Data Analysis Assignment #1 (50 points total)

Desilets, Steve

Submit both the .Rmd and .html files for grading. You may remove the instructions and example problem above, but do not remove the YAML metadata block or the first, “setup” code chunk. Address the steps that appear below and answer all the questions. Be sure to address each question with code and comments as needed. You may use either base R functions or ggplot2 for the visualizations.

The following code chunk will:

a. load the “ggplot2”, “gridExtra” and “knitr” packages, assuming each has been installed on your machine,
b. read-in the abalones dataset, defining a new data frame, “mydata,”
c. return the structure of that data frame, and
d. calculate new variables, VOLUME and RATIO.

Do not include package installation code in this document. Packages should be installed via the Console or ‘Packages’ tab. You will also need to download the abalones.csv from the course site to a known location on your machine. Unless a *file.path()* is specified, R will look to directory where this .Rmd is stored when knitting.

```
## 'data.frame':    1036 obs. of  8 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 ...
```

Test Items starts from here - There are 6 sections - Total 50 points

Section 1: (6 points) Summarizing the data.

(1)(a) (1 point) Use *summary()* to obtain and present descriptive statistics from mydata. Use *table()* to present a frequency table using CLASS and RINGS. There should be 115 cells in the table you present.

```
##  SEX         LENGTH           DIAM            HEIGHT          WHOLE       
##  F:326   Min.   : 2.73   Min.   : 1.995   Min.   :0.525   Min.   :  1.625  
##  I:329   1st Qu.: 9.45   1st Qu.: 7.350   1st Qu.:2.415   1st Qu.: 56.484  
##  M:381   Median :11.45   Median : 8.925   Median :2.940   Median :101.344  
##          Mean   :11.08   Mean   : 8.622   Mean   :2.947   Mean   :105.832  
##          3rd Qu.:13.02   3rd Qu.:10.185   3rd Qu.:3.570   3rd Qu.:150.319  
##          Max.   :16.80   Max.   :13.230   Max.   :4.935   Max.   :315.750  
##      SHUCK             RINGS        CLASS        VOLUME       
##  Min.   :  0.5625   Min.   : 3.000   A1:108   Min.   :  3.612  
##  1st Qu.: 23.3006   1st Qu.: 8.000   A2:236   1st Qu.:163.545  
##  Median : 42.5700   Median : 9.000   A3:329   Median :307.363  
##  Mean   : 45.4396   Mean   : 9.993   A4:188   Mean   :326.804  
##  3rd Qu.: 64.2897   3rd Qu.:11.000   A5:175   3rd Qu.:463.264  
##  Max.   :157.0800   Max.   :25.000            Max.   :995.673  
##      RATIO       
##  Min.   :0.06734  
##  1st Qu.:0.12241  
##  Median :0.13914  
##  Mean   :0.14205  
##  3rd Qu.:0.15911  
##  Max.   :0.31176
```

```
##      rings
## class   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
##      rings
## class  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
```

Question (1 point): Briefly discuss the variable types and distributional implications such as potential skewness and outliers.

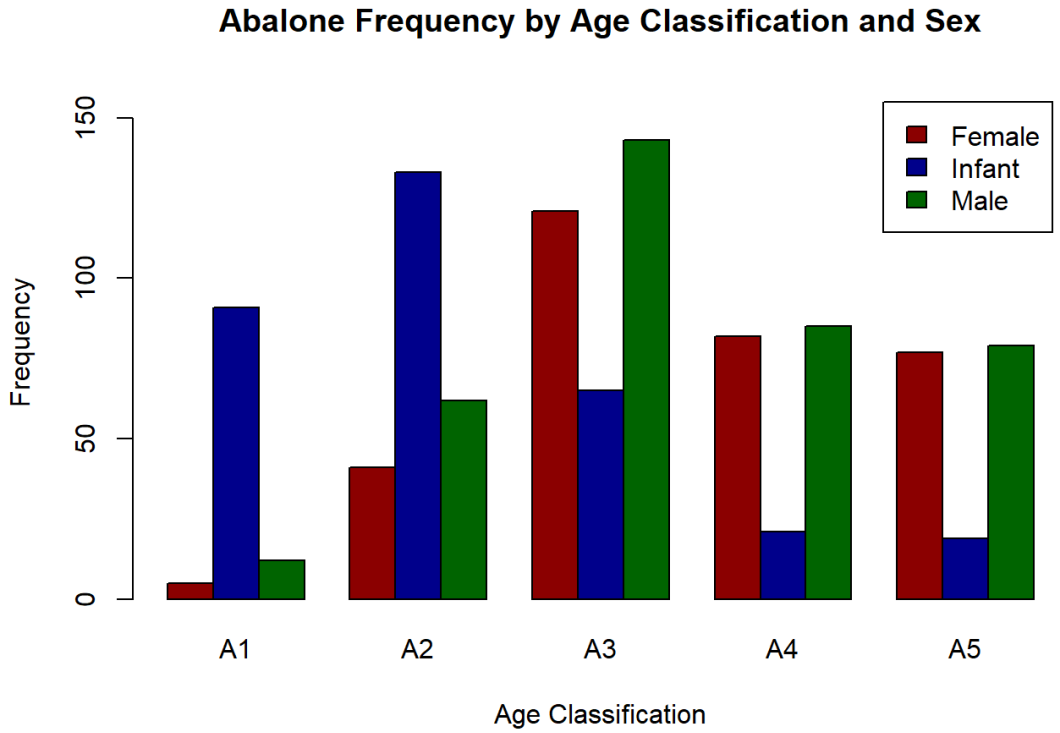
Answer: (Enter your answer here.)

