Fisheries Data Analysis using Regression and Classification

First Project Report, MATH 8050, Fall 2022

Raghavendra Niteesh Ganugapati, C52595319
Adithya Ravi, C09059838
Srivatsa Kandalam, C75941323
Ashwini Balasubramanian, C31547239

Due October 10, 11:59PM

Abstract

Using the kaggle dataset "Fish Market" to predict the species and weight of fish with the help of other predictors in the dataset.

1 Introduction

The kaggle fish market dataset was chosen for the study. The purpose of using this dataset is to determine the fish's breed and to see if the length, width, and height are relative conditions, which is significant in fishery assessment studies since it offers information on the fish's growth, wellbeing, and fitness in sea habitat. Understanding what species exist and how to identify them is critical for biologists and the general public. When species become extinct, biological variation is lost, and only by understanding species can we change the social, political, and economic processes that drive conservation efforts. The most important takeovers are improvements to fish research in every body of water. Estimation of individual fish species populations in a given water mass. Getting additional information about fish without causing any harm to them.

2 Data

2.1 The Dataset:

The dataset chosen is Fish Market dataset which is in csv form, and taken from Kaggle and the data was updated in 2019. This dataset is a record of 7 common different fish species in fish market sales. The total number of observations is 159. It consists of 7 columns/variables: 1. Species : Species name of the fish 2. Weight : Weight of the fish in g 3. Length1 : Vertical length in cm 4. Length2 : Diagonal length in cm 5. Length3 : Cross length in cm 6. Height : Height in cm 7. Width : Diagonal width in cm Thus the dimensions of this dataset is 159x7. Here Species is a qualitative data while the rest are quantitative data.

2.2 Predictors and Response:

In this project we are trying to predict the weight and species of the fish using the length, height and width of the fish. Thus the predictor variables are Length1, Length2, Length3, Height and Width. The response variables are Species and Weight.

2.3 Access to the Dataset:

The Fish Market Dataset is available in Kaggle. Below is a link to it.

Fish Market Dataset

3 Exploratory Data Analysis

3.1 Exploratory Data Analysis (EDA)

```
library(dplyr) #import dplyr
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)


df_fish <-read.csv("Fish.csv",header = TRUE, sep=",", fileEncoding = "UTF-8-BOM")
head(df_fish)
```

```
##   Species Weight Length1 Length2 Length3  Height  Width
## 1   Bream    242    23.2    25.4    30.0 11.5200 4.0200
## 2   Bream    290    24.0    26.3    31.2 12.4800 4.3056
## 3   Bream    340    23.9    26.5    31.1 12.3778 4.6961
## 4   Bream    363    26.3    29.0    33.5 12.7300 4.4555
## 5   Bream    430    26.5    29.0    34.0 12.4440 5.1340
## 6   Bream    450    26.8    29.7    34.7 13.6024 4.9274
```

```
rows<-nrow(df_fish)
columns<-ncol(df_fish)
print(paste(rows,",",columns,"The number of rows and columns respectively"))
```

```
## [1] "159 , 7 The number of rows and columns respectively"
```

```
summary(df_fish)
```

```
##    Species              Weight          Length1         Length2
##  Length:159         Min.   :   0.0   Min.   : 7.50   Min.   : 8.40
##  Class :character   1st Qu.: 120.0   1st Qu.:19.05   1st Qu.:21.00
```

```
##  Mode  :character   Median : 273.0   Median :25.20   Median :27.30
##                     Mean   : 398.3   Mean   :26.25   Mean   :28.42
##                     3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.50
##                     Max.   :1650.0   Max.   :59.00   Max.   :63.40
##     Length3          Height           Width
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048
##  1st Qu.:23.15   1st Qu.: 5.945   1st Qu.:3.386
##  Median :29.40   Median : 7.786   Median :4.248
##  Mean   :31.23   Mean   : 8.971   Mean   :4.417
##  3rd Qu.:39.65   3rd Qu.:12.366   3rd Qu.:5.585
##  Max.   :68.00   Max.   :18.957   Max.   :8.142
```

The Quantitative in the given data set is Weight, Length1 ,Length2 ,Length3 , Height, Width.
Here Length1 is Vertical Length in cm.
Length2 is Diagonal Length in cm.
Length3 is Cross Length in cm.
Height and Width are the dimensions of the fish.

To Check whether the data set has any NA values.

```
cbind(lapply(lapply(df_fish, is.na),sum))
```

```
##         [,1]
## Species 0
## Weight  0
## Length1 0
## Length2 0
## Length3 0
## Height  0
## Width   0
```

We didn't get any NA values.

