upGrad

iit-b
ज्ञानमुत्तमम

Lending Club Case Study

Introduction

One of the leading Consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.

2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Objective

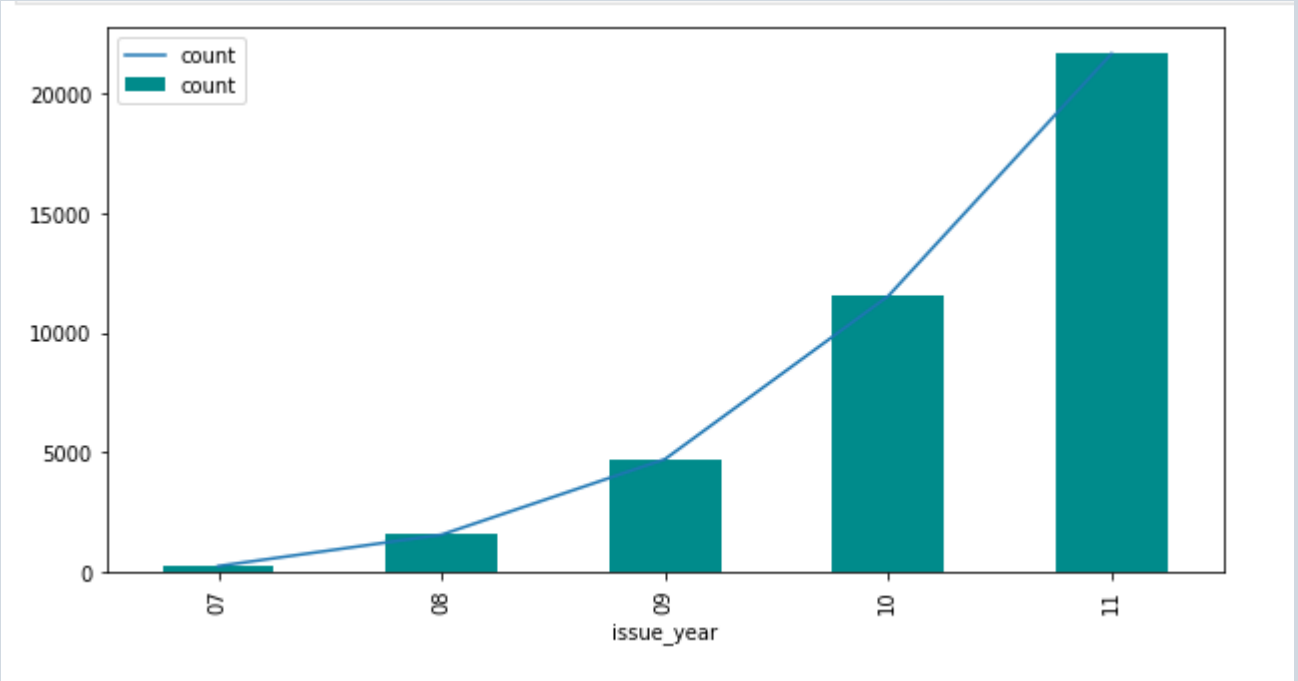
To find the risky applicants by identifying the patterns and corelations in the given data which will be helpful for the organization to minimize the business loss.

EDA

After analyzing the data we performed following

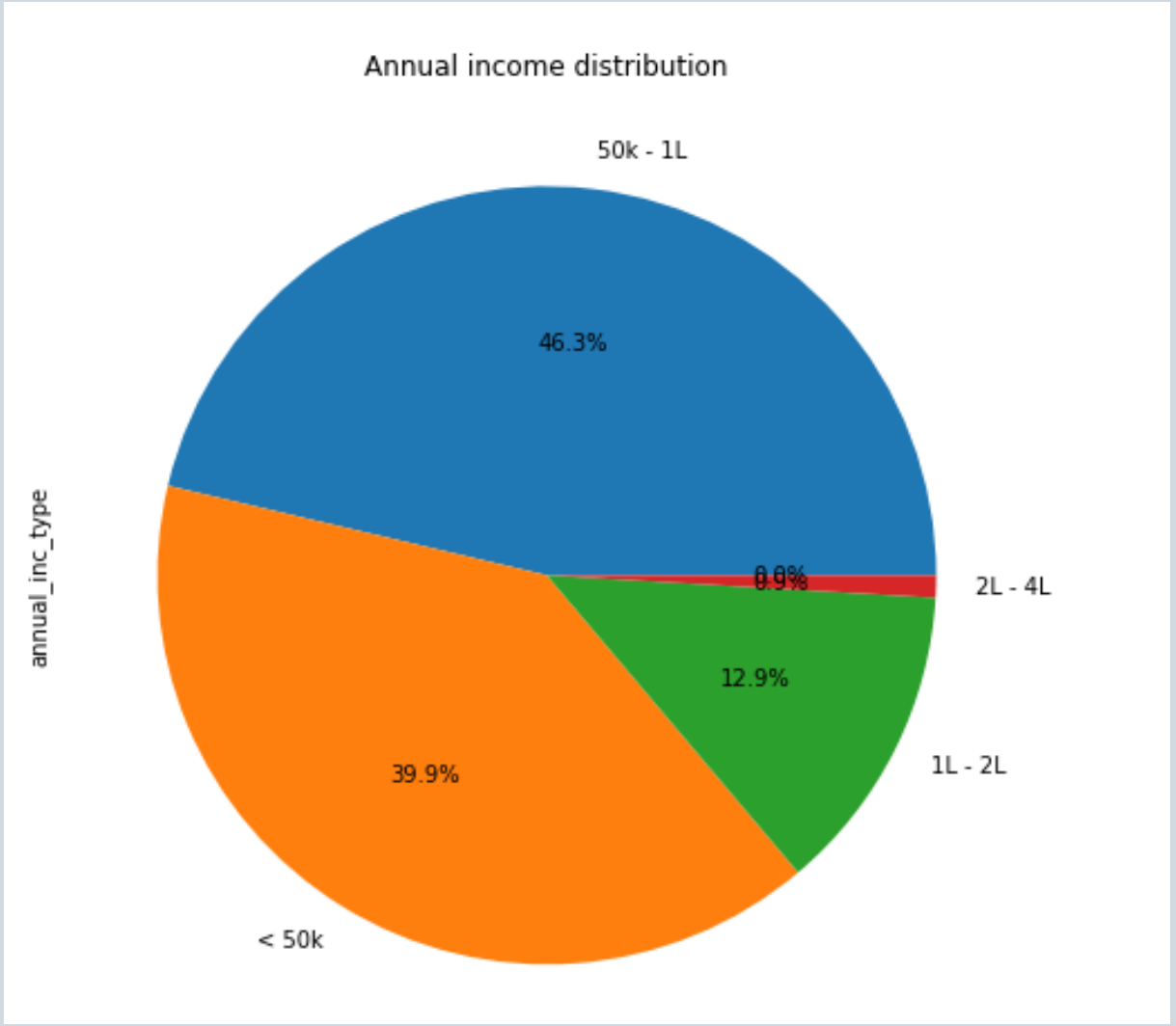
1. Data with 100% of null values are removed
2. Data cleaning is performed
3. Missing value treatment is done
4. Data imputation is performed
5. Outliers detection and treatment is performed
6. Various columns derived from existing data
7. Correlation Metrix is created to find the correlation between the different attributes

After Plotting a graph of loan_status vs issue_year



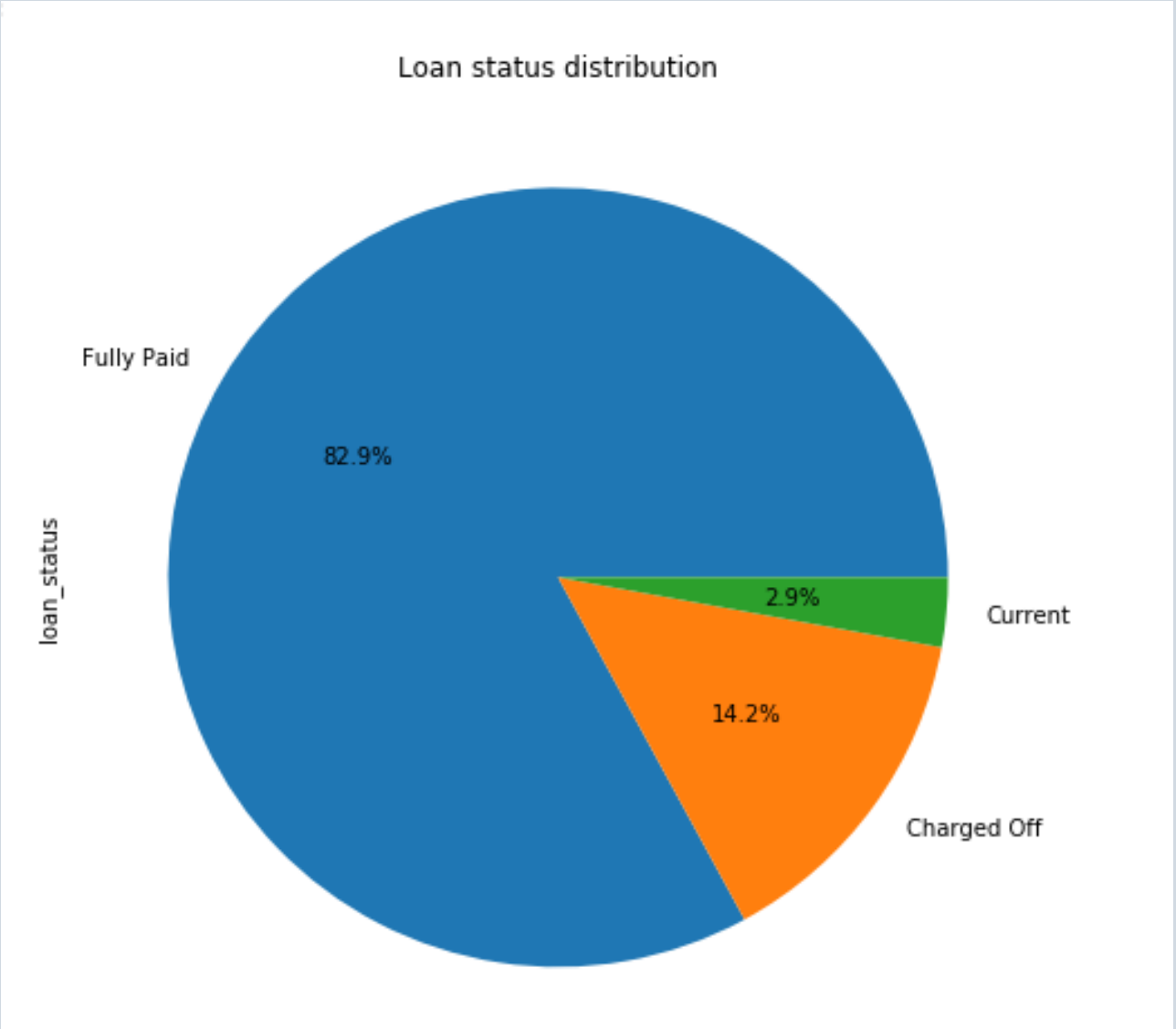
Inference:

There is a steady increase in the issuance of loan



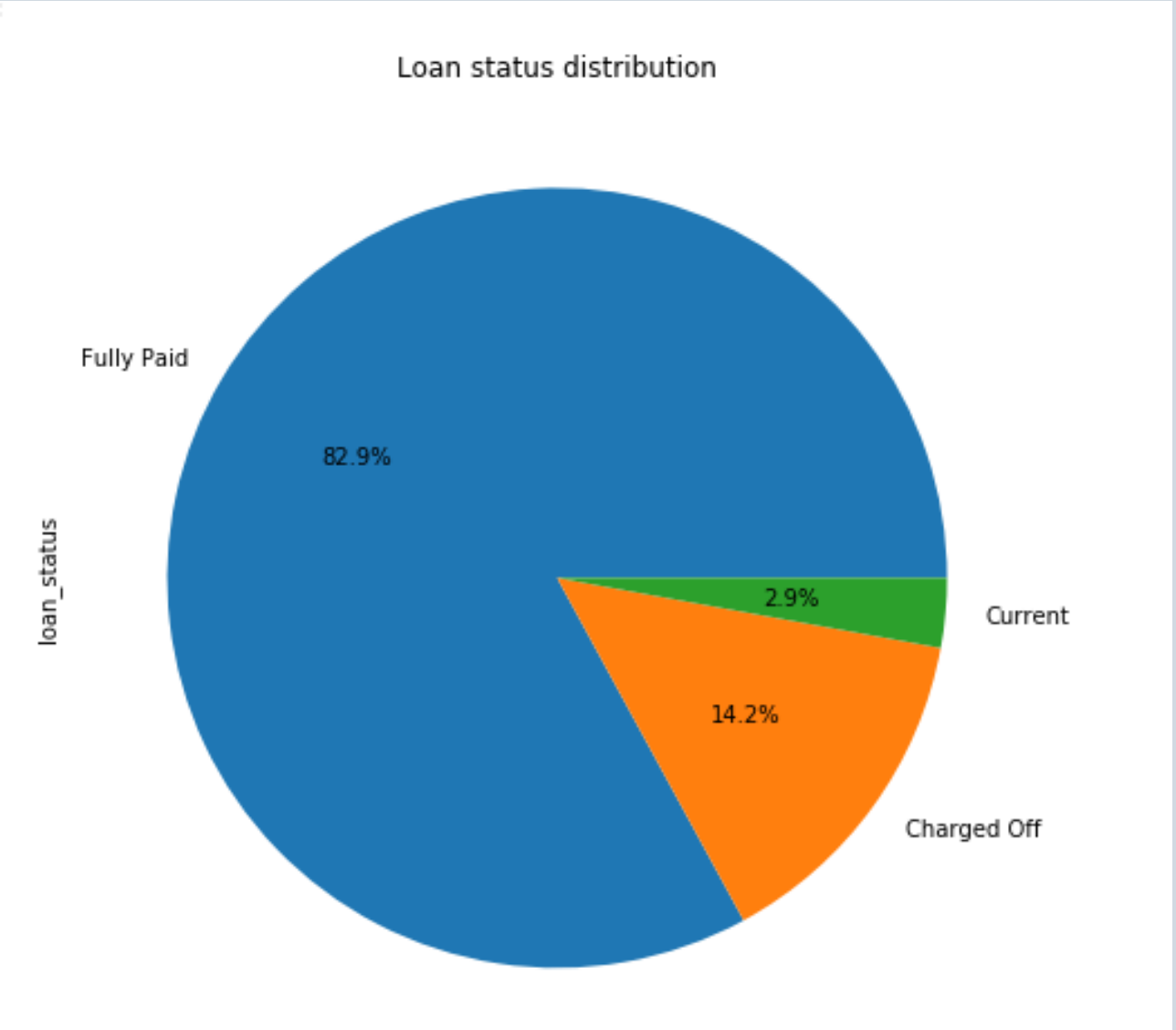
Inference:

People with Annual income below 2 lakhs have taken more loans



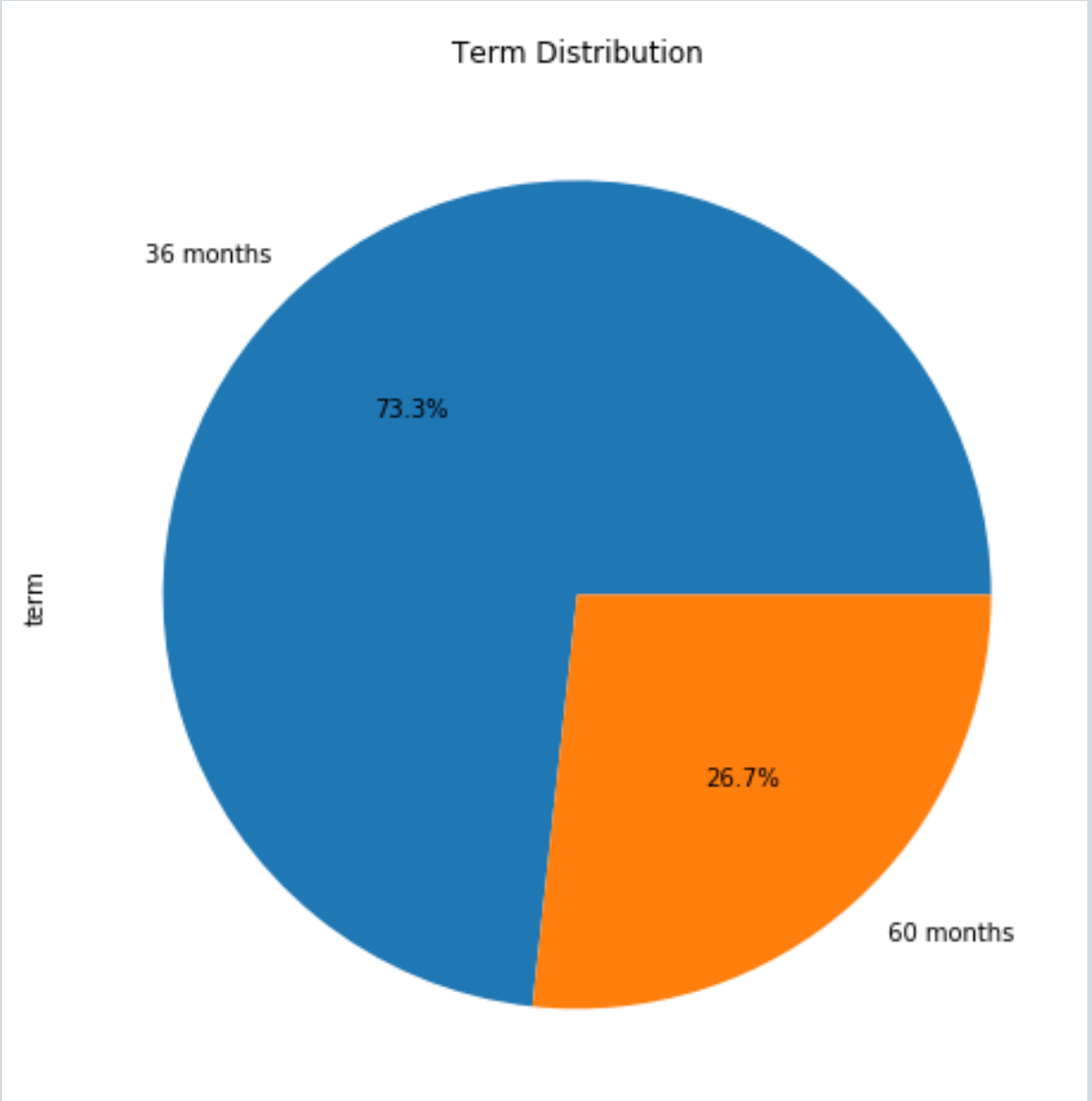
Inference:

14.2% persons are labelled as 'Charged Off'



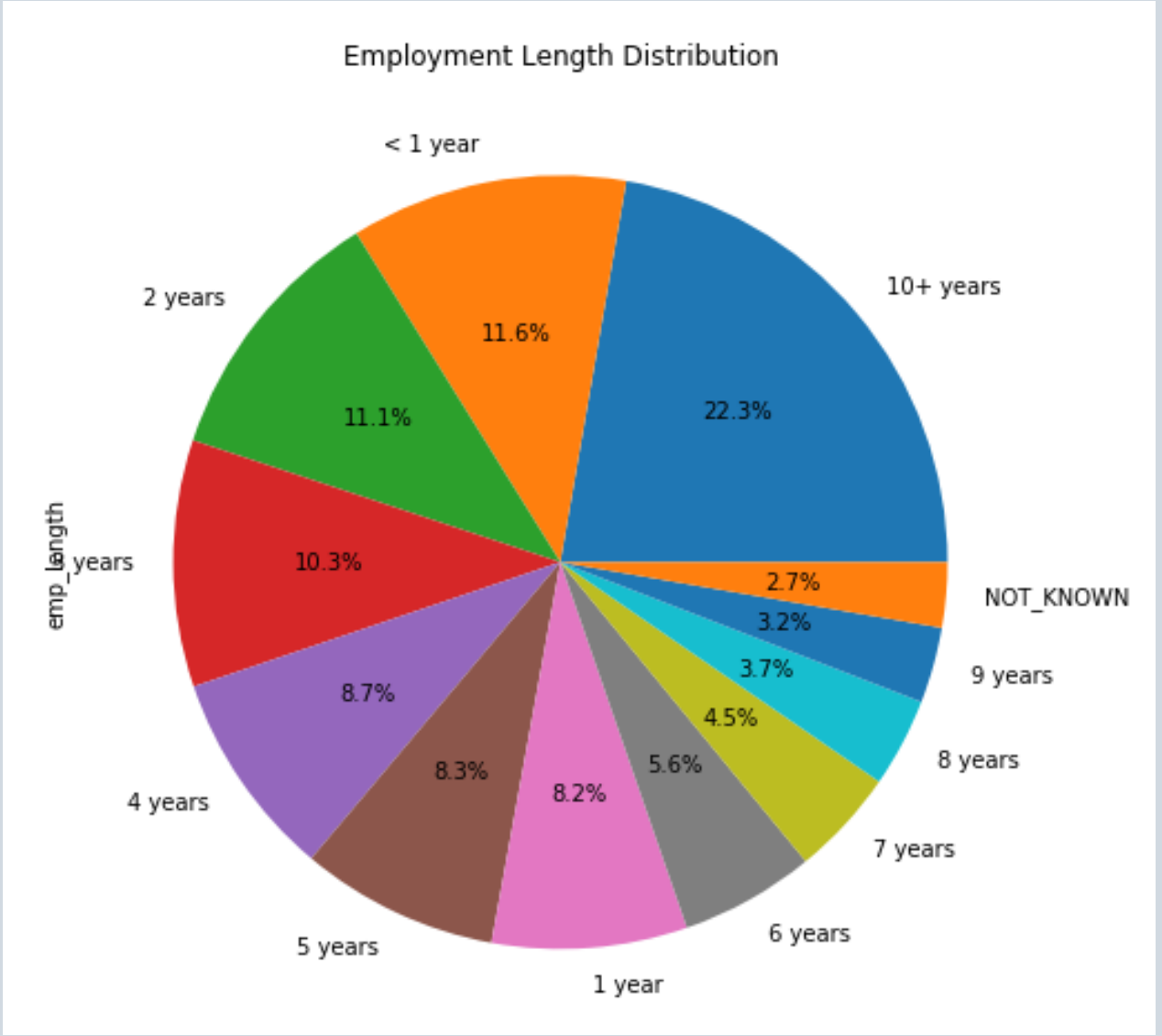
Inference:

14.2% persons are labelled as 'Charged Off'



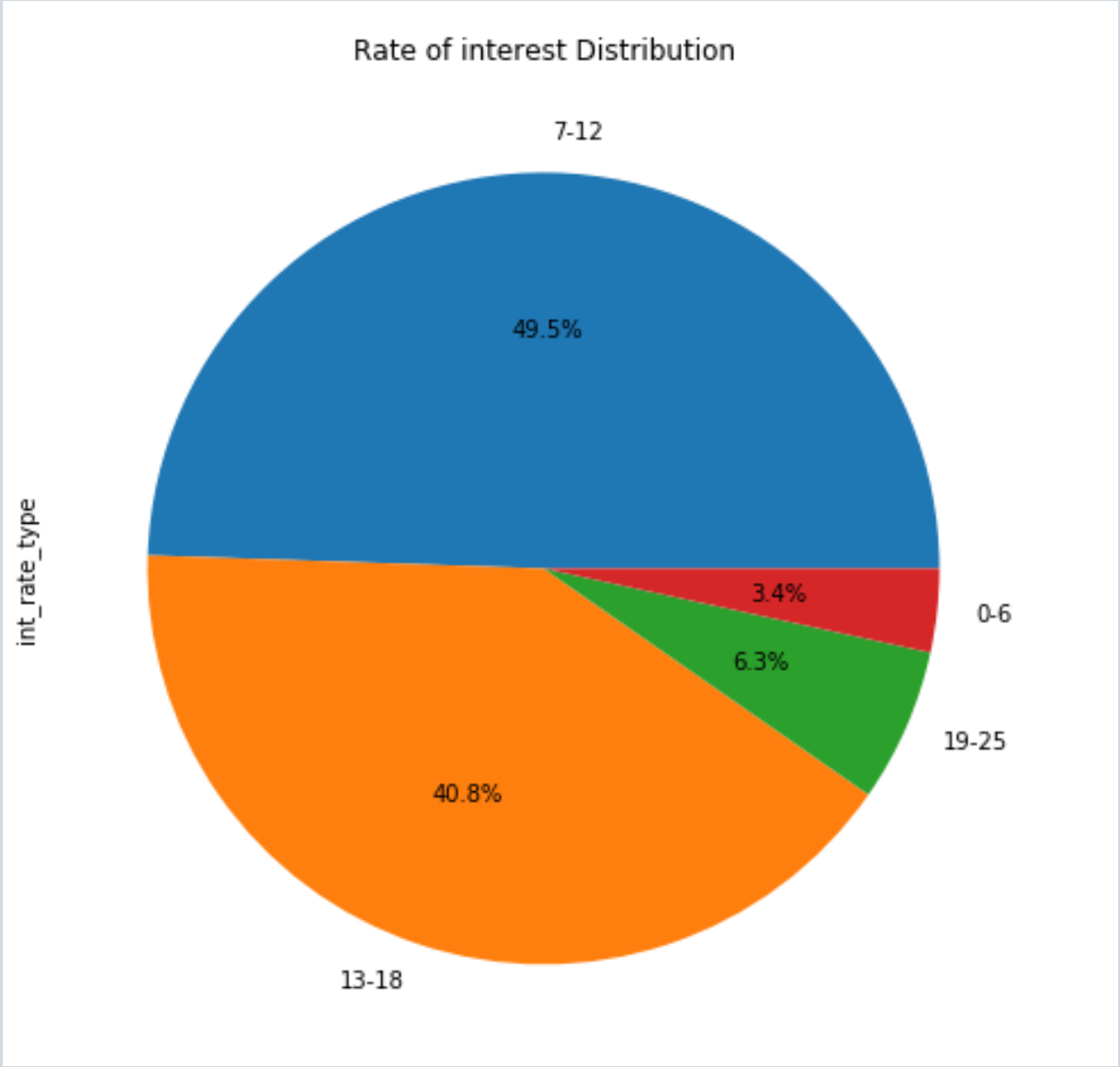
Inference:

14.2% persons are labelled as 'Charged Off'



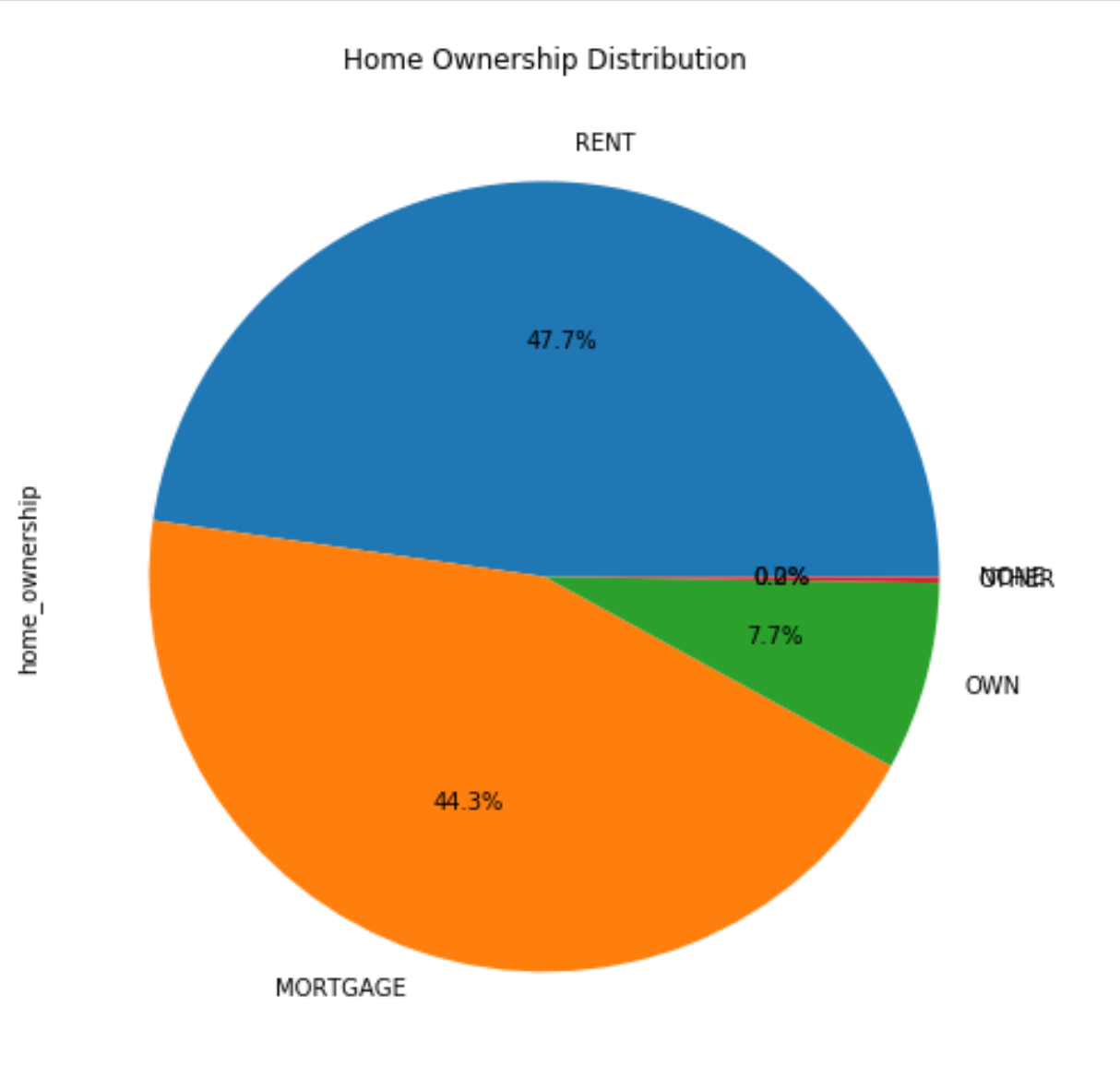
Inference:

People with Employment Length > 10 years have taken more loan



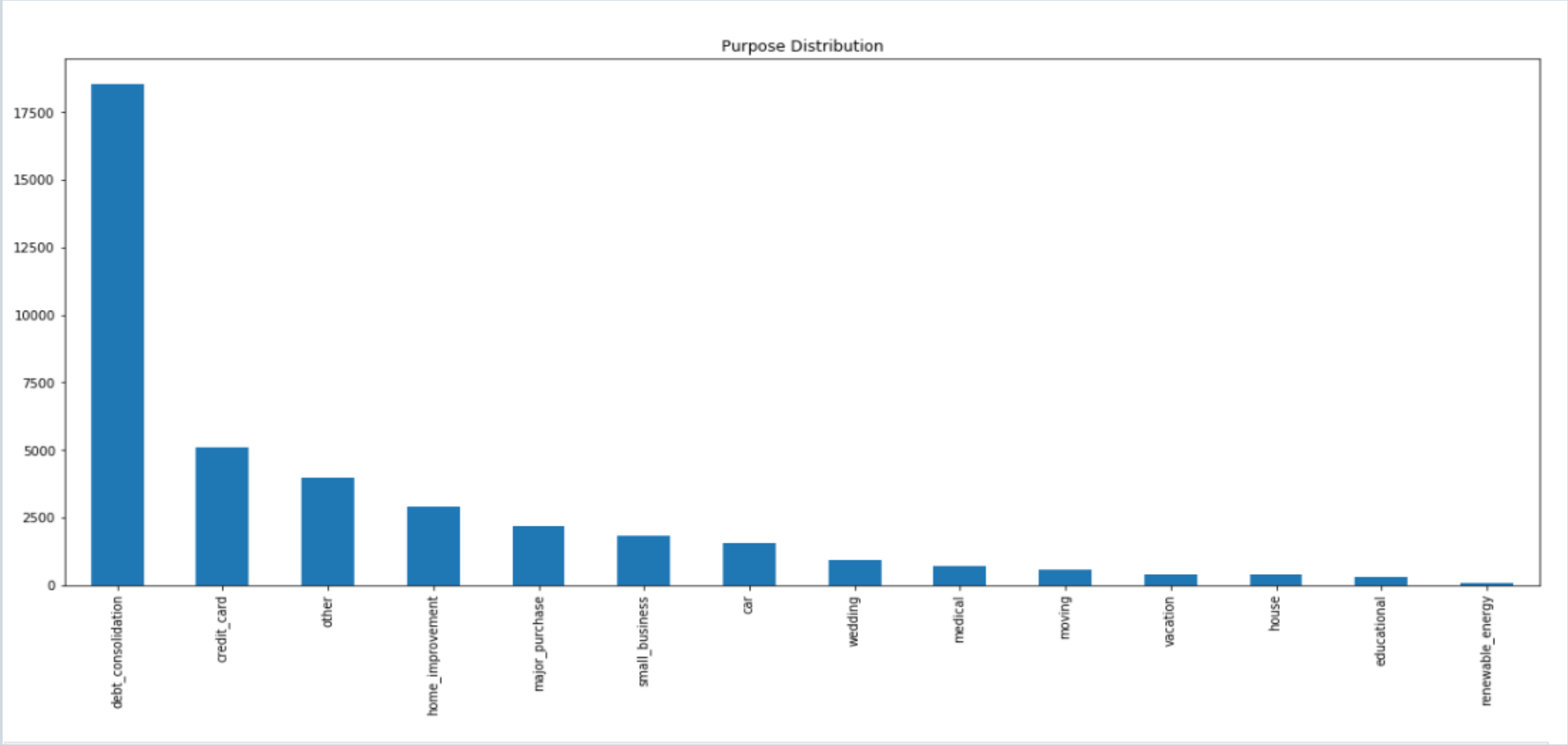
Inference:

People have 7%-12% of rate of interest have opted more loans



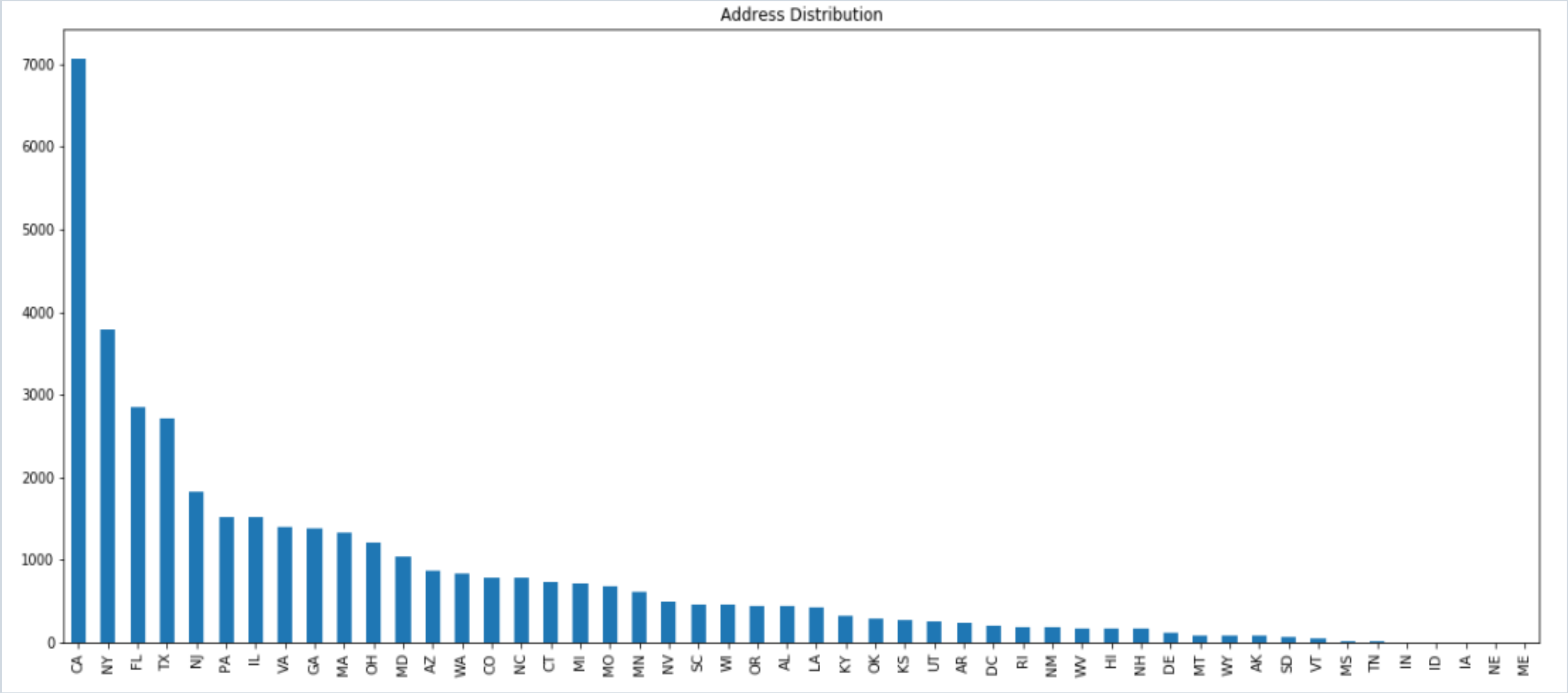
Inference:

People who has RENT or MORTGAGE takes loan



Inference:

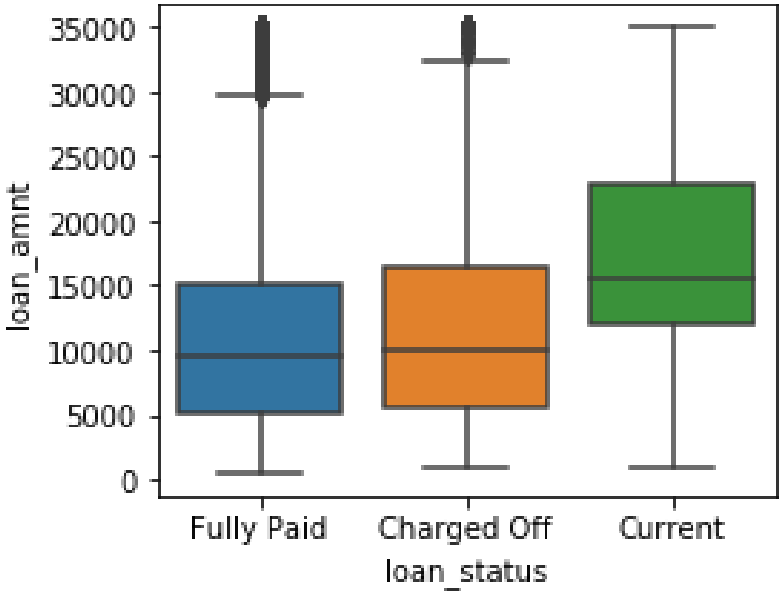
Pepole get loans more for the category 'debt_consolidation'

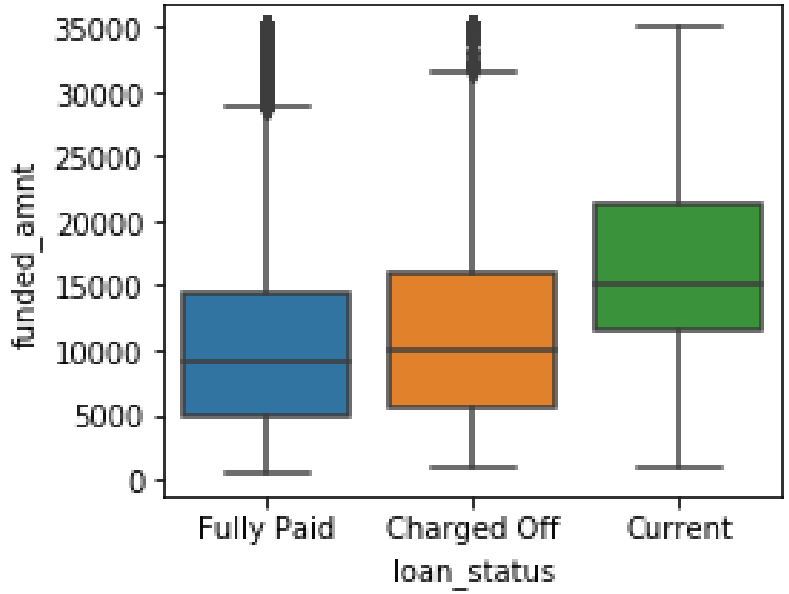


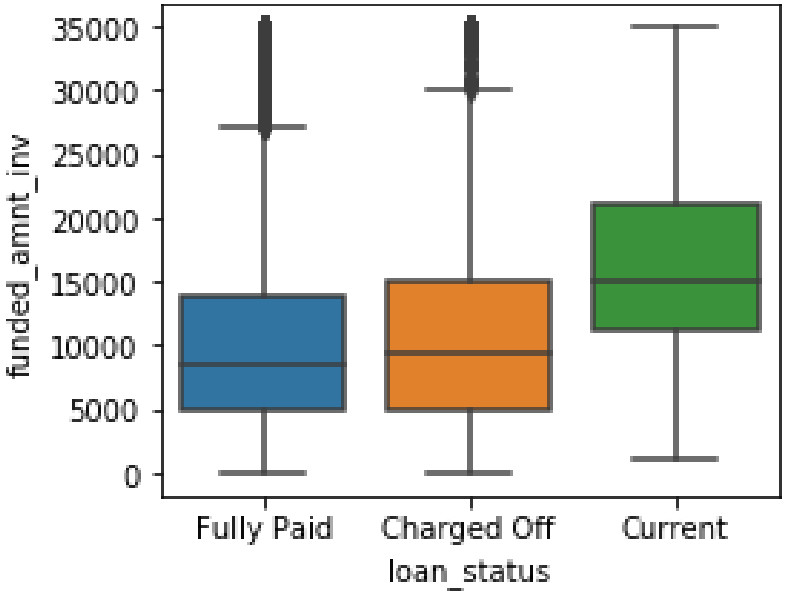
Inference:

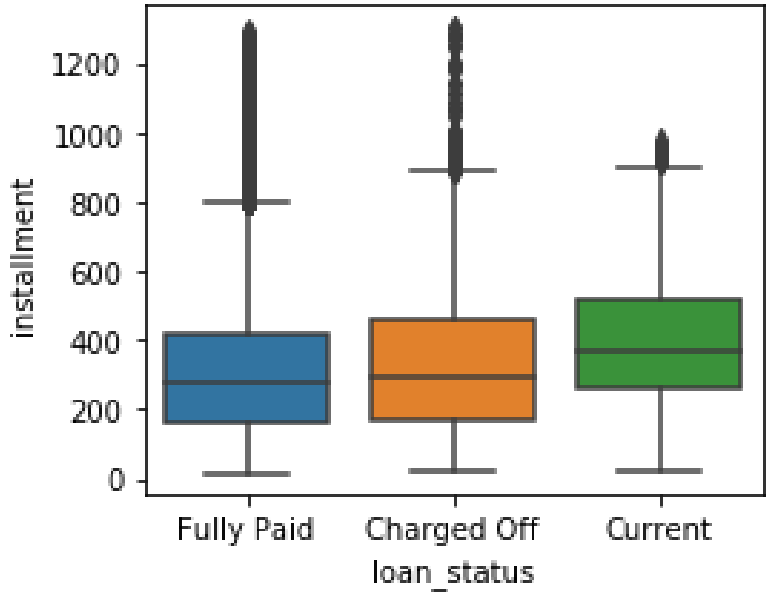
People in 'CA' have taken more loans

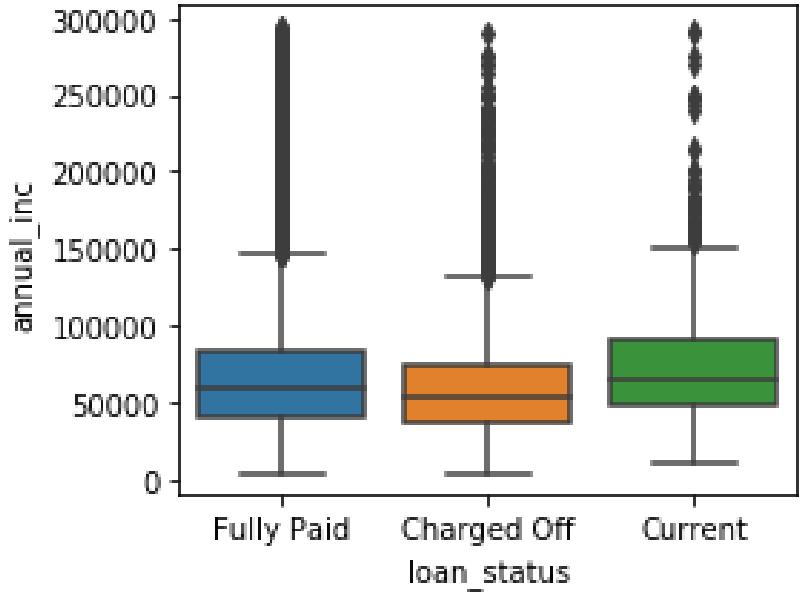
Univariate Analysis - Overall snapshot of numerical variables with Loan Status

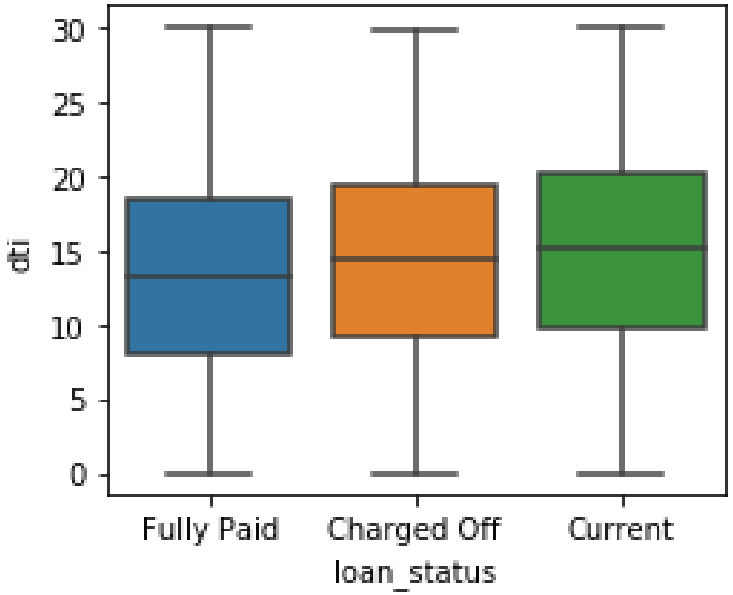

loan_amnt
35000
30000
25000
20000
15000
10000
5000
0
Fully Paid
Charged Off
Current
loan_status

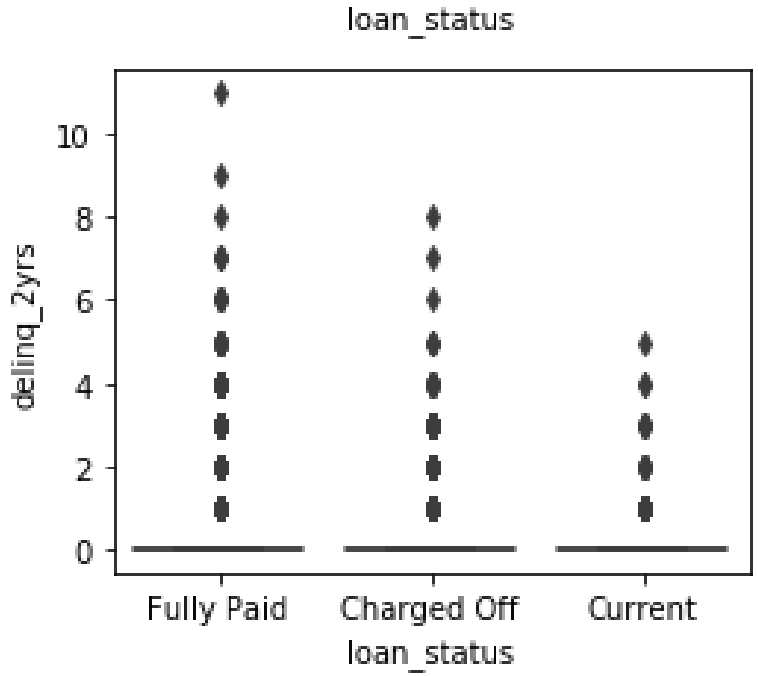

funded_amnt
35000
30000
25000
20000
15000
10000
5000
0
Fully Paid
Charged Off
Current
loan_status

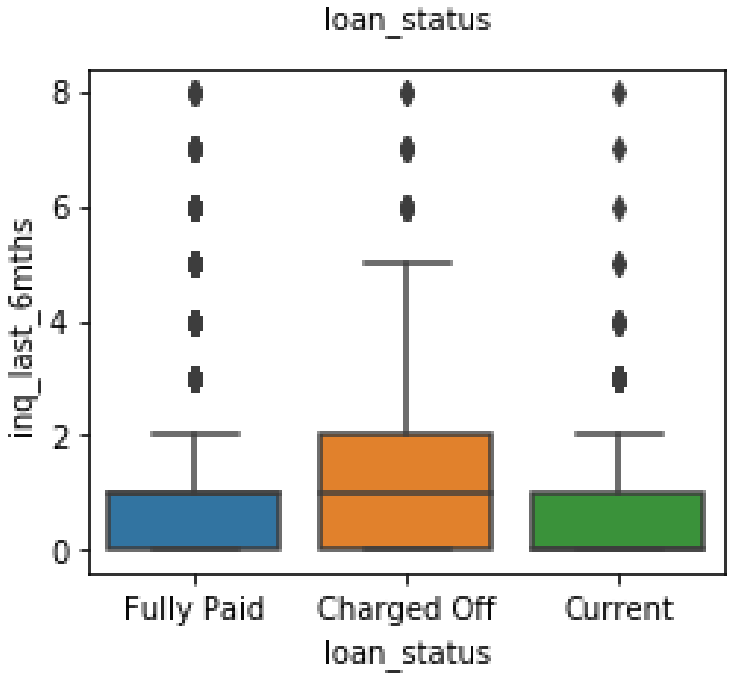

funded_amnt_inv
35000
30000
25000
20000
15000
10000
5000
0
Fully Paid
Charged Off
Current
loan_status

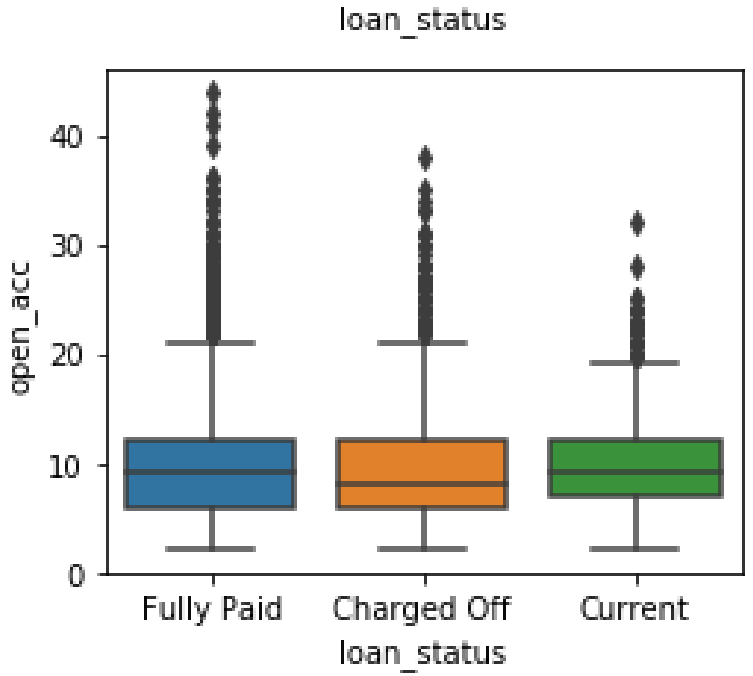

installment
1200
1000
800
600
400
200
0
Fully Paid
Charged Off
Current
loan_status

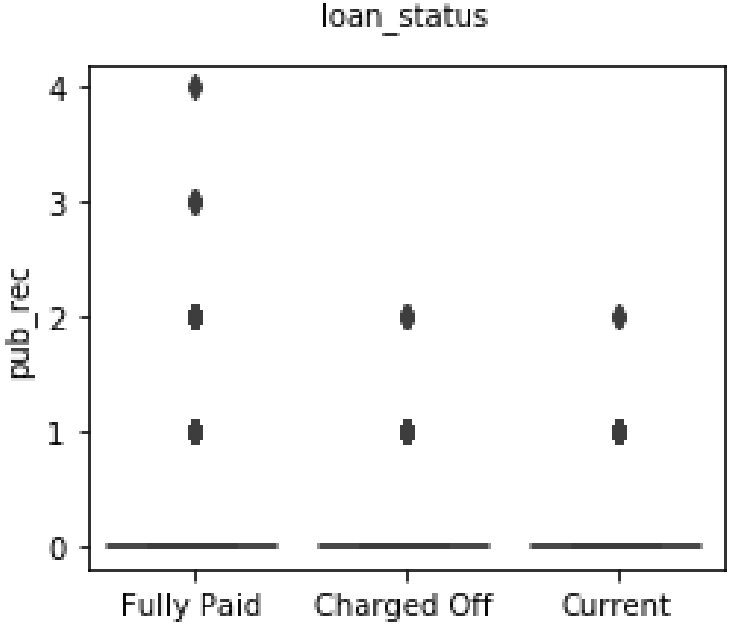

annual_inc
300000
250000
200000
150000
100000
50000
0
Fully Paid
Charged Off
Current
loan_status

