A logo with black text

Description automatically generated

CS770 Machine Learning

Assignment1: Linear Regression, Ridge& lasso Regression

09/27/2024

Submitted by,

Rajeet Chaudhary

J992Y875

**Introduction**

Linear regression is a type of supervised machine learning algorithm.

**1. Simple linear regression:** This involves predicting a dependent variable based on a single independent variable.

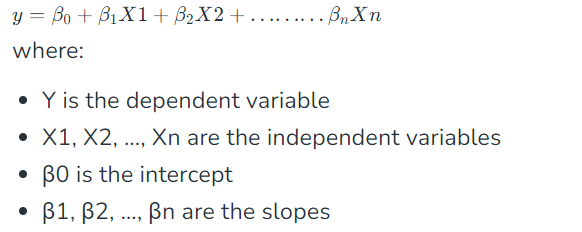
The equation of the regression line is represented as:

*h*(*xi*​)=*β*0​+*β*1​*xi*​

Here,

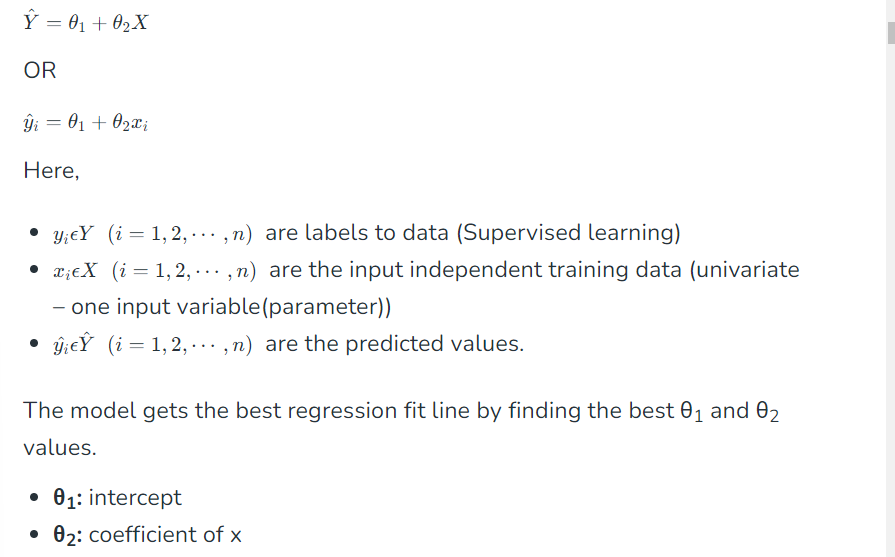
* h(x\_i) represents the **predicted response value** for ith observation.
* b\_0 and b\_1 are regression coefficients and represent the **y-intercept** and **slope** of the regression line respectively.

**2. Multivariate linear regression:** This involves predicting a dependent variable based on multiple independent variables.



The **Best Fit Line** equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

****Here Y is called the target variable and X is known as the predictor of Y(independent variable)



To achieve the best-fit regression line, it is very important to update the θ1 and θ2 values, to reach the best value that minimizes the error between the predicted y value (pred) and the true y value (y).



**Cost function for Linear Regression**

The [cost function](https://www.geeksforgeeks.org/what-is-cost-function/) or the[loss function](https://www.geeksforgeeks.org/ml-common-loss-functions/) is nothing but the error or difference between the predicted value  and the true value.



where,



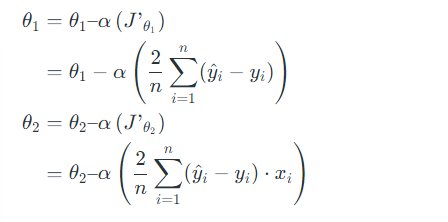
**Gradient Descent Algorithm**

A linear regression model can be trained using the optimization algorithm [gradient descent](https://www.geeksforgeeks.org/gradient-descent-algorithm-and-its-variants/)by iteratively modifying the model’s parameters to reduce the[mean squared error (MSE)](https://www.geeksforgeeks.org/python-mean-squared-error/) of the model on a training dataset.

To update θ1 and θ2 values in order to reduce the Cost function (minimizing RMSE value) and achieve the best-fit line the model uses Gradient Descent.

After differentiating w.r.t θ1 and θ2, final values:

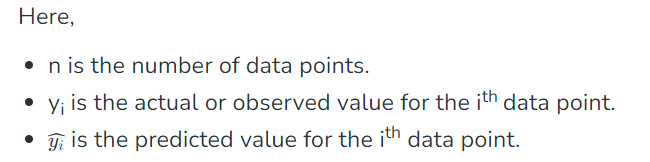


**Evaluation Metrics for Linear Regression**

**Mean Square Error (MSE)**

Mean Squared Error (MSE) calculates the average of the squared differences between the actual and predicted values.





**Root Mean Squared Error (RMSE)**

It describes how well the observed data points match the expected values, or the model’s absolute fit to the data.

In mathematical notation:



**R-Squared**

It isastatistical measure which represents the goodness of fit of a regression model with value from 0 to 1. It is also called as coefficient of determination.

Mathematically,



**RSS -** [**Residual sum of Squares**](https://www.geeksforgeeks.org/residual-sum-of-squares/#:~:text=Residual%20sum%20of%20squares%20is%20used%20to%20calculate%20the%20variance,squares%2C%20the%20better%20the%20model.)  


**TSS- Total Sum of Squares**  


**Methods**

**Data Preparation**

1. The load\_boston dataset was **loaded** from sklearn.datasets.
2. **Exploratory Data Analysis on given Dataset**: Initial exploration was done which includes
   * Checking for missing values
   * Filling Missing values with mean, median, mode
   * Finding and handling outliers
   * Understanding the distribution of values of different independent feature and dependent target values with different plots
   * Visualizing relationships between feature and target variable with different plots
3. **Feature Selection**: With findings of correlation value features having more than 0.5 were selected:
   * RM
   * LSTAT
   * PTRATIO
4. **Target Variable**: Here, the MEDV is the target variable.

After data preparation, the following steps should be taken:

**Regression Models Building Steps**

1. **Split the Data into Training and Testing data**

Divide randomly the dataset into two sets: training sets and testing sets. 25 percent of data will be used as test data or target value and 75 percent of data will be used as training data or input features value to predict the target value which mean test\_size = 0.25

2. **Fit the dataset into the Linear Regression Model**

* **Model Initialization**
* **Model Training**

3. **Evaluation of Model**

* **Predictions**
* **Calculation of r-square value**
* **Calculation of MSE value**

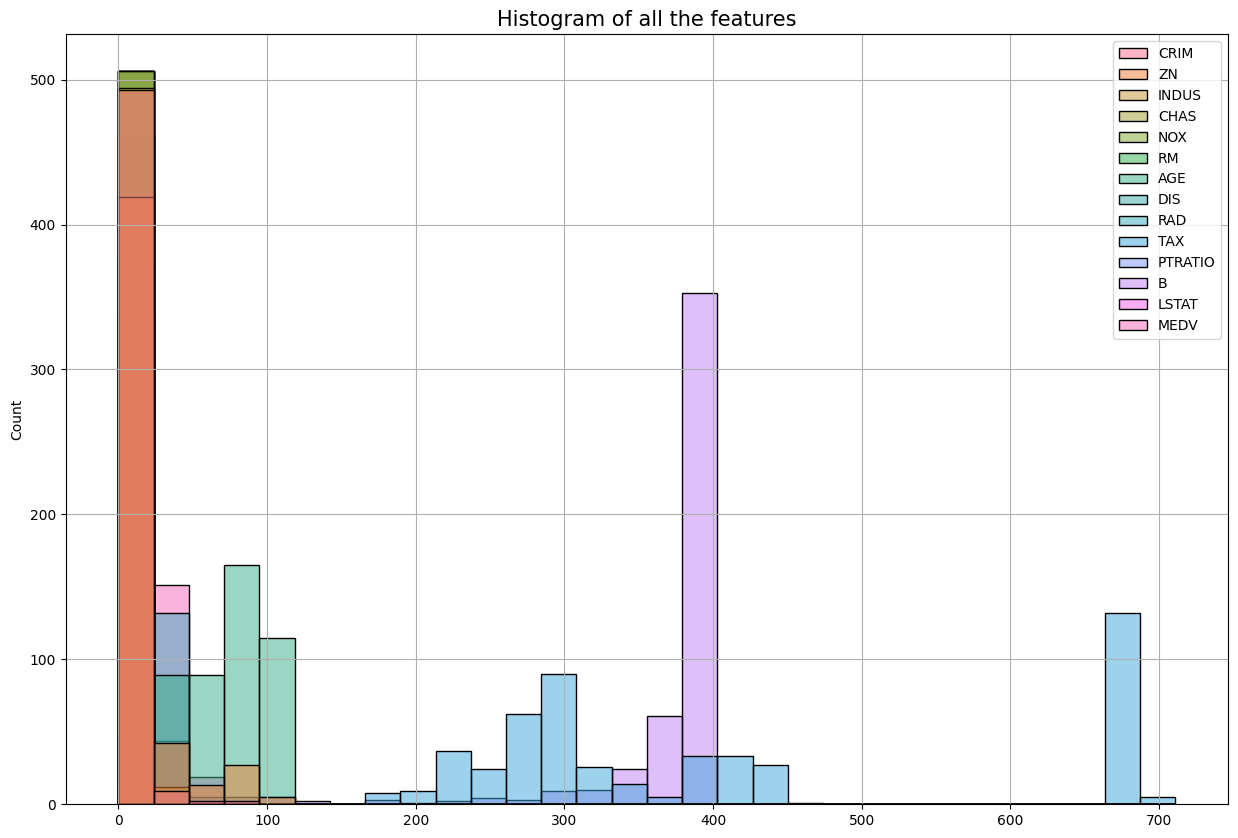
**4. Visualization**

A scatter plot was generated to compare the actual price of house against the predicted price from the different types of linear regression model.

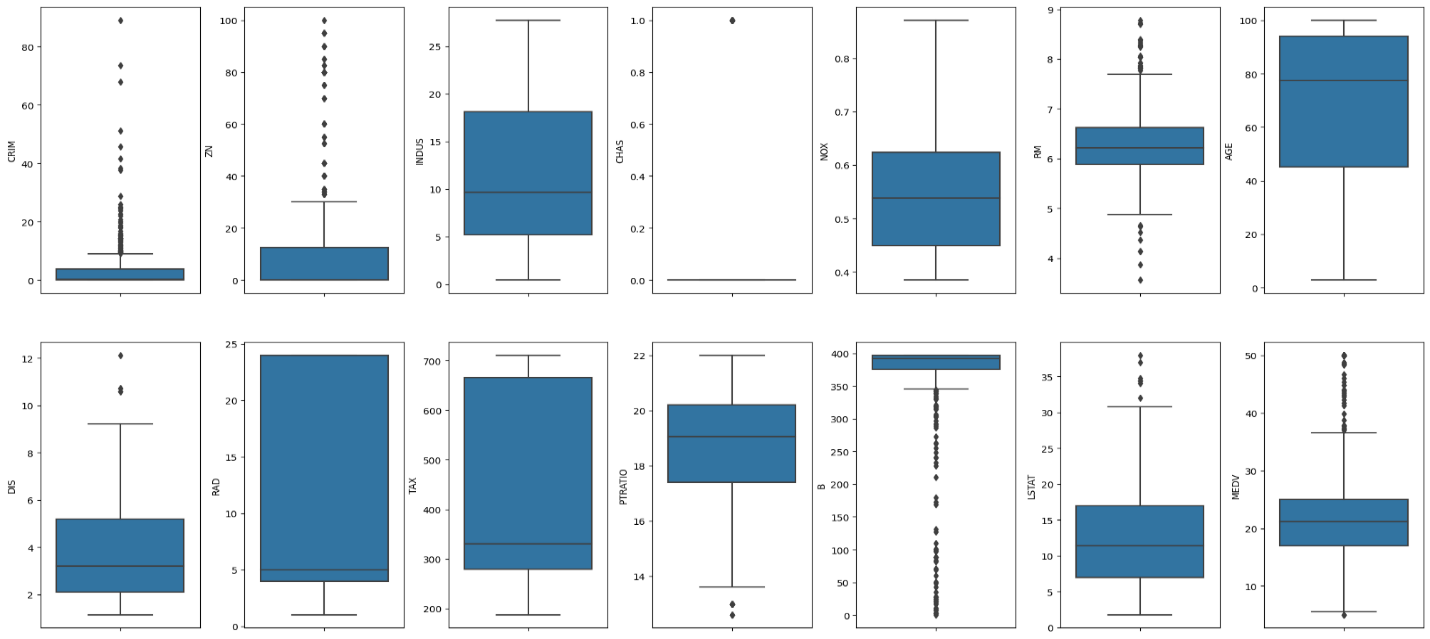
**1. BOSTON HOUSING DATASET**

**DATA VISUALISATION**

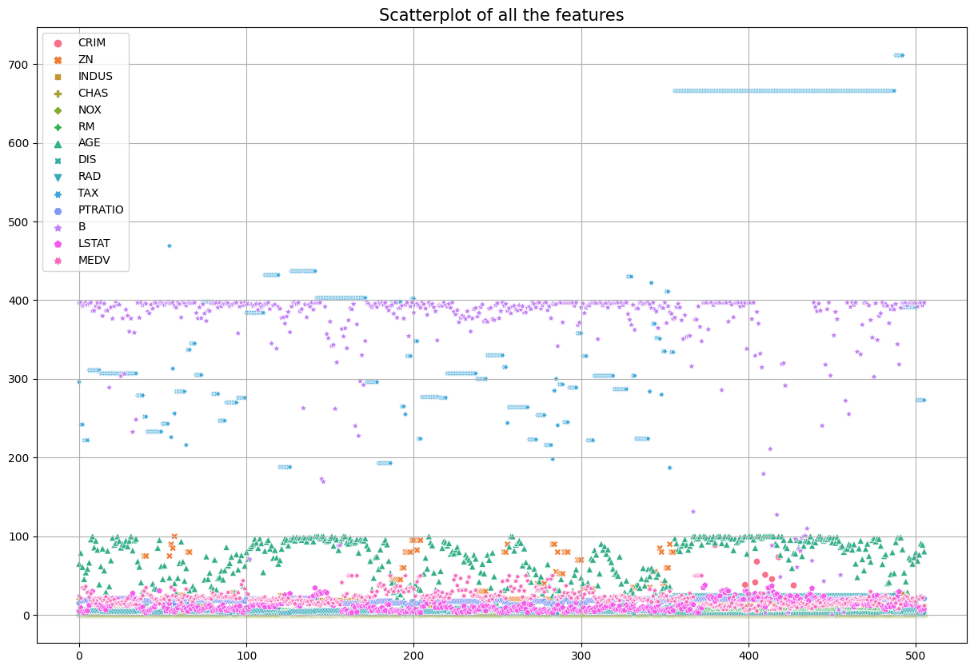
**1. Histogram of all features**



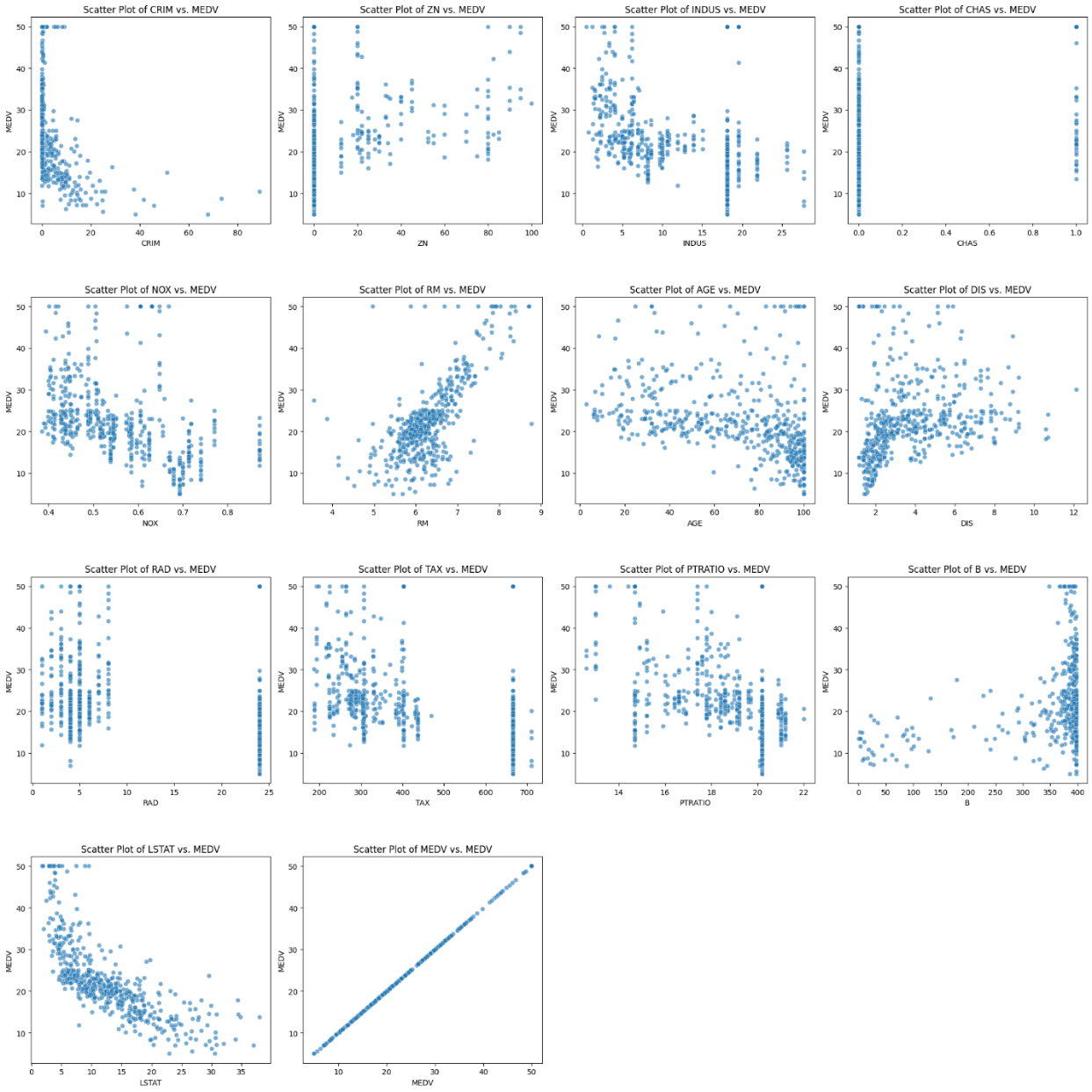
**2. Boxplot**



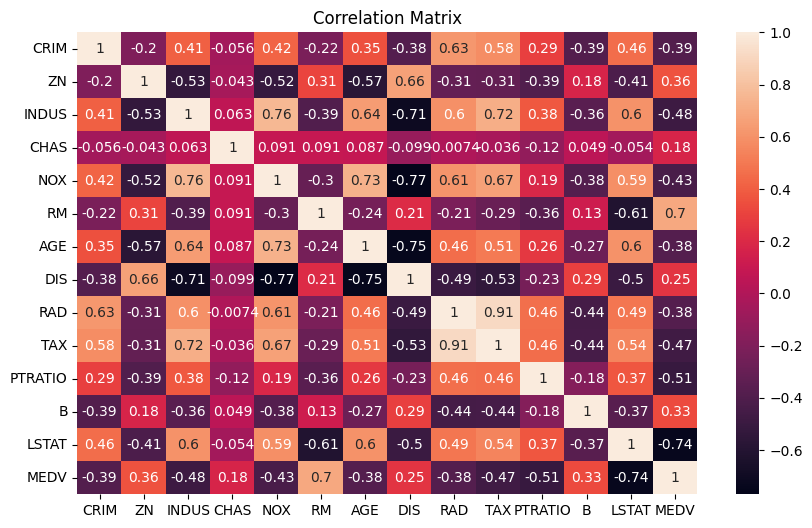
**3. Scatter Plot**



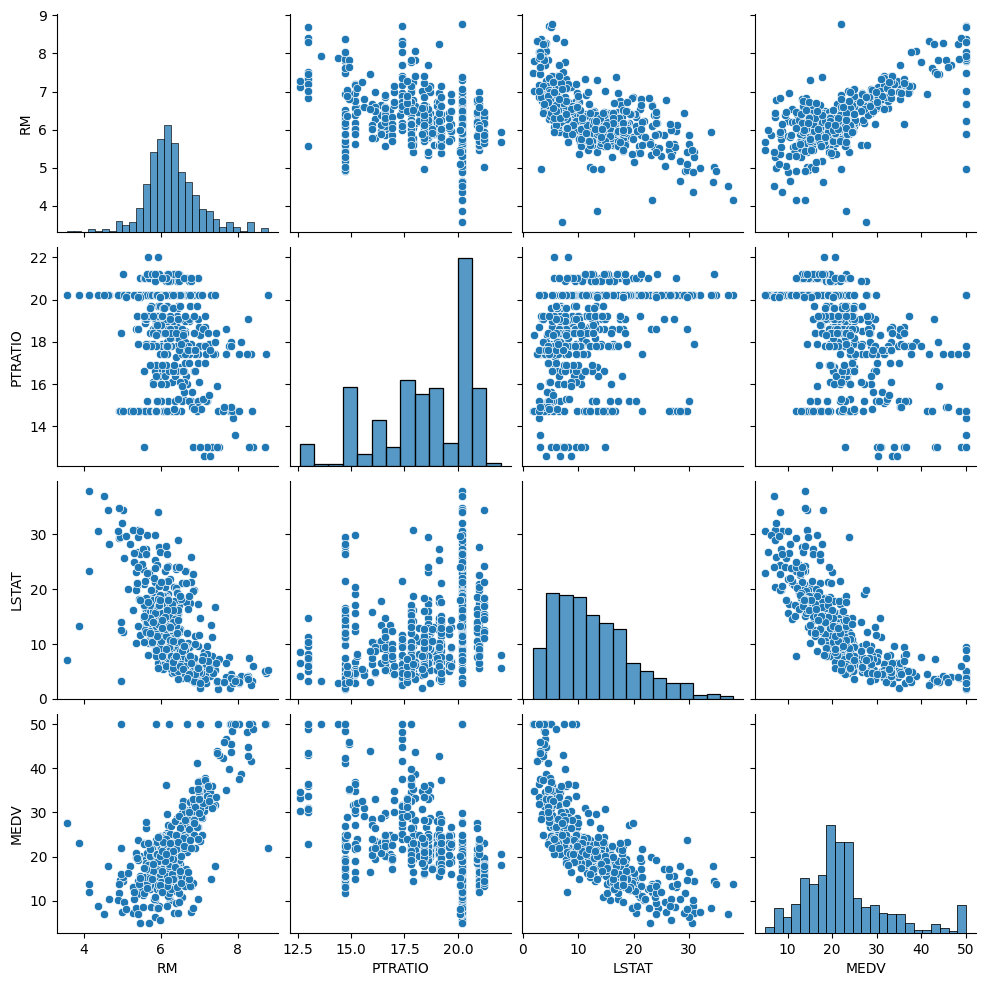
**4. Scatter Plot against MEDV**



**5. Heatmap**



**6. PAIRPLOT RM, PTRATIO, LSTAT, MEDV**

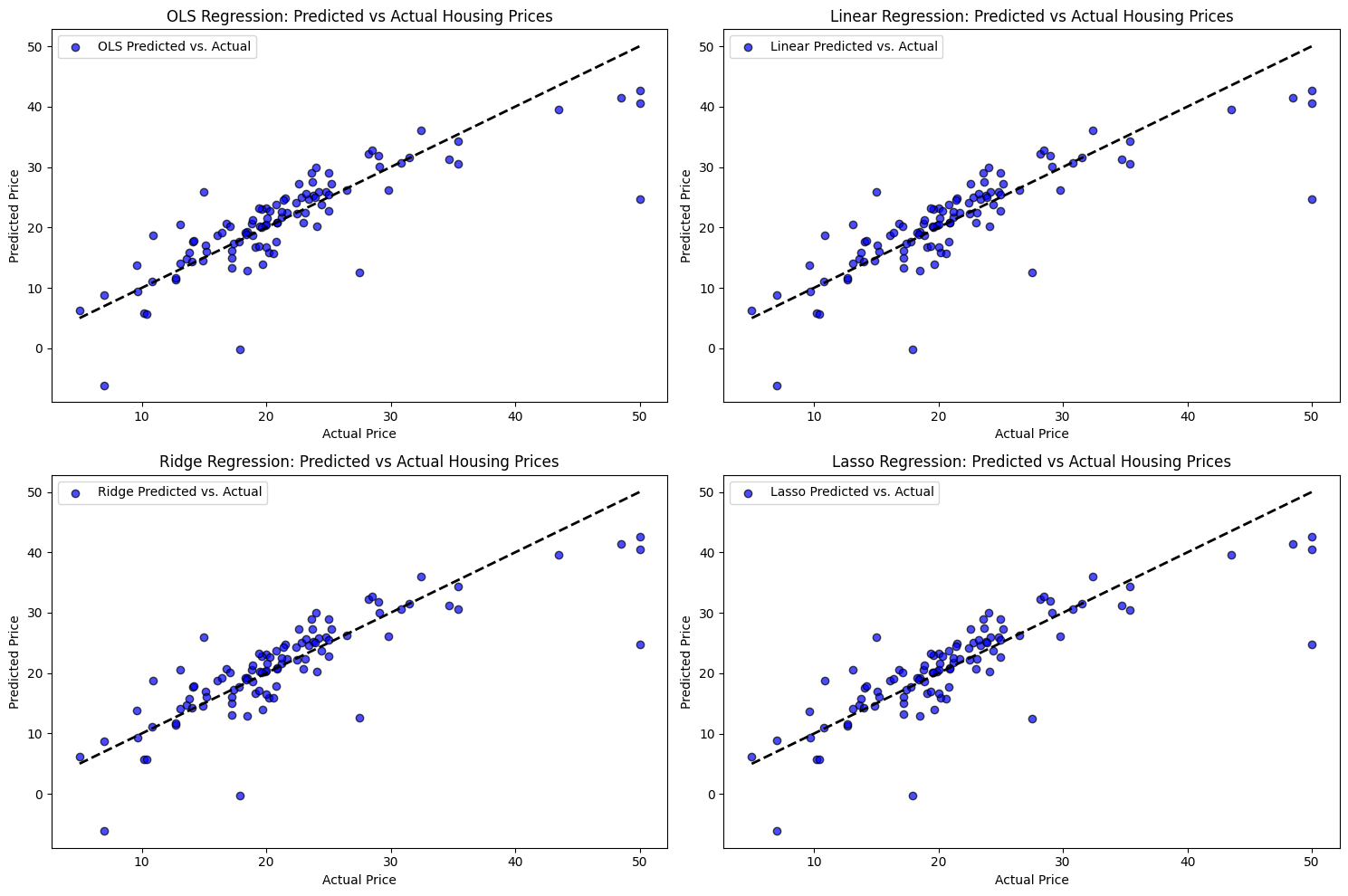


**Results and Discussion**

**Case 1: Without outlier removal and without feature scaled**

Error Calculation Table

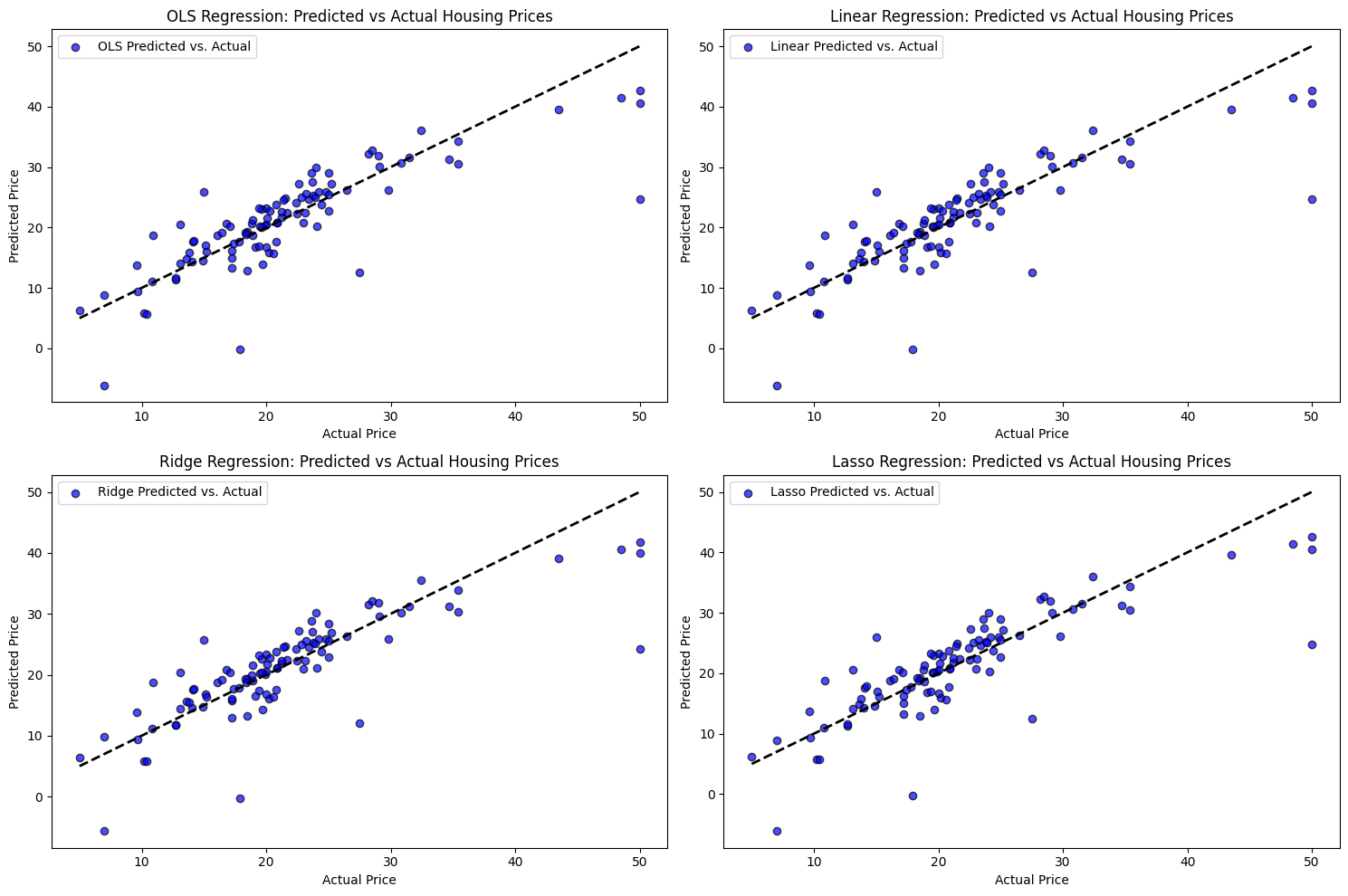
|  |  |  |
| --- | --- | --- |
| Regression | R^2 square | MSE |
| Ordinary Least Square (OLS) Regression | 0.6688 | 24.2911 |
| Linear Regression | 0.6688 | 24.2911 |
| Ridge Regression | 0.6686 | 24.3010 |
| Lasso Regression | 0.6688 | 24.2888 |



**Case 2: Without outlier removal and with feature scaling**

Error Calculation Table

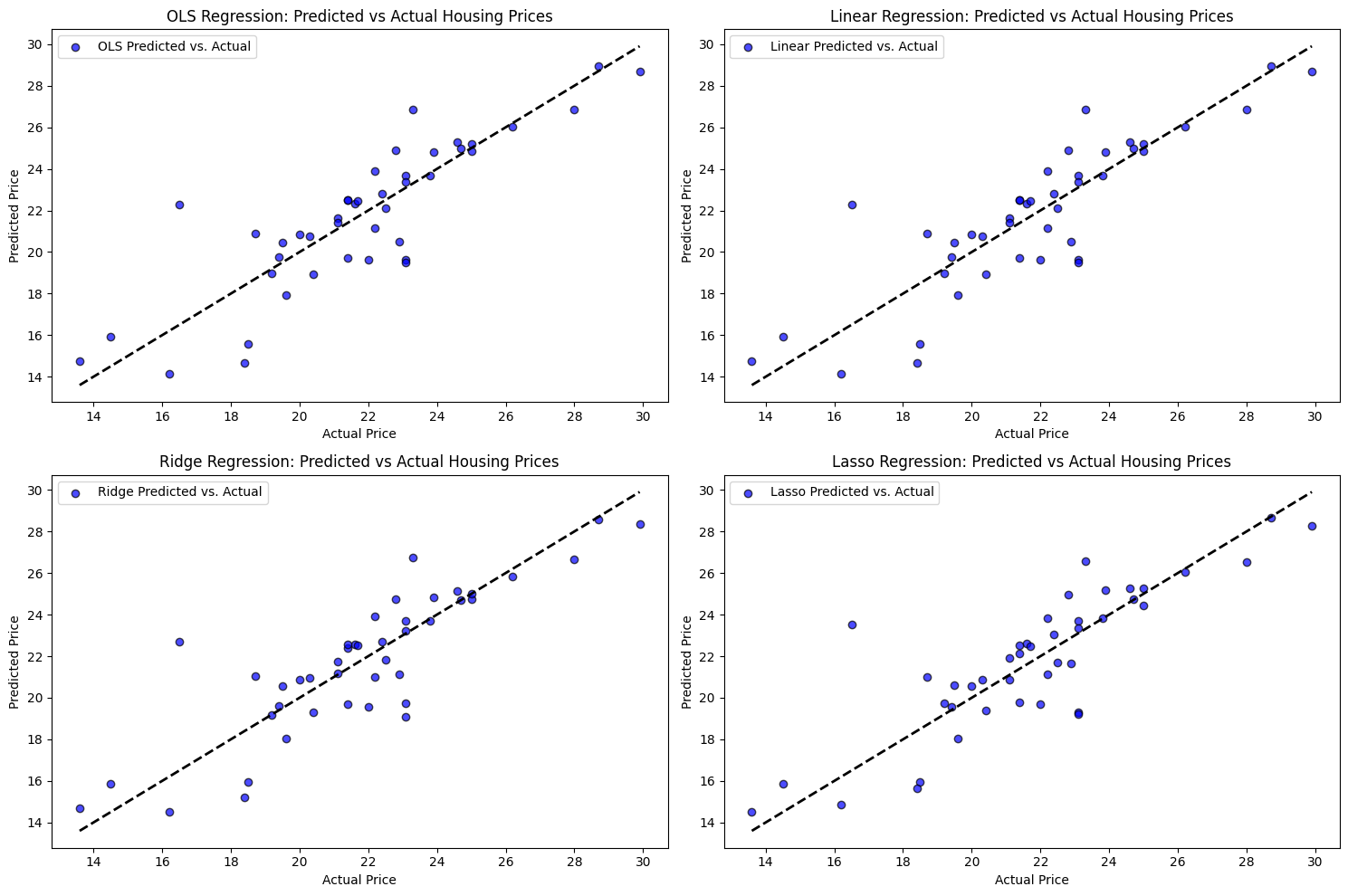
|  |  |  |
| --- | --- | --- |
| Regression | R^2 square | MSE |
| Ordinary Least Square (OLS) Regression | 0.6688 | 24.2911 |
| Linear Regression | 0.6688 | 24.2911 |
| Ridge Regression | 0.6660 | 24.4958 |
| Lasso Regression | 0.6687 | 24.2945 |



**Case 3: With outlier removal and feature scaled**

Error Calculation Table

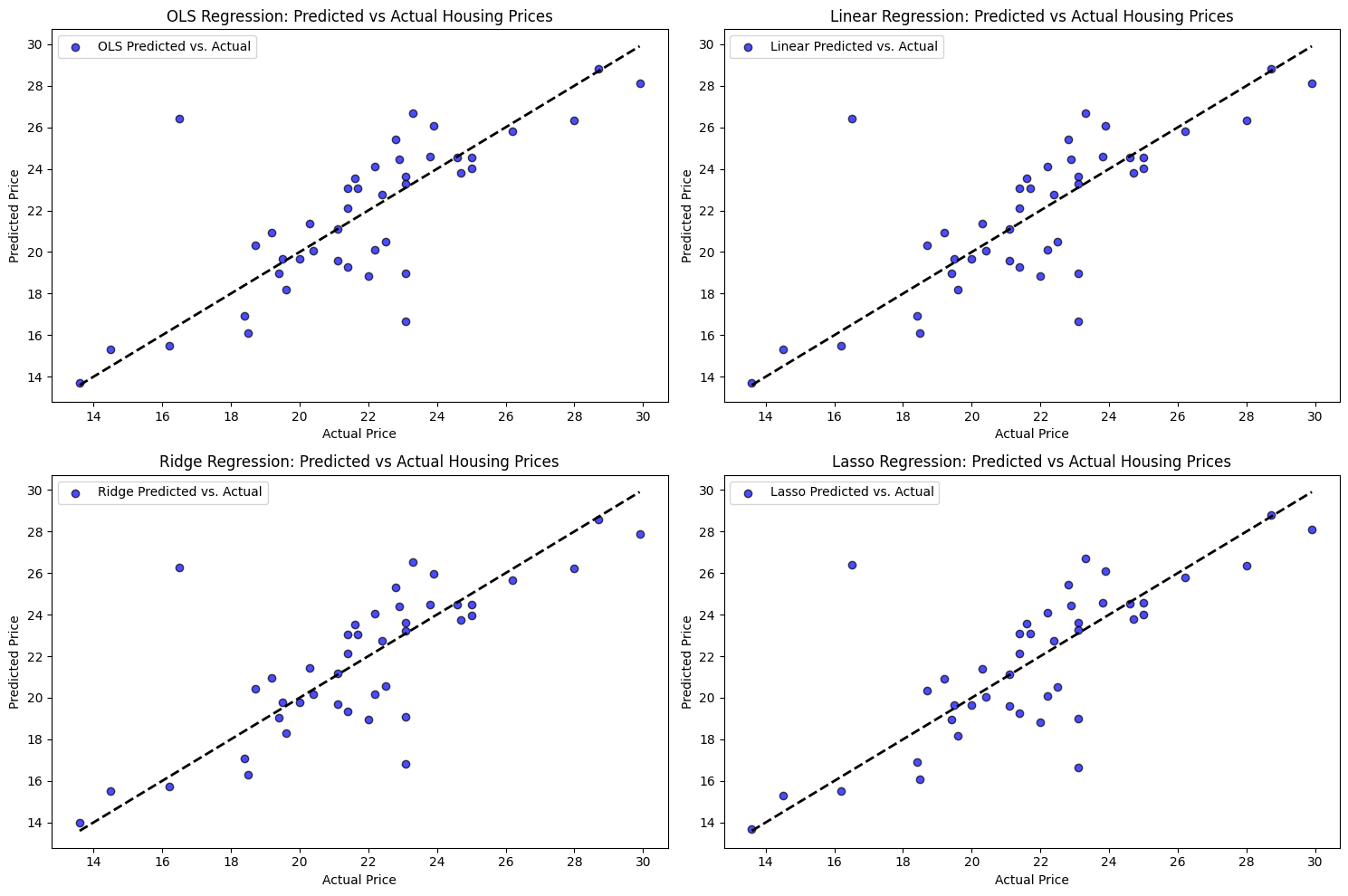
|  |  |  |
| --- | --- | --- |
| Regression | R^2 square | MSE |
| Ordinary Least Square(OLS) Regression | 0.6938 | 3.3628 |
| Linear Regression | 0.6938 | 3.3628 |
| Ridge Regression | 0.6970 | 3.3275 |
| Lasso Regression | 0.6830 | 3.4815 |



**Case 4: With feature having greater than 0.5, RM, PTRATIO, LSTAT and with feature scaled and outlier removed**

Error Calculation Table

|  |  |  |
| --- | --- | --- |
| Regression | R^2 square | MSE |
| Ordinary Least Square (OLS) Regression | 0.4758 | 5.7565 |
| Linear Regression | 0.4758 | 5.7565 |
| Ridge Regression | 0.4985 | 5.5075 |
| Lasso Regression | 0.4761 | 5.7537 |



**Conclusion**

From all the above cases, it can be concluded that

i. There is much improvement in MSE and r square value after outlier removal

ii. Feature scaling little bit decreased the MSE and r square value for each different types of regression.

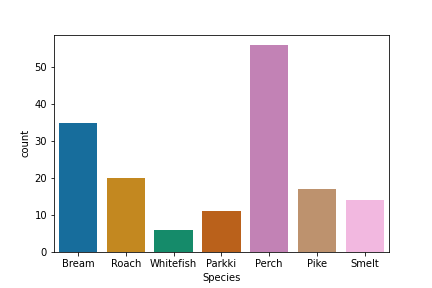
iii. For each different types of Linear Regression the best line is plotted to fit the actual and predicted values.

**2.** **FISH.CV DATASET**

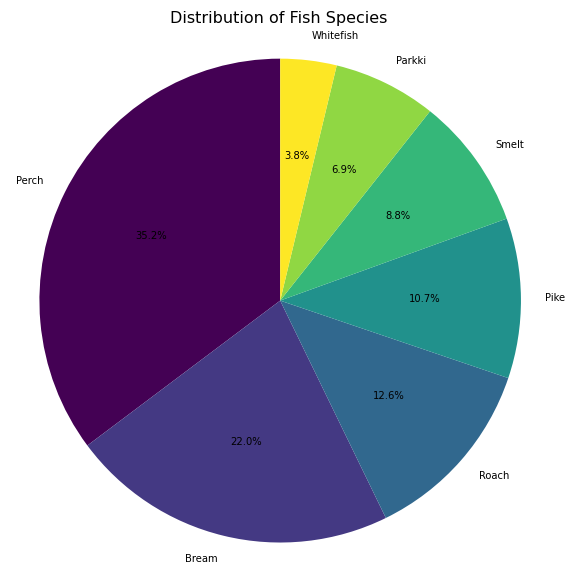
**This dataset consists of:**

* Species = Species name of fish
* Weight = Weight of fish in Gram g
* Length1 = Vertical length in cm
* Length2 = Diagonal length in cm
* Length3 = Cross length in cm
* Height = Height in cm
* Width = Diagonal width in cm

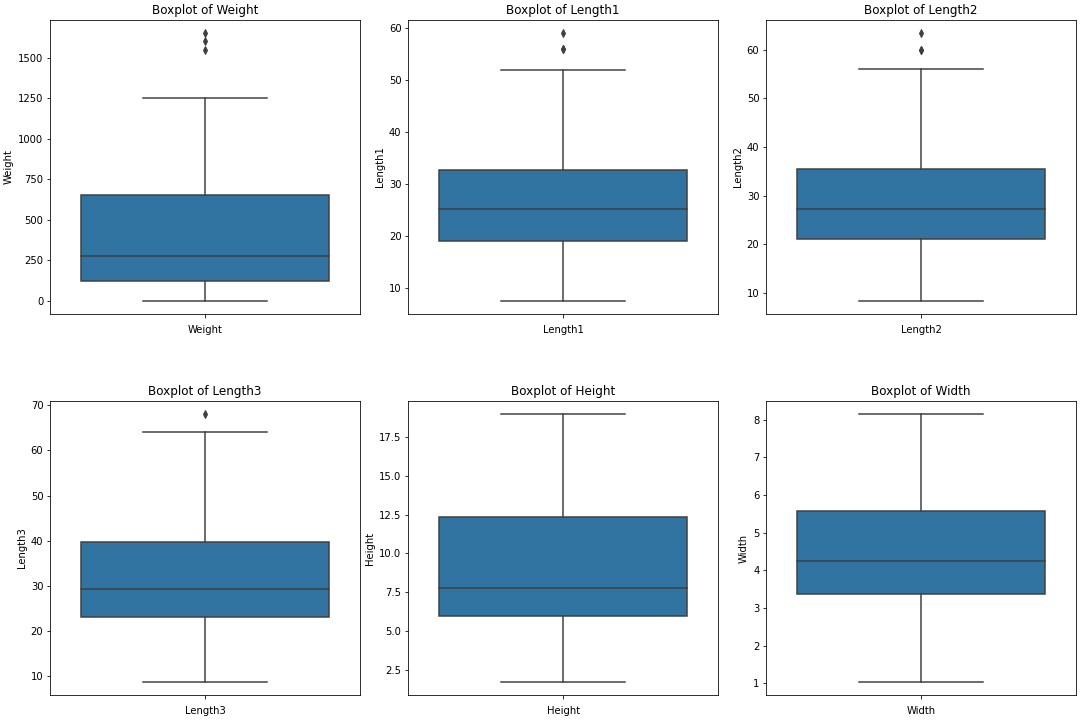
**VISUALIZATION:  
1. Histogram**

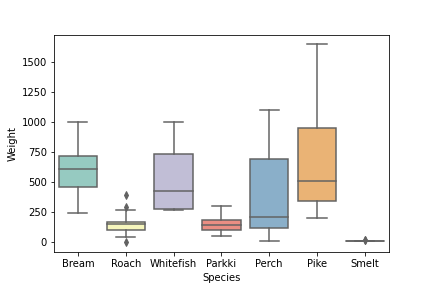
****

**2. Pie chart**

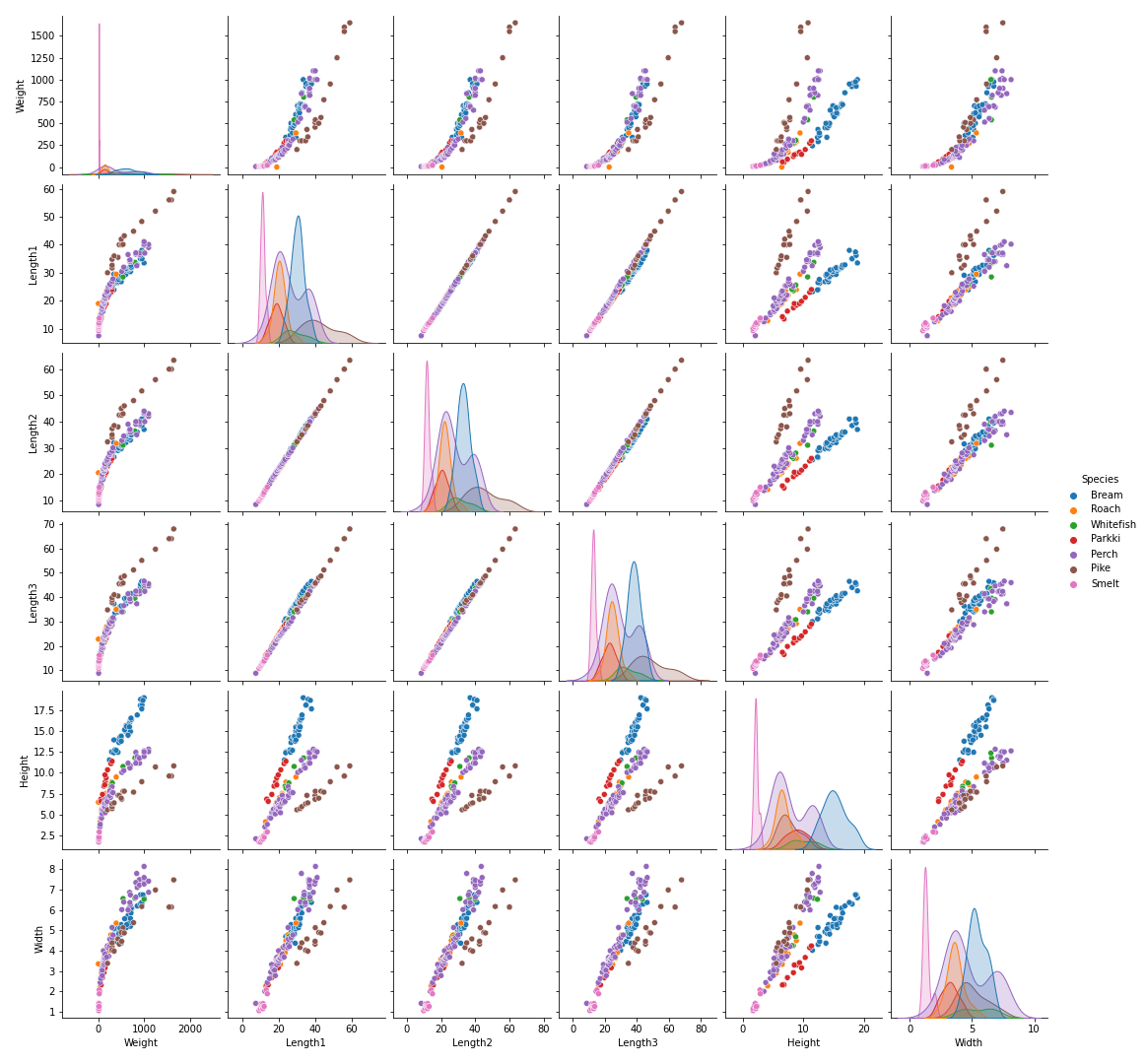
****

**3. BOX PLOT**

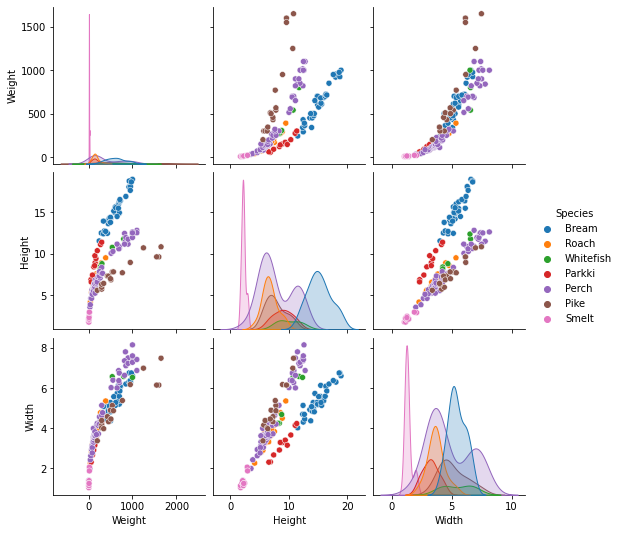
****

**4. Species wise boxplot   
**

**5. PAIR PLOT**

****

**5.1. PAIR PLOT FOR HEIGHT, WIDTH, WEIGHT**



**6. Heatmap**

**Results and Discussion**

R^2 square and MSE value are listed below.  
With this, coefficients are also calculated and scatter plots of actual vs predicted values are generated with the best fit line.

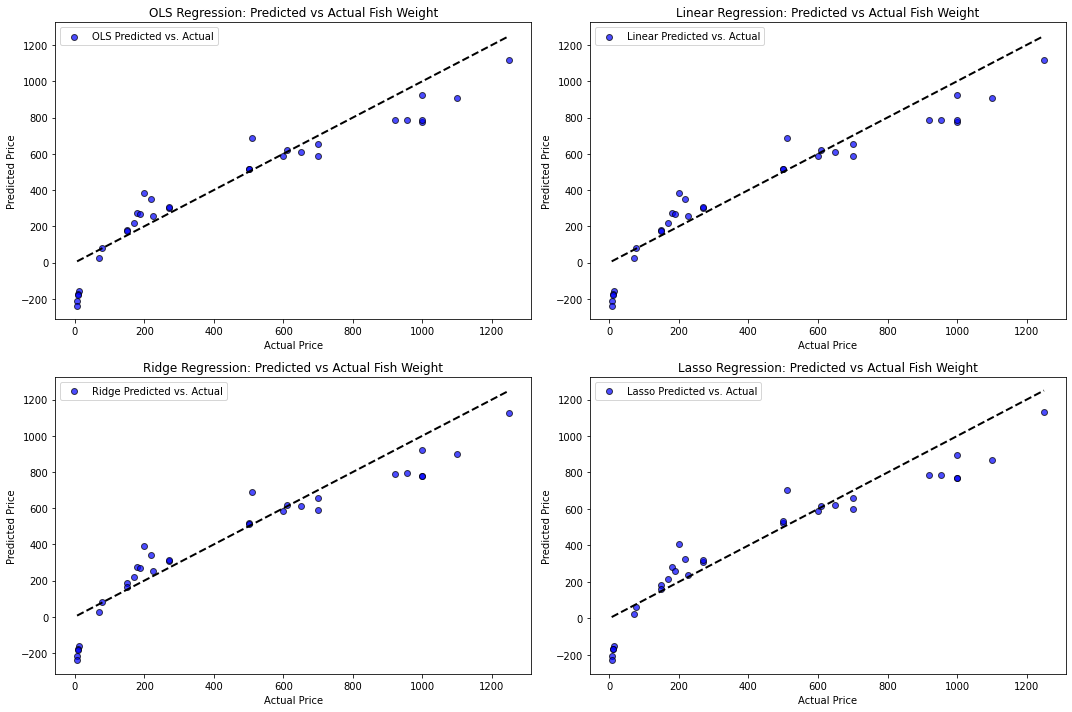
**Case 1: Without outlier removal**

Error Calculation Table

|  |  |  |
| --- | --- | --- |
| Regression | R^2 square | MSE |
| Ordinary Least Square (OLS) Regression | 0.8821 | 16763.8872 |
| Linear Regression | 0.8821 | 16763.8872 |
| Ridge Regression | 0.8803 | 17022.0223 |
| Lasso Regression | 0.8774 | 17431.9637 |

Coefficient Calculation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Features | OLS Regression Coefficients | Linear Regression Coefficients | Ridge Regression Coefficients | Lasso Regression Coefficients |
| Const | -515.305651 |  |  |  |
| Length1 | 43.535265 | 43.5352649 | 27.03858344 | 24.91762755 |
| Length2 | 7.821796 | 7.82179624 | 17.69642626 | 0.1063255 |
| Length3 | -25.256701 | -25.25670105 | -19.51672787 | -0. |
| Height | 23.228912 | 23.2289123 | 20.3581969 | 13.44083754 |
| Width | 27.066493 | 27.06649294 | 27.51531904 | 29.06666741 |



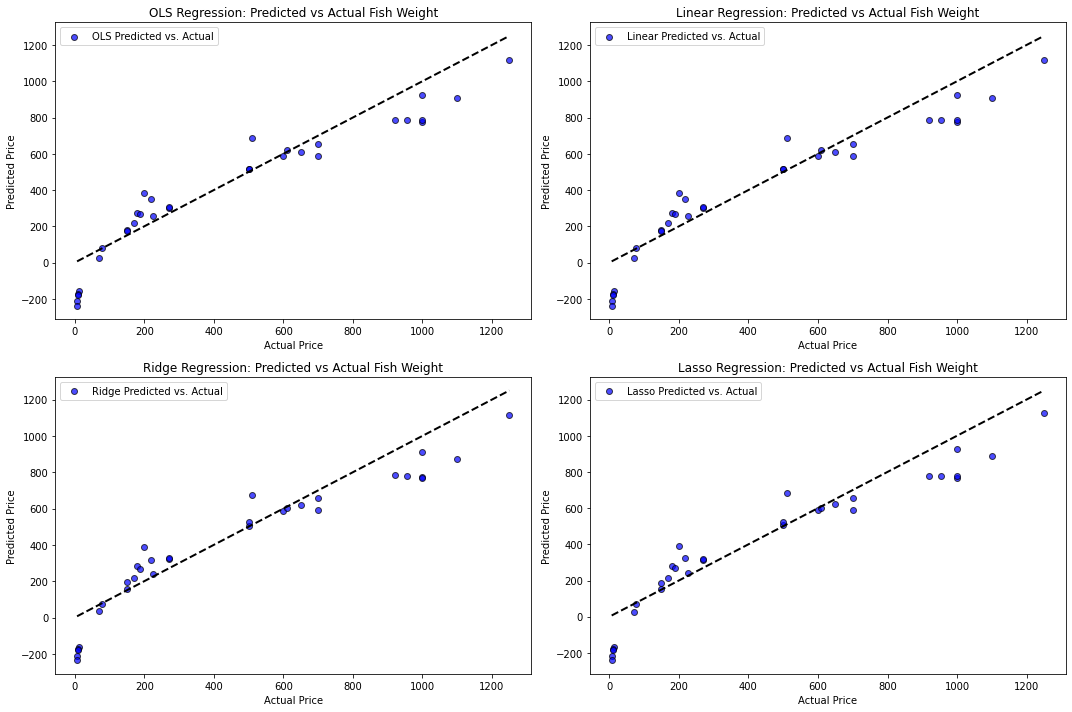
**Case 2: Without outlier removal and with no feature scaling**

Error Calculation Table

|  |  |  |
| --- | --- | --- |
| Regression | R^2 square | MSE |
| Ordinary Least Square(OLS) Regression | 0.8821 | 16763.8872 |
| Linear Regression | 0.8821 | 16763.8872 |
| Ridge Regression | 0.8770 | 17488.5753 |
| Lasso Regression | 0.8773 | 17458.6532 |

Coefficient Calculation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Features | OLS Regression Coefficients | Linear Regression Coefficients | Ridge Regression Coefficients | Lasso Regression Coefficients |
| Const | -515.305651 |  |  |  |
| Length1 | 43.535265 | 432.27472554 | 83.29023683 | 229.96344143 |
| Length2 | 7.821796 | 83.01304108 | 80.60503657 | 0. |
| Length3 | -25.256701 | -288.56797575 | 64.91737168 | 0. |
| Height | 23.228912 | 92.52321581 | 35.7986499 | 44.01533366 |
| Width | 27.066493 | 44.06740907 | 77.01514167 | 75.78051359 |



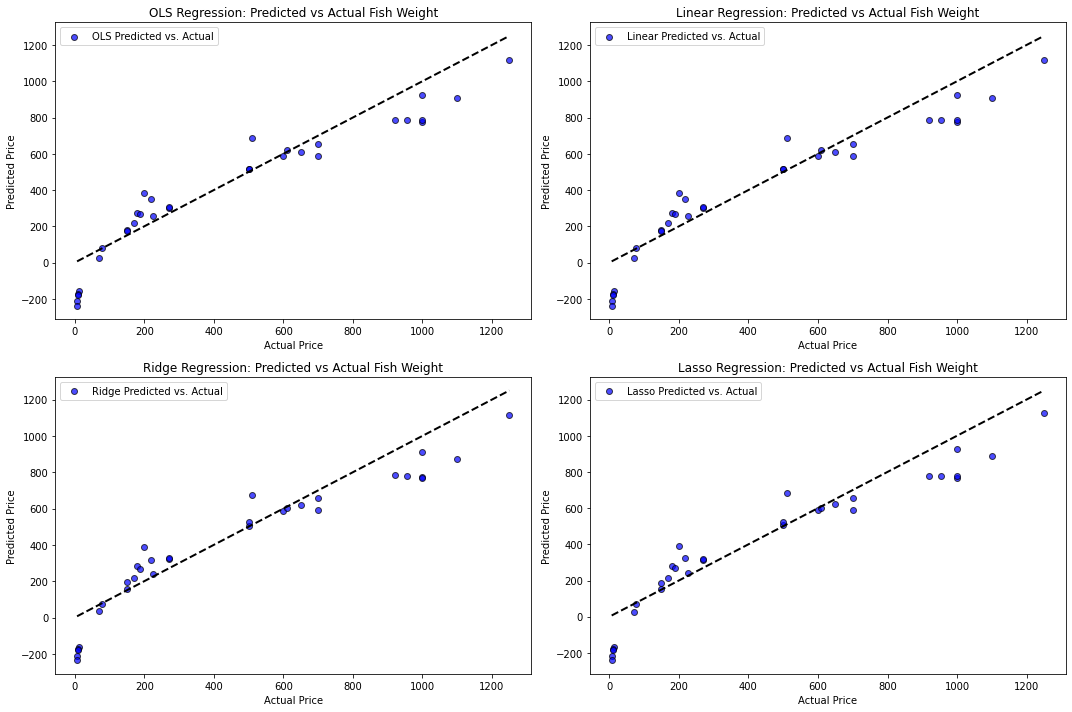
**Case 3: With outlier removal and feature scaled**

Error Calculation Table

|  |  |  |
| --- | --- | --- |
| Regression | R^2 square | MSE |
| Ordinary Least Square(OLS) Regression | 0.8821 | 16763.8872 |
| Linear Regression | 0.8821 | 16763.8872 |
| Ridge Regression | 0.8770 | 17488.5753 |
| Lasso Regression | 0.8773 | 17458.6532 |

Coefficient Calculation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Features | OLS Regression Coefficients | Linear Regression Coefficients | Ridge Regression Coefficients | Lasso Regression Coefficients |
| Const | -515.305651 |  |  |  |
| Length1 | 43.535265 | 432.27472554 | 83.29023683 | 229.96344143 |
| Length2 | 7.821796 | 83.01304108 | 80.60503657 | 0. |
| Length3 | -25.256701 | -288.56797575 | 64.91737168 | 0. |
| Height | 23.228912 | 92.52321581 | 35.7986499 | 44.01533366 |
| Width | 27.066493 | 44.06740907 | 77.01514167 | 75.78051359 |



**Conclusion**

From all the above cases, it can be concluded that

i. There is not much difference in MSE and r square value even after outlier removal as the outlier count was very small.

ii. Feature scaling little bit increased the MSE and r square value for each different types of regression.

iii. Coefficient value increased very much after feature scaling for Linear, Lasso and Ridge Regression while in OLS coefficient, Feature scaling has no effect.