PROJECT REPORT ON LOAN PREDICTION

Problem Definition

Loan Application Status Prediction

This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.

Independent Variables:

- Loan_ID
- Gender
- Married
- Dependents
- Education
- Self_Employed
- ApplicantIncome
- CoapplicantIncome
- Loan_Amount
- Loan_Amount_Term
- Credit History
- Property_Area

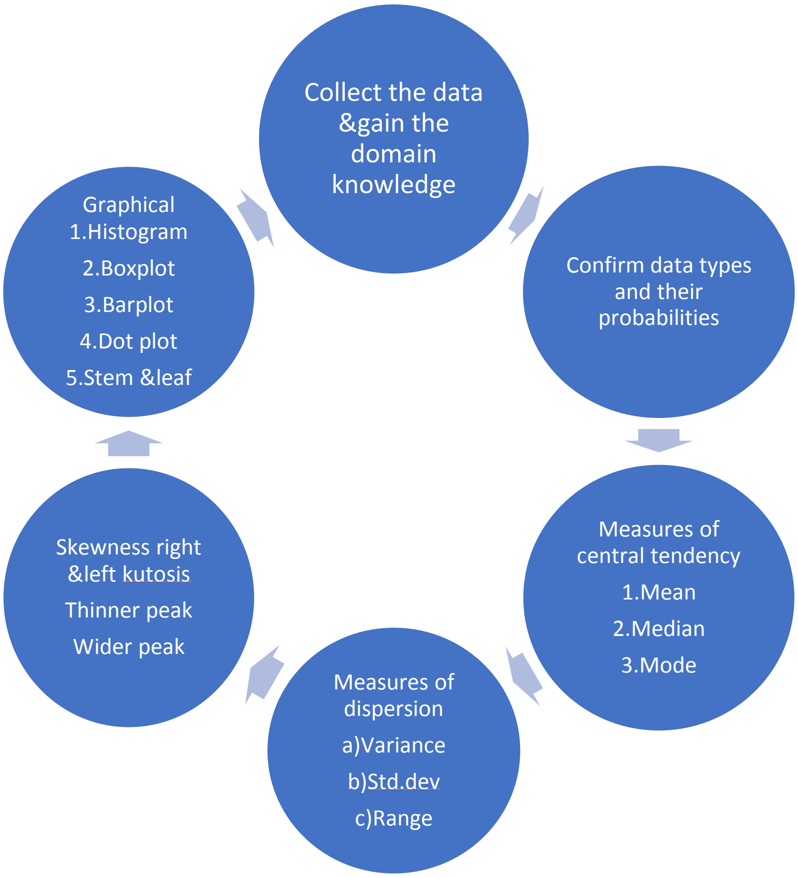
Dependent Variable (Target Variable):

- Loan_Status

You have to build a model that can predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset.

Data Analysis

Below are the steps for Exploratory Data Analysis



Collect the data &gain the domain knowledge

Confirm data types and their probabilities

Measures of central tendency
1.Mean
2.Median
3.Mode

Measures of dispersion
a)Variance
b)Std.dev
c)Range

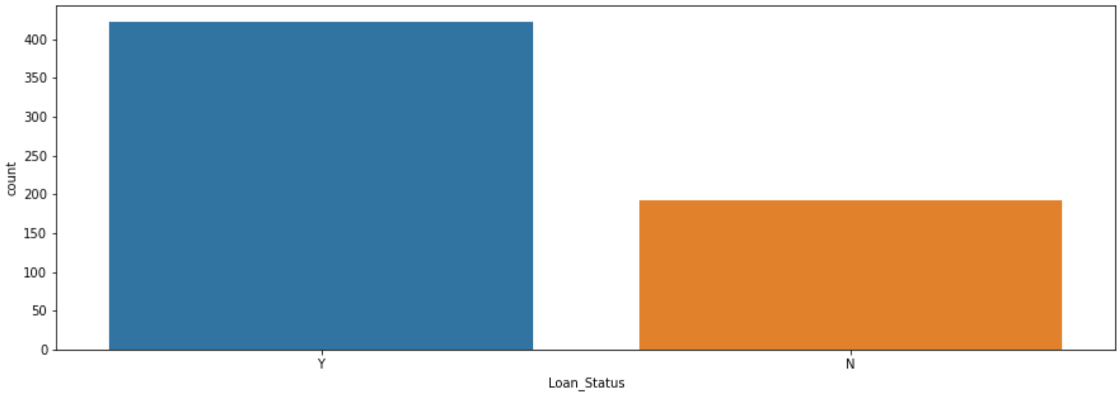
Skewness right &left kutosis
Thinner peak
Wider peak

