Lending Club Case Study

Submitted by –

Drishty Arora
Angad Singh

Contents

- Problem Statement
- Data Description
- Data Understanding & Cleaning
- Univariate-Bivariate-Multivariate Analysis
- Key-Takeaways

Problem Statement

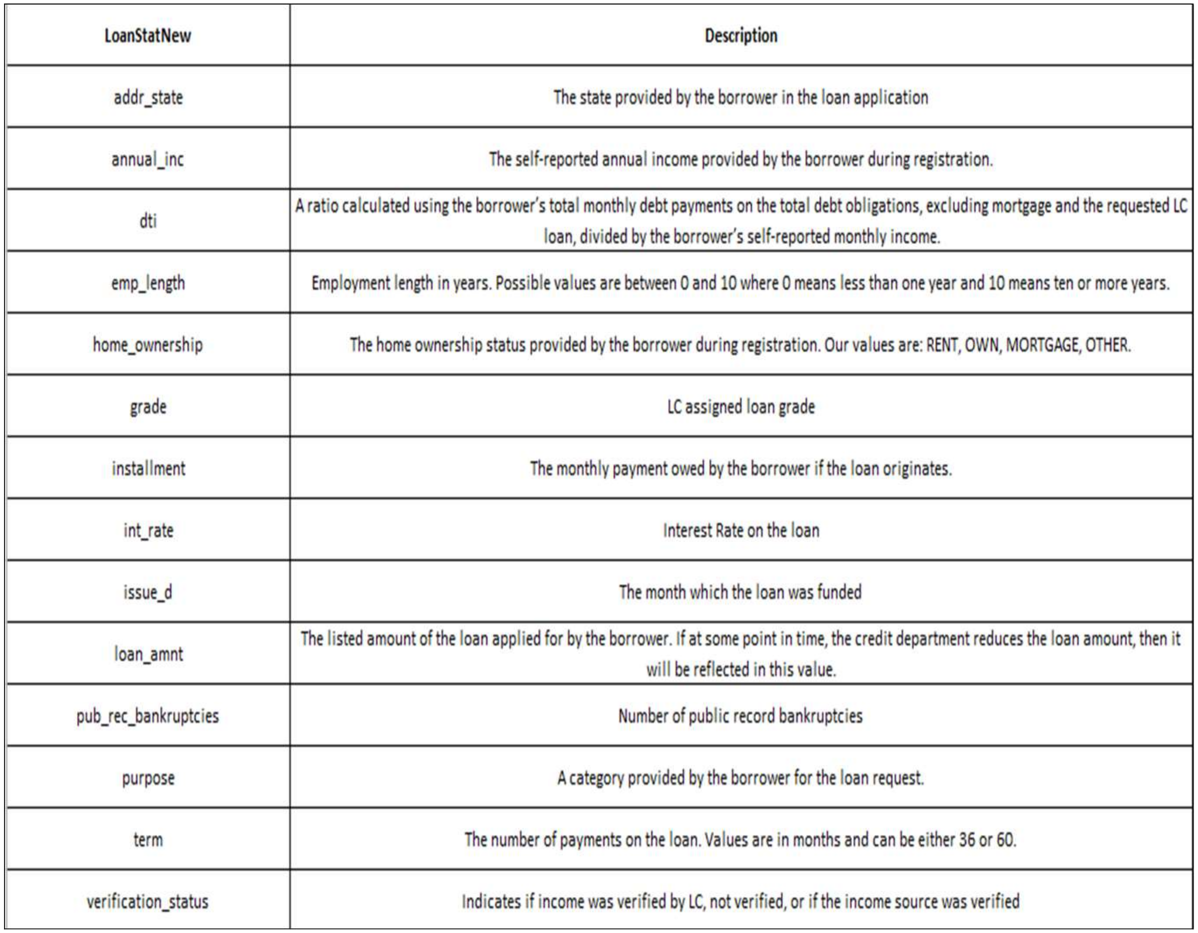
Lending Club, a Consumer Finance marketplace specializing in offering a variety of loans to urban customers, faces a critical challenge in managing its loan approval process. When evaluating loan applications, the company must make sound decisions to minimize financial losses, primarily stemming from 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:
1.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.
2.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.

Data Description



LoanStatNew	Description
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
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.
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.
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
grade	LC assigned loan grade
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified

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.

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 Row & Column Analysis Conclusions :

Row Analysis :

- Identifying and Removing Duplicate Rows: Duplicate rows in the dataset were not found.
- Dropping Loan Status Rows with "Current" status: Rows with a "loan_status" of "Current" are dropped as they represent loans in progress and do not contribute to loan approval decisions. This step also helps clean up unnecessary columns related to "Current" loans.

Column Analysis :

- Columns contain NA values are removed.
- Columns contain only zero values are dropped.
- Columns (id, membColumns (emp_title, desc, title) are dropped as they contain descriptive text (nouns) and do not contribute to the analysis.
- The redundant column (url) are dropped.
- 660 records for pub_rec_bankruptcies are dropped due to missing values.

Column Conversions :

The term column have the "months" text stripped and converted to an integer.

Percentage columns like (int_rate) are currently in object format. These columns have the "%" character stripped and is converted to float.

Columns (loan_amnt, funded_amnt, funded_amnt_inv) are currently in object format and converted to float.

Columns (int_rate, installment, dti) are currently in object format and converted to float.

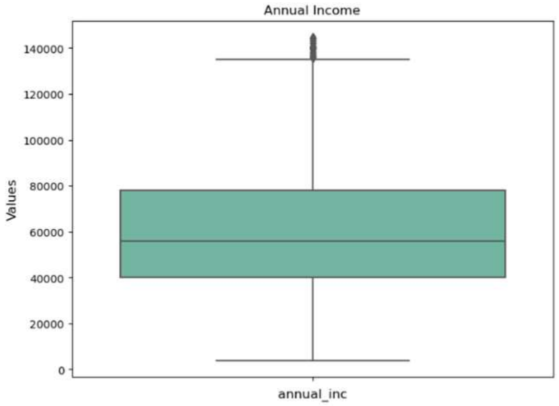
The issue_d column are converted to datetime format with date in YYYY-mm-dd format.

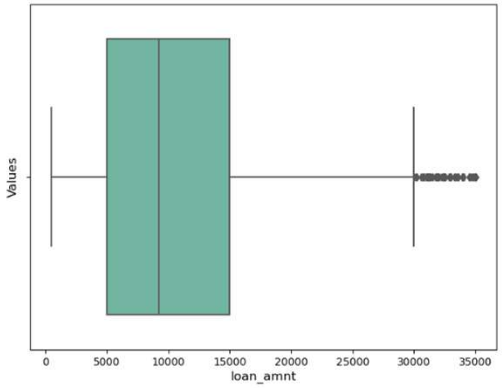
The emp_length column are converted to an integer with the following logic:

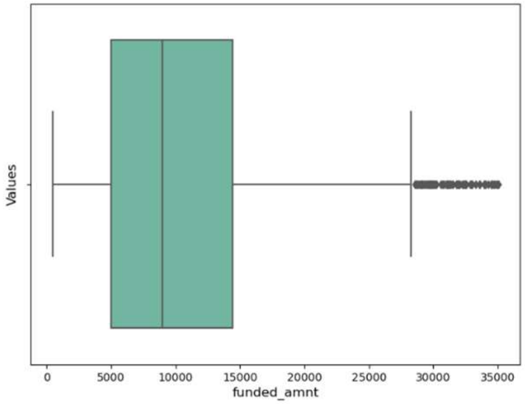
< 1 year: 0
1 year: 1
2 years: 2
3 years: 3
4 years: 4
5 years: 5
6 years: 6
7 years: 7
8 years: 8
9 years: 9
10+ years: 10

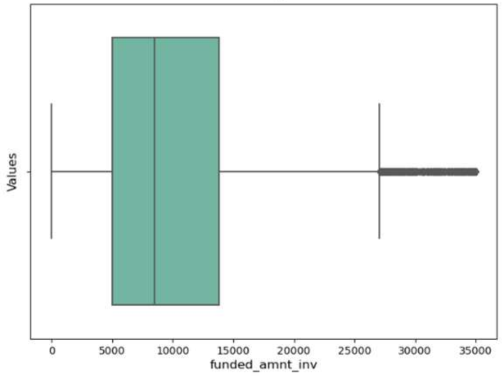
Understanding the outliers :

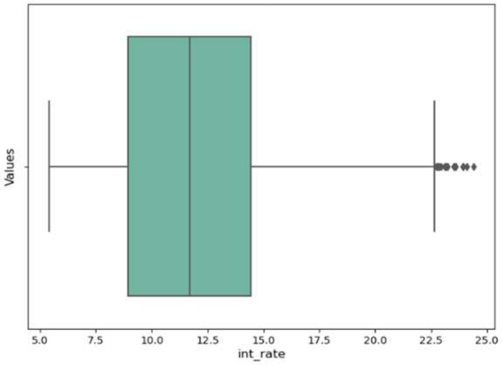
The analysis for outliers is performed on the mentioned columns: loan_amnt, funded_amnt, funded_amnt_inv, int_rate, installment, annual_inc, dti.

