# Module 3 - Assignment 2

## Angelo, Elizabeth

### Exploratory Data Analysis

library(tidyverse)

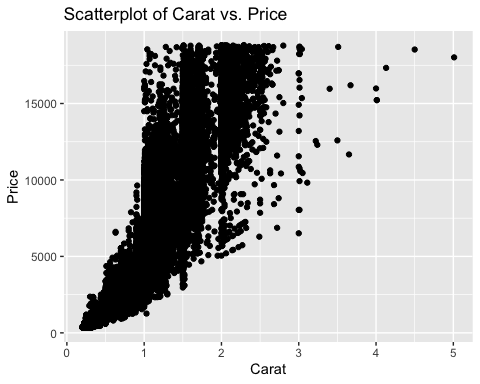
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.5 ✔ purrr 0.3.4  
## ✔ tibble 3.1.6 ✔ dplyr 1.0.7  
## ✔ tidyr 1.1.4 ✔ stringr 1.4.0  
## ✔ readr 2.1.1 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

#### Diamond Color and Price

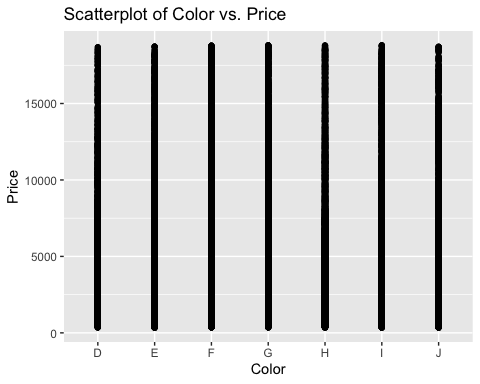
library(ggplot2)  
ggplot(diamonds, aes(x = carat, y = price)) +   
 geom\_point() +   
 labs(title = "Scatterplot of Carat vs. Price",   
 x = "Carat",   
 y = "Price")



1.) What do you notice from the scatter-plot as the carat size increases?  
As carat size increases, the data points become sparsely populated and more spread out. After 4 carats, the price is consistently above $15,000.

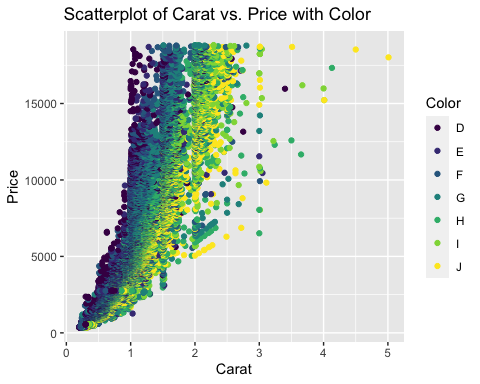
2.) From the scatter-plot, what carats are most represented within the diamonds dataset?  
Carats between 0.5-2.5 appear to be represented the most in the dataset.

ggplot(diamonds, aes(x = color, y = price)) +   
 geom\_point() +   
 labs(title = "Scatterplot of Color vs. Price",   
 x = "Color",   
 y = "Price")



The scatter-plot appears to show the same results for all colors and price. All colors have the same vertical line of plot points starting from Price $0 to just under $2,000.

ggplot(diamonds, aes(x = carat, y = price, color = color)) +   
 geom\_point() +   
 labs(title = "Scatterplot of Carat vs. Price with Color",   
 x = "Carat",   
 y = "Price",   
 color = "Color")

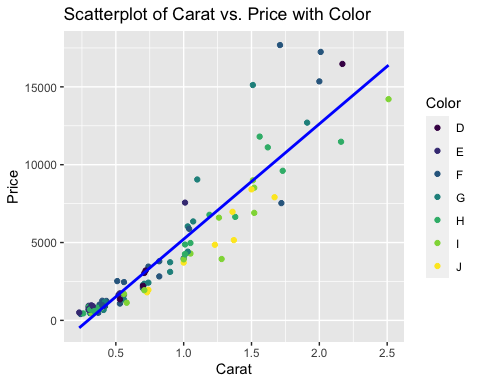


1.) Does color impact Price?  
Based off the scatter-plot, color does not appear to have an impact on price at a significant level. The scatter-plot shows the colors all range in price depending on the carat size, although there are two points that show both a yellow and green 4 carat diamond with green being just slightly more expensive. Yet, most of the points appear to be uniform with carat size impacting price more than color.

2.) Are certain colors associated with carat size?  
It does appear that as carat size increases, the most prominent colors become yellow(J), light green (I), and green(H), with a very sparse population of blue(F) and purple (D, E).

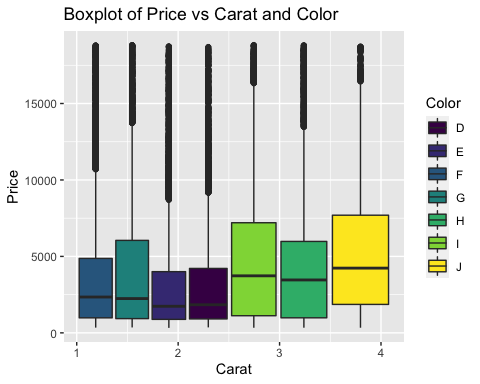
dsample <- diamonds[sample(nrow(diamonds), 100), ]  
ggplot(dsample, aes(x = carat, y = price, color = color)) +   
 geom\_point() +   
 geom\_smooth(method = "lm", se = FALSE, color = "blue") +   
 labs(title = "Scatterplot of Carat vs. Price with Color",   
 x = "Carat",   
 y = "Price", color = "Color")

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



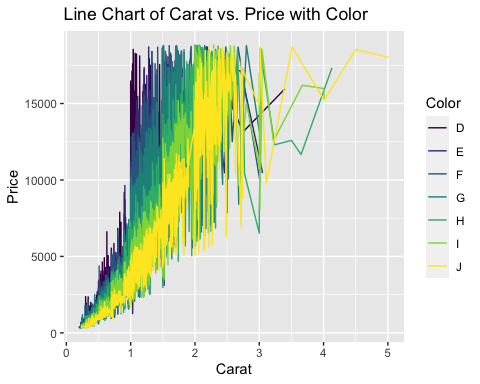
As noticed above, the results regarding color, carat, and price still show that as carat size increases, the color purple decreases after 1 carat. The more prominent colors at higher carat size and price are yellow, teal, and green. Unlike the results earlier, this sample shows more blue with larger carat size and higher prices.

ggplot(diamonds, aes(x = carat, y = price, fill = color)) +   
 geom\_boxplot() +   
 labs(title = "Boxplot of Price vs Carat and Color",   
 x = "Carat",   
 y = "Price",   
 fill = "Color")



The box-plot shows that light green, green, and yellow or strongly correlated to larger carat sizes from 2.5-4 carats. This matches the scatter-plot with color analysis above. These three colors are also associated with higher prices ranging from $13,000 - $15,000+, and account for a higher percentage of the sample population. The colors blue, teal, light purple, and purple range from 1-2.5 in carat size, with prices varying between $8,000 - $15,000+.

ggplot(diamonds, aes(x = carat, y = price, color = color)) +   
 geom\_line() +   
 labs(title = "Line Chart of Carat vs. Price with Color",   
 x = "Carat",   
 y = "Price",   
 color = "Color")



After reviewing the line graph, the results further confirm the analysis above. When carat size reaches 2.5, the prominent colors are yellow, green, and light green, and associated with larger carat size and higher prices. Purple is slightly associated with larger carat size and higher prices up until around 3.4 carat size. It seems all colors are prominent within the 0.2-2.5 carat size range. After a carat size of 2.5, some colors seem to show up less, or not at all.