Home Work 2

R and Statistical analysis

Dr.Shadman

Mina Kanaani
FERDOWSI UNIVERSITY OF MASHHAD

List of Contents

HM2

Mina Kanaani

2022-04-15

Abstract

HM2

This is the Report on the second R and Statistical Analysis's homework.

In this report you will find Questions related to EDA and Statistical part done on "diamonds" data set form "ggplot2" library. almost each question contains one or more plots with a short analysis on the result.

I hope you find this report useful.

```
library(knitr)
library(ggplot2)
kable(head(diamonds,10))
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
0.23	Very Good	H	VS1	59.4	61	338	4.00	4.05	2.39

Part1: EDA

۱. متغیر قیمت (price) چه نوع متغیری است؟ کمی یا کیفی؟ با استفاده از معیارهای عددی، تمایل مرکزی و پراکندگی آن را محاسبه و تحلیل کنید.

1.

The **"price"** variable is ratio and quantitative.

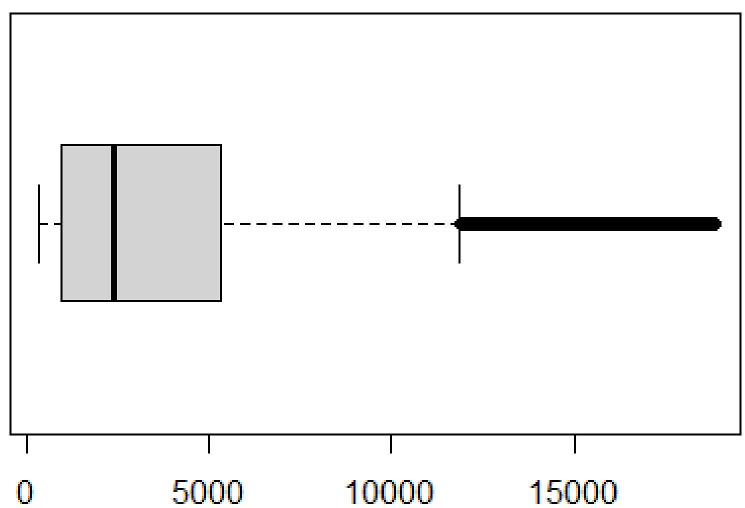
```
attach(diamonds)
getmode=function(v){
  dfv=as.data.frame(table(v))
  #print(dfv[,2])
  #print(which.max(dfv[,2]))
  dfv[which.max(dfv[,2]),1:2]}
getmode(price)

##       v Freq
## 261 605  132
```

The code above,we use Modes to find out which Price is the most frequent for measuring the central tendency . as the result is shown, the price "605" is the most frequent price with 132 times of repeating.

```
fivenum(price)

## [1]   326.0   950.0  2401.0  5324.5 18823.0
```

In this part, to understand the dispersion better, I use “fivenum” function to create the minimum,first quartile,median , third quartile and maximum. as a result, we can see that price 605 with most frequency , is closest between the minimum and first quartile, so the central tendency can be confirmed.

```
max(price)-min(price)

## [1] 18497

sd(price)

## [1] 3989.44

boxplot(diamonds$price,horizontal=T)
```



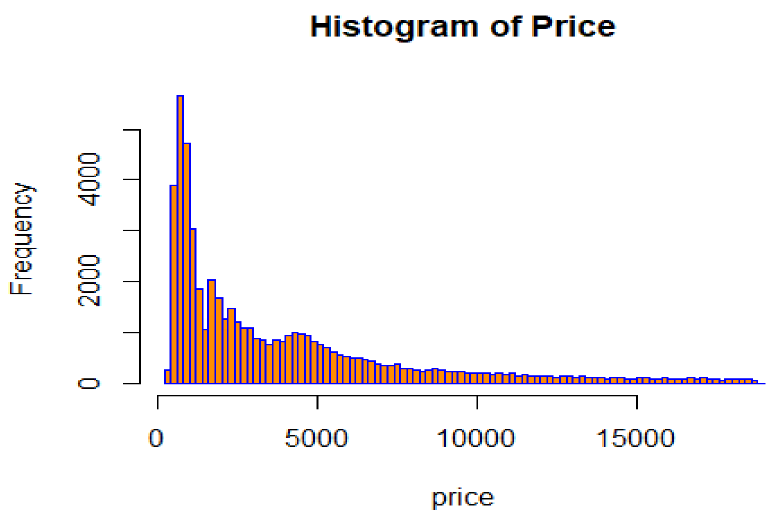
As it is shown in the boxplot , the central tendency is closer to minimum and first quartile than maximum. ad the result of range and standard deviation shows, the average distance from the mean is almost 3990, which is a great number and it shows data points are spread out over a large range of values.

۲. یک هیستوگرام برای متغیر قیمت الماس تهیه کنید و با توجه به آن شکل توزیع این متغیر را تحلیل کنید. تعداد بازه های هیستوگرام را چند بار تغییر دهید و سعی کنید یک مقدار مناسب برای تعداد بازه ها پیدا کنید.

2.

```
hist(diamonds$price,
     xlab='price',
     main='Histogram of Price',
     breaks=75,
     col='darkorange',
     border='blue',
     freq=TRUE)
```

The skewness from normal distribution is visible in the plot above, plus this plot is not symmetrical and certainly "price" factor does not have a normal distribution.



۳. متغیر رنگ (color) چه نوع متغیری است؟ آیا تمام رنگ های الماس (که در شکل بالا نشان داده شده است) در مجموعه داده وجود دارد؟ کدام رنگ بیشـتر در مجموعه داده وجود دارد؟ با اسـتفاده از یک جدول فراوانی، درصـد هر یک از رنگ ها را در مجموعه داده مشخص کنید و با استفاده از یک نمودار میله ای آن را تشریح و تحلیل کنید.

3.

The "**Color**" is qualitative and nominal variable.

```
coltable=as.data.frame(table(color))
coltable[which.max(coltable[,2]),1]
```

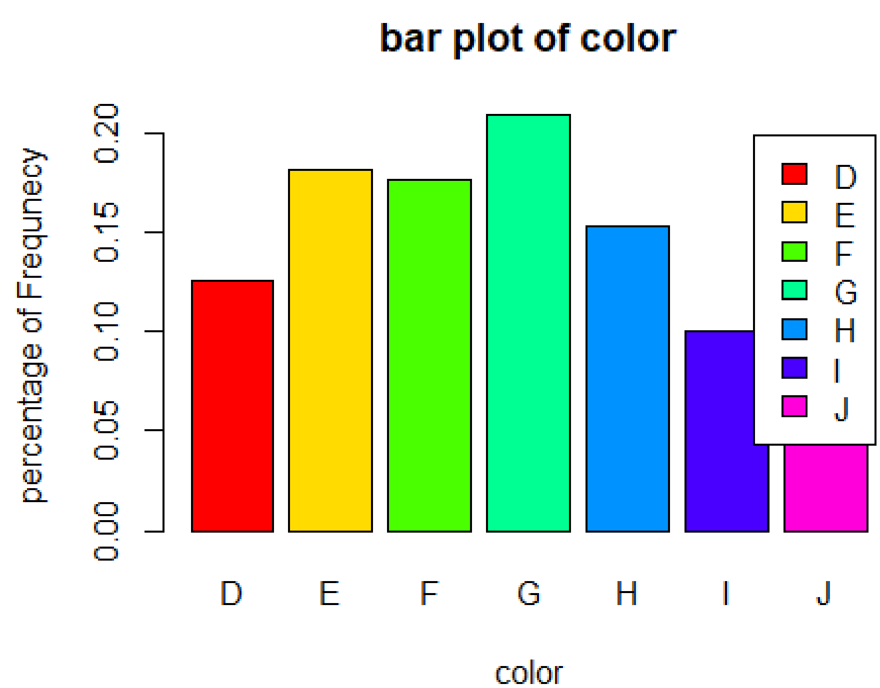
```
## [1] G
## Levels: D E F G H I J
```

Not all the Color types are used in this data set and the color "**G**" is the most frequent.

```
mysum=sum(coltable["Freq"])
mycolor=table(color)
newcolor=mycolor[]/mysum
newcolor*100

## color
##         D         E         F         G         H         I         J
## 12.560252 18.162773 17.690026 20.934372 15.394883 10.051910  5.205784
```

Using "table" function to make a frequency table and then changed the data to percentage.

```
barplot(newcolor,legend.text = T,col=rainbow(7),
        ylab="percentage of Frequncy",
        xlab="color",main="bar plot of color")
```



In this plot, each color is identified with specific color , showing in the legend. same as the result of frequency table in the previous chunk, the color "G" with the highest bar is the most frequent, the other colors frequency is comparable to each other in this plot as well.

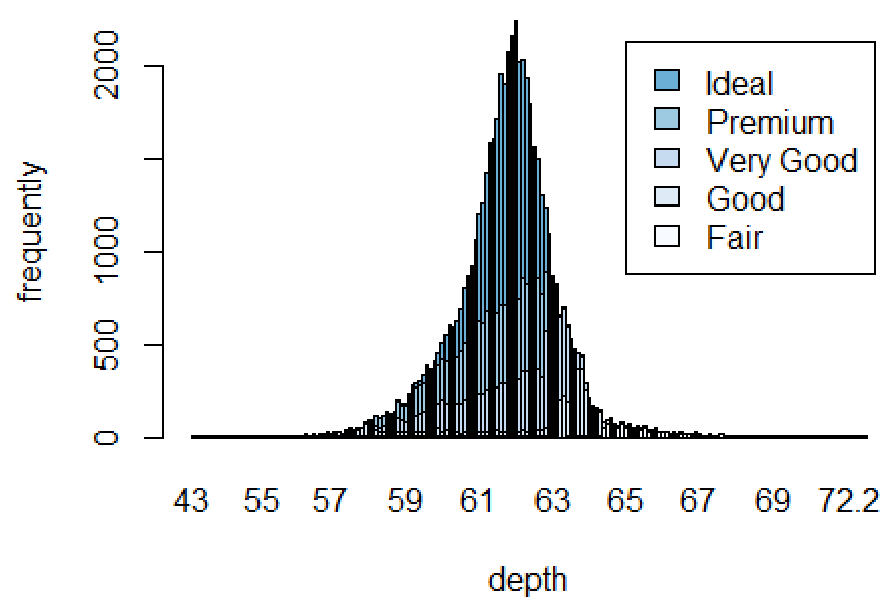
۴. الماس های کدام درجه از تراش، بیشترین عمق (depth) را دارند؟ به طور متوسط، با افزایش یا کاهش درجه تراش، عمق افزایش یا کاهش می یابد؟

4.

```
cut[which.max(depth)]

## [1] Fair
## Levels: Fair < Good < Very Good < Premium < Ideal
```

The "Fair" cut is the cut category with most depth.

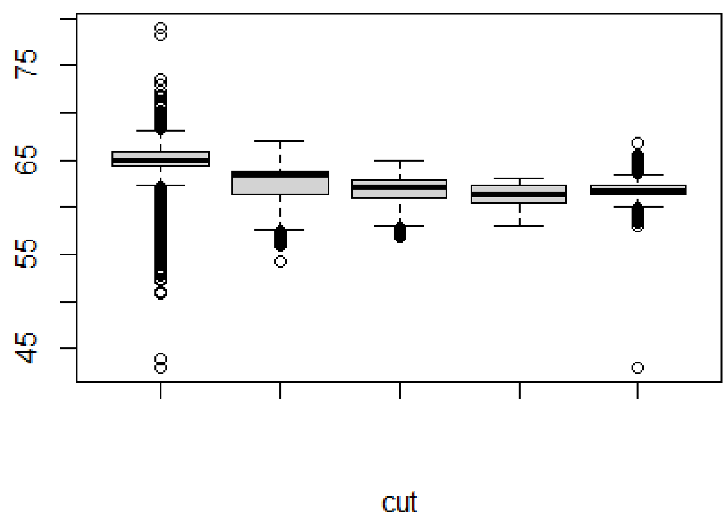
```
barplot(table(cut,depth),legend.text = T,col=blues9,xlab='depth',ylab='freque
ntly')
```



it seems like the relation between depth and frequency has a normal distribution.

up until the depth is around average ,with increase in depth,we have more cut levels and from average to max depth this decrease.

```
s1=depth[cut=="Fair"]
s2=depth[cut=="Good"]
s3=depth[cut=="Very Good"]
s4=depth[cut=="Premium"]
s5=depth[cut=="Ideal"]
boxplot(s1,s2,s3,s4,s5,Horizontal=T,xlab='cut')
```

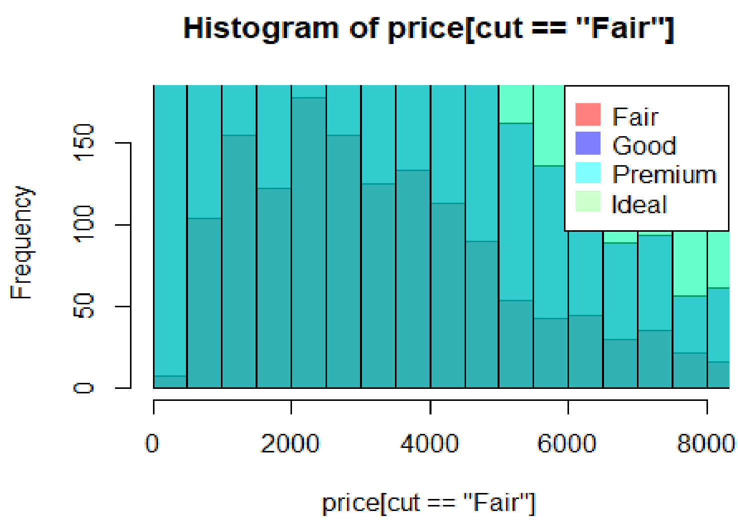


As a result of using boxplot, it can be seen that as the cut category chagnes from the lowest level “Fair” to highest level “Ideal” the mean of depths decrease slowly. so we can say in average the change in cut category can affect depth.

۵. توزیع متغیر قیمت را برای درجه های تراش مختلف مقایسه کنید. آیا چیزی غیرعادی به نظر می رسد؟ توضیح دهید.

5.

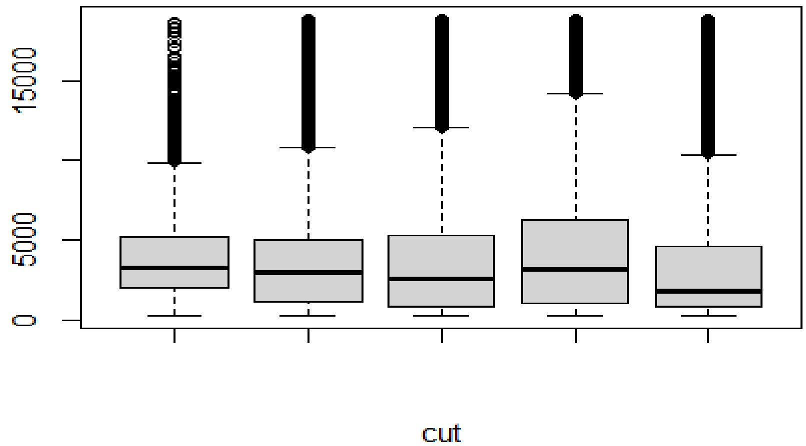
```
hist(price[cut=="Fair"],breaks=50,col=rgb(1,0,0,0.5),xlim=c(0,8000))
hist(price[cut=="Good"],breaks=50,col=rgb(0,0,1,0.5),add=T)
hist(price[cut=="Premium"],breaks=50,col=rgb(0,1,1,0.5),add=T)
hist(price[cut=="Ideal"],breaks=50,col=rgb(0,1,0,0.2),add=T)
legend("topright",legend=c("Fair","Good","Premium","Ideal"),col=c(rgb(1,0,0,0
.5),rgb(0,0,1,0.5),rgb(0,1,1,0.5),rgb(0,1,0,0.2)),
      pt.cex=2,pch=15)
```




```
s1=price[cut=="Fair"]
s2=price[cut=="Good"]
s3=price[cut=="Very Good"]
s4=price[cut=="Premium"]
s5=price[cut=="Ideal"]
boxplot(s1,s2,s3,s4,s5,Horizontal=T,xlab='cut')
```

As it is shown in the box plot above, the most frequent prices in all 5 categories of cut are amost equal and below 5000, so most of our diamonds' prices in every kind of cut are below 5000.

۶. با استفاده از یک جدول توافقی و یک نمودار میله ای رابطه بین دو متغیر رنگ و درجه تراش را توصیف کنید.

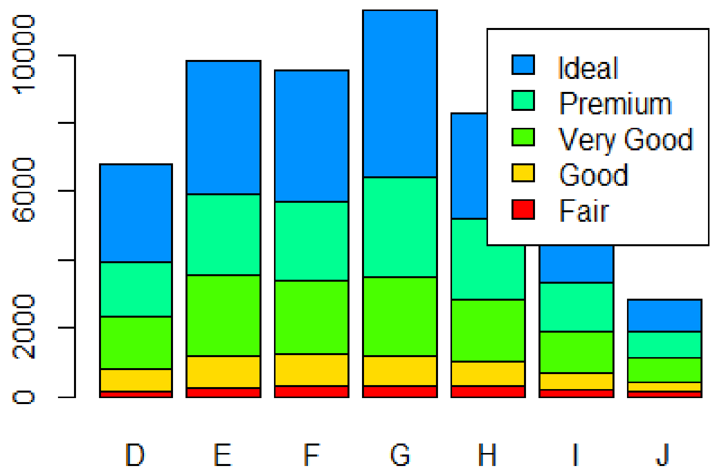


6.