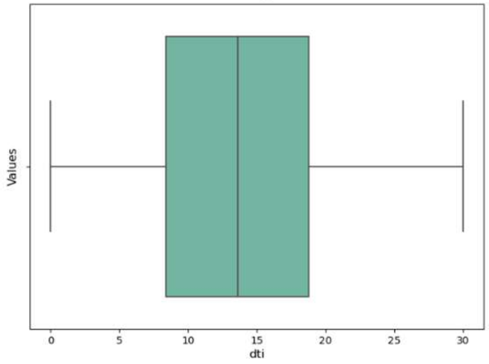












Conclusions from above plots :

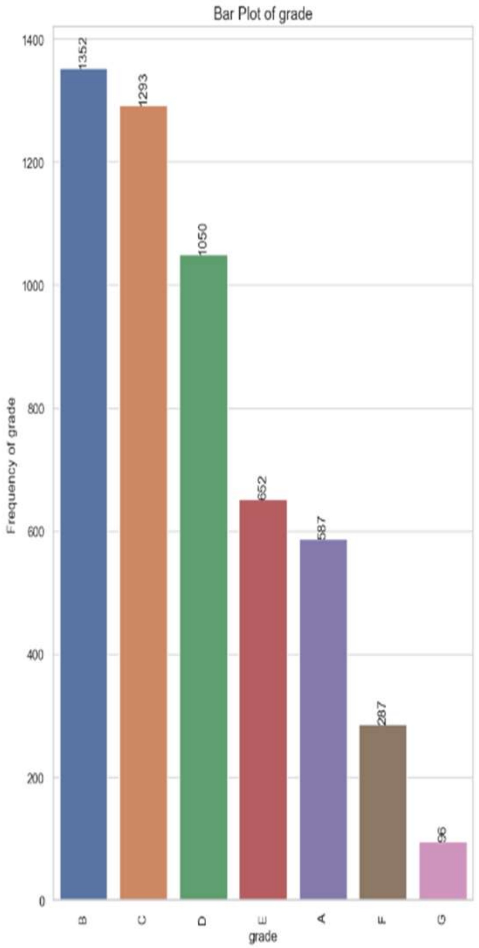
- The annual income of most of the loan applicants is between 40, 000 - 75, 000 USD.
- The loan amount of most of the loan applicants is between 5, 000 - 15, 000.
- The funded amount of most of the loan applicants is between 5, 000 - 14, 000 USD.
- The funded amount by investor for most of the loan applicants is between 5, 000 - 14, 000 USD.

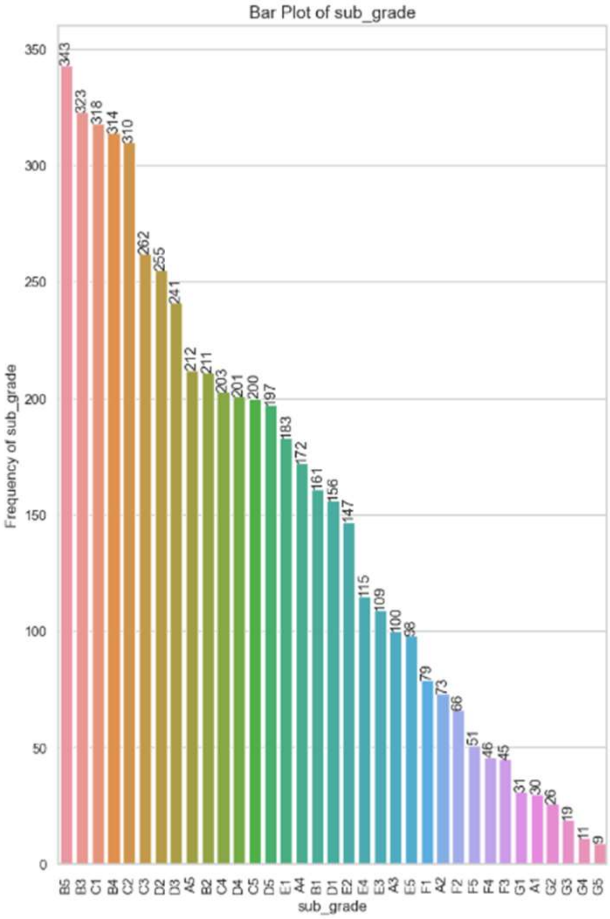
Univariate analysis:

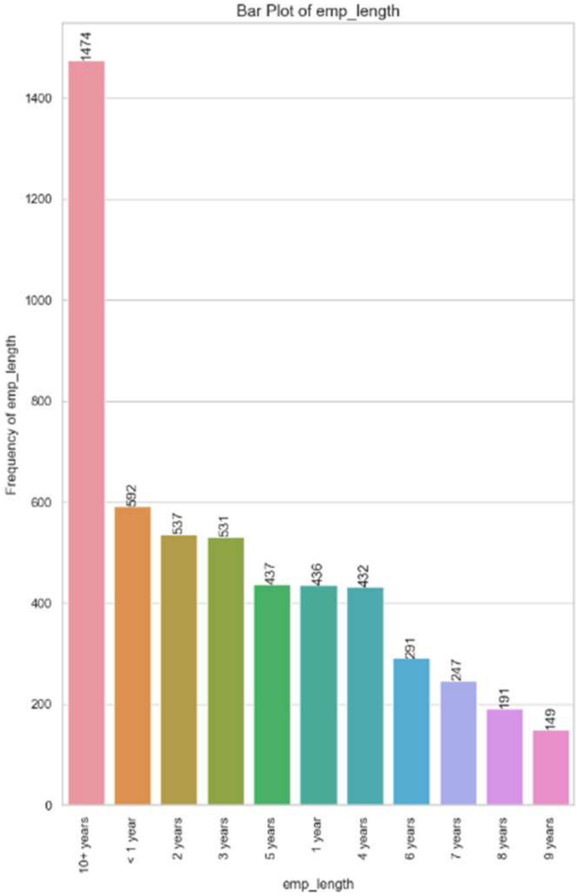
Univariate analysis is a statistical method used to analyze and summarize data sets consisting of one variable. It deals with the analysis of a single variable, rather than multiple variables, to understand its distribution, central tendency and dispersion.

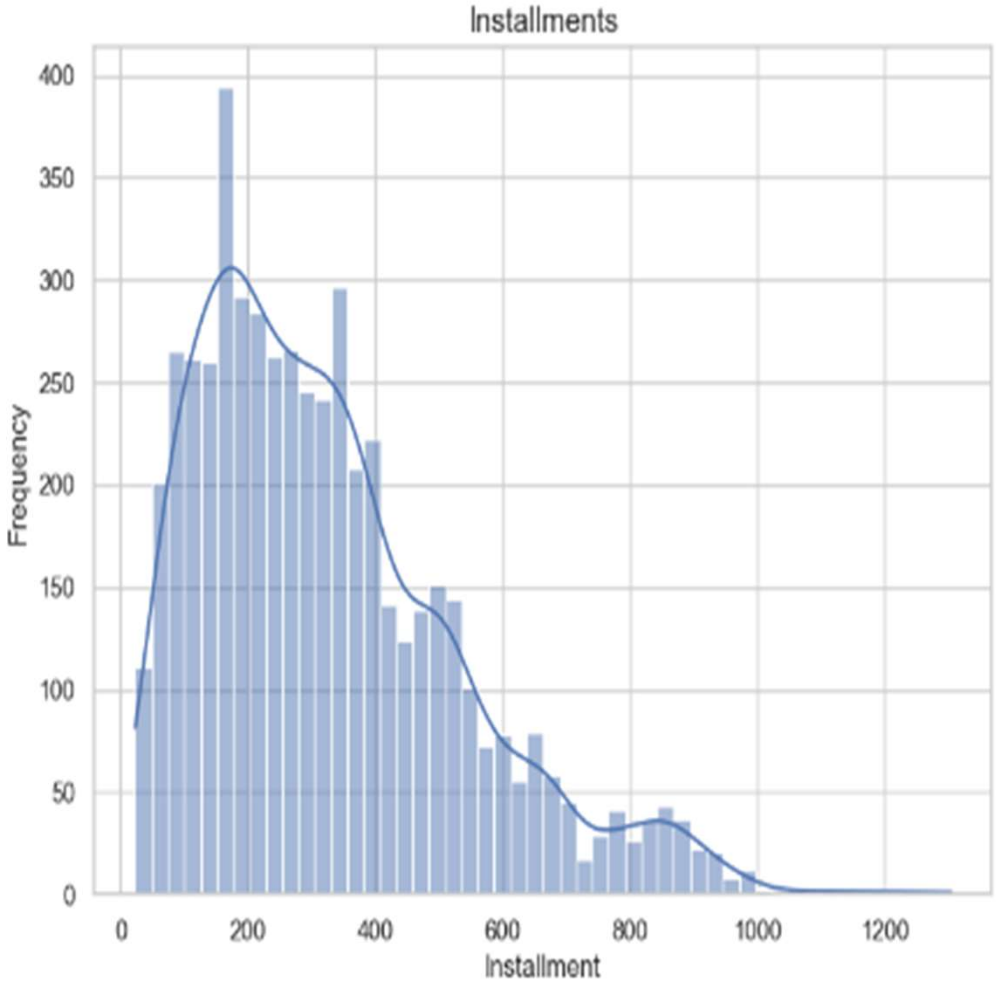
It was carried out for both Categorical and Quantitative Variables.

