Gramener Case Study using EDA

SUBMISSION

Group Name: ***Fantastic 4***

1. Abinash Sahoo
2. Lekha Priya
3. Sushree Sweekruti Das
4. Pradeep Kumar Pingili

<Abstract>

Context:

A **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:

- **If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company**
- **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:

- **Identification of *Loan Applicant Patterns* which indicate if a person is likely to *Loan Default* *(Charge-Off)*.**
- **Understand the 'Consumer Attributes', 'Loan Attributes', 'Driving Factors' behind *Loan Default* criteria.**
- **Company may choose to utilize this knowledge for its portfolio and risk assessment of new loan applicants.**

<Problem solving methodology>

UpGrad

Business Understanding ↔ Understanding the business context of Banking and Financial Services – **LOAN APPROVAL/REJECTION** process along with **RISK ANALYTICS.**

Data Understanding ↔ Perform clear understanding of all the variables and identify which are essential attributes of **CUSTOMER & LOAN** in this analysis using given Data Dictionary.

Data Preparation ↔ Perform required **Data Cleaning, Treating NAs, Single Values, Addressing Outliers** of key relevant columns for analysis

Data Analysis ↔ Using EDA techniques **(Univariate, Bivariate & Segmentation),** Identify various data patterns along with its relationship and visually represent them in PLOTs using quantitative and categorical variables.

Recommendations ↔ Identify and Present all the required Insights captured from our analysis. Provide recommendations and factors that are strong indicators of **LOAN CHARGED-OFFs.**

Cycle process between understanding, preparation & analysis

Data Observations:

We've observed the following –

- Total **39717** records in the given loan dataset.
- Total **111** variables in the given loan dataset.
- Majority of variables contains a single value or more number of NAs

Customer & Loan Attributes:

Customer

- Employment Length
- Annual Income
- City, State, Zip code
- Description
- Loan Purpose
- Home Ownership
- Application Type
- Delinquency Type 2

Loan

- Loan Amount
- Loan Status
- Funded Amount
- Interest Rate
- Loan Grade
- Verification Status
- Term

Column Variable	Category	Unordered Categorical Variable(UOCV)/Ordered Categorical Variable(OCV)/Quantitative Variable(QV)
dti	Input Factors	QV
earliest_cr_line	Input Factors	OCV
inq_last_6mnths	Input Factors	QV
mnths_since_last_record	Input Factors	QV
open_acc	Input Factors	QV
revol_bal	Input Factors	QV
revol_util	Input Factors	QV
total_acc	Input Factors	QV
acc_now_delinq	Input Factors	QV
chargeoff_within_12_mths	Input Factors	QV
delinq_amnt	Input Factors	QV
pub_rec_bankruptcies	Input Factors	QV
Grade	Customer Demographics	OCV
Sub-Grade	Customer Demographics	OCV
home ownership	Customer Demographics	UOCV
annual_inc	Customer Demographics	QV
zip_code	Customer Demographics	UOCV
addr_state	Customer Demographics	UOCV
ID	Customer Information	UOCV
member_id	Customer Information	UOCV
verification_status	Customer Information	UOCV
issue_d	Customer Information	OCV
loan_status	Customer Information	UOCV
emp_title	Customer Information	UOCV
revol_bal	Input Factors	QV
revol_util	Input Factors	QV
total_acc	Input Factors	QV
acc_now_delinq	Input Factors	QV
chargeoff_within_12_mths	Input Factors	QV
delinq_amnt	Input Factors	QV
pub_rec_bankruptcies	Input Factors	QV
Grade	Customer Demographics	OCV
Sub-Grade	Customer Demographics	OCV
home ownership	Customer Demographics	UOCV
annual_inc	Customer Demographics	QV
zip_code	Customer Demographics	UOCV
addr_state	Customer Demographics	UOCV
ID	Customer Information	UOCV
member_id	Customer Information	UOCV
verification_status	Customer Information	UOCV
issue_d	Customer Information	OCV

<Data Understanding>

Below is the List of required columns used for analysis

Col name	Description
id	It is a unique ID for the loan listing.
loan_amnt	This is the amount applied by the borrower for the loan process, ranging from 5000 to 35000 in this dataset. It has been binned with 5000 interval for analysis
funded_amnt	This is the approved loan amont. This data is similar with loan_amnt and treated similarly for analysis.(binned with 5000 interval,ranging from 5000 to 35000)
term	this is the number of payments on the loan, either could be 36 or 60.
int_rate	This is interest rates applied to the loan, ranging from 5%-20%. It has been binned with interval of 5
installment	This is the monthly payment owed by the borrower ranging from 20 to 1300 . After dealing with the outlayers, it has been ranged from 200 to 800, with 200 interval.
grade	there are 7 types of grade (a b c d e f g)
sub_grade	each grade is further catagorised in 5 sub grade making it a1,a2,a2,a4,a5,b1,b2,b3 etc… till g5
emp_length	This is the duration of the employment of borrower ranging from 0 to 10+ years.
home_ownership	there are 4 type of homw ownersghip in this dataset i.e. own, rent, mrtgage, other
annual_inc	this is the annual income declaired by the borrower. In this data set we had data starting from 4000 to 6000000, after treating the outlayers, it has a range of 20000 to 1200000 an d is binned with an inter val of 20000.
verification_status	As per this dataset, there are 3 verification status i.e verified , not verified, source verified.
issue_d	This is the loan issue date, we have considered each year from 2007 to 2011.
loan_status	As per this dataset, loan status is the main factor of analysis . There is 3 type of status fully paid, current and chared off. Chared off is considered as default.
purpose	purpose of the loan is consider as one of the factor of analysis. These purpose are like (credit card, car, home renovation,medical etc) . Total 14 are defined in this dataset.
zip_code	
addr_state	there are 51 state in this dataset.
dti	this is the Debt-to-income ratio, raning from 0 to 30 in this dataset. For analysis purpose we have binned it from less >5 to more<25 with a interval of 5.
delinq_2yrs	This is the number of incidences of delinquency in the borrower's credit file, ranging from 0 to 11
inq_last_6mths	this is the number of inquiries in past 6 months, ranging from 0 to 8
open_acc	after dealing with the outlayed the open account associated with the is ranging fron >5 to 25<
pub_rec	
revol_bal	
revol_util	
total_acc	after dealing with the outlayed the open account associated with the is ranging fron >5 to 25<
pub_rec_bankruptcies	the public record of bankruptcies ranging from 0 to 2.

<Data Preparation>

Data Cleaning & Manipulations :

- In given loan dataset ***54 Variables*** contain all the observations as **NAs**, which are removed.
- ***14 Variables*** have more than **70% of 0s**, so they are removed.
- ***Date*** is converted into standard format and ***%*** is removed from columns wherever required.
- Removed the columns having more than **50% of NAs**.
- Making all the required text columns to **LOWER CASE** (Grade, Purpose, Loan_Status, Sub_grade, verification_status, home_ownership & addr_state)

<Data Preparation>

UpGrad

Data Cleaning & Manipulations :

There are 111 number of columns out of which there are 26 columns which we think that are uselful for our analyis , So will be reporting the issues for those columns which we are going to use.

1.term : It is object format and has a string attached to it. So converted it to numeric format

2.int_rate : It is also in object format and has percentage attached to it. So converted it to numeric format

3.emp_length: There are many issues present in it. Issue that has been solved are missing value treatment. The missing value are assigned as 0. And the values which is present as 10 > years is taken as 10 years and similary for < 1 year is taken as 1.

these assumptions are done for our convinent in calculation. And it is also converted to float format.

4.annual_inc: Here the annual income is divided by 1000 to convert them into thousand format.

5.zip_code: Zip code is reported in object format. It has XXX attached to it. So we removed the XXX and made it to numeric.

6.revol_util: The value has percentage attached to it, so it is reported in the object format. So removing the percentage and converting into the numeric format.

7.issue_d: Date and time format are reported in a wrong way which also cannot be used for analyis. So the issue_d is converted to right date format so that it can be easily used with python.

8.loan_amnt: The loan_amnt data is highly skewed , so outlier treatment is done

9.int_rate: The int_rate data is highly skewed , so outlier treatment is done

10.installment: The installment data is highly skewed, So outlier treatment is done

11.annual_inc: The annual_inc data is highly skewed, So outlier treatment is done

12.open_acc: The open_acc data is highly skewed, So outlier treatment is done.

13.revol_bal: The revol_bal data is highly skewed , So outlier treatment is done.

14.total_acc: The total_acc data is highly skewed , So outlier treatment is done.

15.revol_util: The revol_util has lots of null values, but the revol_bal is zero for all field. So the fields are assigned as 0 for the null values. If the revol_bal is zero for a person definitely the revol_util will be zero for him.

16.funded_amnt: The funded_amnt data is highly skewed, So outlier treatment is done

<Data Analysis>

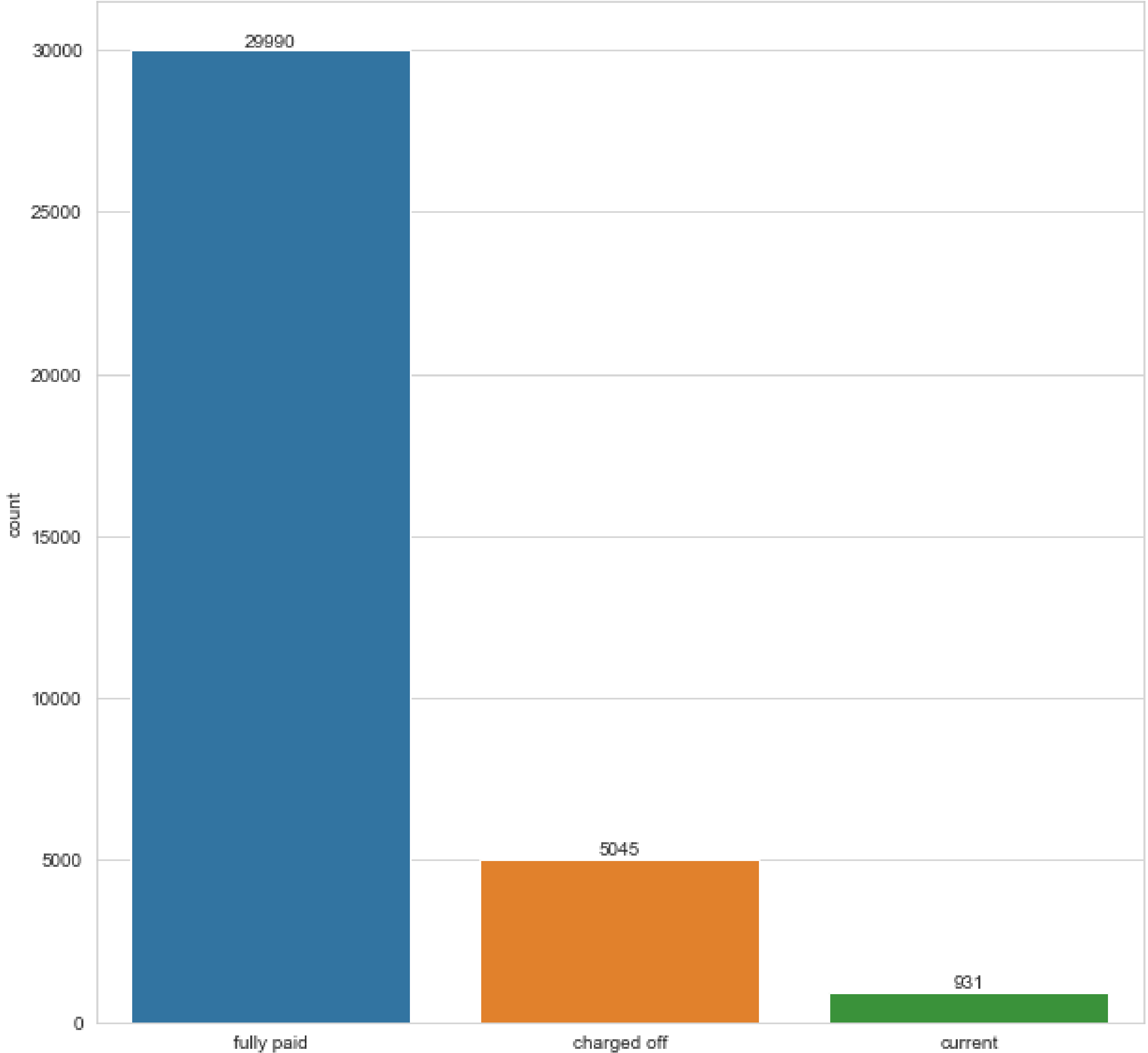
Further, the following have been accomplished:

- Univariate Analysis
  - For Categorical Variables
  - For Numeric Variables
  - For Segmented Variables
- Segmented Univariate Analysis
- Bi-variate Analysis
  - Keeping loan_status fixed in one of the columns
  - Scatter plots
- Provide insights based on the results of the above

<Plot 1 - Total Loans by Status (Univariate) >

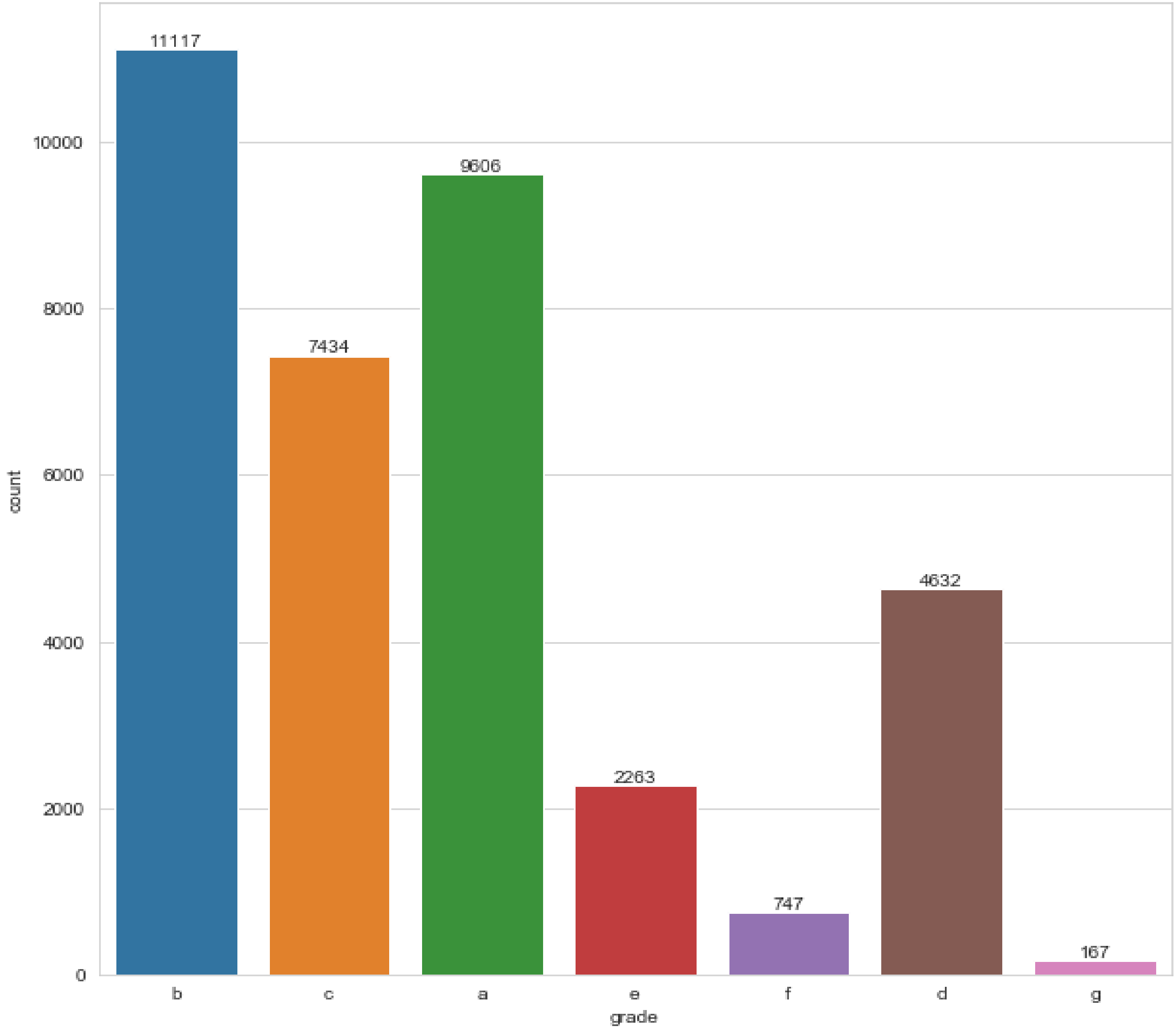
UpGrad

- It is clear that there is significant amount of ***‘Charged Off’*** loans which account for about ***14%*** of the total loan amount.
- A reduction in the total number of ***‘Charged Off’*** i.e. Defaulted loans can result in the bank avoiding financial loss and should therefore be assessed further



