# LENDING CLUB CASE STUDY

- RIDDHIMAN GHOSH ROY

# PROBLEM STATEMENT:

We are provided with the data of 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 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.

The company can take the decision to accept the loan proposal or reject the loan proposal. In the event the company accepts the loan proposal the loan can progress into three states: Fully Paid, Current and Charged Off.

The company incurs loss if the loan is approved but the borrowers default while paying it back. The company also incurs a loss if they don't approve the loan and thus losing the opportunity to gain capital.

***Objective:*** Use Exploratory Data Analysis (EDA) to understand how customer attributes and loan attributes influence the tendency to default.

# Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data. EDA is an important first step in any data analysis.

In this dataset we will perform EDA.

EDA consists of 4-5 parts or steps namely:

1. Understanding Data
2. Data Cleaning
3. Data Shaping and deriving.
4. Data Analysis: Univariate Analysis ,Segmented Univariate Analysis , Correlation Analysis and Bivariate Analysis

+

# Understanding Data:

This is the first step in the analysis. In this step we observe the skeletal structure or the shape of the dataset by obtaining the rows count and columns count of the dataset as well as we take a look at the number of null values , duplicate values and unique values in the dataset.

```
#Number of Rows and Columns in the dataset
print('Number of Rows:',loan_dataset.shape[0])
print('Number of Columns:',loan_dataset.shape[1])
#Number of Missing values in the dataset
print('Number of Missing values in the dataset:',loan_dataset.isnull().sum().sum())
#Number of unique values in the dataset
print('Number of Unique values in the dataset:',loan_dataset.nunique().sum())
#Number of Duplicates in the dataset
print('Number of Duplicate data in dataset:',loan_dataset.duplicated().sum())
```

```
Number of Rows: 39717
Number of Columns: 111
Number of Missing values in the dataset: 2263366
Number of Unique values in the dataset: 416800
Number of Duplicate data in dataset: 0
```

```
#Shape of the dataset before cleaning
print(loan_dataset.shape)
```

```
(39717, 111)
```

```
#Columns present in the dataset
print(loan_dataset.columns)
```

```
Index(['id', 'member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv',
       'term', 'int_rate', 'installment', 'grade', 'sub_grade',
       ...
       'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq',
       'percent_bc_gt_75', 'pub_rec_bankruptcies', 'tax_liens',
       'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit',
       'total_il_high_credit_limit'],
      dtype='object', length=111)
```

# Data Cleaning:

Data Cleaning is the process of getting rid of impurities and redundancies present in the data , which might cause errors in analysis process in later stages. In data cleaning we get rid of any null values , duplicate values and any variables which might not contribute to the data analysis.

In the picture below we can see the removal of the unnecessary columns from the dataset:

```
#Removing url and zip_code as it is not necessary in EDA
loan_dataset=loan_dataset.drop(['url'],axis=1)
#Removing member_id as it is not necessary for EDA
loan_dataset=loan_dataset.drop(['member_id'],axis=1)
#Removing desc,emp_title,title as it is not necessary in EDA
loan_dataset=loan_dataset.drop(['desc','emp_title','title'],axis=1)
#Removing zip_code as it is not necessary in EDA
loan_dataset=loan_dataset.drop(['zip_code'],axis=1)
#Removing funded_amnt_inv as it is not necessary in EDA
loan_dataset=loan_dataset.drop(['funded_amnt_inv'],axis=1)
```

```
#Removing Customer behavourial variables which dont factor in for loan approval process
behavourial=['delinq_2yrs','earliest_cr_line','last_pymnt_amnt','inq_last_6mths','open_acc','pub_rec','revol_bal','revol_util','t
             'total_pymnt_inv','total_rec_prncp','total_rec_int','total_rec_late_fee','recoveries','collection_recovery_fee','app
             'last_credit_pull_d']
loan_dataset=loan_dataset.drop(behavourial,axis=1)
```

# Data Cleaning(Eliminating Null Values):

We need to identify which columns contain the most amount of Null values and based on the numbers we can either remove the records entirely or we can impute the values. By checking the dataset for Null values we saw that some variables contain a 100% Null values thus we subsequently removed the columns which have more than 50% Null values as shown below in (1) along with the null value percentage after removing the columns as shown in (2):

```
#Checking percentage of null values present in each column
print((loan_dataset.isnull().sum()/loan_dataset.shape[0]*100).round(2).sort_values(ascending=True))
```

```
id                            0.0
policy_code                   0.0
acc_now_delinq                0.0
dti                           0.0
addr_state                    0.0
                             ...
avg_cur_bal                 100.0
bc_open_to_buy              100.0
bc_util                     100.0
mo_sin_old_rev_tl_op        100.0
total_il_high_credit_limit  100.0
Length: 84, dtype: float64
```

```
#Removing the columns having more than 50% of null values
#loan_dataset=loan_dataset.dropna(axis=1,how="all")
loan_dataset=loan_dataset.loc[ :,loan_dataset.isnull().sum()/loan_dataset.shape[0]*100<50]
```

(1)

```
#Checking the percentage of null values present after removing them
print((loan_dataset.isnull().sum()/loan_dataset.shape[0]*100).round(2).sort_values(ascending=True))
```

```
id                            0.00
delinq_amnt                   0.00
acc_now_delinq                0.00
policy_code                   0.00
total_pymnt                   0.00
initial_list_status           0.00
dti                           0.00
addr_state                    0.00
purpose                       0.00
pymnt_plan                    0.00
issue_d                       0.00
loan_status                   0.00
annual_inc                    0.00
home_ownership                0.00
sub_grade                     0.00
grade                         0.00
installment                   0.00
int_rate                      0.00
term                          0.00
funded_amnt                   0.00
loan_amnt                     0.00
verification_status           0.00
tax_liens                     0.10
collections_12_mths_ex_med    0.14
chargeoff_within_12_mths      0.14
pub_rec_bankruptcies          1.75
emp_length                    2.71
dtype: float64
```

(2)

# Data Cleaning(Eliminating Unique Values):

Unique values in this dataset which are not a lot in count , won't really factor in while doing analysis of the dataset. So we removed the columns which have only 1 unique record as shown below:

```
print(loan_dataset.nunique().sort_values(ascending=True))
```

```
tax_liens                     1
delinq_amnt                   1
chargeoff_within_12_mths      1
acc_now_delinq                1
policy_code                   1
collections_12_mths_ex_med    1
initial_list_status           1
pymnt_plan                    1
term                          2
pub_rec_bankruptcies          3
verification_status           3
loan_status                   3
home_ownership                5
grade                         7
emp_length                   11
purpose                      14
sub_grade                    35
addr_state                   50
issue_d                      55
int_rate                    371
loan_amnt                   885
funded_amnt                1041
dti                        2868
annual_inc                 5318
installment               15383
total_pymnt               37850
id                        39717
dtype: int64
```

(1)

```
#Removing the variables having only 1 unique entry as they will not be that significant in EDA
for column in loan_dataset.columns:
    if loan_dataset[column].nunique(dropna=True)==1:
        print(column)
        loan_dataset=loan_dataset.drop(column,axis=1)
```

```
pymnt_plan
initial_list_status
collections_12_mths_ex_med
policy_code
acc_now_delinq
chargeoff_within_12_mths
delinq_amnt
tax_liens
```

(2)

# Data Cleaning(Eliminating Duplicate values)

This dataset doesn't contain any duplicate values. As we can see below:

```
#Checking duplicate rows present in the dataset
print(loan_dataset.duplicated().sum())
```

```
0
```

# Data Cleaning(Dropping the rows based on loan_status):

Since we need to analyse the likelihood of defaulting from loan , therefore we don't need the records with loan_status as 'Current'. Thats why we filtered out those rows as shown below:

```
#Filtering out the entries having loan_status as 'Current' as we only need to analyse the entries having loan_status as 'Fully P
loan_dataset=loan_dataset[loan_dataset['loan_status']!='Current']
```

# Data Cleaning(Imputing the remaining missing values):

As we can see for the column emp_length and pub_rec_bankruptcies we need to get rid of the missing values by imputing/dropping the rows. Thus we dropped the rows of pub_rec_bankruptcies and imputed the rows of emp_length as shown below:

(1)

```
#Checking the missing values still present in the dataset
print(loan_dataset.isnull().sum().sort_values(ascending=False))
```

```
emp_length              1075
pub_rec_bankruptcies     697
annual_inc                 0
total_pymnt                0
dti                        0
addr_state                 0
purpose                    0
loan_status                0
issue_d                    0
verification_status        0
id                         0
loan_amnt                  0
sub_grade                  0
grade                      0
installment                0
int_rate                   0
term                       0
funded_amnt                0
home_ownership             0
dtype: int64
```

(2)

```
#Removing the null values from emp_length as fixing them could lead to data loss and fixing the pub_rec_bankruptcies by fillin
loan_dataset=loan_dataset.dropna(subset=['emp_length'])
loan_dataset.pub_rec_bankruptcies.fillna(0,inplace=True)
```

# Data Shaping and Deriving:

Now that we cleaned the dataset from all the null values,duplicate values,unique values and unnecessary variables, our next step is to shape the data in a uniform manner and derive new columns from the existing columns so that we can have an in depth understanding of the data with new factors.

First we shaped the data by checking the datatypes of all the columns and then turning the columns into proper datatype and also shaping all of the data of those columns into that datatype as shown below with an example of the column 'term'. For this column we converted the datatype from string to int and trimmed the data to be a proper int64 type of data:

```
#Checking the datatypes of the variables
loan_dataset.dtypes
```

```
id                       int64
loan_amnt                int64
funded_amnt              int64
term                    object
int_rate                object
installment            float64
grade                   object
sub_grade               object
emp_length              object
home_ownership          object
annual_inc             float64
verification_status     object
issue_d                 object
loan_status             object
purpose                 object
addr_state              object
dti                    float64
total_pymnt            float64
pub_rec_bankruptcies   float64
dtype: object
```

```
#Checking the values of term variable to identify the datatype
loan_dataset['term'].value_counts()
```

```
term
 36 months    28287
 60 months     9257
Name: count, dtype: int64
```

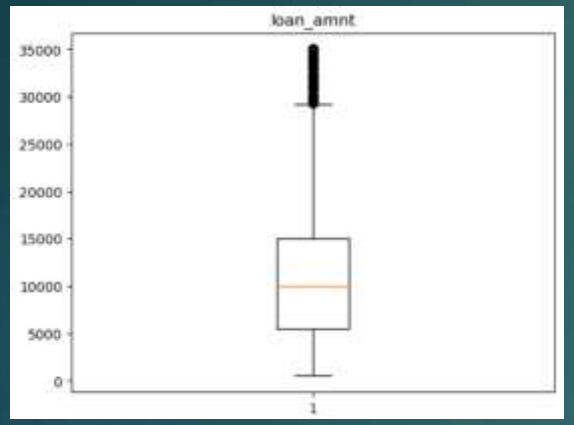
```
#Converting term from string type to int type
loan_dataset.term=loan_dataset.term.apply(lambda x: int(x.replace(' months','').astype(int)
```

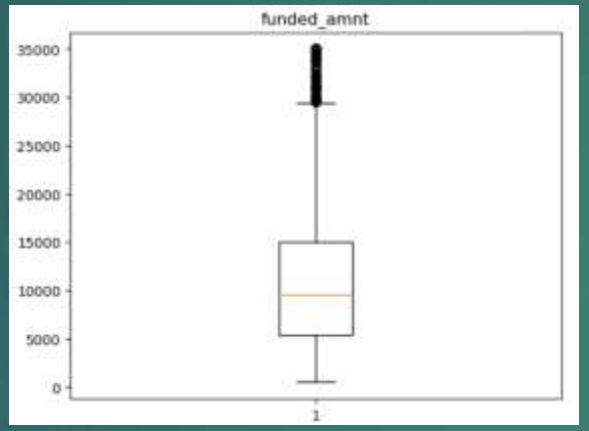
```
#Checking the values of term variable after conversion and modification
loan_dataset['term'].value_counts()
```

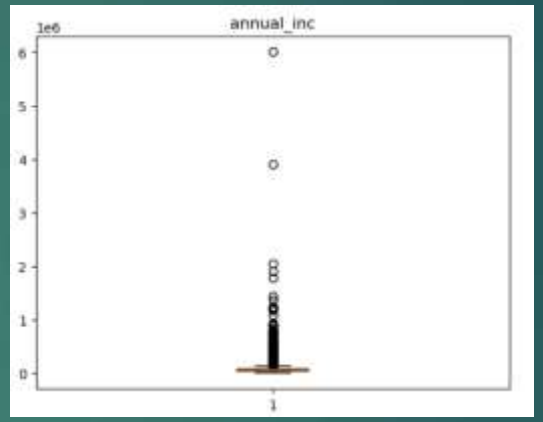
```
term
36    28287
60     9257
Name: count, dtype: int64
```

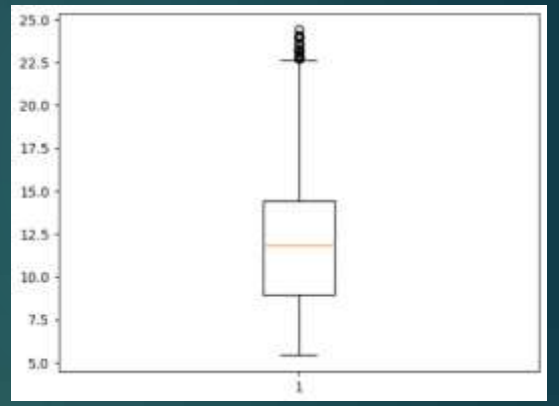
# Data Shaping(Removing Outliers):

Before heading into analysis , its also imperative that we remove all the outliers present in the dataset(if any). Thus we charted the boxplots of all the numerical columns to check for outliers as we can see below. From the boxplots we inferred that there were outliers present for the columns : loan_amnt,funded_amnt,annual_inc and int_rate.



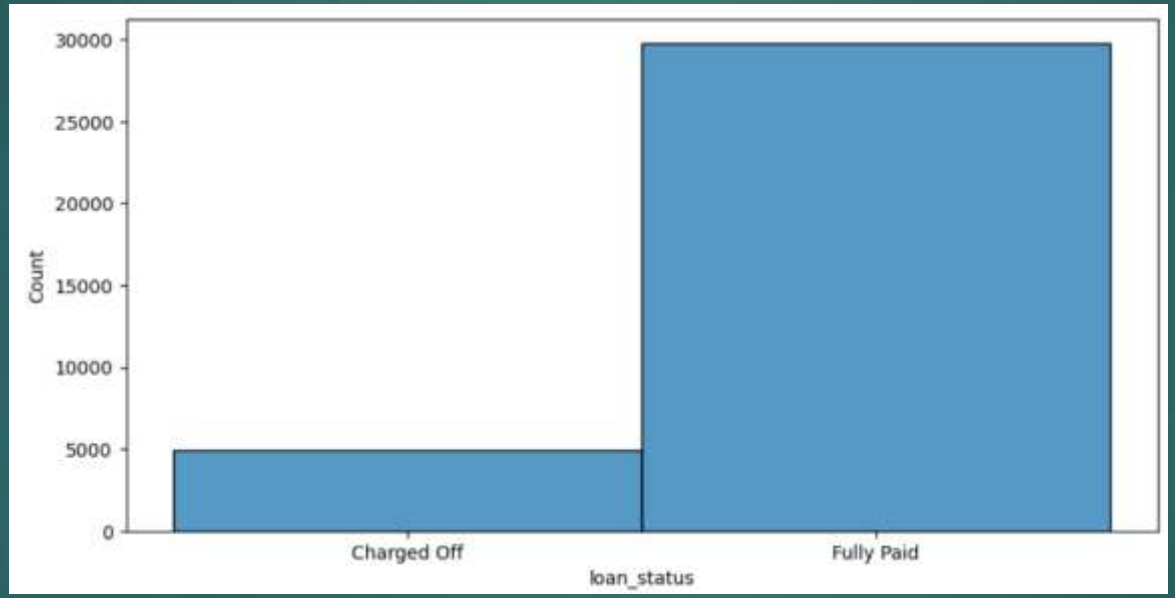




Thus we fixed the Outliers with the help of IQR as shown below:

```
#Removing the outliers
set=['loan_amnt','funded_amnt','int_rate','annual_inc']
def outliers(data,columns,threshold):
    for col in columns:
        Q1=data[col].quantile(0.25)
        Q3=data[col].quantile(0.75)
        IQR=Q3-Q1
        lower_bound=Q1-threshold*IQR
        upper_bound=Q3+threshold*IQR
        data=data[(data[col]>=lower_bound)&(data[col]<=upper_bound)]
    return data
loan_dataset=outliers(loan_dataset,set,1.5)
```

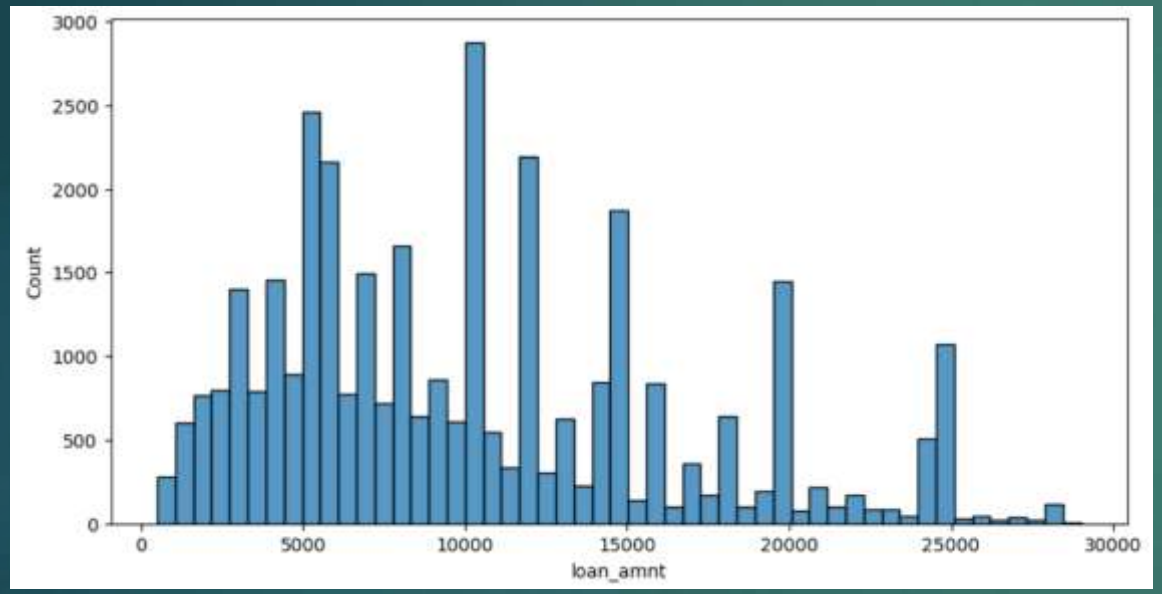
# Univariate Analysis:

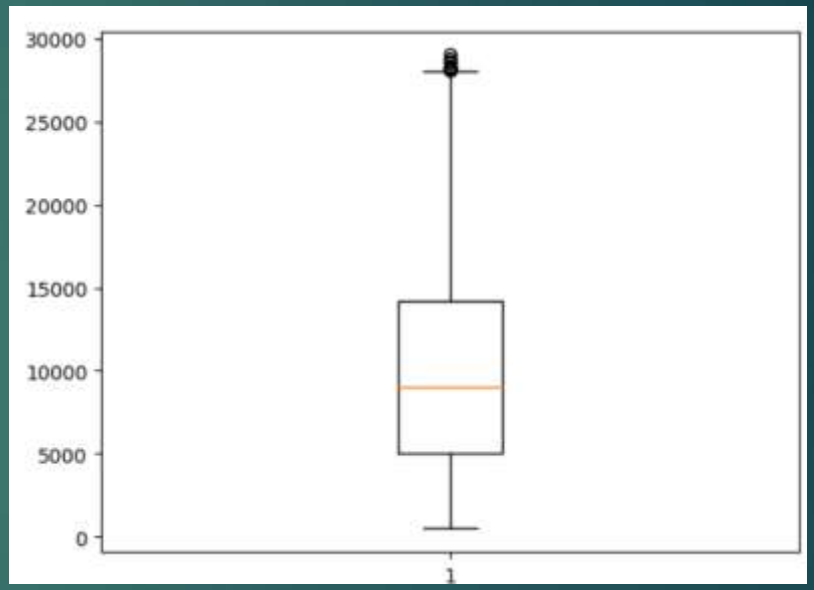
**Analysis of Loan Status:** The analysis of loan status showed that the number of people who have defaulted is way less compared to the people who have fully paid their loan back as we can see below:



# Analysis of Loan Amount:

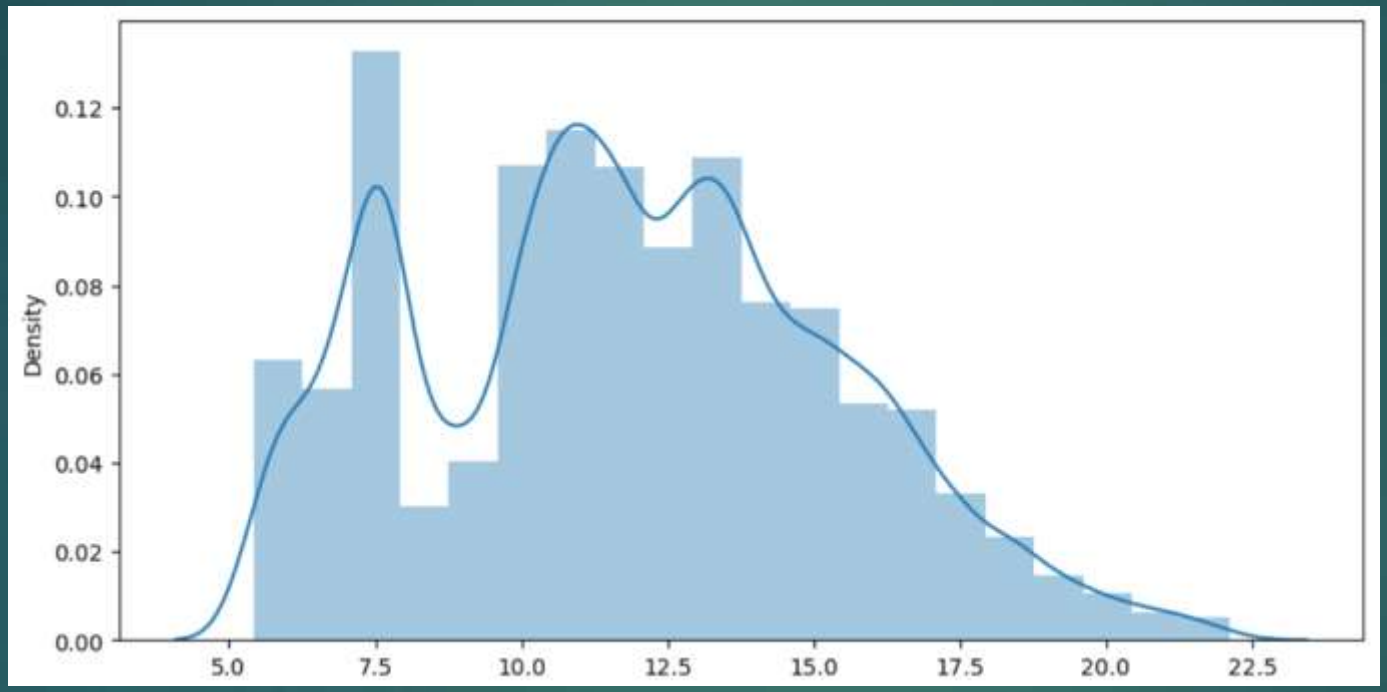
The analysis of loan amount showed that the lowest amount anybody has borrowed is 500 and the highest is 29000. It also inferred that the loan amount of majority of the loans is in the range of 5000-14000 as we can see below:



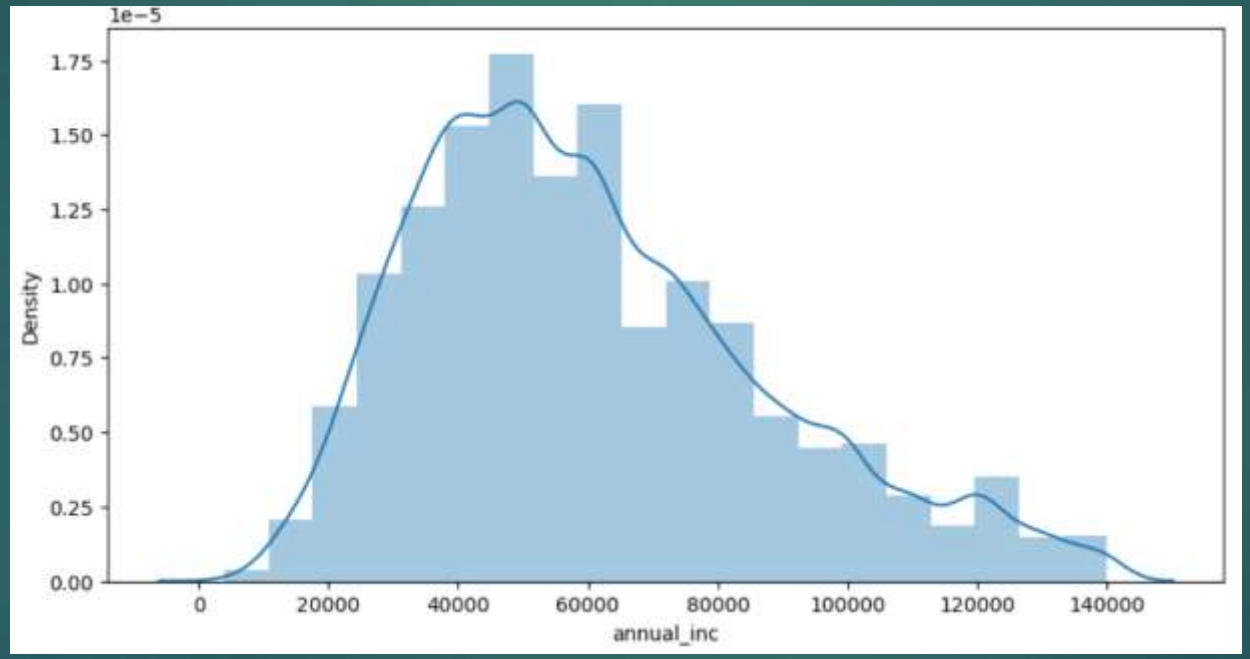
# Analysis of Interest Rate:

The analysis of interest rate showed that the distribution of interest rates from 5-10 is less dense than 10-15 which infers that majority of the loans offer 10-15 % interest rate.



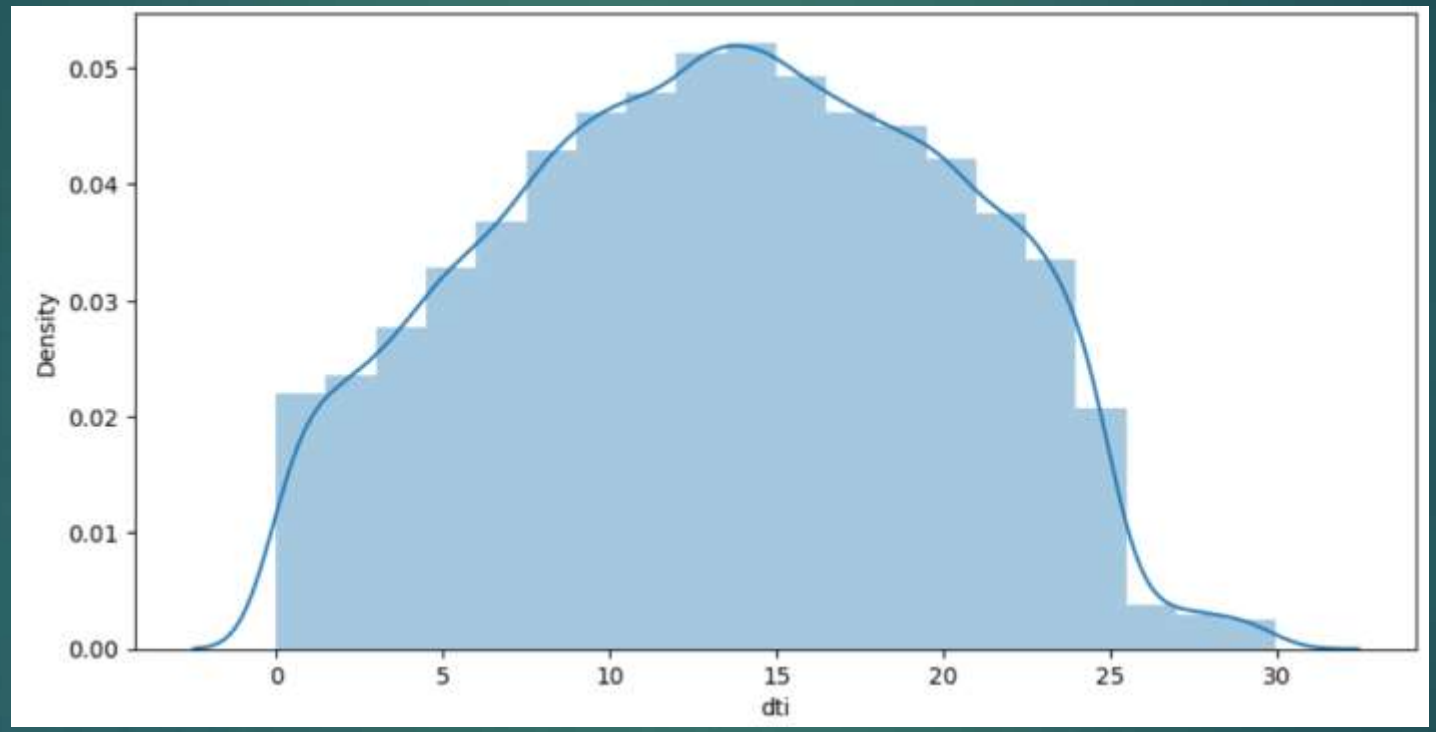
# Analysis of Annual Income:

The analysis of annual income showed that most of the borrowers have an annual income within the range of 40000-60000 and the frequency of people who have borrowed money and have an annual income over 70000 is very low. This suggests that most people who have high annual income have a low tendency to borrow money as shown below:



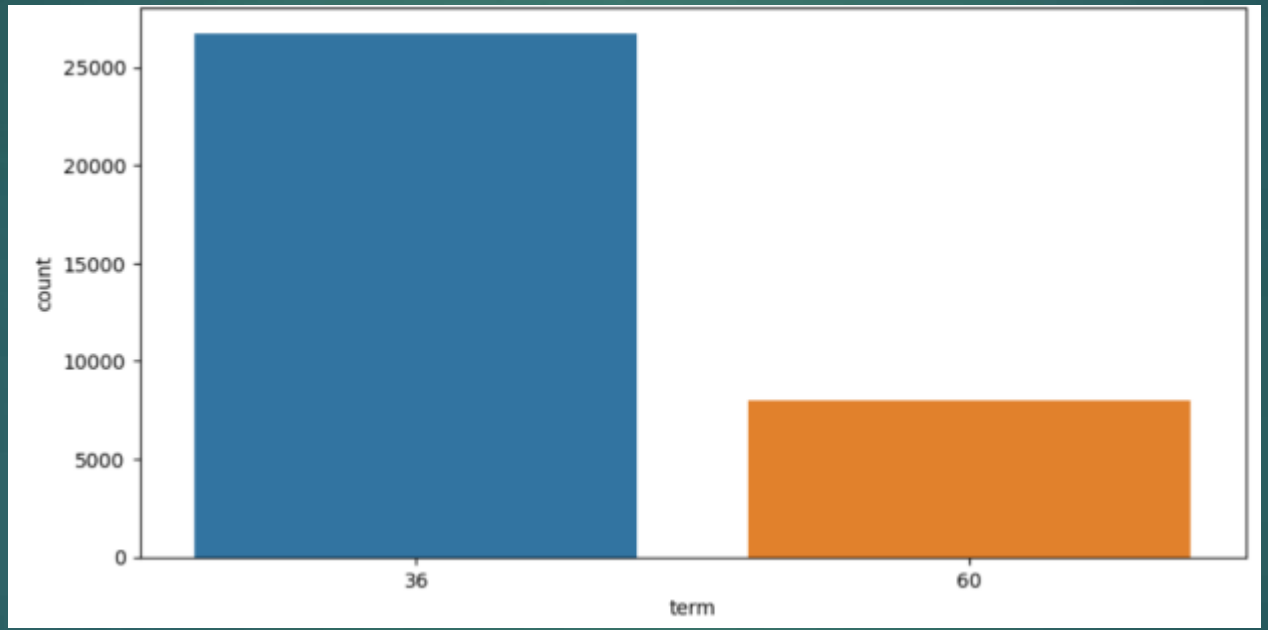
# Analysis of DTI:

The analysis of DTI showed that most borrowers have a DTI within the range of 10-15. This suggests that most borrowers have a moderate DTI ratio as shown below:



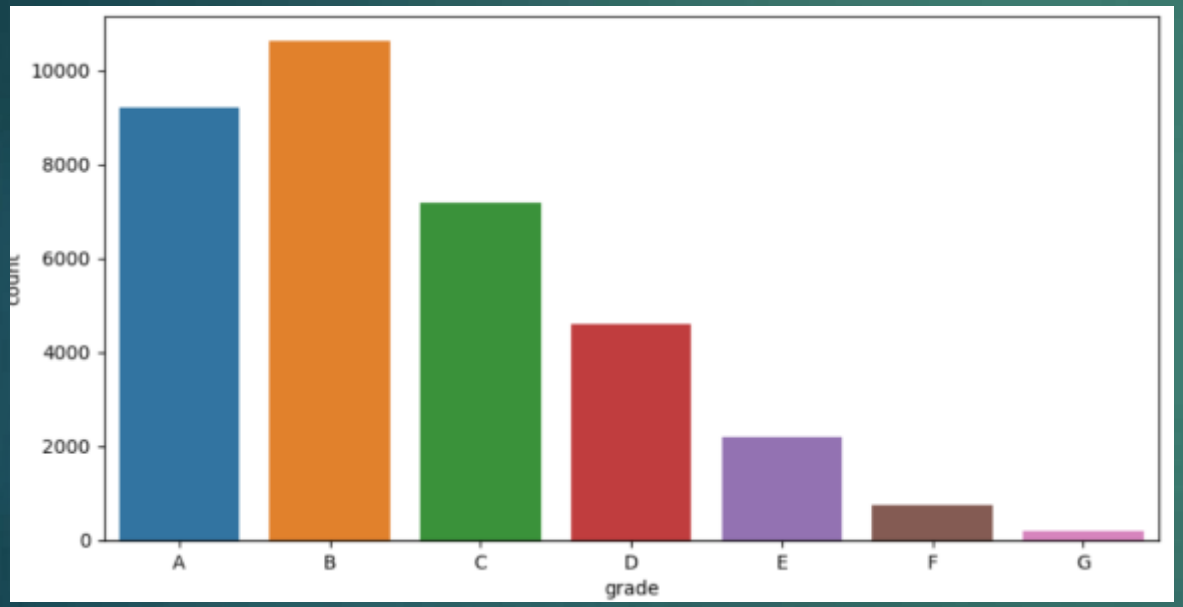
# Analysis of loan term:

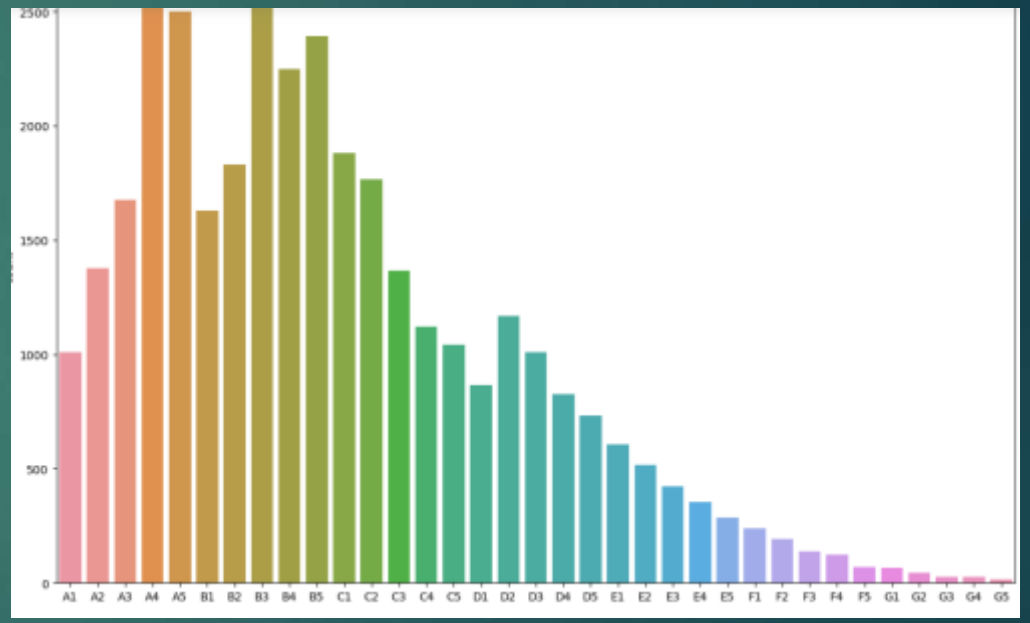
The analysis of loan term showed that majority of loans have a 36 month term. This suggests that most loans which were borrowed, the borrowers opted for the 36 month plan as shown below:



# Analysis of Grade and Subgrade:

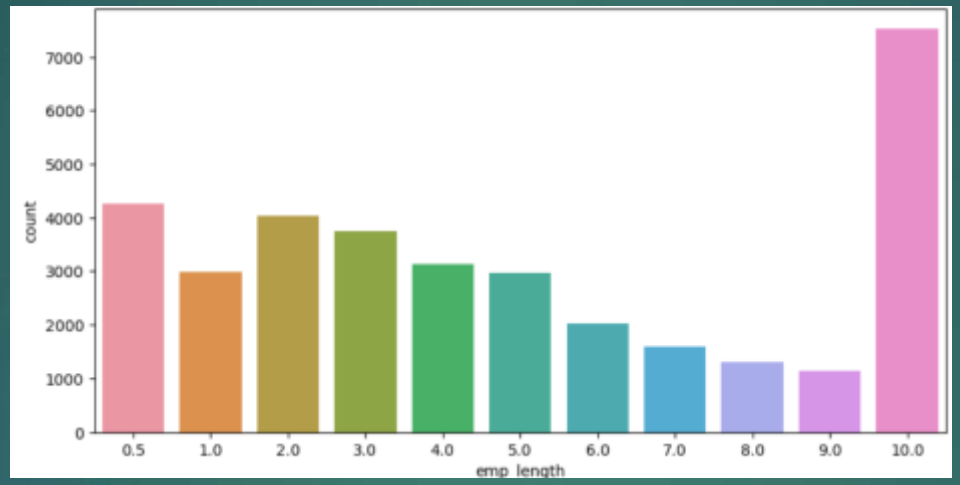
The analysis of grade and subgrade showed that majority of the loans which were borrowed were in the range of grade A-B. Furthermore it showed that majority of the loans were in the subgrade A4,A5 and B3-B5. This suggests that majority of the loans were low risk loans falling in the grades of A4-B5 as shown below:



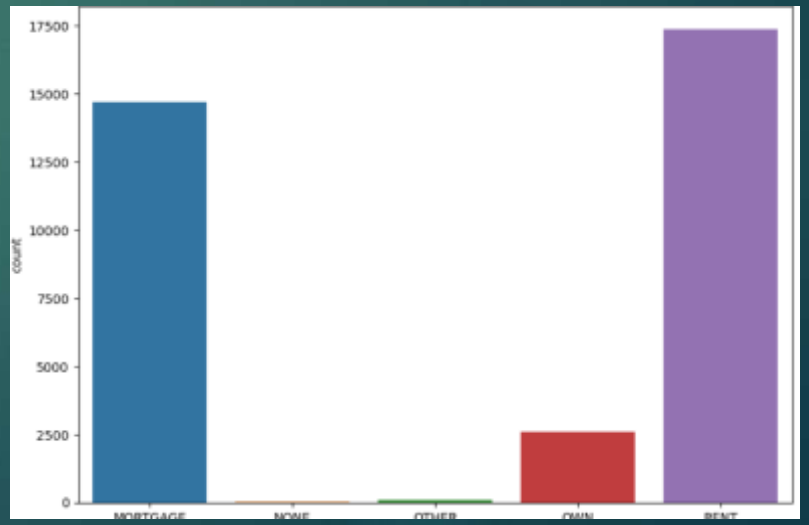
## Analysis of Employment Length:

The analysis of employment length showed that majority of the borrowers have a work experience of over 10 years as shown below:




## Analysis of Home Ownership:

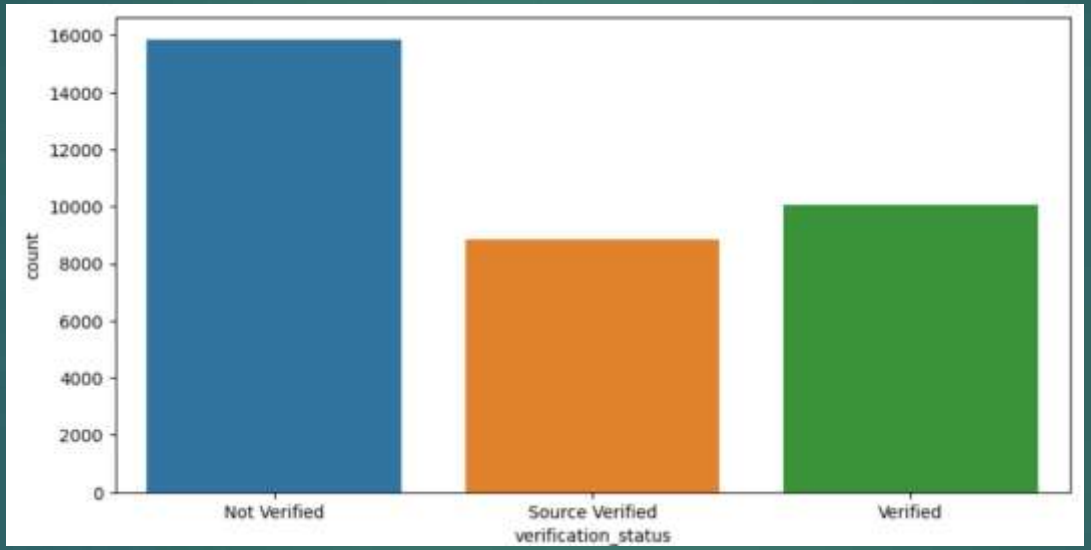
The analysis of home Ownership column showed that majority of the borrowers live in rented properties followed by borrowers who are on mortgage.This suggests that people who have borrowed money, majority of them don't own a home as shown below:



# Analysis of Verification Status:

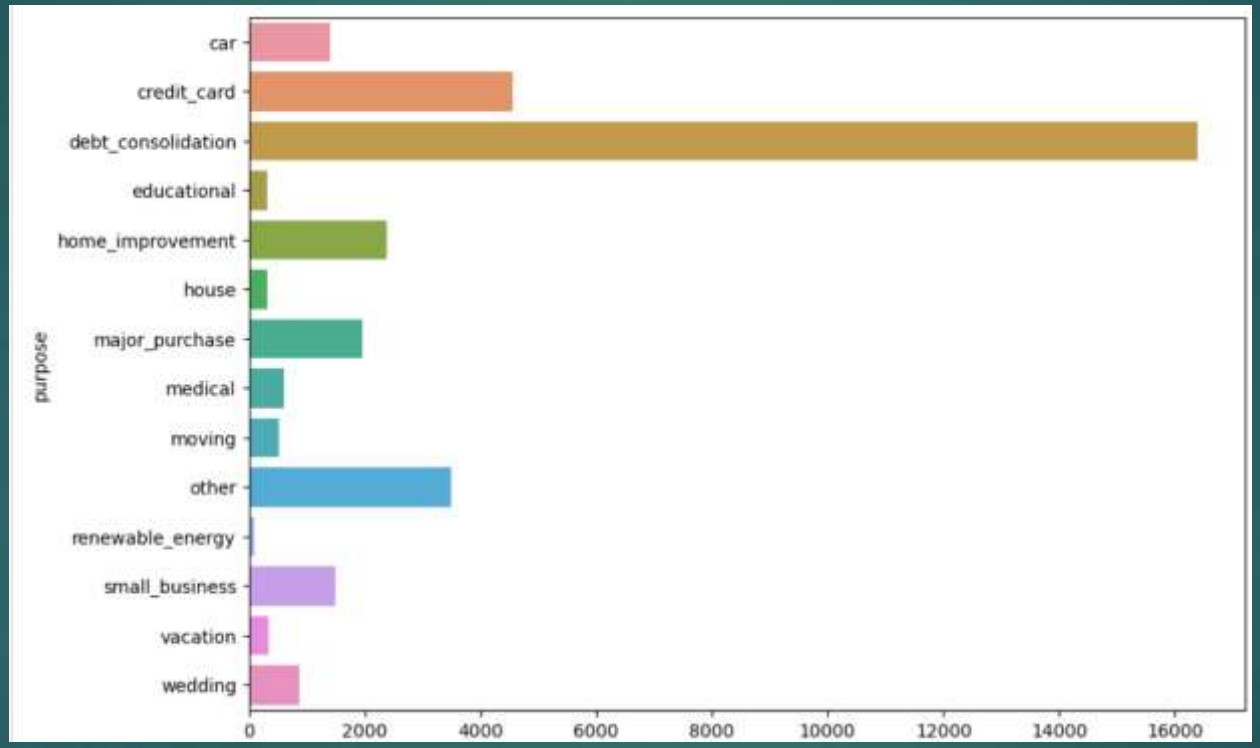
The analysis of verification status column showed that majority of the loans given out were not verified.

This suggests that majority of the loans are given out without doing any source or background verification as shown below:



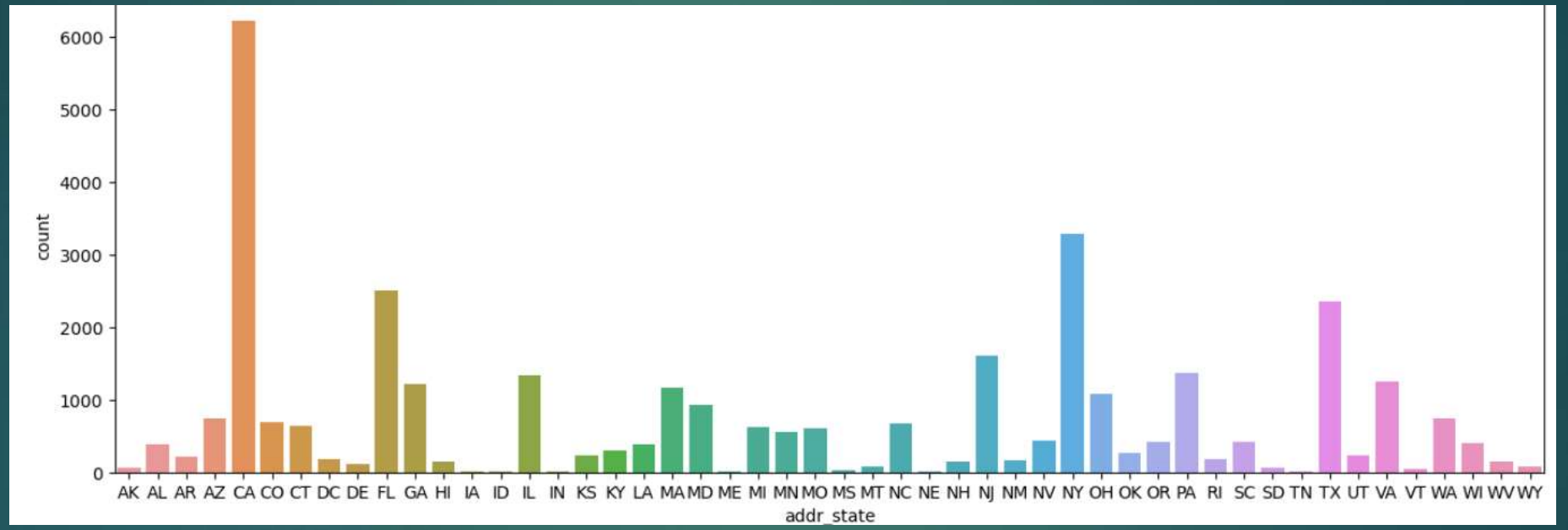
# Analysis of Purpose of Loan:

The analysis of purpose showed that the majority of the loans are being borrowed for the purpose of debt consolidation followed by credit card debt as shown below:



# Analysis of Address State:

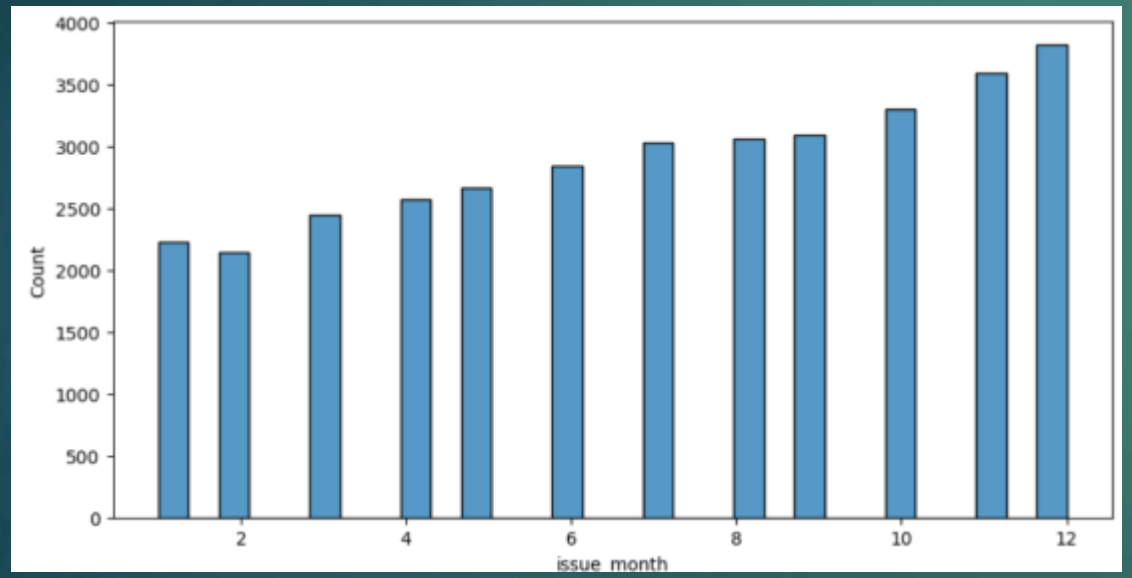
The analysis of address state showed that majority of the loans are being borrowed by the residents of California followed by New York, Florida ,Texas as shown below:

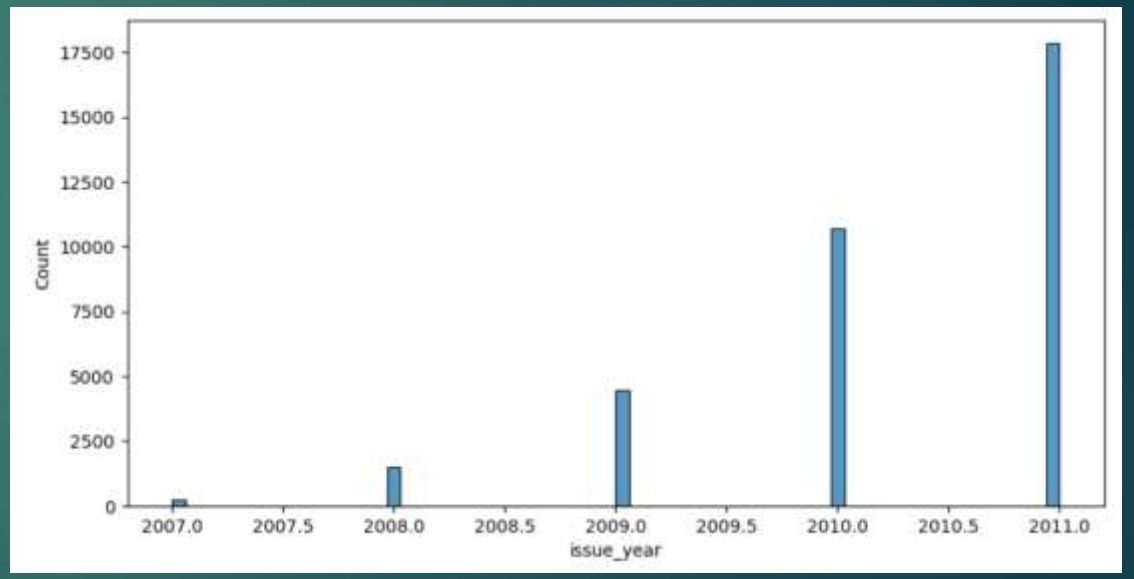


# Analysis of Issue Year and Issue Month:

The analysis of issue year and month showed that the majority of the loans are being approved towards the last quarter of the year and the trend of loan approval is increasing with each quarter in a year, also the amount of loans being provided is increasing in an exponential trend each year since 2007 to 2011.

All these observations suggests that number of loans being provided and number of loans being borrowed is exponentially increasing each year and each month with a large concentration towards the last quarter of each year as shown below:

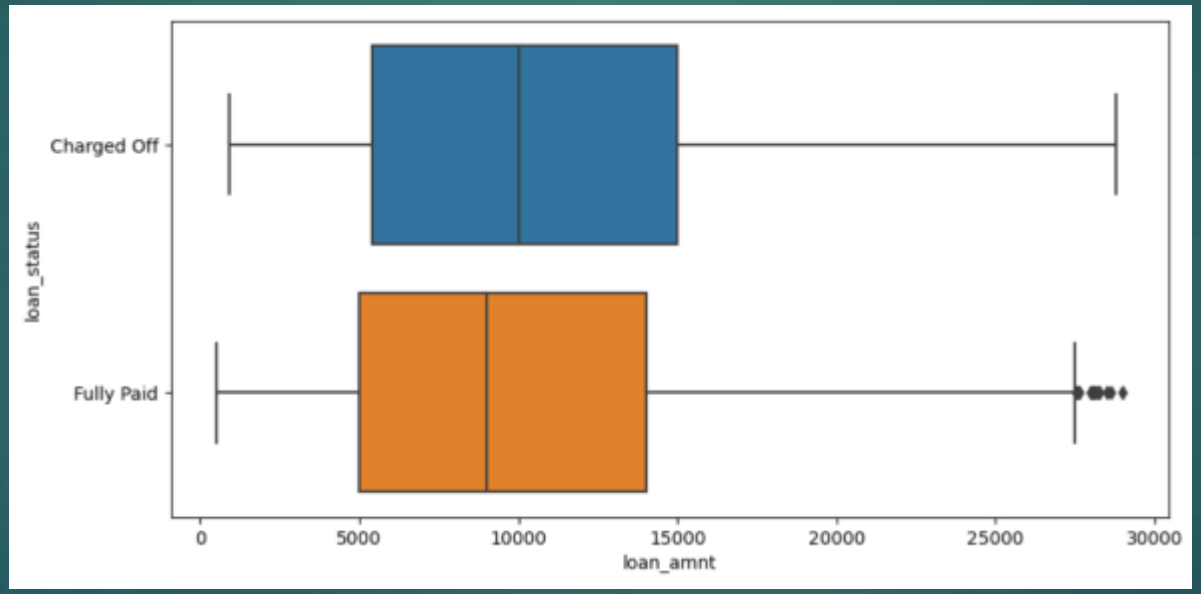


# Segmented Univariate Analysis:

**Analysis of Loan Amount based on Loan Status:** The analysis of loan amount based on loan status showed that for lesser loan amounts, i.e 5000-10000 , there is no significant change in the number of defaulters and full payers,but as the loan amount increases the number of defaulters increases as well.

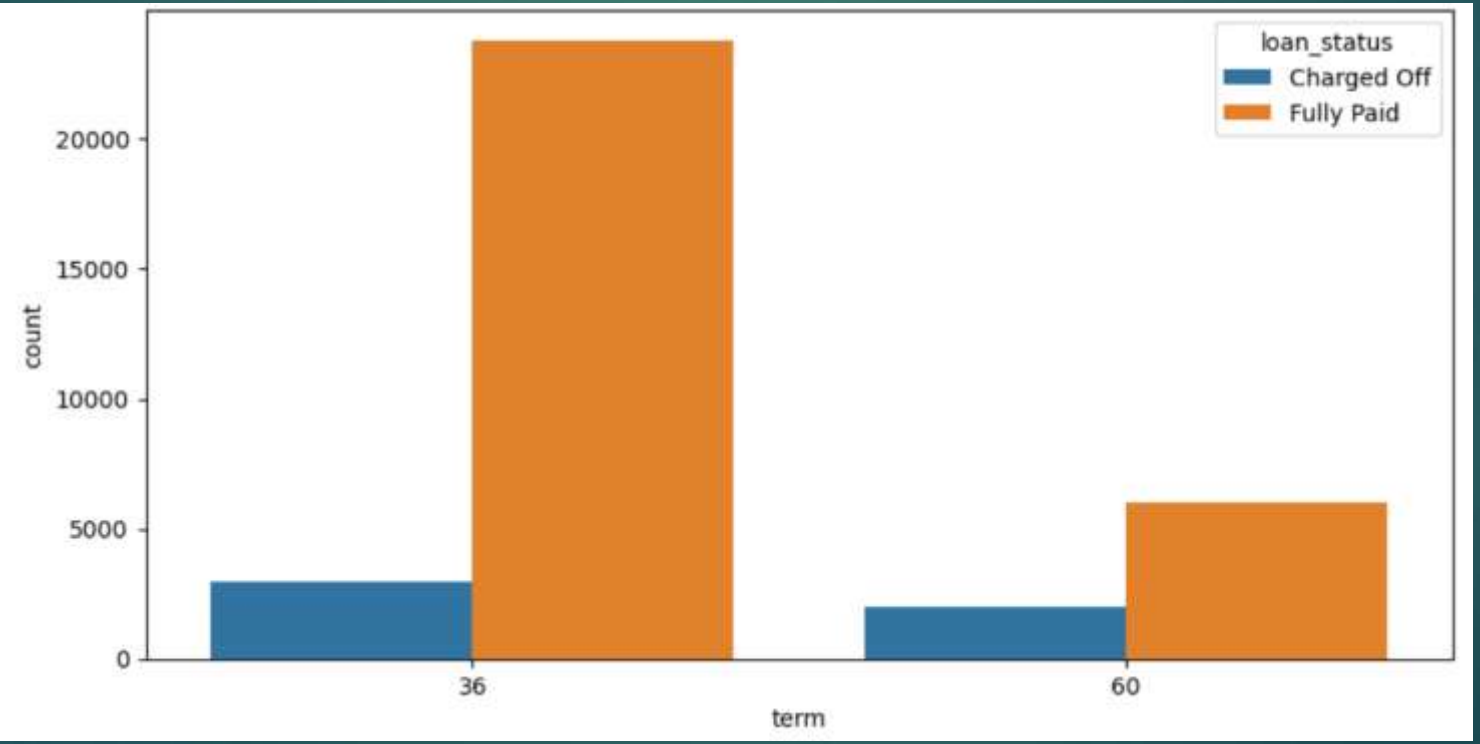
This suggests that the people who have borrowed loans of high amount have a high chance of defaulting as shown below:



# Analysis of Loan Term based on Loan Status:

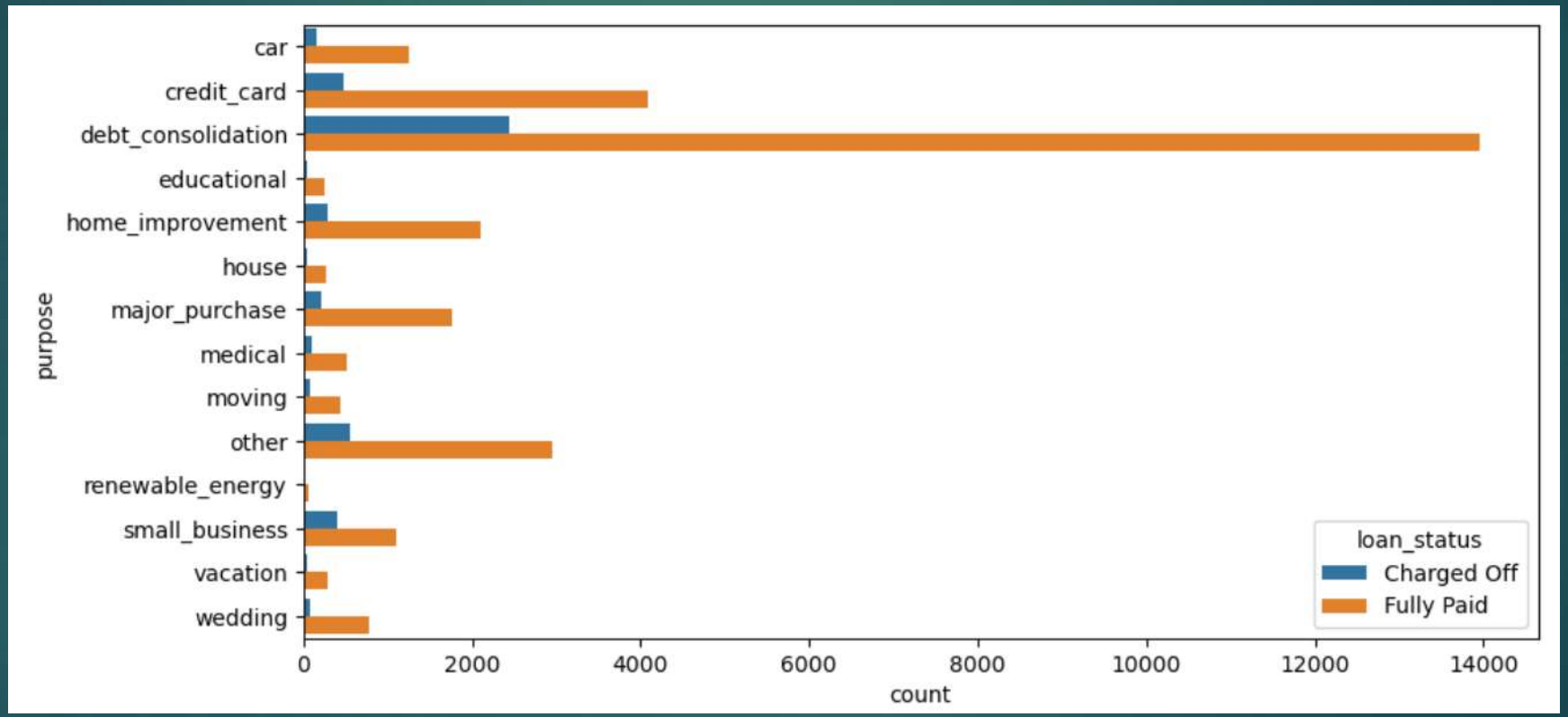
The analysis of loan term based on loan status showed that for loans of term period 36 months, the likelihood of getting paid in full is higher than the loans of term period 60 months.

This suggests that the loans which have a term period of 60 months have a high chance of getting defaulted as shown below:



# Analysis of Purpose based on Loan Status:

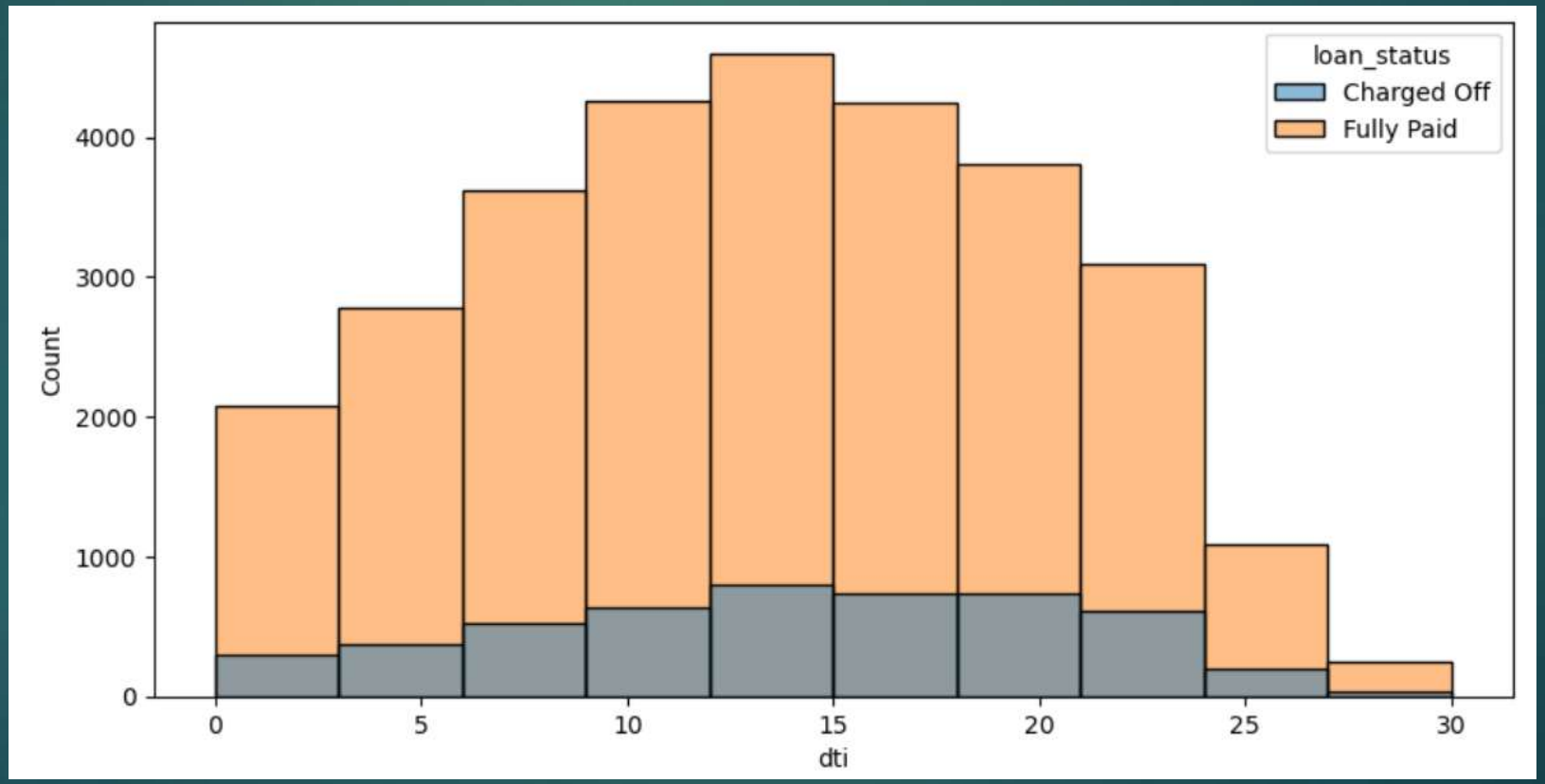
The analysis of purpose based on loan status showed that the most used purpose for borrowing money is debt consolidation and thus it is also the category for which the loans are most defaulted and are fully paid as shown below:



# Analysis of DTI based on Loan Status:

The analysis of DTI based on loan status showed that the number of loans being repaid in full and defaulted increases with increasing DTI upto 15, but starts decreasing for DTI's higher than 15.

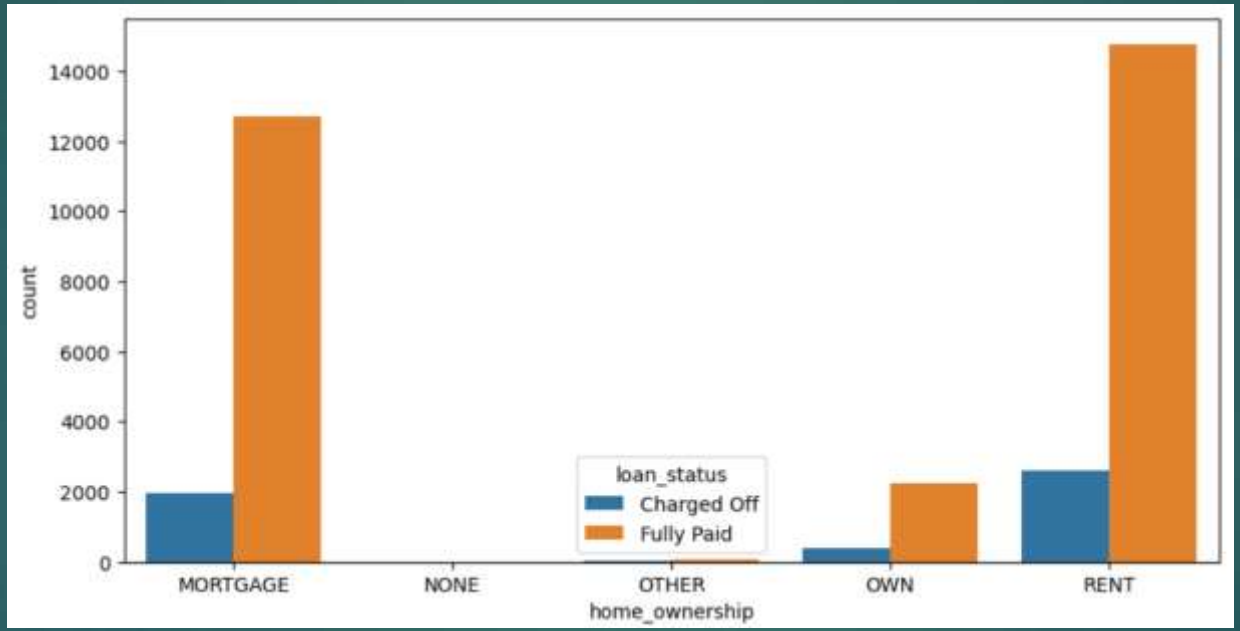
This suggests that the borrowers with DTI within 10-15 have a higher chance of defaulting and paying in full as shown below:



# Analysis of Home Ownership based on Loan Status:

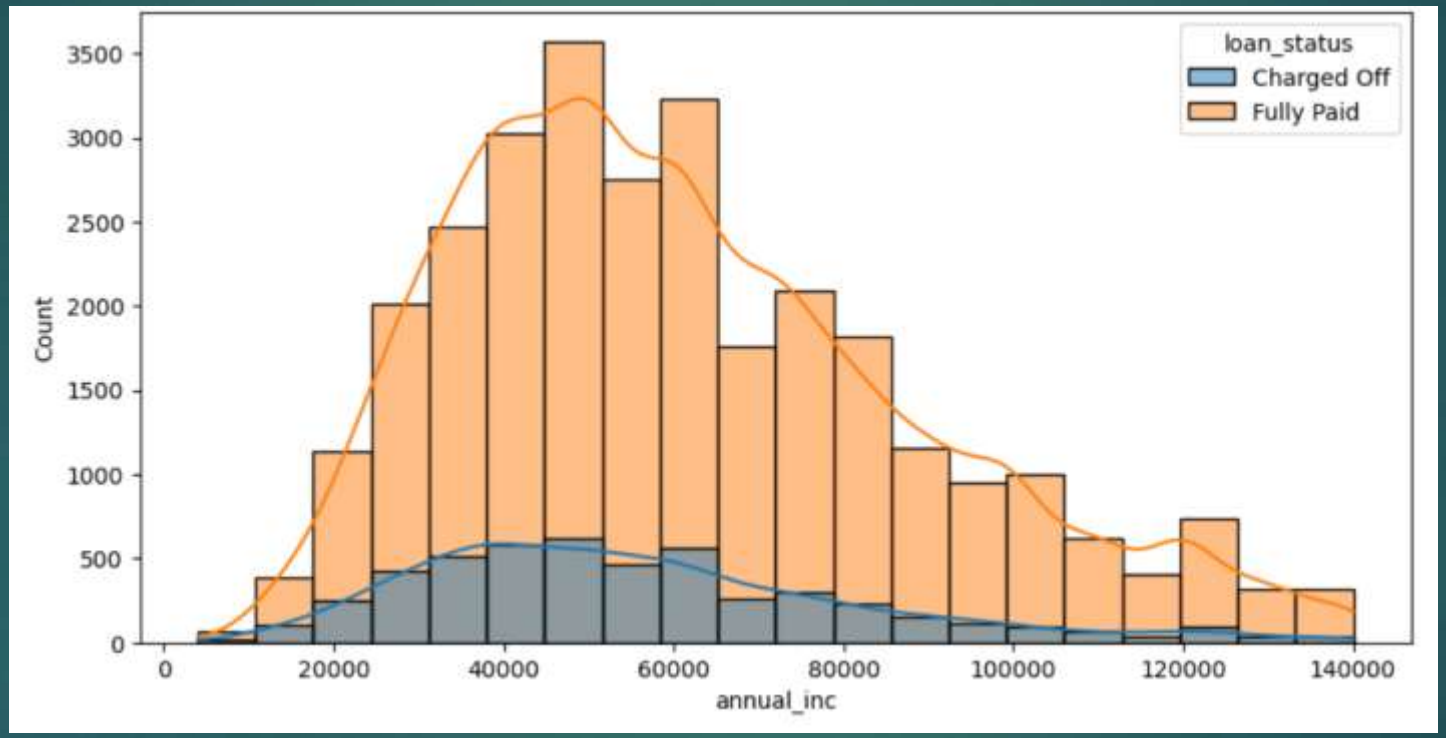
The analysis of home ownership based on loan status showed that the borrowers who own a home have very low count of defaulting loans compared to people who rent or are on mortgage.

