# Module 3 - Assignment 2

## Pierce, Michaela

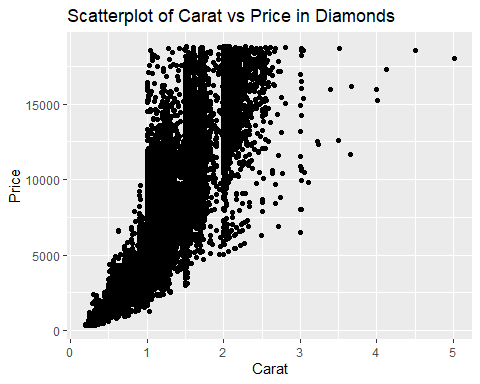
### Exploratory Data Analysis

# Load the tidyverse package  
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 3.5.0 ✔ 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**

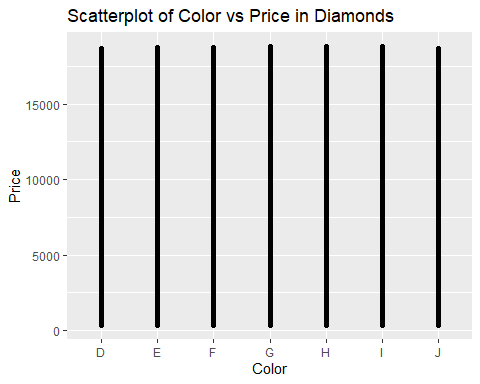
library(ggplot2)  
# Create scatterplot using diamonds dataset  
ggplot(diamonds, aes(x = carat, y = price)) +  
 geom\_point() +  
 labs(title = "Scatterplot of Carat vs Price in Diamonds",  
 x = "Carat",  
 y = "Price")



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

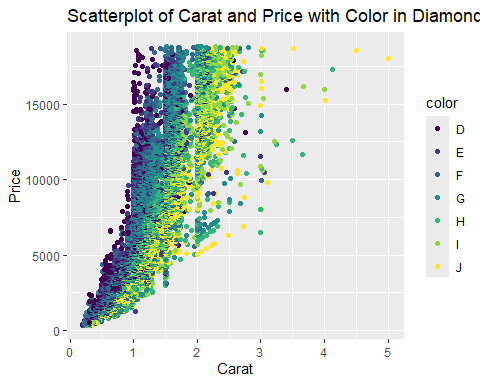
**2.) From the scatterplot, what carats are most represented within the diamonds dataset?**  
1 to 2 carat size are most represented within the diamonds dataset.

# Create scatterplot using diamonds dataset  
ggplot(diamonds, aes(x = color, y = price)) +  
 geom\_point() +  
 labs(title = "Scatterplot of Color vs Price in Diamonds",  
 x = "Color",  
 y = "Price")



I can see it both ways but overall I would say not useful. There is no specific data it gives. The only way I could see it would be useful is by showing color does not influence price. Realistically I am sure it does influence price but based on this plot it is hard to determine. It may be that there are diamonds of each color and varied prices so that there are diamonds of the same price with different colors. This plot is not useful for analyzing the data. I do not see any difference based solely on color.

# Create scatterplot with color using diamonds dataset  
ggplot(diamonds, aes(x = carat, y = price, color = color)) +  
 geom\_point() +  
 labs(title = "Scatterplot of Carat and Price with Color in Diamonds",  
 x = "Carat",  
 y = "Price")



**1.) Does color impact the price?**  
It looks like it may have some impact but carat has more impact.

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

* Color D seems to be associated with 2 carats or less. There are a few that are associated but all are less than 4 carats.
* Color E seems to be associated with 2 carats or less. Looks like there is 1 that is 3 carats.
* Color F seems to be associated with 3 carats or less.
* Color G seems to be associated mainly with 3 carats or less.
* Color H seems to be associated mainly with 3 carats or less. There are a few that are between 3 and 4 carats.
* Color I seems to be associated with 4 carats or less.
* Color J seems to be mostly associated with 2 carats or less. However, there are a few that are 3 carats. A couple are also about 4 carats and there is one that is 5 carats. It looks like this color is less associated with 1 carat or less.

# Create a random sample of 100 rows from the diamonds dataset  
dsample <- diamonds[sample(nrow(diamonds), 100), ]

# Create scatterplot with color using dsample dataset  
ggplot(dsample, aes(x = carat, y = price, color = color)) +  
 geom\_point() +  
 geom\_smooth() +  
 labs(title = "Scatterplot of Carat and Price with Color in Diamonds (Sample)",  
 x = "Carat",  
 y = "Price")

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

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : at 0.3299

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : radius 1e-08

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : pseudoinverse used at 0.3299

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : neighborhood radius 0.0001

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : at 0.3501

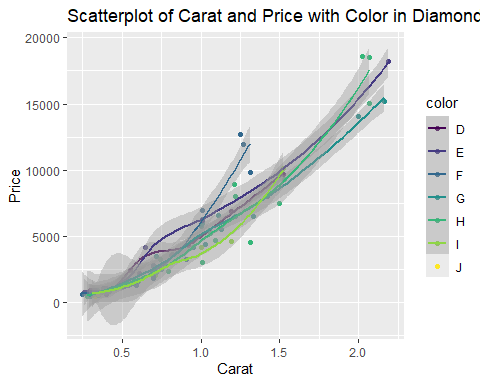
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : radius 1e-08

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : There are other near singularities as well. 1e-08

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : zero-width neighborhood. make span bigger  
  
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,  
## : zero-width neighborhood. make span bigger

## Warning: Failed to fit group 7.  
## Caused by error in `predLoess()`:  
## ! NA/NaN/Inf in foreign function call (arg 5)



The scatterplot above represents a random sample of 100 diamonds. It is interesting to note that color D has the largest price and largest carat size. It also looks like there are only certain colors that have the a carat size of more than 1.5, such as D, G, H, and J.

I wanted specifics on the sample dataset so I used extra code to better understand it.

# Convert color to character to ensure proper printing  
dsample$color <- as.character(dsample$color)  
  
# Find the row with the largest price  
max\_price <- dsample[which.max(dsample$price), ]  
  
# Find the row with the smallest price  
min\_price <- dsample[which.min(dsample$price), ]  
  
# Find the row with the largest carat size  
max\_carat <- dsample[which.max(dsample$carat), ]  
  
# Find the row with the smallest carat size  
min\_carat <- dsample[which.min(dsample$carat), ]  
  
# Print the results  
cat("Color with the largest price:", max\_price$color, "\n")

## Color with the largest price: H

cat("Price:", max\_price$price, "\n")

## Price: 18565

cat("Carat size:", max\_price$carat, "\n\n")

## Carat size: 2.02

cat("Color with the smallest price:", min\_price$color, "\n")

## Color with the smallest price: G

cat("Price:", min\_price$price, "\n")

## Price: 432

cat("Carat size:", min\_price$carat, "\n\n")

## Carat size: 0.3

cat("Color with the largest carat size:", max\_carat$color, "\n")

## Color with the largest carat size: E

cat("Price:", max\_carat$price, "\n")

## Price: 18232

cat("Carat size:", max\_carat$carat, "\n\n")

## Carat size: 2.19

cat("Color with the smallest carat size:", min\_carat$color, "\n")

## Color with the smallest carat size: F

cat("Price:", min\_carat$price, "\n")

## Price: 633

cat("Carat size:", min\_carat$carat, "\n")

## Carat size: 0.25

The code above provides specifics as to what the sample dataset contains. This can also be depicted visually in the scatterplot and line chart in this document.

After studying the sample data set, using the code and the scatterplot, I found some interesting trends/patterns in the data.

I observed the highest-priced diamond, valued at $18,407, representing a grade D stone with a substantial carat size of 2.11. I did extra research to better understand the color grades of diamonds so that the price points and carat sizes were easier to understand. I found that grade D diamonds are fully colorless, are rare and valuable (highest grade/highest quality), which aligns well with the findings from this random data sample.

Conversely, I also noticed the lowest-priced diamond, priced at $449, which is a grade H stone with a modest carat size of 0.24. Additional investigation revealed that grade H diamonds are near-colorless and offer excellent value for their quality. The smaller carat size of 0.24 contributes to its lower price point, which is consistent with expected pricing trends.

Another intriguing observation I made is the presence of another grade D diamond with the largest carat size, being the same diamond with the highest-price. This shows the premium value associated with grade D diamonds, demonstrating that larger carat sizes correspond to higher prices.

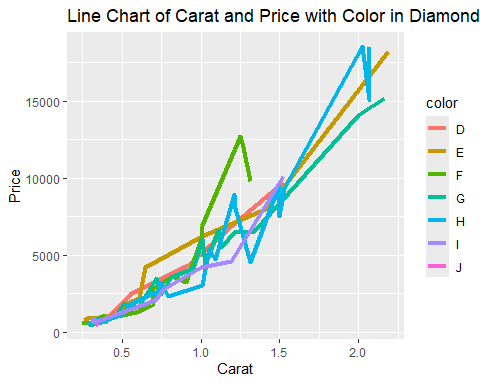
Conversely, I also identified a grade E diamond with a slighter carat size of 0.23, priced at $472, which is just marginally smaller than the lowest-priced grade H diamond. Despite its diminutive size, being just below grade D, this colorless diamond is still of high quality and has that pure icy look. Although its price slightly exceeds that of the lowest-priced diamond, its classification as a grade E gem, indicative of superior quality, justifies its higher pricing. Therefore, while it may initially seem perplexing that the smallest carat size doesn’t correspond to the lowest price, considering the higher quality associated with grade E diamonds clarifies this pricing discrepancy.

Overall, this was very interesting to analyze and was nice learning about diamonds as well. I think doing the research to better understand the diamond color classes helped me understand the data better. I was able to find patterns/trends easier. In short the data shows what is commonly represented in the real world. Higher quality usually has higher prices and larger carat size has a higher price.

Here is the source I used to better understand the diamond color grades. <https://www.bluenile.com/education/diamonds/color>

# Create line chart with color using dsample dataset  
ggplot(dsample, aes(x = carat, y = price, color = color)) +  
 geom\_line(size = 1.5) + # Adjust size to make lines larger  
 labs(title = "Line Chart of Carat and Price with Color in Diamonds (Sample)",  
 x = "Carat",  
 y = "Price")

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



I explained the sample data before doing this chart but still wanted to provide a brief explanation of this line chart. In the sample size it shows that color/grade D has the highest price in the sample. It also shows that as the carat size increases so does the price. It is interesting to look as the other grades below grade D. The pattern of increase in carat size increase in price remains the same. There are some questionable spikes but it could be due to quality as well.