DS5110 Homework 1 - Due Jan. 22

Kylie Ariel Bemis

1/11/2019

Instructions

Create a directory with the following structure:

- `hw1-your-name/hw1-your-name.Rmd`
- `hw1-your-name/hw1-your-name.pdf`

where `hw1-your-name.Rmd` is an R Markdown file that compiles to create `hw1-your-name.pdf`.

Do not include data in the directory. Compress the directory as `.zip`.

Your solution should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. Missing values and overplotting should be handled appropriately. Axes should be labeled clearly and accurately.

To submit your solution, create a new private post of type "Note" on Piazza, select "Individual Student(s) / Instructor(s)" and type "Instructors", select the folder "hw1", go to Insert->Insert file in the Rich Text Editor, upload your `.zip` homework solution. Title your note "[hw1 solutions] - your name" and post the private note to Piazza. **Be sure to post it only to instructors**

Part A

Problems 1–2 ask you to write some basic R functions that may be useful for data manipulation and visualization. You may need to review commonly-used base R functions from the "Vocabulary" chapter of the *Advanced R* textbook.

Problem 1

Write a function of the following form:

`selectCols(data, ...)`

- `data` A `data.frame` to subset by column
- `...` Additional arguments giving the names or indices of columns as strings or integers, respectively.

The function should return a new `data.frame` (never a vector) that has the selected columns.

Test it on the `mpg` dataset. (You do not need to handle errors or exceptions.)

Hint: You can use `list(...)` to turn the `...` arguments into a list that can be more easily manipulated.

Problem 2

Write a function of the following form:

`plotCols(data)`

- `data` A `data.frame` to plot each column

The function should loop through each column of the `data.frame` and plot the distribution of each variable. If it's a continuous variable (numeric), create a histogram. If it's a categorical variable (character or factor), create a bar plot.

Test it on the `mpg` dataset. (You do not need to include the plot output.)

Hint: You can use `aes_string()` to map aesthetics when the variable name is given as a string. You may need to use `print(g)` to print the plot inside a loop, where `g` is the result of a call to `ggplot()`.

Part B

Problems 3–5 use the `diamonds` dataset from the `ggplot2` package, which includes the prices of almost 54,000 round cut diamonds.

Problem 3

Use side-by-side boxplots to visualize the distribution of `price` for each level of `color`. What do you notice about the relationship between price and color? Does it make sense?

(Check the documentation for the dataset to help understand the levels of diamond color.)

Problem 4

Use side-by-side boxplots to visualize the distribution of `carat` for each level of `color`. What do you notice about the relationship between carat and color? Could this help make sense of the previous plot?

Problem 5

Create a scatter plot of `carat` versus `price`, using either an additional aesthetic or faceting to visualize the relationship between carat and price for each level of `color`. Overlay smooth lines for each level of color.

Comment on what you notice about the relationship between carat, price, and color.