LENDING CLUB CASE STUDY

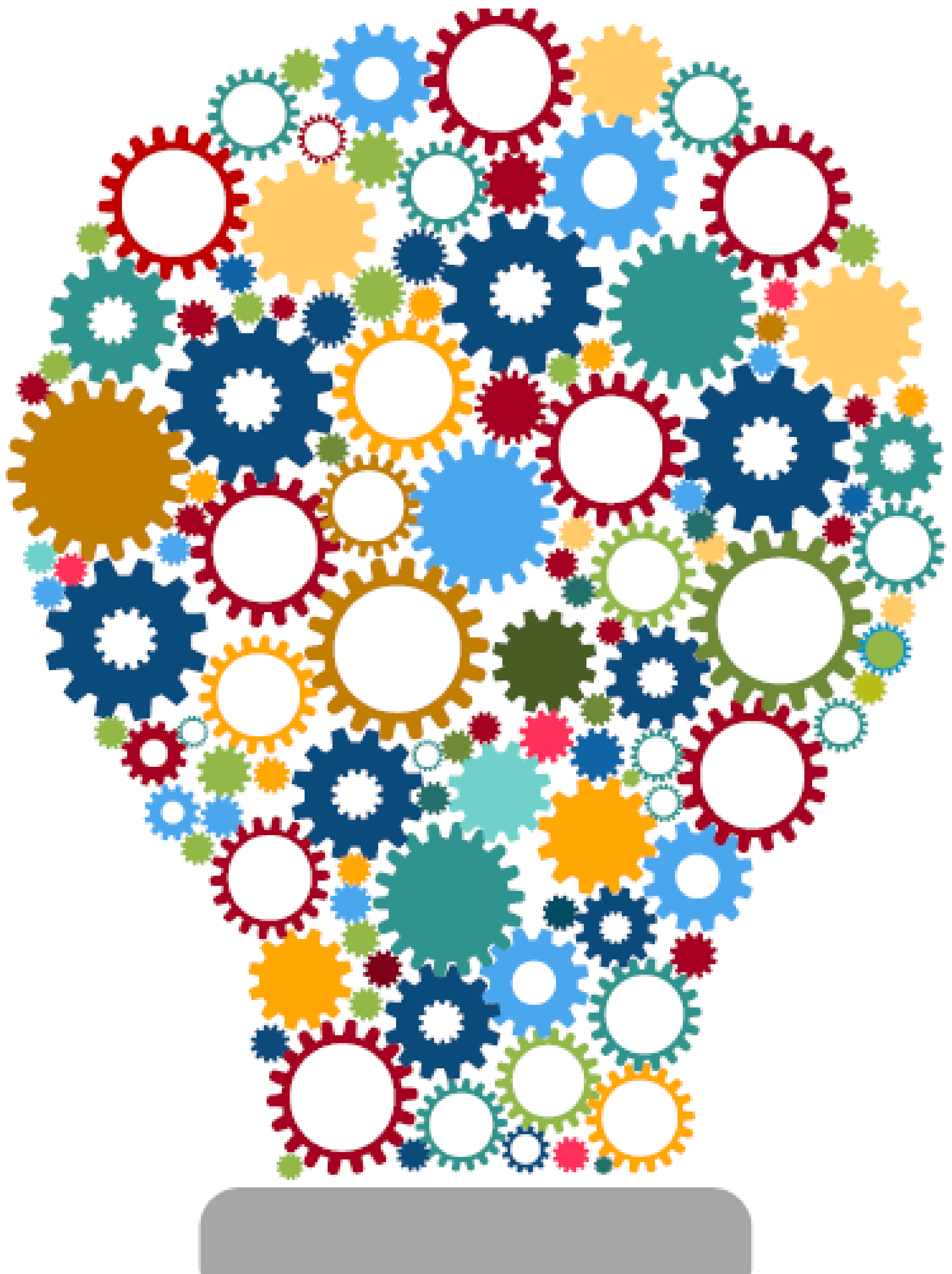
Aditi Agrawal

PROBLEM STATEMENT

APPROACH

ANALYSIS

SUMMARY



PROBLEM STATEMENT

You work for a consumer finance company which specialises 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. 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.

APPROACH

DATA CLEANING

Remove unwanted columns, handle missing values, convert fields to required datatypes

UNIVARIATE ANALYSIS

Categorical variables are analyzed.
Bar plots are used
Continuous variables are analyzed.
Box plots and histograms are used.

SEGMENT ANALYSIS

Funded amount is plotted against verification_status, sub grade, state for individual loan_status

BIVARIATE ANALYSIS

Continuous variable vs continuous variables
Correlation table created for numerical data

CORRELATION ANANLYSIS

DATA CLEANING

Checked for columns that has NA/ null values more than 50% which was for 57 columns out of 111, and removed those 57 columns as no significant analysis can be made from them.

[3]:
```
# Determining the shape of the datset

loan_dataset.shape
```

[3]: (39717, 111)

[4]:
```
# Cleaning the missing data
# listing the null values columns having more than 50%

emptycol=loan_dataset.isnull().sum()
emptycol=emptycol[emptycol.values>(0.5*(loan_dataset.shape[0]))]
len(emptycol)
```

[4]: 57

So, there are 57 columns having null values greater than 50% in the dataset

[5]:
```
# Removing those 58 columns

emptycol = list(emptycol[emptycol.values>=0.3].index)
loan_dataset.drop(labels=emptycol,axis=1,inplace=True)
print(len(emptycol))
```

57

[6]:
```
# Determining the shape of the datset

loan_dataset.shape
```

[6]: (39717, 54)

DATA CLEANING

Removing categorical columns that doesn't help in analysis.

Column_Name	Reason
emp_title	Not used
title	Not used
application_type	Same value for all rows
url	Not used
pymnt_plan	Same value for all rows
initial_list_status	Same value for all rows
policy_code	Same value for all rows

```
loan_dataset.drop(labels=['emp_title', 'title', 'application_type', 'url', 'pymnt_plan',
                          'initial_list_status', 'policy_code', 'desc'], axis=1, inplace=True)
loan_dataset.shape
```

```
(39717, 46)
```

DATA CLEANING

Checking the columns having less null percentage

loan_dataset.isnull().sum()/len(loan_dataset)*100

Fill the missing values and fix formats

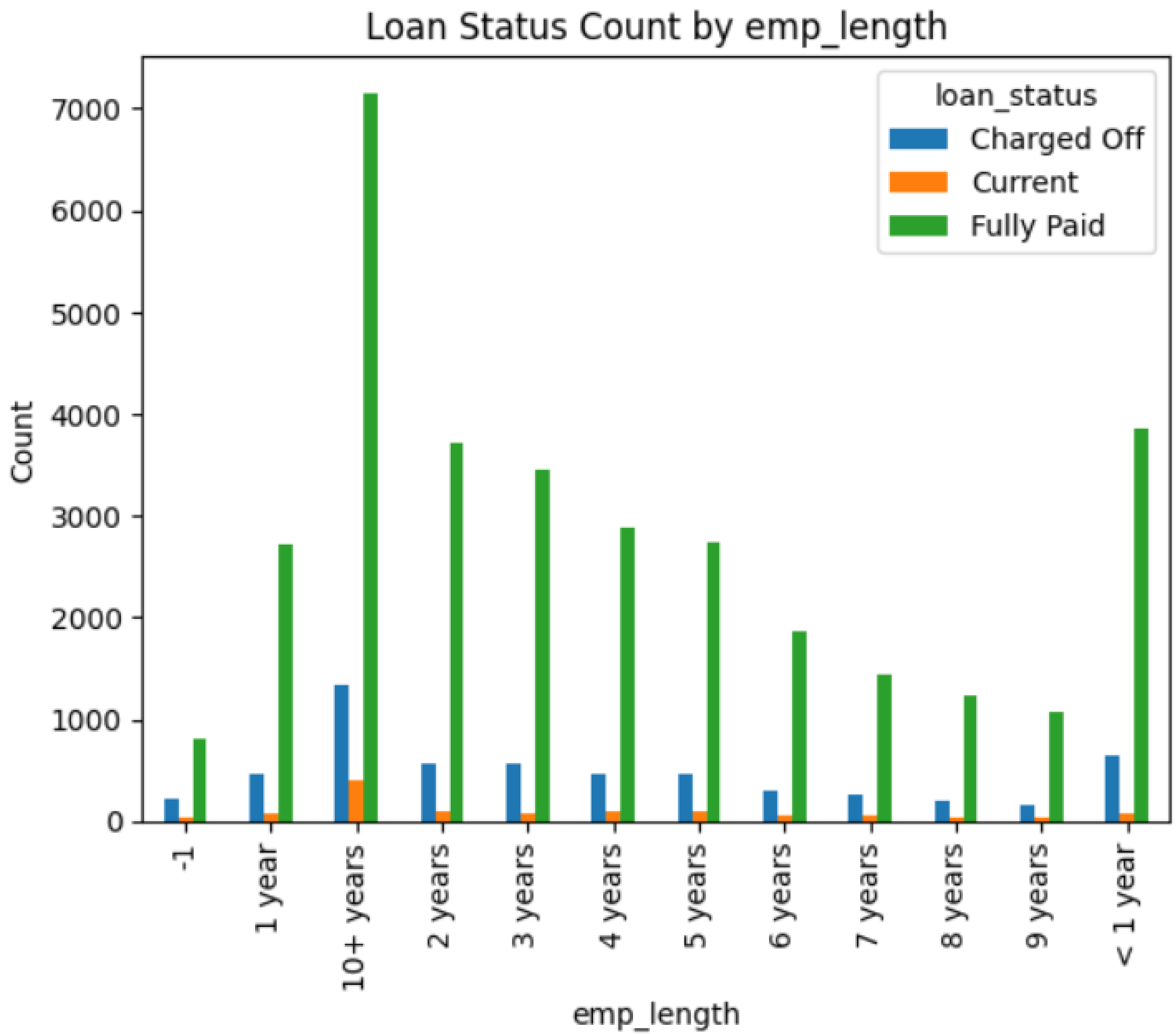
```
# Determining the shape of the datset

loan_dataset.shape
```

(39665, 46)

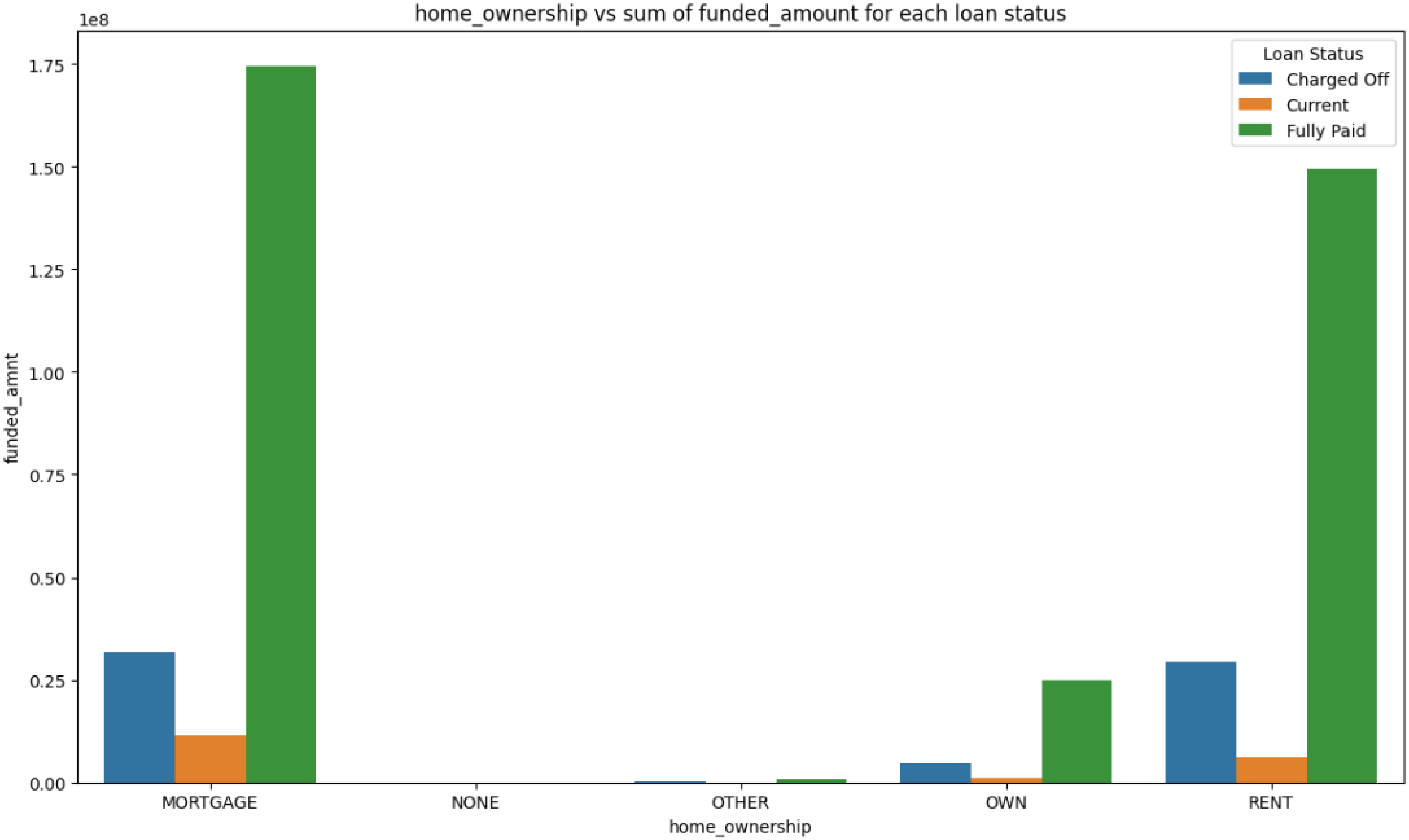
Column_name	Value Filled/Changed
emp_length	Fill NA values with -1
annual_inc	Convert scientific notation to regular numbers
dti	Convert to numerical data
pub_rec_bankruptcies	fill NA values with -1
tax_liens	fill NA values with -1
chargeoff_within_12_mths	fill NA values with -1
collections_12_mths_ex_med	fill NA values with -1
last_pymnt_d	fill missing values with Not Applicable
last_credit_pull_d	only 2 rows with missing value, so dropping those rows
revol_util	less than 0.5% rows with null value, so dropping those rows
funded_amnt	Convert to numerical values
int_rate	Convert to numerical values
installment	Convert to numerical values
revol_util	Convert to numerical values
issue_d	Convert to datetime
open_acc	Convert to numerical value
total_acc	Convert to numerical value

