DATA 612 Practice Test

Andrew Erickson (DATA 612 Thurs)

9/30/2021

## All comments, code, and analysis included below:

### Setup:

#Load in Tidyverse:   
  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

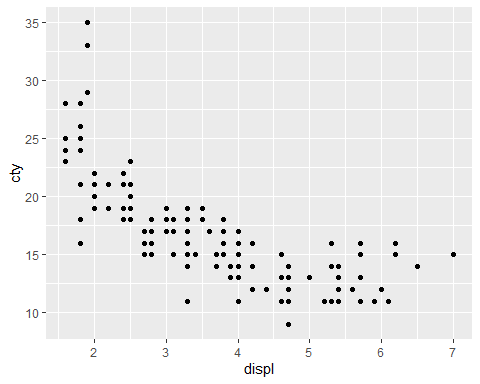
## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.2 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

## Part 1: GGPLOT PRACTICE:

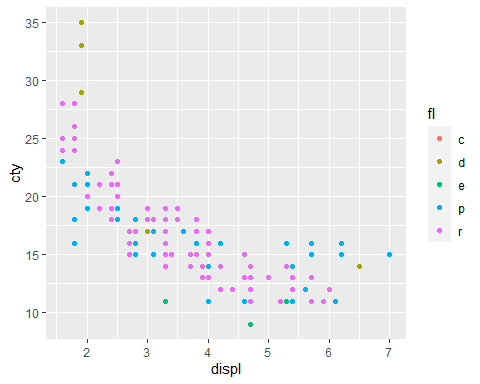
### Question 1:

#Simple scatter plot with explanatory variable displacement and response variable city miles per gallon:  
ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = cty))



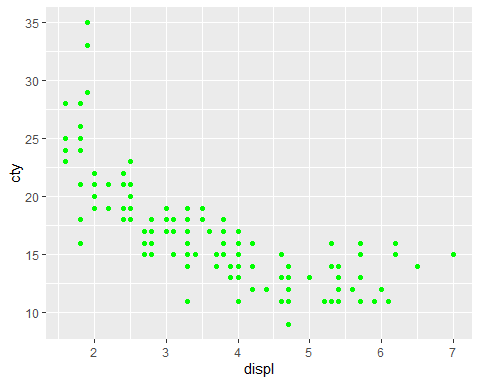
### Question 2:

#Scatter Plot with displ as x and cty as y, with fl as the color argument:   
  
ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = cty, color = fl))



### Question 3:

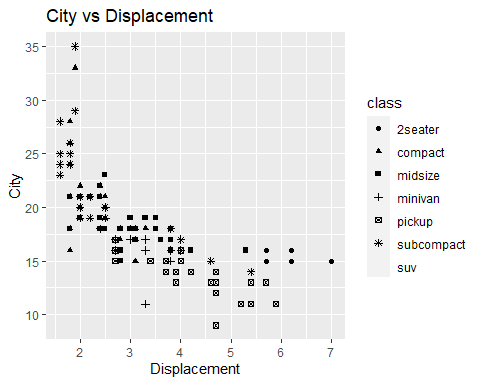
#Scatter plot showing displ as x, cty as y, with green scatter plot points:   
ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = cty), color = "green")

 ### Question 4:

#using the shape argument by class; x = displ, y = cty  
ggplot(data = mpg) +  
 geom\_point(mapping = aes(x = displ, y = cty, shape = class)) +   
 ggtitle("City vs Displacement") +  
 ylab("City") +   
 xlab("Displacement")

## Warning: The shape palette can deal with a maximum of 6 discrete values because  
## more than 6 becomes difficult to discriminate; you have 7. Consider  
## specifying shapes manually if you must have them.

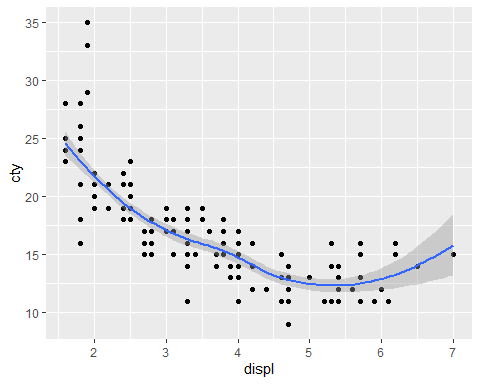
## Warning: Removed 62 rows containing missing values (geom\_point).



