DATA 100 Summer 2020 Midterm 1

FINAL 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 `suraj.rampure@berkeley.edu`. 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 11:59PM 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:

1. (5 points)

The City of Berkeley wants to hear from its homeowners on issues related to zoning laws.

(For the purposes of this question, homeowners are individuals who own their home, instead of leasing or renting from someone else.)

(a) (1 pt) One method of surveying would be to have city workers come to UC Berkeley's campus and ask passing by students and faculty members for their thoughts. Suppose for now that the question "Are you a homeowner?" is not asked.

What type of sample is this?

◯ Convenience sample

◯ Probability sample, but not simple random sample

◯ Simple random sample

◯ Quota sample

(b) (1 pt) Many students and faculty members aren't homeowners, but will be surveyed anyways.

What form of bias or error is this?

◯ Response bias

◯ Non-response bias

◯ Chance error

◯ Selection bias

(c) (1 pt) The City of Berkeley has a list of all the homeowners' email addresses. Instead of the previous surveying technique, now suppose they take the list of all homeowners' email addresses, shuffle it, and send a survey to every other email address. That is, from the shuffled list, they email the first, third, fifth, seventh, and so on.

(You may assume that the shuffling is done uniformly at random, meaning that each email address has the same probability of landing in any particular position. You may also assume that the City of Berkeley has the email address for every single homeowner, and that every single homeowner has a unique email address.)

What type of sample is this?

◯ Quota sample

◯ Convenience sample

◯ Probability sample

(d) (1 pt) Fill in the blank: In this new sampling technique, the sampling frame is _____ the population of interest.

◯ equal to

◯ greater than

◯ smaller than

(e) (1 pt) In this new sampling technique, some homeowners may see the survey and choose not to respond. True or False: The only form of bias or error in this new surveying technique is non-response bias.

◯ True

◯ False

2. (6 points)

Consider a bag of 50 balls, 30 of which are red, and 20 of which are not red. Suppose we sample at random with replacement 10 times from our bag.

(a) (1 pt) What is the probability that there are exactly 7 red balls in our sample? Choose one.

- ◯ $\left(\frac{3}{5}\right)^7$
- ◯ $\binom{50}{30}\left(\frac{3}{5}\right)^7\left(1-\frac{3}{5}\right)^3$
- ◯ $\binom{10}{7}\left(\frac{3}{5}\right)^7\left(1-\frac{3}{5}\right)^3$
- ◯ $1-\left(\frac{2}{5}\right)^7$

(b) (1 pt) What is the expected value of the number of red balls in our sample? Choose one.

- ◯ 4
- ◯ 5
- ◯ 6
- ◯ 8

(c) (1 pt) What is the variance of the number of red balls in our sample? Choose one.

- ◯ $\frac{12}{5}$
- ◯ $\sqrt{\frac{12}{5}}$
- ◯ 3
- ◯ $\frac{12}{10}$
- ◯ None of the above

(d) (2 pt) What is the probability that there are at most 6 red balls in our sample? Select all that apply.

- ☐ $\sum_{k=0}^{6}\binom{10}{k}\left(\frac{3}{5}\right)^k\left(\frac{2}{5}\right)^{10-k}$
- ☐ $1-\binom{50}{6}\left(\frac{7}{10}\right)^6\left(1-\frac{3}{5}\right)^{42}$
- ☐ $1-\sum_{k=7}^{10}\binom{10}{k}\left(\frac{3}{5}\right)^k\left(\frac{2}{5}\right)^{10-k}$
- ☐ $6\cdot\frac{3}{5}$

(e) (1 pt) Let's suppose our sample is now a simple random sample of two balls from the bag. (That is, we are now sampling without replacement, and only 2 balls instead of 10.)

What is the probability that we get two red balls?

- ◯ $\frac{3}{5}\cdot\frac{29}{49}$
- ◯ $1-\frac{3}{5}\cdot\frac{29}{49}$
- ◯ $\left(\frac{3}{5}\right)^2$
- ◯ $1-\left(1-\frac{3}{5}\right)\left(1-\frac{29}{49}\right)$

3. (2 points)

We have a population of 100 individuals, 40 of whom are Canadian.

Suppose we sample at random with replacement 50 times from our population. Let C_i be an indicator variable for if person i in our sample is Canadian or not (where $i = 1, 2, 3, ..., 50$). That is, if person i is Canadian, $C_i = 1$, and otherwise, $C_i = 0$.

Recall from lecture that $E[C_i]$ is 0.4 and $SD[C_i]$ is roughly 0.48.

(a) (1 pt) What is $SD[1 - C_i]$? Select the closest answer.

