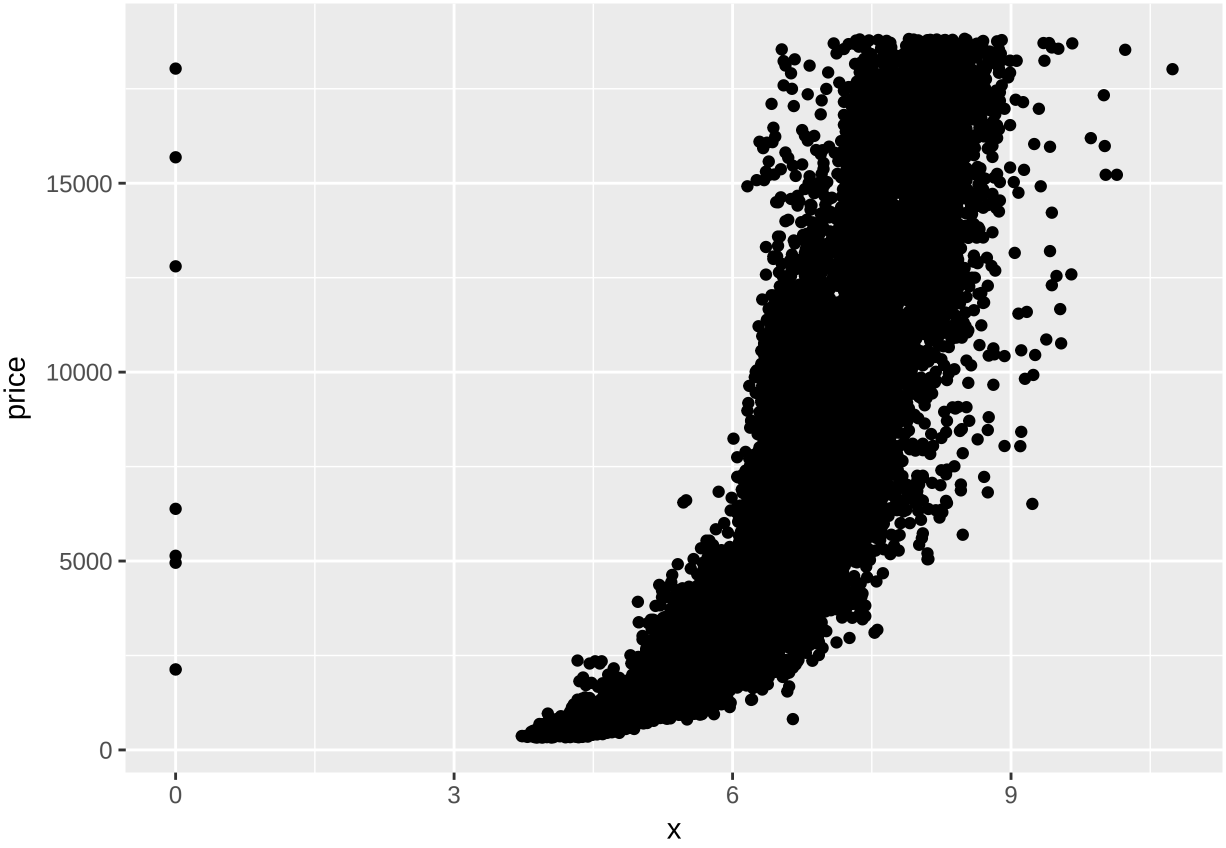
lesson4_problem_set

Dusty P

May 14, 2018

Problem 1 Price vs. x

```
ggplot(aes(x = x, y = price), data = diamonds) +
  geom_point()
```



2. Findings - Price vs. x

There is a general trend towards an increase in price at what appears to be an exponential rate as x increases. But there are a few outliers at x = 0

3. Correlations

```
with(diamonds, cor.test(price, x))
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  price and x
## t = 440.16, df = 53938, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8825835 0.8862594
## sample estimates:
##       cor
## 0.8844352
```

```
with(diamonds, cor.test(price, y))
```

```
##
##  Pearson's product-moment correlation
##
## data:  price and y
## t = 401.14, df = 53938, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8632867 0.8675241
## sample estimates:
##       cor
## 0.8654209
```

```
with(diamonds, cor.test(price, z))
```

```
##
##  Pearson's product-moment correlation
##
## data:  price and z
## t = 393.6, df = 53938, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8590541 0.8634131
## sample estimates:
##       cor
## 0.8612494
```

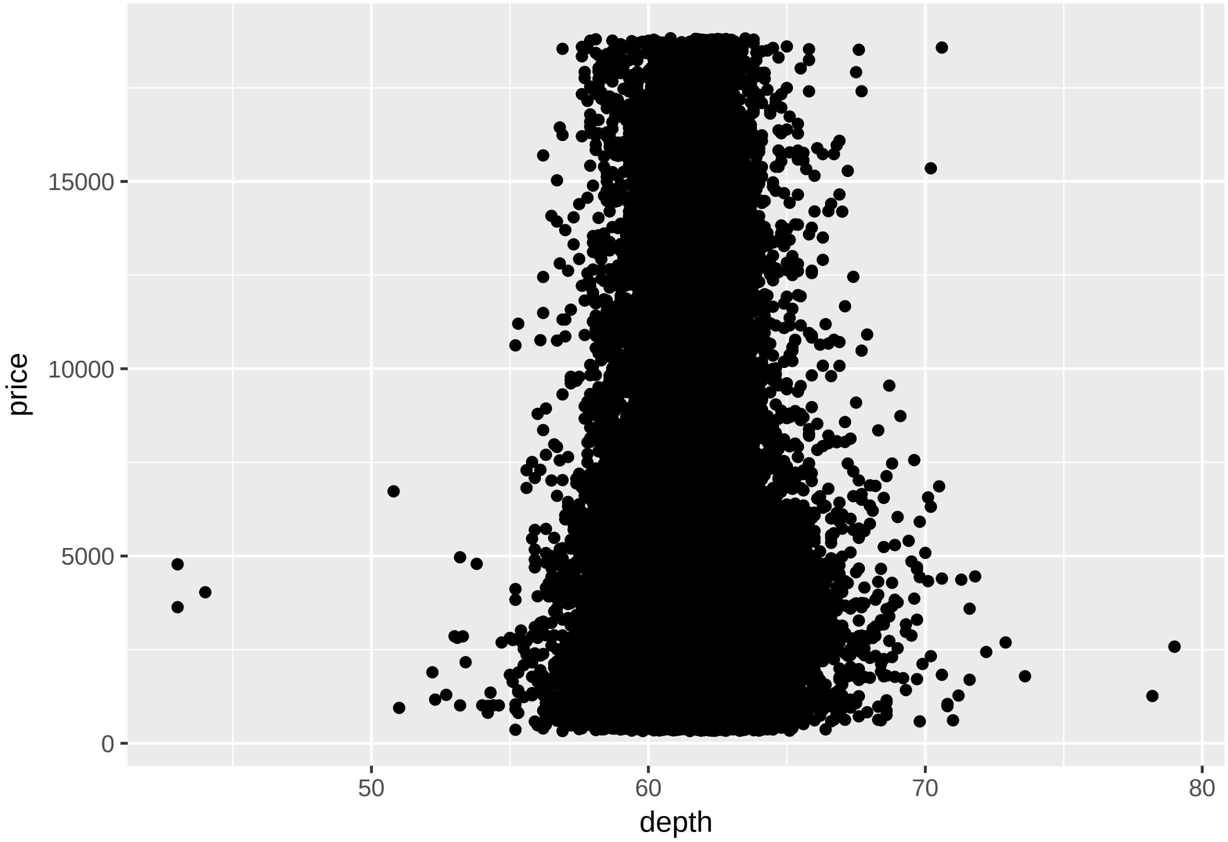
What is the Correlation between price and x? 0.88

What is the correlation between price and y? 0.87

What is the correlation between price and z? 0.86

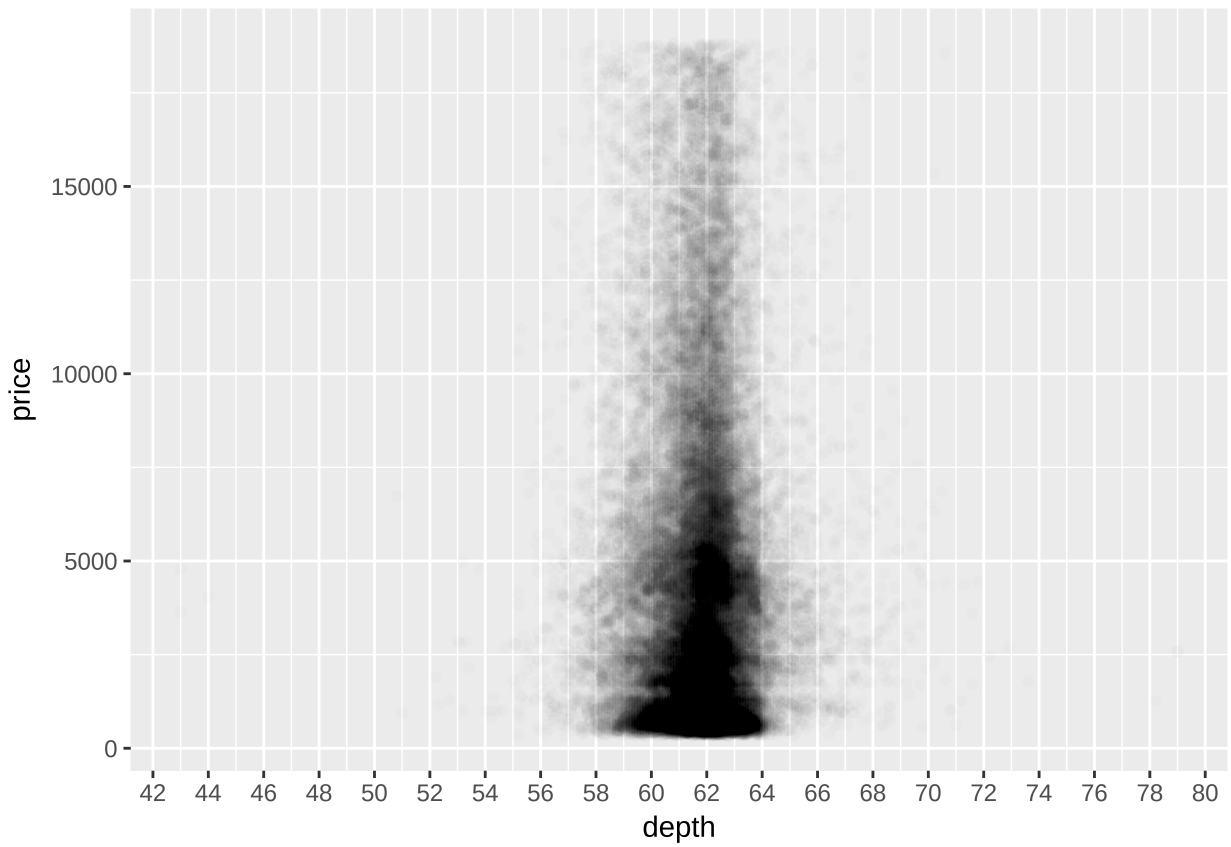
4. Price vs. Depth

```
ggplot(aes(x = depth, y = price), data = diamonds) +
  geom_point()
```



5. Adjustments - Price vs. depth

```
ggplot(data = diamonds, aes(x = depth, y = price)) +
  geom_point(alpha = 1/100) +
  scale_x_continuous(breaks = seq(0, 80, 2))
```



6. Typical Depth Range

Based on the scatterplot of depth vs. price, most diamonds are between what values of depth? 60 - 64

7. Correlation - Price and Depth

```
with(diamonds, cor.test(price, depth))
```

```
##
##  Pearson's product-moment correlation
##
## data:  price and depth
## t = -2.473, df = 53938, p-value = 0.0134
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.019084756 -0.002208537
## sample estimates:
##        cor
## -0.0106474
```

What is the correlation of depth vs. price? -0.01

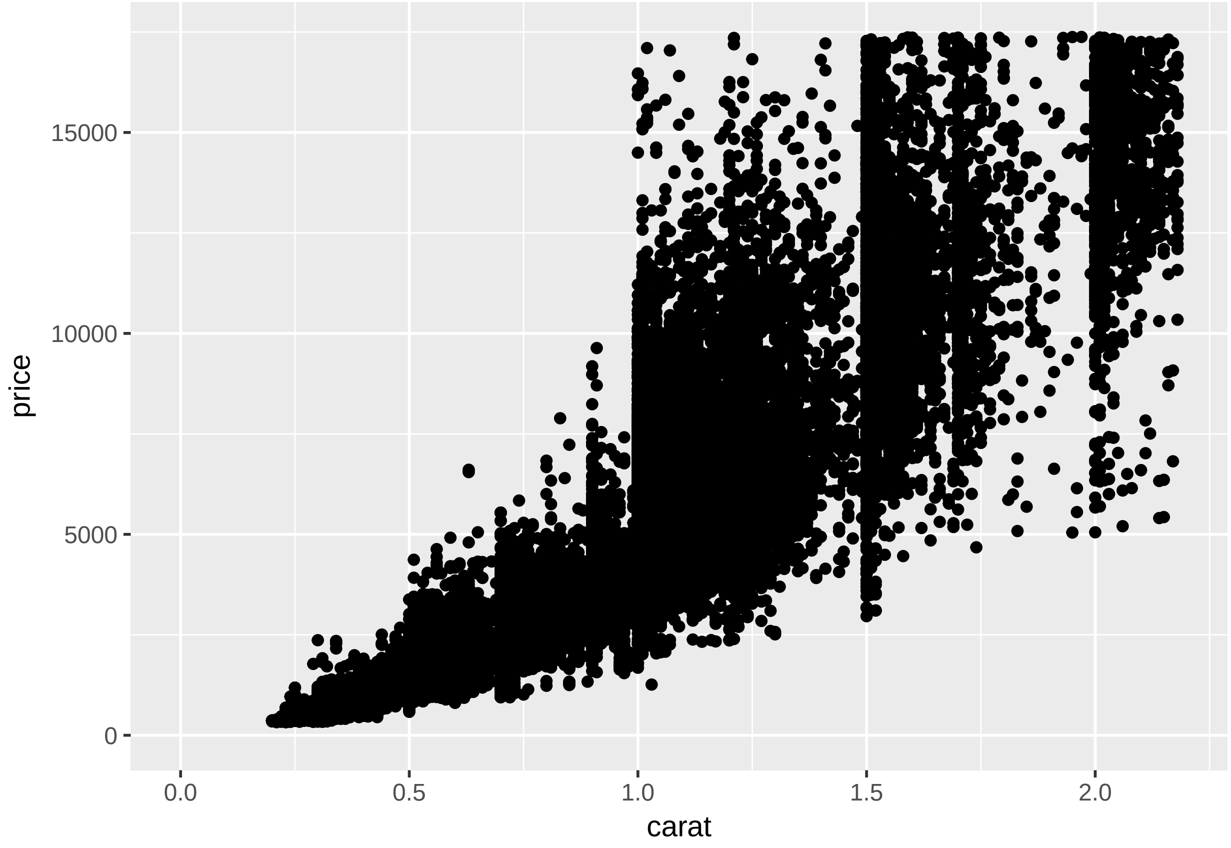
Based on the correlation coefficient woul dyou use depth to predict the price of a diamond? No

Why? Because a lower coefficient inidcates that the two variables are not closely linked.

8. Price vs. Carat

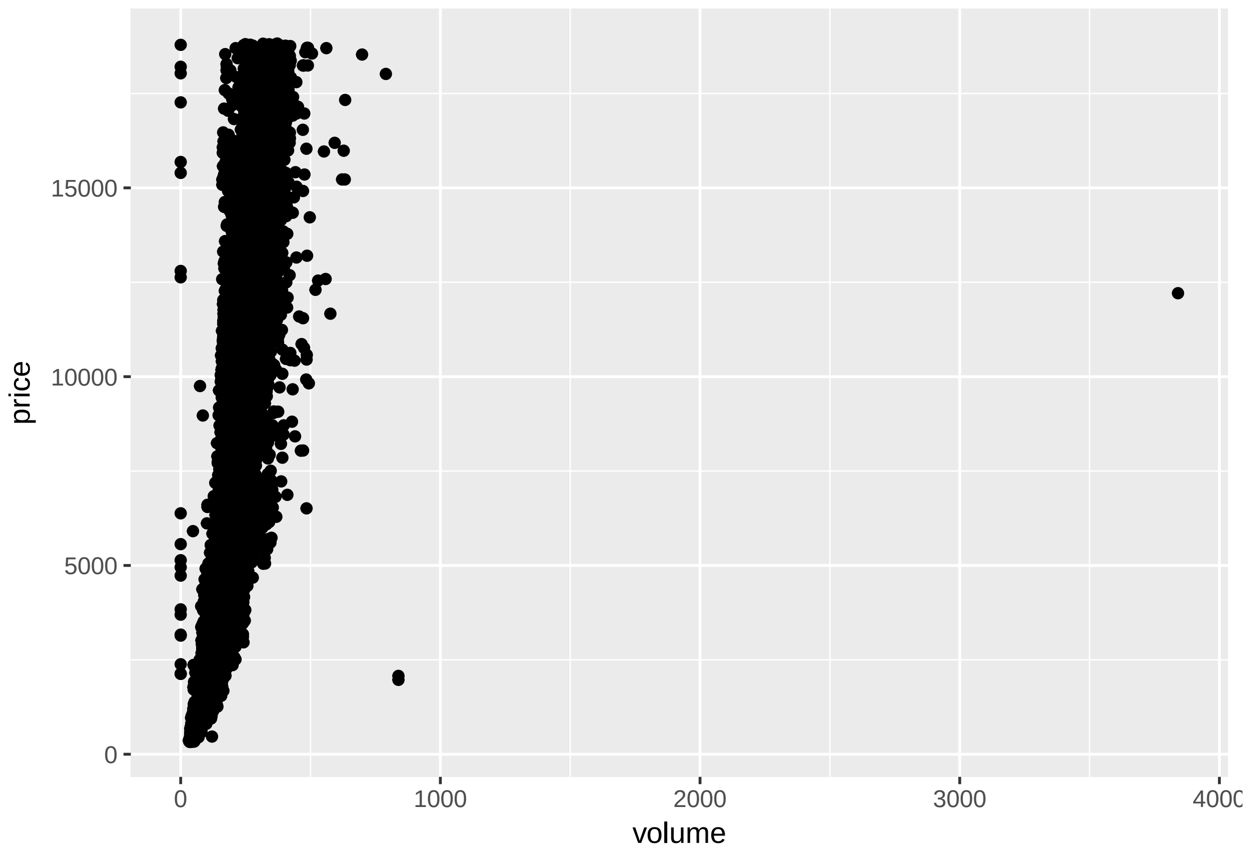
```
ggplot(aes(x = carat, y = price), data = diamonds) +
  geom_point() +
  xlim(0, quantile(diamonds$carat, 0.99)) +
  ylim(0, quantile(diamonds$price, 0.99))
```

```
## Warning: Removed 926 rows containing missing values (geom_point).
```



9. Price vs. Volume

```
diamonds$volume = (diamonds$x * diamonds$y * diamonds$z)

ggplot(aes(x = volume, y = price), data = diamonds) +
  geom_point()
```



10. Findings - Price vs. Volume

What are your observations from the price vs. volume scatterplot? There are some major outliers on the volume scale. Other than that the trend at least appears to be exponential price increase as volume increases.

11. Correlations on Subsets

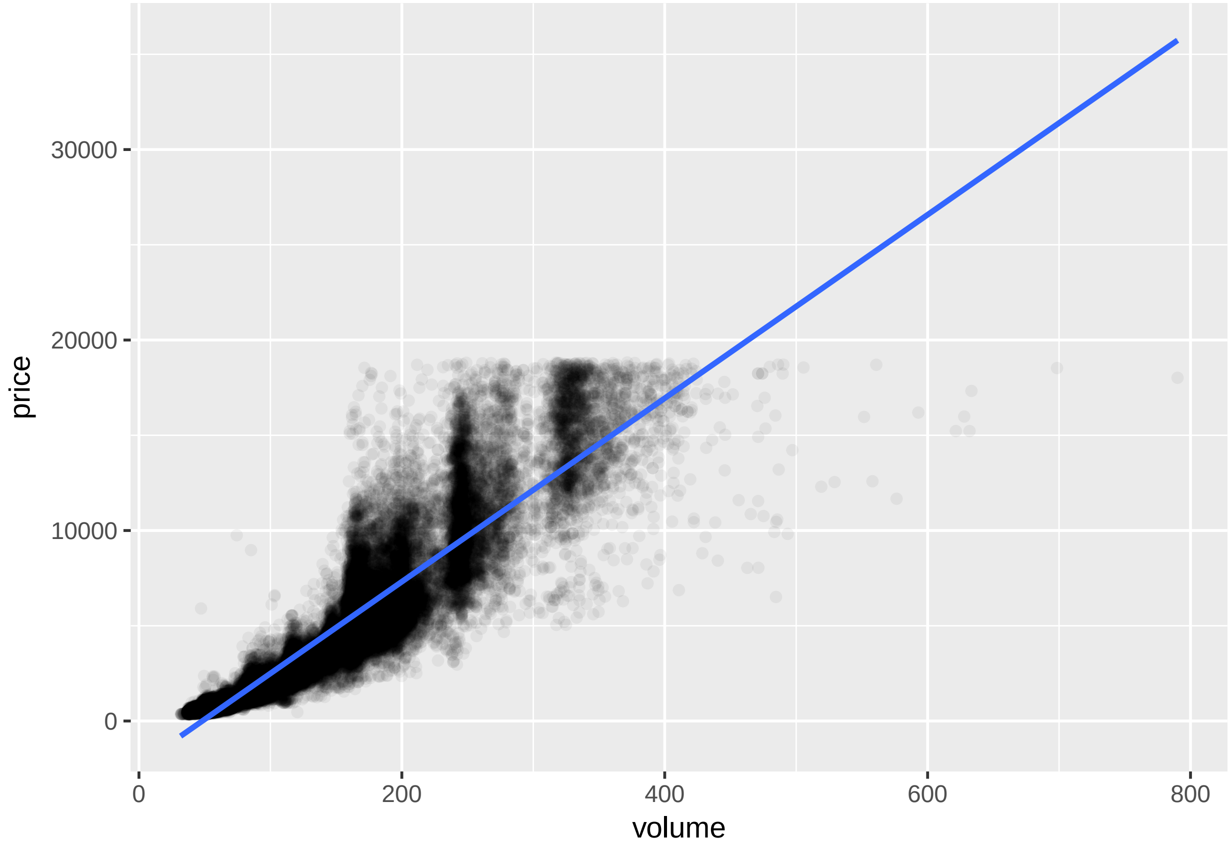
What's the correlation of price and volume? Exclude diamonds that have a volume of 0 or that are greater than or equal to 800.

```r
with(subset(diamonds, volume != 0 & volume < 800), cor.test(price, volume))
```

```
##
##  Pearson's product-moment correlation
##
## data:  price and volume
## t = 559.19, df = 53915, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9222944 0.9247772
## sample estimates:
##       cor
## 0.9235455
```

12. Adjustments - Price vs. Volume

```
ggplot(aes(x = volume, y = price), data = subset(diamonds, volume != 0 & volume < 800)) +
  geom_point(alpha = 1/20) +
  geom_smooth(method = 'lm')
```



No it is not helpful to look at the linear smooth in this case because it does not fit the data very well.

13. Mean Price by Clarity

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

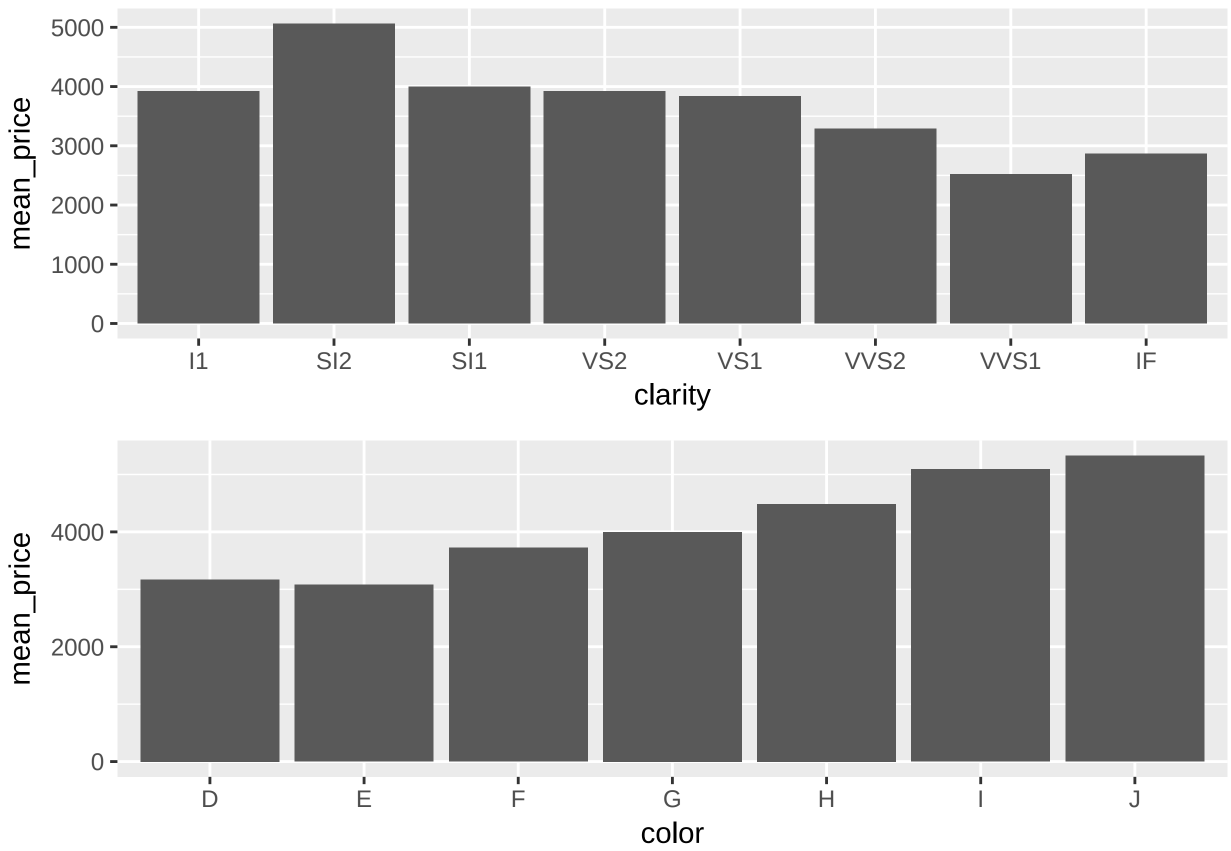
```
d_by_clarity <- group_by(diamonds, clarity)
diamondsByClarity <- summarize(
  d_by_clarity,
  mean_price = mean(price),
  median_price = median(price),
  min_price = min(price),
  max_price = max(price),
  n = n()
)
```

14. Bar Charts of Mean Price

```
data(diamonds)
library(dplyr)

diamonds_by_clarity <- group_by(diamonds, clarity)
diamonds_mp_by_clarity <- summarise(diamonds_by_clarity, mean_price = mean(price))

diamonds_by_color <- group_by(diamonds, color)
diamonds_mp_by_color <- summarise(diamonds_by_color, mean_price = mean(price))

library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
p1 <- ggplot(aes(x = clarity, y = mean_price), data = diamonds_mp_by_clarity) +
  geom_bar(stat = "identity")
p2 <- ggplot(aes(x = color, y = mean_price), data = diamonds_mp_by_color) +
  geom_bar(stat = "identity")
grid.arrange(p1, p2)
```



15. Trends in Mean Price

What do you notice in each of the bar charts for mean price by clarity and mean price by color?

In the clarity chart there is a downward trend from SI2 to WS1 but both of the end clarities. (I1 is lower than SI2 and IF is higher than WS1) In the color chart there is a gradual upwards trend from D to J with a slight dip at E.

16. Gapminder Revisited

The Gapminder website contains over 500 data sets with information about the world's population. Your task is to continue the investigation you did at the end of Problem Set 3 or you can start fresh and choose a different data set from Gapminder.

If you're feeling adventurous or want to try some data munging see if you can find a data set or scrape one from the web.

In your investigation, examine pairs of variable and create 2-5 plots that make use of the techniques from Lesson 4.

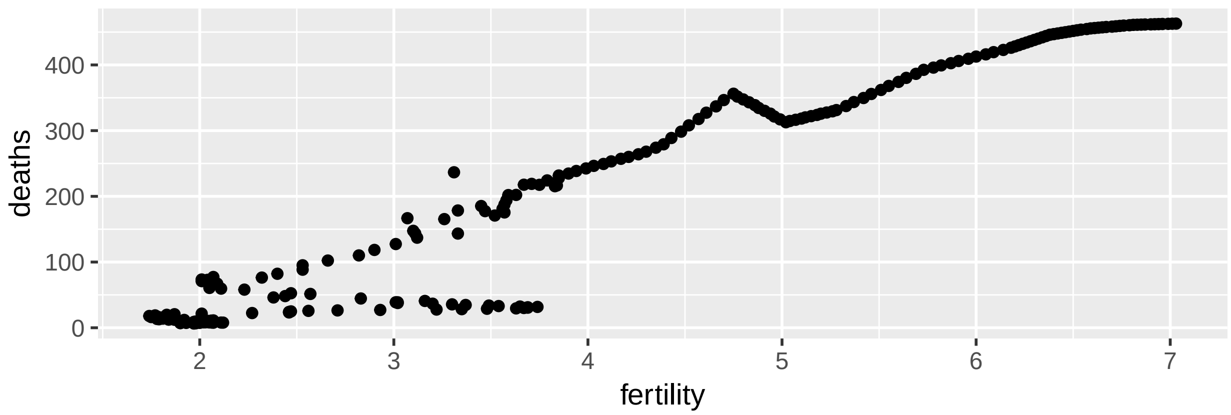
You can find a link to the Gapminder website in the Instructor Notes.

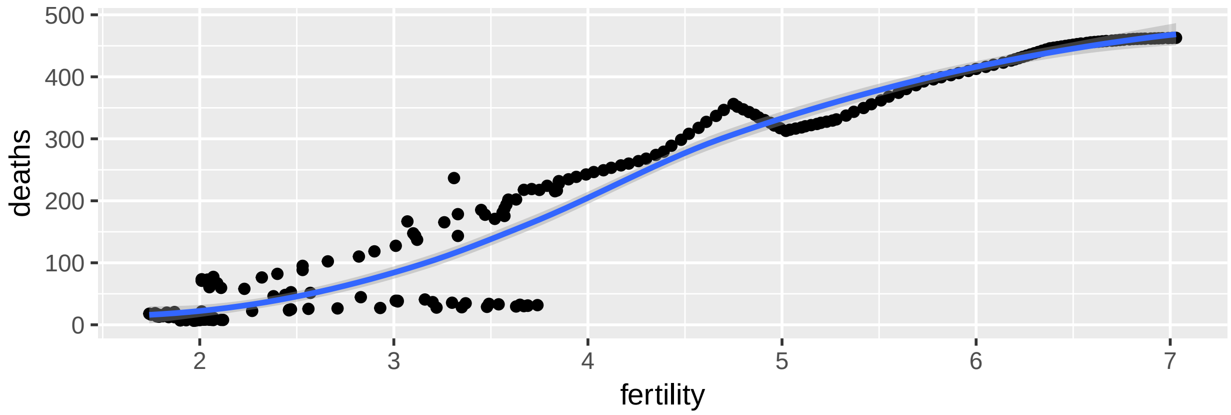
```r
data <- read.csv('indicator gapminder under5mortality.csv')
fertility <- read.csv('total_fertility.csv')
library(tidyr)
library(gridExtra)
library(reshape)
```

```
##
## Attaching package: 'reshape'

## The following objects are masked from 'package:tidyr':
##
##     expand, smiths

## The following object is masked from 'package:dplyr':
##
##     rename
```

```r
data <- melt(data, id = ("X"))
data <- cast(data, variable ~ X, mean)
data <- data[1:216,]
#data

fertility <- melt(fertility, id = ("X"))
fertility <- cast(fertility, variable ~ X, mean)
#fertility

us_data <- data.frame(
  year = fertility$variable,
  fertility = fertility$`United States`,
 deaths = data$`United States`
)

p1 <- ggplot(aes(x = fertility, y = deaths), data = us_data) +
  geom_point()
p2 <- ggplot(aes(x = fertility, y = deaths), data = us_data) +
  geom_point() +
  geom_smooth()

grid.arrange(p1, p2)
```

```
## `geom_smooth()` using method = 'loess'
```


deaths
0
100
200
300
400
2
3
4
5
6
7
fertility


deaths
0
100
200
300
400
500
2
3
4
5
6
7
fertility