R_Markdown_Practice

Naphon

2024-09-04

Download package

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'tibble' was built under R version 4.3.3
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

```
## Warning: package 'purrr' was built under R version 4.3.3
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
## Warning: package 'stringr' was built under R version 4.3.3
```

```
## Warning: package 'forcats' was built under R version 4.3.3
```

```
## Warning: package 'lubridate' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Check data

Using diamonds dataset

```
head(diamonds)
```

```
## # A tibble: 6 × 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
```

```
tail(diamonds)
```

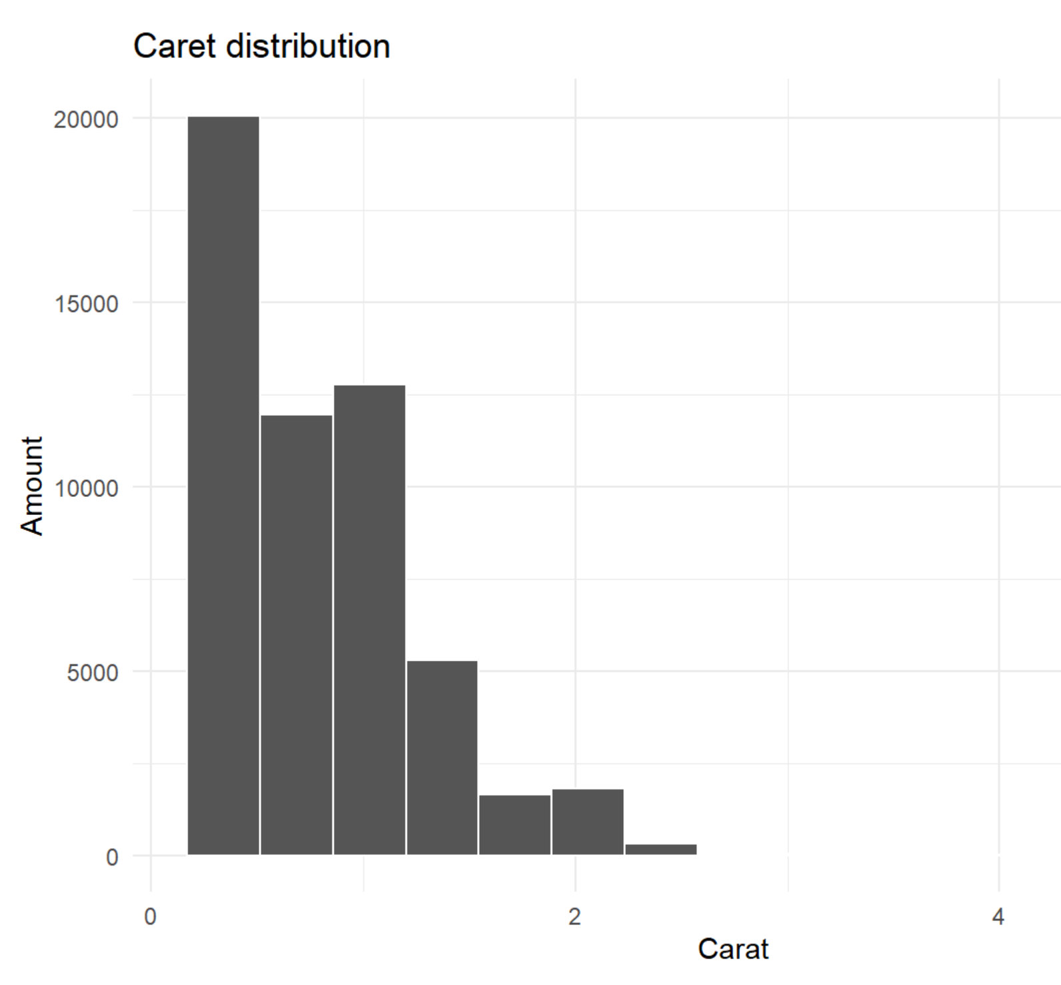
```
## # A tibble: 6 × 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.72 Premium   D     SI1      62.7    59  2757  5.69  5.73  3.58
## 2  0.72 Ideal     D     SI1      60.8    57  2757  5.75  5.76  3.5
## 3  0.72 Good      D     SI1      63.1    55  2757  5.69  5.75  3.61
## 4  0.7  Very Good D     SI1      62.8    60  2757  5.66  5.68  3.56
## 5  0.86 Premium   H     SI2      61      58  2757  6.15  6.12  3.74
## 6  0.75 Ideal     D     SI2      62.2    55  2757  5.83  5.87  3.64
```

```
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, 0.23, 0.…
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver…
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,…
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS1, …
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4, 64…
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62, 58…
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340, 34…
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00, 4.…
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05, 4.…
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39, 2.…
```

Histogram (01)

```
ggplot(diamonds, aes(carat)) +
  geom_histogram(bins=15, col="white")+
  theme_minimal()+
  labs(title = "Caret distribution",
       x = "Carat",
       y = "Amount")
```



The histogram chart (01) above shows positive skewness.

```
diamonds$carat %>%
  mean()
```

```
## [1] 0.7979397
```

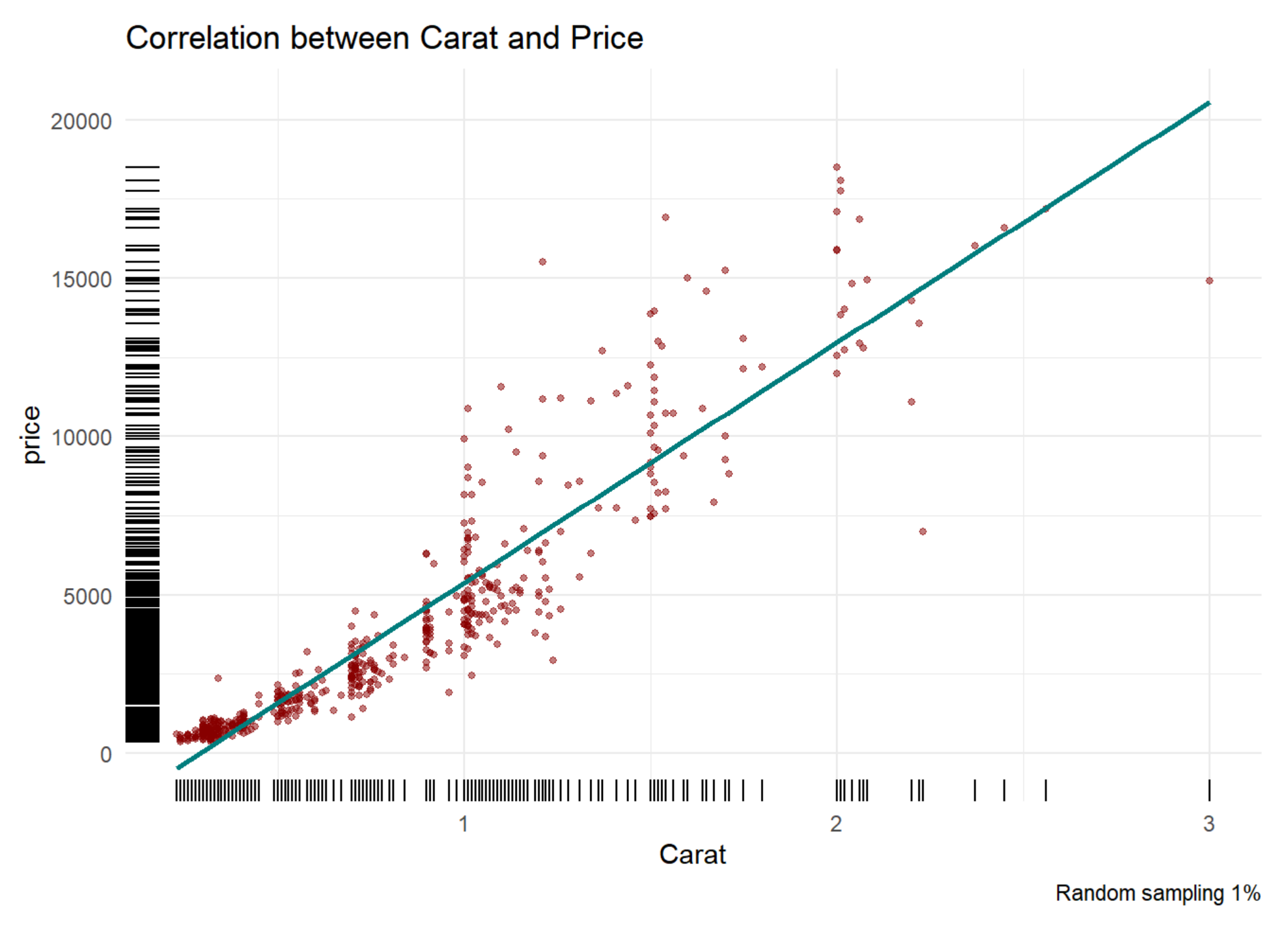
```
diamonds$carat %>%
  median()
```