- ◯ 0.008
- ◯ 0.4
- ◯ 0.48
- ◯ 0.52

(b) (1 pt) What is $E[\frac{1}{50}\sum_{i=1}^{50} C_i]$? Select the closest answer.

- ◯ 0.008
- ◯ 0.4
- ◯ 0.48
- ◯ 0.52

4. (7 points)

```
CREATE TABLE items_owned (
    name TEXT PRIMARY KEY,
    quantity INT,
    color TEXT,
    location TEXT
);
CREATE TABLE recipes (
    item_name TEXT,
    ingredient TEXT REFERENCES items_owned(name),
    quantity INT
);
CREATE TABLE market (
    item_name TEXT,
    price INT
);
```

Here, we have three tables. The `items_owned` table contains the items we own. An item, specifically, is anything we can pick up in the game. The `recipes` table is a list of recipes to make new items. The **item_name** column is the item created with the recipe. In order to make this item, you need an ingredient. The quantity needed of that one ingredient is in the **quantity** column. **Ingredients are items themselves**. Lastly, the `market` table is how much money you'd get (the **price** column) if you were to buy/sell the item (denoted in the **item_name** column) on the market.

Here are the first few rows of the tables defined above:

items_owned

name	quantity	color	location
table	1	red	house
fishing rod	2	blue	beach
wood	20	brown	house
clay	15	brown	airport

recipes

item_name	ingredient	quantity
table	wood	10
fishing rod	tree branch	5
chair	wood	6
statue	clay	20

market

item_name	price
table	100
swimsuit	30
chair	76
statue	1000
wood	2
tree branch	1
clay	3

To make things extra clear, the example tables above tell us:

- We own two blue fishing rods and they're located at the beach
- We can make one fishing rod using 5 tree branches
- If we wanted to buy a swimsuit from the market, it would cost 30 bells (where bells is the currency in Animal Crossing)

Answer all questions using SQL.

(a) (4 pt) Write a query that outputs a table that shows how many bells it would take if one were to buy all the ingredients needed to make each recipe in the recipes table. To be clear: the final table should have a column where each value is a necessary ingredient, and the second column is the cost of buying the needed amount of that ingredient.

If we assume the `recipes` table is only the four rows shown, then the query will output the following table:

col1	col2
wood	32
tree branch	5
clay	60

Note: The column names in this example may or may not be the same as your answer.

Note: If it wasn't clear before, recipes only take one type of ingredient, but the quantity of that ingredient is denoted by the quantity column.

(b) (3 pt) The blue colored items are some of the rarer items in the game. As a result, we want to write a query that outputs the same table as the items owned table, but with a new column called 'rare' that denotes if an item is rare or not based on the criteria above (if it's blue or not). This column will have two possible values: rare or not rare. These are the only two possible values.

5. (6 points)

Suppose Sally has a SQL table for the letters in her library.

```
CREATE TABLE letters (
    ISBN TEXT PRIMARY KEY,
    name TEXT ,
    author TEXT ,
    year_published INT ,
    publisher TEXT
);
```

However, some of this data is incomplete.

Note: The ISBN number acts as a unique identifier for the letters in the library (sort of like an index). Assume the ISBN column has no NULL values.

(a) (2 pt) Sally wants to count how many of the values in the name column are NULL. Select all of the following that does this. If you select None of the Above, make sure to remember what query you believe is correct, as it will be used for part b.

☐ `SELECT COUNT(name) FROM letters WHERE name IS NULL GROUP BY name;`

☐ `SELECT name FROM letters WHERE name IS NULL;`

☐ `SELECT COUNT(*) FROM letters WHERE name IS NULL;`

☐ `SELECT COUNT(name) FROM letters WHERE name IS NULL;`

☐ None of the Above

(b) (2 pt) Using the query above, Sally determines the number of null values in the name column to be 600. Now, Sally wants to make something called a mini-table. Let's make a table with just two columns: the ISBN column and the name column. However, she doesn't want any NULL values in this table. Using your answer from part a, pick the areas that need to be changed to achieve this:

Note: If you selected more than one answer in part a, pick one of them to use for this part. There may be more than one correct set of answers.

```
SELECT (a) FROM (b)
    WHERE (c) GROUP BY (d)
    HAVING (e) ,
);
```

For example, if you need to change the column you wish to group by, then select (d). If your answer in part a didn't use a WHERE clause, then don't select (c). If your answer in part a didn't use a HAVING clause, but you wish to add one now, select (e). Pick the minimal number of areas to be changed.

