Midterm Exam SOLUTIONS – DSC 10, Winter 2021

Q1 RCT

4 Points

Which of the following questions could not be answered by running a randomized controlled experiment?

○ Does eating citrus fruits increase the risk of heart disease?

○ Do exams with integrity pledges have fewer reported cases of academic dishonesty?

○ Does rewarding students for good grades improve high school graduation rates?

● Does drug abuse lead to a shorter life span?

Q2 Sports

4 Points

You are given a table called `sports`, indexed by `Sport` containing one column, `PlayersPerTeam`. The first few rows of the table are shown below:

Sport	PlayersPerTeam
baseball	9
basketball	5
field hockey	11

Which of the following evaluates to "basketball"?

○ `sports.loc[1]`

○ `sports.iloc[1]`

● `sports.index[1]`

○ `sports.get('Sport').iloc[1]`

Q3 Covid Vaccine News

4 Points

The following is a quote from The New York Times' *The Morning* newsletter.

As Dr. Ashish Jha, the dean of the Brown University School of Public Health, told me this weekend: “I don’t actually care about infections. I care about hospitalizations and deaths and long-term complications.”

The data

By those measures, all five of the vaccines — from Pfizer, Moderna, AstraZeneca, Novavax and Johnson & Johnson — look extremely good. Of the roughly 75,000 people who have received one of the five in a research trial, not a single person has died from Covid, and only a few people appear to have been hospitalized. None have remained hospitalized 28 days after receiving a shot.

To put that in perspective, it helps to think about what Covid has done so far to ***a representative group of 75,000 American adults****: It has killed roughly 150 of them and sent several hundred more to the hospital. The vaccines reduce those numbers to zero and nearly zero, based on the research trials.*

Zero isn’t even the most relevant benchmark. A typical U.S. flu season kills between five and 15 out of every 75,000 adults and hospitalizes more than 100 of them.

Why does the article use *a representative group of 75,000 American adults*?

○ Convention. Rates are often given per 75,000 people.

● Comparison. It allows for quick comparison against the group of people who got the vaccine in a trial.

○ Comprehension. Readers should have a sense of the scale of 75,000 people.

○ Arbitrary. There is no particular reason to use a group of this size.

Q4 Employee Database

7 Points

Suppose you are given a table of employees for a given company. The table, called `employees`, is indexed by `employee_id` (string) with a column called `years` (int) that contains the number of years each employee has worked for the company.

Q4.1 Sorting

3 Points

Suppose that the code `employees.sort_values(by='years', ascending=False).index[0]` outputs "2476".

True or False: The number of years that employee "2476" has worked for the company is greater than the number of years that any other employee has worked for the company.

○ True

● False

Q4.2 Start

4 Points

What will be the output of the following code?

```
employees.assign(start=2021-employees.get('years'))
employees.sort_values(by='start').index.iloc[-1]
```

○ the employee id of an employee who has worked there for the most years

○ the employee id of an employee who has worked there for the fewest years

○ an error message complaining about `iloc[-1]`

● an error message complaining about something else

Q5 Boolean Array

3 Points

Suppose `df` is a DataFrame and `b` is any boolean array whose length is the same as the number of rows of `df`.

True or False: For any such boolean array `b`, `df[b].shape[0]` is less than or equal to `df.shape[0]`.

● True

○ False

Q6 Books Grouping

8 Points

You are given a table called `books` that contains columns `author` (string), `title` (string), `num_chapters` (int), and `publication_year` (int).

Suppose that after doing `books.groupby('Author').max()`, one row says

author	title	num_chapters	publication_year
Charles Dickens	Oliver Twist	53	1838

Q6.1 Yes/No 1

2 Points

Based on this data, can you conclude that "Charles Dickens" is the alphabetically last of all names listed in this dataset?

○ Yes

◉ No

Q6.2 Yes/No 2

2 Points

Based on this data, can you conclude that Charles Dickens wrote "Oliver Twist"?

◉ Yes

○ No

Q6.3 Yes/No 3

2 Points

Based on this data, can you conclude that "Oliver Twist" has 53 chapters?

○ Yes

◉ No

Q6.4 Yes/No 4

2 Points

Based on this data, can you conclude that Charles Dickens wrote a book with 53 chapters that was published in 1838?

○ Yes

● No

Q7 Open-Ended: Grouping with Subgroups

4 Points

Give an example of a dataset and a question you would want to answer about that dataset which you would answer by grouping with subgroups (using multiple columns in the `groupby` command). Explain how you would use the `groupby` command to answer your question.

Creative responses that are different than ones we've already seen in this class will earn the most credit.

Q8 Apply

4 Points

Which of the following best describes the input and output types of the `.apply` Series method?

