Exploratory Data Analysis using ggplot

Wanicha Mueangcharoen

Import libraries

```
## -- 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()
##
## Attaching package: 'lubridate'
##
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Data Preparation and Cleaning

- **Overview**

```
head(diamonds)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

- **Cleaning**

```
#find columns with NA's
apply(diamonds, 2, anyNA)
```

```
##   carat     cut   color clarity   depth   table   price       x       y       z
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
```

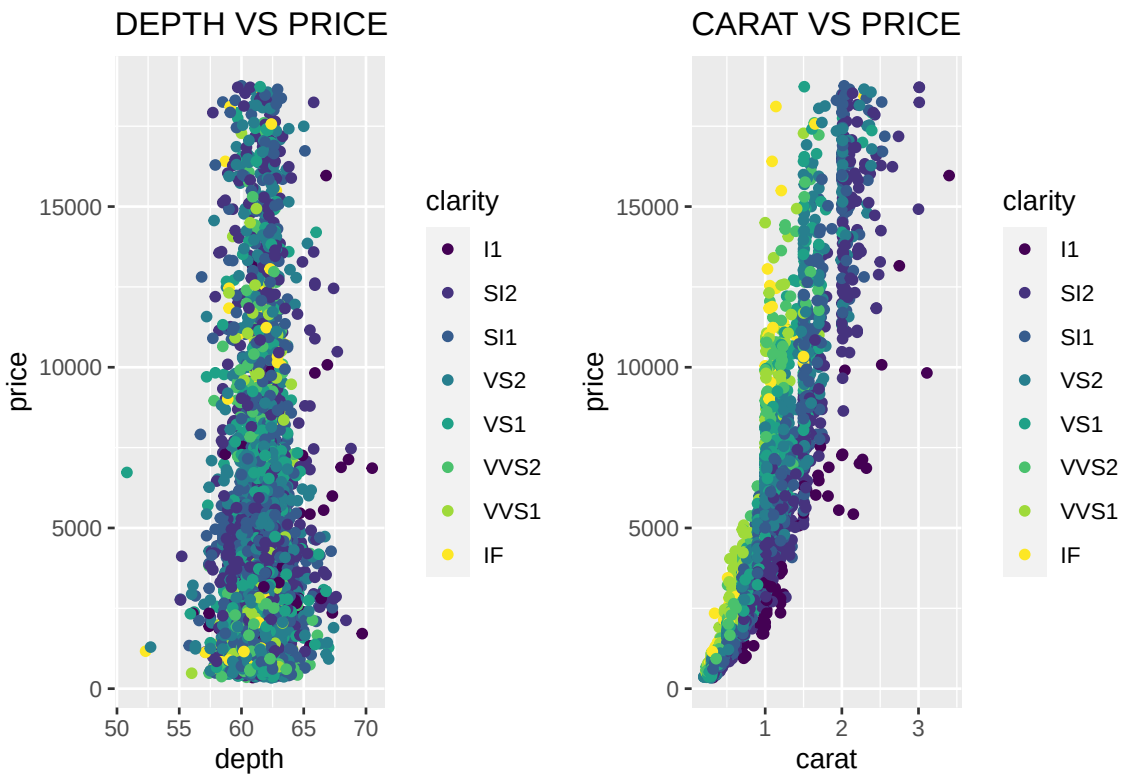
```
# Take 3000 samples for faster execution
set.seed(42) # the recommended way to specify seeds.
df_diamonds <- diamonds %>%
                  sample_n(size = 5000)
```

EDA

1. **Relationships between Price, Carat and Depth of Diamonds (based on Clarity):**

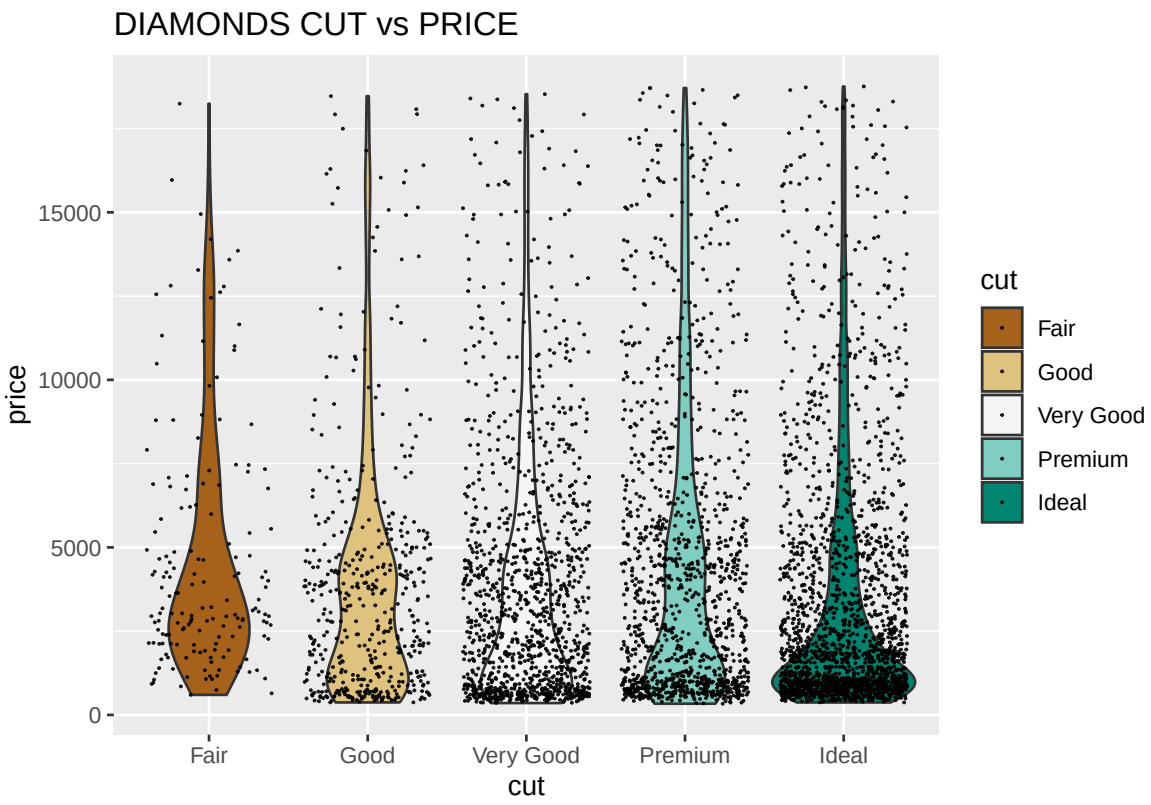
```
d1 <- df_diamonds %>%
  ggplot(aes(depth,price, color = clarity)) +
  geom_point() +
#       theme_clean() +
  labs( title = "DEPTH VS PRICE")


d2 <- df_diamonds %>%
  ggplot(aes(carat,price, color = clarity)) +
  geom_point() +
#       theme_clean() +
  labs( title = "CARAT VS PRICE")
```

```
(d1 + d2)
```



2. **Relationships between Price and Cut**

```
df_diamonds %>%
  ggplot(aes(cut,price, fill = cut)) +
  geom_violin() +
  geom_jitter(color="black", size=0.05, alpha=0.9) +
#       theme_clean() +
    labs(
    title = "DIAMONDS CUT vs PRICE") +
  scale_fill_brewer(palette = 'BrBG')
```

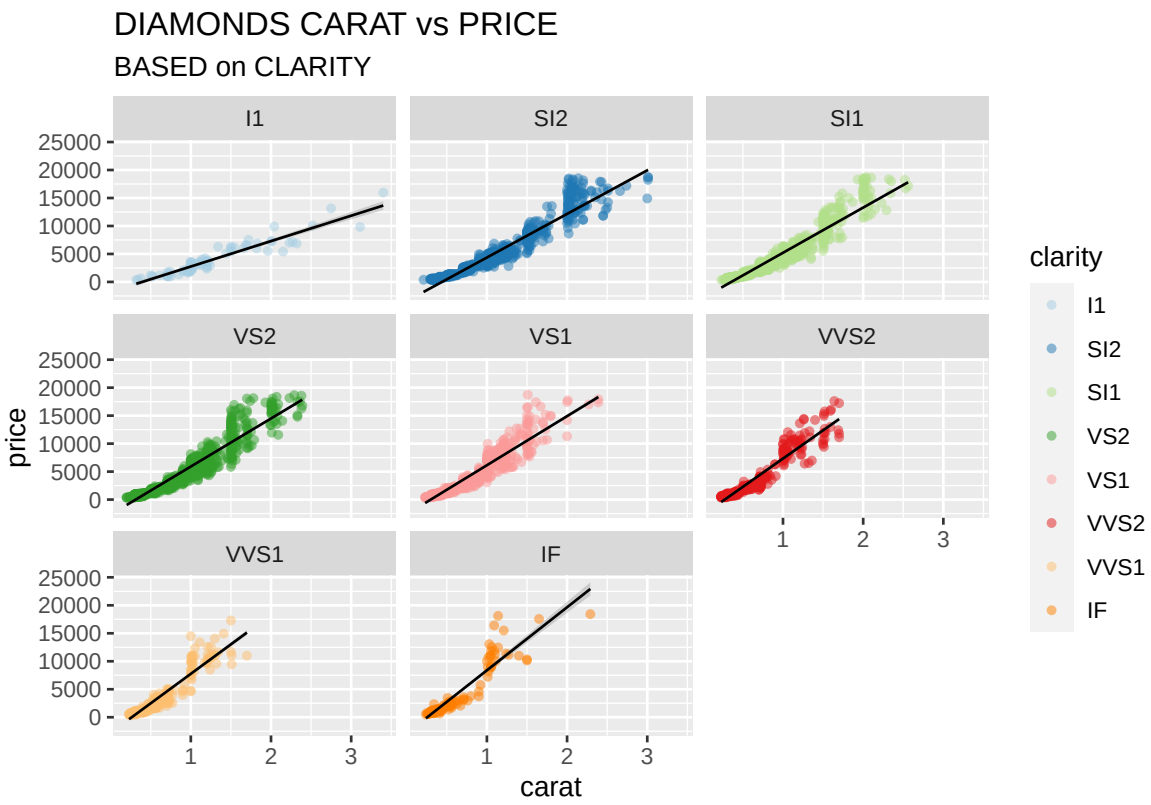


Now we see how the price of a diamond changes across different cut categories. According to the figure, there is not enough information to conclude that the price can rise and fall depending on a cut grade alone, even though a diamond cut is the quality that most significantly impacts its beauty.

3. **Relationship between Carat and Price in Different Depths and Clarity Categories**

```
df_diamonds %>%
  ggplot(aes(carat, price, color = clarity)) +
  geom_point(alpha = 0.5, pch=16)  +
  geom_smooth(method="lm",se = TRUE, color = "black",size=0.5) +
#       theme_clean() +
  facet_wrap(~clarity, nrow = 3) +
  scale_color_brewer(palette = "Paired") +
  labs(
    title = "DIAMONDS CARAT vs PRICE",
    subtitle = "BASED on CLARITY"
  )
```

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



From the figure above, it clearly shows that diamond price increases with carat weight. I1 is sold at very wide range of price and weight (carat), but not the most expensive clarity and popular. VVS2, VVS1 and IF are the most expensive based on the prices and carats that they are sold.

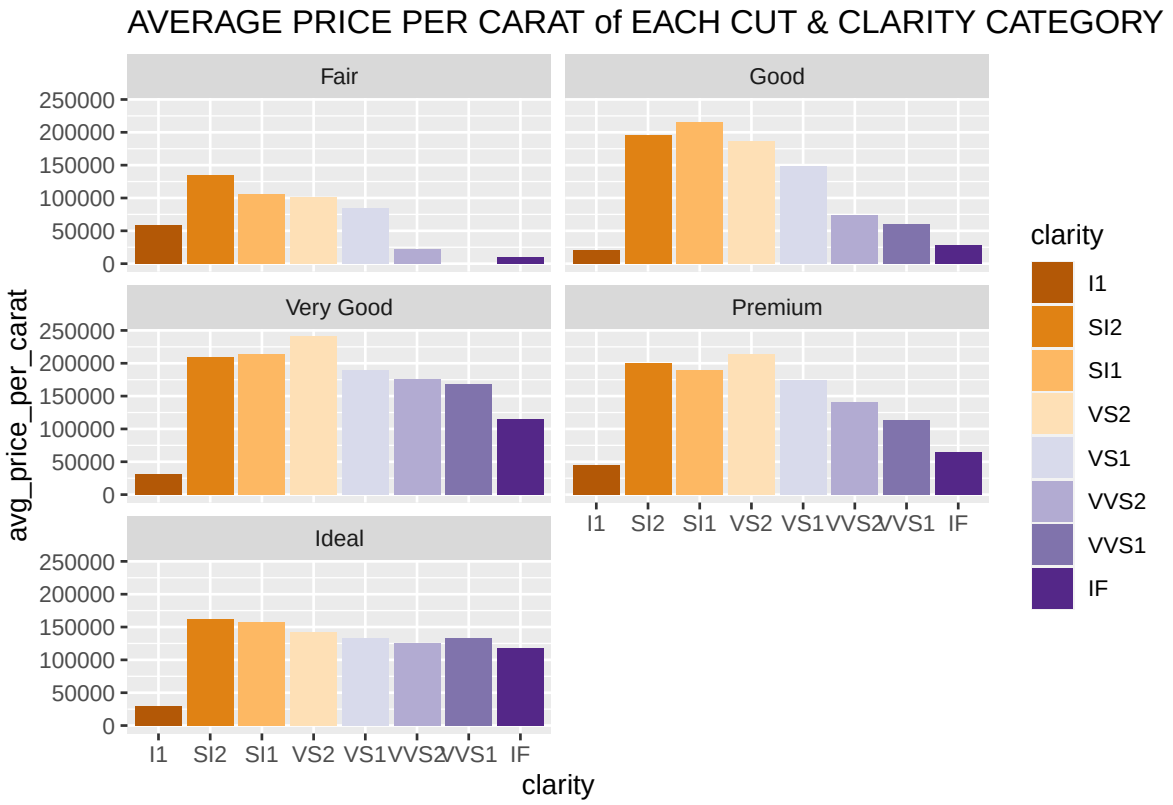
4. **Average Price per Carat of each Cut & Clarity**

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_diamonds %>%
  mutate(clarity,cut,depth, price_per_carat = price/carat) %>%
  group_by(cut,clarity, depth) %>%
  summarise(
    avg_price_per_carat = mean(price_per_carat)
  )
```

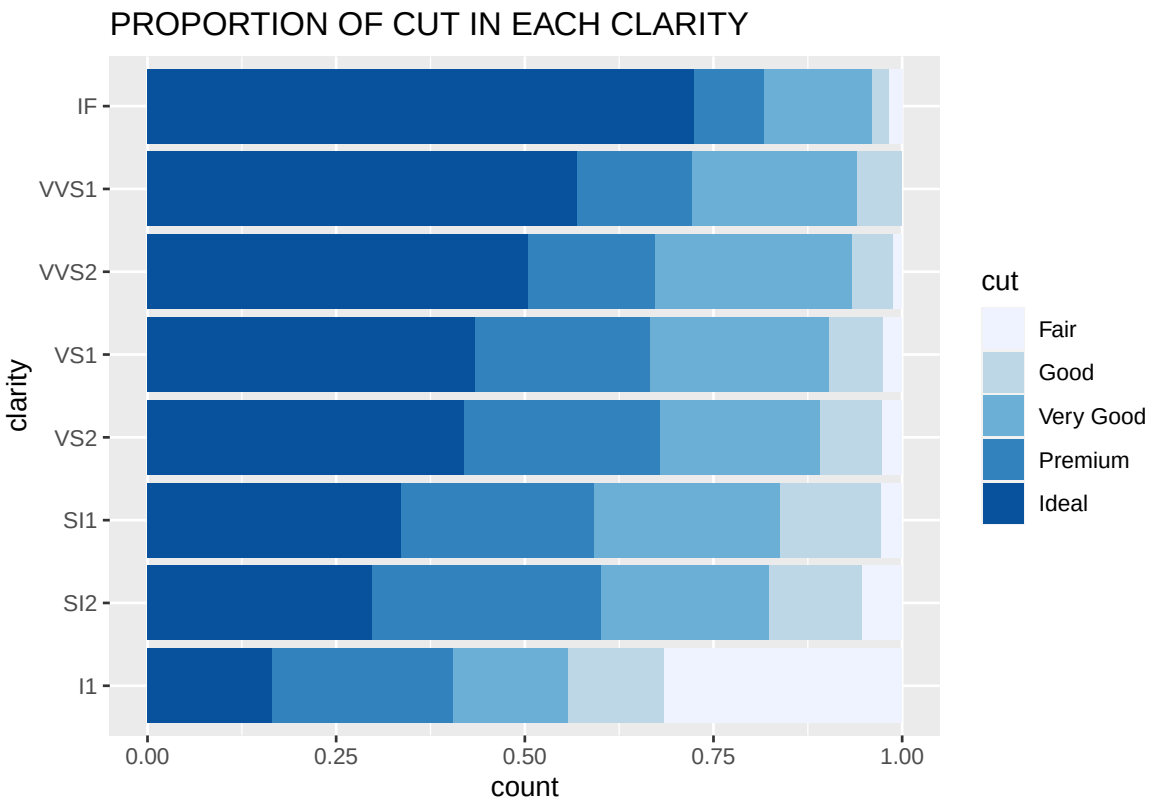
```
## `summarise()` has grouped output by 'cut', 'clarity'. You can override using
## the `.groups` argument.
```

```
price_per_carat %>%
          ggplot(aes(x = clarity, y = avg_price_per_carat, fill=clarity)) +
          geom_col() +
          facet_wrap(~cut, ncol=2) +
#       theme_clean() +
          labs(title = "AVERAGE PRICE PER CARAT of EACH CUT & CLARITY CATEGORY")+
          scale_fill_brewer(palette = 'PuOr')
```



5. **Proportion of Cut in each Clarity**

```
df_diamonds %>%
          ggplot(aes(y = clarity, fill=cut)) +
          geom_bar(position = "fill") +
#       theme_clean() +
          labs(title='PROPORTION OF CUT IN EACH CLARITY') +
          scale_fill_brewer(palette = 'Blues')
```


PROPORTION OF CUT IN EACH CLARITY
clarity
IF
VVS1
VVS2
VS1
VS2
SI1
SI2
I1
0.00
0.25
0.50
0.75
1.00
count
cut
Fair
Good
Very Good
Premium
Ideal