Mamma mia, a *pizza* visualization 🤌🇮🇹

Dataset

Find a dataset (or multiple) that you will explore. Assess the quality of the data it contains and how much preprocessing / data-cleaning it will require before tackling visualization. We recommend using a standard dataset as this course is not about scraping nor data processing.
We have found 3 different datasets that we will be using:

- The main dataset contains data about pizza restaurants in the USA that contains precise locations, price ranges and menus for each restaurant. (here is the metadata that we care about for this dataset:
- We found another dataset for the pizzas. This one contains a set of pizza orders to a fictional restaurant that includes the name of the pizza, the ingredients and the price. Since it is fictional, we won't use it to extract actual information ; however, since there isn't always data for the pizza ingredients in the other pizza dataset, we may use the pizza ingredients for a given pizza as defined in this dataset.
- The last dataset is the cost of living all around the world such as rent, restaurants, groceries etc. by city. We will keep only the cities in the USA from this dataset. We plan to use this data to see the correlation between pizza prices and the cost of living. (here is the metadata for this dataset :

Thankfully these datasets do not require a lot of preparing (ingredients).

Problematic

Frame the general topic of your visualization and the main axis that you want to develop.
• What am I trying to show with my visualization?
• Think of an overview for the project, your motivation, and the target audience.

The main goal is to show the readers insightful data about all the different pizzas in the United States and how it is related to the living cost of each region, through the analysis of pizza restaurants across all the states.

In our first visualization, we want to show a Pizza Index that would work the same way as the Big Mac Index (which means that we want to correlate pizza prices to the cost of living in the different US states).

In our second main visualization, we want to isolate pizza specialities or original creations by states/cities. This way, the audience could build their own pizza and could easily find where they could buy a pizza with those particularities. The goal would be to have a nice interaction where the user selects and drags/drops ingredients and cooking methods onto the pizza ; the website would then show possible matching pizzas all over the United States. We could also have a showcase of some of the most interesting American pizza variants and a bit of description for them.

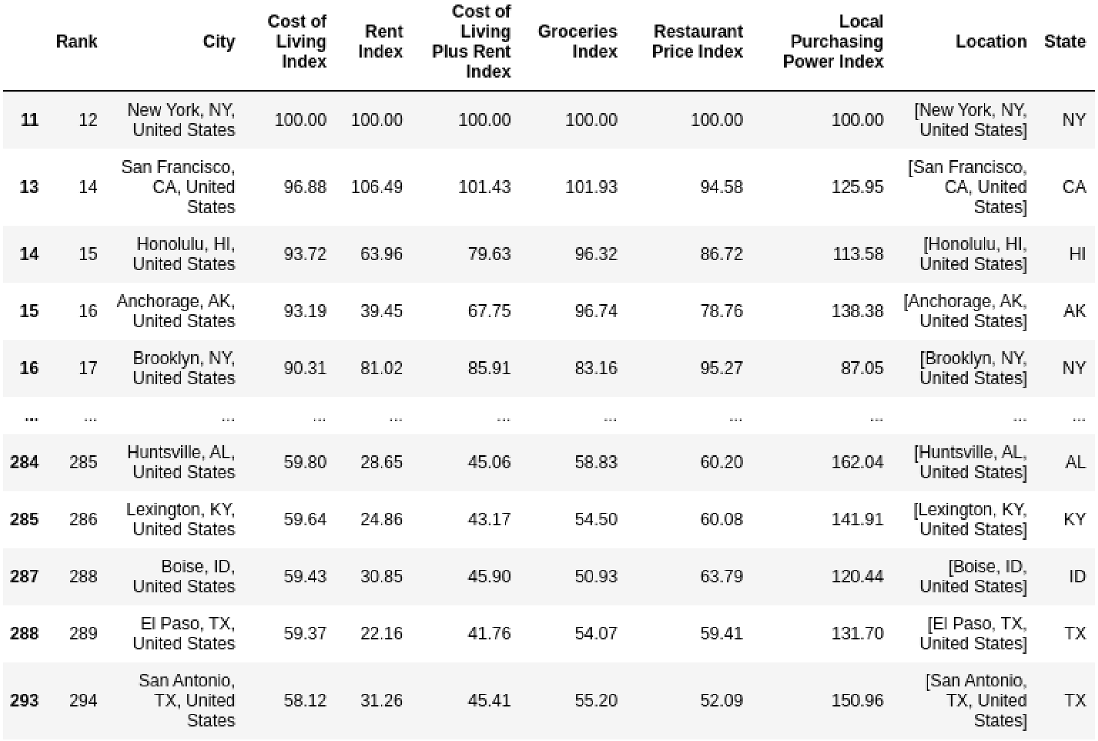
We find the topic fun and lighthearted and at the same time, we see the potential it has for more serious economy-related questions like the pizza index.

