Assessed Coursework Assignment 1

CS5250 Visualisation and Exploratory Analysis

Part -1

1. The data consist of 53940 observations with ten 10 variables. The str() function show the data structure in depth. From the function we note that variables; cut, color and clarity are categorical variables.

Each of the variables is made of different levels as shown below.

```
> levels(cut)

[1] "Fair"      "Good"      "Very Good" "Premium"   "Ideal"

> levels(color)
[1] "D" "E" "F" "G" "H" "I" "J"
> levels(clarity)
[1] "I1"   "SI2"  "SI1"  "VS2"  "VS1"  "VVS2" "VVS1" "IF"
```

cut refers to how well-proportioned the dimensions of a diamond are, and how these surfaces, or facets, are positioned to create sparkle and brilliance. The ideal cut is the most expensive while Fair come in last.

The color variable represent the grading of diamonds. It can be broken down as the following;

D-F The rarest and highest quality with a pure icy look.

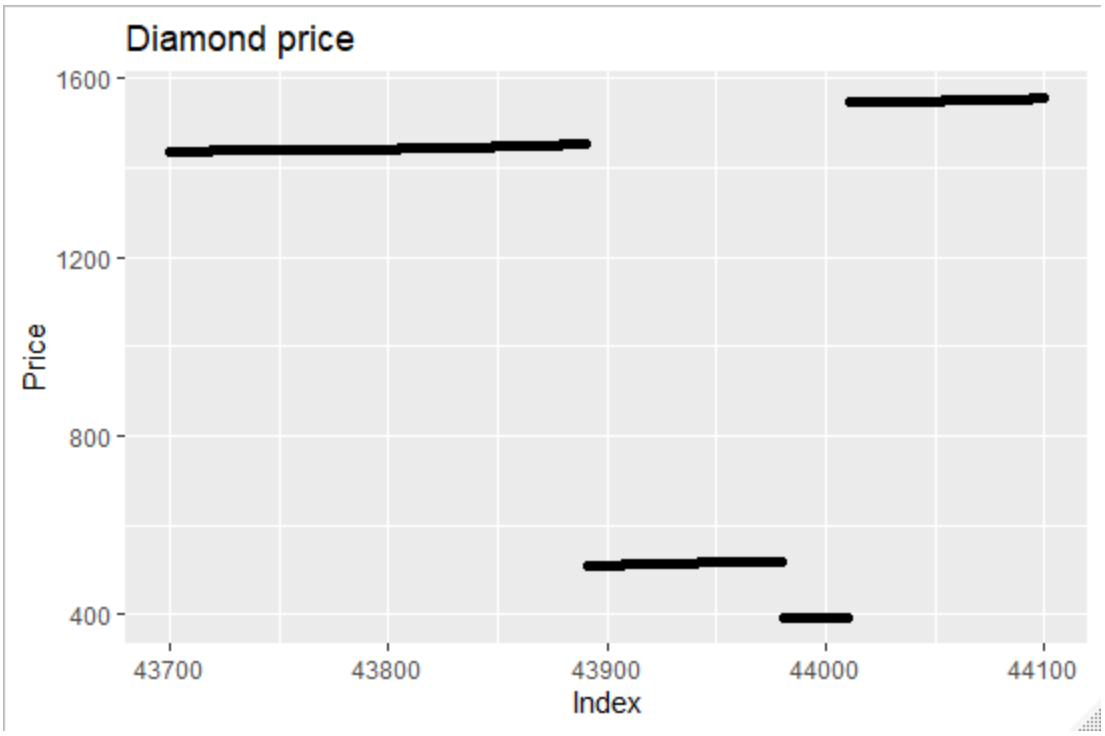
G-H & I-J Near-colourless diamonds: No discernible colour; great value for the quality.

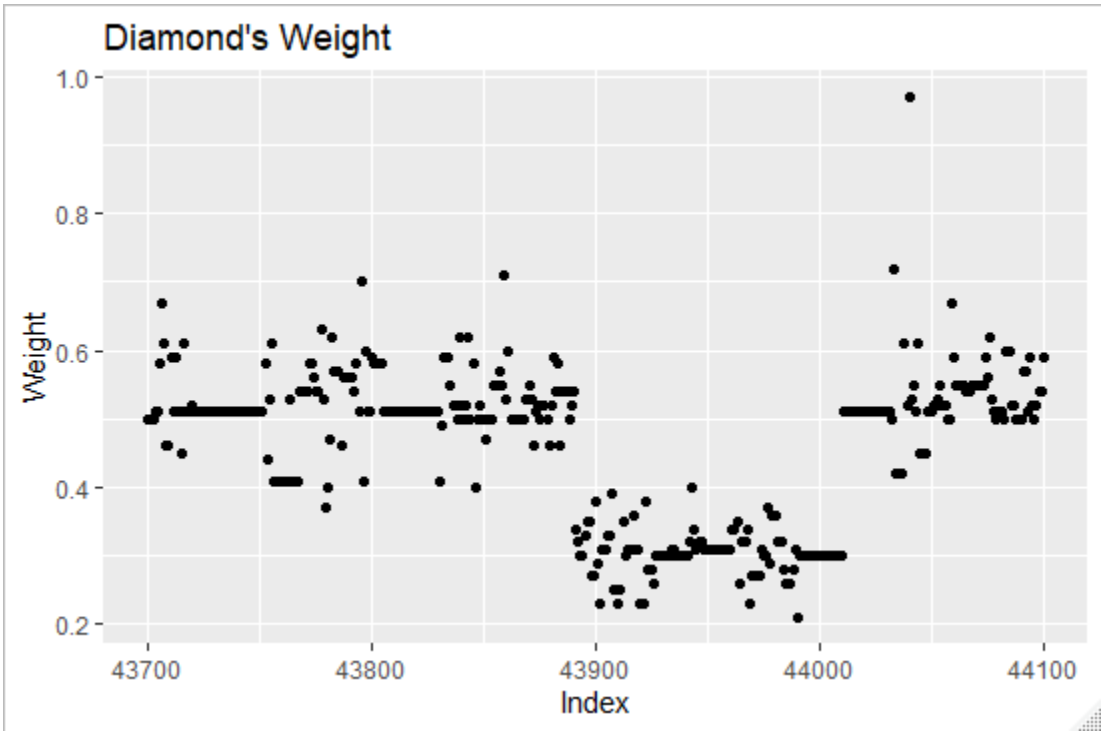
K - Faint colour diamonds: Budget-friendly pick; pairs beautifully with yellow gold.

Hence D and F would represent the most expensive.

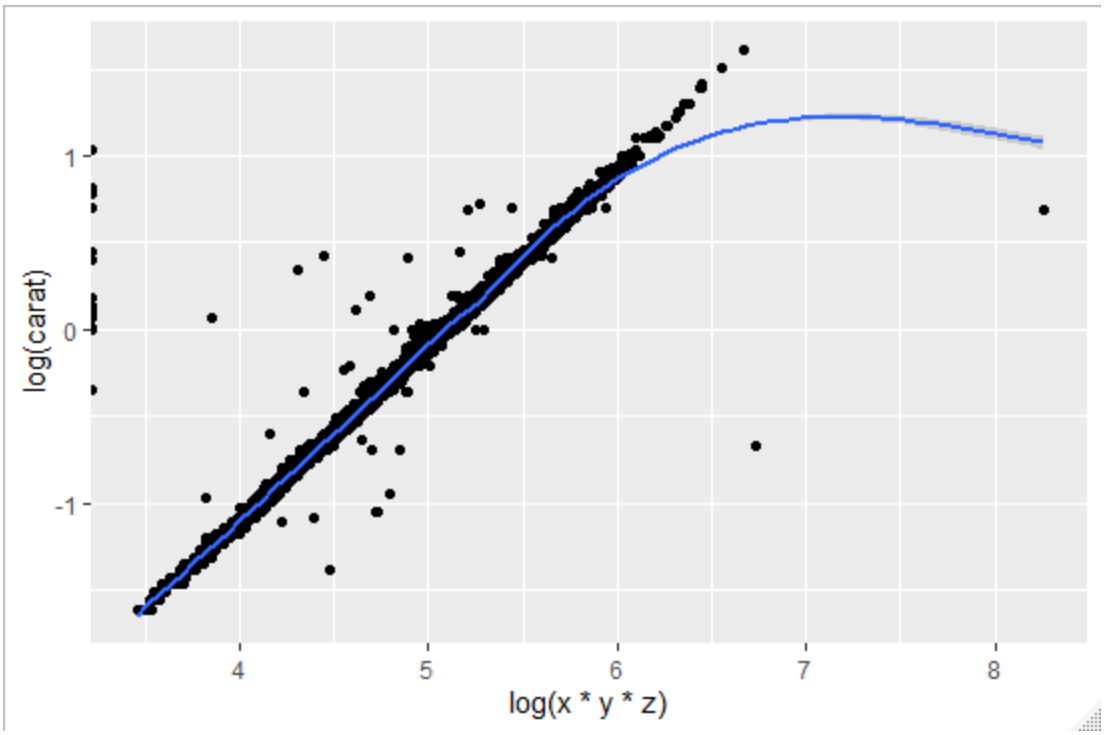
Diamond clarity can be defined as the assessment of small imperfections on the surface and within the stone. By clarity, VVS1, VVS2 are the most expensive. Diamond that appears the same to the naked eyes will require experts and technology to Identify them hence their value will be higher.

2.


Diamond price
1600
1200
800
400
Price
43700
43800
43900
44000
44100
Index


Diamond's Weight
1.0
0.8
0.6
0.4
0.2
Weight
43700
43800
43900
44000
44100
Index

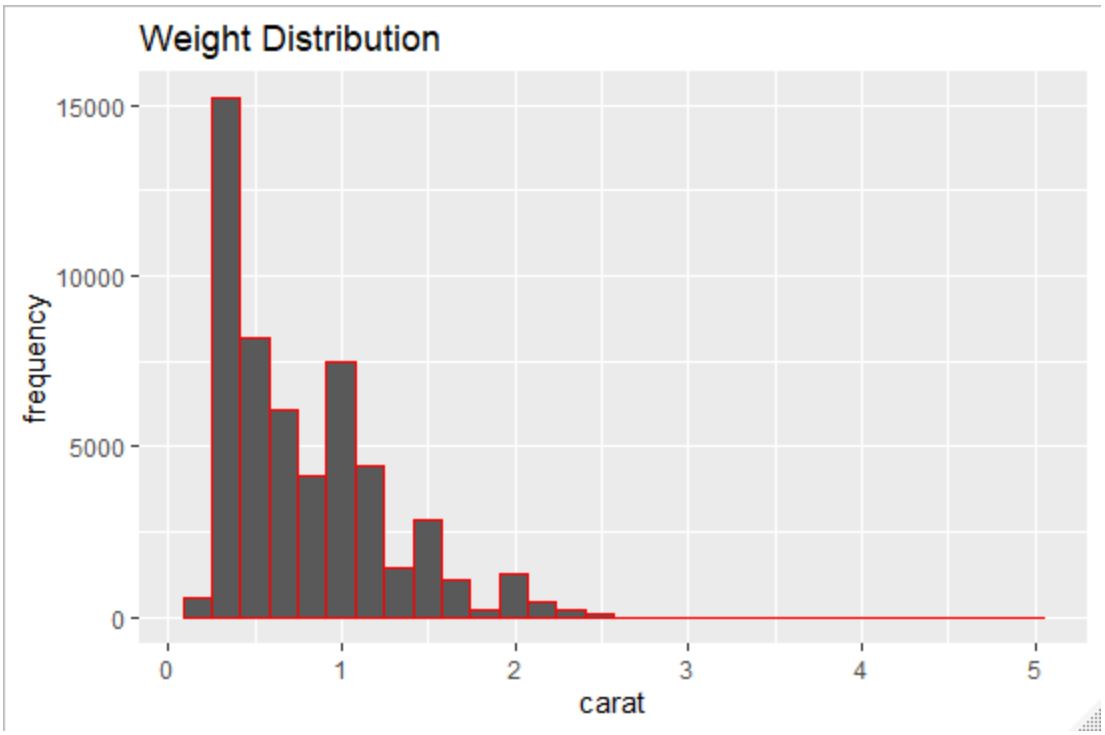
b)



