# Loan Status Prediction

## Introduction

## This is a classification problem in which we need to classify whether the loan will be approved or not. Classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.

## Problem Statement

As mentioned above this is a Binary Classification problem in which we need to predict our Target label which is “Loan Status”.Loan status can have two values: Yes or NO.

Yes: if the loan is approved

NO: if the loan is not approved

So using the training dataset we will train our model and try to predict our target column that is “Loan Status” on the test dataset.

## About the dataset

## So train and test dataset would have the same columns except for the target column that is “Loan Status”.

## Train Dataset:

## https://lh4.googleusercontent.com/Toh_V2T2hPde7_CwbNHbO4oiAk01LykBuVEd86fSQgqiLWu4CHGavONhvBHfXBMCW9Q8S32WmiI5bge5uIZ_70KWbhu_x8Vl-VCF4PC1BkyDn7v4m1q0hfRiF4NfL4U9mdpGSdoA

**Data Preprocessing**

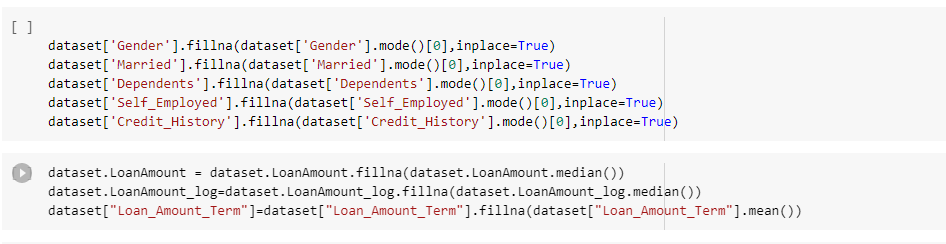
Identify missing values:



Imputing the missing values:

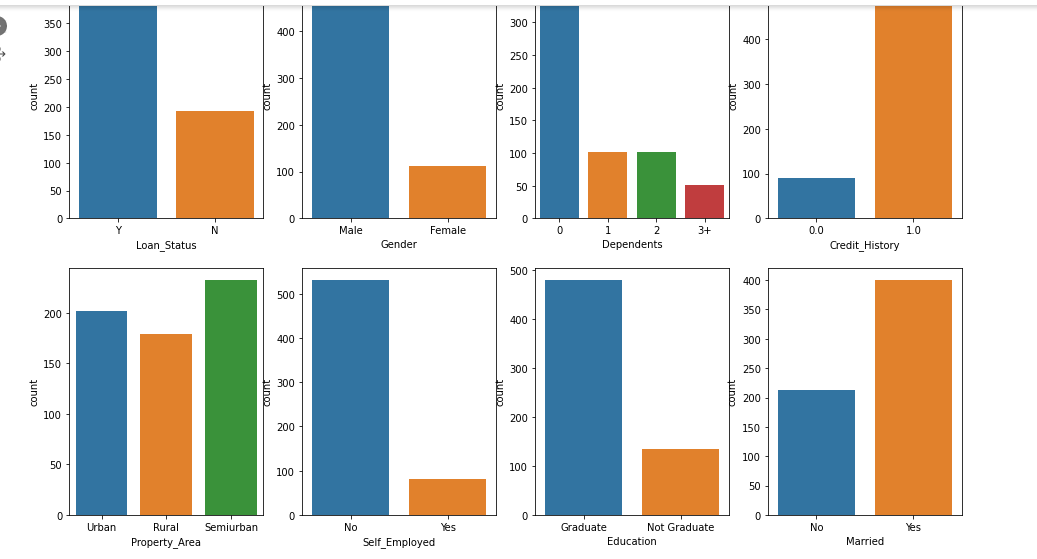
As Gender, Married, Dependents, Self\_Employed, Credit\_History are categorical variables we imputed the missing values with mode.

LoanAmount, Loan\_Amount\_Term features missing values are replaced with median and mean.



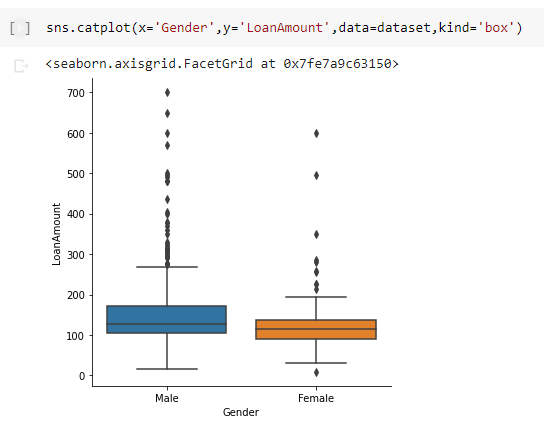
**Exploratory Data Analysis(EDA)**

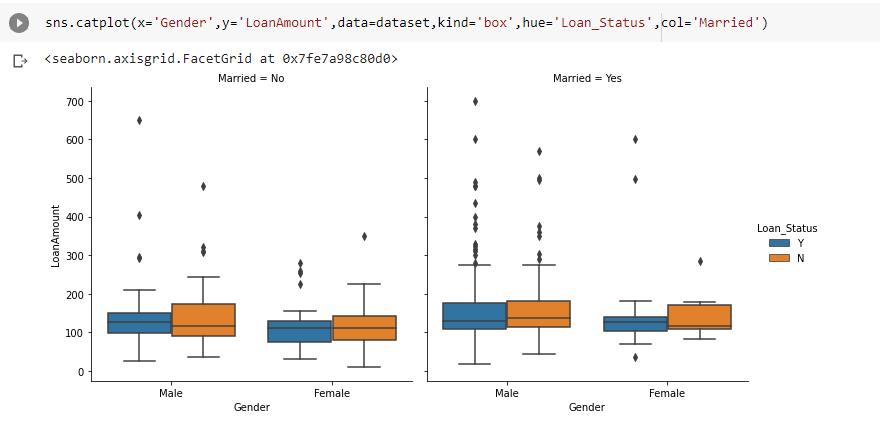
Univariate Analysis on categorical features using countplot:

****

**Univariate Analysis Observations**

1. More Loans are approved Vs Rejected
2. Count of Male applicants is more than Female
3. Count of Married applicant is more than Non-married
4. Count of graduate is more than non-Graduate
5. Count of self-employed is less than that of Non-Self-employed
6. Maximum properties are located in Semiurban areas
7. Credit History is present for many applicants
8. The count of applicants with several dependents=0 is maximum

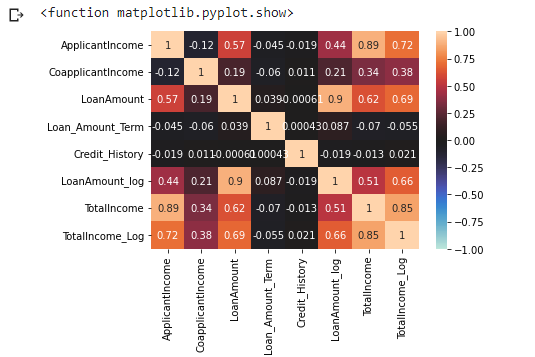
**Bivariate Analysis**



The mean value of Loan Amount applied by males is slightly higher than Females.

If you are married then the loan amount requested is slightly higher than non-married.

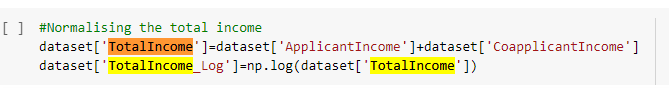
**Correlation matrix**



**Feature Engineering**

**pandas.get\_dummies()** is used for data manipulation. It converts categorical data into dummy or indicator variables.

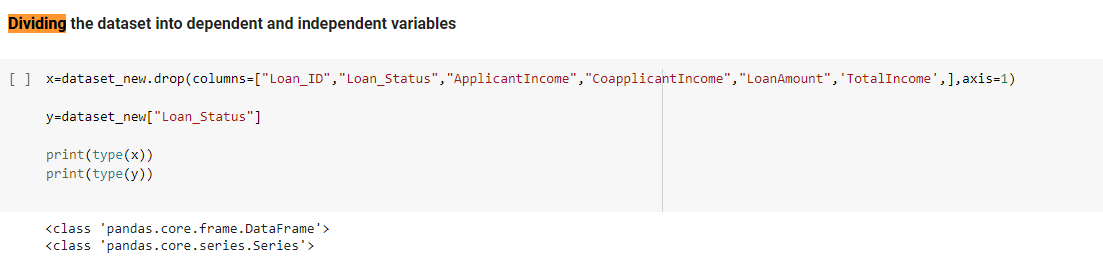
****



TotalIncome= ApplicantIncome+ CoapplicantIncome

Dropped the unwanted columns from train dataset.

Creating x (input variables) and y(Target Variable) from the new\_train data.

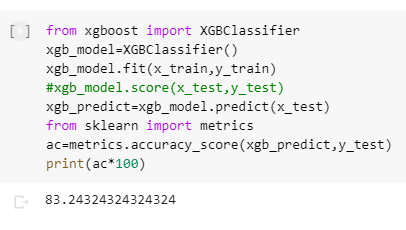


Using train test split on the training data for validation



We have a (70:30) split on the training data.

**Using ML algorithm for training**



We have used multiple algorithms for training purposes like Decision Tree, Random Forest, SVC, Logistic Regression, XGB Regressor, etc.

Among all the algorithms **XGBOOST**  performs best on the validation data with an accuracy score of **83.2%**.

**Insights**

1.After using pdummies instead of my own encoding for the categorical features accuracy increased from

78 % to 80% .

2. I have performed scaling on few features to reduce the range but the accuracy remained same. It has increased to 82% when the features are not scaled.

3.After splitting the data to train and test in the ratio of 70:30 from 80:20 my accuracy increased from 82% to 83.2%.