USA Restaurant Mania!

Jeremy Erickson

Introduction:

When a business is thinking about creating either a new chain of restaurants or adding more locations of an existing chain, it is a great strategic idea to know the lay of the land already for a given city or groups of cities. Knowledge on what types of restaurants are already present and the frequency of those in general is a key factor in making a decision. Additionally, understanding any patterns or relationships between the cities could help drive smart decisions on either which cuisines the company wants to introduce to those cities or which cuisines they should focus on for those closely related cities.

Business Plan:

The top 300 USA cities are listed on Wikipedia. However, it does not include the restaurant venue information along with any additional analytics such as the most common cuisines for each city. Further, one cannot easily see any relationships at all between the cities. If this information was available, the company could make more objective and informed decisions on where to create additional restaurants and what type of cuisine they should offer and serve.

The main purpose of this project is to create the dataset described above. First, a dataset will be created by web scraping the top 300 USA cities with their respective locations. Next, another dataset will be created listing the closest 100 restaurants from the each city center (if available) by leveraging the coordinates. Finally, a common clustering algorithm will be used to visually show relationships between the cities based upon the most frequently available cuisine types based upon the information found.

Data Description:

This section contains descriptions of the data that will be used to analyze the problem of determining where to create new restaurants and what type of cuisine should be severed. The data is to be collected from two main sources.

1. **Top 300 USA Cities by Population:**

   First the table is extracted from the website (https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population) by using a web scraping technique. This dataset includes the City name, the state, the population, the latitude and longitude coordinates.

2. **100 Closest Restaurants around each city:**

   In order to retrieve the closest named restaurants (up to 100) for each of the cities, the coordinates from the Top 300 USA Cities dataset are leveraged to trigger the Foursquare API for each city. The query is to bring back the results in a JSON file format. The results are then parsed further and reshaped in order to display the city, the restaurant name, type, and location for each data point.

3. **Clustering Results:**

   The above dataset described as 100 Closest Restaurants around each city is massaged and manipulated further in such a way to create the top 10 most frequent types of restaurants for each city. This dataset is then used to perform the k-means clustering algorithm in order to generate a new dataset containing the cluster assignments that will be used to visually establish relationships between the cities based upon the top 10 most frequent restaurant types. The visual distribution is displayed by using the USA Map with different colored dots representing how the cities are related to one another based upon restaurant cuisine type distributions. The resulting clustering results dataset will contain the City name, the cluster it belongs to, and the top 10 restaurant types. Deeper analysis of this dataset will provide insight on both how similar each city is to one another and what the top 10 most frequent restaurant types there are within or around the city.

Methodology:

Development Environment:

The main development environment is Windows 10 64 bit edition. Anaconda Python (2019-03 release) version 3.7 64 bit was downloaded and installed following default directions. A Jupyter ipython notebook was created containing all of the code used that is available in the same GitHub repository. Several additional libraries are needed in order to replicate the analysis and are listed below:

> numpy, pandas, json, geopy, matplotlib, folium, random, scikit-sklearn, scipy, bs4, xml, requests, re, and mpl_toolkits

Algorithm:

1. Data was scraped using the Beautiful Soup library from the Wikipedia URL: https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population. Beautiful Soup is a Python library that provided methods for performing Web Scraping. Most information available out on the internet is in the form of unstructured data. When the data is at least Semi-structured such as in the form of a XML document, then it can be parsed.

