# Module 3 - Assignment 2

## Lugo, Alex

### Exploratory Data Analysis

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()

diamonds

## # A tibble: 53,940 x 10  
## carat cut color clarity depth table price x y z  
## <dbl> <ord> <ord> <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  
## 2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31  
## 3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31  
## 4 0.290 Premium I VS2 62.4 58 334 4.2 4.23 2.63  
## 5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75  
## 6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48  
## 7 0.24 Very Good I VVS1 62.3 57 336 3.95 3.98 2.47  
## 8 0.26 Very Good H SI1 61.9 55 337 4.07 4.11 2.53  
## 9 0.22 Fair E VS2 65.1 61 337 3.87 3.78 2.49  
## 10 0.23 Very Good H VS1 59.4 61 338 4 4.05 2.39  
## # ... with 53,930 more rows

#### Diamond Color and Price

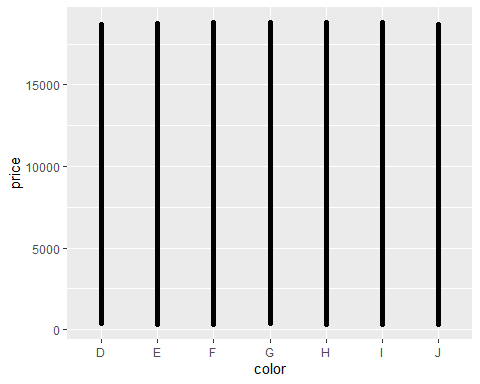
ggplot(diamonds, aes(carat, price))+  
 geom\_point()



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

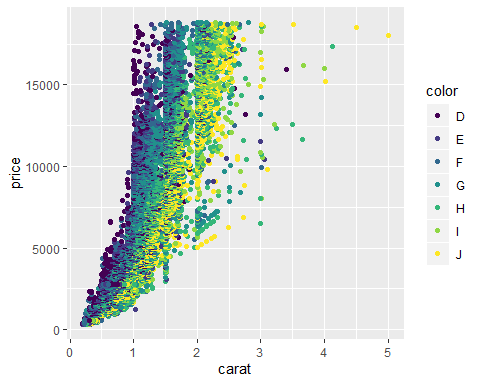
2.) From the scatterplot, what carats are most represented within the diamonds dataset? 1-1.5 contain the most points in the dataset.

ggplot(diamonds, aes(color, price))+  
 geom\_point()



The chart above is not very useful. There are points at just about every price for each color making the scatterpoint look like a bar graph. There seems to be no difference on price based soley on color.

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

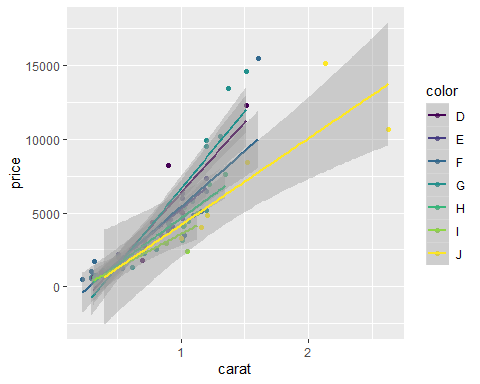


The chart above can now help use answer the questions below.

1.) Does color impact the price? It does seem color has a slight impact on price. We can see this mostly between 0-2 carats. Color “J,” which are the yellow points, tend to be less expensive than their “D” colored counterparts.

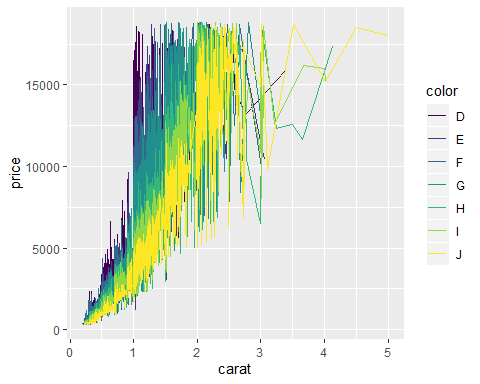
2.) Are certain colors associated with carat size? Provide an example. I would say it is likely color is associated with carat size. The reasoning behind this is that once carat size hits 2, the dark points, or the better colored diamonds, don’t show up on the plot as much. So, as carat size increases, the color of the diamonds worsen. See how the two carats above 4.5 are only color “J.”

dsample <- diamonds[sample(nrow(diamonds), 100), ]   
ggplot(dsample, aes(carat, price, color= color))+  
 geom\_point()+  
 geom\_smooth(method= "lm")



The above plot shows a fitted line for each color in the sample, from the code above. The sample includes on carat sizes between 0 and 2.5. From this, we say that as carat size increases, so does price. However, it becomes harder to tell if color is a factor. Based on the plot, color “J,” the worst colored diamonds, are more prevalent as carat size increases and that color “D,” the best colored diamonds, are not as common as carat size increases. This is because of the fitted line stopping before the end of the plot.

ggplot(diamonds, aes(carat, price, color= color))+  
 geom\_line()



Above we have a line chart of the diamonds dataset. Each color has its own line connecting the points based on carat size and price. I chose this chart as it still helps to show the volume of of each color and how each color affects price at each interval. It is clear to tell that price increases as carat size increases and that at the same carat size, better colored diamonds are worth more.