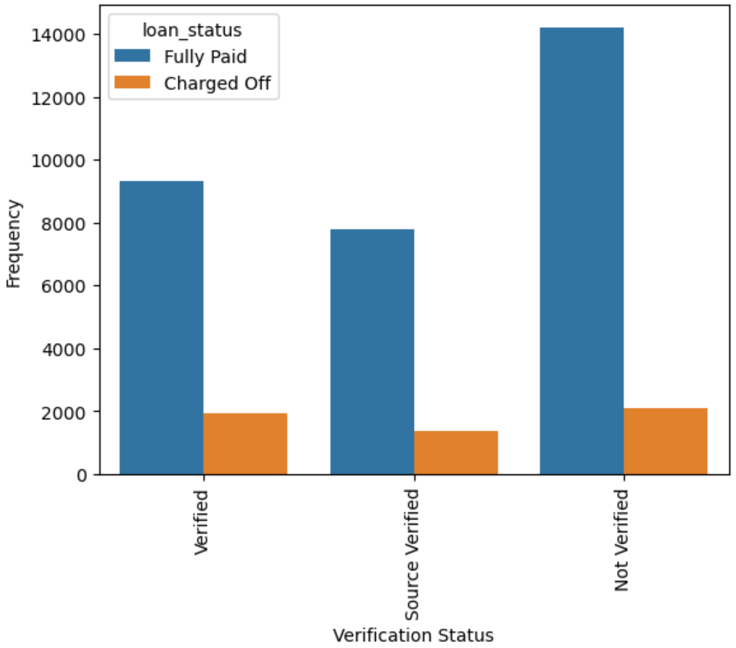
•High Debt to Income Ratio (20-25) have high default rates

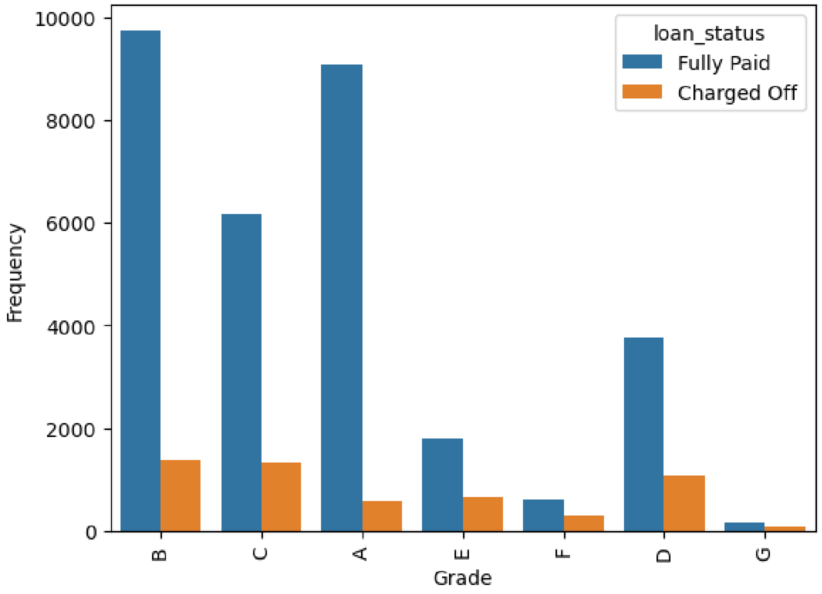
Bivariate and Multivariate Analysis

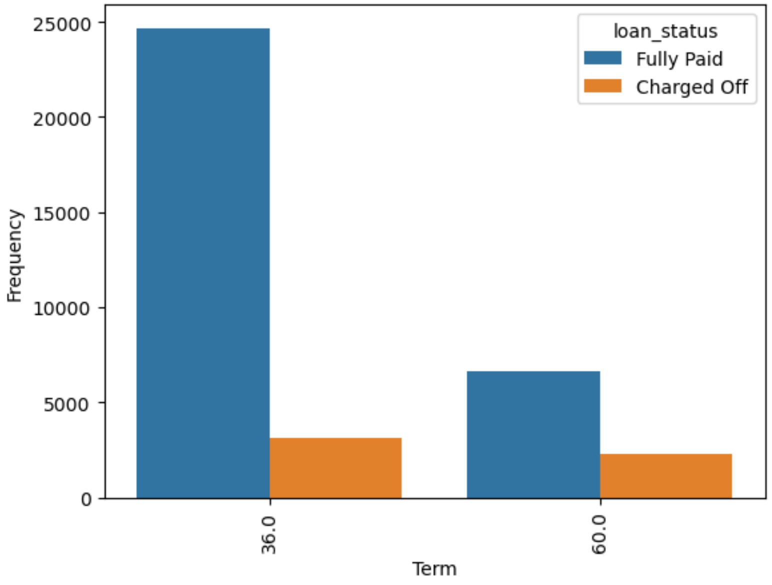
Analysing Loan Status across different Variables

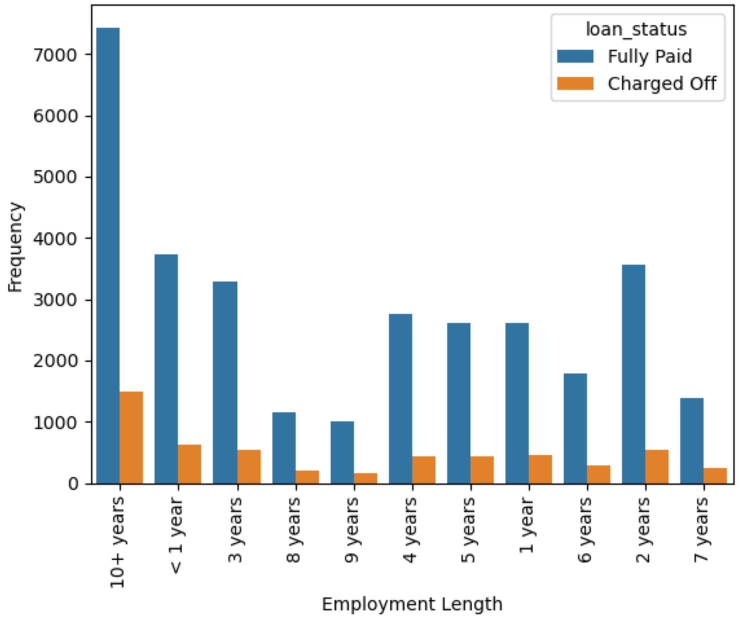






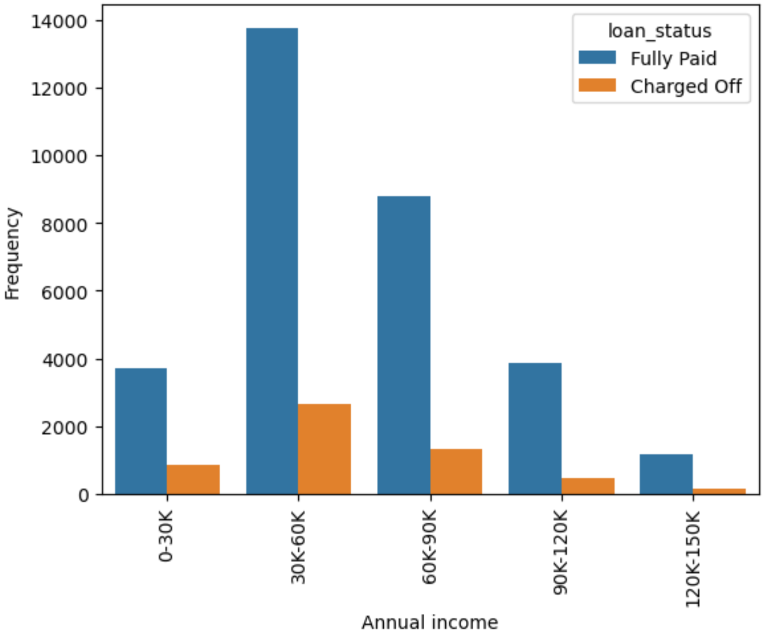


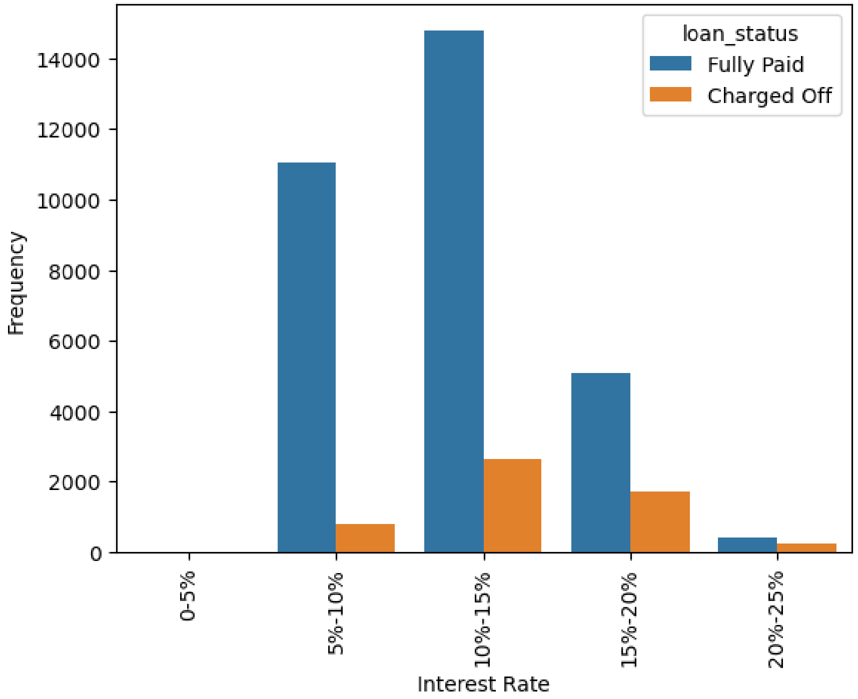


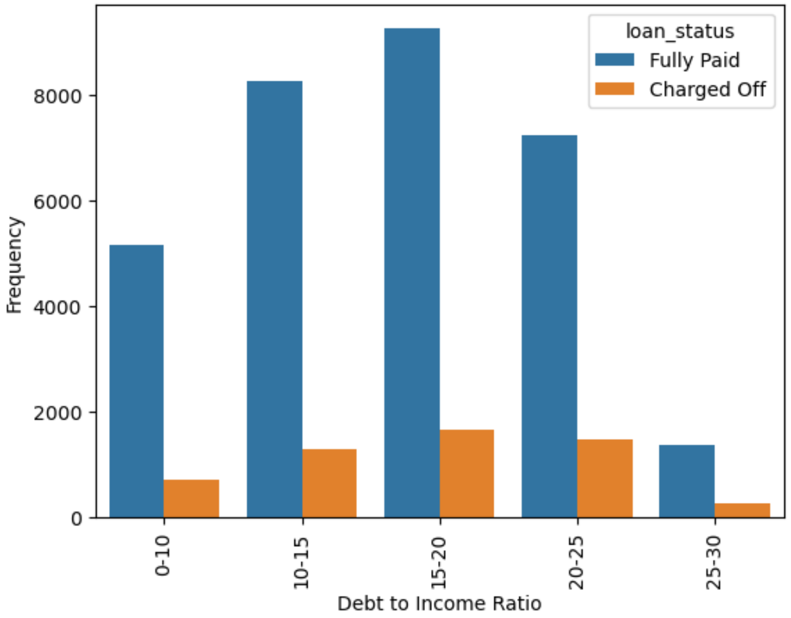


Bivariate and Multivariate Analysis

Analysing Annual Income across Loan Status and Other Variables







Observations and Inferences

•Debt consolidation category has the highest number of loans and defaults

•Rent and Mortgage have high number of loans and defaults

•Non-verified sources of Income have high number of loans and defaults

•California has high number of loans and defaults -Grade B have high number of loans and defaults

•Shorter loan term (36 months) have high number of loans and defaults

•Applicants with high work experience (10 + years) have high number of observations and defaults

•Applicants with 30-60 K income get more loans and default more

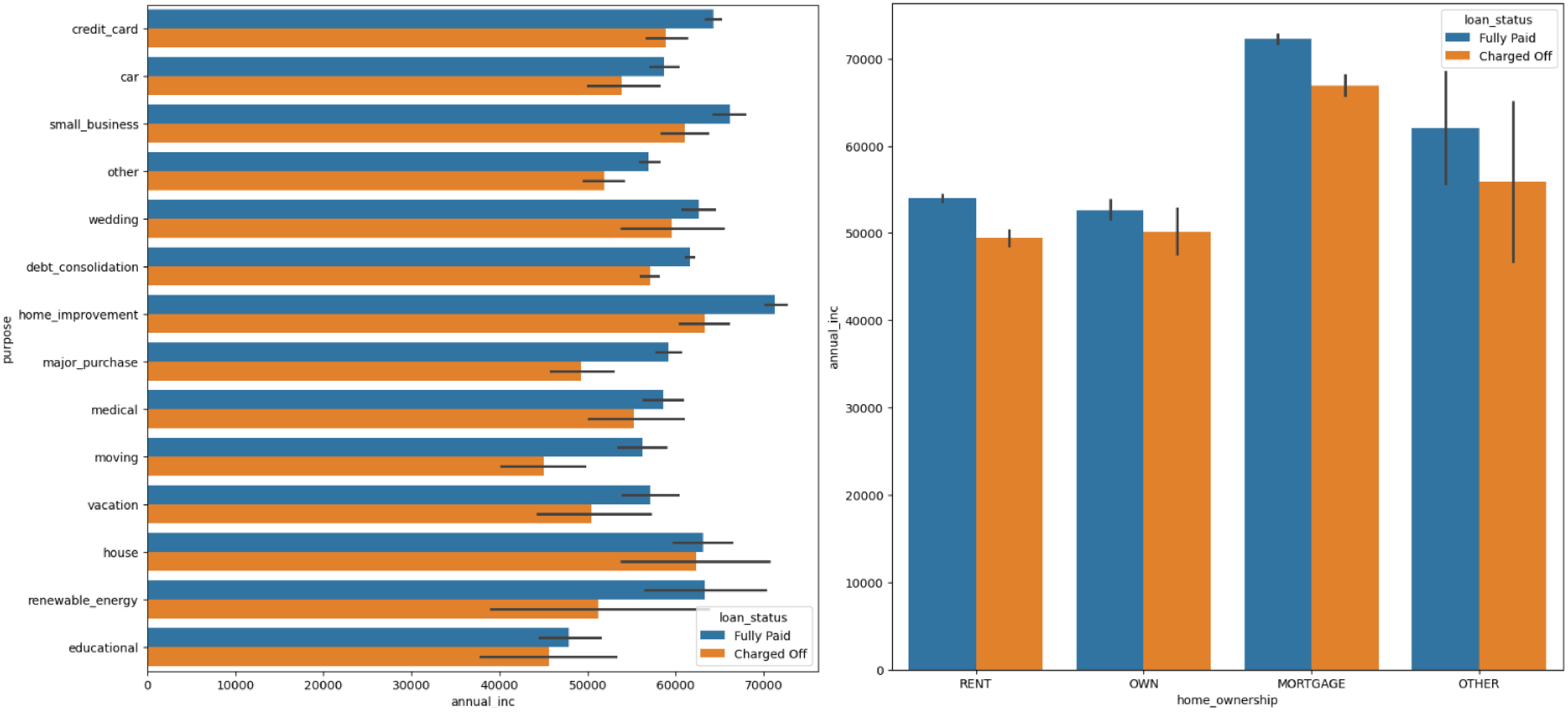
•Funded loans in 5K-10 K have more loans and defaults

•Loans with 5-10 % interest rate have more observations and defaults

•Loans wit 15-20 DTI have more loans and defaults

Bivariate and Multivariate Analysis

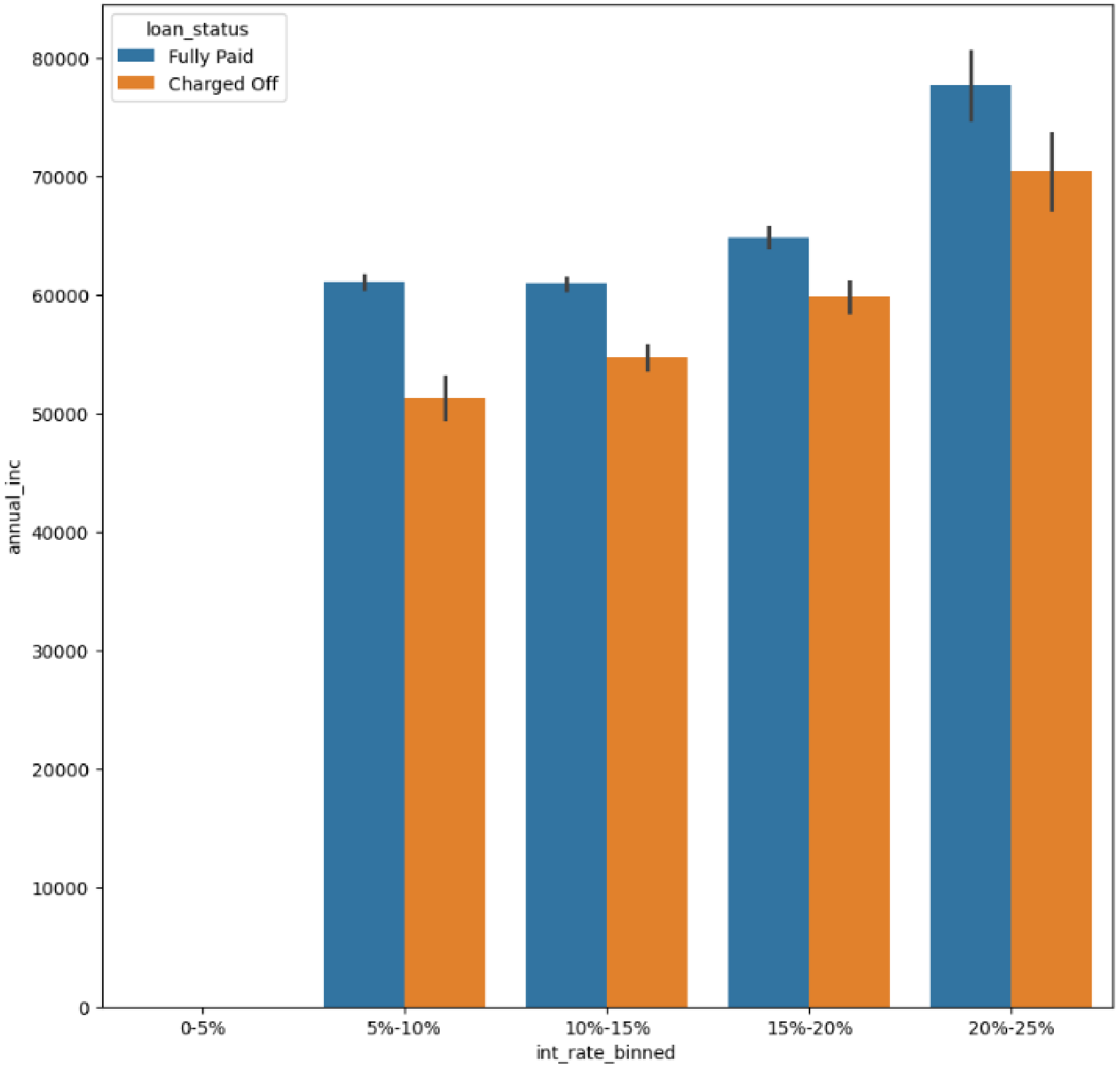
Analysing Annual Income across Loan Status and Other Variables

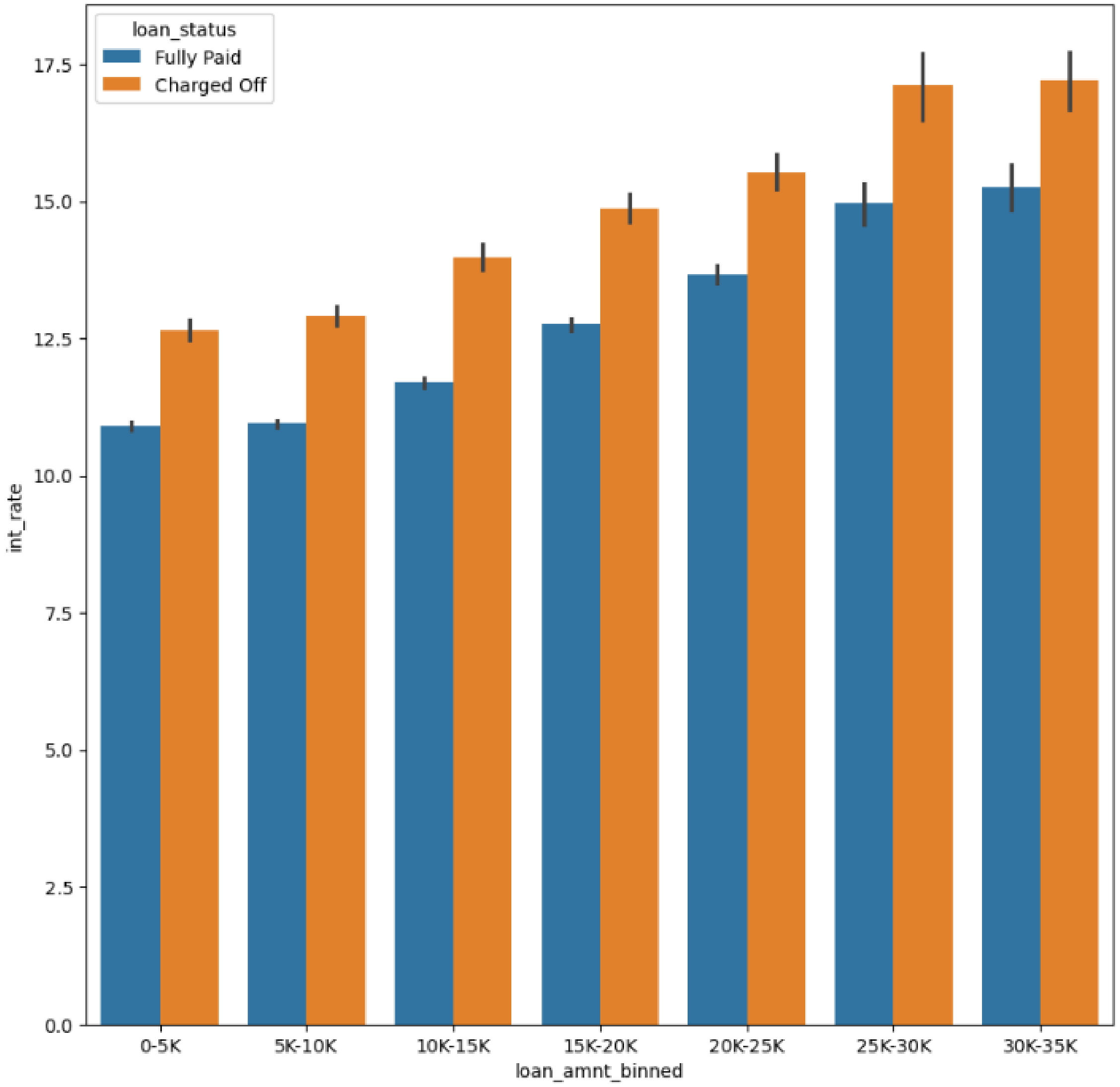


Annual Income across Purpose/ Home Ownership for different loan Status

Bivariate and Multivariate Analysis

Analysing Annual Income and Loan Amount across Loan Status and Interest rate



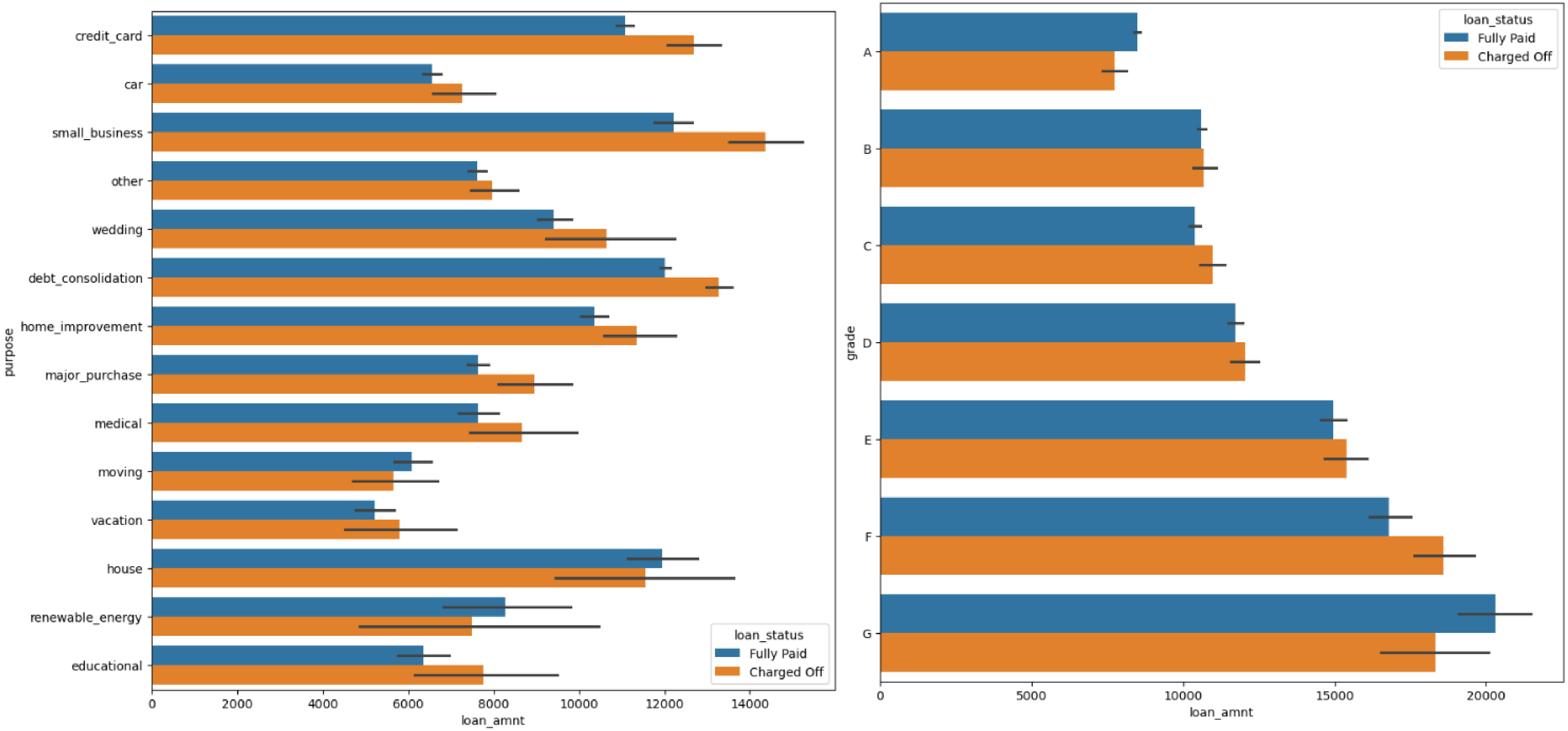


Annual Income across Interest Rate for different loan Status

Loan Amount across Interest Rate for different loan Status

Bivariate and Multivariate Analysis

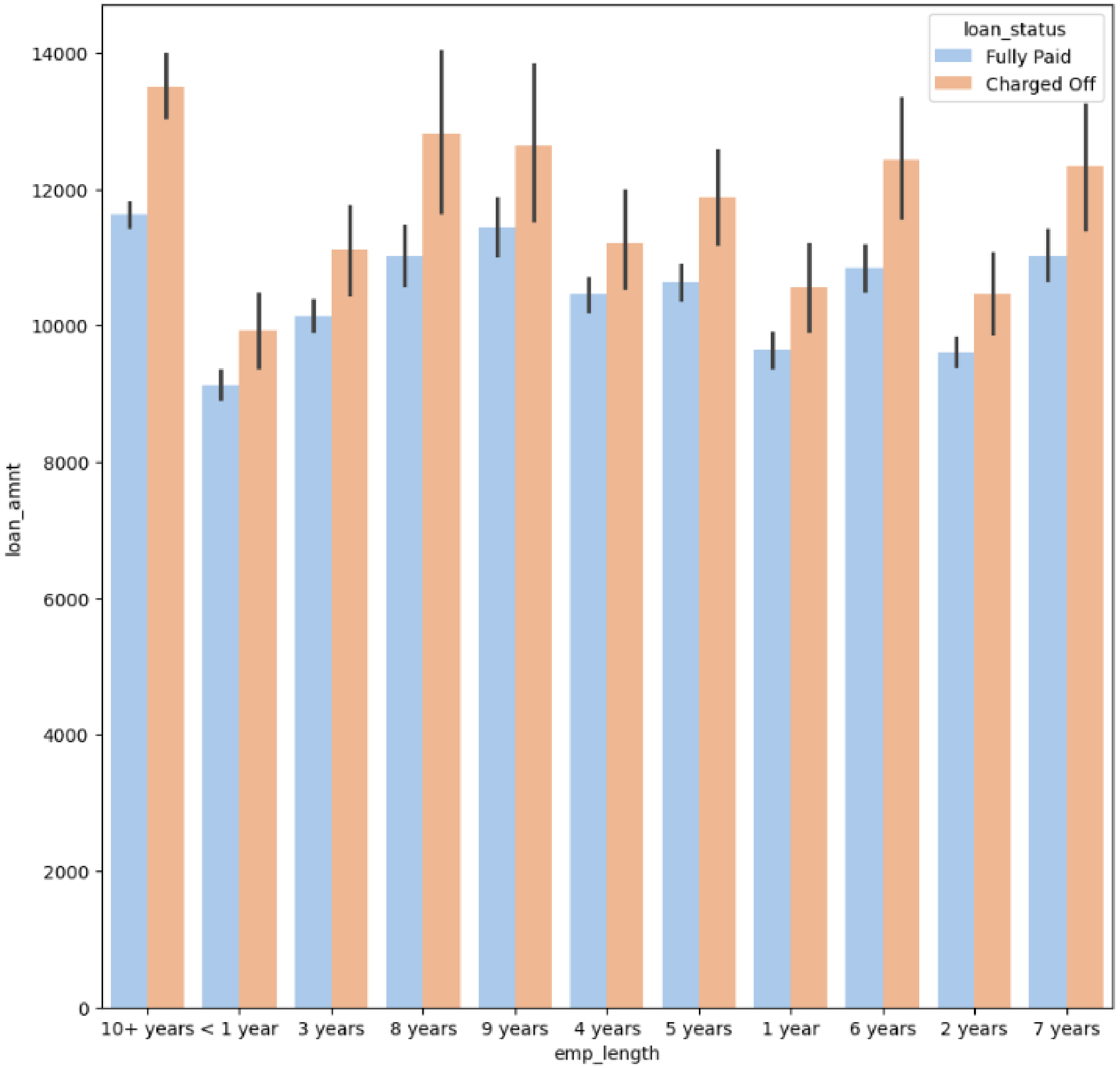
Analysing Loan amount across Loan Status and Other Variables

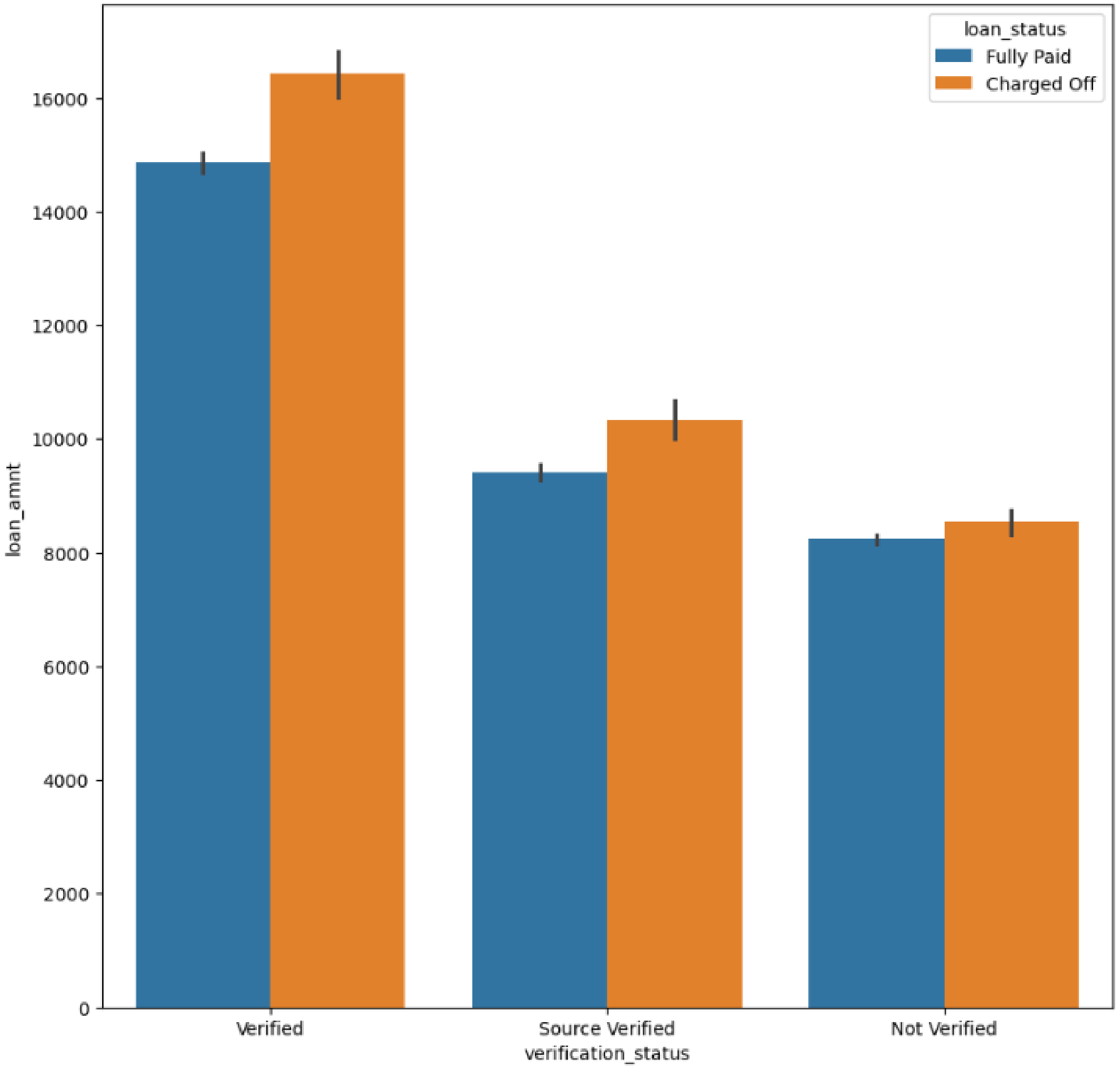


Loan Amount across Purpose/Grade for different loan Status

Bivariate and Multivariate Analysis

Analysing Loan amount across Loan Status and Other Variables





Loan Amount across Employment Length/Verification Status for different loan Status

Bivariate and Multivariate Analysis

Observations and Inferences

- Applicants taking loans for home improvement have high incomes (>= 60 K).
- Applicants with mortgage home ownership have high incomes (>= 60 K).
- Applicants with high interest rates of 20-25 % fall in high income bracket (>=70 K).
- Applicants with high loan amount in the range 30k - 35k are charged high interest rates (>=15 %).
- Applicants with small business have loan requirements of higher amounts (>=12K).
- Applicants with mortgaged houses have loan requirements of higher amounts (>=10 K).
- Applicants with worse grades have higher loan amount.
- Applicants with more experience have higher loan amount.
- Applicants with verified income have higher loan amount.- Applicants with worse grades have higher interest rate.

Conclusions and Recommendations

- **Loan Purpose and Loan Default Risk**: Debt consolidation loans are risky loans and more susceptible to default (based on number of defaults). Small Business loans have high default rates. The lending company should be cautious when approving loans for these purpose and consider high pricing for the riskier loans.
- **Housing Status and Default Risk**: Rented and Mortgaged houses are more likely to default (based on number of defaults).Other category has high default rates This fact needs to be taken into account when underwriting the loans and careful credit analysis and pricing should be done when extending such loans.
- **Verification Status and Default Risk**: Non-verified incomes are more likely to default(based on number of defaults) whereas Verified Loans have high default rate (logically makes less sense). The lending company can invest more effort and resources in verifying the incomes effectively before extending the loans.
- **Geography and Default Risk**: Loan applicants in California, Florida and New York are more likely to default (based on number of defaults). Nebraska has high default rates. The lending company can implement stricter credit policies and adjust the pricing based on loan riskiness.
