Project 1: Collecting and Analyzing Data

The dataset used for this project depicts the weight of different species of fishes. It is called the Fish Market dataset and it has 159 samples of data. The dataset mainly focuses on weights of seven different species of fishes namely Perch, Bream, Roach, Pike, Smelt, Parkki and Whitefish. The weight of the fish is the dependent variable in this data set whereas the vertical length, diagonal length, cross - sectional length, height and width are the independent variables which determine the weight of the fish. The link to the dataset is given below it is also present in the project folder.
Dataset: https://www.kaggle.com/datasets/aungpyaeap/fish-market?resource=download

Step 1 : Installing the libraries
Installing all the necessary libraries required for the project. Went on installing libraries throughout the project whenever required.

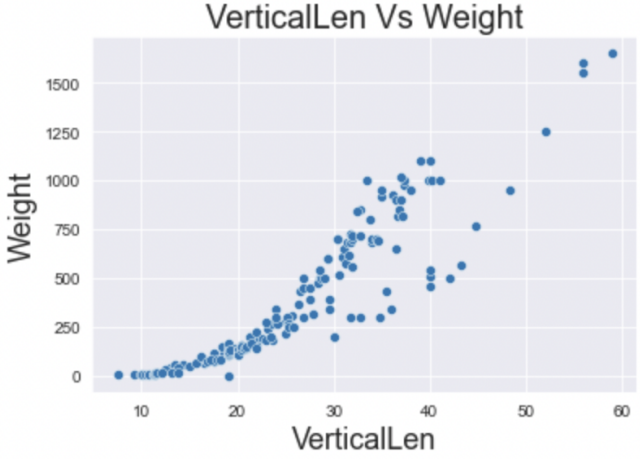
Step 2 : Reading and printing the dataset
Reading the csv file using the pandas library which stores the data in a tabular format which makes it easy to carry out the rest of the work like data cleaning and preprocessing on the dataset.

Step 3 : Renaming the column names
All the three lengths were named as length 1, length 2, length 3. To avoid confusion and to have a better picture, the three lengths were renamed to vertical length, diagonal length and cross - sectional length respectively.

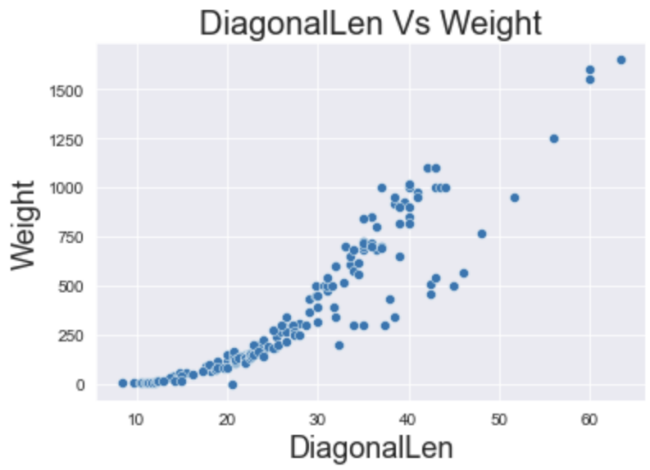
Step 4 : Checking the count and types of Species
Checking the different types of species present in the dataset and the unique number of each species present. Plotting a bar graph with the count of species on the Y- axis against the different types of species on the X-axis.

Step 5 : Plotting independent variables
The next step was to use the scatterplot module from seaborn to plot each independent variable against the dependent variable on a scatter plot and the pyplot module from matplotlib to show the plot diagram. The different scatter plots were :

