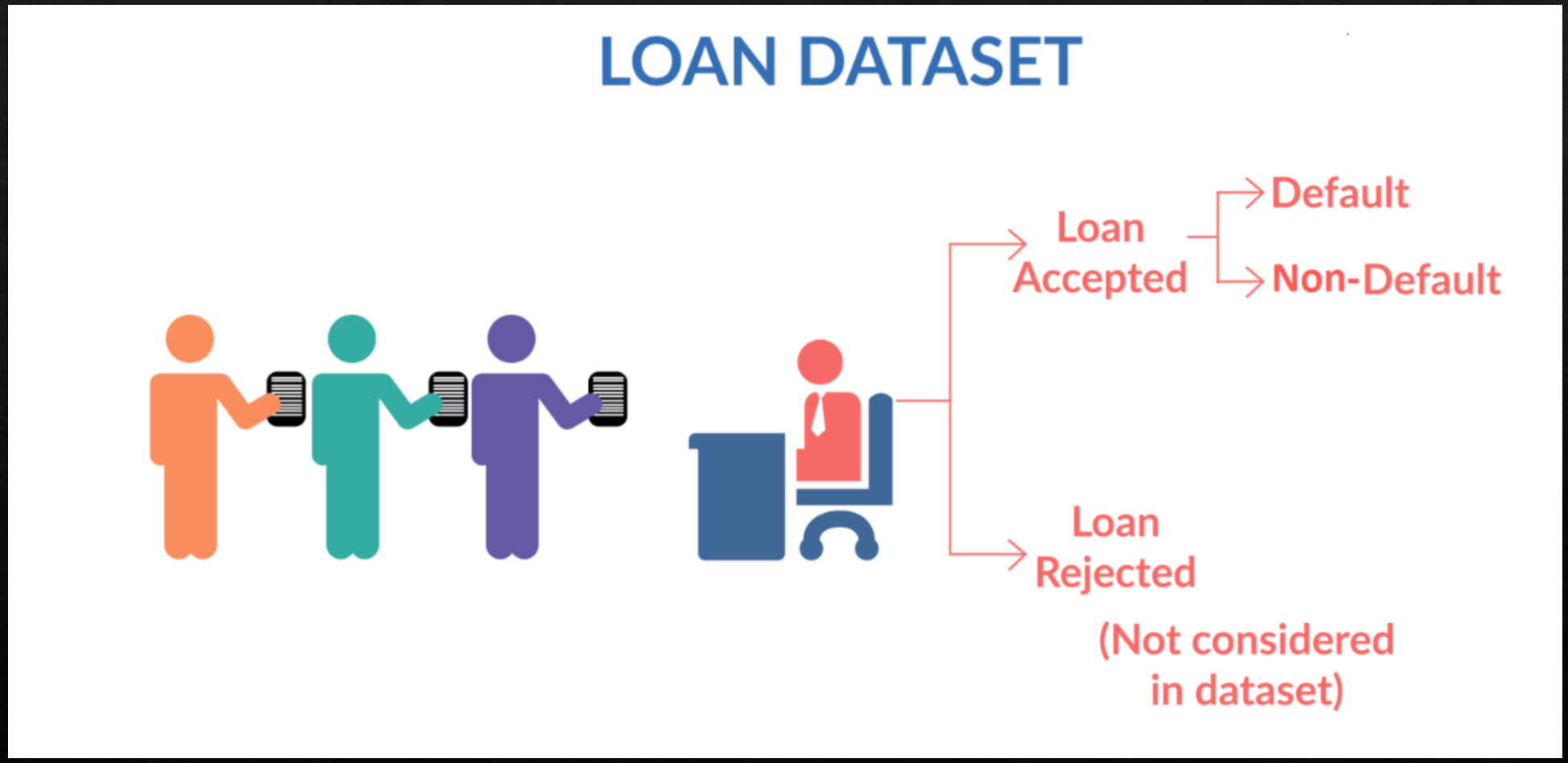
Lending Club Case Study

Suman Sourav Sahoo

Problem Statement

- 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

The Loan Dataset



- When a person applies for a loan, there are two types of decisions that could be taken by the company:
  - **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
    1. Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
    2. Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
    3. Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan
  - **Loan rejected:** The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Business Objective

- This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.
- Lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters’.
- **The Business Objective of this study is to identify Driving Factors (or driver variables) which are strong indicators of default and potentially use the insights in Approval/ Rejection decision making.**
- The Company can utilize this information for its Portfolio and Risk Assessment and can reduce the Risky Loan Approvals and hence Cutting Down the amount of its Credit Loss or Financial Loss.

Exploratory Data Analysis(EDA)

- We have been provided with one Excel File (loan) contains the complete loan data for all loans issued through the time period 2007 to 2011 and the Data Dictionary of the Data Fields, which describes the meaning of each field in the Loans Data File.
- The data given contains the information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicates if a person is likely to default, which can be used by the Lending Club for taking actions such as listed below or more :

    1. Denying the loan
    2. Reducing the amount of loan
    3. Lending (to risky applicants) at a higher interest rate

- EDA is used to understand how consumer attributes and loan attributes influence the tendency of default

Understanding the Given Data Set

- The Data Set which provided consists 3 types of data :-

  I. Customer (Applicant) Demographic Data (Applicant's Specific / Personal information)

  II. Loan Specific Data (Loan Information)

  III. Customer Behavioural Data (Once the Loan is Approved/ Granted)

Types Of Data

Customer Demographic Data :	Loan Specific Data :	Customer Behavioural Data :
• Employment Title • Employment Length • Annual Income • Description • Zip Code • State Address • Home Ownership	• Loan Amount • Funded Loan Amount • Loan Interest Rate • Issue Date • Loan Status • Grade • Term	• No. of Delinquent Accounts • Delinquency 2yrs. • Earliest Credit Line • Instalment • Revolving Balance • Recoveries • Application Type • Loan Purpose

Data Analysing & Cleaning Strategy

- Fully Empty Columns to be Removed
- Columns with values have no relevance with the Business Objective to be Removed
- Columns have only 1 value or all values unique for all rows, and especially a string columns, then it need to be Removed as we can not categories in groups
- Columns will few nulls will be imputed with Median if the column have any significance in the analysis else will get Removed
- Some of the Categorical Columns Values need to be Corrected for the ease of Analysis
- Some Columns need to be derived from existing Columns for doing the Analysis

The Final Data Set

- The Initial Data Set was with a size of 39717 Rows and 111 Columns.
- The Final Data Set is having a count of 39319 Rows and 46 Columns
- Total of 65 fully empty columns were removed
- Out of some Partial Null Columns kept the required 1 column which was imputed with the median and removed the rest
- Customer Behavior Columns are removed as they have no relevance in the Business Objective
- Cleaned and standardized the data of some columns like emp_length , term, interest_rate etc.
- Date Types was corrected to Date Columns
- Some additional Columns like year, month, were derived and added to analyze the Business Objective

Data Analysis & Variable Identification

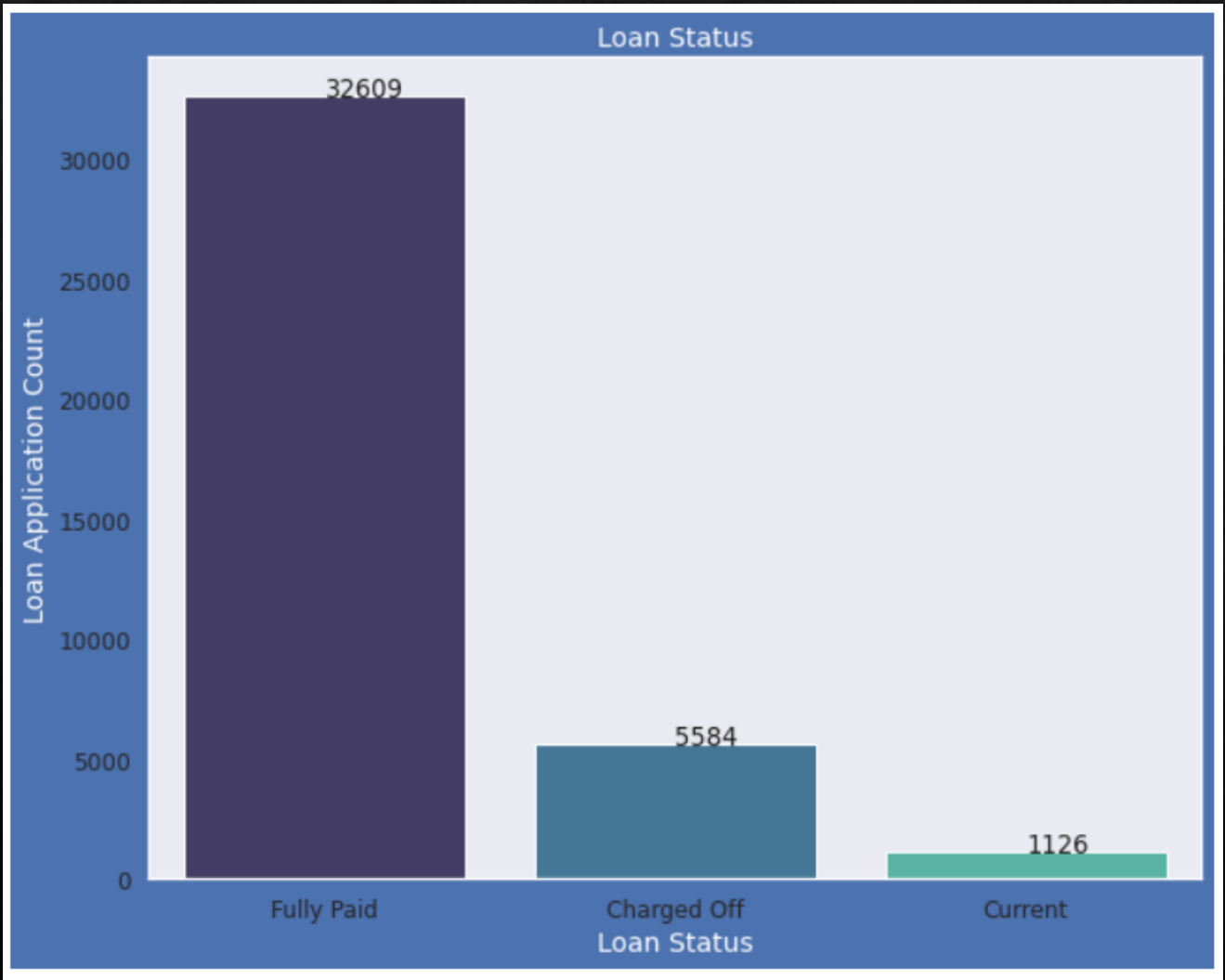
- The target Variable Identified as Loan Status
- The dataset has some categorical variables and continuous Numeric Variables
- The categorical variables identified are: term, emp_length, grade, sub_grade, home ownership, and verification status
- Some continuous variables like loan_amount, funded_amount, inter_rate etc. are binned and create new Categorical Variables
- The variable purpose is used for sector analysis

Data Visualization

- The records taken for Data Analysis and Visualization are the Fully Paid and Defaulters list as the records with Current status is about ongoing transactions and so we can not used them for a history analysis which is the Business Objective.

