**Midterm Exam - DSC 10, Fall 2022**

---

**Instructions:**

- This exam consists of 9 questions, worth a total of 70 points.
- Write your PID in the top right corner of each page in the space provided.
- Please write **clearly** in the provided answer boxes; we will not grade work that appears elsewhere. Completely fill in bubbles and square boxes; if we cannot tell which option(s) you selected, you may lose points.

  ◯ 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 Reference Sheet that was provided to you. Other than that, you may not refer to any other resources or technology during the exam (including calculators).

---

Full Name:

PID:

Lecture Section: ◯ A00 (10-10:50AM) ◯ B00 (11-11:50AM) ◯ C00 (1-1:50PM) ◯ D00 (12-12:50PM)

Seat you are in:

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

Signature:

# Version B

Please do not open your exam until instructed to do so.

**Make sure you have read the data description attached to the Reference Sheet!**

## Question 1 (4 pts)

TritonTire, a UCSD alumni-run car repair shop, has an electronic database of all services it has ever provided, dating back to 1987. The owner, Dasha, says:

"*Our collective records from all 50 states show that, over the years, our repair shops have performed more and more repairs on EVs. After discussing this with our employees, we believe this is because newer EVs are not made to the same standards as the gasoline cars before them.*"

In **at most two sentences**, identify the largest confounding factor that Dasha has not considered in her argument. Your answer must fit in the box below.

PID: _____

## Question 2 (10 pts)

**a)** (2 pt) Which type of visualization should we use to visualize the distribution of `"Range"`?

◯ Bar chart ◯ Histogram ◯ Scatter plot ◯ Line plot

**b)** (4 pt) Teslas, on average, tend to have higher `"Range"`s than BMWs. In which of the following visualizations would we be able to see this pattern? Select all that apply.

☐ A bar chart that shows the distribution of `"Brand"`

☐ A bar chart that shows the average `"Range"` for each `"Brand"`

☐ An overlaid histogram showing the distribution of `"Range"` for each `"Brand"`

☐ A scatter plot with `"TopSpeed"` on the $x$-axis and `"Range"` on the $y$-axis

**c)** (4 pts) Gabriel thinks `"Seats"` is a categorical variable because it can be used to categorize EVs by size. For instance, EVs with 4 seats are small, EVs with 5 seats are medium, and EVs with 6 or more seats are large.

Is Gabriel correct? ◯ Yes ◯ No

Justify your answer in **one sentence**. Your answer must fit in the box below.

## Question 3 (8 pts)

Suppose we've run the following two lines of code.

```
first = evs.get("Brand").apply(max)
second = evs.get("Brand").max()
```

Note:

- The length of a value `v` is defined as `len(v)`, unless `v` is a DataFrame, in which case its length is `v.shape[0]`.
- If `s` is a string, then `max(s)` also evaluates to a string.

**a)** (4 pts) Fill in the blanks: `first` is a __(i)__ of length __(ii)__.

(i) ◯ list ◯ array ◯ string ◯ Series ◯ DataFrame

(ii) [ ]

**b)** (4 pts) Fill in the blanks: `second` is a __(i)__ of length __(ii)__.

(i) ◯ list ◯ array ◯ string ◯ Series ◯ DataFrame

(ii) [ ]

## Question 4 (6 pts)

Suppose we've run the following line of code.

```
counts = evs.groupby("Brand").count()
```

**a)** (3 pts) What value does `counts.get("Range").sum()` evaluate to?

[ ]

**b)** (3 pts) What value does `counts.index[3]` evaluate to?

[ ]

PID: _____

## Question 5 (11 pts)

Consider the following incomplete assignment statement.

```
result = evs______.mean()
```

In each part, fill in the blank above so that `result` evaluates to the specified quantity.

**a)** (3 pts) A DataFrame, indexed by `"Brand"`, whose `"Seats"` column contains the average number of `"Seats"` per `"Brand"`. (The DataFrame may have other columns in it as well.)