☐ (a)

☐ (b)

☐ (c)

☐ (d)

☐ (e)

☐ None of the Above

(c) **(2 pt)** Let's say Sally makes mini-tables for each of the four columns in her data. Each of these mini-tables, like the one she made for name in part b, do not have any NULL values in them.

Sally then asks Quentin to combine all four mini-tables using an OUTER JOIN on the ISBN column. This is done by joining each of the mini-tables one by one sequentially. Sally wants to check Quentin's work, so she counts all the NULL values in the name column of Quentin's new table using the query from part a. She determines the number of NULL values in the name column of Quentin's new table is 0.

Assuming Quentin's OUTER JOIN is correct, what MUST be true based on this information. Select all that apply.

☐ The number of NULL values in the author column of Quentin's new table is 600.

☐ This result is not possible if Quentin's OUTER JOIN was correct.

☐ There aren't any NULL values in any column of Quentin's new table.

☐ For the original data, if a row had a NULL name value, then all other values were NULL for that row (except the ISBN column).

☐ None of the Above

6. (6 points)

A customer gives us a coupon, but we want to determine if we should be using the coupon. We have a scanner that can extract important information from the coupon. We want to write a program using this information to verify if the coupon is correct.

Note: For both questions, if the Python code shown runs correctly for your answer, you will receive full points (even if it doesn't work **in general**).

(a) (2 pt) We work for a company called Mart that has many stores, but they all end with 'mart' (eg. Kmart, Foodmart, etc). Any store specific coupon can be used at any Mart store. Assume no store not owned by Mart ends with 'mart'. Complete the Python function below that takes in a name and returns True if the store name is owned by the parent company and False if not.

```
names = ['kmart', 'toymart', 'foodmart', \
            'walmart', 'martyr', 'martini', 'martmart']

def is_correct_company(name):
    pattern = r'_____________'
    return len(re.findall(pattern, name)) > 0

>>> [is_correct_company(name) for name in names]
[True, True, True, True, False, False, True]
```

(b) (4 pt) We want to use the same scanner to scan both coupons and Mart (the company we work for) rewards member cards. Rewards member cards are interesting in that they have a bunch of characters, but the only ones the scanner cares about (the actual member id) are the ones between two 'm's. The member id must not contain an 'm'.

Note: A reward member card could have multiple member ids.

```
members = ['kfhmaf9whmv', 'mhahahamamhahahaham', 'delaware', \
            'member1andmember2', 'decorumhere', 'kfhmamhellomwhmv']
pattern = r'_________________'
>>> for member in members:
...    print(re.findall(pattern, member))

['af9wh']
['hahaha', 'hahahaha']
[]
['e', 'e']
[]
['a', 'wh']
```

7. (19 points)

Suppose we are provided with a DataFrame `grades` that contains students' grade information from this class last semester. Here are the first four rows of `grades`:

	SID	midterm	final	score_UG	score_GR	type
0	552	69	93	95	94	undergrad
1	129	86	96	93	98	grad
2	412	91	63	87	82	undergrad
3	623	95	93	92	94	undergrad

Specifically,

- `SID` is a student's ID number.
- `midterm` and `final` contain a student's midterm and final exam score, as a percentage.
- `score_UG` contains a student's final grade in the class according to the undergraduate (Data 100) grading scheme. `score_GR` contains their final grade according to the graduate (Data 200) grading scheme.
- `type` tells us if a student is an undergraduate (enrolled in Data 100) or graduate (enrolled in Data 200) student.

(a) (1 pt) True or **False**: The primary key is at the same level of granularity as the entire table.

◯ False

◯ True

(b) (1 pt) What type of variable is `SID`?

◯ Qualitative ordinal

◯ Quantitative discrete

◯ Quantitative continous

◯ Qualitative nominal

(c) (2 pt) We define a student's "weighted exam score" as

$$\frac{3 \cdot (\text{final exam score}) + (\text{midterm exam score})}{4}$$

Write a line of Pandas code that determines the average weighted exam score for all students in the class. Your result should be a single number, not an array, Series, or DataFrame. Feel free to use numpy methods.

(d) (4 points)

The provided DataFrame contains both the undergraduate (`score_UG`) and graduate (`score_GR`) score for a student, but each student only belongs to one of the two categories, as given by the `type` column.

Fill in the blanks below to create a column `score_correct` in `grades` that contains the correct grade for each student. That is, if a student is an undergrad, their value in `score_correct` will be their undergraduate score, and if they are a grad student, their value in `score_correct` will be their graduate score.

```
def get_score_correct(r):
    if ___(1)___:
        return ___(2)___
    else:
        return r['score_GR']

df['score_correct'] = grades.___(3)___(get_score_correct, ___(4)___)
```

i.

- ◯ `r == 'undergrad'`
- ◯ `grades.loc[grades['type'] == 'undergrad', 'score_UG'] == r['score_UG']`
- ◯ `r['type'] == 'undergrad'`
- ◯ `r['type'].str.contains('grad')`

ii.

- ◯ `np.max(r['score_UG'], r['score_GR']`
- ◯ `'score_UG'`
- ◯ `grades['score_UG']`
- ◯ `r['score_UG']`

iii.

- ◯ `filter`
- ◯ `aggregate`
- ◯ `groupby`
- ◯ `apply`

iv.

- ◯ `axis = 1`
- ◯ `'score_UG'`
- ◯ `descending = True`
- ◯ `axis = 0`

(e) **(2 pt)** Now suppose we've computed students' final letter grades in the class correctly, as shown below.

	SID	midterm	final	score_correct	type	letter
0	552	69	93	95	undergrad	P
1	129	86	96	98	grad	A+
2	412	91	63	87	undergrad	B+
3	623	95	93	92	undergrad	A

We want to know the highest scores students got on the final, for each assigned grade bin and grade.

For example, we want to know the highest score undergrads who received an A in the course got on the final, and the highest scores that grad students who received a B in the course got on the final.

Assign `final_scores_by_grade` to a multi-indexed Series that contains this information. For instance, `final_scores_by_grade['Undergrad', 'A+']` should give us the highest score an undergrad who got an A+ in the class got on the final.

Select one.

◯ `final_scores_by_grade = grades.groupby('type').max().groupby('letter').max()['final']`

◯ `final_scores_by_grade = grades.pivot(['type, letter'], 'letter', 'final')`

◯ `final_scores_by_grade = grades.sort_values('letter', ascending = True).groupby(['type', 'letter']).last()['final']`

◯ `final_scores_by_grade = grades.groupby(['type', 'letter']).max()['final']`

(f) **(2 pt)** Which of the following expressions returns a Series of the SIDs of all students who performed below mean on the midterm, and didn't receive an A-, A, or A+ in the class? Select one of the following, or none of them:

Choice 1:

```
grades[(grades['midterm'] < grades['midterm'].mean())  & ('A' in list(grades['letter']))]
```

Choice 2:

```
grades[(grades['midterm'] < grades['midterm'].mean()) & (grades['letter'] < 'A+')]['SID']
```

Choice 3:

```
grades.loc[(grades['midterm'] < grades['midterm'].mean()) \
    & (~grades['letter'].str.contains('A')), 'SID']
```