### Question 5:

#Smooth line fitted over scatter plot; here, we are keeping error bands:   
  
ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = cty)) +   
 geom\_smooth(mapping = aes(x = displ, y = cty))

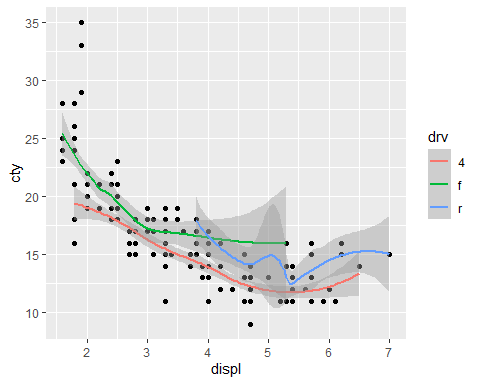
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



### Question 6:

#To create this effect of varying by drv, we include drv in the group parameter in both geom\_point and geom\_smooth:   
  
ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = cty)) +   
 geom\_smooth(mapping = aes(x = displ, y = cty, color = drv))

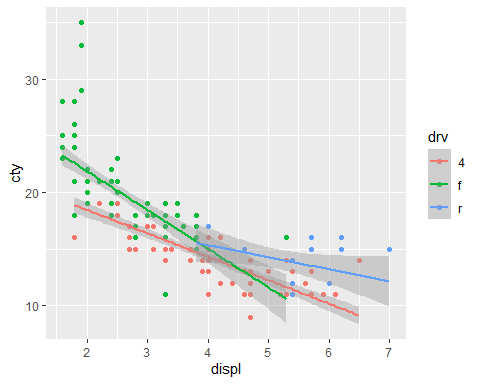
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



### Question 7:

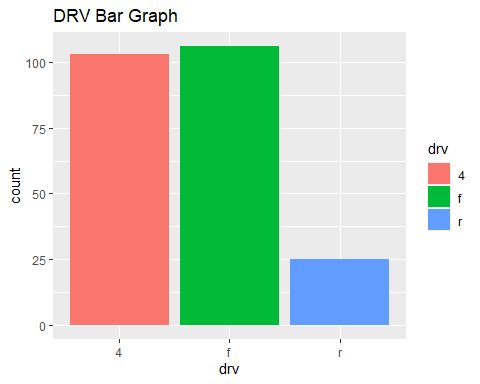
#Scatter plot with overlaid regression lines based on drv levels   
ggplot(data = mpg) +  
 geom\_point(mapping = aes(x = displ, y = cty, color = drv)) +   
 geom\_smooth(mapping = aes(x = displ, y = cty, color = drv), method = "lm")

## `geom\_smooth()` using formula 'y ~ x'



### Question 8:

ggplot(data = mpg) +   
 geom\_bar(mapping = aes(x = drv, fill = drv)) +   
 ggtitle("DRV Bar Graph")



### Question 9:

#We have been instructed to skip this question for now.

### Question 10:

#We have been instructed to skip this question for now.

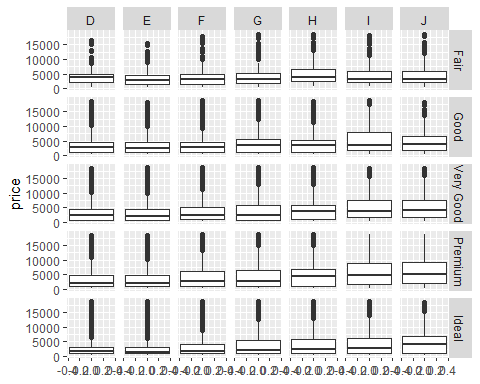
### Question 11:

ggplot(data = diamonds) +   
 geom\_boxplot(mapping = aes(x = cut, y = price))

### 

### Question 12:

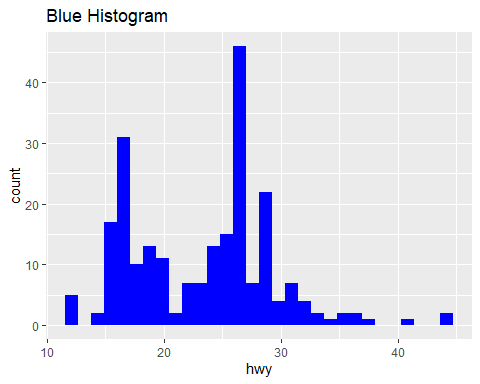
#Facet grid of boxplots, varied by cut and color:   
ggplot(data = diamonds) +   
 geom\_boxplot(mapping = aes(y = price)) +   
 facet\_grid(cut ~ color)



### Question 13:

#Histogram colored blue, variable is hwy:  
  
ggplot(data = mpg) +   
 geom\_histogram(mapping = aes(x = hwy), fill = "blue") +   
 ggtitle("Blue Histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## Part II: Dplyr Practice

### Question 1:

mpg\_q1 <- mpg %>%  
 filter(manufacturer == "dodge", model == "durango 4wd")  
  