<Plot 2 - Total Loans by Grade (Univariate) >

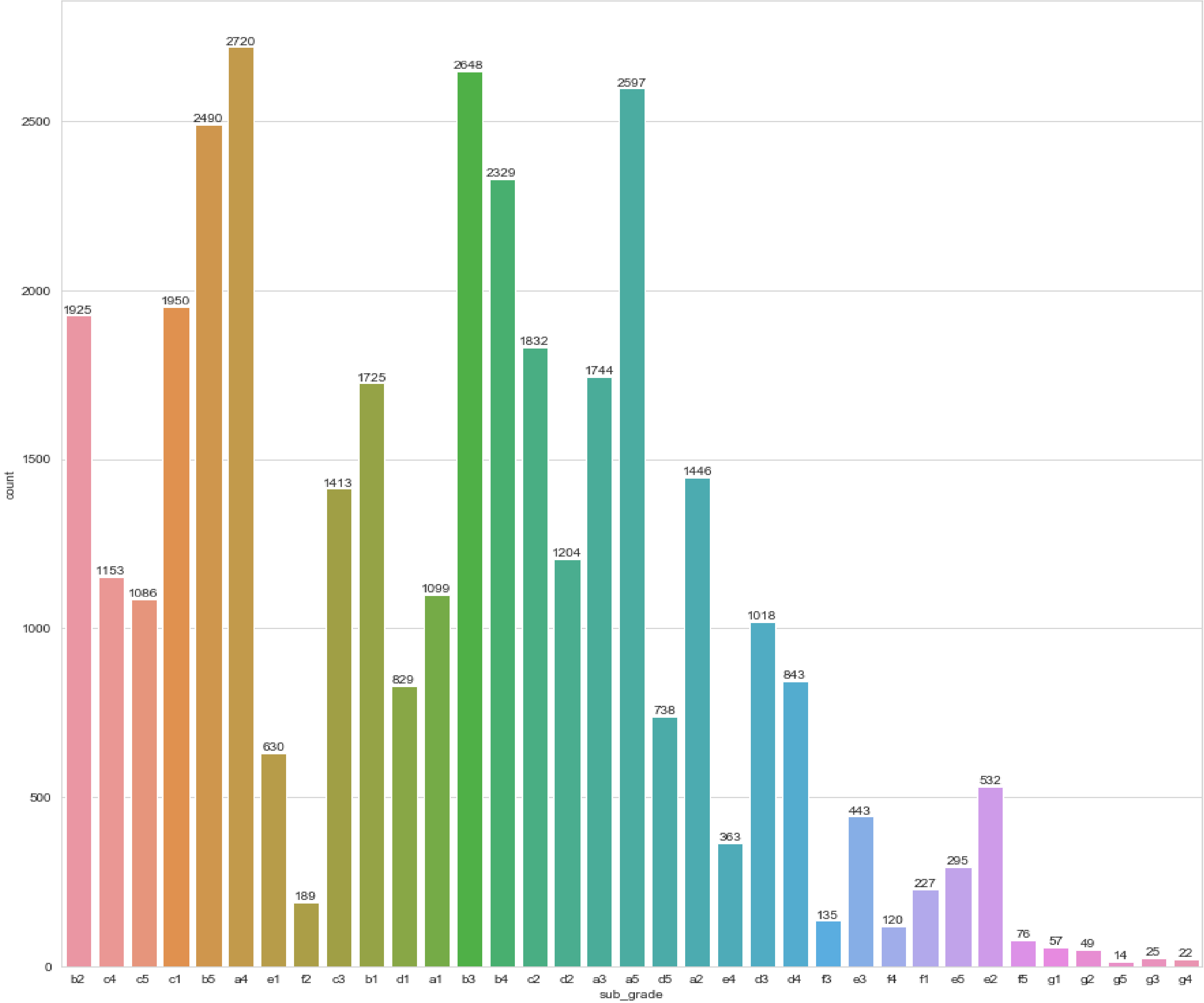
- It is clear that majority of loan(s) are under LOAN GRADE – A & B which is > 9000



<Plot 3 - Total Loans by Sub_Grade (Univariate) >

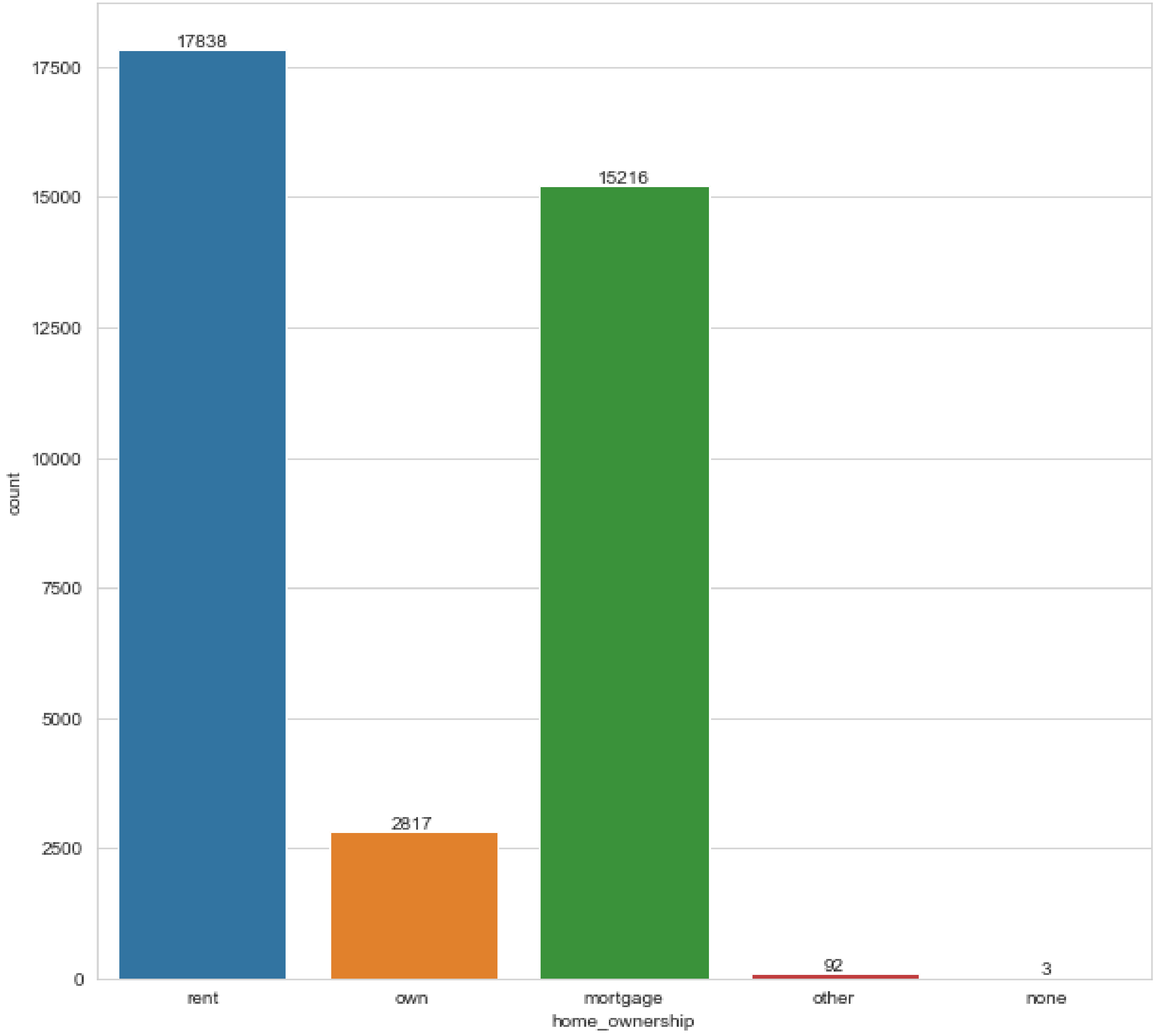
UpGrad

- It is clear that majority of loan(s) are under LOAN SUB GRADE – A4, B3, & A5 which is > 2500



<Plot 4 - Total Loans by Home_Ownership (Univariate) >

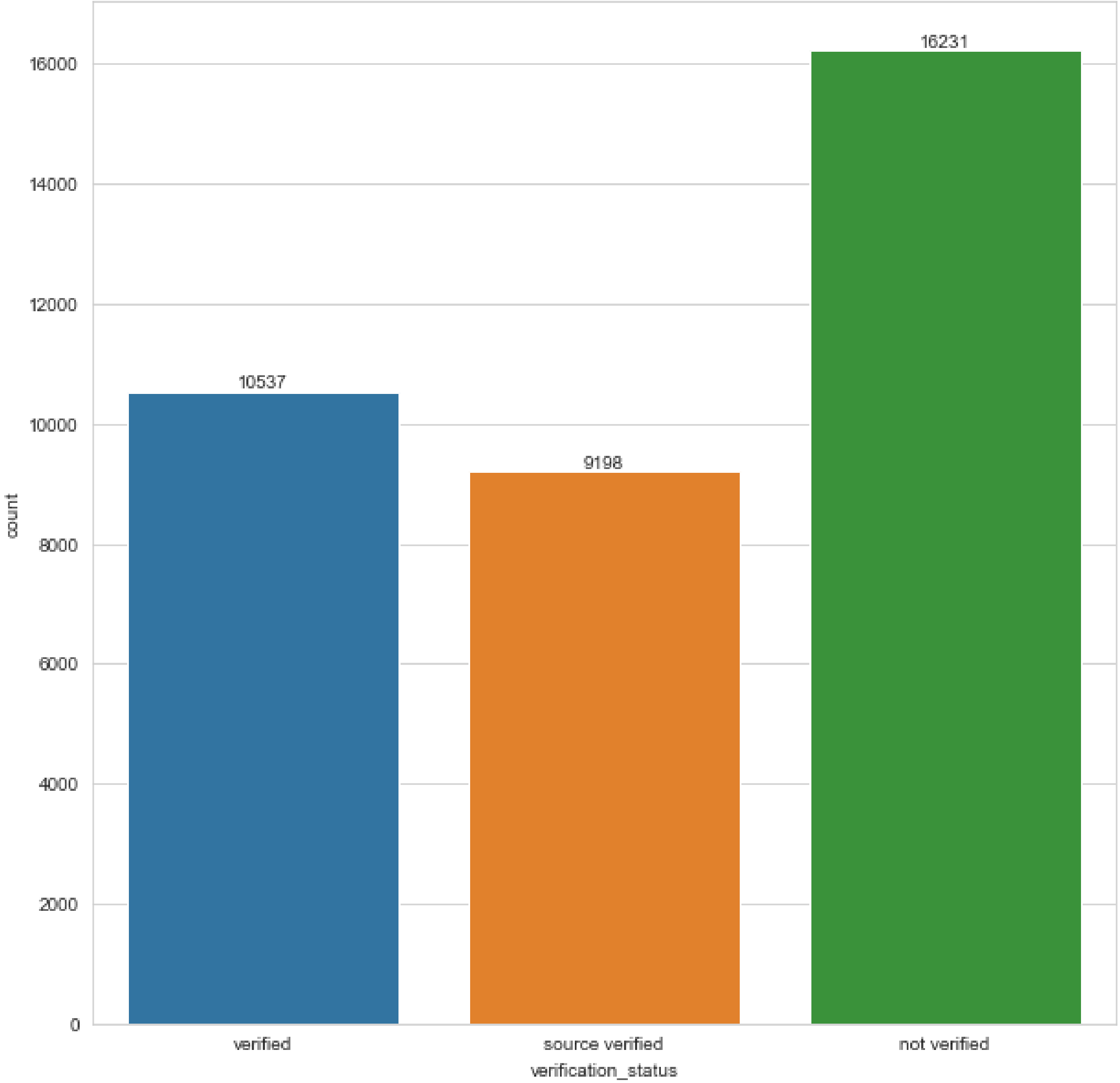
- It is clear that majority of loan(s) are under HOME_OWNERSHIP – RENT & MORTGAGE which is > 15000



<Plot 5 - Total Loans by Verification_Status (Univariate) >

UpGrad

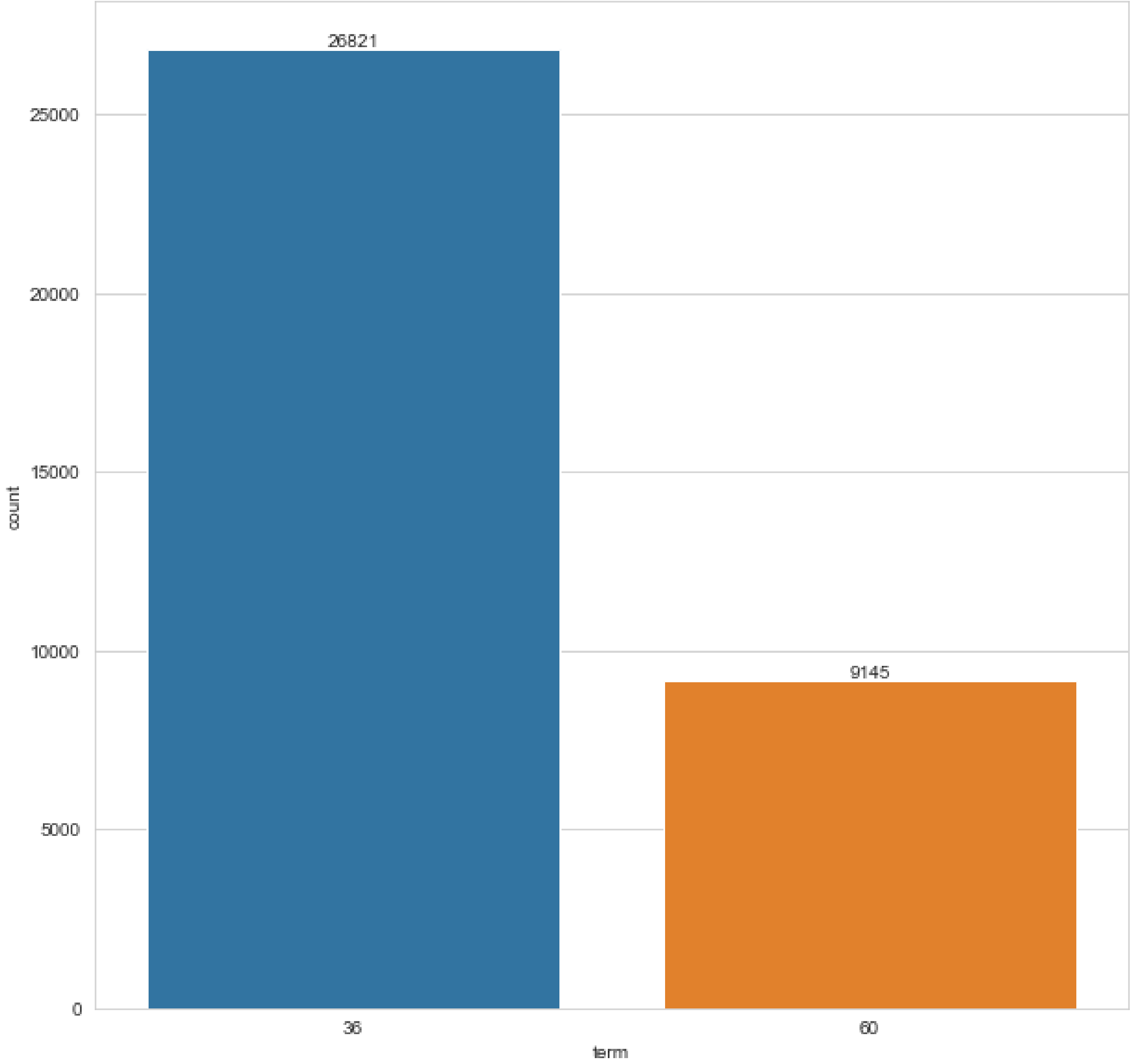
- It is clear that majority of loan(s) are under Verification_Status – NOT VERIFIED which is > 16000



<Plot 6 - Total Loans by Term (Univariate) >

UpGrad

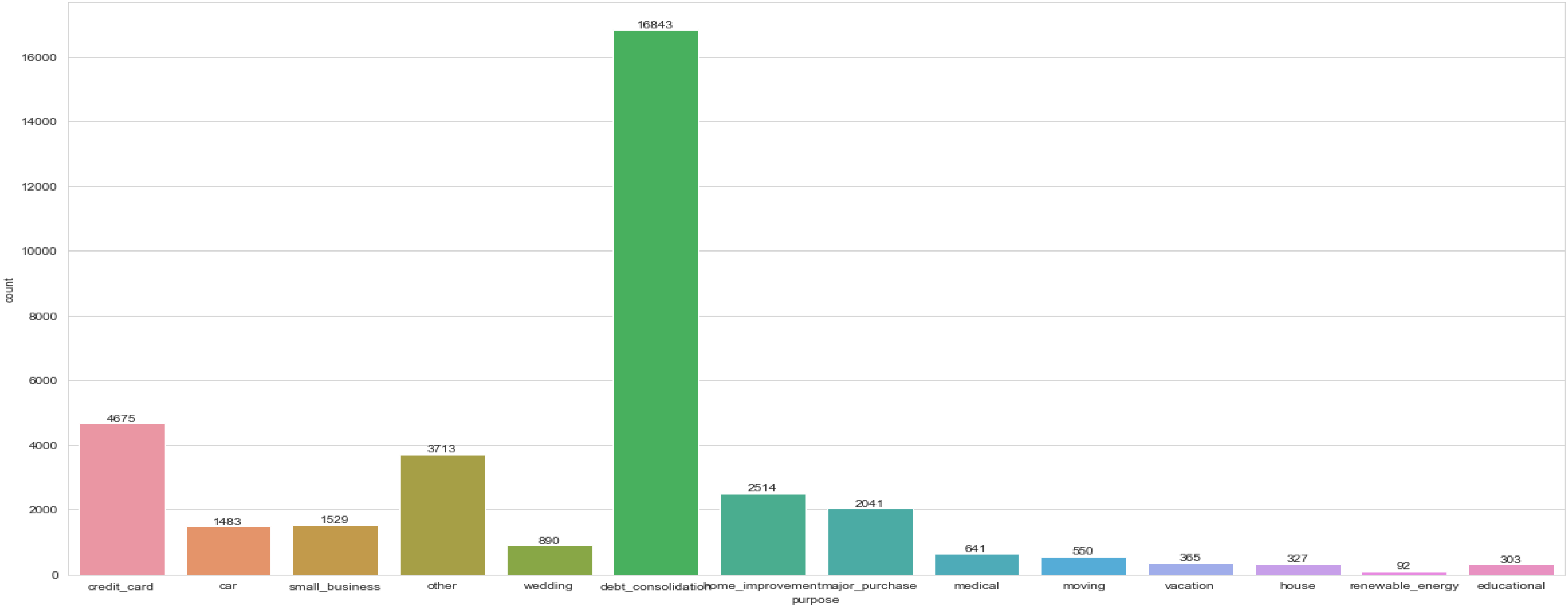
- It is clear that majority of loan(s) are under Term – 36 months > 26000



<Plot 7 - Total Loans by Purpose (Univariate) >

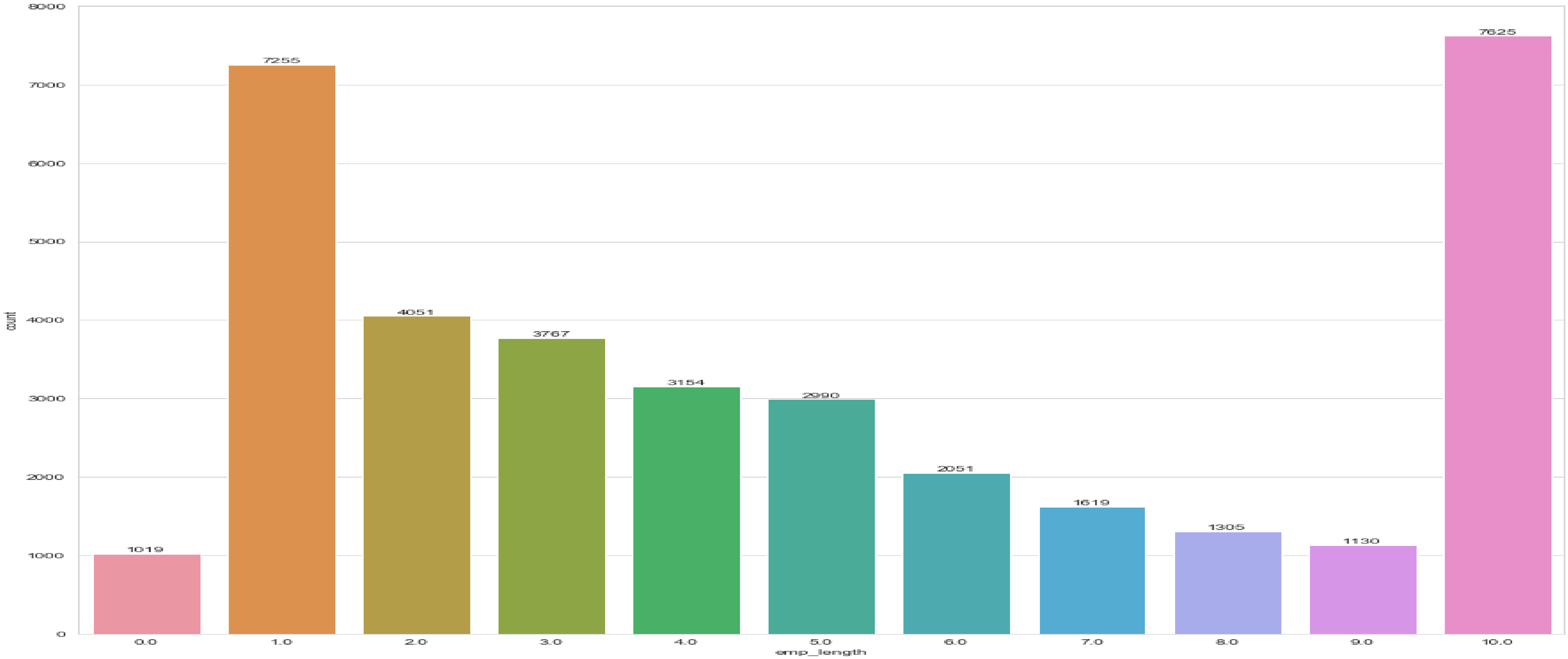
UpGrad

- It is clear that majority of loan(s) are under Purpose – Debt_Consolidation, Credit_card and Other which is > 3700



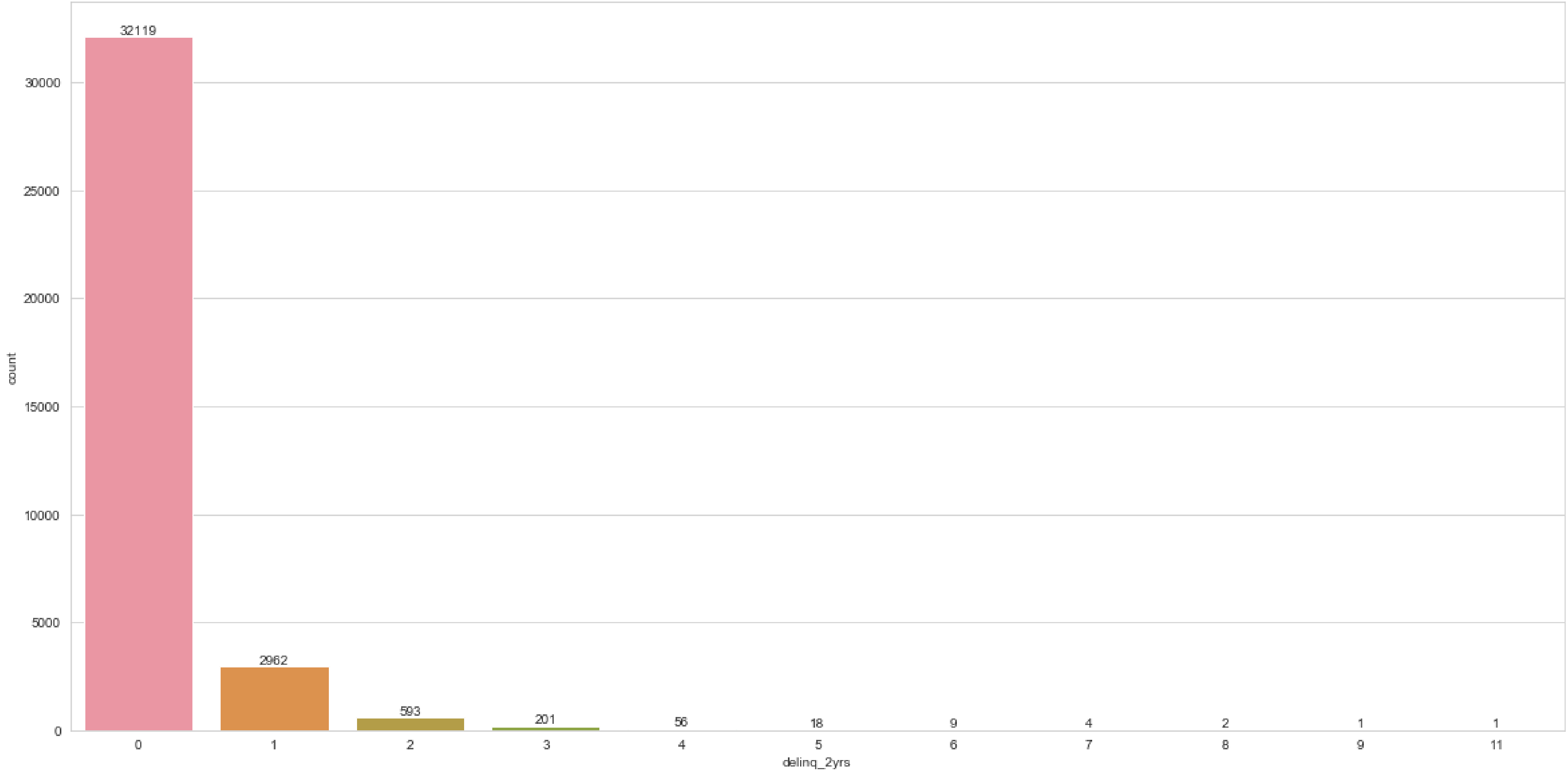
<Plot 8 - Total Loans by Employee_Length (Univariate) >

- It is clear that majority of loan(s) are under Employee_Length – 10 years & 1 year which is > 7200



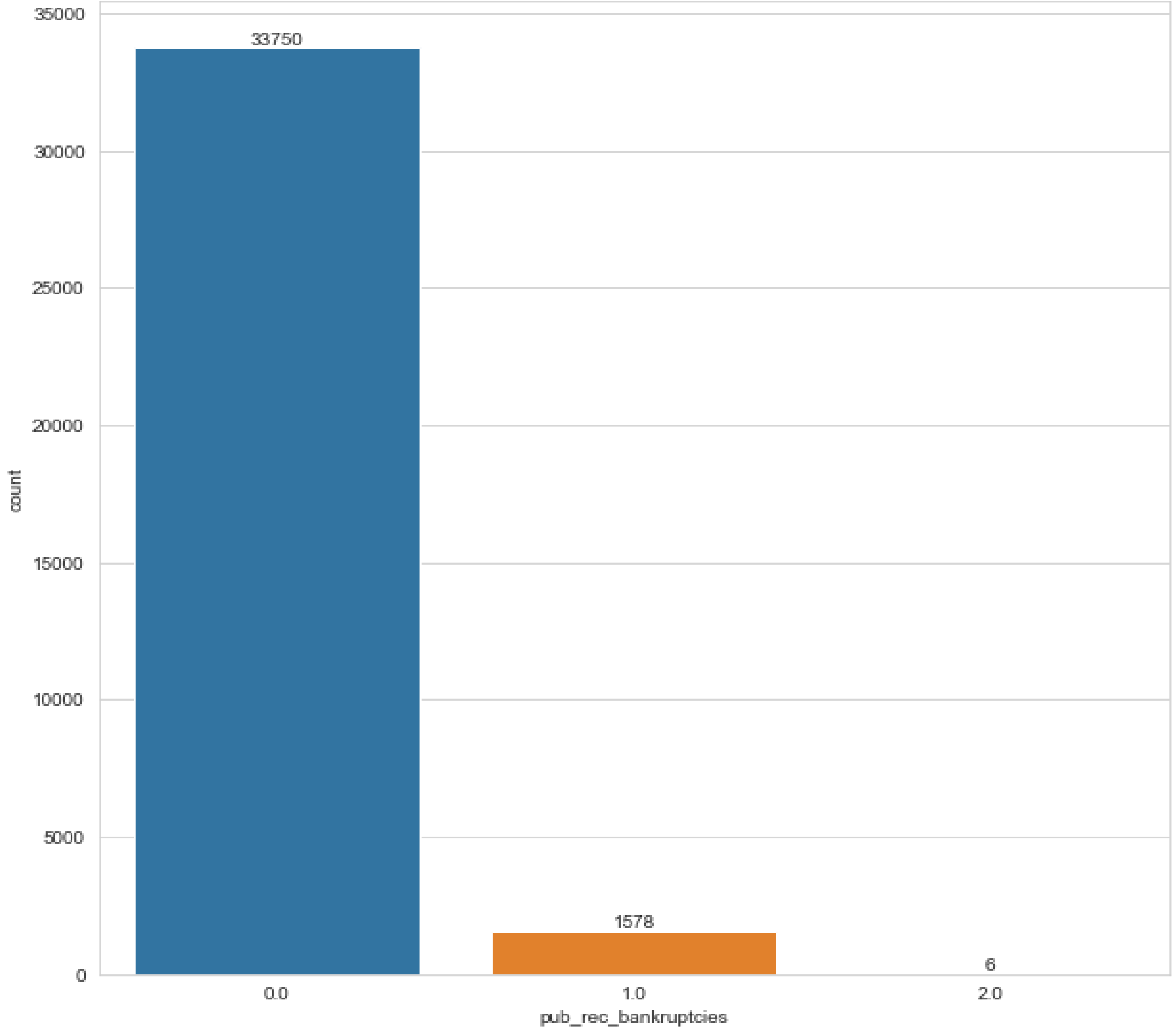
<Plot 9 - Total Loans by Delinq_2yrs (Univariate) >

- It is clear that majority of loan(s) are under Delinq_2yrs – 0 & 1 which is > 2900



<Plot 10 - Total Loans by Pub_rec_bankruptcies (Univariate) >

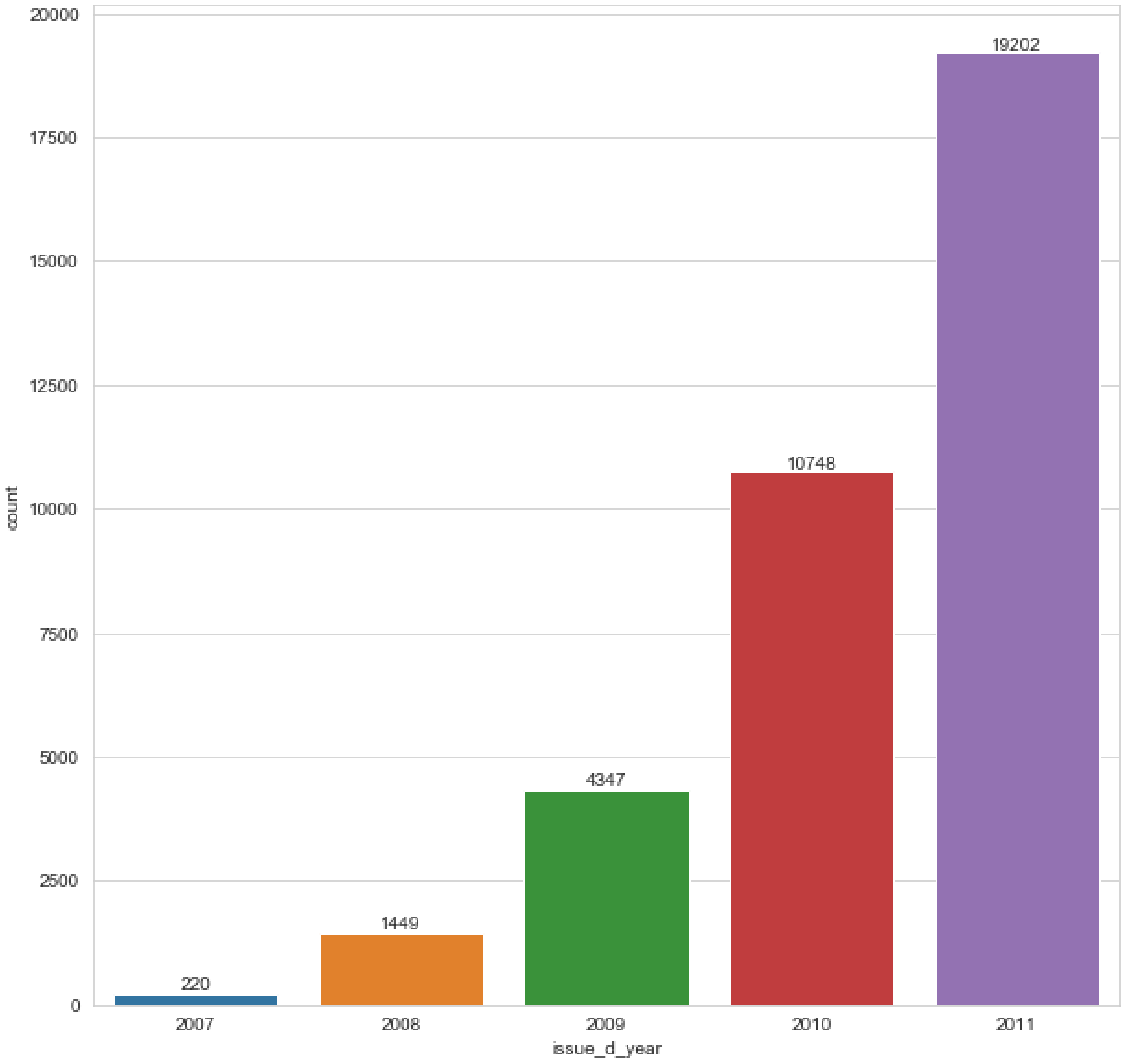
- It is clear that majority of loan(s) are under pub_rec_bankruptcies – 0, which is = 33750



<Plot 11 - Total Loans by Issue_Year(Univariate)>

UpGrad

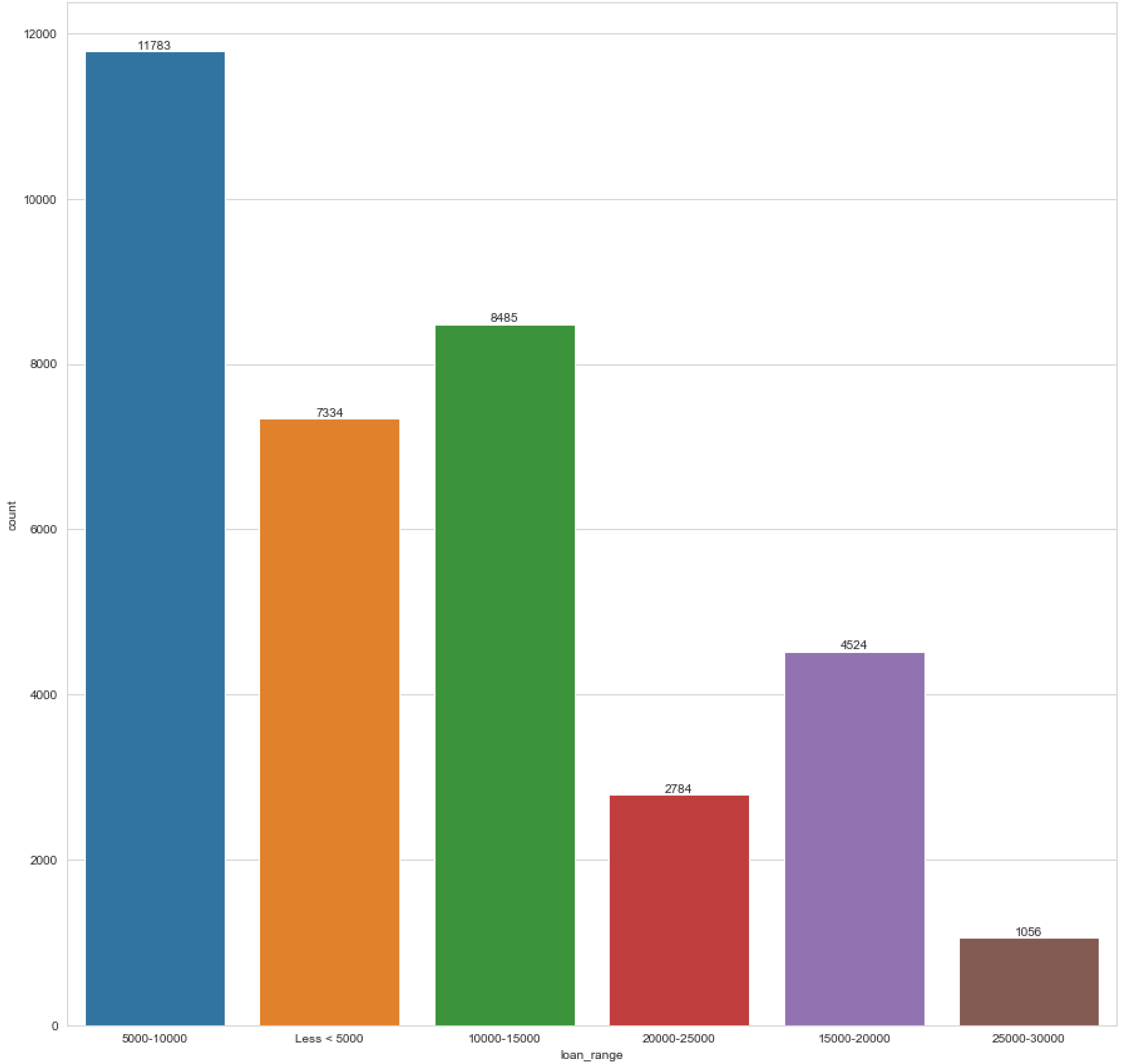
- It is clear that majority of loan(s) are gradually increasing from year-year



<Plot 1 - Total Loans by Loan_Amount range (Segmented)>

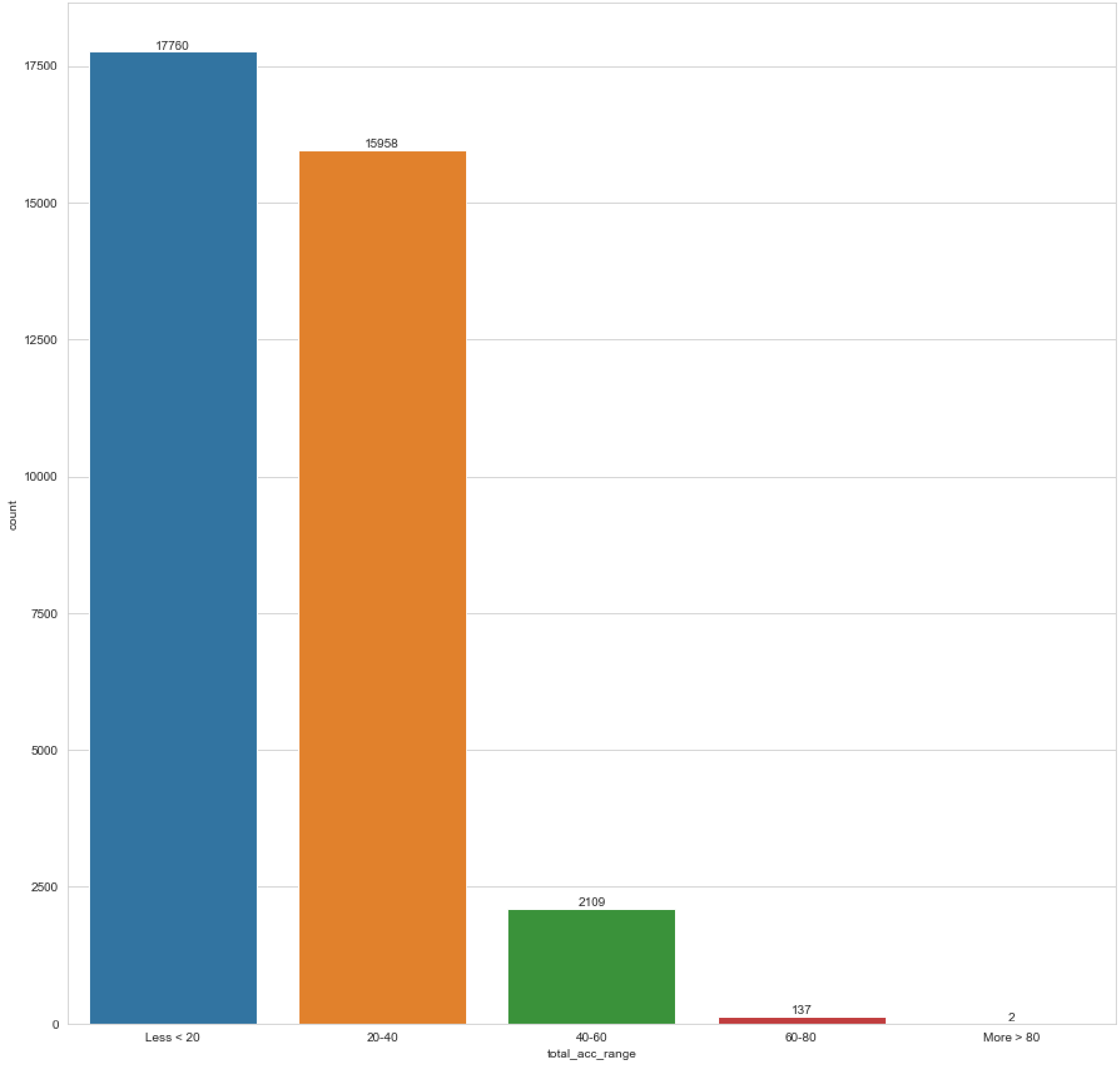
UpGrad

- It is clear that majority of loan(s) are categorized under top 3 ***Loan_Amount*** range between **5000-10000, 10000-15000 and Less < 5000** which is **11783, 8485 & 7334**



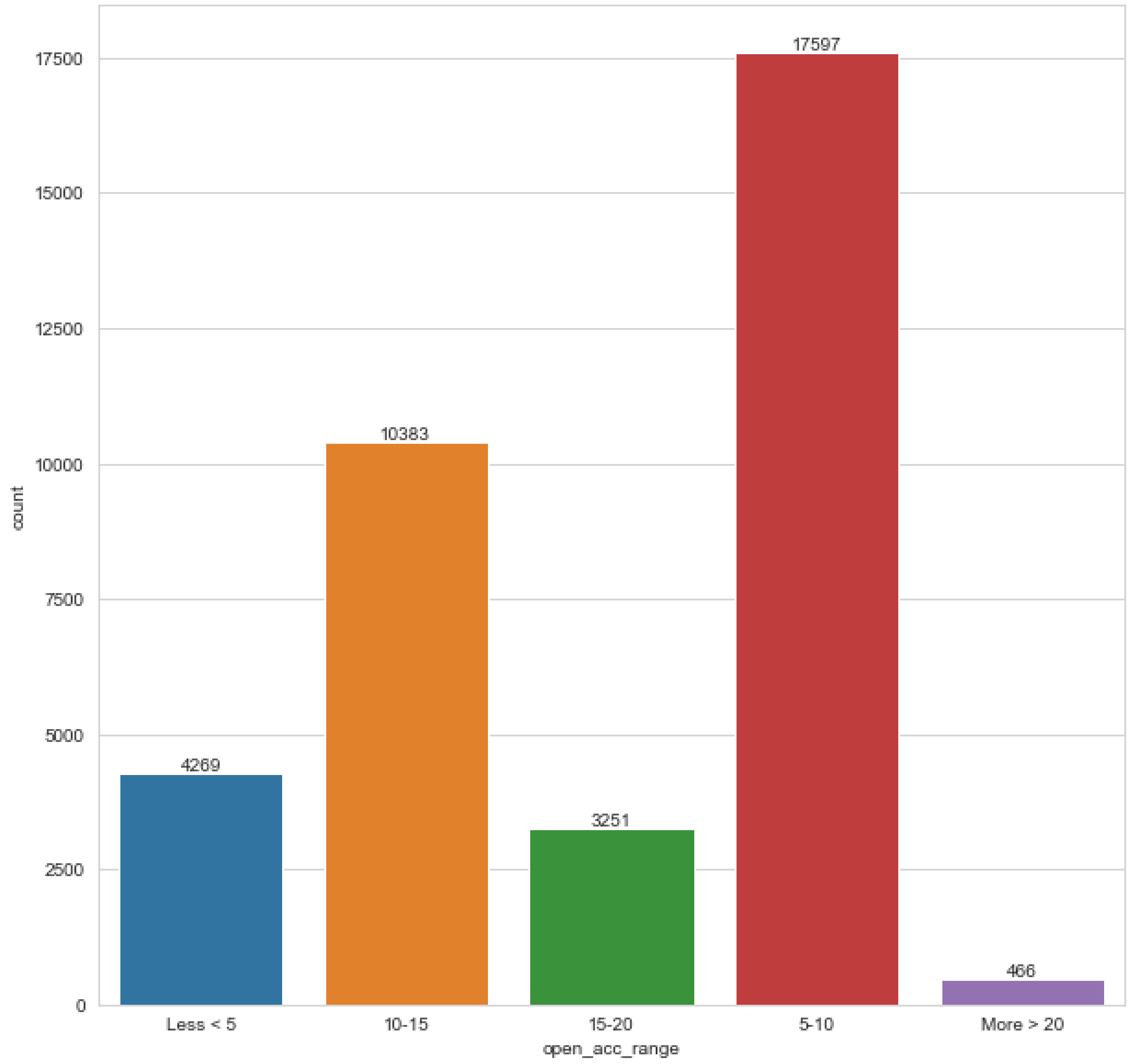
<Plot 2 - Total Loans by Accounts range (Segmented)>

- It is clear that majority of loan(s) are categorized under top 3 ***Accounts*** range between **Less<20, 20-40 and 40-60** which is **17760, 15958 & 2169**



