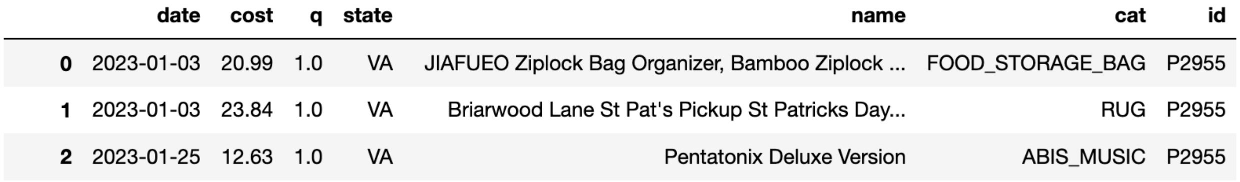
Final Exam Dataframe Reference Sheet - DSC 80, Spring 2024

This page contains information about the dataframes that you will use on the exam. **Only the first few rows are shown for each table.**

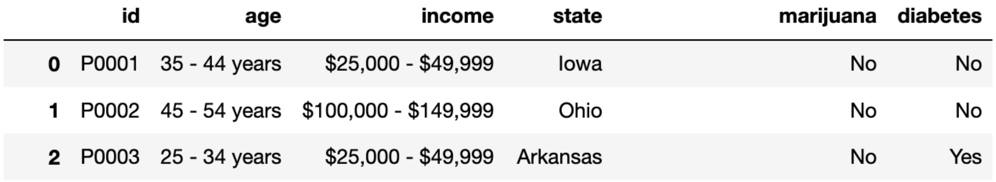
The Open e-commerce dataset contains data about people's Amazon.com purchases[1]. To collect the data, the researchers asked participants to fill out a survey. Only participants who completed the survey were recorded in the data. In one part of the survey, participants were given instructions to download their Amazon purchase history and share it with the researchers. Since this step was not required, not all participants shared their Amazon purchase history with the researchers. The dataset contains two tables, `df` and `survey`. The `df` table was created from participants' purchase history and records individual items purchased from Amazon. The data dictionary below describes each column.



	date	cost	q	state	name	cat	id
0	2023-01-03	20.99	1.0	VA	JIAFUEO Ziplock Bag Organizer, Bamboo Ziplock ...	FOOD_STORAGE_BAG	P2955
1	2023-01-03	23.84	1.0	VA	Briarwood Lane St Pat's Pickup St Patricks Day...	RUG	P2955
2	2023-01-25	12.63	1.0	VA	Pentatonix Deluxe Version	ABIS_MUSIC	P2955

`date` The date that the purchase was made (`pd.Timestamp`).
`code` Cost of one item in US dollars (`float`)
`q` Quantity of items purchased in the order (`float`).
`state` US state where order was shipped (`str`). If the item was a electronic gift card, the researchers recorded `NaN`.
`name` Name of item (`str`).
`cat` Category of item (`str`).
`id` Participant ID (`str`).

The `survey` table records the survey results. The data dictionary below describes each column.



	id	age	income	state	marijuana	diabetes
0	P0001	35 - 44 years	$25,000 - $49,999	Iowa	No	No
1	P0002	45 - 54 years	$100,000 - $149,999	Ohio	No	No
2	P0003	25 - 34 years	$25,000 - $49,999	Arkansas	No	Yes

`id` Participant ID (`str`). Some values in this column don't appear in the `id` column of `df`.
`age` Age of participant (`str`).
`income` Income of participant (`str`).
`state` US state where order was shipped (`str`).
`marijuana` Whether the participant reported that they smoke marijuana (Yes) or don't smoke marijuana (No). (`str`).
`diabetes` Whether the participant reported that they have diabetes (Yes) or don't have diabetes (No). (`str`).

[1] The original dataset was just published on May 13, 2024 and would make a nice dataset for an independent data analysis project! We will use a slightly modified version of this data for this exam. The full citation is:
Berke, A., Calacci, D., Mahari, R. et al. Open e-commerce 1.0, five years of crowdsourced U.S. Amazon purchase histories with user demographics. Sci Data 11, 491 (2024). https://doi.org/10.1038/s41597-024-03329-6