MA117 - Worksheet 10
Two Numerical Variables
Week 3, Tuesday

Problem 1. The `tidyverse` packages come equipped with a data frame called `diamonds`. It's a big data frame, recording 10 variables for 53490 different diamonds. If you type `diamonds` into the console, you'll see the beginning of this data frame. The 10 variables include the following:

- `price`: price in US dollars
- `x`: length in mm
- `y`: width in mm
- `z`: depth in mm
- `depth`: total depth percentage (ie, `z / mean(x, y)`)

(a) Make a plot of a diamond's `depth` against its *price* using the following.

```
ggplot(diamonds, aes(price, depth)) +
    geom_point() +
    geom_smooth(method = "lm")
```

Describe the relationship you see. You can also calculate an explicit formula for the best fit line, together with lots of other useful information, using:

```
summary(lm(depth ~ price, data = diamonds))
```

(b) What is the (multiple[1]) R^2 value? Interpret this R^2 value.

(c) What is the slope of the best fit line? What is the value of the t test statistic associated to the slope of this best fit line?

(d) What is the p-value of the data under the hypothesis that the best fit line has slope zero? Interpret this p-value. Does this p-value make sense alongside your scatterplot?

Problem 2. Here's a data set about species diversity on several Southeast Asian islands:

`https://sagrawalx.github.io/teaching/data/speciesdiversity.csv`

There are five variables.

- `Name` of the island
- `Area` of the island in sq km
- `Species` is the number of mammal species
- `logArea` is the natural log (base e) of `Area`
- `logSpecies` is the natural log (base e) of `Species`

(a) Do `Area` and `Species` seem like they have a linear relationship? What about `Area` and `logSpecies`? `logArea` and `Species`? `logArea` and `logSpecies`?

(b) (Challenging) Use your answers to part (a) to derive a formula that predicts the number of mammal species on an island based on the area of that island.

[1] The "multiple" R^2 is what we learned about earlier in this class. The "adjusted" R^2 is a slightly different number that's sometimes more useful.