**Lab 1: Loading data into R and basic descriptive statistics**

Geog 4300/6300-Shannon

Fall 2016

**Due**: Sunday, Sept 6th

**Value**: 20 points

**Overview**: This lab is intended to assess your ability to use R to load data and to generate basic descriptive statistics. You’ll be using monthly weather data from the Daymet climate database (http://daymet.ornl.gov) from four cities: Athens, GA; Columbus, OH; Eugene, OR; and Los Angeles, CA. These data are on the course ELC site in the CSV file labeled “Daymet\_FourCities”. The following variables are provided:

* year: Year of observation
* yday: Day of the year (Jan. 6 = 6, Feb. 6 = 37, etc.)
* month: Month of observation
* tmax: Maximum recorded temperature on that day (Celsius)
* tmin: Minimum recorded temperature on that day (Celsius)
* VapPres: Vapor pressure (indicative of humidity)
* Precip: Recorded precipitation (mm)
* Day\_leng: Length of daylight (in seconds)
* Elevation: Elevation of the weather station (meters above sea level)
* Lat: Latitude of the weather station
* Lon: Longitude of the weather station
* City: City name

These labs are meant to be done collaboratively, but your final submission should demonstrate your own original thought (don’t just copy your classmate’s work or turn in identical assignments). Your answers to the lab questions should be typed in a separate Word document and turned in using the Assignment Dropbox on the ELC site.

**Procedure:**

1. Download the csv file from the course ELC site. Load the file into R and closely examine each variable.

***Question 1*** *(4 points)****:*** *Provide an example of nominal, ordinal, interval, and ratio data within this dataset.. Explain why each fits in the level of measurement you chose in a sentence or two . If you cannot find an example for one of these four data types (no nominal variables, for example), given an example of climate data that would fit this type.*  
  
There are a lot of observations here, 51,100 to be exact. One easy way to simplify it is to use the aggregate function in R to get simple descriptive statistics. Here’s an example:

aggregate(tmax~City, data=Daymet\_FourCities, mean)

This command should return the mean value of maximum temperatures for each city during the timeframe in which data was available. You can replace “mean” with “sd” to get a similar summary of standard deviation.

Calculate the mean and standard deviation of the monthly MINIMUM temperatures for each city, adapting the command shown above. Copy and paste both your command and the resulting figures into your lab document.

Now try a VERY simple model of climate change. Let’s say that 100 years from now, temperatures in these cities will be warmer by exactly 2 degrees Celsius. You can create a new variable showing the projected new minimum temperatures using this format:

Daymet\_FourCities$tmin\_new<-Daymet\_FourCities$tmin+2

When you reload the table (click on it in the “Environment tab”), the new variable will appear. Calculate the mean and standard deviation for this new variable for each city and copy and paste your command and result into your lab document as well.  
  
***Question 2*** *(3 points)****:*** *What is the mean and standard deviation for the original minimum temperature variable and the new one? How do these compare? Explain any similarities or differences you find.*

1. Now pick one variable to explain: tmax, tmin, or VapPres. We want to create a new dataset where these values are summarized by each year. You can aggregate it this way:

* Year.sum<-aggregate(tmax~year+City,data=Daymet\_FourCities, FUN=mean)

Notice how much smaller this dataset is already. Plot it out using ggplot:

* ggplot(Year.sum, aes(x=year,y=tmax, group=City, colour=City))+geom\_line() + facet\_wrap(~City)

You can also do a boxplot using ggplot:

* ggplot(Year.sum, aes(x=City,y=tmax))+geom\_boxplot()

And also a histogram:

* ggplot(Year.sum, aes(x=tmax,group=City,fill=City))+geom\_histogram()

We can use the subset command to further specify things, select only a single month for comparison over this timeframe.

* July.data<-subset(Daymet\_FourCities, month==”July”)

You can then summarize and graph out this dataset as well using the aggregate and ggplot commands listed above. Adapt this command to create a one month subset, changing “July” to a month of your choosing.

Now write an R script that does the following:

* Creates a single box plot comparing these four cities for the variable you’ve chosen for a single month using the ggplot2 package.
* Creates separated line charts for each of these four cities using the same variable/month, using ggplot2 and the facet\_wrap command as shown above.
* Creates a single histogram for this variable/month showing the distribution of the variable using ggplot2 and the same facet wrap command used for the line chart as shown above.

Don’t worry about things like column names or customization for now—these will be addressed in lab 2.

***Question 3*** *(7 points): Copy and paste your script and resulting graphics your lab responses. In a paragraph or two, describe what these three graphing approaches tell you about the distribution of data for this variable and/or its change over time. Is the data normally distributed or skewed? How do the cities compare to each other? What other overall patterns do you see? Does one graphing approach show these patterns more effectively than another?*

You can also use R to calculate multiple descriptive statistics for these data. You’ve already calculated the mean and standard deviation above using the aggregate function. You can also calculate median and interquartile range (IQR) using this method.

Two other commonly used measures of distribution are skewness and kurtosis. Skewness and kurtosis require you to install the psych package in R. Using the subset command, create separate datasets for each city. Using the “skew” and “kurtosi” functions (part of the psych package) on each subset will return the values you need. Your textbook explains these values on p. 52-53.

***Question 4*** *(6 points):**For the summary data you created for a single month, calculate the following for each of the cities in this dataset: mean, median, standard deviation, interquartile range, skewness, and kurtosis. Create a table listing the values of these for each city for the variable and month you’ve chosen. This table should have a single row for each city showing the above summary statistics for the monthly values (mean/median average temp, mean total rainfall, etc.). In 2-3 paragraphs, explain what you think these results tell you about the central tendency and distribution of monthly temperatures for each city.*