Homework 5–DS 700

Abra Brisbin

12/17/2020

## Problem 1: R as a scientific calculator

### 1a.

Use the c function to make a vector containing the numbers -1, 0, 2, 5, 17. Store your vector in the variable x.

x = c(-1, 0, 2, 5, 17)

Use square brackets [ ] to tell R to display the 3rd element of x.

x[3]

## [1] 2

Display the 6th element of x. What happens?

x[6]

## [1] NA

Try running the following code.

x[-1]

## [1] 0 2 5 17

x[-2]

## [1] -1 2 5 17

x[-3]

## [1] -1 0 5 17

What does the negative sign inside square brackets do?

### 1b.

*Without using c*, make a vector of numbers from -3 to 13, in increments of 4. Store your vector in the variable y, and display the contents of y.

y = seq(-3, 13, by = 4)  
y

## [1] -3 1 5 9 13

Add 5 to each element of y and store the results in the variable z. Display your results.

z = y + 5  
z

## [1] 2 6 10 14 18

Take the square root of each element of y. (Hint: You can Google “square root r” to find the function to take the square root.)

sqrt(y)

## Warning in sqrt(y): NaNs produced

## [1] NaN 1.000000 2.236068 3.000000 3.605551

Add each element of x to the corresponding element of y. Store your results back in the variable y. (This will overwrite the old contents of y.) Display your results.

y = x + y  
y

## [1] -4 1 7 14 30

### 1c.

Run the following code to create a vector in a new way:

my\_subset = 2:4  
my\_subset

## [1] 2 3 4

What does the : do? What are two other ways we could create the same vector?

Use square brackets to extract elements 2, 3, and 4 of x. Your code in the next code cell should not include any numbers.

x[my\_subset]

## [1] 0 2 5

Use square brackets to extract elements 1 and 5 of x. Your code in the next code cell should not include any numbers.

x[-my\_subset]

## [1] -1 17

## Problem 2: Analyzing a data set

This problem will guide you through reading a data set into R and understanding it.

### 2a.

Install the following packages, if you do not already have them installed:

* ggformula
* dplyr
* readr

Remember to do this in the console (lower-left panel of RStudio) or the Packages tab (in the lower-right panel), not in your .Rmd document. Trying to install packages from the .Rmd document can result in an error about “trying to use CRAN without a mirror”.

After installing the packages, use the library function to load them in the code cell below. (Normally, I like to load all my packages at the start of an .Rmd file. That way, if I need to run the .Rmd file on a new computer, it’s easy to see what I need to install.)

library(ggformula)

## Warning: package 'ggformula' was built under R version 4.0.3

## Loading required package: ggplot2

## Loading required package: ggstance

## Warning: package 'ggstance' was built under R version 4.0.3

##   
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':  
##   
## geom\_errorbarh, GeomErrorbarh

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(readr)

Which objects are masked by objects in dplyr? Should we be concerned about these?

### 2b.

Read Wisconsin\_income.csv into R. Call the data frame wi.

wi = read\_csv("C:/Users/brisbia/OneDrive - UW-Eau Claire/CS 491 Data Mining/Wisconsin\_income.csv")

## Parsed with column specification:  
## cols(  
## AGEP = col\_double(),  
## COW = col\_double(),  
## JWMNP = col\_double(),  
## LANX = col\_double(),  
## MAR = col\_double(),  
## SEX = col\_double(),  
## WKHP = col\_double(),  
## DIS = col\_double(),  
## PERNP = col\_double(),  
## Hispanic = col\_double(),  
## Stem = col\_character(),  
## Education = col\_double(),  
## Region = col\_character(),  
## CIT2 = col\_double(),  
## RAC = col\_double()  
## )

Which variables have been read in as categorical?

Use a text editor (such as Notepad++) to open the data dictionary. Which variables should be categorical, but have been read in as numeric?

Tell R to treat those variables as categorical.

wi <- wi %>%  
 mutate(COW = factor(COW),  
 LANX = factor(LANX),  
 MAR = factor(MAR),  
 SEX = factor(SEX),  
 DIS = factor(DIS),  
 RAC = factor(RAC),  
 CIT2 = factor(CIT2),  
 Hispanic = factor(Hispanic))

### 2c.

Sometimes you’ll need to create your own code cells. Start a code cell below by typing 3 backticks (on the key to the left of the number 1) and then {r}. End the cell on a new line by typing 3 backticks again.

dim(wi)

## [1] 13243 15

Inside the cell you created, write a line of code to display the number of rows and columns of the data.

