EDA
Case Study
on
Bank Loans

By:

Shiva Chandra Kante

Karnam Abinay Goud

Loan

A loan is the lending of money by one or more individuals, organizations, or other entities to other individuals, organizations. The recipient that is the borrowers incurs a debt and is usually liable to pay interest on that debt until it is repaid as well as to repay the principal amount borrowed.

About the Dataset

- The data set has 39717 Rows and 111 Columns.

Data Cleaning

```
list_of_columns_to_drop=['mths_since_last_major_derog','annual_inc_joint','dti_joint','verification_status_joint',
                         'tot_coll_amt','tot_cur_bal','open_acc_6m','open_il_6m','open_il_12m','open_il_24m',
                         'mths_since_rcnt_il','total_bal_il','il_util','open_rv_12m','open_rv_24m','max_bal_bc','all_util',
                         'total_rev_hi_lim','inq_fi','total_cu_tl','inq_last_12m','acc_open_past_24mths','avg_cur_bal',
                         'bc_open_to_buy','bc_util','mo_sin_old_il_acct','mo_sin_old_rev_tl_op','mo_sin_rcnt_rev_tl_op',
                         'mo_sin_rcnt_tl','mort_acc','mths_since_recent_bc','mths_since_recent_bc_dlq','mths_since_recent_inq',
                         'mths_since_recent_revol_delinq','num_accts_ever_120_pd','num_actv_bc_tl','num_actv_rev_tl',
                         'num_bc_sats','num_bc_tl','num_il_tl','num_op_rev_tl','num_rev_accts','num_rev_tl_bal_gt_0','num_sats',
                         'num_tl_120dpd_2m','num_tl_30dpd','num_tl_90g_dpd_24m','num_tl_op_past_12m','pct_tl_nvr_dlq',
                         'percent_bc_gt_75','tot_hi_cred_lim','total_bal_ex_mort','total_bc_limit','total_il_high_credit_limit']
```

- There are 54 Columns in which all the rows are Null
- Those 54 Columns to be dropped are in the above Figure.
- In the Dataset 10 columns have Unique Values. So as we don’t need unique Values we are dropping them.

```
#Dropping unique values
df.drop(['id','member_id','url','pymnt_plan','initial_list_status','acc_now_delinq','application_type',
         'chargeoff_within_12_mths','delinq_amnt','tax_liens','collections_12_mths_ex_med'],inplace=True,axis=1)
```

- Two Columns have more than 90% Null Values i.e (mths_since_last_record, next_pymnt_d). So we are dropping those Columns.
- One column named “mths_since_last_delinq” has 64% Null values and we don’t have any use with that column so we can drop that column as well.
- There are no Duplicated rows in the dataset.

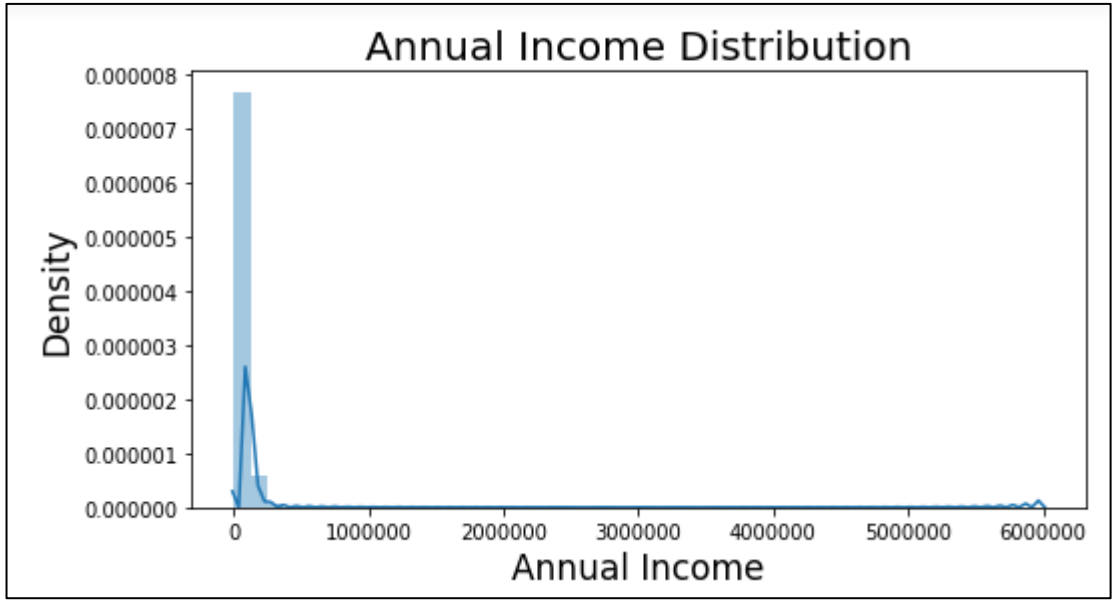
```
df.duplicated().any()
```

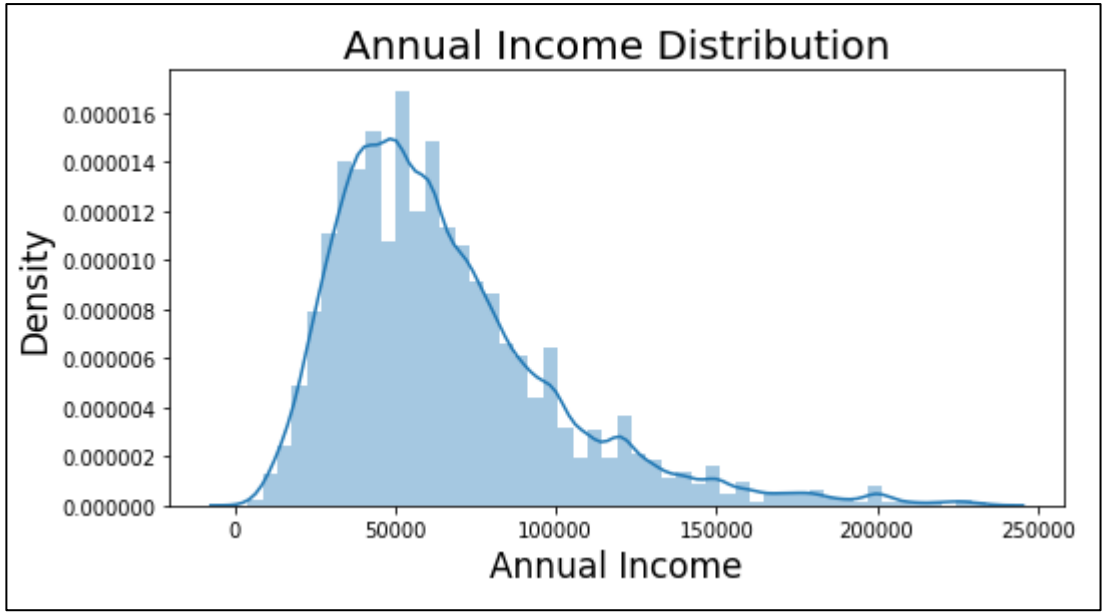
```
False
```

- The column Loan Status has three types of Values Fully Paid, Charged Off and Current. We don’t have use with the Current, so we are dropping the rows which has loan status as current.
- The Column Employee Title has 2386 Null Values. Instead of filling it with mode I filled it as ‘UnKnown’.
- The same above process has been done with the Employee length column as well.
- The Column ‘'pub_rec_bankruptcies‘’ has 687 Null Value and we have no use with that column, so we are dropping that column.
- The Column 'last_pymnt_d’, 'revol_util’ and 'last_credit_pull_d’ have less Number of null values so I am dropping the rows which have null values.

```
df.dropna(subset=['last_pymnt_d','revol_util','last_credit_pull_d'],inplace=True)
```

- The Interest rate is in the string format we have to convert it in to Float type.
- The issue date is the format object so we have to covert it in to the datetime object.
- The column annual income has outliers because mean is 68,835 and the maximum value is 6,00,000. so we need to remove that outlier's. We can see in the below figures how the annual income was spread before removing outliers and after removing outliers.





Univariate Analysis

Univariate analysis is perhaps the simplest form of statistical analysis. Like other forms of statistics, it can be inferential or descriptive. The key fact is that only one variable is involved.

- In figure 1 we can see the power log graph for the Annual Income, from that graph we can say that most people have almost same range of income.

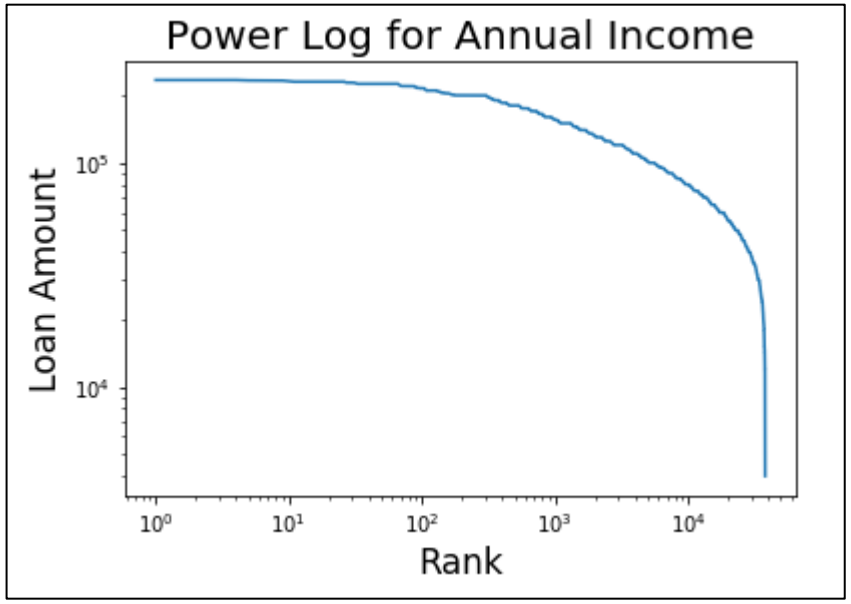


Fig 1

❖ From figure 2 we can infer that most people took loan by interest rate between 10-15.

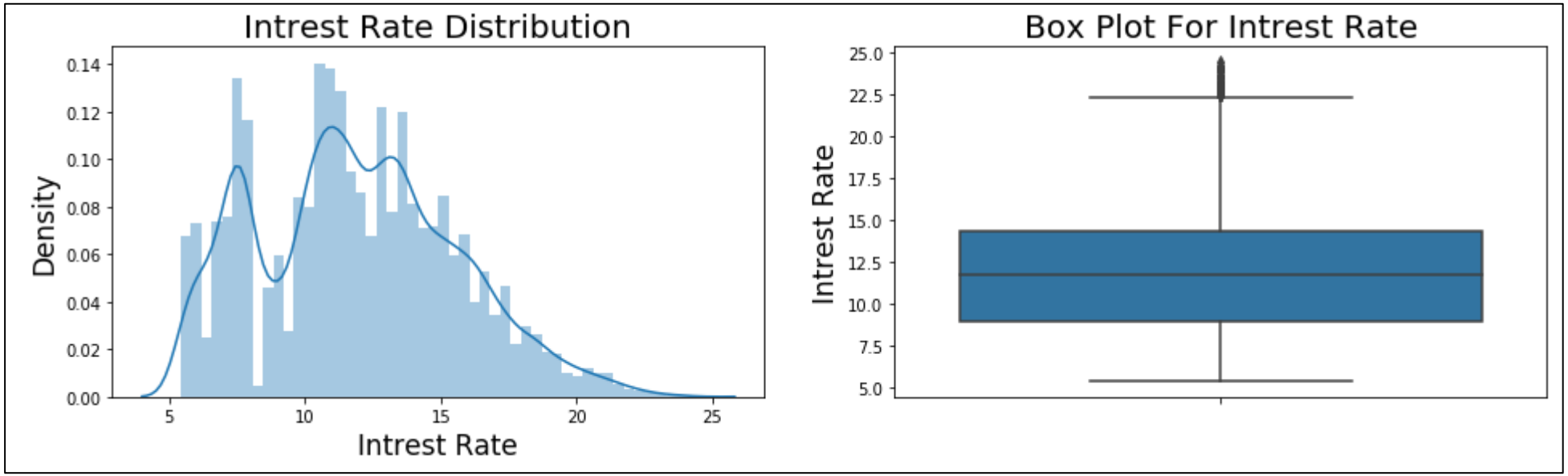


Fig 2

❖ we can infer from figure 3 that the Loan Amount is spread evenly.

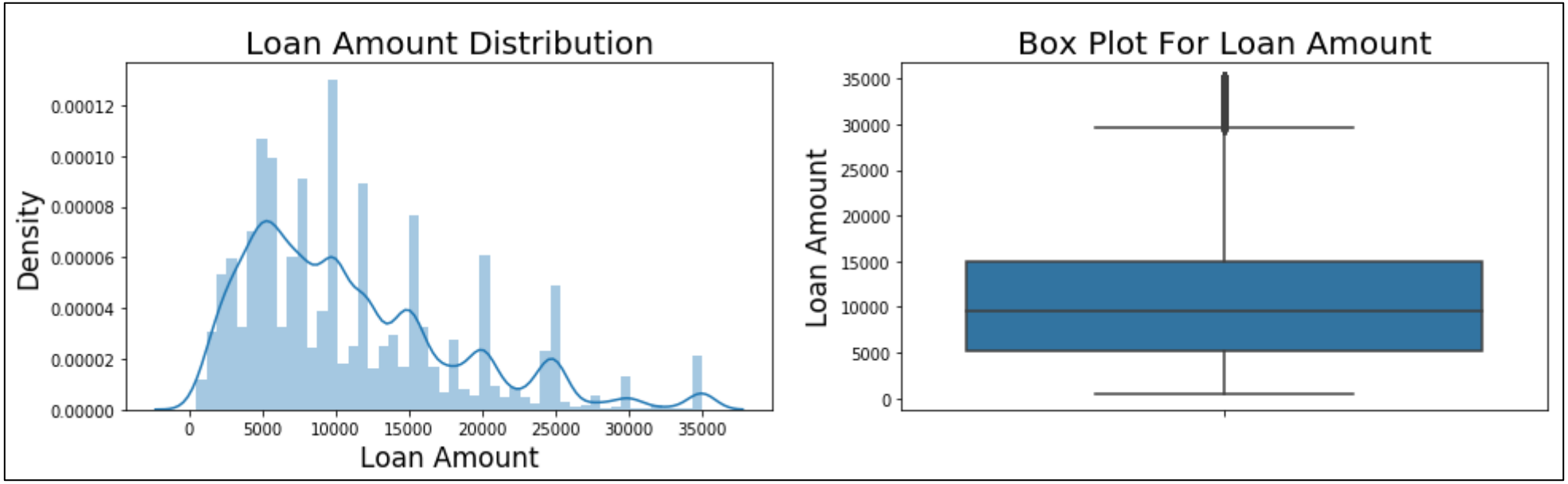


Fig 3

- From figure 4 we can Infer that the Dataset has most Fully Paid People list compared to the charged Off.

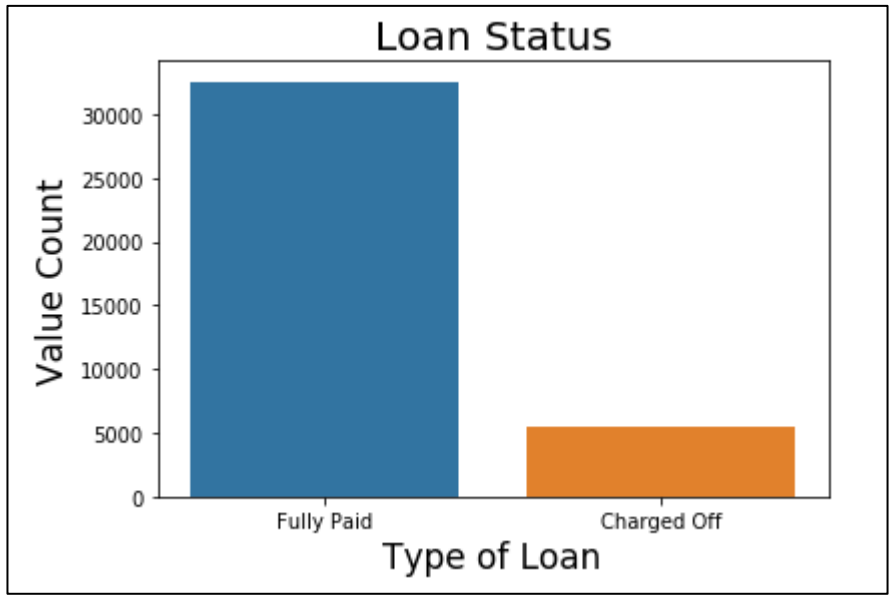


Fig 4

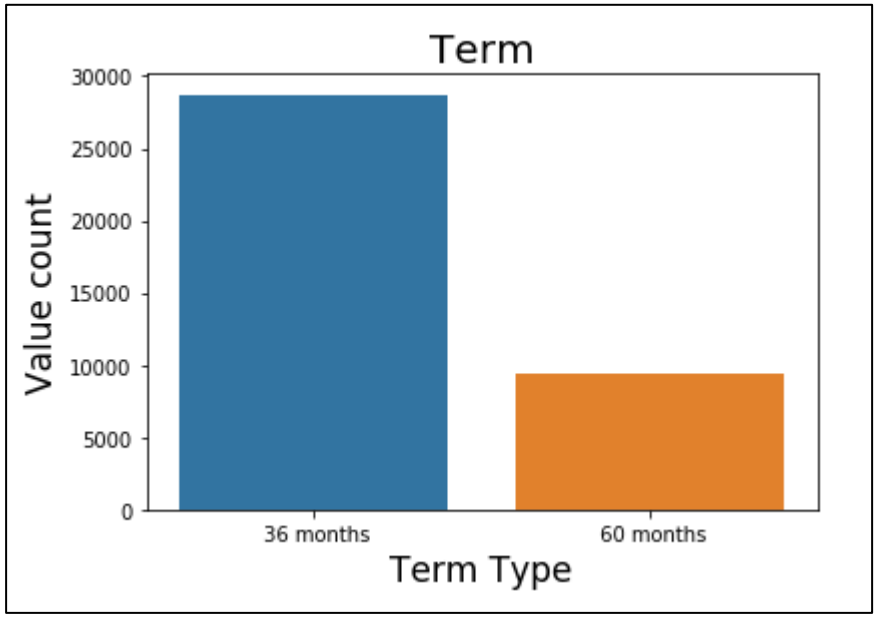


Fig 5

- From figure 5 we can say that most people prefer to take for 3 years loan rather than 5 years.

❖ From figure 6 we can say that most people are paying instalment in range 150-350.

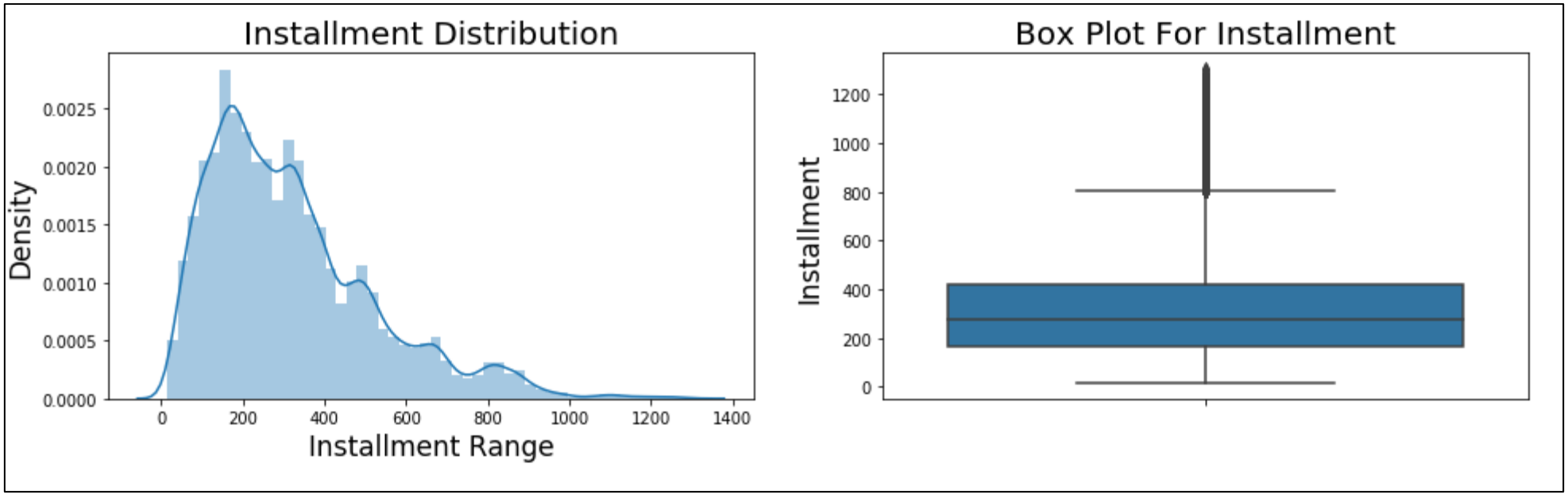


Fig 6

Bivariate And Multivariate Analysis

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them.

- From Figure 7 we can say that the median of Loan Amount of charged-off people is high i.e. the people who are not paying the loan are taking more amount of money.

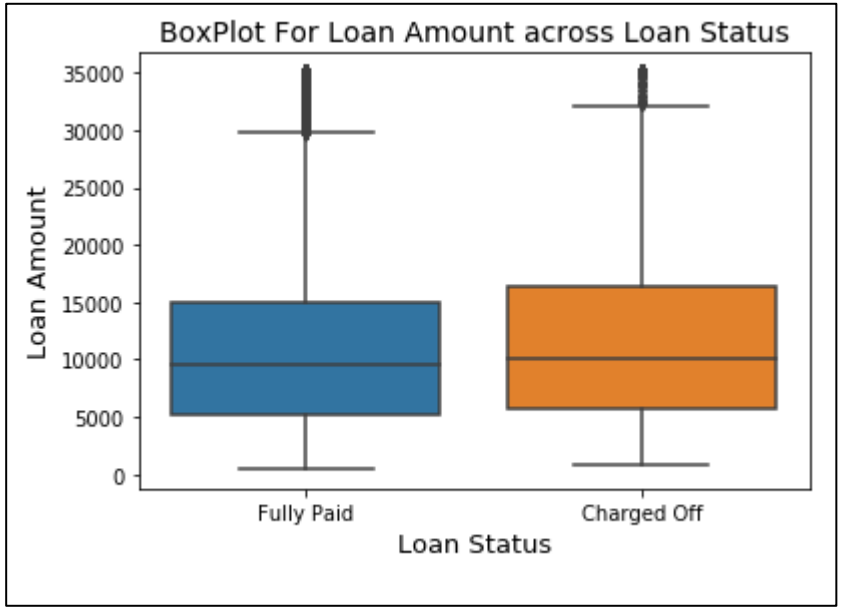


Fig 7

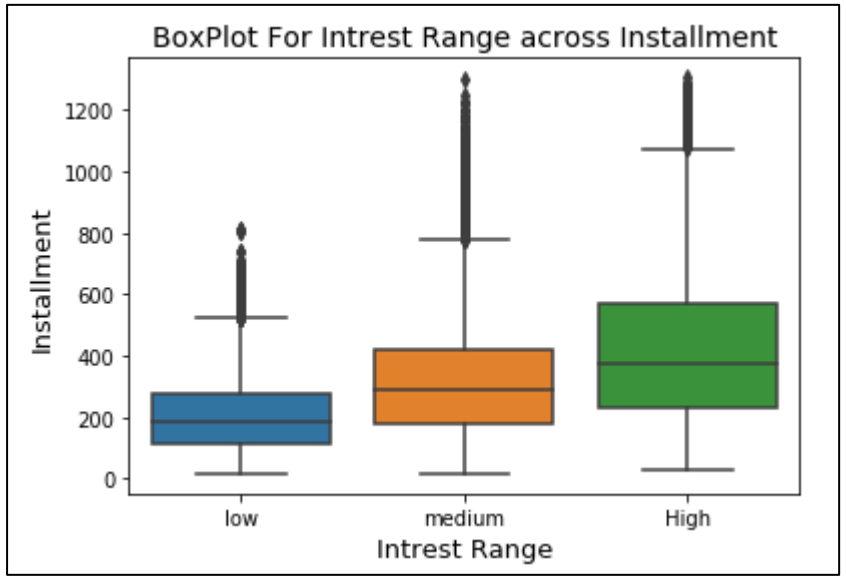


Fig 8

- We can see that the installment is rising as the Income Range is getting high from figure 8.

❖ From the Figure 9 we can say that the median of Installment of charged-off is high.

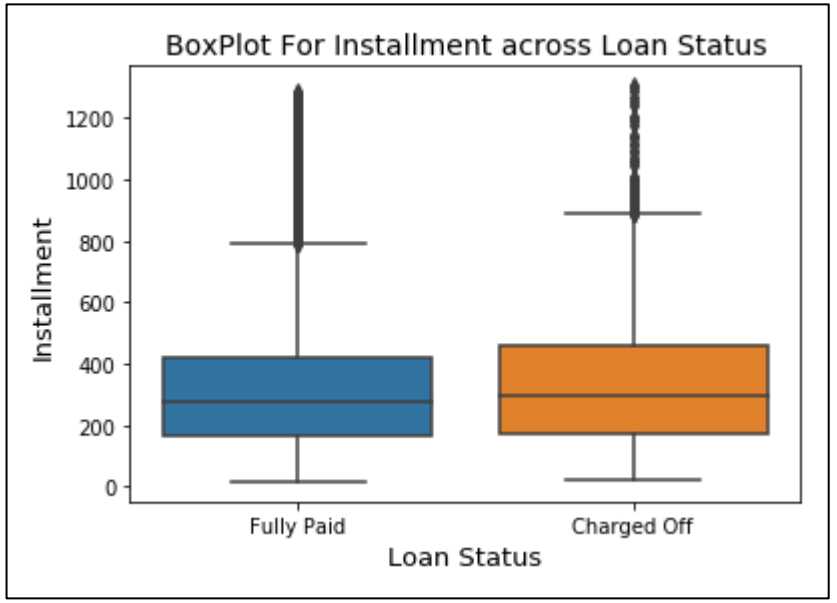


Fig 9

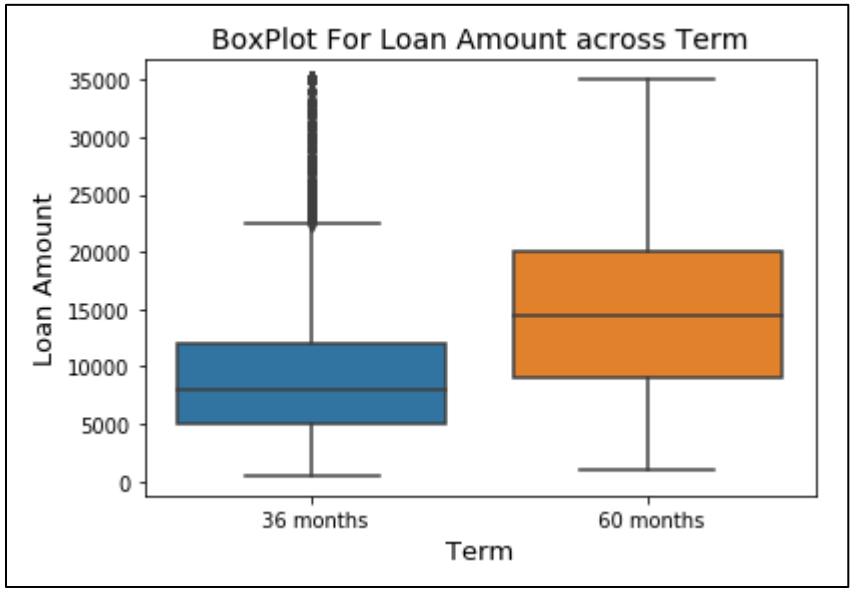


Fig 10

❖ We can see infer that the average Loan Amount of people whose term is 5 years is more from the figure 10.

❖ We can infer from the figure 11 that the median of interest is gradually increasing from Grade A to G continuously.

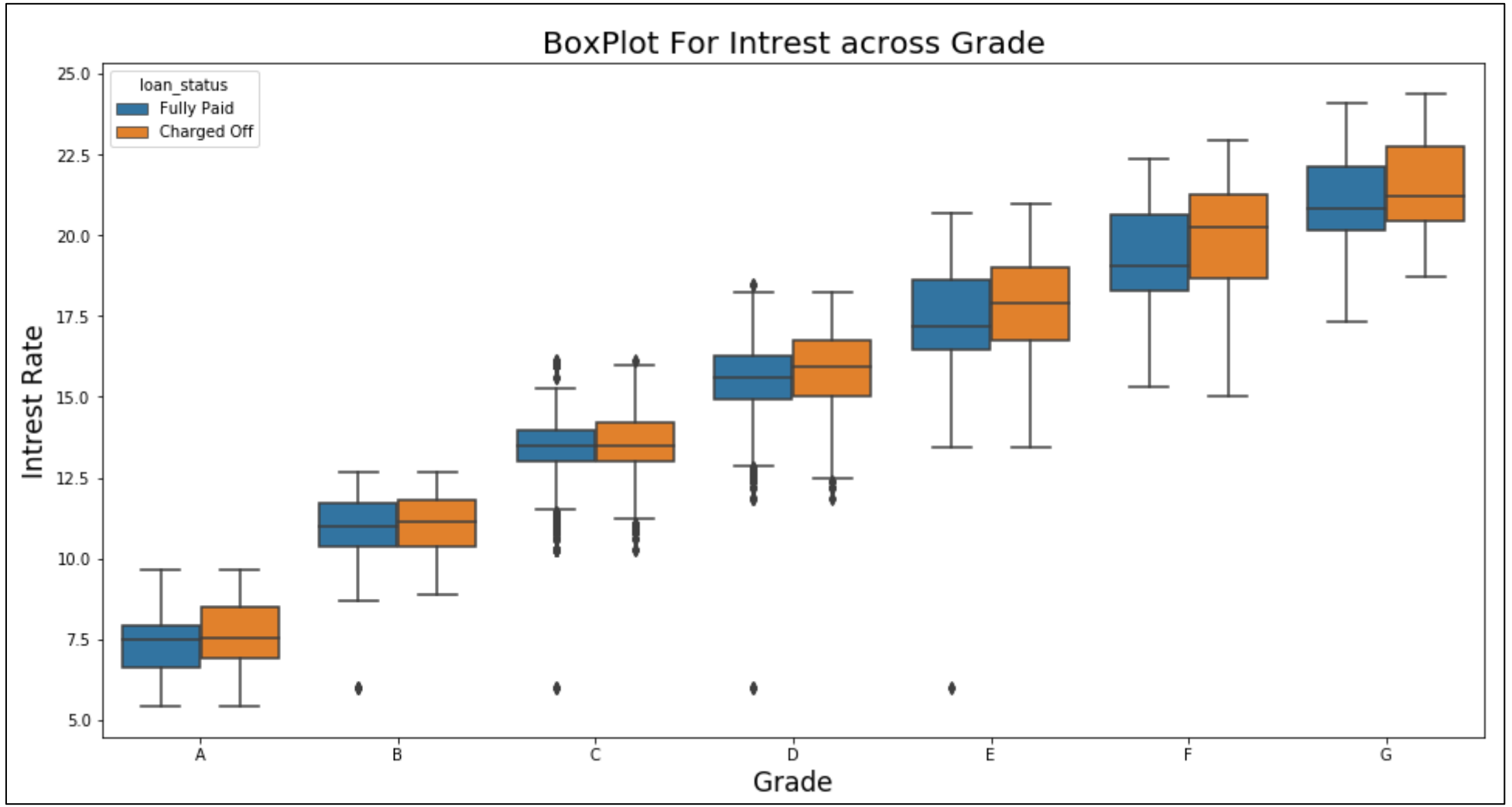


Fig 11

❖ We can infer that the median of interest is gradually increasing from A1 to G5 from the figure 13.

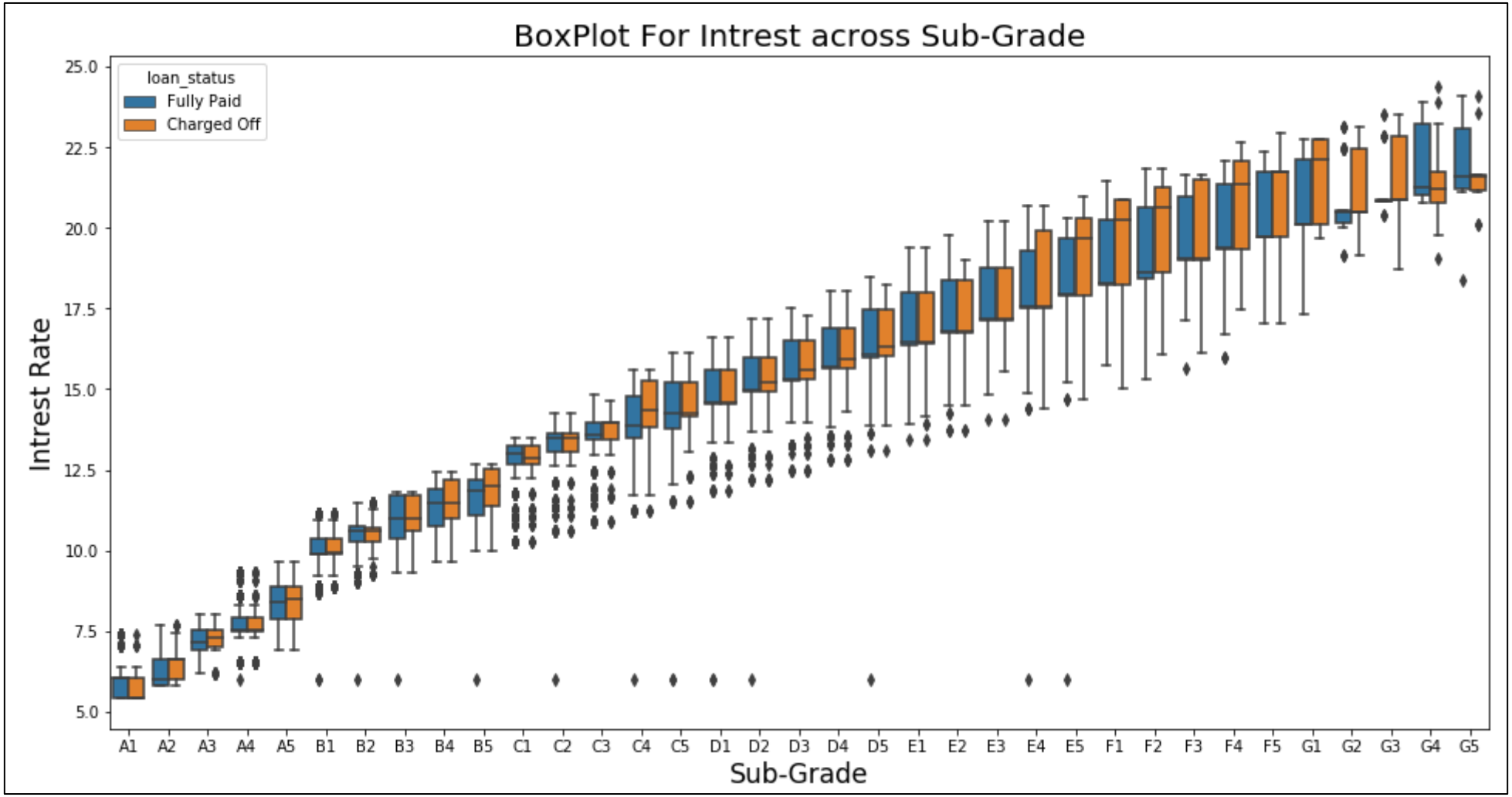


Fig 13

❖ We can see that the people whose home ownership are RENT and MORTAGE are most like to be charged-off from the figure 14.

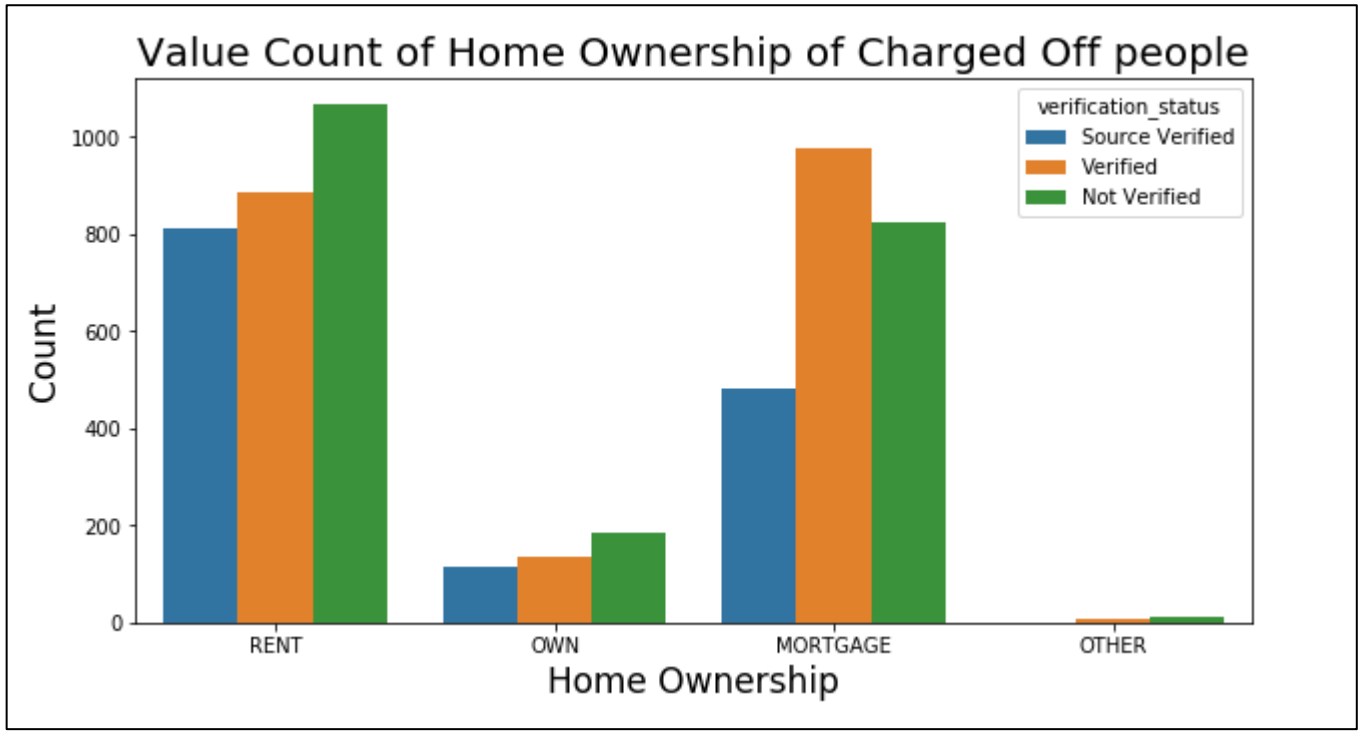


Fig 14

❖ From the figure 15(line plot) we can see that after the steady increase of interest from the year 2007 the interest rate got decreased from 2009 to 2010. The main reason would be Recession.

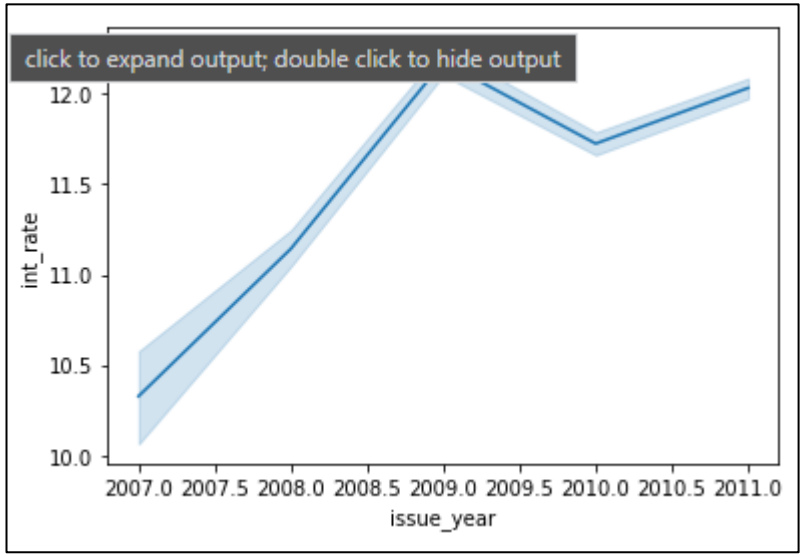


Fig 15

- From the figure 16 we can say that in both the cases i.e. (Fully Paid, Charged Off) the Mortgage people are paying the high Instalments.

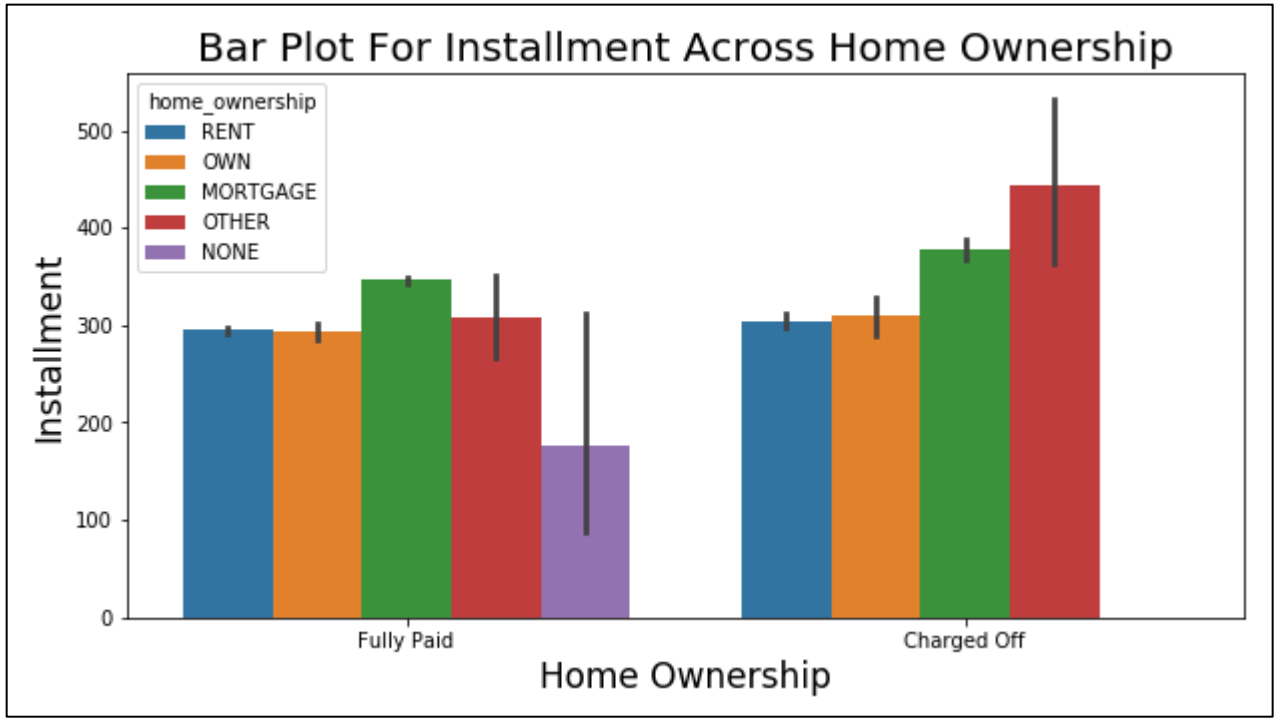


Fig 16

Thank You

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise

- John W.Tukey