Our target audience would be curious-minded people who seek to know more about pizzas in the United States and that are interested in the link between cost of life and pizza price. We also want to keep visitors invested in the website with a fun game of finding the best restaurant that would accommodate their needs of flavors and ingredients (for a pizza of course).

Exploratory Data Analysis

Cost of living dataset :

After the preprocessing step that includes isolating the United States cities (using the second element of [City]) and creating the [State] section to facilitate the merge with other datasets, we obtain a dataset containing 110 rows (=cities) and 10 columns :

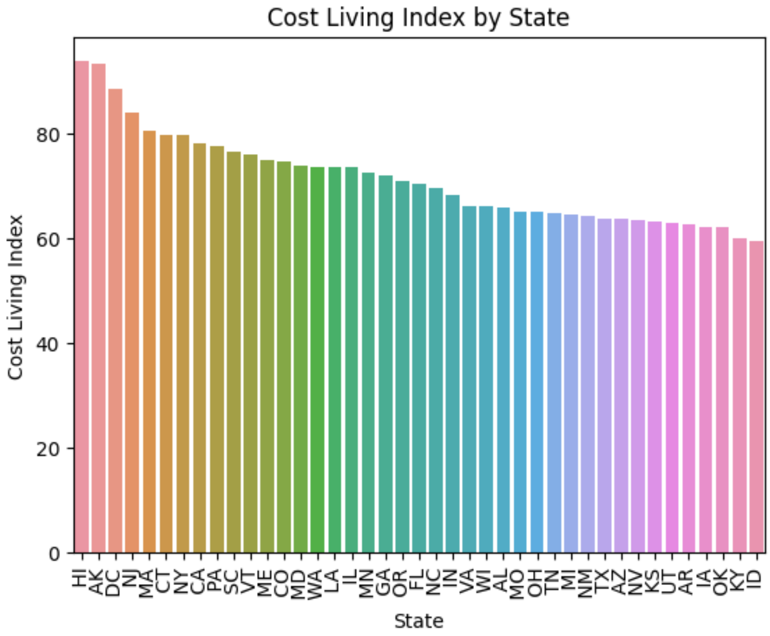


	Rank	City	Cost of Living Index	Rent Index	Cost of Living Plus Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index	Location	State
11	12	New York, NY, United States	100.00	100.00	100.00	100.00	100.00	100.00	[New York, NY, United States]	NY
13	14	San Francisco, CA, United States	96.88	106.49	101.43	101.93	94.58	125.95	[San Francisco, CA, United States]	CA
14	15	Honolulu, HI, United States	93.72	63.96	79.63	96.32	86.72	113.58	[Honolulu, HI, United States]	HI
15	16	Anchorage, AK, United States	93.19	39.45	67.75	96.74	78.76	138.38	[Anchorage, AK, United States]	AK
16	17	Brooklyn, NY, United States	90.31	81.02	85.91	83.16	95.27	87.05	[Brooklyn, NY, United States]	NY
...	...	...	...	...	...	...	...	...	...	...
284	285	Huntsville, AL, United States	59.80	28.65	45.06	58.83	60.20	162.04	[Huntsville, AL, United States]	AL
285	286	Lexington, KY, United States	59.64	24.86	43.17	54.50	60.08	141.91	[Lexington, KY, United States]	KY
287	288	Boise, ID, United States	59.43	30.85	45.90	50.93	63.79	120.44	[Boise, ID, United States]	ID
288	289	El Paso, TX, United States	59.37	22.16	41.76	54.07	59.41	131.70	[El Paso, TX, United States]	TX
293	294	San Antonio, TX, United States	58.12	31.26	45.41	55.20	52.09	150.96	[San Antonio, TX, United States]	TX

We can note that this dataset is centered around NewYork which means that NewYork is the central reference and all its attributes are *100*. This way, a Rent Index that is *80* is cheaper than the one in NewYork.

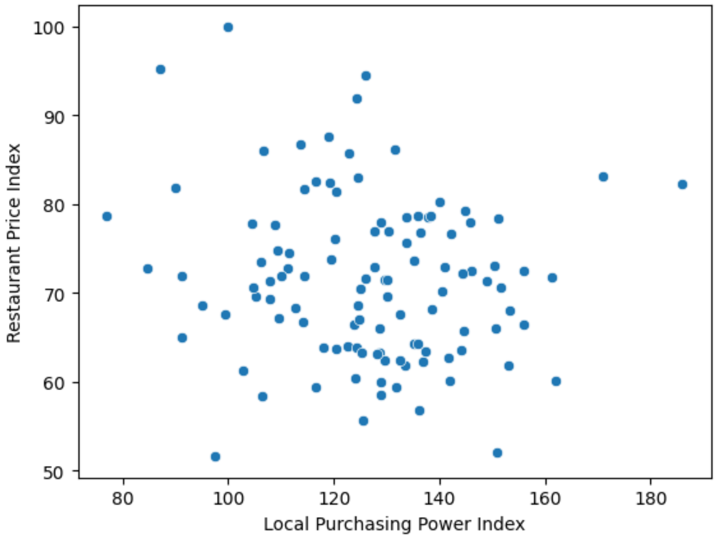
The mean of the Local Purchasing Power Index is 126.72 which is above 100, showing that the Purchasing power is in general higher than in NewYork.

Here is the mean cost of living by state : Note that Hawaii and (surprisingly) Alaska are the most cost living states.



Let's compare some attributes :
We can note that the pearson correlation between the Rent Index and the Cost of Living Index is of 0.80 and the p-value is very small (<< 0.05)
and of 0.70 with a p-value << 0.05 also, between the Groceries Index and the Restaurant Price Index, which makes a lot of sense.

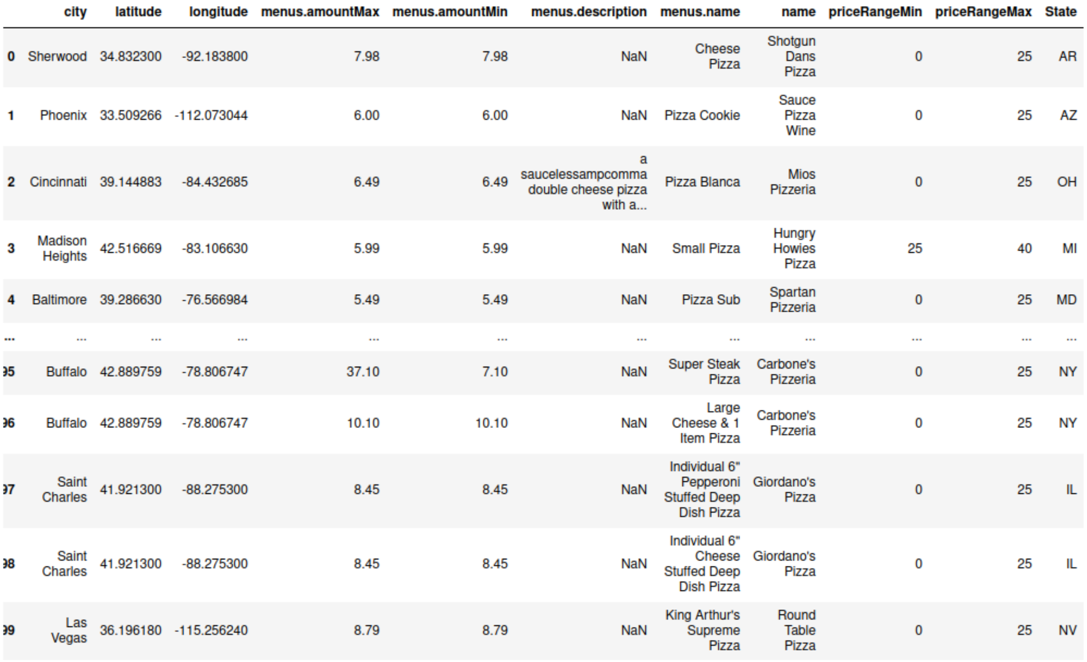
A little bit less subtle, let's try to see how the Local Purchasing Power Index behaves compared to other attributes.



We can see that it is much harder to conclude anything based on this and the pearson correlation which is of -0.12 with a high p-value (=0.176…).

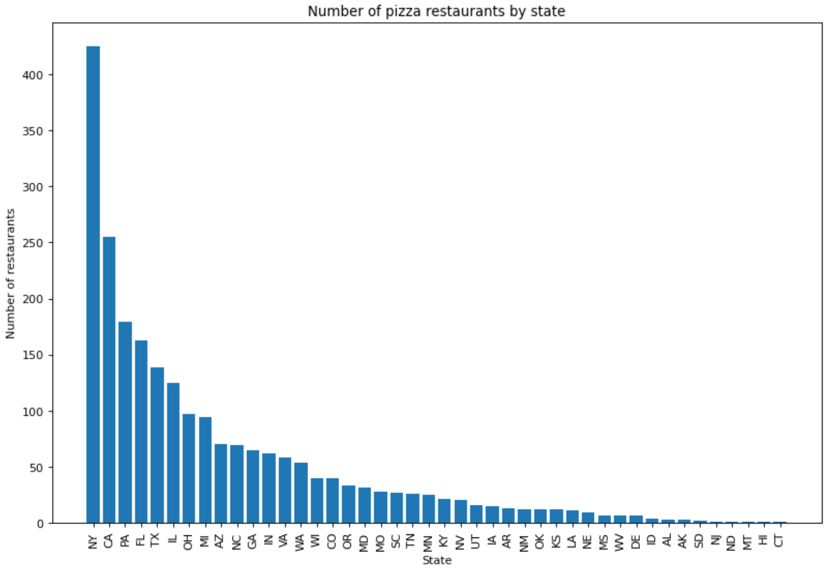
This would require some deeper analysis that we may conduct later on the project

