Loan defaulter prediction

A data driven approach to spot out the defaulters

Done by
Roger Arnold H

Table of contents

Need for this case study ?

There are several lenders who face loss avoiding to understand risky applicants . So , it will be very much helpful for both the lender and borrower to save their money instead to waste it after years. Our aim in this case study is to apply data driven decisions for predicting a loan defaulter

Univariate Analysis

Analyzing single variable to obtain insights

7.78%

This shows that the major need for money is maximum for the borrowers who rent or mortgage houses.

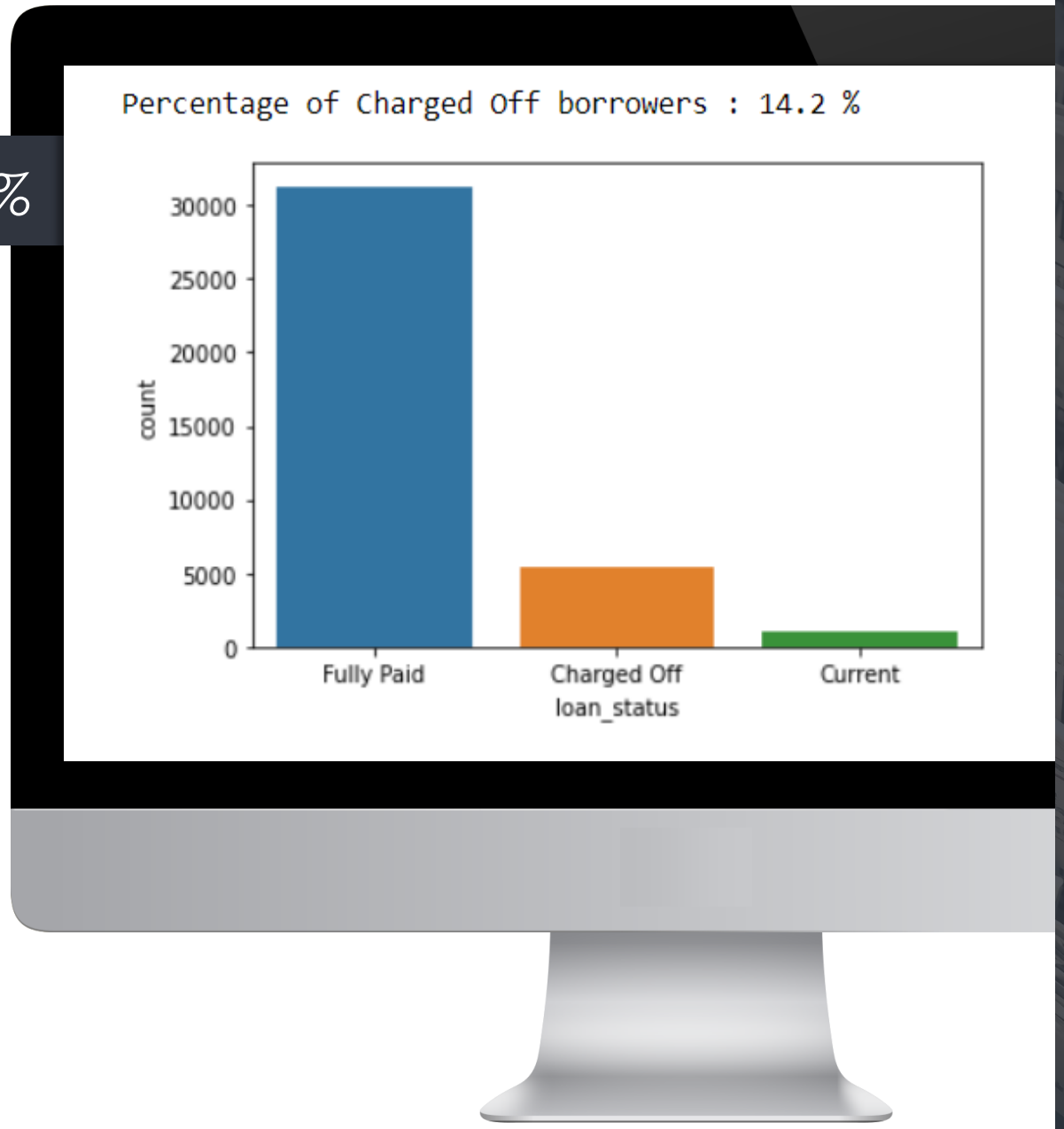
Your logo

44%

Record shows that round 44% of loan lenders are not verifying the borrowers source of income

