LENDING CLUB CASE STUDY

EXPLORATORY DATA ANALYSIS

I N T R O D U C T I O N

- We have a company which 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.
- Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (Credit Loss) i.e borrowers who default cause the largest amount of loss to the lenders.
- The company wants to understand the driving factors behind loan default so that they can evaluate these parameters in the loan application and effectively decide if they can Approve or Reject loan

PROBLEM STATEMENT

- Whenever 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:
    1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
    2. 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
- We need to analyze the dataset provided by the company and perform an Exploratory Data Analysis to highlight the list of driving factors which influence the possibility of an applicant repay the loan or defaulting on it.

ABOUT THE DATASET

- The data is provided in the form of a CSV file (**loan.csv**).
- The data has list of all the loans issued between 2007 and 2011. Some of the specifications are as follows:
    1. There are **111 columns** in the dataset.
    2. There are **39717 records** in the dataset.
    3. The target variable here is the **loan_status** column.
    4. Loan_Status column is an unordered categorical variable with 3 different categories:
        a) Fully Paid
        b) Charged Off
        c) Current
    5. Of the 3 categories, we are interested in **Fully Paid** and **Charged Off**

DATA CLEANING

- Data Preparatory or the data cleaning activities are one of the first and the key set of activities that we need to perform before start analyzing the data.
- In the **Loan.csv** dataset, we had to perform quite a few data wrangling activities. Below are the list of activities:

    1. Remove Columns which had completely *NULL* values.

       There were around 55 columns which had completely NULL values. Since they wont yield any useful insights, had to delete these columns.

    2. Remove records with loan status as *Current.*

       Our target column *loan_status* has 3 categories of which we need only *Fully paid* and *Charged Off*. Hence we had to remove all the records which had loan_status as *Current*

DATA CLEANING

3. Handling the Null values for the rest of the columns

   There were around 3 columns which had 93%, 65% and 35% missing data. Had to delete those columns.

   Had 3 columns which had 6%, 3% and 2% missing data. Had to impute these values with relevant values based on business logic

   Had around 5 columns which had 0.5% to 0.2% missing data. Have deleted those records.

4. The issue with duplicate values

   Around 9 columns had same values for all columns. Deleted those columns.

   Around 7 columns had 90% of its data with same value. Deleted those columns as they are highly skewed

DATA CLEANING

5. Handling Outlier records

   There are around 7 columns which had outlier data. Deleted around 0.5% of the outlier data.

6. Transforming the date fields into DateTime format
7. Transforming the potential numeric columns from their current object type format
8. Transforming the object column into a ordered categorical numeric column
9. Handling the inconsistent data

   There was a column where few records had unrealistic values. Since it constituted 0.25% of the data, removed all those records.

10. Derived columns: Extracting the Month and the Year data from the DateTime column

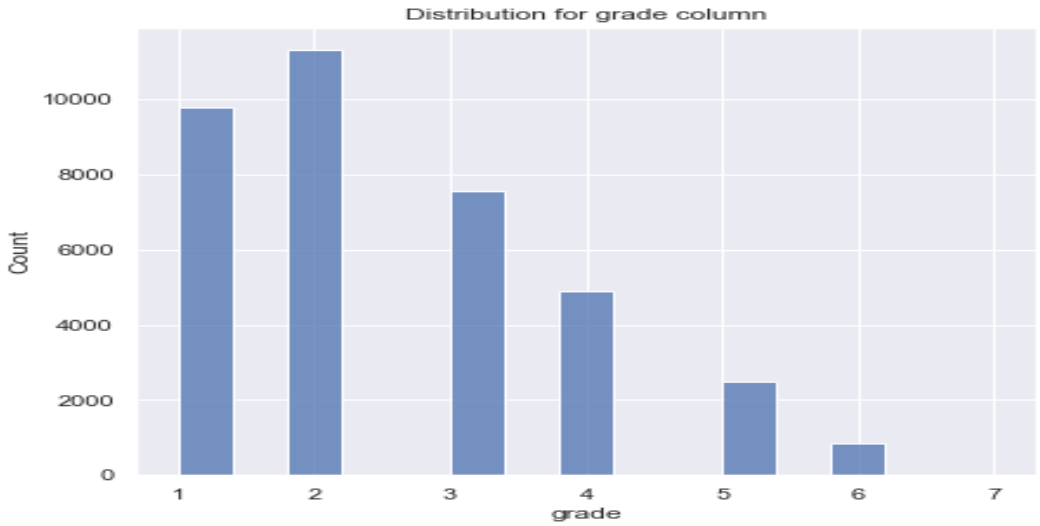
After performing the above activities, we have arrived at a dataset which has 38 columns and 37138 records.

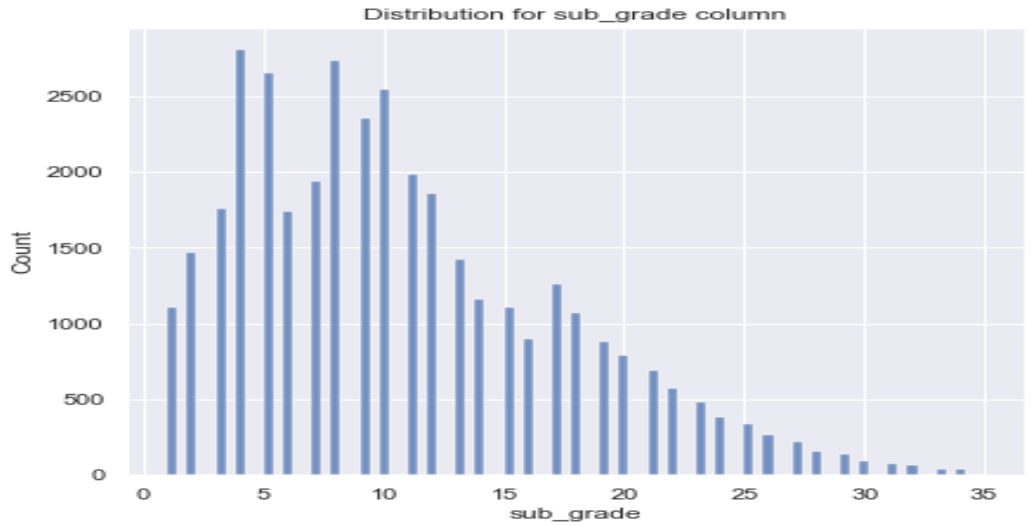
DATA ANALYSIS

- We will divide the data analysis into 2 phases
    1. **Univariate Analysis**
    2. **Bivariate Analysis**
- *Univariate Analysis* is one wherein we will analyze columns individually, and understand the underlying pattern. We will further divide this univariate analysis into below 2 sections:
    1. **Categorical Univariate Analysis**
    2. **Continuous Univariate Analysis**
- *Bivariate Analysis* is one wherein we will analyze the interaction between 2 variables and understand if there exists any relation between the 2 variables. We will further divide this univariate analysis into below 2 sections:
    1. **Categorical Bivariate Analysis**
    2. **Continuous Bivariate Analysis**