Graphical
1.Histogram
2.Boxplot
3.Barplot
4.Dot plot
5.Stem &leaf

Variable analysis

1. Target Variable - Loan Status

```
Y    422
N    192
Name: Loan_Status, dtype: int64
```

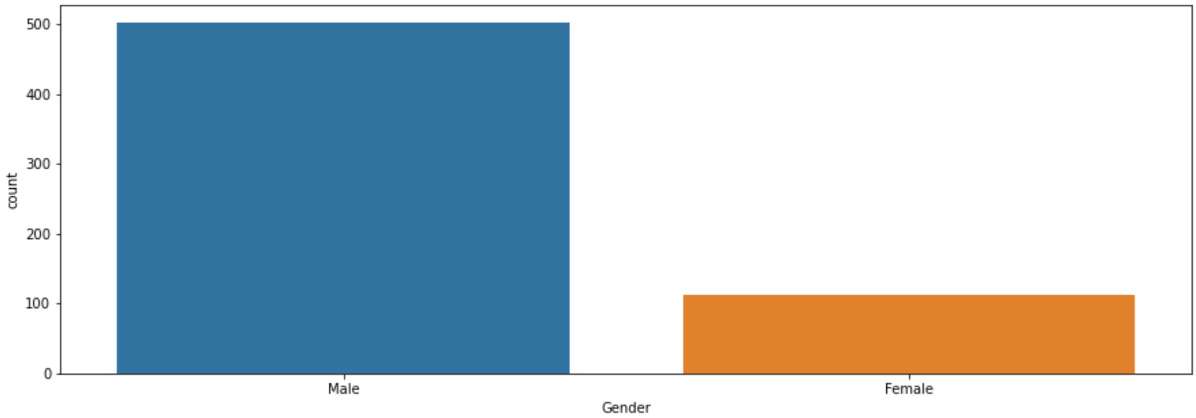


Predictor Variables

There are 3 types of Independent Variables: Categorical, Ordinal & Numerical.
Here the categorical variables are

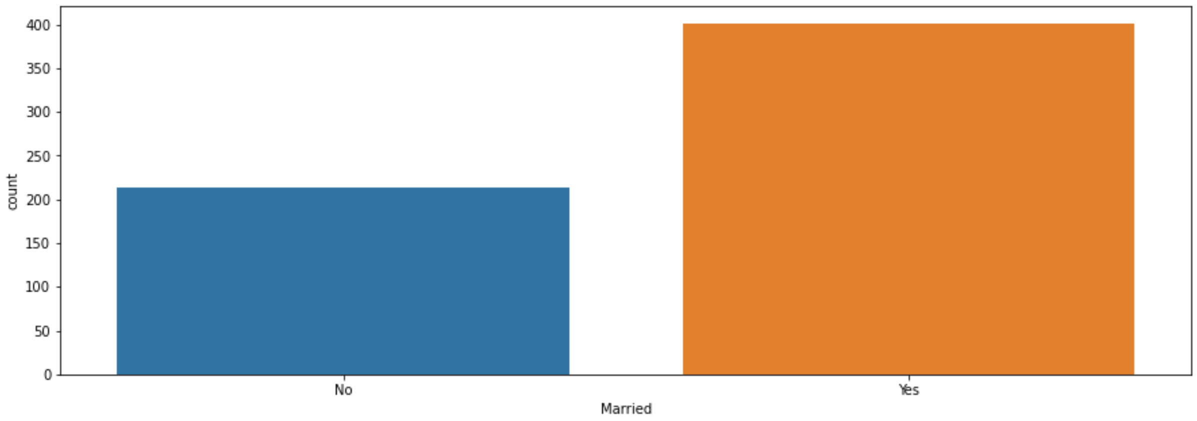
- Gender

```
Male      502
Female    112
Name: Gender, dtype: int64
```



- Marital Status

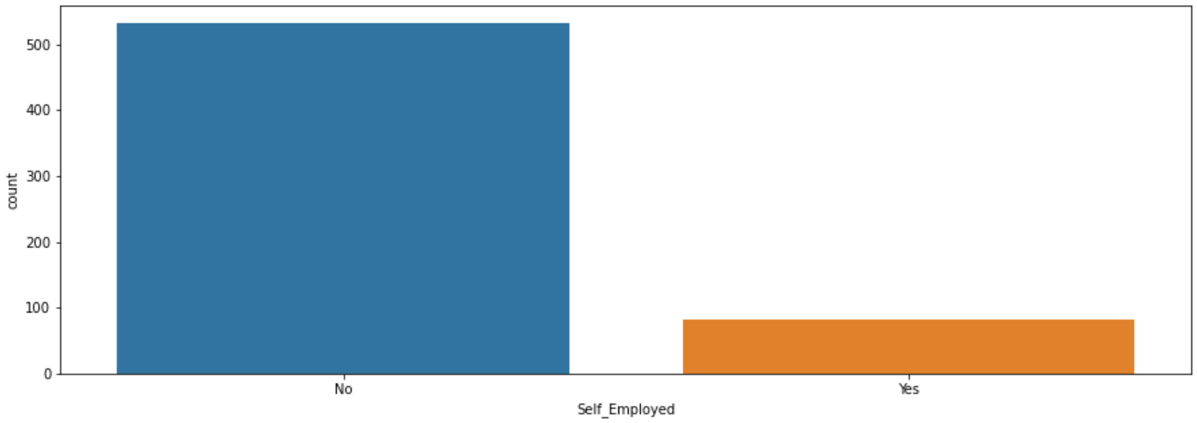
```
Yes    401
No     213
Name: Married, dtype: int64
```



- Employment Type

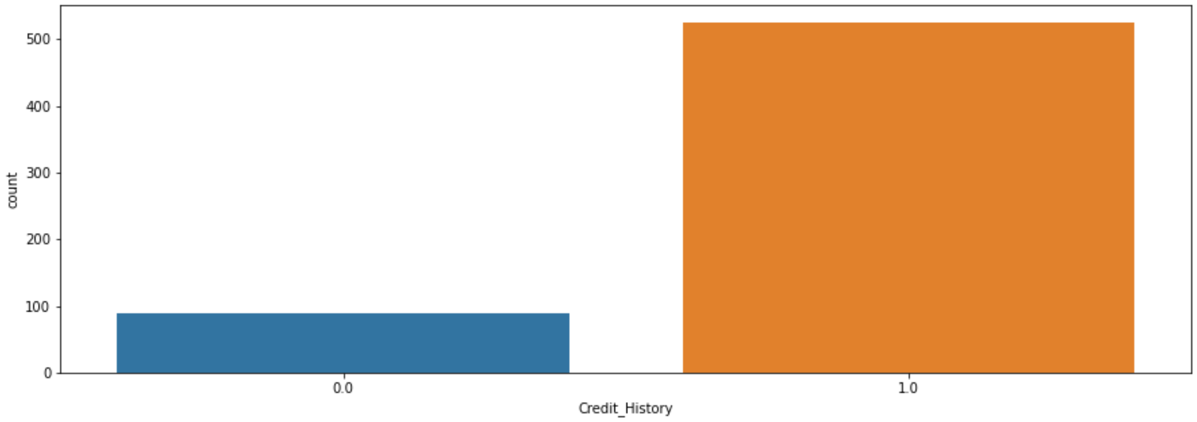
```
No     532
Yes     82
Name: Self_Employed, dtype: int64
```