ANALYSIS

Higher number of the defaulters have either worked for too long 10+ years or are fairly newly employed (<5 years)

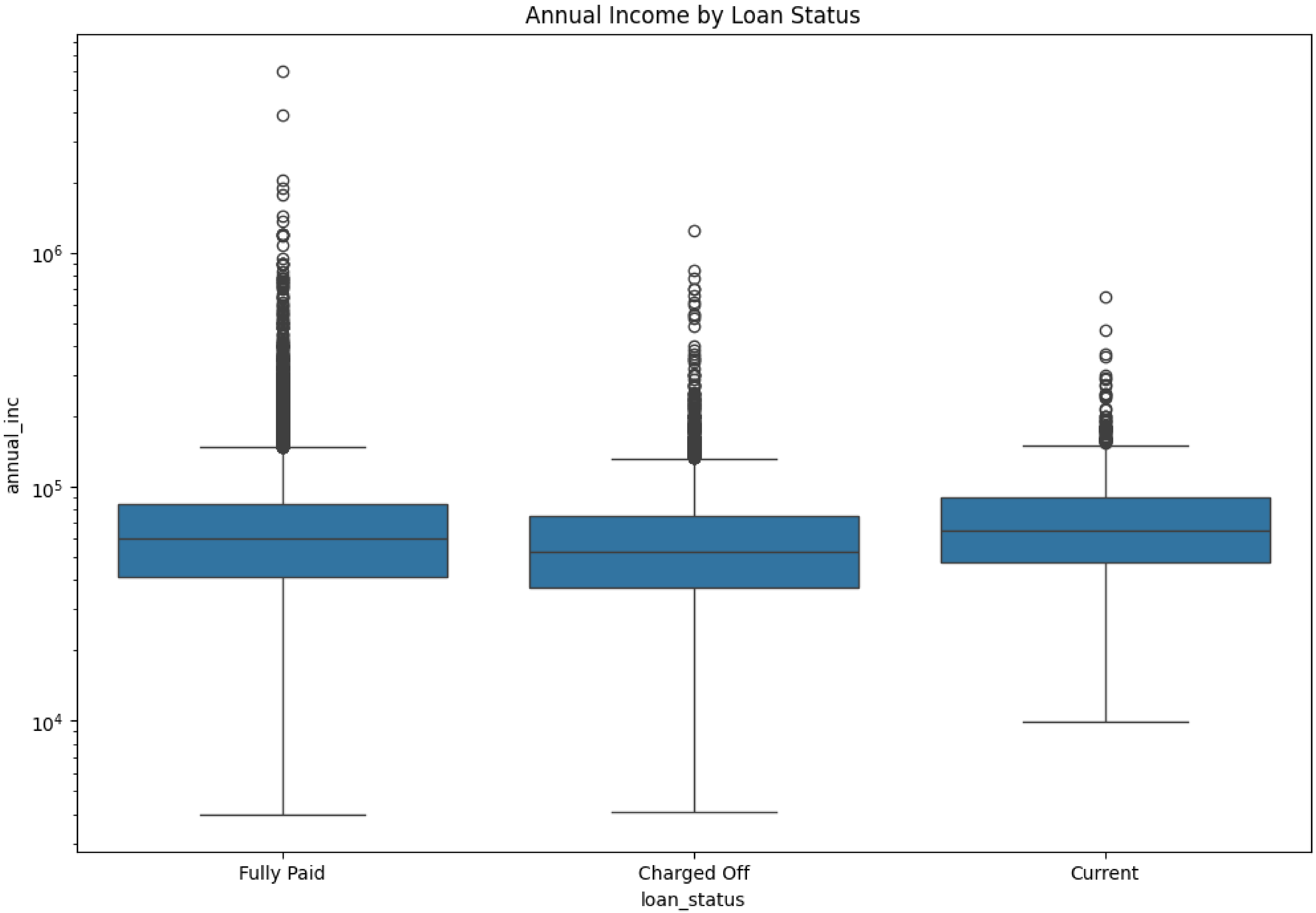


ANALYSIS



Most of the defaulted people are either living in a rented home or servicing a mortgage

