# Data 100, Midterm

Summer 2021

Name: _____

Email: _____@berkeley.edu

Student ID: _____

Exam Time: _____

*All work on this exam is my own* ***(please sign)***: _____

# Honor Code [1 Pt]

1. *As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I will not communicate with any other individual during the exam, current student or otherwise. All work on this exam is my own.*
   (a) Please confirm your agreement with the above statement by writing your name in the space below.

   _____

# This is Prob[ability?] Sampling [13 Pts]

2. A recent study by Women In Data Science (WiDS) estimates that 80% of the world's data scientists identify as male, 15% identify as female, and 5% who do not identify as either. Terrible, right! Here at Data 100, we want to see how our course's gender distribution compares to that of the study conducted by WiDS.

   Per Data 100 enrollment statistics, our class has 400 students. During our first live lecture, we asked all 250 students who attended whether they identified as female. We received a 97% response rate, and found that 43% of respondents were female, and 57% were not female.

   (a) [1 Pt] What is the population of interest?

   ◯ All data scientists in the world
   ◯ All data scientists sampled in the WiDS study
   ◯ **All students in Data 100**
   ◯ All students who attended Lecture 1

   (b) [1 Pt] What is the sampling frame?

   ◯ All data scientists in the world
   ◯ All data scientists sampled in the WiDS study
   ◯ All students in Data 100
   ◯ **All students who attended Lecture 1**

   (c) [2 Pts] Suppose we believe this sample is representative of the population of students taking Data 100. In other words, we expect around 43% of all students in Data 100 to be female. Select one of the following reasons why our assumption may, or may not, hold true. Assume any individual female is just as likely to attend live lecture as any given non-female.

   ◯ It will not hold true. We may have selection bias, as the people who attend live lectures are self-selecting.
   ◯ It will not hold true. We have significant sampling bias due to our sample being different from our population.
   ◯ **It will hold true. Our sample size is large enough, so there is likely little variability from our sample estimates in the population.**
   ◯ It will hold true. Our response rate among our sample was high, so the response rate among the Data 100 population must be similar.

   (d) [1 Pt] What type of sampling technique are we using in the problem above? Choose one.

   ◯ Replacement sampling
   ◯ Simple random sampling
   ◯ Quota sampling
   ◯ **Convenience sampling**

(e) [1 Pt] Say we now want to draw a sample that mimics the study conducted by WiDS - that is, we want our sample to include 80% males, 15% females, 5% non-binary identifying individuals. In this case, what sampling technique are we using?

- ◯ Replacement sampling
- ◯ Simple random sampling
- ◯ **Quota sampling**
- ◯ Convenience sampling

3. Suppose we know that our class consists of 50% males, 45% females, and 5% who do not identify as either male or female. We now draw a random sample **with replacement** of 45 individuals from our course. Let $M$, $F$, $O$ denote random variables for the number of male, female, and neither female nor male identifying students in this sample.

(a) [1 Pt] Which of the following distributions does the random variable $O$ most closely follow? Select one.

- ◯ Binomial(350, 0.05)
- ◯ Binomial(350, 0.95)
- ◯ **Binomial(45, 0.05)**
- ◯ Binomial(45, 0.95)

(b) [2 Pts] What is the probability of selecting exactly 25 non-female students in our sample? Select all that apply.

- ☐ $\binom{45}{25}(.55)^{20}(.45)^{25}$
- ☐ $\binom{45}{25}(.55)^{25}(.45)^{20}$
- ☐ $\binom{45}{20}(.45)^{20}(.55)^{25}$
- ☐ $\binom{45}{20}(.55)^{20}(.45)^{25}$

(c) [2 Pts] What is the probability of selecting at least 25 males in our sample? Select all that apply.

- ☐ $\sum_{k=25}^{45}\binom{45}{k}(.5)^{k}(.5)^{45-k}$
- ☐ $1-\sum_{k=25}^{45}\binom{45}{k}(.5)^{k}(.5)^{45-k}$
- ☐ $1-\binom{45}{25}(.55)^{25}(.45)^{20}$
- ☐ $\sum_{k=0}^{20}\binom{45}{k}(.5)^{k}(.5)^{45-k}$

(d) [2 Pts] Suppose we draw our sample and obtain 20 males, 20 females, and 5 individuals who identify as neither male nor female. From this sample, we want to draw 3 students **without replacement**. What is the probability we do not draw any students who identify as neither male nor female? Choose one.

- ◯ $(\frac{40}{45})^3$
- ◯ $(\frac{40}{45})(\frac{39}{45})(\frac{38}{45})$
- ◯ $(\frac{40}{45})(\frac{39}{44})(\frac{38}{43})$
- ◯ None of the above.

