

Mispriced Diamonds

Poro Burman

7/8/2020

```
load the 'mispriced_diamonds_data.txt' file
```

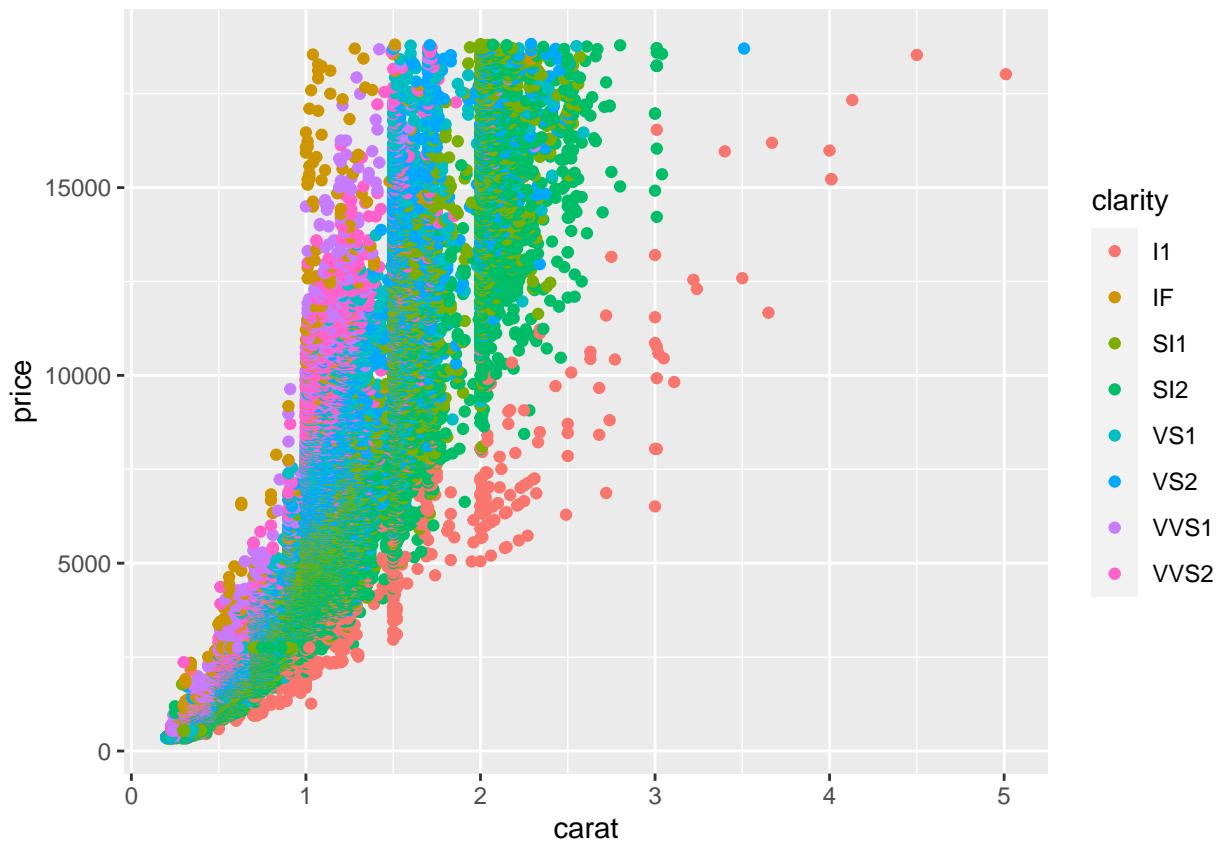
```
mispriced_diamonds <- read.csv("Mispriced_Diamonds_data.txt")
```

```
load 'ggplot2 package' to visualize the data
```

```
library("ggplot2")
```