- Credit History



- Credit History

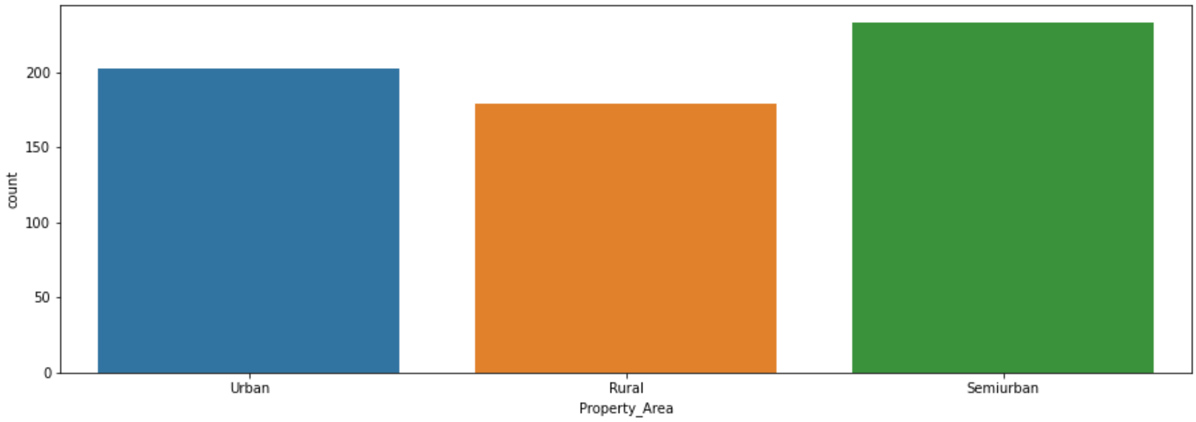
```
1.0    525
0.0     89
Name: Credit_History, dtype: int64
```



Ordinal Features

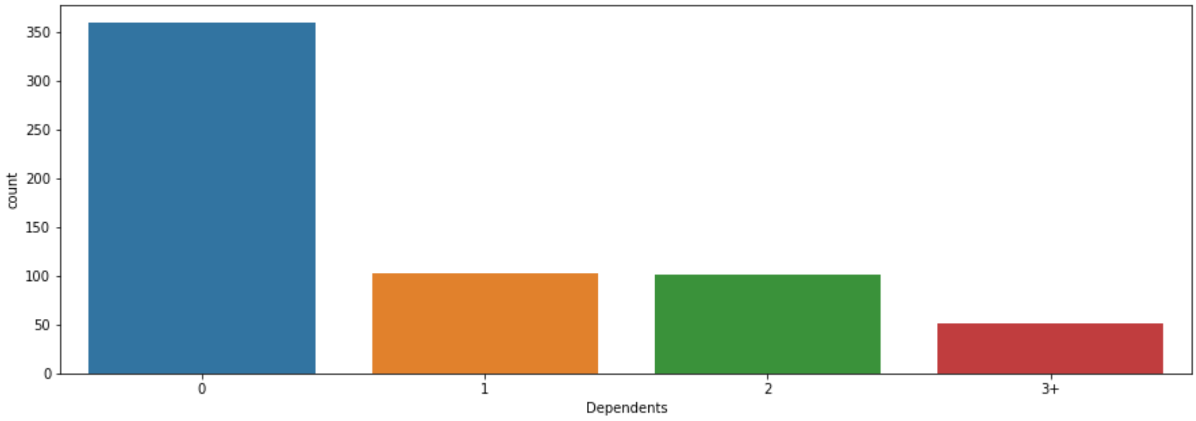
- Property or Area Background

```
Semiurban    233
Urban        202
Rural        179
Name: Property_Area, dtype: int64
```



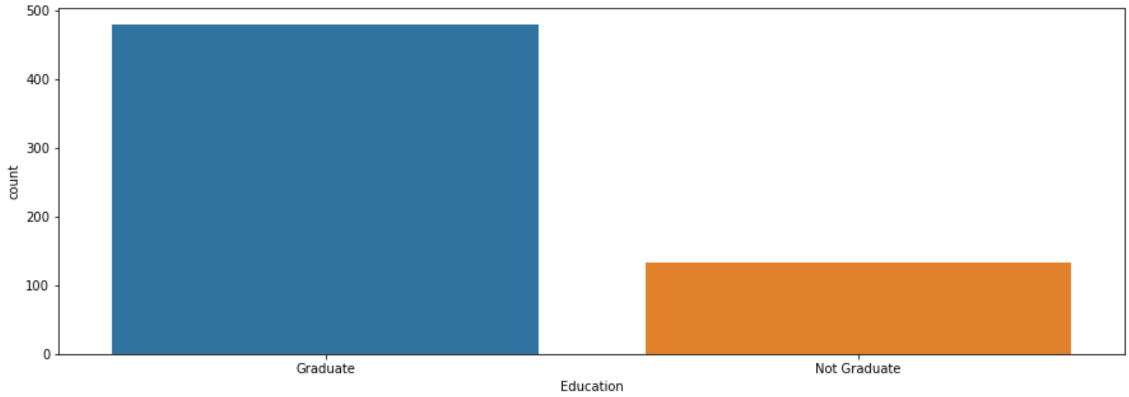
- Number of Dependents

```
0     360
1     102
2     101
3+     51
Name: Dependents, dtype: int64
```



