Diamonds Dataset Visualization

By Siriwut Thongmak

2023-02-25

What's am I doing?

Explore Dataset `diamonds` and create 5 visualizations.

Explore the data

```
library(tidyverse)
head(diamonds)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

Dataset Description:

1. `carat`: weight of the diamond (0.2–5.01)
2. `cut`: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
3. `color`: diamond color, from J (worst) to D (best)
4. `clarity`: a measure of the number and size of inclusions within the diamond, from I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)
5. `depth`: total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
6. `table`: width of top of diamond relative to widest point (43–95)

There are x, y, z data, so let's calculate the volume

Transforming Data

```
library(tidyverse)
diamonds <- diamonds %>%
  mutate(volume = x * y * z) %>%
  select(-c(x, y, z))
head(diamonds)
```

```
## # A tibble: 6 x 8
##   carat cut       color clarity depth table price volume
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int>  <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326   38.2
## 2  0.21 Premium   E     SI1      59.8    61   326   34.5
## 3  0.23 Good      E     VS1      56.9    65   327   38.1
## 4  0.29 Premium   I     VS2      62.4    58   334   46.7
## 5  0.31 Good      J     SI2      63.3    58   335   51.9
## 6  0.24 Very Good J     VVS2     62.8    57   336   38.7
```

Chart 1 - carat, dept, clarity and color Vs price

```
library(patchwork)
p1 <- ggplot(diamonds %>% sample_frac(0.1),aes(carat,price, col=cut))+
        geom_point(size = 1, alpha=0.5) +
        geom_rug() +
        theme_minimal() +
        labs(title = "Relationship between carat and price",
        x = "Carat",
        y = "Price USD",
        caption = "Datasource: Diamonds") +
        scale_color_viridis_d(direction = 1)

p2 <- ggplot(diamonds %>% sample_frac(0.1),aes(depth,price, col=cut))+
        geom_point(size = 1, alpha=0.5) +
        theme_minimal() +
        labs(title = "Relationship between Depth and price",
        x = "Depth",
        y = "Price USD",
        caption = "Datasource: Diamonds") +
        scale_color_viridis_d(direction = 1)

p3 <- ggplot(diamonds %>% sample_frac(0.1),aes(clarity,price, col=cut))+
        geom_col() +
        theme_minimal() +
        labs(title = "Relationship between Clarity and Price",
        x = "Clarity",
        y = "Price USD",
        caption = "Datasource: Diamonds") +
        scale_color_viridis_d(direction = 1)

p4 <- ggplot(diamonds %>% sample_frac(0.1),aes(color,price, col=cut))+
        geom_col() +
        theme_minimal() +
        labs(title = "Relationship between Color and Price",
        x = "Color",
        y = "Price USD",
        caption = "Datasource: Diamonds") +
        scale_color_viridis_d(direction = 1)

(p1+p2)/p3/p4
```

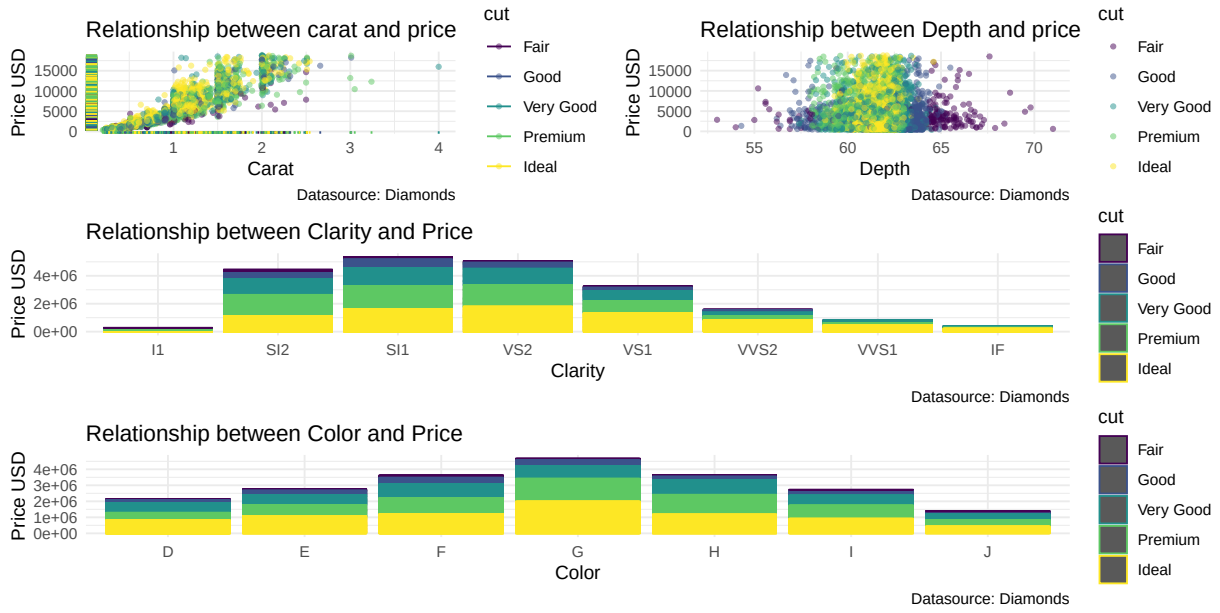


Chart 2 - Count of Cut type

