**Data Science And Technologies**

**Section: W2**

**Project Report**

**Project Name: Real Estate Evaluation(Regression)**

**Instructor: Mazhar Javed**

**Semester: Fall 2019**

**Members: Muhammad Faizan Subhani 15026020016**

**Aamina Khan 150260297**

**Warda Murtaza 15026020336**

**Abstract**

In this project we have predicted house price according to the data given in the dataset. According to the data I’ve received, there are 414 observations. The dataset is composed by 7 columns and 414 rows. All the explanatory variables are continuous as well as the dependent variable. And our goal is to build a regression model to predict the average of Y(house price) by a given X(X1,X2,X3,X4,X5,X6). When we begin are regression analysis, we found the correlation among different factors and finally begin to find the predicted values.

And here I have 6 independent variable (from X1 to X6) and one dependent variable(Y).

**Hypothesis**

The problem of predicting house price from the dataset was a regression problem because we have to predict continues values in it. If we had to predict labels from our dataset, then it would have been a classification problem.

Regression analysis is a powerful statistical method that allows you to examine the relation between two or more variable of interest. In order to conduct a regression analysis, you’ll need to define a dependent variable that you hypothesize is being influenced by one or several independent variables.

You’ll need to establish a comprehensive dataset to work with.

**Applications for Good result**

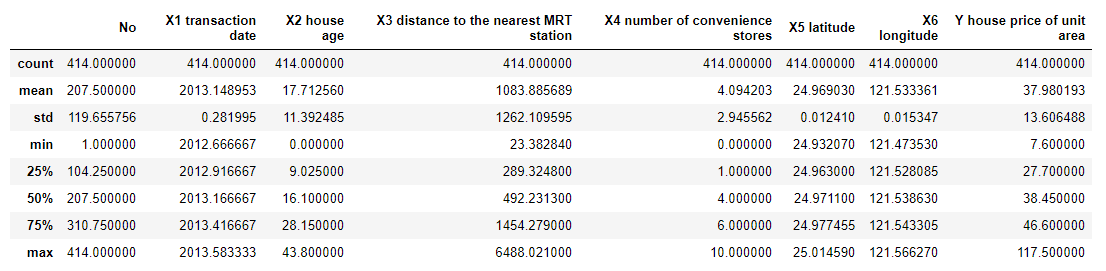
We have used different regression models like Linear regression, Lasso regression, Gradient Boosting, Random Forest regression, Ridge Regression and Decision Tree Regression for our analysis and to compare results so that we can find the best suitable model for our dataset.

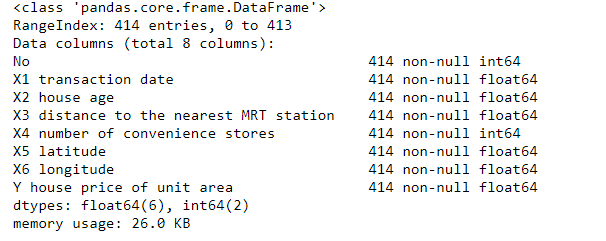
**Analysis**

Firstly, I separated all the data Y into training set (70%) and test set(30%). Trained and fitted the data with different regression models like linear regression, lasso regression, ridge regression, Support Vector Regression, KNN, Decision Tree Gradient Boosting and Random Forest.

**Step 1 check description and info of our dataset.**

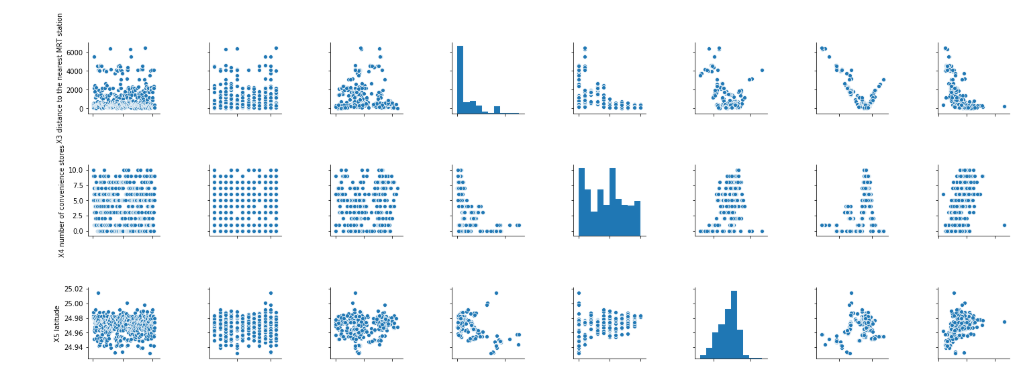
Results

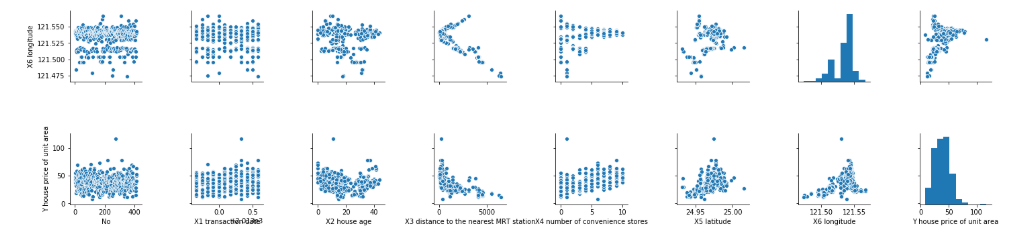




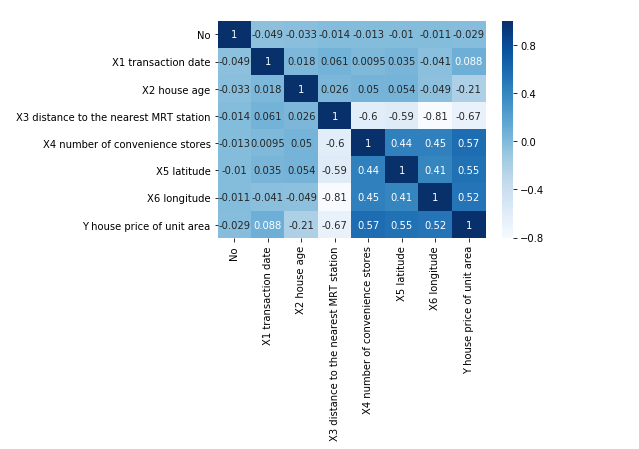
**Step 2 PairPlots to check relationship between different attributes**

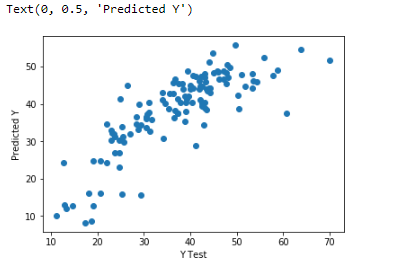




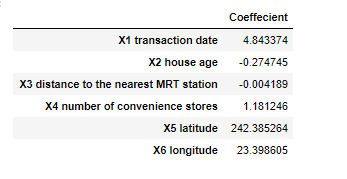


**Step 3 HeatMap**

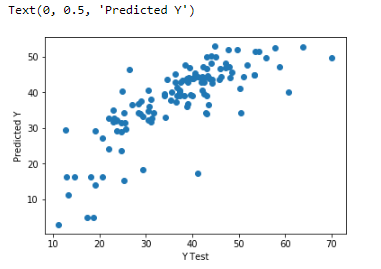


**Step 4 Linear Regression Model (Result)**

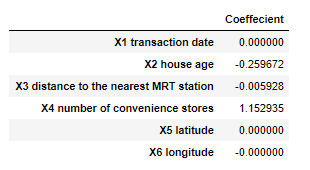




**Step 5 Lasso Regression (Results)**







**Step 6 Random Forest Regression(Results)**

Best cross validation score: 0.61

Best parameters: {'max\_depth': 14}

Test score: 0.7132790271508155

**Step 7 Gradient Boosting Regression(Results)**

Best cross validation score: 0.65

Best parameters: {'max\_depth': 1, 'n\_estimators': 100}

Test score: 0.7851603326730534

**Step 8 KNN(Results)**

Best cross validation score: 0.57

Best parameters: {'n\_neighbors': 8}

Test score: 0.666107852354969

**Step 9 Ridge Regression(Results)**

Best cross validation score: 0.54

Best parameters: {'alpha': 0.01}

Test score: 0.6538831712385077

**Step 10 Decision Tree(Results)**

Best cross validation score: 0.58

Best parameters: {'max\_depth': 2}

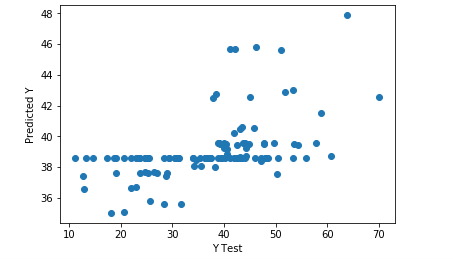
Test score: 0.6355645969097079

**Step 11 Support Vector(Results)**

score: 0.11244115186484294

Mean Squared Error: 117.49177069134497

Root Mean Squared Error: 10.83936209799013



**Conclusion and Comparison of Results**

We found out that our Linear Regression model gave us the best results with

**Mean Absolute Error: 5.392382800671546**

**Mean Squared Error: 46.21083888823276**

**Root Mean Squared Error: 6.797855462440545**

And Gradient Boosting has the highest cross-Validation score

**Best cross validation score: 0.65**

**Best parameters: {'max\_depth': 1, 'n\_estimators': 100}**

**Test score: 0.7851603326730534**