Midterm Exam - DSC 10, Summer 2022

Instructions:

- You have 50 minutes to complete this exam.
- This exam consists of 13 questions, worth a total of 60 points.
- Write your name in the top right of each page in the space provided.
- Please write neatly in the provided answer boxes. We will not grade work that appears elsewhere.
- Completely fill in bubbles and square boxes.
  - ◯ A bubble means that you should only **select one choice**.
  - ☐ A square box means you should **select all that apply**.
- You may refer to the DSC 10 reference sheet and one hand-written page of notes. No other resources or technology (including calculators) are permitted.

Full Name:

PID:

Name of student to your **left**:

Name of student to your **right**:

(Write "N/A" if a wall/aisle is to your left/right.)

By signing below, you are agreeing that you will behave honestly and fairly during and after this exam.

Signature:

Do not open your exam until instructed to do so.

This page is intentionally left blank. You can use it as scratch paper.

Name: _____

Question 1 (1 point)

What is the result of: `df.shape[1]`?

◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 20 ◯ 100

Question 2 (2 points)

What is the result of: `df.shape[0]`?

◯ 3 ◯ 5 ◯ 20 ◯ 100 ◯ 150 ◯ Need more information

Question 3 (4 points)

The code below computes the **mean minutes late** (a `float`) using all the rows in the data.

```
df.__________.__________
       (a)        (b)
```

a) Fill in blank (a).

b) Fill in blank (b).

Question 4 (8 points)

The code below computes **the bus route of the latest bus on July 5th** (an `int`).

```
df[_________]._________._________._________
       (a)         (b)         (c)         (d)
```

a) Fill in blank (a).

b) Fill in blank (b).

c) Fill in blank (c).

d) Which of these can fill in blank (d)? **Select all that apply.**

☐ `loc[0]` ☐ `iloc[0]` ☐ `max()` ☐ `min()`

Question 5 (4 points)

Select all the expressions that evaluate to `True`. Note that `groupby` automatically sorts the index in ascending order.

☐ `df.groupby("day").count() == 20`

☐ `df.groupby("day").mean().shape[0] == 20`

☐ `df.groupby("day").min().index[0] == 0`

☐ `df.groupby("day").max().index[19] == 20`

☐ None of the above.

Name: _____

Question 6 (6 points)

We ran the following code:

```
mystery = df.groupby(["day", "route"]).max().reset_index()
```

Below are the first three rows of `mystery` (there are many more rows in the actual DataFrame).

	day	route	mins_late
0	1	30	2
1	1	201	5
2	2	30	3

a) Which of the following statements is true based on the first three rows of `mystery`?

◯ On July 1, the latest 30 bus was 2 minutes late.

◯ On July 1, at least one 30 bus was 5 minutes late.

◯ None of the above because we used `max()`.

◯ None of the above because we used `reset_index()`.

b) We can't determine the number of rows in `mystery` without more information about `df`. Which single additional piece of information will let us do so?

◯ The number of rows in `df`.

◯ The number of times each day appears in `df`.

◯ The number of unique bus routes in `df`.

◯ The chronological order of bus arrivals in `df`.

◯ None of the above.

Question 7 (2 points)

Which plot would most clearly show that some bus routes were very late on average while others were very punctual?

◯ Scatter plot ◯ Line plot ◯ Bar plot ◯ Histogram

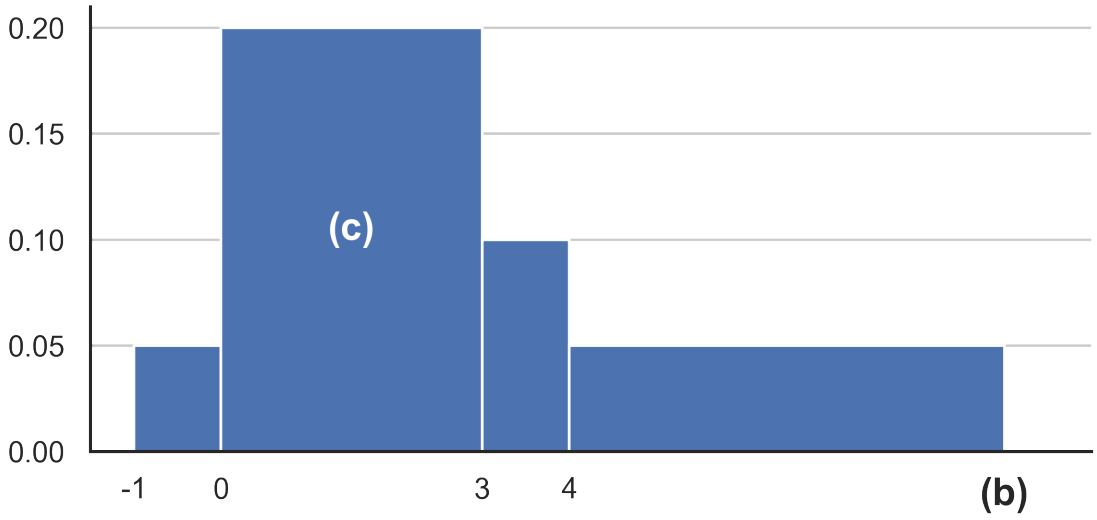
Question 8 (2 points)

Which plot would most clearly show that buses generally arrived earlier as July went on?

◯ Scatter plot ◯ Line plot ◯ Bar plot ◯ Histogram

Question 9 (6 points)

The density histogram below shows the distribution of the `"mins_late"` column in the `df` DataFrame. The x-axis units are in minutes. The y-axis units are in proportion per minute.



a) What percent of buses were between [3, 4) minutes late?

◯ 5% ◯ 10% ◯ 15% ◯ 20% ◯ 25% ◯ 30%

b) The x-axis label marked (b) is missing. What should it be?

◯ 6 ◯ 7 ◯ 8 ◯ 9 ◯ 10

c) Suppose that we split the bar marked (c) into two bars, one for the bin [0, 2) and one for the bin [2, 3). Using only the information in the histogram, complete the sentence:

The [0, 2) bar will have a height of _____.

◯ at least 0 and at most 0.15

◯ at least 0 and at most 0.2

◯ at least 0 and at most 0.3

◯ at least 0.2 and at most 0.3

◯ at least 0.2 and at most 0.4

◯ exactly 0.2

◯ exactly 0.3

◯ Can't tell based on information given.

Name: _____

Question 10 (8 points)

Let's say Sam wants to know: do buses arrive later on weekends? He has another DataFrame called `days` that has two columns:

- `"day" (int)`: the day in July (1-31).
- `"weekend" (bool)`: `True` if the day was a weekend, `False` if not.

a) Which of the follow expressions produces a DataFrame called `merged` that Sam can use to determine the answer? **Select all that apply.**

☐ `merged = df.merge(days, left_on="day", right_on="day")`

☐ `merged = df.merge(days, left_on="route", right_on="day")`

☐ `merged = days.merge(df, left_on="day", right_on="mins_late")`

☐ `merged = days.merge(df, left_on="weekend", right_on="mins_late")`

b) Sam uses `merged` to find that buses arrive 5.7 minutes later on weekdays compared to weekends on average. What can Sam conclude? **Select all that apply.**

☐ There's an association between being on a weekday and buses arriving late.

☐ Sam can expect the buses to be less late on weekends compared to weekdays.

☐ Buses arrive earlier on weekends because there are less people taking the bus.

☐ Buses arrive earlier on weekends because there are fewer buses running on weekends.

c) Suppose that the data in `df` was collected as follows: each time a bus arrives, Sam flips a coin. If it lands heads, he records down the arrival. Otherwise, he doesn't. Sam waits at the bus stop until he records 5 arrivals each day.

Again, Sam uses `merged` to find out that buses arrive 5.7 minutes later on weekdays compared to weekends. Can Sam conclude a causal relationship between weekends and bus arrival times?

◯ Yes ◯ No ◯ Can't say for sure

Question 11 (8 points)

Suppose that Sam wants to compute an "inconvenience" score for each bus arrival. Bigger scores mean more inconvenient. His scoring system works as follows:

- If the bus is ≥ 10 minutes late, the score is the square of the minutes late.
- If the bus is at least 1 and less than 10 minutes late, the score is 2 times the minutes late.
- If the bus is exactly on time, the score is 0.
- If the bus is early, the score is 1000. (Because Sam will definitely miss it.)

The `score` function below takes the minutes a bus was late as input (an `int`) and returns the inconvenience score (an `int`) for that bus arrival.

```
def score(mins):
    if ___(a)___:
        return mins * 2
    elif ___(b)___:
        return 1000
    else:
        return mins ** 2
```

a) What goes in blank (a)?

b) What goes in blank (b)?

c) The following line produces a scatter plot of the inconvenience scores, with the bus lateness on the x-axis and the score on the y-axis.

```
df.assign(_______).plot(kind='scatter', x='mins_late', y='inc')
```

What goes in the blank?

Name: _____

Question 12 (6 points)

The code below creates an array `all_scores` with the inconvenience scores for each arrival in `df`. Assume that the `score` function is implemented correctly.

```
all_scores = ___(a)___

for val in ___(b)___:
    s = ___(c)___
    all_scores = np.append(all_scores, s)
```

a) What goes in blank (a)?

b) What goes in blank (b)?

c) What goes in blank (c)?

Question 13 (3 points)

Suppose that the probability that a bus arrives late is p.

a) What's the probability that the first bus you see arrives late, but the second one is early?

◯ p^2 ◯ $p(1-p)$ ◯ $2p(1-p)$ ◯ $(1-p)^2$

b) If you take one bus every day in July, what's the probability that at least one arrives early? There are 31 days in July.

◯ p^{31} ◯ $1-p^{31}$ ◯ $(1-p)^{31}$ ◯ $1-(1-p)^{31}$

Before turning in your exam, please make sure that your PID or name is on every page.

Feel free to leave us a drawing here!