○ input: string, output: Series

○ input: Series, output: function

● input: function, output: Series

○ input: function, output: function

Q9 Restaurant

8 Points

You are given a table called `restaurants` that contains information on a variety of local restaurants' daily number of customers and daily income. There is a row for each restaurant for each date in a given five-year time period.

The columns of `restaurants` are `name` (string), `year` (int), `month` (int), `day` (int), `num_diners` (int), and `income` (float).

Assume that in our data set, there are not two different restaurants that go by the same `name` (chain restaurants, for example).

Q9.1 Visualization 1

4 Points

What type of visualization would be best to display the data in a way that helps to answer the question "Do more customers bring in more income?"

● scatterplot

○ line plot

○ bar chart

○ histogram

Q9.2 Visualization 2

4 Points

What type of visualization would be best to display the data in a way that helps to answer the question "Have restaurants' daily incomes been declining over time?"

○ scatterplot

● line plot

○ bar chart

○ histogram

Q10 Grocery Store Prices

22 Points

You have a table of data called `prices` that contains information about food prices at 18 different grocery stores. There is column called `broccoli` that contains the price in dollars for one pound of broccoli at each grocery store. There is also a column called 'ice_cream` that contains the price in dollars for a pint of store-brand ice cream.

Q10.1 Data Type

4 Points

What should `type(prices.get('Broccoli').iloc[0])` output?

○ int

● float

○ array

○ Series

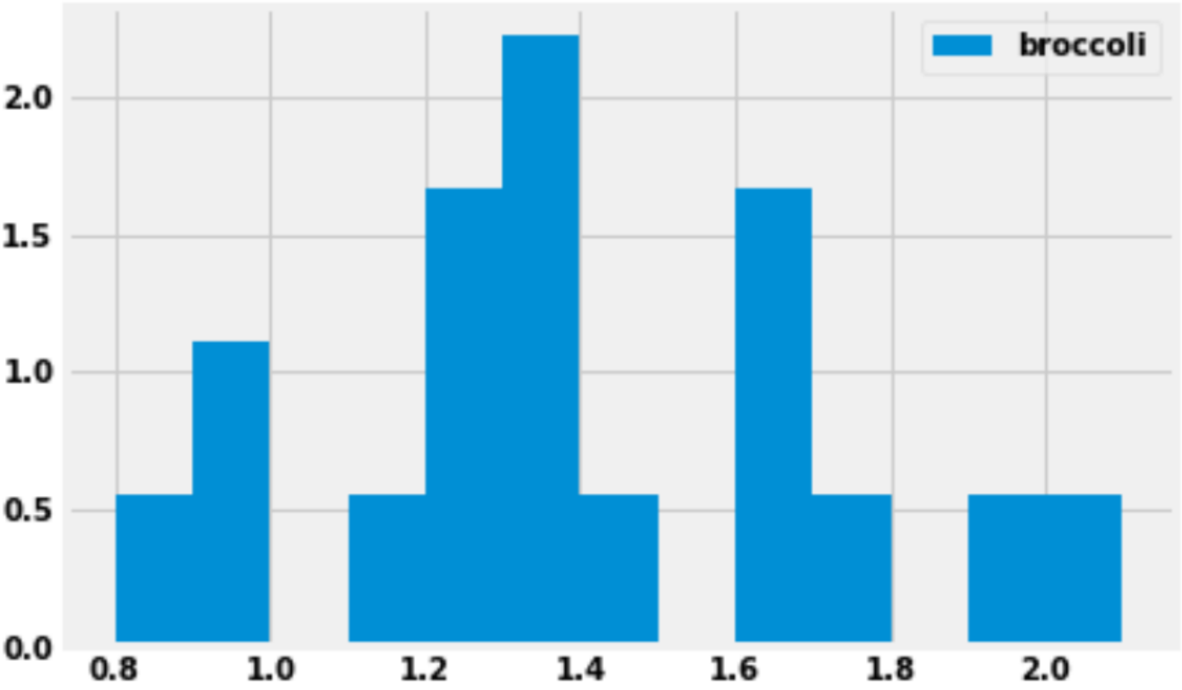
Q10.2 Histogram 1

4 Points

The code

```
prices.plot(kind='hist', y='broccoli', bins=np.arange(0.8, 2.11, 0.1), density=
```

produces the histogram below.



How many grocery stores sold broccoli for a price greater than or equal to $1.30 per pound, but less than $1.40 per pound (the tallest bar)? Input your answer below.

=4+-0

Q10.3 Histogram 2

4 Points

Suppose we now plot the same data with different bins, using the code

```
prices.plot(kind='hist', y='broccoli',
        bins=[0.8, 1, 1.1, 1.5, 1.8, 1.9, 2.5], density=True)
```
.

What would be the height on the y-axis for the bin corresponding to the interval $[\$1.10, \$1.50)$? Input your answer below.

=1.25+-0

Q10.4 Broccoli vs. Ice Cream 1

3 Points

You are interested in finding out the number of stores in which a pint of ice cream was cheaper than a pound of broccoli. Will you be able to determine the answer to this question by looking at the plot produced by the code below?

```
prices.get(['broccoli', 'ice_cream']).plot(kind='barh')
```

● Yes

○ No

Q10.5 Broccoli vs. Ice Cream 2

3 Points

You are interested in finding out the number of stores in which a pint of ice cream was cheaper than a pound of broccoli. Will you be able to determine the answer to this question by looking at the plot produced by the code below?

```
prices.get(['broccoli', 'ice_cream']).plot(kind='hist')
```

○ Yes

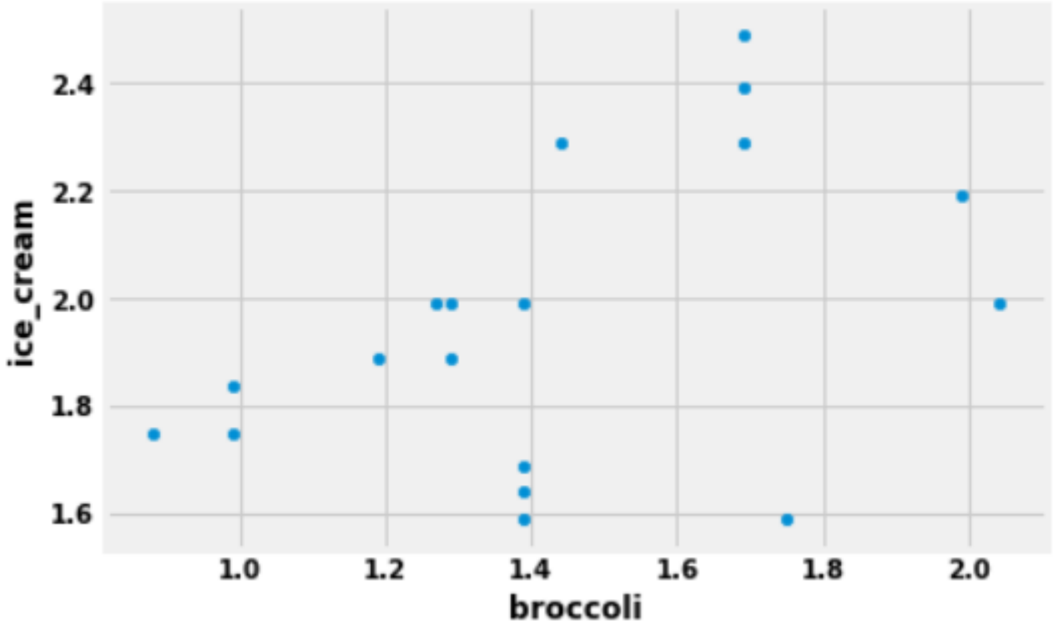
● No

Q10.6 Open-Ended: Broccoli vs. Ice Cream 3

4 Points

The scatterplot below was produced by the code

```
(prices.get(['broccoli', 'ice_cream'])
        .plot(kind='scatter', x='broccoli', y='ice_cream'))
```



Can you use this plot to figure out the number of stores in which a pint of ice cream was cheaper than a pound of broccoli?

If so, say how many such stores there are and explain how you came to that conclusion.

If not, explain why this scatterplot cannot be used to answer the question.

Q11 Random Number

12 Points

You generate a three-digit number by randomly choosing each digit to be a number 0 through 9, inclusive. Each digit is equally likely to be chosen.

Q11.1 Probability 1

4 Points

What is the probability you produce the number **027**? Input your answer below, as a number between 0 and 1 with no rounding.

=0.001+-0

Q11.2 Probability 2

4 Points

What is the probability you produce a number with an odd digit in the middle position? For example, **250**. Input your answer below, as a number between 0 and 1 with no rounding.

=0.5+-0

Q11.3 Probability 3

4 Points

What is the probability you produce a number with a **7** in it somewhere? Input your answer below, as a number between 0 and 1 with no rounding.

=0.271+-0

Q12 Results Array

4 Points

```
results = np.array([])
for i in np.arange(10):
    result = np.random.choice(np.arange(1000), replace=False)
    results = np.append(results, result)
```

After this code executes, `results` contains:

○ a simple random sample of size 9, chosen from a set of size 999 with replacement

○ a simple random sample of size 9, chosen from a set of size 999 without replacement

● a simple random sample of size 10, chosen from a set of size 1000 with replacement

○ a simple random sample of size 10, chosen from a set of size 1000 without replacement