LOAN PREDICTION

BY- SHREYA SHARMA

OBJECTIVE

To Predict the people those are eligible for loan amount based on the customers of the company.

PROBLEM

A Company wants to automate the loan eligibility process (real time) based on customer de-tail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a data set.

LIBRARY IMPORT

PANDAS

NUMPY

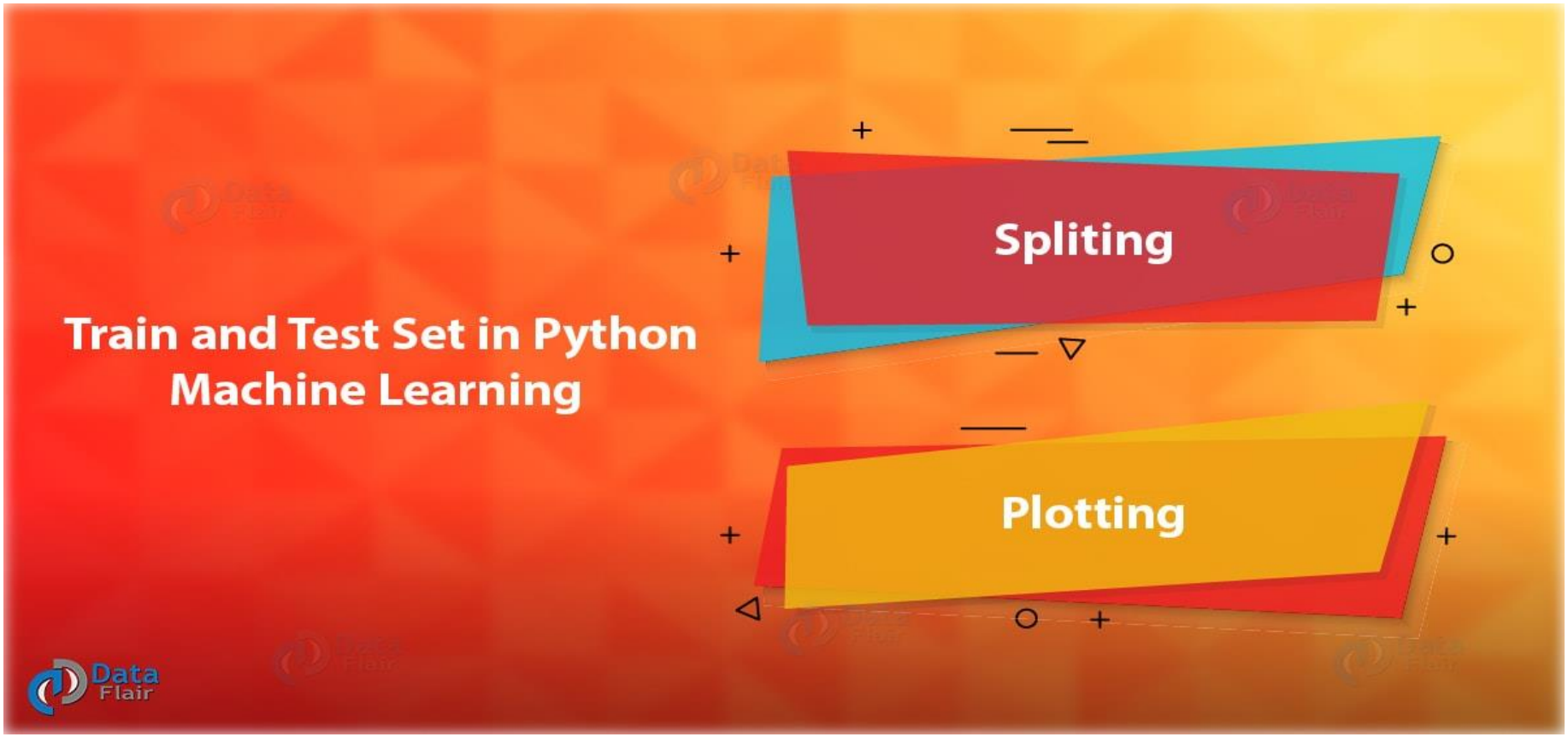