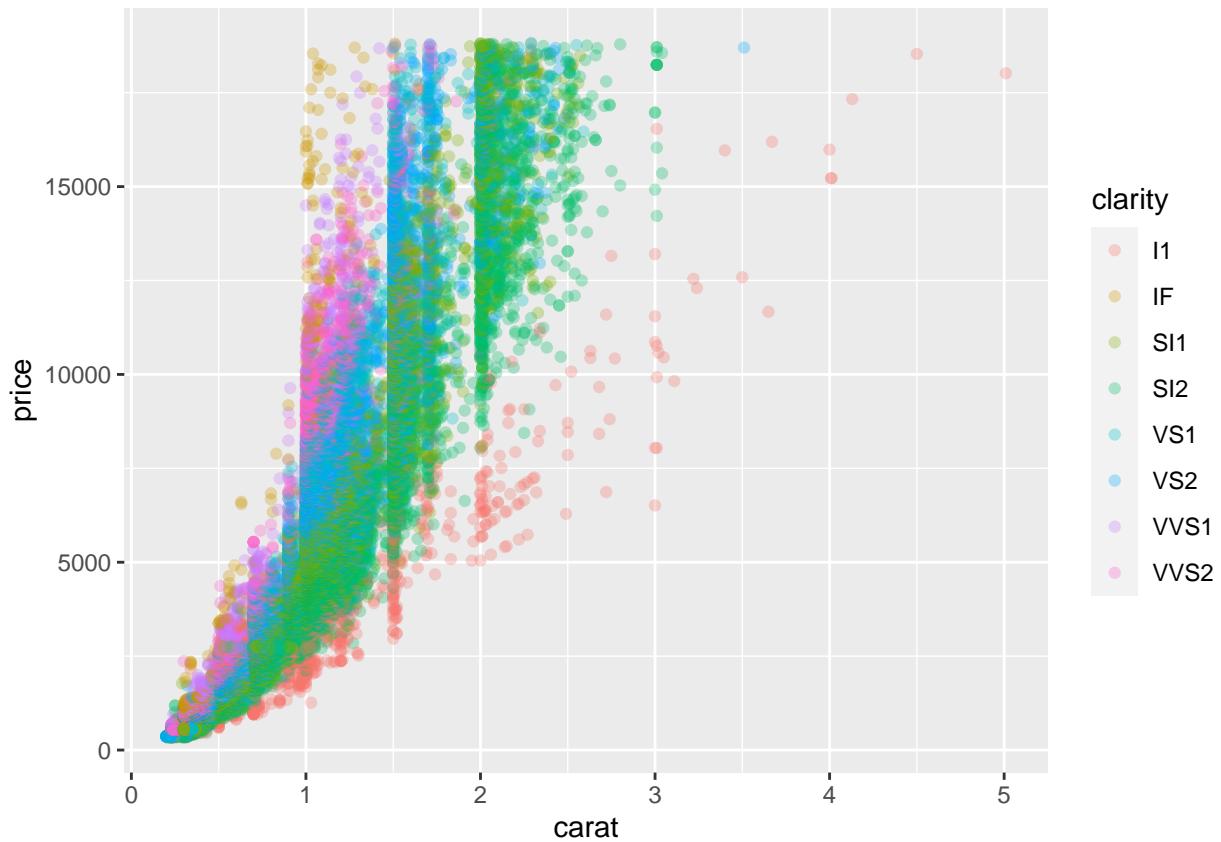
```
create a scatter plot between the 3 different variables of the dataset
```

```
ggplot(data = mispriced_diamonds, aes(x=carat, y=price, colour=clarity)) +  
  geom_point()
```



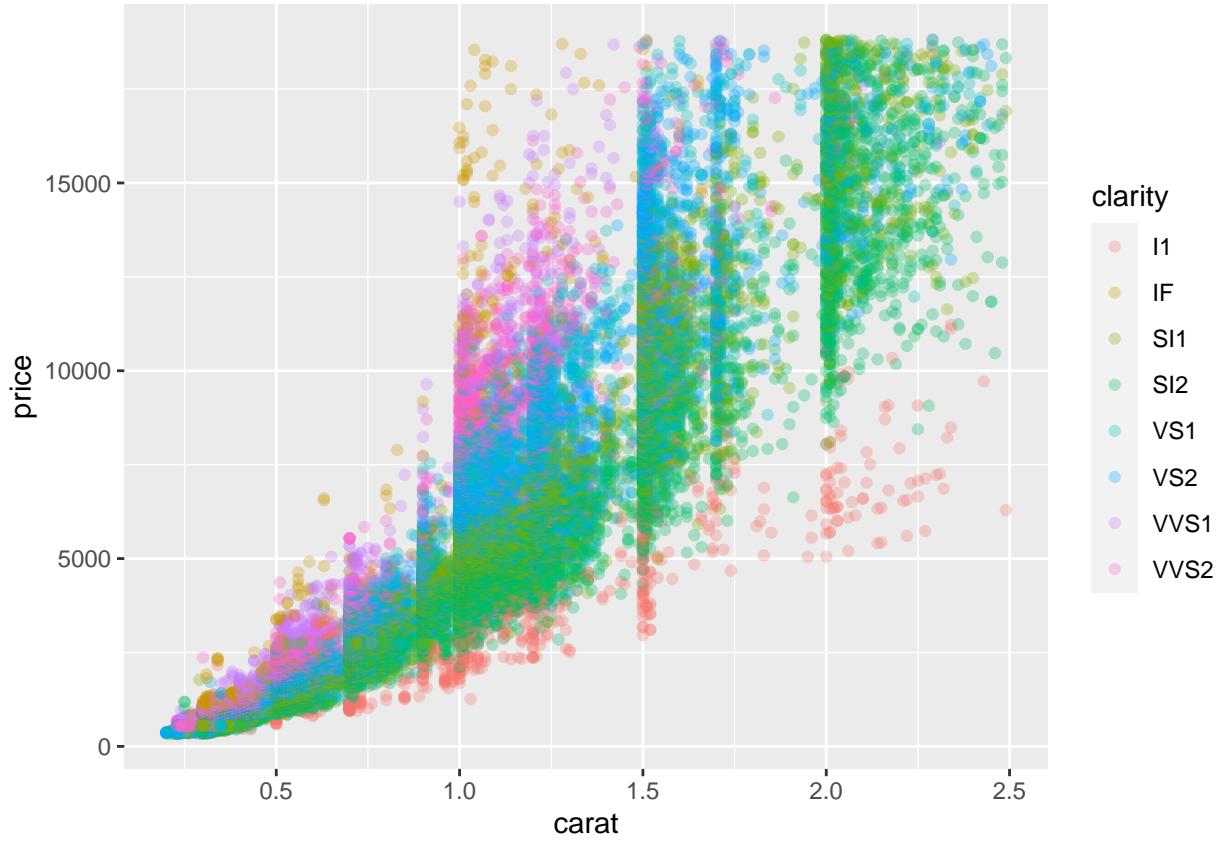
the image looks to cluttered so going to reduce alpha to 0.3 to get a better view

```
ggplot(data = mispriced_diamonds, aes(x=carat, y=price, colour=clarity)) +
  geom_point(alpha=0.3)
```



looking at the image now, it seems that there are a lot of outliers beyond 2.5 carat. These values aren't required so now I'll plot the data without those points

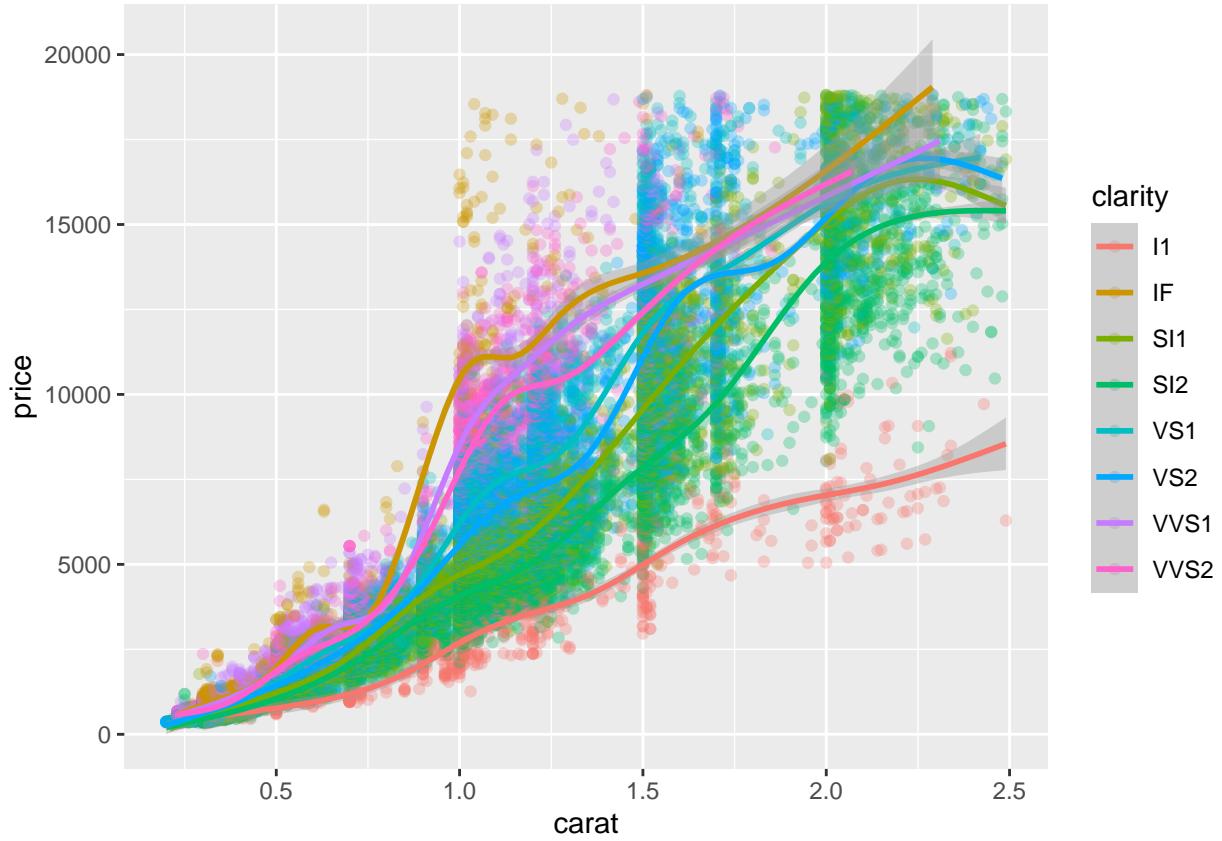
```
ggplot(data = mispriced_diamonds[mispriced_diamonds$carat < 2.5,], aes(x=carat,
  y=price, colour=clarity)) + geom_point(alpha=0.3)
```



I'm going to add the 'geom_smooth' to this graph so that I can get a view of the lines representing averages for each of the clarities.

```
ggplot(data = mispriced_diamonds[mispriced_diamonds$carat < 2.5], aes(x=carat,
y=price, colour=clarity)) + geom_point(alpha=0.3) + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Now when one looks at the figure, one will realize that there seems to be some intersections between the clarities. The pricing of different clarities shouldn't intersect. Looking at carat = 0.7, there seems to be diamonds of different clarities which might be priced together. This is also observed at carat = 1.7 and carat = 2.1. This shows that there might be a possibility that the prices of the diamonds might be incorrectly labelled for these carats.