HW1

Henrique Rio

9/10/2020

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyv
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4
## -- Conflicts ------------------------------------------------------------------------ tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data(diamonds)



help(diamonds)
```

top 10 most expensive diamonds

```
diamonds %>% arrange(desc(price)) %>% head(10)
```

```
## # A tibble: 10 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  2.29 Premium   I     VS2      60.8    60 18823  8.5   8.47  5.16
##  2  2    Very Good G     SI1      63.5    56 18818  7.9   7.97  5.04
##  3  1.51 Ideal     G     IF       61.7    55 18806  7.37  7.41  4.56
##  4  2.07 Ideal     G     SI2      62.5    55 18804  8.2   8.13  5.11
##  5  2    Very Good H     SI1      62.8    57 18803  7.95  8     5.01
##  6  2.29 Premium   I     SI1      61.8    59 18797  8.52  8.45  5.24
##  7  2.04 Premium   H     SI1      58.1    60 18795  8.37  8.28  4.84
##  8  2    Premium   I     VS1      60.8    59 18795  8.13  8.02  4.91
##  9  1.71 Premium   F     VS2      62.3    59 18791  7.57  7.53  4.7
## 10  2.15 Ideal     G     SI2      62.6    54 18791  8.29  8.35  5.21
```

top 10 least expensive diamonds

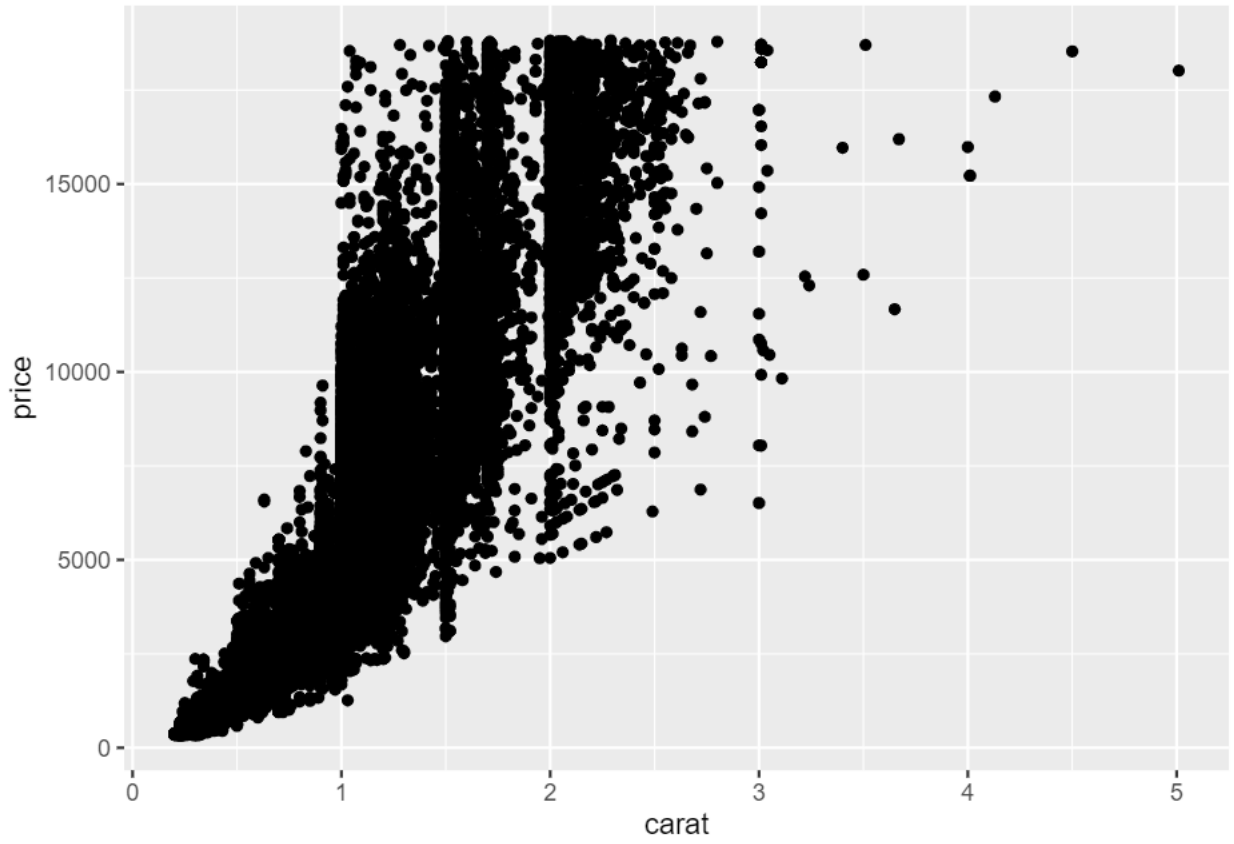
```
diamonds %>% arrange(price) %>% head(10)
```

```
## # A tibble: 10 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 0.290 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
```

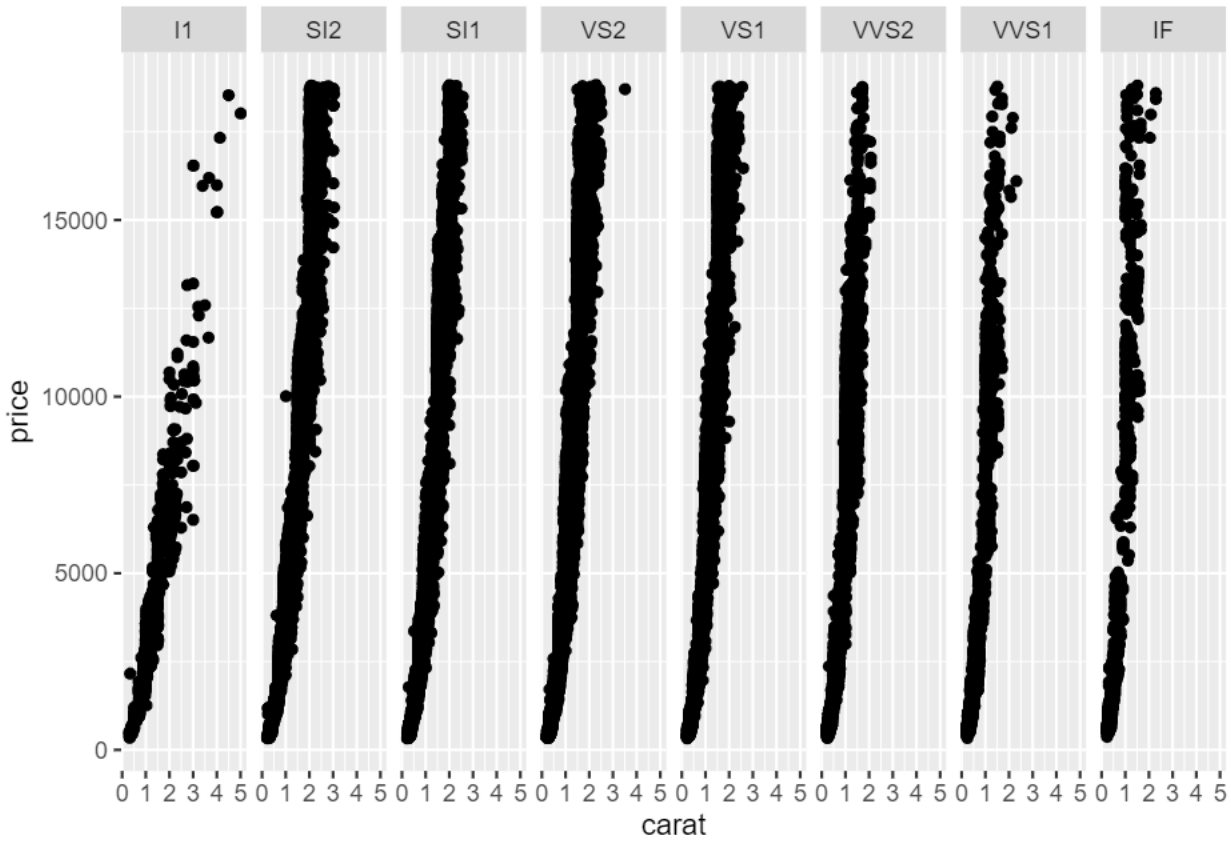
if we look at the most and least expensive diamonds we can see that there is a lot of overlap between the two, which we can further examine by looking at a scatter plot. In the first scatter plot we will look at Carat x Price.

