Lending Club Case Study

Atanu Dutta – Dec 23 – CL60 Batch

IIIT – Bangalore UpGrad
Submission Date: 7th Feb 2024

Problem Statement

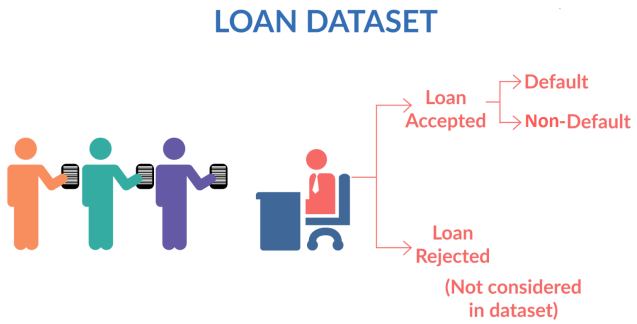
Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). 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'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

In other words, 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 utilise this knowledge for its portfolio and risk assessment

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.



When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

1.Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

1. **Fully paid:** Applicant has fully paid the loan (the principal and the interest rate)
2. **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'.
3. **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

Objective

1. Understanding the dataset
2. Data Cleaning and Manipulation
3. Handling missing data & outliers
4. Data analysis
5. Presentation and Recommendations for Leading Club
6. Conciseness and readability of the code (Jupyter Notebook)

1

Understanding the data set

Loan Dataset Shape **Rows: 39717, Cols: 111**	Loan Dataset info **dtypes: float64(74), int64(13), object(24)**
Data Dictionary Shape **Rows: 117, Cols: 2**	Loan Dataset info **dtypes: object(2)**

Two data files are provided for this case study, will be attached in github

- Loan.csv
- Data_Dictionary.xlsx

We will use Jupiter Notebook and below python packages to complete the EDA exercise

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Warnings
- summarytools

2

Data Cleansing & manipulation

Review the data loaded in data frame to understand which all columns are relevant, if we need to change data types or derive new columns for better analysis of the data. Remove any data which will not help us take any decision

# of NULL Columns Dropped **54 (100%), 1 (97%), 1 (95%), 1 (65%)**	# of Columns with Unique Values = Total Rows Dropped **3 –** ID, URL, Member ID
# of Columns with only 1 Unique Values Dropped **9**	# of Columns with long Desc Dropped **1 -** desc
Check for duplicate rows None **0**	Filtered unrelated data Loan status=Current **1140 rows**
Datatype Conversion Term, Int Rate, issue dt, Employee Length **4**	Derived Values Issue Year, Month, Loan Approved Amt Ratio, Loan Amount Bucket **4**

3

Univariate Analysis

- **Univariate analysis** is a statistical method used to analyse 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

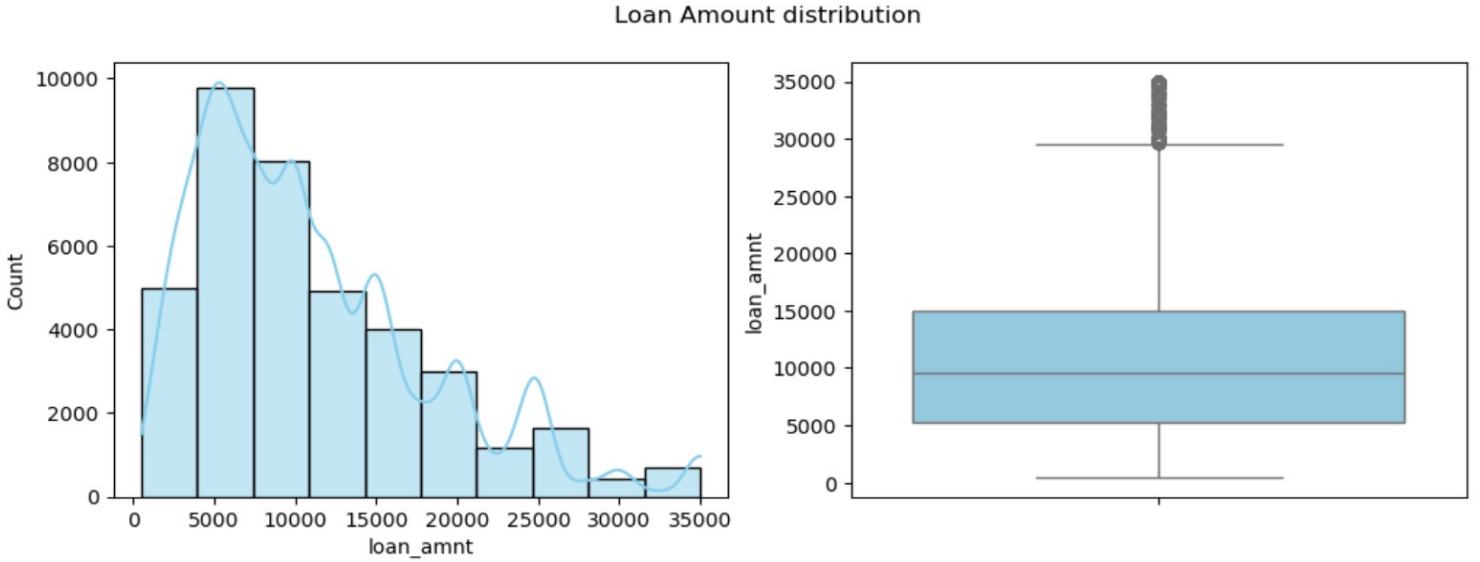
Ordered	Unordered
Grade (A to G)	Address State
Term (36 or 60 Months)	Loan Purpose
Employee Length	Home Ownership
Issue Year & Month	Loan Status

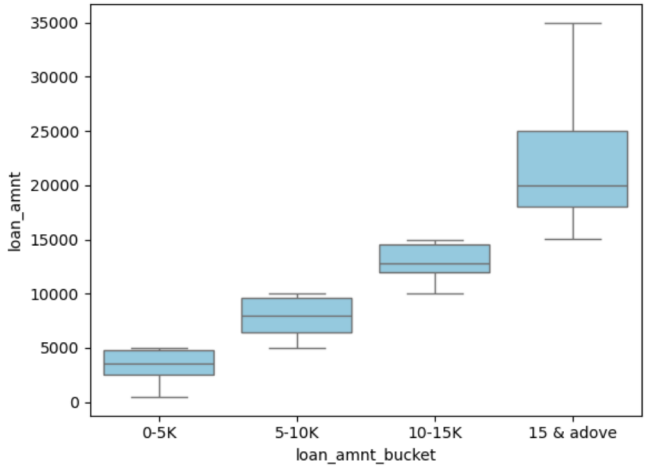
Quantitative Variables
Interest Rate Bucket
Loan and Funded Amount
Annual Increment
DTI
Pub rec bankruptcies
Loan Amount Bin

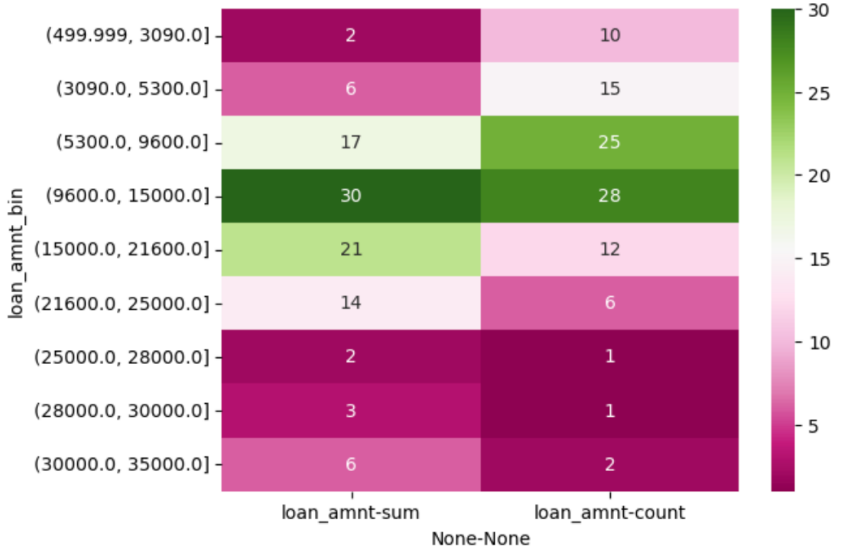
Univariate Analysis

Observations

- More number of people have taken loan in range of **10K as mean is 9600**
- Sum total of loan amount is in range of **9600 to 15000**
- **Very less percentage** of people have taken loan amount **more than 30K**



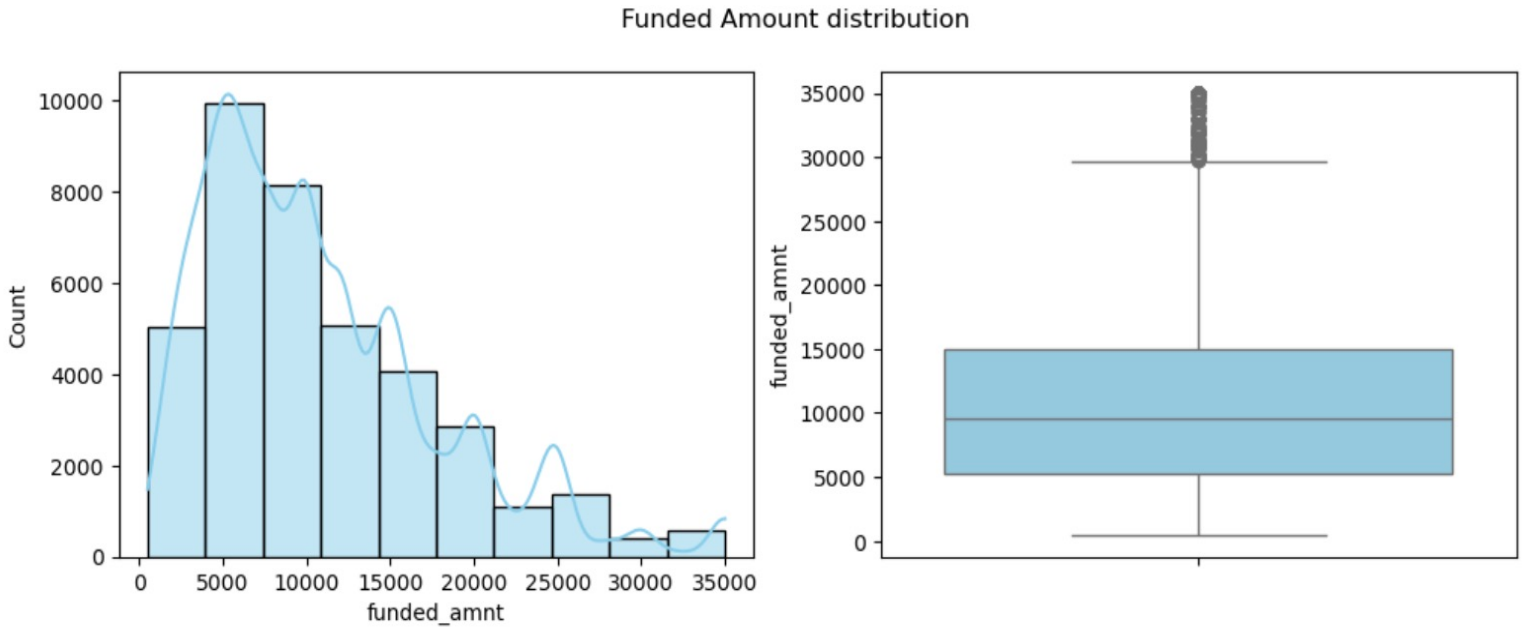




Univariate Analysis

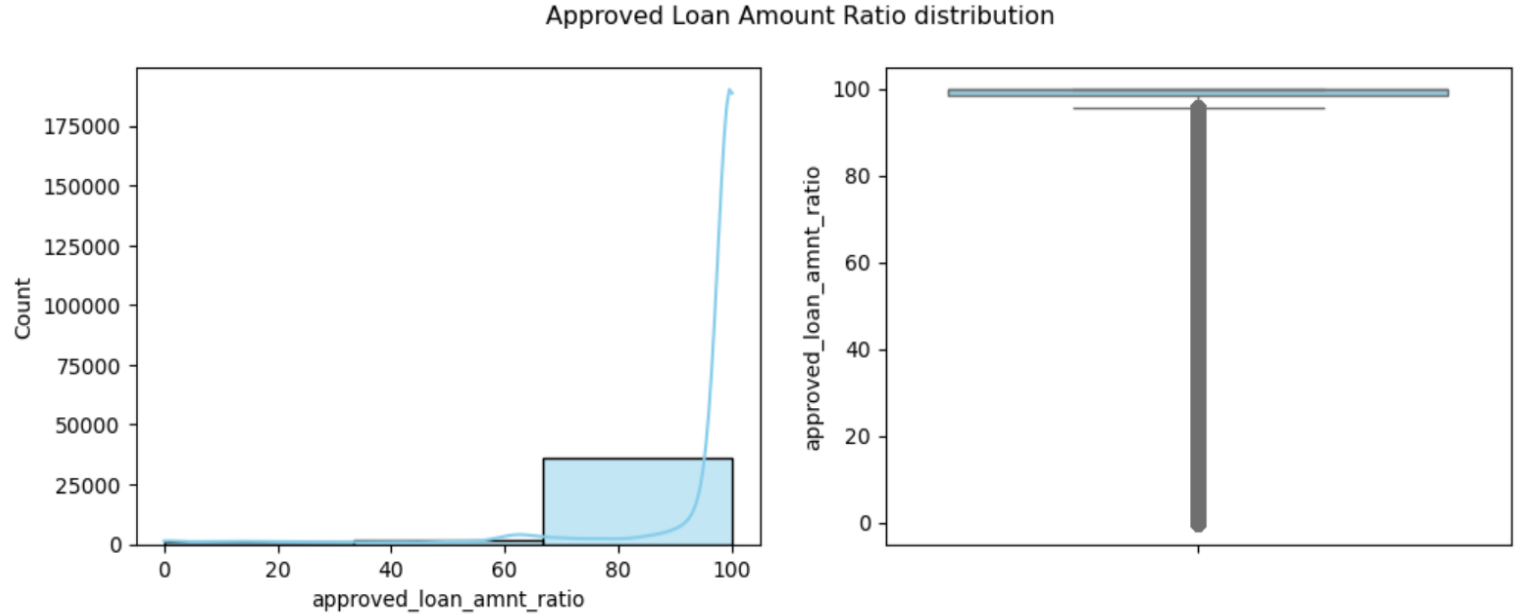
Observations

- behaviours of funded_amnt is same as loan_amnt
- **51 Percentile** of loan application amounts are **100% approved**. In below stats it shows 99.93 is 50 percentile. So, find the exact number have written a reusable function




```
loan.approved_loan_amnt_ratio.describe()
```

