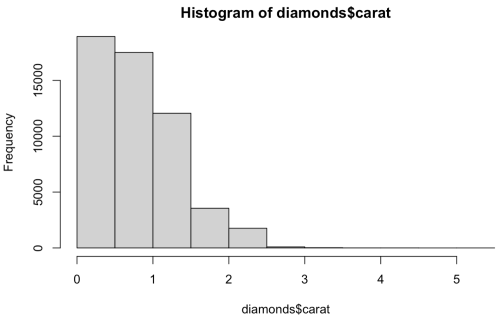
In-Class Assignment 11

For this assignment you will be looking at the diamonds dataset. You can pull this in with the tidyverse library.

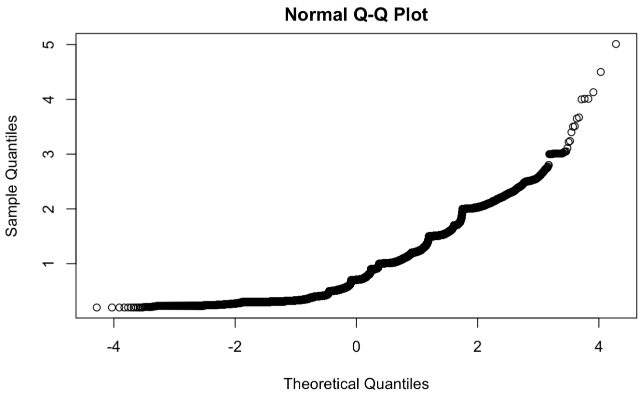
1. Take a look at the fields. Which ones are numerical?
    a. Carat, depth, table, price, x,y, z
2. Which numerical field has the largest relative spread?
    a. price
        i. sd(diamonds$carat)/mean(diamonds$carat)
        ii. sd(diamonds$depth)/mean(diamonds$depth)
        iii. sd(diamonds$table)/mean(diamonds$table)
        iv. sd(diamonds$price)/mean(diamonds$price)
        v. sd(diamonds$x)/mean(diamonds$x)
        vi. sd(diamonds$y)/mean(diamonds$y)
        vii. sd(diamonds$z)/mean(diamonds$z)
            1. [1] 0.5940439
            2. [1] 0.02320057
            3. [1] 0.03888966
            4. [1] 1.014402
            5. [1] 0.1957302
            6. [1] 0.1991681
            7. [1] 0.1994213
3. What are the deciles of carat? Deciles are the quantiles for every tenth from 0.1 to 0.9.
    a. quantile(diamonds$carat, c(0.1, 0.9))
        i. 10% 90%
        ii. 0.31 1.51
4. How does the median value for carat compare to the average value?
    a. mean(diamonds$carat) = 0.7979397
    b. median(diamonds$carat) = 0.7
5. Produce a histogram for carat. How would you describe the distribution?
    a. hist(diamonds$carat)



b.

The distribution is positively skewed (greater on the right and decreasing in frequency from right to left)

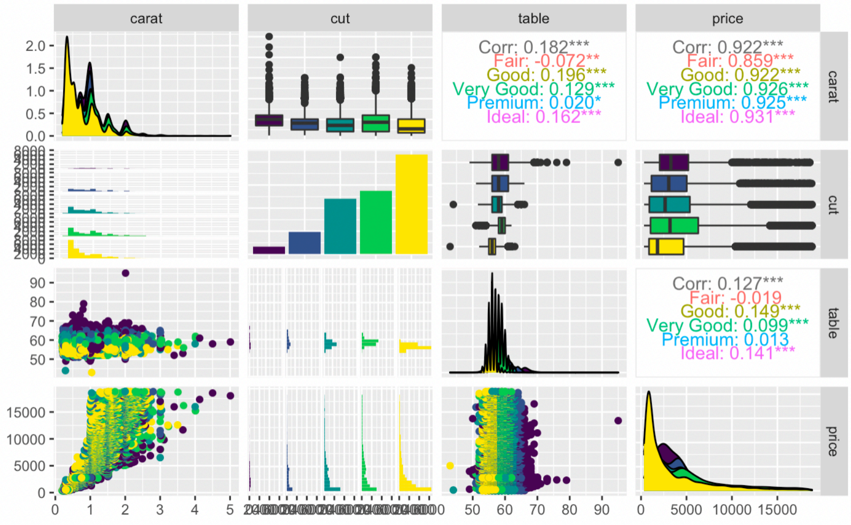
6. Produce a normal probability plot for carat. Does it look like it is normally distributed?



a.hist(diamonds$carat)

No the graph is not normally distributed, the x values start at -4 while the y start at 0. The slope is also exponentially curved which is not the y=x line that we are looking for.

7. Produce the covariance and correlation matrices for the numerical fields. Which field (other than itself) has the highest correlation to price? Note: you can select specific fields in diamonds by stating diamonds[, c(list of field numbers)]. So for example to get the cut and color fields only, you can use diamonds[, c(2, 3)]
   a. cov(diamonds[,c(1,5,6,7,8,9,10)])
   b. cor(diamonds[,c(1,5,6,7,8,9,10)])
8. Produce a pair plot for the fields carat, table, and price, coloring by cut. (Note: this will take a few minutes to produce). Make some comments on what you see.
   a. I see that for the table and carat correlation there is a poor correlation. There is also a poor correlation between price and table. There is a very good correlation between the price and the carat of the diamonds, which is what we found earlier with our coding. There are a few outliers but most of each dataset is similar which we can see clearly from the box plots.

carat
cut
table
price
Corr: 0.182***
Fair: -0.072**
Good: 0.196***
Very Good: 0.129***
Premium: 0.020*
Ideal: 0.162***
Corr: 0.922***
Fair: 0.859***
Good: 0.922***
Very Good: 0.926***
Premium: 0.925***
Ideal: 0.931***
Corr: 0.127***
Fair: -0.019
Good: 0.149***
Very Good: 0.099***
Premium: 0.013
Ideal: 0.141***
2.0
1.5
1.0
0.5
0.0
90
80
70
60
50
15000
10000
5000
0
0
1
2
3
4
5
50
60
70
80
90
0
5000
10000
15000