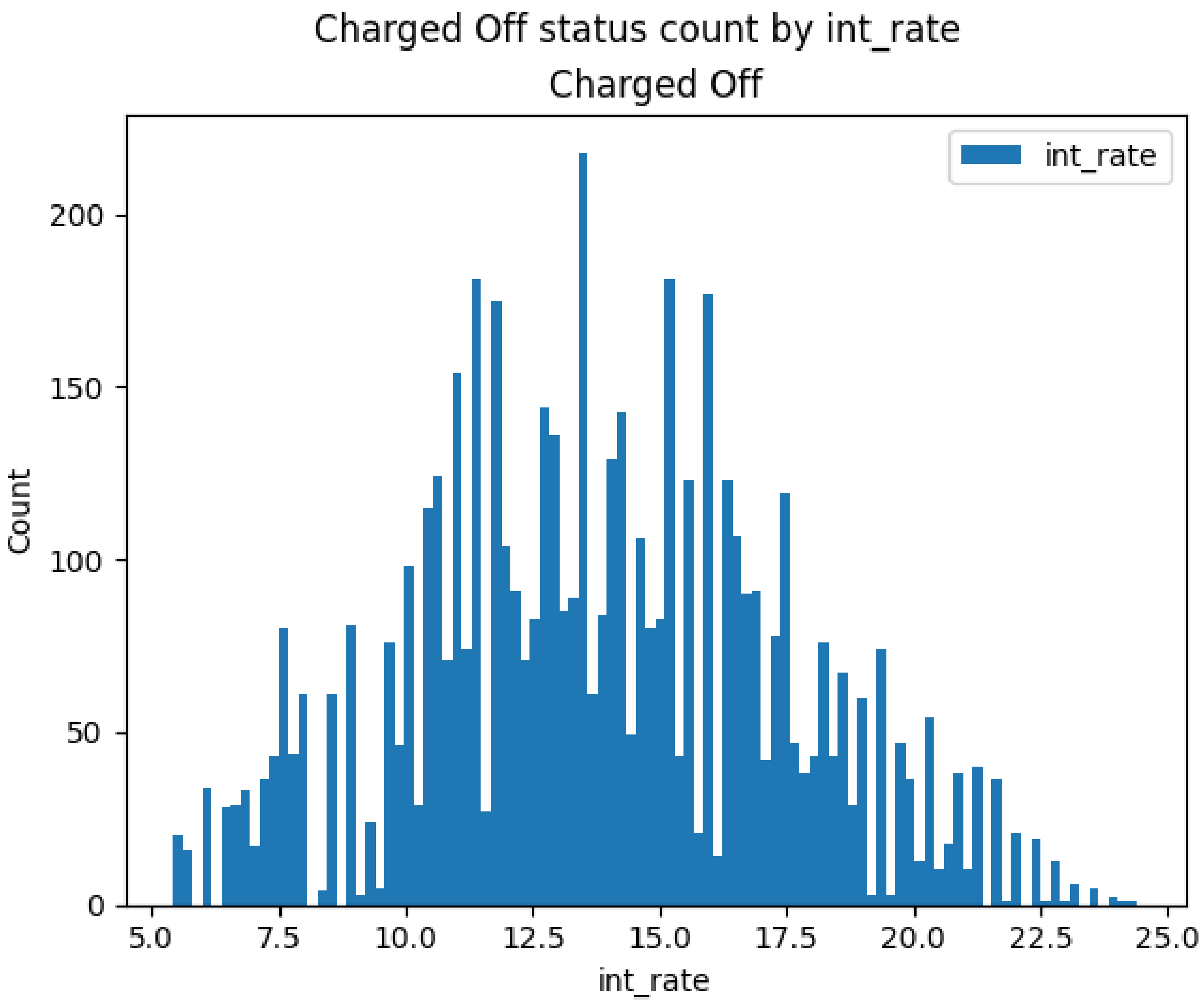
ANALYSIS



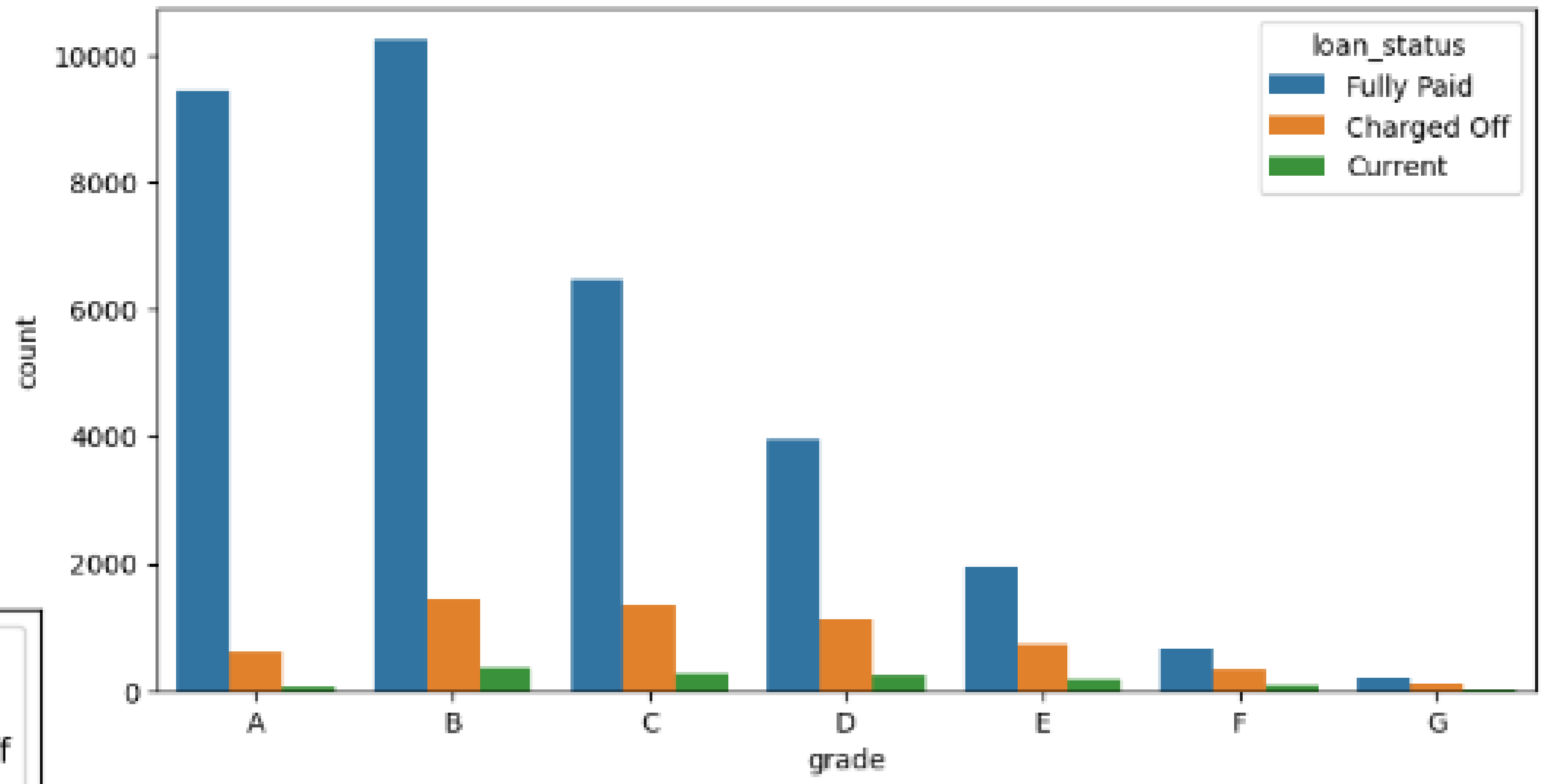
There isn't much significant difference in annual income of people who have fully paid the loan vs the ones who have defualted

ANALYSIS

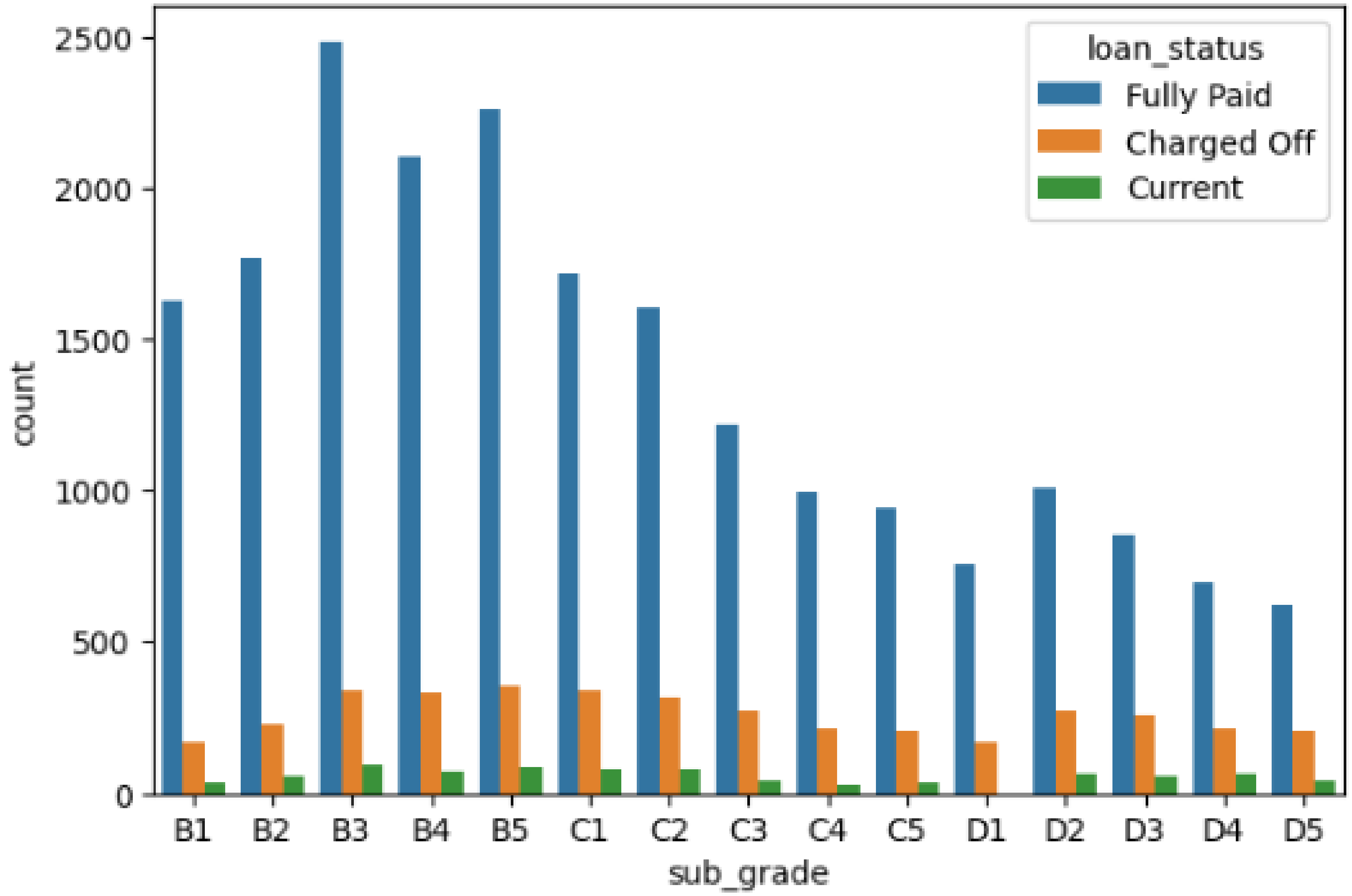
Higher number of the defaulters are within the interest rate 10 – 17%



ANALYSIS

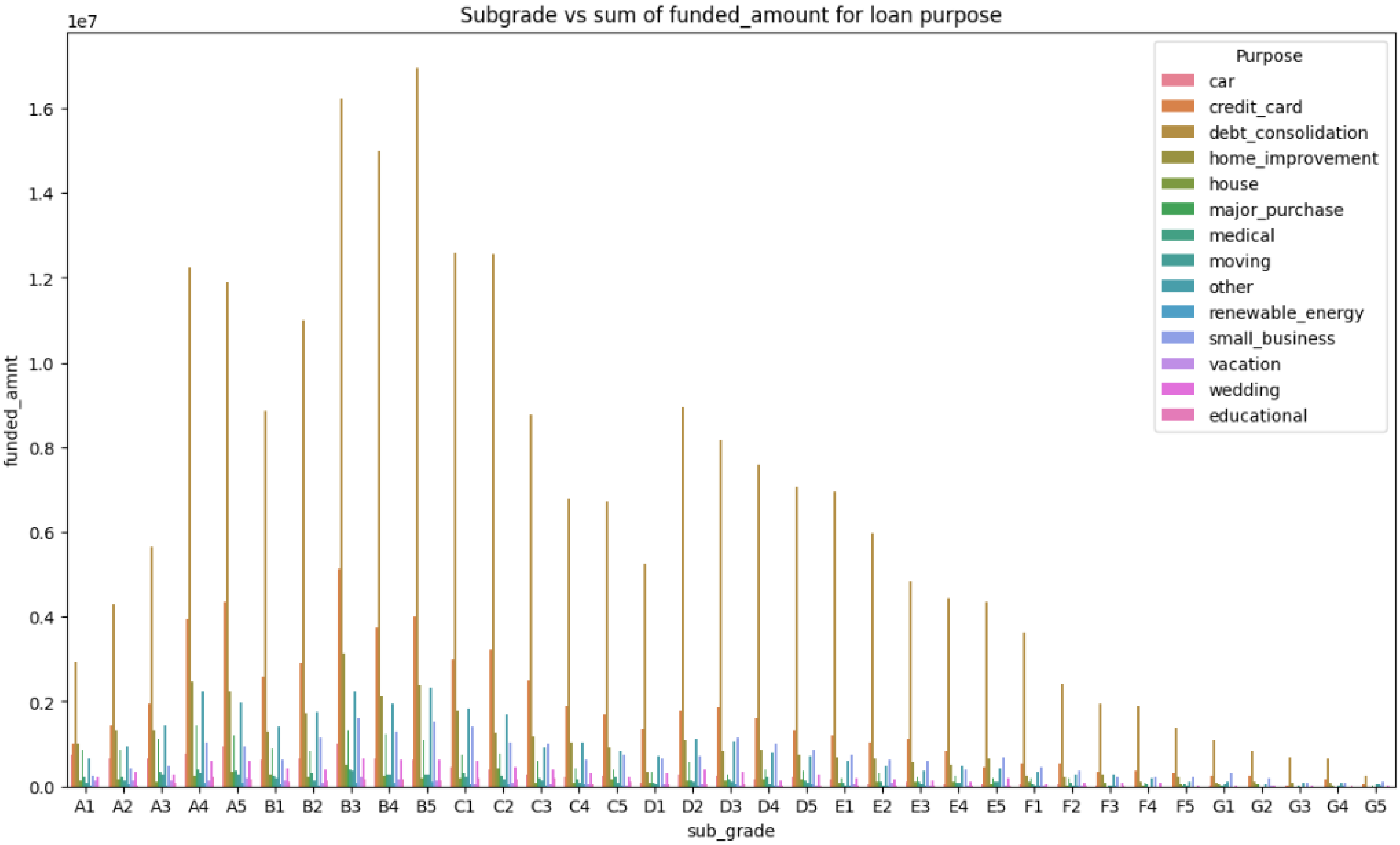


Most of the Charged Off loans belong to B, C and D categories



Isolating Grades B, Cand D and recreating the countplot for subgrades. Highest count is between B3-C2 sub grades and then D2-D3