This dataset contains 10 variables. One of these variables - sex - is a nominal variable. One of these variables - class - is an ordinal variable. The other eight variables (length, diameter, height, whole weight, shucked weight, rings, volume, and density ratio) are all ratio variables. Notably, the distribution of ages (as determined by rings) appeared to be skewed right. That is because, visually it seems that there may be several abalones with a high number of rings that are outliers (in terms of the rings variable). Based on the summary statistics, I would guess that other variables in this dataset will also prove to have outliers. (For me, the most immediately apparent outlier is the abalone with a shucked weight of 157 grams, which is well above the outlier threshold given that the 3rd quartile for shucked weights is 64 grams and the IQR is 41 grams, which would make the outlier threshold roughly 126 grams). Also, notably, the table displaying class and rings is a nice data visualization since it explains clearly how the researchers transformed the rings data into five age classifications.

(1)(b) (1 point) Generate a table of counts using SEX and CLASS. 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.

```
## Margins computed over dimensions
## in the following order:
## 1: sex
## 2: class
```

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



Essay Question (2 points): Discuss the sex distribution of abalones. What stands out about the distribution of abalones by CLASS?

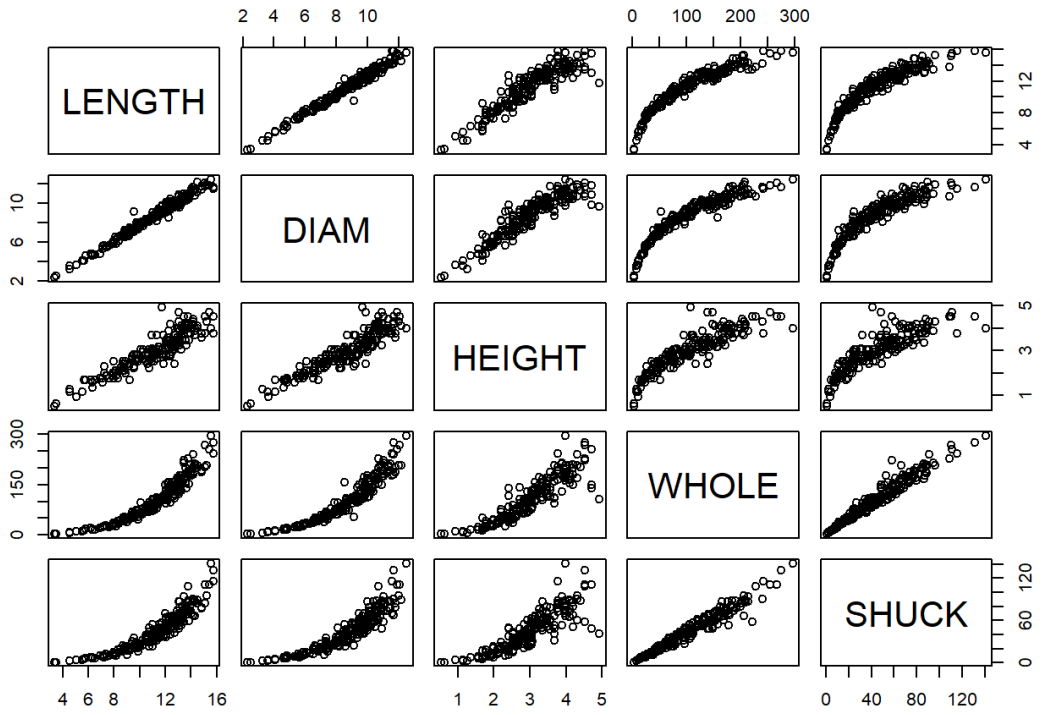
Answer: (Enter your answer here.)

The sex distribution of abalones in this sample is somewhat close to being evenly split between the three categories, with female, infant, and male abalones accounting for 31, 32, and 37 percent of the sample, respectively.

The finding that stands out is that infant abalones represent very high proportions of the first two age classes of abalones in our sample. In fact, as the age class increases, the percentage of the age class comprised of infants decreases. Specifically, infants account for 84%, 56%, 19%, 11%, and 11% of the A1, A2, A3, A4, and A5 age classes, respectively. This finding makes sense since infants should be expected to be relatively young. It is surprising that infants appear in some of the highest age classes in this chart, though, which raises some concerns for me regarding the quality of this data.