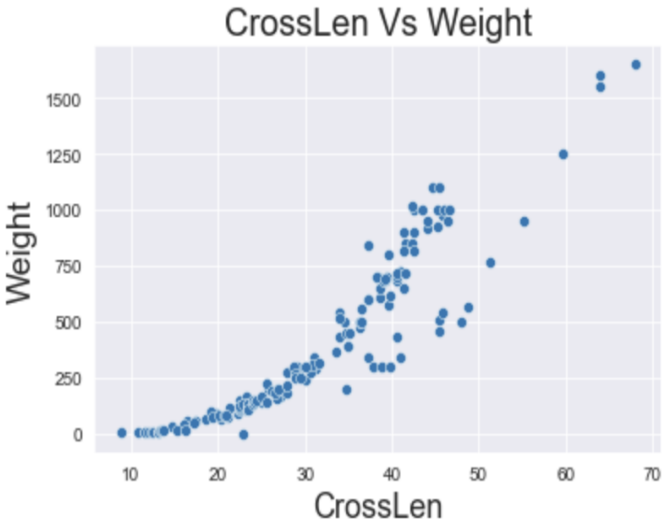
1) **Vertical Length v/s Weight**



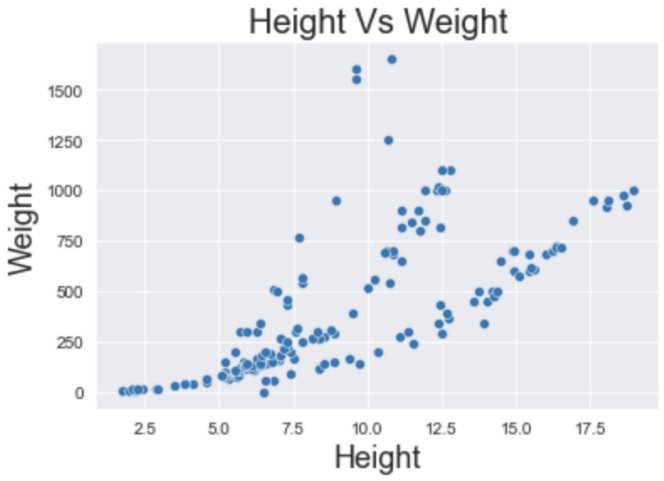
2) Diagonal Length v/s Weight



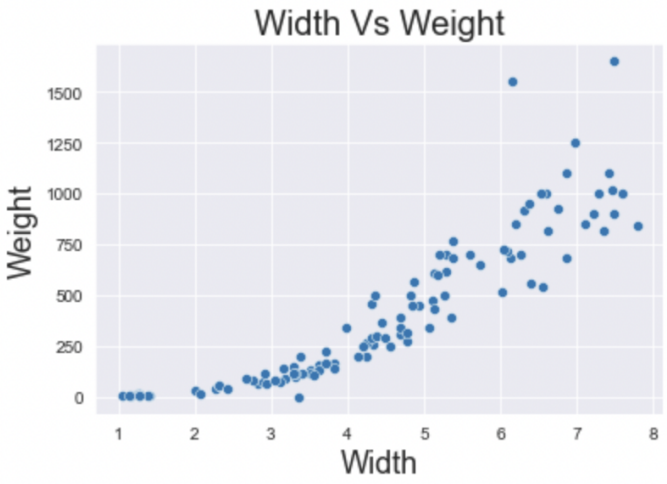
3) Cross-Sectional Length v/s Weight



4) Height v/s Weight



5) **Width v/s Weight**



Step 6 : Data Cleaning

Data cleaning is an important part of a machine learning project. This makes sure that the data used to train and test the model is accurate and it is also important to make accurate predictions and generalize well on the new and unseen data. The steps taken for data cleaning are as follows :

1) Checking and removing any **null values** present in the dataset.
2) Next we describe the data using the **describe()** function in pandas. It gives the count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum of the data. This helps to get a quick overview of the distribution of data and helps in identifying the outliers and missing values.
3) The next step would be removing the outliers present in the dataset. Each column of the dataset is represented using a boxplot graph. A box plot represents the distribution of the data. The line inside the box represents the median and the whiskers represent the maximum and minimum values excluding the outliers. So, anything outside the whiskers is considered an outlier. The outliers are identified using the IQR(interquartile range) method. The Height and Width had no outliers and rest all the outliers were dropped from the dataset.

