Lending Club Case Study

## Exploratory Data Analysis

**Abhishek Dixit**

**Adesh Pathak**

* Problem Statement
* Data Summary
* Data Cleaning
* Data conversions And Derived Columns
* Dropping/Imputing the Rows
* Univariate Analysis
* Bivariate Analysis
* Conclusions

**Problem Statement**

**Problem:**

* You work for a consumer finance company which specialises 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:**

* + Use EDA to understand how consumer attributes and loan attributes influence the tendency of default

**Constraints:**

* 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:
    - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
    - **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’.
    - **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 notavailable with the company (and thus in this dataset)

**Data Summary**

* Loan.csv file contains 39717 rows and 111 columns.
* There is no sub-headers or sub-Footer in the given data set.

# Data Cleaning

* + There were no duplicates rows found.
  + There were 1140 rows present of loan\_status, ‘current’ which has been deleted as loan\_status ‘current’ does n’t participate in analysis.
  + There were 55 columns which is having all the rows values as null/blank and doesn’t participate in analyse has been removed.
  + ‘mths\_since\_last\_record’ and ‘mths\_since\_last\_delinq’ has more than 60% has null values so these columns has been deleted.
  + ‘desc’ and ‘title’ text/description values and doesn’t participate has been dropped from analysis.
  + 11 columns whose values were 1, and is uniqueness in nature has been dropped from analysis.
  + ‘id’, ‘member\_id’, ‘url’, ‘zip\_code’, are unique in nature and ‘funded\_amnt\_inv’, ‘total\_pymnt\_inv’ are same as ‘funded\_amnt’, ‘total\_pymnt’ so these columns has been deleted.
  + After all the Data cleaning process we are left with 35367 rows and 36 columns.

# Data Conversions And Derived Columns

* + Additional string value has been trimmed from ‘term’ column and has been converted to int data types.
  + ‘int\_rate’ and ‘revol\_util’ has been converted from string to int. Additional ‘%’ has been trimmed.
  + Column ‘loan\_funded\_amnt’ and ‘funded\_amnt’ converted to float.
  + ‘emp\_length’ has been converted to number from 0(represent 0-1 year) to 10(represent greater than 9 years).
  + ‘issue\_d’, ‘last\_pymnt\_d’, ‘last\_credit\_pull\_d’ , ‘earliest\_cr\_line’ have been converted to date type.
  + Creating derived columns for ‘issue\_year’ and ‘issue\_month ’ from ‘issue\_d’, ‘last\_pymnt\_d\_month’ and ‘last\_pymnt\_d\_year’ from ‘last\_pymnt\_d’, ‘last\_credit\_pull\_d\_month’ and ‘last\_credit\_pull\_d\_year’ from ‘last\_credit\_pull\_d’, ‘earliest\_cr\_line\_month’ and ‘earliest\_cr\_line\_year’ from ‘earliest\_cr\_line’.
  + ‘profit\_loss’ column is created for each borrower which represent whether the company gain profit (‘+’ sign) or lose (‘-’ sign) their money.
  + ‘ratio’ column is derived which represent the ratio of borrower annual income to the loan amount.

# Dropping/Inputing the rows

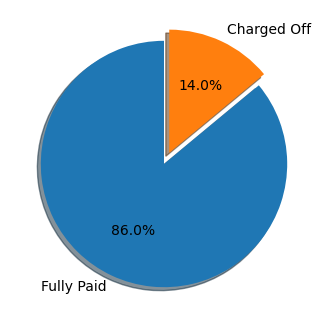
* + ‘emp\_title’,‘emp\_lenght’ and pub\_rec\_bankruptcies contains 6.18%, 2.67% and 1.80% of rows as null, which is very small percetnage of data which we can drop it.
  + Total % of rows deleted: 8.32%,
  + Outliers exits for numeric data 'loan\_amnt', 'funded\_amnt','int\_rate', 'installment', ‘annual\_inc' so these are removed while analysing.
  + Outliers treatment has been done for above fields using Upper Fence and Lower Fence mechanism.

Univariate And Segmented Univariate Analysis

# Loan Status

# 

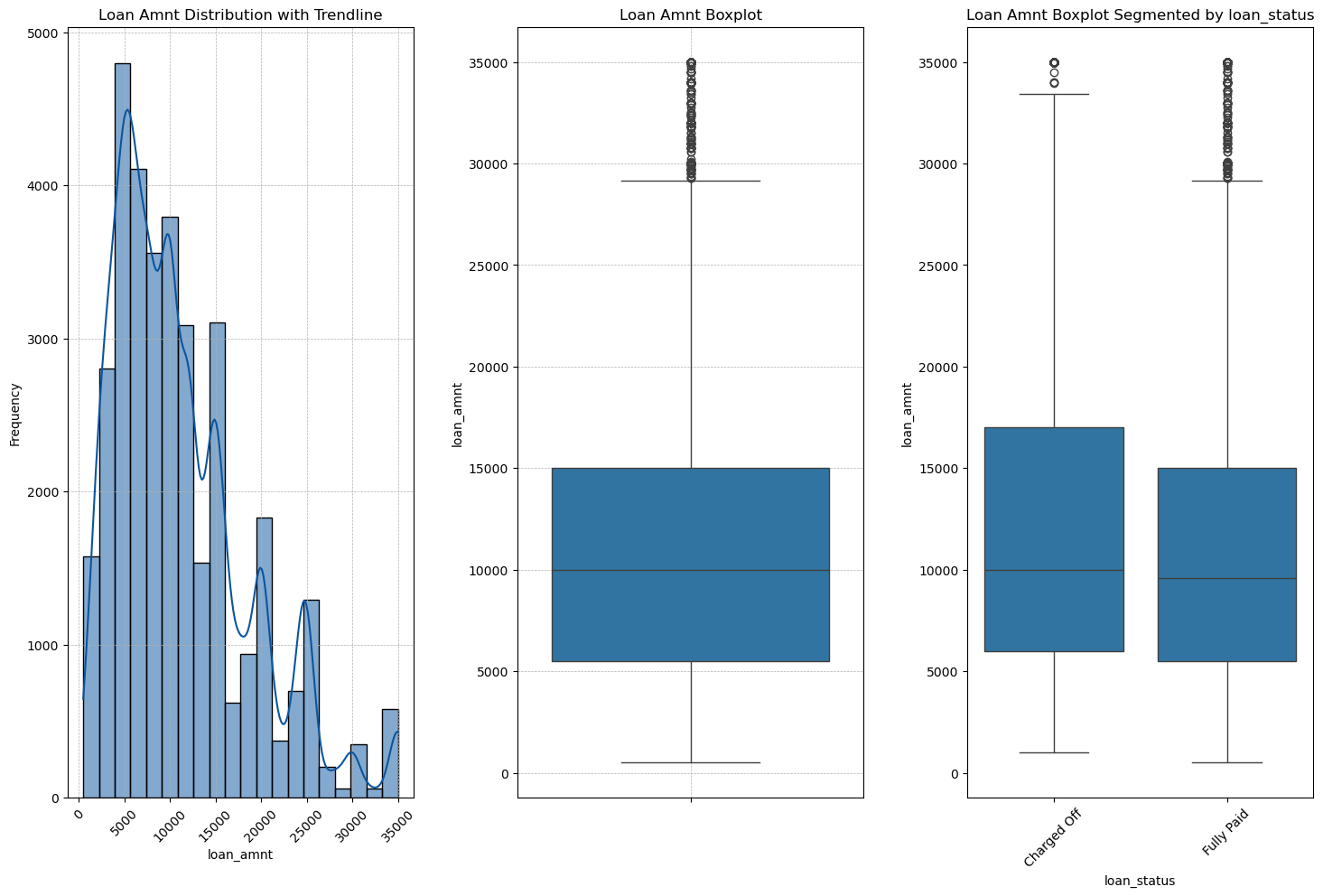
* **Observations:**
  + Approximate 14% borrower are defaulted in the data set.

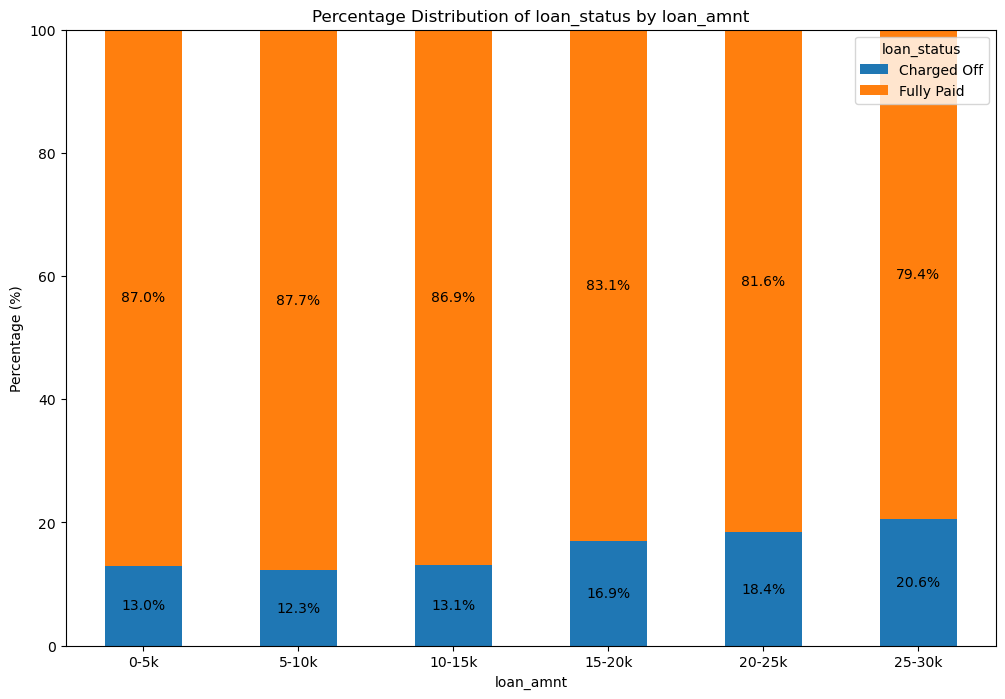


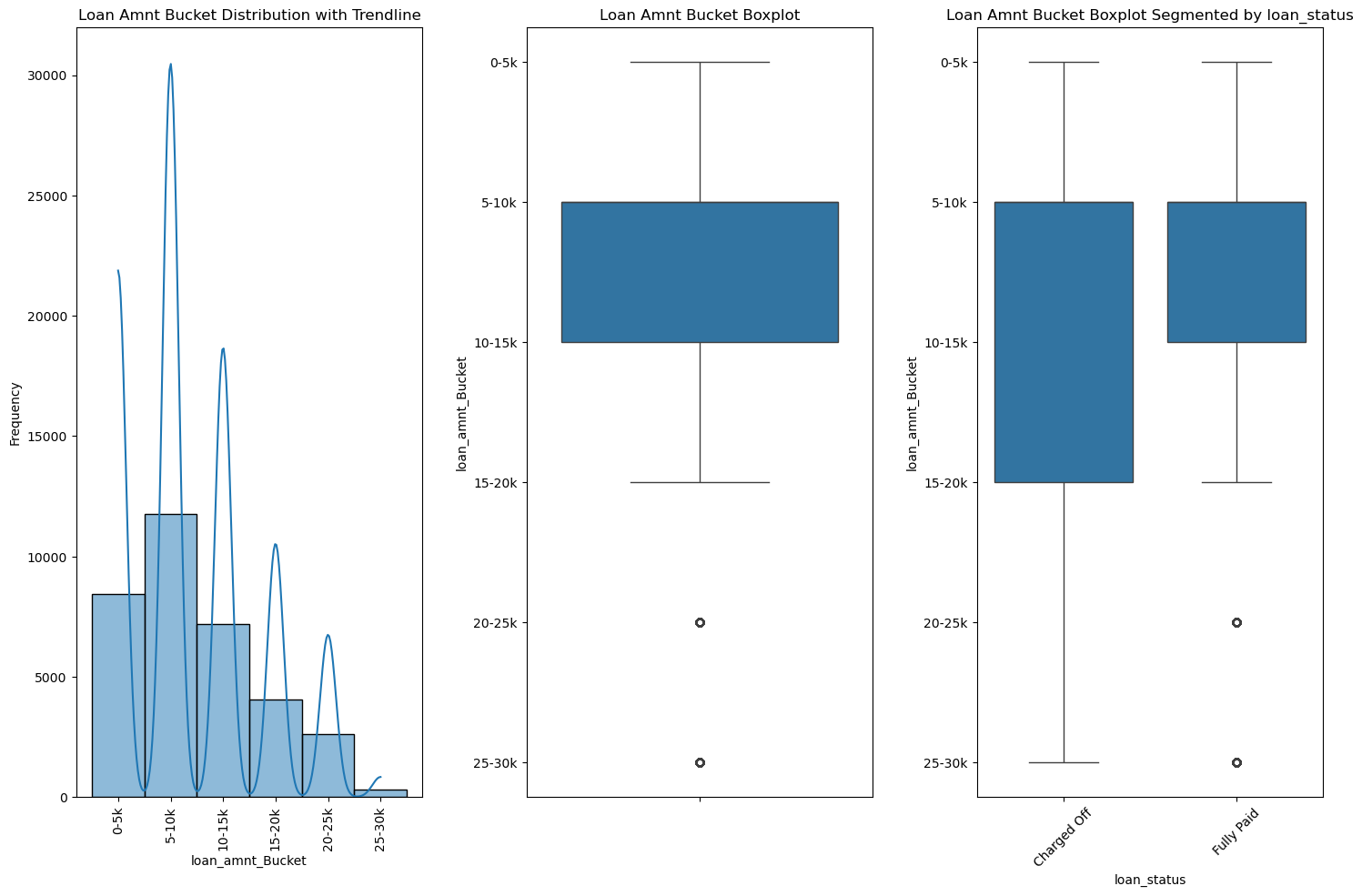
# Loan Amount

### Observations:

* + Overall loan amount varies from 500 to 35000.
  + Charged off loans have high std, mean, median than fully paid.
  + There are Outliers in the Loan Amount So we remove the outliers First
  + After removing outliers, we have bucketed the loan amount in the gap of 5000
  + Most of the loan amount are between 5000 to 10000
  + Fully Paid loans have low IQR in comparison to charged off.
  + Charged Off has high average and std in comparison to Fully Paid
  + % default increasing with increase in loan amount range



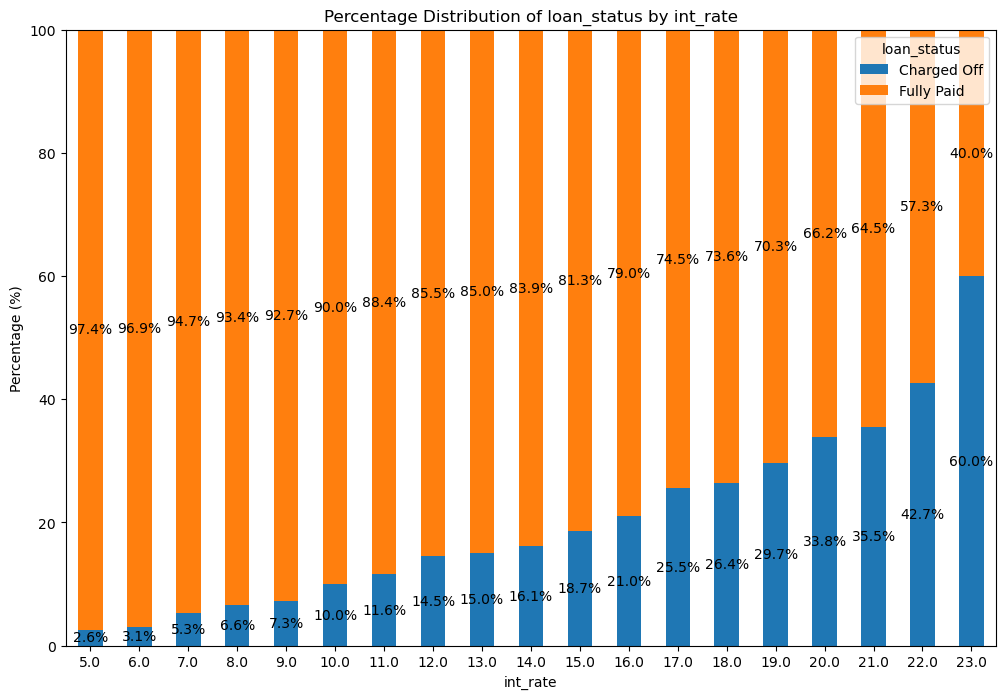


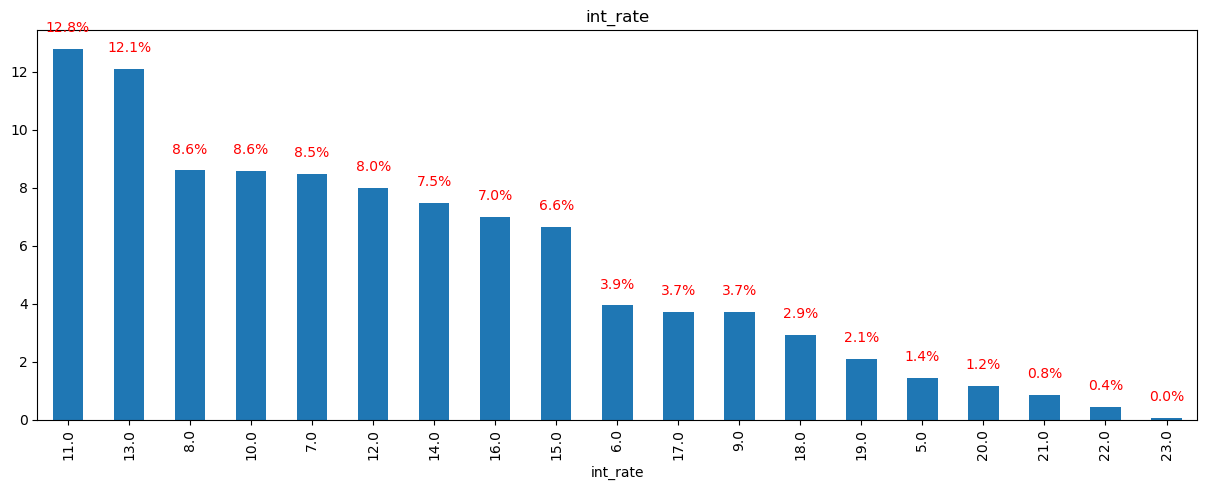


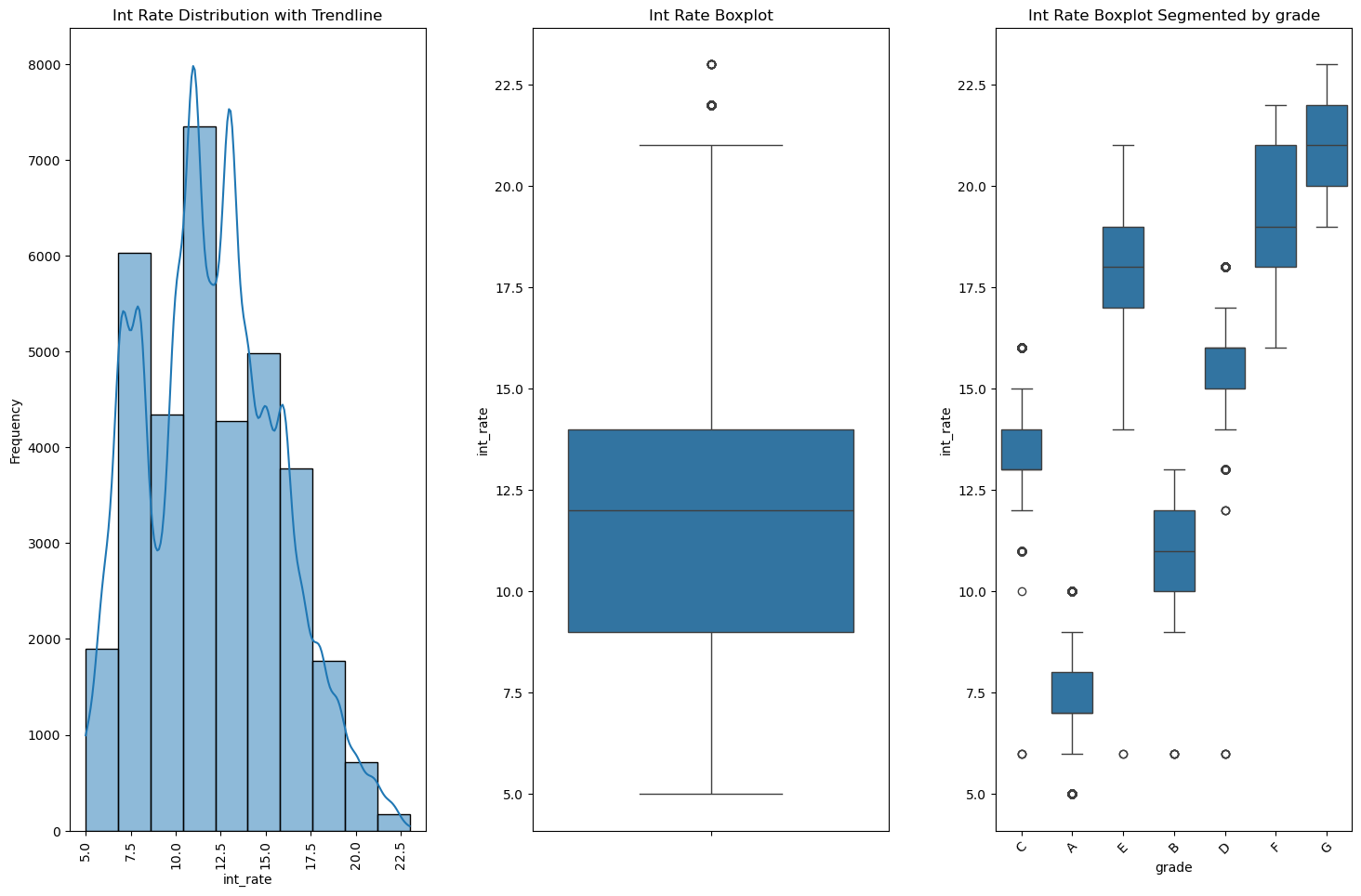
# Interest Rate

* **Observations:**

* First of all, we have removed outliers.
* After removing outliers, we have round off interest rate in difference of 1.
* overall interest rate varies from 5% to 23% after removing outliers
* 11 and 13% interest rate are more frequent in complete data set.
* The interest rate for Charged Off loans appear to be higher than for Fully paid. This is naturally expected. As, the risk increases the rate of interest imposed on the loan also increases
* LC provided grade E,F,G grade has the highest rate of interest.
* on increasing interest rate, the chance of being default also increases.





I

# Installment

* **Observations:**

* First of all, we have removed outliers.
* overall installment varies from 16.08 to 1305.19.
* Most of the Installment fall between 0 to 400.
* loans Charged Off have high installment on average

