GCDS Homework #01

Due: 09-08-2022

Table of Contents

[1 Load the readr and ggplot2 libraries. 1](#_Toc113027402)

[2 What is the main functionality of readr? 1](#_Toc113027403)

[3 How to read a file 1](#_Toc113027404)

[4 Load a local data frame 2](#_Toc113027405)

[5 Examine a data frame 2](#_Toc113027406)

[6 Create a plot 2](#_Toc113027407)

[7 Modify code 2](#_Toc113027408)

[8 Write data to disk 3](#_Toc113027409)

[9 Interpreting data 4](#_Toc113027410)

[10 Thinking of regression 4](#_Toc113027411)

For this homework, you will create an R Markdown document file that will output as an HTML file. Name it **“GCDS\_HW\_01\_yourname”** and add your name to it. The homework will need to be submitted as a knit HTML file.

You should create a separate code block in R Markdown for each of the problems. Type text responses as text and code within code blocks/chunks.

**Questions:**

# 1 Load the readr and ggplot2 libraries.

# 2 What is the main functionality of readr?

# 3 How to read a file

read\_csv() requires at least one argument. What is that argument? Is the type of that object you pass to that argument a numeric object or a string/character object?

# 4 Load a local data frame

Load the pressure data set that is part of base R (hint: we did this with cars and mtcars data).

# 5 Examine a data frame

Use the head function to look at the first several cases/observations of the pressure data.

# 6 Create a plot

Determine what the variable names/columns are and plot any two variables as a scatterplot. You do not know how to plot a scatterplot, so I’ll give you some code to help you. You should remember from previous courses that scatter plots plot the x,y coordinates for data points.

The following code uses the ggplot2 library. If you run this code as is, you will get an error because no arguments are specified. Fix the code so it executes.

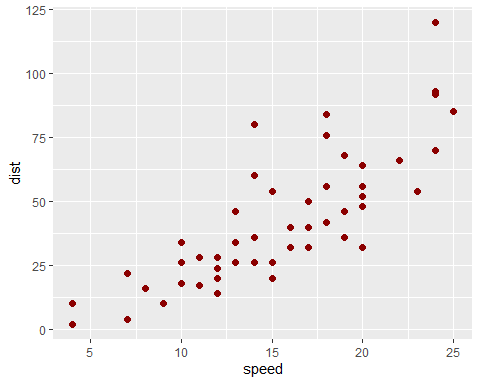
ggplot2::ggplot(pressure, aes(x = ?, y = ?)) + geom\_point()

# 7 Modify code

Take a look at the following code and the plot it produces. Modify the code below to plot green circles for the xy points.

Also, make a second plot that swaps the x and y variables, changes the size of the points, and plots a square instead of a circle for each data point.

ggplot(cars, aes(x = speed, y = dist)) +  
 geom\_point(shape = "circle",   
 color = "darkred",   
 size = 2  
 )



# 8 Write data to disk

Use write\_csv() to write a data frame with file extension .csv (to your to your local “GCDS/data” directory on your computer). You will see that we can use data.frame() to create a silly data frame, which you can save. When saving data files, remember that you need to specify both the *data frame* object and the *name of the file* as arguments. Importantly, the file name you chose will also need to include the *path* to the subdirectory in which you need to save the file. In other words, you’ll need **directorypath/filename + file extension**

If you query R using ?readr::write\_csv you’ll see an example at the bottom on the help page. Note, however, that there is no subdirectory in that function call.

write\_csv(mtcars, "mtcars.csv")

Check also whether you need to pass the data frame object or the file name first. Remember that R will by default save to your *working directory*. We set this the other day as “GCDS”. You may wish to see whether your working directory includes a “/” at the end of the name by checking what your working directory by checking getwd(). The “/” is used to separate directories.

Data Frame:

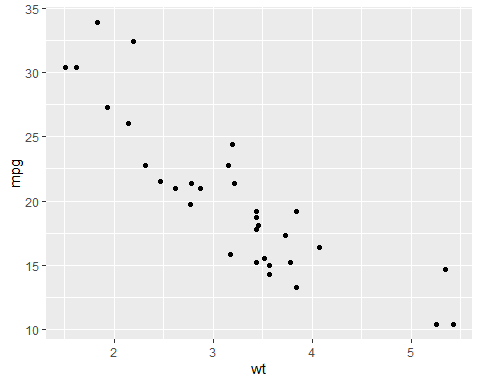
my\_first\_data\_frame <- data.frame(  
 student = c("Bill", "Sally", "Tanya"), # this is a string vector with 3 elements`   
 age = c(22, 21, 19)  
)

If you wanted to write this data frame to your “data” directory as a .csv file, what would your code look like? Type it as text below.

# 9 Interpreting data

Based on the following plot, what can you tell about the slope?

library(magrittr) # loading magrittr for using %>%  
  
mtcars %>%  
 ggplot(., aes(x = wt, y = mpg)) +  
 geom\_point()



# 10 Thinking of regression

Notice how changing the *aesthetics*, aes() in the following code illustrates how cars that have 4, 6, or 8 cylinder engines are represented differently in the plot. Do you believe a single statistical model might fit the data if broken out by cylinder size? Why or why not? Hint: Consider mentally fitting regression lines through the separate points.

Type your answer as text.

mtcars %>%  
 dplyr::mutate(cyl = as.factor(cyl)) %>% # changing cylinder var to a factor   
 ggplot(., aes(x = wt, # x variable  
 y = mpg)) + # y variable  
 geom\_point(aes(color = cyl)) # make geom\_point colors = cyl variable

