LENDING CLUB CASE STUDY

[Data Analysis and Insights]

Author: Gouranga Seth and Nitin Ngaich

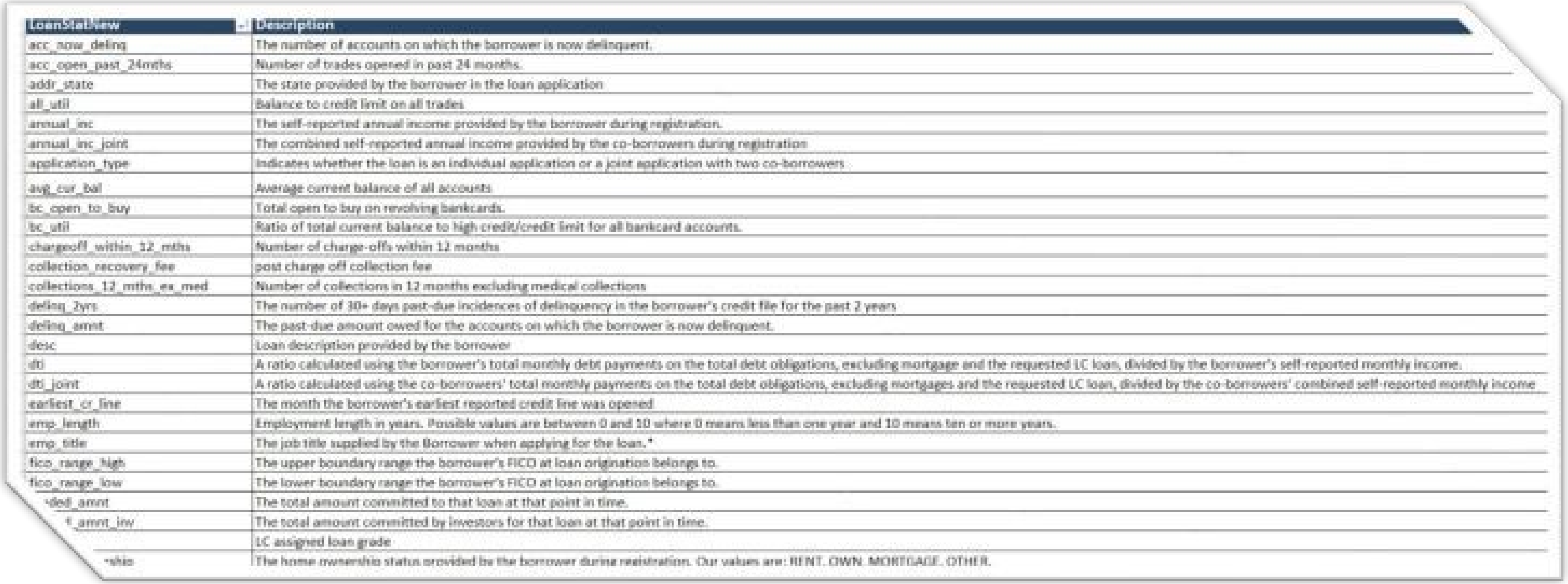
AGENDA

- Introduction
- Problem Statement
- Data Understanding
- Data Cleaning & Pre-processing
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Correlation Analysis
- Suggestions
- References & Useful Links

INTRODUCTION

- **Problem Statement**
  - Minimizing financial losses from loan approval process.
  - Losses occur when borrowers default on loans.
- **Objective:**
  - Reduce credit losses by identifying risky applicants Data Cleaning & Pre-processing
  - Approving loans for likely-to-repay applicants generates profit.
  - Approving loans for likely-to-default applicants results in losses. Univariate Analysis
- **EDA**
  - Exploratory Data Analysis to understand driving factors behind loan default.
  - Knowledge used for portfolio and risk assessment.

DATA DESCRIPTION



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.
f_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.

DATA UNDERSTANDING

- Primary Attribute: Loan Status
  - Fully_Paid: Customers who have successfully repaid their loans.
  - Charged_Off: Customers who have defaulted on their loans.
  - Current: Customers whose loans are presently in progress
- Decision Matrix: Loan Acceptance Outcome
  - Fully Paid: Applicants who have successfully repaid both the principal and the interest rate of the loan.
  - Current: Applicants in the process of making loan installments
  - Charged-off: Applicants who have failed to make timely installments.
- Key Columns of Significance
  - Customer Demographics: Annual Income, Home Ownership, Employment Length, Debt to Income, State
- Excluded Columns: Customer Behavior Columns

DATA UNDERSTANDING

- Granular Data
  - Columns with excessive detail will be omitted.
  - Example: 'sub grade' column
- Columns with NA values
  - 54 columns contain only NA values.
  - These columns will be removed.
- Columns with only 0 values
  - These columns will also be dropped.

DATA CLEANING & PRE-PROCESSING

- Loading data from loan CSV
  - Conversion of mixed data types
- Checking for null values in the dataset
  - 48% of columns with null values were dropped.
- Checking for unique values
  - 9 columns with single unique values were removed.
- Checking for duplicated rows in data
- Dropping Records & Columns
- Common Functions
- Data Conversion
- Outlier Treatment
- Imputing values in Columns

DATA CLEANING & PRE-PROCESSING

- Null Values for pub_rec_bankruptcies
  - 660 null values dropped.
  - Cannot be imputed.
- Post Data Cleaning and Preprocessing
  - 36094 rows × 18 columns left.

UNIVARIATE ANALYSIS INSIGHTS

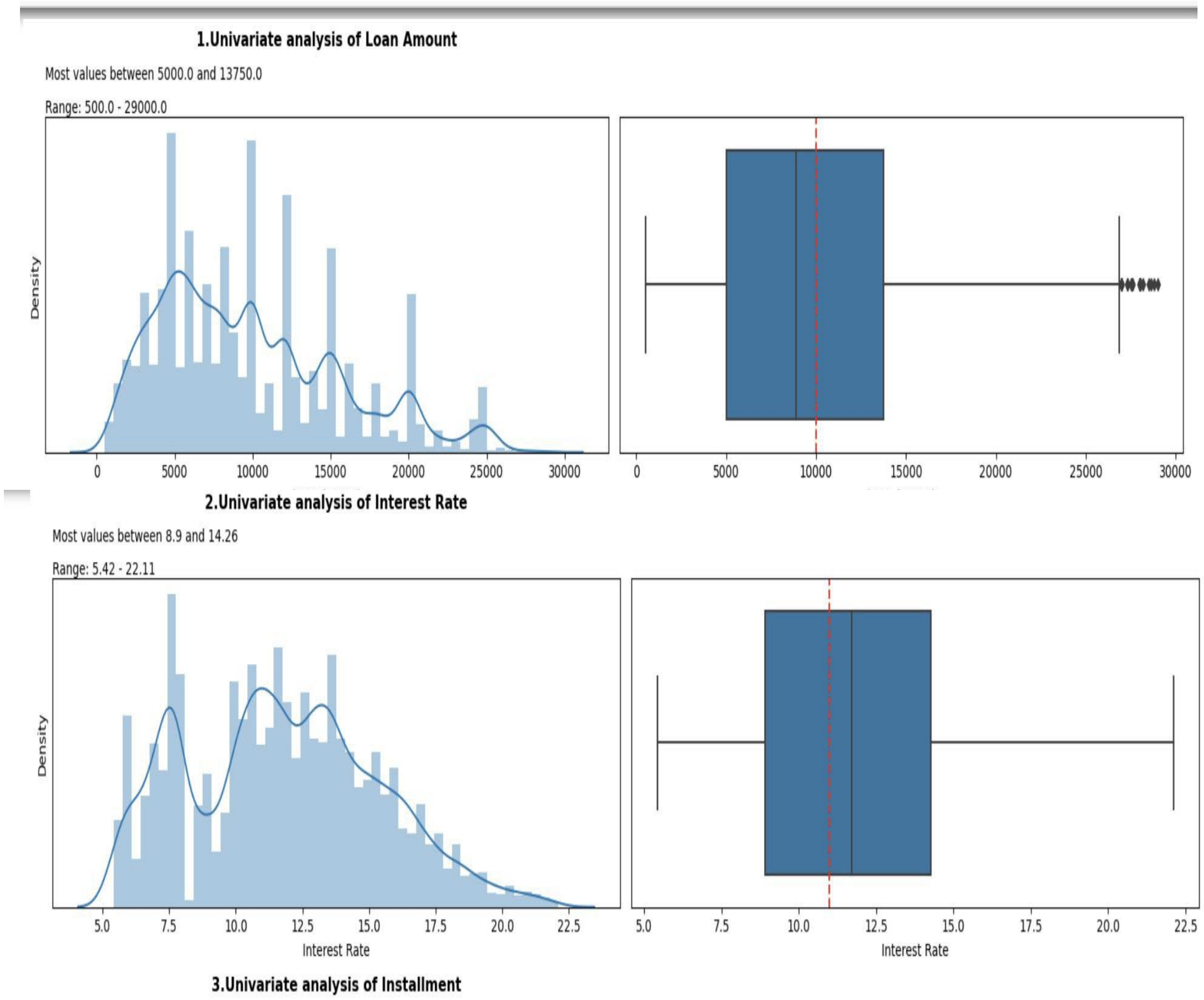
- Target customer segments: Focus on customers with annual income between 0-40K, debt-to-income ratio of 0-20%, and employment length of 10+ years or 0-2 years.

- Analyze low categories: Investigate why other loan purpose categories, such as credit card and major purchase, have lower application counts.

- Prepare for Q4 volume: Anticipate high loan application volume in Q4 due to financial challenges, festive seasons, and debt consolidation.

- Target other quarters for sales growth: Develop strategies to increase loan applications in Q1, Q2, and Q3 to achieve sales growth throughout the year.

UNIVARIATE ANALYSIS:



1.Univariate analysis of Loan Amount

Most values between 5000.0 and 13750.0

Range: 500.0 - 29000.0

2.Univariate analysis of Interest Rate

Most values between 8.9 and 14.26

Range: 5.42 - 22.11

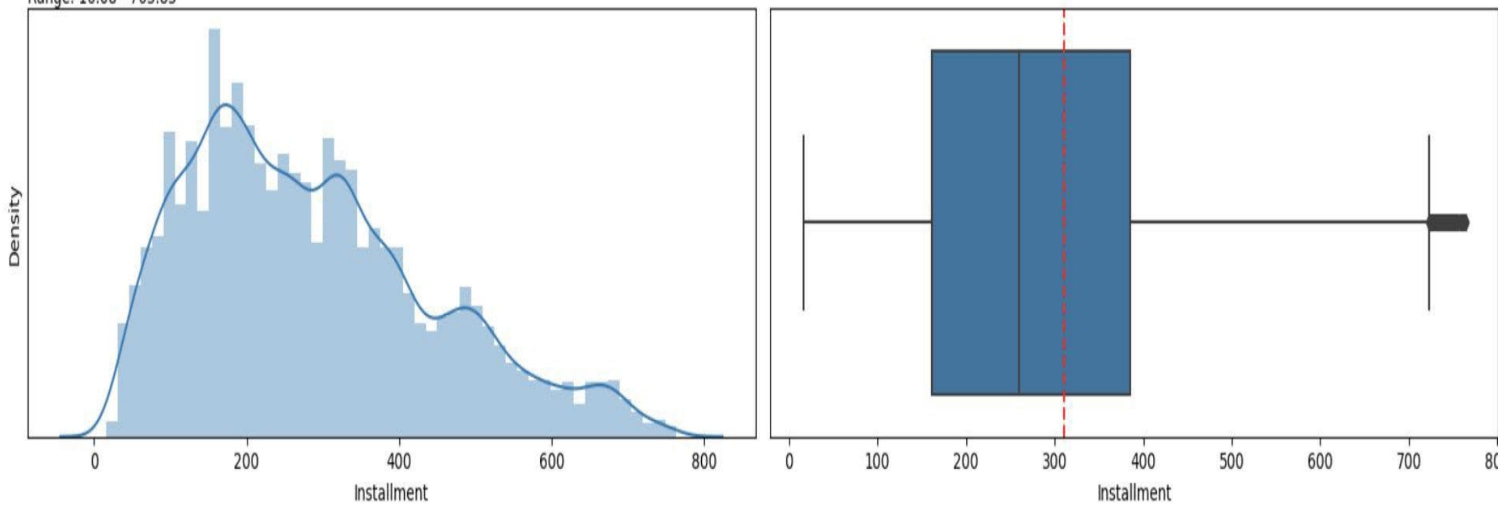
3.Univariate analysis of Installment

UNIVARIATE ANALYSIS:

3.Univariate analysis of Installment

Most values between 161.01500000000001 and 385.78

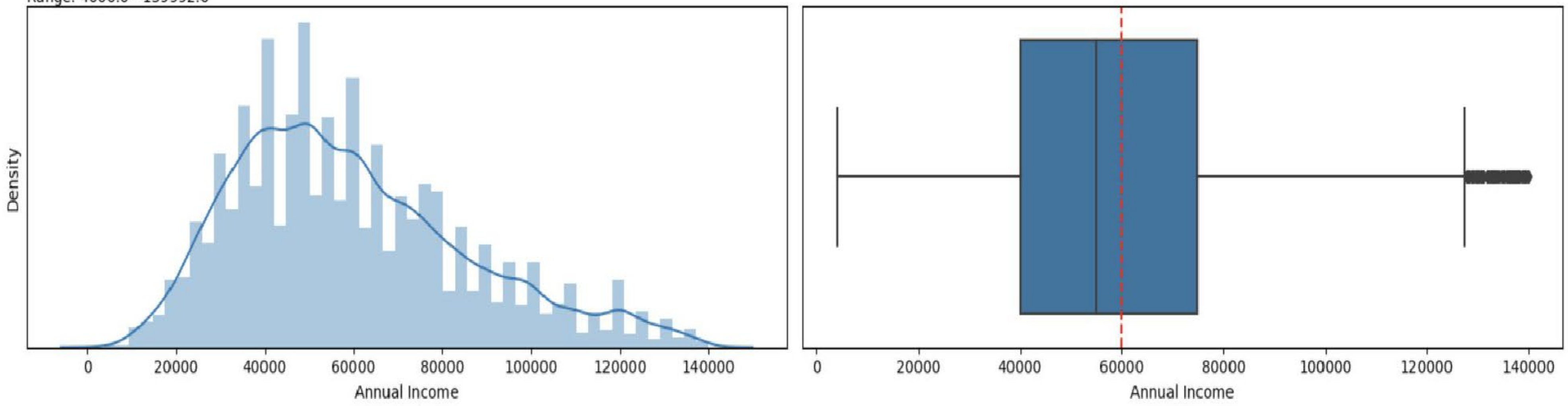
Range: 16.08 - 763.83



4.Univariate analysis of Annual Income

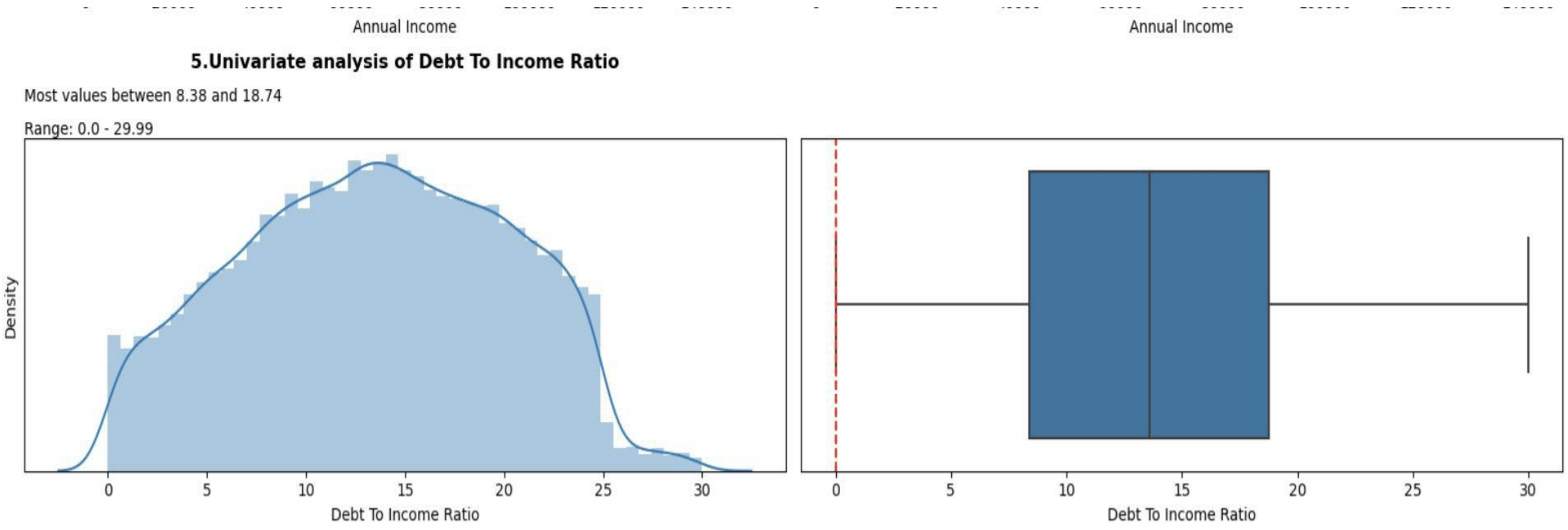
Most values between 40000.0 and 75000.0

Range: 4000.0 - 139992.0



5.Univariate analysis of Debt To Income Ratio

UNIVARIATE ANALYSIS:

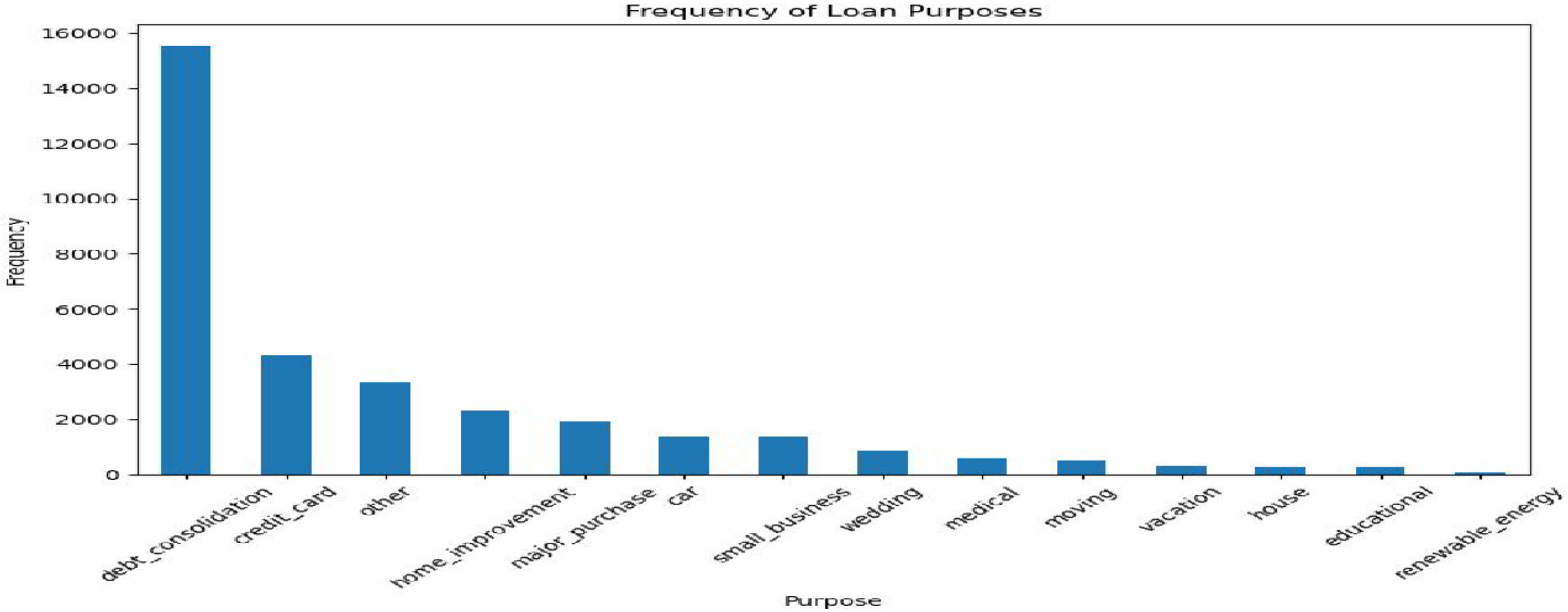


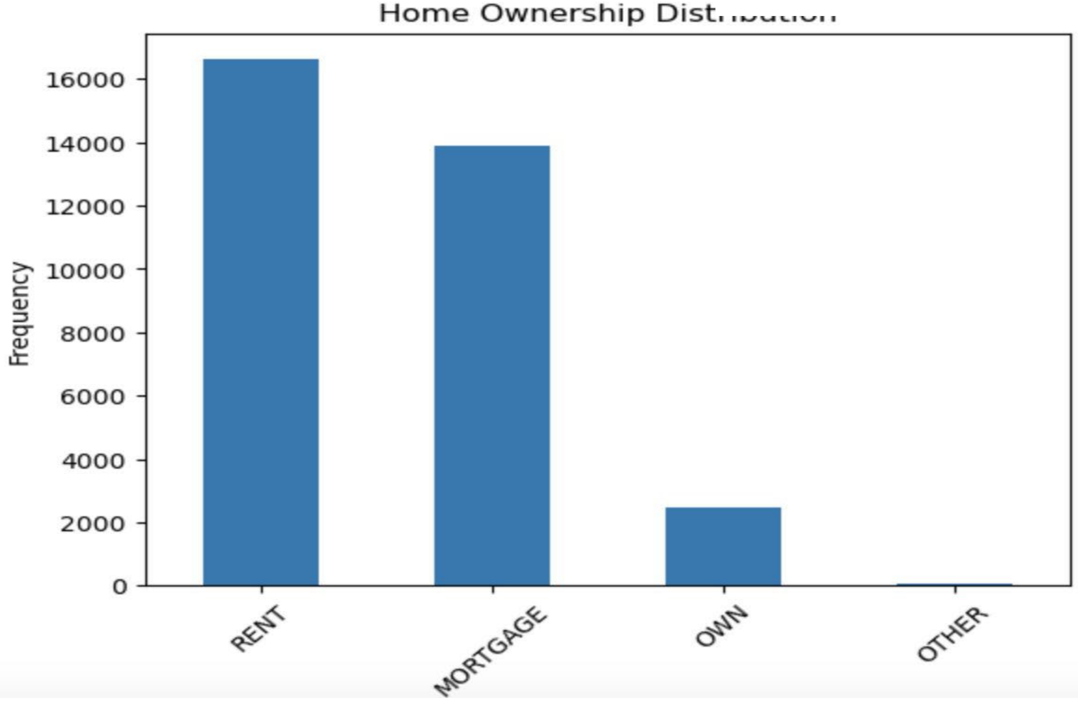
5.Univariate analysis of Debt To Income Ratio

