4. In order to effectively use R to analyze a file of data, you need to learn several basic commands. You can view the manual for a specific command, for instance `subset`, by typing `? subset` at the R shell's prompt. In this phase of the assignment, you should learn how to use several R commands and write a short description of their input, output, and behavior, including one concrete example of the command in action with the data set discussed at a later stage of the assignment. Please study the following commands:

Commands:

attach – Places a data set in the search path

Input: A database, an integer of where to attach in search(), and a name to use for the database.

Output: None

Example: attach(algae)

names – Gets or sets the name of an object

Input: An R object

Output: Header names

Example: names(algae)

```
[1] "season" "size"   "speed"  "mxPH"   "mnO2"   "Cl"     "NO3"    "NH4"
[9] "oPO4"   "PO4"    "Chla"   "a1"     "a2"     "a3"     "a4"     "a5"
[17] "a6"     "a7"
```

head – Returns the first part of data

Input: An object

Output: The first part of data

Example: head(algae)

```
 season  size  speed mxPH mnO2     Cl    NO3     NH4    oPO4     PO4 Chla   a1
1 winter small medium 8.00  9.8 60.800  6.238 578.000 105.000 170.000 50.0  0.0
2 spring small medium 8.35  8.0 57.750  1.288 370.000 428.750 558.750  1.3  1.4
3 autumn small medium 8.10 11.4 40.020  5.330 346.667 125.667 187.057 15.6  3.3
4 spring small medium 8.07  4.8 77.364  2.302  98.182  61.182 138.700  1.4  3.1
5 autumn small medium 8.06  9.0 55.350 10.416 233.700  58.222  97.580 10.5  9.2
6 winter small   high 8.25 13.1 65.750  9.248 430.000  18.250  56.667 28.4 15.1
   a2   a3  a4   a5   a6  a7
1  0.0  0.0 0.0 34.2  8.3 0.0
2  7.6  4.8 1.9  6.7  0.0 2.1
3 53.6  1.9 0.0  0.0  0.0 9.7
4 41.0 18.9 0.0  1.4  0.0 1.4
5  2.9  7.5 0.0  7.5  4.1 1.0
 14.6  1.4 0.0 22.5 12.6 2.9
```

mean – Returns the arithmetic mean of the object supplied

Input: An R object

Output: Returns the mean of the dataset

Example:
```
mean(a1)
[1] 16.9235
```

median – Returns the median of the dataset

Input: An object containing the values to be used to compute the median

Output: Returns the median of the dataset.

Example:
```
median(a1)
[1] 6.95
```

subset – Returns a section of data from a set which meet given conditions.

Input: An object, a logical expression indicating what rows to keep, and the columns to keep data from.

Output: Returns the set of data as determined by the conditions provided.

Example:
```
subset(algae, size == "medium" & season == "spring", select = c(a1, a2, a3))
      a1   a2   a3
75   1.6  8.0 17.6
76   2.2  9.6  5.0
79  14.4  0.0 11.8
81  10.8  0.0  0.0
91   0.0  5.5  3.3
94   0.0  3.1  3.5
98   1.2 16.2  0.0
102  7.0  0.0 13.5
112 46.6  0.0  2.5
114  3.7  1.4  1.1
119  1.9 12.7 25.9
122  1.6  8.9  6.6
125  1.7  0.0 10.3
128  0.0 16.4 10.1
131  4.1  0.0 25.3
134  2.4  1.7  4.2
136 10.3 26.5  6.1
142 19.0  0.0 22.0
145  4.4 11.2  6.8
150  1.9 25.4 21.7
154  3.4 21.5 14.0
```

ls – Gives the names of the objects in the environment

Input: Takes the environment to use in listing the object names.

Output: Returns the names of the objects in the provided environment.