```
## [1] 0.7
```

Scatter plot (02)

```
set.seed(42)
ggplot(diamonds %>% sample_frac(0.01), aes(carat, price))+
  geom_point(size = 1, col="#8B0000", alpha=0.5)+
  geom_smooth(col="#008080",method = "lm", se=FALSE)+
  geom_rug()+
  theme_minimal()+
  labs(title = "Correlation between Carat and Price",
       caption = "Random sampling 1%",
       x = "Carat",
       y = "price")
```

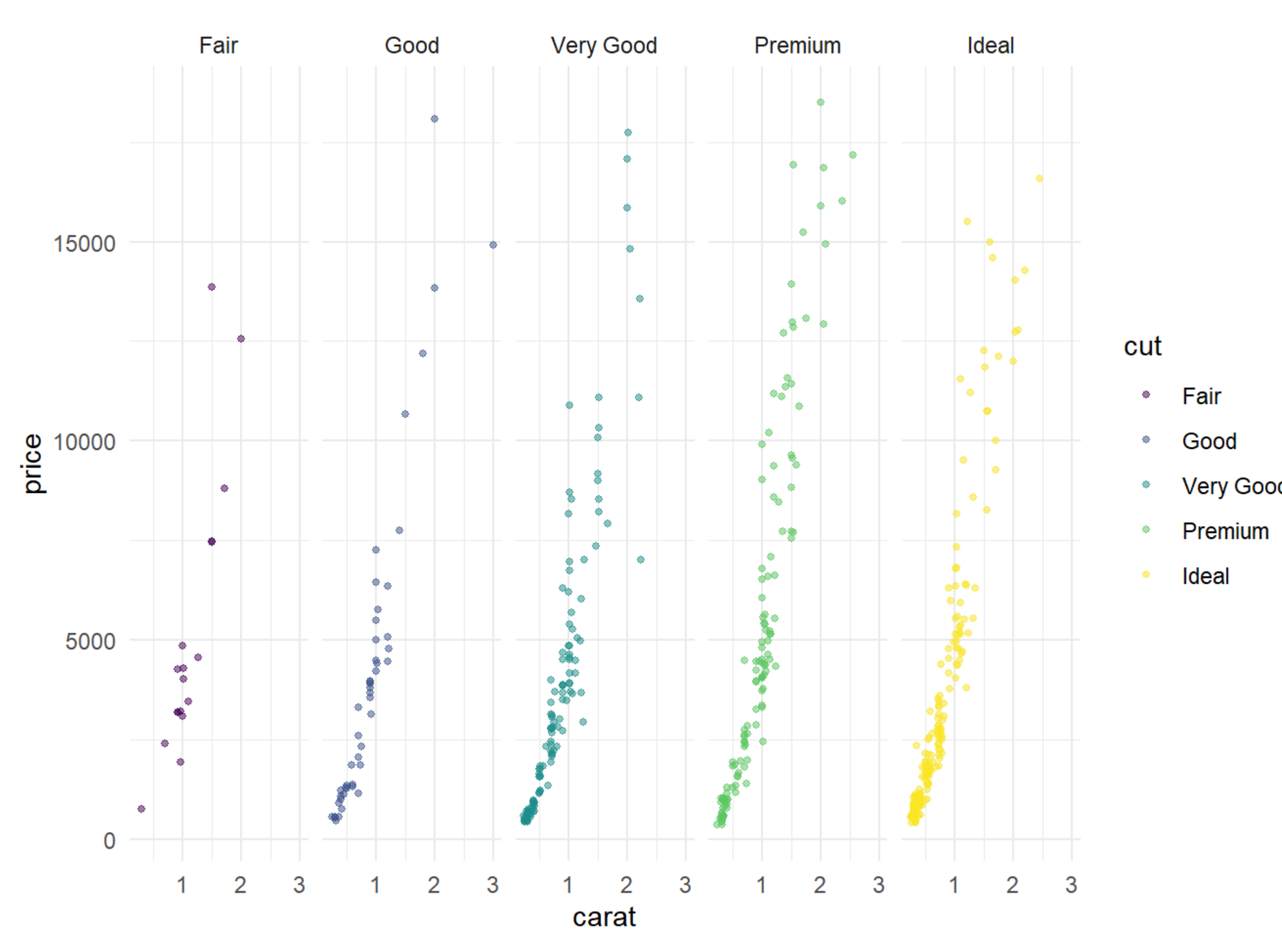
```
## `geom_smooth()` using formula = 'y ~ x'
```



The chart (02) above illustrates positive correlation between Carat and Price. In other words, the value of a diamond rises as its carat weight increases.

Scatter plots for each cut category (03)

```
set.seed(42)
ggplot(diamonds %>% sample_frac(0.01), aes(carat, price, col=cut))+
  geom_point(size = 1, alpha=0.5)+
  theme_minimal()+
  facet_wrap(~cut, nrow = 1)
```



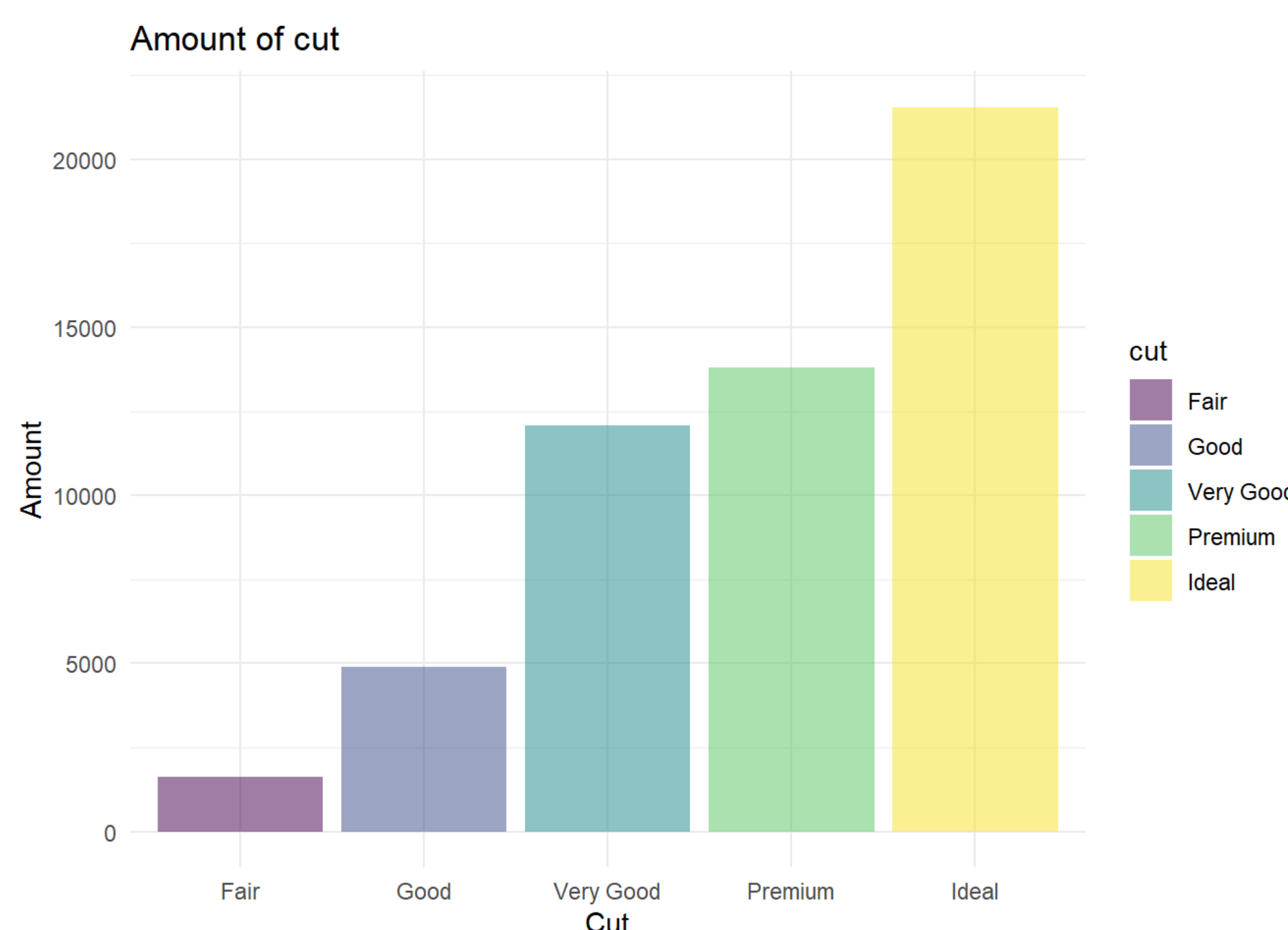

