Visualizing Data Using ggplot2

Last name: Khambaita
First name: Sunil
Student ID: 1000285924
Course section: STA303H1S-L0101

Jan 2nd, 2017

Q1 - Factor variables analysis

Question: (1 point) How many factor variables in this data set? Use R command *str(diamonds)* to find it. For each factor variable, find the one-way frequency table for it. An example of cut variable is given in the solution template.

Answer: We have 3 factor varibles. They are cut (with 5 levels), color (with 7 levels) and clarity (with 8 levels).

```r
# Installing ggplot2 R packages
# install.packages("ggplot2") <--- I ready installed it, hence commented out.

# Loading gg2plot2 and reading in diamonds data
library(ggplot2)
data(diamonds)

# check the type of variables in this data
str(diamonds)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    53940 obs. of  10 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 ...
```

```r
# observations for each level of "cut" variable
table(diamonds$cut) # or summary(diamonds$cut)
```

```
##
##      Fair      Good Very Good   Premium     Ideal
##      1610      4906     12082     13791     21551
```

```
# find the level frequency of "cut" varible
prop.table( table(diamonds$cut) ) # or summary(diamonds$cut)/nrow(diamonds)
```

```
## 
##       Fair       Good  Very Good    Premium      Ideal 
## 0.02984798 0.09095291 0.22398962 0.25567297 0.39953652
```

```
# find the level frequency of "color" varible
prop.table( table(diamonds$color) )
```

```
## 
##          D          E          F          G          H          I 
## 0.12560252 0.18162773 0.17690026 0.20934372 0.15394883 0.10051910 
##          J 
## 0.05205784
```

```
# find the level frequency of "clarity" varible
prop.table( table(diamonds$clarity) )
```

```
## 
##         I1        SI2        SI1        VS2        VS1       VVS2 
## 0.01373749 0.17044865 0.24221357 0.22725250 0.15148313 0.09391917 
##       VVS1         IF 
## 0.06776047 0.03318502
```

Q2 -Produce plot and give comment

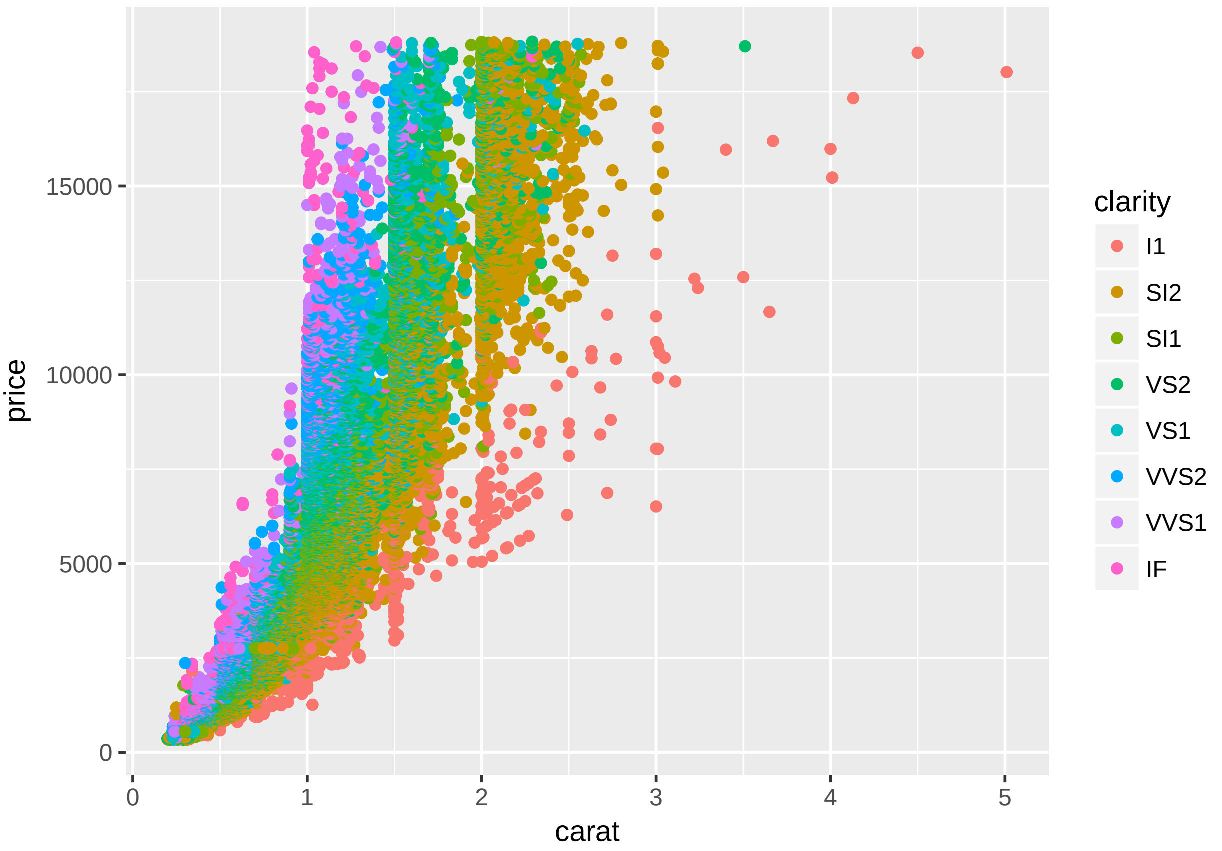
Scatter plot

```r
# Loading the ggplot2 library
library(ggplot2)

# Reading in the data file
data(diamonds)

# Creating a ggplot
myplot <- ggplot(diamonds, aes(x = carat, y = price, color = clarity)) + geom_point()

# print the plot
myplot
```



Comments:

There is few things to observe from just looking at this graph:
- We notice that the clearer the diamond (clarity) the higher the price.
- We notice that the heavier the diamond (carat) the higher the price.
- We notice that as the heavier the diamond gets (carat), the less likely we are to find it being of the highest clarity. As we can see from 3 carats and above, there's mainly SI2 and I1 clarity type diamonds available.