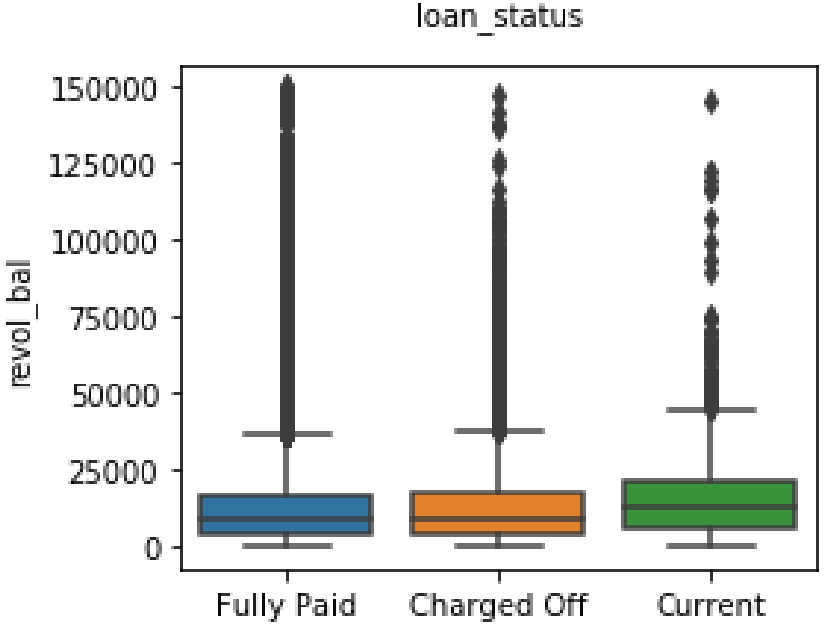

dti
30
25
20
15
10
5
0
Fully Paid
Charged Off
Current
loan_status

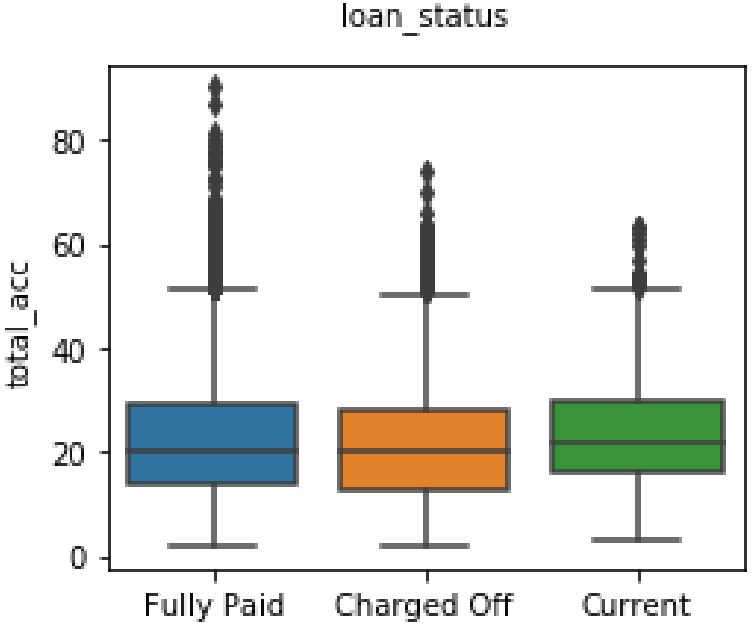

loan_status
delinq_2yrs
0
2
4
6
8
10
Fully Paid
Charged Off
Current

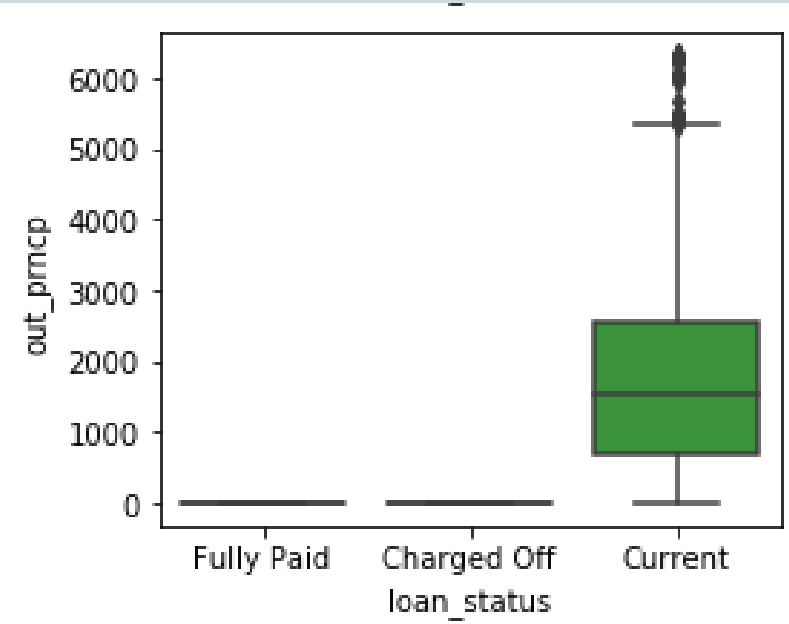

loan_status
inq_last_6mths
0
2
4
6
8
Fully Paid
Charged Off
Current

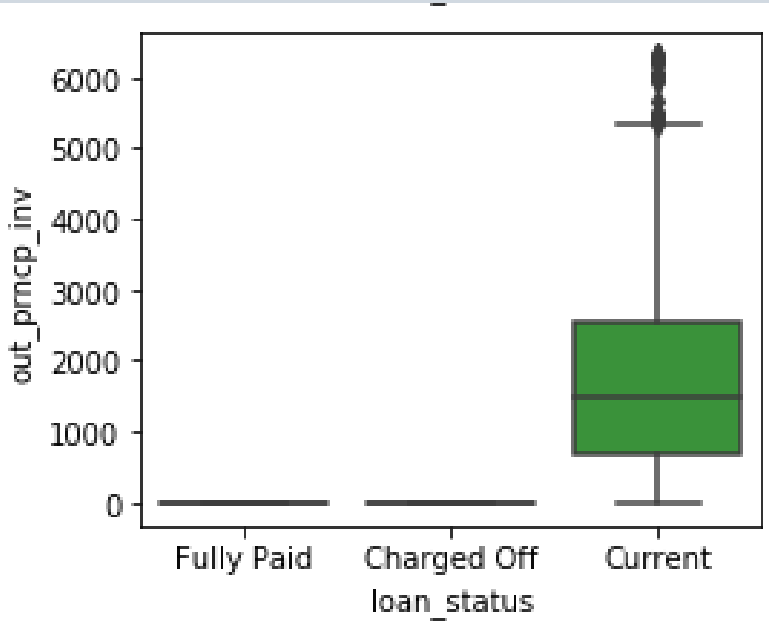

loan_status
open_acc
0
10
20
30
40
Fully Paid
Charged Off
Current

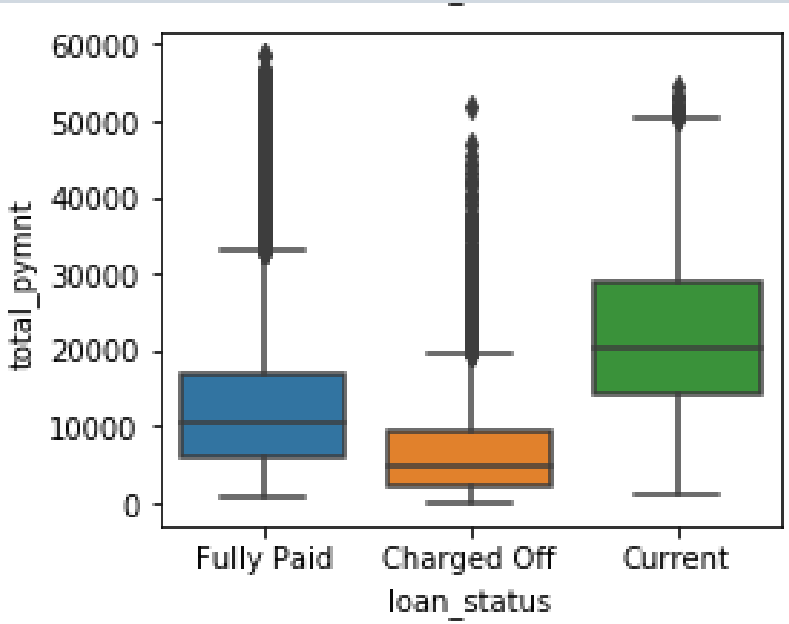

loan_status
pub_rec
0
1
2
3
4
Fully Paid
Charged Off
Current

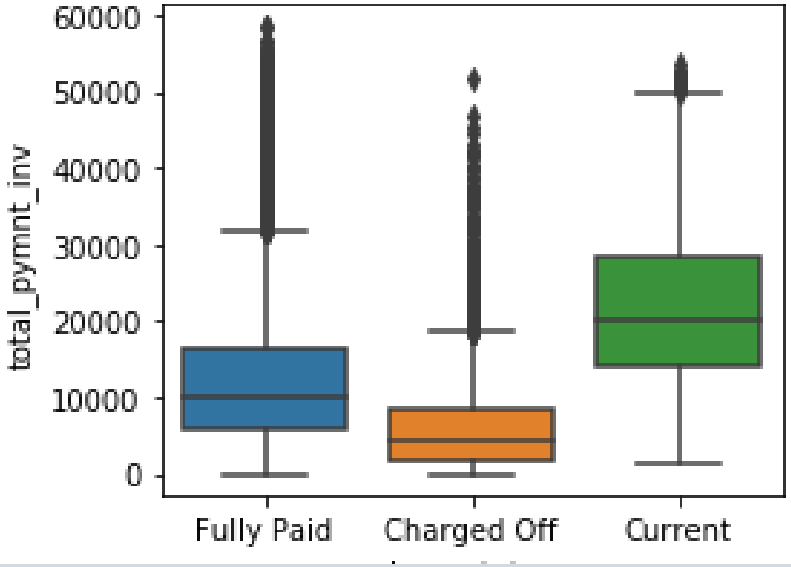

loan_status
revol_bal
0
25000
50000
75000
100000
125000
150000
Fully Paid
Charged Off
Current

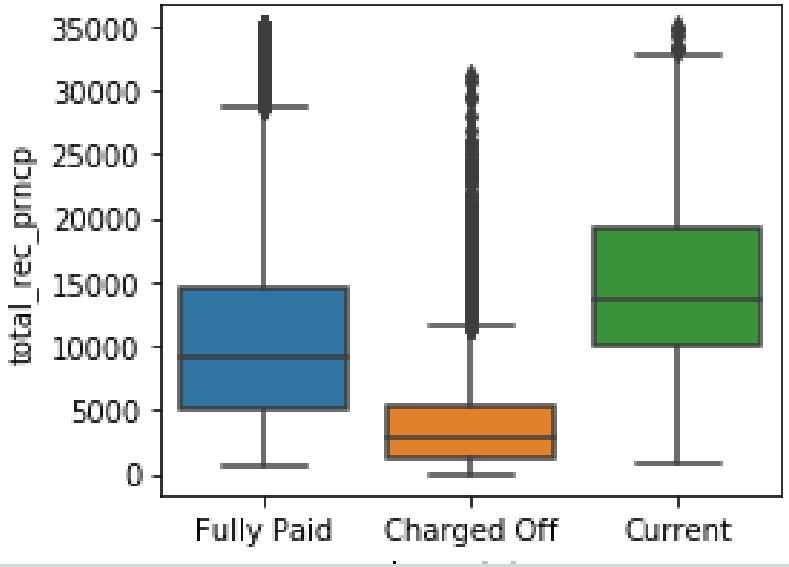

loan_status
total_acc
0
20
40
60
80
Fully Paid
Charged Off
Current

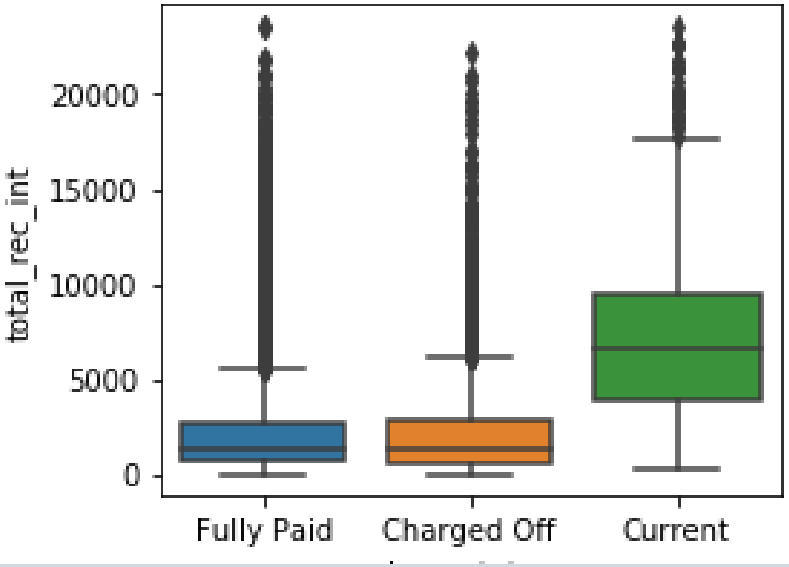

out_prncp
6000
5000
4000
3000
2000
1000
0
Fully Paid
Charged Off
Current
loan_status

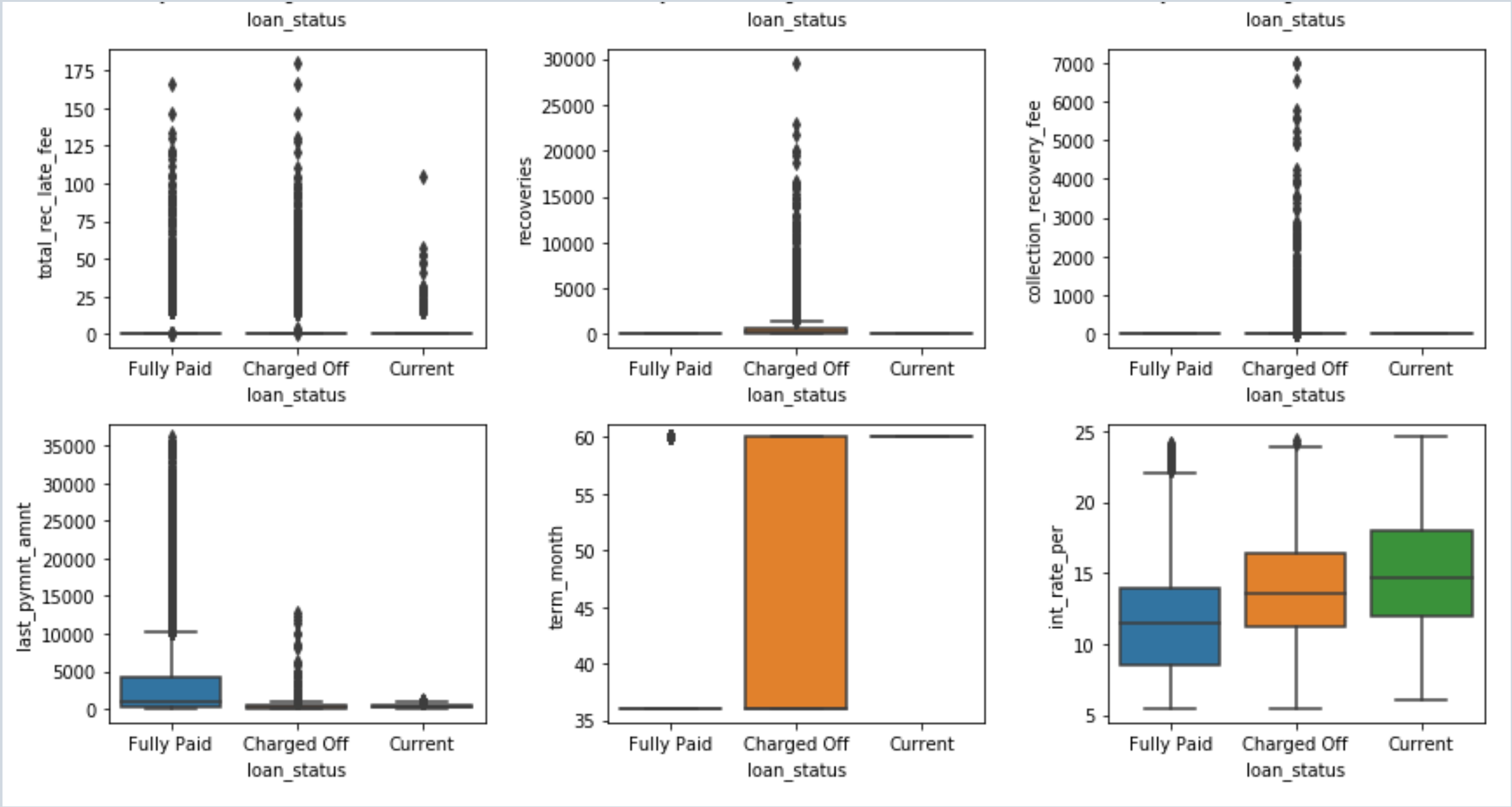

out_prncp_inv
6000
5000
4000
3000
2000
1000
0
Fully Paid
Charged Off
Current
loan_status


total_pymnt
60000
50000
40000
30000
20000
10000
0
Fully Paid
Charged Off
Current
loan_status

