Lending Club EDA Case Study

By:
Ankush Kshirsagar
Vishakha Itankar

Batch: ML C50

Approach for EDA

Data Understanding

- loan.csv contains the input source data for EDA Lending Club case study
- The initial data count of this dataset is 39717 rows and 111 column variables
- The dataset mainly contains the loan characteristics such as loan status, interest rate, term, purpose, issue date
- It also contains the loan applicants demographic and behavioural variables such as employment length, state, revolving balance, next payment date etc
- It contains many variables which are completely blank i.e. 100% missing data
- It also contains variables which has datatype mismatch and needs correction. One such example is interest rate where it should be a numeric column but as it has % symbol it is has datatype as object
- There are also variables which would need to be split to get the derived variables. Example: Issue date to get year and month

Data Cleaning : Variables

1. Identify the missing data by getting the percentage of missing values
   - There are 54 variables which has 100% data missing
   - Also there are 3 more variables which has missing data more than 60%
   - Dropping these variables 54 + 3 = 57 variables
2. Dropping few more columns based on the following analysis
   - Variables which has single value based on above analysis
   - Variables acting as an indexing as it has unique values
   - Variables which are descriptive in nature
3. Dropping variables 'out_prncp','out_prncp_inv' as it has no correlation

Data Cleaning : Rows

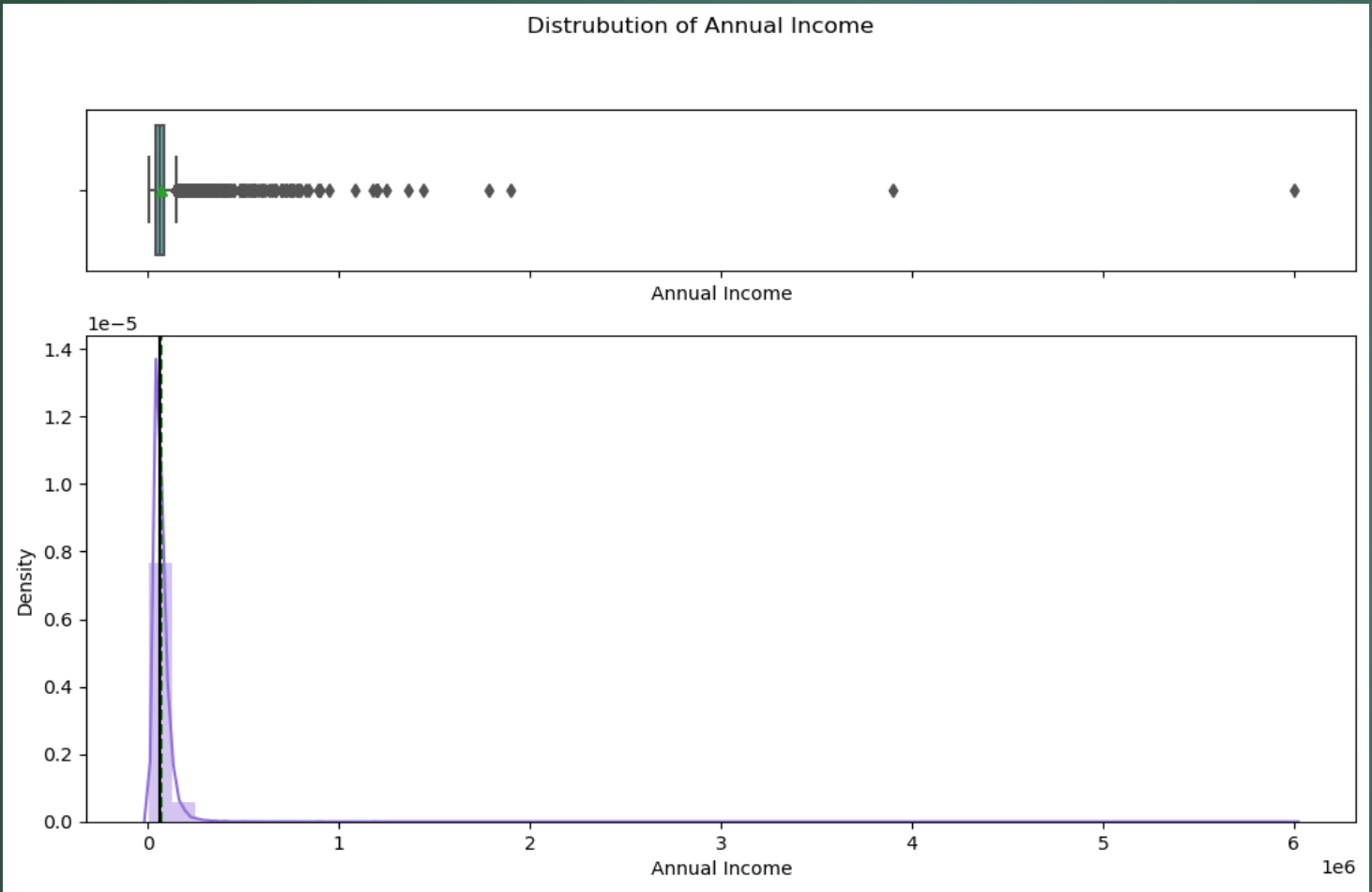
1. Dropping rows based on the NA columns which has percentage less than 0.2%
2. Loan status ='Current' has very less data and won't be helpful in identifying defaulter, so focusing on 'Fully Paid' and 'Charged off' status only
3. emp_length variable has 1024 missing values so need to drop them

Fixing data:

1. Removing '%' symbol from 'revol_util' and 'int_rate' variables
2. Changing its datatypes to float
3. Extracting numeric value from 'emp_length' and 'term' variable
4. Changing its datatypes to number

Univariate Analysis

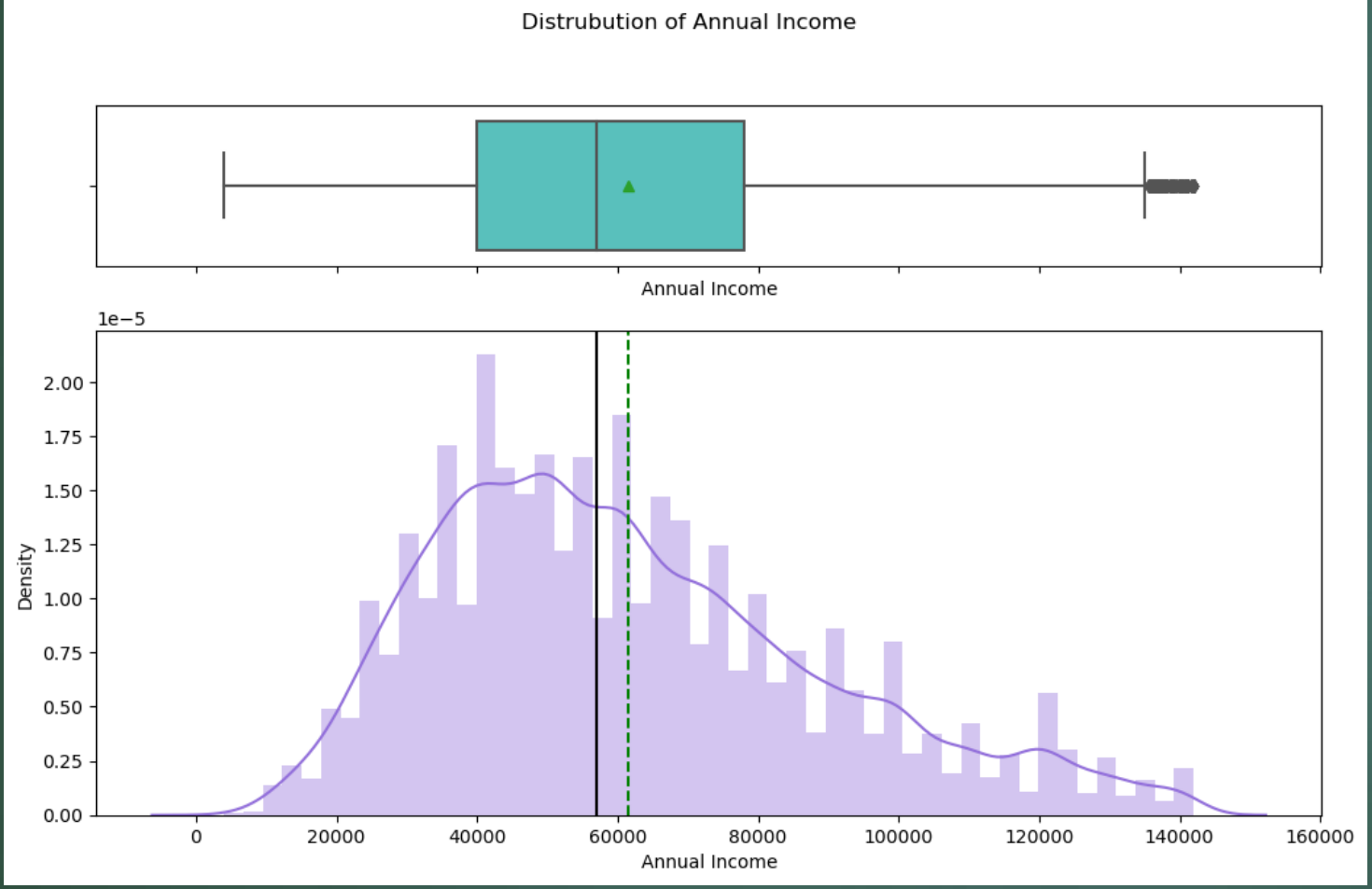
Annual Income:



1. After plotting the variable and analyzing the quantiles it was observed that there outliers with huge values
2. Quantile analysis and plot shows that the outliers after 0.95 quantiles can be removed