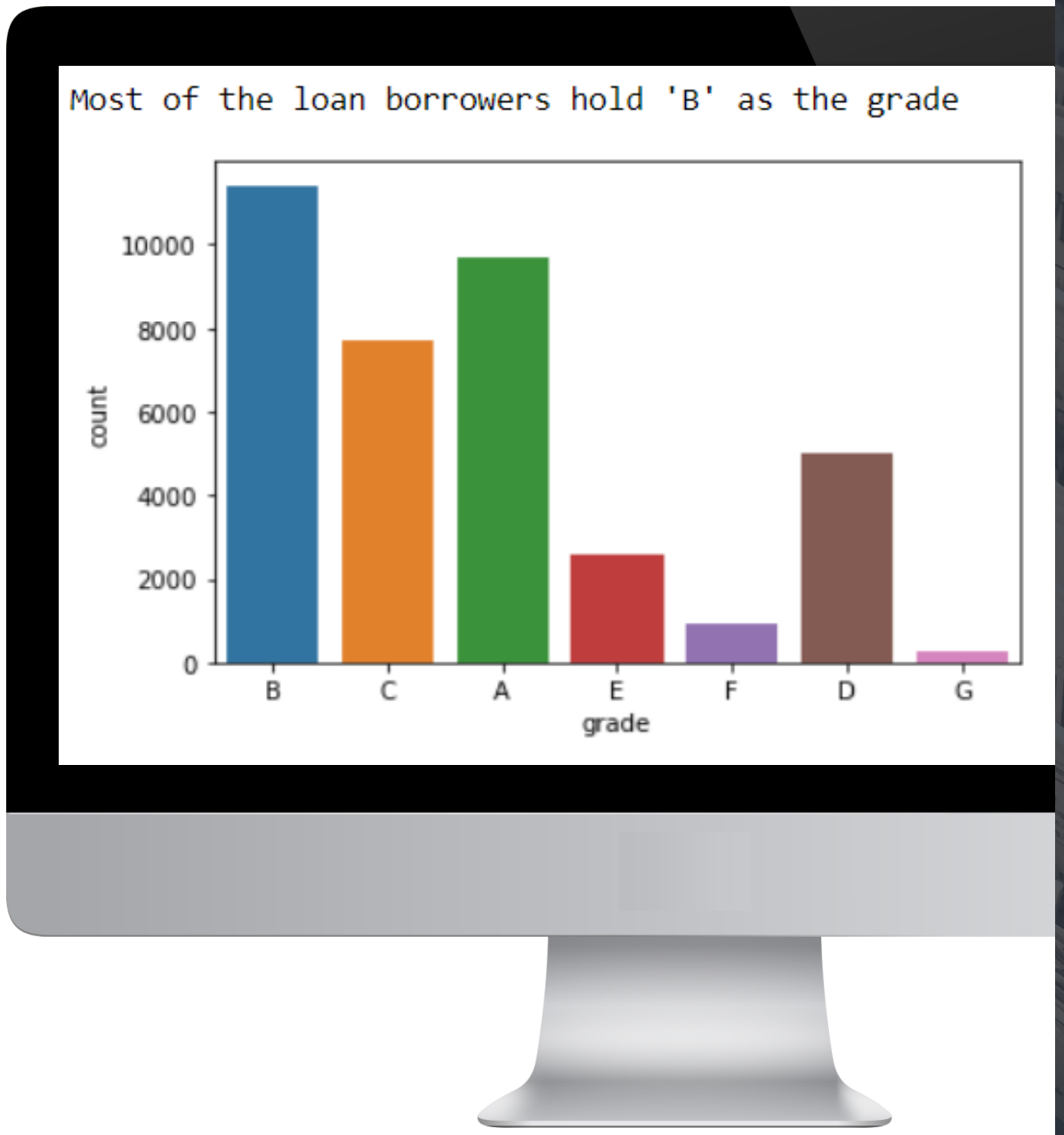
Your logo

14.2%



Charged Off Borrowers make up 14% of our data

Your logo



Majority of the loan borrowers have “B” as the loan grade

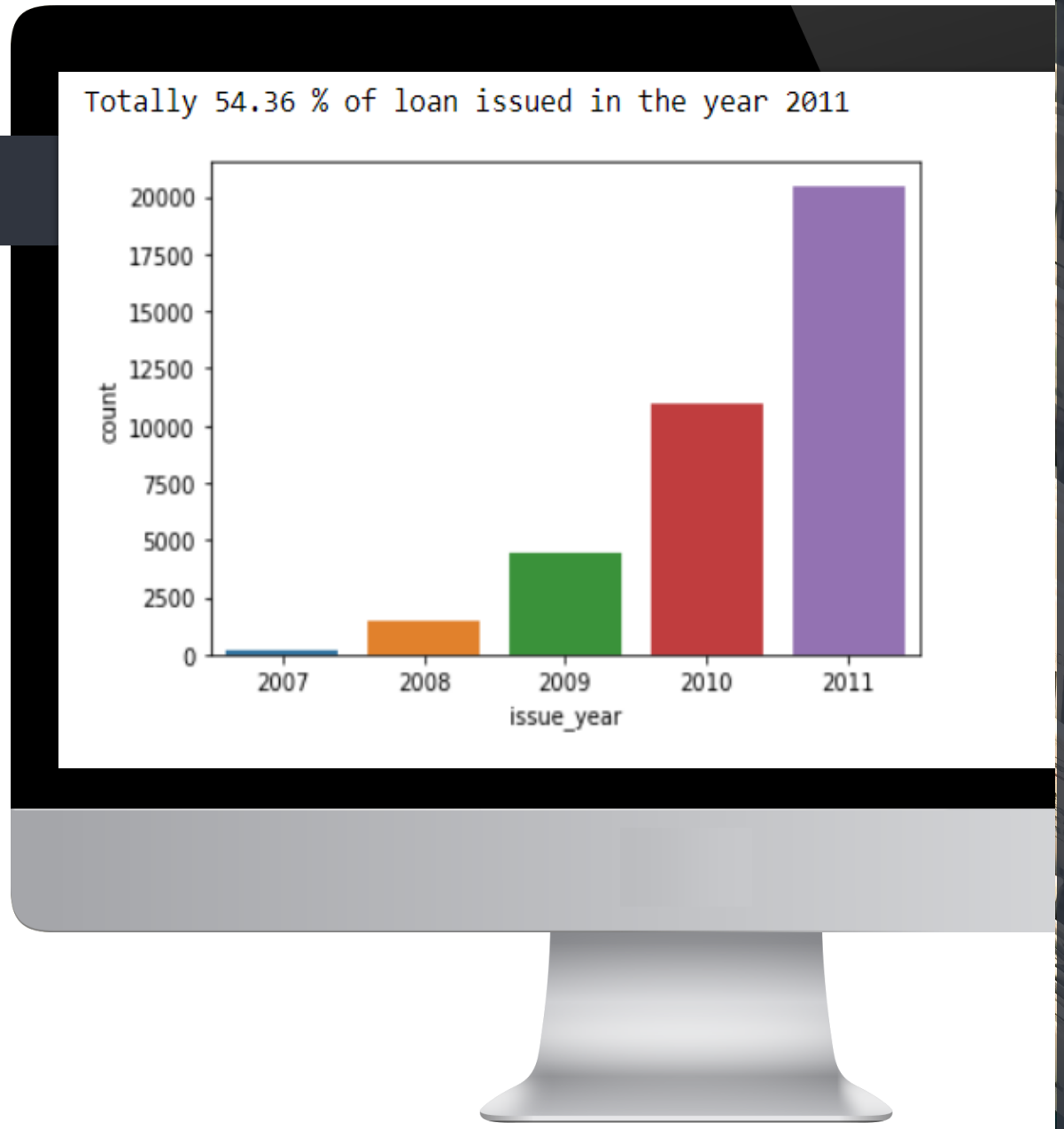
Your logo

25%

Most of the working professionals who has 10 years of professional experience show more interest to borrow loans than the others

Your logo

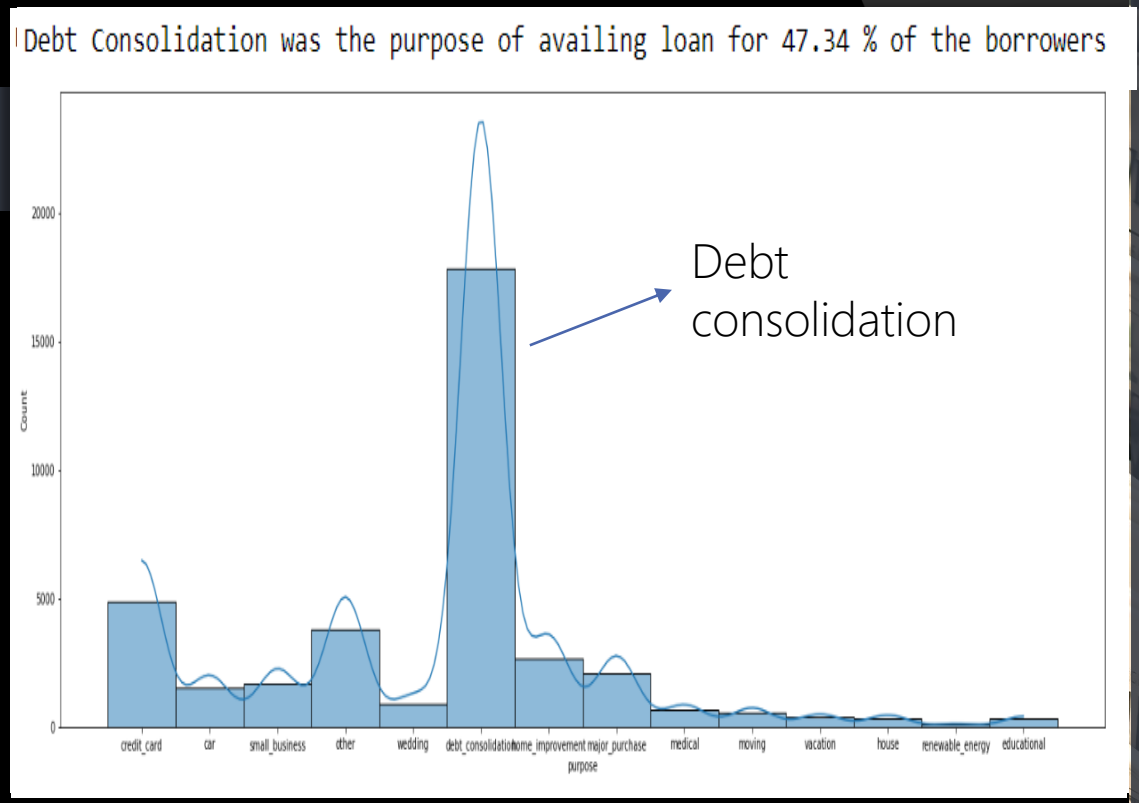
54%



More number of loan was issued in the year 2011. Nearly 54% lending is recorded. US Debt Ceiling Crisis is the main reason for this loan records . It affected markets till Dec 2012.

