DATA C8
Summer 2026

FOUNDATIONS OF DATA SCIENCE

FINAL

11:10AM-2:00PM FRIDAY, AUGUST 14TH 2026

PRINT Your Name: Data C8

PRINT Your Student ID: _____

PRINT Your Exam Room: _____

PRINT the Name of Person to your Left: _____

PRINT the Name of Person to your Right: _____

PRINT Your GSI's Name : _____

INSTRUCTIONS

You have **170 minutes** to complete the exam. There are **6 questions** and **16 pages** on this exam, including this cover page.

Question	1	2	3	4	5	6	Total
Points	12	24	15	18	30	1	100

- This exam is closed book, closed computer and closed calculator, except the Reference Sheet provided for you.
- You may only have with you: a pencil, an eraser and your student ID, unless you have pre-approved accommodations.
- If you need to use the restroom, bring your phone, exam, reference sheet and student ID to the front of the room.
- For written questions:
  - answers written outside the boxes provided will not be graded;
  - if your answer is ambiguous or you provide multiple answers, the worst interpretation will be graded.
- For coding questions:
  - blank spaces may include multiple arguments or functions per blank, but your solution must use every blank available;
  - you may assume the `datascience` and `numpy` libraries are imported, as seen in class;
  - the use of **any code** which has not been taught in this offering of the course is not allowed and will result in zero credit.
- For multiple choice questions, see question types and instructions below.

Questions with **circular bubbles**: you may select only **1 choice**. Questions with **square boxes**: you may select **1 or more choices**.

○ Unselected option (completely unfilled)

● Single option selected (completely filled)

■ You may select multiple squares

■ as long as they are completely filled

You must fill in the bubbles **completely**. Ticks, crosses, or other check marks will **not** receive credit.

HONOR CODE

"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."

SIGN Your Name: _____

1. [12.0 points] General

Read each question carefully and answer accordingly.

(a) [6.0 pts] For which of the following tasks that we have performed throughout the semester has a `for` loop proven useful?

- [x] Creating a bootstrapped confidence interval for the slope of the true line relating two numerical variables
- [x] Calculating the Euclidean distance between a testing point and each of the points in the training set
- [ ] Finding the average of a numerical variable for each category of a categorical variable
- [x] Simulating 10,000 rolls of a fair, six-sided die
- [x] Generating test statistics under a null hypothesis
- [ ] Creating a line plot to visualize two numerical variables in a table

(b) [4.0 pts] Consider the Python function `f` below. What does the call expression `f(make_array(1,0,-1))` evaluate to?

```
def f(arr):
    if sum(arr) == True:
        return 'A'
    elif sum(arr) == False:
        return 'B'
    else:
        return 'C'
```

○ `A` ● `B` ○ `C` ○ The call expression will produce an error.

(c) [2.0 pts] Oski has been getting a lot of attention on his SoundCloud account but is unsure whether the traffic is coming from actual music listeners or bot (fake) accounts. He consults historical data to help him.

- 70 percent of SoundCloud accounts have been found to be bot accounts.
  - Of these accounts, 80 percent send a message directly after following.
- 30 percent of SoundCloud accounts belong to actual music listeners.
  - Of these accounts, 10 percent send a message directly after following.

Suppose Oski gets a new, random follower today, and this follower did not send him a message. Should Oski predict that this follower is a real music listener or a bot account? Give an answer and justify it by showing your work in the box below.

● Real music listener

○ Bot account

P(real | no message) ∝ 0.30 × 0.90 = 0.27 ✓

P(bot | no message) ∝ 0.70 × 0.20 = 0.14

2. [24.0 points] Licensed Squirrels

The squirrels on the UC Berkeley campus are required to take an exam every year to renew a squirrel license that allows them to live on campus. Below are historical statistics of the performance on the exam by version.

Exam	Mean	Standard Deviation
A	75	5
B	80	6

(a) [5.0 pts] A studious squirrel named Andrew knows the material so well that he takes the exams mostly for fun. He takes the exam three times in a row, hoping to receive both versions A and B. Each squirrel has an equal chance of receiving a given exam version on any sitting, independent of other sittings. What's the probability that Andrew only gets to take version A?

- ○ $\frac{1}{3}$
- ○ $\frac{1}{2}$
- ○ $\frac{1}{4}$
- ● $\frac{1}{8}$
- ○ $\frac{1}{6}$
- ○ The correct probability is not here.