Univariate Analysis

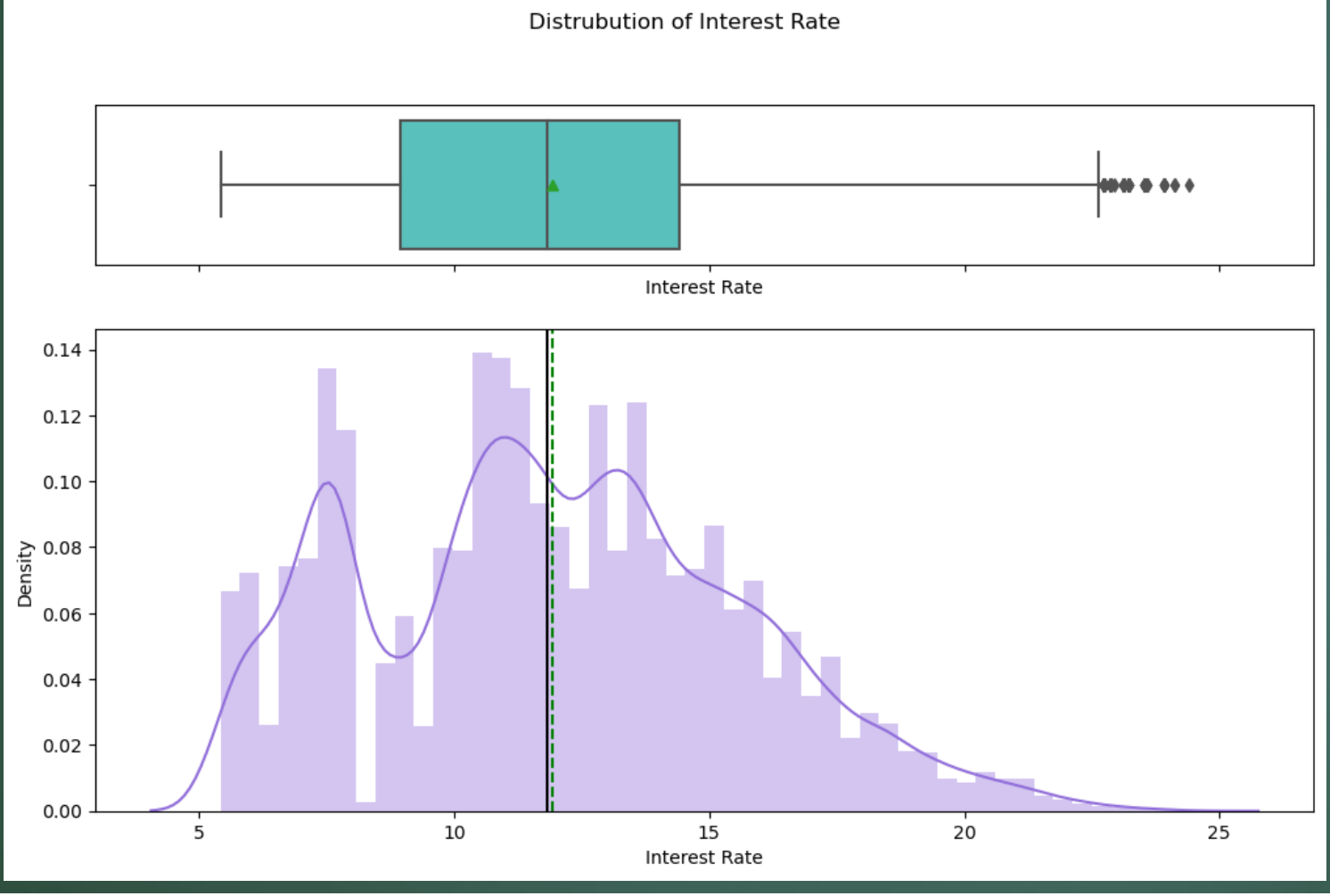
Annual Income: continued



3. After dropping the outliers ,it is significant that majority of values are in the range of 39000 to 79000

Univariate Analysis

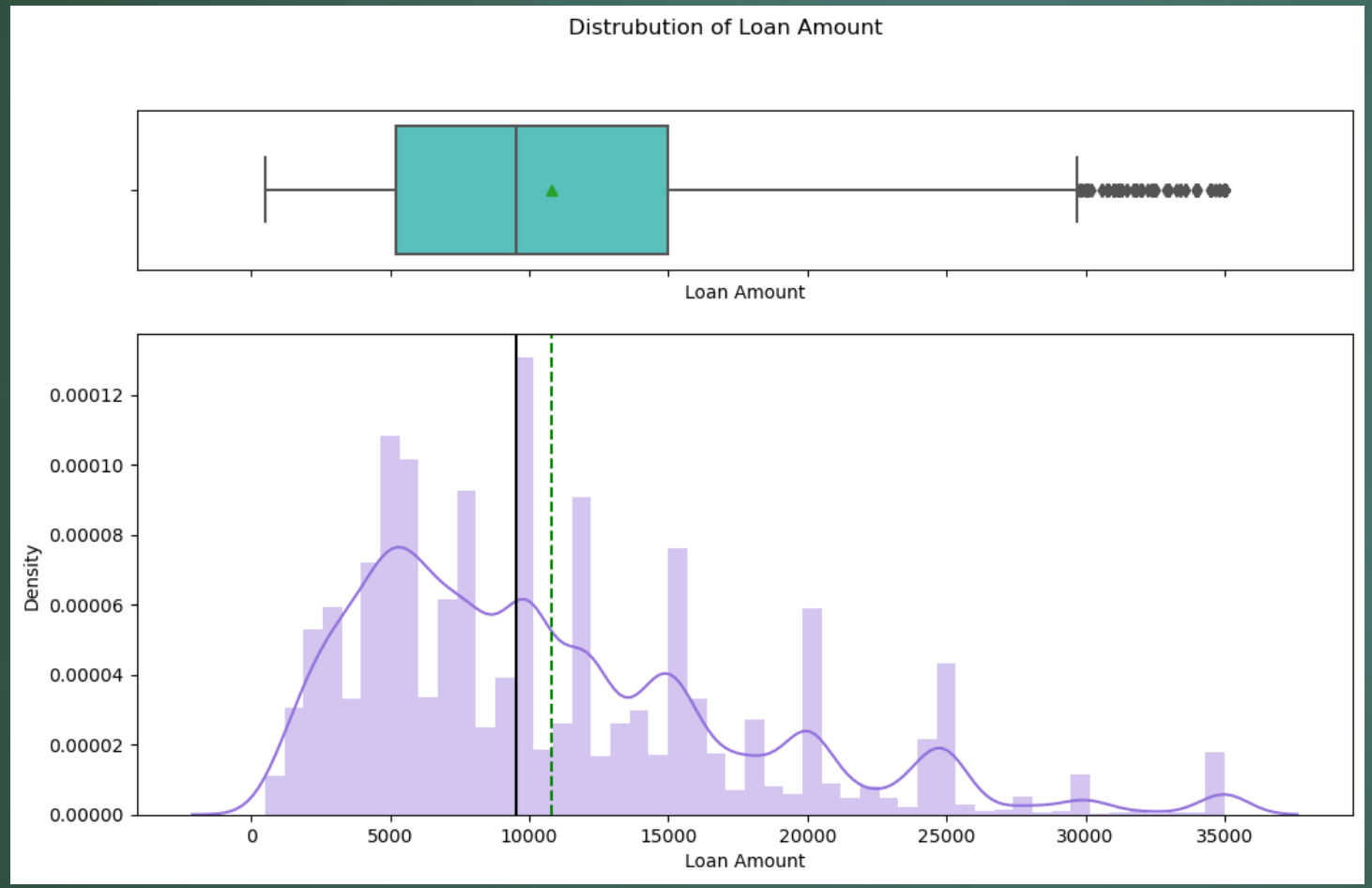
Interest Rate:



1. Majority of the loans are in the interest rate range of 9-15%

Univariate Analysis

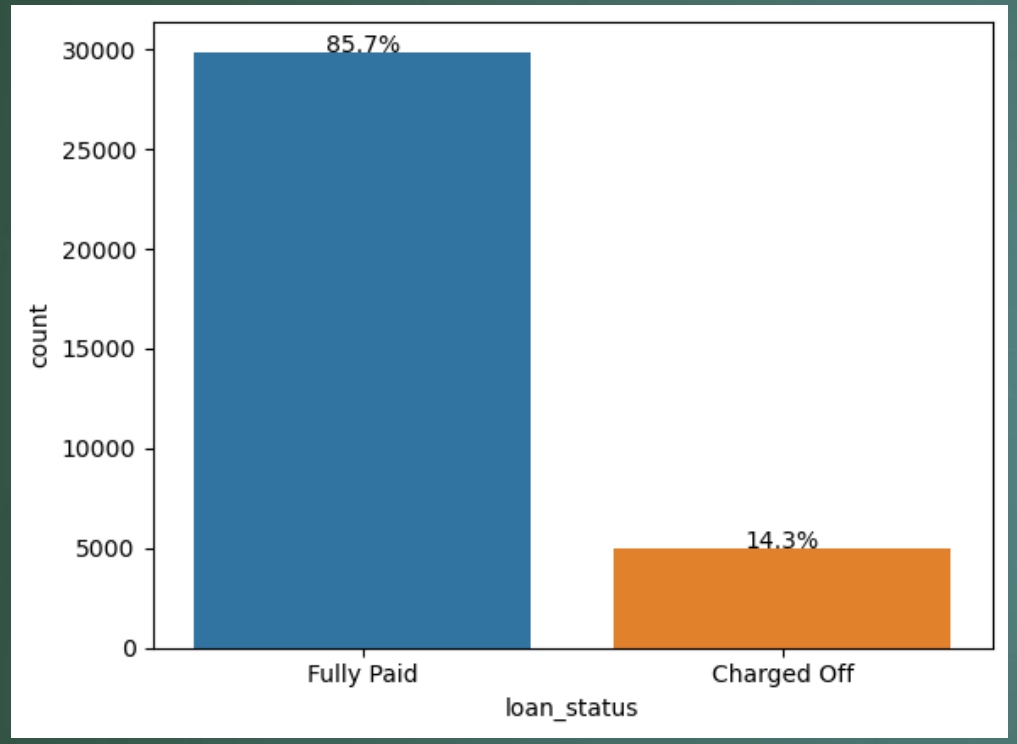
Loan Amount, Funded Amount and Funded Amount Inv:



1. Loan Amount, Funded Amount and Funded Amount Inv has similar distribution and observations
2. There are not many outliers having significantly huge values
3. Most of the loans has loan amount which lies in the range of 5000 to 15000

