Problem Statement

We have to analyse the loan structure of a largest online loan provider. This company also facilitates personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

identifying these risky loan applicants, such that loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants' using EDA is the aim of this case study.

Thought Process

The idea is to analyse the data in 3 steps

1. Cleaning the Data extracted from the CSV file
2. Converting the data into favourable format in order for analysis
3. Explaining the results of univariate and bivariate analysis.

Data Cleaning:

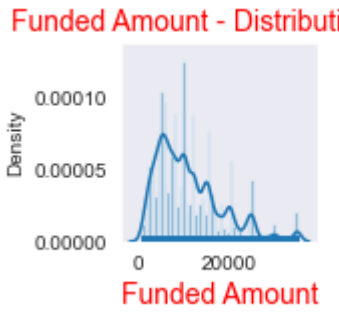
- Identification of data shape
- Identification of NULL columns and dropping them
- Identification of NULL Rows and dropping them
- Removing unnecessary columns further that do not impact our analysis.

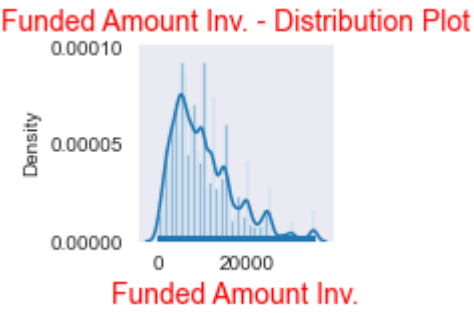
Modification of Data:

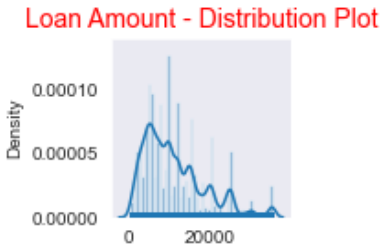
- Converting the data of loan_amnt','funded_amnt','int_rate','funded_amnt_inv','installment','annual_inc','dti','emp_length','total_pymnt' to numeric value
- Removing special characters (‘%, &, -’) from the column data
- Formatting the given time and date format to a single format, so that we can segregate the year and month for better analysis

Data Analysis using univariate and bivariate

1 Let's see distribution of three loan amount fields using distribution plot I.e with respect to Quantitative Variables





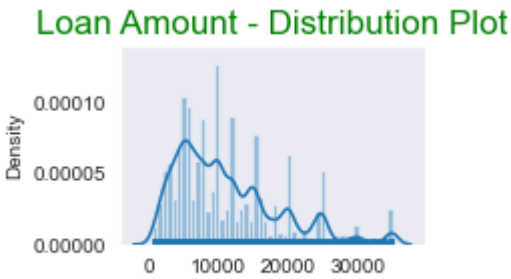


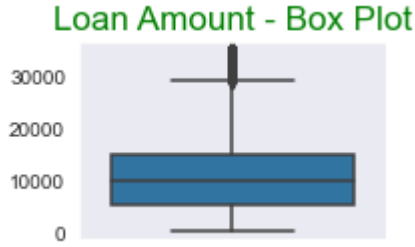
Observation:

Distribution of amounts for all three looks very much similar.

We will work with only loan amount column for rest of our analysis.

2. Univariate Analysis on Loan Amount-Quantitative Variables

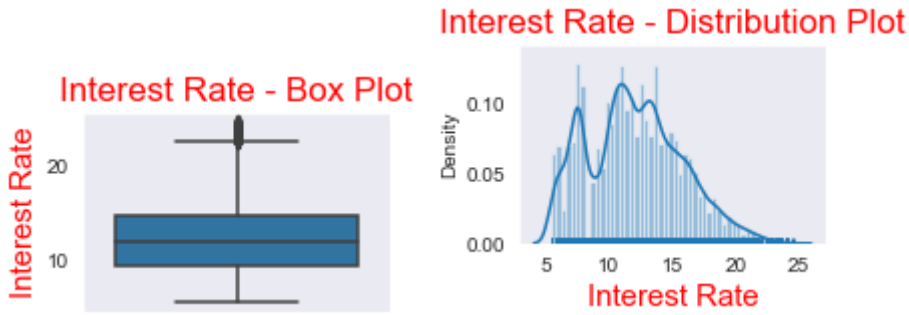




Observations:

Below plots show that most of the Loan amounts are in range of 5000 – 15000

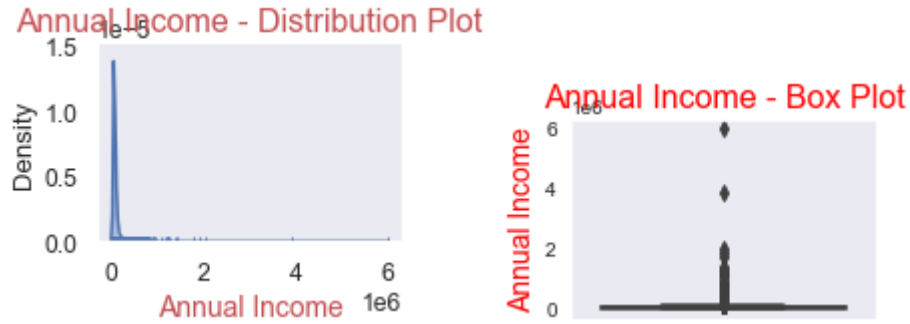
3. Univariate Analysis on Interest Rate-Quantitative Variables



Observations:

The Graph is showing that most of the Interest Rates on loans are in between 10% - 15%

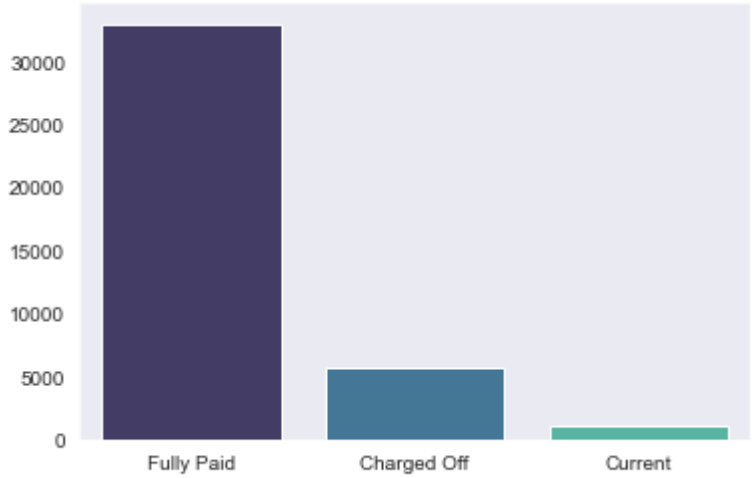
4. Univariate Analysis on Annual Income - Quantitative Variables