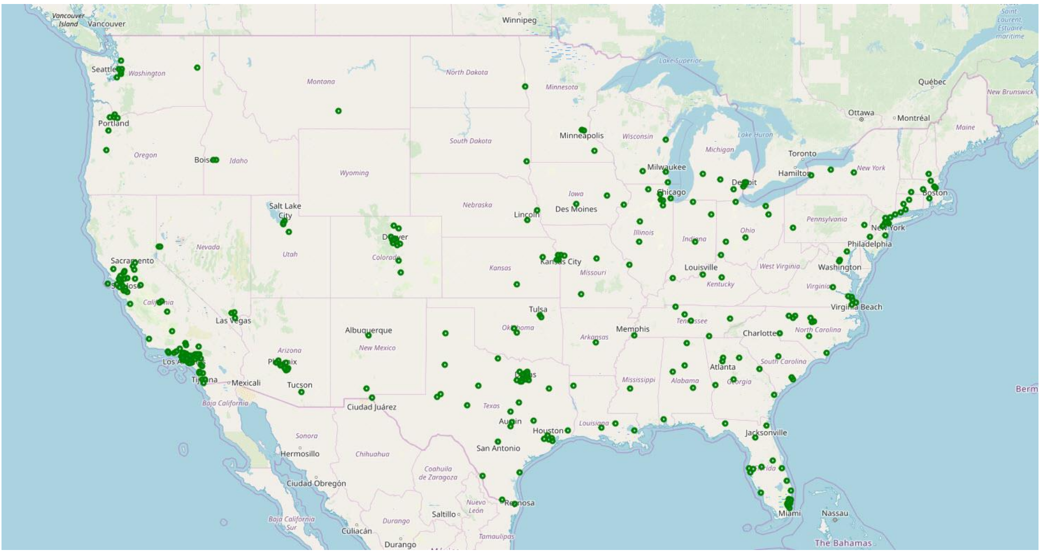
2. The data was placed into a Pandas data frame. A Pandas data frame is one of the more common frameworks to manipulate date further in. A lot of functionality is built into the Pandas library to visualize, manipulate and summarize the data efficiently. The data frame is made up of columns and rows and each data point having a row and column location. A portion of the Pandas data frame table is shown below:

Rank	City	State	Population	Latitude	Longitude
1	New York City	New York	8,398,748	40.6635	-73.9387
2	Los Angeles	California	3,990,456	34.0194	-118.4108
3	Chicago	Illinois	2,705,994	41.8376	-87.6818
4	Houston	Texas	2,325,502	29.7866	-95.3909
5	Phoenix	Arizona	1,660,272	33.5722	-112.0901
6	Philadelphia	Pennsylvania	1,584,138	40.0094	-75.1333
7	San Antonio	Texas	1,532,233	29.4724	-98.5251
8	San Diego	California	1,425,976	32.8153	-117.135
9	Dallas	Texas	1,345,047	32.7933	-96.7665
10	San Jose	California	1,030,119	37.2967	-121.8189

3. The Python library geopy provides a method to retrieve the longitude and latitude coordinates of most addresses. Leveraging geopy, the longitude and latitude was retrieved for The United States address. This is needed to create the base map of the United States for visualization.

4. Next, the folium library is used to visualize the locations of all 300+ cities. Given the base latitude and longitude coordinates of the United States, a loop function is created to add each city to the existing map where each green dot represents a city location. The folium library provides this capability of using an interactive map where one could zoom in and out and provide additional metadata for the city. Here below is just a screenshot and does not provide any of that additional functionality. Please refer to the Jupyter Notebook for that additional capability. See Figure 1 below:

Figure 1: Top 300+ USA Cities based on Population (Anchorage, Alaska and Honolulu, Hawaii not pictured)



5. Foursquare provides various types of information centered on a certain geographical location and/or longitude and latitude coordinates. Some of the categories available are trending, top picks, food, nightlife, coffee, fun, shopping and breakfast. It provides a modern user experience in which the user can quickly gather information around a given or current location. Foursquare provides a developer API to gain backend access to that data that they have collected. Leveraging the Foursquare API, requests were issued to return the 100 (if available) closest restaurant names, and types based upon the longitude and latitude scraped from the website for each city. The result was a Pandas data set. A snippet of the Pandas data frame is shown below:

Rank	City	State	Population	Venue Name	Type	Latitude	Longitude
1	New York City	New York	8,398,748	PLG Coffee House and Tavern	Café	40.66000671	-73.95336194
1	New York City	New York	8,398,748	The Food Sermon	Caribbean Restaurant	40.66458836	-73.9537351
1	New York City	New York	8,398,748	Barboncino	Pizza Place	40.672104	-73.95741216
1	New York City	New York	8,398,748	Hunky Dory	Bistro	40.67313911	-73.95702881
1	New York City	New York	8,398,748	Silver Rice	Sushi Restaurant	40.67418665	-73.95703711
1	New York City	New York	8,398,748	Saraghina	Pizza Place	40.68359	-73.93534
1	New York City	New York	8,398,748	The Islands	Caribbean Restaurant	40.67703588	-73.96356322
1	New York City	New York	8,398,748	Dough	Donut Shop	40.689042	-73.956978
1	New York City	New York	8,398,748	Puerto Viejo	Latin American Restaurant	40.67892483	-73.96196022
1	New York City	New York	8,398,748	Speedy Romeo	Pizza Place	40.68739667	-73.95987869
1	New York City	New York	8,398,748	Olmsted	New American Restaurant	40.67717553	-73.96893144
1	New York City	New York	8,398,748	Chilo's	Taco Place	40.68841796	-73.95698084
1	New York City	New York	8,398,748	Emily	Pizza Place	40.68341995	-73.96655064
1	New York City	New York	8,398,748	Der Pioneer	Bakery	40.64591088	-73.97202797
1	New York City	New York	8,398,748	Evelina Restaurant	Italian Restaurant	40.68958329	-73.97108254
1	New York City	New York	8,398,748	Olea	Tapas Restaurant	40.68771612	-73.97059433

6. Exploration of the ~30,000 row Pandas dataset is performed next to look for trends and potentially cleaning up the data set further before downstream analysis. A few key discoveries were made.
    a. The City variable was not a unique field. There were at least 2 instances of the same city name being found in different states. Therefore, a new unique variable was created called Address that was a concatenation of the City and State variables
    b. It was observed that a couple of cities had less than 100 restaurants returned. There was no correction made and these cities were still included in the analysis.
    c. It was determined that there were 129 unique Types of restaurants across the total data set. Pizza Place, American, Mexican, Italian, and a Sandwich Place were the top 5 types. The 11th most common type was "Restaurant". Therefore, those records were eliminated from the analysis. The type "Restaurant"

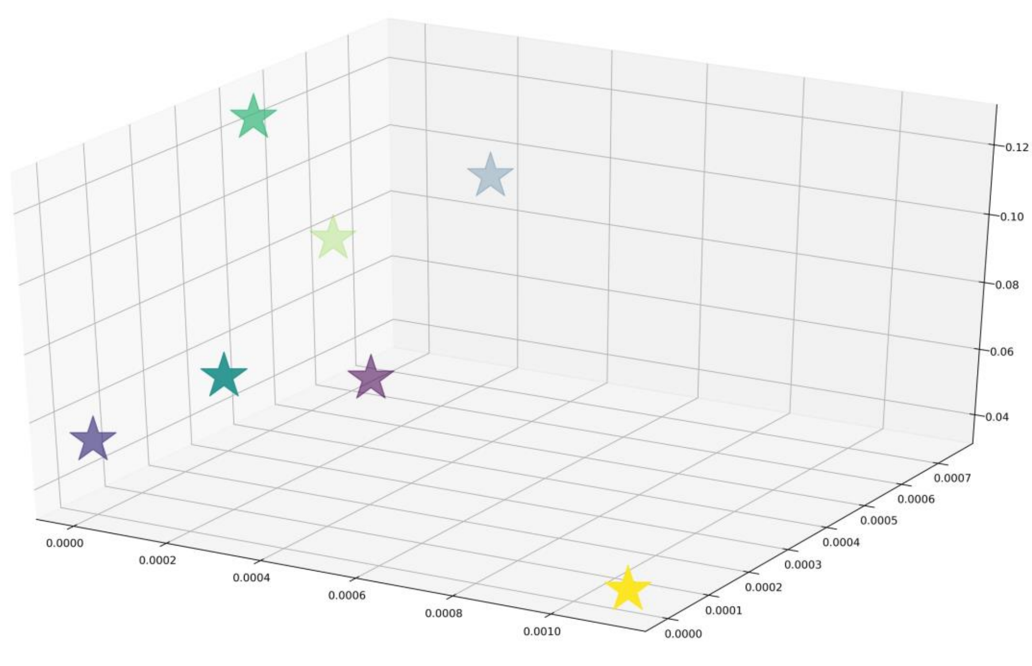
7. Next, a new Pandas data frame was created by calculating the frequency of each restaurant type for each city. This dataset will be used for the K-means clustering. Clustering typically does not work on categorical data. Transforming the data set from the categories to a distribution of values meets the requirements needed to perform the K-means clustering.

