Data Visualization for predicting dimonds'prices based on their clarity

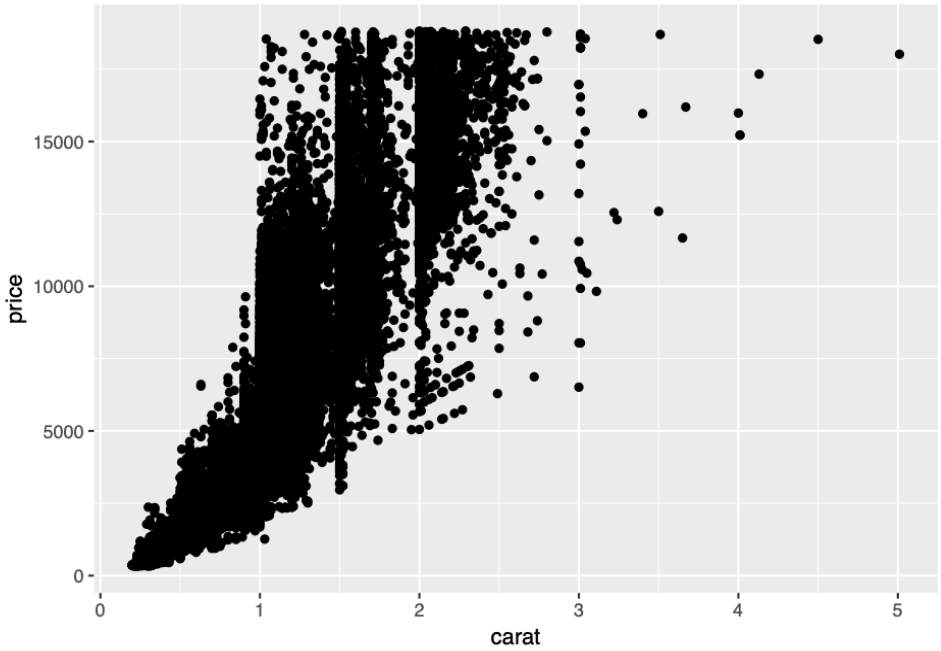
Peiyun Zhou

This is an example using ggplot to visualize the data The Research Question is "Does a better clarity always mean a higher sale price?""

```
library(ggplot2) # load the library
# one way to load the dimond data
# mydata<-read.csv((file.choose()))
# Another way to load the data
# setwd("The directory of where the file is located") # Set working directory
mydata<-read.csv("Mispriced-Diamonds.csv",header=T,sep=",")
```

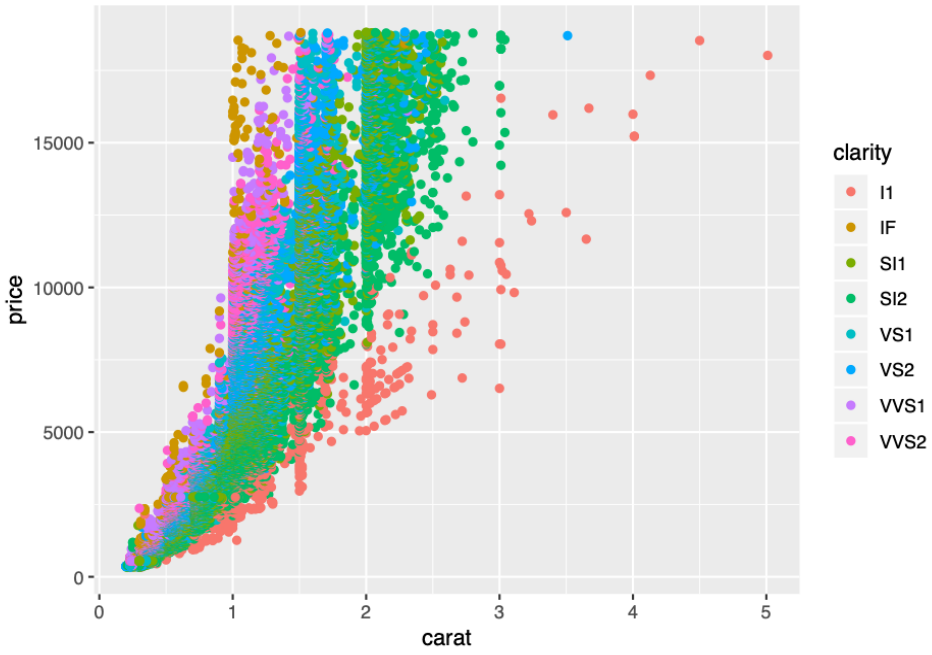
First, let's plot the relationship between the size of carat and the sale prices

```
ggplot(data=mydata,aes(x=carat,y=price))+
  geom_point()
```



