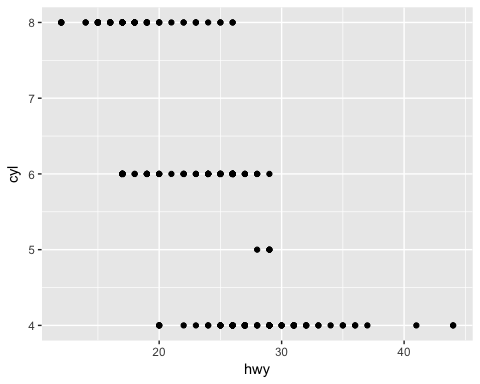
Homework 1

# 3.2.4 Exercises

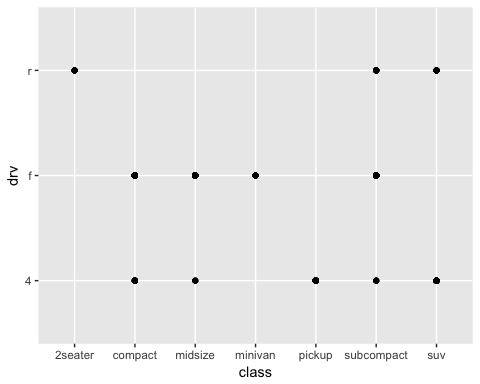
## (a) For the mpg dataset, make a scatterplot of hwy vs cyl. What does this plot tell you about these two variables?

ggplot(data=mpg) + geom\_point(mapping=aes(x=hwy, y=cyl))

 It looks as though this graph is telling me that the more cylinders a vehicle contains, the less highway miles per gallon it achieves. It looks like highway miles per gallon is a continuous variable and number of cylinders is limited to whole numbers and, for the most part whole even numbers.

## (b) What happens if you make a scatterplot of class vs. drv. Is this plot useful? Why or why not?

ggplot(data=mpg) + geom\_point(mapping=aes(x=class, y=drv))

 This graph would be useful if one was looking for information on how drivetrain is related to type of vehicle within this dataset. Neither of these variables appears continous based on this graph. Thus, the points are concentrated on the intersections of the categories within each variable.

# 3.3.1 Exercises

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

The line below will give us basic information about the mpg dataset.

?mpg

Most of this data we can seperate into categorical variables and continuous variables. There are also a few that, I believe, could arguably fit into either categorical or continuous classification.

### Categorical

*Manufacturer (manufacturer)* Year of Manufacture (year) *Numbers of Cylinders (cyl)* Type of Transmission (trans) *Drivetrain (drv)* Fuel Type (fl) \*"Type" of Car (class)

### Continuous

*Model Name (model)* Engine Displacement in Litres (displ) *City Miles Per Gallon (cty)* Highway Miles Per Gallon (hwy)

### Ambiguous

*Model Name (model) - could be considered categorical since there are only certain models listed for this example data set. It could also be considered continuous since the data could be considered to be on a continuous scale.* Engine Displacement in Litres(displ) - could also be considered categorical since there is only a limited range for this variable in this dataset. It could also be considered continuous since the data is on a continuous scale. Number of cylinders (cyl) could be considered continuous since it is an integer, but it has a finite number of options, so I have classified it as categorical.

These two categories are not mutually exclusive. I have seperated them into the two categories in a mutually exclusive way as I saw most fitting.

If we want to view the structure for a specific dataset we can type:

str(mpg)

## 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" ...

And we can add this to the end of an r plot as we have been doing with geom\_point or geom\_smooth.

We can also run this line to see some of the data that is in the dataset

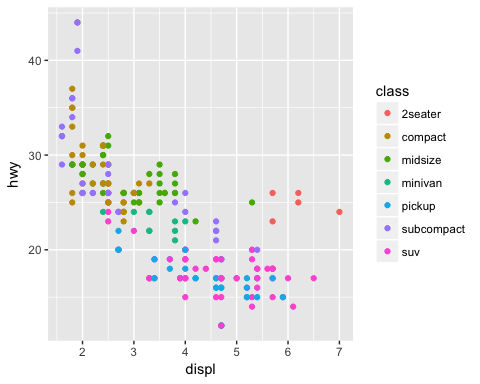
mpg

## # A tibble: 234 × 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>

I'm not sure that answers the question. It is a tad bit ambiguous.

