Analyse diamonds data with ggplot

Code

Hide

```
library(dplyr)
library(tidyverse)
library(patchwork)
library(RColorBrewer)
```

Data Information

Hide

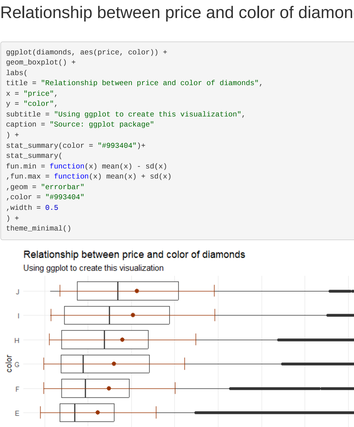
```
glimpse(diamonds)
```

```
Rows: 53,940
Columns: 10
$ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22,…
$ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very …
$ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J,…
$ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, S…
$ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1,…
$ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 5…
$ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339…
$ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87,…
$ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78,…
$ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49,…
```

plot cut vs color

Hide

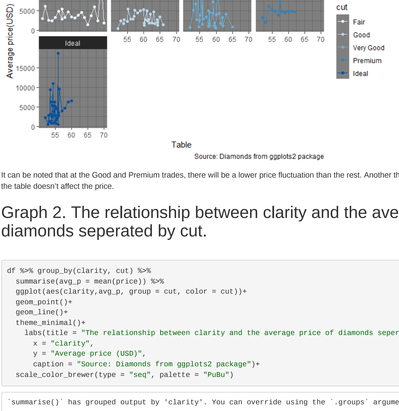
```
ggplot(diamonds, aes(cut,
                     fill=color)) +
  geom_bar(position = "dodge")
```



Relationship between price and color of diamonds with boxplot

Hide

```
ggplot(diamonds, aes(price, color)) +
geom_boxplot() +
labs(
title = "Relationship between price and color of diamonds",
x = "price",
y = "color",
subtitle = "Using ggplot to create this visualization",
caption = "Source: ggplot package"
) +
stat_summary(color = "#993404")+
stat_summary(
fun.min = function(x) mean(x) - sd(x)
,fun.max = function(x) mean(x) + sd(x)
,geom = "errorbar"
,color = "#993404"
,width = 0.5
) +
theme_minimal()
```



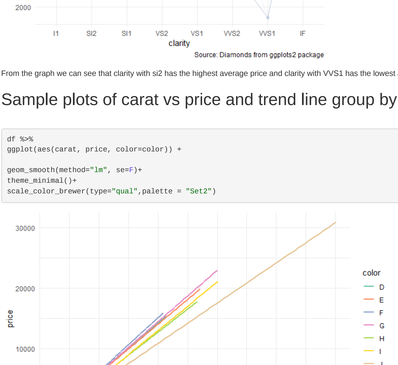
We find that color j has the highest average price and the color j diamonds have the largest interquartile range and they have a median of around 4700, which is the highest. On the other hand, the color j diamonds have the lowest median, which is a have a median of around 3000

The relationship between table and the average price of diamonds separated, by cut

Hide

```
df %>% group_by(table, cut) %>%
 summarise(avg_p = mean(price)) %>%
 ggplot(aes(table, avg_p, group = cut, color = cut)) +
 geom_point()+
 geom_line()+
 theme_dark()+
 facet_wrap(~ cut, ncol = 4)+
 labs(title = "The relationship between table and the average price of diamonds separated, by cut ",
  x = "Table",
  y = "Average price(USD)",
  caption = "Source: Diamonds from ggplots2 package")+
scale_color_brewer(type = "seq", palette = "Blues")
```

```
`summarise()` has grouped output by 'table'. You can override using the `.groups` argument.
```



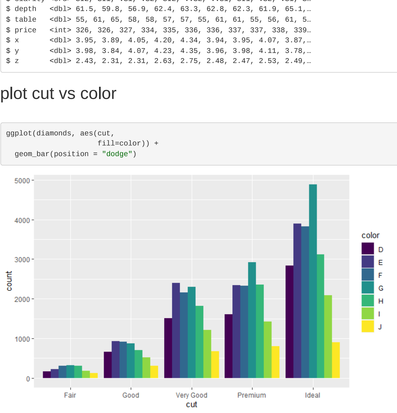
It can be noted that at the Good and Premium trades, there will be a lower price fluctuation than the rest. Another thing to notice is that the size of the table doesn't affect the price.

Graph 2. The relationship between clarity and the average price of diamonds seperated by cut.

Hide

```
df %>% group_by(clarity, cut) %>%
  summarise(avg_p = mean(price)) %>%
  ggplot(aes(clarity,avg_p, group = cut, color = cut))+
  geom_point()+
  geom_line()+
  theme_minimal()+
    labs(title = "The relationship between clarity and the average price of diamonds seperated by cut",
       x = "Clarity",
       y = "Average price (USD)",
      caption = "Source: Diamonds from ggplots2 package")+
  scale_color_brewer(type = "seq", palette = "PuBu")
```

```
`summarise()` has grouped output by 'clarity'. You can override using the `.groups` argument.
```



From the graph we can see that clarity with si2 has the highest average price and clarity with VVS1 has the lowest average price.

Sample plots of carat vs price and trend line group by color

Hide

```
df %>%
ggplot(aes(carat, price, color=color)) +

geom_smooth(method="lm", se=F)+
theme_minimal()+
scale_color_brewer(type="qual",palette = "Set2")
```


From the graph where carat is equal to 3, the colors that get the average price from high to low can be g , e , i, h, j.