Quantitative	Categorical
‘annual_inc', 'dti', 'funded_amnt', 'funded_amnt_inv', 'installment', 'int_rate', 'issue_d', 'loan_amnt', 'pub_rec_bankruptcies', 'term', 'issue_m', 'issue_y'	'addr_state', 'emp_length', 'grade', 'home_ownership', 'loan_status', 'purpose', 'sub_grade', 'verification_status', 'issue_q', 'loan_paid', 'loan_amnt_bucket', 'int_rate_bucket', 'annual_inc_bucket', 'dti_bucket'


Bar Plot of grade
Frequency of grade
grade
0
200
400
600
800
1000
1200
1400
B 1352
C 1293
D 1050
E 652
A 587
F 287
G 96


Bar Plot of sub_grade
Frequency of sub_grade
sub_grade
0
50
100
150
200
250
300
350
B5 343
B3 323
C1 318
B4 314
C2 310
C3 262
D2 255
D3 241
A5 212
B2 211
C4 203
D4 201
C5 200
D5 197
E1 183
A4 172
B1 161
D1 156
E2 147
E4 115
E3 109
A3 100
E5 98
F1 79
A2 73
F2 66
F5 51
F4 46
F3 45
G1 31
A1 30
G2 26
G3 19
G4 11
G5 9


Bar Plot of emp_length
Frequency of emp_length
emp_length
0
200
400
600
800
1000
1200
1400
10+ years 1474
< 1 year 592
2 years 537
3 years 531
5 years 437
1 year 436
4 years 432
6 years 291
7 years 247
8 years 191
9 years 149

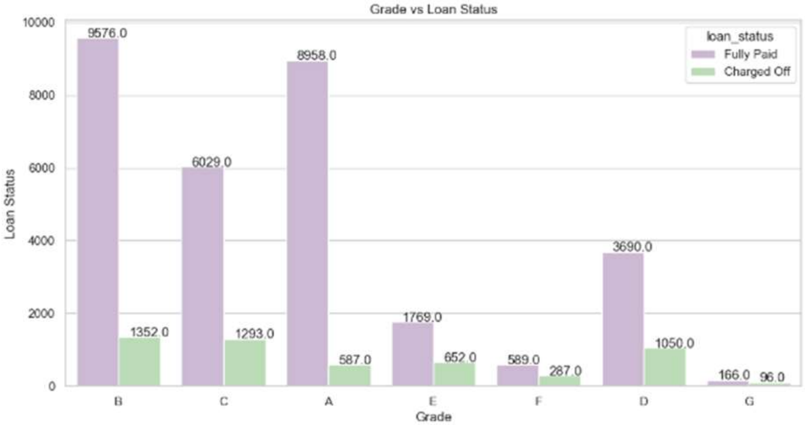


Highlights from univariate analysis :