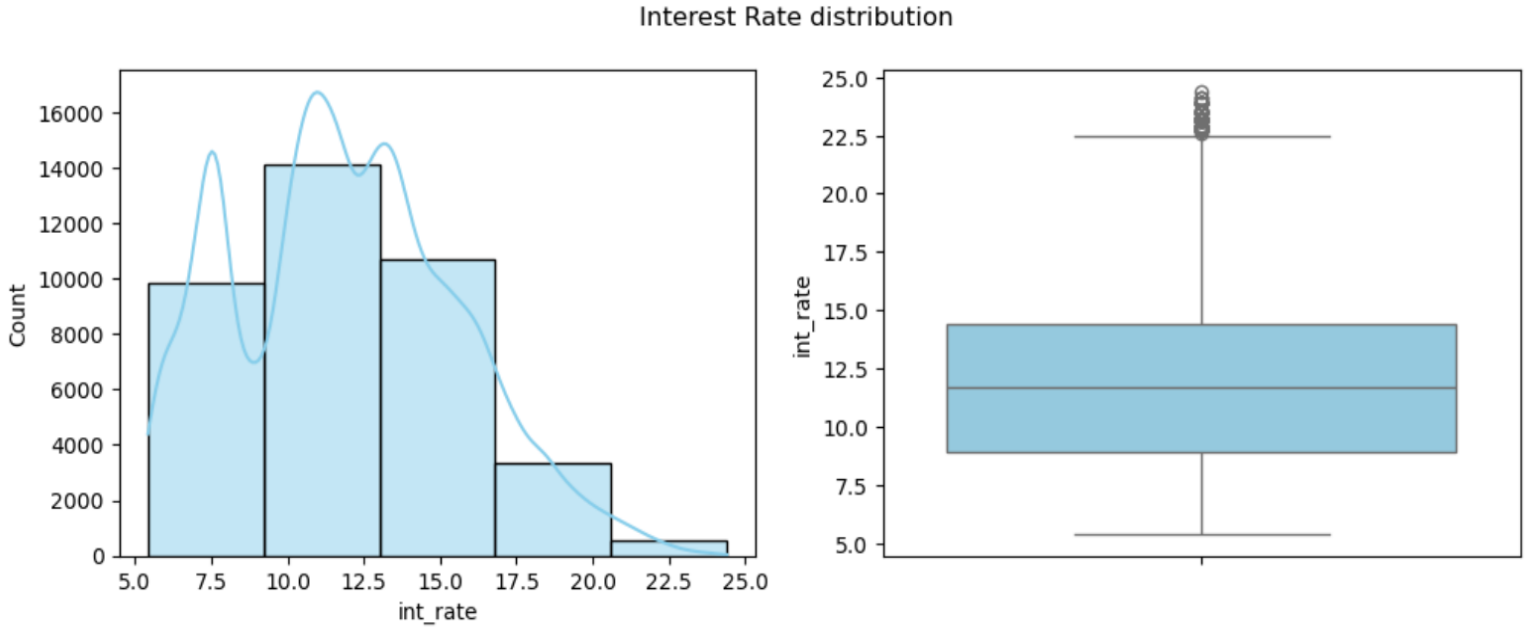
```
count    38577.000000
mean        93.787685
std         17.331127
min          0.000000
25%         98.250000
50%         99.930000
75%        100.000000
max        100.000000
Name: approved_loan_amnt_ratio, dtype: float64
```

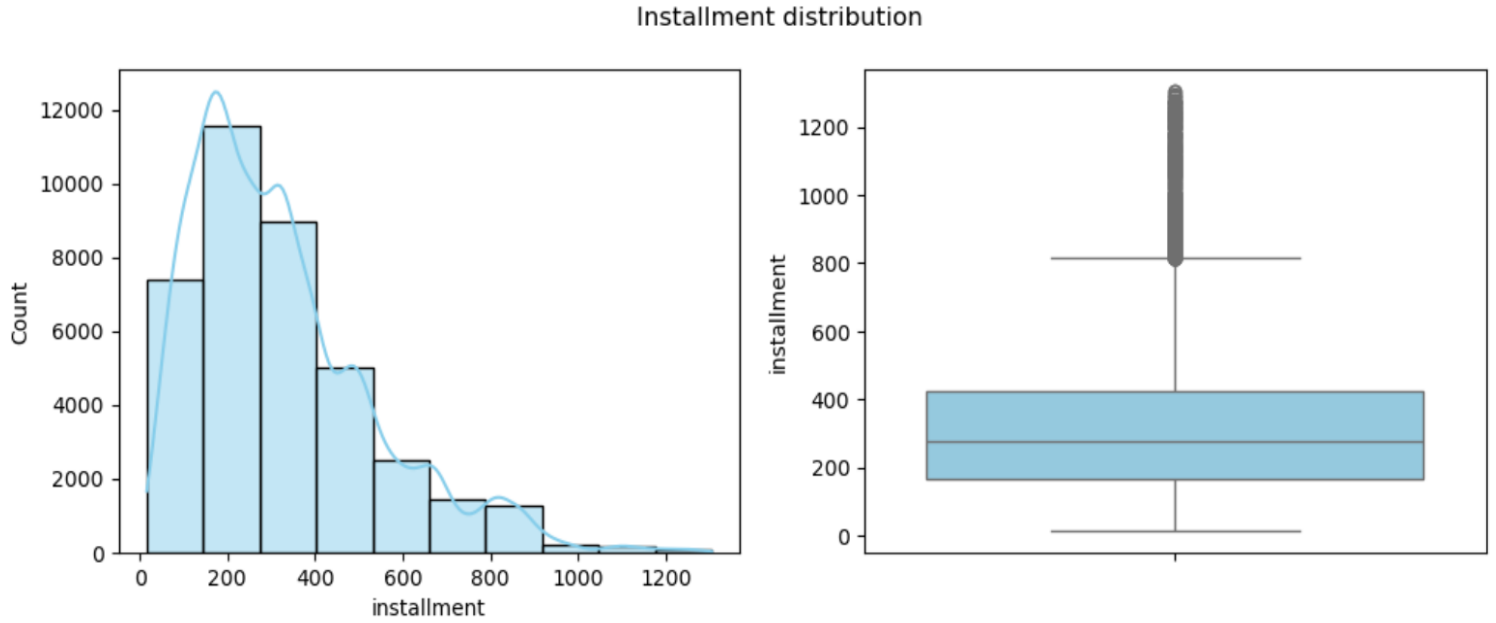


Univariate Analysis

Observations

- Most of the interest rates lies between **9% to 14.5%** Some people took loan at higher rates of interest, **22.5%**
- Mostly people have taken instalment of **277 months means 23 Years**, we see many outliers like 800+ which means more than 66 years





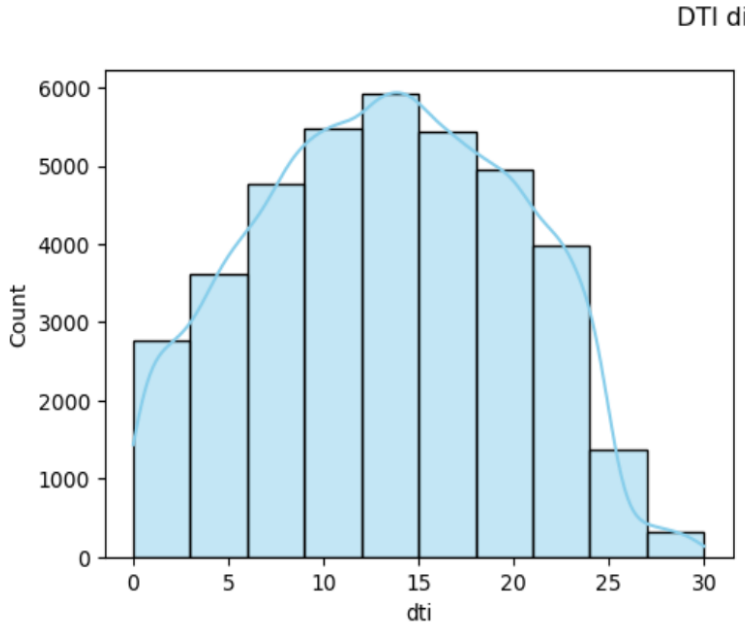
Univariate Analysis

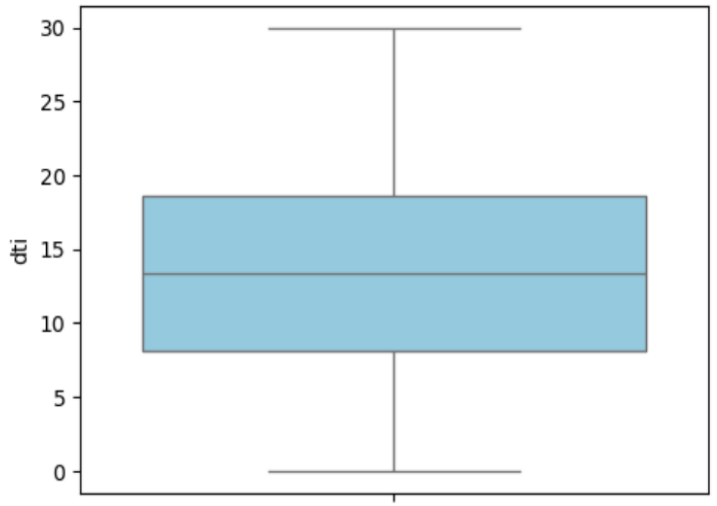
Observations

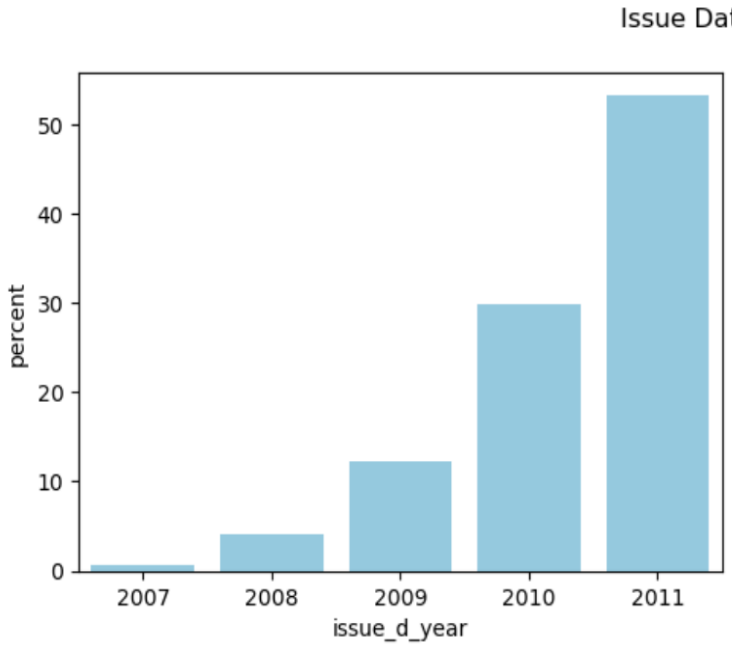
- There are **no outlier in DTI** and it looks like normal distribution curve. Means loans are given to people with good DTI ration
- over a period of time lending club has **almost doubled** the loan disbursement very year
- last 3 months of the year **Q4 - Oct to Dec the issue rate is high**

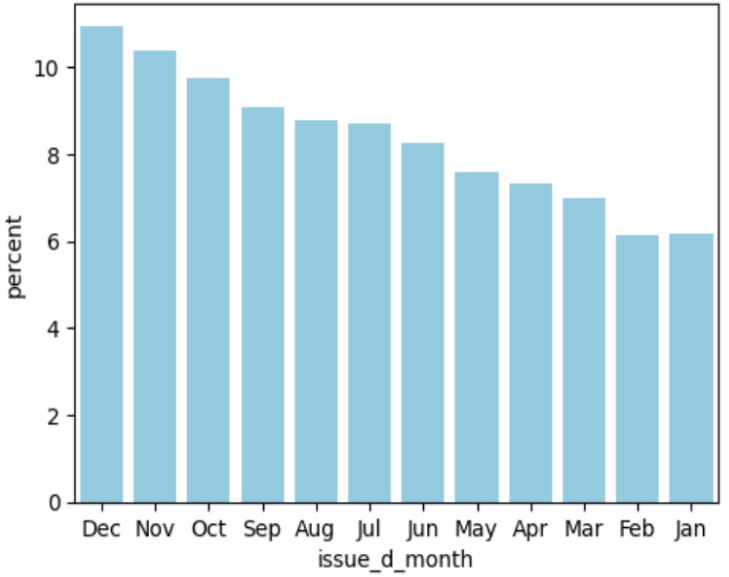
```
loan.pub_rec_bankruptcies.value_counts(normalize =True)*100
```

```
0.0    95.665259
1.0     4.321542
2.0     0.013200
Name: pub_rec_bankruptcies, dtype: float64
```





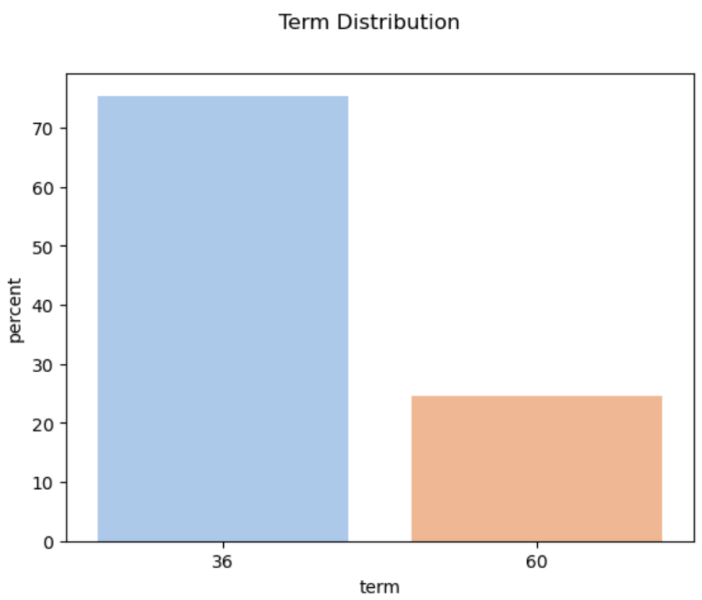


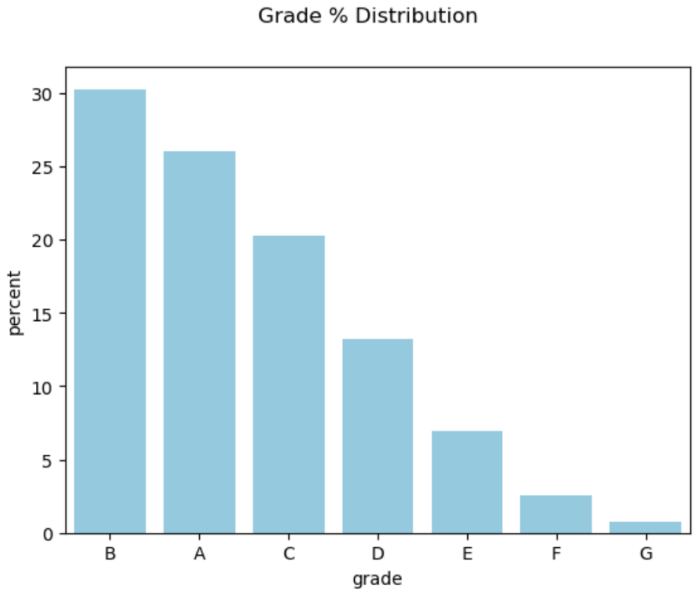


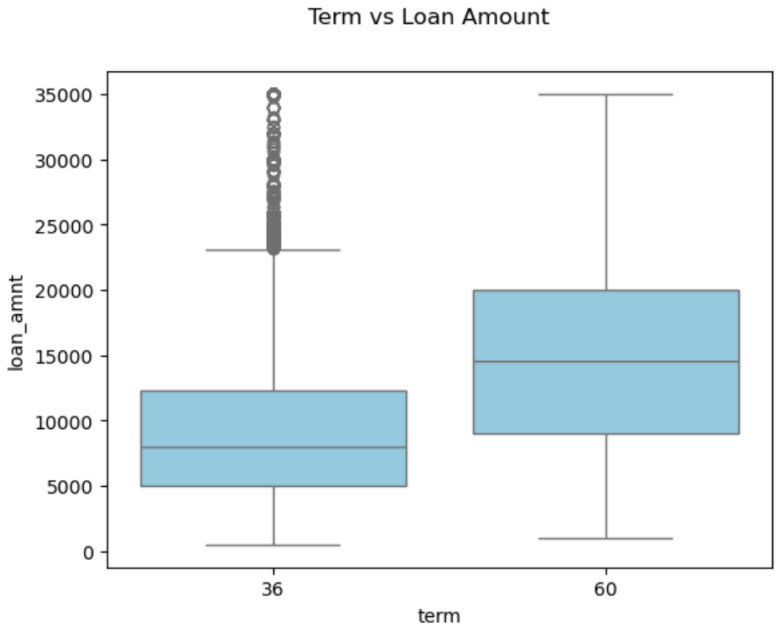
Univariate Analysis

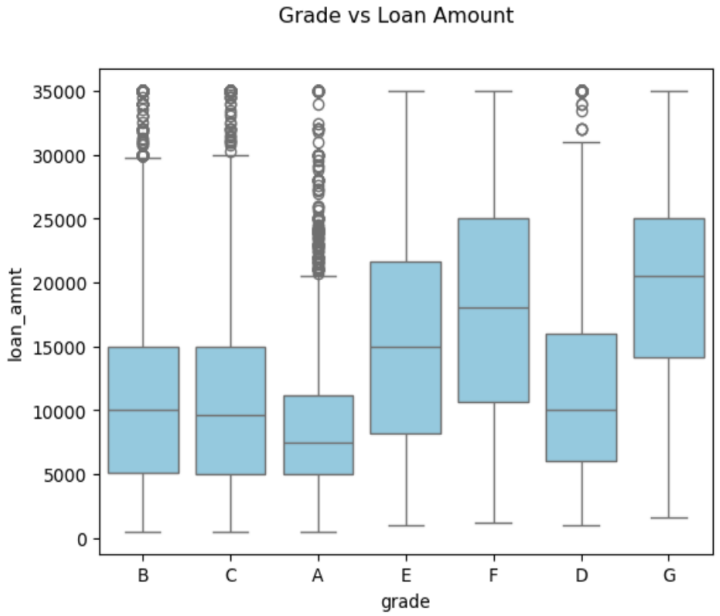
Observations

- **75% of issue loan** is for **36 months** term & **25%** is for **60 months** term
- **Higher amount** loans have high term of **60 months**
- **56%** borrower are from **Grade A & B**
- **Grade F & G** has taken **higher amount/range of loan**
- **Grade A, B, C have taken lower range of loans**





