Master of Science
in Machine Learning
& AI

LENDING CLUB CASE STUDY

Mr. J. Naveen Kumar

INTRODUCTION

Lending club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return. This Company is the largest online marketplace , facilitating personal loans, business loans and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

How Lending Club Works

- Borrower apply for loans to a financial institution or companies.
- Then the Financial institution or companies will analyze the loan request and has to make a decision for approval
- Then once it is approved , it will be let to the investors who will provide the loan amount

PROBLEM STATEMENT

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:

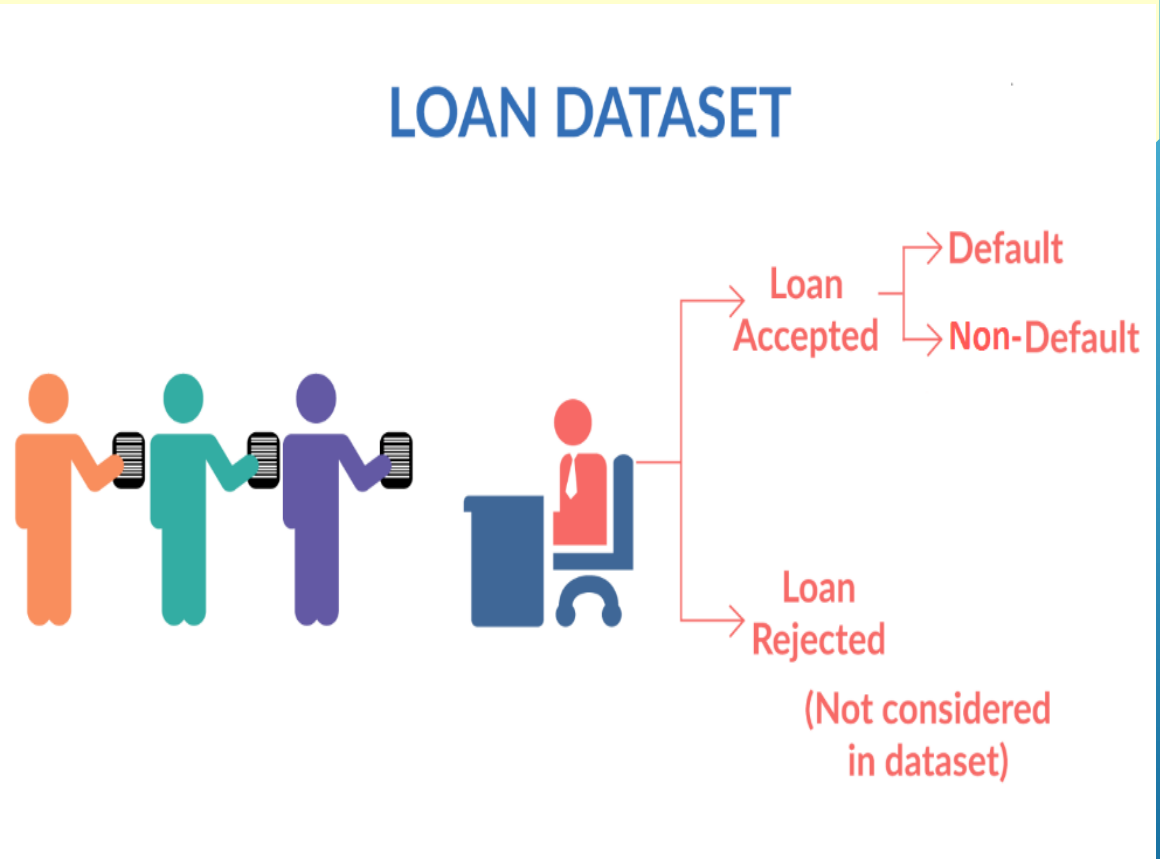
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial loss for the company

BUSINESS UNDERSTANDING

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



OVERVIEW

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Steps :

1. Data understanding
2. Data Cleaning and Manipulation
3. Data Analysis (univariate , bivariate and tri variate)
4. Conciseness And Readability Of The Code
5. Observations and Recommendations

DATA UNDERSTANDING

How the data is understood :

1. First had a look with the dataset which is being shared
2. Then Identify all the data quality issues.
3. Then reported those issues
4. Provided the detailed meanings for the variables used.

