Lab Sheet 9

STAT260, University of Otago

Sem. 2, 2021

Task 1

The data that we will use today is a random sample of 1000 observations from the "diamonds" data set which is part of the `ggplot2` package. Today we will use these 1000 observations across a subset of just three of the variables:

- **price:** price in US dollars
- **carat:** weight of the diamond
- **cut:** quality of the cut (Fair, Good, Very Good, Premium, Ideal)

These data are in the file `diamonds1000.csv`, labelled "Diamonds data set (Lab 9)" on the Resources page for this course. Load this data into R and use `str` to examine the structure of the data. Also show the first 10 rows of the data.

Answer:

```
diamonds <- read.csv('./diamonds1000.csv')
str(diamonds)
```

```
## 'data.frame':    1000 obs. of  3 variables:
##  $ price: int  1332 2062 675 402 8596 439 829 984 941 1815 ...
##  $ carat: num  0.53 0.7 0.3 0.23 1.2 0.3 0.33 0.42 0.45 0.52 ...
##  $ cut  : chr  "Good" "Good" "Premium" "Good" ...
```

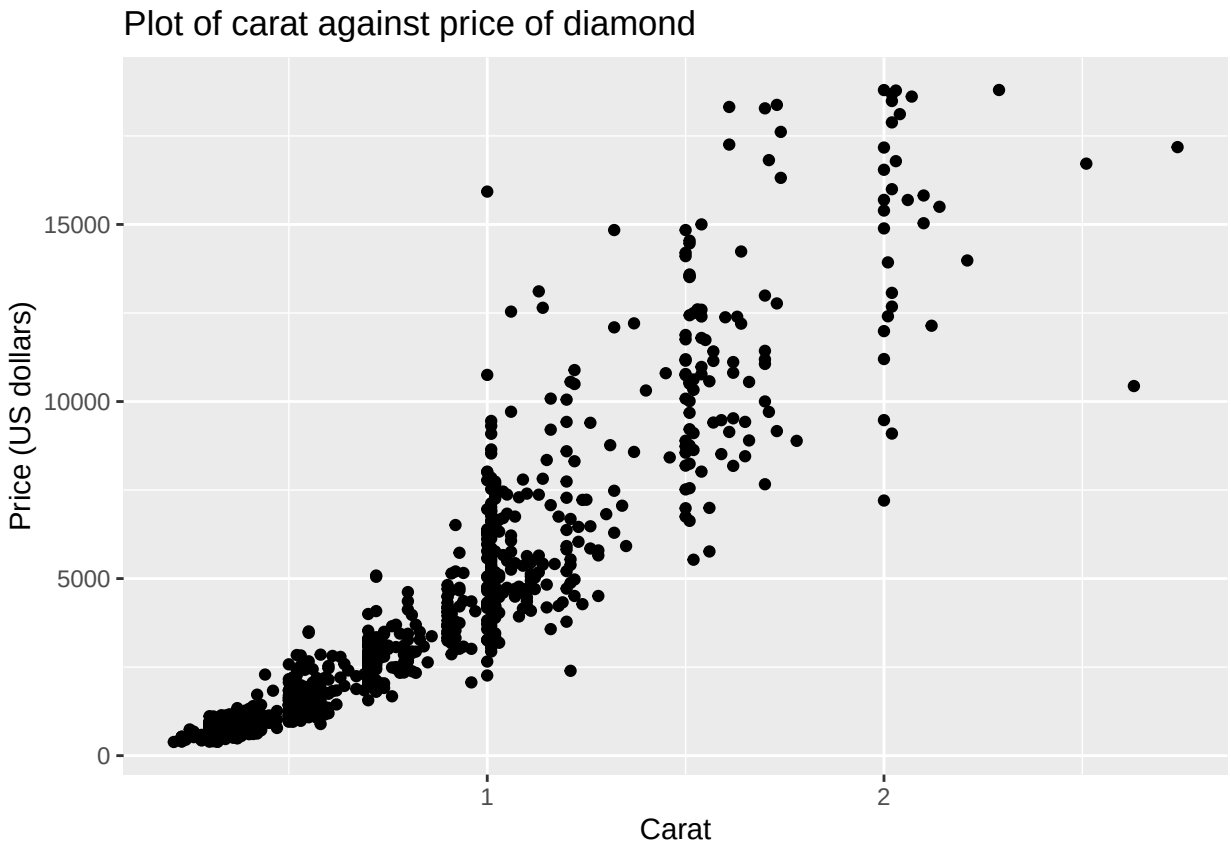
Task 2

Make a scatterplot of carat *versus* price. Are there any problems with the way the data are displayed? Make a second plot, using jitter and alpha levels to improve how the data are displayed.

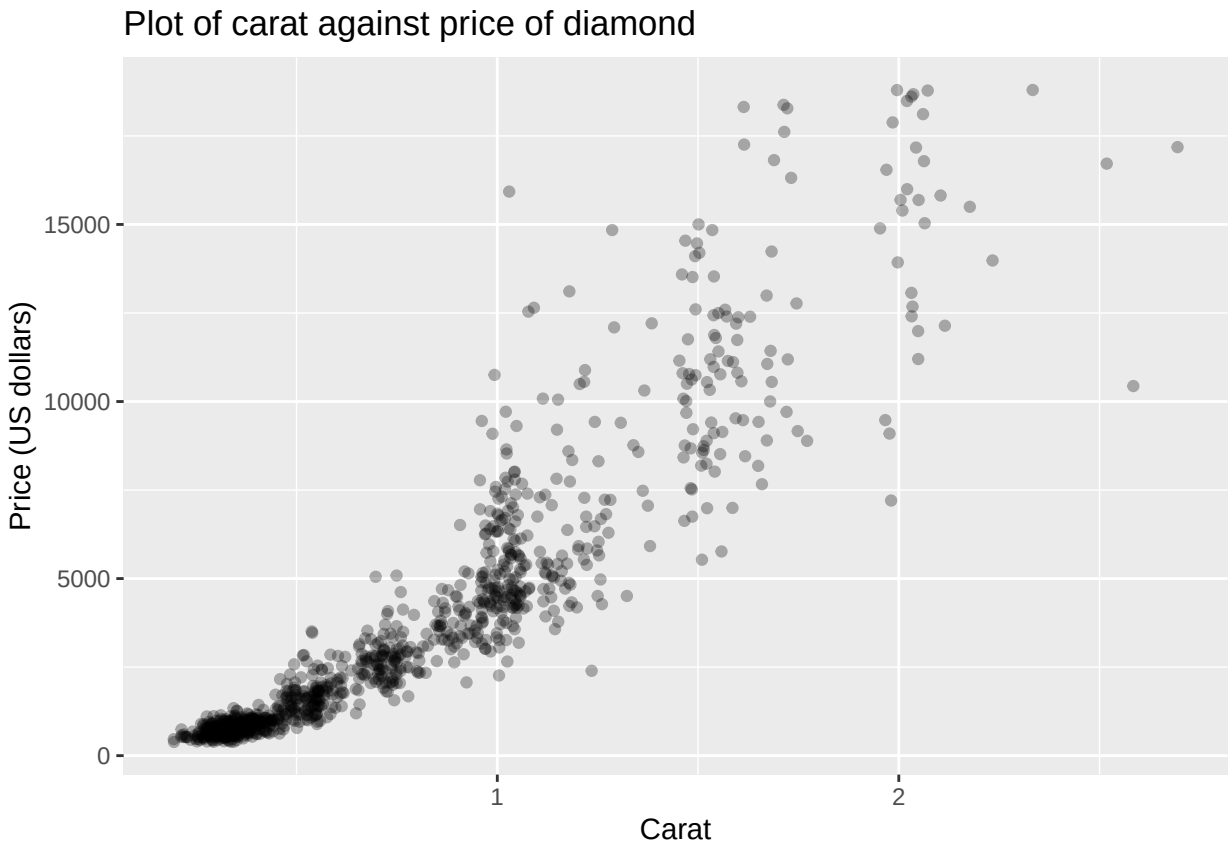
Between carat 0 and 1 there are a lot of overlaping data points, whicch makes it hard to clearly see the distribution of the data points.

Answer:

```
diamonds %>%
  ggplot(aes(x=carat, y=price))+
  geom_point()+
  ggtitle('Plot of carat against price of diamond')+
  ylab('Price (US dollars)')+
  xlab('Carat')
```




```
diamonds %>%
  ggplot(aes(x=carat, y=price))+
  geom_jitter(width = 0.05, height = 0.1, alpha = 0.3)+
  ggtitle('Plot of carat against price of diamond')+
  ylab('Price (US dollars)')+
  xlab('Carat')
```



******

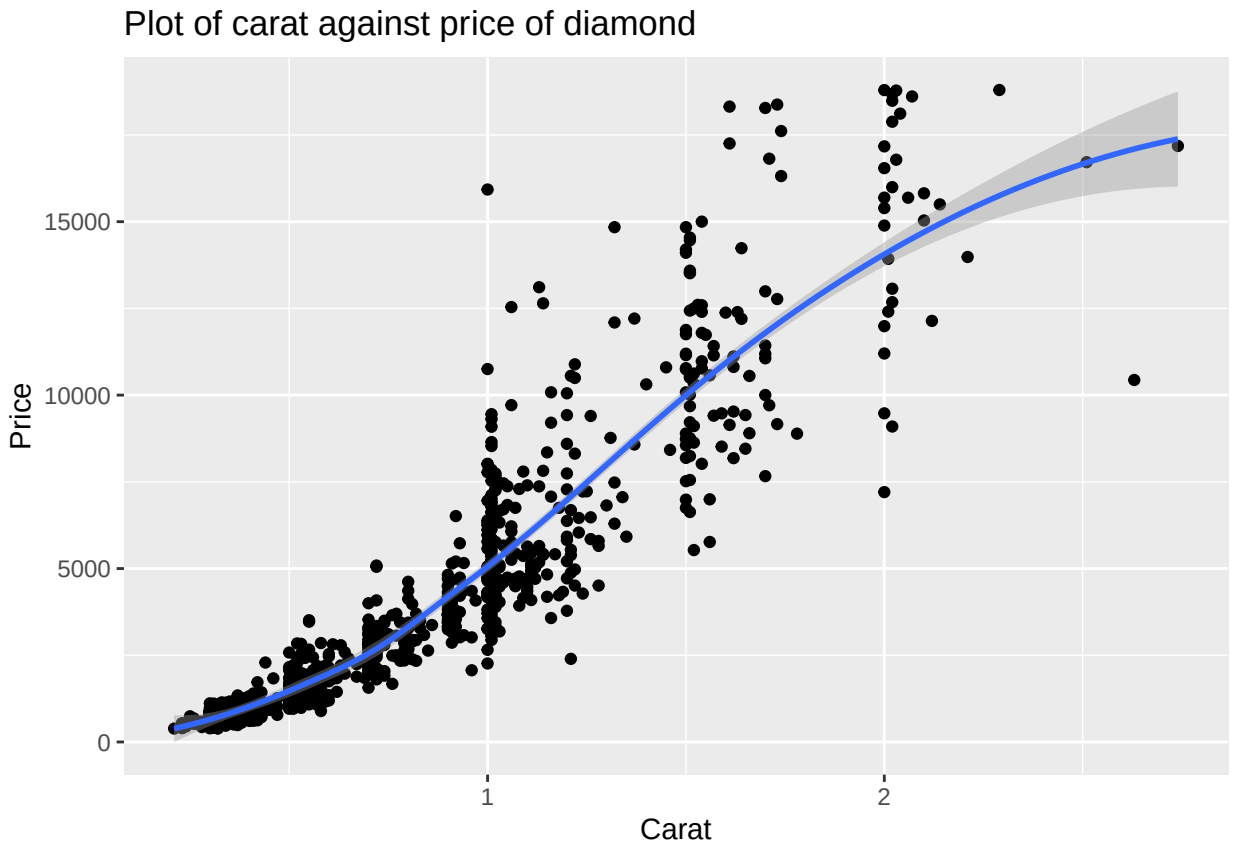
Task 3

Add a smoth line to your scatterplot from Task 1, using the `loess` method to fit the line. Describe the relationship between carat and price. Is it linear? (i.e., is it a straight-line relationship?)

Answer:

The relationship between carat and price is a positive non-linear relationship.

```
diamonds %>%
  ggplot(aes(x=carat, y=price))+
  geom_point()+
  geom_smooth(method = 'loess', formula = y~x)+
  ggtitle('Plot of carat against price of diamond')+
  ylab('Price')+
  xlab('Carat')
```



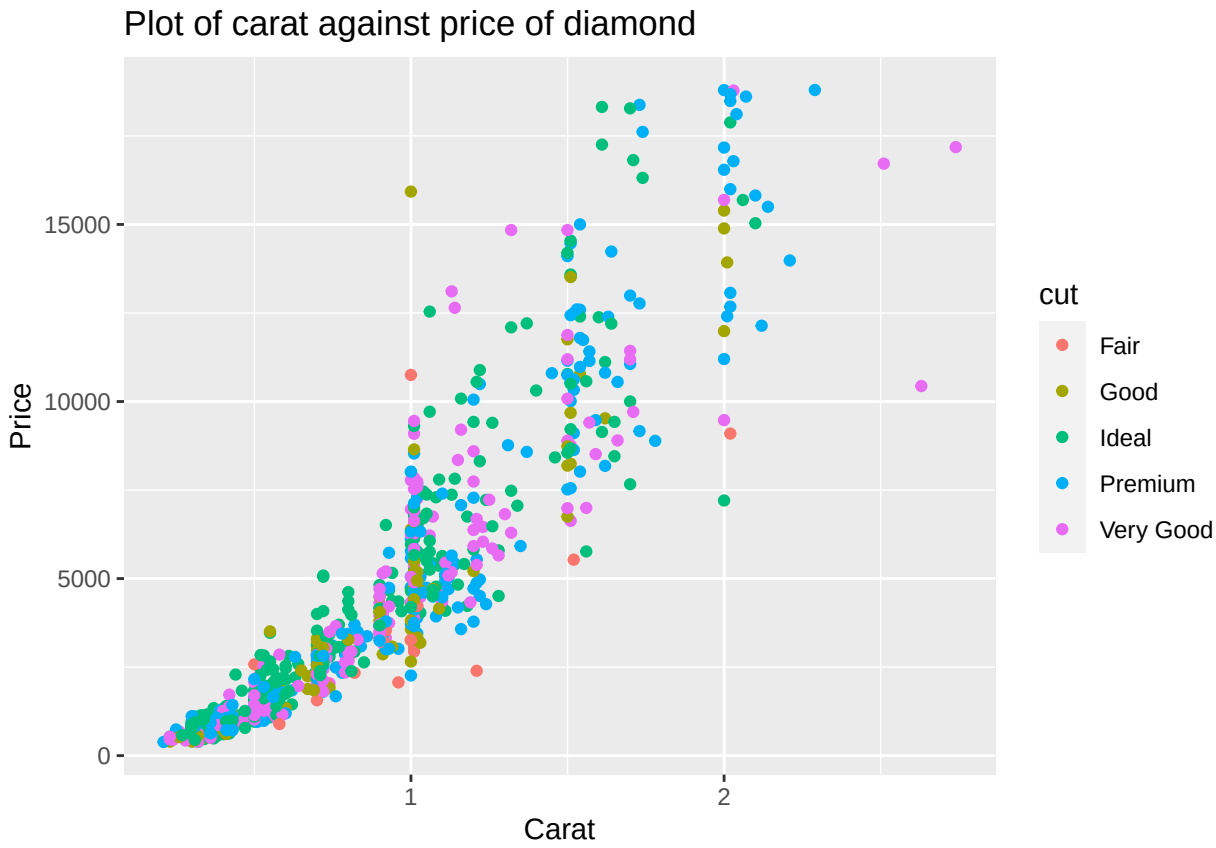
Task 4

Add information about the diamonds' cut to the scatterplot, using colour. Repeat this, with cut now used to define the size of the points in the plot. In your opinion, are these approaches a good way to include information about cut in the plot?

Answer:

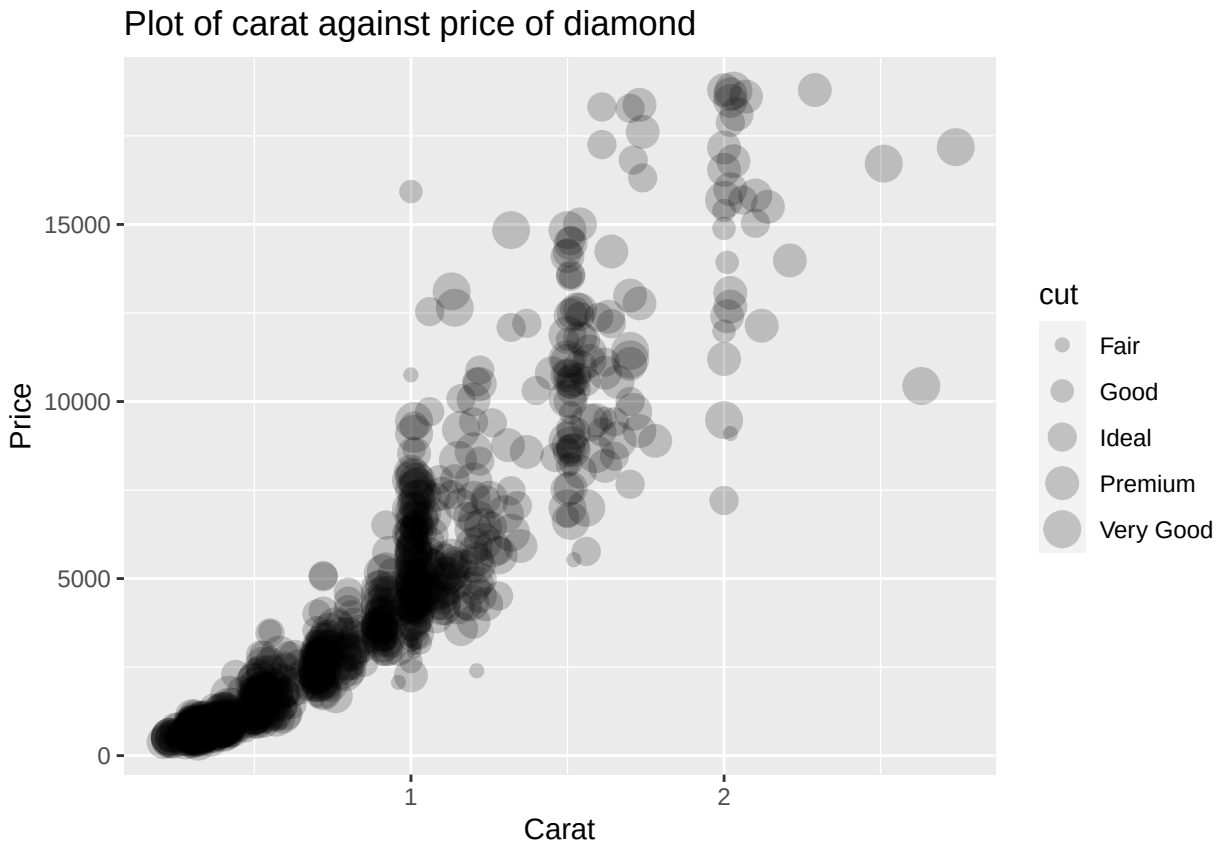
No we should not include information about diamond size in the graph because it does not give us any further insign to the relatioship between carat adn diamond price.

```
diamonds %>%
  ggplot(aes(x=carat, y=price, color = cut))+
  geom_point()+
  ggtitle('Plot of carat against price of diamond')+
  ylab('Price')+
  xlab('Carat')
```




```
diamonds %>%
  ggplot(aes(x=carat, y=price, size = cut))+
  geom_point(alpha = 0.2)+
  ggtitle('Plot of carat against price of diamond')+
  ylab('Price')+
  xlab('Carat')
```

```
## Warning: Using size for a discrete variable is not advised.
```

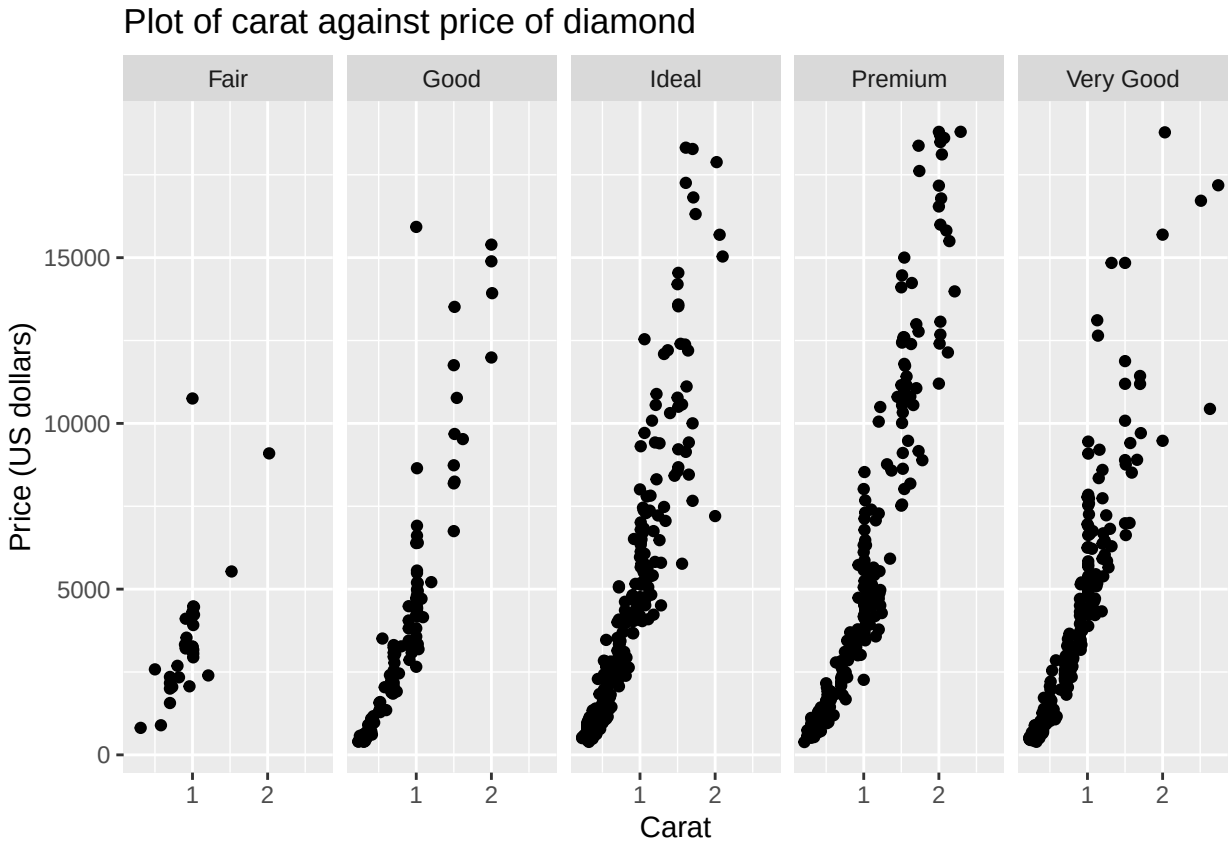


Task 5

Now take your scatterplot and facet it by cut (you should get five separate scatterplots). Which approach (of the three you have just used) was the most effective at adding the information about cut?

Answer:

```
diamonds %>%
  ggplot(aes(x=carat, y=price))+
  geom_point()+
  facet_grid(~cut)+
  ggtitle('Plot of carat against price of diamond')+
  ylab('Price (US dollars)')+
  xlab('Carat')
```



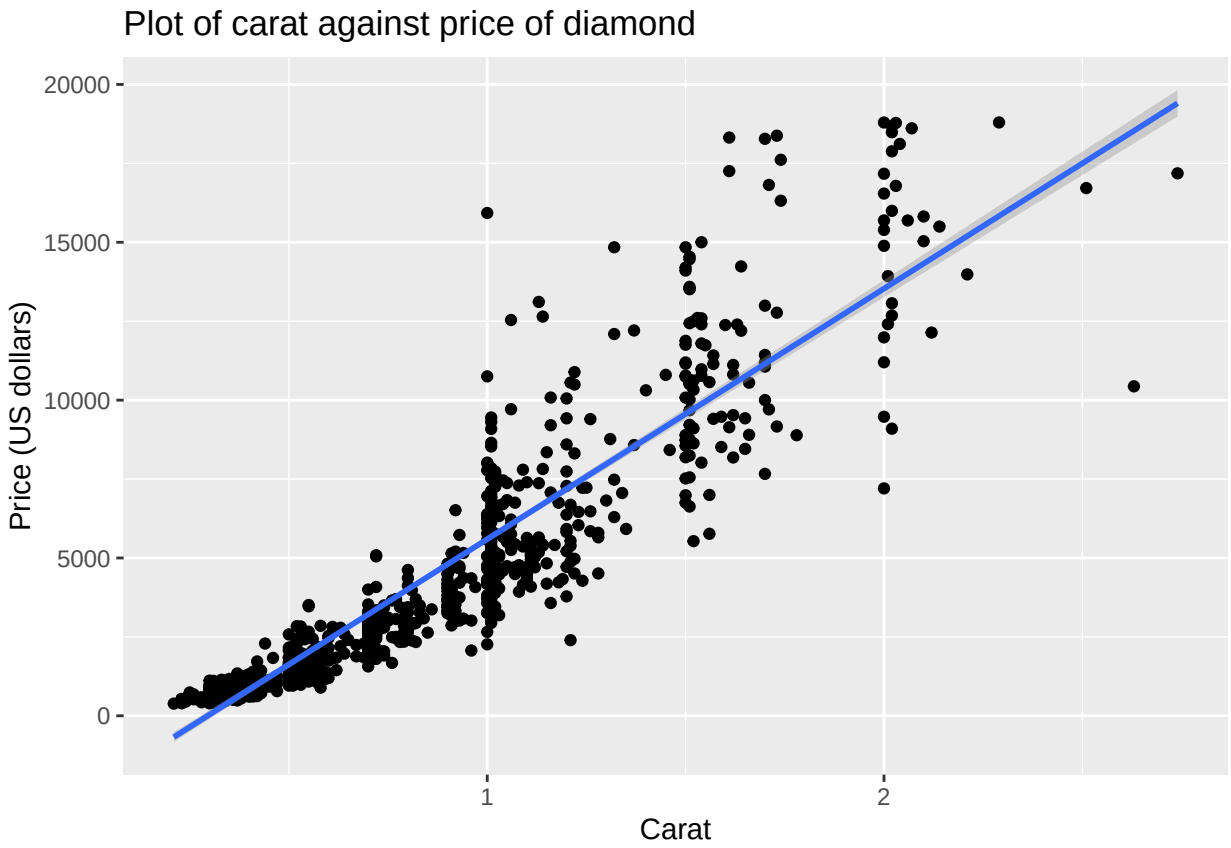
Task 6

Add a straight line of best fit to each of the scatterplots from Task 5. Then make a second plot which includes both a straight line, and also a line fit via loess that is coloured red. Do not include any information about the standard error of the lines in the plots. Which line type (straight-line or loess) do you think provides a better fit to the data?

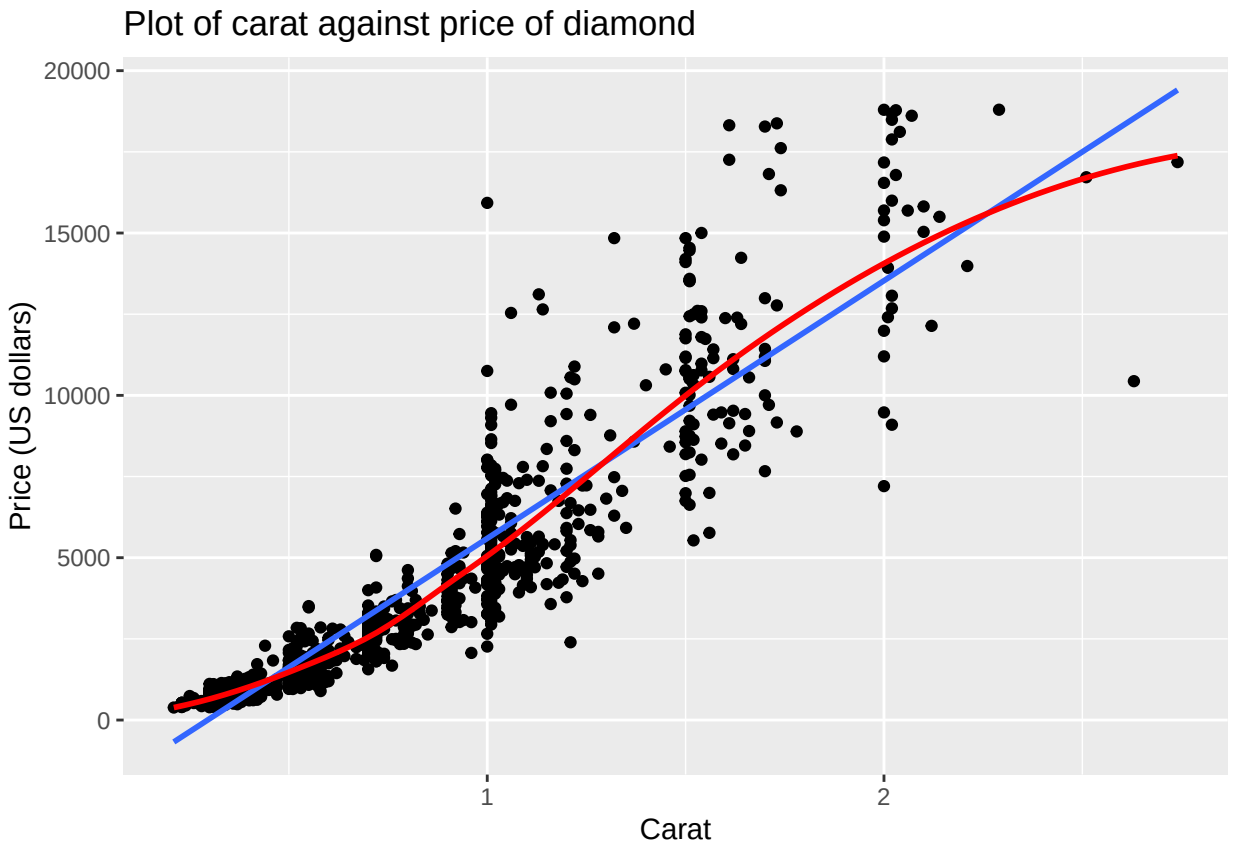
Answer:

geom_smooth() width loess method provide a better git to the data.

```
diamonds %>%
  ggplot(aes(x=carat, y=price))+
  geom_point()+
  geom_smooth(method = 'lm', formula = y~x)+
  ggtitle('Plot of carat against price of diamond')+
  ylab('Price (US dollars)')+
  xlab('Carat')
```




```
diamonds %>%
  ggplot(aes(x=carat, y=price))+
  geom_point()+
  geom_smooth(method = 'lm', formula = y~x, se = FALSE)+
  geom_smooth(method = 'loess', formula = y~x, se = FALSE, color = 'red')+
  ggtitle('Plot of carat against price of diamond')+
  ylab('Price (US dollars)')+
  xlab('Carat')
```


Plot of carat against price of diamond
20000
15000
10000
5000
0
Price (US dollars)
1
2
Carat