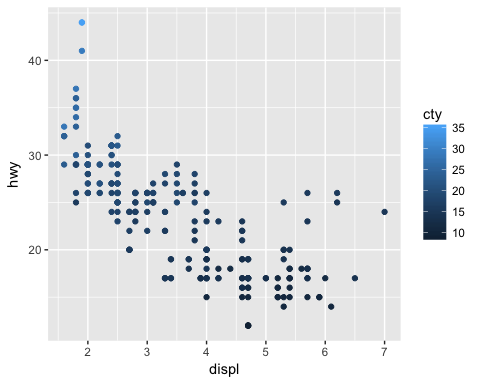
## (b) Consider the scatterplot of displacement vs miles per gallon that we have been studying:

ggplot(data=mpg) + geom\_point(mapping=aes(x=displ, y=hwy, color=class))

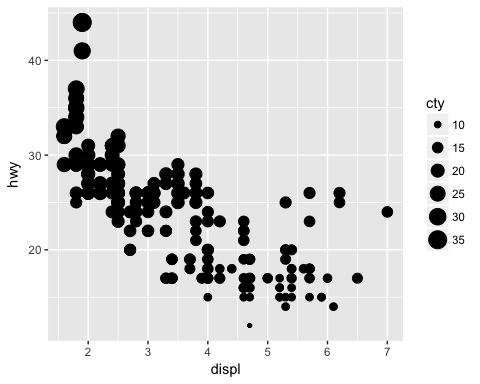


Map a continuous variable to color, size, and shape. How do these aesthetics behave differently for categorical vs. continuous variables?

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))

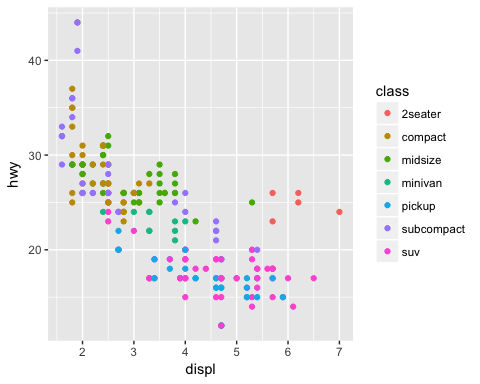


Cannot plot: 'ggplot(data=mpg) + geom\_point(mapping=aes(x=displ, y=hwy, shape=cty))'

Shape only allows for 6 variations. Therefore, if we try to map any variable with more than 6 options, we are presented with warnings because not all of the data will be represented by the plot. Also r prevents a continuous variable to be mapped to shape. If we map a continuous variable to any of these aesthetics, the data creates a spectrum instead of seperating a finite amount of shapes, sizes, or colors as it does for categorical data.

We can see the differences when continuous variables versus categorical variables are assigned to these aesthetics by comparing the below graph of class, a categorical variable, being mapped to color and the above graphs of continuous variables being mapped to the aesthetics mentioned.

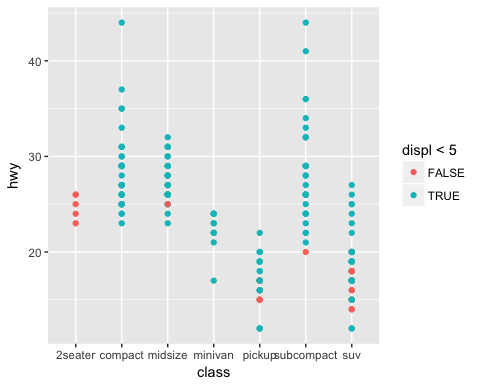
ggplot(data=mpg) + geom\_point(mapping=aes(x=displ, y=hwy, color=class))



