COMPSCIX 415.2 Homework 2

Sanatan Das

February 11, 2018

Load tidyverse package

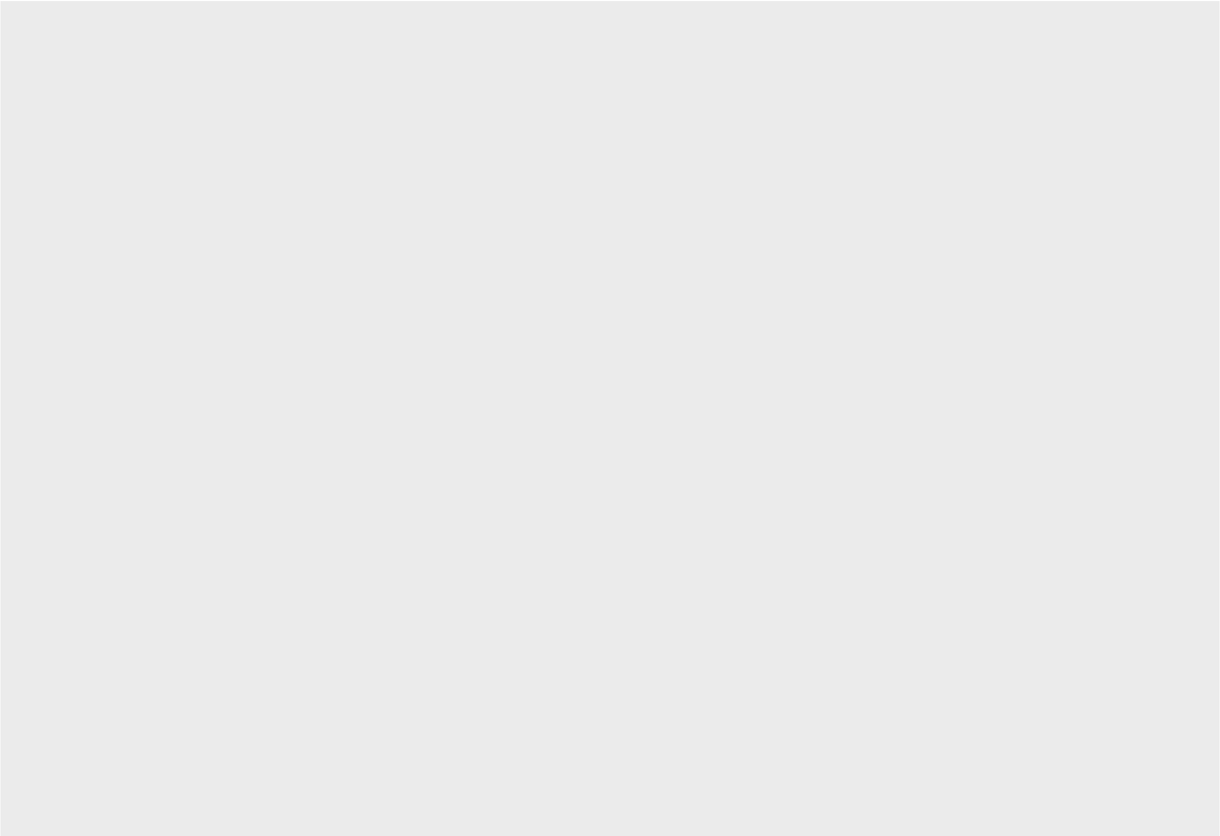
```
library(tidyverse)
```

3.2.4 Exercises

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

ANSWER 1: ggplot(data = mpg) just creates the coordinate system. Layers need to be added to it to visualize graphs.

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



QUESTION 2. How many rows are in mpg? How many columns?

ANSWER 2: Number of rows - 234, Number of columns - 11

```
glimpse(mpg)
```

```
## Observations: 234
## Variables: 11
## $ manufacturer <chr> "audi", "audi", "audi", "audi", "audi", "audi", "...
## $ model        <chr> "a4", "a4", "a4", "a4", "a4", "a4", "a4", "a4 qua...
## $ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1.8, 1.8, 2.0,...
## $ year         <int> 1999, 1999, 2008, 2008, 1999, 1999, 2008, 1999, 1...
## $ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6...
## $ trans        <chr> "auto(l5)", "manual(m5)", "manual(m6)", "auto(av)...
## $ drv          <chr> "f", "f", "f", "f", "f", "f", "f", "4", "4", "4",...
## $ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20, 19, 15, 1...
## $ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28, 27, 25, 2...
## $ fl           <chr> "p", "p", "p", "p", "p", "p", "p", "p", "p", "p",...
## $ class        <chr> "compact", "compact", "compact", "compact", "comp...
```

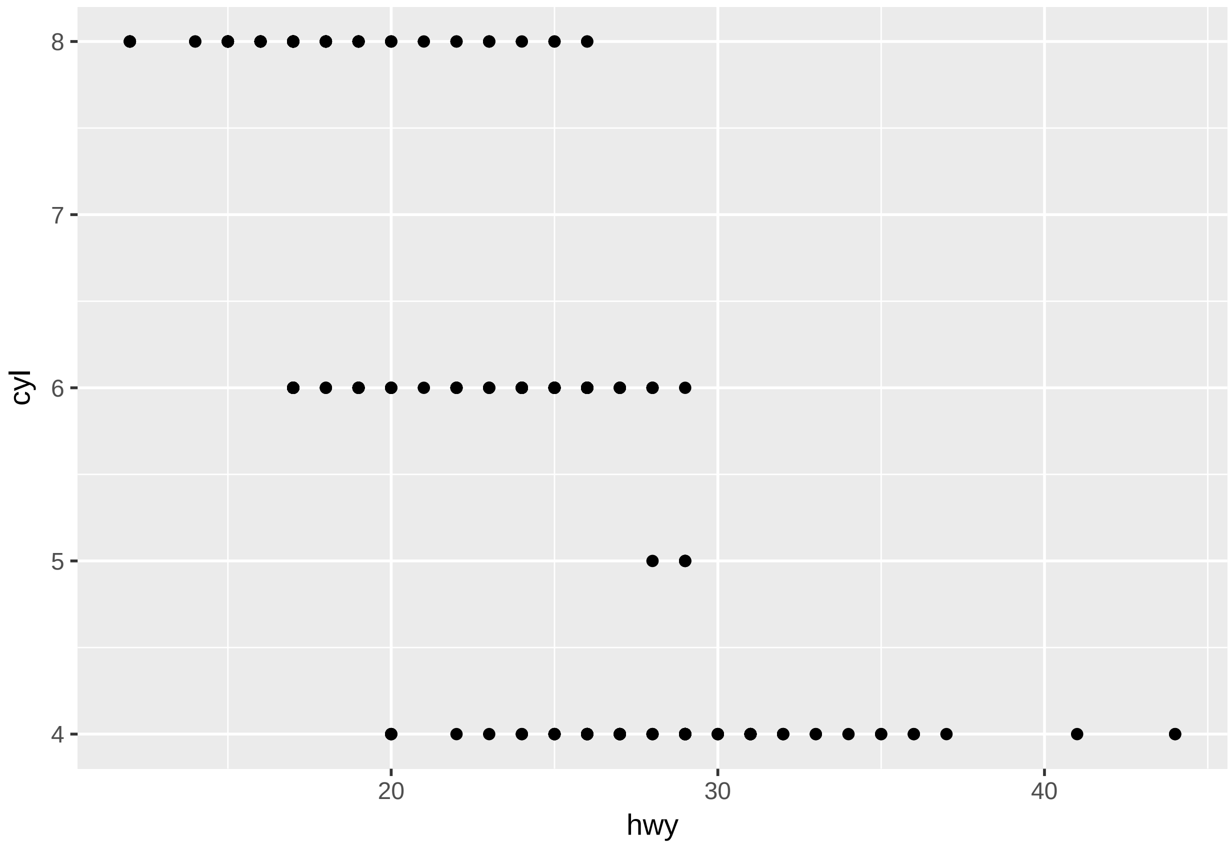
QUESTION 3. What does the drv variable describe? Read the help for ?mpg to find out.

