**Machine learning project in python to predict loan approval**

We have the dataset with the loan applicants data and whether the application was approved or not. we will build a machine learning model to predict the loan approval probability.

## Steps involved in this machine learning project:

Following are the steps involved in creating a well-defined ML project:

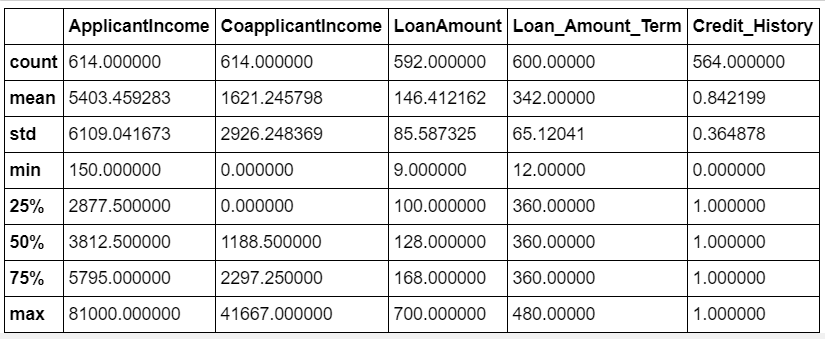
* Understand and define the problem
* Analyze and prepare the data
* Apply the algorithms
* Reduce the errors
* Predict the result

## Our Project : Predict if the loan application will get approved

We have the loan application information like the applicant's name, personal details, financial information and requested loan amount and related details and the outcome (whether the application was approved or rejected). Based on this we are going to train a model and predict if a loan will get approved or not.

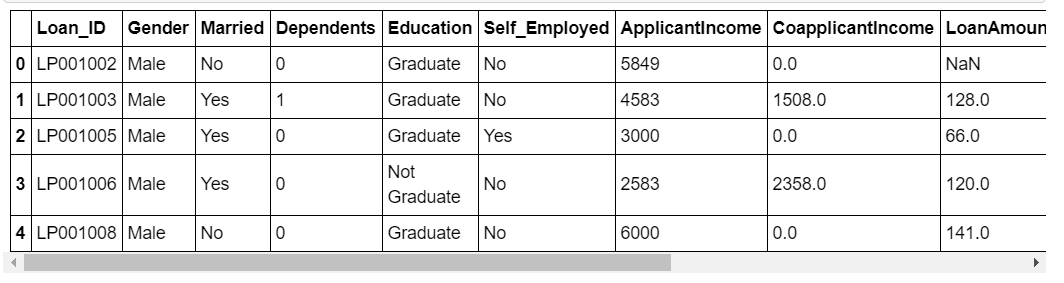
**Describe**

Pandas**dataframe.info()** function is used to get a concise summary of the dataframe. It comes really handy when doing exploratory analysis of the data. To get a quick overview of the dataset we use the dataframe.info() function.



**Head**

Pandas head() method is used to return top n (5 by default) rows of a data frame or series.



**Value\_counts**

Pandas**Index.value\_counts()**function returns object containing counts of unique values. The resulting object will be in descending order so that the first element is the most frequently-occurring element. Excludes NA values by default.

**Result:**

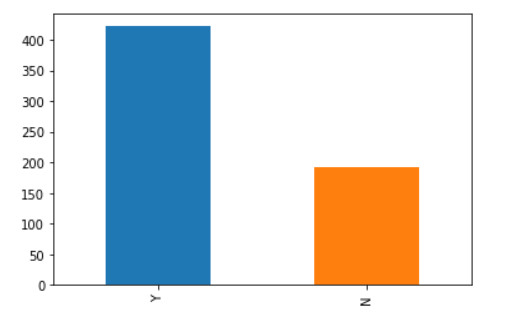
Y 422

N 192

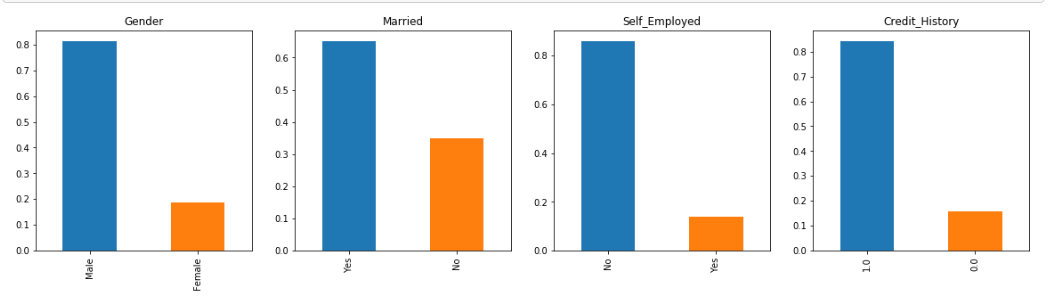
Name: Loan\_Status, dtype: int64

**Plot.bar**

Pandas **DataFrame.plot.bar()** plots the graph vertically in form of rectangular bars



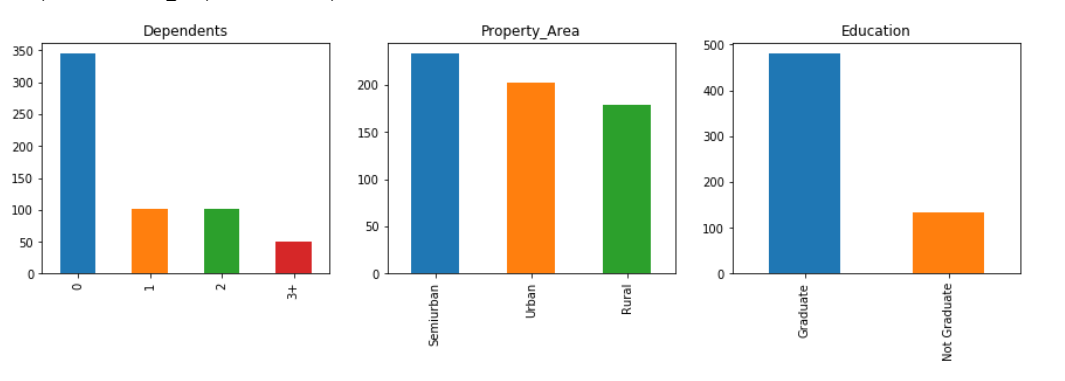
#### Visualizing Categorical variable



It can be inferred from the above bar plots that:

* 80% applicants in the dataset are male.
* Around 65% of the applicants in the dataset are married.
* Around 15% applicants in the dataset are self employed.
* Around 85% applicants have repaid their debts.

#### Visualizing Ordinal variable



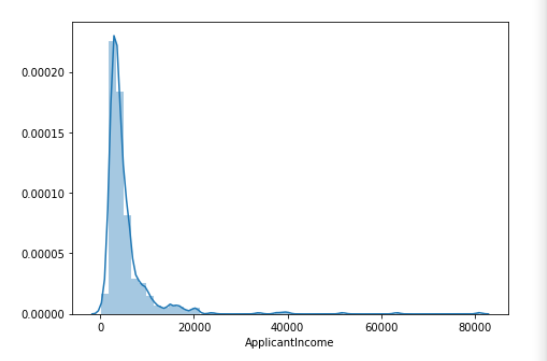
**Following inferences can be made from the above bar plots:**

* Most of the applicants don’t have any dependents.
* Around 80% of the applicants are Graduate.
* Most of the applicants are from Semiurban area.

#### Visualizing Numerical Variable

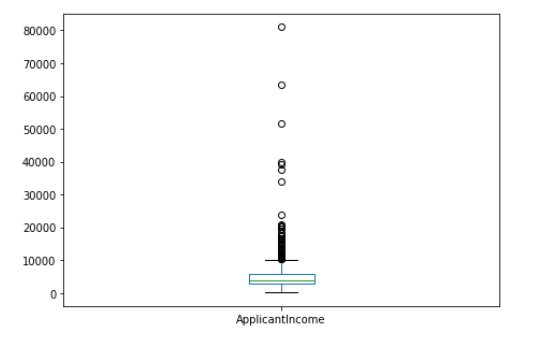
**Distplot**

# **This function combines the matplotlib hist function (with automatic calculation of a good default bin size) with the seaborn kdeplot() and rugplot() functions. It can also fit scipy.stats distributions and plot the estimated PDF over the data.**



**Box Plots**

 Boxplots are a measure of how well distributed the data in a data set is. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile and third quartile in the data set.



## Splitting the Data set

As we have seen already, In Machine learning we have two kinds of datasets

* Training dataset - used to train our model
* Testing dataset - used to test if our model is making accurate predictions

Our dataset has 480 records. We are going to use 80% of it for training the model and 20% of the records to evaluate our model.

Though our dataset has lot of columns, we are only going to use the Income fields, loan amount, loan duration and credit history fields to train our model.

## Evaluating the model and training the Model

We are going to apply the below four algorithms to this problem and evaluate its effectiveness. And finally choose the best algorithm and train it.

* **Logistic Regression :**
* Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables

Result: 0.7708333333333334

**Decision tree :**

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables.

Result: 0.6458333333333334

**Random forest :**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Result: 0.7604166666666666

So, Regression algorithm works fine for our use case. We can change / play around with the variables that we used for training the model till we get better accuracy.

## Summary

We built an end-to-end project and tested different algorithms in this Project . This concludes this project on machine learning. The course gave us a good primer to the machine learning concepts and boosted our overall confidence with machine learning.