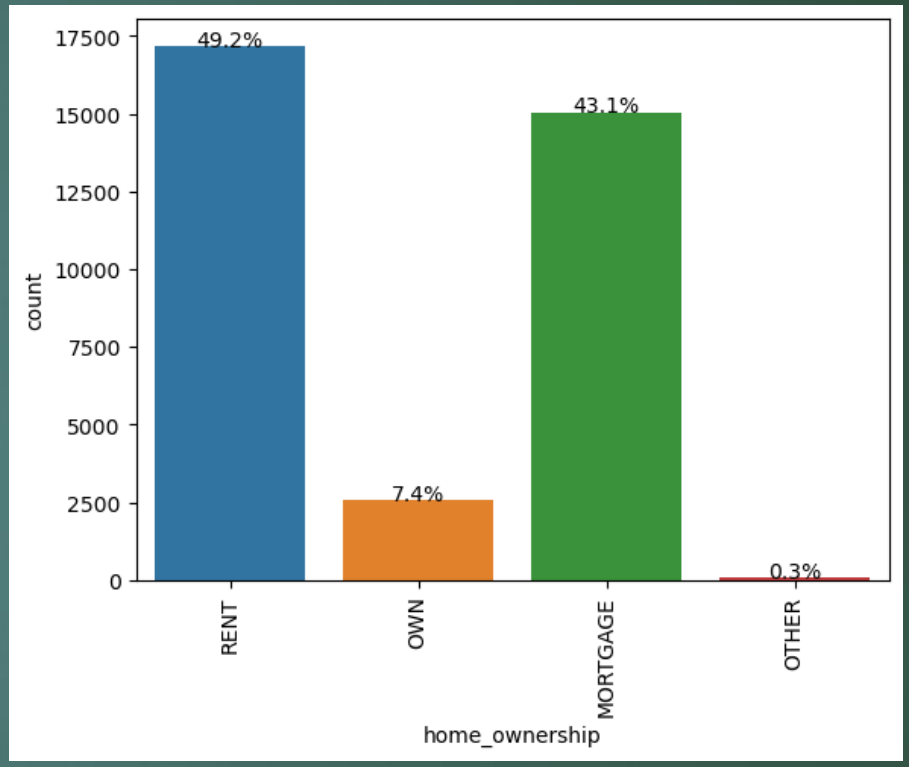
Segmented Univariate Analysis

Loan Amount



Majority of the applicants are able to pay the loan amount fully

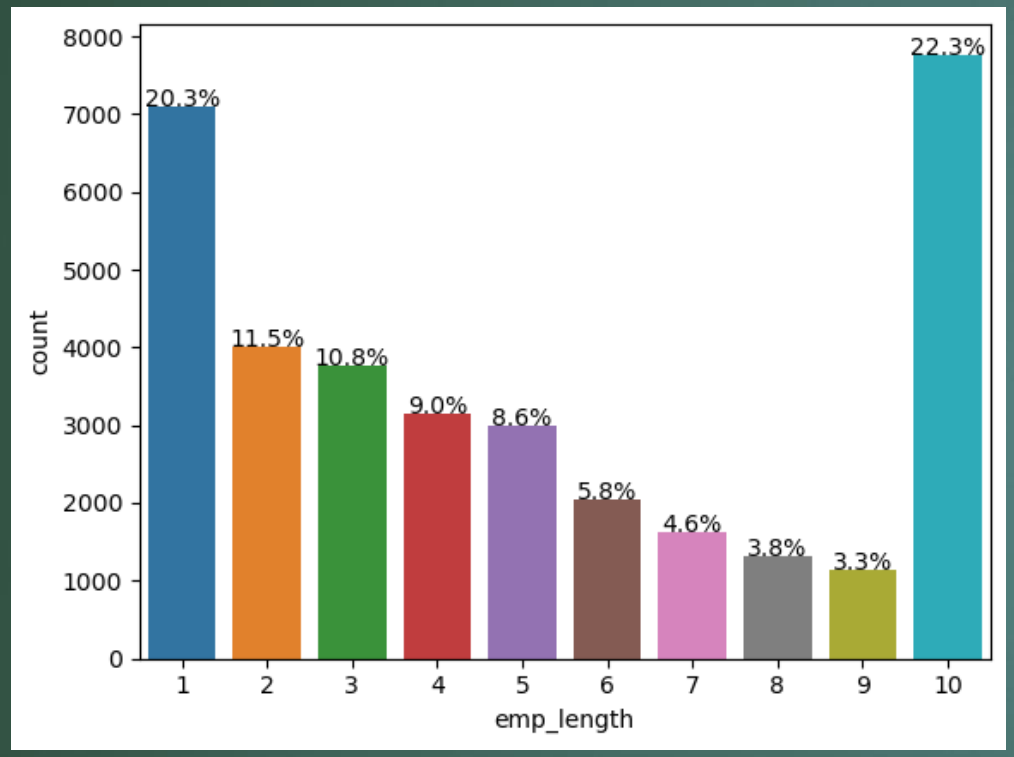
Home Ownership



Most of the loan holders have rented house or has mortgage on their home

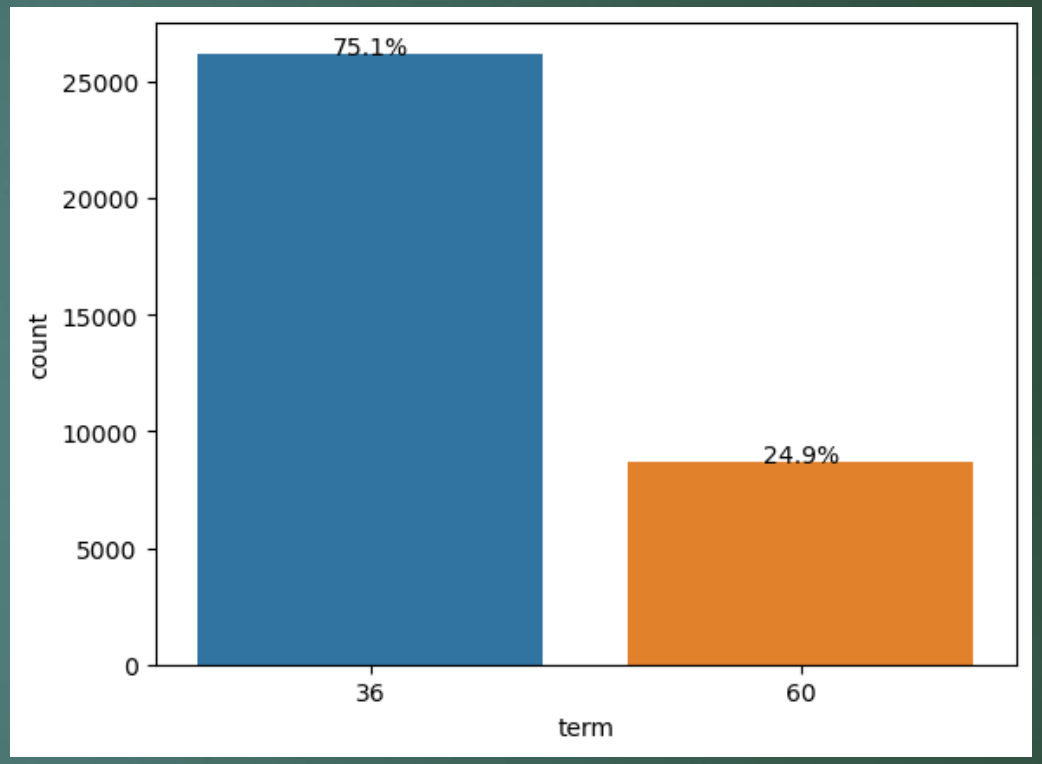
Segmented Univariate Analysis

Employment Length



Maximum loans are taken by the applicants who are employed for a tenure of 10+ years or 1 year

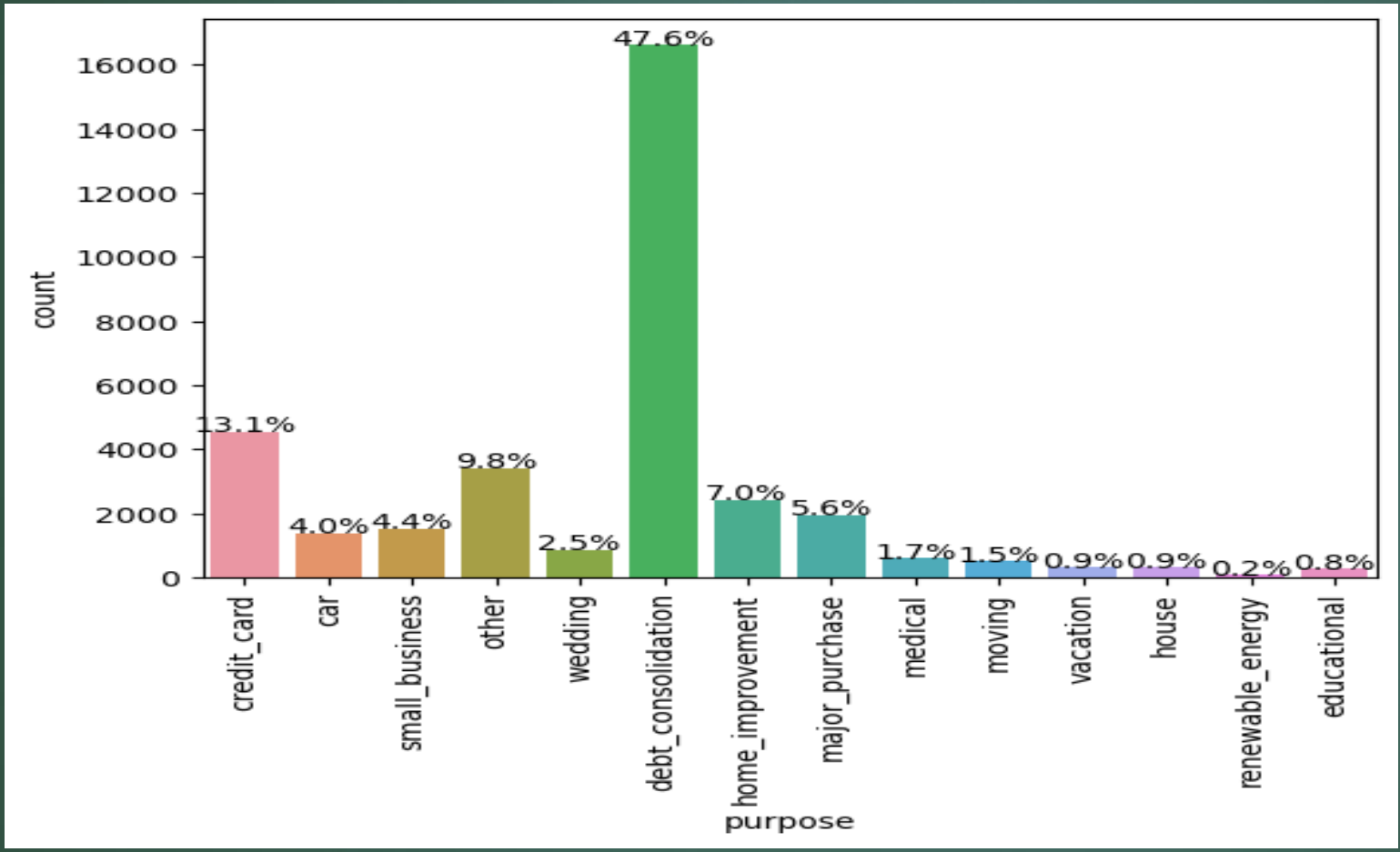
Term of the loan



There are only 2 terms(36 months and 60 months) for which the applicants has applied for the loan