CATEGORICAL UNIVARIATE ANALYSIS

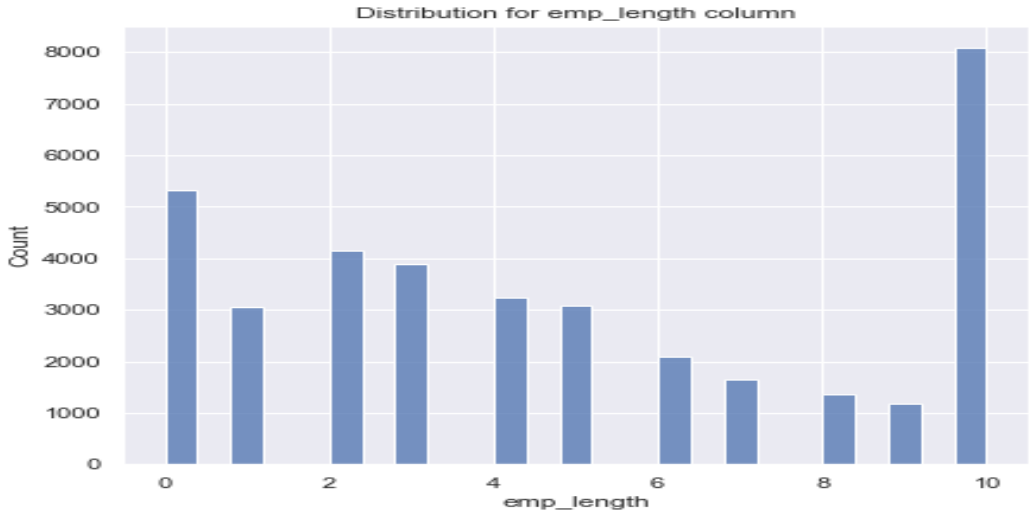
- **Grade and Sub-Grade column:**
  - Lower the grade lesser the borrowers. For sub-grade, the number of borrowers increase with decrease in sub-grade i.e from A1 to A5. This might be because the borrowers tend to maintain a higher grade (even though lower sub-grade) than a lower grade.

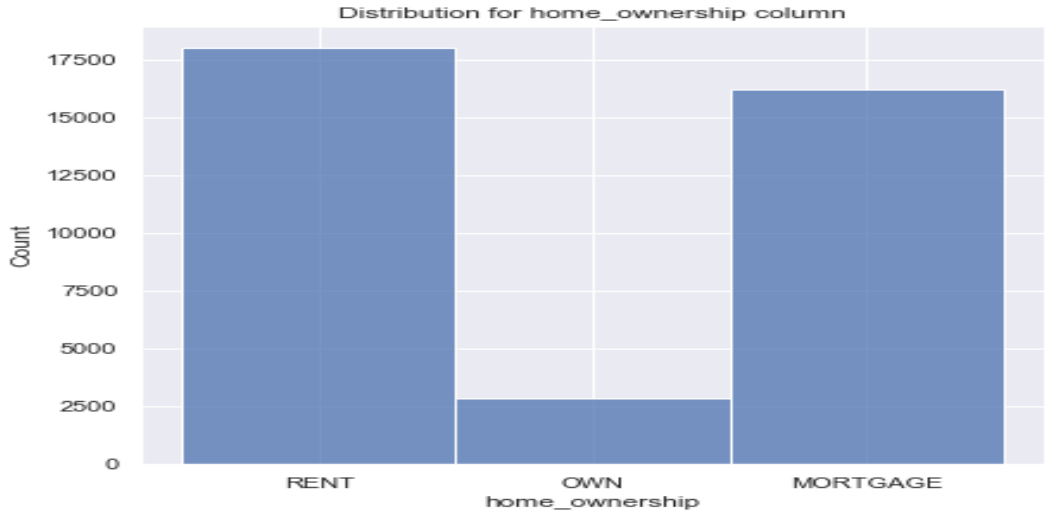




CATEGORICAL UNIVARIATE ANALYSIS

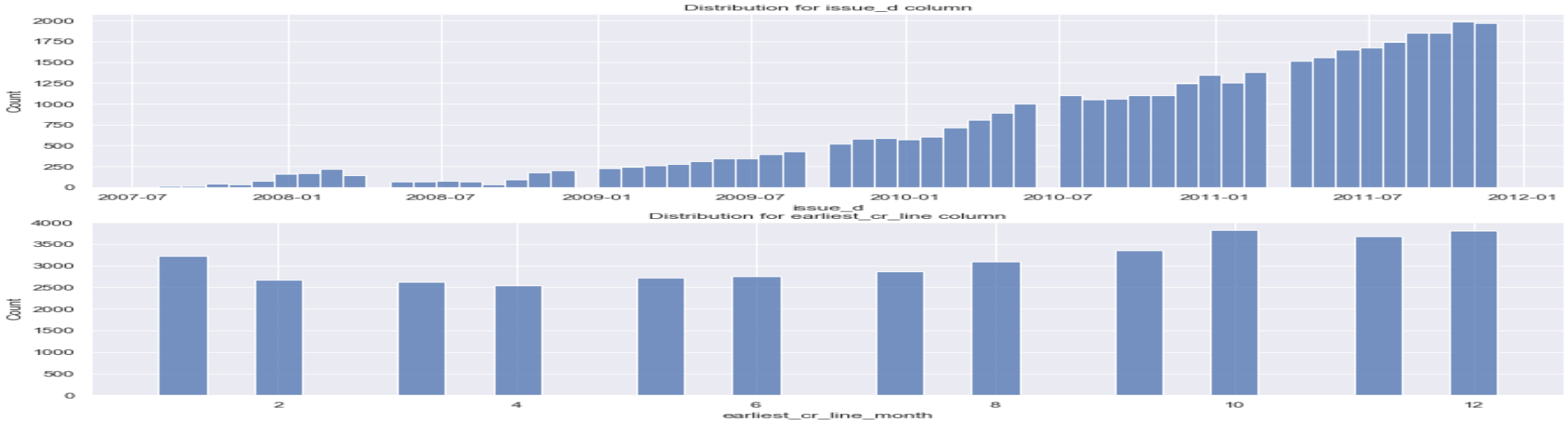
- **Employee Length and Home Ownership**
  - **Employee Length**: More the experience, lesser the number of users who tend to borrow. But for people who are less than 1 year experience tend to borrow high compared to 1-2 years. This is because the sense of getting a job, compels many users to apply for loan, and since many would have applied for load in the initial year of getting a job, they tend to refrain for the next year.
  - **Home Ownership**: Most of the borrowers are either on rent or already have a mortgage. This shows that people who tend to borrow have some or the other prior obligations





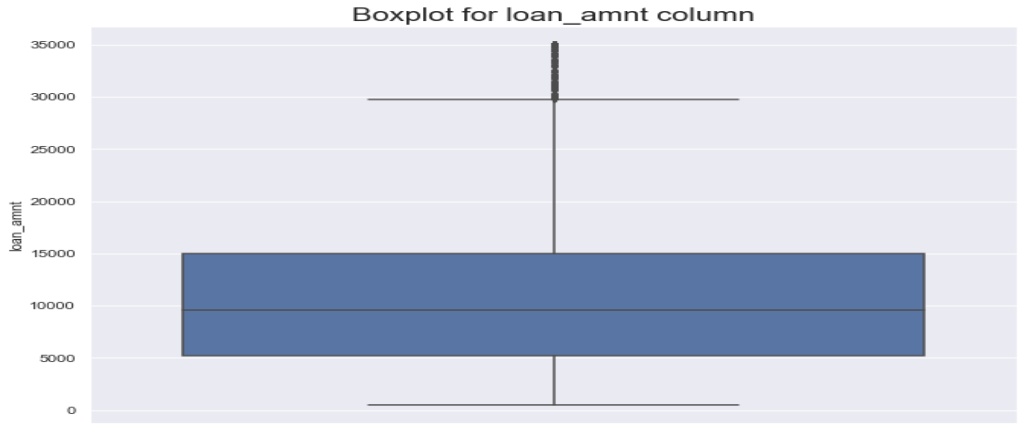
CATEGORICAL UNIVARIATE ANALYSIS

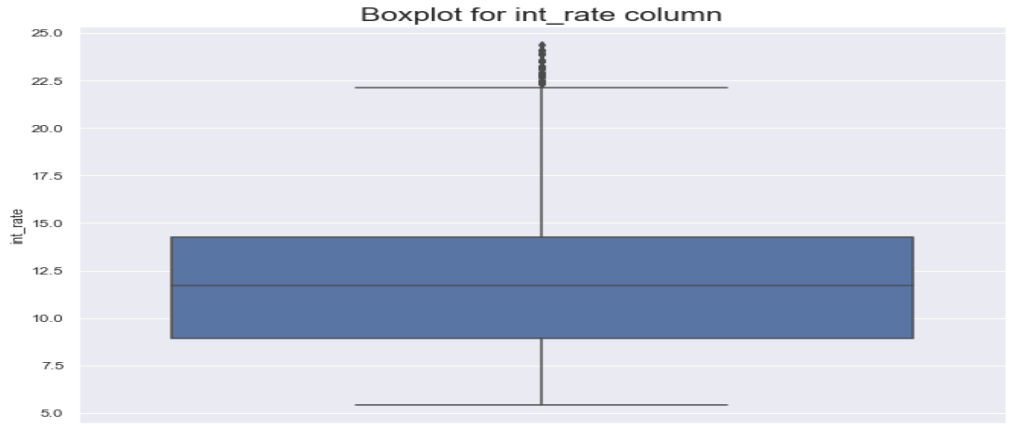
- **Issue Date and Earliest Credit Line**
  - I**ssue_d**: Between March 2008 till Oct 2008, the issuance of loans decreased. This is due to the 2008 financial crisis. But start of 2009, there was an increase in disbursement of loan and it gradually increased all the way till 2012.
  - **earliest_cr_line_month**: Its interesting to note that more numbers of borrowers have their credit line opened during the month of Oct, Nov, Dec and Jan. This is specially the time of Thanks Giving, Black Friday and Christmas (Holiday time in the US).



CONTINUOUS UNIVARIATE ANALYSIS

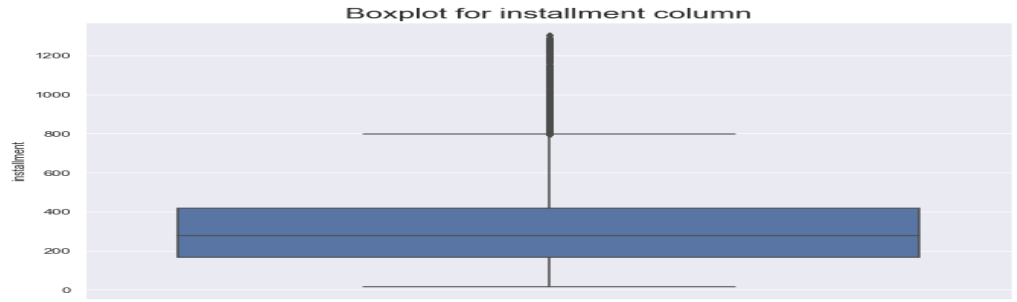
- Here we will try to explore the Continuous columns in depth and highlight any interesting pattern that we see in them with relevant graphs
- **Loan_amnt and Int_rate**
  - **loan_amnt**: Its interesting to find that the values are more concentrated around the multiples of 5000. Also, we can see prominent spikes around multiple of 1000. One of the reason for this might be the fact that people take loan amount in multiple of 1k or 5k
  - **int_rate**: 50% of the values are between 9% and 14%. A good number of values are centered around 7.5% (Close to 12% of values)