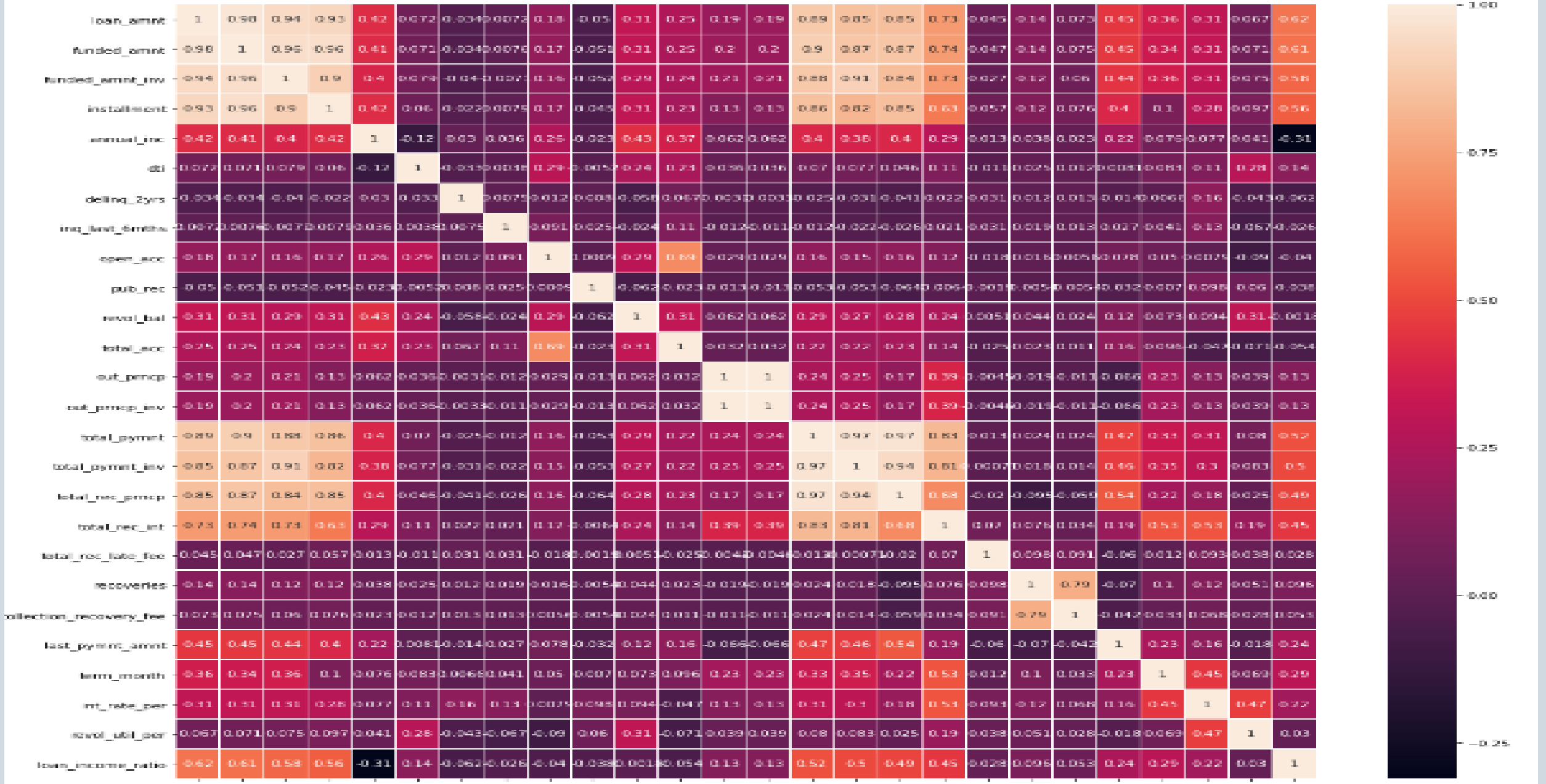

total_pymnt_inv
60000
50000
40000
30000
20000
10000
0
Fully Paid
Charged Off
Current


total_rec_prncp
35000
30000
25000
20000
15000
10000
5000
0
Fully Paid
Charged Off
Current


total_rec_int
20000
15000
10000
5000
0
Fully Paid
Charged Off
Current

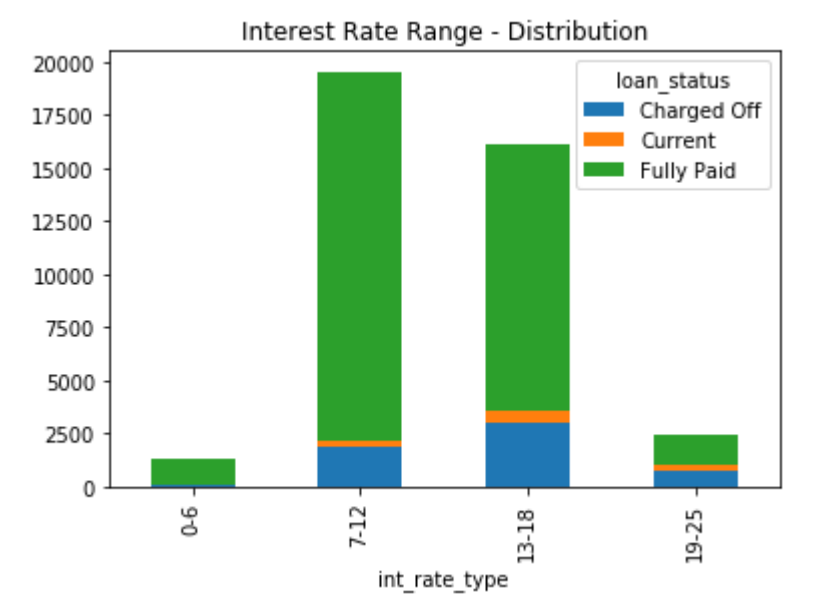

loan_status
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_amnt
term_month
int_rate_per
Fully Paid
Charged Off
Current

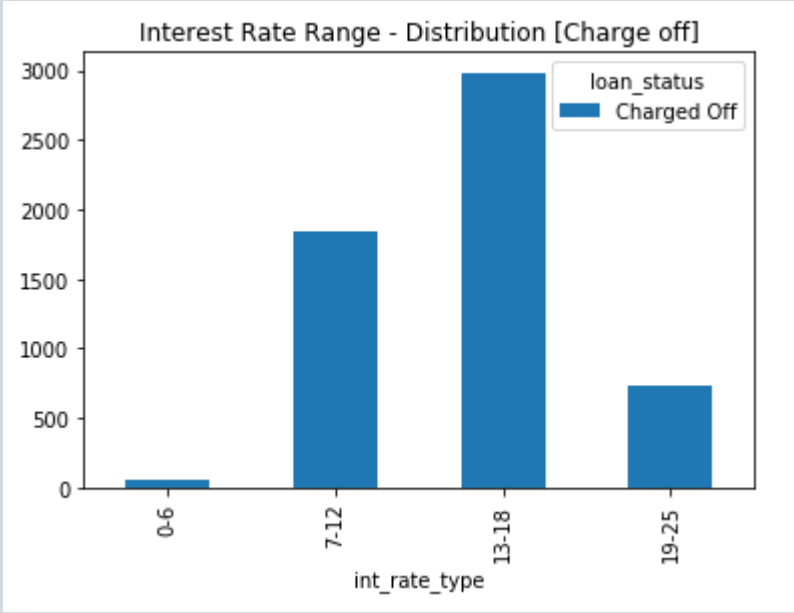
Univariate Analysis - Overall correlation snapshot of numerical variables


loan_amnt
funded_amnt
funded_amnt_inv
installment
annual_inc
dti
delinq_2yrs
inq_last_6mths
open_acc
pub_rec
revol_bal
total_acc
out_prncp
out_prncp_inv
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_amnt
term_month
int_rate_per
revol_util_per
loan_income_ratio
1.00
0.75
0.50
0.25
0.00
−0.25

Inference:

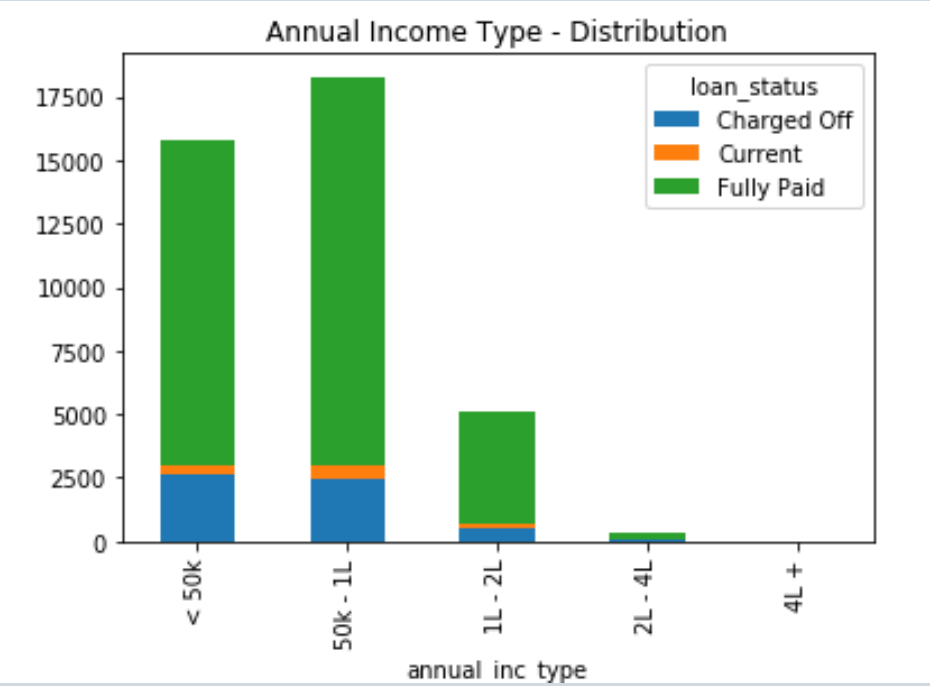
- Correlation between variable that are observed on heatmap
- loan amount, funded amount,funded_amnt_inv and installments show high correlation amongst each other, hence any one column can be taken for consideration for analysis.
- total_payment,total payment_inv,total_rec_prncp and total_rec_int exhibit the same feature.
- pub_rec_bankruptcy and pub_rec show 100% correlation.
- recoveries and collection_recovery show a high correlation of 80% hence one of them can be discarded.
- current applicants show a high correlation to out_prncp and out_prncp_inv

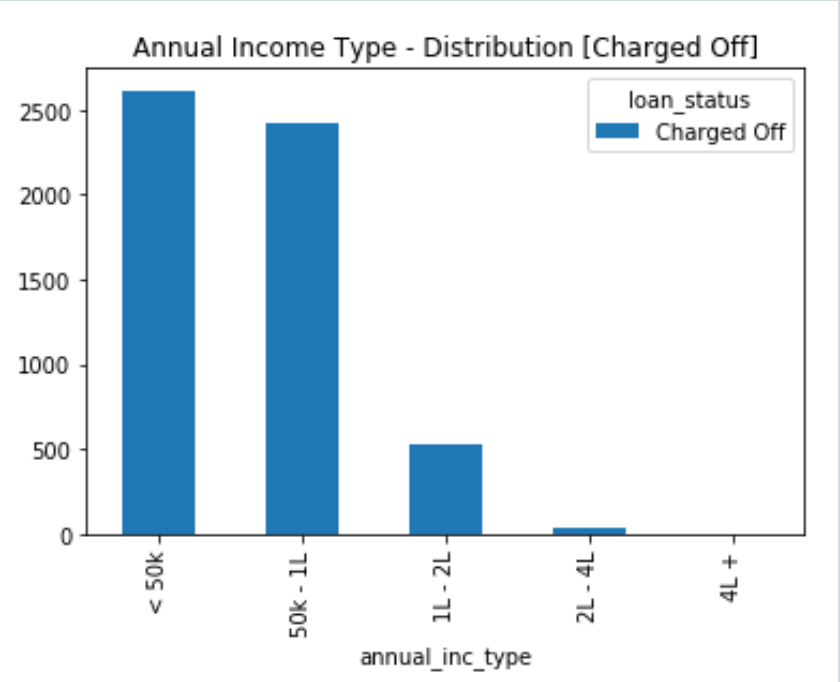




Inference on Rate of Interest (Univariate Analysis) ::

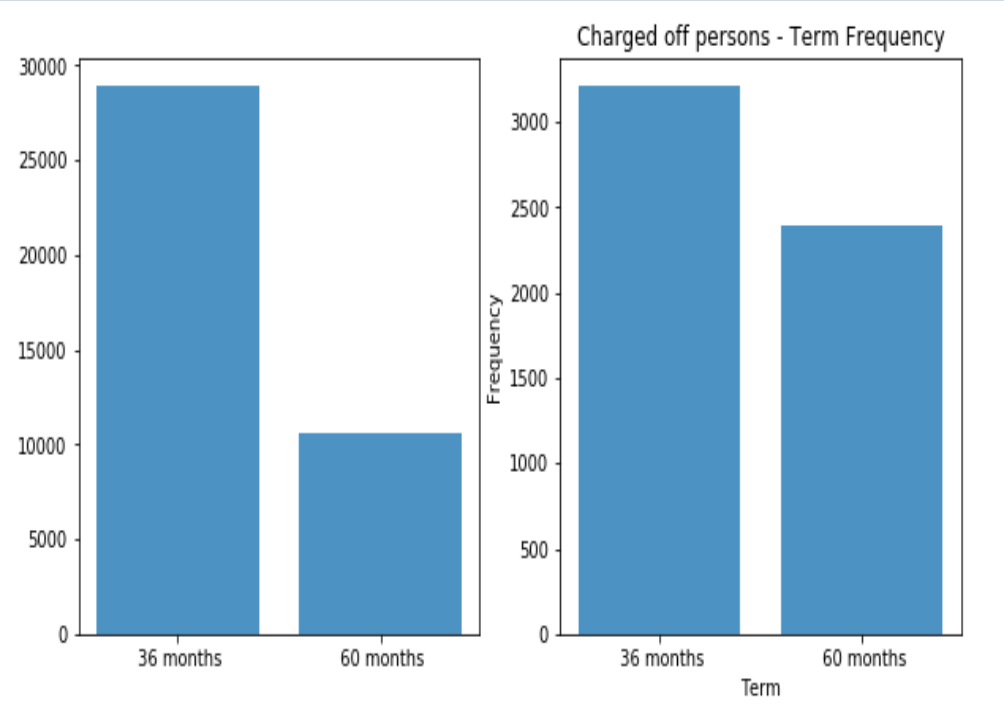
It is clear that People have 7%-12% of rate of interest have opted more loans. But interestingly, people with 13%-18% of rate of interest have more defaulters (i.e Charge off ratio is more in the range of 13-18%)





Inference on Annual Income (Univariate Analysis) :

It is clear that People who have 50k-1L of annual income have opted more loans. But interestingly, people who have annual income <50k have more loan defaulters than people who earns 50k-1L(majority).

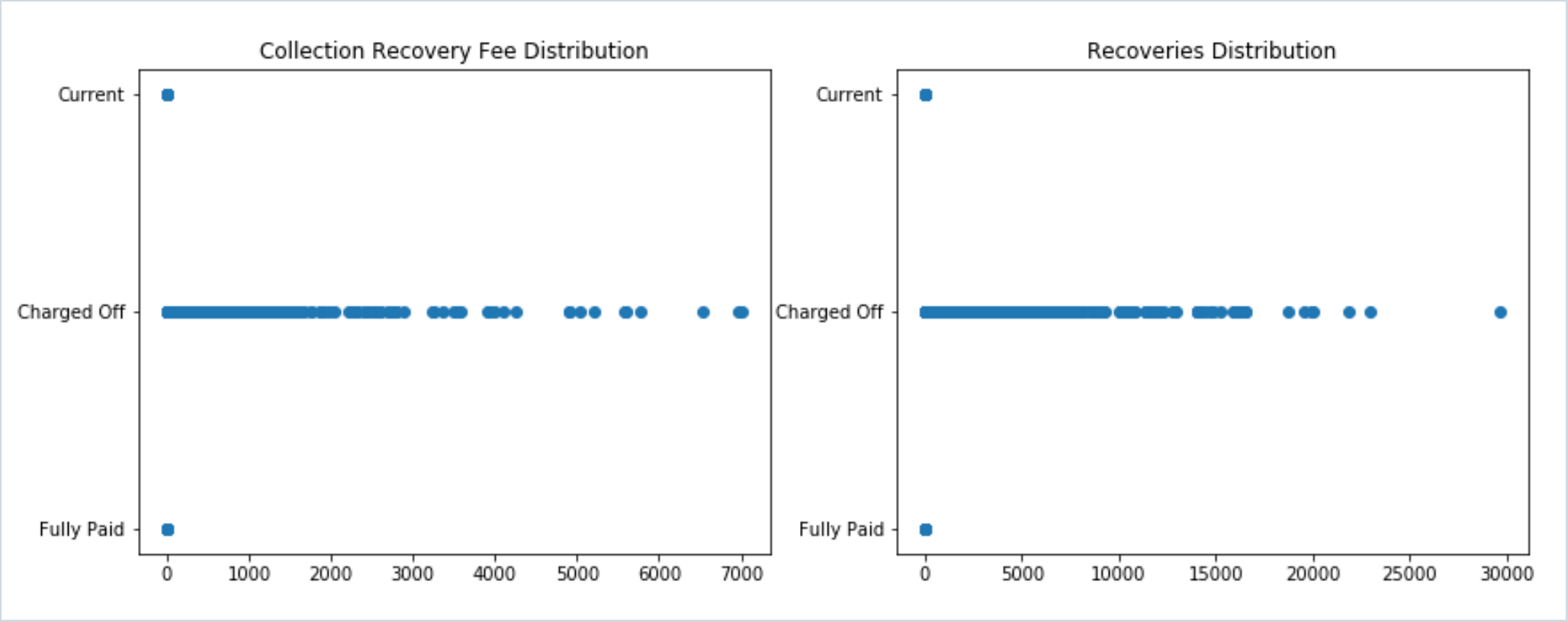


Inference on TERM (Univariate Analysis) :

1) From the graph - 'Charged off persons - Term Frequency', In both 36 months and 60 months, persons who are labelled as 'Charged Off' are available.

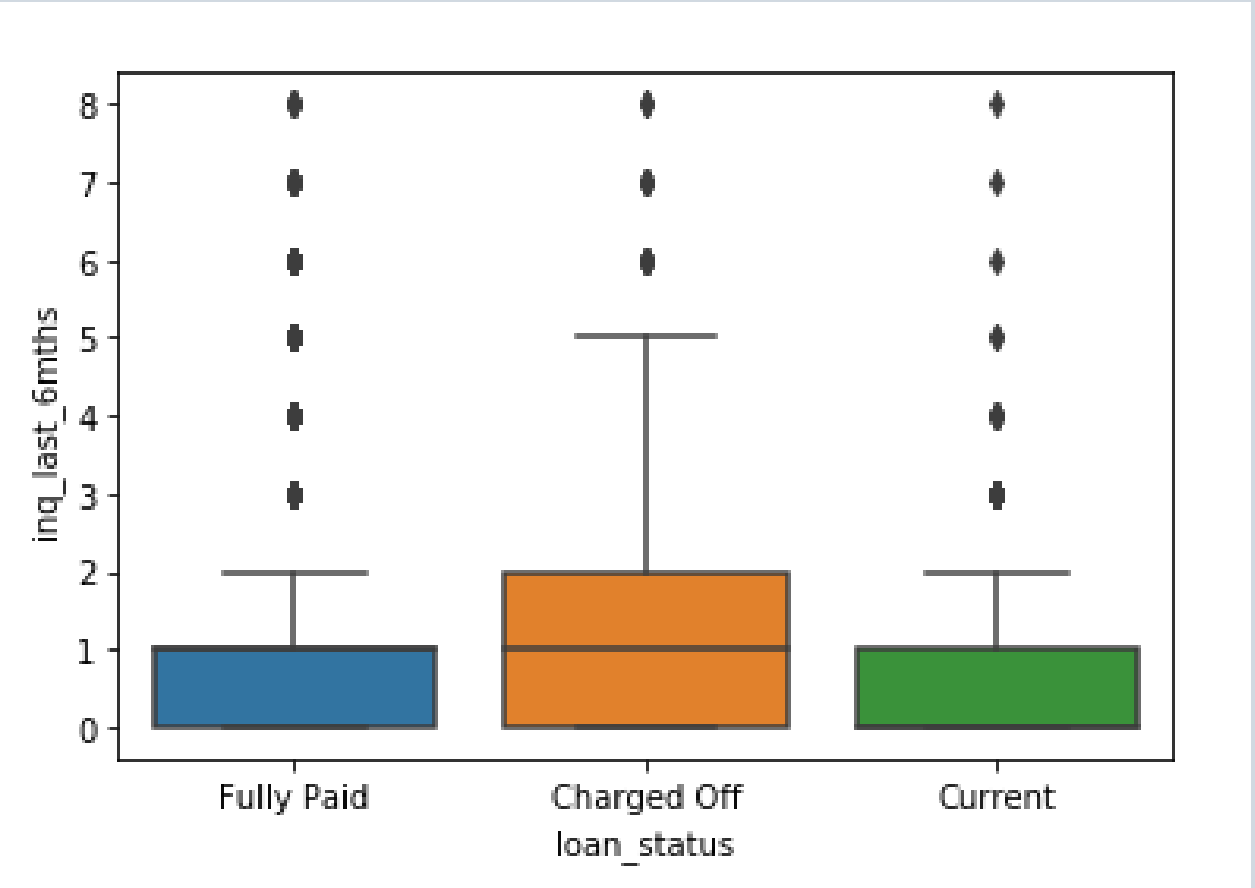
2) On comparing the term ratio (Number of persons who opt for loan in 36 months to 60 months) for the entire dataset with the charged off dataset is 2.739 and 1.344. **Both the ratio are not almost equal and that gives us the interesting pattern.**

3)Visually we can see from two graphs that, at 60 months, there is a raise in the second graph. It implies, defaulters are available in both terms. **But for the term 60 months, more defaulters are found even though the entire dataset has less number of persons who opted for 60 months**



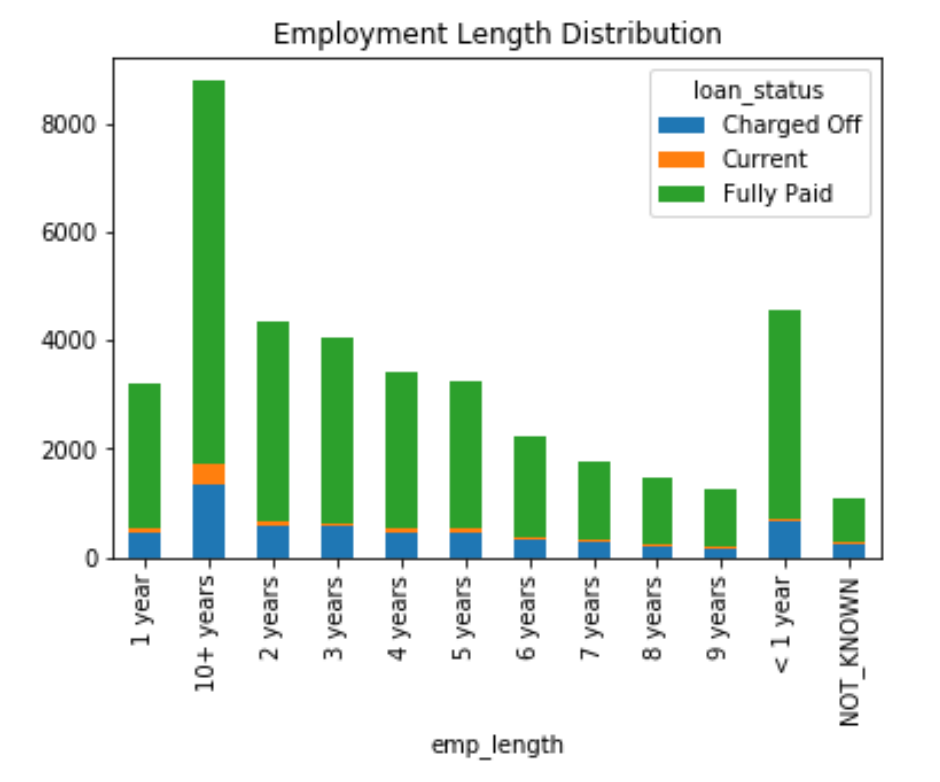
Inference on Collection Recovery Fee & Recoveries (Univariate Analysis) :

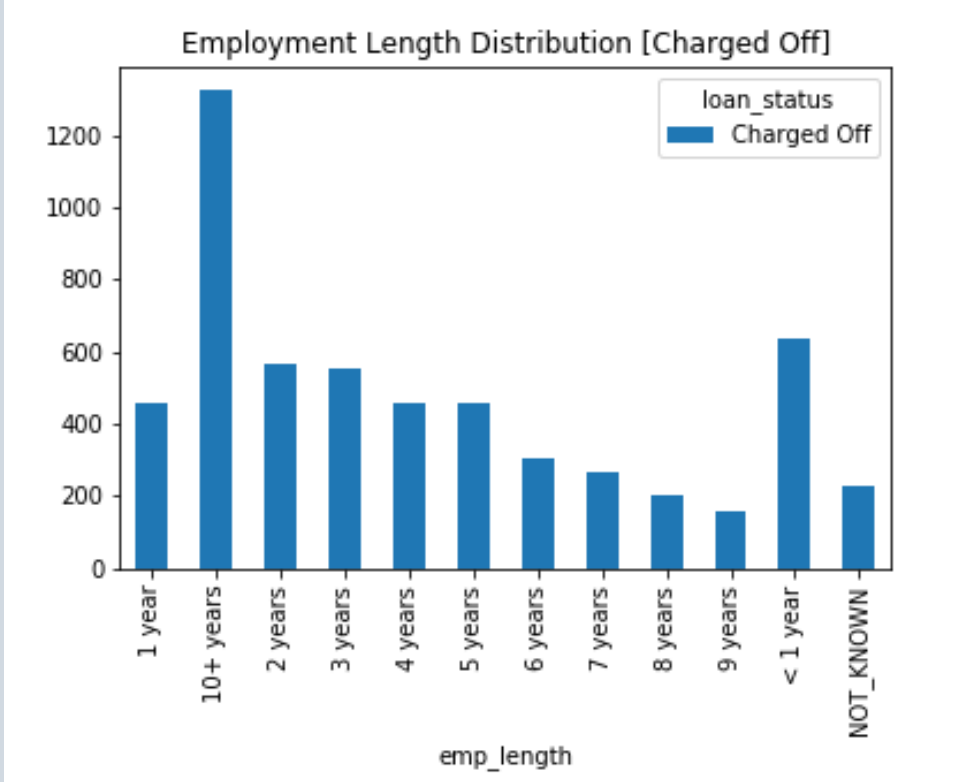
It is very obivious that the Charged off persons will be charged with Collection Recovery Fee and Recoveries.So, it is evident that these 2 columns strongly contribute to the defaulters



Inference on inq_last_6mths (Univariate Analysis):

It is evident from the graph that, frequency of the people who enquired 2 times or more in the last 6 months, tend to default more.