**b)** (4 pts) A number, corresponding to the average `"TopSpeed"` of all EVs manufactured by Audi in `evs`.

**c)** (4 pts) A number, corresponding to the average of the natural logarithm of the `"TopSpeed"` of all EVs in `evs`. *(Hint: The function* ***`np.log`*** *computes the natural logarithm of a single number.)*

## Question 6 (9 pts)

The DataFrame below shows the distribution of `"BodyStyle"` for all `"Brands"` in `evs`, other than Nissan. We will call EVs made by a `"Brand"` other than Nissan "non-Nissan EVs"; there are 24 non-Nissan EVs in `evs`.

| Brand | Hatchback | Liftback | Pickup | SUV | Sedan |
|---|---|---|---|---|---|
| Tesla | 0 | 2 | 3 | 4 | 3 |
| BMW | 1 | 0 | 0 | 1 | 1 |
| Audi | 0 | 0 | 0 | 8 | 1 |

Use the information above to answer the following questions.

**a)** (3 pts) Suppose we randomly select one of the non-Nissan EVs and it is either an SUV or a sedan. What is the most likely `"Brand"` of the randomly selected non-Nissan EV?

◯ Tesla ◯ BMW ◯ Audi

**b)** (3 pts) Suppose we randomly select two of the non-Nissan EVs without replacement. The probability that both are BMWs is equal to $\frac{1}{k}$, where $k$ is a positive integer. What is $k$?

◯ 8 ◯ 56 ◯ 64 ◯ 84 ◯ 92 ◯ 108

**c)** (3 pts) Suppose we randomly select one of the non-Nissan EVs and it is an SUV. What is the probability that it is made by Tesla? Give your answer as a simplified fraction.

PID: _____

## Question 7 (3 pts)

Below, we provide the same DataFrame as shown at the start of Question 6, which contains the distribution of `"BodyStyle"` for all `"Brands"` in `evs`, other than Nissan.

| Brand | Hatchback | Liftback | Pickup | SUV | Sedan |
|---|---|---|---|---|---|
| **Tesla** | 0 | 2 | 3 | 4 | 3 |
| **BMW** | 1 | 0 | 0 | 1 | 1 |
| **Audi** | 0 | 0 | 0 | 8 | 1 |

Suppose we've run the following few lines of code.

```
tesla = evs[evs.get("Brand") == "Tesla"]
bmw = evs[evs.get("Brand") == "BMW"]
audi = evs[evs.get("Brand") == "Audi"]

combo = tesla.merge(bmw, on="BodyStyle").merge(audi, on="BodyStyle")
```

How many rows does the DataFrame `combo` have?

◯ 21 ◯ 24 ◯ 35 ◯ 65 ◯ 72 ◯ 96

## Question 8 (9 pts)

TritonTrucks is an EV startup run by UCSD alumni. Their signature EV, the TritonTruck, has a subpar battery (the engineers didn't pay attention in their Chemistry courses).

A new TritonTruck's battery needs to be replaced after 500 days, unless it fails first, in which case it needs to be replaced immediately. On any given day, the probability that a given TritonTruck's battery fails is 0.02, independent of all other days.

Fill in the blanks so that `average_days_until_replacement` is an estimate of the **average number of days a new TritonTruck's battery lasts without needing to be replaced**.

```
def days_until_replacement(daily_conditions):
    days = 0
    for i in __(a)__:
        if daily_conditions[i] == True:
            __(b)__
        else:
            return days
    return days

total = 0
repetitions = 10000
for i in np.arange(repetitions):
    # The first element of the first argument to np.random.choice is
    # chosen with probability 0.98
    daily_conditions = np.random.choice(__(c)__, 500, p=[0.98, 0.02])
    total = total + days_until_replacement(daily_conditions)
average_days_until_replacement = total / repetitions
```

