Week 1 Assignment: Exploring Data

Instructions

Each question is broken into several parts. You should provide R code, a written description, or both as called for in each part. Upload your solutions to the dropbox. Please include your netID and a description in the name of each file you upload (this makes it much easier for us to organize the submissions).

Unlike the other datasets and materials used in the course, the UW weather dataset used in question 2 is not available through the Catalyst site. It is larger than is allowed there. Instructions for downloading the file are given in question 2.

1. Abalone Data

We used summary statistics, histograms, and box plots to explore the Abalone data in class. Now we’ll use aggregation to continue to explore that dataset. We know the height of each abalone, but we can more clearly identify some trends if we map the large number of unique heights to just a few height groups. We’ll also take a look at another kind of aggregation (averaging) to reveal a relationship between weight and age. Some of the solutions to this problem are contained in the slides we didn’t have time to cover in the first lecture. Please try them on your own first.

* 1. Use the command “quantile” to find the deciles (10 groups) for height from the complete data set. Hint: you may find the command “seq” helpful.
  2. Use the command “cut” to assign each height value to the corresponding decile (e.g., the smallest values are assigned to the first decile and get mapped to the value, 1). Hint: use “as.numeric” to get integer values instead of ranges.
  3. Now create a table of age vs. height decile. Examine the table and describe what you observe.
  4. Another way to aggregate the data is averaging. Let’s compute the average whole weight of abalone as a function of age and plot the relationship.
     1. Use the commands “unique” and “sort” to find the unique values of Age and store the values in ascending order to a variable named “ua”.
     2. Use the command “sapply” to apply a function to each value in “ua”. The function should return the mean whole weight of all abalone of a given age. Hint: type “help(‘function’)” to find out more about user defined functions. The quotes inside the parentheses are important.
     3. Finally, use the “plot” command to plot mean weight vs. age. Describe the relationship revealed by the plot. Include an explanation for the behavior seen in the abalone of the 25-30 year age group.

1. Weather Data

The atmospheric science department at UW has been collecting [weather data from a station set up on their rooftop (in Seattle) since 1998](http://www.atmos.washington.edu/cgi-bin/list_uw.cgi). The station records time, temperature, humidity, and similar measurements once every minute. Let’s explore this dataset in order to become familiar with the data, sanity check to verify data quality, and finally look for some simple relationship and trends. Use the same data set we used in class (weather\_data\_2000\_2014.csv). If you need to download the dataset again, a download link is available on the Catalyst site.

* 1. Using summary statistics, histograms, boxplots, or other means identify and describe at least one data quality issue in the dataset.
  2. Filter the data to remove the questionable data you identified in part a. How much of the data is affected? Hint: some of the functions “length”, “nrow”, “ncol”, “is.na”, and “which” may be helpful.
  3. Look for and describe a monthly trend in the data.