

SOC 4930/5050: Problem Set 01 - Initial Data Cleaning

Christopher Prener, Ph.D.

September 4th, 2017

Directions

Complete all of the following questions using the data from the testDriveR package. Your well-formatted R Notebook source (the .Rmd file) and html output should be uploaded to your assignments repository by 4:15PM on Monday, September 18th, 2017.

Part 1: Cleaning Data

Use the gss16 data frame saved in the testDriveR package and make the following changes using “piped” code:

1. Keep only the following variables - ID_, SEX, HRS1, WRKSTAT, INCOME16
2. Rename all of the variables except SEX. These are what each of the other variables refer to:
 - ID_ - identification number
 - HRS1 - number of hours worked last week
 - WRKSTAT - work status (full time, part time, etc.)
 - INCOME16 - income last year, categorized
3. Create a string version of the SEX variable that is equal to “male” when SEX == 1 and is otherwise equal to “female”
4. Remove the original SEX variable
5. How many men and women are in the data set?

Part 2: Plotting Data

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

6. Create a bar plot of the string “sex” variable you created
7. Create a scatter plot of last year’s income by hours worked that (a) only shows data for fulltime workers (when the “work status” variable is equal to 1)¹ and (b) colors points using the “sex” variable you created

¹ Hint: if you use a piped set of functions you can include the appropriate dplyr function for extracting observations before you create your plot