DSC 80 Discussion 5 Worksheet

Name: ______________________________

1 FA23 Midterm Problem 4

In this question, we will continue to work with the `donkeys` dataset from Problem 3. The first few rows of the table column descriptions are shown again below for convenience.

	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.

Alan wants to see whether donkeys with `BCS` ≥ 3 have larger `Weight` values on average compared to donkeys that have `BCS` < 3. To generate a single sample under his null hypothesis, Alan should **(choose one)**:

- Resample 744 donkeys with replacement from donkeys.
- Resample 372 donkeys with replacement from donkeys with `BCS` < 3, and another 372 donkeys with `BCS` ≥ 3.
- Randomly permute the `Weight` column.

Doris wants to use multiple imputation to fill in missing values in `WeightAlt`. She knows that `WeightAlt` is MAR on `BCS` and `Age`, so she will perform multiple imputation conditional on `BCS` and `Age` – each missing value will be filled in with values from a random `WeightAlt` value from a donkey with the same `BCS` and `Age`. Assume that all `BCS` and `Age` combinations have observed `WeightAlt` values. Fill in the blanks in the code below to estimate the median of `WeightAlt` using multiple imputation conditional on `BCS` and `Age` with 100 repetitions.

```
def impute(col):
    col = col.copy()
    n = __________________
    fill = np.random.choice(____________________)
    col[__________________] = fill
    return col
results = []
for i in range(__________):
    imputed = (donkeys.__________(___________)['WeightAlt'
        ].____________(_____________))
    results.append(imputed.median())
```

2 WI23 Final Exam Problem 1

The DataFrame sat contains one row for ***most*** combinations of `Year` and `State`, where `Year` ranges between 2005 and 2015 and `State` is one of the 50 states (not including the District of Columbia). Assume `sat` does not contain any duplicate rows — that is, there is only one row for every unique combination of `Year` and `State` that is in `sat` – and that `sat` does not contain any null values.

	Year	State	# Students	Math	Verbal
0	2014	Washington	41277	519	510
1	2013	Arizona	22283	529	522
2	2006	Kansas	2545	591	582
3	2011	North Dakota	219	612	586
4	2009	New Mexico	2209	548	553

The data description stated that there is one row in `sat` for most combinations of `Year` (between 2005 and 2015, inclusive) and `State`. It turns out that there are 11 rows in sat for all 50 states, except for one state. Fill in the blanks below so that `missing_years` evaluates to an array, sorted in any order, containing the years for which that one state does not appear in `sat`.

```
state_only = sat.groupby("State").filter(_____________________)
merged = sat["Year"].value_counts().to_frame().merge(
    state_only,
   _______________________________________________________________
   )
missing_years =
   __________________________________________________________.
   to_numpy()
```

The following DataFrame contains summary statistics for all SAT takers in New York and Texas from 2005 to 2015. Suppose we want to run a statistical test to assess whether the distributions of the number of students between 2005 and 2015 in New York and Texas are significantly different.

	mean	median	std
State			
New York	157950.818182	157989.0	3430.986500
Texas	155035.909091	148102.0	22509.092685

Given the above DataFrame, which test statistic is **most likely** to yield a significant difference?

- A. mean number of students in Texas – mean number of students in New York
- B. |mean number of students in Texas – mean number of students in New York|
- C. |median number of students in Texas – median number of students in New York|
- D. The Kolmogorov-Smirnov statistic