Segmented Univariate Analysis

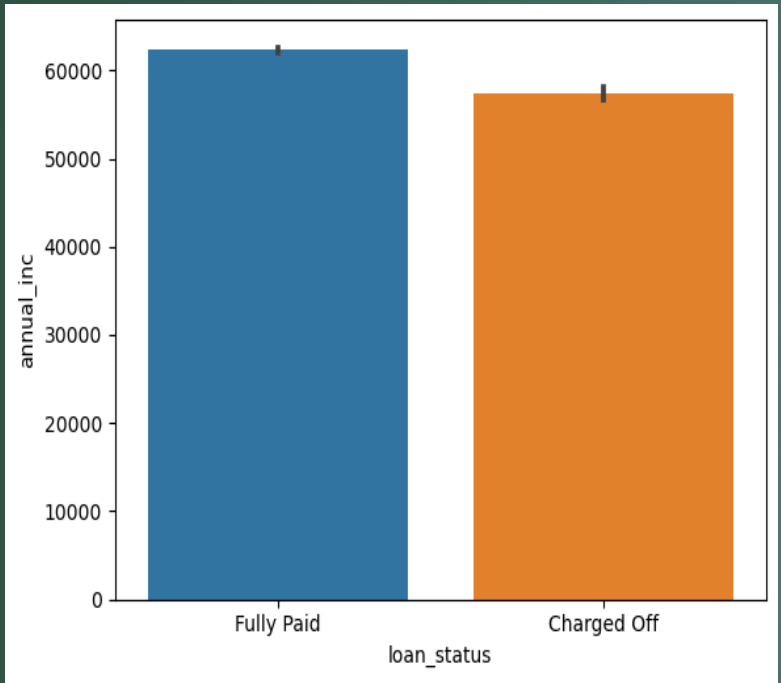
Purpose of loan



Maximum percentage of applicants has opted to take loans for debt consolidation

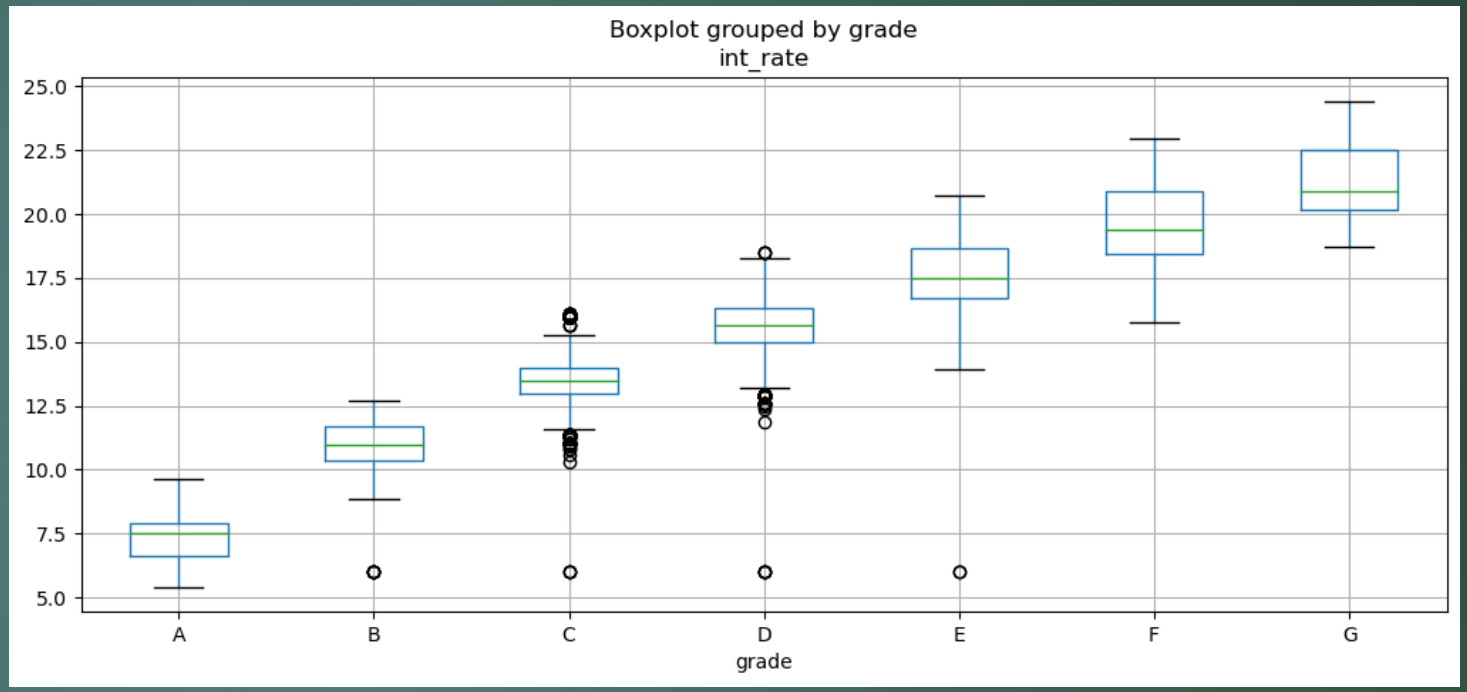
Bivariate Analysis

Loan Status Vs Annual Income



Applicants having higher **average** annual income(above 58000) are able to pay the loan and there are no defaulters in that range

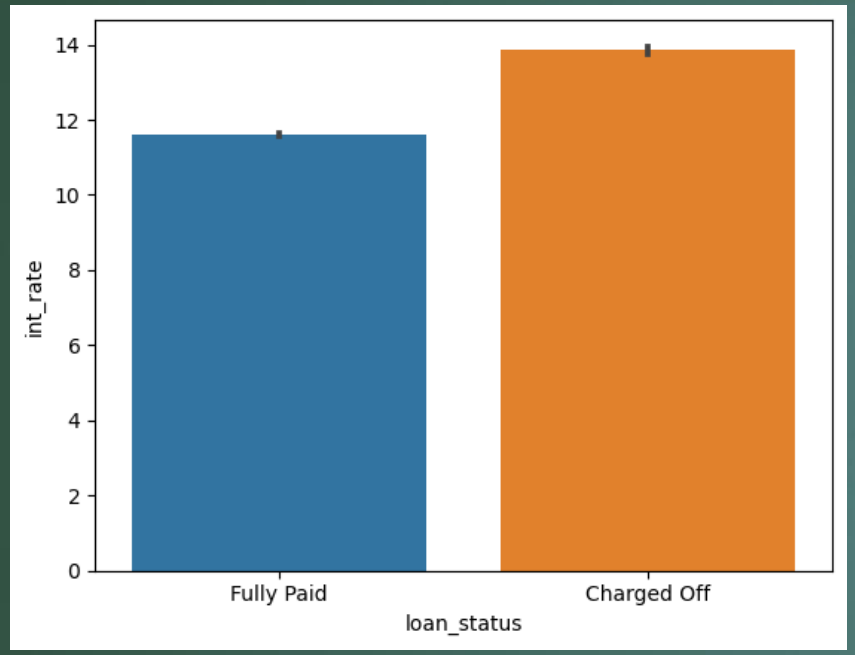
Interest Rate Vs Grade



There is an evident trend of increasing interest rate as the Grade decreases, excluding few outliers

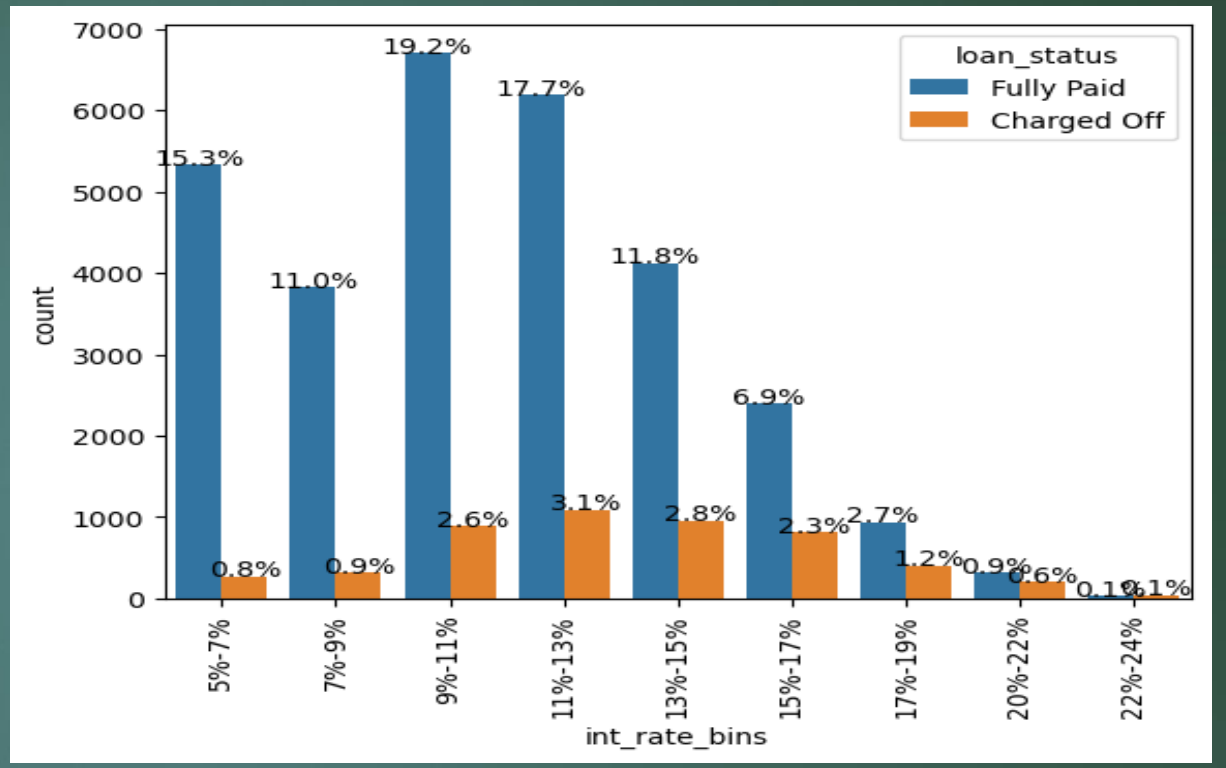
Bivariate Analysis

Loan Status Vs Interest Rate



- **Average** interest rate(mean) of Fully Paid loans is around ~11.6% and of Charged Off loan is ~12.6%
- Which implies there is a possibility that higher interest loans have more defaulters

