Diamonds Dataset

Laddawan Poonpipat

2024-03-21

Content

- carat: weight of the diamond (0.2–5.01)
- cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- color: diamond colour, from J (worst) to D (best)
- clarity: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- depth total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
- table width of top of diamond relative to widest point (43–95)
- price price in US dollars ($326–$18,823)
- x: length in mm (0–10.74)
- y: width in mm (0–58.9)
- z: depth in mm (0–31.8)

Package:

```
library(rmarkdown)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Head of daimond data frame

```
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
```

Descriptive statistics

```
summary(diamonds)
```

```
##      carat               cut        color        clarity          depth      
##  Min.   :0.2000   Fair     : 1610   D: 6775   SI1    :13065   Min.   :43.00  
##  1st Qu.:0.4000   Good     : 4906   E: 9797   VS2    :12258   1st Qu.:61.00  
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80  
##  Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171   Mean   :61.75  
##  3rd Qu.:1.0400   Ideal    :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50  
##  Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00  
##                                     J: 2808   (Other): 2531                  
##      table           price             x                y         
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000  
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720  
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710  
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735  
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540  
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900  
##                                                                   
##        z         
##  Min.   : 0.000  
##  1st Qu.: 2.910  
##  Median : 3.530  
##  Mean   : 3.539  
##  3rd Qu.: 4.040  
##  Max.   :31.800  
## 
```

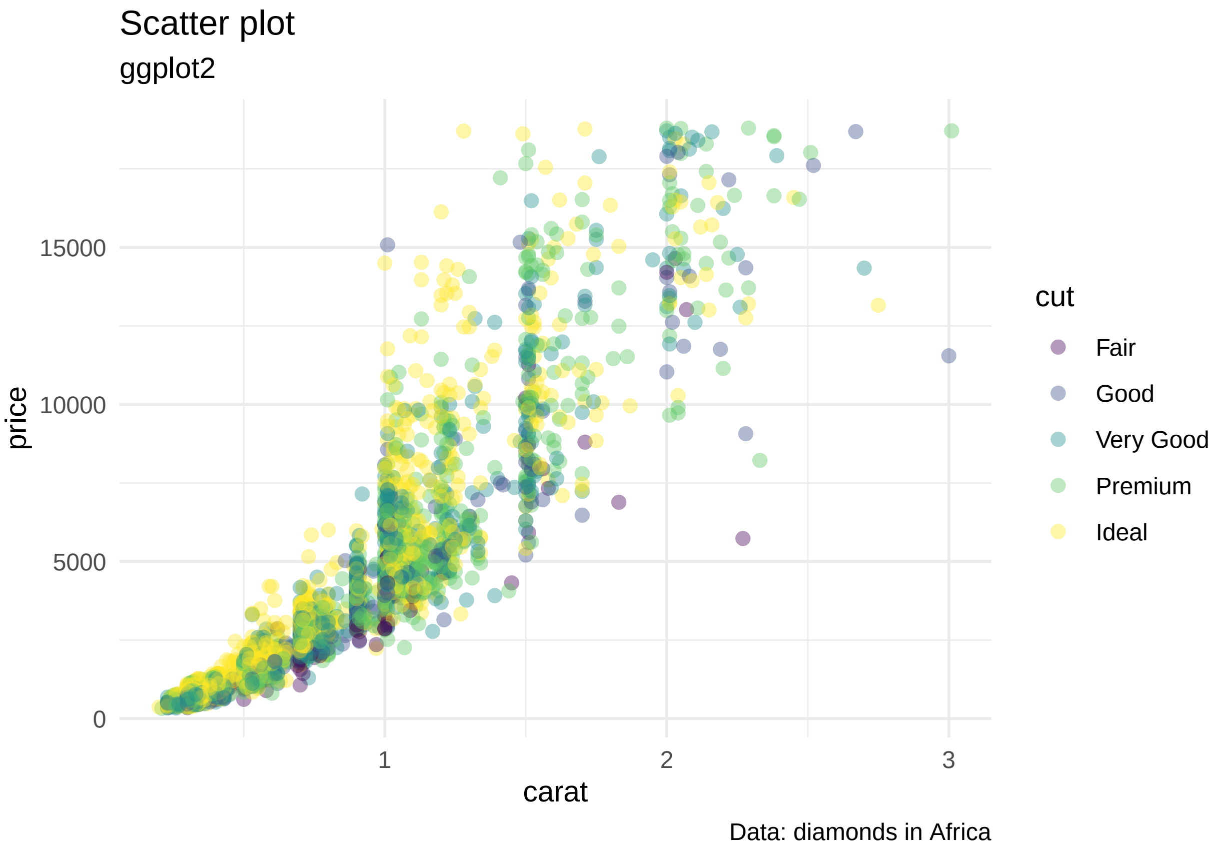
Scatter plot

- carat: weight of the diamond (0.2–5.01)
- price price in US dollars ($326–$18,823)

```
set.seed(13)
ggplot(diamonds %>% sample_n(3000),
       mapping = aes(x=carat, y=price,
                     color = cut))+
  geom_point(alpha=0.4, size= 2)+
  theme_minimal()+

  labs(
    title="Scatter plot",
    subtitle = "ggplot2",
    caption= "Data: diamonds in Africa",

  )
```



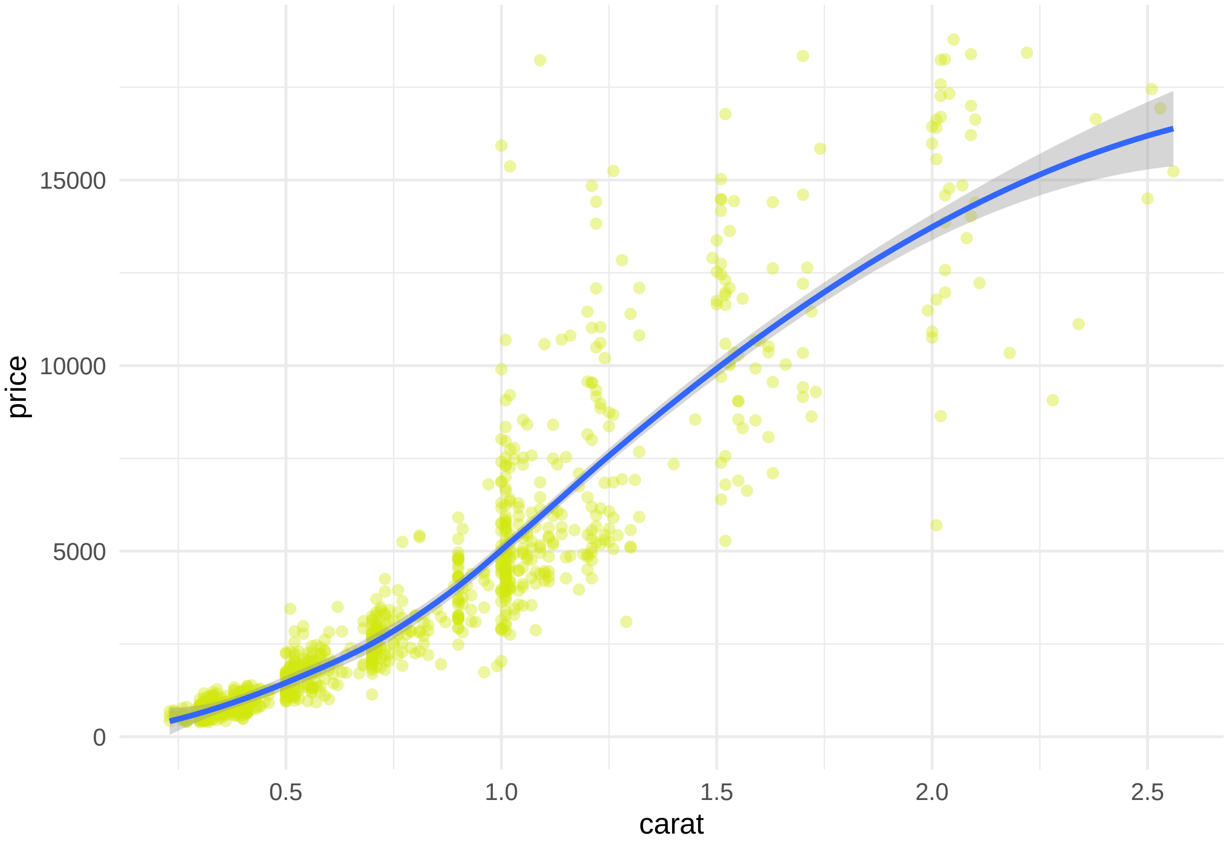
Scatter plot

- geom_smooth()
- carat: weight of the diamond (0.2–5.01)
- price price in US dollars ($326–$18,823)

