Explore diamonds with ggplot

Kavinddd

Import library and prepare data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     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)
# Check Null
diamonds %>% is.na() %>% sum()
```

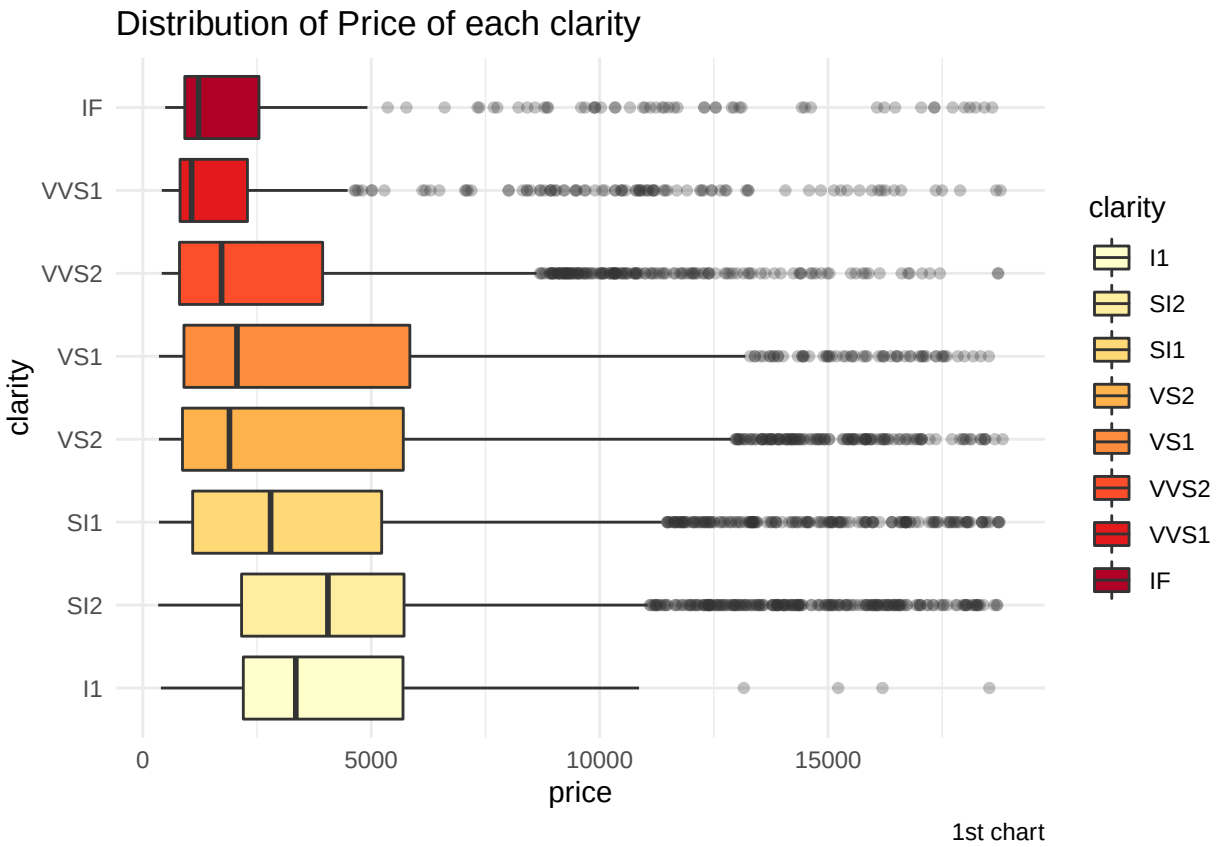
```
## [1] 0
```

```
# So the data is already cleaned
# Take 5000 sample of the data to allow faster execute
set.seed(2001)
df <- diamonds %>% sample_n(size = 10000)
```

EDA

Distribution of price of each clarity

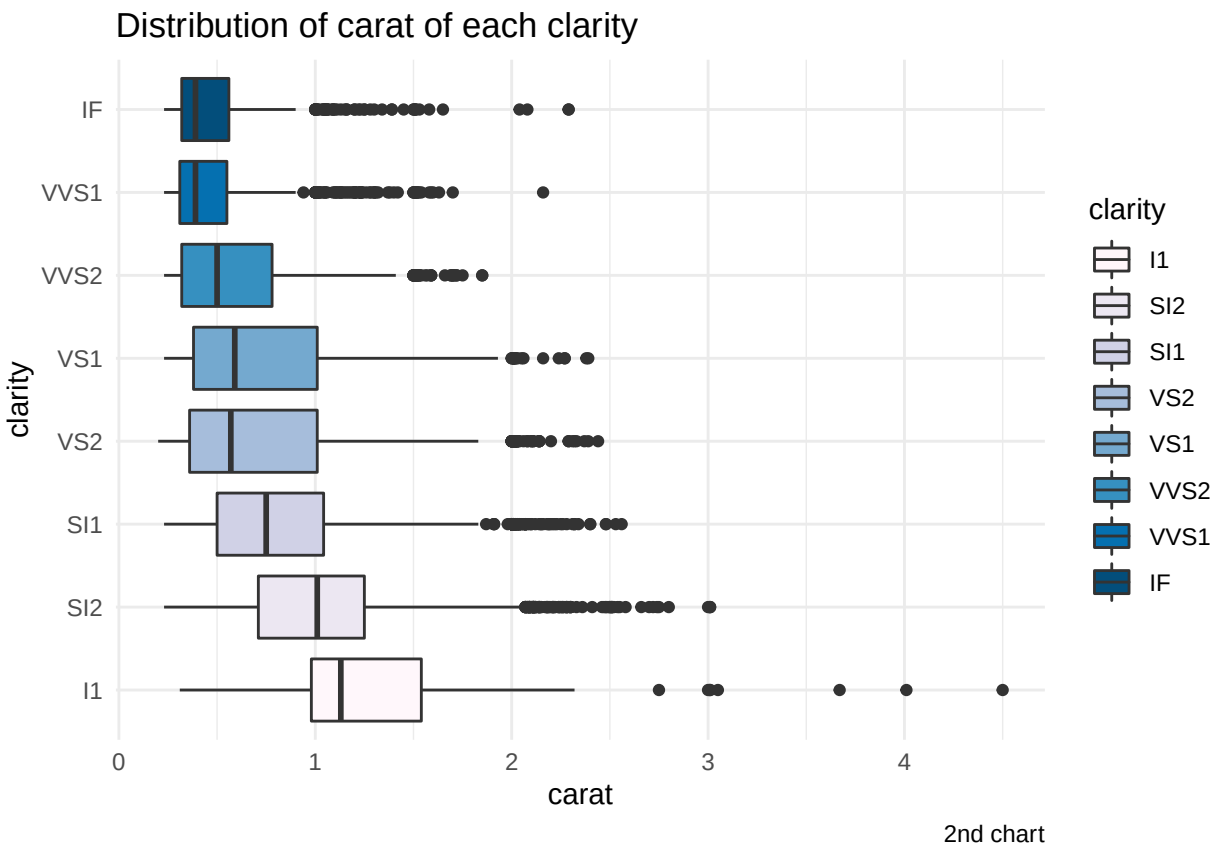
```
df %>%
  ggplot(aes(price, clarity, fill=clarity)) +
  geom_boxplot(outlier.alpha=0.3) +
  labs(title="Distribution of Price of each clarity",
       caption='1st chart') +
  theme_minimal() +
  scale_fill_brewer(palette='YlOrRd', type='seq')
```



VVS1 has the lowest median, so this type of clarity tends to be sold for a lowest price among all clarity, but SI2 has the highest median, so this type is the most expensive in average. VS1 and VS2 are selling in a very large range of prices, so there must be other factors influenced the price. Anyway, this chart shows something that make no sense at all, the IF clarity level is the rarest and most expensive one in reality, but this shows something opposite because the IF diamonds are might very small and have a light weight naturally.

The distribution of carat (weight of diamonds) of each clarity

```
df %>%
  ggplot(aes(y=clarity, x=carat, fill=clarity)) +
  geom_boxplot() +
  theme_minimal() +
  labs(
    title="Distribution of carat of each clarity",
    caption = "2nd chart"
    )+
  scale_fill_brewer(palette = 'PuBu')
```



As expected, the weight of the IF is very small, so the weight of diamonds is a very important factor of the price of diamonds.

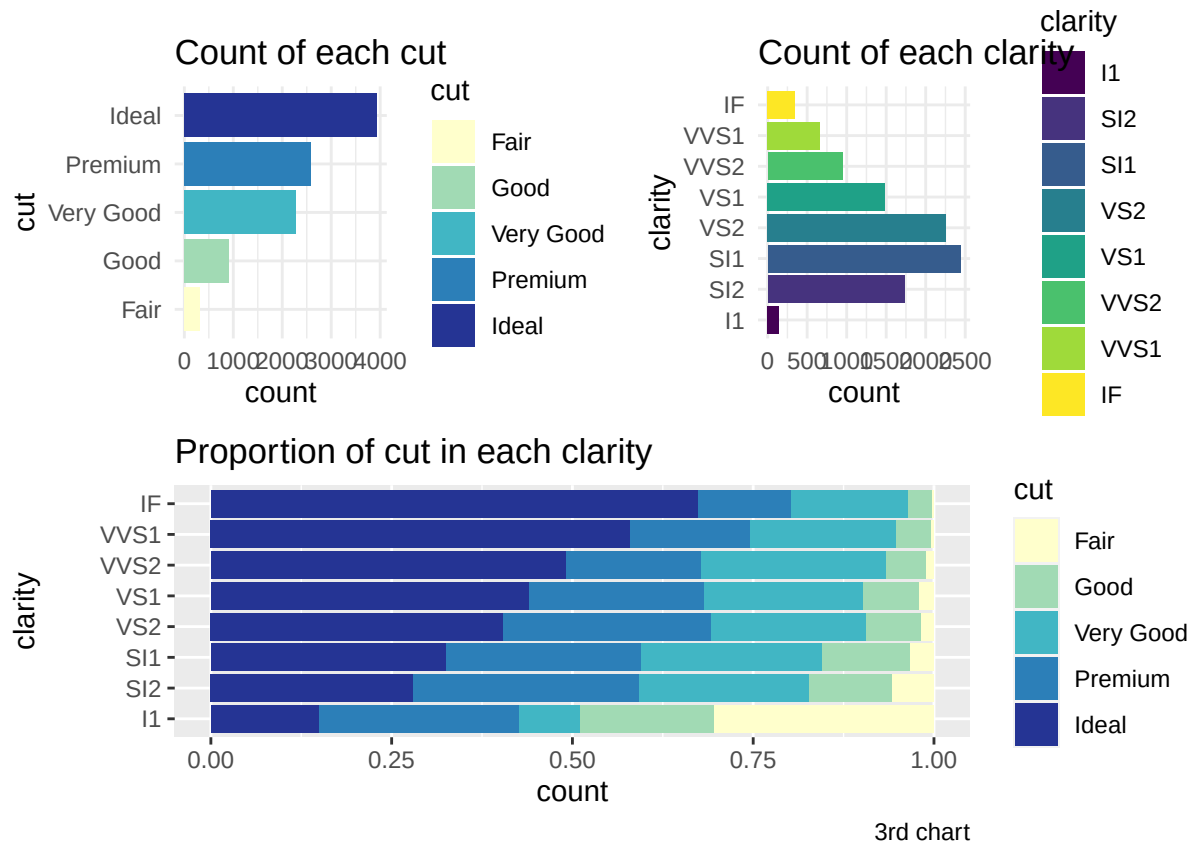
Count plot of each cut and clarity

```
p1 <- df %>%
  ggplot(aes(y=cut, fill=cut)) +
  geom_bar() +
  labs(title="Count of each cut")+
  theme_minimal() +
  scale_fill_brewer(palette = 'YlGnBu',type = "seq")

p2 <- df %>%
  ggplot(aes(y=clarity, fill=clarity)) +
  geom_bar() +
  labs(title="Count of each clarity")+
  theme_minimal()


p3 <- df %>%
  ggplot(aes(y=clarity,fill=cut)) +
  geom_bar(position = 'fill') +
  labs(title='Proportion of cut in each clarity',
       caption='3rd chart') +
  scale_fill_brewer(palette='YlGnBu', type='seq')
```

```
(p1 + p2) / p3
```



The proportion of cut of each clarity - So, we can see that type cut is also another factor that influenced high price, I1 and SI2 are the clarity having the highest proportion in fair cuts, and they are also the clarity that have the highest selling price.

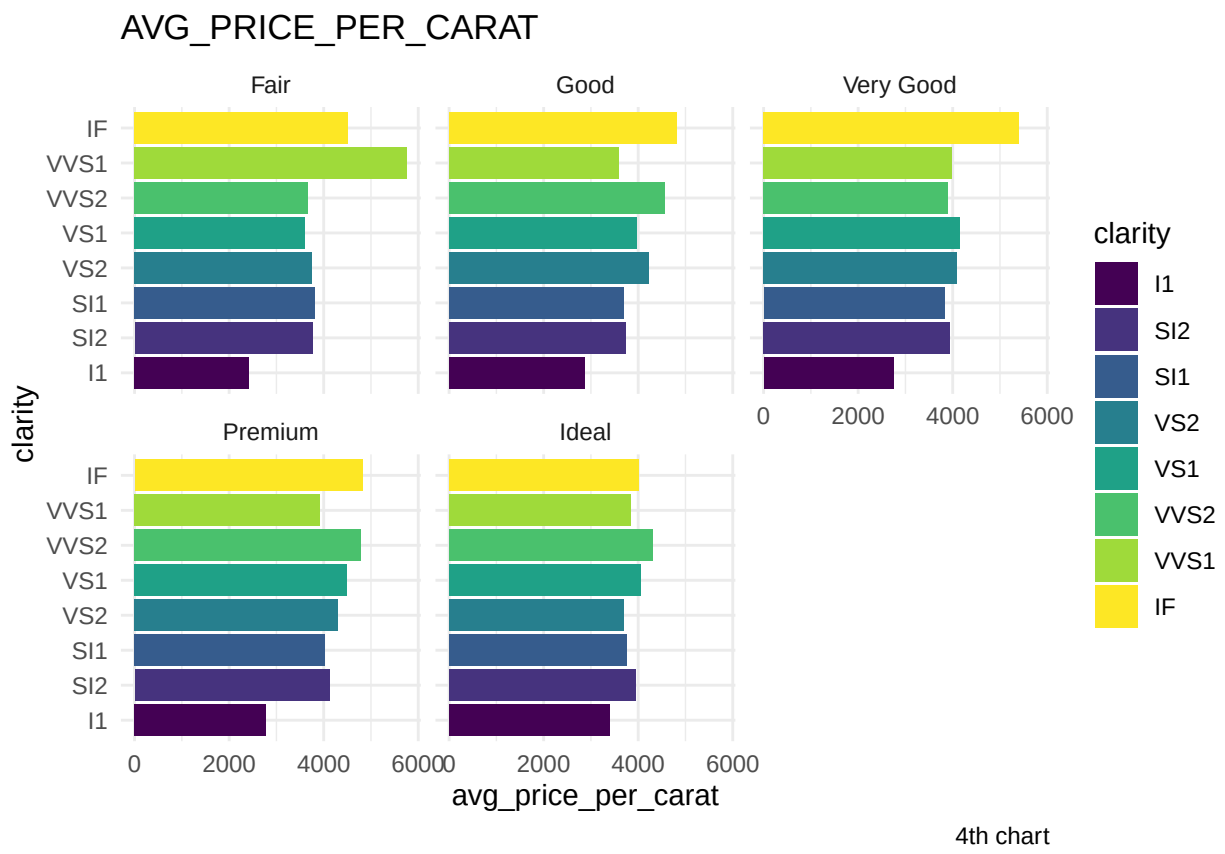
Price/Carat

We cannot use price to measure which one is expensive directly, so we will divide price by carat.

```
# Create price_per_carat column
price_per_carat <- df %>% transmute(
  clarity,
  cut,
  price_per_carat = price/carat
) %>%
  group_by(cut,clarity) %>%
  summarise(
    avg_price_per_carat = mean(price_per_carat)
  )
```

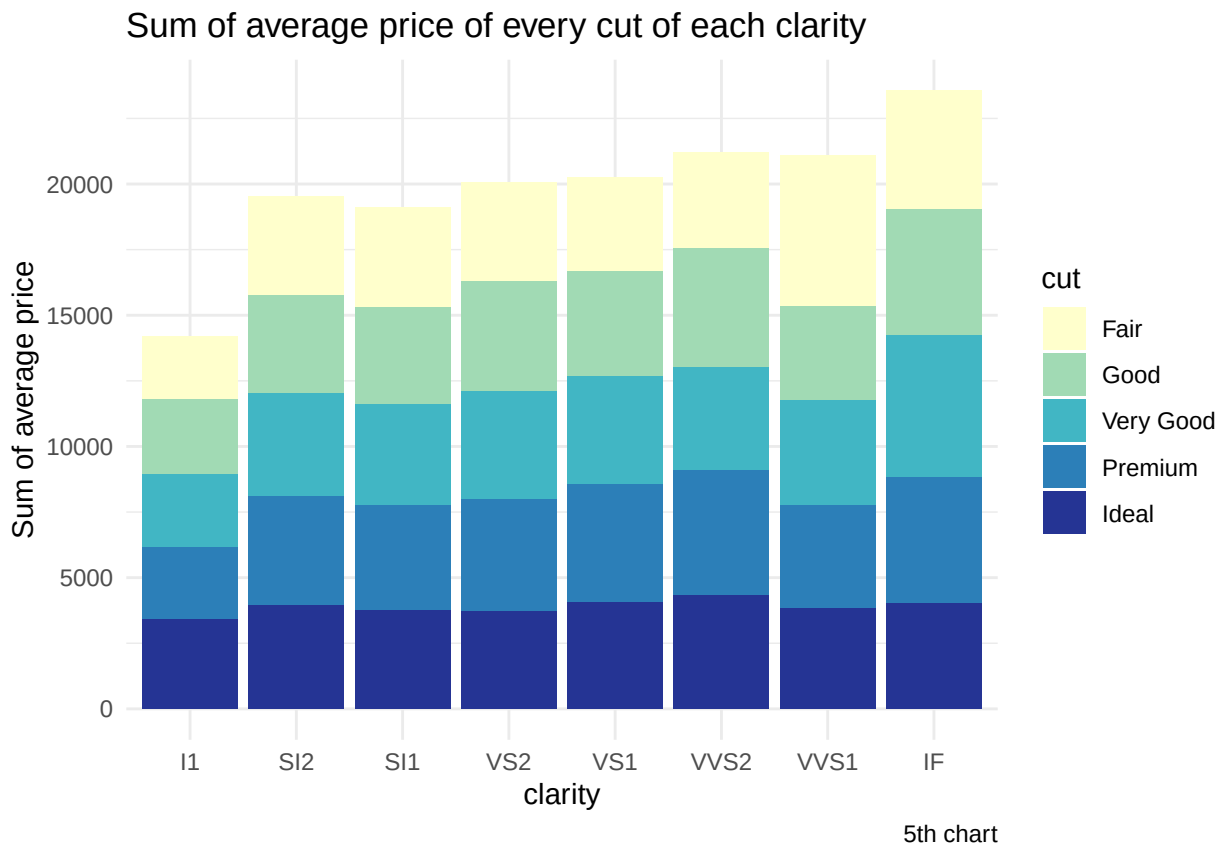
```
## `summarise()` has grouped output by 'cut'. You can override using the `.groups`
## argument.
```

```
# Plot
price_per_carat %>%
  ggplot(aes(y=clarity, x=avg_price_per_carat, fill=clarity)) +
  geom_col() +
  facet_wrap(~cut) +
  theme_minimal() +
  labs(title = 'AVG_PRICE_PER_CARAT',
       caption='4th chart')
```



The chart makes more sense now, let's see if we aggregate this by sum and see which one is the most expensive (Stacked bar chart)

```
p4 <- price_per_carat %>%
  ggplot(aes(clarity,avg_price_per_carat, fill=cut))
p4 +
  geom_col() +
  labs(
    title= 'Sum of average price of every cut of each clarity ',
    caption='5th chart')+
  ylab("Sum of average price") +
  theme_minimal()  +
  scale_fill_brewer(palette = 'YlGnBu',type = "seq")
```



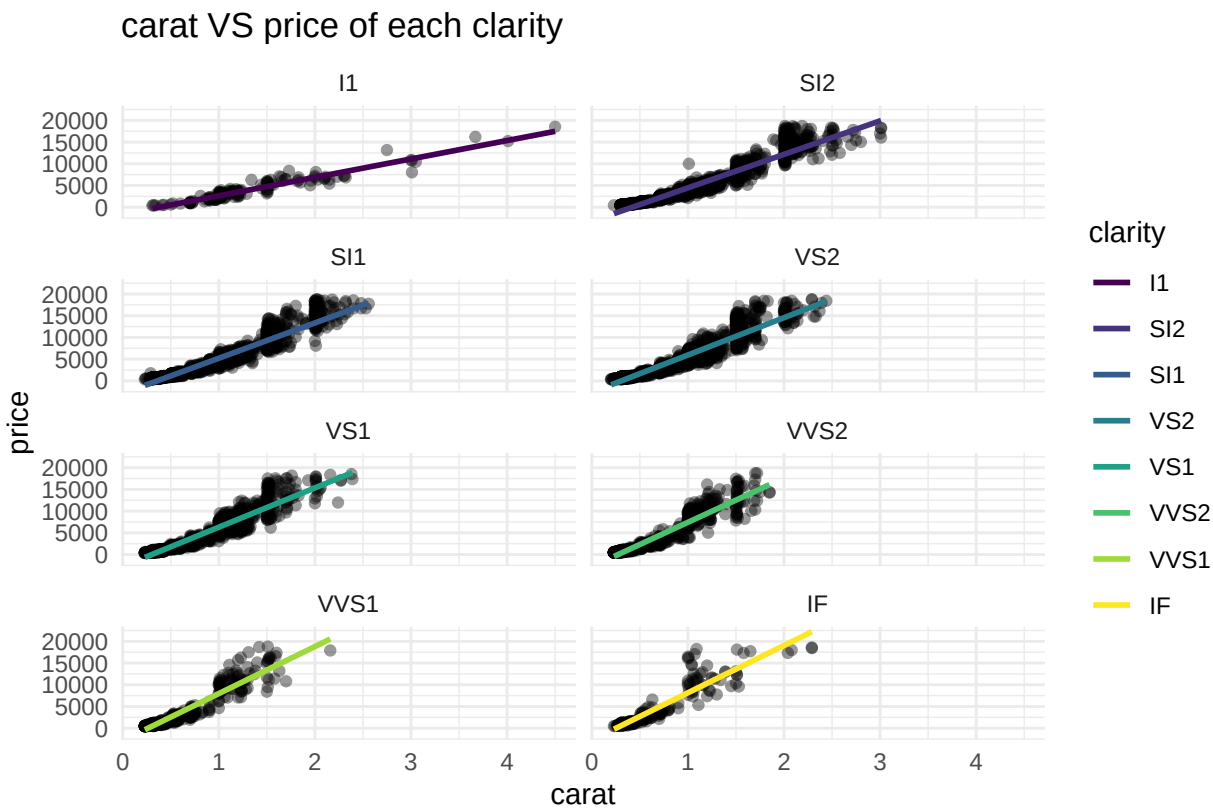
There we go! the IF clarity type is the most valuable and rarest, so this makes sense now.

Relationship between weight and price in each clarity types

More valuable the clarity, more steep the regression line.

```
df %>%
  ggplot(aes(carat, price, color=clarity)) +
  geom_point(color='black', alpha=0.4) +
  geom_smooth(method='lm', se=F) +
  facet_wrap(~clarity,nrow = 4) +
  theme_minimal() +
  labs(title="carat VS price of each clarity",
       caption='6th chart')
```

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


carat VS price of each clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
price
carat
0
5000
10000
15000
20000
0
1
2
3
4
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF

6th chart