ANSWER 3: It describes the vehicle drive type. f = front-wheel drive, r = rear wheel drive, 4 = 4wd

QUESTION 4. Make a scatterplot of hwy vs cyl

ANSWER 4:

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

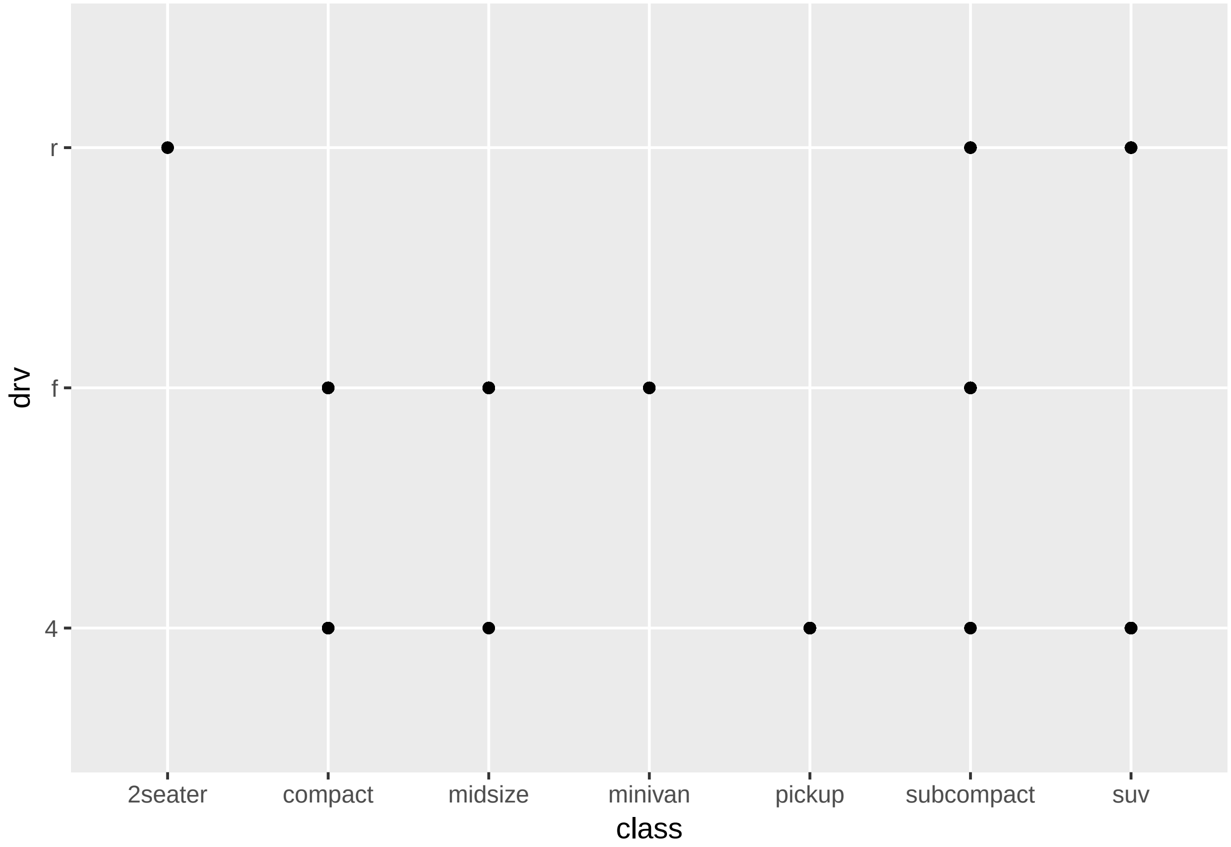


QUESTION 5. What happens if you make a scatterplot of class vs drv? Why is the plot not useful?

ANSWER 5:

The variables class and drv are two independent variables. They don't impact each other, i.e. no relationship between the two.