Step 7 : Organize your data

The next step was to organize the dataset. The dataset which was a single csv file was divided into two different csv files. One for the training set and one for the testing set. 70% of the data is for training (equivalent to 109 samples) whereas 30% of the data is for testing (which is equivalent to 47 samples). The function divide_data() takes in the new_data (which is the dataset got after data preprocessing and cleaning). In the function ‘n’ depicts the length of the dataset. The number of rows to be included in the training set is n multiplied by the training percentage. The data is shuffled so that the training and testing sets are made up with random shuffled data. And then split into training and testing sets and stored in different variables. Then two csv files are created separately namely train.csv file and test.csv file which store the appropriate data. Then

we check if the data is present in the operating system and the output can be viewed using the pandas read_csv function.

1) Train.csv

```
In [345]: pd.read_csv("train.csv")
```

Out[345]:

	Species	Weight	VerticalLen	DiagonalLen	CrossLen	Height	Width
0	Perch	120.0	20.0	22.0	23.5	6.1100	3.4075
1	Smelt	7.5	10.0	10.5	11.6	1.9720	1.1600
2	Bream	340.0	29.5	32.0	37.3	13.9129	5.0728
3	Smelt	7.0	10.1	10.6	11.6	1.7284	1.1484
4	Perch	300.0	26.9	28.7	30.1	7.5852	4.6354
...	...	...	...	...	...	...	...
104	Bream	680.0	31.8	35.0	40.6	15.4686	6.1306
105	Pike	300.0	31.7	34.0	37.8	5.7078	4.1580
106	Smelt	19.9	13.8	15.0	16.2	2.9322	1.8792
107	Roach	180.0	23.6	25.2	27.9	7.0866	3.9060
108	Perch	225.0	22.0	24.0	25.5	7.2930	3.7230

109 rows × 7 columns

2) Test.csv

```
In [377]: pd.read_csv("test.csv")
```

Out[377]:

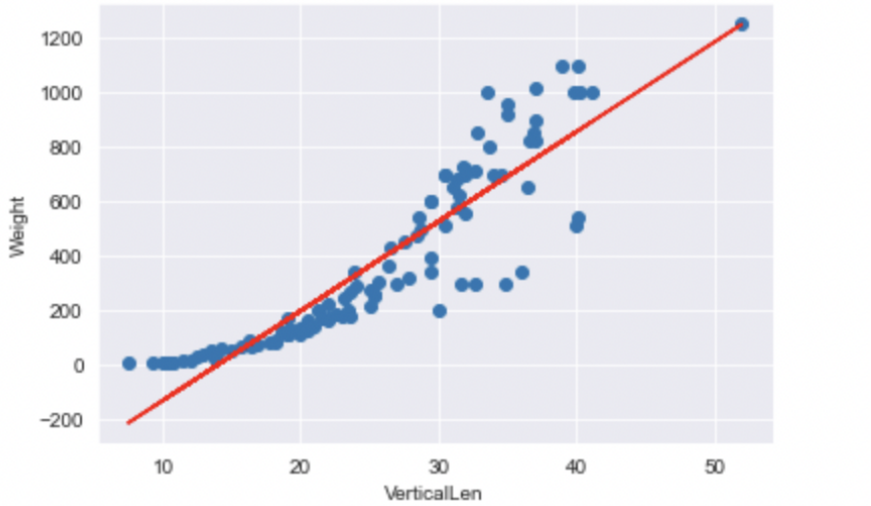
	Species	Weight	VerticalLen	DiagonalLen	CrossLen	Height	Width
0	Perch	840.0	32.5	35.0	37.3	11.4884	7.7957
1	Perch	78.0	16.8	18.7	19.4	5.1992	3.1234
2	Whitefish	800.0	33.7	36.4	39.6	11.7612	6.5736
3	Bream	700.0	30.4	33.0	38.3	14.8604	5.2854
4	Perch	690.0	34.6	37.0	39.3	10.5717	6.3666
5	Whitefish	270.0	24.1	26.5	29.3	8.1454	4.2485
6	Bream	950.0	38.0	41.0	46.5	17.6235	6.3705
7	Bream	680.0	31.8	35.0	40.6	15.4686	6.1306
8	Pike	200.0	30.0	32.3	34.8	5.5680	3.3756
9	Smelt	10.0	11.3	11.8	13.1	2.2139	1.2838
10	Parkki	170.0	19.0	20.7	23.2	9.3960	3.4104
11	Perch	320.0	27.8	30.0	31.6	7.6156	4.7716
12	Bream	430.0	26.5	29.0	34.0	12.4440	5.1340
13	Perch	225.0	22.0	24.0	25.5	7.2930	3.7230
14	Bream	475.0	28.4	31.0	36.2	14.2628	5.1042

Step 8 : Execute a linear regression (Include graphs)

The next step includes applying the linear regression model. Linear regression is applied with the dependent variable on the Y-axis and the independent variable on the X-axis. Linear regression is applied separately for each independent variable along with the dependent variable. The relationship between each independent variable and the dependent variable is as follows in regard to the slope and R coefficient :

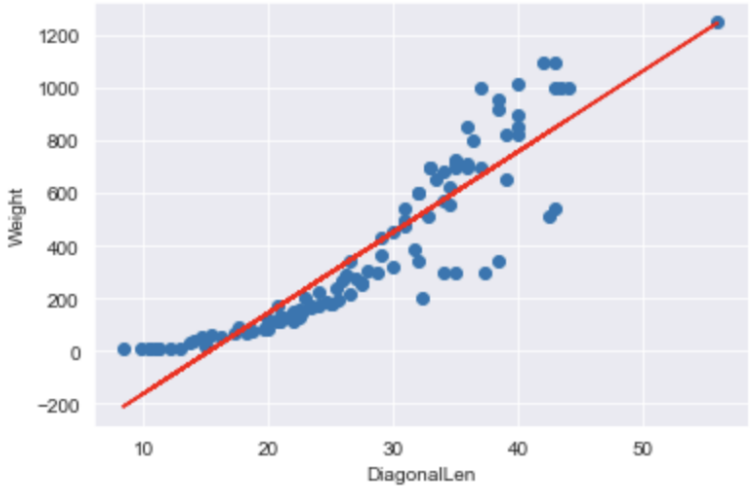
The slope of a linear regression model represents the change in the dependent variable for one-unit change in the independent variable and the R-coefficient depicts how well the regression model fits the data and it ranges from 0 to 1.

1) Vertical Length v/s Weight



Slope: 29.433247469788316
The value of slope is positive which means as the vertical length of the fish increases the weight also increases. It is a positive correlation.
R-Coefficient: 0.8010509163171013
The R value is not 1 but it is closer to 1 depicting that the linear regression model fits the data well.

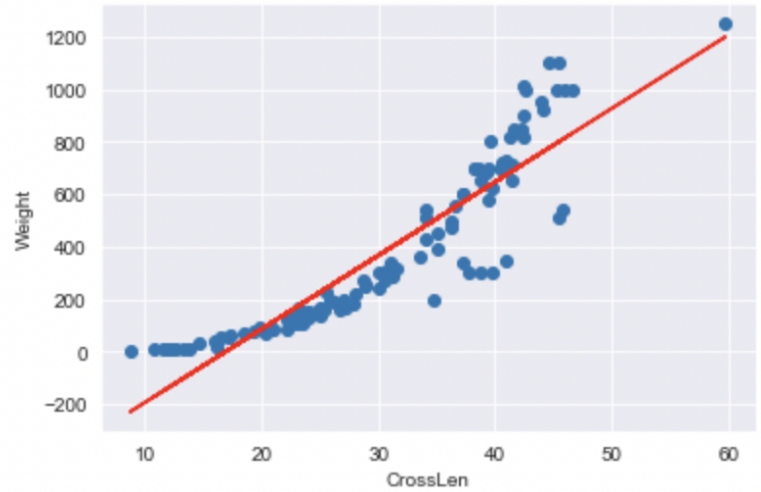
2) Diagonal Length v/s Weight



Slope: 27.564481236203843
The value of slope is positive which means as the diagonal length of the fish increases the weight also increases. It is a positive correlation.

R-Coefficient: 0.8071567503584482
The R value is not 1 but it is closer to 1 and slightly greater than the R value of vertical length depicting the model fits slightly better for diagonal length.