Distribution of Loan Status over Interest Rate

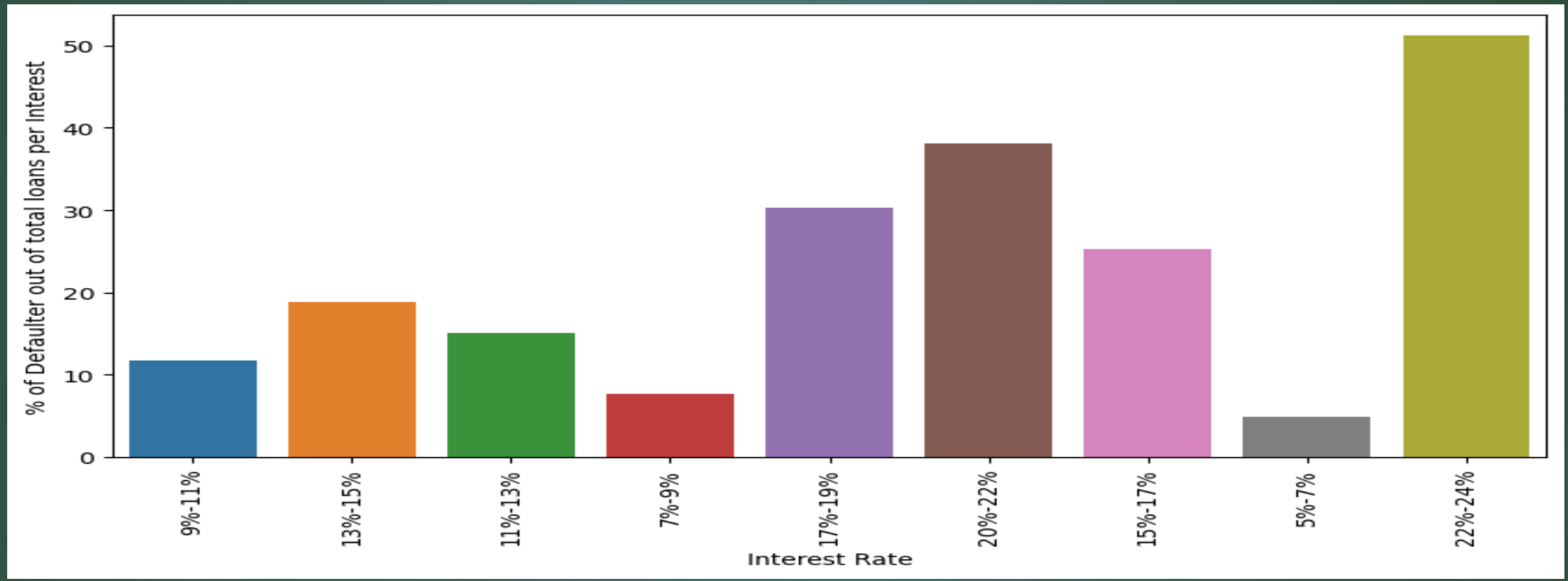


- Maximum defaulted loans are in the range of 11-13% followed by 13-15%
- Approximately equal percentage of loans were fully paid and defaulted at the higher range of interest rate i.e. 20-24%
- Very less applicants has defaulted at the lower rates 5-9%

Bivariate Analysis

Analysis Charge Off Loan Status vs Interest rate

Note:
Percentage default = Number of loans defaulted at an interest rate range divided by Total(defaulted + fully paid) number of loans given at an interest range* 100

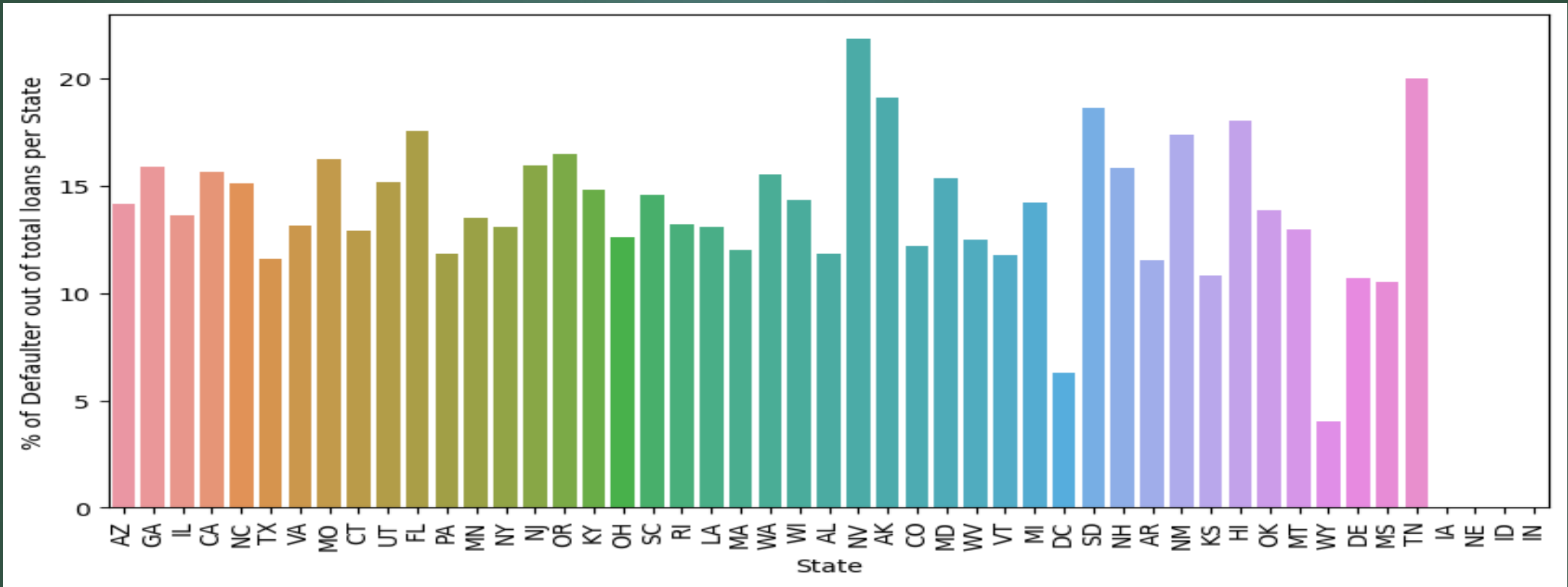


- Loans with interest range 22%-24%, 20%-22% and 17%-19% have seen more defaulters

Bivariate Analysis

Note: Percentage default = Number of loans defaulted in a state divided by Total(defaulted + fully paid) number of loans taken in a State * 100