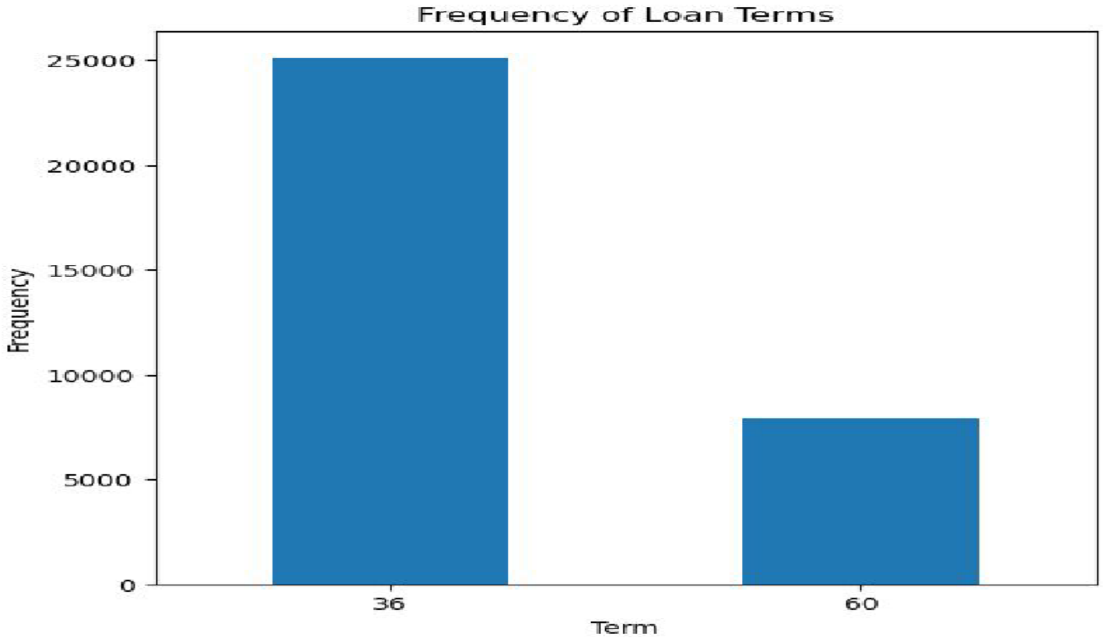
Most values between 8.38 and 18.74

Range: 0.0 - 29.99

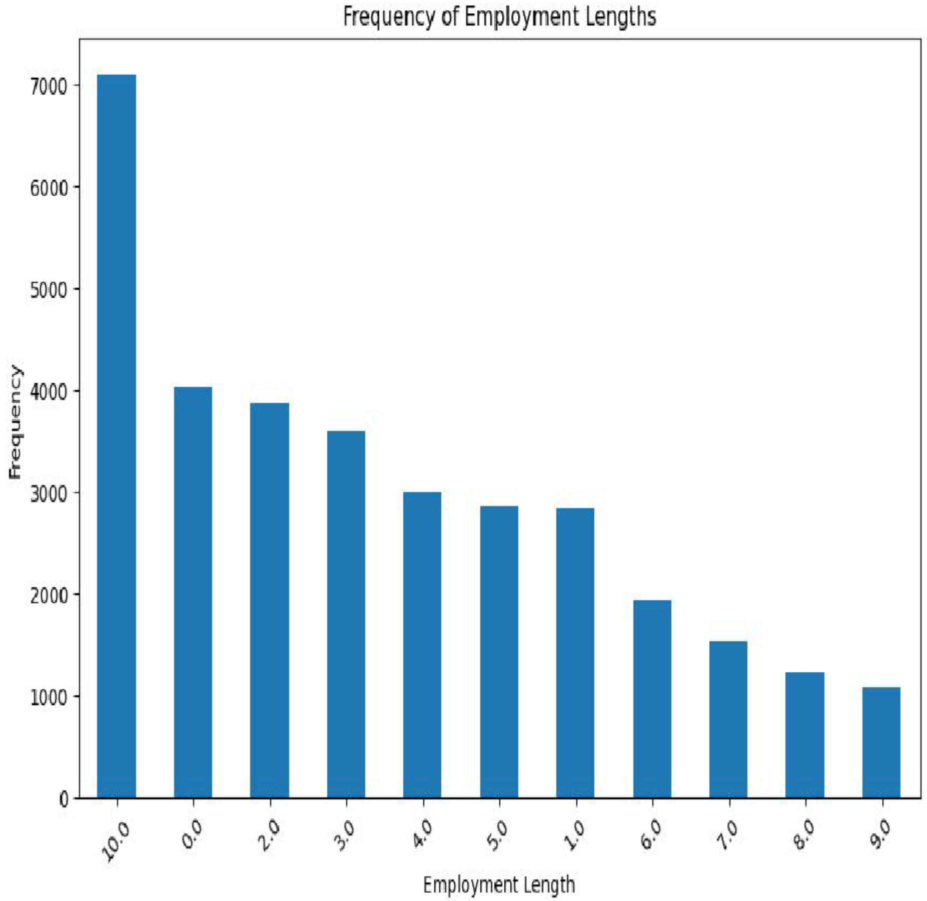
UNIVARIATE ANALYSIS:

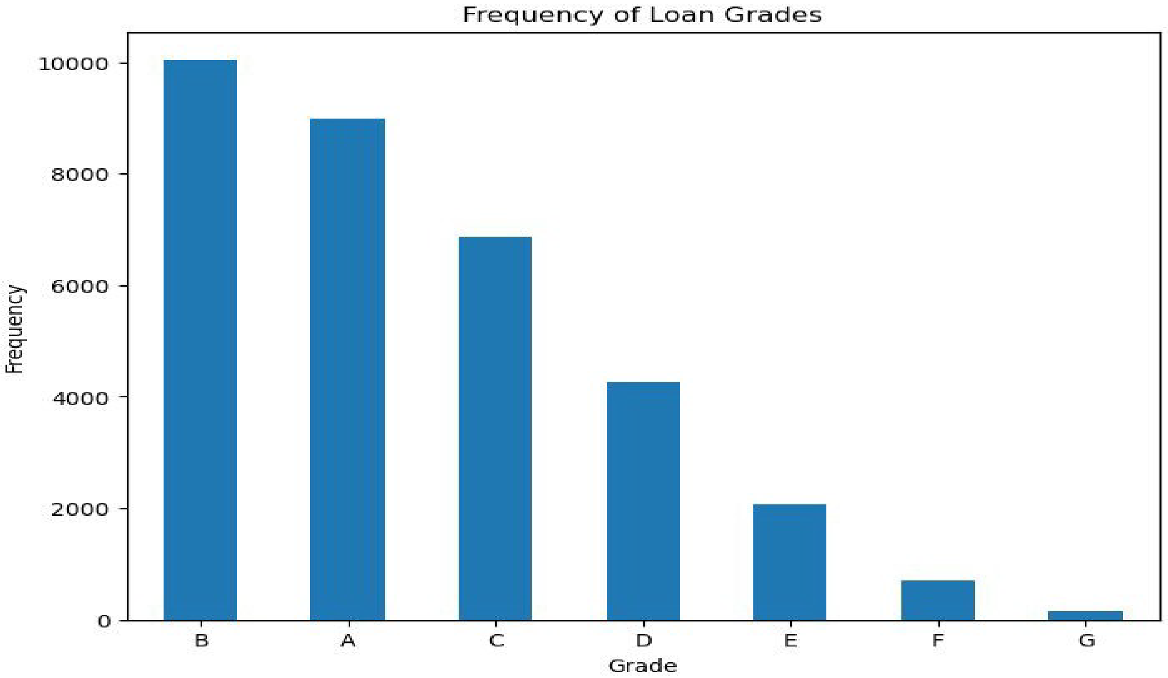


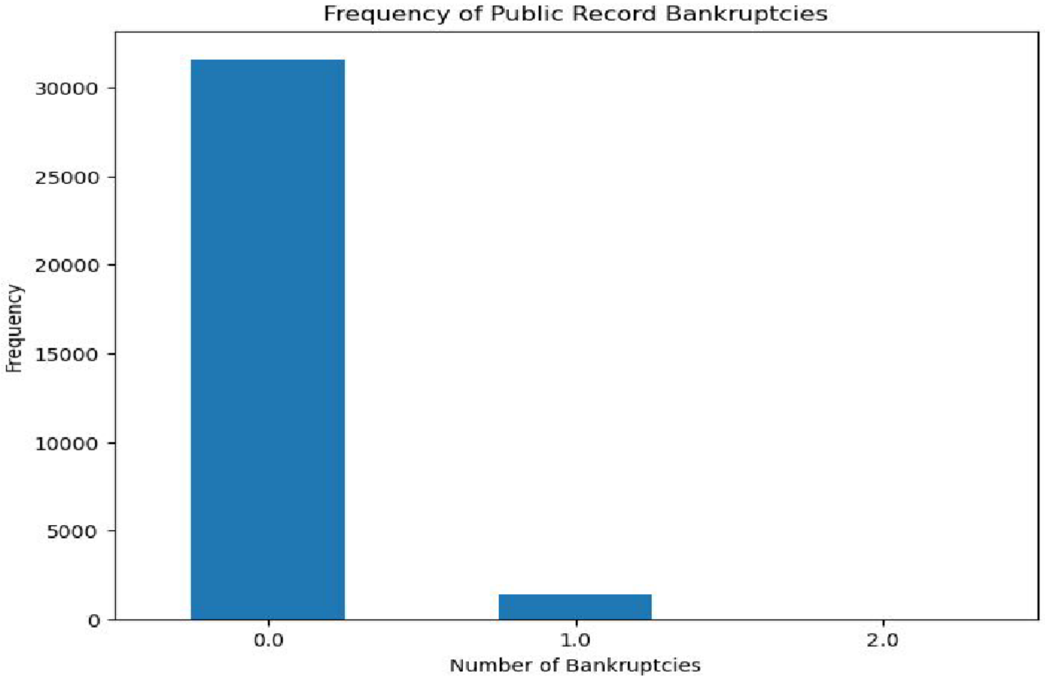




UNIVARIATE ANALYSIS:



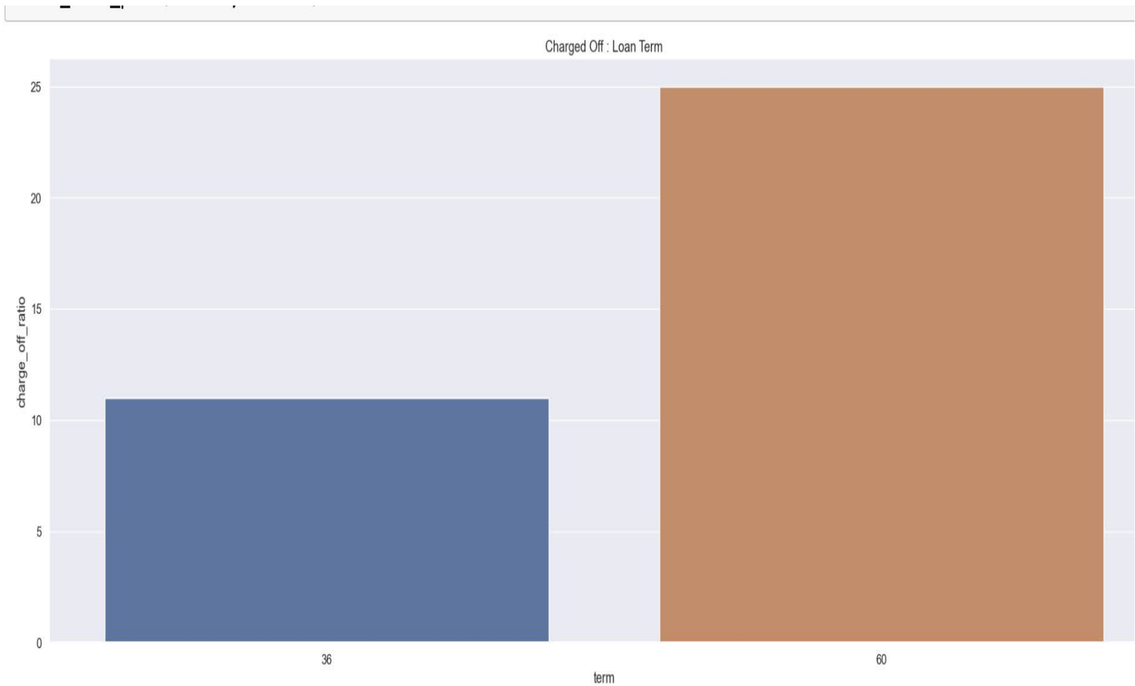


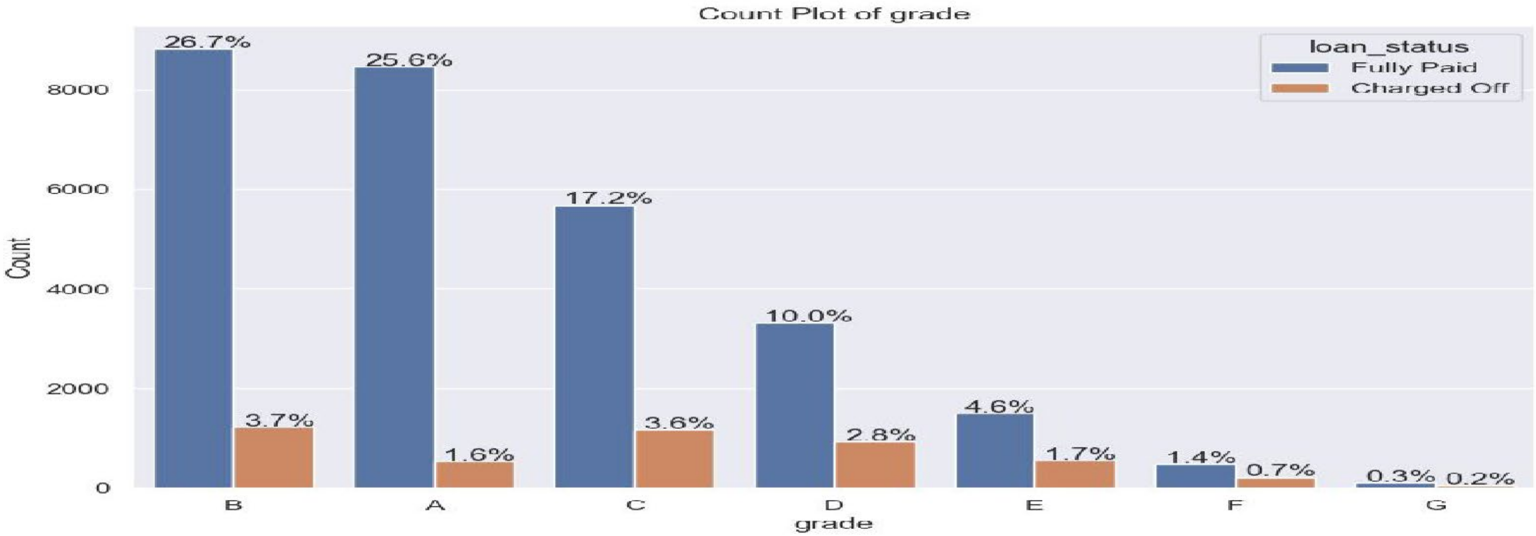


BIVARIATE ANALYSIS

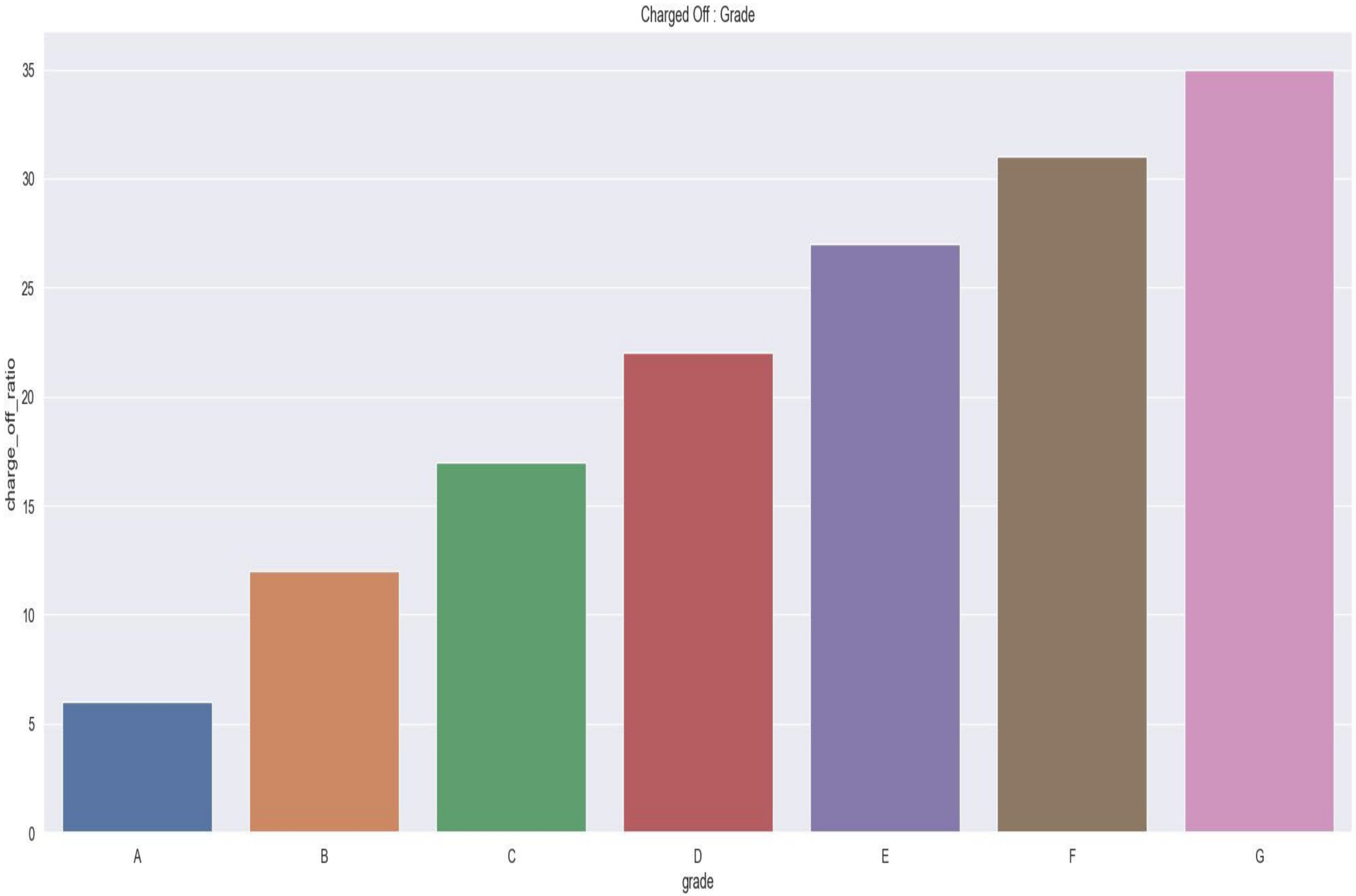
- Bivariate analysis is a statistical method that involves the simultaneous analysis of two variables (factors).
  - It aims to determine the empirical relationship between them.
  - The analysis can be used to test hypotheses, identify patterns, or explore relationships between the variables.
- It was carried out for both Categorical and Quantitative Variables
  - Categorical Variables: Ordered and Unordered
  - Quantitative Variables: Int Rate Bucket, Debt to Income Bucket, Annual Income Bucket, Funded Amount Bucket, Loan Amount Bucket
- Bivariate Analysis Observations
  - Ordered Categorical Variables: The loan applicants belonging to Grades B, C, and D contribute to most of the

BIVARIATE ANALYSIS:

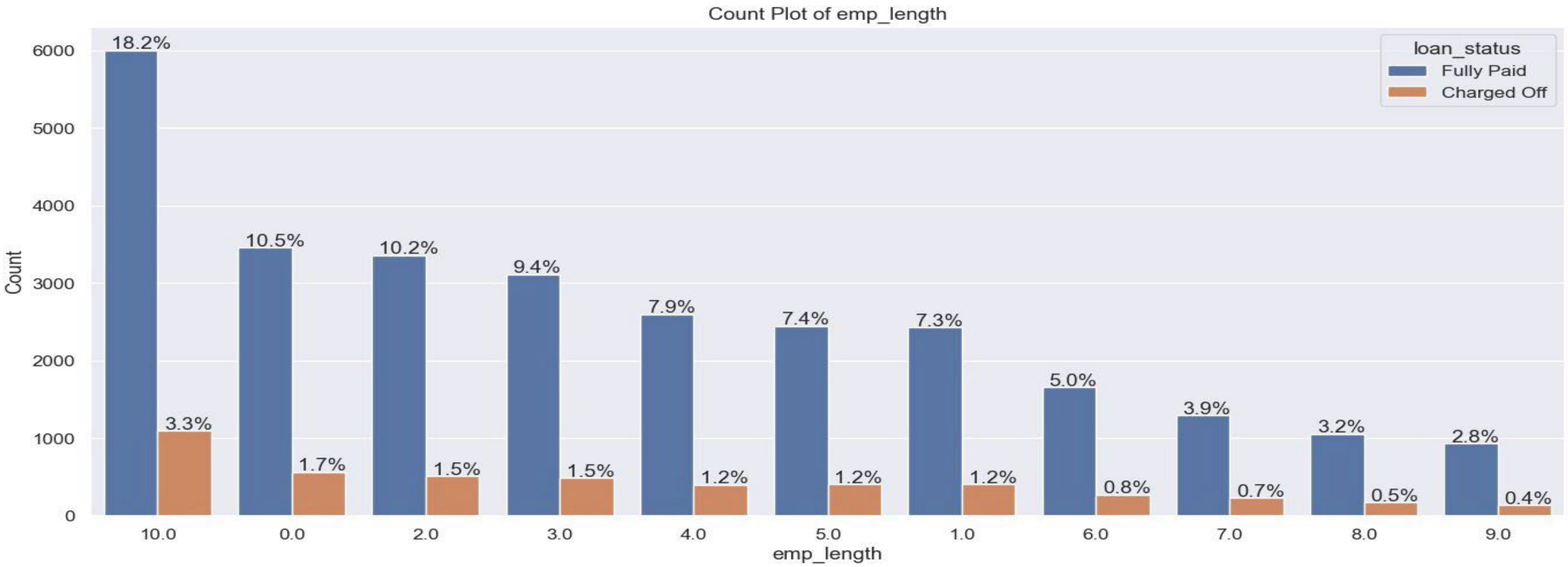




BIVARIATE ANALYSIS:



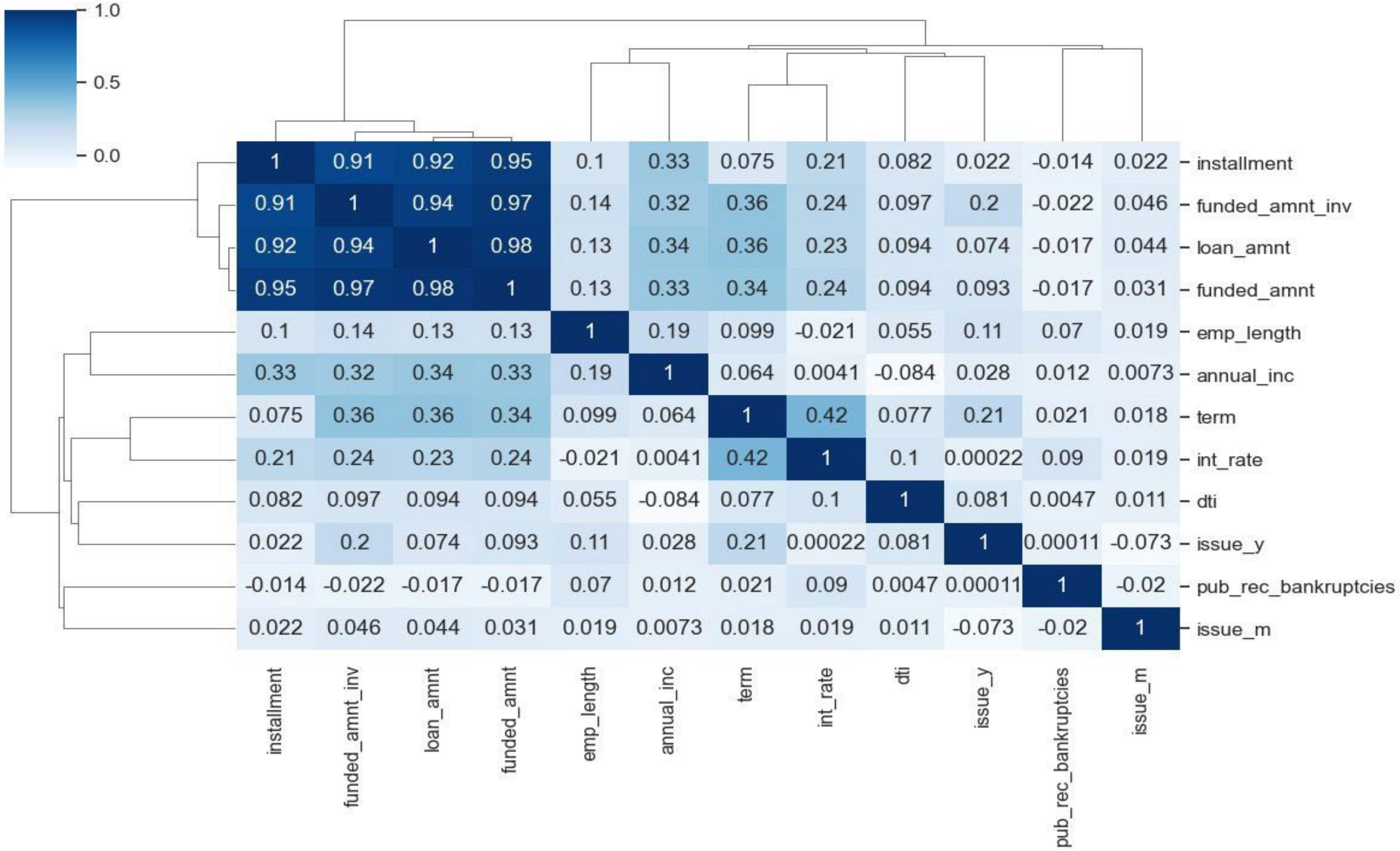
BIVARIATE ANALYSIS:



MULTIVARIATE ANALYSIS

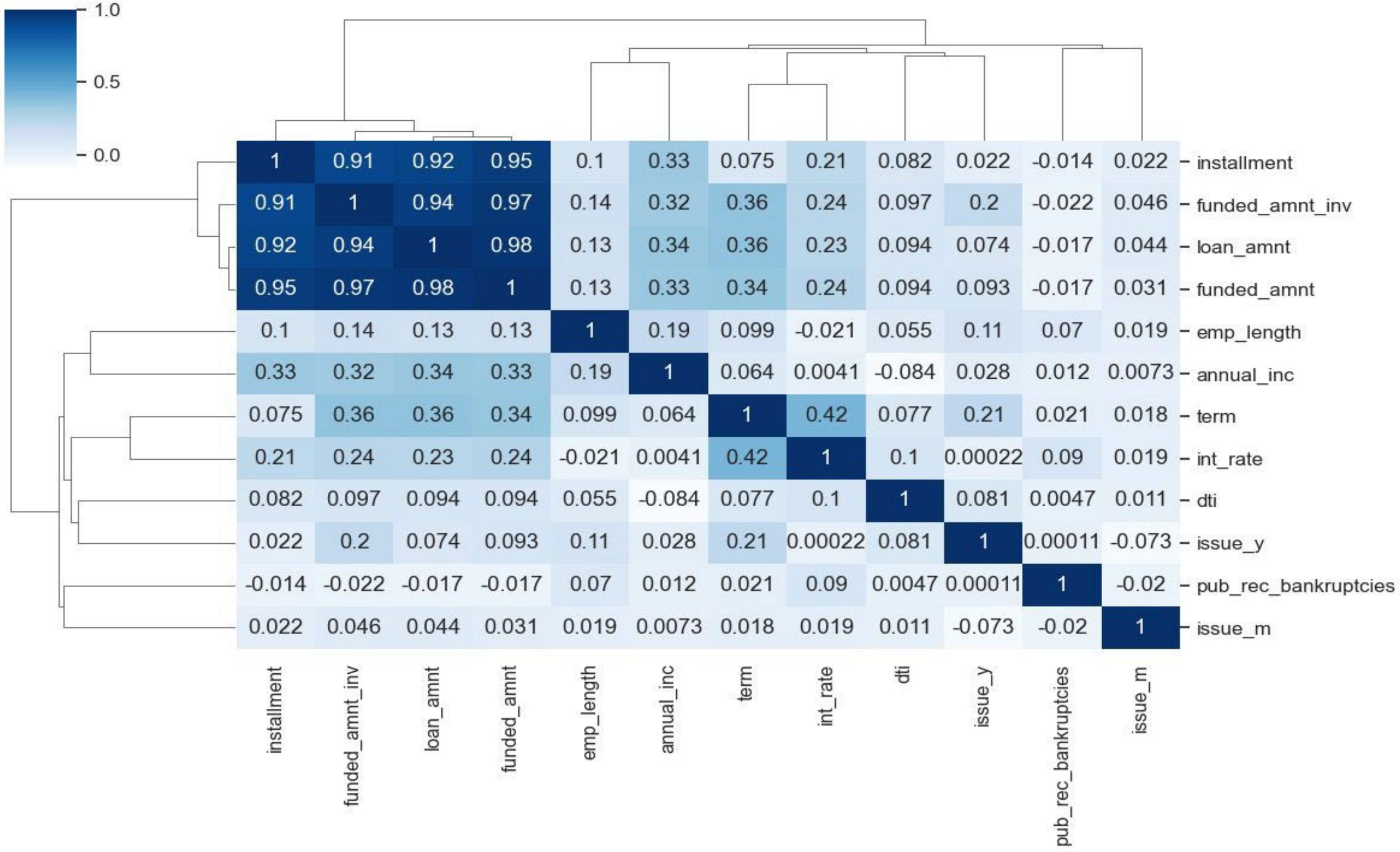
- Statistical technique used to analyze data involving more than two variables.
  - Examines relationships between multiple variables simultaneously.
- Widely used in various fields
  - Economics, social sciences, biology, marketing, and environmental science
- Can include different types of variables.
  - Categorical, numerical, or a combination of both
- Observations and 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.

MULTIVARIATE ANALYSIS:



	installment	funded_amnt_inv	loan_amnt	funded_amnt	emp_length	annual_inc	term	int_rate	dti	issue_y	pub_rec_bankruptcies	issue_m
installment	1	0.91	0.92	0.95	0.1	0.33	0.075	0.21	0.082	0.022	-0.014	0.022
funded_amnt_inv	0.91	1	0.94	0.97	0.14	0.32	0.36	0.24	0.097	0.2	-0.022	0.046
loan_amnt	0.92	0.94	1	0.98	0.13	0.34	0.36	0.23	0.094	0.074	-0.017	0.044
funded_amnt	0.95	0.97	0.98	1	0.13	0.33	0.34	0.24	0.094	0.093	-0.017	0.031
emp_length	0.1	0.14	0.13	0.13	1	0.19	0.099	-0.021	0.055	0.11	0.07	0.019
annual_inc	0.33	0.32	0.34	0.33	0.19	1	0.064	0.0041	-0.084	0.028	0.012	0.0073
term	0.075	0.36	0.36	0.34	0.099	0.064	1	0.42	0.077	0.21	0.021	0.018
int_rate	0.21	0.24	0.23	0.24	-0.021	0.0041	0.42	1	0.1	0.00022	0.09	0.019
dti	0.082	0.097	0.094	0.094	0.055	-0.084	0.077	0.1	1	0.081	0.0047	0.011
issue_y	0.022	0.2	0.074	0.093	0.11	0.028	0.21	0.00022	0.081	1	0.00011	-0.073
pub_rec_bankruptcies	-0.014	-0.022	-0.017	-0.017	0.07	0.012	0.021	0.09	0.0047	0.00011	1	-0.02
issue_m	0.022	0.046	0.044	0.031	0.019	0.0073	0.018	0.019	0.011	-0.073	-0.02	1

MULTIVARIATE ANALYSIS:



	installment	funded_amnt_inv	loan_amnt	funded_amnt	emp_length	annual_inc	term	int_rate	dti	issue_y	pub_rec_bankruptcies	issue_m
installment	1	0.91	0.92	0.95	0.1	0.33	0.075	0.21	0.082	0.022	-0.014	0.022
funded_amnt_inv	0.91	1	0.94	0.97	0.14	0.32	0.36	0.24	0.097	0.2	-0.022	0.046
loan_amnt	0.92	0.94	1	0.98	0.13	0.34	0.36	0.23	0.094	0.074	-0.017	0.044
funded_amnt	0.95	0.97	0.98	1	0.13	0.33	0.34	0.24	0.094	0.093	-0.017	0.031
emp_length	0.1	0.14	0.13	0.13	1	0.19	0.099	-0.021	0.055	0.11	0.07	0.019
annual_inc	0.33	0.32	0.34	0.33	0.19	1	0.064	0.0041	-0.084	0.028	0.012	0.0073
term	0.075	0.36	0.36	0.34	0.099	0.064	1	0.42	0.077	0.21	0.021	0.018
int_rate	0.21	0.24	0.23	0.24	-0.021	0.0041	0.42	1	0.1	0.00022	0.09	0.019
dti	0.082	0.097	0.094	0.094	0.055	-0.084	0.077	0.1	1	0.081	0.0047	0.011
issue_y	0.022	0.2	0.074	0.093	0.11	0.028	0.21	0.00022	0.081	1	0.00011	-0.073
pub_rec_bankruptcies	-0.014	-0.022	-0.017	-0.017	0.07	0.012	0.021	0.09	0.0047	0.00011	1	-0.02
issue_m	0.022	0.046	0.044	0.031	0.019	0.0073	0.018	0.019	0.011	-0.073	-0.02	1

CORRELATION ANALYSIS

Strong Positive Correlations:

- Loan amount (loan_amnt) is highly correlated with funded amount (funded_amnt) and funded amount inverse (funded_amnt_inv), indicating a strong relationship between these variables.
- Loan amount is also highly correlated with installment, suggesting that loan amount and monthly payments are closely related.

Weak Positive Correlations:

- Loan amount has a moderate positive correlation with term, annual income (annual_inc), and employment length (emp_length), indicating some relationship between these variables.
- Loan amount has a weak positive correlation with interest rate (int_rate) and debt-to-income ratio (dti).

Weak Negative Correlations:

- Loan amount has a weak negative correlation with public record bankruptcies (pub_rec_bankruptcies), indicating that loan amount and bankruptcy history are inversely related.
- Annual income has a weak negative correlation with debt-to-income ratio, suggesting that higher income is associated with lower debt-to-income ratios.

Other Observations:

- The issue year (issue_y) and issue month (issue_m) variables have weak correlations with other variables, indicating that they may not be strongly related to loan characteristics.
- The pub_rec_bankruptcies variable has weak correlations with most other variables, suggesting that it may be an independent factor in loan decisions.

SUGGESTIONS

- Implement Stricter Criteria for Grades B, C, and D
  - Minimize default risks with stricter risk assessment and underwriting criteria.
- Focus on Subgrades B3, B4, and B5
  - Consider additional risk mitigation measures or lower loan amounts.
- Evaluate and Limit 60-Month Loans
  - Decrease likelihood of defaults by limiting maximum term or adjusting interest rates
- Comprehensive Credit Scoring System
  - Incorporate various risk-related attributes for gauging creditworthiness.
- Capitalizing on Market Growth
  - Maintain competitive edge while ensuring robust risk management practices.
- Anticipate Peak Periods
  - Ensure efficient processing to meet customer demands during busy seasons.

REFERENCES & USEFUL LINKS

Technology Package	Version	Documentation
Python	3.11.4	https://www.python.org
Matplotlib	3.7.1	https://matplotlib.org/
NumPy	1.24.3	https://numpy.org/
Pandas	1.5.3	https://pandas.pydata.org/
Seaborn	0.12.2	https://seaborn.pydata.org/

- **GitHub Repository Link:** https://github.com/gseth2004/Lending_Club_Case_Study