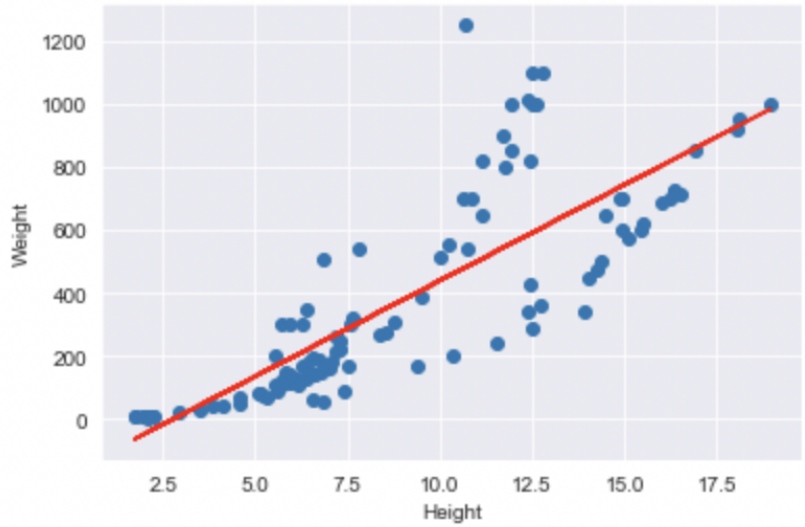
3) Cross-Sectional Length v/s Weight



Slope: 25.520962444240723
The value of slope is positive which means as the cross length of the fish increases the weight also increases. It is a positive correlation. But it is less than both the vertical and diagonal length showing a less stronger relationship between X and Y compared to the above two variables.

R-Coefficient: 0.821662902921837
The R value is greater than the diagonal and vertical length which means the model fits the data well.

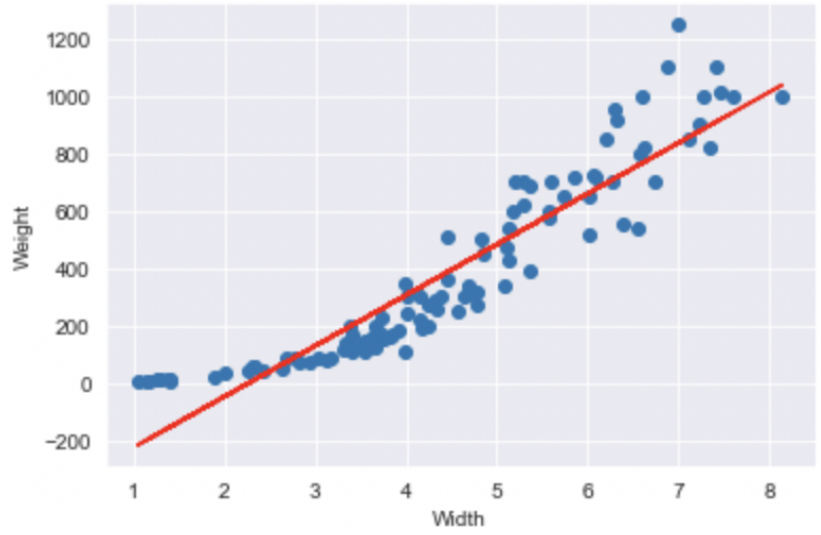
4) Height v/s Weight



Slope: 57.236739324999384
The value of slope is positive which means as the height of the fish increases the weight also increases. It is a positive correlation. It is greater compared to the above three lengths which depicts a stronger relationship between X and Y.

R-Coefficient: 0.6109836532230448
The R value is less compared to the length discussed above which shows that the model does not fit the data very well comparatively.

5) Width v/s Weight [The best fit line]



Slope: 167.77158382823765
The slope is positive and is the highest as compared to all the other independent variables which clearly shows that the relationship between the Width and Weight is very strong and as the width increases the weight also increases.

R-Coefficient: 0.8497451128141986
It also has the highest R value which shows the best fit for the linear regression model compared to all other independent variables.

