Exploratory Data Analysis in R Programing

Sirinthip Ngamchaluay

Using Diamonds dataset in R package.

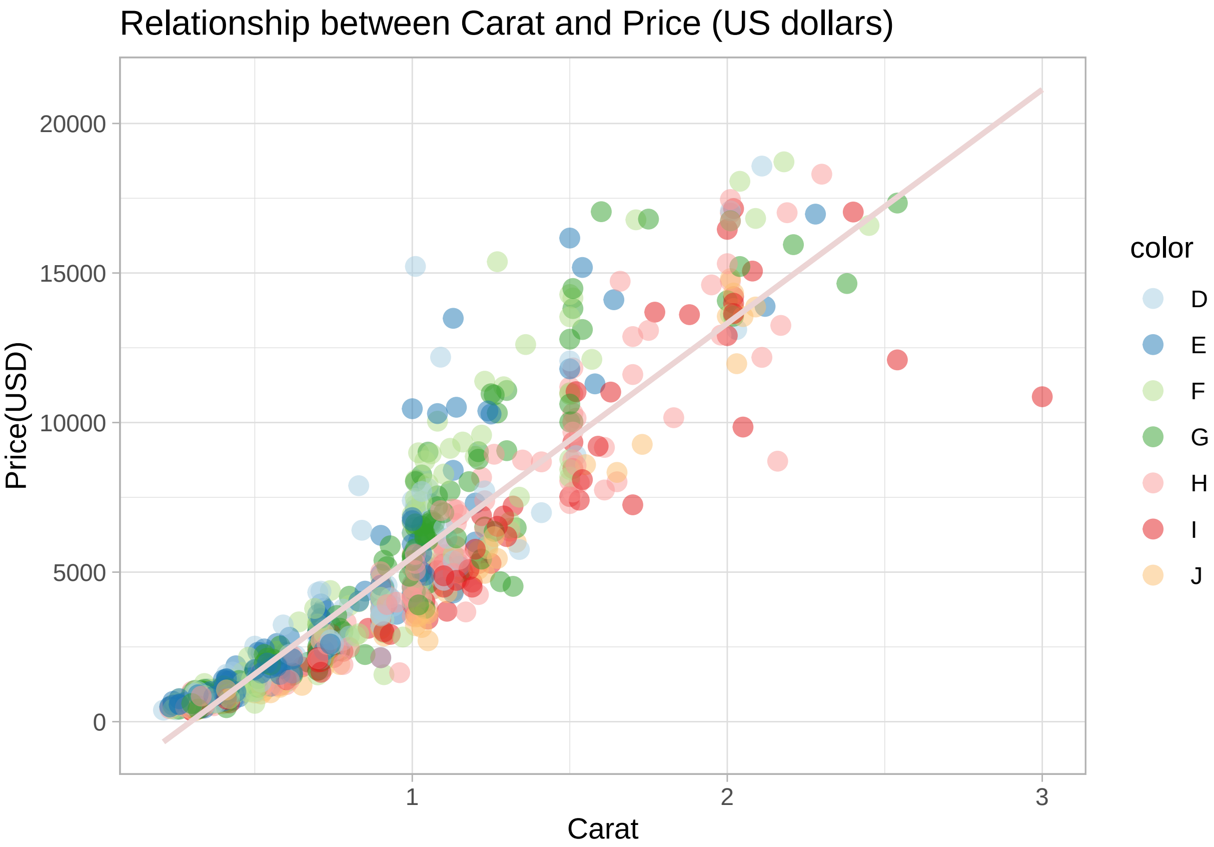
- ## Relationship between Carat Weight and Price

```r
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()
```

```r
library(dplyr)
library(ggplot2)
library(ggthemes)
library(RColorBrewer)
set.seed(99)
small_diamonds <- sample_n(diamonds,1000)
ggplot(small_diamonds, aes(carat,price))+
  geom_point(size = 3, alpha = 0.5, aes(color = color))+
  geom_smooth(method = "lm",color = "#ECD4D4",se=F)+
  labs(title = "Relationship between Carat and Price (US dollars)",
          x = "Carat",
          y = "Price(USD)")+
  theme_light()+
  scale_color_brewer(type = "qual", palette = 3)
```

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

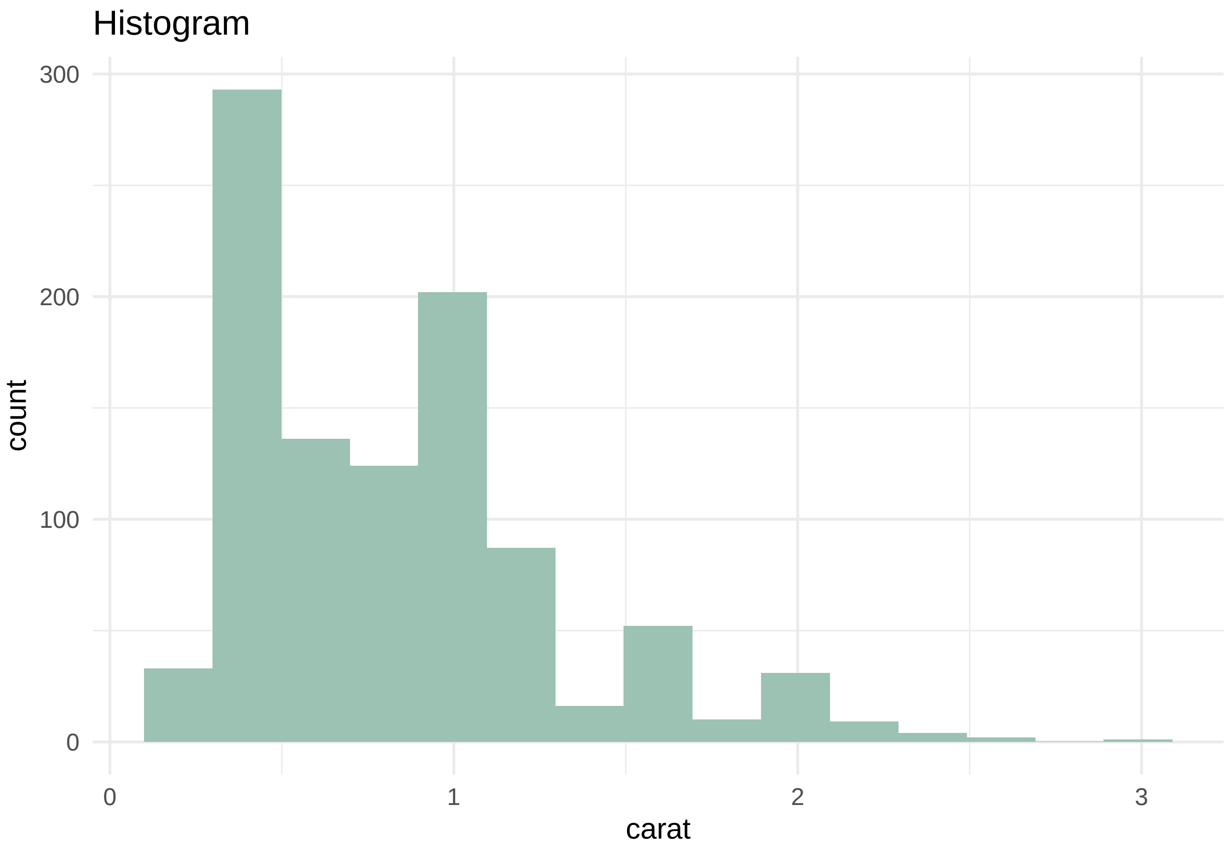


•

Price and carat correlate in the same direction, if the carat weight increases pricing will also be high.

- **Histogram of Carat Weight**

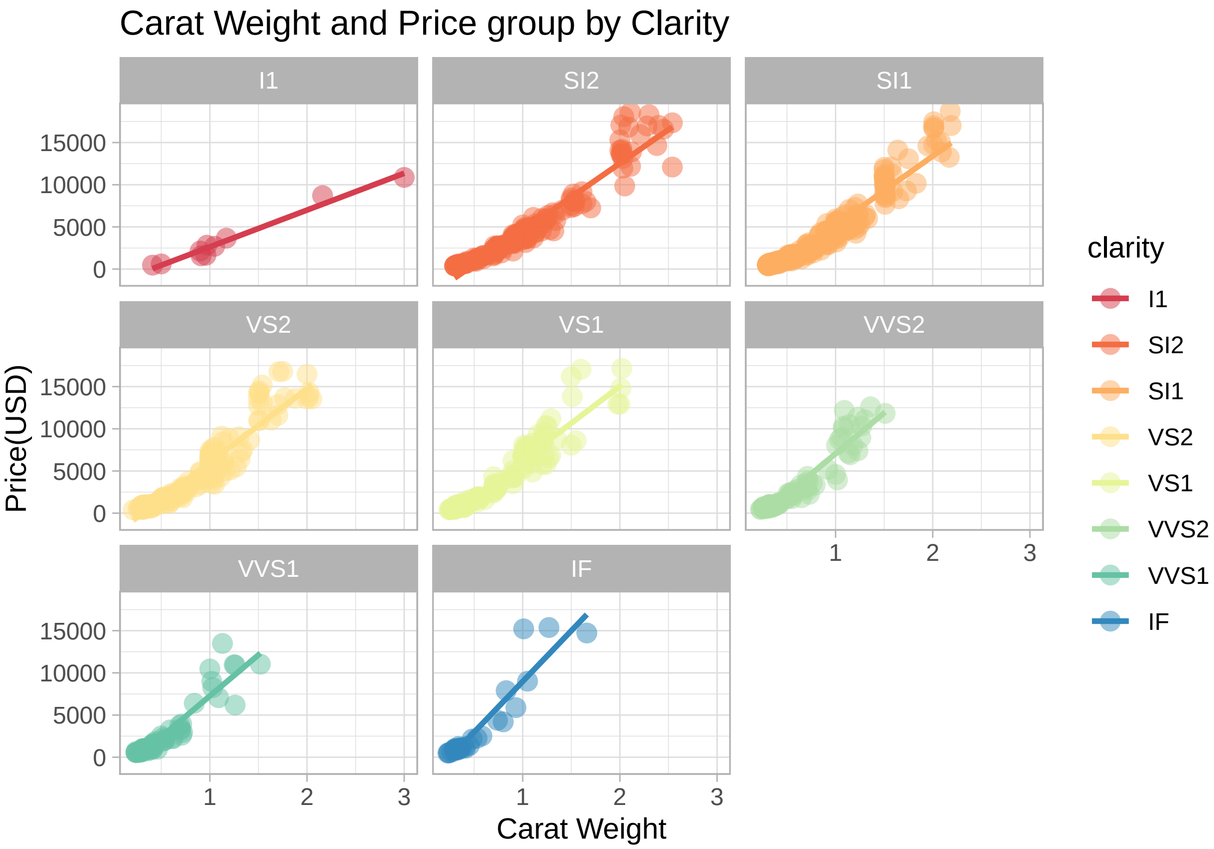
```
ggplot(small_diamonds, aes(carat))+
  geom_histogram(bins = 15, fill = "#9BC2B2")+
  labs(title = "Histogram")+
  theme_minimal()
```



- **Carat Weight vs. Price group by Clarity**

```
ggplot(small_diamonds, aes(carat,price,color=clarity))+
  geom_point(size=3,alpha=0.5)+
  geom_smooth(method = "lm",se=F)+
  labs(title = "Carat Weight and Price group by Clarity",
          x = "Carat Weight",
          y = "Price(USD)")+
  facet_wrap(~clarity,ncol = 3)+
  theme_light()+
  scale_color_brewer(palette = "Spectral")
```

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

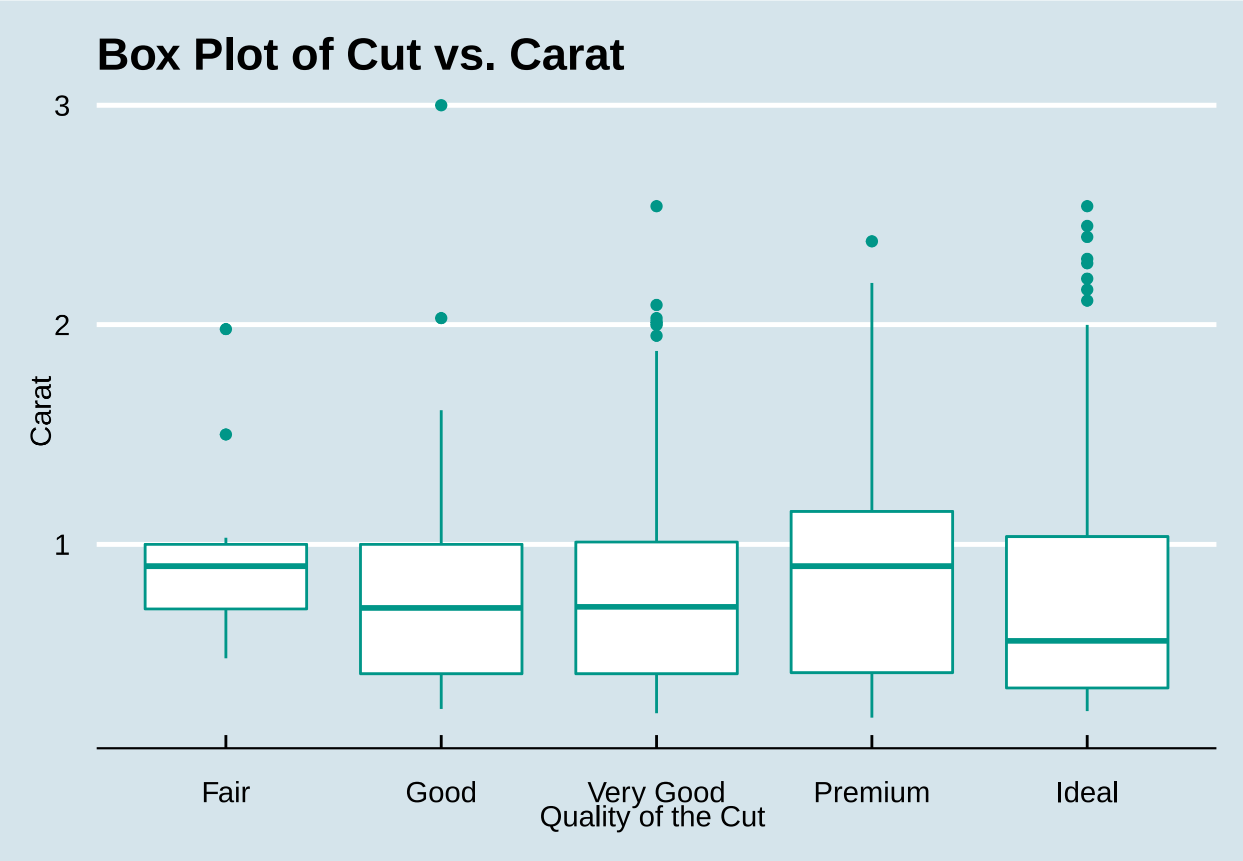


-

Price varies according to clarity and carat weight. IF (best) and I1 (worst).

- **Box Plot of Carat Weight and Quality of the Cut**

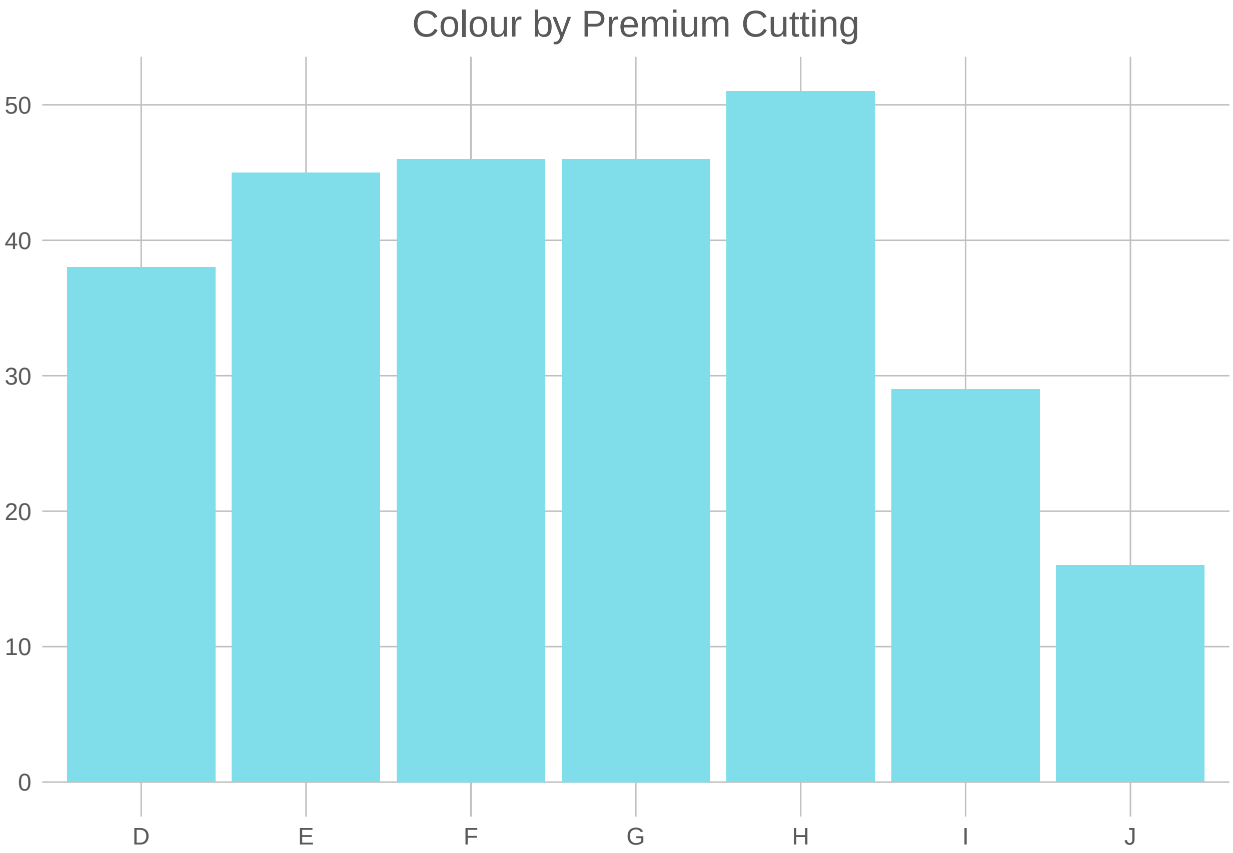
```
ggplot(small_diamonds, aes(cut,carat))+
  labs(title = "Box Plot of Cut vs. Carat",
         x = "Quality of the Cut ",
         y = "Carat")+
  geom_boxplot(color = "#009688")+
  theme_economist()
```



- Diamond Weight (between 0.2 to 5.01), the measure of positions of Premium cut greater than others cut.

- **Premium Cutting Diamond by Colour**

```
filter_data <- small_diamonds %>% filter(cut == "Premium")
ggplot(filter_data, aes(color))+
  labs(title = "Colour by Premium Cutting",
           x = "Diamond Colour ",
           y = "n")+geom_bar(fill = "#80deea")+
  theme_excel_new()
```



- The most frequency is H colour. However, the overall of frequency approaches the colour (D,E,F) are expensive and rare diamonds.