YELP BUSINESS AND REVIEWS

The dataset used is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge for students to conduct research or analysis on Yelp's data . In the dataset you'll find information about businesses across 8 metropolitan areas in the USA and Canada. It contains 5 json files but for the purpose of this project I will be using only the business and reviews data files.

I came across this data set on Kaggle a platform where we can find data science resources necessary for the community. The first Business data set contains around 170,000 rows and and 14 columns. The second Reviews data set consists of 700,000 rows and 9 columns.

Features:

Business:

- business_id : Id of the business
- Name: name of the business
- address : location/address
- city : city where it is present
- state : state in which it is present
- postal_code : zip code
- latitude : latitudinal location of the business
- longitude : longitudinal location of the business
- stars : customer ratings
- review_count : total number of reviews
- is_open : status whether it is open or close
- attributes : services offered at the place
- categories : Categories the business fall under

	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	attributes	categories	hours
0	Pns2l4eNsfO8kk83dixA6A	Abby Rappoport, LAC, CMQ	1616 Chapala St, Ste 2	Santa Barbara	CA	93101	34.426679	-119.711197	5.0	7	0	{'ByAppointmentOnly': 'True'}	Doctors, Traditional Chinese Medicine, Naturop...	None
1	mpf3x-BjTdTEA3yCZrAYPw	The UPS Store	87 Grasso Plaza Shopping Center	Affton	MO	63123	38.551126	-90.335695	3.0	15	1	{'BusinessAcceptsCreditCards': 'True'}	Shipping Centers, Local Services, Notaries, Ma...	{'Monday': '0:0-0:0', 'Tuesday': '8:0-18:30', ...
2	tUFrWirKiKi_TAnsVWINQQ	Target	5255 E Broadway Blvd	Tucson	AZ	85711	32.223236	-110.880452	3.5	22	0	{'BikeParking': 'True', 'BusinessAcceptsCredit...	Department Stores, Shopping, Fashion, Home & G...	{'Monday': '8:0-22:0', 'Tuesday': '8:0-22:0', ...
3	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	80	1	{'RestaurantsDelivery': 'False', 'OutdoorSeati...	Restaurants, Food, Bubble Tea, Coffee & Tea, B...	{'Monday': '7:0-20:0', 'Tuesday': '7:0-20:0', ...
4	mWMc6_wTdE0EUBKIGXDVfA	Perkiomen Valley Brewery	101 Walnut St	Green Lane	PA	18054	40.338183	-75.471659	4.5	13	1	{'BusinessAcceptsCreditCards': 'True', 'Wheelc...	Brewpubs, Breweries, Food	{'Wednesday': '14:0-22:0', 'Thursday': '16:0-2...

Reviews:

- business_id: Id of the business

- stars : customer ratings
- Useful : total number of useful reviews
- funny : total number of funny reviews
- Cool : total number of cool reviews
- Text : review text
- Date : date of the review
- Review id: id of the review
- User id : id of the user who posted it

	review_id	user_id	business_id	stars	useful	funny	cool	text	date
0	KU_O5udG6zpxOg-VcAEodg	mh_-eMZ6K5RLWhZyISBhwA	XQfwVwDr-v0ZS3_CbbE5Xw	3	0	0	0	If you decide to eat here, just be aware it is...	2018-07-07 22:09:11
1	BiTunyQ73aT9WBnpR9DZGw	OyoGAe7OKpv6SyGZT5g77Q	7ATYjTIgM3jUlt4UM3IypQ	5	1	0	1	I've taken a lot of spin classes over the year...	2012-01-03 15:28:18
2	saUsX_uimxRlCVr67Z4Jig	8g_iMtfSiwikVnbP2etR0A	YjUWPpI6HXG530lwP-fb2A	3	0	0	0	Family diner. Had the buffet. Eclectic assortm...	2014-02-05 20:30:30
3	AqPFMleE6RsU23_auESxiA	_7bHUi9Uuf5_HHc_Q8guQ	kxX2SOes4o-D3ZQBkiMRfA	5	1	0	1	Wow! Yummy, different, delicious. Our favo...	2015-01-04 00:01:03
4	Sx8TMOWLNuJBWer-0pcmoA	bcjbaE6dDog4jkNY91ncLQ	e4Vwtrqf-wpJfwesgvdgxQ	4	1	0	1	Cute interior and owner (?) gave us tour of up...	2017-01-14 20:54:15

Data Reading, Cleaning and Formatting:

We begin by importing the necessary libraries and reading the json files of our data in our Google colab notebook. For this analysis we will use numpy and pandas for linear algebra, data processing and json for reading our file. We will also use seaborn and matplotlib.pyplot for our graphical representation and analysis. Since our files are huge and all the data inside the files was divided into multiple dictionaries we first parse each line into a python dictionary and then append it to the list which will then be converted to pandas data frame.

However in case of the reviews file it contains a lot of rows which I feel will not be required hence I proceed to divide the file to chunks of 100 rows each and create a list of chunks that will only add rows till it reaches the limit of 70000. This list is then converted to pandas data frame

I then check for null values present in both the data frames and drop these null values that were present in the attributes, categories and hours column. I then delete the columns that I will not be using for analysis. This results in the final data sets with no null values where the business data set contains136601 rows and 13 columns and the reviews data set contains 70000 rows and 7 columns.

```
business_id     0
stars           0
useful          0
funny           0
cool            0
text            0
date            0
dtype: int64
(70000, 7)
business_id      0
name             0
address          0
city             0
state            0
postal_code      0
latitude         0
longitude        0
stars            0
review_count     0
is_open          0
attributes       0
categories       0
dtype: int64
(136601, 13)
```

