COMPSCIX 415.2 Homework 2

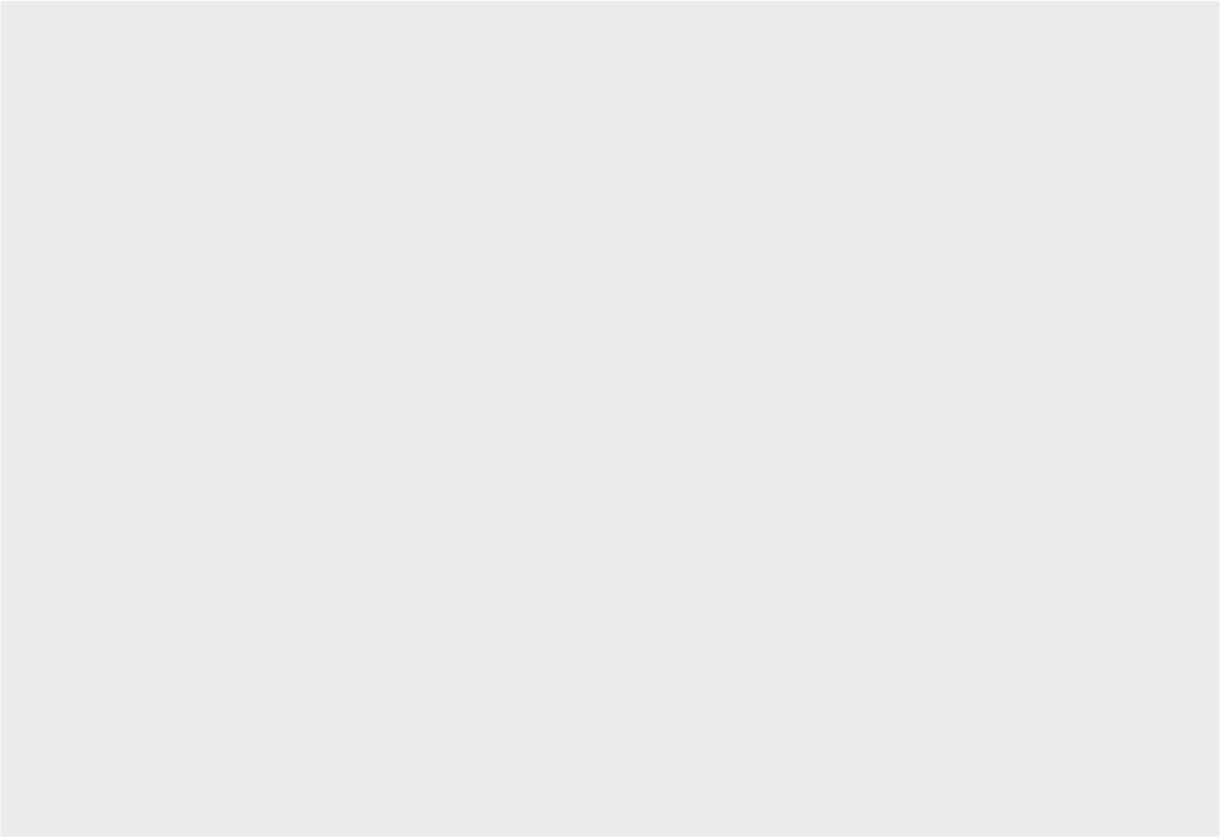
```
library(tidyverse)
data("mpg")
glimpse(mpg)
```

```
## 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...
```

EXERCISE 3.2.4

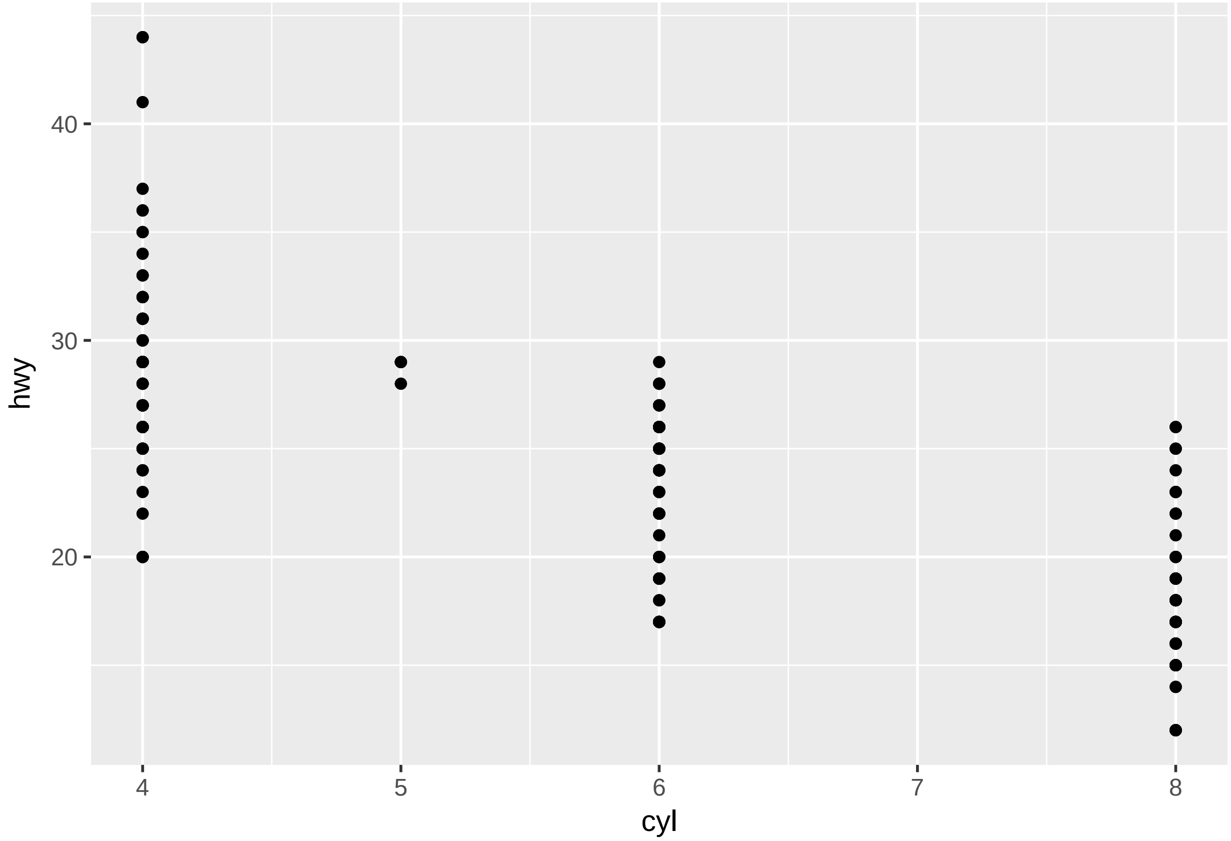
1. All i see is a blank canvas or axis-set. See code below:

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



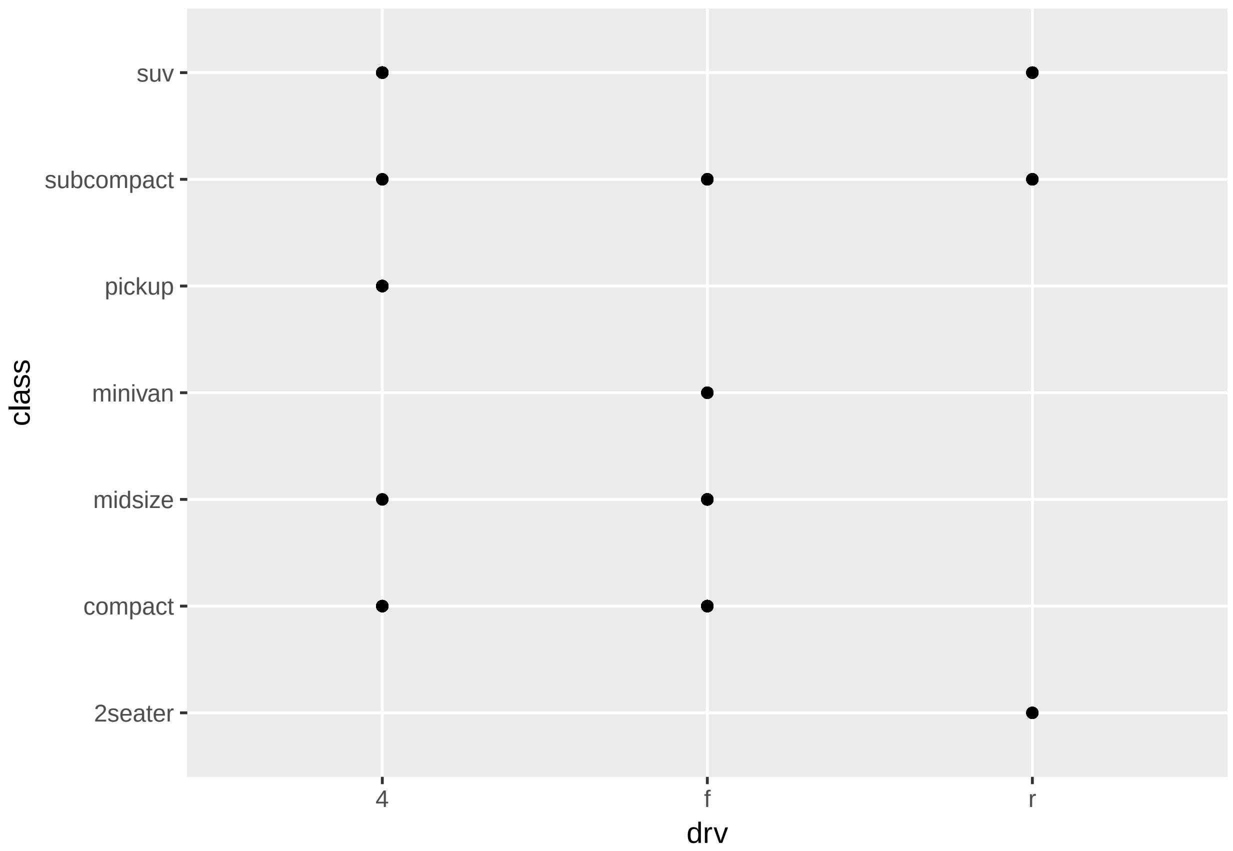
2. There are 234 rows and 11 columns in the mpg dataset
3. The drv variable describes the drive type of the car. i.e. whether the car is front-wheel drive, rear-wheel drive, or 4-wheel drive
4. See scatterplot of hwy vs. cyl below:

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



5. The graph class vs. drv is not very useful because we are graphing discrete 1-to-1 categorical data as a scatter plot. It doesn't allow an easy visual understanding of how many classes belong to each drive type.See graph below:

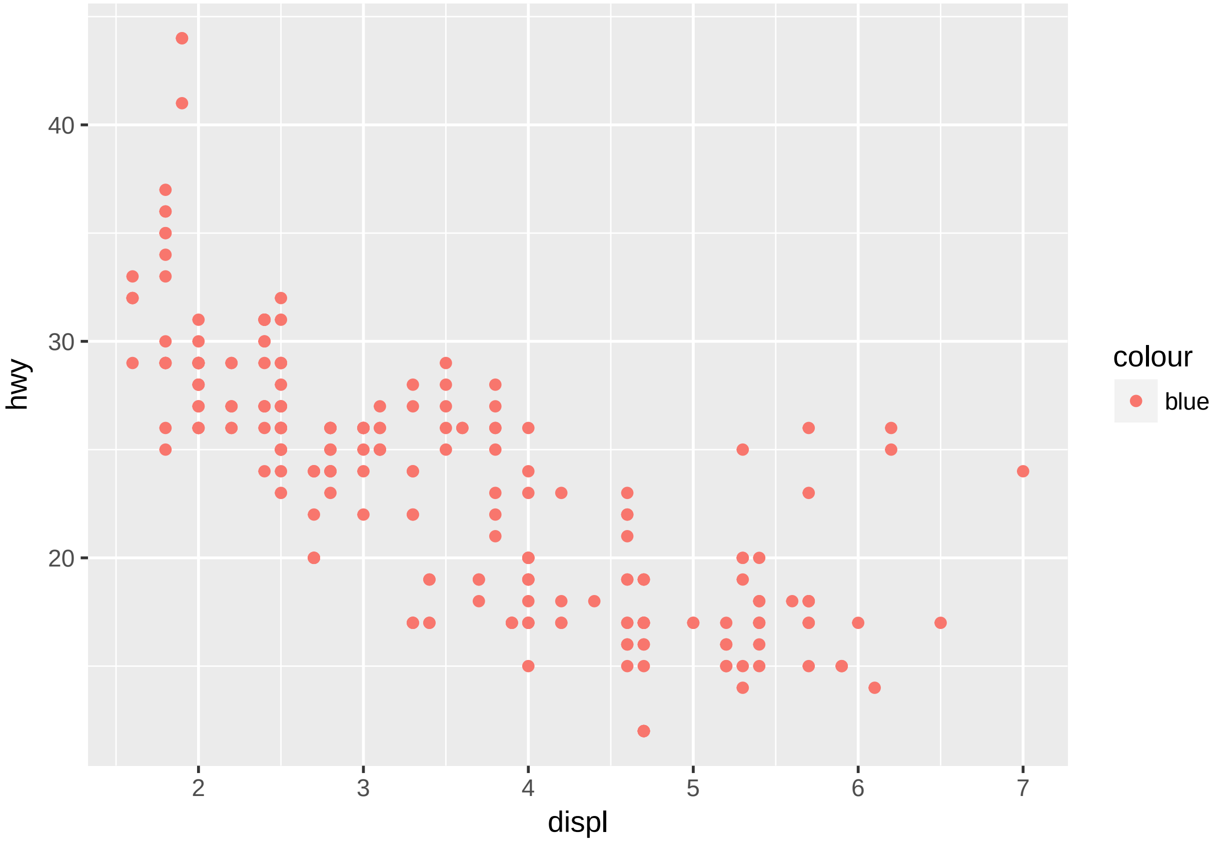
```
ggplot(data = mpg) + geom_point(mapping = aes(x = drv, y = class))
```



