Homework-1

##Loading Libraries

knitr::opts\_chunk$set(echo=TRUE, cache=TRUE, fig.asp=0.65, fig.width=5.5)  
library(ggplot2)  
library(tidyverse)  
theme\_set(theme\_bw(base\_size = 14)) # Tested others, put 14 since it is best for your pc  
head(mpg)

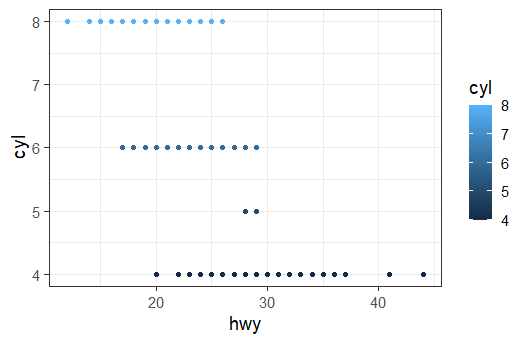
## # A tibble: 6 x 11  
## manufacturer model displ year cyl trans drv cty hwy fl class   
## <chr> <chr> <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>   
## 1 audi a4 1.8 1999 4 auto(l5) f 18 29 p compa~  
## 2 audi a4 1.8 1999 4 manual(m5) f 21 29 p compa~  
## 3 audi a4 2 2008 4 manual(m6) f 20 31 p compa~  
## 4 audi a4 2 2008 4 auto(av) f 21 30 p compa~  
## 5 audi a4 2.8 1999 6 auto(l5) f 16 26 p compa~  
## 6 audi a4 2.8 1999 6 manual(m5) f 18 26 p compa~

### 3.2.4 Exercises

#### (1) 3.2.4 Exercise 4 (10 pts)

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

plot1 <- ggplot(mpg, aes(x = hwy, y = cyl, color = cyl)) +  
 geom\_point()  
  
plot1

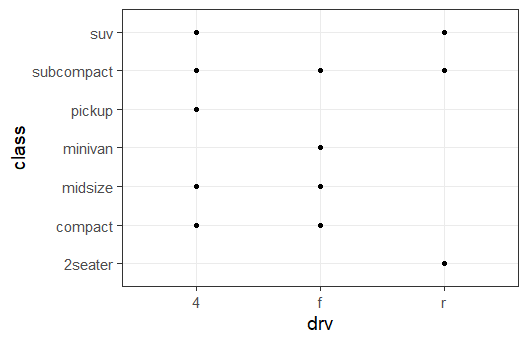


# From this scatter plot it would be reasonable to assume negative correlation between the amount of cylinders and the highway miles per gallon.

#### (2) 3.2.4 Exercise 5 (10 pts)

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

plot2 <- ggplot(mpg, aes(x = drv, y = class)) +  
 geom\_point()  
  
plot2



# This plot can be useful when determining the all the possible different drive trains of each type of vehicle, but the information could be given in a better way other than a scatter plot.

### 3.3.1 Exercises

#### (3) 3.3.1 Exercise 2 (10 pts)

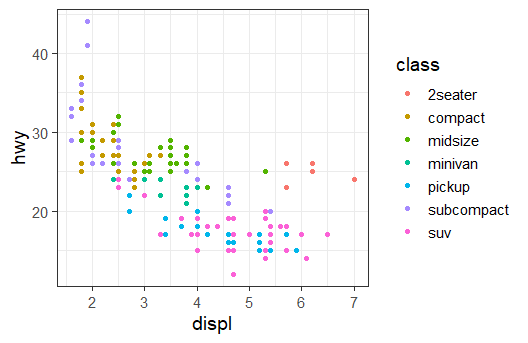
Which variables in mpg are categorical? Which are continuous? (Hint: type ?mpg to read the documentation for the dataset). How can you see this information when you run mpg?

