Regression Analysis Project

Statistical Data Analysis and Prediction of cost of homes using different Regression models

(in Python)

Sanket Bhale

Contents

1. Introduction: 3
2. Methodology 3
3. Summarizing data: 4
4. Exploratory Data Analysis: 5
5. Modeling and Analysis by Ordinary Least Square (OLS) Linear Regression models: 9
6. Modeling and Analysis by Machine learning based Regression models: 12
7. Performance comparison of Machine learning based Regression models: 12

Conclusions: 14

1. Introduction:

The primary aim of this project is to analyze and predict cost of houses from the given dataset. The dataset is sourced from Kaggle.com and its primary objective is to help in implementing and practicing various statistical data analysis practices. The primary focus is on exploring various regression algorithms and analyzing their performance for the given dataset. The dataset is derived from the information collected by the U.S. Census Services concerning housing. The analysis is performed using Python language in Jupyter notebook.

Description of Dataset: The dataset consists of 506 observations and 13 variables(columns) with 12 being predictor variables and 1 response variable.
Predictor variables:

1. CRIM - per capita crime rate
2. ZN - proportion of residential land zoned for lots over
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five employment centers
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. LSTAT - % lower status of the population

Response variable:

13. MEDV - Median value of owner-occupied homes in $1000's

2. Methodology

1. **Summarizing dataset (Descriptive data analysis).**
   Import dataset, perform descriptive and 5-number summary to summarize distribution of data variables. Check for missing values.
2. **Pre-processing and Exploratory Data Analysis.**
   Check and explore for outliers.
   Remove outliers.
   Calculate correlation matrix.
3. **Modeling and Analysis by Ordinary Least Square Regression Models**
   Fit simple regression models based on Ordinary Least Squares.
   Analyze various statistics: ANOVA, R-squared, correlation, F-statistics, skewness and others.
   Build and analyzed 3 models through stepwise regression (Backward elimination).
4. **Modeling and Analysis by machine learning based regression models**
   Prepare data for training and testing the machine learning-regression models.
   Fit and cross validate regression models.
   a. *Simple Learning Regression model*
   b. *Polynomial Regression model*
   c. *Ridge Regression model*

   d. *Lasso Regression model*
   e. *Decision Regression Tree model*
   f. *Random Forest Regression*
5. **Compare performance of regression models.**
   Assess performance of models through visualizations and statistics.
6. **Conclusion**

3. Summarizing data:

The Descriptive Statistics and 5-number summary is shown in the diagram below to assess distribution of dataset.

	CRIM	ZN	INDUS	CHAS	NOX	RM
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500
75%	3.677082	12.500000	18.100000	0.000000	0.624000	6.623500
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000

	AGE	DIS	RAD	TAX	PTRATIO	LSTAT
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	68.574901	3.795043	9.549407	408.237154	18.455534	12.653063
std	28.148861	2.105710	8.707259	168.537116	2.164946	7.141062
min	2.900000	1.129600	1.000000	187.000000	12.600000	1.730000
25%	45.025000	2.100175	4.000000	279.000000	17.400000	6.950000
50%	77.500000	3.207450	5.000000	330.000000	19.050000	11.360000
75%	94.075000	5.188425	24.000000	666.000000	20.200000	16.955000
max	100.000000	12.126500	24.000000	711.000000	22.000000	37.970000

	MEDV
count	506.000000
mean	22.532806
std	9.197104
min	5.000000
25%	17.025000
50%	21.200000
75%	25.000000
max	50.000000

Diagram: Data Summary

The above diagram assists in concluding that all variables are numerical and there are **no abnormal entries** in any variables.

The next step is to **check if there are any missing values.**

```
#Check missing values
data.isnull().sum()
```

```
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
AGE        0
DIS        0
RAD        0
TAX        0
PTRATIO    0
LSTAT      0
MEDV       0
dtype: int64
```

Diagram: No missing values in the dataset.