Observations:

The Graph is showing that most of the borrower's Annual incomes are in between 40000-80000

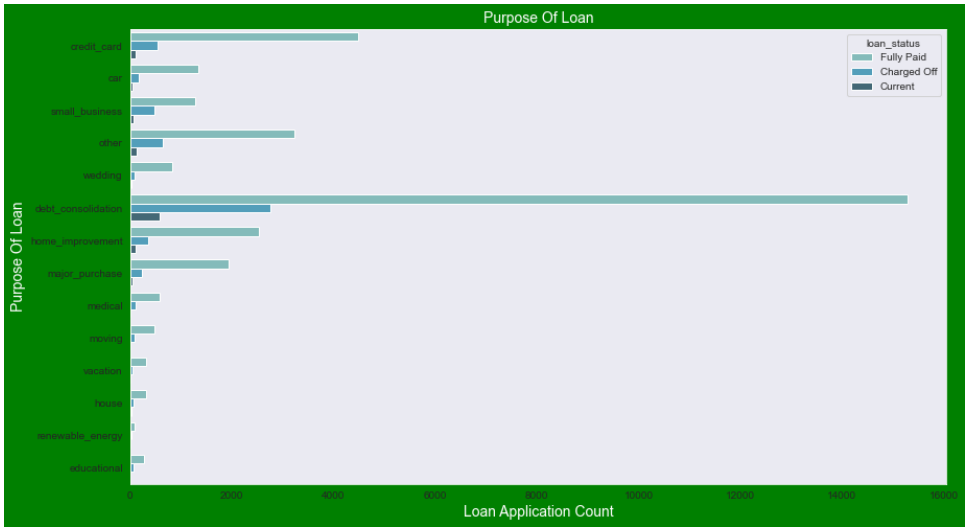
5. Univariate Analysis - Unordered Categorical Variables - Loan Status



Observations:

Graph shows that close to 14% loans were charged off out of total loan issued.

6. Bivariate Analysis on purpose of Loan against Charge off Proportion.

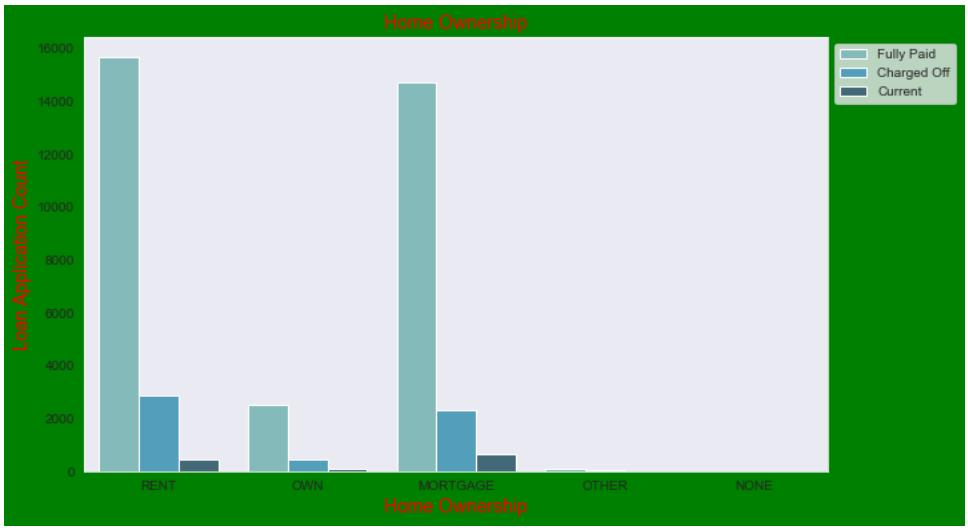


Observations:

Graph shows that most of the loans were taken for the purpose of debt consolidation & paying credit card bill.

Number of charged off count also high too for these loans.

7, Univariate Analysis - Unordered Categorical Variables - Home Ownership

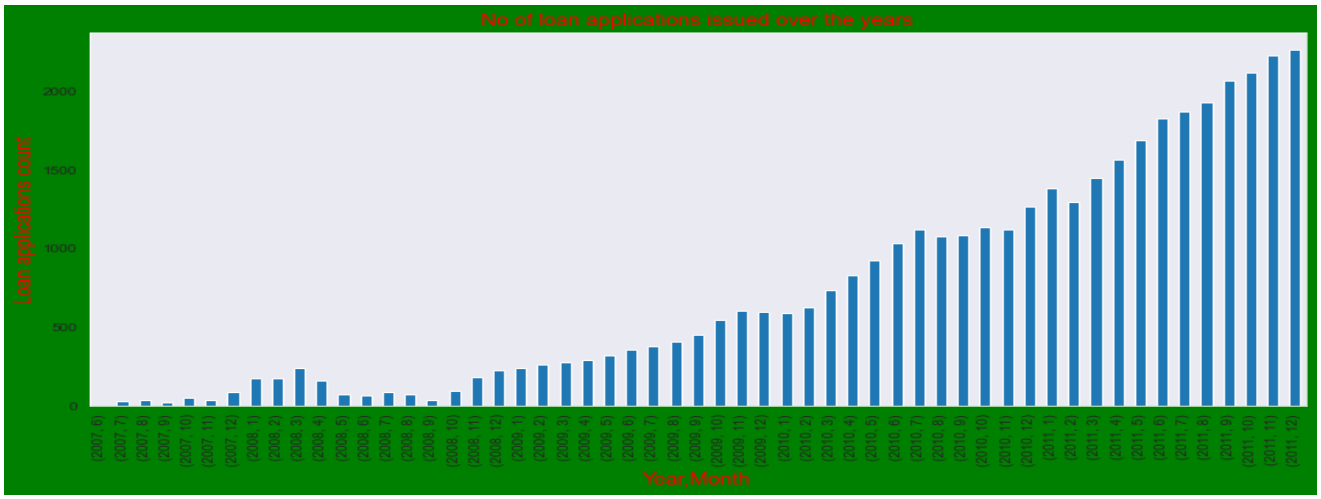


Observations:

Graph shows that most of them living in rented home or mortgaged their home.

Applicant numbers are high from these categories so charged off is high too.

8.Derived Column - Ordered Categorical Variables Let us look into number of loans which were approved every year/month Let's use derived column year to check pattern of loan issuing over the years.



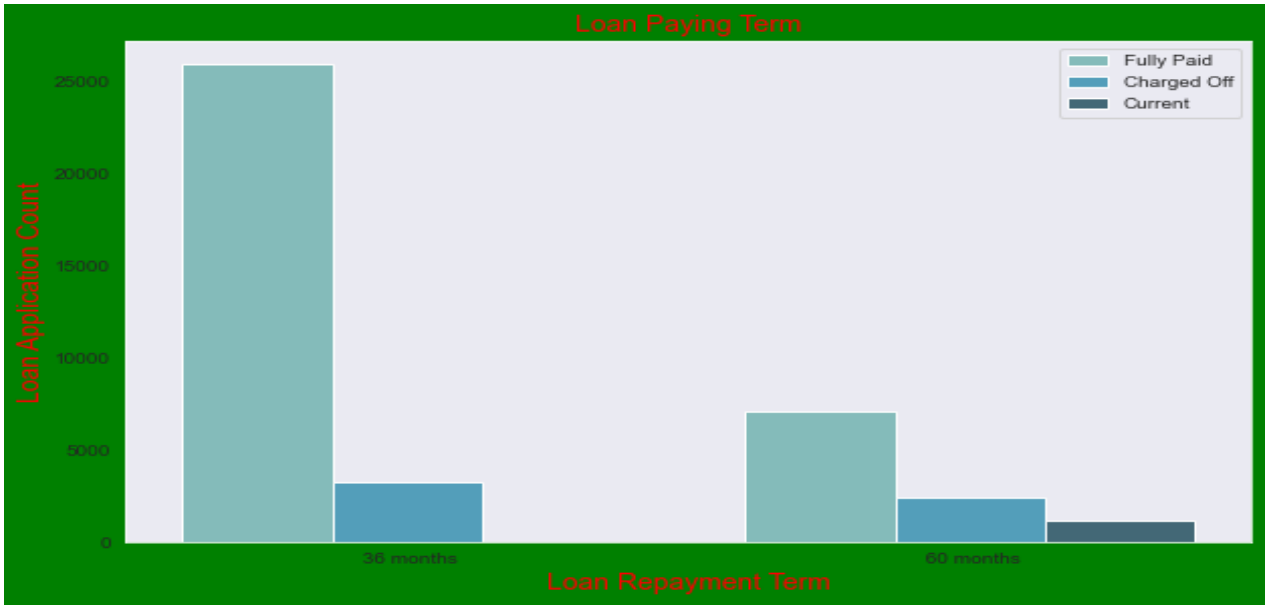
Observation

Count of loan application is increasing every passing year.

So, increase in number of loan applications are adding more to number of charged off applications.

Number of loans issued in 2008(May-October) got dipped, may be due to Recession.

9. Univariate Analysis - Ordered Categorical Variables- Loan Paying Term

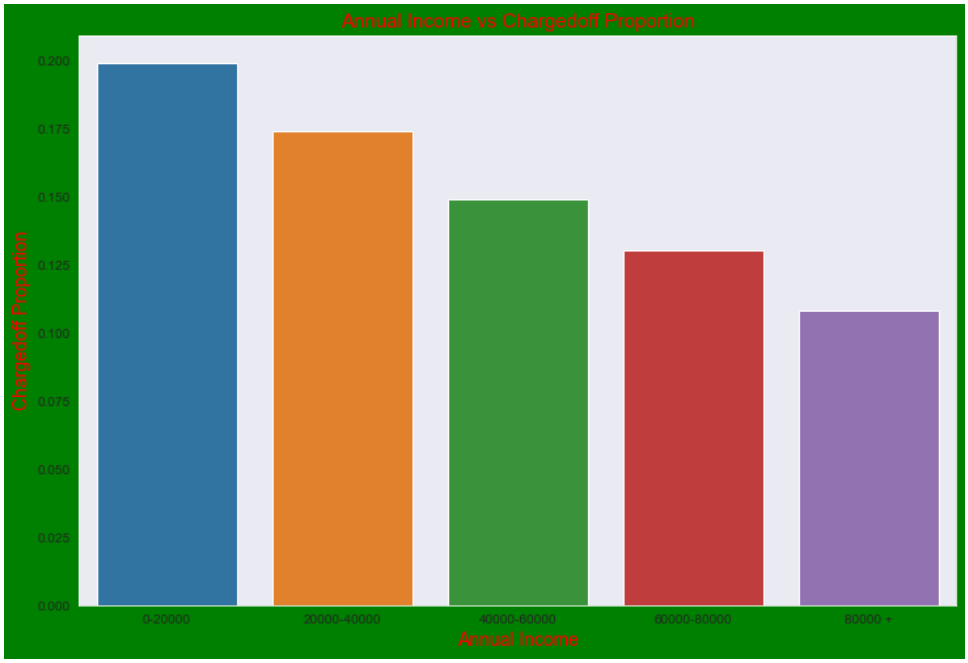


Observations:

Plot shows that those who had taken loan to repay in 60 months had more % of number of applicants getting

Charged off as compared to applicants who had taken loan for 36 months.

10. Bivariate Analysis on annual income against Chargedoff_Proportion -

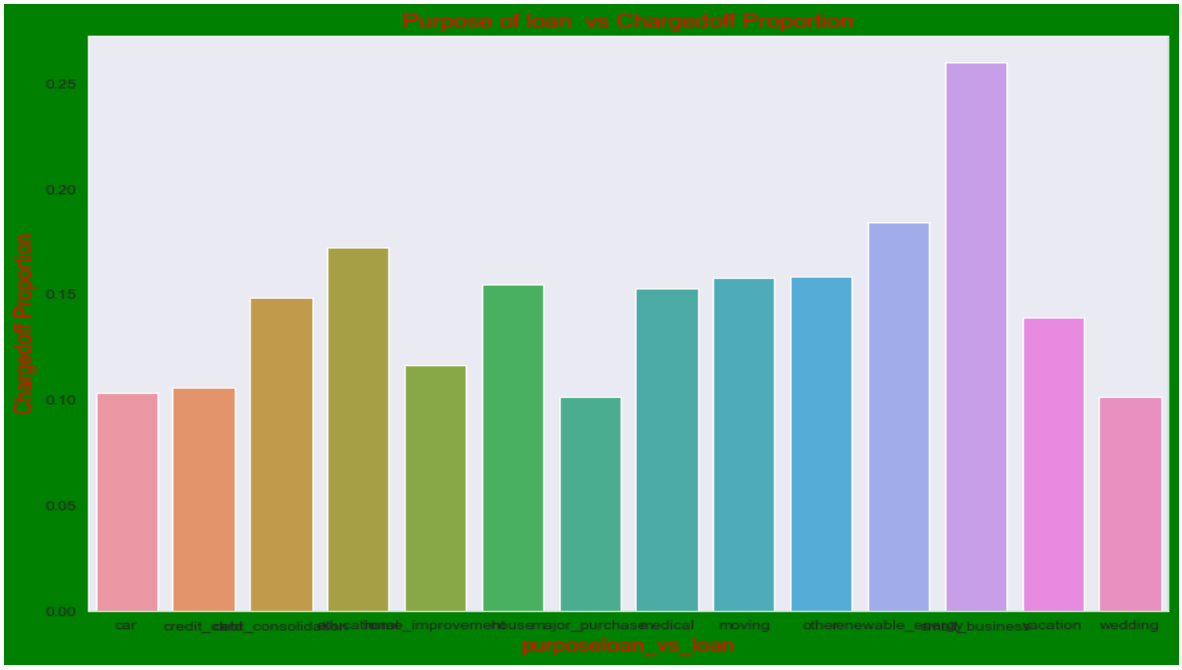


Observations:

Income range 80000 and more has less chances of charged off.

