Lending Club Case Study

9th March 2022

- Prashant Mohan Sinha

Agenda

- Introduction
- Problem Statement
- Exploratory Data Analysis
- Tools Used
- Input File and Packages Used
- Data Sourcing and Understanding
- Data Cleaning
- Data Analysis – Univariate and Bivariate
- Conclusion

Introduction

- Lending Club is a facilitator of Loans which connects borrowers seeking a loan with investors who provide them and expect a return.

- When an applicant submits a loan application, the company has to decide whether to approve a loan by considering the factors of applicant's profile.

Problem Statement

The decision of the company is associated with two business risks:

- If the applicant is likely to repay the loan, then the company's decision to not approve the loan is a loss.
- If the applicant is likely to default or not to repay the loan, then the company's decision to approve the loan is a loss.

Problem Statement

- It is important for a company to understand the factors or variables of loan default, which in turn is helpful in reducing Risky Loan Applicants.
- The aim of this Case Study is to analyze the factors or variables of Risky Loan Applicants using Exploratory Data Analysis technique.

Exploratory Data Analysis

Exploratory Data Analysis is a Statistical Technique of analyzing data sets to extract the summary of their characteristics. It uses Statistics, Graphs and Visualization methods. The EDA process involves the following steps:

- Data Sourcing
- Data Cleaning
- Data Analysis – Univariate and Bivariate

Tools Used

- Python Programming Language.
- Anaconda Packages for Machine Learning and Data Science.
- Jupyter Notebook.
- Microsoft Excel

Input File and Packages Used

Input File :

- loan.csv

Packages Used :

- Numpy
- Pandas
- Matplotlib
- Plotly
- Seaborn

Data Sourcing and Understanding

The input file is read and the some of the observations are obtained.

There are total 39717 entries in the given file ('*loan.csv*') and 111 attributes for each id.

The summary of data type of the attribute are as follows:

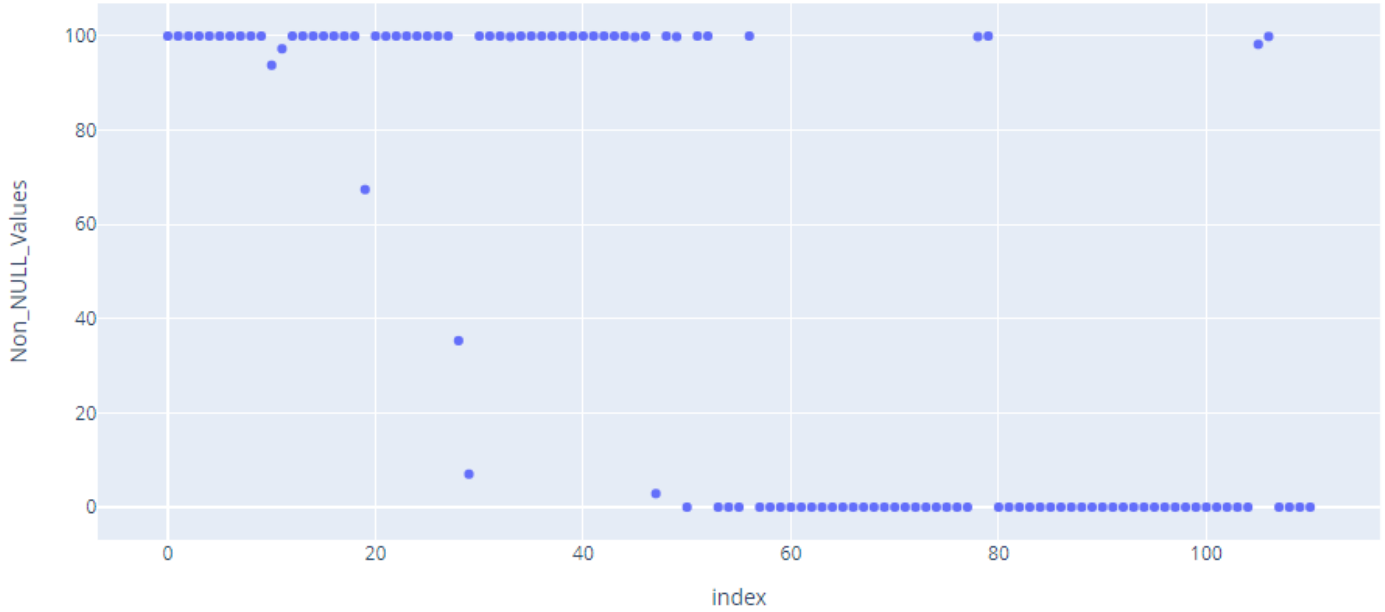
- **Numerical Variables**: 74 columns of float64 type and 13 columns of int64 type
- **Categorical Variable**: 24 columns

Data Sourcing and Understanding

- The above summary is based on just after merely loading the given file. It may likely that some of columns have partial data or complete lack of any data.
- A Scatter Plot is used to understand the percentage of values a column or a variable contains. The following slide represents the same.

Data Sourcing and Understanding

Scatter Plot of Non-NULL Data Column



Data Sourcing and Understanding

- There are 54 Columns with very few or No Missing Values and 57 columns having less than 40% of total entries.
- There is one column name 'desc' (data type object), which is approx. 67.42% of total entries. For the sake of keeping the information provided in description, we will fill it with 'Not_Mentioned'.
- With the above observation, we can divide the data columns in two parts:
    - Accepted columns >= 60%
    - Rejected Columns <60 %

Data Sourcing and Understanding

- We can now filter the Dataframe with accepted columns.
- Data understanding will continue as we forward with data cleaning & formatting.

Data Cleaning

Data Cleaning is a process of handling Data Quality Issues involving the following steps :

- Fix rows and columns
- Fix missing values
- Standardize values
- Fix invalid values
- Filter data

Data Cleaning

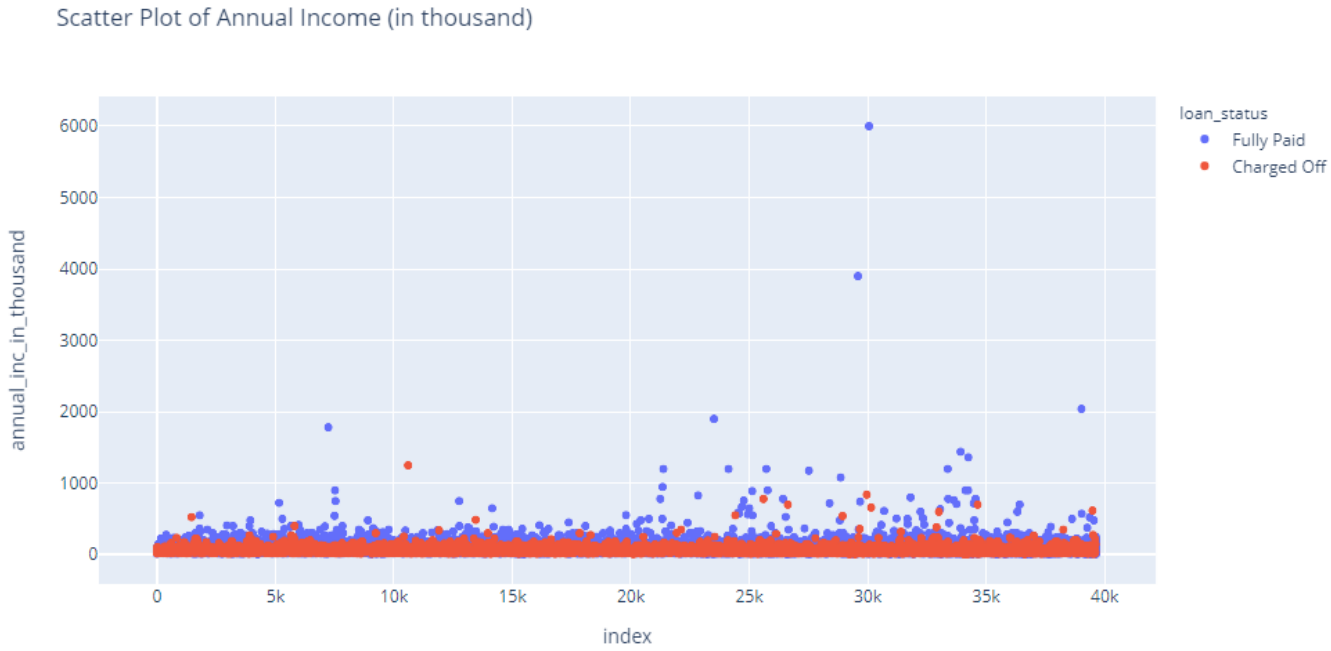
Data Cleaning process in this Case Study involves the following steps :

- Drop the No Use Columns (having Missing Values in all entries) from DataFrame.
- Find the number of distinct values in each columns & drop the columns having only one unique value.
- Drop the Rows in **loan_status** column having values as ‘Current’.
- Type conversion of columns: **'term'** & **'int_rate’**. (Removing ‘months’ and ‘%’ string in the values of respective columns and change their format to integer and float respectively).

Data Cleaning

- Drop the rows where **emp_length** is Missing.
- Type conversion of column **'emp_length'** (Removing ‘years’ and ‘+’ string and converting the values to Integer).
- Conversion of values of ‘**annual_inc**’ column from Million Units to Thousand Units.
- Imputation of ‘**pub_rec_bankruptcies**’ column by filling the missing values with median value.
- Copy the DataFrame as Master Data.

Data Sourcing and Understanding



Data Cleaning

Observations obtained when '**annual_inc**' column is converted from Million to Thousand Units :

- 75 % Quantile of Annual Income is 83000.
- From the above chart, the annual income above 1000k are all have paid the loan amount except one outlier case of charged off case where the income is 1200k and loan is taken for debt consolidation.
- There is big gap between in general average annual income & high income people. To make data distibution not biased towards high income, rows(14) having annual income 1000K are removed.

Data Analysis

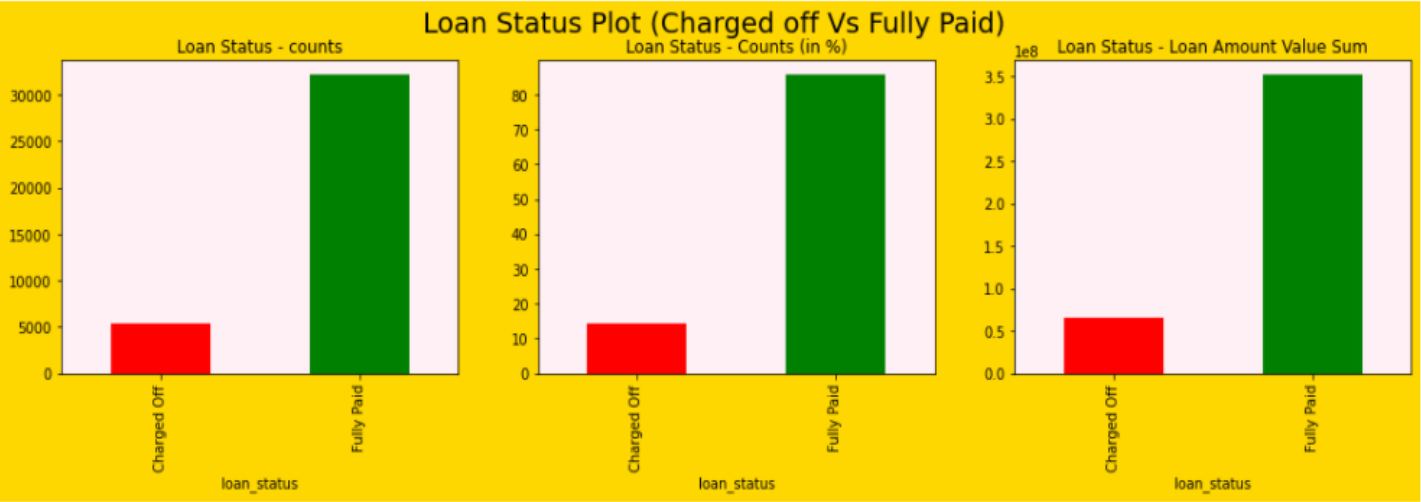
Data Analysis is a technique of inspecting the obtained data in order to derive useful information out of it.

In this Case Study, we used Univariate and Bivariate Analysis.

- **Univariate Analysis** – Analysis of one variable or column at a time.
- **Segmented Univariate Analysis** – Analysis of one variable or column at a time by grouping them based on various segments of data.
- **Bivariate Analysis** – Analysis of two variables or columns at a time.

Data Analysis – Univariate and Segmented Univariate

Loan Status of Fully Paid vs Charged Off




```
By percentage :  loan_status
Charged Off    14.0
Fully Paid     86.0
Name: loan_amnt, dtype: float64
```

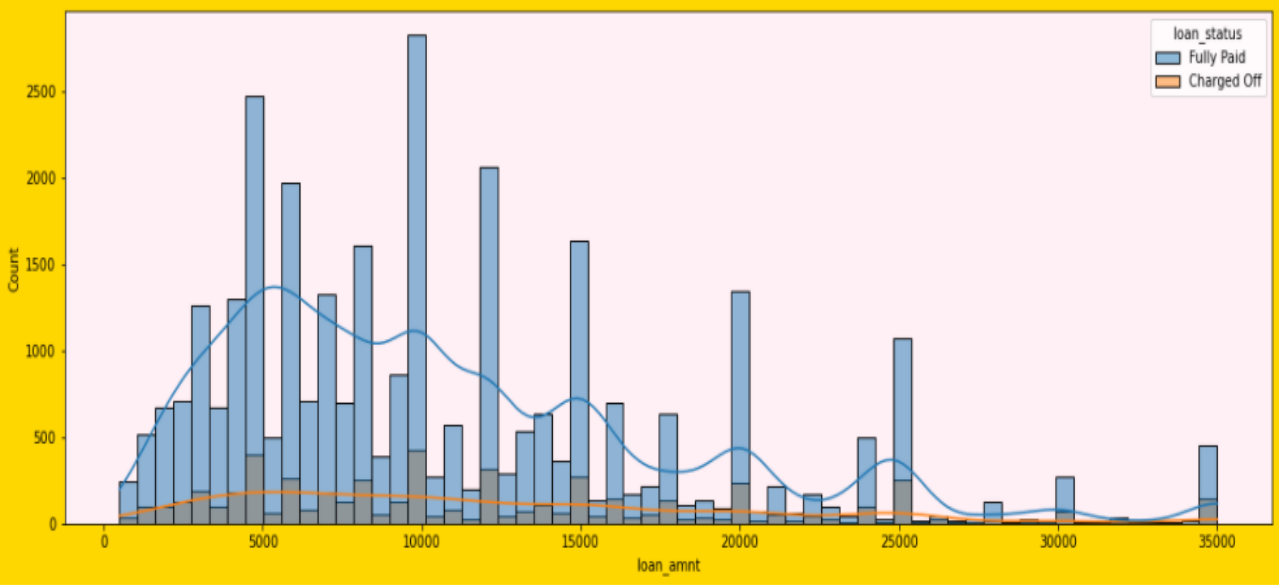
Data Analysis – Univariate and Segmented Univariate

Observation :

- Approximately **14%** of loans in the dataset are **defaulted**.

Data Analysis – Univariate and Segmented Univariate

Distribution of Loan Amount Disbursed



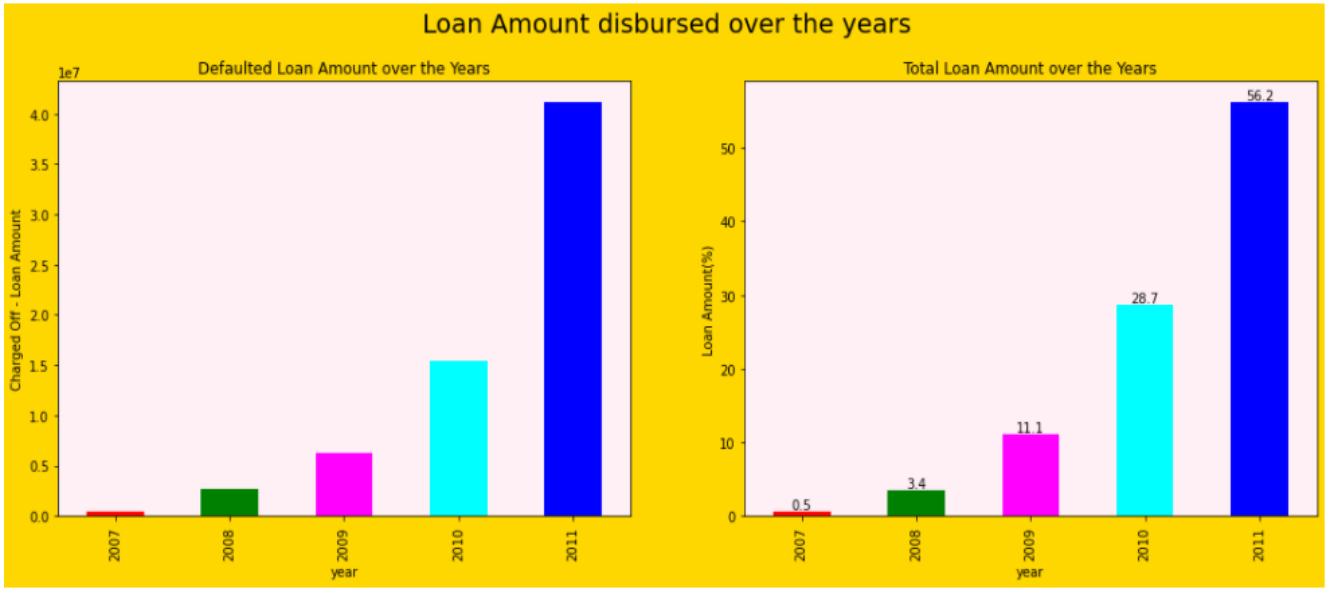
Data Analysis – Univariate and Segmented Univariate

Observation:

- Maximum number of loan disbursed is from 5000 to 10000. There are some distribution amounting to 15000, 20000, & 25000 relatively in high numbers.
- More than 25000, it is very minimal in numbers.