In the code cell below, write code to display the first 6 rows of data.

head(wi)

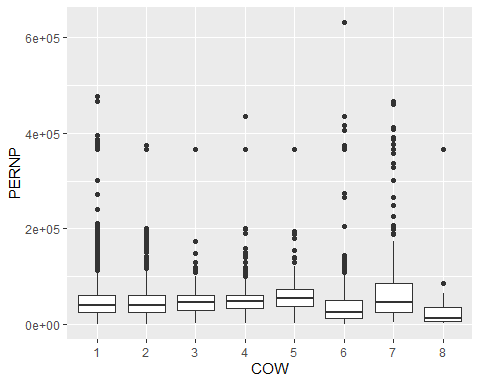
## # A tibble: 6 x 15  
## AGEP COW JWMNP LANX MAR SEX WKHP DIS PERNP Hispanic Stem Education  
## <dbl> <fct> <dbl> <fct> <fct> <fct> <dbl> <fct> <dbl> <fct> <chr> <dbl>  
## 1 34 1 15 2 3 2 40 2 28000 0 NoDeg 14  
## 2 46 1 15 2 1 2 45 2 65000 0 No 16  
## 3 42 1 3 2 1 1 45 2 52000 0 NoDeg 12  
## 4 34 1 6 2 5 1 37 2 22000 0 NoDeg 12  
## 5 33 1 20 2 3 2 40 2 27500 0 NoDeg 13  
## 6 33 1 5 2 3 2 40 2 29200 1 NoDeg 12  
## # ... with 3 more variables: Region <chr>, CIT2 <fct>, RAC <fct>

Use the first 6 lines of data and the data dictionary to help you answer the question: In this data set, what does 1 observational unit represent?

### 2d.

Make a boxplot of earnings, grouped by class of worker.

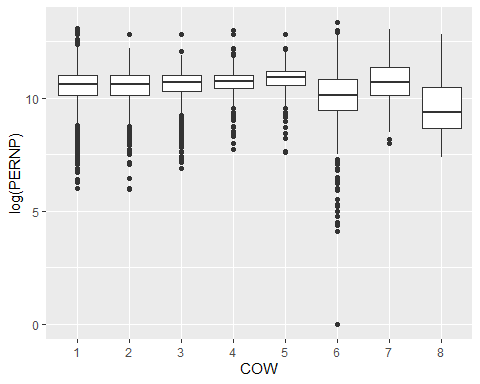
gf\_boxplot(PERNP ~ COW, data = wi)



When most observations of a variable are clustered around a low value, and most of the outliers are on the high end of the distribution, we say the variable is “right-skewed”. When a variable is right-skewed, it can sometimes be helpful to reduce the skew by taking the log.

Run the following code cell to make a boxplot of log(earnings), grouped by class of worker.

gf\_boxplot(log(PERNP) ~ COW, data = wi)



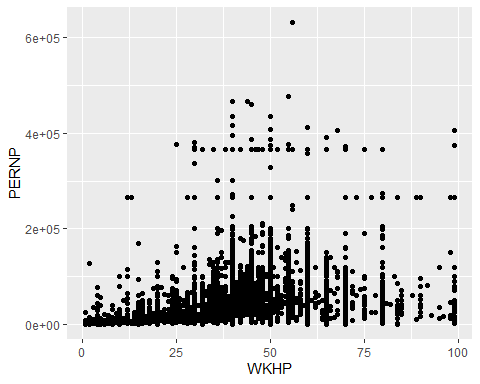
Which class of worker has the lowest-paid person in the data set?

Which class of worker has the lowest median income? (Note that in a boxplot, the median is denoted by the thick line in the middle of the box.)

## 2e.

Make a scatterplot of earnings as a function of hours worked per week.

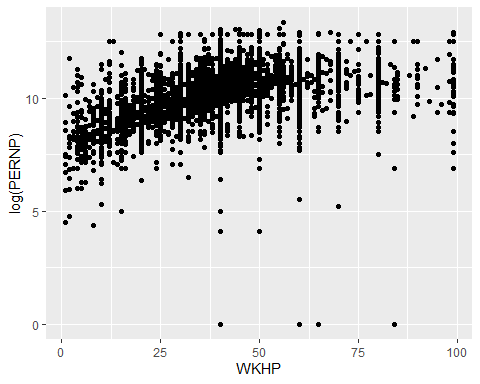
gf\_point(PERNP ~ WKHP, data = wi)



Comment on the relationship between the variables.

Let’s see if we can get a more informative graph by reducing the skew in earnings. Make a scatterplot of the relationship between log(earnings) and hours worked per week.

gf\_point(log(PERNP) ~ WKHP, data = wi)



Describe the relationship between the variables.

Add a line of best fit to the graph. Is its slope positive or negative? What does this tell you, in the context of this data set?

gf\_point(log(PERNP) ~ WKHP, data = wi) %>%  
 gf\_smooth(log(PERNP) ~ WKHP, method = "lm", data = wi)

