Data visualization of diamonds dataset

Supattra P.

2023-02-26

Load library

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

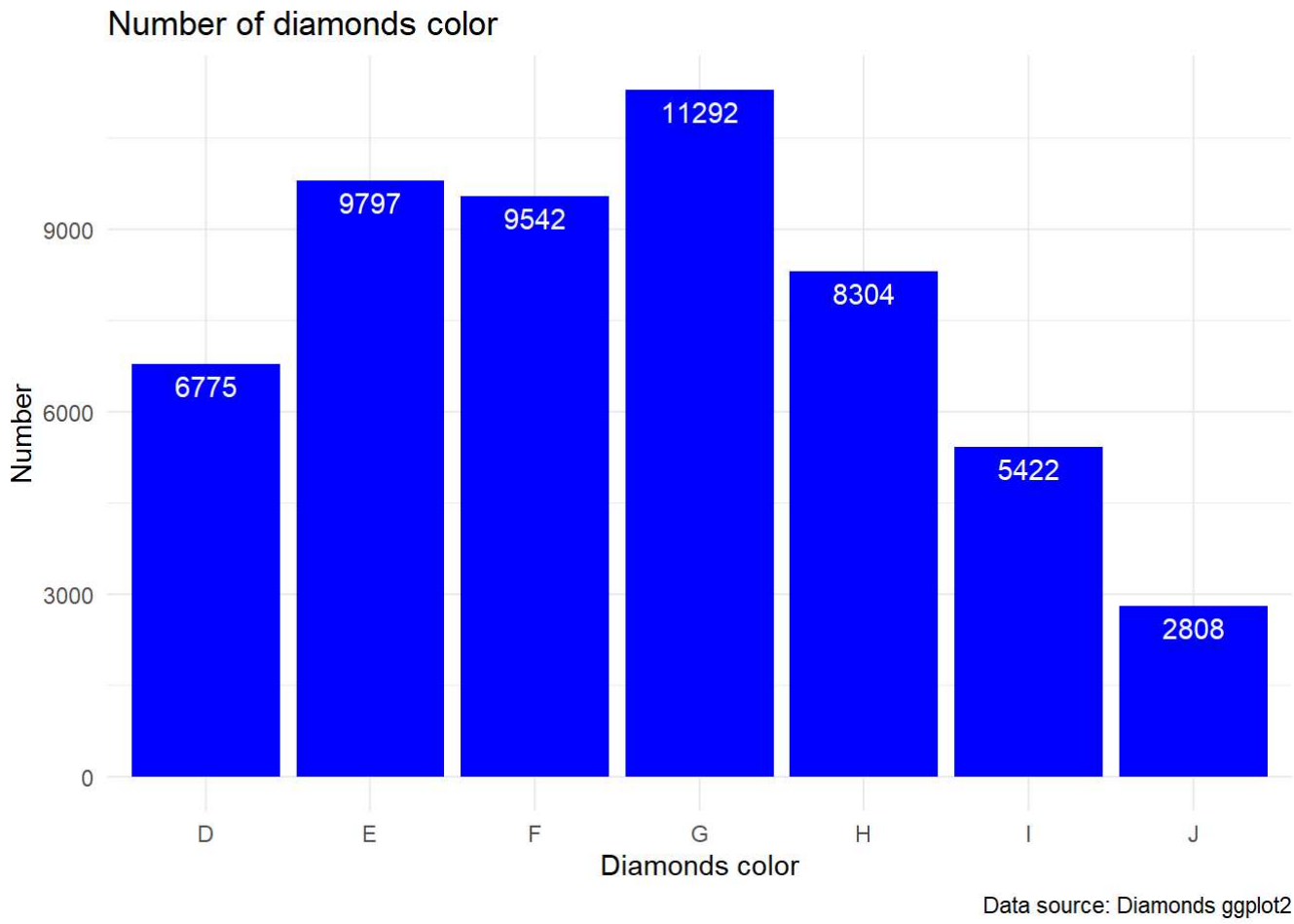
```
## Warning: package 'tibble' was built under R version 4.2.3
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
library(dplyr)
library(scales)
library (ggplot2)
```

Chart 1: Summary number of each diamonds color

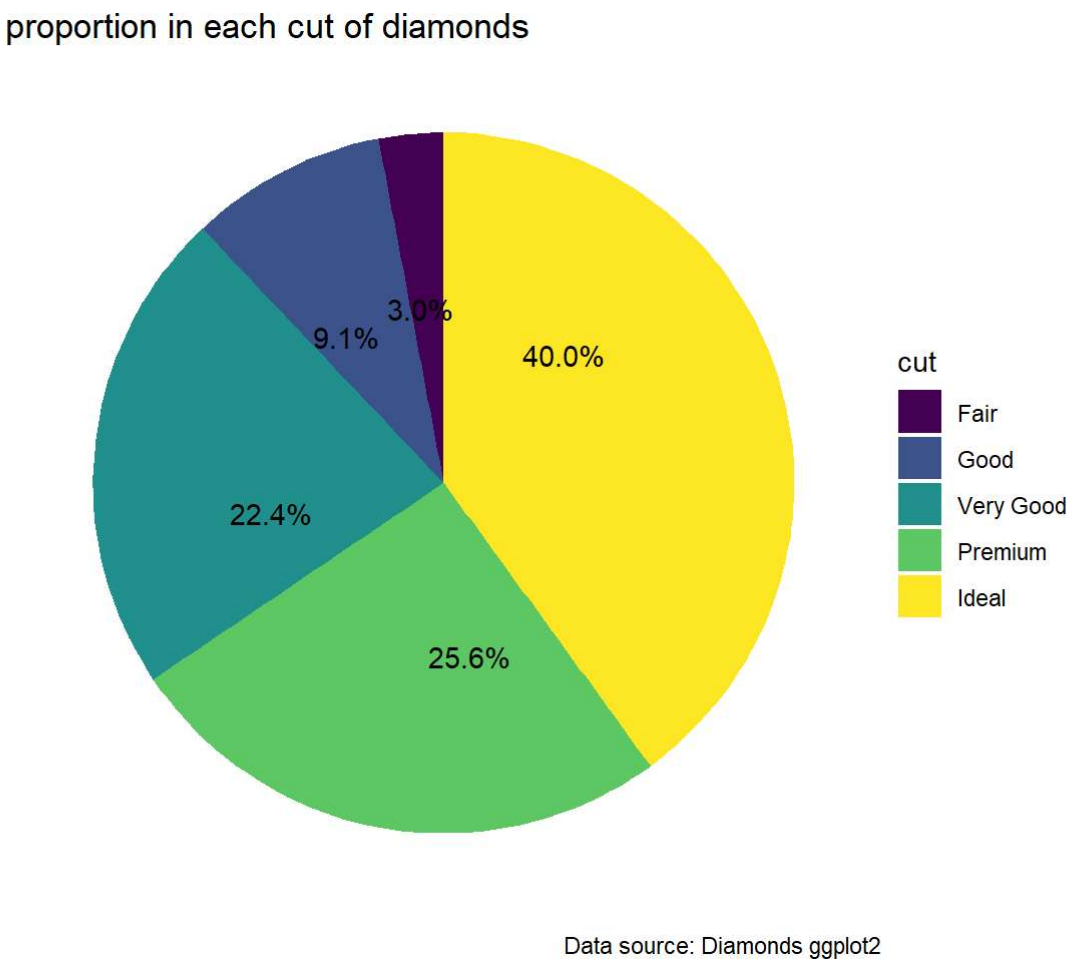
```
ggplot(diamonds, aes(x=color)) +
  geom_bar(fill = "blue") +
  geom_text(aes(label = after_stat(count)), stat = "count", vjust = 1.5, colour = "white") +
  theme_minimal() +
  labs(
    title = "Number of diamonds color",
    x = "Diamonds color",
    y = "Number",
    caption = "Data source: Diamonds ggplot2")
```



Conclusion: Found that top 3 of number of diamonds color; 1.color G 2.color E and 3. color F

Chart 2: Summary proportion in each cut of diamonds

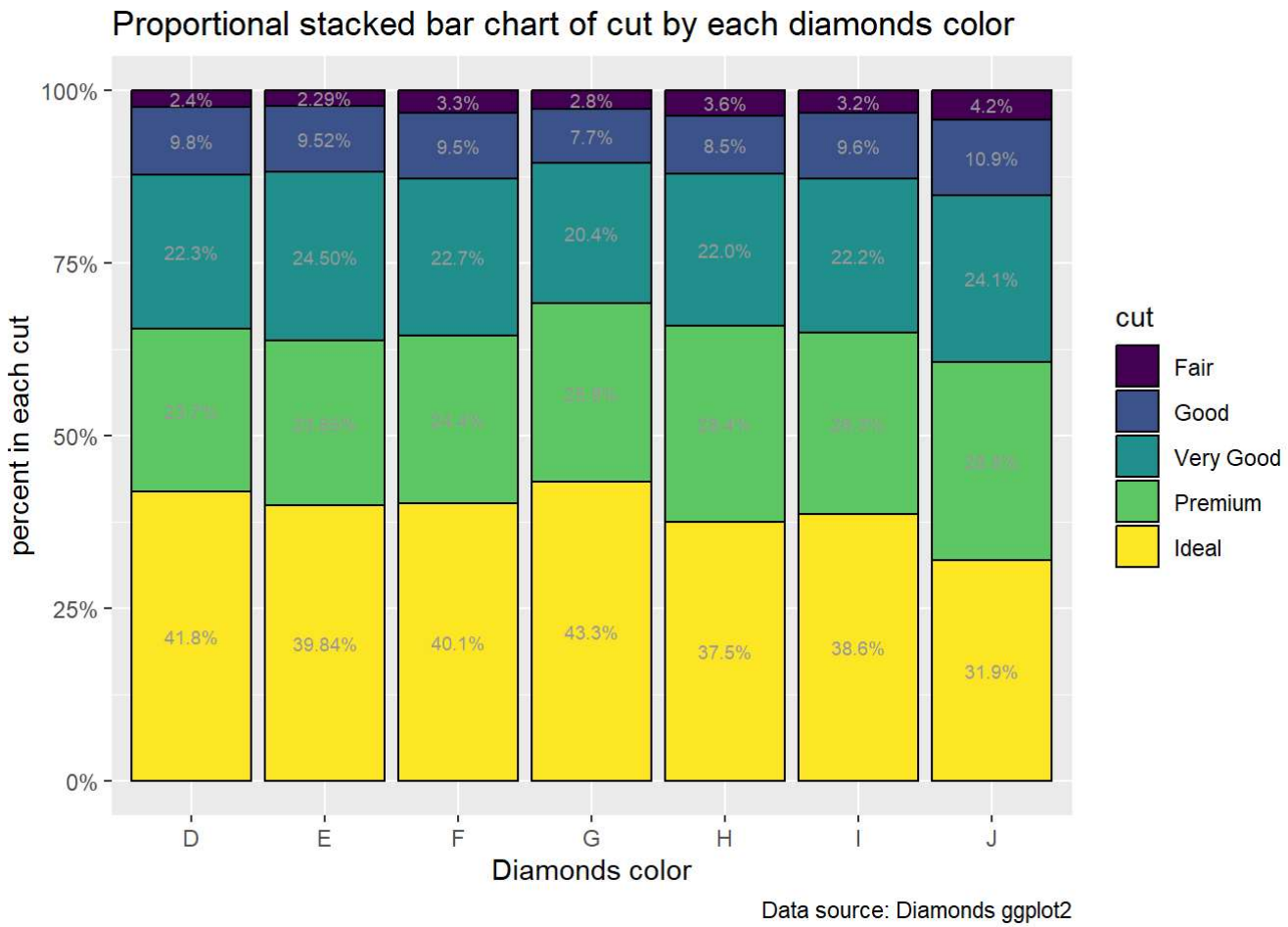
```
df <- diamonds %>%
  group_by(cut) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

ggplot(df, aes(x = "", y = perc, fill = cut)) +
  geom_col() +
  geom_text(aes(label = labels),
            position = position_stack(vjust = 0.3)) +
  coord_polar(theta = "y") +
  theme_void() +
  labs(
    title = "proportion in each cut of diamonds",
    caption = "Data source: Diamonds ggplot2")
```



Conclusion: Ideal is the most proportion that 40% of all

Chart 3: Proportional stacked bar chart of cut by each diamonds color

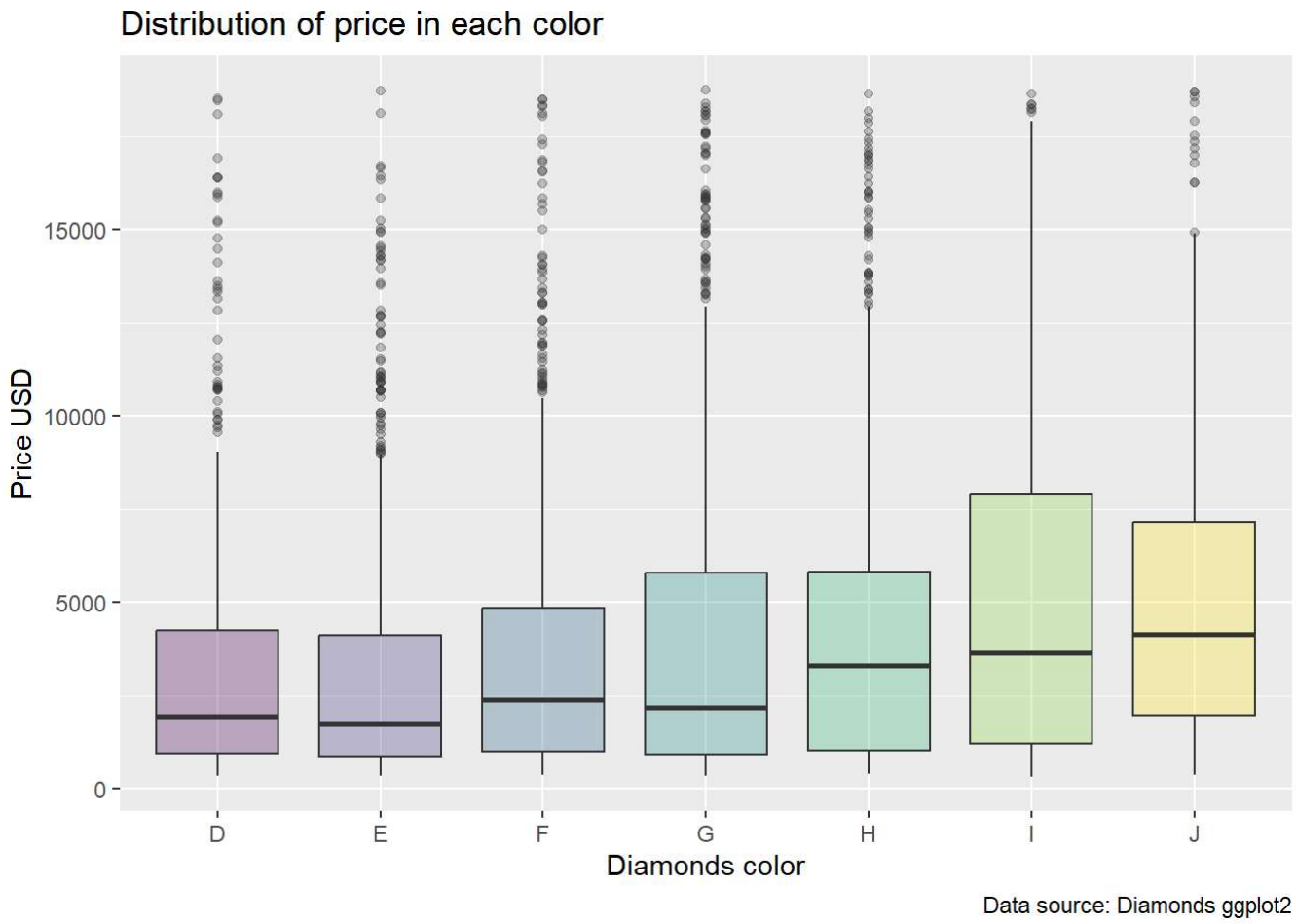
```
#create summary table
t1 <- diamonds %>%
  group_by(color, cut) %>%
  tally() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  mutate(labels = scales::percent(perc))

ggplot(t1, aes(x = color, y = perc, fill = cut)) +
  geom_bar(position = "fill", stat = "identity",color='black',width=0.9) +
  scale_y_continuous(labels = scales::percent) +
  geom_text(aes(label = labels),
            position = position_stack(vjust = 0.5), size = 2.5, colour = "#999999") +
  labs(
    title = "Proportional stacked bar chart of cut by each diamonds color",
    x = "Diamonds color",
    y = "percent in each cut",
    caption = "Data source: Diamonds ggplot2")
```



Conclusion: In each diamonds color is contain of most Ideal cut and color G has percentage of Ideal cut more than other colors

Chart 4: Distribution of price in each color

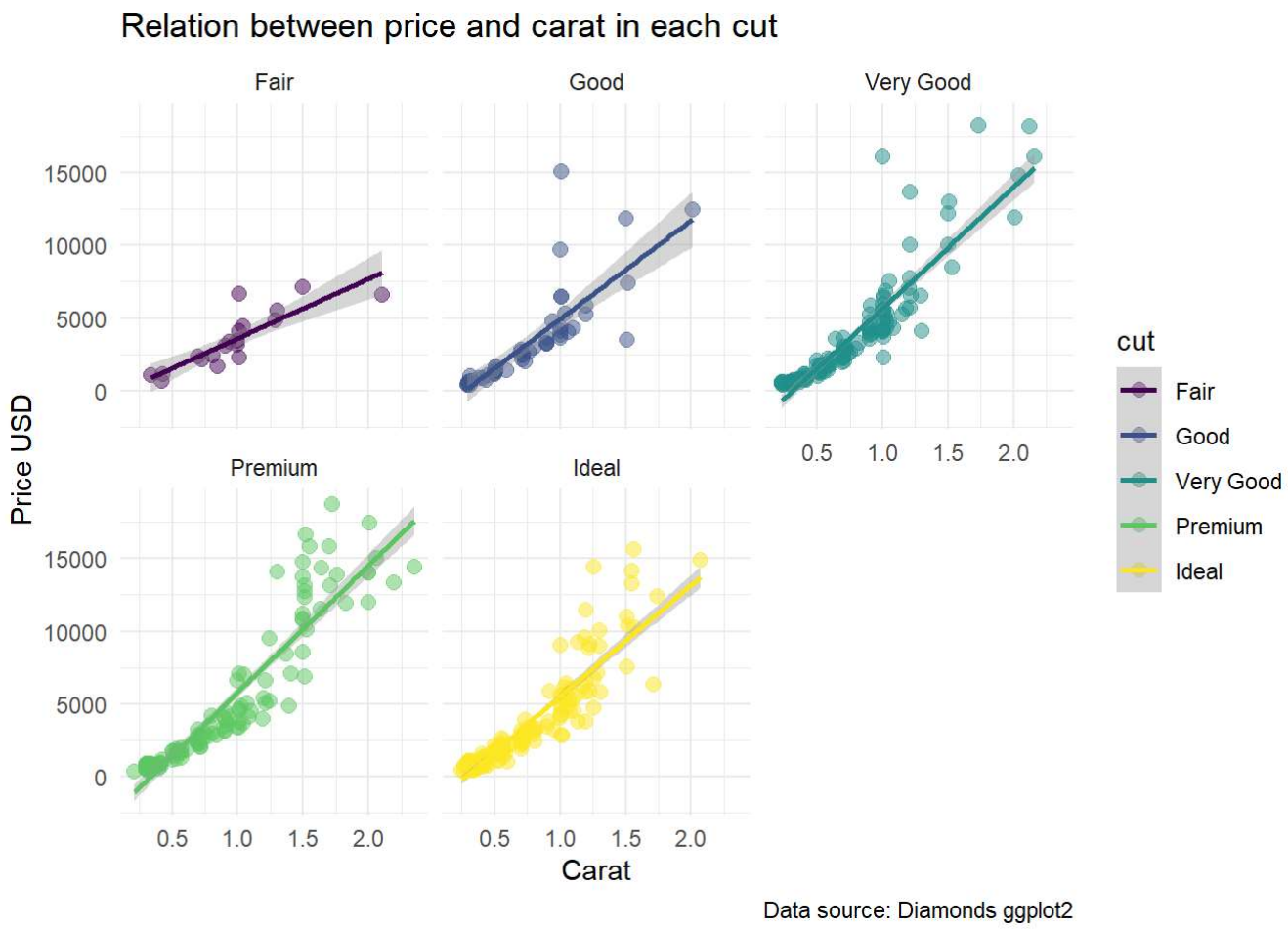
```
set.seed(42)
ggplot(diamonds %>% sample_n(5000), aes(color, price, fill = color)) +
  geom_boxplot(alpha=0.3) +
  theme(legend.position="none") +
  labs(
    title = "Distribution of price in each color",
    x = "Diamonds color",
    y = "Price USD",
    caption = "Data source: Diamonds ggplot2")
```



Conclusion: color I is the most variance when compare with other colors

Chart 5: Relation between price and carat in each cut

```
ggplot(diamonds %>% sample_n(500), aes(carat, price, col = cut)) +
  geom_point(size = 2.5, alpha = 0.5) +
  theme_minimal() +
  labs(
    title = "Relation between price and carat in each cut",
    x = "Carat",
    y = "Price USD",
    caption = "Data source: Diamonds ggplot2") +
  geom_smooth(method = lm) +
  facet_wrap(~ cut)
```



Conclusion: In each cut found that when carat is increased, then price is increased