Assignment 1
CIS 335
Monica Klosin

1. **Dataset downloaded from https://www.kaggle.com/shivam2503/diamonds**
2. **Dataset downloaded, coding in R.**

3. **Identify the different types of attributes**

   The response variable will be the price of the diamond.
   The explanatory variables and their attributes are listed in the table below.

Variable Name	Variable Type
Carat	Quantitative
Cut	Categorical
Color	Categorical
Clarity	Categorical
Depth	Quantitative
Table	Quantitative
x (length of diamond)	Quantitative
y (width of diamond)	Quantitative
z (depth of diamond)	Quantitative

4. **Identify the mean, median, mode of the numeric attributes. Also identify which attributes are positively or negatively skewed.**

To find the mean and median in R:

```
mean(diamond$variable)
median(diamond$ variable)
```

To find the mode I used a function to calculate it:

```
Mode <- function(x) {
    ux <- unique(x)
    ux[which.max(tabulate(match(x, ux)))]
}
Mode(diamond$ variable)
```

To figure out if an attribute is positively or negatively skewed, in R:

```
skewness(diamond$variable)
```

If the value returned is positive, the attribute has a positive skew. If the value is negative, the attribute has a negative skew.

Carat:

mean: 0.7979

median: 0.7
mode: 0.3
skew: Positive (1.11)

Depth:
- mean: 61.74
- median: 61.8
- mode: 62
- skew: Negative (-0.08)

Table:
- mean: 57.45
- median: 57
- mode: 56
- skew: Positive (0.79)

x:
- mean: 5.73
- median: 5.7
- mode: 4.37
- skew: Positive (0.37)

y:
- mean: 5.73
- median: 5.71
- mode: 4.34
- skew: Positive (2.43)

z:
- mean: 3.54
- median: 3.53
- mode: 2.7
- skew: Positive (1.52)

5. **Compute the IQR for numerical attributes. Based on the IQR, determine the outliers, and decide if we keep the outliers or not.**

To find the quartiles of the numerical attributes, in R:

```
summary(diamond)
```

To find the IQR value, and the lower and upper bounds of the numerical attributes, in R:

```
IQRvalue = (diamonds$value)
lowervalue = Q1 - (IQRvalue * 1.5)
print(lowery)
highervalue = Q3 + (IQRvalue* 1.5)
```

```
     carat            cut                  color                clarity
 Min.   :0.2000   Length:53940       Length:53940       Length:53940
 1st Qu.:0.4000   Class :character   Class :character   Class :character
 Median :0.7000   Mode  :character   Mode  :character   Mode  :character
 Mean   :0.7979
 3rd Qu.:1.0400
 Max.   :5.0100
     depth           table           price             x
 Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 0.000
 1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710
 Median :61.80   Median :57.00   Median : 2401   Median : 5.700
 Mean   :61.75   Mean   :57.46   Mean   : 3933   Mean   : 5.731
 3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540
 Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.740
       y                z
 Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.710   Median : 3.530
 Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :58.900   Max.   :31.800
```

A data point is an outlier if it more than 1.5*IQR above Q3 or below Q1.

Carat:
For Carat, the IQR is 0.64.
With this calculation, a diamond with a carat value above 2 or below -0.54 (which isn't possible) then that diamond carat value is an outlier.
From this, we see that 2154 out of 53940 diamonds fall into this outlier category. Even though they are considered outliers by the calculations, they should be kept in the data set since these records are not measurement errors or made by any sampling problems – in the world there are just a few diamonds that are heavier than average.

x (length):

For x, the IQR is 1.83.
With this calculation, a diamond with a length below 1.965mm or above 9.285mm is an outlier. From this, we see that 32 out of 53940 diamonds fall into this outlier category. Even though they are considered outliers by the calculations, they should be kept in the data set since these records are not measurement errors or made by any sampling problems – in the world there are just a few diamonds that are bigger than average.

z (depth):
For z, the IQR is 1.13.
With this calculation, a diamond with a depth below 1.215mm or above 5.735mm is an outlier. From this, we see that 49 out of 53940 diamonds fall into this outlier category. Even though they are considered outliers by the calculations, they should be kept in the data set since these records are not measurement errors or made by any sampling problems – in the world there are just a few diamonds that are bigger than average.

6. **Use scatter plots to determine if there's correlation between the numeric attributes**

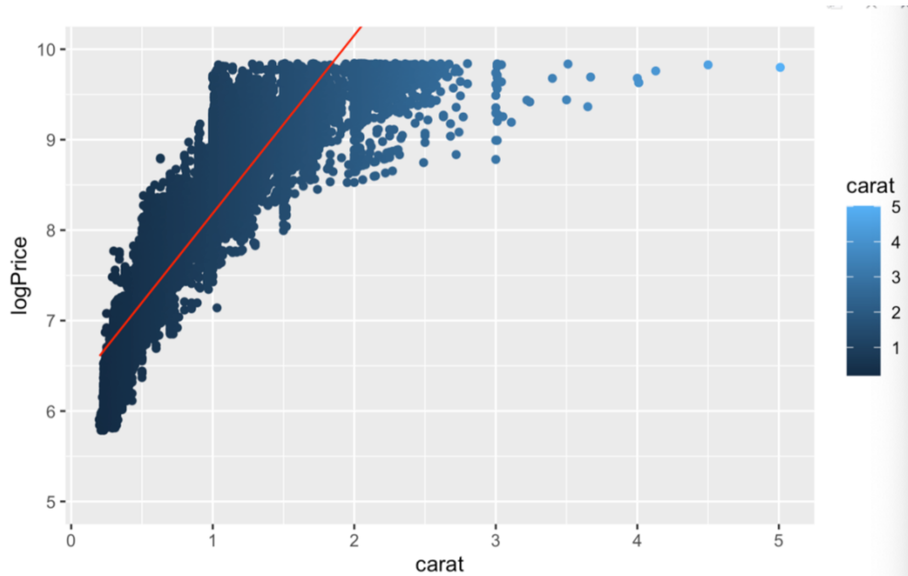
Since the data is very positively skewed right ,it would be difficult to see the data as is in the scatterplot, I took the log of the response variable Price to better display the data in a scatterplot.

In R,

To create the scatterplot:

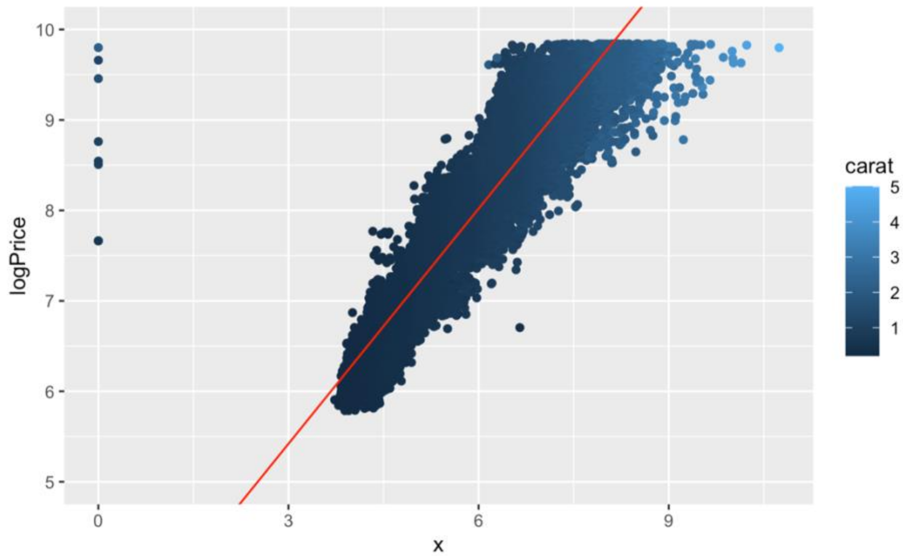
```
diamond %>% ggplot(aes(variable, logPrice)) +
geom_point(aes(color=carat)) +
coord_cartesian(ylim = c(5, 10)) +
        geom_smooth(method='lm', formula= y~x, color="red",
```

Correlation between cut and carat?



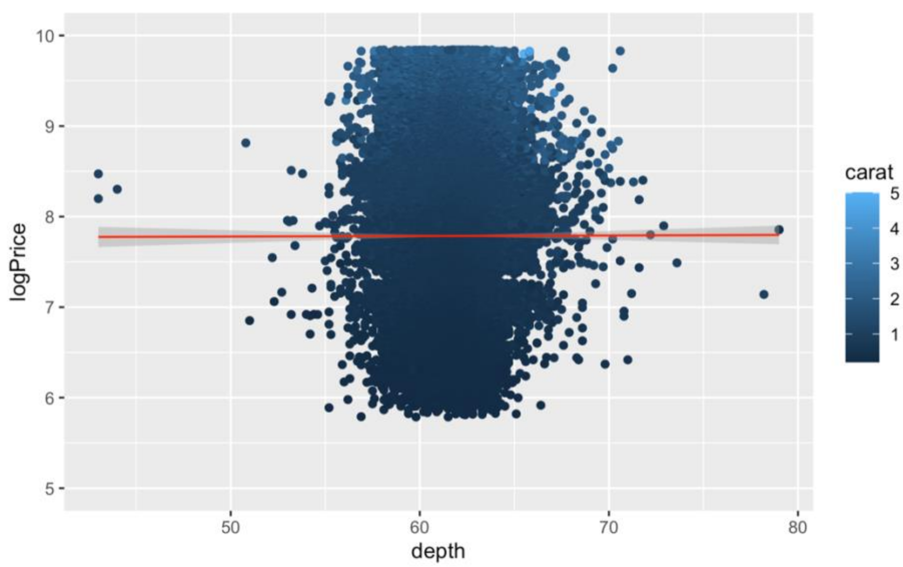
There is a positive correlation between the carat of a diamond and its price.

Correlation between price and x (length of diamond)?



There is a positive correlation between the length of a diamond and its price.

Correlation between price and depth?



There is no correlation between a diamond's total depth percentage and its price.

7. **Draw boxplots for the numeric attributes**

In R:

```
ggplot(diamond, aes(x=carat)) +
   geom_boxplot(fill="slateblue", alpha=0.2, type="count") +
   xlab("Carat") + ggtitle("Boxplot of Carat weight of
```