Step 9 : Execute multiple linear regression

Multiple linear regression is a method that allows us to predict the dependent variable based on multiple independent variables. After applying multiple linear regression the training data the following results were obtained for the model coefficients :

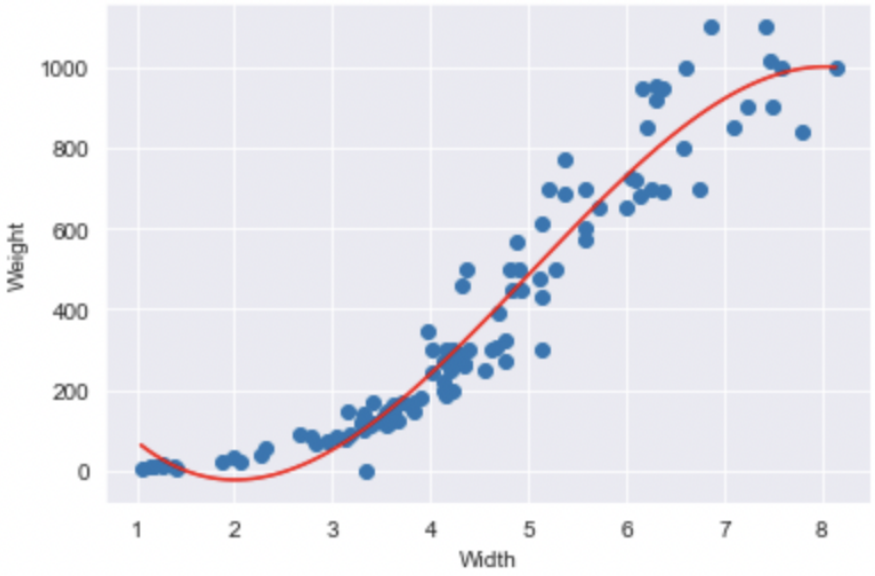
VerticalLen: 79.0812529017546
DiagonalLen: -20.49036551661035
CrossLen: -36.87156860149436
Height: 33.566067111151035
Width: 45.76376378689991

The VerticalLen variable is strongly related to the dependent variable Weight which means as the weight increases the vertical length increases and this is the strongest relationship compared to all the other independent variables and it is a positive correlation. The Diagonal length and the Cross length gives negative values which clearly shows negative correlation that is as the diagonal length and cross length increases the weight decreases.Width and Height also show positive correlation but come after the vertical length in regard to a stronger relationship between the dependent and independent variables. Cross length is the least related as it has a value of -36.87 which is the smallest compared to all.

Step 10 : Execute linear regression with polynomial model

Linear regression with a polynomial model refers to using a linear regression to fit a polynomial function to the data. The independent variables are raised to a power in order to fit a polynomial equation to the data. Here the linear, squared, and cubic versions of the independent variable are used. The correlation matrix is measured to find the relationship between the dependent and the independent variables. The best fit for an independent variable is the one with the higher value of correlation matrix which is the Width(0.916947) against the Weight which is the dependent variable. In the function that is executed to perform the linear regression polynomial model the independent column (Width) is stored in a variable called 'var'. Three new features are created called linear, square and cubic which are raised to 1, 2, 3 respectively. A new dataframe is created which has all the three features. The model is then fitted and values are predicted. The results are plotted on a graph and the best fit 3rd order polynomial is depicted in the graph.

Best fit 3rd order polynomial



Step 11 : Implement Principal Components Analysis

PCA is used to convert a large set of variables into a smaller set of principal components that capture as much of the variation in the original data as possible. The process of PCA involves finding the eigenvectors of the covariance matrix of the original data and they are used to transform the data into a new coordinate system with the eigenvectors as the new axes. Whitening refers to a technique that is used to decorrelate the data by transforming it whereas non-whitened data is the data that has not been transformed to decorrelate the features. The Eigenvalues for non- whitened data are larger than the eigenvalues of whitened data. This is because eigenvalues are a measure of the amount of variation. In a whitened dataset, the eigenvalues will be equal to the variances of the individual features, since the data has been decorrelated transformed to have zero mean and unit variance it has a lower value. The eigenvalues of non-whitened data is larger because it is not transformed to decorrelate the features and value is a combination of both variance and covariance of the feature

Step 12 : Apply PCA on the Fish Market dataset

After applying the PCA function to the fish market dataset the results obtained had significant dimensions. significant dimensions means principal components that have a large amount of the variation in the data. The eigenvalues of the PCA represent the amount of variation in the data. The result has two significant dimensions. The first principal component has an eigenvalue of 4.07921073e-01 in the whitened PCA results and 2.92646905e+02 in the non-whitened PCA results, which is much larger than the other eigenvalues. Also the second principal component has an eigenvalue of 1.55936100e-02 in the whitened PCA and 7.88296547e+00 in the non-whitened PCA results which is larger than the other eigenvalues but it is smaller than the first principal component.