8. Then, another Pandas data frame was created to show “Top 25 Restaurant Types per City” for visualization. These were determined off of the frequencies found in the Pandas data frame from the above step.

Address	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
New York City, New York	Pizza Place	Bakery	Italian Restaurant	Café	Donut Shop
Los Angeles, California	Sushi Restaurant	American Restaurant	Italian Restaurant	Café	Mexican Restaurant
Chicago, Illinois	New American Restaurant	Italian Restaurant	Pizza Place	Sandwich Place	Donut Shop
Houston, Texas	Mexican Restaurant	Burger Joint	Pizza Place	Café	American Restaurant
Phoenix, Arizona	Pizza Place	Burger Joint	American Restaurant	Mexican Restaurant	Italian Restaurant
Philadelphia, Pennsylvania	Pizza Place	Italian Restaurant	American Restaurant	Café	Breakfast Spot
San Antonio, Texas	Mexican Restaurant	Burger Joint	Pizza Place	Bakery	American Restaurant
San Diego, California	Seafood Restaurant	American Restaurant	Pizza Place	Mexican Restaurant	Sandwich Place
Dallas, Texas	New American Restaurant	American Restaurant	Steakhouse	Pizza Place	Burger Joint
San Jose, California	Sandwich Place	Pizza Place	Mexican Restaurant	Korean Restaurant	Breakfast Spot
Austin, Texas	Pizza Place	Taco Place	Sandwich Place	Burger Joint	Food Truck

9. Next, the scikit-learn library was used to call the K-means method. The scikit-learn library provides many machine learning algorithms available for Data scientists and Python programmers to use. For this analysis, K-means clustering was used to group the cities into clusters based upon the restaurant type similarities. The main function K-means clustering provides is to partition a certain number of observations into a defined number of clusters (k). Each observation will be classified as being part of a certain cluster where its value or distribution is the closest to the nearest mean of the clusters. One does not know the optimal number of clusters to choose. Therefore, one could assign too many or too few k clusters. Visualization of those clusters and observing the distances between the cluster means is key towards defining the right number of clusters. Several rounds of clustering was perform to determine the optimal number of clusters. The optimal number of clusters was determined to be 7.

Figure 2: The Visual Inspection of the 7 Cluster Means



10. Now each city was assigned to one of the seven clusters by the K-means clustering algorithm belonged to a cluster, the next step was to visualize the cluster assignment on the map of The United States. Another folium map was created in order to provide this visualization. The map is shown in the results section.

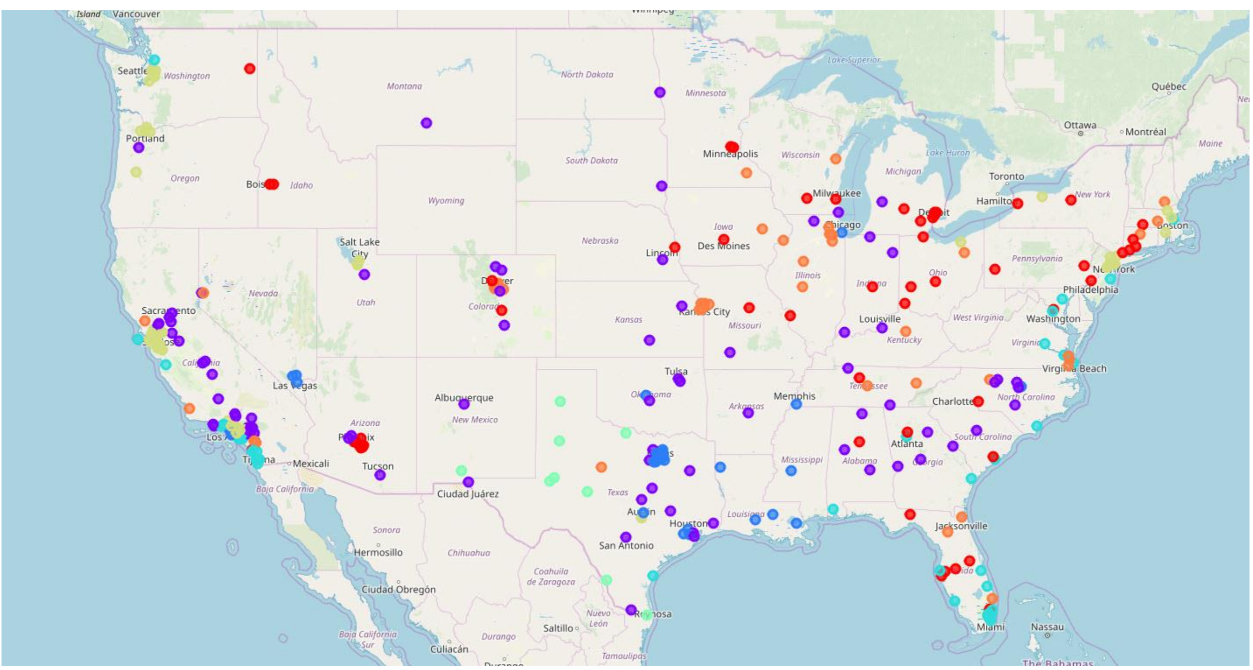
11. The final step was providing a visual distribution of each cluster by using pie charts for the top 10 restaurant types per cluster. These visuals are show in the results section. Interpretations of these results will be reported in the Results section and Discussed in the discussion section.

