R Notebook

Loading Libraries

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

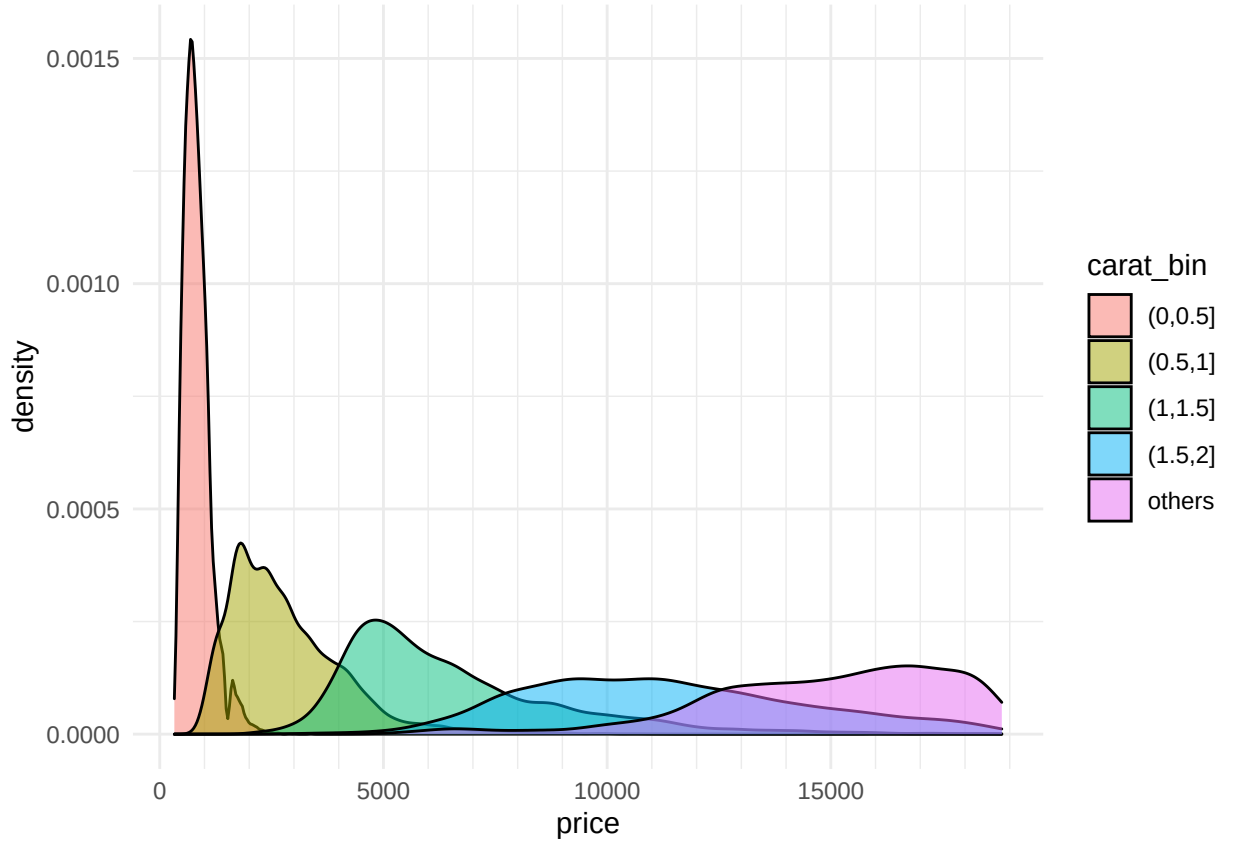
```
library(patchwork)

set.seed(7)
sample_diam <- sample_n(diamonds, 5000)
```

Visualizing The Data

This dataset consists of diamonds with the price less than $19,000. The more carat, the more diversity in diamond prices. For example, the price of diamonds with less than 0.5 carat tend to cluster around $1,000 regardless of other properties while the price of diamonds with 1.5 - 2 carat is spread around $5,000 - $19,000.

```
diamonds %>%
  mutate(carat_bin = cut(carat, seq(0, 2, 0.5))) %>%
  mutate(carat_bin = replace_na(fct_expand(carat_bin, 'others'), 'others')) %>%
  ggplot(aes(price, fill = carat_bin)) +
  geom_density(alpha = .5) +
  scale_x_continuous(minor_breaks = function(x) seq(0, max(x), 1000)) +
  theme_minimal()
```



Diamond Color

- the D, E, F, and G colors have the highest price growth as the carat increases despite their low median price
- the H and I colors have moderate effect on the price
- the J color, even though it has the highest median price, it has the least effect on diamond price supported by the fact that the prices of diamonds with this color are grouped around $2,500 - $7,500 with a few outliers in the box plot.

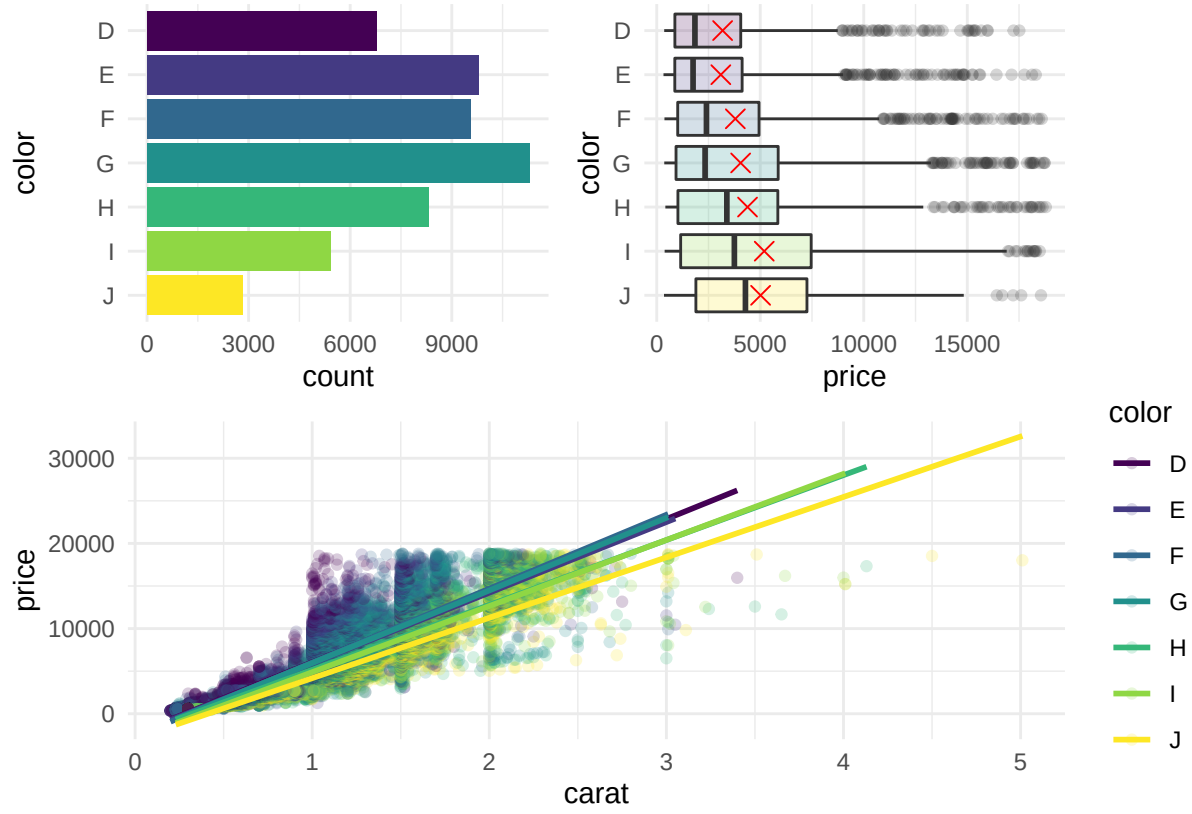
```
p1 <- ggplot(diamonds, aes(y = color, fill = color)) +
  geom_bar() +
  guides(fill="none") +
  scale_y_discrete(limits = rev) +
  theme_minimal()

p2 <- ggplot(sample_diam, aes(y = color, price, fill = color)) +
  geom_boxplot(alpha = .2) +
  stat_summary(fun = mean, shape = 4, size = 3, color = 'red', geom = 'point') +
  guides(fill="none") +
  scale_y_discrete(limits = rev) +
  theme_minimal()

p3 <- ggplot(diamonds, aes(carat, price, color = color)) +
  geom_point(alpha = .2) +
  geom_smooth(method = 'lm', se = FALSE) +
```

```
  theme_minimal()

(p1 + p2) / p3
```



Diamond Cut Quality

- The ideal cut quality improves diamond price by the most reflected through the steepest slope in the carat-price plot
- The premium and very good cut quality has almost the same effect on the price, slightly less than the ideal quality. Also, the good cut quality still has a moderate effect on diamond price
- The fair cut quality has a significantly lower price improvement than other cut qualities