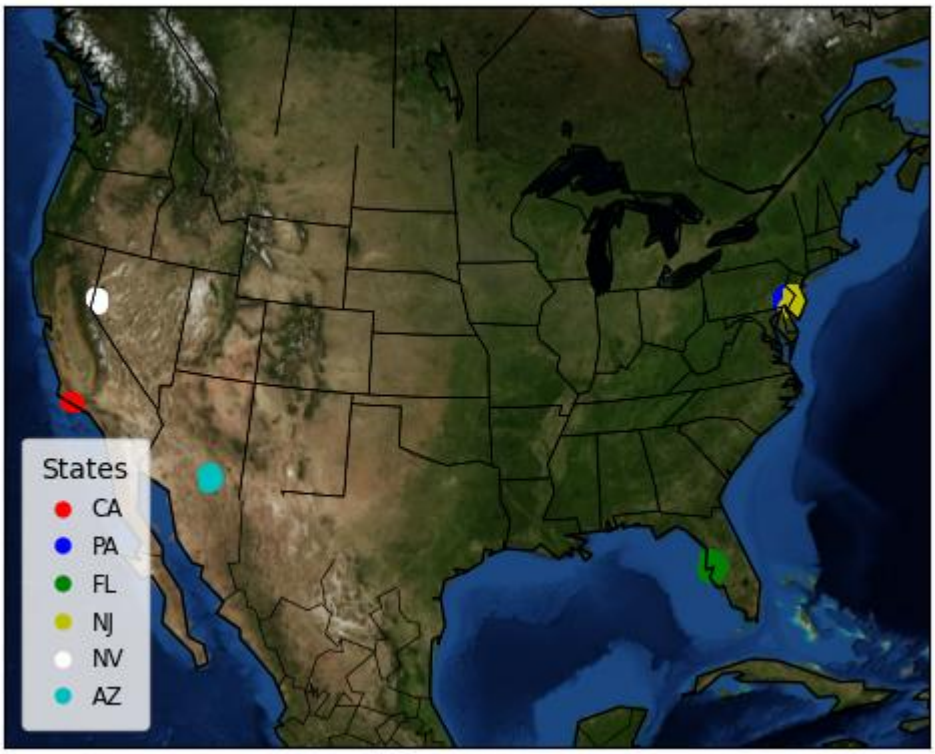
I was curious to know the location of these business specifically restaurants on the united states map. Upon filtering only restaurants from the entire data set I created a new datframe which will contain all these restaurants and data related to them. There were around 51703 restaurants.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 51703 entries, 3 to 150340
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   business_id   51703 non-null  object
 1   name          51703 non-null  object
 2   address       51703 non-null  object
 3   city          51703 non-null  object
 4   state         51703 non-null  object
 5   postal_code   51703 non-null  object
 6   latitude      51703 non-null  float64
 7   longitude     51703 non-null  float64
 8   stars         51703 non-null  float64
 9   review_count  51703 non-null  int64
 10  is_open       51703 non-null  int64
 11  attributes    51703 non-null  object
 12  categories    51703 non-null  object
 13  restaurant    51703 non-null  bool
dtypes: bool(1), float64(3), int64(2), object(8)
memory usage: 5.6+ MB
```

I further created different data frames of a few popular states and the restaurants present in these states and included only those whose review count us greater than 100. Combined all of these data frames together and included only a few necessary columns. The states that I decided to look into were California, Pennsylvania, Arizona, Florida, Nevada and New Jersey. A total of 7150 restaurants were found that were frequently visited in these states

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7150 entries, 85 to 150275
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   name        7150 non-null   object
 1   stars       7150 non-null   float64
 2   categories  7150 non-null   object
 3   latitude    7150 non-null   float64
 4   longitude   7150 non-null   float64
 5   city        7150 non-null   object
 6   attributes  7150 non-null   object
 7   state       7150 non-null   object
```

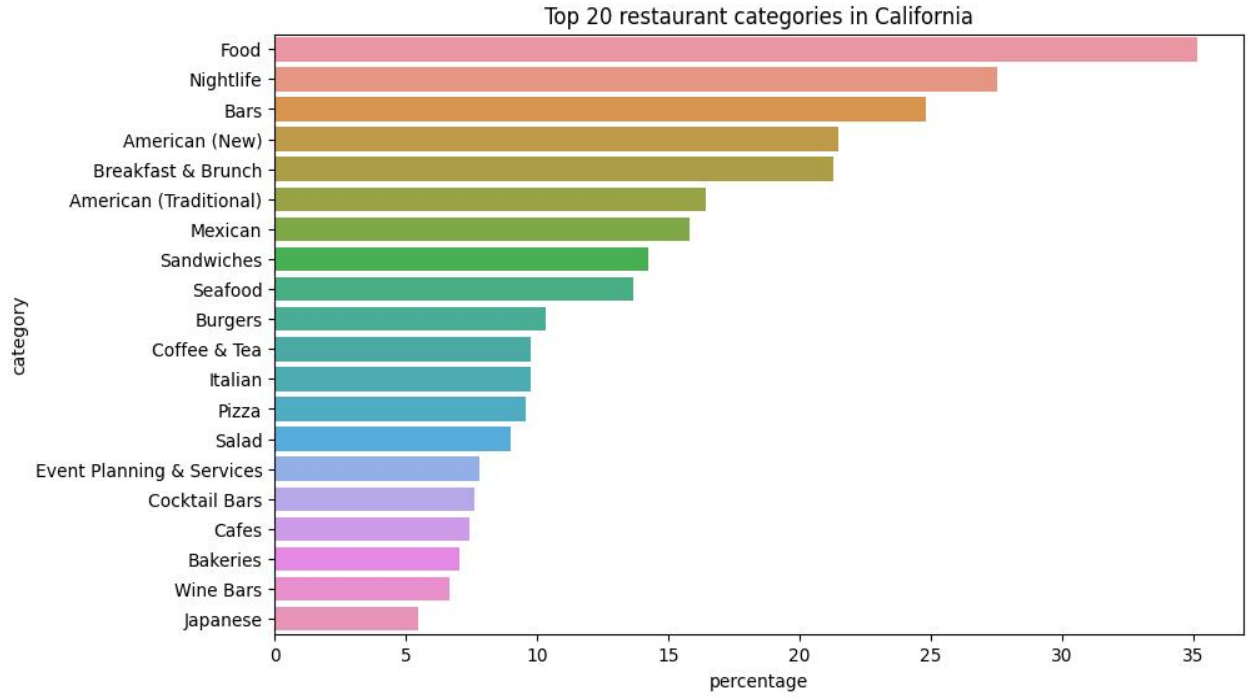
The map below represents the geographical locations of these restaurants on the united states map. I used the basemap package present in matplotlib to get a satellite image of the states.



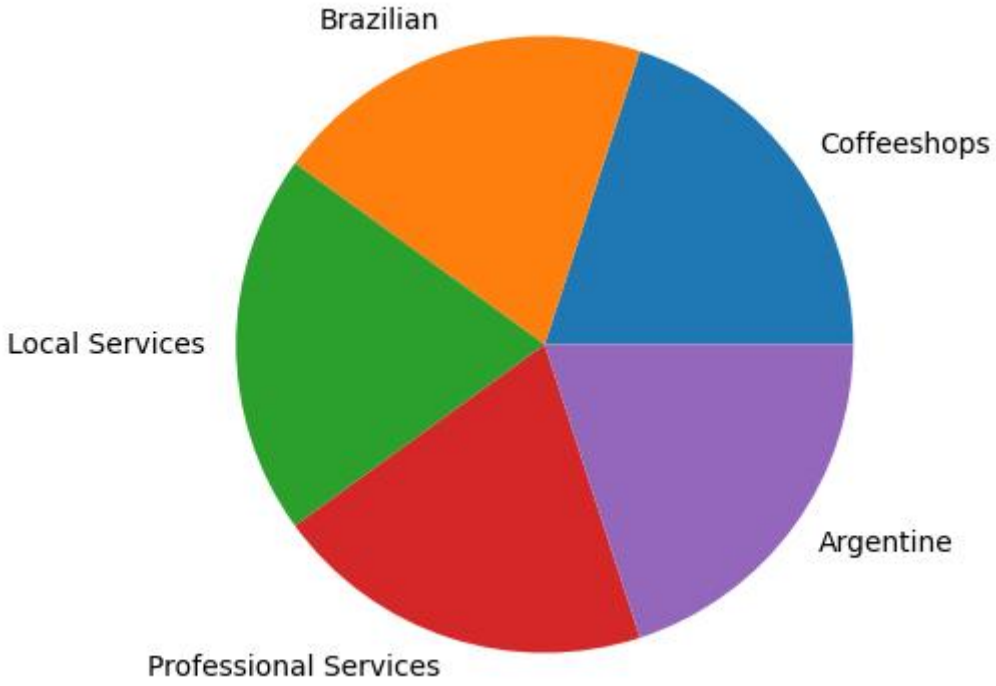
I then decided to conduct analysis for only one particular state initially.
As I will be visiting California this summer I was curious to know the top 20 restaurant categories I could visit there. I start of by defining a function called category which will count the frequency, percentage and average ratings for all the restaurant categories present in the categories column of the California restaurants data frame which are split by a , and return a data frame listing all the categories and their respective data. For California below picture represents the new category data frame created:

category	frequency	rating_sum	average_rating	percentage
Food	180	716.5	3.980556	35.156250
Nightlife	141	557.0	3.950355	27.539062
Bars	127	502.5	3.956693	24.804688
American (New)	110	441.5	4.013636	21.484375
Breakfast & Brunch	109	425.5	3.903670	21.289062
American (Traditional)	84	324.5	3.863095	16.406250
Mexican	81	314.5	3.882716	15.820312
Sandwiches	73	293.5	4.020548	14.257812
Seafood	70	270.5	3.864286	13.671875
Burgers	53	207.0	3.905660	10.351562

Using seaborn we plot the top 20 restaurant categories. As you can observe from the graph below Food, Nightlife and Bars seem to be very popular categories out of all the 20



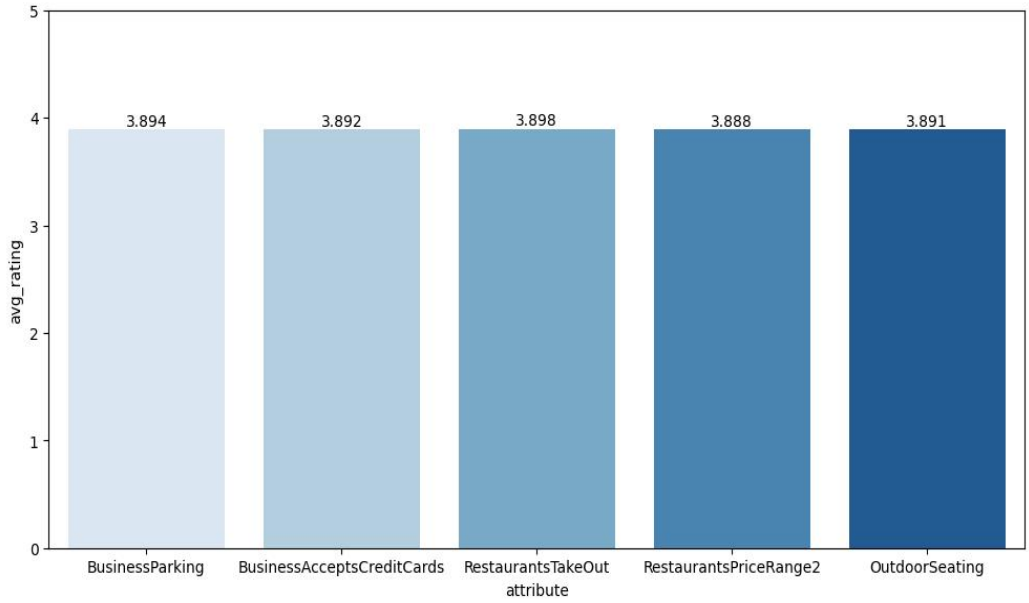
Similarly what are the least popular categories of restaurants in California. I used to a pie chart that Brazilian, Argentine & Coffee shops seems to be the least popular



