# Module 3 Assignment 2

## Graybill, Gil

### Exploratory Data Analysis

#### Diamond Color and Price

# bring in the library  
library(tidyverse)

## -- Attaching packages ------------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

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

ggplot(data = diamonds, aes(x = carat, y = price))+  
 geom\_point()



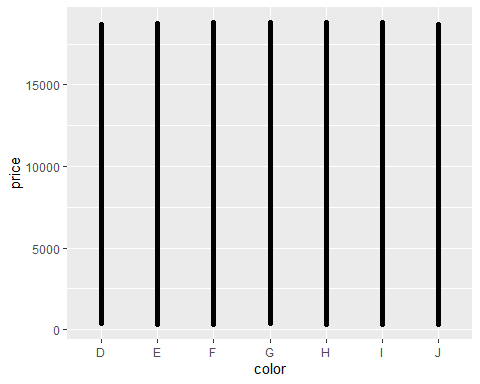
1.) What do you notice from the scatterplot as the carat size increases?

ANSWER: The price increases as the carat size increases.

2.) From the scatterplot, what carats are most represented within the diamonds dataset?

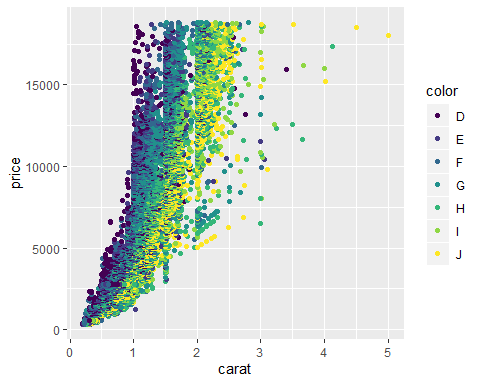
ANSWER: There seem to be more diamonds clustered to the right, or positive side, of a whole carat boundary.

ggplot(data = diamonds, aes(x = color, y = price))+  
 geom\_point()



From this graph it’s hard to see any difference in price between the colors. Since there are 50,000+ entries and we’re comparing categorical data, the points are getting clustered.

ggplot(data = diamonds, aes(x = carat, y = price, color = color))+  
 geom\_point()



1.) Does color impact the price?

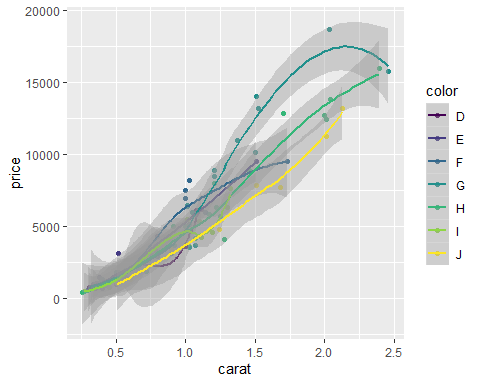
ANSWER: It looks like the color represented by D are more expensive than the others, with the color represented by J being the least expensive.

2.) Are certain colors associated with carat size? Provide an example.

ANSWER: H, I, and J look tend to be larger carat sizes, with D being smaller carat sizes.

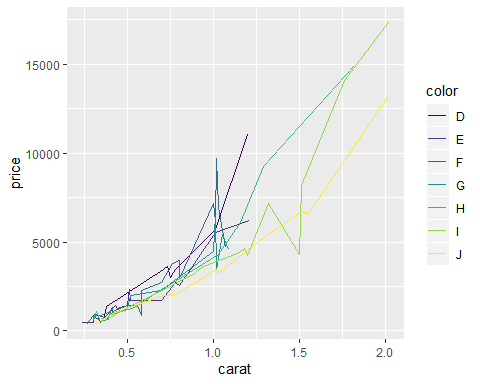
dsample <- diamonds[sample(nrow(diamonds), 100), ]  
ggplot(data = dsample, aes(x = carat, y = price, color = color))+  
 geom\_point()+  
 geom\_smooth()

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



With a sample size of 100 it’s easier to see that there is a relationship between color and price. D seems to be the most expensive per carat, but they don’t tend to be bigger than 1 carat. E is the next most expensive, and it looks like they go up to 1.5 carats. At sizes of less than 0.5 carats it’s hard to tell a difference between price and color.

dsample <- diamonds[sample(nrow(diamonds), 100), ]  
ggplot(data = dsample, aes(x = carat, y = price, color = color))+  
 geom\_line(mapping = aes())



It definately looks like the color represented by J is the least expensive and that D is the most expensive on a per carat basis.