# DATA 100 Fall 2020 Midterm-Exam



**INSTRUCTIONS**

This is your exam. Complete it either at exam.cs61a.org or, if that doesn't work, by emailing course staff with your solutions before the exam deadline.

This exam is intended for the student with email address `<EMAILADDRESS>`. If this is not your email address, notify course staff immediately, as each exam is different. Do not distribute this exam PDF even after the exam ends, as some students may be taking the exam in a different time zone.

For questions with **circular bubbles**, you should select exactly *one* choice.

- ◯ You must choose either this option
- ◯ Or this one, but not both!

For questions with **square checkboxes**, you may select *multiple* choices.

- ☐ You could select this choice.
- ☐ You could select this one too!

**You may start your exam now. Your exam is due at <DEADLINE> Pacific Time.** Go to the next page to begin.

**Preliminaries**

You can complete and submit these questions before the exam starts.

**(a)** What is your full name?

**(b)** What is your Berkeley email?

**(c)** What is your student ID number?

**(d)** When are you taking this exam?

- ◯ Thursday 7pm PDT
- ◯ Friday 8am PDT
- ◯ Other

**(e)** Honor Code: *All work on this exam is my own.*

By writing your full name below, you are agreeing to this code:

**(f)** Important: You must copy the following statement exactly into the box below. Failure to do so may result in points deducted on the exam.

"I certify that all work on this exam is my own. I acknowledge that collaboration of any kind is forbidden, and that I will face severe penalties if I am caught, including at minimum, harsh penalties to my grade and a letter sent to the Center for Student Conduct."

**1. (a) (8.0 points)**

The Japanese Government would like to survey its 130 million citizens as part of its census. There are 125 million people currently in Japan, and 120 million of them are Japanese citizens. Somehow, the Japanese Government is able to survey all 125 million people currently in Japan, but isn't able to survey anyone outside of Japan. (Note that this is **not** a probability sample.) Everyone that the government surveys responds. For this question, terms like Japanese, American, Indian, etc. refer to people's citizenships, and we assume there is no dual citizenship.

We begin by categorizing the people we will and will not be surveying. For this question, each letter corresponds to the following category:

A: In the sampling frame and part of the population of interest.

B: In the sampling frame but *not* part of the population of interest.

C: *Not* in the sampling frame but part of the population of interest.

D: *Not* in the sampling frame and *not* part of the population of interest.

E: Insufficient information.

**(2.0 points)**

Choose the appropriate category for each individual listed below.

**(0.5 pt)** Japanese Prime Minister Yoshihide Suga, who is currently in Japan.

**i.**
- (●) A
- ( ) B
- ( ) C
- ( ) D
- ( ) E

Someone who is Japanese is in the population of interest, and someone who is in Japan is in the sampling frame.

**B. (0.5 pt)** A Canadian tourist visiting Japan.

- ( ) A
- (●) B
- ( ) C
- ( ) D
- ( ) E

Since they are Canadian, they are not Japanese (by the assumptions stated in the question), thus they're not in the population of interest. Since they're in Japan, they're in the sampling frame.

**C. (0.5 pt)** A Japanese businesswoman who has recently returned to Japan after a 10 year trip.

- (●) A
- ( ) B
- ( ) C
- ( ) D
- ( ) E

They are Japanese and currently in Japan, thus they are in both.

**D. (0.5 pt)** The Japanese Ambassador to Peru, situated in its capital Lima.

- ( ) A
- ( ) B
- (●) C
- ( ) D
- ( ) E

They are Japanese, so they are in the population of interest. However, they are in Peru, which is not Japan, so they are not in the sampling frame.

**(2.0 pt)** Select all possible sources of bias and error in this method of surveying. (Remember our assumptions at the start.)

- [x] Selection bias
- [x] Response bias
- [ ] Non-response bias
- [ ] Random error
- [ ] None of the above

There is selection bias since our population of interest and sampling frame don't entirely overlap. There are many Japanese citizens who are not in the sampling frame, and there are many non-Japanese citizens who are in the sampling frame.

There is response bias in any survey.

There is not non-response bias, since we are told to assume that everyone who is given the survey responds.

There is not random error, because this is not a probability sample (we are sampling everyone in the sampling frame).

**iii.** **(2.0 pt)** Suppose we select a single member of the sampling frame uniformly at random. What is the probability they are not in the population of interest?

- ○ 0
- ○ 1/13
- ○ 1/24
- ● 1/25
- ○ 1
- ○ None of the above

There are 125 million people in the sampling frame. 120 million of them are Japanese (hence, in the population of interest), and 5 million of them are not. Thus, the probability of selecting someone not in the population of interest is $\frac{5}{125} = \frac{1}{25}$.

**iv.** **(2.0 pt)** Suppose we select a single member of the population of interest uniformly at random. What is the probability that they are not in the sampling frame?

- ○ 0
- ● 1/13
- ○ 2/13
- ○ 2/25
- ○ 1
- ○ None of the above

There are 130 million people in the population of interest. 120 million of them are in Japan (hence, in the sampling frame), and 10 million of them are not. Thus, the probability of selecting someone not in the sampling frame is $\frac{10}{130} = \frac{1}{13}$.

**(8.0 points)**

The Japanese Government would like to survey its 125 million citizens as part of its census. There are 120 million people currently in Japan, and 115 million of them are Japanese citizens. Somehow, the Japanese Government is able to survey all 120 million people currently in Japan, but isn't able to survey anyone outside of Japan. (Note that this is **not** a probability sample.) Everyone that the government surveys responds. For this question, terms like Japanese, American, Indian, etc. refer to people's citizenships, and we assume there is no dual citizenship.

We begin by categorizing the people we will and will not be surveying. For this question, each letter corresponds to the following category:

A: In the sampling frame and part of the population of interest.

B: In the sampling frame but *not* part of the population of interest.

C: *Not* in the sampling frame but part of the population of interest.

D: *Not* in the sampling frame and *not* part of the population of interest.

E: Insufficient information.

**(2.0 points)**

Choose the appropriate category for each individual listed below.

**(0.5 pt)** Japanese Prime Minister Yoshihide Suga, who is currently in Japan.

**(b) i.**
- (●) A
- ( ) B
- ( ) C
- ( ) D
- ( ) E

Someone who is Japanese is in the population of interest, and someone who is in Japan is in the sampling frame.

**B. (0.5 pt)** A Canadian tourist visiting Japan.

- ( ) A
- (●) B
- ( ) C
- ( ) D
- ( ) E

Since they are Canadian, they are not Japanese (by the assumptions stated in the question), thus they're not in the population of interest. Since they're in Japan, they're in the sampling frame.

**C. (0.5 pt)** A Japanese businesswoman who has recently returned to Japan after a 10 year trip.

- (●) A
- ( ) B
- ( ) C
- ( ) D
- ( ) E

They are Japanese and currently in Japan, thus they are in both.

**D. (0.5 pt)** The Japanese Ambassador to Peru, situated in its capital Lima.

- ( ) A
- ( ) B
- (●) C
- ( ) D
- ( ) E

They are Japanese, so they are in the population of interest. However, they are in Peru, which is not Japan, so they are not in the sampling frame.

**(2.0 pt)** Select all possible sources of bias and error in this method of surveying. (Remember our assumptions at the start.)

- [x] Selection bias
- [x] Response bias
- [ ] Non-response bias
- [ ] Random error
- [ ] None of the above

There is selection bias since our population of interest and sampling frame don't entirely overlap. There are many Japanese citizens who are not in the sampling frame, and there are many non-Japanese citizens who are in the sampling frame.

There is response bias in any survey.

There is not non-response bias, since we are told to assume that everyone who is given the survey responds.

There is not random error, because this is not a probability sample (we are sampling everyone in the sampling frame).

**iii.** **(2.0 pt)** Suppose we select a single member of the sampling frame uniformly at random. What is the probability they are not in the population of interest?

- ○ 0
- ● 1/24
- ○ 1/25
- ○ 23/25
- ○ 1
- ○ None of the above

There are 120 million people in the sampling frame. 115 million of them are Japanese (hence, in the population of interest), and 5 million of them are not. Thus, the probability of selecting someone not in the population of interest is $\frac{5}{120} = \frac{1}{24}$.

**iv.** **(2.0 pt)** Suppose we select a single member of the population of interest uniformly at random. What is the probability that they are not in the sampling frame?

- ○ 0
- ○ 1/13
- ○ 1/24
- ● 2/25
- ○ 1
- ○ None of the above

There are 125 million people in the population of interest. 115 million of them are in Japan (hence, in the sampling frame), and 10 million of them are not. Thus, the probability of selecting someone not in the sampling frame is $\frac{10}{125} = \frac{2}{25}$.

**2. (10.0 points)**

Suppose you are a famous country music artist. Every time you release a song, it lands on the Spotify Top 200 chart ("the charts") with a probability of $\frac{2}{3}$, independent of all other factors.

For the first four parts of this question, assume that you choose to release 12 songs in 2020.

**(a) (2.0 pt)** What is the probability that exactly half of the songs you release end up on the charts in 2020? Select all that apply.

- ☐ $\frac{1}{2}$
- ■ $\binom{12}{6}\left(\frac{2}{3}\right)^6\left(\frac{1}{3}\right)^6$
- ☐ $\binom{12}{6}\left(\frac{1}{3}\right)^6$
- ■ $1-\sum_{k=0}^{5}\binom{12}{k}\left(\frac{2}{3}\right)^k\left(\frac{1}{3}\right)^{12-k}-\sum_{k=7}^{12}\binom{12}{k}\left(\frac{2}{3}\right)^k\left(\frac{1}{3}\right)^{12-k}$
- ☐ $1-\sum_{k=0}^{5}\binom{12}{k}\left(\frac{2}{3}\right)^k\left(\frac{1}{3}\right)^{12-k}$

Let $Y$ be a random variable representing the number of songs you release in 2020 that end up on the charts. $Y$ follows the Binomial distribution with $n=12$ and $p=\frac{2}{3}$. Then in general, $P(Y=k)=\binom{12}{k}\left(\frac{2}{3}\right)^k\left(\frac{1}{3}\right)^{12-k}$ and

$$P(Y=6)=\binom{12}{6}\left(\frac{2}{3}\right)^6\left(\frac{1}{3}\right)^6$$

This gives one correct answer. The other comes from realizing $\sum_{k=0}^{12}P(Y=k)=1$, and hence

$$P(Y=6)=1-\sum_{k=0,k\neq 6}^{12}P(Y=k)=1-\sum_{k=0}^{5}P(Y=k)-\sum_{k=7}^{12}P(Y=k)$$

**(b) (2.0 pt)** What is the expected number of songs that you release in 2020 that **will not** end up on the charts? Round your answer to the closest integer.

**From above, $\mathbb{E}[Y]=np=12\cdot\frac{2}{3}=8$. This means the expected number of songs that you release in 2020 that will end up on the charts is 8, and so the expected number that will not end up on the charts in 2020 is $12-8=4$ (which is also $12\cdot\frac{1}{3}$).**

**(c) (3.0 pt)** In 2021, you somehow determine that each song you release will chart with probability $\frac{2}{7}$. How many songs should you release in 2021 such that the expected number of songs that end up on the charts in 2021 is the same as it is in 2020? Give your answer as an integer.

**From above, $\mathbb{E}[Y]=8$. Let $Z$ be the random variable representing the number of songs you release in 2021 that end up on the charts, and let $n_Z$ be the number of songs we release in 2021 total (i.e. the answer we are looking for). We need to choose $n_Z$ such that $\mathbb{E}[Y]=\mathbb{E}[Z]=\frac{2}{7}n_Z$. Solving for $n_Z$ gives $n_Z=28$.**

**(d) (1.0 pt)** Let $Y$ be the random variable representing the number of songs you release in 2020 that end up on the charts (some number between 0 and 12), and let $Z$ be the random variable representing the number of songs you release in 2021 that end up on the charts (some number between 0 and your answer to the previous part).

Which of the following is true about the relationship between $Y$ and $Z$? Select all that apply.

☐ $Y$ and $Z$ are equal

☐ $Y$ and $Z$ are identically distributed

■ None of the above

Neither of the provided options are correct. The range of values that $Z$ takes on is different than the range of values that $Y$ takes on. For example, $Z$ can be up to 28, whereas $Y$ can only be up to 12. Thus, they don't have identical distributions, and certainly are not equal.

(If you want a tangible example, $P(Z = 28) = (\frac{2}{7})^{28}$, but $P(Y = 28) = 0$.)

**(e) (2.0 pt)** Now suppose you decide to release $n$ songs in 2020, where $n$ is some even integer. Which of the following functions $f(n)$ correctly return the probability that **at least half** of the songs you release in 2020 end up on the charts?

○
$$f(n) = \binom{n}{\frac{n}{2}} \frac{2^{\frac{n}{2}}}{3^n}$$

●
$$f(n) = \sum_{k=\frac{n}{2}}^{n} \binom{n}{k} \frac{2^k}{3^n}$$

○
$$f(n) = 1 - \sum_{k=0}^{\frac{n}{2}} \binom{n}{k} (\frac{2}{3})^k (\frac{1}{3})^{n-k}$$

○
$$f(n) = \sum_{k=\frac{n}{2}+1}^{n} \binom{n}{k} (\frac{2}{3})^k (\frac{1}{3})^{n-k}$$

○ None of the above

Let $X$ be the number of songs we release in 2020. $X$ follows the Binomial distribution with parameters $n$ and $p = \frac{2}{3}$. Then,

$$P(X \geq \frac{n}{2}) = \sum_{k=\frac{n}{2}}^{n} \binom{n}{k} (\frac{2}{3})^k (\frac{1}{3})^{n-k} = \sum_{k=\frac{n}{2}}^{n} \binom{n}{k} \frac{2^k}{3^n}$$

This answer may not immediately appear to be correct, since it doesn't look like it contains a factor of the form $p^k(1-p)^{n-k}$, however it is just slightly simplified. There is something wrong with each of the other remaining options (e.g. the indices are wrong).

**3. (13.0 points)**

The table `billionaires` contains information about the ten most wealthy individuals in year1 and year2. Note that an individual can appear twice in the table if they appeared in both the year1 and year2 lists.

`billionaires` table schema:

```
CREATE TABLE billionaires (
    namereplace TEXT, /* name of individual */
    year INT, /* corresponding year */
    netWorth FLOAT, /* net worth in billions */
    company TEXT, /* primary company */
)
```

Below are 8 random rows from the `billionaires` table:

| namereplace | year | netWorth | company |
|---|---|---|---|
| Jeff Bezos | year2 | 113 | Amazon |
| Bill Gates | year2 | 98 | Microsoft |
| Bill Gates | year1 | 86 | Microsoft |
| Bernard Arnault | year2 | 76 | LVMH |
| Mark Zuckerberg | year2 | 54.7 | Facebook |
| Jeff Bezos | year1 | 72.8 | Amazon |
| Amancio Ortega | year1 | 71.3 | Zara |
| Warren Buffett | year2 | 67.5 | Berkshire Hathaway |

**(a) (4.0 points)**

**i. (1.0 pt)** Write a SQL query that returns the `namereplace`, `year`, and `netWorth` for all entries that have a `netWorth` greater than or equal to num1 and less than num2. Only show the first 5 rows.

Your query **must** follow the following structure.

```
SELECT ___________(a)____________
FROM billionaires
_____________(b)_________________
______(c)_______;
```

What goes in blank (a)?

```
namereplace, year, netWorth
```

**ii. (2.0 pt)** What goes in blank (b)?

```
WHERE billionaires.netWorth >= num1 AND billionaires.netWorth < num2
```

**iii. (1.0 pt)** What goes in blank (c)?

```
LIMIT 5
```

**(b) (9.0 points)**

**i. (2.0 pt)** Write a SQL query to find the names of all billionaires who appear in the table for both year1 and year2. Your returned table should also have a column, `avgNetWorth`, that contains the average of their year1 and year2 net worth. The billionaire(s) with the highest average net worth should be at the top of the table.

Your answer **must** follow the following structure. *Hint: You cannot use a `GROUP BY`. Instead, think about how you can calculate `avgNetWorth` in your `SELECT` statement. Also think about how joining the `billionaires` table with itself could help.*

```
SELECT ___________(a)_____________ AS avgNetWorth
FROM ____________(b)_____________
WHERE _____(c)______
AND _____________(d)_____________
ORDER BY ____(e)_____ DESC;
```

What goes in blank (a)?

```
b1.namereplace, (b1.netWorth + b2.netWorth)/2
```

**ii. (2.0 pt)** What goes in blank (b)?

```
billionaires as b1, billionaires as b2
```

**iii. (2.0 pt)** What goes in blank (c)?

```
b1.namereplace = b2.namereplace
```

**iv. (2.0 pt)** What goes in blank (d)?

```
b1.year = year1 AND b2.year = year2
```

**v. (1.0 pt)** What goes in blank (e)?

```
avgNetWorth
```

**4. (24.0 points)**

Throughout this question, we are dealing with pandas DataFrame and Series objects. All code for this question, where applicable, must be written in Python. You may assume that pandas has been imported as `pd`.

The DataFrame `eb_hikes` provided below shows different hikes in the East Bay.
The information contained in the DataFrame includes:

- Trail: trail name which is unique (string)
- Elevation Gain: total elevation gain (ft) of a trail (int)
- Length: length of trail in miles (float)
- Location: City, State where the trails are located (string)

The first four rows of `eb_hikes` are shown below.

| | Trail | Elevation Gain | Length | Location |
|---|---|---|---|---|
| **0** | Lake Temescal Loop | 95 | 1.1 | Oakland, CA |
| **1** | Inspiration Point | 452 | 4.0 | Berkeley, CA |
| **2** | Fire Trails | 1496 | 4.2 | Berkeley, CA |
| **3** | Piedmont Park Loop | 137 | 0.8 | Piedmont, CA |

**(a) (2.0 pt)** Since all the trails are located in the East Bay, we know that they are all located in California. Write a line of Pandas code that changes `eb_hikes['Location']` to only display the city in which the trail is located. (For instance, the first four elements of `eb_hikes['Location']` would be `['Oakland', 'Berkeley', 'Berkeley', 'Piedmont']`.)

```
eb_hikes['Location'] = eb_hikes['Location'].str.split(',', expand=True)[0]
```

**(b) (3.0 pt)** We like to go on hikes with large Elevation Gains, and we also like to go on hikes in cityname. Write a line of Pandas code to create a `Series` containing the names of the trails who satisfy at least one of the following conditions:

- They are located in cityname
- Their Elevation Gain is at least elevnum ft

Assign your result to the variable `varname`. You may assume that `eb_hikes['Location']` has already been modified according to the previous part.

```
varname = eb_hikes.loc[(eb_hikes['Elevation Gain'] >= elevnum) |
(eb_hikes['Location'] == 'cityname'), 'Trail']
```

**(c) (9.0 points)**

Now suppose we have another DataFrame, `user_hikes`, that shows trails people have hiked with the following information:

- User: unique user name (string)
- Trail: trail name (string)
- Rating: user rating of the trail out of 5 (float)
- Difficulty: difficulty of trail (Easy, Medium, Hard, Very Hard) (string)
- Time Taken: time taken to complete the hike in minutes (int)

Note that users may appear in `user_hikes` more than once, if they have gone on more than one hike.

| | User | Trail | Rating | Difficulty | Time Taken |
|---|---|---|---|---|---|
| **0** | Bob Honey | Piedmont Park Loop | 2 | Easy | 10 |
| **1** | Susie Thomas | Fire Trails | 4.3 | Very Hard | 255 |
| **2** | Josh Loop | Piedmont Park Loop | 4 | Easy | 20 |
| **3** | Susie Thomas | Inspiration Point | 3 | Medium | 95 |
| **4** | Bob Honey | Fire Trails | 4 | Hard | 150 |

Say we want to filter `user_hikes` to only show rows for trails with an average rating among all users of num1 or higher as well as a maximum time taken among all users of num2 minutes or lower. Fill in the blanks below to create this DataFrame. Make sure to sort the output DataFrame first by the rating and then by the time taken to hike the trail. Your result should have the same columns as `user_hikes`.

```
def best_short_hikes(df):
    short_hikes = ___(a)___
    best_hikes = ___(b)___
    return short_hikes and best_hikes

user_hikes.____(c)____.filter(best_short_hikes).sort_values(____(d)____)
```

**i. (2.0 pt)** What goes in blank (a)?

```
df['Time Taken'].max() <= num2
```

**ii.** **(2.0 pt)** What goes in blank (b)?

```
df['Rating'].mean() >= num1
```

**iii.** **(2.0 pt)** What goes in blank (c)?

```
groupby('Trail')
```

**iv.** **(2.0 pt)** What goes in blank (d)?

```
['Rating', 'Time Taken']
```

**v.** **(1.0 pt)** Why might Bob Honey's Piedmont Park Loop rating still be in the output DataFrame you specified above even though his rating is less than num1 for that hike?

- ◯ The time taken for his Piedmont Park Loop hike was less than num2 minutes so we don't care that his rating was less than num1.
- ● Even with his rating of 2, the Piedmont Park Loop had an average rating of num1 or higher.
- ◯ Bob Honey hiked the Fire Trails which had a rating of 4.
- ◯ None of the above

The individual rows for any given trail don't matter, as long as the aggregate statistics for that trail satisfy the conditions we specified.

**(d) (1.0 pt)** What type of variable is Difficulty in the `user_hikes` DataFrame?

- ○ Quantitative discrete
- ○ Quantitative continuous
- ○ Qualitative nominal
- ● Qualitative ordinal

"Very hard", for example, isn't a number, so this isn't quantitative. The different levels of Difficulty have some sense of ordering, and so this is ordinal.

**(e) (7.0 points)**

As a reminder, here are the DataFrames `eb_hikes` and `user_hikes`, respectively:

`eb_hikes`

| | Trail | Elevation Gain | Length | Location |
|---|---|---|---|---|
| **0** | Lake Temescal Loop | 95 | 1.1 | Oakland, CA |
| **1** | Inspiration Point | 452 | 4.0 | Berkeley, CA |
| **2** | Fire Trails | 1496 | 4.2 | Berkeley, CA |
| **3** | Piedmont Park Loop | 137 | 0.8 | Piedmont, CA |

`user_hikes`

| | User | Trail | Rating | Difficulty | Time Taken |
|---|---|---|---|---|---|
| **0** | Bob Honey | Piedmont Park Loop | 2 | Easy | 10 |
| **1** | Susie Thomas | Fire Trails | 4.3 | Very Hard | 255 |
| **2** | Josh Loop | Piedmont Park Loop | 4 | Easy | 20 |
| **3** | Susie Thomas | Inspiration Point | 3 | Medium | 95 |
| **4** | Bob Honey | Fire Trails | 4 | Hard | 150 |

The DataFrame below shows the average minutes per mile that it takes each user to complete a trail.

| Trail | Fire Trails | Inspiration Point | Piedmont Park Loop |
|---|---|---|---|
| **User** | | | |
| **Bob Honey** | 35.714286 | NaN | 12.5 |
| **Josh Loop** | NaN | NaN | 25.0 |
| **Susie Thomas** | 60.714286 | 23.75 | NaN |

```
temp = user_hikes.___(a)___(eb_hikes, ______b______)
temp['Mins per Mile'] = ______(c)______
df = temp.____________________(d)___________________________
```

Fill in the blanks such that we output the DataFrame `df` above.

**i.** **(1.0 pt)** What goes in blank (a)?

```
merge
```

**ii.** **(1.0 pt)** What goes in blank (b)?

```
on='Trail'
```

**iii.** **(2.0 pt)** What goes in blank (c)?

```
temp['Time Taken'] / temp['Length']
```

**iv.** **(3.0 pt)** What goes in blank (d)?

```
pivot_table(index='User', columns='Trail', values='Mins per Miles',
aggfunc=np.mean)
```

**(f) (2.0 pt)** Consider the DataFrame `df` you created in the previous part. Suppose someone told you that User "Michael James" was the fastest recorded individual, as measured by average minutes per mile, to finish the Piedmont Park Loop (with no ties). Which of the following is guaranteed to correctly determine Michael James' average minutes per mile for the Inspiration Point trail? Select all that apply.

- [x] `df.loc[df['Piedmont Park Loop'] == df['Piedmont Park Loop'].min(), 'Inspiration Point']`
- [ ] `df.loc[~(df['Piedmont Park Loop'] <= df['Piedmont Park Loop'].max()), 'Inspiration Point']`
- [x] `df.sort_values('Piedmont Park Loop').loc[:, 'Inspiration Point'].iloc[0]`
- [ ] `df.sort_values('Piedmont Park Loop').loc[:, 'Inspiration Point'].loc[0]`
- [ ] None of the above

`df.loc[df['Piedmont Park Loop'] == df['Piedmont Park Loop'].min(), 'Inspiration Point']` correctly computes the result.

`df.loc[~(df['Piedmont Park Loop'] <= df['Piedmont Park Loop'].max()), 'Inspiration Point']` is wrong, and is a distractor. It will return an empty Series, since "not less than or equal to the max" equates to "greater than the max", which is nonsensical.

`df.sort_values('Piedmont Park Loop').loc[:, 'Inspiration Point'].iloc[0]` also correctly computes the result.

`df.sort_values('Piedmont Park Loop').loc[:, 'Inspiration Point'].loc[0]` is also wrong, because `.loc` requires us to pass in an index label, and 0 is not in our index labels'.

**5. (8.0 points)**

HTTP is a response-reply internet protocol and is used by web-browsers to request content from servers to display to the user. The first few lines of an HTTP request have the following format:

```
POST /fa20/syllabus HTTP/1.1
User-Agent: Mozilla/4.0 (compatible; MSIE5.01; Windows NT)
Host: ds100.org
```

The first line contains an HTTP "verb", usually `GET` or `POST`, the path on the host being requested (`/fa20/syllabus` above), and the HTTP version being used to send the request. The lines below that are HTTP request headers that define the required parameters so that the server can process the request. (There are typically more, but we've omitted them since they're not relevant to this problem.)

**(a) (4.0 pt)** In the string `extract_verb`, write a regular expression that extracts the HTTP verb from a request.

For example:

```
>>> request_1 = """POST /fa20/syllabus HTTP/1.1
User-Agent: Mozilla/4.0 (compatible; MSIE5.01; Windows NT)
Host: ds100.org"""
>>> re.findall(extract_verb, request_1)[0]
'POST'

>>> request_2 = """GET /su19/syllabus HTTP/1.1
User-Agent: Safari/13.1 (Macintosh; Intel Mac OS X 10_10)
Host: data8.org"""
>>> re.findall(extract_verb, request_2)[0]
'GET'

>>> request_3 = """GARBAGE /useless HTTP/1.1
User-Agent: Garbage/0.0 (garbage)
Host: garbage.ca"""
>>> re.findall(extract_verb, request_3)[0]
'GARBAGE'
```

Your regex should work on requests that follow the format of the examples above. Assume that all HTTP requests have "HTTP" in them. Please write the regex as you would in Python with the form `extract_verb = r"..."`.

- Hint: Use capturing groups.
- Hint: Part of your regex may be the string `"\sHTTP/[\d.]+"`.

```
extract_verb = r"([A-Z]+)\s\/.*\sHTTP\/[\d.]+"
```

**(b) (4.0 pt)** In the string `extract_browser`, write a regular expression that extracts the name of the web browser (without its version) from a request.

For example:

```
>>> request_1 = """POST /fa20/syllabus HTTP/1.1
User-Agent: Mozilla/4.0 (compatible; MSIE5.01; Windows NT)
Host: ds100.org"""
>>> re.findall(extract_browser, request_1)[0]
'Mozilla'

>>> request_2 = """GET /su19/syllabus HTTP/1.1
User-Agent: Safari/13.1 (Macintosh; Intel Mac OS X 10_10)
Host: data8.org"""
>>> re.findall(extract_browser, request_2)[0]
'Safari'

>>> request_3 = """GARBAGE /useless HTTP/1.1
User-Agent: Garbage/0.0 (garbage)
Host: garbage.ca"""
>>> re.findall(extract_browser, request_3)[0]
'Garbage'
```

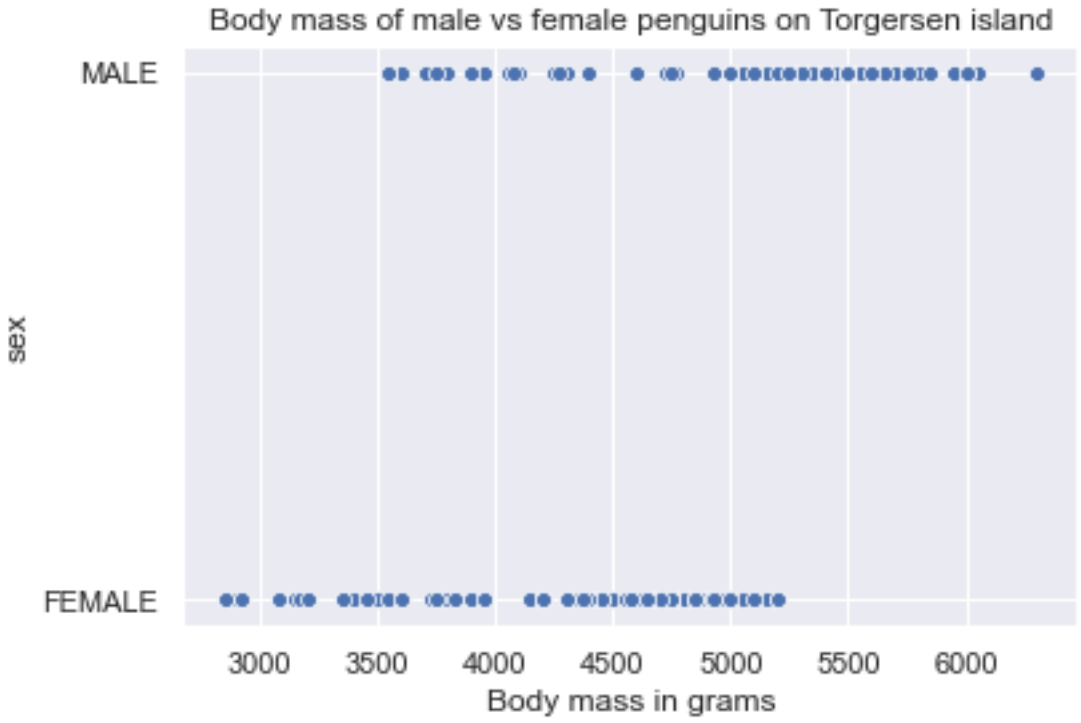
Again, please write the regex as you would in Python with the form `extract_browser = r"..."`.

```
extract_browser = r"User-Agent: (\w+)\/.*"
```

**6. (6.0 points)**

Answer each of the following questions regarding visualizations of a dataset collected about penguin species in the Palmer Archipelago in Antarctica.

Plot of body mass of male vs female penguins:




**(a) (2.0 pt)** The above visualization suffers from overplotting. Which of the following would be more appropriate types of visualization for this data? Select all that apply.

- ☐ Side-by-side line plots
- ■ Overlaid density curves
- ■ Side-by-side boxplots
- ☐ Bar chart

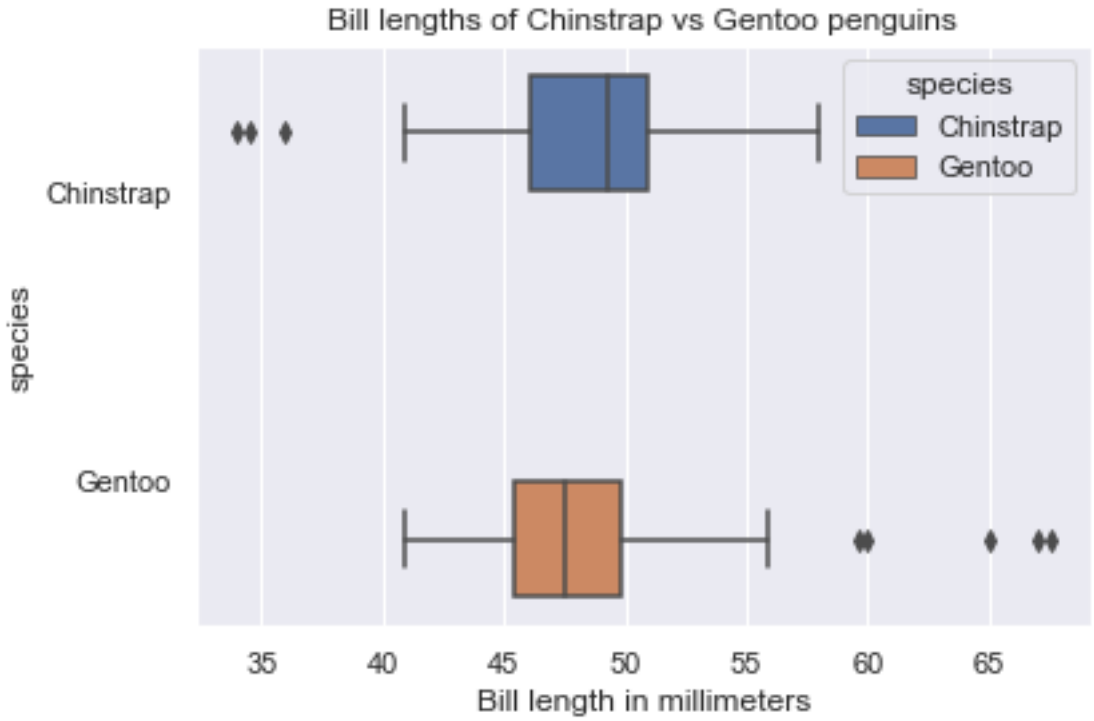
There is no sense of time in this data, so a line plot would not be appropriate. Instead, we want to see the distribution of the masses of both male and female penguins. To visualize a distribution we use density curves and boxplots, which is why overlaid density curves and side-by-side boxplots are appropriate. Bar charts would not be appropriate as they would only show one number for male penguins and one number for female penguins, which is not the entire distribution. (Bar charts are not to be confused with histograms.)

**(b) (2.0 pt)** Suppose instead we wanted to visualize bill length vs. body mass for all female penguins on the Archipelago. Which of the following would be appropriate types of visualization for this data? Select all that apply.

- ☐ Histograms
- ■ 2D density curves
- ☐ Side-by-side boxplots
- ■ Scatter plots

We now want to visualize the relationship between two continuous variables. 2D density curves and scatter plots are appropriate for this task.

**(c) (2.0 pt)** Plot of bill lengths of Chinstrap vs Gentoo penguins:




We are now using a subset of the data, and visualizing the bill lengths of two penguin species using side-by-side box plots.

Which of these numbers is closest to the inter-quartile range of the Gentoo penguin bill lengths visualized above, in millimeters?

- ○ 2.5
- ● 5
- ○ 10
- ○ 15

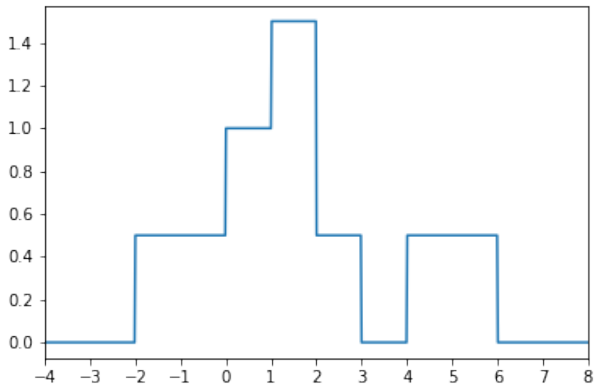
The inter-quartile range is the length of the box in a boxplot. The Gentoo box spans from roughly 45 mm to 50 mm, so the IQR is roughly 50 mm - 45 mm = 5 mm.

**7. (5.0 points)**

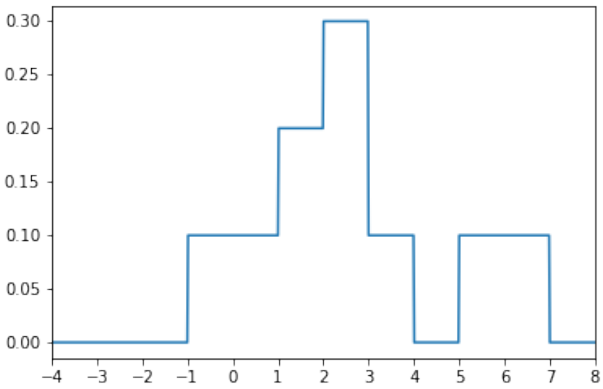
**(a) (3.0 pt)** For this question, suppose we have a dataset $X = [x_1, x_2, x_3, x_4, x_5] = [1, 1, -1, 5, 2]$.

Which of the following is the estimated density of $X$ with the Boxcar kernel and bandwidth parameter $\alpha = 2$? Recall that the boxcar kernel is

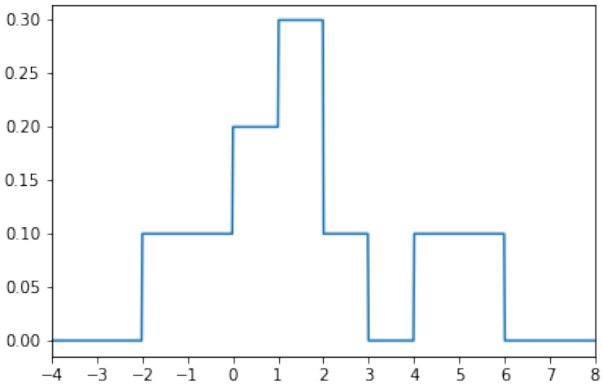
$$K_\alpha(x, x_i) = \begin{cases} \frac{1}{\alpha} & |x - x_i| \leq \frac{\alpha}{2} \\ 0 & \text{otherwise} \end{cases}$$



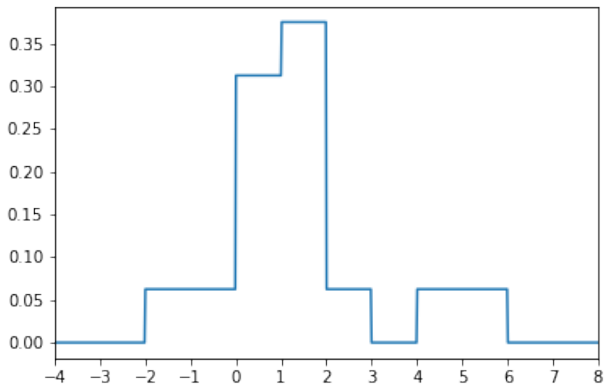

○




○




●




○

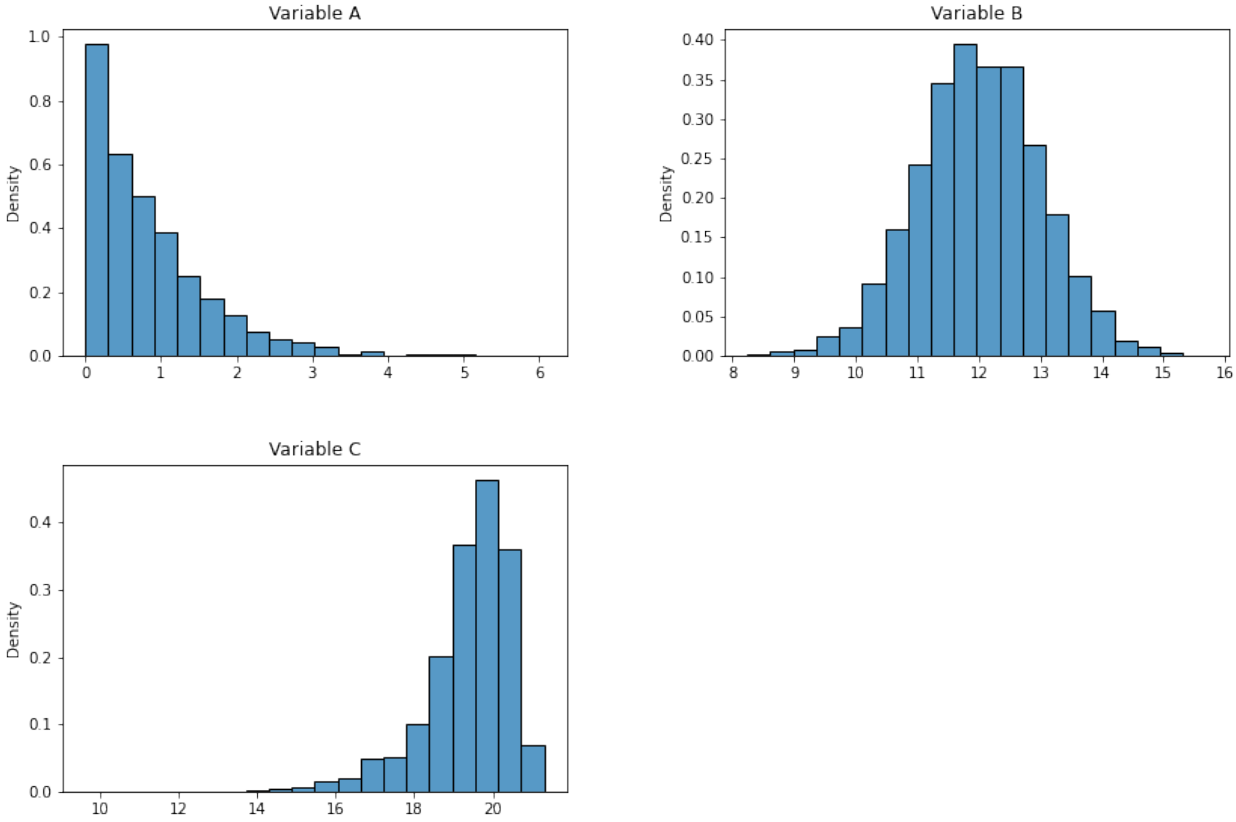
The first option has an incorrect y-axis. The area under this KDE is much larger than 1, so it is invalid.

The second option has an incorrect x-axis. We know that there is one observation at 5, meaning there should be a box of width 2 centered at $x = 5$. However, the box on the far right is centered at $x = 6$.

The third option is correct.

The fourth option's shape is incorrect. Since there are 2 observations at $x = 1$, and also an observation at $x = 2$, the "height" of the box at $x = 1$ should be three times the height of the box at $x = 5$ (this is because the bandwidth parameter is 2, and so at a given point on the x-axis, any observations that are within one unit on either side will contribute to the height). Notice this is true in the correct KDE, but not in this one.

**(b) (2.0 pt)** Below, we show the distributions of three variables.




Which of the above distributions would be made more symmetric by applying a log transformation to the x-axis? Select all that apply.

- [x] Variable A
- [ ] Variable B
- [ ] Variable C
- [ ] None of the above

Data that is right-skewed can be made more symmetric with a log transformation on the x-axis. Only Variable A's distribution is right-skewed. Applying a log transformation to either of the other variables would make their distributions more left-skewed.

**8. (14.0 points)**

Suppose we want to fit a constant model, $\hat{y} = \theta$, to a dataset $\{y_1, y_2, ..., y_n\}$.

**(a) (3.0 pt)** For the first two parts of this question, suppose we use the following loss function:

We choose to use the following loss function:

$$L(y_i, \theta) = -\ln\left(\frac{1}{\theta}\exp\left(-\frac{|y_i|}{\theta}\right)\right)$$

What is the derivative of the average loss $R$ (i.e. empirical risk) for this choice of model and loss function with respect to $\theta$?

- Hint: $\exp(x) = e^x$, and $\ln(x) = \log_e(x)$.
- Hint: $\ln(a\exp(b)) = \ln(a) + b$.

○ $\frac{dR}{d\theta} = 0$

● $\frac{dR}{d\theta} = \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{\theta} - \frac{|y_i|}{\theta^2})$

○ $\frac{dR}{d\theta} = \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{\theta} + \frac{|y_i|}{\theta^2})$

○ $\frac{dR}{d\theta} = -\frac{1}{n}\sum |y_i| \exp(-\frac{|y_i|}{\theta})$

○ $\frac{dR}{d\theta} = \frac{1}{n}\sum \frac{|y_i|}{\theta} \exp(-\frac{|y_i|}{\theta})$

First, we use the hint to re-write $L(y_i, \theta)$. Then, we take the derivative of it with respect to $\theta$ to find the derivative of the loss for a single observation.

$$\begin{aligned} L(y_i, \theta) &= -\ln\left(\frac{1}{\theta}\right) + \frac{|y_i|}{\theta} \\ \frac{d}{d\theta}L(y_i, \theta) &= -\frac{1}{\frac{1}{\theta}} \cdot \frac{-1}{\theta^2} - \frac{|y_i|}{\theta^2} \\ &= \frac{1}{\theta} - \frac{|y_i|}{\theta^2} \end{aligned}$$

Then, since $R = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \theta)$, we have

$$\frac{dR}{d\theta} = \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{\theta} - \frac{|y_i|}{\theta^2})$$

**(b) (3.0 pt)** Determine the value of $\hat{\theta}$ that minimizes average loss for the above choice of model and loss function.

- ○ $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- ○ $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$
- ● $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} |y_i|$
- ○ $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} e^{|y_i|}$
- ○ $\hat{\theta} = \frac{1}{n}\frac{1}{\sum_{i=1}^{n} |y_i|}$
- ○ $\hat{\theta} = \frac{1}{n}\frac{1}{\sum_{i=1}^{n} y_i}$

Setting the above quantity equal to 0 gives

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\theta} - \frac{|y_i|}{\theta^2}\right) = 0$$

$$\sum_{i=1}^{n}\left(\frac{1}{\theta} - \frac{|y_i|}{\theta^2}\right) = 0$$

$$\frac{n}{\theta} - \frac{1}{\theta^2}\sum_{i=1}^{n}|y_i| = 0$$

$$n\theta - \sum_{i=1}^{n}|y_i| = 0$$

$$\implies \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}|y_i|$$

as desired.

**(c) (2.0 pt)** For the remaining parts of this question, consider the following two loss functions.

$$L_A(y_i, \theta) = 2(y_i - \theta)^2$$

$$L_B(y_i, \theta) = 3|y_i - \theta|$$

Let $\hat{\theta_A}$ and $\hat{\theta_B}$ be the values of $\theta$ that minimize average $L_A$ loss and average $L_B$ loss, respectively, for the constant model and a given dataset.

Which of the following is true regarding the relationship between $\hat{\theta_A}$ and $\hat{\theta_B}$ for the dataset $y = [1, 2, 3, 4, 5]$?

- ○ $\hat{\theta_A} > \hat{\theta_B}$
- ● $\hat{\theta_A} = \hat{\theta_B}$
- ○ $\hat{\theta_A} < \hat{\theta_B}$
- ○ Impossible to tell

Note, $L_A$ is essentially squared loss, and $L_B$ is essentially absolute loss. Scaling the entire loss by a constant does not change the optimizer at all; the value of $\theta$ that minimizes $\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$ is the same value that minimizes $\frac{1}{n}\sum_{i=1}^{n} c(y_i - \theta)^2 = c\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$ for any positive constant $c$.

As covered in class, the value of $\theta$ that minimizes average squared loss for the constant model is the mean, and the value of $\theta$ that minimizes average absolute loss for the constant model is the median.

For the set of values provided, the mean and median are both 3, and hence $\hat{\theta_A} = \hat{\theta_B}$.

**(d) (2.0 pt)** Select the correct pair of optimal $\theta$ values for the dataset $y = [2, 2, 2, 2, 3, 3]$.

- ○ $\hat{\theta_A} = \frac{14}{3}, \hat{\theta_B} = 6$
- ○ $\hat{\theta_A} = \frac{7}{6}, \hat{\theta_B} = 2$
- ○ $\hat{\theta_A} = \frac{7}{6}, \hat{\theta_B} = \frac{2}{3}$
- ● $\hat{\theta_A} = \frac{7}{3}, \hat{\theta_B} = 2$
- ○ None of the above

The mean of these values is $\frac{14}{6} = \frac{7}{3}$, and the median is 2.

**(e) (2.0 pt)** Consider the dataset $y = [1, 2, 3, 4, c]$, where $c$ is some constant greater than 4. As $c \to \infty$, what happens to $\hat{\theta_A}$ and $\hat{\theta_B}$?

- ○ $\hat{\theta_A}$ and $\hat{\theta_B}$ both increase
- ● $\hat{\theta_A}$ increases, but $\hat{\theta_B}$ remains constant
- ○ $\hat{\theta_A}$ remains constant, but $\hat{\theta_B}$ increases
- ○ Both $\hat{\theta_A}$ and $\hat{\theta_B}$ remain constant

As we increase $c$, the mean of $[1, 2, 3, 4, c]$ will increase. The median, however, will always be 3.

**(f) (2.0 pt)** Now consider another new loss function, $L_C(y_i, \theta) = (y_i - 2\theta)^2$.

True or False: For every dataset consisting of positive real numbers $y = [y_1, y_2, ..., y_n]$, the value of $\theta$ that minimizes average $L_A$ loss is equal to the value of $\theta$ that minimizes average $L_C$ loss (in other words, $\hat{\theta_A} = \hat{\theta_C}$).

◯ True

● False

False. A little bit of algebra shows that the value of $\theta$ that minimizes average $L_C$ loss is $\frac{1}{2}\bar{y}$. This is not the same as scaling the entire loss by some constant, as we did in earlier parts of this question.

**9. (12.0 points)**

**(a) (2.0 pt)** Consider the simple linear regression model $\hat{y} = \theta_0 + \theta_1 x$.

Which of the following expressions evaluate to $\hat{\theta_1}$, the value of $\theta_1$ that minimizes average squared loss for the simple linear regression model? (Note, $r$ is the correlation coefficient. You can assume $\hat{\theta_0}$ is already defined, and that $\bar{x}, \bar{y}, \sigma_x, \sigma_y \neq 0$.)

○ $\hat{\theta_1} = \frac{r}{\sigma_x^2}$

○ $\hat{\theta_1} = \bar{y} - \hat{\theta_0}\bar{x}$

● $\hat{\theta_1} = \frac{\bar{y} - \hat{\theta_0}}{\bar{x}}$

○ $\hat{\theta_1} = \frac{1}{n\sigma_x\sigma_y} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

Many of these options are intentionally tricky!

In class, we saw that $\hat{\theta_0} = \bar{y} - \hat{\theta_1}\bar{x}$. Rearranging for $\hat{\theta_1}$ yields $\hat{\theta_1} = \frac{\bar{y} - \hat{\theta_0}}{\bar{x}}$ as required.

**(b) (3.0 pt)** Now consider two models:

- Model A: $\hat{y} = \theta_0 + \theta_1 x$
- Model B: $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

We fit both models using the same dataset, and we fit both models using all data available to us. The $x_1$ feature in the second model is the same as the $x$ feature in the first model. As usual, we determine optimal coefficients by minimizing average squared loss.

Let $\text{RMSE}_A$ and $\text{RMSE}_B$ represent the root mean squared error on our dataset for Model A and Model B, respectively. Furthermore, let $R_A^2$ and $R_B^2$ represent the Multiple $R^2$ coefficient values for Model A and Model B, respectively. Lastly, let $\hat{\theta_{1,A}}$ and $\hat{\theta_{1,B}}$ represent the optimal values of $\hat{\theta_1}$ for Model A and Model B, respectively.

Which of the following statements are guaranteed to be true? Select all that apply.

■ $\text{RMSE}_A \geq \text{RMSE}_B$

☐ $\text{RMSE}_A \leq \text{RMSE}_B$

☐ $R_A^2 \geq R_B^2$

■ $R_A^2 \leq R_B^2$

☐ $\hat{\theta_{1,A}} = \hat{\theta_{1,B}}$

☐ $\hat{\theta_{1,A}} \neq \hat{\theta_{1,B}}$

☐ None of the above

As we add features, RMSE on our training data either stays the same or goes down, it cannot go up. Similarly, as we add features, our $R^2$ either stays the same or goes up, it cannot go down. In general, there is no direct relationship between the coefficient on a particular feature in two different models. They may or may not be different.

**(c) (3.0 pt)** For the remainder of this question, we will use the multiple linear regression model, which is of the form

$$\hat{y} = x \cdot \theta = \sum_{j=0}^{p} \theta_j x_j$$

As in class, assume:

- $\mathbb{Y}$ is a vector containing our observed responses (i.e. true $y$'s).
- $\mathbb{X}$ is a design matrix whose first column is 1 (i.e. $x_0 = 1$ for all observations), and $\mathbb{X}_i$ represents the $i$th row of $\mathbb{X}$.
- We determine $\hat{\theta}$ by minimizing average squared loss.

$\hat{\theta}$ is the minimizer of which of the following quantities? Select all that apply.

- ☐ $\sum_{i=1}^{n}(y_i - \theta)^2$
- ■ $\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{X}_i \cdot \theta)^2$
- ■ $\sum_{i=1}^{n}(y_i - \mathbb{X}_i \cdot \theta)^2$
- ■ $\frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$
- ■ $||\mathbb{Y} - \mathbb{X}\theta||_2^2$
- ☐ $\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{X}_i \cdot \theta)$
- ☐ $\sum_{i=1}^{n}(y_i - \mathbb{X}_i \cdot \theta)$
- ☐ None of the above

The correct answers are all equivalent to either average squared loss or total squared loss (which is average squared loss without the $\frac{1}{n}$).

**(d) (2.0 pt)** Suppose we have $n = 4$ observations. What are possible valid values for the residual vector $e = \mathbb{Y} - \mathbb{X}\hat{\theta}$ given our model and $\hat{\theta}$? Select all that apply.

- ■ $e = [-1, 2, 3, -4]^T$
- ☐ $e = [-1, 2, 3, 4]^T$
- ■ $e = [1, -1, 1, -1]^T$
- ☐ $e = [0, 0.1, 0, 0.1]$
- ☐ None of the above

Since our model has an intercept term as stated above, we know that the residuals must sum to 0.

**(e) (2.0 pt)** For this part, assume we have a design matrix and true response vector defined as follows:

$$\mathbb{X} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & -1 & 4 \\ 1 & 2 & 0 \end{bmatrix}, \mathbb{Y} = \begin{bmatrix} 3 \\ 4 \\ -1 \end{bmatrix}$$

Suppose $\hat{\theta} = [-1, 4, 1]^T$. What is the squared loss for the second row in our dataset? Give your answer as an integer.

$$\hat{y_2} = \mathbb{X}_2^T \hat{\theta} = [1, -1, 4]^T [-1, 4, 1] = -1$$
$$L_2 = (y_2 - \hat{y_2})^2 = (4 - (-1))^2 = 25$$

**No more questions.**