Problem Set 1

Stats 506, Fall 2018

Due: Monday October 1, 5pm

Instructions

- Submit the assignment by the due date via canvas. If you intend to utilize late days, please send an email to the GSI and cc me *before* the assignment is due. In your email please indicate how many days you intend to use.
- Use Rmarkdown to create and submit a single pdf with your answers to each question along with supporting evidence in the form of tables and graphs.
- All tables and graphs should be neatly labeled and appear polished. Code and object names should not appear in polished tables or graphs.
- For questions 2 and 3 do data manipulation and analyses in separate `.R` files named `ps1_q2.R` and `ps2_q3.R` and then use `source()` to execute these in your Rmarkdown file *from the same directory*. Commands to produce tables and graphs should go in the `Rmd` file.
- You should also submit the `ps1.Rmd` file and any sourced supporting scripts. Please name the files as follows: `ps1.Rmd` , `ps1_q1.sh` , `ps1_q2.R` , `ps1_q3.R` , etc. All files should be executable without errors.
- All files read, sourced, or referred to within scripts should be assumed to be in the same working directory (`./`).
- Your code should be clearly written and it should be possible to assess it by reading it. Use appropriate variable names and comments. Your style will be graded using the style rubric (./StyleRubric.html) [15 points].
- Some of these exercises may require you to use commands or techniques that were not covered in class or in the course notes. You can use the web as needed to identify appropriate approaches. Part of the purpose of these exercises is for you to learn to be resourceful and self sufficient. Questions are welcome at all times, but please make an attempt to locate relevant information yourself first.
- You may wish to review the formulas for standard errors and confidence intervals on page 2 of the document available here (https://open.umich.edu/sites/default/files/downloads/f12-stats250-bgunderson-statsfullyellowcard_0.pdf).

Question 1

In this question you will use command line tools to answer question about the 2015 Residential Energy Consumption Survey (RECS 2015) data set (https://www.eia.gov/consumption/residential/data/2015/index.php?view=microdata).

In addition to your Rmd file, please submit a shell script `ps1_q1.sh` written in *Bash* using the “shebang” `#!/bin/bash`. Your script should assume the file `recs2015_public_v3.csv` is in the same directory and be executable as `bash ps1_q1.sh`.

Part A [5 points; 2.5 each]

In part A, your solution to each question should be a Linux “one-liner”, i.e. a series of one or more commands connected by pipes “|”. Please provide both your solution and the result. Your solution must be written in text so that it can be copied and pasted if needed.

i. How many rows are there for region 3 in the RECS 2015 data set?

ii. Write a one-liner to create a compressed data set containing only the variables: DOEID, NWEIGHT, and BRRWT1-BRRWT96.

Part B [10 points; 5 each]

i. Write a Bash `for` loop to count and print the number of observations within each region.

ii. Produce a file `region_division.txt` providing a sorted list showing unique combinations of values from `REGIONC` and `DIVISION`. Include the contents of that file in your solution. *Hint:* See `man uniq`.

Question 2 [25 pts]

In this question, you will use **R** to answer questions about flights originating in New York City, NY (NYC) in 2013 and 2014. Data for 2013 can be found in the `nycflights2013` **R** package. Data through October 2014 is available here (https://raw.githubusercontent.com/wiki/arunsrinivasan/flights/NYCflights14/flights14.csv). Your answers should be submitted as nicely formatted tables produced using Rmarkdown.

a. Which airlines were responsible for at least 1% of the flights departing any of the three NYC airports between January 1 and October 31, 2013?

b. Among the airlines from part “a”, compare the number and percent of annual flights in the first 10 months of 2013 and the first 10 months of 2014. Your table should include: the airline name (not carrier code), a nicely formatted number (see `format()`), percents for each year with 95% CI, and change in percent with 95% CI. Which airlines showed the largest increase and decrease? Why do some airlines show an increase in the percent of flights but a decrease in the number of flights?

c. Among of the three NYC airports, produce a table showing the percent of flights each airline is responsible for. Limit the table to the airlines identified in part a and include confidence intervals for your estimates. Which airline is the largest carrier at each airport?

Question 3 [45 pts; 15 pts each]

In this question, you will use **R** to answer questions about the RECS 2015 data. You should read the section on computing standard errors available here (https://www.eia.gov/consumption/residential/data/2015/pdf/microdata.pdf). For each question, produce a nicely formatted table and graph to support you answer. In your tables and graphs please provide standard errors for all point estimates.

a. What percent of homes have stucco construction as the *major outside wall material* within each division? Which division has the highest proportion? Which the lowest?

b. What is average total electricity usage in kilowatt hours in each division? Answer the same question stratified by urban and rural status.

c. Which division has the largest disparity between urban and rural areas in terms of the proportion of homes with internet access?