Analysis Charge Off Loan Status vs State

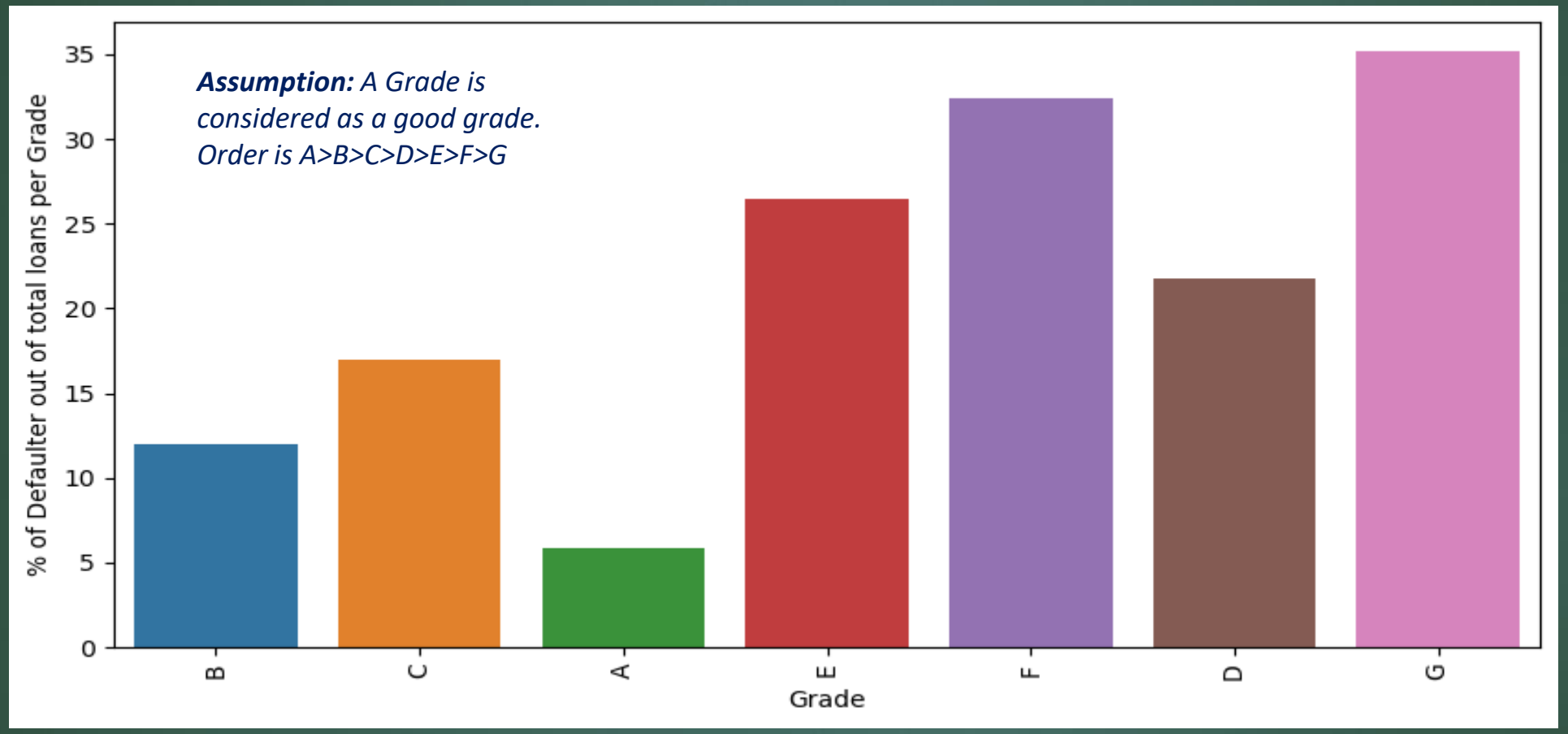


- Percentage of defaulter against the total loan applications is higher in NV,TN,AK and SD
- Applicants from NV,TN,AK,SD states has a tendency to default more

Bivariate Analysis

Note: Percentage default = Number of loans defaulted for a grade divided by Total(defaulted + fully paid) number of loans given for a grade * 100

Analysis Charge Off Loan Status vs Grade

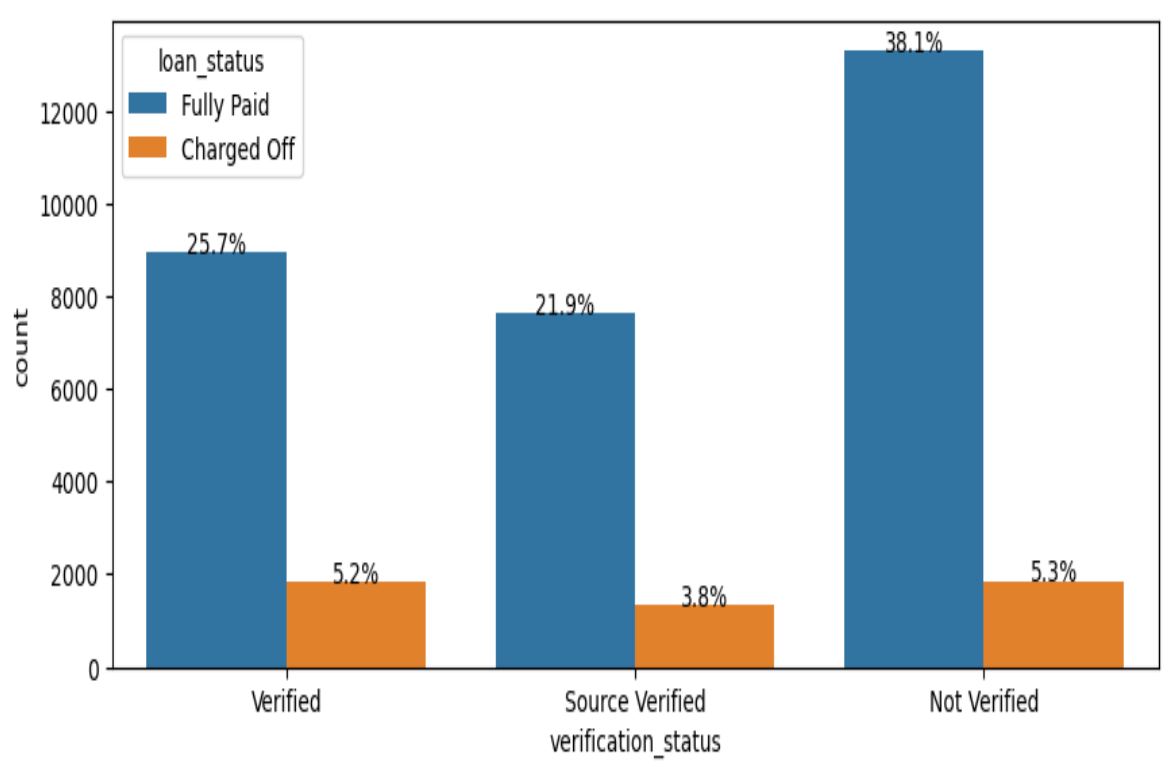


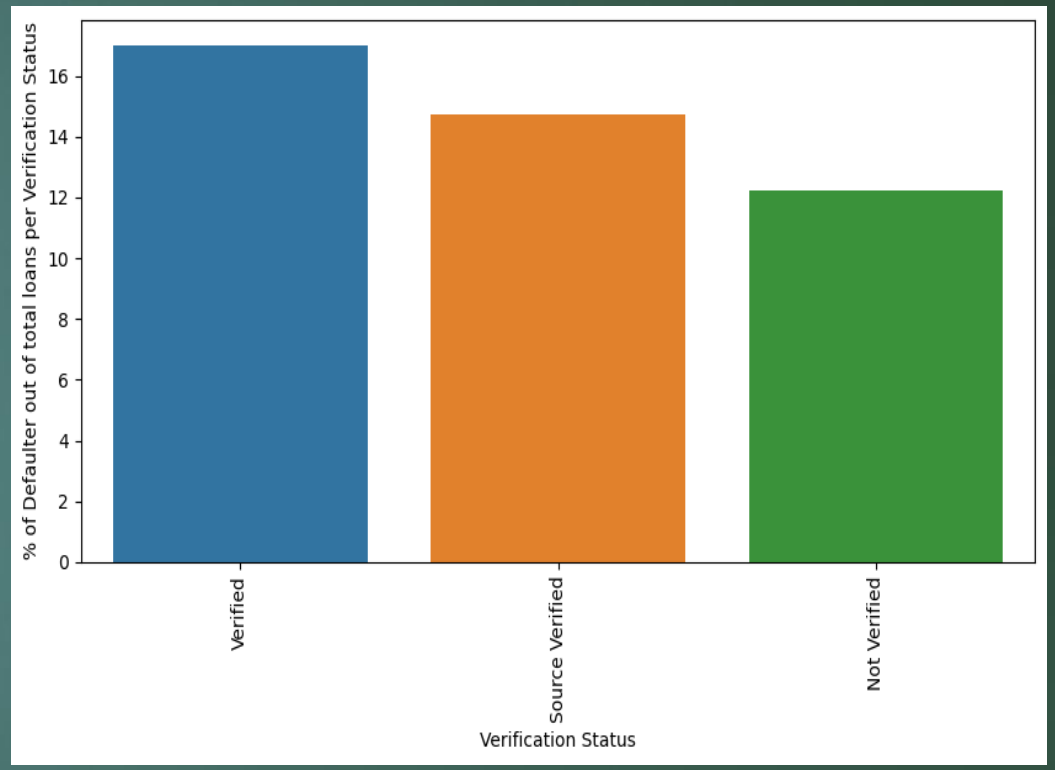
- Percentage of defaulter against the total loan applications is higher for Grade G followed by F

Bivariate Analysis

Analysis Charge Off Loan Status vs Verification Status

Note:
Percentage default = Number of loans defaulted for a Verification Status divided by Total(defaulted + fully paid) number of loans given for a Ver status * 100