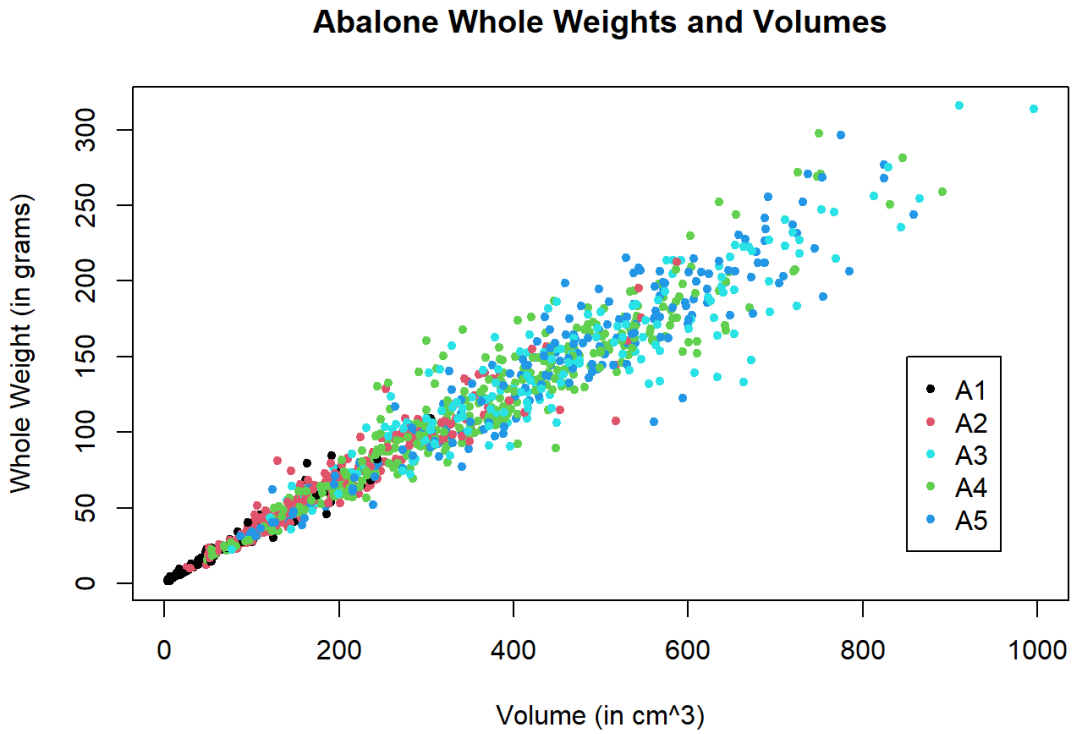
(1)(c) (1 point) Select a simple random sample of 200 observations from “mydata” and identify this sample as “work.” Use *set.seed(123)* prior to drawing this sample. Do not change the number 123. Note that *sample()* “takes a sample of the specified size from the elements of x.” We cannot sample directly from “mydata.” Instead, we need to sample from the integers, 1 to 1036, representing the rows of “mydata.” Then, select those rows from the data frame (Kabacoff Section 4.10.5 page 87).

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.



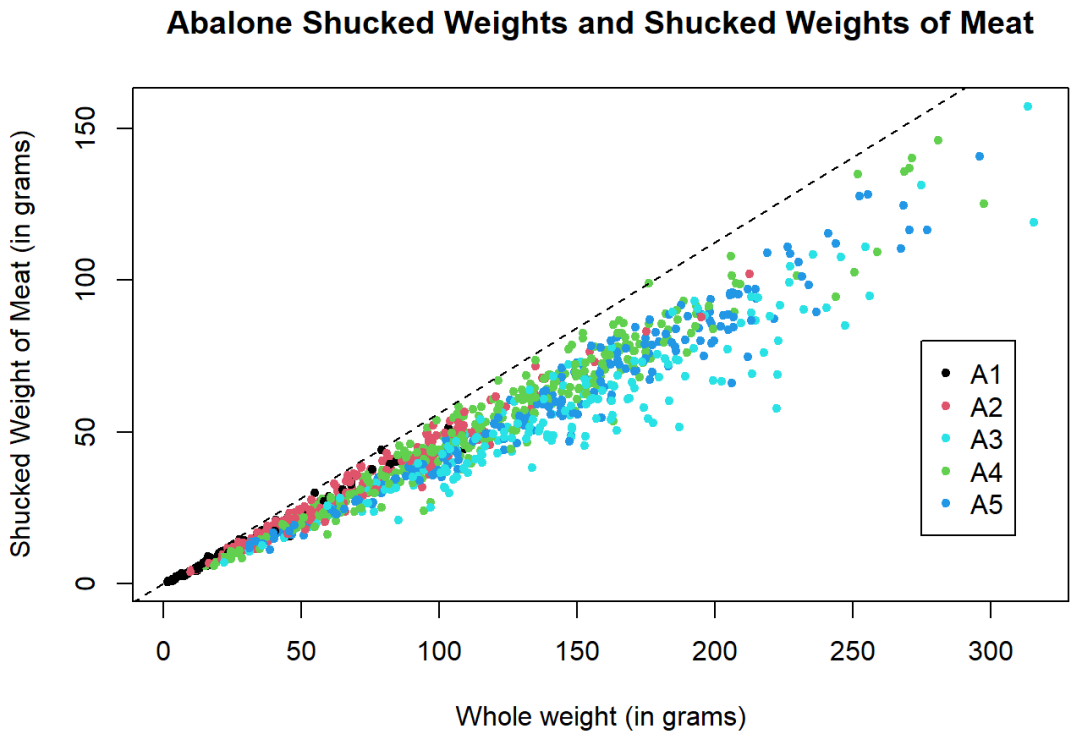
Section 2: (5 points) Summarizing the data using graphics.

(2)(a) (1 point) Use “mydata” to plot WHOLE versus VOLUME. Color code data points by CLASS.



(2)(b) (2 points) 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.



