Lending Club Case Study

Submitted By:

Udit Shinghal

Utkarsh Sanwal

Contents

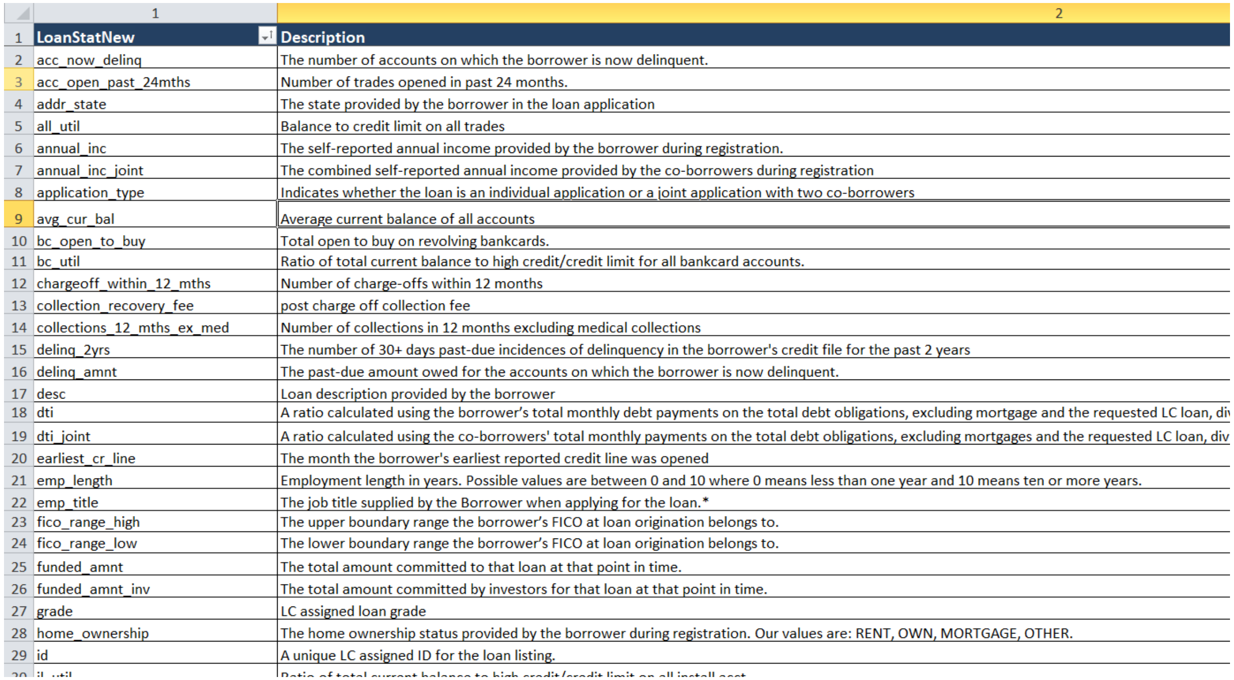
- Problem Statement
- Data Description
- Data Understanding
- Data Cleaning & Pre-processing
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Correlation Analysis

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.
- 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 39717records 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.



	1	2
1	LoanStatNew	Description
2	acc_now_delinq	The number of accounts on which the borrower is now delinquent.
3	acc_open_past_24mths	Number of trades opened in past 24 months.
4	addr_state	The state provided by the borrower in the loan application
5	all_util	Balance to credit limit on all trades
6	annual_inc	The self-reported annual income provided by the borrower during registration.
7	annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
8	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
9	avg_cur_bal	Average current balance of all accounts
10	bc_open_to_buy	Total open to buy on revolving bankcards.
11	bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
12	chargeoff_within_12_mths	Number of charge-offs within 12 months
13	collection_recovery_fee	post charge off collection fee
14	collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
15	delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
16	delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
17	desc	Loan description provided by the borrower
18	dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, di
19	dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, div
20	earliest_cr_line	The month the borrower's earliest reported credit line was opened
21	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.
22	emp_title	The job title supplied by the Borrower when applying for the loan.*
23	fico_range_high	The upper boundary range the borrower’s FICO at loan origination belongs to.
24	fico_range_low	The lower boundary range the borrower’s FICO at loan origination belongs to.
25	funded_amnt	The total amount committed to that loan at that point in time.
26	funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
27	grade	LC assigned loan grade
28	home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
29	id	A unique LC assigned ID for the loan listing.
30	il_util	Ratio of total current balance to high credit/credit limit on all install acct

Data Cleaning & Pre-processing

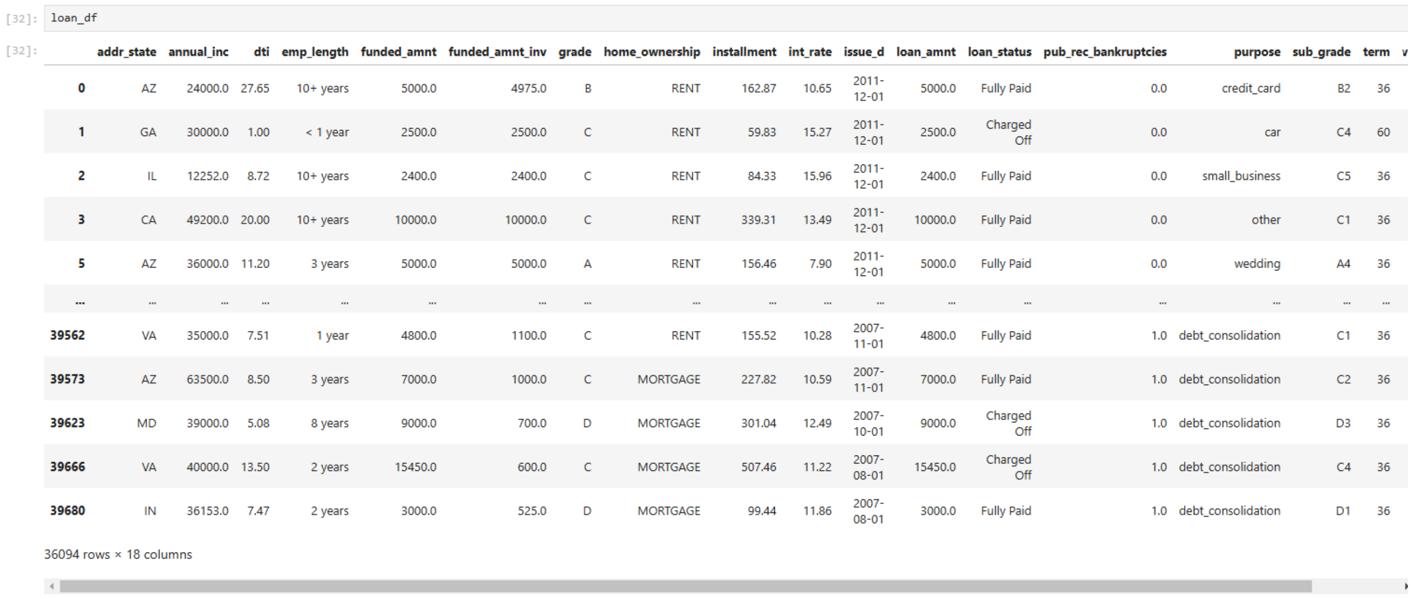
- Loading data from loan CSV
- Checking for null values in the dataset
- Checking for unique values
- Checking for duplicated rows in data
- Dropping Records & Columns
- Common Functions
- Data Conversion
- Outlier Treatment
- Imputing values in Columns

Data Cleaning and Pre-processing

- Importing Required Libraries
- **Checking and removing Null Values** : 48% of columns were dropped
- **Checking for Unique Values**: Unique Values Does provide any help in data analysis, these values were dropped, a total of 9 columns had null values.
- **Checking for Duplicate Rows:** No Duplicate Rows were found
- **Dropping Records and columns:**
    - Dropped Records with Loan status as "Current"
    - Dropped columns with missing Data as >=65% these columns with pollute our data
    - Dropping columns which does not help our analysis like : Member_id, zip_code, 21 such columns were removed
- **Common Functions:** Common functions were created for repeating common operations like plotting bar graphs, box plots, histograms, countplots, binning etc.
- **Outlier Treatment:** Calculated and filtering out the outliers outside of lower and upper bound.
- **Imputing values in Columns: Replaced missing values of annual_inc with the corresponding mode value of annual_inc of the emp_length annual_inc field:** They Employment length has **1015** missing values, which means either they are **not employed or self-employed (business owners).** Considering they have a decent average annual income, we have assumed that these are business owners and we have added their employment duration with the mode value of **emp_length** which is **10+ years**.
    - Mapped employment length with the respective number of years in int.
    - Imputed **NONE** values as **OTHER** for **home_ownership.**
    - Replaced the **'Source Verified'** values as **'Verified'** since both values mean the same thing i.e. the loan applicant has some source of income which is verified.
    - There are **660 null values**

Post Data cleaning and Pre-processing of dataset, we were left with **36094 rows × 18 columns.**

Clean Data



```
[32]: loan_df
```

[32]:

	addr_state	annual_inc	dti	emp_length	funded_amnt	funded_amnt_inv	grade	home_ownership	installment	int_rate	issue_d	loan_amnt	loan_status	pub_rec_bankruptcies	purpose	sub_grade	term	v
0	AZ	24000.0	27.65	10+ years	5000.0	4975.0	B	RENT	162.87	10.65	2011-12-01	5000.0	Fully Paid	0.0	credit_card	B2	36	
1	GA	30000.0	1.00	< 1 year	2500.0	2500.0	C	RENT	59.83	15.27	2011-12-01	2500.0	Charged Off	0.0	car	C4	60	
2	IL	12252.0	8.72	10+ years	2400.0	2400.0	C	RENT	84.33	15.96	2011-12-01	2400.0	Fully Paid	0.0	small_business	C5	36	
3	CA	49200.0	20.00	10+ years	10000.0	10000.0	C	RENT	339.31	13.49	2011-12-01	10000.0	Fully Paid	0.0	other	C1	36	
5	AZ	36000.0	11.20	3 years	5000.0	5000.0	A	RENT	156.46	7.90	2011-12-01	5000.0	Fully Paid	0.0	wedding	A4	36	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
39562	VA	35000.0	7.51	1 year	4800.0	1100.0	C	RENT	155.52	10.28	2007-11-01	4800.0	Fully Paid	1.0	debt_consolidation	C1	36	
39573	AZ	63500.0	8.50	3 years	7000.0	1000.0	C	MORTGAGE	227.82	10.59	2007-11-01	7000.0	Fully Paid	1.0	debt_consolidation	C2	36	
39623	MD	39000.0	5.08	8 years	9000.0	700.0	D	MORTGAGE	301.04	12.49	2007-10-01	9000.0	Charged Off	1.0	debt_consolidation	D3	36	
39666	VA	40000.0	13.50	2 years	15450.0	600.0	C	MORTGAGE	507.46	11.22	2007-08-01	15450.0	Charged Off	1.0	debt_consolidation	C4	36	
39680	IN	36153.0	7.47	2 years	3000.0	525.0	D	MORTGAGE	99.44	11.86	2007-08-01	3000.0	Fully Paid	1.0	debt_consolidation	D1	36	

