### R and RStudio Refresher

## Fernandez, Miguel

## BAN-502

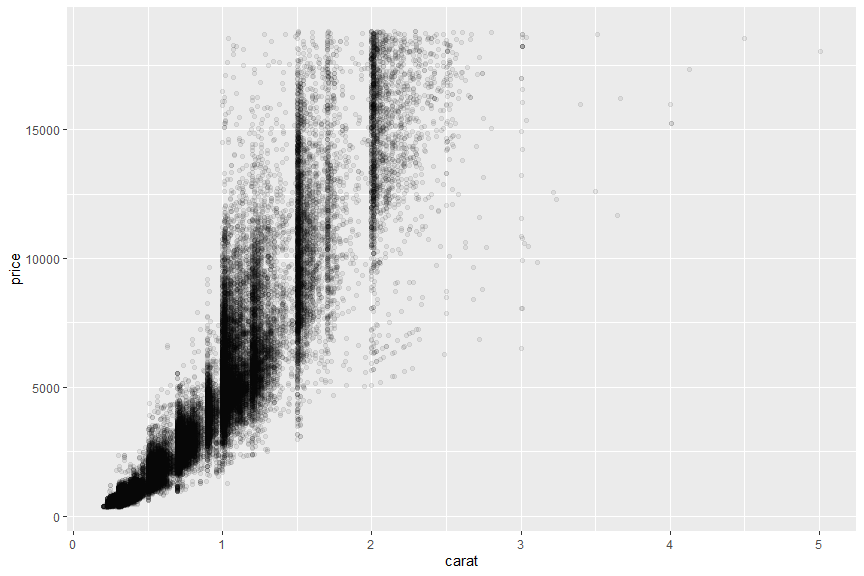
#### Task 1 & 2

# Import libraries  
library(tidyverse)  
library(knitr)  
  
# Create dataframe  
diamonddata = diamonds  
# str(diamonddata)

There are 53,940 rows and 10 columns in the diamonddata data set.

#### Task 3

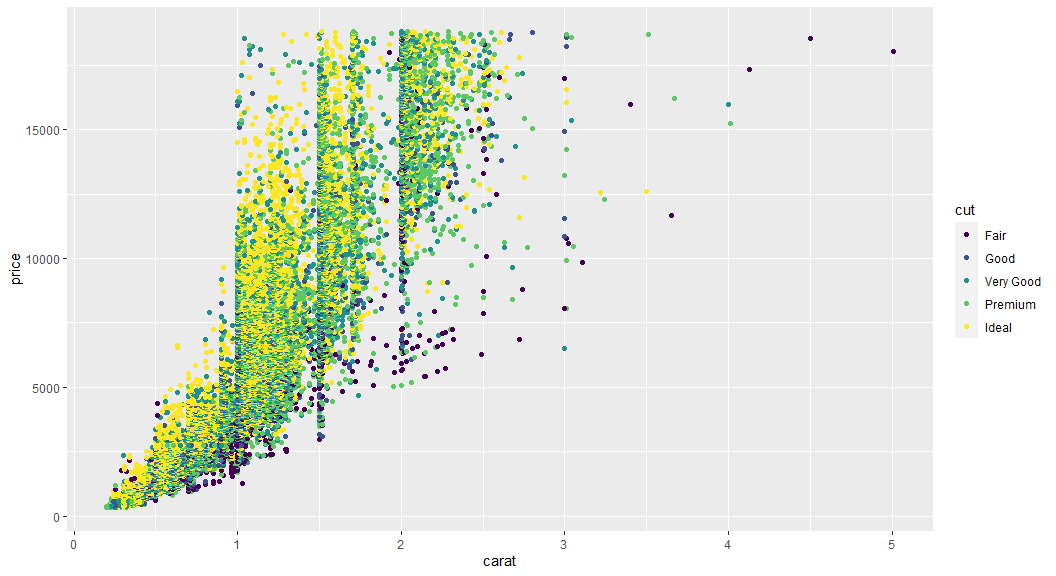
# Create scatterplot  
ggplot(diamonddata, aes(x=carat, y=price)) +  
 geom\_point(alpha=1/15)



The plot above displays the relationship between prices of diamonds and the weight, or carat, of a diamond. The plot reveals a strong positive relationship between carat and price meaning that as the carat increases, so to does the price. Looking at the plot, we can conclude that diamonds between approximately 0.2 carats and 1.25 carats are most represented in the data set. Using the alpha parameter in the plot, we can see that area appears to be densely clustered.

#### Task 4

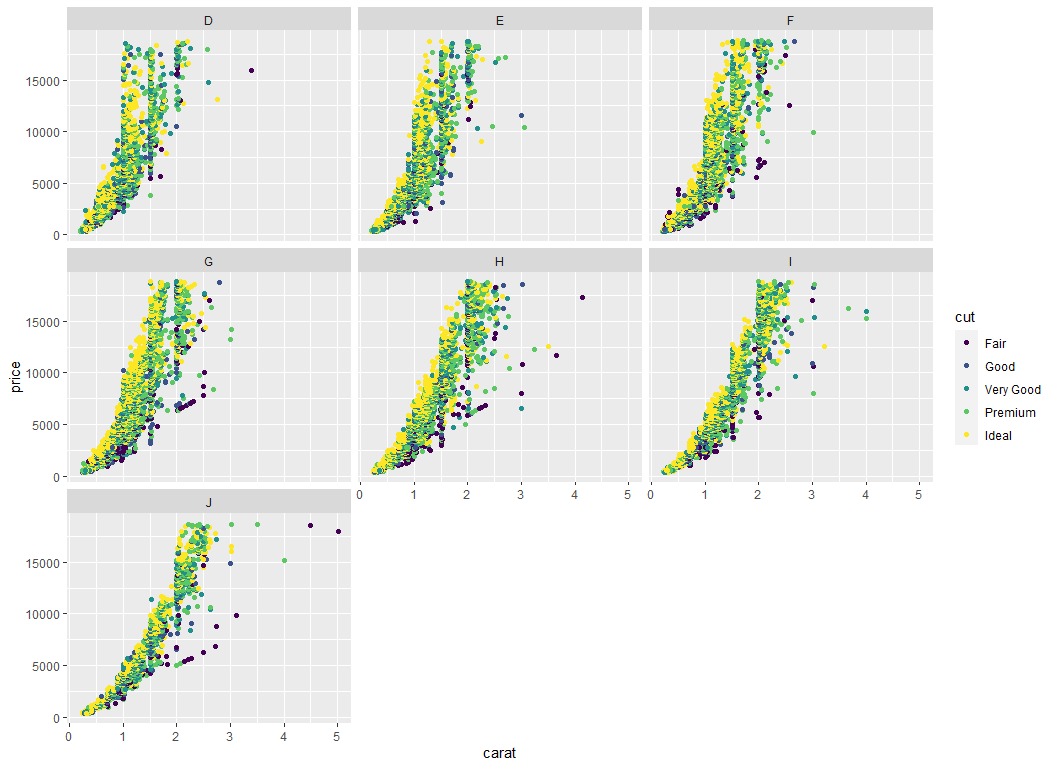
# Create second scatterplot  
ggplot(diamonddata, aes(x = carat, y = price, color = cut)) +  
 geom\_point()



The plot above examines the relationship between carat, price and cut. We can see that ideal and premium cuts often are the most expensive. However, there are a few outliers present in the data. If a diamond is larger than 3.5 carats, the quality of the cut decreases but the price is very expensive. We can determine that size has more of an impact on price than cut on these diamonds. Diamonds that are between 0.5 carats and 3 carats with a cut of ideal or premium tend to be more expensive than those diamonds with fair, good or very good cuts.

#### Task 5

# Create third scatterplot  
ggplot(diamonddata, aes(x = carat, y = price, color = cut)) +  
 geom\_point() +  
 facet\_wrap(~ color)



Looking at the facet grid above, we can see similar relationships among price, size, cut by the color of diamonds in the data set. What is interesting is that diamonds with ideal cuts seem to appear on the left side of the cluster while diamonds with premium and very good cuts are in the middle of the cluster and finally, diamonds with good and fair cuts appear on the right side of the cluster. From the plot, we can assume that ideal cut diamonds tend to be smaller and more expensive while fair cut diamonds are larger and less expensive. Colors D, E, F and G are more tightly clustered while colors H, I and J have larger outliers. The distribution in the plots also appears to be non-linear.

#### Task 6 & 7

# Read in data  
inventory <- read\_csv("InventoryData.csv")  
  
# Check structure and generate summary stats  
# str(inventory)  
# summary(inventory)  
  
# Filter data  
inventoryA <- inventory %>%  
 filter(Supplier == 'A')  
  
# Check structure of inventoryA  
# str(inventoryA)

There are 3,695 rows in the data frame inventoryA.

#### Task 8

inventoryA = mutate(inventoryA, OnHandRatio = `On Hand` / `Annual Demand`)

The code above takes the newly created data frame and creates a new variable called OnHandRatio. This is calculated by taking existing variables, On Hand and Annual Demand, and creating a ratio by dividing them.

#### Task 9

# Avg cost data frame  
avg\_cost <- inventory %>%  
 group\_by(`Item SKU`) %>%  
 summarize(SKUAvgCost = round(mean(`Cost per Unit ($)`),2))  
   
# Preview avg\_cost  
kable(head(avg\_cost))

|  |  |
| --- | --- |
| Item SKU | SKUAvgCost |
| 0100 | 73.90 |
| 011 | 10.18 |
| 0113 | 164.24 |
| 0122 | 105.49 |
| 013 | 17.59 |
| 0133 | 142.42 |

#### Task 10

I just completed MIS-503 which introduced me to R. I am also working through an R for Data Science course on Udemy. I have been programming in Python for more than a year which introduced me to data analysis concepts from a programming perspective. Having said all that, I have not experienced any real difficulty with the topics or concepts. These are all things I have implemented in Python at work and home. I am really enjoying R, however there are things that Python is better at and conversely, there are things that R is better at. But it’s best to have as many tools in your toolbelt as possible.