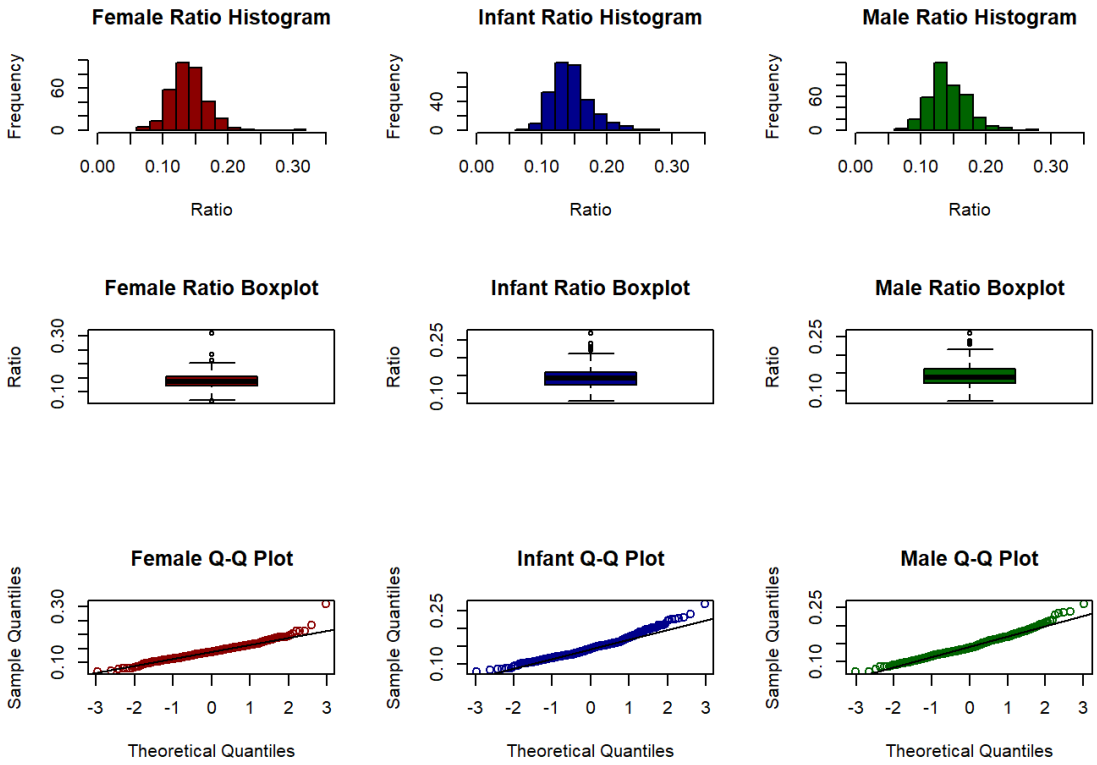
Essay Question (2 points): How does the variability in this plot differ from the plot in (a)? Compare the two displays. Keep in mind that SHUCK is a part of WHOLE. Consider the location of the different age classes.

Answer: (Enter your answer here.)

There exist several similarities between the two plots above. First, for the abalones in our dataset, whole weight appears to have a positive, relatively linear relationship with both shucked weight of meat and with volume. In fact, whole weight has a correlation of 0.97 with volume and with shucked weight. Second, there exists similar heteroskedasticity in both plots above. Specifically, if a linear regression model were to be constructed for these plots, the residual variances would be quite small for abalones with lower whole weights than the residual variances for abalones with higher whole weights. Third, similar age classification patterns appear in both of the plots of above. Specifically, abalones in the first age class tend to have some of the lowest whole weights, shucked weights, and volumes. As abalones move up in age class though, the distinction between abalones by whole weight, shucked weight, and volume becomes more negligible.

Section 3: (8 points) Getting insights about the data using graphs.

(3)(a) (2 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.



Essay Question (2 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: (Enter your answer here.)

For the abalones in our sample, the distributions of ratios (meaning the ratio of shucked weight in grams to volume in centimeters cubed) appear to be somewhat similar for male, female, and infant abalones. The median ratio for each of these 3 categories is approximately 0.14. Meanwhile, the IQR for female abalones ranges from 0.12 to 0.15, and the IQR for male and infant abalones each range from 0.12 to 0.16. In the histograms, all three distributions appear to be skewed right, and in the boxplots, all three distributions have several outliers in their right tails. Due to the right skew, it's likely that none of these samples are normally distributed. The nonlinearity of the points in the Q-Q Plots underscore the finding that these data are not quite normally distributed.

(3)(b) (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. 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.

	SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO	sex	shuck_to_whole
3	I	10.080	7.350	2.205	79.37500	44.00000	6	A1	163.364040	0.2693371	Infant	0.5543307
37	I	4.305	3.255	0.945	6.18750	2.93750	3	A1	13.242072	0.2218308	Infant	0.4747475
42	I	2.835	2.730	0.840	3.62500	1.56250	4	A1	6.501222	0.2403394	Infant	0.4310345
58	I	6.720	4.305	1.680	22.62500	11.00000	5	A1	48.601728	0.2263294	Infant	0.4861878
67	I	5.040	3.675	0.945	9.65625	3.93750	5	A1	17.503290	0.2249577	Infant	0.4077670
89	I	3.360	2.310	0.525	2.43750	0.93750	4	A1	4.074840	0.2300704	Infant	0.3846154
105	I	6.930	4.725	1.575	23.37500	11.81250	7	A2	51.572194	0.2290478	Infant	0.5053476
200	I	9.135	6.300	2.520	74.56250	32.37500	8	A2	145.027260	0.2232339	Infant	0.4341995
350	F	7.980	6.720	2.415	80.93750	40.37500	7	A2	129.505824	0.3117620	Female	0.4988417
379	F	15.330	11.970	3.465	252.06250	134.89812	10	A3	635.827846	0.2121614	Female	0.5351773

	SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO	sex	shuck_to_whole
420	F	11.550	7.980	3.465	150.62500	68.55375	10	A3	319.365585	0.2146560	Female	0.4551286
421	F	13.125	10.290	2.310	142.00000	66.47062	9	A3	311.979938	0.2130606	Female	0.4681030
458	F	11.445	8.085	3.150	139.81250	68.49062	9	A3	291.478399	0.2349767	Female	0.4898748
586	F	12.180	9.450	4.935	133.87500	38.25000	14	A5	568.023435	0.0673388	Female	0.2857143
746	M	13.440	10.815	1.680	130.25000	63.73125	10	A3	244.194048	0.2609861	Male	0.4892994
754	M	10.500	7.770	3.150	132.68750	61.13250	9	A3	256.992750	0.2378764	Male	0.4607254
803	M	10.710	8.610	3.255	160.31250	70.41375	9	A3	300.153640	0.2345924	Male	0.4392281
810	M	12.285	9.870	3.465	176.12500	99.00000	10	A3	420.141472	0.2356349	Male	0.5621008
852	M	11.550	8.820	3.360	167.56250	78.27187	10	A3	342.286560	0.2286735	Male	0.4671205