Changing the column names

```
colnames(df_fish)<-c("Species","Weight","Verticallength","Diagonallength",
                     "Crosslength","Height","Width")
head(df_fish)
```

```
##   Species Weight Verticallength Diagonallength Crosslength  Height  Width
## 1   Bream    242           23.2           25.4        30.0 11.5200 4.0200
## 2   Bream    290           24.0           26.3        31.2 12.4800 4.3056
## 3   Bream    340           23.9           26.5        31.1 12.3778 4.6961
## 4   Bream    363           26.3           29.0        33.5 12.7300 4.4555
## 5   Bream    430           26.5           29.0        34.0 12.4440 5.1340
## 6   Bream    450           26.8           29.7        34.7 13.6024 4.9274
```

Removing 0 values from the data set.

```
length(which(df_fish == 0))
```

```
## [1] 1
```

We found one zero value, by using filter we are removing the zero value

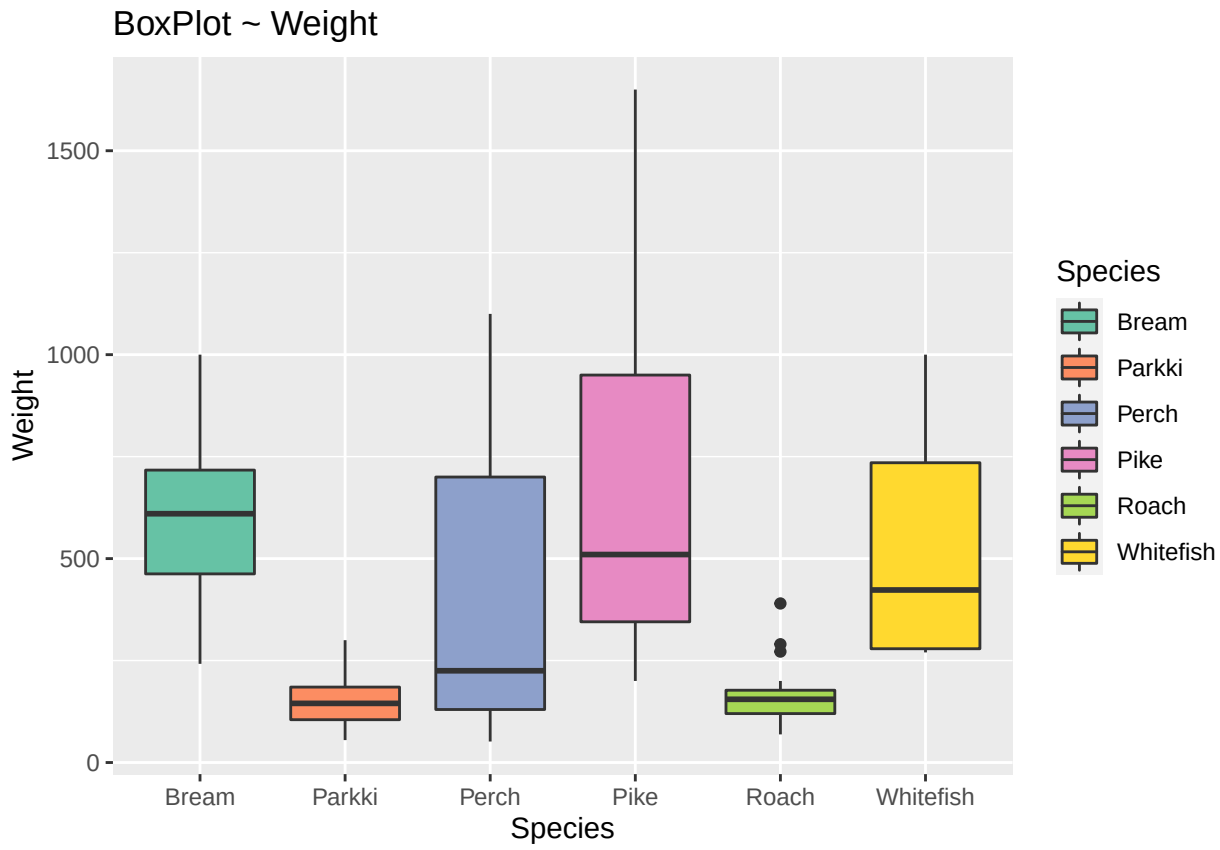
```
df_fish<-filter(df_fish,Weight>0,Verticallength>0,Diagonallength>0,Crosslength>0)

length(which(df_fish == 0))
```