MATPLOTLIB

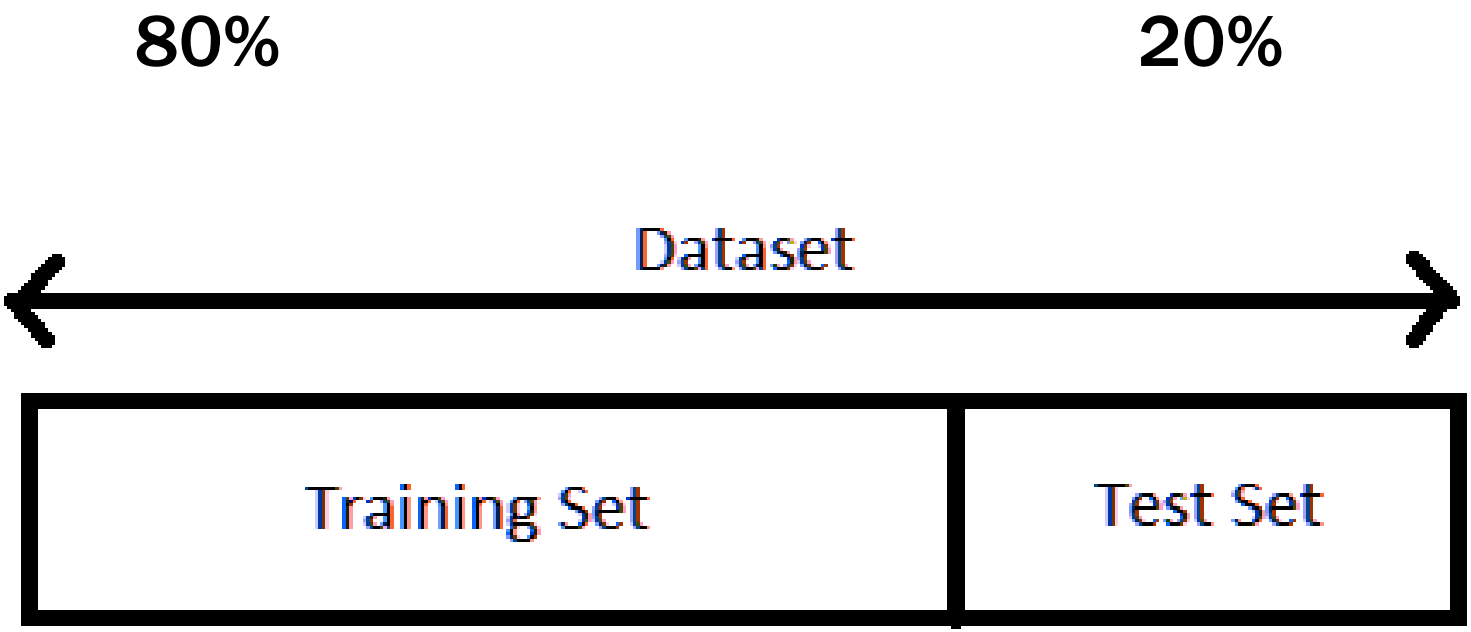
SKLEARN

Data

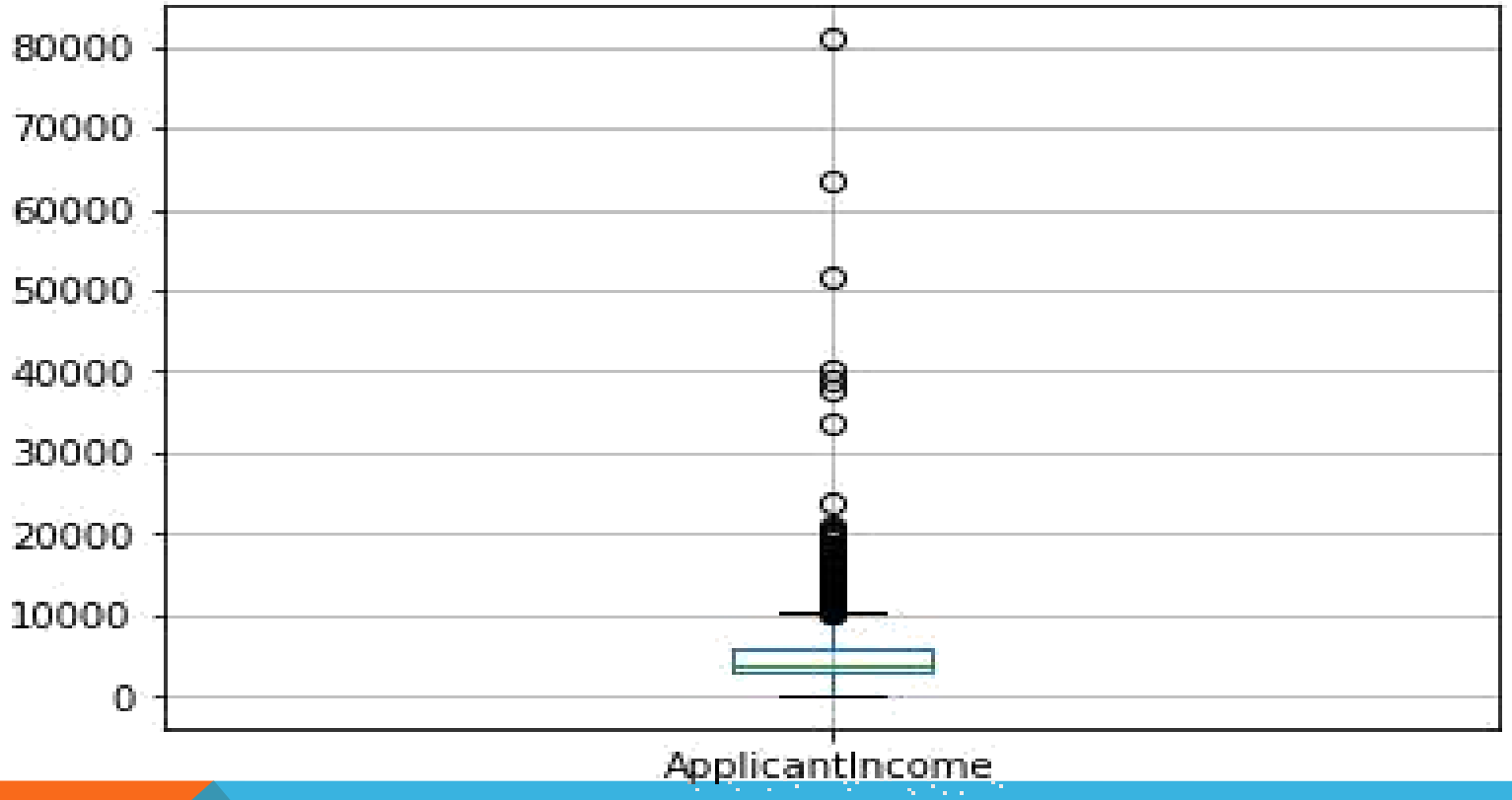
- Variable Descriptions:

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)

