DSC 80 Discussion 3 Worksheet

Name: ________________________

1 WI23 Midterm Problem 6

The DataFrame tv_excl (right) contains information about a group of TV shows, and DataFrame counts (left) contains the number of TV shows for *every* combination of "Age" and "Service" in tv_excl.

Service	Disney+	Hulu	Netflix	Prime Video
Age				
13+	NaN	4.0	2.0	1.0
16+	13.0	405.0	320.0	147.0
18+	NaN	223.0	445.0	134.0
7+	91.0	246.0	245.0	149.0
all	116.0	97.0	151.0	144.0

	Title	Year	Age	IMDb	Rotten Tomatoes	Service
0	Jersey Shore	2009	16+	3.6	54	Hulu
1	Henry Hugglemonster	2013	all	5.3	42	Disney+
2	Fast & Furious Spy Racers	2019	7+	5.5	62	Netflix
3	Atlanta	2016	18+	8.6	84	Hulu
4	Played	2013	NaN	6.4	45	Prime Video

Given the above information, what does the following expression evaluate to?
`tv_excl.groupby(["Age", "Service"]).sum().shape[0]`

Tiffany would like to compare the distribution of Age for Hulu and Netflix. Specifically, she'd like to test the following hypotheses:

- **Null Hypothesis:** The distributions of Age for Hulu and Netflix are drawn from the same population distribution, and any observed differences are due to random chance.
- **Alternative Hypothesis:** The distributions of Age for Hulu and Netflix are drawn from different population distributions.

Is this a hypothesis test, or a permutation test? Why?

Consider the DataFrame distr, defined below.

```
hn = counts[["Hulu", "Netflix"]]
distr = (hn / hn.sum()).T  # Note that distr has 2 rows and 5 columns.
```

To test the hypotheses above, Tiffany decides to use the total variation distance as her test statistic. Which of the following expressions **DO NOT** correctly compute the observed statistic for her test?

```
[] distr.diff().iloc[-1].abs().sum() / 2
[] distr.diff().sum().abs().sum() / 2
[] distr.diff().sum().sum().abs() / 2
[] (distr.sum() - 2 * distr.iloc[0]).abs().sum() / 2
[] distr.diff().abs().sum(axis=1).iloc[-1] / 2
```

2 WI23 Final Problem 2.5

Suppose $\vec{a} = \begin{bmatrix} a_1 & a_2 & \ldots & a_n \end{bmatrix}^T$ and $\vec{b} = \begin{bmatrix} b_1 & b_2 & \ldots & b_n \end{bmatrix}^T$ are both vectors containing proportions that add to 1. As we've seen before, the TVD is defined as follows:

$$\text{TVD}(\vec{a}, \vec{b}) = \frac{1}{2} \sum_{i=1}^{n} |a_i - b_i|$$

The TVD is not the only metric that can quantify the distance between two categorical distributions. Here are three other possible distance metrics:

- $\text{dis1}(\vec{a}, \vec{b}) = \vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n$
- $\text{dis2}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} = \frac{a_1 b_1 + a_2 b_2 + \ldots + a_n b_n}{\sqrt{a_1^2 + a_2^2 + \ldots + a_n^2}\sqrt{b_1^2 + b_2^2 + \ldots + b_n^2}}$
- $\text{dis3}(\vec{a}, \vec{b}) = 1 - \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$

Of the above three possible distance metrics, only one of them has the same range as the TVD (i.e. the same minimum possible value and the same maximum possible value) *and* has the property that smaller values correspond to more similar vectors. Which distance metric is it?

[] dis1 [] dis2 [] dis3

3 FA23 Midterm Problem 4

In this question, we will work with the DataFrame `donkeys`, about the health of various donkeys. (Don't worry about the `WeightAlt` column for now.)

	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. Select all the possible test statistics that Alan could use to conduct this hypothesis test. Let μ_1 be the mean weight of donkeys with `BCS` ≥ 3 and μ_2 be the mean weight of donkeys with `BCS` < 3.

[] μ_1 [] $\mu_1 - \mu_2$ [] $2\mu_2 - \mu_1$ [] $|\mu_1 - \mu_2|$ [] Total variation distance