Pizza dataset :
This dataset has 10000 lines and 24 columns. It provides menu information for 2285 different restaurants in 1028 different cities in 44 of the 50 states. We chose to keep the following columns : [city, latitude, longitude, menus_amountMax, menus_amountMin (price of the pizza), menus_description, menus_name, name (name of the restaurant), priceRangeMin, priceRangeMax (price range of the restaurant), state] :



	city	latitude	longitude	menus.amountMax	menus.amountMin	menus.description	menus.name	name	priceRangeMin	priceRangeMax	State
0	Sherwood	34.832300	-92.183800	7.98	7.98	NaN	Cheese Pizza	Shotgun Dans Pizza	0	25	AR
1	Phoenix	33.509266	-112.073044	6.00	6.00	NaN	Pizza Cookie	Sauce Pizza Wine	0	25	AZ
2	Cincinnati	39.144883	-84.432685	6.49	6.49	a saucelessampcomma double cheese pizza with a...	Pizza Blanca	Mios Pizzeria	0	25	OH
3	Madison Heights	42.516669	-83.106630	5.99	5.99	NaN	Small Pizza	Hungry Howies Pizza	25	40	MI
4	Baltimore	39.286630	-76.566984	5.49	5.49	NaN	Pizza Sub	Spartan Pizzeria	0	25	MD
...	...	...	...	...	...	...	...	...	...	...	...
95	Buffalo	42.889759	-78.806747	37.10	7.10	NaN	Super Steak Pizza	Carbone's Pizzeria	0	25	NY
96	Buffalo	42.889759	-78.806747	10.10	10.10	NaN	Large Cheese & 1 Item Pizza	Carbone's Pizzeria	0	25	NY
97	Saint Charles	41.921300	-88.275300	8.45	8.45	NaN	Individual 6" Pepperoni Stuffed Deep Dish Pizza	Giordano's Pizza	0	25	IL
98	Saint Charles	41.921300	-88.275300	8.45	8.45	NaN	Individual 6" Cheese Stuffed Deep Dish Pizza	Giordano's Pizza	0	25	IL
99	Las Vegas	36.196180	-115.256240	8.79	8.79	NaN	King Arthur's Supreme Pizza	Round Table Pizza	0	25	NV

Below is the distribution of the number of restaurants by state :



The price of pizzas ranges from 0.25 dollars to 243 dollars.

2nd Pizza dataset :

This dataset has 48k lines and it has the following metadata : [order_details_id, order_id, pizza_id, quantity, order_date, order_time, unit_price, total_price, pizza_size, pizza_category, pizza_ingredients, pizza_name]

As we said before we might only use this one if we need to map pizza_name to pizza ingredients.

Related work

We took some inspiration from the concept of the Big Mac index. This is a concept used in economics to quantify the cost of living in a given country by how much the McDonald's Big Mac costs in that country. There have been many studies on this ; here is one of them. It compares the calculated cost of living with the cost of the Big Mac to see whether the correlation is as strict as is sometimes said. Our approach differs for several reasons. First, we are looking only at the USA, so our geographical entities will be regional rather than national. Second, we are using a more abstract value than Big Mac cost : since we are looking at Pizza cost by restaurant, this will also change within cities based on the quality of the restaurant. Finally, we are exploring more data than just the cost : we are looking at the types of available pizzas by region, pizza ingredients and more.

There are also studies on the pizza industry in the USA, like this one. However, these are often quite boring, don't have much of a visualization aspect and don't go much into detail. We would improve on all these points.

Visually, we would like to create a data story that takes the user on a metaphorical journey, using the data. A big source of inspiration for this type of visualization was the Wine101 project from last year's Data Visualization course. This website is very interactive, which makes it a lot more fun than just looking at data representations. We were specifically interested in using a similar presentation template (as opposed to the animations).

We also found this representation of regional foods all over the world, represented as a map. It is quite buggy and slow, but it allows you to navigate from more to less detail by zooming in and out of certain regions. This may be an idea that we would incorporate into our project.