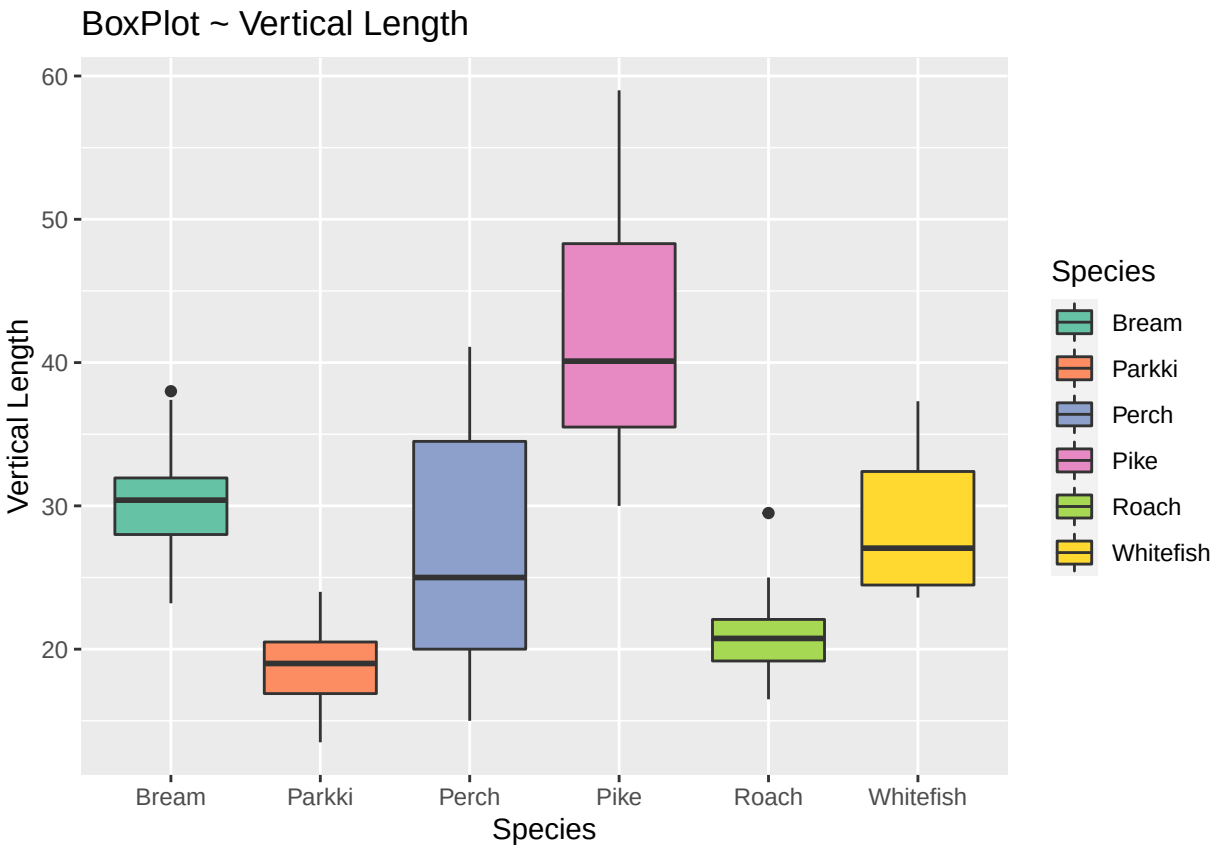
```
## [1] 0
```

```
df_fish<-filter(df_fish,Weight>=50)
```

