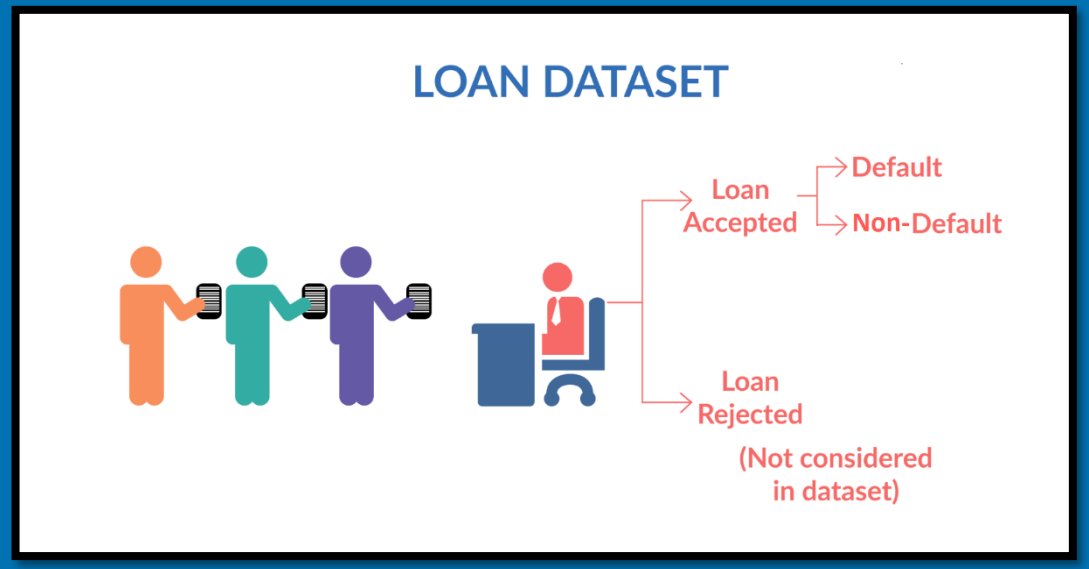
Lending Club Case Study

Exploratory Data Analysis Project



By:
Leander Antony
Venkata Saikrishna Dussa

Index:

Overview

Background

Consumer Finance Company lends various types of loans to Urban Customers. When the company receives a loan application, the company has to make a decision on loan approval based on applicant's profile. We will solve this decision problem using **Exploratory Data Analysis** using **the past Customers Data** (Provided in the CSV File)

Problem Statement:

We need to minimize the below risk to avoid loss of Business for Consumer Finance Company.

- Approving loans to applicant **"Likely to repay the Loan"**
- Rejecting loans to applicant **"Not Likely to repay the Loan"**

Solution:

We need to analyze the Provided Data and draw the insights for various Features.

Tools Used

Tools

- Python
- Jupyter Notebook
- Git

Libraries:

- pandas
- numpy
- matplotlib
- seaborn

Approach Used

- Data Understanding
  - Overlooking the Dataset
  - Removing Total missed Columns

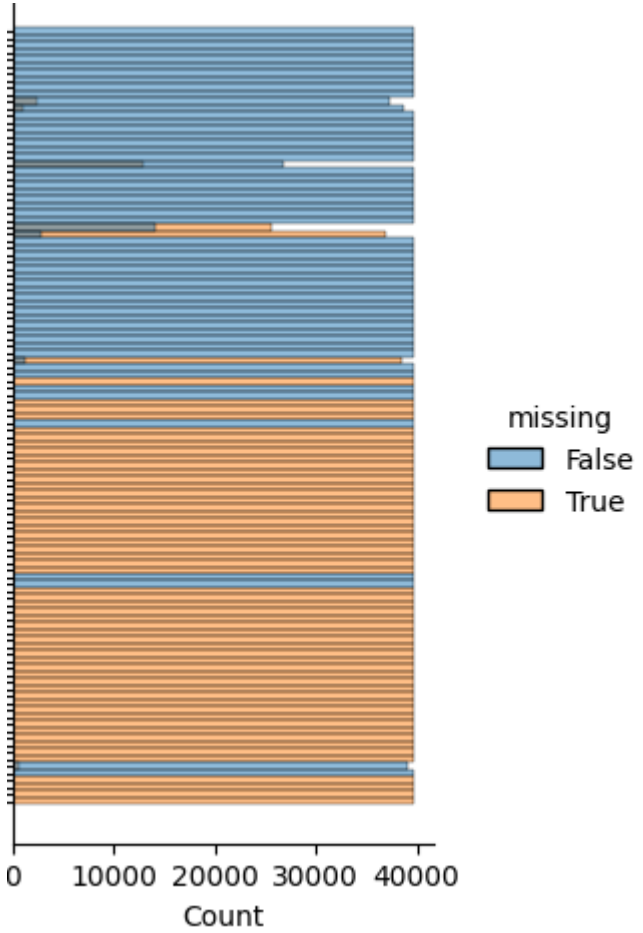
- Data Cleaning and Manipulation
  - Outlier Detection
  - Removing Outliers Data
  - Extracting New Columns

