Abalone Population Analysis:
Data Analysis Project #1

BY

Sada Borrego, Juan Javier

Class 401-DL-Sec39

Introduction to Statistical Analysis

MASTER IN PREDICTIVE ANALYTICS

At

NORTHWESTERN UNIVERSITY

2017

Abstract

The purpose of this paper is to make an exploratory data analysis of failed Abalone population observational study, the goal is to determine probable reasons why the original study was not successful, or if any other variables should be considered in predicting age based on physical characteristics. This exploratory data analysis intend to identify possible relationships between the physical characteristics and other variables observed in Abalone data collection, and how this would be significant to understand the different underlying relations among the variables, improve future observational studies and conclusions in the second delivery.

Introduction

Blacklip Abalone, according to Wild Fisheries Research Program, is large flattened marine mollusk use mainly as a food source for humans; the populations of this mollusk are distributed from Cabo San Lucas in Bajada California, Mexico all the way up to Oregon. According to the Center for Biological Diversity, the current Blacklip Abalone population is only 1% of the population that existed in 1985. The combination of many factors, that will be mentioned below, with low reproductive rates makes the Blacklip Abalone an endangered species. According to Fishtech Abalones and the Center for Biological Diversity, Blacklip Abalone has decreased its population to nearly 45%, from 18,000 metrics tons to 10,000 metric tons. The factors often cited as the most relevant for such decline in the populations are:

1. Predation. Sea otters (major abalone predators) expanded their range in central California virtually eliminating recreational and commercial fisheries for abalone and other invertebrates. Commercial abalone harvesting is now primarily concentrated in southern California.
2. Mortality of small abalone for many reasons.
3. Over harvesting. Abalone are easily over harvested because of slow growth and variable reproductive success.
4. Competition. Sea urchins and other species, utilizing abalone food and living space.
5. Illegal harvesting, the most important reason in declining.
6. Loss of habitat. Coastal "development" and pollution have ruined large areas of abalone habitat.
7. Environmental factors, such global warming, pollution or changing environmental factors.

8. Diseases, such as withering syndrome.

Due to the factors mentioned and the rapid decrease in abalone population , the probability of a successful reproduction is very questionable and the chances not just of increasing populations but of species survival are very slim; that's the reason why in 2006 the Center for Biological Diversity requested "*that the Secretary of Commerce, through the National Marine Fisheries Service ("NMFS" or "NOAA Fisheries"), list the Black Abalone (Haliotis cracherodii) as Endangered under the federal Endangered Species Act ("ESA"), 16 U.S.C. § 1531 – 1544. The Center also requests that black abalone critical habitat be designated concurrently with its listing*" (Center for Biological Diversity, 2006).

Making sustainable efforts to understand the Blacklip abalone is very important for species subsistence; one intent of a group of investigators was to predict the age of abalone from physical measurements thus avoiding the necessity of counting growth rings for aging. Ideally, a growth ring is produced each year of age. Currently, age is determined by drilling the shell and counting the number of shell rings using a microscope. This is a difficult and time consuming process. Ring clarity can be an issue. At the completion of the breeding season sexing abalone can be difficult.

Similar difficulties are experienced when trying to determine the sex of immature abalone referred to as infants. As the study was inconclusive the investigators determined additional information would be required such as weather patterns and location which affect food availability; the objective of this first delivery is to determine plausible reasons why the original study was not successful in predicting age based on physical characteristics.

This paper uses the statistical and graphical methods of exploratory data analysis on abalone data collection and focuses on the approach followed by the scientist to gather the information. The main purpose is trying to understand the different variables or methods used during gathering process, and provide a systematic scheme for looking at data and extracting the patterns that are contained in the data. The work on this paper is the foundation for a second delivery which involves statistical inference using analysis of variance and linear regression.

The tools that will be used in this exploratory data analysis are:

1. Making exploratory statistical analysis in the whole data set, regardless of differentiators such as sex or class, to have an overview of the data and understand each variable as separate entity.

2. Making exploratory graphs for singles variables to recognize the structure, distribution, identify outliers or any other atypical characteristic.
3. Understand the possible relationship between variables, more specifically between categorical and numerical variables, to comprehend possible patterns and structures.
4. Apply principles of analytic graphics, to understand the relations among variables, type of distribution, and variation in the different variables, central tendency measurements, data characteristics and identify outliers. Plotting the data permits the analyst to determine the extent to which the assumptions are valid and to catch obvious errors in data entry.

The application of exploratory data analysis (EDA) prior to the actual abalone regression and variance analysis is crucial as it identifies the problems with the data collection, problematic samples, problematic data and poorly examined abalone data collected, if undetected, can compromise the outcome of the experiment. This exploratory data analysis of such large data sets thus requires specialized numerical and graphical strategies, to understand the variables, their relationships and set up a solid base to prepare a deeper statistical analysis, as proposed for the second delivery.

Results

This first attempt statistical analysis is based on the research made by scientist to determine the age of abalone based on psychical measurements; after Black lip abalone research was finish and data gathered (*n* = 1036 observations into 8 different variables), before starting the descriptive statistical analysis is important to understand the different variables include and their type. The variables are divided into 2 categories:

A. Nominal variables:
    1) *Sex*; divided into female, male and infant (specimens that haven't reach sexual maturity).
    2) *Class*; Age classification based on RINGS (A1= youngest,., A6=oldest)
B. Ratio variables:
    1) *Length*; longest shell length in cm.
    2) *Diam*; diameter perpendicular to length in cm.
    3) *Height*; height perpendicular to length and diameter in cm.
    4) *Whole*; whole weight of abalone in grams.
    5) *Shuck*; Shucked weight of meat in grams.

6) *Rings*; Age (+1.5 gives the age in years)

Aside from the variables mentioned, another 2 were add: 1) Abalone volume, which is the multiplication of the length, diam and height, and 2) the ratio, obtained dividing shuck against volume.

The first step is getting the basic summary statics, the procedure of summarizing the general data per variable, provides statistics that detail the central tendency, spread, extremes and distributional characteristics of the data, Table 1 below shows the results of this first analysis and help us to shed some light into data distribution. The summary values are the next ones:

SEX	LENGTH	DIAM	HEIGHT
Length:1036	Min. : 2.73	Min. : 1.995	Min. :0.525
Class :character	1st Qu.: 9.45	1st Qu.: 7.350	1st Qu.:2.415
Mode :character	Median :11.45	Median : 8.925	Median :2.940
	Mean :11.08	Mean : 8.622	Mean :2.947
	3rd Qu.:13.02	3rd Qu.:10.185	3rd Qu.:3.570
	Max. :16.80	Max. :13.230	Max. :4.935

WHOLE	SHUCK	RINGS	CLASS
Min. : 1.625	Min. : 0.5625	Min. : 3.000	Length:1036
1st Qu.: 56.484	1st Qu.: 23.3006	1st Qu.: 8.000	Class :character
Median :101.344	Median : 42.5700	Median : 9.000	Mode :character
Mean :105.832	Mean : 45.4396	Mean : 9.984	
3rd Qu.:150.319	3rd Qu.: 64.2897	3rd Qu.:11.000	
Max. :315.750	Max. :157.0800	Max. :25.000	

VOLUME	RATIO
Min. : 3.612	Min. :0.06734
1st Qu.:163.545	1st Qu.:0.12241
Median :307.363	Median :0.13914
Mean :326.804	Mean :0.14205
3rd Qu.:463.264	3rd Qu.:0.15911
Max. :995.673	Max. :0.31176

Table 1. Summary data from Abalone Study

As mentioned above, Table 1 shows the resulting basic descriptive statistics for each variable, except for Sex and Class variables which are nominal: mean, median, first and third quantiles, as well as the minimum and maximum value for individually variable.

Despite of Table 1 displays central tendency measures and a good approach to understand the variables, is necessary to address every numeric variable itself to understand its distribution and shape. Starting with Length variable, as we can see on the Figure number 1, data distribution has a negative skewness because has several outliers in the lower part of the measurements, in this specific variable there are 12 lower outlier values that pull the tail into the left part of the Histogram. In the case of Length after making computations the skew for this data is -0.67, as the value is negative the meaning is that median is larger than the mean and the values concentration is on the right hand of the histogram. Analyzing the kurtosis the value for this variable is 0.16, so we have departures from normality.