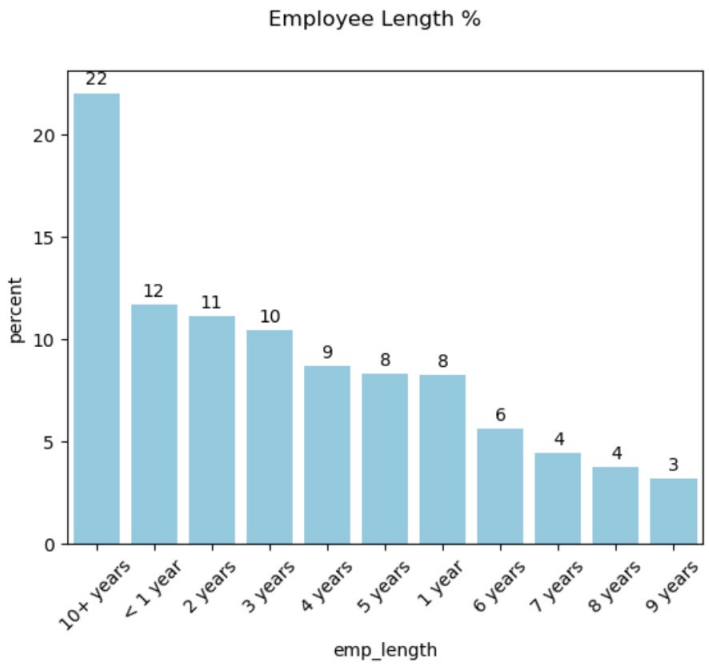


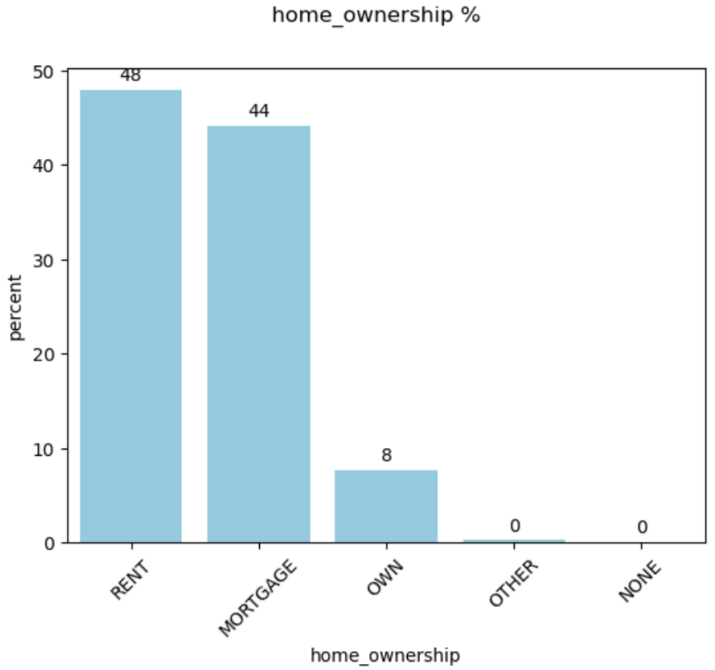


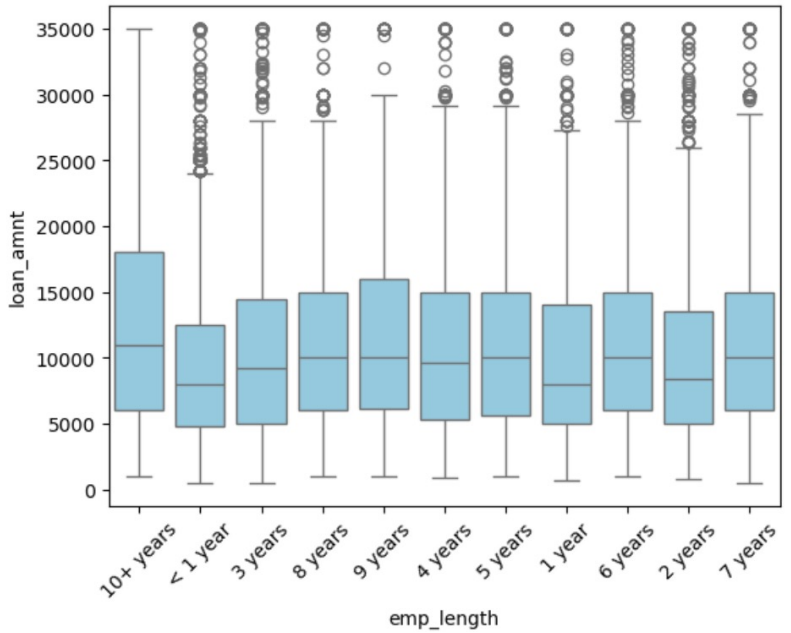
Univariate Analysis

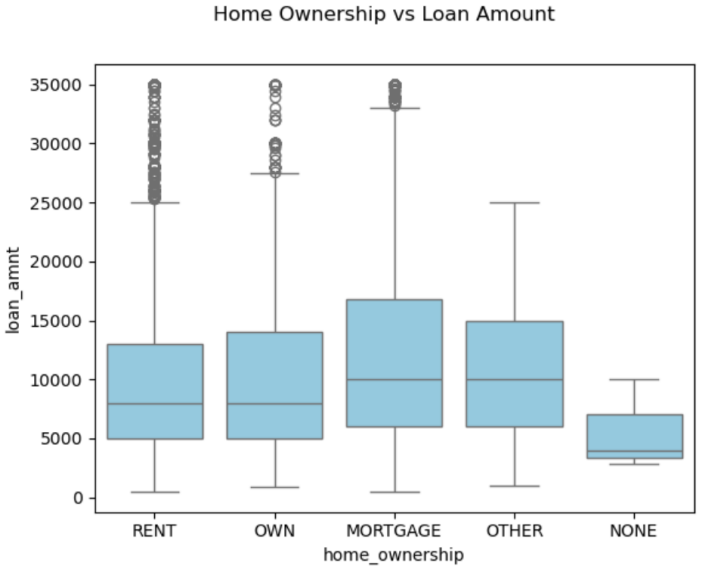
Observations

- **Most** of the borrower are of **10+ Years** Length
- **92%** of the borrower are have having **Rent & Mortgage**
- **Employee with 10+ Years** have taken **higher** loan amount
- Borrower with **Mortgage** has taken **higher amount** of loans followed by **Others**





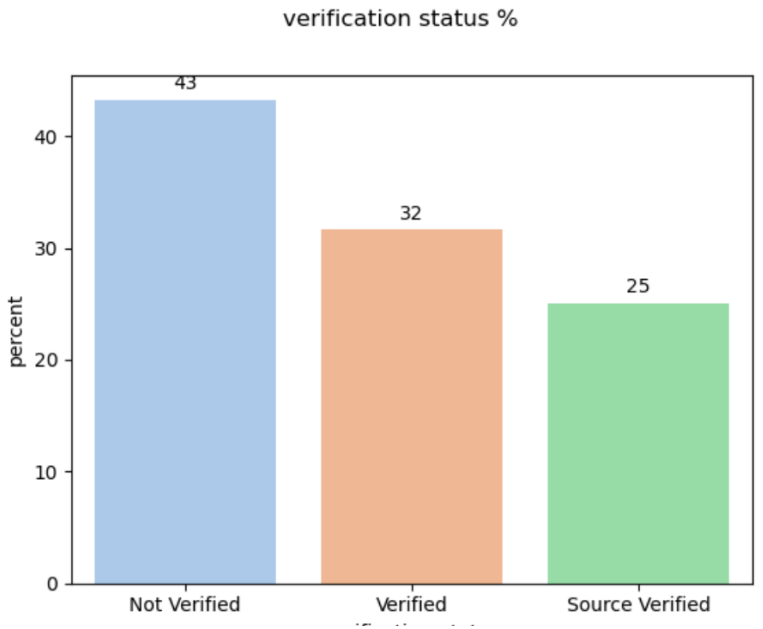


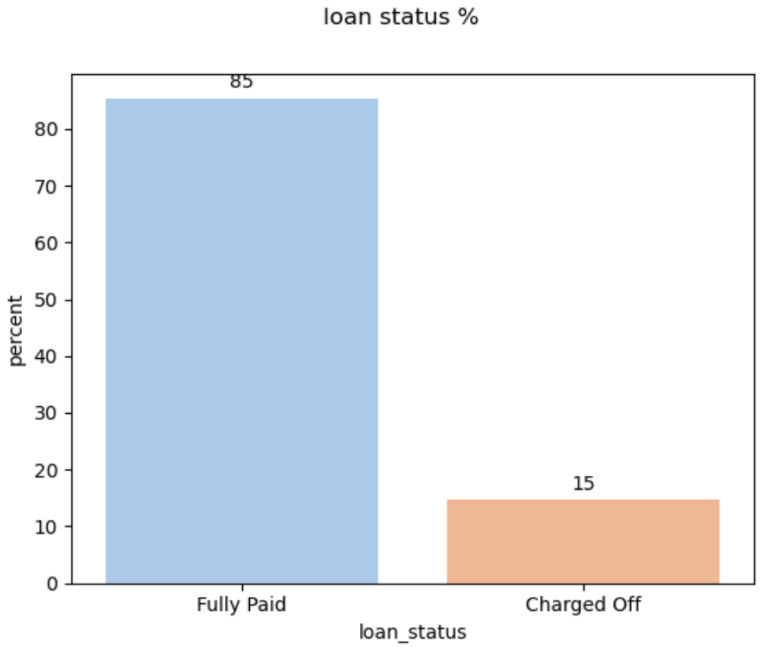


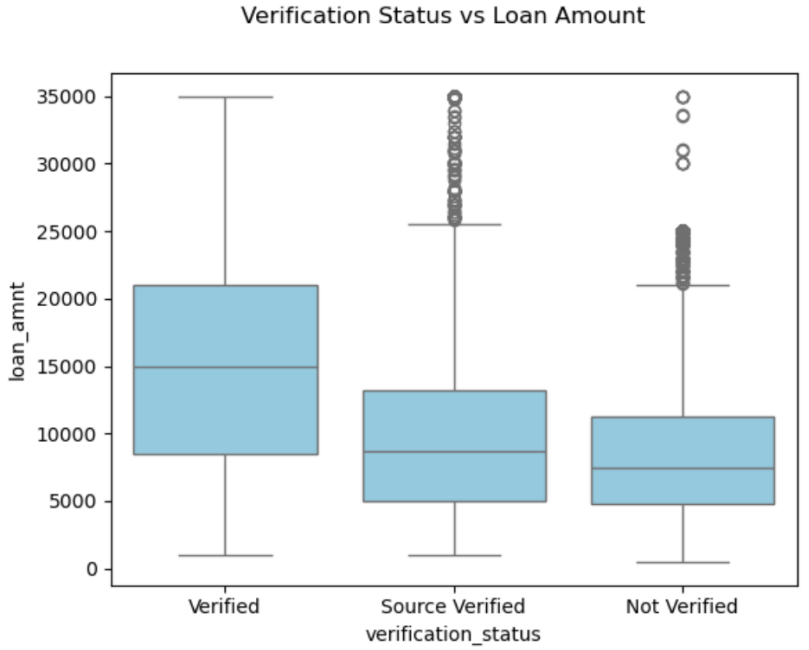
Univariate Analysis

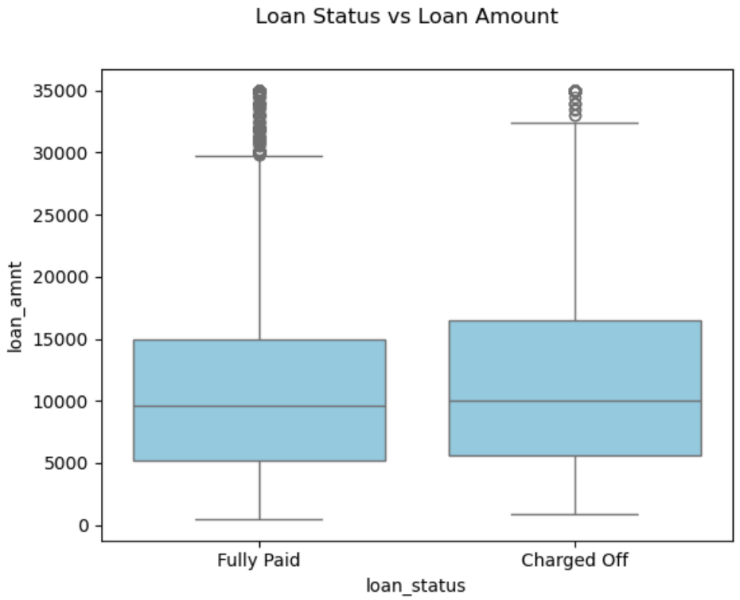
Observations

- **43%** of the loans are issued **without Verification**
- **85%** of the borrower **fully paid** the load and **15%** of the **Defaulted**
- **47%** of people took loan for Debt Consolidation
- Followed by **13%** for Credit Card Payment
- **Higher loan** amount is ONLY issued to **Verified Borrower**
- There is not much impact on loan amount in terms Fully Paid & Charged off





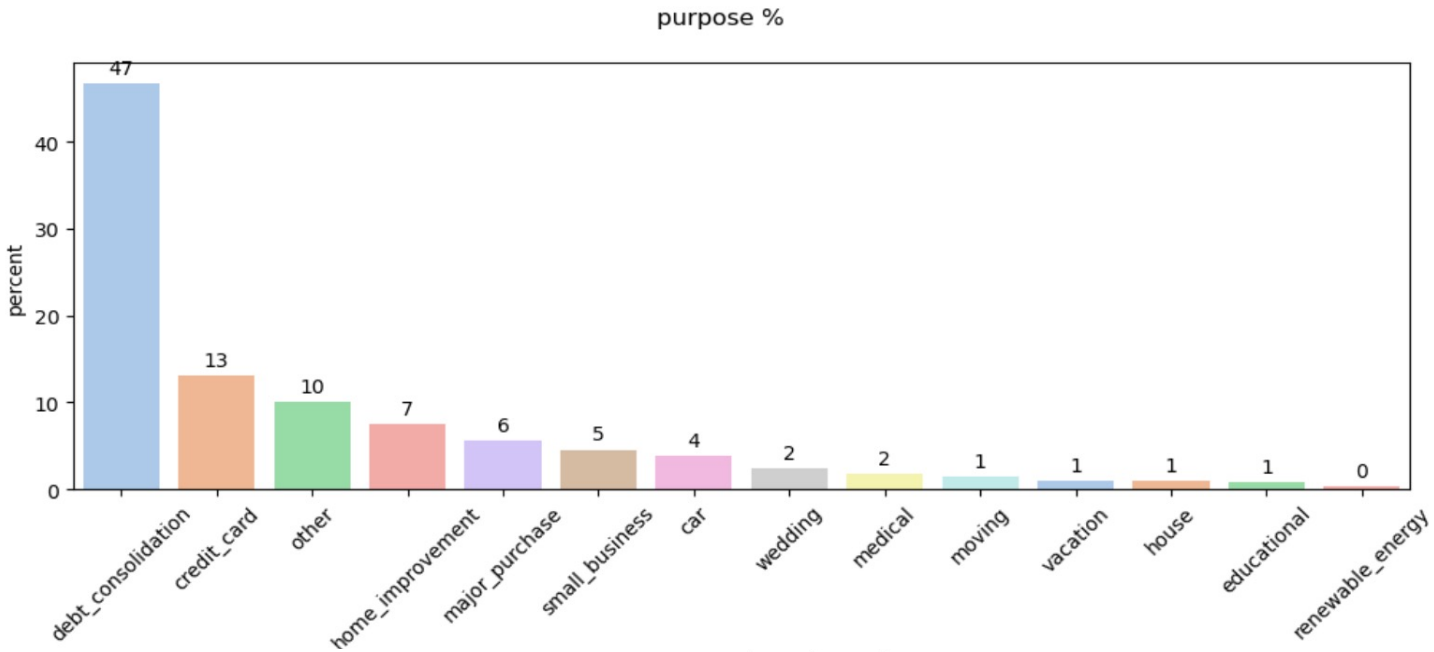


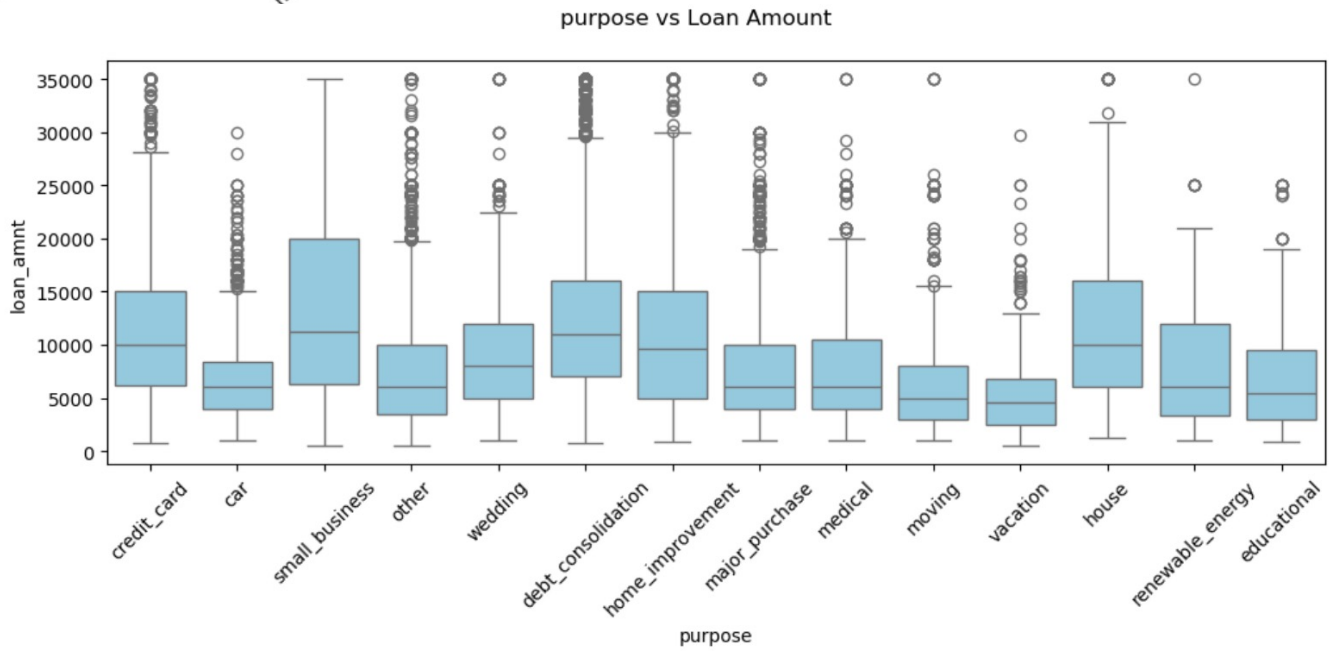


Univariate Analysis

Observations

- **47%** of people took loan for **Debt Consolidation**
- Followed by **13%** for Credit Card Payment
- **Small business is** taking **higher loan amount**

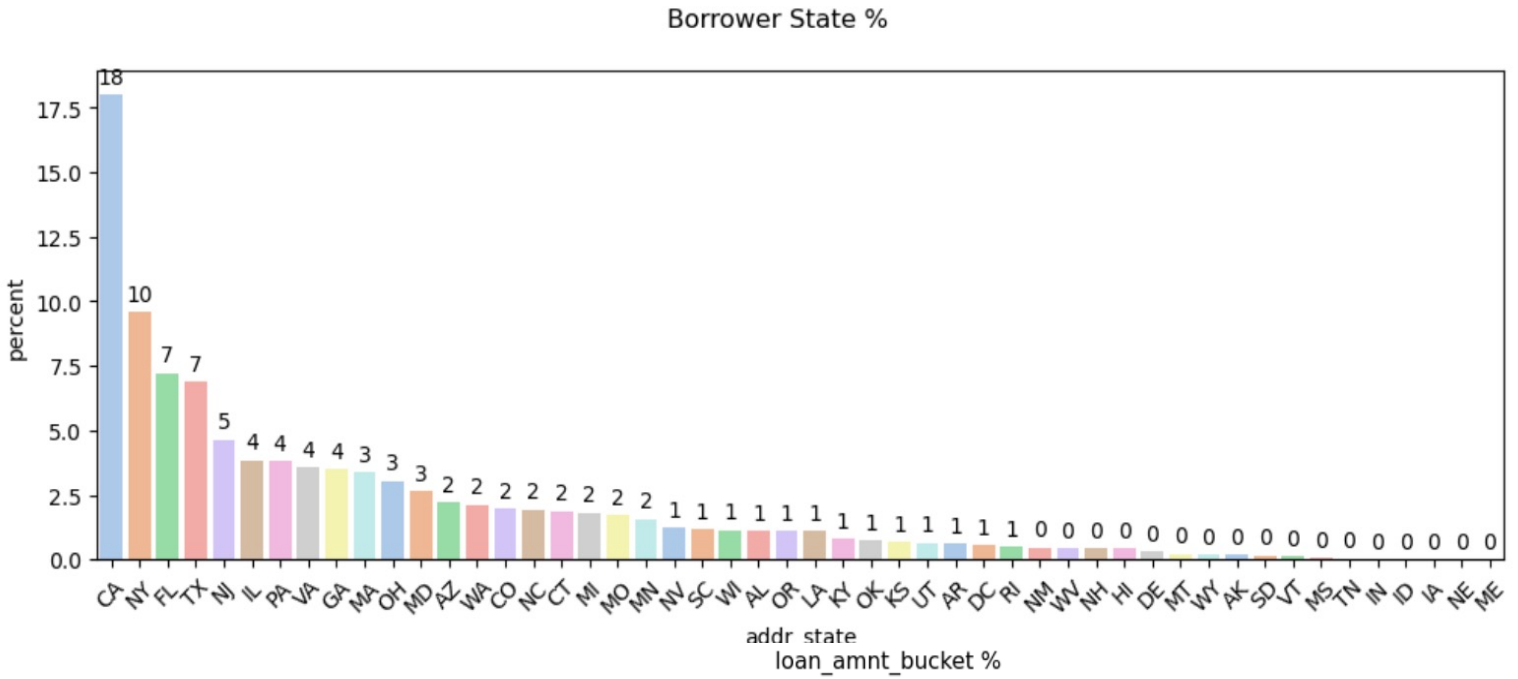


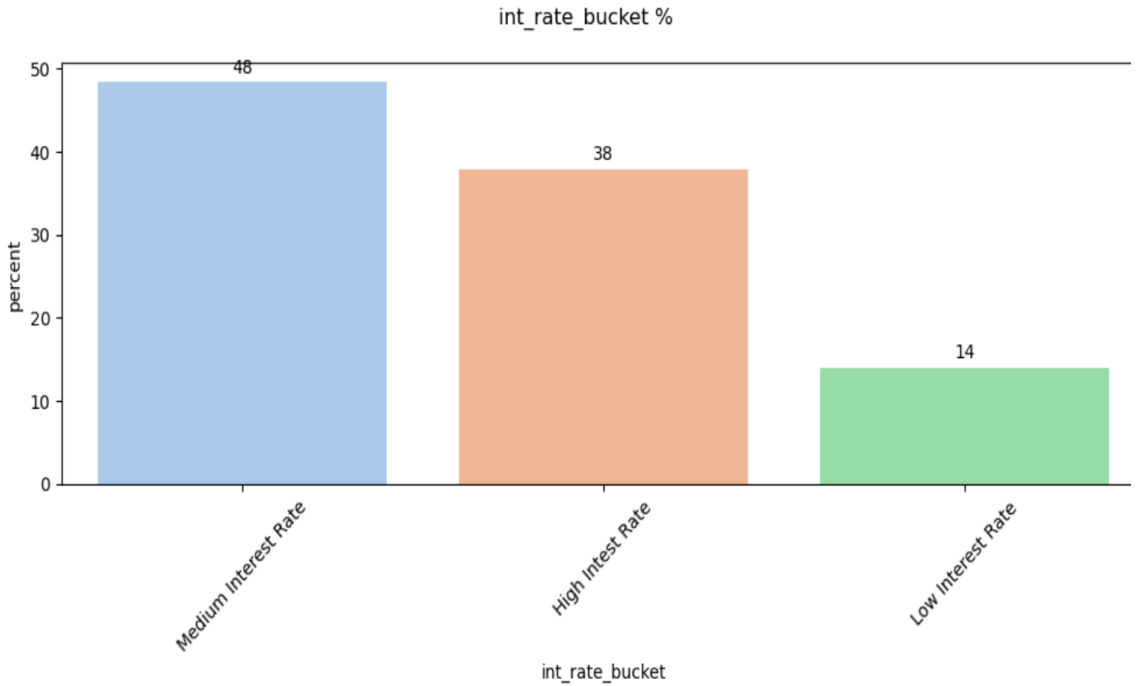


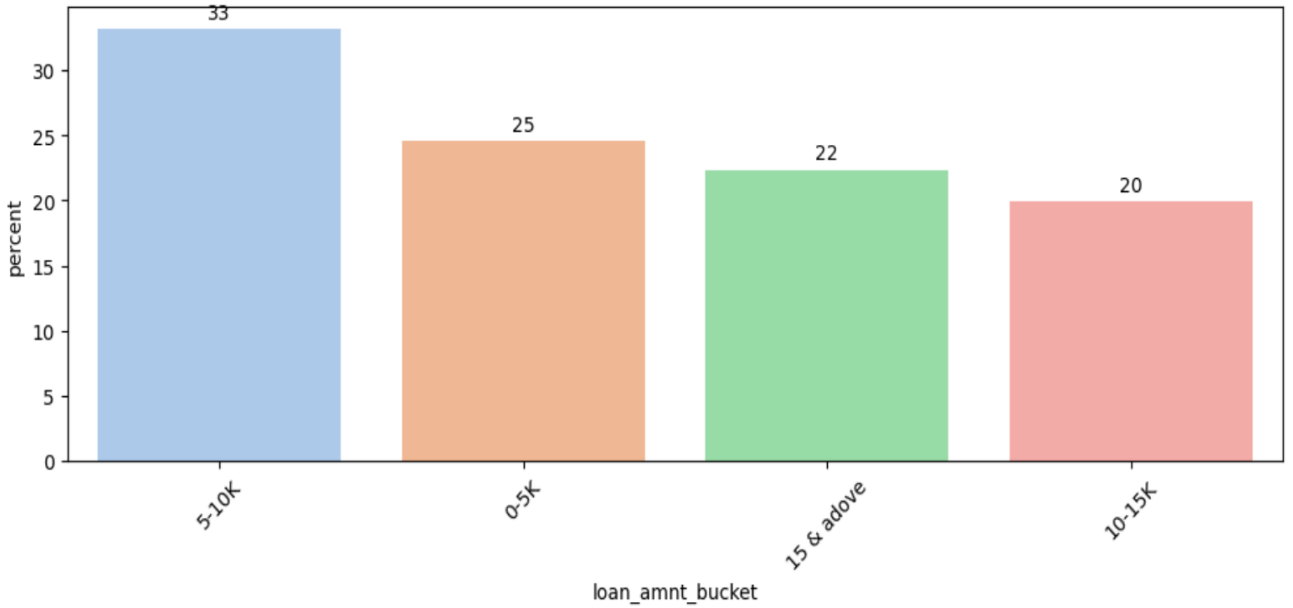
Univariate Analysis

Observations

- **48%** of people have taken medium interest rate loan
- Only **14%** have borrower has taken low interest rate
- **58%** of people have taken loan in range of **0-10K**
- Only **22%** have taken loan above 15K







4

Bivariate Analysis

- Bivariate analysis is a statistical method that involves the simultaneous analysis of two variables. It aims to determine the relationship between them to identify patterns
- It was carried out for both Categorical and Quantitative Variables

Ordered	Unordered
Grade (A to G)	Address State
Term (36 or 60 Months)	Loan Purpose
Employee Length	Home Ownership
Issue Year & Month	Loan Status

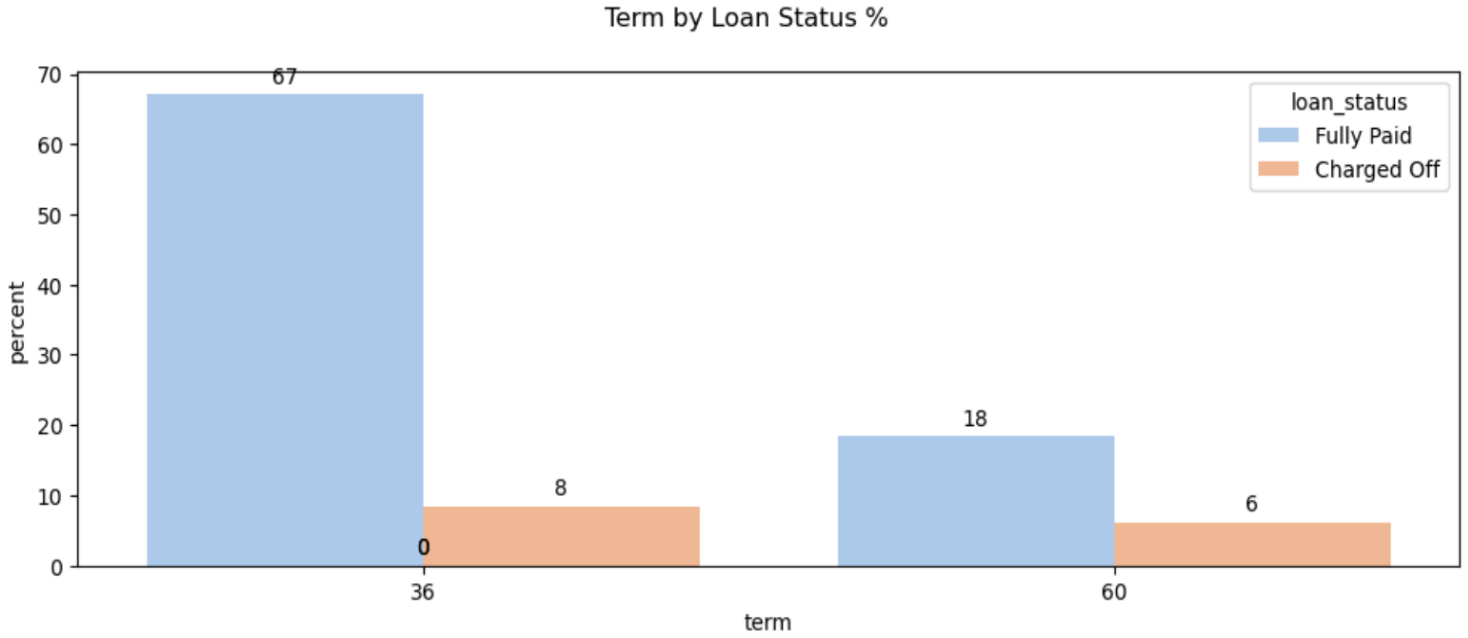
Quantitative Variables
Interest Rate Bucket
Loan and Funded Amount
Annual Increment
DTI
Pub rec bankruptcies
Loan Amount Bin

