R Notebook

```
#Arjun Bhan
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

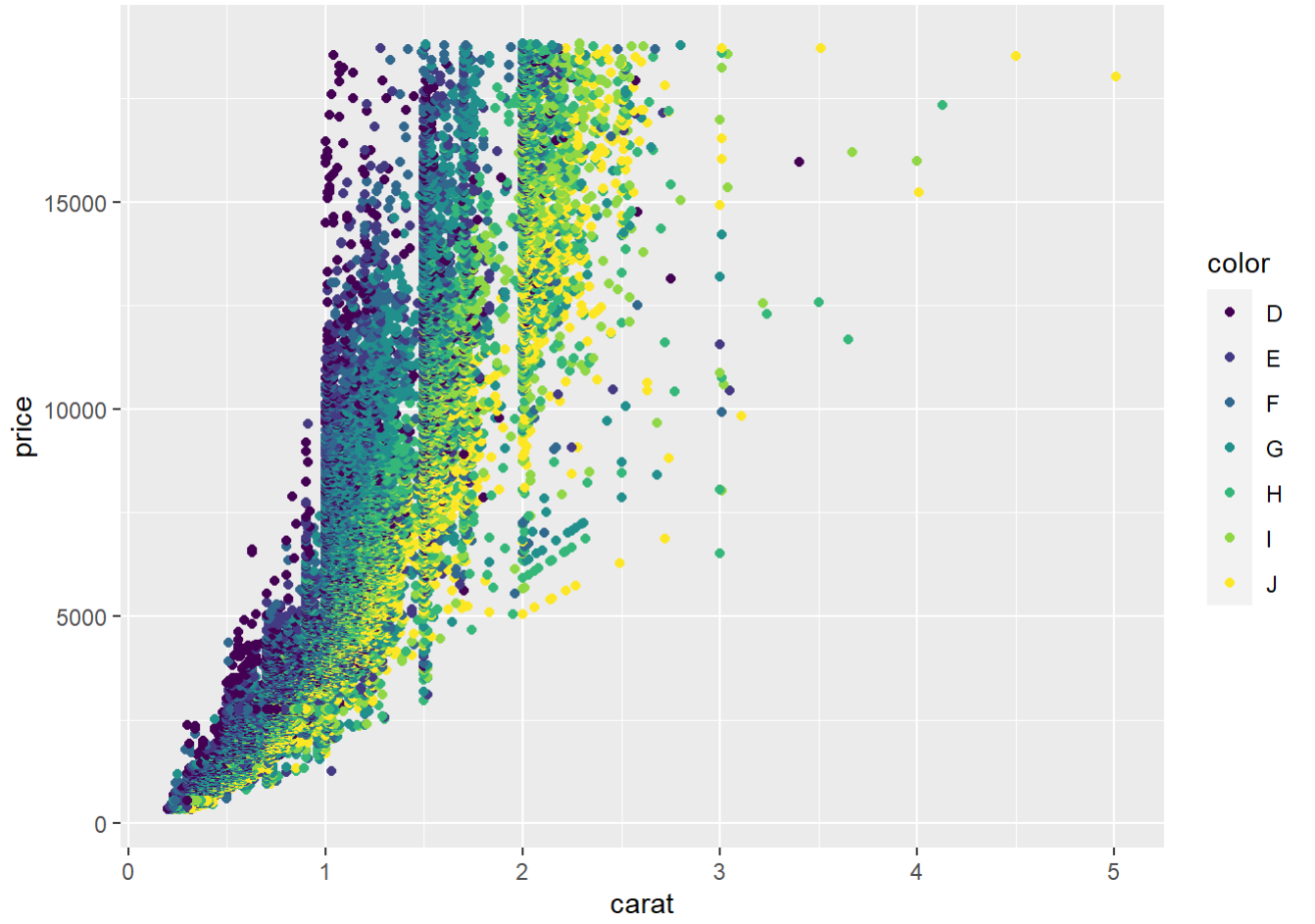
```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(RColorBrewer)
diamonds
```

carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39

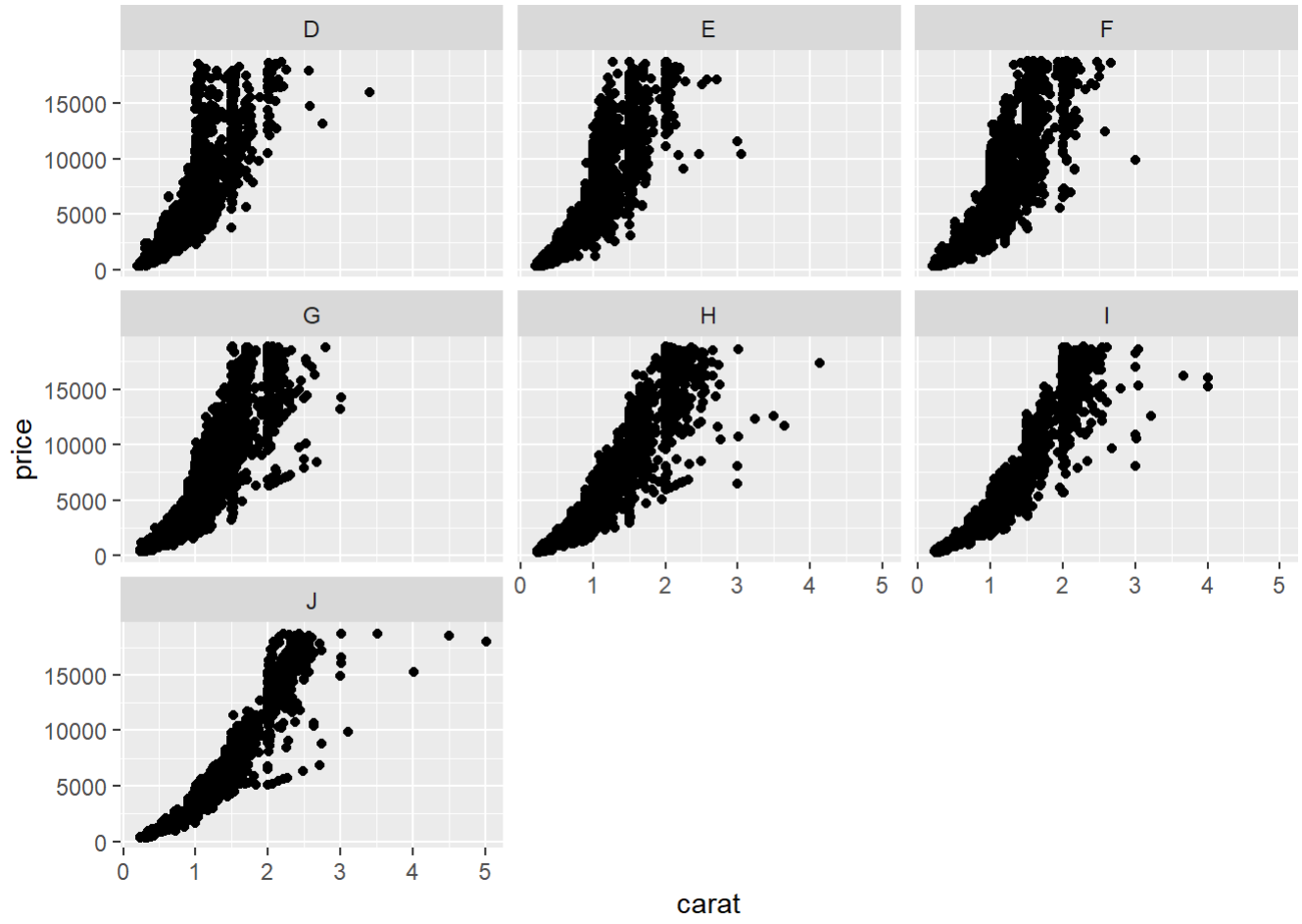
1-10 of 10,000 rows Previous **1** 2 3 4 5 6 … 1000 Next

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




```
# A reason why the data is so hard to evaluate is that there is too many indivshual data point f
or a dot plot to be used effectivly.
```

```
ggplot(diamonds, aes(carat,price))+geom_point()+facet_wrap(~color)
```



 # I feel that faceting the data makes it easier to interpret. It breaks the indvishual color dat a by its values. This helps us see the price and carat values for each of the 7 color groups. Th e data of each color is so simular that it can be hard to compare them visually. The data visual ization can be improved by showing the average price for each of the groups. This is because it is simpiler to see that some groups have a higher average price than others. It also helps solv e the issue of there being too many data points and variables (factors that impact price) to int erpret.

```
diamonds$CarType=cut_width(diamonds$carat, .2)
```

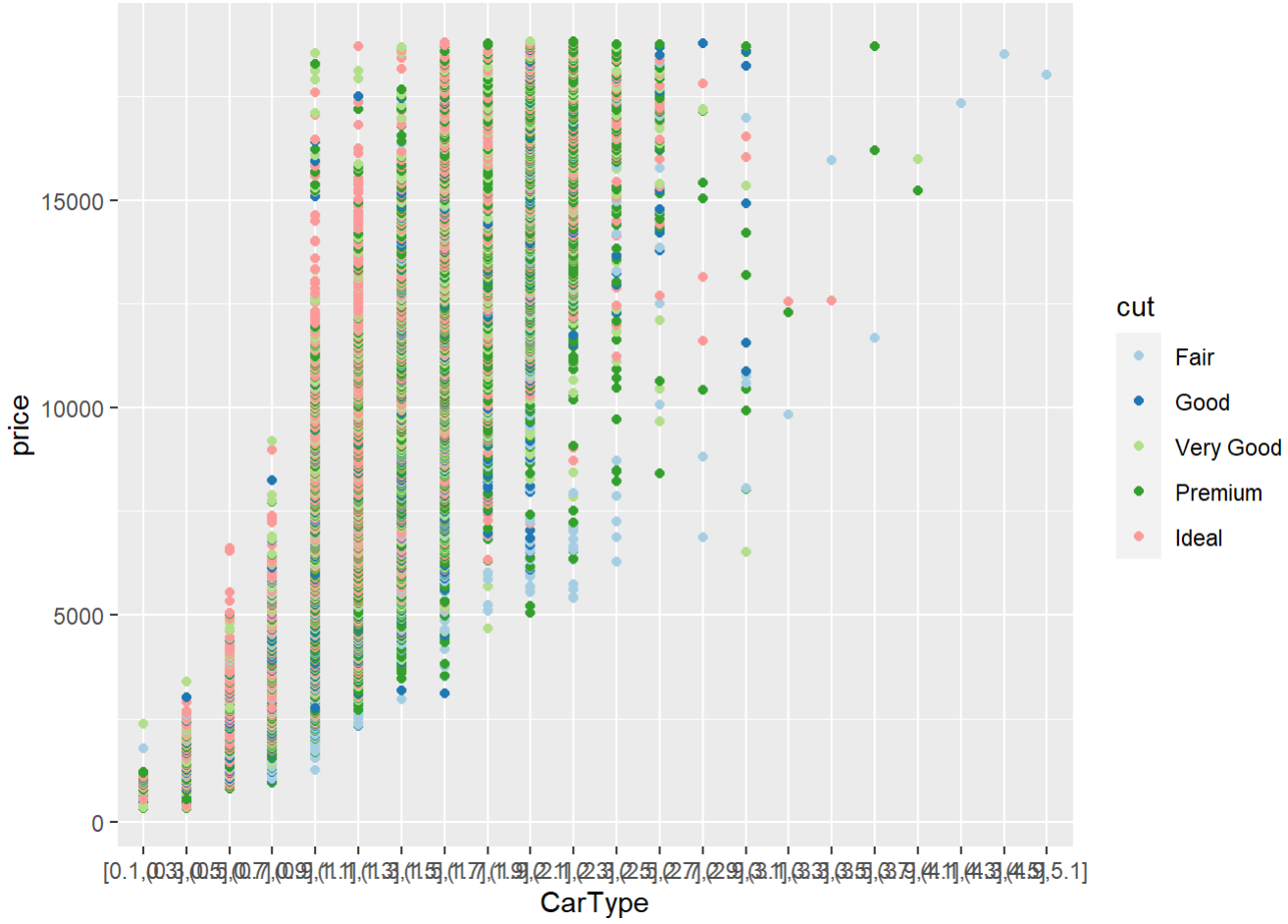
```
NewDim<-diamonds%>% group_by(cut,CarType)
NewDim
```

carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39

1-10 of 10,000 rows | 1-10 of 11 columns Previous **1** 2 3 4 5 6 … 1000 Next

```
ggplot(NewDim, aes(CarType,price,color=cut))+geom_point()+scale_color_brewer(palette="Paired")
```



#All of these distributions tell me that there is a postive correlation between the carat rating and the price of diamonds. This is true for all cuts. I feel that the second graph tells this story the clearest by dividing each diamond by their color group.