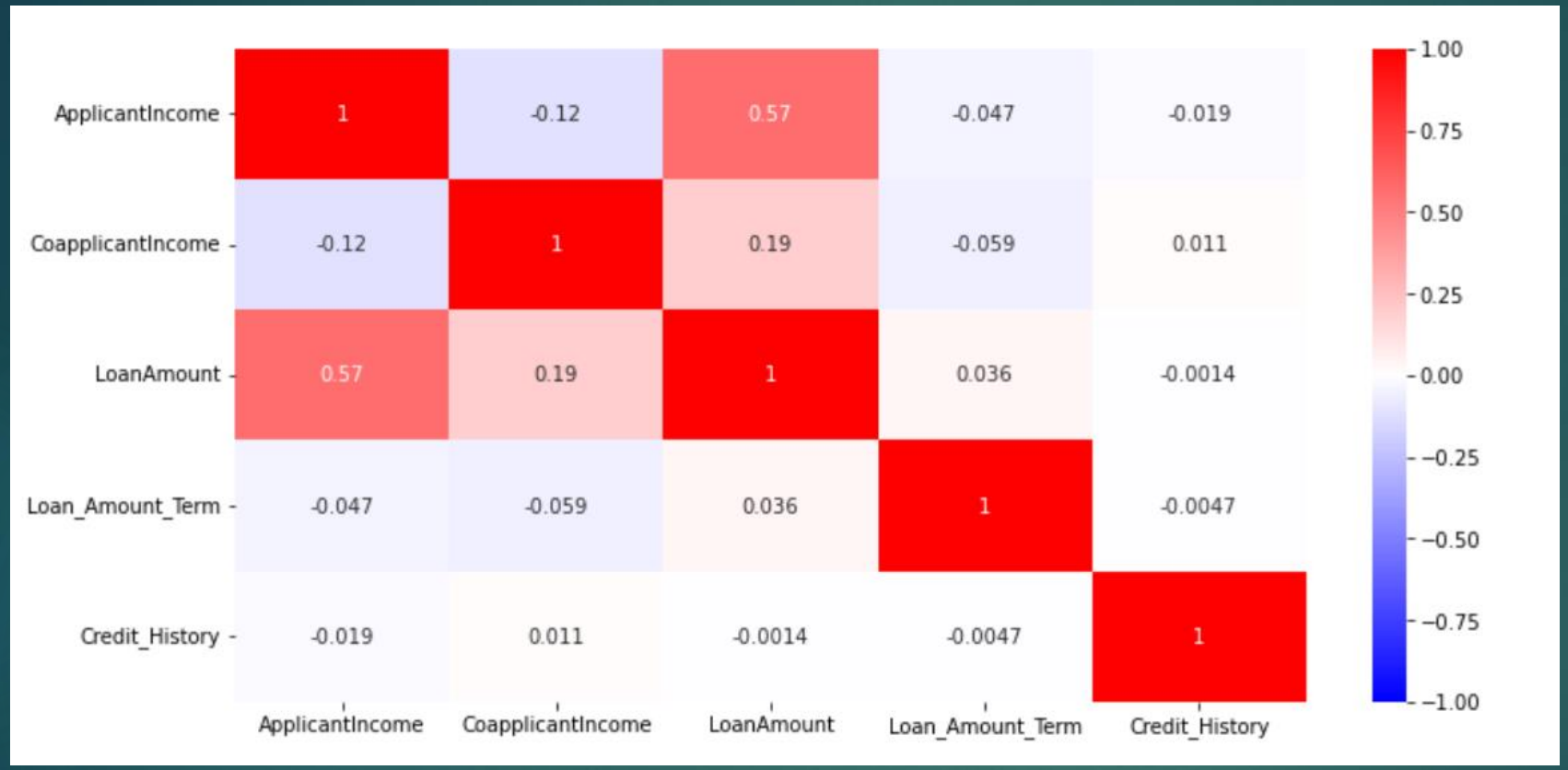
Capstone Project

RAM CHARAN SINGH S9135097A

Problem Statement

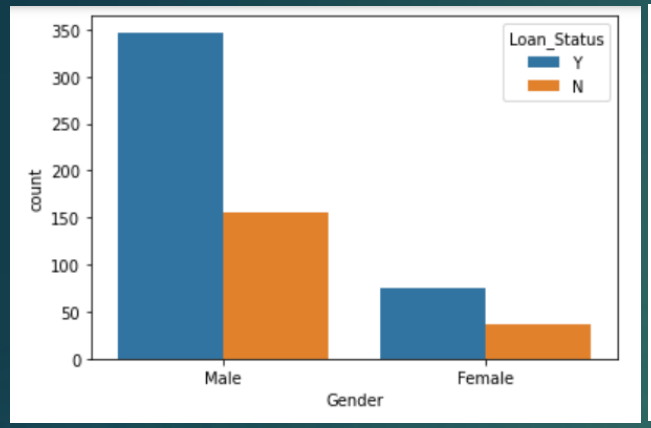
This project aims to predict which customers of the bank are likely to have their loan applications approved based on the details provided by the customer from their submitted online application.

