# Module 3 - Assignment 2

## Black, Tyler

### Exploratory Data Analysis

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

#### Diamond Color and Price

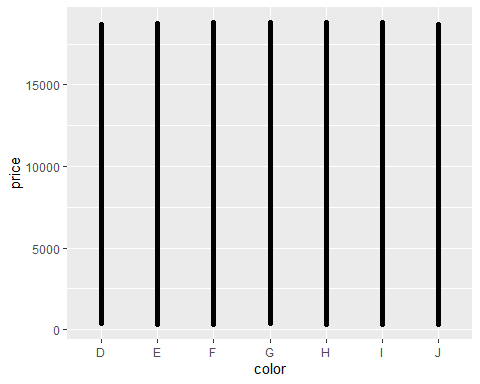
ggplot(data=diamonds, aes(x=carat,y=price)) +  
 geom\_point()



1.) What do you notice from the scatterplot as the carat size increases?  
**When the carat size increases, the price also increases. However, after a certain carat size, the price isn’t as dependent on the carat size**

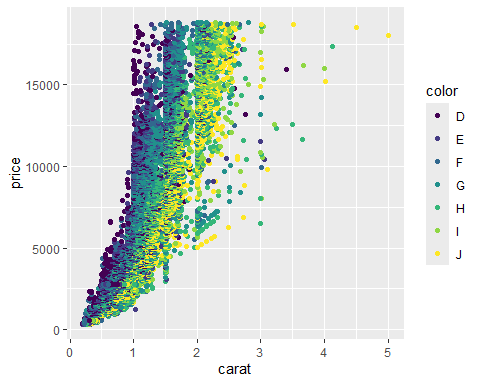
2.) From the scatterplot, what carats are most represented within the diamonds dataset?  
**Most of the diamonds are between the .5 to 2.5 carat size per the scatter plot**

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



Do you see any difference based solely on color? **I do not**

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

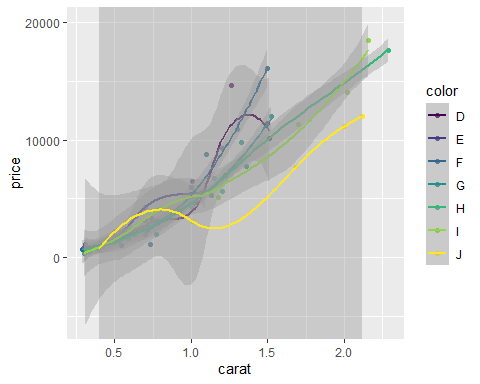


1.) Does color impact the price?  
**Yes, color impacts price. A diamond can have a low carat size, but have the best color (‘D’) and be just as expensive as the diamonds with the biggest carat size. However, if a diamond has a big carat size, it can offset the diamond having the worst color (‘J’) and be as expensive as diamonds with better colors**

2.) Are certain colors associated with carat size? Provide an example.  
**The best color diamonds (‘D’) seem to be associated with a low carat size. For example, diamonds with color (‘D’) and a carat size of 1 have many observations in the $15,000 price range. In contrast, the worst color diamonds (‘J’) seem to be associated with a bigger carat size. There are a few observations of diamonds with the worst color (‘J’) and a carat size of over 4 in the price range of over $15,000.**

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'

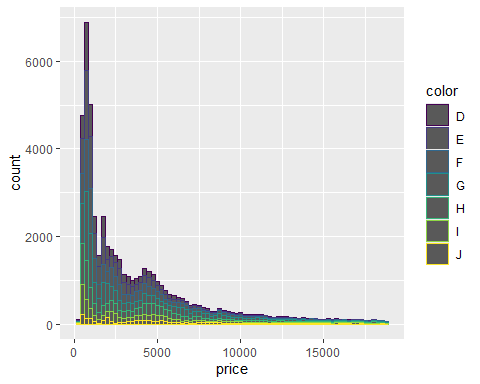


**PLEASE UPDATE IN .DOCX - DR.CUMMINGS IGNORE THIS LINE**

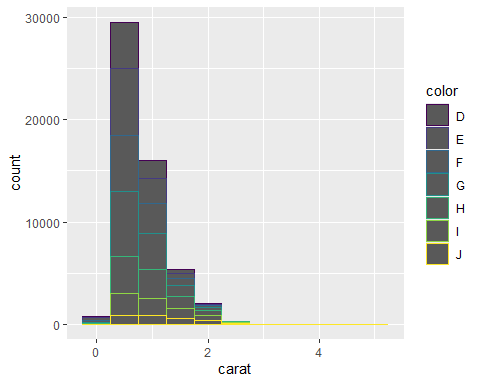
**I think this chunk of code is harder to analyze than when using the entire data set. Additionally, certain colors or a non-representative number of observations of each color class may be chosen since the sample is random and only 100 observations.**

#### Additional Plots/Graphs

ggplot(data=diamonds,aes(x=price, color = color)) +  
 geom\_histogram(binwidth = 250)



ggplot(data=diamonds,aes(x=carat, color = color)) +  
 geom\_histogram(binwidth = .5)



**I think these histograms do a great job of showing the distribution of diamond colors per carat size and price. These histograms provide great evidence for the conclusions purported earlier**