Data Analysis – Univariate and Segmented Univariate

Distribution of Loan Amount Disbursed over the years



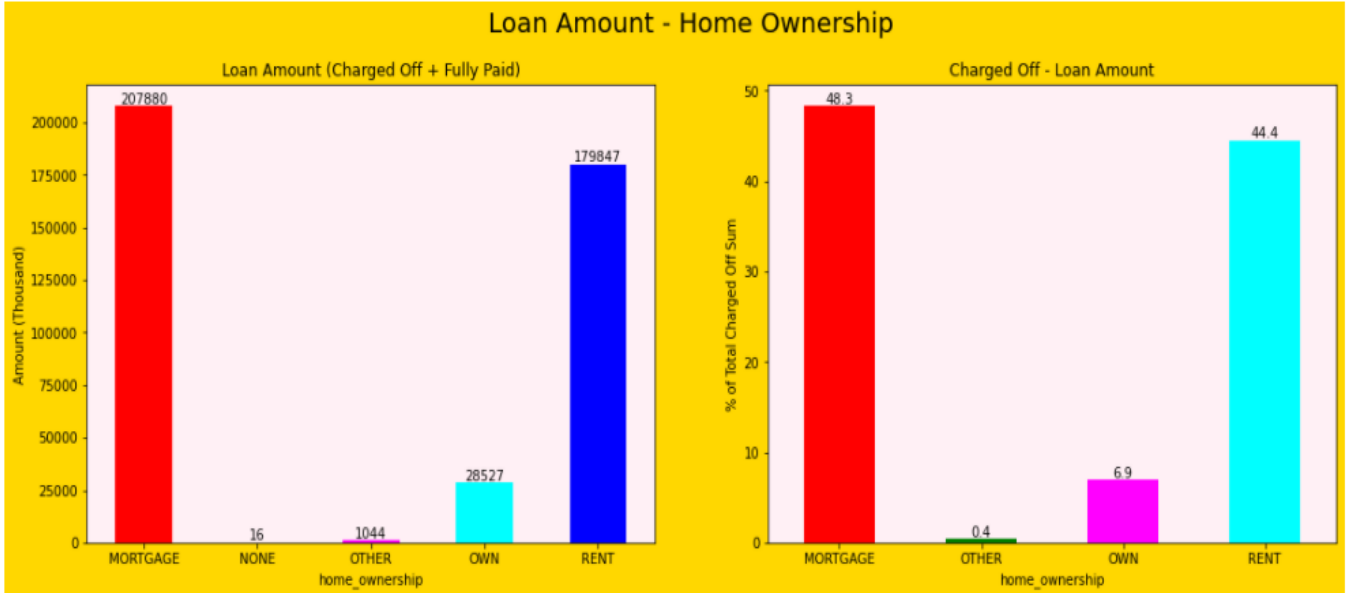
Data Analysis – Univariate and Segmented Univariate

Observation:

- Maximum number of loan disbursed in the year 2011 alone which is more than 50%.
- More the loan amount, more is the case of default

Data Analysis – Univariate and Segmented Univariate

Categorical Plots : Home Ownership Vs Loan Amount



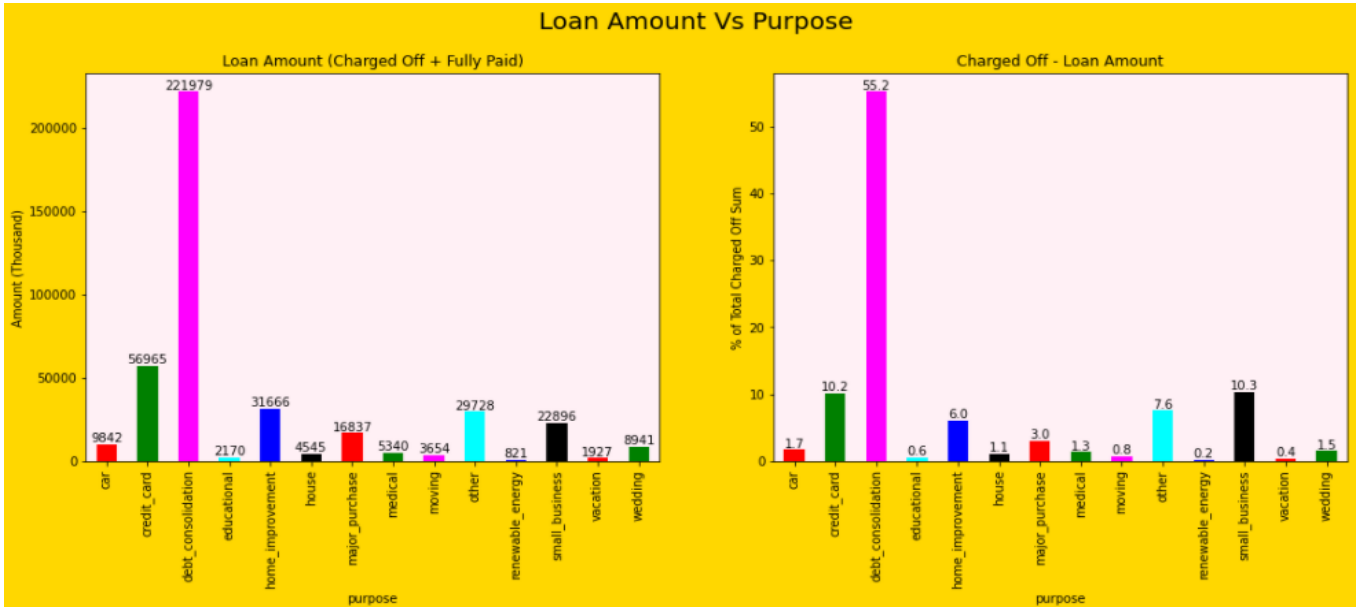
Data Analysis – Univariate and Segmented Univariate

Observation :

- Predominantly, Category ‘Mortgage’ and ‘Rent’ had taken the loan, which thus reflects on the defaulters as same.

Data Analysis – Univariate and Segmented Univariate

Categorical Plots : Purpose Vs Loan Amount



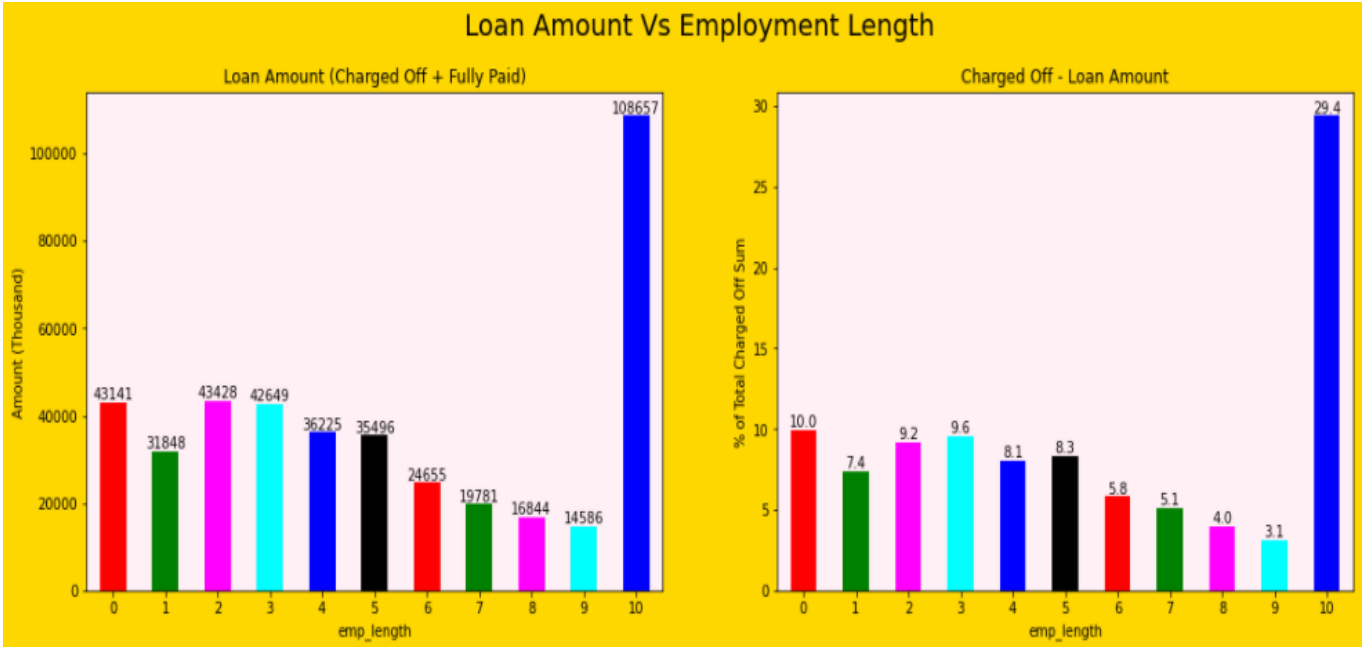
Data Analysis – Univariate and Segmented Univariate

Observation :

- Debt consolidation is the major factor for the loan application which is followed by credit card and home improvement.
- Defaulters are showing the same term except most of small business case not able to repay the loan. Defaulting form Debt consolidation category is upto 55%.

Data Analysis – Univariate and Segmented Univariate

Categorical Plots : Employment Length Vs Loan Amount



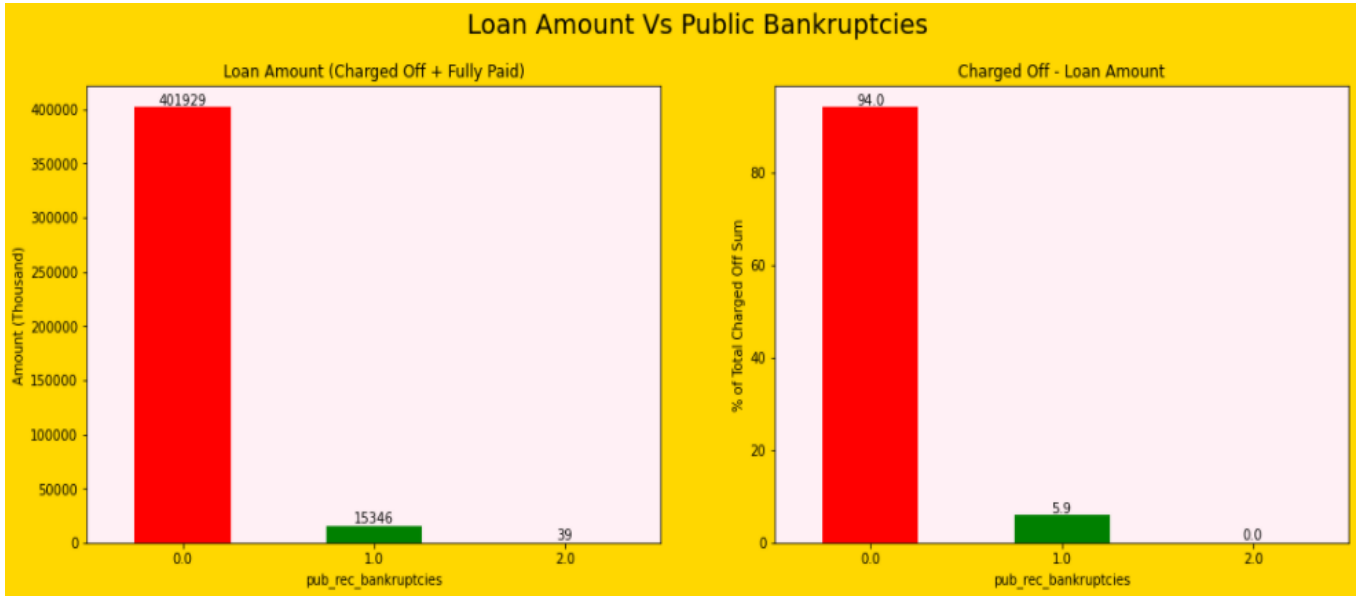
Data Analysis – Univariate and Segmented Univariate

Observation :

- Individual either in the early phase of job or 10th years onward takes the loan and by the rate they do get defaulted.
- 10+ years employees contribute for loan defaulting up to 30%.

Data Analysis – Univariate and Segmented Univariate

Categorical Plots : Public Bankruptcies Vs Loan Amount



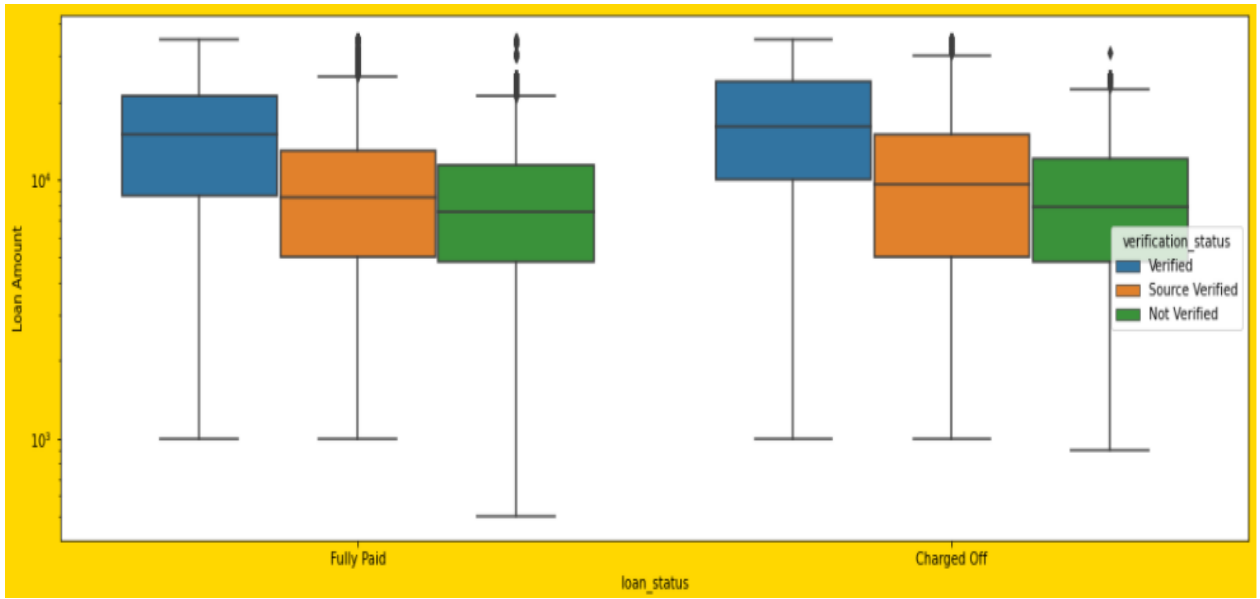
Data Analysis – Univariate and Segmented Univariate

Observation :

- 94% have no Public derogatory records. 6% have 1 derogatory public record.
- Having even 1 derogatory record increases the chances of Charge Off significantly.

Data Analysis – Univariate and Segmented Univariate

Categorical Plots : Verification Status Vs Loan Amount



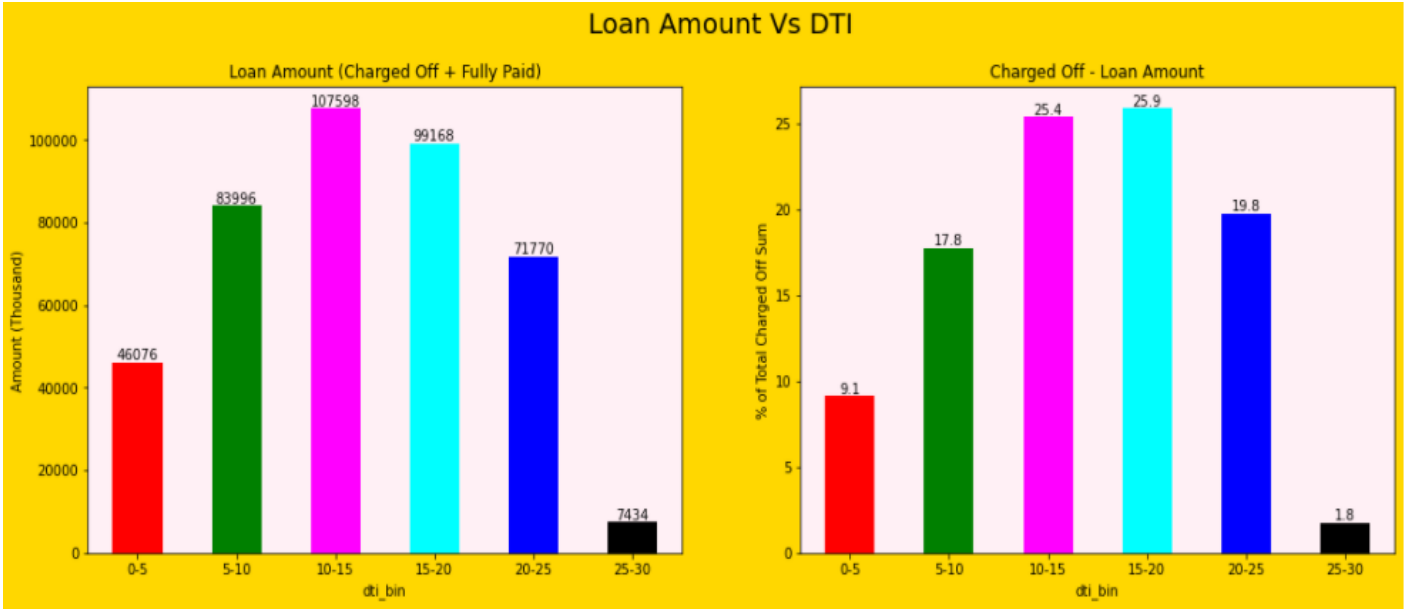
Data Analysis – Univariate and Segmented Univariate

Observation :

- Higher Loan amounts are verified more often but it does not exclude for defaulting to pay even though it is less in total counts.

Data Analysis – Univariate and Segmented Univariate

Categorical Plots : Debt To Income Ratio Vs Loan Amount



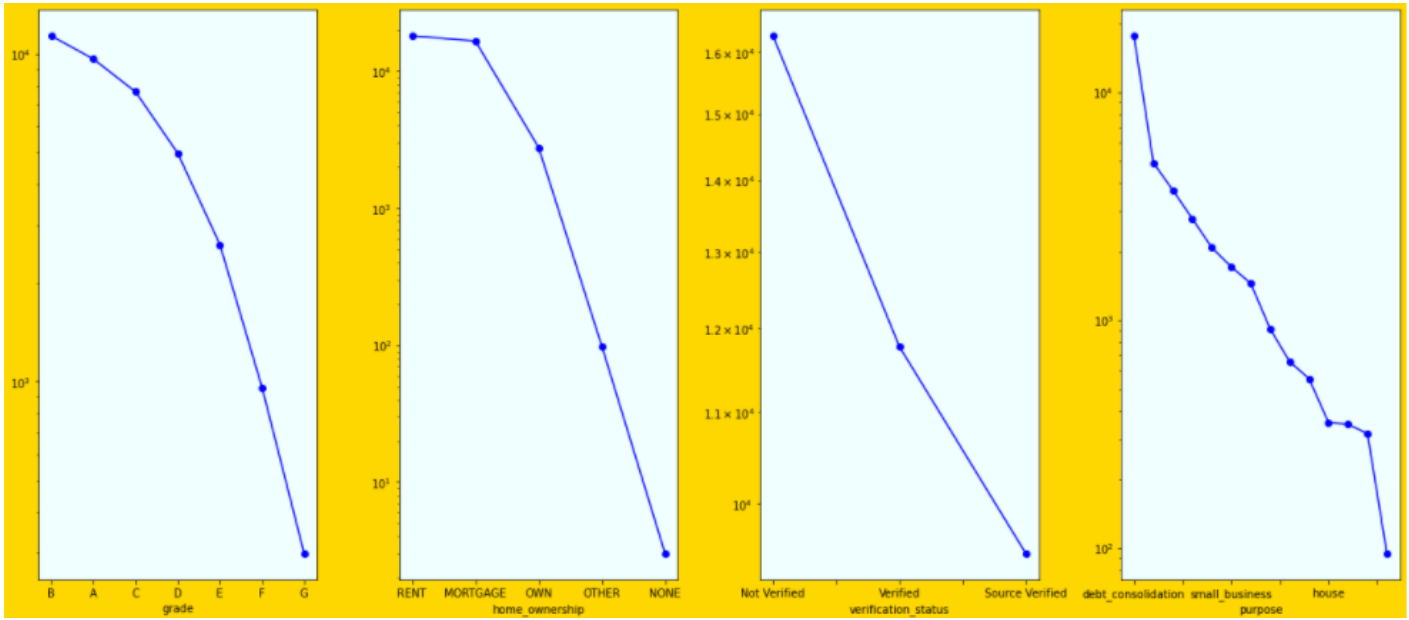
Data Analysis – Univariate and Segmented Univariate

Observation :

- When the debt payment to income ratio is higher than 15, higher percentage of loans are Charged Off.
- Higher the dti higher the chances of loan being Charged Off.

Data Analysis – Univariate and Segmented Univariate

Power Law Distribution : Categorical Variable – Grade, Home Ownership, Verification Status, Purpose



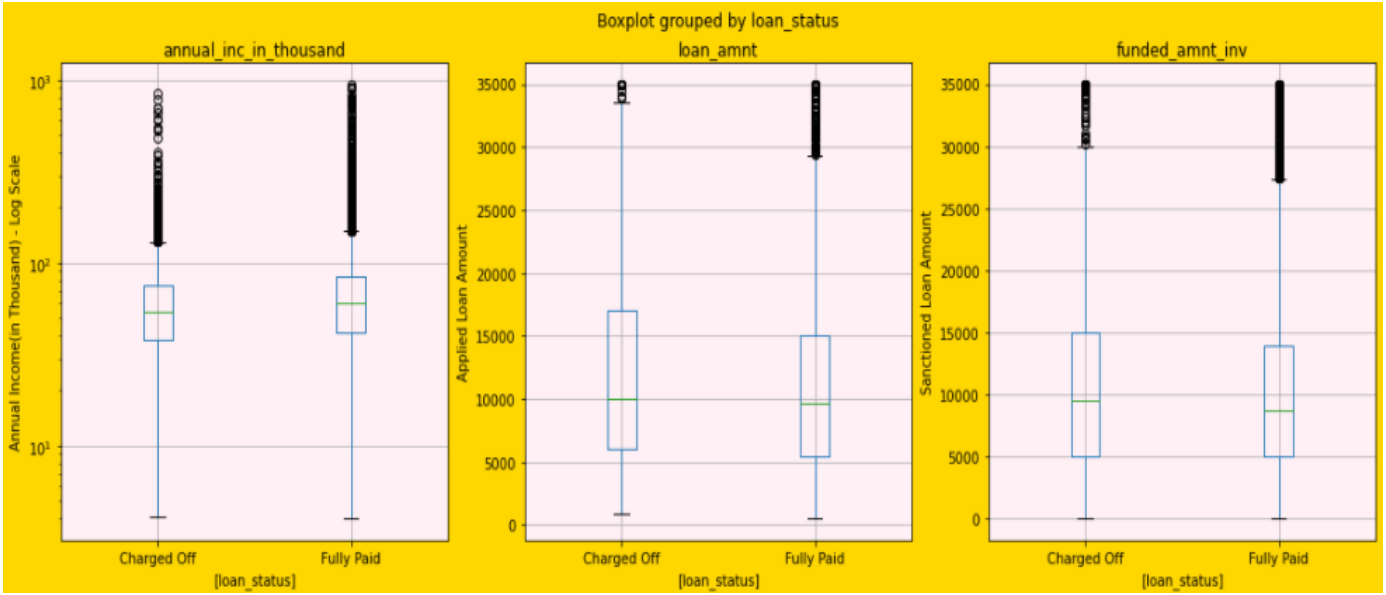
Data Analysis – Univariate and Segmented Univariate

Observation :

- A relative change in the categorical quantity purpose, verification status & home ownership can proportionally change in the loan amount.
- Purpose, verification status and somewhat home ownership follows the power law distribution.

Data Analysis – Univariate and Segmented Univariate

Box Plots of Annual Income, Loan Amount, Sanctioned Loan Vs Loan Status



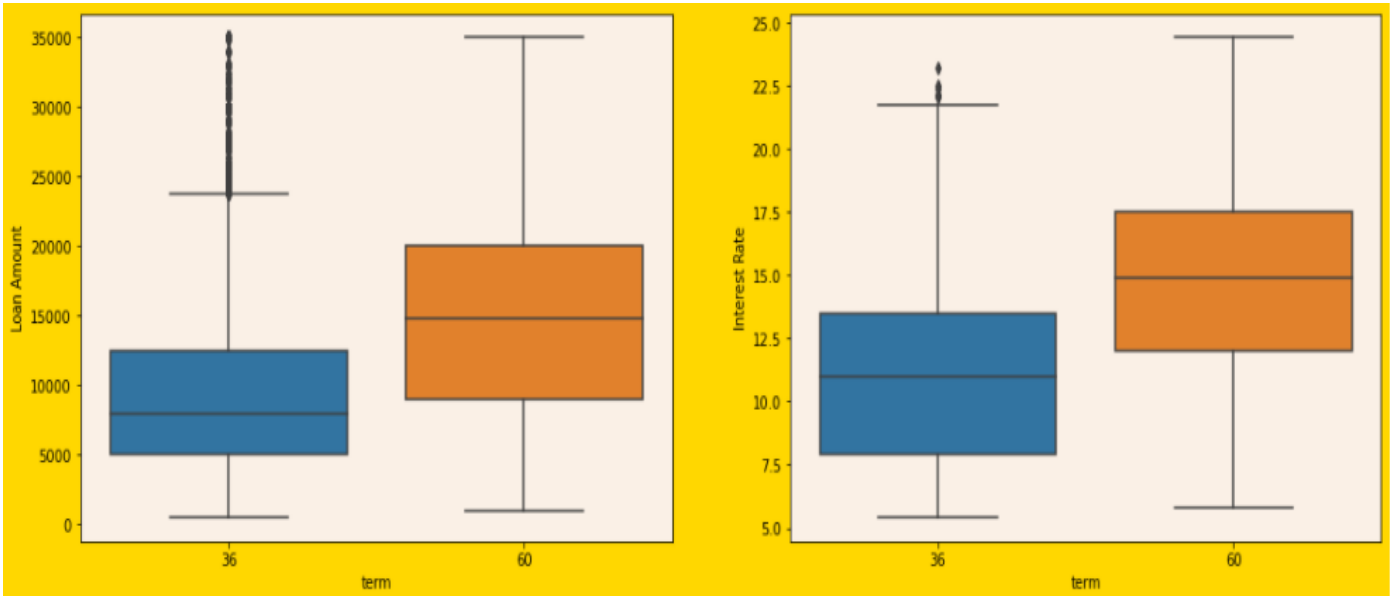
Data Analysis – Univariate and Segmented Univariate

Observation :

- **Defaulted** and **Fully Paid Loan** are independent of their annual income.
- Applied loan amount & sanctioned loan amount with respect to **defaulted** category is somewhat higher than **Fully Paid** category.

Data Analysis – Univariate and Segmented Univariate

Interest Rate Vs Loan Amount Vs Term



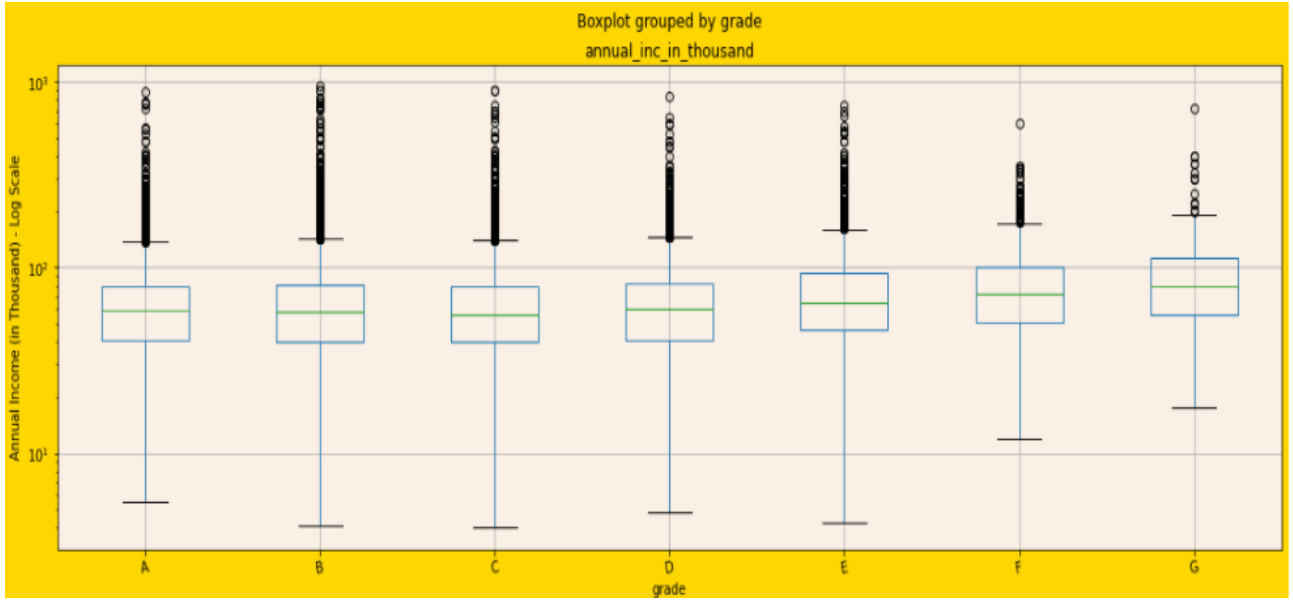
Data Analysis – Univariate and Segmented Univariate

Observation :

- Bigger the Loan amounts, higher the term and higher the interest rate.

Data Analysis – Univariate and Segmented Univariate

Box Plot : Annual Income for Different Grade



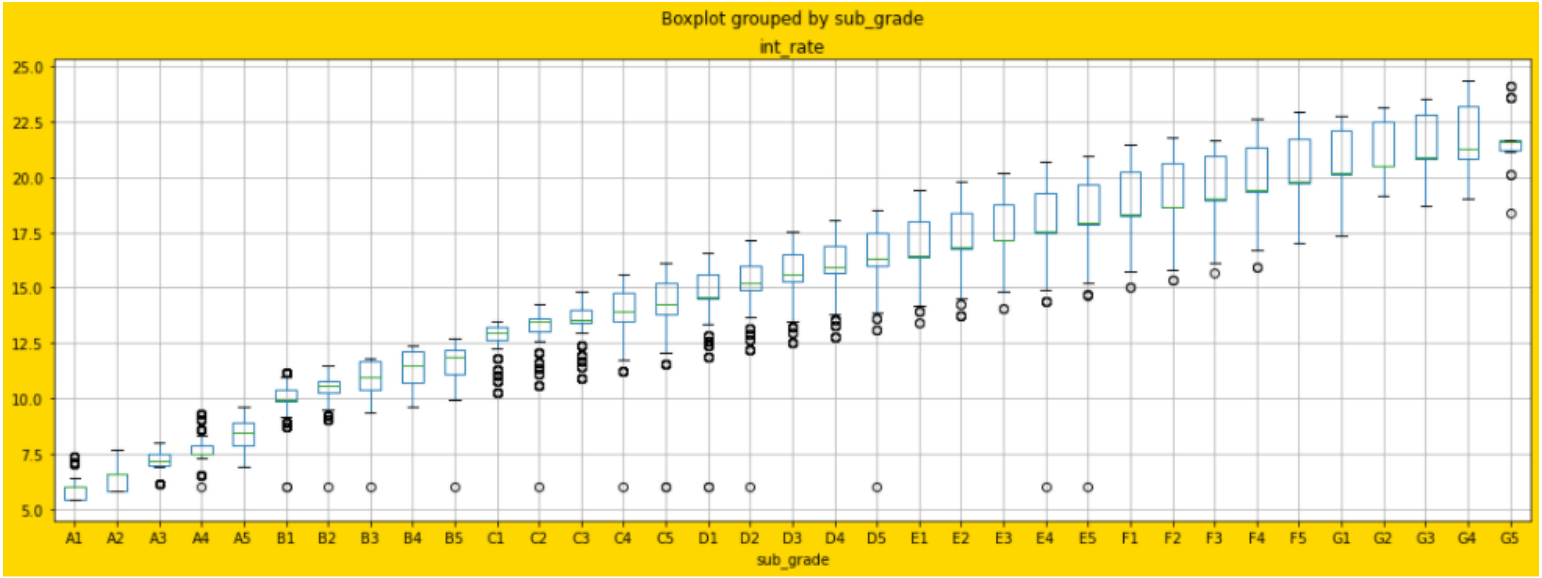
Data Analysis – Univariate and Segmented Univariate

Observation :

- Grade F & G are relatively better off position for lowest most income. In general, bracket from 25% to 75% percentile of annual income of all grades are more or less same, not much big difference.

Data Analysis – Univariate and Segmented Univariate

Box Plot : Interest Rate for Different Sub Grade



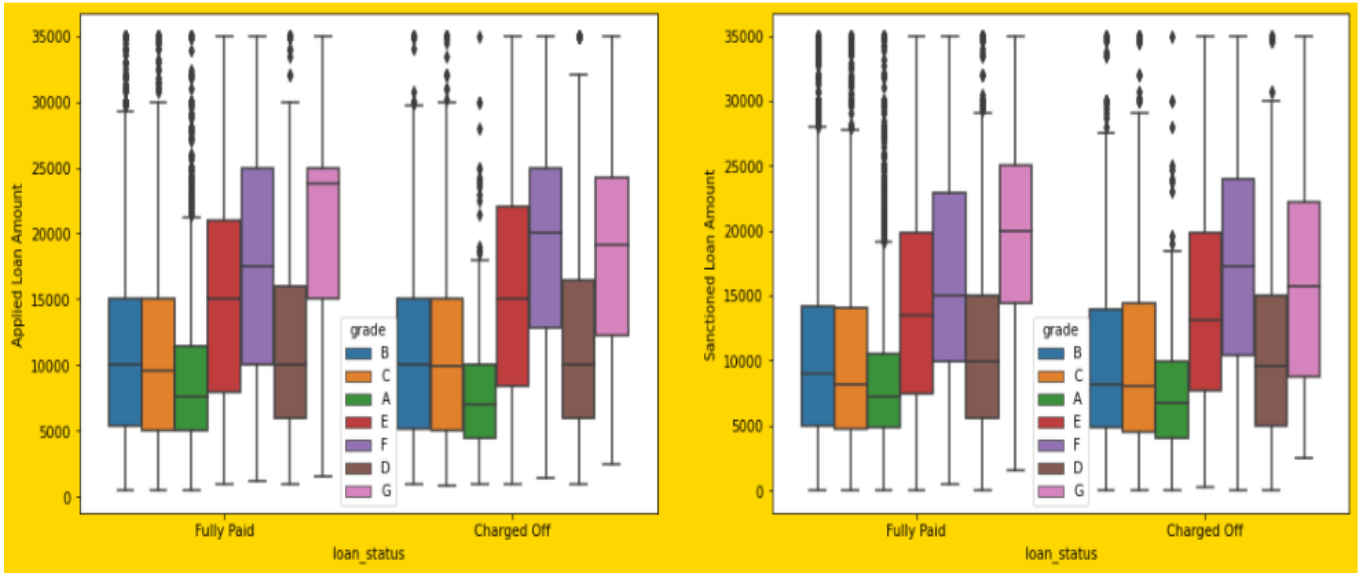
Data Analysis – Univariate and Segmented Univariate

Observation :

- Interest rates are directly proportional to the subgrade. Larger or lower the sub grade, higher or lower are the rate of interest for the loan.

Data Analysis – Bivariate

Box Plot : Applied Loan Amount and Sanctioned Loan Amount grouped by Grade.



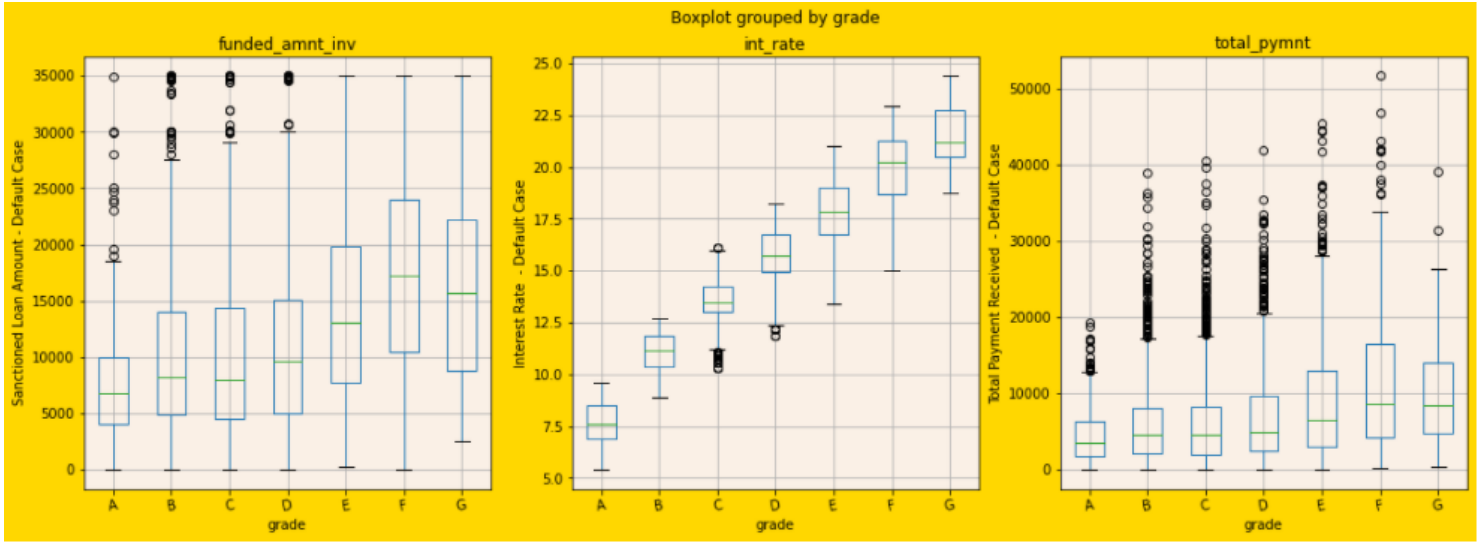
Data Analysis – Bivariate

Observation :

- Grade E, F & G have applied for higher loan amount and they have been sanctioned the higher loan amount.
- Large proportion to default is also being contributed by Grade E, F & G.

Data Analysis – Bivariate

Box Plot : Sanctioned Loan Amount, Interest Rate and Total Payment for each Grade and Sub Grade (Default/Charged Off Case)



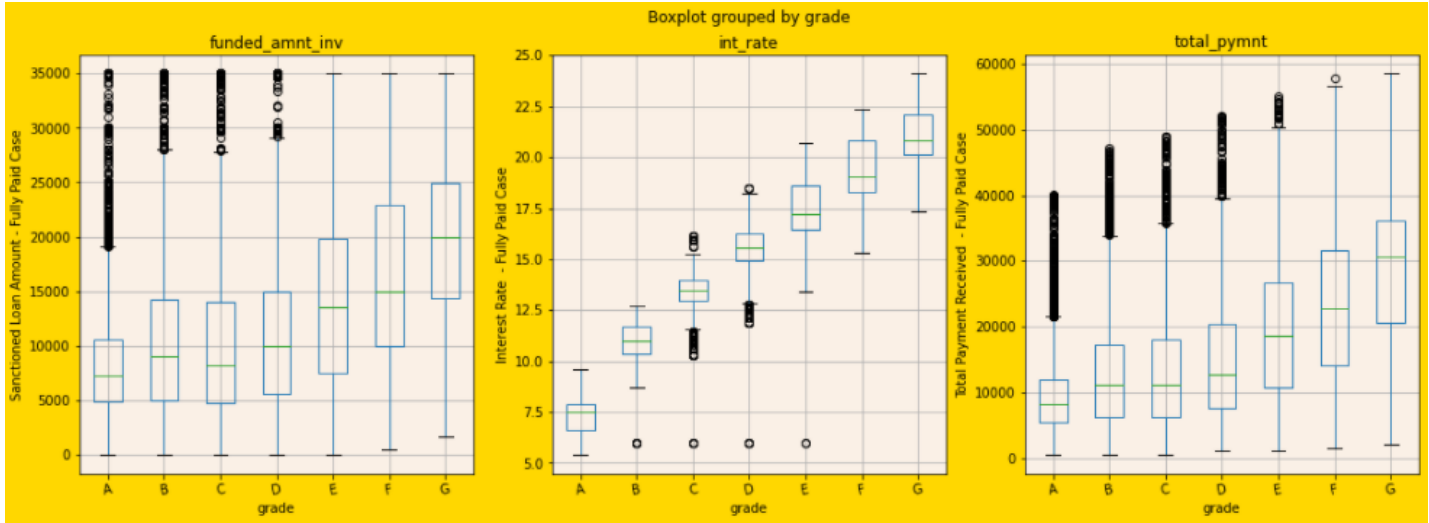
Data Analysis – Bivariate

Observation :

- Sanctioned loan amount is more general relatively increasing from Grade A to Grade G.
- Interest Rate from Grade A to Grade G is significantly different to each other and increased from Grade A to Grade G. Grade A is having 6% to 9% interest rate whereas Grade G is paying the high interest from 18% to 24% approx.
- Even though having the higher interest Rate of 15-22% paid by the Grade F, they are less riskier among all grades even in case of default.

Data Analysis – Bivariate

Box Plot : Sanctioned Loan Amount, Interest Rate and Total Payment for each Grade and Sub Grade (Fully Paid Case).



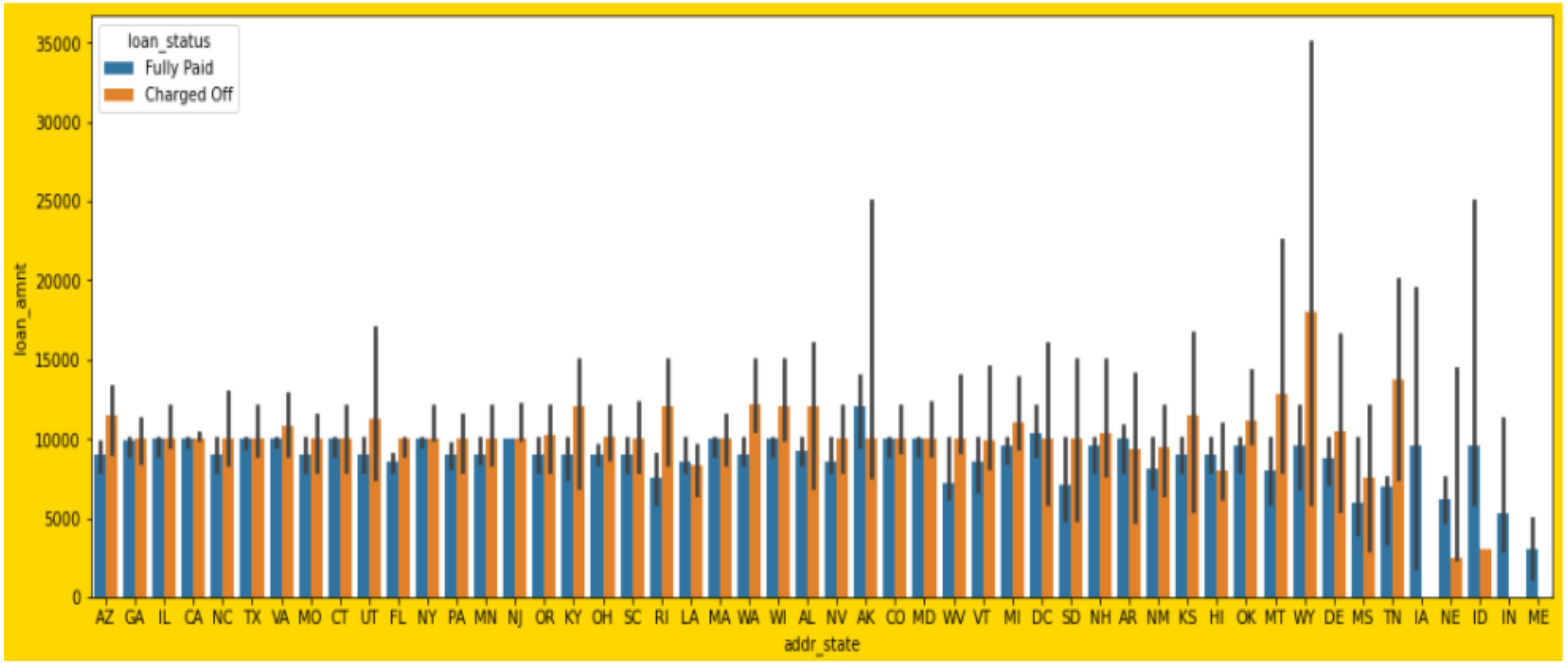
Data Analysis – Bivariate

Observation :

- Grade E, F & G have given the most profits to the lending company.

Data Analysis – Bivariate

Loan Amount Vs Address State grouped by Loan Status.



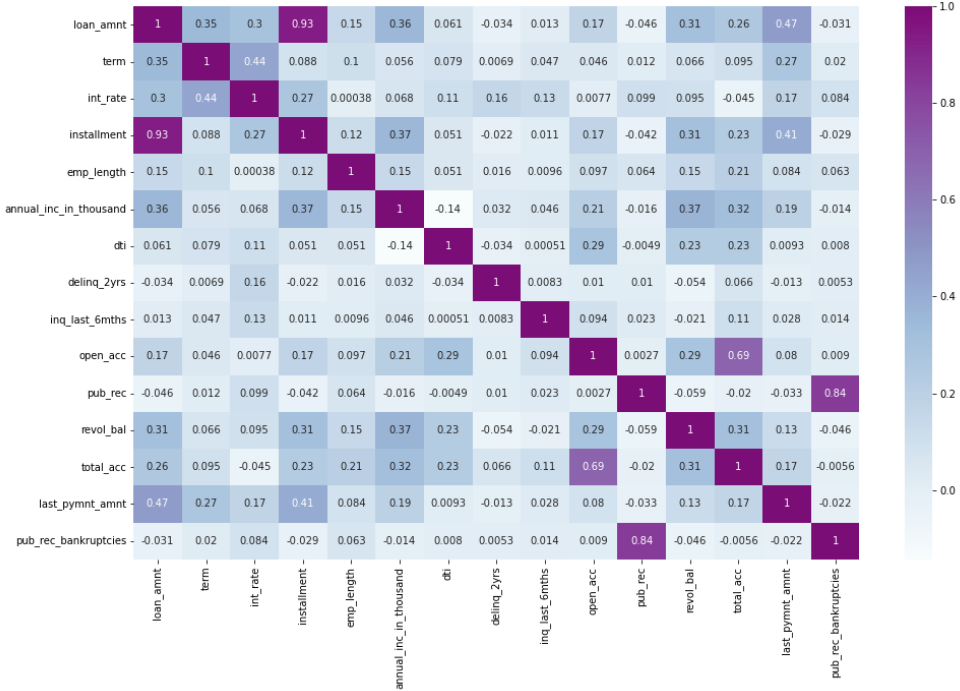
Data Analysis – Bivariate

Observation :

- Interesting to observe, States WY has the highest average(median) loan amount that was charged off. State must be looked into by the LC for further investigation.

Data Analysis – Bivariate

The Heat Map of various Variables.



Data Analysis – Bivariate

Observation :

- Loan amount is highly correlated positively with number of instalment & last payment by 93% and 47% respectively.
- Loan amount is correlated positively with the revolving balance amount by 31%, this means riskier loans are getting approved.
- Revolving balance correlated positively with Annual income & Instalment by 37 % to 31%.
- Employment length not greatly correlatable to loan amount, instalment or interest rate.

Data Analysis – Bivariate

Observation :

- Public Bankruptcies is negatively correlated with loan amount; this is a good sign.
- Low value correlation suggest that of many variables suggest, they are not correlatable and hence will produce no trend to each other.

Conclusion

- Approximately **14%** of loans in the provided dataset are **defaulted**.
- Maximum number of loan disbursed is from 5000 to 10000.
  Maximum number of loan disbursed in the year 2011 alone which is more than 50%.
- Predominantly, most of the loan applicant either have **mortgage** something or living in a **rented** place, and the same category reflects for the defaulters, combining both of them are **93% of default case**.

Conclusion

- Giving loan for **debt consolidation purpose** is a **riskier business**, more than half of them (out of total default case) can default.
- **10+ years employees** contribute for loan defaulting upto **30%**, i.e., one third of total default cases.
- **Avoid giving loan** to the applicant who has even **one public derogatory record**.
- When **dti**, dept payment to income ratio is **higher than 15**, they are **likely to default**.

Conclusion

- Other than very income group, annual income does not affect the loan defaulting.
- **Grade employee (E, F & G)** have given the **most profits** to the lending company with **high interest rate** suitably with **15-22%** rate.
- **States WY** has the highest average(median) loan amount that was **charged off**. This state must be looked into by the LC for further investigation.

Thank You !

9th March 2022

- Prashant Mohan Sinha