# PROJECT REPORT OF

**REAL ESTATE-**

**HOUSE PRICE PREDICTION**

**SUBMITTED BY:**

**Himanshu Gupta**

### Abstract

Various Regression techniques are used in this project to predict the House Price based on several features. But before applying those techniques firstly the data is cleaned up by detecting outliers (using Scatter Plot), removing redundant features and extracting cities name from different addresses. After cleaning the data Exploratory Data Analysis has been done for visualization and checked multicollinearity in our data set and fitted different regression technique like Multiple Linear Regression, Ridge & Lasso Regression, PCA, and finally accuracy of each model are checked.

### Introduction

The problem falls under the category of supervised learning algorithms. The dataset is downloaded from Kaggle website. The dataset comprises 12 input features and one target feature. The input features include features that may or may not impact the price. My goal for this project is to build an end to end solution or application that is capable of predicting the house prices better than individuals.

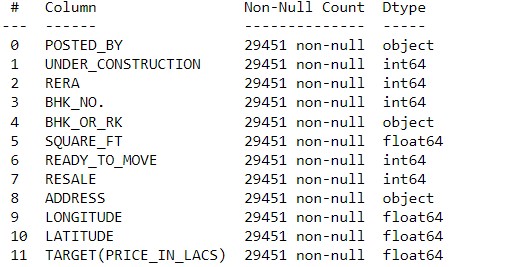
We will be building models to predict house prices of different cities which are located in all over the India. Dataset consist of different features like number of BHK, square feet, under construction, ready to move, resale, longitude, latitude, price etc.

We know that this is supervised learning problem as our data set consists of labelled observations and it does looks like multivariate regression should be our got to option but we will explore multiple ways of building the model and finally pick the one with lowest error rate RMSE (Root Mean Square Error) or MAE (Mean Absolute Error)

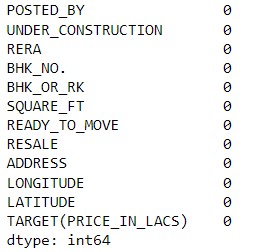
### Data Information

### This dataset is downloaded from kaggle website which consist a total of non-null 29451 observations with 12 features. Dataset looks like

Description of dataset is given below-



Missing Value Information of dataset -



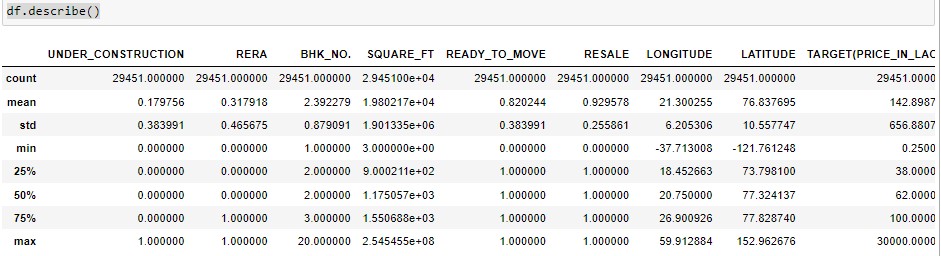
We can see that in this dataset there is no missing values are available.

### Objective

Objective of this project is to **predict the House price** based on number of BHK, square feet, under construction, ready to move, resale, longitude, latitude, posted by, BHK or RK features. We will identify the features which are most important in diagnosis.

### Data Pre-processing

### Statistical summary of dataset :-



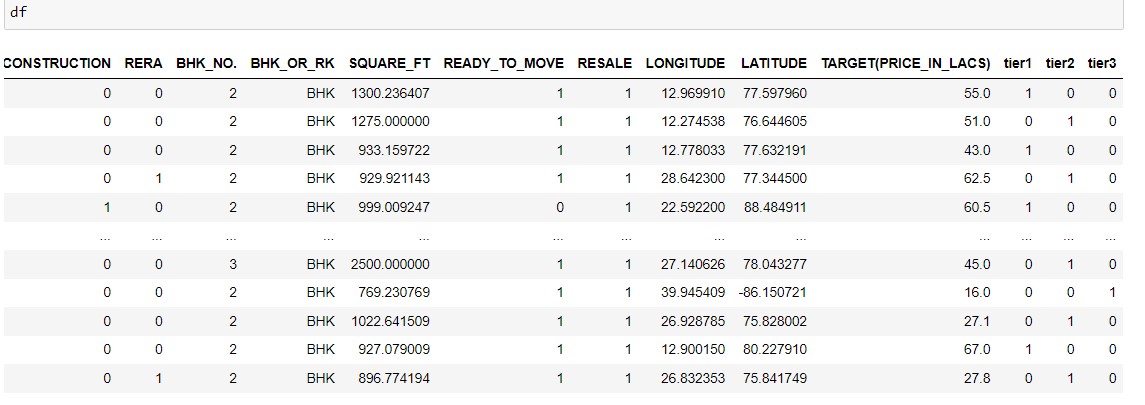
**Extraction of city name from address and combined into 3 tier:**

