Homework 5

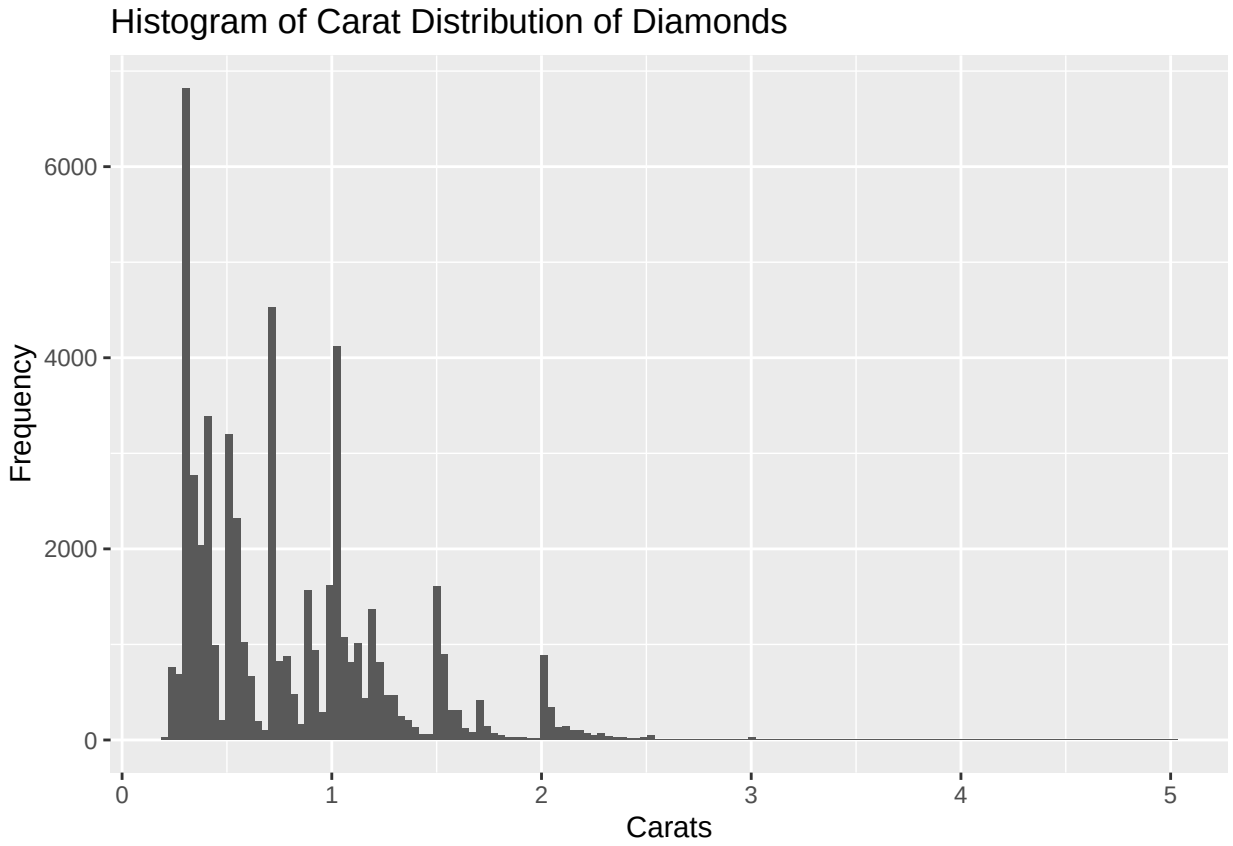
Kevin Jin

3/28/2022

Problem 1

Part (a)

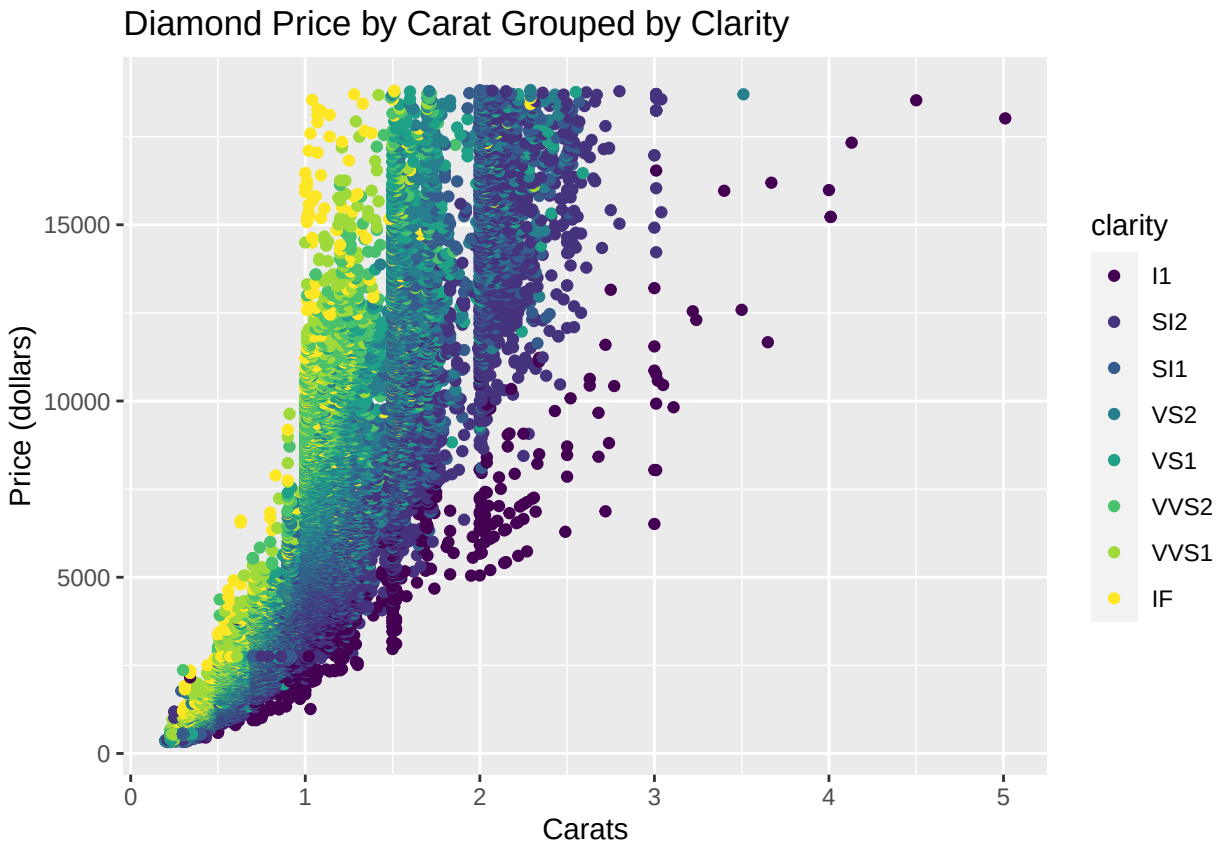
```
library(ggplot2)
data(diamonds)
x <- diamonds$carat
n <- length(x)
# Use Freedman-Diaconi to calculate bin width since there are many outliers
h <- 2 * IQR(x) / n ^ (1 / 3) # bin width
k <- ceiling((max(x) - min(x)) / h) # number of bins

ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = carat), bins = k) +
  labs(title = "Histogram of Carat Distribution of Diamonds",
       x = "Carats",
       y = "Frequency")
```



As there are many outliers in the distribution of carats, it is best to use Freedman-Diaconi's choice to calculate the binwidth and number of bins. This histogram shows that the distribution of carats is left-skewed towards a smaller number of carats; the dataset tends to have diamonds of smaller mass.

Part (b)

```
ggplot(data = diamonds) +
  geom_point(mapping = aes(x = carat, y = price, color = clarity)) +
  labs(title = "Diamond Price by Carat Grouped by Clarity",
       x = "Carats",
       y = "Price (dollars)")
```

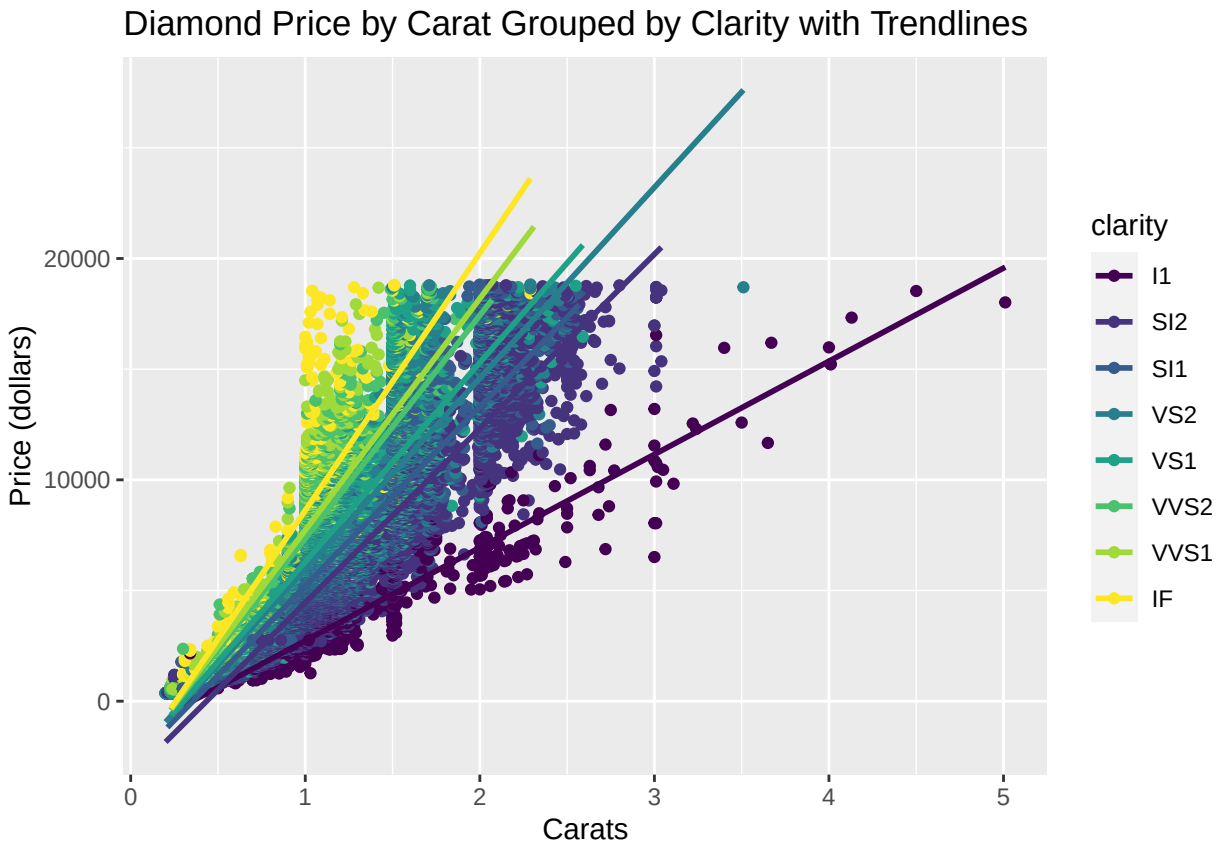


It is difficult to discern much information about the dataset from this scatter plot because it is too busy and the data points tend to overlap, making it hard to spot patterns or trends. However, it is possible to see a general positive correlation between carats and price for all diamond clarities.

Part (c)

```
ggplot(data = diamonds) +
  geom_point(mapping = aes(x = carat, y = price, color = clarity)) +
  geom_smooth(mapping = aes(x = carat, y = price, color = clarity),
              method = "lm",
              se = FALSE) +
  labs(title = "Diamond Price by Carat Grouped by Clarity with Trendlines",
       x = "Carats",
       y = "Price (dollars)")
```

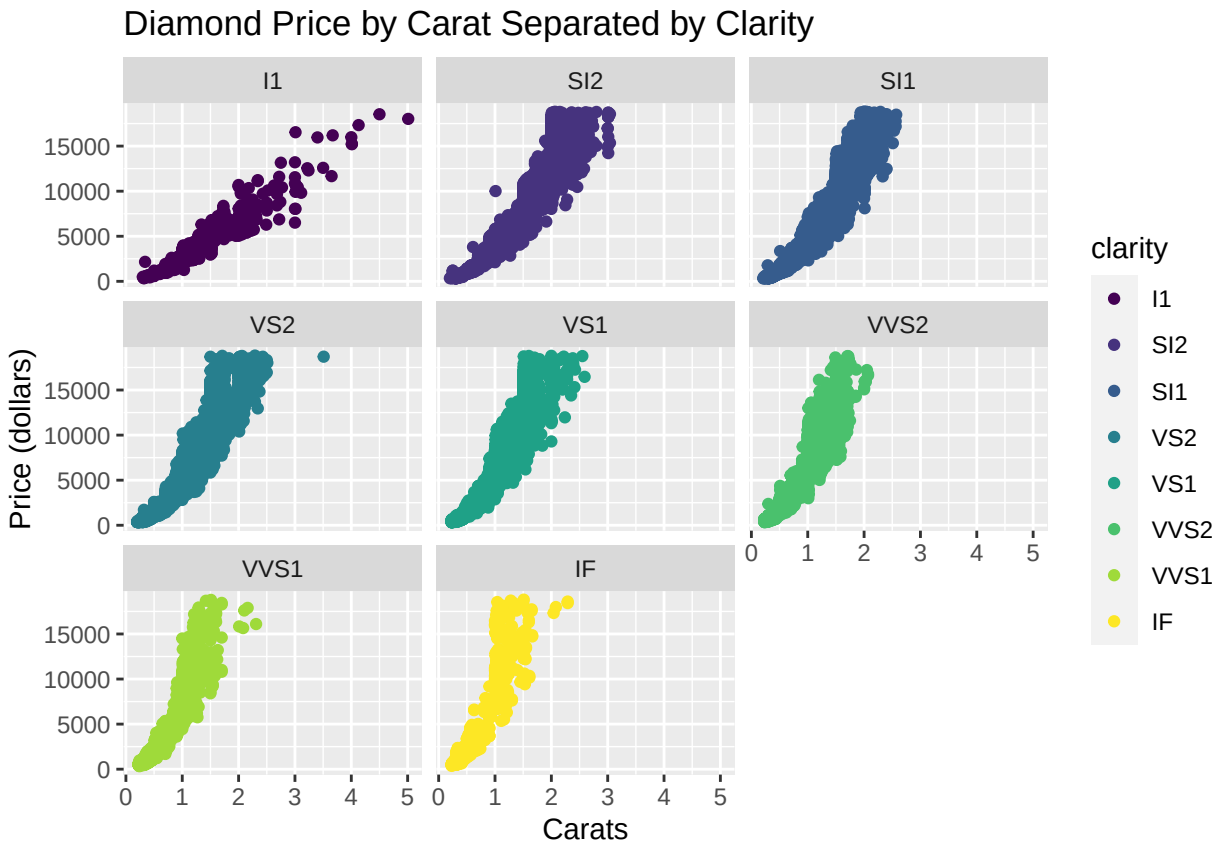
```
## `geom_smooth()` using formula 'y ~ x'
```



Adding trendlines to the scatter plot makes it much easier to see that there is a positive correlation between carats and price. It is also clear that the correlations are roughly linear.

Part (d)

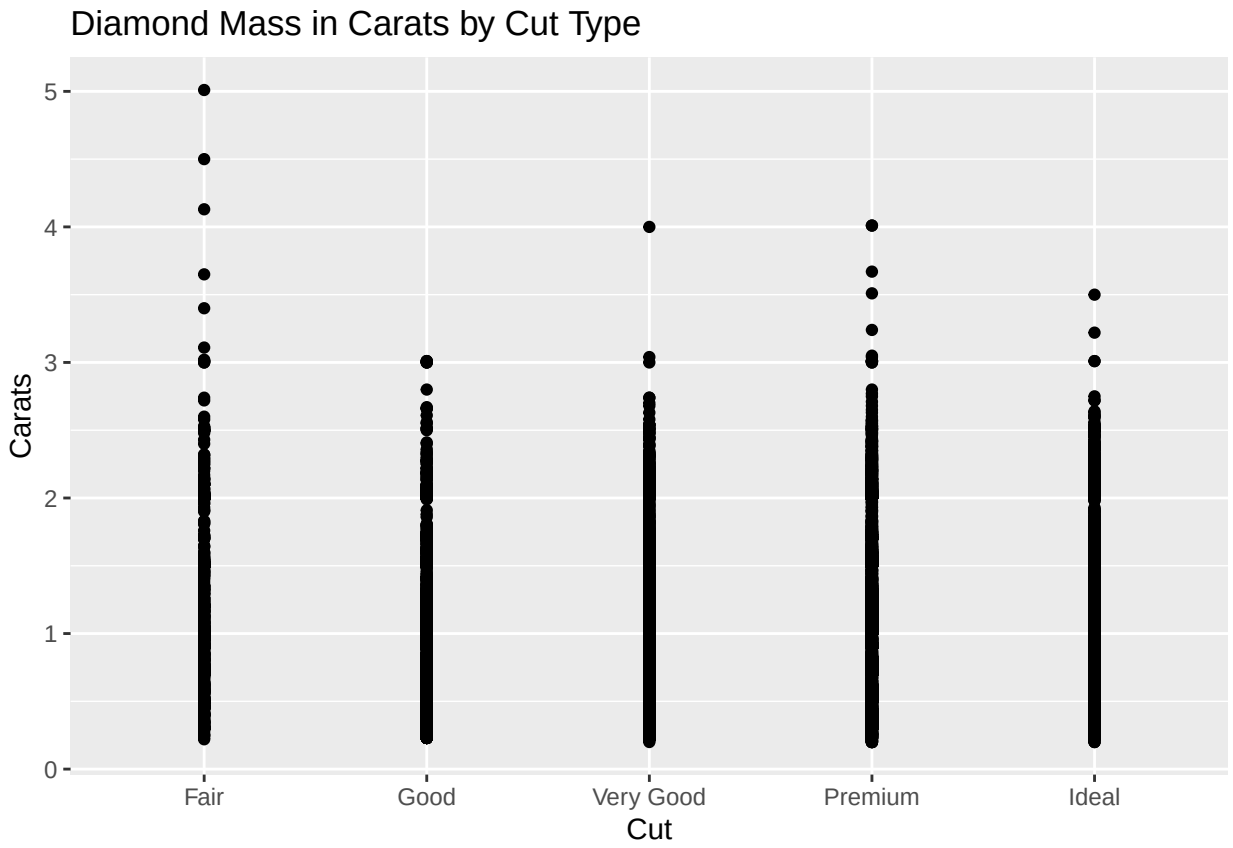
```
ggplot(data = diamonds) +
  geom_point(mapping = aes(x = carat, y = price, color = clarity)) +
  labs(title = "Diamond Price by Carat Separated by Clarity",
       x = "Carats",
       y = "Price (dollars)") +
  facet_wrap(~clarity, nrow = 3)
```



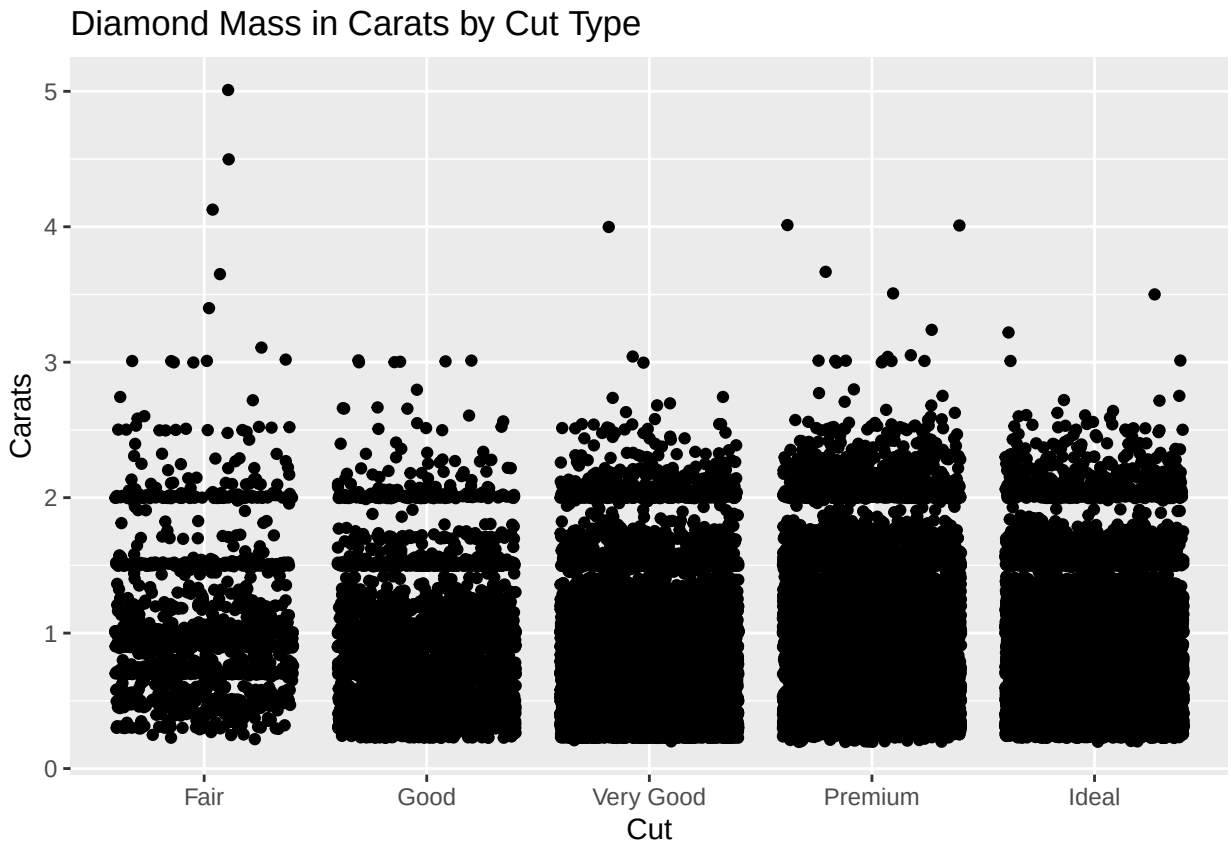
Faceting by clarity makes it much easier to view the trend of each clarity on its own. It is possible to see that I1 clarity has a less steep correlation compared to the other clarities.

Part (e)

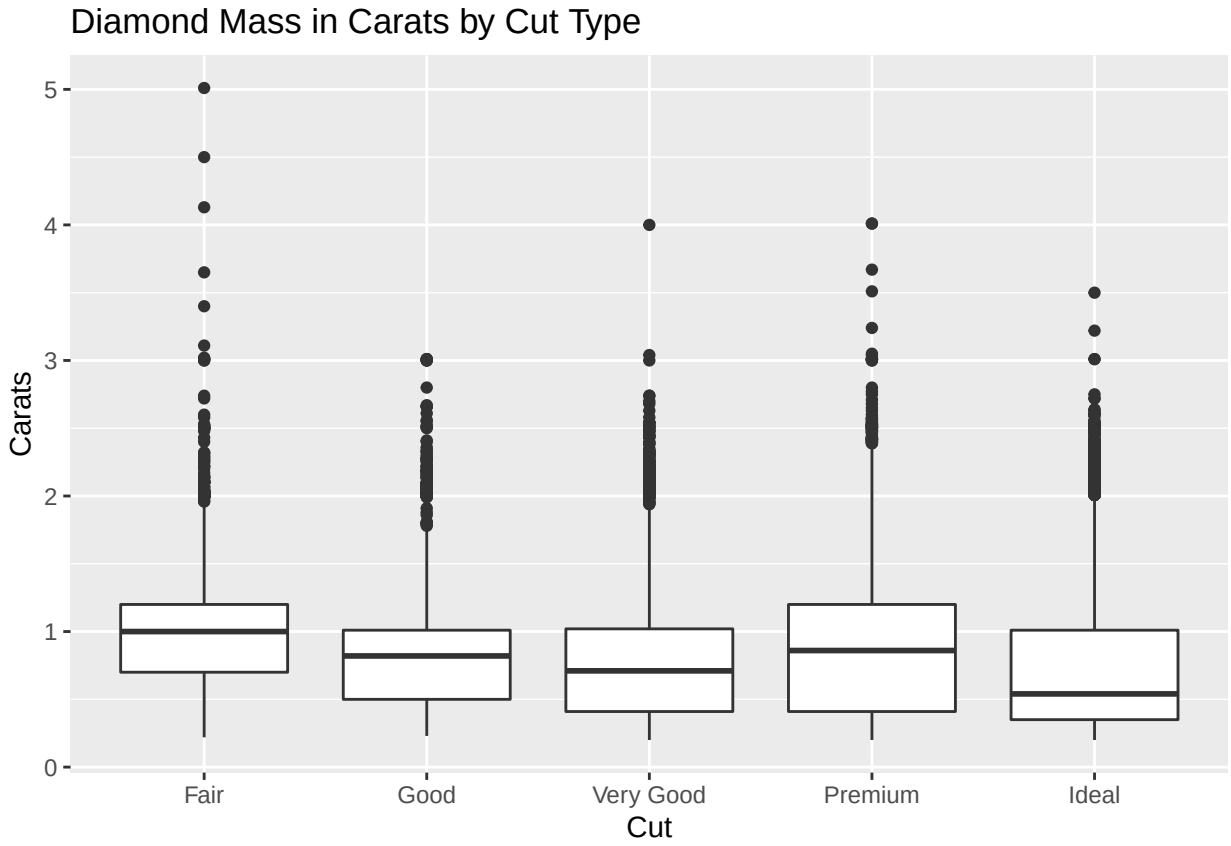
```
ggplot(data = diamonds) +
  geom_point(mapping = aes(x = cut, y = carat)) +
  labs(title = "Diamond Mass in Carats by Cut Type",
       x = "Cut",
       y = "Carats")
```



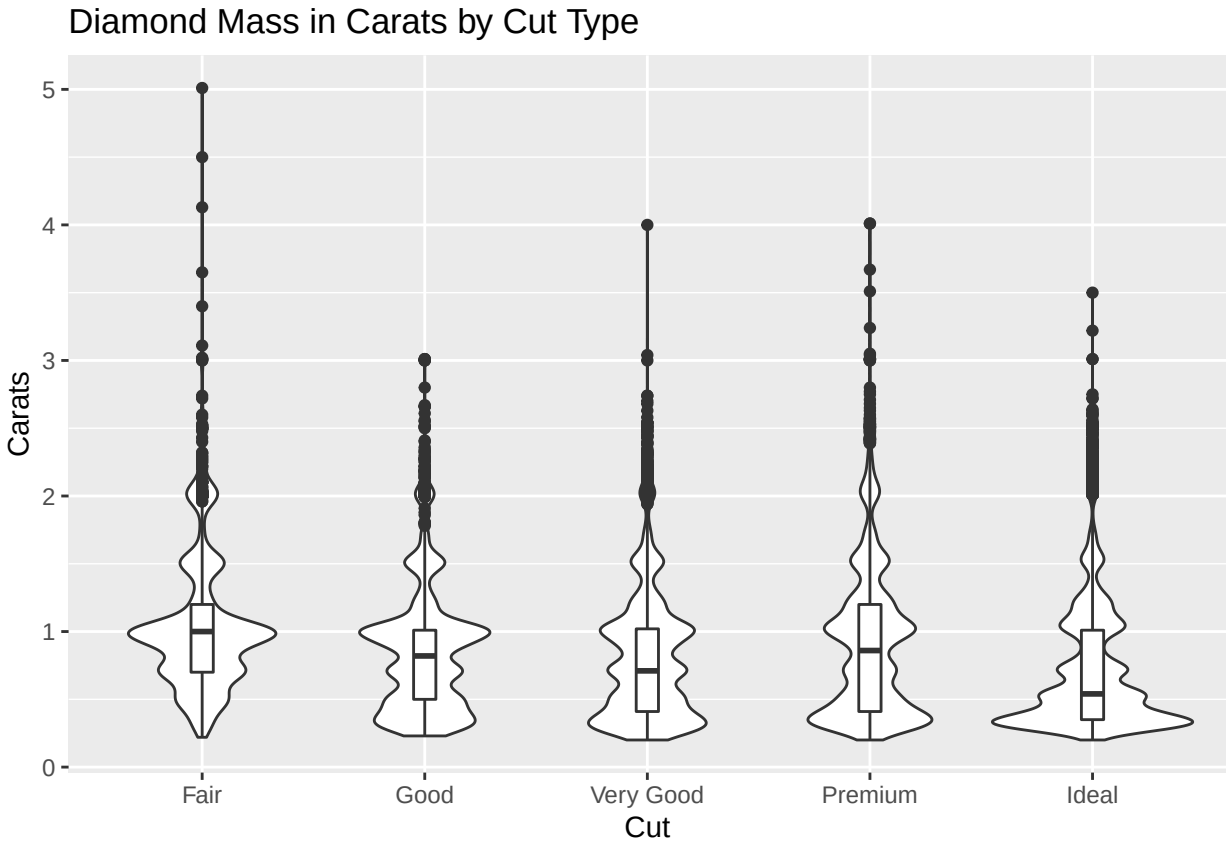

```
ggplot(data = diamonds) +
  geom_jitter(mapping = aes(x = cut, y = carat)) +
  labs(title = "Diamond Mass in Carats by Cut Type",
       x = "Cut",
       y = "Carats")
```




```
ggplot(data = diamonds) +
  geom_boxplot(mapping = aes(x = cut, y = carat)) +
  labs(title = "Diamond Mass in Carats by Cut Type",
       x = "Cut",
       y = "Carats")
```




```
ggplot(data = diamonds) +
  geom_violin(mapping = aes(x = cut, y = carat)) +
  geom_boxplot(mapping = aes(x = cut, y = carat), width = 0.1) +
  labs(title = "Diamond Mass in Carats by Cut Type",
       x = "Cut",
       y = "Carats")
```

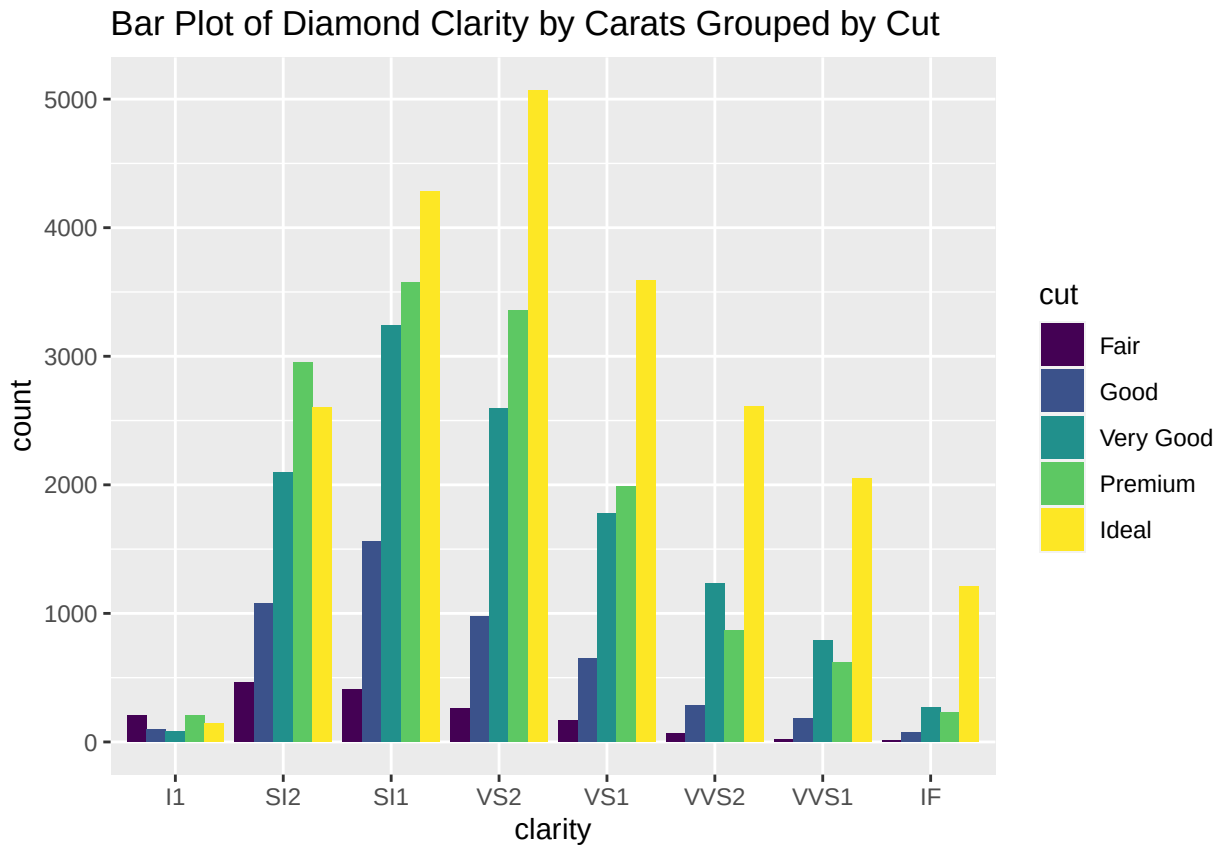


A violin plot is best for visualizing the distributions of carat sizes among all cut types. Carat is a continuous quantity and the distributions of carat differs along x, so a violin plot provides more information and is the best to show the differing densities, since a boxplot is not affected by differing distributions if the median and inter-quartile range are the same.

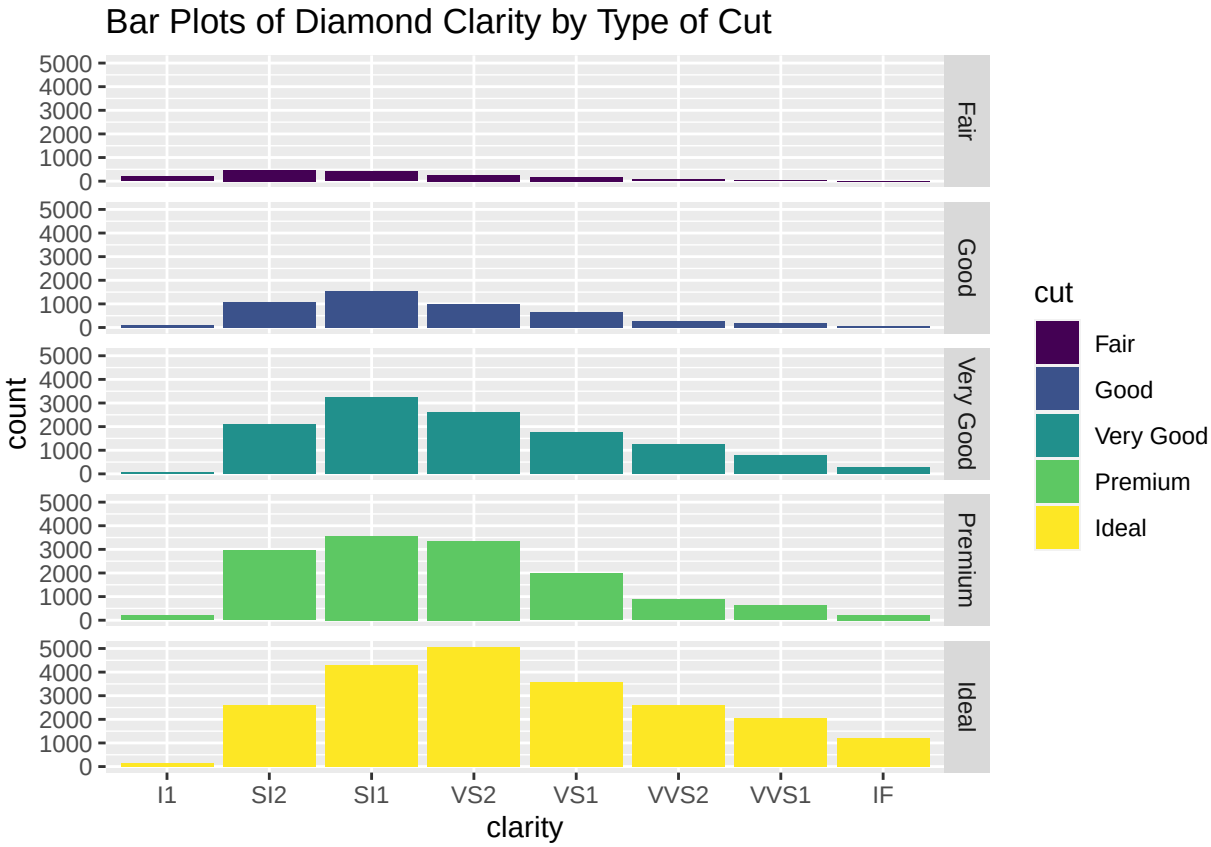
Problem 2

Part (a)

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = clarity, fill = cut),
           position = "dodge") +
  labs(title = "Bar Plot of Diamond Clarity by Carats Grouped by Cut")
```




```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = clarity, fill = cut)) +
  facet_grid(cut~.) +
  labs(title = "Bar Plots of Diamond Clarity by Type of Cut")
```

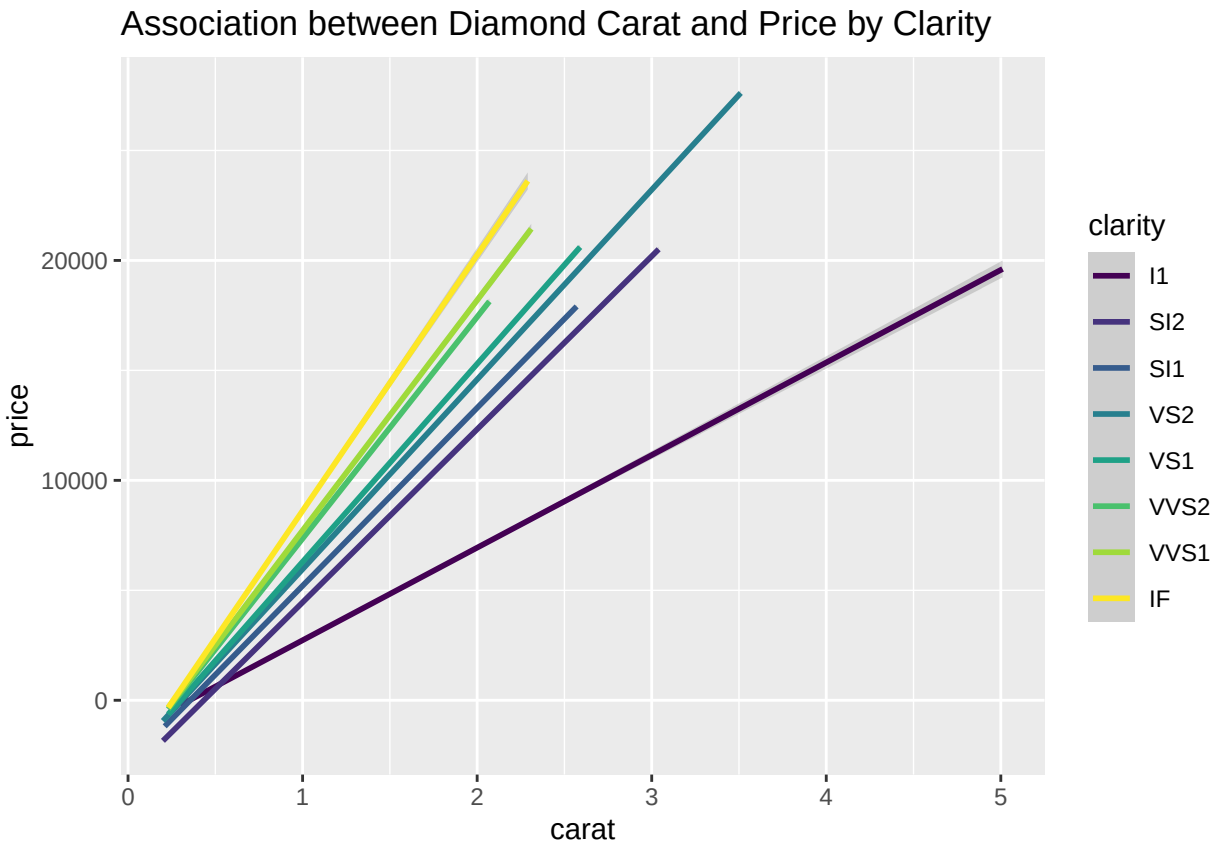


The grouped bar plot is better for visualizing differences between clarity types and is more visually attractive, while the faceted bar plots are better for visualizing the individual distributions of each clarity type.

Part (b)

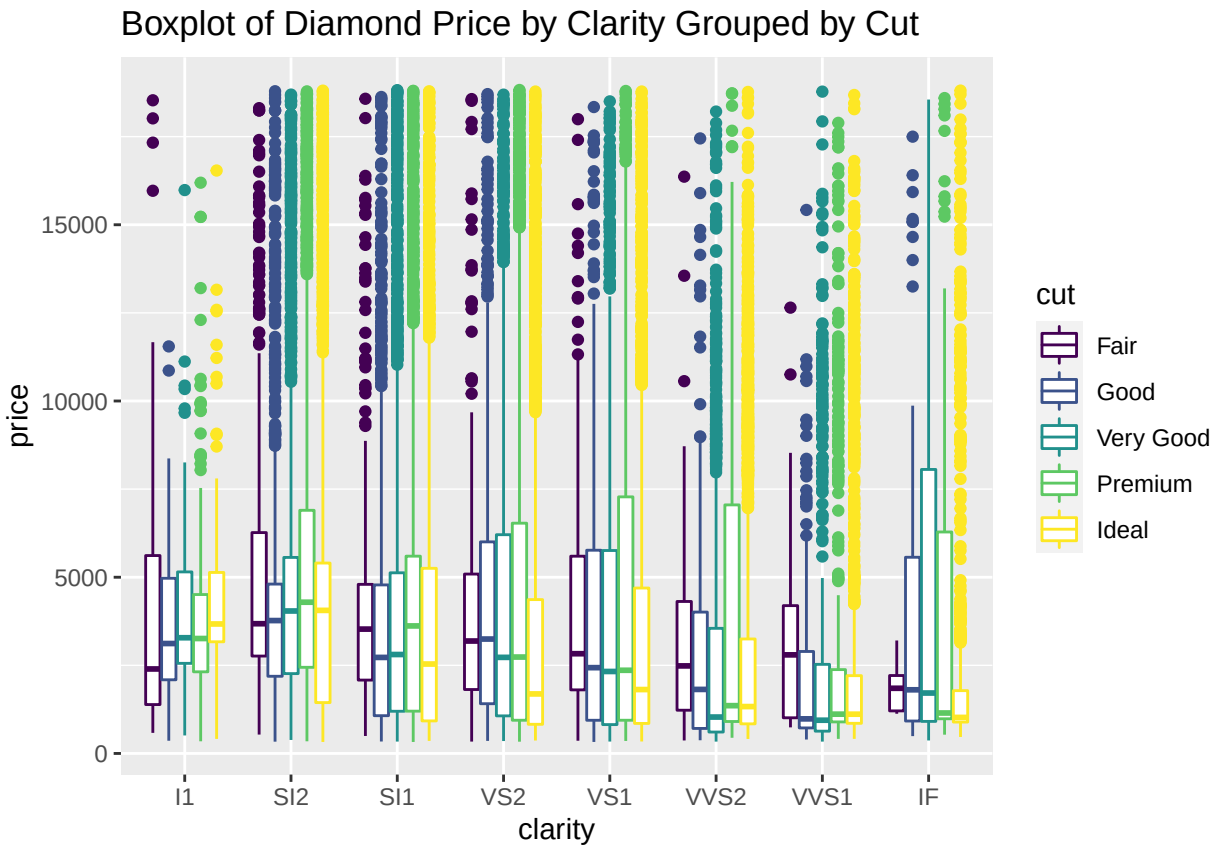
```
ggplot(data = diamonds) +
  geom_smooth(mapping = aes(x = carat, y = price, color = clarity),
              method = "lm") +
  labs(title = "Association between Diamond Carat and Price by Clarity")
```

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



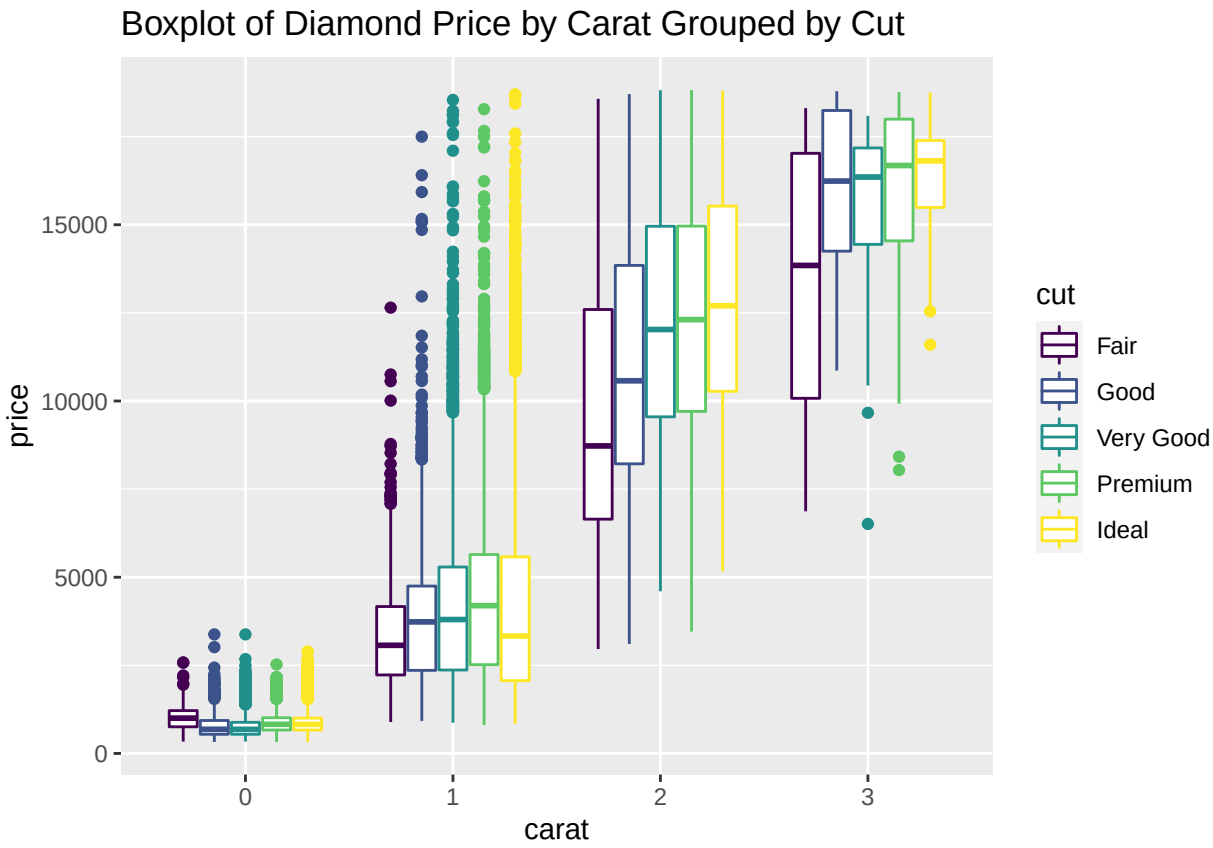
Part (c)

```
ggplot(data = diamonds) +
  geom_boxplot(mapping = aes(x = clarity,
                             y = price,
                             color = cut)) +
  labs(title = "Boxplot of Diamond Price by Clarity Grouped by Cut")
```



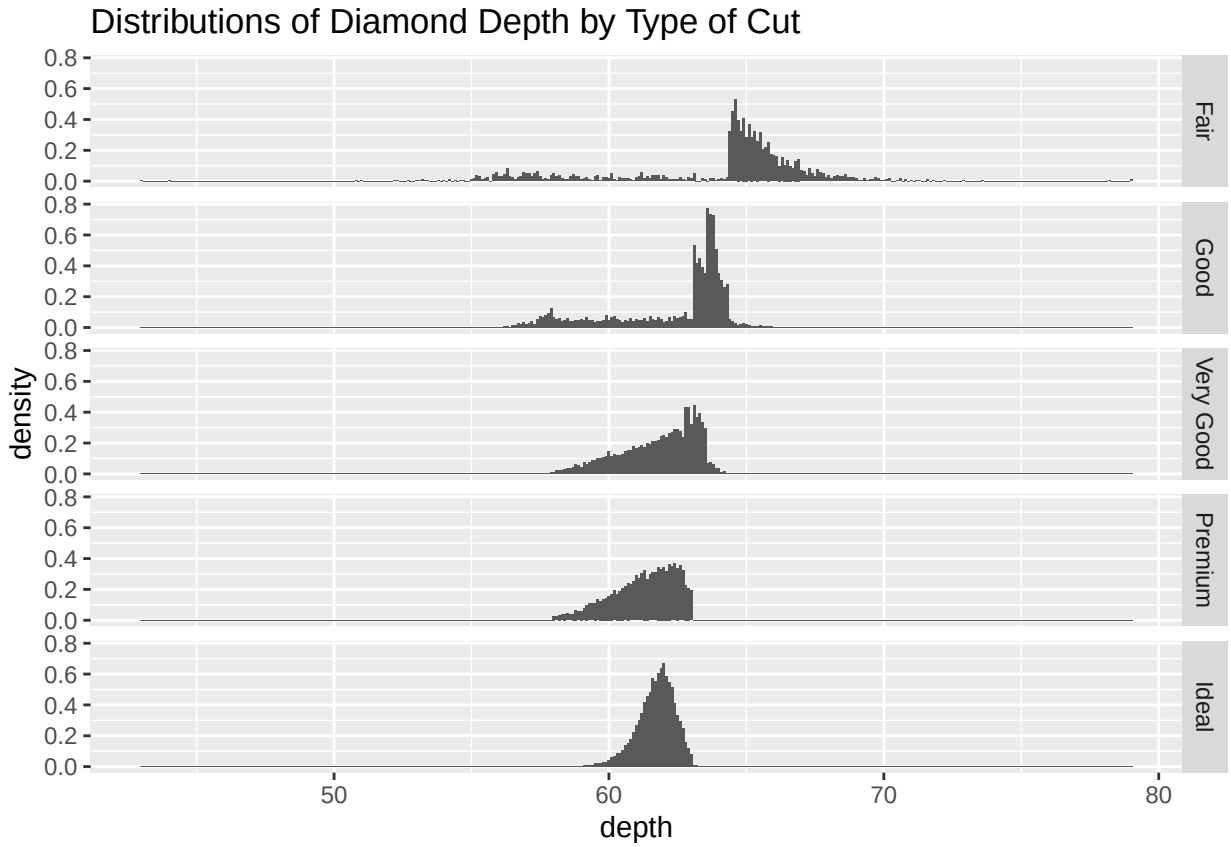
Part (d)

```
ggplot(data = subset(diamonds, carat < 3.5)) +
  geom_boxplot(mapping = aes(x = as.factor(round(carat)),
                             y = price,
                             color = cut)) +
  labs(title = "Boxplot of Diamond Price by Carat Grouped by Cut",
       x = "carat")
```



Part (e)

```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = depth, y = ..density..),
                 binwidth = 0.1) +
  facet_grid(cut~.) +
  labs(title = "Distributions of Diamond Depth by Type of Cut")
```


Distributions of Diamond Depth by Type of Cut
density
depth
Fair
Good
Very Good
Premium
Ideal
0.0
0.2
0.4
0.6
0.8
50
60
70
80