Data Visualisation Bootcamp Homework

Parn

2023-07-07

Instruction

Use diamonds dataset to create 5 charts. knit pdf and submit in discord.

Prepare the environment by install & library packages

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

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

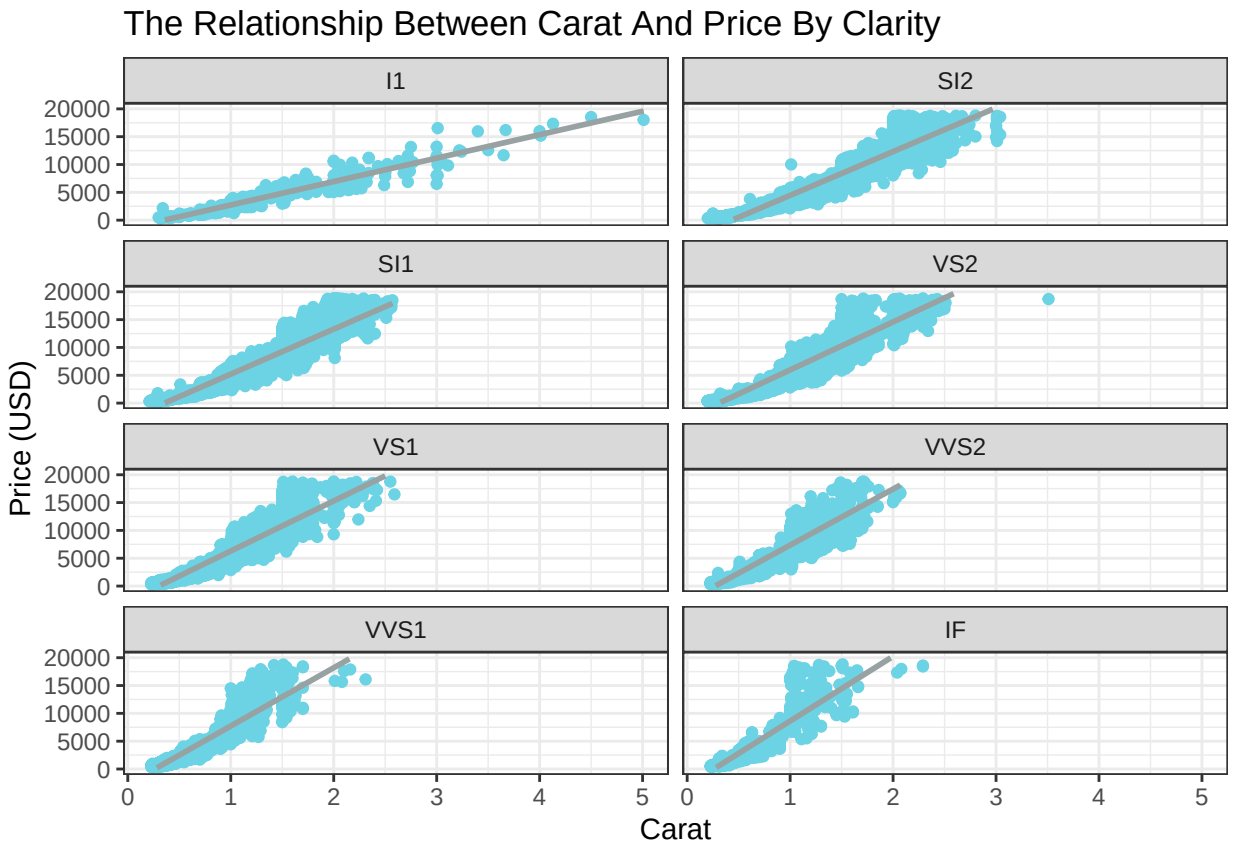
The Relationship Between Carat and Price By Clarity

Two continuous variables

```
ggplot(diamonds, aes(carat, price)) +
  geom_point(color = "#6bd3e3") +
  geom_smooth(method = "lm", color = "#98a2a3") +
  ylim(0, 20000)+
  facet_wrap(~clarity, ncol = 2) +
  theme_bw() +
  labs(title = "The Relationship Between Carat And Price By Clarity",
       x = "Carat",
       y = "Price (USD)")
```

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

```
## Warning: Removed 70 rows containing missing values (`geom_smooth()`).
```



The relationship between carat and price is positive in every group of clarity. It is also confirmed by the positive correlation of these factors (correlation = 0.92).

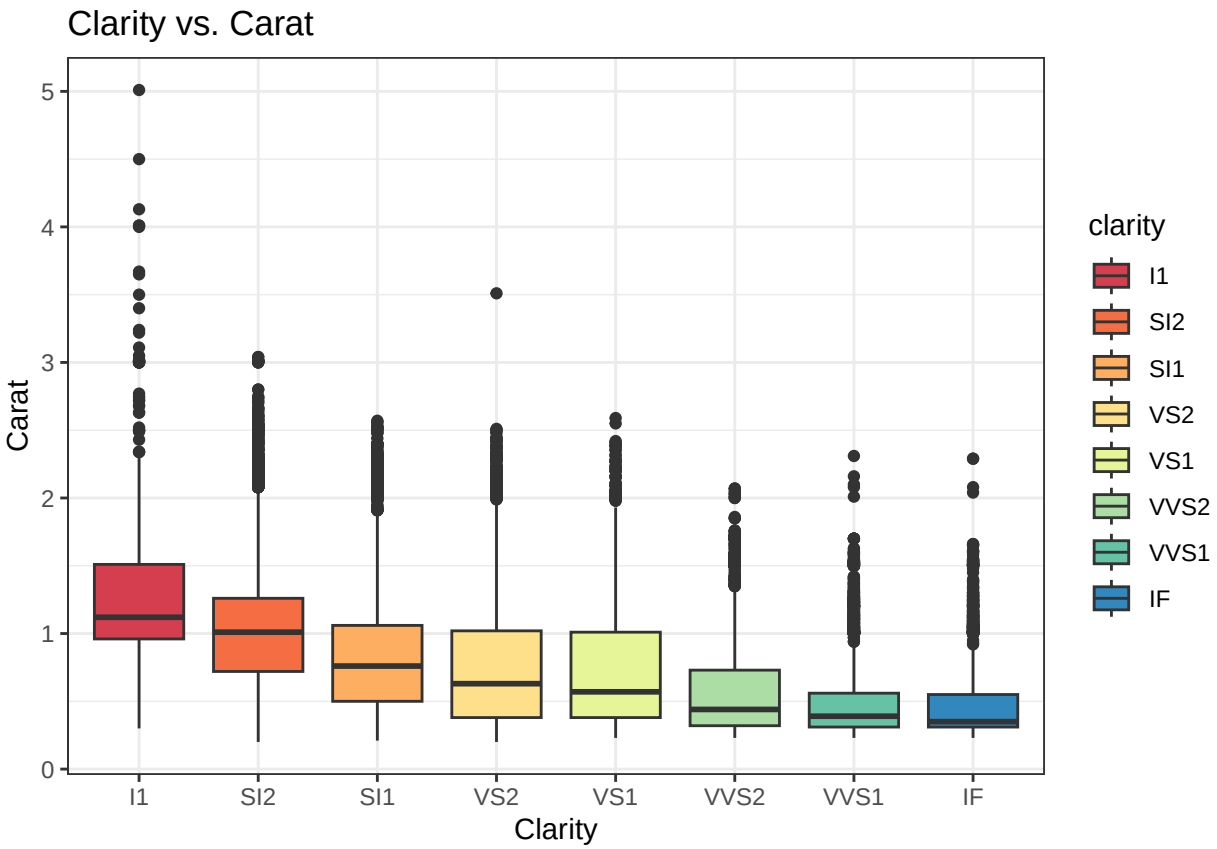
```
cor(diamonds$carat, diamonds$price)
```

```
## [1] 0.9215913
```

Calrity VS. Carat

Two variables: one discrete and one continuous

```
ggplot(diamonds, aes(clarity, carat, fill = clarity)) +
  geom_boxplot() +
  theme_bw() +
  scale_fill_brewer(type = "div", palette = 9) +
  labs(title = "Clarity vs. Carat",
       x = "Clarity",
       y = "Carat")
```



The graph shows that the majority of the diamond clarity groups has carat values less than 3. However, the values within the I1 clarity (the worst) are more spread out than others. This is confirmed by the highest standard deviation of this group.

Statistic Summary By Clarity

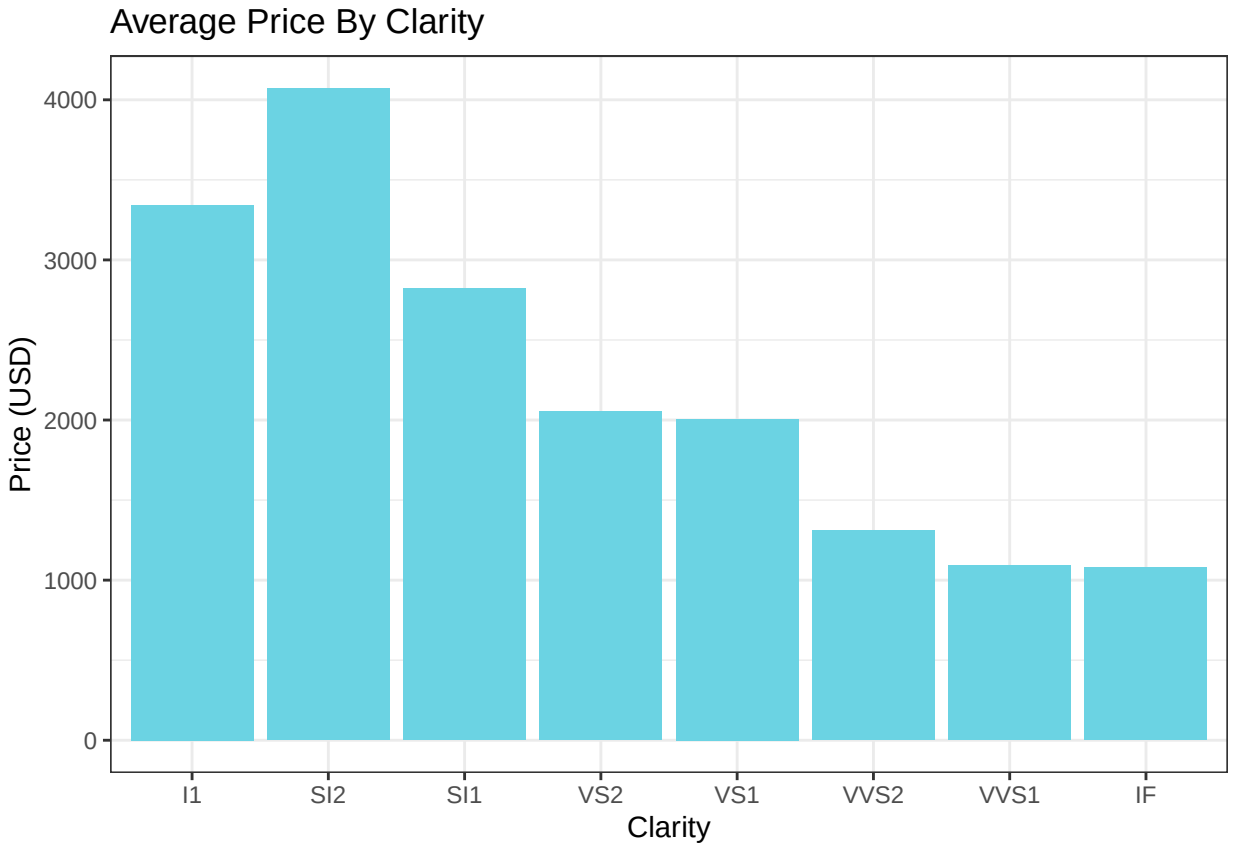
```
## # A tibble: 8 x 2
##   clarity sd_carat_by_clarity
##   <ord>                 <dbl>
## 1 I1                    0.632
## 2 SI2                   0.517
## 3 SI1                   0.450
## 4 VS2                   0.446
## 5 VS1                   0.424
## 6 VVS2                  0.360
## 7 VVS1                  0.300
## 8 IF                    0.313
```

The Average Price By Clarity

Two variables: one discrete and one continuous

```
diamonds %>% group_by(clarity)%>%
  summarise(avg_price_by_clarity = median(price)) %>%
  ggplot(aes(clarity, avg_price_by_clarity)) +
  geom_col(fill = "#6BD3E3") +
  theme_bw() +
```

```
  labs (title = "Average Price By Clarity",
        x = "Clarity",
        y = "Price (USD)")
```

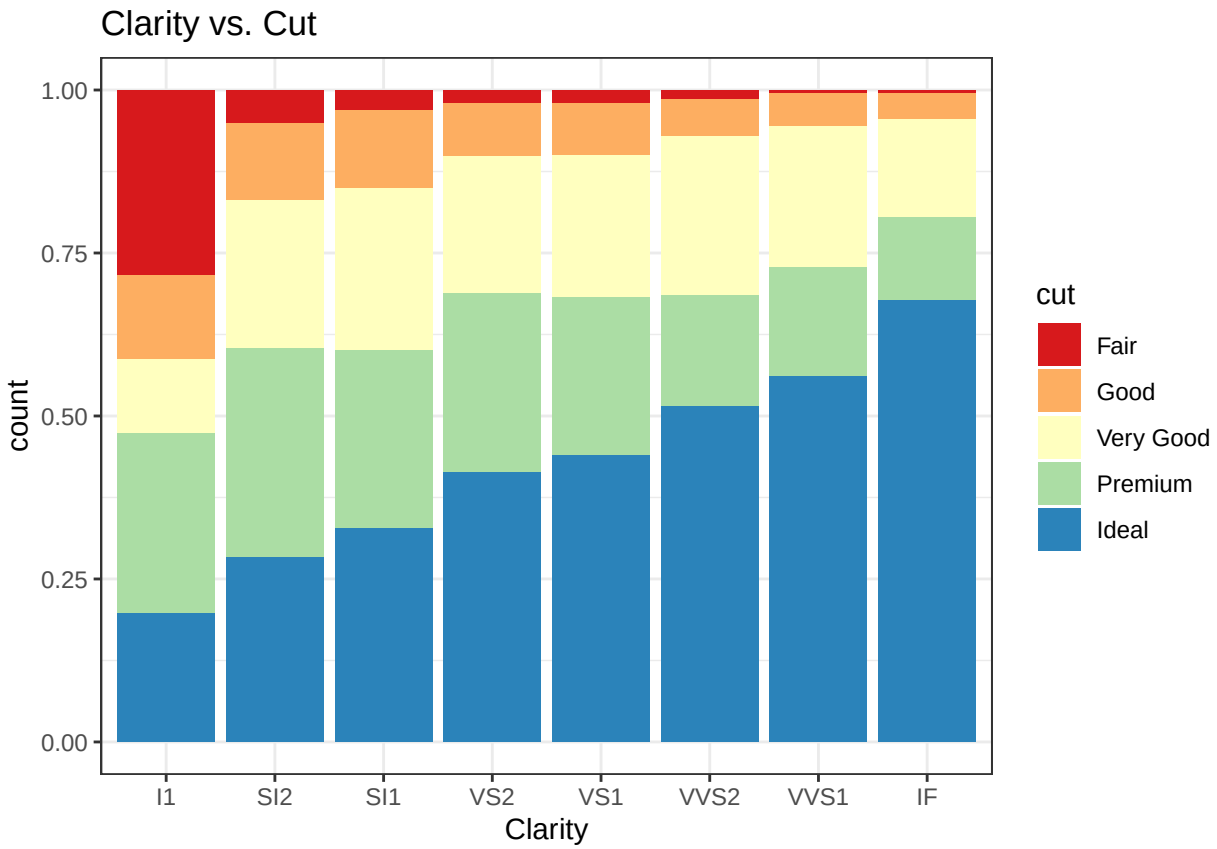


This graph shows that the SI2 group has the highest average price even though it is not the best clarity of diamonds. It is likely that clarity of diamonds does not define the price.

Clarity VS. Cut

One discreted variable

```
ggplot(diamonds, aes(clarity, fill = cut)) +
  geom_bar(position = "fill") +
  scale_fill_brewer(type = "div", palette = 9) +
  theme_bw() +
  labs(title = "Clarity vs. Cut",
       x = "Clarity")
```

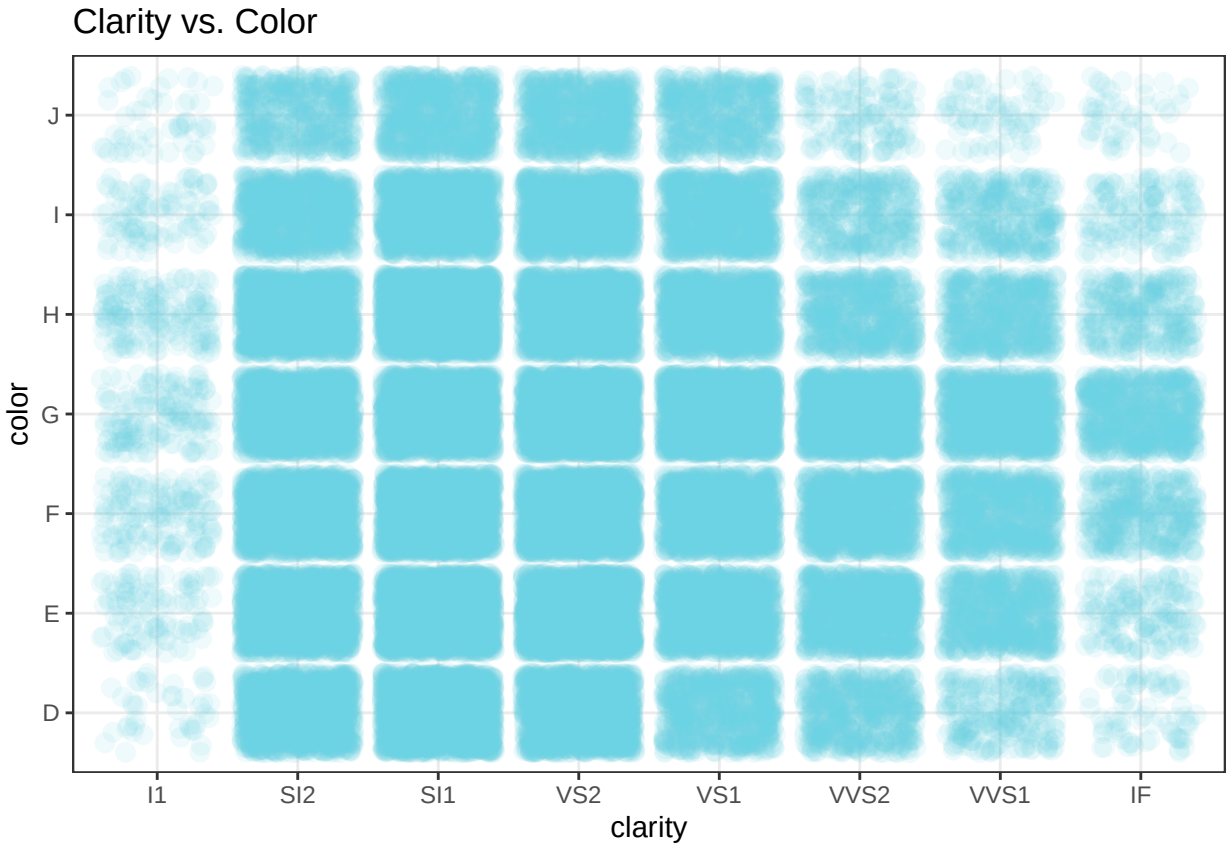


This graph shows that the diamonds with higher clarity tend to have the higher quality of the cut as well.

Clarity VS. Color

Two discreted variables

```
ggplot(diamonds, aes(clarity, color)) +
  geom_jitter(col = "#6BD3E3", alpha = 0.1, size = 3) +
  theme_bw() +
  labs(title = "Clarity vs. Color")
```



This graph shows the population of the diamonds dataset by clarity and color.