This suggests that the people who own a home are less likely to default as shown below:



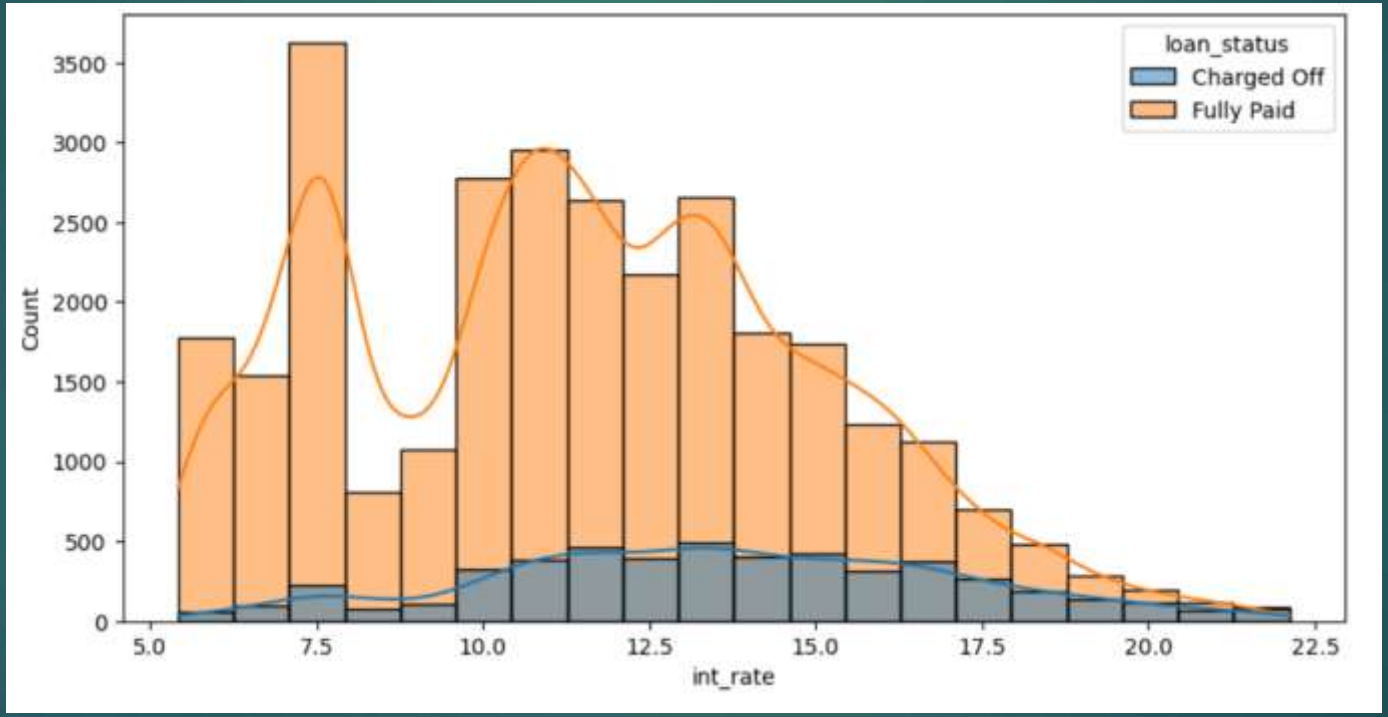
# Analysis of Annual Income based on Loan Status:

The analysis of annual income based on loan status showed that the borrowers who earn 40000-60000 annually have a high likelihood to default as shown below:



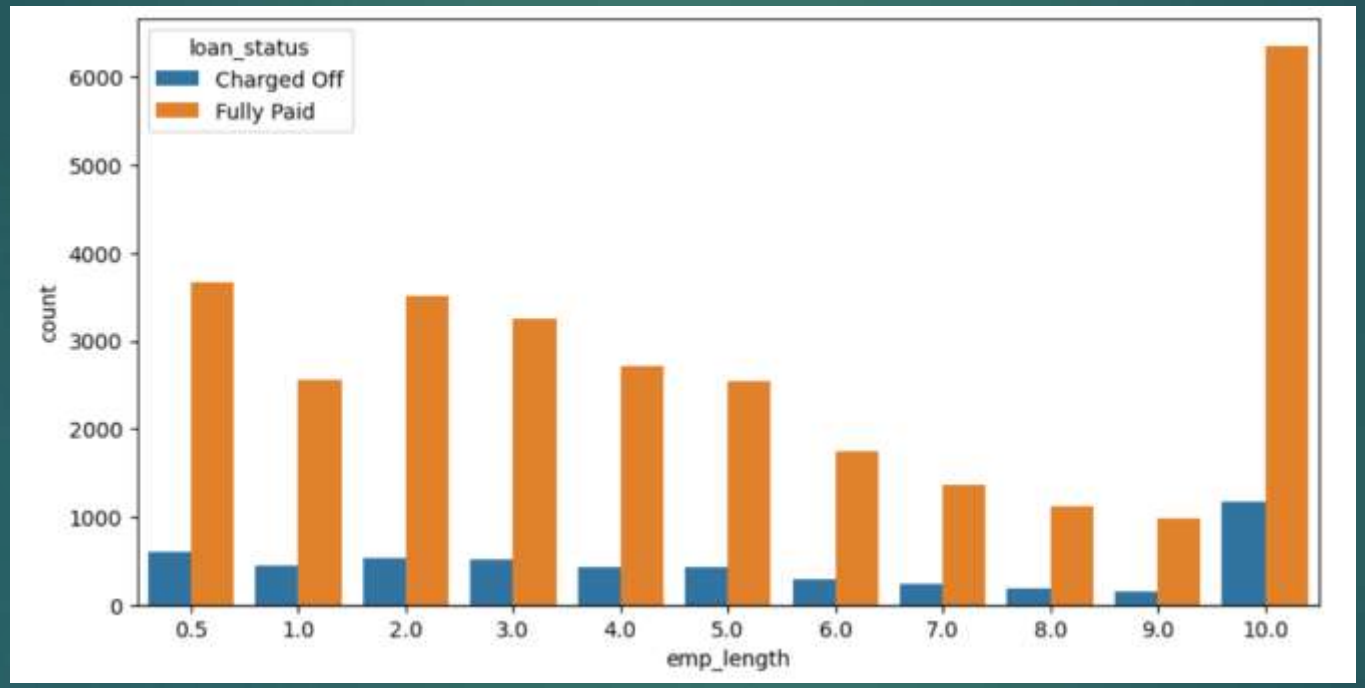
# Analysis of Interest Rate based on Loan Status:

The analysis of interest rate based on loan status showed that the borrowers who have borrowed a loan with interest rate of 11-17.5 have a high likelihood to default. We can also infer that with increasing interest rate the number of defaulters increases upto 17.5 post which it decreases following a stable trend as shown below:



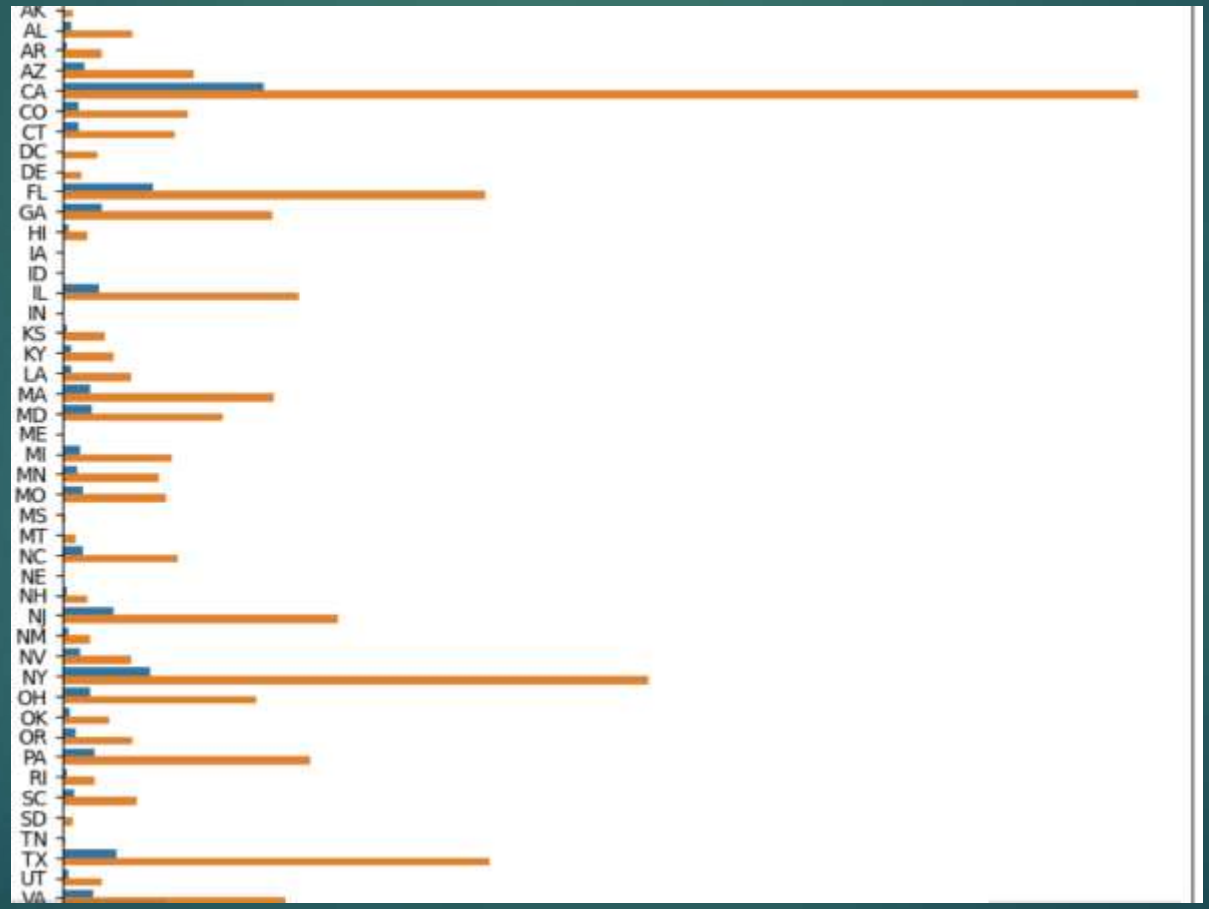
# Analysis of Employment Length based on Loan Status:

The analysis of employment length based on loan status showed that the borrowers who have more than 10 years of working experience have a high likelihood to fully pay as well as default on their loans as shown below:



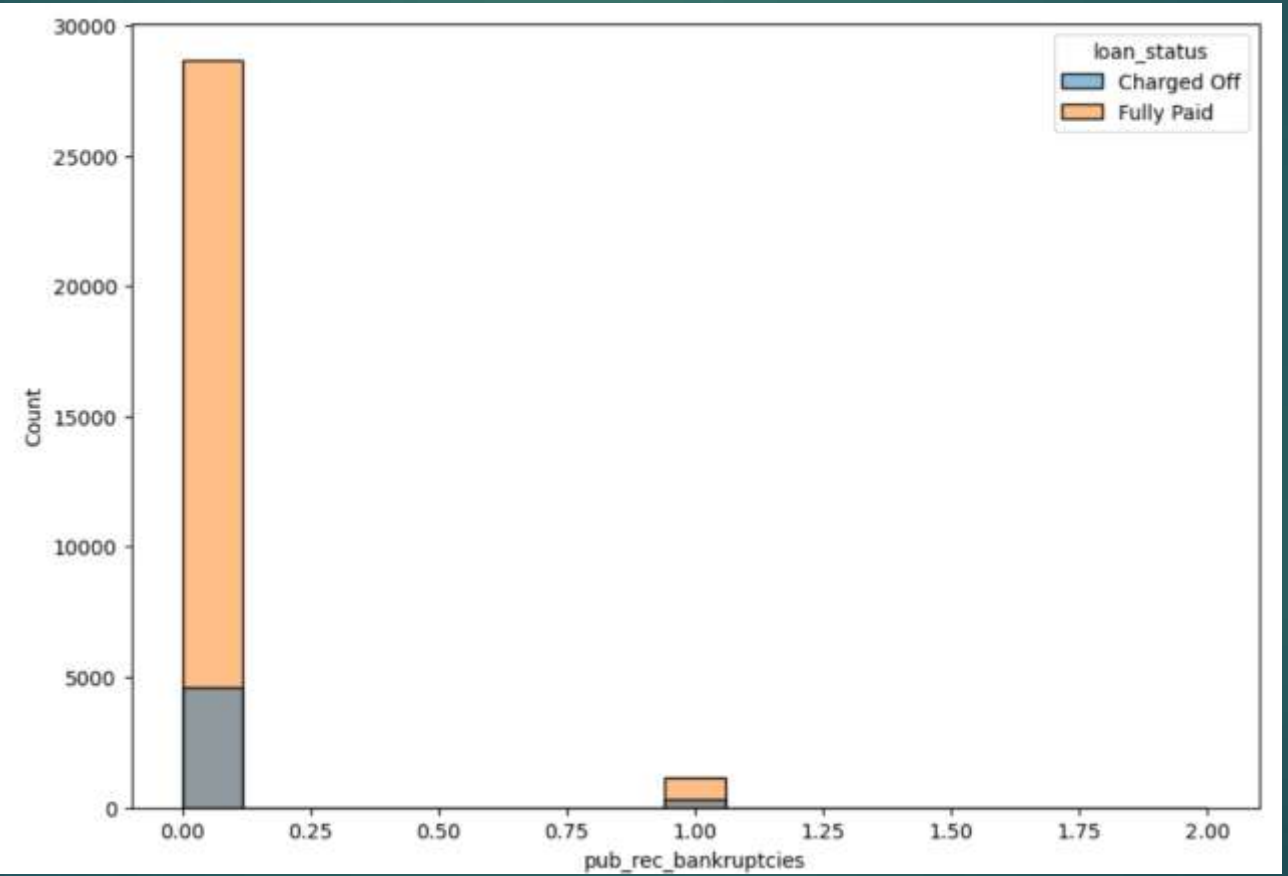
# Analysis of Address State based on Loan Status:

The analysis of address state based on loan status showed that the borrowers who are from California , New York and Florida have a high likelihood to pay their loan back in full as well as default on their loan as shown below:



# Analysis of Public Record Bankruptcies based on Loan Status:

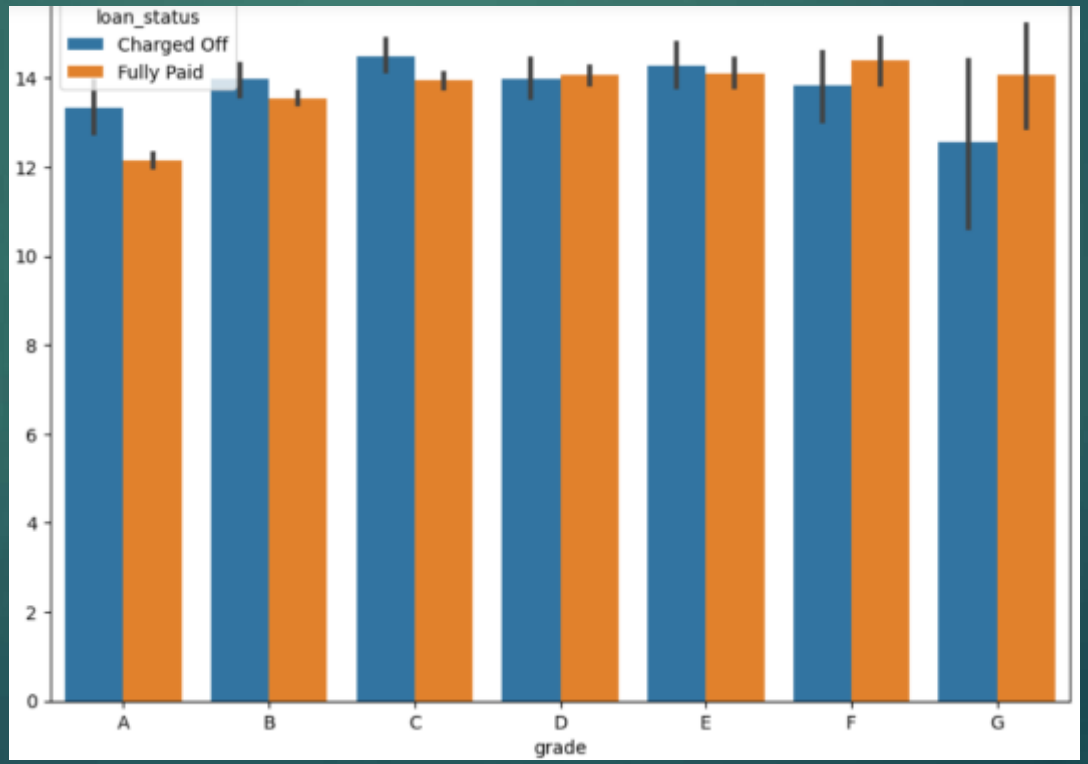
The analysis of public record bankruptcies based on loan status showed that the borrowers who have 0 public record bankruptcies have the highest number of fully paid as well as defaulted loans as shown below:



# Bivariate Analysis:

**Analysis of DTI over Grade for Loan Status:** The analysis of DTI over grade in terms of loan status showed that for loans of grade A-C the higher the DTI the higher the chance of the loan being defaulted, where as for higher grade loans the value of DTI and fully paid and charged off loans hold a similar ratio and even at times the fully paid loans for higher grades have even DTI.

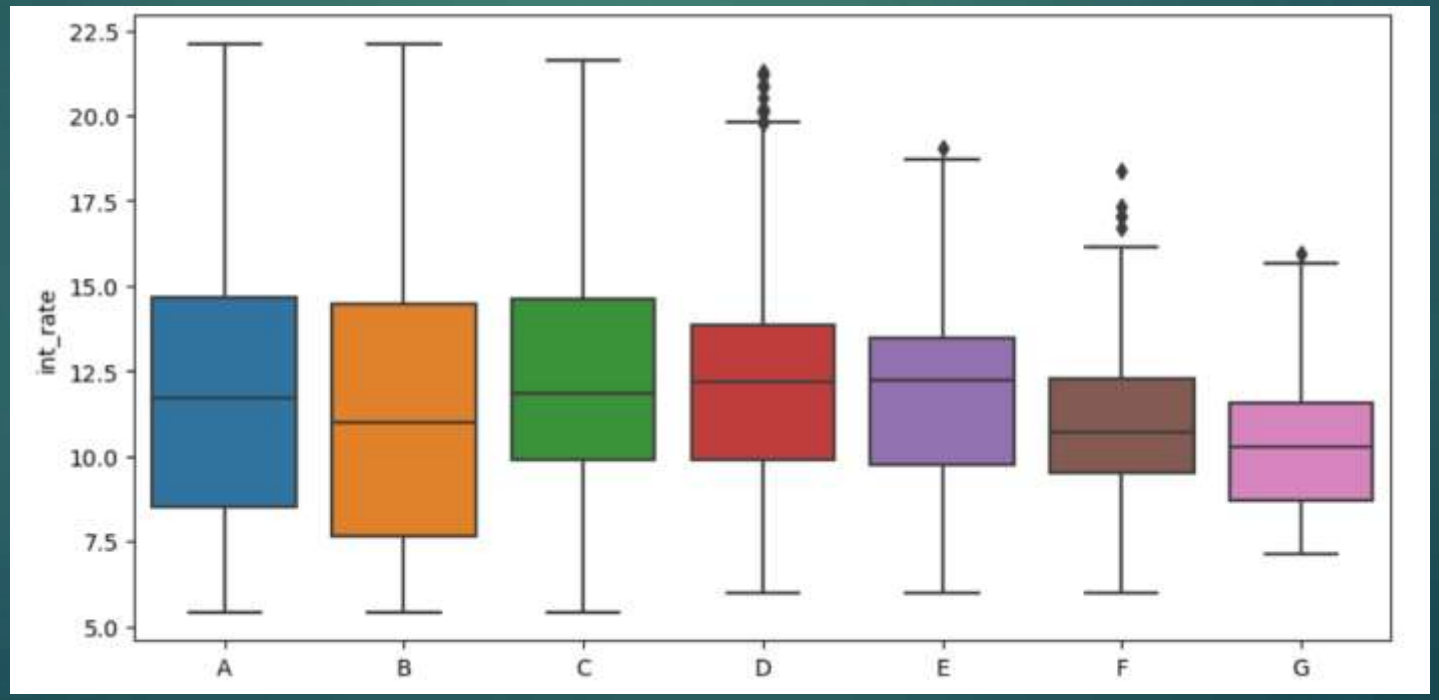
This suggests that for low grade loans with borrowers having high DTI in comparison are more likely to default on their loans ,as shown below:



# Analysis of Interest Rate based on Grade:

The analysis of interest rate based on grade showed that the median Interest rate decreases with each grade after grade E. The 75% percentile follows a declining curve with increase in grade. The 25% percentile drops for grade B but rises again for grade C after which it follows a declining curve with increasing grade.

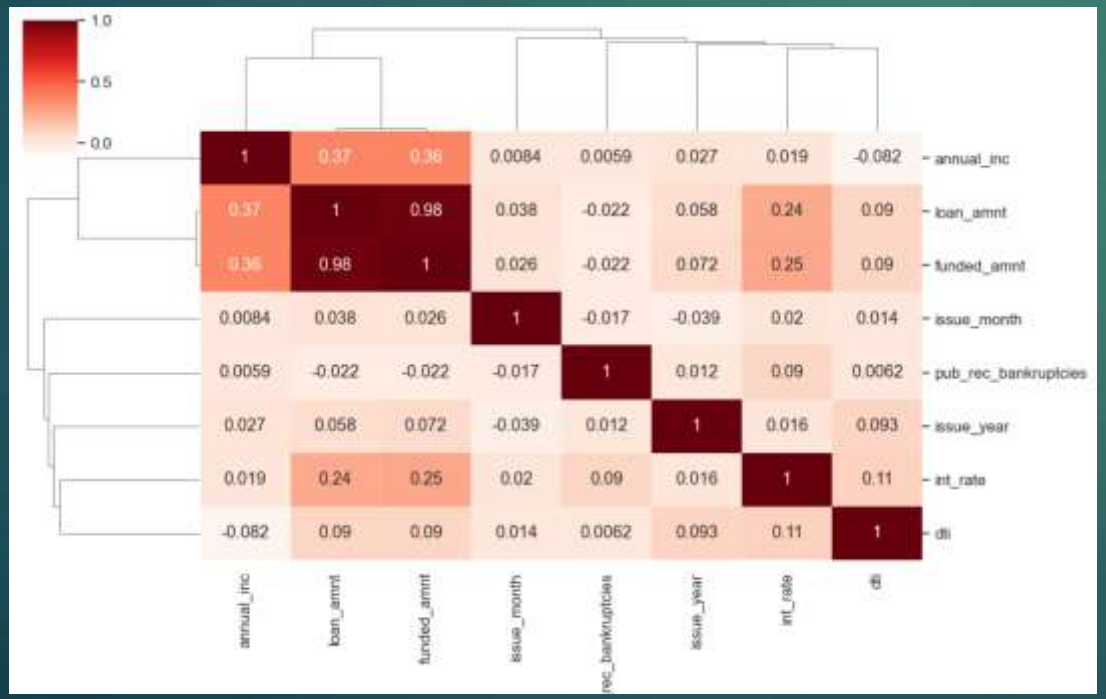
All these observations infer that the higher the loan grade the lower the interest rate and grade B has the broadest Inter Quartile range and thus contains a large spectrum of interest rates as shown below:



# Correlation Analysis:

The Correlation analysis of the numerical columns showed that:

- There is a high correlation between funded_amnt and loan_amnt.
- There is a moderate correlation between annual_inc and loan_amnt.
- There is a moderate correlation between annual_inc and funded_amnt.
- There is a moderate correlation between int_rate and loan_amnt.
- There is a moderate correlation between int_rate and funded_amnt.



# Recommendations:

Major driving factors which can be used to predict loan defaulters are:

1) DTI
2) Loan Term Period
3) Loan Amount
4) Annual Income
5) Grade
6) Home Ownership
7) Interest Rate

We can refer to the comparative analysis of all these variables to figure out and predict loan defaulters and then come up with strategy to minimize defaulting by addressing all these factors.