Data 100 Checkpoint Assignment

Spring 2020

Name: _____

Email: _____@berkeley.edu

Student ID: _____

Instructions:

- This checkpoint assignment consists of **145 points** and must be completed in the **24 hour** time period ending at **8:00pm on Tuesday (3/10)**, unless you have accommodations supported by a DSP letter.
- For all questions on this assignment, there is **exactly one correct choice.**
- Please submit your answers electronically under "Checkpoint Assignment" through Gradescope.
- **Policy**: We have modified the policy for the checkpoint. **We encourage students to work individually. You can discuss the questions with your peers; however, you cannot share answers.** This aligns with the current assignment policy: you may collaborate, but you must submit your answers individually.
- **Piazza**: We will keep Piazza open for the duration of the checkpoint; however, **you may only make private posts to ask clarifying/logistical questions.** To mitigate answer sharing, you cannot use Piazza to get help on the checkpoint questions. Sharing answers and asking non-clarifying/non-logistical questions on Piazza constitutes cheating.

1 Sampling

1. [2 Pts] A bootstrap sample consists of n draws made uniformly at random with replacement from an original set of n individuals.

   Person A is an individual in the original set. The chance that Person A appears in the bootstrap sample is:

   ◯ $\frac{1}{n}$ ◯ $1 - \frac{1}{n}$ ◯ $\left(\frac{1}{n}\right)^n$ ◯ $1 - \left(\frac{1}{n}\right)^n$ ◯ $1 - \left(1 - \frac{1}{n}\right)^n$

2. The website of the Superior Court of California in Alameda County describes the population that is eligible to serve on juries, as follows. "You are eligible to serve as a juror if you are 18 years old, a U.S. citizen and a resident of the county or district where summoned. You must be able to understand English, and be physically and mentally capable of serving. In addition, you must not have served as any kind of juror in the past 12 months, must not be currently incarcerated in any prison or jail, and must not have been convicted of a malfeasance in office for which your civil rights have not been restored."

   Next, the website describes the process for random selection. "The names of jurors are selected at random from lists of registered voters. The law also allows that courts may use the names of all persons who have drivers licenses or identification cards issued by the Department of Motor Vehicles."

   (a) [1 Pt] This method of sampling leads to non-response bias because the sampling frame is smaller than the eligible population.

   ◯ True ◯ **False**

   (b) [1 Pt] This method of sampling leads to selection bias because the sampling frame is different from the eligible population.

   ◯ **True** ◯ False

   (c) [1 Pt] This method of sampling leads to response bias because the sampling frame is bigger than the eligible population.

   ◯ True ◯ **False**

   (d) [1 Pt] The described process of random selection is carried out. The data show that the average educational level of the selected jurors is higher than the average educational level of the eligible population. This is due to

   ◯ chance error but not bias

   ◯ bias but not chance error

◯ neither bias nor chance error

◯ **both bias and chance error**

Solution: See Lecture 2 Slide 16. Random samples have both kinds of error.

3. A university wants to study the experience of students enrolled in its big classes, defined as classes with enrollments of 500 or more. There are 20 such classes. From each of these classes, one enrolled student is chosen uniformly at random to take part in the university's survey. You can assume that the selection from each class is performed independently of the selections in the other classes. In this scenario:

   (a) [1 Pt] There are students in the population of interest who are not in the sampling frame.

   ◯ True ◯ **False**

   (b) [1 Pt] There are students in the sampling frame who are not in the population of interest.

   ◯ True ◯ **False**

   (c) [1 Pt] The method of sampling produces a probability sample of students enrolled in the big classes.

   ◯ **True** ◯ False

   (d) [1 Pt] The method of sampling produces a simple random sample of students enrolled in the big classes.

   ◯ True ◯ **False**

   (e) [1 Pt] Because a student is chosen from each class, all students in the big classes have the same chance of being selected.

   ◯ True ◯ **False**

   (f) [1 Pt] Because a student is chosen from each of 20 big classes, there will be 20 students in the sample.

   ◯ True ◯ **False**

4. A randomized controlled experiment has 100 participants. Each participant will be randomly assigned to the treatment or control group as follows: A fair coin will be tossed; if the coin lands heads the participant will be assigned to Treatment, and if it lands tails the participant will be assigned to Control.

   (a) [2 Pts] The chance that at least 45 participants get assigned to Treatment is:

- ◯ $\frac{45}{100}$
- ◯ $\binom{100}{45}0.5^{45}0.5^{55}$
- ◯ $\binom{100}{46}0.5^{46}0.5^{54}$
- ◯ $\sum_{k=45}^{100}\binom{100}{k}0.5^{k}0.5^{100-k}$
- ◯ $1-\sum_{k=0}^{45}\binom{100}{k}0.5^{k}0.5^{100-k}$

(b) [2 Pts] The chance that both groups have at least 45 participants is:

- ◯ The square of the answer to Part i
- ◯ $\sum_{k=45}^{55}\binom{100}{k}0.5^{k}0.5^{100-k}$
- ◯ $\sum_{k=90}^{100}\binom{100}{k}0.5^{k}0.5^{100-k}$

5. [2 Pts] A university offered two linear algebra classes last semester. Class I had 200 students of whom 30% received an A grade. Class II had 800 students and a tougher curve: only 20% of its students got an A. You can assume that no student was in both classes.

   If a student picked randomly from the 1000 students in the two classes got an A, the chance that the student took Class I is

   - ◯ 30%
   - ◯ 30% of $20\% = 6\%$
   - ◯ $\frac{30}{30+20} = 60\%$
   - ◯ $\frac{60}{60+160} \approx 27\%$
   - ◯ 20%
   - ◯ $\frac{20+30}{2} = 25\%$
   - ◯ $\frac{160}{60+160} \approx 73\%$

2 SQL

Inspired by the City of Berkeley's recent efforts to improve the quality of local roads, the state of California has begun to collect and analyze road quality around the whole state. To help with the analysis, the state has created the following database tables, and it's your job to help analyze the data. Keep in mind that there can only be one city in each state with a given name.

```
CREATE TABLE cities (
    name TEXT,
    population INT,
    year_founded INT,
    cars_per_capita FLOAT
);
CREATE TABLE roads (
    city_name TEXT FOREIGN KEY REFERENCES cities(name),
    road_name TEXT,
    road_length INT, // measures road length in feet
    road_quality FLOAT, // on a scale from 0-1
    PRIMARY KEY (city_name, road_name)
);
CREATE TABLE highways (
    highway_id INT PRIMARY KEY,
    start_city TEXT FOREIGN KEY REFERENCES cities(name),
    end_city TEXT FOREIGN KEY REFERENCES cities(name),
    highway_quality FLOAT // on a scale from 0-1
);
CREATE TABLE highway_cities (
    city_name TEXT FOREIGN KEY REFERENCES cities(name)
    highway_id INT FOREIGN KEY REFERENCES highways(highway_id)
);
```

Note: The `highways` table only captures the highway's endpoints—the cities the highway begins and ends in. The `highway_cities` table captures *all* the cities the highway passes through.

6. [3 Pts] There is a missing `PRIMARY KEY` constraint on the `cities` table. Which of the following should be the primary key for `cities`?

   ◯ `name, population`

   ◯ `name, population, year_founded`

   ◯ **`name`**

   ◯ `name, year_founded`

7. [4 Pts] The first question the state asks you is to calculate the total number of cars in the state based on the data they have collected about each city. Which of the following queries answers that question?

◯ **`SELECT SUM(cars_per_capita * population) FROM cities;`**

◯
```
SELECT cars_per_city
FROM (
     SELECT cars_per_capita * population AS cars_per_city
     FROM cities
);
```

◯
```
SELECT *
FROM cities
WHERE cars_per_capita * population;
```

◯
```
SELECT cars_per_capita * population
FROM cities JOIN roads
    ON cities.name = roads.city_name;
```

8. Next, the state wants to know which cities in the state have the sub-standard road quality. According to the state's guidelines, a city has substandard road quality if the average quality of all the roads in the city is below `0.4`. Fill in the blanks below to write a query that calculates which cities in the state have sub-standard road quality, sorted in increasing order of road quality (i.e., the first row in the result should be the city with the worst road quality). The query should return the name of the city along with its average road quality.

```
SELECT _____(1)_____, _____(2)_____ as avg_quality
FROM cities JOIN roads
     ON cities.name = roads.city_name
GROUP BY _____(3)_____
HAVING _____(4)_____ < 0.4
ORDER BY avg_quality _____(5)_____;
```

(a) [1 Pt] What option should go in blank (1)?

◯ `population` ◯ **`name`** ◯ `year_founded`

(b) [1 Pt] What option should go in blank (2)?

◯ **`AVG(road_quality)`** ◯ `road_quality` ◯ `AVG(road_length)`

(c) [2 Pts] What option should go in blank (3)?

◯ `avg_quality` ◯ `road_name` ◯ **`name`**

(d) [2 Pts] What option should go in blank (4)?

◯ `road_quality` ◯ **`avg_quality`** ◯ `SUM(road_quality)`

(e) [1 Pt] What option should go in blank (5)?

◯ **ASCENDING** ◯ DESCENDING

Solution:

```
SELECT name, AVG(road_quality) as avg_quality
FROM cities JOIN roads
     ON cities.name = roads.city_name
GROUP BY name
HAVING avg_quality < 0.4
ORDER BY avg_quality ASCENDING;
```

9. [4 Pts] Now that the state knows which cities are culprits for having bad roads, they also want to investigate the relationship between cities and highways. However, they first want to know if there is a relationship between a city's population and the number of highways running through it. Which of the following queries will return each city's name as well as the number of highways it has running through it? Note that we are interested in every city a highway passes through, not just cities where the highways start or stop.

◯
```
SELECT name, highway_id
FROM cities JOIN highways
     ON cities.name = highways.start_name;
```

◯
```
SELECT name, COUNT(highway_id)
FROM cities, highway_cities
WHERE cities.name = highway_cities.city_name
GROUP BY cities.name;
```

◯
```
SELECT start_city, COUNT(*)
FROM highways
GROUP BY start_city;
```

◯
```
SELECT * FROM highway_cities;
```

3 Data Cleaning and Pandas

In this question, we will be looking at the `contest` dataframe which contains data from a math contest in 2019. In the contest, each participant had a total of five questions. The participants submit each question separately and each row of the DataFrame records a particular submission of one of the contestants by some participant. The `Timestamp` column specifies the time a given problem is submitted by a participant; each timestamp is discretized to the minute and has been properly converted to a Pandas `datetime` object with `pd.to_datetime`. The `Contestant` column contains the id-name pair of each participant. The `Question` column contains the question that was submitted. The `Correct` column tells us if the answer given in the submission is correct (1) or not (0). **Assume each participant can have several submissions for the same problem, but they can only submit one question per minute.**

	Timestamp	Contestant	Question	Correct
0	2019-11-17 14:09:00	1132E - Joe	1	0
1	2019-11-17 14:10:00	1362C - Bob	2	1
2	2019-11-17 14:10:00	0049A - Fred	2	1
3	2019-11-17 14:11:00	1362D - Ethan	1	1
4	2019-11-17 14:11:00	1362A - Steve	1	1
5	2019-11-17 14:11:00	0049E - David	1	1
6	2019-11-17 14:12:00	0027A - Michelle	4	1
7	2019-11-17 14:12:00	1362D - Ethan	2	1
8	2019-11-17 14:12:00	0016C - Grace	1	0
9	2019-11-17 14:12:00	0049E - David	2	1

10. [1 Pt] What is the granularity of the dataframe?

◯ **Submission** ◯ `Participant` ◯ `Timestamp`

11. Answer each of the following True/False questions:

(a) [1 Pt] `Timestamp` should be the primary key of this dataframe.

◯ True ◯ **False**

(b) [1 Pt] To best visualize the number of submissions for each question we should use a box plot with a separate box for each question.

◯ True ◯ **False**

(c) [1 Pt] `Contestant` is a nominal variable.

◯ **True** ◯ False

12. [3 Pts] Assuming that the column `"Question"` is of type `int` in python. Which of the following lines of code computes the total number of submissions for question 4 in the contest?

◯ `contest.groupby('Question').count().loc[4]`

◯ **`contest[contest['Question'] == 4].shape[0]`**

◯ `contest.iloc[contest['Question'] == 4].size`

◯ `contest.groupby('Question').filter(lambda x: x['Question'] == 4).shape[0]`

13. [3 Pts] Each value in the `"Contestant"` column contains both the name and the id of each contestant. Which of the following lines of code creates a new column `id` that contains the id of each contestant? Assume all ids are of length 5 and each entry of the `Contestant` column are formatted the same way and there are no spaces before or after any of the id-name pairs.

◯ `contest['id'] = contest['Contestant'].str[1:5]`

◯ `contest['id'] = contest['Contestant'].str.split('-')[0]`

◯ **`contest['id'] = contest['Contestant'].str[:5]`**

◯ `contest['id'] = contest['Contestant'].str[:, 5]`

14. [4 Pts] Notice that each participant may have several submissions for a problem. Which one of the following lines of code returns the most recent submission by each participant on Question 1? Larger timestamps correspond to more recent submissions.
Note: The solutions here may take more than one line. The symbol ";" indicates the end of a statement.

◯ **`contest[contest['Question'] == 1].sort_values('Timestamp', ascending = False).groupby('Contestant').agg('first')`**

◯ `temp = contest.groupby('Contestant').agg('max'); temp[temp['Question'] == 1]`

◯ `contest.groupby('Contestant').filter(lambda x: x['Question'].min() == 1)`

◯ `temp = contest.sort_values('Timestamp', ascending = False) .groupby('Contestant').agg('first'); temp[temp['Question'] == 1]`

15. [4 Pts] Some questions may be harder than others. Which of the following lines of code returns the question number of the question that has the **lowest** average score. Each student may have multiple submissions, we only want to include the latest submission in our calculation of the average score. **Note: The solutions here may take more than one line. The symbol ";" indicates the end of a statement.**

◯
```
contest.groupby('Question').mean().sort_values('Timestamp')
.index[0]
```

◯
```
 (contest.sort_values('Timestamp').groupby(['Contestant',
'Question'])
.filter(lambda x:  x['Correct'].min()).reset_index(1)
.sort_values().index[0])
```

◯
```
temp = contest.pivot_table(index = 'Contestant', columns
=
'Question', values = 'Correct');

temp['avg_score'] = temp.mean(axis = 1);

temp.sort_values('avg_score').index[0]
```

◯
```
 (contest.sort_values('Timestamp', ascending=False)
.groupby(['Contestant','Question']).first()
.reset_index(1)
.groupby('Question')['Correct']
.mean().sort_values().index[0])
```

4 Regex

The Data 100 TAs want to use regex to match student answers to determine if they should receive points or not.

16. Suppose the TAs want to award partial credit to students who placed a Kleene closure after a `\d`. A Kleene closure defines how many occurrences of something must be matched; see the code below for examples of Kleene closures. Which of the following regular expressions would work such that the following Python expression outputs correctly?

```
>>> answers = ["^\d{16}$", "\d{14}|\d", "\d+", "[3]\d*",
        "\d", "\d\w\dd+"]
>>> lst = [bool(re.findall(__________, answers[i])) for i
        in range(len(answers))]
>>> lst
[True, True, True, True, False, False]
```

Remember, `bool([])` is False and `bool(["strings"])` is True.
For the following regular expressions, select True if it leads to the correct Python output and False otherwise.

(a) [1 Pt] `r'\d[{+*]'` ◯ True ◯ **False**

(b) [1 Pt] `r'\\d[{+*]'` ◯ **True** ◯ False

(c) [1 Pt] `r'd[{+*]'` ◯ True ◯ **False**

(d) [1 Pt] `r'\d.+'` ◯ True ◯ **False**

(e) [1 Pt] `r'\\d.+'` ◯ True ◯ **False**

(f) [1 Pt] `r'\d.*'` ◯ True ◯ **False**

(g) [1 Pt] `r'\\d.*'` ◯ True ◯ **False**

17. Now, suppose the TAs want to award partial credit to students who correctly escape a dollar sign at the beginning of their regular expression. Which of the following regular expressions would work such that the following Python expression outputs correctly?

```
>>> answers = ["\$(.*?)", "\$\d+\.\d{2}", "\d+", "^\d+\$"]
>>> lst = [bool(re.findall(_____________, answers[i])) for i
        in range(len(answers))]
>>> lst
[True, True, False, False]
```

For the following regular expressions, select True if it leads to the correct Python output and False otherwise.

(a) [1 Pt] `r'\\\$.*'` ◯ True ◯ **False**

(b) [1 Pt] `r'\\$.*'` ◯ True ◯ **False**

(c) [1 Pt] `r'\$.*'` ◯ True ◯ **False**

(d) [1 Pt] `r'^\\\$.*'` ◯ **True** ◯ False

(e) [1 Pt] `r'^\\$.*'` ◯ True ◯ **False**

(f) [1 Pt] `r'^\$.*'` ◯ True ◯ **False**

18. Finally, the TAs want to see how many groupings a certain solution has. They only want to count groups which don't have groups inside of it. Which of the following regular expressions would work such that the following Python expression outputs correctly?

```
>>> answers = ["\$(.*?)", "\$((\d+\.*\d*)|(\d+))", "\d+",
        "^(\d+)|(\$)", "^((\d+)|(\$\))"]
>>> lst = [len(re.findall(_____________, answers[i])) for i
        in range(len(answers))]
>>> lst
[1, 2, 0, 2, 2]
```

For the following regular expressions, select True if it leads to the correct Python output and False otherwise.

(a) [1 Pt] `r'\(.*\)'` ◯ True ◯ **False**

(b) [1 Pt] `r'\('` ◯ True ◯ **False**

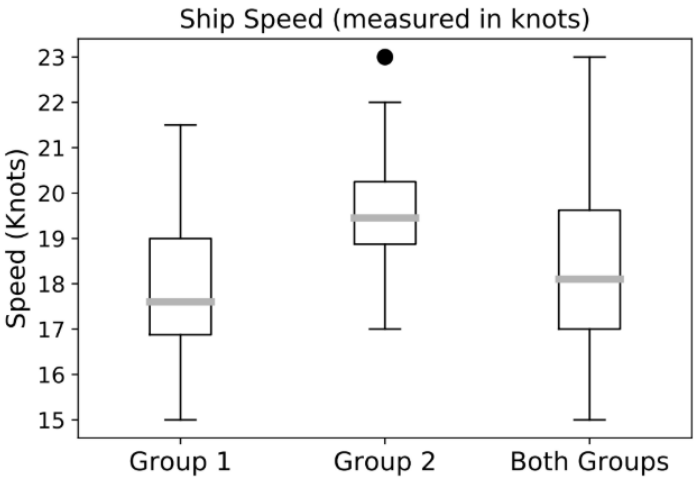
(c) [1 Pt] `r'(\(.*?\))'` ◯ **True** ◯ False

(d) [1 Pt] `r'(\(.*\)|\(.*?\))'` ◯ True ◯ **False**

(e) [1 Pt] `r'(\([^)]*\))'` ◯ **True** ◯ False

5 Visualization

19. The plot below summarizes the distributions of speeds of two groups of ships. The box plot on the extreme right is for all the ships. The other two box plots are for the individual groups.



For parts a through c, consider the box plot for Group 1. Based on this box plot alone, what can you conclude about the % of speeds in Group 1 that are 18 knots or more?

(a) [1 Pt] $< 50\%$ of the speeds in Group 1 are 18 knots or more

- ◯ True
- ◯ **False**

(b) [1 Pt] $\leq 50\%$ of the speeds in Group 1 are 18 knots or more

- ◯ **True**
- ◯ False

(c) [1 Pt] $\geq 50\%$ of the speeds in Group 1 are 18 knots or more

- ◯ True
- ◯ **False**

(d) [1 Pt] Which one of the following statements can we conclude from the boxplots?

- ◯ Since the box in the Group 1 plot is bigger than the box in the Group 2 plot, we can conclude that there are more ships in Group 1 than in Group 2.
- ◯ **Since the distribution in the Both Groups plot is much closer to that of Group 1 than of Group 2, we can conclude that there are more ships in Group 1 than in Group 2.**
- ◯ Based on these box plots, it is not possible to determine whether there are more ships in Group 1 than in Group 2.

Solution: The underlying data points in a boxplot are not guaranteed to be evenly spaced throughout the boxplot distribution. We cannot assume that there are, or are not, speeds between the median and 18 knots.

Given their distributions, we would expect the median of ”Both Groups” to be higher if both Group 1 and Group 2 were evenly sized. Similarly, we would also expect the IQR of ”Both Groups” to have a higher spread if Group 1 and Group 2 were evenly sized.

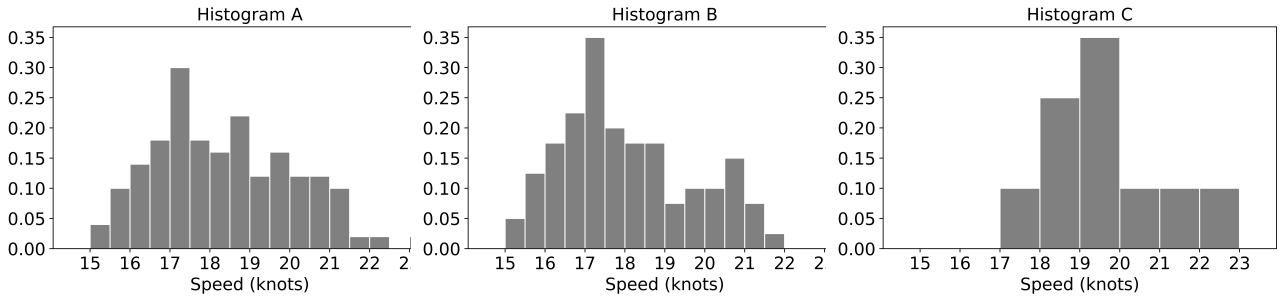
A. False

B. True

C. False

D. Group 1 has more ships than group 2.

20. All of the histograms below are based on the box plots above, and are drawn to the density scale.



For parts a through c, match the histograms to their corresponding box plot.

(a) [1 Pt] Histogram A: ◯ Group 1 ◯ Group 2 ◯ **Both Groups**

(b) [1 Pt] Histogram B: ◯ **Group 1** ◯ Group 2 ◯ Both Groups

(c) [1 Pt] Histogram C: ◯ Group 1 ◯ **Group 2** ◯ Both Groups

Solution:

A. Both groups. Median is around 17, and the max reaches over 23.

B. Group 1. Median is around 17, and the max is around 22.

C. Group 2. Median is around 20.

D. 10%. Height of the bin is 10%, width of the bin is 1.

E. Between 20% and 25%. The heights of the bins are around 30% and 17%, and the widths are $.5$.

(d) [1 Pt] About _____% of the speeds in Histogram C are in the [17, 18) bin.

◯ 5 ◯ **10** ◯ 15 ◯ 20 ◯ 25 ◯ 30 ◯ 35

(e) [1 Pt] In Histogram A, the percent of speeds in the [17, 18) range is:

◯ between 30 and 50
◯ **between 20 and 25**
◯ less than 10

21. Recall the Gaussian kernel and boxcar kernel functions given below. For the following questions, assume $\alpha > 0$.

Gaussian Kernel:

$$K_\alpha(x, z) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{(x-z)^2}{2\alpha^2}\right)$$

Box Car Kernel:

$$B_\alpha(x, z) = \begin{cases} \frac{1}{\alpha} & \text{if } -\frac{\alpha}{2} \leq x - z \leq \frac{\alpha}{2} \\ 0 & \text{else} \end{cases}$$

(a) [1 Pt] Suppose we want to find the kernel density estimation for 5 data points x_1 through x_5. Then, the kernel density estimate with a Gaussian kernel is $f_\alpha(x) = \sum_{i=1}^{5} K_\alpha(x, x_i)$.

- ◯ True
- ◯ **False**

(b) [1 Pt] If we wanted to visualize 5 data points, it is generally a good idea to use a rug or dot plot instead of a KDE.

- ◯ **True**
- ◯ False

(c) [1 Pt] Increasing α for a boxcar kernel increases the smoothness of the KDE.

- ◯ **True**
- ◯ False

(d) [1 Pt] If the height of the boxcar kernel is 5, then the width of the boxcar kernel should be $\frac{1}{5}$.

- ◯ **True**
- ◯ False

Solution:

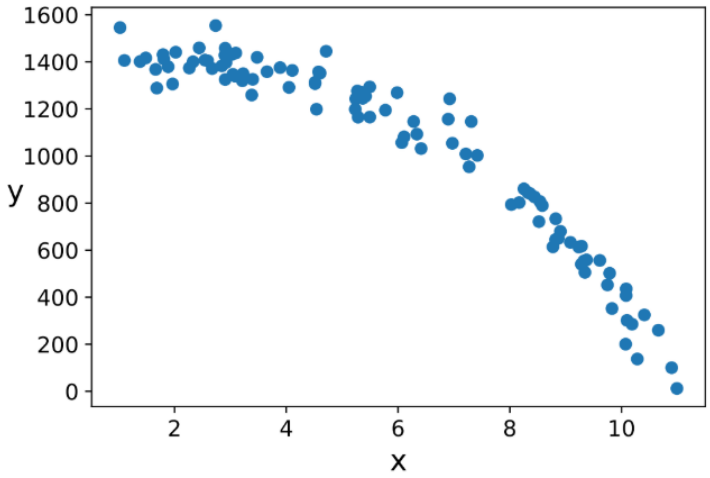
A. False; $f_\alpha(x) = \frac{1}{5} \sum_{i=1}^{5} K_\alpha(x, x_i)$

B. True

C. True

D. True; the area under the boxcar kernel should be 1.

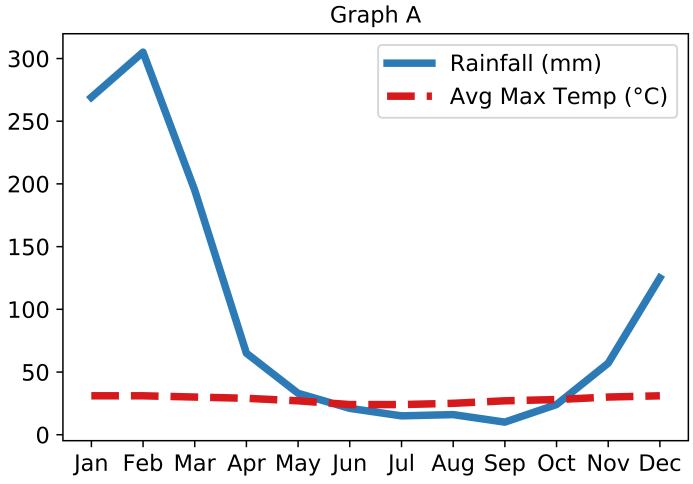
22. [2 Pts] Which one of the following transformations could help make more linear the relationship shown in the plot below?

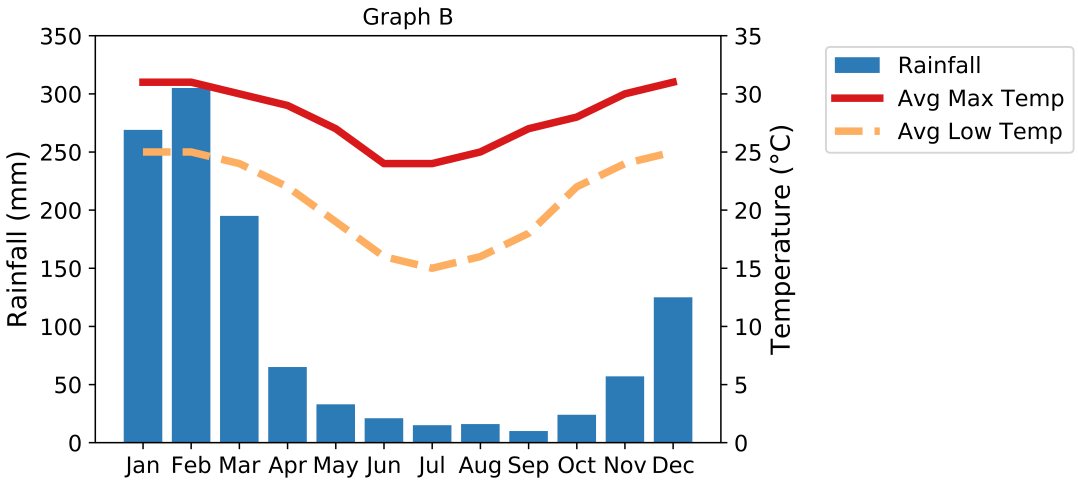


◯ e^x ◯ x^2 ◯ $x^{0.5}$ ◯ $\log(x)$ ◯ $\log(y)$

23. The following two graphs (A and B) show monthly temperature and rainfall trends in Townsville, Australia, where summer is from December to February. You can assume that both graphs represent data from a typical year in Townsville.

(°C * 9/5) + 32 = °F	
°C	°F
0	32
25	77
30	86





(a) [1 Pt] How many data variables are represented in Graph B?

◯ 1 ◯ 2 ◯ 3 ◯ **4**

(b) [1 Pt] In Graph A, the average max temperature is almost constant across the months. In Graph B, it has a clear dip during the winter months. Which of the following is the main reason for this apparent discrepancy?

- ◯ At least one of the graphs is based on incorrect data.
- ◯ Graph A is OK but Graph B has a problem because it has two different vertical scales for temperature and rainfall.
- ◯ **Graph B is OK but Graph A has a problem because it uses the same vertical scale for both variables.**

(c) [1 Pt] We can conclude that the temperature in Townsville is always above 10°C.

- ◯ True
- ◯ **False**

(d) [1 Pt] Select the reason for using a bar chart to represent rainfall and lines to represent temperature in Graph B.

- ◯ Rainfall is measured in units of length (mm) and so it has to be encoded by length, whereas many different encodings work for temperature.
- ◯ It was just the choice of the creator of the graph; it would have been equally good to use a double bar chart for the mean minimum and maximum temperatures and a line plot for rainfall.
- ◯ **Using a bar chart for rainfall and line plots for temperature results in chart that is easier to read than a double bar chart for temperature with a line plot for rainfall.**
- ◯ It was just the choice of the creator of the graph; it would have been equally good to draw overlaid scatter plots consisting of a point for each month, with x equal to rainfall for both plots and y equal to the mean maximum temperature for one plot and the mean minimum temperature for the other.

Solution:

A. 4. Date, rainfall, avg max temp, and avg low temp.

B. Graph B is OK, but Graph A has a problem. Combining the rainfall and temperature scales causes temperature trends to be lost since rainfall has a much higher range than temperature.

C. False. We only have information on average temperatures, so we cannot draw this conclusion.

D. Using a bar chart for rainfall is easier to read. Using a double bar chart for temperature would result in either a very wide graph or very thin bars. In addition, it would be harder to exclusively foucs on max temp or min temp, since the bars would be mixed together.

6 Modeling

24. In each situation below, say whether the stated conclusion is True or False.

(a) [2 Pts] Suppose our model has a parameter w, and denote our loss function by L_1. Now suppose we construct a different loss function L_2 defined by $L_2(w) = 2L_1(w) + 2$. Then the values of w that minimize L_1 and L_2 are different.

◯ True ◯ **False**

(b) [2 Pts] Data scientists are considering two different models for a data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Model 1 is $\hat{y}_1(w) = w_0 + w_1 x + w_2 x^2$ where $w = [w_0, w_1, w_2]$. Model 2 is $\hat{y}_2(\theta) = \theta x^2$ for a one-dimensional parameter θ. With both models, the data scientists will use average squared loss. Let L_1 be the loss function for Model 1 and L_2 the loss function for Model 2. Then $\min_w L_1(w) \leq \min_\theta L_2(\theta)$.

◯ **True** ◯ False

25. In order to estimate the value of a response variable y based on a predictor variable x, a data scientist decides to use a function of the form $\hat{y} = cx$ for some constant c, along with squared loss.

Let the data consist of n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

(a) [3 Pts] Let $L(c)$ be the average loss in using the line $\hat{y} = cx$ as the estimate. Find a formula for $L(c)$.

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

(b) [4 Pts] Find c^*, the value of c that minimizes the average loss in part (a).

◯ $\frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$
◯ $\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$
◯ $r\frac{\sigma_y}{\sigma_x}$ where r is the correlation between x and y, σ_y is the standard deviation of y, and σ_x is the standard deviation of x
◯ $\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^2}$

Solution:

$$L(c) = \frac{1}{n}\sum_{i=1}^n (y_i - cx_i)^2$$

$$\frac{\partial}{\partial c} L(c) = -\frac{2}{n}\sum_{i=1}^n (y_i - cx_i) \cdot x_i$$

Setting this equal to 0, we have:

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{c}x_i)\cdot x_i = 0 \implies \sum_{i=1}^{n} x_i y_i - \hat{c}\sum_{i=1}^{n} x_i^2 = 0 \implies \hat{c} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

26. A data scientist is trying to estimate the value of a numerical variable y by a constant θ. That is, no matter what the actual value of y happens to be, the data scientist will estimate it to be θ.

    For an observed value y and estimated value θ, the data scientist measures loss as follows:

$$l(\theta, y) = \begin{cases} 0 & \text{if } \theta = y \\ 1 & \text{if } \theta \neq y \end{cases}$$

    For observations $y_1, y_2, \ldots, y_n$, let $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, y_i)$.

    Suppose the data scientist has 10 observations 26, 26, 26, 26, 29, 29, 32, 32, 35, 38. The mean of the data is 29.9.

    (a) [3 Pts] For the given data, find $L(32)$.

    ◯ 0 ◯ 0.2 ◯ 0.4 ◯ 0.5 ◯ 0.6 ◯ **0.8** ◯ 1

    (b) [4 Pts] Let θ^* be the value of θ that minimizes $L(\theta)$. For the given data, find θ^*.

    ◯ **26** ◯ 29 ◯ 29.9 ◯ 32 ◯ 35 ◯ 38

7 Regression

27. Consider a data set of numerical (x, y) pairs. Let $\bar{x}$ and $\bar{y}$ denote the means of x and y respectively, let σ_x and σ_y denote the standard deviations of x and y respectively, and let r denote the correlation between x and y. Select whether each of the option below is the equation of the regression line for estimating y based on x.

(a) [2 Pts] $\hat{y} = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

◯ **True** ◯ False

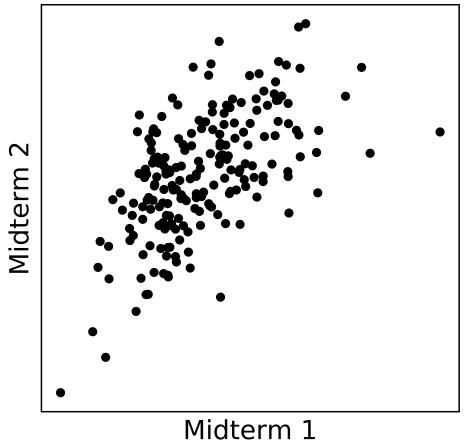
(b) [2 Pts] $\hat{y} = \bar{y} - r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

◯ True ◯ **False**

(c) [2 Pts] $\frac{\hat{y}-\bar{y}}{\sigma_y} = r\frac{x-\bar{x}}{\sigma_x}$

◯ **True** ◯ False

28. [3 Pts] A class has two midterms. A scatter plot of the scores on the two tests is shown below.



A student scores 1.2 standard deviations above average on Midterm 1. Select the correct option.

The least squares linear prediction of this student's Midterm 2 score based on the Midterm 1 score

◯ is more than 1.2 standard deviations above average

◯ is 1.2 standard deviations above average

◯ **is less than 1.2 standard deviations above average**

◯ may differ from the average by more than 1.2 standard deviations, or not; it is not possible to tell

Solution: Less than 1.2 SDs above average. This is an application of 27.3:

$\frac{\hat{y}-\bar{y}}{\sigma_y} = r\frac{x-\bar{x}}{\sigma_x}$

On the left hand side you have the regression estimate of the Midterm 2 score, in standard units of Midterm 2 scores. On the right hand side you have the given Midterm 1 score in standard units, times a positive fraction r. The product is smaller than the given Midterm 1 score in standard units.

By looking at the scatter diagram you can see that r is positive but not 1, so the inequality is strict.

The scatter diagram is slightly curved, but that doesn't matter. The least squares linear predictor is given by the equation of the same regression line you used in 27.

29. A data scientist studying a population of newborns is trying to estimate birth weight (measured in ounces) based on the number of gestational days (that is, the number of days the mother was pregnant). The equation of the regression line is

estimated birth weight = 0.47*(gestational days) - 10.75

Determine the correctness of each of the statements below based on the equation. You can assume that the mothers and babies referred to below are the same as or similar to those on whom the regression was performed.

(a) [2 Pts] The unit of measurement of the intercept is ounces.

◯ **True** ◯ False

(b) [2 Pts] Longer pregnancies are associated with higher birth weights.

◯ **True** ◯ False

(c) [2 Pts] If a mother's pregnancy is extended by one day, her baby is estimated to gain an additional 0.47 ounces in birth weight.

◯ True ◯ **False**

Solution: False. The slope estimates the average difference in y between two different groups separated by one unit in x, and therefore also the difference between individuals in the two different groups. But it doesn't make claims about y for a single individual as their x-value changes. The data are cross-sectional, not longitudinal. We are not watching an individual woman as her pregnancy progresses.

(d) [2 Pts] If Mother A's pregnancy lasted 10 more days than that of Mother B, then Mother A's newborn is estimated to weigh 4.7 ounces more than Mother B's newborn.

◯ **True** ◯ False

8 Gradient Descent

30. [1 Pt] Gradient descent is used in modeling to

- ◯ compute the loss.
- ◯ compute the gradient of the loss.
- ◯ maximize the loss.
- ◯ **minimize the loss.**

31. [1 Pt] For a model with k parameters and a dataset with n data points in d dimensions, how many dimensions is the gradient of the loss?

◯ 1 ◯ **k** ◯ k+1 ◯ d ◯ n

32. [1 Pt] Each iteration of stochastic gradient descent

- ◯ strictly increases the loss.
- ◯ strictly decreases the loss.
- ◯ **may increase or decrease the loss.**

33. [1 Pt] The gradient of the loss in stochastic gradient descent is

- ◯ is always negative (all entries in the gradient are negative).
- ◯ is always equal to zero (all entries in the gradient are 0).
- ◯ **computed on a sample or subset of the dataset.**
- ◯ computed on the entire dataset.

34. [1 Pt] Which of the following settings would most justify the use of stochastic gradient descent instead of gradient descent.

- ◯ All entries in the gradient of the loss are negative.
- ◯ **The dataset is large.**
- ◯ The data was constructed from a simple random sample.
- ◯ The loss function is not defined.

35. For the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ and the loss function:

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sin\left(w_0 + w_1 x_i\right)\right)^2$$

answer each of the following questions.

(a) [3 Pts] Which of the following is the gradient of the loss function with respect to $w = [w_0, w_1]$

- ◯ 0
- ◯ $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sin\left(w_0 + w_1 x_i\right)\right)^2$
- ◯ $\frac{1}{n}\sum_{i=1}^{n}\sin\left(y_i - \cos\left(w_0 + w_1 x_i\right) w_1\right)$
- ◯ $-\frac{2}{n}\sum_{i=1}^{n}\left(y_i - \sin\left(w_0 + w_1 x_i\right)\right)\cos\left(w_0 + w_1 x_i\right)[1, x_i]$
- ◯ $\left[-\frac{2}{n}\sum_{i=1}^{n}\cos\left(w_0 + w_1 x_i\right), -\frac{2}{n}\sum_{i=1}^{n}\cos\left(w_0 + w_1 x_i\right) x_i\right]$

(b) [2 Pts] If we consider a dataset consisting of three data points: $\mathcal{D} = \{(0,0), (2\pi, 1), (4\pi, 3)\}$, the loss value at an optimum w^* will be:

- ◯ 0
- ◯ Less than 0
- ◯ **Greater than 0**

(c) [2 Pts] If we consider a dataset consisting of three data points: $\mathcal{D} = \{(0,0), (\pi/2, 1), (4\pi, 0)\}$, the loss value at an optimum w^* will be:

- ◯ **0**
- ◯ Less than 0
- ◯ Greater than 0

36. Which of the following could affect whether or not gradient descent converges to the global minimum?

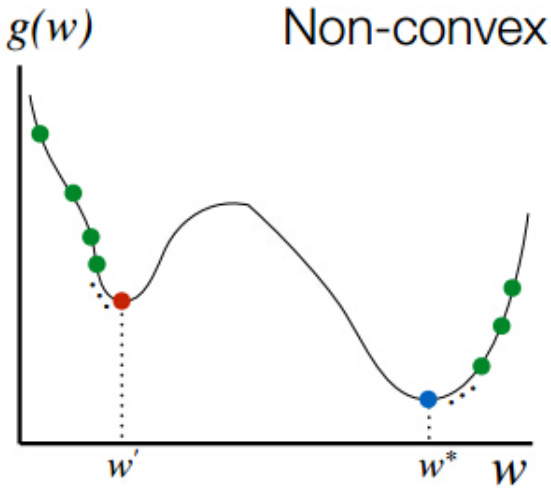
(a) [1 Pt] Learning Rate

◯ **True** ◯ False

(b) [1 Pt] Initialization of parameters

◯ **True** ◯ False

Solution: The initialization of parameters affects the starting parameters of gradient descent. If the loss function is not convex, the choice of starting parameters will affect whether the gradient descent algorithm converges to a local minimum or global minimum. The graph below shows a non-convex loss function with respect to parameter w.

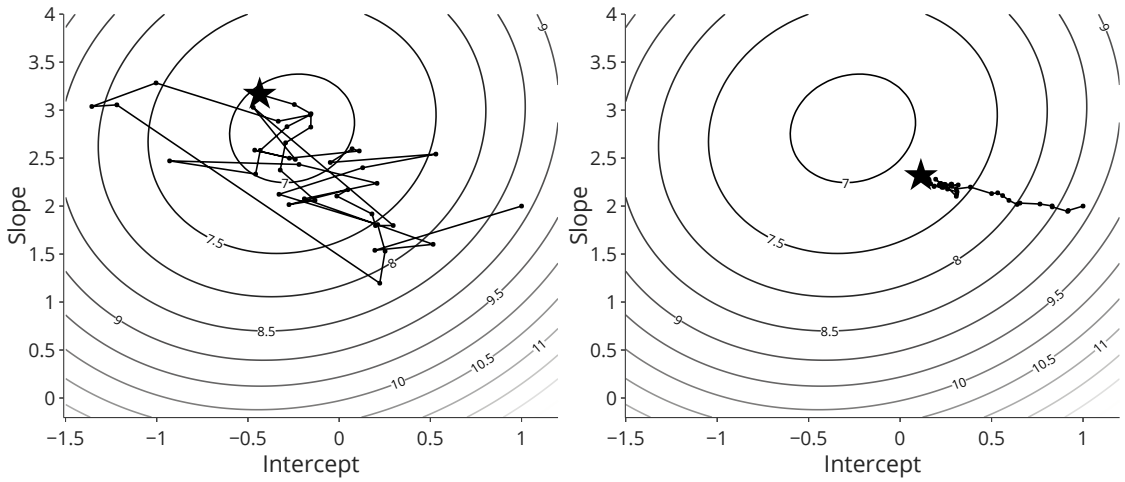


If we initialize w to be very small, then gradient descent will converge to a local minimum (given a reasonable learning rate). If we initialize w to be very large, then gradient descent will converge to a global minimum (given a reasonable learning rate).

(c) [1 Pt] The ordering of the data (ignoring issues related to numerical precision).

◯ True ◯ **False**

37. The following figures depict the contours of the loss surface as well as the solution path of parameters obtained by running an optimization algorithm. The star corresponds to the final parameter value. Use the following figures to answer each of the following questions.



(a) Path A (b) Path B

(a) [1 Pt] How many parameters does this model have?

◯ 1 ◯ **2** ◯ 3 ◯ 4

(b) [1 Pt] The sequence of parameters in **Path A** more likely came from an execution of:

◯ Gradient Descent ◯ **Stochastic Gradient Descent**

(c) [1 Pt] Which optimization path got closer to the optimal solution?

◯ **Path A** ◯ Path B

(d) [1 Pt] Which optimization path likely has the lower learning rate?

◯ Path A ◯ **Path B**