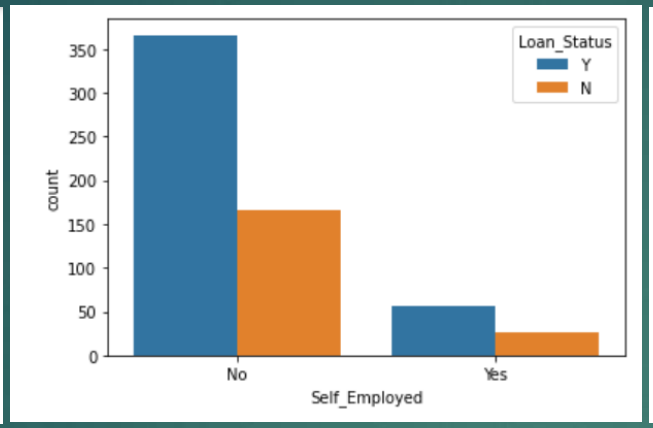
Data Exploration

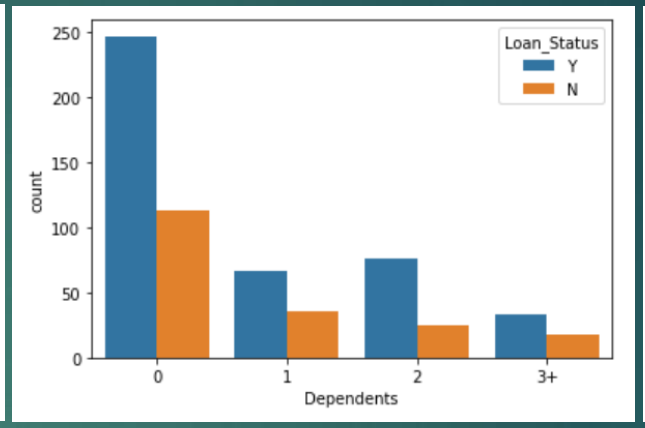


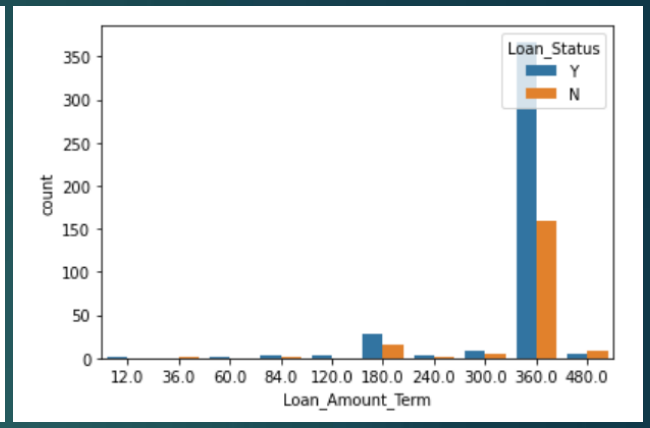
- There is little to no correlations for most variables except between LoanAmount and ApplicantIncome or CoapplicantIncome (0.57 & 0.19 respectively).