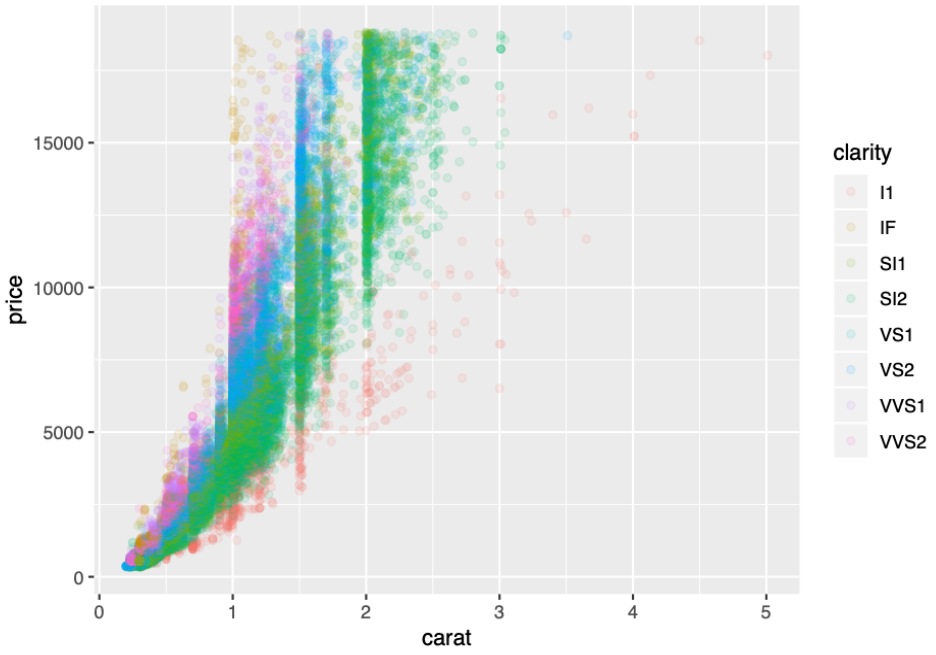
Then, let's use different colors to show the clarity

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



We can adjust the transparency of the dots by changing the alpha value.

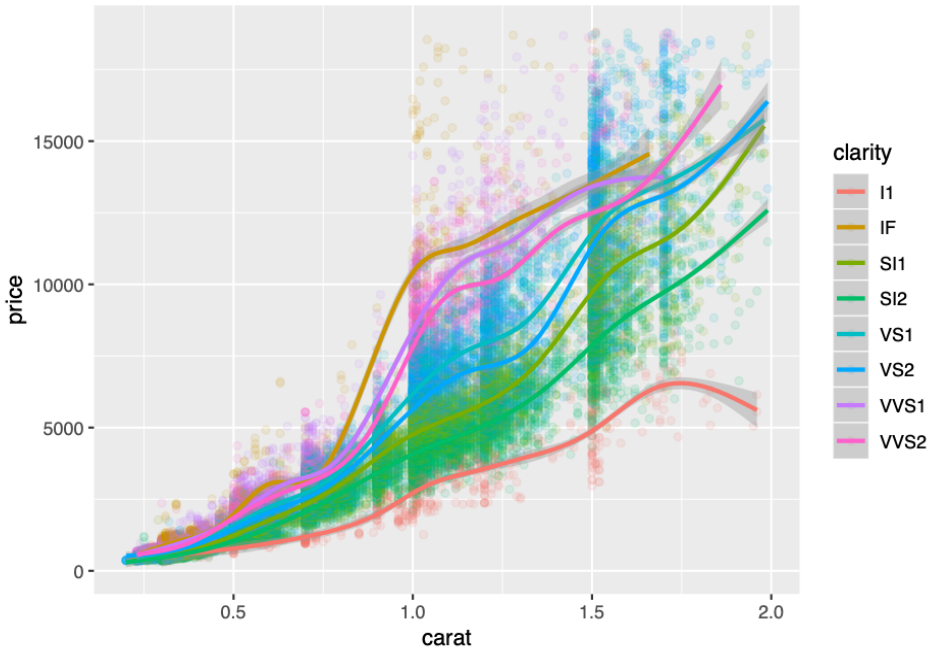
```
ggplot(data=mydata,aes(x=carat,y=price,colour=clarity))+
  geom_point(alpha=0.1)
```



We will look at the diamonds that are smaller than 2 carets and add smooth lines to illustrate the patterns. Based on the plot below, we found that clarity and price is not always a positive linear relationship. Depends on how big the diamond is, the price differ based on its clarity.

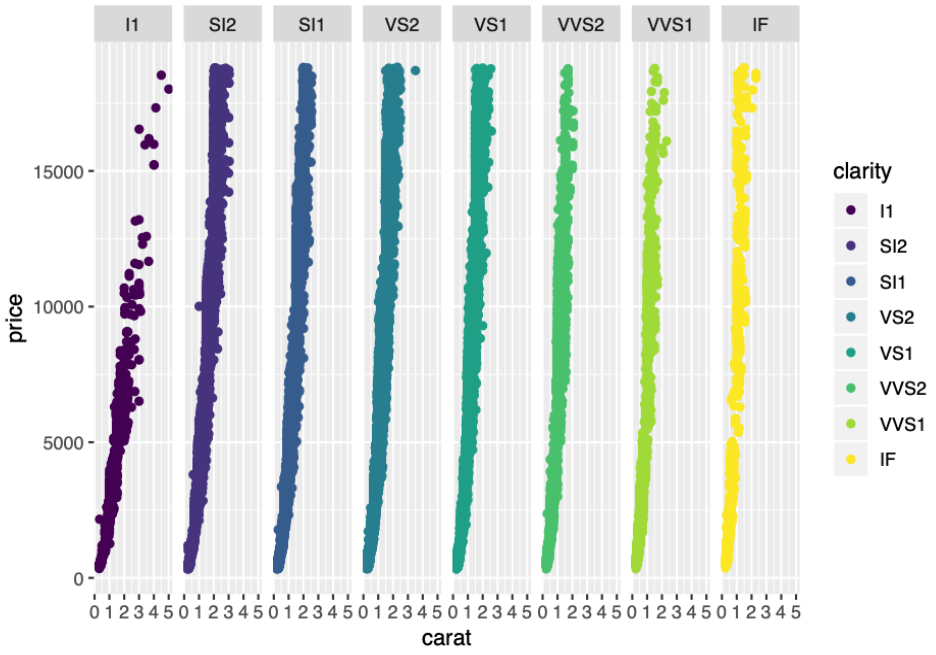
```
ggplot(data=mydata[mydata$carat<2,],aes(x=carat,y=price,colour=clarity))+
  geom_point(alpha=0.1)+
  geom_smooth()
```

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



We can also look the price for each dimond based on their clarity and carat size

```
qplot(data=diamonds,carat,price,colour=clarity,facets=.~clarity)
```


I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
price
15000
10000
5000
0
0 1 2 3 4 5
carat
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF