Data Analysis - Worksheet 4

Jiancong Zhu

2024-10-17

Worksheet Week 1

Please review the following document which provides useful background information pertinent to this assignment.Data Analysis Project Instruction.pdf

Instructions for the first assignment appear in the Rmd template below.The required abalones.csv data are given below.

Save the abalones data on your computer without opening the file in Excel.Do not change the order of the observations in the original file. If you do, your answers may not correspond to the answer sheet and sample report.

Part 1 Enter your data into a DataFrame and set up your datatypes and values. *** Section 0 *** (16 points)

1. (a) (2 points) Use the getwd() to find your working directory.

```
getwd()
```

```
## [1] "/Users/jiancongzhu/Downloads"
```

2. (b) (2 points)Use the setwd(. . .) to set your working directory, ensure that your R program and your csv file are in the same directory.

```
setwd("/Users/jiancongzhu/Desktop/MSDS401")
```

3. (c) (2 points) Import the following packages in your R studio ggplot2, gridExtra, psych, knitr, moments

```
#install.packages("ggplot2")
#install.packages("gridExtra")
#install.packages("psych")
install.packages("knitr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/gy/2k9pcch950b5p98c4__0z5fw0000gn/T//Rtmp33Febk/downloaded_packages
```

```r
#install.packages("moments")
```

4. (d) (2 points) Load the “ggplot2”, “gridExtra”,”psych” and “knitr” packages

```r
library(ggplot2)
library(gridExtra)
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(knitr)
```

5. (e) (2 points) Read-in the abalones dataset, defining a new data frame, “mydata,”

```r
mydata <- read.csv(file.path("~/Desktop/MSDS401", "abalones.csv"), sep=",", stringsAsFactors = TRUE
```

6. (f) (2 points) calculate new variables, VOLUME and RATIO.

```r
mydata$VOLUME <- mydata$LENGTH * mydata$DIAM *mydata$HEIGHT
mydata$RATIO <- mydata$SHUCK / mydata$VOLUME
```

7. (g) (2 points) Run str(mydata) and ensure that your data have been read properly. You should expect 1036 observations and 8 variables.

```r
str(mydata)
```

```
## 'data.frame':    1036 obs. of  10 variables:
##  $ SEX   : Factor w/ 3 levels "F","I","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38 4.69 21.19 ...
##  $ SHUCK : num  4.31 1.19 44 2.25 9.88 ...
##  $ RINGS : int  6 4 6 3 6 6 5 6 5 6 ...
##  $ CLASS : Factor w/ 5 levels "A1","A2","A3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ VOLUME: num  28.7 8.1 163.4 12.2 59.7 ...
##  $ RATIO : num  0.15 0.147 0.269 0.185 0.165 ...
```

8. (h) (2 points) Change the values of “SEX”, there are many ways to do this. We will change the levels of SEX from I,M,F to Infant,Male and Female.

```r
levels(mydata$SEX)
```

```
## [1] "F" "I" "M"
```

```r
levels(mydata$SEX) <-c("Infant","Male","Female")
```

7. Use the function levels(mydata$SEX), you should first run the command to see the order of the levels (values), then create a vector with the new levels, in the correct order. To do this use the c(,,,) and insert the Values: INFANT, MALE, FEMALE, in the right order.

##Part 2 Review the descriptive statistics. ***##### Section 1: (19 points) Summarizing the data.***

```r
#(a)    (2 points) Use *summary()* to obtain and present descriptive statistics from mydata, specifical
summary(mydata)
```

```
##      SEX          LENGTH           DIAM            HEIGHT     
##  Infant:326   Min.   : 2.73   Min.   : 1.995   Min.   :0.525  
##  Male  :329   1st Qu.: 9.45   1st Qu.: 7.350   1st Qu.:2.415  
##  Female:381   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   A1:108  
##  1st Qu.: 56.484   1st Qu.: 23.3006   1st Qu.: 8.000   A2:236  
##  Median :101.344   Median : 42.5700   Median : 9.000   A3:329  
##  Mean   :105.832   Mean   : 45.4396   Mean   : 9.993   A4:188  
##  3rd Qu.:150.319   3rd Qu.: 64.2897   3rd Qu.:11.000   A5:175  
##  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  
```

```r
#(b)    (2 points) Use *describeBy()* to obtain all the descriptive statistics.
describeBy(mydata)
```

```
## Warning in describeBy(mydata): no grouping variable requested
```

```
##         vars    n   mean    sd median trimmed   mad  min    max  range  skew
## SEX*       1 1036   2.05  0.82   2.00    2.07  1.48 1.00   3.00   2.00 -0.10
## LENGTH     2 1036  11.08  2.51  11.45   11.26  2.49 2.73  16.80  14.07 -0.67
## DIAM       3 1036   8.62  2.08   8.93    8.76  2.02 2.00  13.23  11.23 -0.62
## HEIGHT     4 1036   2.95  0.81   2.94    2.96  0.78 0.52   4.93   4.41 -0.23
## WHOLE      5 1036 105.83 61.92 101.34  102.39 69.64 1.62 315.75 314.12  0.47
## SHUCK      6 1036  45.44 27.72  42.57   43.61 30.23 0.56 157.08 156.52  0.64
## RINGS      7 1036   9.99  3.32   9.00    9.67  2.97 3.00  25.00  22.00  1.24
## CLASS*     8 1036   3.08  1.22   3.00    3.10  1.48 1.00   5.00   4.00  0.05
```

```
## VOLUME    9 1036 326.80 194.71 307.36  316.24 218.44 3.61 995.67 992.06  0.44
## RATIO    10 1036   0.14   0.03   0.14    0.14   0.03 0.07   0.31   0.24  0.71
##        kurtosis   se
## SEX*      -1.52 0.03
## LENGTH     0.16 0.08
## DIAM       0.00 0.06
## HEIGHT    -0.18 0.03
## WHOLE     -0.29 1.92
## SHUCK      0.20 0.86
## RINGS      2.65 0.10
## CLASS*    -0.91 0.04
## VOLUME    -0.48 6.05
## RATIO      1.67 0.00
```

(c) (2 points) Compare the information received in the two steps above. Explain your answer here

*** describe(mydata) offers more detailed statistics compared to summary(mydata). It provides insights into the distribution of the data (like skewness and kurtosis) and the spread of values (standard deviation and range), which are not available in the simple summary.What's more, summary(mydata) is useful for a quick overview and to identify any missing values. On the other hand, describeBy(mydata) is more informative when we need to understand the shape and distribution of our data, making it more suitable for detailed exploratory data analysis. ***

(d) (2 points) Identify which of the variables have the highest dispersion. What does this mean about the specific characteristic of the abalohi fnes?

*** The variable "WHOLE" has the highest dispersion since it has the largest standard deviation of 61.9.This indicates a wide range of variability in the whole weights of abalones in the dataset. This variability shows that the abalones differ significantly in size and weight, and likely due to factors such as age, growth conditions, or genetic diversity. ***

```r
#(e)    (2 points) Consider the variable that has the highest dispersion and analyze further by SEX. To
#Use the following function
#   aggregate(cbind(VOLUME,WIDTH,LENGTH,DIAMETER,RATIO)~SEX,data = mydata,sd)
    aggregate(cbind(VOLUME,LENGTH,DIAM,RATIO)~SEX,data=mydata,sd)
```

```
##      SEX   VOLUME   LENGTH     DIAM      RATIO
## 1 Infant 178.1521 1.761329 1.451918 0.02803902
## 2   Male 124.3015 2.355896 1.896151 0.03047035
## 3 Female 178.1719 2.126713 1.742809 0.02926084
```

```r
#Repeat the steps for mean and median.
    aggregate(cbind(VOLUME,LENGTH,DIAM,RATIO)~SEX,data=mydata,mean)
```

```
##      SEX   VOLUME    LENGTH     DIAM     RATIO
## 1 Infant 431.9845 12.379049 9.718298 0.1395934
## 2   Male 175.2994  9.099648 6.927820 0.1448629
## 3 Female 367.6330 11.675394 9.146024 0.1417126
```

```r
    aggregate(cbind(VOLUME,LENGTH,DIAM,RATIO)~SEX,data=mydata,median)
```

```
##      SEX   VOLUME LENGTH  DIAM     RATIO
```

```
## 1 Infant 425.4799 12.705 9.870 0.1380895
## 2   Male 148.4515  9.240 7.035 0.1415999
## 3 Female 366.4346 12.075 9.555 0.1377640
```

```r
#(f)    (2 points) Use *table()* to present a frequency table using CLASS and RINGS. There should be 11
    table(mydata$CLASS, mydata$RINGS)
```

```
##
##        3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   A1   9   8  24  67   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##   A2   0   0   0   0  91 145   0   0   0   0   0   0   0   0   0   0   0   0
##   A3   0   0   0   0   0   0 182 147   0   0   0   0   0   0   0   0   0   0
##   A4   0   0   0   0   0   0   0   0 125  63   0   0   0   0   0   0   0   0
##   A5   0   0   0   0   0   0   0   0   0   0  48  35  27  15  13   8   8   6
##
##       21  22  23  24  25
##   A1   0   0   0   0   0
##   A2   0   0   0   0   0
##   A3   0   0   0   0   0
##   A4   0   0   0   0   0
##   A5   4   1   7   2   1
```

```r
#(g)    (2 points) Repeat and Create another table() to present a frequency table using CLASS, SEX. Do y
    table(mydata$CLASS, mydata$SEX)
```

```
##
##      Infant Male Female
##   A1      5   91     12
##   A2     41  133     62
##   A3    121   65    143
##   A4     82   21     85
##   A5     77   19     79
```

```r
#(h)    (2 points) Repeat and Create another table() to present RINGS and SEX.
    table(mydata$RINGS, mydata$SEX)
```

```
##
##      Infant Male Female
##   3       0    8      1
##   4       0    7      1
##   5       0   22      2
##   6       5   54      8
##   7       7   64     20
##   8      34   69     42
##   9      63   46     73
##   10     58   19     70
##   11     55   14     56
##   12     27    7     29
##   13     20    6     22
##   14     16    4     15
##   15     10    2     15
##   16      5    1      9
```

```
## 17      7   2     4
## 18      4   0     4
## 19      5   1     2
## 20      1   3     2
## 21      2   0     2
## 22      1   0     0
## 23      4   0     3
## 24      1   0     1
## 25      1   0     0
```