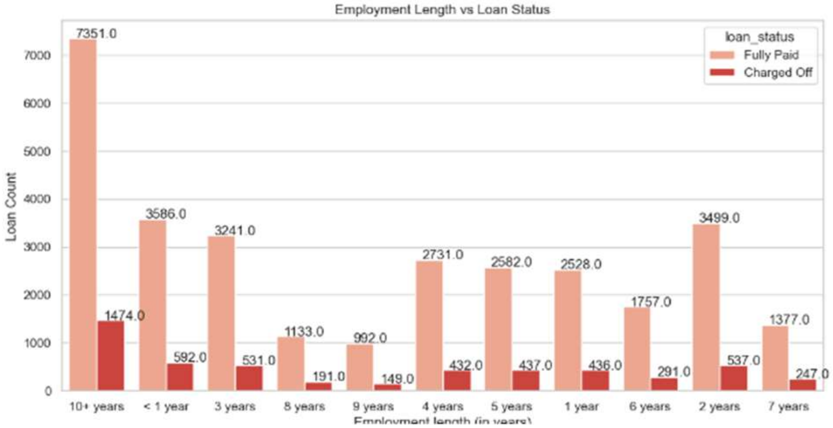
- Applicants who had been employed for more than 10 years are for the highest number of "Charged off" loans, totaling 1,474.
- This indicates that long-term employment history did not necessarily guarantee successful loan repayment.
- The majority of "Charged off" loan participants, totaling 2,715 individuals, 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.
- Among loan participants who charged off, 1,178 loan applicants had very high debt-to-income ratios.

Bivariate Analysis :

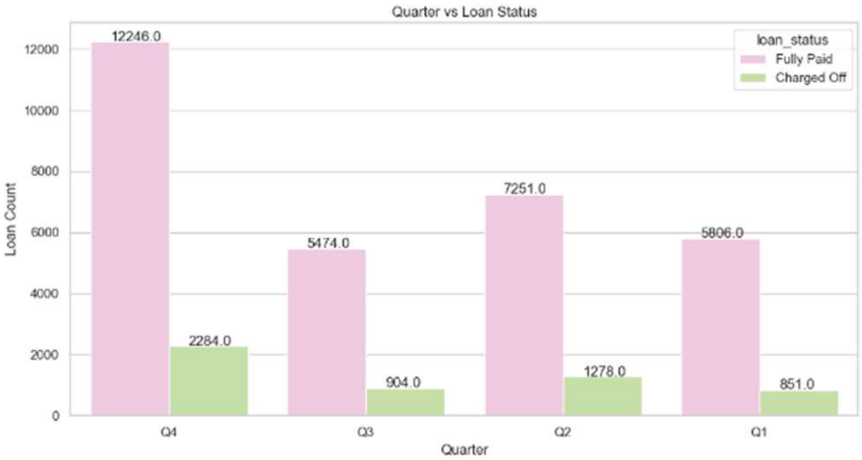
Bivariate analysis is a statistical method that involves the simultaneous analysis of two variables (factors).
It aims to determine the empirical relationship between them. The analysis can be used to test hypotheses, identify patterns, or explore relationships between the variables. It was carried out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



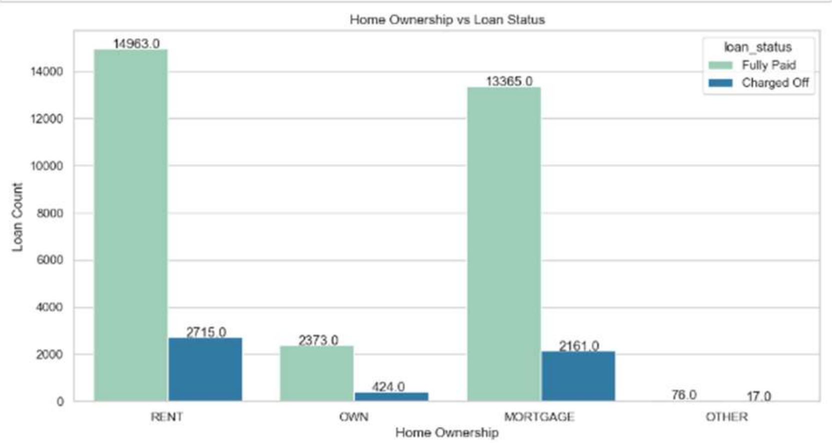
As per the above graph we can see ,the loan applicants belonging to Grades B, C and D contribute to most number of "Charged Off" loans



