Lending Club Case Study

Discussion topics

- Problem statement – Introduction
- Analysis of the data frame
- Data cleaning process
- Treating outliers
- Analysis of data elements
- Closing Comments

Problem statement – Introduction

- Identify attributes which are having influence on the loan repayment
- Bad loans can be avoided by looking at the trends

Analysis of the data frame

- Data set size => ~33MB
- Number of Columns => 111
- Number of Rows => 39717
- Number of Columns having more than 40% of NaN values => 57 => Dropped them
- Number of Columns having only Unique values => 9 => Dropped them
- Post cleanup left out with 45 columns
- Walk through of the data present => 6 more columns dropped
- Number of Columns having missing values => 6
  - Missing values imputed for 5 columns with mode (Categorical type)
  - Dropped rows for missing values in 1 column

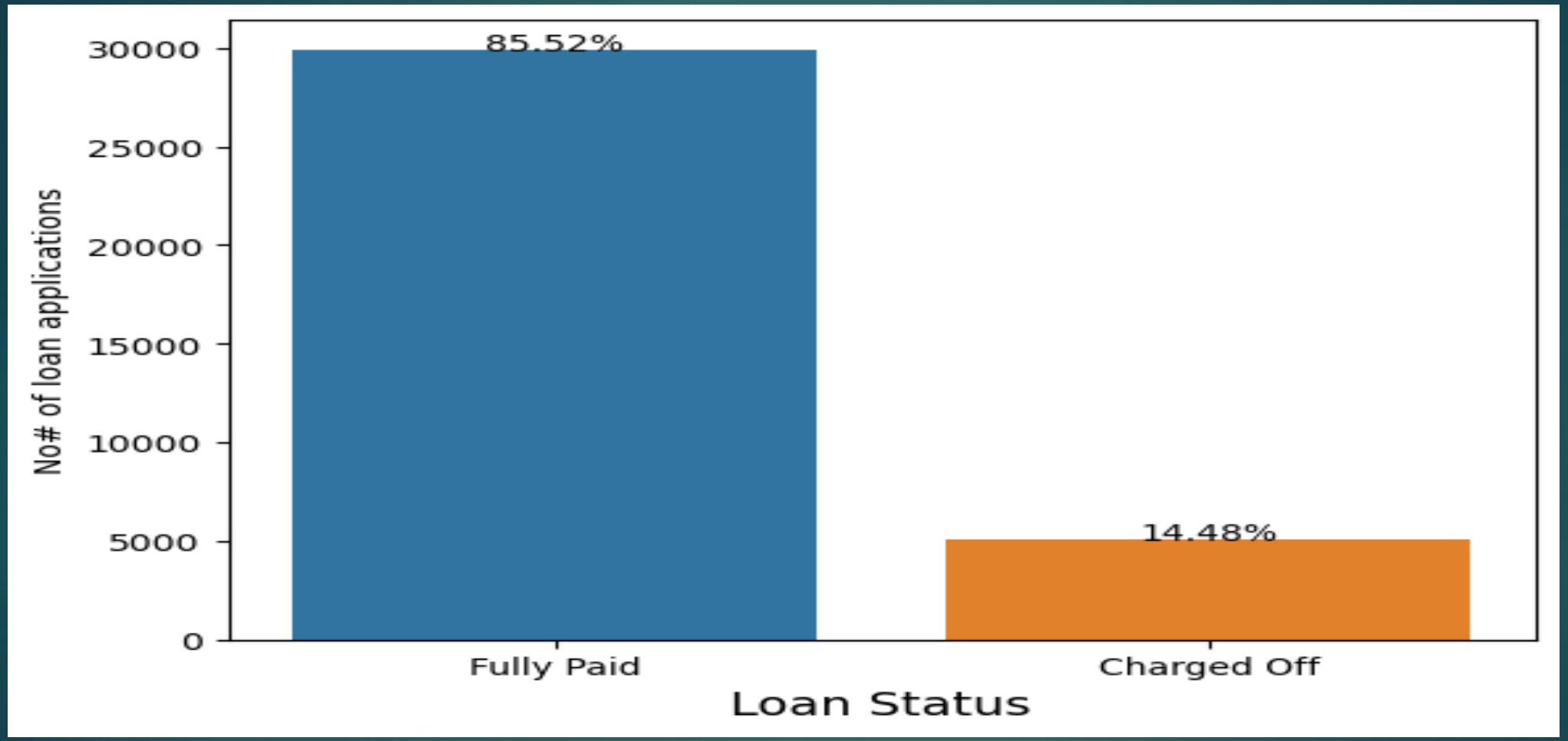
Data cleaning process & Treating Outliers - 1

- Out of 30 Columns, Picked few columns for Numerical type
  - loan_amnt
  - funded_amnt
  - funded_amnt_inv
  - annual_inc
- Out of 30 Columns, Picked few columns for Categorical type
  - Loan_status
  - grade
  - term
  - Verification status
  - emp_length
  - purpose