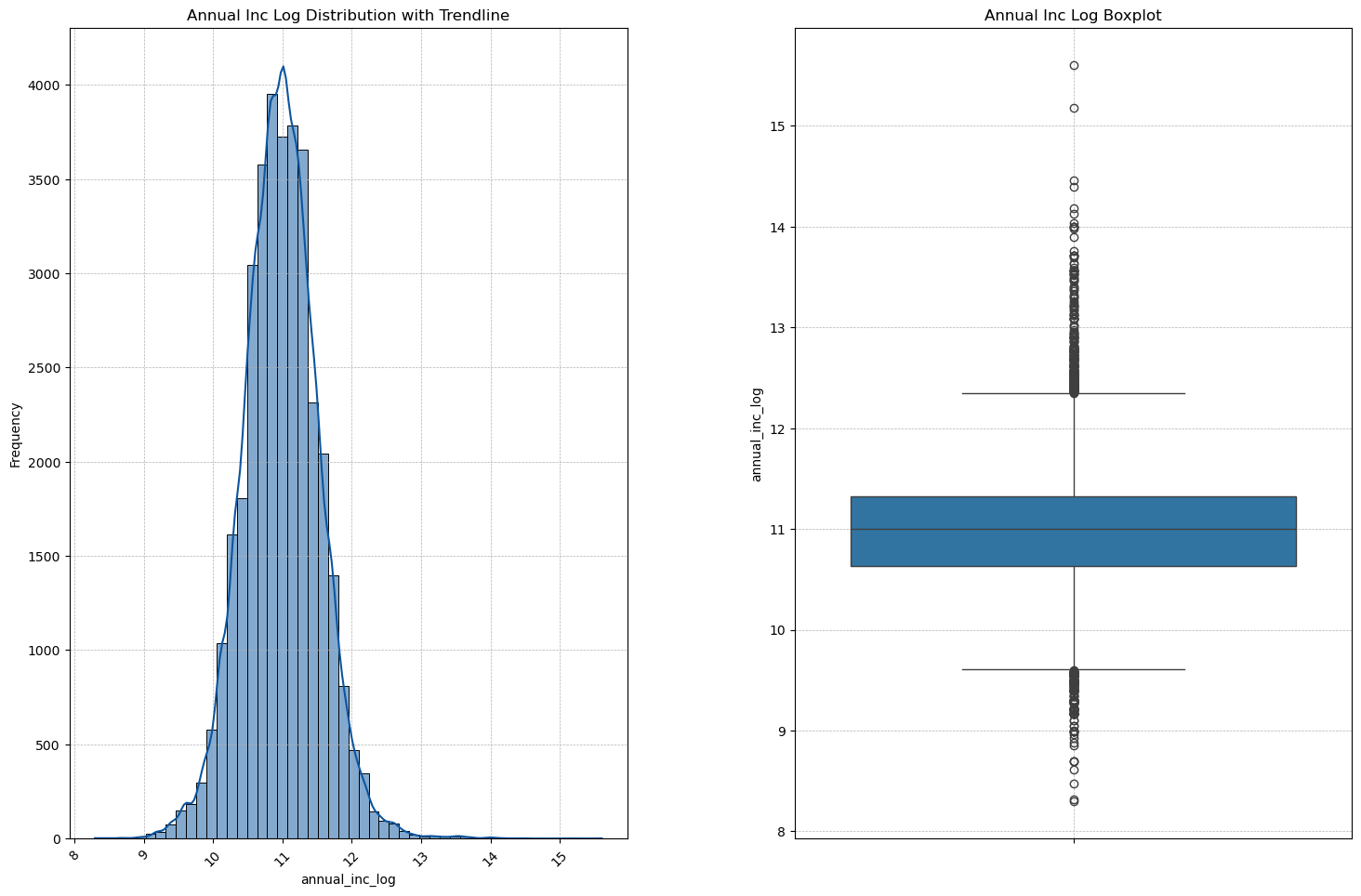
# 

# 

# Annual Income

### Observations:

* The Log plot of annual income show normal distribution means symmetric about it means.
* To further Analysis we have remove outliers.
* After removing outliers, we have created bucket of 20K difference.
* most of the people in the data set fall in income range 30-70k after removing outliers
* Loan defaults are higher for lower income, and progressively reduce as incomes go up rate of interest imposed on the loan also increases
* On lower Salery bucket range higher the chance of default.