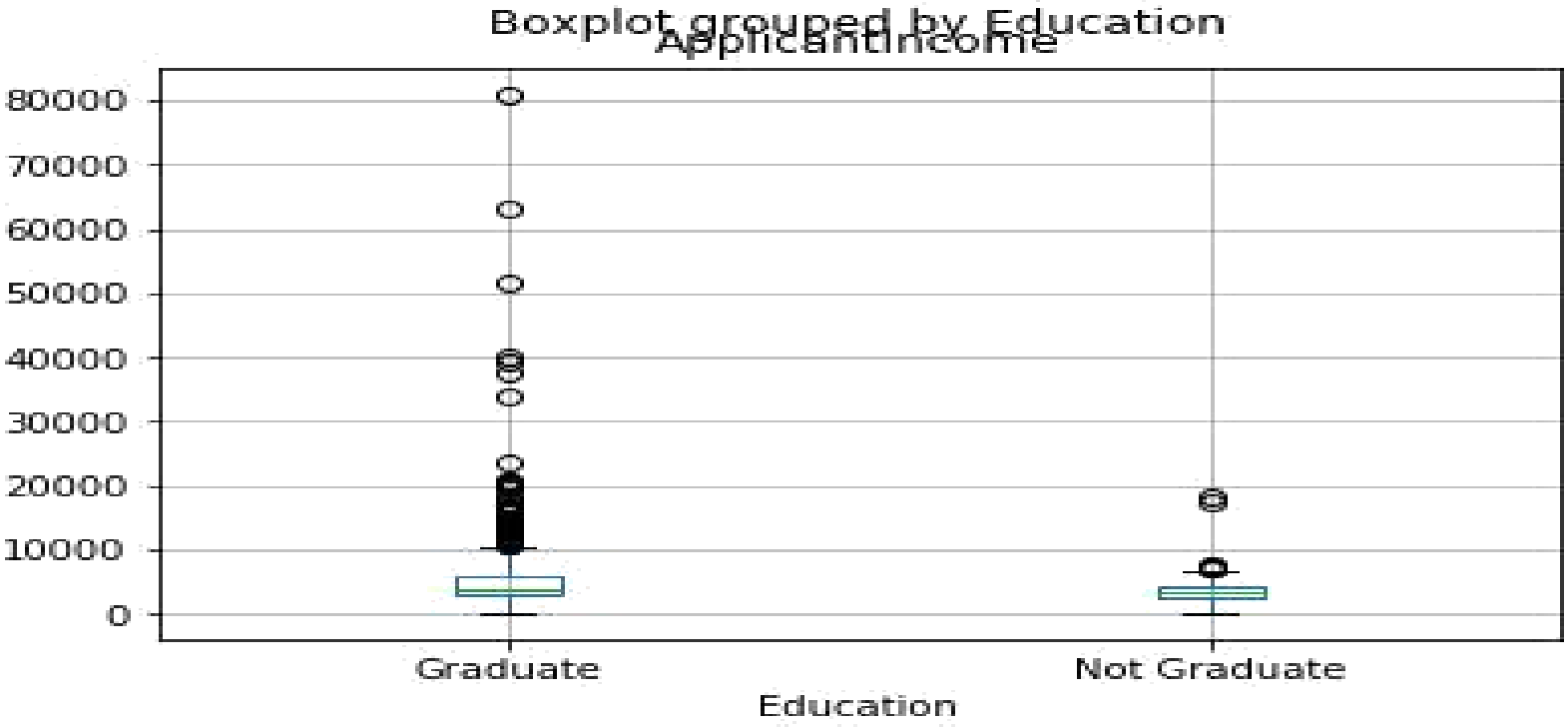

Train and Test Set in Python
Machine Learning
Spliting
Plotting
Data
Flair


80%
20%
Dataset
Training Set
Test Set

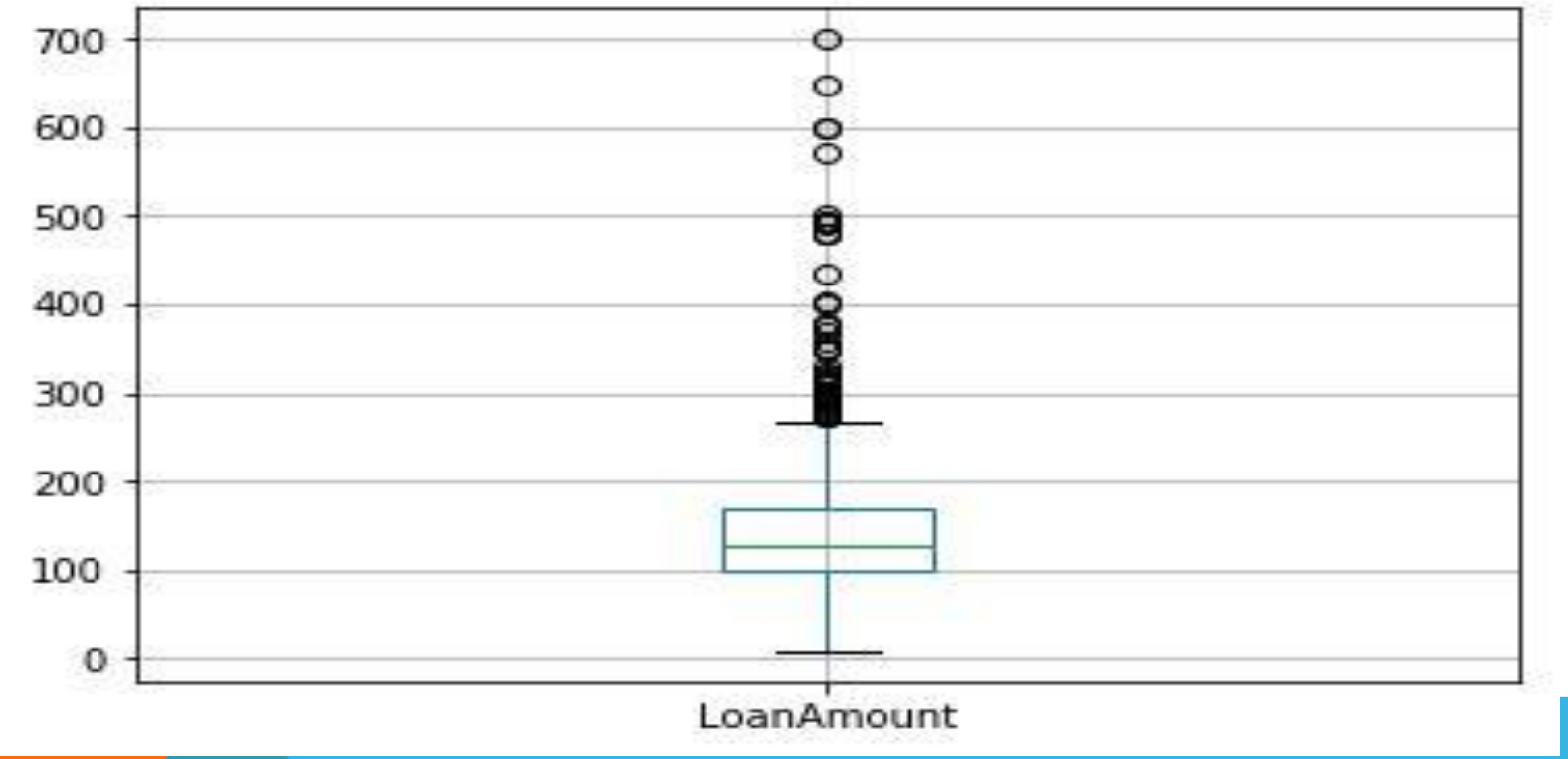
APPLICANT INCOME OF TRAINING DATA SET



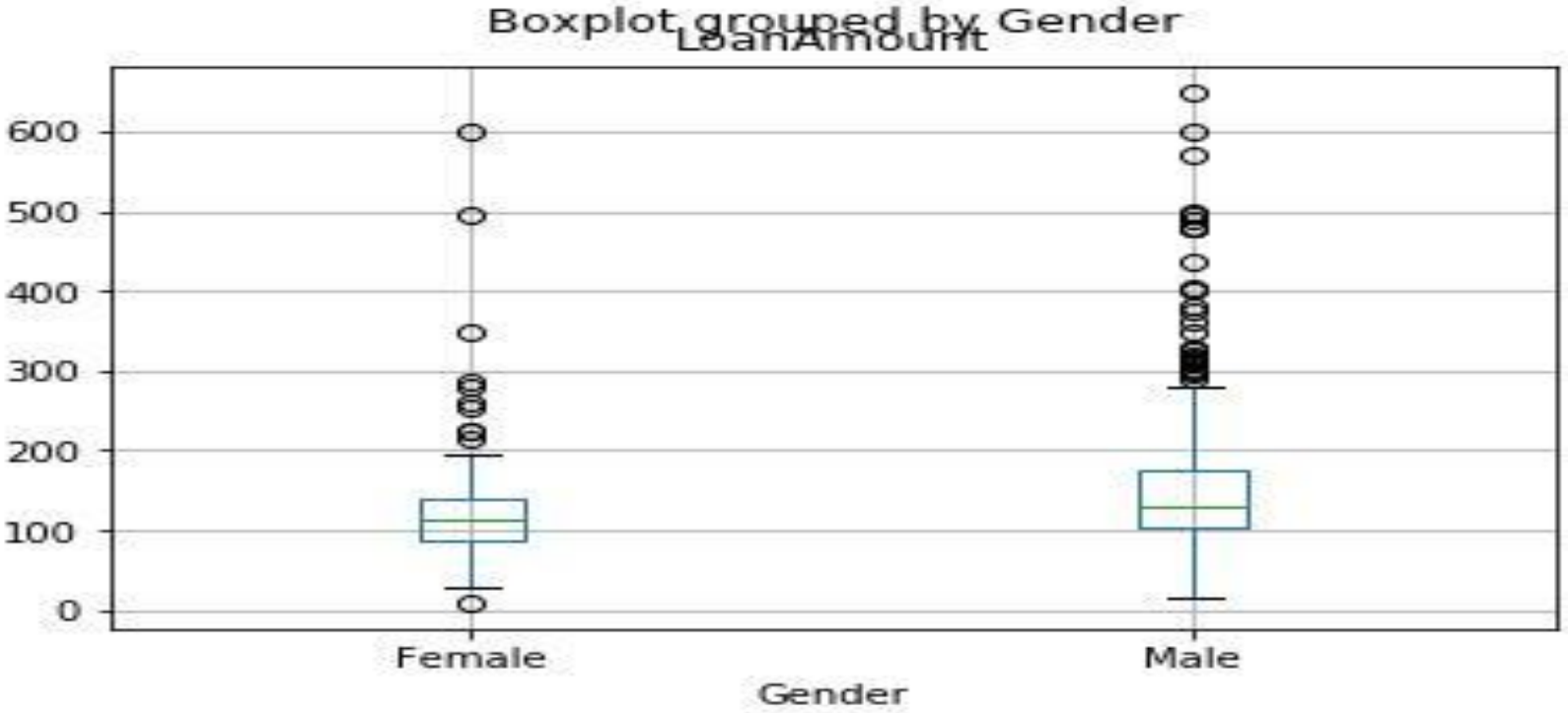
APPLICANT INCOME BY VARIABLE EDUCATION OF TRAINING DATA SET



LOAN AMOUNT OF TRAINING DATA SET



LOAN AMOUNT BY VARIABLE GENDER OF TRAINING DATA SET



UNDERSTANDING DISTRIBUTION OF CATEGORICAL VARIABLES

]: # Loan approval rates in absolute numbers

loan_approval = df['Loan_Status'].value_counts()['Y'] print(loan_approval)

- 422 number of loans were approved.

