## Week 1 Assignment 3

### Name - DEY, Sankha

#### Refresher Assignment on R

##### Task 1

#install.packages("tidyverse")  
# Used 'message =FALSE' to silence the ouput at final knit  
library(tidyverse)

##### Task 2

diamonddata = diamonds  
nrow(diamonddata)

## [1] 53940

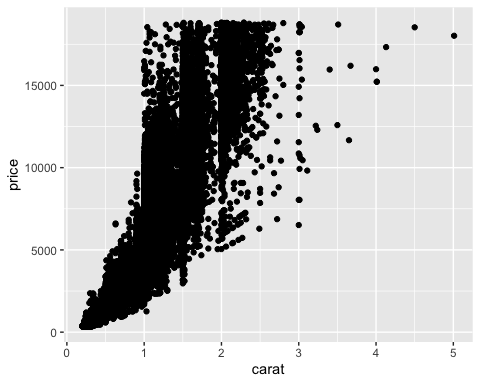
ncol(diamonddata)

## [1] 10

We have data of 53940 rows and 10 columns.

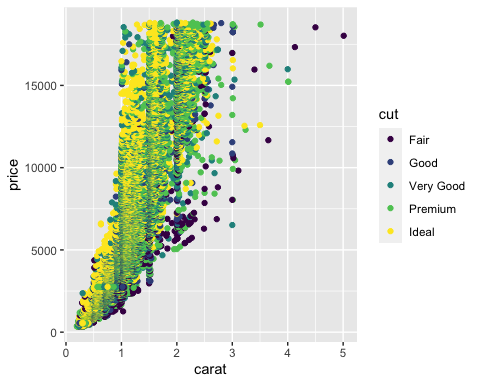
##### Task 3

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

 Diamond price increases when carat size gets bigger (from 0 to 2.5). Diamonds with carat size closer to 0.2 have the least price. Carat size 3 is scarce and their prices are not higher than the same of carat size 2.Overall there is a positive relationship between these two variables.

##### Task 4

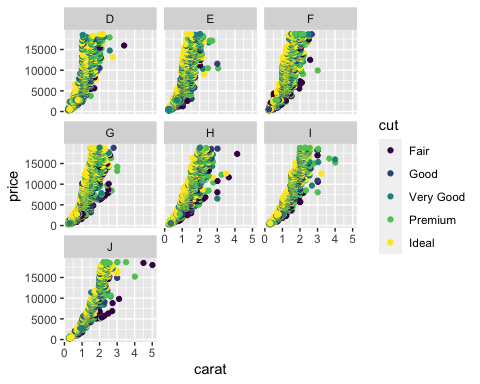
ggplot(diamonddata,aes(x=carat,y=price,color= cut))+  
 geom\_point()



**Carat VS Price:** Diamond price increases when carat size gets bigger (from 0 to 2.5). Diamonds with carat size closer to 0.2 have the least price. Carat size 3 is scarce and their prices are not higher than the same of carat size 2. There is a positive relationship between these two variables.  
**Cut VS Price:** From this diagram, it is inconclusive if cut determines price. Because we can see each cut has scatterred equally across the price range (except few outliers for cut fair).  
**Cut VS Carat:** Cut Fair is associated with bigger carat size. Above 4 carats size diamond are all cut into Fair.There is a negative relationship.

##### Task 5

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

 Considering color D is the best and J is the worst, there is a kind of negative relationship between color and price.There is also a slight negative relationship between carat and cut. The largest diamond is cut Fair.  
Color D and E diamonds are generally small sizes (carat 2 and less). Carat size 3 and biggers dimaonds are mostly comprised of color H,I and J. Color J and Fair Cut diamonds are comparitively costly.

##### Task 6

inventory <- read\_csv("InventoryData.csv")

## Parsed with column specification:  
## cols(  
## `Item SKU` = col\_character(),  
## Store = col\_character(),  
## Supplier = col\_character(),  
## `Cost per Unit ($)` = col\_double(),  
## `On Hand` = col\_double(),  
## `Annual Demand` = col\_double()  
## )

str(inventory)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 13561 obs. of 6 variables:  
## $ Item SKU : chr "0100" "0100" "0100" "0100" ...  
## $ Store : chr "003480" "01611" "01611" "020109" ...  
## $ Supplier : chr "A" "B" "D" "B" ...  
## $ Cost per Unit ($): num 125.32 115.12 53.61 2.26 60.51 ...  
## $ On Hand : num 159 40 174 176 74 48 6 129 82 17 ...  
## $ Annual Demand : num 1693 351 1691 1559 733 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. `Item SKU` = col\_character(),  
## .. Store = col\_character(),  
## .. Supplier = col\_character(),  
## .. `Cost per Unit ($)` = col\_double(),  
## .. `On Hand` = col\_double(),  
## .. `Annual Demand` = col\_double()  
## .. )

summary(inventory)

## Item SKU Store Supplier Cost per Unit ($)  
## Length:13561 Length:13561 Length:13561 Min. : 0.0   
## Class :character Class :character Class :character 1st Qu.: 137.0   
## Mode :character Mode :character Mode :character Median : 377.5   
## Mean : 504.4   
## 3rd Qu.: 775.5   
## Max. :1982.3   
## On Hand Annual Demand   
## Min. : 0.0 Min. : 0.0   
## 1st Qu.: 50.0 1st Qu.: 483.0   
## Median :101.0 Median : 965.0   
## Mean :100.5 Mean : 966.2   
## 3rd Qu.:151.0 3rd Qu.:1448.0   
## Max. :200.0 Max. :2150.0

##### Task 7

inventoryA <- inventory %>%  
 filter(Supplier == "A")  
nrow(inventoryA)

## [1] 3695

In inventoryA dataframe, there are 3695 rows.

##### Task 8

It will create a new variable at the right side of the dataframe inventoryA. New variable name will be OnHandRatio. This variable will be calculated by the formula - (On Hand)/ (Annual Demand).

##### Task 9

avg\_cost <- inventoryA %>%  
 group\_by(`Item SKU`) %>%  
 summarize(SKUAvgCost= mean(`Cost per Unit ($)`))  
glimpse(avg\_cost)

## Observations: 1,720  
## Variables: 2  
## $ `Item SKU` <chr> "0100", "011", "0113", "0122", "013", "0133", "0137", "014…  
## $ SKUAvgCost <dbl> 125.32000, 12.33000, 197.36500, 81.70667, 14.31667, 203.47…

##### Task 10

I felt the most challenging part was the formatting of visuals graphs at ggplot. There are so many themes and parameters that I felt very chellenging to remember. I came across a few of such functions like- labs(), theme\_dark(), scale\_y\_continuous(labels = scales::comma) , element\_text(angle = , vjust=, size=). This would need more practice.