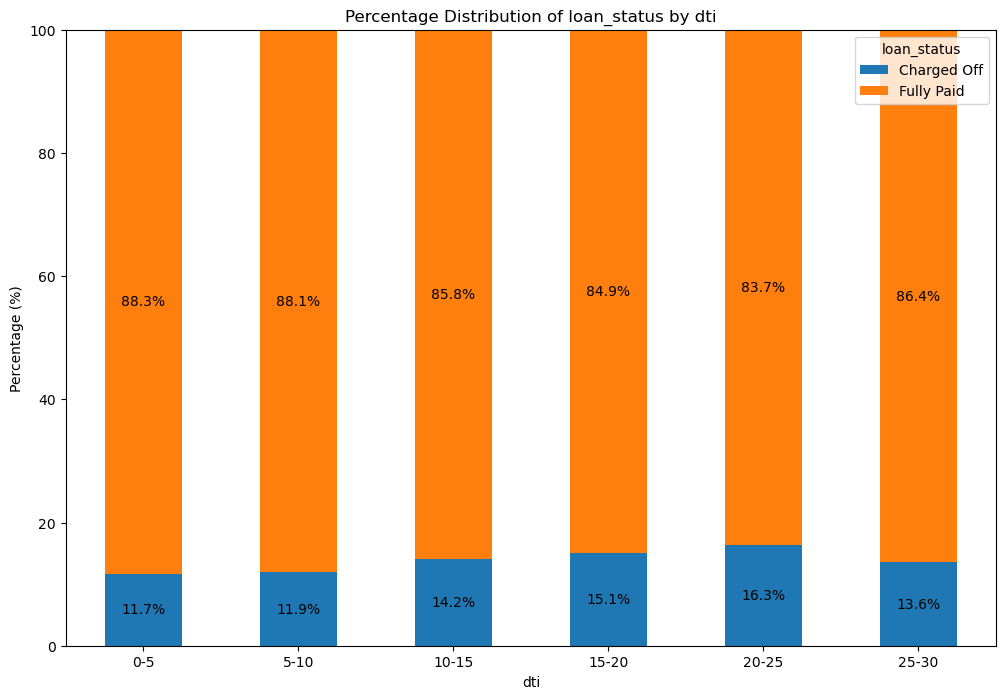


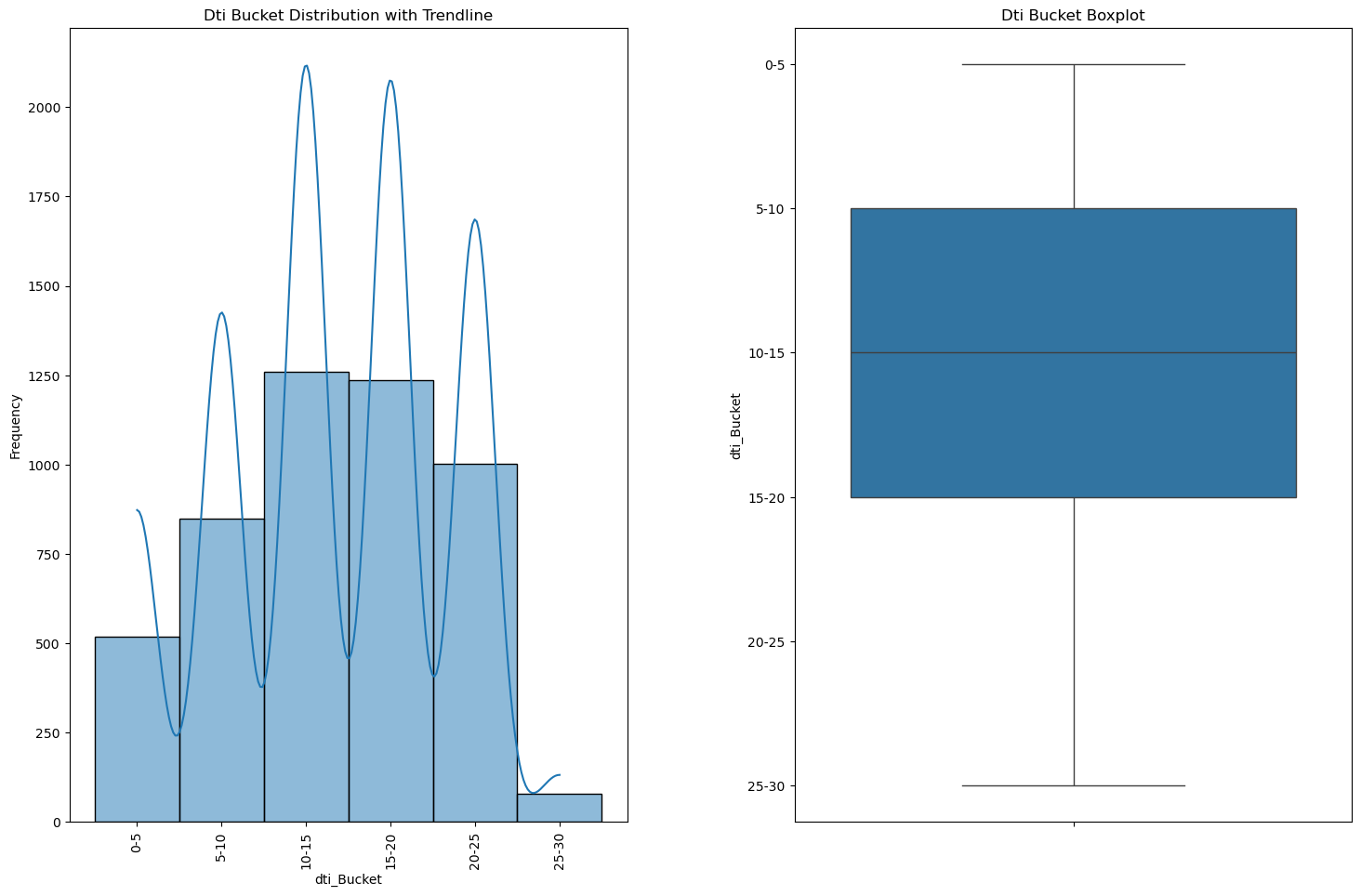
# 

# DTI(Debt to Income Ratio)

* **Observations:**

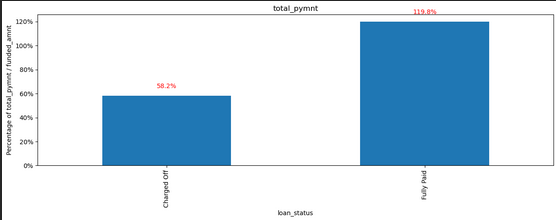
* First of all, we have removed outliers.
* After removing outliers, we have created bucket of 5-5 difference.
* The dti index varies from min 0 to max of 30. The median dti is of 13.5
* Dti value has high frequency between 0-25.
* As the debt-to-income ratio increases the percent of being default of the borrower also increase.