Out[37]: Loan_Status	N	Y	All
Credit_History			
0.0	82	7	89
1.0	97	378	475
All	179	385	564

output percentage

- **79.58 % of the applicants whose loans were approved have Credit_History equals to 1.**

df['Y']

Credit_History	
0.0	0.078652
1.0	0.795789
All	0.682624

Name: Y, dtype: float64

DATA PREPARATION FOR MODEL BUILDING

SKLEARN REQUIRES ALL INPUTS TO BE NUMERIC, WE SHOULD CONVERT ALL OUR CATEGORICAL VARIABLES INTO NUMERIC BY ENCODING THE CATEGORIES. BEFORE THAT WE WILL FILL ALL THE MISSING VALUES IN THE DATASET.

MODEL BUILDING

Look at the available missing values in the dataset

fullData.isnull().sum()

Out[186]: ApplicantIncome	0
CoapplicantIncome	0
Credit_History	29
Dependents	10
Education	0
Gender	11
LoanAmount	27
LoanAmount_log	389
Loan_Amount_Term	20
Loan_ID	0
Loan_Status	367
Married	0
Property_Area	0
Self_Employed	23
Type	0
dtype: int64	

Identify categorical and continuous variables

Imputing Missing values with mean for continuous variable

Imputing Missing values with mode for categorical variables

Create a new column as Total Income

LOGISTIC REGRESSION MODEL

We make our model with ‘Credit_History’, ’Education’ & ’Gender’

Create logistic regression object

Predict Output

Store it to test dataset

Accuracy : 80.945%

Cross-Validation Score : 80.946%