DATA CLEANING

How the data is cleaned:

1. Removed all the null values in all the column and rows
2. Removed unwanted columns which is not relevant for the analysis
3. Outlier Treatment done for the columns having outliers
4. Column's datatype are converted to a convenient format - Standardization
5. Created new derived columns for extensive analysis
6. Imputation done for few columns which will not impact any other columns for analysis
7. String and date manipulation is done correctly

OBSERVATIONS FOR DATA CLEANING

Before Cleaning

1. First the given dataset contains 39717 rows and **111 columns**

After Cleaning

1. First removed all the columns which has all rows as null values
2. It was around 54 columns.
3. Then Identified which are the columns has same value in all the rows.
4. After identified removed all those columns which was 20 columns
5. Then Removed all the Customer behavior modules which will not provide any information for the credit approving. So removed 18 columns
6. **Finally started the analysis with 19 columns**
7. Removed the certain rows by neglecting the current loan approvals rows since it is not required.
8. Data Analysis will be for **defaulter's vs Non defaulters**

DATA ANALYSIS

How The Data Is Analyzed:

- Univariate And Segmented Univariate Analysis Is Done Correctly And Appropriate Realistic Assumptions Are Made Wherever Required. The Analyses Successfully Identify At Least The 5 Important Driver Variables are identified.
- Business-driven, Type-driven And Data-driven Metrics Are Created For The Important Variables And Utilized For Analysis. The Explanation For Creating The Derived Metrics Is Mentioned And Is Reasonable.
- Bivariate Analysis Is Performed Correctly And Is Able To Identify The Important Combinations Of Driver Variables. The Combinations Of Variables Are Chosen Such That They Make Business Or Analytical Sense.
- The Most Useful Insights Are Explained Correctly In The Jupyter Notebooks.
- Appropriate Plots Are Created To Present The Results Of The Analysis.
- The Choice Of Plots For Respective Cases Is Correct. The Plots Clearly Present The Relevant Insights And Easy To Read. The Axes And Important Data Points Are Labelled Correctly.

DATA ANALYSIS

Univariate Analysis:

Univariate Analysis for Ordered and unordered Categorial variables

Highlights

- 1. Only a smaller number of loan applications gets charged off compared to fully paid. Almost 30000 applications are fully paid, remaining 5000+ applications are charged off.
- 2. Most of the loan Applications fall in the 36 terms and less in 60 terms.25000+ are in 36 months term and 60 months term lies in the range of (5000 - 10000)
- 3. 15000 Loan Applications are not verified , 10000+ applications are verified, and source verified lies in the range of (5000 - 10000)
- 4. Public Bankruptcies with 0 has maximum loan applications. Around 30000+ applications are with 0 value.(clean records)
- 5. Rent is the highest homeownership applied for the loan applications around 15000+. The next is mortgage with 15000 and (own, others) are very minimal
- 6. Emp length who has 10 years of experience has the maximum loan applications which is around 8000
- 7. Grade with B has the higher loan applications around 10000 and it goes on next with A,C,D.
- 8. Northeast and West are the areas where more amount of loan applications received.

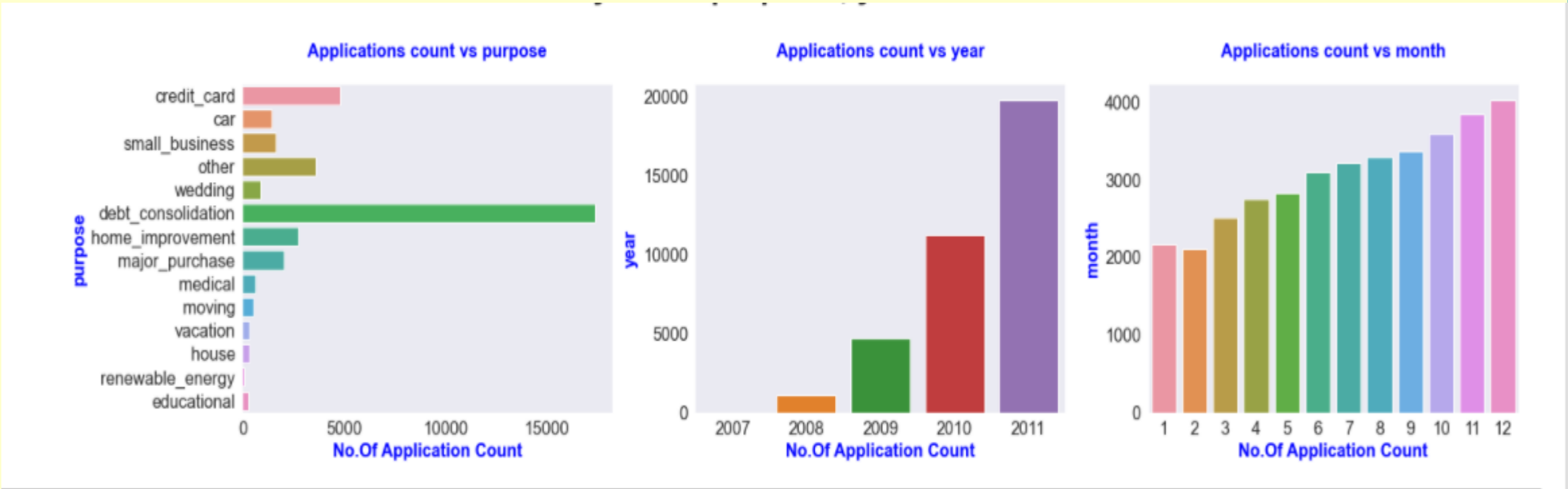
DATA ANALYSIS

Univariate Analysis:

Univariate Analysis for Ordered and unordered Categorial variables

Highlights

- Debt Consolidation is the highest chosen purpose among all the loan applications. The next is credit card
- 2011 is the year having more no of loan applications.
- Loan Application count is gradually increasing with the increase of the year.
- Loan Application count is gradually increasing with the increase of/; the month.
- Dec is the month having more no of loan applications may be due to more festival seasons, they might have opt the loan..