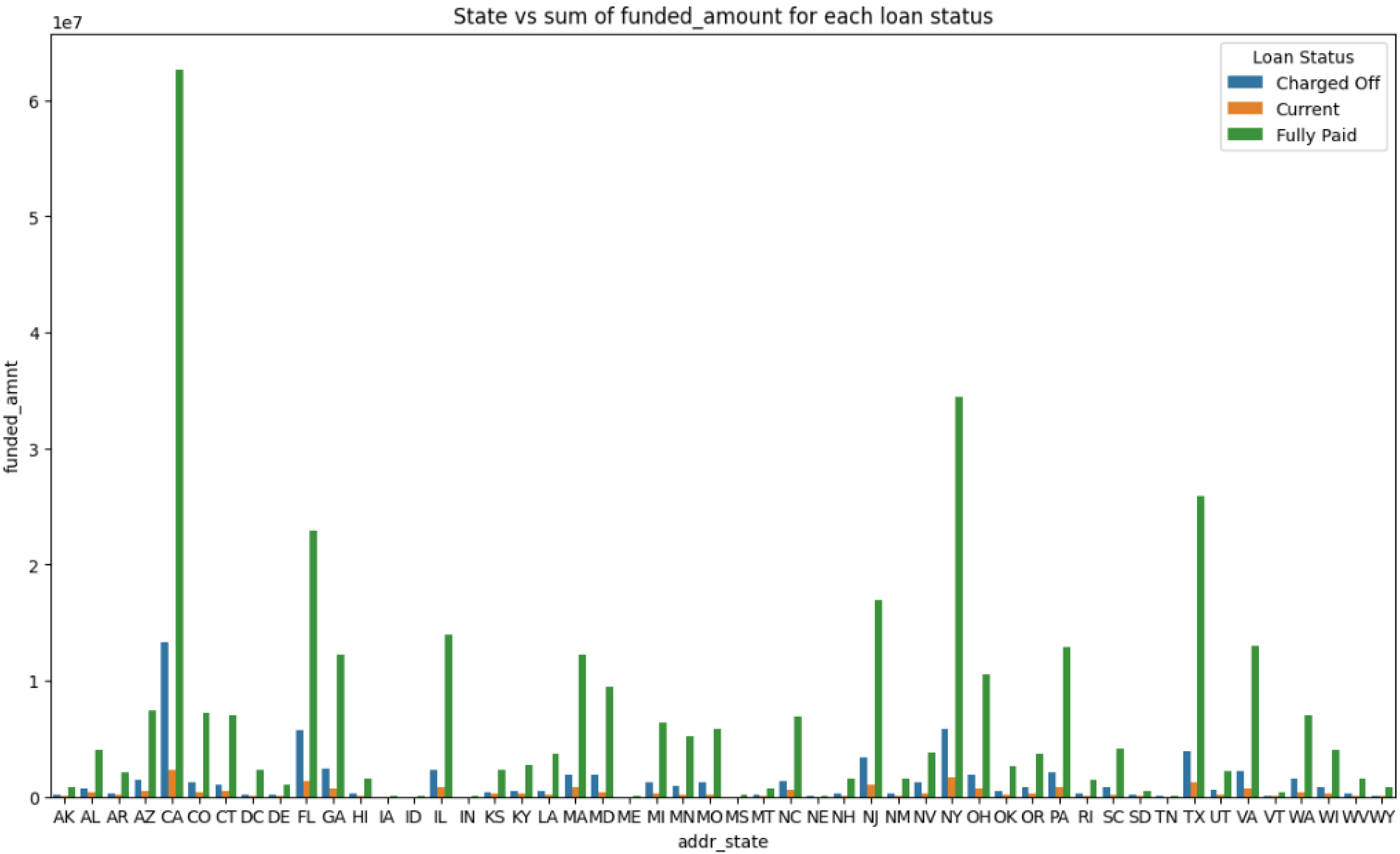
ANALYSIS



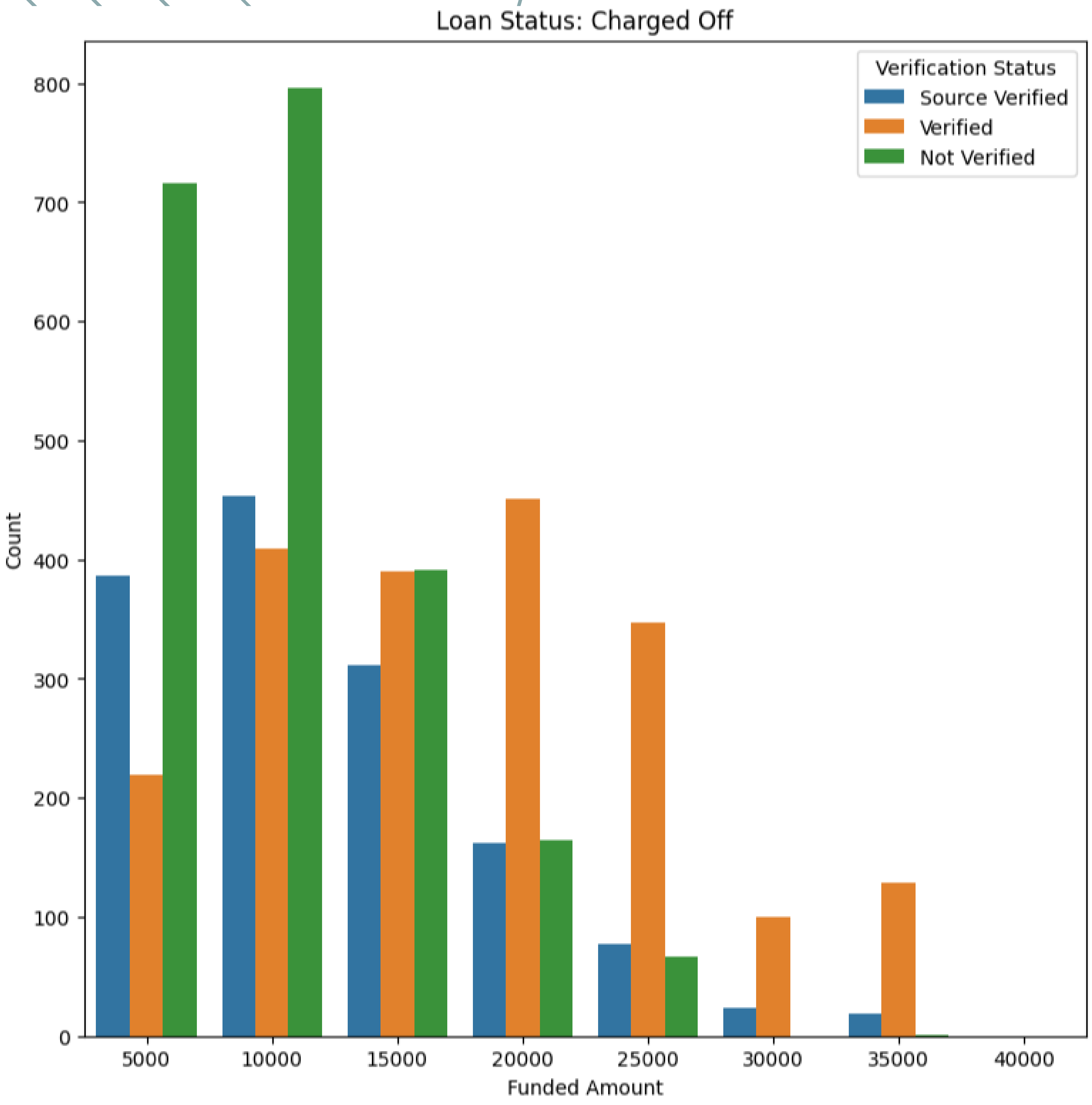
Bulk of the loan is taken for debt_consolidation. This is true for all sub-grades.

ANALYSIS

CA has the most number of Charged Off loan status followed by NY and FL



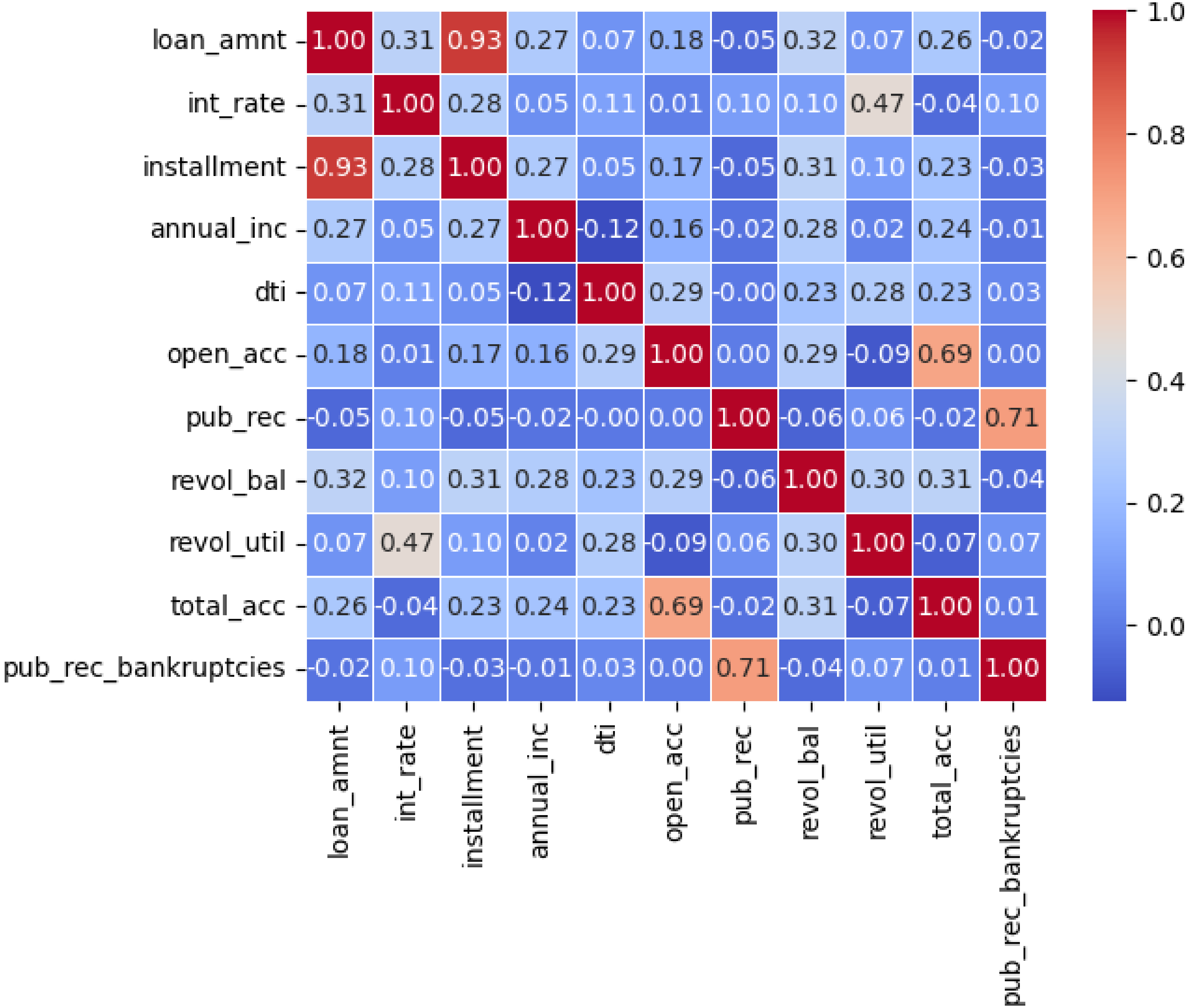
ANALYSIS



For Charged Off status, a very high count of loans of amount up to 10000 are funded without verification. And for funded amount between 10000-15000 too the count of not verified loans is significant.

Lack of Verification is resulting to charge offs???

