# Module 3 - Assignment 2

## Kiser, Stephen

### Exploratory Data Analysis

library(tidyverse)

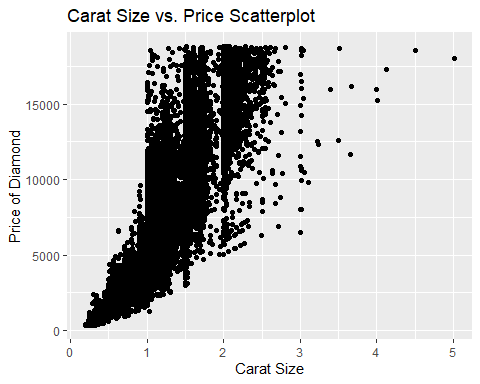
## -- Attaching packages -------------------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

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

#### Diamond Color and Price

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



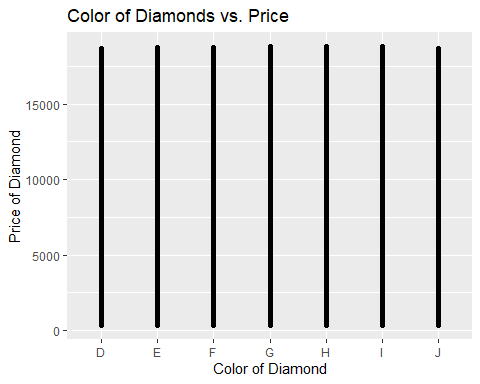
1.) What do you notice from the scatterplot as the carat size increases?

As the carat size increases the price increases. Also there are less data collected for higher carat sized diamonds.

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

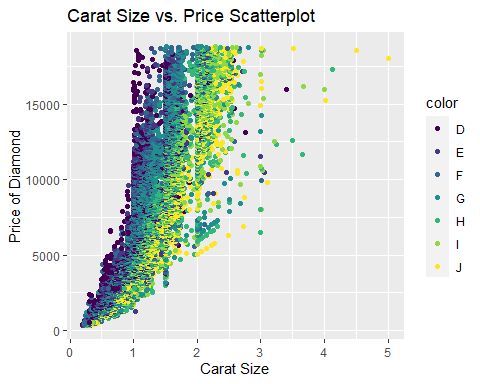
I would say the the carat size between 0 and 2.5 are the most represented for.

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



This scatterplot has made vertical lines for each color going up to the max price. This plot shows that color has no effect on the price of the diamond because each color reached the same max price. I do not see any difference solely on color.

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



After adding in the color function of this scatterplot we are able to see what impact color has on diamonds. Price is not effected by the color because every color has both low priced diamonds and high priced diamonds. However, color is associated with the carat size of the diamond. We see more yellow colored dots (color J) the higher the carat size is, and see more purple colored dots (color D) in the lower carat size

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'

## 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, : pseudoinverse used at 0.3417

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : neighborhood radius 0.6983

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : There are other near singularities as well. 0.95707

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : span too small. fewer  
## data values than degrees of freedom.

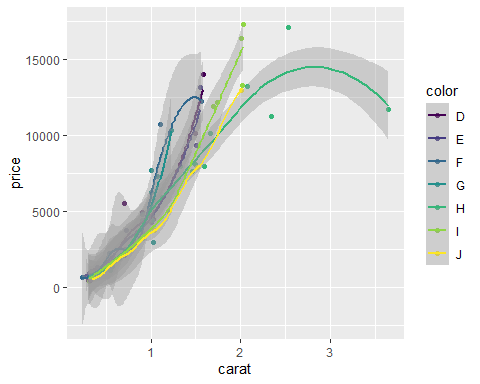
## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at  
## 0.3417

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius  
## 0.6983

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition  
## number 0

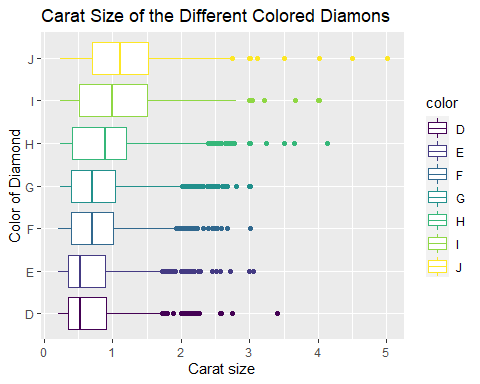
## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : There are other near  
## singularities as well. 0.95707

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -  
## Inf



What I can see from this random sample is that as carat size increases for all diamonds then price also increases. Color I has a slower rate of increase compared to the other diamonds, and then starts to plateau around 2.0 carats. Color G increases in price in the smaller carat size, however it starts to plateau around 2.0 carats. Color J price is much cheaper than the other colors in the smaller carat sizes.

ggplot(data = diamonds, aes(x= carat, y= color, color=color)) +  
 geom\_boxplot()+  
 labs(title = "Carat Size of the Different Colored Diamons", x= "Carat size", y= "Color of Diamond")



I created a boxplot to see how color and carat size are related. In this chart we can see what the average carat size of the different colored diamonds. As we can see color J has an average of over 1 carat, and have outliers all the way to 5 carats. Color D has the smallest average carat size, and only has an outlier of just less than 3.5 carats. This graph further shows that diamonds that are colored J are going to be bigger carat size while diamonds colored D are going to be a smaller carat size.