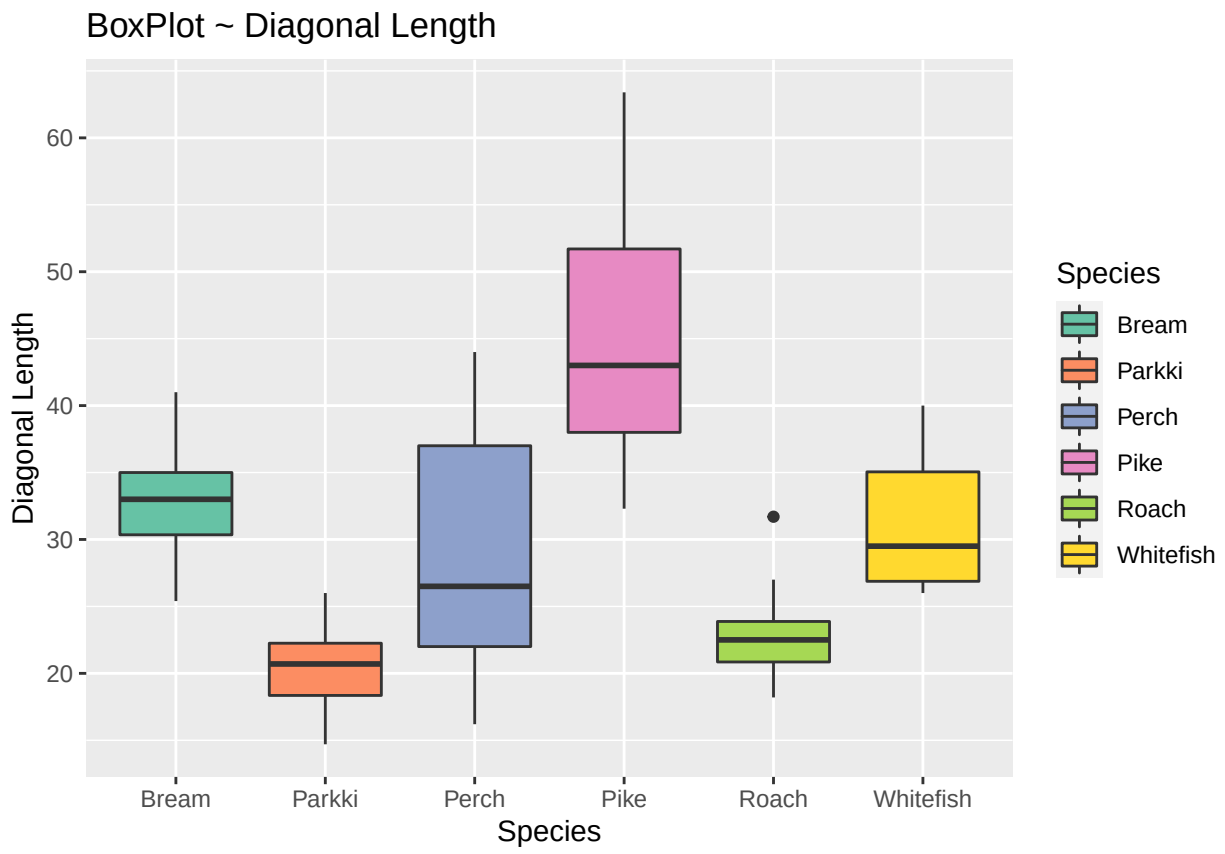
```
ggplot(df_fish,aes(x=Species,y=Weight,fill=Species))+geom_boxplot()+
  scale_fill_brewer(palette = "Set2")+ggtitle("BoxPlot ~ Weight")
```



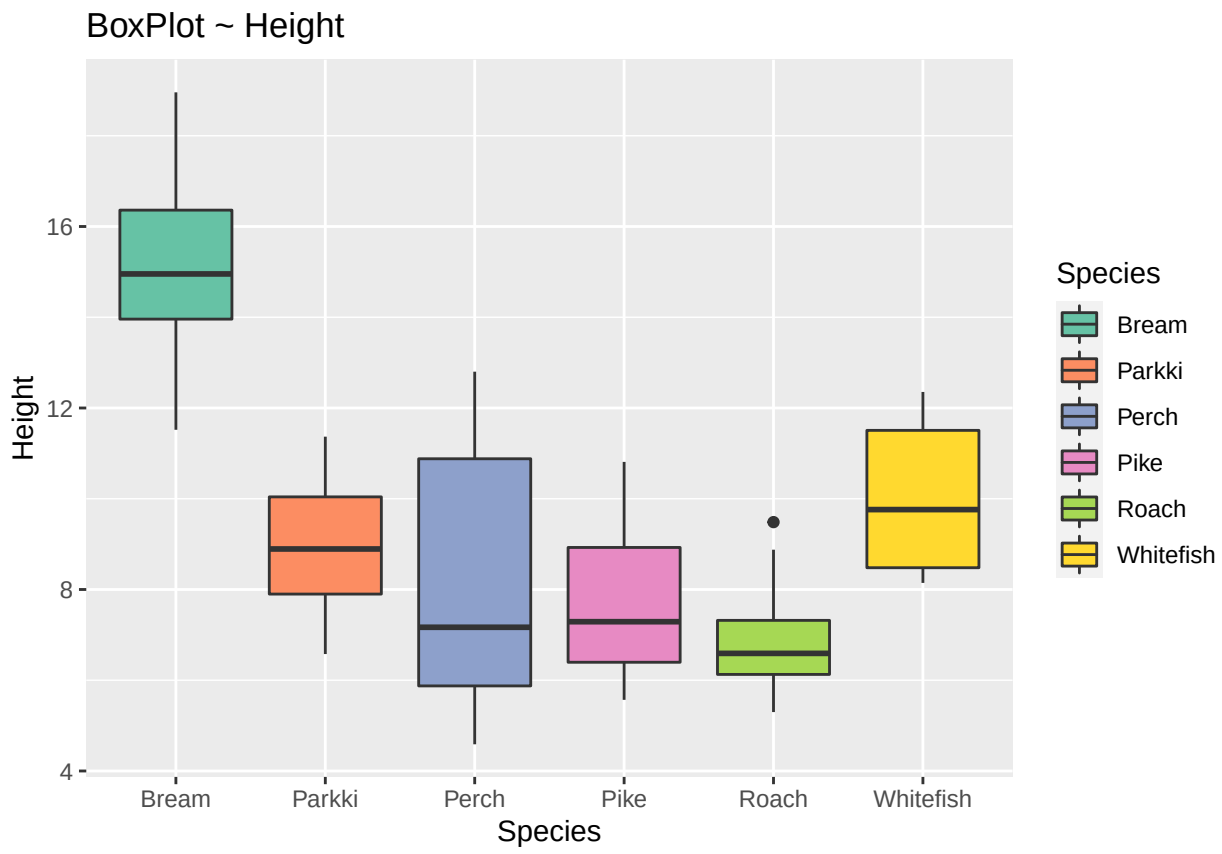

```
ggplot(df_fish,aes(x=Species,y=Verticallength,fill=Species))+geom_boxplot()+
  scale_fill_brewer(palette = "Set2")+ggtitle("BoxPlot ~ Vertical Length")+
  ylab("Vertical Length")
```