Your logo

47%

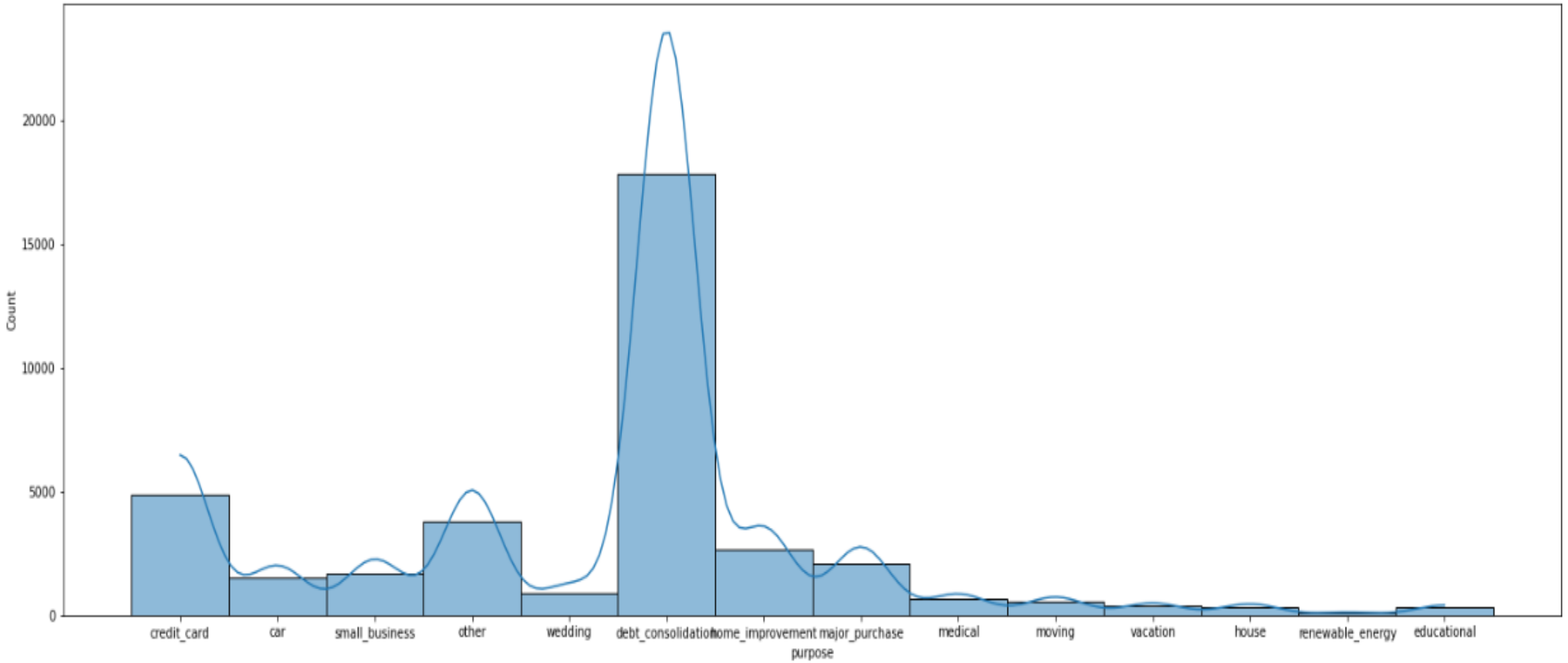


About 47.34% of borrowers avail loan for debt consolidation as their major purpose

Your logo

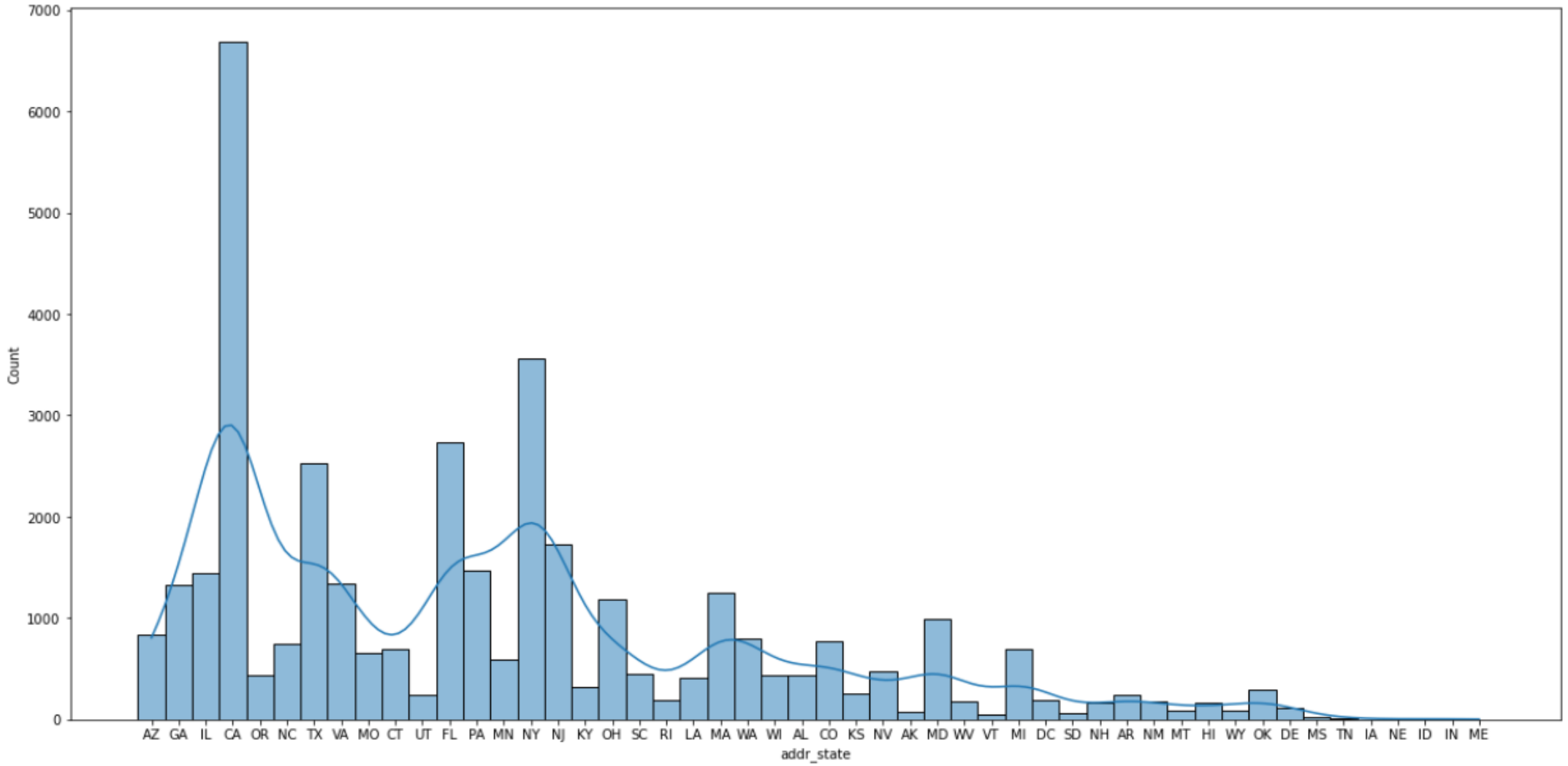
Debt Consolidation as the major purpose

Debt Consolidation was the purpose of availing loan for 47.34 % of the borrowers



California shows good strength of loan borrowers all time

```
Californian borrowers make up 18.0 % of our data
```



Univariate Insights:

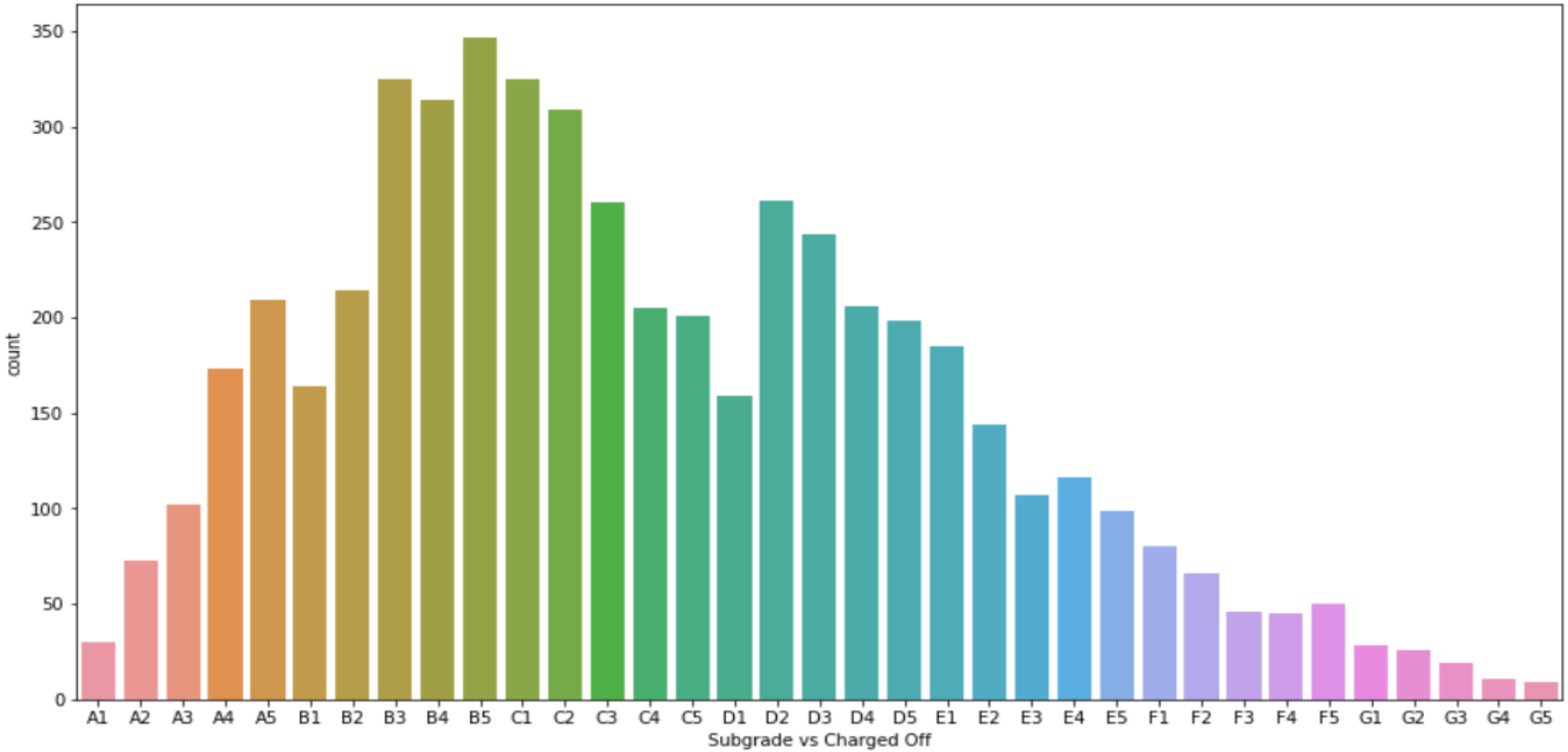
- Majority of the loan borrowers either rent the house or mortgage it. Own house loan borrowers are very few
- 44% of the borrowers income source is not verified
- Debt consolidation is the major purpose for availing loans . 47.34% of borrowers loans for this specific reason
- California shows the highest record of loan borrowers. Around 17.74% of total borrowers are from California. Which in turn gives us 3% of total charged off borrowers from this very same state
- Charged Off borrower's make up 14% of our data.
- 24.65% of the borrowers have 10 years of work experience.
- 54.36% of loan issues happened in the year of 2011

Bivariate Analysis

Analyzing double variables to obtain insights

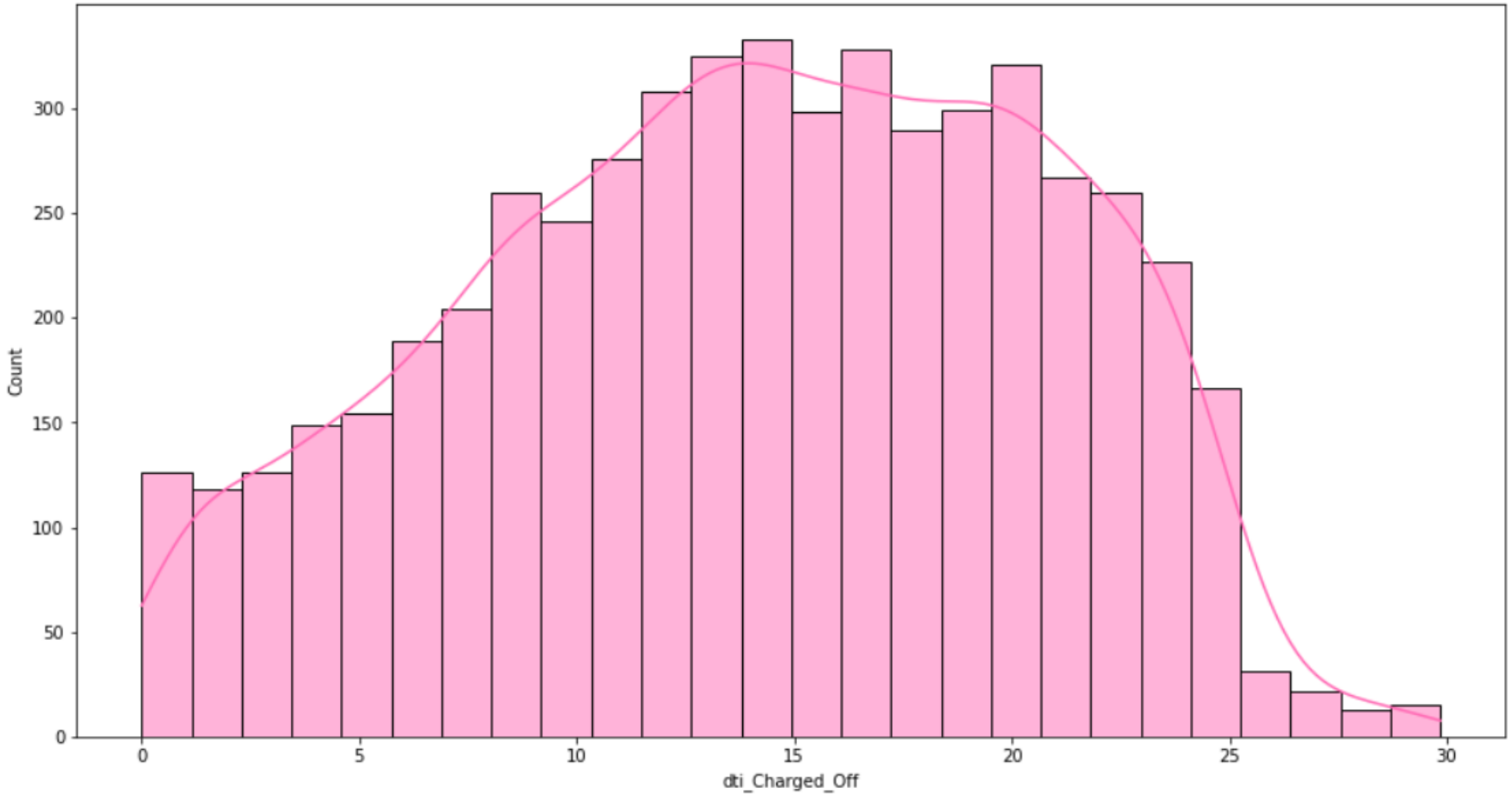
Behavior of sub grades in charged off borrowers

Grades with B3, B4, B5 C1, C2 shows high records for charged Off loans
They make up 30.28 % of the total charged off data



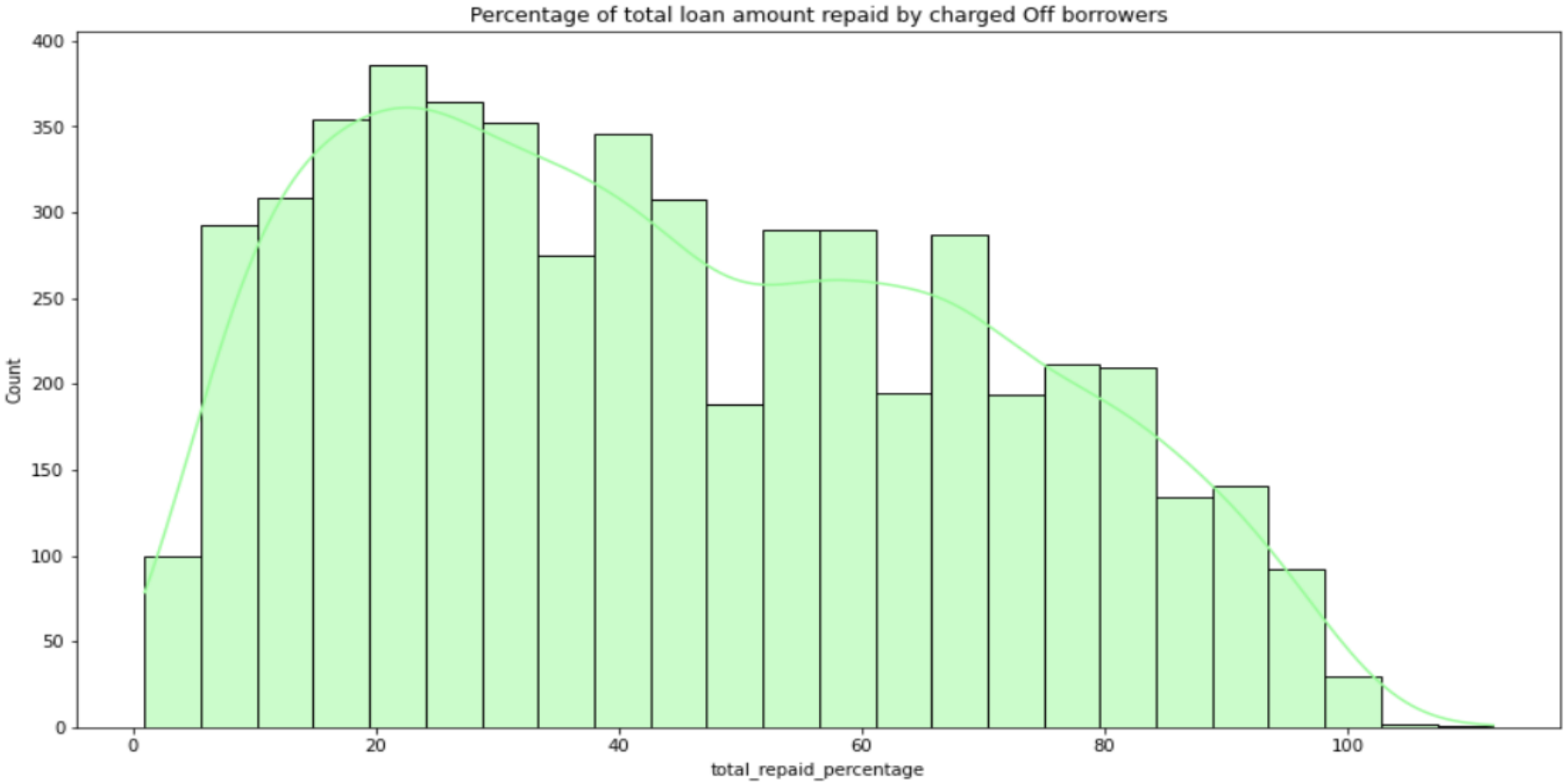
Behavior of Debt to income ratio with charged Off Borrowers

Safe debt to income ratio is desired to be less than 12%
Borrowers who got charged off with DTI ratio more than 12% is , 62.85981308411215%



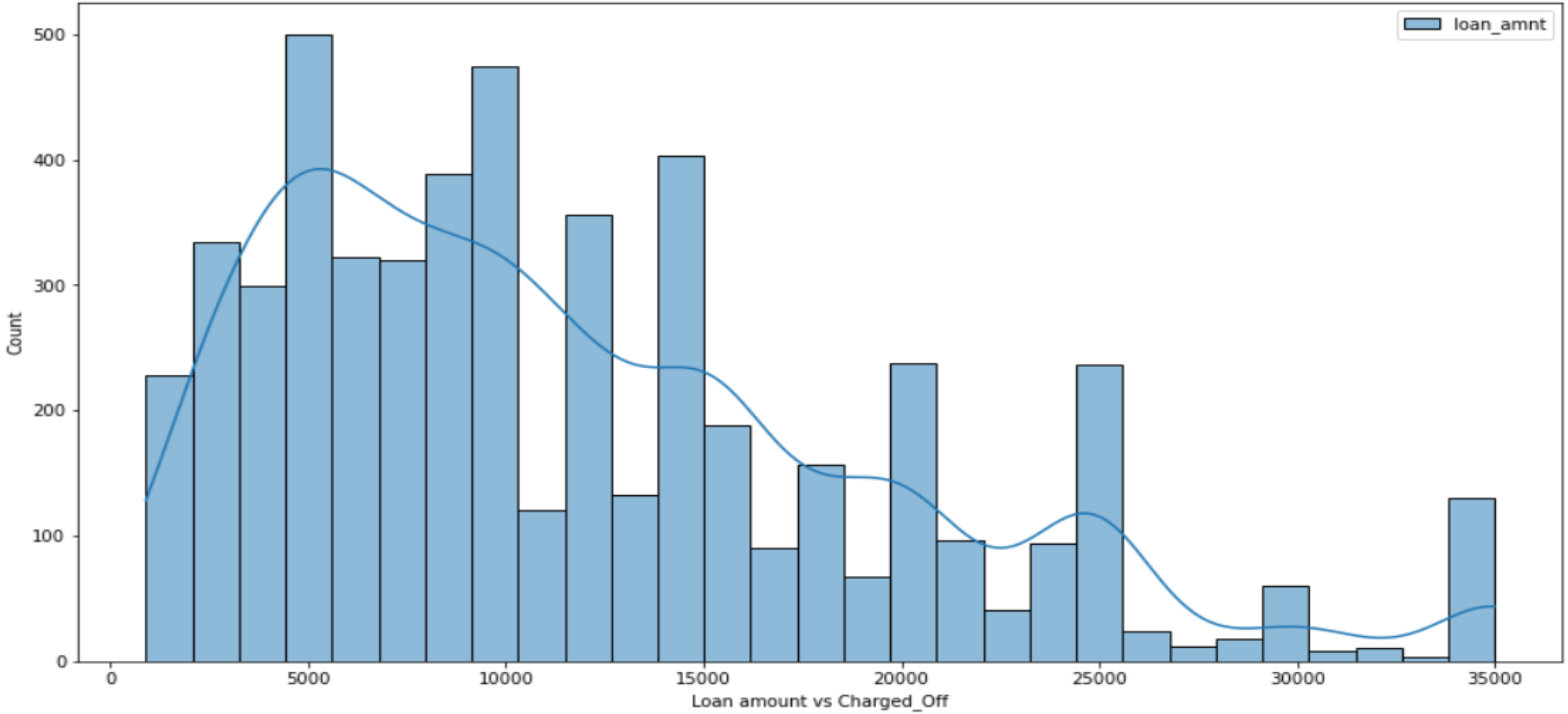
Total amount repaid by charged Off borrowers

```
Average loan amount percentage repaid by the borrowers : 44.33 %
```



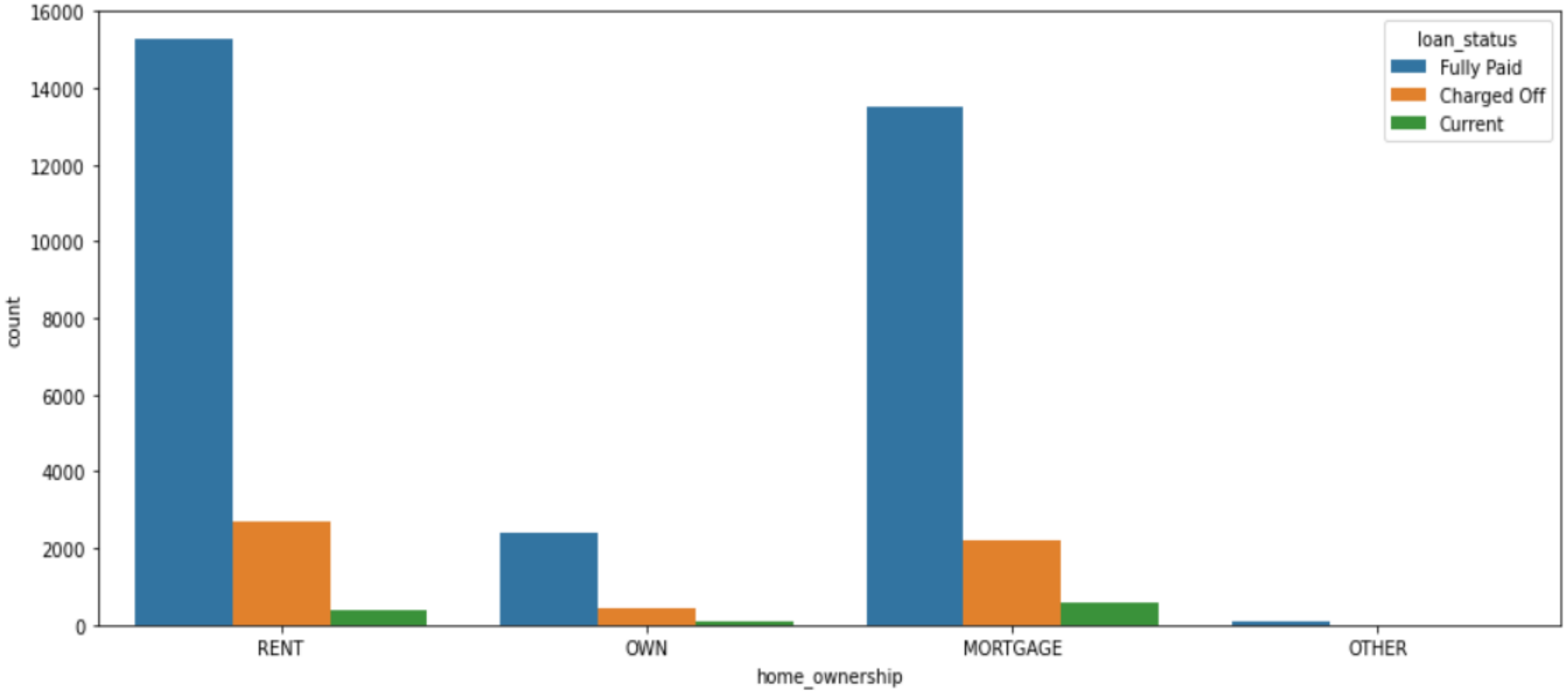
Charged Off loan amounts

Loan amounts of range 5000$ to 15000$ shows more charged Off records
About 67.5 % of borrowers are charged off for this loan range



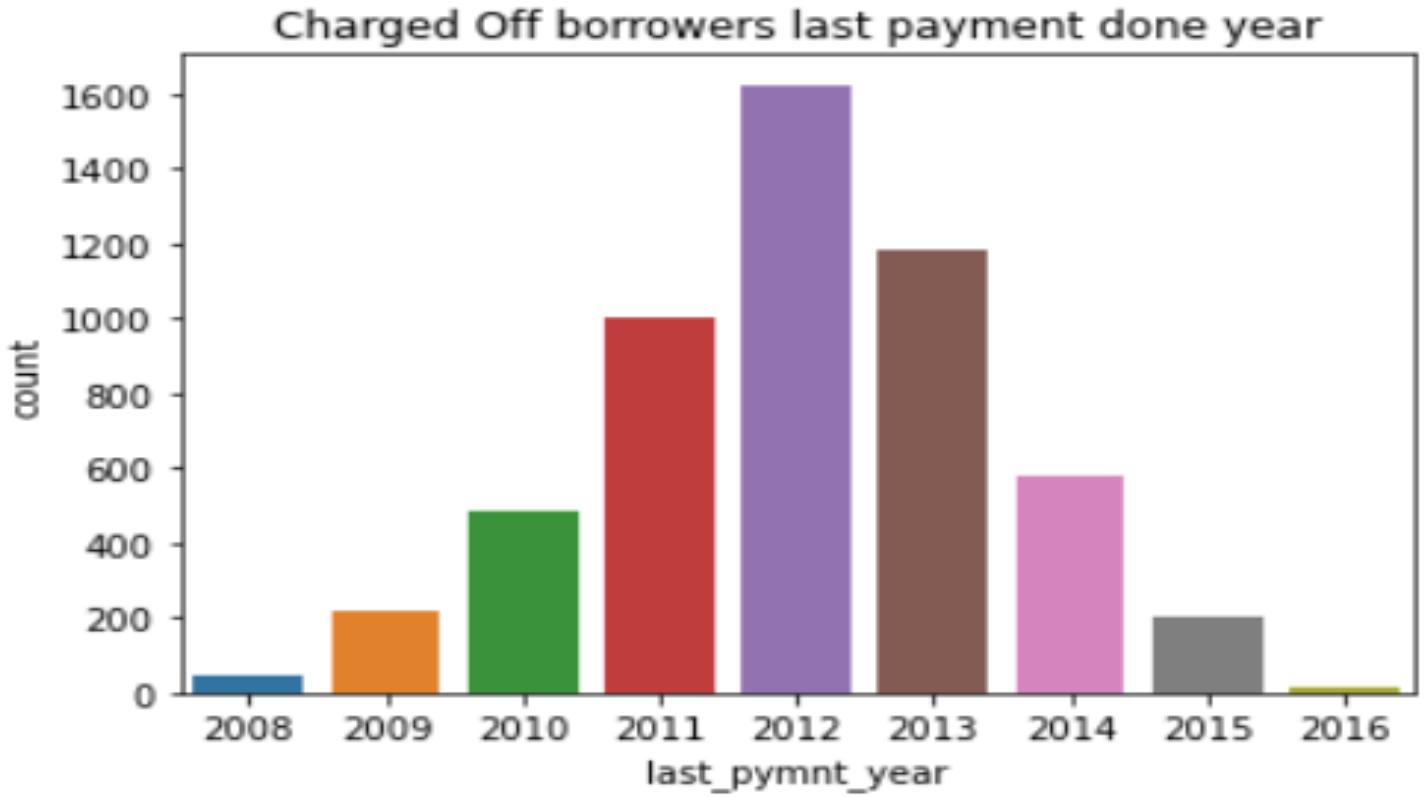
House owners are very less likely to get charged Off

```
Percentage of Charged Off house owners : 8.02 %
```



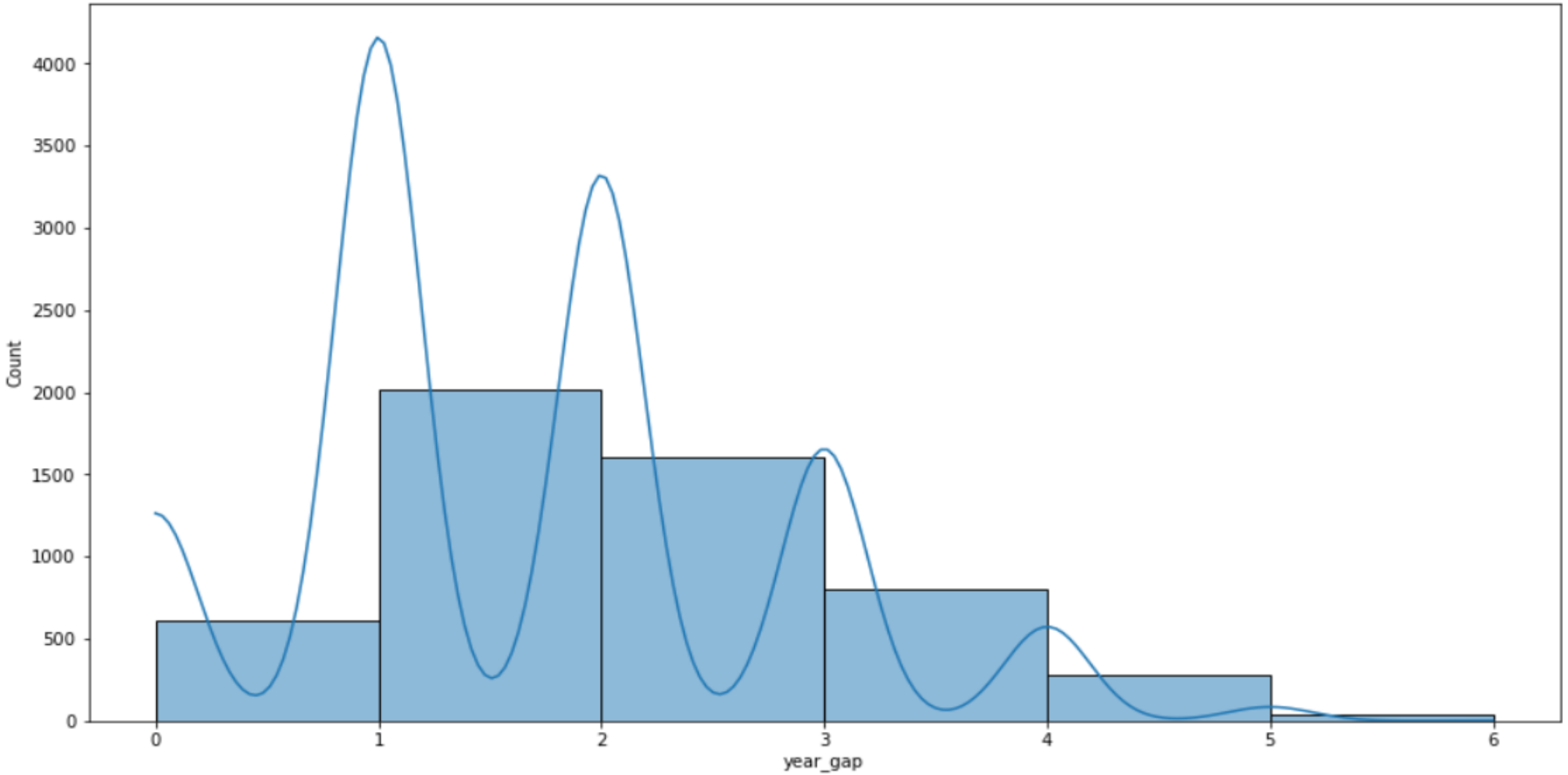
Years which has maximum number of Charged Off records

The years 2011,2012,2013 has more charged Off records
Over 71.21 % of loan payment suspended on these years



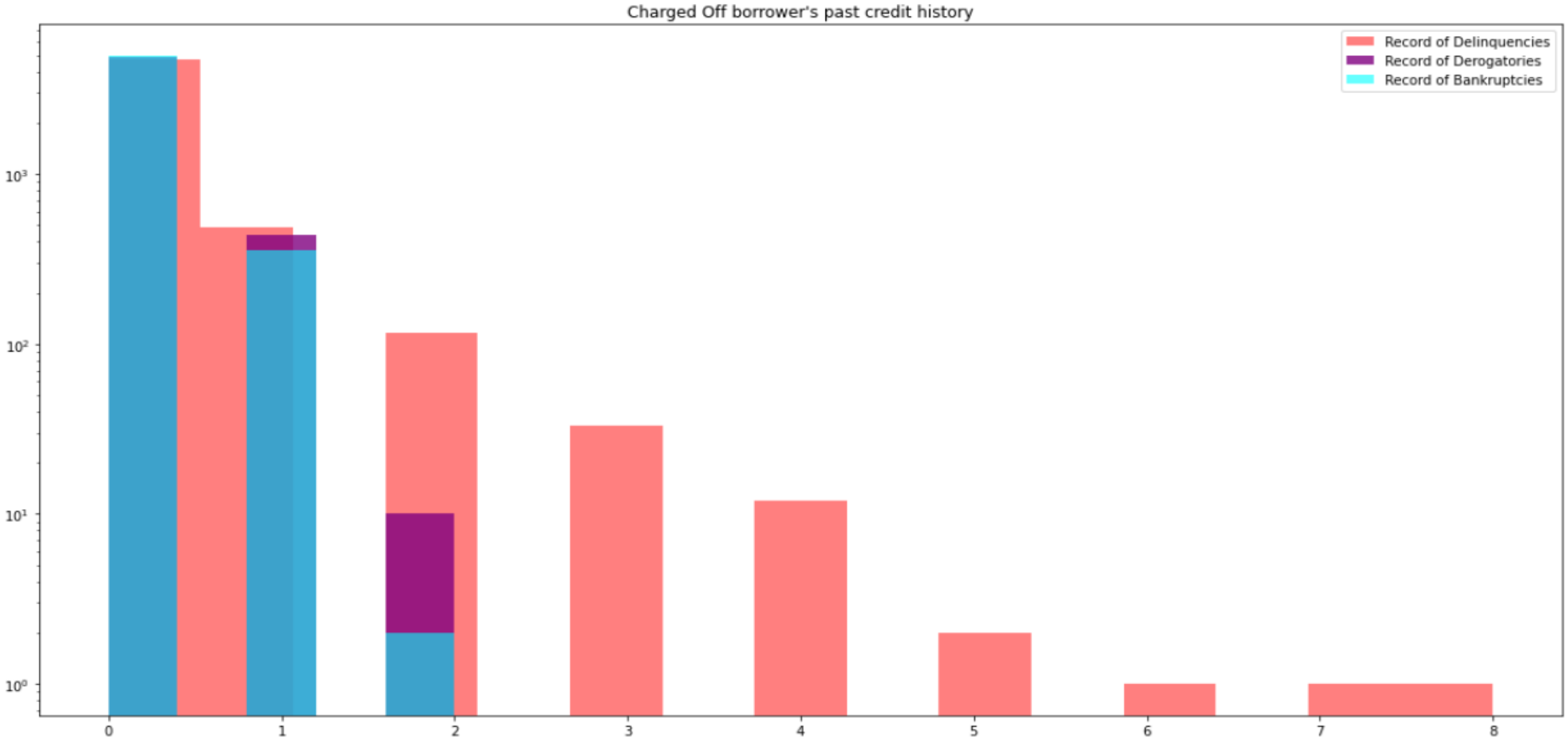
Loan payment consistency

Consistent loan payment is found even for the charged Off borrowers.
Over 67.64 % of the borrowers paid installment maximum 2 years



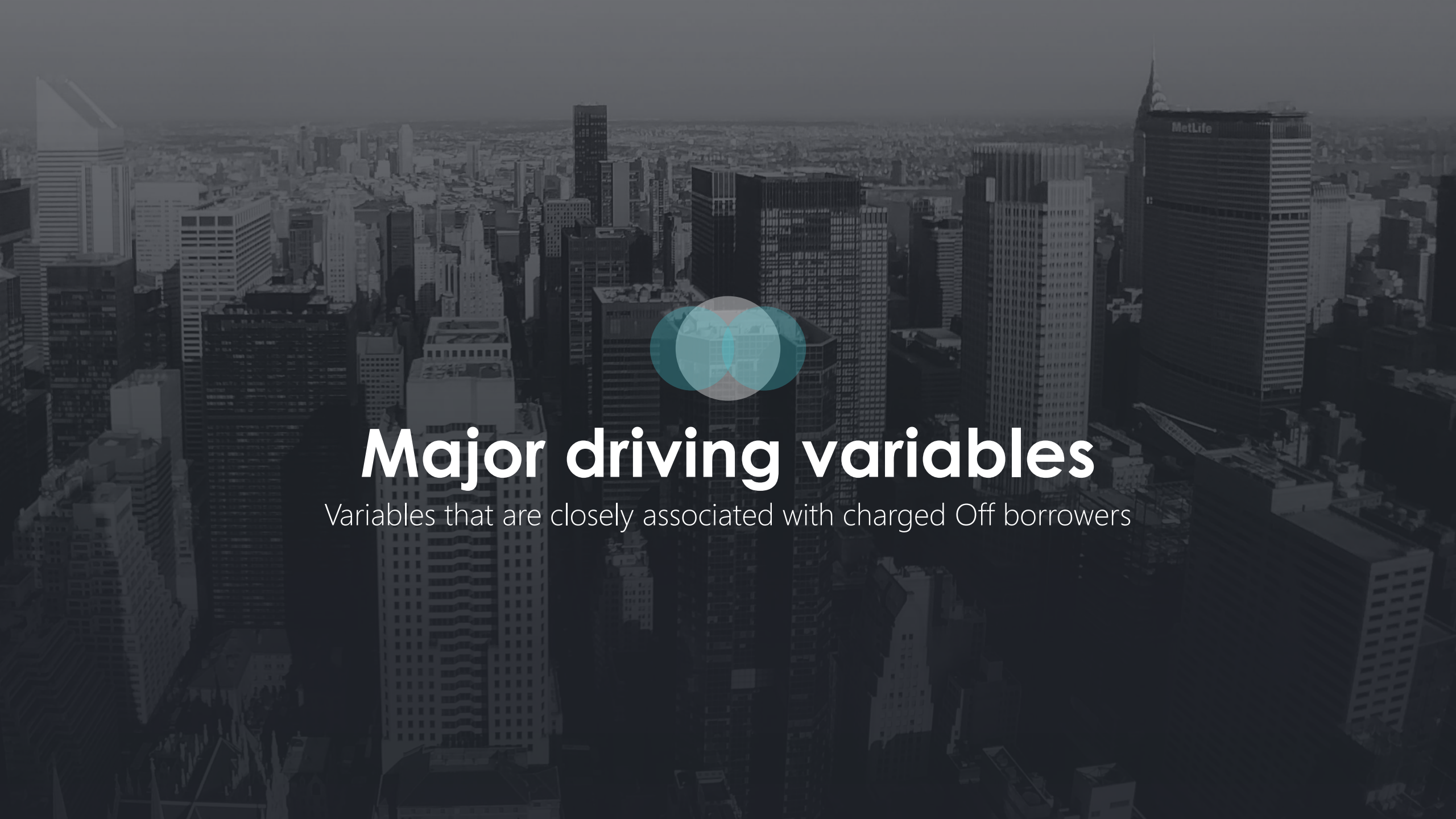
This shows charged Off borrowers past credit history

Most of the charged of borrowers have records for delinquency



Bivariate Insights

- 30.28% of the charged off borrowers belong to these grades : B3,B4,B5,C1,C2
- 62.86% of the charged off borrowers has dti more than 12
- On average, Charged Off borrowers repay 44.32% of the total amount
- Percentage of house owners charged Off = 8.01%
- Percentage of Charged Off borrowers with Not verified income sources = 38.50%
- Average loan amount lent for charged off borrowers is 11,800 . About 67% of charge Off happened for the loan amount in the range of 5000 to 15000
- The years 2011 to 2013 has the most number of charged off records . About 71.2% of charge offs happened in these years
- If we analyze the loan commitment duration, the data shows that 67.64% of borrowers are charged off within first 2 years of tenure
- It's hard to predict the loan defaulter using these parameters , "open_acc" , "revol_util" , "addr_state" . Their distribution varies uniformly throughout the data
- Most of the charged Off borrowers have records of delinquency



Major driving variables

Variables that are closely associated with charged Off borrowers

Major Driving Variables associated closely with loan defaulter

- Debt to income ratio
- Home ownership
- Income source verification status of a borrower
- Total number of years borrower paid installment
- Total percentage of amount repaid
- Sub grade of a borrower
- Records of delinquency

Prediction Constraints

Number of checks lenders need to validate before issuing loan

Checks we need to carry out to predict a loan defaulter.

- House owners are the safest borrowers. Lenders can be confident on them.
- dti ratio should be less than 12
- Bankers should critically monitor all borrowers once they paid 40% of total payment . Chances of defaulters are more after the paid this ratio of amount
- B3, B4, B5, C1, C2 grade borrowers are more likely to default.
- It's not a good idea to lend money without verifying borrower's income source. So bankers need to eliminate this activity completely
- The high risky loan amounts which are more likely to default is mostly in the range of 5000$ to 15000$
- Lenders cannot predict the defaulter within the first year. So it's always good to monitor due date and payment consistency after a year from loan issue
- There should not be any records of delinquency. As this pattern is more likely to be repeated by the borrowers

Prediction accuracy

Obedience of existing defaulter data for our constraints

Accuracy of our constraint

- The existing data has totally 5, 350 charged off borrowers
- Using our constraint we predicted 5, 054 of the borrowers out of these which in turn gives us *96% prediction accuracy*.

Let's take a look on the analysis we made (reference jupyter notebook cell #72)

iv. Testing our constraints with past charged Off borrowers

In [72]:

```
# Using our constraints we can compare how much data can be identified as charged off

print("Total number of charged off Borrowers in our data : "+str(loan[loan["loan_status"]=="Charged Off"].shape[0]))

print("We have predicted "+str((loan[loan["loan_status"]=="Charged Off"].shape[0])-
                                (loan[ (loan["dti"]<=12) & (loan["total_repaid_percentage"]>40) &
                                       (loan["verification_status"]=="Source Verified") &
                                       (loan["loan_status"]=="Charged Off") & (loan["year_gap"]>1) &
                                       (loan["delinq_2yrs"]==0)].shape[0]))+" charged off borrowers using our constraint")

prediction_perentage = round(((loan[loan["loan_status"]=="Charged Off"].shape[0])-
                              (loan[ (loan["dti"]<=12) & (loan["total_repaid_percentage"]>40) &
                                     (loan["verification_status"]=="Source Verified") &
                                     (loan["loan_status"]=="Charged Off")& (loan["year_gap"]>1) &
                                     (loan["delinq_2yrs"]==0)].shape[0]))/(loan[loan["loan_status"]=="Charged Off"].shape[0]))

print("\nPrediction accuracy :"+str(prediction_perentage)+" %")
```

```
Total number of charged off Borrowers in our data : 5350
We have predicted 5124 charged off borrowers using our constraint

Prediction accuracy :96.0 %
```

INFERENCE : As we have seen in the final analysis, 96% of the charged Off data lies within our constraint

We can have this 96% as our prediction accuracy

Conclusion

The variables dti, total_repaid_percentage , verification_status , home_ownership , last_pymnt_year are three major parameters that can be used to predict the loan defaulter at the maximum of 96% accuracy.

Of course only these variables can't be the major parameter of loan defaulters every time . As we see our data, the loan issue rate went peak in 2011, the reason for this is " *2011 US Debt Ceiling Crisis* " . It affected many of the people's financial status which lead them to borrow.

The major influencer variable might be a different one other year, but these five driver variables which we have selected is going to be there all time. Thus we conclude that, our data-driven analysis will definitely help to spot the loan-defaulters.


THANK YOU