BHSEC Queens
Statistics Spring 2016
Class 12: March 2

Name: _____

LAB 2: EXAMINING COLLEGE DATA

The goal of this lab is to examine college statistics for a sample of four-year colleges that BHSEC alumni now attend and produce a well-written lab report communicating your conclusions.

IMPORT THE DATA

- Go to the course website at `classroom.google.com` and find today's post. The PDF you see is a digital copy of this handout. Download the other two files, `lab2-source-code.R` and `lab2-data.csv` and save them to the Statistics folder. *Do not open either of them!*
- In RStudio, click the "Open" button in the top left (looks like a folder with an arrow) and select the file `lab2-source-code.R`.
- Now, click the "Source" button in the top left window. Let me know if after a minute or so you do NOT get the `>` symbol back in the bottom left.
- Click the "Import Dataset" button in the top right window, and select "from text file."
- Select the file "lab2-data.csv."
- Name it `lab2.data`.
- Make sure under the Heading option that "Yes" is selected and click "Import."
- The data should appear as a table in the top left and also by name in the top right.
- Finally, in the console (bottom left), type `attach(lab2.data)`. (This step means that we can refer to the variables by their names more easily.) NOTE: you will get an error ("The following object is masked…"); that doesn't matter.

REMEMBER: SOURCE, IMPORT, ATTACH.

WRITING LAB REPORTS

- Use the "Open" button in the top left to open the `name-lab2-template.Rmd` file.
- The R Markdown document I've given you is the template for your lab report, into which you will type your code. *It must be in the same folder with your data and source code to work correctly.*
- When working on labs, work on the exercises in the console until you are satisfied with the result and then copy and paste your code into the appropriate place in the template. Then add text analysis in the R Markdown document.
- A line of white space adds a line break.
- Click the "Knit HTML" button to create your lab report as an HTML file. You must submit the HTML file for all labs.
- When you are done, make sure you change the "author" field to include your name and your partner's. Then rename your HTML file to include your last name and first initial, like so: `thomsonv-lab2.html` or `westk-lab2.html`.

VARIABLES:

We will be looking at data from colleges currently attended by BHSEC alumni, including the top 20 most popular colleges among BHSEC graduates. There are 61 colleges in the dataset and all the data is from 2013 ± one year. All data is for the colleges *overall*, not just for BHSEC students who go there. The variables are as follows:

`name`	Name of college/university
`control`	Public or Private
`location`	Urban or Rural (note that suburban is included in Urban)
`in.state`	Total one-year cost for in-state students
`out.of.state`	Total one-year cost for out-of-state students
`percent.admitted`	Acceptance rate
`admission.yield`	What percent of accepted students actually attended?
`tests.required`	Are tests (like SAT or ACT) required for application?
`sat.cr.Q1`	Lower quartile of SAT Critical Reading scores
`sat.math.Q1`	Lower quartile of SAT Math scores
`finaid`	Percentage of students receiving financial aid (including loans)
`enrollment.undergrad`	Number of undergraduate students
`white`	Percent of students who are non-Hispanic white
`women`	Percent of students who are women
`grad.rate.4`	Percent of undergrads who graduate within 4 years.
`grad.rate.6`	Percent of undergrads who graduate within 6 years.

LAB 2 EXERCISES:

WARNING: Each time you open up RStudio, you must SOURCE the source code (including clicking the "source" button, IMPORT the data, and ATTACH the data. Your lab report will not knit unless all the files are in the same folder with their original names.

1. Use a frequency table or bar plot to determine how many schools in this list are private and how many are public. Make an observation in context.
2.
   - Compute the mean, median, and range of out-of-state cost for these colleges.
   - Make a histogram of out-of-state cost of these colleges.
   - Make some observations in context. Why do you think the distribution looks the way it does?
3.
   - Compute the average in-state cost for private colleges and the average in-state cost for public colleges.
   - Compute the standard deviation for both groups using the `sd` command. Which group has greater variation in cost?
   - Make a pair of side-by-side box plots comparing in-state cost for private and public colleges.
   - Do in-state cost and institutional control seem to be associated? Make some observations in context.
4. Do the same for out-of-state costs. Note any similarities and differences.
5. Are public or private colleges in this dataset more likely to make tests like the SAT optional? Answer this question with either a table, a mosaic plot, or a segmented bar plot.
6. Compare the number of undergraduates by type of institutional control using boxplots. In other words, do `num.undergrads` and `inst.control` seem to be associated? Make some observations.

7. Pick a quantitative variable in the dataset. *This question may not be a repeat of a question above.*
   - Examine its shape, center, and spread using appropriate commands.
   - Make a visual (either a histogram or a boxplot).
   - Make some observations in context.
   - Repeat this question for urban schools only. How do the results differ or stay the same?
8. Pick two categorical variables from the dataset to investigate. Why do expect there to be an association between the two variables, or do you expect no association?
   - Make a table of the two variables.
   - Make a barplot or mosaic plot with appropriate title.
   - Make a short observation in context about the two variables.
   - Do these two variables seem to be associated with each other? Explain.
9. Pick a quantitative variable and a categorical variable that you predict are associated. Why do you predict that the two variables are associated with each other.
   - Compute the average and the range of the quantitative variable.
   - Make a side-by-side boxplot or two separate histograms to show the two variables.
   - Make some observations in context.
   - Do these two variables seem to be associated with each other? Explain.
10. Now that you are done, write a short conclusion paragraph describing some of your findings. You should not try to squeeze in the answer to every question, but instead summarize some of what you have done in plain English. You should keep technical language to a minimum, but any claims should be supported by the rest of your lab work, so while you are allowed to speculate on the cause of an association that you found, it should be clearly labeled as speculative or hypothetical.

RSTUDIO REFERENCE: MANIPULATING AND VISUALIZING DATA

Here are the commands from Lab 1, which you may find useful when figuring out Lab 2.

- To find the number of observations in a variable: `length(gender)`.
- One-way table: `table(gender)`.
- Single-variable barplot: `barplot(table(year))`.
- Two categorical variables: `table(gender,year)`, `plot(table(gender,year))`, and `barplot(table(gender,year))`.[1]
- One quantitative variable: `summary(hrs.sleep)`, `range(hrs.sleep)`, `hist(hrs.sleep)`, `sd(hrs.sleep)`.
- Boxplots: `boxplot(hrs.sleep)` and `boxplot(hrs.sleep~gender)`.
- Title: `plot(table(gender,year), main="Year of Stats students according to gender")`
- Axis labels: `boxplot(hrs.sleep~gender, xlab="Gender of students", ylab="Hours of sleep")`
- Subsetting a *whole table:* `females.only <- subset(lab1data, gender == "Female")`. Watch out for the `==`.
- Subsetting a *single variable:* `female.commutes <- subset(commute, gender=="Female")`.

[1] PRO TIP: to get a legend on your segmented barplot, you need the `legend.text` argument: `barplot(table(gender,year), legend.text=TRUE)`.