Income ranges from 0 to 20000 has high chances of charged off.

11. Bivariate Analysis on purpose of Loan against Chargedoff_Proportion.

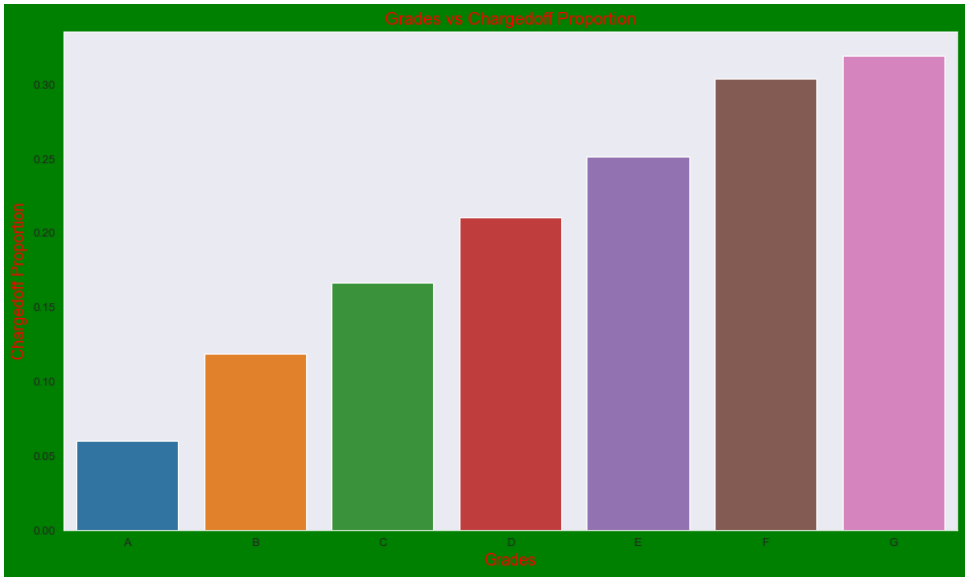


Observations:

Small Business applicants have high chances of getting charged off.

Renewable energy was charged off proportion is better as compare to other categories.

12. Bivariate Analysis on grade against Charged off_Proportion

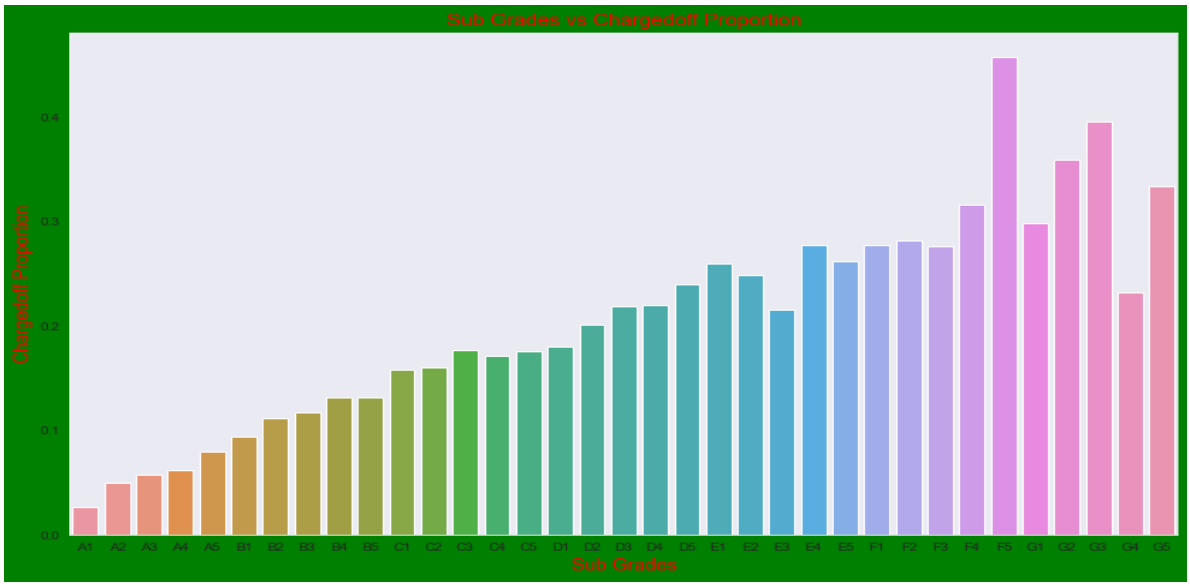


Observations:

Grade "A" has very less chances of charged off.

Grade "F" and "G" have very high chances of charged off.

13.Understanding grades better using Bivariate Analysis on sub grade against Chargedoff_Proportion