**Total Payment**

* **Observations:**
* First of all, we have removed outliers.
* The average payment received to date for the Charged Off loan is comparatively less than Fully Paid loans
* Without removing outliers in total payment LC get approximate 59% of its payment in charged off and in fully paid it gets 20% profit.



# 

**Term**

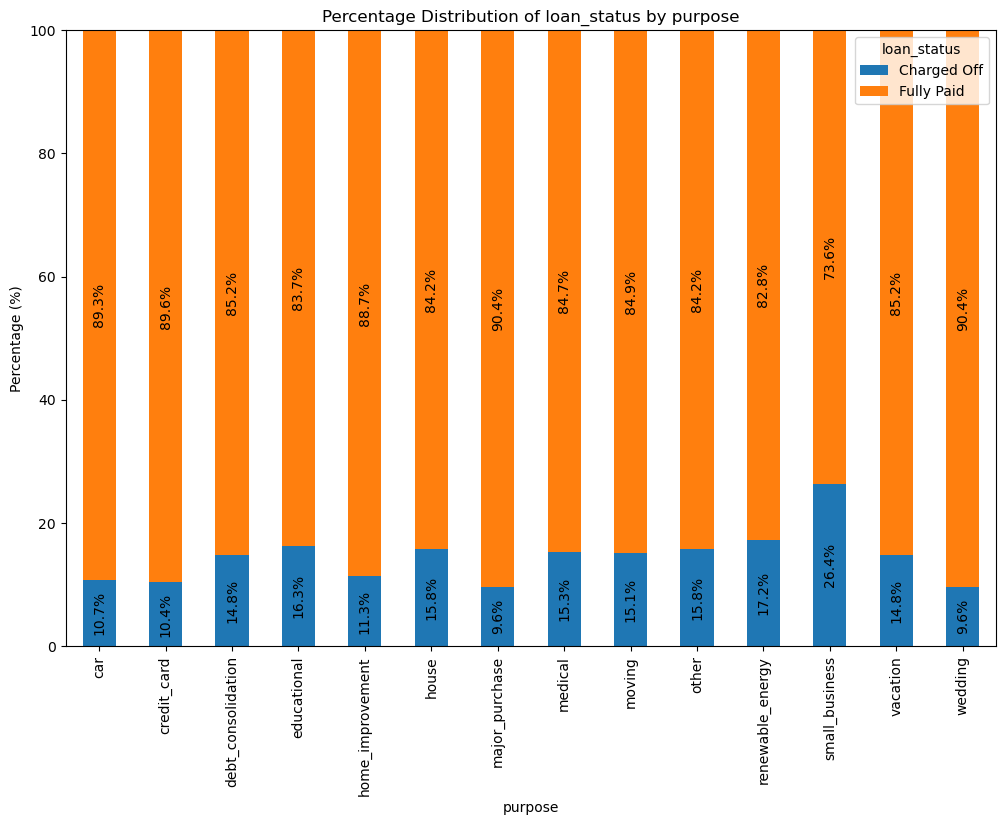
* **Observations:**
* Higher the term higher chance of default.

# 

# Purpose

* **Observations:**

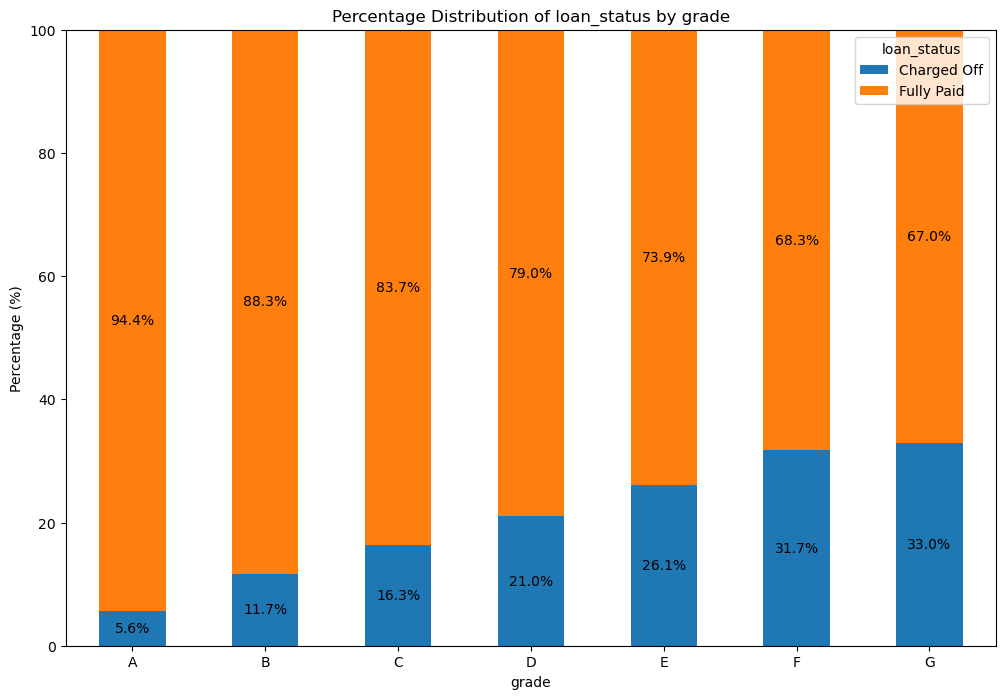
* Small business have higher chance of default.



# Grade

* **Observations:**

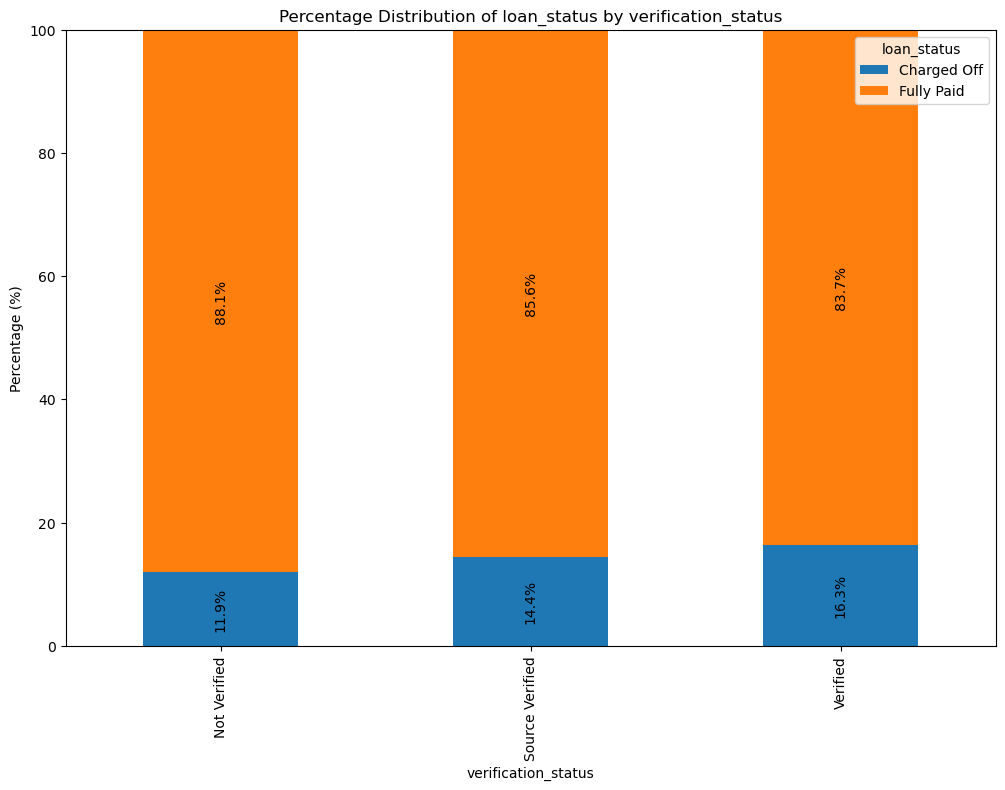
* As the grade provided by lending club increases the chance of being default also increases



# Verification Status

* **Observations:**

* Source verified and verified has higher % of default



# Ratio of annual Income to the Loan Amount Taken

* **Observations:**

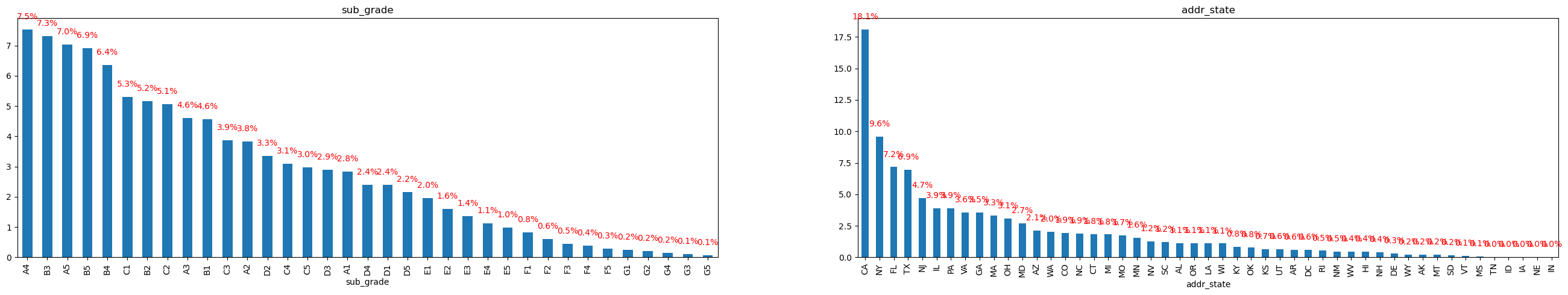
* Lower the ratio higher the chance of being default

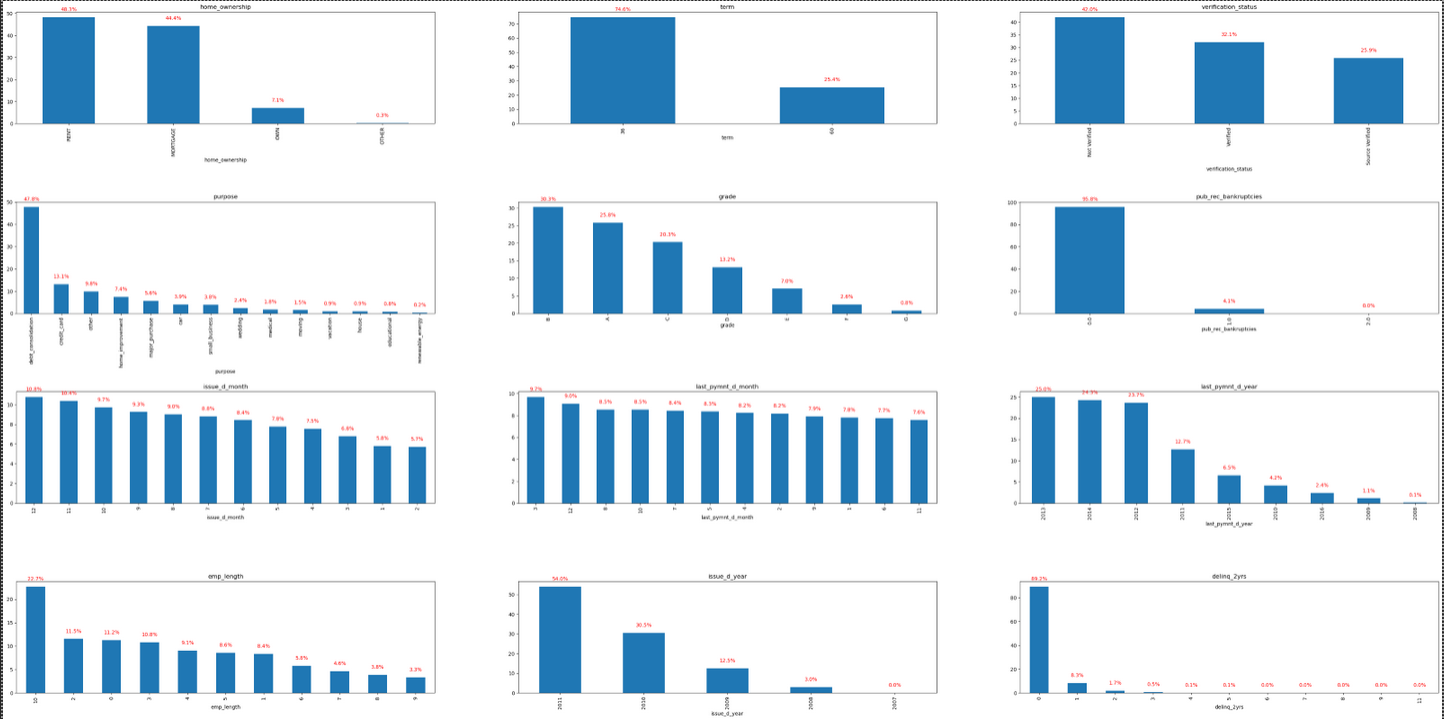
# 

# Categorical Variables

* **Observations:**

* Most of the Borrower has home ownership either rental or mortgage
* Most of the borrower has loan term 3 year
* There is a power law distribution in varification\_status and most of the borrower are not verified
* The loan purpous debt\_consolidation is quite high among the borrowers
* Borrower Grade count trend B>A>C>D>E>F>G
* Most of the borrower has zero public bankruptcies records
* 10+ years of employment is the most common length, followed by 2 years, 3 years and less than 1 year.
* The distribution shows a finer breakdown of the grades, with sub-grades like B4, B5, A5, and B3 being the most common. The frequency decreases as you move to lower sub-grades, particularly in grades like F and G.
* Over the years organization has been giving more loans. From Aug,2007 to 2011 the loans issued have risen significantly.
* Within a year, the number of loans issued rises over the month from jan to Dec. Dcember is the month of the year where the maximum number of loans are being issued.
* Overall, around 75% of the last payment dates are during 2012-2014 period.
* Maximum loans ~18% are from California following that is 9.5% from New york state and Florida with 7%. This is to be expected as these are also the three most populous US states
* The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years values range between 0 and 11.





Bivariate Analysis

# Conclusions

* + - Income range between 0-20000 has high chances of charged off.
    - Interest rate more than 16% has good chances of charged off as compared to other category interest rates.
* Those who are not owning the home is having high chances of loan defaulter.
* Those applicants having loan for small business is having high chances for loan defaults.
* High DTI value having high risk of defaults.
* Higher the Bankruptcies record higher the chance of loan defaults.
* DE States is holding highest number of loan defaults.
* The Loan applicants with loan Grade G is having highest Loan Defaults.