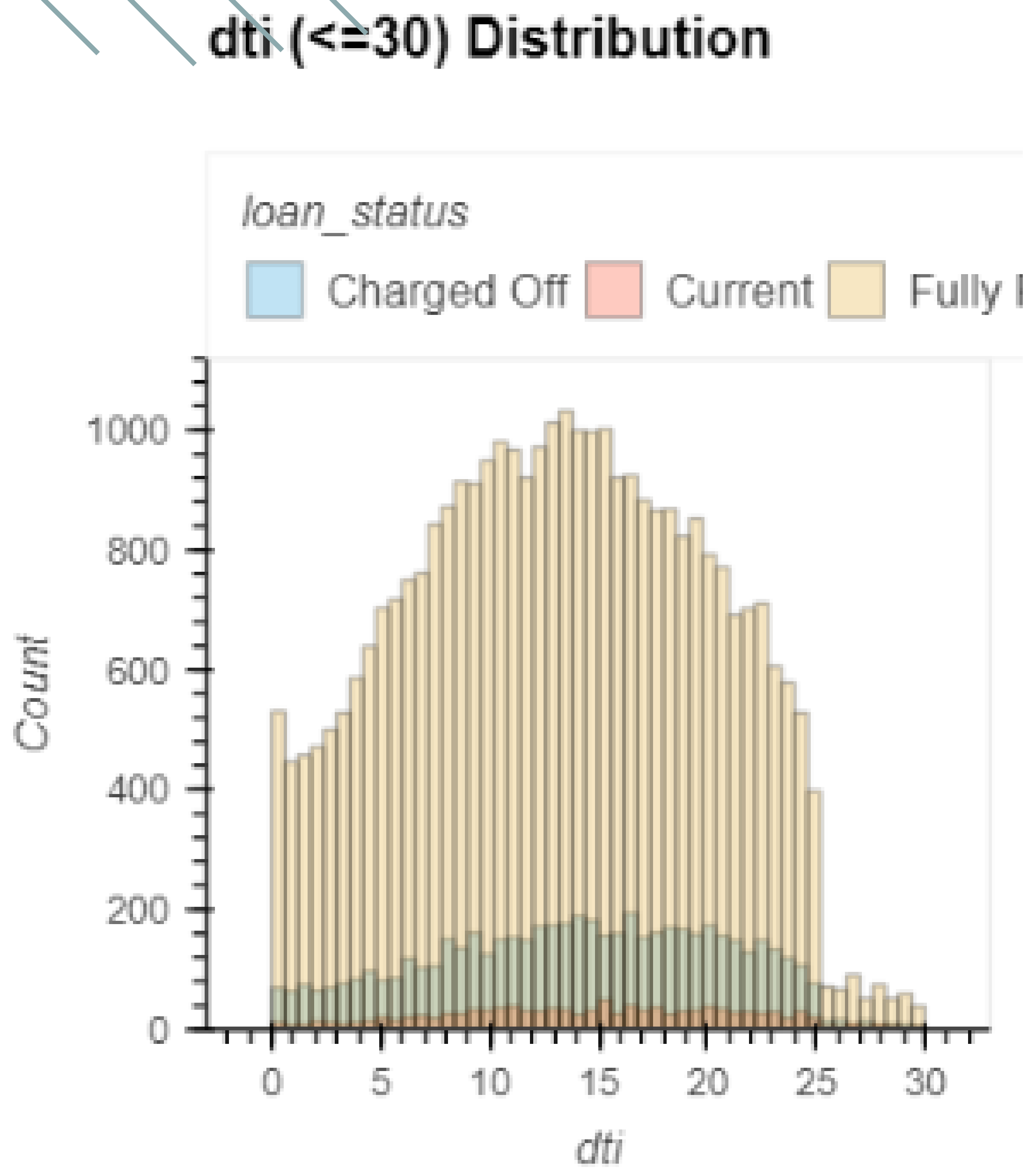
ANALYSIS



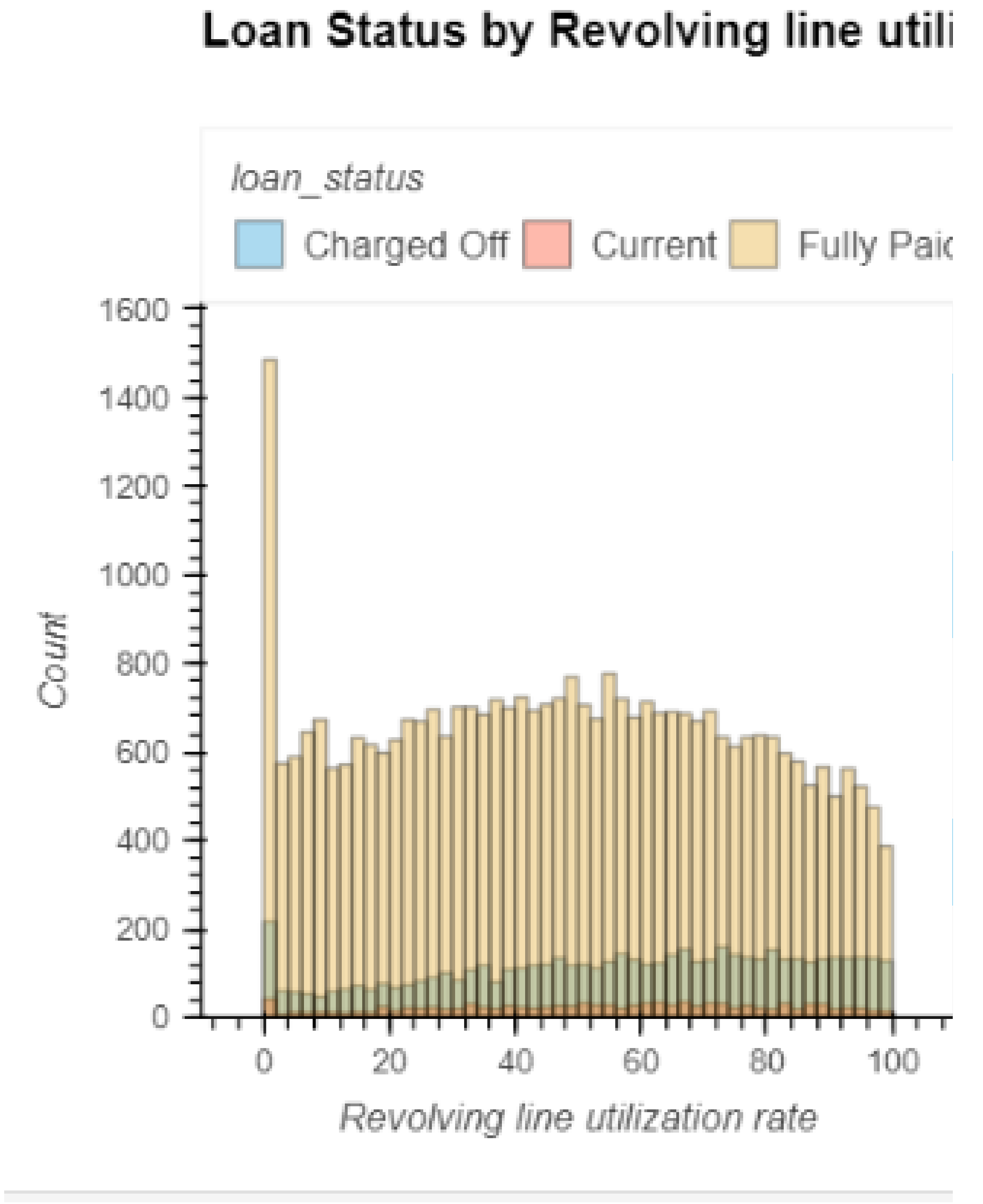
We noticed almost perfect correlation between "loan_amnt" and "installment".

int_rate and revol_util are partially correlated.

ANALYSIS



As both revolving utilization % increases and debt to income ratio increases the tendency to charge of increases.revolving utilization % increases and debt to income ratio increases the tendency to charge of increases.



SUMMARY

- Top charge off's happening between B3-C2 sub grades and then D2-D3.
- Bulk of the loans are taken for debt consolidation and that is true across the sub grades.
- Bulk of the loans are taken by people living on rent or currently servicing a mortgage.
- 10+ years of employment has the maximum loans disbursed, followed by less than one year, two years.
- High counts of amounts less than 10000 are issued with out verification. And these are charging off.
- As both revolving utilization % increases and debt to income ratio increases the tendency to charge of increases.
- The interest rate is also an indicator for defaulters as the charged off count is higher for middle range interest rate.

IMPORTANT DRIVER VARIABLES

- DTI
- Interest Rate
- Revolving utilization ratio
- Verification status
- Sub grade

THANK YOU