tier1= "Ahmedabad, Bangalore, Chennai, Delhi, Hyderabad, Kolkata, Mumbai ,Pune"

tier2= "Agra, Ajmer, Aligarh, Amravati, Amritsar, Asansol, Aurangabad, Bareilly, Belgaum… Vijayawada, Visakhapatnam, Vellore , Warangal"

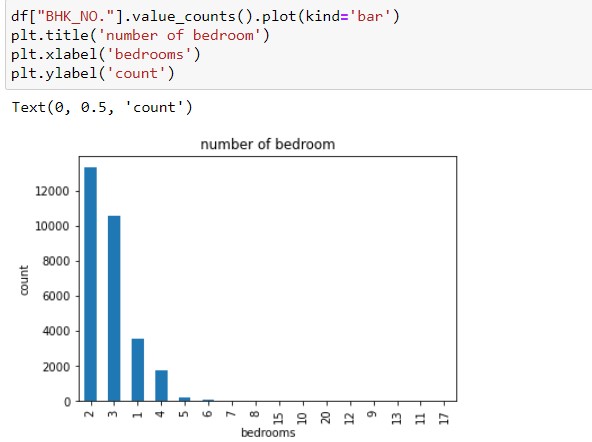
tier3= all remaining cities

#Resource- https://en.wikipedia.org/wiki/Classification\_of\_Indian \_cities



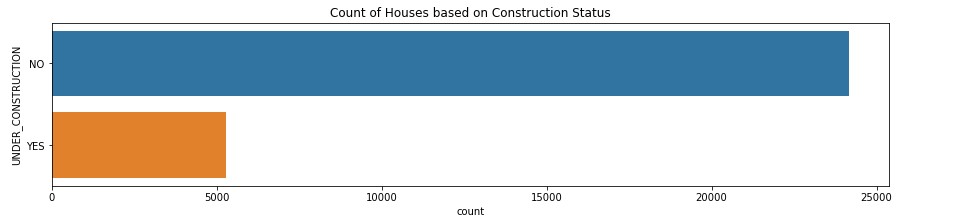
**Basic Exploratory Data Analysis**

In EDA we graphically visualize the different features.



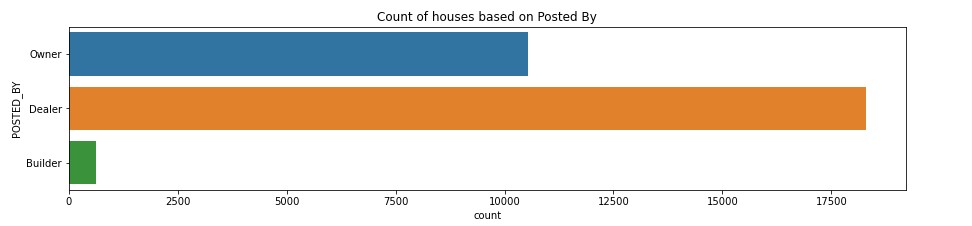
Maximum houses belong to in the range of 2-4 number of rooms.

**Count of houses based on construction:**



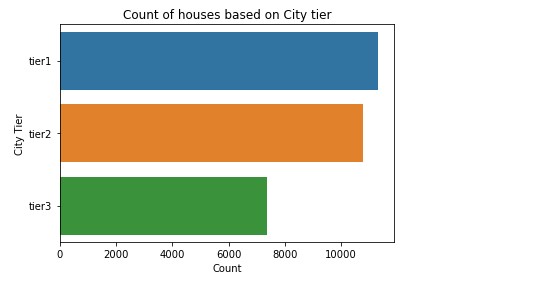
Maximum houses are in under construction.

**Count of houses based on Posted by:**

****

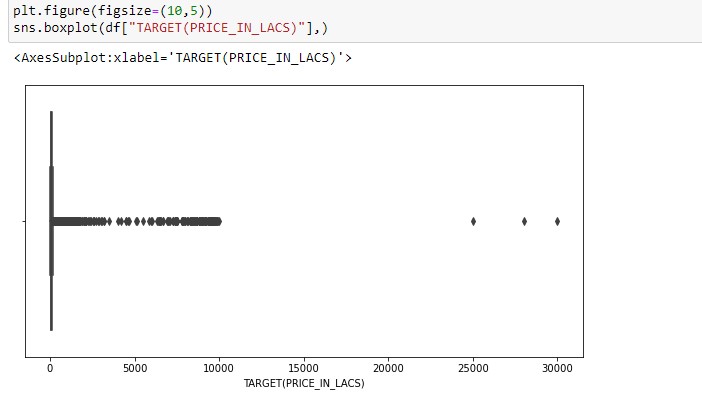
Maximum houses are posted by dealer**.**

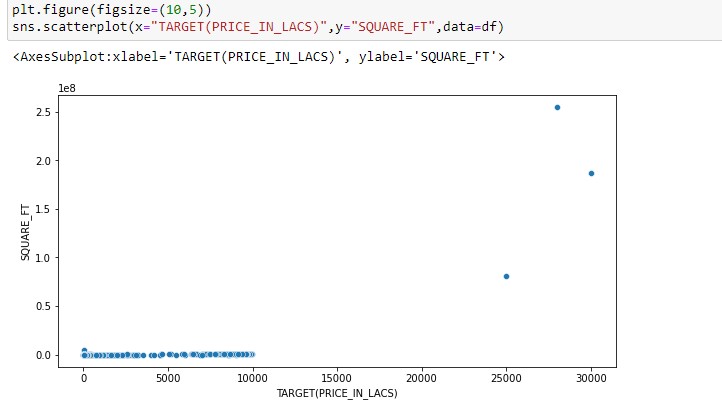
**Count of houses tier wise:**

****

Maximum houses belong to tier1 cities.

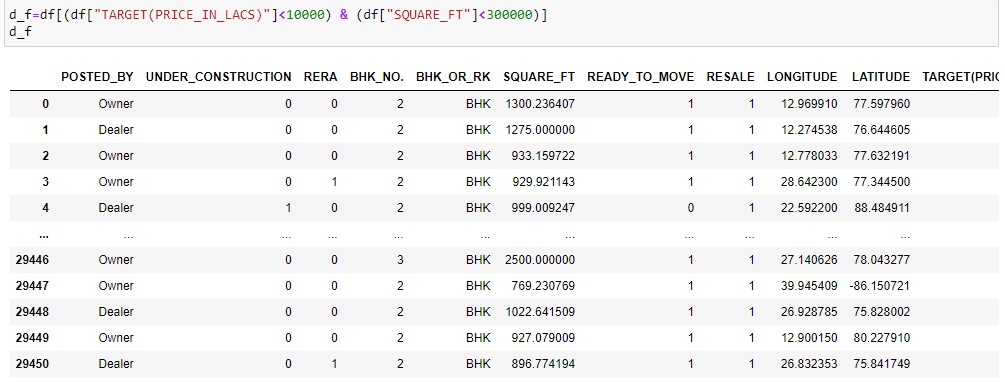
**Detection of outliers:**

****

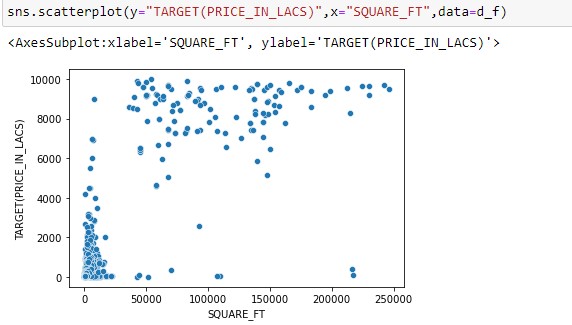
****

From above boxplot and scatter plot we can easily remove outliers from the data set.

