Estimating Shucked Weight for selling Commercial Crabs

Overview

A project submitted due to partial fulfillment of the requirement for the degree of B.Sc Statistics Honours from University Of Calcutta

Submitted by

Aniruddha Mukherjee

CU Roll No: 213115-21-0120

Registration No: 115-1111-0591-21

Semester: VI

Paper: DSE-B2

Under the supervision of

Debjit Majumder

DECLARATION

I, Aniruddha Mukherjee, a student of Semester 6, entitled to the programme B. Sc. Statistics Hons. at Surendranath College , University of Calcutta with CU Regn. No. 115-1111-0591-21 and CU Roll No. 213115-21-0120, solemnly declare that the project entitled "Estimating Shucked Weight for selling Commercial Crabs" is a genuine and original work undertaken by me under the supervision of Sir Debjit Majumder as a part of my Bachelor's programme.

I further affirm that:

● All the data collected for this project is authentic and sourced from reliable and verifiable sources.

● The analysis, interpretations, and conclusions presented in this project are the results of my own research and statistical analysis.

● This project has not been submitted for any other academic degree or assessment at any institution.

● I understand the academic integrity requirements of my university, and I have adhered to all guidelines and regulations regarding plagiarism and research ethics.

● In case of any unintentional oversight or error, I take full responsibility for the same and will cooperate with the necessary revisions or corrections as advised by my faculty or supervisor.

I am sincerely committed to the successful completion of this project and the pursuit of knowledge in the realm of spiritual intelligence assessment.

Signature: ________
Date: ________

❖ Contents

❖ Introduction

- Crabs are primarily aquatic animals, inhabiting, oceans, freshwater habitats, and occasionally venturing onto land. Crabs play important roles in marine ecosystems as scavengers, predators and prey.
- There are many types of species of crab. They are found in a wide range of habitats across the globe. Crabs are very tasty, and it is consumed by people worldwide as a seafood dish.
- In the Boston area, particularly in Massachusetts and the wider New England region, crab is a popular seafood choice and is commonly consumed in various dishes. So, commercial crab farming business is developing the lifestyle of the people of coastal areas. By proper care and management, we can earn more from crab farming business than shrimp farming.

❖ OBJECTIVES

- For a commercial crab farmer knowing the right age of the crab helps them decide if and when to harvest the crabs. Beyond a certain age, there is negligible growth in crab's physical characteristics and hence, it is important to know the time of the harvesting to reduce cost and increase profit.
- The goal of this study is to understand how our target features which Shucked weight is reacting in different situations.
- For a consumer perspective also, it is helpful to know the shucked weight of a crab when given certain measurements to understand better the pricing of the crab or to determine if or if not to buy that crab.

❖ The Dataset

The dataset contains estimated age based on the physical attributes of farmed crab from Boston area. The dataset consists of information on 9 variables for 3893 individual crabs namely.

- **Sex**: Female/ Male /Indeterminate.
- **Length**: Length of the Crab (in Feet; 1 foot =30.48cms)
- **Diameter**: Diameter of the Crab (in Feet; 1 foot=30.48cms)
- **Height**: Height of the Crab (in Feet; 1 foot =30.48cms)
- **Weight**: Weight of the Crab (in ounces; 1 Pound = 16 ounces)
- **Shucked Weight**: Weight without the shell (in ounces; 1 Pound =16 ounces)
- **Viscera Weight**: Weight that wraps around the abdominal organs deep inside body (in ounces; 1 Pound =16 ounces)
- **Shell Weight**: Weight of the Shell (in ounces; 1 Pound =16 ounces)
- **Age**: Age of the Crab (in months)

The source of the dataset is:

https://www.kaggle.com/datasets/sidhus/crab-age-prediction?resource=download

- The column named Sex is divided into three parts: Male, Female, and Indeterminate. An indeterminate sex crab is a crab whose sex cannot be easily identified based on external characteristics.
- All the other remaining columns are continuous in nature where the Length of a crab typically refers to the measurement of the crab’s body from one end to the other.

- The diameter of a crab refers to the measurement across the widest part of the crab’s body.
- The Height of a crab typically refers to the vertical measurement of the crab’s body. This measurement is taken from the bottom of the crab’s body (ventral side) to the top of its carapace (dorsal side).
- The weight of a crab refers to the total mass of the crab. This measurement encompasses the entire crab, including its carapace(shell),legs, claws, and internal organs.
- The weight of crabs is crucial for pricing and selling. Crabs are often sold by weight, and larger, heavier crabs generally fetch higher prices, where the shucked weight (which is in the Shucked weight column) refers to the weight of the crab’s edible meat after it has been removed from the shell. This measurement excludes the weight of the shell, gills, and other non-edible parts.
- The column named as Viscera weight refers to the weight of the internal organs and other non-edible parts inside the crab’s body cavity. This includes components such as the digestive system, gills, reproductive organs and other internal tissues. Higher the viscera weight means lower yield of edible meat from the total weight of the crab.
- The Shell weight (which is in the shell weight column) of a crab refers to the weight of the crab’s exoskeleton, also known as the carapace, along with the legs and claws, excluding the internal organs and edible meat.
- Lastly the column Age refers to the age of a crab in months.

❖ METHODOLOGY

Step 1:

In this analysis, we will examine the dataset if there is some missing value present or not. We will examine the response variable in our study. Also, we will look onto the fact that if any other way of regressing our target is possible or not. By using various data visualization techniques such as Bar Diagram, scatter-plot, heat map, polynomial fitting, trend curve as well as some descriptive measure such as correlation, we will gain an understanding of the features and how they relate to one another.

Step 2:

The next step involves modification of our dataset through several many actions, after getting a fair amount of idea which we have done in step 1; such as getting our action area or in which area of the dataset we are going to work out our analysis.

Step 3:

This step involves another way of looking at our dataset. This involves if there is some other thing which is affecting our response variable. This also involves another process of our analysis which will help our target of study.

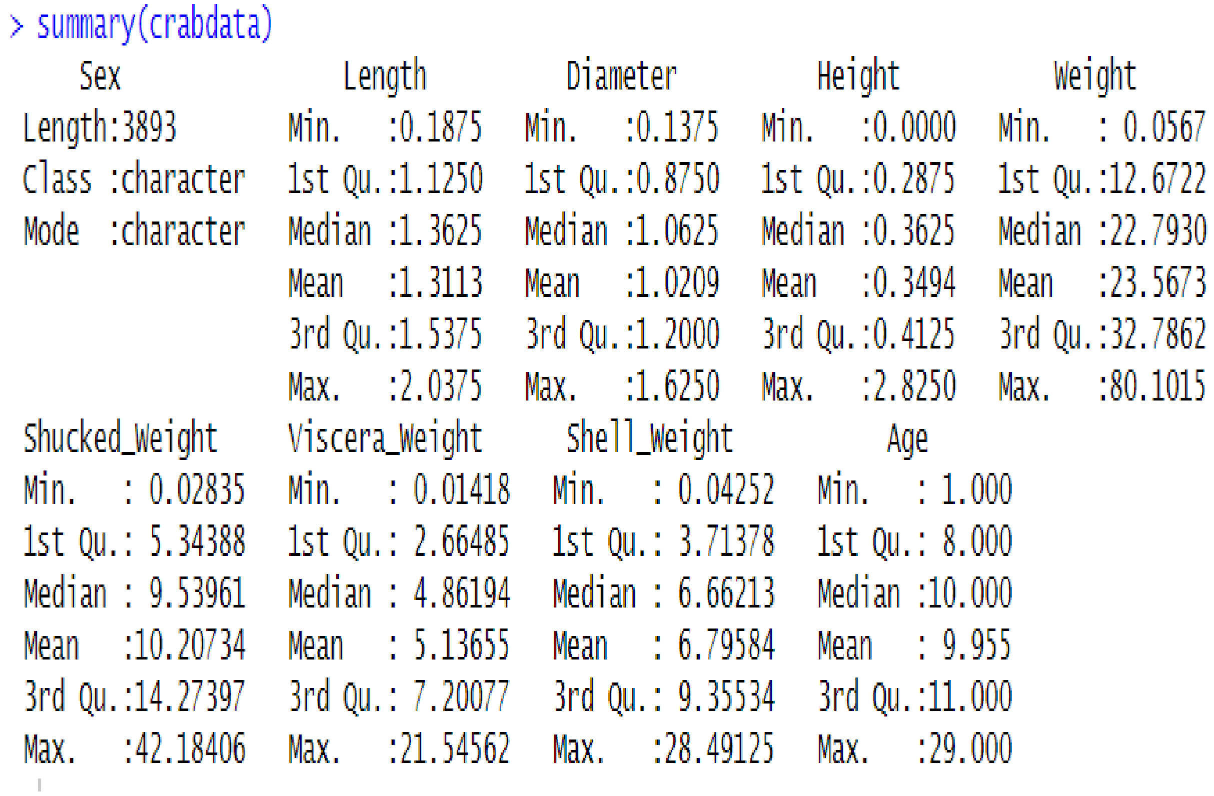
Step 4:

This step involves getting the answer of some real-life problem. Or to simplify the real problem into some simpler form so that in real life process while firming it is going to help.

➢ Summarizing the Dataset

We can observe that there are 3893 rows in the data set and there are 9 columns in the dataset.

- The summary of the dataset is given below.



```
> summary(crabdata)
     Sex                Length          Diameter          Height           Weight       
 Length:3893        Min.   :0.1875   Min.   :0.1375   Min.   :0.0000   Min.   : 0.0567  
 Class :character   1st Qu.:1.1250   1st Qu.:0.8750   1st Qu.:0.2875   1st Qu.:12.6722  
 Mode  :character   Median :1.3625   Median :1.0625   Median :0.3625   Median :22.7930  
                    Mean   :1.3113   Mean   :1.0209   Mean   :0.3494   Mean   :23.5673  
                    3rd Qu.:1.5375   3rd Qu.:1.2000   3rd Qu.:0.4125   3rd Qu.:32.7862  
                    Max.   :2.0375   Max.   :1.6250   Max.   :2.8250   Max.   :80.1015  
 Shucked_Weight     Viscera_Weight      Shell_Weight           Age        
 Min.   : 0.02835   Min.   : 0.01418   Min.   : 0.04252   Min.   : 1.000  
 1st Qu.: 5.34388   1st Qu.: 2.66485   1st Qu.: 3.71378   1st Qu.: 8.000  
 Median : 9.53961   Median : 4.86194   Median : 6.66213   Median :10.000  
 Mean   :10.20734   Mean   : 5.13655   Mean   : 6.79584   Mean   : 9.955  
 3rd Qu.:14.27397   3rd Qu.: 7.20077   3rd Qu.: 9.35534   3rd Qu.:11.000  
 Max.   :42.18406   Max.   :21.54562   Max.   :28.49125   Max.   :29.000  
```

- Here we can observe the median and mean are almost same for each of the variables. The range of variation of these variables are very different. To comment about skewness and other features we need to plot these variables.

➢ Checking For Missing Values

Let us now plot the data in the following graph and check for any missing observations.

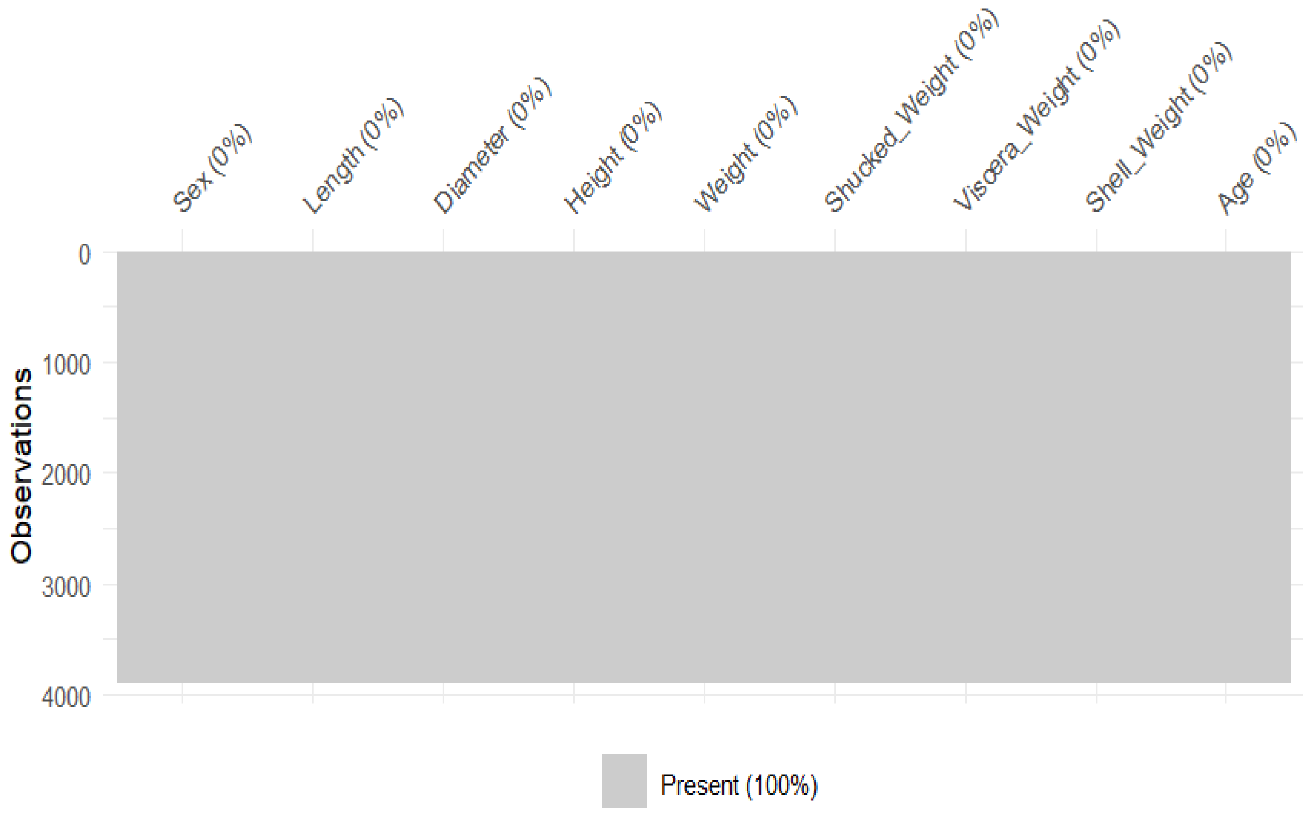


Figure 1: Checking for missing observations in the data.

This plot provides a very specific visualization about the amount of missing data, showing in black the location of missing values and providing the percentage of missing values on the overall data and in each variable. From figure 1 it is evident that there are no missing values in the dataset, so we do not need to change any part of the dataset. Now we will plot these variables to analyze relationship between them.

❖ Exploratory Data Analysis

➢ Counting number of crabs with respect to their sex

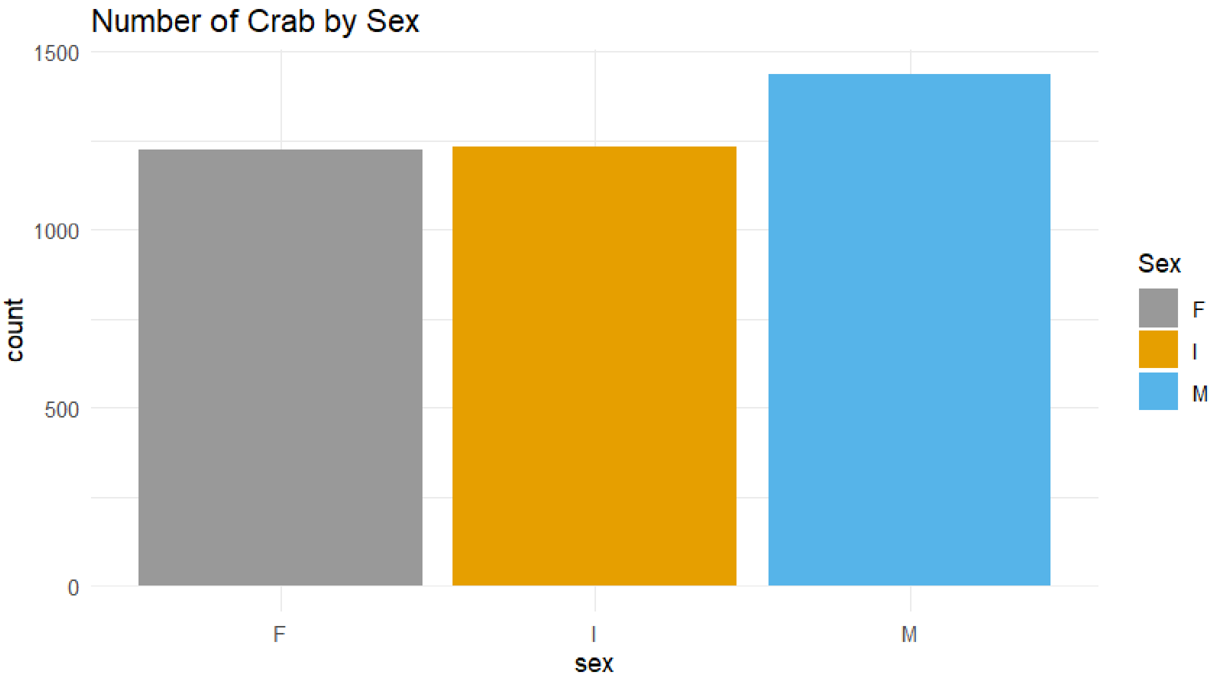


Figure 2: Bar plot counting number of crabs by sex

➢ Observations:

I) In our dataset we can observe that the number of male crabs is highest and the number of female crabs is least.

II) The difference between the number of female crabs and the number of indeterminate crabs is very less. To be precise there are 1225 female and 1233 indeterminate crabs.

➢ Weight at different Ages with respect to gender

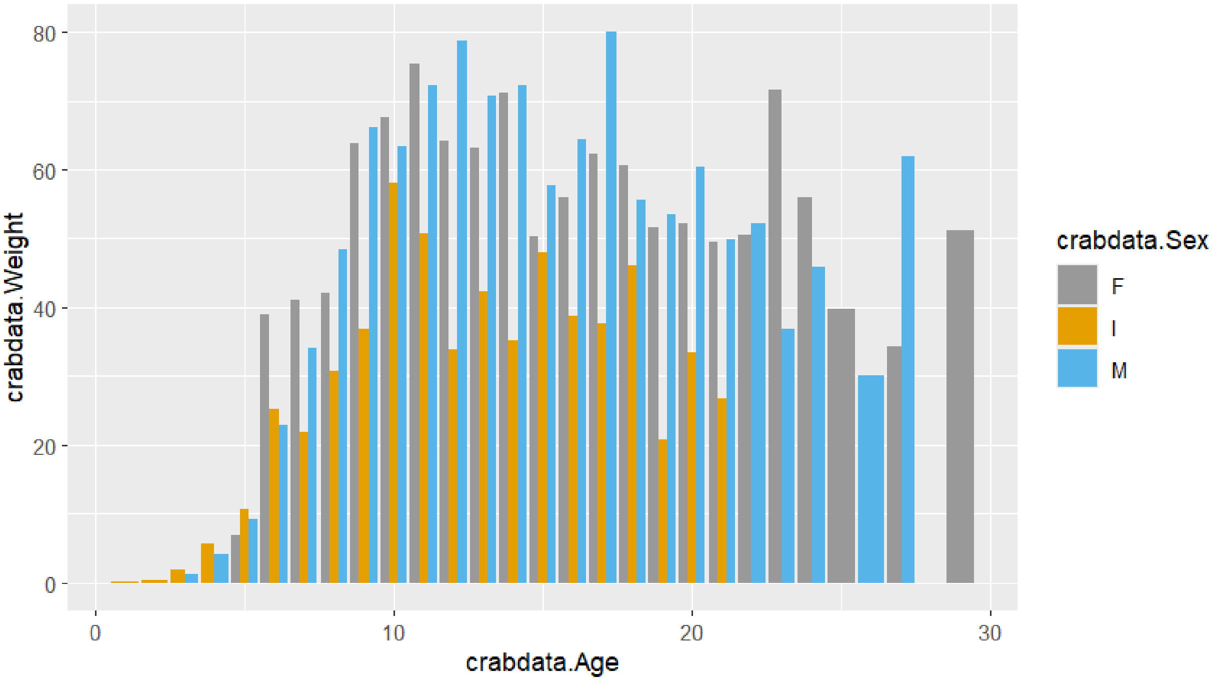


Figure 3: Multiple Bar Diagram Showing growth of weight in different ages.

- ➢ From the figure we can observe that the weight is more for male crabs than female crabs in higher age groups and in majority of the cases and female crabs have more weight in the small age groups
- ➢ Male crabs also have better weight than indeterminate crabs in later age group and in the beginning indeterminate crabs have more weight.

➢ Studying Relationship between Shucked weight and age

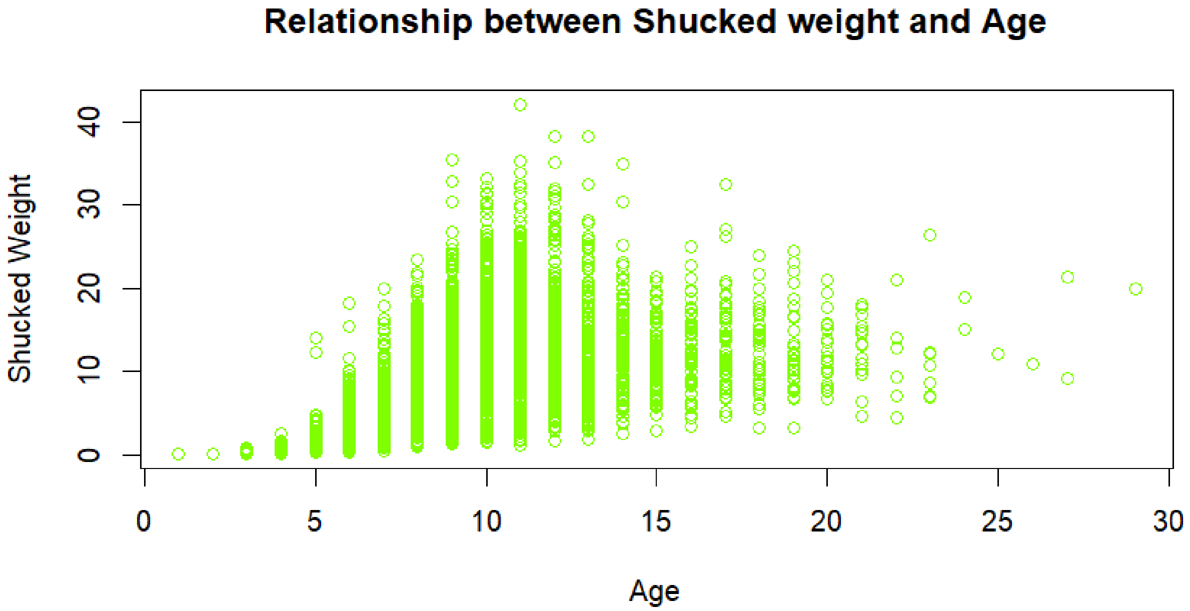


Figure 4: Scatter-plot of shucked weight and age

- ➢ From the scatter-plot we can observe that when the age increases the shucked weight also increases but after a certain time the rate of increase in shucked weight decreases and the increase in shucked weight is not significant enough so after that point keeping that crab will not be profitable, selling that crab in that optimal age will be ideal scenario for the farmer.

Trend Line fitting

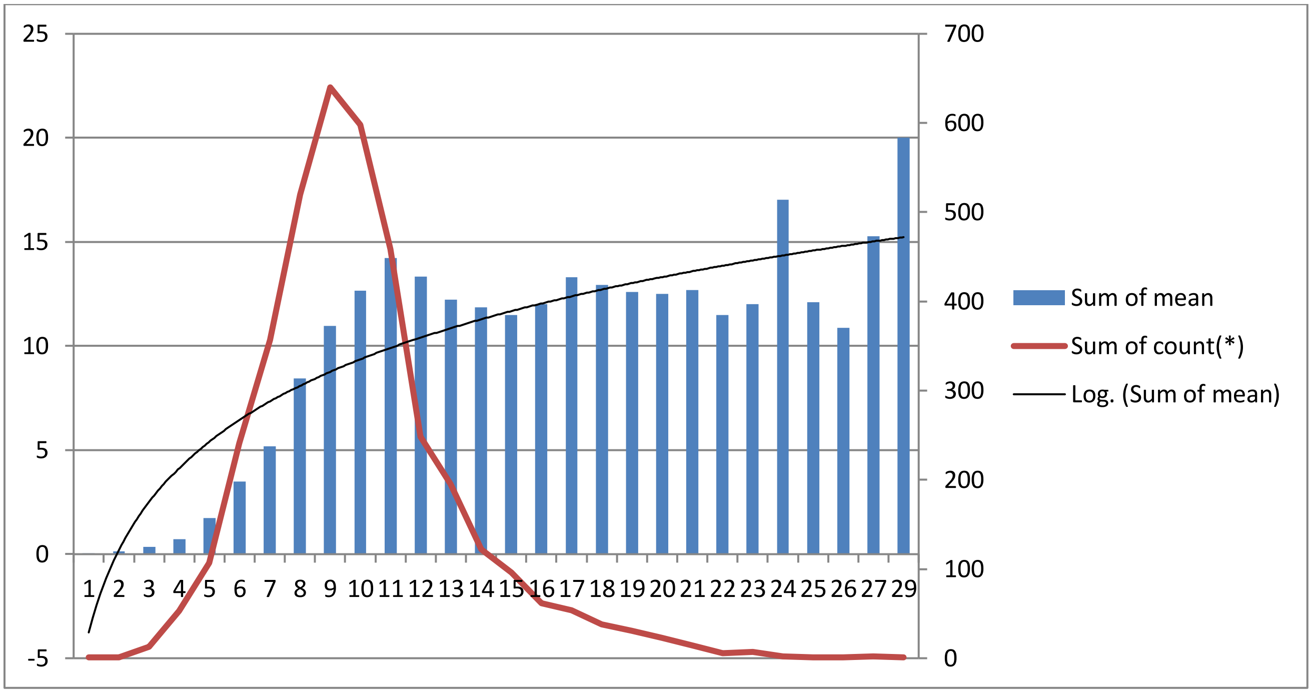


Figure 5: Trend line of average shucked weight corresponding to different age.

- For this figure we grouped the dataset by age and then calculated total shucked weight for a particular age and number of crabs on that particular age and then calculated mean shucked weight for each age and the blue bar represents that mean or average shucked weight for that particular age group.
- Here the brown line represents count of crabs on particular age, we can observe that count of crabs is highest when age is 9 months and count of crabs for age group 6-13 months is more than that of the other age groups.
- From the figure we can observe that average shucked weight is increasing as the age increases but after reaching a certain age the average shucked weight does not increase significantly and we can observe more or less same shucked weight for older crabs also so this tells us about when the crab is young then the shucked weight is rapidly increasing but for older crabs the average shucked weight is not that much increasing.

- We have fitted a trend line over the bar with the help of excel and we can observe that the variables shucked weight and age are not linearly related. The fitted trend line is logarithmic so shucked weight is related to age as Log(age).

➢ Relationship between Diameter and Shucked Weight

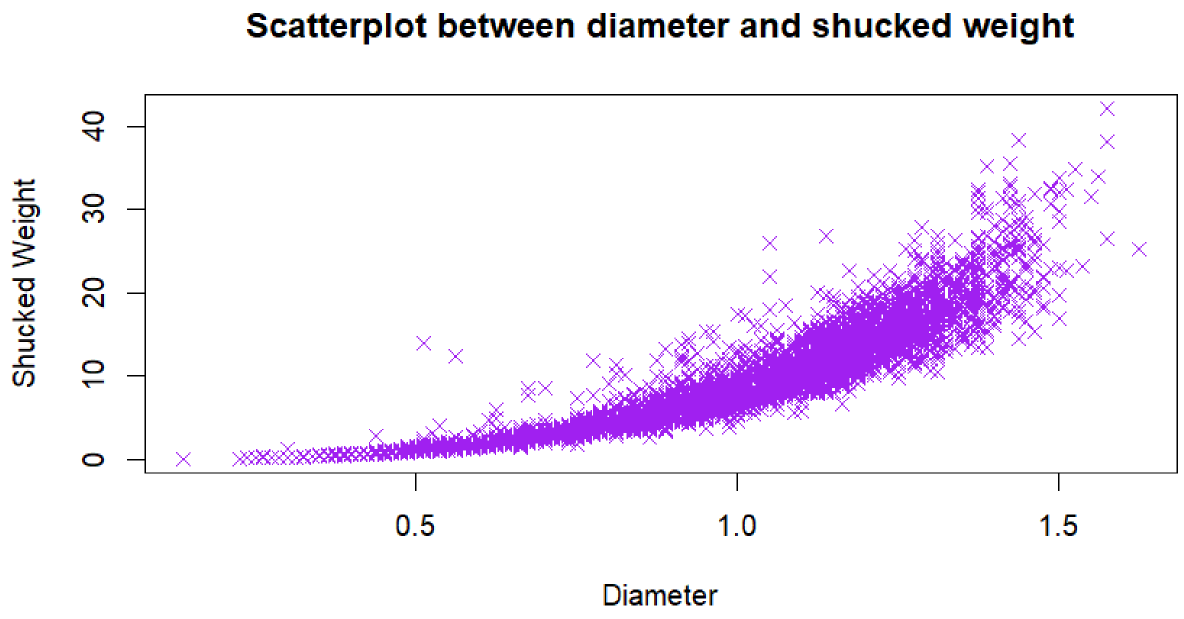


Figure 6: Scatter-plot showing relationship between diameter and shucked weight.

- ➢ From this figure we can observe that as diameter increases shucked weight also increases that if the shucked weight of a crab is high then the diameter of the crab is likely to be high too.

➢ Trend Line Fitting of Diameter and Shucked Weight

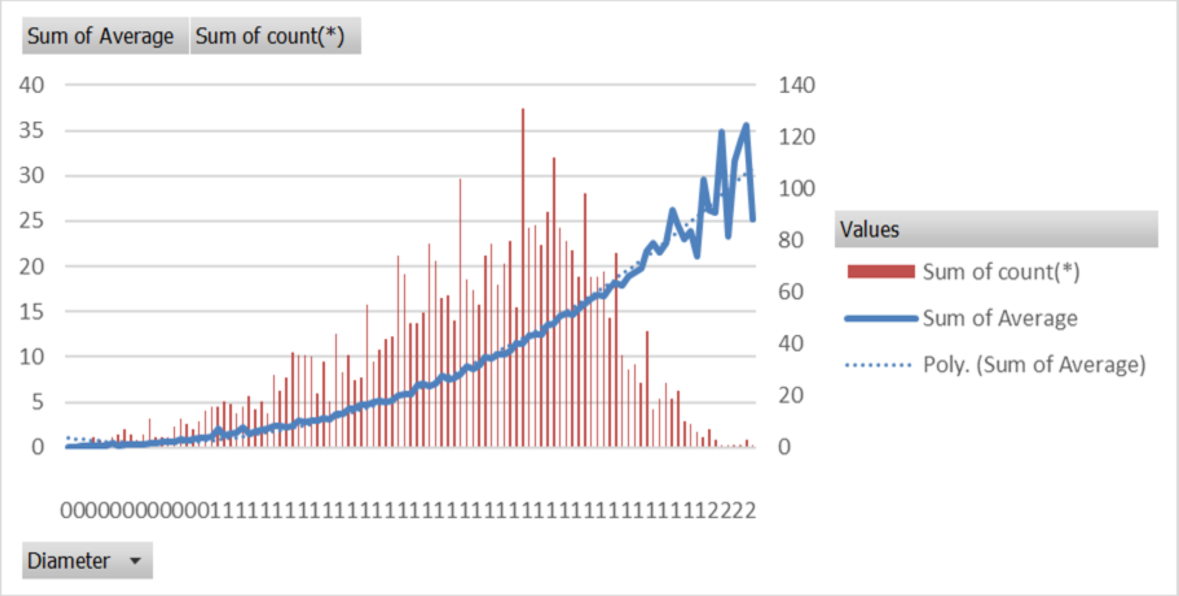


Figure 7: Trend line of average shucked weight and diameter

- ➢ For this figure we grouped the dataset by diameter and then calculated total shucked weight for a particular diameter and number of crabs on that particular diameter and then calculated mean shucked weight for each diameter and the green line represents that mean or average shucked weight for that particular diameter of a crab.
- ➢ Here X axis denotes diameter and Y axis denotes average shucked weight. As we have seen earlier through scatter-plot also as diameter increases average shucked weight also increases and there is little variation towards the end.
- ➢ We have fitted a trend line to the data and here trend line is a polynomial of degree 2 so length is related to shucked weight as $(Diameter)^2$.

➢ Relationship between Shucked weight and length

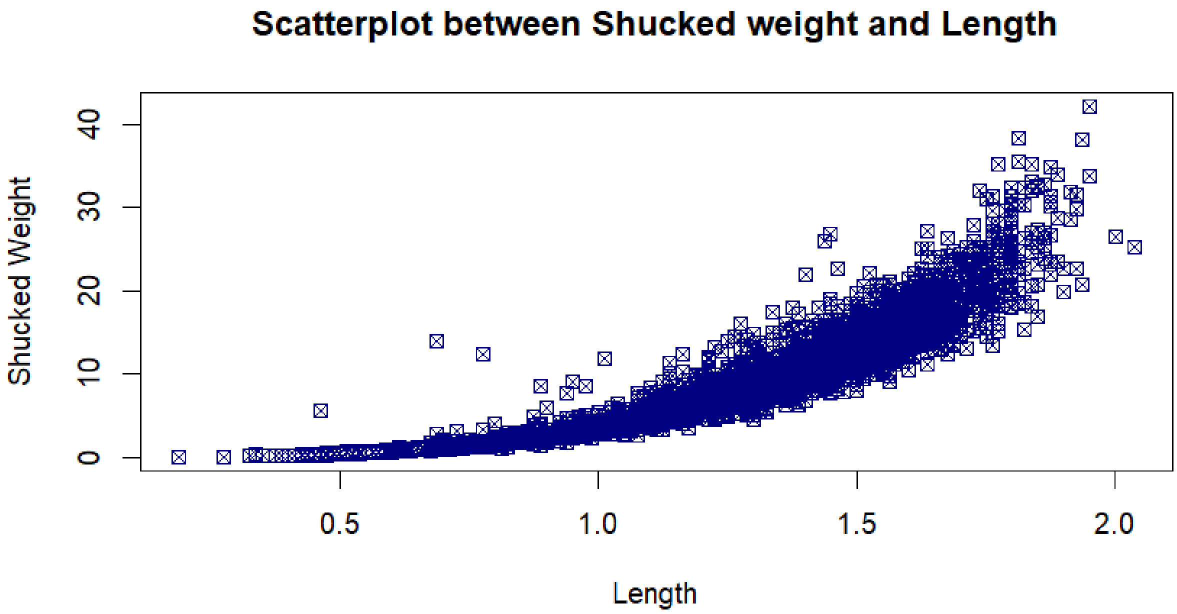


Figure 8: Scatter-plot showing relationship between Length and Shucked Weight

- From this figure we can observe as Length increases Shucked weight also increases.

➢ Trend Line Fitting

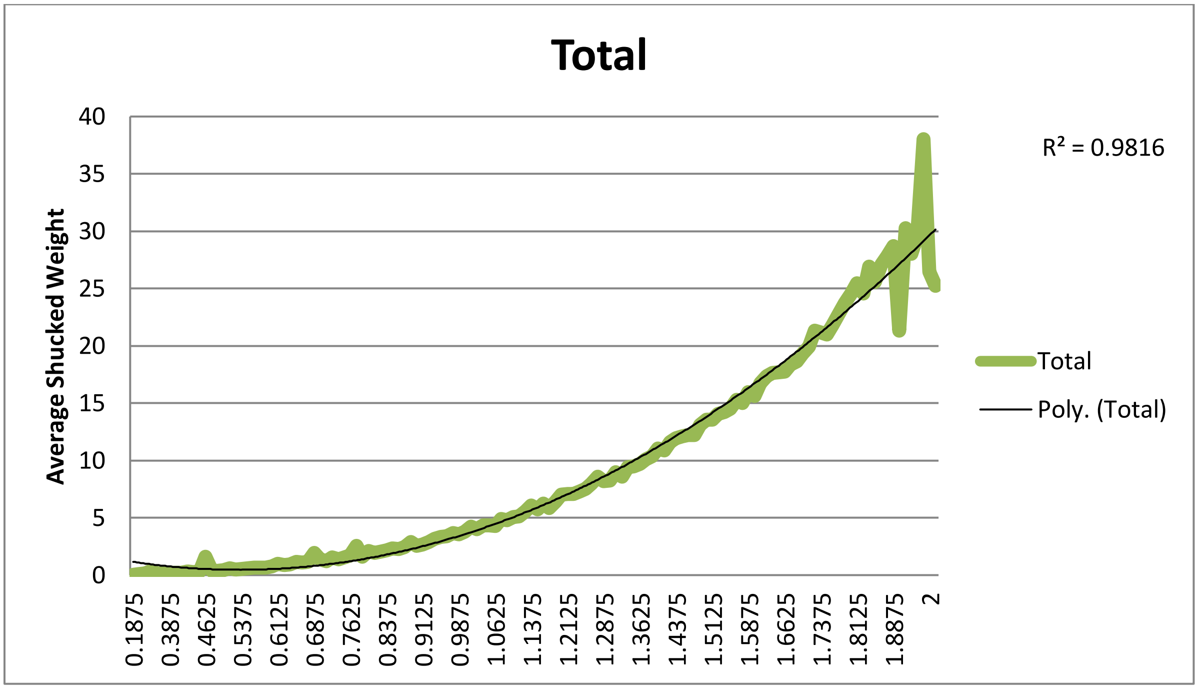


Figure 9: Trend line of average shucked weight and length

- For this figure we grouped the dataset by length and then calculated total shucked weight for a particular length and number of crabs on that particular length and then calculated mean shucked weight for each length and the green line represents that mean or average shucked weight for that particular length of a crab.
- Here X axis denotes length and Y axis denotes average shucked weight. As we have seen earlier through scatter-plot also as length increases average shucked weight also increases and there is little variation towards the end.
- We have fitted a trend line to the data and here trend line is a polynomial of degree 2 so length is related to shucked weight as $(Length)^2$

➢ Relationship between Height and Shucked Weight

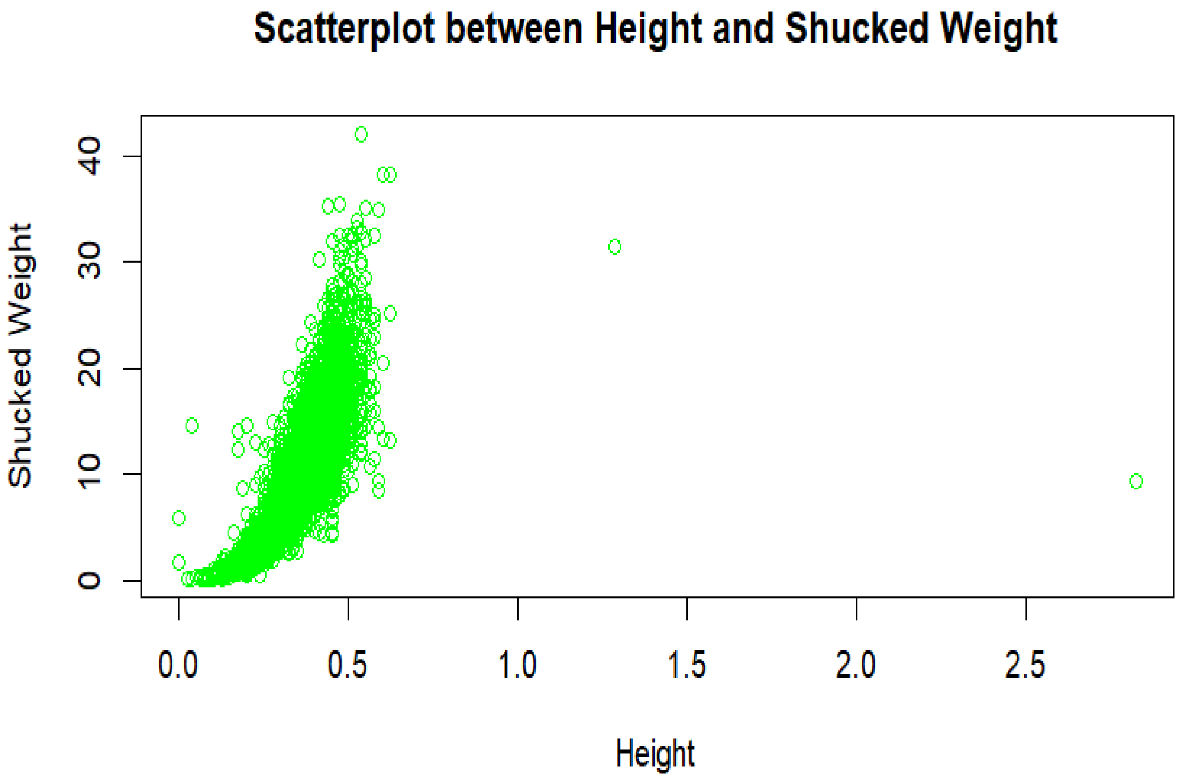


Figure 10: Scatter-plot showing relationship between Height and Shucked Weight

- ➢ From the figure we can observe as height increases the shucked weight increases very rapidly. We can also observe there are some outliers which are far away from other data points. Some crabs have unusually high height and for them the shucked weight is not so high.

Trend Line Fitting

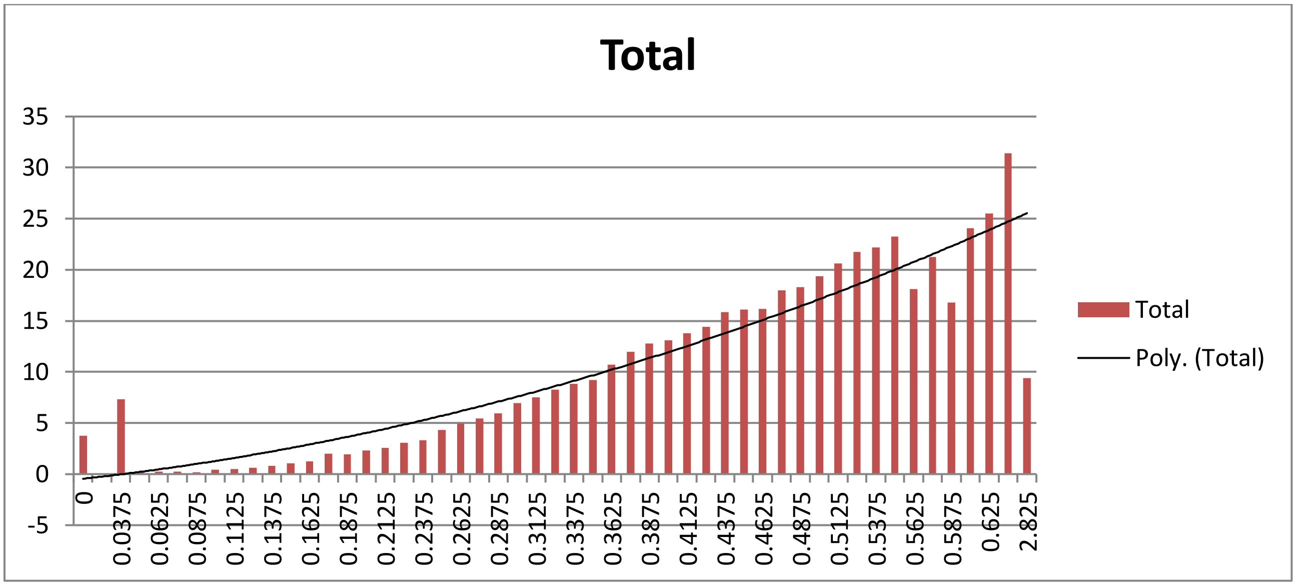
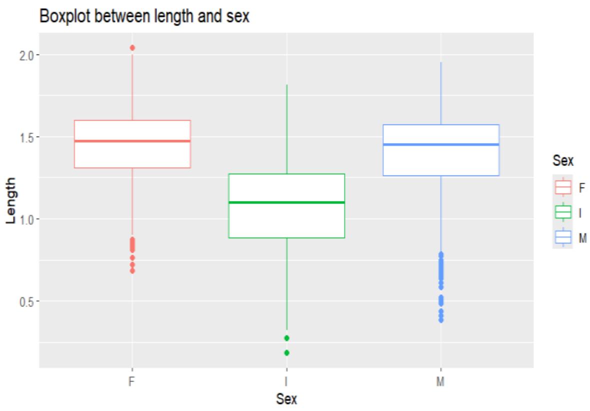
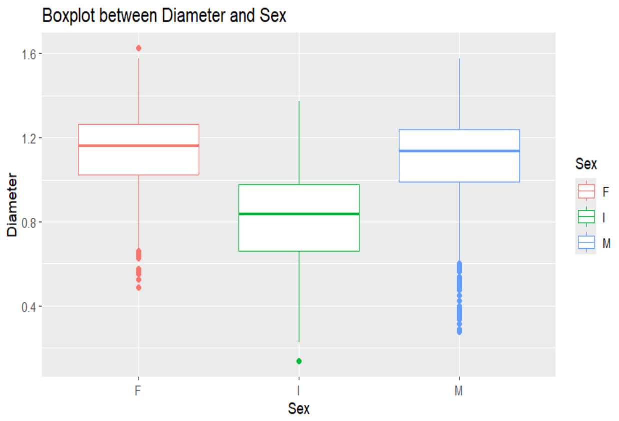


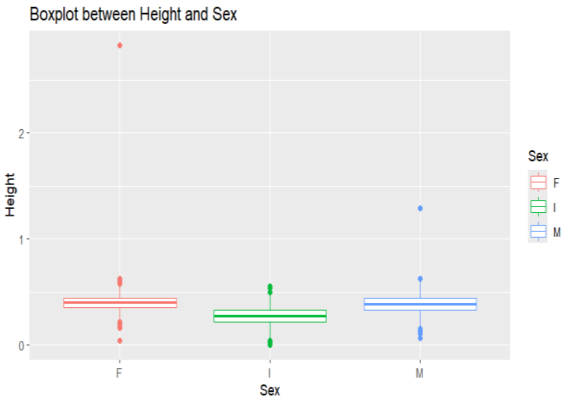
Figure 11: Trend Line of height and average shucked weight

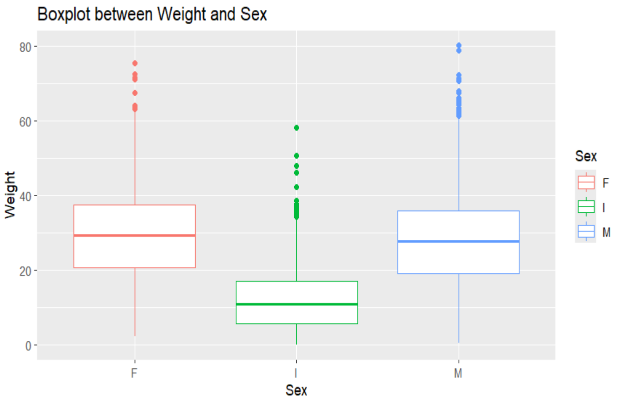
- For this figure we grouped the dataset by height and then calculated total shucked weight for a particular age and number of crabs on that particular age and then calculated mean shucked weight for each age and the brown bar represents that mean or average shucked weight for that particular height.
- Here the Y axis represents average shucked weight and X axis represents height. As we have seen earlier also by scatter-plot this graph also reveals that as height increases average shucked weight also increases.
- Here the black line is fitted trend line, which is a polynomial of degree 2, so we can say that height is related to shucked weight as $(Height)^2$
- Now we should check if there are outliers present in the dataset with the help of box-plots.

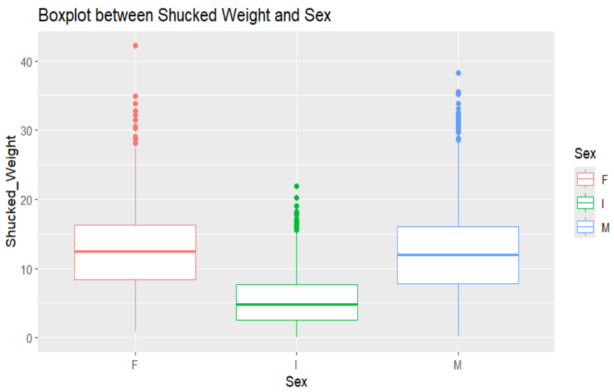
➢ Boxplot

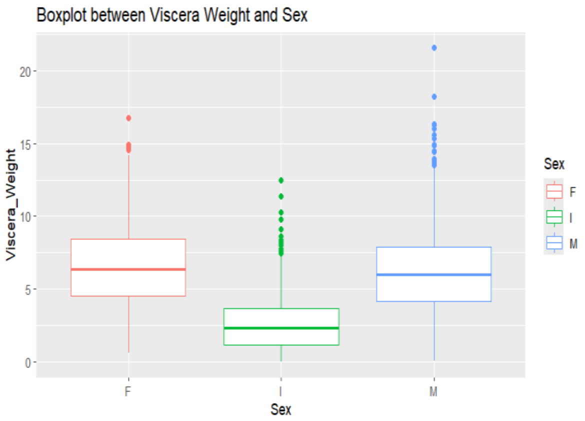

Boxplot between length and sex
Length
2.0
1.5
1.0
0.5
F
I
M
Sex
Sex
F
I
M


Boxplot between Diameter and Sex
Diameter
1.6
1.2
0.8
0.4
F
I
M
Sex
Sex
F
I
M


Boxplot between Height and Sex
Height
2
1
0
F
I
M
Sex
Sex
F
I
M


Boxplot between Weight and Sex
Weight
80
60
40
20
0
F
I
M
Sex
Sex
F
I
M


Boxplot between Shucked Weight and Sex
Shucked_Weight
40
30
20
10
0
F
I
M
Sex
Sex
F
I
M


Boxplot between Viscera Weight and Sex
Viscera_Weight
20
15
10
5
0
F
I
M
Sex
Sex
F
I
M

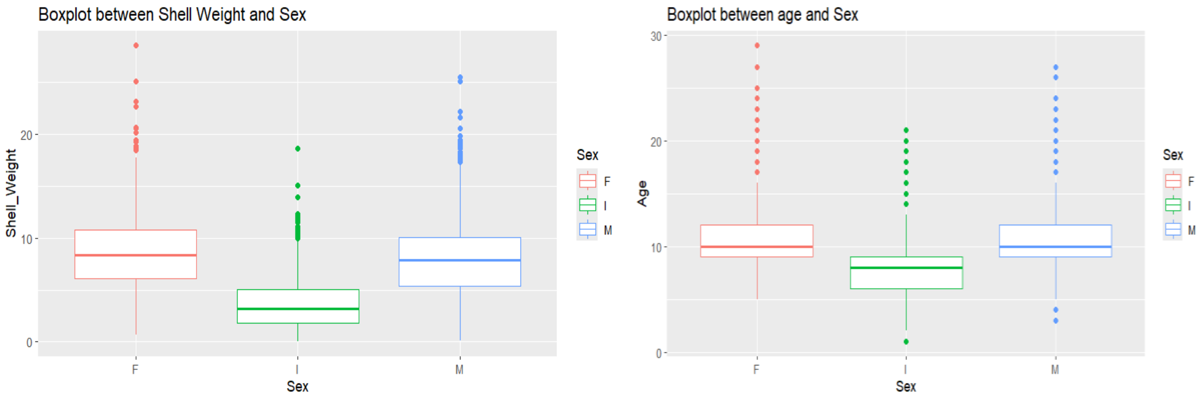


Figure 12: Box-plots of 8 variables and sex

- Here we can observe from the figures that all figures suggest that there are outliers present.
- We can also observe that in all the figures the male and female box-plots are almost same suggesting they have almost same median and 1st quartile and 3 quartile, so they are almost similarly distributed, however Indeterminate crabs have lower median than male and female also 1st and 3rd quartiles are lower as compared to male and female crabs.

➢ Testing Correlation among variables

- Now we should check if a variable is related to some other variable.

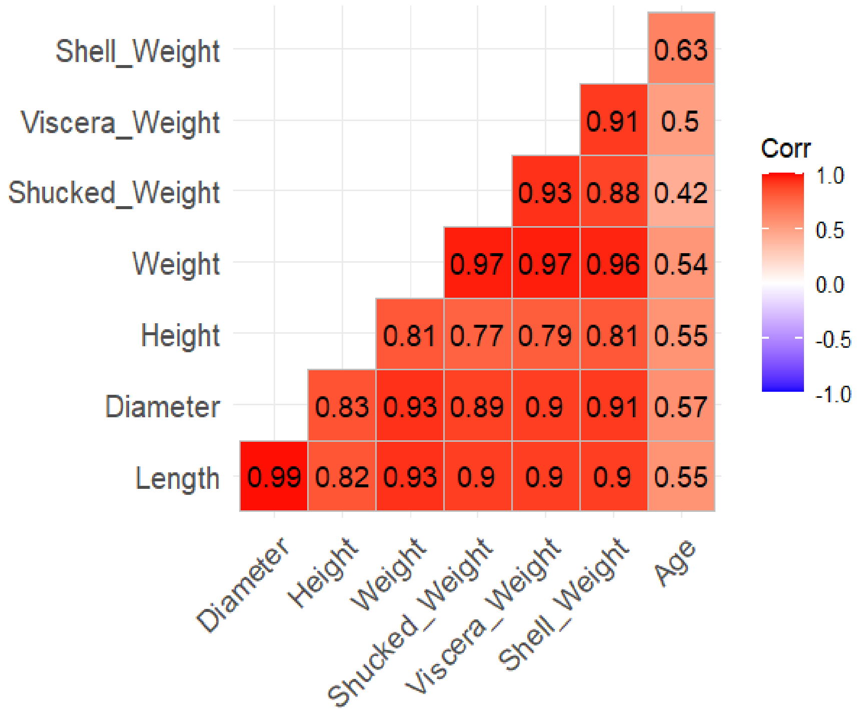


Figure 13: Correlation plot showing relationship among variables

- Here we can observe Age is not directly related to any of the other variables.
- Height is also not strongly correlated with other variables.
- Length and diameter are related to all other variables except age and height, although the relationship with height is good.
- As expected, weight is related with shucked weight, viscera weight and shell weight which is obvious.

Summarization based on the above exploratory data analysis.

From the pictures above we can observe that the relationship between shucked weight and other variables are increasing in nature for small crabs and after reaching a certain age the growth in shucked weight is not as much as for the small crabs. There may be some outlier value but for the majority portion of the dataset after reaching a threshold the growth in shucked weight flattens. Now for better understanding based on above graphs and their conclusions we will divide out dataset into two parts based on length and diameter.

Also, we can see that Age cannot explain Shucked Weight alone in a great extent. So, we have to consider another variable or the possible combination of variables to see how it is affecting our Shucked Weight.

As we observe that for the crab which is small the Shucked Weight is increasing in nature but for big crab the trend line of the Shucked weight id not increasing in nature, but the line became flatter. Considering this we will make our analysis different for two types of crabs. One is for small another is for big.

For this, the crabs for which the length is less than median(length) we will call them small crabs and if length is greater than median(length) we will call then large crab and the terminology will be same based on diameter also.

- At first, we divided our dataset in two portions with respect to length. We calculated the median length of all the crabs, and it comes out to be 1.3625. Now we are calling the crabs whose length is smaller than 1.3625 is small crabs and the crabs for which the length is greater than 1.3625 are big crabs.
- Now we will plot the data for the small crabs and try to explore the relationship between age and shucked weight for these small crabs and to visualize this we will use some graphs.

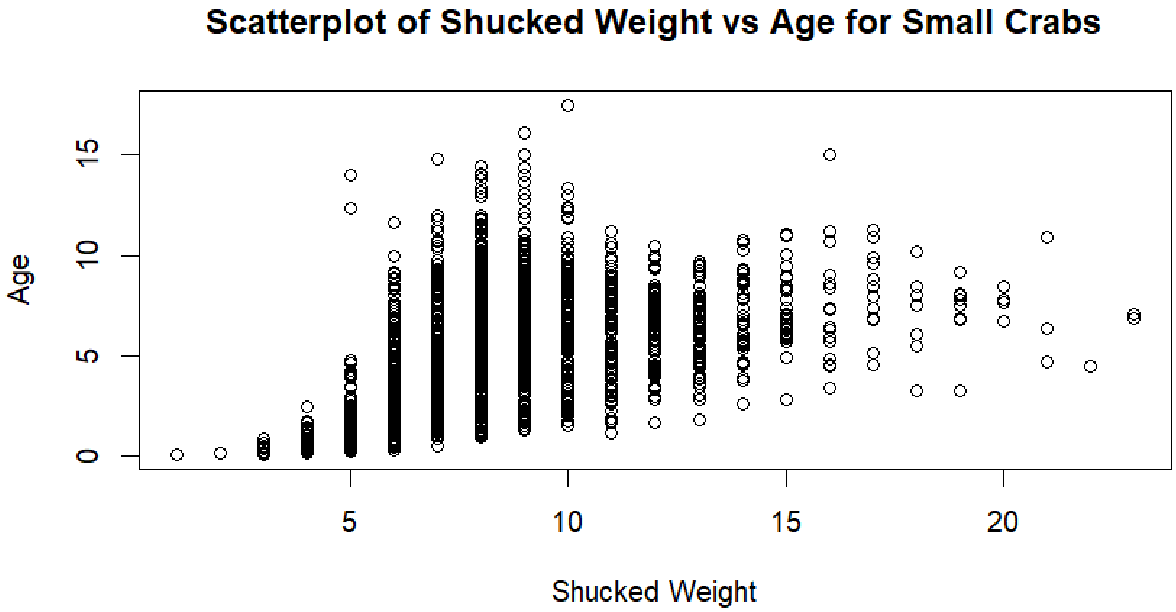


Figure 14: Relationship between shucked weight and age for small crabs

- We can observe for small crabs as the age increases the shucked weight also increases and they have a positive correlation as it seems.
- Now we will fit a trend line to observe it more clearly.

➢ Fitting a trend line

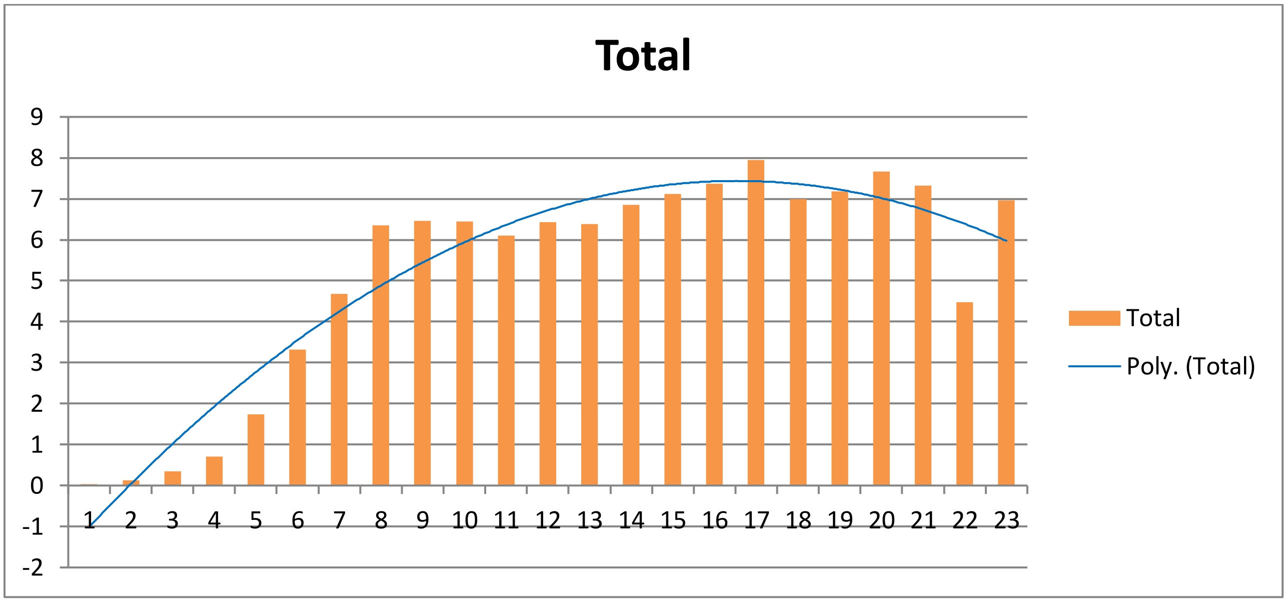


Figure 15: Trend Line fitting on average shucked weight vs age

- For small crab data we have at first calculated average shucked weight grouping by age and then plotted the data and added a trend line.
- Here the x axis denotes the age of crab in months and y axis denotes the average shucked weight.
- Now let us check the same for big crabs also how it changes by plotting same graphs.

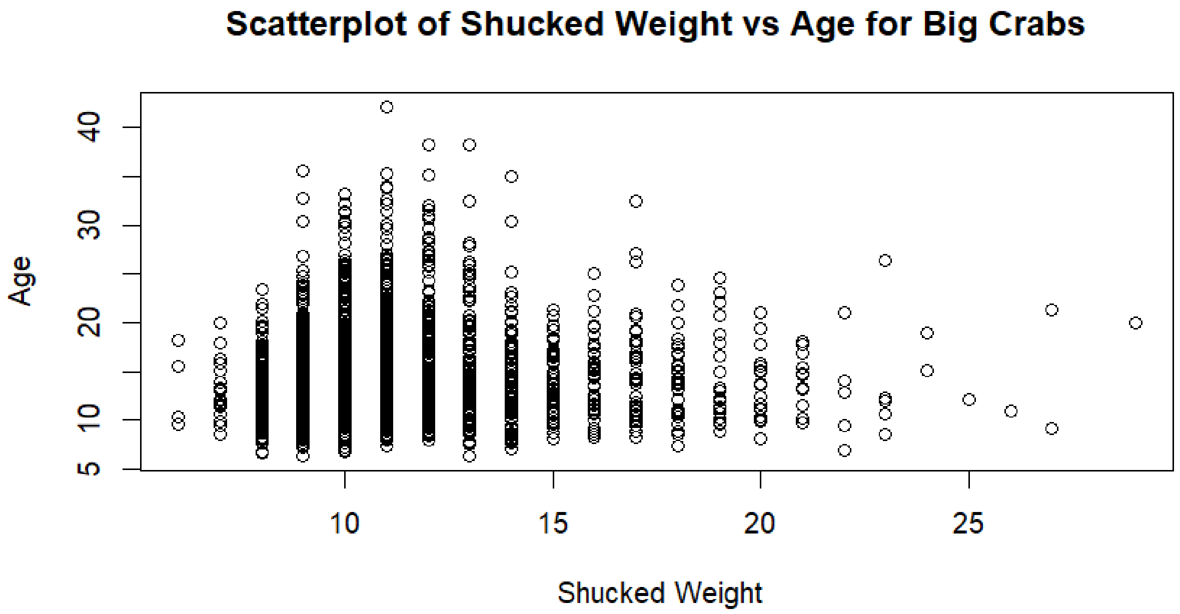


Figure 16: Relationship between shucked weight and age for big crabs

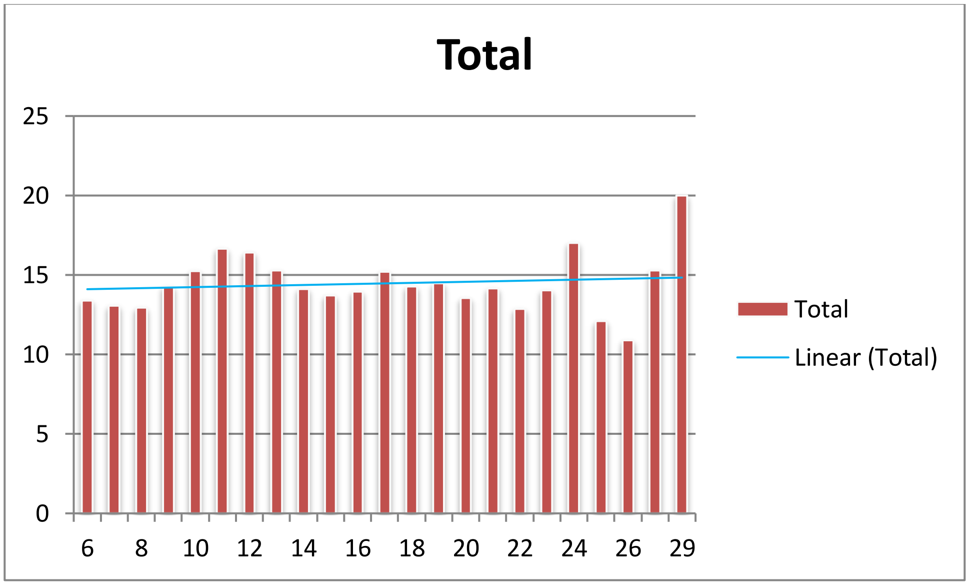


Figure 17: Trend line fitting on average shucked weight vs age

- Here for big crabs at first grouping by age we calculated average shucked weight and then we plotted average shucked weight vs age and added a trend line.

- By the above figures we can observe that for big crabs as the age increases the shucked weight remains more or less the same. So as the age increases the increase in shucked weight stabilizes for big crabs.
- Now we will check these same characteristics by dividing the dataset similarly by diameter. Now we will say those crabs as small crabs whose diameter is less than the median diameter and those crabs whose diameter is greater than the median diameter, we will call them big crabs. Now we will plot similar figures as above to analyze the relationship of shucked weight with age for these new two datasets and make some comment on them.

➢ Scatter plot of Shucked Weight Vs Age for Small Crabs

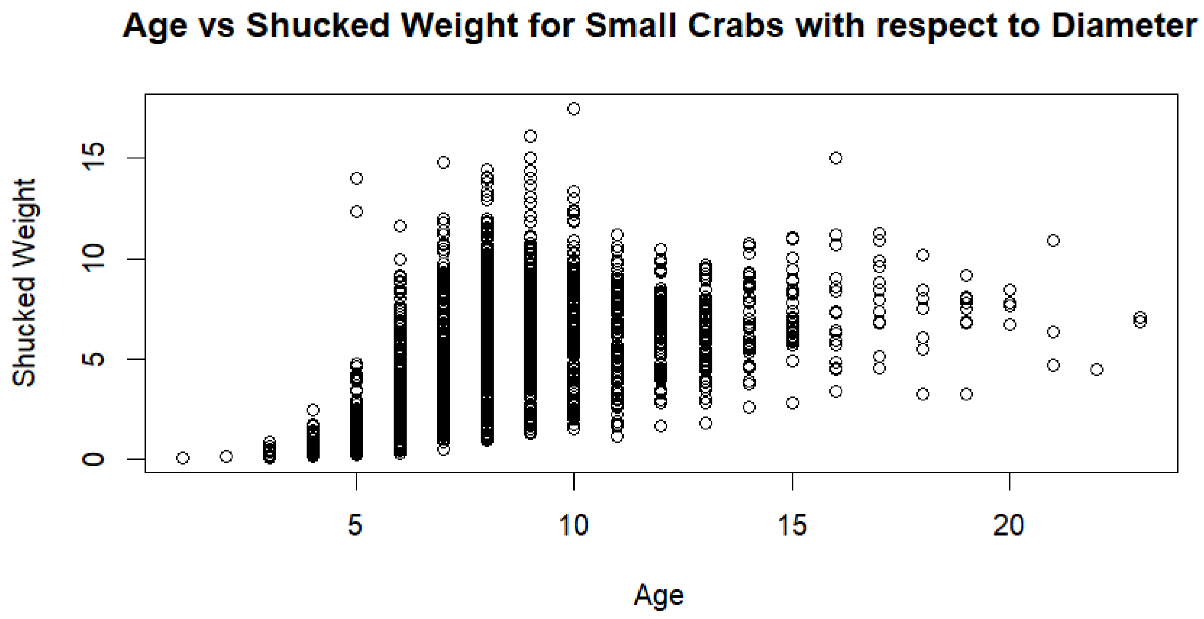


Figure 18: Age Vs Shucked Weight of crabs with diameter< Median(diameter)

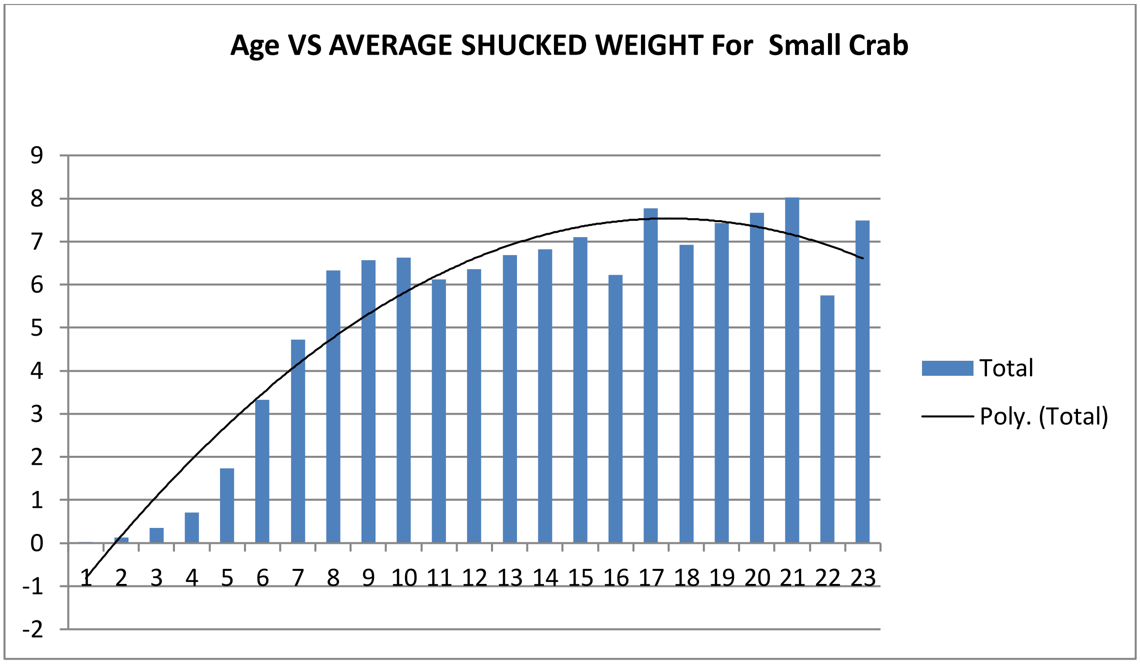


Figure 19: Trend line fitting for small crabs whose diameter is less than the median diameter

- Here the x axis denotes the age in months and y axis denotes the average shucked weight for crabs whose diameter is less than the median diameter.
- From the above figures it is evident that as the age increases the shucked weight also increases rapidly for small crabs. A polynomial of degree 2 was fitted.
- Now we shall check the same for big crabs that is the crabs whose diameter is greater than the median diameter.

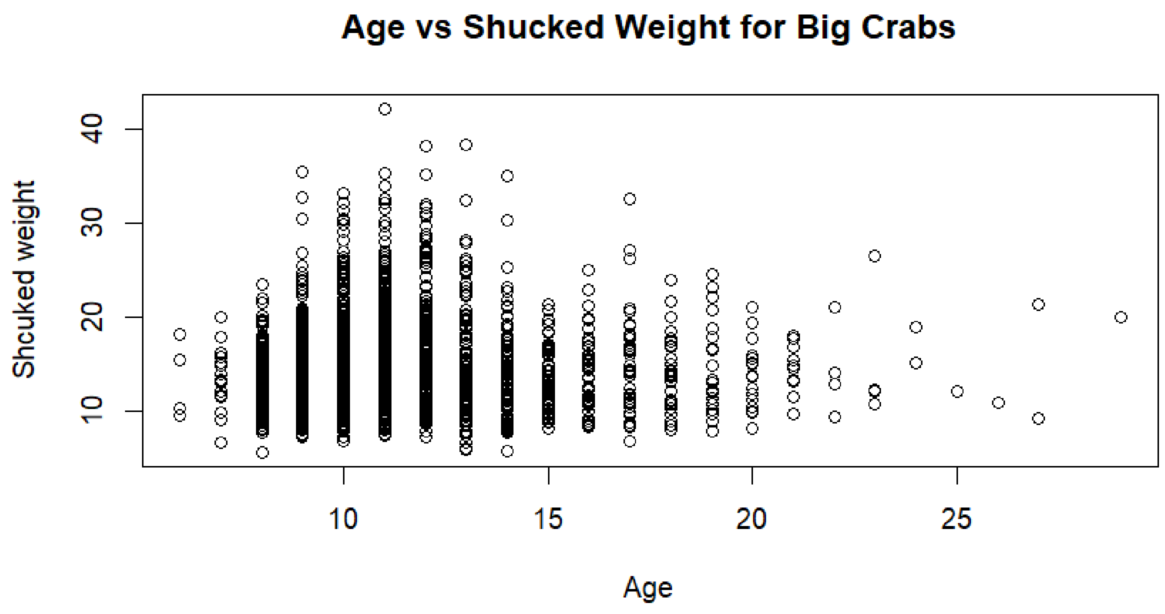


Figure 20: Age Vs Shucked Weight of crabs with diameter> Median(diameter)

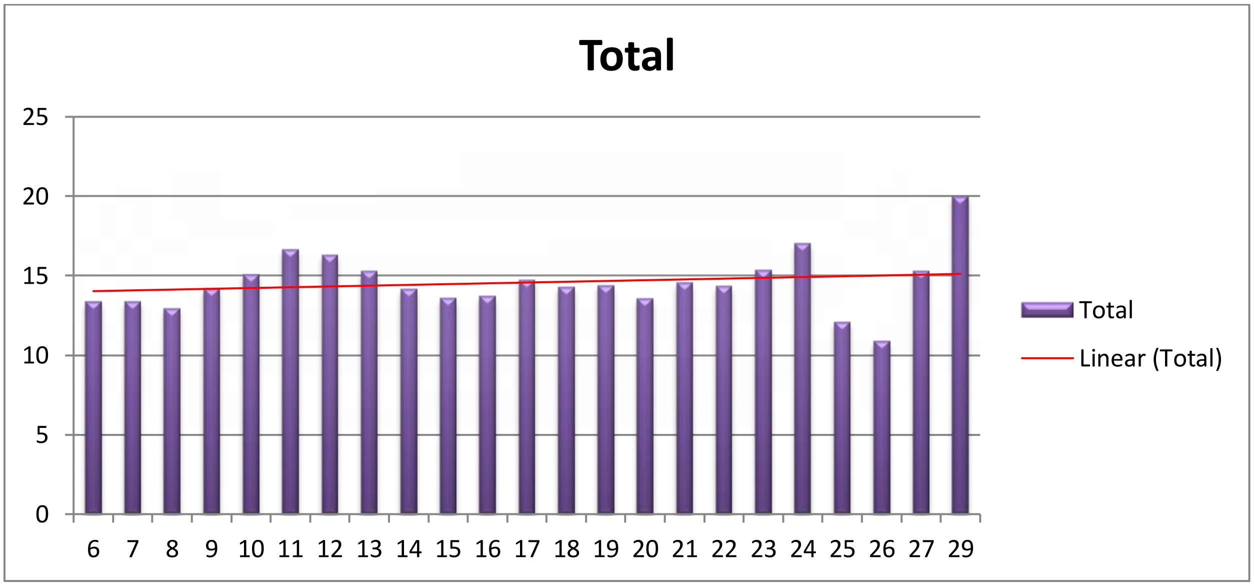


Figure 21: Trend line fitting for Big crabs whose diameter is greater than the median diameter

- Here the x axis denotes the age in months and y axis denotes the average shucked weight for crabs whose diameter is greater than the median diameter.

- From the above figures we can observe that for those big crabs that is whose diameter is greater than median diameter the shucked weight remains more or less stable as the age increases. Here we have fitted a linear trend line.
- So as expected in both cases we observe the same thing as before.
- From figure 13 (correlation plot) we can observe that the correlation between length and diameter is 0.99 which is quite high so we can consider length only primarily.

Data Modification and Transformation

Our response variable (shucked weight) is not linearly related to the predictor (Age). The response variable is linearly related to the log transformation of the response variable. So we will use the **log transformation: Log(Age)** , for building up the model.

Also, our response variable is not linearly related to the other predictor.

such as; Length, Height, Diameter. It is related with the polynomial of order 2 of the predictors. So, in this case our transformation will be **(predictor)^2**.

- The correlation between the Length and Diameter is 0.9795329, so we can consider only Length primarily.

▪ Model Building and Model Modification

As we can see that our predictors are related to our response variable through different transformation of them so we will build up our model by taking that suitable transformation of that particular predictor variable. Also suppose for a normal linear regression the model will look like

Response=(intercept)+(coefficient1)*predictor1+(coefficient2)*predictor2+

(**coefficient3)*predictor3+….+error term.**

Now, our predictors are basically the physical features of crabs. So, if say all of the predictors are taking value 0 i.e., predictor1=predictor2=predictor3=…=0 then

Response=(intercept)+(coefficient1)*0+(coefficient2)*0+

(coefficient3)*0+….+error term.

i.e. our response in taking some value which means that even if the predictors are not available alternatively we can say if the crab is not there at all(because if a crab has no length , height , diameter , age ; there is no crab at all) still we have the value of our response variable which is absurd.

So, we will take the model as a without intercept model.

Now we check for the impact of each of the predictor variable on our response (Shucked weight), also whether the coefficients are significant or not . In this case we will take the transformed predictor accordingly.

➢ Model1: Shucked weight = β_1 log(Age)+u_1 ; where β_1 = coefficient of log(Age)

u_1= error term

➢ Model2: Shucked weight = β_2 (Length)^2+u_2; where β_2 = coefficient of (Length)^2

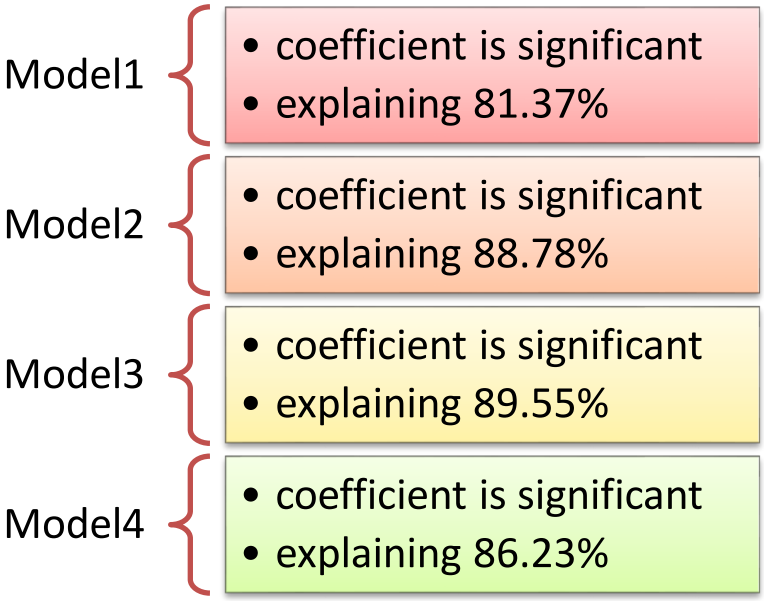
u_2= error term

- Model3: Shucked weight = β_3 (Diameter)^2+u_3; where β_3 = coefficient of (Diameter)^2

u_3= error term

- Model4: Shucked weight =β_4 (Height)^2+u_4; where β_4= coefficient of (Height)^2

u_4= error term



Model1
- coefficient is significant
- explaining 81.37%

Model2
- coefficient is significant
- explaining 88.78%

Model3
- coefficient is significant
- explaining 89.55%

Model4
- coefficient is significant
- explaining 86.23%

Observation:

As we can see that the Age is explaining 81% alone. Also, Length and Diameter is explaining almost in same amount (although we already mentioned that we can work with the Length primarily because their correlation is too high). But Diameter is explaining more than Height.

Height is explaining least among them.

So, now let us take the combination of them to see how they react together.

As we have to keep the Age in predictor because we actually want to see how shucked weight is changing with respect to age along with others so we will keep Age as a predictor, and we will exchange other predictors as well.

- Model5: Shucked weight = $\alpha_1$log(Age)+α_{12} (Length)^2+v$_1$

  where α_1 = coefficient of log(Age)
  α_{12}= coefficient of (Length)^2

  v$_1$ = error term

- Model6: Shucked weight = $\alpha_2$log(Age)+α_{22}(Diameter)^2+v$_2$

  where α_2 = coefficient of log(Age)
  α_{12}= coefficient of (Diameter)^2

v$_2$ = error term

- Model7: Shucked weight = $\alpha_3$log(Age)+α_{23} (Height)^2+v$_3$

  where α_3 = coefficient of log(Age)
  α_{23} = coefficient of (Height)^2

v$_3$ = error term

Model 5

- coefficients are significant
- explaining 92.36%

Model 6

- coefficients are significant
- explaining 93.38%

Model 7

- coefficients are significant
- explaning 86.29%

Observation:

From the above model with Age the Height is explaining least and there is no significant difference between model 5 and model 6 but still Diameter is explaining more that Length. But we can also consider the model with Length as Length and Diameter is highly correlated.

Now, let us see how both models with Age interact themselves.

➢ Model8: Shucked weight = $\alpha_{18}\log(\text{Age})+\beta_{18}(\text{Length})^2+\gamma_{18}(\text{Diameter})^2+v_{18}$

where α_{18} = coefficient of log(Age)

β_{18} = coefficient of $(\text{Length})^2$

γ_{18}= coefficient of $(\text{Length})^2$

v_{18} = error term

➢ Model9: Shucked weight = $\alpha_{19}\log(\text{Age})+\beta_{19}(\text{Length})^2+\gamma_{19}(\text{Height})^2+v_{19}$

where α_{19} = coefficient of log(Age)

β_{19} = coefficient of $(\text{Length})^2$

γ_{19}= coefficient of $(\text{Height})^2$

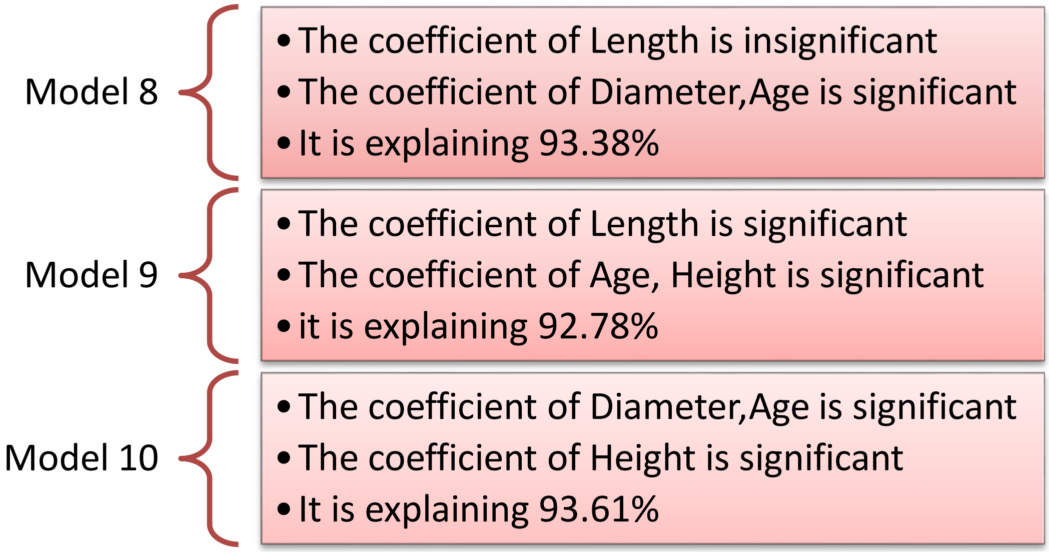
v_{19} = error term

- Model10: Shucked weight = $\alpha_{110}\log(\text{Age})+\beta_{110}(\text{Diameter})^2+\gamma_{110}(\text{Height})^2+v_{110}$

where α_{110} = coefficient of log(Age)
β_{110}= coefficient of $(\text{Diameter})^2$

γ_{110}= coefficient of$(\text{Height})^2$

v_{110}=error term



Model 8

- The coefficient of Length is insignificant
- The coefficient of Diameter,Age is significant
- It is explaining 93.38%

Model 9

- The coefficient of Length is significant
- The coefficient of Age, Height is significant
- it is explaining 92.78%

Model 10

- The coefficient of Diameter,Age is significant
- The coefficient of Height is significant
- It is explaining 93.61%

=

Observation:

In model 8 the coefficient of Length becomes insignificant.

In model 9 and 10 we took Age and Height with Length and Diameter respectively. In this case also we observed that in Model 10 where we took Diameter with Age and Height , it is explaining more than Model 9.

Conclusion:

- ♦ Comparing all of the model we have seen that --------

    i. In Model 9; where we consider a non-intercept model with Age, Diameter and Height as response, the multiple $R^2$is 0.9361 . Which means that this is explaining 93.61% of the data.

    ii. In Model 6; where we consider a non-intercept model with Age and Diameter as response, the multiple $R^2$is 0.9338. means that it is explaining 93.38% of the data.

    iii. Which Comparing these two we can say that we can take Model 6 with two predictor variables. There is no need to take more than two predictor or model 9 because the multiple $R^2$is not increasing significantly.

- ➢ We have done the above analysis based on the small crab only.
- ➢ From the above diagram we have seen that the shucked weight

is increasing in nature but for the large crab the shucked weight is stable, so we can say that for small size and big size crabs the analysis are different.

- ➢ Now, we will consider the whole dataset and we will consider the

interaction between the size to see how our predictor is reacting, because in real life we have to deal with mixed sizes of crab not only the small one.

Including Interaction in Regression

Let's say X1 and X2 are features of a dataset and Y is the class label or output that we are trying to predict. Then, If X1 and X2 interact, this means that the effect of X1 on Y depends on the value of X2 and vice versa then where is the interaction between features of the dataset. As we know that our dataset contains interaction. We should take interaction into account in our model for better precision or accuracy.

As we have seen that for small crab our dataset is showing a certain pattern of line when we are plotting shucked weight in the y axis and Length (Diameter) and some another pattern for big crabs, so clearly the interaction of size is affecting our analysis. That is why we are taking account of the interaction as mentioned above.

In this case we are considering the whole dataset and therefore our model looks like:

$Y=D1*x1 + D2*Indicator_{size1}+D3*Indicator_{size2}+$ Interaction

UNDERSTANDING OF THE FITTED MODEL:

- **SUPPOSE I HAVE ONE OBSERVATION WHICH IS OF SIZE1,AND LET US ASSUME THAT THE PREDICTOR X1 IS HAVING AN INTERACTION WITH SIZE1. THEREFORE , IN THAT CASE OUR MODEL WILL BE**

Y= D1*x1 + D2+ Interaction.

- **NOW IF I HAVE ONE OBSERVATION WITH SIZE2 THEN OUR MODEL WILL BE**

  Y= D1*x1 +D3

- **NOW LET US FIT THE FOLLOWING MODEL AS PER OUR ANALYSIS AND TO SEE HOW IT IS AFFECTING OUR RESPONSE IN MANY WAYS:**

WE HAVE FITTED WITH INTERACTION MODEL ACCORDINGLY TO SEE HOW IT IS EXPLAINING OUR DATASET.

HERE WE WILL ALSO CONSIDER NON-INTERCEPT MODEL AS WELL.

➢ MODEL_I1:

Shucked weight = D1 log(Age) + Interaction with age.*

➢ MODEL_I2:

Shucked weight = D1 log(Age) + D2* (Length)^2 + Interaction with age.*

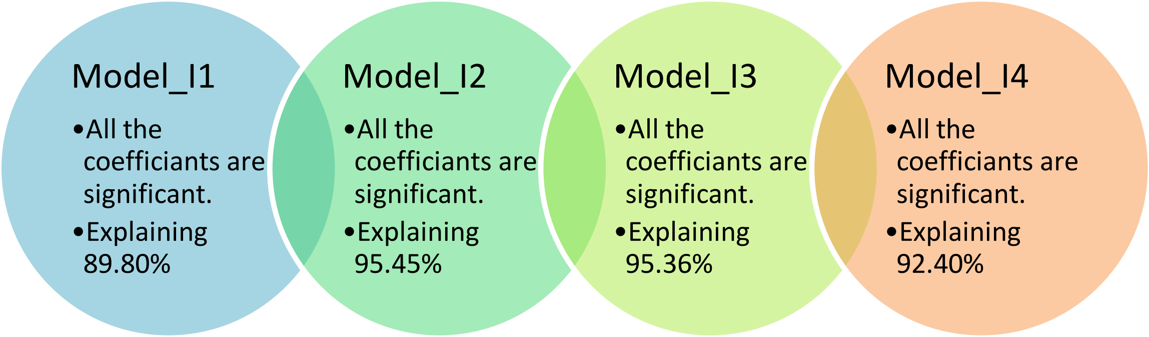
➢ MODEL_I3:

Shucked weight = D1 log(Age) +D2* (Diameter)^2 + Interaction with age.*

➢ MODEL_I4:

Shucked weight = D1 log(Age) + D2*(Height)^2+ Interaction with age.*

DIAGRAMMATIC REPRESENTATION OF BETTER UNDERSTANDING TO COMPARE THE ABOVE MODELS



Model_I1

- All the coefficiants are significant.
- Explaining 89.80%

Model_I2

- All the coefficiants are significant.
- Explaining 95.45%

Model_I3

- All the coefficiants are significant.
- Explaining 95.36%

Model_I4

- All the coefficiants are significant.
- Explaining 92.40%

We have also done some other combination but the multiple R^2 is not Increasing significantly.

Observation:

- When we are considering the interaction regression with Age only it is explaining 89.80% of the data.
- On the other hand, when we are considering any other predictor variable with Age then; Model_I2 i.e., Length with Age and the interaction is giving us a best fit among them.

Conclusion:

- Whenever we are taking account of the interaction effect of sizes, the Length along with Age is giving us the best model.

- But when we are not taking account of the interaction effect of sizes the Diameter along with Age is giving us the best fit.

Acknowledgement

I would like to express my heartfelt gratitude and extend my sincerest appreciation to the individuals who have played a significant role in the successful completion of my project. Their guidance, support, and expertise have been invaluable throughout my journey.

First and foremost, I am deeply indebted to my project supervisor, **Debjit Majumder**, for his unwavering dedication, insightful feedback, and continuous encouragement. His profound knowledge in the field of statistics and his willingness to share his expertise have been instrumental in shaping the direction of my project. I am truly grateful for his mentorship and guidance, which have been invaluable assets in this endeavour.

I would also like to extend my sincere thanks to the Head of the Statistics Department, Prof. Soma Nag, for her constant support and encouragement. His vision and leadership have provided me with a conducive environment to pursue my research interests. I am grateful for his valuable insights and the opportunities he has provided for my intellectual growth.

Additionally, I would like to express my gratitude to the other esteemed professors of the Statistics Department, Sir Shilpak Mukherjee, Madam Champa Chakraborty. Their expertise and willingness to share their knowledge have been of immense help in refining my research ideas and broadening my understanding of statistical concepts. Their constructive criticism and valuable suggestions have played a pivotal role in shaping my project.

Appendix

- Codes for this project are uploaded on https://drive.google.com/file/d/1rpbPwjUcuAjgw5N50yzCpWREJ39lmQyP/view?usp=drive_link
- References for this project are

  i) The dataset obtained from Kaggle

  ii) Wikipedia

  iii) Youtube

  iv) Geeks for geeks for helps in code

Thank You