Essay Question (3 points): Briefly discuss the variable types and distributional implications such as potential skewness and outliers.

*** The dataset includes categorical variables like SEX and CLASS, and numeric variables such as RINGS, LENGTH, and WHOLE. The distribution of RINGS might be skewed, with certain age groups more common, indicating a potential concentration around specific ages. Outliers could exist in variables like LENGTH or WHOLE, reflecting significant size variations among abalones. The distribution across SEX shows that some age groups are more prevalent in specific sexes, suggesting underlying biological or environmental influences.***

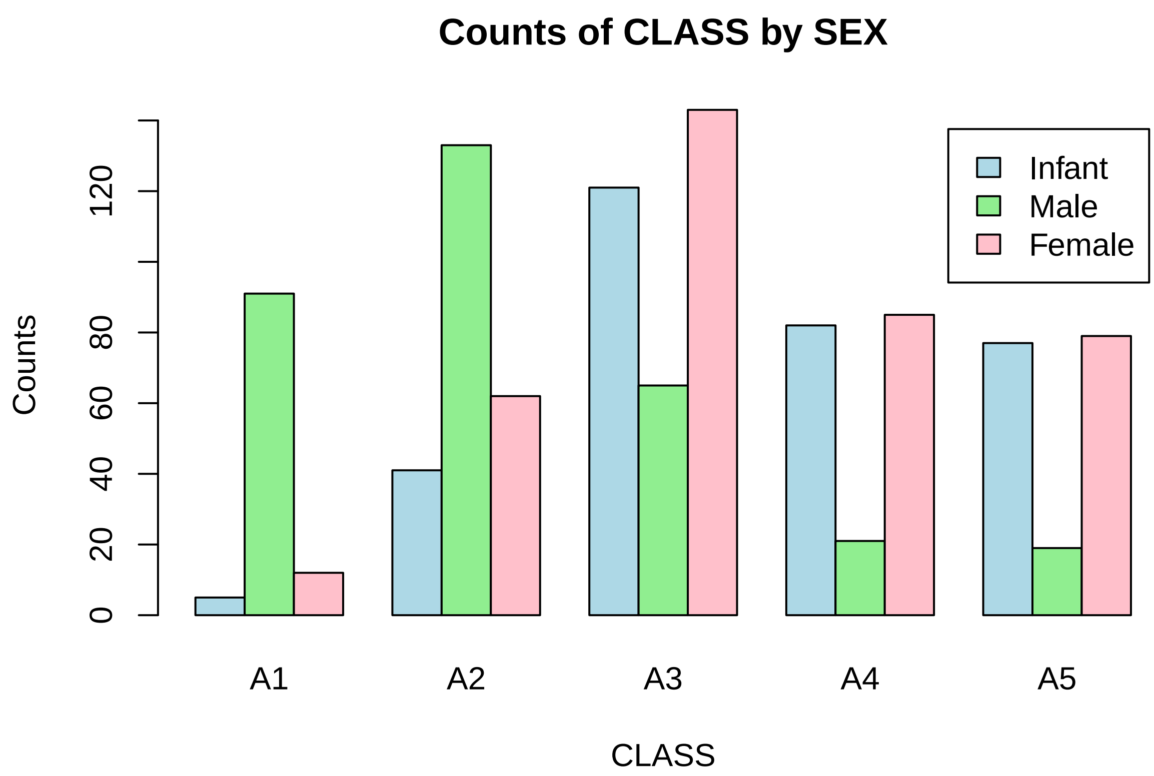
Section 3: (9 points) Getting insights about the data using graphs.

(2 point) Consider the table of counts using SEX and CLASS, you created above. Add margins to this table (Hint: There should be 15 cells in this table plus the marginal totals. Apply *table()* first, then pass the table object to *addmargins()* (Kabacoff Section 7.2 pages 144-147)). Lastly, present a barplot of these data; ignoring the marginal totals.

```r
# Create the table of counts for SEX and CLASS
sex_class_table <- table(mydata$SEX, mydata$CLASS)

# Add margins to the table
sex_class_table_margins <- addmargins(sex_class_table)
print(sex_class_table_margins)
```

```
##
##           A1   A2   A3   A4   A5  Sum
##   Infant   5   41  121   82   77  326
##   Male    91  133   65   21   19  329
##   Female  12   62  143   85   79  381
##   Sum    108  236  329  188  175 1036
```

```r
# Create the barplot for the data, ignoring marginal totals
barplot(sex_class_table, beside = TRUE, legend = TRUE,
        col = c("lightblue", "lightgreen", "pink"),
        main = "Counts of CLASS by SEX", xlab = "CLASS", ylab = "Counts")
```

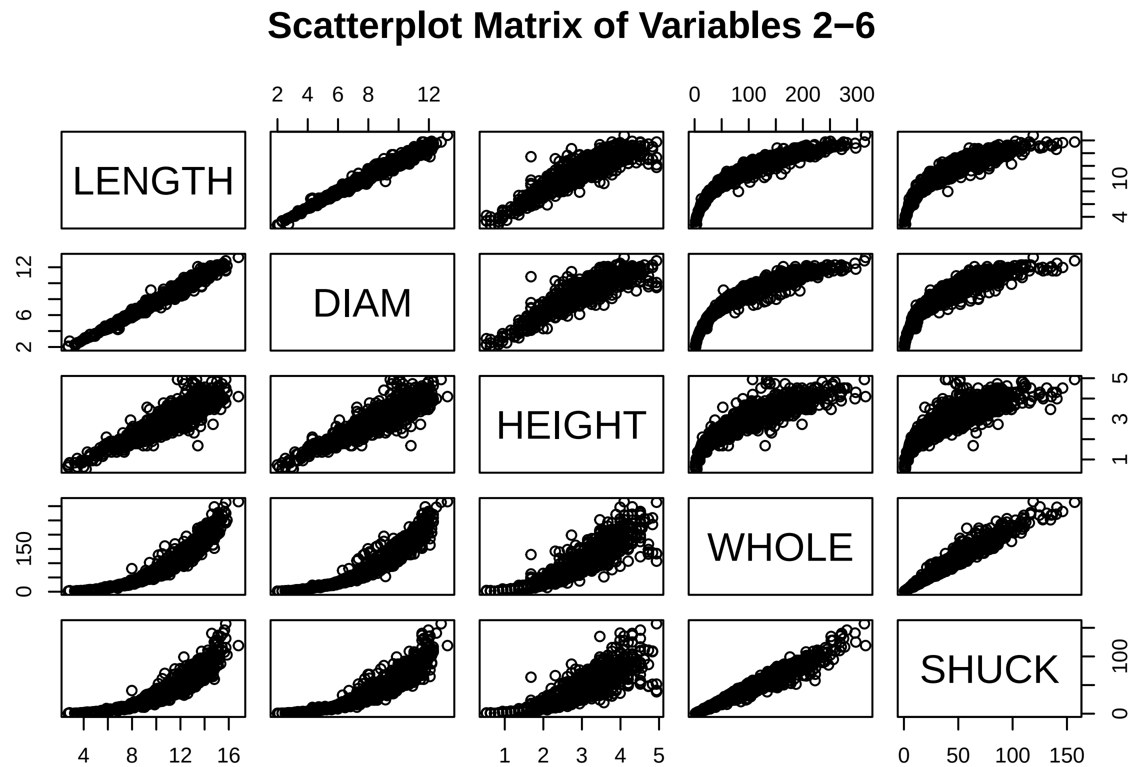


Essay Question (3 points): Discuss the sex distribution of abalones. What stands out about the distribution of abalones by CLASS?

Answer:

(2 points) Using "work", construct a scatterplot matrix of variables 2-6 with *plot(work[, 2:6])* (these are the continuous variables excluding VOLUME and RATIO). The sample "work" will not be used in the remainder of the assignment. **Essay Question (2 points): Discuss the relationships of the variables in your matrix. What information provides to you if any?

```
# Construct a scatterplot matrix of variables 2-6 using "work"
work <- mydata
plot(work[, 2:6], main = "Scatterplot Matrix of Variables 2-6")
```




```
# Essay Question: Discuss the relationships of the variables in your matrix.
# Answer : The scatterplot matrix provides a visual representation of the relationships between the con
```

Section 4: (16 points) Summarizing the data using graphics.

(a) (2 points) Use "mydata" to plot WHOLE versus VOLUME. Color code data points by CLASS.

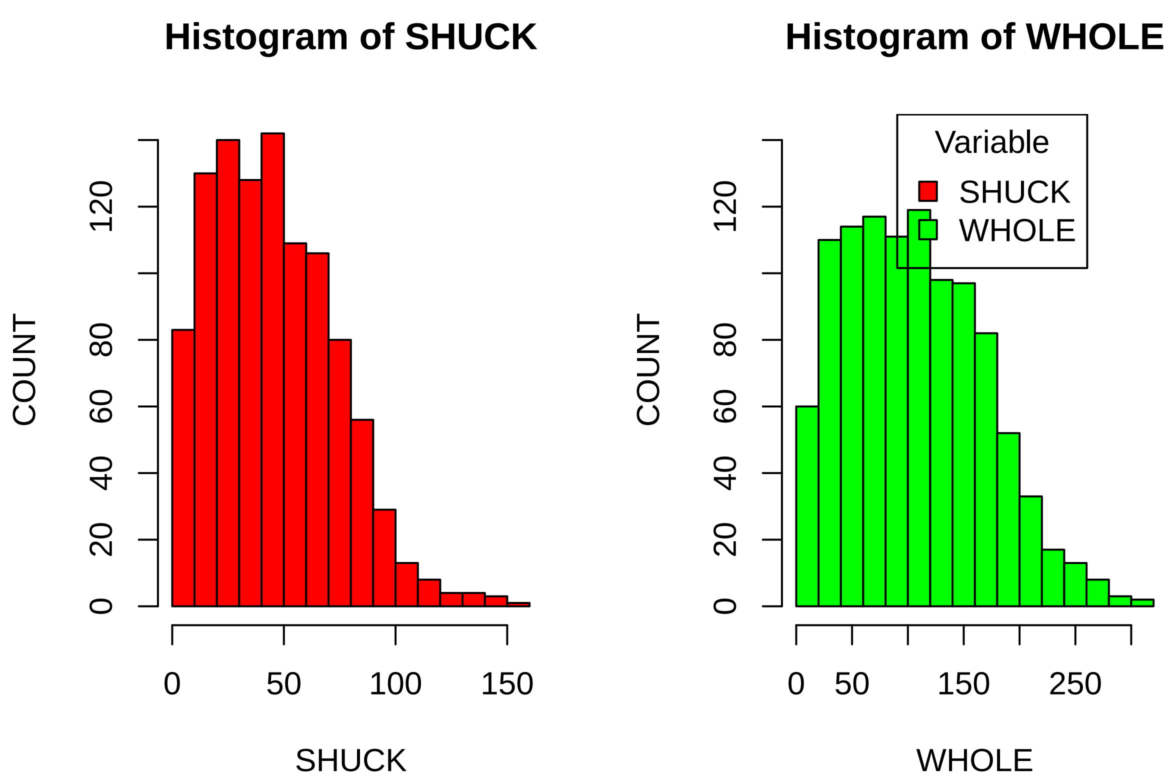
(2 points) Use "mydata" to plot a histogram of SHUCK and a histogram of WHOLE. Present the two histograms in parallel, the one next to the other and ensure that the y-axis units of the two histograms are the same so you can compare the results. Use title, legend, labels for the axis and different colors.

```
# Plot WHOLE versus VOLUME, color coded by CLASS
par(mfrow=c(1,2))
shuck_hist <- hist(mydata$SHUCK, plot = FALSE)
whole_hist <- hist(mydata$WHOLE, plot = FALSE)
y_max <- max(shuck_hist$counts, whole_hist$counts)
y_max
```

```
## [1] 142
```

```
hist(mydata$SHUCK,
     col  = "red",
     main = "Histogram of SHUCK",
     xlab = "SHUCK",
```

```
     ylab = "COUNT",
     ylim =  c(0,y_max))

hist(mydata$WHOLE,
     col  = "green",
     main = "Histogram of WHOLE",
     xlab = "WHOLE",
     ylab = "COUNT",
     ylim =  c(0,y_max))
par(mfrow=c(1,1))
legend("topright", legend = c("SHUCK", "WHOLE"), fill = c("red", "green"), title = "Variable")
```

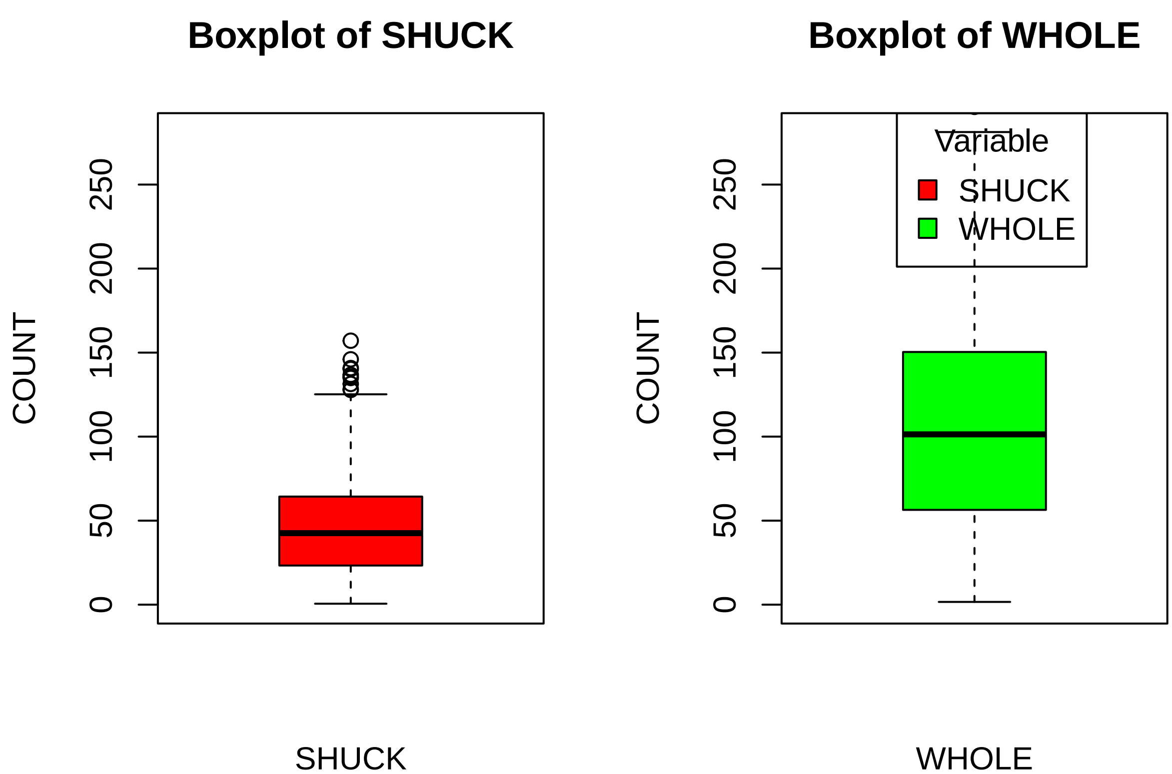


(2 points) Use "mydata" to plot a boxplot of SHUCK and a boxplot of WHOLE. Present the two histograms in parallel, the one next to the other, and ensure that the y-axis units of the two boxplots are the same so you can compare the results. Use title, legend, labels for the axis, and different colors.

```
par(mfrow=c(1,2))
shuck_box <- boxplot(mydata$SHUCK, plot = FALSE)
whole_box <- boxplot(mydata$WHOLE, plot = FALSE)
y_max <- max(shuck_box$stats, whole_box$stats)
y_max
```

```
## [1] 281.25
```

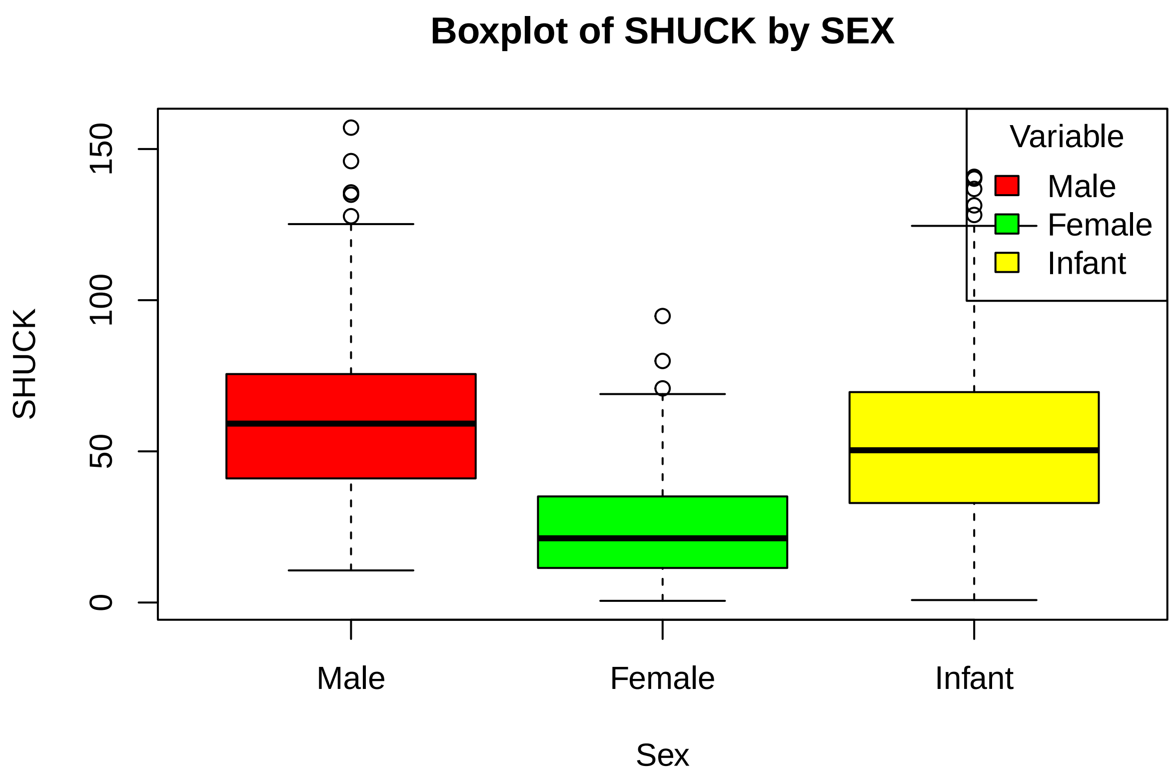
```r
boxplot(mydata$SHUCK,
     col  = "red",
     main = "Boxplot of SHUCK",
     xlab = "SHUCK",
     ylab = "COUNT",
     ylim =  c(0,y_max))

boxplot(mydata$WHOLE,
     col  = "green",
     main = "Boxplot of WHOLE",
     xlab = "WHOLE",
     ylab = "COUNT",
     ylim =  c(0,y_max))

par(mfrow=c(1,1))
legend("topright", legend = c("SHUCK", "WHOLE"), fill = c("red", "green"), title = "Variable")
```



(2 points) Now, plot parallel vertical boxplots of SHUCK by SEX. You should have three boxplots in one graph. Change the colors, titles, legends etc...

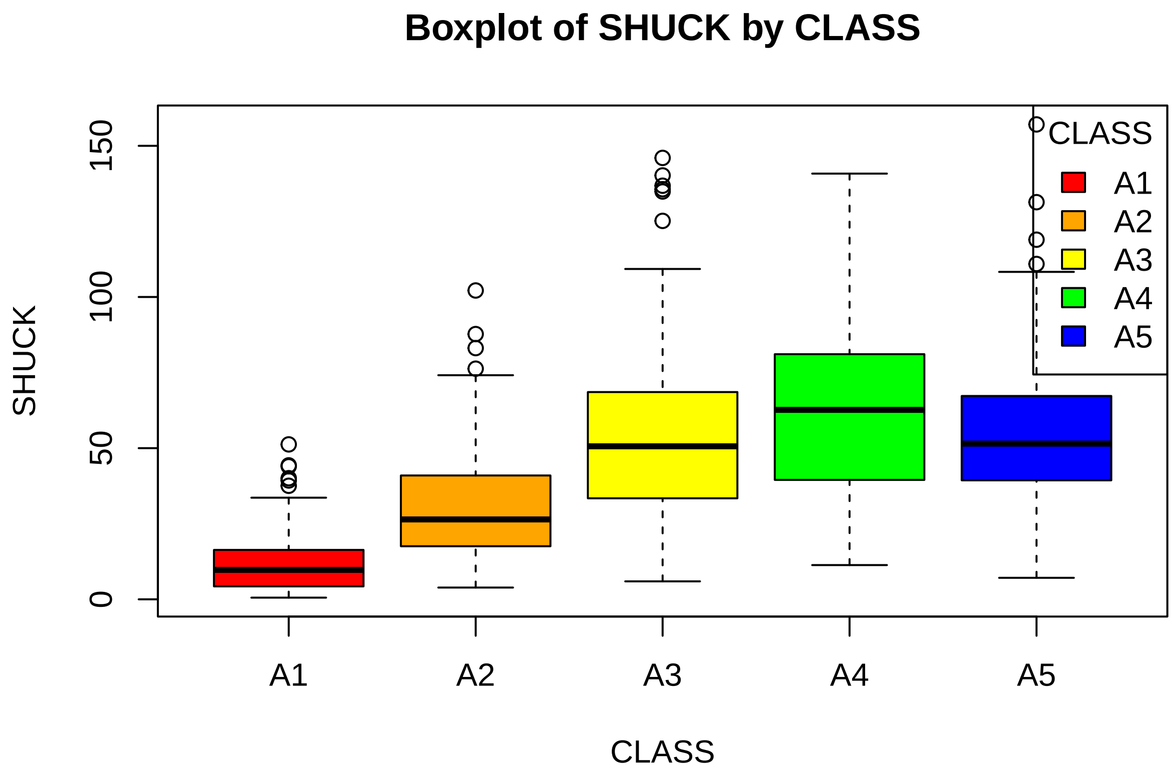
```r
boxplot(SHUCK ~ SEX, data = mydata,
        col = c("red", "green", "yellow"),
        main = "Boxplot of SHUCK by SEX",
        xlab = "Sex",
        ylab = "SHUCK",
        names = c("Male", "Female", "Infant"))
```

```r
legend("topright", legend = c("Male", "Female", "Infant"), fill = c("red", "green", "yellow "), title =
```



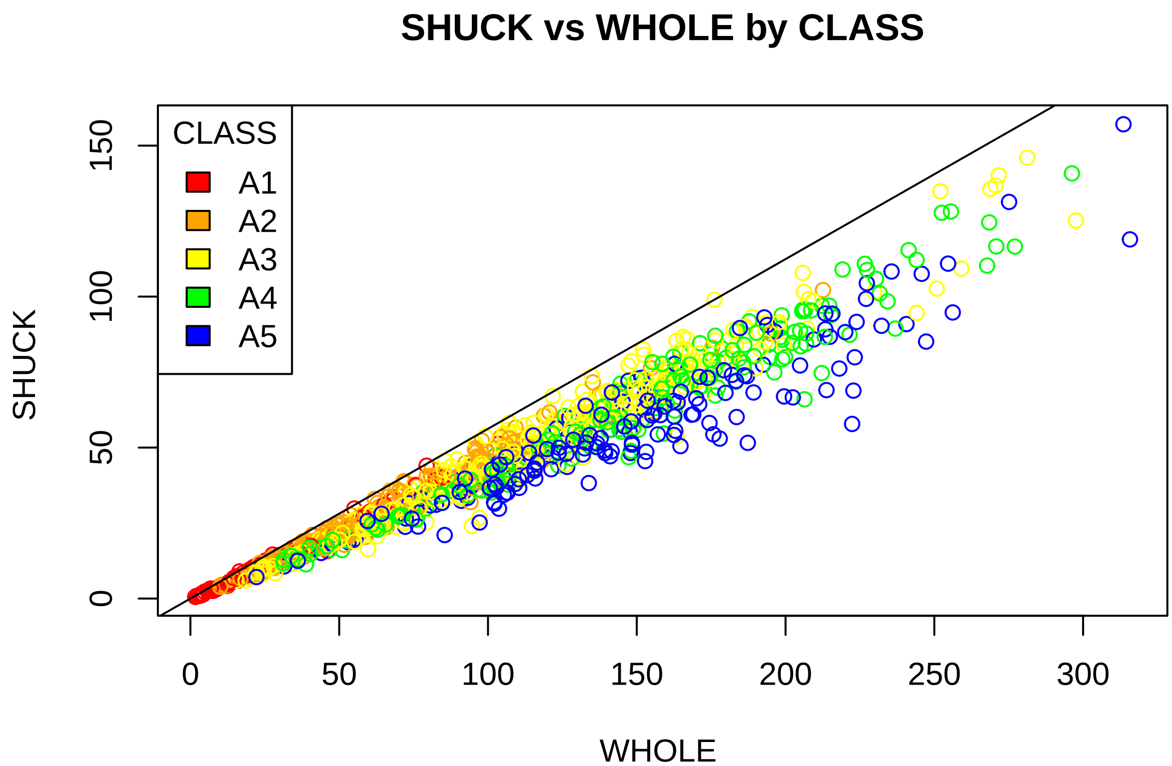
(2 points) And finally plot parallel vertical boxplots of SHUCK by CLASS. You should have three boxplots in one graph. Change the colors, titles, legends etc...

```r
colors <- c("red", "orange", "yellow", "green", "blue")
boxplot(SHUCK ~ CLASS, data = mydata,
        col = colors,
        main = "Boxplot of SHUCK by CLASS",
        xlab = "CLASS",
        ylab = "SHUCK",
        names = c("A1", "A2", "A3", "A4", "A5"))

legend("topright", legend = c("A1", "A2", "A3", "A4", "A5"), fill = colors, title = "CLASS")
```



(4 points) Now, let's compare SHUCK and WHOLE. Use "mydata" to plot SHUCK versus WHOLE with WHOLE on the horizontal axis. Color code data points by CLASS. As an aid to interpretation, determine the maximum value of the ratio of SHUCK to WHOLE. Add to the chart a straight line with zero intercept using this maximum value as the slope of the line. If you are using the 'base R' *plot()* function, you may use *abline()* to add this line to the plot. Use *help(abline)* in R to determine the coding for the slope and intercept arguments in the functions. If you are using ggplot2 for visualizations, *geom_abline()* should be used.

```
colors <- c("red", "orange", "yellow", "green", "blue")
max_ratio <- max(mydata$SHUCK / mydata$WHOLE, na.rm = TRUE)
fclass <- as.factor (mydata$CLASS)
plot(mydata$WHOLE, mydata$SHUCK,
     col = colors[as.numeric(fclass)],
     xlab = "WHOLE ",
     ylab = "SHUCK ",
     main = "SHUCK vs WHOLE by CLASS")

abline(a = 0, b = max_ratio, col = "black",)
legend("topleft", legend = c("A1", "A2", "A3", "A4", "A5"), fill = colors, title = "CLASS")
```



Essay Question (2 points): By now, you have enough data to review the relation of SHUCK and WHOLE and the dispersion of data. How does the variability in the last plot differ from the plot in (2a)? Compare the two displays. Keep in mind that SHUCK is a part of WHOLE. Consider the location of the different age classes.

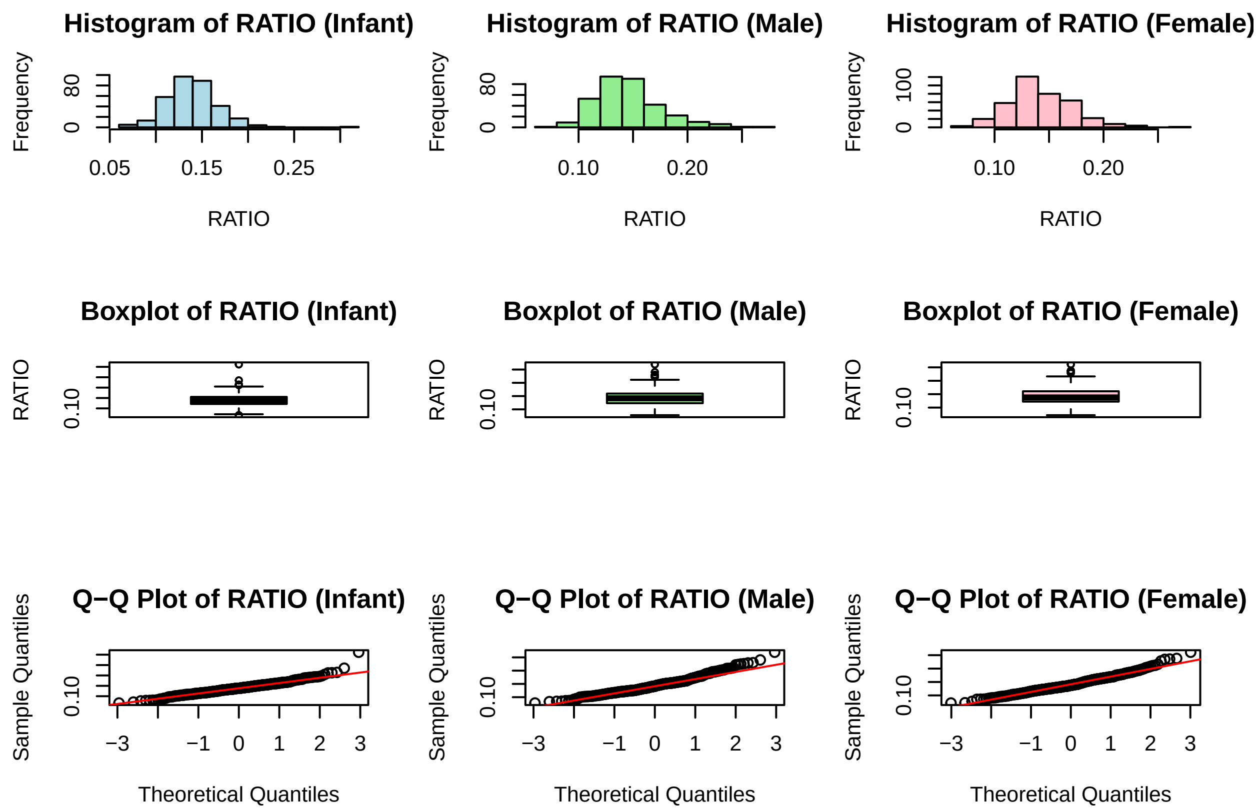
Answer :The bar plot emphasizes the variability in population based on sex within different classes, highlighting the total counts of each sex group rather than focusing on any continuous weight measure. In contrast, the scatter plot is more focused on individual variability, suggesting that as abalones grow larger, the variation in the proportion of weight represented by the shucked portion increases. The bar plot shows differences in the number of abalones of each sex per class, while the scatter plot highlights the increasing variation in shucked weight with increasing abalone size.

Section 5: (15 points) Getting insights about the data using graphs.

(6 points) Use "mydata" to create a multi-figured plot with histograms, boxplots and Q-Q plots of RATIO differentiated by sex. This can be done using *par(mfrow = c(3,3))* and base R or *grid.arrange()* and ggplot2. The first row would show the histograms, the second row the boxplots and the third row the Q-Q plots. Be sure these displays are legible.

```
# Create multi-figured plot with histograms, boxplots, and Q-Q plots of RATIO differentiated by SEX
par(mfrow = c(3, 3))
sex_levels <- levels(mydata$SEX)
colors <- c("lightblue", "lightgreen", "pink")

# First row: Histograms of RATIO by SEX
for (i in 1:length(sex_levels)) {
  ratio_data <- as.numeric(mydata$RATIO[mydata$SEX == sex_levels[i]])
```

```r
  if (length(ratio_data) > 0) {
    hist(ratio_data, breaks = 10, col = colors[i],
         main = paste("Histogram of RATIO (", sex_levels[i], ")", sep = ""),
         xlab = "RATIO", ylab = "Frequency")
  }
}

# Second row: Boxplots of RATIO by SEX
for (i in 1:length(sex_levels)) {
  ratio_data <- as.numeric(mydata$RATIO[mydata$SEX == sex_levels[i]])
  if (length(ratio_data) > 0) {
    boxplot(ratio_data, col = colors[i],
            main = paste("Boxplot of RATIO (", sex_levels[i], ")", sep = ""),
            ylab = "RATIO")
  }
}

# Third row: Q-Q plots of RATIO by SEX
for (i in 1:length(sex_levels)) {
  ratio_data <- as.numeric(mydata$RATIO[mydata$SEX == sex_levels[i]])
  if (length(ratio_data) > 0) {
    qqnorm(ratio_data, main = paste("Q-Q Plot of RATIO (", sex_levels[i], ")", sep = ""))
    qqline(ratio_data, col = "red")
  }
}
```



```r
# Reset plotting layout
par(mfrow = c(1, 1))
```

Essay Question (3 points): Compare the displays. How do the distributions compare to normality? Take into account the criteria discussed in the sync sessions to evaluate non-normality.

Answer: (I chose the Q-Q plot to assess normality. The male distribution is nearly normal, showing only slight deviations from symmetry and lighter tails. The female distribution, on the other hand, has higher kurtosis and exhibits left skewness. The infant distribution displays a strong positive skew along with excess kurtosis. Overall, I can conclude that males and females exhibit more normality compared to infants.)

(2 points) The boxplots in (3)(a) indicate that there are outlying RATIOs for each sex. *boxplot.stats()* can be used to identify outlying values of a vector.

```r
male_data <- mydata[mydata$SEX == "Male", "RATIO"]
female_data <- mydata[mydata$SEX == "Female", "RATIO"]
infant_data <- mydata[mydata$SEX == "Infant", "RATIO"]
male_outliers <- boxplot.stats(male_data)$out
female_outliers <- boxplot.stats(female_data)$out
infant_outliers <- boxplot.stats(infant_data)$out
```

(6 points) Present the abalones with these outlying RATIO values along with their associated variables in "mydata". Display the observations by passing a data frame to the kable() function. Basically, we want to output those rows of "mydata" with an outlying RATIO, but we want to determine outliers looking separately at infants, females and males.

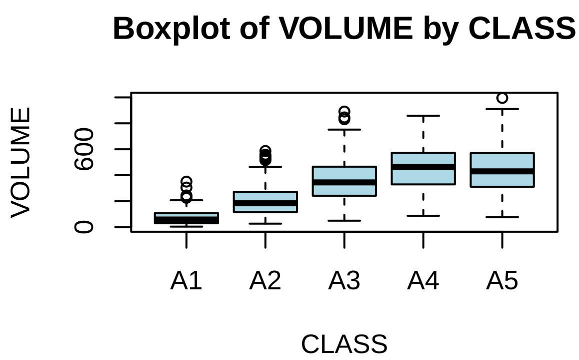
```r
male_data <- subset(mydata, SEX == "Male")
female_data <- subset(mydata, SEX == "Female")
infant_data <- subset(mydata, SEX == "Infant")
male_outliers <- boxplot.stats(male_data$RATIO)$out
female_outliers <- boxplot.stats(female_data$RATIO)$out
infant_outliers <- boxplot.stats(infant_data$RATIO)$out
male_outliers_data <- male_data[male_data$RATIO %in% male_outliers, ]
female_outliers_data <- female_data[female_data$RATIO %in% female_outliers, ]
infant_outliers_data <- infant_data[infant_data$RATIO %in% infant_outliers, ]
outliers_data <- rbind(male_outliers_data, female_outliers_data, infant_outliers_data)
kable(outliers_data)
```

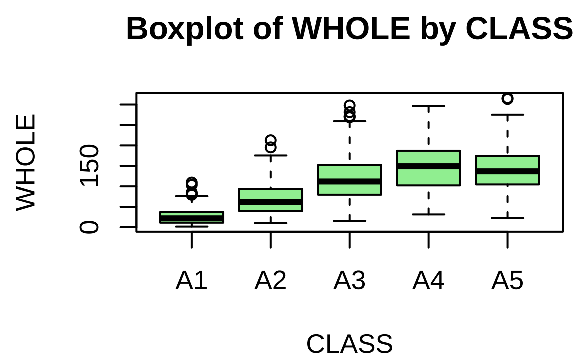
	SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
3	Male	10.080	7.350	2.205	79.37500	44.00000	6	A1	163.364040	0.2693371
37	Male	4.305	3.255	0.945	6.18750	2.93750	3	A1	13.242072	0.2218308
42	Male	2.835	2.730	0.840	3.62500	1.56250	4	A1	6.501222	0.2403394
58	Male	6.720	4.305	1.680	22.62500	11.00000	5	A1	48.601728	0.2263294
67	Male	5.040	3.675	0.945	9.65625	3.93750	5	A1	17.503290	0.2249577
89	Male	3.360	2.310	0.525	2.43750	0.93750	4	A1	4.074840	0.2300704
105	Male	6.930	4.725	1.575	23.37500	11.81250	7	A2	51.572194	0.2290478
200	Male	9.135	6.300	2.520	74.56250	32.37500	8	A2	145.027260	0.2232339
746	Female	13.440	10.815	1.680	130.25000	63.73125	10	A3	244.194048	0.2609861
754	Female	10.500	7.770	3.150	132.68750	61.13250	9	A3	256.992750	0.2378764
803	Female	10.710	8.610	3.255	160.31250	70.41375	9	A3	300.153640	0.2345924

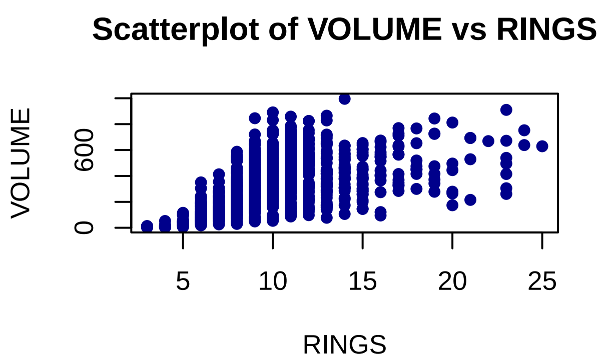
	SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
810	Female	12.285	9.870	3.465	176.12500	99.00000	10	A3	420.141472	0.2356349
852	Female	11.550	8.820	3.360	167.56250	78.27187	10	A3	342.286560	0.2286735
350	Infant	7.980	6.720	2.415	80.93750	40.37500	7	A2	129.505824	0.3117620
379	Infant	15.330	11.970	3.465	252.06250	134.89812	10	A3	635.827846	0.2121614
420	Infant	11.550	7.980	3.465	150.62500	68.55375	10	A3	319.365585	0.2146560
421	Infant	13.125	10.290	2.310	142.00000	66.47062	9	A3	311.979938	0.2130606
458	Infant	11.445	8.085	3.150	139.81250	68.49062	9	A3	291.478399	0.2349767
586	Infant	12.180	9.450	4.935	133.87500	38.25000	14	A5	568.023435	0.0673388

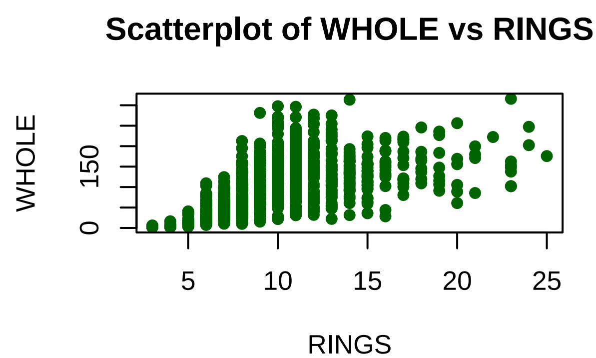
Part 4 ***### Section 4: (8 points) Getting insights about possible predictors.*** (4)(a) (3 points) With "mydata," display side-by-side boxplots for VOLUME and WHOLE, each differentiated by CLASS There should be five boxes for VOLUME and five for WHOLE. Also, display side-by-side scatterplots: ?VOLUME and WHOLE versus RINGS. Present these four figures in one graphic: ?the boxplots in one row and the scatterplots in a second row. Base R or ggplot2 may be used.

```r
par(mfrow = c(2, 2)) # 2 rows and 2 columns for side-by-side plots

# Boxplots for VOLUME and WHOLE by CLASS
boxplot(VOLUME ~ CLASS, data = mydata, main = "Boxplot of VOLUME by CLASS",
        xlab = "CLASS", ylab = "VOLUME", col = "lightblue")
boxplot(WHOLE ~ CLASS, data = mydata, main = "Boxplot of WHOLE by CLASS",
        xlab = "CLASS", ylab = "WHOLE", col = "lightgreen")

# Scatterplots for VOLUME and WHOLE versus RINGS
plot(mydata$RINGS, mydata$VOLUME, main = "Scatterplot of VOLUME vs RINGS",
     xlab = "RINGS", ylab = "VOLUME", col = "darkblue", pch = 16)
plot(mydata$RINGS, mydata$WHOLE, main = "Scatterplot of WHOLE vs RINGS",
     xlab = "RINGS", ylab = "WHOLE", col = "darkgreen", pch = 16)
```









Essay Question (5 points) How well do you think these variables would perform as predictors of age? ?Explain. ***Answer: (Based on the plots, volume (VOLUME) and whole weight (WHOLE) show some potential as predictors of age, as seen in the scatterplots with RINGS, which is an indicator of age. Both volume and whole weight appear to have a positive correlation with RINGS, suggesting that as these values increase, age also tends to increase. However, there is considerable overlap and variability in the data points, indicating that the relationship is not strictly linear or highly consistent. Therefore, while these variables could provide some insight into predicting age, their predictive accuracy might be limited due to noise and variability.)*** —– PART 5 ***### Section 5: (12 points) Getting insights regarding different groups in the data.*** (5)(a) (2 points) Use *aggregate()* with "mydata" to compute the mean values of VOLUME, SHUCK and RATIO for each combination of SEX and CLASS. Then, using *matrix()*, create matrices of the mean values. Using the "dimnames" argument within *matrix()* or the *rownames()* and *colnames()* functions on the matrices, label the rows by SEX and columns by CLASS. Present the three matrices (Kabacoff Section 5.6.2, p. 110-111). The *kable()* function is useful for this purpose. ?You do not need to be concerned with the number of digits presented.

```
mean_values <- aggregate(cbind(VOLUME, SHUCK, RATIO) ~ SEX + CLASS, data = mydata, mean)

# Create matrices of the mean values using *matrix()*
volume_matrix <- matrix(mean_values$VOLUME, nrow = length(unique(mydata$SEX)), byrow = TRUE,
                        dimnames = list(unique(mydata$SEX), unique(mydata$CLASS)))
shuck_matrix <- matrix(mean_values$SHUCK, nrow = length(unique(mydata$SEX)), byrow = TRUE,
                       dimnames = list(unique(mydata$SEX), unique(mydata$CLASS)))
ratio_matrix <- matrix(mean_values$RATIO, nrow = length(unique(mydata$SEX)), byrow = TRUE,
                       dimnames = list(unique(mydata$SEX), unique(mydata$CLASS)))

# Display the matrices using *kable()*
```

```r
library(knitr)
kable(volume_matrix, caption = "Mean VOLUME by SEX and CLASS")
```

Table 2: Mean VOLUME by SEX and CLASS

	A1	A2	A5	A3	A4
Male	255.2994	66.51618	103.7232	276.8573	160.3200
Infant	245.3857	412.60794	270.7406	358.1181	498.0489
Female	316.4129	442.61552	486.1525	318.6930	440.2074

```r
kable(shuck_matrix, caption = "Mean SHUCK by SEX and CLASS")
```

Table 3: Mean SHUCK by SEX and CLASS

	A1	A2	A5	A3	A4
Male	38.90000	10.11332	16.39583	42.50305	23.41024
Infant	38.33855	59.69121	37.17969	52.96933	69.05161
Female	39.85369	61.42726	59.17076	36.47047	55.02762

```r
kable(ratio_matrix, caption = "Mean RATIO by SEX and CLASS")
```

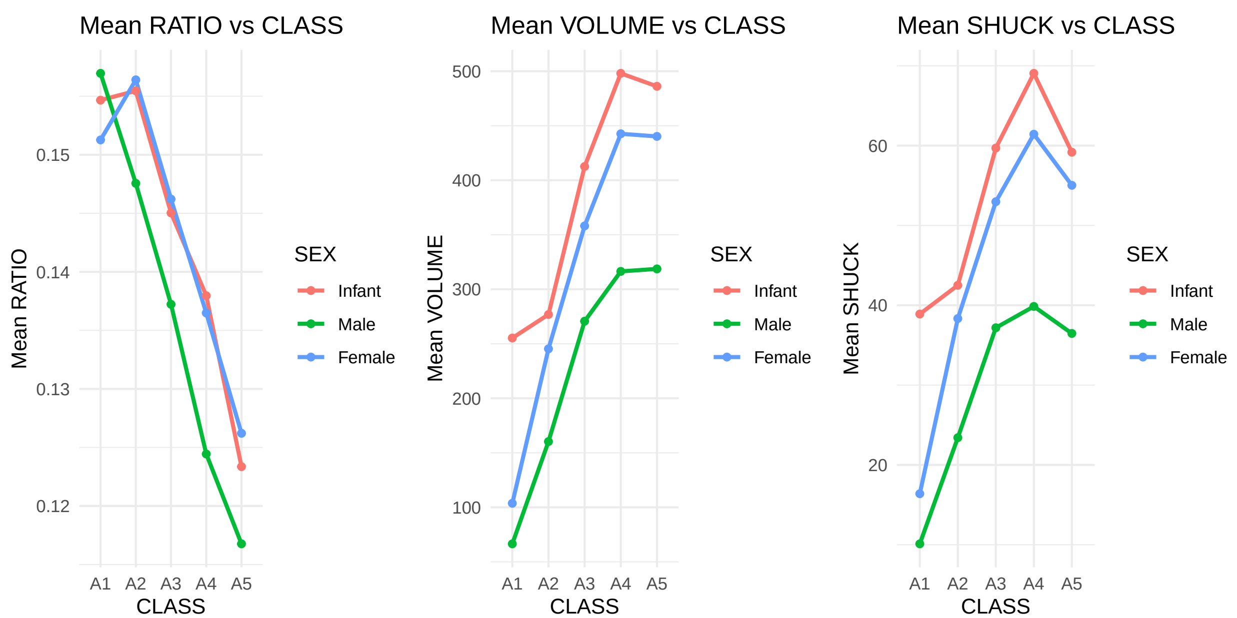
Table 4: Mean RATIO by SEX and CLASS

	A1	A2	A5	A3	A4
Male	0.1546644	0.1569554	0.1512698	0.1554605	0.1475600
Infant	0.1564017	0.1450304	0.1372256	0.1462123	0.1379609
Female	0.1244413	0.1364881	0.1233605	0.1167649	0.1262089

(5)(b) (3 points) Present three graphs. Each graph should include three lines, one for each sex. The first should show mean RATIO versus CLASS; the second, mean VOLUME versus CLASS; the third, mean SHUCK versus CLASS. This may be done with the 'base R' *interaction.plot()* function or with ggplot2 using *grid.arrange()*.

```r
library(gridExtra)

# Plot mean RATIO vs CLASS
g1 <- ggplot(mean_values, aes(x = CLASS, y = RATIO, color = SEX, group = SEX)) +
  geom_line(linewidth = 1) +
  geom_point() +
  labs(title = "Mean RATIO vs CLASS", x = "CLASS", y = "Mean RATIO") +
  theme_minimal()

# Plot mean VOLUME vs CLASS
g2 <- ggplot(mean_values, aes(x = CLASS, y = VOLUME, color = SEX, group = SEX)) +
  geom_line(linewidth = 1) +
  geom_point() +
  labs(title = "Mean VOLUME vs CLASS", x = "CLASS", y = "Mean VOLUME") +
```

```r
  theme_minimal()

# Plot mean SHUCK vs CLASS
g3 <- ggplot(mean_values, aes(x = CLASS, y = SHUCK, color = SEX, group = SEX)) +
  geom_line(linewidth = 1) +
  geom_point() +
  labs(title = "Mean SHUCK vs CLASS", x = "CLASS", y = "Mean SHUCK") +
  theme_minimal()

# Arrange the three plots in a grid
grid.arrange(g1, g2, g3, nrow = 1)
```

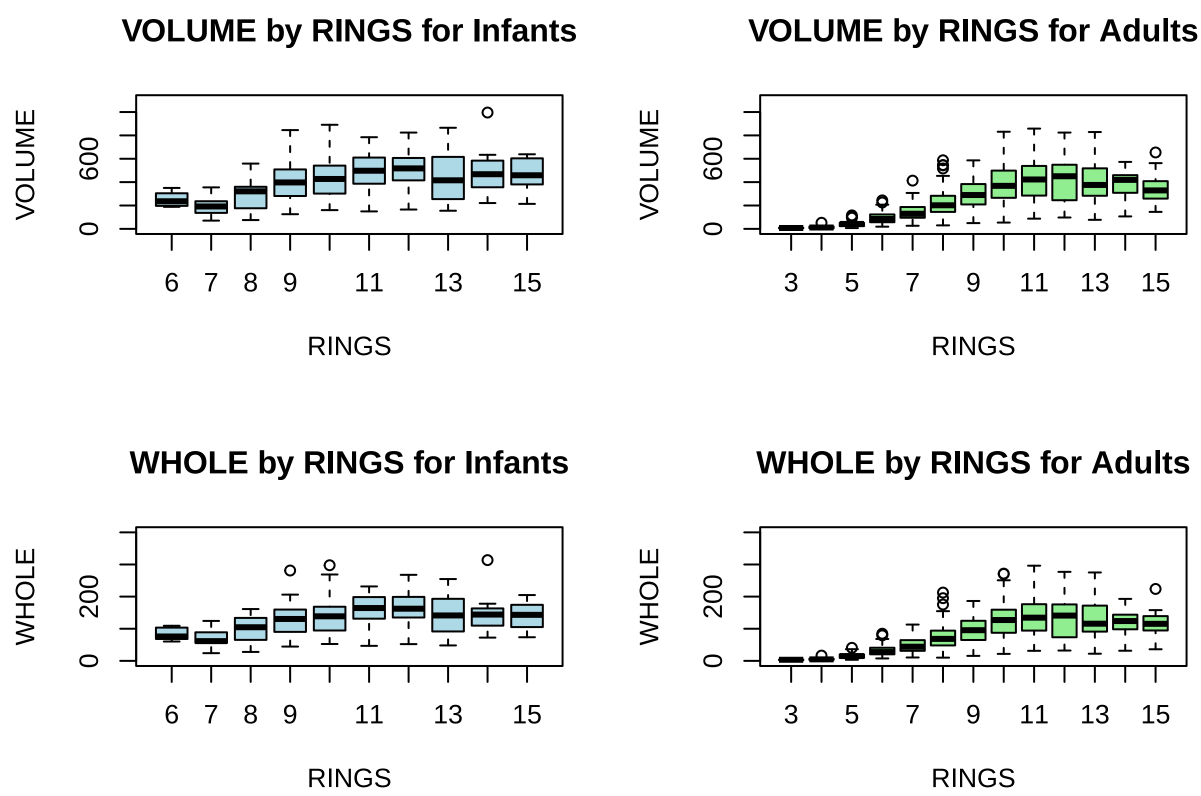


Essay Question (2 points): ?What questions do these plots raise? ?Consider aging and sex differences. ***Answer: (These plots raise questions regarding the relationship between aging and sex differences across different classes. Specifically, why do infants consistently have higher mean volume and shuck weight compared to males and females in certain classes, and why does this pattern change as we move to higher classes? Additionally, there is a notable decline in mean ratio across all sexes with increasing class; it prompts the question of whether age or class is negatively correlated with ratio across different sexes and what biological factors may account for these variations. These observations suggest possible age and sex-specific growth or development patterns that require further investigation.)***

5(c) (3 points) Present four boxplots using *par(mfrow = c(2, 2)* or *grid.arrange()*. The first line should show VOLUME by RINGS for the infants and, separately, for the adult; factor levels "M" and "F," combined. The second line should show WHOLE by RINGS for the infants and, separately, for the adults. Since the data are sparse beyond 15 rings, limit the displays to less than 16 rings. One way to accomplish this is to generate a new data set using subset() to select RINGS $<$ 16. ?Use ylim $=$ c(0, 1100) for VOLUME and ylim $=$ c(0, 400) for WHOLE. ?If you wish to reorder the displays for presentation purposes or use ggplot2 go ahead.

```r
subset_data <- subset(mydata, RINGS < 16)

# Create boxplots for VOLUME and WHOLE by RINGS for infants and adults
par(mfrow = c(2, 2))
```

```r
# VOLUME by RINGS for infants (SEX = "Infant")
boxplot(VOLUME ~ RINGS, data = subset(subset_data, SEX == "Infant"),
        main = "VOLUME by RINGS for Infants", xlab = "RINGS", ylab = "VOLUME",
        ylim = c(0, 1100), col = "lightblue")

# VOLUME by RINGS for adults (SEX = "M" or "F")
boxplot(VOLUME ~ RINGS, data = subset(subset_data, SEX %in% c("Male", "Female")),
        main = "VOLUME by RINGS for Adults", xlab = "RINGS", ylab = "VOLUME",
        ylim = c(0, 1100), col = "lightgreen")

# WHOLE by RINGS for infants (SEX = "I")
boxplot(WHOLE ~ RINGS, data = subset(subset_data, SEX == "Infant"),
        main = "WHOLE by RINGS for Infants", xlab = "RINGS", ylab = "WHOLE",
        ylim = c(0, 400), col = "lightblue")

# WHOLE by RINGS for adults (SEX = "M" or "F")
boxplot(WHOLE ~ RINGS, data = subset(subset_data, SEX %in% c("Male", "Female")),
        main = "WHOLE by RINGS for Adults", xlab = "RINGS", ylab = "WHOLE",
        ylim = c(0, 400), col = "lightgreen")
```



Essay Question (2 points): ?What do these displays suggest about abalone growth? ?Also, compare the infant and adult displays. ?What differences stand out?

Answer: (I think most of the abalone's growth seems to occur during the first half of its life span, specifically from 1 ring to 10-11 rings, where both volume and weight increase rapidly. In contrast, after reaching 10-11 rings, growth appears to either level off or even decrease. Interestingly, while it is logical that infants have lower volumes and weights compared to adults, there is a surprising amount of overlap between them. For instance, an abalone with a volume of 500 cm^3 or a weight of 150g is likely an adult, but anything below this threshold could just as easily be either an infant or an adult. Many abalones classified as "infants" share the same size and weight as some smaller or lighter adults. Additionally, infants tend to exhibit less variability in volume and weight, staying closer to each other in these measures. Except for the group of highly variable 12-ring infants, infants generally have smaller interquartile ranges (IQRs) and standard deviations compared to adults, who display more variability in volume and weight within each ring grouping.)

Section 6: (11 points) Conclusions from the Exploratory Data Analysis (EDA).

Conclusions

Essay Question 1) (5 points) ? ?Based solely on these data, what are plausible statistical reasons that explain the failure of the original study? Consider to what extent physical measurements may be used for age prediction. ***Answer: (The likely reason for the original study's lack of success lies in the failure to account for the misclassification of infants. Since immature abalone are challenging to correctly determine in terms of sex, distinguishing between an immature abalone and a small adult abalone can be problematic. This sex information could be critical when attempting to predict age using physical measurements. Furthermore, abalone growth in terms of size and weight appears to taper off once they reach 10-11 rings, meaning that weight and volume no longer increase proportionally with additional rings beyond that point. For young abalone, lower weight and volume correlate more strongly with fewer rings, and the sex of the abalone plays a significant role in their size and weight—females tend to be heavier and larger compared to males and infants, as shown in the 5b displays. However, in the later years, as depicted in the 5c displays, the size and weight of abalone no longer increase proportionally with the number of rings. This is compounded by the already weak correlation between volume and rings, as well as weight and rings, shown in the 4b displays. Therefore, it is nearly impossible to accurately determine the age of an A3-A5 abalone based solely on physical measurements. In conclusion, inaccurate sex classifications make it challenging to predict the age of young abalone from physical measurements, while the slowing growth in weight and volume in A3-A5 abalone complicates the age prediction for older individuals.)***

Essay Question 2) (3 points) Do not refer to the abalone data or study. ?If you were presented with an overall histogram and summary statistics from a sample of some population or phenomenon and no other information, what questions might you ask before accepting them as representative of the sampled population or phenomenon? ***Answer: (I would like to learn more about the sample and population. For instance, what is the sample size, and how does it relate to the population size? Are any population parameters known, or is it feasible to estimate them? Which sampling methods were employed? Was it a simple random sample, or was a non-random approach used? How was the sampling frame defined? When was the sample collected, and is the data current or outdated? Are there any known biases or influential factors that could impact the representativeness of the sample compared to the population? These factors are crucial in determining whether the sample accurately represents the target population.)***

Essay Question 3) ?(3 points) ? ?Do not refer to the abalone data or study. ?What do you see as difficulties analyzing data derived from observational studies? Can causality be determined?

?What might be learned from such studies? ***Answer: (I feel like analyzing data from observational studies is challenging due to the numerous factors involved. Observational studies are especially prone to human error and bias, which can arise during the study's construction, data collection, or result analysis, significantly impacting the findings. In the absence of a control group, the strongest outcome achievable is establishing correlation, not causation. It is easy for such studies to overlook important explanatory variables that might influence the response variable. While determining causality is not possible through an observational study, these studies can help identify variables of interest that may be further investigated to establish causal relationships.)***