◯ Choice 1

◯ Choice 2

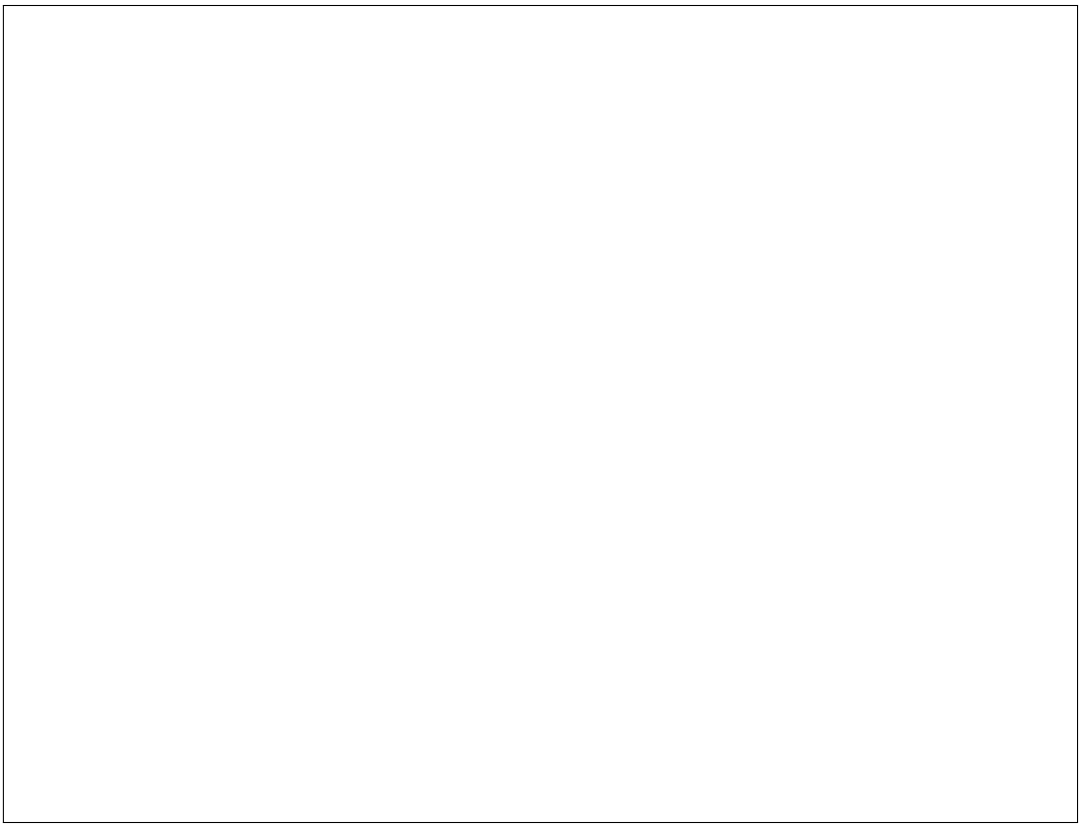
◯ Choice 3

◯ None of the above

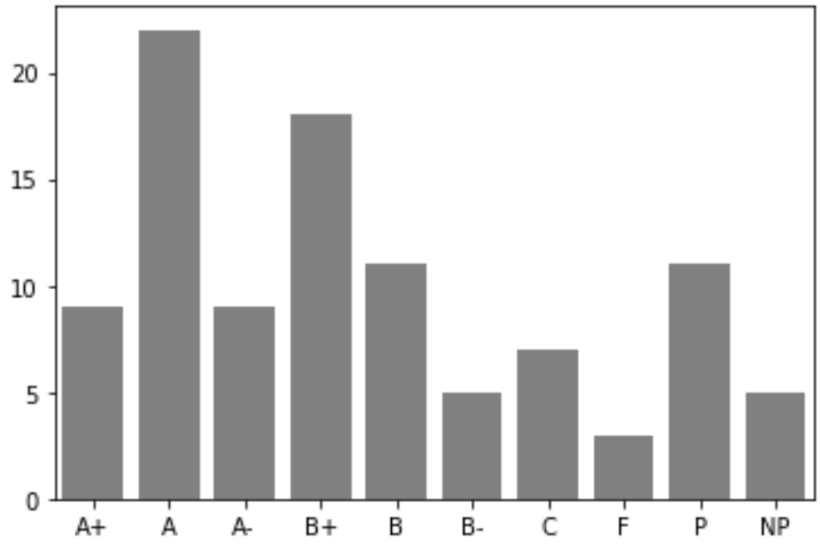
(g) **(4 pt)** Below, assign `null_grades` to a Series whose index is grade, and whose values are the total number of null midterm and final exam scores for students who had that grade. Take as many lines as you need.

To be clear: We are asking for the number of times the midterm or final column is null for each given grade.

```
letter
A     3
A+    6
B     0
B+    1
C     0
C+    0
NP    0
P     5
S     0
```



(h) (2 pt) Consider the following visualization.



Ignoring the order of the bars, which of the following lines of code could have resulted in this visualization? Select all that apply.

☐ `sns.countplot(grades['letter'])`

☐ `plt.hist(grades['letter'], bins = grades['letter'].index)`

☐ `plt.bar(grades['letter'].value_counts().index, grades['letter'].value_counts().values)`

☐ `sns.boxplot(grades['letter'])`

(i) (1 pt) Which of the following visualizations would **not** produce meaningful information? (Recall, `sns.distplot` computes a histogram of our data, along with an overlaid Kernel Density Estimate.) Choose one.