```r
#contigency Table
table(color,cut)

##      cut
## color Fair Good Very Good Premium Ideal
##     D  163  662      1513    1603  2834
##     E  224  933      2400    2337  3903
##     F  312  909      2164    2331  3826
##     G  314  871      2299    2924  4884
##     H  303  702      1824    2360  3115
##     I  175  522      1204    1428  2093
##     J  119  307       678     808   896

#it's obvious the better the cut degree gets, the more frequent is the use of that color.
```

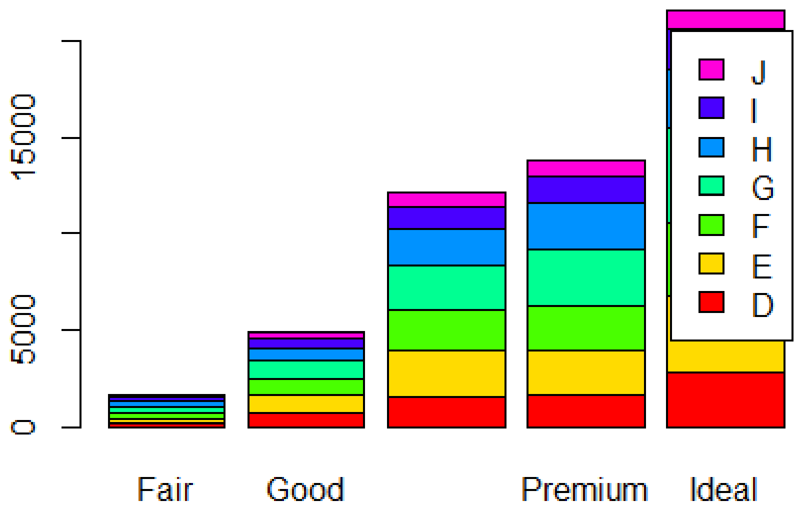
It is obvious the better the cut degree gets, the more frequent is the use of that color.

```r
barplot(table(cut,color),col=rainbow(7),legend.text = T)
```



In this bar plot,we can see that there is more frequency of each color at Ideal cut and it decrease as cut degrees getting to lower levels.

```
barplot(table(color,cut),col=rainbow(7),legend.text = T)
```



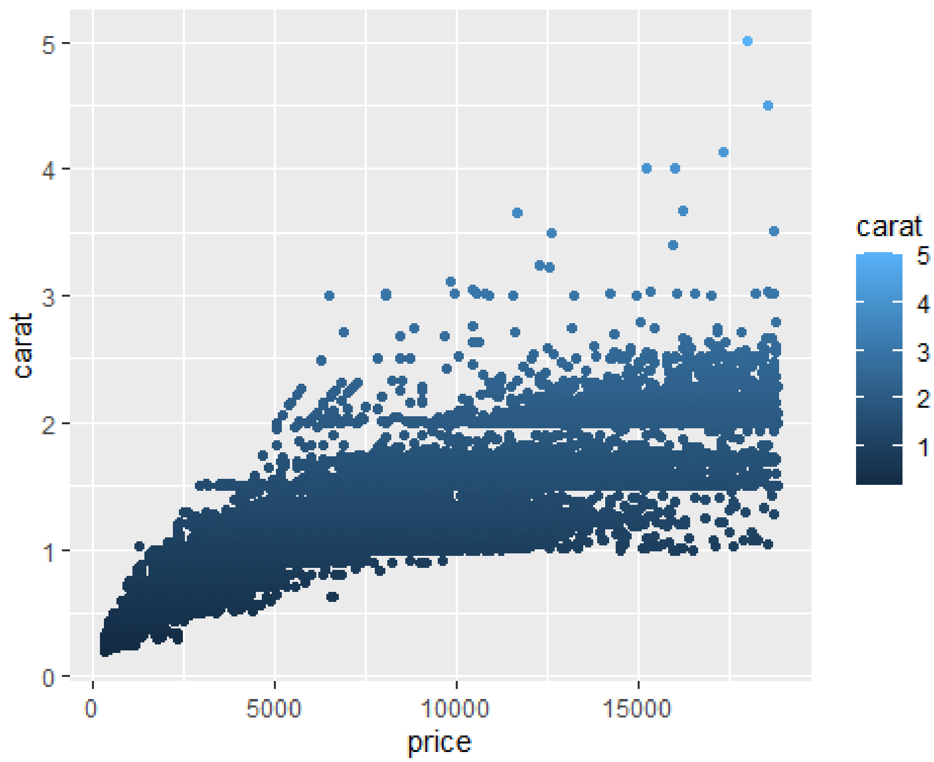
As it shows , all the degree of cuts include all 7 color but the frequency of color in

higher level of cut is more than basic level.

۷. با استفاده از یک نمودار پراکندگی رابطه بین دو متغیر قیمت و قیراط را تشریح کنید. آیا الگویی در نمودار دیده می شود؟ سپس این نمودار را فقط برای الماس های دارای قیمت کوچکتر یا مساوی ۱۰۰۰ دلار، با و بدون استفاده از jittering رسم کنید. آیا استفاده از jittering سبب تغییر در شکل ظاهری نمودار و تحلیل آن می شود؟ توضیح دهید.

7.

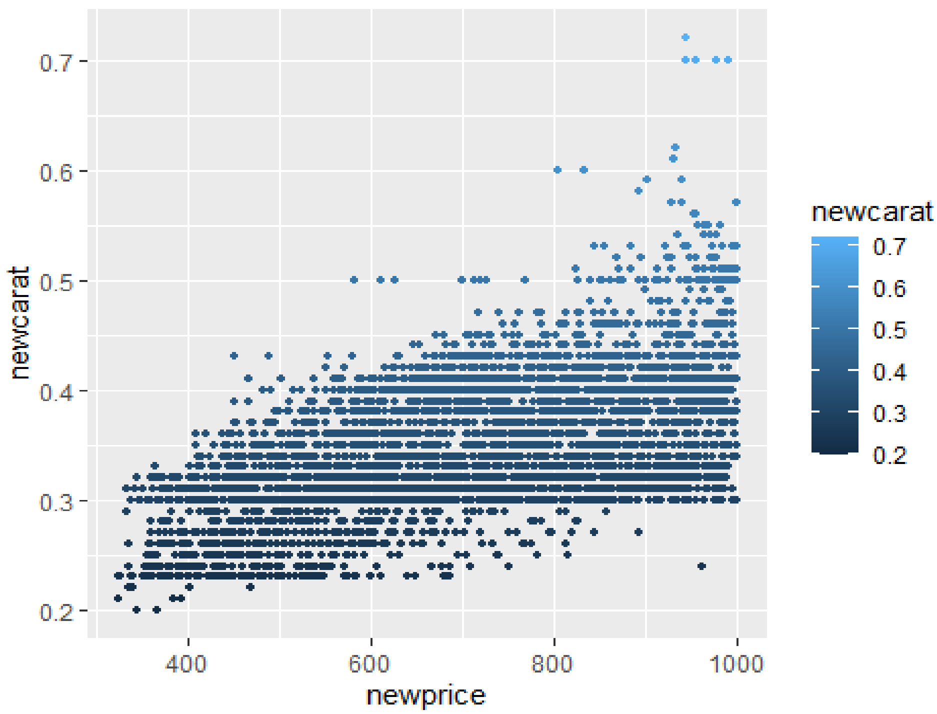
```
ggplot(data=diamonds)+
geom_point(mapping=aes(x=price,y=carat,color=carat))
```



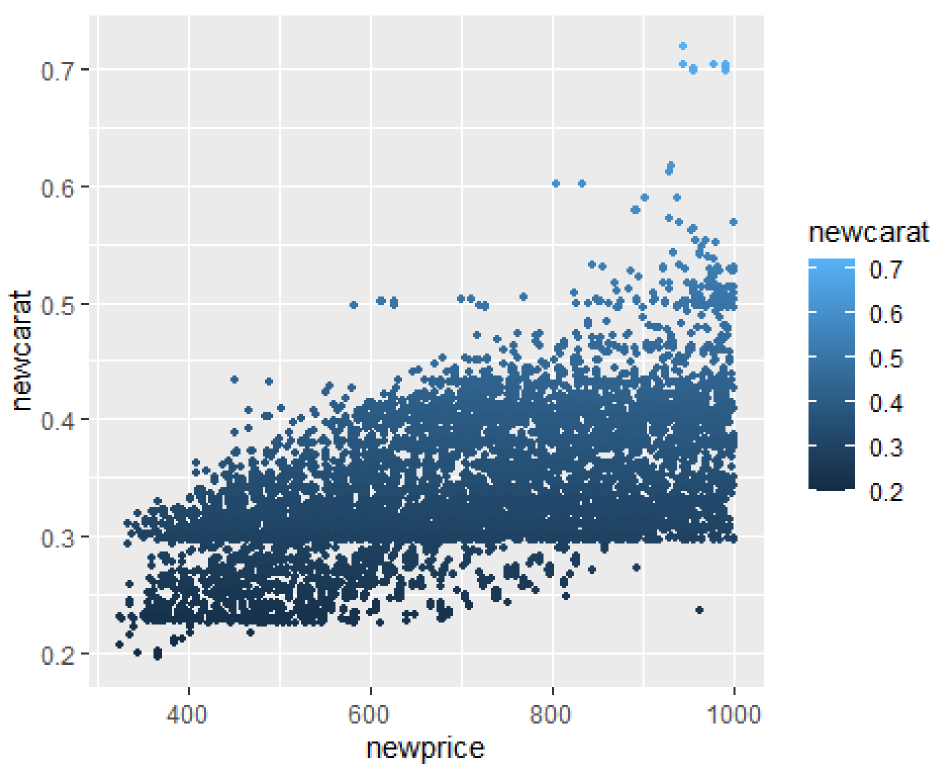
As it's shown in the plot, the density near the price 0 to 10000 is too much bt it's almost steady with the rate of carat between 0 to less than 2. as the price goes up and carat increase, the density lowers down and we can have visible scatter plots. so with the higher price and carat between 0 to 2, the density is high.

```
newprice=price[price<=1000]
#table(newprice)
newcarat=carat[price<=1000]
#table(newcarat)
newdiamonds=subset(diamonds,price<=1000)
ggplot(data=newdiamonds)+
geom_point(mapping=aes(x=newprice,y=newcarat,color=newcarat),size=1)
```

Now with this new data consists of only prices less than 1000, we have a less dense plot, as it is shown, the density in parts where carat is 0.2 to 0.4 is stable and a lot more than other parts. As price goes higher and so does the carat the density and frequency lower down.we can conclude that in almost any price less than 1000 , when the carat is between 0.2 to 0.45 , the frequency is more than other time and it's stable.




```
ggplot(data=newdiamonds)+
geom_point(mapping=aes(x=newprice,y=newcarat,color=newcarat),size=1,position
= "jitter")
```

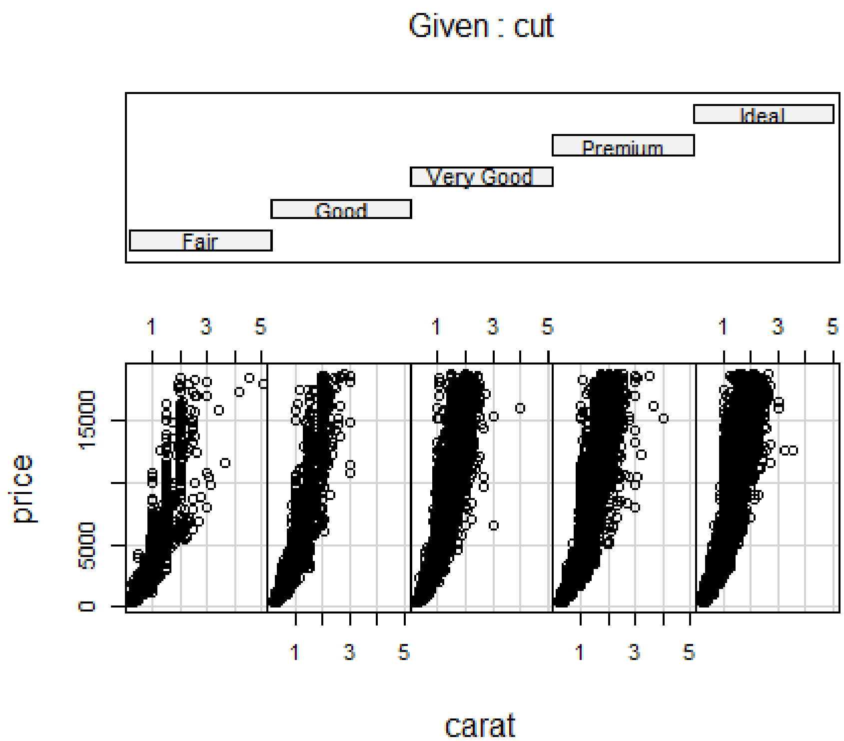


Using jitters did not really change the plot, the same high density and frequency is shown.

۸. رابطه بین دو متغیر قیمت و قیراط را بر اساس متغیر درجه تراش تشریح کنید. سعی کنید از ویژگی های ظاهری مختلف برای رسم نمودارها استفاده کنید.

8.

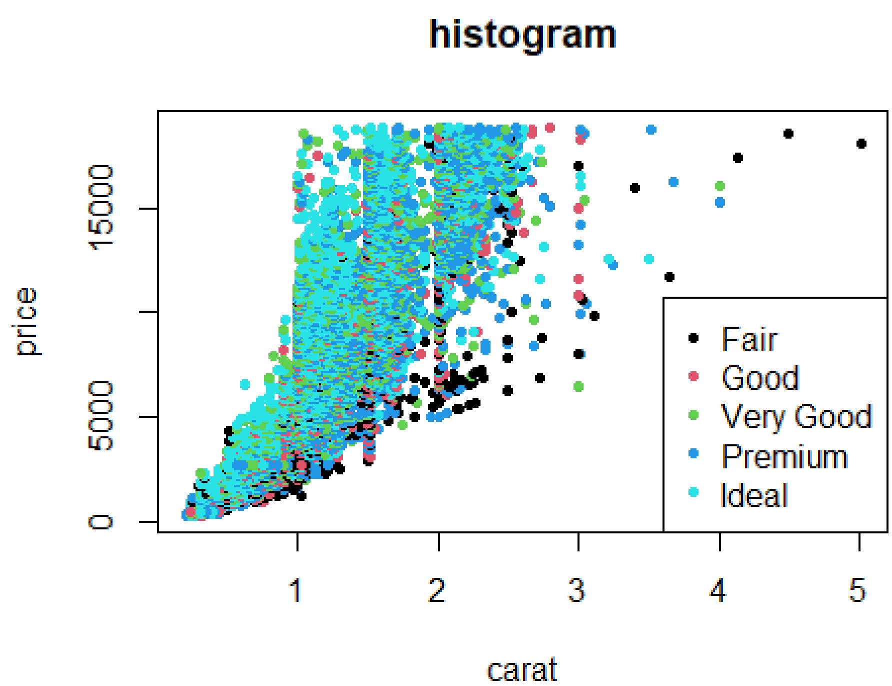
```
par(mfrow=c(1,1))
coplot(price~carat|cut,rows=1)
```



In almost in every cut degree, there is high density of diamonds with price between 0 to 10000 and carat between 0 to 3. at some cut degree such as Fair and Good the density in lower prices is higher than higher prices whereas in very Good , Premium and Ideal cut the density between carat 0 to 3 is high in almost every price.

```
plot(price~carat,
     xlab = "carat",
     ylab = "price",
     main = "histogram",
     pch  =20,
     cex  = 1,
```

```
     col   = factor(cut, labels = 1:5))
legend("bottomright",
       legend = levels(factor(cut)),
       pch   = 20,
       col   = 1:5)
```

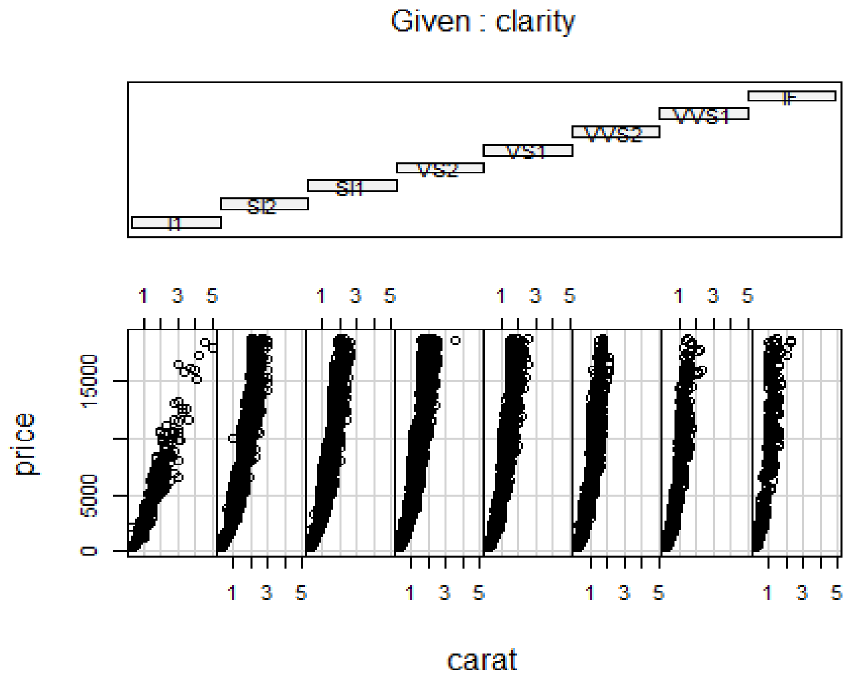


This plot like the previous one shows high density of three high cut degree “Very good(green)”,“Premium(dark blue)”,“Ideal(light blue)” in almost each price in carat in the range of 0 to less than 3. also cut such as “Good” and “Fair are in higher density when the price is lower than 10000 , the same result as”coplot”.

۹. با استفاده از نمودار پراکندگی، رابطه بین دو متغیر قیمت و قیراط را بر اساس متغیر درجه شفافیت تشریح کنید. سپس، متغیرهای قیمت و قیراط را در مقیاس لگاریتمی در نمودار پراکندگی وارد کنید و نمودارها را تحلیل کنید.

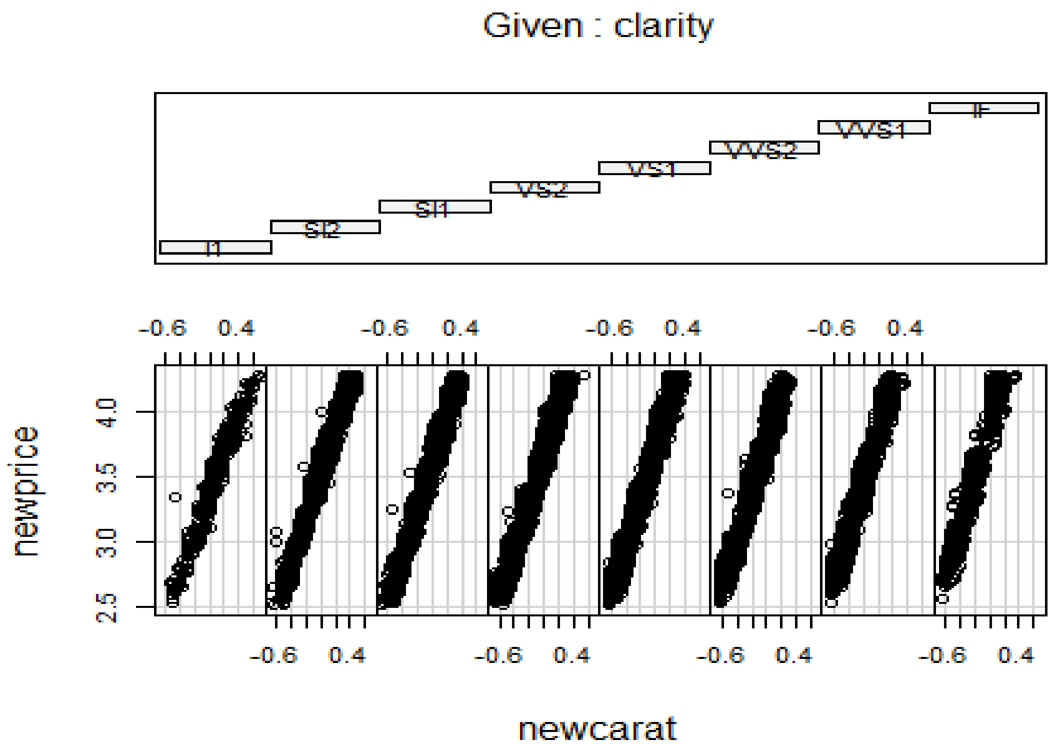
9.

```
coplot(price~carat|clarity,rows=1)
```

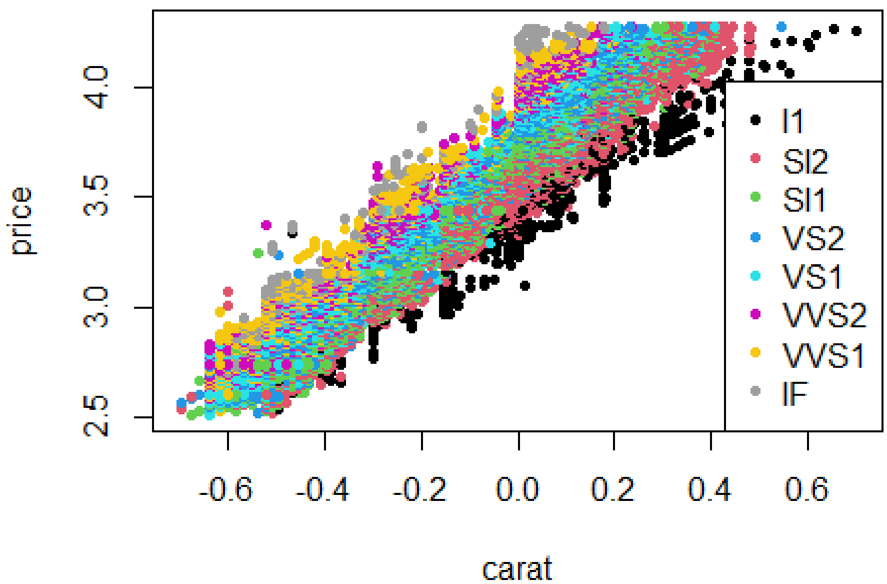


As it is shown in the chart, the density of diamonds with carat between 0 to3 is high in almost every clarity specially in SI2 ,SI1,VS2,Vs1,VVS2 this high density is illustrated better. this density is distributed almost evenly in each price.there is visible that in I1 clarity there is only high density in carat between 0 to 3 in prices only from0 to 10000. In clarity of VVS1,IF we can see that the density is lower comparing to other 5 previous clarity and from prices higher than 15000 , the density starts to decline gradually.

```
newcarat=log10(carat)
#newcarat
newprice=log10(price)
#newprice
coplot(newprice~newcarat|clarity,rows=1)
```




```
plot(newprice~newcarat,
     xlab="carat",
     ylab="price",
     pch=20,
     cex=1,
     col=factor(clarity,labels=1:8))
legend("bottomright",
       legend = levels(factor(clarity)),
       pch   = 20,
       col   = 1:8)
```

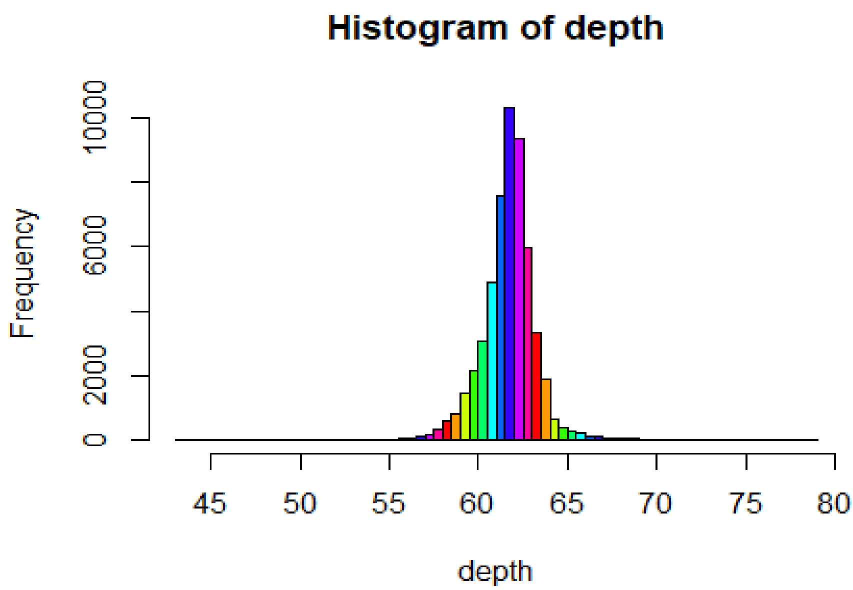


As same as the result with coplot, there is high density of SI2 ,SI1,VS2,Vs1,VVS2 in carat higher than -0.2 t0 0.4(with log10) each illustrated with distinctive color,as it's shown there is almost an even distribution of density in lower and higher prices with this specific clarity, so we can conclude that price is not an effective factor on density of clarity with different carat.

۱۰. **(تمرین تشویقی)** سعی کنید ۲ مورد تحلیل دیگر با استفاده از ابزارهای مختلف EDA بر روی این مجموعه داده ارائه دهید. خلاق باشید و در هر یک از تحلیل ها حداقل ۳ متغیر را دخیل کنید.

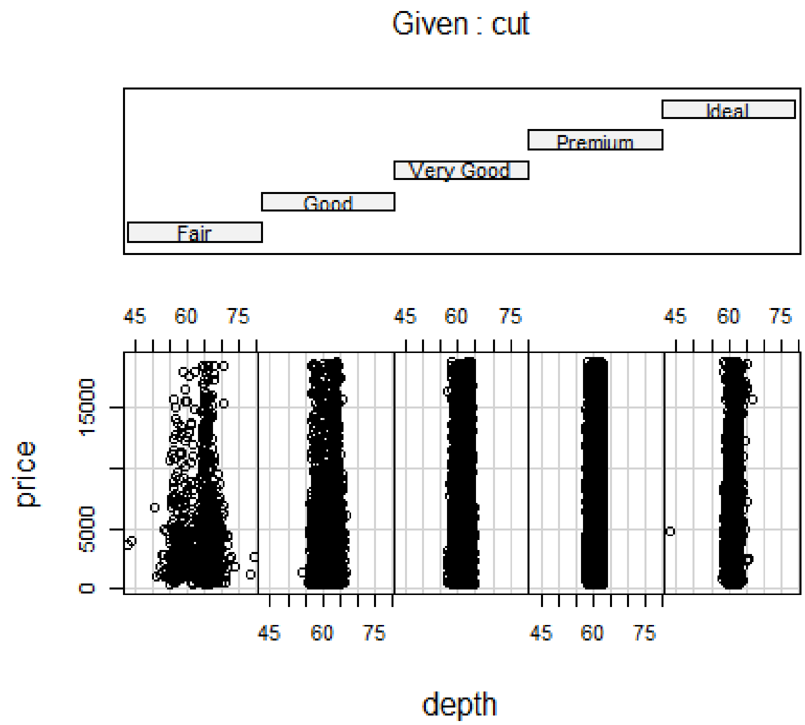
10-1

```
m=as.data.frame(table(depth))

hist(depth,breaks=100,col=rainbow(10))
```



There is obvious that the depth of diamonds can have normal distribution.

```
coplot(price~depth|cut,rows=1)
```

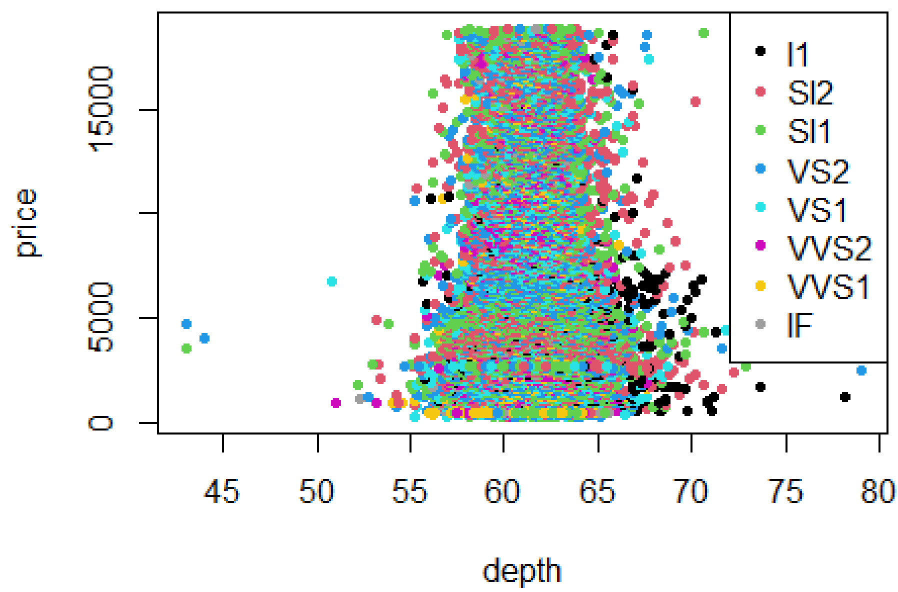


As it is illustrated in the charts, we have 5 cut category and except for the Fair categoyr, there density for depth between 55 to 65 in any price is high for any cut category. this means that the better the cut category gets, the depth doesn't change and probably the depth in range of 55 to 65 is the ideal for all sorts of cut and price. in Fair category, because it's not as good as others, the density has spread around 50 to 70 in depth and there is higher density in 65 to 70 in almost each price. it has to be mentioned that price is not the key factor in this chart because depth is not related to it in any cut category.

10-2

```
plot(price~depth,
          xlab="depth",
          ylab="price",
          pch=20,
          cex=1,
          col=factor(clarity,labels=1:8))
legend("topright",
       legend = levels(factor(clarity)),
       pch  = 20,
       col  = 1:8)
```

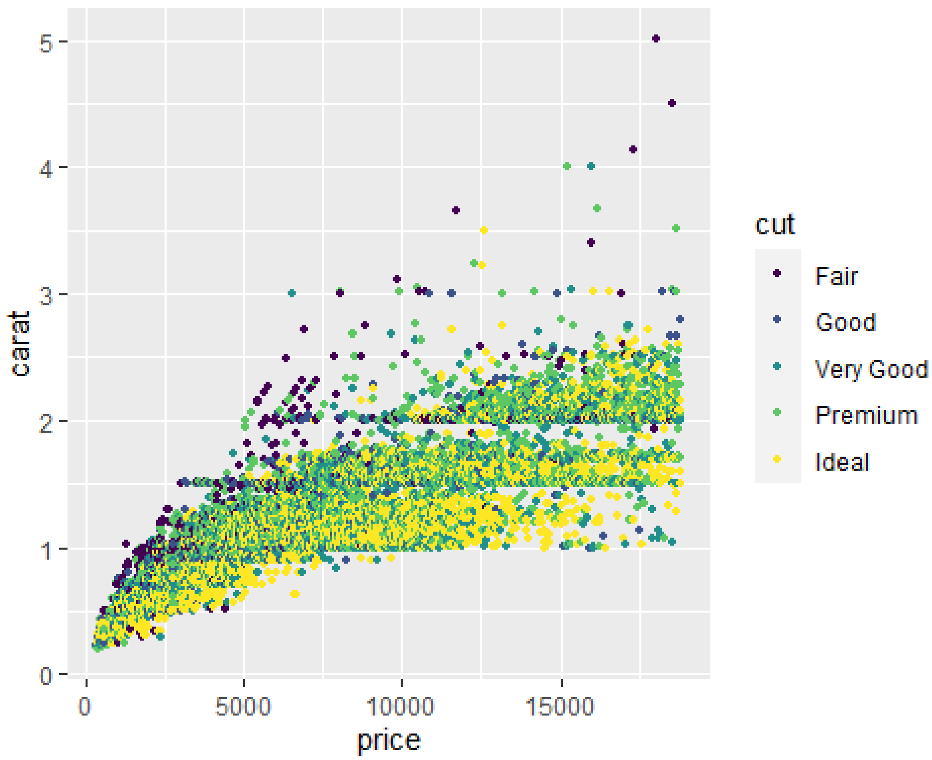
Like the previous chart , this one resulted the same, all the clarity



categories,have high density in depth between 55 to 65, this range of depth is the most frequent and it has nothing to do with price factor cause the density is evenly distributed, regardless of price.

10-3

```
ggplot(data =diamonds)+
  geom_point(mapping=aes(price,carat,color=cut),size=1)
```



As we can see, the frequency of ideal cut is overall more than other kind of cut, also the high density of diamonds with carat between 0.2 to 0.45 is is stable in almost all price. we can conclude that between carat 0.2 to 0.45, there is more diamonds with cut of "prumium" and " ideal" , the more the price goes up, the more these two categories frequency gets. As we go further in x(price) axis,we can see color blue is getting fade. it shows that diamonds with same carat but diffrent cut, can have different prices.

Part2 : Statistic analysis

۱۱. آیا رابطه معناداری میان دو متغیر رنگ و درجه تراش وجود دارد؟

11.

{H0: there is not a meaningful relationship between color and cut

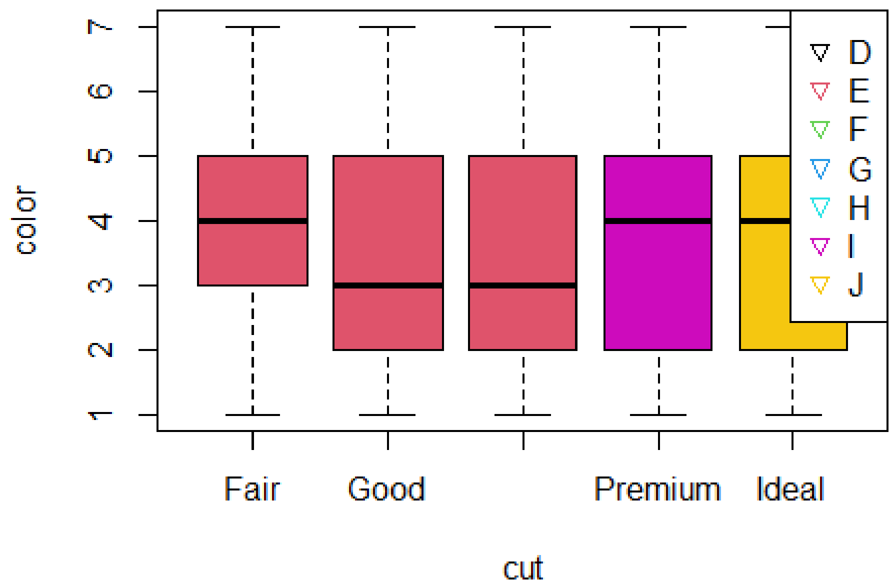
{H1: otherwise.

```
table1=table(cut,color)
chi2=chisq.test(table1)
chi2

##
##  Pearson's Chi-squared test
##
## data:  table1
## X-squared = 310.32, df = 24, p-value < 2.2e-16

kable(chi2$expected>5)
```

	D	E	F	G	H	I	J
Fair	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Good	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Very Good	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Premium	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Ideal	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

All the expected frequencies are greater than 5 so we can say this test is working fine. since the p_value is less than 0.05 ,we can conclude that the h0 is not acceptable and so there is a relationship between cut and color.

```
boxplot(color~cut,
     xlab="cut",
     ylab="color",
     pch=20,
     cex=1,
     col=factor(color,labels=1:7))
legend("topright",
       legend = levels(factor(color)),
       pch  = 25,
       col  = 1:7)
```



There is also visible that first 3 cut category are mostly color E and as the category gets better the color changes from I to J.

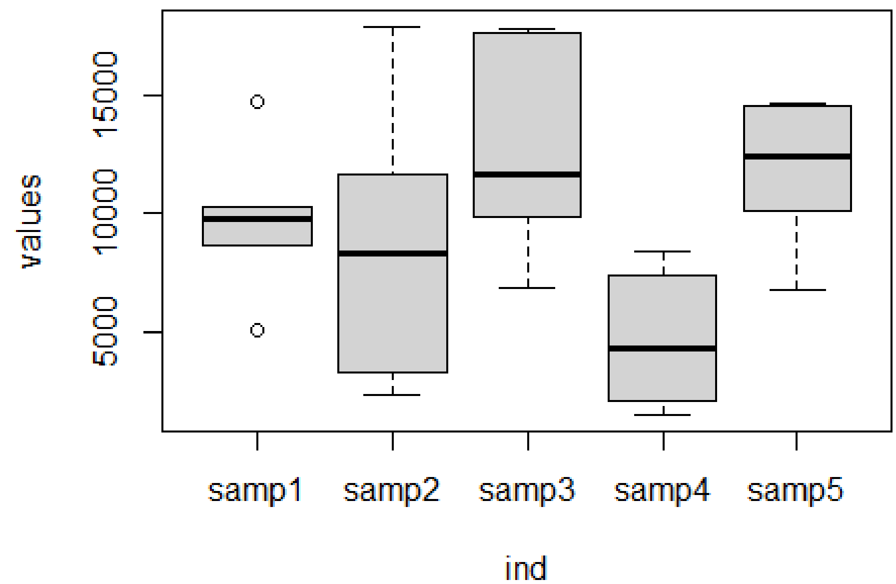
۱۲. آیا رابطه معناداری میان دو متغیر قیمت و درجه تراش وجود دارد؟ اگر بله، مشـــخص کنید که قیمت بین کدام ســـطوح متغیر درجه تراش متفاوت است؟

12.

{H0: No releation ship.(m1!=m2!=m3!=m4!=m5)

{H1: otherwise(m1=m2=m3=m4=m5)

```
price1=(price[cut=="Fair"])
price2=(price[cut=="Good"])
price3=(price[cut=="Very Good"])
price4=(price[cut=="Premium"])
price5=(price[cut=="Ideal"])
samp1=runif(5,min(price1),max(price1))
samp2=runif(5,min(price2),max(price2))
samp3=runif(5,min(price3),max(price3))
samp4=runif(5,min(price4),max(price4))
samp5=runif(5,min(price5),max(price5))
table2=stack(data.frame(samp1,samp2,samp3,samp4,samp5))
#table2
boxplot(values~ind,data=table2)
```



There is shown in the plots that the means of this 5 cuts are not same, now we can do ANOVAL test to make sure of the result.

```
oneway.test(values~ind,data=table2,var.equal = T)

##
##  One-way analysis of means
##
## data:  values and ind
## F = 2.5159, num df = 4, denom df = 20, p-value = 0.07381
```

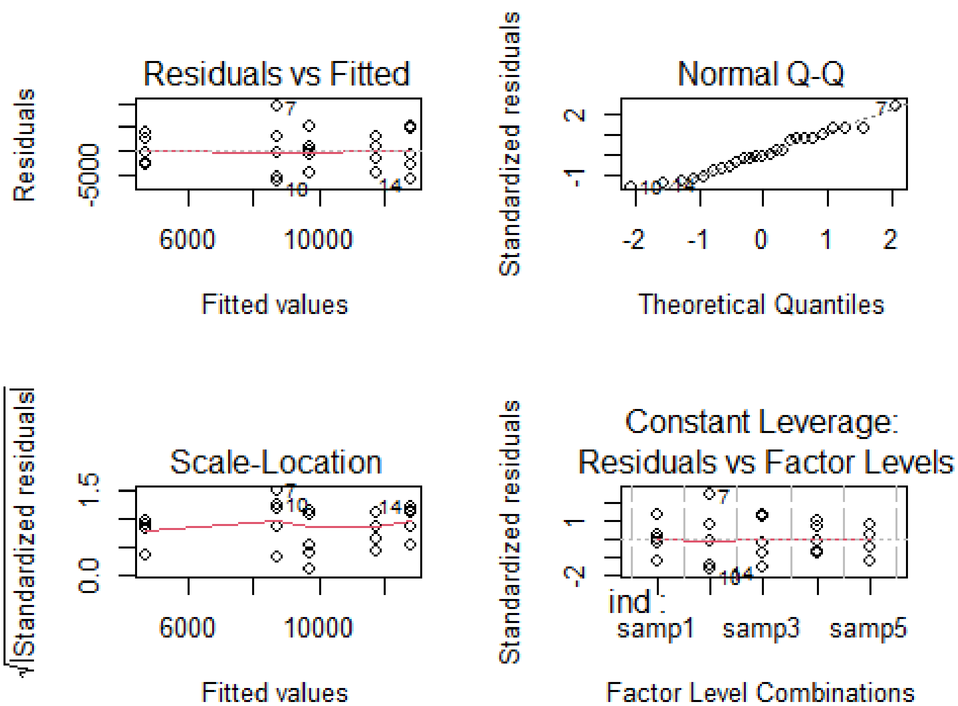
p_value is wa greater than 0.05 so we can say that H0 as all the means of samples are equal is acceptable.

```
anova(lm(values~ind,data=table2))

## Analysis of Variance Table
##
## Response: values
##           Df    Sum Sq  Mean Sq F value  Pr(>F)
## ind        4 195186083 48796521  2.5159 0.07381 .
## Residuals 20 387911314 19395566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
AN=aov(values~ind,data=table2)
AN

## Call:
##    aov(formula = values ~ ind, data = table2)
##
## Terms:
##                       ind Residuals
## Sum of Squares  195186083 387911314
## Deg. of Freedom         4        20
##
## Residual standard error: 4404.04
## Estimated effects may be unbalanced
```

This test gives us the same result as well.

```r
par(mfrow=c(2,2))
plot(AN)
```




```r
par(mfrow=c(1,1))
```

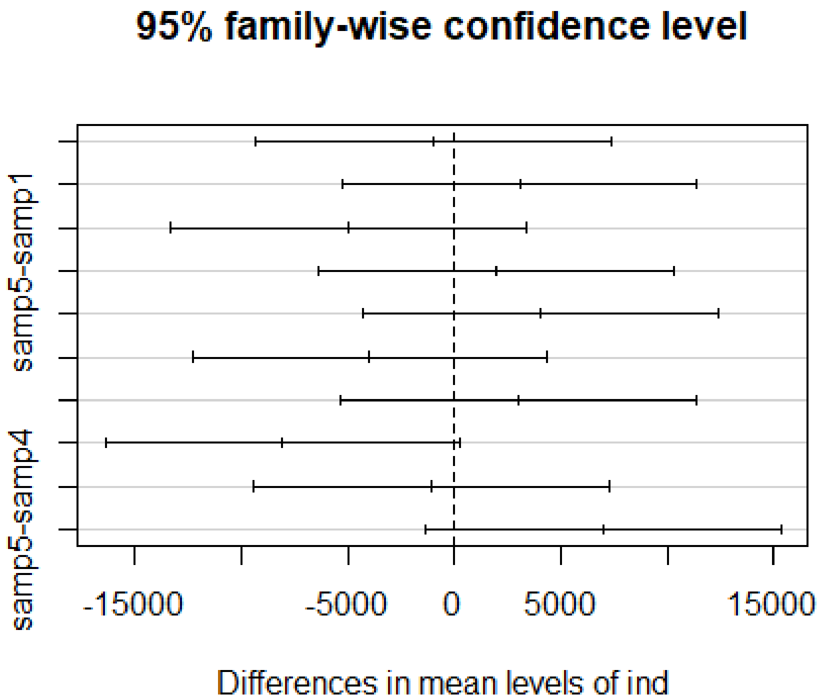
We can see that the hypothesis about variances equality is correct ,also the hypothesis about normal distribution is correct too.

```r
kruskal.test(values~ind,data=table2)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  values by ind
## Kruskal-Wallis chi-squared = 8.1009, df = 4, p-value = 0.08795
```

Because the p-value is greater than 0.05, we accept the h0 in this test,so all the means are equal.

```
TukeyHSD(AN)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = values ~ ind, data = table2)
##
## $ind
##                  diff        lwr        upr     p adj
## samp2-samp1 -1011.940  -9346.778  7322.8986 0.9959648
## samp3-samp1  3060.385  -5274.453 11395.2236 0.8051345
## samp4-samp1 -4981.484 -13316.322  3353.3543 0.4070314
## samp5-samp1  2001.410  -6333.429 10336.2478 0.9497348
## samp3-samp2  4072.325  -4262.513 12407.1632 0.5971413
## samp4-samp2 -3969.544 -12304.383  4365.2938 0.6194717
## samp5-samp2  3013.349  -5321.489 11348.1874 0.8136012
## samp4-samp3 -8041.869 -16376.708   292.9689 0.0618832
## samp5-samp3 -1058.976  -9393.814  7275.8624 0.9951938
## samp5-samp4  6982.894  -1351.945 15317.7317 0.1287236

plot(TukeyHSD(AN))
```



The P adj of all sample's pairs are greater than 0.05 so H1 is not acceptable in any of them , it can be mentioned that non of the means are similar.plus according to the plot, 0 is in all the confidence interval of pairs compared. so it is proven the means are equal and there is a relationship between price and cut.

برای حل قســـمت های بعدی، زیر مجموعه ای از داده ها را در نظر بگیرید که مقدار متغیر قیراط الماس ها کوچکتر از ۰.۵ باشـــد و برای این زیرمجموعه به قسمت های بعدی پاسخ دهید.

۱۳. آیا میانگین متغیر قیمت برای الماس های دارای درجه تراش Fair و Good اختلاف معناداری دارد؟

13.

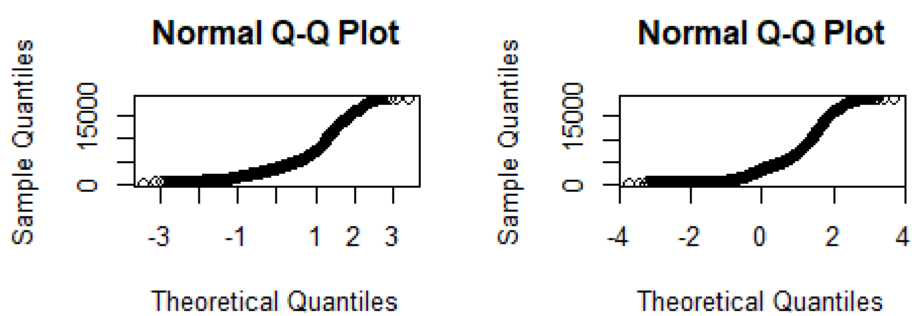
{H0: m(Fair)=m(Good)

{H1: otherwise

```
newdiamonds=subset(diamonds,carat<0.5)
newdiamonds

## # A tibble: 17,674 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
##  2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
```

```
## 3 0.23 Good      E     VS1      56.9    65    327  4.05   4.07   2.31
## 4 0.29 Premium   I     VS2      62.4    58    334  4.2    4.23   2.63
## 5 0.31 Good      J     SI2      63.3    58    335  4.34   4.35   2.75
## 6 0.24 Very Good J     VVS2     62.8    57    336  3.94   3.96   2.48
## 7 0.24 Very Good I     VVS1     62.3    57    336  3.95   3.98   2.47
## 8 0.26 Very Good H     SI1      61.9    55    337  4.07   4.11   2.53
## 9 0.22 Fair      E     VS2      65.1    61    337  3.87   3.78   2.49
## 10 0.23 Very Good H    VS1      59.4    61    338  4      4.05   2.39
## # ... with 17,664 more rows
```

```
par(mfrow=c(2,2))
qqnorm(price[cut=="Fair"])
qqnorm(price[cut=="Good"])
x1=price[cut=="Fair"]
y1=price[cut=="Good"]
```



Both of them are not normal but have the same distribution shape, so we use Mann-Whitney test.

```
wilcox.test(x,y,alternative="two.sided")

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 1450451080, p-value = 0.3993
## alternative hypothesis: true location shift is not equal to 0
```

p-value is less than 0.05, so we reject the H0 and there is meaningful difference between the tow means.

۱۴. آیا نسبت الماس هایی که دارای قیمت بیشتر از ۱۰۰۰ دلار هستند، برای درجه تراش های Fair و Good اختلاف معناداری دارد؟

14.

{H0: P(Fair)=P(Good)

{H1: otherwise.

```
p1=sum(price[cut=="Fair"]>1000)/length(price)
p2=sum(price[cut=="Good"]>1000)/length(price)
n=c(length(price[cut=="Fair"]),length(price[cut=="Good"]))
x=c(sum(price[cut=="Fair"]>1000),sum(price[cut=="Good"]>1000))
prop.test(x,n,alternative = "two.sided",correct = F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x out of n
## X-squared = 197.77, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1406757 0.1747586
## sample estimates:
##    prop 1    prop 2
## 0.9310559 0.7733388
```

As the result shows, p-value is less than 0.05, so we reject the H0 and we can conclude there is a meaningful difference between p1 and p2.

Conclusion

This Report is created by Rmarkdown and aims to illustrate the analysis and reasons behind the chunks of R codes and statistical calculations.

I hope you a have found this Report appropriate to your use.

Thank you for your attention.

Mina Kanaani.