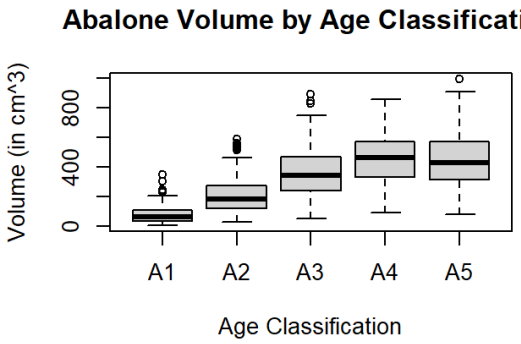
Essay Question (2 points): What are your observations regarding the results in (3)(b)?

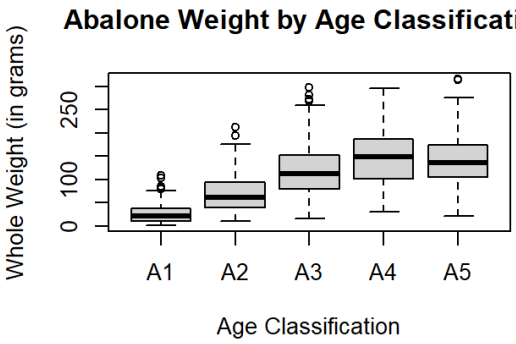
Answer: (Enter your answer here.)

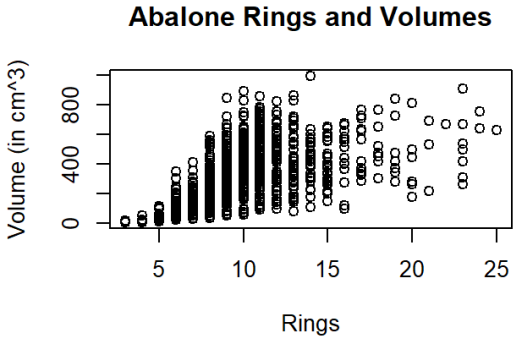
After disaggregating our abalone dataset by sex, there are 19 outliers for ratio (meaning the ratio of shucked weight in grams to volume in centimeters cubed). There is one outlier for the female distribution that falls in the left tail. The other 18 outliers all fall in the right tails of their respective distributions. The fact that 95% of these outliers are in the right tails reinforces our finding that these distributions appear to be skewed right. Also, notably, if these data were normally distributed, then we would expect approximately 0.70% of the data to be outliers (per the standard definition of outliers being 1.5 standard deviations below the first quartile or 1.5 standard deviations above the third quartile). By that definition, in a sample of 1,036 draws from a standard normal distribution, we would expect to find roughly 7 outliers. Since our samples of abalone ratios disaggregated by sex result in 19 outliers (not 7), it's likely that the abalone ratios are not quite normally distributed.

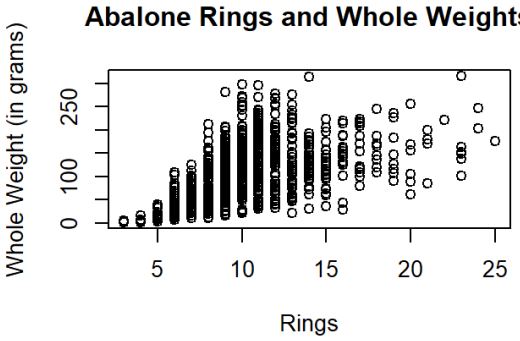
Section 4: (8 points) Getting insights about possible predictors.

(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.









Essay Question (5 points) How well do you think these variables would perform as predictors of age? Explain.

Answer: (Enter your answer here.)

The boxplots indicate that volume and whole weight might be decent (though not ideal) predictors of age classification for the first two age classes. For example, the IQR for the volumes of class A1 abalones ranges from roughly 31 to 108 centimeters cubed. The middle 50% of volumes displayed by the boxes for the other four classes have no overlap with that range. However, the volume boxplots for A3, A4, and A5 abalones (and to some extent A2 abalones) overlap with one another quite a lot. The whole weight boxplot displays very similar trends to those just described for the volume boxplots.

While volume and whole weight appear to be mediocre predictors of age classification, the scatterplots indicate that volume and whole weight are potentially even weaker predictors of specific ages (as determined by rings). While a regression line could show a positive relationship between rings and volume or whole weight, the residual variance of that regression line would be very high, which would weaken the ability of volume and whole weight to serve as effective predictors of numeric ages on their own.

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.

Volume Matrix

	A1	A2	A3	A4	A5
Female	255.29938	276.8573	412.6079	498.0489	486.1525
Infant	66.51618	160.3200	270.7406	316.4129	318.6930
Male	103.72320	245.3857	358.1181	442.6155	440.2074

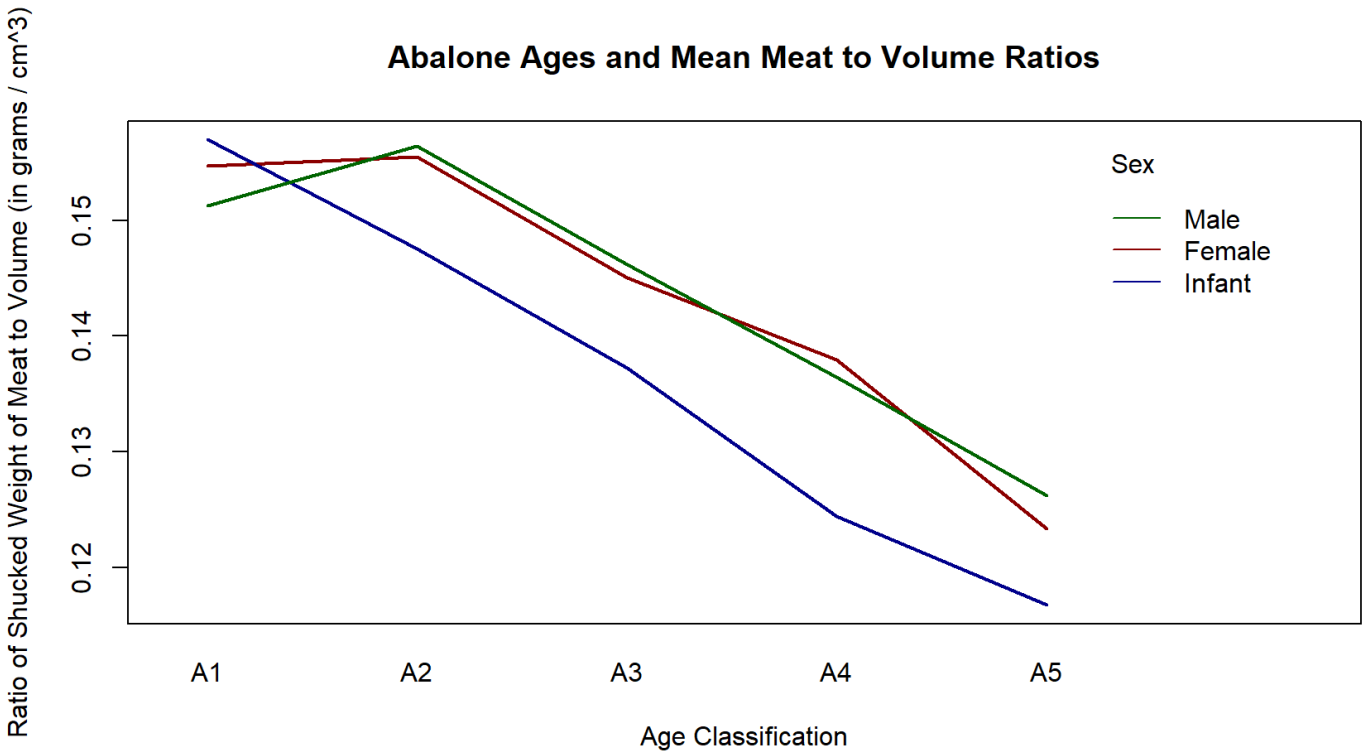
Shucked Weight of Meat Matrix

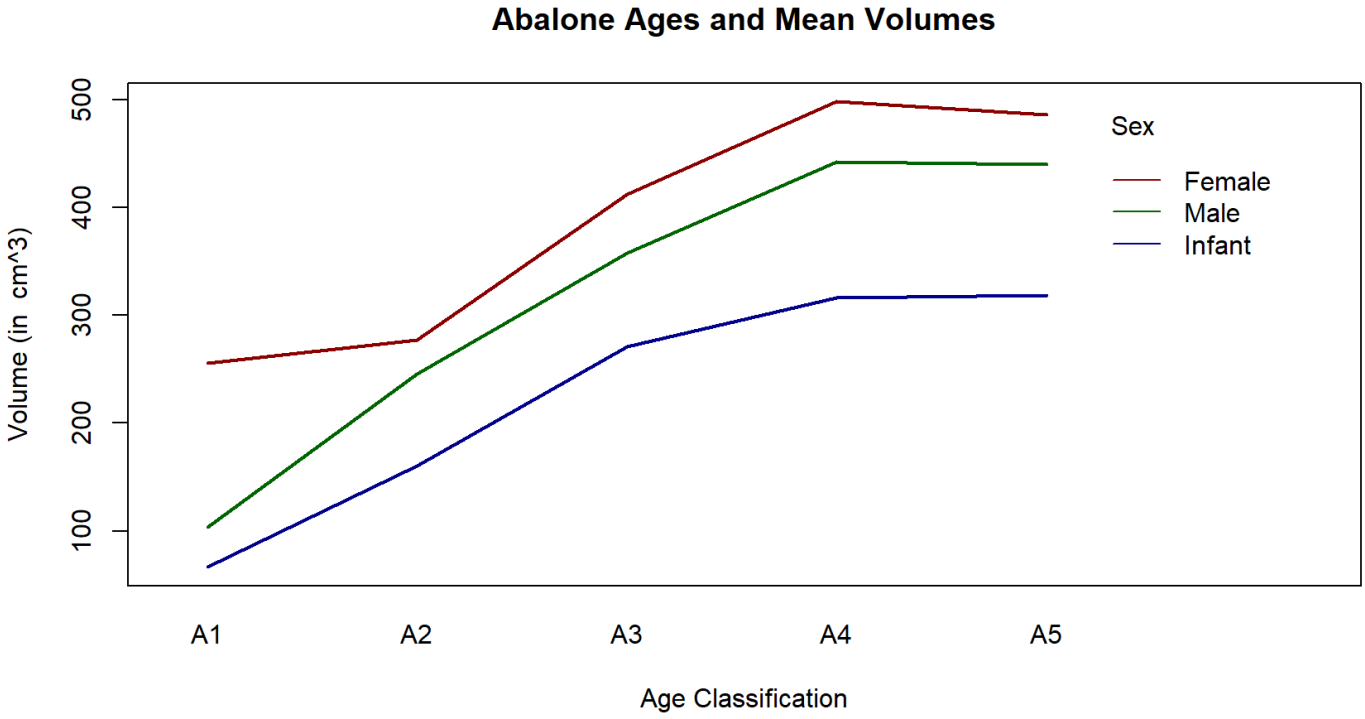
	A1	A2	A3	A4	A5
Female	38.90000	42.50305	59.69121	69.05161	59.17076
Infant	10.11332	23.41024	37.17969	39.85369	36.47047
Male	16.39583	38.33855	52.96933	61.42726	55.02762

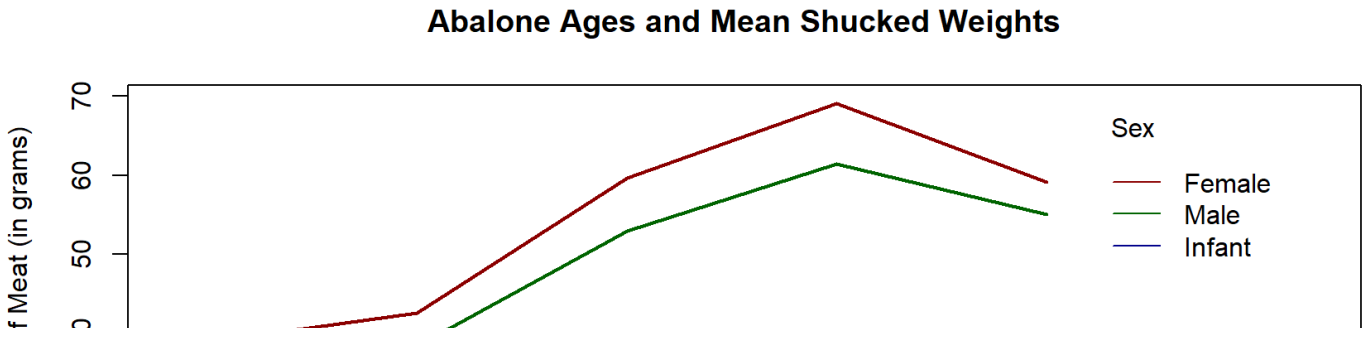
Matrix of Ratio of Shucked Weight of Meat to Volume

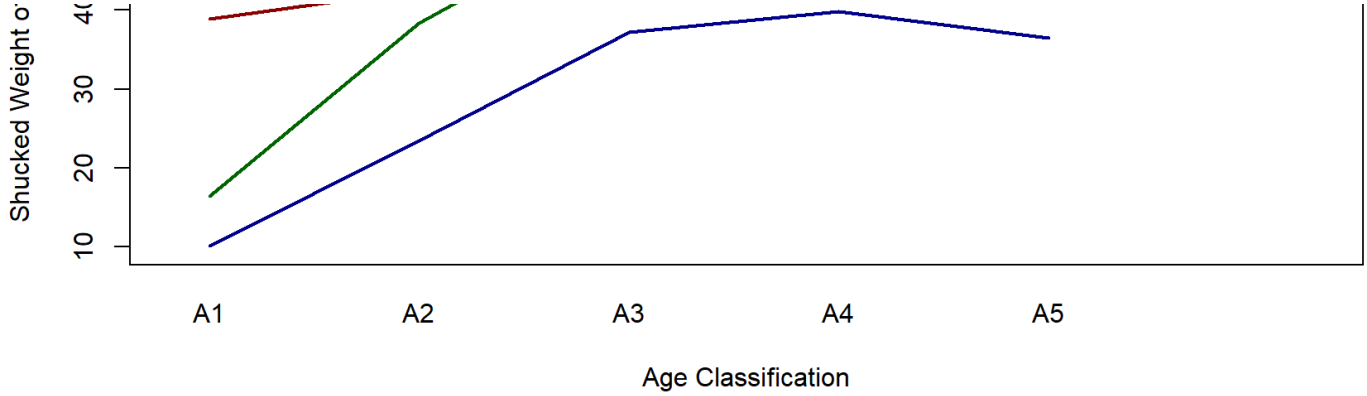
	A1	A2	A3	A4	A5
Female	0.1546644	0.1554605	0.1450304	0.1379609	0.1233605
Infant	0.1569554	0.1475600	0.1372256	0.1244413	0.1167649
Male	0.1512698	0.1564017	0.1462123	0.1364881	0.1262089

(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()*.


Abalone Ages and Mean Meat to Volume Ratios
Ratio of Shucked Weight of Meat to Volume (in grams / cm^3)
0.12
0.13
0.14
0.15
A1
A2
A3
A4
A5
Age Classification
Sex
Male
Female
Infant


Abalone Ages and Mean Volumes
Volume (in cm^3)
100
200
300
400
500
A1
A2
A3
A4
A5
Age Classification
Sex
Female
Male
Infant


Abalone Ages and Mean Shucked Weights
f Meat (in grams)
50
60
70
Sex
Female
Male
Infant



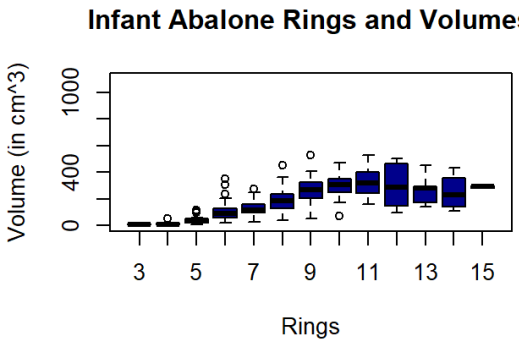
Essay Question (2 points): What questions do these plots raise? Consider aging and sex differences.

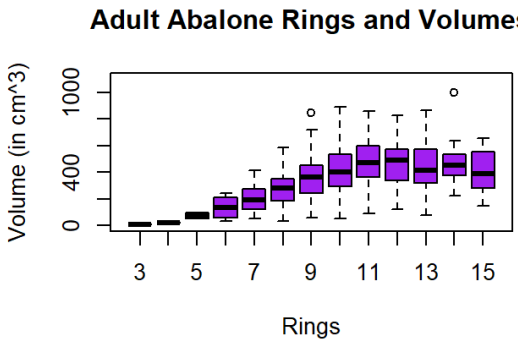
The plots provide some interesting insights that raise some intriguing questions. First, sex seems like it could play an important role in helping to predict abalone age. Specifically, on average, young female abalones seem to be over two times larger than young male or young infant abalones. This trend of female abalones being larger on average than other abalones persists for all five age classes, so it's possible that sex could be a helpful factor for predicting age class in a hypothetical regression model.

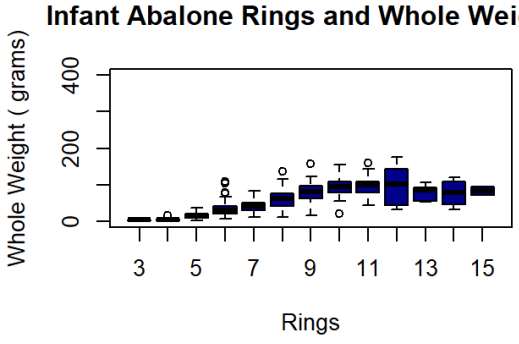
Another interesting finding is that for each sex, average abalone shucked weight seems to increase as abalones age for the first four age classes as abalones get older, but then seems to decrease around age class five. Combined with the fact that abalone volumes, on average, tend to stop getting larger around the same time, a very interesting trend appears as a result in the ratio plot. The ratio plot shows that when abalones are disaggregated by sex, a relatively linear, negative relationship between age classification and density ratio appears. This raises the question of whether sex and density ratio could be a helpful predictor of age (though if scientists cannot determine this ratio without killing the abalone to weigh the meat, then this finding may not be helpful in practice).

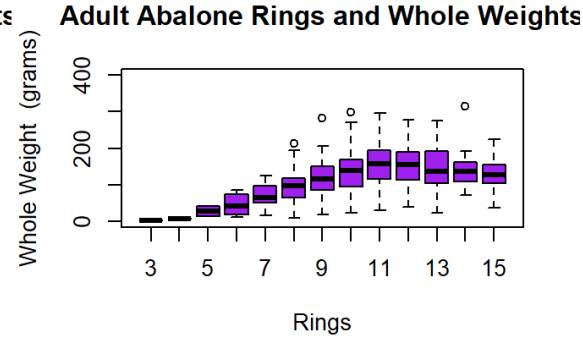
Answer: (Enter your answer here.)

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.









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

Answer: (Enter your answer here.)

These plots suggest that for both adults and for infants, abalone whole weights and volumes tend to increase as abalones age from having three rings to nine rings. However, as abalones age past having nine rings, their whole weights and volumes tend to level off, and the abalones don't appear to grow very much more.

Also, the plots seem to show that infant abalones tend to have lower volumes and whole weights than adult abalones do. For example, the plots suggest that on average, adult abalones stop growing at around 150 grams and 450 centimeters cubed. However, the plots suggest that, on average, infant abalones stop growing around 100 grams and 300 centimeters cubed.

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: (Enter your answer here.)

One reason that the original abalone study may have failed is because the physical attributes of the abalones in the sample may not have been particularly effective predictors of abalone age. Part of the reason for this is that abalones seem to stop growing in terms of whole weight and volume around the third or fourth age class. For that reason, size and whole weight may not be very strong predictors of age execpt for the youngest abalones. Notably, it does seem possible that, when combined with sex information, the ratio of shucked weight to volume may be a decent predictor of abalone age. However, it's likely that the researchers had to kill the abalone in order to determine the shucked weight, so even if this model was effective, it wouldn't provide answer the researchers question: "How do we determine an abalone's age without killing the abalone?"

Another reason that the original abalone study may have failed is due to data quality. It doesn't entirely make sense to me why there would be 238 infant abalones in age classes A2 - A5. This raises concerns regarding the quality of the data measurements for sex, rings, and age class. If the quality of the data is sub-optimal, then the ability of researchers to draw conclusions may be undermined.

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: (Enter your answer here.)

If I were presented with a histogram and summary statistics for a population, I would be interested in learning a bit more information prior to accepting the observations as representative of the entire population. First, I would be curious to know what the sample size and population sizes are. If the sample size is quite small compared to the population size, then I might be skeptical that the sample size is representative of the population. Second, I would be interested in knowing what sampling technique the researcher had used to select observations. If the researcher did not use an established random sampling technique, such as simple random sampling, then the sampling methodology may have been biased. Third, I would be interested in whether the researcher experienced any issues with measurement that could be impacting the quality of the data. If the researcher had a difficulty measuring certain variables and had to guess or leave values blank in the dataset, then the quality of the data may be diminished and the sample data may not reflect the population data.

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: (Enter your answer here.)

One of the primary difficulties with analyzing data derived from observational studies is determining causality. That's because experiments generally aim to randomly assign treatments to certain observations prior to measuring the effect, so that researchers can aim to determine whether the treatment caused a measurable difference in the measured effect. Observational studies, however, generally just measure observations without applying treatments. Therefore, while observational studies can be used to identify correlations between variables, they are not particularly effective at determining causality.