DATA ANALYSIS

Univariate Analysis:

Univariate Analysis for Ordered and unordered Categorial variables

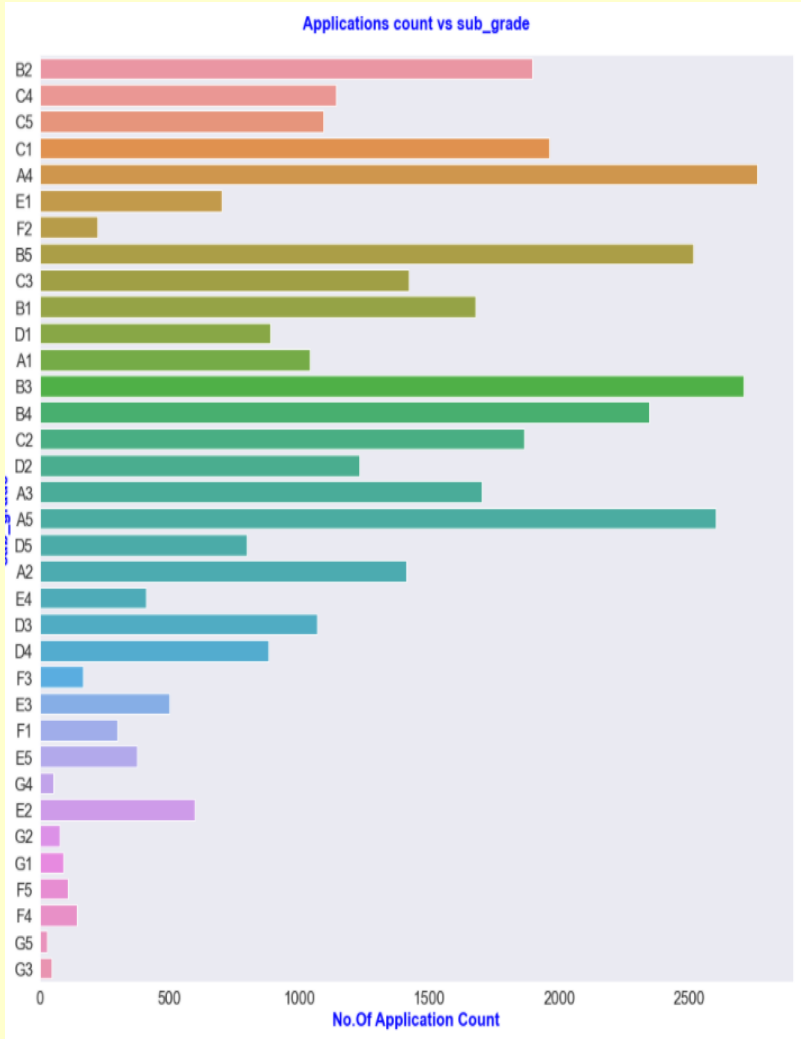
Highlights

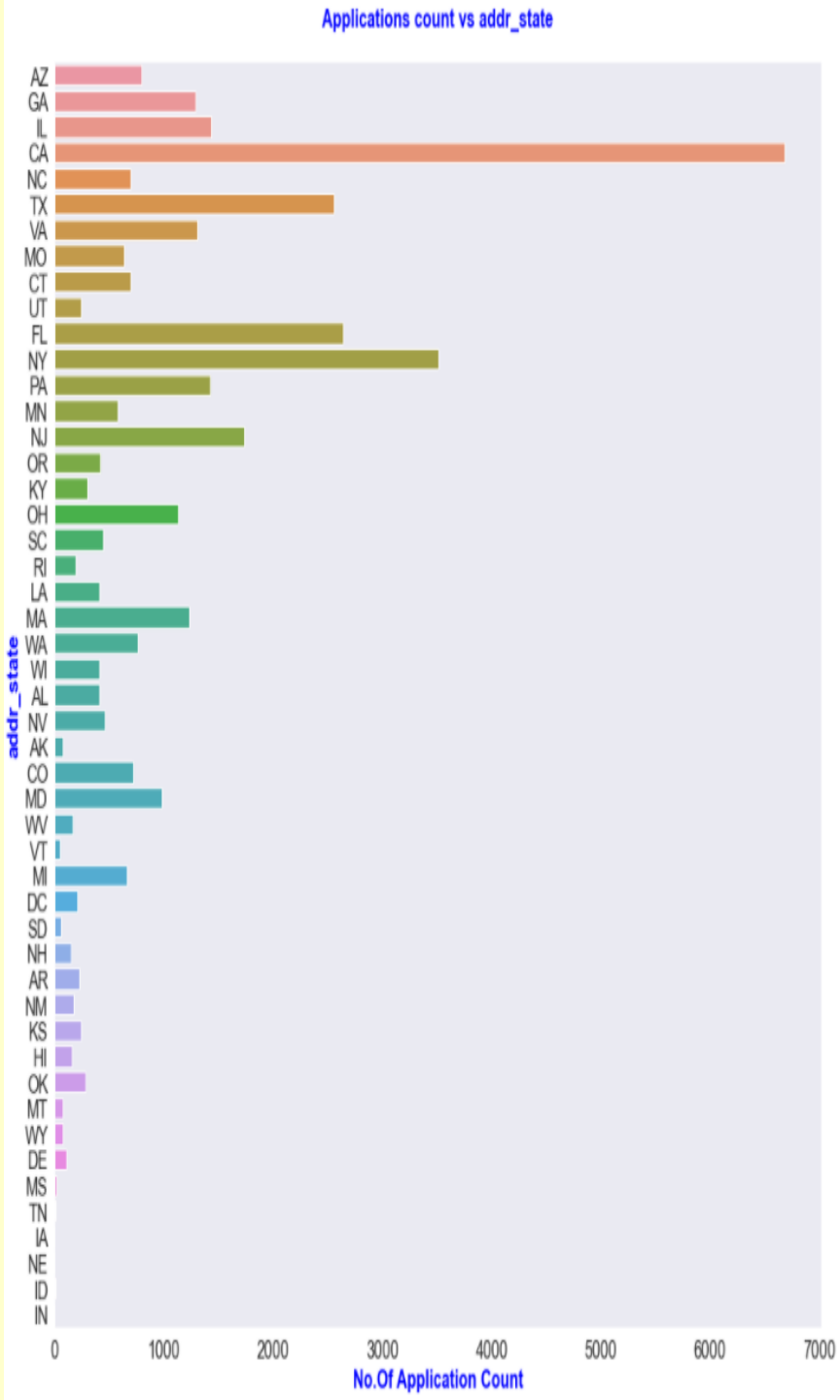
Addr_state

- CA(California) is the highest addr_state to have more no of loan applications.
- Next is NY (New York) is second
- TX and FL are the next in the order

Grade

- 1A4, B3, A5,B5,B4 are the grades having more no of loan applications.
- Most of the A and B Grades are the ones having more no of loan applications





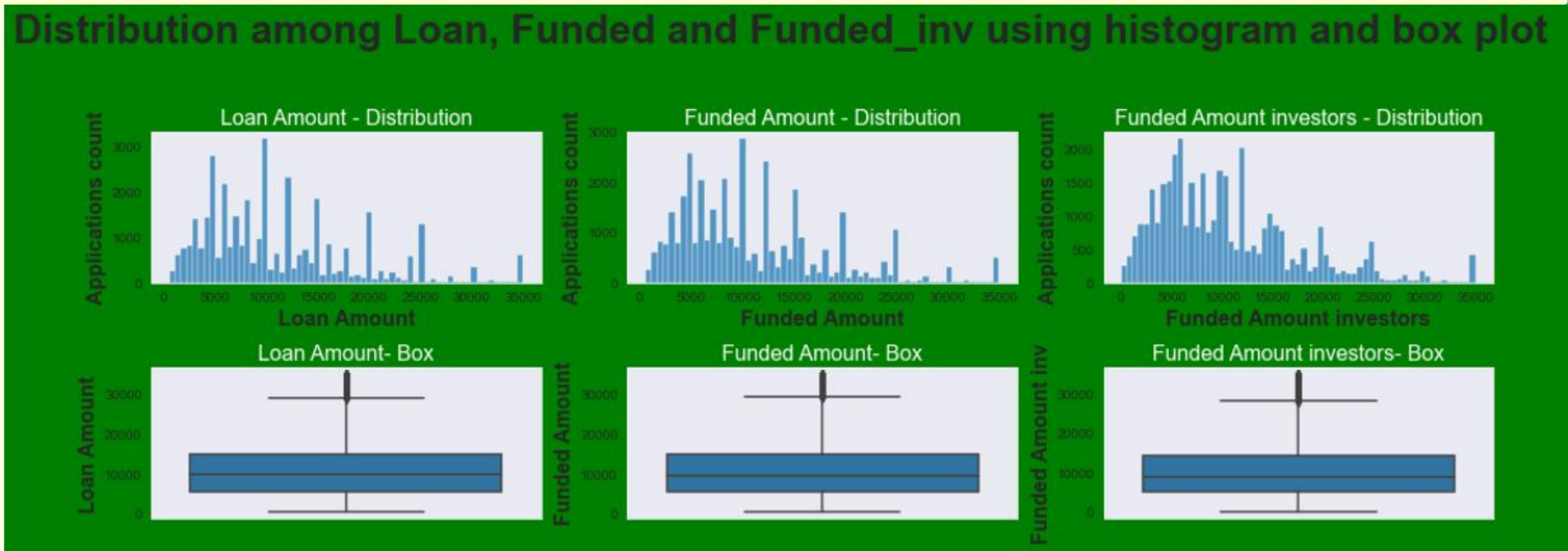
DATA ANALYSIS

Univariate Analysis:

Univariate Analysis for quantitative variables

Highlights

- 1. All the three-histogram chart are more similar to each other
- 2. So Only Loan Amount can be taken into consideration for future analysis
- 3. More amount of loan applications lies in the range of 5000 to 10000 for all the three columns
- 4. Also all the three columns has not much outliers to be treated.