Results:

The main goal of the analysis was to provide objective, useful information to a potential restaurant investor on what are the most frequent types of cuisines available for each of the top 300 American cities (population size based) and if there are any similarities between the cities based upon similar cuisine distributions. The investor could then create more informed decisions on where to invest and what types of cuisines they should provide as part of their restaurant chains.

Below is the United States map view locations for the cities used in this analysis. The color represents the K-Means cluster that each city belongs to. The clusters are based upon the most frequent cuisine type.

Figure 3: The Map of The United States illustrating the Clustering membership and distribution



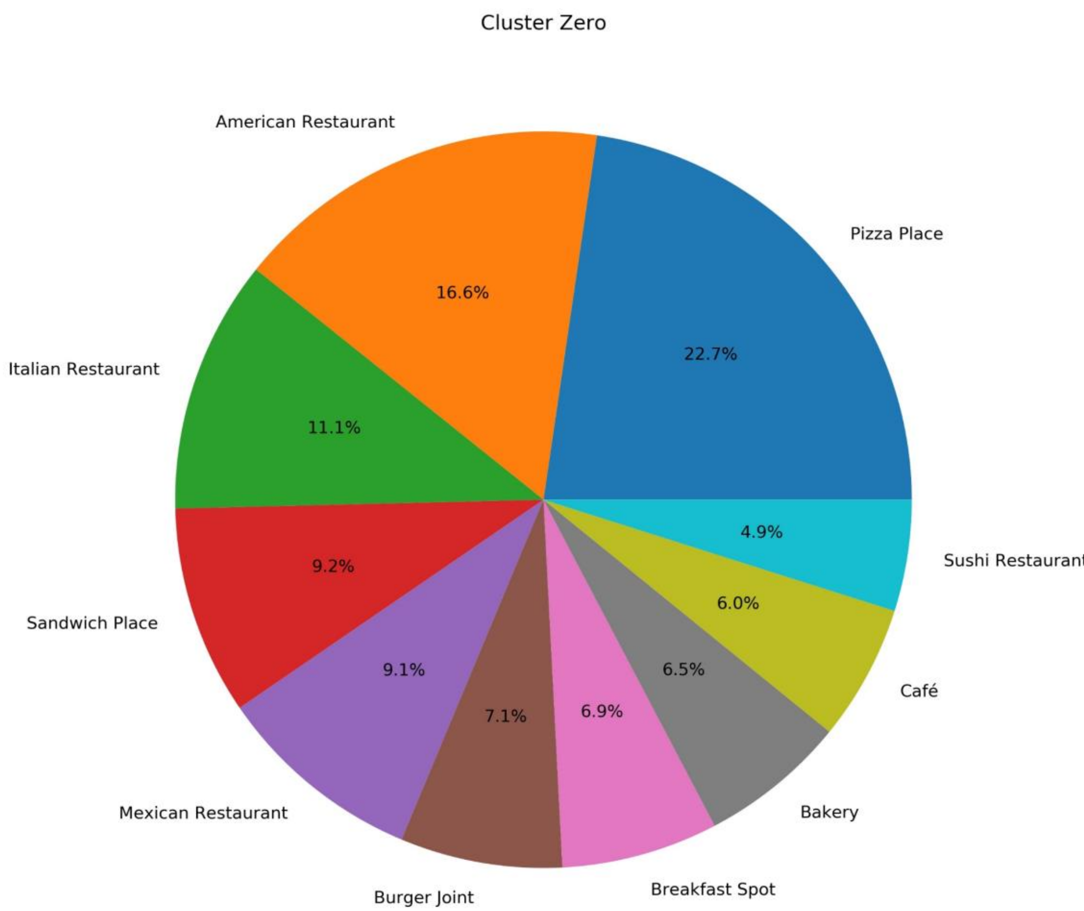
The legend is below:

Cluster 0 Cluster 1 Cluster 2 Cluster 3

Cluster 4 Cluster 5 Cluster 6

The pie charts below report the top 10 most frequent cuisine type for each of the 7 clusters. The list of the cities are also available for each cluster in the GitHub Repository.

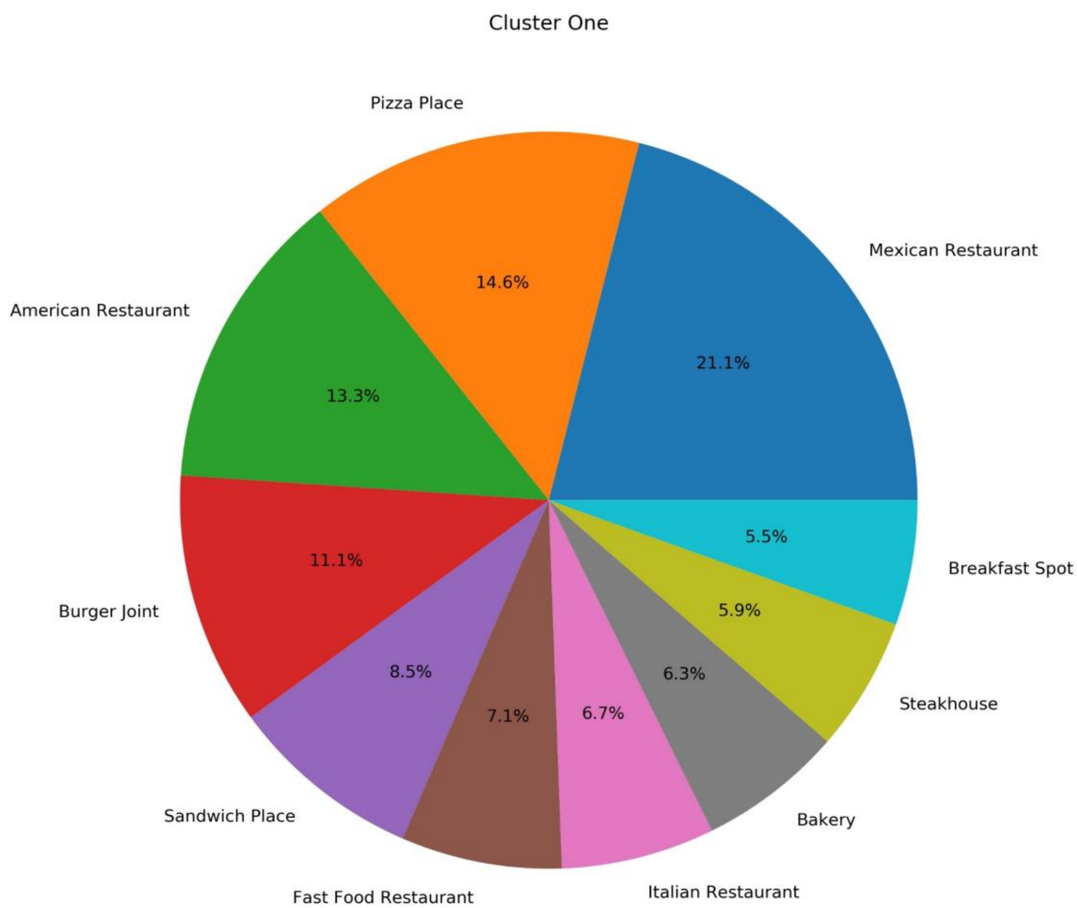
Figure 4: The Pie Chart Distribution of Restaurant Types for Cluster 0



