Abalone EDA

Herman, Scott
Predict 401 Section 58
TITLE: Data Assignment 1

INTRODUCTION

The purpose of the assignment below is to review the study of abalone data to determine why a pervious study was unsuccessful in predicting age based upon their characteristics. The first step in exploring the abalone data set consists of defining and understanding the given variables and the relationship between each. The data set includes ten different variables defined by various abalone physical characteristics and consists of 4,141 observations. A random sample of 500 observations was utilized throughout this analysis.

RESULTS

Figure 1 below presents the summary statistics of our sample.

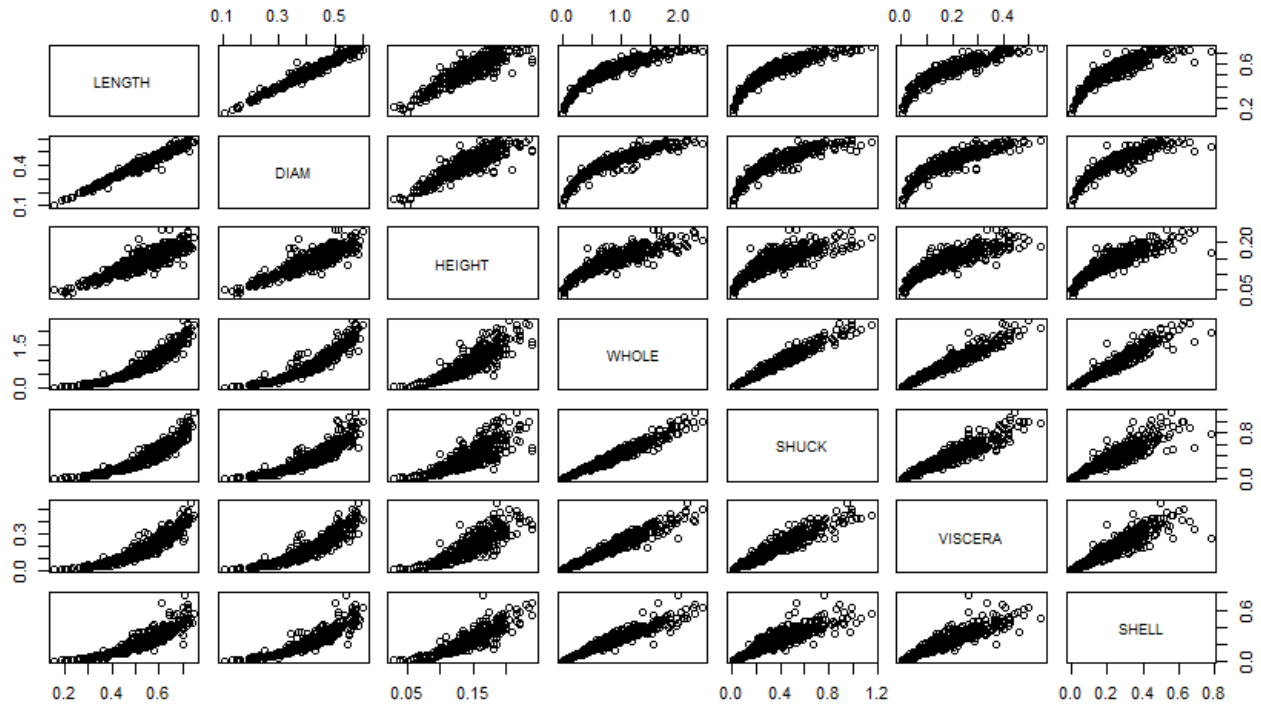
Figure 1: Summary of the Data

```
SEX         LENGTH           DIAM            HEIGHT           WHOLE
F:157   Min.   :0.1550   Min.   :0.1050   Min.   :0.0300   Min.   :0.0175
I:153   1st Qu.:0.4600   1st Qu.:0.3538   1st Qu.:0.1150   1st Qu.:0.4740
M:190   Median :0.5400   Median :0.4200   Median :0.1400   Median :0.7865
        Mean   :0.5263   Mean   :0.4088   Mean   :0.1392   Mean   :0.8296
        3rd Qu.:0.6100   3rd Qu.:0.4763   3rd Qu.:0.1650   3rd Qu.:1.1405
        Max.   :0.7450   Max.   :0.6000   Max.   :0.2400   Max.   :2.3810
     SHUCK            VISCERA           SHELL            RINGS          CLASS
Min.   :0.0050   Min.   :0.00350   Min.   :0.0050   Min.   : 3.000   A1: 49
1st Qu.:0.1979   1st Qu.:0.09988   1st Qu.:0.1350   1st Qu.: 8.000   A2:114
Median :0.3285   Median :0.17150   Median :0.2362   Median : 9.000   A3:176
Mean   :0.3617   Mean   :0.18223   Mean   :0.2354   Mean   : 9.786   A4: 91
3rd Qu.:0.5091   3rd Qu.:0.25662   3rd Qu.:0.3200   3rd Qu.:11.000   A5: 35
Max.   :1.1565   Max.   :0.55000   Max.   :0.7800   Max.   :21.000   A6: 35
```