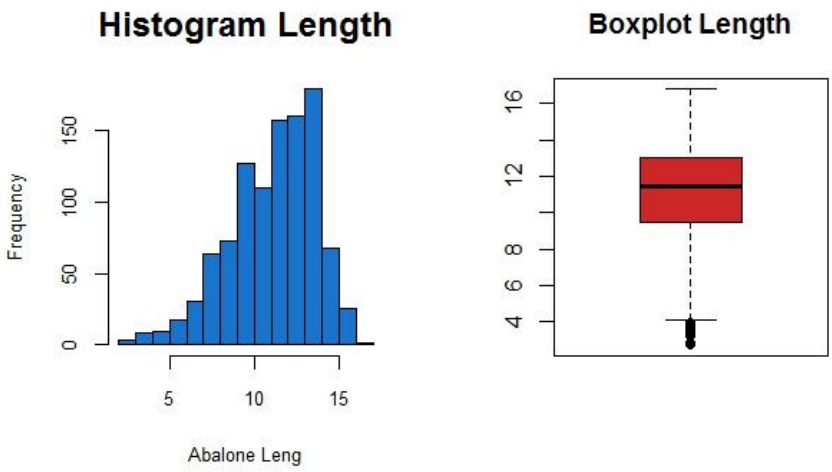

Figure 1. Abalone Length Histogram and Boxplot

For variables Diam and Height data displays, in Figure 2, the same pattern that Length, both have negative skewness, -0.62 for Diam and -0.23 for Height. The left tail in Histogram indicates that median is larger than mean and the majority of values concentrate in the right part of the graph, this is because by the outliers in the lower part of the values, Diam has 13 outliers and Height 6 values. The kurtosis for

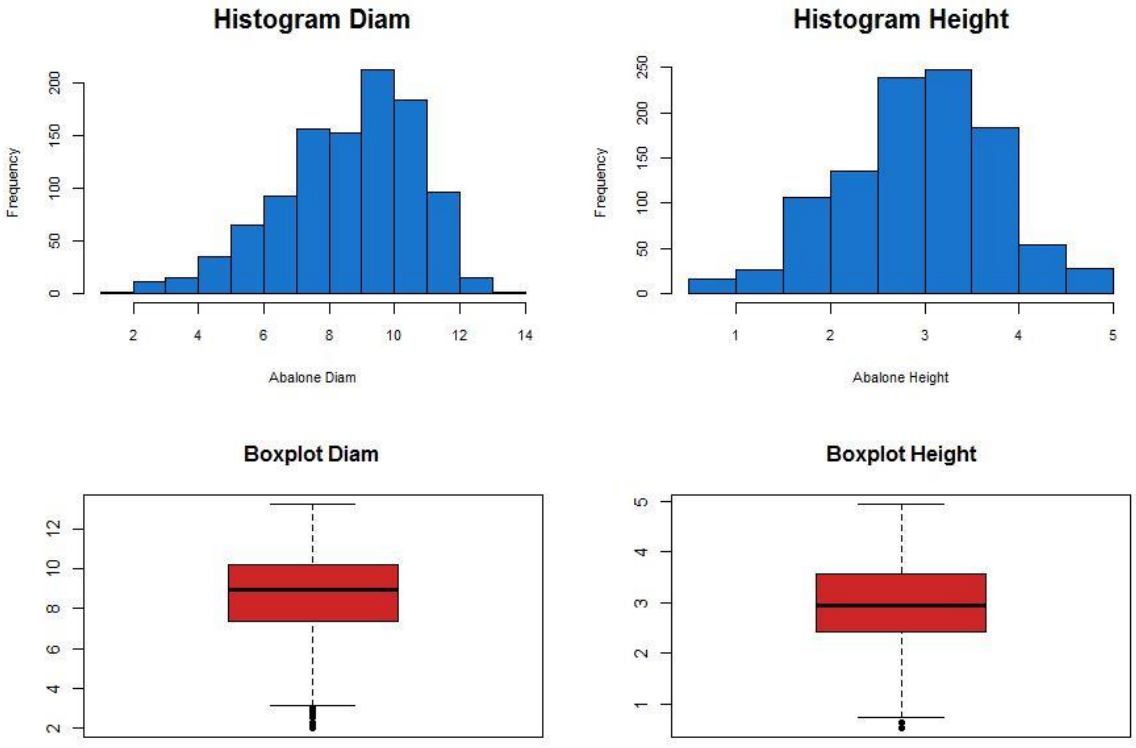

Figure 2. Diam & Height Histogram and Boxplot

Diam is 0.00030 so is near to the normal distribution, but the kurtosis for Height is -0.18 resulting in lighter tails with a flat peak. The skewness level of each graph can be understand because of quantity and values of outliers on each one; in the case of Diam variable data presents 13 outliers in the lower values and for Height there are only 6 outliers existent.

About the remaining 3 variables, Whole, Shuck and Rings, the figure number 3 shows a right and positive skewness, implicating that the mean is larger than the median and the outliers are in the upper values of the data set, as can be seen in the boxplot graphs. While in Whole and Shuck variables the skewness is lighter, 0.47 and 0.63 values, in the case of Rings the value for skew is 1.23 not just with more outliers but with more separations beyond the extremes of the whisker. Computing kurtosis values for the variables we obtained that Whole = -0.28, Shuck = 0.20 and Rings = 2.68, for the first 2 variables the values are close to zero indicating that data are close to normality, but in the Rings case kurtosis value is big enough and close to 3 indicating that distribution could be not normally distributed or the presence of heavy outliers.

