Capstone Project 1- Predicting Loan Default

Exploratory Data Analysis

Contents

[The Lending Club Data Set 2](#_Toc3654903)

[Libraries Used 2](#_Toc3654904)

[Problem Statement 2](#_Toc3654905)

[Exploratory Data Analysis 2](#_Toc3654906)

[a. The Count plot of Loan Grades and default loans and no default loans 3](#_Toc3654907)

[b. The Distribution plot of Interest Rate vs FICO Score 3](#_Toc3654908)

[c. The Distribution plot of Loan amount vs Interest rate 4](#_Toc3654909)

[d. FICO Score Statistics for default and No Default loans 4](#_Toc3654910)

[e. The Regplot for FICO Score and Loan Default 5](#_Toc3654911)

[f. The Whisker plot for Annual Income Statistics 5](#_Toc3654912)

[g. The Whisker plot for Interest Rate Statistics 6](#_Toc3654913)

[h. The Distribution graph for debt to interest ratio 6](#_Toc3654914)

[i. The Count plot of verification status 7](#_Toc3654915)

[j. The Count plot of loan terms for defaulted loans 7](#_Toc3654916)

[k. The Count plot of home ownership 8](#_Toc3654917)

[l. The Distribution plot and Boxplot for loan amount 9](#_Toc3654918)

[m. Pearson’s Correlation Coefficient 9](#_Toc3654919)

[Conclusion 9](#_Toc3654920)

# The Lending Club Data Set

The Dataset used in this project is The Lending Club dataset available from kaggel website. (https://www.kaggle.com/wordsforthewise/lending-club/home). It is a real world data set which contains 2004126 rows of loan listing and 150 columns (attributes) of the each loan listing from year 2007 to 2018 Q2. Out of 150 columns, some columns have missing data and some columns are not needed for the analysis. After applying several data wrangling method, data set is cleaned and saved in clean\_loan.csv. Visual EDA is done on data loaded from clean\_loan.csv.

# Libraries Used

Besides Python standard libraries, Pandas, Matplotlib.pyplot, Numpy and Seaborn libraries are used.

# Problem Statement

At The Lending Club, Investors can access each of the loan applications and accept or reject it based on the credit history, FICO score, income, current job status and many other attributes mentioned in the application. The purpose of this exercise is to find a general trend of how different features (columns) in the dataset are related to each other. By plotting different graphs, we try to find some insights and figure out which properties are related to loan status.

# Exploratory Data Analysis

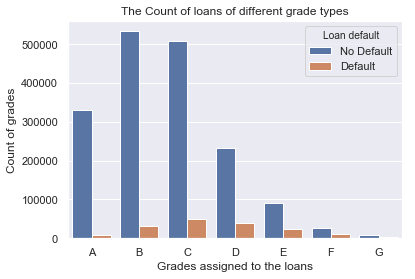
Exploratory data analysis was first defined by John Tukey in 1961 as *"Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."*

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. We can say that EDA is statisticians’ way of *storytelling* where you explore data, find patterns and tells insights. Often you have some questions in hand you try to validate those questions by performing EDA.

Let’s analyze different properties.

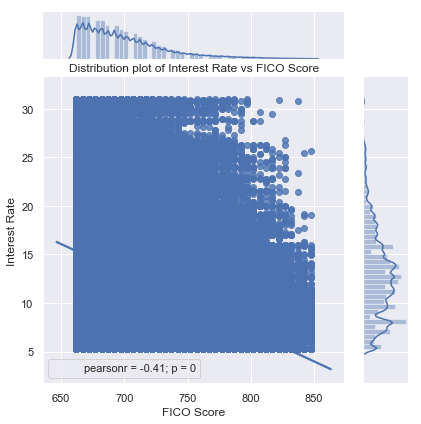
## The Count plot of Loan Grades and default loans and no default loans

###### *The Count plot below shows that Grade A type loans have lower default count. Loans with Grades B, C, and D have higher default count. This could be because there are high number of loans graded as B, C, and D.*



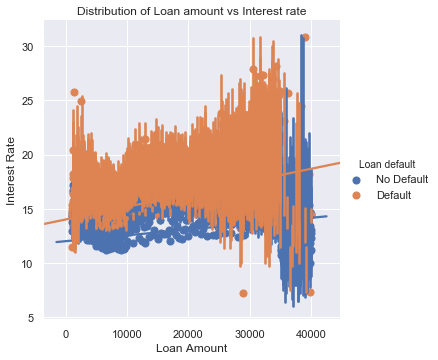
## The Distribution plot of Interest Rate vs FICO Score

*The Distribution plot below shows that borrowers with lower FICO Score got high interest rate. The plot shows the distribution of FICO score on marginal x axis and distribution of Interest rate on marginal y axis. Pearson coefficient is -ve which means increasing FICO score corresponds to lower interest rate.*

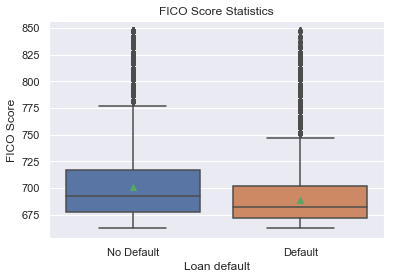


## The Distribution plot of Loan amount vs Interest rate

*The Distribution plot below shows that the same range of loan amounts have higher interest rate for loans which are defaulted. To reduce the number of points, x\_estimator is used which plots the average y value of all the points with same x value and draws a line to indicate the range of values****.***



## FICO Score Statistics for default and No Default loans

*The Whisker plots for FICO Score for default and no default loans show that mean and median value of FICO score is less for loans which are defaulted. That means borrowers with less FICO score have higher chances defaulting the loan.*

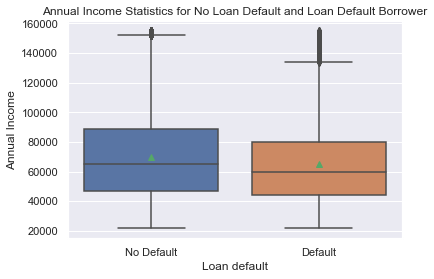
## The Regplot for FICO Score and Loan Default

*The regression line for FICO Score vs Loan Default is linear which means the FICO Score and default rate are inversely proportional to each other.*



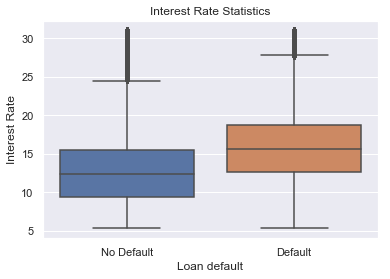
## The Whisker plot for Annual Income Statistics

*The plot shows that the mean and median of annual income of borrowers with defaulted loan are lower than the mean and median of the annual income of borrowers with no default. But this difference is not significantly big that we can draw a clean conclusion based on it.*



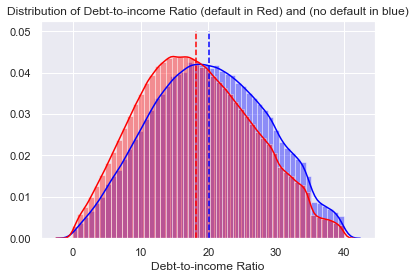
## The Whisker plot for Interest Rate Statistics

*The plot shows that minimum and maximum values of interest rates are similar but the mean and median of interest rates of loans with default are significantly higher than the mean and median of the interest rates of loans with no default.*



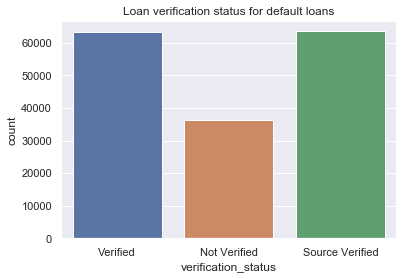
## The Distribution graph for debt to interest ratio

*The Distribution plot below shows the distribution is normal for both default and no default borrowers. The default borrowers tend to have higher debt to income ratio than the no default borrowers.*



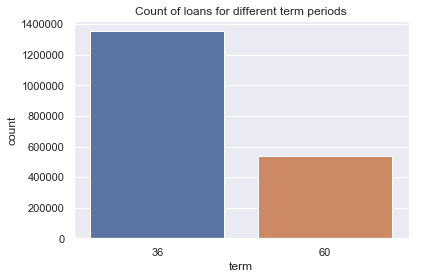
## The Count plot of verification status

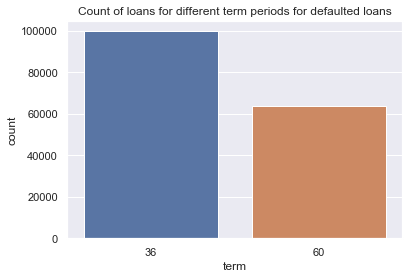
*Verification status indicates if income was verified by [Lending Club], not verified, or if the income source was verified. The plot below shows that there are more chances of loan default when borrowers' income is not verified.*



## The Count plot of loan terms for defaulted loans

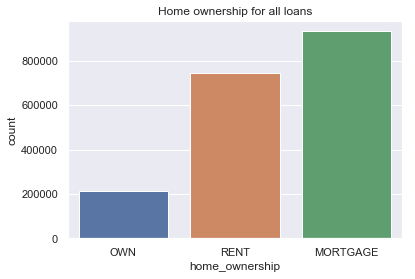
*The comparison shows that the loans with 60 months term defaulted more than 36 months loans.*

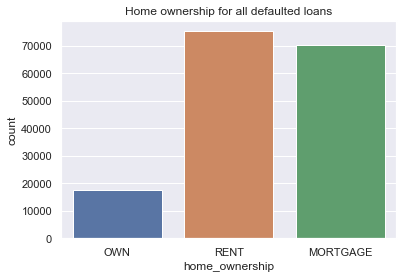
**



## The Count plot of home ownership

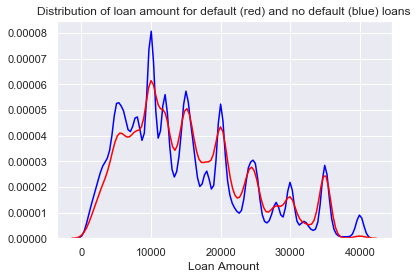
*The plots show that borrowers with default loans tend to RENT the house.*





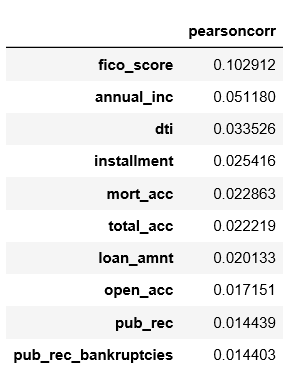
## The Distribution plot and Boxplot for loan amount

*The distribution plot of loan amount shows that loans with higher loan amount have higher chances of default.*

**

## Pearson’s Correlation Coefficient

###### *Pearson's correlation coefficient for each predictor to find the dependency on response (default):*

**

# Conclusion

*Above analysis shows that FICO score, interest rate, loan amount, annual income, debt to income ratio have high impact on the loan defaults. The Pearson Correlation Coefficient also shows that FICO score, annual income and debt to income ratio are 3 bigger impact on loan defaults.*