#To check that the filtering occurred properly:   
  
head(mpg\_q1)

## # 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 dodge durango ~ 3.9 1999 6 auto(l~ 4 13 17 r suv   
## 2 dodge durango ~ 4.7 2008 8 auto(l~ 4 13 17 r suv   
## 3 dodge durango ~ 4.7 2008 8 auto(l~ 4 9 12 e suv   
## 4 dodge durango ~ 4.7 2008 8 auto(l~ 4 13 17 r suv   
## 5 dodge durango ~ 5.2 1999 8 auto(l~ 4 11 16 r suv   
## 6 dodge durango ~ 5.7 2008 8 auto(l~ 4 13 18 r suv

### Question 2:

mpg\_q2 <- mpg %>%  
 filter(cty < 10, hwy < 16)   
  
#To show that our filtered data set is displaying properly   
head(mpg\_q2)

## # A tibble: 5 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 dodge dakota pi~ 4.7 2008 8 auto(~ 4 9 12 e pick~  
## 2 dodge durango 4~ 4.7 2008 8 auto(~ 4 9 12 e suv   
## 3 dodge ram 1500 ~ 4.7 2008 8 auto(~ 4 9 12 e pick~  
## 4 dodge ram 1500 ~ 4.7 2008 8 manua~ 4 9 12 e pick~  
## 5 jeep grand che~ 4.7 2008 8 auto(~ 4 9 12 e suv

### Question 3:

mpg\_q3 <- mpg %>%  
 arrange(desc(displ))  
  
#To show that displacement is arranged in descending order  
head(mpg\_q3)

## # 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 chevrolet corvette 7 2008 8 manual~ r 15 24 p 2sea~  
## 2 chevrolet k1500 ta~ 6.5 1999 8 auto(l~ 4 14 17 d suv   
## 3 chevrolet corvette 6.2 2008 8 manual~ r 16 26 p 2sea~  
## 4 chevrolet corvette 6.2 2008 8 auto(s~ r 15 25 p 2sea~  
## 5 jeep grand ch~ 6.1 2008 8 auto(l~ 4 11 14 p suv   
## 6 chevrolet c1500 su~ 6 2008 8 auto(l~ r 12 17 r suv

### Question 4:

mpg\_q4 <- mpg %>%  
 arrange(cty) %>%   
 print(n = 30)

## # A tibble: 234 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 dodge dakota p~ 4.7 2008 8 auto(~ 4 9 12 e pick~  
## 2 dodge durango ~ 4.7 2008 8 auto(~ 4 9 12 e suv   
## 3 dodge ram 1500~ 4.7 2008 8 auto(~ 4 9 12 e pick~  
## 4 dodge ram 1500~ 4.7 2008 8 manua~ 4 9 12 e pick~  
## 5 jeep grand ch~ 4.7 2008 8 auto(~ 4 9 12 e suv   
## 6 chevrolet c1500 su~ 5.3 2008 8 auto(~ r 11 15 e suv   
## 7 chevrolet k1500 ta~ 5.3 2008 8 auto(~ 4 11 14 e suv   
## 8 chevrolet k1500 ta~ 5.7 1999 8 auto(~ 4 11 15 r suv   
## 9 dodge caravan ~ 3.3 2008 6 auto(~ f 11 17 e mini~  
## 10 dodge dakota p~ 5.2 1999 8 manua~ 4 11 17 r pick~  
## 11 dodge dakota p~ 5.2 1999 8 auto(~ 4 11 15 r pick~  
## 12 dodge durango ~ 5.2 1999 8 auto(~ 4 11 16 r suv   
## 13 dodge durango ~ 5.9 1999 8 auto(~ 4 11 15 r suv   
## 14 dodge ram 1500~ 5.2 1999 8 auto(~ 4 11 15 r pick~  
## 15 dodge ram 1500~ 5.2 1999 8 manua~ 4 11 16 r pick~  
## 16 dodge ram 1500~ 5.9 1999 8 auto(~ 4 11 15 r pick~  
## 17 ford expediti~ 4.6 1999 8 auto(~ r 11 17 r suv   
## 18 ford expediti~ 5.4 1999 8 auto(~ r 11 17 r suv   
## 19 ford f150 pic~ 5.4 1999 8 auto(~ 4 11 15 r pick~  
## 20 jeep grand ch~ 6.1 2008 8 auto(~ 4 11 14 p suv   
## 21 land rover range ro~ 4 1999 8 auto(~ 4 11 15 p suv   
## 22 land rover range ro~ 4.6 1999 8 auto(~ 4 11 15 p suv   
## 23 lincoln navigato~ 5.4 1999 8 auto(~ r 11 17 r suv   
## 24 lincoln navigato~ 5.4 1999 8 auto(~ r 11 16 p suv   
## 25 toyota land cru~ 4.7 1999 8 auto(~ 4 11 15 r suv   
## 26 chevrolet c1500 su~ 6 2008 8 auto(~ r 12 17 r suv   
## 27 dodge ram 1500~ 4.7 2008 8 manua~ 4 12 16 r pick~  
## 28 dodge ram 1500~ 4.7 2008 8 manua~ 4 12 16 r pick~  
## 29 ford expediti~ 5.4 2008 8 auto(~ r 12 18 r suv   
## 30 land rover range ro~ 4.2 2008 8 auto(~ 4 12 18 r suv   
## # ... with 204 more rows