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

If we map an aesthetic to something other than a variable name, then the aesthetic will be mapped to the options within this other object. In this case, the variable (displ < 5) is a boolean that will result in either true or false. Thus, this is the resulting graph.

ggplot(data=mpg) + geom\_point(mapping=aes(x=class, y=hwy, color=displ < 5))

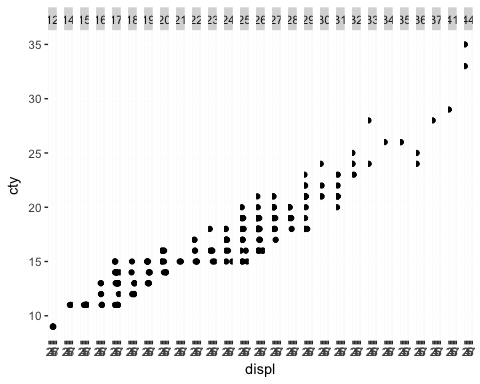


If we wanted to map an aesthetic to another object that was not a variable name, then the aesthetic would map the options within this other object onto the graph.

# 3.5.1 Exercises

## (a) What happens if you facet on a continuous variable?

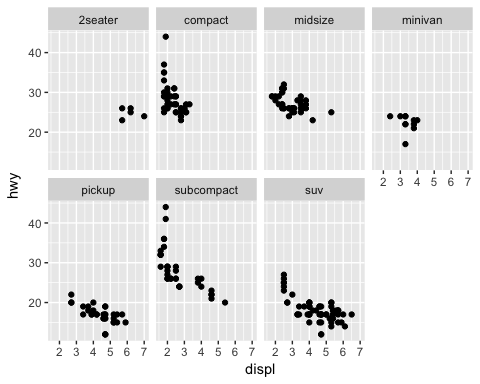
ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = cty)) +   
 facet\_grid(~hwy)



We are presented with numerous graphs that are very difficult to interpret since there are so many of them within a finite viewing space. Faceting on a continuous variable is attempting to make a new plot for each individual value of that continuous variable, which is very difficult since it is continuous. If we do the same thing with a categorical variable, we will be presented with a plot for each individual value of a categorical variable, which is much easier since this will be a finite, pre-defined number.

## (b) Consider the first faceted plot in this section:

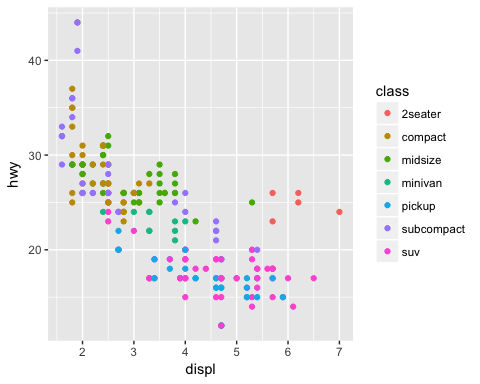
ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = hwy)) +   
 facet\_wrap(~ class, nrow = 2)



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

We can see by comparing the graph below, using color to map class, to the ones from above, using faceting to map class, that it is much more difficult to seperate the class from the below graph. This is because all of the points are placed in the same area. Thus, it is very difficult to compare the classes to one another. The plots from above are much easier to compare to one another and see individual classes mapped based on displacement and highway miles per gallon. The above plots are advantageous over using color to distinguish class of automobile in the graph below. A disadvantage to using faceting is that it may be difficult to spot trends accross all data if we are seperating the data into facets.

ggplot(data = mpg) +  
 geom\_point(mapping = aes( x = displ, y = hwy, color = class))



We can imagine that color can become more advantageous based on the dataset. If we have an immensely large dataset, then color may be better in distinguishing trends in that dataset. We can also imagine a dataset that has isolated variables as easily distinguishable using color or faceting.