Data cleaning process & Treating Outliers - 2

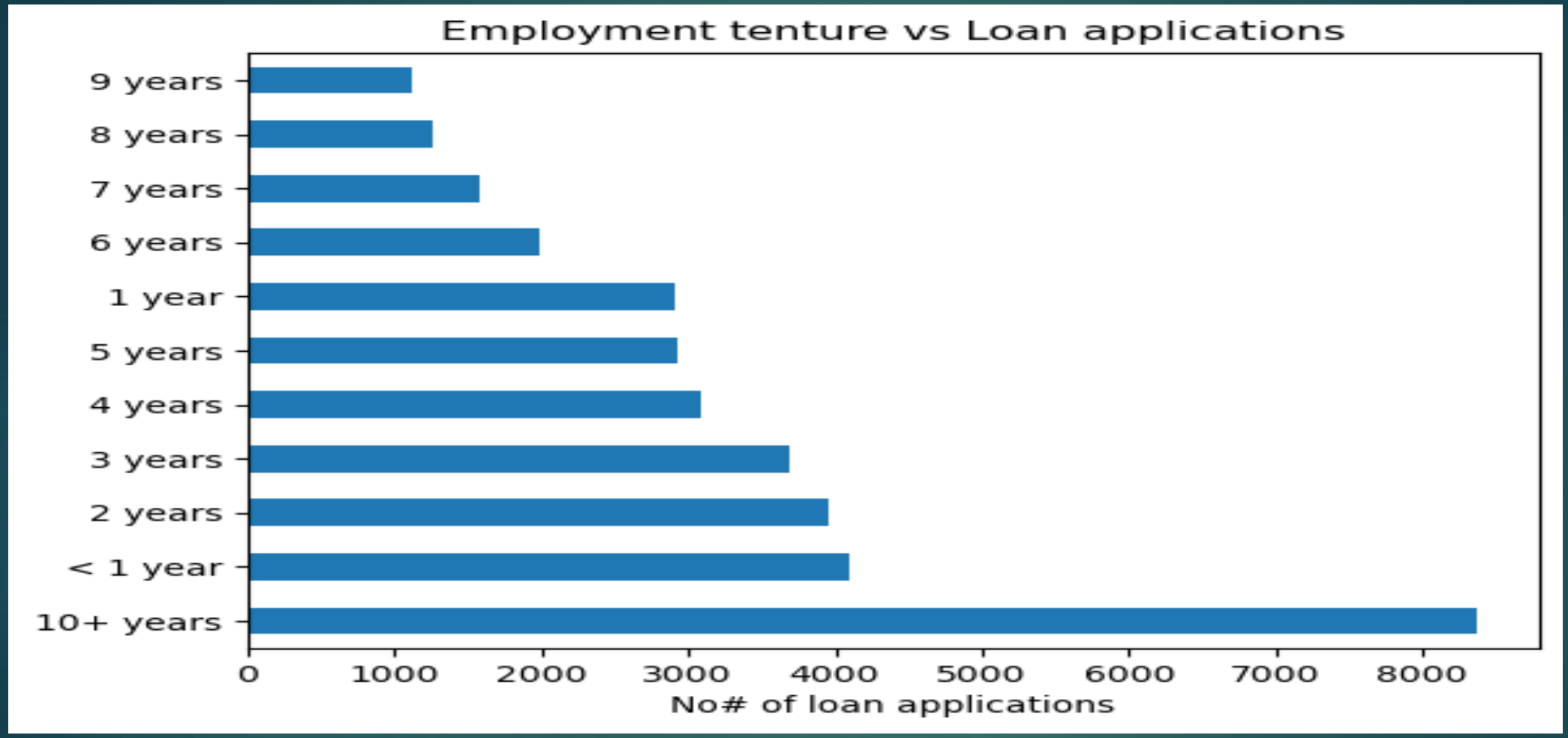
- Used Boxplot to find the outliers and solved this problem using IQR method.
- Created a derived columns for capturing issue year and month
- At the time of final clean up
  - No of columns => 41
  - No of rows => 37880

Univariate Analysis - 1



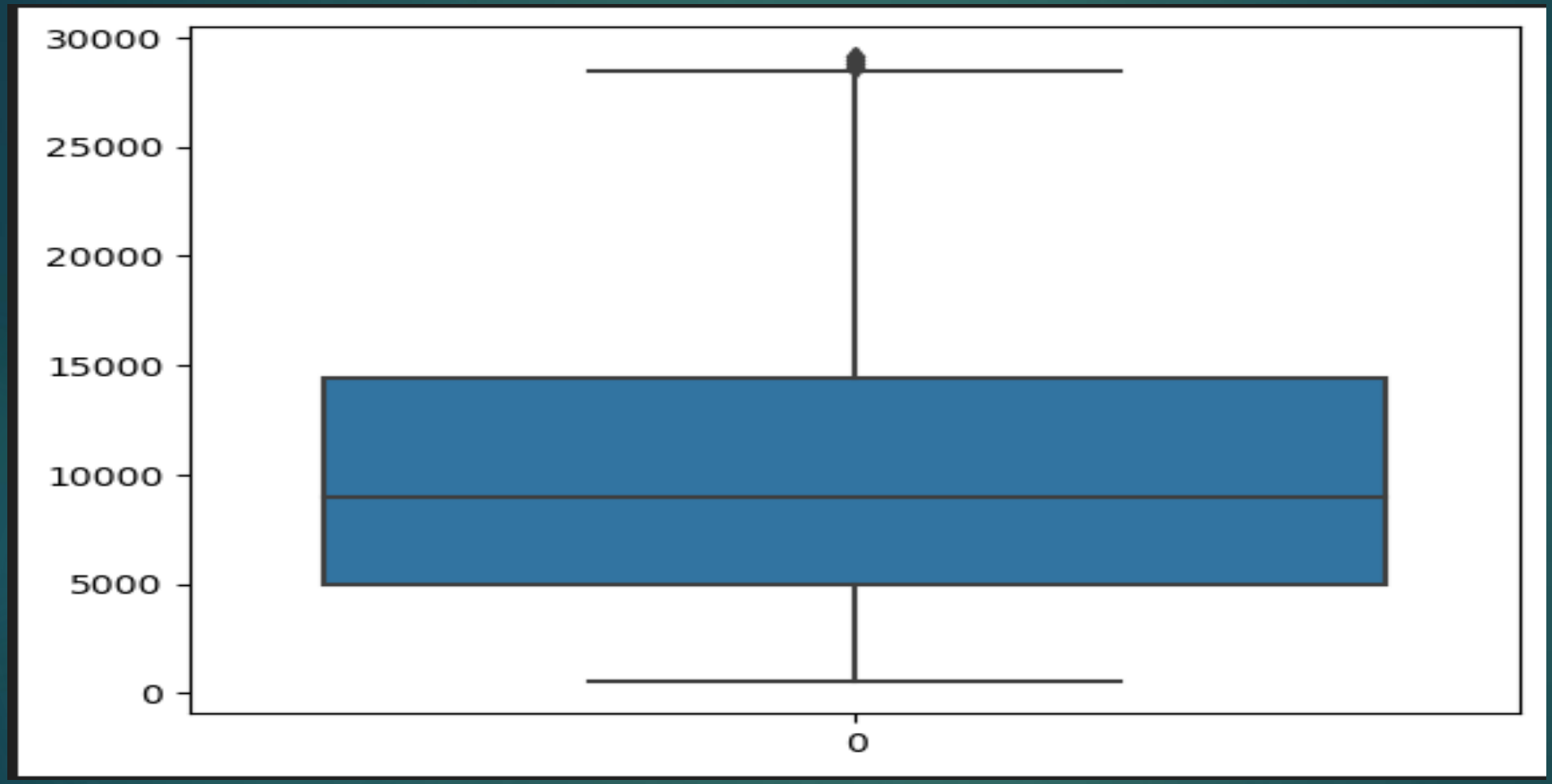
From the current dataset we have 14% of people have defaulted or in chargedoff.