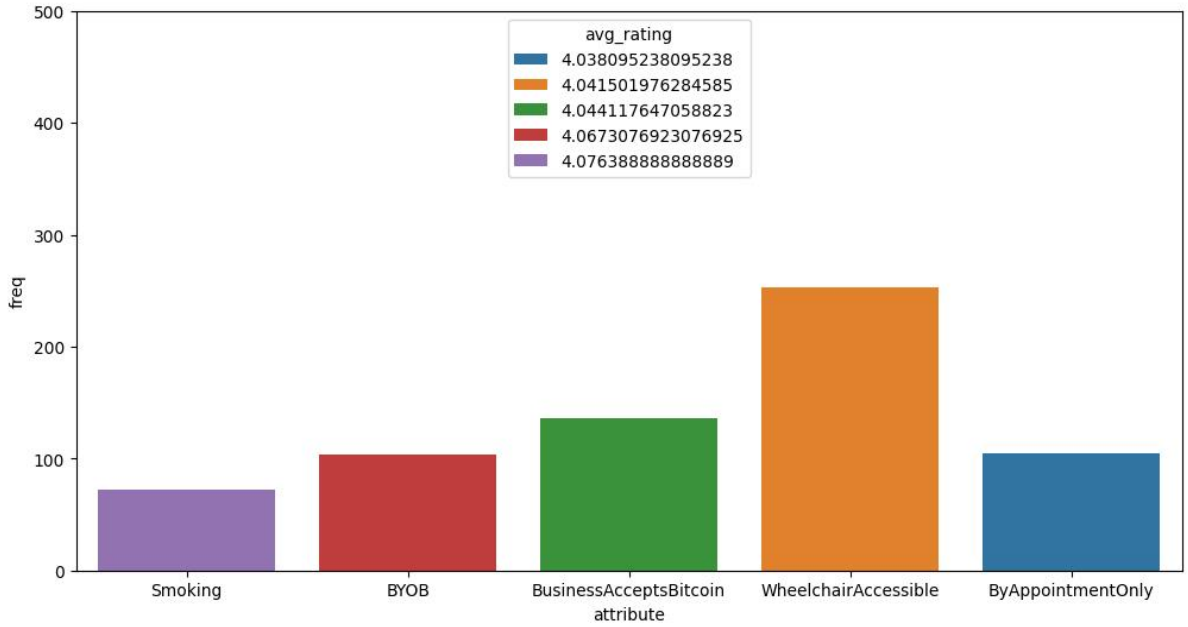
Another interesting task would be knowing the attributes or services provided by these frequently visited restaurants. I again defined a function called att which will count the frequency and average ratings for all the restaurant attributes present in the California restaurants data frame and return a data frame listing all the attributes and their respective data. For California below picture represents the new category data frame created:

	attribute	freq	avg_rating
19	BusinessParking	510	3.894118
5	BusinessAcceptsCreditCards	509	3.891945
0	RestaurantsTakeOut	508	3.897638
8	RestaurantsPriceRange2	504	3.887897
10	OutdoorSeating	504	3.890873
11	RestaurantsDelivery	503	3.895626
21	WiFi	499	3.885772
3	Ambience	492	3.889228
12	HasTV	492	3.887195
14	Alcohol	492	3.883130

The plot below showcases the attributes of the restaurants that are frequently visited. Almost all these attributes seem to have a similar rating and do attract customers.

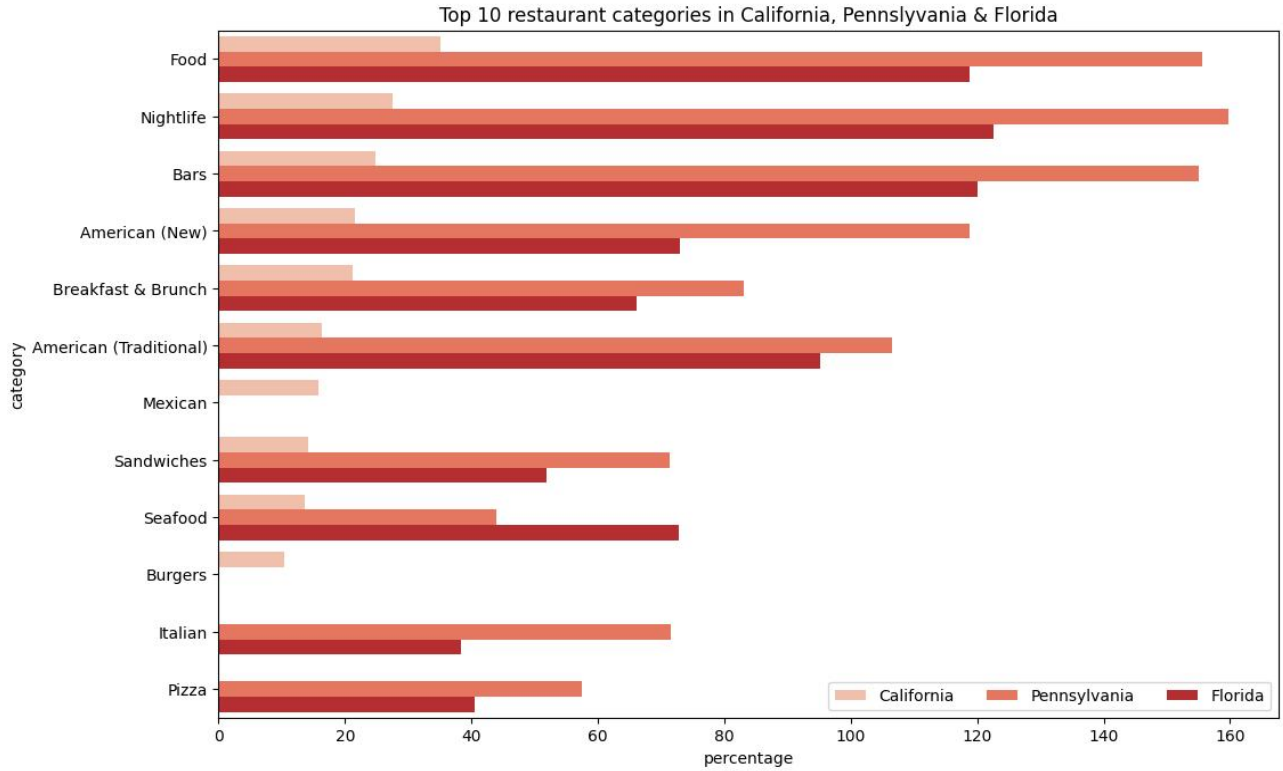


However, what are the attributes of the highest rated restaurants. Another barplot below showcases that restaurants that allow smoking might not be frequently visited but are indeed rated high and the ones that provide wheelchair accessibility are certainly frequently visited. The restaurants that allow seating by appointment only are the highest though they are no frequently visited as they could probably fine dining restaurant and a little on the expensive end.

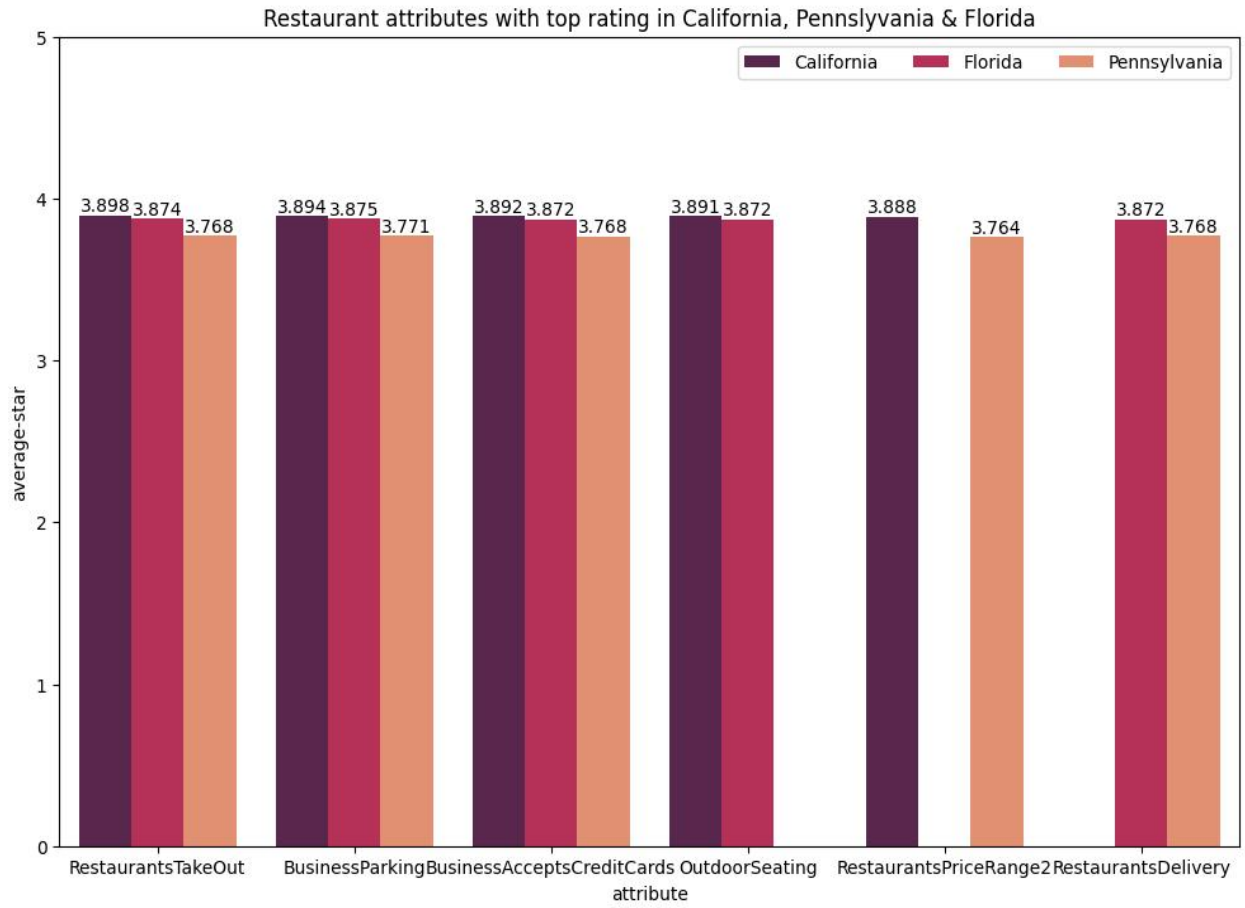


How would this data for California compare against two other states though?
For that I again create category and attribute data frame for Pennsylvania and Florida state using the functions defined earlier. And combine the top 10 categories for the three states into one data frame and combine the top 5 attributes for the three states into another data frame.