◯ `sns.distplot(grades['score_correct'])`

◯ `sns.distplot(grades['SID'])`

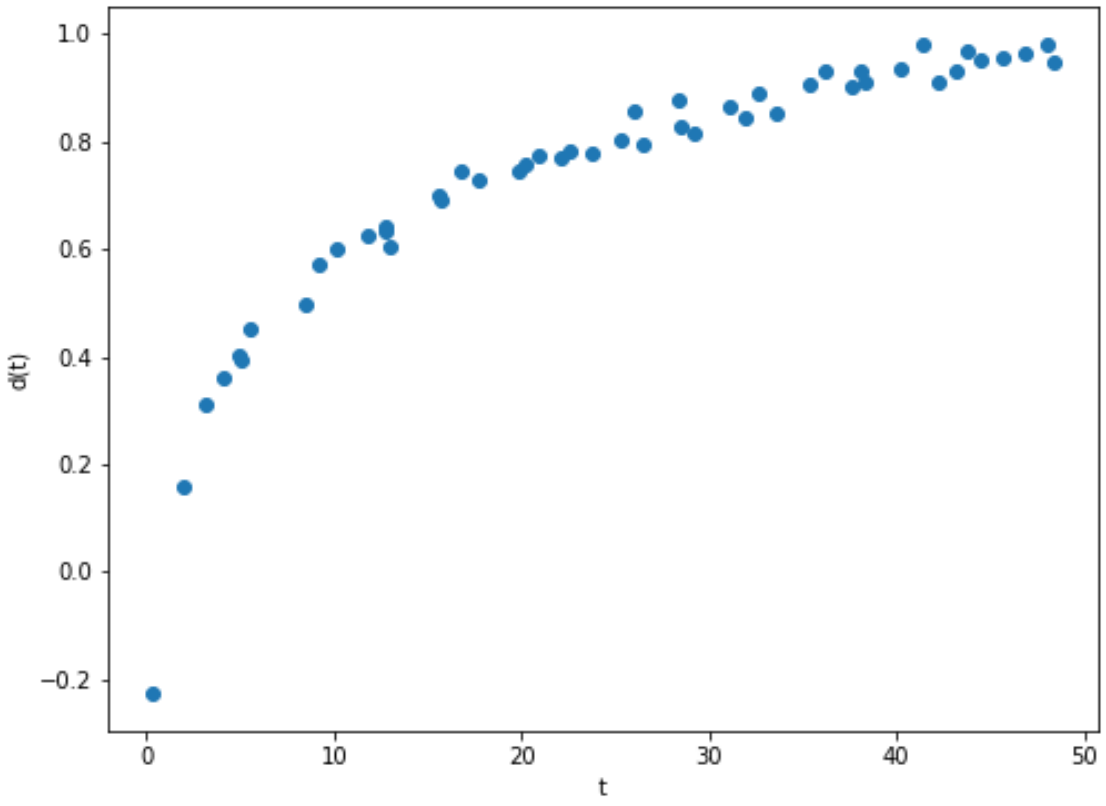
◯ `sns.distplot(grades['midterm'])`

◯ `sns.distplot(grades['final'])`

8. (3 points)

Consider the following scatter plot, generated by running

```
plt.scatter(t, d)
```



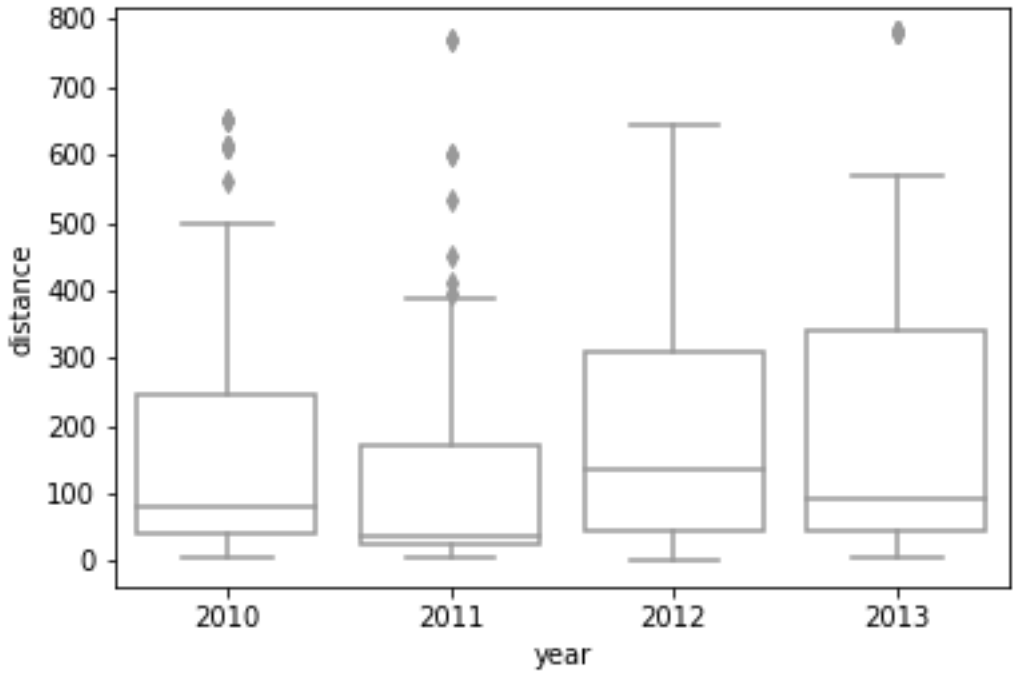
(a) (3 pt) Which of the following scatter plots would show a linear relationship? Select all that apply.

- ☐ `plt.scatter(t**2, d**2)`
- ☐ `plt.scatter(t**2, d)`
- ☐ `plt.scatter(t, d**2)`
- ☐ `plt.scatter(np.log(t), np.log(d))`
- ☐ `plt.scatter(t, np.log(d))`
- ☐ `plt.scatter(t, d**16)`
- ☐ `plt.scatter(np.log(t), d)`

9. (11 points)

We have a dataset, that contains a log of flight data. Specifically, it contains two columns; one containing the year in which a flight was taken, and another containing the distance traveled on that flight.

The following boxplot depicts the distributions of flight distances for the years 2010, 2011, 2012, and 2013.



(a) (1 pt) Which of the four years has the lowest median flight distance?

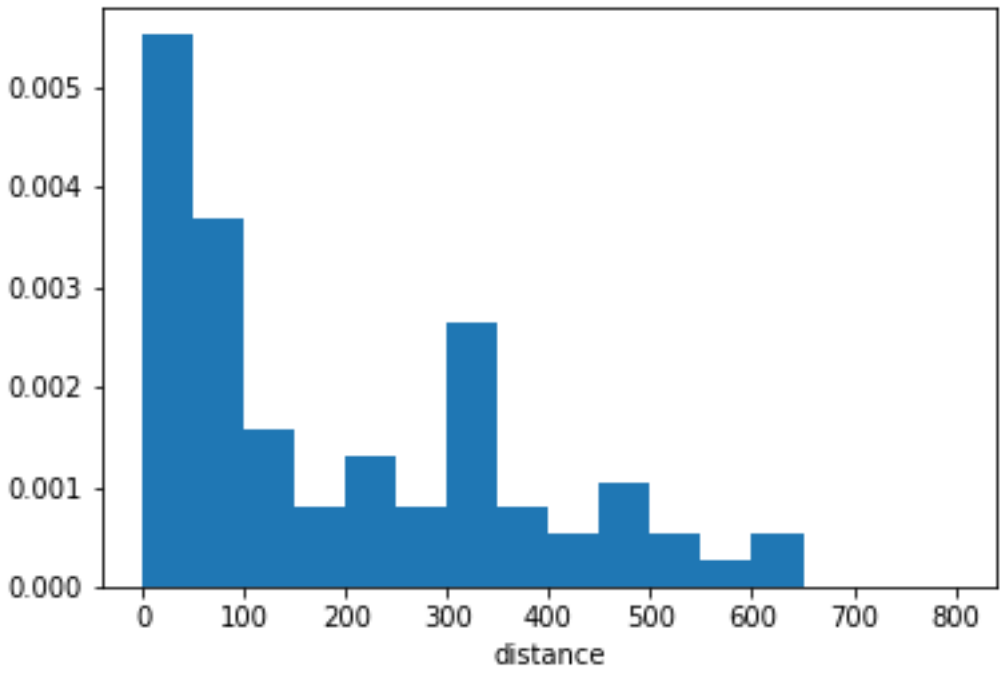
- ◯ 2010
- ◯ 2011
- ◯ 2012
- ◯ 2013

(b) (1 pt) Which year has a lower mean flight distance, 2012 or 2013?

- ◯ 2012
- ◯ 2013
- ◯ Their means are equal
- ◯ Impossible to tell

(c) (6 points)

Now, consider this histogram.



It shows the distribution of flight distances for one of the four years above. Which year's distribution does it show?

i. (2 pt)

- ◯ 2011
- ◯ 2012
- ◯ 2013

ii. (2 pt) Each bin in the histogram above has width 50. For the sake of simplicity, suppose bin [0, 50) has height 0.006, and bin [50, 100) has height 0.004. If there are 200 flight observations for this particular year, how many of them had a distance less than 100?

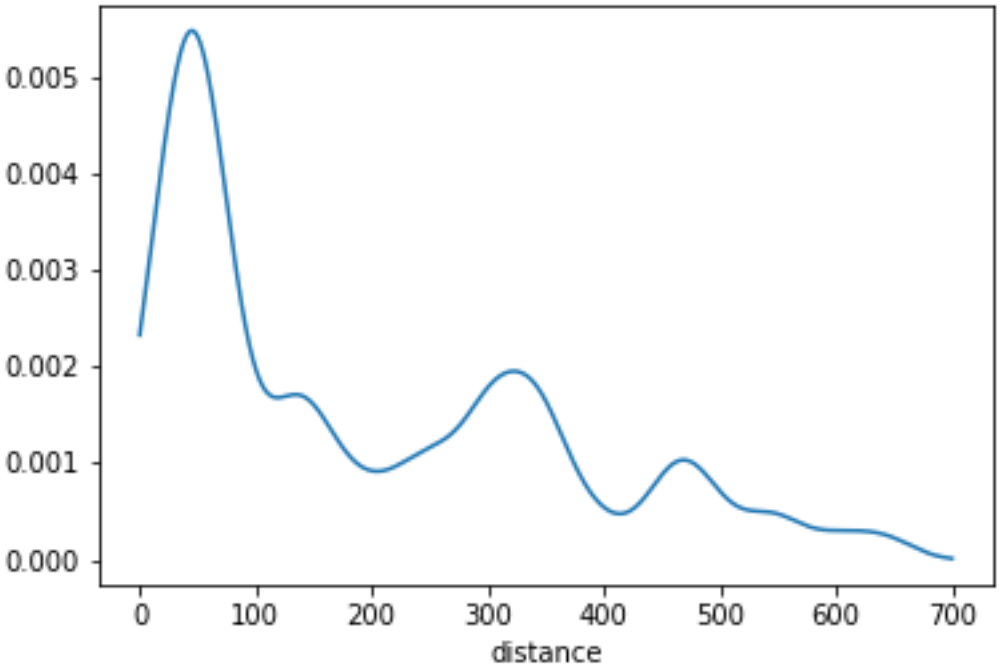
- ◯ 30
- ◯ 40
- ◯ 60
- ◯ 100

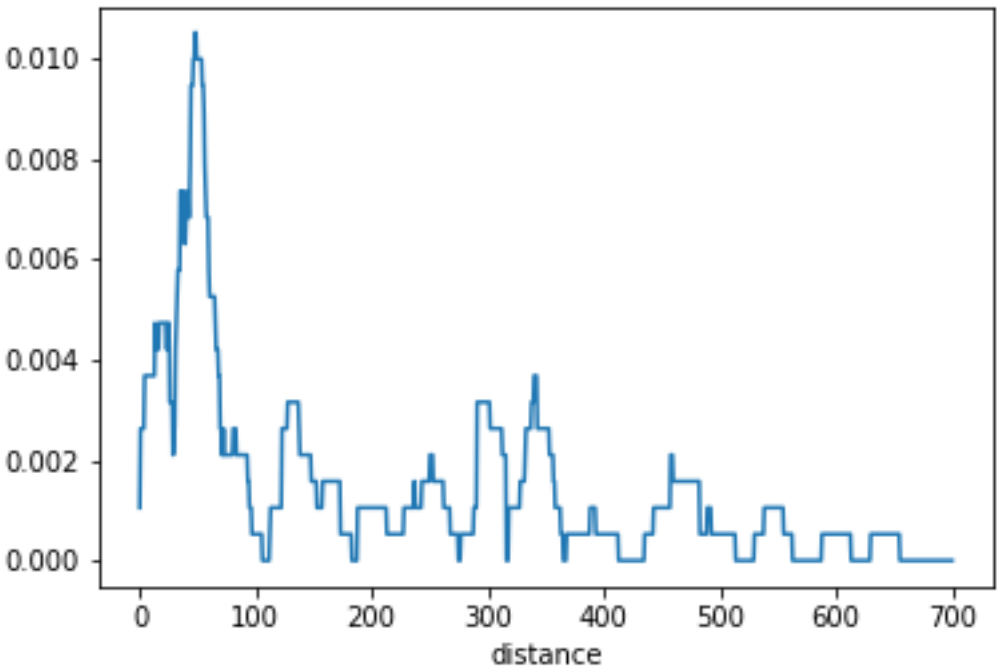
iii. (2 pt) Which of the following descriptions are true of the values described by the histogram above? Select all that apply.

- ☐ Left-skewed
- ☐ Right-skewed
- ☐ Left-tailed
- ☐ Right-tailed
- ☐ Mean is likely less than the median
- ☐ Mean is likely equal to the median
- ☐ Mean is likely greater than the median
- ☐ There are many outliers

(d) (3 points)

Two Kernel Density Estimates were created using the data from the previous histogram.





i. **(1 pt)** Which of the following Kernels was used to create the first (top) KDE?

◯ Gaussian kernel

◯ Boxcar kernel

ii. **(1 pt)** Which of the following Kernels was used to create the second (bottom) KDE?

◯ Gaussian kernel

◯ Boxcar kernel

iii. **(1 pt)** Which of the following is the most likely value of the bandwidth parameter, α, used to create the first KDE?

◯ 0.25

◯ 2.5

◯ 25

◯ 250

No more questions.