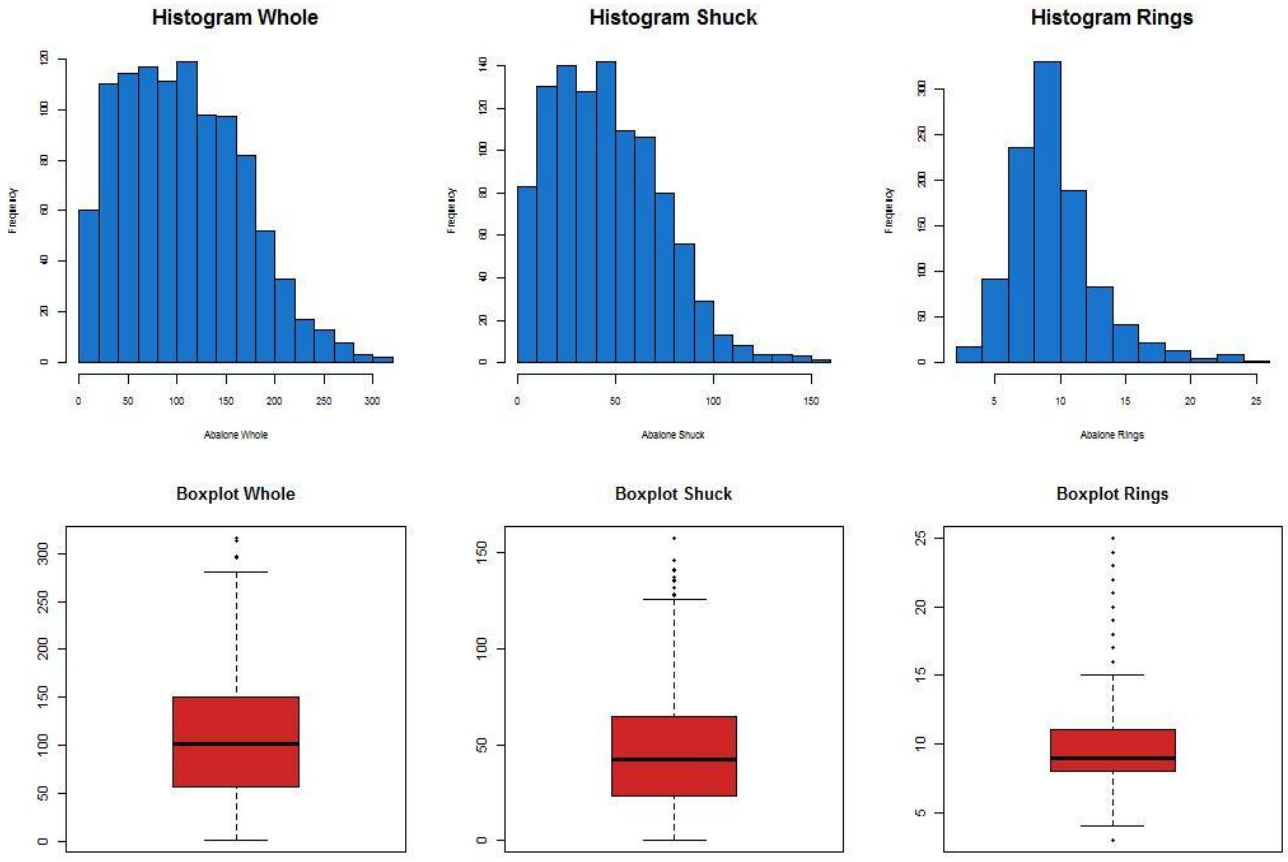

Figure 3. Whole, Shuck & Rings Histograms and Boxplots

Now after having analyzed the basic descriptive statistics for all the variables, is in our interest to address the relation between the sex and the class variables to understand how both variable behave in relation to each other and which patterns can be uncover or any special futures are existent. As we can see in Table 2 and Figure 4, summarized the relation between both variables.

	A1	A2	A3	A4	A5	A6	Sum
Female	5	41	121	82	36	41	326
Infant	91	133	66	21	10	8	329
Male	12	62	143	85	37	42	381
Sum	108	236	330	188	83	91	1036

Table 2. Sex vs Class comparison

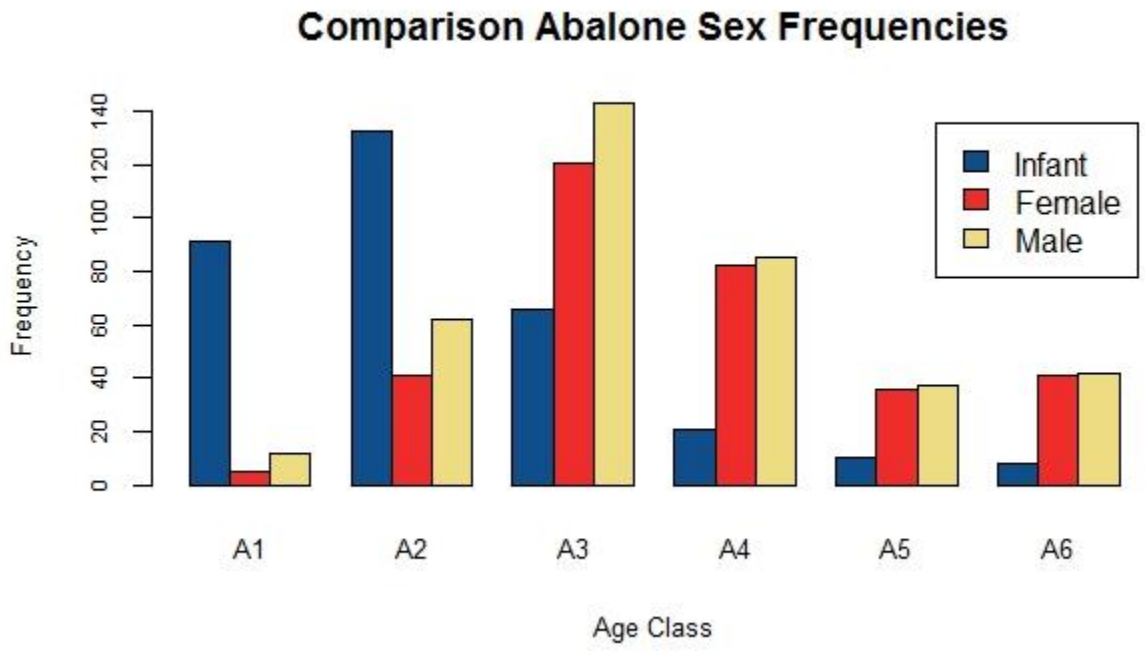

Figure 4. Sex vs Class frequency comparison

The first thing we can notice is that the total observations sum across the 6 different classes for the male, female and infant is distributed evenly, the probabilities at this point would range from 0.31 females, 0.32 infants and to 0.37 for males, likewise there are things to point out here: number one, we can see logically that infant participation in the numbers decrease across de class, but also the numbers in general decline as we move to older classes suggesting a depleted population or at least that is being eroded in the time and fewer adults reach the older classes, with the implication about sexual maturity for breeding and multiply the population.

The number two, in older classes, from particularly from number A4 to A6, there are still infant observations when the reasonable process for this observation should be as we move into older classes

the infant participation should decrease into the point of disappearing but as we can see infants maintain a presence in all classes regardless of the age; likewise, the major infant participation should be in class A1 and no A2; all this reinforced the problem stated in the introduction, "*study was not successful in predicting age based on physical characteristics*", determining abalone age with physical characteristics seems to be subject to investigators appreciation in the best case or the rings are not a sure characteristics to classify the abalones. Finalizing and going deeper into the class analysis, the major observations concentration is class A2 and A3, actually between those 2 classes amount to 53% of the values.

In order to additionally explore the relationship between variables we sampled 200 observations from the whole population and summarized the result a matrix plots, which allow us to see the different interactions within the data. Figure 5 exhibited those facts and interactions:

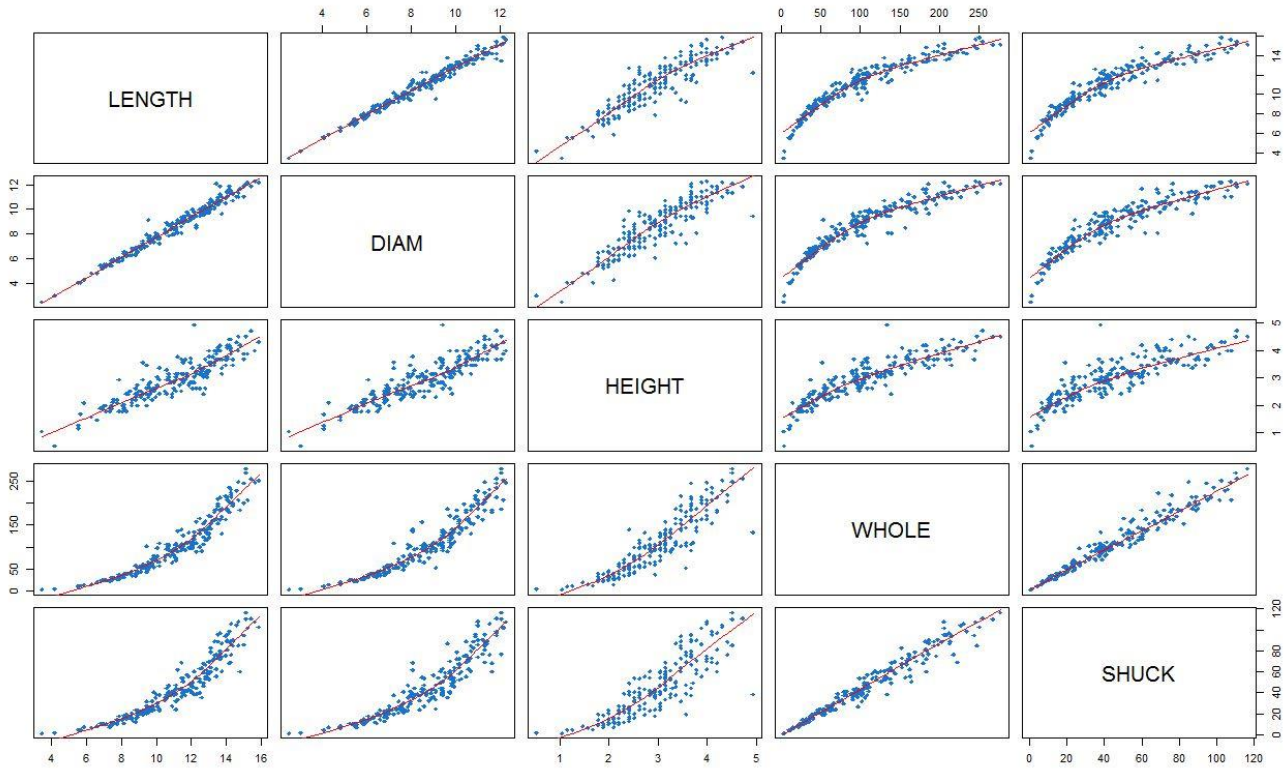

Figure 5. Variables Plot Matrix

Opening the analysis let's investigate the general type of relation we can see in the matrix above; the first conclusion we get is that all associations between variables are positive, with different strong degrees of relation all variables relate positively and with no negative values the inference is that if one variable increase the other one too.