**Cleaned Data:**

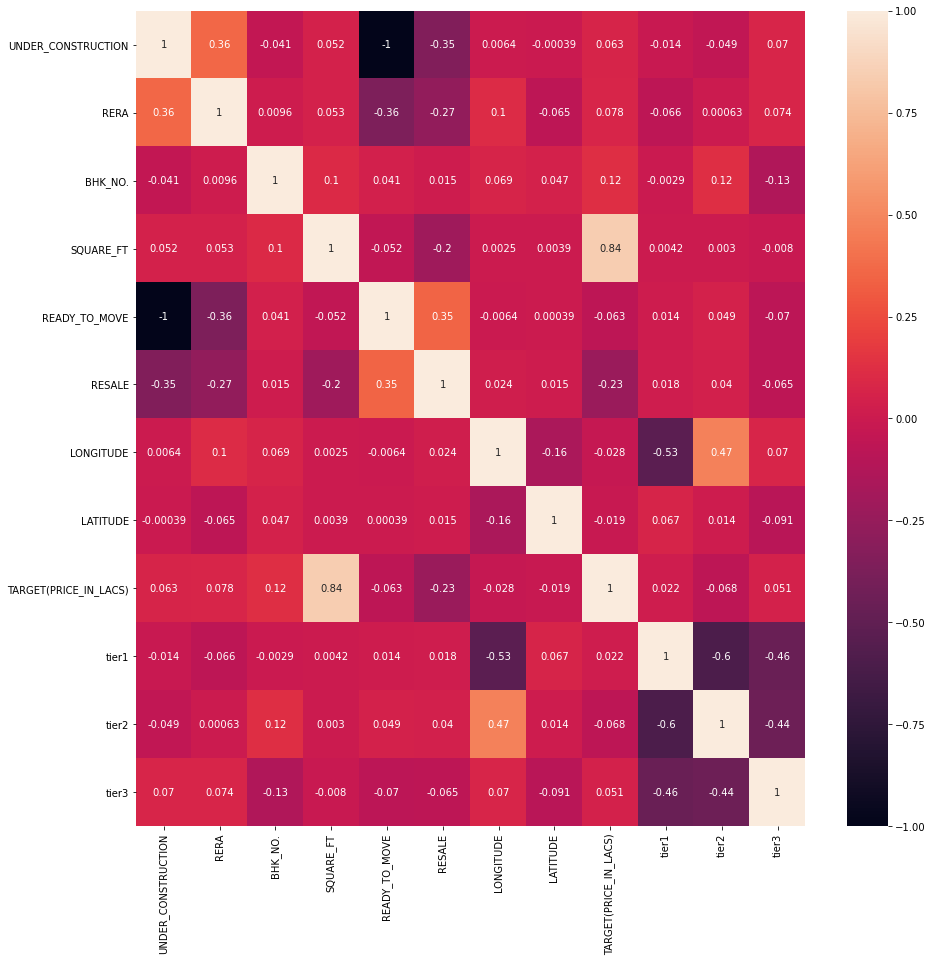
****

**We can verify our cleaned data set from below scatter plot.**

****

**Correlation coefficient:**

To calculate correlation coefficient between different pairs of features we used Heat map to graphically visualize:



From above heat map we can easily see that “Under construction” and “ready to move” features are strongly negatively correlated (i.e. -1) to each other so we can drop any one of them. So we have dropped “Under construction” feature from the data set.



**To create Dummy variable:**

In our data set “Posted by” and “BHK\_RK” has categorical feature so we convert these features to numerical feature using One hot Encoding.



Now the final shape of the data set is



Separate our regressor variables and target variable from the data set:



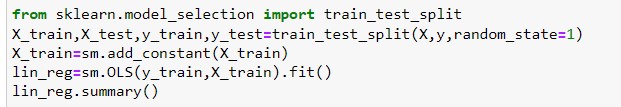
**Fitting regression model**

**Multiple regression model:**

[Multiple regression](https://www.sciencedirect.com/topics/economics-econometrics-and-finance/multiple-regression) is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of [multiple regression](https://www.sciencedirect.com/topics/economics-econometrics-and-finance/multiple-regression) analysis is to use the independent variables whose values are known to predict the value of the single dependent value. Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction.