EXERCISE 3.3.1

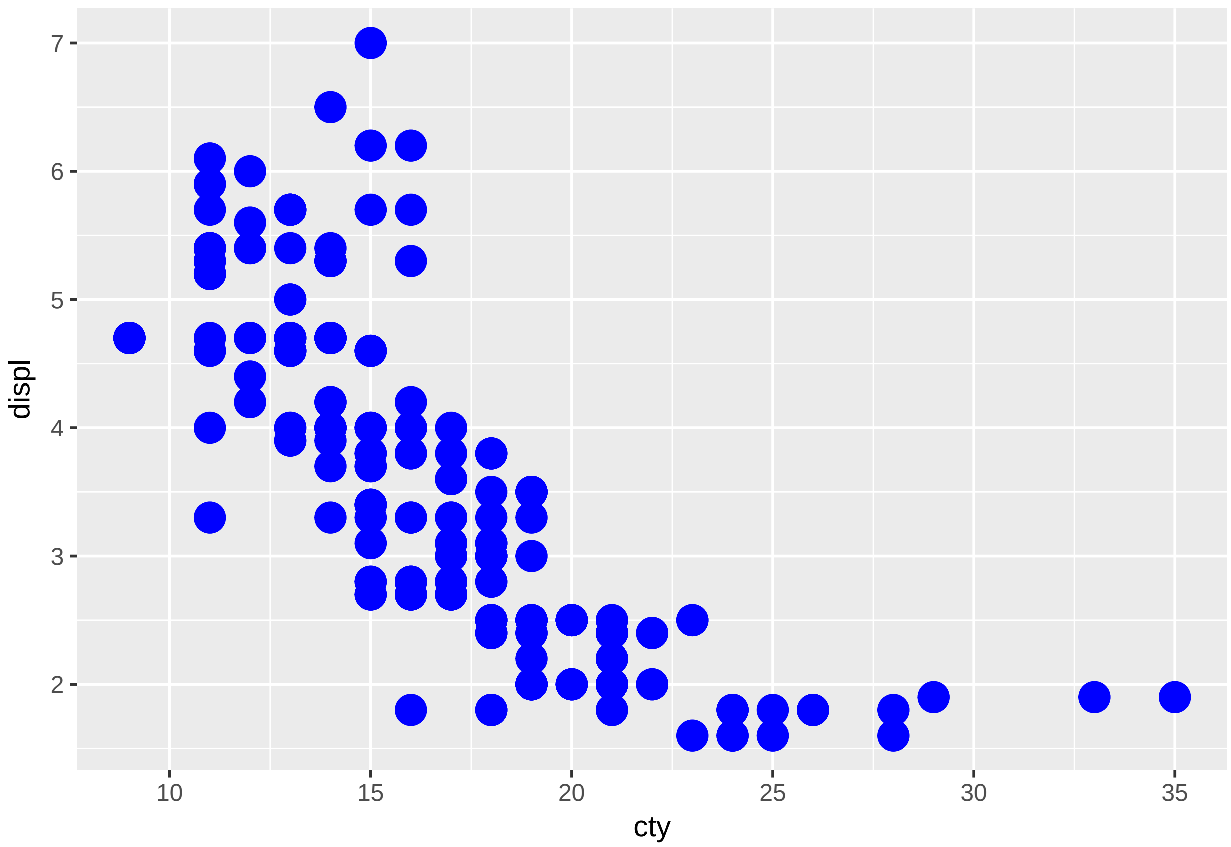
1. The following code does not plot the mpg dataset hwy vs. disp in the color blue because the code "color =”blue“" is included in the aes() function rather than outside aes() but still inside geom_point()

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



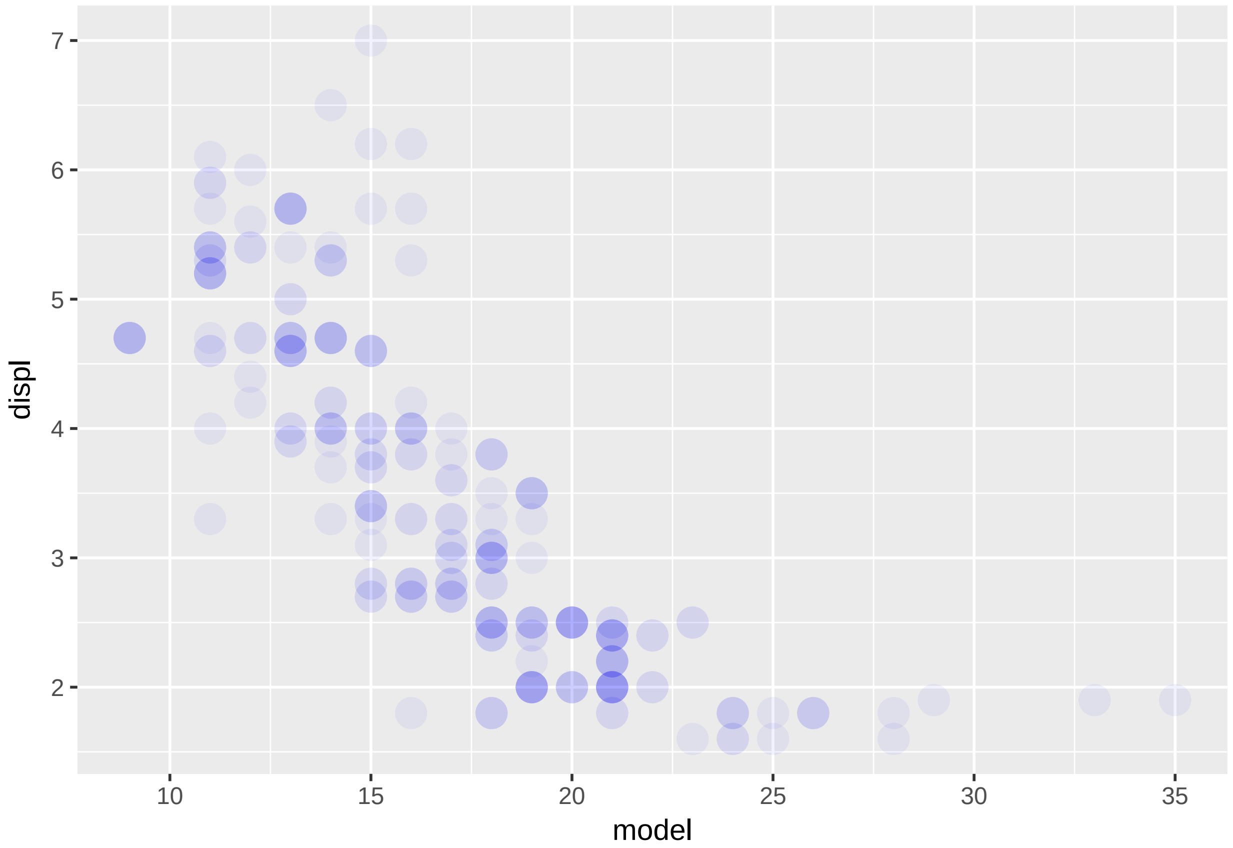
2. The categorical variables included in the dataset mpg, include: manufacturer, model, trans, drv, fl, and class. The continuous variables included are: displ, year, cyl, cty, and hwy
3. A plot of displ vs. cty in blue color with a size of 5 is shown below. Mapping continuous variables allows for overlapping/same data points, whereas every point in a categorical datat set, when plotted, is discrete (i.e. no overlapping)

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = cty, y = displ), color = "blue", size = 5)
```



4. Adding another aes() changes the name of the axes but does not change the actual plotted data

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = cty, y = displ), color = "blue", size = 5, alpha = 1/20) +
  aes(x = model, y = displ)
```



5. The stroke aesthetic allows the user to modify the width of the border of plotted shapes. It only works for shapes that have a border.

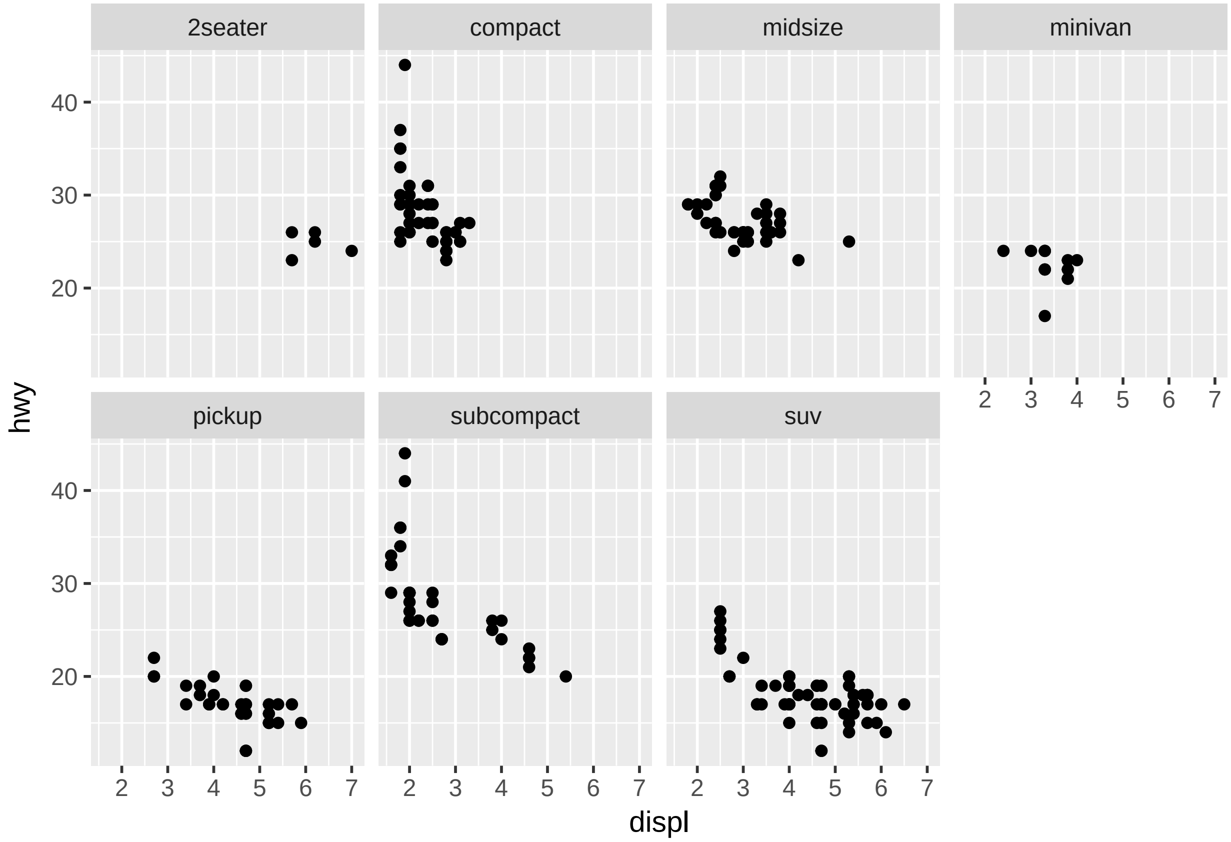
6. An error message occurs when you try to map an aesthetic other than a variable name.

EXERCISE 3.5.1

4. An advantage of using the aesthetic of faceting is that it is an easy way to visually understand the quantity of data in category of the plotted variable (in this case class). It is also useful in getting a more intimate understanding of how each category of the plotted variable (class) is distributed across the independent variable. However, it is not a great way of understanding how each category compares to the others (i.e. understanding the distribution of the entire plotted dataset). If the dataset is larger, it is a less effective visual tool to communicate the differences between the categories, if each category tends to have similar distributions as their respective sample sizes grow. However, it is much more useful if each category tends to have a unique distribution (such as hwy vs. displ). This is because the greater the size of the dataset, the "noisier" a plot gets, so disagregating the plotted data into the categories gives the viewer more granularity.

5.

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



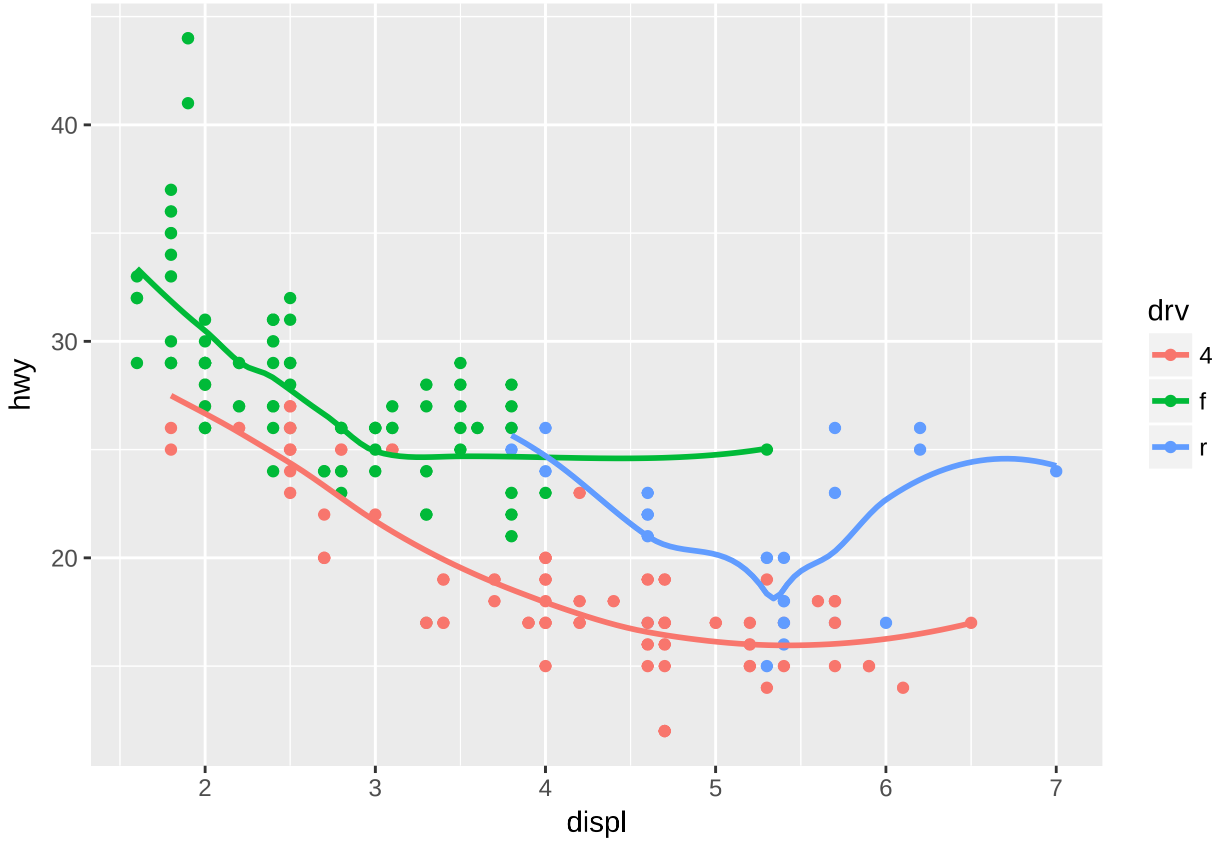
In facet_wrap, nrow defines the number of rows and ncol defines the number of columns of subplots to display. It allows you to break up the subplots into a more digestable display. I.e. if you were to graph the plot above with nrow = 1, you would display the seven categories in one row (which would be the same as plotting ncol = 7.

EXERCISE 3.6.1

1. The following geoms are what i would use to plot different types of charts: line chart = geom_line boxplot = geom_boxplot histogram = geom_histogram area chart = geom_area
2. I would guess that the plot provided in the R for Data Science exercise will plot hwy vs displ first as a scatter plot, overlayed with a second layer of a smooth geom with the drv category broken up up into 3 different line colors and the light-gray range removed:

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

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



The aspect of the graph that i did not foresee was that the scatterplot would also segregate the data by color.

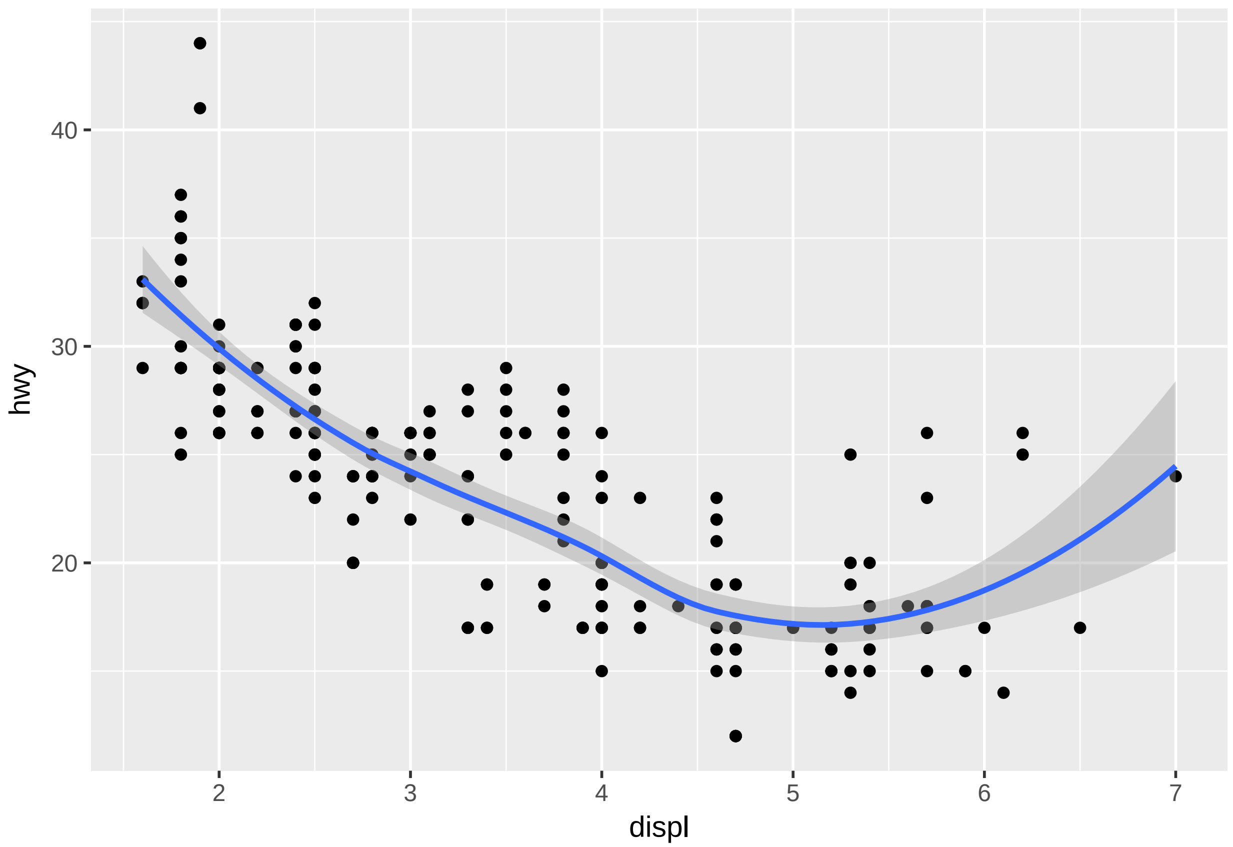
3. show.legend = FALSE removes the legend from the plot. I think that it is used in "R for Data Science" earlier in the chapter because the point of the graphing 3 plots next to each other was not necessarily to understand the data (which the legend provides more information for doing) but to visually compare the graphs. By removing the whitespace and redundant legends, it makes it easier to do so.

4. The se argument (which is provided by default in geom_smooth) provides a gray, semi-transparent confidence interval around the plotted line.

5. The two graphs provided in the exercise should not be any different. The first provides the variables in a global location of the code chunk, whereas the second provides the variables locally which is more redundant.

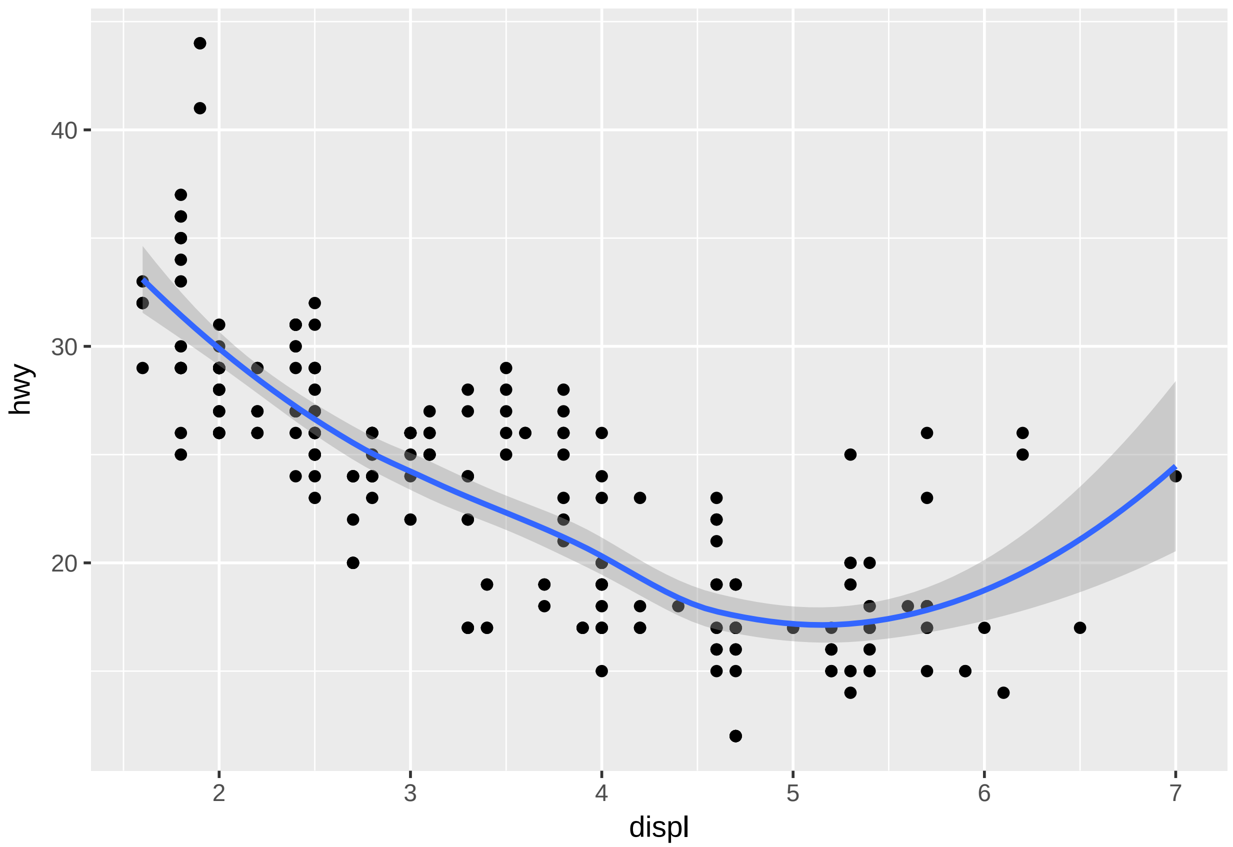
```r
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()
```

```
## `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))
```

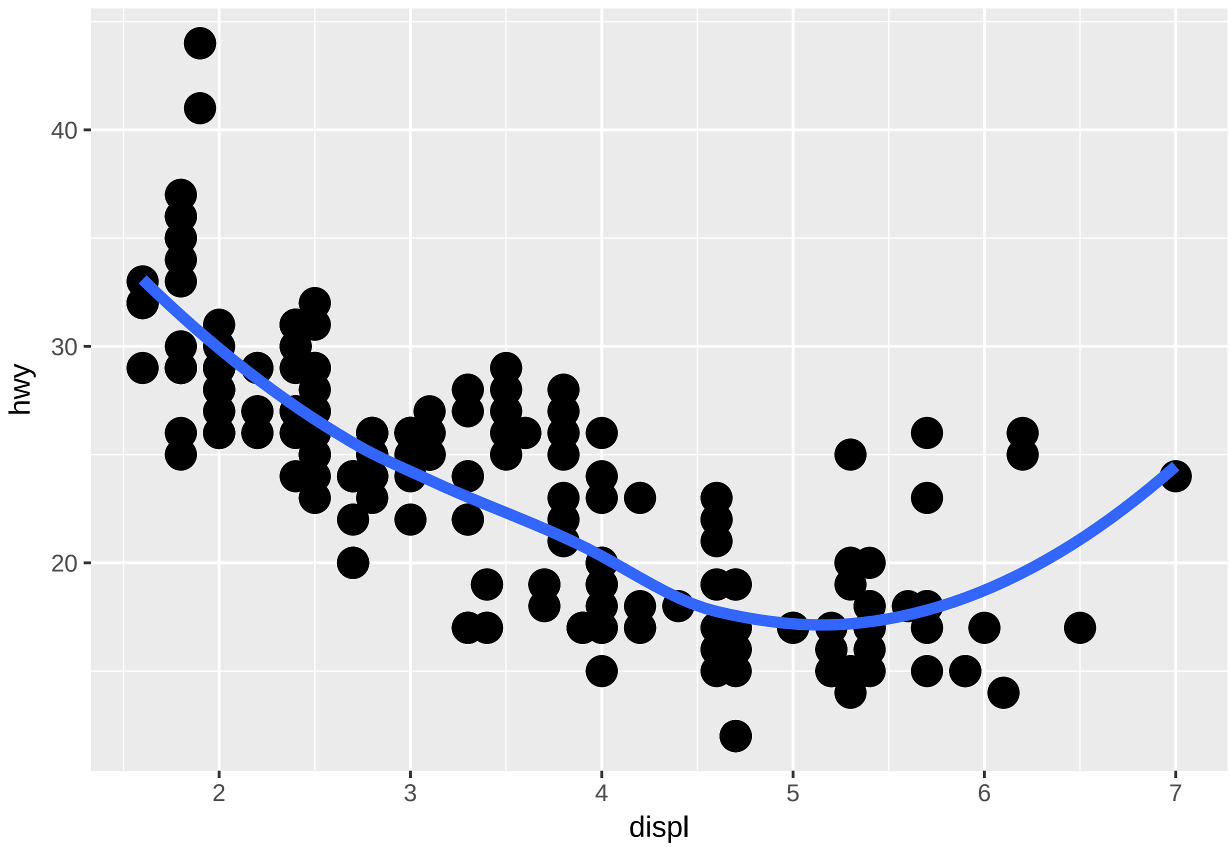
```
## `geom_smooth()` using method = 'loess'
```



6. See graphs below:

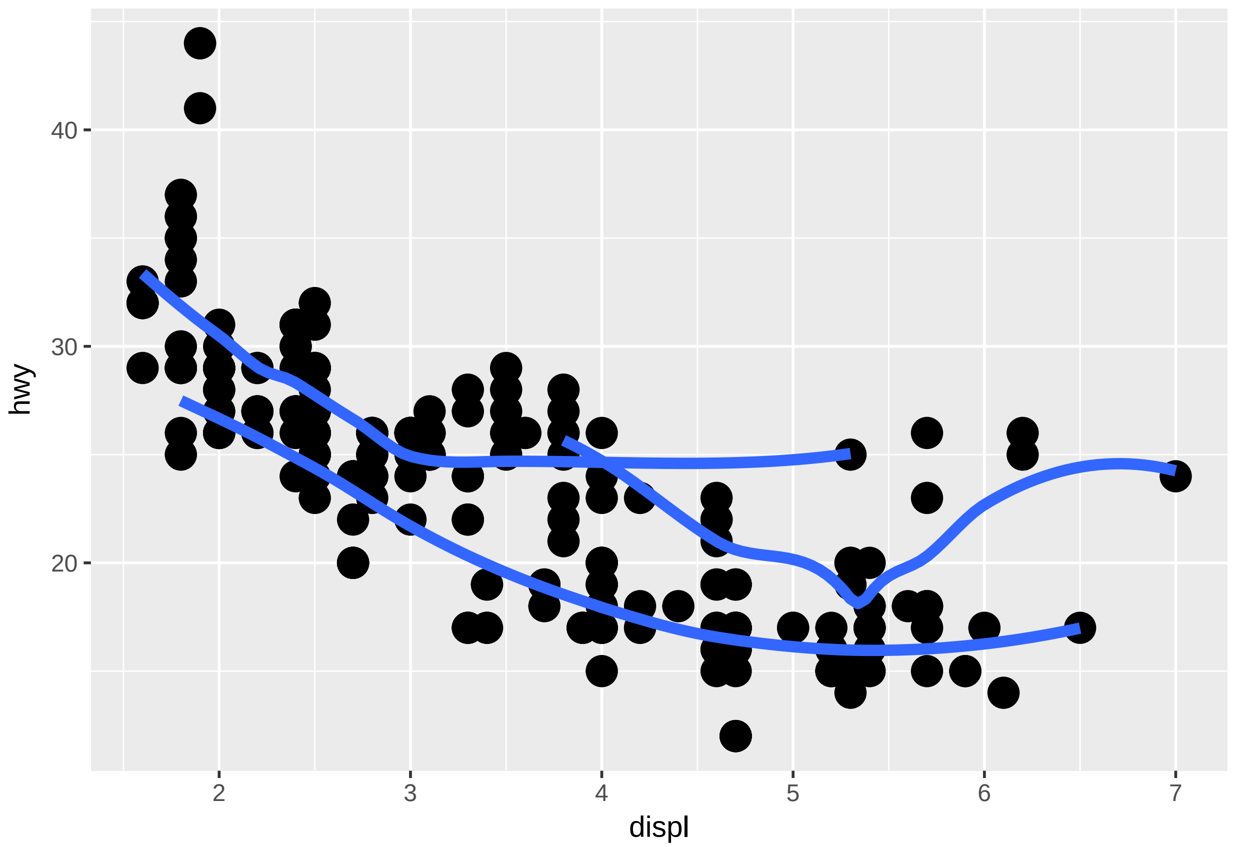
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(size = 5) +
  geom_smooth(se = FALSE, size = 2)
```

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



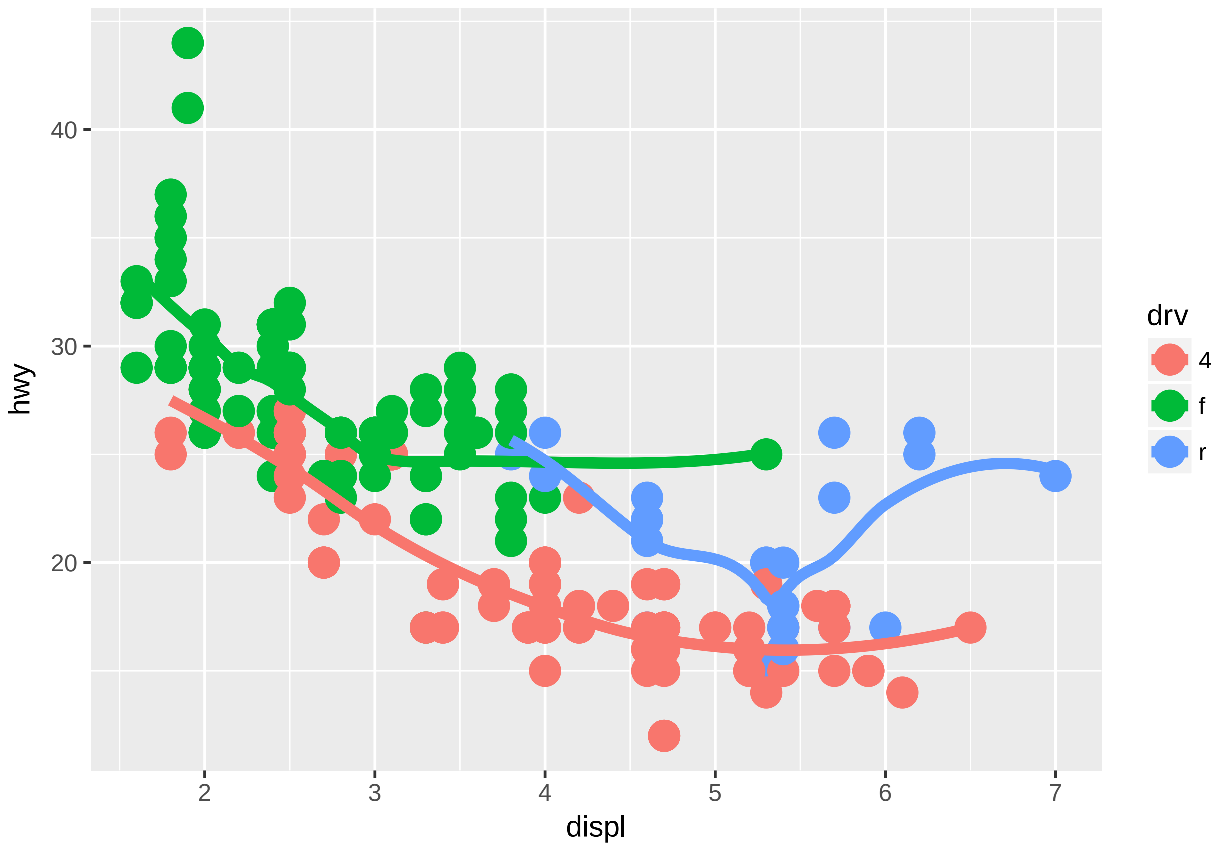
```r
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, group = drv)) +
  geom_point(size = 5) +
  geom_smooth(se = FALSE, size = 2)
```

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



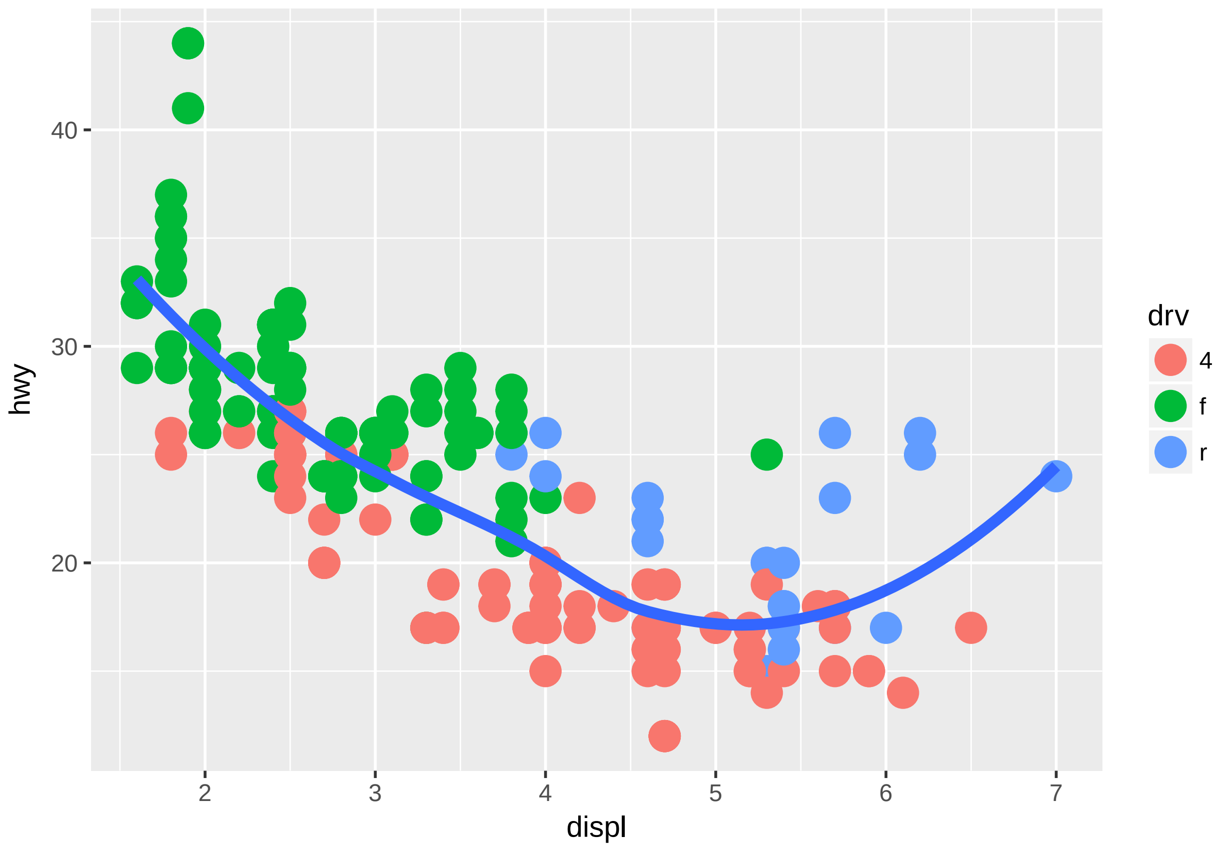
```r
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, group = drv, color = drv)) +
  geom_point(size = 5) +
  geom_smooth(se = FALSE, size = 2)
```

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



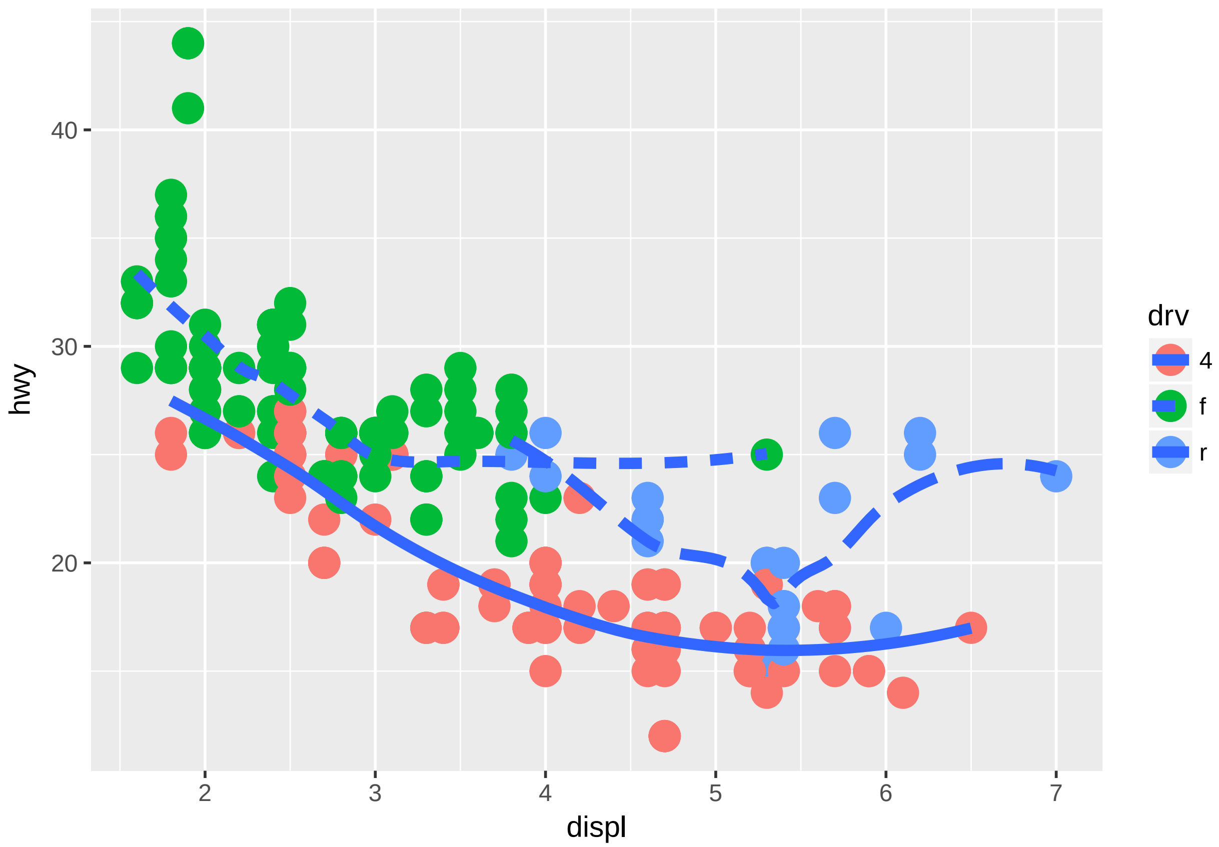
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(mapping = aes(color = drv), size = 5) +
  geom_smooth(se = FALSE, size = 2)
```

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



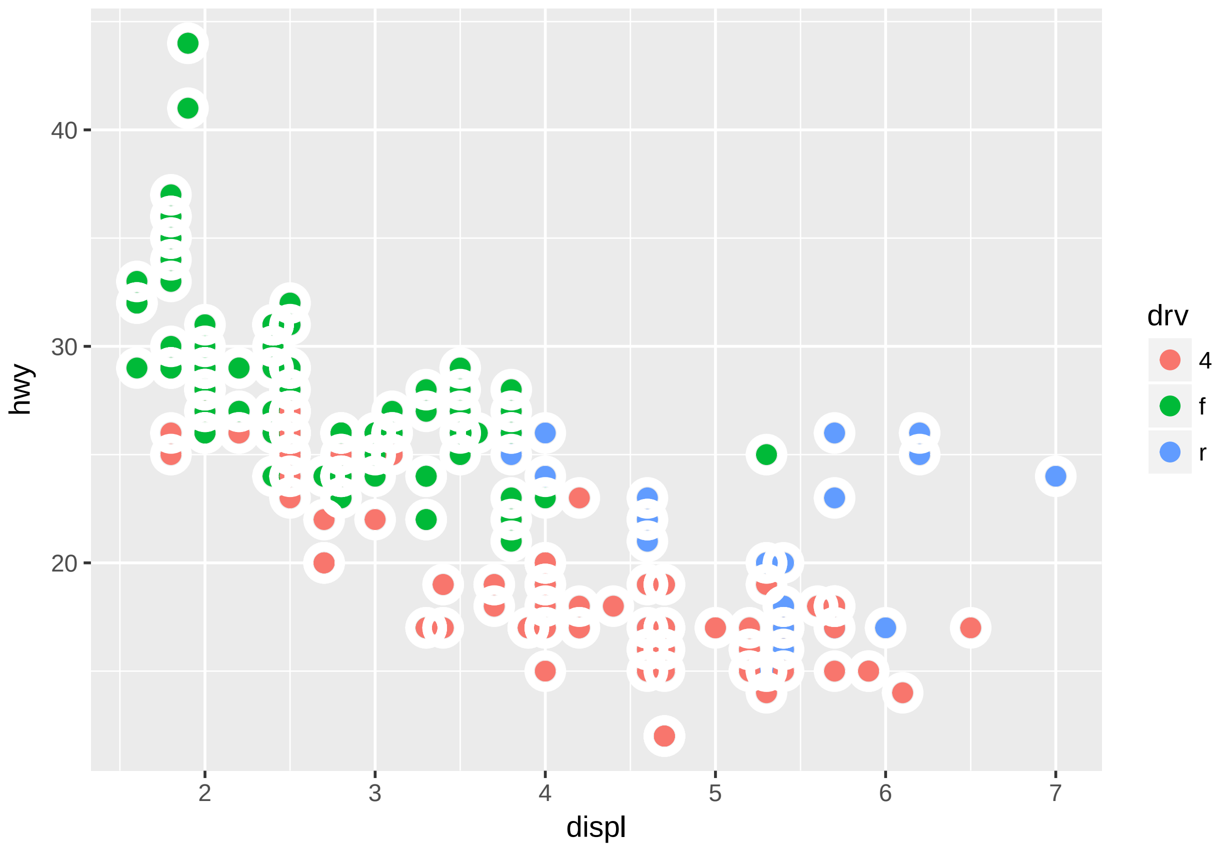
```r
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, group = drv, linetype = drv)) +
  geom_point(size = 5, mapping = aes(color = drv)) +
  geom_smooth(size = 2, se = FALSE)
```

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



Not quite the last one but attempting it!:

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point(size = 3) +
  geom_point(shape = 21, color = 'white', size = 4, stroke = 2.5)
```



EXERCISE 3.7.1

2. geom_bar() makes a bar chart that uses the height of the bars in proportionality to the number of cases in each grouping. It is different from geom_col() in that geom_col() uses the values of the data as the height of the bars instead.

What is a Data Scientist?

Graph 1: The usual challenge in using a donut chart of not being able to fully convey the sizes of each categorical variable are mitigated by the use of called out percentages! However, since the data is ordinal, the use of a linear chart might have been more effective.

Graph 2: I think this is an effective way of communicating where the lion's share of "best new source of data science talent" lies: students studying compuer science. However, if i were to replot this graph i would have made the order of the variables go in an ascending order (smallest on the left) rather than descending (smallest on the right) since we read left to right i think it would lead the reader a bit and garner more engagement in the graph.

Graph 3: I like the use of decreasing opacity in this graph to hammer home the differences between categorical variables (yay alpha!).

Graph 4: The graph is trying to communicate that it is significantly more likely that data scientists have advanced degrees than BI professionals do. I personally think that this graph does a terrible job of doing this. The creator of the graph uses faceting which isolates each category of the categorical variable being plotted (highest level of schooling). However, it took me some time to catch on to the trend of both bins (BI and Data Scientist) as you move from left to right. In my opinion the creator could have overlayed a line graph for each "bin" with a point at the max of each bar and color coordinated it with the bars. That would link the different facet subplots together and draw trend lines. An alternative is to filter out some of the uninformative data. This could be done by only graphing the facet subplots: Masters/Professional Degree, and Doctoral Degree.

Graph 5: I think this graph achieves its purpose in communicating that Business is overwhelmingly the major which BI professionals studied in college. The size of the large circle immediately shows scale.

Graph 6: Again, i like the addition of the percentage call outs on this donut chart. However, I was a bit confused on the descriptions for both of the dichotomous options. It didn't necessarily relate to new technology as the graph title mentioned but to the work that data scientists do.

Graph 7: I think that the characteristics of data scientists are appropriately represented. There is a bit of disconnect for me between the dichotomous options lists and the ordinal bar at the bottom of the graphic. I am not quite sure how this could be overcome though.

Graph 8: Again faceting the graph loses the visual of seeing trends quickly. I found this graphic to be extremely confusing since the percentages of the variable distribtions did not add up to 100%. It is unclear on what the actual data is (time spent, team's reliance on DS/BI, project involvement breakdown, etc.). I would have to know more about the data to be able to better display the graphics but i think that using an area chart might be a good option for this dataset.

Graph 9: Again providing percentages that do not add up to 100% is extremely confusing for me. I also think that a donut plot with called out percentages (assuming they add up to 100%) or a shape size graph (same as graph 2, sorry not sure on the terminology) would be much better to relate the relative involvements of the data science team members.