4. Exploratory Data Analysis:

Analyze distribution of data through plotting Boxplots. **Check for outliers** as they may influence the regression line. Outliers need to be removed from data before doing any analysis. Outliers are calculated based on the Inter Quantile Range formula (IQR). Any observation of a variable that lies above and below 1.5*IQR of maximum value and minimum value of variable are said to be outliers. (IQR=Q3-Q1)

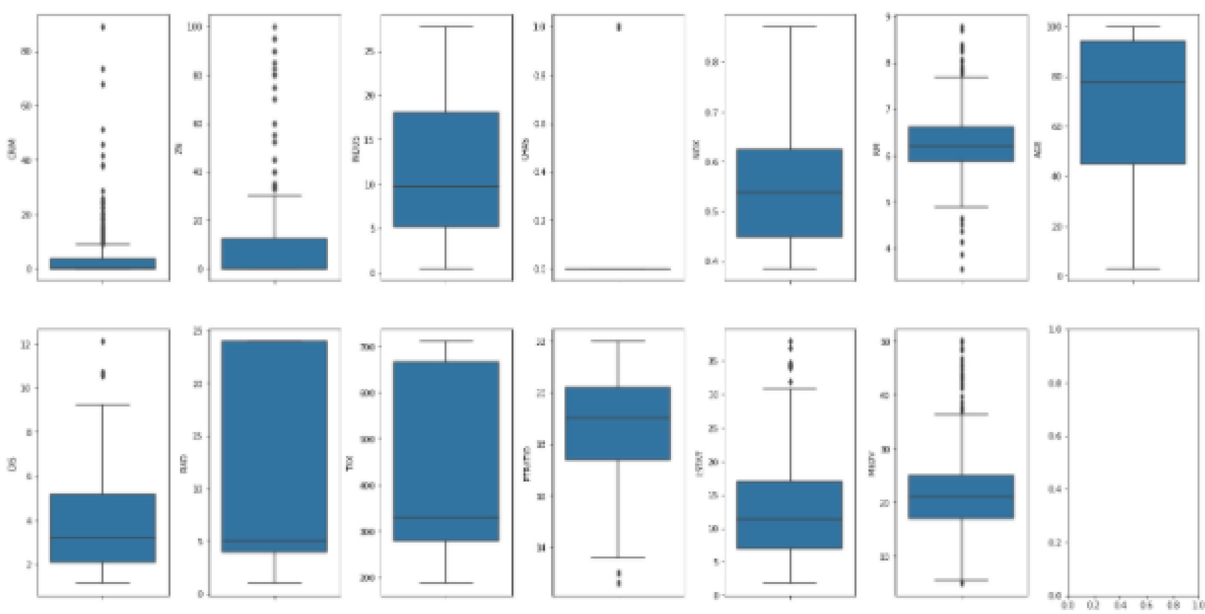


Diagram: Checking for outliers

As can be seen from the above diagram there are **outliers present in the CRIM, ZN, RM and MEDV columns**. Let's calculate percentage of outlier observations in the variables.

```
Column CRIM outliers = 13.04%
Column ZN outliers = 13.44%
Column INDUS outliers = 0.00%
Column CHAS outliers = 100.00%
Column NOX outliers = 0.00%
Column RM outliers = 5.93%
Column AGE outliers = 0.00%
Column DIS outliers = 0.99%
Column RAD outliers = 0.00%
Column TAX outliers = 0.00%
Column PTRATIO outliers = 2.96%
Column LSTAT outliers = 1.38%
Column MEDV outliers = 7.91%
```

Diagram: Percentage of outliers in the data for all variables

There are more than 10% outliers in the CRIM and ZN columns. Removing these outliers may result in loss of excess data. It is recommended to remove only limited number of observations from the data as it may highly affect the performance of the regression models. Therefore, **all observation corresponding to outliers from the MEDV columns are removed**. The data distribution is analyzed to further check presence of outliers.

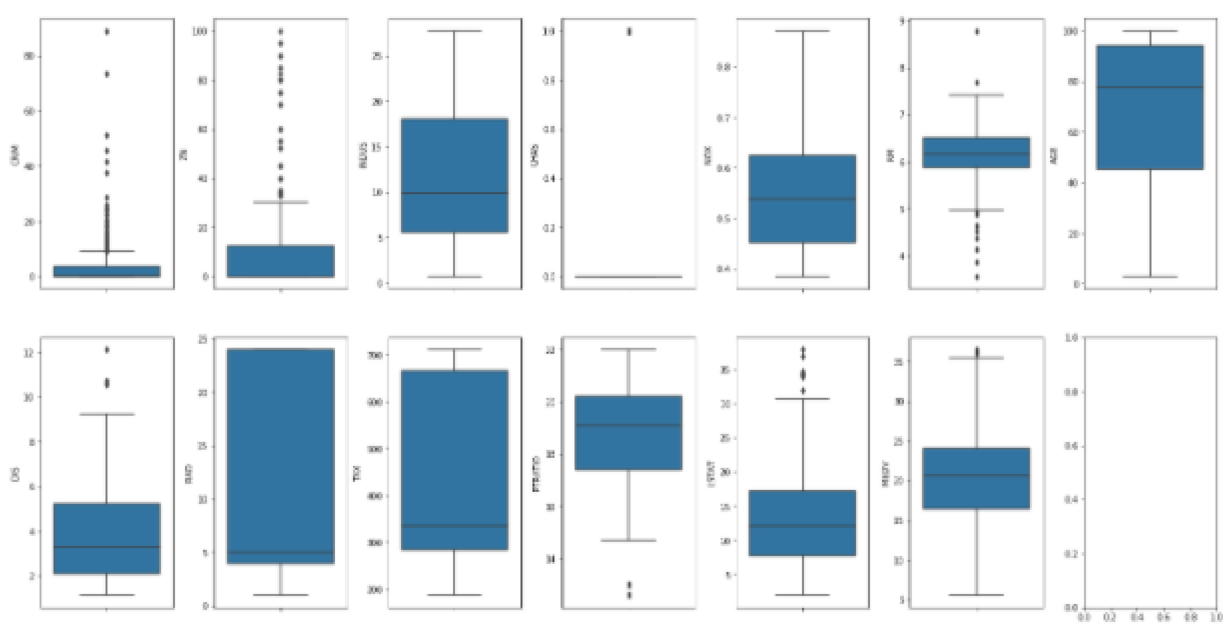


Diagram: Checking for outliers after removing outliers from the MEDV column.

```
Column CRIM outliers = 13.30%
Column ZN outliers = 12.88%
Column INDUS outliers = 0.00%
Column CHAS outliers = 100.00%
Column NOX outliers = 0.00%
Column RM outliers = 3.00%
Column AGE outliers = 0.00%
Column DIS outliers = 1.07%
Column RAD outliers = 0.00%
Column TAX outliers = 0.00%
Column PTRATIO outliers = 2.15%
Column LSTAT outliers = 1.50%
Column MEDV outliers = 1.29%
```

Diagram: Percentage of Outliers in variables.

After removing observations corresponding to outliers in MEDV column using IQR formula, we can still see that the percentage of outliers in CRIM and ZC column is still greater than 10%. Deleting all observations based on these outliers will result in loss of more than 15% of data (as we already deleted approx. 7 % data in the previous step). It is high recommended to carry our analysis with removing any more observations.

The next step is to calculate correlation matrix that will help us in understanding relations between different variables. The correlation matrix consists of values between 0 and 1. The closer the value to 1 , higher is the correlation. These values are based on Pearson's correlation coefficient. Highly corelated variables need to be treated before continuing analysis.

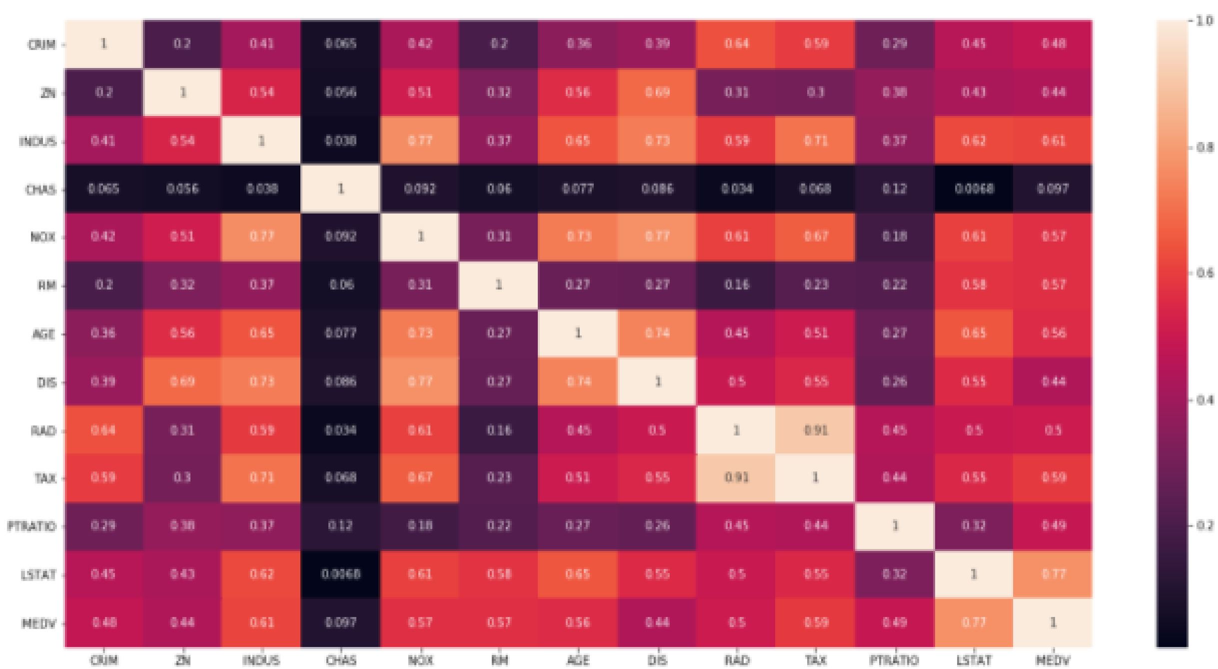


Diagram: Correlation Matrix

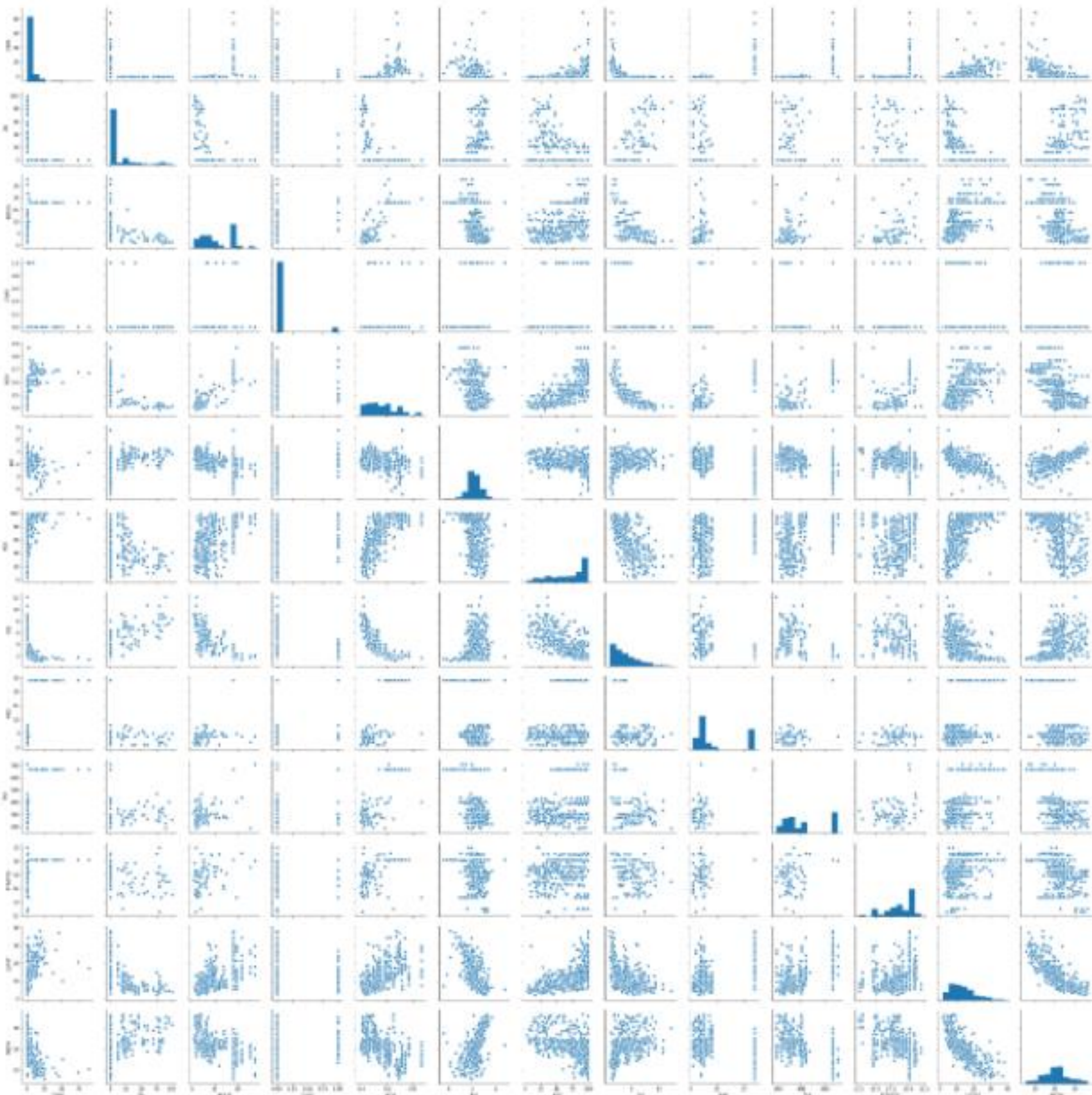


Diagram: Relation between predictor variables and response variables.

The above diagram shows value of response variable (MEDV) with each value of the variables. It shows the distribution and relation of the response variable to predictor variables.

5. Modeling and Analysis by Ordinary Least Square (OLS) Linear Regression models:

For modeling, we use stepwise regression method to fit OLS regression models. The initial model is built by fitting all predictor variables and response variable. The models are assessed based on various summary statistics like Rsq, adj. Rsq, p-values, F-statistics, t-statistics for each predictor variable and Durbin -Watson test statistics. We will basically assess the model based on Rsq adj value that explains how well the model can adjust and explain the variability. Then the predictor variables are assessed based on the p-values. The variables with p value>0.05 are ignored while fitting the next model. P-values less than the 0.05 signifies that the variable is insignificant and can be ignored while fitting next model. Each time a new model is built until p values for all variables are less than 0.05 or if the improvement Rsq adj value stops. This follows the backward elimination methodology of stepwise Regression.

5.a Initial model:

OLS Regression Results

Dep. Variable:	MEDV	**R-squared:**	0.971
Model:	OLS	**Adj. R-squared:**	0.970
Method:	Least Squares	**F-statistic:**	1261.
Date:	Mon, 18 Jan 2021	**Prob (F-statistic):**	0.00
Time:	02:52:15	**Log-Likelihood:**	-1271.2
No. Observations:	466	**AIC:**	2566.
Df Residuals:	454	**BIC:**	2616.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
CRIM	-0.1061	0.028	-3.754	0.000	-0.162	-0.051
ZN	0.0370	0.012	3.071	0.002	0.013	0.061
INDUS	-0.0798	0.050	-1.582	0.114	-0.179	0.019
CHAS	1.3300	0.756	1.760	0.079	-0.155	2.815
NOX	4.9350	2.601	1.897	0.058	-0.177	10.047
RM	4.7379	0.276	17.167	0.000	4.196	5.280
AGE	-0.0292	0.011	-2.683	0.008	-0.051	-0.008
DIS	-0.4093	0.156	-2.623	0.009	-0.716	-0.103
RAD	0.0338	0.051	0.664	0.507	-0.066	0.134
TAX	-0.0092	0.003	-3.037	0.003	-0.015	-0.003
PTRATIO	-0.0071	0.088	-0.081	0.935	-0.179	0.165
LSTAT	-0.2550	0.040	-6.298	0.000	-0.335	-0.175

Omnibus:	76.694	**Durbin-Watson:**	1.013
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	230.704
Skew:	0.765	**Prob(JB):**	8.00e-51

Diagram: Simple Linear regression analysis model statistics (with all variables), ANOVA

The initial model is fitted and built based on all the variables. As seen from the diagram, the p-values for PTRATIO, RAD, INDUS and few other variables is more than 0.05. Therefore, we will ignore the PTRATIO variable (as it has the highest p-value) while fitting the next model. Also, the model has high Rsq values(0.97) that fairly explains the variability.

5.b Second model:

OLS Regression Results

Dep. Variable:	MEDV	**R-squared:**	0.971
Model:	OLS	**Adj. R-squared:**	0.970
Method:	Least Squares	**F-statistic:**	1379.
Date:	Mon, 18 Jan 2021	**Prob (F-statistic):**	0.00
Time:	02:56:36	**Log-Likelihood:**	-1271.2
No. Observations:	466	**AIC:**	2564.
Df Residuals:	455	**BIC:**	2610.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CRIM	-0.1062	0.028	-3.759	0.000	-0.162	-0.051
ZN	0.0373	0.011	3.348	0.001	0.015	0.059
INDUS	-0.0803	0.050	-1.605	0.109	-0.179	0.018
CHAS	1.3359	0.751	1.778	0.076	-0.141	2.813
NOX	4.9388	2.598	1.901	0.058	-0.166	10.044
RM	4.7244	0.220	21.485	0.000	4.292	5.157
AGE	-0.0292	0.011	-2.693	0.007	-0.051	-0.008
DIS	-0.4140	0.144	-2.867	0.004	-0.698	-0.130
RAD	0.0339	0.051	0.667	0.505	-0.066	0.134
TAX	-0.0093	0.003	-3.097	0.002	-0.015	-0.003
LSTAT	-0.2559	0.039	-6.567	0.000	-0.332	-0.179

Omnibus:	76.151	**Durbin-Watson:**	1.012
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	226.952
Skew:	0.763	**Prob(JB):**	5.22e-50

Diagram: Simple Linear regression analysis model statistics (all variables except PTRATIO)

The new model is fitted and built based on all the variables excluding PTRATIO. As seen from the diagram, the p-values for RAD, INDUS and few other variables is more than 0.05. Therefore, we will ignore the RAD variable (as it has the highest p-value) while fitting the next model. Also, the model has high Rsq values(0.97) that is equal to performance of the initial model.

5.c Third model:

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.971
Model:	OLS	Adj. R-squared:	0.970
Method:	Least Squares	F-statistic:	1519.
Date:	Mon, 18 Jan 2021	Prob (F-statistic):	0.00
Time:	02:57:34	Log-Likelihood:	-1271.5
No. Observations:	466	AIC:	2563.
Df Residuals:	456	BIC:	2604.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CRIM	-0.1006	0.027	-3.731	0.000	-0.154	-0.048
ZN	0.0363	0.011	3.289	0.001	0.015	0.058
INDUS	-0.0886	0.048	-1.830	0.068	-0.184	0.007
CHAS	1.3861	0.747	1.855	0.064	-0.082	2.855
NOX	4.8960	2.595	1.886	0.060	-0.204	9.996
RM	4.7021	0.217	21.649	0.000	4.275	5.129
AGE	-0.0297	0.011	-2.744	0.006	-0.051	-0.008
DIS	-0.4255	0.143	-2.970	0.003	-0.707	-0.144
TAX	-0.0076	0.002	-4.374	0.000	-0.011	-0.004
LSTAT	-0.2572	0.039	-6.612	0.000	-0.334	-0.181

Omnibus:	77.818	Durbin-Watson:	1.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	229.871
Skew:	0.782	Prob(JB):	1.21e-50

Diagram: Simple Linear regression analysis model statistics (all variables except PTRATIO and RAD)

The third model is fitted and built based on all the variables excluding PTRATIO and RAD . As seen from the diagram, the p-values for INDUS and few other variables is more than 0.05.

This model similar to previous initial models has high and approximately identical Rsq values(0.97). This indicates that removing any variable further will not affect the performance of the model considerably. Therefore, we can use any model for our prediction purposes.

However, **the initial linear regression model should be preferred as it has all the original variables for predicting the response variable(MEDV) i.e the median cost of houses.**

6. Modeling and Analysis by Machine learning based Regression models:

The Regression models based on following algorithms are modeled and analyzed:

a. Simple Learning Regression model
b. Polynomial Regression model
c. Ridge Regression model
d. Lasso Regression model
e. Decision Regression Tree model
f. Random Forest Regression

Before fitting data models, the data is randomly split into training and testing datasets. The training dataset is 70% of the entire data and remaining 30% data is used for testing purposes.

Statistics like Coefficient of Variance (CV), R^2 scores for train and test data, and Mean Square Errors (RMSE) for each model are calculated. These statistics are used to compare and assess the performance of all models. The following diagram shows the comparison of these statistics for all the models.

	Model	RMSE	R2_Score(training)	R2_Score(test)	Cross-Validation
0	Linear Regression	2.959144	0.719575	0.808205	0.679218
1	Polynomial Regression (2nd)	2.971015	0.867566	0.806663	0.679218
2	Ridge Regression	2.558112	0.884665	0.856667	0.791249
3	Lasso Regression	2.549891	0.884339	0.857587	0.796092
4	Decision Tree Regression	3.240414	1.000000	0.770011	0.623908
5	Random Forest Regression	2.503292	0.974582	0.862745	0.799710

Diagram: Comparing statistics for different Regression models

7. Performance comparison of Machine learning based Regression models:

The performance of models is analyzed visually bases on statistics.

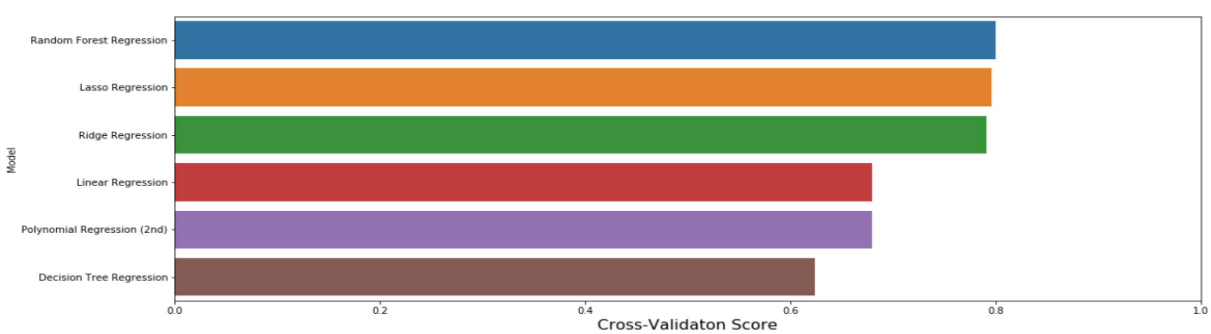
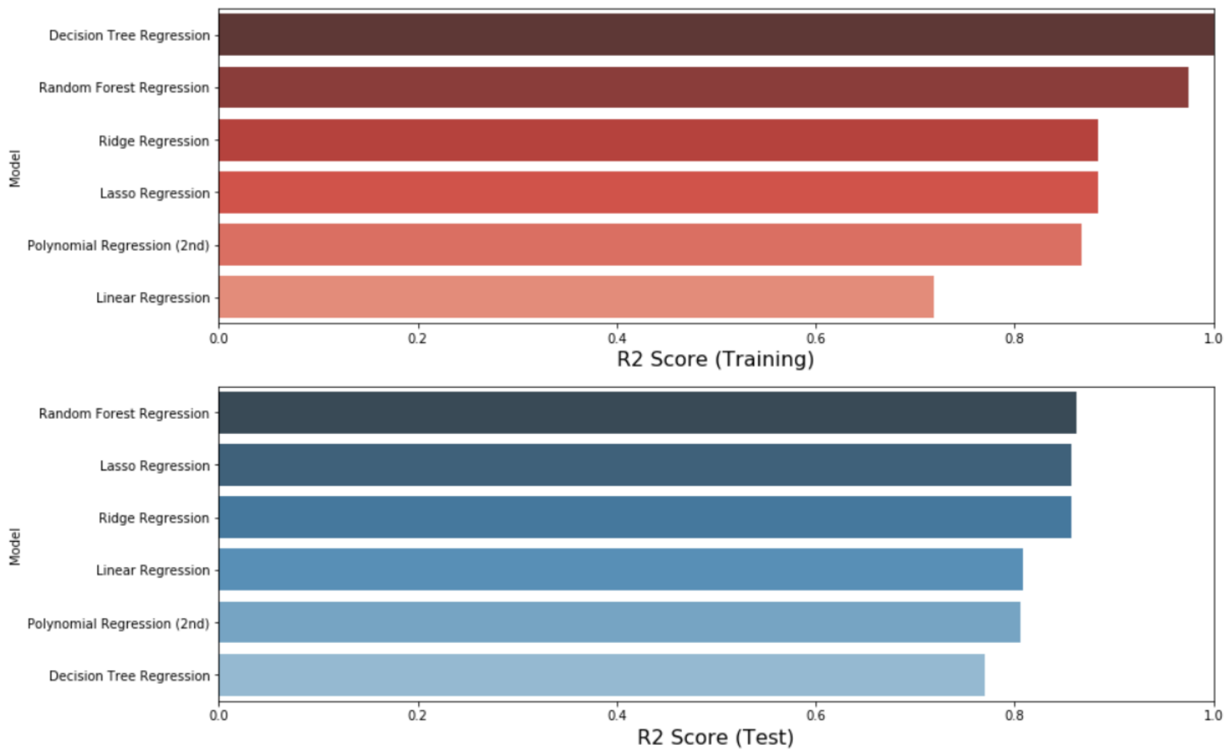
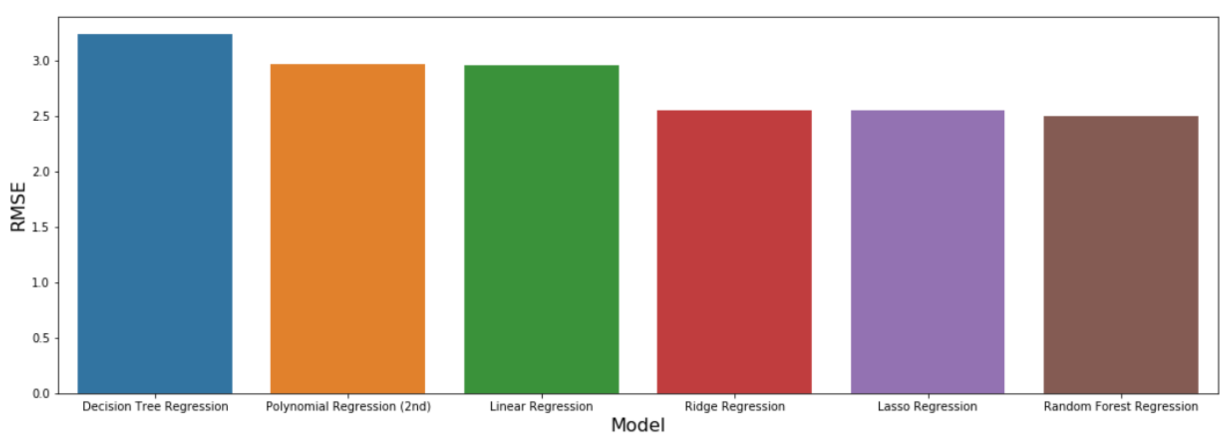


Diagram: Cross validation score for Regression models



<u>Diagram: R2 statistics for Regression models on trained and tested data.</u>



<u>Diagram: RMSE for Regression models</u>

Following all the visuals there is no straight conclusion on the best performing model. The Random Forest Regressor model has the least Mean square error, best R^2 (test) and second best R^2 (train). The Decision Tree Regression model has R^2 (train)=1 which is misleading as can be seen its test statistics. Based on these, we can conclude that **Random Forest Regression model is the best performing model for this dataset**.

Conclusions:

1. The best performing Linear Regression model based on OLS was the initial model that was modeled by fitting all the variables.
2. The best performing model based on the machine learning based regression models was the Random Forest Regression model for the given data.
3. The worst performing model among the machine learning based regression models was the Decision tree Regression model for the given data.
4. Only linear regression models were fitted based on the OLS method. There is scope to fit higher degree models that could better explain the variability in fitted model.
5. The median cost of houses can be predicted based on the above models.
6. Successfully modeled and statistically analyzed different regression models.
7. Various statically inferences (t-tests, ANOVA, p-values, others) were made to distinguish factors(predictor variables that may affect the outcome (response variables).