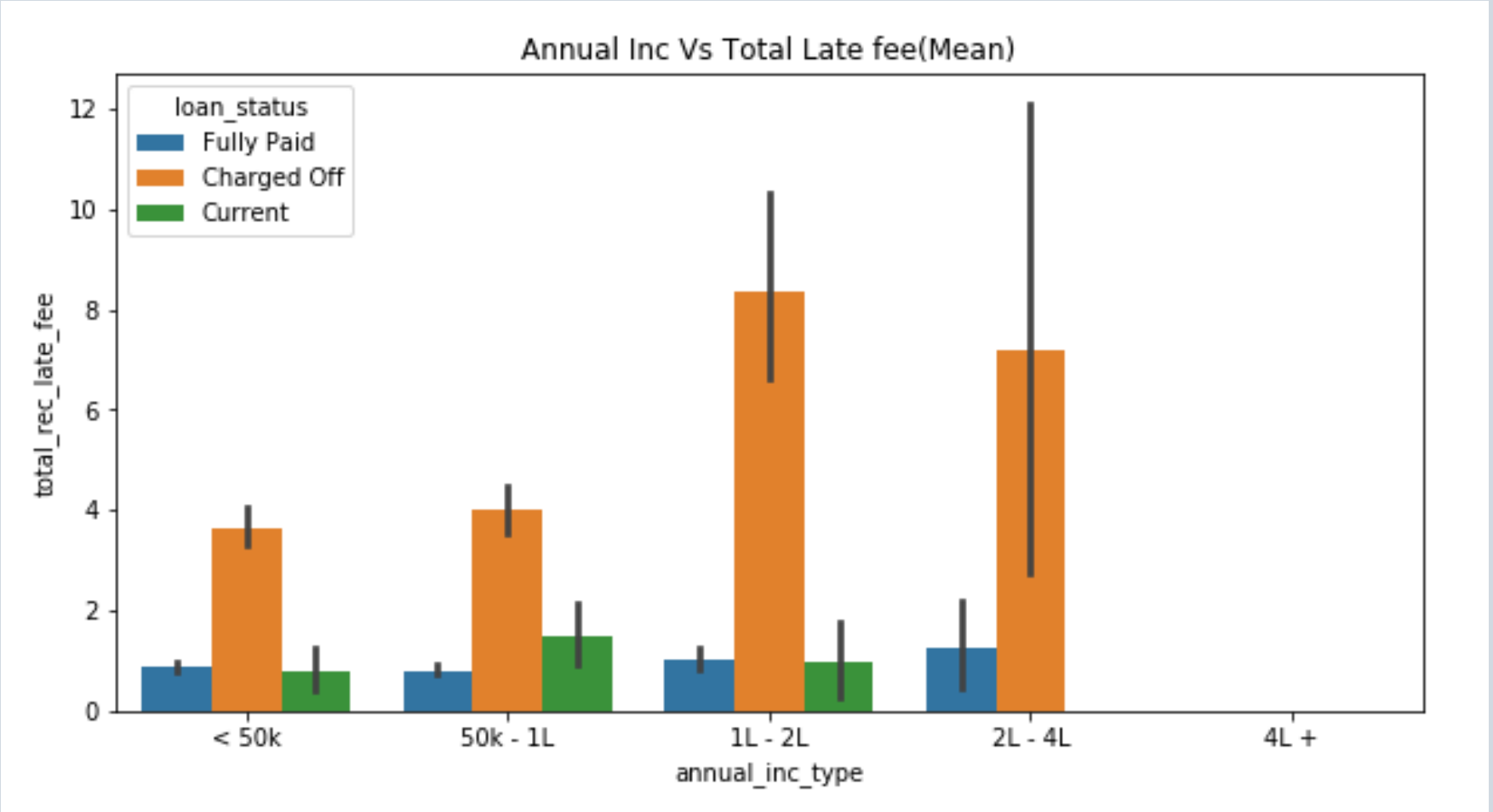




Inference on emp_length (Univariate Analysis):

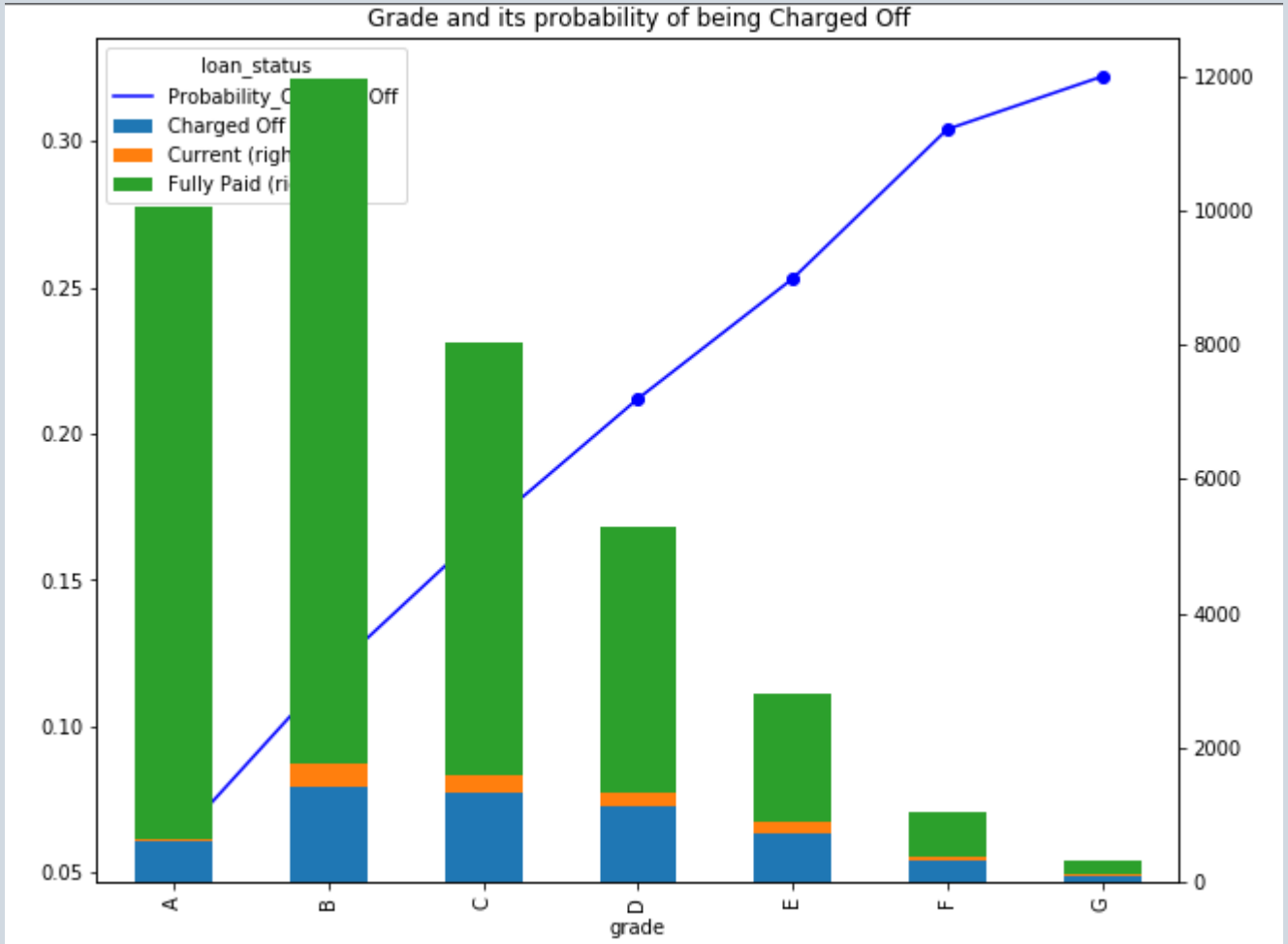
1) It is evident from the graph that, probability of the frequency of the people who charged off are those who are less than a year experienced

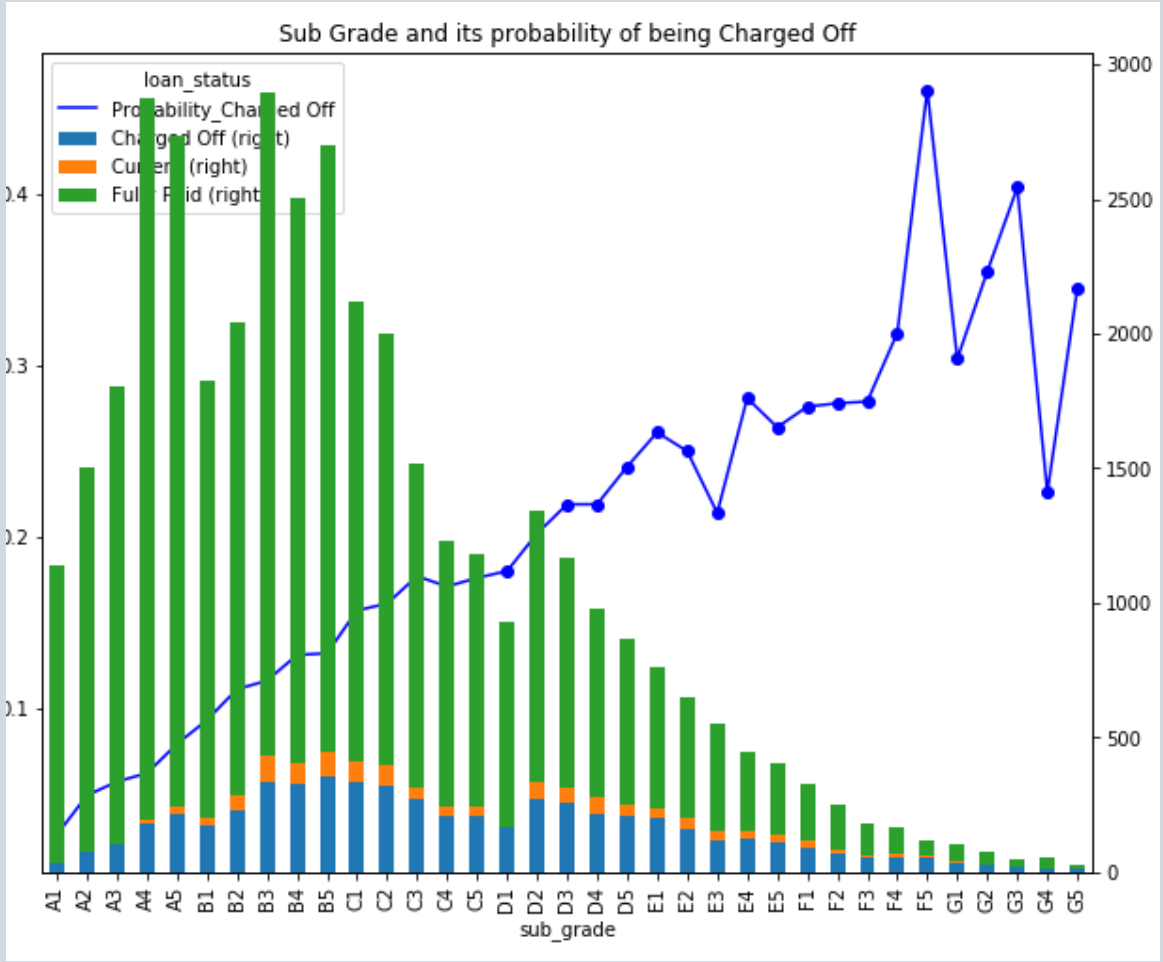
2) We cannot count in 10+ experienced. Why? - As the frequency of the loan is more for 10+ experienced people, charge off count is more. With respect to the ratio of "ChargeOff/Total", person whose employment <1 year tends to be a defaulter



Inference on Annual Income with LateFee on Loan Status: (Bivariate Analysis):

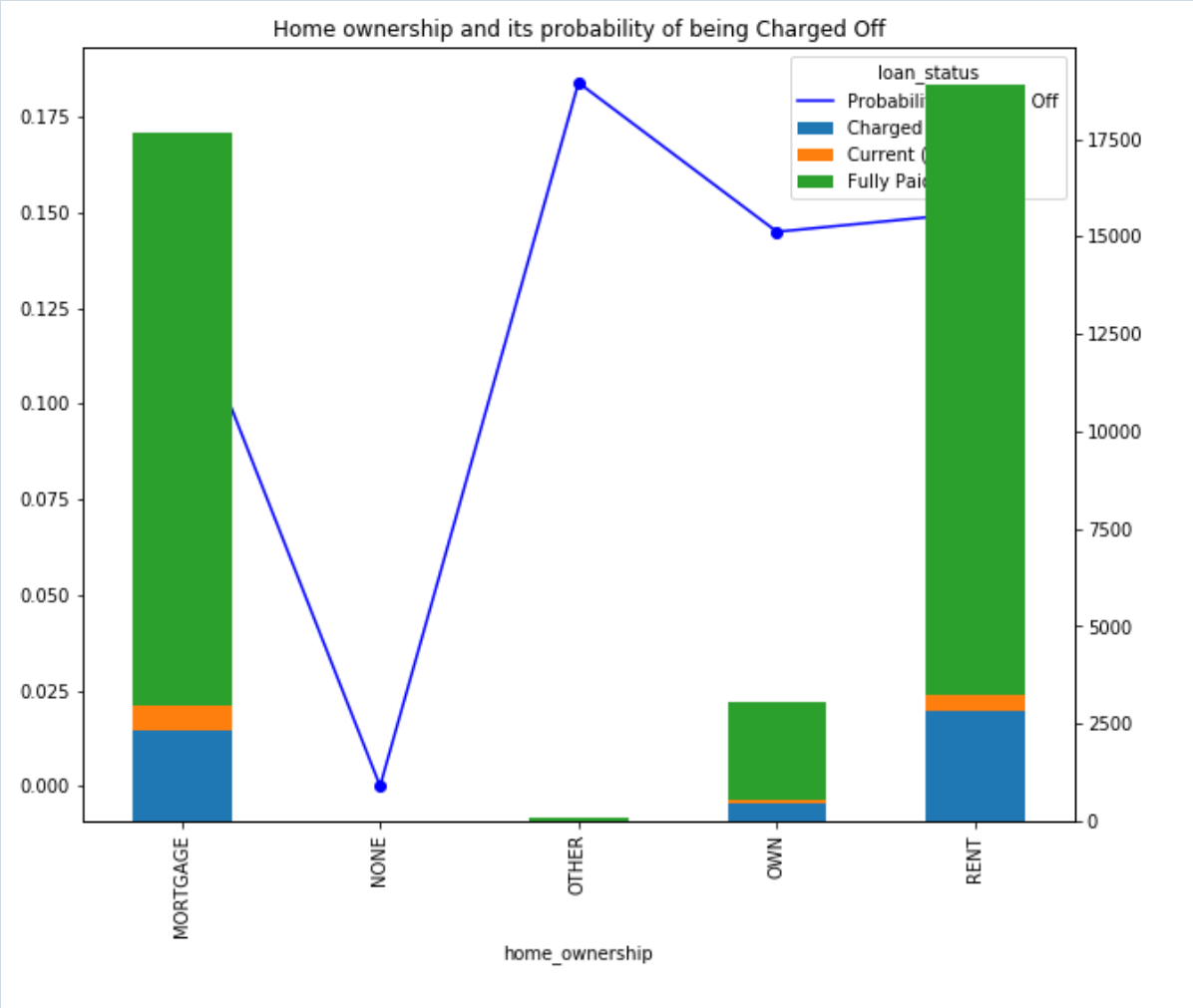
From the graph on 'Annual Inc Vs Total Late fee (Mean)', it is known that people whose income falls between 1 Lakh - 4 Lakh and those who avail late fee tends to be a defaulter more.





Inference - Grade/SubGrade (Bivariate Analysis):

As the Grade increases, the probability of charge off is also increased.



Inference - Home Ownership (Bivariate Analysis):

1) People who stated 'OTHER' as their Home ownership tend to charge off more.
2) Next to 'OTHER', people who stated 'RENT' as their ownership tend to charge off more