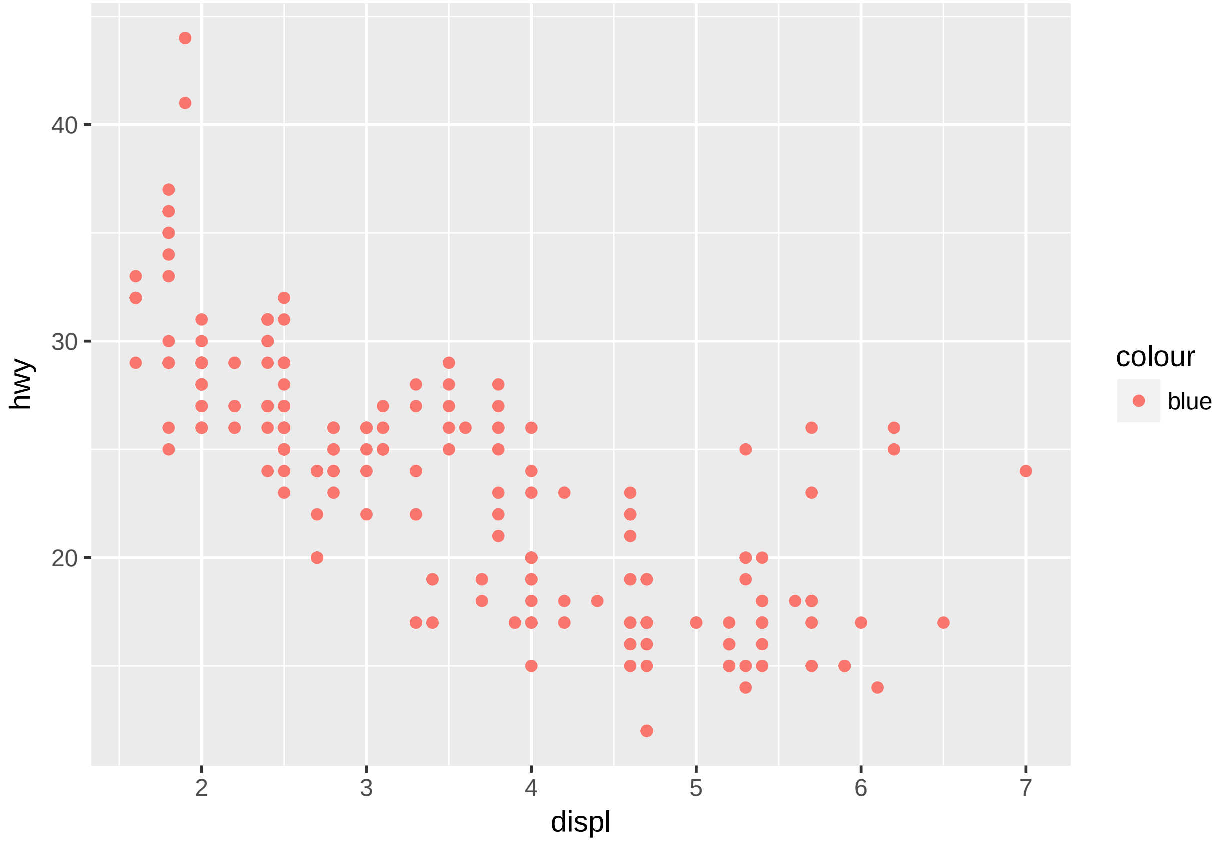
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = class, y = drv))
```



3.3.1 Exercises

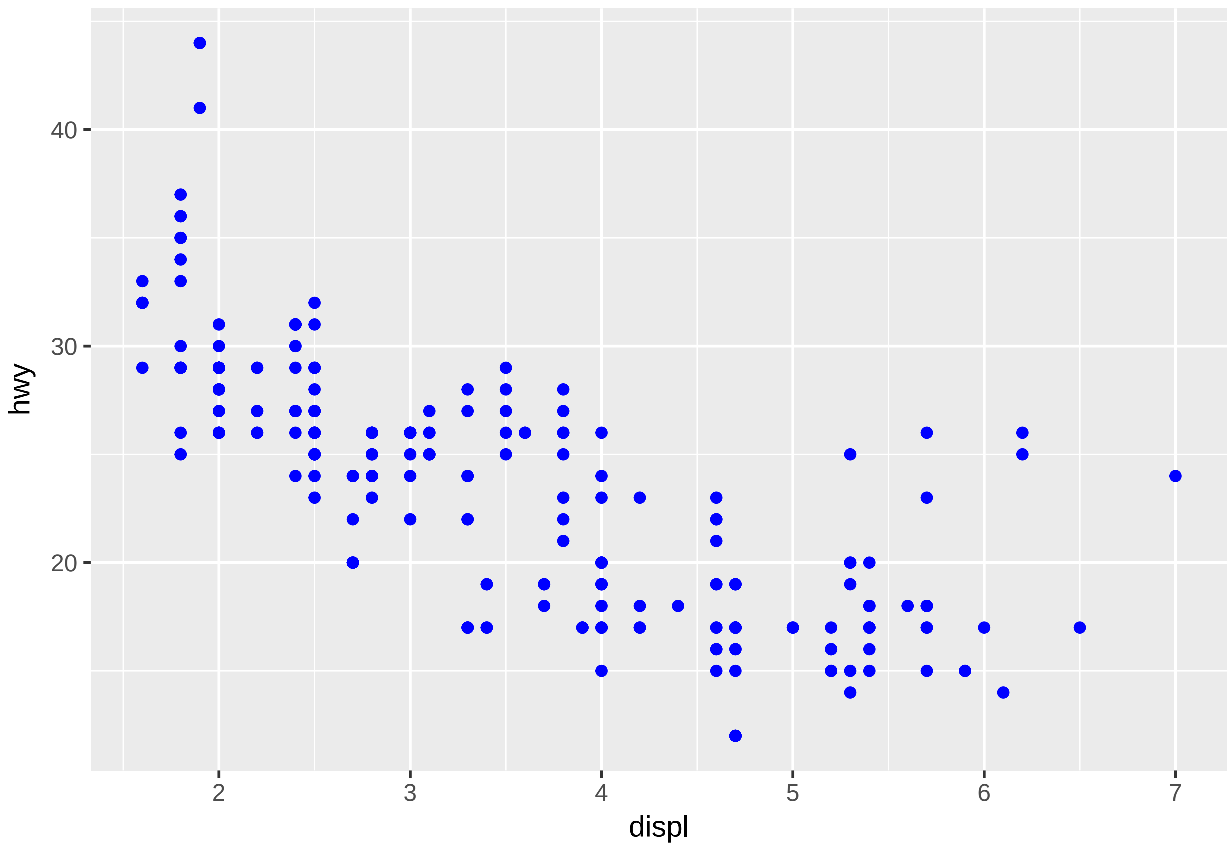
QUESTION 1 : What's gone wrong with this code? Why are the points not blue?

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



ANSWER 1: The code syntax is not correct. The color should not be the part of aesthetic. The correct code should be as below.

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



QUESTION 2: Which variables in mpg are categorical? Which variables are continuous? (Hint: type ?mpg to read the documentation for the dataset). How can you see this information when you run mpg?

ANSWER 2: Categorical variables:trans, drv, fl, class

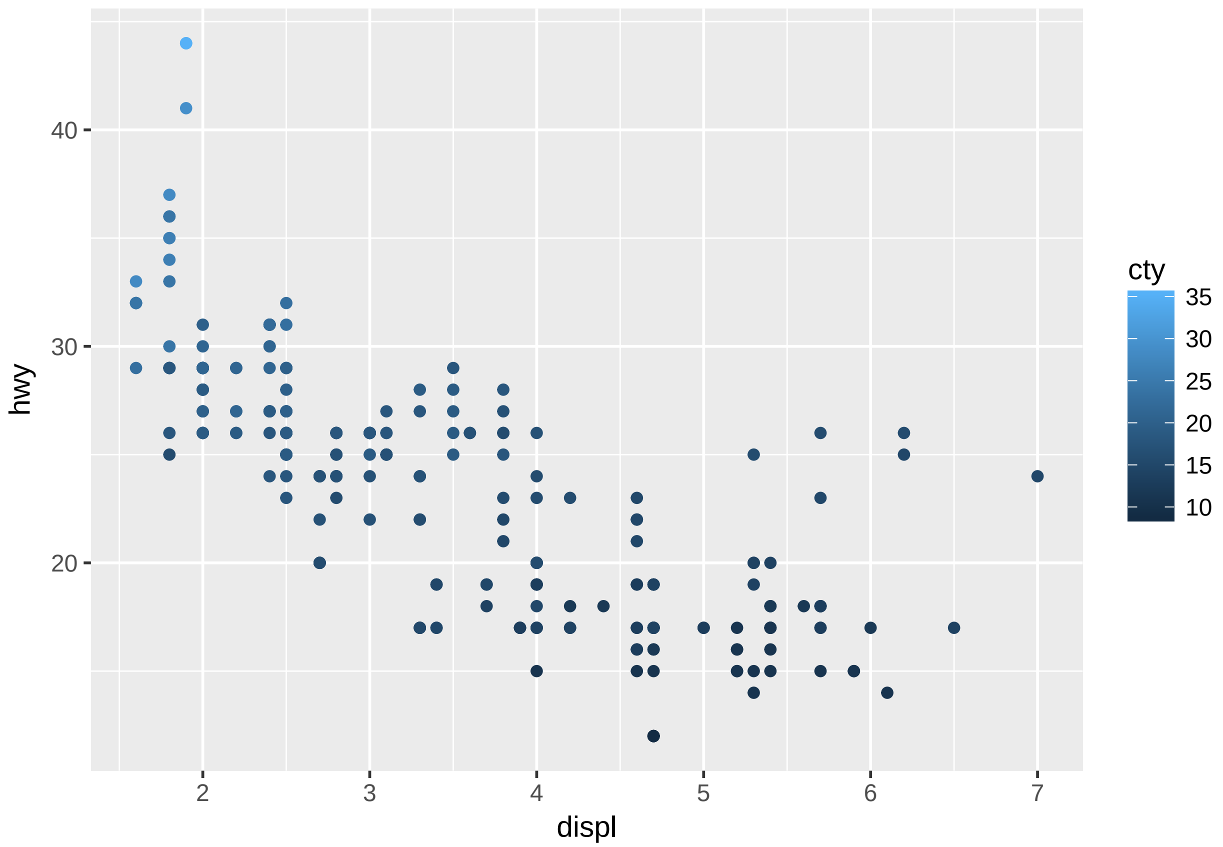
Contunuous variables: manufacturer, model, displ, year, cyl, cty, hwy

The categorical variables have discrete of values (types) whereas continuous variables have a series of values.

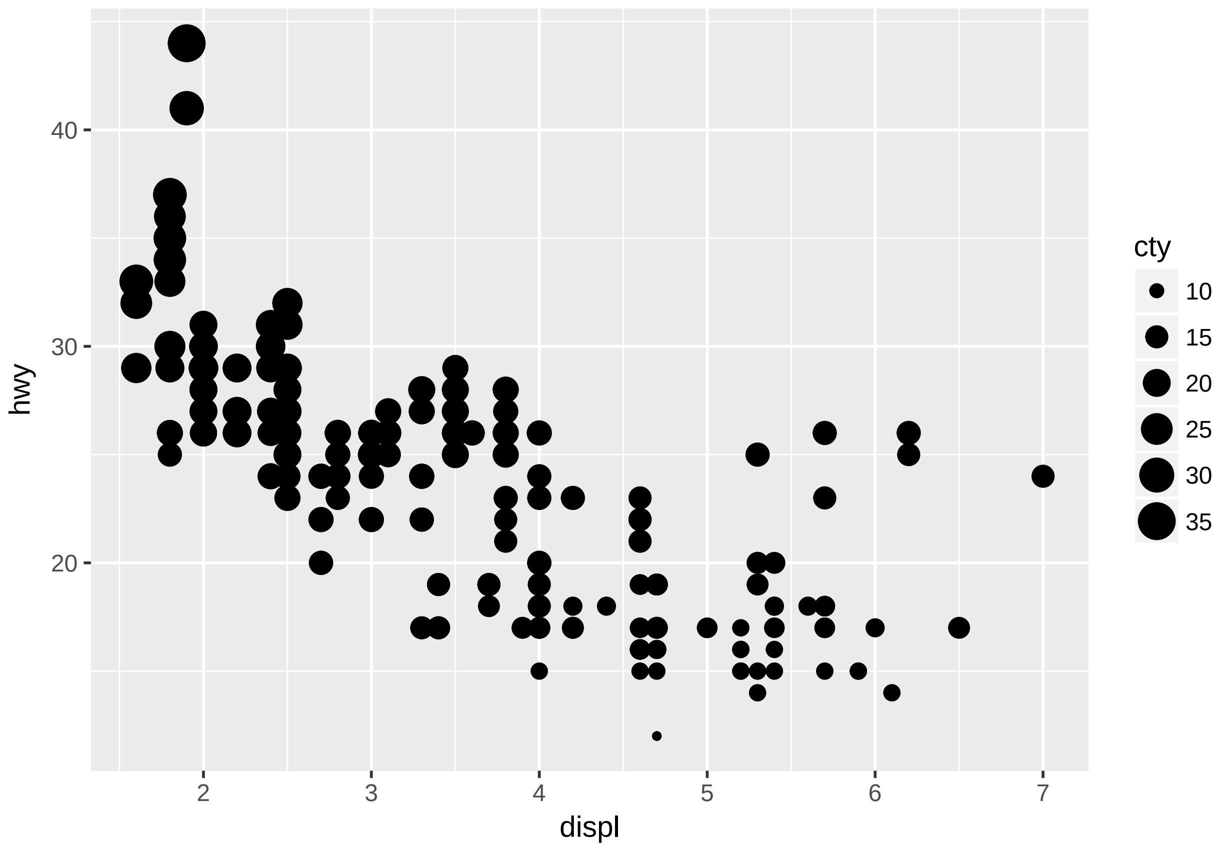
QUESTION 3: Map a continuous variable to color, size, and shape. How do these aesthetics behave differently for categorical vs. continuous variables?

ANSWER 3:

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



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



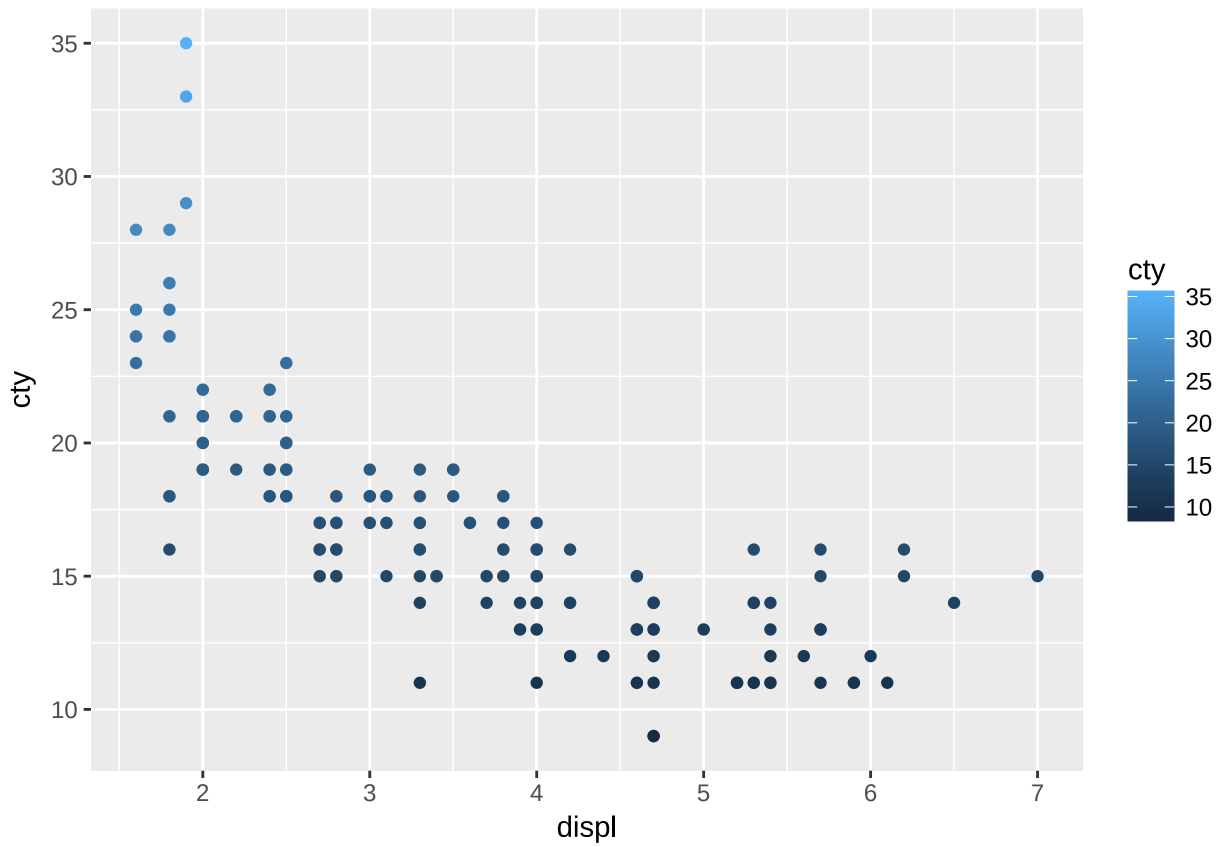
shape cannot be applied to cintinuous variable.Aesthetics provides a continuous(incremental) graph for continuous variables(as in above for variable cty)

QUESTION 4: What happens if you map the same variable to multiple aesthetics?

ANSWER 4:

It displays the grapph for both as below.

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



QUESTION 5: What does the stroke aesthetic do? What shapes does it work with?

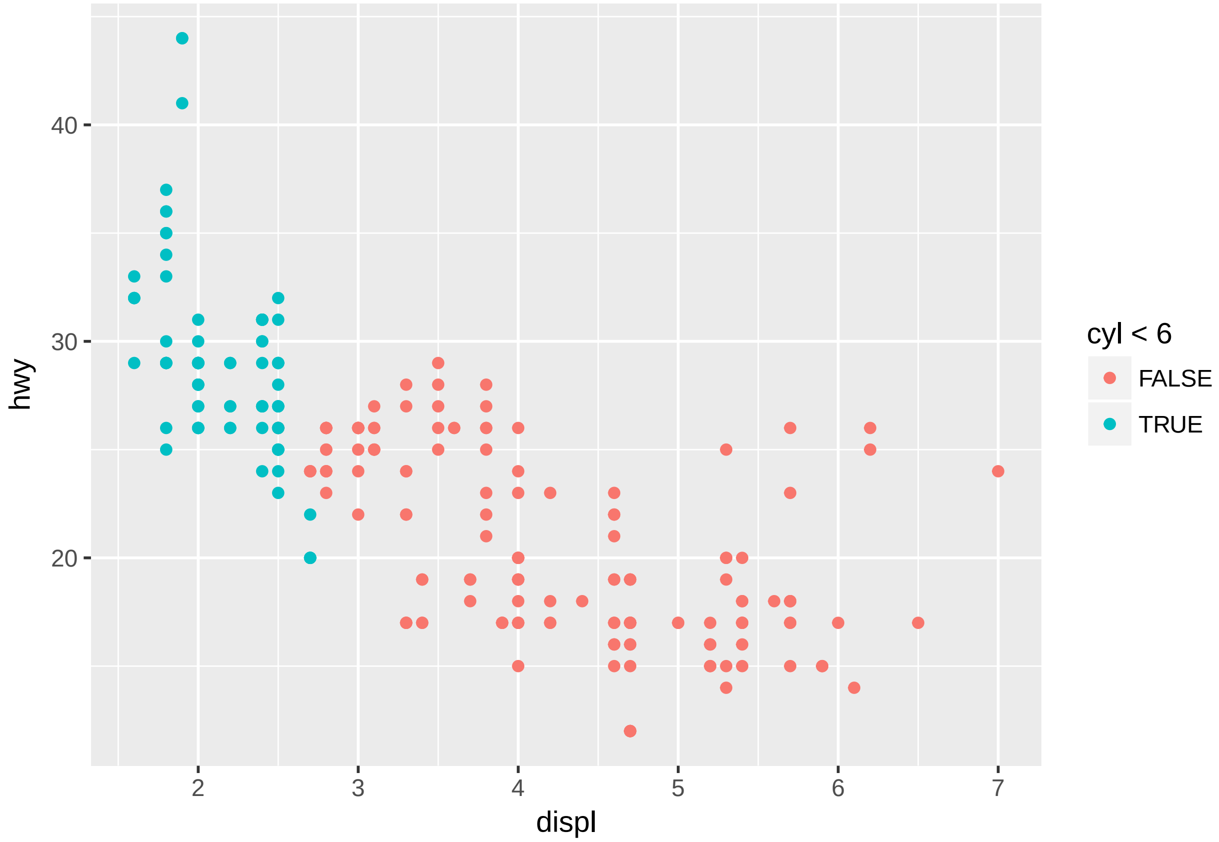
ANSWER 5: stroke is used to control the thickness of the border of the shapes. Used for shape 21-25

QUESTION 6: What happens if you map an aesthetic to something other than a variable name, like aes(colour = displ < 5)?

ANSWER 6:

The graph is divided in two boolean sections of the condition as below.

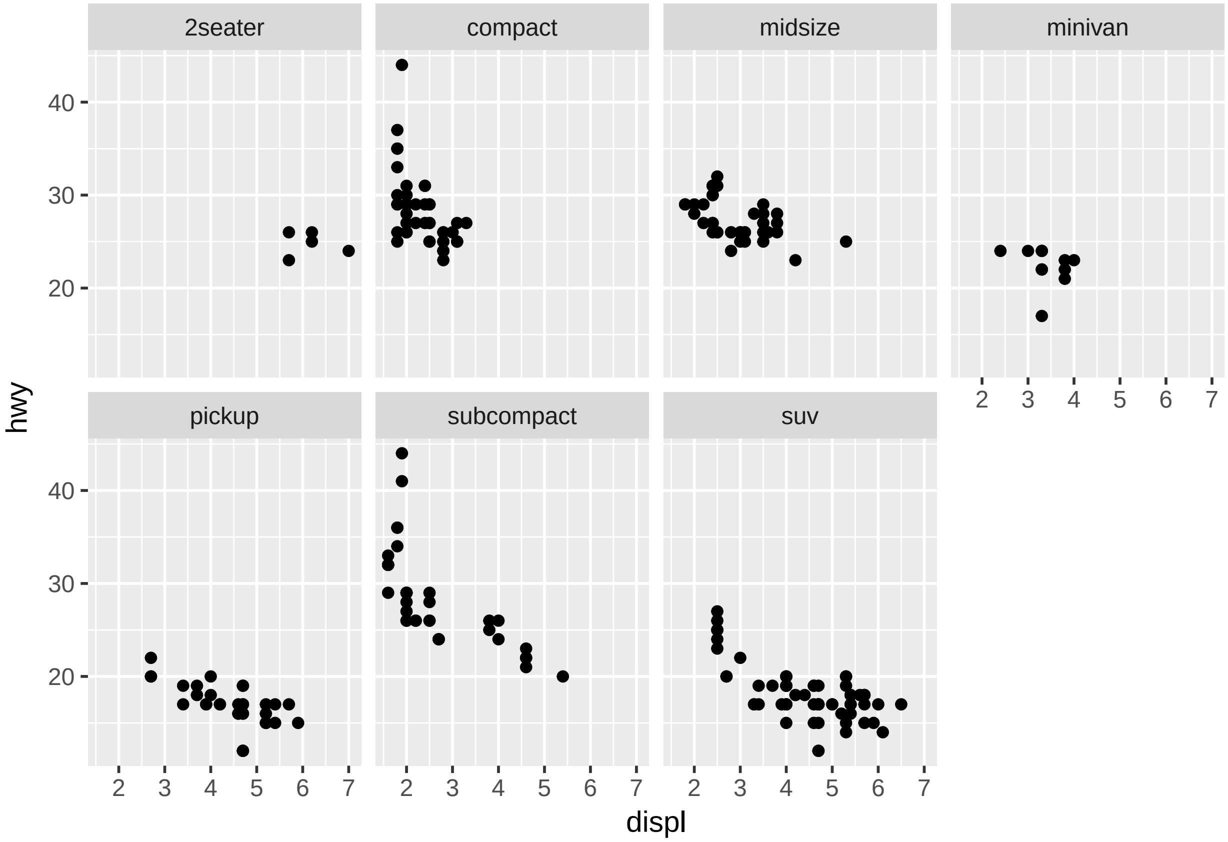
```r
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = cyl<6))
```



3.5.1 Exercises

QUESTION 4 : Take the first faceted plot in this section:

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)
```