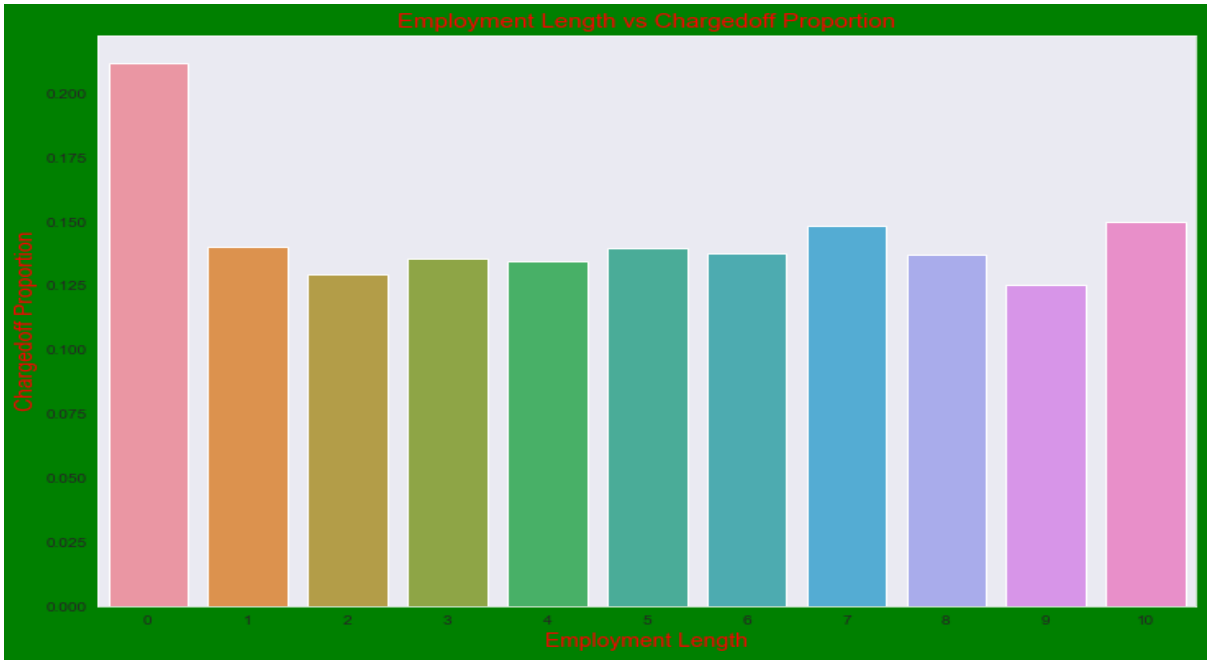
Observations:

Sub Grades of "A" has very less chances of charged off.

Sub Grades of "F" and "G" have very high chances of charged off.

Proportion of charged off is increasing with sub grades moving from sub grades of "A" towards sub grades of "G"

14. Bivariate Analysis on employment length against Chargedoff_Proportion



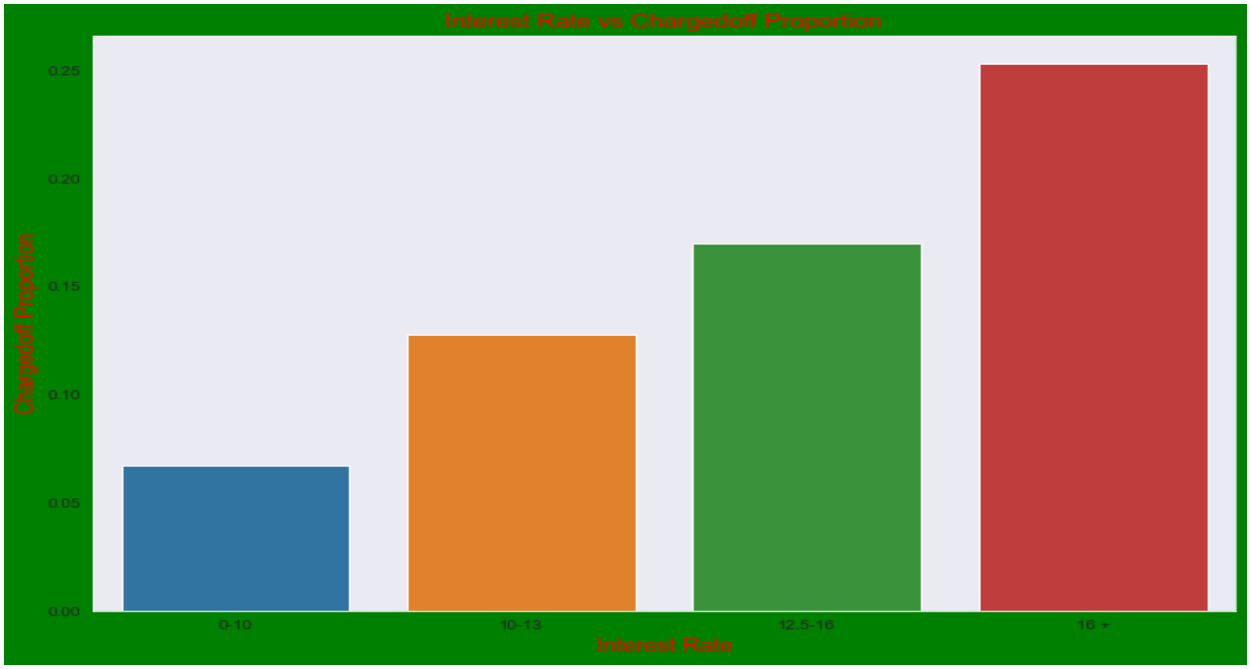
Observations:

Those who are not working or have less than 1 year of work experience have high chances of getting charged off.

It makes sense as with less or no experience they don't have source of income to repay loan.

Rest of the applicants have more or less same chances of getting charged off.

15. Bivariate Analysis on interest rate against Chargedoff_Proportion



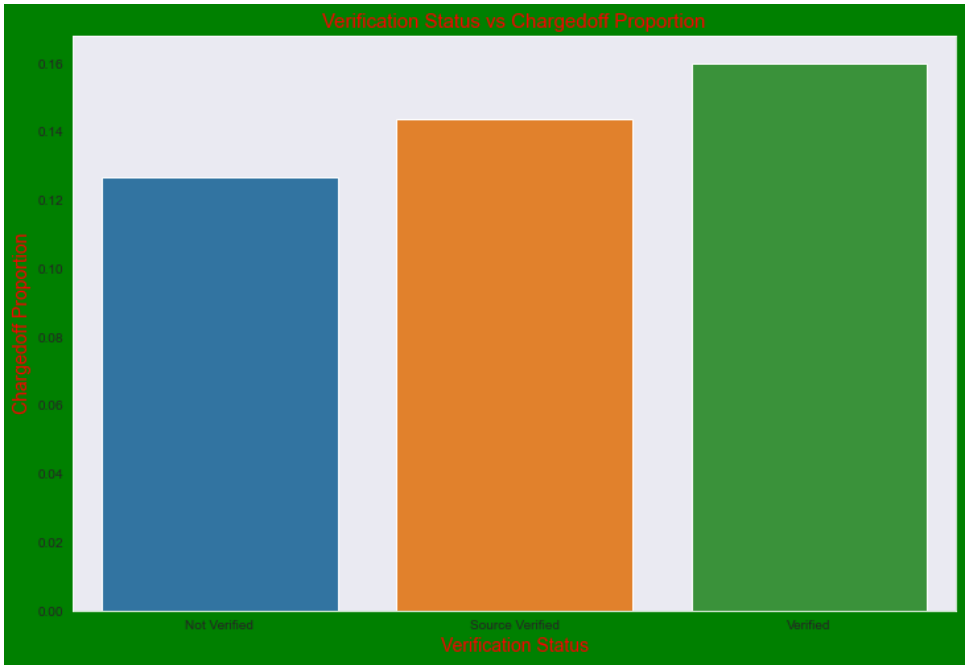
Observations:

Interest rate less than 10% has very less chances of charged off. Interest rates are starting from minimum 5 %.

