ETL Project Plan - Pizza!

Collaborators: Meera Shah and Monte Rodriguez

Datasets to Extract, Transform, and Load:

Monte:

- https://github.com/angelddaz/pizza_delivery/blob/master/RawDelData.csv
- https://data.world/sdhilip/pizza-datasets

Meera:

- https://www.kaggle.com/datafiniti/pizza-restaurants-and-the-pizza-they-sell
- https://data.world/makeovermonday/2020w13-does-pineapple-belong-on-a-pizza

Questions for Analysis (Bonus):

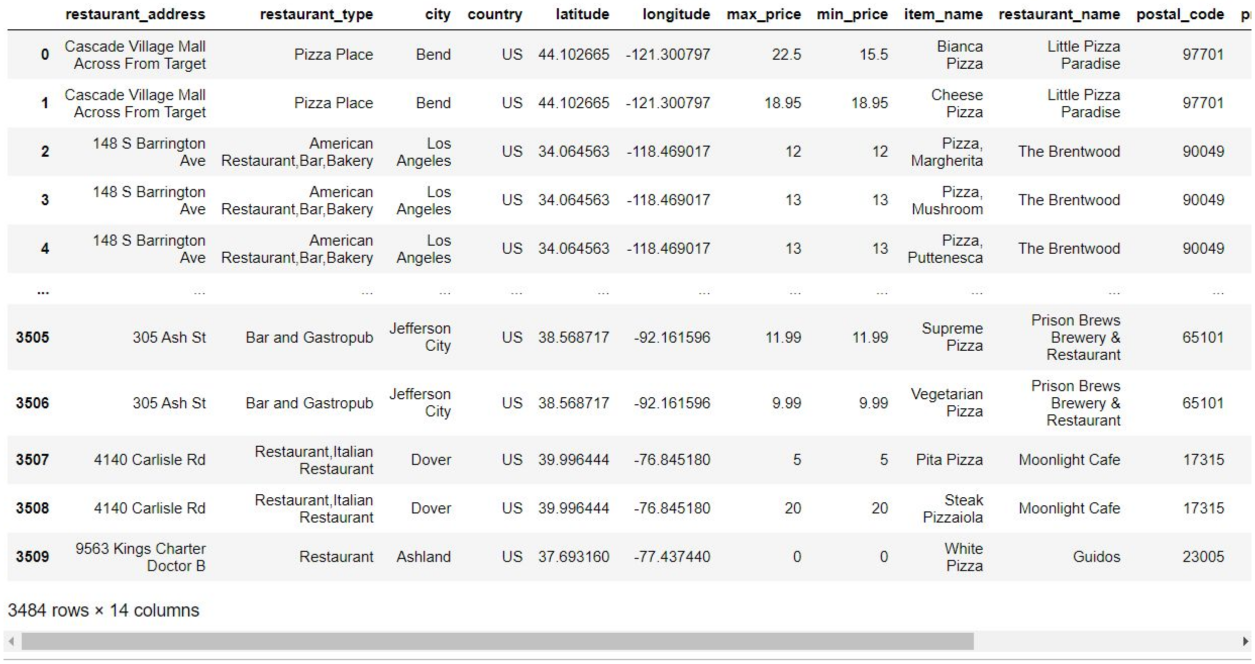
- Which pizza places are visited the most?
- Which restaurants are the most affordable?
- Which toppings are the most popular?
- How many people said pizza was their favorite?
- Which state has the most pizza places?
- The different types of pizza (deep dish, NY-style, etc)?
- Which location has which types of pizzas, and which has the most different types?
- Which company has the most businesses nationwide?
- How many people prefer pineapple?
- Cold pizza vs hot pizza - Which is more favorable?

ETL Project Report
Submitted by: Meera Shah and Monte Rodriguez

1. **Extract:**
    a. In this project, we looked at the popularity of pizza and the different restaurants that served pizza throughout the United States. We found some datasets with restaurant information, topping popularity information, nutritional information, and delivery information. Because the timeframe of this project was short, we prioritized the locations and price ranges of each restaurant.

    b. We began by importing all data into our Jupyter notebooks and renaming the columns headers so that they were more descriptive of the information that they provided.
        - The pizza delivery data only had data for one state and our other datasets were on a national level.
        - The nutritional facts data did not provide enough detail with regards to the brands of pizza. Instead, the brands were assigned random letters, which was too vague to include in our database.
        - After discussing the above, we agreed as a team to exclude these datasets since we already had four other datasets to work.

Clean Restaurant Dataframe:



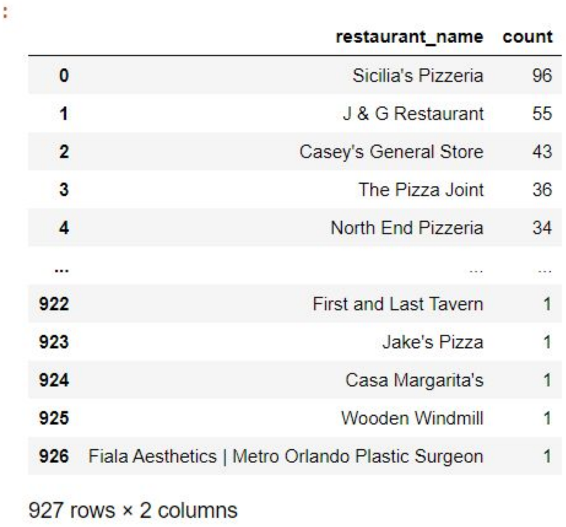
	restaurant_address	restaurant_type	city	country	latitude	longitude	max_price	min_price	item_name	restaurant_name	postal_code	p
0	Cascade Village Mall Across From Target	Pizza Place	Bend	US	44.102665	-121.300797	22.5	15.5	Bianca Pizza	Little Pizza Paradise	97701	
1	Cascade Village Mall Across From Target	Pizza Place	Bend	US	44.102665	-121.300797	18.95	18.95	Cheese Pizza	Little Pizza Paradise	97701	
2	148 S Barrington Ave	American Restaurant,Bar,Bakery	Los Angeles	US	34.064563	-118.469017	12	12	Pizza, Margherita	The Brentwood	90049	
3	148 S Barrington Ave	American Restaurant,Bar,Bakery	Los Angeles	US	34.064563	-118.469017	13	13	Pizza, Mushroom	The Brentwood	90049	
4	148 S Barrington Ave	American Restaurant,Bar,Bakery	Los Angeles	US	34.064563	-118.469017	13	13	Pizza, Puttenesca	The Brentwood	90049	
...	...	...	...	...	...	...	...	...	...	...	...	
3505	305 Ash St	Bar and Gastropub	Jefferson City	US	38.568717	-92.161596	11.99	11.99	Supreme Pizza	Prison Brews Brewery & Restaurant	65101	
3506	305 Ash St	Bar and Gastropub	Jefferson City	US	38.568717	-92.161596	9.99	9.99	Vegetarian Pizza	Prison Brews Brewery & Restaurant	65101	
3507	4140 Carlisle Rd	Restaurant,Italian Restaurant	Dover	US	39.996444	-76.845180	5	5	Pita Pizza	Moonlight Cafe	17315	
3508	4140 Carlisle Rd	Restaurant,Italian Restaurant	Dover	US	39.996444	-76.845180	20	20	Steak Pizzaiola	Moonlight Cafe	17315	
3509	9563 Kings Charter Doctor B	Restaurant	Ashland	US	37.693160	-77.437440	0	0	White Pizza	Guidos	23005	

3484 rows × 14 columns

2. Transform

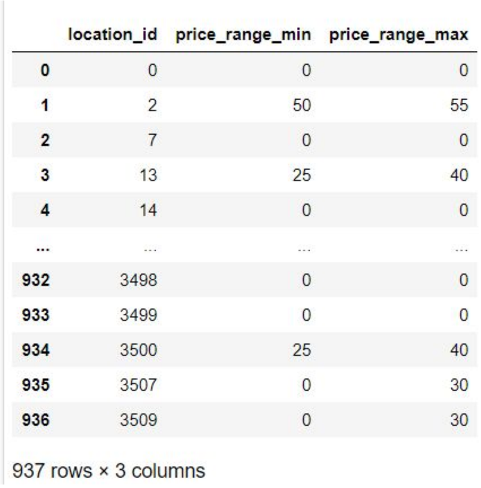
a. The toppings dataset did not require much cleaning, so we left that one as is. The restaurant dataset, however, provided us with a lot of extra information that we did not need. To clean the restaurant data we started by creating a dataframe that only contained the particular columns we were interested in.

- Our “clean_restaurant_df” only consisted of the restaurant name, anything pertaining to restaurant location, and the price range columns. We proceeded to drop any duplicate values, and remove any null values that created problems in our datasets. We split the major dataframe into multiple smaller dataframes; one for how many of each restaurant exists in the United States, one for the location of each restaurant, and one for the price ranges of the menus.



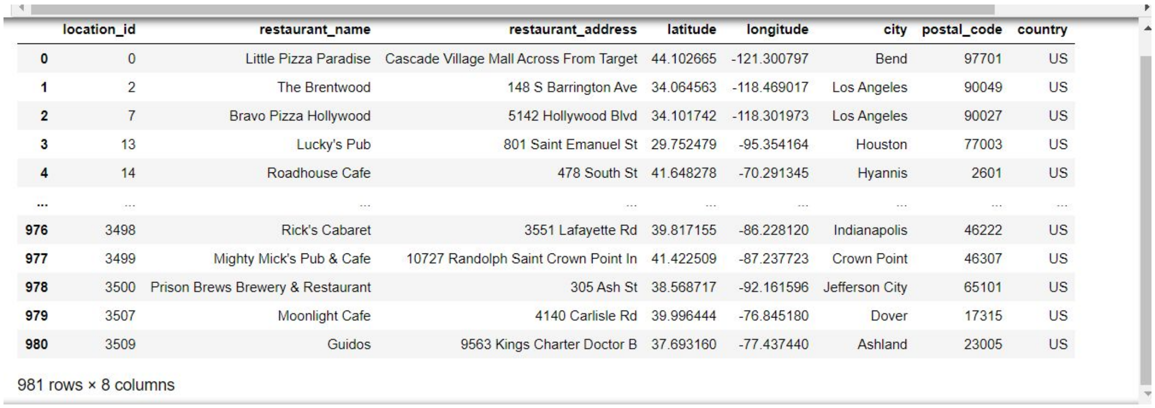
	restaurant_name	count
0	Sicilia's Pizzeria	96
1	J & G Restaurant	55
2	Casey's General Store	43
3	The Pizza Joint	36
4	North End Pizzeria	34
...	...	...
922	First and Last Tavern	1
923	Jake's Pizza	1
924	Casa Margarita's	1
925	Wooden Windmill	1
926	Fiala Aesthetics | Metro Orlando Plastic Surgeon	1

927 rows × 2 columns



	location_id	price_range_min	price_range_max
0	0	0	0
1	2	50	55
2	7	0	0
3	13	25	40
4	14	0	0
...	...	...	...
932	3498	0	0
933	3499	0	0
934	3500	25	40
935	3507	0	30
936	3509	0	30

937 rows × 3 columns



	location_id	restaurant_name	restaurant_address	latitude	longitude	city	postal_code	country
0	0	Little Pizza Paradise	Cascade Village Mall Across From Target	44.102665	-121.300797	Bend	97701	US
1	2	The Brentwood	148 S Barrington Ave	34.064563	-118.469017	Los Angeles	90049	US
2	7	Bravo Pizza Hollywood	5142 Hollywood Blvd	34.101742	-118.301973	Los Angeles	90027	US
3	13	Lucky's Pub	801 Saint Emanuel St	29.752479	-95.354164	Houston	77003	US
4	14	Roadhouse Cafe	478 South St	41.648278	-70.291345	Hyannis	2601	US
...	...	...	...	...	...	...	...	...
976	3498	Rick's Cabaret	3551 Lafayette Rd	39.817155	-86.228120	Indianapolis	46222	US
977	3499	Mighty Mick's Pub & Cafe	10727 Randolph Saint Crown Point In	41.422509	-87.237723	Crown Point	46307	US
978	3500	Prison Brews Brewery & Restaurant	305 Ash St	38.568717	-92.161596	Jefferson City	65101	US
979	3507	Moonlight Cafe	4140 Carlisle Rd	39.996444	-76.845180	Dover	17315	US
980	3509	Guidos	9563 Kings Charter Doctor B	37.693160	-77.437440	Ashland	23005	US

981 rows × 8 columns

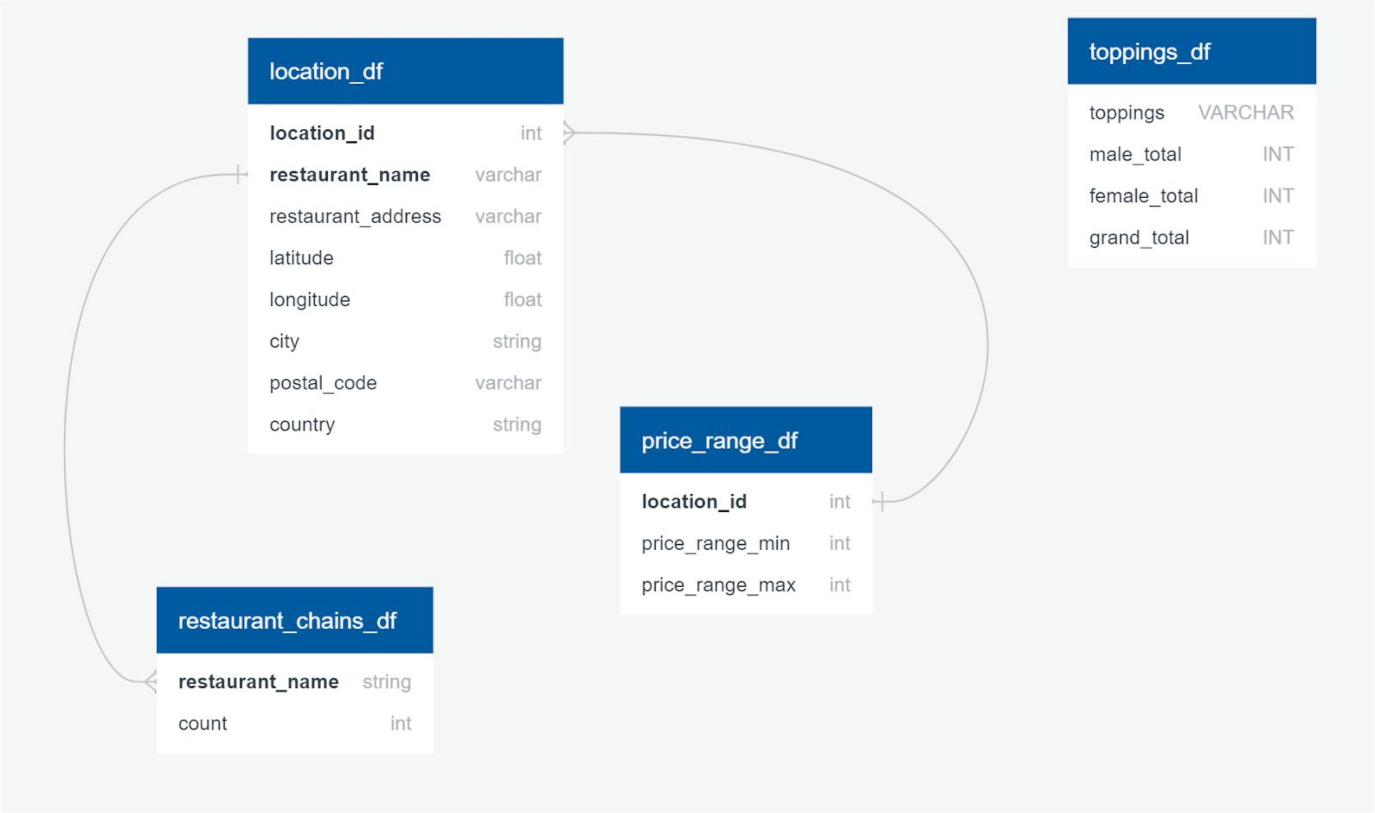
3. Load

a. While loading our tables, we identified that “location_df” had city names under the “province” column. By using the “postal_code” column, Meera came up with the idea to use Google APIs to request the province names and replace the city names accordingly.

        - This proved to be quite difficult since the dataset was composed of approximately 3,500 columns, with a large majority having city names under the province columns.
        - After spending much time attempting to resolve this matter both individually and as a team, with TAs, and even Tutors with no success, we made the decision to take a different approach.
        - Working as a team, we identified that it would be a risk if we continued to try and find ways to resolve this. We ultimately made the decision to exclude this information since the final dataset would still provide key information for our pizza database.

    b. With our four datasets and column selections finalized, we proceed in loading the data. We mainly used PostGres to structure our final data set by joining the three tables “restaurant_chains_df” , “location_df”, and “price_range_df”. The “toppings_df” would be used as a supplementary dataset to our restaurant database (for identifying favorite toppings).
    c. With the datasets loaded, we had our final data set and supplementary data set!

ERD:



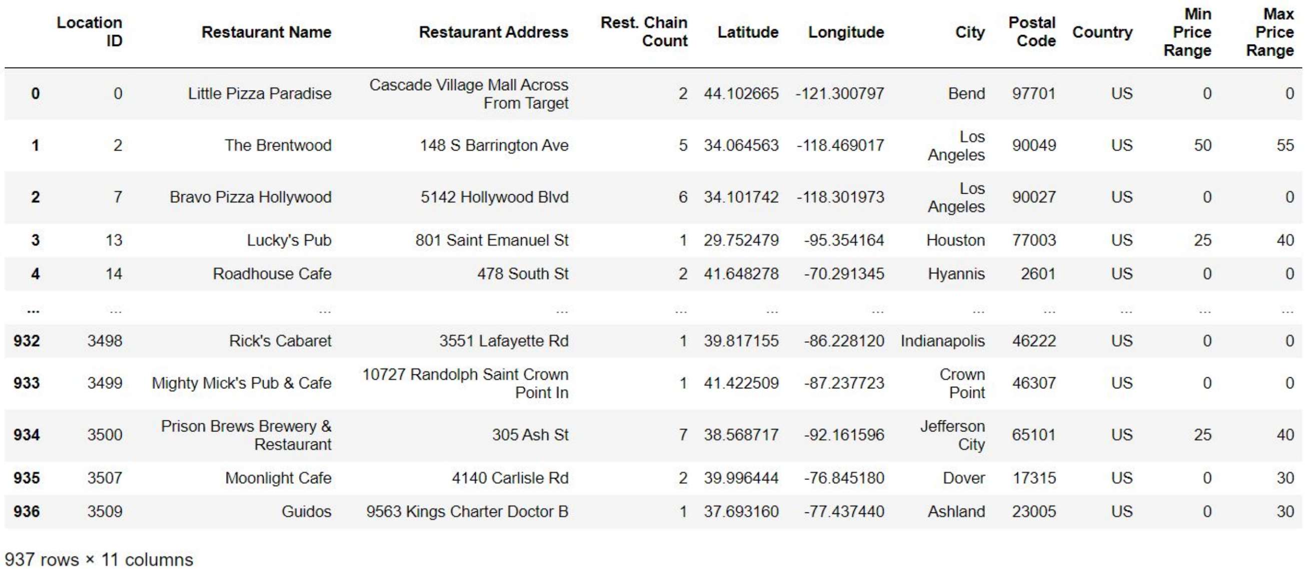
location_df	
location_id	int
restaurant_name	varchar
restaurant_address	varchar
latitude	float
longitude	float
city	string
postal_code	varchar
country	string

toppings_df	
toppings	VARCHAR
male_total	INT
female_total	INT
grand_total	INT

price_range_df	
location_id	int
price_range_min	int
price_range_max	int

restaurant_chains_df	
restaurant_name	string
count	int

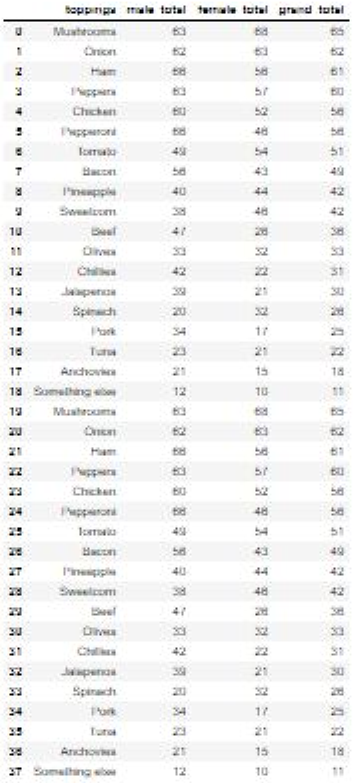
Final Database



	Location ID	Restaurant Name	Restaurant Address	Rest. Chain Count	Latitude	Longitude	City	Postal Code	Country	Min Price Range	Max Price Range
0	0	Little Pizza Paradise	Cascade Village Mall Across From Target	2	44.102665	-121.300797	Bend	97701	US	0	0
1	2	The Brentwood	148 S Barrington Ave	5	34.064563	-118.469017	Los Angeles	90049	US	50	55
2	7	Bravo Pizza Hollywood	5142 Hollywood Blvd	6	34.101742	-118.301973	Los Angeles	90027	US	0	0
3	13	Lucky's Pub	801 Saint Emanuel St	1	29.752479	-95.354164	Houston	77003	US	25	40
4	14	Roadhouse Cafe	478 South St	2	41.648278	-70.291345	Hyannis	2601	US	0	0
...	...	...	...	...	...	...	...	...	...	...	...
932	3498	Rick's Cabaret	3551 Lafayette Rd	1	39.817155	-86.228120	Indianapolis	46222	US	0	0
933	3499	Mighty Mick's Pub & Cafe	10727 Randolph Saint Crown Point In	1	41.422509	-87.237723	Crown Point	46307	US	0	0
934	3500	Prison Brews Brewery & Restaurant	305 Ash St	7	38.568717	-92.161596	Jefferson City	65101	US	25	40
935	3507	Moonlight Cafe	4140 Carlisle Rd	2	39.996444	-76.845180	Dover	17315	US	0	30
936	3509	Guidos	9563 Kings Charter Doctor B	1	37.693160	-77.437440	Ashland	23005	US	0	30

937 rows × 11 columns

Supplementary Database:



	toppings	male total	female total	grand total
0	Mushrooms	63	68	65
1	Onion	62	63	62
2	Ham	66	56	61
3	Peppers	63	57	60
4	Chicken	60	52	56
5	Pepperoni	66	46	56
6	Tomato	49	54	51
7	Bacon	56	43	49
8	Pineapple	40	44	42
9	Sweetcorn	38	46	42
10	Beef	47	28	38
11	Olives	33	32	33
12	Chillies	42	22	31
13	Jalapenos	39	21	30
14	Spinach	20	32	26
15	Pork	34	17	25
16	Tuna	23	21	22
17	Anchovies	21	15	18
18	Something else	12	10	11
19	Mushrooms	63	68	65
20	Onion	62	63	62
21	Ham	66	56	61
22	Peppers	63	57	60
23	Chicken	60	52	56
24	Pepperoni	66	46	56
25	Tomato	49	54	51
26	Bacon	56	43	49
27	Pineapple	40	44	42
28	Sweetcorn	38	46	42
29	Beef	47	28	38
30	Olives	33	32	33
31	Chillies	42	22	31
32	Jalapenos	39	21	30
33	Spinach	20	32	26
34	Pork	34	17	25
35	Tuna	23	21	22
36	Anchovies	21	15	18
37	Something else	12	10	11