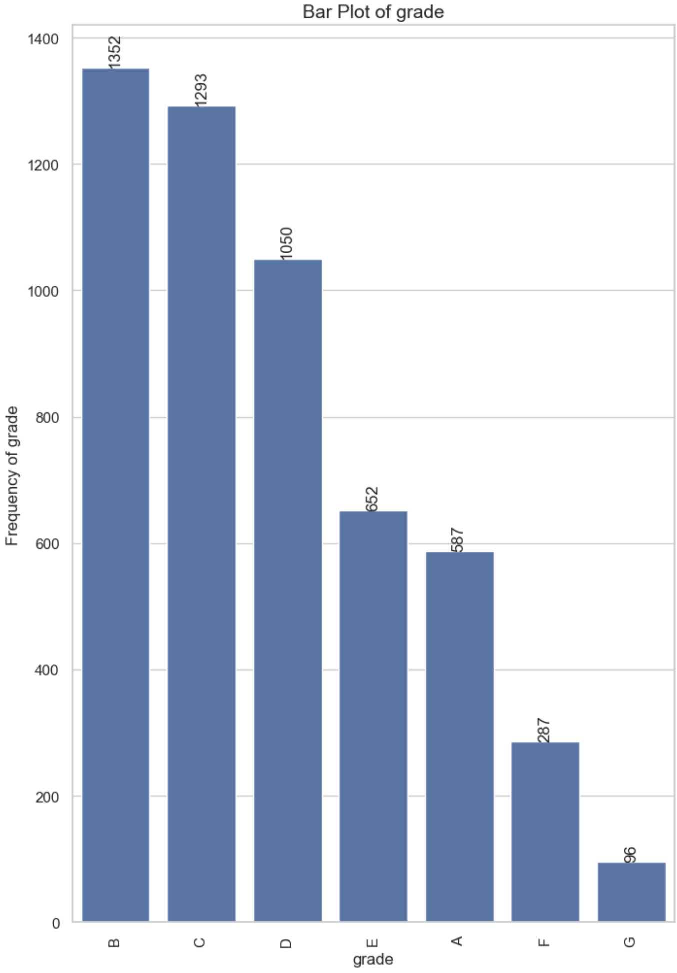
36094 rows × 18 columns

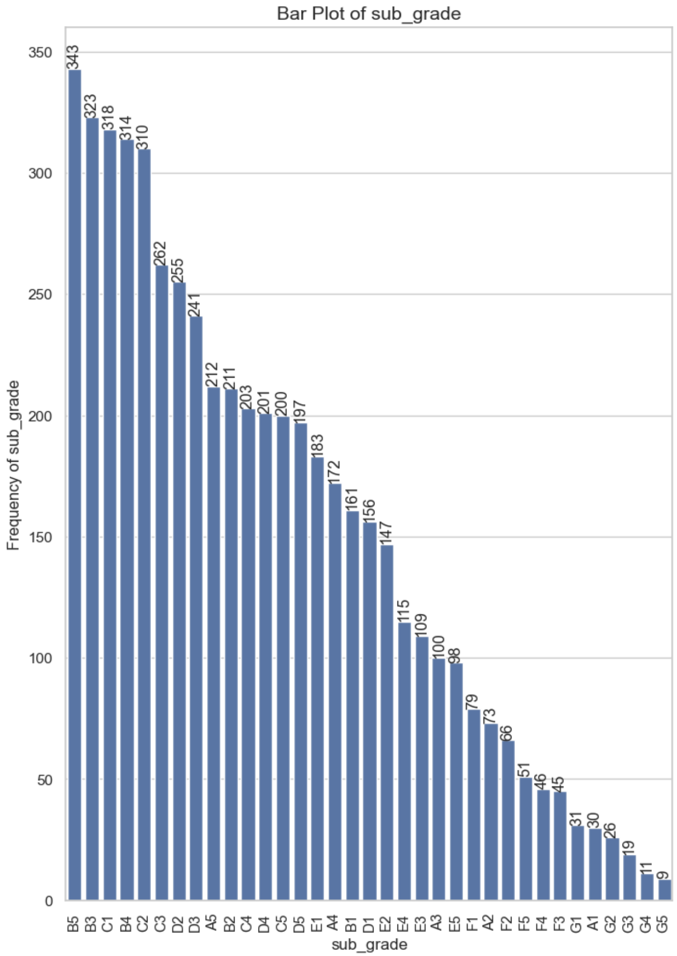
Univariate Analysis

- Univariate analysis is a statistical method used to analyze and summarize datasets 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

Univaraite Analysis (Unordered Categorically)

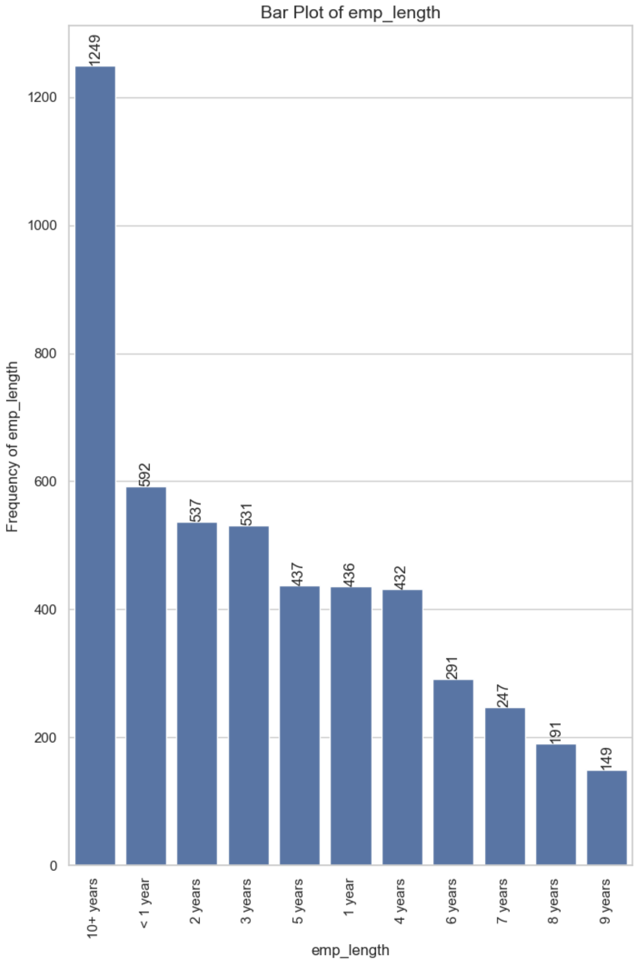
Grade & Sub Grade

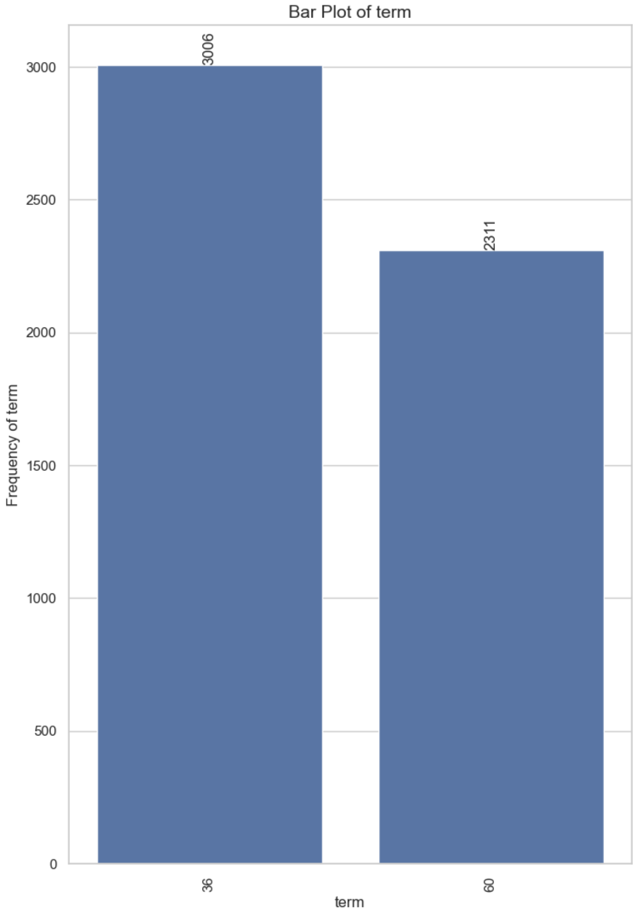




Univaraite Analysis (Unordered Categorically)

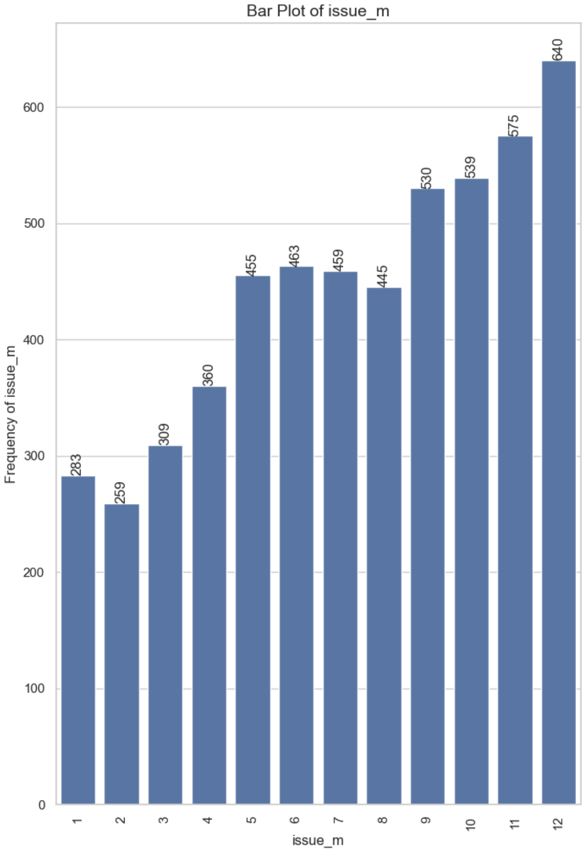
Bar Plot of Item & Bar Plot of Emp_length

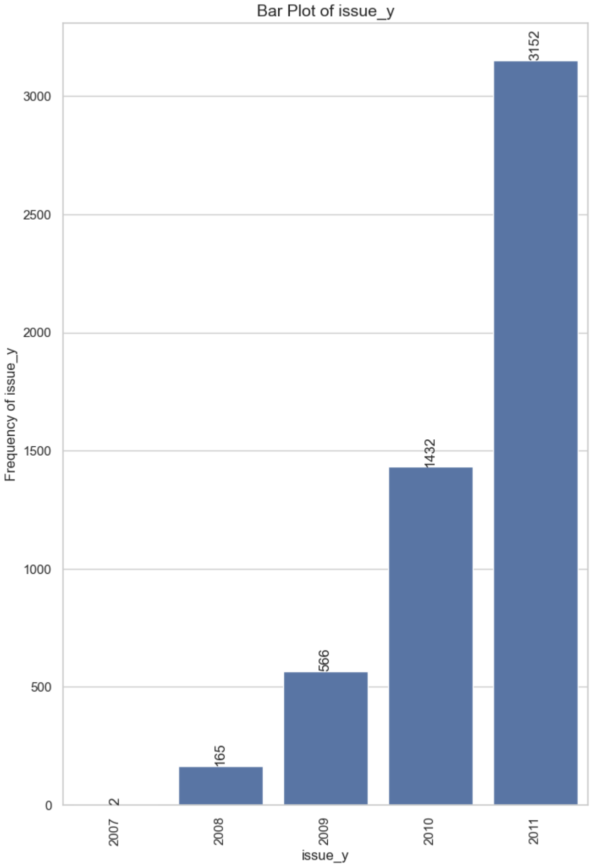




Univaraite Analysis (Unordered Categorically)

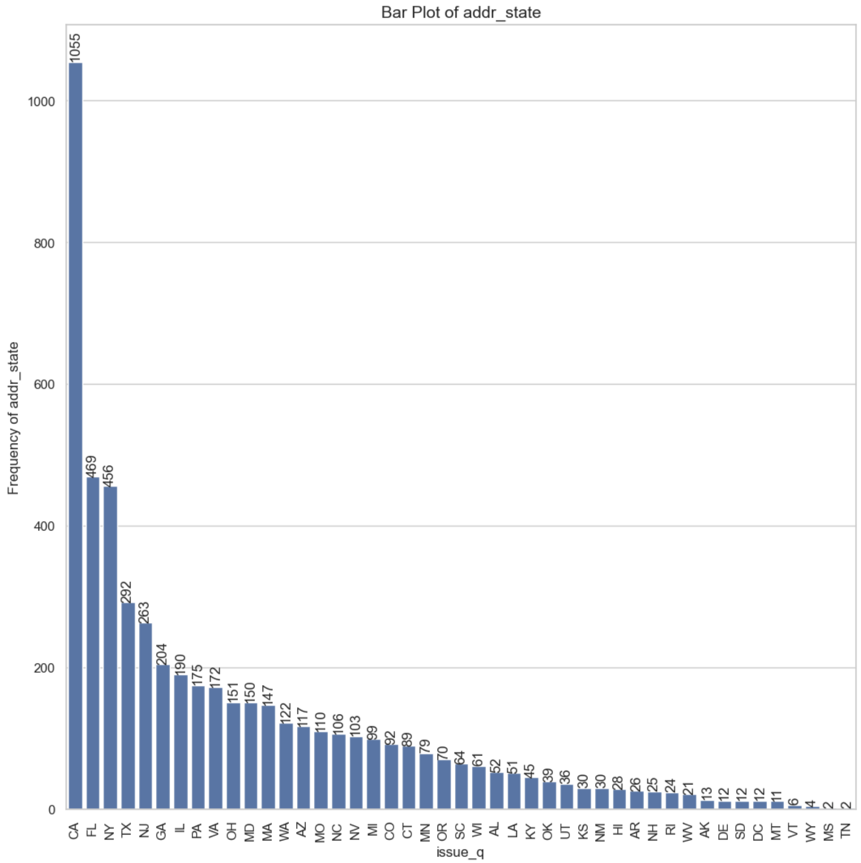
Term and Employee Length

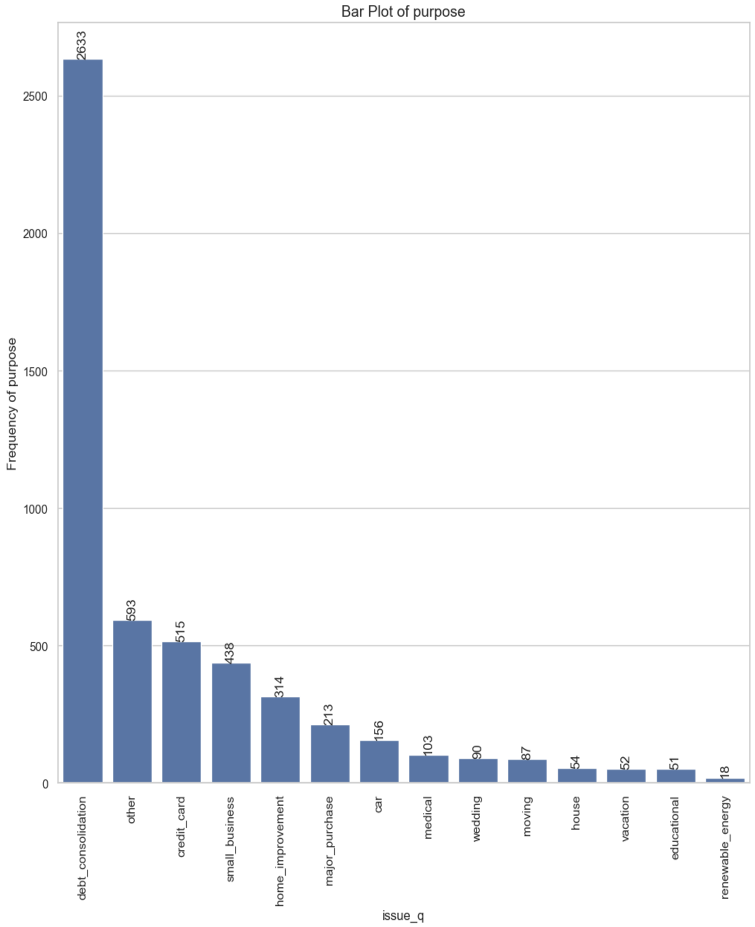




Univaraite Analysis (Unordered Categorically)

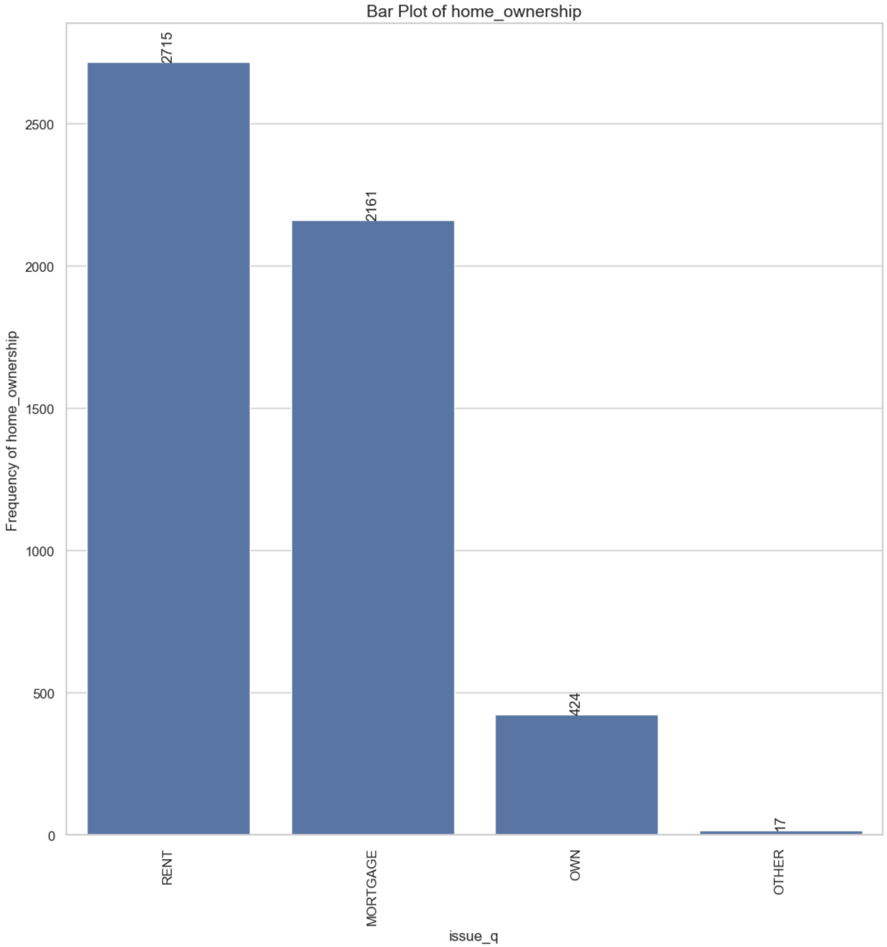
Address State and Purpose of Loan





Univaraite Analysis (Unordered Categorically)