Example:
```
ls(algae)
 [1] "a1"   "a2"   "a3"   "a4"   "a5"   "a6"   "a7"   "Chla"
 [9] "Cl"   "mnO2" "mxPH" "NH4"  "NO3"  "oPO4" "PO4"  "season"
[17] "size" "speed"
```

summarize / summary – A generqic function that produces an overview of the object provided based on the "class" of the given object.

Input: The object to provide the summary for

Output: The summary of the object provided – output varies dependent on the class of the object.

Example:
```
summary(algae$a1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.50    6.95   16.92   24.80   89.80

summary(algae$a1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.50    6.95   16.92   24.80   89.80
```

5. After you have learned more about how each one of these commands works, you should learn more about the `algae` data set that is part of the DMwR library. What are the names of the attributes in this data set? How many attributes are in the data set? How many rows are in the data set? While you must use the R programming language to answer these questions, you can check your answers by visiting the Algae Data Set Description in the UCI Machine Learning Repository and the Predicting Algae Blooms Description in the DMwR book. Of course, you can also consult the DMwR book that is on reserve for our class.

Strategy:
To find this information I am going to use the head() function to give me the attribute names and the number of attributes. To get the number of rows I am going to just type the name of the dataset and look at the number of listed rows.

Code:
```
> head(algae)
> algae
```

Answers:

The attribute names are: a1, a2, a3, a4, a5, a6, a7, Chla, Cl, mn02, mxPH, NH4, NO3, oPO4, PO4, season, size, speed

How many attributes in the dataset: There are 18 attributes.

How many rows are in the data set: There are 200 Rows of data.

6. The `a1` attribute gives the frequency number of a harmful algae known as "a1" in the data set. Please note that the data set does not contain any information about the name or the characteristics of this type of alga. Each value in the `a1` attribute corresponds to a frequency for this algae being found in the specified environment. In this case, small values are better since they indicate the presence of less of this harmful algae. Using the mean and the median values of the frequencies for alga `a1`, is this algae more likely to bloom in small, medium, or large rivers?

Strategy:
To determine whether algae 1 prefers small, medium, or large rivers, I will look at the summary of the subset of data from the algae data collection after restricting the data to only the a1 column and showing it in three samples using the river size of small, medium, or large as a condition.

Code:
```
> summary(subset(algae, size == "small", select = a1))
> summary(subset(algae, size == "medium", select = a1))
> summary(subset(algae, size == "large", select = a1))
```

Applicable data returned by the above code:

a1	**Small**	Medium	Large
Mean	**27.15**	11.268	11.35
Median	**19.40**	3.650	2.40

Complete Answer:
After reviewing the above data, I determined that algae 1 prefers a small river in order to thrive.

7. The `a2` and `a3` attributes respectively give the frequency number of a harmful algae known as "a2" and "a3" in the data set. Using the mean and the median values of the frequencies for algae `a2` and `a3`, are these algae more likely to bloom in small, medium, or large rivers?

Strategy:
To determine whether algae 2 and 3 prefer small, medium, or large rivers, I will look at the summary of the subset of data from the algae data collection after restricting the data to the a2 and a3 columns while showing it in three samples using river sizes of small, medium, or large as conditions.

Code:
```
> summary(subset(algae, size == "small", select = a2))
> summary(subset(algae, size == "medium", select = a2))
> summary(subset(algae, size == "large", select = a2))
>
> summary(subset(algae, size == "small", select = a3))
> summary(subset(algae, size == "medium", select = a3))
> summary(subset(algae, size == "large", select = a3))
```

Applicable data returned by the above code:

a2	Small	Medium	**Large**
Mean	5.283	7.862	**10.14**
Median	0.000	2.750	**9.70**

a3	Small	**Medium**	Large
Mean	3.23	**5.537**	3.722
Median	1.00	**2.000**	1.800

Complete Answer:
After reviewing the above data, I determined that algae 2 prefers a large river in order to thrive, while algae 3 favors a medium sized river to thrive.