****

Here Y is the dependent variable, and X1,…,Xn are the n independent variables. In calculating the weights, a, b1,…,bn, regression analysis ensures maximal prediction of the dependent variable from the set of independent variables. This is usually done by least squares estimation.





Accuracy of the model:



Calculate RMSE:

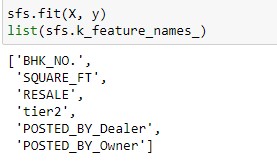


**Feature Selection:**

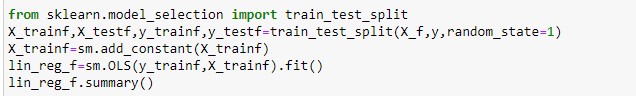
**1.Forward Selection Technique:**

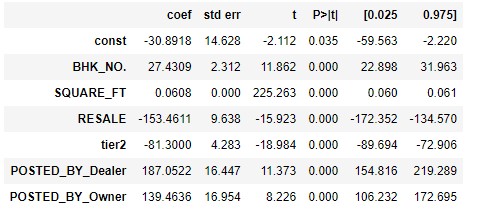
**Forward selection** is a type of [stepwise regression](https://www.statisticshowto.com/stepwise-regression/) which begins with an empty model and adds in [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/)one by one. In each forward step, you add the one variable that gives the single best improvement to your model.We do not delete the already added feature. in every iteration, we add only those features which increase the overall model fit.

**Extract features by Forward selection:**

****

Model building with selected features (Forward Selection):



****

Accuracy of the model:

****

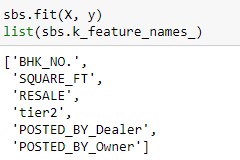
Calculate RMSE:

****

**2.Backward Elimination Technique:**

In backward elimination in the first step we include all predictors and in subsequent steps, keep on removing the one which has the highest p-value (>.05 the threshold limit). after a few iterations, it will produce the final set of features which are enough significant to predict the outcome with the desired accuracy.

**Extract features by Forward selection:**

****

Model building with selected features (Backward Elimination):



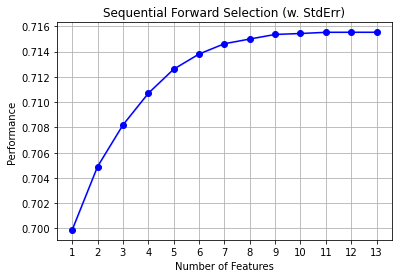
Accuracy of the model:

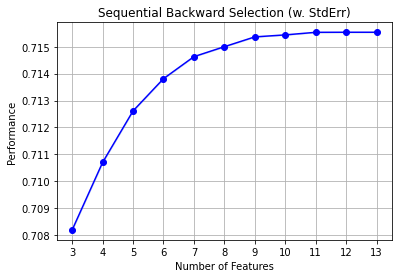


Calculate RMSE:

****

Graphical visualization of Forward and Backward selection technique:





**Detection of Multicollinearity:**

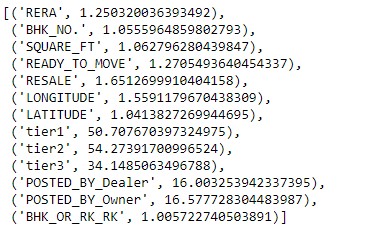
Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a [regression model](https://courses.analyticsvidhya.com/courses/Fundamentals-of-Regression-Analysis?utm_source=blog&utm_medium=what-is-multicollinearity). For example, height and weight, household income and water consumption, mileage and price of a car, study time and performance of students etc.

**VIF (Variance Inflation Factor)** score of an independent variable represents how well the variable is explained by other independent variables. R^2 value is determined to find out how well an independent variable is described by the other independent variables. A high value of R^2 means that the variable is highly correlated with the other variables. This is captured by the VIF which is denoted below:

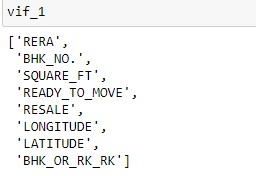
**[VIF formula](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/03/multicollinearity_VIF.png)**

So, the closer the **R^2** value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable.

Several Features and corresponding VIF Value:

****

Selecting columns whose VIF values <10:



Model building after VIF:



Accuracy of the model:



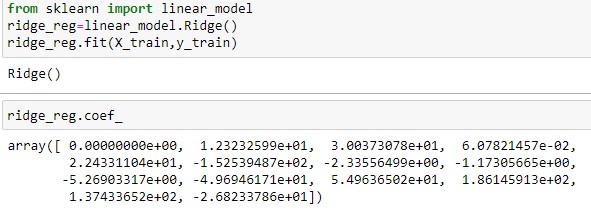
Calculate RMSE:



**Ridge Regression**

Ridge regression is a method of estimating the [coefficients](https://en.wikipedia.org/wiki/Coefficient) of multiple-[regression models](https://en.wikipedia.org/wiki/Regression_model) in scenarios where linearly independent variables are highly correlated. Ridge regression was developed as a possible solution to the imprecision of least square estimators when linear regression models have some multicollinear (highly correlated) independent variables—by creating a ridge regression estimator.

Model fitting:



Accuracy of the model:



Calculate RMSE:



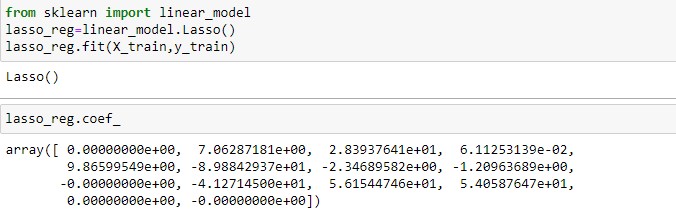
**Lasso Regression**

The word “LASSO” stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. It is a statistical formula for the regularisation of data models and feature selection.

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection or parameter elimination.

The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.

Model fitting:



Accuracy of the model:



Calculate RMSE:

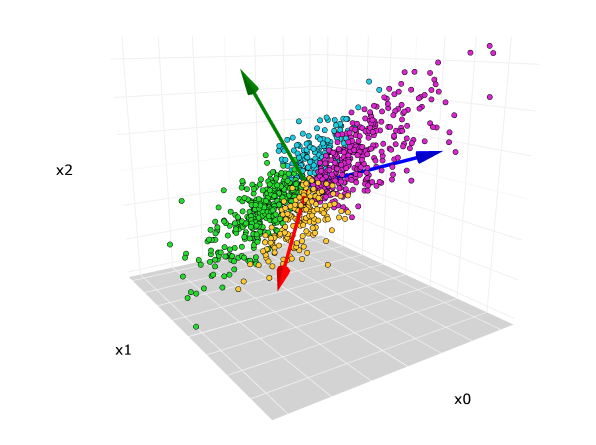


**Principal Component Analysis**

Principal component analysis (PCA) is a technique that transforms high-dimensions data into lower-dimensions while retaining as much information as possible.Principal component analysis (PCA) is the process of computing the principal components and using them to perform a [change of basis](https://en.wikipedia.org/wiki/Change_of_basis) on the data, sometimes using only the first few principal components and ignoring the rest.

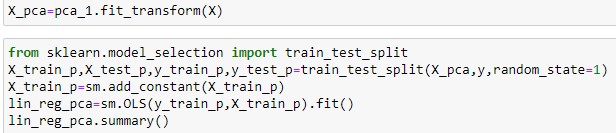
PCA is used in [exploratory data analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis) and for making [predictive models](https://en.wikipedia.org/wiki/Predictive_modeling). It is commonly used for [dimensionality reduction](https://en.wikipedia.org/wiki/Dimensionality_reduction) by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible.

PCA is extremely useful when working with data sets that have a lot of features and these features are correlated to each other i.e. there is multicollinearity present in the data set.



Model fitting:







Accuracy of the model:

****

Calculate RMSE:

****

**Conclusion:**

After applying the different regression models, we have got below accuracies with different models:

1.Multiple Linear Regression (Using all features)- 0 .724605805

2.Multiple Linear Regression (Using feature selection)-

* + 1. Forward Selection Technique- 0.722369118
    2. Backward Elimination Technique- 0.722369118

3.Ridge Regression- 0.724605735

4.Lasso Regression- 0.723602351

5.Principal Component Regression- 0.724605805

Every regression model shows accuracy more than 72%, We can conclude that we are able to build strong models with respect to our data. Overall, we can say that our models are pretty good.

**References:**

UCI machine Learning Respiratory,

Gareth, J., et al. (2014). Introduction to Statistical Learning.