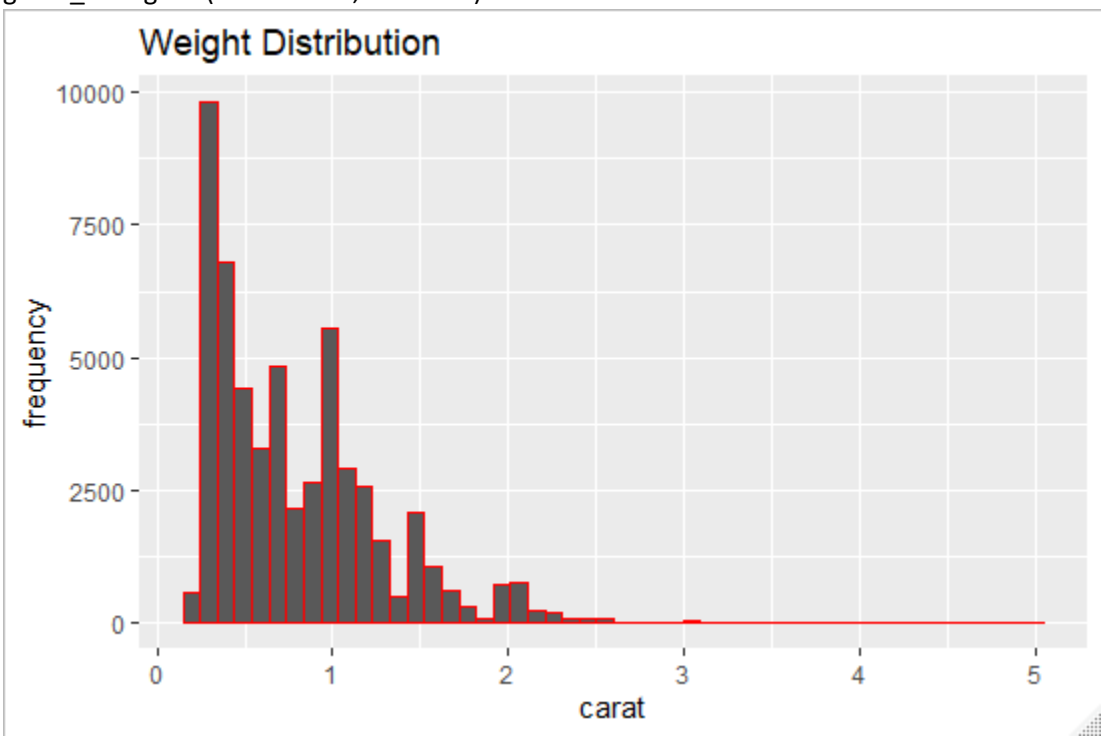
There seems to be some values that are extremely large. In the x*y*z value. By using the which.max() function we get and index of 24068. If we subset the data at this point we note that y has a value of 58.9 which is an outlier. The true value could have been 5.89 but it was entered as 58.9.

c) Distributions of the variables.

```
#histogram for weights
p <- ggplot(diamonds, aes(x=carat)) + geom_histogram(color="red", bins = 30)
p+ggtitle("Price Distribution")+xlab("carat")+ylab("frequency")
```

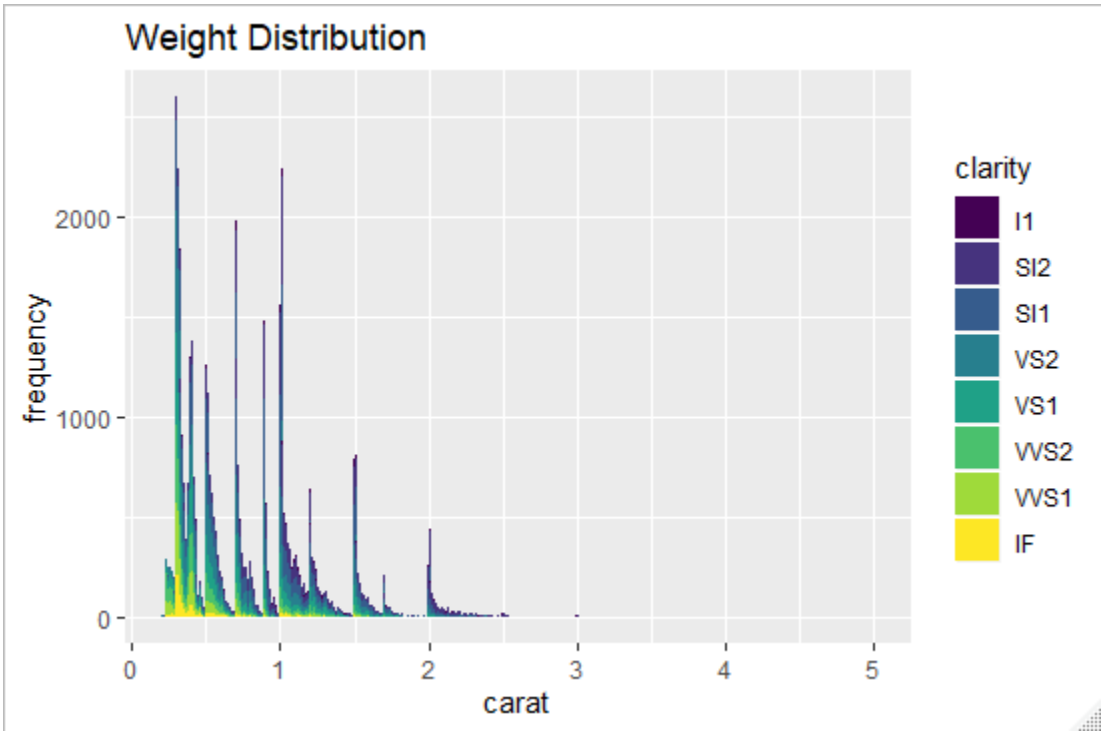


The weight is strongly skewed to the right.
We change the bin by bin parameter in geom_histogram()
geom_histogram(color="red", bins = 50)

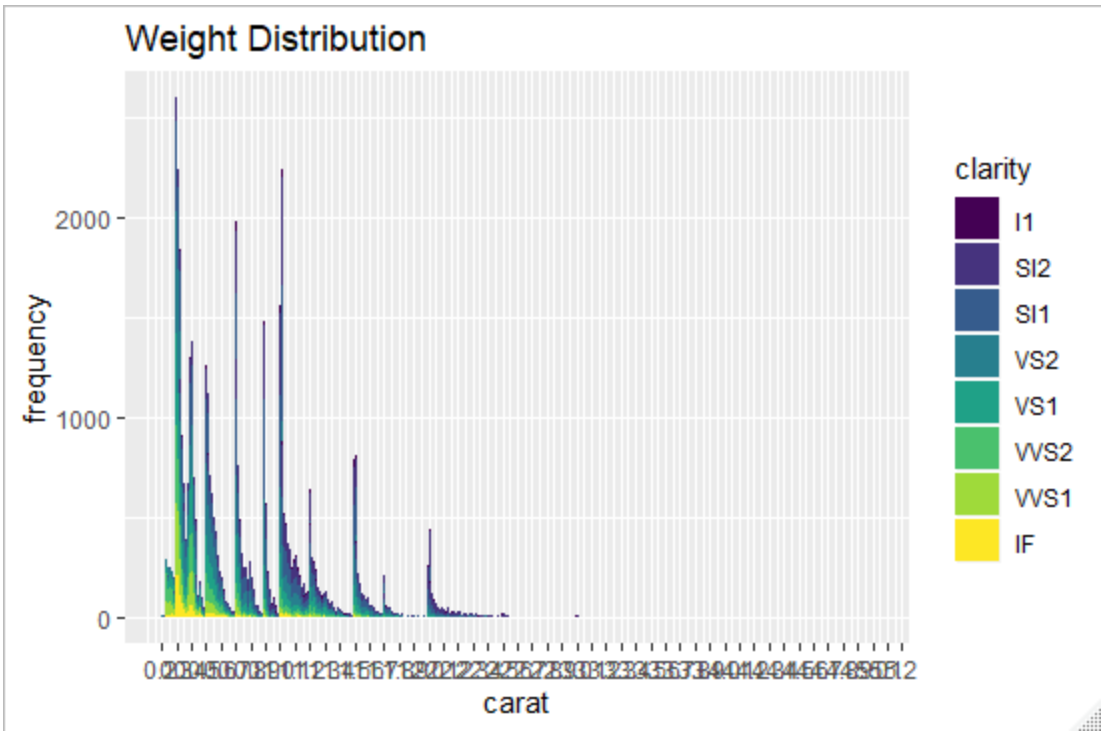


Another way could be as follows;

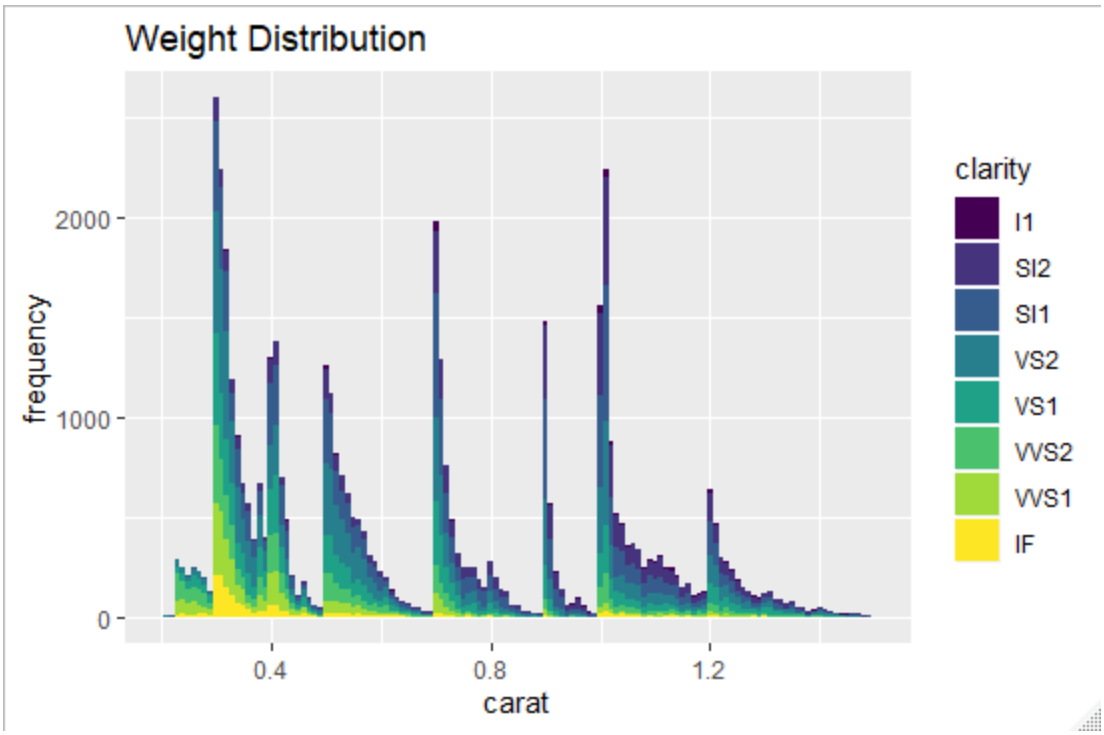
```
p <- ggplot(diamonds,aes(x=carat, fill=clarity)) +
  geom_histogram(center=0.01,binwidth=0.01)
p+ ggtitle("Weight Distribution")+xlab("carat")+ylab("frequency")
```




```
#showing the peaks
p<- p + scale_x_continuous(breaks=caratbreaks)
p+ggtitle("Weight Distribution")+xlab("carat")+ylab("frequency")
```

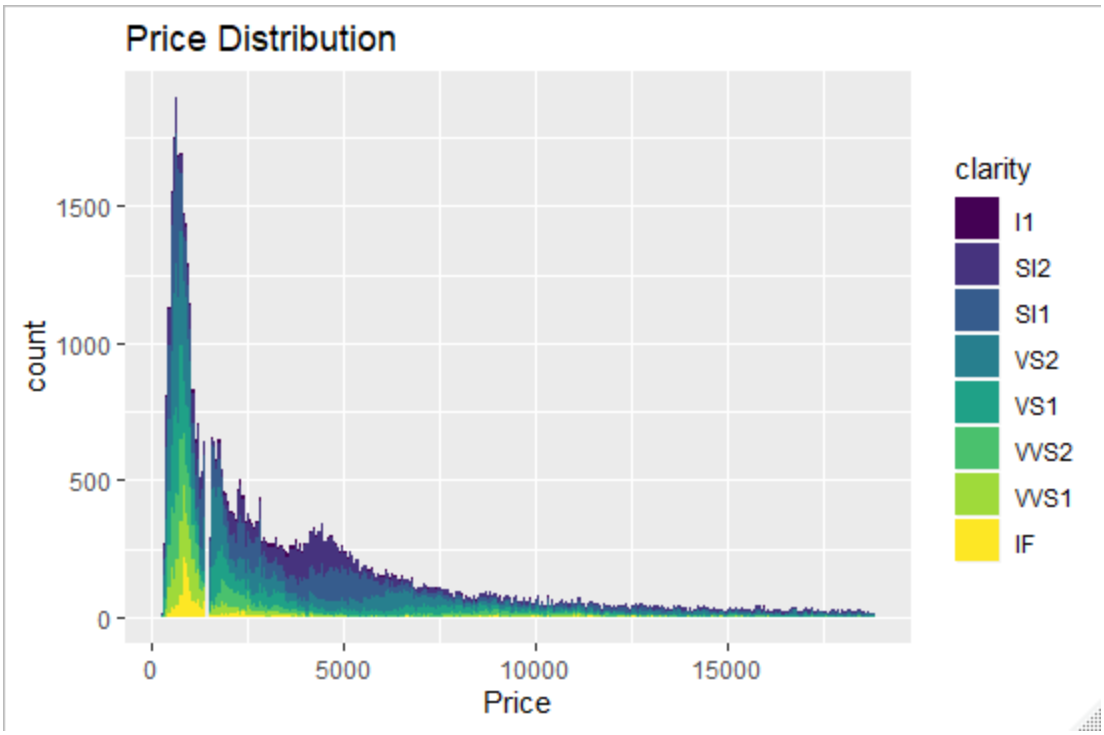



```
#limiting the scate to 0.2,1.5
p <- ggplot(diamonds,aes(x=carat, fill=clarity)) +
  geom_histogram(center=0.01,binwidth=0.01)  +
  scale_x_continuous(limits = c( 0.2,1.5))
p+ggtitle("Weight Distribution")+xlab("carat")+ylab("frequency")
```

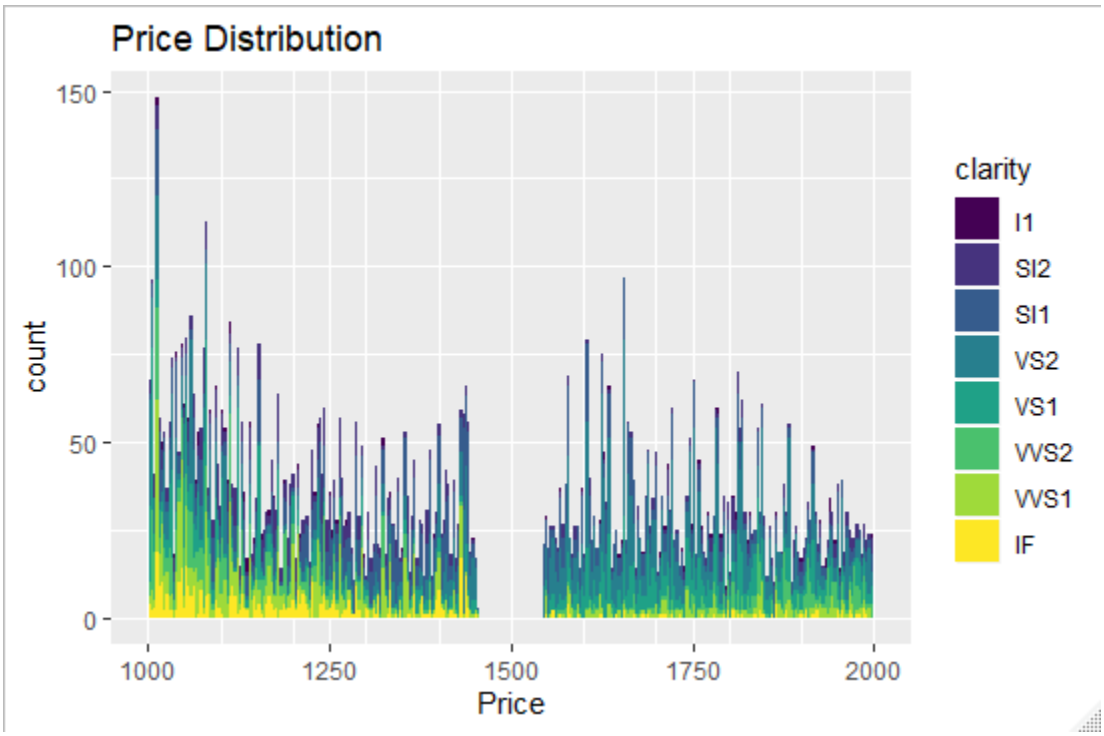


Checking the distribution of prices

```
#price distribution
p <- ggplot(diamonds,aes(x=price, fill=clarity)) + geom_histogram(bins = 300)+
  ggtitle("Price Distribution")+xlab("carat")+ylab("count")
p
```

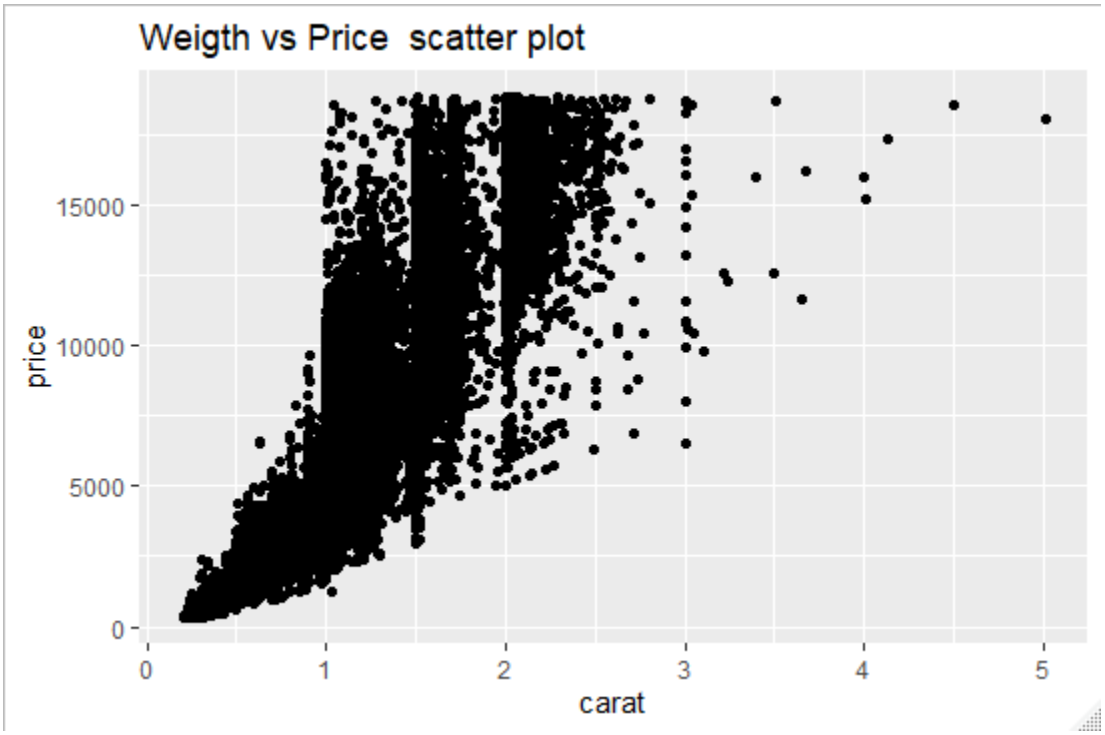


#examining the gap at 1500
p + scale_x_continuous(minor_breaks=seq(500,15000, by=100), limits=c(1000,2000))

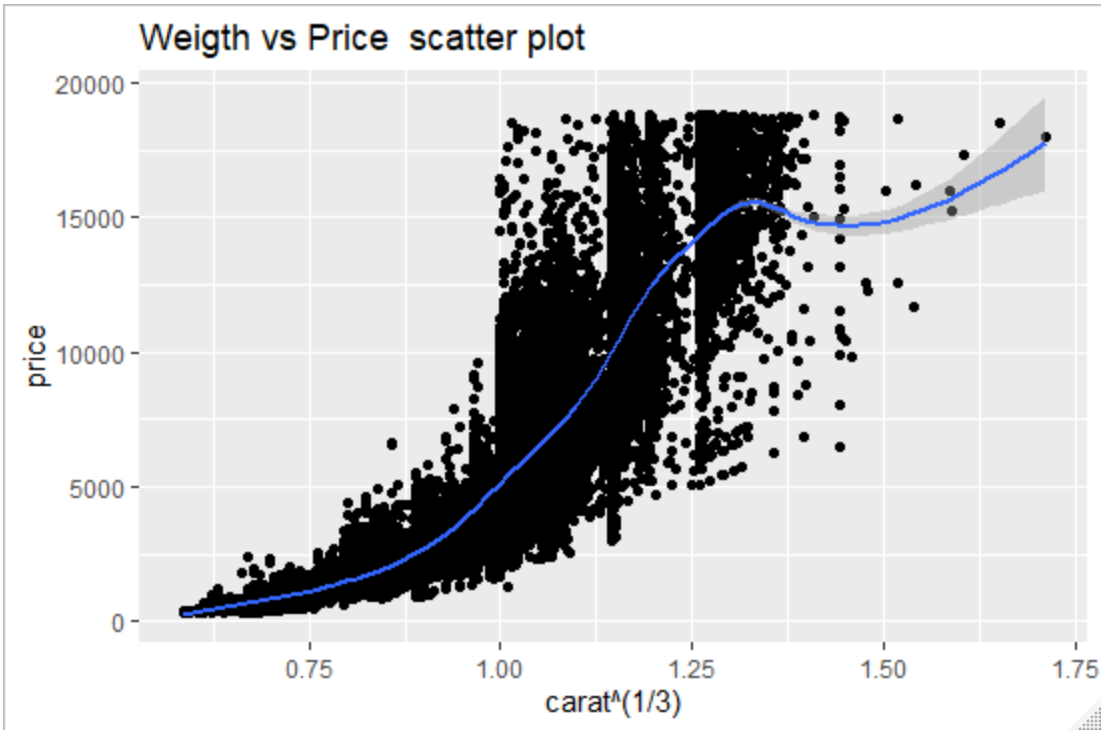


Q3. How is weight (in carats) related to price?

```
#how weight relate to price
p <- ggplot(diamonds,aes(x=carat, y=price))+ geom_point()+
ggtitle("Weigth vs Price  scatter plot")
p
```



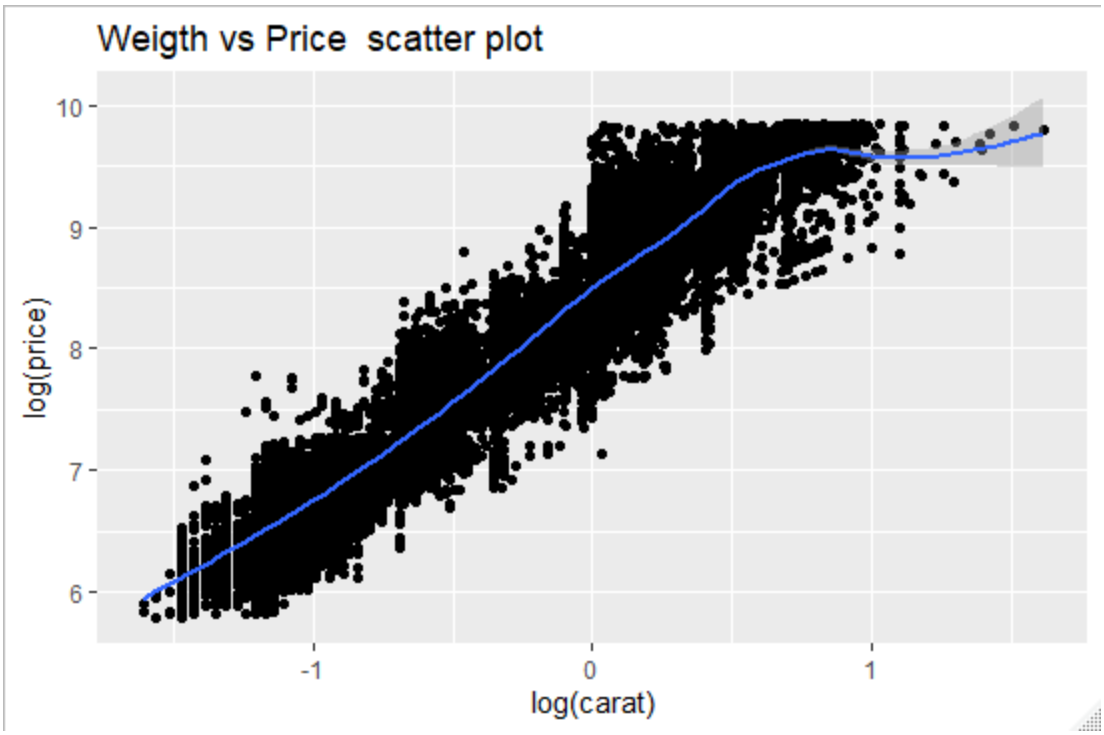

```
#using the function aes(x=carat^(1/3), y=price)
p <- ggplot(diamonds,aes(x=carat^(1/3), y=price))+ geom_point()+
  ggtitle("Weigth vs Price  scatter plot")
p + geom_smooth()
```



The point lies more on a staright line than a curve.

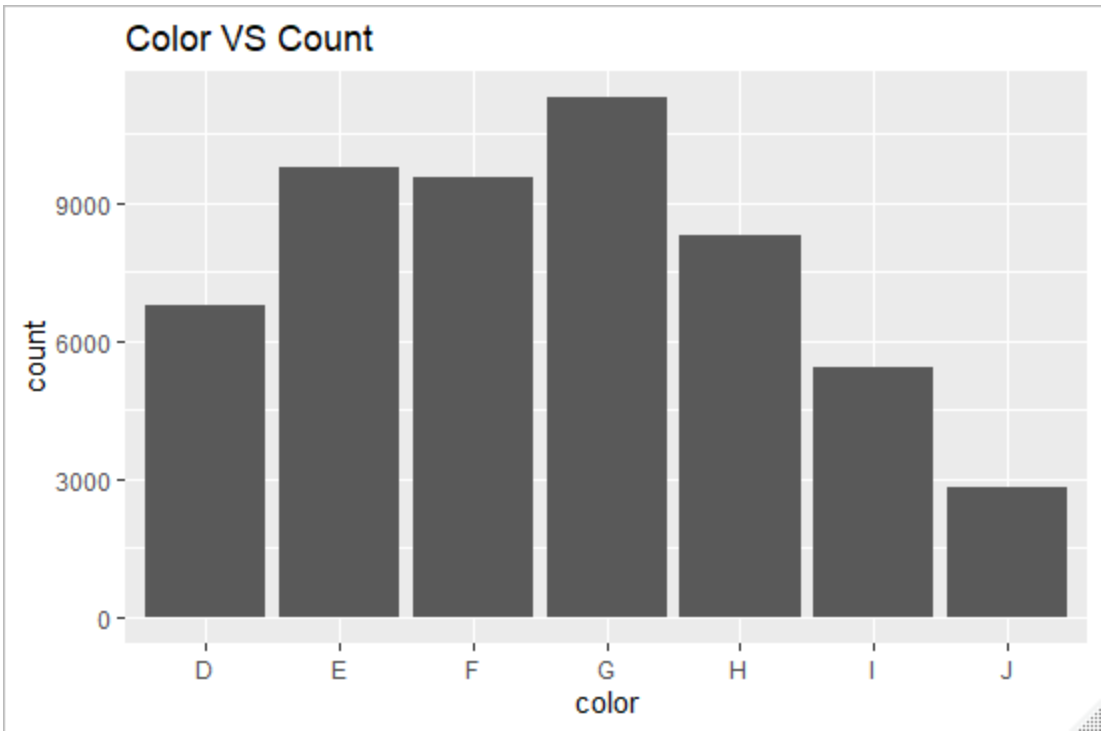
#repeating wit log scale

```
p <- ggplot(diamonds,aes(x=log(carat), y=log(price)))+ geom_point()+
  ggtitle("Weigth vs Price  scatter plot")
p + geom_smooth()
```

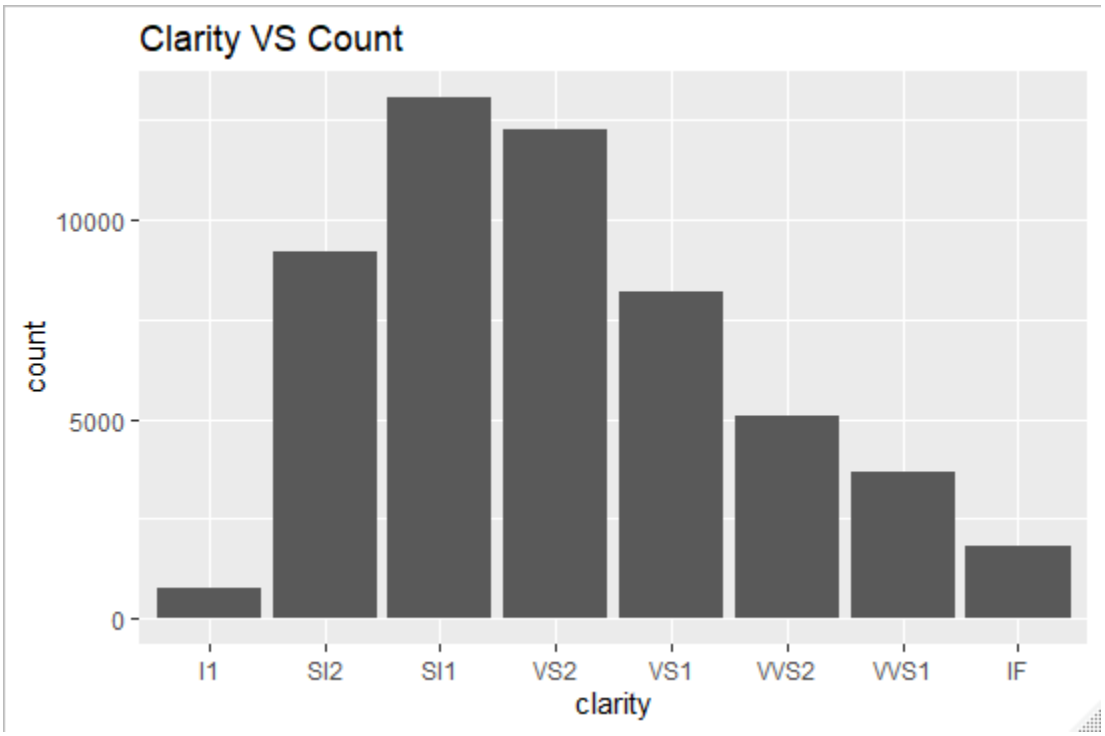


Q4. What are the distributions of clarity, colour, and cut

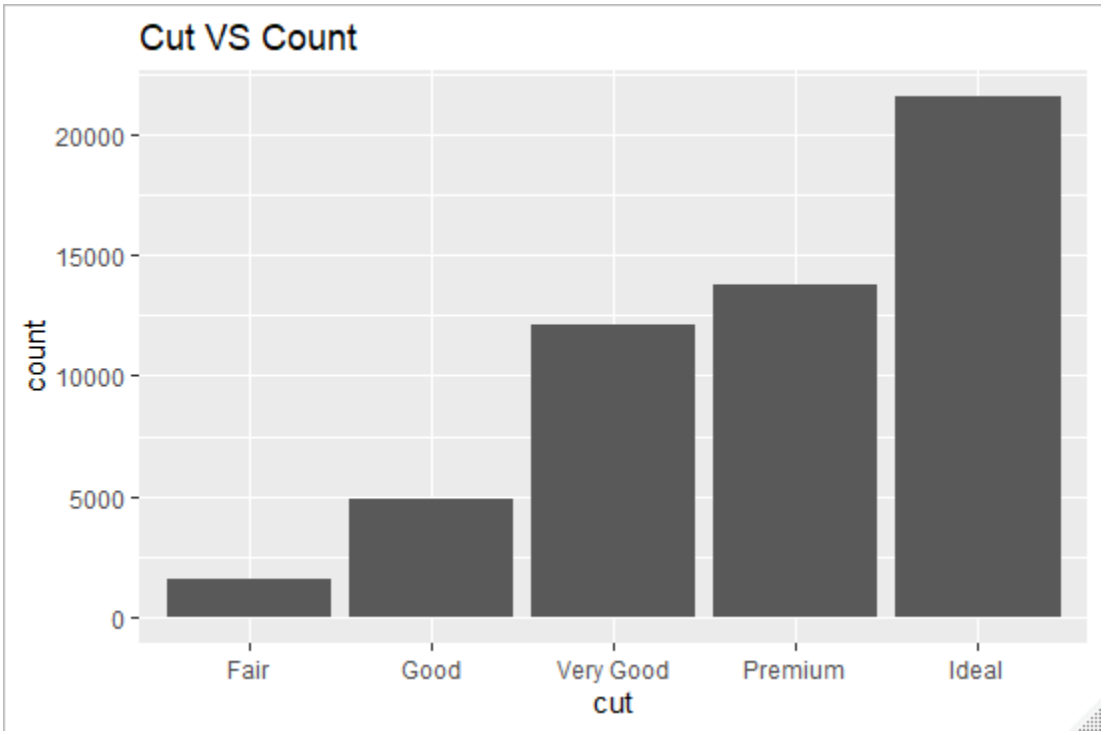
#color
p<- ggplot(diamonds, aes(x=color)) + geom_bar()
p+ ggtitle("Color VS Count")



#clarity
p<- ggplot(diamonds, aes(x=clarity)) + geom_bar()
p+ ggtitle("Clarity VS Count")

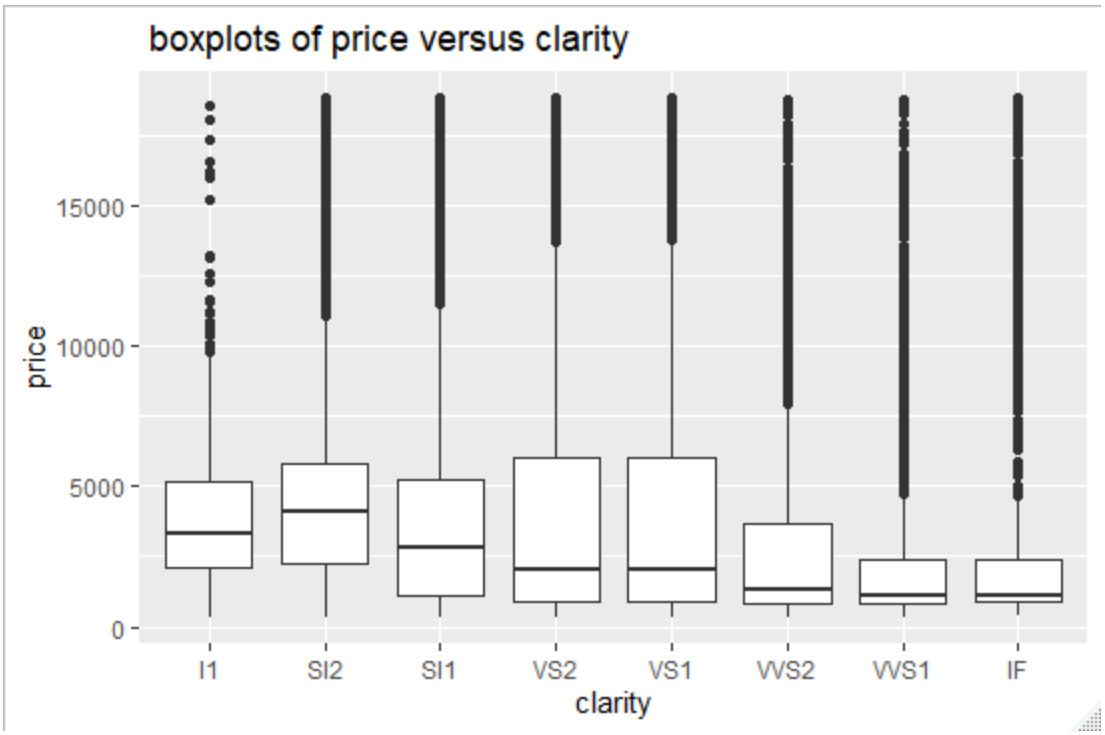


#cut
p<- ggplot(diamonds, aes(x=cut)) + geom_bar()
p+ ggtitle("Cut VS Count")



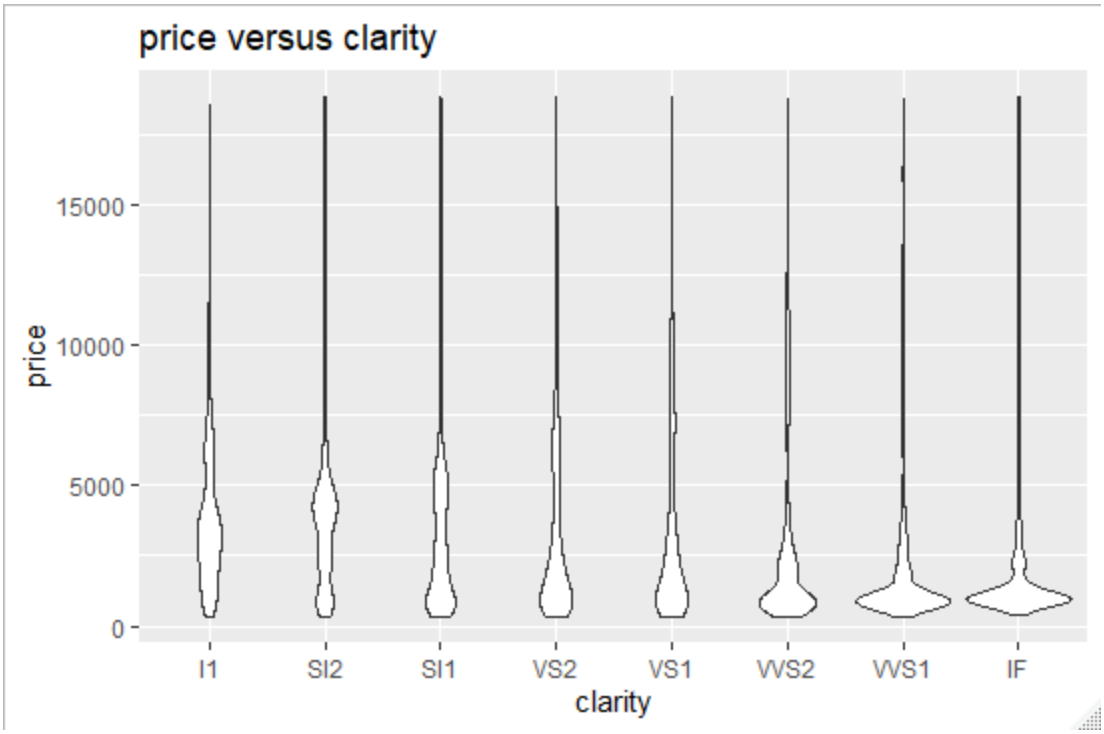
Q5. How do clarity, colour, and cut affect price?

#boxplots of price versus clarity
p<-ggplot(diamonds, aes(x=clarity, y=price)) + geom_boxplot()
p+ggtitle(" boxplots of price versus clarity")



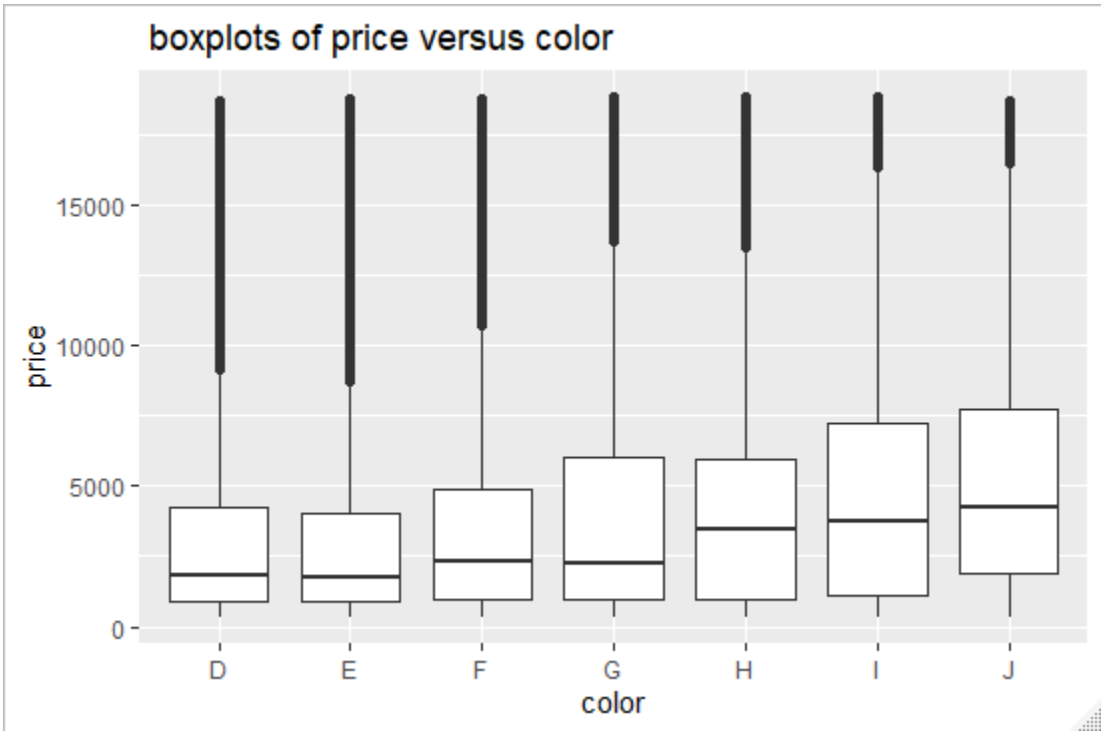
The graph is useful since it demonstrate the distribution of price vs clarity. The price median lies between 0 and 5000 and it is different for each clarity. S12 has the highest and VVs1 and IF being the lowest. This follows the description given on the site about clarity and it is what was expected.

```
p<-ggplot(diamonds, aes(x=clarity, y=price)) + geom_violin()
p+ggtitle("price versus clarity")
```

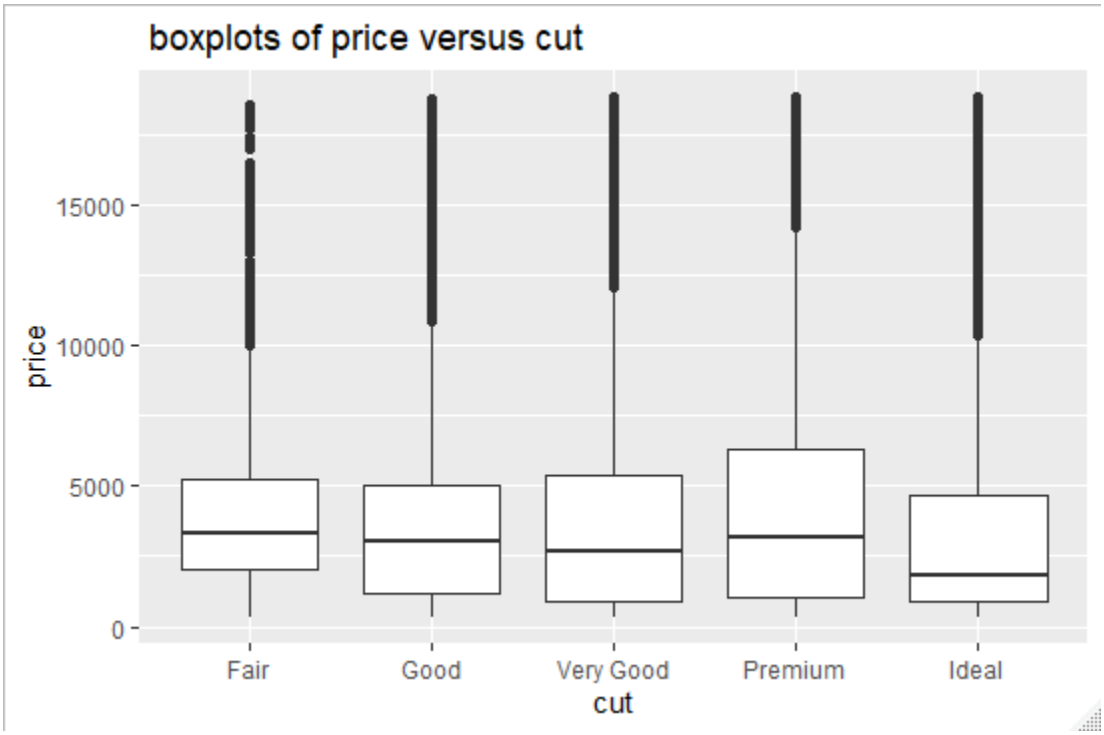


The box plot is much easier to read .

#boxplot graphs to show the distributions prices for different levels of colour
p<-ggplot(diamonds, aes(x=color, y=price)) + geom_boxplot()
p+ggtitle(" boxplots of price versus color")



#boxplot graphs to show the distributions prices for different levels of cut
p<-ggplot(diamonds, aes(x=cut, y=price)) + geom_boxplot()
p+ggtitle(" boxplots of price versus cut")



Ideal has the lowest median yet it is the most expensive. Ideal is a rare type of diamond and this can explain the fact as to where it has the lowest median.

5.2 Does the distribution of clarity, colour, and cut change with carat-weight?

```
> quantile(carat)
  0%  25%  50%  75% 100%
0.20 0.40 0.70 1.04 5.01
```

#add caratfactor to the data

diamonds$caratfactor <- caratfactor

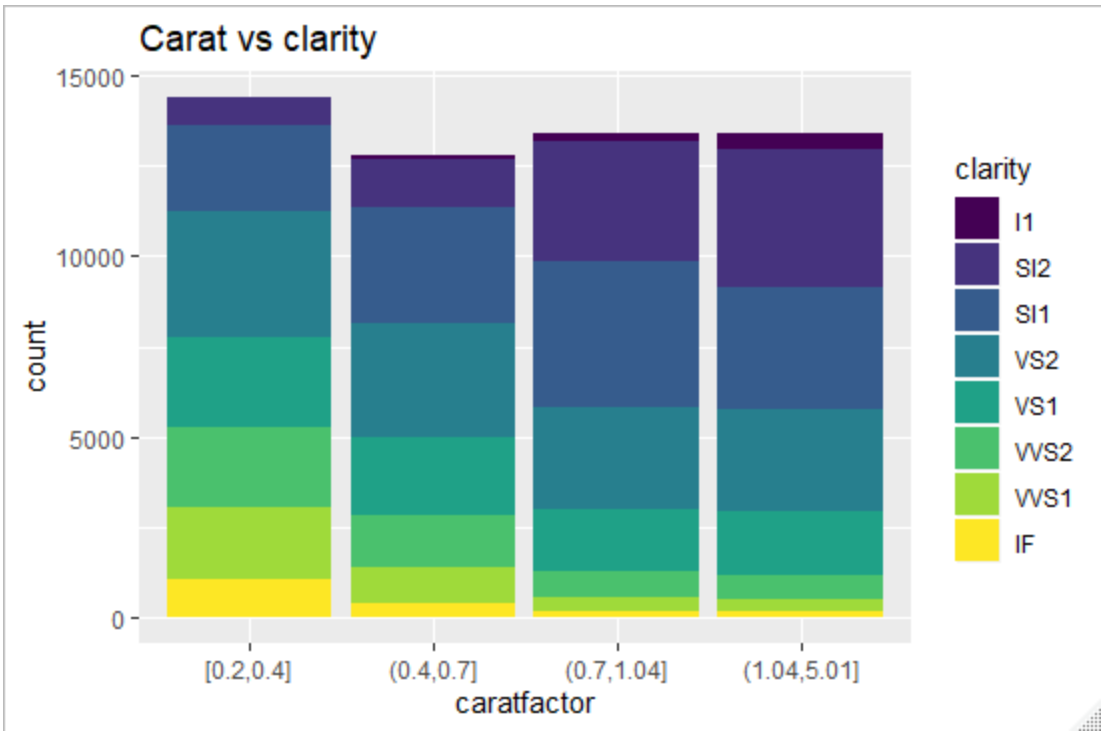
#check the data stucture

str(diamonds)

Classes ‘tbl_df’, ‘tbl’ and 'data.frame': 53940 obs. of 11 variables:

```
$ carat      : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
$ cut        : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
$ color      : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
$ clarity    : Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
$ depth      : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
$ table      : num  55 61 65 58 58 57 57 55 61 61 ...
$ price      : int  326 326 327 334 335 336 336 337 337 338 ...
$ x          : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
$ y          : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
$ z          : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
$ caratfactor: Factor w/ 4 levels "[0.2,0.4]","(0.4,0.7]",..: 1 1 1 1 1 1 1 1 1 1 ...
```

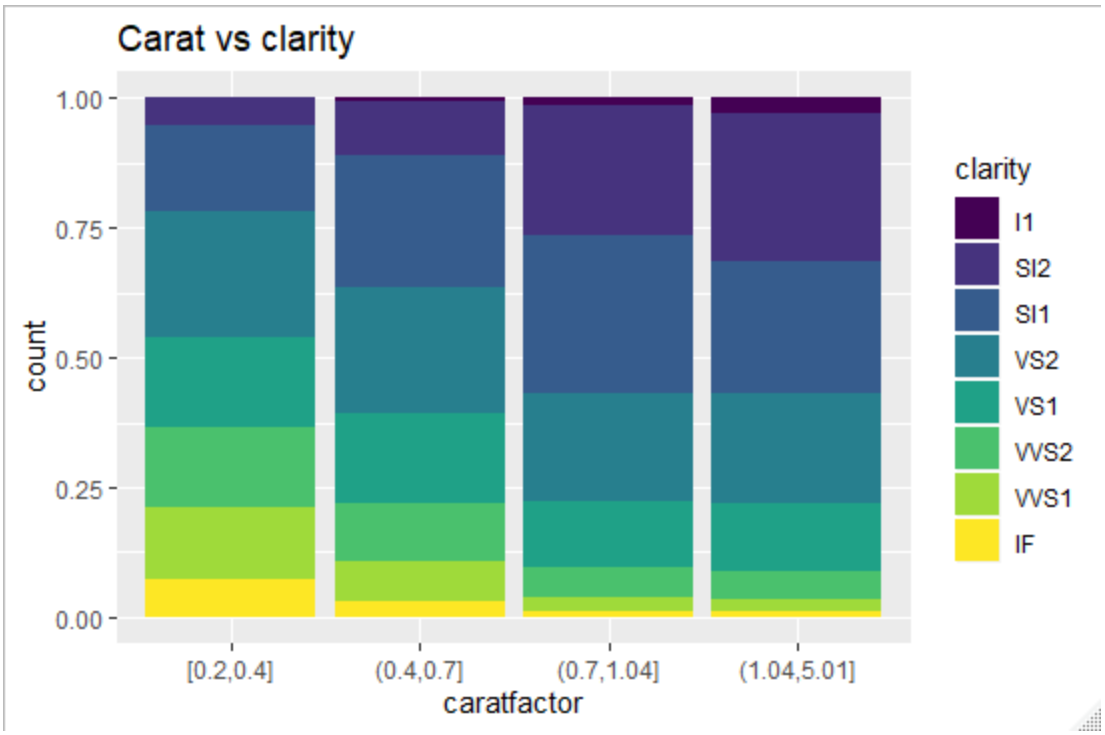
```
#carat vs clarity
p<- ggplot(diamonds, aes(x=caratfactor, fill=clarity)) + geom_bar()
p+ggtitle("Carat vs clarity")
```



Making some adjustments.

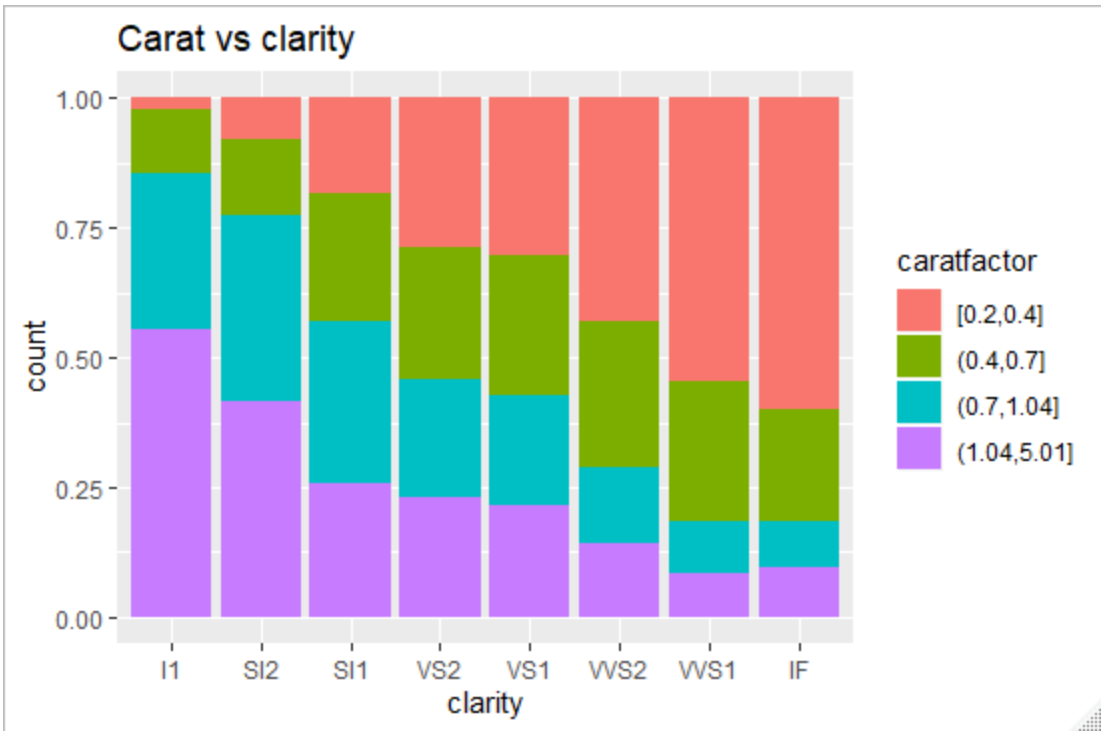
p<-ggplot(diamonds, aes(x=caratfactor, fill=clarity)) + geom_bar(position='fill')

p+ggtitle("Carat vs clarity")




```
p<-ggplot(diamonds, aes(x=clarity, fill=caratfactor)) + geom_bar(position="fill")

p+ggtitle("Carat vs clarity")
```



#find the number of diamond in eac data frame

nrow(halfcaratdiamonds)

4536

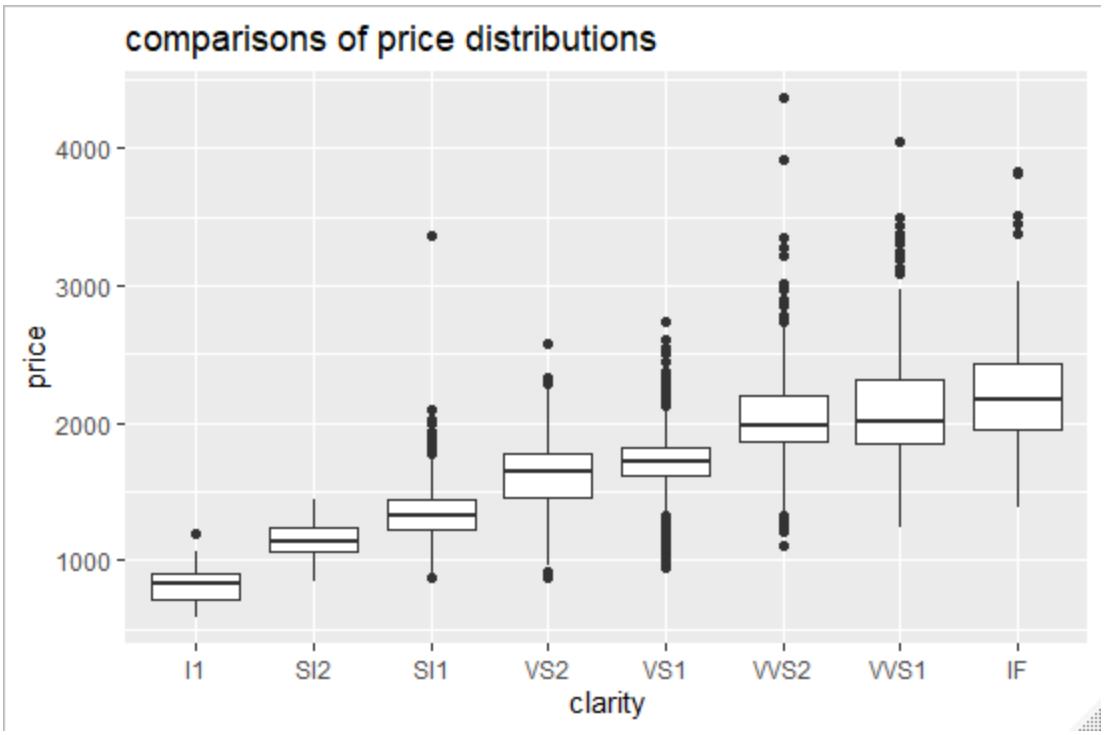
nrow(onecaratdiamonds)

9260

#boxplot comparisons of price distributions

p<- ggplot(halfcaratdiamonds,aes(x=clarity,y=price)) + geom_boxplot()

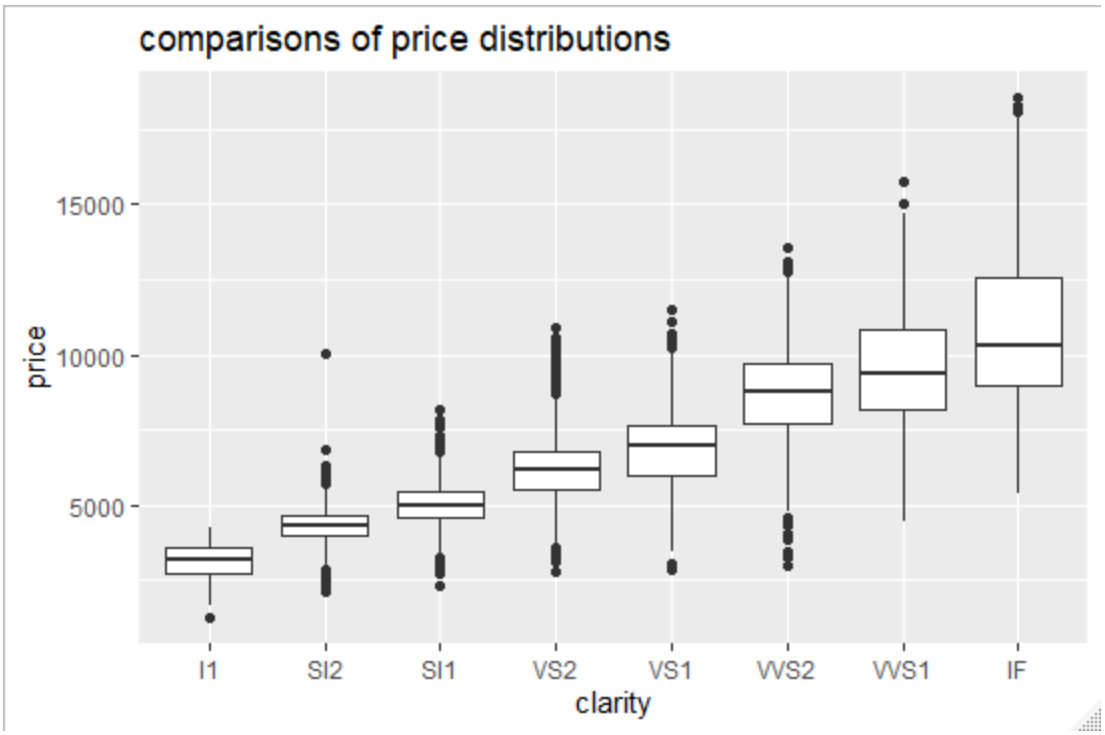
p+ggtitle("comparisons of price distributions ")



#boxplot comparisons of price distributions (onecaratdiamonds)

p<- ggplot(onecaratdiamonds,aes(x=clarity,y=price)) + geom_boxplot()

p+ggtitle("comparisons of price distributions ")



Do cut, clarity, and colour affect price now? Which seems to have the most powerful effect? Label your graphs, add them to your worksheet.

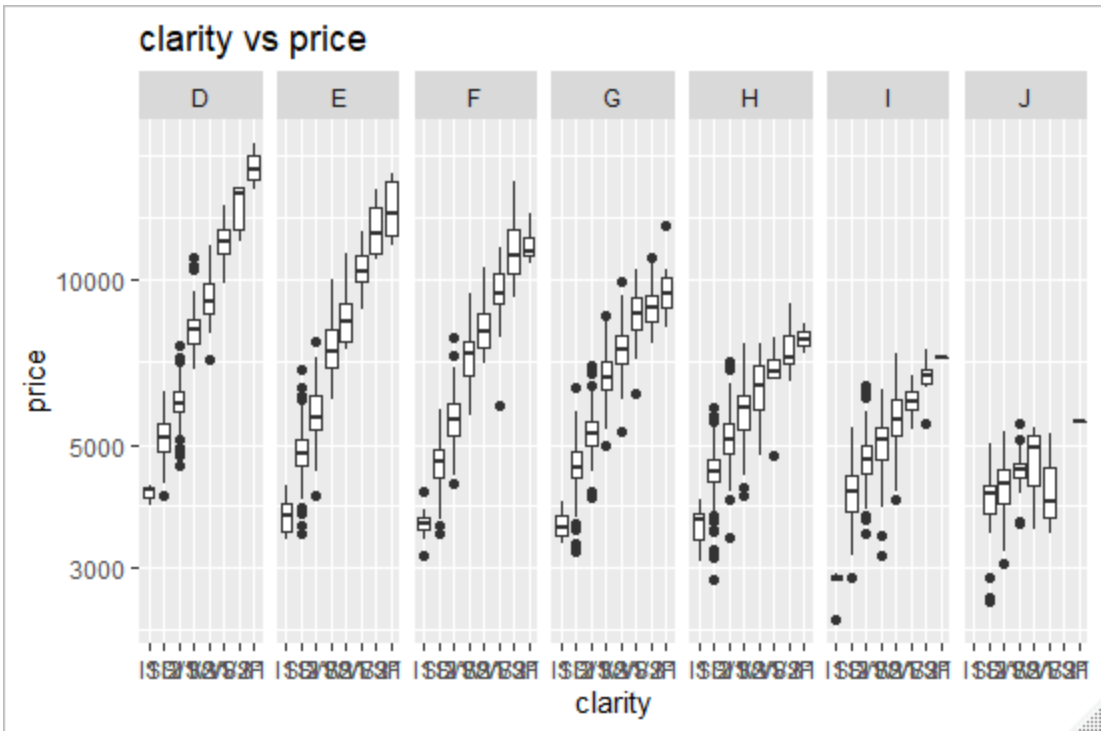
It is evident that cut, clarity and color affect the price. Clarity has a bigger effect than color and cut.

Q6. Is there any interaction between clarity, colour and cut in determining the price?

#facetting

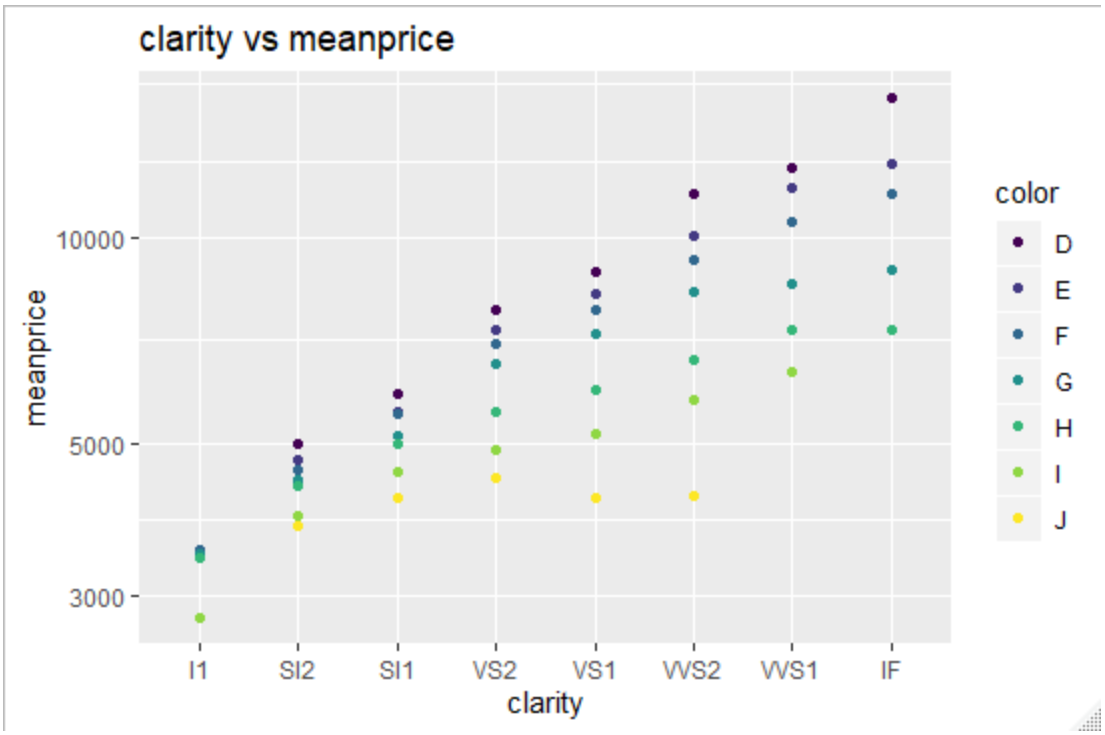
?Facet

```
p<-ggplot(subset(diamonds, carat >0.99 & carat < 1.2 & cut =="Ideal"), aes(x=clarity, y=price)) +
  geom_boxplot() + facet_grid(. ~ color) +scale_y_log10( )
p+ggtitle("clarity vs price ")
```

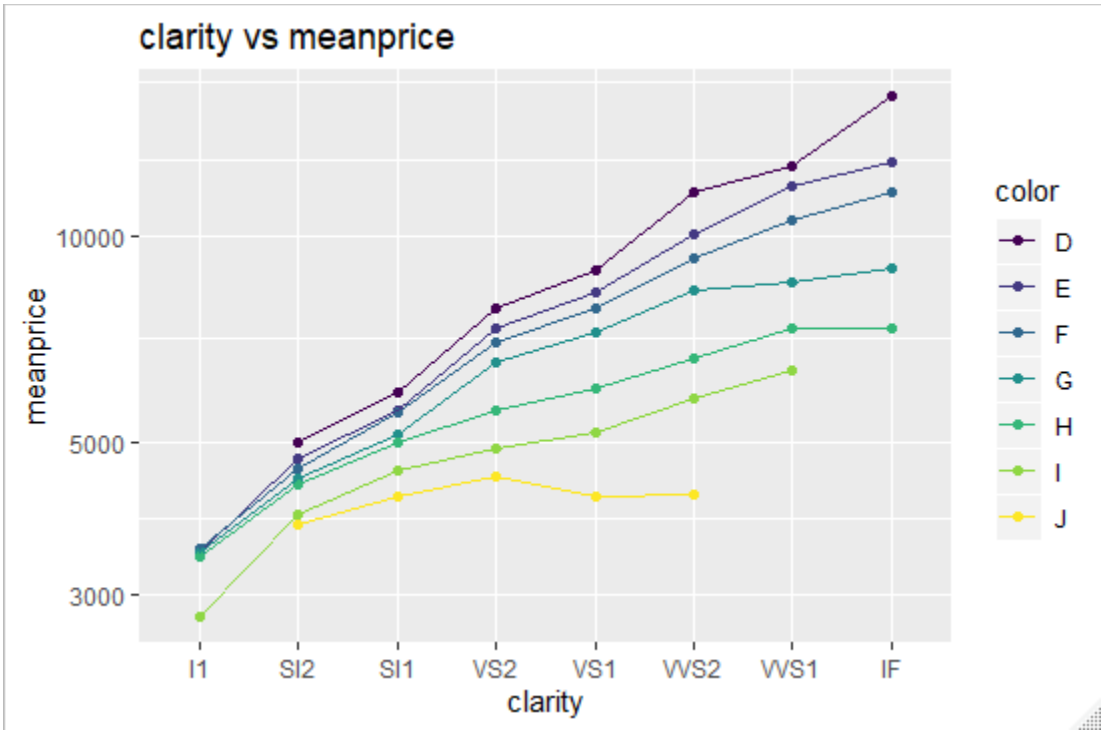


p<-ggplot(d1cmeans, aes(x=clarity, y=meanprice, color=color)) + geom_point() + scale_y_log10()

p+ggtitle("clarity vs meanprice ")




```
p<-ggplot(d1cmeans, aes(x=clarity, y=meanprice, color=color, group=color)) + geom_point() +
  scale_y_log10() + geom_line()
p+ggtitle("clarity vs meanprice ")
```



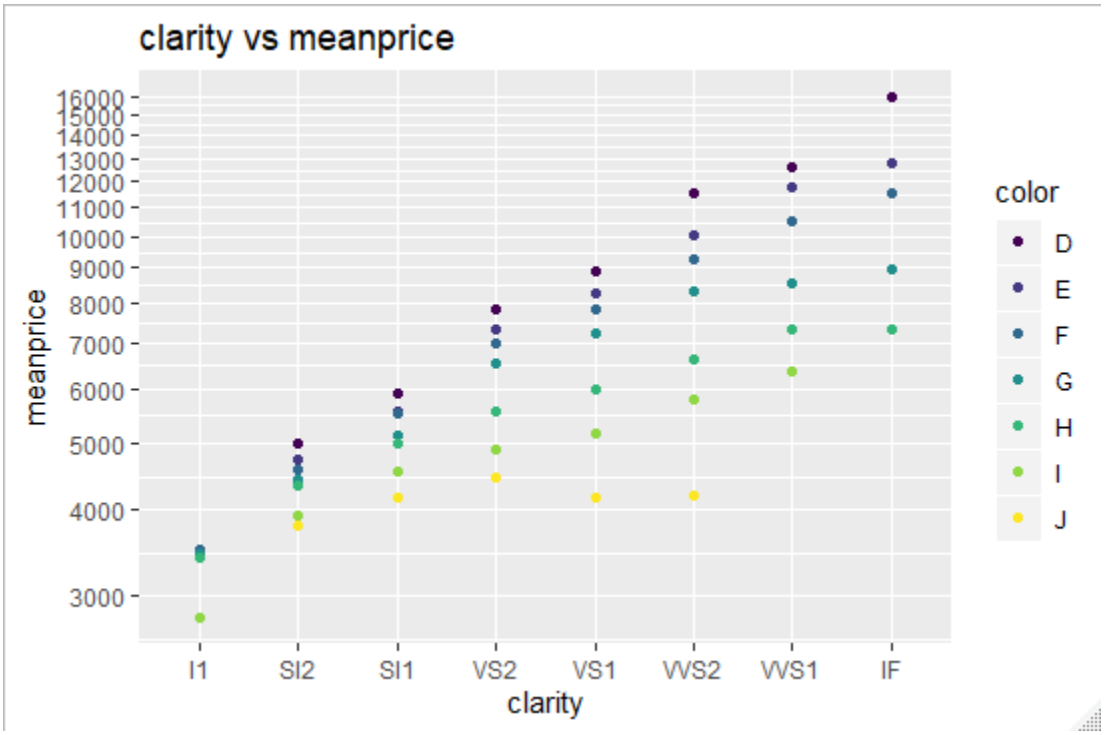
#Improving our graph using scale_y_log10()

pricebreaks <- seq(3000,16000,by=1000)

p<-ggplot(d1cmeans, aes(x=clarity, y=meanprice, color=color, group=color)) + geom_point() +
  scale_y_log10(breaks=pricebreaks)

p+ggtitle("clarity vs meanprice ")

p+ggtitle("Weight Distribution")+xlab("carat")+ylab("frequency")

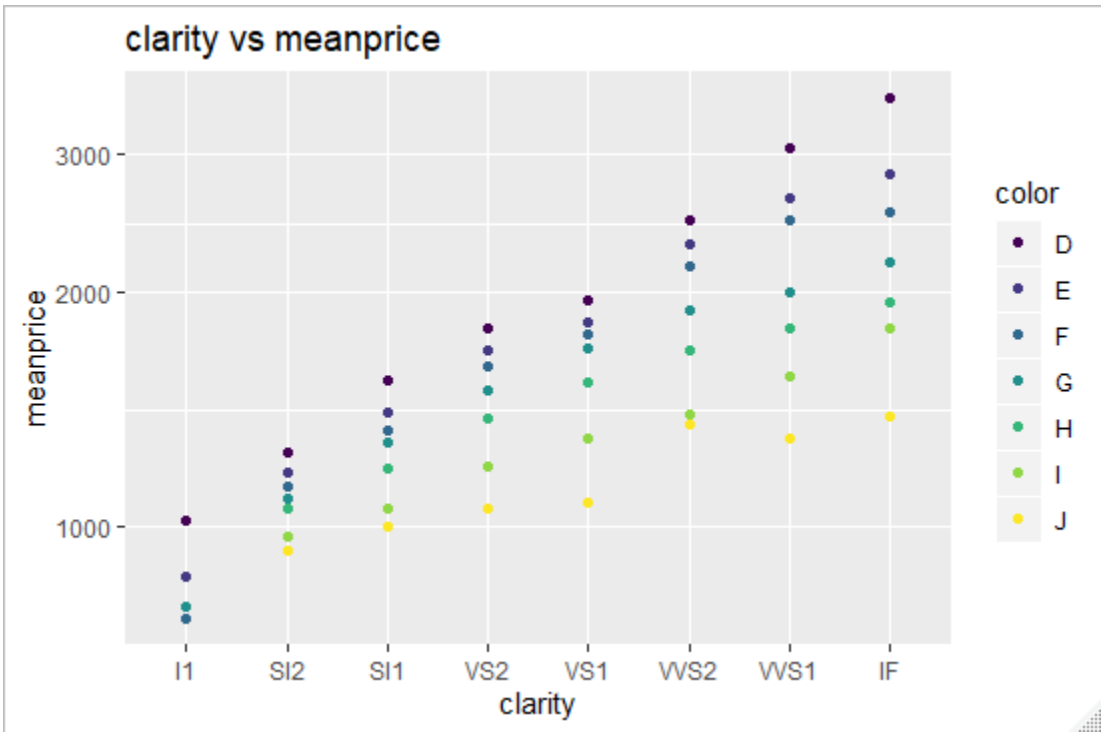


half-carat diamonds.

d1cmeans2 <- ddply(halfcaratdiamonds, .(clarity, color), summarize, meanprice=mean(price))

p<-ggplot(d1cmeans2, aes(x=clarity, y=meanprice, color=color, group=color)) + geom_point() +
  scale_y_log10()

p+ggtitle("clarity vs meanprice ")



Q7. Write down two questions that you could answer with these data, and use appropriate visualisations and summary statistics to answer them.

The data heightweight is data set that represents the height and weight of school children.

We can examine the data using the str command

'data.frame': 236 obs. of 5 variables:
$ sex : Factor w/ 2 levels "f","m": 1 1 1 1 1 1 1 1 1 1 ...
$ ageYear : num 11.9 12.9 12.8 13.4 15.9 ...
$ ageMonth: int 143 155 153 161 191 171 185 142 160 140 ...
$ heightIn: num 56.3 62.3 63.3 59 62.5 62.5 59 56.5 62 53.8 ...
$ weightLb: num 85 105 108 92 112 ...

The data is made up 236 observations and 5 variable. The sex variable which is a categorical variable wit two level.

levels(heightweight$sex)

"f" "m"

F is for female and m represent male.

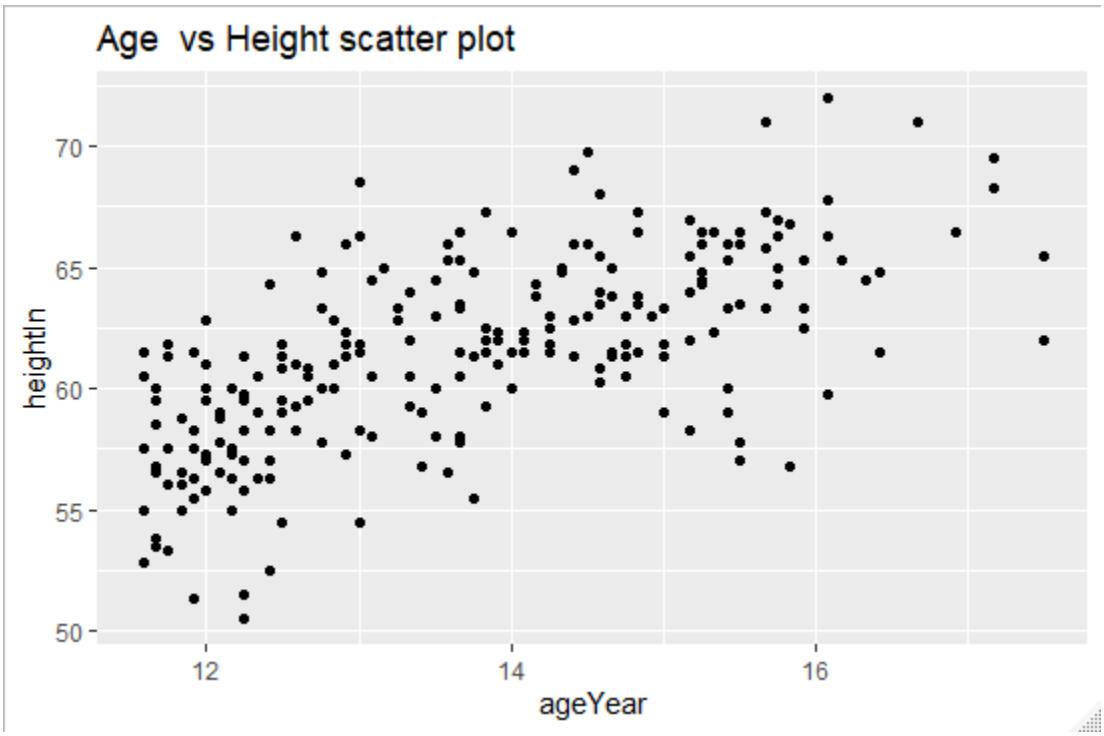
We can use the data to try and answer the following questions.

1. How is age of student related to height

We will use ggplot to show the relationship between the two variables.

#relationship between age and height using scatter plot.

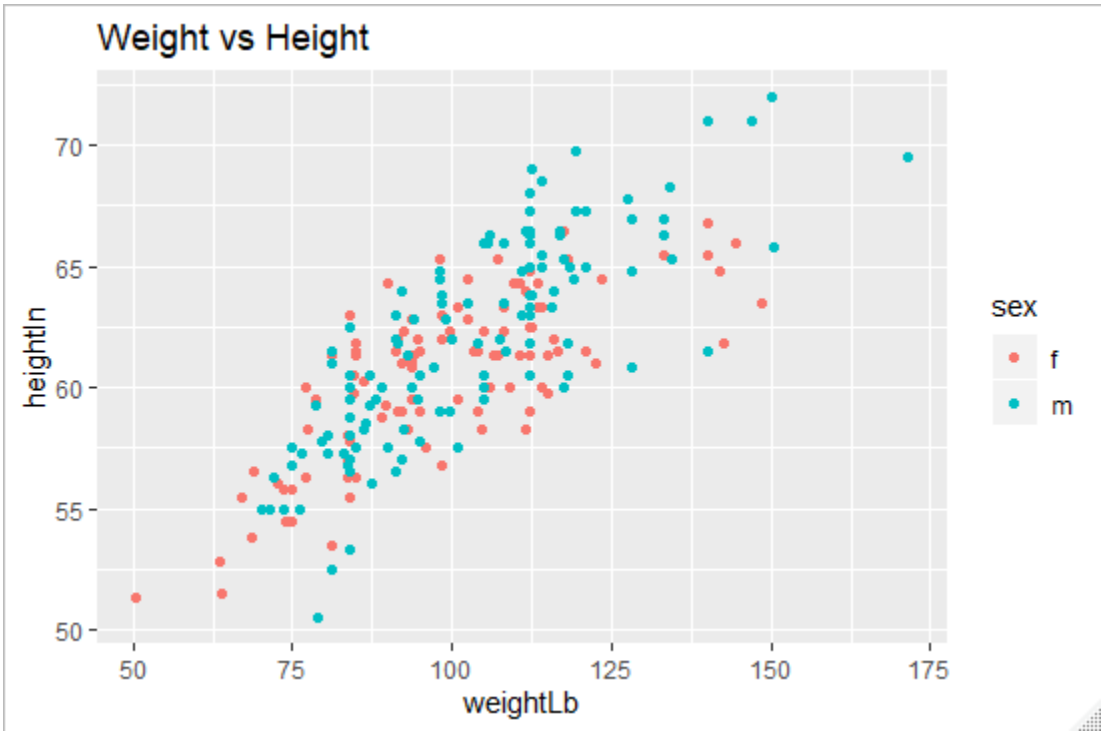
```
p <- ggplot(heightweight,aes(x=ageYear, y=heightIn))+ geom_point()+
  ggtitle("Age  vs Height scatter plot")
p
```



It is a positive relationship, meaning as age increase the height also increase. However it is not perfectly linear.

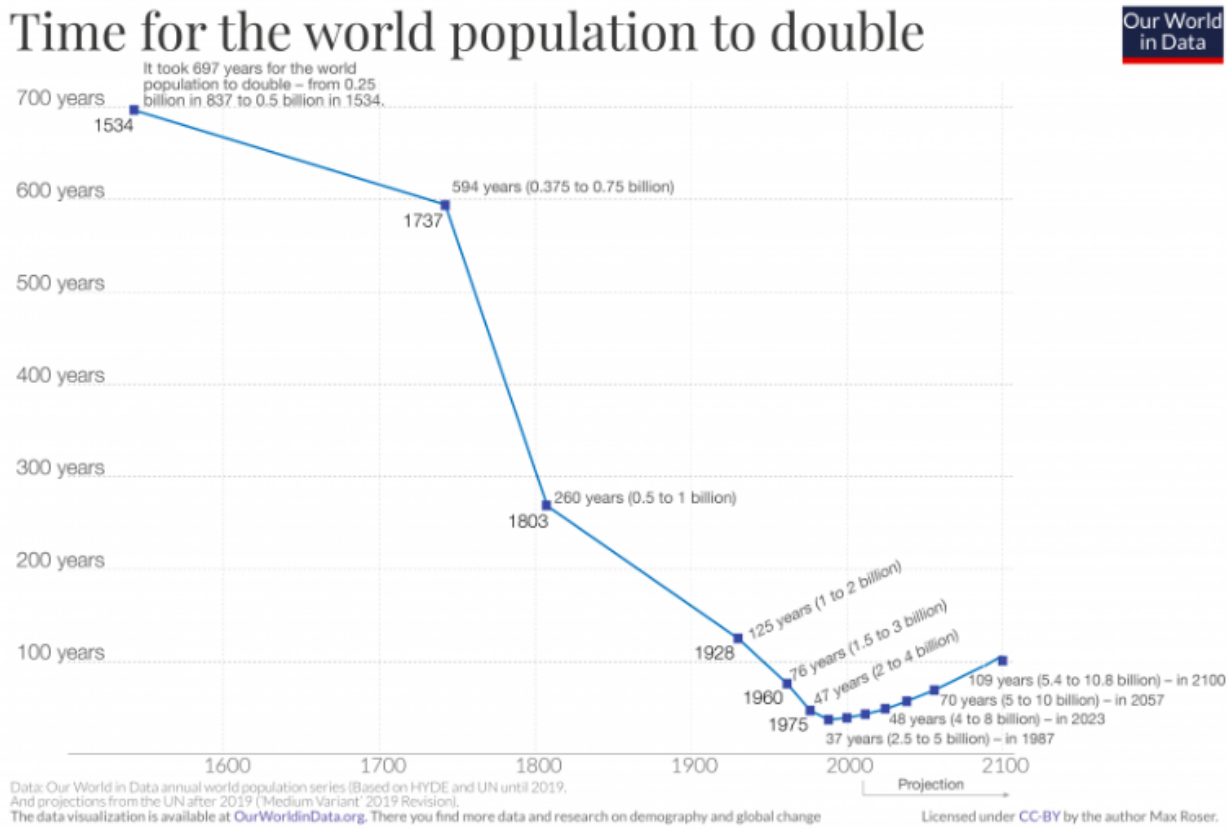
2. We can use the data to check how sex affects weight, and height.

#How sex affect weigth and height
p<-ggplot(heightweight, aes(x=weightLb, y=heightIn, color= sex)) + geom_point()
p+ggtitle("Weight vs Height ")+xlab("weight")+ylab("height")



We can see that boys leads when it comes to height and weight.

Part -2



The visualization shows how strongly the growth rate of the world population changed over time.

We can see from chart that in the past the population grew slowly,it took nearly seven centuries for the population to double from 0.25 billion (in the early 9th century) to 0.5 billion in the middle of the 16th century.As the growth rate slowly climbed, the population doubling time fell.The fastest doubling of the world population happened between 1950 and 1987,a doubling from 2.5 to 5 billion people in just 37 years

We can see that population growth has been slowing, and along with it the doubling time.It can be predicted that by 2100, it will once again have taken approximately 100 years for the population to double to a predicted 10.8 billion.

This visualisation used is the UN projections to show how the doubling time is projected to change until the end of this century.

Is the aim fulfilled?

I feel that data chosen for visualization was precise on point, the main aim for the visualization is to compare how to that taken for the world population has been varying in the past years. We observed a minima in the above chart showing that 1950-1987 took least time to double. But we cannot predict from this chart whether in near future will there be time when the population will increase at a higher rate than it has been. Will it take less time than 37 years?? The question remains.

The graphical display suits best for the given data. Also it is very clear and easy to extract information from the given visualization both visually and statistically. We can say it made a good use of given data and graphics.

Modifications

Despite the fact that it has achieved the aim,I think there could have been more modification to the given visualisation.

- As we discussed in the previous point it would have been much more convenient if the graph could predict if in the near future there will be a point when the time taken to double the population will be less than the time taken in the period of 1950 - 1987 or will this period will remain lowest.

A different way

- A different way in which I would have made in the given visualisation is why only visualize the number of births?I would have taken the number of deaths also in account.it would be really interesting to see how time taken for death to double will align with time taken to double the population.

This way would have been a lot better as it will take into account both the number of deaths and the number of births.
And maybe the reason for the decrease in time for doubling the population is the slow death rate.

From the given graph we cannot say what is the reason, but if we take into account both death and birth then we will definitely reach to the answer to the question “what is the reason for decrease in the time to double the population”