```
base <- ggplot(diamonds %>%
                 sample_n(1000) %>%
                 filter(carat <= 2.8),
               aes(x=carat, y=price))

base +
  theme_minimal()+
  geom_point(alpha = 0.4, color = "#d2e80e")+
  geom_smooth(method = "loess", se=TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



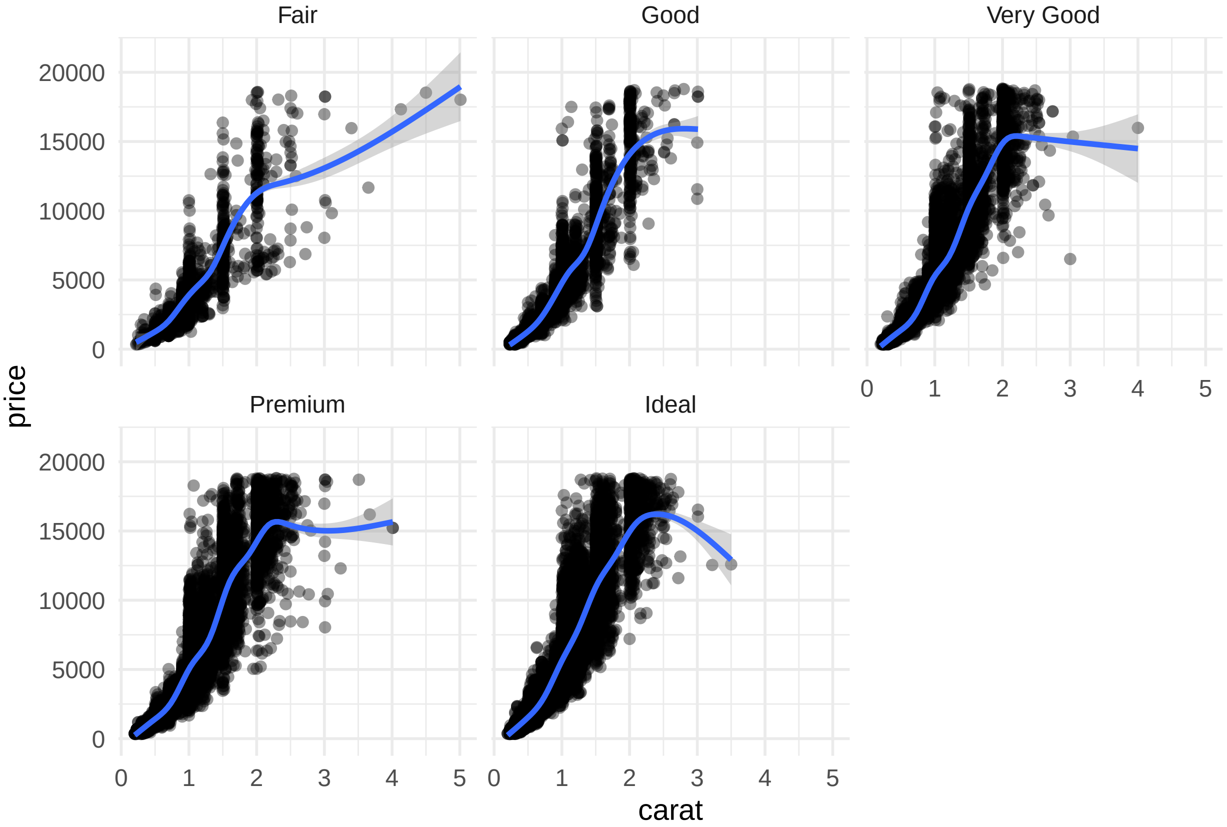
Scatter plot

- acet_wrap()
- quality of the cut (Fair, Good, Very Good, Premium, Ideal)

```
ggplot(diamonds, aes(carat, price))+
  geom_point(alpha=0.4, suze = 0.5)+
  geom_smooth()+
  theme_minimal()+
  facet_wrap(~cut, ncol=3)
```

```
## Warning in geom_point(alpha = 0.4, suze = 0.5): Ignoring unknown parameters:
## `suze`
```

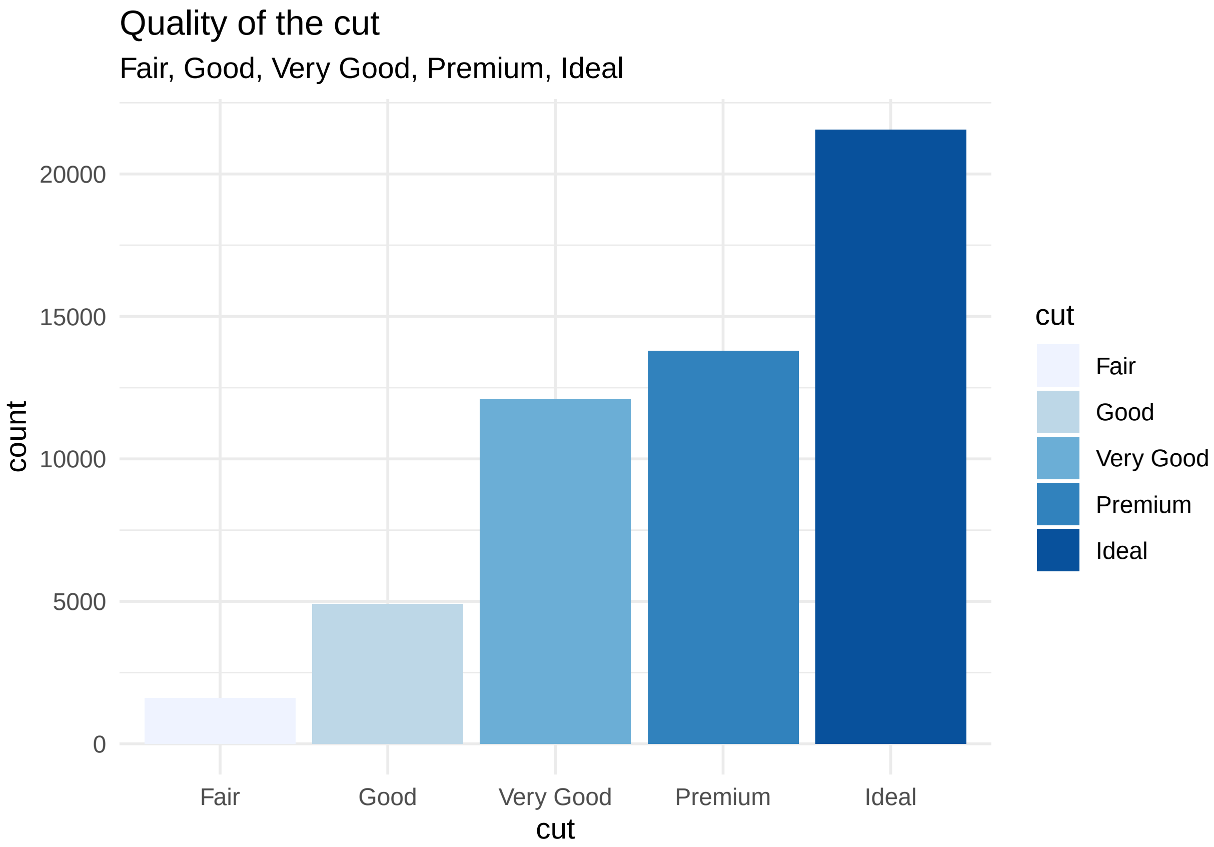
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Bar plot

- quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- price price in US dollars ($326–$18,823)

