Sample 1

The City of Feishu 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 Feishu Univ'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 Feishu 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 Feishu 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

Sample 2

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 following DataFrame `ath` contains the names of athletes who participated in the Olympic Games, including all the Games from Athens 1896 to Tokyo 2020. The first 5 lines of the table are shown below. You may assume that the `ID` column is the primary key of the table.

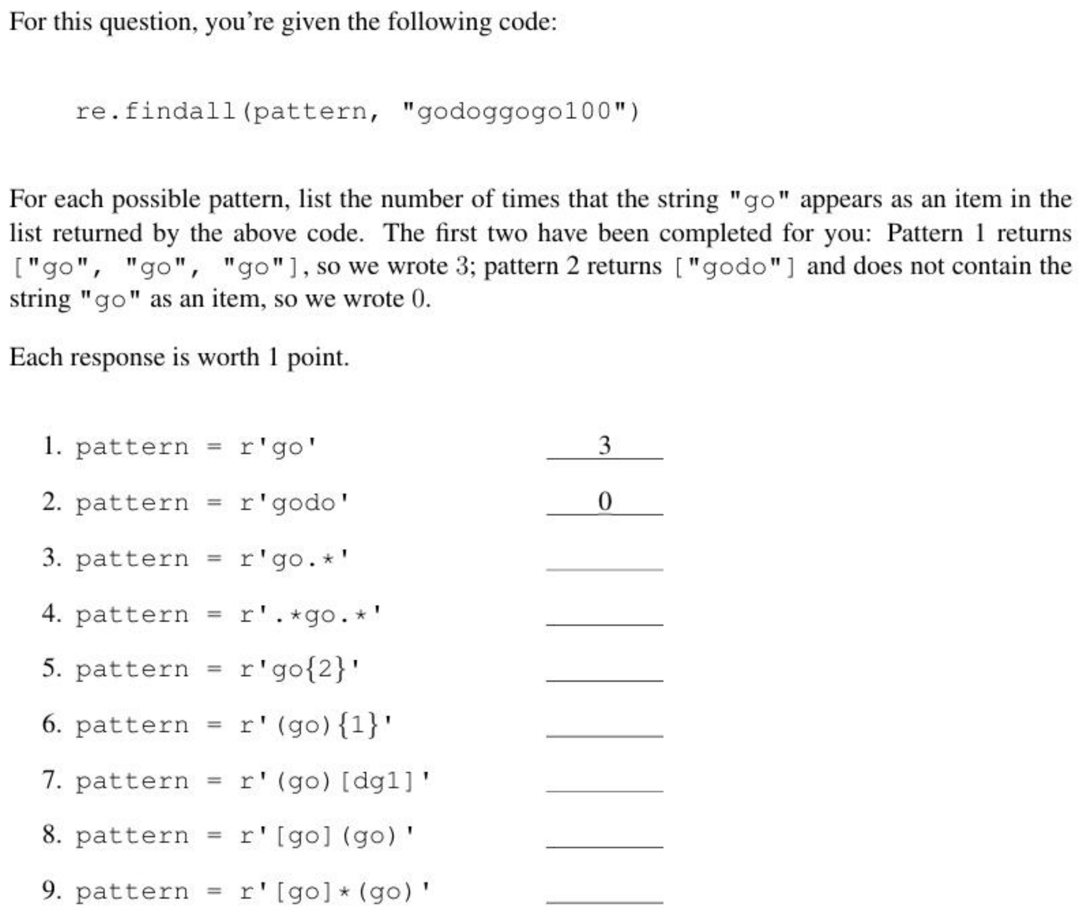
(a) [2 Pts] Choose the line of code that correctly sorts the number of unique Olympic events per year, in descending order of events. The result should show a series with the year and the number of events.

```
○ ath.groupby(’Year’)[’Event’].value_counts()
  .sort_values(ascending=False)

○ ath.groupby(’Event’)[’Year’].unique()
  .sort_values(ascending=False)

○ ath.groupby(’Year’)[’Event’].unique().agg(len)
  .sort_values(ascending=False)

○ ath.groupby(’Year’)[’Event’].unique()
  .sort_values(ascending=False)

○ ath.groupby(’Year’)[’Event’].unique().sort_values()

○ ath.groupby([’Year’,’Event’]).value_counts()
```

(b) [2 Pts] Choose the line of code that correctly identifies the athlete with the most medals of all time. The result should show a series with one row of the name of the athlete and the corresponding number of medals.

```
○ ath.groupby([’Name’,’Medal’]).count()
  .sort_values(ascending=False)

○ ath.groupby(’Name’)[’Medal’].count()
  .sort_values(ascending=False).head(1)

○ ath.groupby(’Medal’)[Name’].count()
  .sort_values(ascending=False).head(1))

○ ath.groupby(’Name’)[’Medal’].value_counts().head(1)

○ ath.groupby(’Medal’)[’Name’].value_counts().head(1)

○ ath.groupby(’Name’)[’Medal’].count().max()
```

Sample 3



For this question, you're given the following code:

```
re.findall(pattern, "godoggogo100")
```

For each possible pattern, list the number of times that the string "go" appears as an item in the list returned by the above code. The first two have been completed for you: Pattern 1 returns ["go", "go", "go"], so we wrote 3; pattern 2 returns ["godo"] and does not contain the string "go" as an item, so we wrote 0.

Each response is worth 1 point.

1. pattern = r'go' ___3___
2. pattern = r'godo' ___0___
3. pattern = r'go.*' ________
4. pattern = r'.*go.*' ________
5. pattern = r'go{2}' ________
6. pattern = r'(go){1}' ________
7. pattern = r'(go)[dg1]' ________
8. pattern = r'[go](go)' ________
9. pattern = r'[go]*(go)' ________

Sample 4

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.

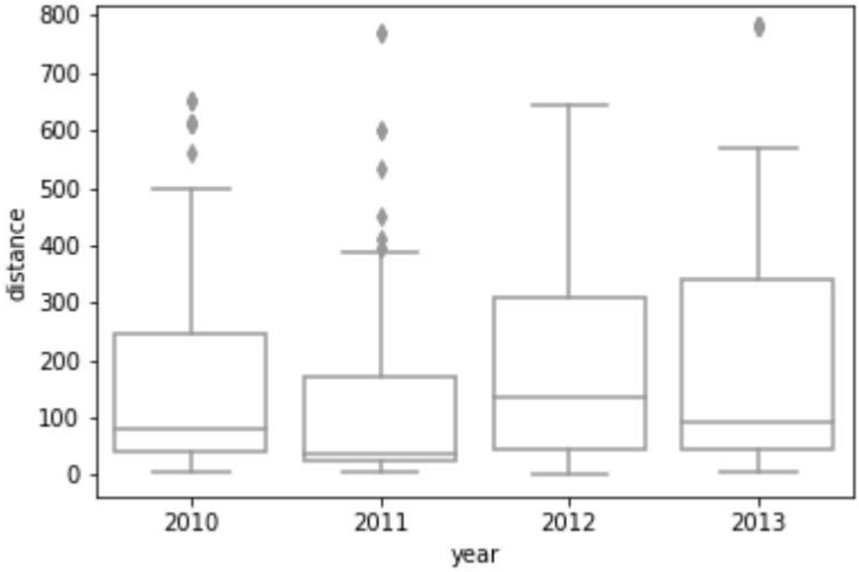
□ For the original data, if a row had a NULL name value, then all other values were NULL for that row (except the ISBN column).

□ None of the Above

Sample 5

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?

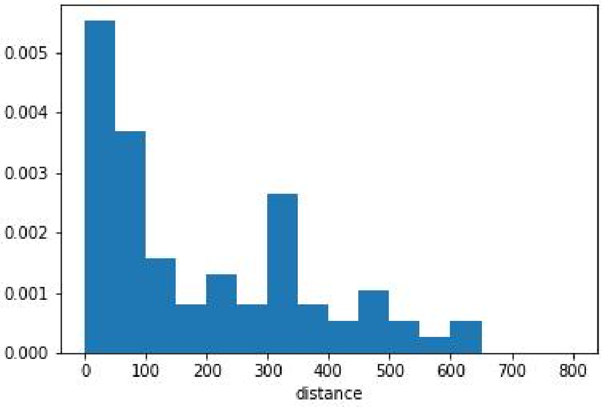
○ 2012

○ 2013

○ Their means are equal

○ Impossible to tell

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

Sample 6

Suppose we have one qualitative variable that that we convert to numerical values using one-hot encoding. We've shown the first four rows of the resulting design matrix below:

a	**b**	**c**
1	0	0
1	0	0
0	0	1
0	1	0

(a) [2 Pts] Say we train a linear model m_1 on these data. Then, we replace all of the 1 values in column **a** with 3's and all of the 1 values in column **b** with 2's and train a new linear model m_2. Neither m_1 nor m_2 have an intercept term. On the training data, the average squared loss for m_1 will be ____ that of m_2.

○ A. greater than
○ B. less than
○ C. equal to
○ D. impossible to tell

(b) [2 Pts] To account for the intercept term, we add a column of all ones to our design matrix from part a. That is, the resulting design matrix has four columns: **a** with 3's instead of 1's, **b** with 2's instead of 1's, **c**, and a column of all ones. What is the rank of the new design matrix with these four columns?

○ A. 1
○ B. 2
○ C. 3
○ D. 4

(c) [1 Pt] Suppose we divide our sampling frame into three clusters of people, numbered 1, 2, and 3. After we survey people, along with our survey results, we save their cluster number as a new feature in our design matrix. Before training a model, what should we do with the cluster column? (Note: This part is independent of parts a and b.)

○ A. Leave as is
○ B. One-hot encode it
○ C. Normalize it
○ D. Use bag of words