HW1_Data_Visualization_Diamonds

Arpagorn Kleawsinak

2023-11-23

Data Visualization

HW: Use diamonds dataset to creat 5 charts

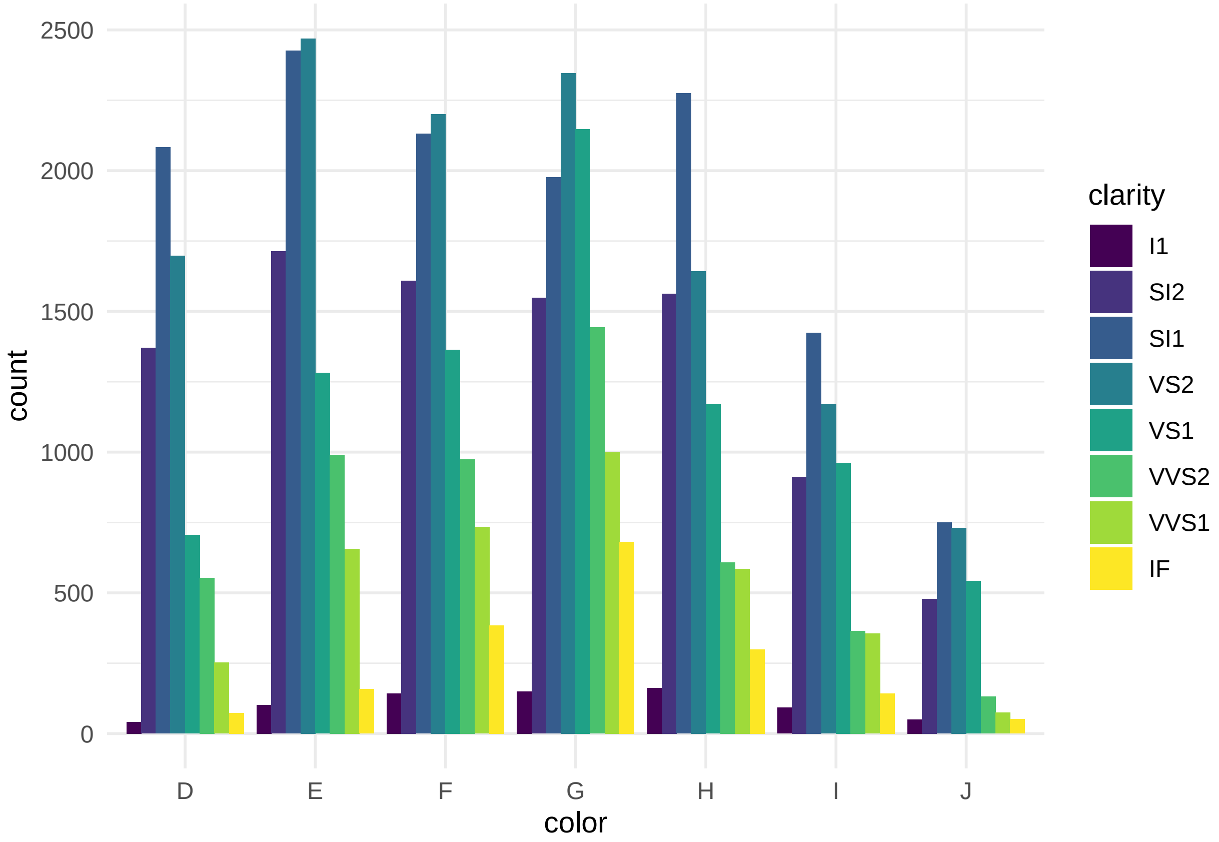
Diamonds table

```
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. Which color of diamond has the most IF clarity?

```
ggplot(diamonds, aes(color, fill = clarity))+
  geom_bar(position = "dodge")+
  theme_minimal()
```



The chart above displays distribution of clarity across all colors of diamonds. The bars for each clarity are stacked within each color. The proportion of **IF** is higher among **G, F**, and **H** colors.

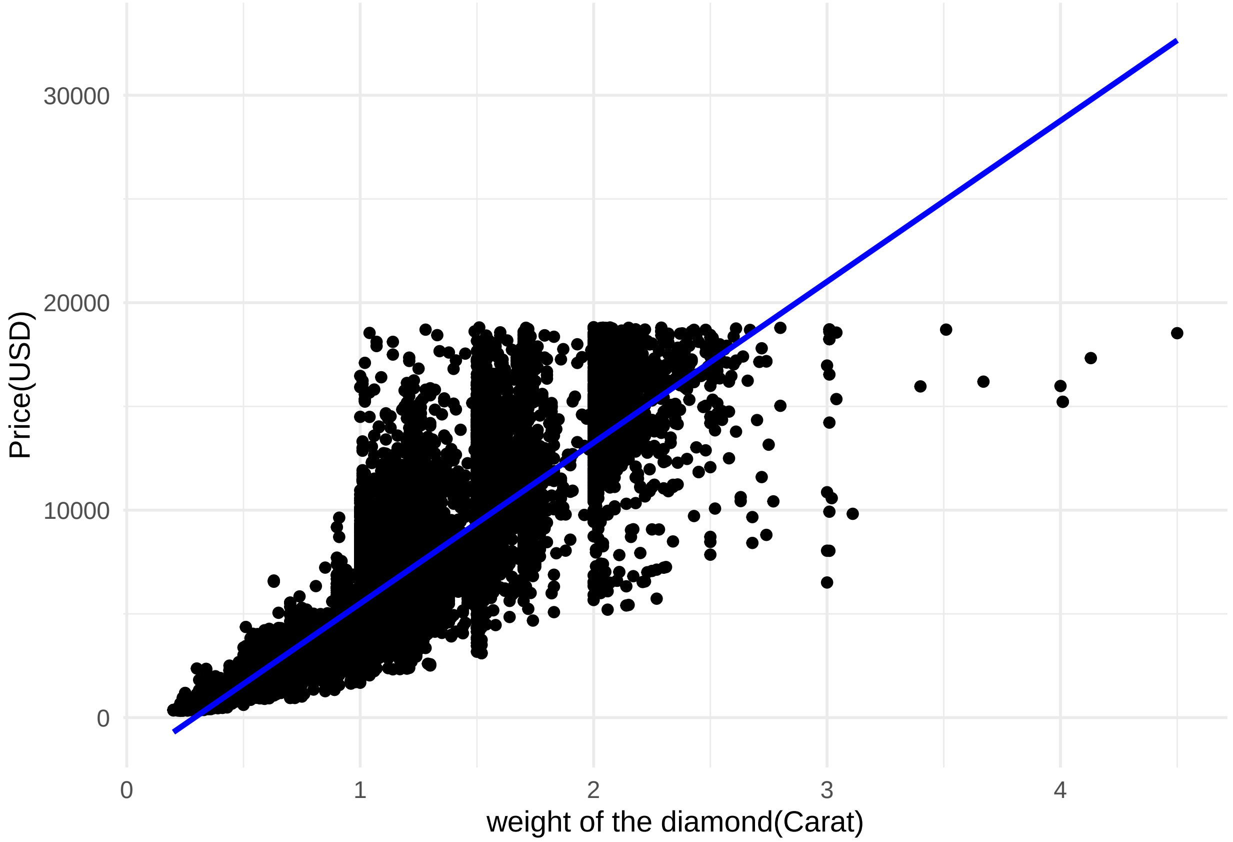
```
diamonds%>%
  filter(clarity == "IF")%>%
  group_by(color)%>%
  count(clarity)%>%
  arrange(desc(n))%>%
  head()
```

```
## # A tibble: 6 x 3
## # Groups:   color [6]
##   color clarity     n
##   <ord> <ord>   <int>
## 1 G     IF        681
## 2 F     IF        385
## 3 H     IF        299
## 4 E     IF        158
## 5 I     IF        143
## 6 D     IF         73
```

2.What is the correlation between carat and the price?

```
set.seed(24)
min_diamonds <- sample_frac(diamonds, 0.6)
```

```
ggplot(min_diamonds, aes(carat, price)) +
  geom_point()+
  geom_smooth(method = lm, col = "blue", formula = "y ~ x")+
  labs(x="weight of the diamond(Carat)", y="Price(USD)")+
  theme_minimal()
```

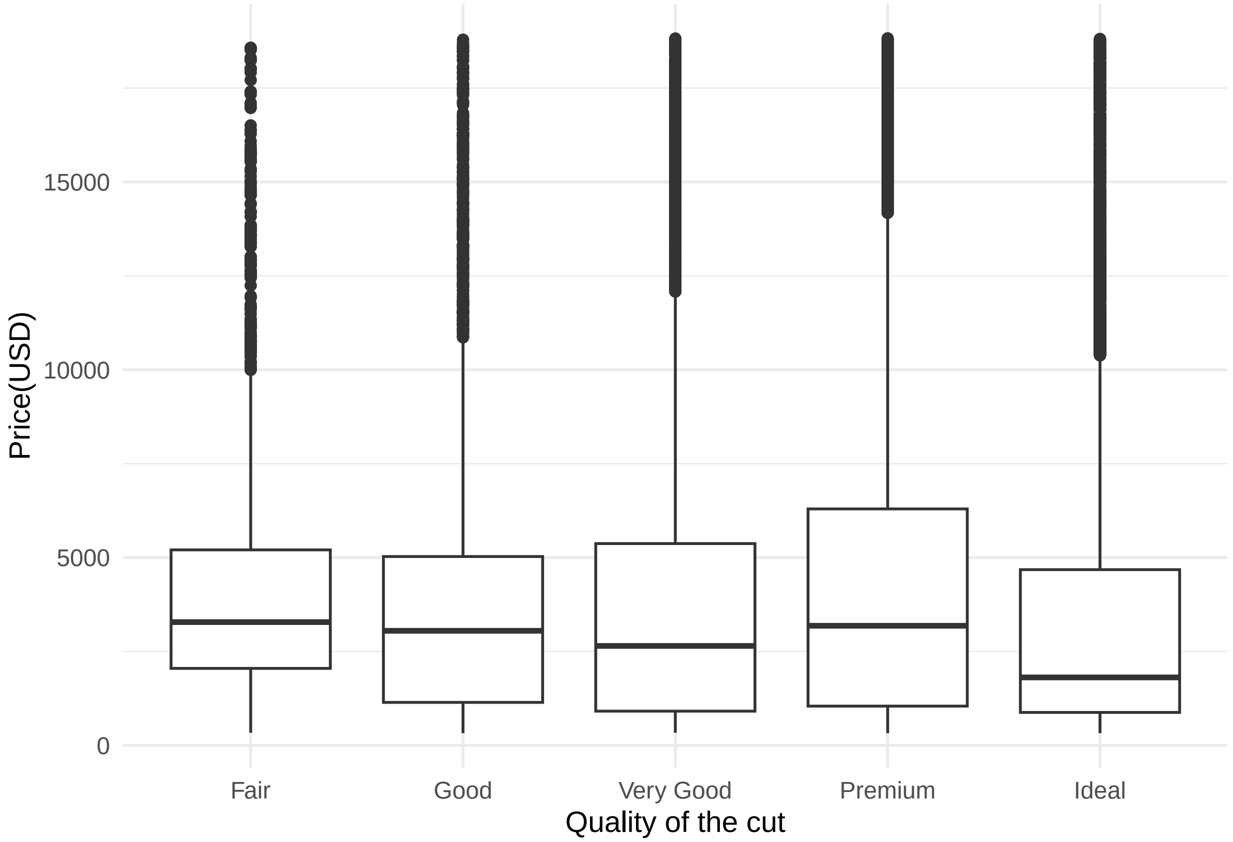


```
## `geom_smooth()` using formula = 'y ~ x'
```

The graph displays a positive correlation between carat and price, indicating that as the weight of the diamonds increases, so does their price.

3. What is the correlation between cut and the price?

```
ggplot(diamonds, aes(cut, price))+
  geom_boxplot()+
  labs(x="Quality of the cut", y="Price(USD)")+
  theme_minimal()
```



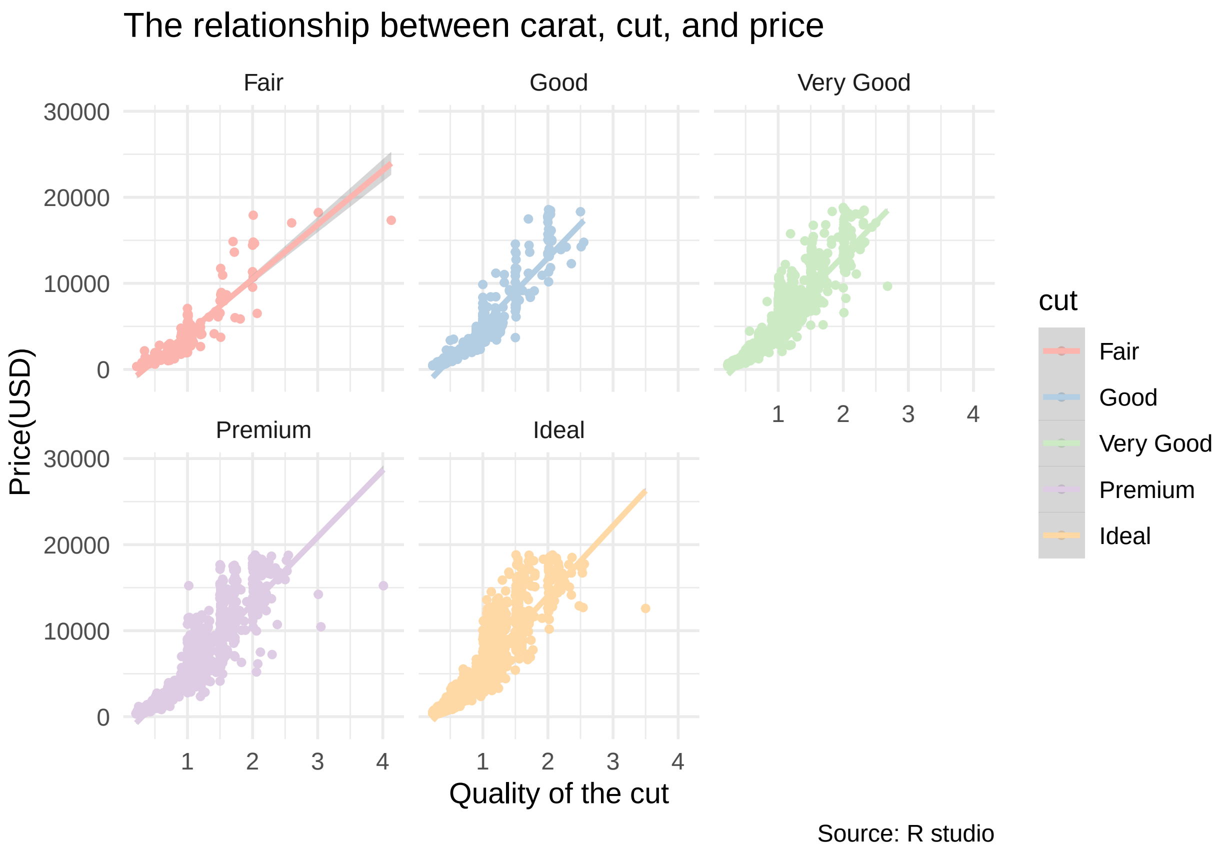
The graph indicates that the premium cut has highest distribution. The median price is highest for fair, premium, and good cuts, respectively. The diamonds with an ideal cut have the lowest median price.

```
diamonds%>%
  group_by(cut)%>%
  summarise(med_price = median(price))%>%
  arrange(desc(med_price))
```

```
## # A tibble: 5 x 2
##   cut       med_price
##   <ord>         <dbl>
## 1 Fair          3282
## 2 Premium       3185
## 3 Good          3050.
## 4 Very Good     2648
## 5 Ideal         1810
```

4.What is the correlation among carat, cut and price?

```
set.seed(72)
min_diamonds <- sample_n(diamonds, 7000)
ggplot(min_diamonds, aes(carat, price, col = cut)) +
  geom_point(size= 1)+
  geom_smooth(method = "lm", formula = "y ~ x")+
  theme_minimal()+
  scale_color_brewer(type ="qua", palette = 4)+
  labs(x = "Quality of the cut",
       y="Price(USD)",
       caption= "Source: R studio",
       title = "The relationship between carat, cut, and price")+
  facet_wrap(~cut)
```



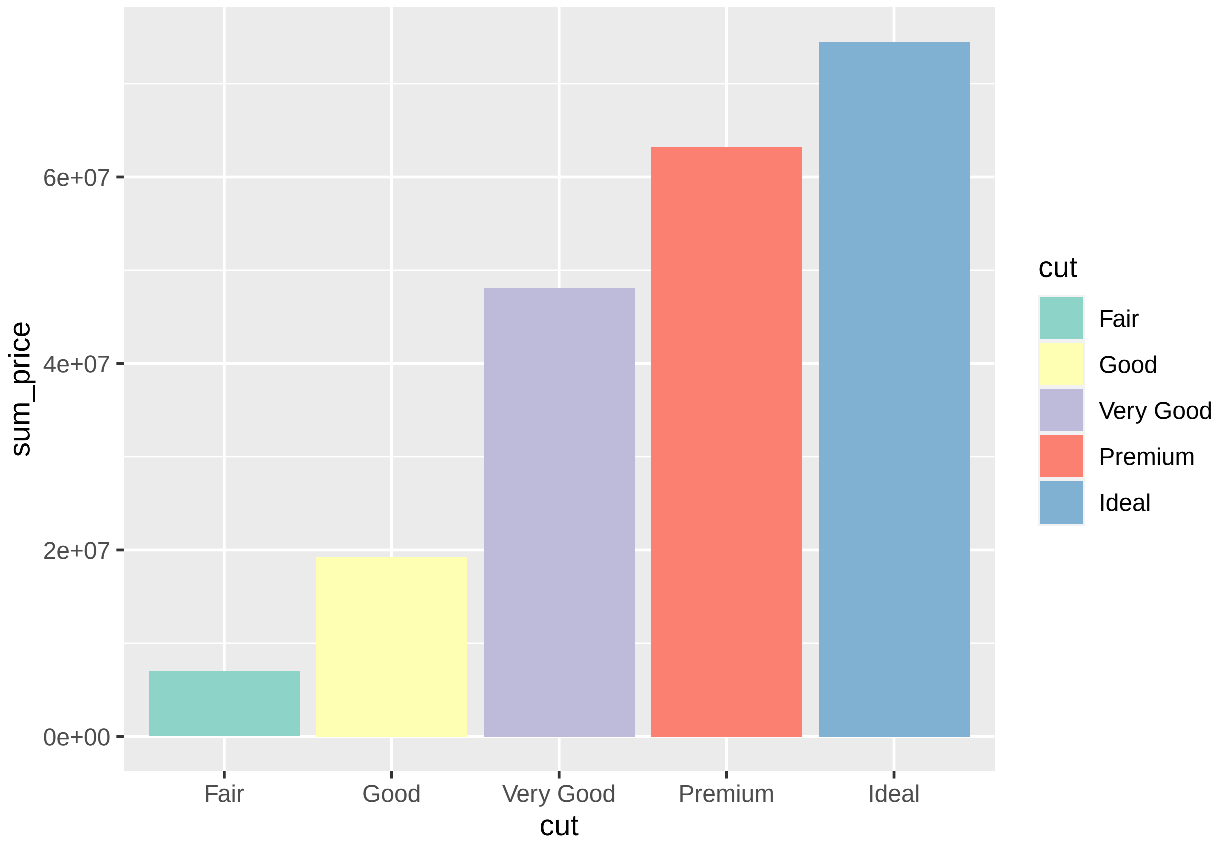
The graph illustrates a positive correlation between carat, cut and price. Then diamonds with higher carat and superior cut quality tend to higher prices.

5.Which the cut type that generates the highest revenue?

```
diamonds%>%
  group_by(cut)%>%
  summarise(sum_price = sum(price),
            sum_carat = sum(carat),
            amount = n(),
            avg_price_per_carat = sum_price/sum_carat)%>%
  arrange(desc(sum_price))
```

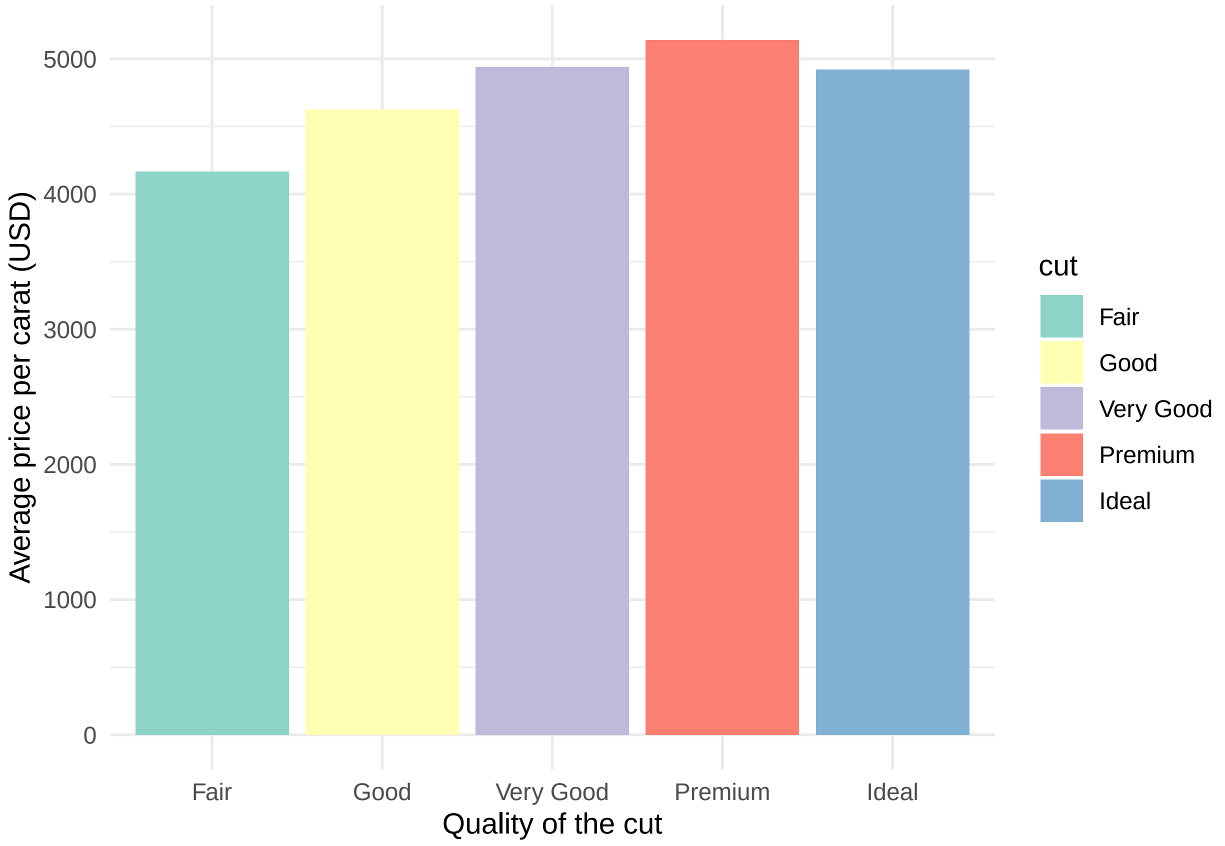
```
## # A tibble: 5 x 5
##   cut       sum_price sum_carat amount avg_price_per_carat
##   <ord>         <int>     <dbl>  <int>               <dbl>
## 1 Ideal      74513487    15147.  21551               4919.
## 2 Premium    63221498    12301.  13791               5140.
## 3 Very Good  48107623     9743.  12082               4938.
## 4 Good       19275009     4166.   4906               4627.
## 5 Fair        7017600     1684.   1610               4167.
```

```
diamonds%>%
  group_by(cut)%>%
  summarise(sum_price = sum(price))%>%
  arrange(desc(sum_price))%>%
  ggplot(aes(cut, sum_price,fill = cut))+
  geom_col()+
  scale_fill_brewer(palette = "Set3")
```



```
  theme_minimal()+
  labs(x= "Quality of the cut",
       y = "Revenue (USD)")
```

```
diamonds%>%
  group_by(cut)%>%
  summarise(sum_price = sum(price),
            sum_carat = sum(carat),
            avg = sum_price/sum_carat)%>%
  arrange(desc(sum_price))%>%
  ggplot(aes(cut, avg, fill = cut))+
  geom_col()+
  scale_fill_brewer(palette = "Set3")+
  theme_minimal()+
  labs(x= "Quality of the cut",
       y = "Average price per carat (USD)")
```



Summary,

- Revenue by cut quality If we focused on price by cut quality reveals that diamonds with an ideal cut generate the highest revenue among different quality grades.

- Revenue per carat If we considering revenue per carat, Premium-cut diamonds outperform Ideal-cut diamonds. This suggests that focusing on producing more Premium-cut diamonds could potentially increase revenue for diamond suppliers.