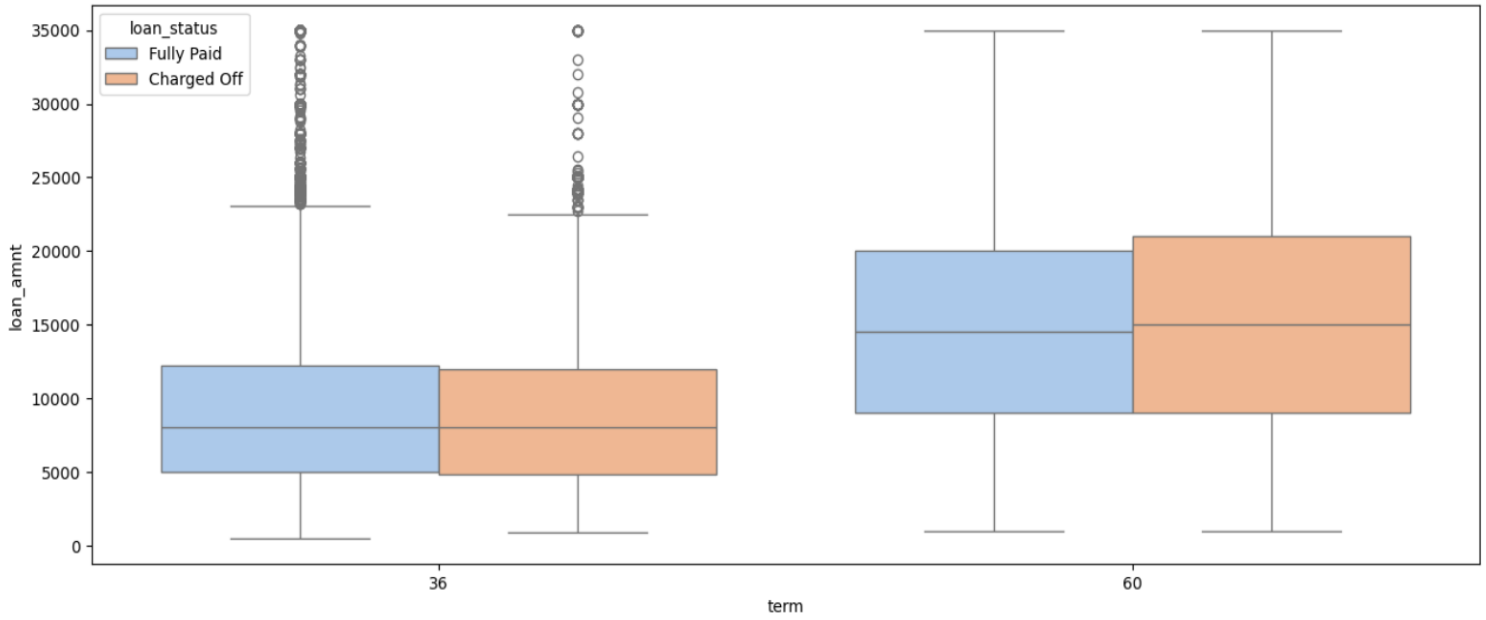
Bivariate Analysis

Observations

- Overall percentage of Defaulters is little higher **8%** for Term **36 months** vs **6%** for **60 Months**

- Borrower with 60 months terms defaulted where they took higher loan amount. But its not so relevant to take a major call

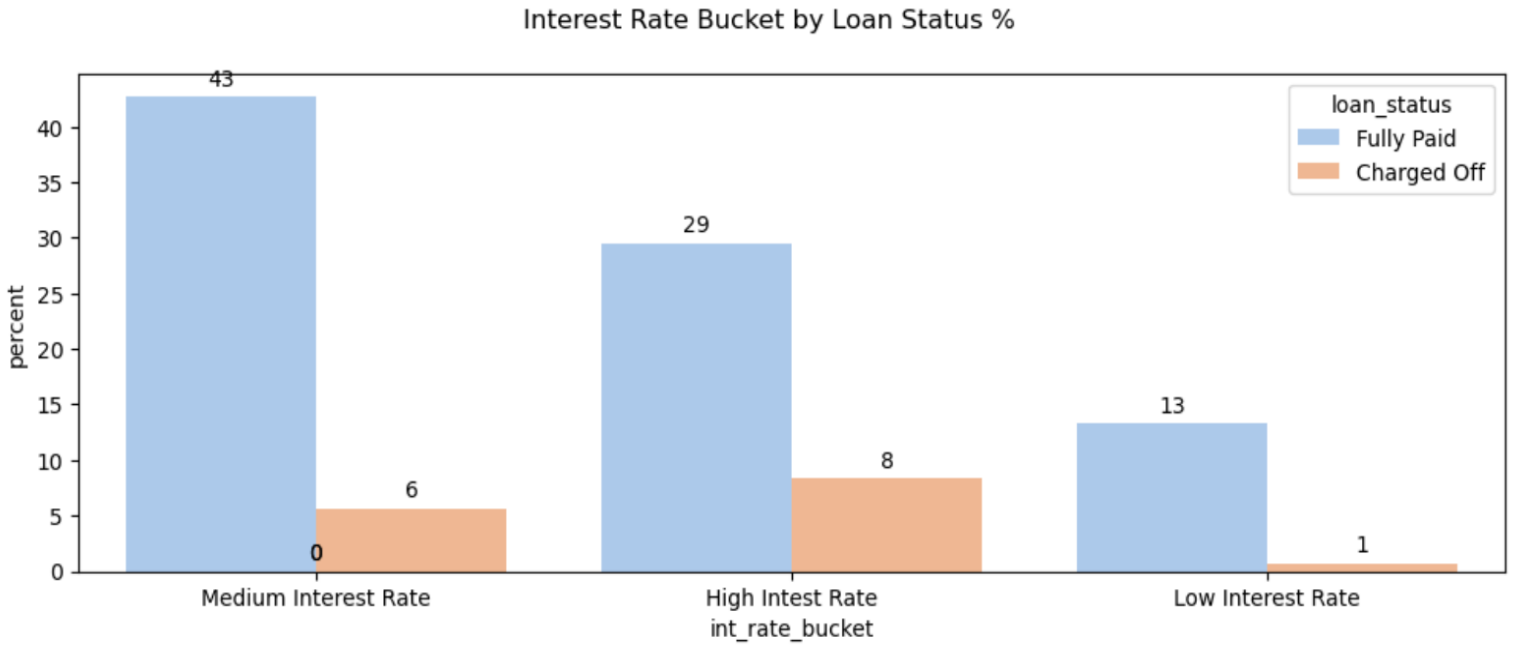


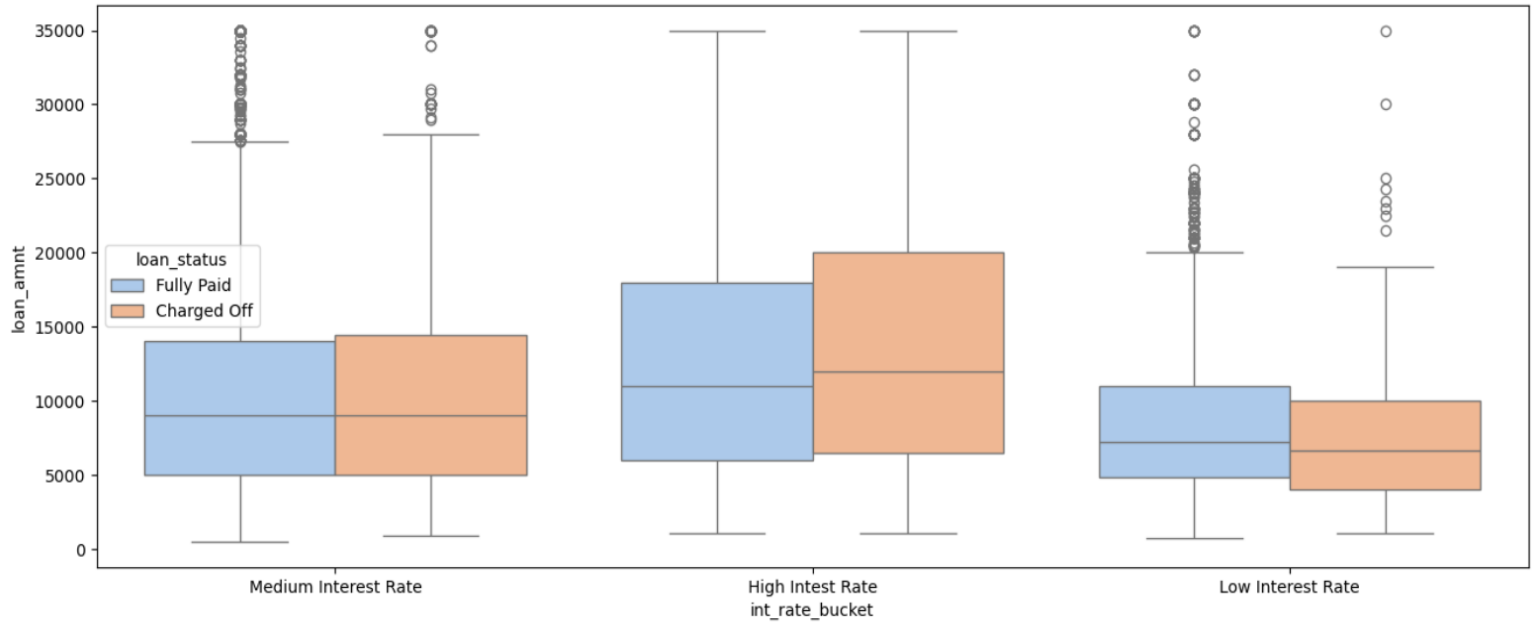


Bivariate Analysis

Observations

- Overall percentage of Defaulters are little higher **8%** for **high interest rate bucket** followed by **Medium - 6%**
- Only **1%** default who took loan in low interest rate bucket, means there is very less risk in giving loans in low interest rate bucket
- Borrower who defaulted in high interest rate category took move loan amounts vs rest in same cohort



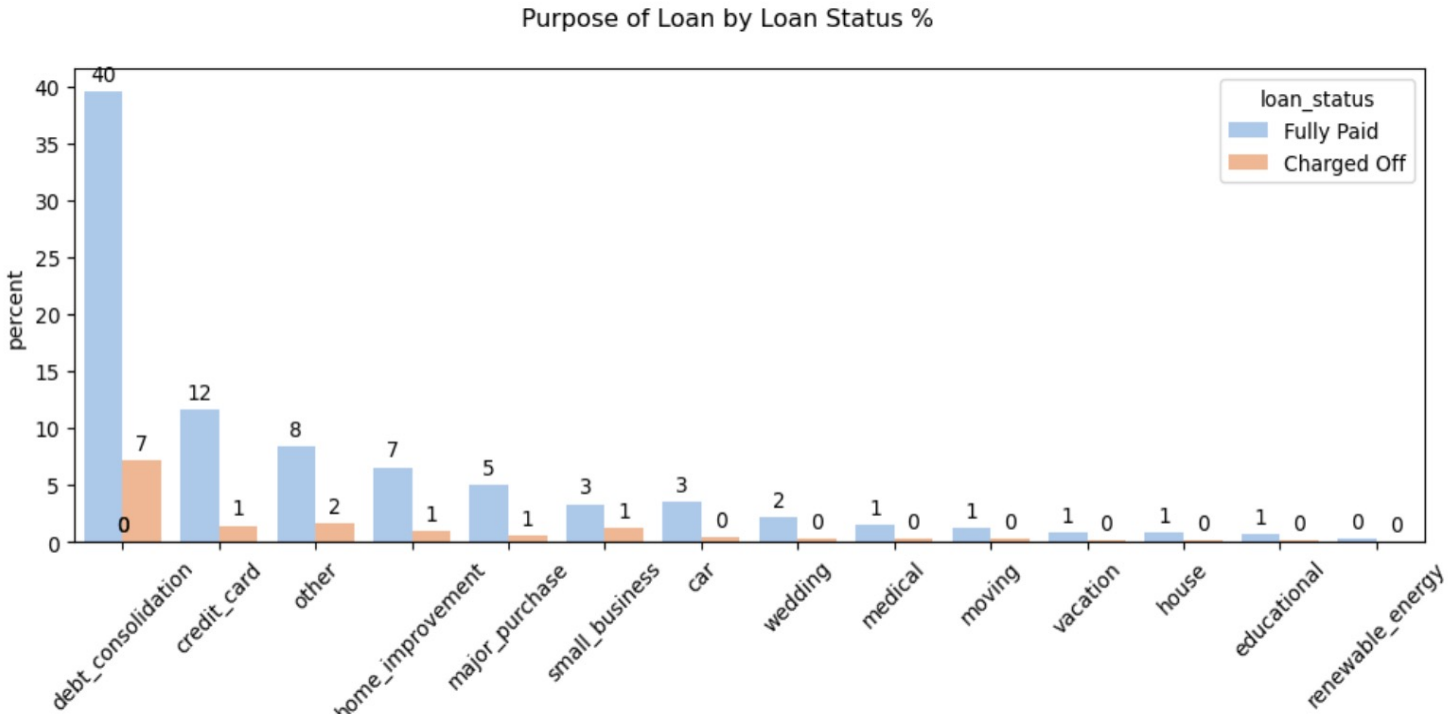


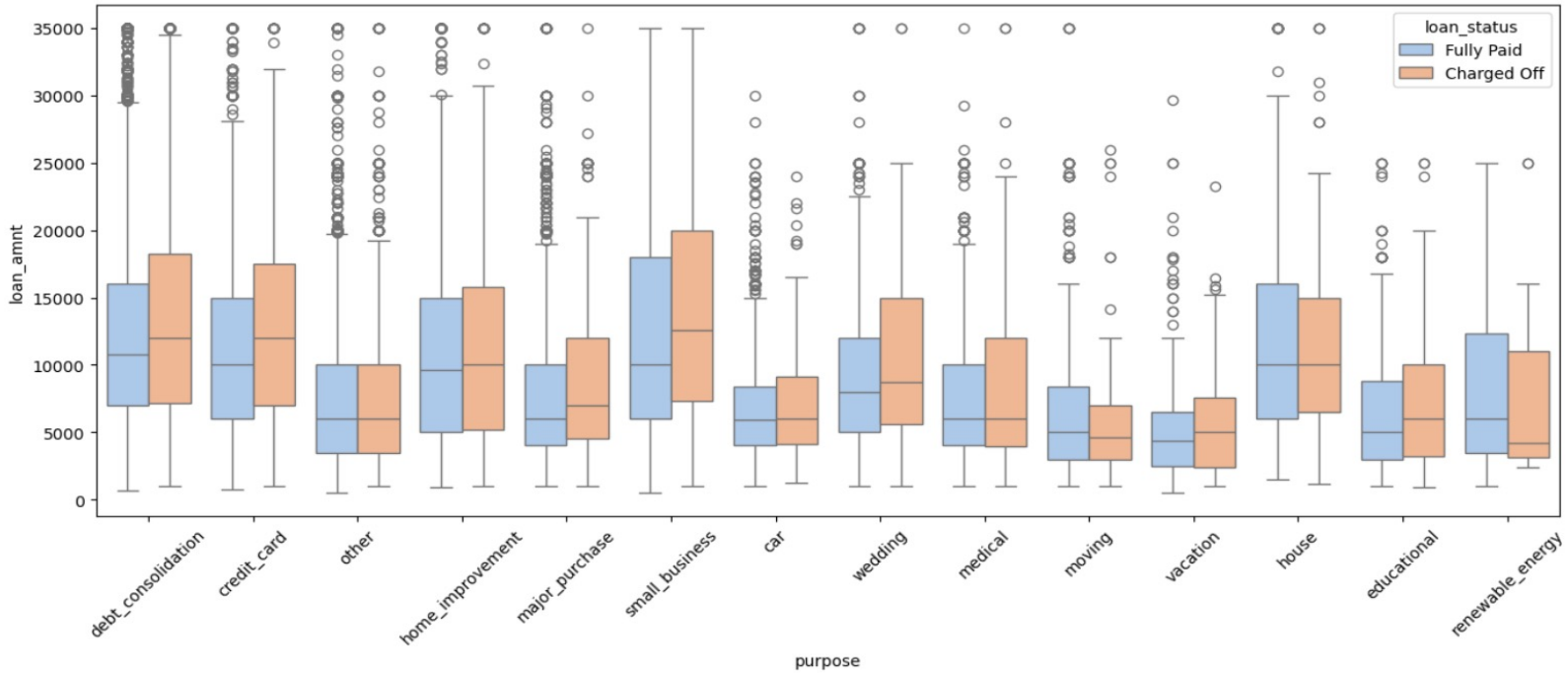
Bivariate Analysis

Observations

1. Out of total **47%**, **7%** of borrower who took loan for **debt consolidation defaulted**
2. highest loan amounts ranges are in **small business**, **debt consolidation & home improvements**

- highest probability of defaulting is **small business** but the volume is very small

