Lab 3

Suixin Jiang

9/15/2019

Diamonds

The goal of the diamonds dataset is to see what characteristics are most influential on price. Perform an exploratory data analysis and come up with some conjectures on what variables impact price. Can some associations be explained by other variables? For example, can the decrease in price as the cut worsens be explained by the carat of the diamond?

```
suppressPackageStartupMessages(library(tidyverse))
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
## Warning: package 'tidyr' was built under R version 3.5.3
```

```
## Warning: package 'purrr' was built under R version 3.5.3
```

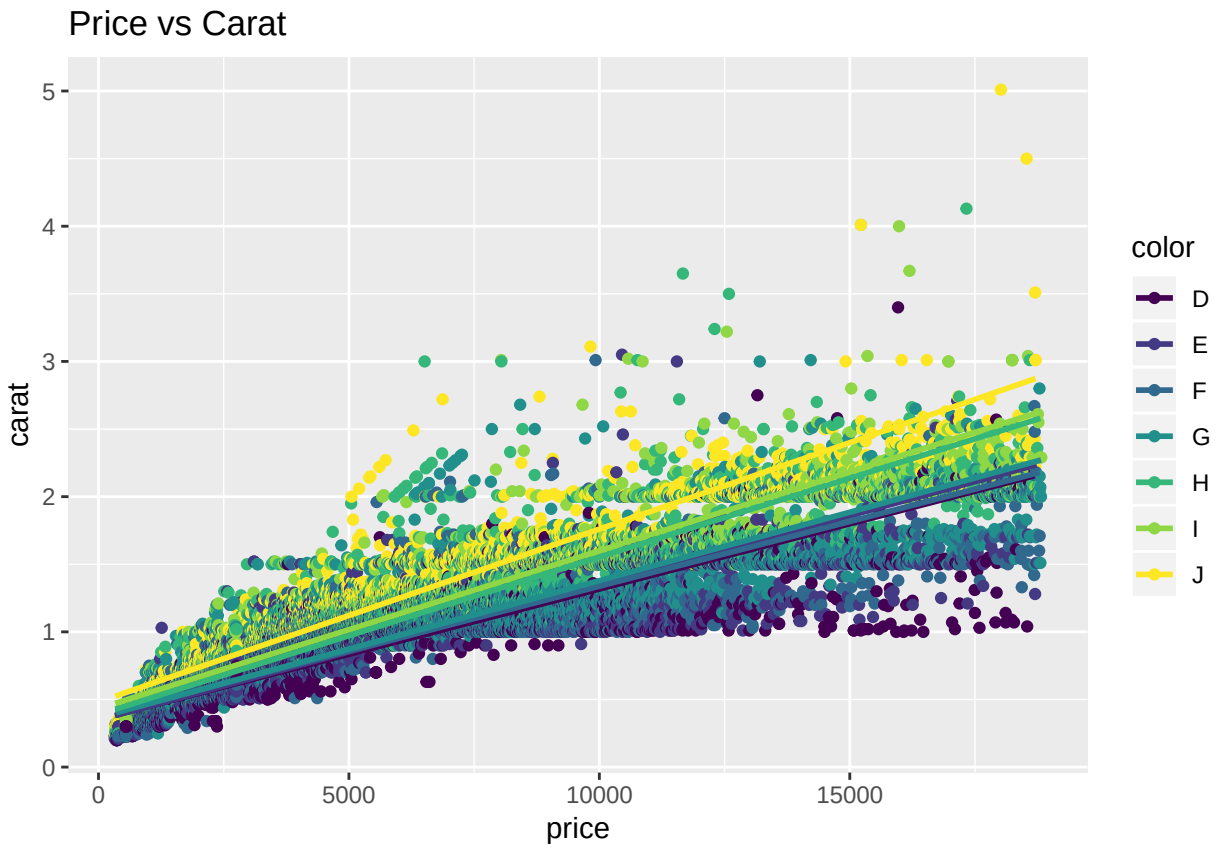
```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
library(ggplot2)
D = diamonds
```

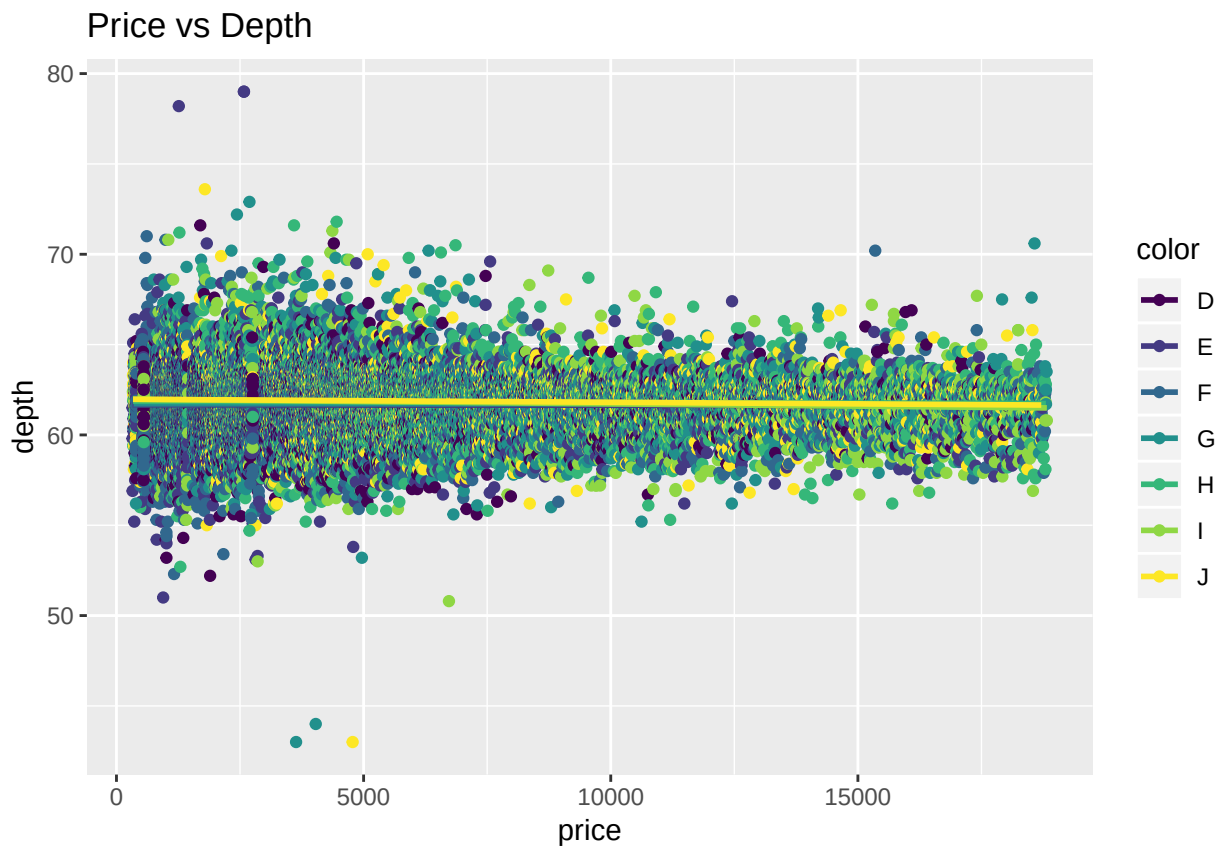
Task

Make a plot of price vs all other variables.