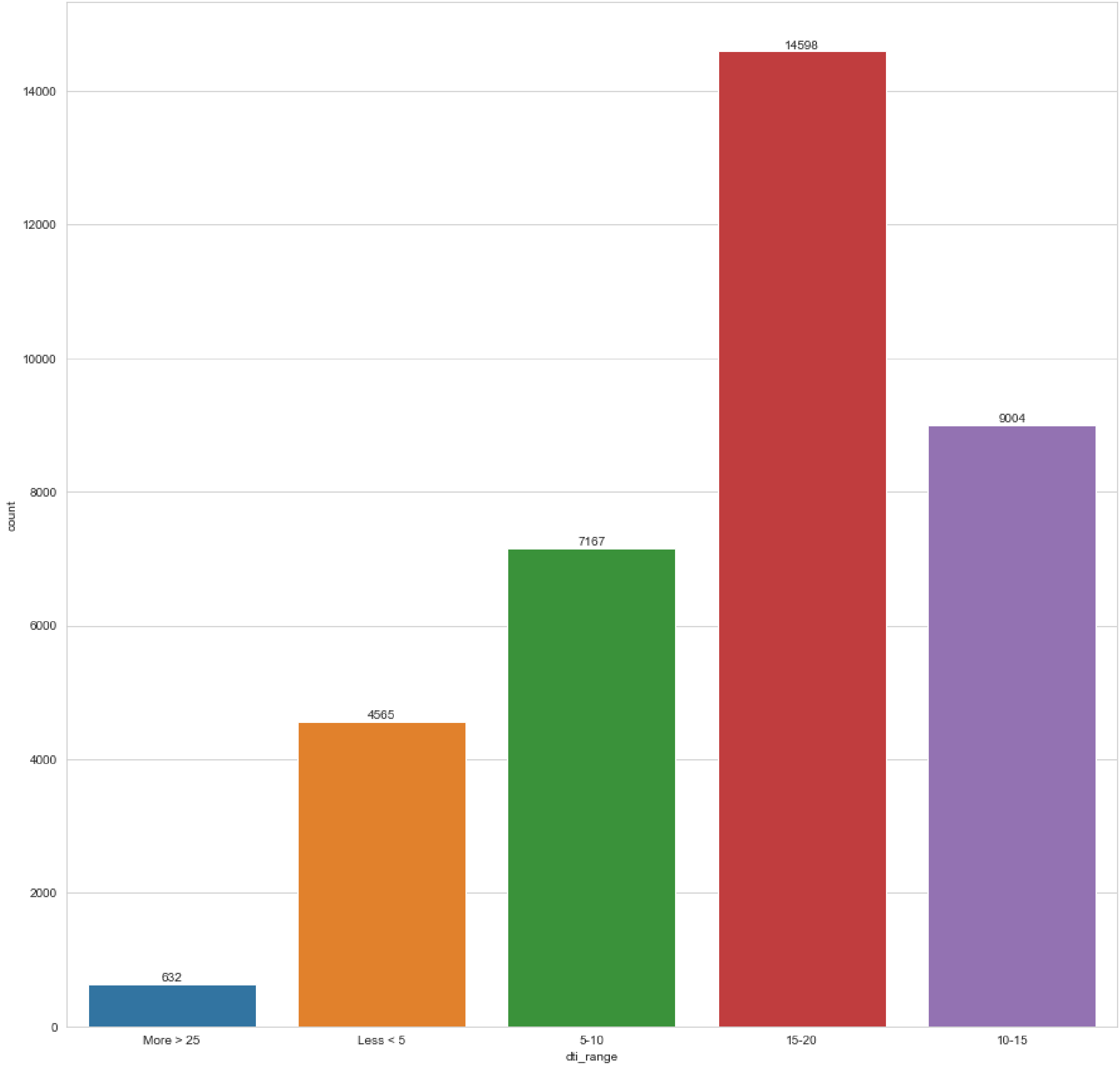
<Plot 3 - Total Loans by Open_Account range (Segmented)>

- It is clear that majority of loan(s) are categorized under top 3 ***Open_Accounts*** range between **5-10, 10-15 and Less<5** which is **17507, 10383 & 4269**



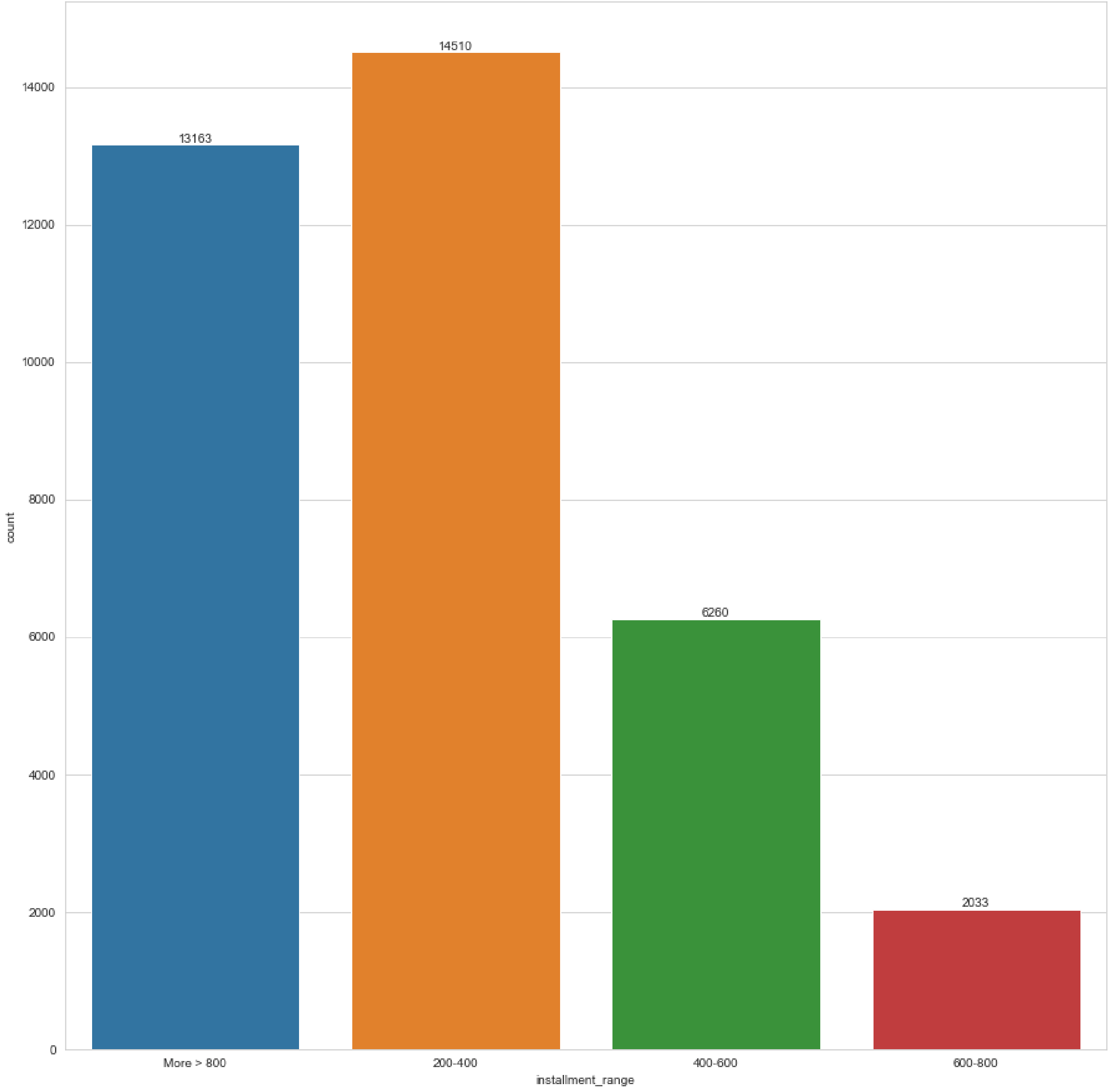
<Plot 4 - Total Loans by DTI range (Segmented)>

- It is clear that majority of loan(s) are categorized under top 3 ***DTI*** range between 1**5-20, 10-15 and 5-10** which is **14598, 9004 & 7167**



<Plot 4 - Total Loans by Installment range (Segmented)>

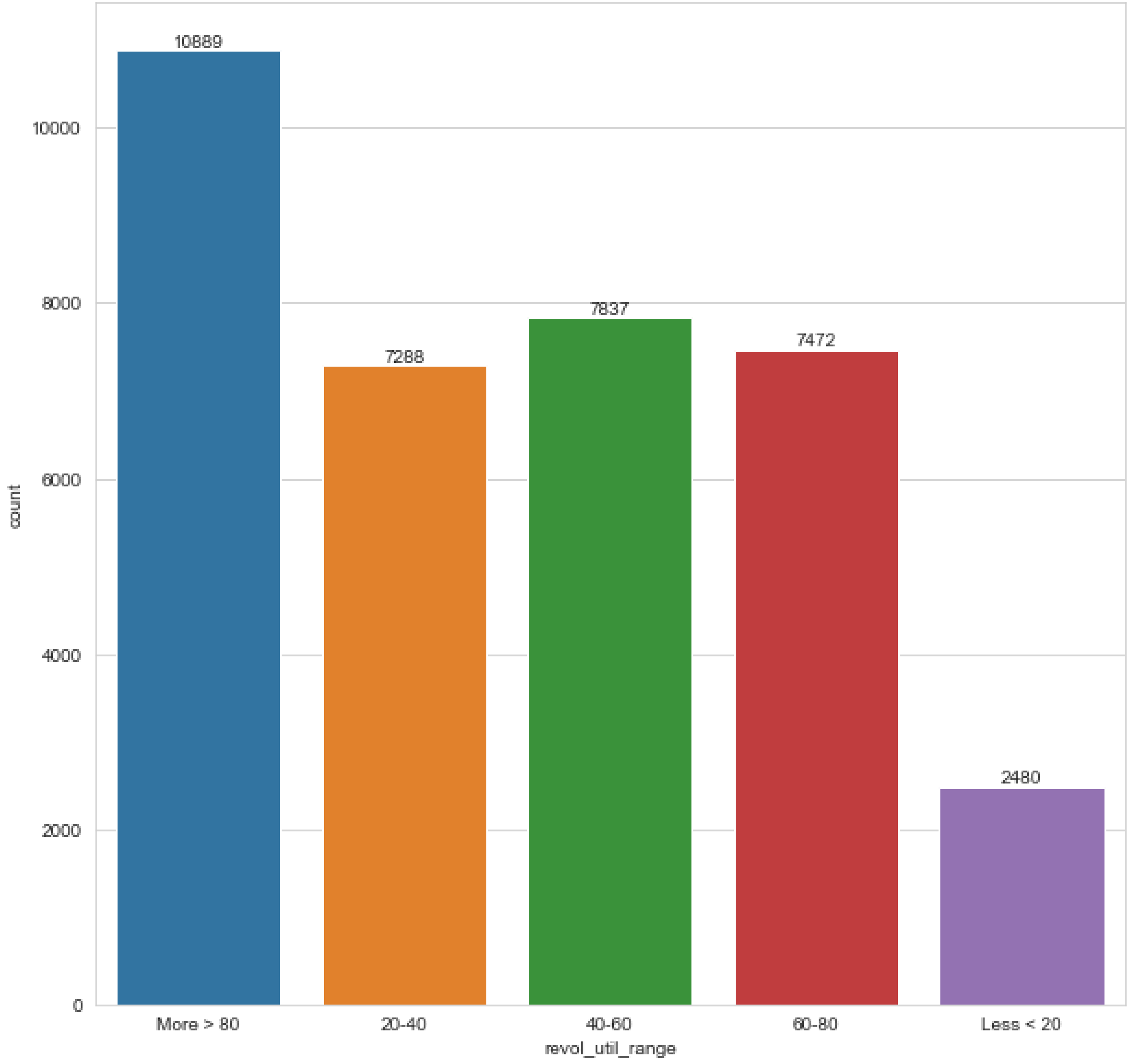
- It is clear that majority of loan(s) are categorized under top 3 ***Installment*** range between 200-**400, more>800 and 400-600** which is **14510, 13163 and 2033**



<Plot 6 - Total Loans by Revol_Util range (Segmented)>

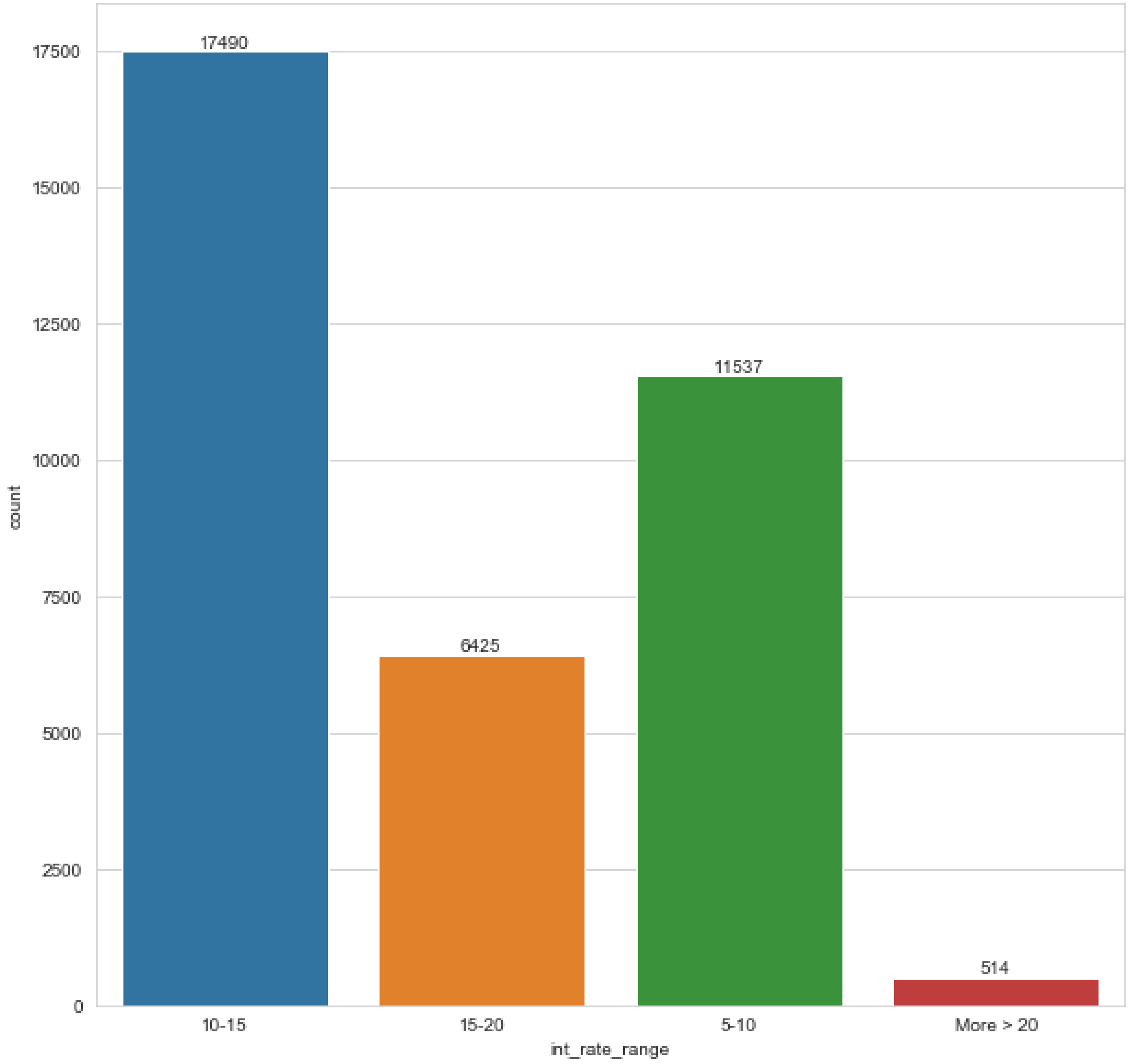
UpGrad

- It is clear that majority of loan(s) are categorized under top 3 ***Revol_util*** range between **more>80, 40-60 and 60-80** which is **10889, 7837and 7472**



<Plot 7 - Total Loans by Int_Rate range (Segmented)>

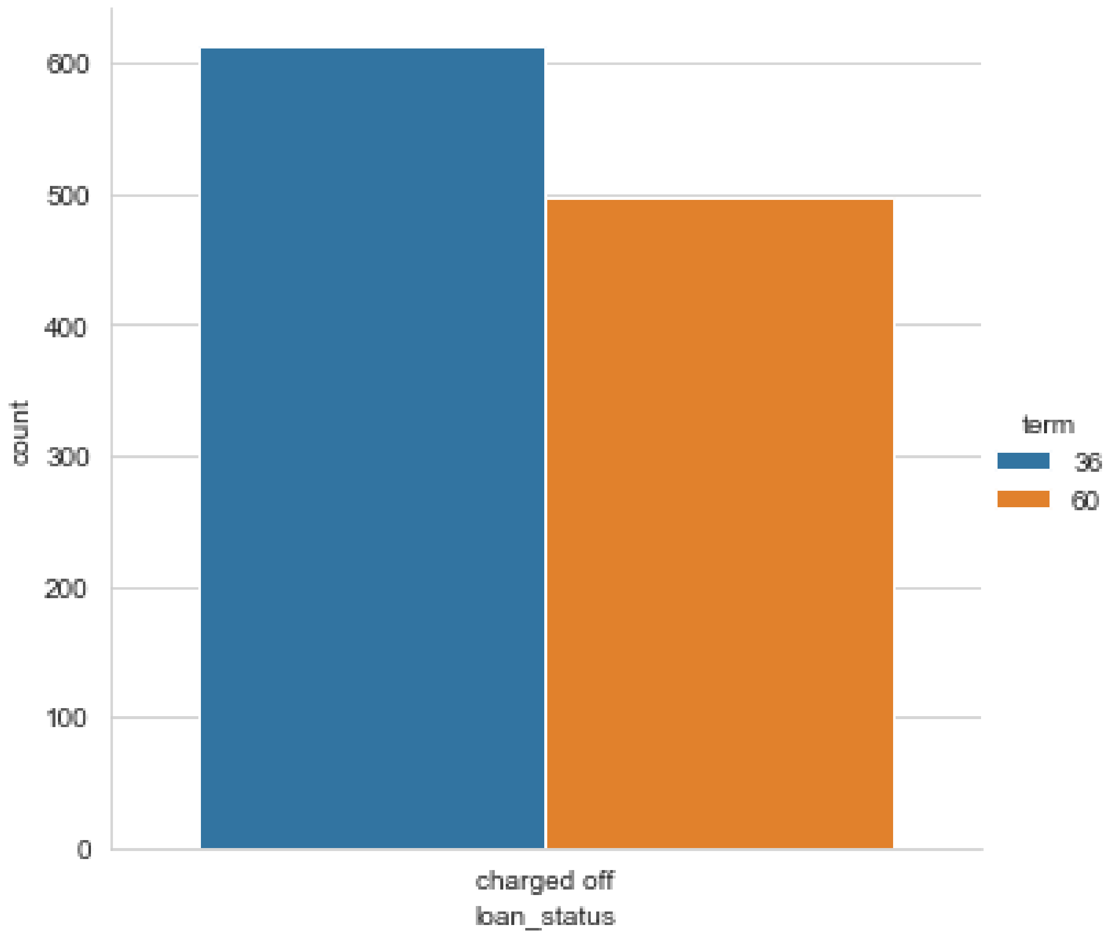
- It is clear that majority of loan(s) are categorized under top 3 ***Int_Rate*** range between **10-15, 5-10 and 15-20** which is **17490, 11537 & 6425**



<Plot 1 – Loan_Status vs Term (BiVariate)>

UpGrad

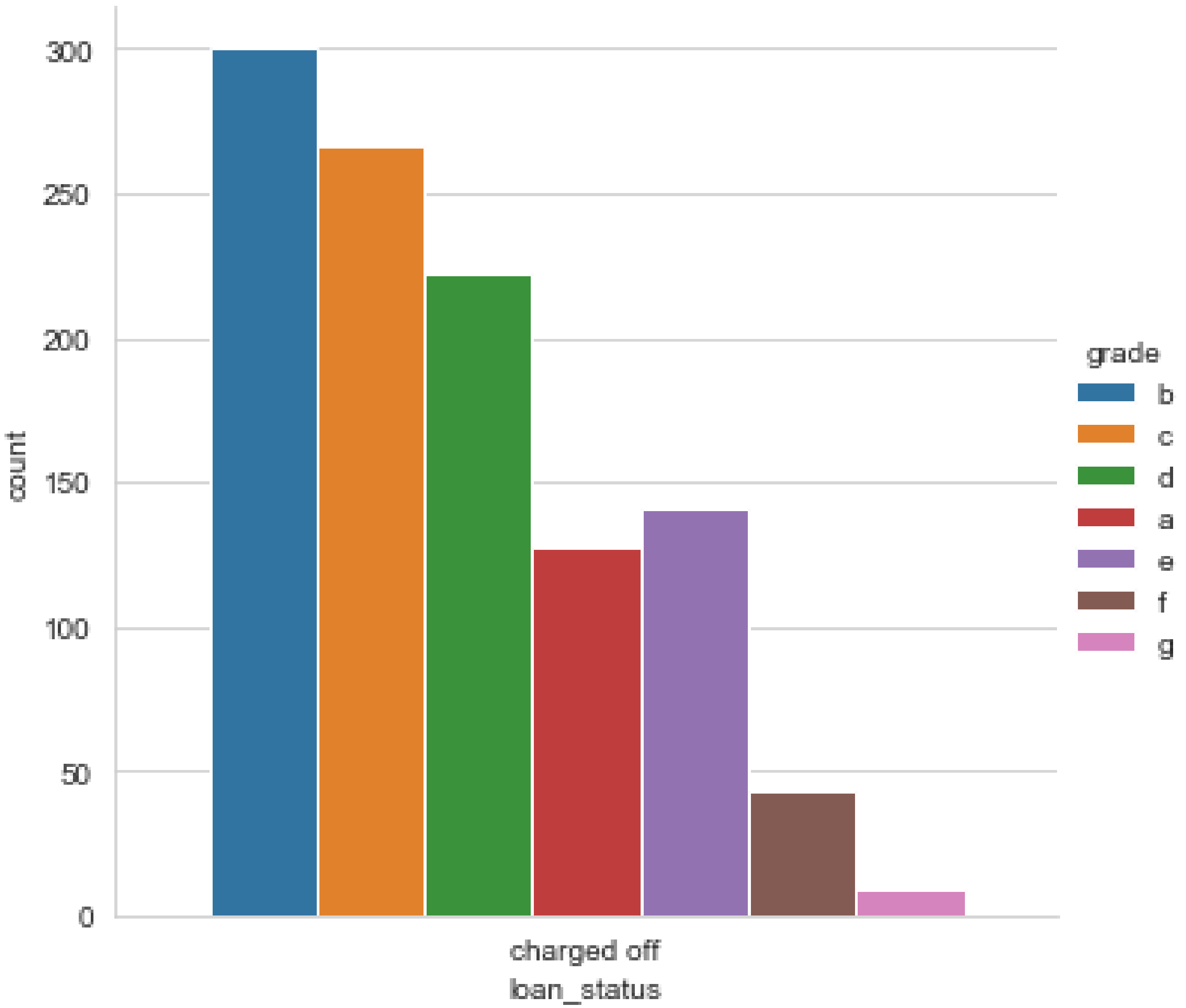
- It is clear that majority of loan(s) are getting Charged-Off with Terms as **36 months**



<Plot 2 – Loan_Status vs Grade (BiVariate)>

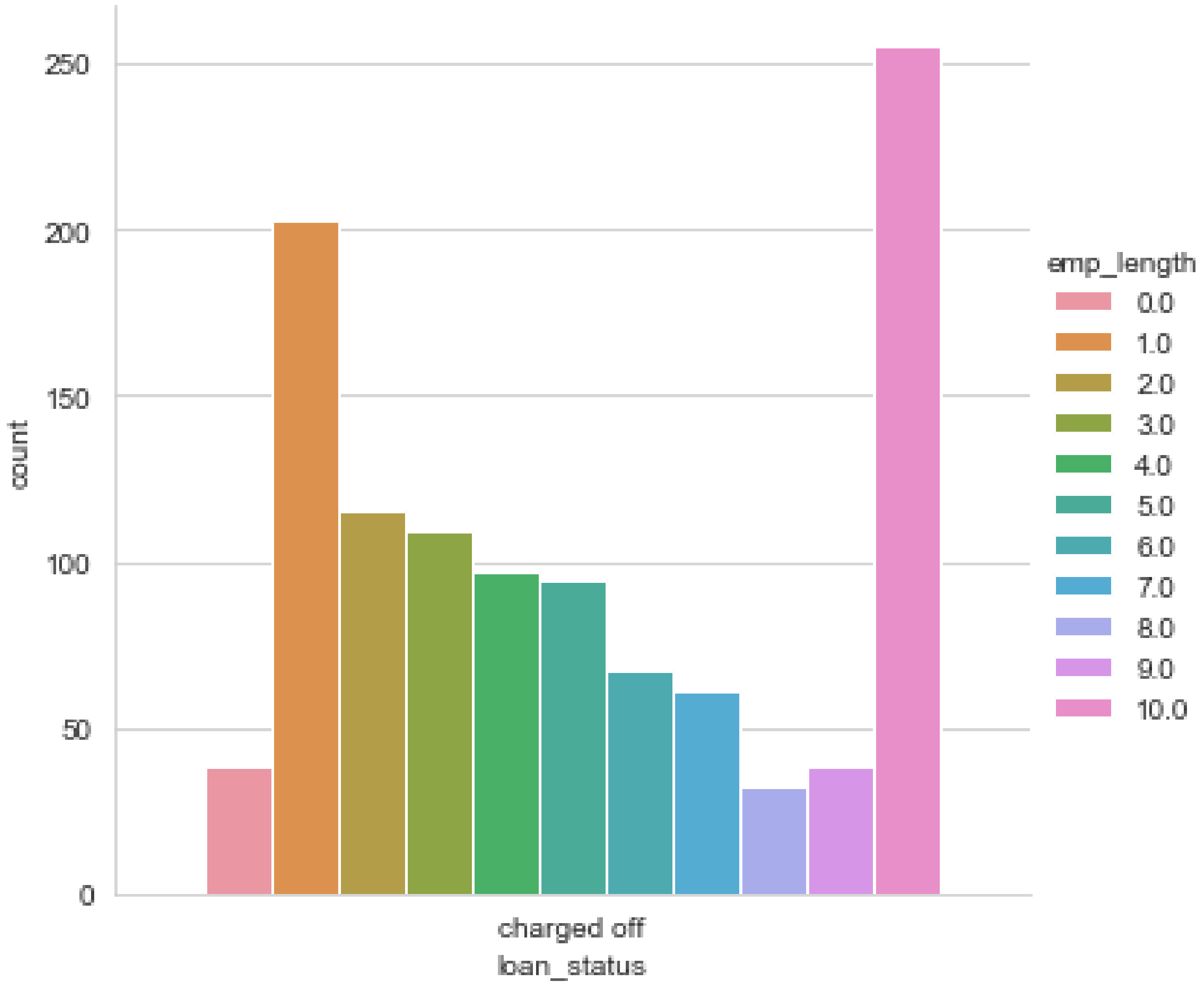
UpGrad

- It is clear that majority of loan(s) are getting Charged-Off with Grade as **B, C & D**



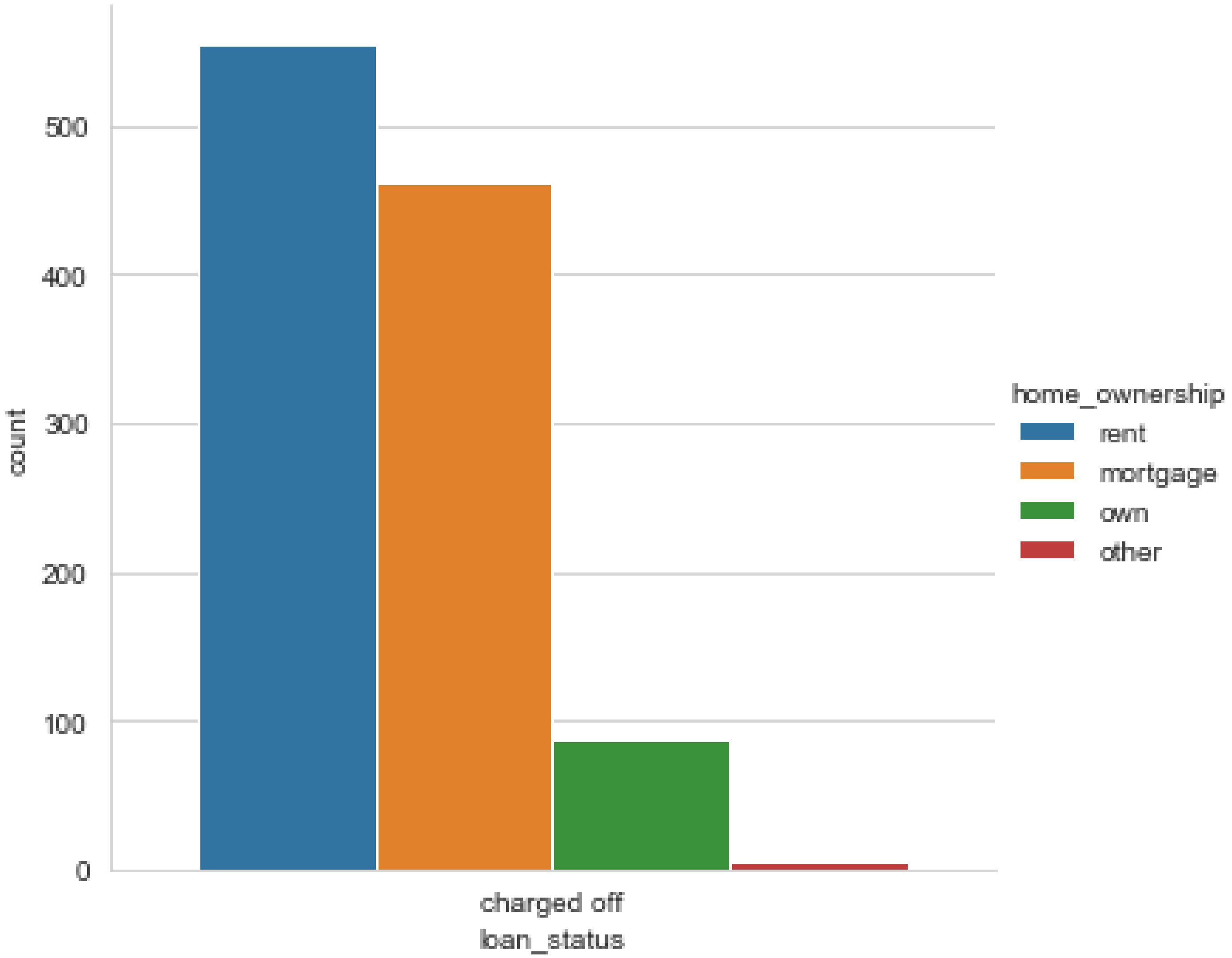
<Plot 3 – Loan_Status vs Emp_length (BiVariate)>

- It is clear that majority of loan(s) are getting Charged-Off with Emp_length is ***unkown*** have a greater chance followed by ***less than one year***



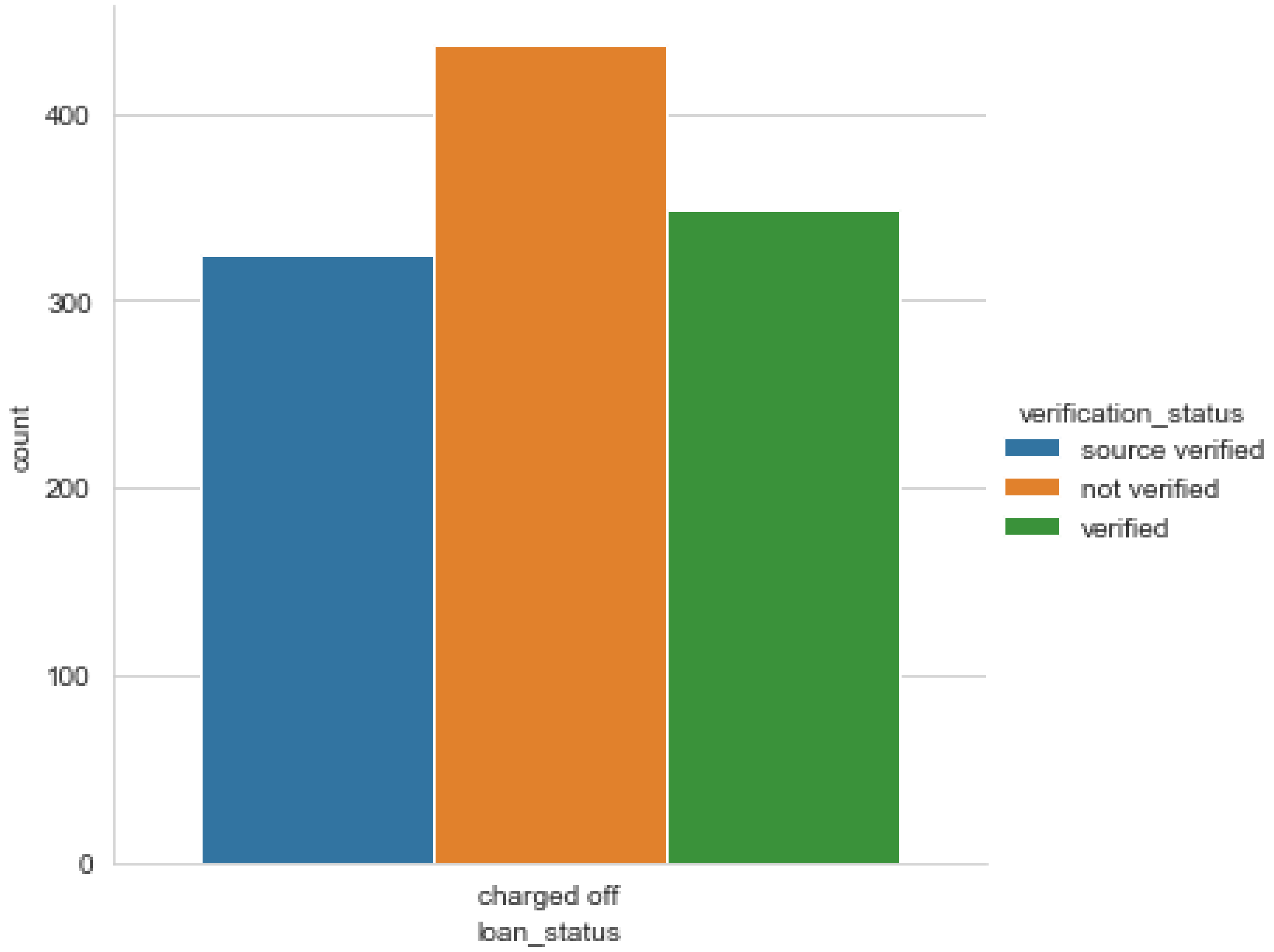
<Plot 4 – Loan_Status vs Home_Ownership (BiVariate)>

- It is clear that majority of loan(s) are getting Charged-Off with Home_Ownership is ***rent*** have a greater chance followed by ***mortgage***



<Plot 5 – Loan_Status vs Verification_Status (BiVariate)>

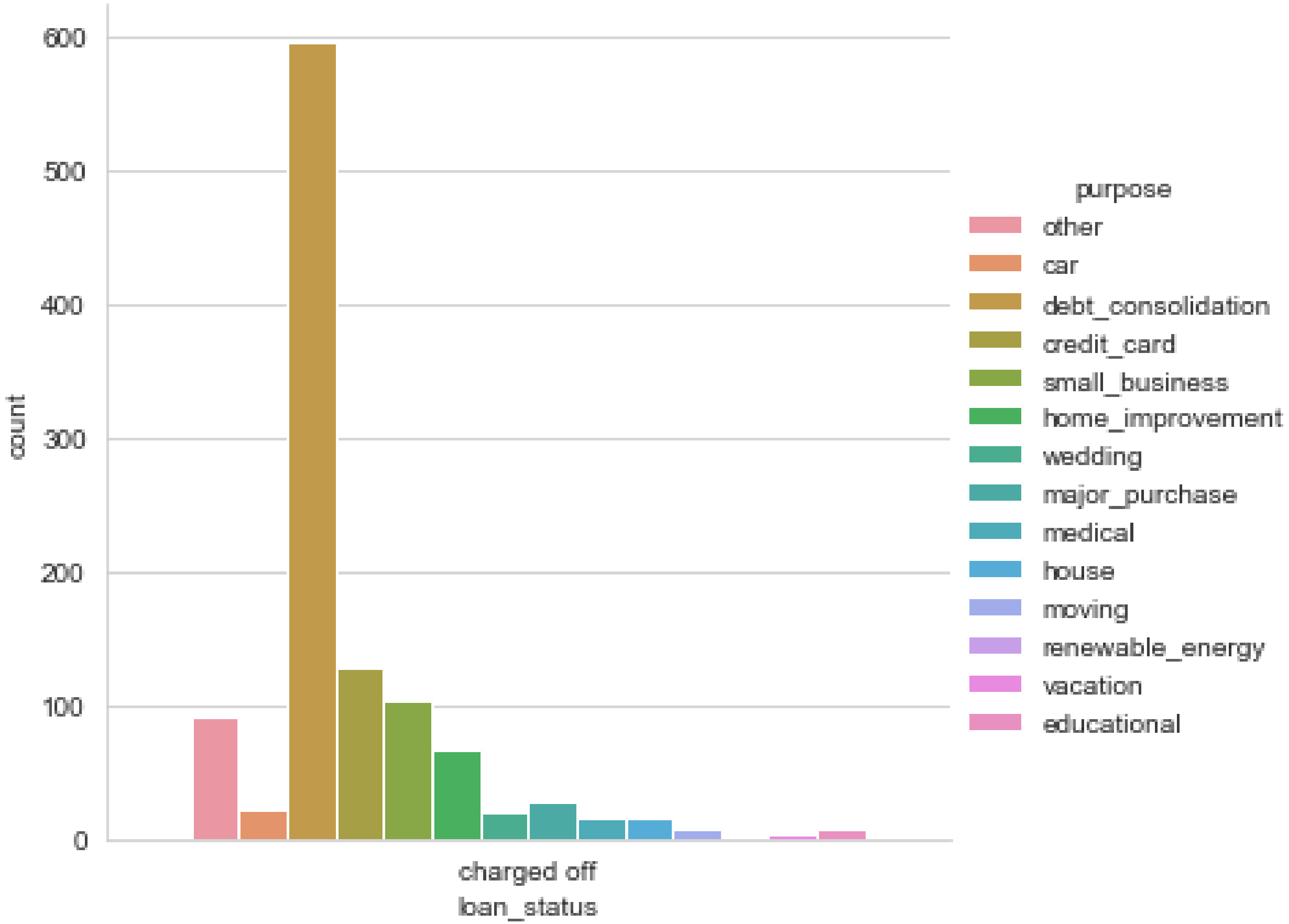
- It is clear that majority of loan(s) are getting Charged-Off with Verification_Status is ***Not Verified*** have a greater chance followed by ***verified***



<Plot 6 – Loan_Status vs Purpose (BiVariate)>

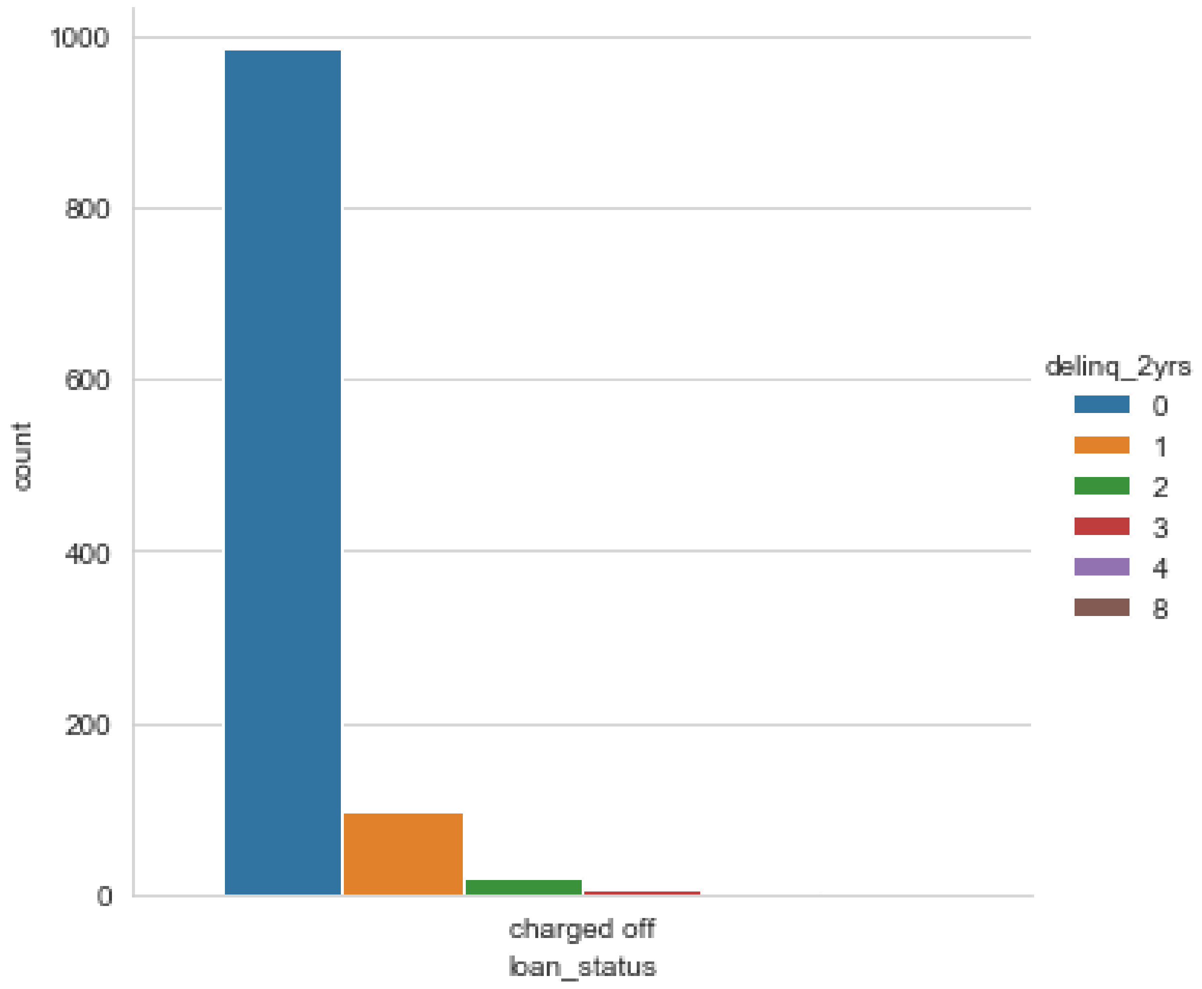
UpGrad

- It is clear that majority of loan(s) are getting Charged-Off with Purpose is ***debt consolidation*** hav a greater chance.



<Plot 6 – Loan_Status vs delinq_in_2 years (BiVariate)>

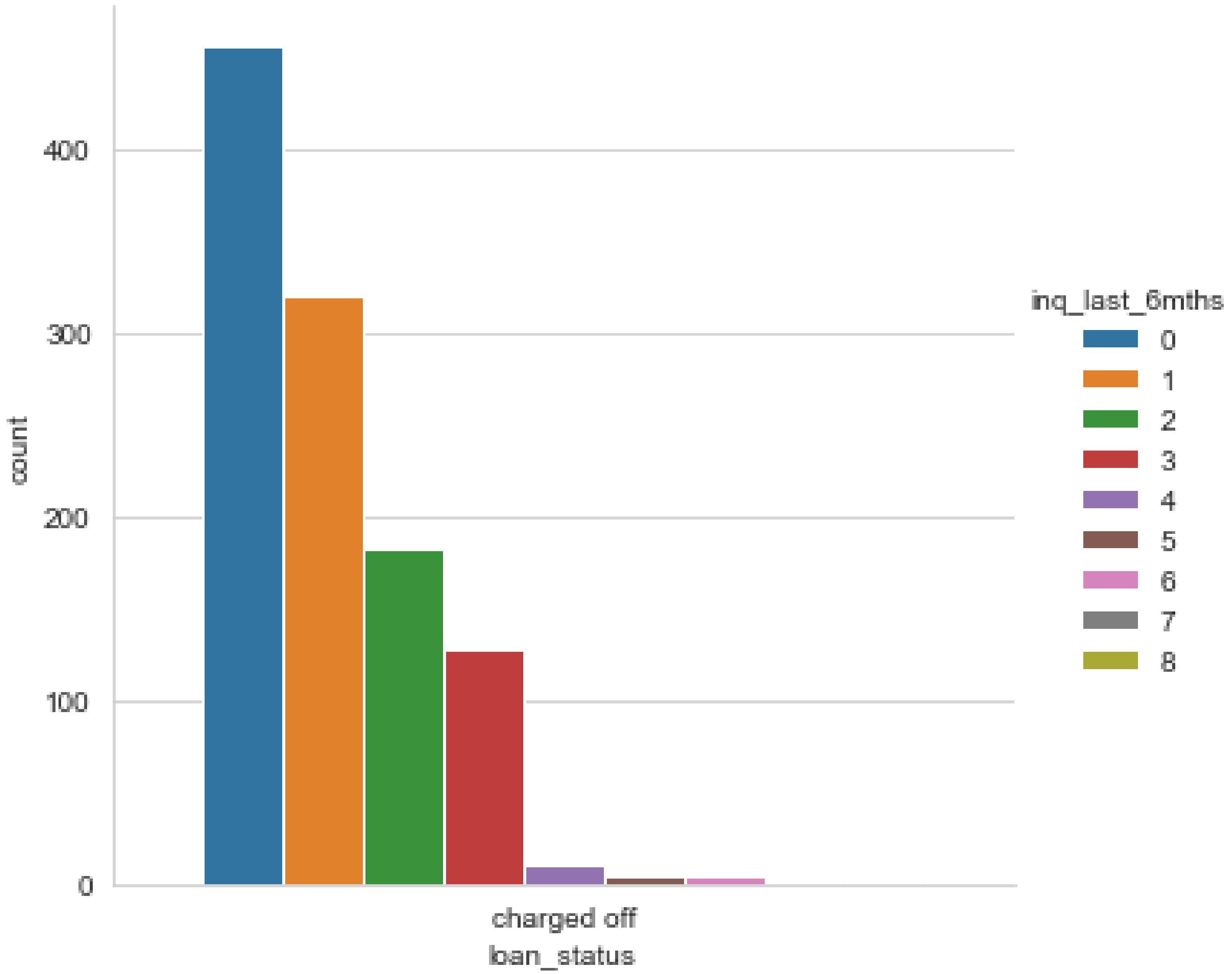
- It is clear that majority of loan(s) are getting Charged-Off with delinq_in_2 years is ***0*** have a greater chance.



<Plot 7 – Loan_Status vs inq_last_6mths (BiVariate)>

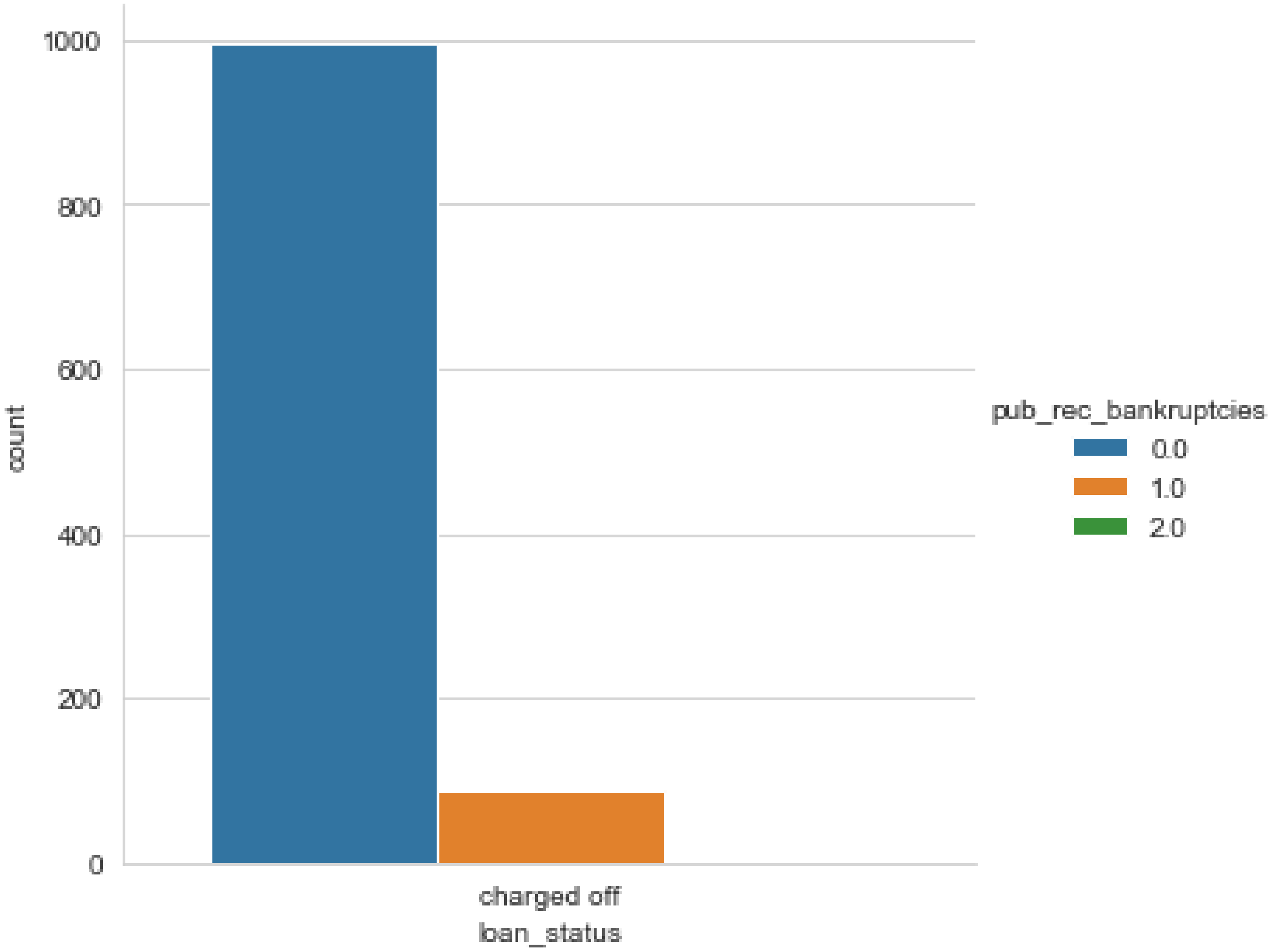
UpGrad

- It is clear that majority of loan(s) are getting Charged-Off with inq_last_6mths is ***0*** have a greate chance.



<Plot 8 – Loan_Status vs pub_rec_bankruptcies (BiVariate)>

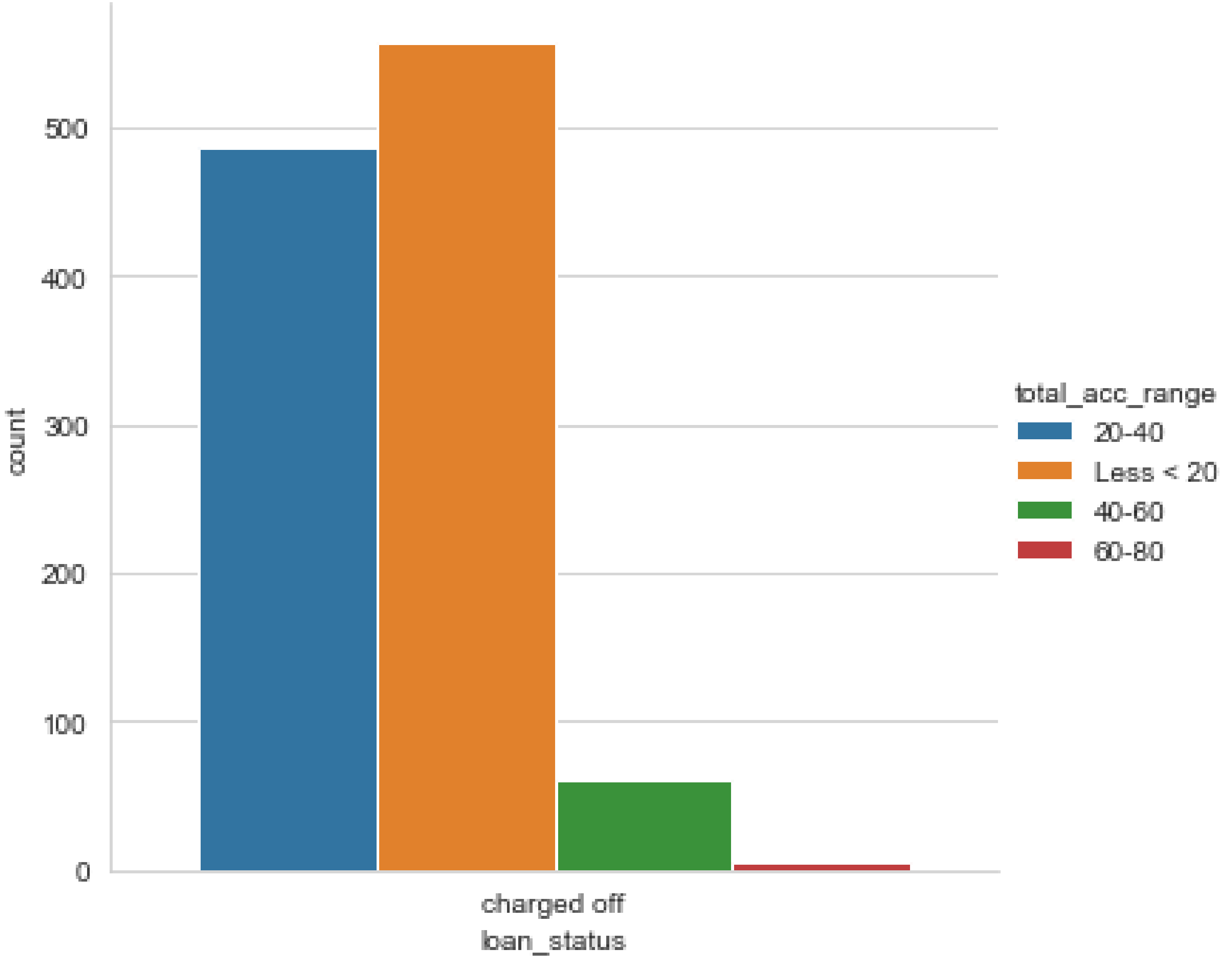
- It is clear that majority of loan(s) are getting Charged-Off with pub_rec_bankruptcies is ***0*** have a greater chance.



<Plot 9 – Loan_Status vs total_acc_range (BiVariate)>

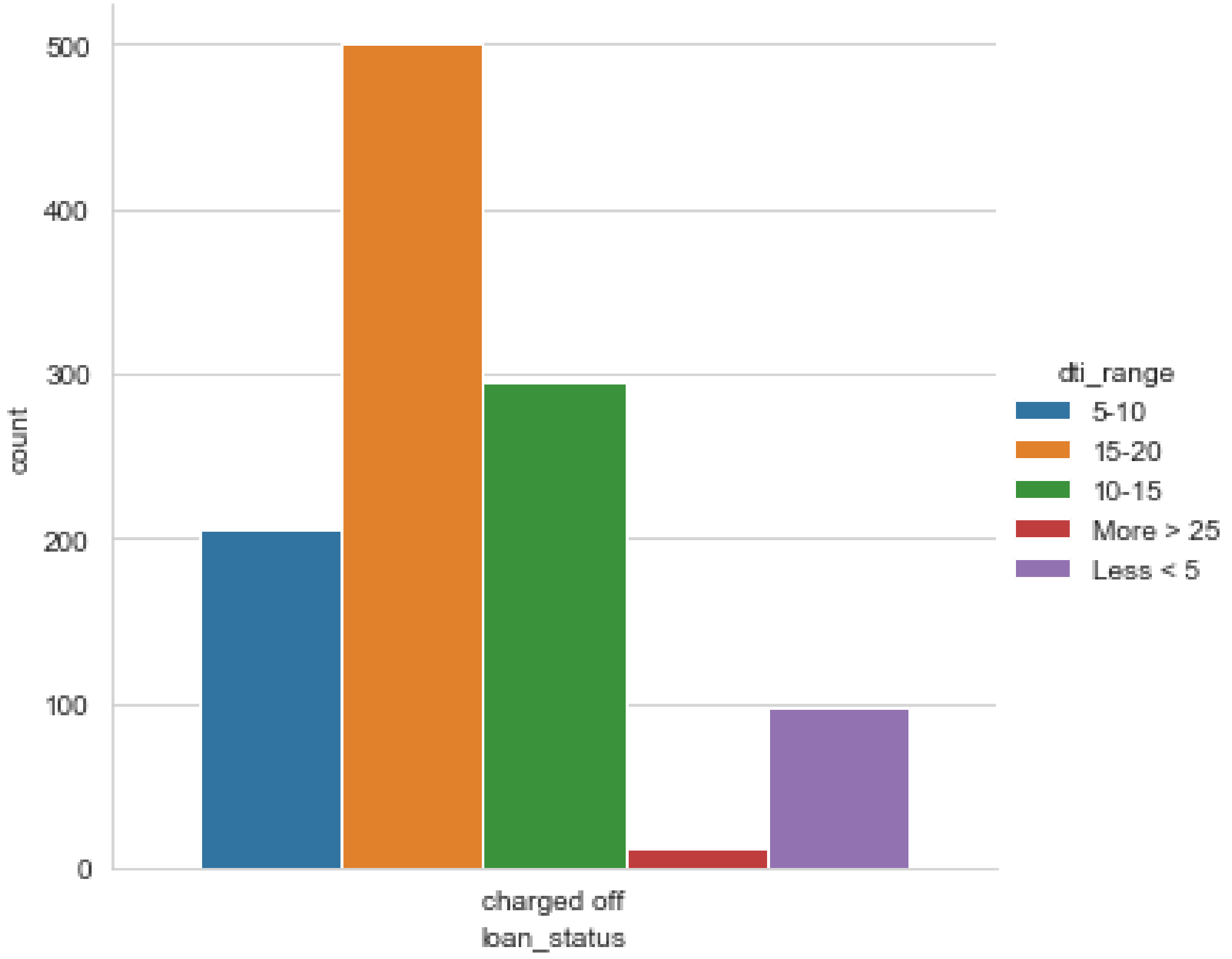
UpGrad

- It is clear that majority of loan(s) are getting Charged-Off with total_acc_range is ***Less <20*** have a greater chance.



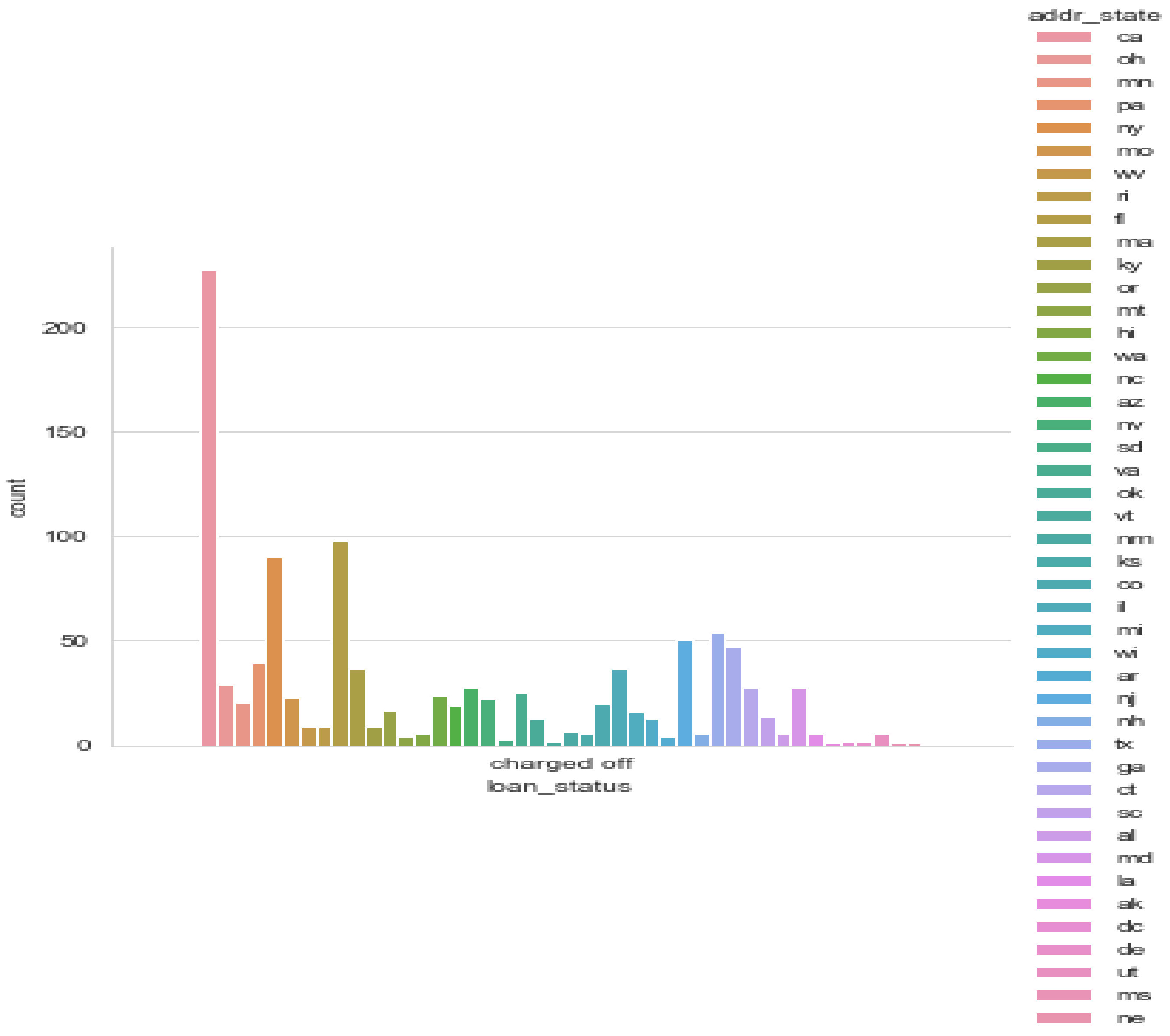
<Plot 10 – Loan_Status vs dti_range (BiVariate)>

- It is clear that majority of loan(s) are getting Charged-Off with dti_range is ***15-20 range*** have a greater chance.



<Plot 10 – Loan_Status vs addr_state (BiVariate)>

- It is clear that majority of loan(s) are getting Charged-Off with addr_state is ***CA*** have a greater chance.



<Recommendations>

- For Extremely high loan amount and extremely high interest rate , Small Business and Debt Consolidation are leading into maximum defaulter.
- For Extremely high loan amount and mid interest rate , Small Business and Debt Consolidation are leading into maximum defaulter
- For Extremely high loan amount and extremely high interest rate , Home Ownership as OTHERS is leading into maximum defaulter.
- For Extremely high loan amount and mid interest rate , Home Ownership RENT is leading into maximum defaulter.

<Conclusions>

- As per as the analysis, we found few deciding factors which ends up in determining the loan defaulter applicant. So, the bank should consider the deciding factors before sanctioning the loan to avoid the credit loss.
- Focus needs to be on reducing the number of loans that can turn into ‘Charged Off’ which automatically results in
- the loans converting to a successfully ‘Fully Paid’ status
- The state of California is where majority of the loans are availed and also defaulted, so there needs to be more attention in California on newer applications while the same ought to be observed in other states
- Thorough verification of the information like the Annual Income quoted by the Customer needs to happen. The total loan amount has to be weighed against the annual income and disbursed
- ‘Debt Consolidation’ as a purpose on the applications needs to verified in terms of the annual income and the total loans already held by the Customer to corroborate the ability to repay
- The term of ’60 months’ i.e. Longer duration of the loan needs to carefully assessed and disbursed to individuals who fairly meet the criteria above
- Also, along with the determination of grade, the loans that are being disbursed at a higher rate of interest have to be screened further in terms of the customer being able to make payments over a longer period consistently