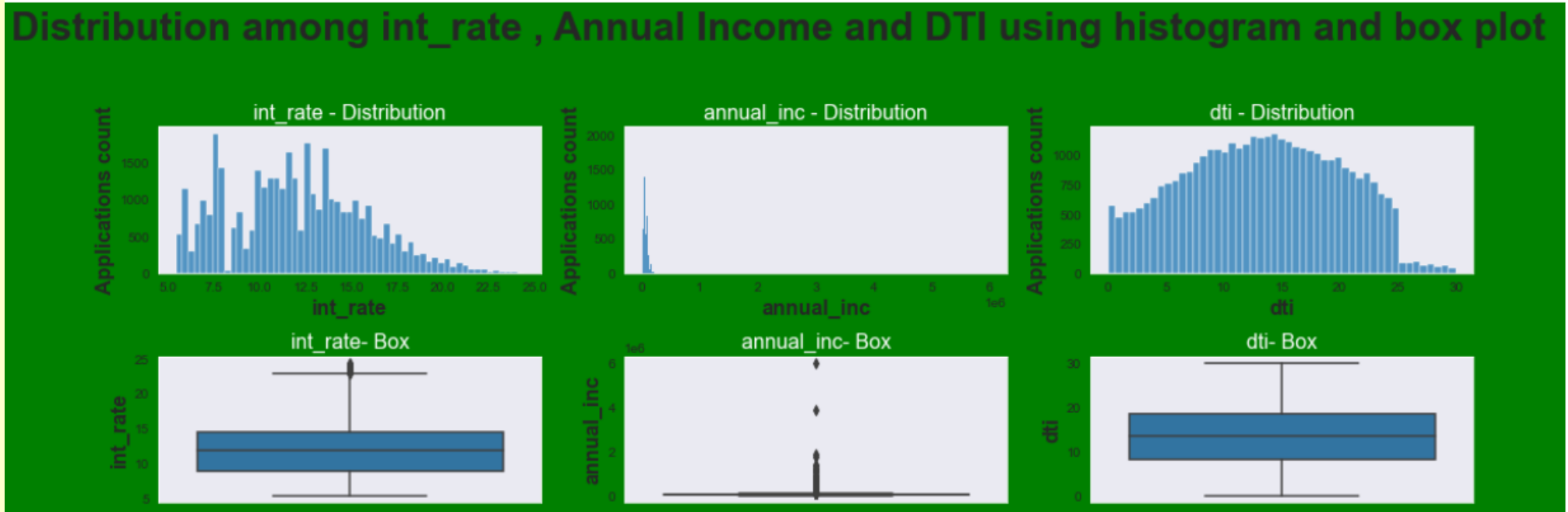
DATA ANALYSIS

Univariate Analysis:

Univariate Analysis for quantitative variables

Highlights

- Loan Applications count for the Dti Column Distribution is gradually getting increased when the dti increases and attains a peak at (12 to 15) range
- Having less no of applications count when dti is more than 25.
- It also stars gradually decreasing after attaining the peak around 15.
- Annual Income Column has more outliers. So outlier treatment is needed to remove the outliers
- Dti and int_ rate has very minimal outliers which can be neglected
- There is a sudden dip in the interest rate after 7.5 % and then it starts gradually increasing



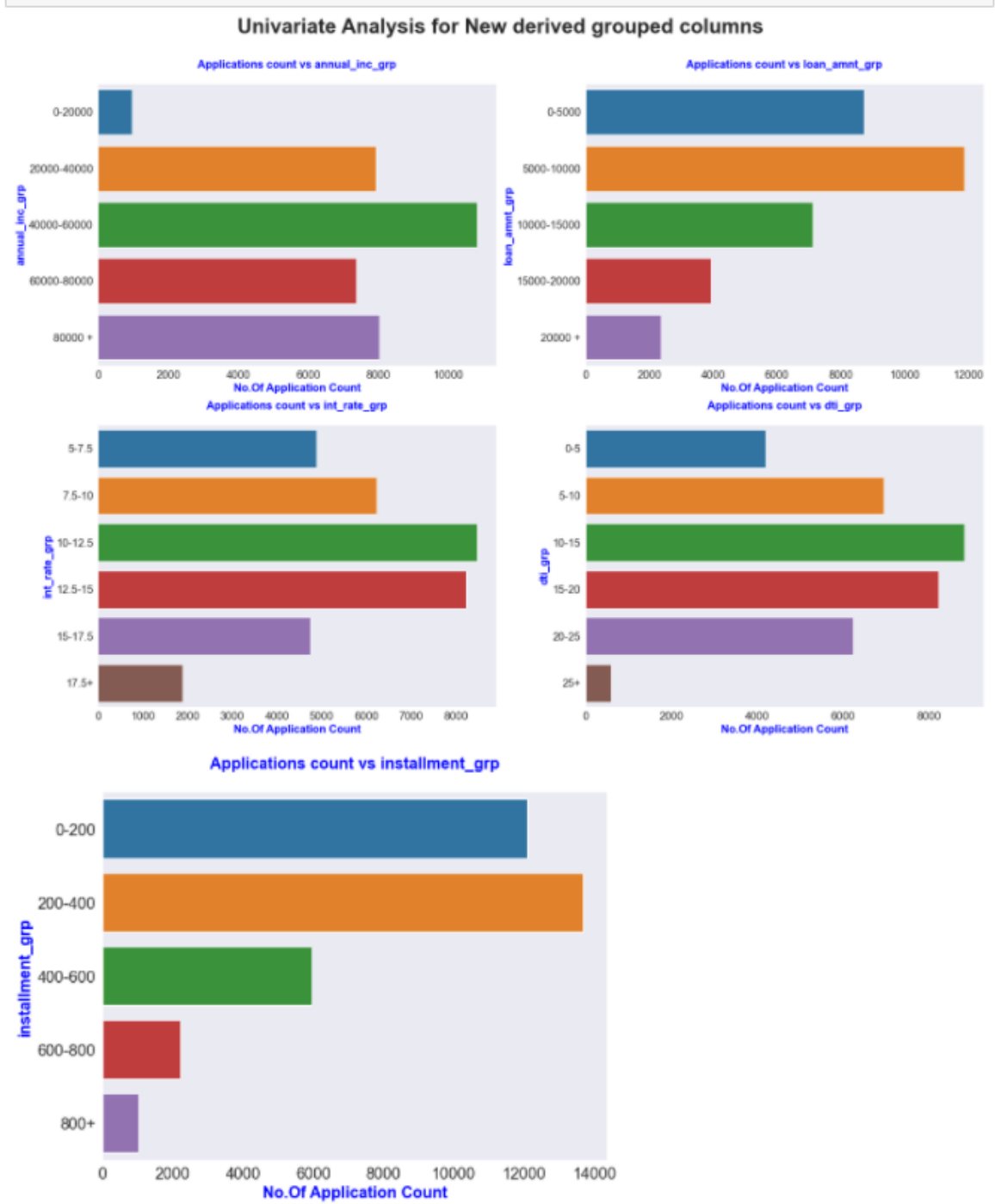
DATA ANALYSIS

Univariate Analysis:

Univariate Analysis for derived columns

Highlights

- Loan Amount grp - The Maximum no of loan applications lies in the range of 5000 - 10000
- Annual income grp - The Maximum no of loan applications lies in the range of 40000 - 60000
- int rate grp - The Maximum no of loan applications lies in the range of 10 - 12.5 %
- dti grp - The Maximum no of loan applications lies in the range of 10 to 15
- Installment - The Maximum no of loan applications lies in the range of 200 to 400