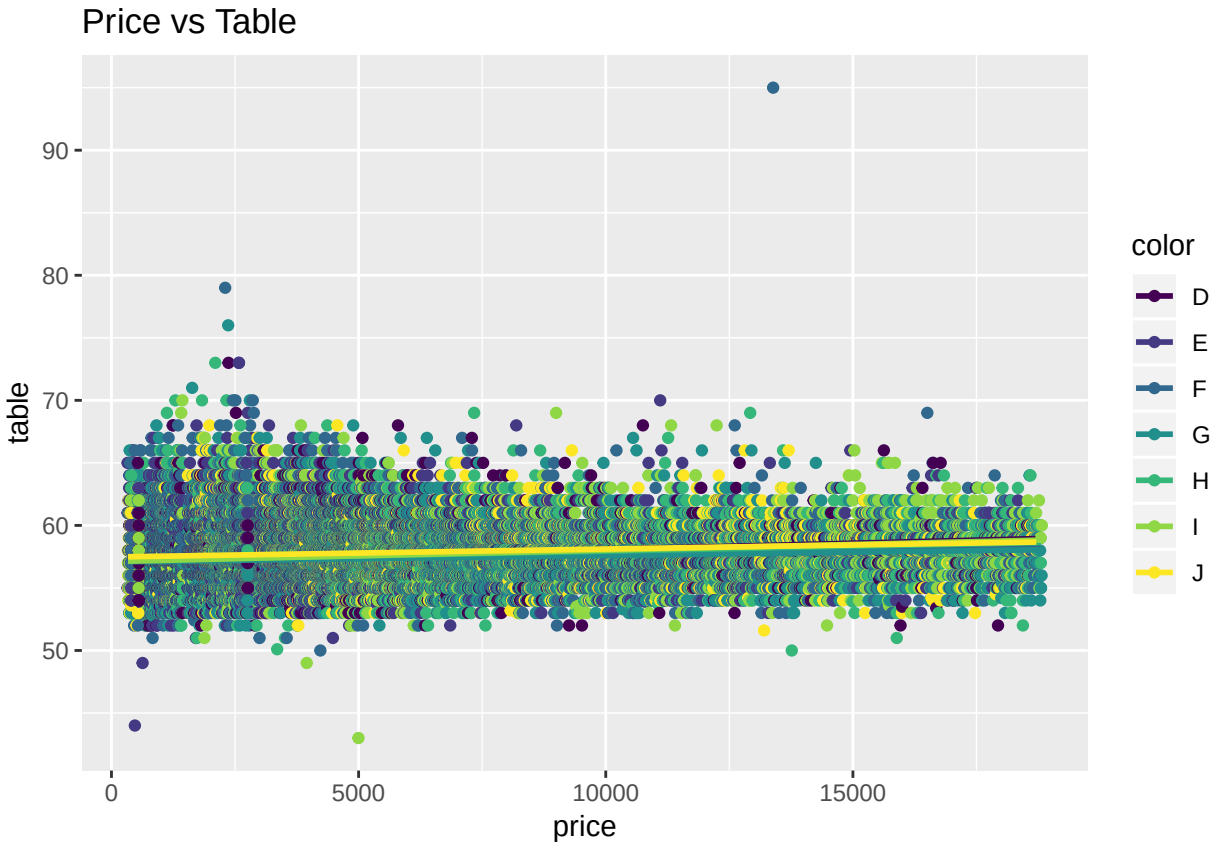
```
# Price vs Carat. Overall, there is a positive relationship between price and
# carat, which the heavier the more expensive. And for same weights diamonds,
# the better color the more expensive.
ggplot(data = D, mapping = aes(x = price, y = carat, color = color)) +
  geom_point() +
  geom_smooth(method = glm, se = F) +
  ggtitle("Price vs Carat")
```



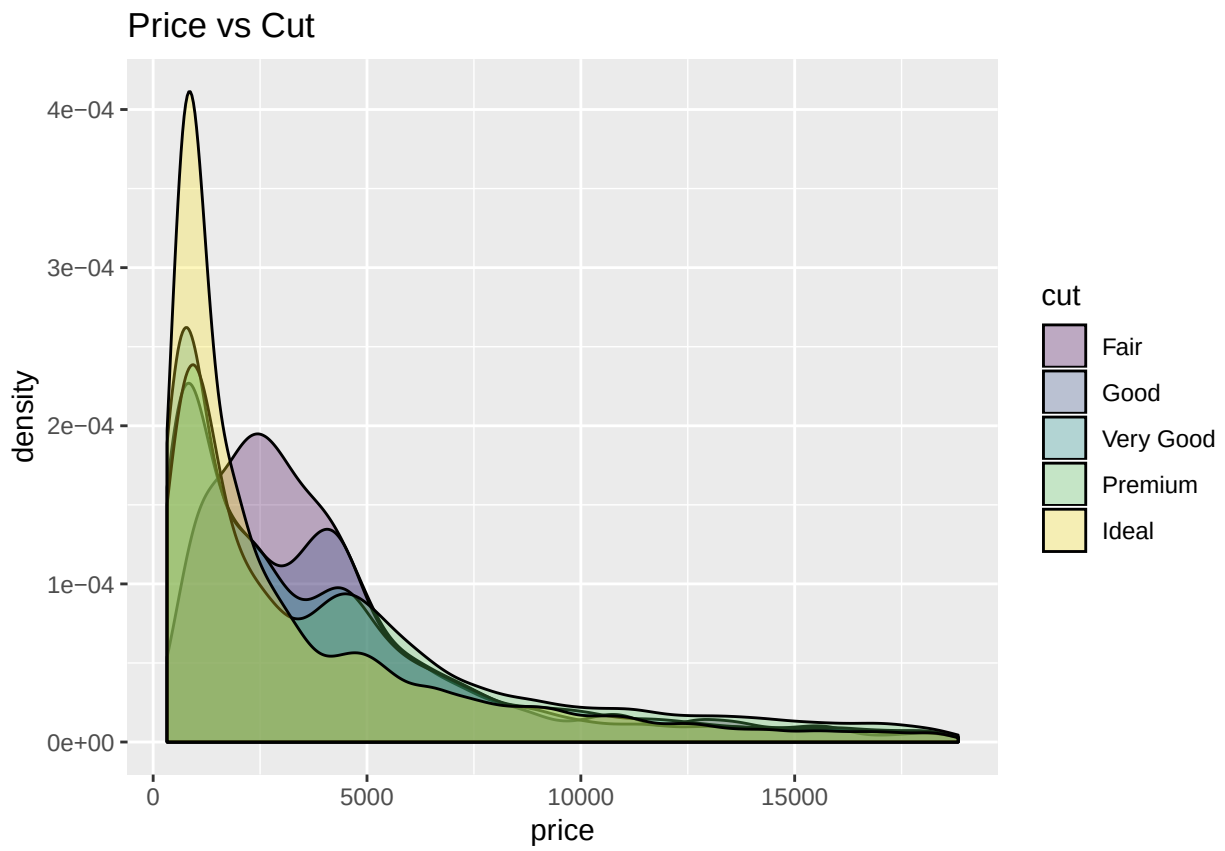

```
# Price vs Depth(& Table). For each color diamonds, depth and table do not have
# a significant influence on price. They are normally distributed.
ggplot(data = D, mapping = aes(x = price, y = depth, color = color)) +
  geom_point() +
  geom_smooth(method = glm, se = F) +
  ggtitle("Price vs Depth")
```



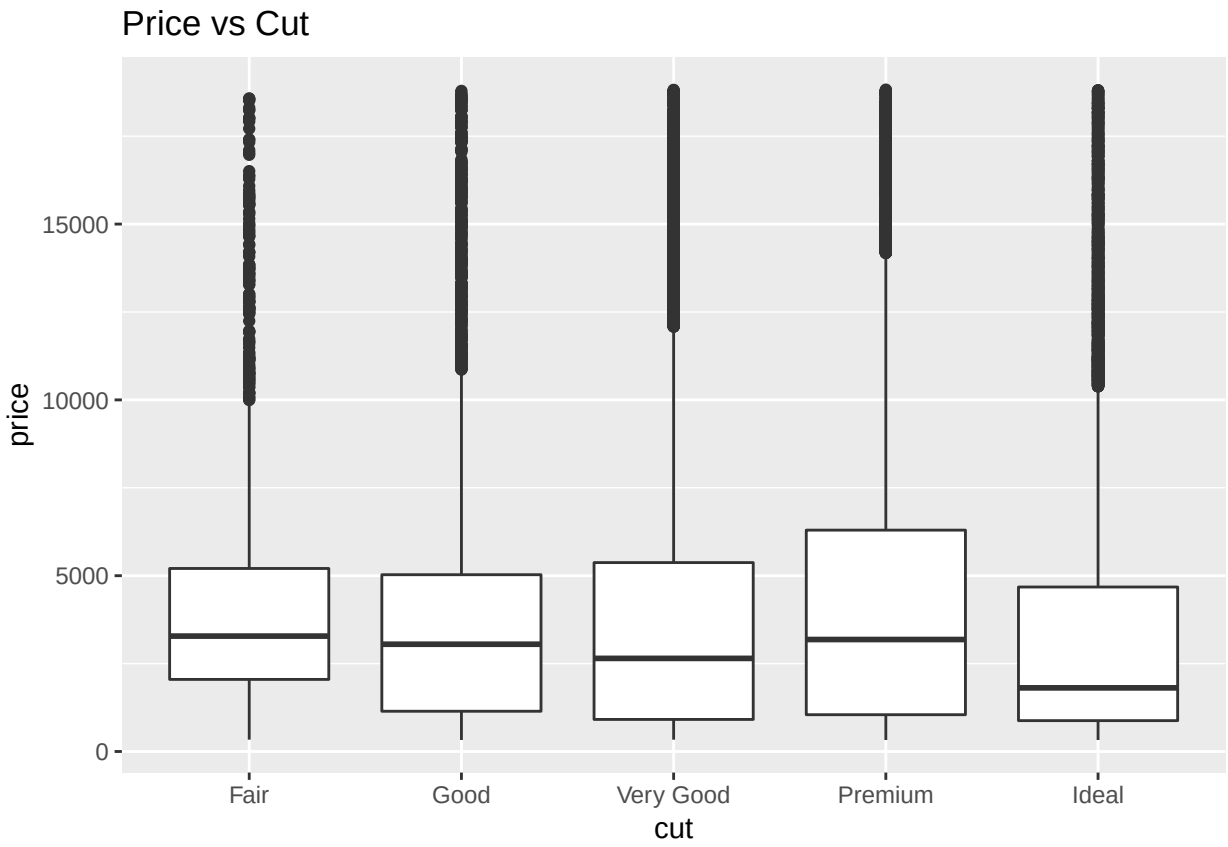

```
ggplot(data = D, mapping = aes(x = price, y = table, color = color)) +
  geom_point() +
  geom_smooth(method = glm, se = F) +
  ggtitle("Price vs Table")
```



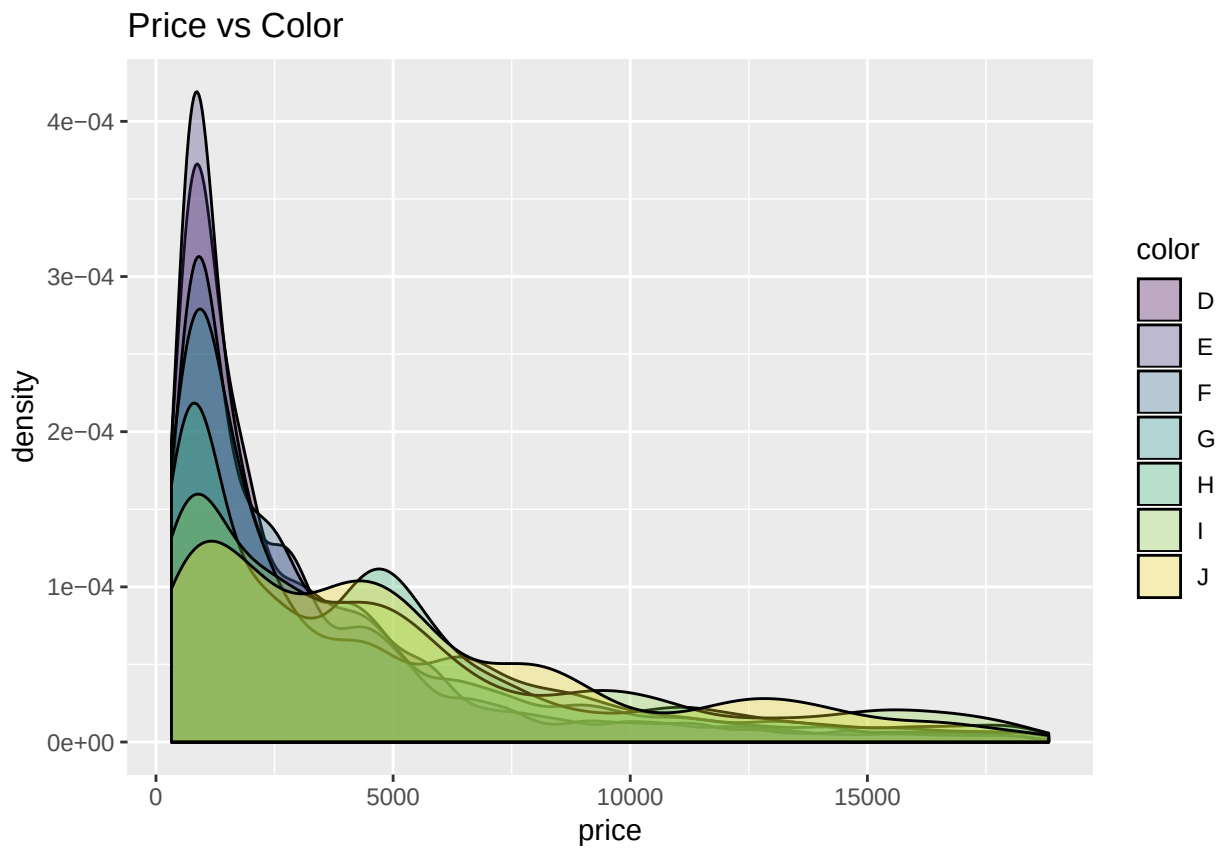

```
# Price vs Cut. Surprisingly, ideal cut diamond has a lower median, but for same
# weights diamonds cut would become a positive factor for the price.
ggplot(data = D, mapping = aes(x = price, fill = cut)) +
  geom_density(alpha = 0.3) +
  ggtitle("Price vs Cut")
```



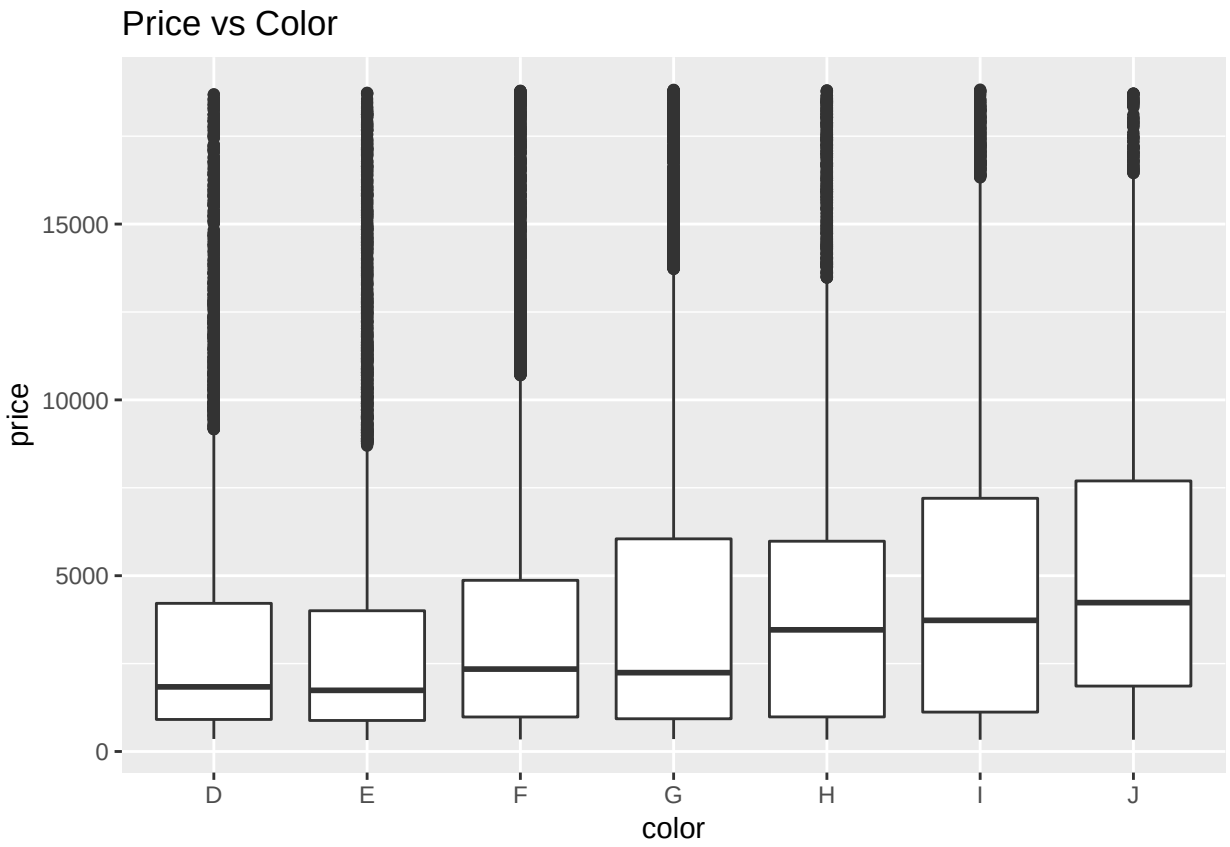

```
ggplot(data = D, mapping = aes(x = cut, y = price)) +
  geom_boxplot() +
  ggtitle("Price vs Cut")
```



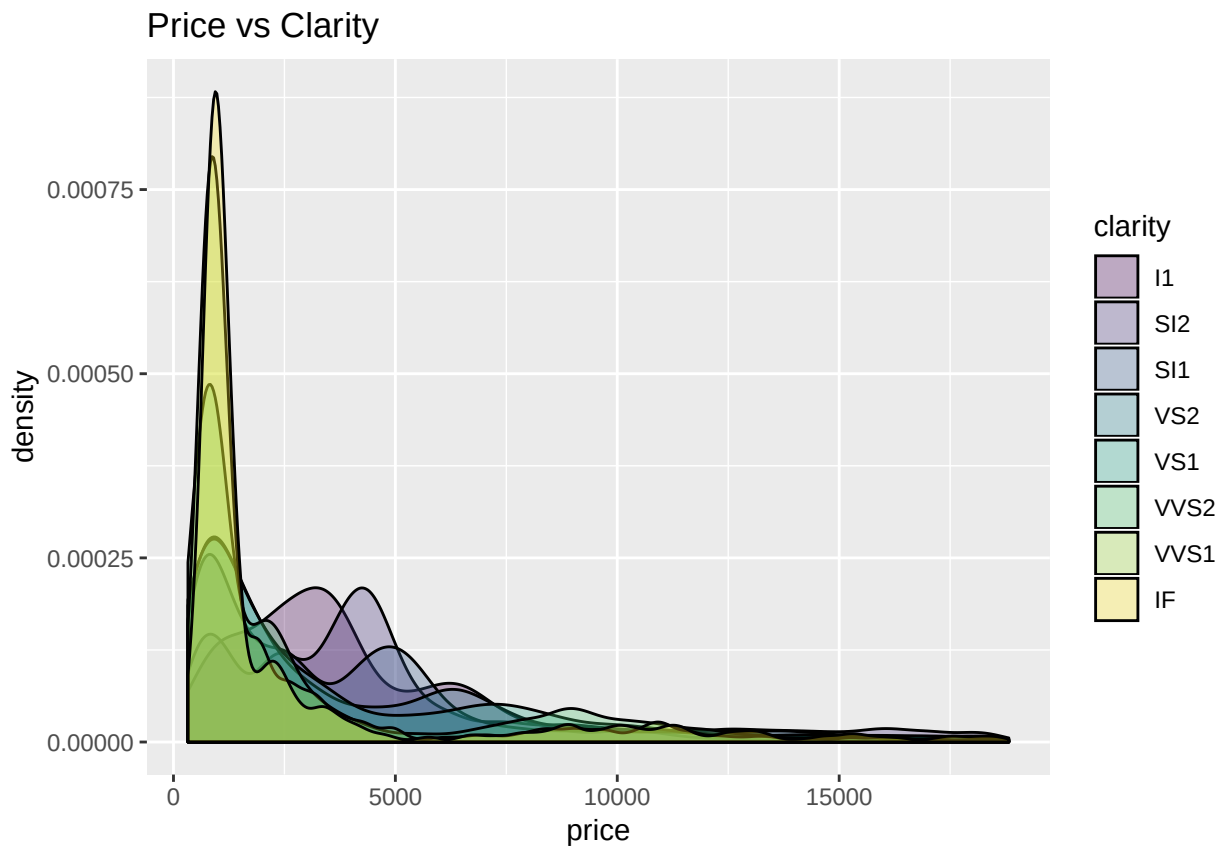

```
# Price vs color. Basically the same to cut, best color diamond has a lower
# median. But as we analyze above, for the same weights diamond the better
# the color the more expensive.
ggplot(data = D, mapping = aes(x = price, fill = color)) +
  geom_density(alpha = 0.3) +
  ggtitle("Price vs Color")
```



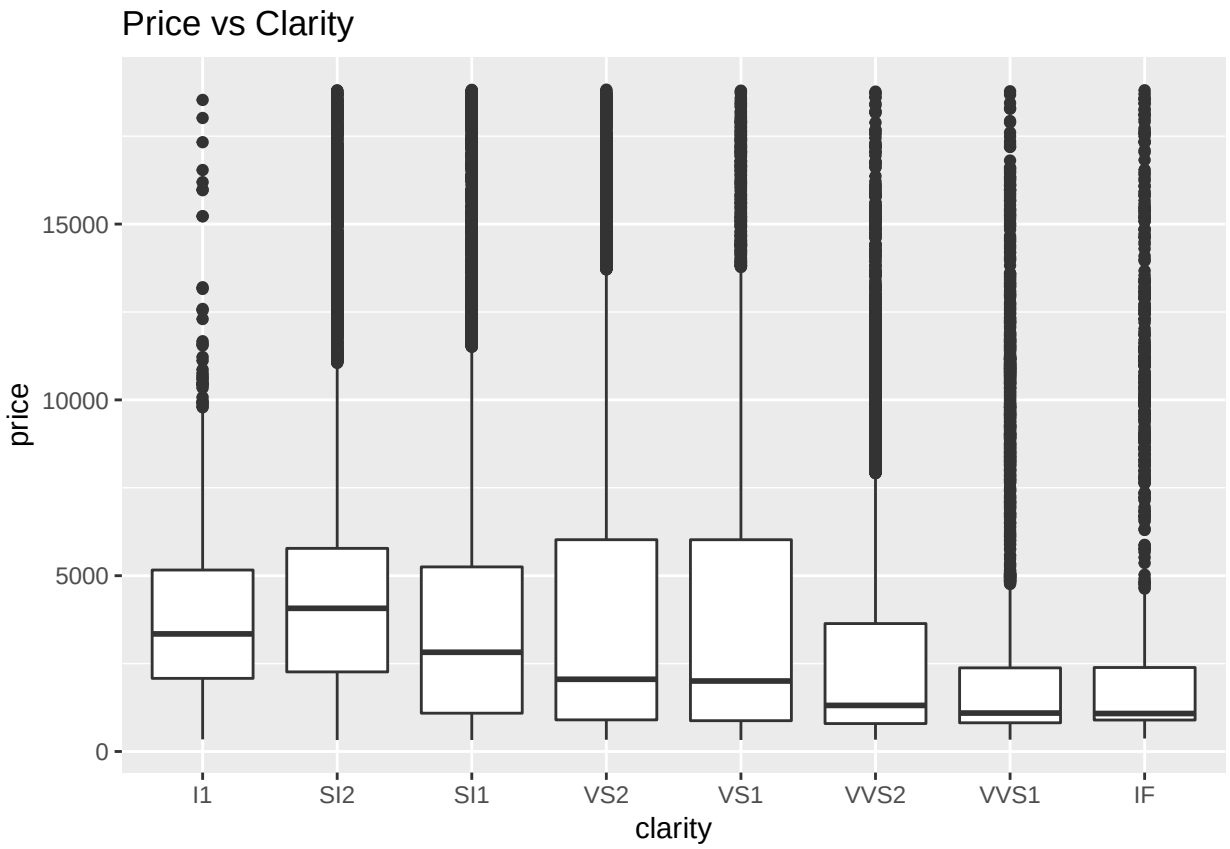

```
ggplot(data = D, mapping = aes(x = color, y = price)) +
  geom_boxplot() +
  ggtitle("Price vs Color")
```




```
# Price vs Clarity. The same as cut and color, the best clear diamond has a lower # median. However, wh
# be an important factor.
ggplot(data = D, mapping = aes(x = price, fill = clarity)) +
  geom_density(alpha = 0.3) +
  ggtitle("Price vs Clarity")
```



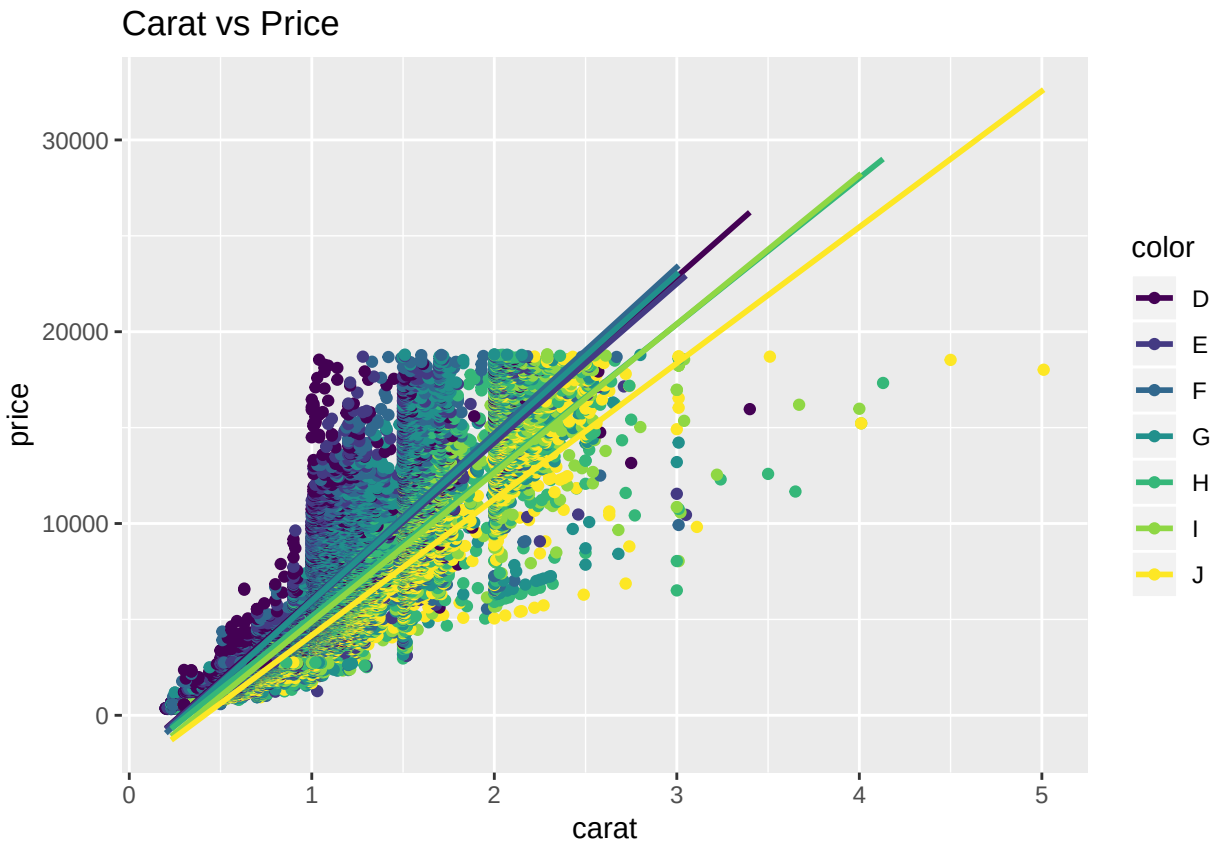

```
ggplot(data = D, mapping = aes(x = clarity, y = price)) +
  geom_boxplot() +
  ggtitle("Price vs Clarity")
```



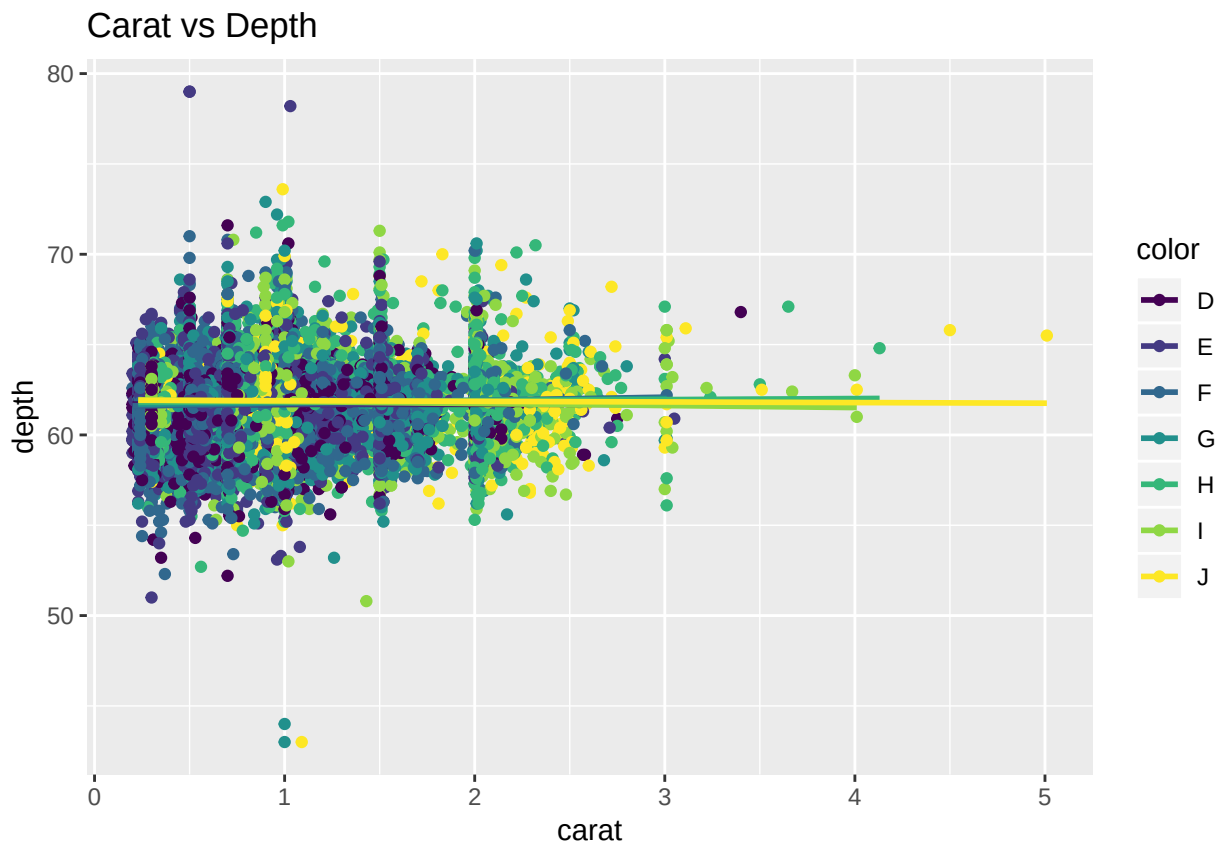
Task

Make a plot of carat vs all other variables.

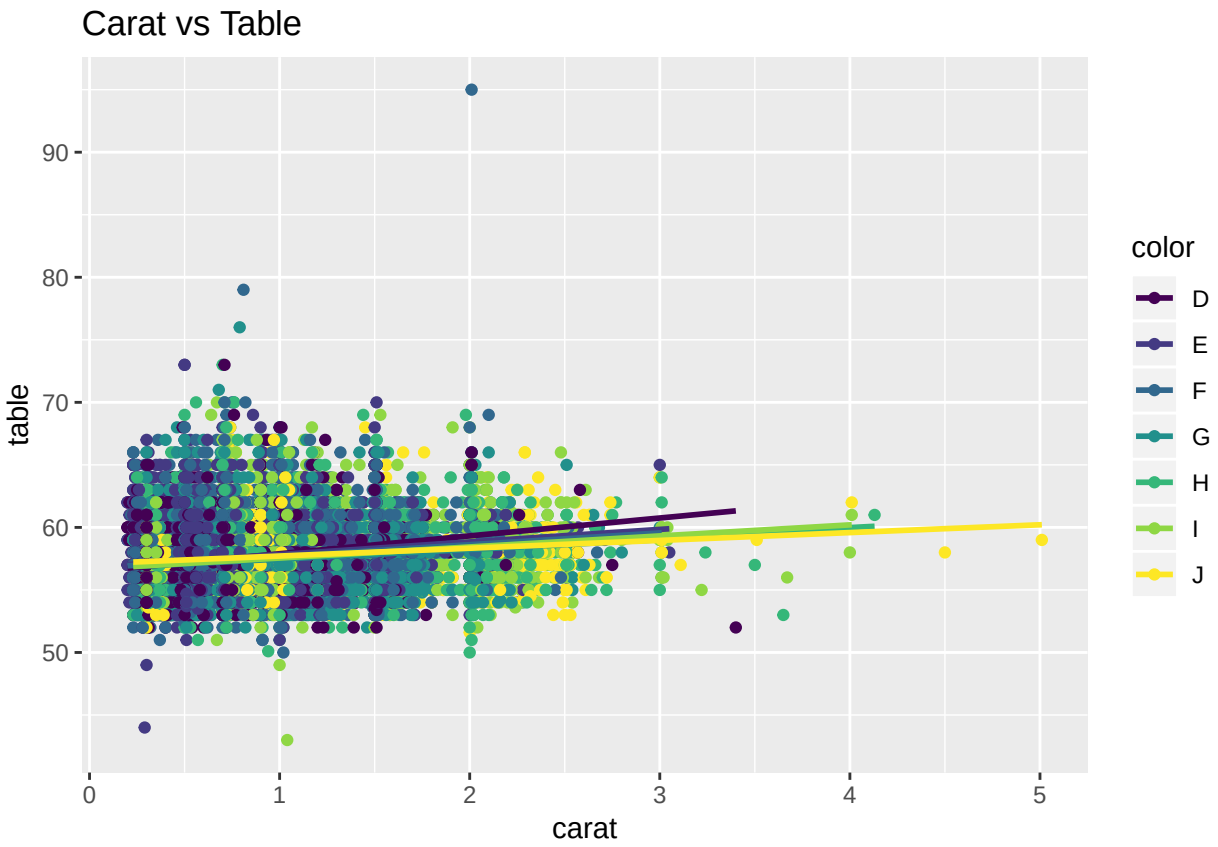
```
# Carat vs Price. There is a very clear positive relationship between carat and price.
ggplot(data = D, mapping = aes(x = carat, y = price, color = color)) +
  geom_point() +
  geom_smooth(method = glm, se = F) +
  ggtitle("Carat vs Price")
```



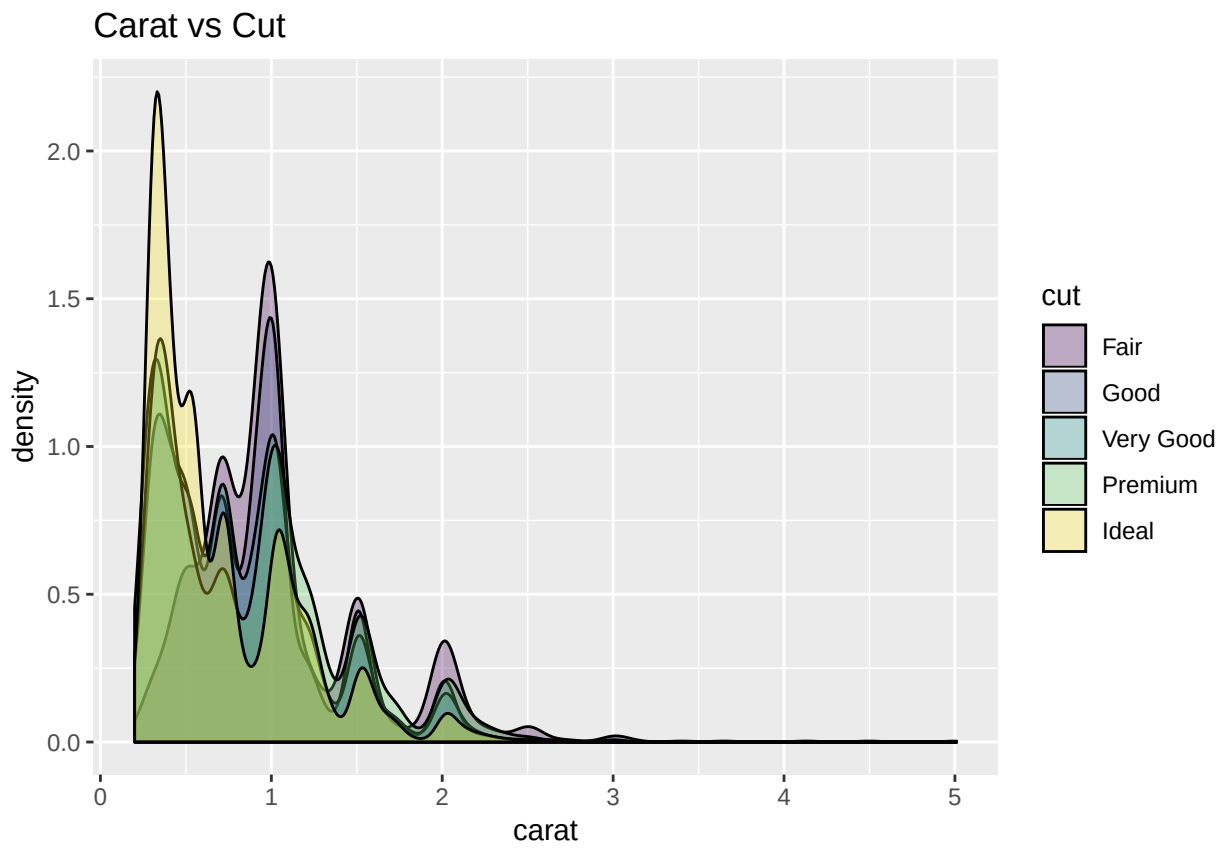

```
# Carat vs Depth(& Table). Depth and table do not influence on the price.
ggplot(data = D, mapping = aes(x = carat, y = depth, color = color)) +
  geom_point() +
  geom_smooth(method = glm, se = F) +
  ggtitle("Carat vs Depth")
```



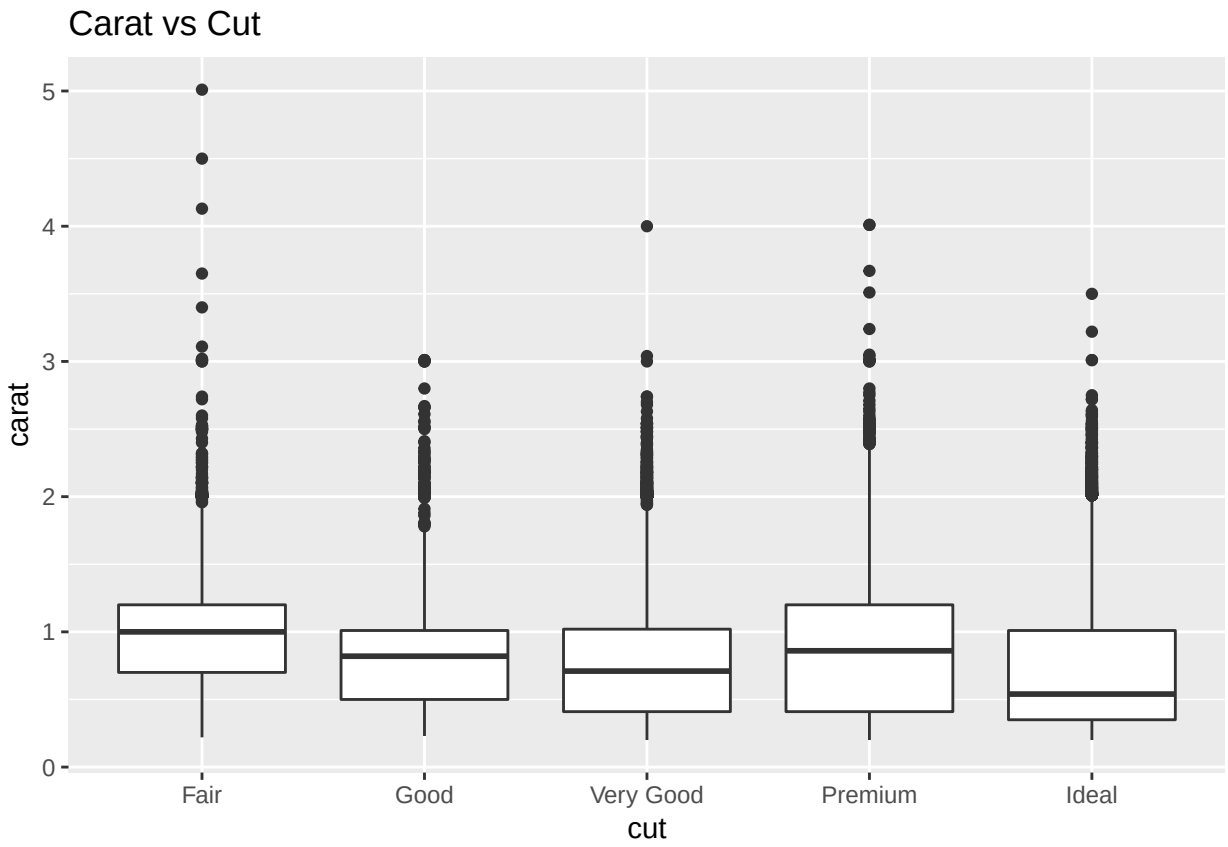

```
ggplot(data = D, mapping = aes(x = carat, y = table, color = color)) +
  geom_point() +
  geom_smooth(method = glm, se = F) +
  ggtitle("Carat vs Table")
```



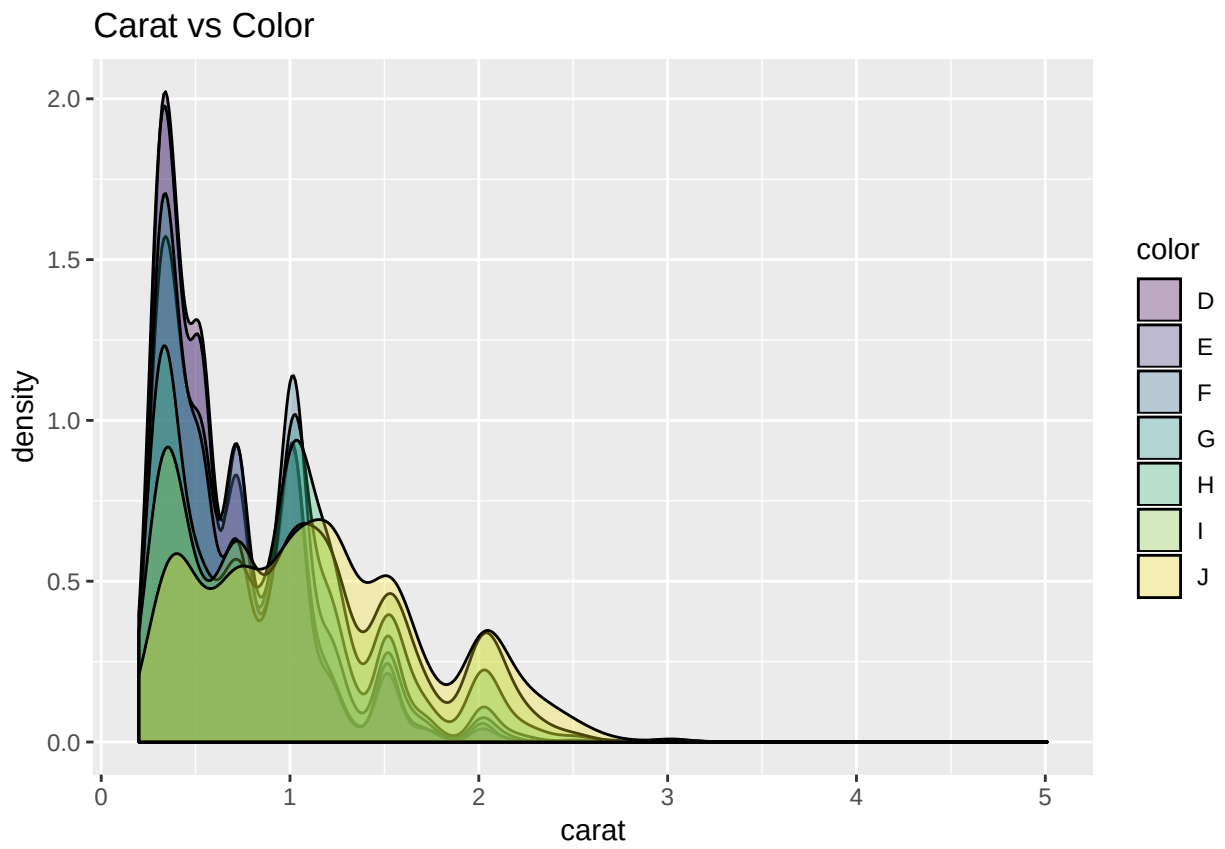

```
# Carat vs Cut. Regardless of the cut, the diamond below 2 carats has the
# majority of sales.
ggplot(data = D, mapping = aes(x = carat, fill = cut)) +
  geom_density(alpha = 0.3) +
  ggtitle("Carat vs Cut")
```



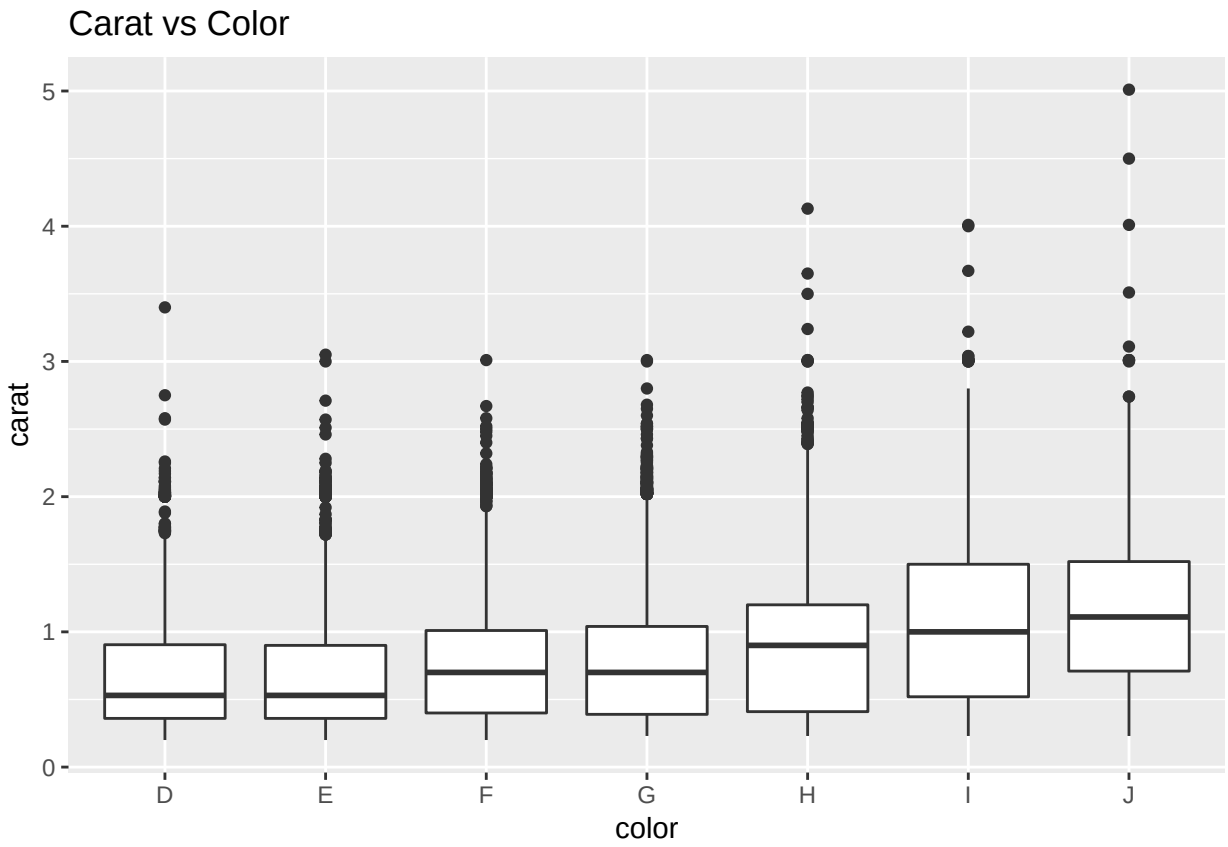

```
ggplot(data = D, mapping = aes(x = cut, y = carat)) +
  geom_boxplot() +
  ggtitle("Carat vs Cut")
```



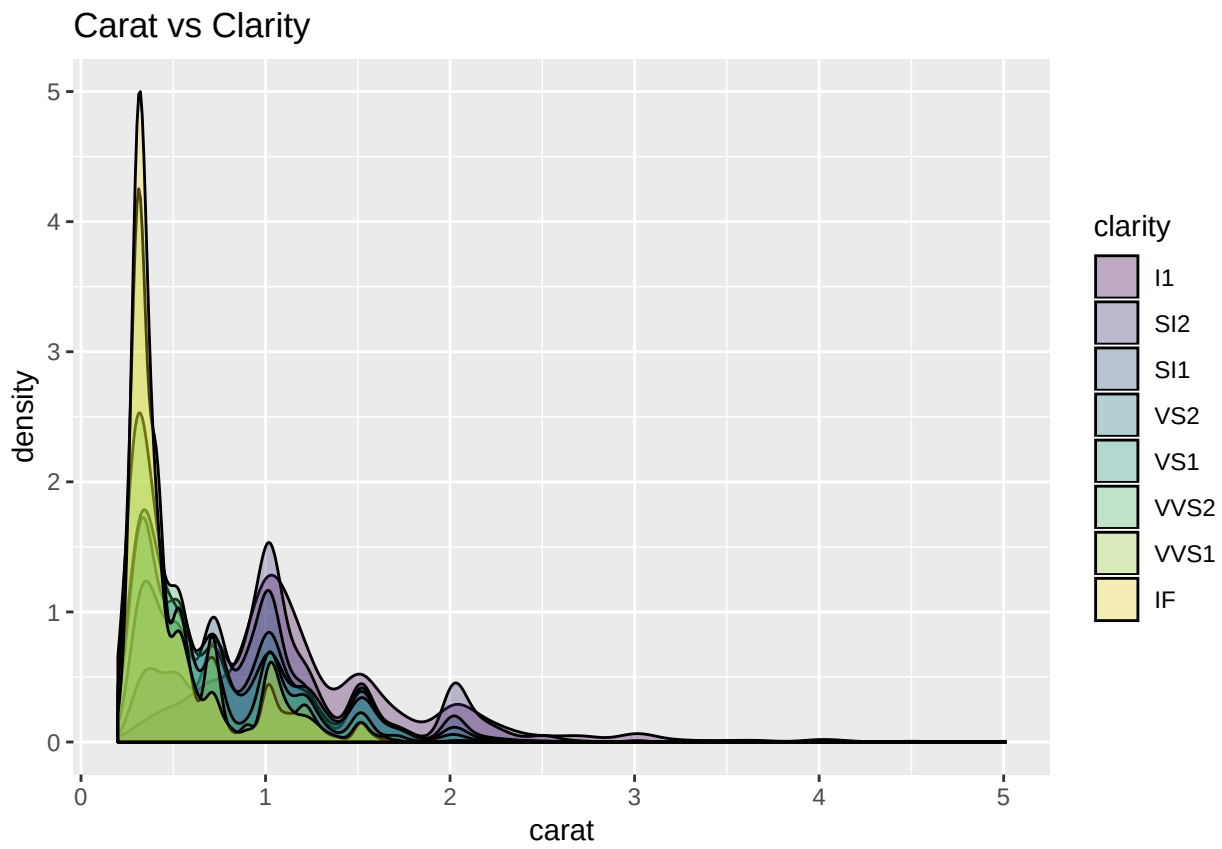

```
# Carat vs Color. Overall, the better the color, the lighter the weight.
ggplot(data = D, mapping = aes(x = carat, fill = color)) +
  geom_density(alpha = 0.3) +
  ggtitle("Carat vs Color")
```



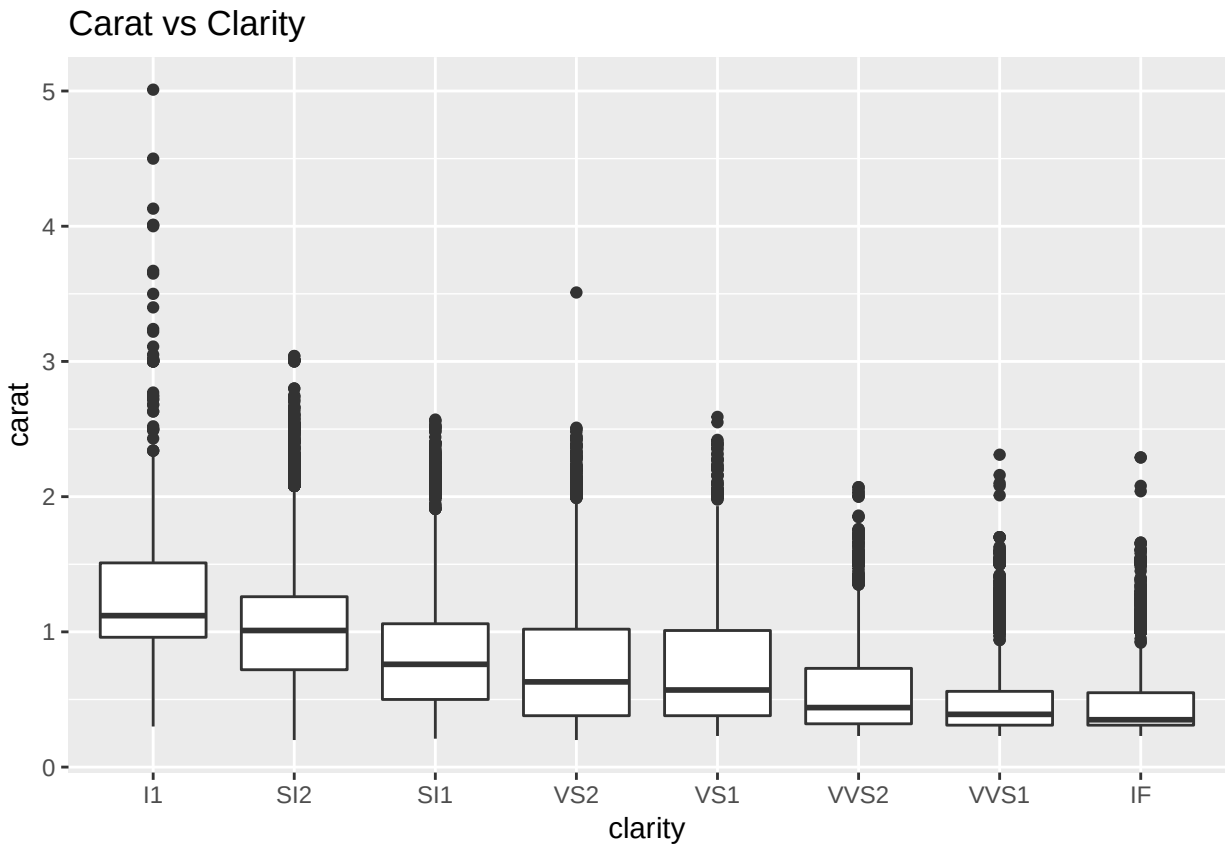

```
ggplot(data = D, mapping = aes(x = color, y = carat)) +
  geom_boxplot() +
  ggtitle("Carat vs Color")
```




```
# Carat vs Clarity. Basically the same to the color, more clear diamonds tend to
# be lighter in weight.
ggplot(data = D, mapping = aes(x = carat, fill = clarity)) +
  geom_density(alpha = 0.3) +
  ggtitle("Carat vs Clarity")
```




```
ggplot(data = D, mapping = aes(x = clarity, y = carat)) +
  geom_boxplot() +
  ggtitle("Carat vs Clarity")
```


Carat vs Clarity
carat
clarity
0
1
2
3
4
5
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF