DSC 80 Discussion 4 Worksheet

Name: _____

1 FA23 Midterm Problem 3

The `donkeys` table contains data from a research study about donkey health. The researchers measured the attributes of 544 donkeys. The next day, they selected 30 donkeys to reweigh. The first few rows of the `donkeys` table are shown below (left), and the table contains the following columns (right):

	id	BCS	Age	Weight	WeightAlt
0	d01	3.0	<2	77	NaN
1	d02	2.5	<2	100	NaN
2	d03	1.5	<2	74	NaN

`id`	A unique identifier for each donkey (`d01`, `d02`, etc.).
`BCS`	Body condition score: from 1 (emaciated) to 3 (healthy) to 5 (obese) in increments of 0.5.
`Age`	Age in years: <2, 2–5, 5–10, 10–15, 15–20, and over 20 years.
`Weight`	Weight in kilograms.
`WeightAlt`	Second weight measurement taken for 30 donkeys. NaN if the donkey was not reweighed.

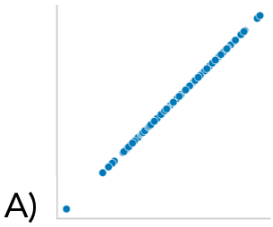
Consider the following scenarios for how the researchers chose the 30 donkeys to reweigh. In each scenario, select if the missing mechanism for the `WeightAlt` column is NMAR, MAR, or MCAR.

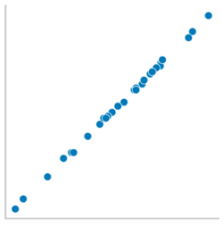
Note: Although the missing data are missing by design from the perspective of the original researchers, since we can't directly recover the missing values from our other data, we can treat the missing data as NMAR, MAR, or MCAR.

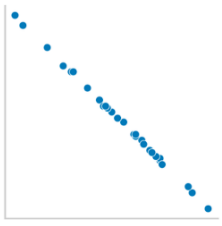
- A. The researchers chose the 30 donkeys with the largest `'Weight'` values to reweigh.
- B. The researchers drew 30 donkeys uniformly at random without replacement from the donkeys with `BCS` score of 44 or greater.
- C. The researchers set `i` as a number drawn uniformly at random between 0 and 514, then reweighed the donkeys in `donkeys.iloc[i:i+30]`.
- D. The researchers reweighed all the donkeys, but deleted all the values in `'WeightAlt'` except for the 30 lowest values.
- E. The researchers split up the donkeys into the 6 different age groups, then sampled 5 donkeys uniformly at random without replacement within each age group.

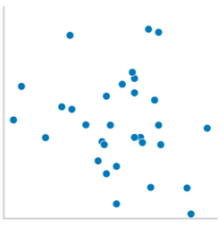
For this next question, assume that the researchers chose the 30 donkeys to reweigh by drawing **a simple random sample of 30 underweight donkeys: donkeys with BCS values of 1, 1.5, or 2.** The researchers weighed these 30 donkeys one day later and stored the results in `'WeightAlt'`.

Which of the following shows the scatter plot of `'WeightAlt' - 'Weight'` on the y-axis and `'Weight'` on the x-axis? Assume that missing values are not plotted.

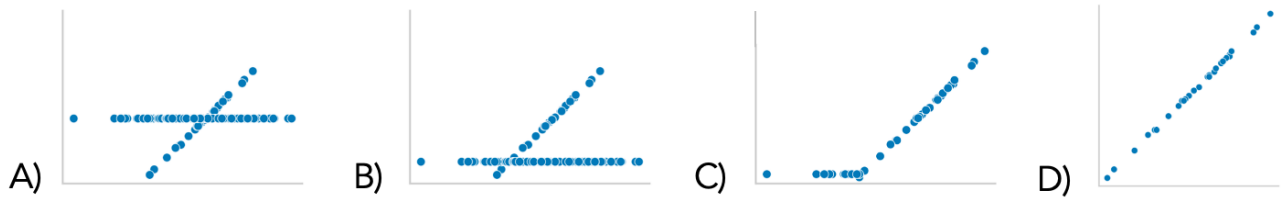
A) 

B) 

C) 

D) 

Suppose we use mean imputation to fill in the missing values in `'WeightAlt'`. Select the scatter plot `'WeightAlt'` on `'Weight'` after imputation.



2 FA23 Final Problem 3

The `bus` table (left) records bus arrivals over 1 day for all the bus stops within a 2 mile radius of UCSD. The data dictionary (right) describes each column.

	time	line	stop	late
0	12pm	201	Gilman Dr & Mandeville Ln	-1.1
1	1:15pm	30	Gilman Dr & Mandeville Ln	2.8
2	11:02am	101	Gilman Dr & Myers Dr	-0.8
3	8:04am	202	Gilman Dr & Myers Dr	NaN
4	9am	30	Gilman Dr & Myers Dr	-3.0

`time`	Time of arrival (`str`). Note that the times are inconsistently entered (e.g. `12pm` vs. `1:15pm`).
`line`	Bus line (`int`). There are multiple buses per bus line each day.
`stop`	Bus stop (`str`).
`late`	The number of minutes the bus arrived after its scheduled time. Negative numbers mean that the bus arrived early (`float`). Some entries in this column are missing.

For each of the following questions, select the correct procedure to simulate a single sample under the null hypothesis, and the correct test statistic for the hypothesis test. Assume that the `'time'` column of the `bus` DataFrame has already been parsed into timestamps.

Are buses equally likely to be early or late?

- [] `np.random.choice([-1, 1], bus.shape[0])`
- [] `np.random.choice(bus['late'], bus.shape[0], replace = True)`
- [] Randomly permute the `'late'` column

Test statistic:

[] Number of values below 0 [] np.mean [] np.std [] TVD [] K-S statistic

Is the `'late'` column MAR dependent on the `'line'` column?

- [] `np.random.choice([-1, 1], bus.shape[0])`
- [] `np.random.choice(bus['late'], bus.shape[0], replace = True)`
- [] Randomly permute the 'late' column

Test statistic:

[] Absolute difference in means [] Absolute difference in proportions [] TVD [] K-S statistic