```
ggplot(diamonds, aes(cut, fill = cut))+
  geom_bar(position = "dodge")+
  theme_minimal() +
  labs(title = "Count of Cut type",
       x = "Cut",
       y = "Count",
       caption = "Datasource: Diamonds")
```

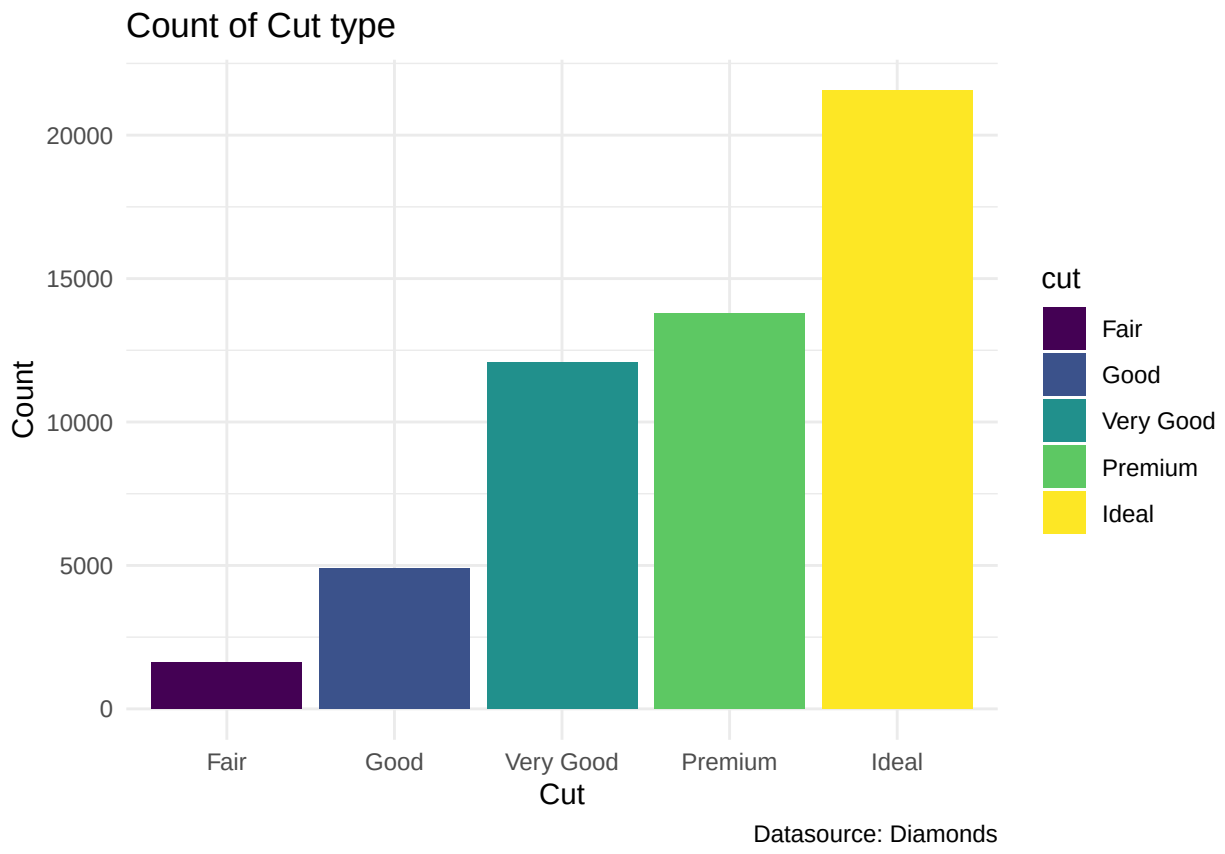


Chart 3 - Density plot of price by color

```
ggplot(diamonds, aes(price, fill = color))+
  geom_density(size = 0, alpha = 0.5, adjust = 2)+
  theme_minimal() +
  labs(title = "Density plot of price by color",
       x = "price",
       caption = "Datasource: Diamonds")
```

Chart 4 - carat vs volume

```
set.seed(42)
ggplot(diamonds %>% sample_n(5000), aes(carat,volume, col = cut))+
  geom_point(size = 3, alpha = 0.3)+
  geom_rug()+
  theme_minimal() +
  labs(title = "carat vs volume",
       x = "carat",
       caption = "Datasource: Diamonds") +
  scale_color_viridis_d(direction = 1)
```

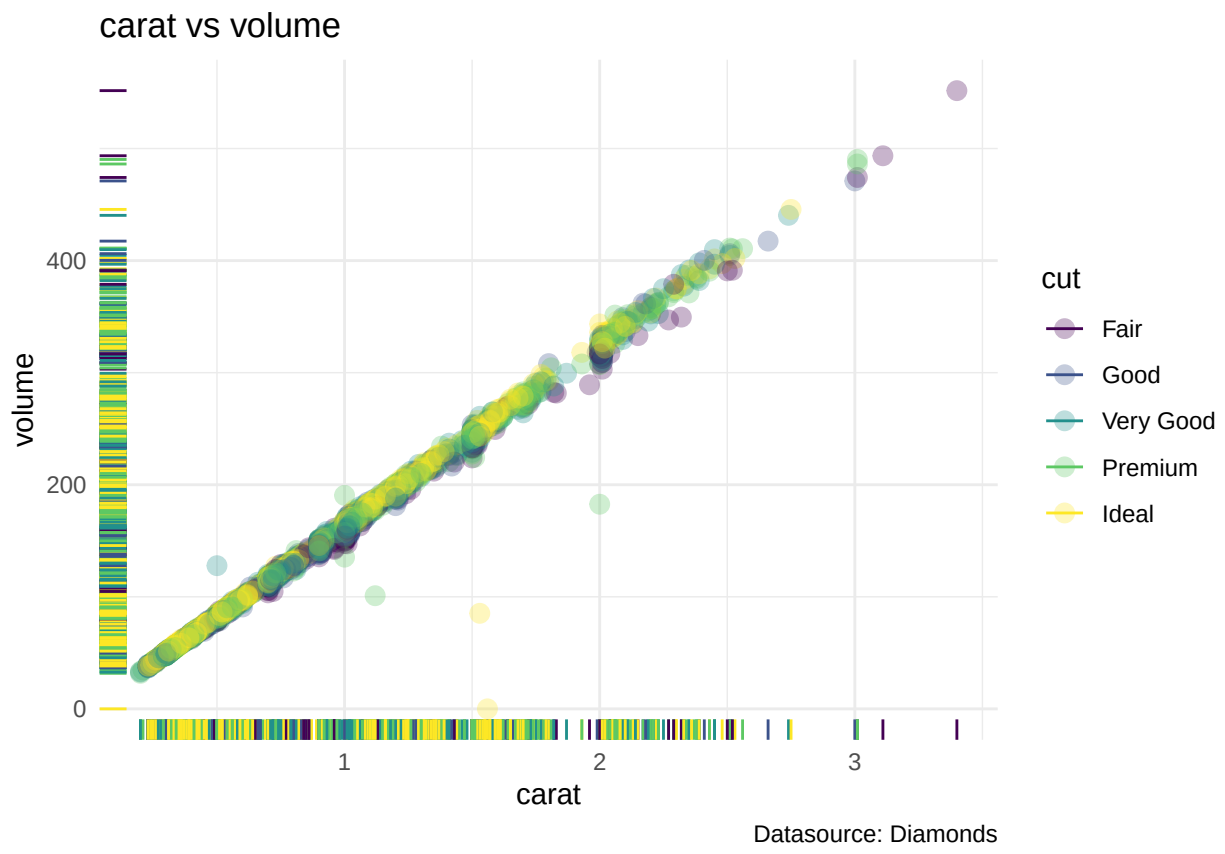
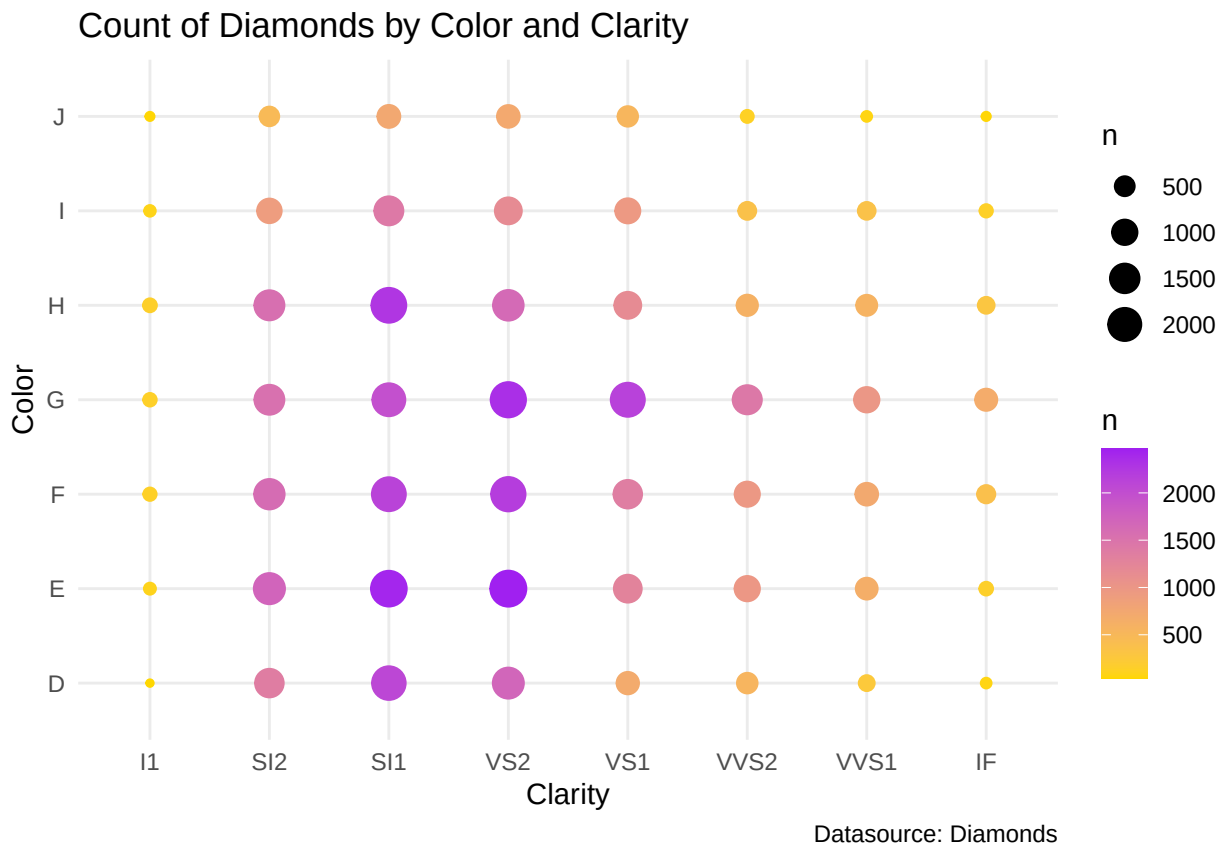


Chart 5 - Count of Diamonds by Color and Clarity

```
ggplot(diamonds, aes(x = clarity, y = color, col = ..n..)) +
  geom_count() +
  theme_minimal()+
  scale_color_gradient(low = "gold",high="purple")+
  labs(title = "Count of Diamonds by Color and Clarity",
       x = "Clarity",
       y = "Color",
       caption = "Datasource: Diamonds")
```


Count of Diamonds by Color and Clarity
Color
J
I
H
G
F
E
D
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
Clarity
n
500
1000
1500
2000
n
2000
1500
1000
500
Datasource: Diamonds