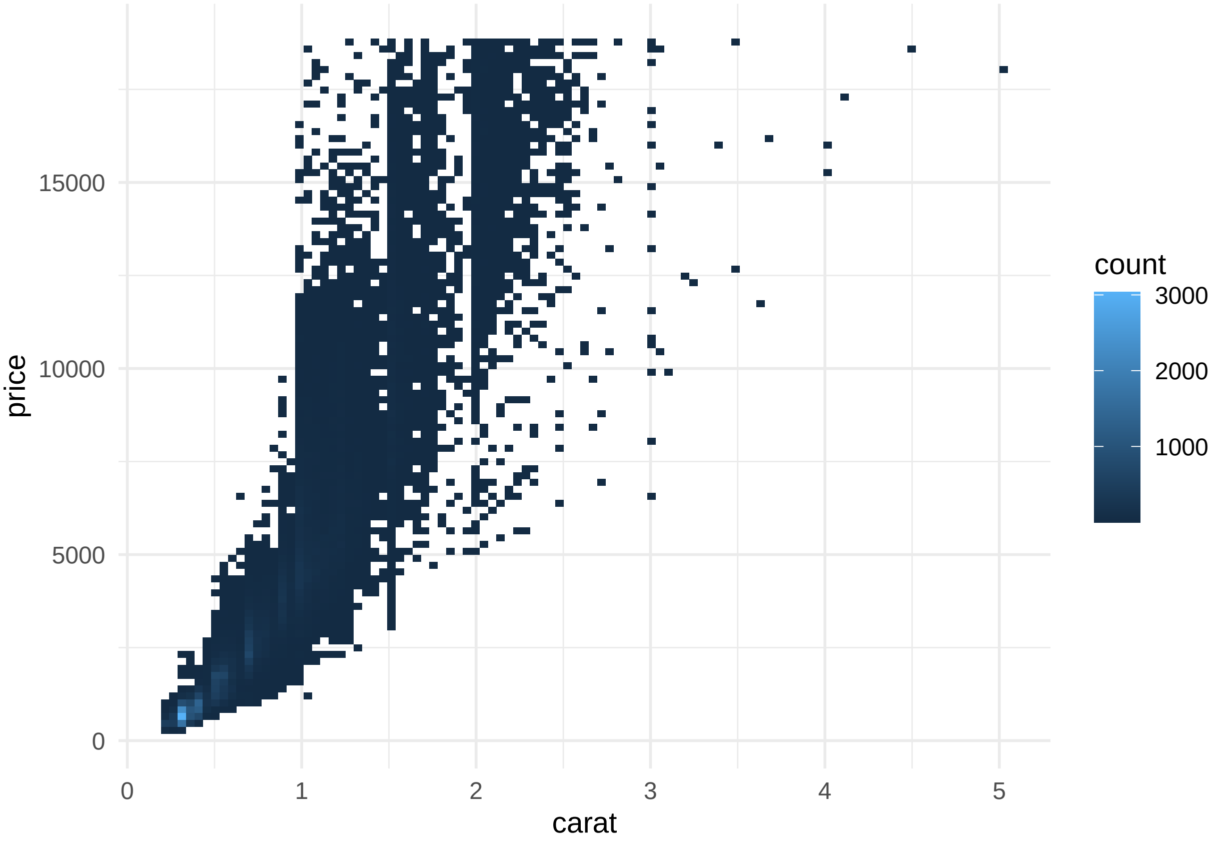
```
ggplot(diamonds, aes(cut, fill=cut))+
  geom_bar()+
  theme_minimal()+
  scale_fill_brewer(palette = "Blues")+
  labs(
    title = "Quality of the cut",
    subtitle = "Fair, Good, Very Good, Premium, Ideal"
  )
```



#

Density plot - quality of the cut (Fair, Good, Very Good, Premium, Ideal) - price price in US dollars ($326–$18,823)

```
ggplot(diamonds,
       aes(carat, price)) +
  geom_bin2d(bins = 100)+
  theme_minimal()
```


count
3000
2000
1000
price
15000
10000
5000
0
0
1
2
3
4
5
carat