- Education Level

```
Graduate        480
Not Graduate    134
Name: Education, dtype: int64
```



Numerical Features

- The Applicant's Income
- The Co-Applicant's Income

EDA Concluding Remarks

- ➢ 80% of loan applicants are male in the training dataset.
- ➢ Nearly 70% are married
- ➢ Nearly 85–90% loan applicants are self-employed
- ➢ The loan has been approved for more than 65% of applicants.
- ➢ Almost 58% of the applicants have no dependents.
- ➢ Highest number of applicants are from Semi Urban areas, followed by urban areas.
- ➢ Around 80 % of the applicants are Graduate.

Pre-processing Pipeline

A. Data Cleaning

a. Finding the Data type of variables

```
Loan_ID                object
Gender                 object
Married                object
Dependents             object
Education              object
Self_Employed          object
ApplicantIncome         int64
CoapplicantIncome     float64
LoanAmount            float64
Loan_Amount_Term      float64
Credit_History        float64
Property_Area          object
Loan_Status            object
dtype: object
```

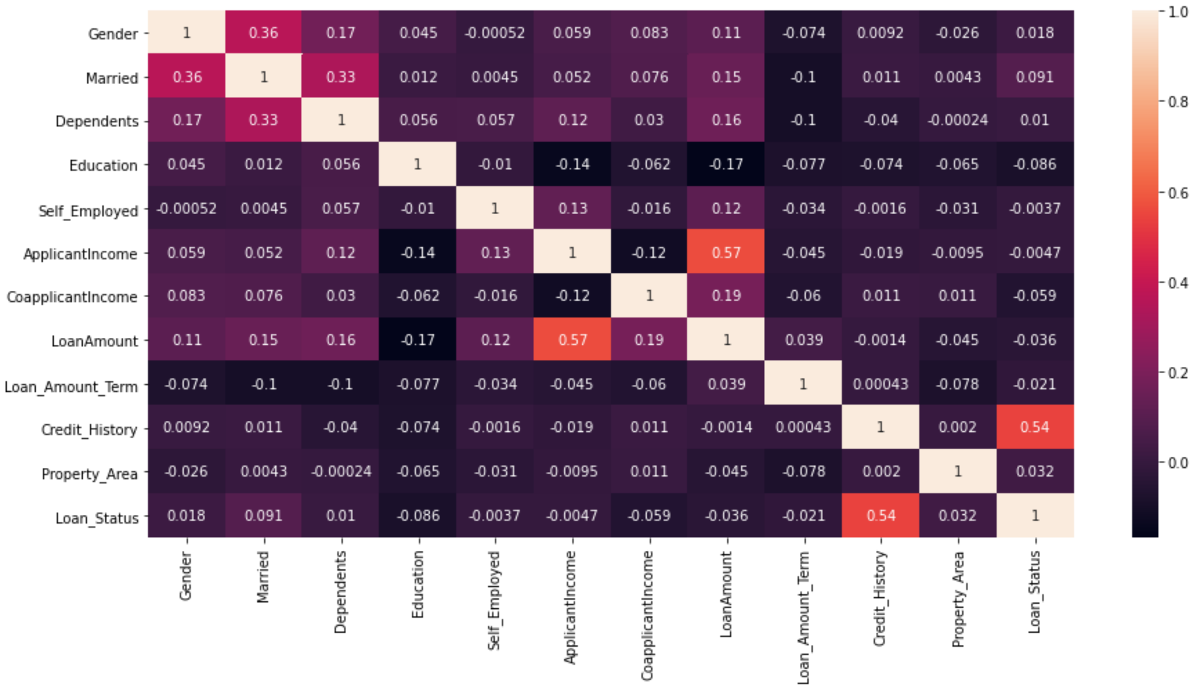
b. Handling missing/null values:

```
Loan_ID               0
Gender               13
Married               3
Dependents           15
Education             0
Self_Employed        32
ApplicantIncome       0
CoapplicantIncome     0
LoanAmount           22
Loan_Amount_Term     14
Credit_History       50
Property_Area         0
Loan_Status          0
dtype: int64
```

Now we know that the missing/null values are replaced by mean or mode of the remaining values depending on the type of data whether float or object type. For float type it is mean, and for object type it is mode. Thus handled the missing values.

c. Dummy variables for categorical variables: Now before model building, we need to encode the object data type variables in 0 or 1 so as to work on the data. Also, we need to drop the columns we don’t need, like here the loan ID is unique thus dropped it off.

d. Correlation between Quantitative Variables
Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation. Here we see the below graph.



