Title: Loan Approval Prediction

Bhargava Gaggainpali  
DSC680 - Applied Data Science  
Project 1 - Milestone 2

[bgaggainpali/bgaggainpali\_DSC680 (github.com)](https://github.com/bgaggainpali/bgaggainpali_DSC680)

**Any surprises from your domain from these data?**

Loan Approval Process is part of Financial industry. I have selected the topic, as I was interested in knowing the process of loan approval and the key factors in it. As I explore more about the domain, I understand that its not same set of rules which is being followed across domain. And each subdomain like Home loan approval, Car loan approval and such classification have basic approval structure, but the factors which influence are different. Bank should not lose business by denying a legitimate customer, who can repay. Also, it should not approve loan to in-eligible customer. Banks are playing important role in challenging times like now, with COVID pandemic across the globe.

**The dataset is what you thought it was?**

Initially when I looked at the dataset, I had questions about the variables as they are more of general kind and was guessing that if would serve my purpose of analysis. when closely observed the stats, it surprised me as the amount of value we can retrieve from such data. I am satisfied with the dataset which I have selected.

<https://www.kaggle.com/premptk/loan-approval-prediction-model/data?select=LoanApprovalPrediction.csv>

**Variable Description**

Loan\_ID Unique Loan ID

Gender Male/ Female

Married Applicant married (Y/N)

Dependents Number of dependents

Education Applicant Education (Graduate/ Under Graduate)

Self\_Employed Self employed (Y/N)

ApplicantIncome Applicant income

CoapplicantIncome Coapplicant income

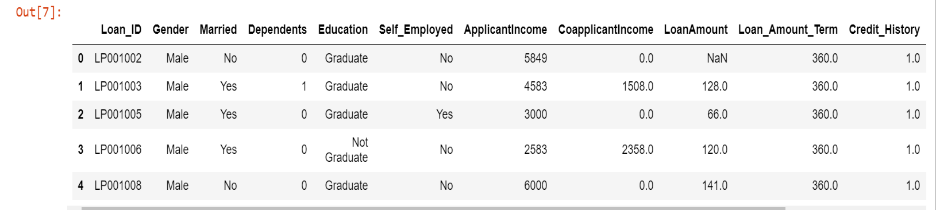
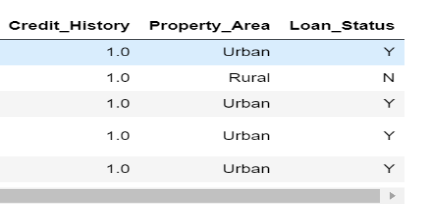
LoanAmount Loan amount in thousands

Loan\_Amount\_Term Term of loan in months

Credit\_History credit history meets guidelines

Property\_Area Urban/ Semi Urban/ Rural

Loan\_Status Loan approved (Y/N)

Categorical Features: Based on the data, (Yes/No or Male/Female) below are categorical variables.

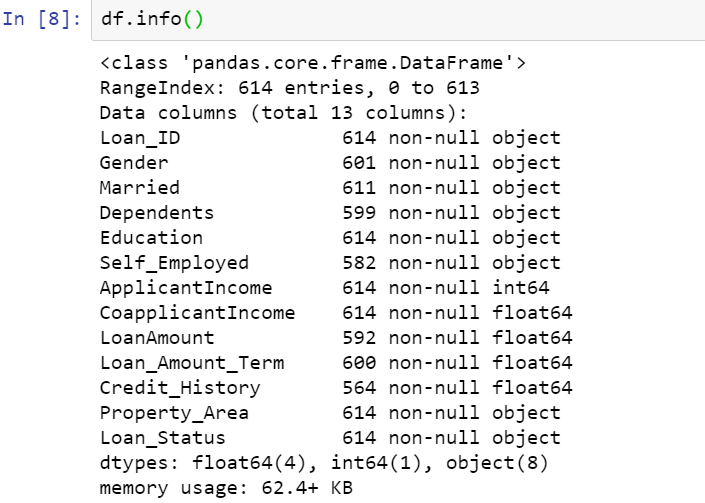
Gender, Married, Self\_Employed, Credit\_History, Loan\_Status.

Ordinal Features: Based on the data with inherent hierarchy, below are ordinal variables.

Dependents, Education and Property\_Area

Numerical Features: Based on the numerical data, below are numerical variables.

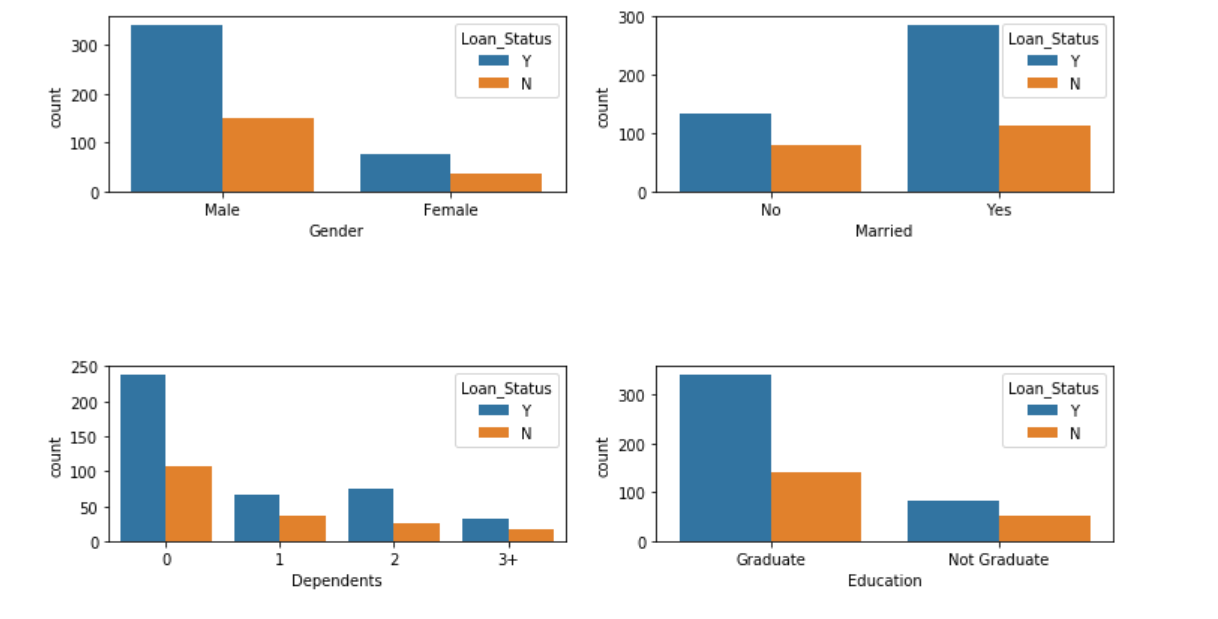
ApplicantIncome, Co-applicantIncome, LoanAmount, Loan\_Amount\_Term

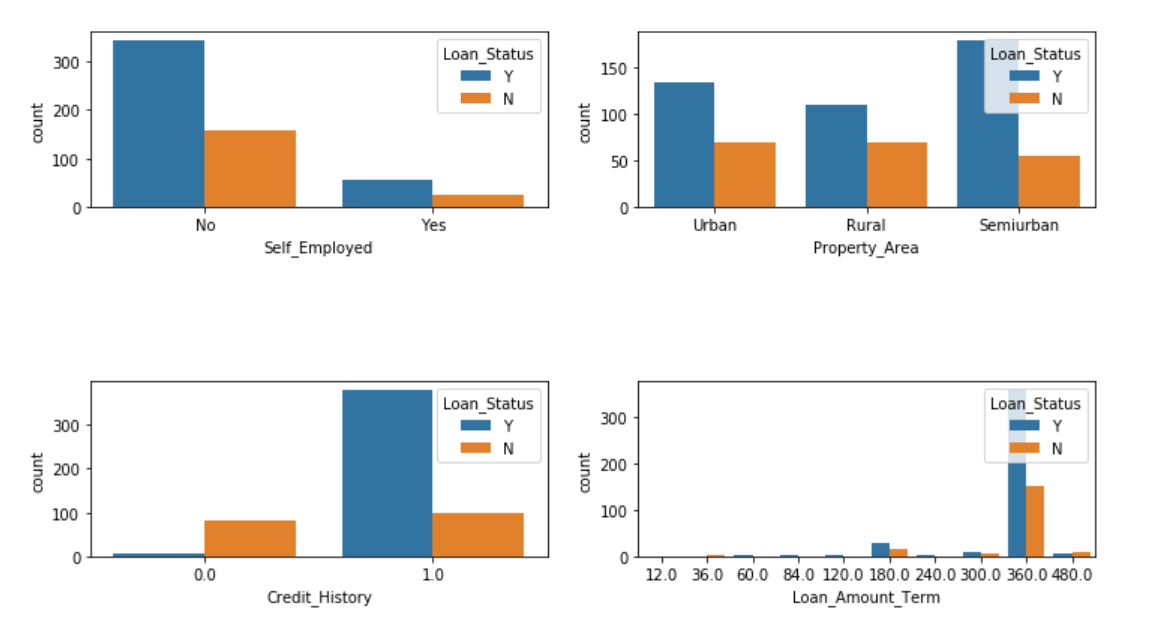


**Have you had to adjust your approach or research questions?**

After initial analysis of looking at the dataset values and the basic stats, I had to change my focus on considering many factors. Initially was under the impression that, Loan approval depends on education, applicant income and limited factors. I saw surprising stats when I used visualizations to give clear idea on how each factor has its effect on the Loan approval process.

Yes, I had to increase my research questions to explore and include more variables, than initially prepared. Its based on the initial analysis using visualization.





**Is your method working?**

Till now, I am comfortable and confident that my methods would work as initially planned. The steps which I am following are giving me good results as expected.

After completing the data acquisition and initial analysis, by running the stats and using visualization, I had to increase in the number of variables to explore and analyze and also consider for model building. If the number of variables are significantly more and if need, I am planning to apply dimensionality reduction methods like PCA to reduce the number of variables for model building.

**What challenges are you having?**

So far, I have not faced any major challenges. The process and the steps I am following are giving me good results as expected. Small challenges I have faced is to convert all the data in to common format using Python functions and syntax, which I expect get better on practice. I am now focusing on building the model and might challenge on implementing it. Based on the accuracy of the results, will have to restructure the code if its not as per expected and also planning to build multiple models to choose a better one which suits the dataset values.