Solutions Manual for Modern Data Science with R, 2e by Benjamin Baumer, Daniel Kaplan, Nicholas Horton (All Chapters)

Chapter 1

Preface

This document includes sample solutions for exercises in the second edition of *Modern Data Science with R*.

Please do not share these solutions online.

Note that the R environment is changing quickly and packages are frequently updated. Instructors should verify that their code works with their setup.

We have marked the exercises as Easy, Medium, or Hard to help distinguish the degree of difficulty. The exercises have been ordered by difficulty level. These are quite arbitrary designations that may be appropriate or not depending on audience.

Additional instructor materials (including freely downloadable sample chapters and related resources) can be found at http://mdsr-book.github.io.

We welcome suggestions and improvements in these solutions.

To be informed about updates please reach out to Nicholas Horton (nicholasjhorton@gmail.com).

chapter	Easy	Hard	Medium	N
data-I	6	1	8	15
appR	11	NA	2	13
reproducible	10	NA	2	12
text	6	1	5	12
data-II	3	NA	8	11
sqlI	4	1	6	11
dataviz-II	2	1	7	10
dataviz-III	1	2	7	10
ethics	1	1	8	10
simulation	NA	NA	9	9
algorithmic	1	1	6	8
sqlII	1	5	2	8
foundations	2	NA	5	7
learning-II	NA	3	4	7
dataviz-I	1	NA	5	6
iteration	2	1	3	6
join	1	1	4	6
learning-I	1	1	4	6
modeling	2	1	3	6
regression	1	NA	4	5
spatial-I	1	3	1	5
spatial-II	NA	2	2	4
netsci	NA	2	1	3
not working	NA	NA	1	1

Here is the summary of the online only exercises:

chapter	Easy	Medium	Hard	N
data-I	6	NA	NA	6
dataviz-I	3	2	NA	5
dataviz-II	4	NA	1	5
join	3	NA	1	4
algorithmic	2	1	NA	3
data-II	3	NA	NA	3
text	NA	2	1	3
dataI	1	NA	1	2
foundations	NA	2	NA	2
simulation	NA	1	1	2
sqlII	1	NA	1	2
dataviz-III	NA	1	NA	1
ethics	NA	1	NA	1
not-working	NA	NA	1	1
regression	NA	1	NA	1
spatial	NA	NA	1	1
sqlI	1	NA	NA	1

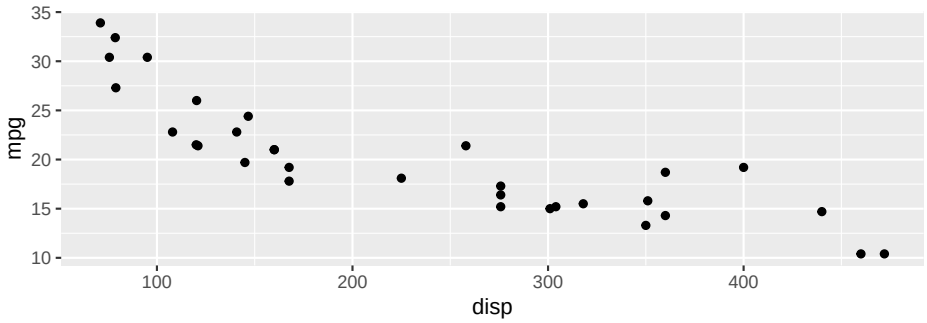
Chapter 2

Data visualization

2.1 Exercises

Problem 1 (Easy): Consider the following data graphic.

```
ggplot(mtcars, aes(x = disp, y = mpg)) +
  geom_point()
```



The `am` variable takes the value `0` if the car has automatic transmission and `1` if the car has manual transmission. How could you differentiate the cars in the graphic based on their transmission type?

SOLUTION:

Map the `color` aesthetic to the `am` variable.

Problem 2 (Medium): Pick one of the Science Notebook entries at https://www.edwardtufte.com/tufte (e.g., "Making better inferences from statistical graphics"). Write a brief reflection on the graphical principles that are illustrated by this entry.

SOLUTION:

Answers will vary.

Problem 3 (Medium): Find two graphs published in a newspaper or on the internet in the last two years.

a. Identify a graphical display that you find compelling. What aspects of the display work well, and how do these relate to the principles established in this chapter? Include a screen shot of the display along with your solution.

b. Identify a graphical display that you find less than compelling. What aspects of the display don't work well? Are there ways that the display might be improved? Include a screen shot of the display along with your solution.

SOLUTION:

Answers will vary.

Problem 4 (Medium): Find two scientific papers from the last two years in a peer-reviewed journal (*Nature* and *Science* are good choices).

a. Identify a graphical display that you find compelling. What aspects of the display work well, and how do these relate to the principles established in this chapter? Include a screen shot of the display along with your solution.

b. Identify a graphical display that you find less than compelling. What aspects of the display don't work well? Are there ways that the display might be improved? Include a screen shot of the display along with your solution.

SOLUTION:

Answers will vary.

Problem 5 (Medium): Consider the two graphics related to *The New York Times* "Taxmageddon" article at http://www.nytimes.com/2012/04/15/sunday-review/coming-soon-taxmageddon.html. The first is "Whose Tax Rates Rose or Fell" and the second is "Who Gains Most From Tax Breaks."

a. Examine the two graphics carefully. Discuss what you think they convey. What story do the graphics tell?
b. Evaluate both graphics in terms of the taxonomy described in this chapter. Are the scales appropriate? Consistent? Clearly labeled? Do variable dimensions exceed data dimensions?
c. What, if anything, is misleading about these graphics?

SOLUTION:

a. Answers will vary. The main take-aways are that tax rates on the rich have fallen, while tax rates on the poor have not. Moreover, in terms of the "cost" to the US Treasury, most tax breaks go to those in the top 20% in income.
b. Answers will vary. In the first graphic, the y-scale is the tax rate, but there is no axis. The labels about the segment of the income distribution are not reflected in any quantity. The color scheme is not immediately helpful. The change in real pre-tax income figures to not have a corresponding quantitative representation. In the second graphic, the y-scale is very confusing, and possibly misleading. The scale is a percentage, and yet the income amounts on the left are *not* on a linear scale.
c. Answers will vary. Most obviously, the y-scales in either data graphic.

Problem 6 (Medium): Consider the data graphic http://tinyurl.com/nytimes-unplanned about birth control methods.

a. What quantity is being shown on the y-axis of each plot?
b. List the variables displayed in the data graphic, along with the units and a few typical values for each.
c. List the visual cues used in the data graphic and explain how each visual cue is linked to each variable.
d. Examine the graphic carefully. Describe, in words, what *information* you think the data graphic conveys. Do not just summarize the *data*—interpret the data in the context of the problem and tell us what it means. (Note: *information* is meaningful to human beings—it is not the same thing as *data*.)

SOLUTION:

- The number of women out of 100 who will have an unplanned preganancy.
- Variables:
  - `years`: units of years: 1, 2, 3, ...
  - `number_of_women`: units of people: 1, 2, 3, ...
  - `method`: categorical: spermicides, etc.
  - `use_type`: categorical: typical, perfect
- Mappings:
  - horizontal position is mapped to `year`
  - vertical position is mapped to `number_of_women`
  - facets are mapped to `method`
  - lines/shapes are mapped to `use_type`
  - color/shade is mapped to `number_of_women`
- Most any birth control methods will likely fail at least once with typical use over a 10 year period. However, Depo-Provera and the pill are the most effective, especially with perfect use.