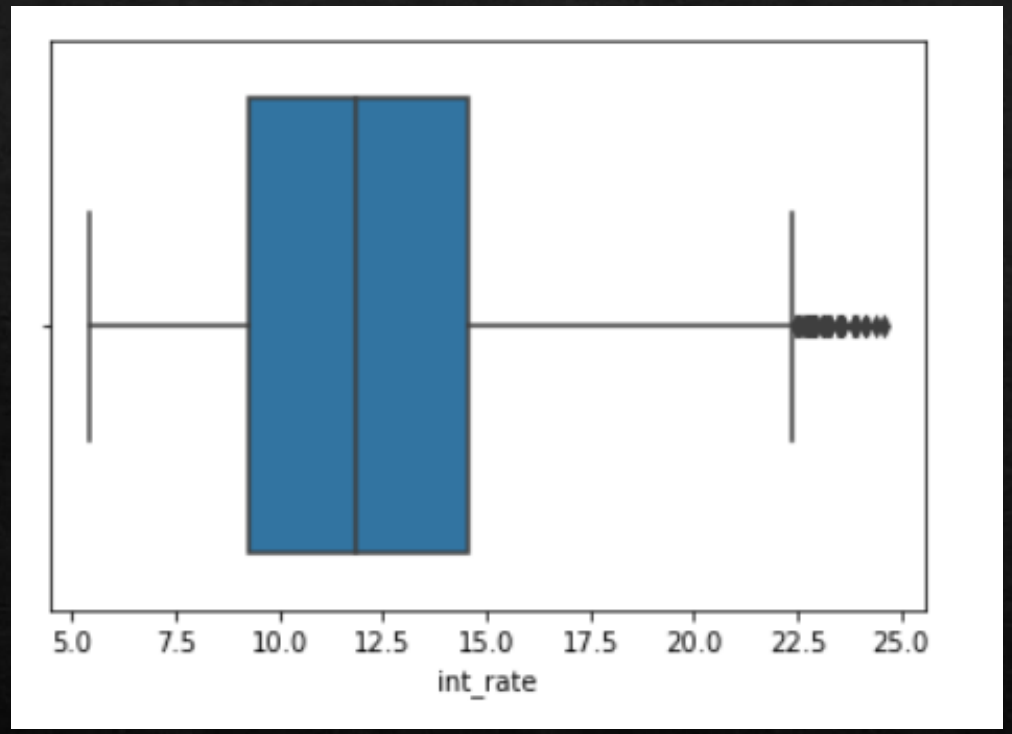
- The Calculated defaulter Status Rate is close to 14%.
- Record Count based on loan –
  - Fully paid – 32609
  - Charged off – 5584

Univariate Analysis

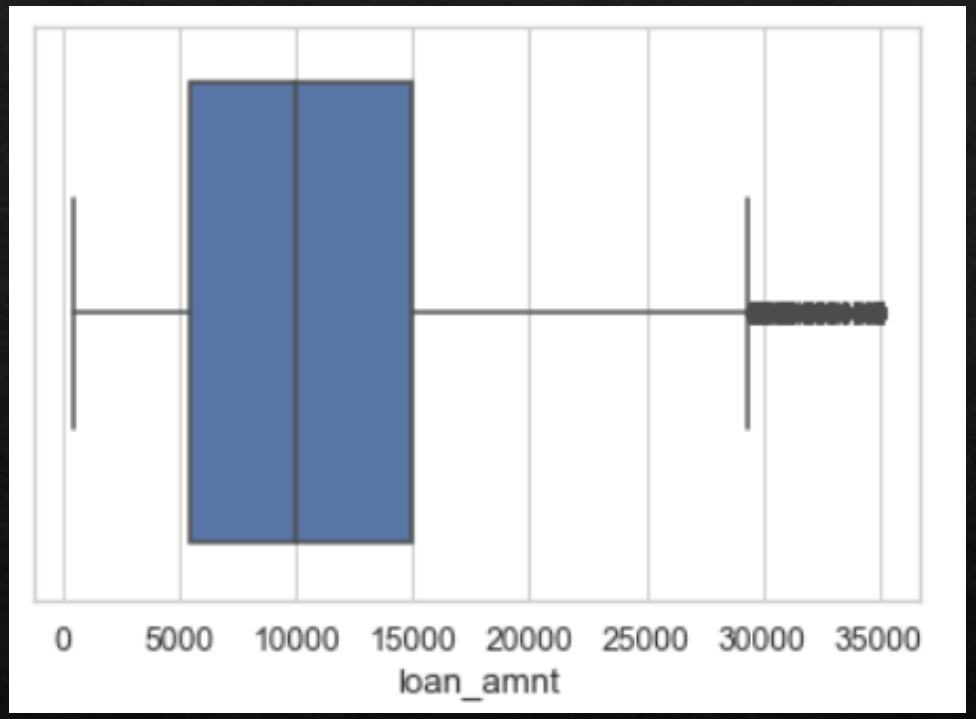


Close to 14% loans were charged off out of total loan issued

Univariate Analysis



Average interest rate is 12 %.
And after 75% percentile interest rate increased to 25% from 15%.



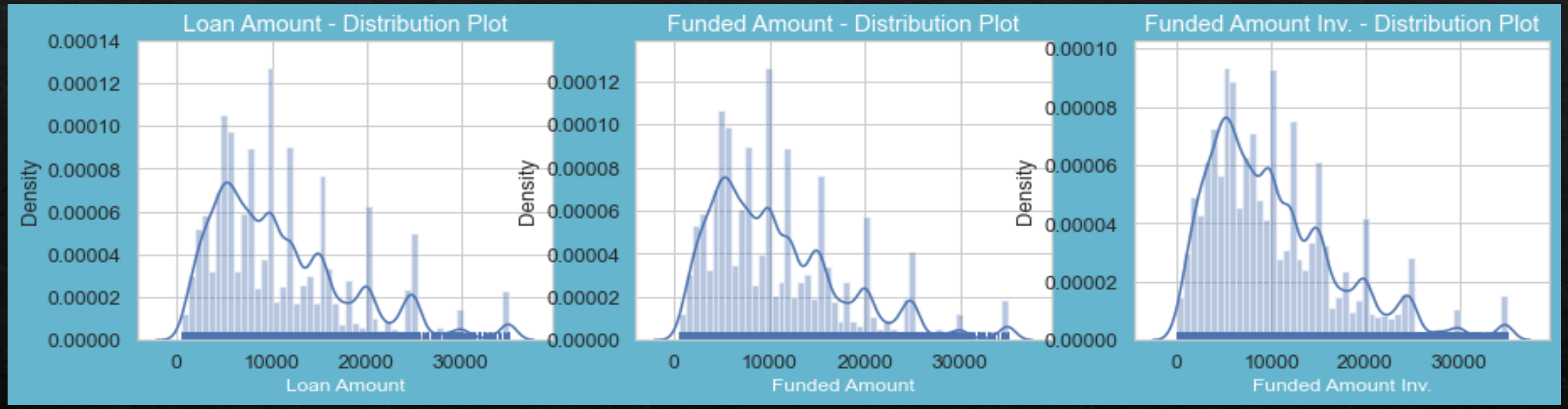
Loan amount column using Box plot: most of the loan_amounts are in range of 5000 - 15000

Bivariate Analysis : Correlation Matrix

- Loan amount, investor amount, funding amount are strongly correlated
- Annual income with DTI(Debt-to-income ratio) is correlated negatively(that is if annual_income is high, dti(Debt-to-income) ratio is low & vice versa)
- Positive correlation between annual income and employment years, that means the income increases with work experience

Univariate Analysis

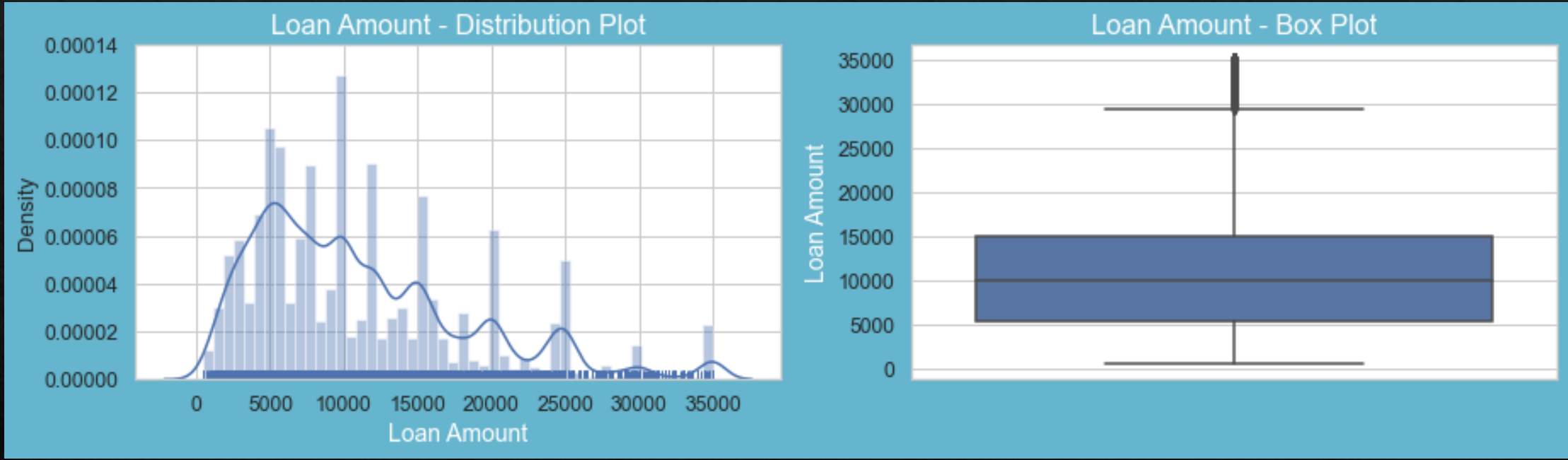
Distribution of the three loan amount fields



- Observation:
  - Distribution of amounts for all three looks very much similar.
  - We will work with only loan amount column for rest of our analysis.

