DSC 80 Discussion 2 Worksheet

1 WI24 Midterm Problem 3

Jasmine is a veterinarian. Below, you'll find information about some of the dogs in her care, separated by district and breed.

	Beagle		Cocker Spaniel	
	Mean Weight	**Count**	**Mean Weight**	**Count**
District 1	25	3	20	2
District 2	45	1	x	y

What is the mean weight of all beagles in the table above, across both districts?

Notice that the table above has two unknowns, x and y. Find **positive integers** x and y such that the mean weight of all beagles is equal to the mean weight of all cocker spaniels, *where x is as small as possible.*

2 FA23 Midterm Problem 1

	date	name	food	weight
0	2023-01-01	Sam	Ribeye	0.20
1	2023-01-01	Sam	Pinto beans	0.10
2	2023-01-01	Lauren	Mung beans	0.25
3	2023-01-02	Lauren	Lima beans	0.30
4	2023-01-02	Sam	Sirloin	0.30

Find the total kg of food eaten for each day and each person in `df` as a Series.

```
df.groupby(_____)[_____].sum()
```

Find all the unique people who did not eat any food containing the word "beans".

```
def foo(x):
    return _______________________________


df.groupby(_____).____________________(foo)['name'].unique()
```

3 FA23 Final Problem 1

The `bus` table (left) records bus arrivals over 11 day for all the bus stops within a 22 mile radius of UCSD.

	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.

The `stop` table (left) contains information for all the bus lines in San Diego (not just the ones near UCSD).

	line	stop	next
0	201	Gilman Dr & Mandeville Ln	VA Hospital
1	201	VA Hospital	La Jolla Village Dr & Lebon Dr
2	30	VA Hospital	Villa La Jolla Dr & Holiday Ct
3	30	UTC	NaN

`line` Bus line (`int`).
`stop` Bus stop (`str`).
`next` The next bus stop for a particular bus line (`str`). For example, the first row of the table shows that after the `201` stops at `Gilman Dr & Mandeville Ln`, it will stop at the `VA Hospital` next. A missing value represents the end of a line.

Compute the number of buses in `bus` whose next stop is `'UTC'`.

```
x = stop.merge(__________, on = _____________________________, how = _________________)

x[_____________________________________________].shape[0]
```

Compute the number of unique pairs of bus stops that are exactly two stops away from each other. For example, if you only use the first four rows of the `stop` table, then your code should evaluate to the number 2, since you can go from `'Gilman Dr & Mandeville Ln'` to `'La Jolla Village Dr & Lebon Dr'` and from `'Gilman Dr & Mandeville Ln'` to `'Villa La Jolla Dr & Holiday Ct'` in two stops. Hint: The `suffixes = (1, 2)` argument to merge appends a `1` to column labels in the left table and a `2` to column labels in the right table whenever the merged tables share column labels.

```
m = __________.merge(__________, left_on = __________, right_on = __________,
      how = ____________, suffixes=(1, 2))
(m[_____________________________].drop_duplicates().shape[0])
```

4 FA22 Midterm Problem 7

	category	completed	minutes	urgency	client
0	work	False	NaN	2.0	NaN
1	work	False	NaN	1.0	NaN
2	work	True	13.5	2.0	NaN
3	work	False	NaN	1.0	NaN
4	relationship	True	5.3	NaN	NaN
...	...	...	...	...	...
9831	consulting	True	71.7	2.0	San Diego Financial Analysts
9832	finance	True	36.4	1.0	NaN
9833	work	True	31.1	1.0	NaN
9834	work	True	24.8	3.0	NaN
9835	work	False	NaN	2.0	NaN

The code below creates a pivot table.

```
pt = tasks.pivot_table(index='urgency', columns='category', values='completed', aggfunc='sum')
```

Which of the below snippets of code will produce the same result as `pt.loc[3.0, 'consulting']`? **Select all that apply.**

☐ Snippet 1:

```
tasks[(tasks['category'] == 'consulting') & (tasks['urgency'] == 3.0)]['completed'].sum()
```

☐ Snippet 2:

```
tasks[tasks['urgency'] == 3].groupby('category')['completed'].sum().loc['consulting']
```

☐ Snippet 3:

```
tasks.groupby('urgency')['completed'].sum().loc[3.0, 'consulting']
```

☐ Snippet 4:

```
tasks.groupby(['urgency', 'category'])['completed'].sum().loc[(3.0, 'consulting')]
```

☐ Snippet 5

```
tasks.groupby('completed').sum().loc[(3.0, 'consulting')]
```

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 $\vec{a} = \begin{bmatrix} a_1 & a_2 & ... & a_n \end{bmatrix}^T$ and $\vec{b} = \begin{bmatrix} b_1 & b_2 & ... & b_n \end{bmatrix}^T$ 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:

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

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