Univariate Analysis - 2



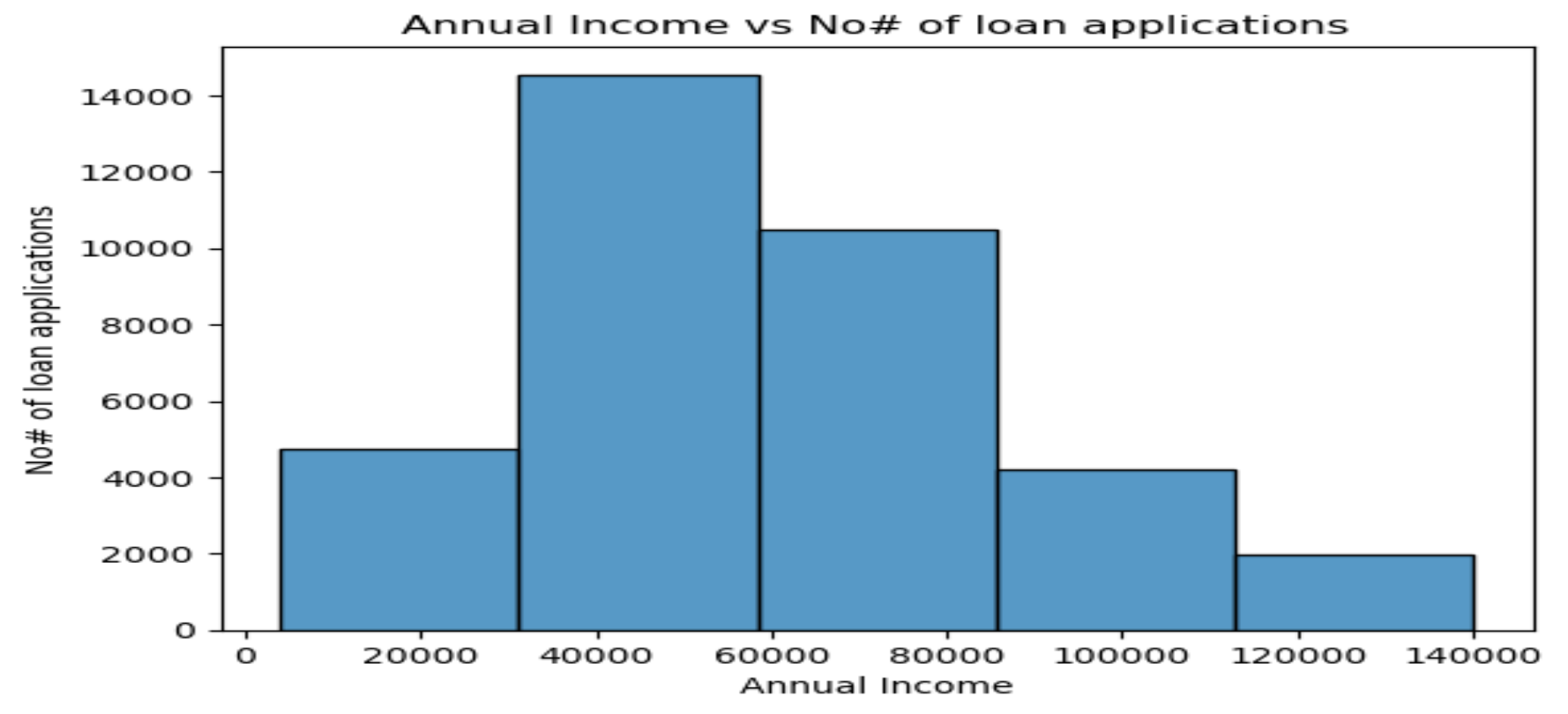
Individuals having experience more than 10+years are taking more loans when compared with other experience levels

Univariate Analysis - 3



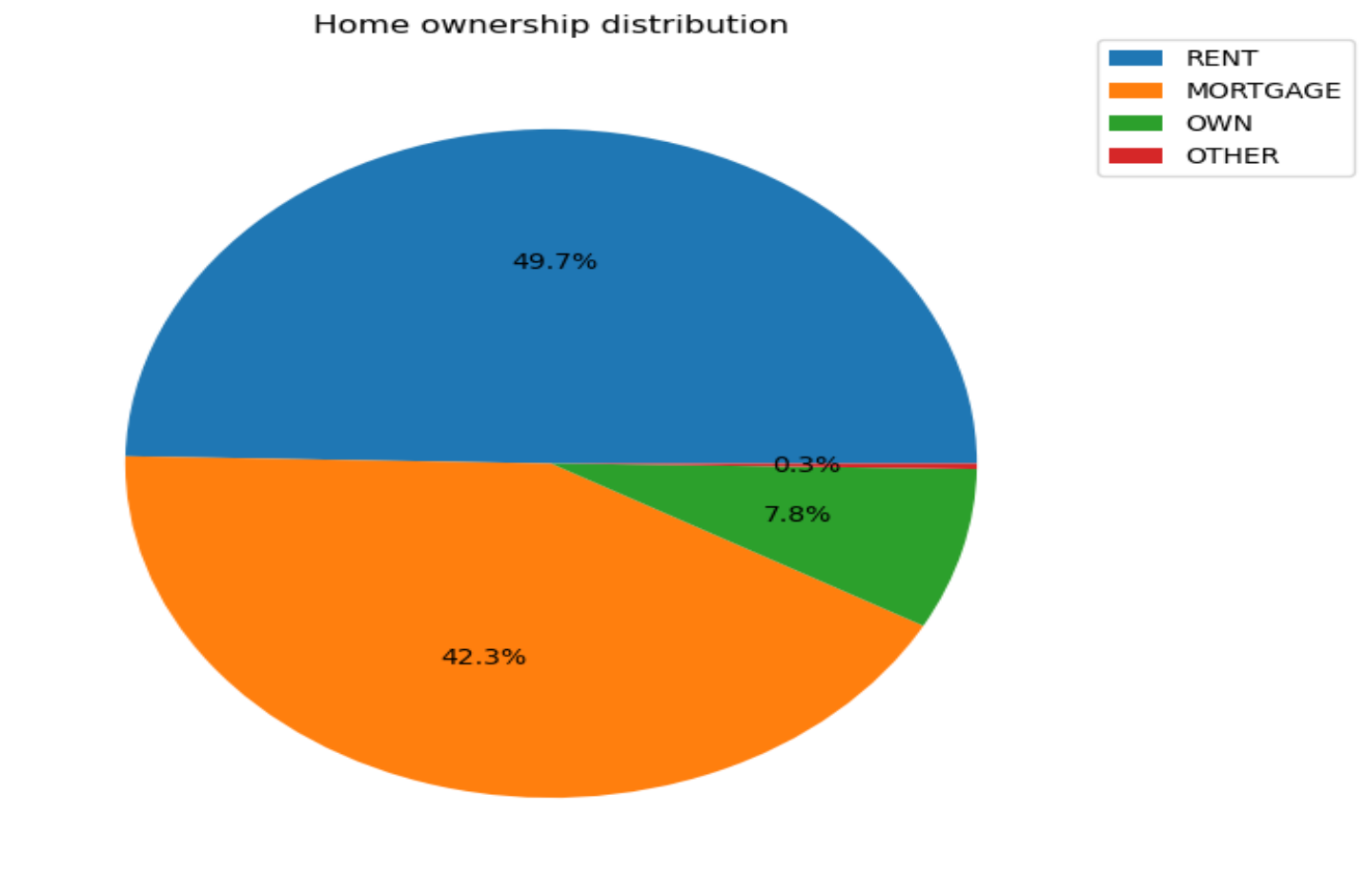
Loan amount is between 500 to ~29,000 (Currency seems to be in USD as the states name and other details in the data set point to US regions)

Univariate Analysis - 4



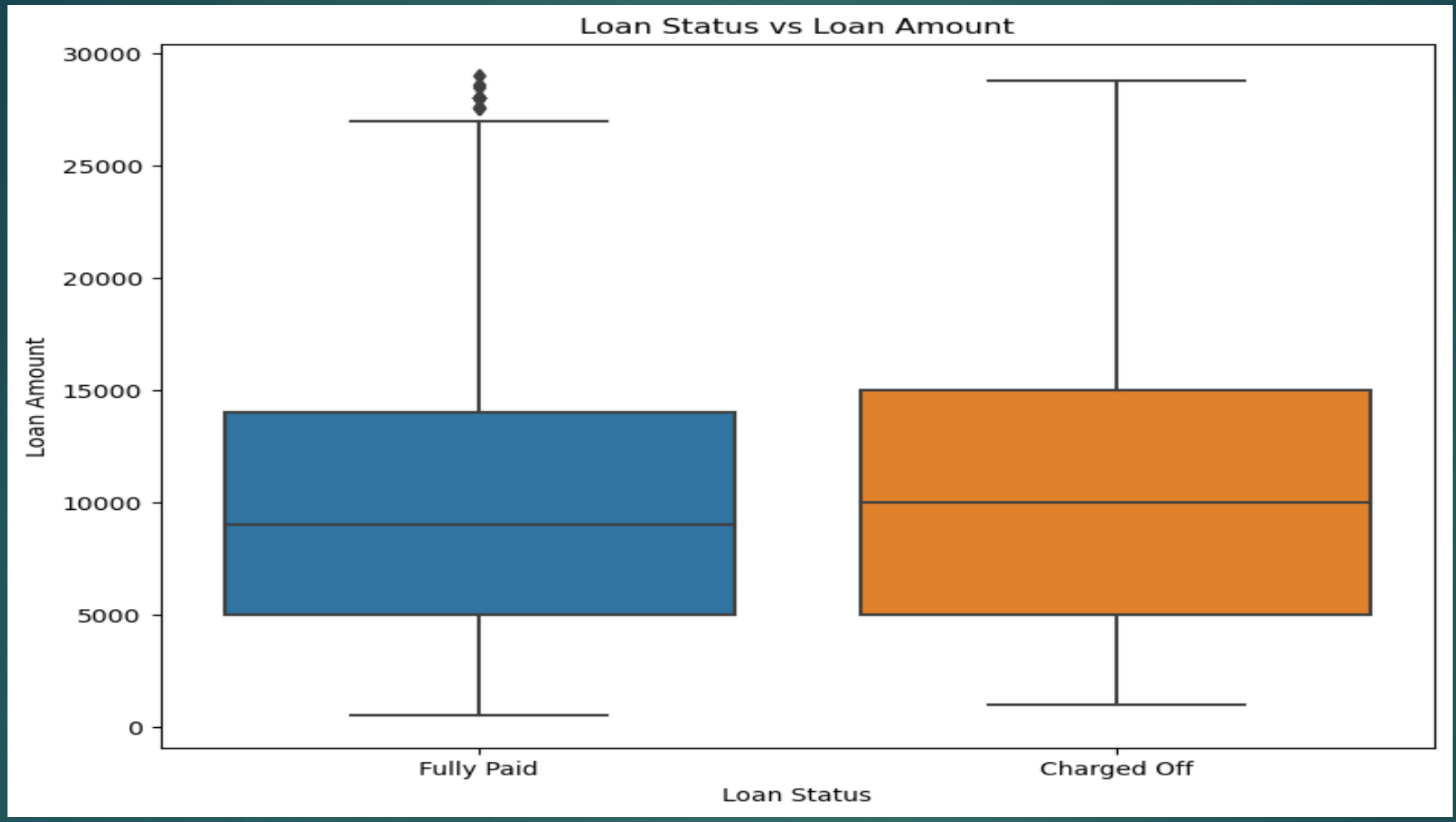
People with Annual Income in range of 30K to 90K are having high chances of getting loan

Univariate Analysis - 5



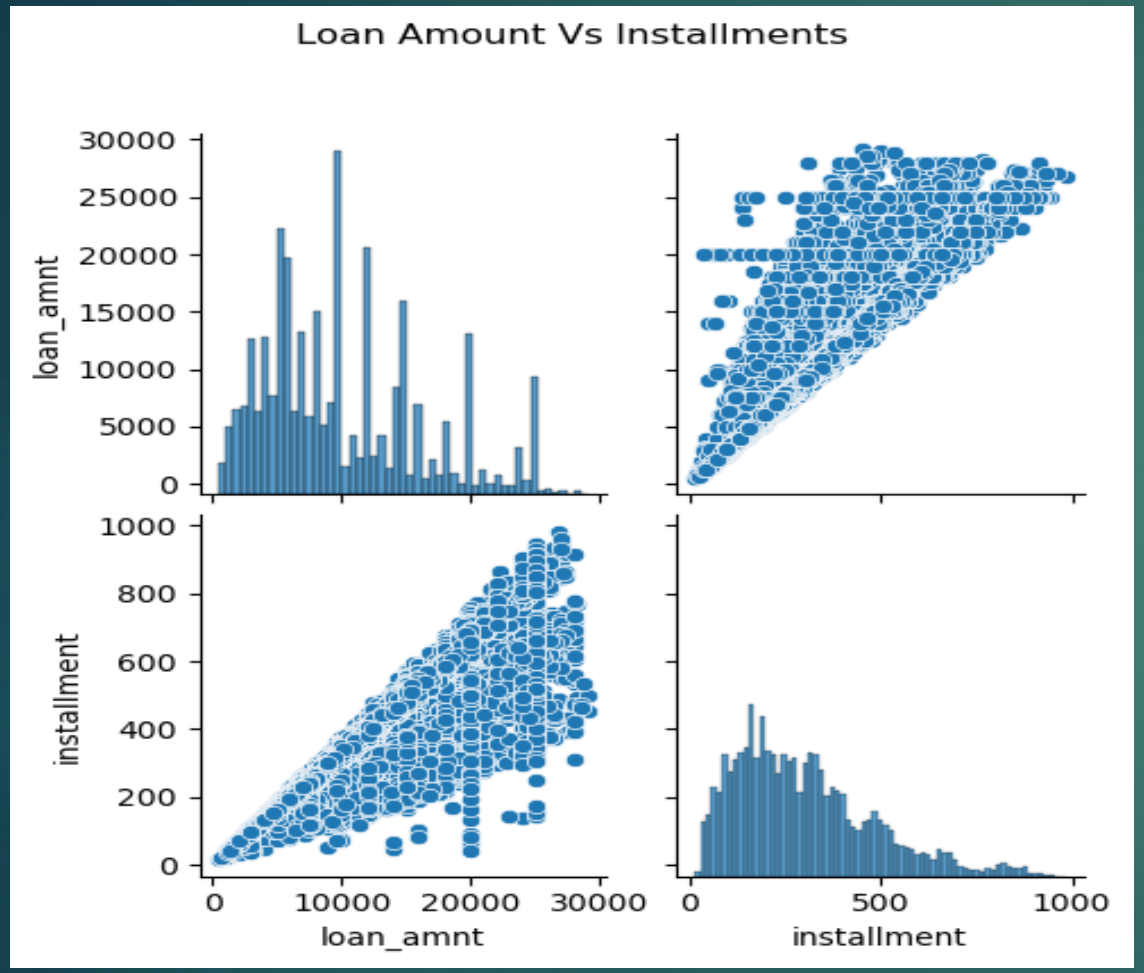
Majority of loan applications are staying in rented house

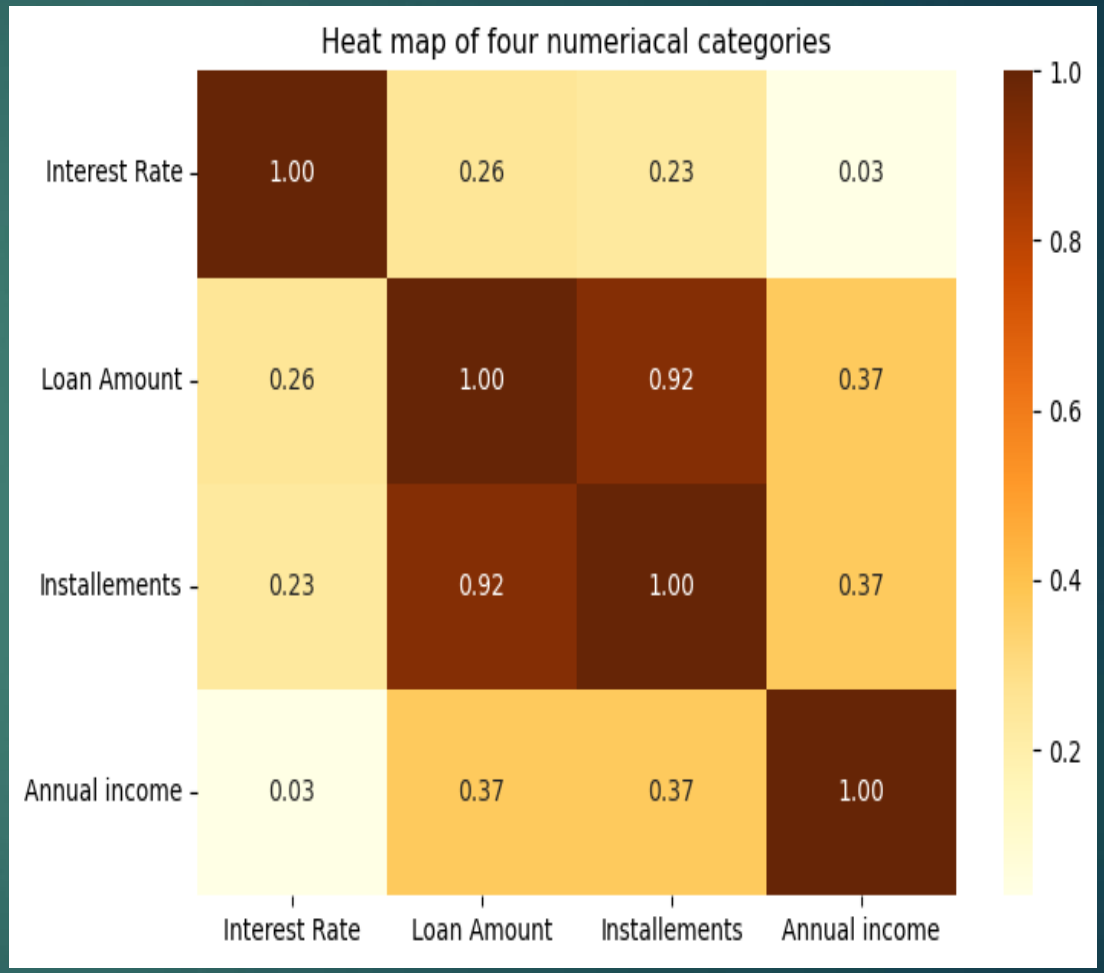
Bivariate Analysis - 1



"Charged_Off" status on Loan seems to high where there is a bigger loan amount

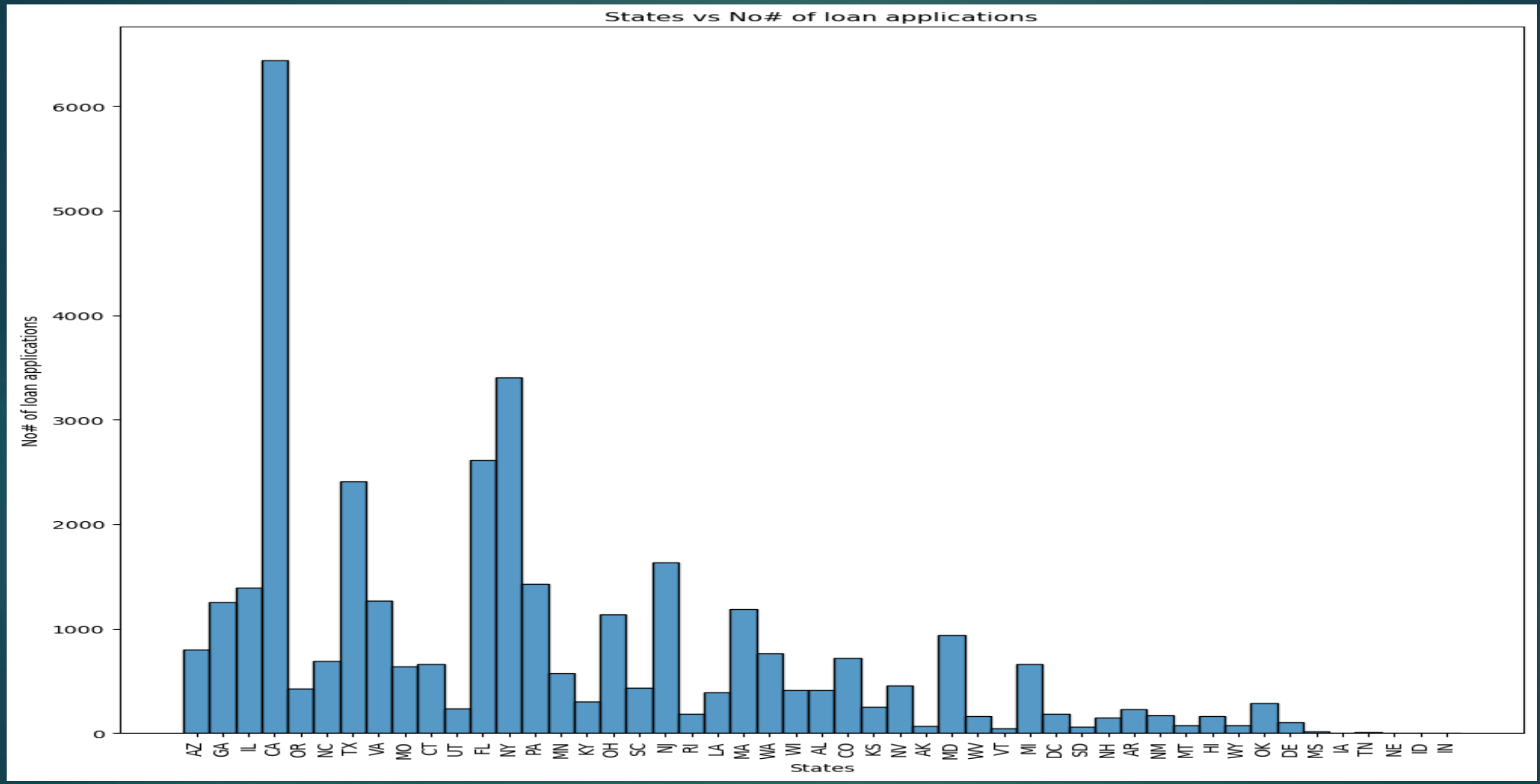
Bivariate Analysis - 2





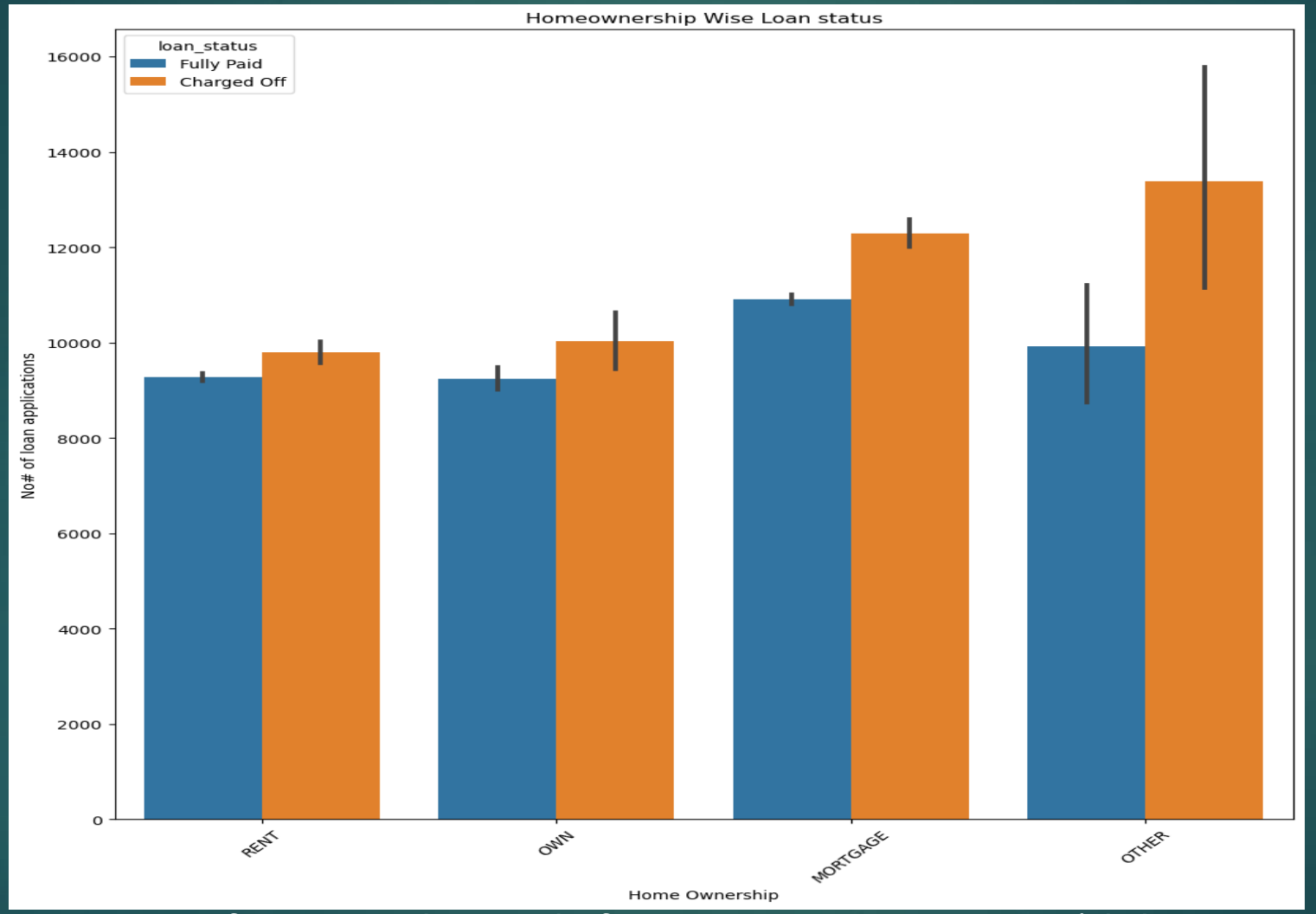
As the Loan amount is increasing installment amount too increasing in linear way