$P(AAA) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 1/8$

(b) [4.0 pts] Two friendly squirrels, Ramisha and Thomas, decided to sit for the exam together.

- Ramisha received exam A and scored a **78**.
- Thomas received exam B and scored an **83**.

Who performed better relative to the exam they took? *Show your work in the box below.*

● Ramisha ○ Thomas ○ The squirrels performed equally well.

R: $\frac{78-75}{5} = \frac{3}{5}$ T: $\frac{83-80}{6} = 1/2$

(c) [3.0 pts] In the box below, calculate the score that Ramisha would have received on exam B based on her performance on exam A. *Show your work.*

$3/5 = \frac{x-80}{6} \Leftrightarrow \frac{18}{5} = x-80 \Rightarrow x = 80 + 18/5$

$= 80 + \frac{15}{5} + \frac{3}{5}$

$= 83.6$

(d) [2.0 pts] Far away from campus, a squirrel named Edwin has hacked into the university database to learn more about the distribution of the license exam scores. Edwin cannot access all of the scores, but finds a way to take a random sample (with replacement) of 10,000 exam A scores. His hope is to use his sample mean to estimate the true mean of exam A scores. Between what values can Edwin expect 95 percent of all possible sample means to lie?

- ○ $75 \pm 2 * \frac{5}{10000}$
- ○ $75 \pm \frac{5}{10000}$
- ○ $75 \pm \frac{5}{100}$
- ○ 75 ± 5
- ● $75 \pm 2 * \frac{5}{100}$
- ○ $75 \pm 2 * 5$
- ○ The correct answer is not here.

(e) Another squirrel named Mariel has also hacked into the university database. Like Edwin, she can't access all of the scores, but she is interested in learning about the mean exam score for version B. She takes a random sample (without replacement) of 10,000 exam B scores.

i. [5.0 pts] Mariel uses her sample and calculates a 99 percent confidence bootstrapped confidence interval as [74,76]. Interpret the interval within the context of the problem.

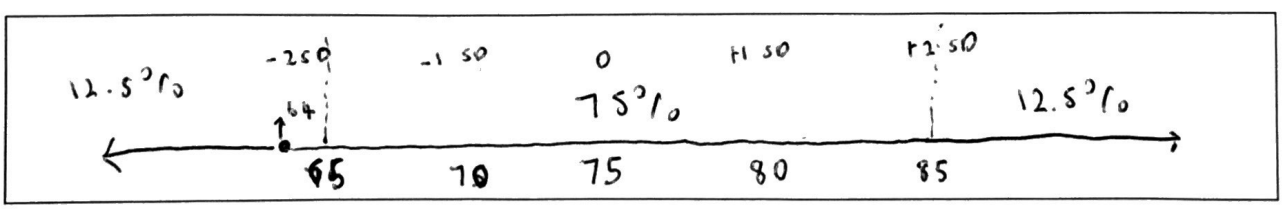
- ○ Mariel is 99 percent confident that all exam B scores in the university database lie between 74 and 76.
- ○ Mariel is 99 percent confident that all exam B scores in her sample lie between 74 and 76.
- ● Mariel is 99 percent confident that the mean exam B score in the university database lies between 74 and 76.
- ○ Mariel is 99 percent confident that the mean exam B score in her sample lies between 74 and 76.
- ○ The correct answer is not here.

ii. [3.0 pts] What would happen to Mariel's 99 percent confidence interval if she increased her sample size?

- ○ The confidence level would increase.
- ○ The confidence level would decrease.
- ○ The width of the interval would increase.
- ● The width of the interval would decrease.
- ○ Both the confidence level and the width of the interval would change in some way.

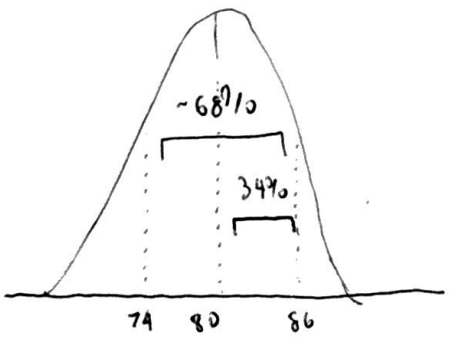
(f) [1.0 pt] This year, a score on exam A is considered passing if it lies in the top 87.5% of all historical data (data from last year and before then). A squirrel named Jeremy took exam A and scored a **64** on his exam this year. Did Jeremy pass his exam and renew his squirrel license? *Show your work in the box below.*

- ○ Yes
- ● No



(g) [1.0 pt] Suppose it is known that the true probability distribution of exam B is Normal. A squirrel named Sangjun doesn't have time to take the exam and asks the university to assign him a random score according to the probability distribution of exam B. What is the probability that the university gives him a score between 80 and 86?

- ○ 0.18
- ○ 0.5
- ○ 0.32
- ○ 0.16
- ● The correct answer is not here.



3. [15.0 points] Prediction

Review the incomplete Python function `my_split` below, which splits a table `tbl` into two new tables: `train` and `test`. Answer the questions that follow.

```
def my_split(tbl, pct):

    tbl = tbl.with_column('Row Number', _____[1]_____)


    train_rows = np.random.choice(tbl._____[2]_____,
                                  size=round(tbl.num_rows*(pct/100)),
                                  replace=_____[3]_____)
    train = tbl._____[4]_____

    test = tbl._____[5]_____('Row Number', are.not_contained_in(_____[6]_____))
    return train, test
```

(a) Fill in the blanks to complete the `my_split` function.

i. [1.0 pt] Fill in blank `[1]`.

np.arange(tbl.num_rows)

ii. [1.0 pt] Fill in blank `[2]`.

column('Row Number')

iii. [1.0 pt] Fill in blank `[3]`.

False

iv. [1.0 pt] Fill in blank `[4]`.

take(train_rows)

v. [1.0 pt] Fill in blank `[5]`.

where

vi. [1.0 pt] Fill in blank `[6]`.

train_rows

(b) [4.0 pts] In two to three sentences, describe the purpose of a function like `my_split` (when is it used and why is it needed).

my_split is used in preparation to fit a predictive model. It is needed because we would like to test the performance of our model on data it hasn't seen yet. Fitting the model on only some of the data and leaving the rest for the purpose of evaluation accomplishes this goal.

(c) Assume that `my_split` has been used to obtain a `train` and `test` set, and that an end goal is to build a multiple linear regression model to predict a numerical outcome y using a set of four numerical predictors, x_1, x_2, x_3 and x_4.

i. [3.0 pts] In the box below, write out the mathematical form of the multiple linear regression model.

$y = a + bx_1 + cx_2 + dx_3 + ex_4$

ii. [2.0 pts] Describe how the `minimize` function in the `datascience` library can be used to find the exact values for the slopes of the model.

- ● Call `minimize` on a function which calculates the RMSE for the training set. Use the slopes that are output to calculate the RMSE on the testing set.
- ○ Call `minimize` on a function which calculates the RMSE for the testing set. Use the slopes that are output to calculate the RMSE on the training set.
- ○ Call a function which calculates the RMSE for the training set on `minimize`. Use the slopes that are output to calculate the RMSE on the testing set.
- ○ Call a function which calculates the RMSE for the testing set on `minimize`. Use the slopes that are output to calculate the RMSE on the training set.

Initials:

4. [18.0 points] The pen is mightier than the word

Read carefully the following excerpt from a research article by Rutchick et al. (2010) and answer the questions that follow.

"One hundred and twenty-nine students [participants] were recruited from introductory psychology courses... participants were given a one-page essay [written by an eighth grader]. The essay contained no grammatical or spelling errors, although (as might be expected of an eighth grader's work) there were a number of suboptimal word choices...

...Participants were instructed to evaluate the essay, indicating flaws (such as points at which the phrasing or word choice could be improved), and to award the essay a grade from zero (worst) to one hundred (best). Participants were... randomly assigned to use either a red or blue pen in evaluating the essay...

Because the color red is implicitly associated with avoidance and failure, and red pens specifically have long been associated with errors, we propose that exposure to a red pen activates the concepts of errors, poor performance, and evaluative harshness... We hypothesized that using a red pen to complete this task would cause participants to assign lower grades than using a blue pen.

"

The `pen` table below contains the approximate data from the study. The first few rows are provided below for reference.

Pen Color	Score Given
Red	72
Blue	97
Blue	98
Red	83

`(125 rows omitted...)`

In addition, some summary statistics are provided.

Pen Color	Count	Score Mean
Red	65	76
Blue	64	80

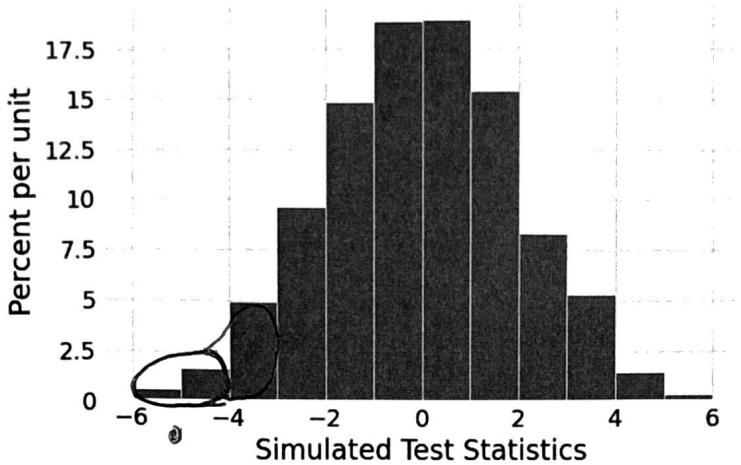
(a) [4.0 pts] Write an appropriate null hypothesis for the test below in at least one sentence.

> The distribution of potential red pen scores is the same as the distribution of potential blue pen scores.

(b) [3.0 pts] How should a simulated test statistic for this hypothesis test be computed under the null hypothesis?

- ○ Bootstrap the Pen Color column and compute the appropriate test statistic with the shuffled table.
- ● Shuffle the Pen Color column and compute the appropriate test statistic with the shuffled table.
- ○ Collect a new set of participants, randomly assign 65 to grade with red pens and 64 to grade with blue pens, and then compute the appropriate test statistic.
- ○ None of these methods are appropriate to simulate the test statistic under the null hypothesis.

(c) [3.0 pts] The following is a visualization of simulated test statistics under the null hypothesis. Shade the area on the histogram corresponding to the p-value.



circling last ~~two~~ three bin also acceptable

(d) [3.0 pts] The researchers obtained a p-value of 0.048. Assuming a p-value cutoff of 0.05, state the conclusion of the hypothesis test.

- ◯ The evidence supports the idea that, in the population of graders, scores given with red pens are lower than scores given with blue pens, on average.
- ◯ The evidence supports the idea that, among the 129 graders on which data was collected, scores given with red pens are lower than scores given with blue pens, on average.
- ◯ The evidence supports the idea that, in the population, scores given with blue pens are lower than scores given with red pens, on average.
- ● An appropriate conclusion is not here.

(e) [5.0 pts] Based on the results and nature of the study, can the researchers make a causal claim (that is, can they say that the color of the pen used to grade *caused* scores to be lower)?

Give an answer and justify it in two to three sentences.

The researchers can make a causal claim. Because participants were randomly assigned a pen color, any other factors which ~~could~~ might impact scores ought to be balanced among pen color groups. Therefore, since the ~~code~~ conclusion implies an association, it can also imply causation

5. [30.0 points] Palmer Penguins

The `penguins` table contains information on 333 penguins that were surveyed in Antarctica, specifically in the Palmer Archipelago. A four row excerpt, along with column descriptions, is below.

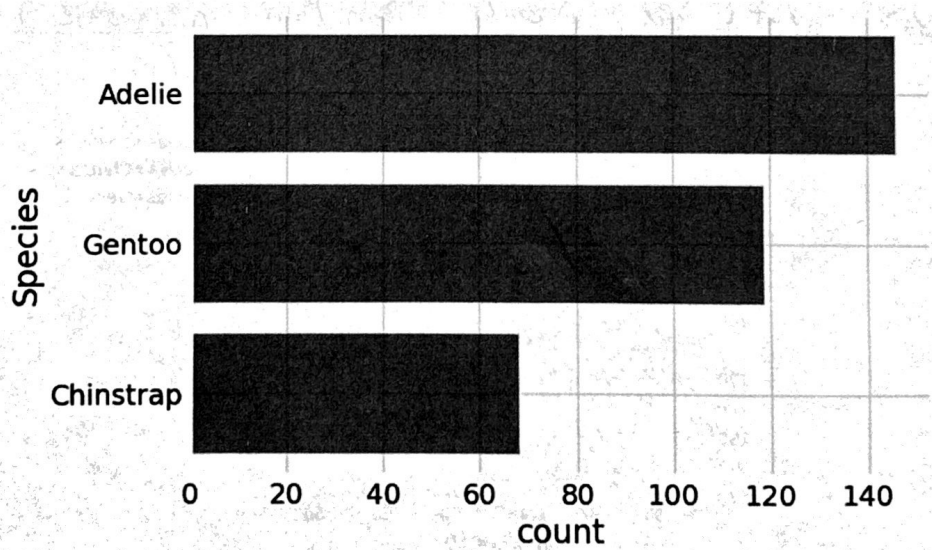
- **Species:** The species of penguin; one of "Adelie", "Chinstrap" or "Gentoo"
- **Island:** The island in the Archipelago on which the penguin was found; one of "Biscoe", "Dream" or "Torgersen".
- **Bill Length:** in millimeters.
- **Bill Depth:** in millimeters.
- **Flipper Length:** in millimeters.
- **Body Mass:** in grams.
- **Sex**: one of "Male" or "Female".

Species	Island	Bill Length	Bill Depth	Flipper Length	Body Mass	Sex
Chinstrap	Dream	49.2	18.2	195	4400	Male
Adelie	Torgersen	37.3	20.5	199	3775	Female
Adelie	Torgersen	39.5	17.4	186	3800	Female
Gentoo	Biscoe	46.1	13.2	211	4500	Female

`(329 rows omitted...)`

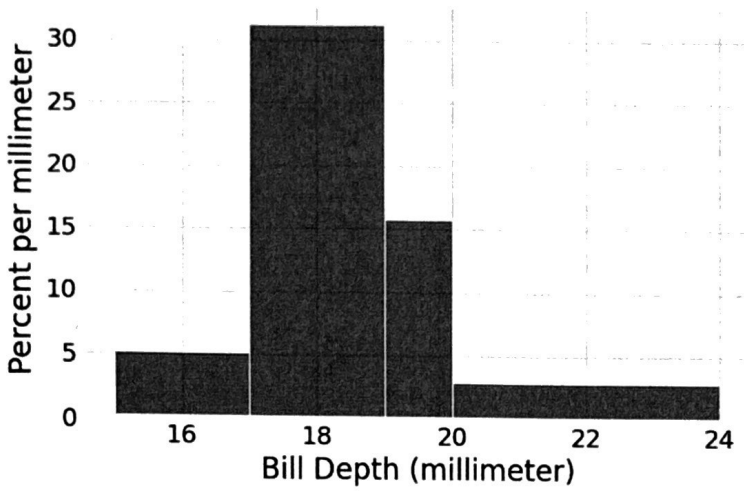
For the following set of questions, write one line of code to produce the exact output you see. Where necessary, additional detail will be given to help you.

(a) [2.0 pts]




```
penguins.group('Species').sort('count', descending=True).barh('Species')
```

(b) [2.0 pts] *The below visualization was created only with Adelie penguins.*




```
penguins.where('Species', 'Adelie').hist('Bill Depth', unit='millimeter', bins=
```

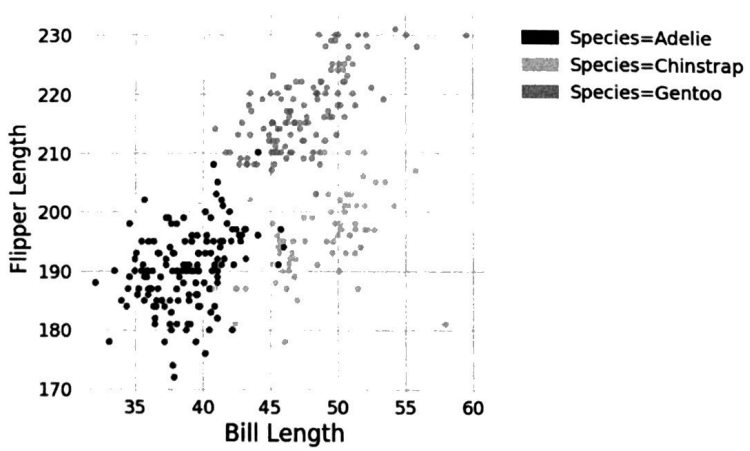
make_array(15, 17, 19, 21, 24))

(c) [2.0 pts] *The numbers in the cells of the below data structure are* ***averages****.*

Species	Male	Female
Adelie	3368	4043
Chinstrap	3527	3938
Gentoo	4679	5484

```
penguins.pivot('Sex', 'Species', 'Body Mass', np.average)
```

(d) [2.0 pts]




```
penguins.scatter('Bill Length', 'Flipper Length', group='Species')
```

(e) We will now perform k-nearest neighbors classification to predict the species of a penguin based on its bill length and flipper length.

43 205

i. [4.0 pts] Consider a penguin named Cody having a bill length of 43 millimeters and a flipper length of 205 millimeters. The table below contains the 10 penguins closest to Cody in the training set based on Euclidean distance. Using k-nearest neighbor classification with $k = 3$, what species of penguin would we predict Cody to be?

Class	Bill Length	Flipper Length
Gentoo	43	208
Gentoo	44	205
Adelie	41	202
Chinstrap	44	202
Gentoo	44	208
Adelie	41	205
Gentoo	44	208
Adelie	41	205
Gentoo	43	208
Gentoo	45	208

$\sqrt{0^2+3^2} = \sqrt{9} = 3$
$\sqrt{1^2+0^2} = \sqrt{1} = 1$
$\sqrt{2^2+3^2} = \sqrt{13}$
$\sqrt{1^2+3^2} = \sqrt{10}$
$\sqrt{1^2+3^2} = \sqrt{10}$
$\sqrt{2^2+0^2} = \sqrt{4} = 2$
$\sqrt{1^2+3^2} = \sqrt{10}$
$\sqrt{2^2+0^2} = \sqrt{4} = 2$
$\sqrt{0^2+3^2} = \sqrt{9} = 3$
$\sqrt{2^2+3^2} = \sqrt{13}$

● Adelie ○ Chinstrap ○ Gentoo

ii. [3.0 pts] The table below records the performance of k-nearest neighbor classification with $k = 3$ on the testing set. Using the table, calculate the value of a metric which can be used to evaluate the performance of the algorithm across all species.

Predicted	Adelie	Gentoo	Chinstrap
Adelie	41	1	2
Gentoo	1	15	1
Chinstrap	1	0	38

6/100

(f) Finally, we will focus specifically on Gentoo penguins. In particular, we will perform linear regression to estimate the flipper length of a Gentoo penguin using its bill length. Some useful summary statistics for the relevant variables among the Gentoo penguins are provided below.

r	Mean x	SD x	Mean y	SD y
0.66	47	3	217	7

i. *Using the summary statistics below,* calculate the slope and intercept of the linear regression line. Show your work in the box provided. *You do not need to simplify.*

1. [2.0 pts] Slope:

$0.66 \times \frac{7}{3} \sim \frac{14}{9}$ ✱

2. [2.0 pts] Intercept

$217 - 14/9 \cdot 47$

ii. [4.0 pts] Interpret the slope of your line in the context of the problem.

- ○ For every one millimeter increase in flipper length, we expect the bill length to increase by the slope value in millimeters.
- ● For every one millimeter increase in bill length, we expect the flipper length to increase by the slope value in millimeters.
- ○ Increasing the flipper length by the slope value in millimeters leads to an expected increase in bill length by one millimeter.
- ○ Increasing the bill length by the slope value in millimeters leads to an expected increase in flipper length by one millimeter.

iii. Use the `gentoo` table below (a four row excerpt is given) to answer the following questions, where necessary.

Bill Length	Flipper Length	Fitted Value	Residual
46.6	210	216	−6
51.5	230	223	7
46.5	213	216	−3
46.1	211	215	−4

`(115 rows omitted...)`

1. [4.0 pts] Write one line of code to create a scatter plot that can be used to assess the fit of the linear model to this data.

```
gentoo.scatter('Bill Length', 'Residual')
```

2. [3.0 pts] What in the scatter plot mentioned directly above would you be looking for to confirm that a linear model is the most appropriate fit to the data?

- ○ We should see a positive linear relationship between the variables in the scatter plot, since this is the relationship between x and y themselves.
- ○ We should see a negative linear relationship between the variables in the scatter plot, to complement the relationship between x and y.
- ● We should see no pattern between the variables in the scatter plot.
- ○ We should see a shape in the scatter other than a linear one.