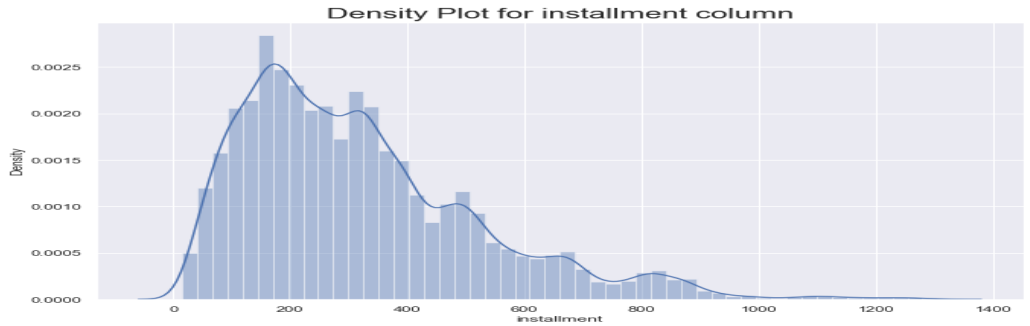


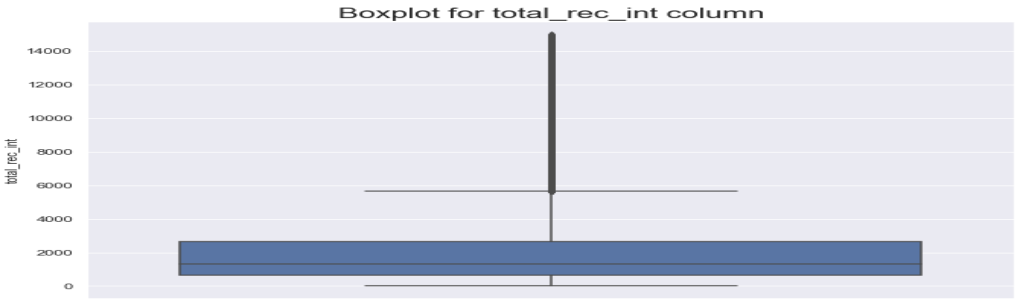


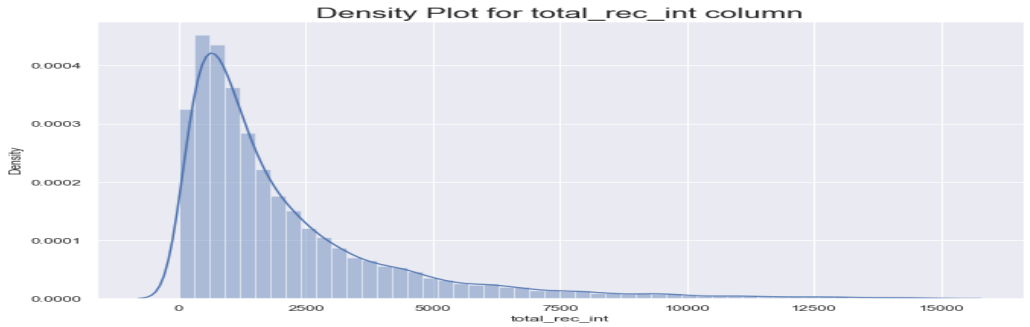
CONTINUOUS UNIVARIATE ANALYSIS

- **Installment and Total Received Interest**
  - **installment**: 50% of the values are between 180 and 420 and 99% of value are between 50 and 820. From 200, the number of borrowers owing the installments decreases as the installments increases. This tends to show that most of the installments are of smaller amounts
  - **total_rec_int**: Lesser borrowers pay more interest and the number of users decreases as the interest increases.



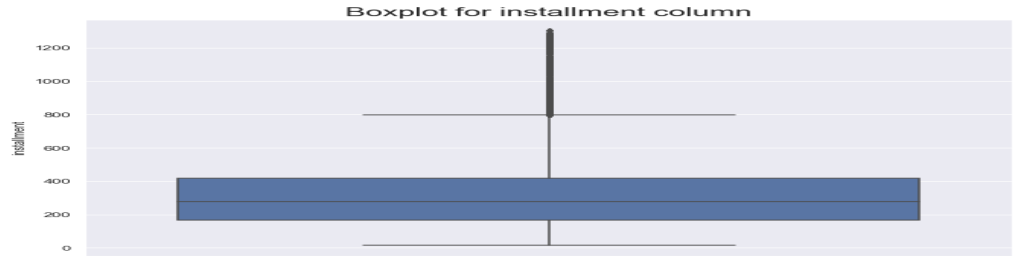


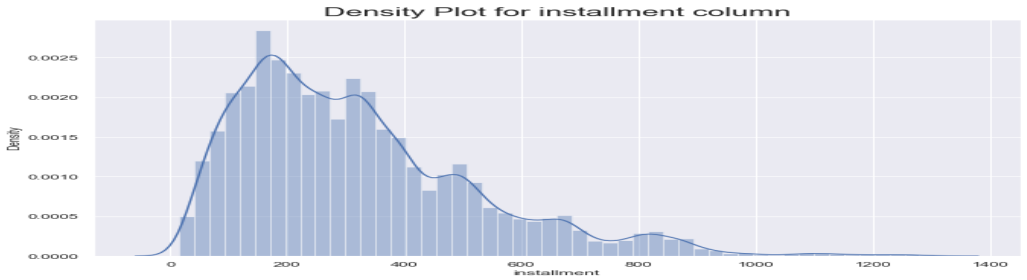


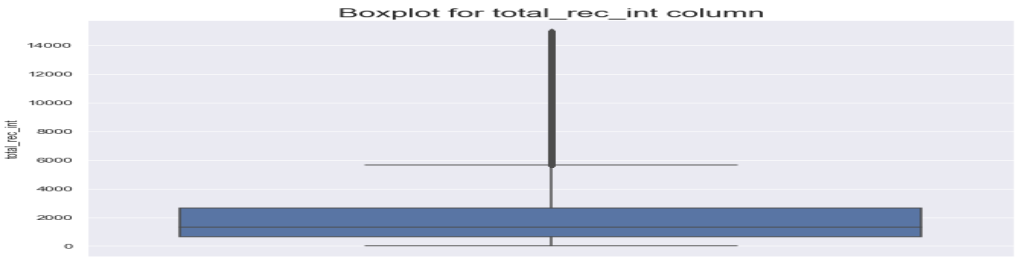


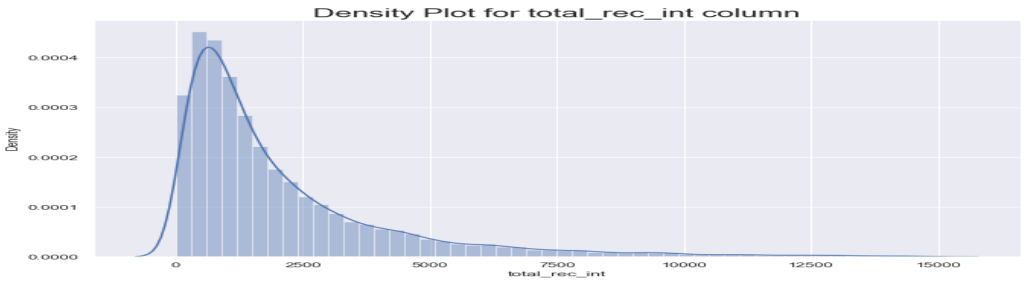
CONTINUOUS UNIVARIATE ANALYSIS

- **Installment and Total Received Interest**
  - **Installment**: 50% of the values are between 180 and 420 and 99% of value are between 50 and 820. From 200, the number of borrowers owing the installments decreases as the installments increases. This tends to show that most of the installments are of smaller amounts
  - **total_rec_int**: total_rec_int has an interesting smooth pattern. Seems like the lesser borrowers pay more interest and the number of users decreases as the interest increases.









BIVARIATE ANALYSIS

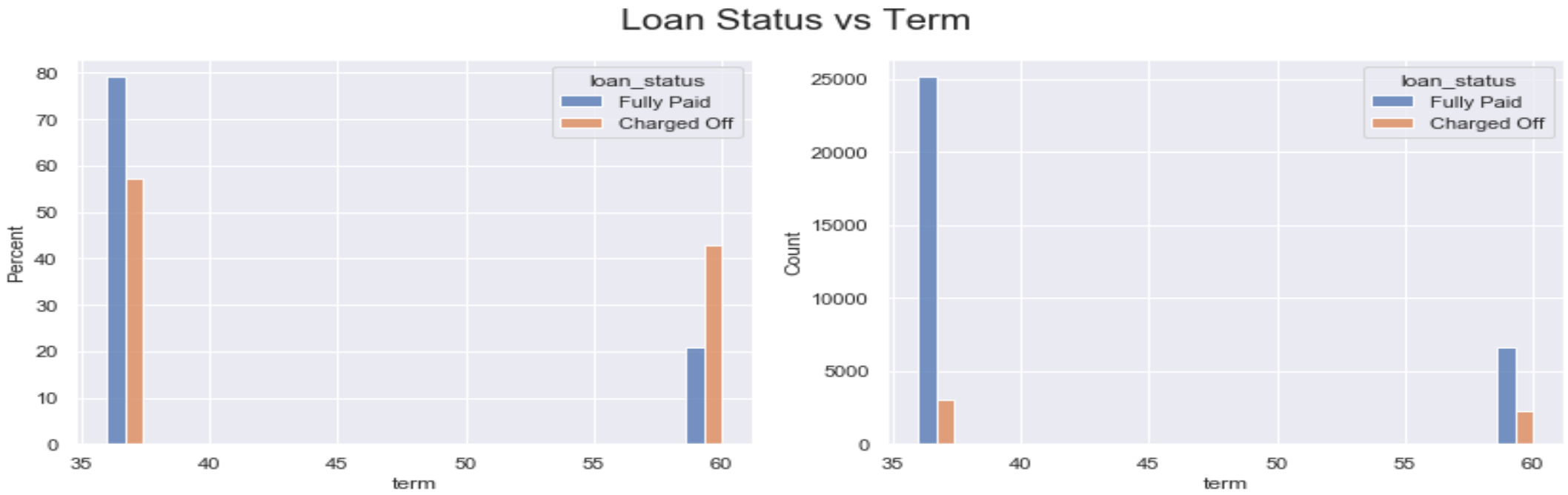
- Bivariate Analysis is one wherein we will analyze the interaction between 2 variables and understand if there exists any relation between the 2 variables. For our Bivariate analysis, we will focus more on interaction of other columns with respect to our target variable *loan_status*. We will further devide this univariate analysis into below 2 sections:
    - Categorical Bivariate Analysis
    - Continuous Bivariate Analysis

CATEGORICAL BIVARIATE ANALYSIS

- We will first focus on the bivariate analysis of the categorical variables with respect to our target variable *loan_status*. The plot we will use is that of *histplot*. Here we will create 2 histplots:
  - First one highlights the *Percent* of each category of *loan_status* for each segment of the analyzed categorical variable. The reason we are plotting this graph is that we the data is skewed towards *Fully Paid* category of *loan_status*. Hence just plotting a histplot of count wont yield any meaningful insights. Also, this will help us understand if the proportion of one particular category of *loan_status* dominates over other.
  - Second one displays the count of category of *loan_status* for each segment of the analyzed categorical variable.
- Of all the variables analyzed, we will only share the findings of those variables which has a meaningful impact on *loan_status*

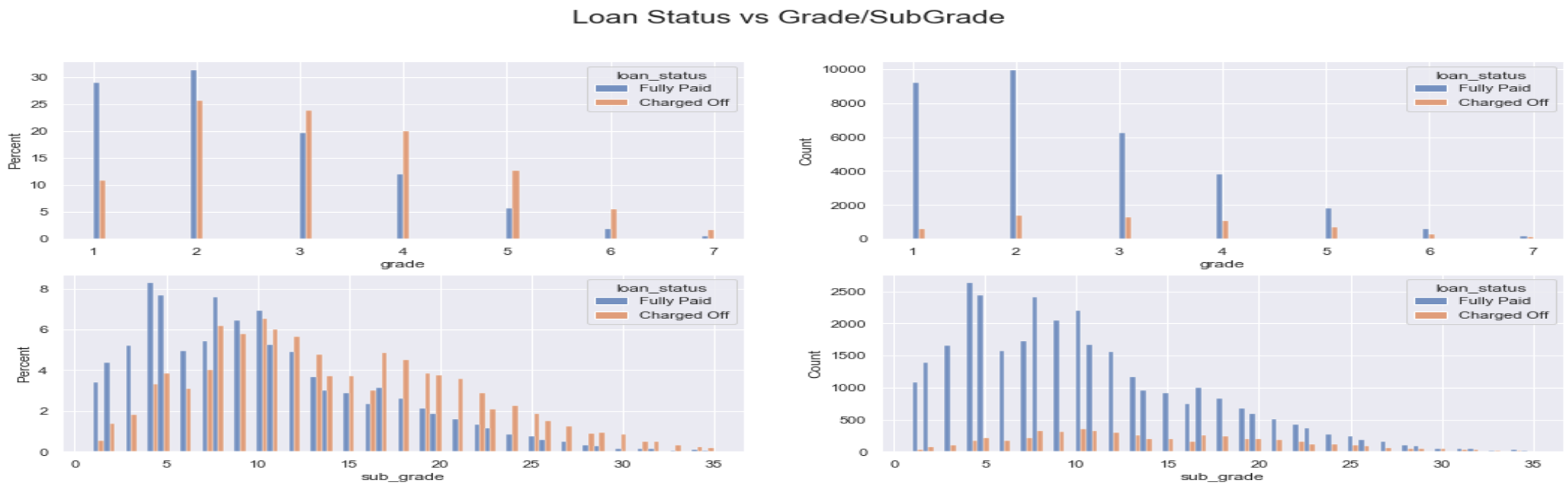
CATEGORICAL BIVARIATE ANALYSIS

- **Loan Status vs Term**
  - What we analyze here is that *Long term loans tend to result in higher proportion of defaulters when compared to short term loan*. One strong reason for this might be the fact that borrowers might be able to keep up with long term commitments and hence might fail to repay the full amount.



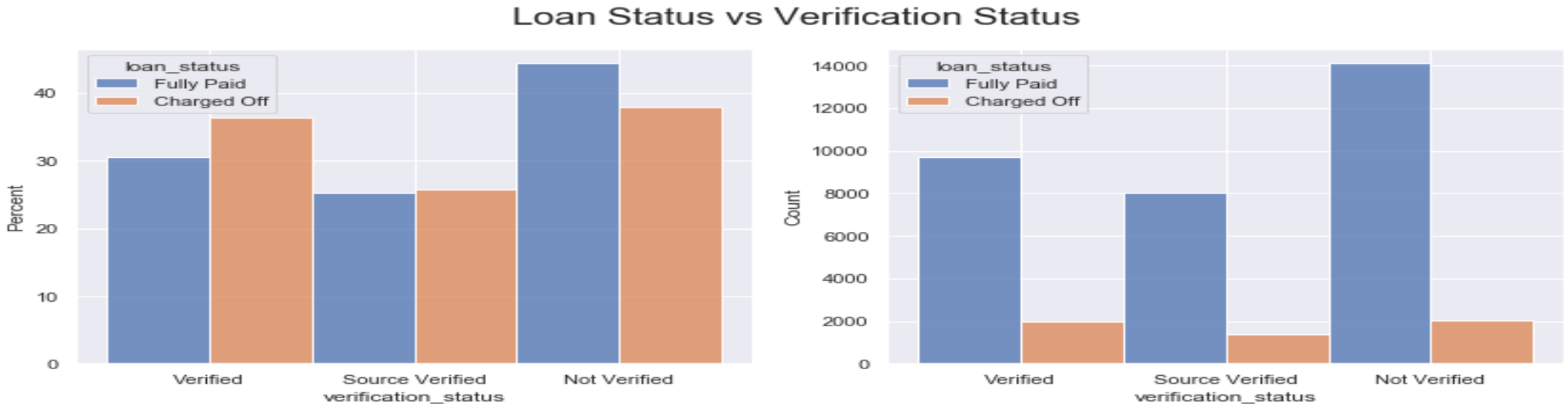
CATEGORICAL BIVARIATE ANALYSIS

- **Loan Status vs Grade/Sub Grade**
  - As per the analysis, it seems like *potential defaulters always tend to maintain a lower grade as compared to the non-defaulters*. This is very evident in the below graph, wherein the lower grades (C, D and E) have higher proportion of defaulters as compared to the higher grades