# A Giant Panda Problem [21 Pts]

4. It's college signing day! Our friend Isaac has gotten into many great universities, but doesn't know what college he wants to attend. Let's help him make his decision!

   Suppose we are provided with a DataFrame `university` with information about universities in the United States.

   Specifically,

   - `School` is the name of the university, as a str
   - `Acceptance Rate` is the acceptance rate of the school, as a float
   - `Tuition` is the tuition the school charges, in intervals of 1000, as an int
   - `Location` is the "City, State" location of the school, as a str.
   - `Degree Program` is the major program Isaac got offered acceptance to, as a str
   - `Accepted` is a binary variable, 1 if Isaac got accepted to that school, else 0, as a str

   We are also given a DataFrame `rankings` with the top national undergraduate universities from the US News.

   Specifically,

   - `School` is the name of the university, as a str
   - `Ranking` is the official US News ranking given to the university, as an int

   Provided are the first few rows of each DataFrame.

   **`University`**

| | School | Acceptance Rate | Tuition | Location | Degree Program | Accepted |
|---|---|---|---|---|---|---|
| 0 | MIT | 0.17 | 50000 | Cambridge, Massachusetts | Data Science | 1 |
| 1 | Stanford | 0.03 | 50000 | Stanford, California | Computer Science | 1 |
| 2 | UC Berkeley | 0.16 | 15000 | Berkeley, California | Data Science | 1 |
| 3 | Harvard | 0.06 | 60000 | Cambridge, Massachusetts | Statistics | 0 |
| 4 | UCLA | 0.12 | 15000 | Los Angeles, California | Statistics | 1 |

`Rankings`

| | University | Ranking |
|---|---|---|
| **0** | UC Berkeley | 1 |
| **1** | Harvard | 2 |
| **2** | Stanford | 3 |
| **3** | MIT | 4 |
| **4** | UCLA | 5 |

(a) [1 Pt] What type of variable is `Ranking`?

- ◯ Qualitative nominal
- ◯ **Qualitative ordinal**
- ◯ Quantitative discrete
- ◯ Quantitative continuous

(b) [1 Pt] What type of variable is `Tuition`?

- ◯ Qualitative nominal
- ◯ Qualitative ordinal
- ◯ **Quantitative discrete**
- ◯ Quantitative continuous

(c) [2 Pts] Isaac decides he will only consider attending a university if its tuition costs are low. Say we wanted to replace our `university` DataFrame with a new DataFrame which only contains schools which have tuition less than 50000. Which of the following line(s) of code accomplishes this? Select all that apply.

- ☐ `university = university["Tuition"] < 50000`
- ☐ **`university = university[university["Tuition"] < 50000]`**
- ☐ `university = university.iloc["Tuition" < 50000, :]`
- ☐ `university = university.loc["Tuition" < 50000]`
- ☐ **`university = university.loc[university["Tuition"] < 50000, ["School", "Acceptance Rate", "Tuition", "Location", "Degree Program", "Accepted"]]`**
- ☐ **`university = university.loc[university["Tuition"] < 50000, :]`**
- ☐ None of the above

5. To this new `university` table, add a new column `State` with the state that each university is located in. This column should not include any extra characters (punctuation, whitespace, letters).

(a) [4 Pts] Write a one line solution to solve this problem - no credit will be given for solutions exceeding one line.

*Hint: Consider using string methods.*

**Solution:**

```
university["State"] = university["Location"]
.str.split(", ").str[1]
```

6. For the following parts of the question, you can assume your answer to the last sub-part is correct. Suppose Isaac is only interested in considering attending schools in states where he was accepted to at least one Computer Science major among all schools in that state. Using our newly transformed table from question 4, fill in the following lines to return an array, named `idealStates`, which contains distinct states that satisfy this condition.

```
def isIdealState(group):
    if _________________A_________________:
        return True

accepted = university[university["Accepted"] == _B_]

idealStates = ____________________C____________________
```

Fill in the blanks in each part as indicated above.

(a) [3 Pts] What goes in the blank indicated by the letter `A`?

**Solution:**

```
any(group["Degree Program"]
.str.contains("Computer Science")):
```

(b) [1 Pt] What goes in the blank indicated by the letter `B`?

**Solution:**

```
1
```

(c) [3 Pts] Select the following line of code that goes in the blank indicated by the letter C.

- ◯ `accepted['State'].groupby("State").filter(isIdealState).unique()`
- ◯ `accepted.groupby("State").filter(isIdealState)["State"]`
- ◯ `accepted["State"].apply(isIdealState)["State"]`
- ◯ **`accepted.groupby("State").filter(isIdealState)["State"].unique()`**
- ◯ `accepted.groupby("State").agg(isIdealState)["State"].unique()`
- ◯ None of the above

7. For the following parts of the question, you can assume your answer to the previous sub-part is correct. That is, `idealStates` is defined correctly. Knowing what states he would like to attend school in (the states listed in `idealStates`), Isaac decides to limit his choices to the highest ranked school from each of these states. Fill in the following lines of code to return these schools.

```
idealSchools = _____A_____

complete = _____B_____

complete.___C___("Ranking").groupby(___D___).first()
```

Fill in the blanks in each part as indicated above.

(a) [1 Pt] What goes in the blank indicated by the letter A?

_____

(b) [3 Pts] What goes in the blank indicated by the letter B?

_____

(c) [1 Pt] What goes in the blank indicated by the letter C?

_____

(d) [1 Pt] What goes in the blank indicated by the letter D?

_____

**Solution:**

```
idealSchools = university[university["State"]
    .isin(idealStates)]
complete = idealSchools.merge(right=rankings,
    how="inner", left_on="School", right_on="University")
omplete.sort_values("Ranking").groupby("State").first()
```

# Tracking Packages [6 Pts]

8. One method of obtaining data is to use an API. APIs often return data in XML or JSON format, as briefly discussed in lecture. For example, the API to obtain package tracking data from the United States Postal Service returns a file containing tracking updates in XML format. Here is an example excerpt from an API call, with XML tags removed:

```
Your item was delivered in or at the mailbox at 11:14 am
    on February 13, 2020 in BERKELEY, CA 94704.
Out for Delivery, 02/13/2020, 7:10 am, BERKELEY, CA 94704
Arrived at Post Office, February 12, 2020, 6:29 pm,
    BERKELEY, CA 94710
USPS in possession of item, February 12, 2020, 2:22 pm,
    BERKELEY, CA 94710
Arrived Shipping Partner Facility, USPS Awaiting Item,
    February 11, 2020, 11:02 pm, UNION CITY, CA 94587
Departed Shipping Partner Facility, USPS Awaiting Item,
    February 10, 2020, 8:39 pm, AVENEL, NJ 07001
Arrived Shipping Partner Facility, USPS Awaiting Item,
    February 10, 2020, 7:03 pm, AVENEL, NJ 07001
```

(a) [4 Pts] Assume that there are only new lines after each status update. That is, there are only 7 lines in the above text. Write a regular expression to extract the state abbreviation from each update. Assume the abbreviation is any two-letter string immediately preceding the ZIP code. Your regex should be able to extract the abbreviation for any state.

Given your regex, the following code should run appropriately:

```
import re
pattern = ...
for line in lines:
    print(re.findall(pattern, line))
```

```
['CA']
['CA']
['CA']
['CA']
['CA']
['NJ']
['NJ']
```

Write your regex below:

```
r'_____________________________________________'
```

**Solution:**

```
r'([A-Z]{2})\s\d{5}'
```

(b) [2 Pts] Suppose we want to extract the date and time from the status updates above, then convert to Pacific Standard Time (PST). What qualities of the data make this a difficult task? Select all that apply.

- ☐ Not all status updates have a reported date and time.
- ☐ **The reported times lack time zone information.**
- ☐ **The reported dates do not follow a single, common format.**
- ☐ The data is not in a rectangular format.
- ☐ None of the above

# The SQL to the Problem [10 Pts]

9. Here, we have two tables. The `courseStaff` table contains a row for every individual on Summer 2021 TikTok Studies 101 course staff.

   Specifically,

   - `name` is the name of the student, as a str.
   - `sid` is a student's ID number, as an int.
   - `role` is the student's position they hold on staff, as a str.
   - `reportsTo` is the sid of the courseStaff member that the student reports to, as an int.
   - `gradeReceived` is the grade the student received when they took TikTok Studies 101, as an int.
   - `semesterTook` is the semester the student took the course, as a str.

   The `prof` table consists of all previous offerings of TikTok Studies 101 along with that semester's *singular lead professor*.

   Specifically,

   - `name` is the name of the professor, as a str.
   - `semesterTaught` is the semester that professor taught TikTok Studies 101, as a str.
   - `avgGradeGiven` is the average grade given to students, as a str.
   - `rateMyProfessorRating` is the average professor rating, as an int.

   Provided are the first few rows of each table. Note that each table shows a random subset of the data it holds. Assume there are no missing values in either table.

`courseStaff`

| | name | sid | role | reportsTo | gradeReceived | semesterTook |
|---|---|---|---|---|---|---|
| 0 | Brock Lee | 12345678 | Instructor | 12345678 | 98 | Spring 2019 |
| 1 | Anita Bath | 87654321 | TA | 12345678 | 92 | Spring 2018 |
| 2 | Joe Mama | 13483929 | Tutor | 87654321 | 88 | Summer 2020 |
| 3 | Bryce Hall | 123134489 | AI | 13483929 | 13 | Spring 2021 |

`prof`

| | name | semesterTaught | avgGradeGiven | rateMyProfessorRating |
|---|---|---|---|---|
| 0 | Prof A. | Spring 2020 | 84% | 4.3 |
| 1 | Prof D. | Spring 2019 | 76% | 3.2 |
| 2 | Prof S. | Fall 2020 | 80% | 2.8 |
| 3 | Prof B | Spring 2021 | 79% | 3.7 |

(a) [1 Pt] Which column(s) can be considered a primary key for the `courseStaff` table above? Assume there are no two individuals on course staff with the same name. Select all that apply.

- ☐ **courseStaff.name**
- ☐ **courseStaff.sid**
- ☐ courseStaff.role
- ☐ courseStaff.semesterTook

(b) [1 Pt] Which column(s) can be considered a primary key for the `prof` table above? Assume that a singular professor can teach multiple semesters of TikTok Studies 101. Select all that apply.

- ☐ prof.name
- ☐ **prof.semesterTaught**
- ☐ prof.avgGradeGiven
- ☐ prof.rateMyProfessorRating

(c) [2 Pts] Note that the `courseStaff` table contains 20 rows, while the `prof` table contains 15 rows. After executing the following query, how many rows will the resulting table have?

```
SELECT * FROM courseStaff, prof;
```

_____

**Solution:**

300

10. Suppose we wanted to write a SQL query that returns the course staff role and average grade received for each role who took the class with Prof A. The resulting table should be sorted in order of decreasing average grade received. Your SQL query must follow the following structure:

```
SELECT role, _____A_____
```

```
FROM courseStaff
INNER JOIN _____B_____
WHERE _____C_____
_____D_____
ORDER BY_____E_____
;
```

Fill in the blanks in each part as indicated above.

(a) [1 Pt] What goes in the blank indicated by the letter `A`?

_____

**Solution:**

```
AVG(gradeReceived) as [insert any alias here]
```

(b) [2 Pts] What goes in the blank indicated by the letter `B`?

_____

**Solution:**

```
prof ON semesterTook = semesterTaught
```

(c) [1 Pt] What goes in the blank indicated by the letter `C`?

_____

**Solution:**

```
prof.name = "Prof A."
```

(d) [1 Pt] What goes in the blank indicated by the letter `D`?

_____

**Solution:**

```
GROUP BY role
```

(e) [1 Pt] What goes in the blank indicated by the letter `E`?

_____

**Solution:**

```
ORDER BY [insert same alias here] DESC
```

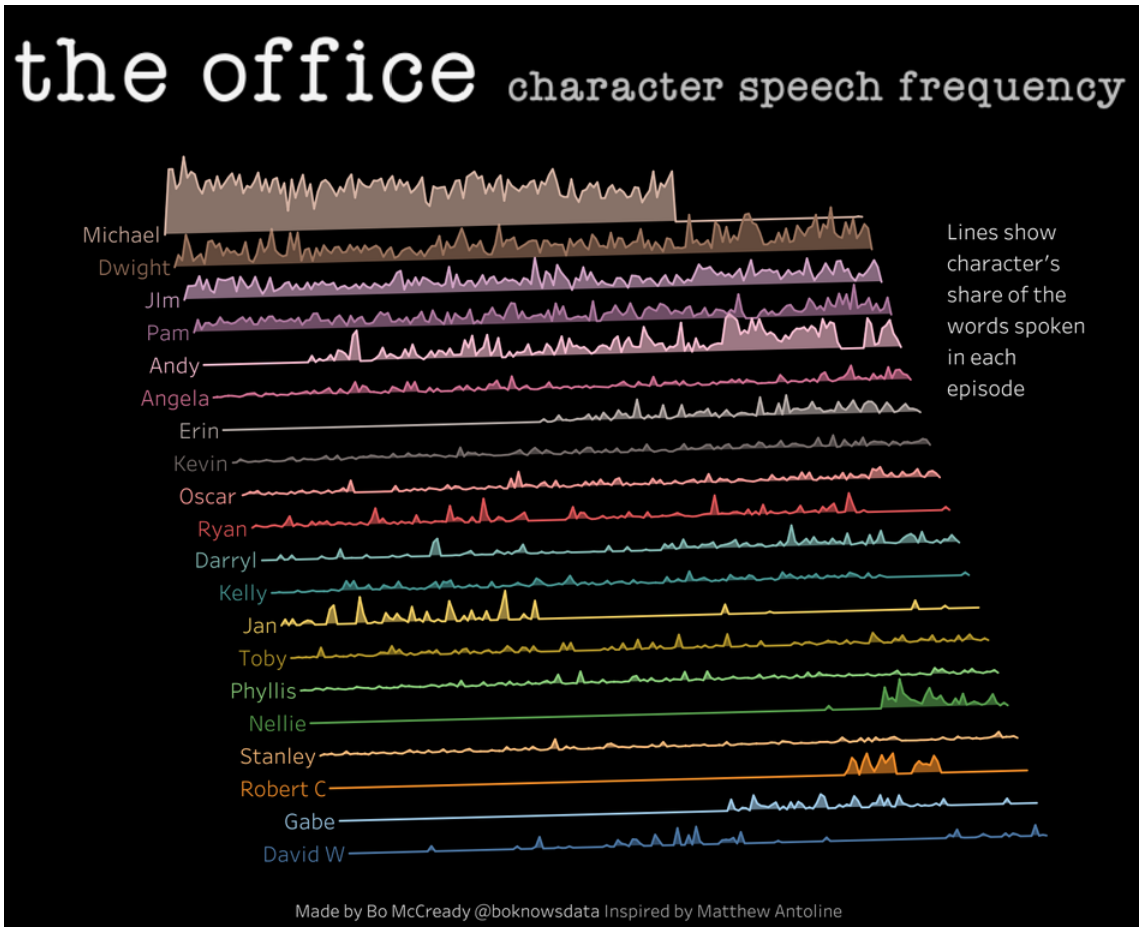
# Data Is Beautiful? [6 Pts]

11. In this question, you will see visualizations selected from among the top posts on r/dataisbeautiful, a subreddit devoted to data visualization. Following each visualization, you will be asked to point out one flaw with the figure, referencing concepts from lecture.

    While we have some clear answers in mind, we will be lenient when grading this question. As long as you point out some aspect of the image, and explain why it is a flaw, you will receive full credit. Answers such as "there is no flaw" or answers about the underlying data (and not the image itself) will receive no credit. **Please keep your answers concise**.

    Note that the titles for these plots are given directly above the image—answers such as "bad title" or "missing title" will also receive no credit.

    (a) [2 Pts] "The Office Character Speech Frequency" by u/BoMcCready




Name a flaw with the visualization:

| **Solution:** |
|---|

- The x-axis is not aligned for each character, making at hard to compare different characters' speech frequency at the same points in time.
- The x-axis is not labeled.

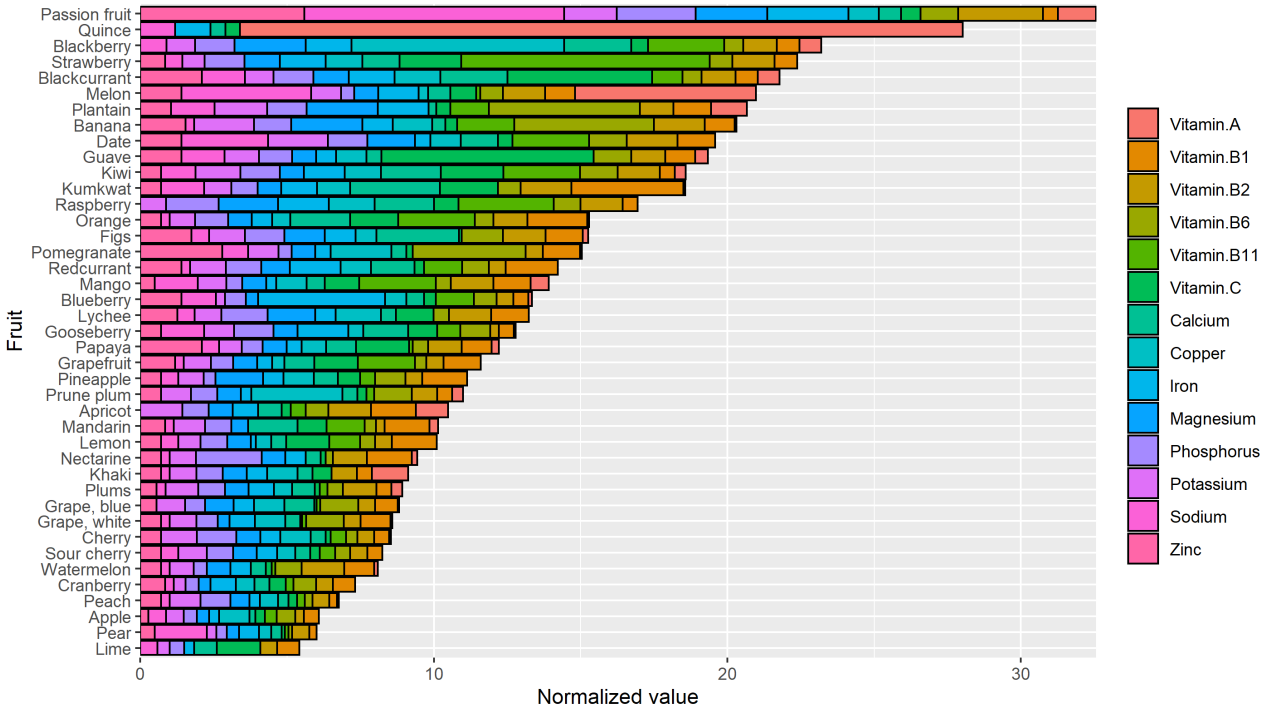
(b) [2 Pts] "Map of Italy showing only the towns and villages that have at least a BC year cited in their Wikipedia article (the darker the shade, the earlier the year cited)" by u/hd189

Name a flaw with the visualization:

**Solution:**

- There is no legend—we know a darker shade corresponds to an earlier time, but we have no idea exactly what times correspond to which colors.
- We are not sure what the white areas represent.

(c) [2 Pts] ”An apple a day, might keep the doctor away, but it definitely doesn’t contain the most vitamins and minerals in comparison to other fruits!” by u/Houses_of_Nick




Name a flaw with the visualization:

**Solution:** Possible answers:

- The baseline is jiggled, so we can not make a comparison of the difference in levels across fruits for a specific vitamin.
- 40 fruits clustered together makes the plot very busy, leading to overplotting
- We have no idea what the ”Normalized value” on the x-axis refers to.
- Colors for neighboring vitamins are so similar, making it hard to differentiate between them.

# Loss Functions [9 Pts]

12. Let's look at some loss functions!

(a) [5 Pts] Which of the following functions are **NOT** reasonable loss functions? Note that $\hat{y}$ is the prediction and $y$ is the actual value. Assume we can find the optimal parameters for each loss function.

**Hint:** Think about how we use the loss function in the modeling process to find the optimal model parameters.

- ☐ $y - \hat{y}$
- ☐ $\frac{1}{y-\hat{y}}$
- ☐ $(y - \hat{y})^2$
- ☐ $\frac{1}{(y-\hat{y})^2}$
- ☐ $e^{(y-\hat{y})^2}$
- ☐ $e^{-(y-\hat{y})^2}$

**Solution:** Since we *minimize* the loss function to find the optimal model parameters, a loss function must output higher values when the prediction $\hat{y}$ is far away from $y$ and lower values when $\hat{y}$ is close to $y$. Since Choice 4 and 6 do not have this property, they are not reasonable loss functions.

Additionally, Choice 1 is not a reasonable loss function because the optimal $\hat{y}$ is $\infty$ since we are minimizing the loss function. Choice 2 is a lot more complicated. The optimal $\hat{y}$ would be very slightly above $y$, as it would lead to a large negative loss. But if $\hat{y}$ is ever-so-slightly less than $y$, we have very large positive loss, and if $y = \hat{y}$, which would be a perfect prediction, the loss is undefined, so this is an unreasonable loss function.

(b) Suppose you are the owner of a restaurant and you have collected the following amount of tips over the last $n = 9$ days: $\{91, 92, 97, 98, 100, 103, 103, 106, 120\}$. You want to predict how much tip you will collect today.

You use the constant model $\hat{t} = \theta$, where $\hat{t}$ is your prediction for the amount of tips you will collect today. For each of the following loss functions, select the correct choice for $\hat{t}$. In the loss functions, $t_i$ signifies the tip on the $i^{\text{th}}$ day.

Note the following statistics regarding $t_1, \ldots, t_9$:

- sum = 910
- mean = 101.11
- median = 100
- mode = 103
- standard deviation = 8.17

i. [1 Pt] $\frac{1}{n}\sum_{i=1}^{n}(t_i - \theta)^2$

◯ 910 ◯ **101.11** ◯ 100 ◯ 103 ◯ 8.17 ◯ None of the above

ii. [1 Pt] $\frac{1}{n}\sum_{i=1}^{n}|t_i - \theta|$

◯ 910 ◯ 101.11 ◯ **100** ◯ 103 ◯ 8.17 ◯ None of the above

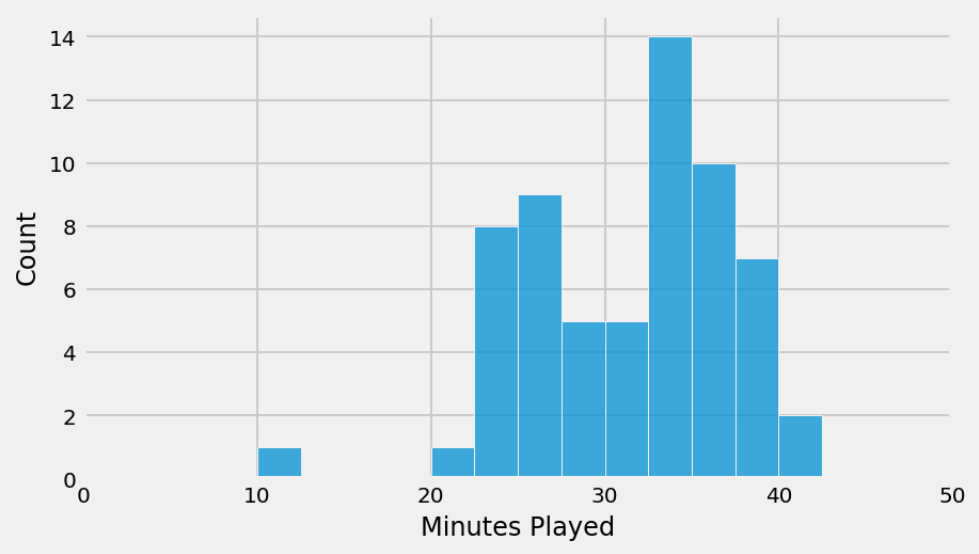
iii. [2 Pts] $\frac{1}{n}\sum_{i=1}^{n}[3(t_i - \theta)^2 + t_i]$

◯ 910 ◯ **101.11** ◯ 100 ◯ 103 ◯ 8.17 ◯ None of the above

**Solution:** The constant factors and the $t_i$ do not affect the optimal choice of $\theta$ since they go to 0 after differentiating with respect to $\theta$.

# Processing KDEs [6 Pts]

13. Below is a histogram displaying Joel Embiid's minutes played for each game throughout the 2020-21 NBA season.

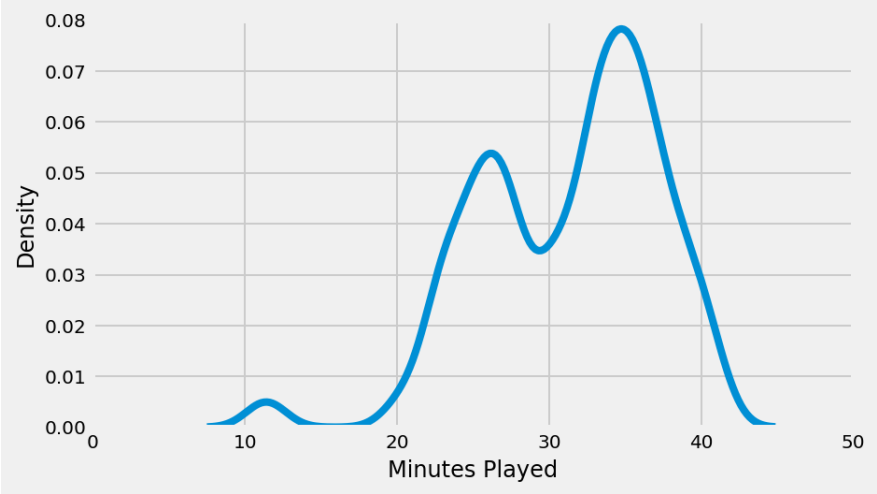



(a) [2 Pts] Which of the following describe the distribution from the histogram above? Select all that apply.

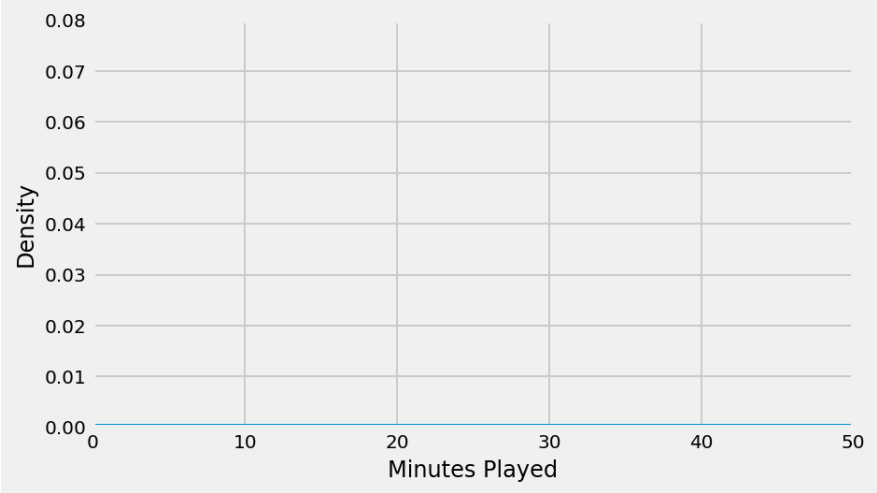
- ☐ Unimodal
- ☐ **Bimodal**
- ☐ **Skewed left**
- ☐ Skewed right

Below are 4 different kernel density estimates (KDEs) for the above distribution. Note that the x-axis and y-axis for all plots have the same scale.

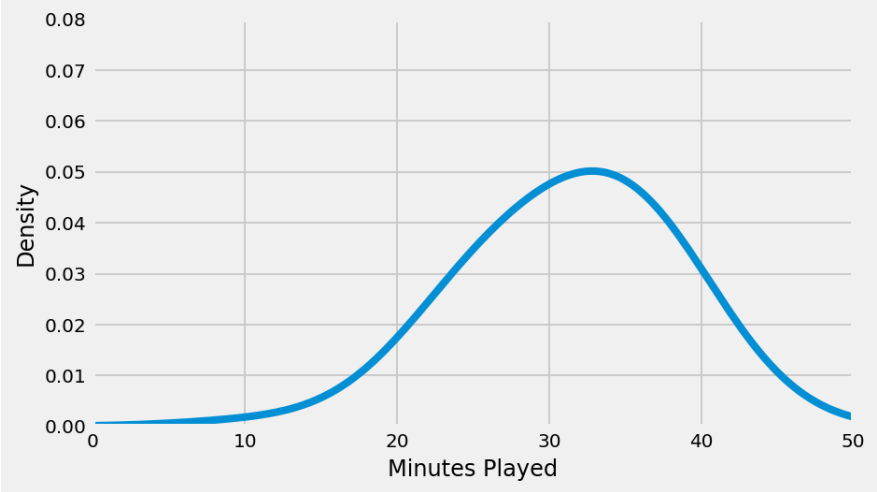
A.



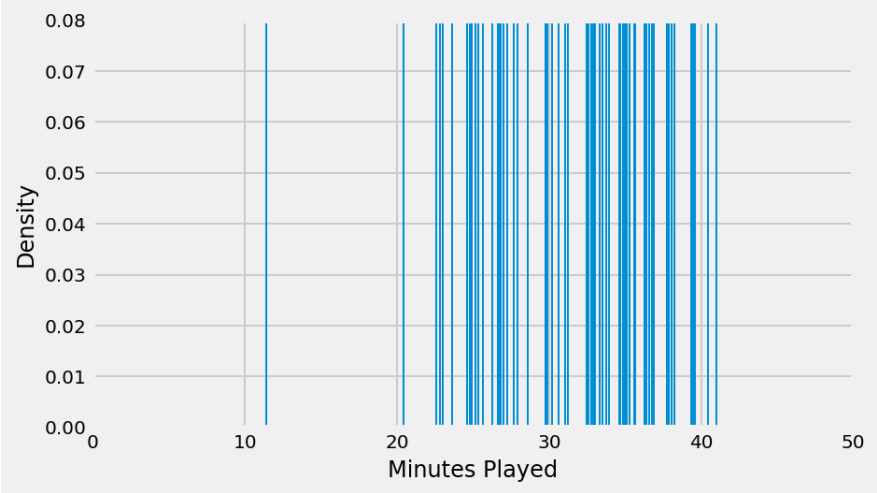
B.


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
0
10
20
30
40
50
Minutes Played

C.


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
0
10
20
30
40
50
Minutes Played

D.


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
0
10
20
30
40
50
Minutes Played

(b) [2 Pts] Which of the above KDEs (A, B, C, or D) best represents the limiting behavior of a KDE as the bandwidth parameter $\alpha$ goes to 0?

◯ A ◯ B ◯ C ◯ **D**

(c) [2 Pts] Which of the above KDEs (A, B, C, or D) best represents the limiting behavior of a KDE as the bandwidth parameter $\alpha$ goes to infinity?

◯ A ◯ **B** ◯ C ◯ D