To identify if the variables are correlated or not there are two methods one is to consider the covariance matrix while other is to consider the eigenvectors. In a covariance matrix if the off-diagonal elements are not equal to zero it means that there is a correlation between the variables.When considering the eigenvectors, if two or more variables have large coefficients in the same principal component then they are considered to be correlated variables.

When the features of the data have different variances, whitening is useful.
Whitening the data scales the features which helps in comparing the importance of each feature therefore it makes sense to use whitening in this dataset. If we do not use whitening the principal components will be influenced by scales of the original features. The components will have maximum variance but they will not be directly comparable. In the output both whitened and non-whitened PCA have the same covariance matrix, so both the variables are correlated.

Step 13 : Implement multiple linear regression on the projected data
Using the projected data and running the multiple linear regression the independent variable that is highly correlated is the one that has large absolute values of model coefficients. Those are likely to be the most strongly related to the dependent variable (Weight) which over here is the Width as it has a value of 152.41 which is a positive correlation. The eigenvectors with the largest eigenvalues are most strongly related to the dependent data. Here the first eigenvector, with an eigenvalue of 2.92646905e+02, is the most strongly related to the dependent data. Each eigenvector represents a linear combination of the original variables. The first eigenvector represents the linear combination of the variables with the highest variance in the data.

Step 14 : Extensions - Lasso and Ridge Regression Models
For extension I have performed both Lasso and Ridge regression on the fish market dataset.

Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso) regression is a type of linear regression. It uses L1 regularization which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients where some coefficients can become zero and get eliminated from the model. The larger penalties result in coefficient values closer to zero which is the ideal. It is able to select a subset of the predictors in the model by shrinking the coefficients of the less important predictors to zero.

Ridge Regression

It uses L2 regularization; it only shrinks the coefficients of the predictors but never sets any of them to zero. It prevents overfitting by adding a penalty term to the cost function. The penalty term is the sum of the squares of the coefficients multiplied by a regularization parameter. The coefficients of the model become smaller which reduces the impact of any single feature on the outcome.

The output results for both Lasso and Ridge regression are similar; it just uses different regularization methods L1 and L2 respectively. The model coefficient output answers are different compared to multiple linear regression. Width is the independent variable that is closely related to the dependent variable weight with a coefficient of 152.4179005523065. The diagonal length and cross length have negative correlation whereas all others have positive correlation.

Things Learned:

1) Data Visualization
2) Data Cleaning
3) Splitting in Train and Test CSV's
4) Simple Linear Regression
5) Multiple Linear Regression
6) Linear Regression with Polynomial Model
7) Principal Component Analysis
8) Lasso and Ridge Regression

Summary of the Project :

This project cleared a lot of concepts of linear regression. The fish market dataset was actually very accurate for this project. It had all numerical values and the models were trained accurately. Data cleaning and removing outliers was a very necessary step so as to get the best results after model training. This project helped me understand the differences between the results of linear regression and multiple linear regression. It shows how the independent and dependent variables are related to each other and how minor changes can affect the accuracy of the results. Principal component analysis was a very interesting topic. Looking at different eigenvalues, eigenvectors gave me the understanding of how things varied. Effect of whitening and non-whitening in PCA was helpful to understand the correlated data. The visualisations and graph helped me better understand the data.

References :

1. https://www.telusinternational.com/insights/ai-data/article/10-open-datasets-for-linear-regression
2. https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html
3. https://www.investopedia.com/terms/m/mlr.asp
4. https://www.geeksforgeeks.org/ml-principal-component-analysispca/
5. https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression