DSB 11 - Data Viz Homework

Moon Pantakarn

2025-06-28

Homework: My ggplot2 Project

1. Distribution of Diamond Carat:

```
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   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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 error
```

```
library(dplyr)
hist_diam <- diamonds %>%
  sample_n(5000)

ggplot(data = hist_diam,
       mapping = aes(carat)) +
  geom_histogram(bins = 10) +
  labs(title = "Distribution of Diamond Carat",
       x = "Carat Weight",
       y = "Count") +
  theme_minimal()
```

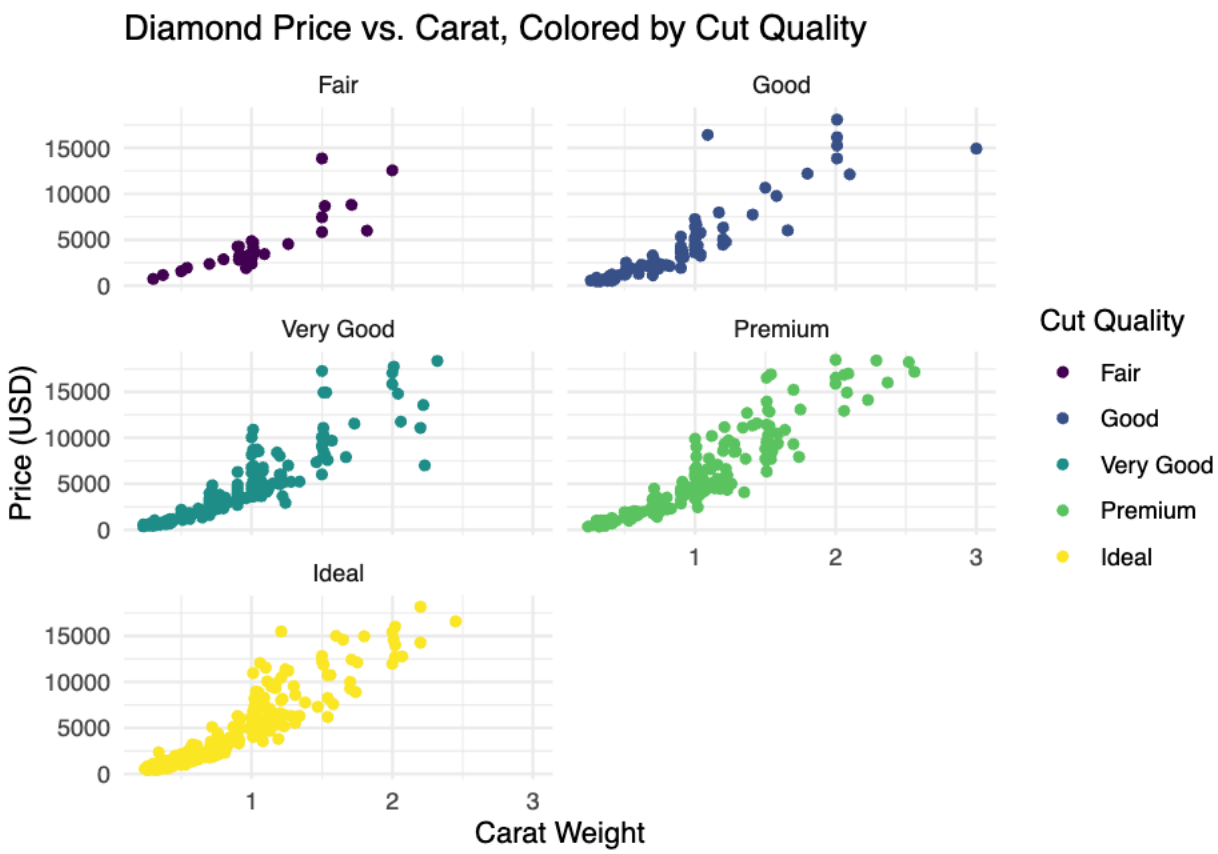
- It shows right skewed distribution.
- Larger diamonds is rarer than smaller diamonds.
- Store stocks carat that below 0.99 because of price advantage. Moreover, carat diamonds between below 0.99 and 1 is not significant different as much.

2. Price vs. Carat with Cut as Color:

```r
library(tidyverse)
library(dplyr)

set.seed(42)
small_diamonds <- diamonds %>%
  sample_n(1000)

ggplot(data = small_diamonds,
       mapping = aes(x = carat, y = price, color = cut)) +
  geom_point() +
  facet_wrap(~ cut, ncol = 2) +
  labs(title = "Diamond Price vs. Carat, Colored by Cut Quality",
       x = "Carat Weight",
       y = "Price (USD)",
       color = "Cut Quality") +
  theme_minimal()
```



- The pattern is when carat weight of diamonds is larger, pricing in every quality will be more expensive.

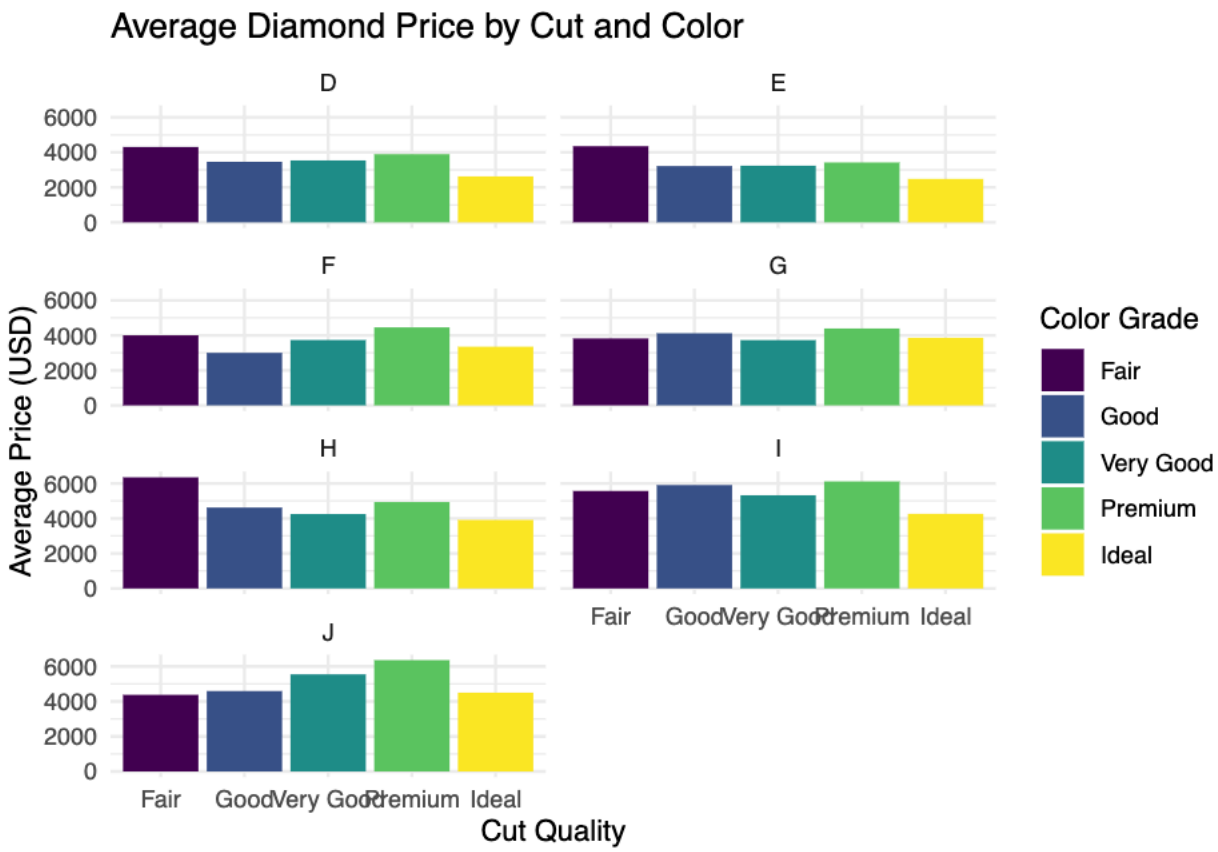
3.Average Price by Cut and Color:

```
library(tidyverse)
library(dplyr)

set.seed(42)
bar_diamonds <- diamonds %>%
  sample_n(5000) %>%
  group_by(cut, color) %>%
  summarise(avg_price = mean(price)) %>%
  arrange(-avg_price)
```

```
## `summarise()` has grouped output by 'cut'. You can override using the `.groups`
## argument.
```

```
ggplot(data = bar_diamonds,
       mapping = aes(x = cut, y = avg_price), fill = color,) +
  geom_col(mapping = aes(fill = cut), position = 'dodge') +
  facet_wrap(~ color, ncol = 2) +
  labs(title = "Average Diamond Price by Cut and Color",
       x = "Cut Quality",
       y = "Average Price (USD)",
       fill = "Color Grade") +
  theme_minimal()
```



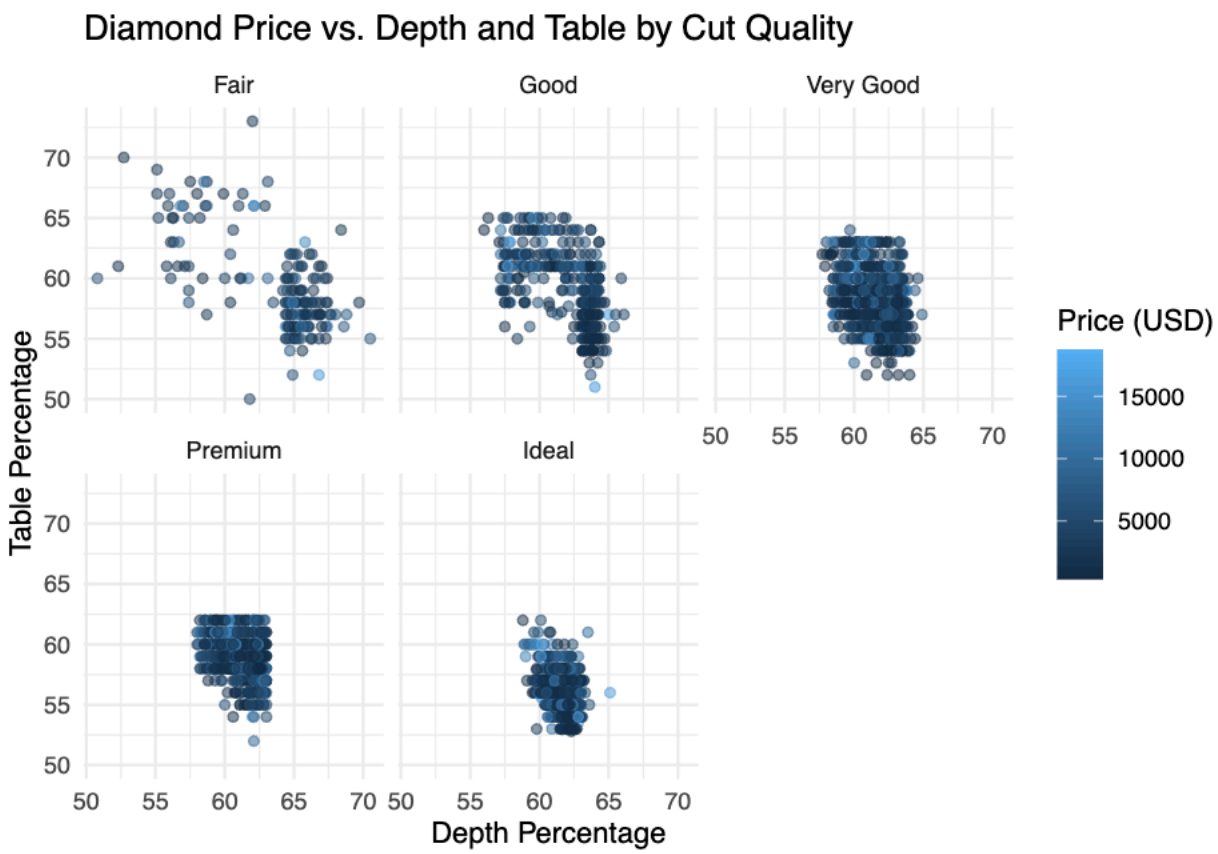
- The combinations of cut and color that tend to have the highest is premium with color J (6,360) and lowest average prices is Ideal with color E (2,483)

4.Relationship between Depth, Table, and Price:

```
library(tidyverse)
library(dplyr)

set.seed(42)
point_diamonds <- diamonds %>%
  sample_n(5000)

ggplot(data = point_diamonds,
       mapping = aes(depth, table))  +
  geom_point(mapping = aes(color = price), alpha = 0.5) +
  facet_wrap(~ cut) +
  labs(title = "Diamond Price vs. Depth and Table by Cut Quality",
       x = "Depth Percentage",
       y = "Table Percentage",
       color = "Price (USD)") +
  theme_minimal()
```



- Price is higher when depth is 59-62% and table 53-60% for round brilliants. These optimal ranges are where diamonds achieve maximum brilliance.

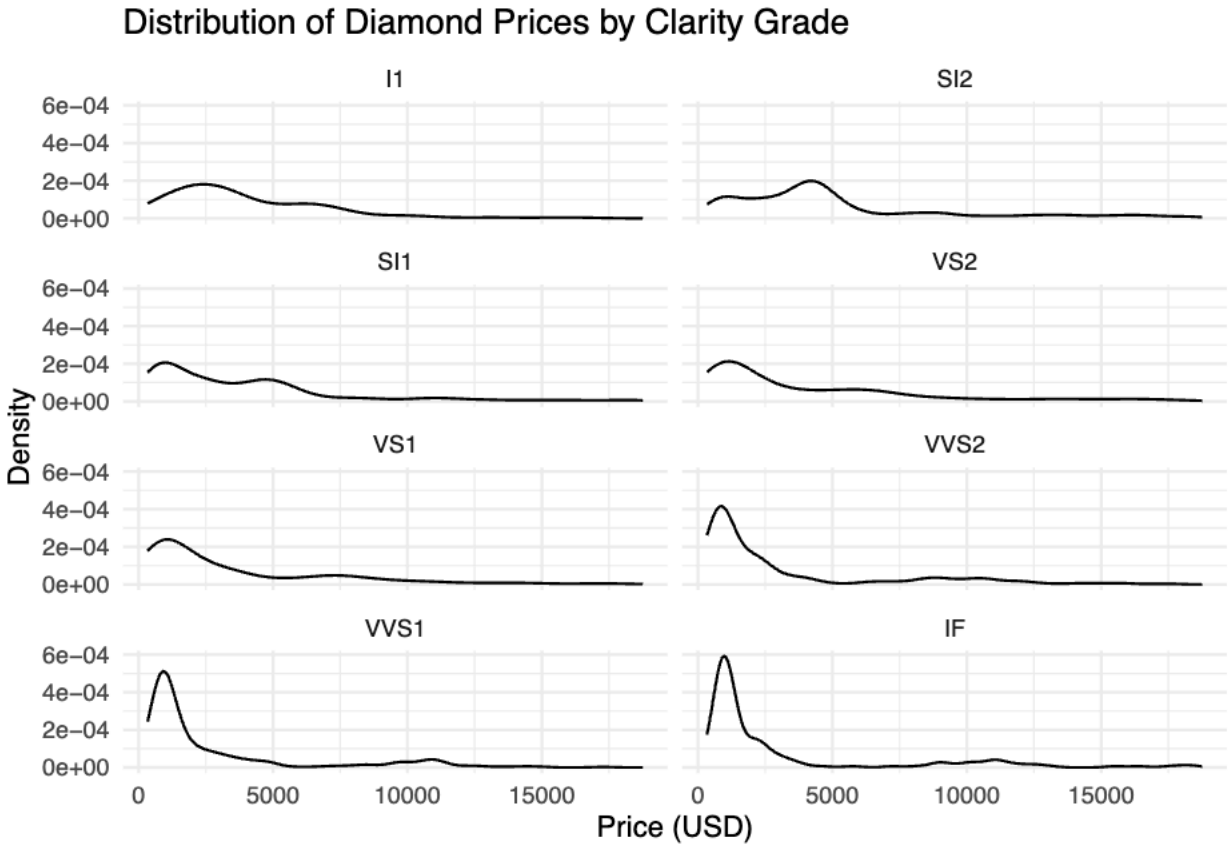
5.Faceting Price Distribution by Clarity:

```
library(tidyverse)
library(dplyr)

set.seed(42)
sam_diamonds <- diamonds %>%
  sample_n(5000)

cla_price <- sam_diamonds %>%
  group_by(clarity) %>%
  summarise(avg_price = median(price))


ggplot(data = sam_diamonds,
       mapping = aes(price)) +
  geom_density() +
  facet_wrap(~ clarity, ncol = 2) +
  labs(title = "Distribution of Diamond Prices by Clarity Grade",
       x = "Price (USD)",
       y = "Density") +
  theme_minimal()
```



- The higher clarity grades have consistently correlate with higher prices. This confirms that fewer inclusions and blemishes lead to a significant price premium, reflecting rarity and visual perfection.
- The higher clarity grades(VS1, VS2, VVS1, VVS2, IF, FL) show a narrow price range. However, the lower clarity grades(I1, SI2) show a wide price range.