Lending Club Case Study

Group Members:

- Mitesh Upadhyay
- Ajith Menon

Contents

Problem Statement

- ✓ Lending Club specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

- ✓ Two types of risks are associated with the bank's decision.

    - ○ If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
    - ○ If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

- ✓ 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.

Business Objective

- ✓ If the company approves the loan, there are 3 possible scenarios described below:

Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
Current: Applicant is in the process of paying the instalments, 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 instalments in due time for a long period of time, i.e. he/she has defaulted on the loan.

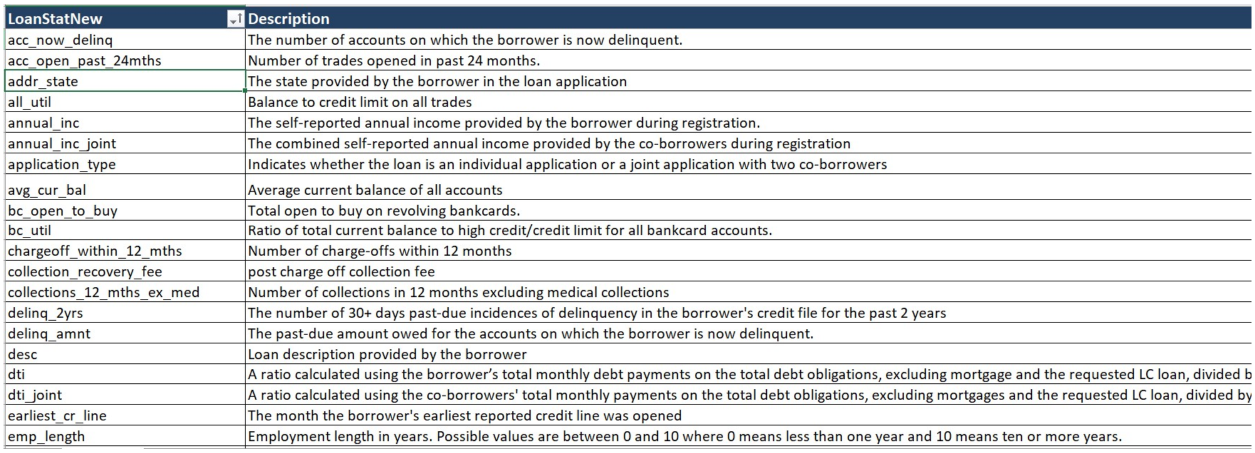
- ✓ Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

- ✓ Company wants to understand the driving factors 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.

Analysis Approach

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 b
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
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.

Data Understanding

Key Columns of Significance:

Customer Perspective:

- ✓ **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.
- ✓ **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 Perspective:

- ✓ **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.
- ✓ **Interest Rate (int_rate):** Represents the annual rate at which the borrower will be charged interest on the loan amount.

Excluded Columns:

- ✓ 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 (emp_title, desc, title) will be dropped as they contain descriptive text (nouns) and do not contribute to the analysis.
- ✓ Columns (id, member_id) will be dropped as they are index variables with unique values and do not contribute to the analysis.

Data Cleaning and Pre Processing

- ✓ **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.

- ✓ **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

- ✓ **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.

- ✓ **Dropping Records & Columns:** 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 Conversion:** Converted columns like interest rate to float and terms to int

- ✓ **Imputing values in Columns** : Mapped employment length with the respective number of years in int.

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

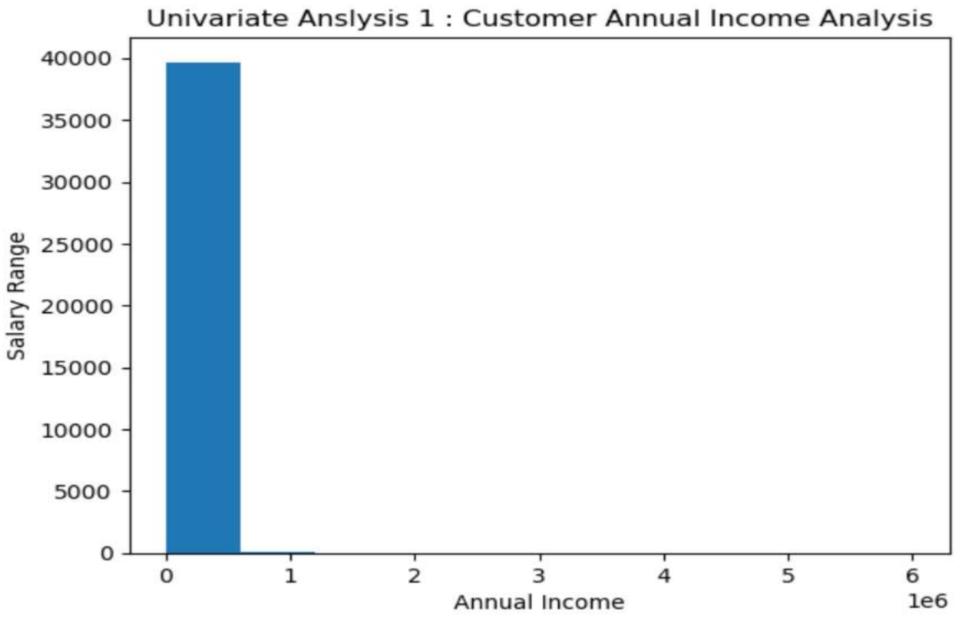
Grade (grade)	Address State (addr_state)
Term (36 / 60 months) (term)	Loan purpose (purpose)
Employment length (emp_length)	Home Ownership (home_ownership)
	Loan status (loan_status)

B. Quantitative Variables:

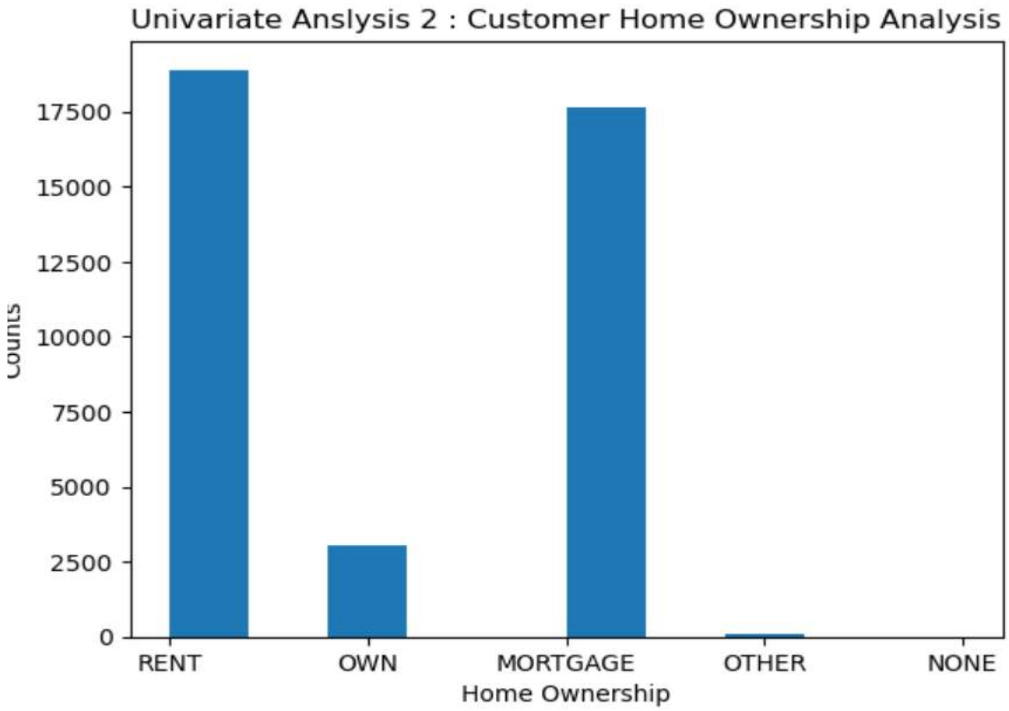
- ✓ Interest rate bucket (int_rate_bucket)
- ✓ Annual income bucket (annual_inc_bucket)
- ✓ Loan amount bucket (loan_amnt_bucket)
- ✓ Funded amount bucket (funded_amnt_bucket)
- ✓ Debt to Income Ratio (DTI) bucket (dti_bucket) ✓ Monthly Installment (installment)

Univariate Analysis (Unordered Categorical)

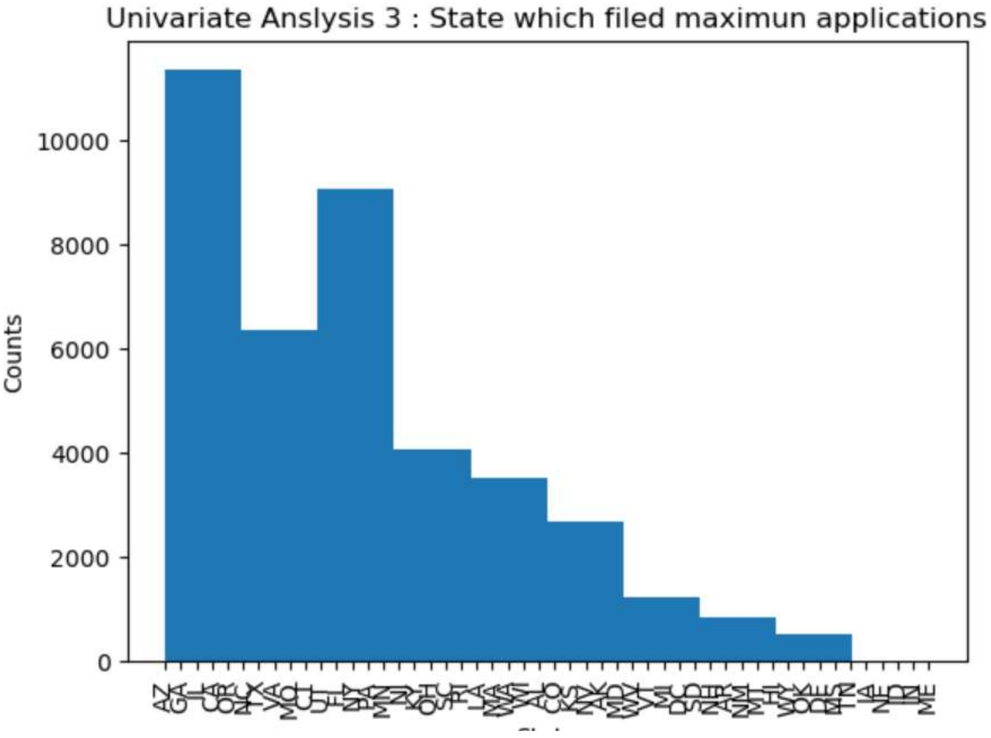
1. annual_inc : Most of the customer has the Income from 0 - 40 k



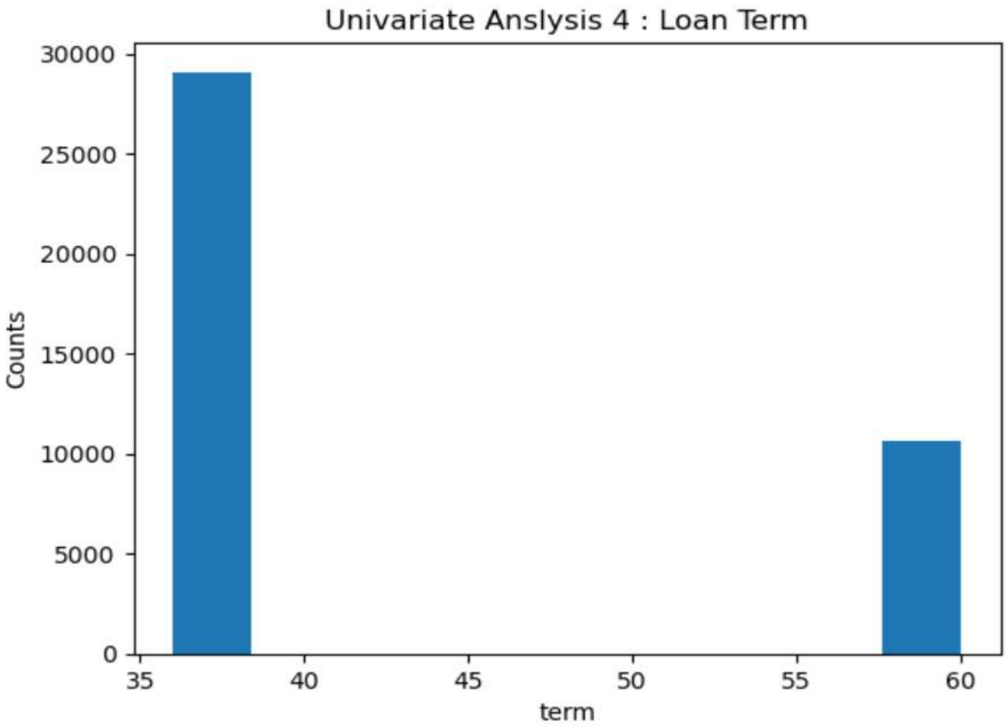
2. Mostly Customers has home status either RENT and MORTGAGE



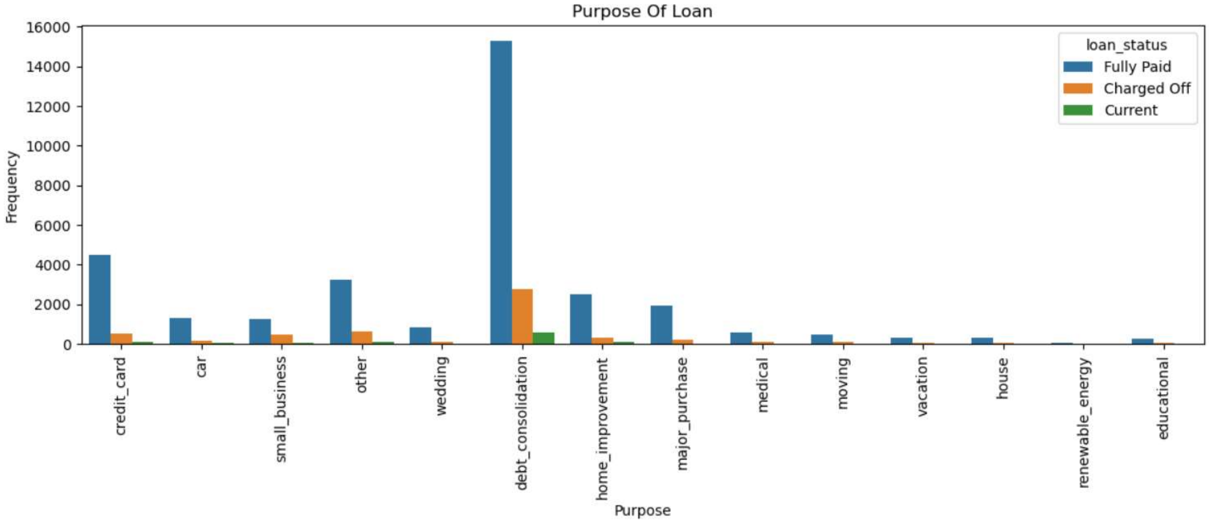
3. State which has highest Applications filed



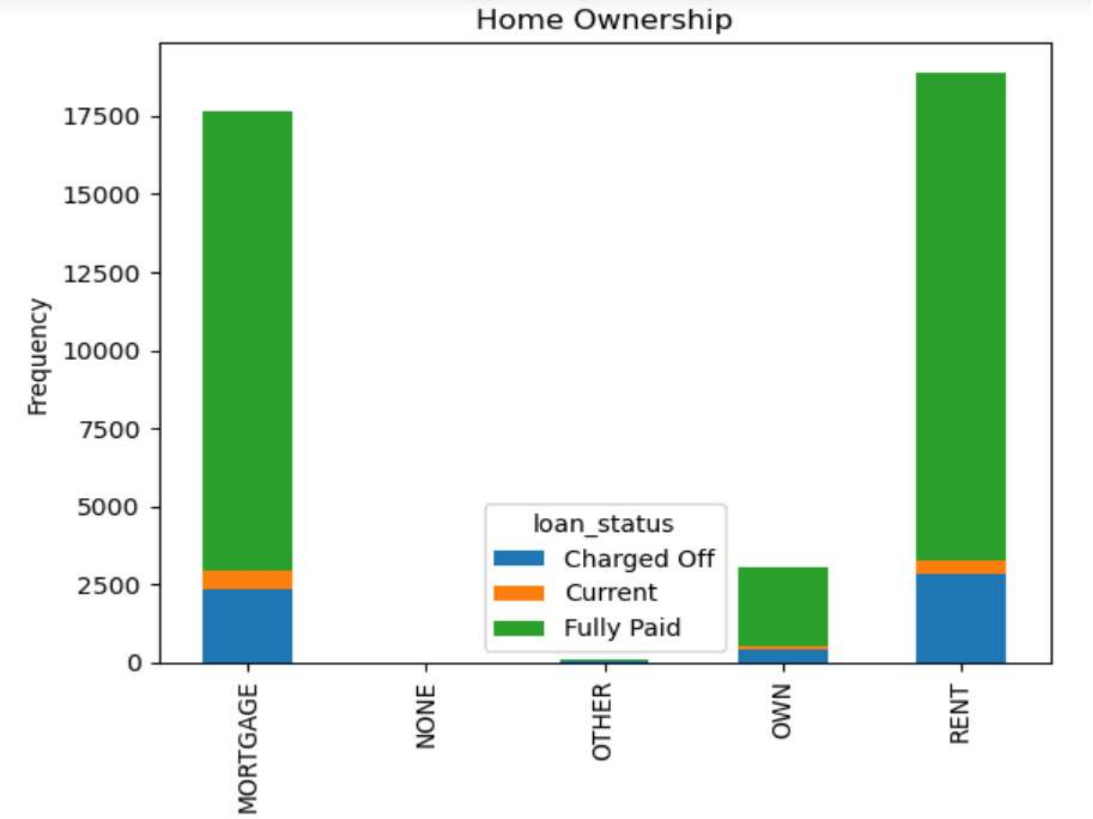
4. Most of the customers has opted for term of 36 months



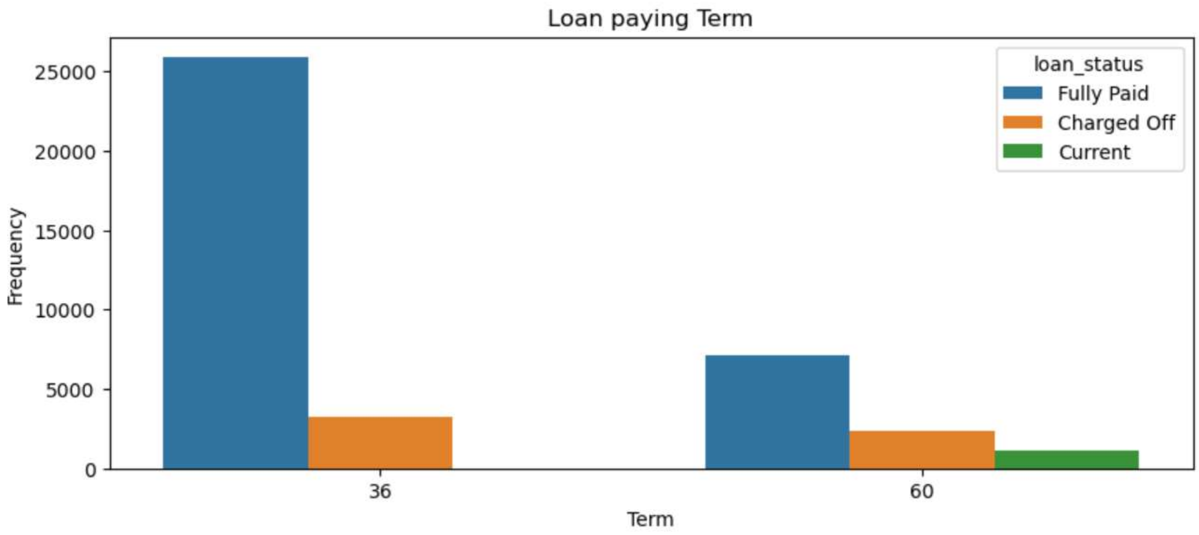
5. Most of Customer opted load for debt consolidation



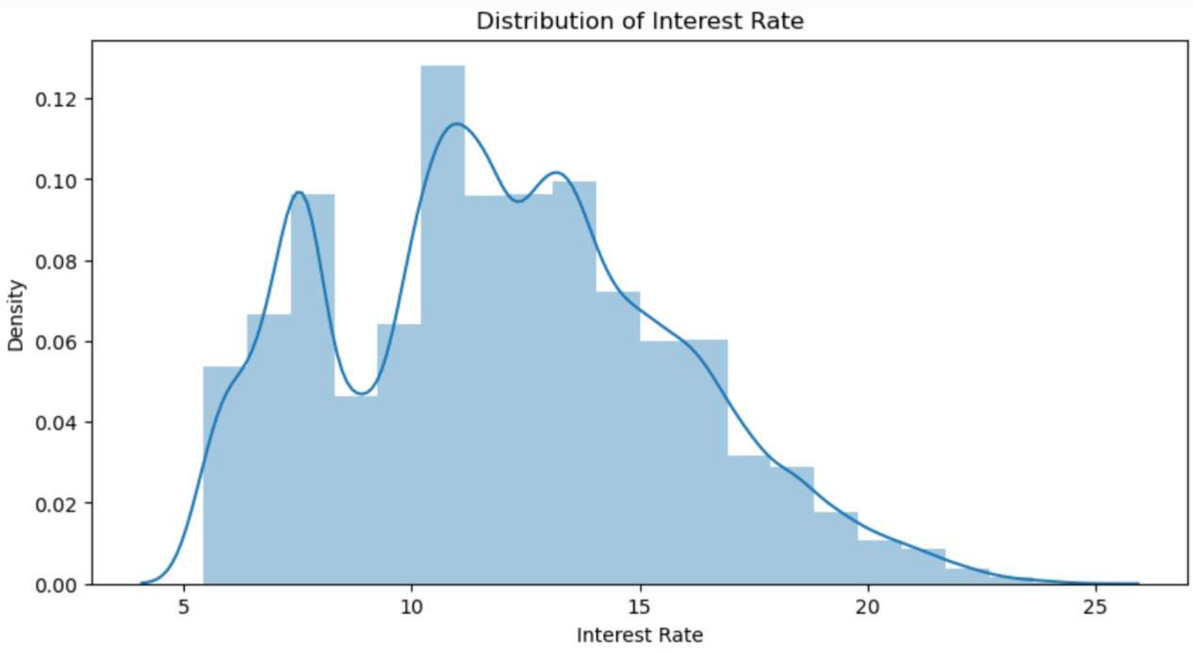
5. Most of them have taken loan who are in rent or mortgage their home



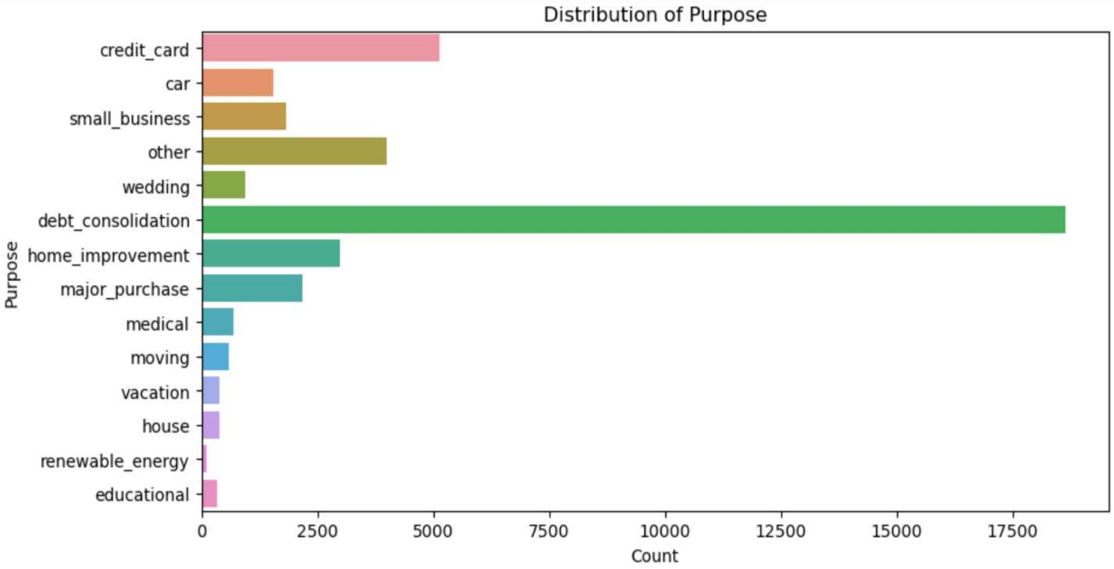
6. Most of customers have taken loan for 36 months as compared to 60 months



7. The interest rate is more crowded around 5-10 and 10-15 with a drop near 10

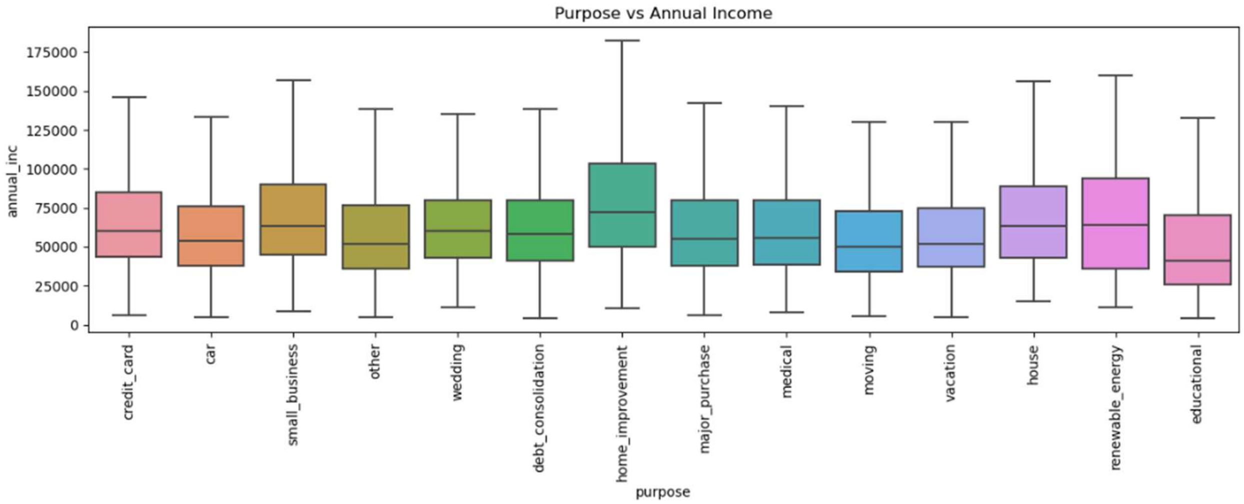


8. Major loans are taken for debt consolidation followed by credit card

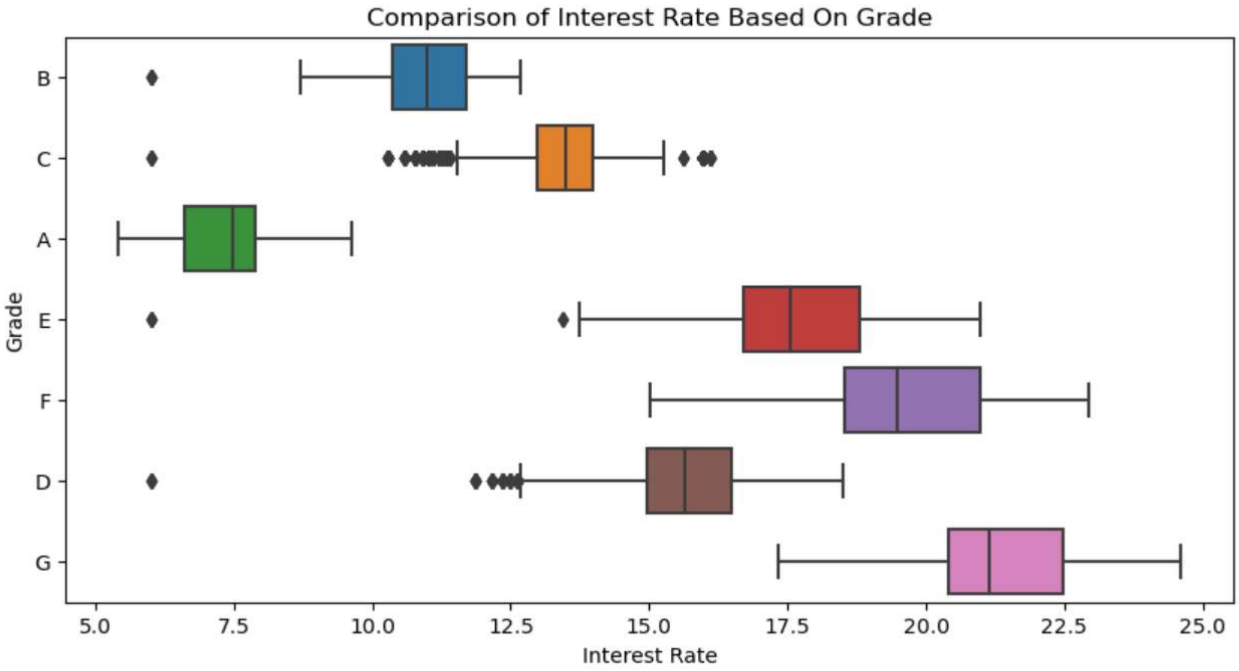


Bivariate Analysis

1. In Annual Income vs Purpose variable we can say that, the borrowers who has high annual income are taking loans mostly for home improvement and small business



2. The Grade represent risk factor thus we can say interst rate increases with the risk.



Thank You

[Type here]