R Notebook

```
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)
library(scales)
```

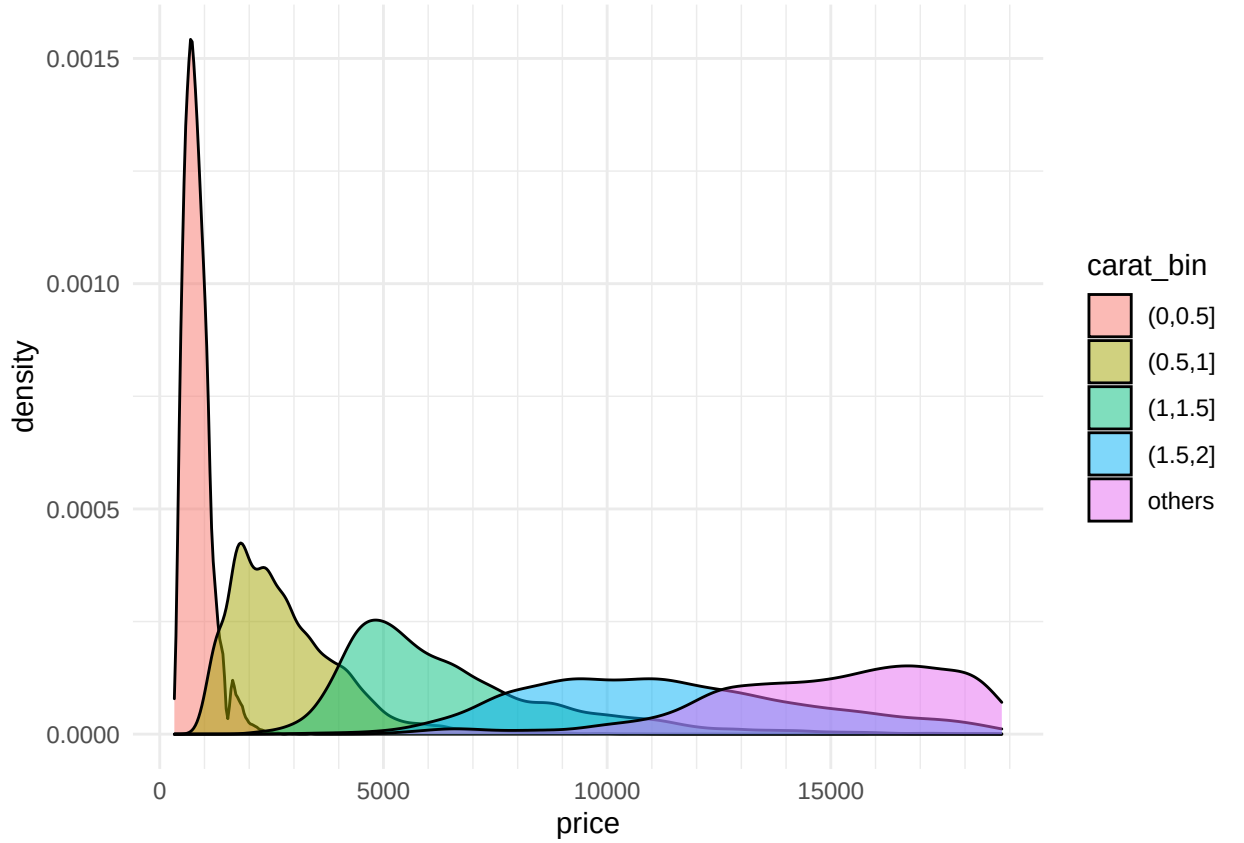
```
## Warning: package 'scales' was built under R version 4.2.2
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

Visualizing The Data

This dataset consists of diamonds with prices less than $19,000. The density plot below shows that diamond prices become more diverse as the carat increases. For example, when the weight of diamonds is less than 0.5 carat, their price tends to cluster around $1,000 regardless of other properties. On the other hand, when the weight of diamonds is between 1.5 and 2 carat, their price range is widened to 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()
```



Since diamond prices start to disperse when their weight is more than 1 carat, the data should be split into two groups before analyzing the effects of each property on a diamond price.

```
diamonds <- diamonds %>% mutate(group = if_else(carat <= 1, '<= 1 carat', '> 1 carat'))
```

The following sections explore the effect of diamond color, cut, and clarity on their price. Each property has three main plots:

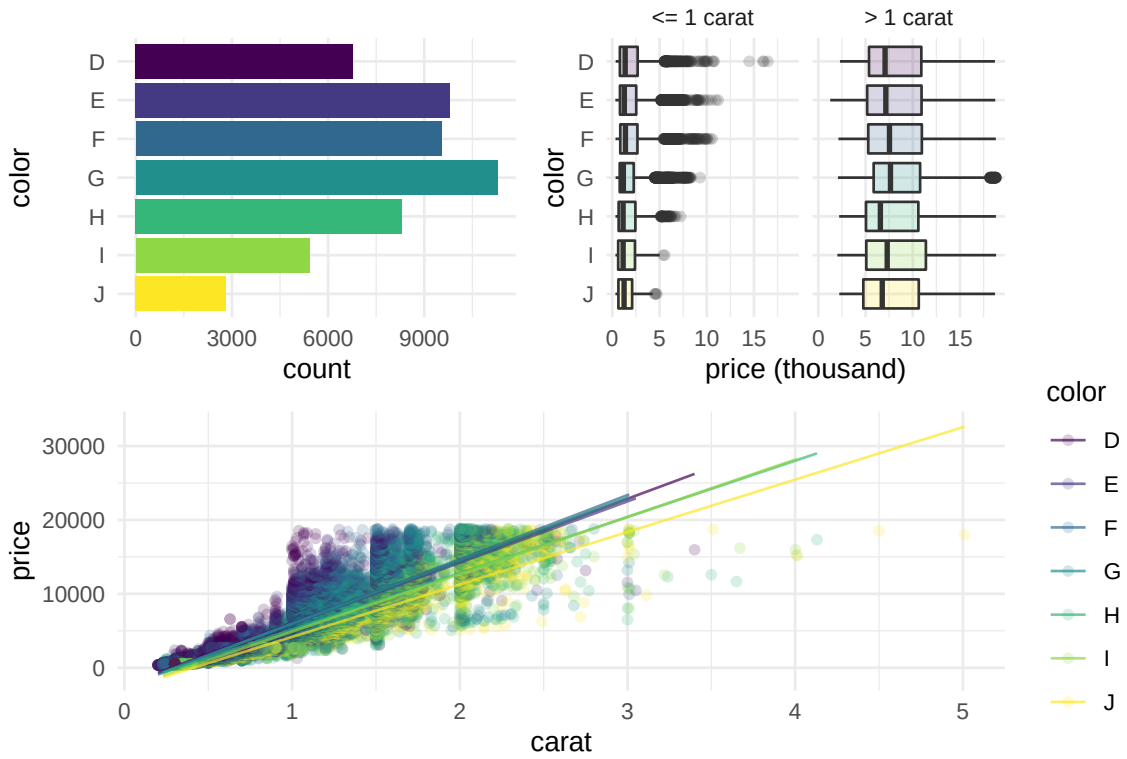
- bar plot shows the distribution of a property across the entire dataset
- box plots show the price distribution of a property for each diamond group
- scatter plot with linear regression lines shows the relation between carat and price for each property's value

Diamond Color

- For diamonds with equal or less than 1 carat, regardless of their color, their median prices are about the same at $1,500, and approximately a third-fourth of them are worth less than $2,500. However, although uncommon, diamonds with colors such as D, E, and F can be significantly more expensive than the others.
- For diamonds with more than 1 carat, the average price of each color is slightly different from each other, falling in the range of roughly $8,000 - $8,500. Color G is the most expensive, while colors H and J are marginally cheaper than the others.
- From the scatter plot, the price growth of each diamond color as the carat increases can be divided into three groups:

  - high: D, E, F, G
  - moderate: H, I
  - low: J

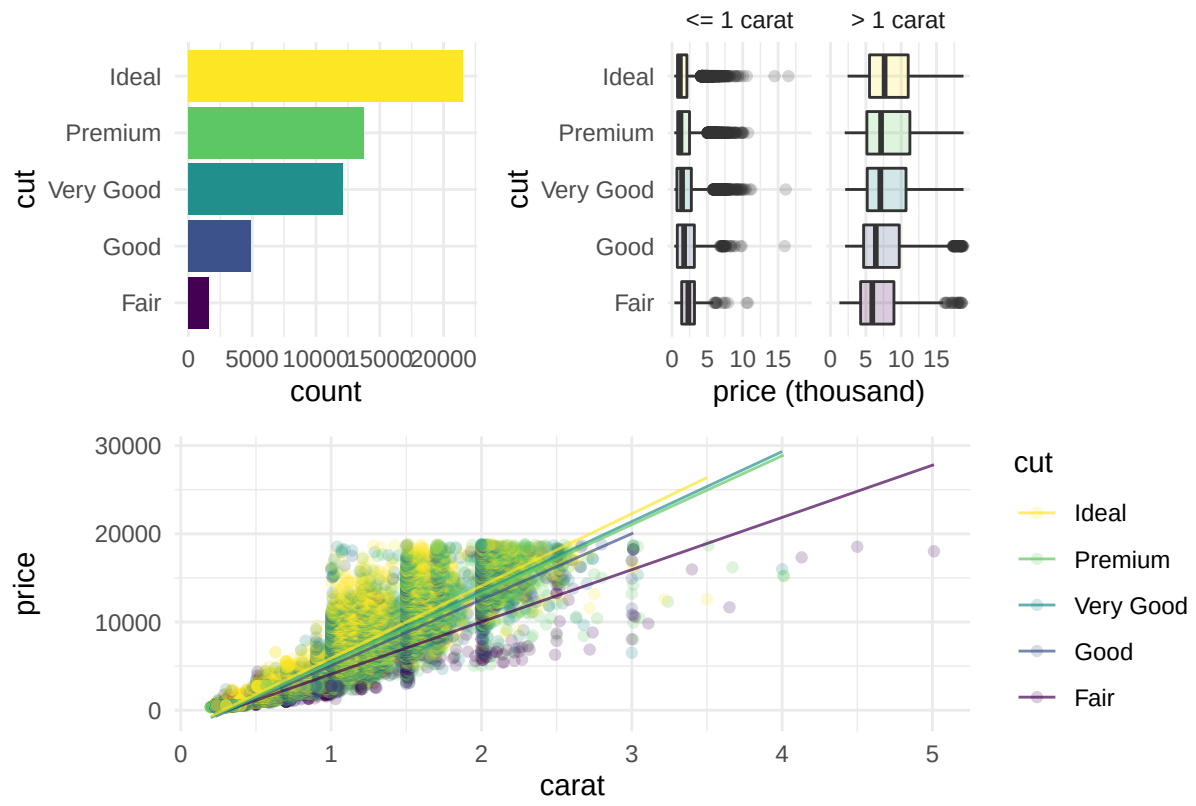
```
bar_plot <- list(geom_bar(),
  guides(fill="none"),
  theme_minimal())

box_plot_by_group <- list(geom_boxplot(alpha = .2),
  guides(fill="none"),
  scale_x_continuous(labels = label_number(scale = .001)),
  labs(x = 'price (thousand)'),
  facet_wrap(~ group),
  theme_minimal())

point_plot <- list(geom_point(alpha = .2),
  geom_line(stat='smooth', method = 'lm', alpha = .7),
  theme_minimal())

p1 <- ggplot(diamonds, aes(y = color, fill = color)) +
  bar_plot +
  scale_y_discrete(limits = rev)

p2 <- ggplot(diamonds, aes(y = color, price, fill = color)) +
  box_plot_by_group +
  scale_y_discrete(limits = rev)

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

(p1 + p2) / p3
```



Diamond Cut Quality

- For diamonds with equal or less than 1 carat, the median diamond price of each cut quality is in the following order, from the most expensive to the cheapest: Fair > Good > Very Good > Premium > Ideal.
- For diamonds with more than 1 carat, the price order is reversed, with the ideal cut being the most expensive and the fair cut being the cheapest.
- From the scatter plot, the price growth of each diamond cut quality as the carat increases is in the following order, from fastest to slowest: Ideal > Premium & Very Good > Good > Fair.

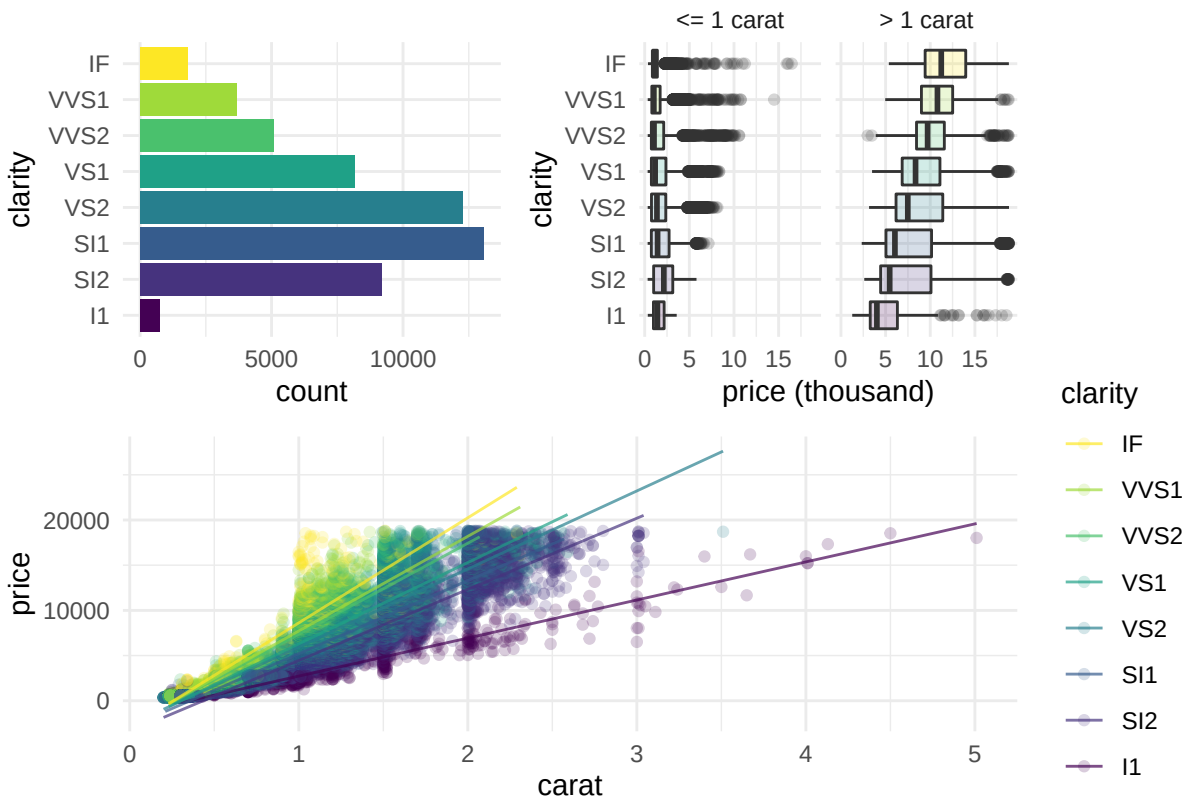
```
p1 <- ggplot(diamonds, aes(y = cut, fill = cut)) +
  bar_plot

p2 <- ggplot(diamonds, aes(y = cut, price, fill = cut)) +
  box_plot_by_group

p3 <- ggplot(diamonds, aes(carat, price, color = cut)) +
  point_plot +
  guides(color = guide_legend(reverse=TRUE))

(p1 + p2) / p3
```



Diamond Clarity

- For diamonds with equal or less than 1 carat, the SI2 clarity has the highest median price at about $2,000. Meanwhile, the VVS2, VVS1, and IF clarity has the lowest median price at about $1,000.
- For diamonds with more than 1 carat, the order of the median price of each clarity is quite clear, from the most expensive to the cheapest: IF > VVS1 > VVS2 > VS1 > VS2 > SI1 > SI2 > I1.
- From the scatter plot, as the carat increases, diamonds with the IF clarity have the fastest price growth, and diamonds with the I1 clarity have the slowest price growth.

```
p1 <- ggplot(diamonds, aes(y = clarity, fill = clarity)) +
  bar_plot

p2 <- ggplot(diamonds, aes(y = clarity, price, fill = clarity)) +
  box_plot_by_group

p3 <- ggplot(diamonds, aes(carat, price, color = clarity)) +
  point_plot +
  guides(color = guide_legend(reverse=TRUE))

(p1 + p2) / p3
```


<= 1 carat
> 1 carat
clarity
IF
VVS1
VVS2
VS1
VS2
SI1
SI2
I1
0
5000
10000
count
0
5
10
15
price (thousand)
price
20000
10000
0
1
2
3
4
5
carat