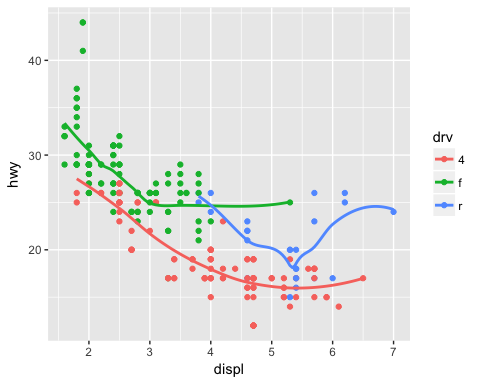
# 3.6.1 Exercises

## (a) Run this code in your head and predict what the output will look like. Then, run the code in R and check your predictions:

This graph will trend in a negative direction, for as displacement continues higher to the right (x-axis), highway miles per gallon will continue downward (y-axis). Adding the aspect of color to drivetrain, will seperate the data out along the x-axis, for as displacement continues higher to the right, drv will seperate into rear wheel drive. It will also seperate the data along the y-axis, for as highway miles per gallon continues to rise upward, drv will seperate into front wheel drive. Four wheel drive is somewhat continuous across the graph, trending in a negative direction for the reasons stated above. We will get a line drawn between the trends of the graph as well as points that reflect the dataset since we are using both "geom\_point" and "geom\_smooth".

ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +   
 geom\_point() +   
 geom\_smooth(se = FALSE)

## `geom\_smooth()` using method = 'loess'

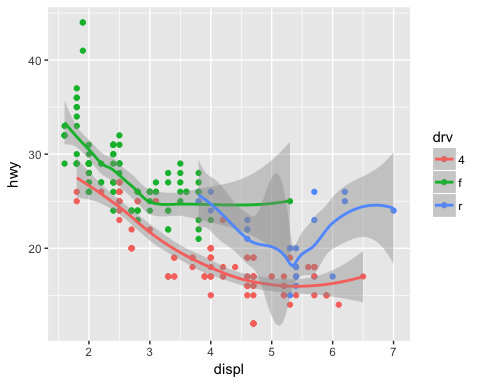


## (b) What does the se argument to geom\_smooth() do?

This puts a gray area around our trend lines from "geom\_smooth". The gray area is referred to as a confidence area, since it gives us a gray area to demonstrate with what confidence the trend line was plotted. The wider the gray area, the less confidence there is in the line at that point.

ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +   
 geom\_point() +   
 geom\_smooth(se = TRUE)

## `geom\_smooth()` using method = 'loess'

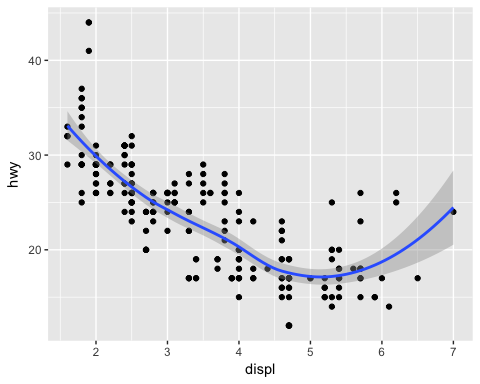


## (c) Will these two graphs look different? Why/why not?

No these graphs will not look different because they will be graphing the exact same variables. In one, we are just defining the plot as a whole then adding in the types of geometries we would like plotted on the graph. On the second, we are defining a graph then adding geometries with specific variables defined within the geometries themselves. These specific variables within these geometries match the variables from the first graph where we define the variables when defining the graph itself.

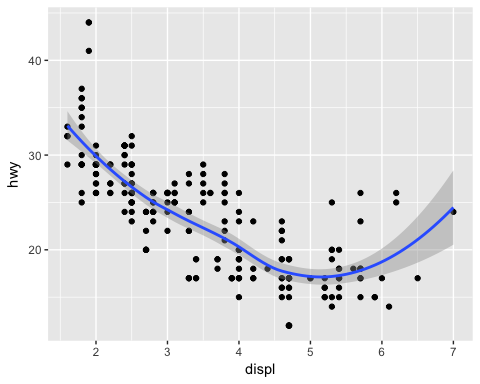
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +   
 geom\_point() +   
 geom\_smooth()

## `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))

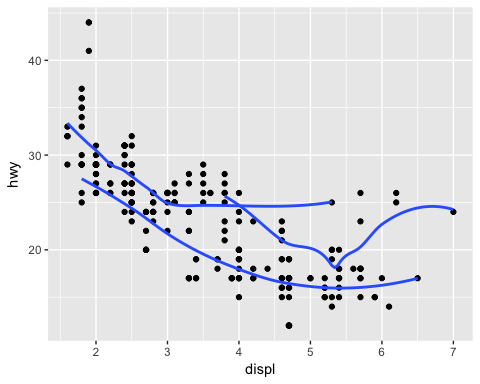
## `geom\_smooth()` using method = 'loess'



## (d) Recreate the R code needed to generate each of the 6 graphs for problem 6.

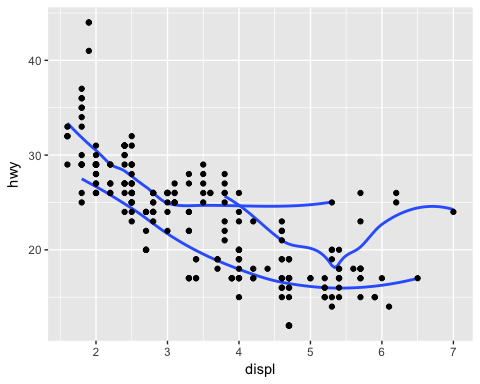
ggplot(data=mpg, mapping=aes(x=displ, y=hwy, group=drv)) + geom\_point() + geom\_smooth(se=FALSE)

## `geom\_smooth()` using method = 'loess'



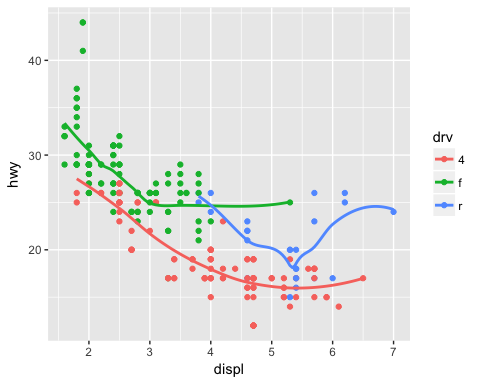
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +   
 geom\_smooth(aes(group = drv), se = FALSE) +  
 geom\_point()

## `geom\_smooth()` using method = 'loess'



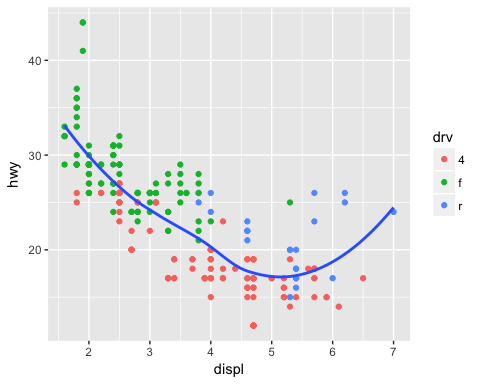
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +   
 geom\_point() +   
 geom\_smooth(se = FALSE)

## `geom\_smooth()` using method = 'loess'



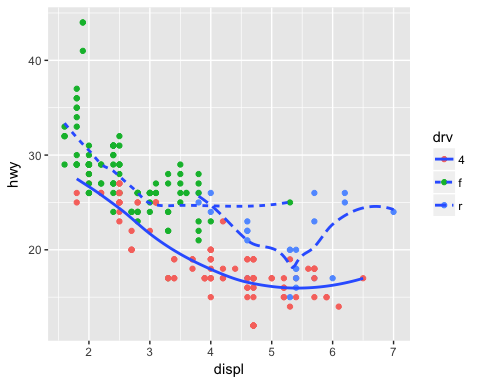
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +   
 geom\_point(aes(color = drv)) +   
 geom\_smooth(se = FALSE)

## `geom\_smooth()` using method = 'loess'



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))