# The categorical variables are: manufacturer, model, cyl, trans, drv, fl, class  
# The continuous variables are: displ, year, cty, hwy  
  
# The function below lets us check the variable type, knowing that integers and numerical based data types are the only data type that allows continuous variables we can check those data types for all the continuous variables. Leaving all other variables to be discrete.  
str(mpg)

## tibble [234 x 11] (S3: tbl\_df/tbl/data.frame)  
## $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...  
## $ model : chr [1:234] "a4" "a4" "a4" "a4" ...  
## $ displ : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...  
## $ year : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...  
## $ cyl : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...  
## $ trans : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...  
## $ drv : chr [1:234] "f" "f" "f" "f" ...  
## $ cty : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...  
## $ hwy : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...  
## $ fl : chr [1:234] "p" "p" "p" "p" ...  
## $ class : chr [1:234] "compact" "compact" "compact" "compact" ...

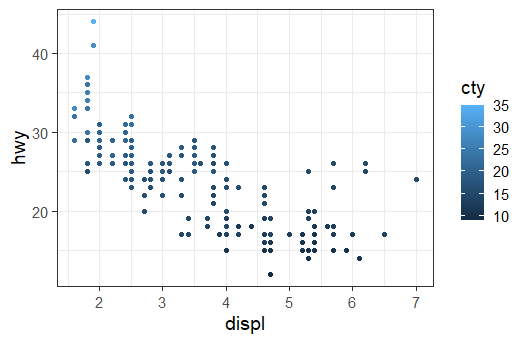
#### (4) 3.3.1 Exercise 3 (10 pts)

Consider the scatterplot of displacement vs miles per gallon that we have been studying:

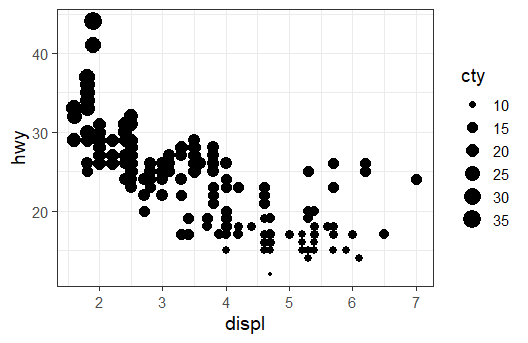


Map a continuous variable to color, size, and shape, and show the resulting graphs. How do these aesthetics behave differently for categorical vs. continuous variables?

color1 <- ggplot(data = mpg) +  
 geom\_point(mapping = aes(x = displ, y = hwy, color = cty))  
  
size1 <- ggplot(data = mpg) +  
 geom\_point(mapping = aes(x = displ, y = hwy, size = cty))  
  
shape1 <- ggplot(data = mpg) +  
 geom\_point(mapping = aes(x = displ, y = hwy, shape = cty))  
  
  
color1 #Color operates as more of a gradient of a single color rather than several different colors



size1 #Size fits the continuous variable more since the changes in size fit more implies a relationship which discrete variables might not have.

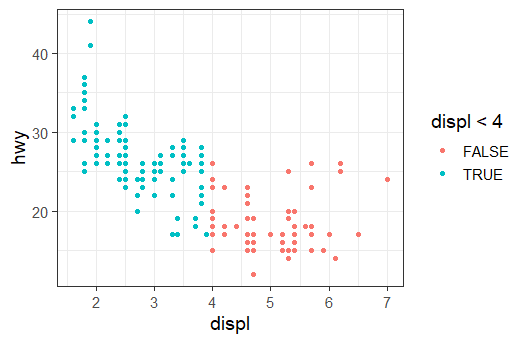


#shape1 #Does not work, since a shape cannot be properly scaled for a continous variable

#### (5) 3.3.1 Exercise 6 (10 pts)

What happens if you map an aesthetic to something other than a variable name, like aes(color = displ < 4)? Note that you will also need to specify and .

ggplot(data = mpg) +  
 geom\_point(mapping = aes(x = displ, y = hwy, color = displ < 4))



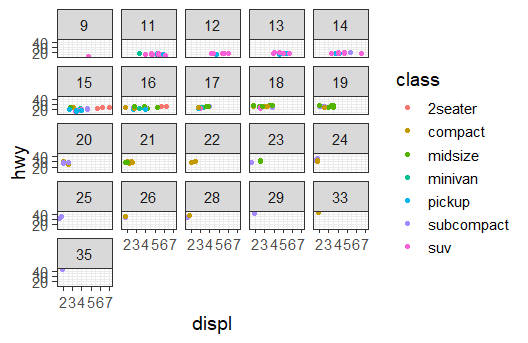
# Two separate colors will represent the results, either true or false.

### 3.5.1 Exercises

#### (6) 3.5.1 Exercise 1 (5 pts)

What happens if you facet on a continuous variable? Show an example. Discuss when this might be useful, and when it is likely not useful.

ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = hwy, color = class)) +  
 facet\_wrap(facets = vars(cty))

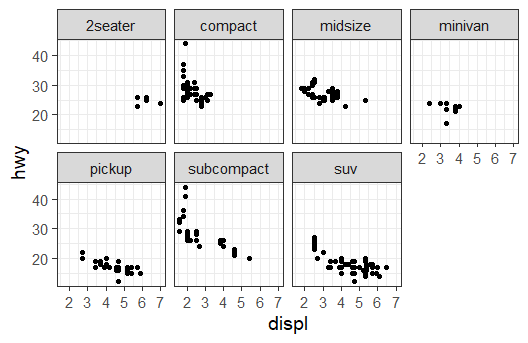


# Several facets pop up for every option, this can be useful when a researcher wants to deep dive on the effect a continuous variable has. However, it is not very useful outside of that since it is a lot of information and hard to digest. Plus it will have a hard time loading all that data in a readable fashion.

#### (7) 3.5.1 Exercise 4 (10 pts)

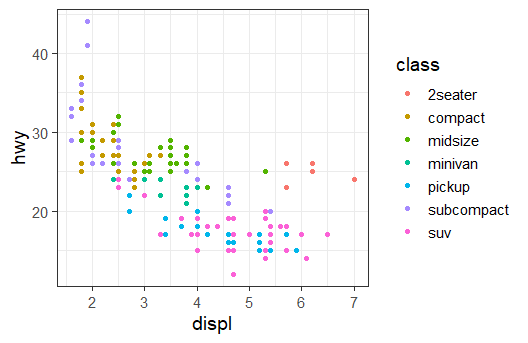
Consider the first faceted plot in this section:

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



Create a comparable graph that uses the color aesthetic. 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?

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



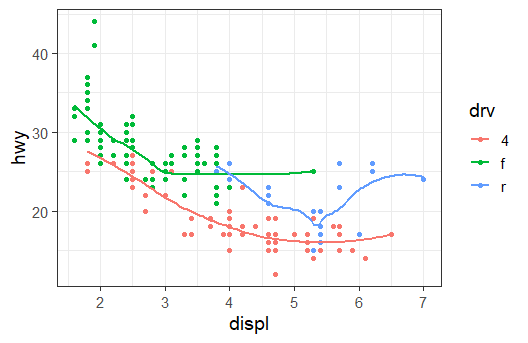
# The advantage of using faceting is that it is much easier to compare how each class interacts with the given variables. The disadvantages would come around if a researcher is comparing each class to itself over the given variables. Colors is more advantageous the more unique classes in the variable since it is easier to compare, despite losing the ability to clearly see patterns.

### 3.6.1 Exercises

#### (8) 3.6.1 Exercises 2 (10 pts)

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

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



# Predictions, a scatter plot and a line graph will appear in the same space and will have the same variables and colors.   
  
# My prediction proved to be true, since each "geom" line of code states the type of graph, each graph will follow the same parameters labeled above.

#### (9) 3.6.1 Exercises 4 (5 pts)

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

# It displays the confidence interval around the machine learning technique used to calculate the smooth line.

#### (10) 3.6.1 Exercises 5 (5 pts)

Will these two graphs look different? Why or why not?

ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +   
 geom\_point() +   
 geom\_smooth()  
  
ggplot() +   
 geom\_point(data = mpg, mapping = aes(x = displ, y = hwy)) +   
 geom\_smooth(data = mpg, mapping = aes(x = displ, y = hwy))  
  
# No since both are saying the same thing. However, the second one is not allowed in any of the 50 states since it rewrites the exact same labeling twice.

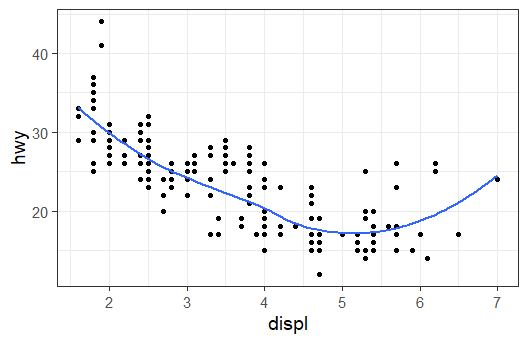
#### 3.6.1 Exercises 6

Recreate the R code needed to generate each of these graphs:

##### (11) (6.1, 5 pts)

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

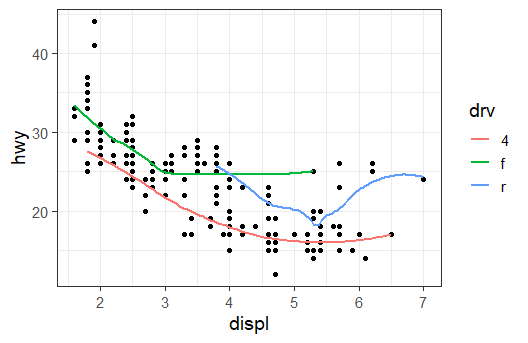
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



##### (12) (6.2, 5 pts)

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

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

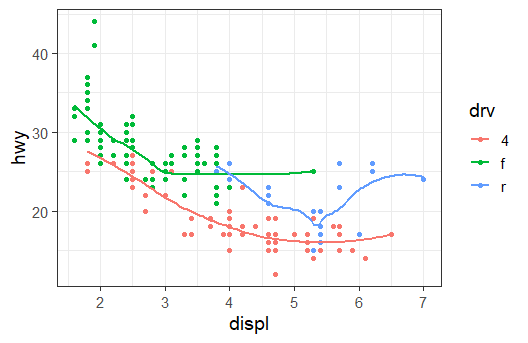


# Could not find an aes function that would split the data based on a variable without using color

##### (13) (6.3, 5 pts)

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

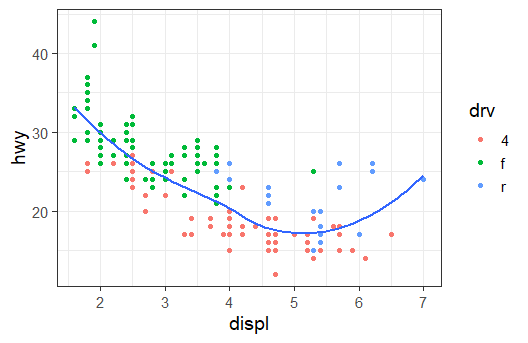
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



##### (14) (6.4, 5 pts)

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

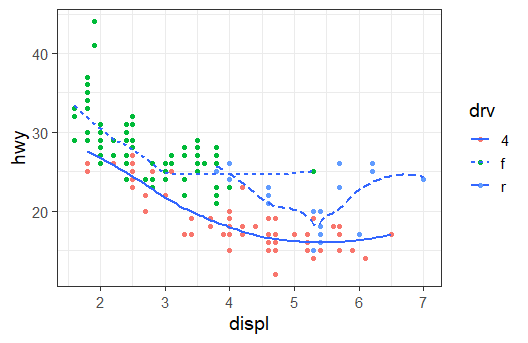
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



##### (15) (6.5, 5 pts)

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' and formula 'y ~ x'



##### (16) (6.6, 5 pts)

ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +   
 geom\_point(size = 3.5, color = "grey") + #White was difficult to see  
 geom\_point(aes(color = drv))

