Assignment 2

BIOSTAT 213: Introduction to Computing in the R Software Environment

Due: Wednesday, February 20, 2019

Collaboration with other students in this course is permitted on this assignment. However, you must write all work—including R code—in your own voice, as much as possible, and acknowledge your collaborators as thoroughly as possible.

The assignment is due online, in the Collaborative Learning Environment, by 9 am, Wednesday, February 20, 2019. You *must* submit a Word-compatible document or a PDF file.

You will be using the `nhanes.csv` and `unps.csv` data files for this assignment. The questions below do *not* assume any of the modifications to the data from lab or slides. In other words, you may begin this assignment by simply re-importing the CSV files.

For each of the below, share your code as well as the output (if there is any output). If your answer to a question is unclear from your R output, or difficult to identify, include an answer in text form (in addition to the code and output).

1. *Use* ***nhanes.csv****.* We saw on slide 63 of Module 3 that ggplot2's `geom_boxplot()` automatically includes missing x values (note the NA group on the far right). A method for removing the missing values from the plot is to use the `data` argument to `ggplot()` to specify the subset of the data with non-missing x values. Using ggplot2, reproduce the plot from slide 63, using the scientific ordering of the levels, and excluding the missing x values (ie, you should no longer see the NA group in the plot).

2. *Use* ***nhanes.csv****.* As mentioned in class, a critique of box plots is that they don't provide a complete visualization of distributions. A proposed solution to this problem is the violin plot. Redo the plot above, but substitute `geom_violin()` for `geom_boxplot()`. Use the `draw_quantiles` argument to add lines for the 25th, 50th, and 75th percentiles, as shown in one of the examples here.

3. *Use* ***nhanes.csv****.* Find the mean body mass index (`bmxbmi`) among individuals with a history of smoking (`hs`). Find the mean body mass index among individuals without a history of smoking.

4. *Use* ***nhanes.csv****.* Restricting analysis to individuals 30 years of age (`ridageyr`) or younger, conduct a t-test, using body mass index as the y variable (`bmxbmi`) and history of military service (`military`) as the x variable. You may consider using the `data` argument to `t.test()`. Use object assignment to save the test. Without copying and pasting numeric values, use the test object to find the estimated military-minus-no mean difference in body mass index.

5. *Use* ***nhanes.csv****.* Use `head()` to list the first few rows of body mass index (`bmxbmi`), weight (`bmxbmi`), and height (`bmxht`), with the rows sorted by body mass index (in increasing order). Your output should look like this:

```
     bmxbmi bmxwt bmxht
1295   14.1  32.8 152.3
5718   14.2  36.5 160.4
1499   15.2  42.2 166.8
2762   15.4  47.4 175.7
3368   15.4  50.8 181.7
3764   15.4  32.3 144.6
```

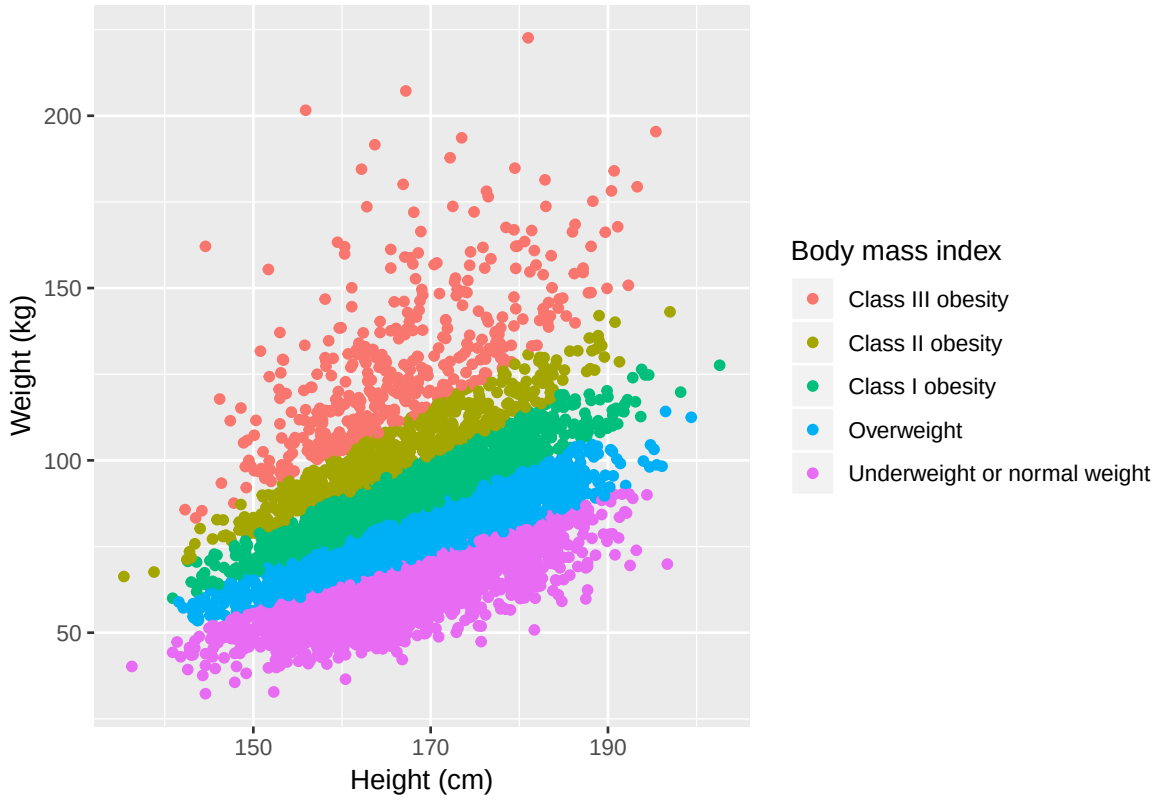
6. *Use* ***nhanes.csv****.* The `with()` function, mentioned in the section on categorization in Module 4, is useful for conducting stratified analysis.

   (a) Use `with()` to cross tabulate history of gender (`gender`) and smoking (`hs`), among individuals 75 years of age (`ridageyr`) or older. Place gender along the rows and smoking along the columns.

(b) Compute the row-wise percents.

7. *Use* ***nhanes.csv.*** How many individuals are missing a value for *any* (ie, one or more) of the following 3 variables: history of smoking (`hs`), history of military service (`military`), and asthma (`asthma`)?

8. *Use* ***nhanes.csv.*** Using ggplot2, plot weight (`bmxwt`) and height (`bmxht`), with weight on the y axis. Use color coding for categories of body mass index (`bmicat`), and exclude individuals with missing body mass index from the plot (ie, you should not see an NA category in the legend). You should also reorder the levels of `bmicat`, using scientific ordering, beginning with the largest group (`Class III obesity`). Optionally, send the plot to your colleagues and encourage them to enroll in the course if it is offered again next year `;-)`! Your plot should look like this:



9. *Use* ***unps.csv.*** Add a variable to the data frame for living within 1 km of a road (`dist_road`). The variable should be a factor with 2 levels. Use `table()` to provide a tabulation of the variable.

10. *Use* ***unps.csv.*** Add a variable to the data frame for `wasted` × `stunted`. The variable should be a factor with 4 levels, in the following order: wasting and stunting, wasting only, stunting only, and neither. Use `table()` to provide a tabulation of the variable. Your output should look like this (you need not use the same variable name or the same level names):

```
> table(uu$ws,useNA='always')

Wasting and stunting         Wasting only         Stunting only
                  77                  132                  1552
             Neither                 <NA>
                3251                    0
```