Data Visualization with ggplot2

Pakakorn Kaeoluan

Contents

Coach: Datarockie

I'm learning R markdown to create document

```
cat("Hello There! my name is Pakakorn\n")
```

```
## Hello There! my name is Pakakorn
```

```
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   1.0.10
## 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()
```

```
library(patchwork)
```

Prep data

1.Review

```
glimpse(diamonds)
```

```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS1, ~
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4, 64~
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62, 58~
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340, 34~
```

```
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00, 4.~
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05, 4.~
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39, 2.~
```

- Dimension: cut,color,clarity
- Numerical: carat,price,depth,table,x,y,z

2.Sampling

```
set.seed(66)
df <- sample_n(diamonds,1000)
df
```

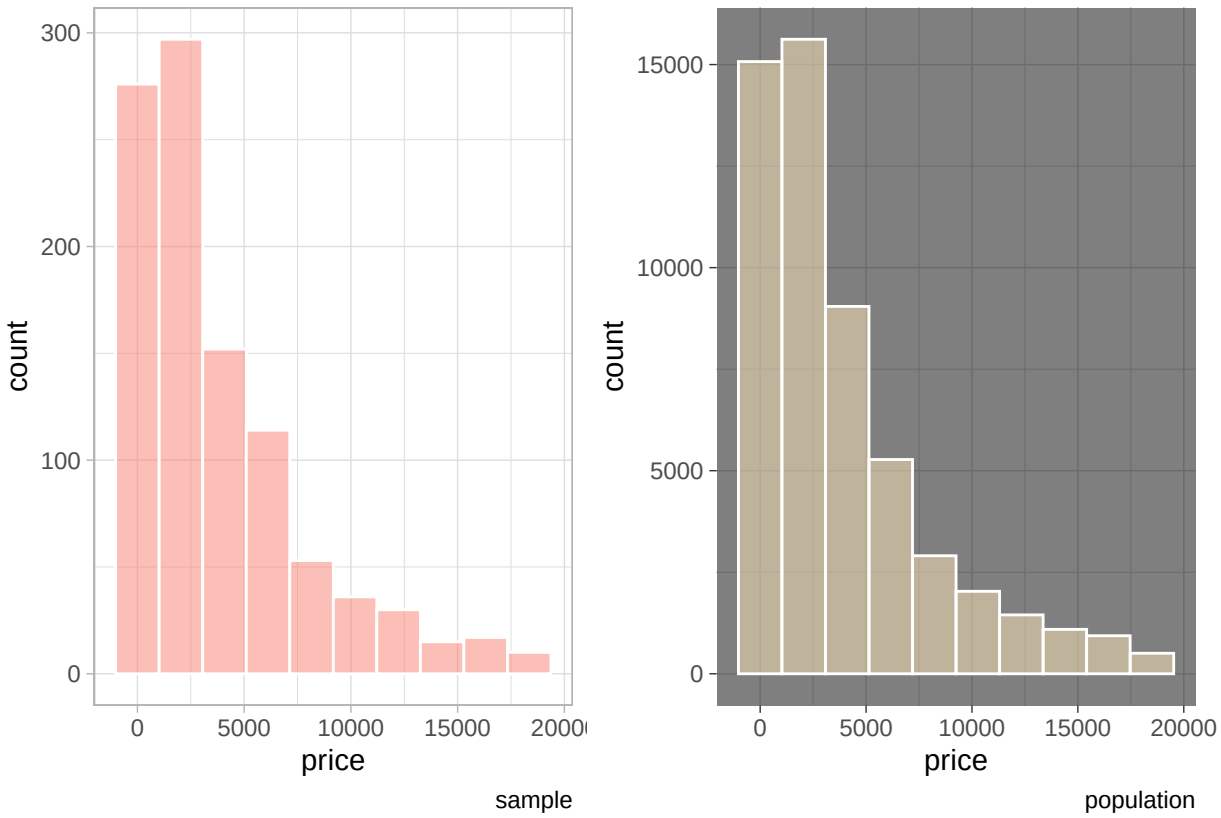
```
## # A tibble: 1,000 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  2.02 Very Good H     VS2      61.5  59   17887  8.08  8.21  5.01
##  2  2.18 Premium   I     SI1      63    58   13263  8.23  8.17  5.22
##  3  1    Premium   E     SI1      61.6  59    5500  6.38  6.41  3.94
##  4  1    Good      G     VVS2     63.1  58    7453  6.33  6.37  4.01
##  5  1.2  Premium   G     VS2      59.6  58    7258  6.94  6.89  4.12
##  6  1.61 Premium   G     SI1      62.6  58   11303  7.48  7.45  4.67
##  7  1.02 Ideal     F     VS1      61    56    8011  6.49  6.52  3.97
##  8  0.39 Ideal     I     VVS1     62.5  53.1   820  4.68  4.7   2.93
##  9  1    Good      D     VS2      62.1  64    5174  6.38  6.35  3.95
## 10  0.9  Ideal     H     SI2      62.8  55    3016  6.18  6.11  3.86
## # ... with 990 more rows
```