Various Types of Home Ownership

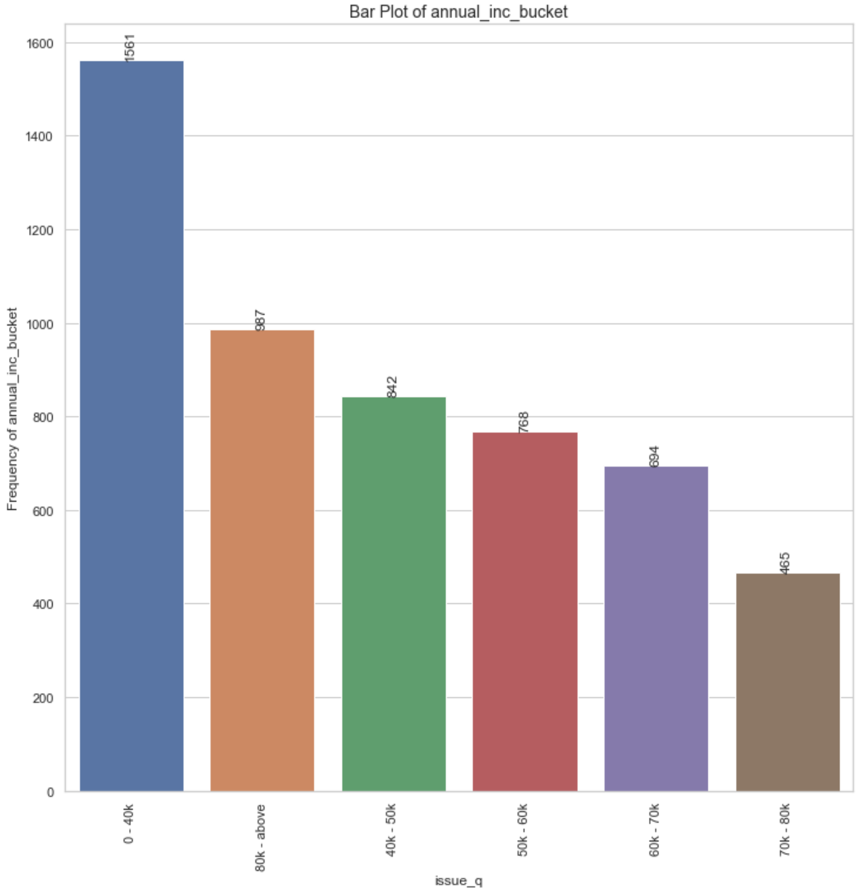


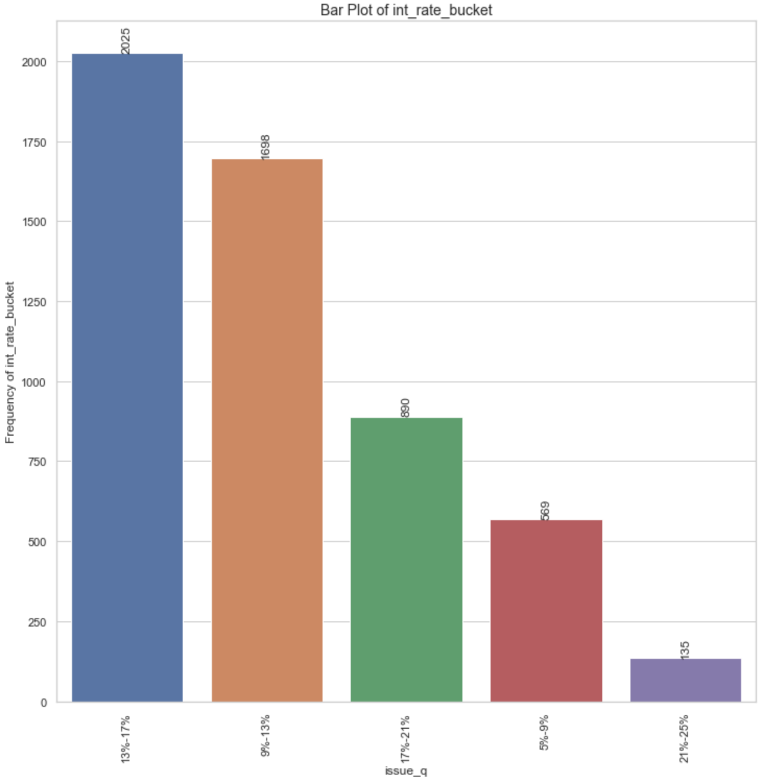
Univariate Analysis (Unordered and Ordered)

- **Observations:**
  - **Ordered Categorical Variable**
    - Grade B had the highest number of "Charged off" loan applicants, with a total of 1,352 applicants, indicating that applicants with this credit grade faced challenges in repaying their loans.
    - Short-term loans with a duration of 36 months were the most popular among "Charged off" applicants, with 3,006 applications. This suggests that a significant portion of applicants who experienced loan default chose shorter repayment terms.
    - Applicants who had been employed for more than 10 years accounted 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.
  - **Unordered Categorical Variable**
    - California had the highest number of "Charged off" loan applicants, with 1,055 applicants. For such applicants, the lending company needs to implement stricter eligibility criteria or credit assessments due to a higher number of "Charged off" applicants from this state.
    - Debt consolidation was the primary loan purpose for most "Charged off" loan applicants, with 2,633 applicants selecting this option. The lending company needs to exercise caution when approving loans for debt consolidation purposes, as it was the primary loan purpose for many "Charged off" applicants.
    - 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. The lending company should enhance risk assessment practices, including stricter credit checks and lower loan-to-value ratios, for applicants with a history of loan defaults. They should offer financial education and support services to help borrowers manage their finances and improve loan repayment outcomes.

Univaraite Analysis (Quantative Variable)

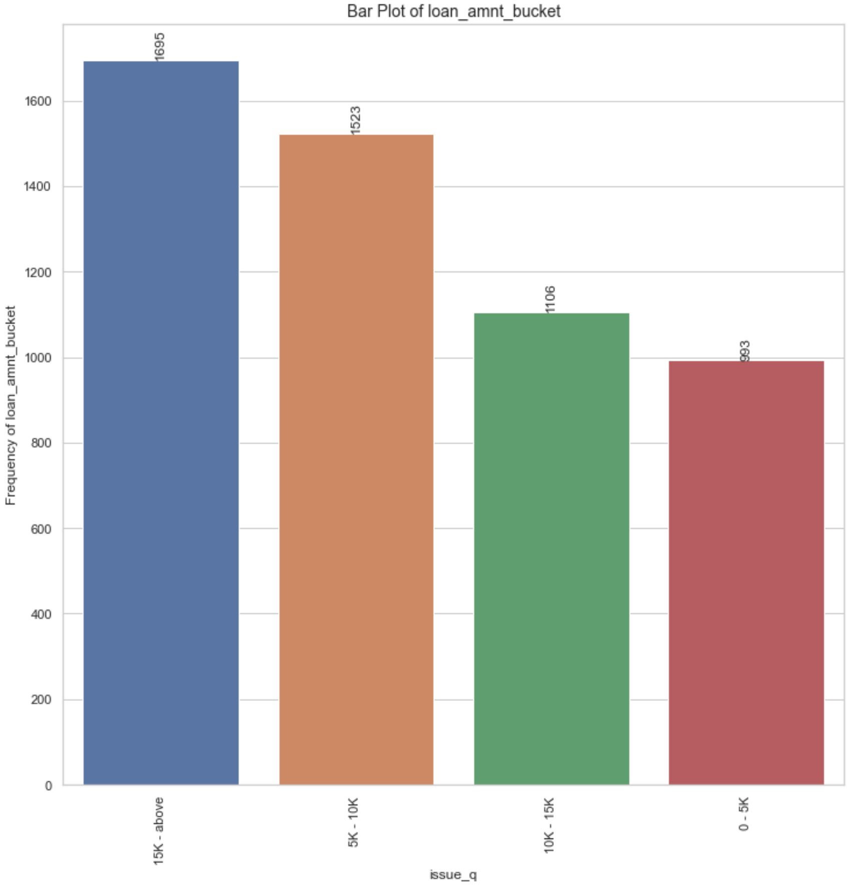
Buckets of Annual Income Status and Loan Interest Rates

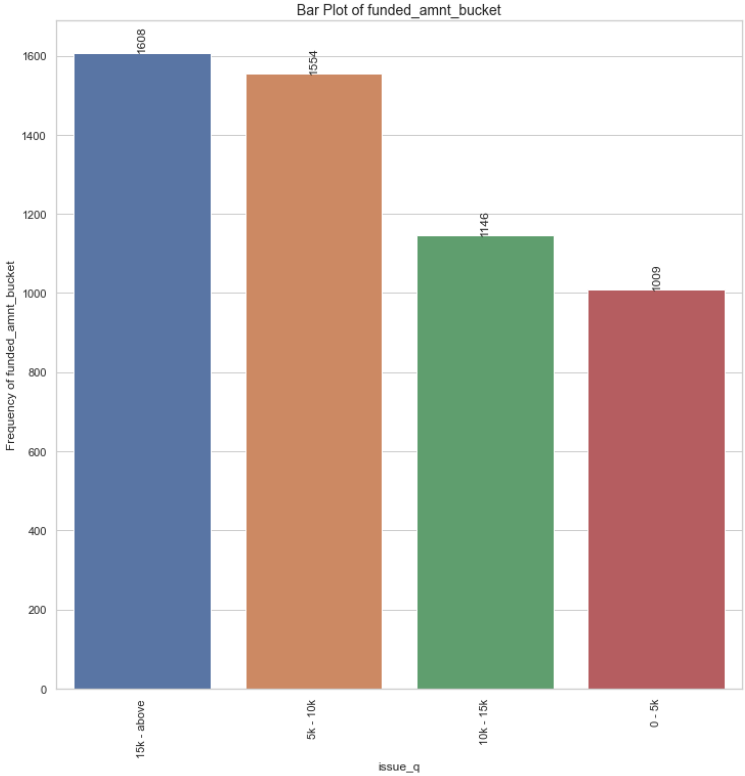




Univaraite Analysis (Quantative Variable)

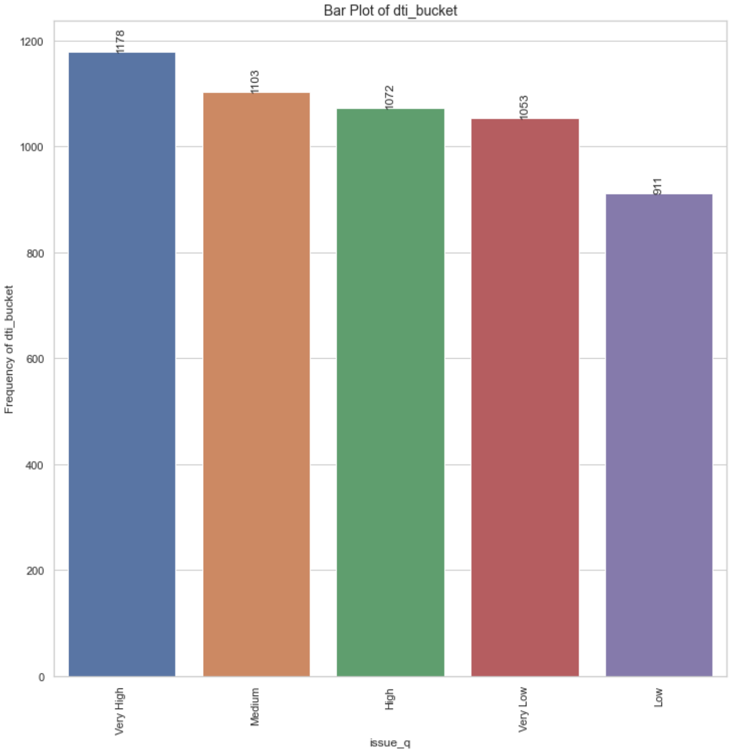
Buckets of Loan Amount and funded Amount

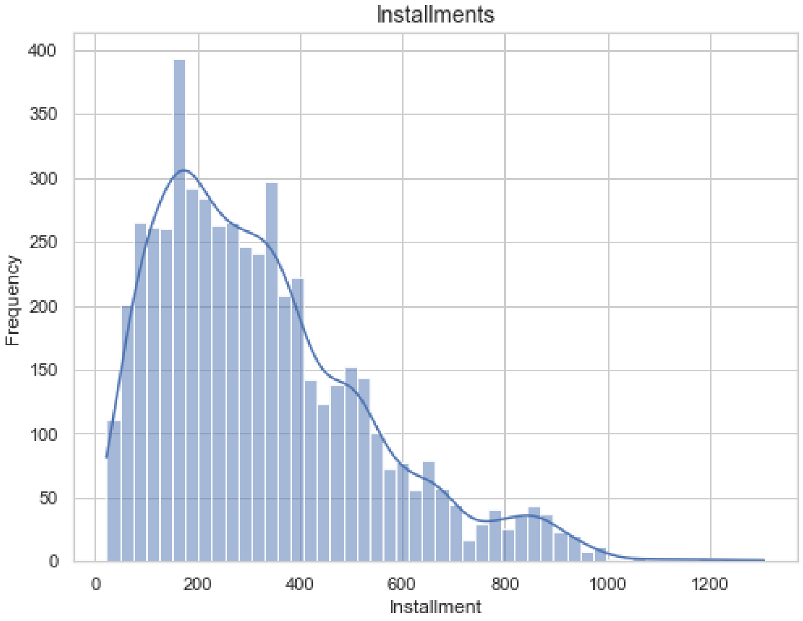




Univaraite Analysis (Quantative Variable)

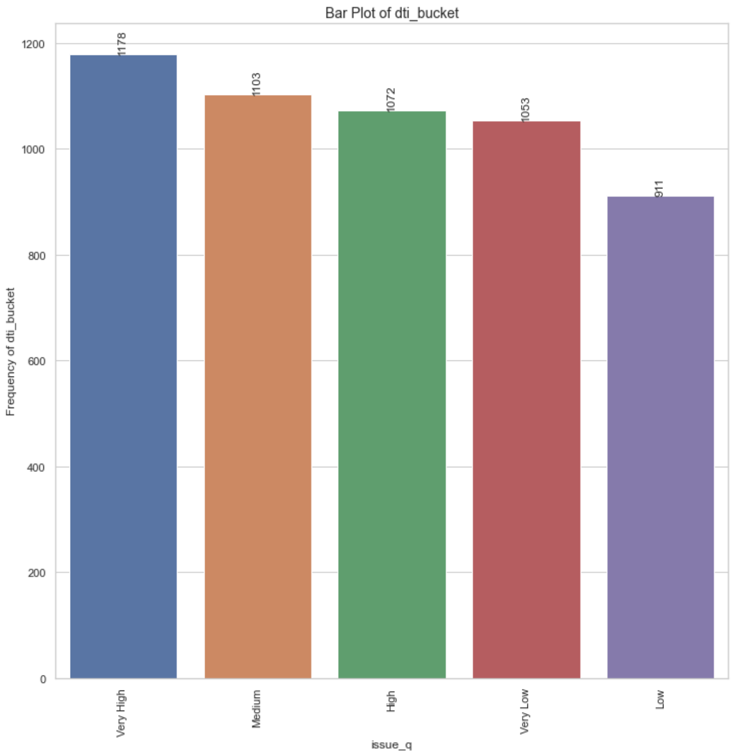
Bucket of DTI and Histogram of Installment

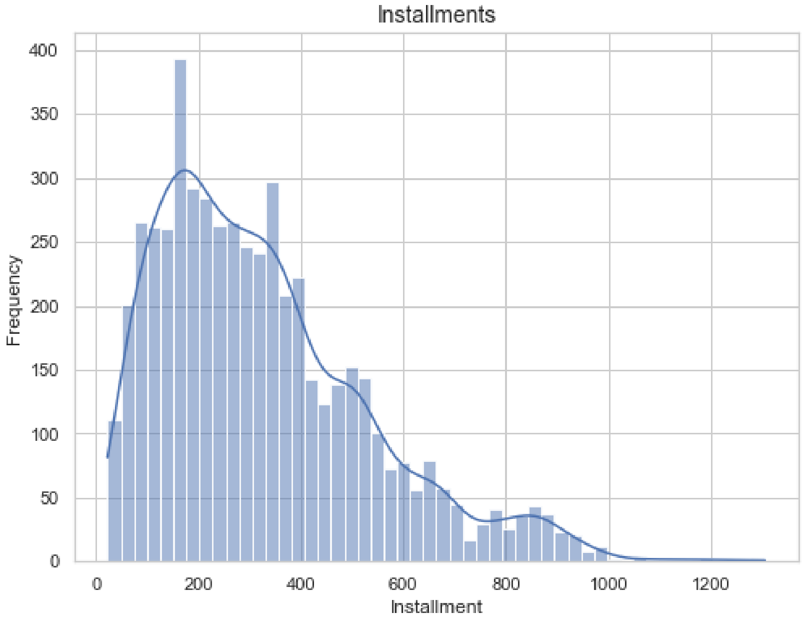




Univaraite Analysis (Quantative Variable)

Bucket of DTI and Histogram of Installment





Univaraite Analysis (Quantative Variable)

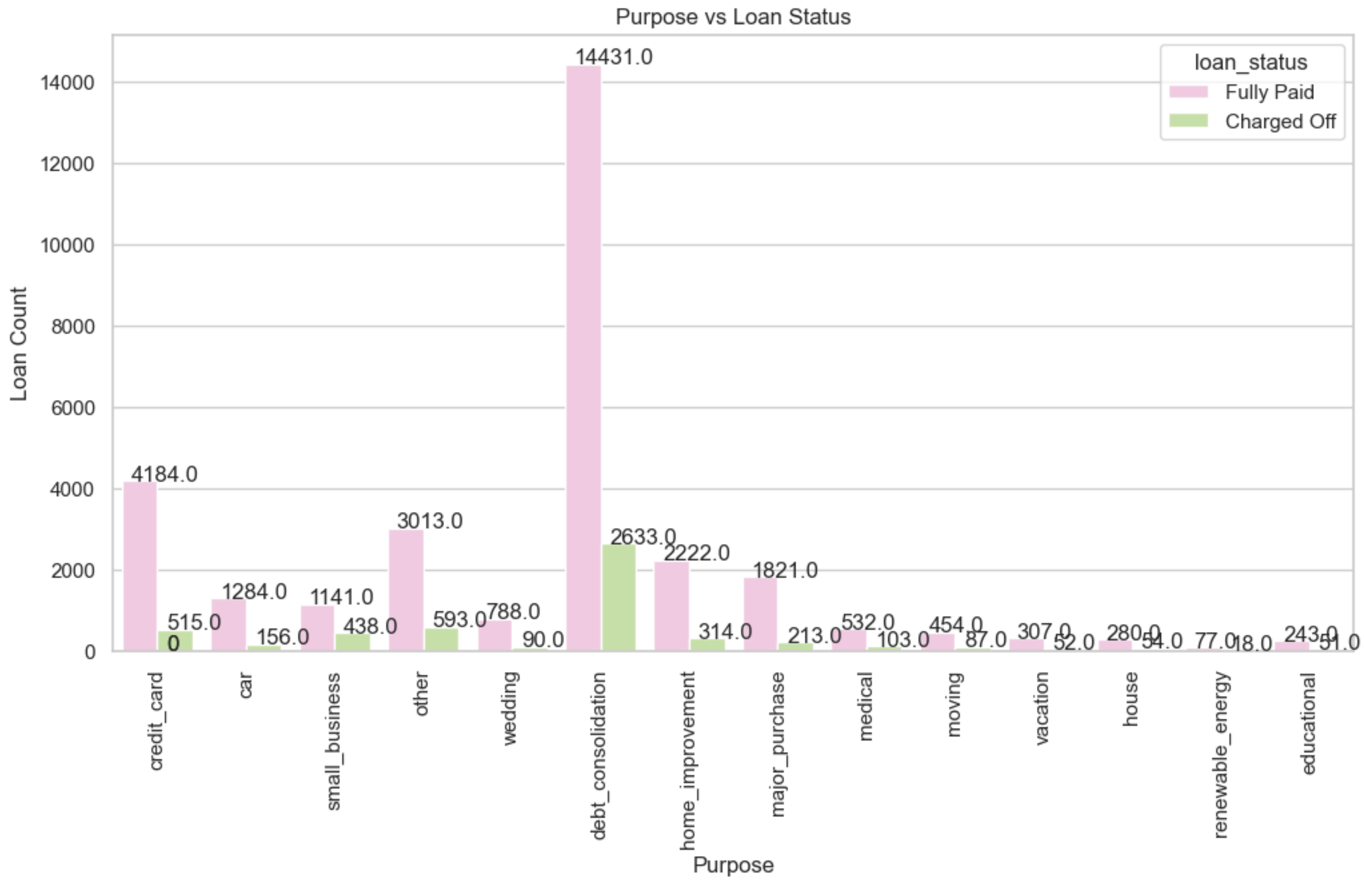
- **Observations:**
  - **Quantative Variable**
    - 1,561 loan applicants who charged off had annual salaries less than 40,000 USD. The lending company should exercise caution when lending to individuals with low annual salaries. They should implement rigorous income verification and assess repayment capacity more thoroughly for applicants in this income bracket.
    - Among loan participants who charged off (2,025), a considerable portion belonged to the interest rate bucket of 13%-17%. To reduce the risk of default, the lending company should consider offering loans at lower interest rates when possible.
    - 1,695 loan participants who charged off received loan amounts of 15,000 USD and above. The lending company should evaluate applicants seeking higher loan amounts carefully. They should ensure the applicants must have a strong credit history and repayment capability to handle larger loans.
    - 1,608 loan participants who charged off received funded amounts of 15,000 USD and above. The lending company should ensure that the funded amounts align with the borrower's financial capacity. They should conduct thorough credit assessments for larger loan requests.
    - Among loan participants who charged off, 1,178 loan applicants had very high debt-to-income ratios. The lending company should implement strict debt-to-income ratio requirements to prevent lending to individuals with unsustainable levels of debt relative to their income.

Bivaraite 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 Ordered Categorical Variable and Unordered Categorical Variable

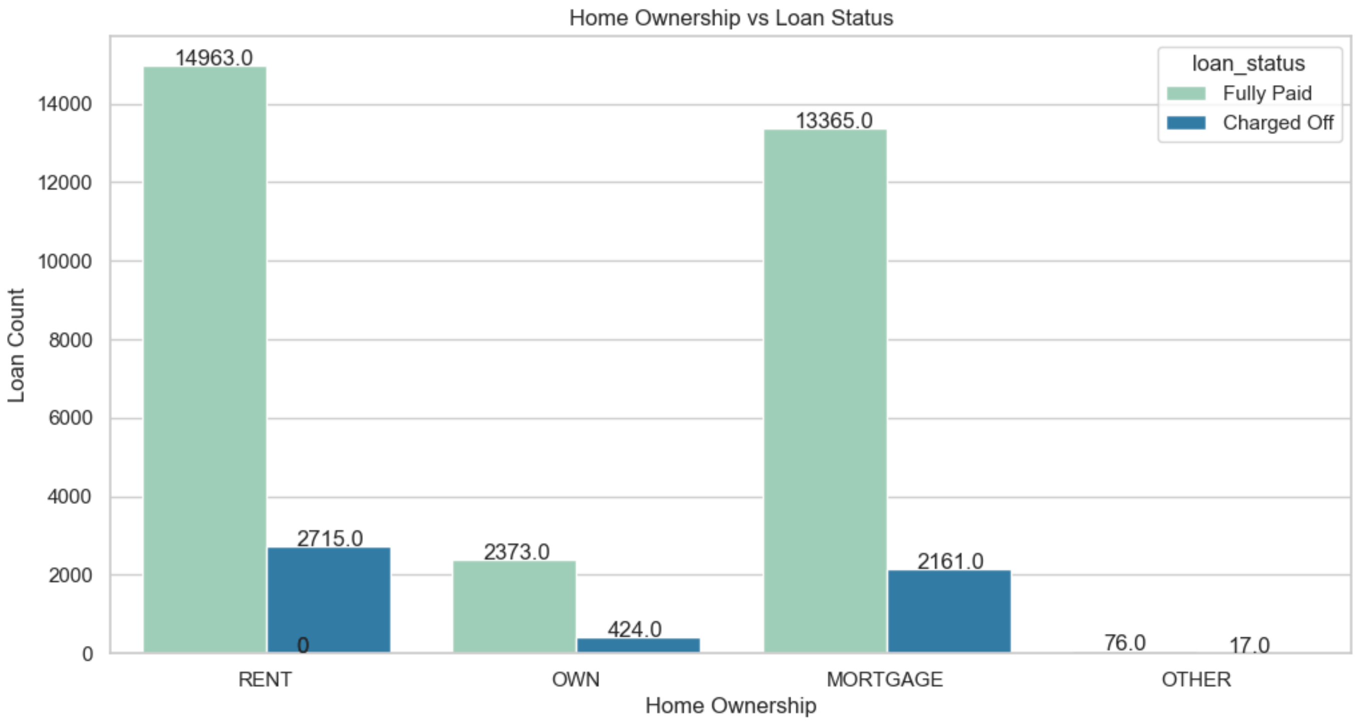
Bivaraite Analysis (Unordered Categorical)

Purpose of Loan vs Status of Loan



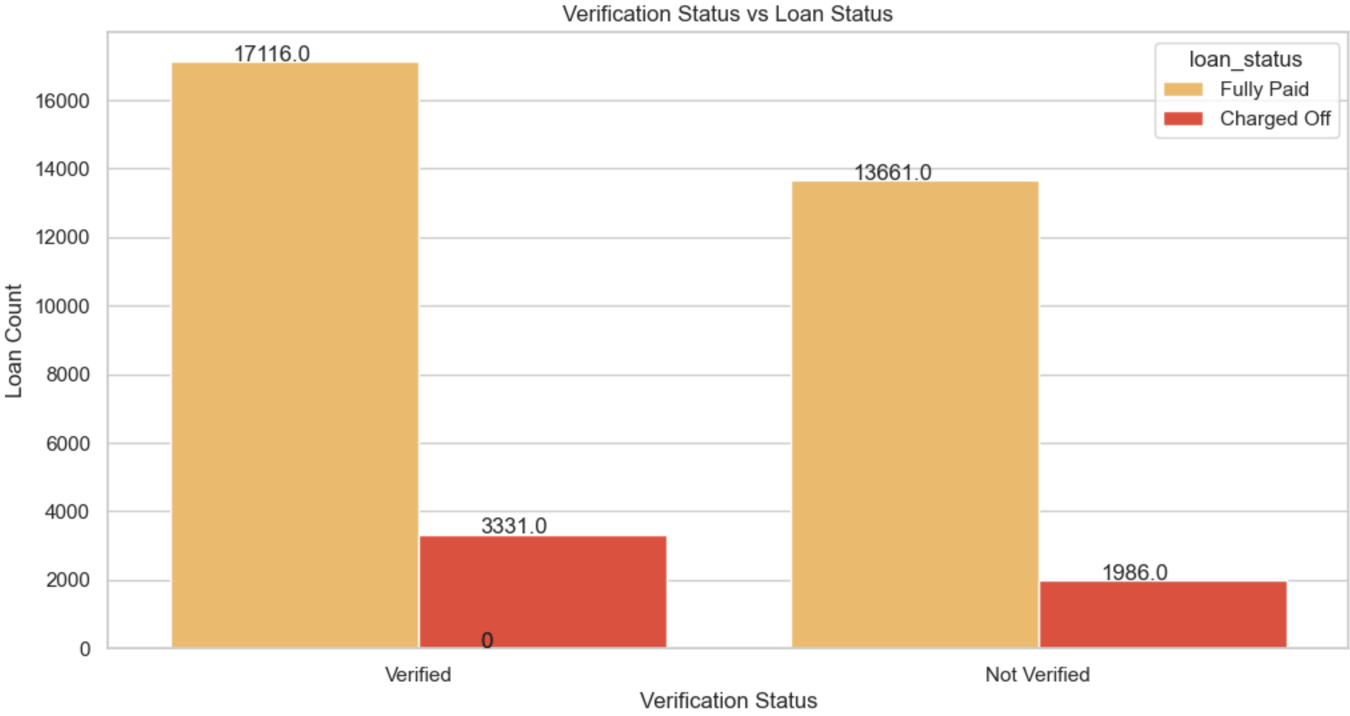
Bivaraite Analysis (Unordered Categorical)

Home Ownership vs Loan Status



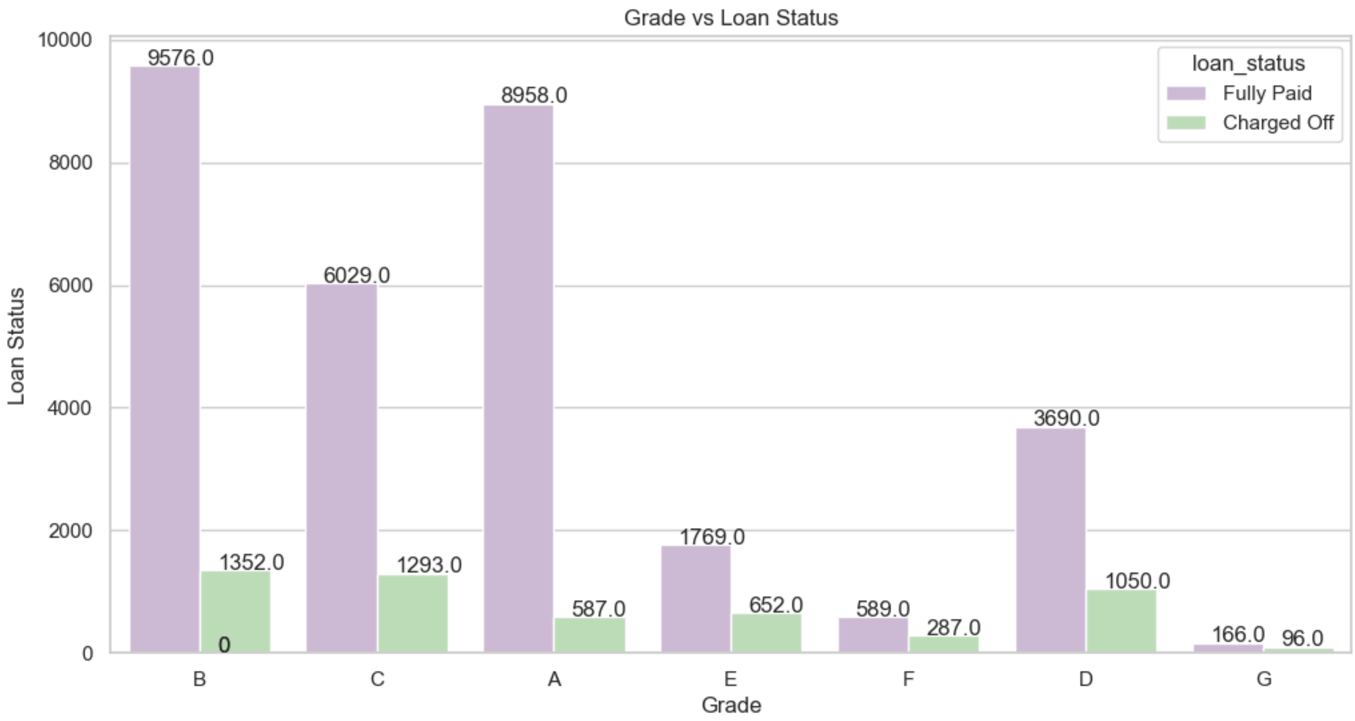
Bivaraite Analysis (ordered Categorical)

Verification Status



Bivaraite Analysis (ordered Categorical)

Loan Grade vs Status of Loan



Bivariate Analysis (Unordered and Ordered)

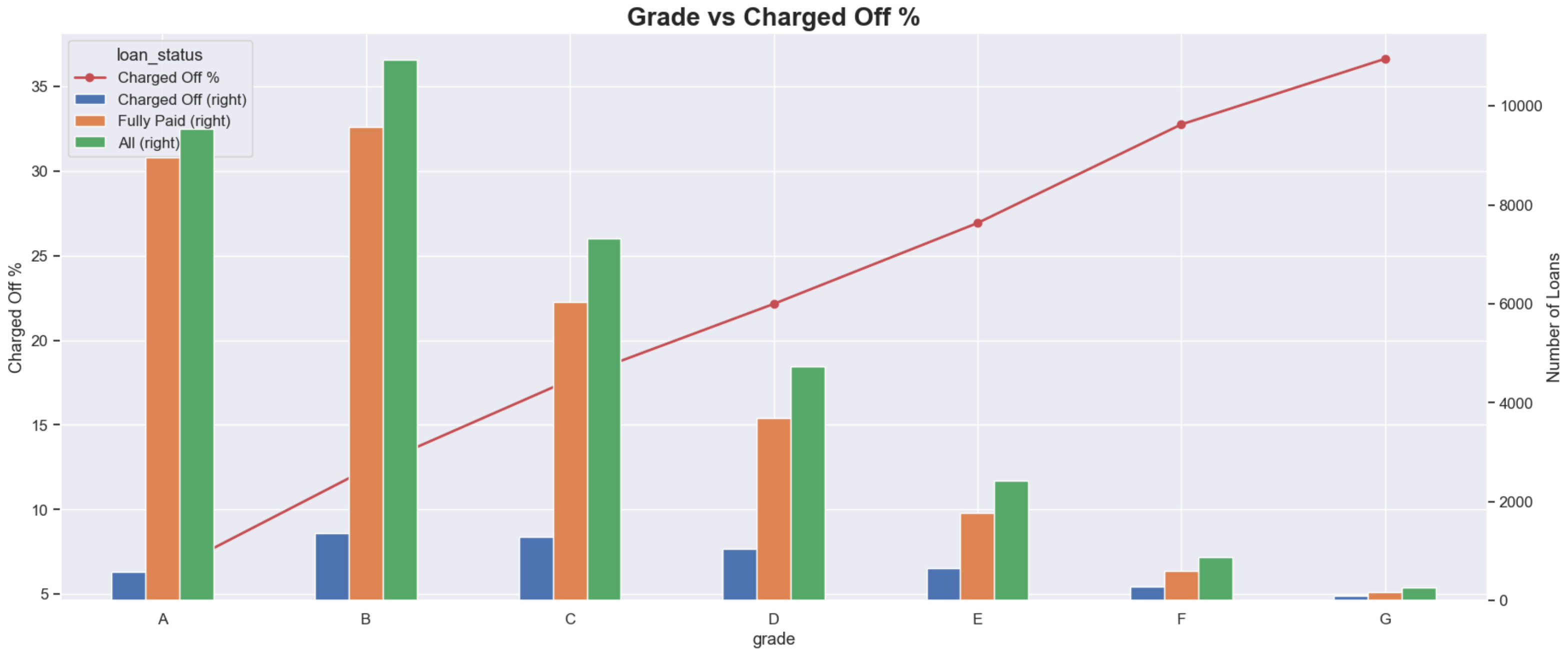
- **Observations:**
  - **UnOrdered Categorical Variable**
    - The Loan Applicants from B C and D category contribute to most "Charged off" loan
    - Loan Applicants Default more with tenure with 60 months then tenure for 36 months.
    - Most loan applicants have ten or more years of Experience and they are less likely to default.
  - **ordered Categorical Variable**
    - Loan Applicants who live in rented or mortgaged houses are more likely to default.
    - Verified Loan applicants are less likely to default on their loans

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

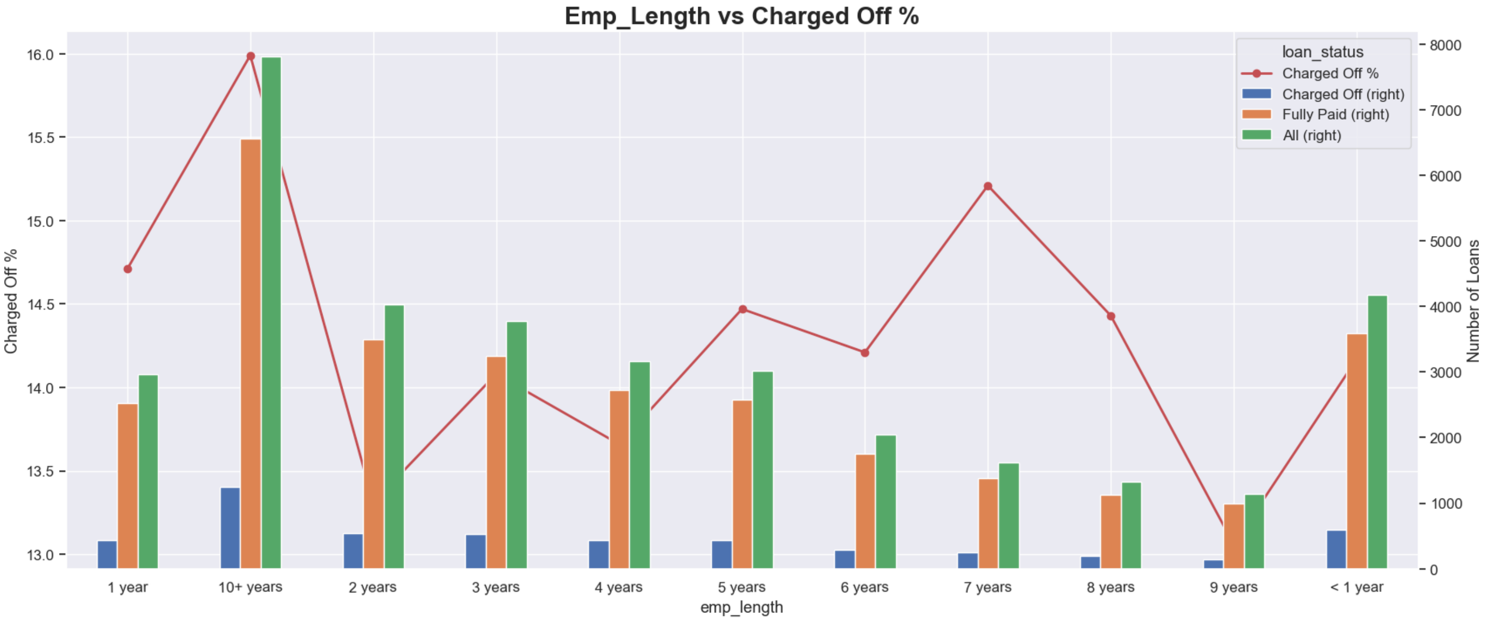
Multivariate Analysis

Graded vs Charged off %



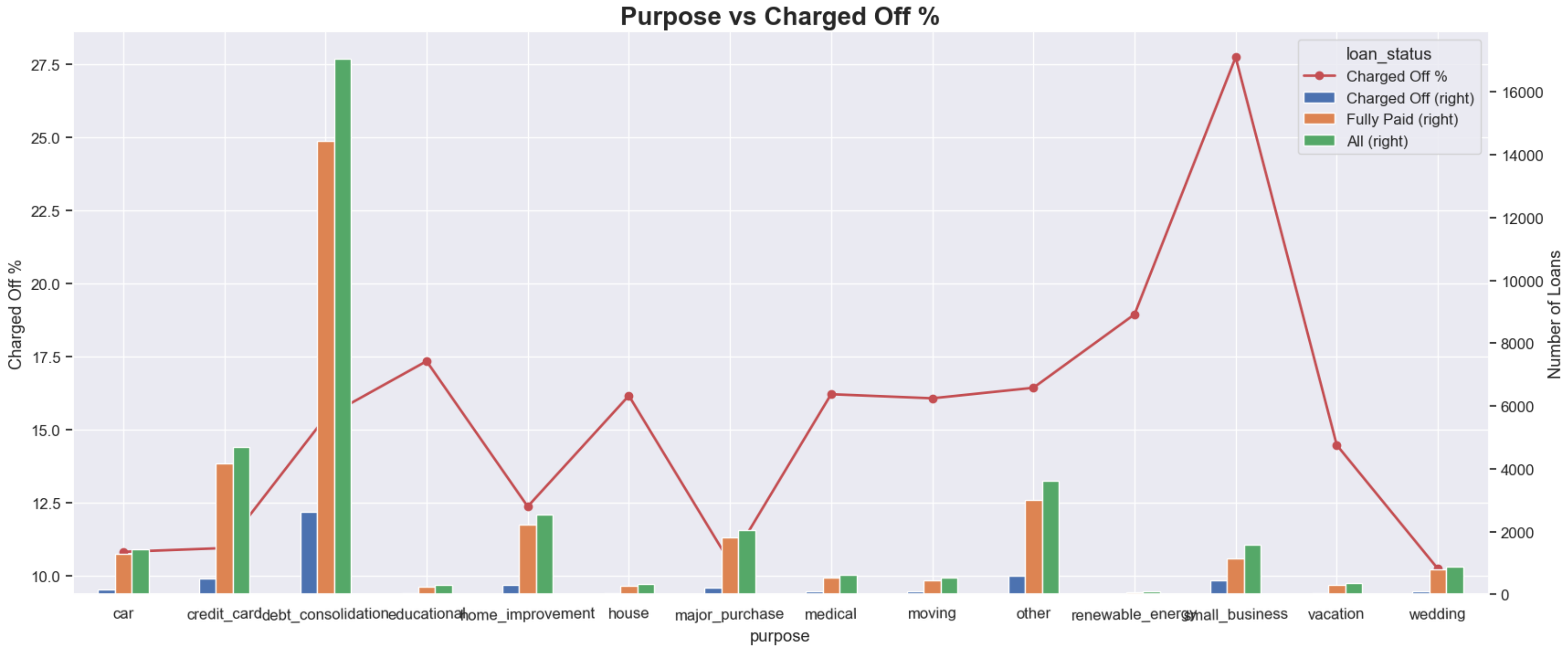
Multivariate Analysis

Employee Experience Length



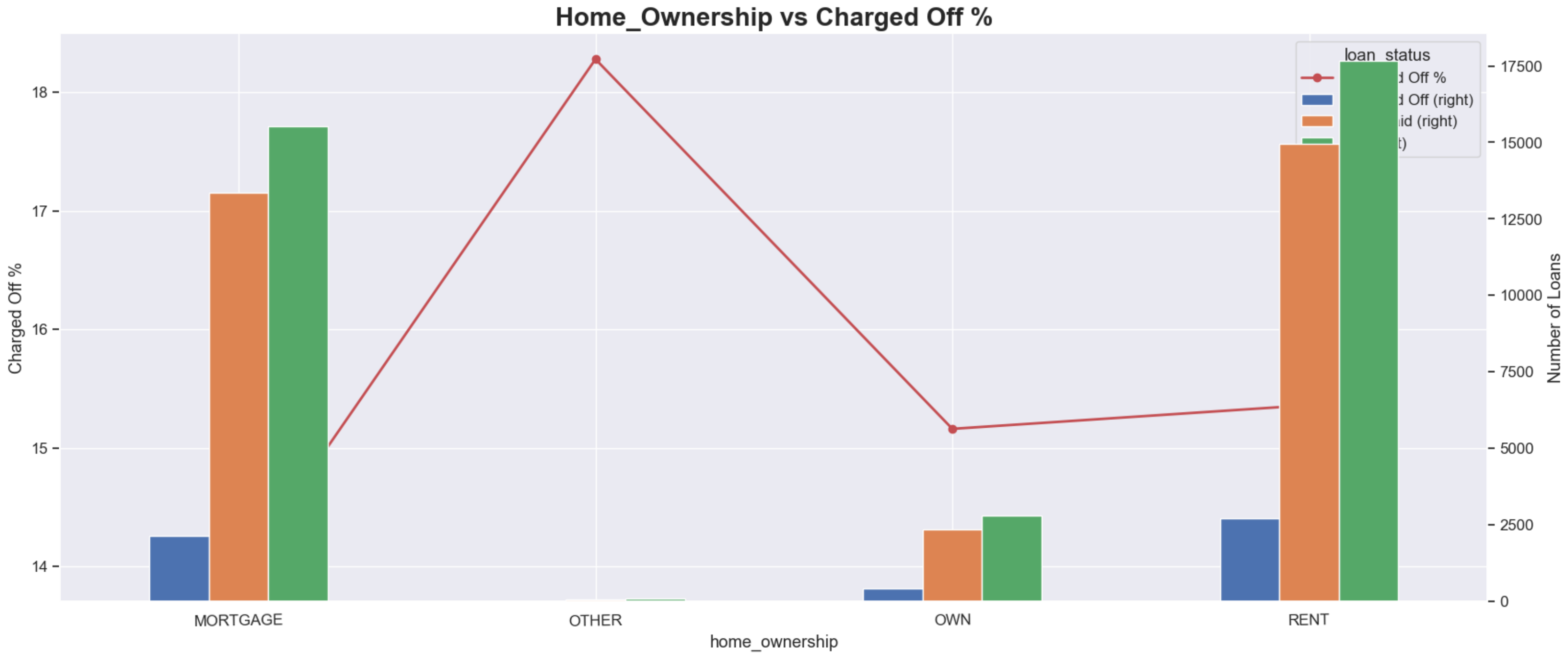
Multivariate Analysis

Purpose



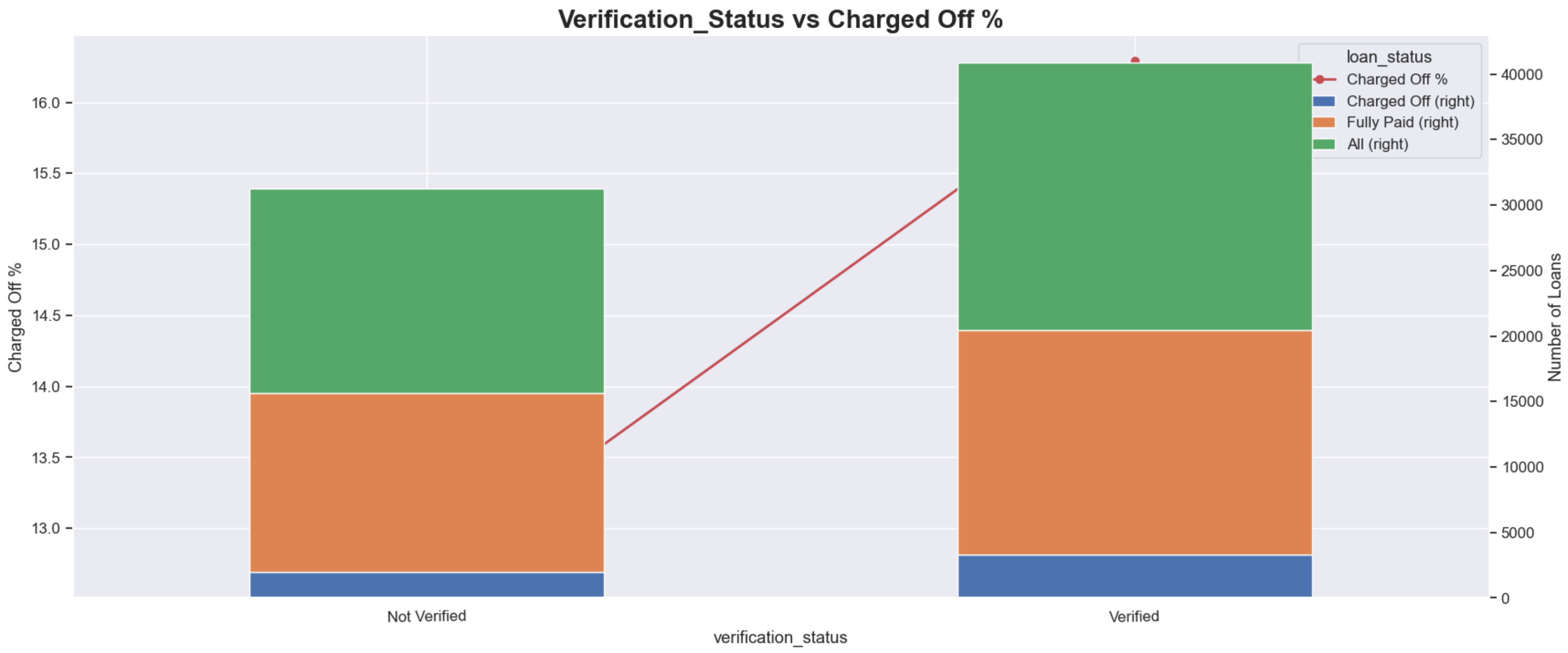
Multivariate Analysis

Home Owned or Rented



Multivariate Analysis

Verification Status



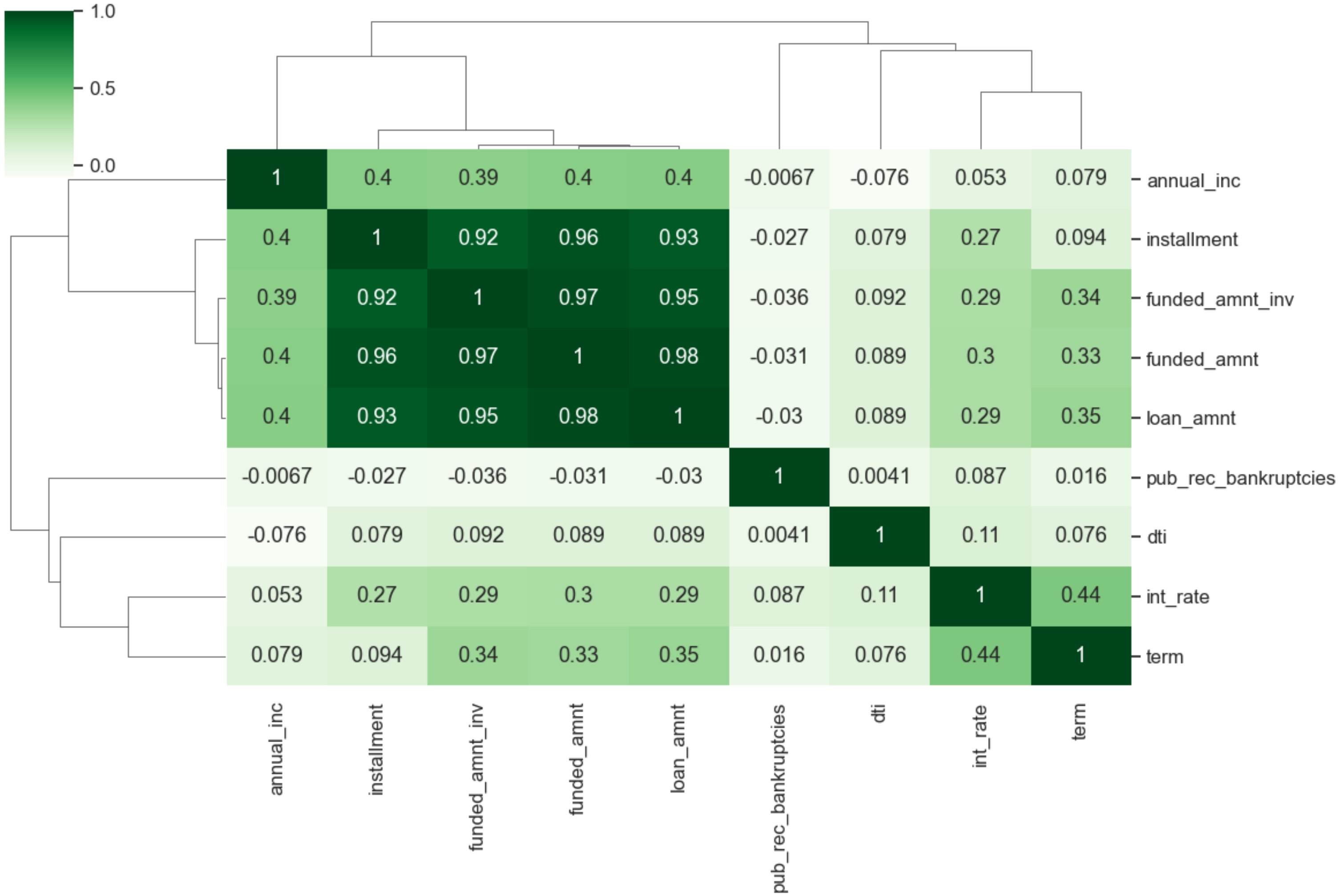
Multivariate Analysis

- **Observations:**
  - Tendency to default the loan is likely with loan applicants belonging to B, C, D grades.
  - Loan applicants with 10 years of experience has maximum tendancy to default the loan.
  - Borrowers from Rented House Ownership have highest tendency to default the loan.
  - The borrowers who are in lower income groups have maximum tendency to default the loan and it generally decreases with the increase in the annual income.
  - The tendency to default the loan is increasing with increase in the interest rate.

Correlation Analysis

- **Understanding:**
  - Correlation analysis is a statistical technique used to measure the strength and direction of the relationship between two or more variable.
  - It quantifies the degree to which change in one variable are associated with changes in another variable.
  - Correlation analysis is widely used in various fields, including finance, economics, biology, psychology, and social sciences, to understand patterns and relationship in data

Correlation Analysis



	annual_inc	installment	funded_amnt_inv	funded_amnt	loan_amnt	pub_rec_bankruptcies	dti	int_rate	term
annual_inc	1	0.4	0.39	0.4	0.4	-0.0067	-0.076	0.053	0.079
installment	0.4	1	0.92	0.96	0.93	-0.027	0.079	0.27	0.094
funded_amnt_inv	0.39	0.92	1	0.97	0.95	-0.036	0.092	0.29	0.34
funded_amnt	0.4	0.96	0.97	1	0.98	-0.031	0.089	0.3	0.33
loan_amnt	0.4	0.93	0.95	0.98	1	-0.03	0.089	0.29	0.35
pub_rec_bankruptcies	-0.0067	-0.027	-0.036	-0.031	-0.03	1	0.0041	0.087	0.016
dti	-0.076	0.079	0.092	0.089	0.089	0.0041	1	0.11	0.076
int_rate	0.053	0.27	0.29	0.3	0.29	0.087	0.11	1	0.44
term	0.079	0.094	0.34	0.33	0.35	0.016	0.076	0.44	1

Correlation Analysis

- **Observation:**

Strong Correlation

- - `installment` has a strong correlation with funded_amnt, loan_amnt, and funded_amnt_inv
- - `term` has a strong correlation with interest rate
- - `annual_inc` has a strong correlation with loan_amount

Weak Correlation

- - `dti` has weak correlation with most of the fields
- - `emp_length` has weak correlation with most of the fields

Negative Correlation

- - `pub_rec_bankrupticies` has a negative correlation with almost every field
- - `annual_inc` has a negative correlation with dti

Thank you !