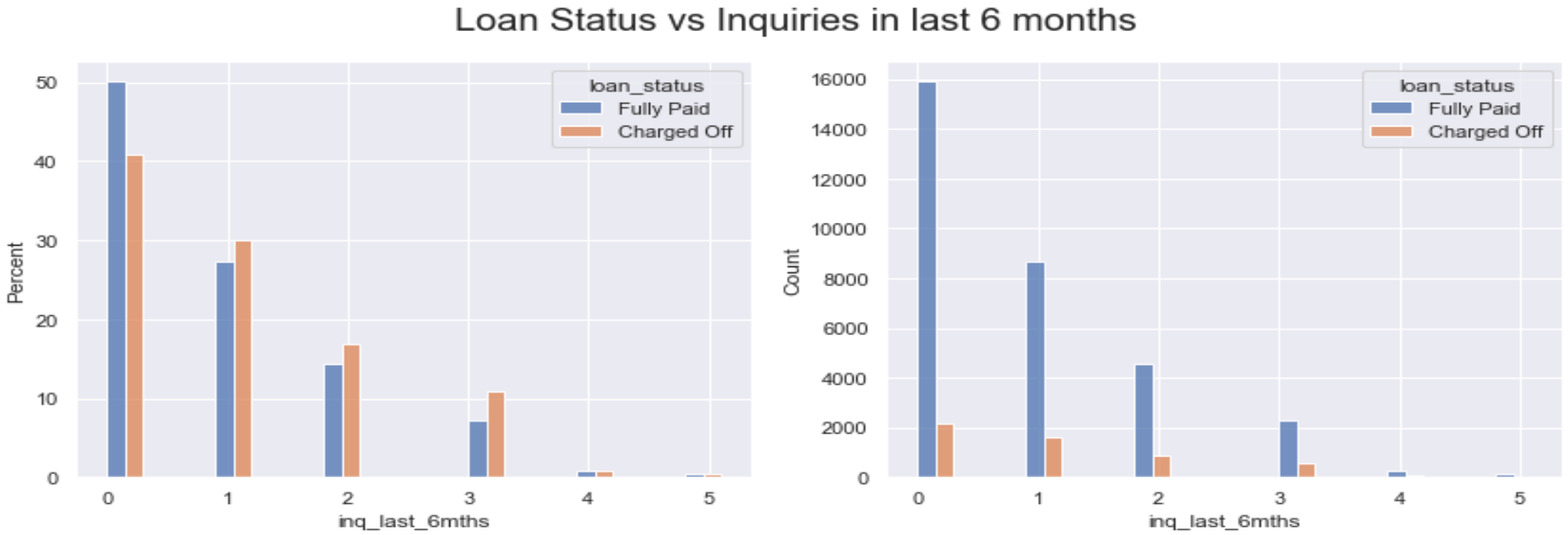
CATEGORICAL BIVARIATE ANALYSIS

- **Loan Status vs Verification Status**
  - A general though is that if the income or source of income is verified, then the proportion of defaulters should be less as compared to Not verified loan applications. But as per the below graph, the *proportion of defaulters in Verified and Source verified category is more than Not Verified*. This might highlight the fact the due-diligence followed in verifying the income needs to be revisited. The difference is not very high, and hence we cannot establish with certainty that there is a issue in the verification process, but surely the process can be reviewed.



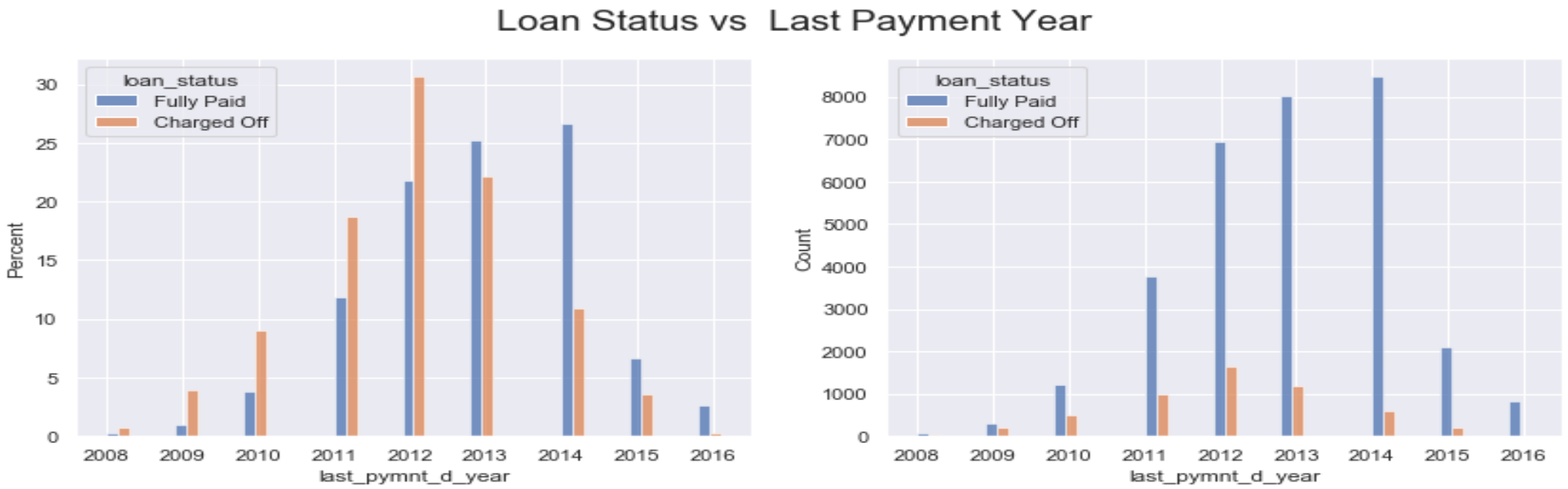
CATEGORICAL BIVARIATE ANALYSIS

- **Loan Status vs Inquiries in last 6 months**
  - Its observed that borrowers who defaulted inquired more number of times as compared to the Fully Paid borrowers. Hence keeping a track on the number of inquiries can be handy.



