MPA 634  
Data Science and R for Administrators  
Homework #2

Aesthetic Mappings and Facets

1. Problems 1 – 6 from Exercises 3.3.1

3.3.1:

1.) *What’s wrong with this code? Why are the points not blue?*

The problem with the code is that the color argument is *inside* the aes function, which is why the points aren’t blue. To correctly change the color of the graph, the color needs to be defined *outside* of the aes function. For example:

ggplot(data = mpg) +

geom\_point(mapping = aes(x = displ, y = hwy), color = "blue")

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?*

Categorical Variables: manufacturer, model, year, cyl, trans, drv, fl, class

Continuous Variables: displ, cty, hwy (displ could be continuous since the engine displacement can theoretically be any volume, and doesn’t necessarily *have* to be certain values)

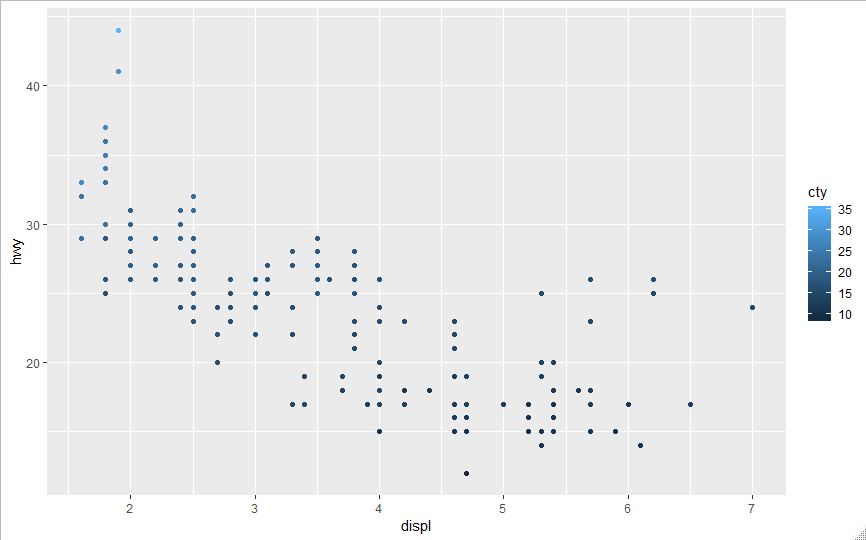
You can see this information by clicking on the data ‘mpg’ in the data window and the table of information will appear. You can also see it in the console when you load the data running the code ‘data(mpg)’.

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

Continuous Variable, cty, to color:

#Map a continuous variable to color

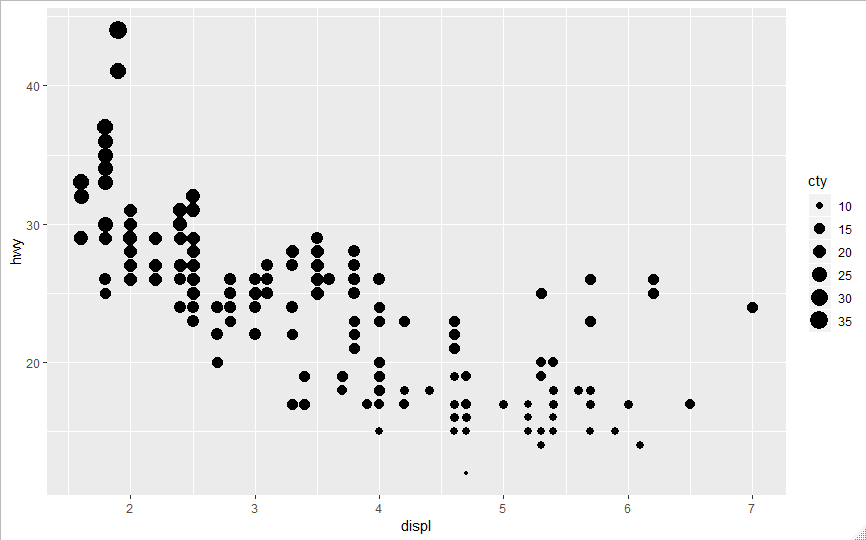
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = cty)) +

 geom\_point()

Continuous Variable, cty, to size:

#Map a continuous variable to size

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

geom\_point()

Continuous Variable, cty, to shape:

#Map a continuous variable to shape

ggplot(data = mpg, mapping = aes(x = displ, y = hwy, shape = cty)) +

geom\_point()

Error: A continuous variable can not be mapped to shape

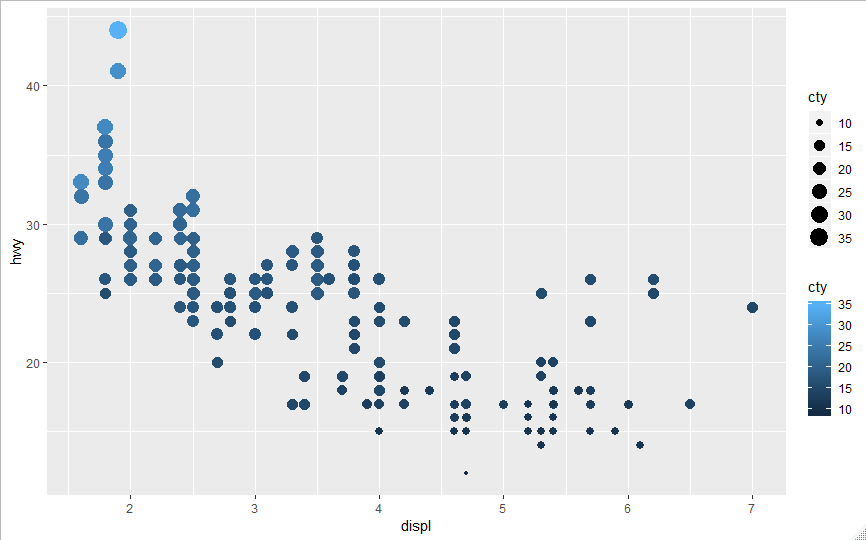
The aesthetics of continuous variables behave differently from the aesthetics of categorical variables because they are not discrete variables. In other words, since there are infinite values possible for a continuous variable, e.g., 1.0, 1.001, 1.000001 etc., the aesthetics behave more as a gradient of colors or sizes etc. to reflect the different potential values of a continuous variable. For example, using cty to define the color for the points in the first graph results in different shades of blue to show the difference in cty values, but if cty was a categorical variable, then entirely different and discrete colors would have been used for the different points. Furthermore, for aesthetics that are not continuous, such as shape, continuous variables cannot be mapped to that aesthetic.

*4.) What happens if you map the same variable to multiple aesthetics?*

Depending on the variable type, the program will either work or not. For example, if one tries to map a continuous variable to an aesthetic that requires a categorical variable, the program will not work. However, if all the variables are of the correct data type, then the plot will feature all the different aesthetics for the same variable. This can be problematic though since, as shown in the example below where cty is mapped to both size and color, the information gained from using the same variable on multiple aesthetics is redundant.

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

geom\_point()



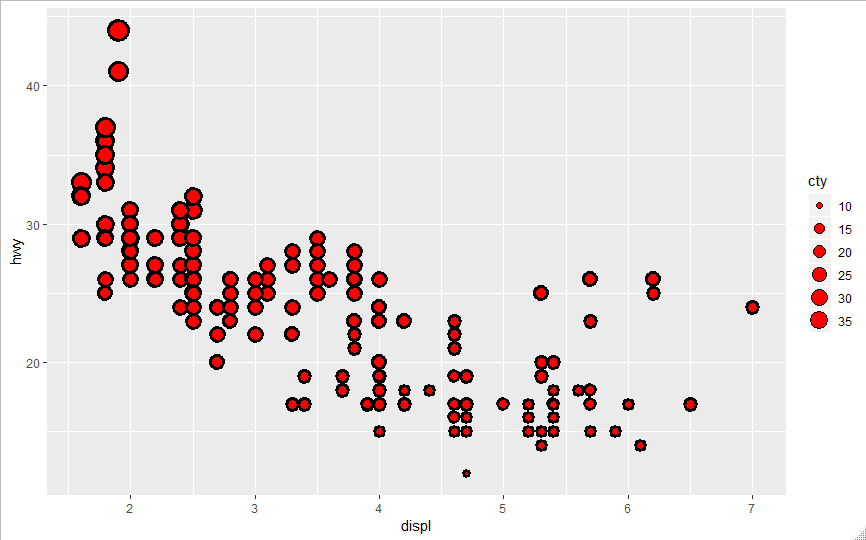
5.) *What does the stroke aesthetic do? What shapes does it work with? (Hint: use ?geom\_point)*

For shapes 21-24, stroke allows you to control the width of the border. An example is shown below.

#Map a continuous variable to size using the stroke aesthetic as well

ggplot(data = mpg, mapping = aes(x = displ, y = hwy, size= cty, stroke = 2)) +

geom\_point(shape = 21, fill = "red")



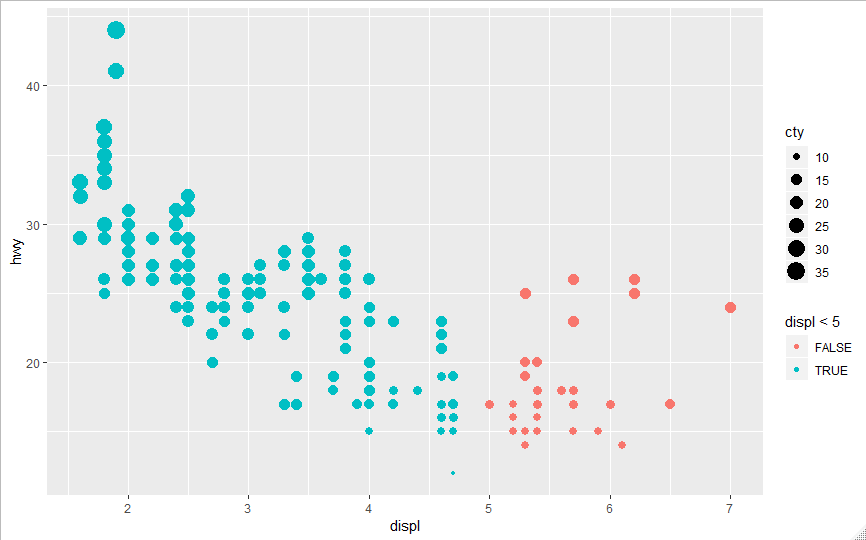
6.) *What happens if you map an aesthetic to something other than a variable name, like aes(colour = displ < 5)? Note, you’ll also need to specify x and y.*

When assigning the inequality “displ < 5” to color, R takes it and splits the data points into two different colors, one color is assigned to all the data points with displ < 5 and the other color for data points with displ >= 5. In other words, it assigns colors based on whether the inequality is satisfied for each data point.

#Map a continuous variable to size and something different to color

ggplot(data = mpg, mapping = aes(x = displ, y = hwy, size= cty, color = displ < 5)) +

geom\_point()



1. Problem 4 from Exercises 3.5.1

3.5.1:

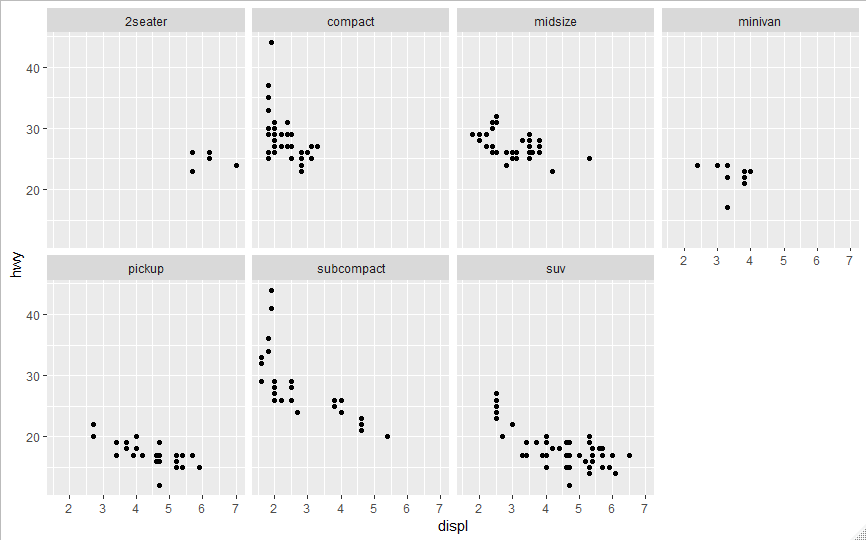
4.) *Take the first faceted plot in this section:*

*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?*



Faceting offers a few advantages vs using the color aesthetic, but also has disadvantages as well. First, one advantage is the clear separation of different types of classes into separate plots. This allows one to easily see different trends for different classes of cars without overlapping data. However, this can also be deemed a disadvantage since then it is visually more difficult to compare these trends (for example a linear fit) between the different classes. In this particular case, where the size of the dataset is not very large, using color to separate the different classes would be acceptable since the single plot would not be overly crowded. Next, using the facets can have the added benefit of using colors to add an additional dimension to the graph. For example, here the different facets are colored by the “trans” variable.

#Use facet

ggplot(data = mpg) +

geom\_point(mapping = aes(x = displ, y = hwy, color = trans)) +

facet\_wrap(~ class, nrow = 2)

If colors were used instead of facet, the additional dimension would have to come from size or something else to avoid a more confusing figure.

For larger datasets, unlike this one, the facet would be much more recommended. This is because the data points would become much too crowded on the plot if color was the only way to distinguish between classes and it would be hard to read/separate them visually. Using facets would break the large datasets into smaller, more manageable chunks.