- sample vs population

```
viz1 <- ggplot(df,aes(price))+
  geom_histogram(bins = 10,alpha = 0.5,fill = "salmon",color = "white")+
  theme_light()+
  labs(caption = "sample")

viz2 <- ggplot(diamonds,aes(price))+
  geom_histogram(bins = 10,alpha = 0.5,fill = "wheat1",color = "white")+
  theme_dark()+
  labs(caption = "population")

viz1+viz2
```

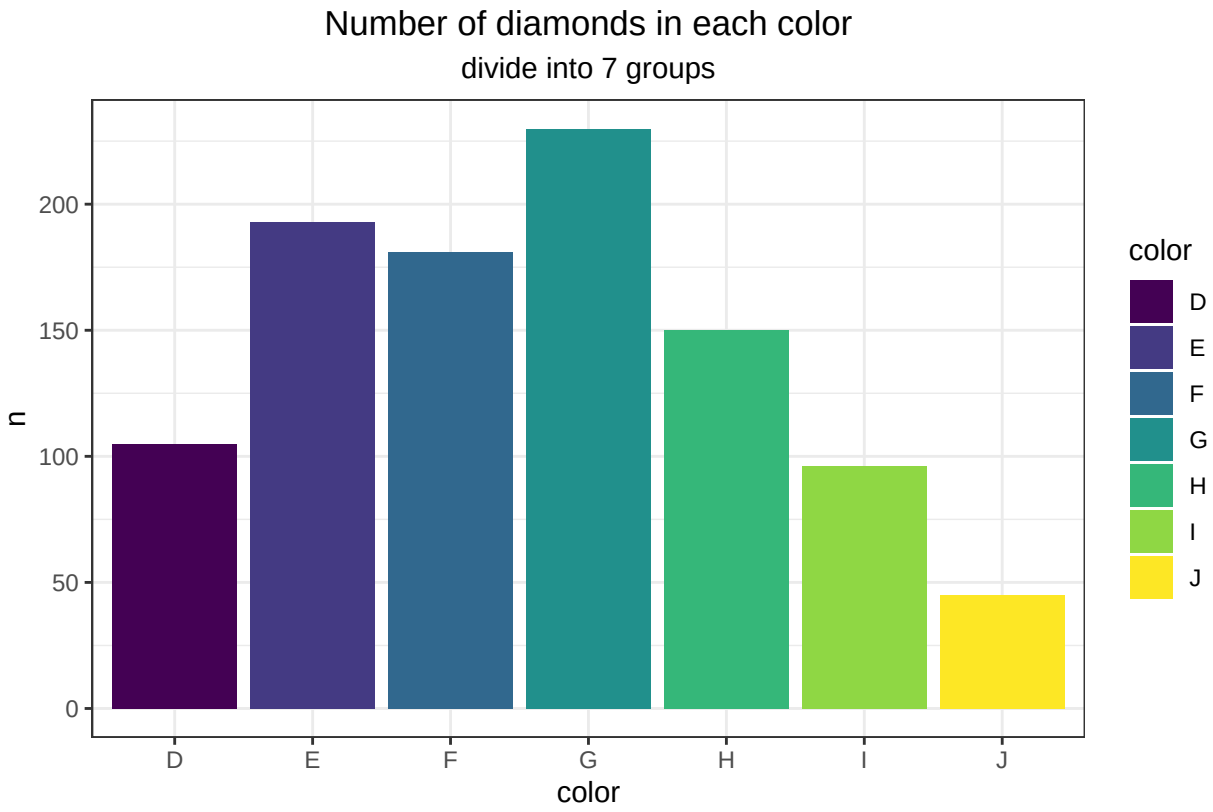


Visualization

1. One variable

- 1.1 One variable - discrete

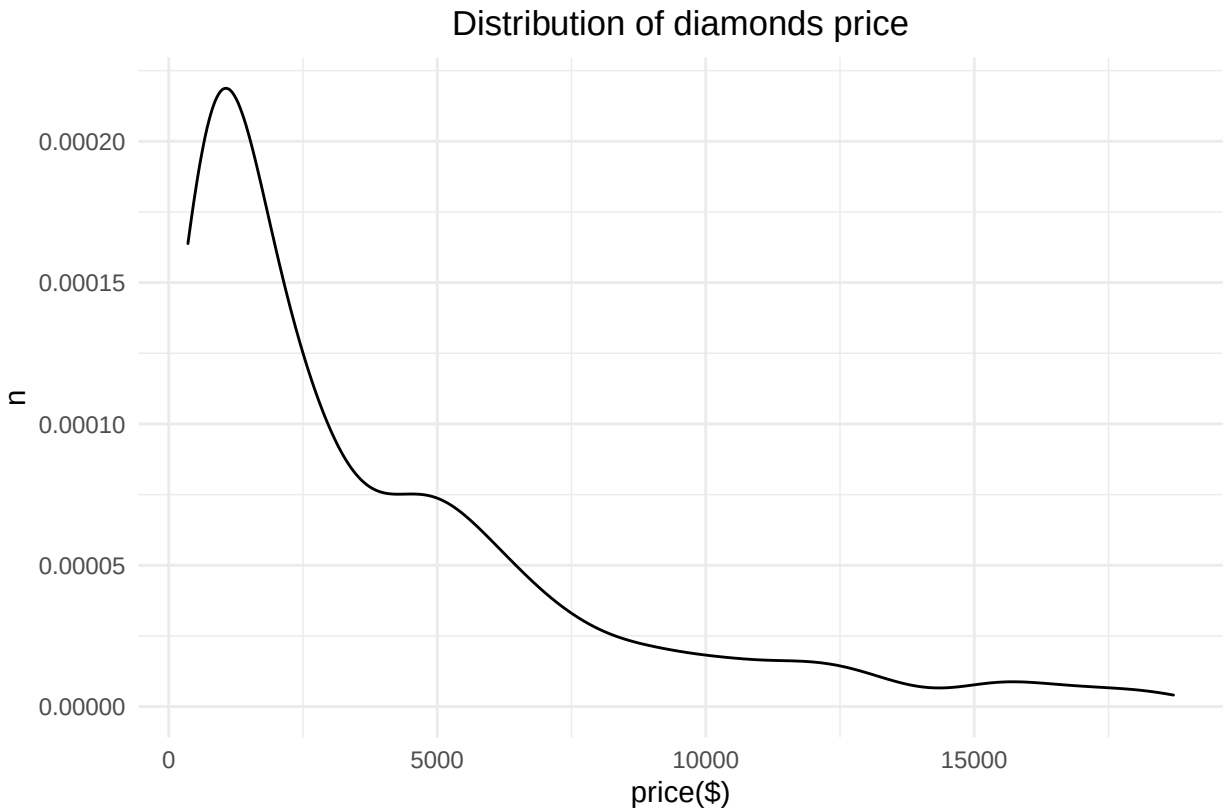
```
df%>%
  ggplot(aes(color,fill=color))+
  geom_bar()+
  theme_bw()+
  labs(x="color",y="n",title = "Number of diamonds in each color",subtitle = "divide into 7 groups",cap
  theme(plot.caption=element_text(hjust=0.5),
        plot.title=element_text(hjust=0.5),
        plot.subtitle=element_text(hjust=0.5))
```



As above figure, The most diamonds is G group

- 1.2 One variable - continuous

```
df%>%
  ggplot(aes(price))+
  geom_density()+
  theme_minimal()+
  labs(x="price($)",y="n",title = "Distribution of diamonds price",caption = "As above figure,it shown s
  theme(plot.caption=element_text(hjust=0.5),
        plot.title=element_text(hjust=0.5))
```



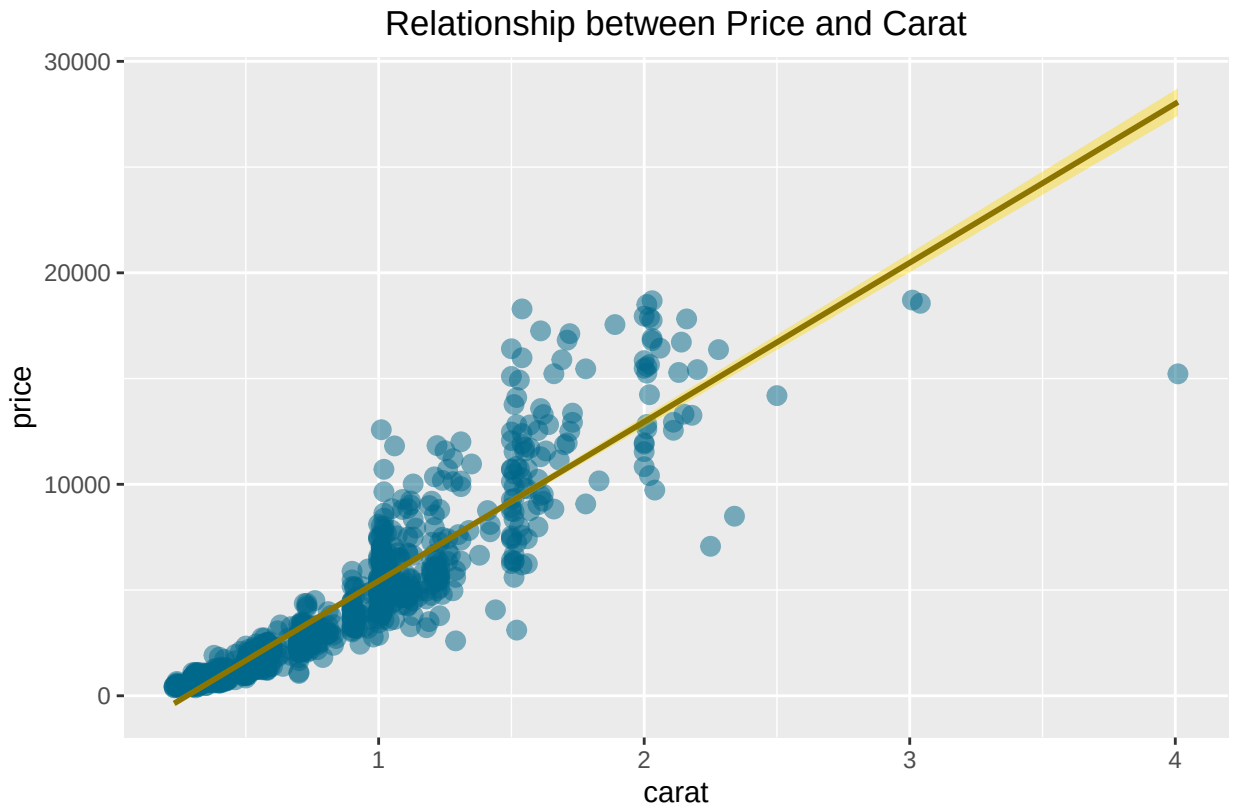
As above figure,it shown skewed distribution

2. Two variable

- 2.1 Two variable - continuous

```
df%>%
  ggplot(aes(carat,price))+
  geom_point(color="deepskyblue4",alpha=0.5,size=3)+
  geom_smooth(color="gold4",fill="gold",method="lm")+
  theme_gray()+
  labs(title = "Relationship between Price and Carat",caption = "As above graph, the relationship betwe
  theme(plot.caption=element_text(hjust=0.5),
        plot.title=element_text(hjust=0.5))
```

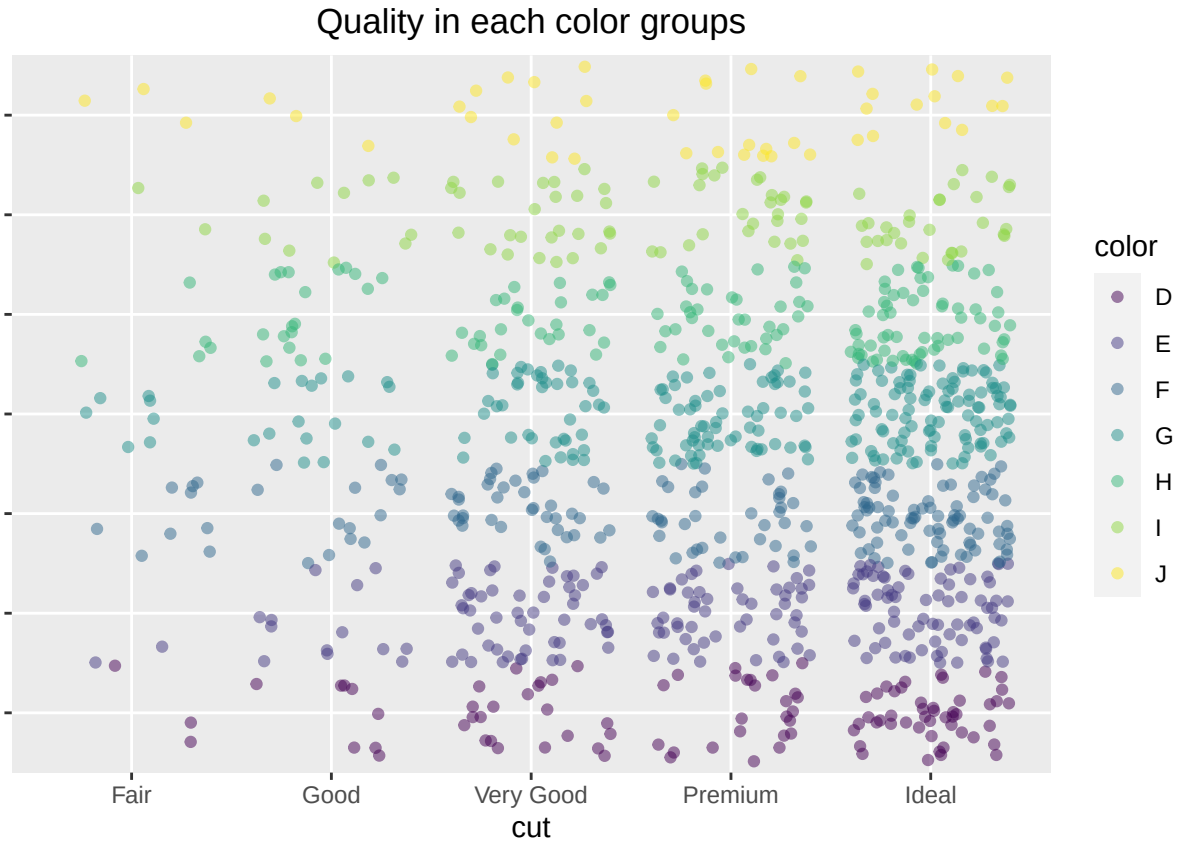
```
## `geom_smooth()` using formula 'y ~ x'
```



As above graph, the relationship between price and carat is positive

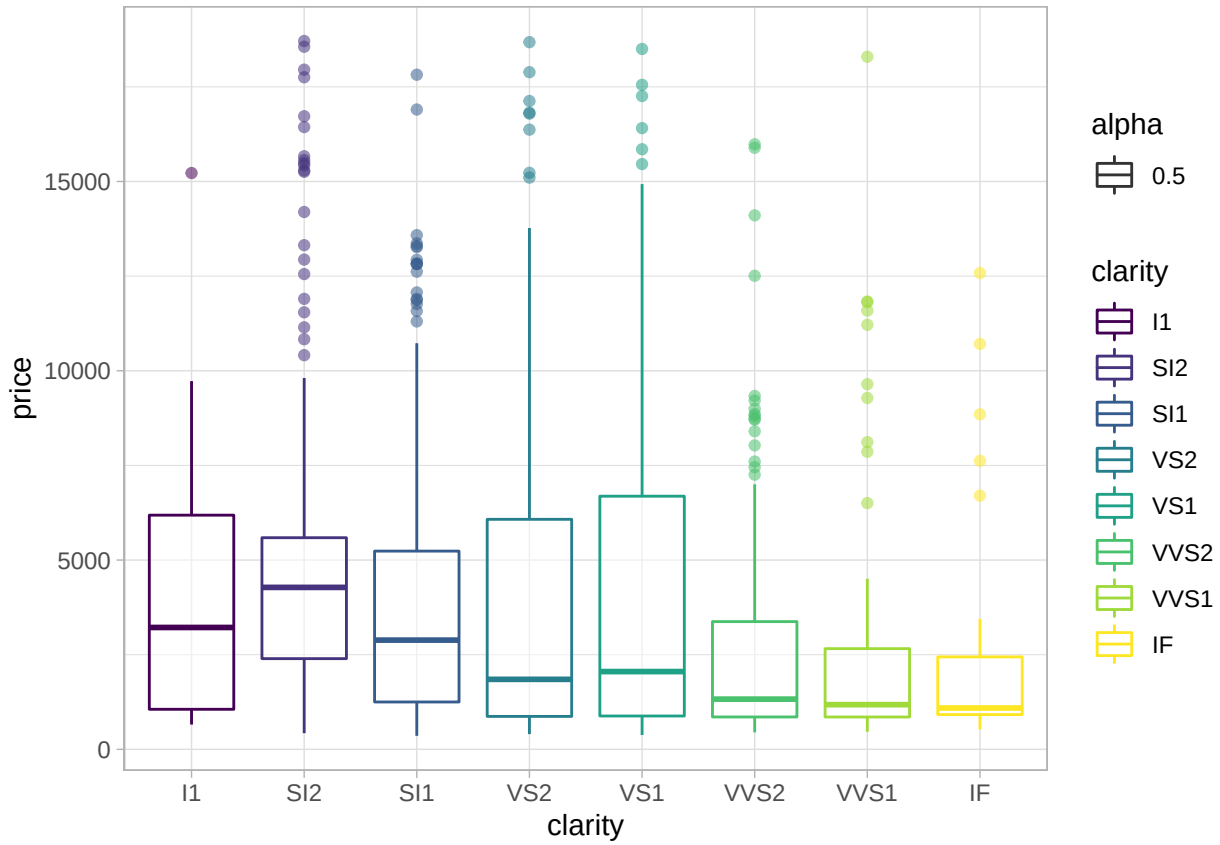
- 2.2 Two variable - discrete

```
df%>%
ggplot(aes(cut,color))+
geom_jitter(alpha=0.5,height = 0.5,aes(color=color))+
theme(axis.text.y=element_blank())+
labs(title = "Quality in each color groups",y = "")+
theme(plot.title=element_text(hjust=0.5))
```



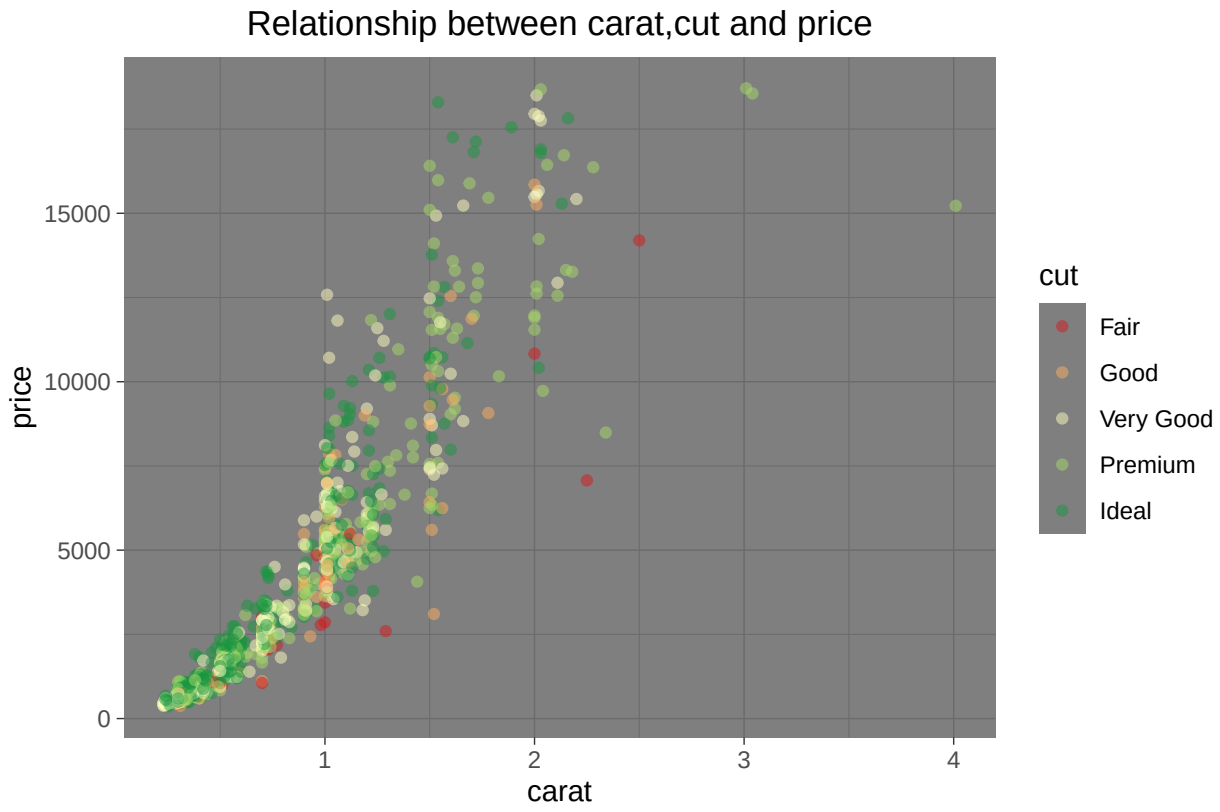
- 2.3 Two variable - continuous/discrete

```
qplot(clarity,price,color=clarity,data = df,geom = "boxplot",alpha=0.5)+
theme_light()
```



3.three var

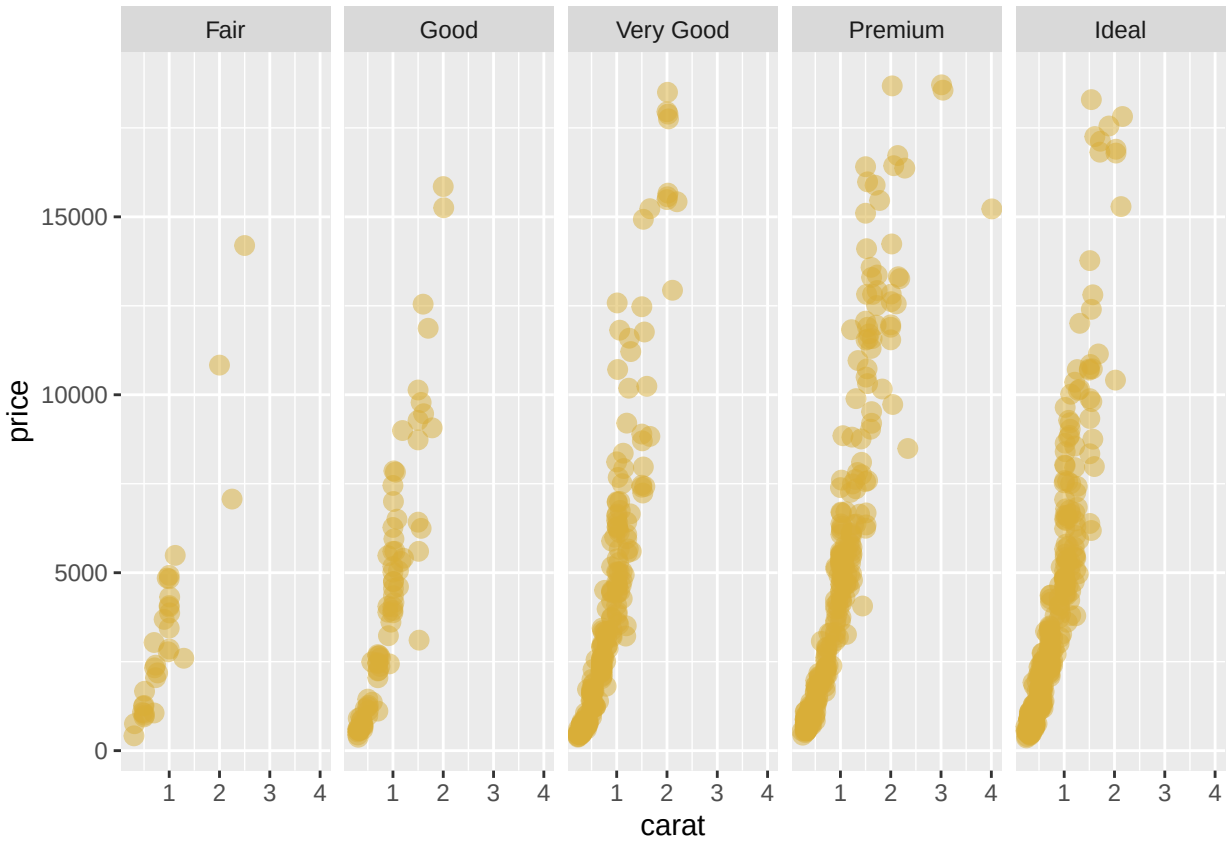
```
df%>%
ggplot(aes(carat,price,color=cut))+
geom_point(alpha=0.5)+
scale_color_brewer(palette = "RdYlGn")+
theme_dark()+
labs(title = "Relationship between carat,cut and price" ,caption ="carat is more important factor thar
theme(plot.caption=element_text(hjust=0.5),
      plot.title=element_text(hjust=0.5))
```



carat is more important factor than cut

4.Face wrap

```
df %>%
  ggplot(aes(carat,price))+
  geom_point(color="#d9ae38",alpha=0.5,size=3)+
  facet_wrap(~cut,ncol=5)
```


Fair
Good
Very Good
Premium
Ideal
price
15000
10000
5000
0
1
2
3
4
carat