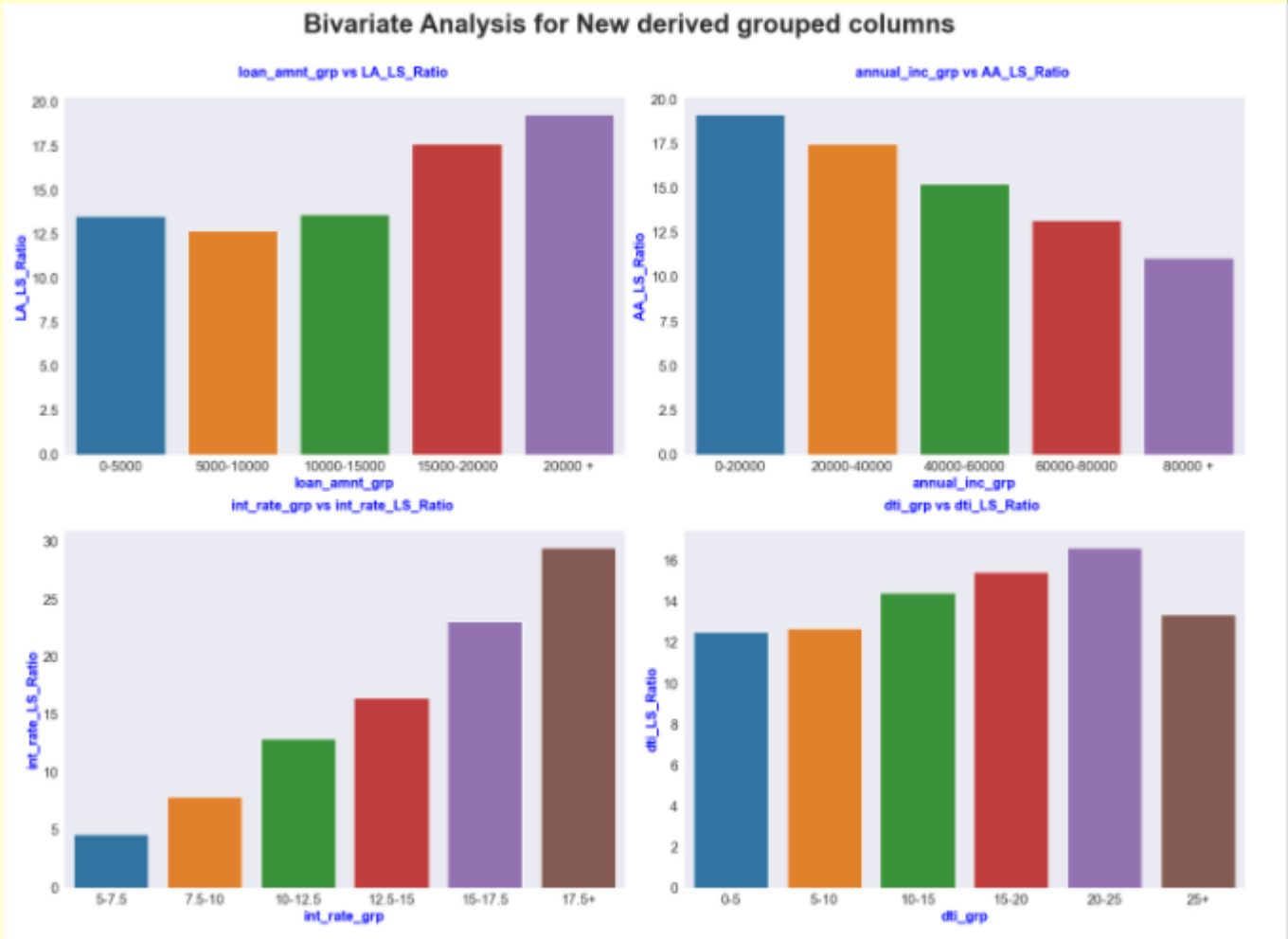
DATA ANALYSIS

Bivariate Analysis:

Bivariate Analysis for derived columns

Highlights

- 1. The borrowers who ask Loan amount around 20000 has more changes of defaulting.
- 2. Higher the loan amount increases , the chances of defaulters is increased
- 3. The borrowers who has low income has more chance of getting defaulted
- 4. Higher the Annual income, the changes of defaulters can be decreased.
- 5. Loan amount and Annual income is negatively correlated when it is compared with the charged off (defaulters)
- 6. The Borrowers who has got high interest rate has more chance of getting defaulted
- 7. Higher the interest rate increases , the chances of defaulters is increased
- 8. Loan amount and interest rate is positively correlated when it is compared with the charged off (defaulters)
- 9. The Borrowers who has dti around (20-25) will have more chances of getting defaulted.



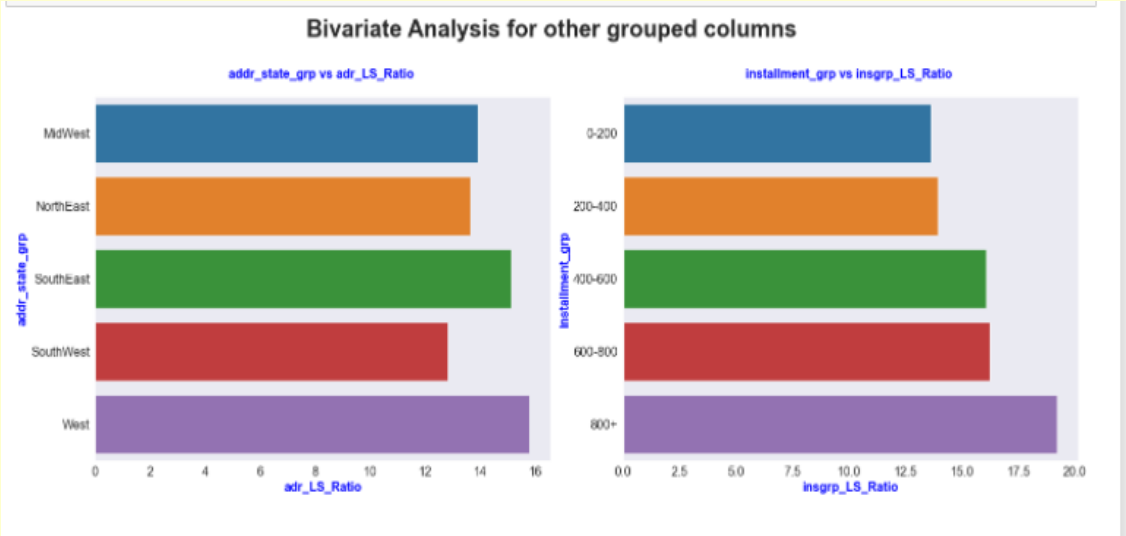
DATA ANALYSIS

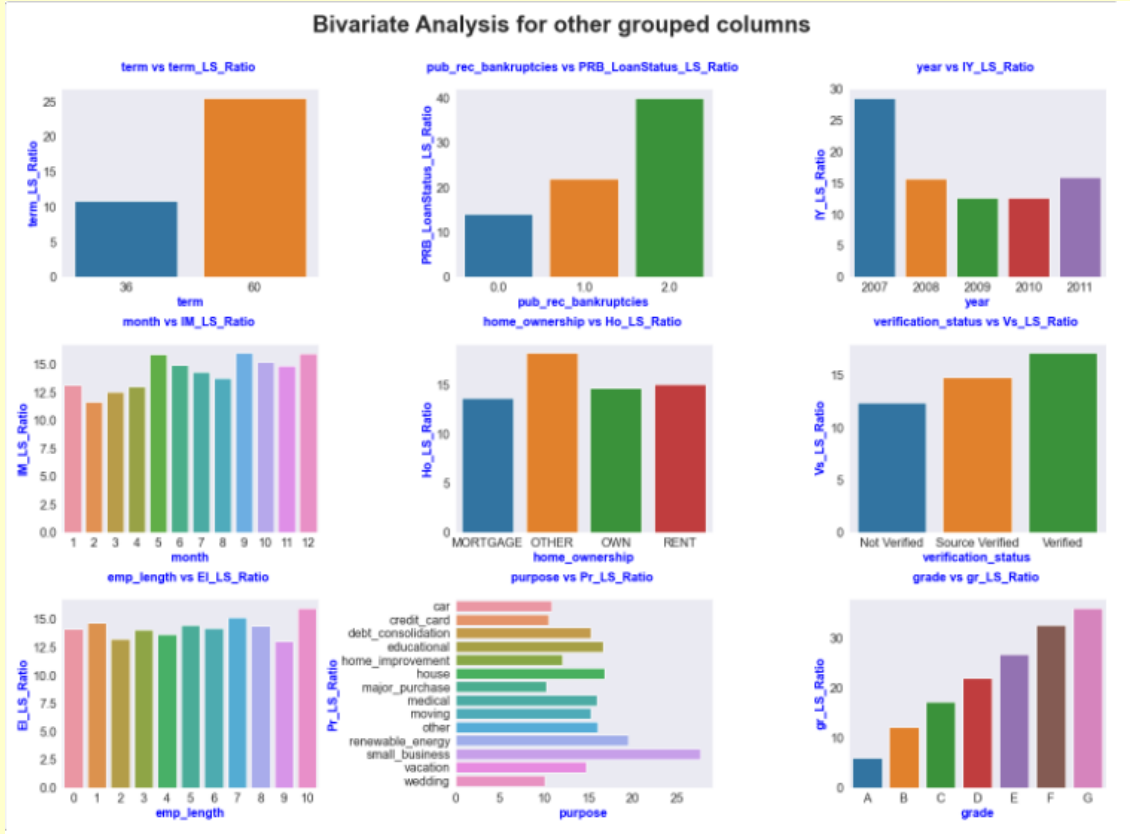
Bivariate Analysis:

Bivariate Analysis for other grouped columns

Highlights

- 1. The borrowers who ask Loan amount around 20000 has more changes of defaulting.
- 2. Higher the loan amount increases , the chances of defaulters is increased
- 3. The borrowers who has low income has more chance of getting defaulted
- 4. Higher the Annual income, the changes of defaulters can be decreased.
- 5. Loan amount and Annual income is negatively correlated when it is compared with the charged off (defaulters)
- 6. The Borrowers who has got high interest rate has more chance of getting defaulted
- 7. Higher the interest rate increases , the chances of defaulters is increased
- 8. Loan amount and interest rate is positively correlated when it is compared with the charged off (defaulters)
- 9. The Borrowers who has dti around (20-25) will have more chances of getting defaulted.





CONCISENESS AND READABILITY OF THE CODE

How The Code is written:

- The Code Is Readable With Appropriately Named Variables And Detailed Comments Are Written Wherever Necessary.

- He Code Is Concise And Syntactically Correct. Wherever Appropriate, Built-in Functions And Standard Libraries Are Used Instead Of Writing Long Code (If-else Statements, For Loops, Etc.).

Observations and Recommendations

1. The term which has 60 months has more defaulters
2. The bankruptcies who has 2 has more chances of defaulting
3. 2007 has the most no of defaulters
4. Then defaulters got decreased after 2007 and in 2011 it starts again increasing
5. 5th,9th,11th and 12th months has more no of defaulters. All these months have more festivals at that time
6. Employee length who has 10 years of experience has more chances of defaulting
7. In home ownership , other grp has the more chances of getting defaulters
8. verified applications has more no of defaulters than not verified
9. Small business has more no of defaulters and next is renewable energy has more no of defaulters
10. Defaulters getting gradually increased when the grades are decreased. G grade has more number of defaulters.

Thank you

Adapt it with your needs and it will capture all the audience attention.