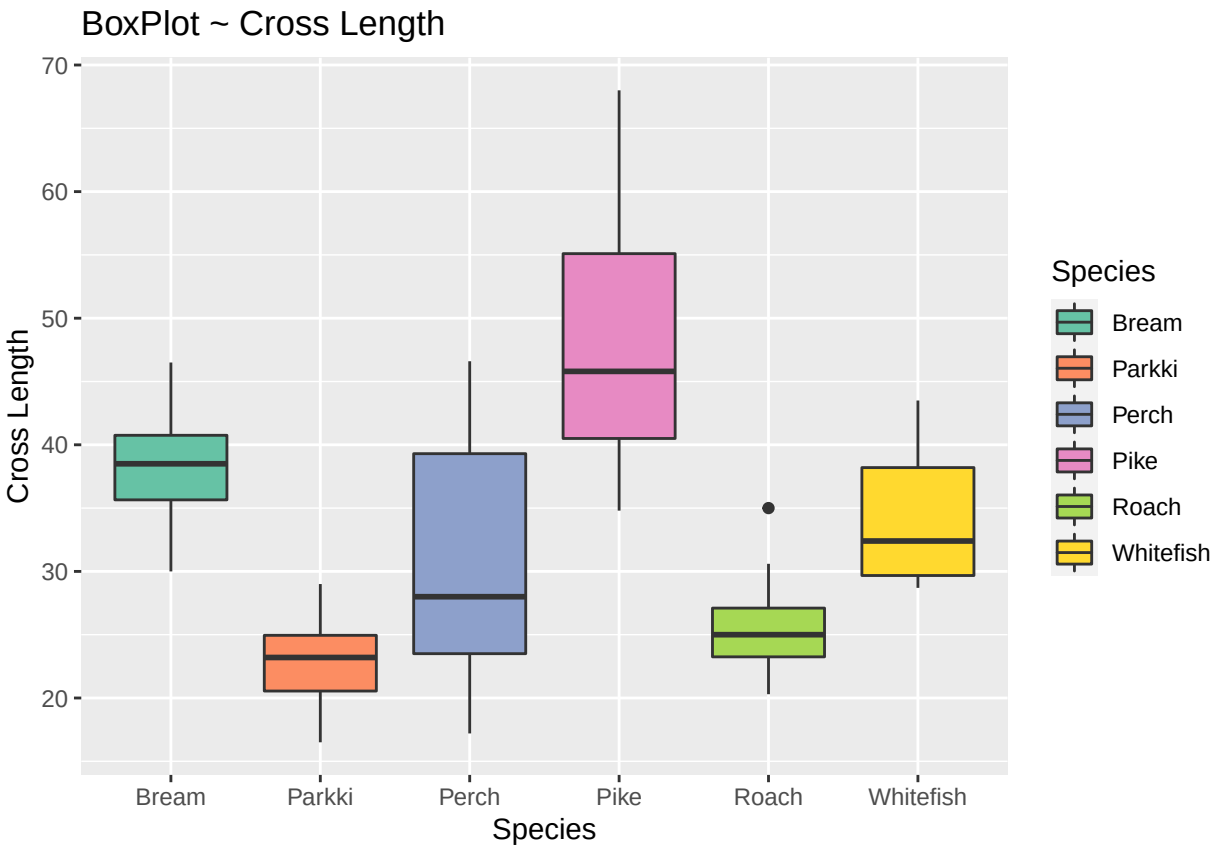

```
ggplot(df_fish,aes(x=Species,y=Diagonallength,fill=Species))+geom_boxplot()+
  scale_fill_brewer(palette = "Set2")+ggtitle("BoxPlot ~ Diagonal Length")+
  ylab("Diagonal Length")
```



