HW 09 Written Work

Graded

Student

Sangwon Ji

Total Points

38 / 38 pts

Question 1

1.3 **5** / 5 pts

✔ **+ 5 pts** Correct graph

+ 0 pts Incorrect/blank

Question 2

1.4 **5** / 5 pts

✔ **+ 2 pts** Correct answer: 0.5

✔ **+ 3 pts** Correct explanation: The two variables are positively associated, so it's not -.5. The data roughly follow a line, so the correlation is probably closer to .5 than to 0.

+ 0 pts Incorrect/blank

Question 3

1.8 **5** / 5 pts

✔ **+ 2 pts** No

✔ **+ 3 pts** Correct explanation: we have absolutely no information on the triple jump distances in any remote region near 18.29 meters, so it's not smart to make an estimate for it based on this data that is outside our observed range.

+ 0 pts Incorrect/blank

Question 4

2.1 **5** / 5 pts

✔ **+ 5 pts** Correct

+ 0 pts Incorrect/blank

Question 5

2.4 **5** / 5 pts

✔ **+ 5 pts** Correct graph

+ 0 pts Incorrect/blank

Question 6

2.5 **5** / 5 pts

✔ **+ 2 pts** Correct answer: no, this is not a good model

✔ **+ 3 pts** Correct explanation: The true data are not even close to being linear.

+ 0 pts Incorrect/blank

Question 7

3.4 **5** / 5 pts

✔ **+ 5 pts** Correct explanation: The new predictor is a horizontal line that passes through the average value for outcome. Therefore it does not minimize least squared error, as only the regression line is the unique straight line that minimizes least squared error among all straight lines.

+ 0 pts Incorrect/blank

Question 8

3.9 **3** / 3 pts

✔ **+ 3 pts** Reasonable explanation: The regression line is the unique straight line (in other words, the unique slope/intercept pair) that minimizes RMSE. Therefore, we can also find the regression line by finding the slope and intercept values that minimize RMSE

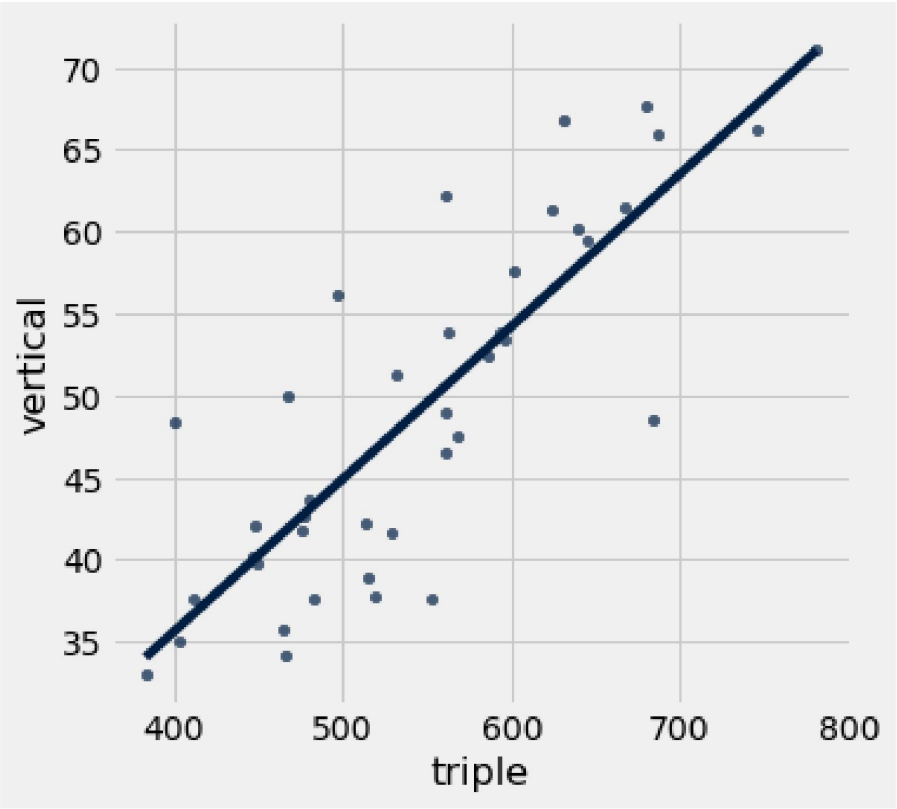
+ 0 pts Incorrect/blank

Question 1.3. Before running a regression, it's important to see what the data looks like, because our eyes are good at picking out unusual patterns in data. Draw a scatter plot, **that includes the regression line**, with the triple jump distances on the horizontal axis and the vertical jump heights on vertical axis. **(5 points)**

See the documentation on `scatter` here for instructions on how to have Python draw the regression line automatically.

Hint: The `fit_line` argument may be useful here!

```
In [8]: jumps.scatter("triple", "vertical", fit_line=True)
```



Question 1.4. Does the correlation coefficient r look closest to 0, .5, or -.5? Explain. **(5 points)**

The correlation coefficient r looks closest to 0.5 because the association between the varaibles is positive also the slope of the line looks its closest to 0.5.

Question 1.8. Do you think it makes sense to use this line to predict Edwards' vertical jump? **(5 points)**

Hint: Compare Edwards' triple jump distance to the triple jump distances in `jumps`. Is it relatively similar to the rest of the data (shown in Question 1.3)?

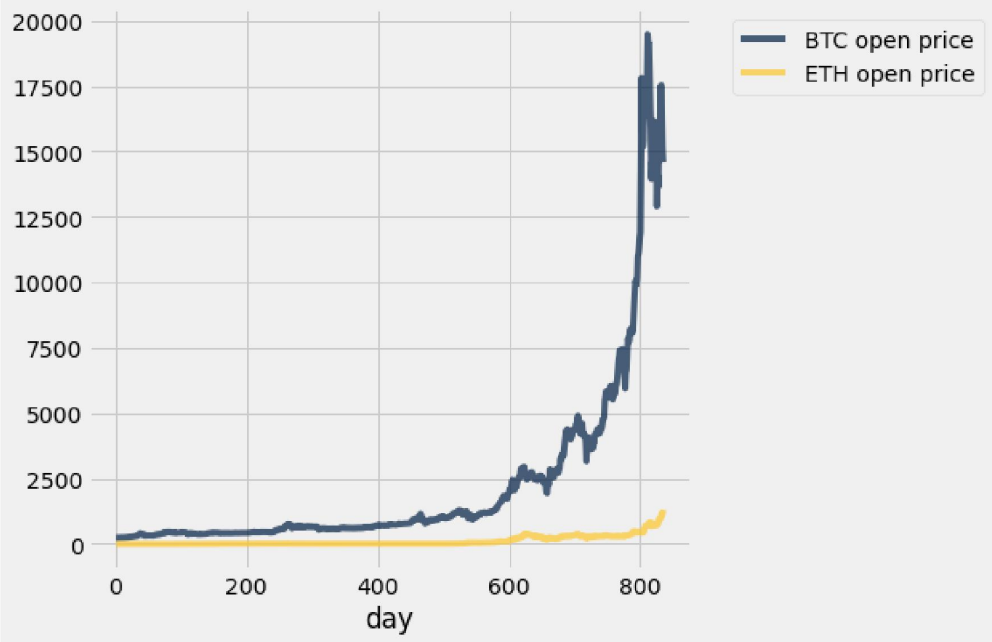
I think it doesn't make sense to use this line to predict Edwards' vertical jump. Compared to the data shown in question 1.3, the jump distance by Edward is not similar to the data we have, so it is not likely to compare them.

Question 2.1. In the cell below, create an overlaid line plot that visualizes the BTC and ETH open prices as a function of the day. Both BTC and ETH open prices should be plotted on the same graph. **(5 points)**

Hint: Section 7.3 in the textbook might be helpful!

```
In [17]: # Create a line plot of btc and eth open prices as a function of time
         BTC_line = btc.select("day", "open").relabeled("open", "BTC open price")
         ETH_line = eth.select("day", "open").relabeled("open", "ETH open price")

         BTC_ETH = BTC_line.join("day", ETH_line)
         BTC_ETH.plot("day")
```




```
In [18]: BTC_ETH.show()
```

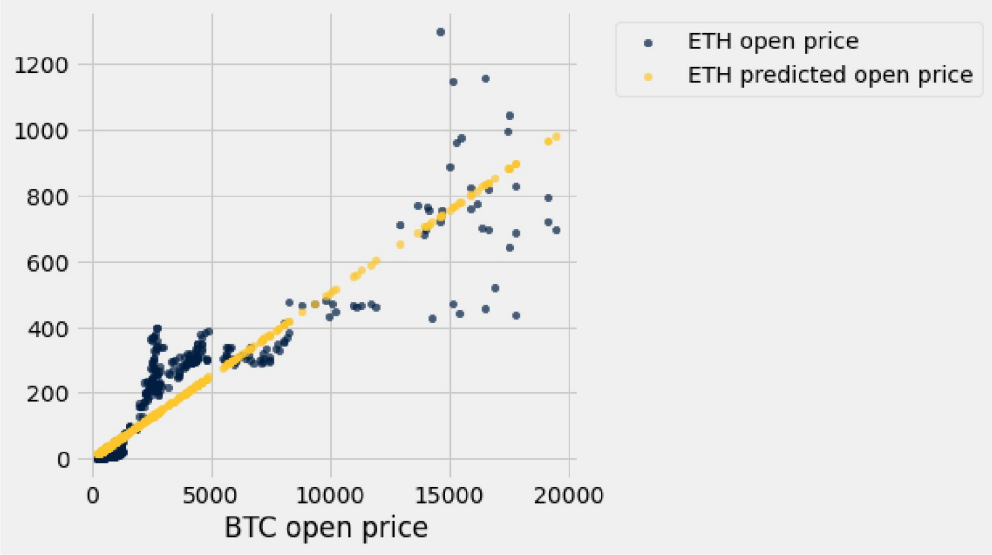
```
<IPython.core.display.HTML object>
```

Question 2.4. Now, using the `eth_predictor` function you just defined, make a scatter plot with BTC prices along the x-axis and both real and predicted ETH prices along the y-axis. The color of the dots for the real ETH prices should be different from the color for the predicted ETH prices. **(5 points)**

Hint 1: An example of such a scatter plot is generated can be found here.

Hint 2: Think about the table that must be produced and used to generate this scatter plot. What data should the columns represent? Based on the data that you need, how many columns should be present in this table? Also, what should each row represent? Constructing the table will be the main part of this question; once you have this table, generating the scatter plot should be straightforward as usual.

```
In [24]: btc_open = BTC_ETH.column("BTC open price")
         eth_pred = BTC_ETH.apply(eth_predictor, "BTC open price")
         eth_pred_actual = BTC_ETH.column("ETH open price")
         predicted_ETH = BTC_ETH_price.with_column("ETH predicted open price", eth_pred)
         predicted_ETH.scatter("BTC open price")
```



Question 2.5. Considering the shape of the scatter plot of the true data, is the model we used reasonable? If so, what features or characteristics make this model reasonable? If not, what features or characteristics make it unreasonable? **(5 points)**

Considering the shape of the catter plot of the ture data, there seems to be a positive relation between them, and the prediction seems to be accurate when the BTC open price is lower than 10000. However, as it goes over, it doesn't seem that accurate to the predcition we had below 10000. So, I think the model is reasonable below 10000, but after that, it doensn't seem reasonable enough. Therefore, even though there is a positive relation between, I think the model in all is not reasonable enough.

Question 3.4. Suppose that we create another model that simply predicts the average outcome regardless of the value for spread. Does this new model minimize the least squared error? Why or why not? **(5 points)**

The new model won't minimize the least squared error. The model that simply predicts the average outcome regardless of the value for spread won't be a straight line but a horizontal one that is not the regression line, wheer the line is the unique straight line.

Question 3.9. The slope and intercept pair you found in Question 3.8 should be very similar to the values that you found in Question 3.3. Why were we able to minimize RMSE to find the same slope and intercept from the previous formulas? **(3 points)**

Borrowing the definition from the question above, Regression line is the unique, the only straight line that minimizes root mean squared error among all possible fit lines. The Slope and Intercept we used in Question 3.3 were to find the regression line. Finding the Slope and Intercept that minimize root mean squared error, we can also find the regression line. That is, giving the same results since we are using the minimize root mean squared error and how we were able to do it from the same slope and intercept from the previous formulas.