R_Code_for_3_1

Echo

8/26/2020

3.1 Introduction

This chapter will teach you how to visualize your data using **ggplot2**.

3.1.1 Prerequisites

This section will focus on ggplot2, one of the core members of the 'tidyverse' package. Firstly, we need to load the *tidyverse* by running this code.

```r
rm(list=ls()) # clear all workspace
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2      v purrr   0.3.3
## v tibble  3.0.3      v dplyr   0.8.5
## v tidyr   1.0.2      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Note if we need to explict about where a function (or dataset) comes from, we will use the special form *package::function*. For instance, *ggplot2::ggplot* tells you clearly that we are using the **ggplot** function from the ggplot2 package.

3.2 First steps

Let's use the first graph to answer a question: Do cars with big engines use more fuel than cars with small engines? Try to make your answer precise. What is the relationship between engine size and fuel efficiency look like? Is it positive? Negative? Linear? Nonlinear?

3.2.1 The mpg data frame

You can test your answer with the *mpg* data frame in the ggplot2 (aka *ggplot2::mpg*). A data frame is a rectangular collection of variables (in the columns) and observations (in the rows). *mpg* contains observations collected by the US Environmental Protection Agency on 38 models of car.

```
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model   displ  year   cyl trans   drv    cty   hwy fl    class
##    <chr>        <chr>   <dbl> <int> <int> <chr>   <chr> <int> <int> <chr> <chr>
##  1 audi         a4        1.8  1999     4 auto(l~ f        18    29 p     comp~
##  2 audi         a4        1.8  1999     4 manual~ f        21    29 p     comp~
##  3 audi         a4        2    2008     4 manual~ f        20    31 p     comp~
##  4 audi         a4        2    2008     4 auto(a~ f        21    30 p     comp~
##  5 audi         a4        2.8  1999     6 auto(l~ f        16    26 p     comp~
##  6 audi         a4        2.8  1999     6 manual~ f        18    26 p     comp~
##  7 audi         a4        3.1  2008     6 auto(a~ f        18    27 p     comp~
##  8 audi         a4 quat~  1.8  1999     4 manual~ 4        18    26 p     comp~
##  9 audi         a4 quat~  1.8  1999     4 auto(l~ 4        16    25 p     comp~
## 10 audi         a4 quat~  2    2008     4 manual~ 4        20    28 p     comp~
## # ... with 224 more rows
```
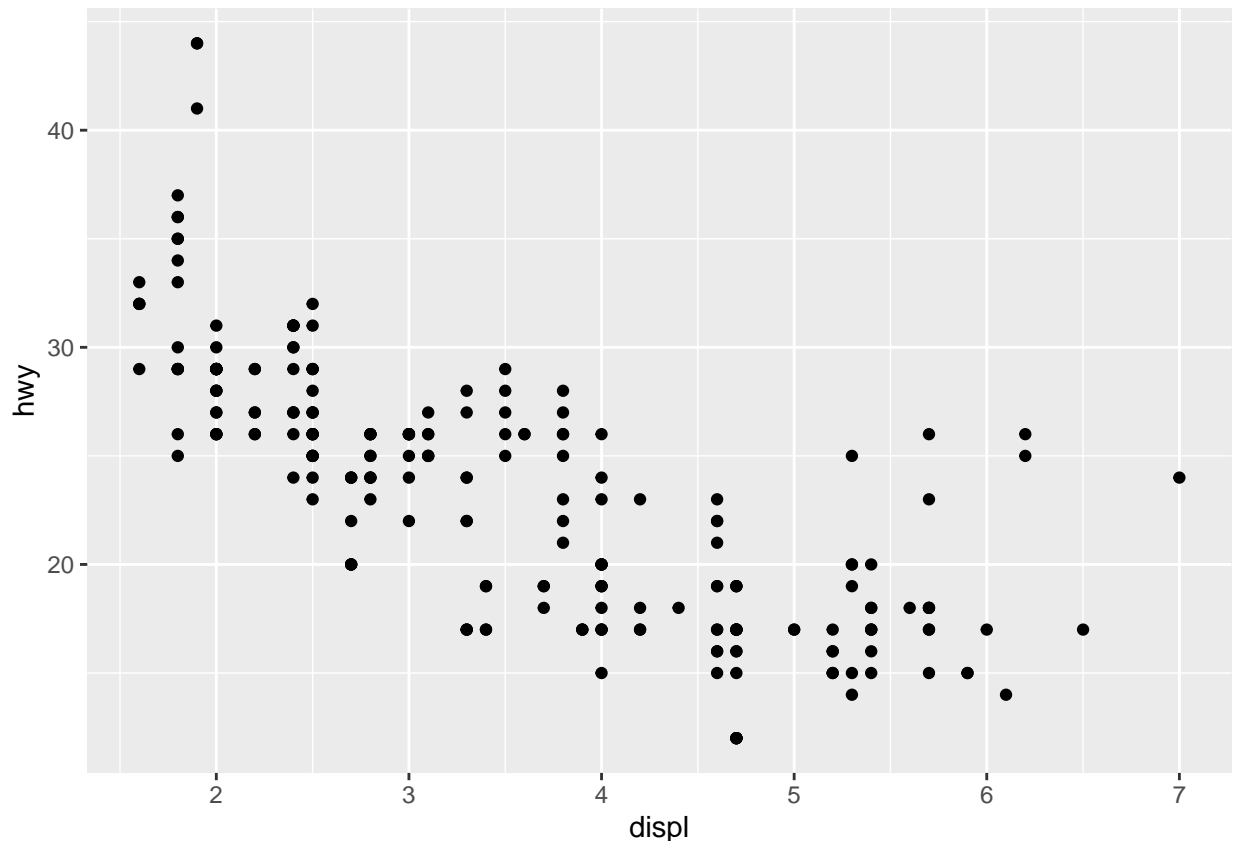
Among the variables in *mpg* are:

1. *displ*, a car's engine size, in litres
2. *hwy*, a car's fuel efficiency on the highway, in miles per gallon (mpg).

3.2.2 Creating a ggplot

To plot *mpg*, run this code to put *displ* on the x-axis and *hwy* on the y-axis:

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y =hwy))
```

The plot clearly shows a negative relationship between engine size *displ* and fuel efficiency *hwy*, i.e. cars with big engines use more fuel.

You complete your graph by adding one or more layers to *ggplot*. The function *geom_point* adds a layer of points to your plot, which creats a scatterplot. Note: *ggplot2* comes with many geom functions that each add a different type of layer to a plot.

Each *geom* function in *ggplot2* takes a *mapping* argument. This defines how variables in your dataset are mapped to visual properties. The *mapping* argument is always paired with *aes*, and the x and y arguments of *aes* specify which variables to map to the x and y axes.

3.2.3 A graphing template

Let's turn this code into a reusable *template* for making graphs with ggplot2.

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

3.2.4 Exercises

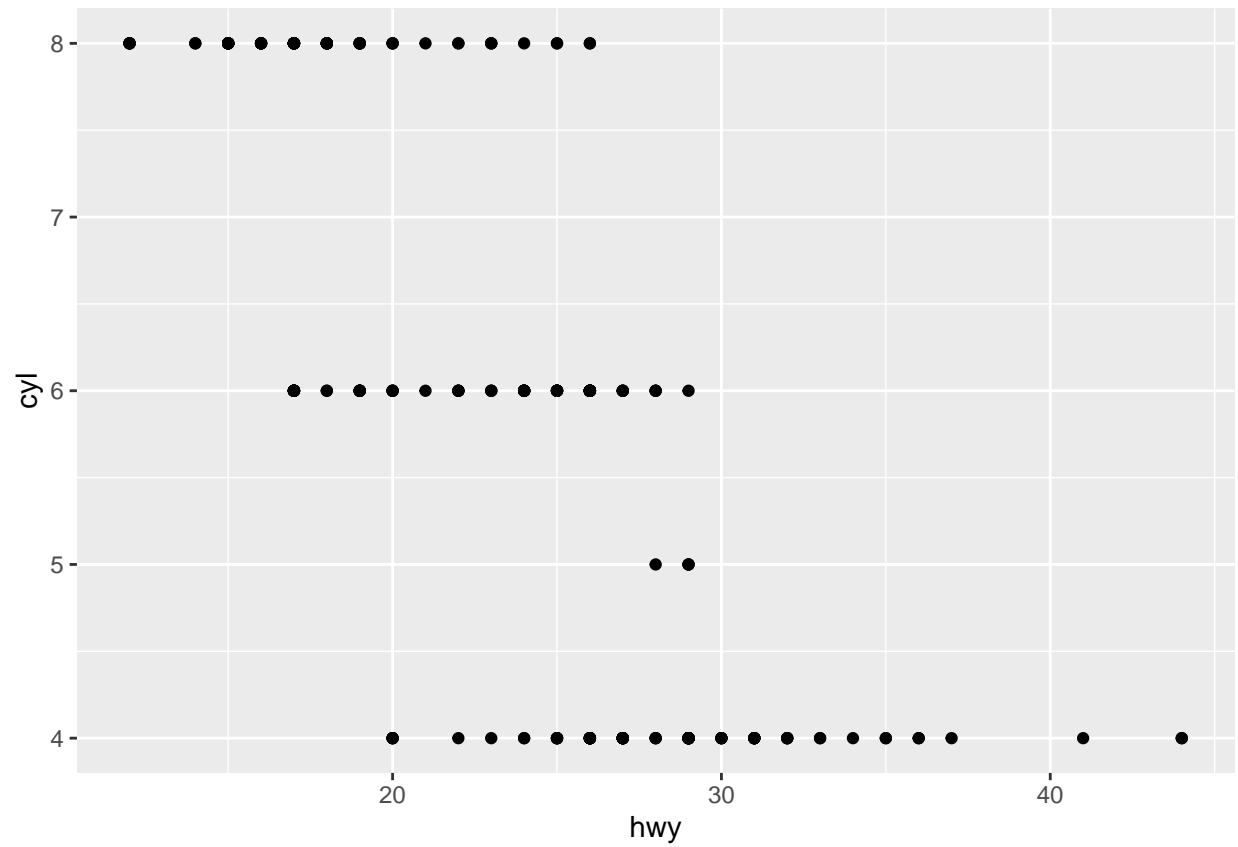1). Run the *ggplot(data = mpg)*. What do you see?

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

I see a blank graph

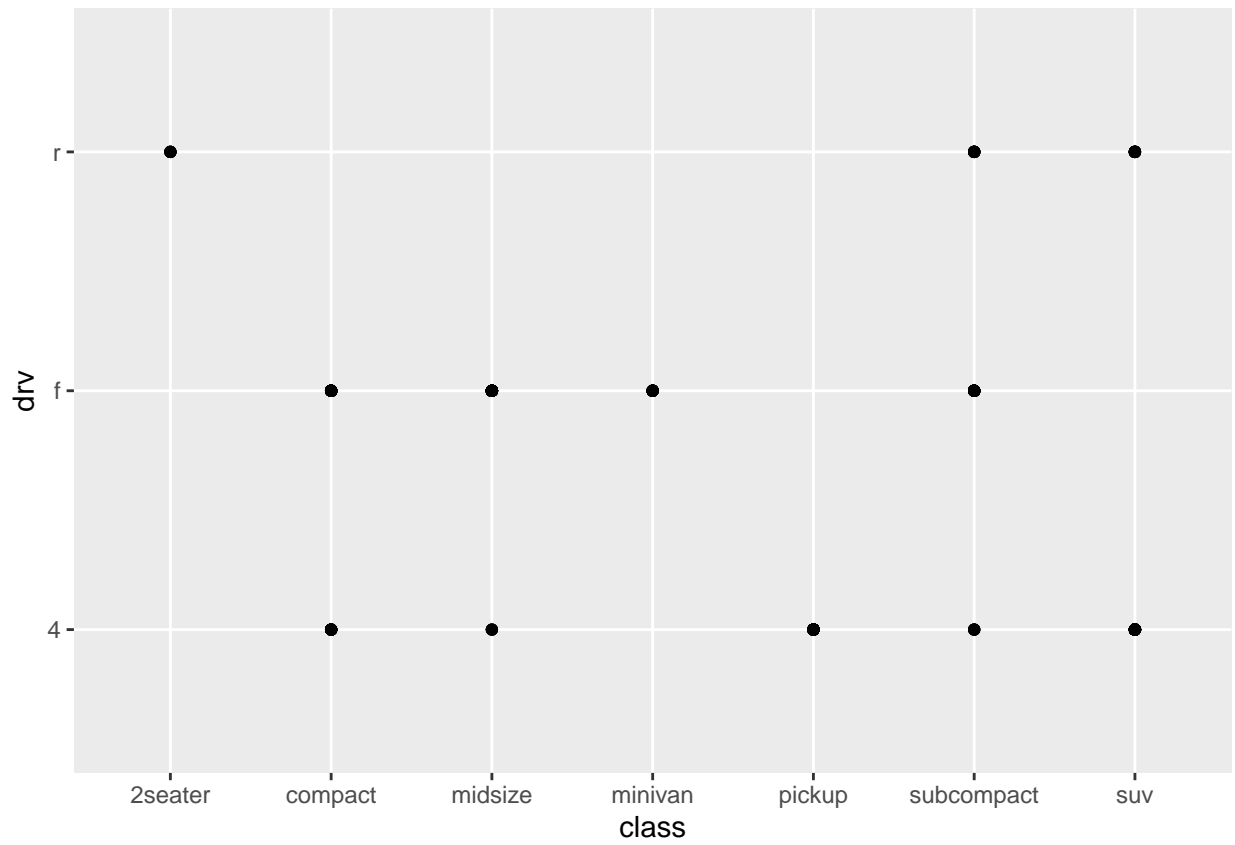**2). How many rows are in _mpg_ and how many rows?**

3). Make a scatterplot of _hwy_ vs _cyl_.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = hwy, y = cyl))
```

4). What happens if you make a scatterplot of *class* and *drv* ? Why is the plot not useful?

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

It didn't show any changing trend w.r.t the variables.