From the summary above, it is difficult to make any initial conclusions on the variables' relationship. What this does tell us is that we have three different variable types. Sex is a nominal measurement, Class is of ordinal level and the remaining eight variables are ratio-level measurements. In order to explore theses further, a scatterplot matrix was utilized to plot the relationships between our eight ratio-level variables, shown below in Figure 2. This scatterplot indicates a mostly positive linear relationship between these variables. Length, Diameter, Height and Whole weight possess the most obvious linear correlation and appear to be evenly dispersed. On the other hand, while the relationship between Shuck, Viscera and Shell are also positive, it does appear that the higher end values appear to have an increased variation among them. Meaning that Shell size that is significantly larger than the mean/median shell size, does not necessarily indicate a Viscera or Shuck value that is equally larger. This tells us that there is more exploring to do among these variables to further understand this relationship.

Figure 2: Scatterplot

Abalone EDA



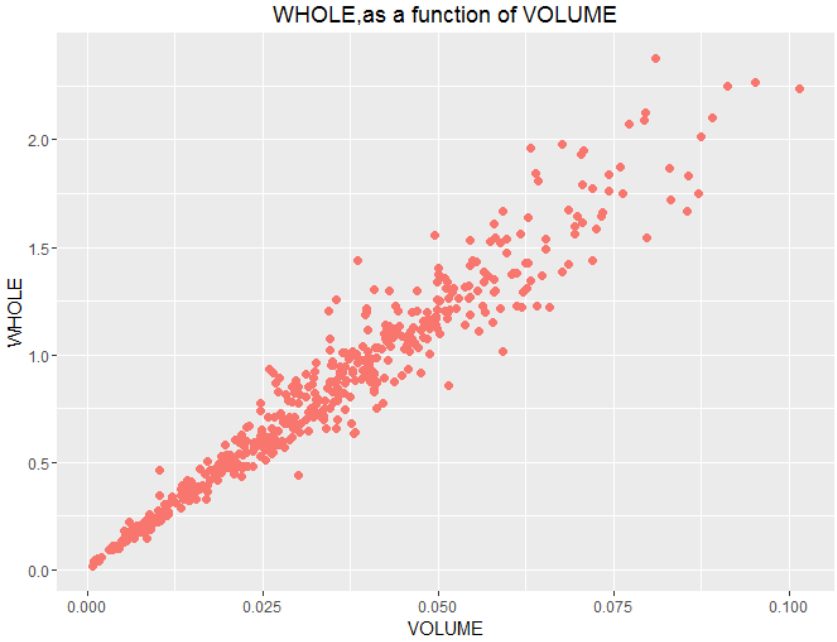
To understand the relationship between our two non-ratio-level variables, a frequency matrix was utilized. Figure 3 displays a table showcasing abalone Sex broken down by each of the six different classes, which signify their age. Here, the last column indicates that we have a nearly equal number of Females and Infants in our sample. Also, in exploring the bottom row, it appears that our sample contains significantly more abalone classified as A2 and A3 than any of the remaining classes. Perhaps this may need some additional exploration.