**a)** (3 pts) What goes in blank (a)?

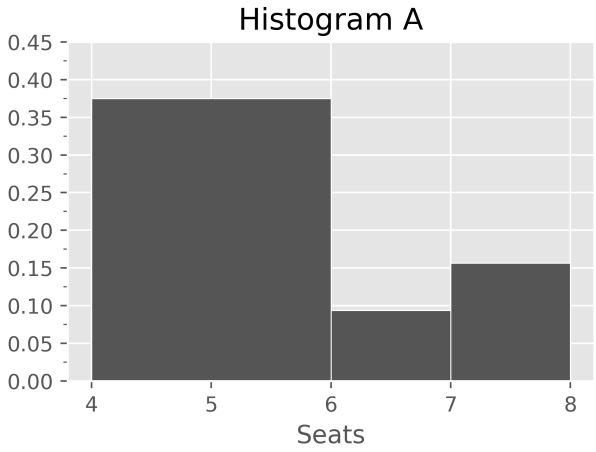
**b)** (3 pts) What goes in blank (b)?

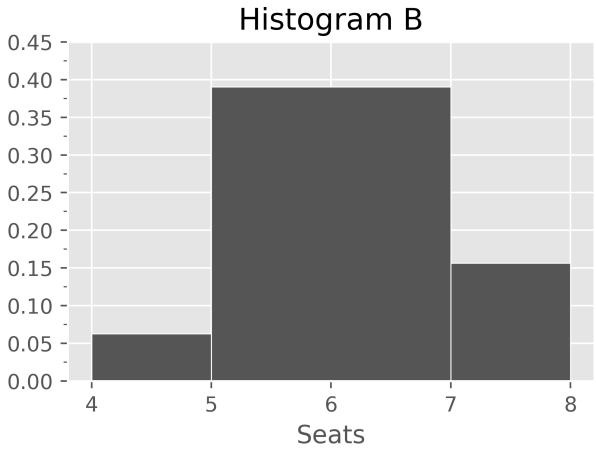
**c)** (3 pts) What goes in blank (c)?

PID: _____

## Question 9 (10 pts)

Histograms A and B below both display the distribution of the `"Seats"` column, using different bins. Each histogram includes all 32 rows of `evs`.




**a)** (3 pts) How many EVs in `evs` have exactly 6 seats?

**b)** (4 pts) How many EVs in `evs` have exactly 5 seats?

**c)** (3 pts) Histogram C (not pictured) also displays the distribution of the `"Seats"` column, but uses just a single bin, $[4, 9]$. What is the height of the sole bar in Histogram C?

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

## The Data

In September 2020, Governor Gavin Newsom announced that by 2035, all new vehicles sold in California must be zero-emissions vehicles. Electric vehicles (EVs) are among the most popular zero-emissions vehicles (though other examples include plug-in hybrids and hydrogen fuel cell vehicles).



The DataFrame `evs` consists of **32** rows, each of which contains information about a different EV model.

- `"Brand" (str)`: The vehicle's manufacturer.
- `"Model" (str)`: The vehicle's model name.
- `"BodyStyle" (str)`: The vehicle's body style.
- `"Seats" (int)`: The vehicle's number of seats.
- `"TopSpeed" (int)`: The vehicle's top speed, in kilometers per hour.
- `"Range" (int)`: The vehicle's range, or distance it can travel on a single charge, in kilometers.

The first few rows of `evs` are shown below (though remember, `evs` has 32 rows total).

| | Brand | Model | BodyStyle | Seats | TopSpeed | Range |
|---|---|---|---|---|---|---|
| 0 | Tesla | Model 3 Long Range Dual Motor | Sedan | 5 | 233 | 450 |
| 1 | BMW | iX3 | SUV | 5 | 180 | 360 |
| 2 | Tesla | Model 3 Standard Range Plus | Sedan | 5 | 225 | 310 |
| 3 | Audi | Q4 e-tron | SUV | 5 | 180 | 400 |
| 4 | Nissan | Leaf | Hatchback | 5 | 144 | 220 |

**Throughout this exam, we will refer to `evs` repeatedly.**

Assume that:

- The only four values in the `"Brand"` column are `"Tesla"`, `"BMW"`, `"Audi"`, and `"Nissan"`.
- We have already run `import babypandas as bpd` and `import numpy as np`.