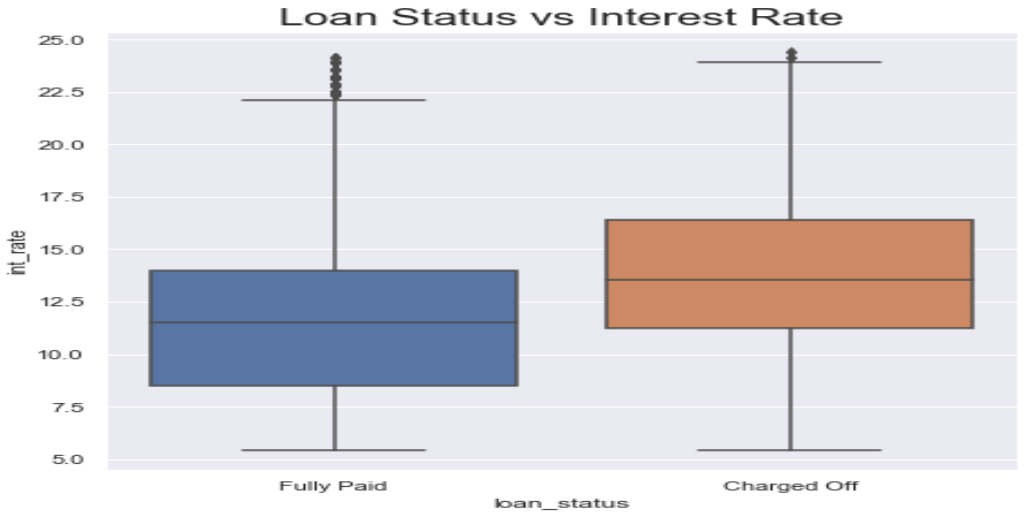
CATEGORICAL BIVARIATE ANALYSIS

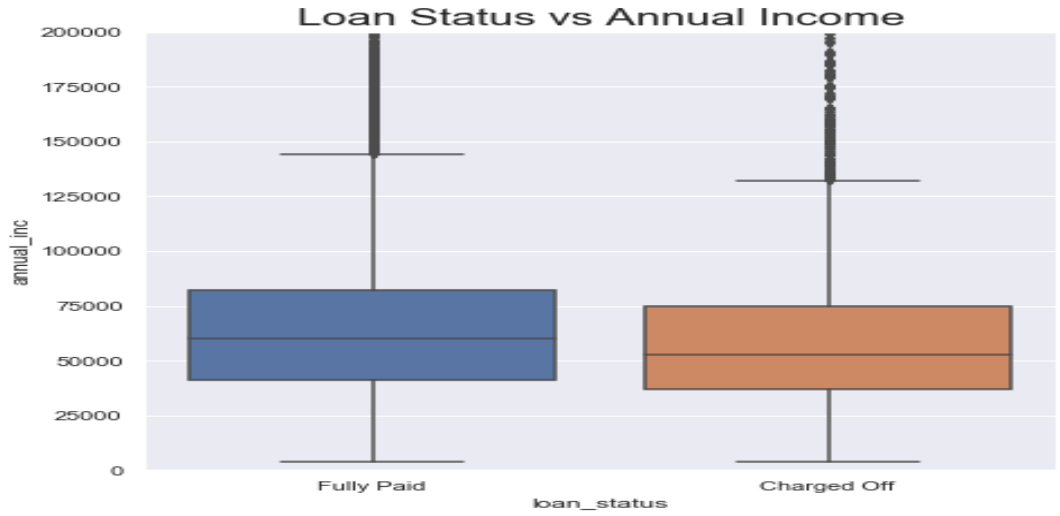
- **Loan Status vs Last Payment Year**
  - From the below graph its observed that defaulters tend to stop paying early as compared to the Fully Paid borrowers. Though this is not the factor which is leading to default, but tracking the last payment date and following up effectively might reduce the number of defaults.



CONTINUOUS BIVARIATE ANALYSIS

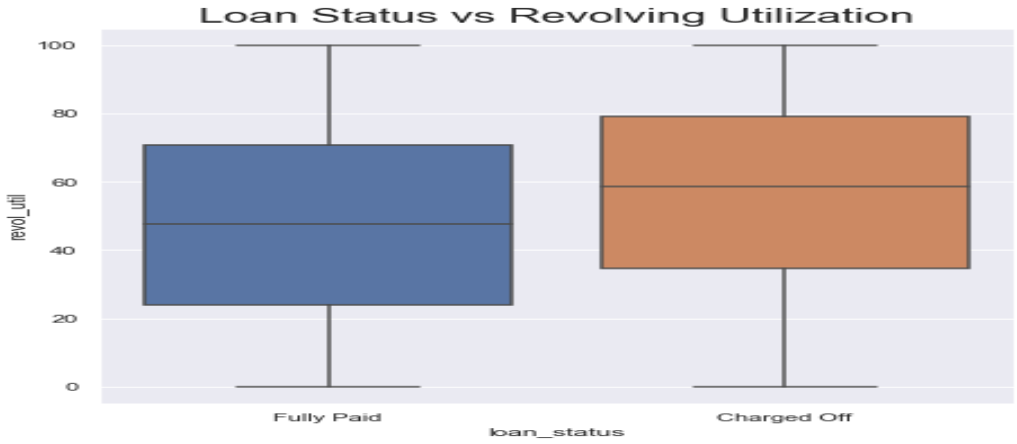
- **loan_status vs int_rate**
  - For charged-off loans, the interest rate seems to be higher when compared to Fully Paid. Charged-off loans have 2-3 points more when compared to Fully paid. This is substantial difference in the interest rate, and borrowers might not be able to repay the loan owing to higher interest rate. This can be a *strong influencer* of the default rate
- **loan_status vs annual_inc**
  - It can be observed that charged-off candidates have lower spread of annual income and lower mean and median as compared to the Fully Paid candidates.