```
ggplot(data=diamonds, aes(carat,price))+geom_point()
```



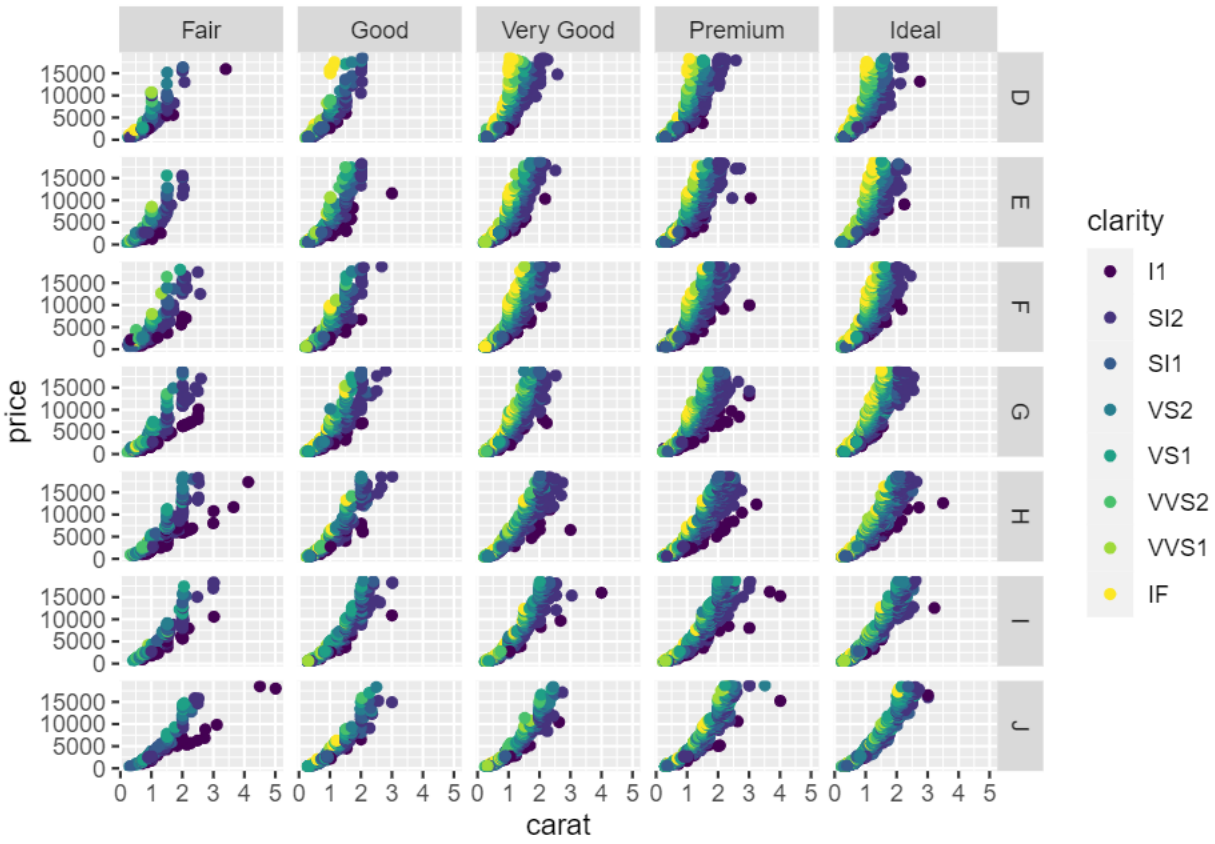
It is clear that diamonds that have more the 1 carat, have other factors influencing the price since there appears to be a lot of variation among those.

```
ggplot(data=diamonds, aes(carat,price))+geom_point()+facet_grid(.~diamonds$clarity)
```



If we separate the diamonds by their clarity, we still do not have a lot of information on what influences the price since they pretty much look the same.

```
ggplot(data=diamonds, aes(carat,price,color=clarity))+geom_point()+facet_grid(diamonds$color~diamonds$c
```



If we separate by colour cut and clarity, we still see almost the same trend between all of the diamonds, which shows that this data is problematic as not only there is not enough data on diamonds above 3 carats, its alsommn hard to determine the factor that influence the price of diamond since almost all of them have cheap and expensive diamonds. This makes me assume that there are other factor not include in the data that have a lot of influence of on the price of the diamonds.