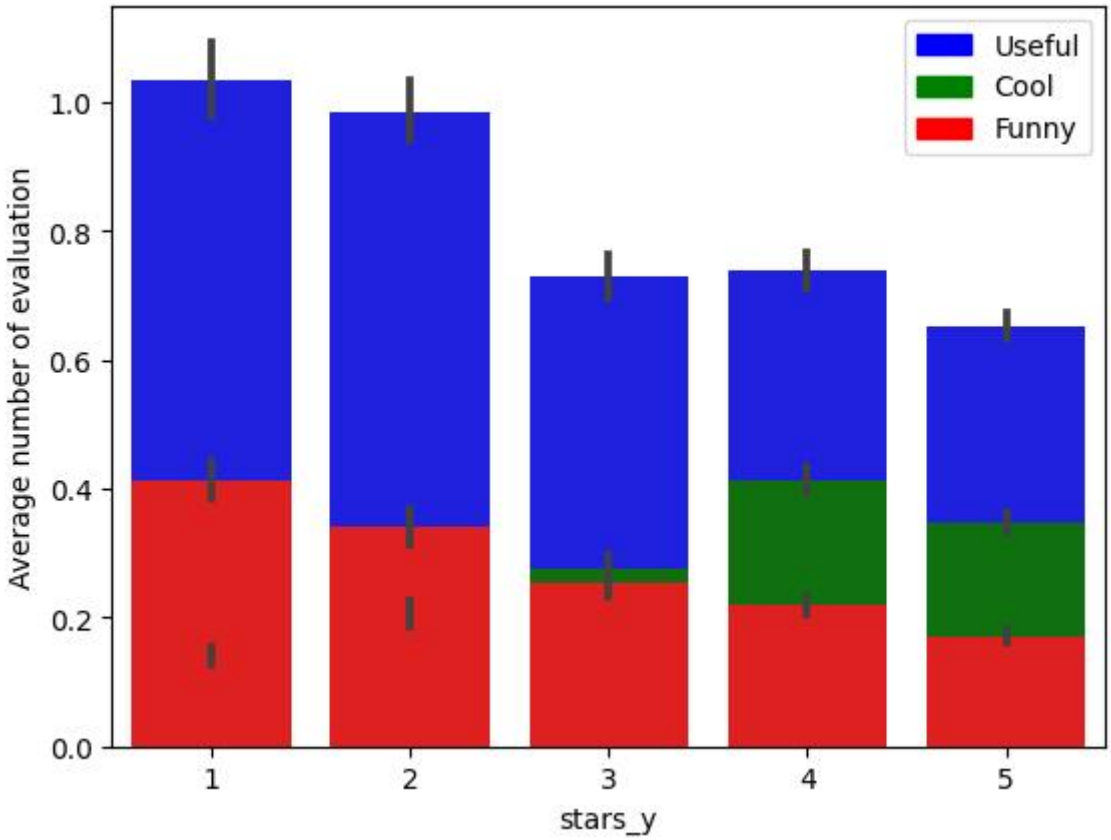
The graph below show cases how food is more popular in Pennsylvania and Florida than in California. Similar could be said for other categories like bars and nightlife. However, only California seems to have frequently visited Mexican restaurants.



And for the attributes Outdoor seating is not popular in Pennsylvania and Restaurants Delivery in California. The rating for all these restaurants seem to be very similar.



I also wanted to understand how reviews are helpful and hence merged the restaurant data with reviews data by business id and decided to plot the rate of useful, cool and sunny reviews for each rating category. Clearly the customers found the reviews the most useful for all the restaurants rated 1 and 2.



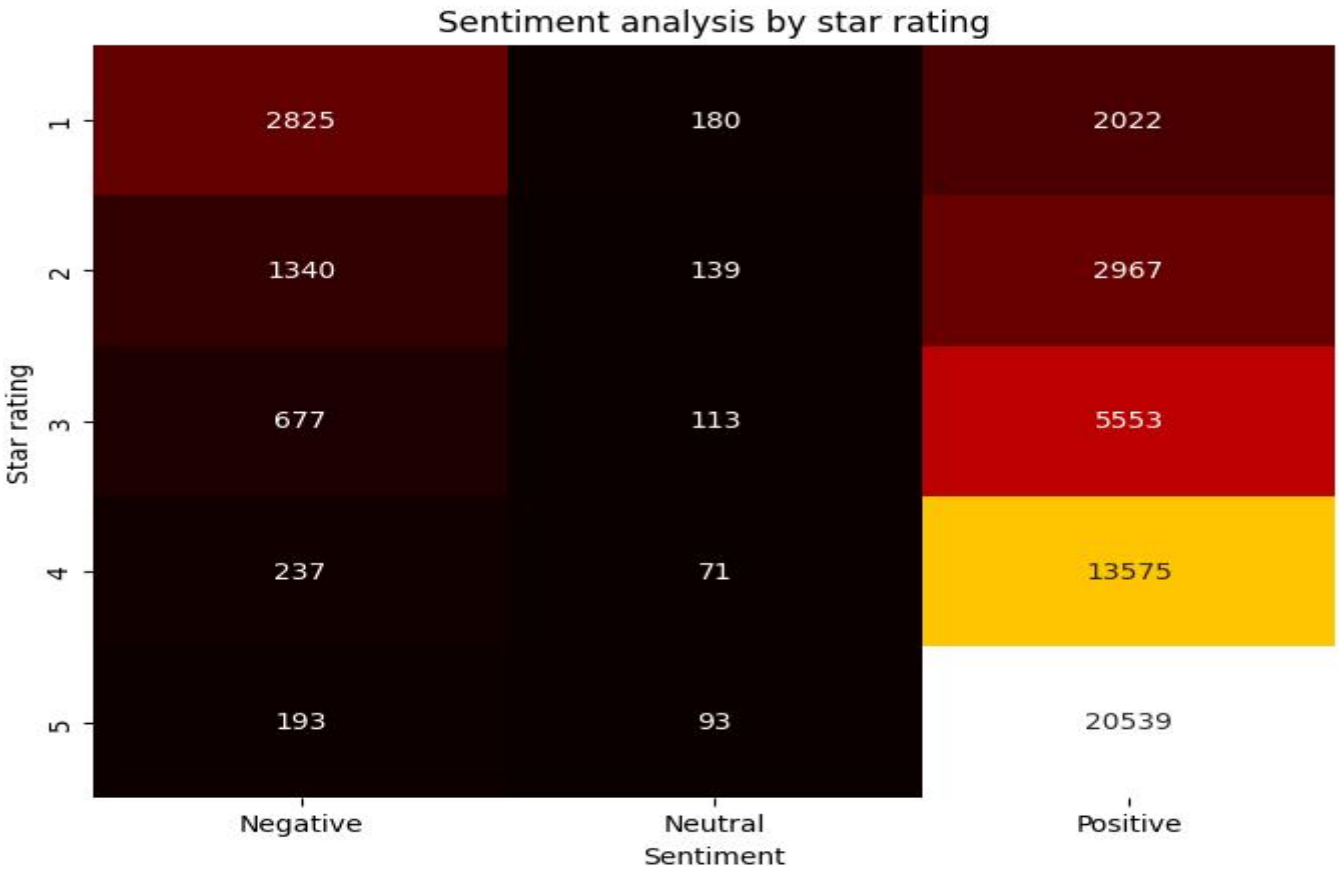
I wanted to understand the sentiment of these reviews and hence decided to perform sentiment analysis on the text. I used Vader sentiment analysis tool which is a a rule-based sentiment analysis tool that uses a lexicon of words and phrases with pre-defined sentiment scores to determine the overall sentiment of the text. I added another column called sentiment which will tell us the sentiment for the restaurant where 1 indicates positive, 0 is neutral and -1 is negative.

restaurant	stars_y	useful	funny	cool	text	date	Sentiment
True	4	0	0	1	This is nice little Chinese bakery in the hear...	2014-05-26	1
True	4	3	1	2	This is the bakery I usually go to in Chinatow...	2013-10-05	1
True	5	0	0	0	A delightful find in Chinatown! Very clean, an...	2013-10-25	1
True	5	5	0	5	I ordered a graduation cake for my niece and i...	2018-05-20	1
True	4	2	1	1	HK-STYLE MILK TEA: FOUR STARS\n\nNot	2013-10-25	-1

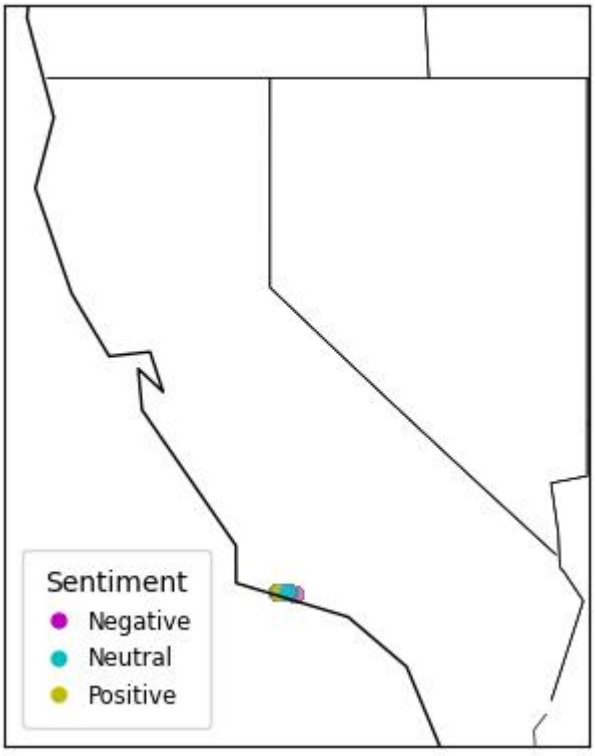
I decided to group by the restaurant rating and the number of sentiment tweets in each rating.

stars_y	Negative	Neutral	Positive
1	2825	180	2022
2	1340	139	2967
3	677	113	5553
4	237	71	13575
5	193	93	20539

The heat map below helps visualize patters in these ratings and sentiments.



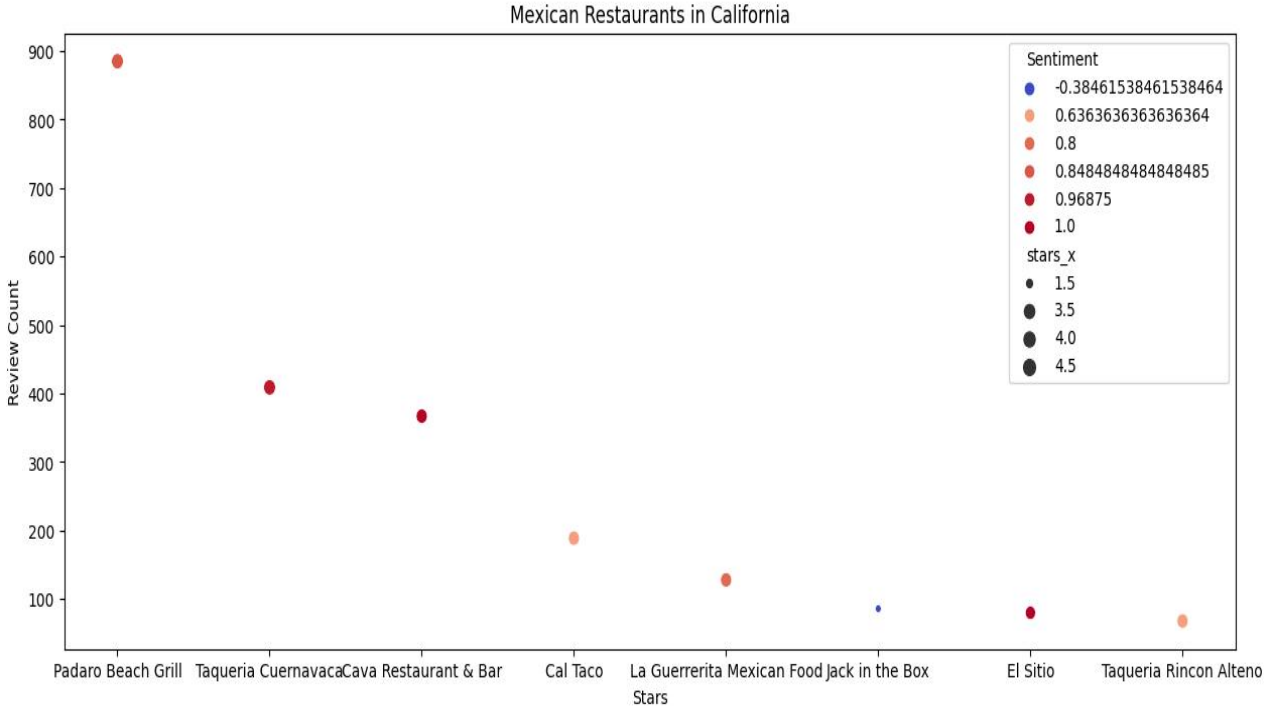
I wanted to specifically try the Mexican restaurants in California and decided to filter them out and ploy them on the bases of the sentiments on yelp.