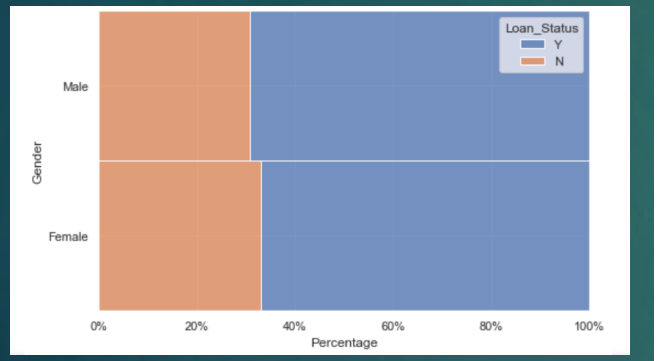
Data Exploration

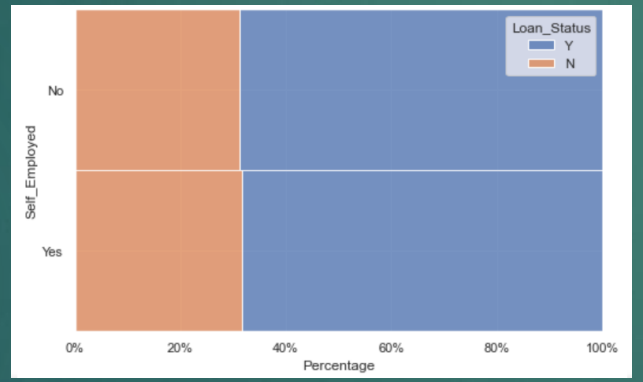


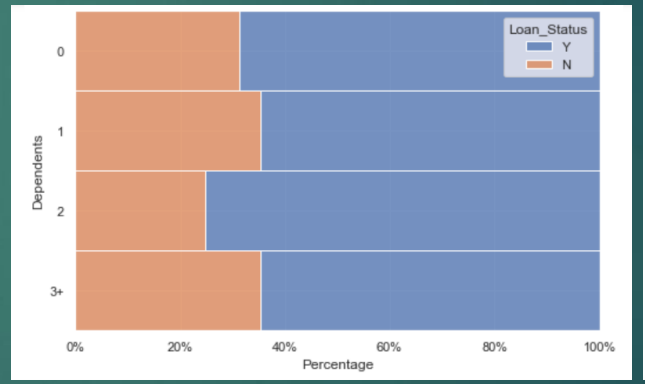


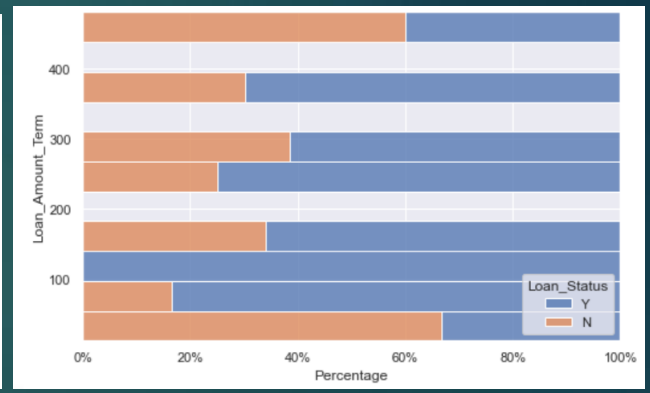






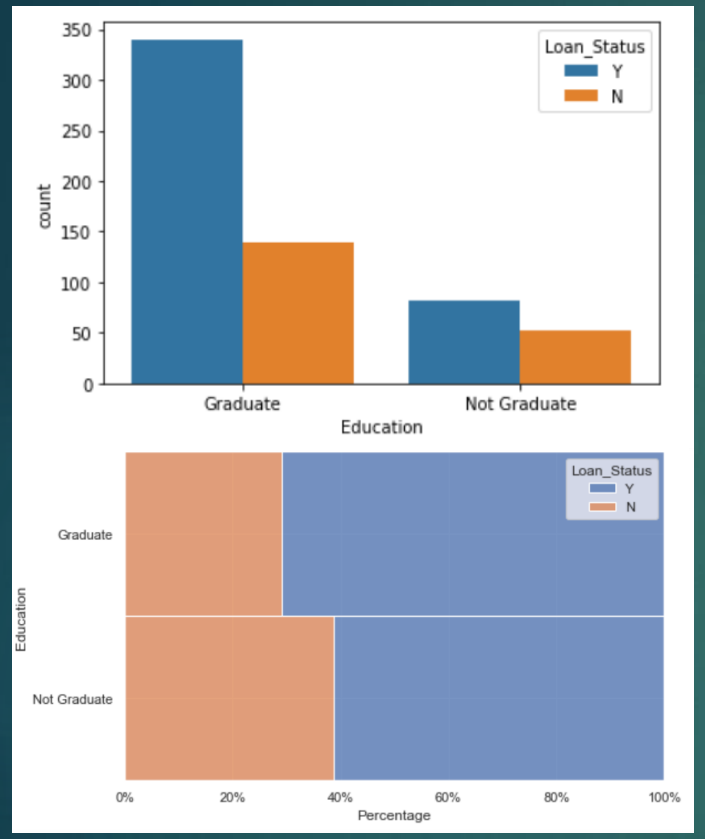


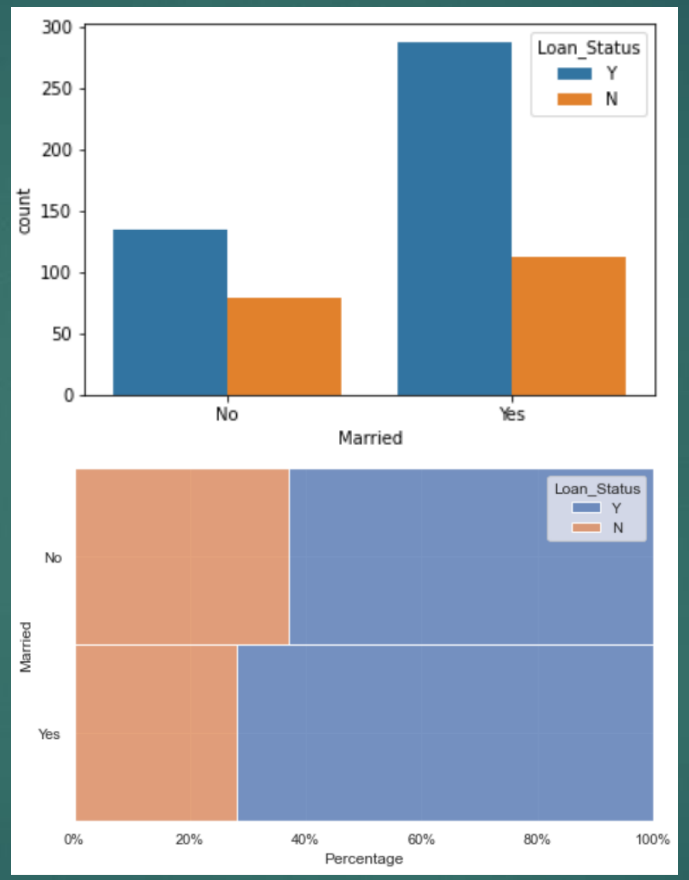


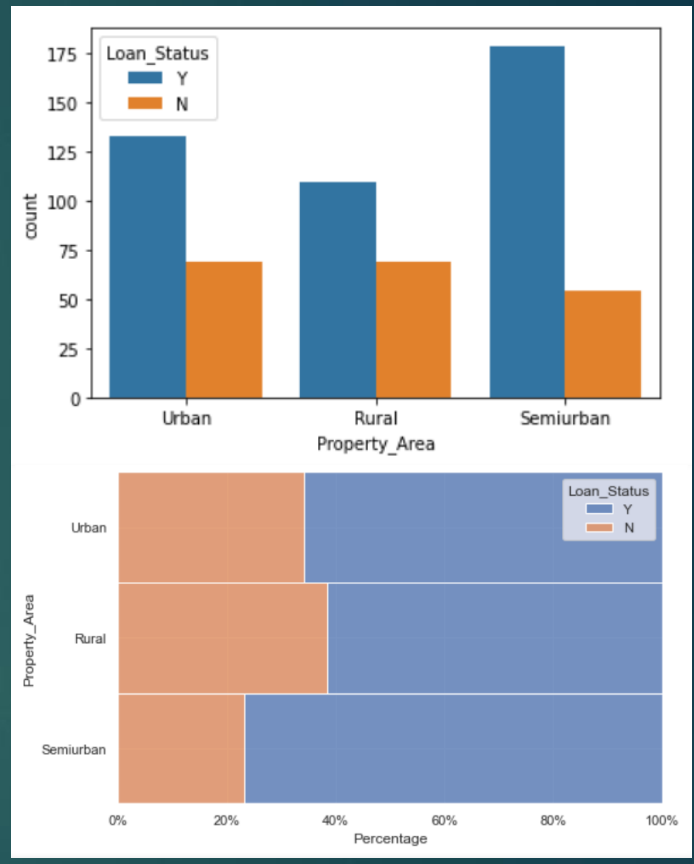


- A person's gender, self employment status, dependents and loan amount term have little impact on being able to get your bank loan approved.