Interest rate more than 16% has good Chances of charged off as compared to other category interest rates.

Charged off proportion is increasing with higher interest rates.

16. Bivariate Analysis on verification status against Chargedoff_Proportion



Observations:

There is not much difference in charged off proportion.

This variable doesn't provide any insights for charged off.

17. Bivariate Analysis on Derogatory Public Record against Chargedoff_Proportion

Observations:

A derogatory item is an entry that may be considered negative by lenders because it indicates risk and hurts

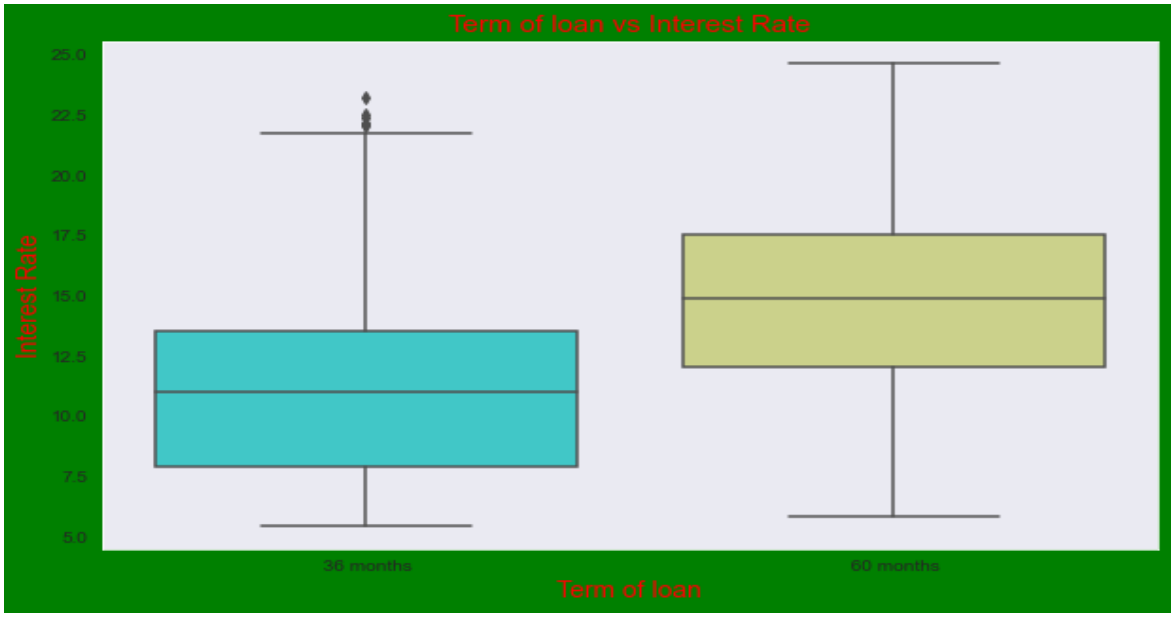
Your ability to qualify for credit or other services. Public records and collections are derogatory items

Because they reflect financial obligations that were not paid as agreed.

Those who already have pub_rec value 1 or 2 have charged off chances higher than who have no Derogatory Public Record.

pub_rec count 3-4 has less numbers so cannot reach on any conclusions.

18. Bivariate Analysis - Term of loan vs Interest Rate

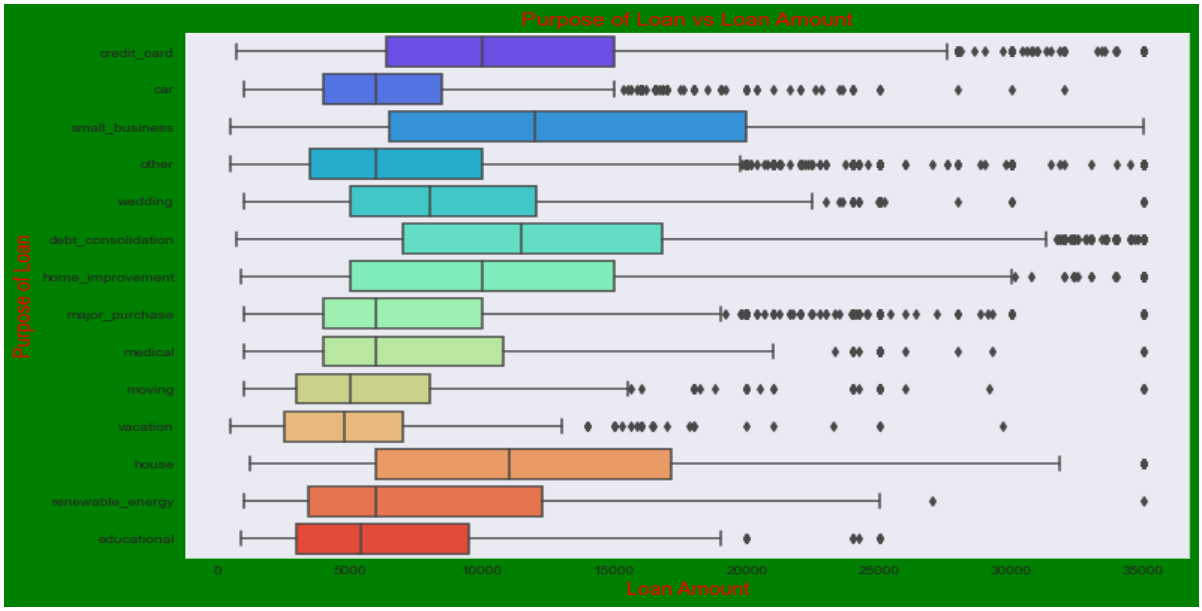


Observations:

It is clear that average interest rate is higher for 60 months loan term.

Most of the loans issued for longer term had higher interest rates for repayment.