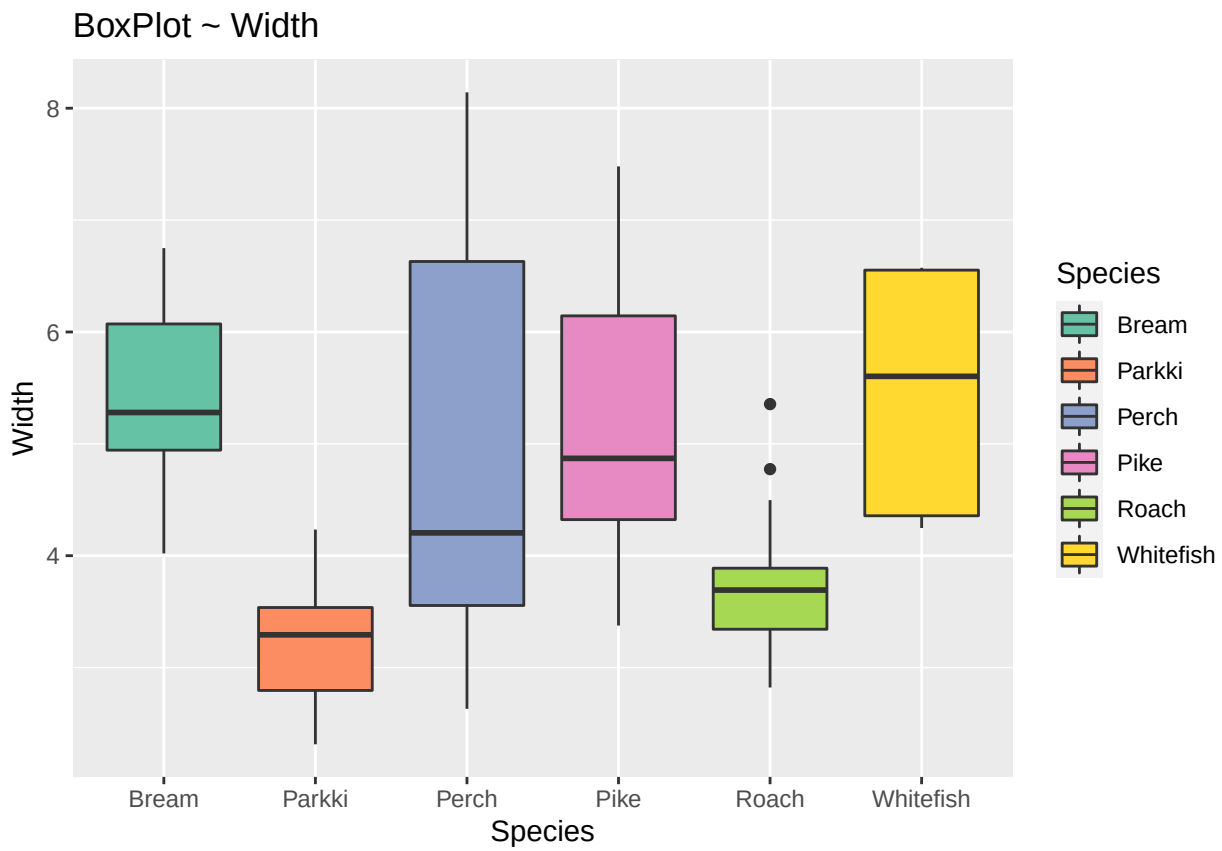

```
ggplot(df_fish,aes(x=Species,y=Height,fill=Species))+geom_boxplot()+
    scale_fill_brewer(palette = "Set2")+ggtitle("BoxPlot ~ Height")+
  ylab("Height")
```




```
ggplot(df_fish,aes(x=Species,y=Crosslength,fill=Species))+geom_boxplot()+
  scale_fill_brewer(palette = "Set2")+ggtitle("BoxPlot ~ Cross Length")+
  ylab("Cross Length")
```




```
ggplot(df_fish,aes(x=Species,y=Width,fill=Species))+geom_boxplot()+
   scale_fill_brewer(palette = "Set2")+ggtitle("BoxPlot ~ Width")
```



We can see that there is at least an outliers for the Roach out of all fish types.
We will use Quantile and try to remove the outliers.

```
df_roach<-df_fish %>% filter(df_fish$Species == 'Roach')
quantile(df_roach$Weight)
```

```
##     0%    25%    50%    75%   100%
##  69.00 120.00 155.00 177.25 390.00
```

We will remove the outliers which above 177.25 in weight.

```
df_roach<-df_roach %>% filter(Weight>177.25)
nrow(df_roach)
```

```
## [1] 5
```

To remove the outliers data.

```
df_fish <-anti_join(df_fish,df_roach)
```

```
## Joining, by = c("Species", "Weight", "Verticallength", "Diagonallength",
## "Crosslength", "Height", "Width")
```

```
nrow(df_fish)
```

```
## [1] 135
```

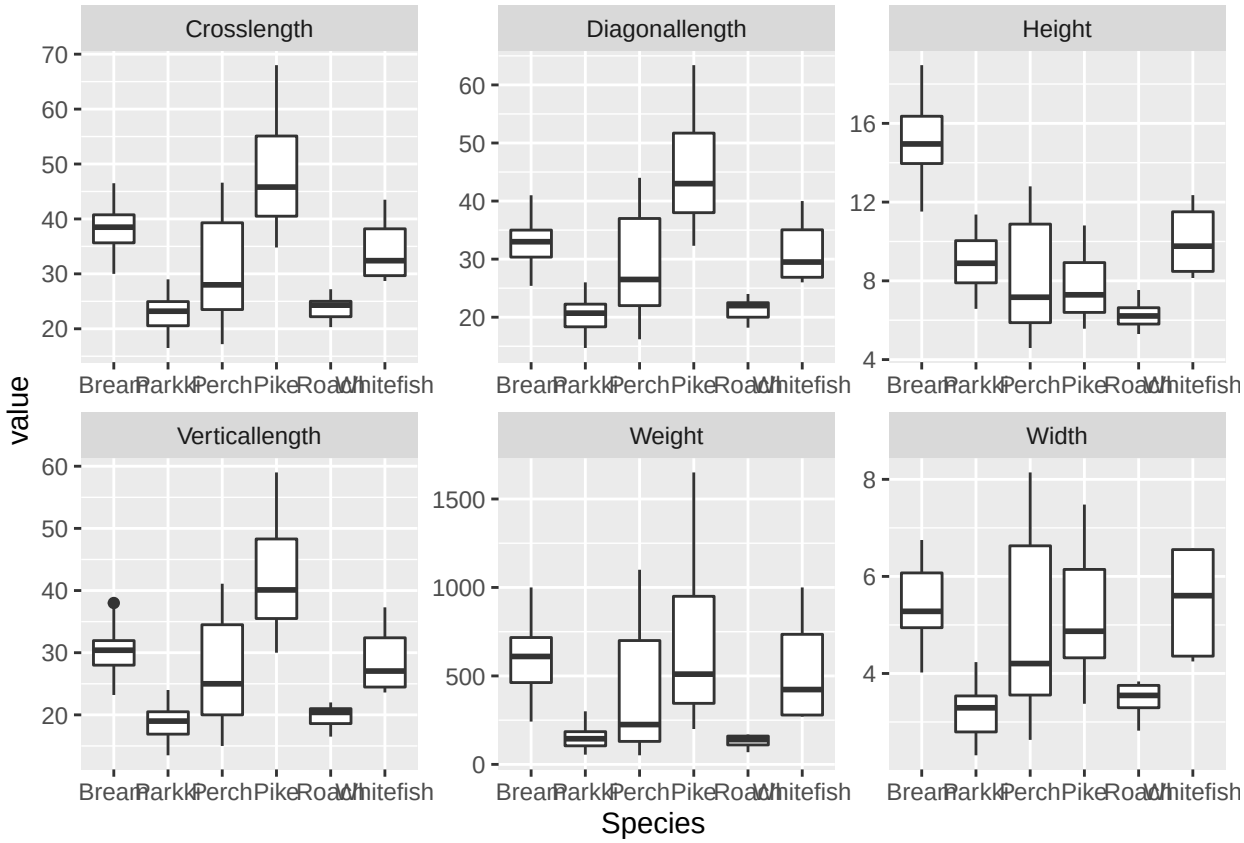
Checking the outliers.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df.boxplot<- df_fish%>% pivot_longer(cols=Weight:Width,names_to="variable",
                                     values_to="value")