Univariate Analysis

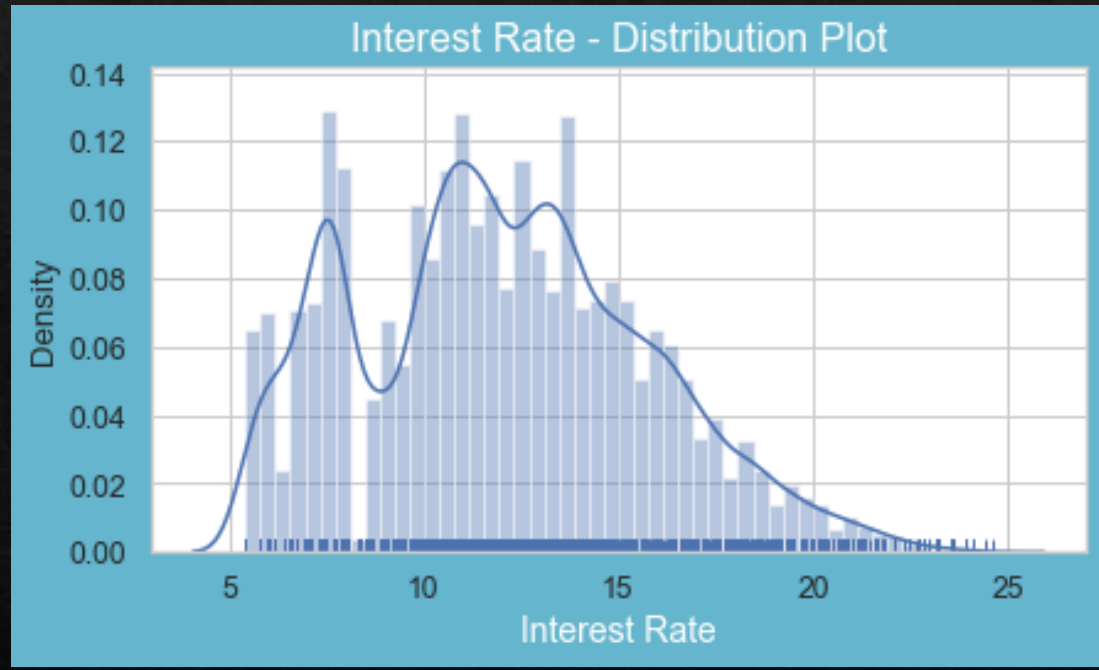
Univariate Analysis using Loan Amount

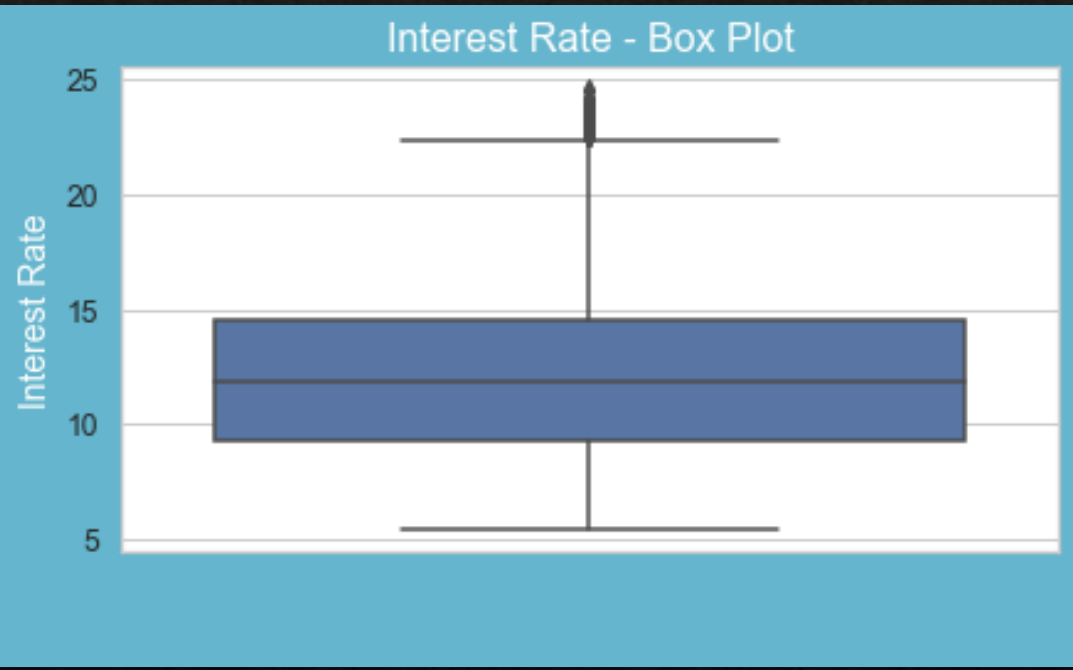


- Observations :
  - From both the plots above, we can know that most of the loan_amounts are in range of 5000 - 15000

Univariate Analysis

Univariate Analysis using Interest Rate

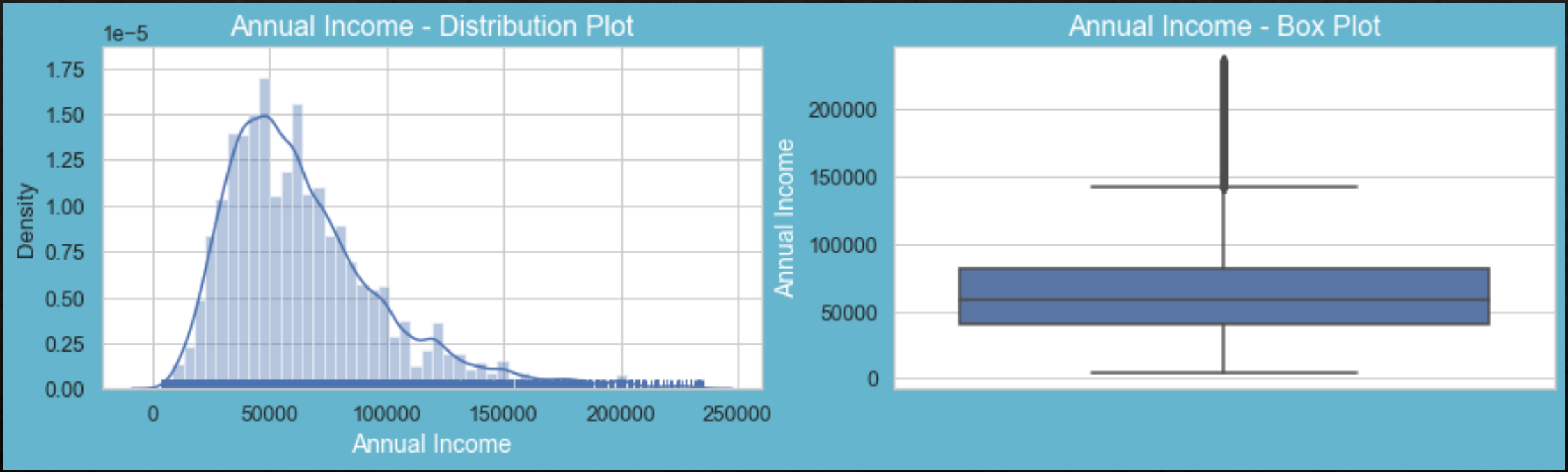




- Observations :
  - From both the plots above, we can know that, most of the Interest Rates on loans are in range of 10% - 15%

Univariate Analysis

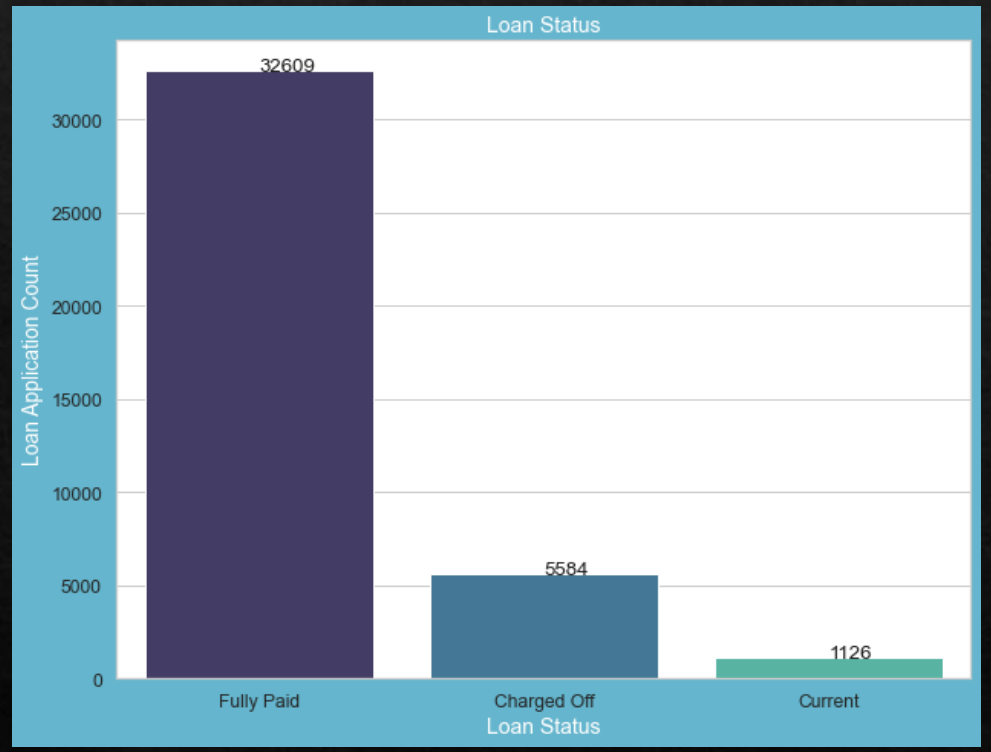
Univariate Analysis using Annual Income



- Observations :
  - The above 2 plots show that, most of the borrower's Annual incomes are in the range of 40000 - 80000.