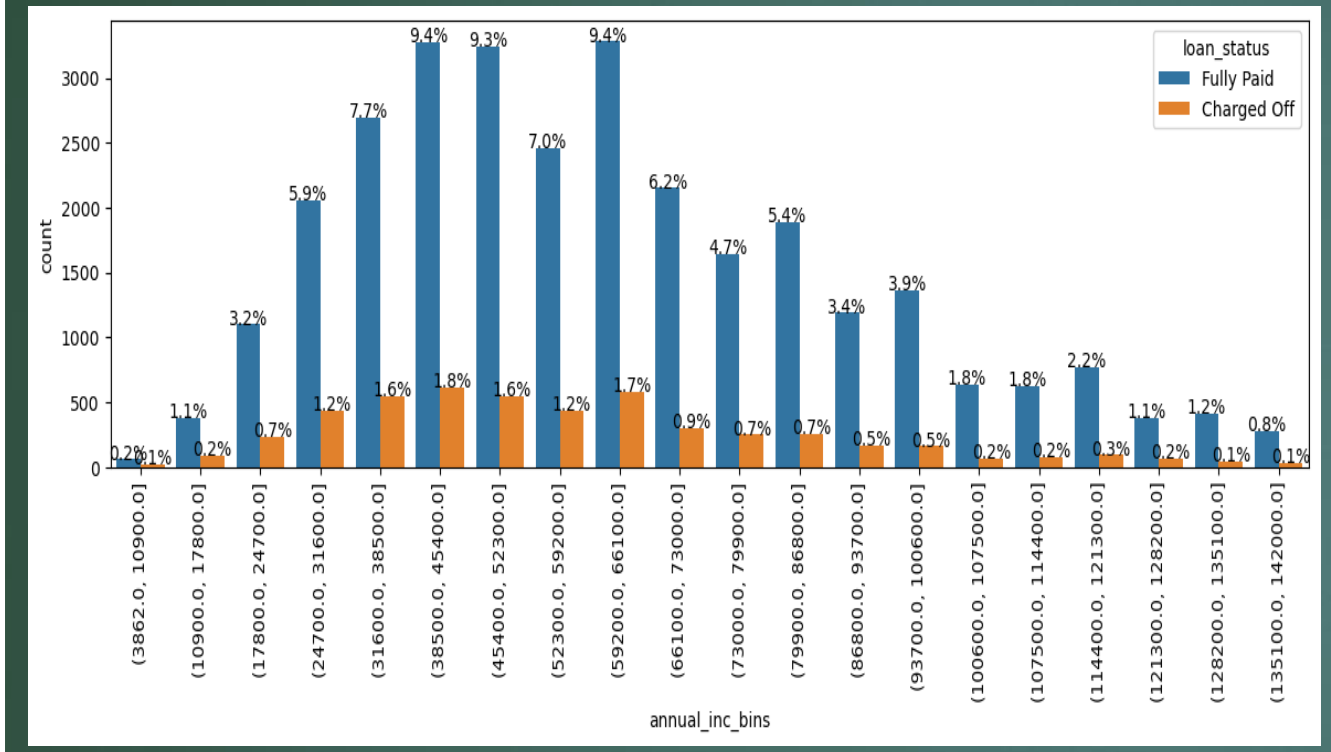


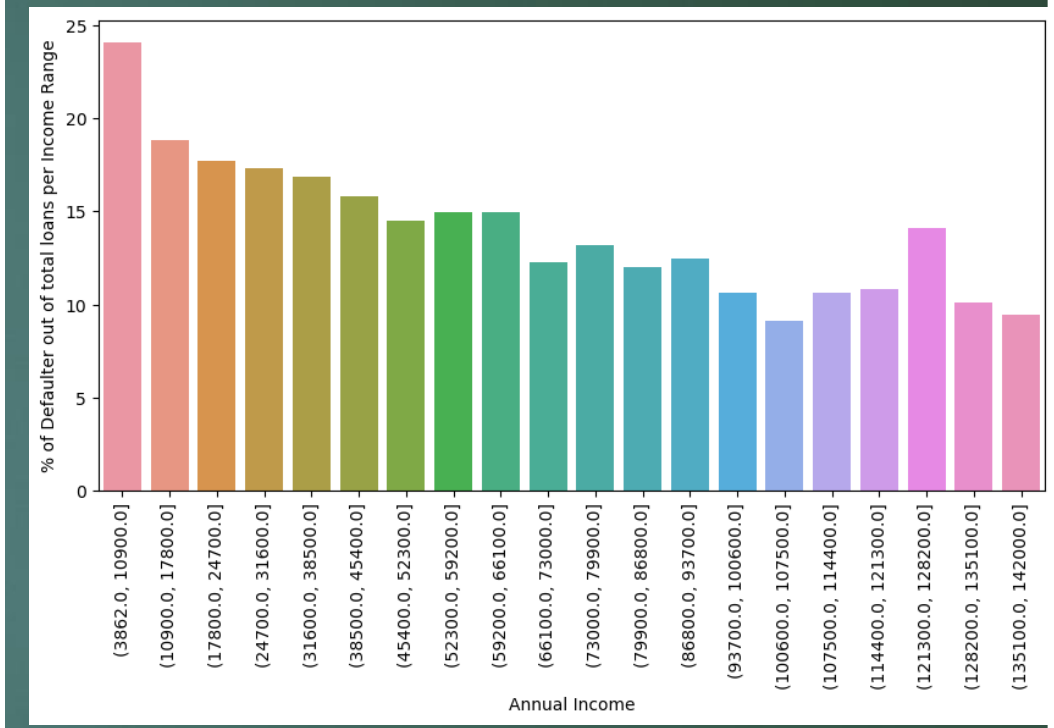
- Verified loans has the most defaulted loans as per the ratio of defaulter per verification status
- When it comes to count of defaulter across all verification status's, then loans with 'Not Verified' status have slightly more defaulter, but there is not significant difference in defaulter count of 'Verified' and 'Not Verified'

Bivariate Analysis

Analysis Charge Off Loan Status vs Annual Income

Note:
Percentage default = Number of loans defaulted for an Income range divided by Total(defaulted + fully paid) number of loans given for an income range* 100





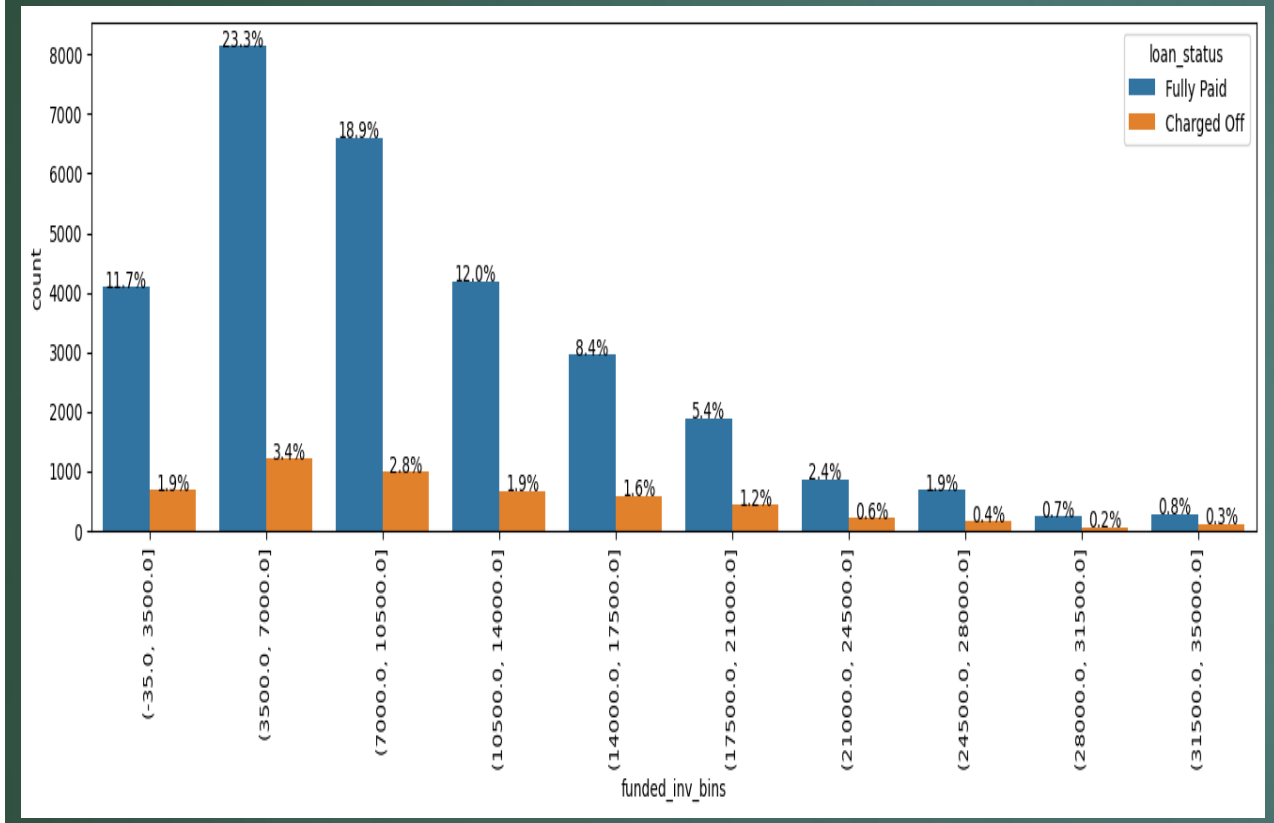
- The spread of loan is dense in the range of 24700-86800.
- The above statement holds true for defaulters as well, as majority of defaulters can also be seen in that income range

- While observing the ratio of defaulters vs total loans, applicants having lower annual income are more likely to default

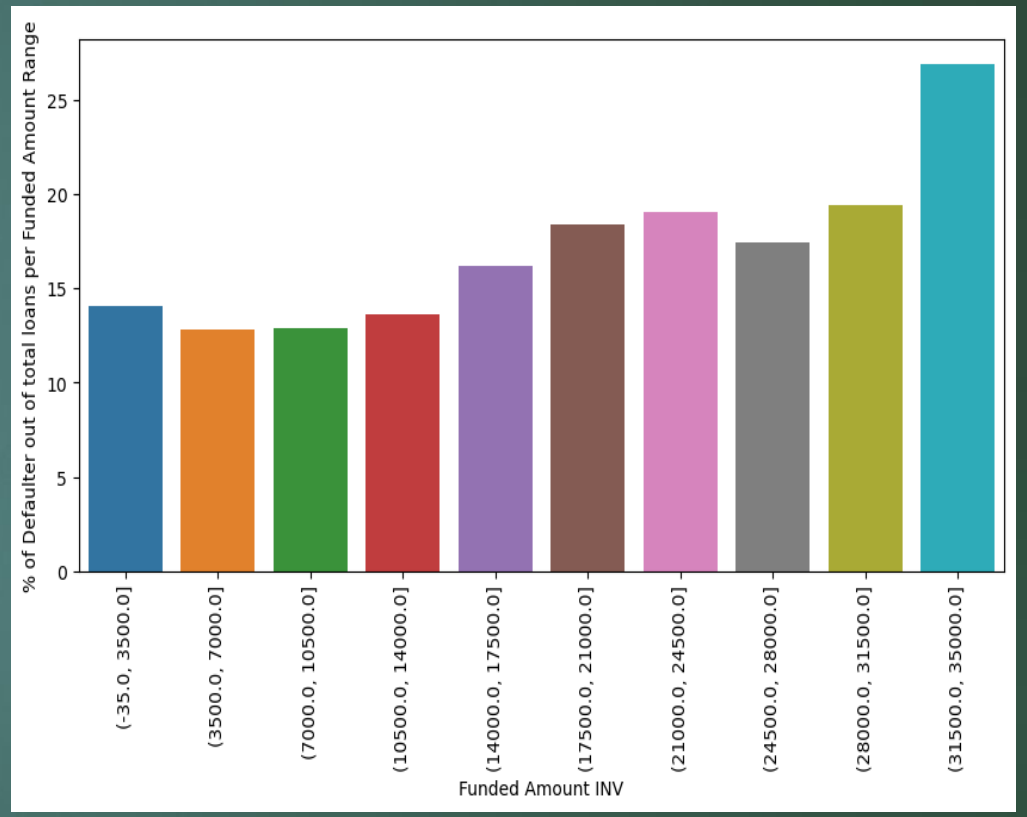
Bivariate Analysis

Note:
Percentage default = Number of loans defaulted for an amount range divided by Total(defaulted + fully paid) number of loans given for that amount range* 100

Analysis Charge Off Loan Status vs Funded Amount Inv



- Majority of Loans are having Funded Amount Inv in the range 3500- 10500
- Defaulters also has the same trend in this range

