SOC 4930/5050: Lab-01 - Initial Data Cleaning

Christopher Prener, Ph.D.

Fall 2018

Directions

Complete all of the following questions using the data from the `testDriveR` package.[1] Your well-formatted R Notebook source (the `.Rmd` file) and `html` output should be *ready* to be uploaded to your assignments repository by 4:15pm on Monday, September 10th, 2017. We'll go through the submission itself together as a class.

Analysis Development: Create a Project Folder System

1. Using RStudio, add an R Project to a new directory named `Lab-01`. To do this, you will want to go to: File ▷ New Project ▷ New Directory ▷ New Project and save your new folder and R Project to your Desktop or another similar location where you can easily find it.
2. RStudio should automatically open your new project. Verify this by looking up at the righthand corner of RStudio's window - you should see a blue box icon with a dark blue `R` in it. Next to that should be the text `Lab-01`.
3. R Projects set something called the working directory, which is a critically important piece of programming that we'll continue to talk about this semester.
4. In the Files tab on the lower righthand side of RStudio's screen, add a New Folder using the New Folder button right below Files. Name this new folder `docs`.
5. Create a new notebook by going to File ▷ New File ▷ R Notebook. Save it within that `docs/` subdirectory you just created.
6. Edit the heading of your notebook so that it looks like so:

```
---
title: "Lab-02 Notebook"
author: "your name"
date: '(‘r format(Sys.time(), "%B %d, %Y")‘)'
output:
  github_document: default
  html_notebook: default
---
```

[1] See course website for latest details on installing `testDriveR`, which can be found on the Course Software page!

7. Use RMarkdown syntax to create your first assignment notebook! Make sure it has an introductory section, a section for loading packages, a section for loading data, and a section for each part below. These sections should be second-level headings (e.g. `## Introduction`). In Both Part 1 and Part 2, use third level headings to designate question numbers (e.g. `### Question 9`).
8. When you are done, "knit" your document by clicking the Knit button in the toolbar at the top of the notebook.

Part 1: Cleaning Data

Use the `auto17` data frame saved in the `testDriveR` package and make the following changes using "piped" code:

9. Extract observations for German cars (those manufactured by Audi, BMW, Mercedes-Benz, Porsche, and Volkswagon).
10. Keep only the following variables: `id, mfrDivision, carLine, combFE, guzzlerStr, displ`
11. Rename the `mfrDivision` and `combFE` variables.[2]
12. Create a new logical variable that is `TRUE` if the vehicle is a guzzler (`guzzlerStr == "G"`) and is `FALSE` otherwise.
13. Re-order the data frame based on your re-named `combFE` variable from high to low.
14. Print the "head" of the data frame - what is the most fuel efficient German car for sale in the United States for model year 2017?
15. How many German cars in total are for sale in the United States for model year 2017?
16. How many German cars are "gas guzzlers"?

Part 2: Plotting Data

Use the your cleaned German car data to produce the following plots:.

17. Create a bar plot of the logical "gas guzzler" variable you created.
18. Create a histogram of the average fuel efficiency variable.

[2] Not sure what these variables measure? Type `?auto17` into the console of RStudio and scroll through the help file for the data set.

19. Create a scatter plot of the average fuel efficiency and `displ` variables that (a) highlights "gas guzzler" vehicles and (b) uses the "jitter" positions adjustment.

Reminders

Remember that a replication file will be posted on GitHub and linked to from the course website. I will also provide some screen shots of the analysis development section to help you navigate around RStudio's user interface.