	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
Gender	1	0.36	0.17	0.045	-0.00052	0.059	0.083	0.11	-0.074	0.0092	-0.026	0.018
Married	0.36	1	0.33	0.012	0.0045	0.052	0.076	0.15	-0.1	0.011	0.0043	0.091
Dependents	0.17	0.33	1	0.056	0.057	0.12	0.03	0.16	-0.1	-0.04	-0.00024	0.01
Education	0.045	0.012	0.056	1	-0.01	-0.14	-0.062	-0.17	-0.077	-0.074	-0.065	-0.086
Self_Employed	-0.00052	0.0045	0.057	-0.01	1	0.13	-0.016	0.12	-0.034	-0.0016	-0.031	-0.0037
ApplicantIncome	0.059	0.052	0.12	-0.14	0.13	1	-0.12	0.57	-0.045	-0.019	-0.0095	-0.0047
CoapplicantIncome	0.083	0.076	0.03	-0.062	-0.016	-0.12	1	0.19	-0.06	0.011	0.011	-0.059
LoanAmount	0.11	0.15	0.16	-0.17	0.12	0.57	0.19	1	0.039	-0.0014	-0.045	-0.036
Loan_Amount_Term	-0.074	-0.1	-0.1	-0.077	-0.034	-0.045	-0.06	0.039	1	0.00043	-0.078	-0.021
Credit_History	0.0092	0.011	-0.04	-0.074	-0.0016	-0.019	0.011	-0.0014	0.00043	1	0.002	0.54
Property_Area	-0.026	0.0043	-0.00024	-0.065	-0.031	-0.0095	0.011	-0.045	-0.078	0.002	1	0.032
Loan_Status	0.018	0.091	0.01	-0.086	-0.0037	-0.0047	-0.059	-0.036	-0.021	0.54	0.032	1

```
Loan_Status          1.000000
Credit_History       0.540556
Married              0.091478
Property_Area        0.032112
Gender               0.017987
Dependents           0.010118
Self_Employed       -0.003700
ApplicantIncome     -0.004710
Loan_Amount_Term    -0.020974
LoanAmount          -0.036416
CoapplicantIncome   -0.059187
Education           -0.085884
```

We see that the factors affecting the loan status are Credit History, Married or not followed by property area and gender. But the most important positively affecting factor is Credit History as Credit History is important in banks to give any person loan. A credit history is the record of how a person has managed his or her credit in the past, including total debt load, number of credit lines, and timeliness of payment. Lenders look at a potential customer's credit history to decide whether or not to offer a new line of credit, and to help set the terms of the loan. As we know it depends a lot on gender as seen in previous scenarios that males are less likely to save and return loan amount than women. Also, the negative factors affecting the loan application status is the education level and Applicant Income.

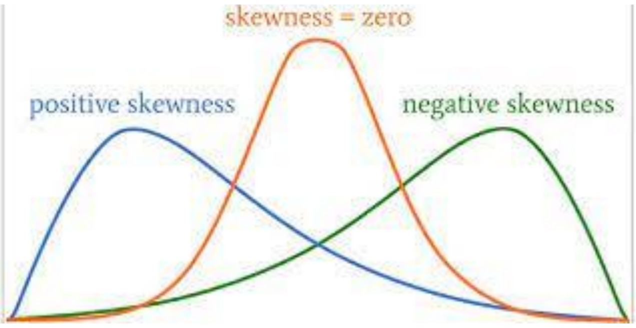
Building Machine Learning Models

For model building, we first need to separate the dependent and independent variables x and y. Here the factors are the independent variables and Loan_Status is the dependent variable as discussed thus dropping off Loan_Status and proceeding.

a. Checking the skewness:

Skewness is a measure of the symmetry of a distribution. The highest point of a distribution is its mode. The mode marks the response value on the x-axis that occurs with the highest probability. A distribution is skewed if the tail on one side of the mode is fatter or longer than on the other: it is asymmetrical.

In an asymmetrical distribution a negative skew indicates that the tail on the left side is longer than on the right side (left-skewed), conversely a positive skew indicates the tail on the right side is longer than on the left (right-skewed). Asymmetric distributions occur when extreme values lead to a distortion of the normal distribution.



In the data we get the below skewness:

```
CoapplicantIncome     7.491531
ApplicantIncome       6.539513
LoanAmount            2.726601
Self_Employed         2.159796
Education             1.367622
Dependents            1.015551
Property_Area        -0.066196
Married              -0.644850
Gender               -1.648795
Credit_History       -2.021971
Loan_Amount_Term     -2.389680
```

We see that the data is skewed and to bring the skewness in the range of (-0.5,0.5) using power transform

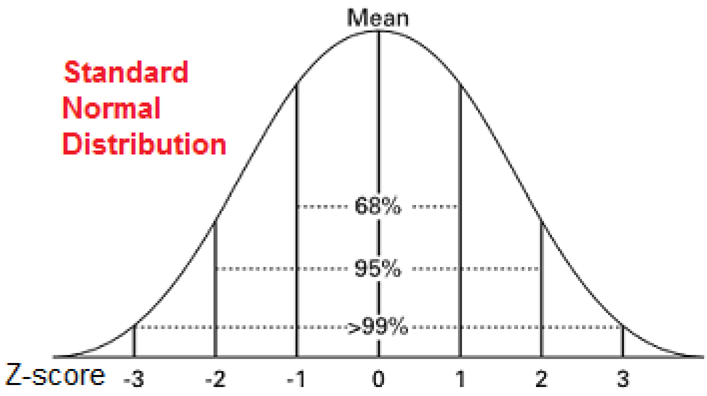
```
Self_Employed        2.159796
Education            1.367622
Loan_Amount_Term     0.490150
Dependents           0.377295
Property_Area       -0.041074
LoanAmount          -0.056384
CoapplicantIncome   -0.102698
```

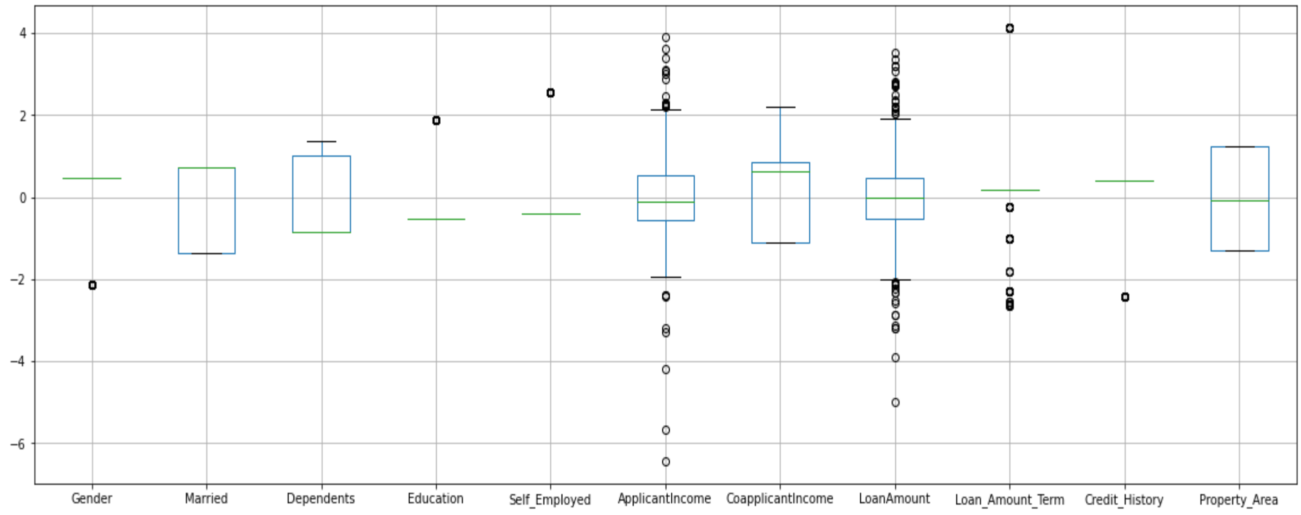
```
ApplicantIncome      -0.284434
Married              -0.644850
Gender               -1.648795
Credit_History       -2.021971
```

Now we see that the data is not skewed, and we have removed the skewness. Thus, we can now proceed.

b. Handling Outliers

Now we need to check whether outliers are present in the data or not. For that we need to check if the z value/z score of all the factors is exceeding the range (-3,3). As we want the data to be normally distributed in the range of (-3,3) as the data be in the 99% domain. Thus this will be the best data to work with.





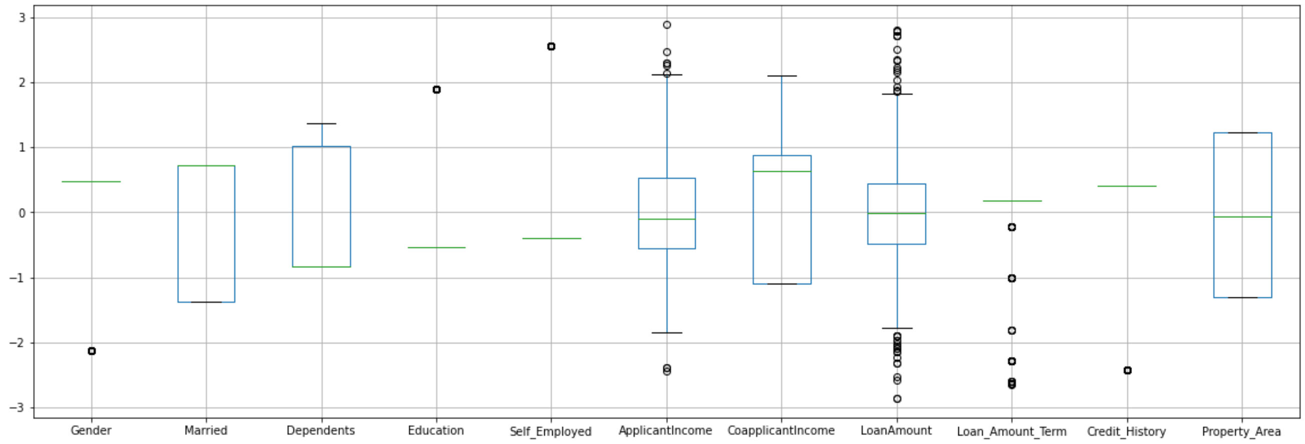
Here we see there are a few outliers, thus removing them and bringing the data in the range of (-3,3)

```
from scipy.stats import zscore

z=np.abs(zscore(new_df))

new_df=new_df[(z<3).all(axis=1)]
```

Using the above code, we were successfully able to remove the outliers and we get the below range of data



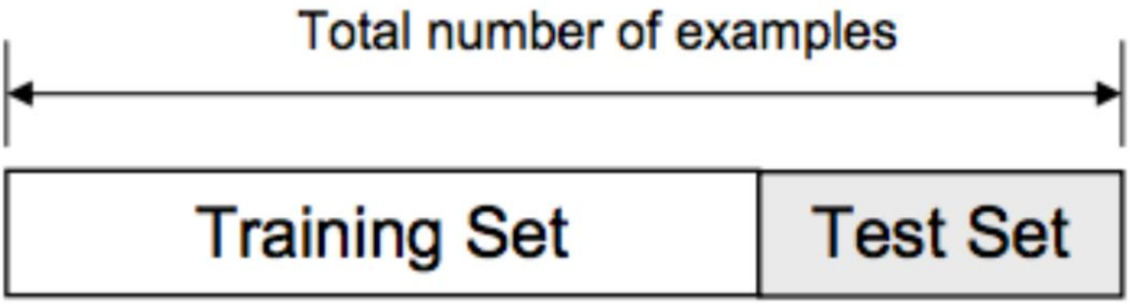
As we see there are no outliers, thus can proceed with modelling

c. Regression model: To predict the Loan Application Status we need to do Logistic regression modelling as the value of Loan_Status is 0 or 1 thus binary mapping depending whether the loan application of a person gets approved or not

Running the regression, we get best accuracy as 0.9230769230769231 on Random state 263.

Train/Test Split

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset



Using the random state, we divide the training and testing data in the ratio of 0.8 and 0.2 meaning 20% of the data is the testing data and training data is 80%

Thus,
Train data= 465 records
Test data= 117 records
As total records are 595 as we removed the outliers.

Now running the regression models, below are the accuracy and results we get:

1. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression)

```
Accuracy 92.3076923076923
[[22  8]
 [ 1 86]]
             precision    recall  f1-score   support

         0.0       0.96      0.73      0.83        30
         1.0       0.91      0.99      0.95        87

    accuracy                           0.92       117
   macro avg       0.94      0.86      0.89       117
weighted avg       0.93      0.92      0.92       117
```

2. Decision Tree Clasifier

```
Accuracy 70.94017094017094
[[21  9]
 [25 62]]
             precision    recall  f1-score   support

         0.0       0.46      0.70      0.55        30
         1.0       0.87      0.71      0.78        87

    accuracy                           0.71       117
   macro avg       0.66      0.71      0.67       117
weighted avg       0.77      0.71      0.73       117
```

3. Random Forest Classifier

```
Accuracy 87.17948717948718
[[23  7]
 [ 8 79]]
             precision    recall  f1-score   support

         0.0       0.74      0.77      0.75        30
         1.0       0.92      0.91      0.91        87

    accuracy                           0.87       117
   macro avg       0.83      0.84      0.83       117
weighted avg       0.87      0.87      0.87       117
```

4. Support Vector Classifier

```
Accuracy 91.45299145299145
[[22  8]
 [ 2 85]]
              precision    recall  f1-score   support

         0.0       0.92      0.73      0.81        30
         1.0       0.91      0.98      0.94        87

    accuracy                           0.91       117
   macro avg       0.92      0.86      0.88       117
weighted avg       0.91      0.91      0.91       117
```

Here we see that the best classifier is Logistic Regression with 92% accuracy.

d. Cross validation

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

We get the below cv scores.

Cross validation score of Logistic Regression Model is 0.8144857058650162
Cross validation score of Random Forest Classifier is 0.79043619216033
Cross validation score of Decision Tree Classifier is 0.7131447096964338
Cross validation score of Support Vector Classifier is 0.8162098437960508

We see that the best accuracy is given by Logistic Regression of 92% and cross validation score as 81% which may be because the data is slightly overfitted, also we have little data to test and train.

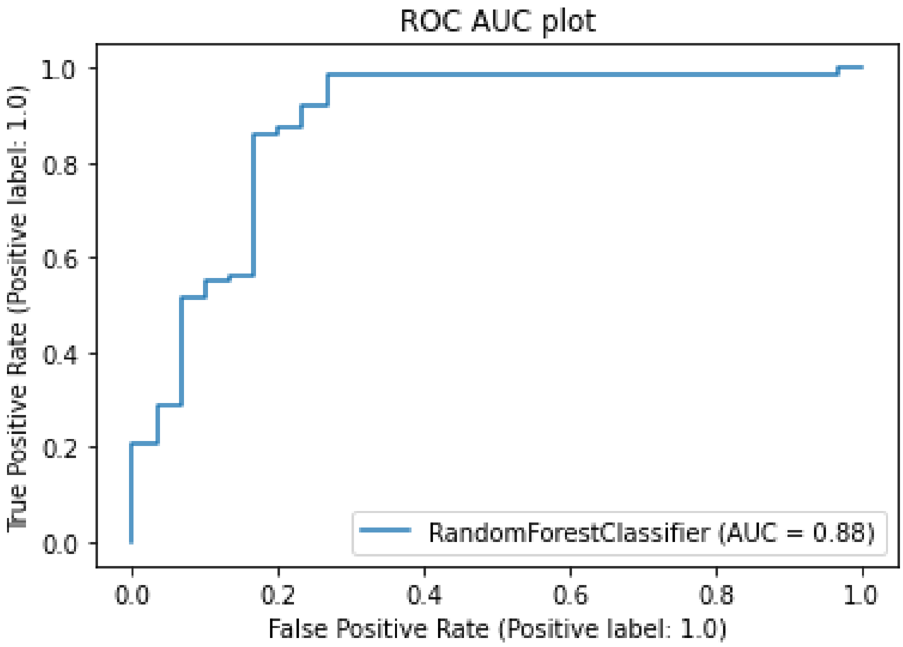
e. Hyper parameter Testing

In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

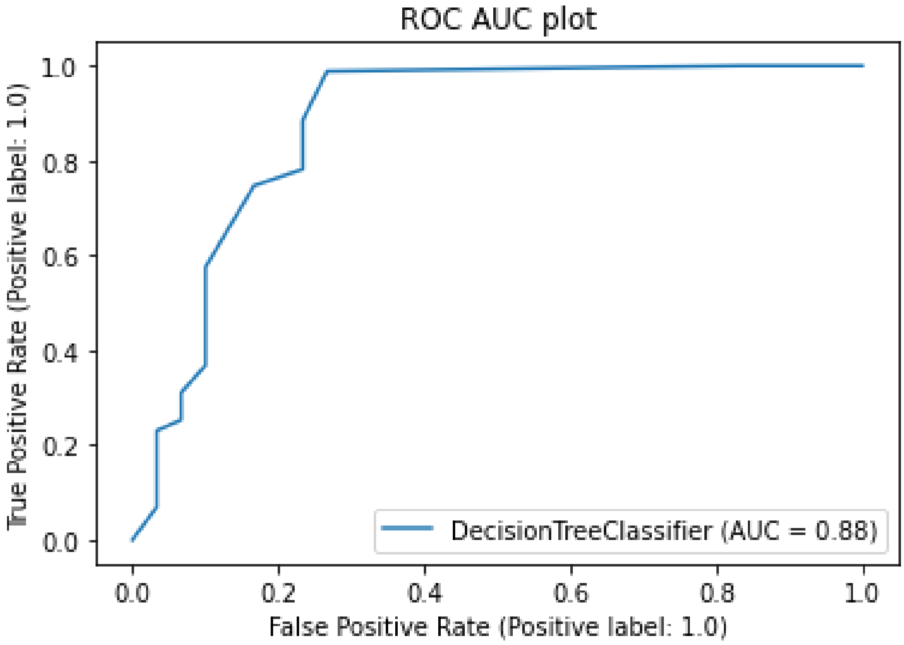
There are two types of search used for hyper parameter testing: Research and RandomizedSearch CV.

With small data sets and lots of resources, Grid Search will produce accurate results. However, with large data sets, the high dimensions will significantly slow down computation time and be very expensive. In this instance, it is advised to use Randomized Search. Thus, here we have used Grid Search CV as the data is too little thus better accuracy.

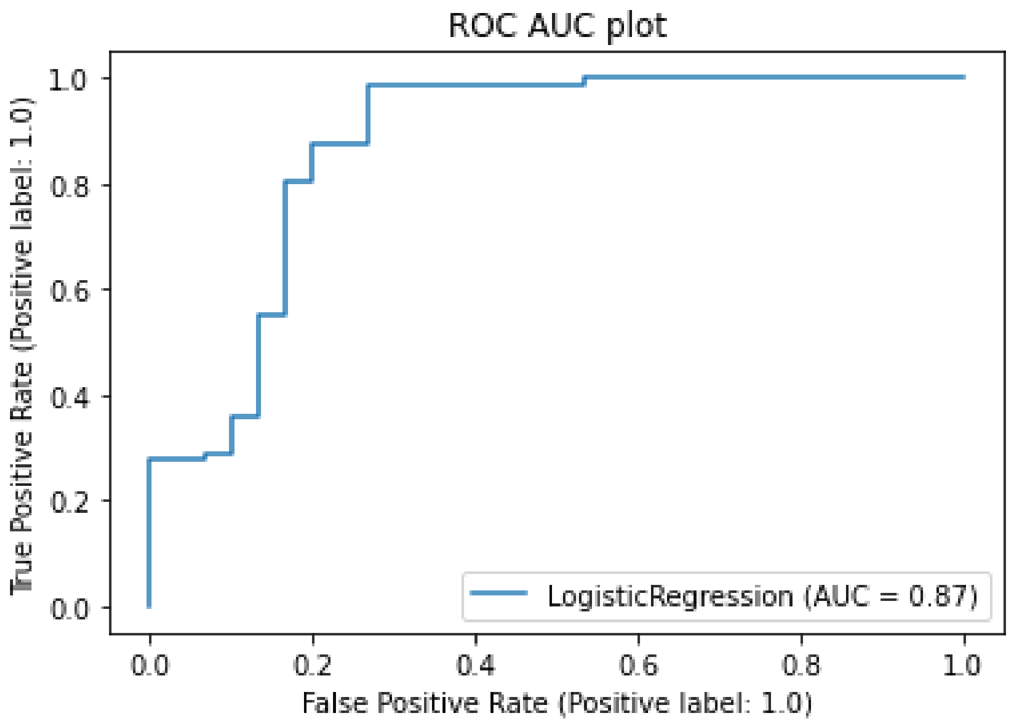
- Used GridSearchCV for RandomForestClassifier- Accuracy is 92.3%



- Used GridSearchCV for DecisionTreeClassifier- Accuracy is 92.3%



- Used GridSearchCV for Logistic Regression- Accuracy is 92.3%



Model accuracy is 92.3% through Logistic Regression and using hyper parameter testing also the accuracy is same. But for Random Forest Classifier and Decision Tree Classifier the accuracy has increased and is same 92.3%. AUC accuracy is 87% using the Logistic Regression and 88 % using the Random Forest Classifier and Decision Tree Classifier

Concluding Remarks

We did Exploratory information Investigation on the highlights of this dataset and saw how each include is distributed.

We dissected each variable to check in the event that information is cleaned and ordinarily distributed. We cleaned the information and evacuated NA values. We tried to find the correlation and based on the outcomes, we accepted whether or not there's a relation between the Loan Application Status and the other factors, we saw that the Credit History is the most important factor positively related followed by the Marital status, which means the higher the credit history the higher chances of loan application to get rejected. and the Education is negatively related which means the higher the candidate’s education the lower chances of his/her loan application to get rejected. We used the logistic regression model to classify the loan application status of whether they will be approved or not or whether the loan will be given or not depending on the various factors.

We got a very good model with a 92.3% accuracy which is very good. Just a disadvantage that the data available to train the model is too less and thus needs to be improved.

Thus, this model will help the banks to identify the loan defaulters. But to improve the model accuracy and the cross-validation score, we need to have more and more data to train and test and then we can move to conclusion.