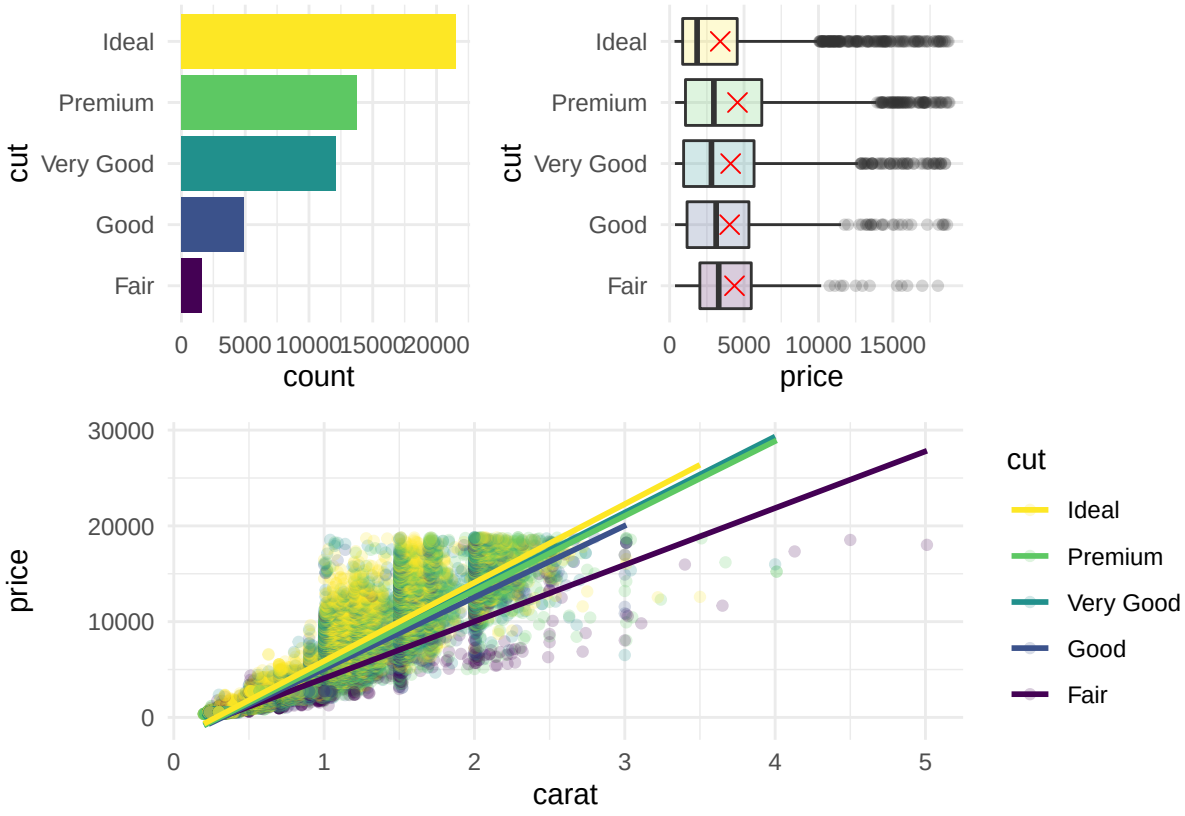
```
p1 <- ggplot(diamonds, aes(y = cut, fill = cut)) +
  geom_bar() +
  guides(fill="none") +
  theme_minimal()

p2 <- ggplot(sample_diam, aes(y = cut, price, fill = cut)) +
  geom_boxplot(alpha = .2) +
  stat_summary(fun = mean, shape = 4, size = 3, color = 'red', geom = 'point') +
  guides(fill="none") +
  theme_minimal()

p3 <- ggplot(diamonds, aes(carat, price, color = cut)) +
```

```
  geom_point(alpha = .2) +
  geom_smooth(method = 'lm', se = FALSE) +
  guides(color = guide_legend(reverse = TRUE)) +
  theme_minimal()

(p1 + p2) / p3
```



Diamond Clarity

- The effect of each diamond clarity on diamond price improvement can be clearly distinguished in the following order, from the best to the worst: IF > VVS1 > VVS2 > VS1 > VS2 > SI1 > SI2 > I1

```
p1 <- ggplot(diamonds, aes(y = clarity, fill = clarity)) +
  geom_bar() +
  guides(fill="none") +
  theme_minimal()

p2 <- ggplot(sample_diam, aes(y = clarity, price, fill = clarity)) +
  geom_boxplot(alpha = .2) +
  stat_summary(fun = mean, shape = 4, size = 3, color = 'red', geom = 'point') +
  guides(fill="none") +
  theme_minimal()

p3 <- ggplot(diamonds, aes(carat, price, color = clarity)) +
```

```
  geom_point(alpha = .2) +
  geom_smooth(method = 'lm', se = FALSE) +
  guides(color = guide_legend(reverse = TRUE)) +
  theme_minimal()

(p1 + p2) / p3
```