Bivariate Analysis - 3



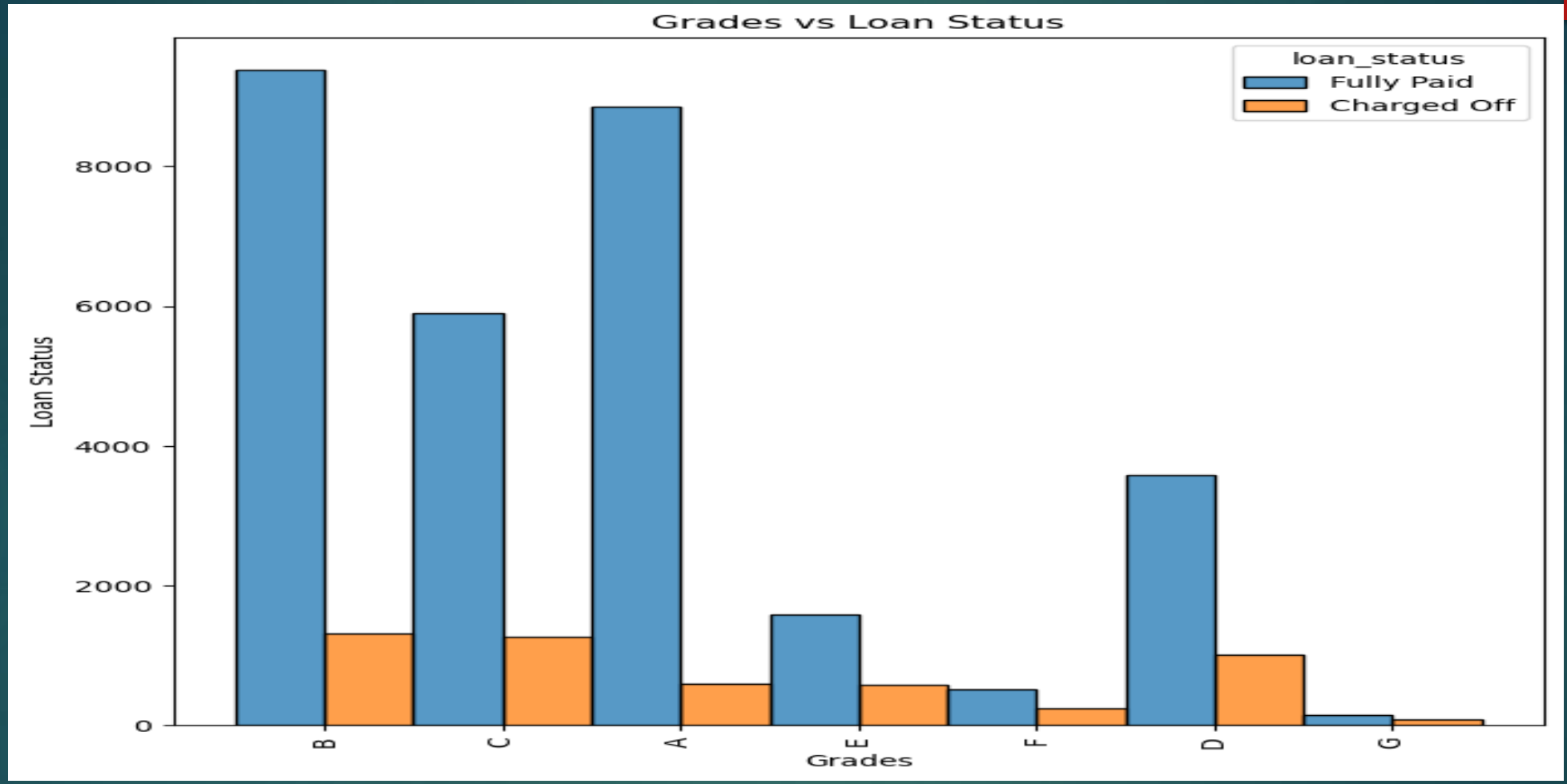
Larger number of loan applications are from 'CA'

Bivariate Analysis - 4



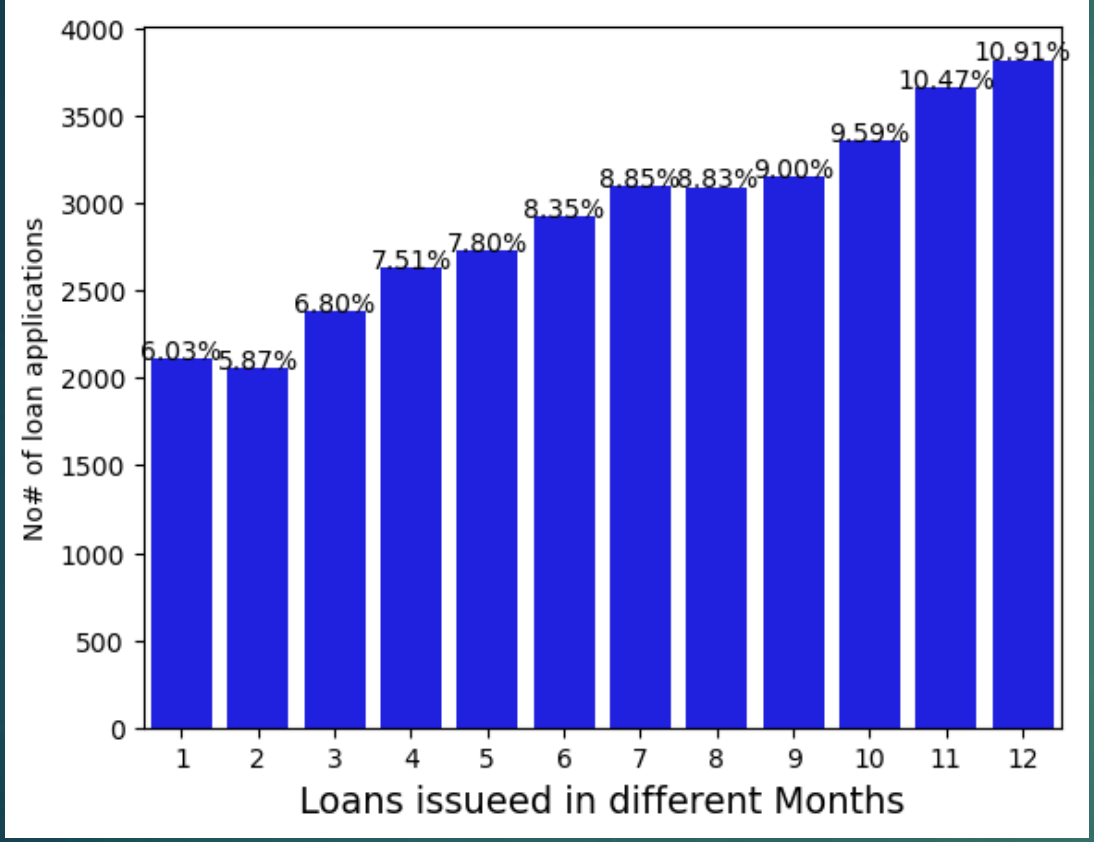
Maximum amount of loans where defaulters are present with home ownership in category mortgage or others.

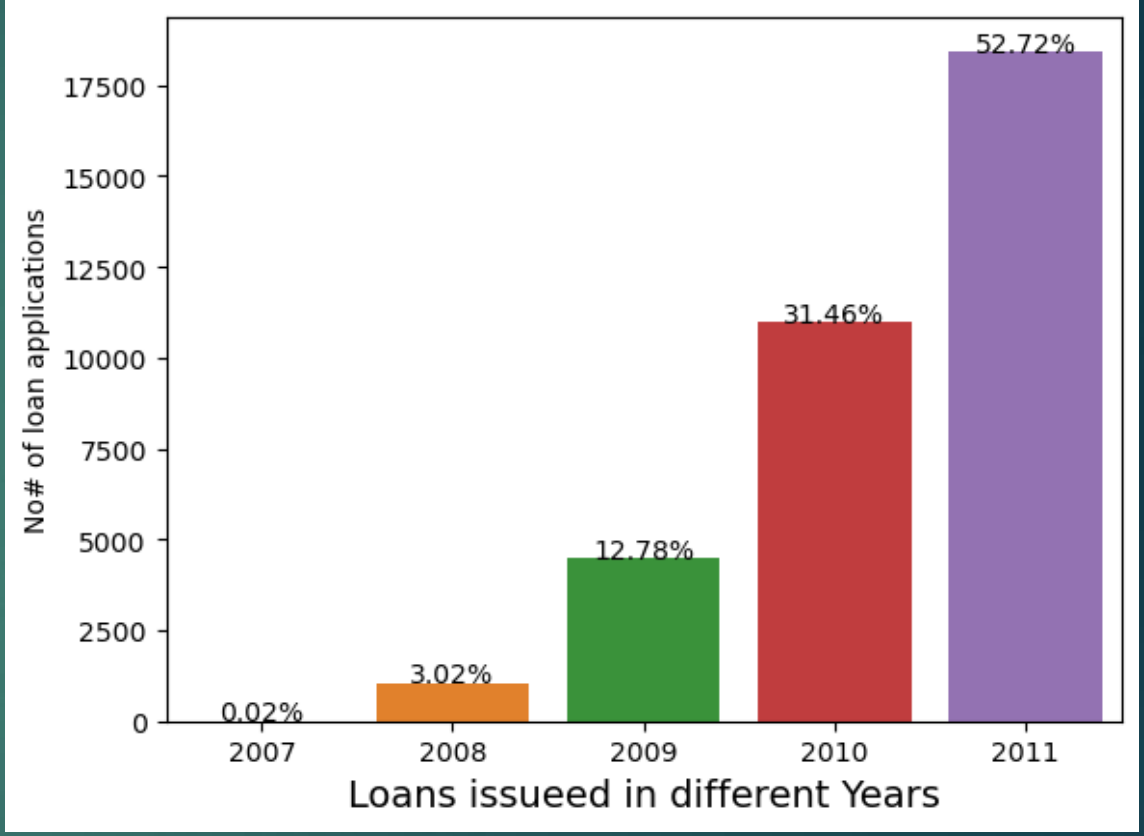
Bivariate Analysis - 5



In Loans of Grade A,B,C there are maximum amount of applications.

Segmented Analysis - 1





In Dec month there are more number of loans got issued.

In Year 2011 % of loans given are more.

Closing Comments

- Columns like Employment tenure, Loan amount, Annual income, Home ownership were taken into consideration for identification of patterns.
- Not all combinations of Numerical and Categorical columns considering for EDA
- No many derived columns where looked out for creation
- Few columns (might be important) got dropped due to lack of domain knowledge