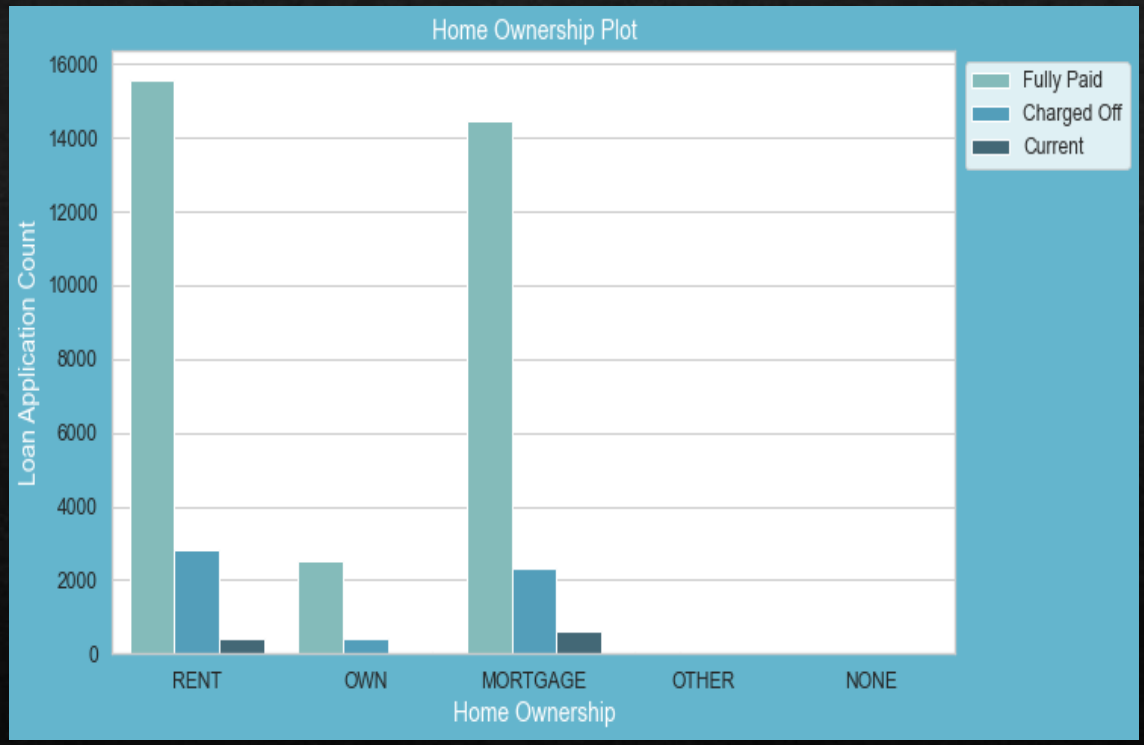
Univariate Analysis

Univariate Analysis - Unordered Categorical Variables - Loan Status



- Observation :
  - The above plot shows that close to 14%(5584 loan applicants) loans were charged off out of total loans issued.

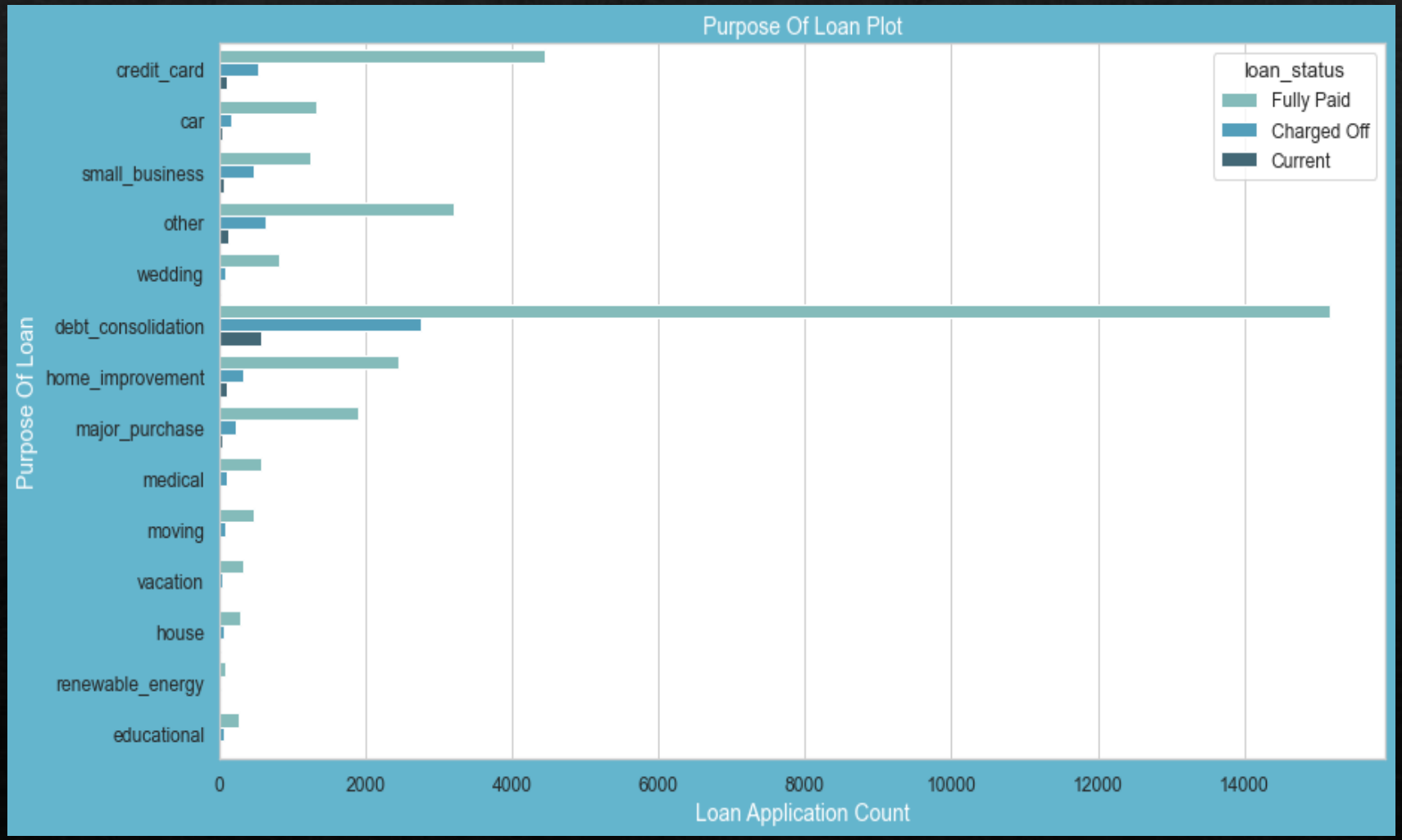
Univariate Analysis - Unordered Categorical Variables - Home Ownership



- Observation :
  - The above plot shows that, most of them living in rented home or in mortgaged home.
  - Applicant numbers are high from these 2 categories so charged-off is high too.

Univariate Analysis

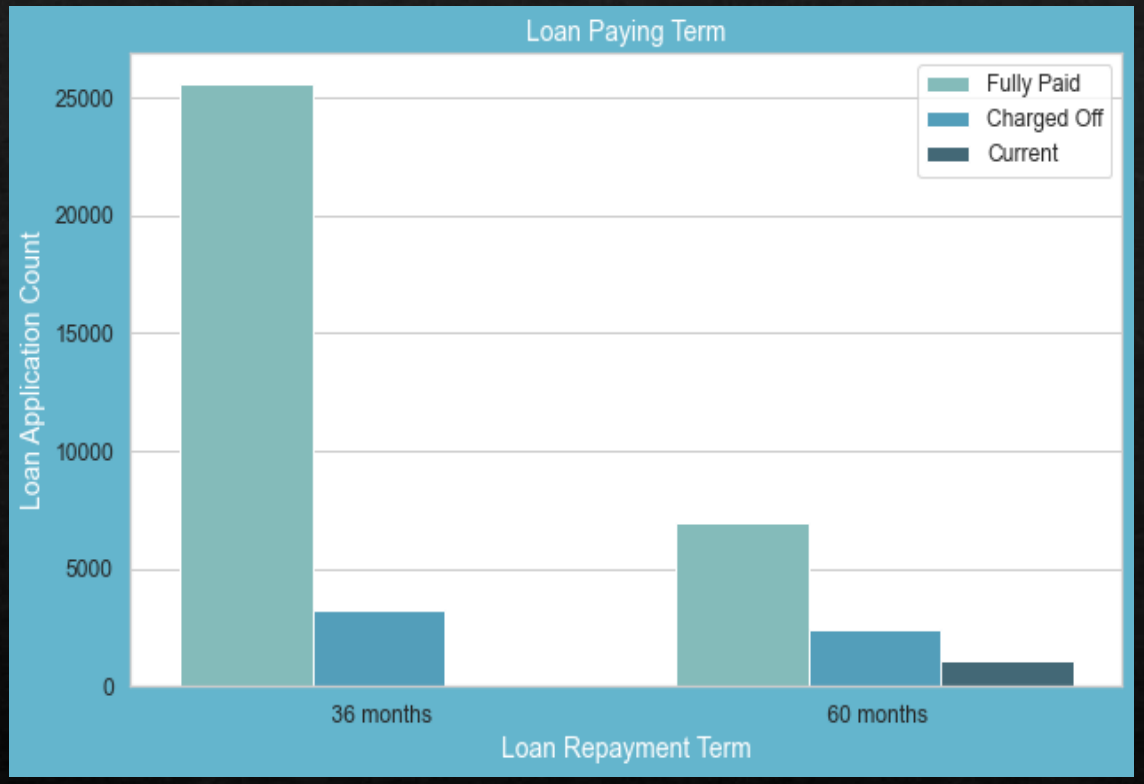
Univariate Analysis - Unordered Categorical Variables - Purpose Of Loan



- Observation :
  - The above plot shows that, most of the loans were taken for the purpose of 'debt consolidation' & 'paying credit card bill'.
  - Number of charged off count is also high too for these loans.

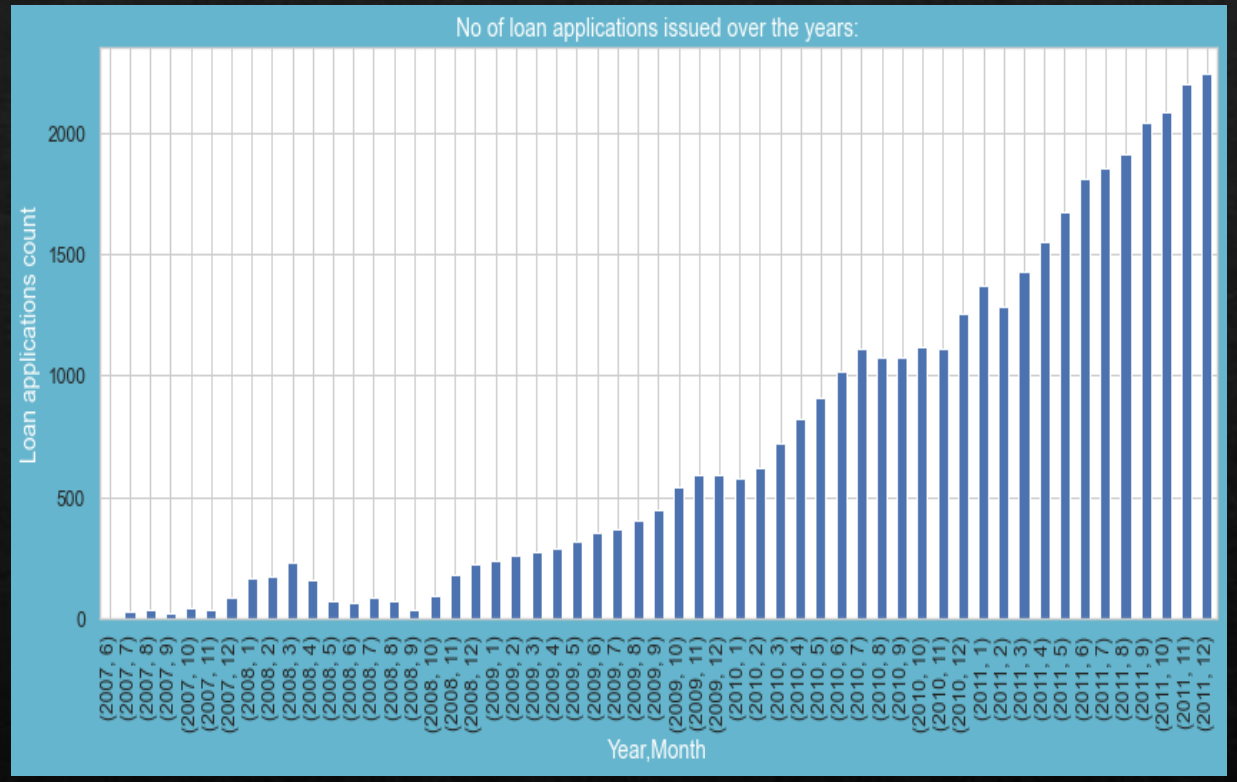
Univariate Analysis

Ordered Categorical Variables- Loan Paying Term



From above plot, applicants who took 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.

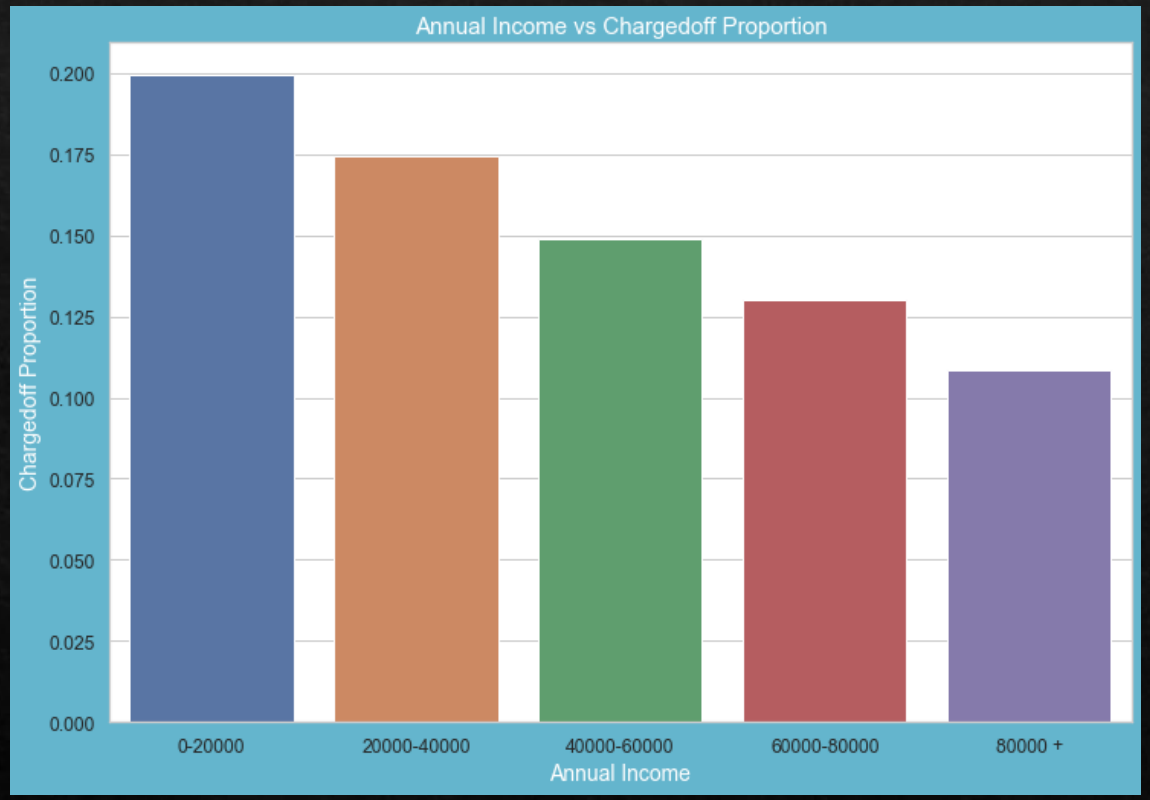
Derived Column - Ordered Categorical Variables – Year and Month



Here, 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) was dropped.

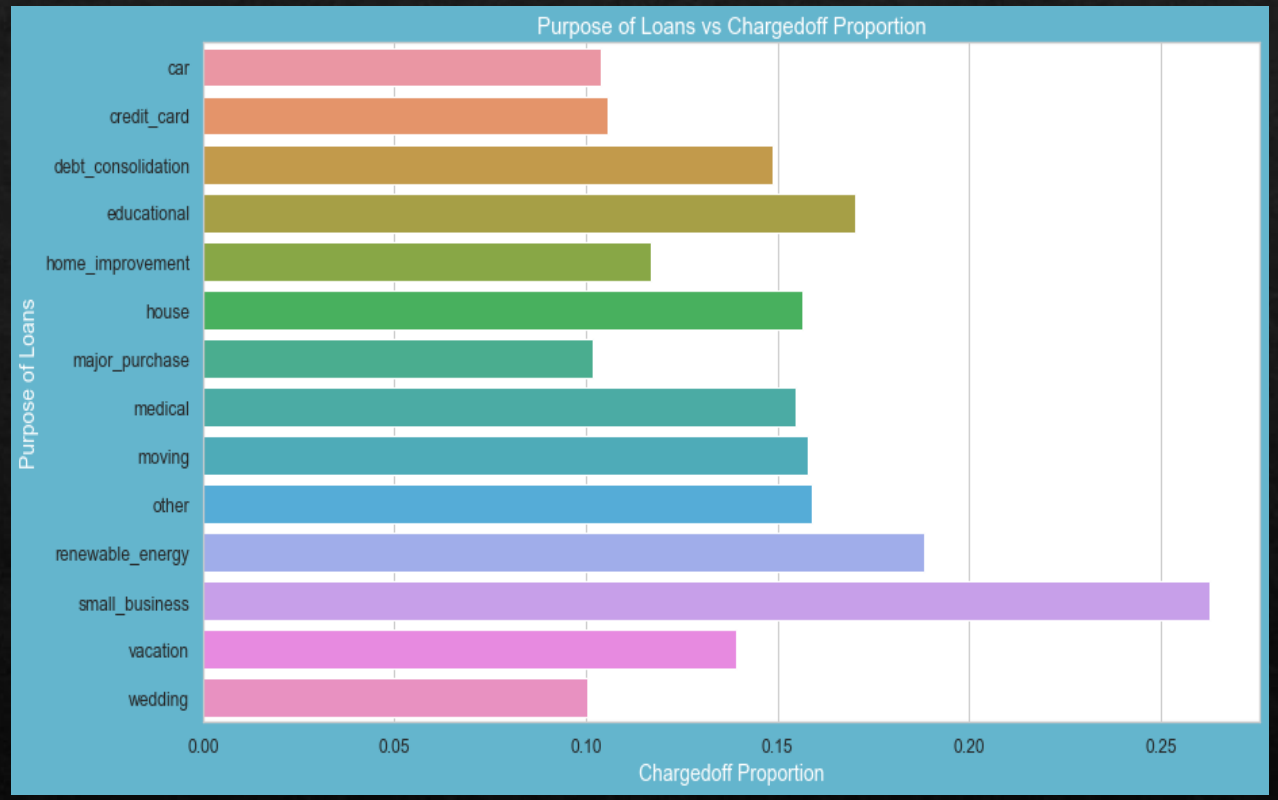
Bivariate Analysis

Annual income v/s Chargedoff_Proportion



- Income range of 80000 above has less chances of loans going charged off.
- Income range of 0-20000 has high chances of loans going charged off.
- With increase in annual income, charged off proportion get decreased.

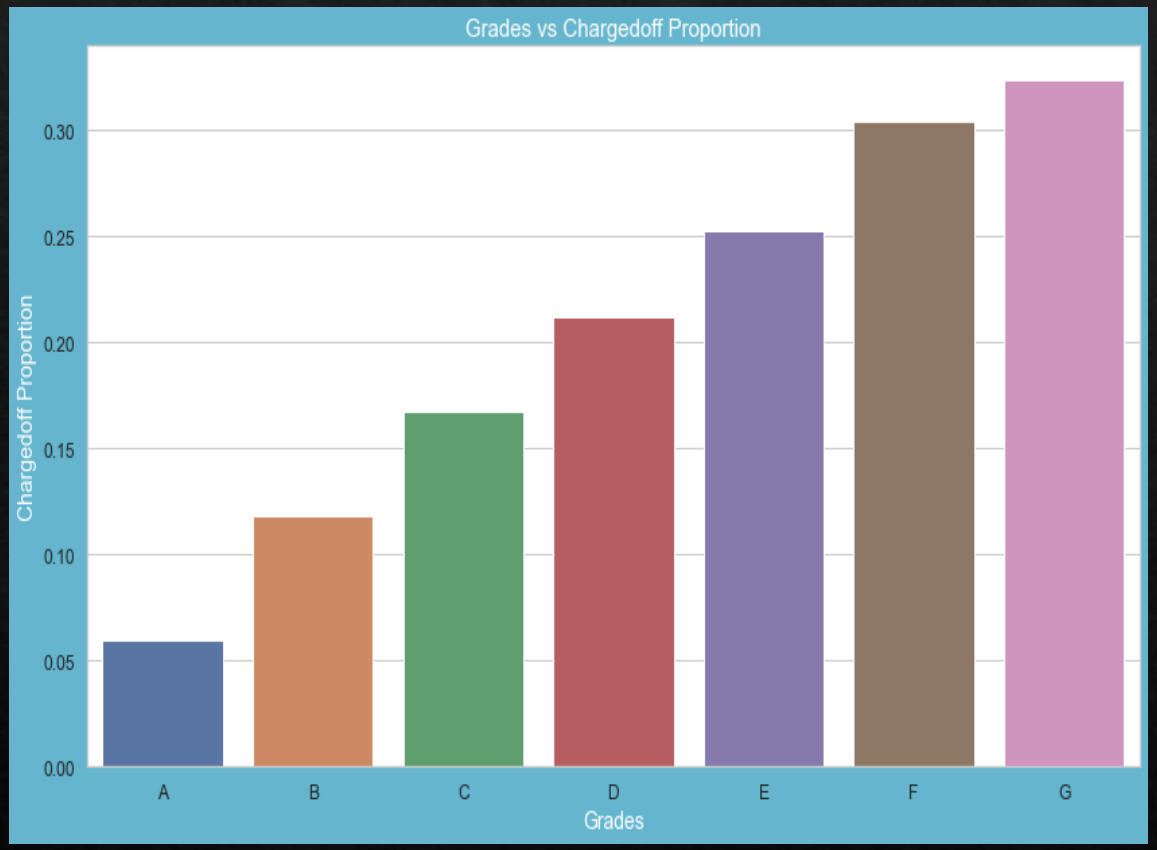
Purpose of Loan v/s Chargedoff_Proportion:



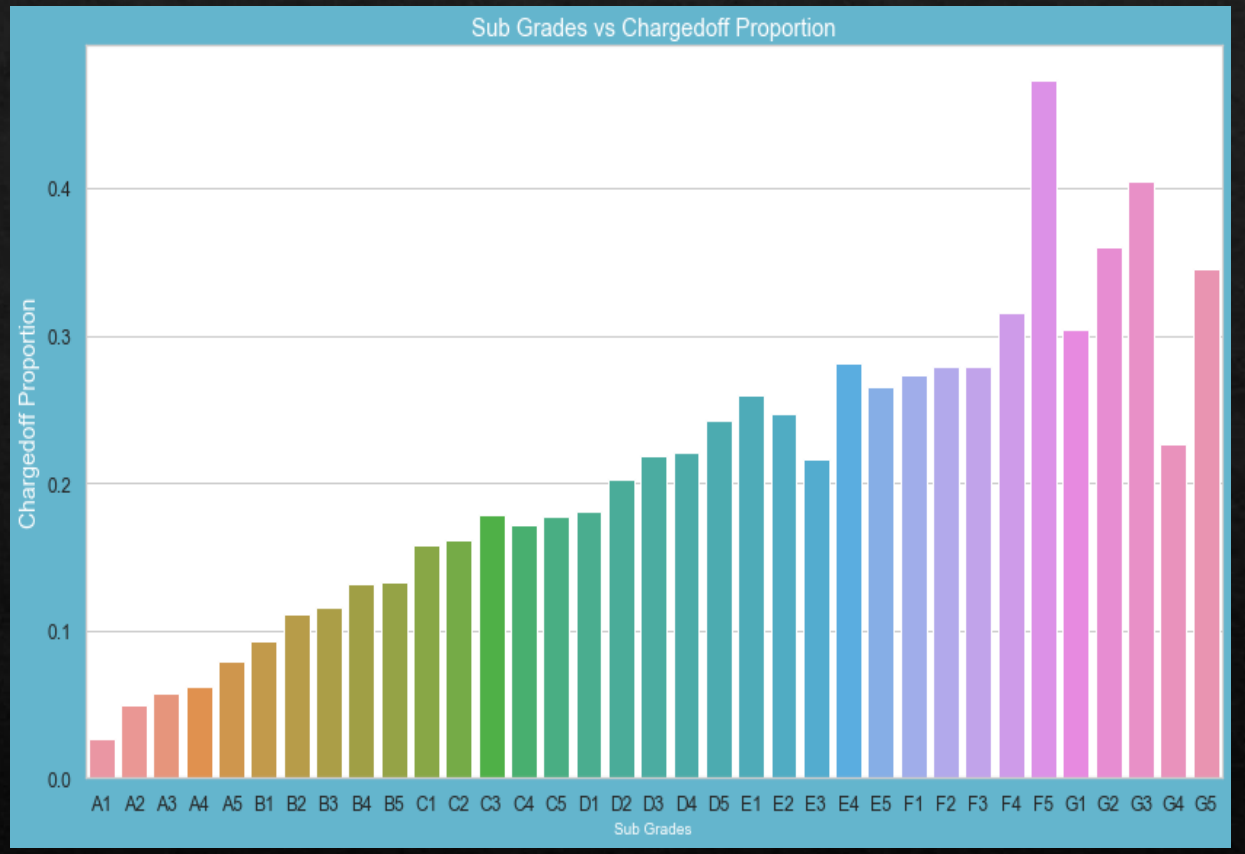
- Small Business applicants have high chances of getting charged off.
- renewable_energy where charged off proportion is better as compare to other categories

Bivariate Analysis

Grade v/s Chargedoff_Proportion



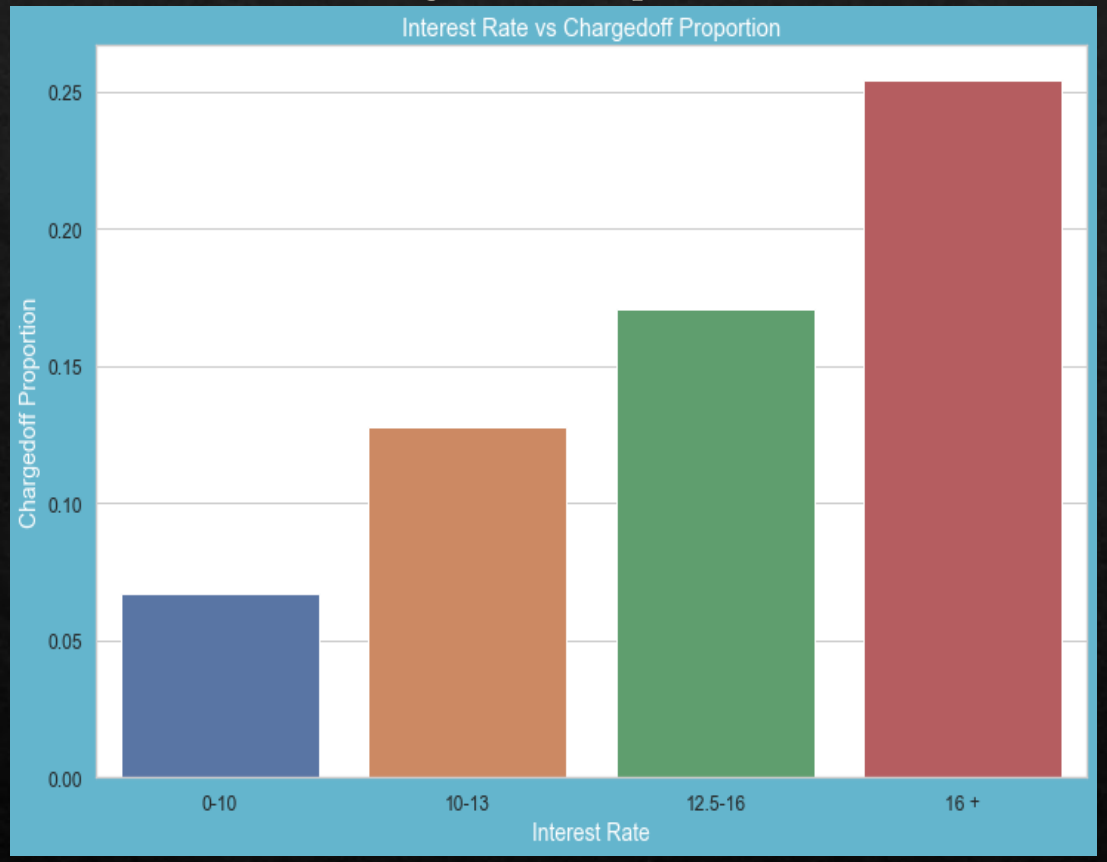
sub_grade v/s chargedoff_proportion



- Grade "A" has very less chances of charged off.
- Grade "F" and "G" have very high chances of charged off.
- Chances of charged off is increasing with grade moving from "A" towards "G"

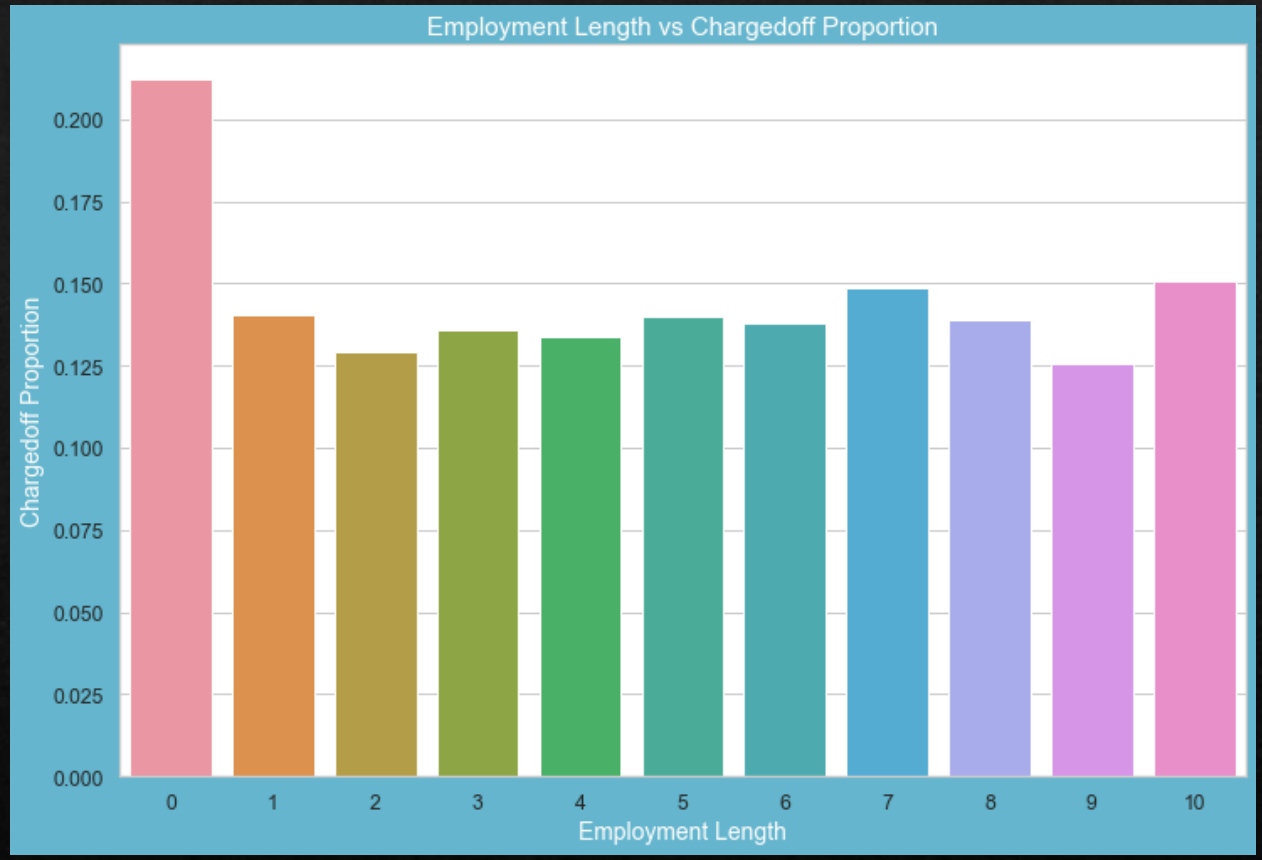
Bivariate Analysis

Interest rate v/s Chargedoff_Proportion



- Interest rate less than 10% has very less chances of charged off and are starting from min 5 %.
- 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.

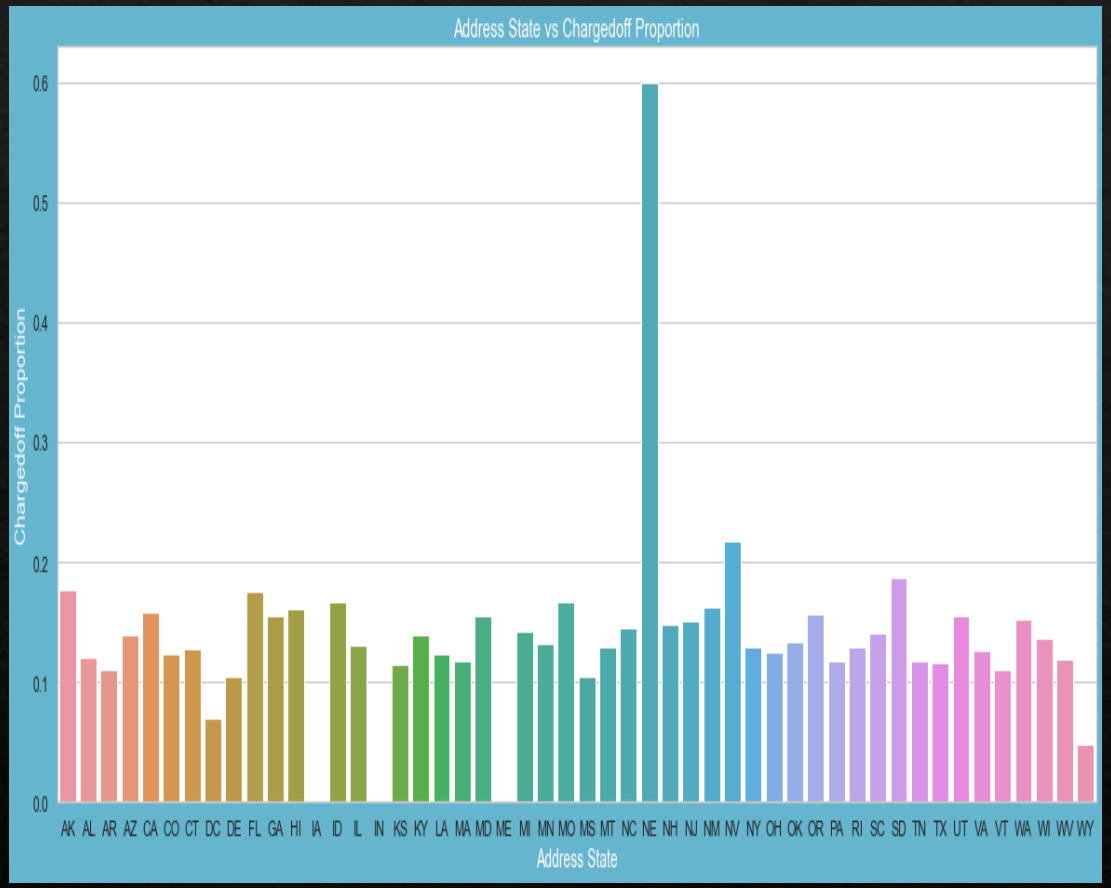
employment length v/s Chargedoff_Proportion



- Those who are not working or have less than 1 year of work experience have high chances of getting charged off as no source of income properly.
- Rest of the applicants have more or less same chances of getting charged off.

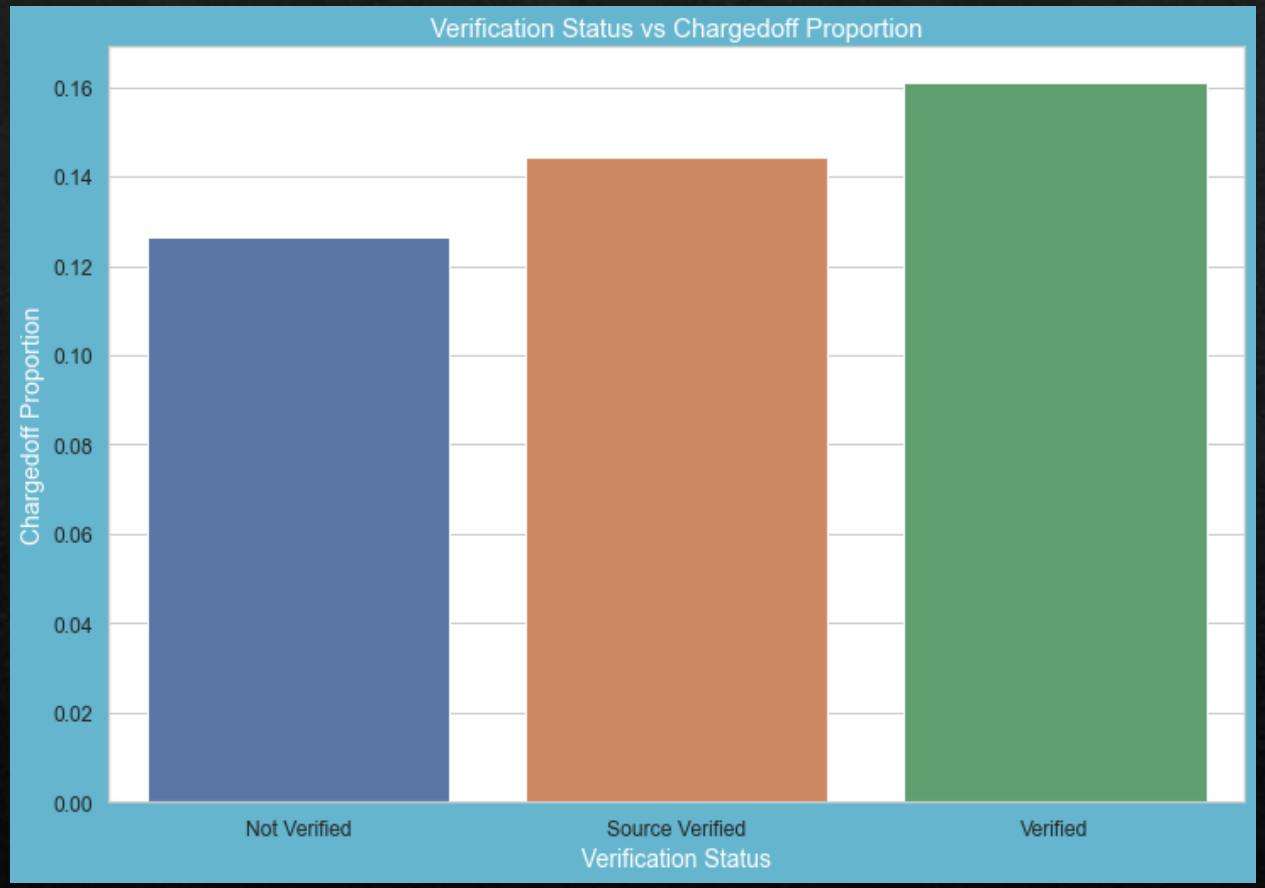
Bivariate Analysis

address state v/s Chargedoff_Proportion



- State 'NE' has very high chances of charged off but number of applications are too low to make any decisions.
- NV,AK and FL states shows good number of charged offs in good number of applications

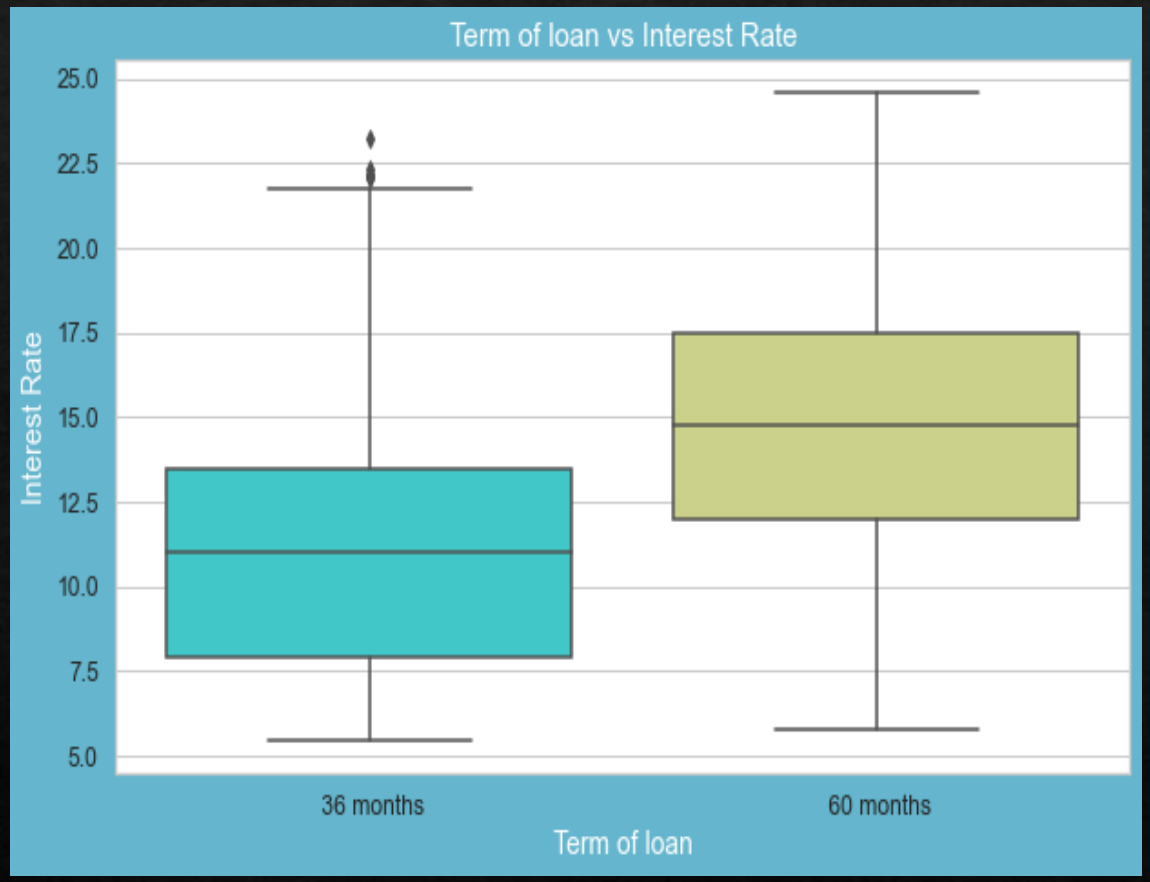
Verification Status v/s Chargedoff_Proportion



- There is not much difference in charged off proportion.
- This variable doesn't provide any insights for charged off.

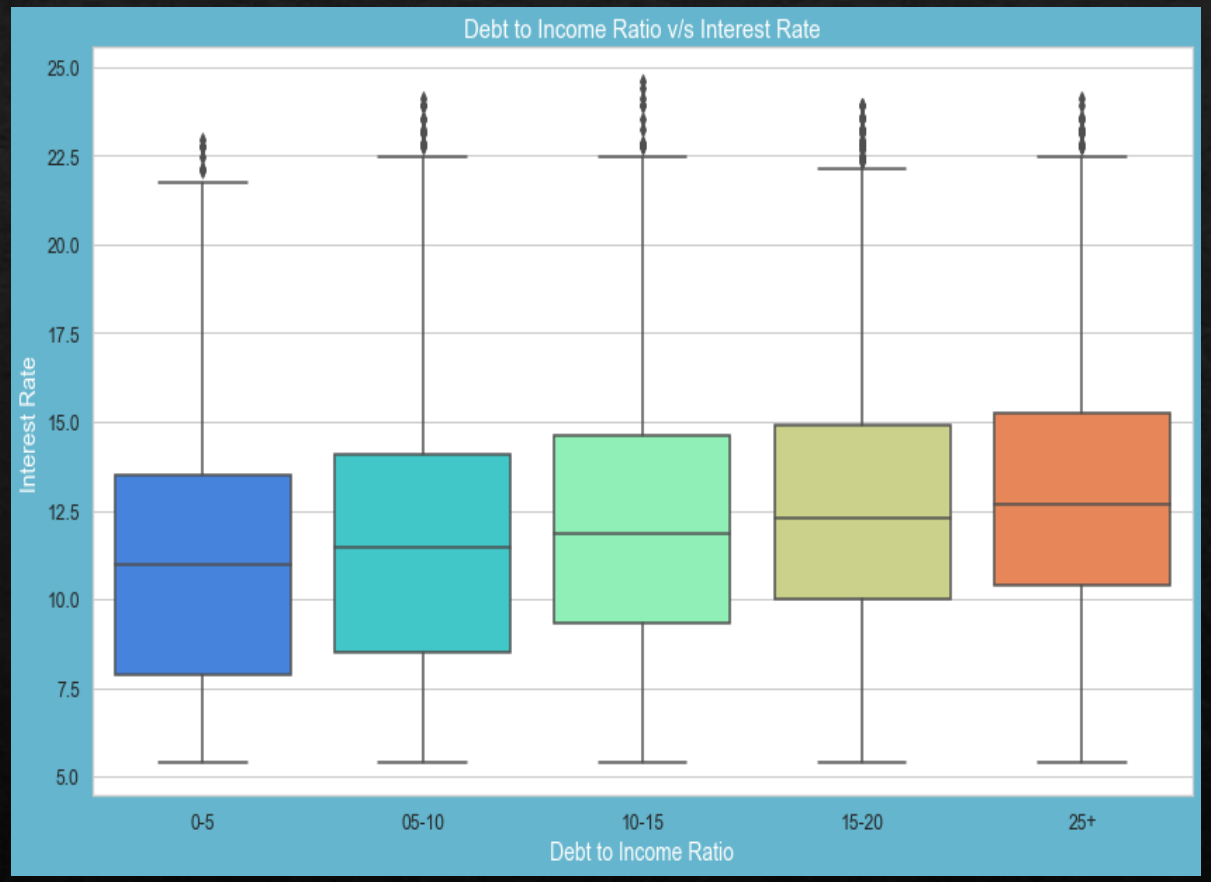
Bivariate Analysis

Term of loan vs Interest Rate - Box Plot



- It is clear that the average interest rate is higher for the 60 months term of loan.
- Most loans issued for longer term had higher interest rates during repayment.

DTI vs Interest Rate - Box Plot



- 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.

Observations and Conclusions

- After conducting a thorough Analysis with the Customer Demographic and Loan Data, the Observations are as below :

- From above Analysis we can see that average default rate across all categories is 15%
- The number of loan applicants is increasing every year
- People with less experience have high chance of default.
- People lying in medium DTI range have high chances of default Higher the DTI ratio, lessen the chances of loan getting accepted
- People who have high loan to annual income ratio are at high risk of defaulting.
- Defaulting changes the Grades from high grade to low grade
- The following are the top 4 categories where maximum loan applications have been received and hence high is the defaulting probability in these categories Debt Consolidation, Credit Card, Other and Home Improvement