**ANSWER** : After arranging the city miles per gallon column in ascending order, we see that five vehicles–Dodge Dakota Pickup 4WD, Dodge Durango 4WD, Dodge Ram 1500 Pickup 4WD (automatic), Dodge Ram 1500 Pickup 4WD (manual), and Jeep Grand Cherokee 4WD – are tied for the lowest city miles per gallon with 9.

### Question 5:

mpg\_q5 <- mpg %>%  
 filter(manufacturer == "ford", year == 1999, cty < 16, hwy < 16) %>%  
 select(manufacturer, model)  
  
#To ensure our chunk of code worked:   
head(mpg\_q5)

## # A tibble: 1 x 2  
## manufacturer model   
## <chr> <chr>   
## 1 ford f150 pickup 4wd

### Question 6:

mpg\_q6 <- mpg %>%  
 filter(manufacturer == "ford", model == "mustang") %>%  
 mutate(difference = hwy - cty) %>%  
 select(manufacturer, model, cty, hwy, difference)   
  
#To display a few rows of our modified data set:   
head(mpg\_q6)

## # A tibble: 6 x 5  
## manufacturer model cty hwy difference  
## <chr> <chr> <int> <int> <int>  
## 1 ford mustang 18 26 8  
## 2 ford mustang 18 25 7  
## 3 ford mustang 17 26 9  
## 4 ford mustang 16 24 8  
## 5 ford mustang 15 21 6  
## 6 ford mustang 15 22 7

### Question 7:

#First, we load in the nycflights13 data set:   
library(nycflights13)

## Warning: package 'nycflights13' was built under R version 4.1.1

#To show just a portion of that data set:   
head(flights)

## # A tibble: 6 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int> <int>  
## 1 2013 1 1 517 515 2 830 819  
## 2 2013 1 1 533 529 4 850 830  
## 3 2013 1 1 542 540 2 923 850  
## 4 2013 1 1 544 545 -1 1004 1022  
## 5 2013 1 1 554 600 -6 812 837  
## 6 2013 1 1 554 558 -4 740 728  
## # ... with 11 more variables: arr\_delay <dbl>, carrier <chr>, flight <int>,  
## # tailnum <chr>, origin <chr>, dest <chr>, air\_time <dbl>, distance <dbl>,  
## # hour <dbl>, minute <dbl>, time\_hour <dttm>

#We are hashing out the help function, but I did explore the data frame  
#?flights()  
  
#We are similarly hashing out View to allow this rmd file to knit:  
#View(flights)  
  
flights\_q7 <- flights %>%  
 select(carrier, year, month, day, arr\_delay) %>%  
 filter(carrier == "AA", year == 2013, month == 3, day == 17) %>%  
 arrange(desc(arr\_delay))  
  
#To ensure our modified data frame is displaying properly:  
head(flights\_q7)

## # A tibble: 6 x 5  
## carrier year month day arr\_delay  
## <chr> <int> <int> <int> <dbl>  
## 1 AA 2013 3 17 67  
## 2 AA 2013 3 17 39  
## 3 AA 2013 3 17 39  
## 4 AA 2013 3 17 36  
## 5 AA 2013 3 17 33  
## 6 AA 2013 3 17 22

### Question 8:

mpg\_q8\_dplyr <- mpg %>%   
 filter(year == 1999) %>%  
 group\_by(manufacturer) %>%  
 summarise(mean\_cty\_mpg = mean(cty, na.rm = TRUE)) %>%  
 arrange(desc(mean\_cty\_mpg))  
  
#To check to see which manufacturer has the highest mpg for 1999:   
head(mpg\_q8\_dplyr)

## # A tibble: 6 x 2  
## manufacturer mean\_cty\_mpg  
## <chr> <dbl>  
## 1 honda 24.8  
## 2 volkswagen 21.2  
## 3 subaru 19   
## 4 hyundai 18.3  
## 5 toyota 18.2  
## 6 nissan 17.7

**ANSWER**: AT 24.8 city miles per gallon, Honda had the best average gas mileage in 1999.