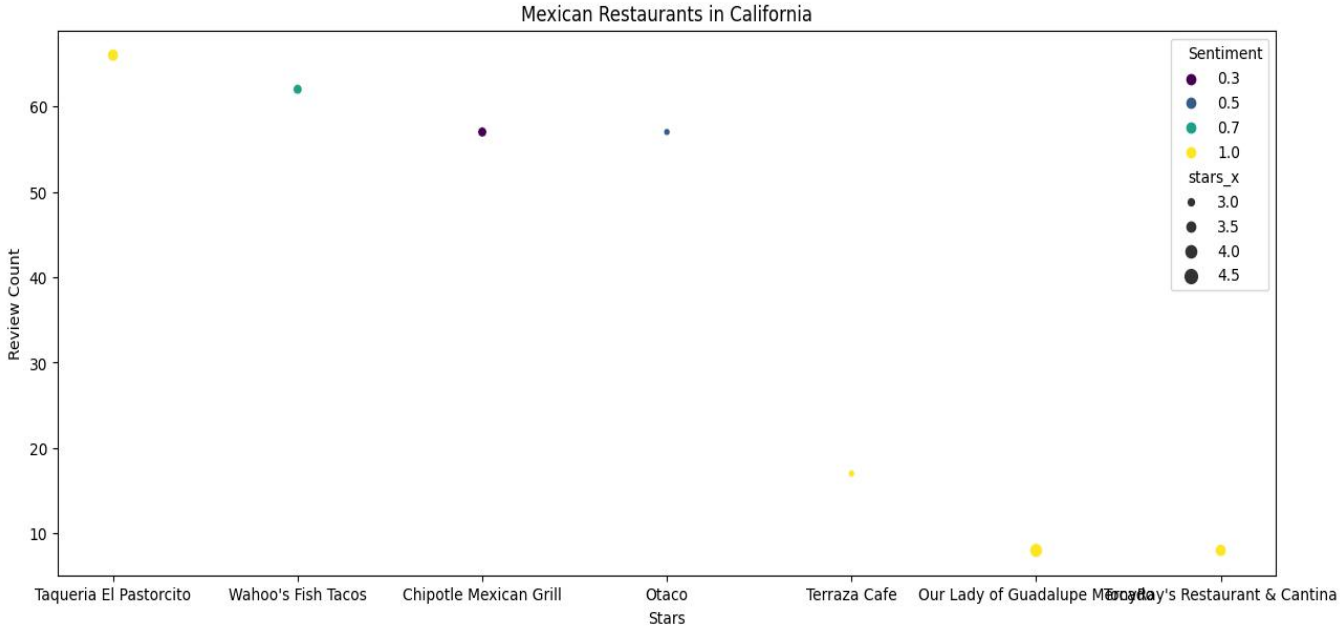


I then grouped all of these restaurants by name and calculated the average rating, review count and sentiment for all these restaurant.

	name	stars_x	review_count	Sentiment
8	Padaro Beach Grill	4.5	885.0	0.848485
9	Taqueria Cuernavaca	4.5	409.0	0.968750
1	Cava Restaurant & Bar	4.0	367.0	1.000000
0	Cal Taco	4.0	189.0	0.636364
5	La Guerrerita Mexican Food	4.0	128.0	0.800000
4	Jack in the Box	1.5	86.0	-0.384615
3	El Sitio	3.5	80.0	1.000000
11	Taqueria Rincon Alteno	4.0	68.0	0.636364
10	Taqueria El Pastorcito	4.0	66.0	1.000000
14	Wahoo's Fish Tacos	3.5	62.0	0.700000
2	Chipotle Mexican Grill	3.5	57.0	0.300000
6	Otaco	3.0	57.0	0.500000
12	Terraza Cafe	3.0	17.0	1.000000
7	Our Lady of Guadalupe Mercado	4.5	8.0	1.000000
13	TonyRay's Restaurant & Cantina	4.0	8.0	1.000000

The scatter plot below showcases the relation between the ratings and frequency for each restaurant. As you can see Jack in the box even though frequently visited has pretty bad reviews and negative sentiments




Mexican Restaurants in California
Review Count
60
50
40
30
20
10
Sentiment
0.3
0.5
0.7
1.0
stars_x
3.0
3.5
4.0
4.5
Taqueria El Pastorcito
Wahoo's Fish Tacos
Chipotle Mexican Grill
Otaco
Terraza Cafe
Our Lady of Guadalupe Mexican
Tequila Ray's Restaurant & Cantina
Stars