Now into a more profound analysis among the different variables; first, analyzing the relation between Length and Diam we can infer that both variables have a strong positive relation as almost all the values are pretty close to the trend line and also because the line is straight indicates that values are not disperse, likewise another conclusion we got is that length increase almost at same rate that diam, increasing together and the major observations concentration is in the upper values of each variable with practically no values in the lower part, this could mean that rates increases and correlates after gaining certain length and diam. The other strong straight relation we can find is between abalone length and shuck, the same pattern is present the trend line is straight and positive indicating a strong connection so the length and meat weight in abalone appears to increase together, but the difference with the first relation is concentrate in the lower observations values this could lead us to conclude that most grams of meat are develop in length early stages.

Regarding at Figure 5 we can identify another two positive relations with a little less strength that those two mentioned above. Comparing abalone height with diam and length in both cases relationship give the impression to be strong and positive, so as length and diam growth the height does it too, but the values are a little more disperse and not tightly concentrated around the trend line as length and diam. Another significant point here is how the interaction values between height vs length and height vs diam are gathered from the mid-section of the values to the upper part and the observations numbers in lower values are pretty scarce with a possible explanation could be that those values are outliers.

Observing for the next interactions between whole vs length, whole diam, shuck vs length and shuck vs diam we will encompass these different variables and their relations in this paragraph, in all those cases we can observe a positive relation with solid bond and the values are spread almost evenly all over the values spectrum, we can notice correspondingly that at lower levels whole and shuck rise at inferior rates than length and diam, but as we move to upper values of length and diam growing rate response related to diam and length increase more rapidly. Finally, the weaker relations height vs whole and height vs shuck, still a positive relations but values are more spread from trend line and with several values far away from this line suggestion an outlier and that in some cases the relation amid these variables are not as simple and direct as we might think.

Utilizing R program to plot the next relation to analyze, we study how abalone whole weight and volume interact within each other, Figure 6 exhibits how each observation of weight behave with its correspond volume value along the whole spectrum.

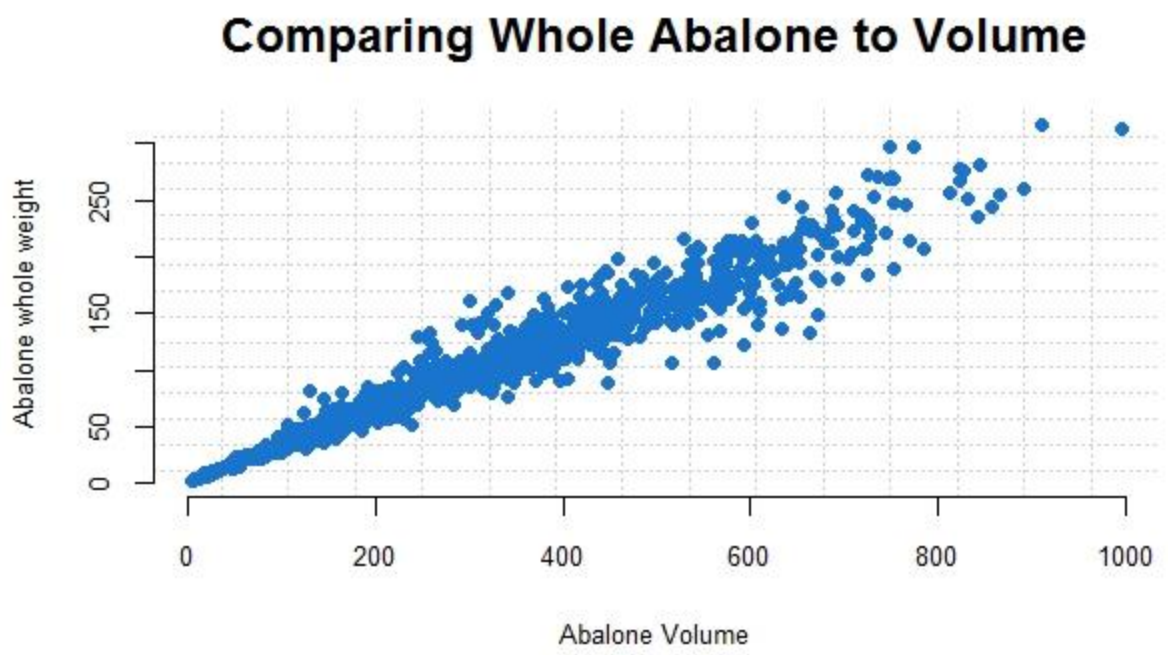

Figure 6. Comparing abalone weight to volume

Figure 6 does a good job of explaining the relationship between the variables, and derivate from it we can observe that all the points on the scatter plot have a positive relation to each other as abalone volume increase so does the weight in grams, actually the association between seems to be strong enough to assume that there's an solid influence from volume to increase abalone weight. Even so there's a solid relation in all values, this particularly true when the lineal interaction can be seen especially in lower observations values, from around 600' values in volume and 200's in weight to below the points are too close and grouped. A second conclusion is about data wedge shaped, this suggest first that as values growth in both elements also variability increased, correspondingly this shape in data suggest that as values move into volume and whole height higher values direction of the graph the variance in these observations rises in and the pattern broadens.

If we compare the results of Figure 6 against results in Figure 5, with this study can be established that first graph shows less variability of its elements and clear growing tendency between abalone volume and weight as wells as the same tendency with the interactions among its constituent elements (height, diam and length), the relations between them are clear positive with less dispersion, so the contribution of these three variables cooperate positively to increase abalone volume and weight.

Continuing with the exploratory data analysis the next step is evaluate the dynamic of the relationship between abalone shuck and the whole weight, to understand variable's behavior R language was used to generate a plot the different observations and the result is presented in Figure 7.

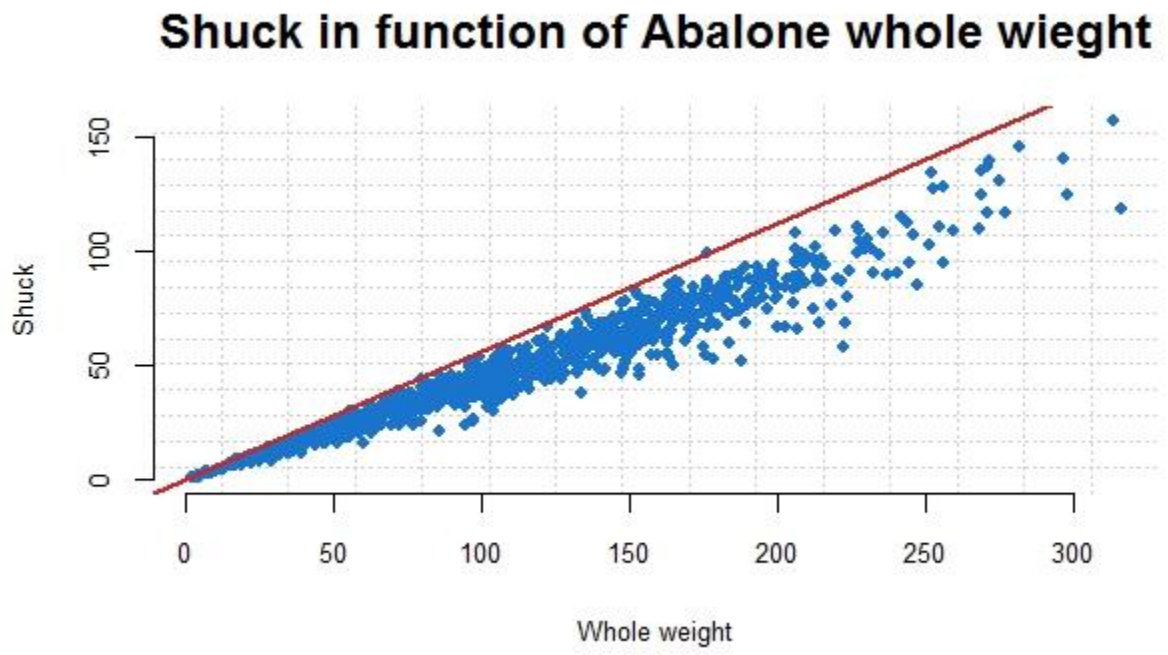

Figure 7. Comparing Abalone shuck vs whole weight

As well as the comparison between abalone whole weight and its volume, the relationships showed in Figure 7 is practically the same; both variables seems to have a strong positive relationship while abalone weight increase the shuck does it to at the same rate, although the presence of values in the upper levels is infrequent; similarly we can see the almost entirely the observations are below the straight line draw in the plot. The main difference among Figure 6 and Figure 7 is that values in in Figure 7 seems to have less variability because of the absence of wedge shape in data.

On the next part of the document we would look into the ratio, abalone shuck divided against abalone volume, but discriminating the analysis into the three different sexes; the idea is look for possible patterns, investigate if data is normal o not as well as identify the outliers of each graph. Figure 8 would help us to examine the different ratios.

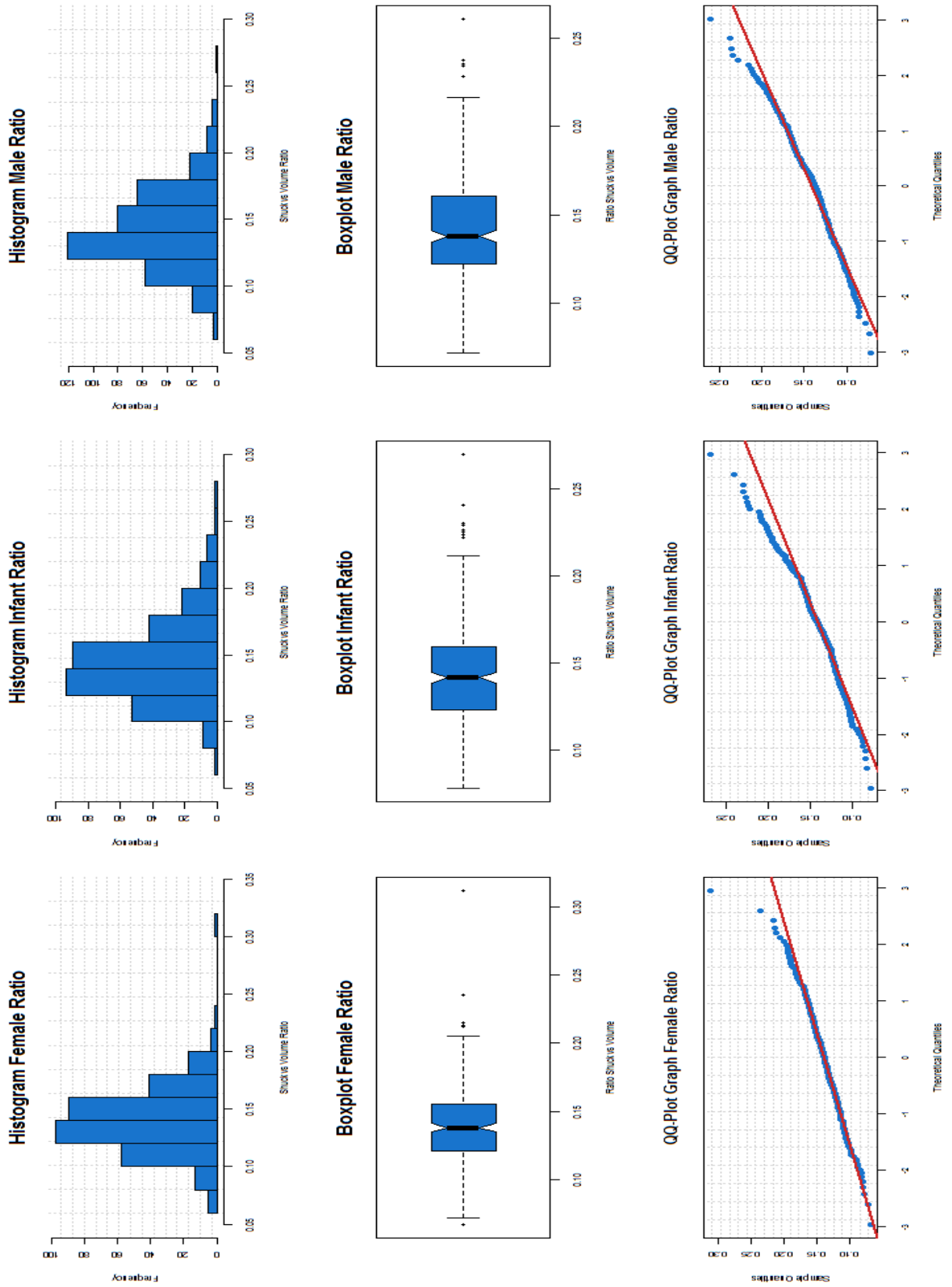

Figure 8. Ratio breakdown by sex

The first conclusion we get from the three ratio histograms is that all are right skewed, meaning that mean is greater than median for three sexes and ratio has few outlier in the higher values with low frequencies, another conclusion from the histograms is major data concentration in each histogram ratio 10 to ratio 20 while the peak is around 15 value; in terms of data behavior seems that regardless of sex data conduct is almost the same, with concentrations and peak around same values and data spread quite similar between them.

Continuing with the analysis commenting the Box plots, in a general overview we can see there is a coincidence about the median all three sexes with value is around 0.13 because of the box shape the graphs have right skewness. In a more deep scrutiny in female case there's not much spread in the data as the box is pretty compact, but we can see outlier's impact in both extremes of the plot is true that most of the outliers are just a few measurements out of the upper whisker (values 0.212, 0.215, 0.213 and 0.235) there are two values that have a good impact, below of the lower whisker we have abalone outlier with a ratio of 0.067 and one way beyond the upper whisker with 0.312 ratio. While In male case, data is more spread that it's female counterpart and all the outliers are above upper whisker of male ratio with values 0.261, .238, 0.235, 0.236, and 0.229 this reinforce the right skewness mentioned in histogram analysis mentioned in the last paragraph. Finally, for infant box plot the data are not spread seems to have a normal behavior but the outliers start farther from upper whisker with the following values: 0.267, 0.222, 0.240, 0.226, 0.225, 0.230, 0.229 and 0.223.

When we start this paper we assume that data has normal distribution and we can use QQ Plot graphs to check this assumption, the graph allows us to see at-a-glance if our normality assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation. Figure 8 shows the ratio QQ Plots discriminated by sex, from all 3 graphs we can see the points are almost all on the diagonal line and even data distribution can be improve the reality is we can assume normal distribution even with outliers present. In the case of female and male ratios QQ Plots, we can identify certain tendency in upper values, from theoretical quantile 1 and above, showing certain distance from the line but this can be outlier presence. The infant ratio case is similar to female and male, we can assume normal data distribution as most observations are on the line or fairly close to it, but the difference is that starting from theoretical quantile 0.5 and further than data start to get away from QQ line indicating that we have outliers presence or the variance in data is increasing.

The next step in our exploratory data analysis is to understand how volume and whole weight contribute in the different class classification and this measurements help to predict abalone age.

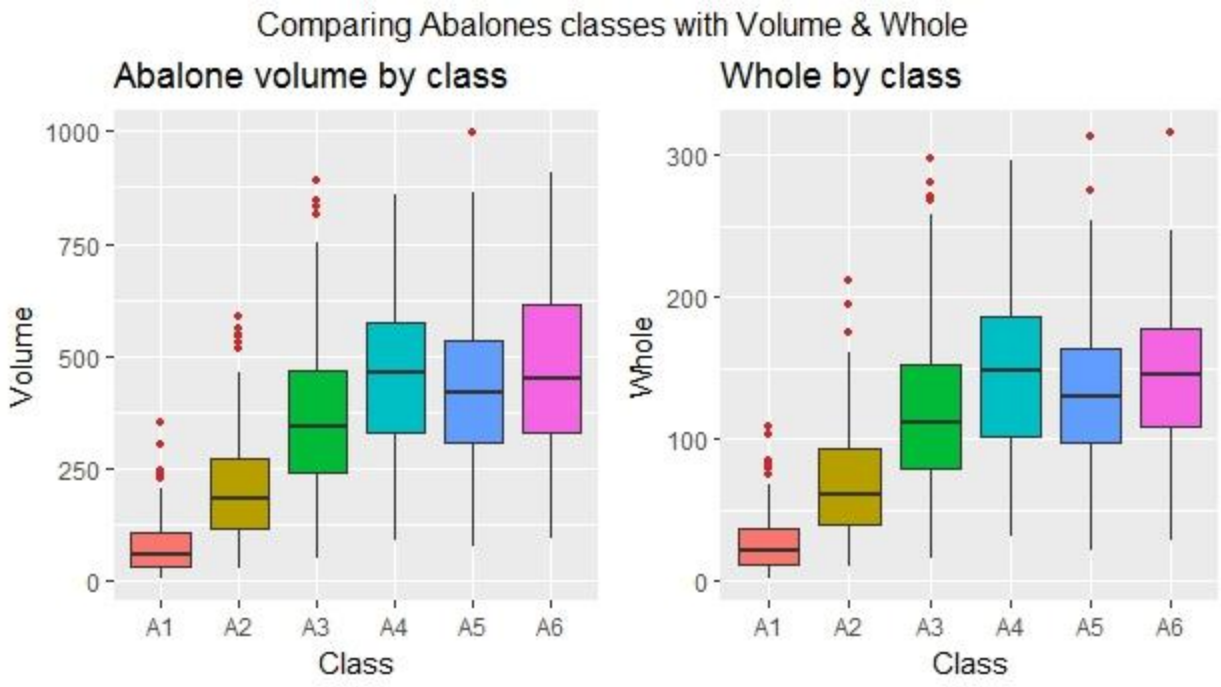

Figure 9. Abalones classes

Figure 9 displays the development between the abalone volume and whole weight in grams against the class. Both graphs showed that volume and whole variables are good abalone age predictors to a certain degree; for example the first 4 classes, from A1 to A4, we see an increase in volume and also gain in weight as this helps to easily differentiate between the classes and ages and hence abalone classification can be easily done. This impact can be seen in how boxplots increase as we move up into the upper classes; the median weight and volume also increase. However for the last two categories both variables don't seem to be good predictors; for instance, the median for volume and weight actually decrease if we compare A5 and A6 with class A4 indicating that variable's impact in prediction is not as certain as in lower classes, and class classification is more complicated and subject to investigator's appreciation.

Another conclusion we can extract from Figure 9 is the outlier's quantities: as we can see there's a good quantity of outliers along the different box plots, especially from classes A1 to A3. This suggests that either age classification base of these 2 factors is not as direct as we might think, and physical development has a variable growing rate making hard to group the observations or else, despite of volume and weight measurements indicate moving observation into the next class, this is subject to investigator's opinion and this can lead to a misclassification inducting errors into data grouping.

The same analysis performed for volume and whole weight in grams against age class would be done for comparing these two factors but with rings, in Figure number 10 we can see this data analysis and their patterns.

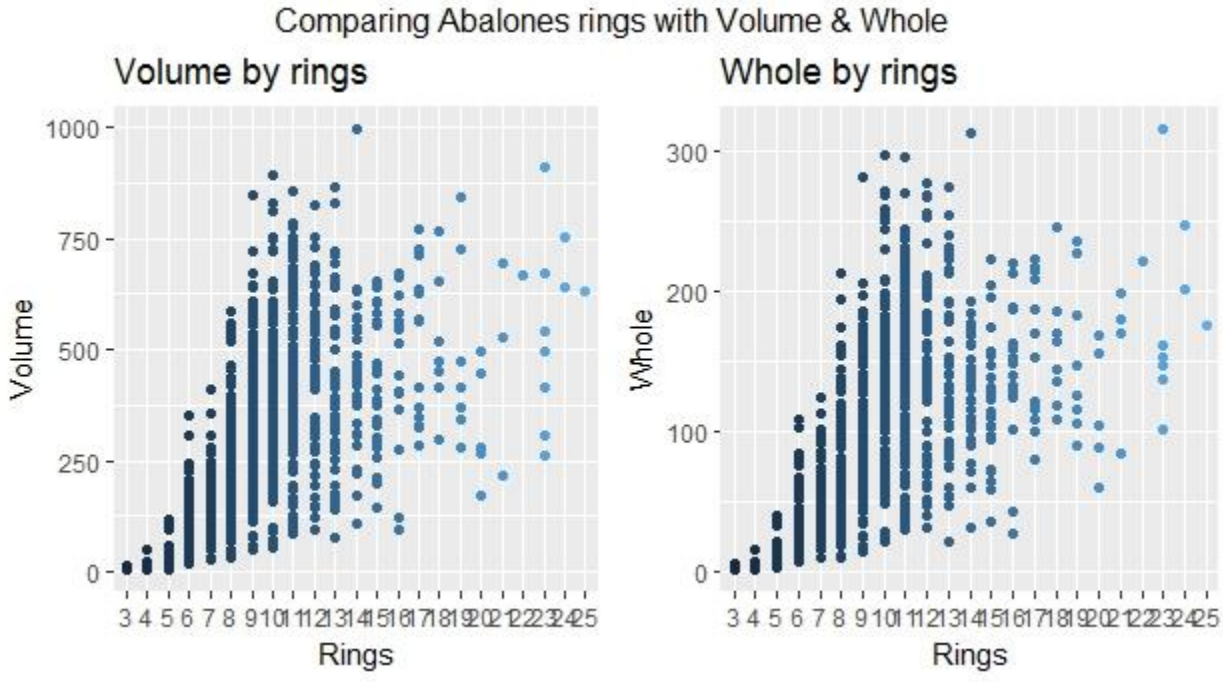

Figure 10. Rings against volume and weight

The general conclusions from Figure 10 is, first, that both graphs present almost the same data outline conduct for volume and weight when comparing with the rings, the increase rates seems to be the same and the peak in data too also data dispersion starts around the same ring value. Correspondingly, as colors in the plots were designed to show darker blue color in major density areas and lighter blue in those areas with less observations, we can realize from Figure 10 that observations densities are seemingly the same for both variables major concentrations are in lower rings values and cero in weight and volume to nearby ring 11 where the peak is located; from there densities started decreasing until ring 41 from where data is completely diffuse into very different values range, when we perceive the presence of outliers with bigger values but most of them are in the similar with same values that mid-range rings, this reinforce the idea that physical characteristics extract from volume and whole weight are partially helpful to predict abalone age and this categorization might be matter to interpretation and appreciation rather than a direct connection.

Given the information examined the last 2 graphs it's important to comprehend the median across the different classes ages and ratios but discriminated by sexes, this would allow us to assets data conduct

in these fields. Table 3 shows volume mean by age class and segregated by sexes while Table 4 present ratio mean across age classes and separated in sexes; there's a pattern in all three sexes the mean increase as we move up in data until A4, but in group A5 the median decrease compare with group A4 this could mean wrong observations capture, heavy outliers with big impact in population's mean, wrong age class sorting according to physical characteristics or classes need redefinition based on data gathered. Another interesting thing to point out is even if by definition infant category shouldn't be in older abalone varieties we see a presence in all age classes and also the mean in both variables are fairly big, contrary to the logic where we would think the mean should have a small value if there were any presence of this sex.

	A1	A2	A3	A4	A5	A6
Female	255.30	276.86	412.61	498.05	454.10	514.30
Infant	66.52	160.32	278.95	316.41	261.75	328.16
Male	103.72	245.39	358.12	442.62	436.15	443.78

Table 4. Volume mean by sex and class

	A1	A2	A3	A4	A5	A6
Female	0.1547	0.1555	0.1450	0.1380	0.1282	0.1191
Infant	0.1570	0.1476	0.1369	0.1244	0.1179	0.1154
Male	0.1513	0.1564	0.1462	0.1365	0.1300	0.1229

Table 3. Ratio mean by age class and sexes

Our final data review for this exploratory data analysis is present graphically the information stated above and in tables 3 & 4; Figure 11 displays this information.

Figure 11. Ratio & Volume means by class and sexes

Figure 11 show how data develop across age classes, as mentioned before the first quick answer we get how mean is growing with each classes we move up but suddenly at A5 class's median values decrease if we compare with lower classes. Also of we see ratio values decrease as we move into upper classes. Along with the graph there are some questions that remain unanswered are:

1. Why median values in volume and weight increase in older ages for abalones, specifically in A5, and then median values grow again.
2. What is a good distributional fit for a set of numbers? We have seen that data is distributed normally but we need to asset how we can improve it and at what point volume and weight are goo predictors of age also if there is any causal relation that would help to predict abalone age and future populations.
3. Another question is why ratio is decreasing as abalones are getting older.
4. Asset if there's any causal relations between volume and weight gain and abalone age.

Conclusions

After analyzing all data and see the different interactions between variables, study could have failed because of the following reasons:

1. Too much weight was put into physical characteristics to determine abalone age, and the result is investigator's appreciation come in play to group observations leading to miss classifications and ergo wrong data to draw conclusions from it.
2. Physical abalone characteristics are partially helpful to predict age, but when physical characteristics are similar between different observations subjective appreciation from investigators is part of data classification.
3. Even so many variables were include into the investigation, there's a weakness in the model because factors such weather, location or government regulations are not part of it.
4. Similarly, failure to discover causal relations between variables (or assume there's one) or the real impact of the different variables in predicting abalone's age, lead to false discoveries and draw incorrect conclusions from it.

If abalone data were presented with no studies, graphs or tables to understand better the data or variable's behavior, we must the following questions:

1. How data was gather, how sampling was done and see if the population is really representative of the species.
2. Determining how variables were chosen and if those variables really help to understand abalones really and future, and if variables would help to state hypothesis.
3. If physical abalones structures would help to assess assumptions on which statistical inference will be based.
4. If the variables help to ask the right questions about data and no overshadow other type of variables that would help to draw more meaningful conclusions.

When there is not a clear relation or correlation between variables the probability to have wrong conclusions is high, as mentioned before "*ring clarity can be an issue. At the completion of the breeding season sexing abalone can be difficult. Similar difficulties are experienced when trying to determine the sex of immature abalone referred to as infants.*", physical features are not as clear as we might think and that implies that individual appreciation is a fact when data is grouped; at this point certain strong criteria with no room for interpretation should be put in place so investigator can classify data with standard features and no in personal terms. Causality can be accepted if statistics process are run into it and determine correlation degree, but just with see data in tables or graphs assuming causality is a great risk to draw mistaken conclusions from it.

Appendix

R code

1. Preliminaries

```
> View(abalones)
> mydata <- abalones
> str(mydata)
> library(ggplot2)
> library(moments)
> library(gridExtra)
> mydata$VOLUME <- mydata$LENGTH * mydata$DIAM * mydata$HEIGHT
> mydata$RATIO <- mydata$SHUCK / mydata$VOLUME
```

Programming code 1(a)

```
> sumwhole <- summary(mydata)
> t2 <- data.frame(mydata$LENGTH, mydata$DIAM, mydata$HEIGHT, mydata$WHOLE, m
ydata$SHUCK, mydata$RINGS)
> skew(t2)
> kurtosi(t2)
> par(mfrow = c(1, 2))
> hist(mydata$LENGTH, main = "Histogram Length", cex.main=1.5, col="dodgerblu
e3", xlab = "Abalone Leng", ylab = "Frequency", cex.lab=0.8, cex.axis=0.8)
> boxplot(mydata$LENGTH, main = "Boxplot Length", col="firebrick3", pch=16)
>
> par(mfrow = c(1,1))
> par(mfrow = c(2,2))
> hist(mydata$DIAM, main = "Histogram Diam", cex.main=1.5, col="dodgerblue3",
xlab = "Abalone Diam", ylab = "Frequency", cex.lab=0.8, cex.axis=0.8)
> hist(mydata$HEIGHT, main = "Histogram Height", cex.main=1.5, col="dodgerblu
e3", xlab = "Abalone Height", ylab = "Frequency", cex.lab=0.8, cex.axis=0.8)
> boxplot(mydata$DIAM, main = "Boxplot Diam", col="firebrick3", pch=16)
> boxplot(mydata$HEIGHT, main = "Boxplot Height", col="firebrick3", pch=16)
>
> par(mfrow = c(1,1))
> par(mfrow = c(2,3))
> hist(mydata$WHOLE, main = "Histogram Whole", cex.main=1.5, col="dodgerblue3
", xlab = "Abalone Whole", ylab = "Frequency", cex.lab=0.8, cex.axis=0.8)
> hist(mydata$SHUCK, main = "Histogram Shuck", cex.main=1.5, col="dodgerblue3
", xlab = "Abalone Shuck", ylab = "Frequency", cex.lab=0.8, cex.axis=0.8)
> hist(mydata$RINGS, main = "Histogram Rings", cex.main=1.5, col="dodgerblue3
", xlab = "Abalone Rings", ylab = "Frequency", cex.lab=0.8, cex.axis=0.8)
> boxplot(mydata$WHOLE, main = "Boxplot Whole", col="firebrick3", pch=16)
> boxplot(mydata$SHUCK, main = "Boxplot Shuck", col="firebrick3", pch=16)
> boxplot(mydata$RINGS, main = "Boxplot Rings", col="firebrick3", pch=16)
>
> skew(mydata$LENGTH)
> skew(mydata$DIAM)
> skew(mydata$HEIGHT)
> skew(mydata$WHOLE)
> skew(mydata$SHUCK)
> skew(mydata$RINGS)
> kurtosi(mydata$LENGTH)
> kurtosi(mydata$DIAM)
> kurtosi(mydata$HEIGHT)
> kurtosi(mydata$WHOLE)
> kurtosi(mydata$SHUCK)
```

```
> kurtosi(mydata$RINGS)
```

Programming code 1(b)

```
> sexvsclass <- table(mydata$SEX, mydata$CLASS)
> rownames(sexvsclass) <- c("Female", "Infant", "Male")
> addmargins((sexvsclass))
> prob <- addmargins(prop.table(sexvsclass))
> round(prob, digits = 2)
> barplot(table(mydata$SEX, mydata$CLASS)[c(2, 1, 3),], beside = TRUE, main =
"Comparison Abalone Sex Frequencies", xlab = "Age Class", ylab = "Frequency",
cex.lab = .8, cex.axis = .7, col = c("dodgerblue4", "firebrick2", "lightgolde
nrod2"), cex.names = .8, legend.text = c("Infant", "Female", "Male"),)
```

Programming cod 1(c)

```
> set.seed(123)
> work <- mydata[sample(1:nrow(mydata), 200, replace = FALSE),]
> plot(work[,2:6], panel=panel.smooth, pch=18, col="dodgerblue3", lwd=1.5)
```

Programming code 2(a)

```
> df1a <- data.frame(mydata$VOLUME, mydata$WHOLE)
> plot(df1a, main = "Comparing Whole Abalone to Volume", cex.main=1.5, xlab =
"Abalone Volume", ylab = "Abalone whole weight", cex.lab=0.8, cex.axis=0.8, p
ch= 16, col="dodgerblue3", panel.first = grid(15), frame.plot = FALSE)
```

Programming code 2(b)

```
> df <- data.frame(mydata$SHUCK, mydata$WHOLE)
> m <- mydata$SHUCK/mydata$WHOLE
> total <- cbind(df,m)
> colnames(total) <- c("Shuck", "Volume", "Ratio")
> max2b <- max(m)
> plot(mydata$WHOLE, mydata$SHUCK, main = "Shuck in function of Abalon
e whole wieght", cex=.8, cex.main = 1.5, xlab = "Whole weight", ylab =
"Shuck", cex.lab = 0.8, cex.axis = 0.8, pch = 16, col = "dodgerblue3",
panel.first = grid(15), frame.plot = FALSE)
> abline(a=0, b=max2b, col="firebrick3", lwd=2, lty=1)
```

Programming code 3(a)

```
> tableq3 <- data.frame(mydata$SEX, mydata$RATIO)
> tf <- subset(tableq3, mydata$SEX =="F")
> ti <- subset(tableq3, mydata$SEX =="I")
> tm <- subset(tableq3, mydata$SEX =="M")
> par(mfrow = c(3, 3))
> hist(tf$mydata.RATIO, main = "Histogram Female Ratio", panel.first=grid(15)
, xlab = "Shuck vs Volume Ratio", ylab = "Frequency", cex.main=1.5, cex.lab=0
.8, cex.axis=0.8, col="dodgerblue3", plot=TRUE, xlim = c(0.05,0.34))
> hist(ti$mydata.RATIO, main = "Histogram Infant Ratio", panel.first=grid(15)
, xlab = "Shuck vs Volume Ratio", ylab = "Frequency", cex.main=1.5, cex.lab=0
.8, cex.axis=0.8, col="dodgerblue3", plot=TRUE, xlim = c(0.05,0.30), ylim = c
(0,100))
> hist(tm$mydata.RATIO, main = "Histogram Male Ratio", panel.first=grid(15),
xlab = "Shuck vs Volume Ratio", ylab = "Frequency", cex.main=1.5, cex.lab=0.8
```

```
, cex.axis=0.8, col="dodgerblue3", plot=TRUE, xlim = c(0.05,0.30), ylim = c(0
,130))
> boxplot(tf$mydata.RATIO, horizontal=TRUE, main="Boxplot Female Ratio", cex.
main=1.5, xlab="Ratio Shuck vs Volume", cex.lab=0.8, cex.axis=0.8, pch=16, lw
d=1.7, col="dodgerblue3", notch = TRUE, staplewex = 0.5)
> boxplot(ti$mydata.RATIO, horizontal=TRUE, main="Boxplot Infant Ratio", cex.
main=1.5, xlab="Ratio Shuck vs Volume", cex.lab=0.8, cex.axis=0.8, pch=16, lw
d=1.7, col="dodgerblue3", notch = TRUE, staplewex = 0.5)
> boxplot(tm$mydata.RATIO, horizontal=TRUE, main="Boxplot Male Ratio", cex.ma
in=1.5, xlab="Ratio Shuck vs Volume", cex.lab=0.8, cex.axis=0.8, pch=16, lwd=
1.7, col="dodgerblue3", notch = TRUE, staplewex = 0.5)
> qqnorm(tf$mydata.RATIO, pch=16, col="dodgerblue3", panel.first=grid(20), ma
in = "QQ-Plot Graph Female Ratio", cex.axis=0.7, cex.lab=0.8)
> qqline(tf$mydata.RATIO, col="firebrick3", lwd=2)
> qqnorm(ti$mydata.RATIO, pch=16, col="dodgerblue3", panel.first=grid(20), ma
in = "QQ-Plot Graph Infant Ratio", cex.axis=0.7, cex.lab=0.8)
> qqline(ti$mydata.RATIO, col="firebrick3", lwd=2)
> qqnorm(tm$mydata.RATIO, pch=16, col="dodgerblue3", panel.first=grid(20), ma
in = "QQ-Plot Graph Male Ratio", cex.axis=0.7, cex.lab=0.8)
> qqline(tm$mydata.RATIO, col="firebrick3", lwd=2)
> par(mfrow = c(1, 1))
```

Programming code 3(b)

```
> outfemale <- boxplot.stats(tf$mydata.RATIO, coef = 1.5, do.conf = TRUE, do.
out=TRUE)
> outinfant <-  boxplot.stats(ti$mydata.RATIO, coef = 1.5, do.conf = TRUE, do
.out=TRUE)
> outmale <-  boxplot.stats(tm$mydata.RATIO, coef = 1.5, do.conf = TRUE, do.o
ut=TRUE)
```

Programming cod 4(a)

```
> grid.arrange(
+ ggplot(
+ mydata, aes(x=factor(mydata$CLASS), y= mydata$VOLUME, fill=mydata$CLASS))+
+ geom_boxplot(outlier.color = "firebrick3", outlier.shape = 16, outlier.size
= 1.2)+
+ labs(title="Abalone volume by class", x="Class", y="Volume")+
+ theme(legend.position = "none"),
+ ggplot(
+ mydata, aes(x=factor(mydata$CLASS), y= mydata$WHOLE, fill=mydata$CLASS))+
+ geom_boxplot(outlier.color = "firebrick3", outlier.shape = 16, outlier.size
= 1.2)+
+ labs(title="Whole by class", x="Class", y="Whole")+
+ theme(legend.position = "none"),
+ nrow=1, top= "Comparing Abalones classes with Volume & Whole"
+ )


> grid.arrange(
+ ggplot(
+ mydata, aes(x=factor(mydata$RINGS), y=mydata$VOLUME, color=mydata$RINGS))+
+ geom_point()+
+ theme(legend.position = "none")+
```

```
+ labs(title="Volume by rings", x="Rings", y="Volume"),
+ ggplot(
+ mydata, aes(x=factor(mydata$RINGS), y=mydata$WHOLE, color=mydata$RINGS))+
+ geom_point()+
+ theme(legend.position = "none")+
+ labs(title="Whole by rings", x="Rings", y="Whole"),
+ nrow=1, top= "Comparing Abalones rings with Volume & Whole"
+ )
```

Programming cod 5(a)

```
> aggdatavolume <- aggregate(mydata$VOLUME, by=list(mydata$SEX, mydata$CLASS)
, FUN=mean, na.rm=TRUE)
> a <- tapply(aggdatavolume$x, list(aggdatavolume$Group.1, aggdatavolume$Grou
p.2), mean)
> rownames(a) <- c("Female", "Infant", "Male")
> colnames(a) <- c("A1", "A2", "A3", "A4", "A5", "A6")
> round(a, digits = 2)
> aggdataratio <- aggregate(mydata$RATIO, by=list(mydata$SEX, mydata$CLASS),
FUN=mean, na.rm=TRUE)
> b <- tapply(aggdataratio$x, list(aggdataratio$Group.1, aggdataratio$Group.2
), mean)
> rownames(b) <- c("Female", "Infant", "Male")
> colnames(b) <- c("A1", "A2", "A3", "A4", "A5", "A6")
> round(b, digits = 4)
```

Programming cod 5(b)

```
> outv <- aggregate(VOLUME ~ SEX + CLASS, data = mydata, mean)
> outr <- aggregate(RATIO ~ SEX + CLASS, data = mydata, mean)

> grid.arrange(
+ ggplot(data = outv, aes(x = CLASS, y = VOLUME, group = SEX, colour = SEX))+
+ geom_line() + geom_point(size = 2)+
+ ggtitle("Plot of Mean VOLUME versus CLASS for Three Sexes"),
+ ggplot(data = outr, aes(x = CLASS, y = RATIO, group = SEX, colour = SEX))+
+ geom_line() + geom_point(size = 2)+
+ ggtitle("Plot of Mean RATIO versus CLASS for Three Sexes")
+ )
```

Reference:

1. Hartwig, F., & Dearing, B. E. (1979). *Exploratory Data Analysis*. Beverly Hills, Calif: SAGE Publications, Inc.
2. Tukey JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.
3. Morgenthaler, S. (2009). *Exploratory data analysis*. Wires Computational Statistics, 1(1), 33-44. doi:10.1002/wics.2
4. Blacklip Abalone (Haliotis rubra). (n.d.). Retrieved April 16, 2017, from http://www.dpi.nsw.gov.au/__data/assets/pdf_file/0009/375858/BlacklipAbalone.pdf
5. Facts About Abalone. (n.d.). Retrieved April 20, 2017, from http://www.fishtech.com/facts.html
6. Abalone Introduction. (n.d.). Retrieved April 20, 2017, from http://www.marinebio.net/marinescience/06future/abintro.htm
7. Diversity, C. B. (n.d.). PETITION TO LIST THE BLACK ABALONE (HALIOTIS CHRACHERODII) AS THREATENED OR ENDANGERED UNDER THE ENDANGERED SPECIES ACT . Retrieved April 22, 2017, from http://www.biologicaldiversity.org/species/invertebrates/black_abalone/pdfs/Black-Ab-Petition-12-21-06.pdf