19. Bivariate Analysis - Purpose of loan vs Loan amount

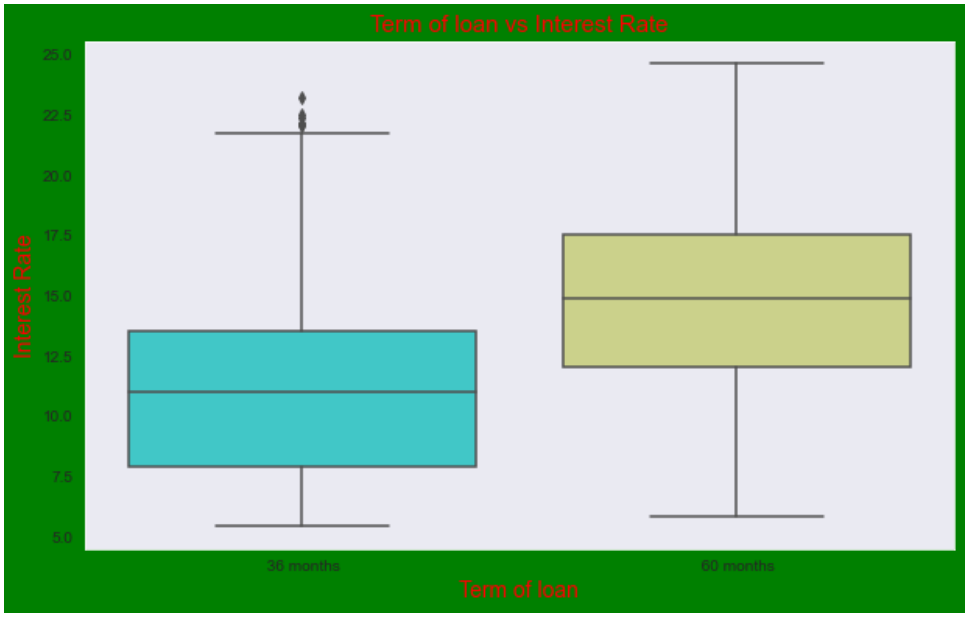


Observations:

Median,95th percentile,75th percentile of loan amount is highest for loan taken for small business purpose among all purposes.

Debt consolidation is second and Credit card comes 3rd.

20. Bivariate Analysis - Term of loan vs Interest Rate

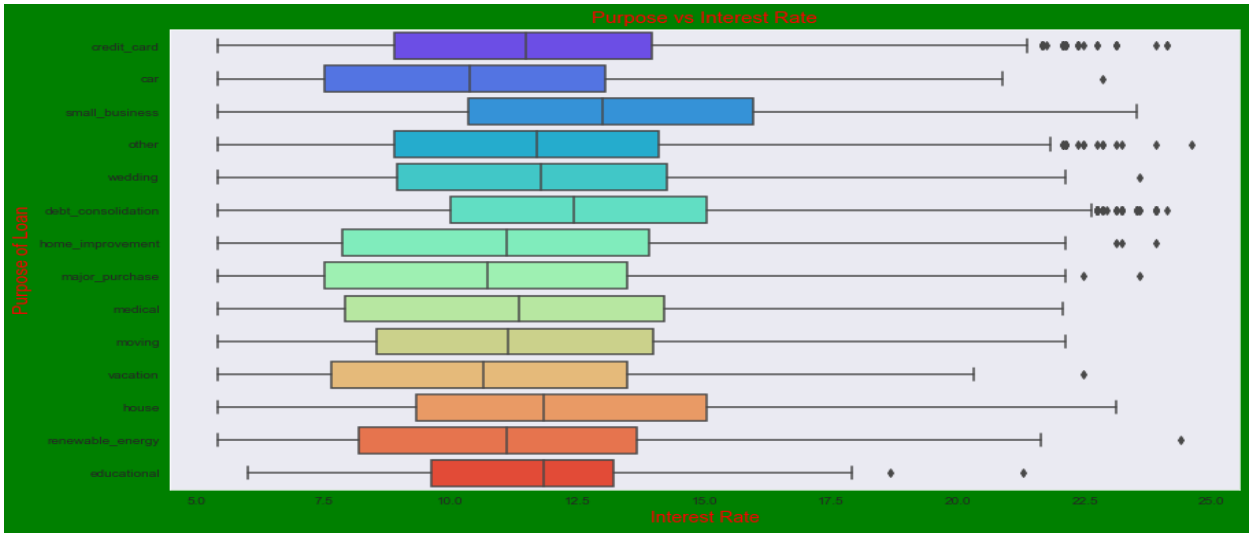


Observations:

It is clear that average interest rate is higher for 60 months loan term.

Most of the loans issued for longer term had higher interest rates for repayment

21. Bivariate Analysis - Purpose vs Interest Rate



.

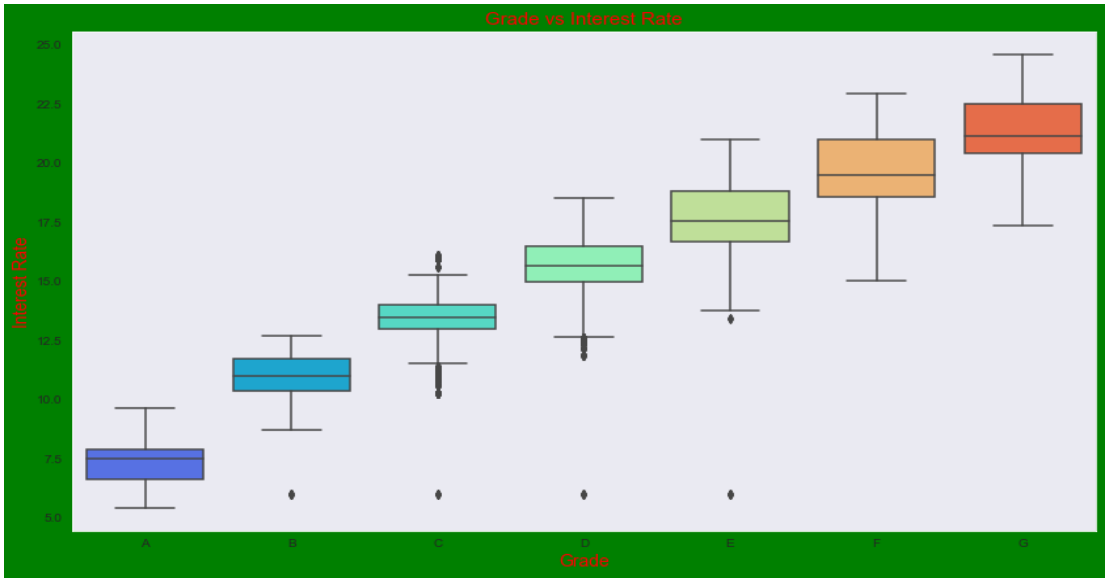
Observations:

It is clear that average interest rate is highest for small business purpose.

Loans taken for small business purposes had to repay the loan with more interest rate as compared to other.

Debt consolidation is 2nd where borrowers had to pay more interest rate.

22. Bivariate Analysis - Grade vs Interest Rate



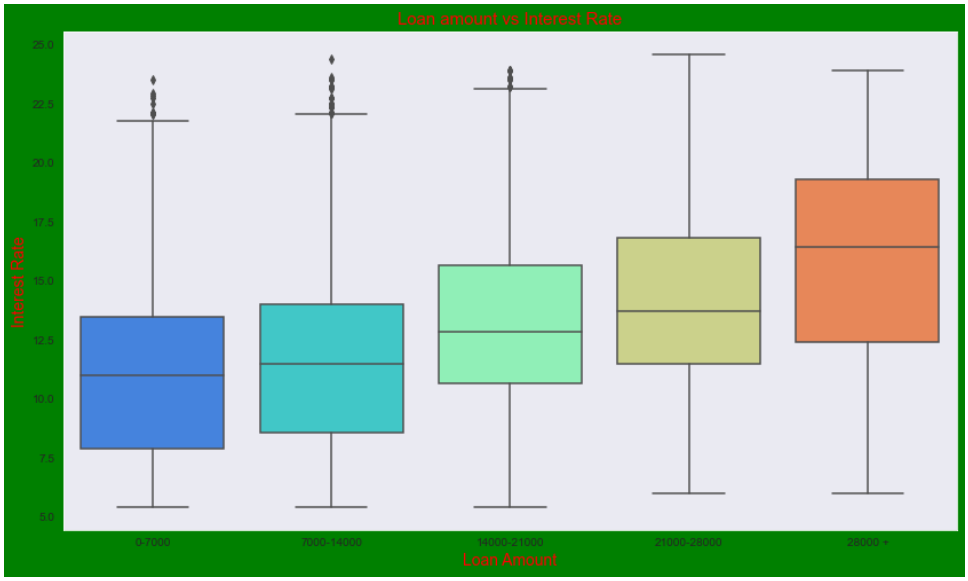
Observations:

A-grade is a top letter grade for a lender to assign to a borrower.

The higher the borrower's credit grade, the lower the interest rate offered to that borrower on a loan.

It is clear that interest rate is increasing with grades moving from A to F.

23. Bivariate Analysis - Loan Amount vs Interest Rate



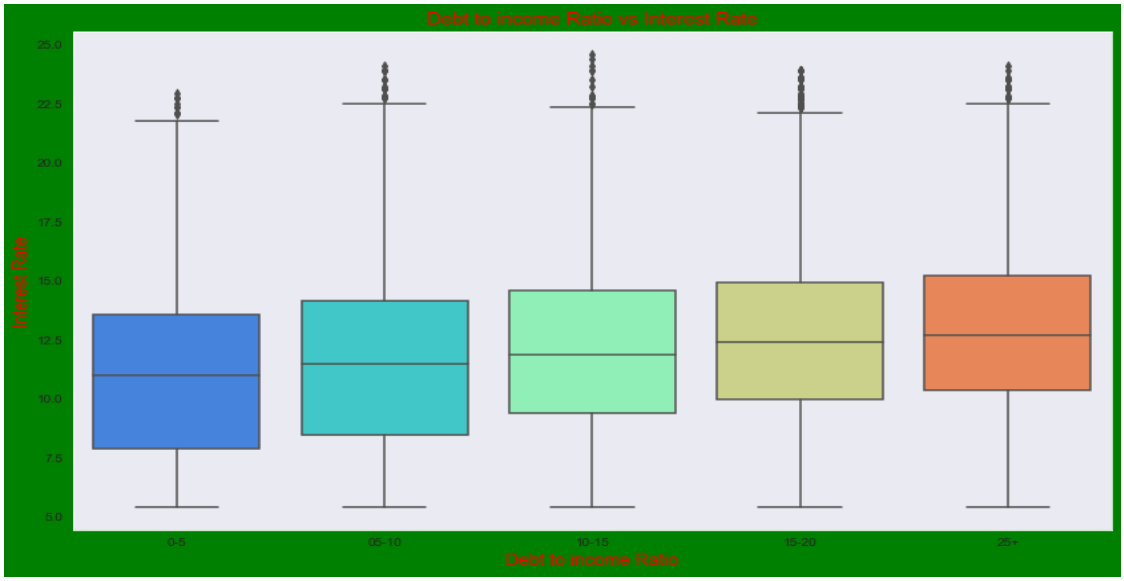
Observations:

It is clear that interest rate is increasing with loan amount increase.

Probably when loan amount is more it is taken for longer loan term, we saw earlier that longer the loan term more the

Interest rate.

24. Bivariate Analysis - dti vs Interest Rate

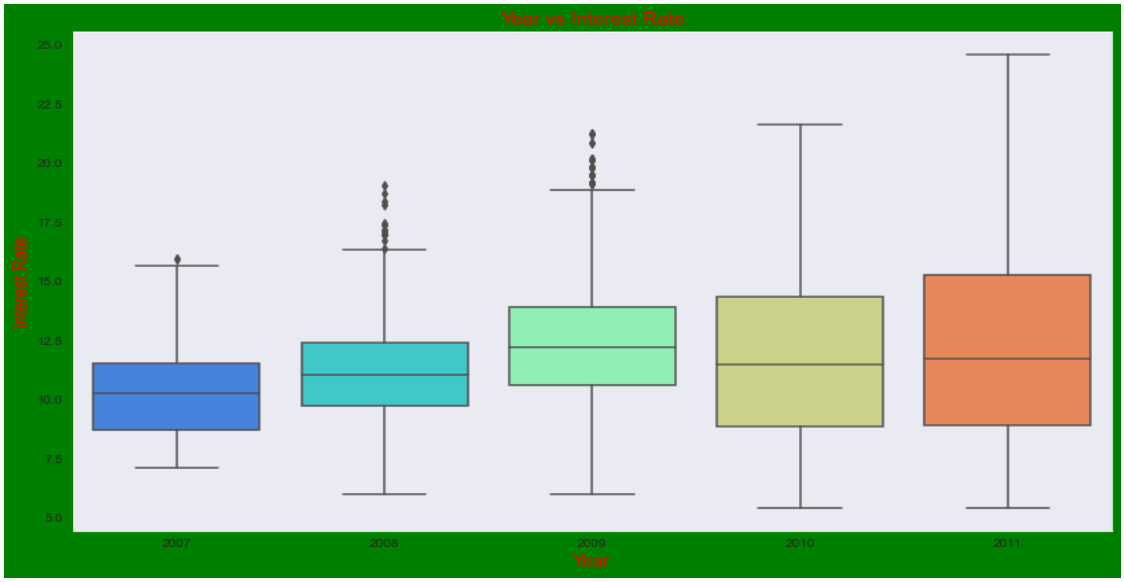


Observations:

If your DTI is low enough you may get a lower interest rate.

Plot shows no significant variation but there is slight increase in interest rate with increase in DTI.

25. Bivariate Analysis - year vs Interest Rate



Observations:

Plot shows interest rate is increasing slowly with increase in year.