Data C100/200 - Midterm 1

Fall 2025

Name: ____

Email: __________________________________@berkeley.edu

Student ID: ____

Name and SID of the person on your left: __________________________

Name and SID of the person on your right: __________________________

Exam Room:______________________ Seat Number: _______________

Instructions:

This exam consists of **37 points** spread out over **7 questions** and the **Honor Code certification**. The exam must be completed in **50 minutes** unless you have accommodations supported by a DSP letter.

- Note that some questions have circular bubbles to select a choice. Please shade in the circle fully to mark your answer.
- Blank answers and incorrect answers are graded identically, so it's in your best interest to answer every question.
- **You MUST write your Student ID number at the top of each page.**
- You should not use a calculator, scratch paper, or notes you own other than the reference sheets distributed at the beginning of the exam.

For all Python questions, you may assume `Pandas` has been imported as `pd`, `NumPy` as `np`, the Python RegEx library as `re`, `matplotlib.pyplot` as `plt`, and `seaborn` as `sns`.

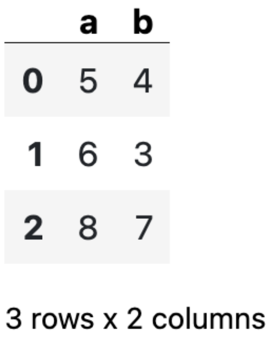
Honor Code [1 Pt]:

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I am the person whose name is on the exam, and I completed this exam in accordance with the Honor Code.

Signature: ____

This page has been intentionally left blank.

1 Pandas? More like Pabndabs [7.5 Pts]



	a	b
0	5	4
1	6	3
2	8	7

3 rows x 2 columns

The `DataFrame` above is called `df`. Consider the following code snippets. If the code snippet returns an integer between 0 and 8, **write the integer**. If it returns something other than an integer between 0 and 8 or results in an error, **write NA**.

Note: Recall that for a Series `s`, `s.iloc[0]` will return the first value of that Series.

(a) [1.5 Pt] `df['a'][1]`

Solution: 6

(b) [1.5 Pt] `df.loc['b', :][1]`

Solution: NA

(c) [1.5 Pt] `df.iloc[1, 2]`

Solution: NA

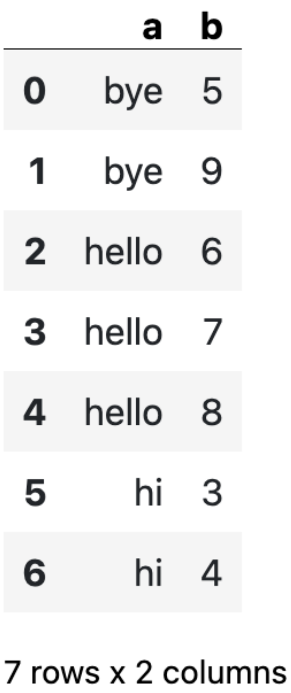
(d) [1.5 Pt] `df[df['b'] == 7]['a'].iloc[0]`

Solution: 8

(e) [1.5 Pt] `df.sort_values('a', ascending=False).head(1)['b'].iloc[0]`

Solution: 7

2 A Group of Pandas [4.5 Pts]



	a	b
0	bye	5
1	bye	9
2	hello	6
3	hello	7
4	hello	8
5	hi	3
6	hi	4

7 rows x 2 columns

The `DataFrame` above is called `df`. Each of the following code snippets returns a `DataFrame`. **Enter the number of rows** in the returned `DataFrame`. If there is not enough information provided to answer the question, write **NA**.

(a) [1.5 Pt] `df.groupby('a')[['b']].agg('max')`

Solution: 3

(b) [1.5 Pt] `df.groupby('a').head(1)`

Solution: 3

(c) [1.5 Pt] `df.groupby('a').filter(lambda sf: np.max(sf['b']) > 4)`

Solution: 5

3 Annual Dose of Regex [4 Pts]

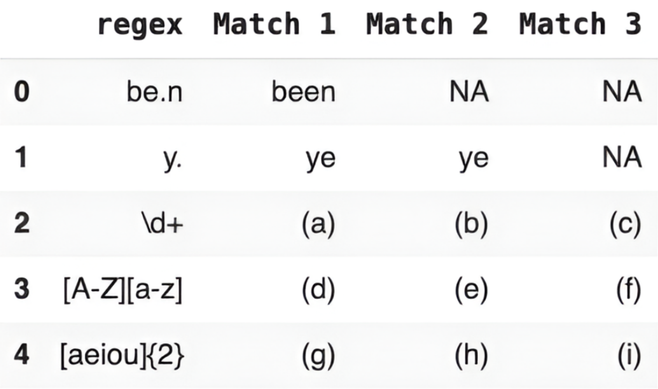
Consider the following sentence:

> *"It has been 71 years since the year 1954."*

For each of the regular expressions in the table below, write the matching substring(s) from the provided sentence. Make sure to write substring matches in the order they appear in the text.

- The first matching substring should go under **Match 1**, the second matching substring under **Match 2**, and the third matching substring under **Match 3**.
- If there are no matches, write **NA** in all three boxes. If there is only one match, write **NA** in the Match 2 and Match 3 boxes. If there are only two matches, write **NA** in the Match 3 box.

For your reference, the sentence is shown again below. We have also included an example table illustrating the format of the answer table.



	regex	Match 1	Match 2	Match 3
0	be.n	been	NA	NA
1	y.	ye	ye	NA
2	\d+	(a)	(b)	(c)
3	[A-Z][a-z]	(d)	(e)	(f)
4	[aeiou]{2}	(g)	(h)	(i)

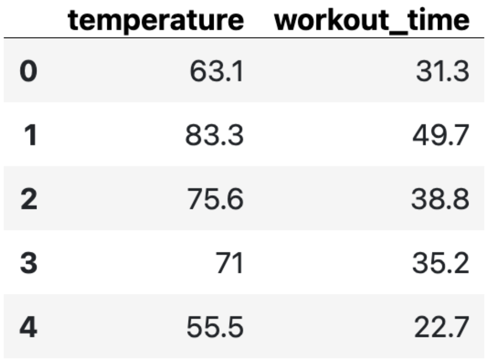
> *"It has been 71 years since the year 1954."*

regex	**Match 1**	**Match 2**	**Match 3**
`\d+`	(a) 71	(b) 1954	(c) NA
`[A-Z][a-z]`	(d) It	(e) NA	(f) NA
`[aeiou]{2}`	(g) ee	(h) ea	(i) ea

4 RSF HistoGains [8 Pts]

Rohan would like to determine whether workouts are longer on hotter or colder days. For several days, Rohan asks students at the gym to record their total workout time (in minutes). He also records the temperature outside during their workout (in °F).

Here are the first five rows of the data Rohan collected. There are 1000 rows in the full `DataFrame`:



	temperature	workout_time
0	63.1	31.3
1	83.3	49.7
2	75.6	38.8
3	71	35.2
4	55.5	22.7

(a) [4 Pts] Rohan wants to visualize the distribution of `workout_time`. Mark **True** if the plot type is appropriate for this visualization task, and **False** otherwise.

○ **True** ○ False **Histogram**
○ True ○ **False** **Scatter plot**
○ **True** ○ False **KDE Plot**
○ **True** ○ False **Boxplot**

(b) [4 Pts] Rohan wants to visualize the relationship between `temperature` and `workout_time`. Mark **True** if the plot type is appropriate for this visualization task, and **False** otherwise.

Note: Assume that the data type cannot be changed before plotting. For example, `temperature` *must be plotted as a quantitative variable. It cannot be converted into a qualitative variable.*

○ **True** ○ False **Hexplot**
○ True ○ **False** **Overlaid histograms**
○ **True** ○ False **Contourplot**
○ True ○ **False** **Side-by-side boxplots**

5 A Sampling of Errors and Biases [4 Pts]

Josh wants to estimate the proportion of UC Berkeley students who prefer cake or pie. He collects a simple random sample (SRS) of **100** emails from the official list of all UC Berkeley students. He emails each selected student and asks them to report whether they prefer cake or pie. **62** students respond to his email within one week. Josh received no additional responses after that.

Jake tells Josh that he could have improved his survey by making two updates:

- Josh should email an SRS of **200** UC Berkeley student emails, instead of **100**.
- Josh should also send a **reminder** email to all students who did not open the initial email within one week.

Assume that Josh could go back in time and make both updates suggested by Jake, and no other changes. Answer the following questions:

Note: You can assume that everyone who fills out either survey will always answer truthfully.

(a) [1 Pt] The chance error of the **updated** survey is __________ the chance error of the **original** survey.

○ Larger than

○ **Smaller than**

○ The same as

(b) [1 Pt] The response bias of the **updated** survey is likely __________ the response bias of the **original** survey.

○ Larger than

○ Smaller than

○ **The same as**

(c) [1 Pt] The non-response bias of the **updated** survey is likely __________ the non-response bias of the **original** survey.

○ Larger than

○ **Smaller than**

○ The same as

(d) [1 Pt] The selection bias of the **updated** survey is likely __________ the selection bias of the **original** survey.

○ Larger than

○ Smaller than

○ **The same as**

6 Loss Function, but Gained Knowledge [6 Pts]

(a) [2 Pts] For a constant model fit to a dataset with at least one datapoint, there is __________ exactly one parameter value that minimizes the **MSE**.

Note: You can assume the dataset is numeric with no missing values.

○ **Always**

○ Never

○ Sometimes, but not always

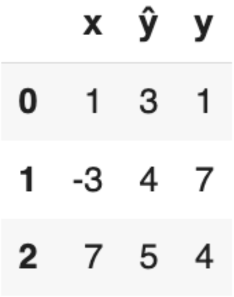
(b) [2 Pts] For a given constant model and dataset, minimizing the **sum** of loss across all data points will __________ result in the same values of the optimal parameter(s), compared to minimizing the **average** loss across all datapoints.

○ **Always**

○ Never

○ Sometimes, but not always

(c) [2 Pts] The three datapoints and predictions below are from a fitted SLR model.



	x	ŷ	y
0	1	3	1
1	-3	4	7
2	7	5	4

Compute the **MAE** of the SLR model. Show your work. Your final answer should be an integer. Draw a clear box around this integer.

Solution:

$$|y - \hat{y}| : \ |1-3| = 2, \quad |7-4| = 3, \quad |4-5| = 1$$

$$\text{MAE} \ = \ \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| = \frac{1}{3}\left(2+3+1\right) \ = \ \boxed{2}$$

7 I'm Partial to Derivatives [2 Pts]

The partial derivative of $\hat{R}(\theta) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\big(y_i - \theta_0 - \theta_1 x_i - \theta_2 x_i^2\big)^2$ with respect to θ_2 is:

- ○ $\frac{1}{n}\sum_{i=1}^{n} 2\big(y_i - \theta_0 - \theta_1 x_i - \theta_2 x_i^2\big)$
- ○ $\frac{1}{n}\sum_{i=1}^{n} 2\big(y_i - \theta_0 - \theta_1 x_i - 2\theta_2 x_i\big)$
- ○ $\frac{1}{n}\sum_{i=1}^{n} 2\big(-\theta_1 - 2\theta_2 x_i\big)$
- ○ $\frac{1}{n}\sum_{i=1}^{n} 2\big(y_i - \theta_0 - \theta_1 x_i - \theta_2 x_i^2\big)\big(-\theta_1 - 2\theta_2 x_i\big)$
- ○ **None of the above**

Solution:

$$
\begin{aligned}
\hat{R}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \theta_0 - \theta_1 x_i - \theta_2 x_i^2\big)^2 \\
\frac{\partial \hat{R}}{\partial \theta_2} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \theta_2}\Big(\big(y_i - \theta_0 - \theta_1 x_i - \theta_2 x_i^2\big)^2\Big) \\
&\overset{\text{chain}}{=} \frac{1}{n}\sum_{i=1}^{n} 2\big(y_i - \theta_0 - \theta_1 x_i - \theta_2 x_i^2\big)\ \frac{\partial}{\partial \theta_2}\big(y_i - \theta_0 - \theta_1 x_i - \theta_2 x_i^2\big) \\
&= \frac{1}{n}\sum_{i=1}^{n} 2\big(y_i - \theta_0 - \theta_1 x_i - \theta_2 x_i^2\big)\ \big(-x_i^2\big)
\end{aligned}
$$

Please state any relevant assumptions in the box below (Optional).

You are done with the midterm- Congratulations!

Draw your favorite DATA 100/200 memory so far!