ggplot(data = df.boxplot, aes(x = Species , y = value)) + geom_boxplot() +
  facet_wrap(facets = ~variable, scales = 'free')
```

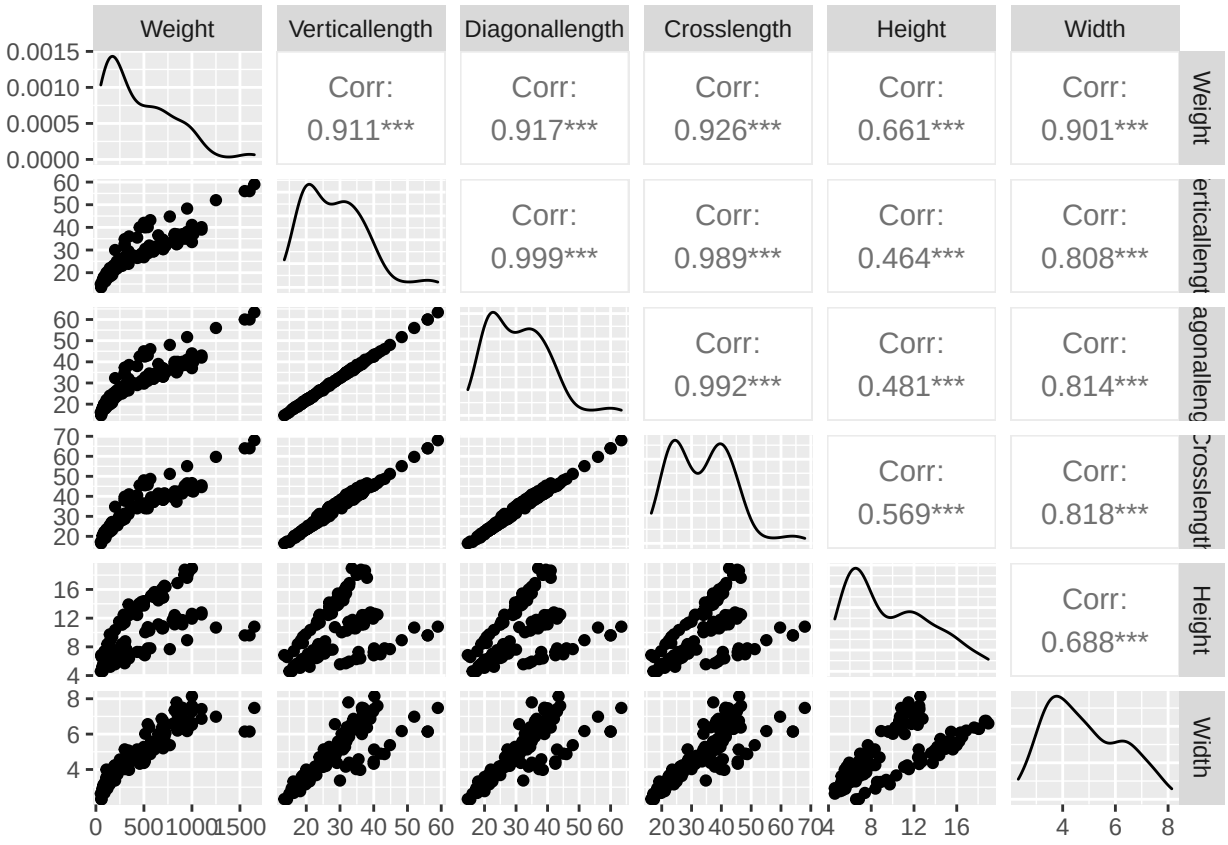


We need to check correlation between all variables.
We are using charts.correlation.

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(df_fish[2:7])
```



From correlation plot we can drop two variables Cross length and Diagonal length.

```
df_fish<-df_fish[-c(4:5)]
str(df_fish)
```

```
## 'data.frame':    135 obs. of  5 variables:
##  $ Species       : chr  "Bream" "Bream" "Bream" "Bream" ...
##  $ Weight        : num  242 290 340 363 430 450 500 390 450 500 ...
##  $ Verticallength: num  23.2 24 23.9 26.3 26.5 26.8 26.8 27.6 27.6 28.5 ...
##  $ Height        : num  11.5 12.5 12.4 12.7 12.4 ...
##  $ Width         : num  4.02 4.31 4.7 4.46 5.13 ...
```

We are left with 135 observations and 5 variables.

3.2 Arguments to support the statistical models/techniques that will be selected to analyze the dataset

For fish market, 2 important information that determine fish price are fish species and fish weight. We would like to predict 1. The species of the fish - Classification 2. The weight of the fish without weighing it - Regression

For the classification problem, we propose to use a naive bayes model. Based on the Bayes theorem, the Naive Bayes Classifier gives the conditional probability of an event A given event B. Some benefits of using the naive bayes model are that is simple, doesn't require much training data, is not sensitive to irrelevant features, is fast and can make real-time predictions.

For the regression problem we use a simple machine learning prediction using Multiple Linear Regression Model based on the information of other columns. MLR allows us to assess the relationship between the outcome and the predictor variables as well as the importance of each predictor to the relationship.

4 Plan for the Second Report

In the second report we are planning to implement two statistical models including the one described at the end of the previous section. This includes using regression to predict the Weight of the fish and classification to predict to Species of the fish. We also plan to refine our model if necessary and test our model by running it with the test data.

5 Explain the contribution of the team member:

1. Niteesh - Data Cleaning
2. Adithya - Outliers Checking after performing the Outlier Elimination based on Weight Variable
3. Srivatsa - Correlation Checking
4. Ashwini - Detecting outliers and checking quantiles

This report was the result of a combined and equal effort from all four team members.