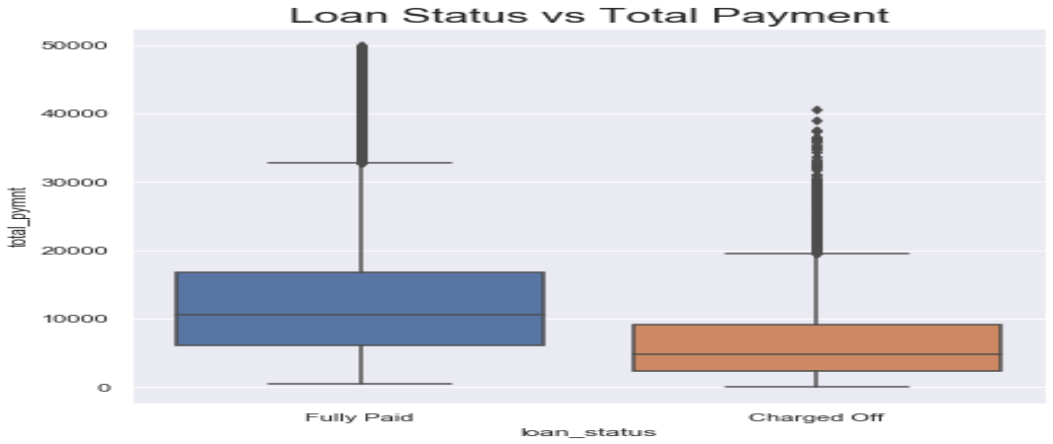




CONTINUOUS BIVARIATE ANALYSIS

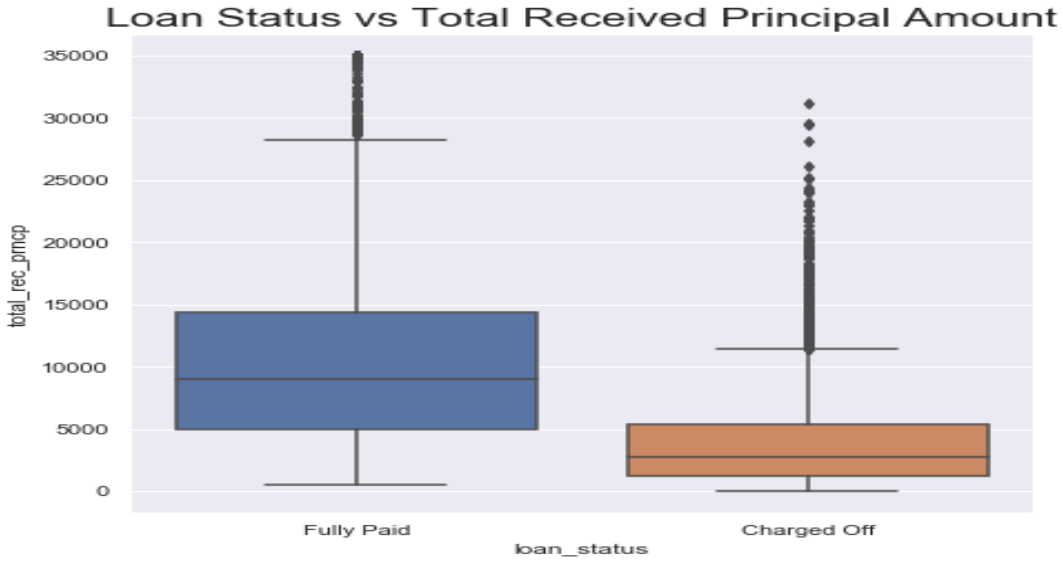
- **loan_status vs revol_util**
  - Revol_util for charged-off loan is significantly higher than fully-paid. While the spread is the same, more than 50% of charged off candidates have 10 points more than Fully-paid candidates. This indicates that charged-off candidates tend to utilize their credit to the limit and hence result in default. This is a *strong influencer* of loan status
- **loan_status vs total_pymnt¶**
  - Total_payment for charged-off is significantly lower than fully paid loans. This resonates with the fact that the charged off loans usually don't pay the full amount. But the difference between charged-off and Fully-paid is very significant. This might not be the driving factor, but instead can be the effect of a borrower who is going to default.

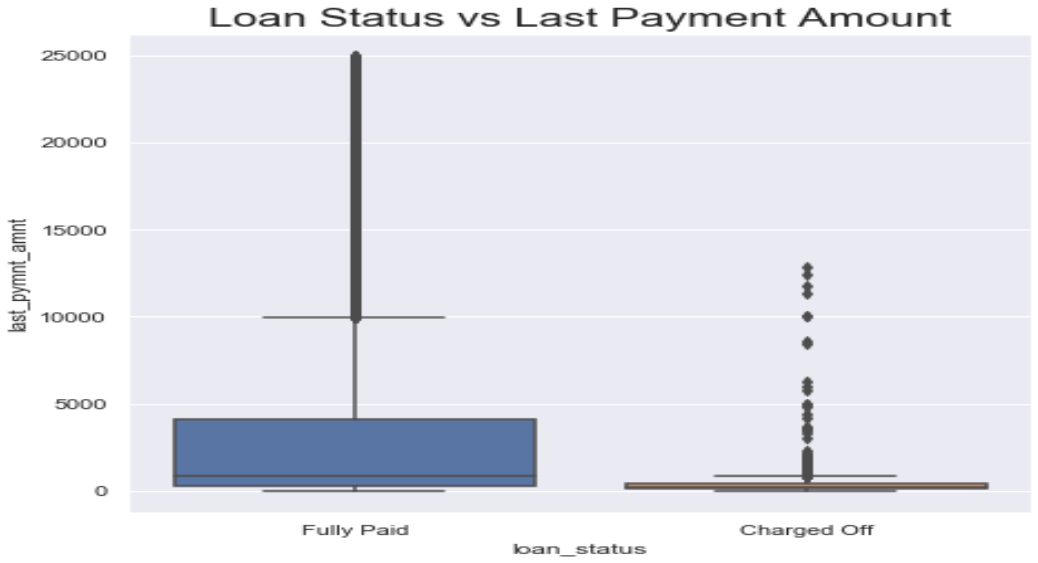




CONTINUOUS BIVARIATE ANALYSIS

- **loan_status vs total_rec_prncp**
  - There is strong correlation between total_rec_prncp & loan status. There is a difference of 60% between them. This might not be the driving factor, but instead can be the effect of a borrower who is going to default
- **loan_status vs last_pymnt_amnt**
  - For the charged-off loans the late payment amount is significantly low compared to fully-paid. The difference here is 4x. This is more of a correlation effect and less of a driving force.





LENDING CLUB EDA SUMMARY

- We will split the summary of our analysis into 2 sections:
  - **Findings**
  - **Recommendations**
- Under the *Findings* section, we will unravel all the interesting analysis that we have come across when we performed *Exploratory Data Analysis (EDA)* on the *loans.csv* dataset.
- Under the *Recommendations* section, we will interpret the findings, and make productive suggestions which can help business take accurate decision on whether they should approve a loan or reject it.

LENDING CLUB EDA: FINDINGS

- Below are the list of points that we have uncovered from the EDA of loans dataset. The below list covers the key parameters which would drive the possibility of loan turning default or not :
  - One of the most compelling find was that the **interest rates** for the borrowers who defaulted was higher when compared to the fully paid borrowers.
  - Drawing from the first point, when the interest rate is high, the amount paid as interest is also high. Hence the *defaulters ended up paying more interest amount but less principal amount*. This was evident in the *loan_status vs total_rec_prncp* bivariate analysis.
  - Its found that longer **duration of loan** resulted in higher percent of defaulters. This is congruous with the fact that longer the duration clubbed with higher interest rate, longer will be the obligations. This will lead to more borrowers to default.
  - Its observed that, lower the **grades** (eg: C , D or E), higher the proportion of borrowers defaulting. This shows that the grades, to a larger extent, reflects the financial health or the loan repayment tendency of a borrower.

LENDING CLUB EDA: FINDINGS

- Defaulters comparatively make higher **enquiries** when compared to the non-defaulters.
- Its observed that **revol_util** for charged-off loan is significantly higher than fully-paid. This indicates that charged-off candidates tend to utilize their credit to the limit and hence result in default.
- The **annual income** of defaulters is relatively lower than non-defaulters.

LENDING CLUB EDA: RECOMMENDATION

- After analyzing the above findings, i have listed down a consolidated list of suggestions which would help the organization in approving good loans and rejecting the bad loans:
  - The probability of a loan defaulting would drastically reduce if we **reduce the interest rates**. Its been observed that higher interest rates combined with longer loan tenure would result in higher interest amount to paid as compared to the principal amount. This would eventually result in the borrower not being able to keep up with the obligation and hence defaulting.
  - Extending point 1, **longer tenure period should be discouraged**, specifically for candidates who have higher interest rate or lower grade. Its observed that longer the tenure higher the probability of loan default.
  - The company should **reconsider approving loan for candidates who have a lower grade** i.e C, D or E.

LENDING CLUB EDA: RECOMMENDATION

- If the candidate, in the recent times, had **made multiple inquiries ,then such applications should be scrutinized**.
- Loan applications from candidates who **tend to utilize their credit to the limit should be re-considered**.
- **Higher income candidates should be given preference in approving loans**, as its observed that higher income candidates tend repay their loans when compared with lower income candidates.

THANK YOU