8. The algae data set also contains information about the speed of the rivers, as stored in the `speed` attribute. Using the mean and the median values for the `a1`, `a2`, and `a3` algae, are these algae more likely to bloom in rivers with a low, medium, or high speed?

Strategy:
To determine whether algae1, 2, and 3 prefer low, medium, or high speed rivers, I will look at the summary of the subset of data from the algae data collection after restricting the data to the a1, a2, and a3 columns while showing it in three samples using river speeds of low, medium, or high as conditions.

Code:
```
> summary(subset(algae, speed == "low", select = a1))
> summary(subset(algae, speed == "medium", select = a1))
> summary(subset(algae, speed == "high", select = a1))
>
> summary(subset(algae, speed == "low", select = a2))
> summary(subset(algae, speed == "medium", select = a2))
> summary(subset(algae, speed == "high", select = a2))
>
> summary(subset(algae, speed == "low", select = a3))
> summary(subset(algae, speed == "medium", select = a3))
> summary(subset(algae, speed == "high", select = a3))
```

Applicable data returned by the above code:

a1	Low	Medium	**High**
Mean	9.209	12.48	**24.345**
Median	3.600	2.80	**17.050**

a2	**Low**	Medium	High
Mean	**10.7**	9.381	4.286
Median	**10.4**	4.600	0.000

a3	Low	Medium	**High**
Mean	3.082	4.375	**4.727**
Median	1.500	1.900	**1.400**

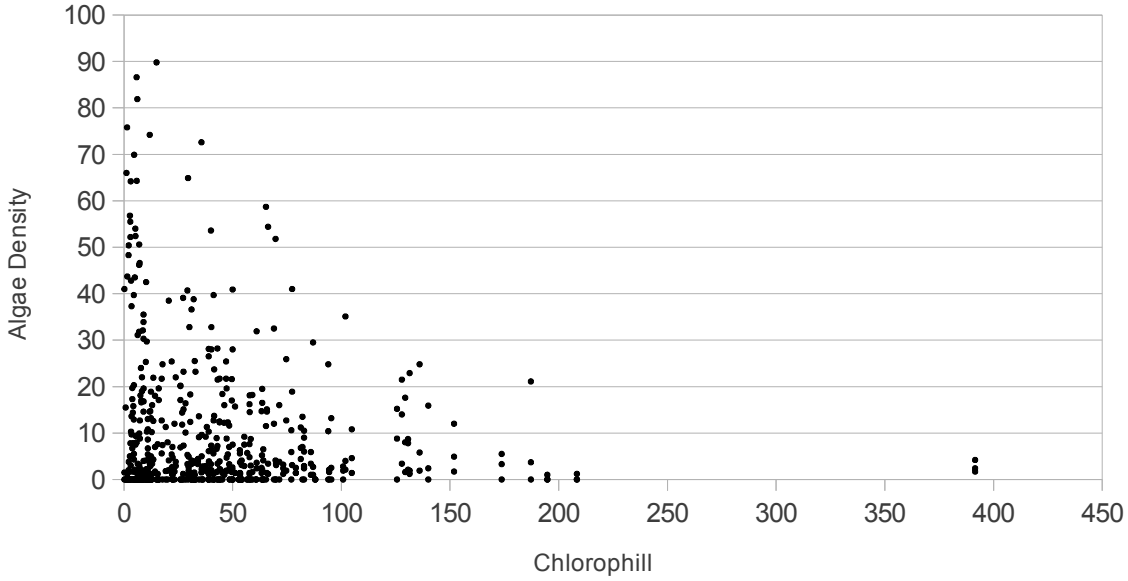
Complete Answer:
After reviewing the above data, I determined that algae 1 prefers a high speed river in order to thrive, while algae 2 favors a low speed river, and algae 3 also prefers a high speed river in order to thrive, similar to algae 1.

9. The `Cl` attribute in the algae data set describes the mean amount of chlorophyll in the rivers. What is the relationship between the mean value of chlorophyll and the amount of the `a1`, `a2`, and `a3` algae on the rivers? For instance, if a river has a high amount of chlorophyll, does this mean that it will contain a high or a low amount of each algae?

Strategy:
To determine whether algae1, 2, and 3 prefer low, medium, or high speed rivers, I will look at the summary of the subset of data from the algae data collection after restricting the data to the Cl, a1, a2, and a3 columns. From here I will find a pattern, and if needed chart the points to determine chlorophyll's impact on algae growth

Code:
```
> subset(algae, select = c(Cl, a1, a2, a3))
```

Applicable data returned by the above code:



Complete Answer:
After reviewing the data returned by the summary of the subset I was unable to bather any useful information. Then I decided to analyze the raw data, and after not being able to find a pattern I graphed the points. There is a weak connection between chlorophyll and algae density, but it appears that the less chlorophyll in the water the more the various algae can thrive, with algae 1 benefiting the most from a minimal amount of chlorophyll.

10. There are many other interesting trends evident in the algae data set. Using any combination of R commands, please identify and try to explain one additional trend. If you are interested, you may also want to explore the visualization of a trend by using the Lattice package that is available by typing `library(lattice)` at the R prompt. For extra credit, you may turn in visualizations, in addition to your code segments, for all of the previous questions.

Strategy:
To determine whether algae1, 2, and 3 prefer spring, summer, autumn, or winter I will use the code below to show the mean algae densities at each of the four seasons by filtering only data for that season for a given algae.

Code:
```
> summary(subset(algae, season == "spring", select = a1))
> summary(subset(algae, season == "summer", select = a1))
> summary(subset(algae, season == "autumn", select = a1))
> summary(subset(algae, season == "winter", select = a1))
>
> summary(subset(algae, season == "spring", select = a2))
> summary(subset(algae, season == "summer", select = a2))
> summary(subset(algae, season == "autumn", select = a2))
> summary(subset(algae, season == "winter", select = a2))
>
> summary(subset(algae, season == "spring", select = a3))
> summary(subset(algae, season == "summer", select = a3))
> summary(subset(algae, season == "autumn", select = a3))
> summary(subset(algae, season == "winter", select = a3))
```

Applicable data returned by the above code:

a1	Spring	Summer	**Autumn**	Winter
Mean	16.650	16.110	**17.750**	17.220

a2	Spring	Summer	**Autumn**	Winter
Mean	6.894	6.433	**9.338**	7.473

a2	**Spring**	Summer	Autumn	Winter
Mean	**6.926**	3.011	1.552	4.794

Complete Answer:
After reviewing the above data, I determined that algae 1 prefers autumn in order to thrive, while algae 2 also favors autumn, but algae 3 prefers spring in order to thrive. This tells me that even though the density was fairly even throughout the year, all three algae prefer more temperate weather to grow.

Lab Review: (Also in on-line lab notebook)
This lab proved to be significantly more challenging than I had originally expected, though once I finished question 6, the lab was much easier to complete.

The most difficult part was learning how to use R and manipulating the data to fit the context of the question I was answering. Once I figured out the set of instructions to give to R, I was able to move fairly quickly through the lab.

The next difficulty I encountered had to do with the dataset itself. After answering question 5, I went to the website to check my answer and found that the number of data rows and columns from the website did not match the dataset I was using. I expected them to be the same, but they are not so I just stuck with the answers I was seeing first hand in my terminal.

The last problem I encountered was the sole fault of my own. When committing to my repository I inadvertently pushed the temporary and recovery files created by open office. When I went back to remove them and add a gitignore exception I deleted the temp file and my lab report itself. At that point I had to revert all of the committed changes and re-clone the repository and try again, this time with success.

Barring the near heart attack mentioned above I enjoyed the lab. Being my first full lab report I hope I understood all of the requirements and tried to lay out the report in a reasonably organized manner.