INFO 2950 Fall 2022 Midterm

Do not turn the page until you are instructed to do so.

Instructions

Some students are taking the exam late due to scheduling constraints. Do not discuss the exam unless you are certain that everyone you are talking to has taken it.

You have 70 minutes to complete this exam. Time will be announced and marked on the board. You may use only a writing utensil and paper. If you use any electronic device for any purpose we will immediately confiscate your exam paper. All calculations have been constructed so that you will not need a calculator.

Write answers only in the assigned space on the answer sheet. When you are done, we will collect your answer sheet ONLY. Graders will see the segment of the answer sheet allocated for each problem and nothing else.

Make sure your name and netid are clearly written on every page of the answer sheet, as we will remove staples to scan it.

The answer sheet is intended to provide more than enough space; don't worry if you don't fill it. Showing your work may allow us to give you partial credit. Do not spend more than 10 minutes on a problem. If you get stuck, move on and come back later.

Raise your hand if you would like to ask a clarifying question.
Good luck!

$$var(X) = \frac{\sum_i (X_i - \bar{X})^2}{N}$$

$$cov(X, Y) = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

$$corr(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$ (not the same σ!)

$\sigma(t)$ is less than 0.5 when t is negative, and greater than 0.5 when t is positive.

if $Y_i = \alpha + \beta X_i + \epsilon_i$, $\beta = \frac{cov(X,Y)}{var(X)}$

A. Multiple Choice (24 points)

1. Match the following expressions to their output.
   I. float(int("2"))
   II. str(int(2.3))
   III. boolean(int("2.3"))
   IV. int(float("2.3"))

   Possible outputs:
   a. 2
   b. “2”
   c. 2.0
   d. NameError

2. If you delete a file on GitHub and someone simultaneously edits their own copy of that file, what happens after you try to merge branches?
   a. Pull request
   b. Merge conflict
   c. Commit and push
   d. Commit and pull

3. For the numpy array given below, what is the value of its `.shape` property?

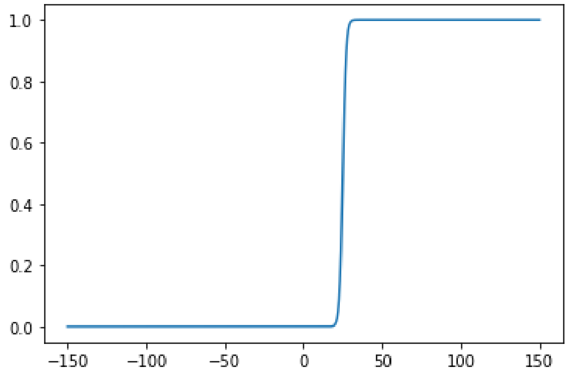
```
[[[8 0]
  [3 8]
  [9 2]
  [6 3]]
 [[3 6]
  [7 6]
  [5 4]
  [6 2]]
 [[4 9]
  [7 8]
  [2 7]
  [6 7]]]
```

   a. (4, 3, 2)
   b. (3, 4, 2)
   c. (12, 2, 1)
   d. (2, 3, 4)

4. Which of these values of a sample has the property that if you convert all the elements less than the value to **-1** and elements greater than the value to **+1**, the sum of all elements of the sample except the value will be **0**?
   a. Mean b. Median c. Mode d. Variance e. None of the above

5. If we remove outliers, which of these necessarily decreases?
    a. Mean
    b. Median
    c. Mode
    d. Variance
    e. None of the above

6. Which of the following is the correct equation for the plot given below? Note: exp(0) = e^0 = 1.0



    a. f(x) = 1 / [1 + exp(-x)]
    b. f(x) = 1 / [1 + exp(-(0.25x))]
    c. f(x) = 1 / [1 + exp(-(x - 25))]
    d. f(x) = 1 / [1 + exp(-(x + 25))]

7. Which of the following is *not* a linear regression?
    a. y ~ ax + b
    b. sqrt(y) ~ $\alpha + \beta x^2$
    c. y ~ $\beta x / (1 + \beta x)$
    d. ln(y) ~ $\alpha + \beta$ log(x)
    e. B,C,D are all *not* linear regressions
    f. None of the above (i.e., A,B,C, D are all linear regressions)

8. The matrix given below is a correlation matrix where a, b, c and d are variables:

```
          a         b         c         d
a  1.000000 -0.235053  0.656181  0.595110
b -0.235053  1.000000 -0.583851 -0.471916
c  0.656181 -0.583851  1.000000  0.871202
d  0.595110 -0.471916  0.871202  1.000000
```

Which pair of variables do you suspect is most likely to be collinear?

a. (a, a)   b. (a, b)   c. (b, c)   d. (c, d)

B. Describe What's Wrong (12 points)

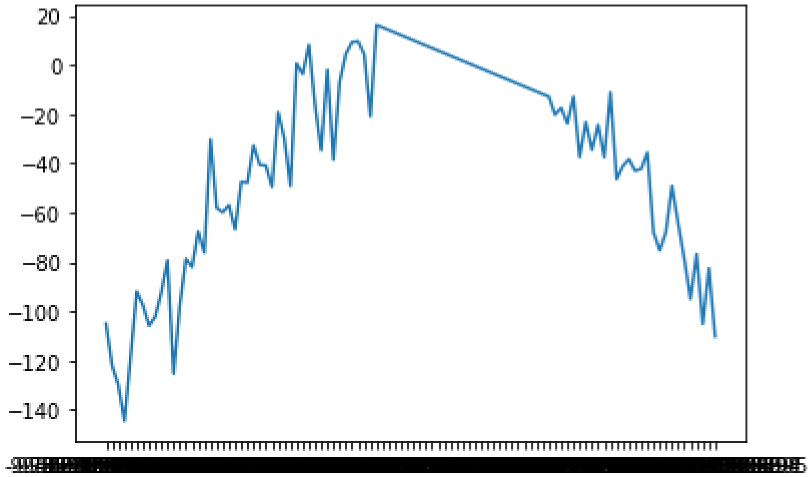
1. Your coworker asks you to analyze the following "correlation matrix". What are two reasons that make you believe this is not a correlation matrix?

```
       a     b
a    3.2  -4.7
b   -2.3   1.8
```

2. What are four things in this dataframe that make you worry?

Fruit	Is_Organic	Cost_USD	Popularity_Percent
Apple	0	1.50	80
Banana	1	1.00	Ninety
Cantaloupe	1	5.99	110
Durian	0	0.000000000001	10
Elderberry	2	7.99	60

3. There are a few things wrong with this plot. Pick one, describe what you observe, and make a guess about what might cause it. Propose an idea about how to fix the issue (assume you are correct in your diagnosis).



C. SQL (16 points)

You are given the following dataframes **df1** and **df2** regarding the INFO2950 pets, their descriptors, and their corresponding visits and costs of going to the vet.

df1

Pet	**Age**	**Species**
Basil	8	Rabbit
Sparky	2	Cat
Angora	1	Cat

df2

VetVisit	**VetCost**	**Pet**
2021-03-15	45	Basil
2021-08-01	115	Sparky
2022-03-20	55	Basil
2022-07-08	85	Sparky
2022-07-08	20	Angora
2022-07-08	0	Plant

1. Write the SQL statement that will generate the following table:

df3

Pet	**TotalCost**
Basil	100
Sparky	200
Angora	20
Plant	0

2. Write the SQL statement (without using WHERE) that will generate the following table:

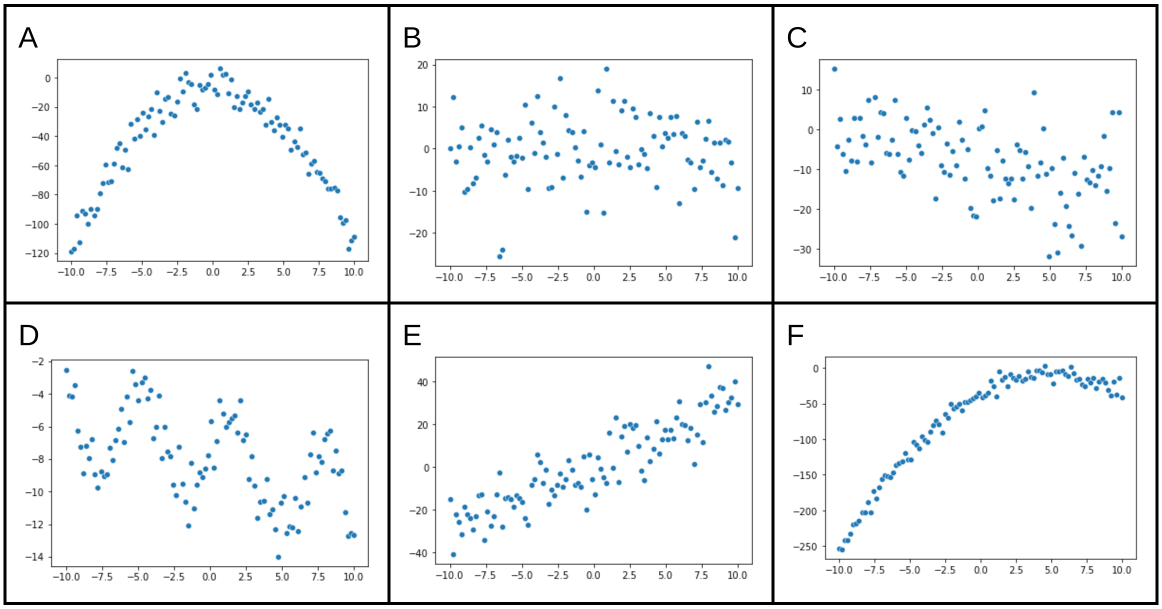
df4

Pet	**Age**	**Species**	**TotalCost**
Basil	8	Rabbit	100
Sparky	2	Cat	200
Angora	1	Cat	20

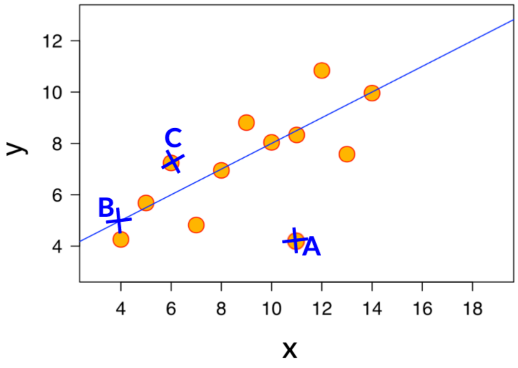
3. Write the SQL statement that is similar to **df4** but restricted only to cats, and sorts rows from lowest to highest total cost. Then, tell us the shape of your resulting dataframe.

D. Statistics and Regression (24 points)

1. For each of the following plots, describe whether the correlation between x and y values is positive, negative, or close to zero.
2.



3. The following figure depicts a scatter plot of (x, y) pairs and a regression line fitted to these data points. There are three points in the plot labeled with A, B and C respectively (in blue). Match the comments listed below (in I, II and III) with points A, B and C:

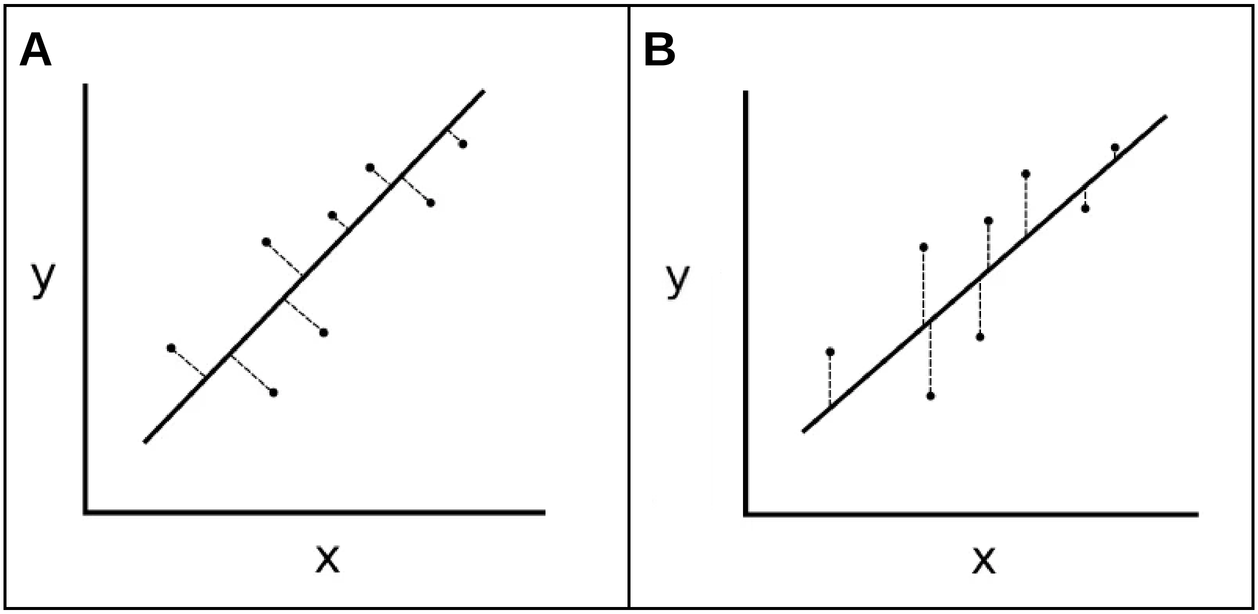


I. The data point with the largest residual ε
II. The point where α + βx = 5
III. The data point when x = 6

Image source: https://en.wikipedia.org/wiki/Linear_regression

4. The variance of *x* is 2.0, and the variance of *y* is 4.0. The slope β of the regression y ~ x is 1.5. What is the covariance of *x* and *y*? What is the slope of the regression x ~ y? Show your work.

5. Which of the two images below best represents the ordinary least squares regression we learned about in class? Explain why, and specify the mathematical concept that regression minimizes.



(Image source: https://subscription.packtpub.com/book/data/9781786462169/3/ch03lvl1sec32/implementing-deming-regression)

6. You make a residual plot based on your data.
   a. What do the x and y axes of a residual plot represent?
   b. Describe one characteristic of a residual plot that would indicate to you that your regression does not have problems.
   c. If the characteristic described in (b) is not in your residual plot, how should you improve your regression model?

7. A logistic regression model has intercept α = 2 and slope β = -3. Is the probability output greater for x = 5 or x = 1? Is the probability for x=1 greater or less than 50%? Why?

E. Interpreting regressions (24 points)

In the following questions, we provide variables for input x (or multiple input x's) and output y, as well as the regression model that describes the relationship between those variables. For each of these questions, interpret using the relevant interpretation strategies:

a. *Summarize the relationship between variables (specify units and reference variables; an exponentiated lookup table for different numbers is provided at the end of this section)*
b. *Predict y-hat (try plugging in values of x that allow you to simplify the resulting y-hat calculations; you may leave predictions as unreduced numeric expressions for this exam)*
c. *Describe outliers and oddities*

1. $y = 4.5 + 2.7x_1 - 4.2x_2$

   y = daily profit from sale of umbrellas ($ USD)
   x_1 = daily average rainfall (millimeters of rain)
   x_2 = weekend binary ($x_2 = 1$ if weekend, $x_2 = 0$ if weekday)

2. $\ln(y) = -2.05 + 1.95x_1$

   y = number of BeReal app users (Approximation from data loosely based on https://www.onlineoptimism.com/blog/bereal-stats-app-figures-data-be-real-numbers-to-know/)
   x_1 = quarters (from Q1-2020 to Q2-2022), where each quarter represents 3 months

3. $y \sim \sigma(4.71 - 0.72x_1)$

   y = whether a college student is not a senior (binary, where y = 1 if freshman/sophomore/junior, and y=0 if senior)
   x_1 = number of jobs the student applied to (numeric)

4. $y = 6.0 + 54.0x_1 + 41.5x_2 - 6.0x_1 * x_2$

   y = customer satisfaction of skincare product (numeric from 0 to 100)
   x_1 = whether Vitamin C is in the skincare product (binary)
   x_2 = whether Vitamin E is in the skincare product (binary)

Exponential lookup table (you may use as many or as few of these as needed):

n	**e^n**	**σ(n)**
-6.0	0.002	0.002
-4.2	0.01	0.01
-2.05	0.13	0.12
-0.72	0.49	0.33
1.95	7.03	0.88
2.7	14.88	0.94
4.5	90.02	0.99
4.71	111.05	0.99
6.0	403.4	1.00