```
  labs(title = "Correlation between Carat and Price",
       caption = "Random sampling 1%",
       x = "Carat",
       y = "price")
```

```
## $x
## [1] "Carat"
##
## $y
## [1] "price"
##
## $title
## [1] "Correlation between Carat and Price"
##
## $caption
## [1] "Random sampling 1%"
##
## attr(,"class")
## [1] "labels"
```

```
diamonds %>%
  count(cut) %>%
  arrange(desc(n))
```

```
## # A tibble: 5 × 2
##   cut           n
##   <ord>     <int>
## 1 Ideal     21551
## 2 Premium   13791
## 3 Very Good 12082
## 4 Good       4906
## 5 Fair       1610
```

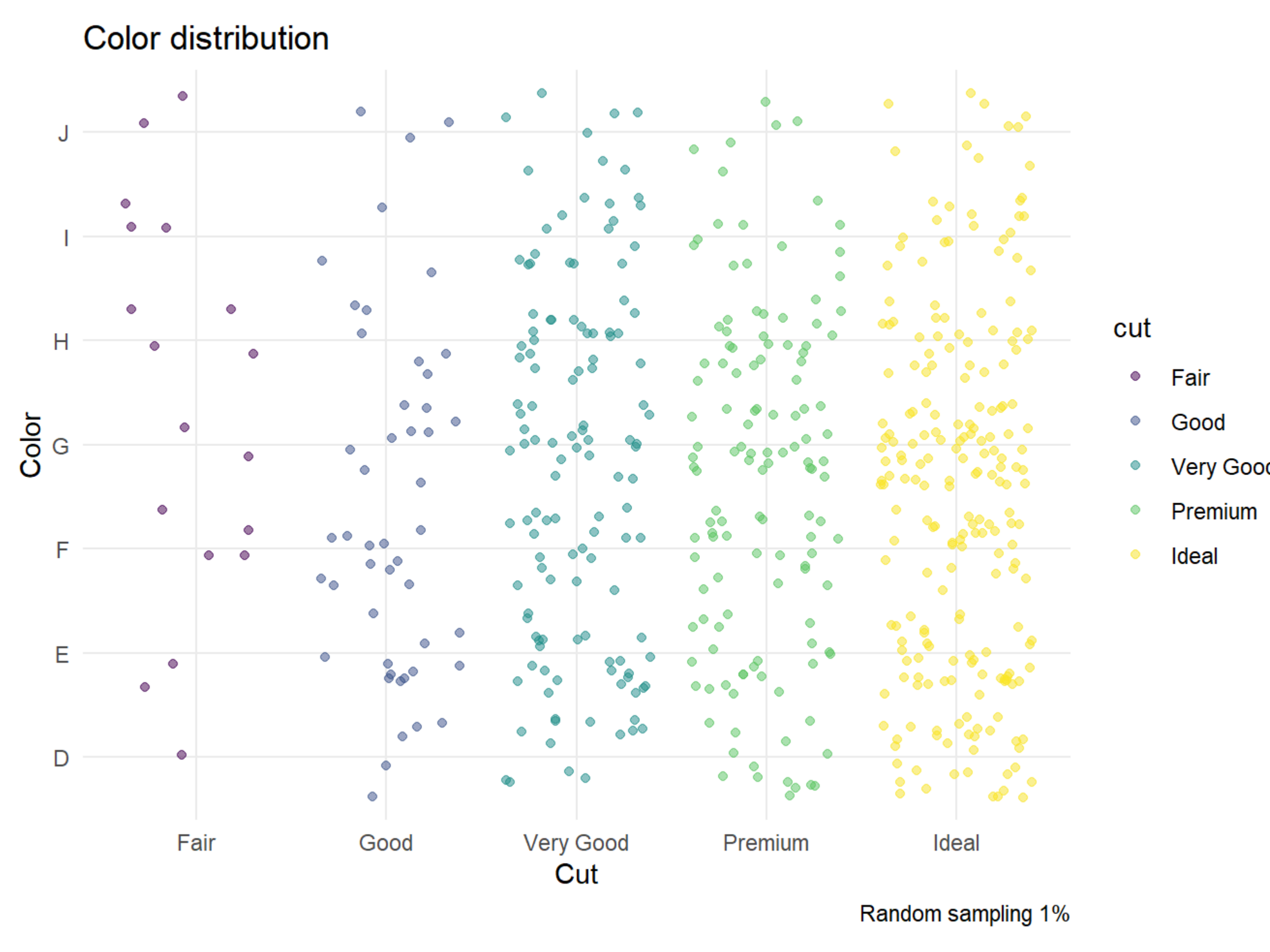
```
ggplot(diamonds, aes(cut, fill=cut)) +
  geom_bar(alpha =0.5)+
  theme_minimal()+
  labs(title = "Amount of cut",
       x = "Cut",
       y = "Amount")
```



There are positive correlations between Carat and Price in each scatter plots (03) above, while both the table and bar chart displays the diamond cuts in an ordinal sequence from 'Ideal' to 'Fair,' providing a clear comparison of the quality rankings.

Jitter plot (04)

```
set.seed(42)
ggplot(diamonds %>% sample_frac(0.01), aes(cut, color, col=cut))+
  geom_jitter(alpha=0.5)+
  theme_minimal()+
  labs(title = "Color distribution",
       caption = "Random sampling 1%",
       x = "Cut",
       y = "Color")
```



The jitter plot (04) above reveals how diamond colors are distributed across various cut categories, highlighting any trends or variations in color quality within each cut type.