Working Through G. Grolemund & H. Wickhams's R for Data Science

Krista DeStasio

7/11/2017

Contents

A Brief Introduction to This File

This R file walks through G. Grolemund & H. Wickhams's online text, "R for Data Science." The code is commented so that the beginning R programmer can read through and understand what each line of code does and compare it to their own as they work through the text. Of course, there is more than one way to write code. This is only one sample way among many, and surely not the *most* elegant. **For the greatest learning benefit, I suggest you attempt each exercise on your own before looking at the code or write-ups provided here.**

For those new to R and RStudio, it may be of additional benefit to knit the document and examine how the code in the Rmd file is visually expressed in the resultant knitted document. For example, see how the `["R for Data Science."](http://r4ds.had.co.nz/index.html)` is expressed as a hyperlink in the preceeding paragraph where it was not surrounded by tick-marks and compare that to how the same text is expressed in this paragraph when surrounded by ticks. See also the difference in appearance when knitting to different document types (HTML, PDF, Word).

Tip: *If you are using RStudio, click the text next to the orange # box at the bottom of the editor window to easily navigate the code chunks.*

Tip: *Use the* `?` *before any command to view the documentation on that function. Do this often. For example, type* `?setwd` *to see a description, usage, arguments, and more for the function* `setwd()`*.*

Begin Work-through

Chapter 3, Data Visualisation

To really understand ggplot2, I highly recommend reading "The Layered Grammar of Graphics" as suggested at the beginning of Chapter 3.

The mpg data frame

```
str(mpg) # Look at the structure of the mpg data frame
```

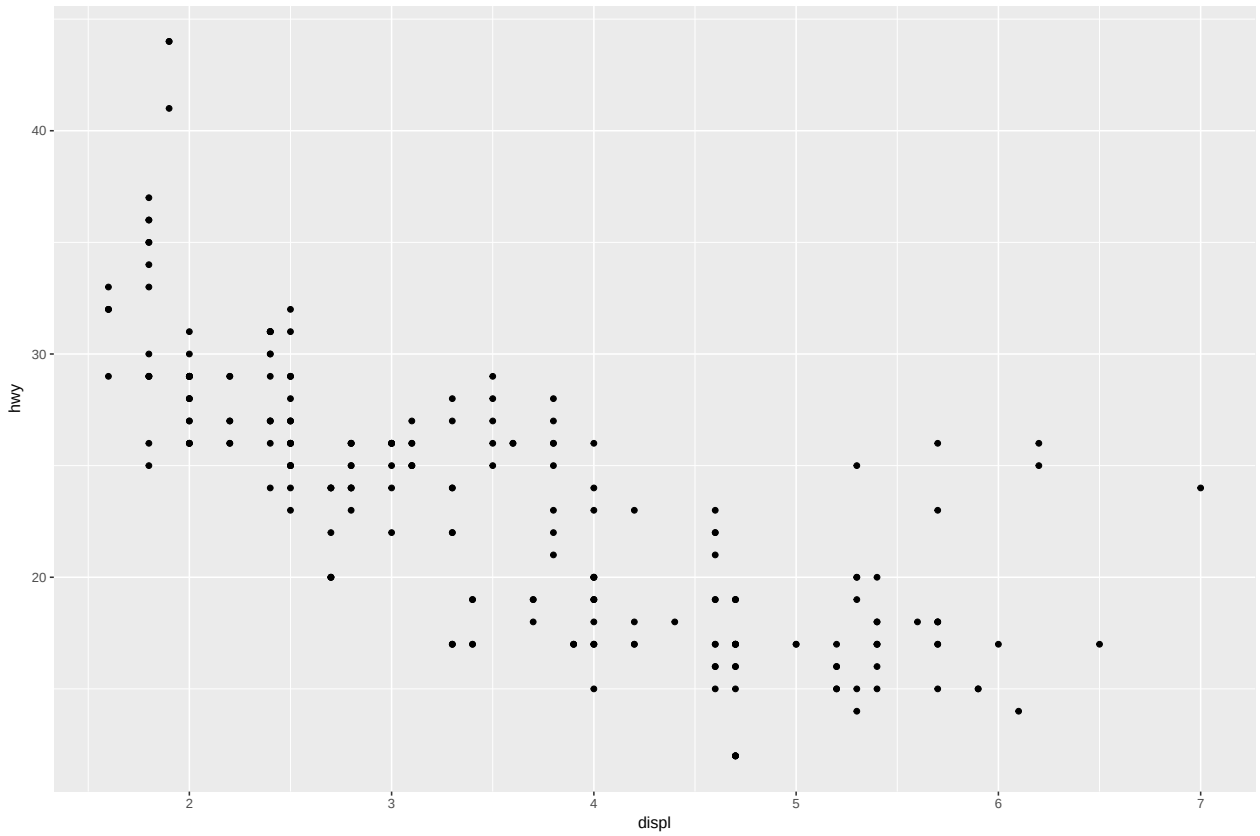
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

```
mpg # Look at the first 10 rows of the mpg data frame
```

```
## # A tibble: 234 x 11
##    manufacturer      model displ  year   cyl      trans   drv   cty   hwy
##           <chr>      <chr> <dbl> <int> <int>      <chr> <chr> <int> <int>
##  1         audi         a4   1.8  1999     4   auto(l5)     f    18    29
##  2         audi         a4   1.8  1999     4 manual(m5)     f    21    29
##  3         audi         a4   2.0  2008     4 manual(m6)     f    20    31
##  4         audi         a4   2.0  2008     4   auto(av)     f    21    30
##  5         audi         a4   2.8  1999     6   auto(l5)     f    16    26
##  6         audi         a4   2.8  1999     6 manual(m5)     f    18    26
##  7         audi         a4   3.1  2008     6   auto(av)     f    18    27
##  8         audi a4 quattro   1.8  1999     4 manual(m5)     4    18    26
##  9         audi a4 quattro   1.8  1999     4   auto(l5)     4    16    25
## 10         audi a4 quattro   2.0  2008     4 manual(m6)     4    20    28
## # ... with 224 more rows, and 2 more variables: fl <chr>, class <chr>
```

Hypothesis: There is a negative linear relationship between engine size and fuel efficiency, such that as engine size increases fuel efficiency decreases.

```
ggplot(data=mpg) + # specify data frame
    geom_point(mapping = aes(x = displ, y = hwy)) # specify that plot is a scatterplot with displ on th
```



The plot confirms the hypothesis that there is a negative relationship between engine size and fuel efficiency.

Template for making graphs with ggplot2

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

Exercises

1. There are no visible results from the code below.

```
ggplot(data = mpg)
```

2. Based on the output from `str(mpg)`, we see that there are 234 rows and 11 columns in the mpg data frame.

```
# Alternative means of finding number of rows and columns
nrow(mpg) # Pring the number of rows
```

```
## [1] 234
```

```
ncol(mpg)
```

```
## [1] 11
```

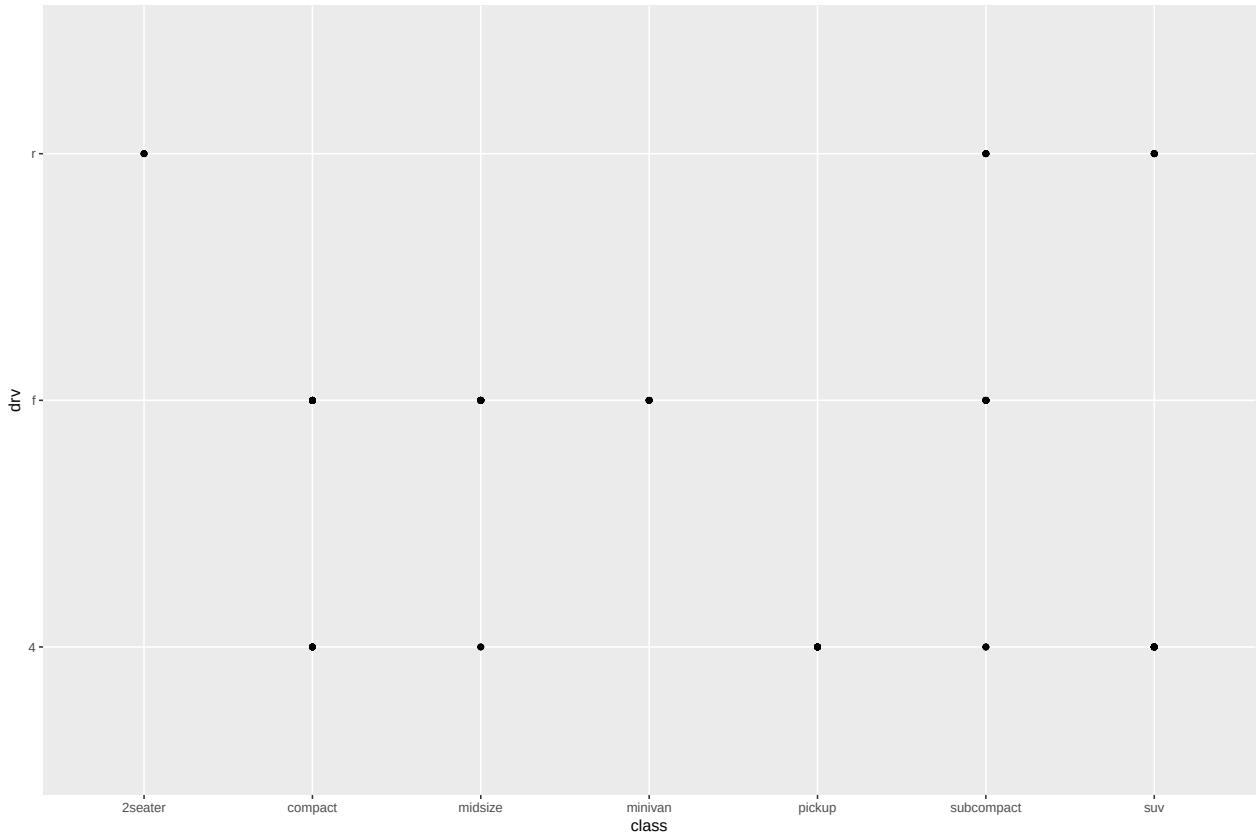
There are 234 rows and 11 columns in the mpg data frame.

3. The `drv` variable describes whether the vehicle is front, rear, or 4-wheel drive.

```
?mpg
```

4. The plot is not useful because the variables are categorical and multiple points are plotted atop one another. We are unable to determine from this plot how many observations there are of each class-drive combination.

```
ggplot(data=mpg) +
    geom_point(mapping = aes(x=class, y=drv))
```


drv
r
f
4
2seater
compact
midsize
minivan
pickup
subcompact
suv
class