- **Loan Term and Default Risk**: Short term loans (term of 36 months) are more likely to default (based on number of defaults). Long term loans have high default rates. The lending company should carefully evaluate the loans.
- **Grade and Default Risk**: Loans with B and C grades are seen to have more defaults (based on number of defaults) .Worse grades (G,F) have higher default rates. Loans with worse grades are riskier and should be priced appropriately.
- **Experience and Default Risk**: Applicants with 10 + years of experience are more likely to default (based on number of defaults and default rate). Although, the total number of employees with 10+ years of experience is significantly higher than in other categories. But still, additional parameters should be considered when evaluating the credit-worthiness of a loan applicant

Conclusions and Recommendations

- **Seasonal Trends and Default Risk**: There are more loan applications in December and Q4.Loans issued in this period are also seen to have higher defaults (based on default number and default rate). Lending company should exercise rigorous credit worthiness analysis in this high peak season by developing their infrastructure and resource management.
- **Lending Amount and Default Risk**: Applicants with loan amount requests and funding of 5K-10K are more likely to default. The lending company should carefully analyse the loan applications with intermediate loan amount requests.
- **DTI and Default Risk**: High Debt to Income Ratio applicants are more likely to default. The lending company should incorporate this information when underwriting the loans and should appropriately price the loans for higher riskiness.
- **Income and Default Risk**: Income Bracket of 30K-60K are more likely to default(based on default count) and 0K-30K based on default rate. The lending company should carefully evaluate the credit worthiness of low-moderate income brackets for their repaying capacity when extending the loan.
- **Credit History and Default Risk**: Applicants with less credit history (less open credit lines, total credit lines, revolving balance or low utilisation rates) are more likely to default. The lending company critically evaluate the loan applications with less or no credit history.