Data Exploration

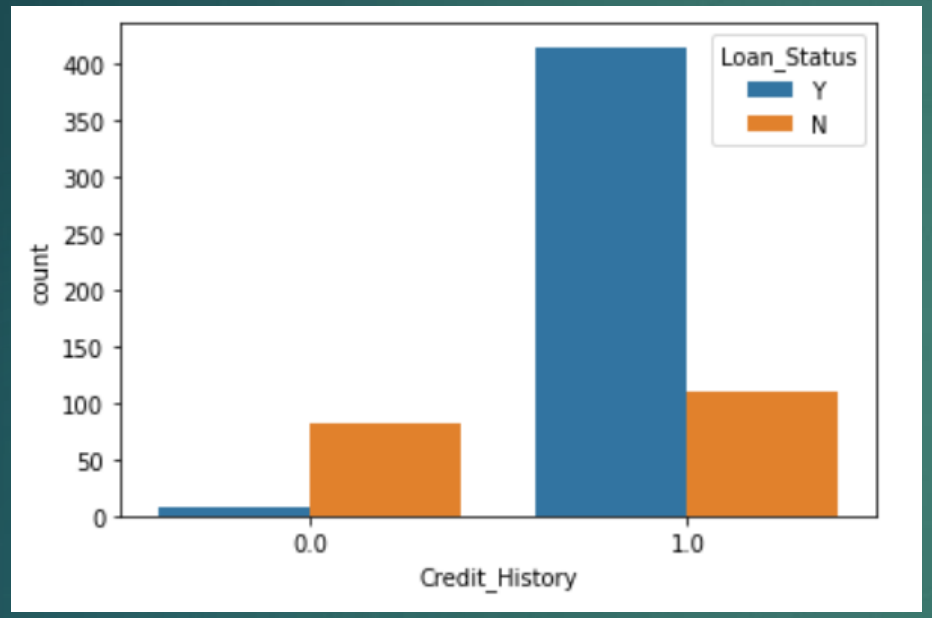


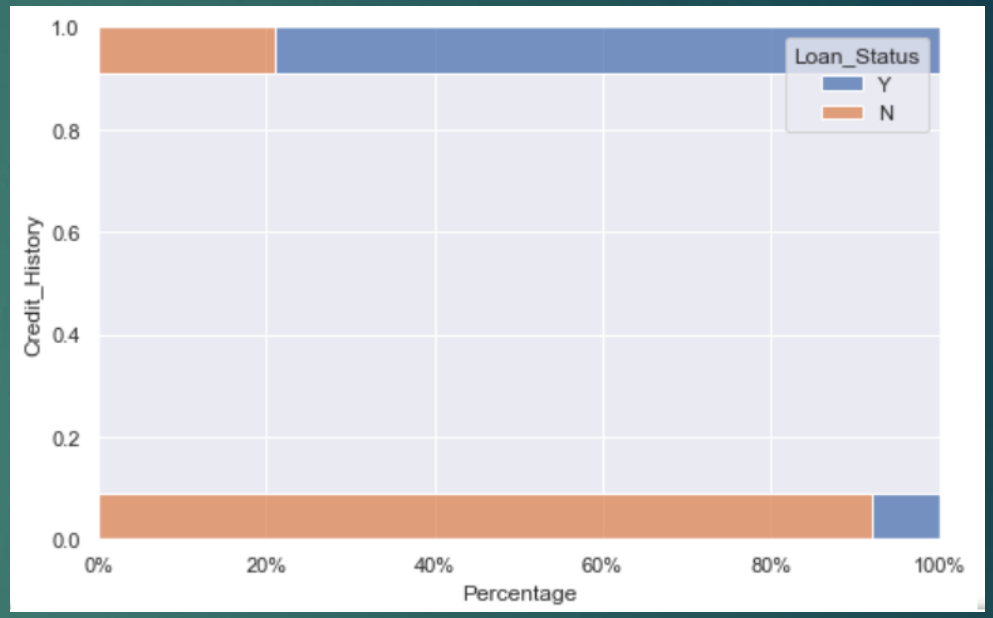




- Being a graduate, married and/or living in a semiurban area increases your likelihood of being able to get your bank loan approved.

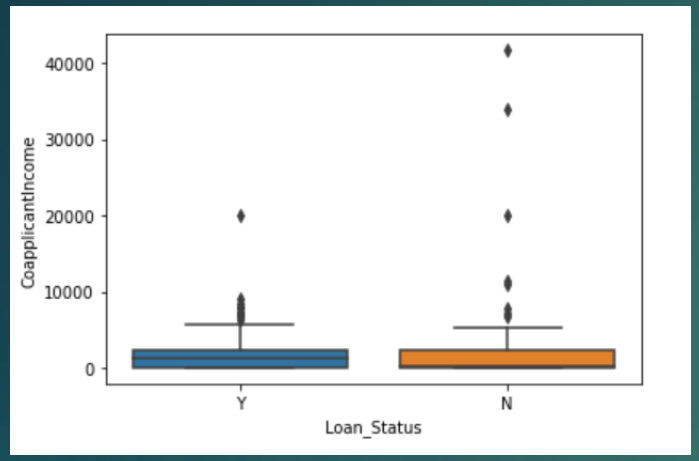
Data Exploration



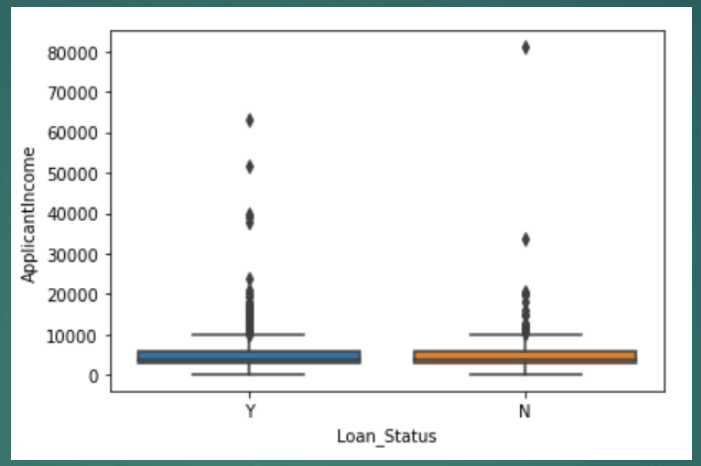


- The best indicator of whether a bank loan is approved is based on having a good credit history.

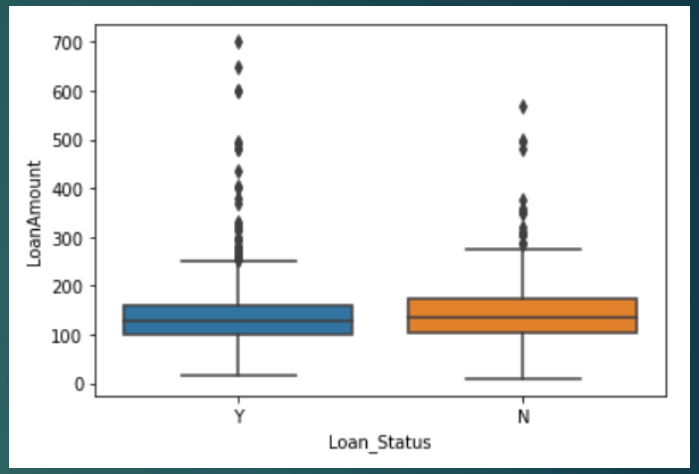
Data Exploration



Mean	Std Dev
N : 5446.078	N : 6819.559
Y : 5384.069	Y : 5765.442



Mean	Std Dev
N : 1877.807	N : 4384.060
Y : 1504.516	Y : 1924.755



Mean	Std Dev
N : 150.945	N : 83.361
Y : 144.350	Y : 84.361

- Looking at the means and standard deviations of ApplicantIncome, CoapplicantIncome and LoanAmount, we can see that these variables have very little effect on the loan outcome as the mean and standard deviations for both yes and no are very close to each other.

Prediction Results

- Three models were used for the predictions.
- Two key metrics were used to determine the capability of the models being Accuracy and ROC.
- With better accuracy and ROC, the model is more capable in predicting which applicants are more likely to be approved.
- As such, Logistic Regression is the most optimal model to use to make predictions regarding bank approval based on a given dataset.

KNearestNeighbors	Logistic Regression	Decision Leaf Tree
Accuracy = 0.67742 ROC = 0.52381	Accuracy = 0.75806 ROC = 0.65447	Accuracy = 0.61290 ROC = 0.57956

Limitations

- The capability of the prediction model is limited by the amount of training data as well as the amount of test data that can be given to it. As this dataset has an amount of data that is only in the hundreds, the data used is very limiting hence reducing the accuracy of the prediction model.