Exploratory Data Analysis

Budsadee Sareerasart

Install Packages

```
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(ggplot2)
library(patchwork)
```

Review Data

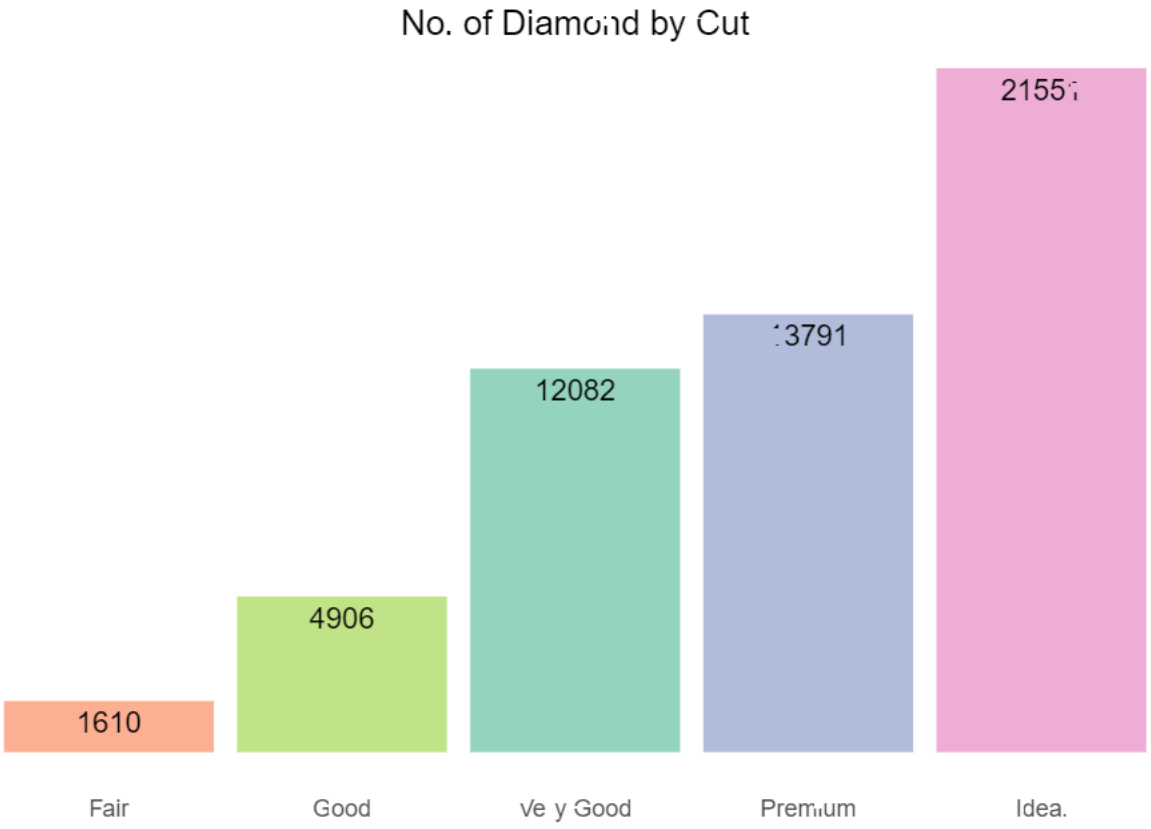
```
head(diamonds)
```

```
## # 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
```

(1) One Variable - Discrete

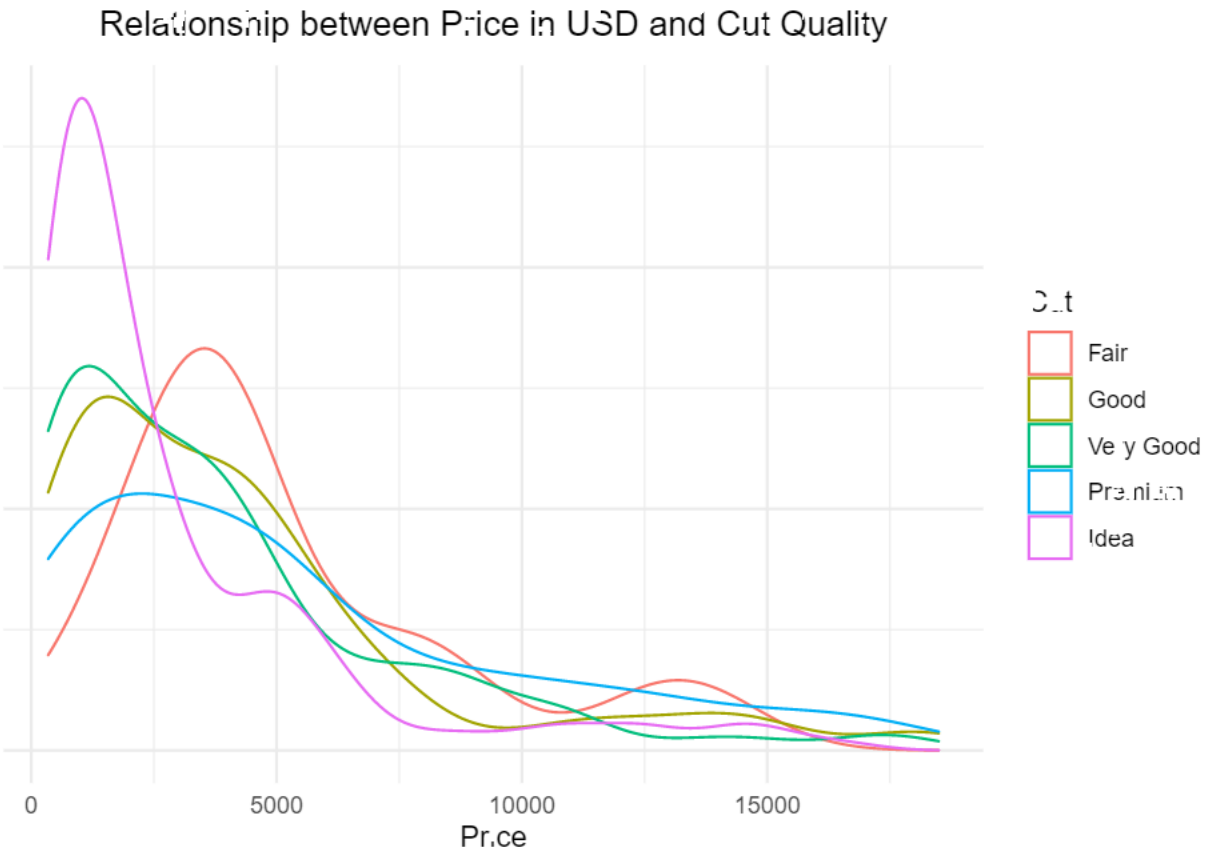
```
## total n_diamonds
ggplot(diamonds, aes(cut, fill = cut)) +
  geom_bar(alpha = 0.7) +
  scale_fill_manual(values = c("#fc8d62",
                               "#a6d854",
                               "#66c2a5",
                               "#8da0cb",
                               "#e78ac3")) +
  geom_text(stat = 'count', aes(label = ..count..),vjust = 1.5) +
  theme_minimal() +
  labs(title = "No. of Diamond by Cut") +
  theme(panel.grid.major = element_blank(),
```

```
        panel.grid.minor = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        legend.position = "none",
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        plot.title = element_text(hjust = 0.5, vjust = -2)) -> a

a
```



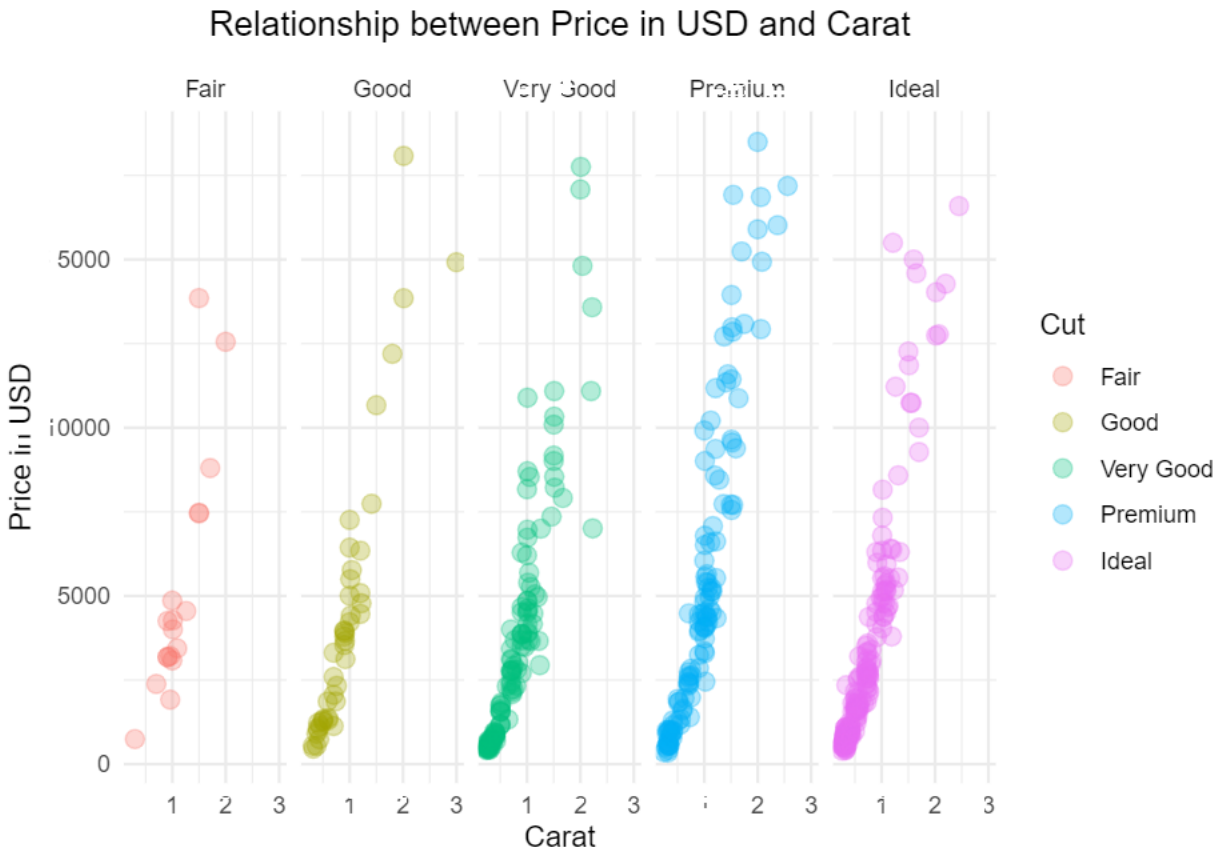
(2) Two Variable - Discrete x Continuous

```
## sample 500_diamonds
set.seed(42)
ggplot(sample_n(diamonds, 500), aes(price, colour = cut)) +
  geom_density() +
  theme_minimal() +
  labs(title = "Relationship between Price in USD and Cut Quality",
       x = "Price") +
  theme(plot.title = element_text(hjust = 0.5, vjust = 2),
        axis.title.y = element_blank(),
        axis.text.y = element_blank()) +
  scale_colour_discrete(name="Cut") -> b

b
```



(3) Two Variable - Continuous x Continuous

```
## sample 500_diamonds
set.seed(42)
ggplot(sample_n(diamonds, 500), aes(carat, price, color = cut)) +
  geom_point(size = 3, alpha = 0.3) +
  facet_wrap(~ cut, ncol = 53) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, vjust = 2)) +
  labs(title = "Relationship between Price in USD and Carat",
       x = "Carat",
       y = "Price in USD") +
  scale_colour_discrete(name="Cut") -> c

c
```

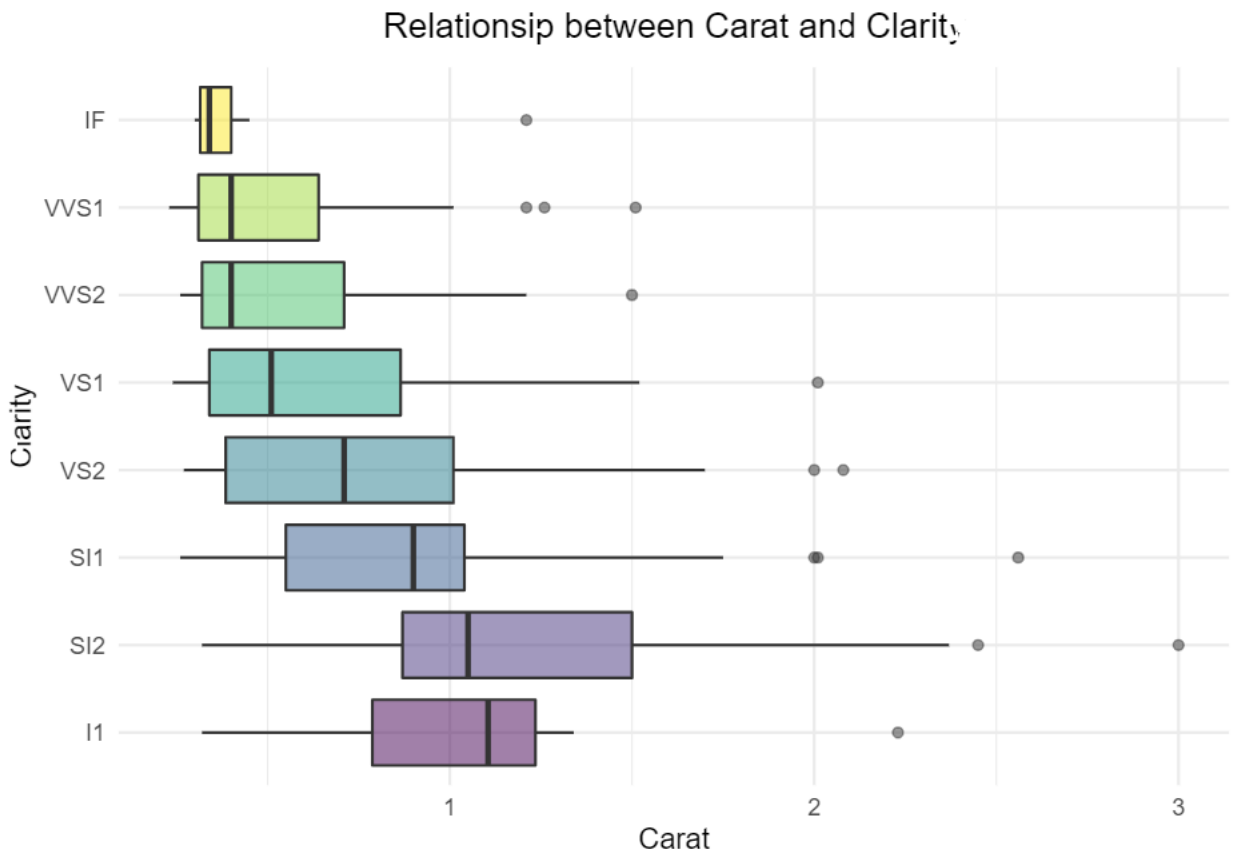


(4) Two Variable

```
## sample 100_diamonds & group price by mean
sample_n(diamonds, 100) %>%
  mutate(group_price = factor(if_else(price > round(mean(diamonds$price)), "Above AVG", "Below AVG"))) 

ggplot(sample_diamonds, aes(group_price, price, fill = group_price)) +
  geom_violin(alpha = 0.5) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, vjust = 2)) +
  labs(title = "Price allocated by mean of price",
       x = "Group_Price",
       y = "Price in USD") +
  theme(legend.position = "none") -> d

d
```

(5) Two Variable

```
## sample 100_diamonds
set.seed(42)
ggplot(sample_n(diamonds, 500), aes(carat, clarity, fill = clarity)) +
  geom_boxplot(alpha = 0.5) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, vjust = 2)) +
  labs(title = "Relationsip between Carat and Clarity",
       x = "Carat",
       y = "Clarity") +
  theme(legend.position = "none") -> e

e
```


Relationsip between Carat and Clarity
IF
VVS1
VVS2
VS1
VS2
SI1
SI2
I1
Clarity
1
2
3
Carat