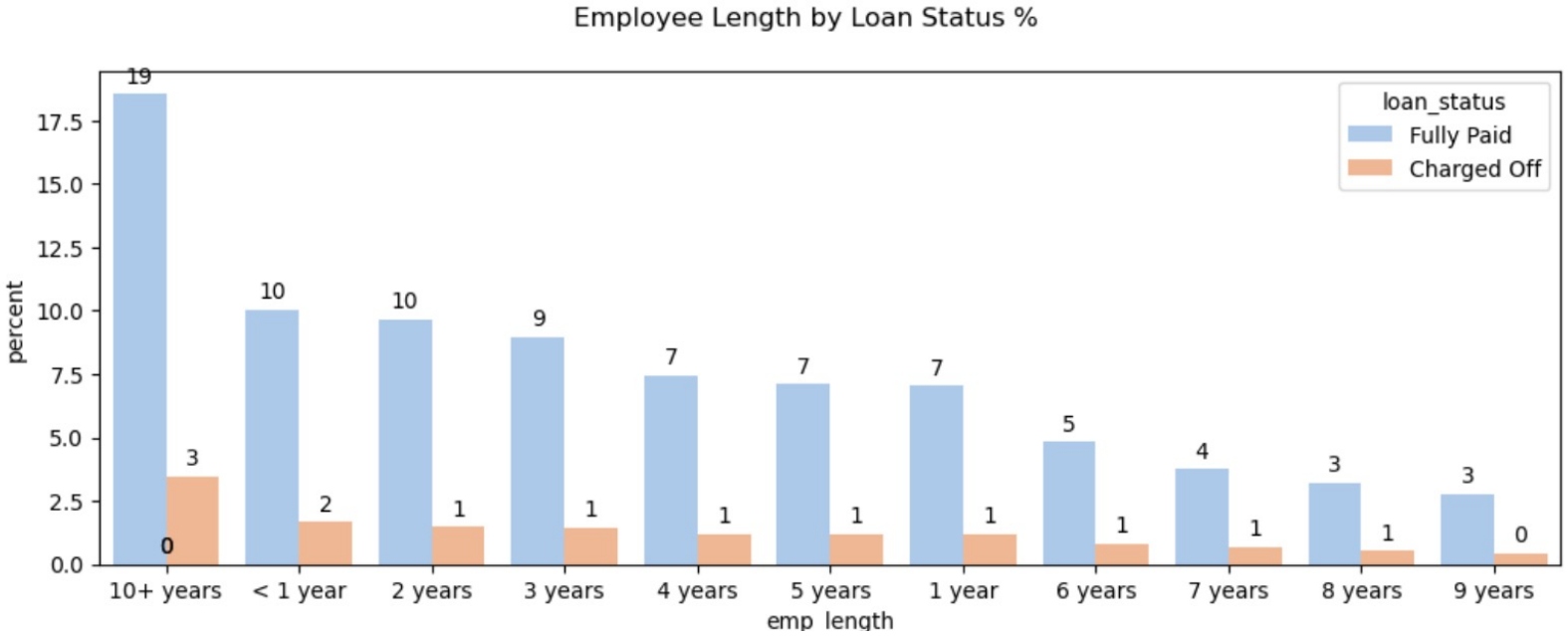


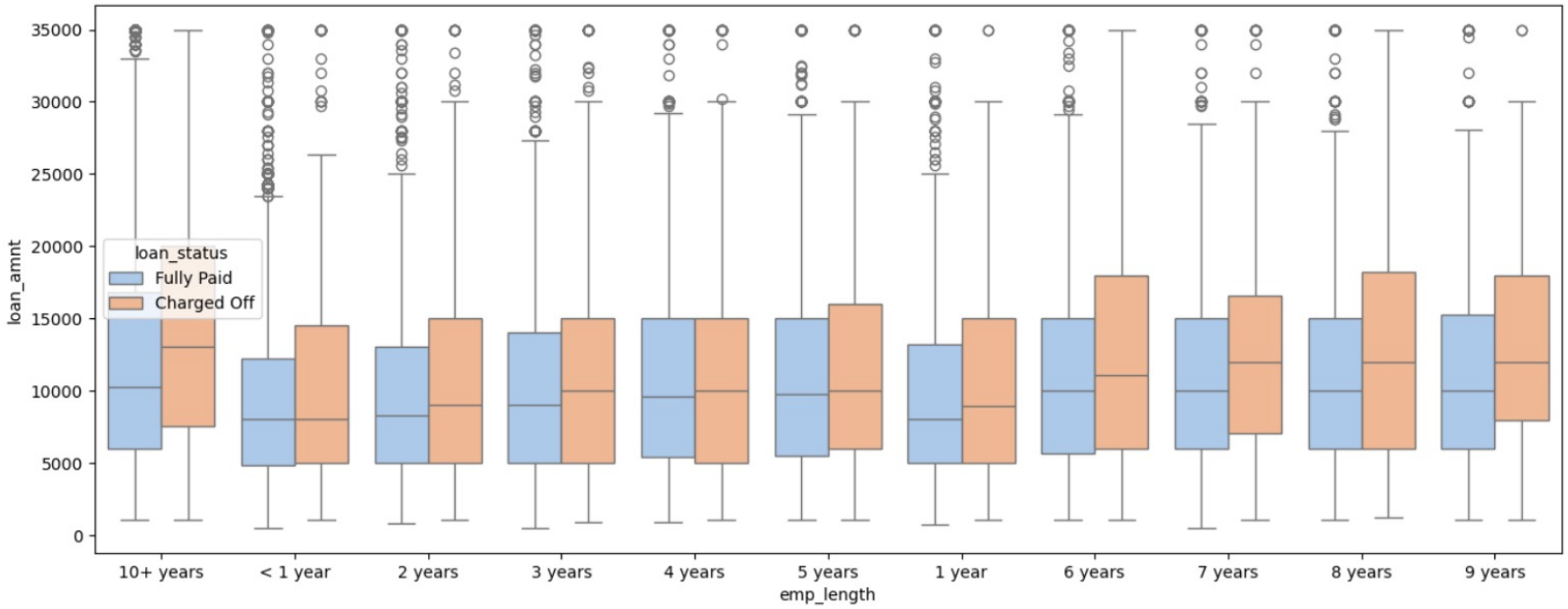


Bivariate Analysis

Observations

1. **Higher percentage-3%** of defaulter are with **10+ years** experience range
2. **Most of the defaulters** took higher loan amount in experience range of **6+ years**



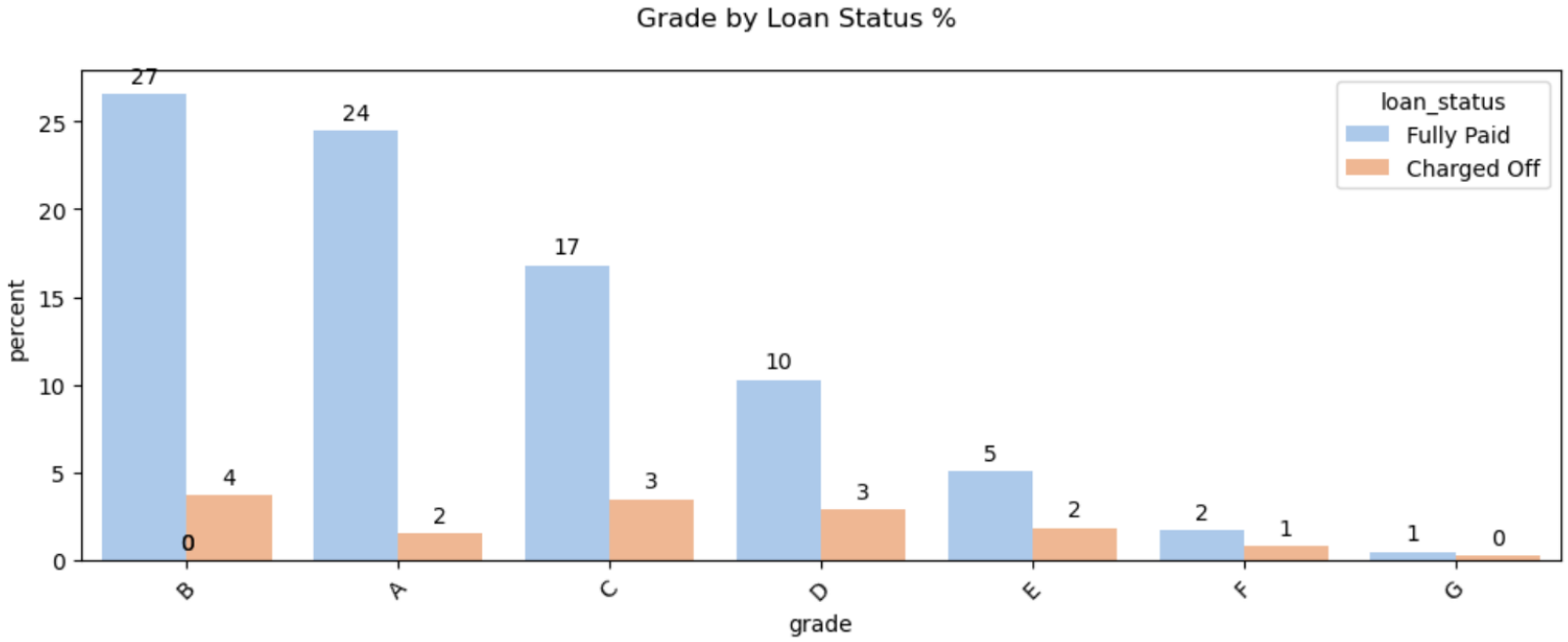


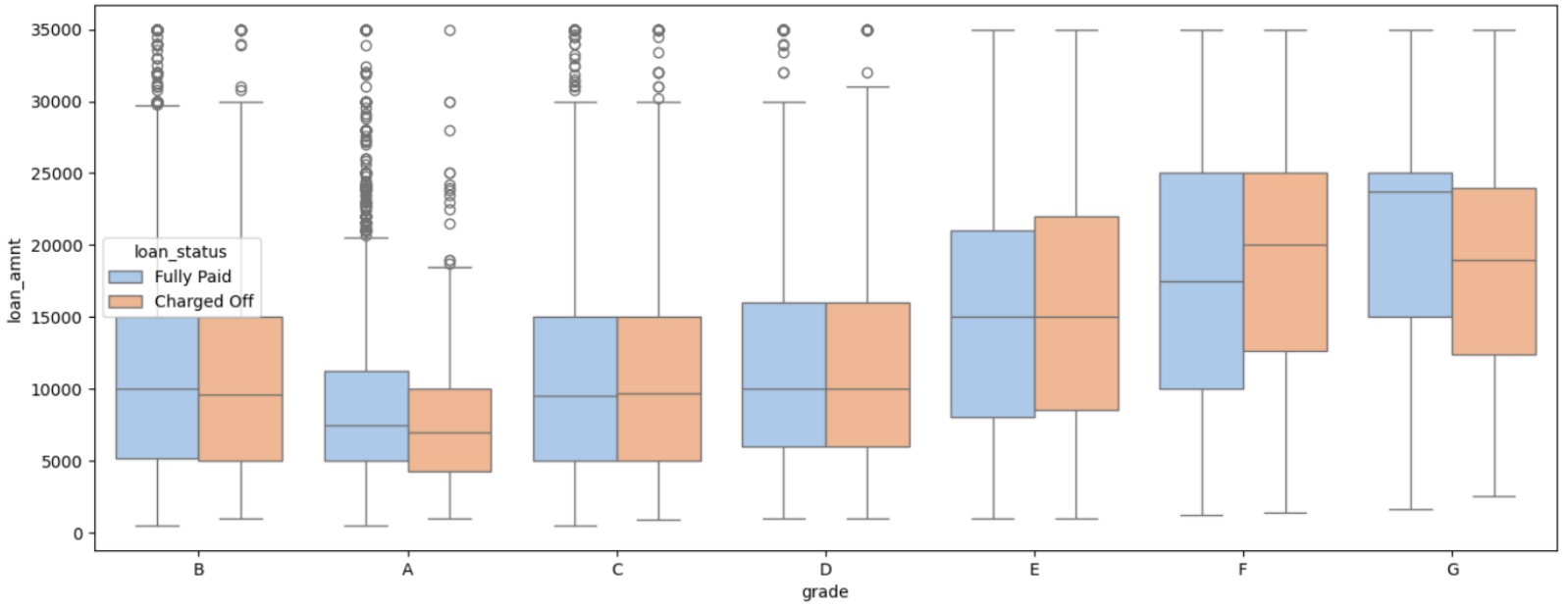
Bivariate Analysis

Observations

1. **Highest** percentage of defaulter are in **grade B followed by C &D**

- **Grade F, E & G's defaulters** are in **high loan** amour range. Means took higher amount of loan and defaulted



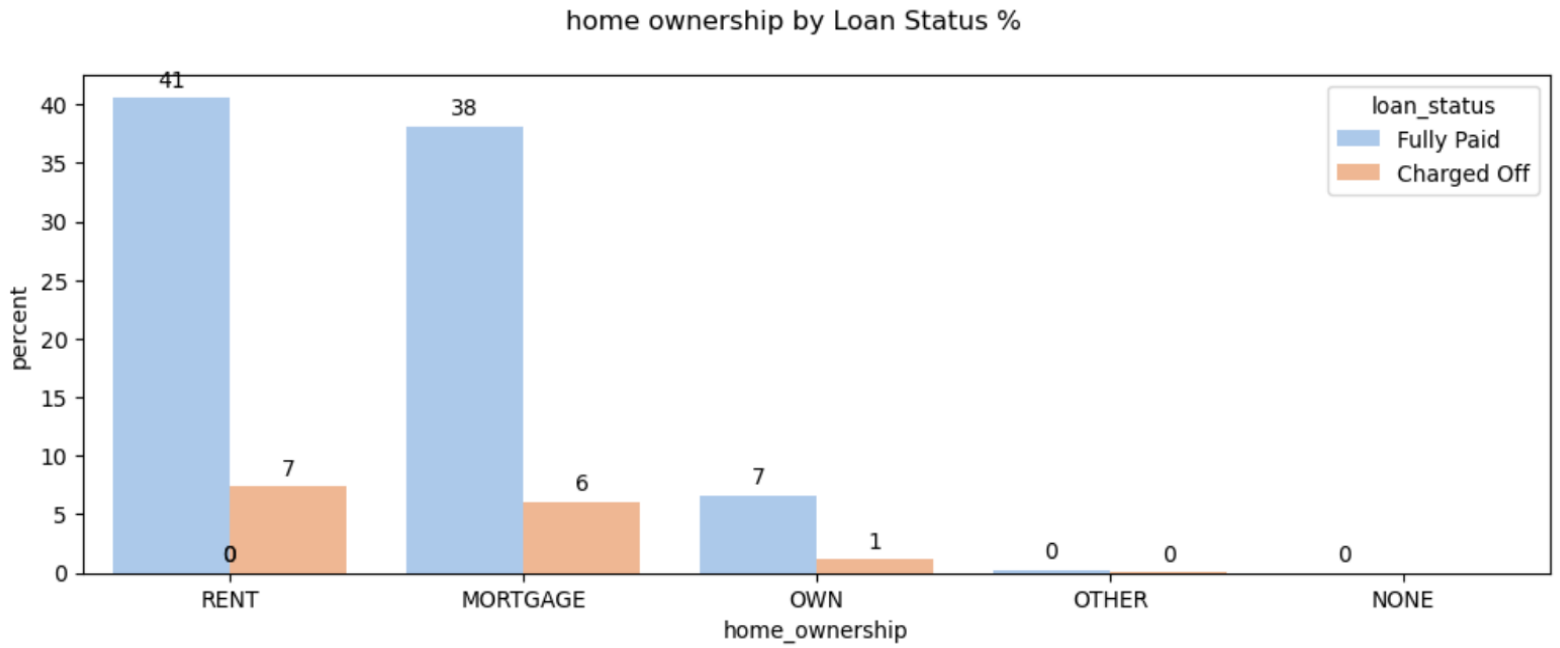


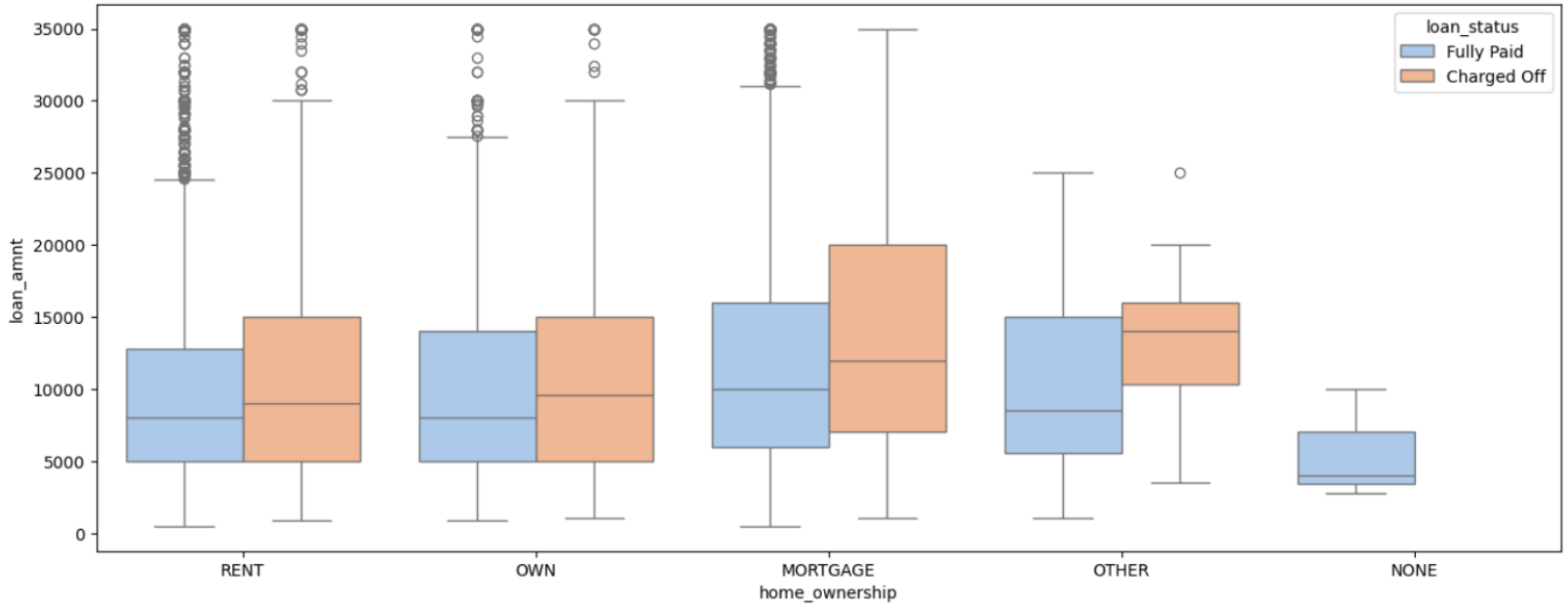
Bivariate Analysis

Observations

1. **Highest** percentage of defaulter are in **rental or mortgage homes category.**

- Defaulter under **mortgage** category took **highest** amount of loans

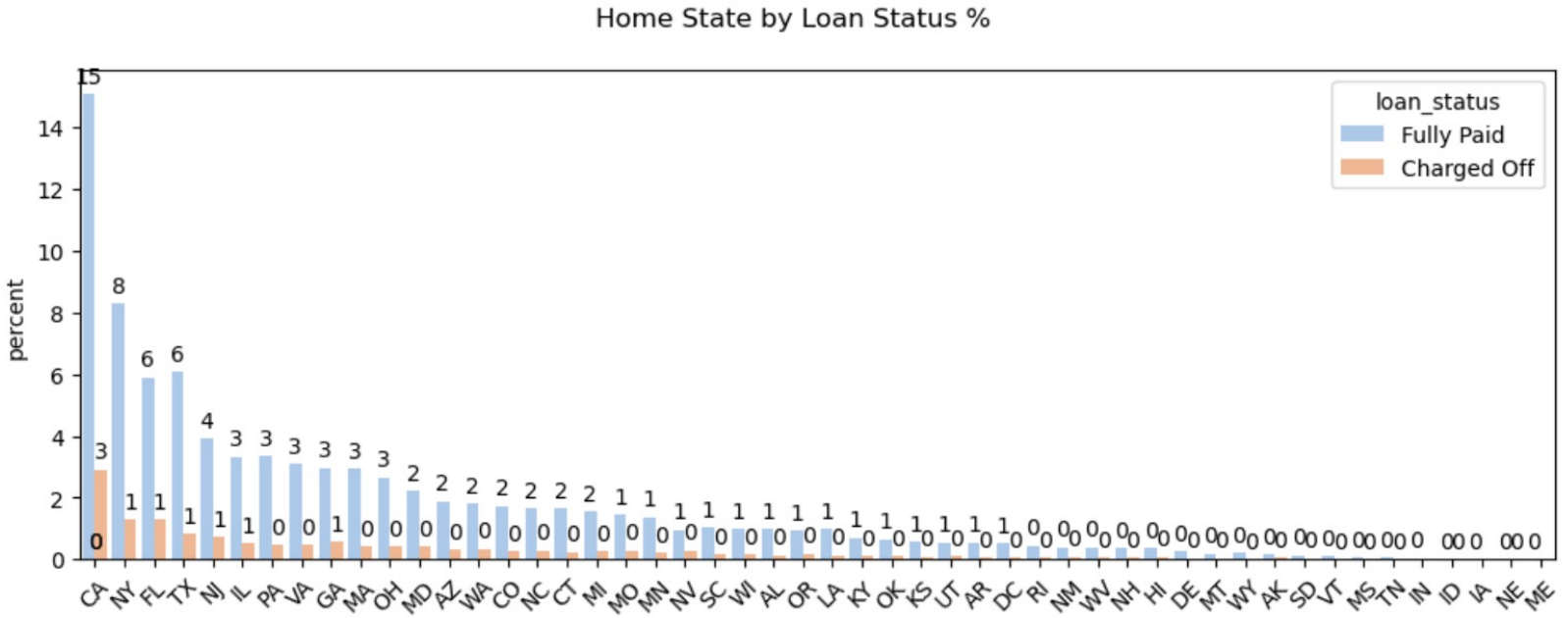


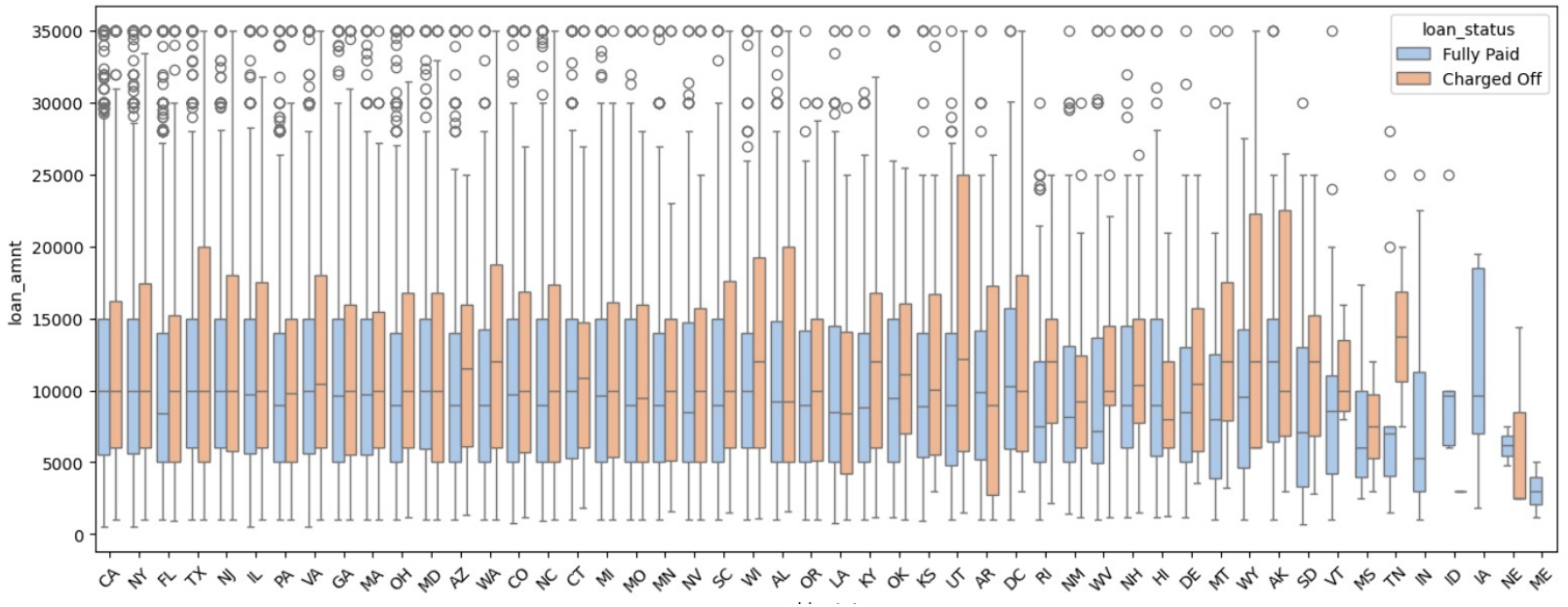


Bivariate Analysis

Observations

1. **Highest** percentage of defaulter stays in CA - California
- The probability of defaulting for a loan is more in following states **UT, AK, WY, AR**
-

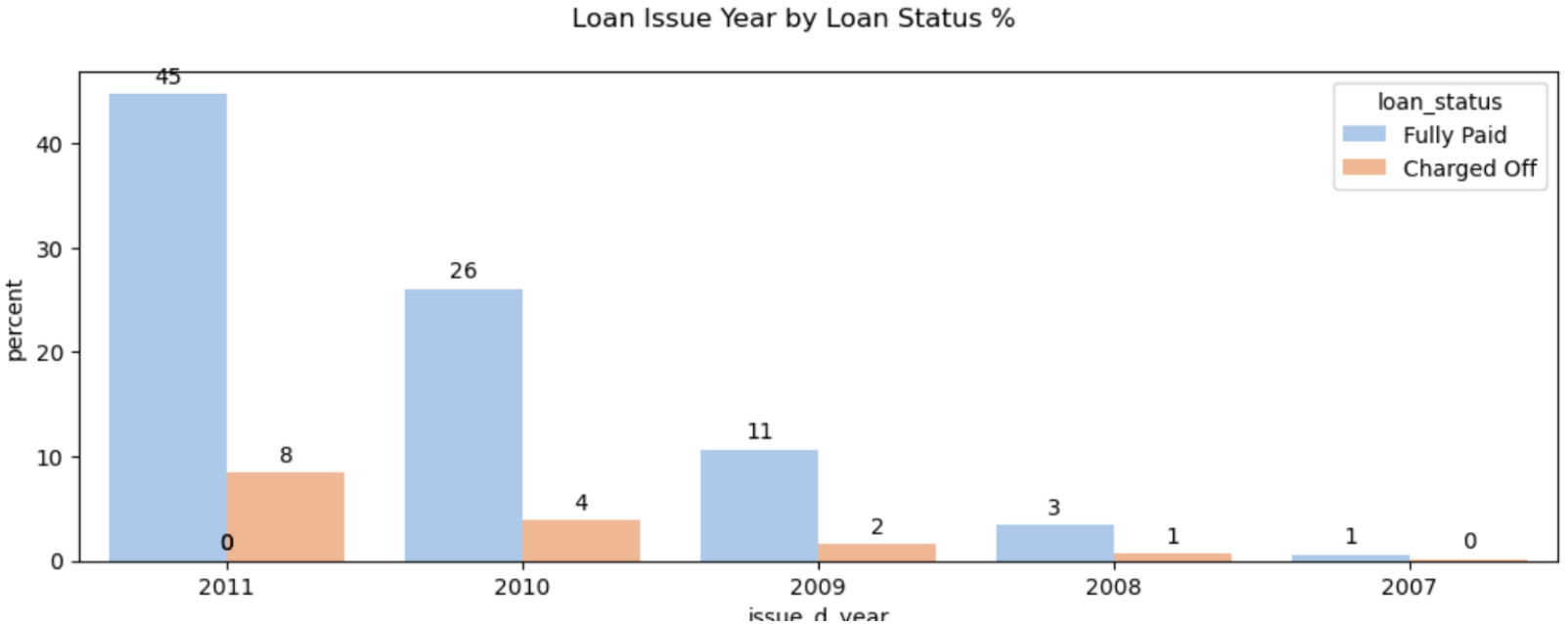


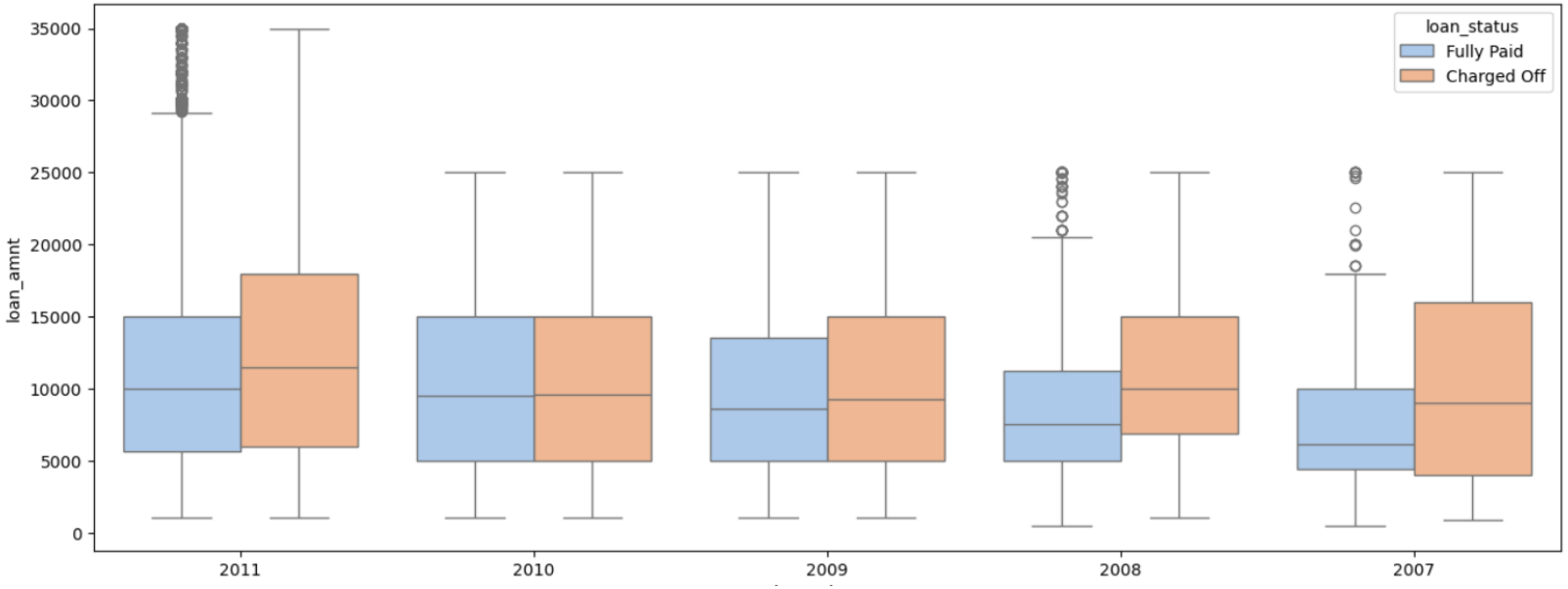


Bivariate Analysis

Observations

1. Most of the **default** happened when loan was taken during **2011 & 2010**, But the volume of load issues during those years were also high
- **Most of the default** happened for the loans taken during **2007 followed by 2011**, This the same duration when Big recession had hit USA
-



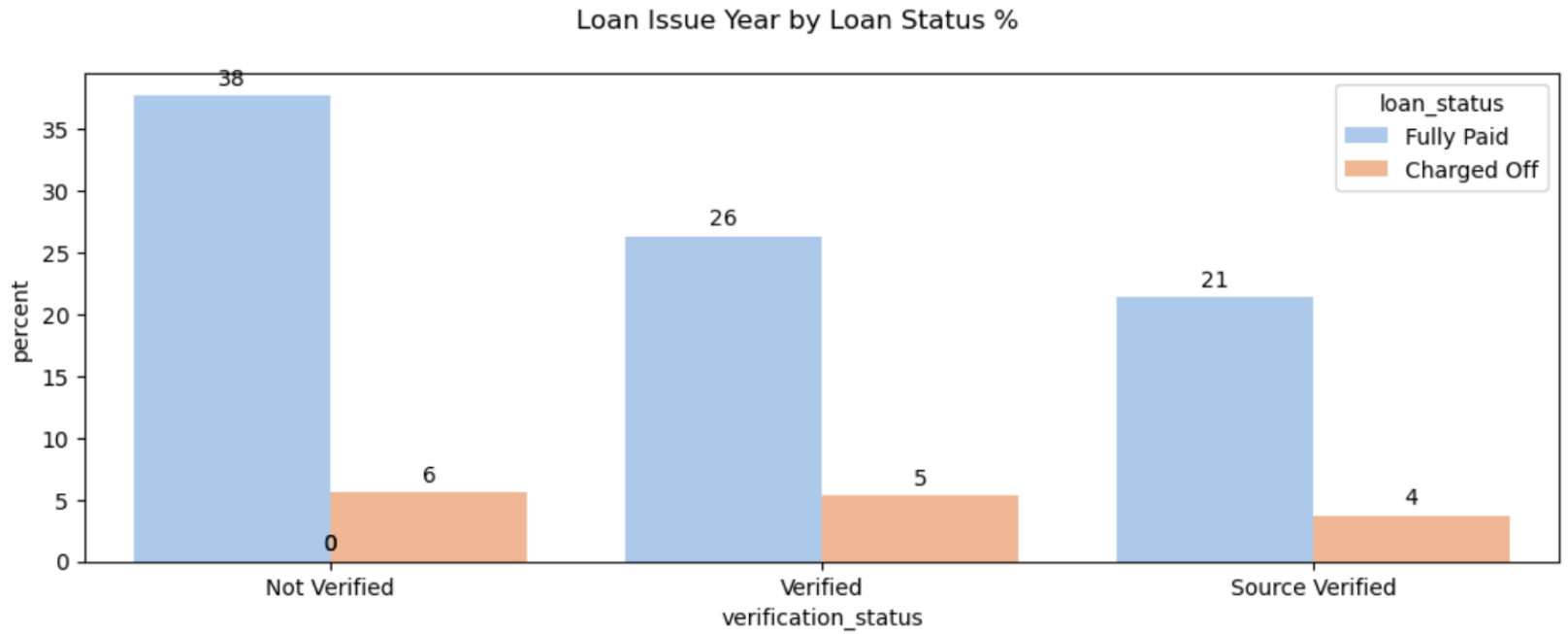


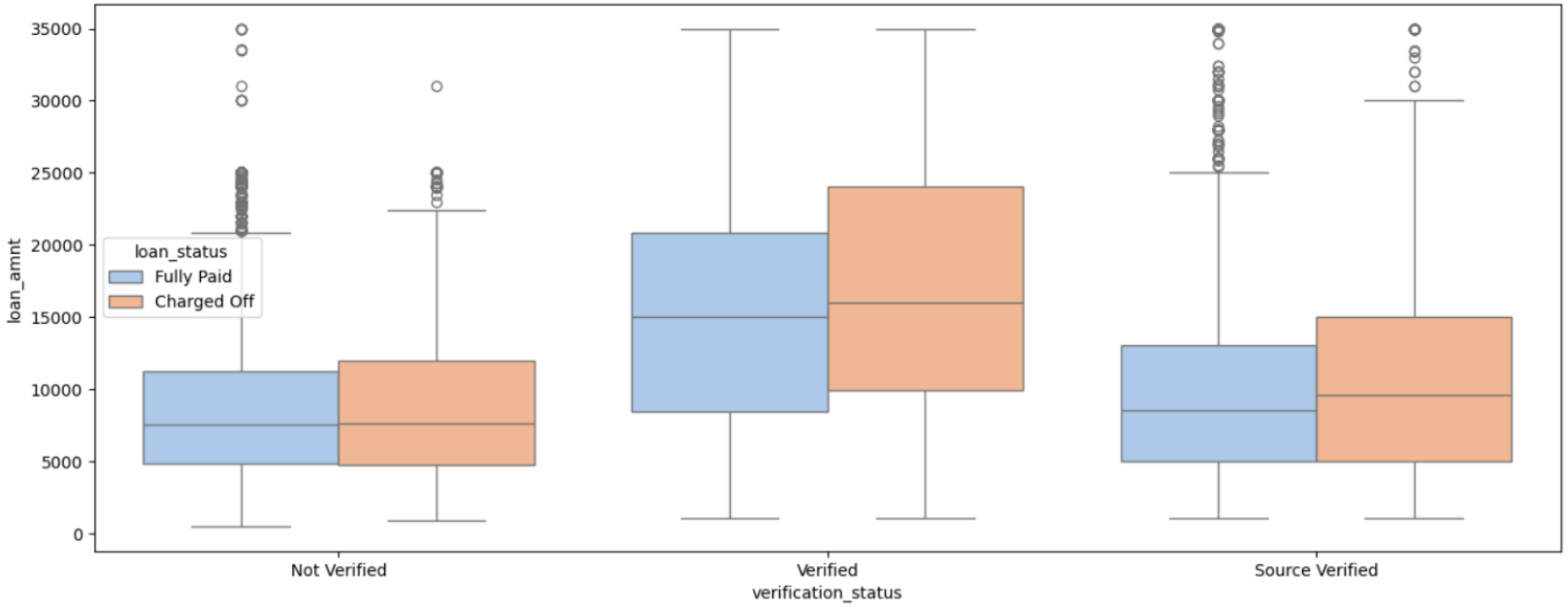
Bivariate Analysis

Observations

1. There is **not much difference in default rate** based on verification status
2. **Verified status** are given **more loan** compare to not verified

•



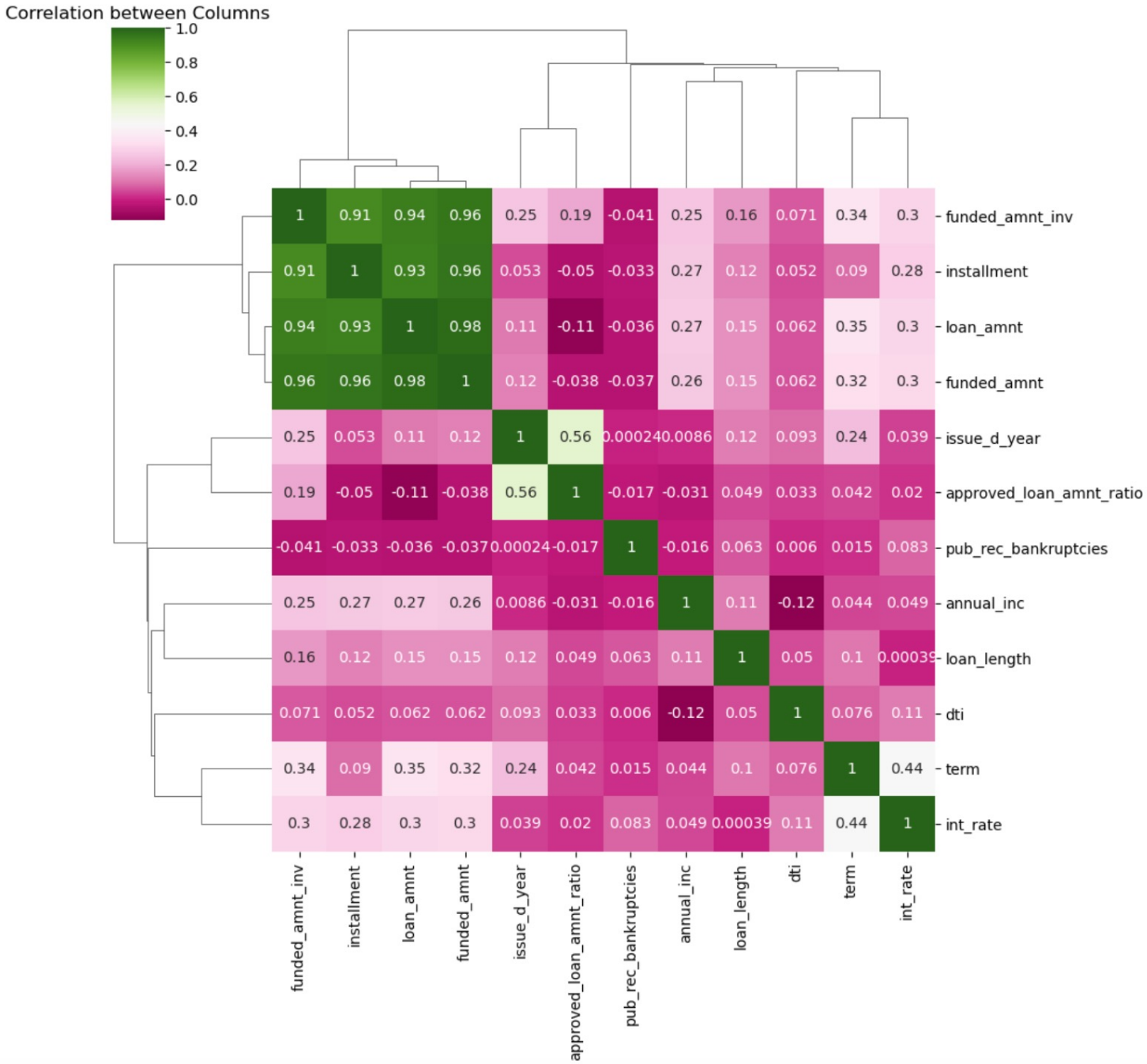


Bivariate Analysis

Observations

1. **Loan Amount** has strong corelation with **Funded Amount** & **Instalment**

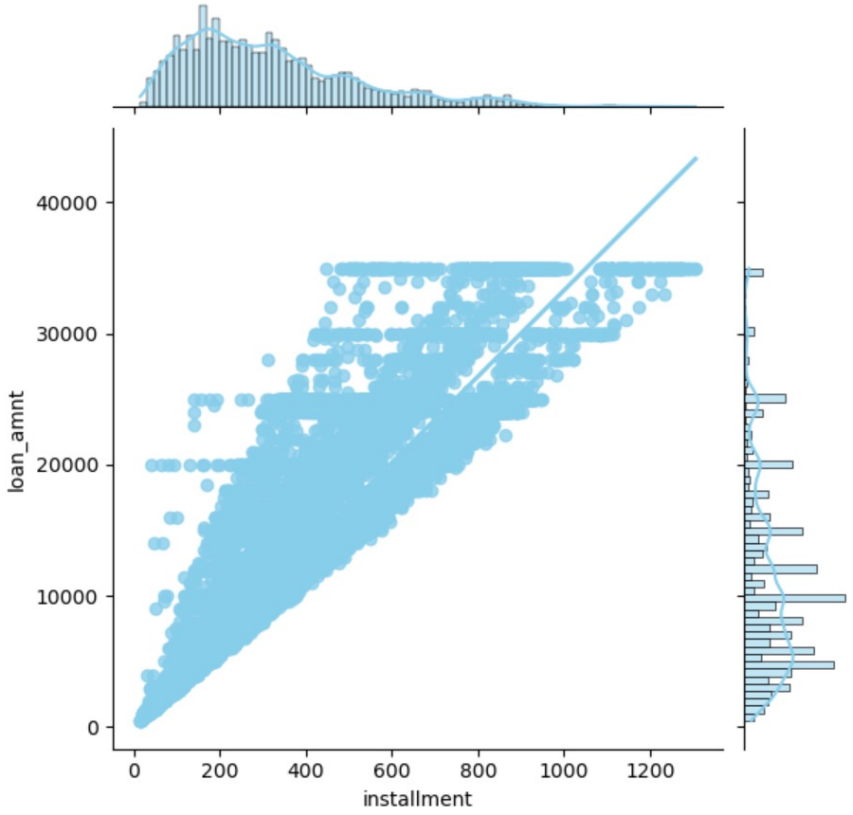
1. Rest of KPI's are not having any strong positive or negative corelations

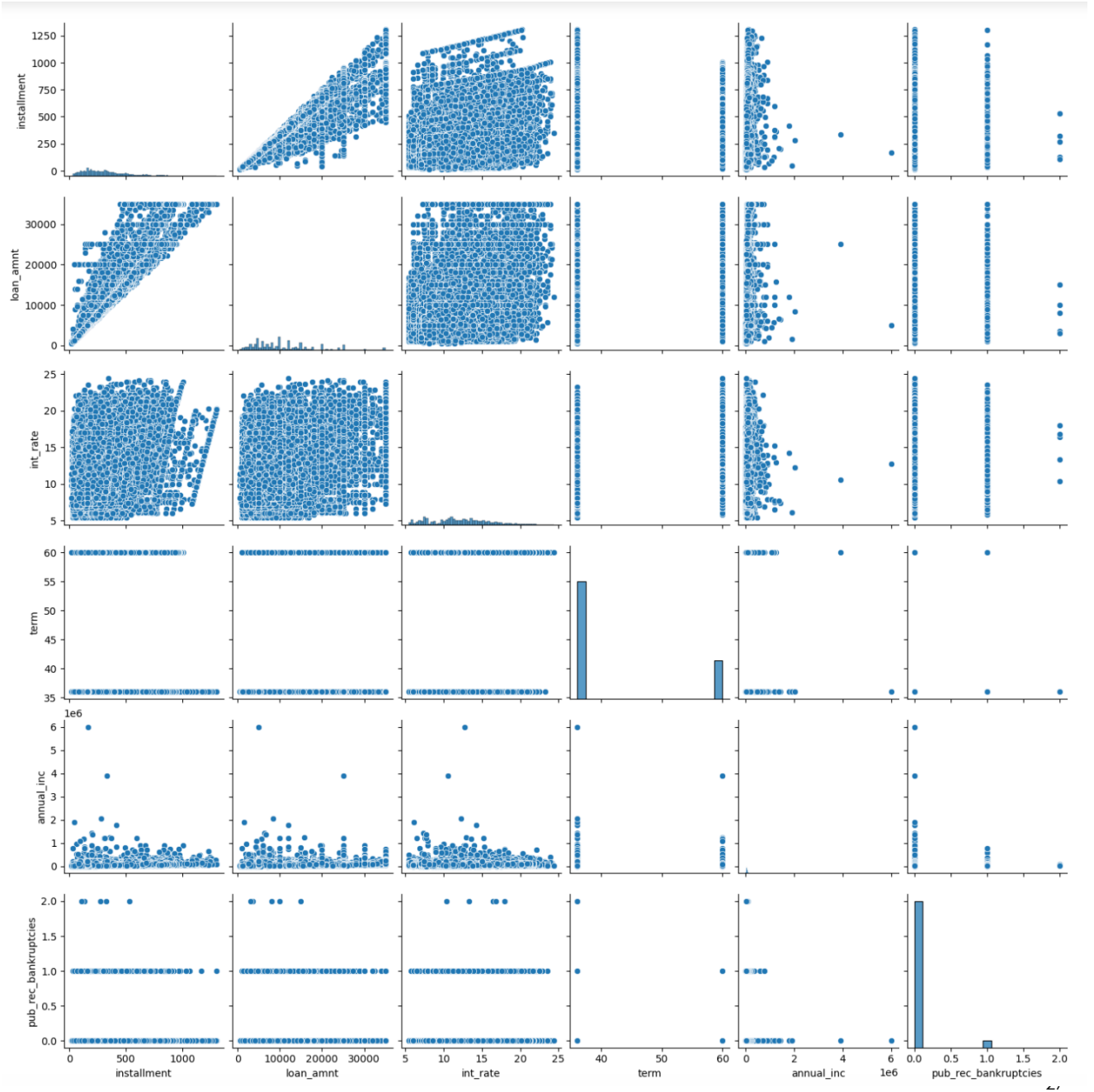


Bivariate Analysis

Observations

- There is high positive corelation between Loan amount and Instalment
- Rest attributes are not showing any strong corelations





Recommendation

There is a more probability of defaulting when

- ✓ Applicants who use the loan to clear other debts, means purpose is debt consolidation
- ✓ Applicants having house ownership as RENT or Mortgage
- ✓ Applicants with employment length of 10+
- ✓ Implement strict verification criteria for grade B, C & D
- ✓ Evaluate the risk associated with 60-month loans with higher loan amount, as there is likelihood to defaults
- ✓ Ensure we are efficient scalable background check process during peak time of Q4
- ✓ Carefully evaluate all debt consolidation loans and keep release only low loan amount for shorter period of time
- ✓ Also take into account which state the loan application are coming from like CA & NY states have higher tendency to default
- ✓ Review applications for borrower who high interest rate category with higher loan amount

Thank you

GitHub link