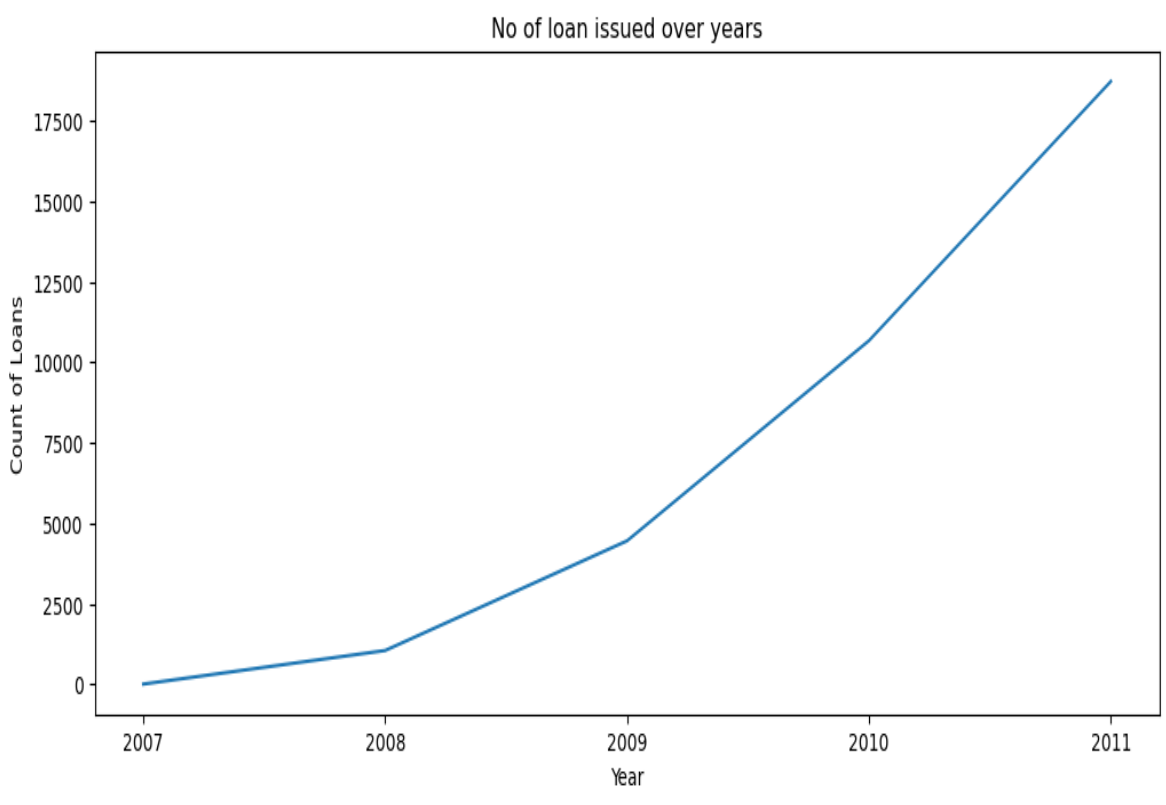


- Applicants having 'Funded Amount Inv' in the range 31500-35000 have more tendency to default

Derived Metrix

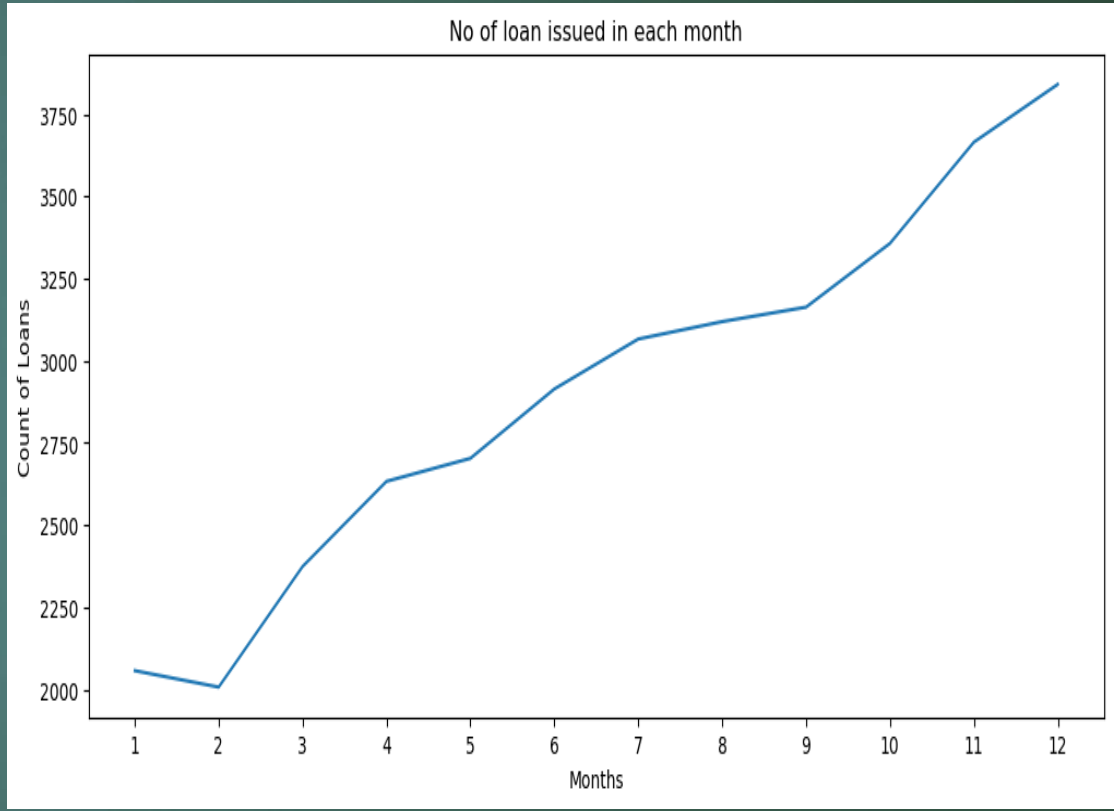
Note: Month and Year variables are derived from Issue_d variable

Analyzing count of loans per issued year



- Count of loans have increased over the years

Analyzing count of loans per issued year

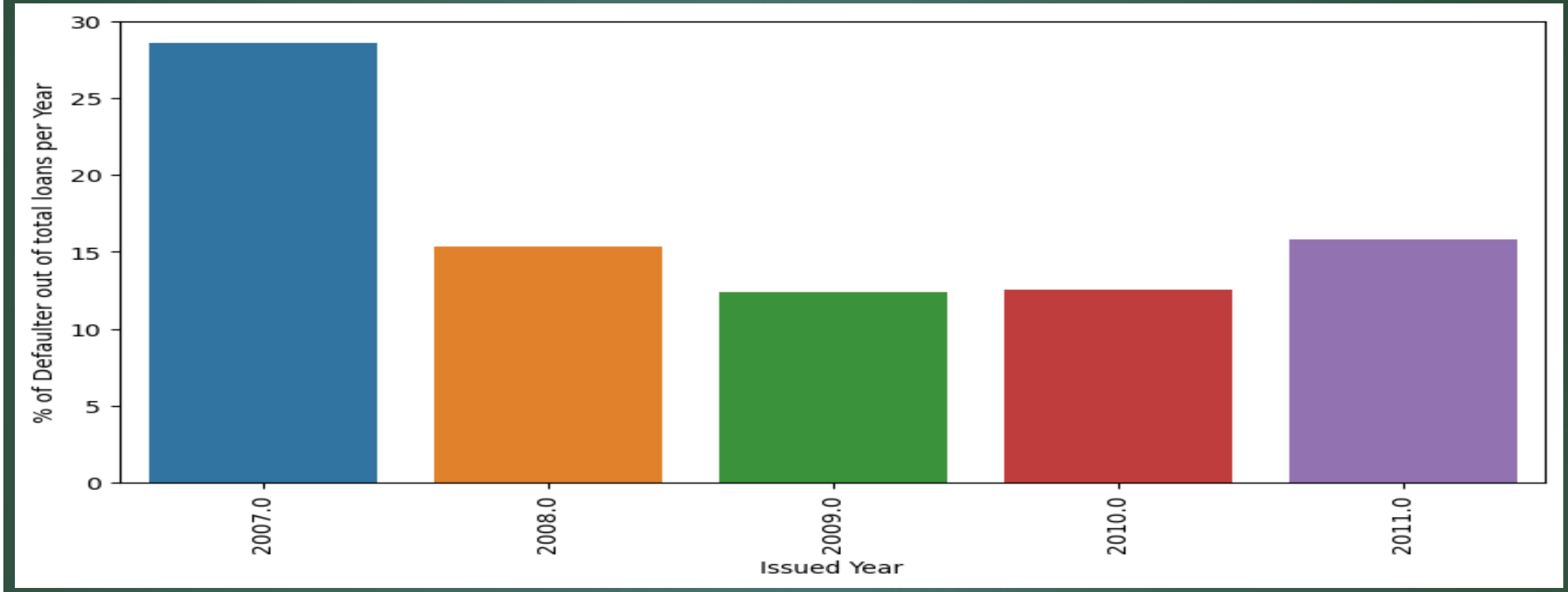


- Count of loans issued in increasing from January to December, with a slight dip in February.
- Most number of Loans are issued in the month of December

Derived Metrix

Percentage of Charge Off Loan Status vs Issued Year

Note:
Percentage default = No of loans defaulted for an issued year divided by Total(defaulted + fully paid) number of loans issued in that year * 100



- More number of defaults observed for loans issued in year 2007
- This could be because of the financial crisis that was observed later in 2008

Recommendations:

1. **Interest Rate :** Loans given at a higher interest rates are defaulted by the applicants frequency. Lending club should restrict the loans to lower rate of interest to avoid loss
2. **Funded Amount Inv**: Loans with higher funded amount inv i.e. >31500$ has a trend to be defaulted by the applicants. Lending club should avoid funding such high value
3. **Grade**: Avoid approving loans with lower Grades F, G
4. **Purpose**: The loan taken for debt consolidation has higher defaulter. Such loans should be evaluated thoroughly and granted with lower value loans
5. **Annual Income:** Applicants with low annual income tend to default more. Lending club should keep threshold on loan values based on the annual income of the applicants
6. **Home Ownership**: Applicants who has rented homes or have mortgages are also more likely to default. Lending club should expedite the loan requests after thoroughly verifying the annual income, credit scores, existing loans etc.

Thank You!!