Figure 3: Sex by Class Frequency Matrix

	A1	A2	A3	A4	A5	A6	Sum
F	2	29	61	40	10	15	157
I	41	61	35	8	3	5	153
M	6	24	80	43	22	15	190
Sum	49	114	176	91	35	35	500

In an attempt to explore this data further, we have created a new variable, Volume, which is defined by the product of an abalone's Length, Height and Diameter. The addition of this variable may allow us to look further into these physical characteristics together, to understand if a notable relationship exists. Figure 4 shows the relationship between an abalone's volume and whole weight. Again it appears that these two variables have a positive linear relationship, though the wedge shape suggests that there is significantly more variation in whole weight as the volume increases. As volume increases the values for whole weight have more variation among them and prove to show more of an uneven dispersion.

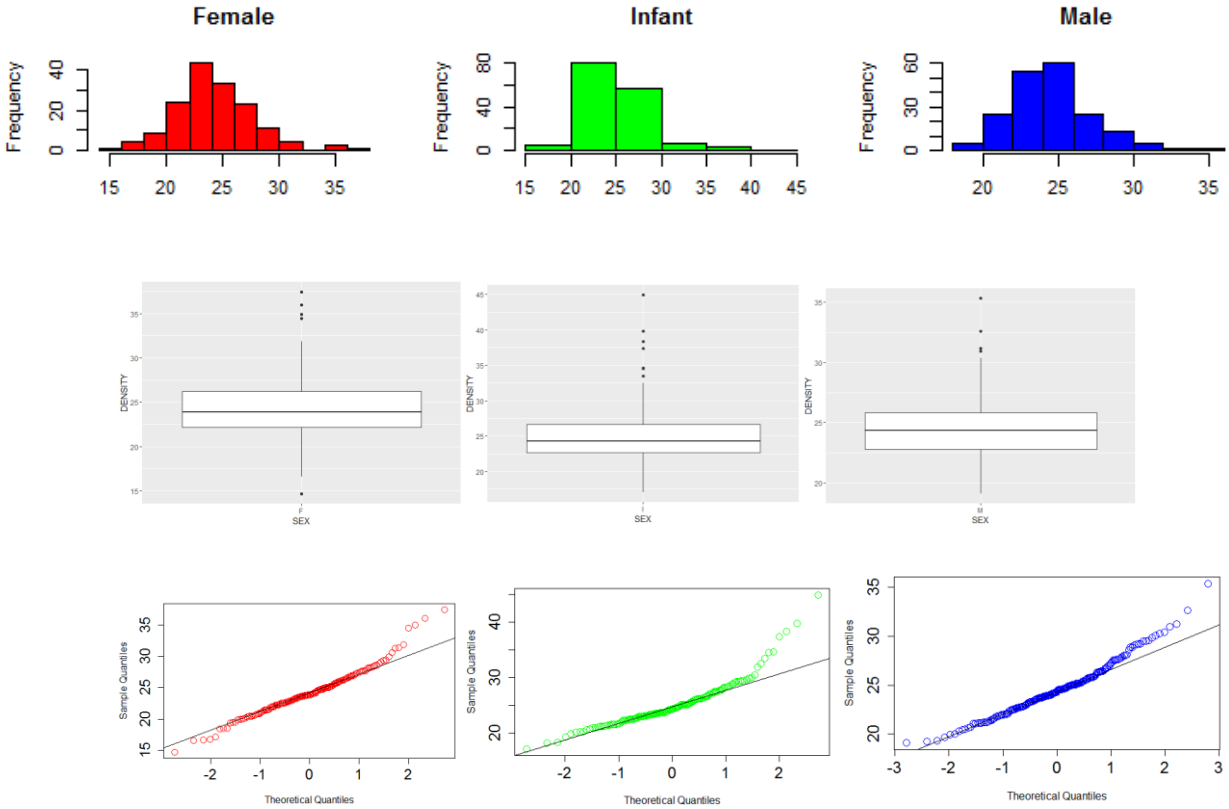
Figure 4: Whole as a Function of Volume



Again, to dig further in an attempt to identify a telling a relationship, we will create one additional variable for Density. Density will be defined by dividing each abalone's Whole weight by Volume. To explore the relationship between abalone Density and Sex, Figure 5 displays three different types of plots to understand how these variables may be related. In the first row, the histograms show us that the Males and Females in the sample are normally distributed in terms of Density, while the infants are skewed slightly to the right. The boxplots, below, tell a similar story, yet the Infants appear to have a greater number of outliers on the higher-end of Density. Lastly, the black line in the QQ Plots reveal where these variables would be normally distributed. In each of the three Sexes, there appears to be greater variance as Density increases. From the figures below, the one thing that stands out are the outliers for Infants in the boxplot. Female and Male abalone appear to have an equal dispersion in terms of Density, so perhaps the number of outliers with Infants, in terms of Volume indicates that physical characteristics may not be the best determinant of abalone age. This will lead to additional exploration.

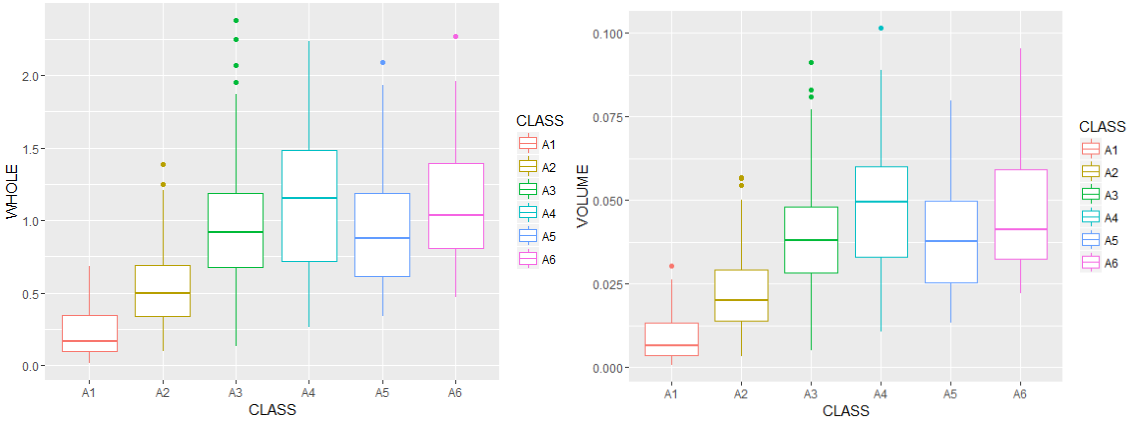
Abalone EDA

Figure 5: Density as a Function of Sex



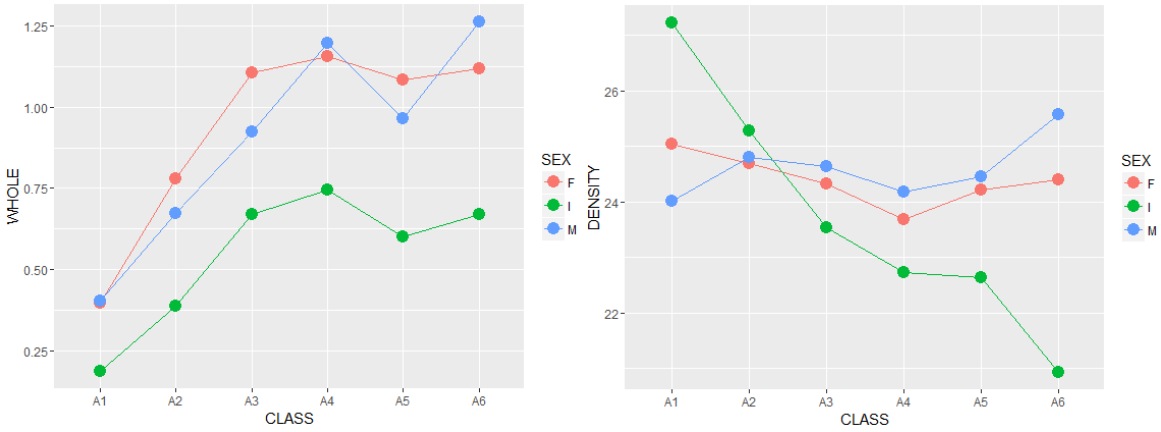
Since Density alone does not give us any clear significance in determining abalone Sex, we're going to look at the relationship between Whole weight and Volume below. Figure 6 shows two boxplots showcasing this below. In looking at the two plots, it appears that most outliers exist in the A2 and A3 class, in terms of both Whole weight and Volume.

Figure 6: Volume and Whole weight considered by Class



Because it does appear that a number of outliers do exist, we've chosen to take the mean of Whole weight and Volume to determine if these outliers are significant in terms of class. Figure 7 below shows this relationship, and what it tells us is that most of the invariability lies within the Infants. This demonstrates that the determination of abalone Sex based on physical characteristics is extremely difficult.

Figure 7: Mean Whole weight and Mean Density per Class, Sex Comparative



CONCLUSION

Based on the analysis and charts displayed above, it appears that determining abalone age is extremely difficult based upon physical characteristics alone. There Is no clear relationship between abalone age and any of the other characteristics displayed above due to the number of outliers and variance that occurs between them. The same can be said about determining abalone Sex. The analysis above also shows no clear indication between abalone Sex and any of the other variables due to the variance and number of outliers. This leads us to conclude that physical characteristics are not a clear indication of abalone age, or sex, and that additional measurements and statistical analyses are required.

Abalone EDA

APPENDIX: ORG = 10 POINTS

```
abalone <- read.csv("abalone.csv", sep = "")
str(abalone)
head(abalone)
set.seed(123)
index<-sample(1:nrow(abalone),500)
mydata<-abalone[index, ]
str(mydata)
head(mydata)
tail(mydata)
summary(mydata)
require(moments)
require(ggplot2)
plot(mydata[,2:8])
addmargins(table(mydata$SEX,mydata$CLASS))
mydata$VOLUME<-mydata[, 2]*mydata[, 3]*mydata[, 4]
head(mydata$VOLUME)
data.frame(mydata$Volume)
ggplot(data = mydata, aes(x = VOLUME, y = WHOLE)) +
  geom_point(aes(color = "blue"),size = 2) + ggtitle("WHOLE,as a function of VOLUME")

mydata$DENSITY<-mydata[, 5]/mydata[, 11]
head(mydata$DENSITY)
data.frame(mydata$DENSITY)

par(mfrow=c(3,3), oma=c(0,0,0,0))
hist(mydata$DENSITY[mydata$SEX == "F"], main = "Female",col="red")
hist(mydata$DENSITY[mydata$SEX == "I"], main = "Infant",col="green")
hist(mydata$DENSITY[mydata$SEX == "M"], main = "Male",col="blue")
ggplot(mydata[mydata$SEX == "F", ], aes(x = SEX, y = DENSITY),col="red") +
  geom_boxplot()
ggplot(mydata[mydata$SEX == "I", ], aes(x = SEX, y = DENSITY)) +
  geom_boxplot()
ggplot(mydata[mydata$SEX == "M", ], aes(x = SEX, y = DENSITY)) +
  geom_boxplot()
qqnorm(mydata[mydata[,1] == "I", 12], col = "green", main = NULL,
      cex.axis = 1.5, cex = 1.5, datax = FALSE)
qqline(mydata[mydata[,1] == "I", 12], datax = FALSE)
qqnorm(mydata[mydata[,1] == "F", 12], col = "red", main = NULL,
      cex.axis = 1.5, cex = 1.5, datax = FALSE)
qqline(mydata[mydata[,1] == "F", 12], datax = FALSE)
qqnorm(mydata[mydata[,1] == "M", 12], col = "blue", main = NULL,
      cex.axis = 1.5, cex = 1.5, datax = FALSE)
qqline(mydata[mydata[,1] == "M", 12], datax = FALSE)
```

Abalone EDA

```
mtext("QQ plot, Density (comparative)", side = 3, line = -2,
     outer = TRUE, cex = 1.5)
par(mfrow = c(1, 1))

par(mfrow=c(1,2))
ggplot(mydata, aes(x = CLASS, y = VOLUME, col = CLASS))+
  geom_boxplot()
ggplot(mydata, aes(x = CLASS, y = WHOLE, col = CLASS)) +
  geom_boxplot()
par(mfrow=c(1,1))

out=aggregate(WHOLE~SEX+CLASS, data=mydata,mean)
ggplot(data=out ,aes(x=CLASS, y=WHOLE, group=SEX, colour=SEX))+
  geom_line()+geom_point(size=4)+
  ggtitle("Plot showing Mean WHOLE versus CLASS for Three Sexes")


out=aggregate(DENSITY~SEX+CLASS, data=mydata,mean)
ggplot(data=out,aes(x=CLASS, y=DENSITY, group=SEX, colour = SEX))+
  geom_line()+geom_point(size=4)+
  ggtitle("Plot showing Mean DENSITY versus CLASS for Three Sexes")
```