What are the advantages to using faceting instead of the colour aesthetic? What are the disadvantages? How might the balance change if you had a larger dataset?

ANSWER 4: Faceting splits the data into separate grids and better visualizes trends within each individual facet.

The disadvantage is that by doing so, it is harder to visualize the overall relationship across facets. The color aesthetic is fine when your dataset is small, but with larger datasets points may begin to overlap with one another.

In this situation with a colored plot, jittering may not be sufficient because of the additional color aesthetic.

QUESTION 5 : Read ?facet_wrap. What does nrow do? What does ncol do? What other options control the layout of the individual panels? Why doesn't facet_grid() have nrow and ncol argument?

ANSWER 5:

nrow sets how many rows the faceted plot will have.

ncol sets how many columns the faceted plot will have.

as.table determines the starting facet to begin filling the plot, and dir determines the starting direction for filling in the plot (horizontal or vertical).

facet_grid forms a matrix of panels defined by row and column facetting variables, nrow and ncol is invalid in this case.

3.6.1 Exercises

QUESTION 1: What geom would you use to draw a line chart? A boxplot? A histogram? An area chart?

ANSWER 1:

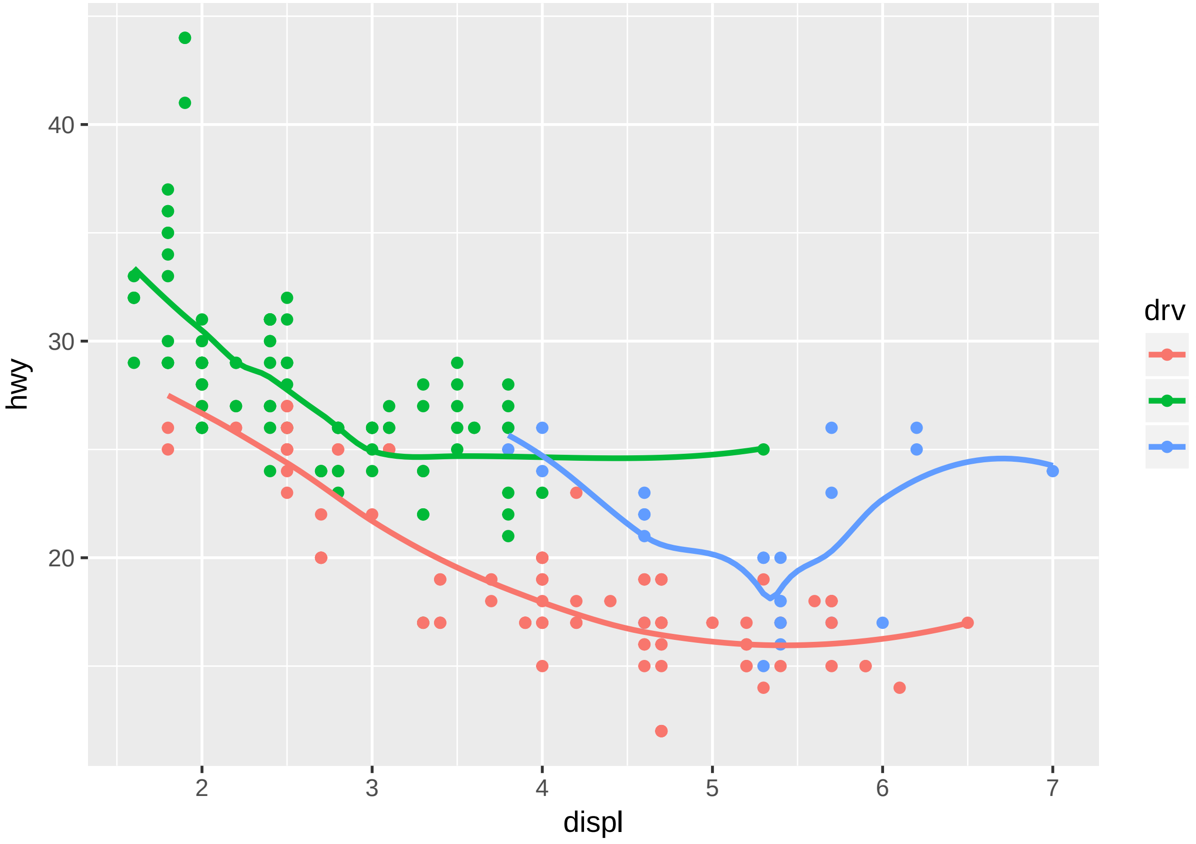
Line chart - geom_line() Boxplot - geom_boxplot() Histogram - geom_histogram() Area chart - geom_area()

QUESTION 2: Run this code in your head and predict what the output will look like. Then, run the code in R and check your predictions.

ANSWER 2:

```r
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```



QUESTION 3:

What does show.legend = FALSE do? What happens if you remove it? Why do you think I used it earlier in the chapter?

ANSWER 3: It removes the legend. The aesthetics are still mapped and plotted, but the key is removed from the graph. Not sure.

QUESTION 4:

What does the se argument to geom_smooth() do?

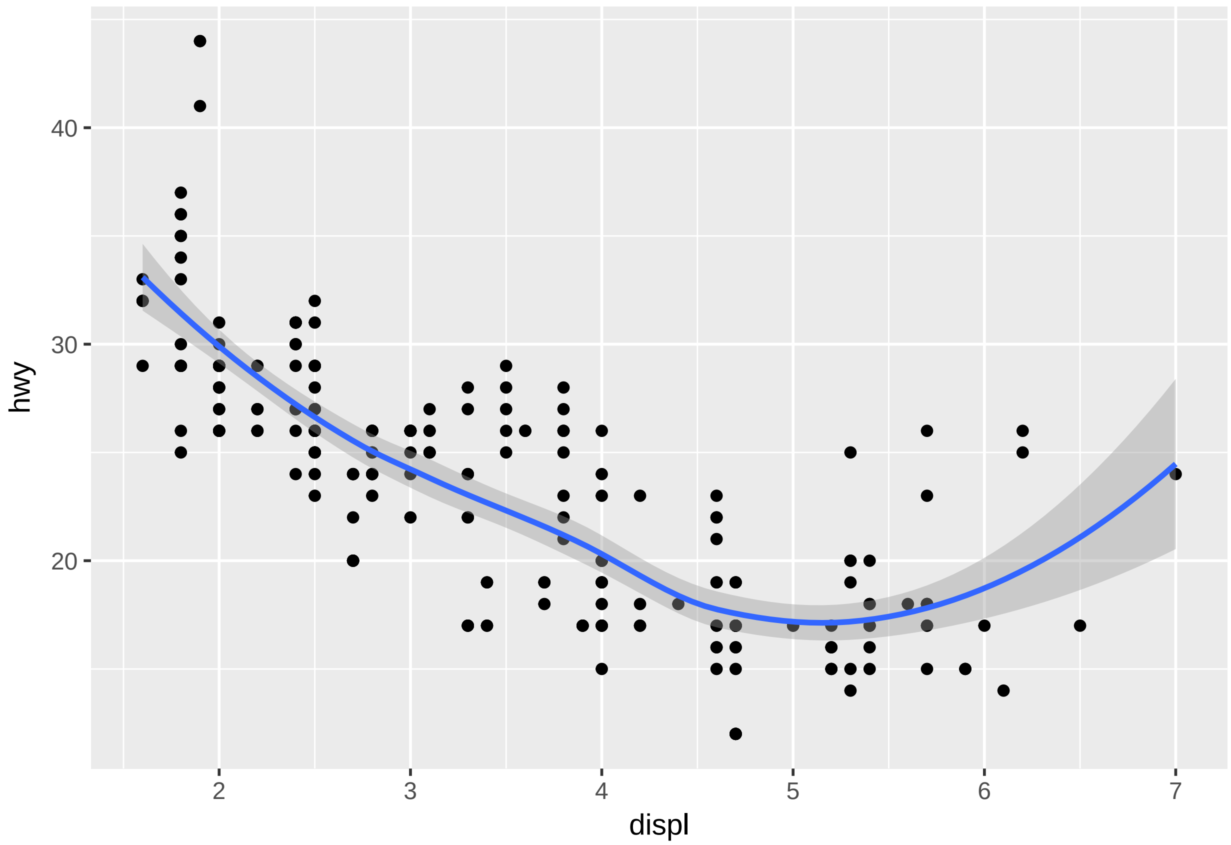
ANSWER 4:

It determines whether or not to draw a confidence interval around the smoothing line.

QUESTION 5:

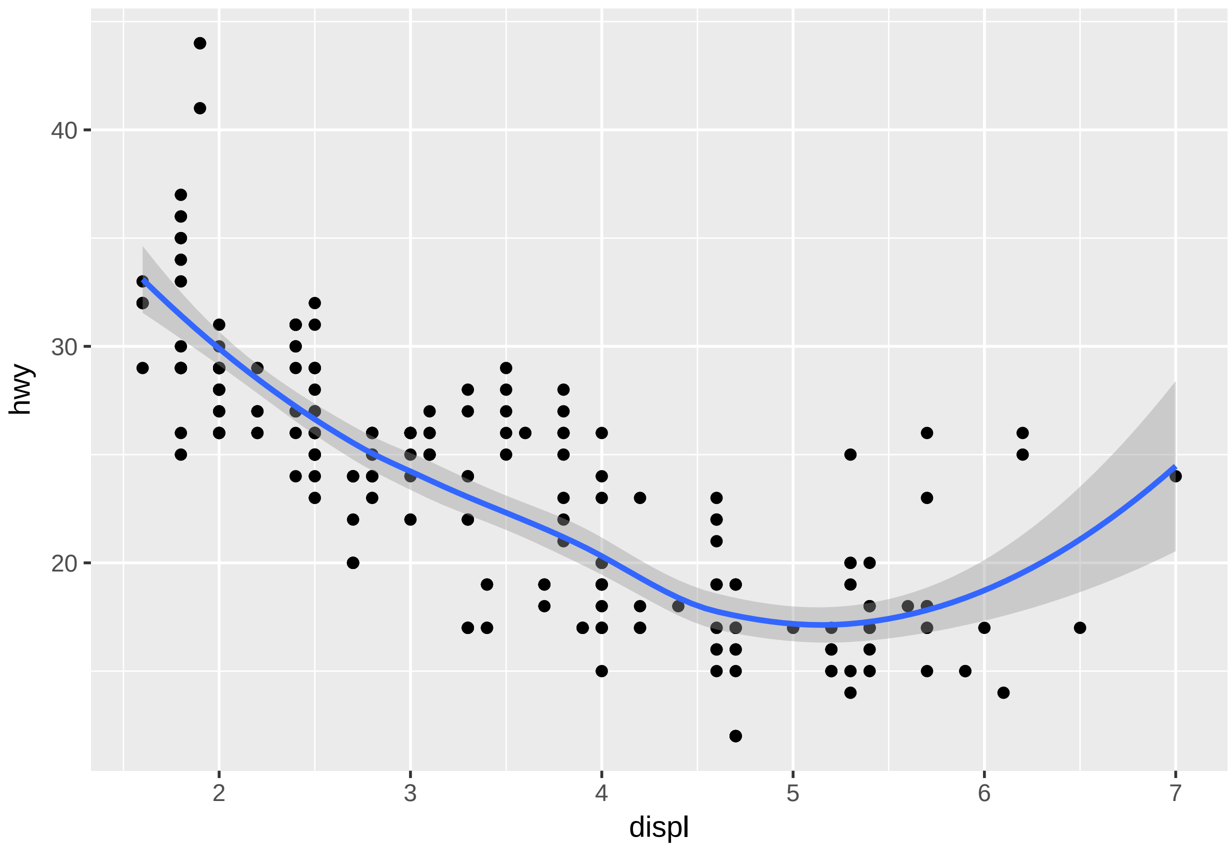
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```



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

```
## `geom_smooth()` using method = 'loess'
```



ANSWER 5:

No because they use the same data and mapping settings. The only difference is that by storing it in the ggplot() function, it is automatically reused for each layer.

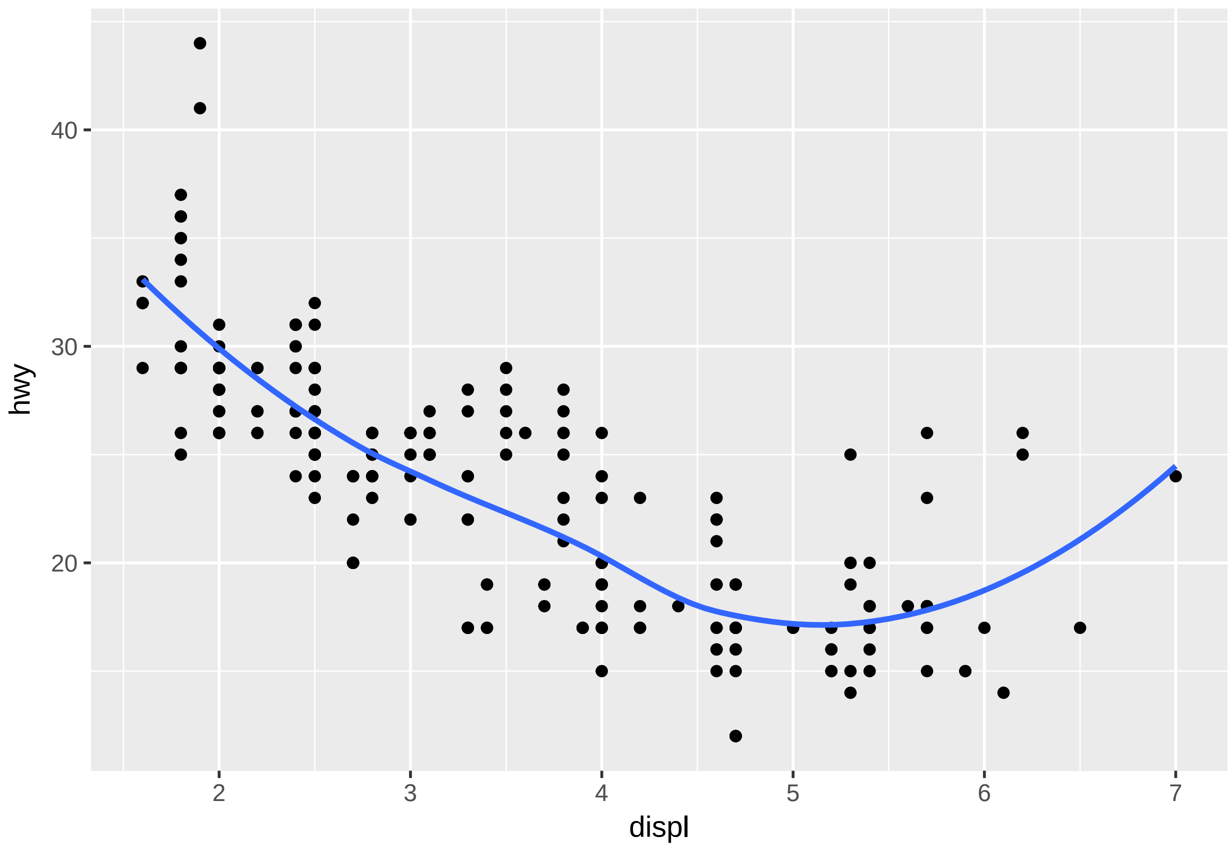
QUESTION 6:

Recreate the R code necessary to generate the following graphs.

ANSWER 6:

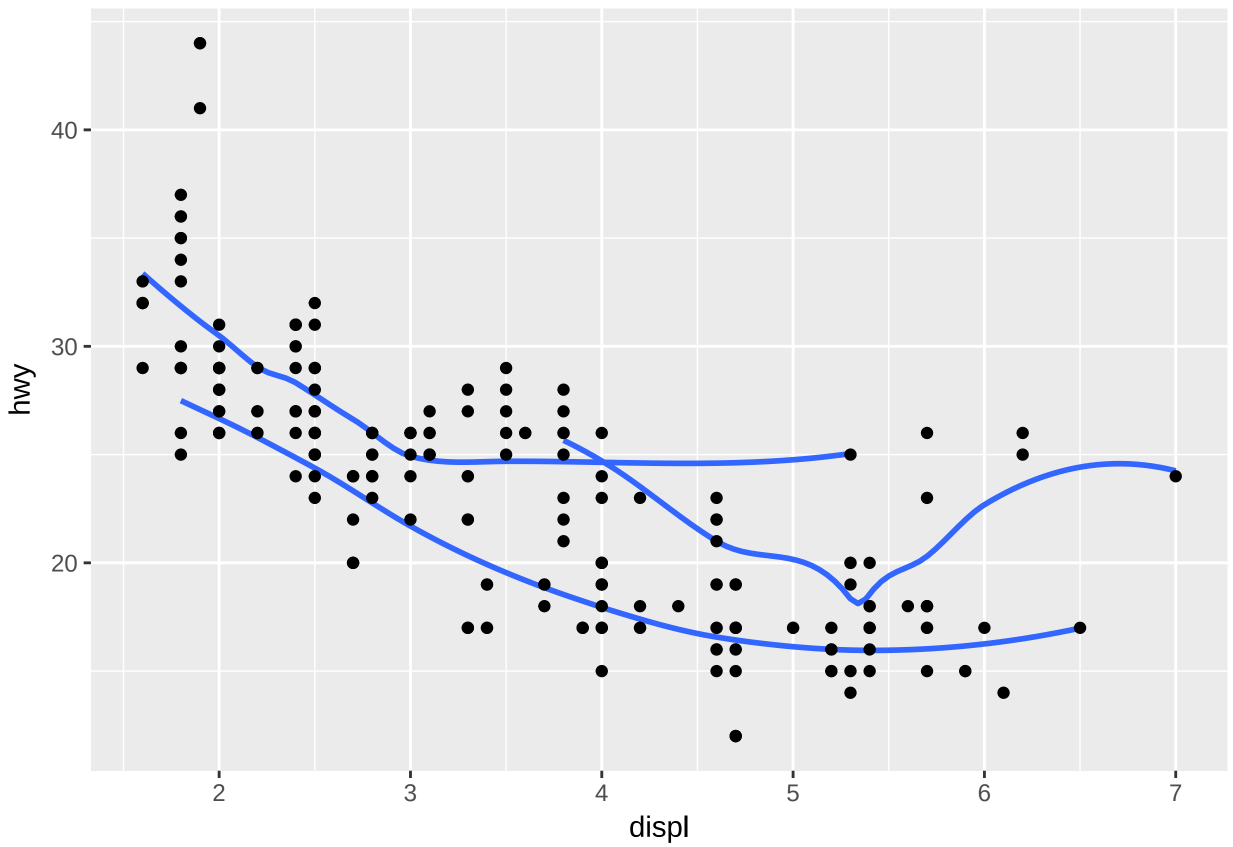
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```



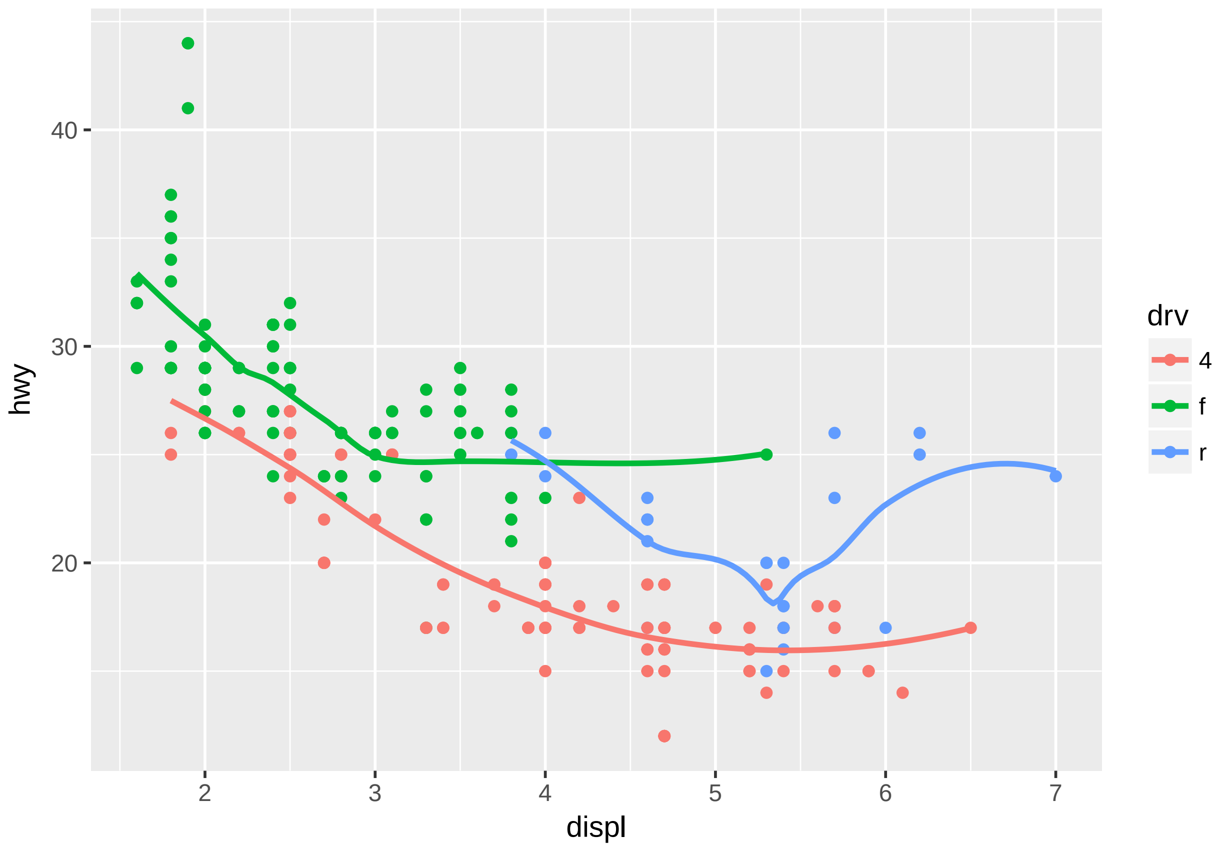
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_smooth(aes(group = drv), se = FALSE) +
  geom_point()
```

```
## `geom_smooth()` using method = 'loess'
```



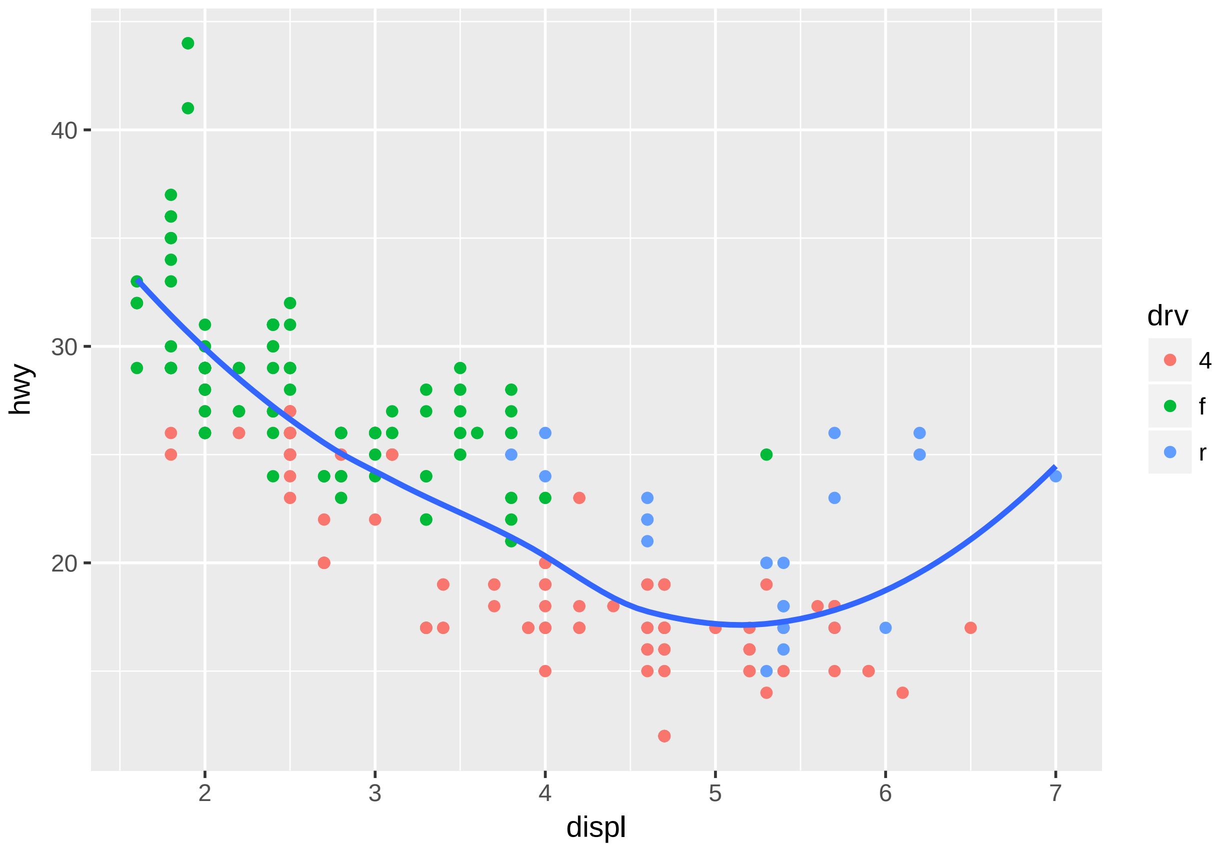
```r
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```



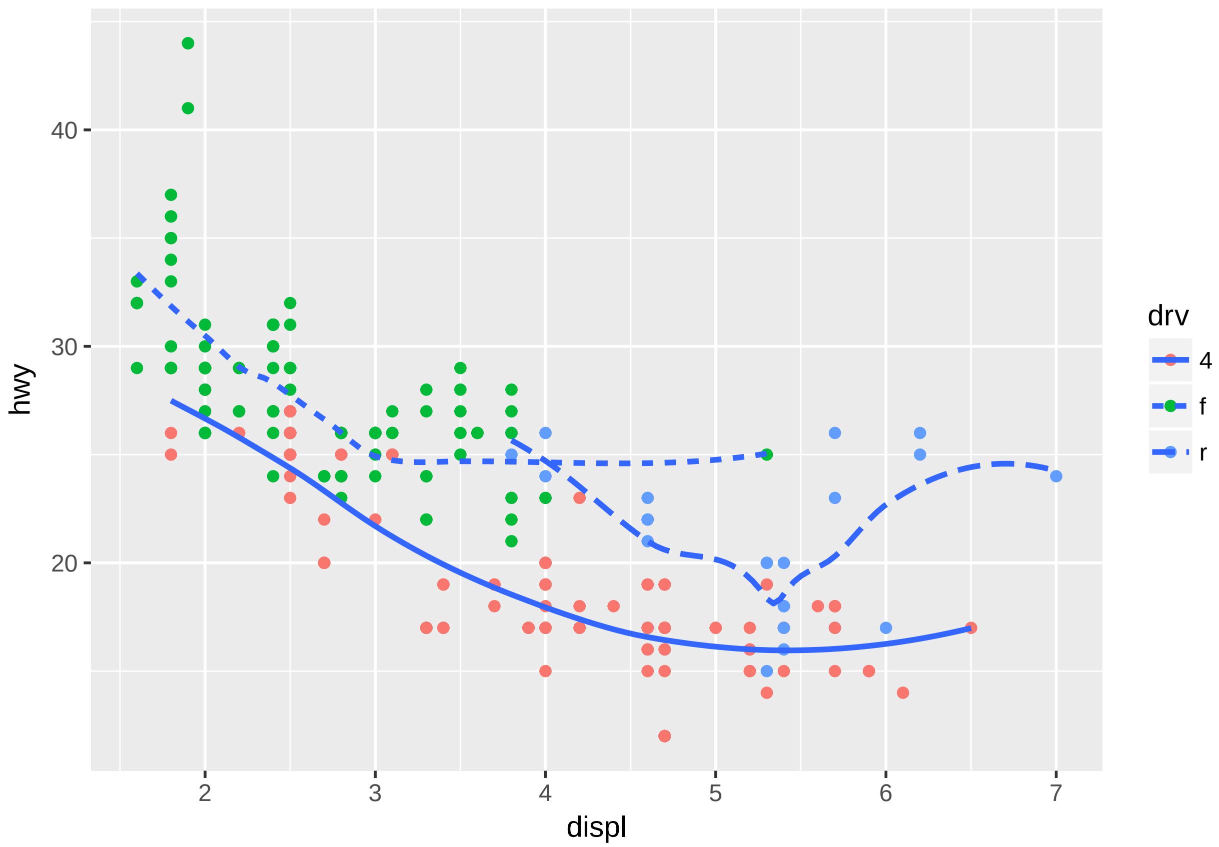
```r
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(aes(color = drv)) +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```

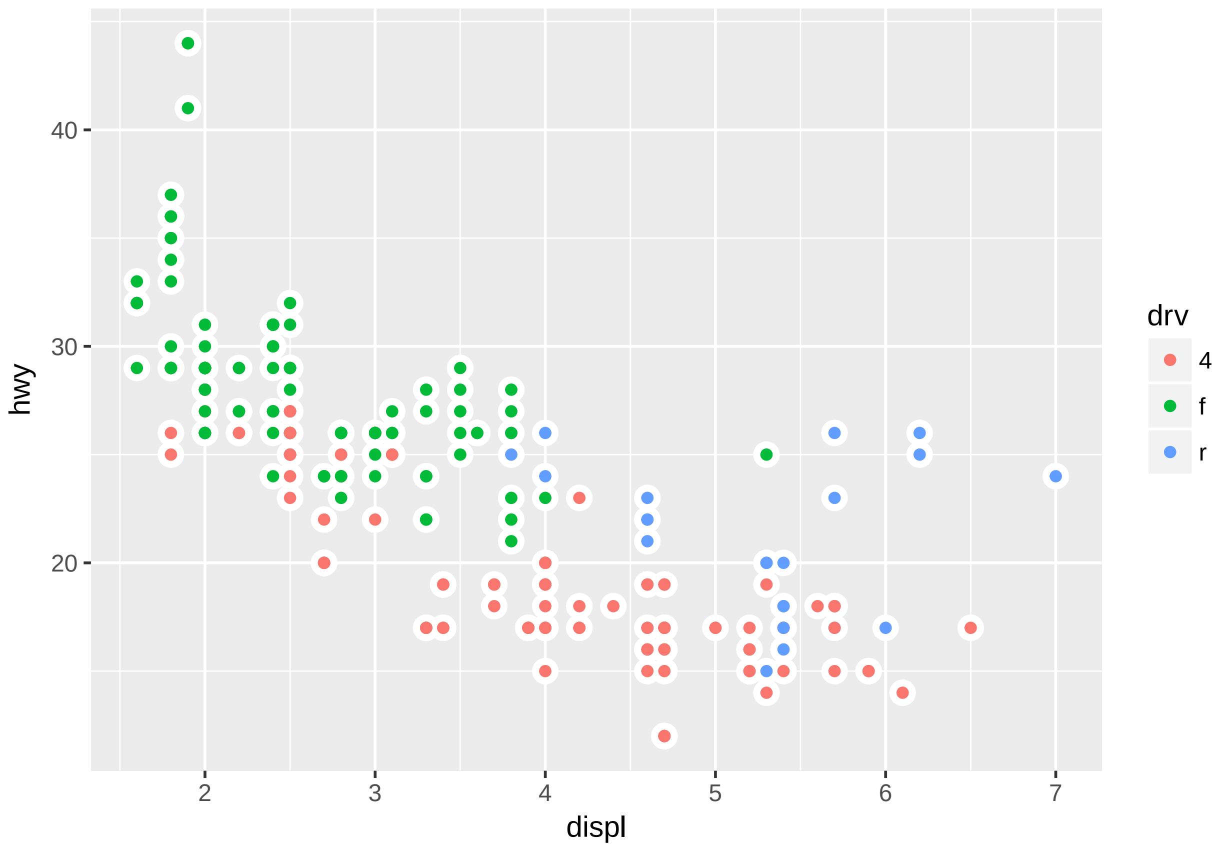


```r
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(aes(color = drv)) +
  geom_smooth(aes(linetype = drv), se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```



```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(size = 4, colour = "white") +
  geom_point(aes(colour = drv))
```



3.7.1 Exercises

QUESTION 2: What does geom_col() do? How is it different to geom_bar()?

ANSWER 2:

In geom_col(), heights of the bars to represent values in the data.

Whereas, geom_bar makes the height of the bar proportional to the number of cases in each group.

Usage:

geom_bar(mapping = NULL, data = NULL, stat = "count", position = "stack", ..., width = NULL, binwidth = NULL, na.rm = FALSE, show.legend = NA, inherit.aes = TRUE)

geom_col(mapping = NULL, data = NULL, position = "stack", ..., width = NULL, na.rm = FALSE, show.legend = NA, inherit.aes = TRUE)

END OF HW2 ASSIGNMENT