Data 8 Foundations of Data Science
Fall 2018

FINAL EXAM SOLUTIONS

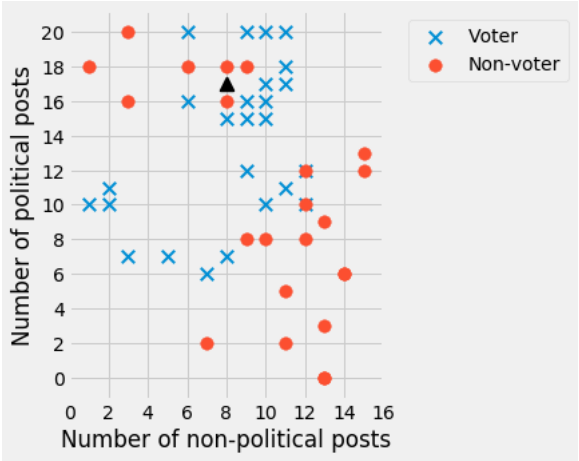
INSTRUCTIONS

- You have 3 hours to complete the exam.
- The exam is closed book, closed notes, closed computer, closed calculator, except one hand-written 8.5" × 11" sheet of notes of your own creation and the official final exam reference guide provided with the exam.
- Mark your answers **on the exam itself**. We will *not* grade answers written on scratch paper.
- For all Python code, you may assume that the statements `from datascience import *` and `import numpy as np` have been executed. Do not use features of the Python language that have not been described in this course.

Last name	
First name	
Student ID number	
Calcentral email (`_@berkeley.edu`)	
Lab GSI	
Your seat number & room	
Name of the person to your left	
Name of the person to your right	
All the work on this exam is my own. **(please sign)**	

1. (8 points) Classification

Candidate A decides to train a classifier to predict whether people will vote in the 2020 U.S. election or not. They gather data on voting records from the 2018 U.S. election and decide to use two features: the number of political and non-political posts on social media that the person made in the month leading up to the election. A scatter plot of their initial sample is shown below:



(a) (2 pt) The candidate is trying to classify the point at $(8, 17)$ shown as a triangle on the graph above. If they use a 3-nearest neighbor classifier, what will their classification be?

- ○ Voter
- ● Non-voter

(b) (4 pt) Suppose the candidate randomly divides the data into test and training sets (both much larger than the set shown above), and finds a test set accuracy of 94%. The candidate decides to apply their trained classifier to a test set from another country with lower rates of internet access. Should they expect the accuracy to be the same, higher or lower? Why?

- ○ Same
- ○ Higher
- ● Lower

Reason:

People might post less, so they will be in the lower-left corner, and accuracy there is lower.

(c) (2 pt) Instead of a k-nearest neighbor classifier, the candidate decides to use a d-distance classifier. In this classifier, instead of choosing the k closest neighbors, we'll instead choose all neighbors within a specified distance d (including points that are exactly d units away). If there are an equal number of points with both labels within that distance, choose whichever class you wish.

If $d = 5$, how would you classify the point at $(8, 17)$ shown as a triangle on the graph above?

- ● Voter
- ○ Non-voter

2. (15 points) Rent control

You find data on rents from a particular city, but the dataset is unlabeled and you don't know what city it is. The table `housing` contains the following data, which represents a random sample of 500 two-bedroom apartments from that city.

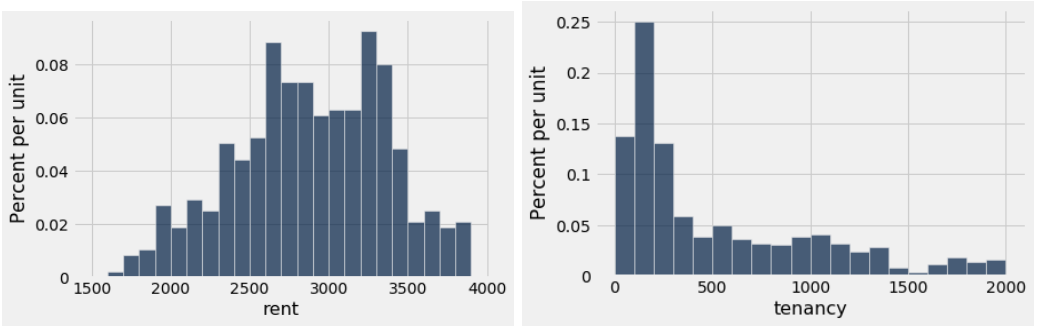
`housing`

tenancy	rent
340	2800
115	2775
1206	2100

`... (997 rows omitted)`

`tenancy` is the number of days that the tenants have lived in the apartment, and `rent` is the rent in dollars.

You produce the following two histograms. Assume that all points are shown in the histogram, and the areas of the bars in each one sum to 100%.



(a) (3 pt) If you only know the information above, which of the following are valid conclusions?

- ● The median rent for the sample is somewhere between $\$2,500$ and $\$3,500$.
- ○ There is a negative correlation between tenancy and rent.
- ● The mean tenancy is higher than the median tenancy.
- ○ The mean tenancy is lower than the median tenancy.

You compute the following quantities in your notebook. Assume you have the function `correlation(tbl, x, y)` from lecture already defined in your notebook.

```
In [236]: round(correlation(housing, 'rent', 'tenancy') ** 2, 3)
Out[236]: 0.214

In [237]: round(np.std(housing.column('rent')), 1)
Out[237]: 549.5

In [238]: round(np.std(housing.column('tenancy')), 1)
Out[238]: 521.5
```

For the remainder of the question, fill in each blank with a mathematical expression that computes the desired quantity. If the question is invalid or the quantity can't be computed from the information given, write "cannot be determined".

Your answers to parts (b)–(d) should **not** contain any Python code other than basic mathematical expressions. Do not use the `slope`, `intercept`, or `correlation` functions.

(b) **(3 pt)** The probability that the correlation between tenancy and rent in the population is 0.

Cannot be determined.

The correlation between tenancy and rent in the population is a fixed number, and so it doesn't make sense to ask about probabilities. We'd also accept 0%.

(c) **(3 pt)** The slope of the regression line for predicting **tenancy** from **rent**.

Cannot be determined.

The information provided isn't enough to determine whether the slope is positive or negative.

We also accepted $\pm\sqrt{0.214}\left(\frac{521.5}{549.5}\right)$, and gave partial credit for a positive or negative version of the above answer.

(d) **(3 pt)** The root mean square error (RMSE) of a regression line predicting **rent** from **tenancy**.

$\sqrt{1-0.214} * 549.5$

(e) **(3 pt)** You want to predict the rent for an apartment where the longest tenant has lived there for 3 years (1095 days). Late one night, you write the following code to bootstrap a 95% confidence interval for your prediction. When you wake up the next morning, you realize that you've made some mistakes. Correct each of the mistakes in the code below. For this question, you may assume that `slope(tbl, x, y)` and `intercept(tbl, x, y)` are already defined as in lecture.

TODO: write solution

```
predictions = make_array()


for i in np.arange(1000):


    samp = housing.sample(with_replacement=True)

              samp
              xxxxxxx
    m = slope(housing, 'tenancy', 'rent')

                  samp
                  xxxxxxx
    b = intercept(housing, 'tenancy', 'rent')

    predictions =
    np.append(predictions, m * 1095 + b)

                  2.5
                  x
left = percentile(5, predictions)


                   97.5
                   xx
right = percentile(95, predictions)
```

3. (10 points) Bandages

A journalist looks for the price that hospitals charge patients for bandages (like Band-Aids). She obtains the following anonymized dataset with information on the size of the hospital (large or small) and how much they charge in dollars. Assume that the data is a random sample from the population of all hospitals in the US.

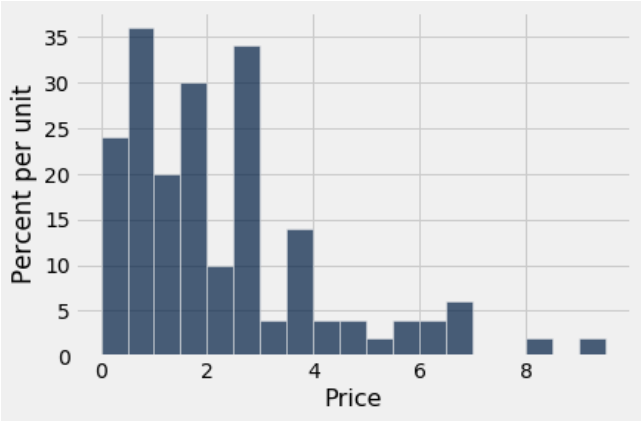
`bandages`

Size	Price
large	7.51
small	4.30
large	9.26

`... (497 rows omitted)`

She uses the following line of code to create the histogram shown below:

```
bandages.where('Size', 'small').hist('Price', bins=np.arange(0, 10, 0.5))
```



The area of the bars sums to 100%, and the highest price in the sample is \$9.71.

(a) (3 pt) What percentage of small hospitals have bandages that cost at least \$1.00 but less than \$2.00?

$0.5(20 + 30) = 25$

(b) (5 pt) Fill in the blanks below so that the array `large_medians` contains 5,000 bootstrap samples of the median price for large hospitals.

```
t = bandages.where('Size', 'large')

large_medians = make_array()

for i in np.arange(5000):

    sample = t.sample().column('Price')

    large_medians = np.append(large_medians, np.median(sample))
```

(c) (2 pt) Fill in the code below so that the array `ci98` contains the endpoints of a 98% confidence interval for the statistic from the previous part.

```
left = percentile(1, large_medians)

right = percentile(99, large_medians)

ci98 = make_array(left, right)
```

4. (19 points) Ice cream

Your two favorite ice cream shops, Game of Cones and Sundae Night Live, start selling mystery pints. Each mystery pint has a random flavor, and the shops put up the following signs with the possible flavors and the probability of getting each one:

Game of Cones:

Flavor	Probability
Vanilla	0.3
Maple	0.15
Raspberry	0.55

Sundae Night Live:

Flavor	Probability
Vanilla	0.7
Maple	0.25
Raspberry	0.05

Answer the following questions. For each one, write your answer as a Python expression or an unsimplified expression—you do not need to evaluate it to get a number.

If there is not enough information provided to answer the question, write "not enough information".

(a) (2 pt) You buy five pints from Game of Cones. What is the probability that they are all vanilla?

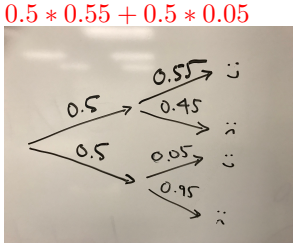
`0.3**5`

(b) (2 pt) You buy five pints from Game of Cones. What is the probability that they are all the same flavor?

`0.3**5 + 0.15**5 + 0.55**5`

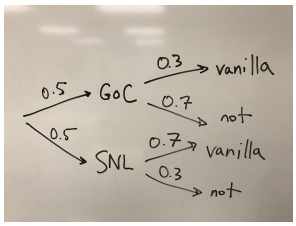
(c) (3 pt) Your roommate randomly picks one shop (each shop is equally likely to be chosen) and buys you a pint. What is the probability that it is raspberry flavored?

$0.5 * 0.55 + 0.5 * 0.05$



(d) (3 pt) Your roommate randomly picks one shop (each shop is equally likely to be chosen) and buys you a pint. If the pint was vanilla flavored, what is the probability they went to Game of Cones?

$$\frac{0.5 * 0.3}{0.5 * 0.3 + 0.5 * 0.7}$$



An ice cream researcher decides that for an entire semester, they'll buy one pint every weekday. Because of their schedule, they go to Game of Cones three days a week and Sundae Night Live two days a week.

(e) (3 pt) The researcher asks their roommates for help eating the ice cream on a randomly chosen weekday, but their roommates hate anything maple flavored. What is the probability that day's pint is **not** maple flavored?

$0.6 * 0.85 + 0.4 * 0.75$

or: $1 - (0.6 * 0.15 + 0.4 * 0.25)$

Suppose the researcher suspects that mystery pints from Game of Cones do not follow the distribution that the shop claims (in the table above). They decide to keep track of the flavors they got from visiting Game of Thrones and then conduct a hypothesis test using the following null and alternative hypotheses:

Null hypothesis: Mystery ice cream pints from Game of Cones are distributed according to the probability distribution specified by the shop.

Alternative hypothesis: Mystery ice cream pints from Game of Cones are not distributed according to the probability distribution specified by the shop.

The researcher decides to use a p-value cutoff of 0.01 for their test.

(f) (2 pt) What is an appropriate choice of test statistic for this test?

Total variation distance between the shop's distribution of flavors and the observed distribution of flavors.

(g) (2 pt) If the null hypothesis is true, what is the probability that their test will fail to reject the null hypothesis?

0.99

(h) (2 pt) The researcher conducts the hypothesis test above using what they learned in Data 8, and obtains a p-value of 0.12. Which of the following are true?

- ○ There is a 0.12 probability that the null hypothesis is true.
- ○ There is a 0.88 probability that the alternative hypothesis is true.
- ● 12% of the simulated values of the test statistic were greater than or equal to the observed value.
- ○ The researcher should reject the null hypothesis.
- ● The researcher should fail to reject the null hypothesis.
- ○ None of the above are true.

5. (19 points) Library books

You are given the following two tables, one describing 330,000 unique books from a library's catalog and one describing authors.

`books`

title	author	binding	year published	times borrowed
Little Women	Louisa M. Alcott	Paperback	1869	314
Weapons of Math Destruction	Cathy O'Neil	Hardcover	2016	27
The Design of Experiments	Ronald A. Fisher	Hardcover	1935	120
Non-Zero Probabilities	Nora K. Jemisin	Paperback	2010	65

`... (329996 rows omitted)`

`authors`

name	year born	books written
Nora K. Jemisin	1972	13
Louisa M. Alcott	1832	31
Jomny Sun	1990	2

`... (38495 rows omitted)`

For parts (a) through (c), fill in the blanks of the Python expressions to compute the described value.

- You must use *only* the lines provided: do not add any new lines.
- The code in the last line should evaluate to the value described.
- You may leave blanks empty if they don't need to be filled in.

(a) (2 pt) The average of the number of times borrowed, over all books published in the 1990s.

```
a = books.where('year published', are.between(1990, 2000))
np.mean(a.column('times borrowed'))
```

(b) (3 pt) The most popular author at this library (i.e., the one whose books, when combined, have been borrowed the highest number of times).

```
x = books.select('author', 'times borrowed')
y = x.group('author', sum)
z = y.sort('times borrowed sum', descending=True)
z.column('author').item(0)
```

(c) **(4 pt)** A table with one row per book, and two columns: the `title` of the book and the author's age at the time it was published. For example, Louisa M. Alcott was born in 1832 and Little Women was published in 1869, so she was 37 years old. The new column should be called `author age`.

```
joined = books.join('author', authors, 'name')

joined = joined.with_column(

            'author age',

            joined.column('year published') - joined.column('year born'))

joined.select('title', 'author age')
```

(d) **(2 pt)** Suppose that the average number of times borrowed for all books is 80, and the standard deviation is 20. Fill in the oval(s) next to **all** statements that must be true.

- ● At least 75% of the books have been borrowed 40 to 120 times.
- ○ About 68% of the books have been borrowed 60 to 100 times.
- ○ There is a 95% chance that the average number of times borrowed is between 40 and 120.

(e) **(3 pt)** One of the librarians asks you for help testing whether there's a difference in the number of times paperback books are borrowed compared to hardcover books. Provide a null and alternative hypothesis the librarian could use to answer their question:

Null hypothesis

The number of times borrowed for paperback books comes from the same distribution as the number of times borrowed for hardcover books.

Note: we're not giving points for saying "Any difference is due to random chance."

Alternative hypothesis

They do not come from the same distribution.

(f) **(3 pt)** Naomi doesn't have the full data we do, but she wants to estimate the percentage of books in the library that are paperback. She samples 250 books at random from the library and, using the Central Limit Theorem, comes up with a 95% confidence interval: $(39.7\%, 52.3\%)$. Fill in the oval(s) next to **all** correct statements.

- ● If she convinces 1000 library patrons to go through the same process she did (with different random books), about 95% of them will produce intervals containing the true percentage of paperback books in the library.
- ● If she samples four times as many books and repeats the process, her new confidence interval will be about half as big.
- ○ If she repeated the same procedure in a library with 1/4 as many books, her interval would be about half as big.

(g) **(2 pt)** She decides to save time by sampling 1000 books at random from a single row of shelves (instead of 250 from the whole library) and then using the same process to get a confidence interval for the percentage of books in the library that are paperback. Is this a good idea? Fill in the oval(s) next to **all** correct statements.

- ○ This is not a good idea, because you need to use the bootstrap rather than the Central Limit Theorem to obtain this confidence interval.
- ● This is not a good idea, because it is not a uniform random sample from the entire library.
- ● This is not a good idea, because it's possible that the books on that shelf might not be representative of the rest of the library.
- ○ This is not a good idea; instead, we should use a 99% confidence interval to make up for the fact that all books came from the same shelf.
- ○ This is a good idea, because the increase in sample size will be more than sufficient to outweigh any bias due to looking at only a single row of shelves.
- ○ This is a good idea, because it is a uniform random sample from all books on that row.

6. (16 points) Cal Bears Football

The Cal football team hired a new coach last season, and he was just signed on for another 5 years. We'd like to test whether his coaching has improved Cal's performance. We have two tables, each representing one season. Each row records the number of points scored by Cal and by our opponent in a single game:

`season18`

Cal	Opponent
24	17
21	18
45	23

`... (9 rows omitted)`

`season16`

Cal	Opponent
51	31
40	45
50	43

`... (9 rows omitted)`

We want to conduct a hypothesis test to determine whether Cal did better under coach Wilcox (in 2018) than under coach Dykes (in 2016).

We'll consider the *spread* for each game, which is the difference between Cal's score and our opponent's score. For example, in a game where Cal scored 40 and our opponent scored 45, the spread would be -5.

(a) **(2 pt)** Fill in the blanks in the following null and alternative hypotheses.

Null hypothesis:

The spreads from the games in 2018, under coach Wilcox, come from the same distribution as the spreads from the games in 2016, under coach Dykes, and any difference in the observed sample is due to chance.

Alternative hypothesis:

The spreads from the games in 2018, under coach Wilcox come from a distribution with a larger average spread than the spreads from the games in 2016, under coach Dykes.

(b) **(3 pt)** For our test statistic, we'll use the difference between the average spread of games coached by Wilcox and the average spread of games coached by Dykes (average Wilcox spread − average Dykes spread). Fill in the blanks to construct the `combined` table, so it looks as follows. It should have as many rows as the `season16` and `season18` tables combined. Don't worry about the order of the rows or columns.

`combined`

Coach	Spread
Wilcox	7
Wilcox	3
...	...
Dykes	20
Dykes	-5
...	...

Hint: Recall that `tbl1.append(tbl2)` returns a table with all the rows from both `tbl1` and `tbl2`, and that `tbl.with_column(colname, val)` returns a table with a new column with values from `val`.

```
coach18 = season18.with_column('Coach', "Wilcox")

coach16 = season16.with_column('Coach', "Dykes")

combined = coach16.append(coach18)

combined = combined.with_column(

    'Spread',

    combined.column('Cal') - combined.column('Opponent'))

combined = combined.drop('Cal', 'Opponent')
```

(c) (1 pt) We typically use a simulation to conduct the hypothesis test. Which hypothesis do we base our simulations on?

- ● Null hypothesis
- ○ Alternative hypothesis

(d) (5 pt) Using the table `combined`, complete the code below so that the array `diffs` will contain 10,000 simulated values of the test statistic. You might not need all the blanks, but don't add extra lines.

```
coaches = combined.column('Coach')

diffs = make_array()

for i in np.arange(10000):

    shuffled = combined.select('Spread').sample(with_replacement=False)

    shuffled = shuffled.with_column('Coach', coaches)

    spread_avgs = shuffled.group('Coach', np.average).column(1)

    spread_diff = spread_avgs.item(1) - spread_avgs.item(0)

    diffs = np.append(diffs, spread_diff)
```

(e) (2 pt) Which of the following lines of code correctly compute the p-value for this test? Assume that the name `obs_diff` has been assigned to the observed value of the test statistic (i.e., the difference between the two seasons' average spreads).

- ○ `np.count_nonzero(diff <= obs_diff) / 10000`
- ● `np.count_nonzero(diff >= obs_diff) / 10000`
- ○ `np.count_nonzero(diff <= 0.05) / 10000`

○ `np.count_nonzero(diff >= 0.05) / 10000`

(f) (3 pt) Suppose you obtain a p-value of 0.00003. If your p-value cutoff is 0.05, can you conclude that changing coaches caused Cal to do better? Why or why not?

○ Yes

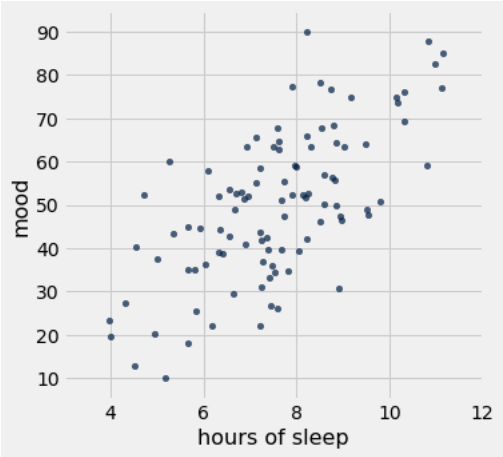
● No

Reason: This test is not enough to determine causation: there could be other confounding factors between 2016 and 2018 (harder/easier opponents, other players on the team, etc.).

7. (13 points) Mood and sleep

Your friend decides to survey 100 randomly selected Berkeley students about their sleep habits and their mood. Their mood is rated on a scale from 1 to 100, with 1 being very sad and 100 being very happy.

For parts (a) and (b), assume your friend obtains the following data:



Your friend tries to fit a line, but their kernel crashes before they can finish their analysis. They can only tell you the following pieces of information:

- The correlation coefficient r between the two variables is 0.71.
- The standard deviation of x is 1.72 hours.
- The standard deviation of y is 15.86 points.

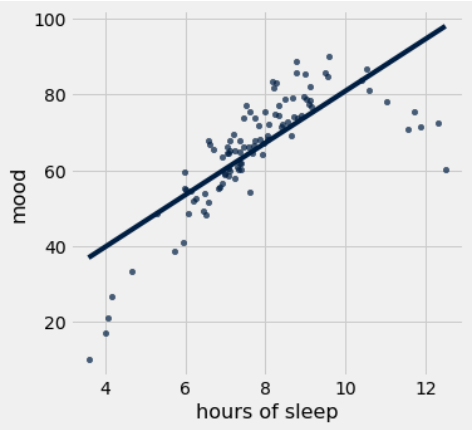
(a) (3 pt) If your friend wants to predict someone's mood from their sleeping habits, what slope would they obtain? Write your answer as a mathematical expression: you do not need to simplify. Do not use any Python code beyond basic math expressions. Do not use the `slope`, `intercept`, or `correlation` functions.

$$0.71 * \frac{15.86}{1.72}$$

(b) (3 pt) Your friend uses the bootstrap to obtain a 90% confidence interval for the correlation coefficient: $(0.62, 0.78)$. Which of the following claims are supported by their results? Mark all that are correct.

- ○ Because the students in the sample were chosen at random and the confidence interval does not contain $r = 0$, we can conclude that increased sleep causes an improvement in mood.
- ● Because the students in the sample were chosen at random and the confidence interval does not contain $r = 0$, we could reject the hypothesis that there is no association in the population between amount of sleep and mood with a p-value cutoff of 0.1.
- ○ If your friend had repeated the sampling process many times and computed a confidence interval for each sample, approximately 90% of them would produce confidence intervals narrower than the one we obtained.
- ○ This confidence interval is invalid because the bootstrap can only be used for the slope or predictions.

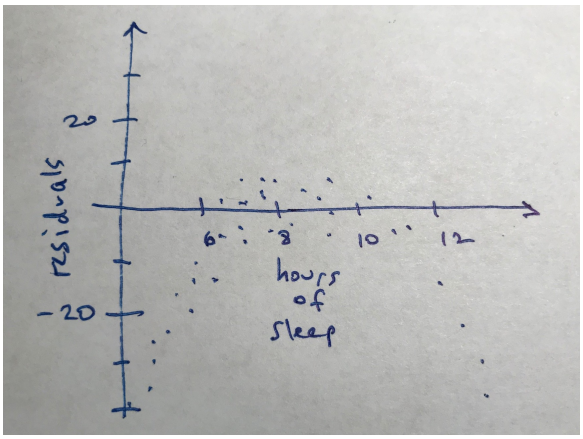
For the remainder of this question, assume instead that your friend had obtained the following data:



Your friend decides to fit a line to the data using the `slope` and `intercept` functions as defined in lecture, and gets values of 6.87 and 12.24 respectively. The resulting line is also shown on the plot above.

(c) (3 pt) Draw a rough sketch of the residual plot for your friend's line. Make sure your axes are labeled (both with axis labels and with numbers for scale).

Hint: Your plot doesn't have to include all or even most of the points, but it must have the correct shape and labels.



(d) (2 pt) Based on your sketch from (c), would linear regression be a good choice for making predictions from this data?

- ○ Yes, linear regression is always a good approach regardless of the data.
- ○ Yes, linear regression is a good model here because there is a strong positive association between mood and hours of sleep.
- ○ Yes, linear regression is a good model here because the residual plot does not show an upward trend or a downward trend.
- ● No, linear regression is not a good model here because the residual plot shows a curved pattern.
- ○ No, linear regression is not a good model here because this was not a controlled experiment and an association does not imply causation.

(e) (2 pt) Your friend uses the slope and intercept above to compute a predicted mood score of 87 for someone who gets 11 hours of sleep. When you suggest trying a quadratic function instead of a line to improve predictions for longer nights of sleep, your friend replies, "The regression line gives you the smallest possible RMSE, so there's no possible way to get a prediction with lower average error!"

Is your friend right or wrong? Select one and briefly explain.

- ○ Right
- ● Wrong

Explanation: The regression line gives the smallest possible RMSE among all linear models, but a quadratic might give even lower RMSE. Or: A quadratic might give lower RMSE than any line.

8. (0 points) Data art (optional) Draw a graph or picture describing your experience in Data 8.

9. (0 points) Write your name in the space provided on one side of every page of the exam. You're done!