Assignment 7
Suicide data set: regression, correlation, hypothesis testing. Wednesday March 19, 2014
Due date: Tuesday April 1, 2014, midnight.

In this assignment, we work with the suicide dataset. You are to read in the dataset, create some new columns, compute some correlations, regressions and some hypothesis tests. However, the problems states will not discuss R. It is up to you to figure out how to accomplish the tasks asked of you. All the datasets can be found in the link provided in the homework assignment given on March 19.

The objective of this assignment is to get you used to using R from a problem description that does not you anything about R commands. I am not concerned with how realistic the lab is compared to an actual experiments. You presumably learn to condut experiments in other courses, so you are free to apply all you have learned in this lab. If you disagree with some of is listed below, please state this in your report and state your reasons. Feel free to change the order of some of the tasks (give your reasons).

Some of the columns, after getting rid of NAs and setting elements to zero, might not have enough data. That is fine. Just state this in your report.

The point of departure is the suicide dataset, downloadable from the course web site. The file name is: *VanOrden_for dist.csv*. Most of the columns relate to some aspect of somebody's state of mind, such as Burden to one's parents, anxiety level, depression, etc. Scores range from 0 to 7 for each column that correspond to a question answered regarding suides. We will combines columns in a given category to get scores that go from 0 to $7 * nb\_scores$. For example, if we combine two columns with scores from 0 to 7, the combined column has a score that goes from 0 to 14.

I am aware that this is not the best approach (in terms of combining columns. A better approach is to take the average of all the columns. Thus, if combining columns BSS1 through BSS7, sum up BSS1 through BSS7 and divide by 7.

Tasks:

1. Please note the number of columns in your report.

2. Combine the columns that relate to burden (bur), loneliness (isel), depression (bdi), anxiety (bai) and suicide ideation (bss) into single columns. For example, there are columns bur1 through bur15, which should be combined into a column labelled burden (or any other name you choose.)

3. This data set contains some "NA" elements, to be removed using the command "na.omit" (or better yet, replace them by zero), and 999, which should be replaced by 0 (use the command "replace"). You'll notice an additional "99" somewhere, which you can also replace by zero. Look at documentation and tutorials on google to do this.

4. Construct a new dataset with the following columns: age, gender, burden,

loneliness, depression, anxiety, and suicide ideation. The labels are up to you, but should be short and descriptive.

5. Calculate the correlation between the following columns and suicide ideation: burden, depression, loneliness, anxiety.

6. Plot scattergrams of each column against each other column (except for age and gender), and identify the plots that are the most meaningful to you. Calculate and plot the linear fit for only the plots that look meaningful. State your finds in your report.

7. Is the mean suicide ideation significantly different for women and for men?

8. Plot the mean suicide ideation as a function of age using line or box plots. (Hint: this involves computing the mean suicide ideation for each age category.)

The report should contain the first 10 lines of your new dataset, the images created, the script used to answer the questions and explanations of the plots you find relevant and why. Give the output to the questions relating to "mean suicide ideation".