Lab Two: An Introduction to R/Rstudio Cont'd

STA 111 (Summer Session I)

Lab Objective

To learn how to do basic data exploration.

Lab Procedures

For this lab, use the RStudio pre-installed on any of the lab computers. Last time we learned how to create new data sets. Create a new data frame for the ten most fatal earthquakes on record. Call the new data, "newdata". Use `" "` for the countries to signify that they have string values. `c("Haiti","China")` would create a string vector with those two countries.

Country	Deaths
Haiti	92,000
China	242,769
Iran	150,000
China	235,502
Indonesia	230,210
Syria	230,000
China	820,000
Iran	200,000
Turkey	240,000
Japan	142,800

After you input all the data, lets try the `which` command. Type `which(newdata$Country=="Turkey")`. Also try `which(newdata$Country=="China" & newdata$Death==820000)`. Now answer the following questions. You don't have to turn in anything for questions A and B. Their purpose is to get you familiar with RStudio.

Questions:

How many earthquakes killed 200,000 or more people?

With ten cases its straightforward to look at the data and get an accurate count. But with a longer dataset, counting the incidences of each number by hand would be cumbersome. In such settings, you can make life easier by sorting the numbers in increasing order, then counting the incidences. Do that by typing `newdata[order(newdata$Deaths),]`. Even better, you can find the exact rows that satisfy `Deaths` $\geq$ 200,000 by typing `newdata[which(newdata$Deaths` $\geq$ `200,000),]`.

If you want to sort the data first by country and then by death toll (i.e. have all countries in alphabetical order with fatalities listed in increasing order by country), which command would you use?

Try typing `newdata[order(newdata$Country,newdata$Deaths),]` .

Can you think of anything else?

Creating a Reproducible Lab Report

We will be using a markdown language, R Markdown, to type up the lab report. This allows you to complete your lab entirely in RStudio as well as ensuring reproducibility of your analysis and results. To help get you started, a template is provided for you. Use the following code to download this template:

```
download.file("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2016/Labs/Lab2.Rmd", destfile ="Lab2.Rmd")
```

You will see a new file called Lab2.Rmd in the Files tab on the pane in the bottom right corner of your RStudio window. We will refer to this as your R markdown file or your report. Click on the file name to open the file. All you need to do to complete the lab is to type up your brief answers and the R code (when necessary) in the spaces provided in the document. Earlier in the lab spaces are provided for you to enter R code chunks. Later in the lab youll need to figure out whether code is needed to answer a particular question, and if so a new chunk can be inserted by clicking on the Insert Chunk button (dropdown menu under Chunks on the upper right corner of your markdown document). Before you keep going type your name. Then click on Knit PDF and youll see your document in a new pop-up window. You can save the pdf and email it to yourself or simply submit directly on Sakai.

Lab Questions:

Load in the data set Forbes94, which contains the 1994 compensation information for Chief Executive Officers (CEOs) of several large companies. You can load the data by typing:

```
Forbes94 = read.table("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2016/Labs/Forbes94.txt",header=T)
```

When you get a data set, the first thing to do is to figure out how many variables and how many units of observation you have to play with. Do you remember how to check the number of variables and individuals? Let's get into some data analyses. Compile your answers in the markdown file. You are permitted and encouraged to talk about questions with your classmates, but write up your lab report with your own words.

1. *R displays missing values with NA.* True or false: There are more than five CEOs whose values of total compensation are missing in the data file. (Hint: You can do this by sorting the data and browsing).

2. What is the salary (not total compensation) of the CEO of Duke Power? (Hint: try the `which` command).

3. Of all CEOs, which has the highest total compensation? Which has the lowest total compensation?

4. Which industry type has the highest average CEO total compensation? Be careful not to read the decimals incorrectly when you answer the question. (Hint: Try the `summary` command).

5. How many of these CEOs got their undergraduate degree from Duke?

6. Highest attained educational degree is in the variable GradDegree. Which degree has the highest total compensation: MBA (business), JD (law), MD (physician), PhD, or no graduate degree? Use highest average total compensation as your criterion, and choose only from these categories.

This ends the lab. Remember to turn in your lab reports on Sakai.