- Data Analysis
  - Univariate Analysis
  - Derived Metrics
  - Bivariate Analysis

- Conclusions

Data Understanding

- Based on shape (39717, 111), we are provided with large Dataset which has **text data** and **numerical data**
- There are more missing values in the Dataset as per “displot”
- We will exclude the columns which has more than 4000 rows

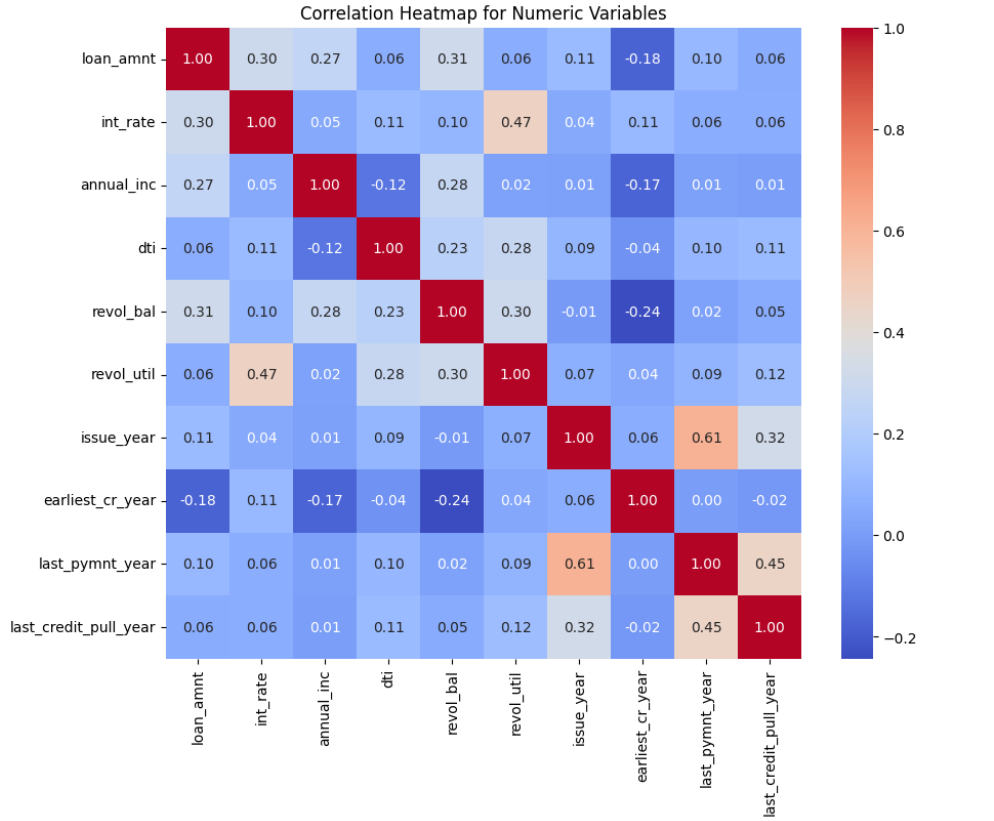


Data Cleaning and Manipulation

- **Interest_Rate** and **Revol_Util** column are having “%” which makes it “String”, we are converting it into Numerical Data.
- Converting Date related columns into proper “datetime” columns.
- Rounded off “installment” and “total_payment” columns.
- Removed unrelated data which are not yet processed “loan_status” = “Current” and “home_ownership” in [“OTHER” and “NONE”]
- Evaluating again “Missing Counts”
- Extracted “year” data from “issue_year”, “earliest__cr_year”, “last_pymnt_year”, “last_credit_pull_year” columns.

Data Cleaning and Manipulation -Correlation

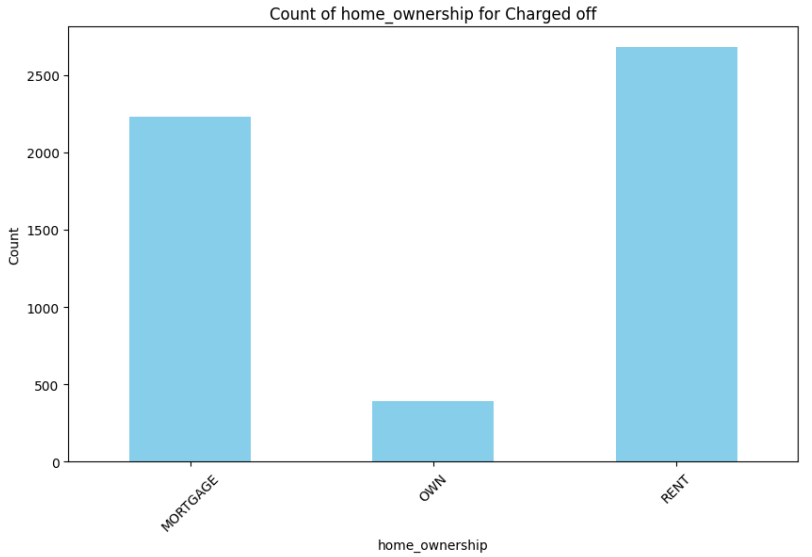
- Evaluated Correlation for Numerical Columns.
- We have observed that very less correlation between taken columns.
- So we will finalize the columns to proceed with Data Analysis.



Data Analysis – Univariate Analysis (Charged Off)

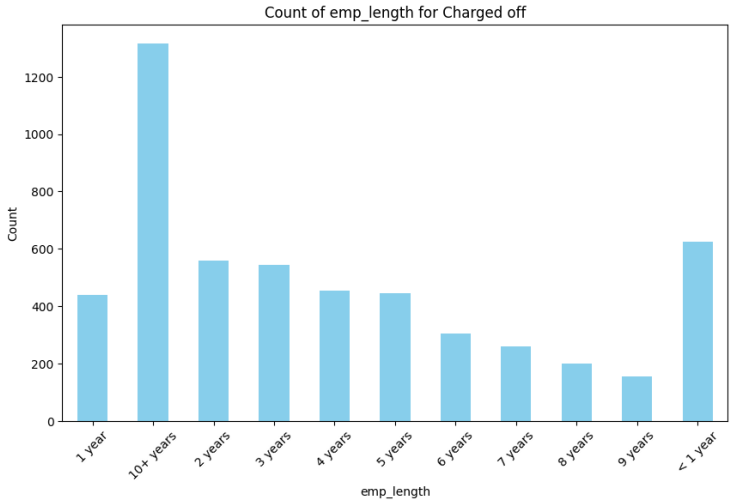
Home Ownership:

- **OWN** Home defaults least.
- **MORTAGE** and **RENT** are taking more loans.
- **RENT** borrowers default more than **MORTAGE**



Employee Length:

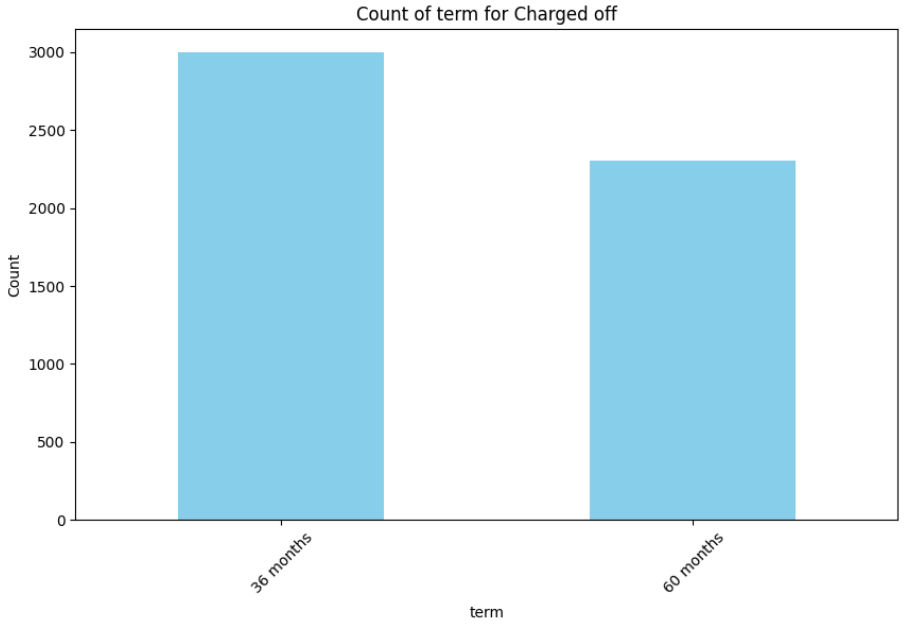
- **<1 Employees** are more **RISK** to give loan.
- As years **increasing**, employees defaults is **DECREASING**.
- **10+ years** contains all more than10 year employees, so it gave cumulatively higher number



Data Analysis – Univariate Analysis (Charged Off)

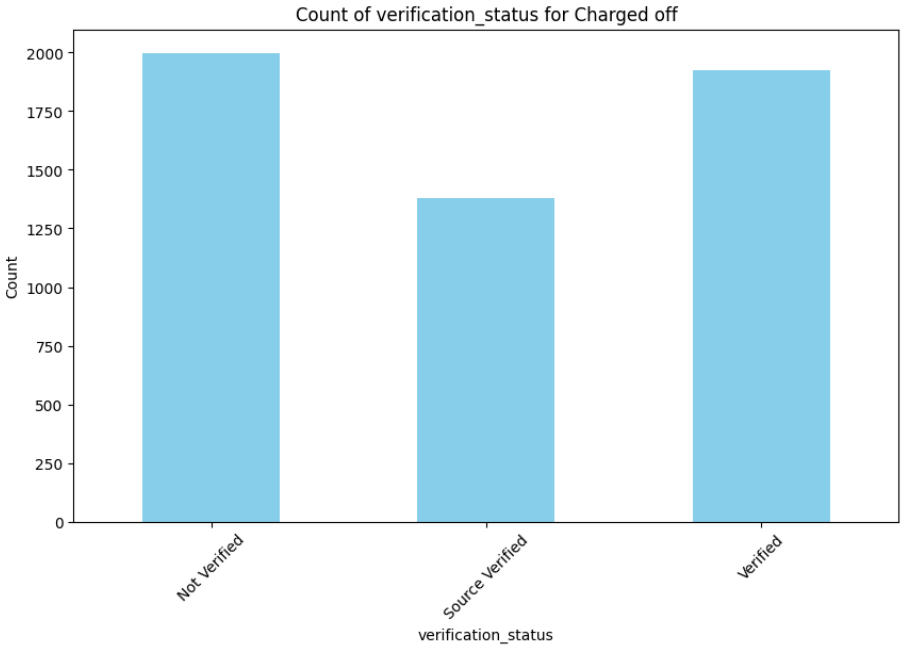
Term:

- **More time Term(60 months)** are defaulting Loan less compared to **Less time Term(36 Months)**



Verification Status:

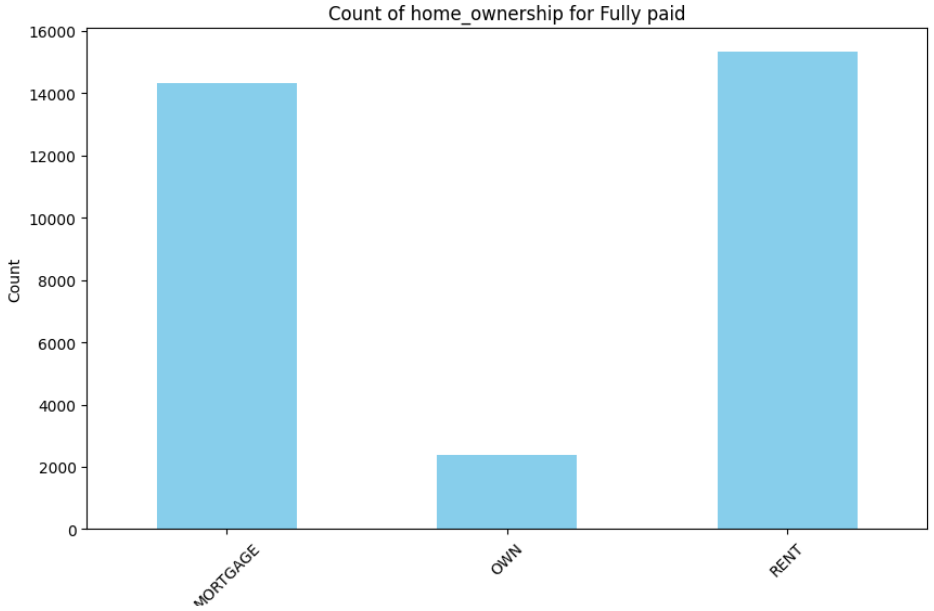
- Lender is doing more verification, so there is no much difference for Loan Default for **Verified and Non Verified** Customers.
- **Source Verified** are defaulting less compared to others.



Data Analysis – Univariate Analysis (Fully Paid)

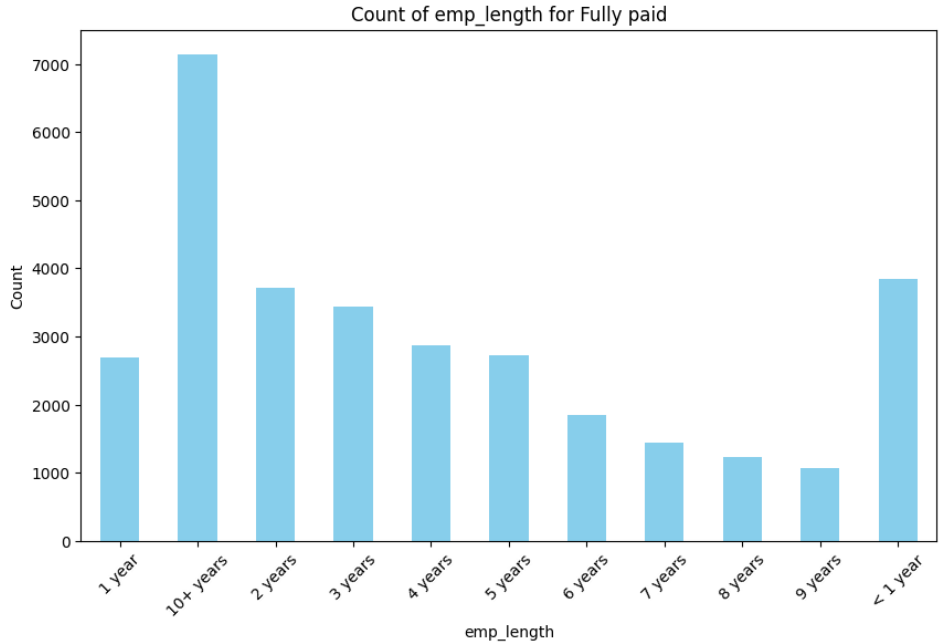
Home Ownership:

- Earlier Assumption is correct, **OWN** house customers are very less, which is making less **Fully Paid**
- **MORTAGE and RENT** are equally **Fully Paying** the loan



Employee Length

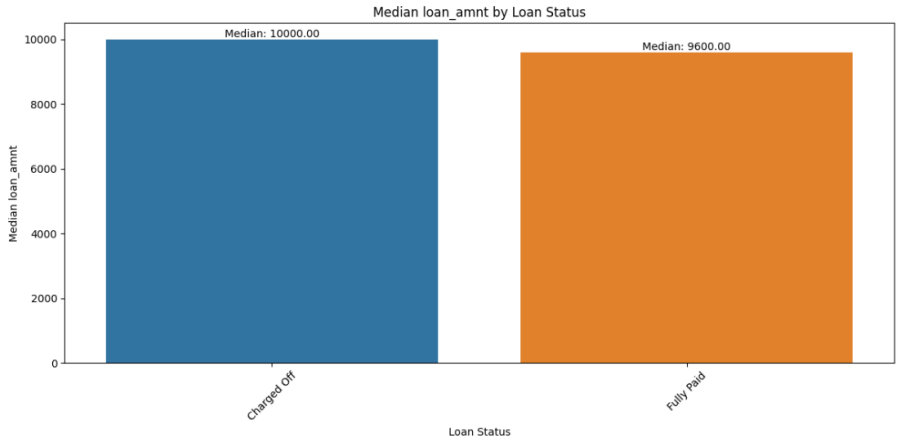
- Earlier assumption is incorrect, **<1 year** are paying more, they are not riskier too.
- More Job Applications from **0-5 years**



Data Analysis – Derived Metrics

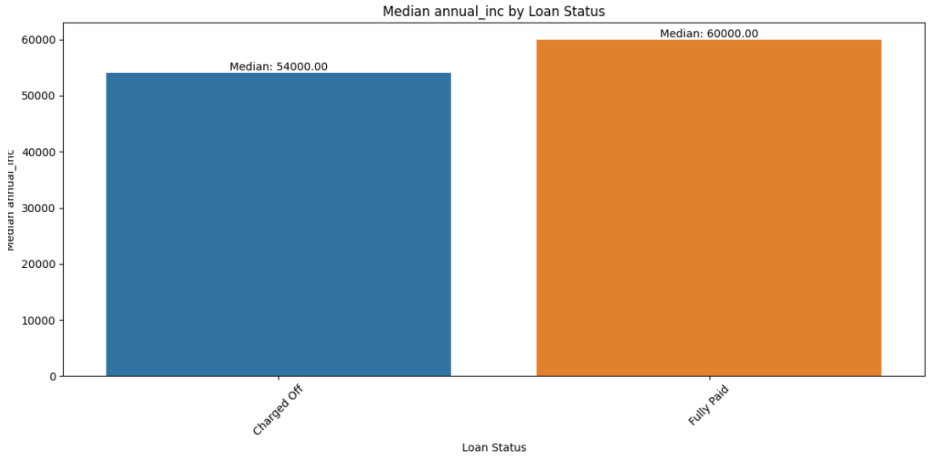
Loan Amount:

- No Significant difference between "**Charged off**" and "**Fully Paid**" for "**Loan Amount**" median. So we assume, *"**Loan Amount**" is not relevant to analyze.*



Annual Income:

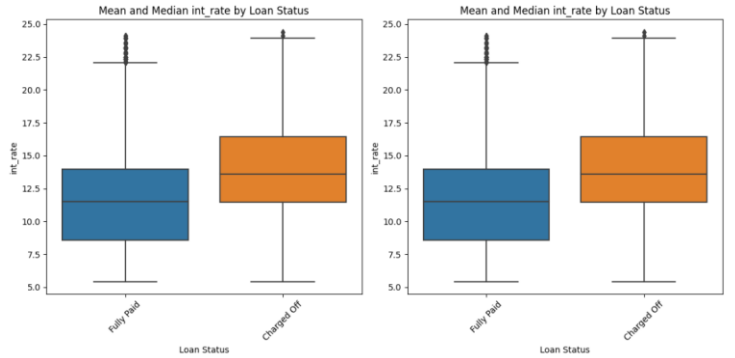
- **Fully Paid** borrowers have a higher income than the Defaulters by around **11%.**



Data Analysis – Derived Metrics

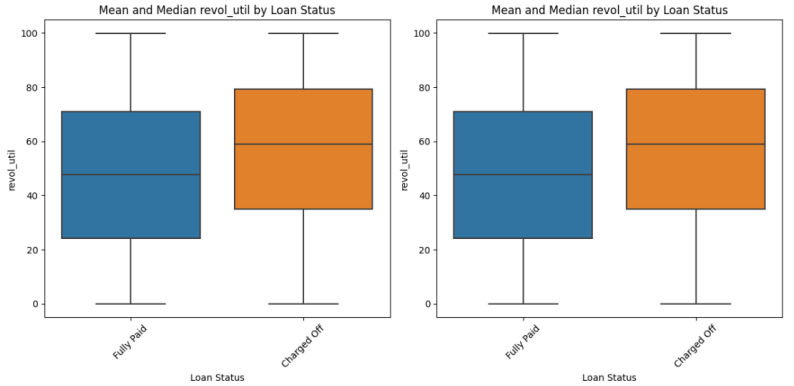
Interest Rate:

- Median Interest Rate of defaulters is nearly 20% more than the interest for the other borrowers



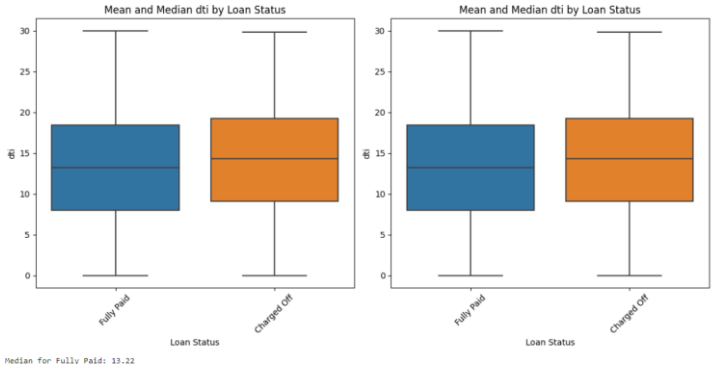
Revol Utilisation:

- The revol_util Median for defaulters is more than 20% higher than those who repaid the loan successfully.
- The defaulters have already used up close to 60% of their existing credit lines



Debt to Income Ratio:

- Not much difference between the defaulters and the other borrowers.
- **Loan Amount and Annual Income** don't play a major role in **identifying potential defaulters**



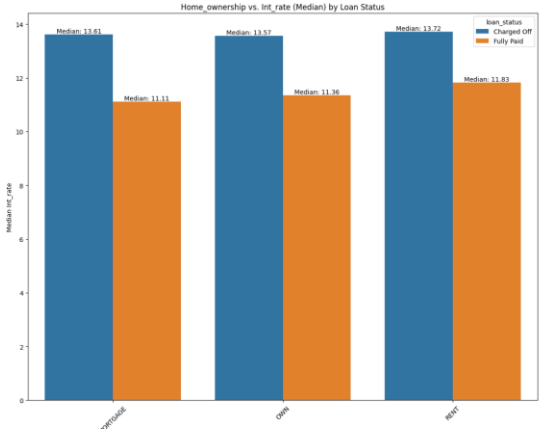
Data Analysis – Bivariate Analysis

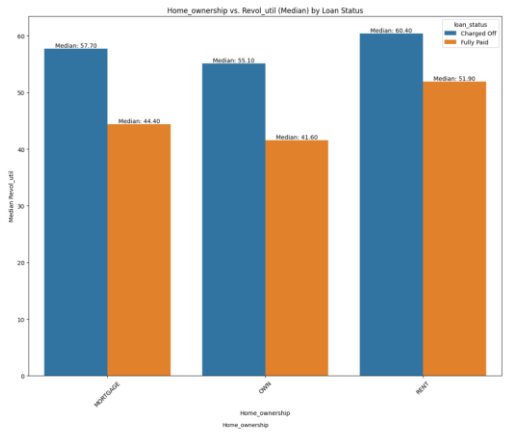
Home Ownership:

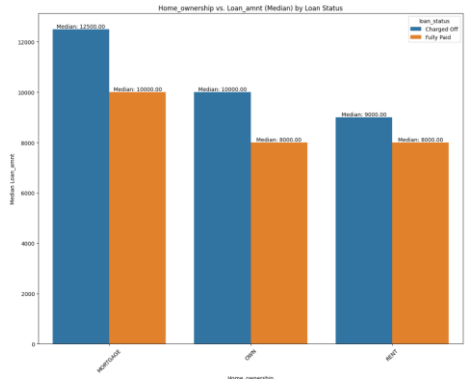
- Defaulters are suffering from Highest Interest Rates in all categories of Home Ownership.

- 55% in Revol Utilisation seems to be the median cutoff point above which the borrowers starts defaulting, also borrowers who rent a house are using more of their available credit lines.

- The loan amnt given to mortgaged defaulters is comparably high coupled with **high interest could be a cause for defaulting**



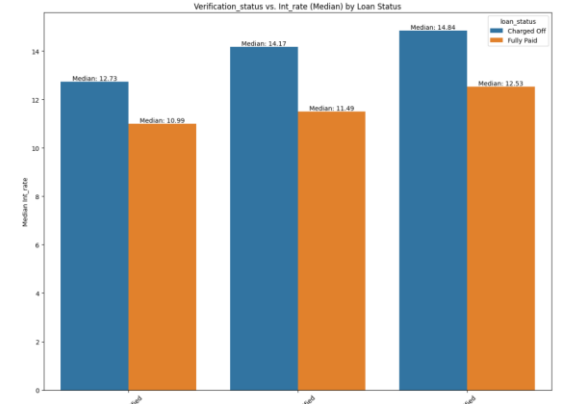


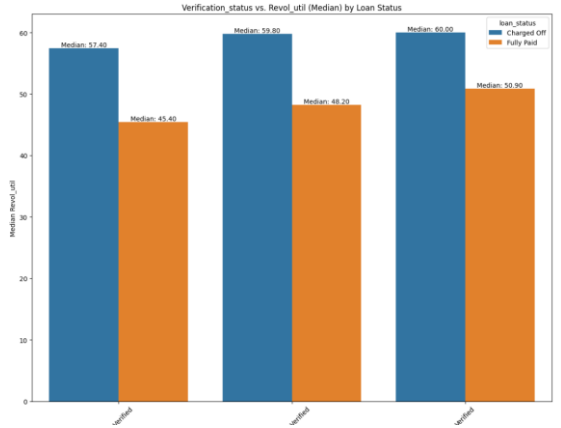


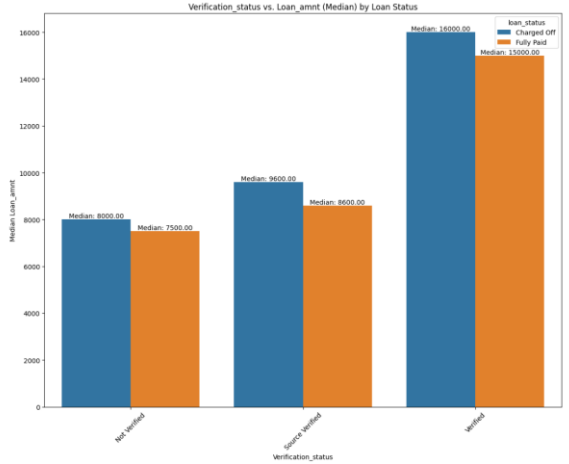
Data Analysis – Bivariate Analysis

Verification Status:

- The verified borrowers who have defaulted are having to face a median interest of 15% which is very high.
- The revol utilisation rate of the defaulters here is more or less the same. So it can be concluded to not play a major role in defaulting.
- The loan amount given to verified borrowers ideally should be comparably high but here its 100% more than the amount sanctioned for Not Verified borrowers.
- Loan Amount coupled with high interest is causing a huge number of defaulters eventhough they are verified.





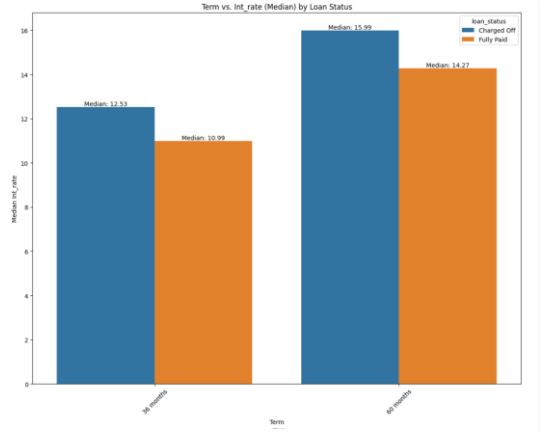


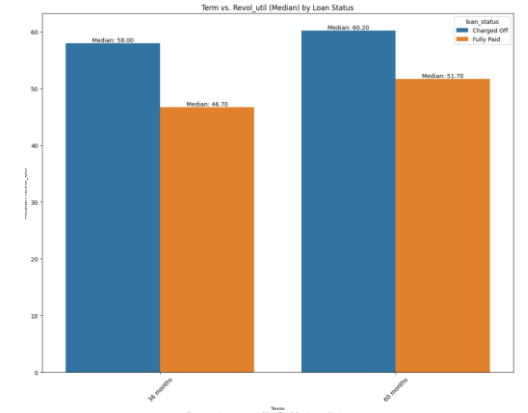
Data Analysis – Bivariate Analysis

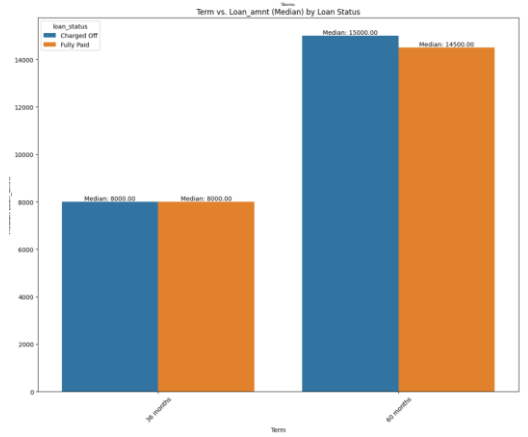
Term:

- The median Interest Rates are 16% for 60 months and the loan amount sanctioned is nearly 90% more than loan tenure of 36 months.

- **The longer termed loans have a higher sanctioned amount coupled with the high interest could be causing default**.







Conclusions

- Major Customers are “**Mortgaged” and “Rent”** House Customers having **1 to 5 years of experience**.
- More customers are preferring **36 Months plan and paying the loan Successfully**.
- Increasing Tenure period and Reducing Loan Amount with Interest has **less chances to be Defaulted.**
- Loan Verification is not showing big impact on Fully Paid.
- As **Interest Rates are increasing, there is a good chance for Default**.
- **55% in Revolving Line Utilization** seems to be the median cutoff point above which the **borrowers starts defaulting.**
- Revolving Line Utilization, Loan Amount and Interest Rates are the driving variables to evaluate the Loan Default.

THANK YOU !