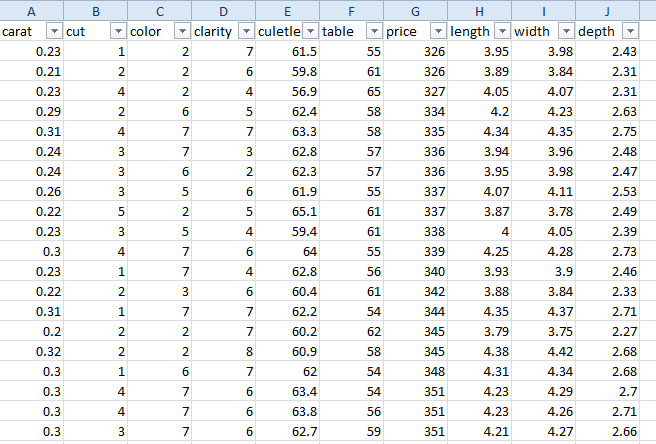
I used a dataset of 53940 Diamonds showing their size details, carats and price. I downloaded the csv file from kaggle.com. The data fields includes: carat, cut, color, clarity, culetlength, table, price, length, width and depth as shown below.



Here, the cut column in this dataset is the clustering field which stands for its quality: 1- Ideal, 2 – Premium, 3 – Very Good, 4 – Good and 5 - Fair. Now, I am going to feed this dataset into k means package with cluster value as 5 (since the varieties of diamond are 5 from Ideal to Fair) and check whether the cut value of the diamonds is clustered based on the other fields are not.

Color from 1 – 7 with 1 being best.

Clarity from 1 – 8 with 1 being best.

Let me load this file into R studio,

#Load the data

diamond <- read.csv("C:/python/diamonds.csv")

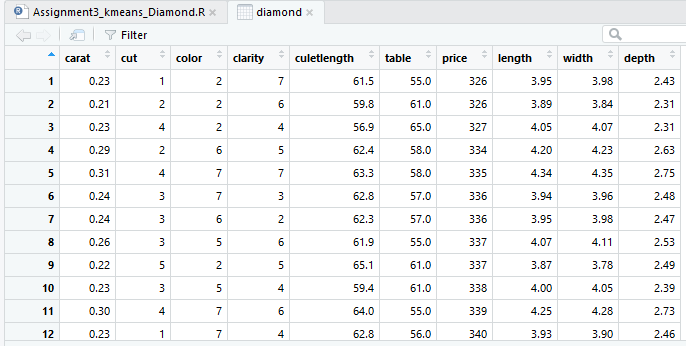
names(diamond)

nrow(diamond)

head(diamond)

[1] "carat" "cut" "color" "clarity" "culetlength" "table"

[7] "price" "length" "width" "depth"



#Remove the variable "cut"

data<-diamond

data$cut<-NULL

head(data)

#Apply the k-means

clus\_output<-kmeans(data,5)

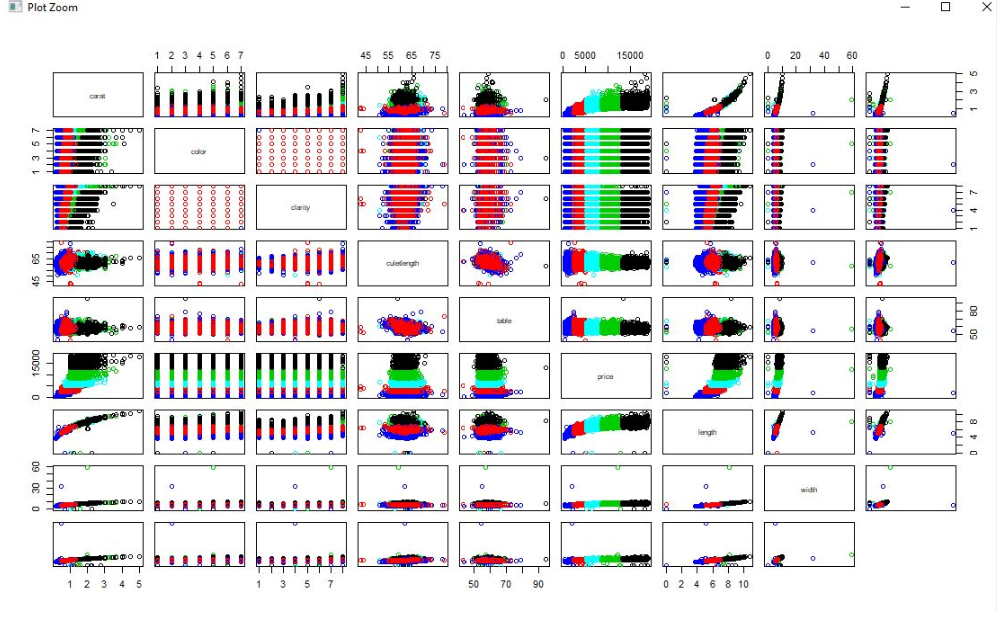
clus\_output

clus\_output$size

clus\_output$cluster

#Plot the clusters in different colors according to our model

plot(data, col=clus\_output$cluster)

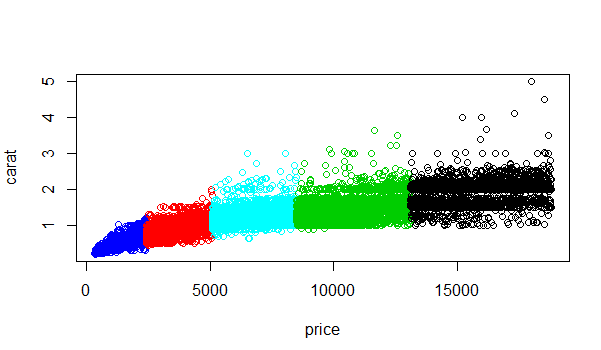


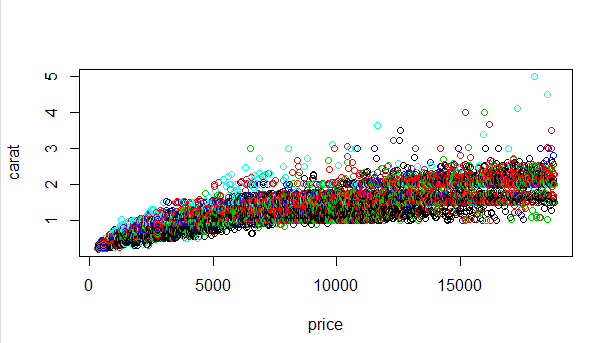
After plotting the dataset, it looks like price Vs other fields shows a pretty good clustering.

So, let me plot price Vs carat w.r.t dataset and cut field and compare the plots.

plot(diamond[c("price", "carat")], col=clus\_output$cluster)

plot(diamond[c("price", "carat")], col=diamond$cut)





table(diamond$cut, clus\_output$cluster)

1 2 3 4 5

1 851 4003 1595 12669 2433

2 984 3182 1343 5985 2297

3 569 3162 953 5676 1722

4 211 1551 300 2151 693

5 74 643 111 550 232

**My inferences on this K means clustering are as follows:**

1. I have used K means package for this dataset because the diamonds can be grouped from Ideal to Fair based on its dimensions, color, clarity and price.
2. The K means clustering algorithm has done a clustering which is NOT fitting with the cut value of the diamonds based on the above two graphs.
3. The cut value of the diamonds is not predominantly depends on this dataset alone. However, the algorithm has predicted a pretty good clustering based on price and carat of the diamonds. Basically, the combination of price and any other field in the dataset gives a fair clustering for this dataset.