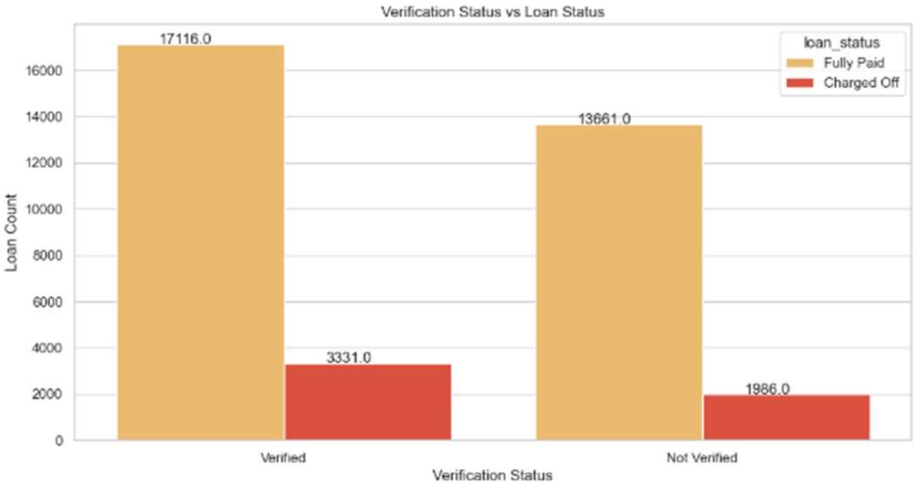
As per the observation above ,most number of loan applicants are 10 or more years of experience. They also are the ones who are most likely to default.



As per the observations , Q4 is the most preferred quarter for taking loans. This is mainly due to the holiday season coming up.



The loan applicants who live in a rented or mortgaged house are more likely to default



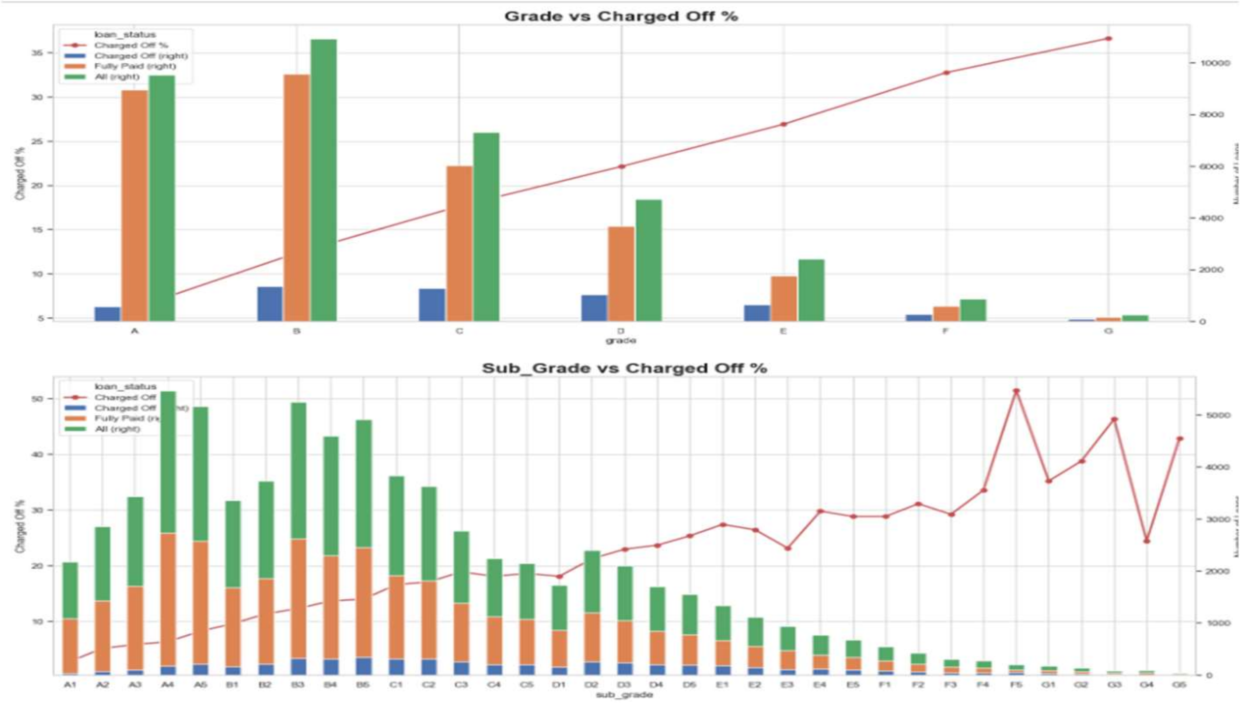
The loan applicants who have been verified are defaulting more than the applicants who are not verified

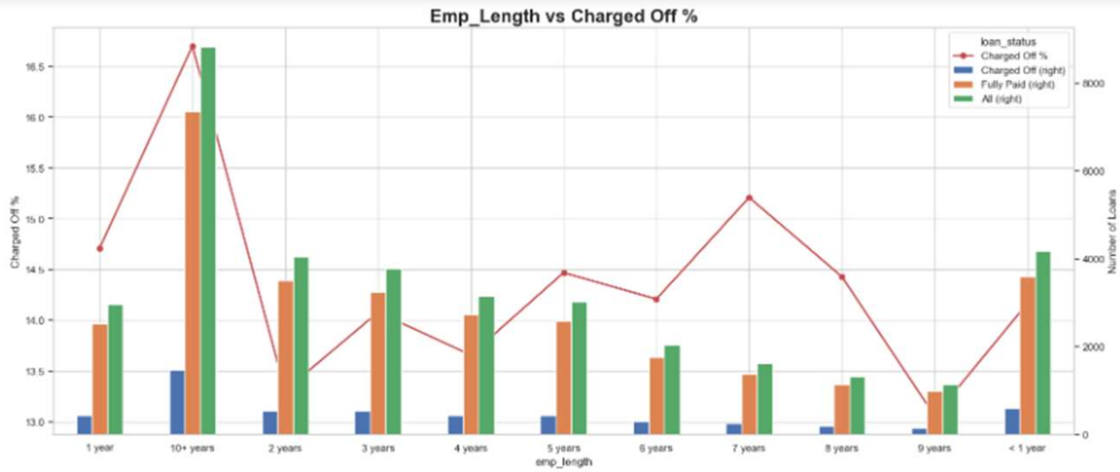
Highlights from bivariate analysis :

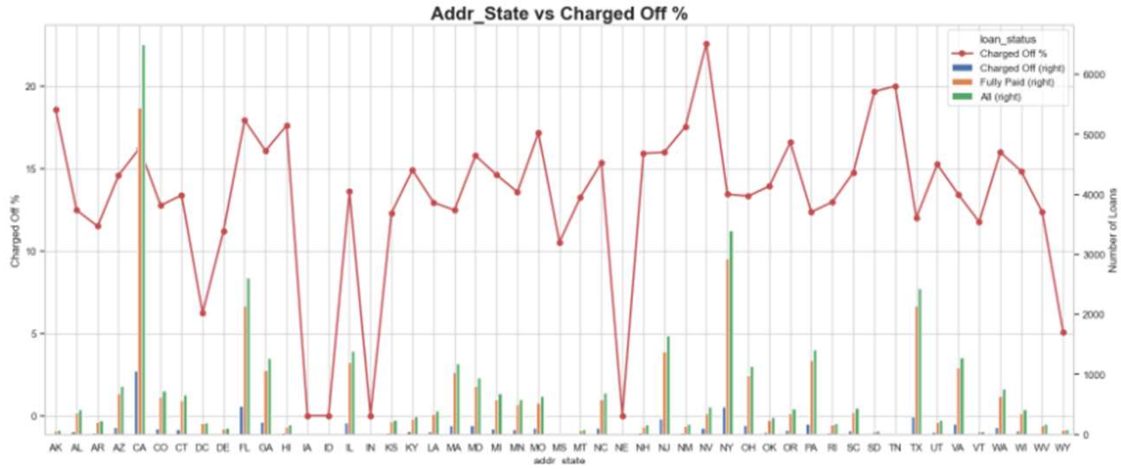
- The loan applicants belonging to Grades B, C, and D contribute to most of the "Charged Off" loans.
- The fourth quarter (Q4) is the most preferred quarter for taking loans, primarily because of the upcoming holiday season.
- Loan applicants who live in rented or mortgaged houses are more likely to default.
- Verified loan applicants are defaulting more than those who are not verified.

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.







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

Key-takeaways:

- **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.

- **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.