Cluster 0 Cities List:

City	State	Population	Latitude	Longitude	City	State	Population	Latitude	Longitude
Birmingham	Alabama	209,880	33.5274	-86.799	Sterling Heights	Michigan	132,964	42.5812	-83.0303
Phoenix	Arizona	1,660,272	33.5722	-112.0901	Lansing	Michigan	118,427	42.7143	-84.5593
Tempe	Arizona	192,364	33.3884	-111.9318	Ann Arbor	Michigan	121,890	42.2761	-83.7309
Glendale	Arizona	250,702	33.5331	-112.1899	Warren	Michigan	134,587	42.4929	-83.025
Scottsdale	Arizona	255,310	33.6843	-111.8611	Detroit	Michigan	672,662	42.383	-83.1022
Chandler	Arizona	257,165	33.2829	-111.8549	Clinton	Michigan	100,800	42.5903	-82.917
Mesa	Arizona	508,958	33.4019	-111.7174	Saint Paul	Minnesota	307,695	44.9489	-93.1041
Gilbert	Arizona	248,279	33.3103	-111.7431	Minneapolis	Minnesota	425,403	44.9633	-93.2683
Oceanside	California	176,080	33.2245	-117.3062	St. Louis	Missouri	302,838	38.6357	-90.2446
Boulder	Colorado	107,353	40.027	-105.2519	Columbia	Missouri	123,180	38.951561	-92.328638
Colorado Springs	Colorado	472,688	38.8673	-104.7607	Omaha	Nebraska	468,262	41.2644	-96.0451
Waterbury	Connecticut	108,093	41.5585	-73.0367	Buffalo	New York	256,304	42.8925	-78.8597
New Haven	Connecticut	130,418	41.3108	-72.925	Syracuse	New York	142,749	43.041	-76.1436
Stamford	Connecticut	129,775	41.0799	-73.546	Charlotte	North Carolina	872,498	35.2078	-80.831
Bridgeport	Connecticut	144,900	41.1874	-73.1958	Columbus	Ohio	892,533	39.9852	-82.9848
Washington, D.C.	District of Columbia	702,455	38.9041	-77.0172	Toledo	Ohio	274,975	41.6641	-83.5819
Tampa	Florida	392,890	27.9701	-82.4797	Cincinnati	Ohio	302,605	39.1402	-84.5058
Tallahassee	Florida	193,551	30.4551	-84.2534	Dayton	Ohio	140,640	39.7774	-84.1996
Coral Springs	Florida	133,507	26.2707	-80.2593	Philadelphia	Pennsylvania	1,584,138	40.0094	-75.1333
Orlando	Florida	285,713	28.4166	-81.2736	Allentown	Pennsylvania	121,433	40.5936	-75.4784
St. Petersburg	Florida	265,098	27.762	-82.6441	Pittsburgh	Pennsylvania	301,048	40.4398	-79.9766
Lakeland	Florida	110,516	28.0555	-81.9549	North Charleston	South Carolina	113,237	32.9178	-80.065
Sandy Springs	Georgia	108,797	33.9315	-84.3687	Nashville	Tennessee	669,053	36.1718	-86.785
Boise	Idaho	228,790	43.6002	-116.2317	Spokane	Washington	219,190	47.6669	-117.4333
Meridian	Idaho	106,804	43.6142	-116.3989	Madison	Wisconsin	258,054	43.0878	-89.4299
Indianapolis	Indiana	867,125	39.7767	-86.1459	Milwaukee	Wisconsin	592,025	43.0633	-87.9667
Des Moines	Iowa	216,853	41.5726	-93.6102	Springfield	Massachusetts	155,032	42.1155	-72.54

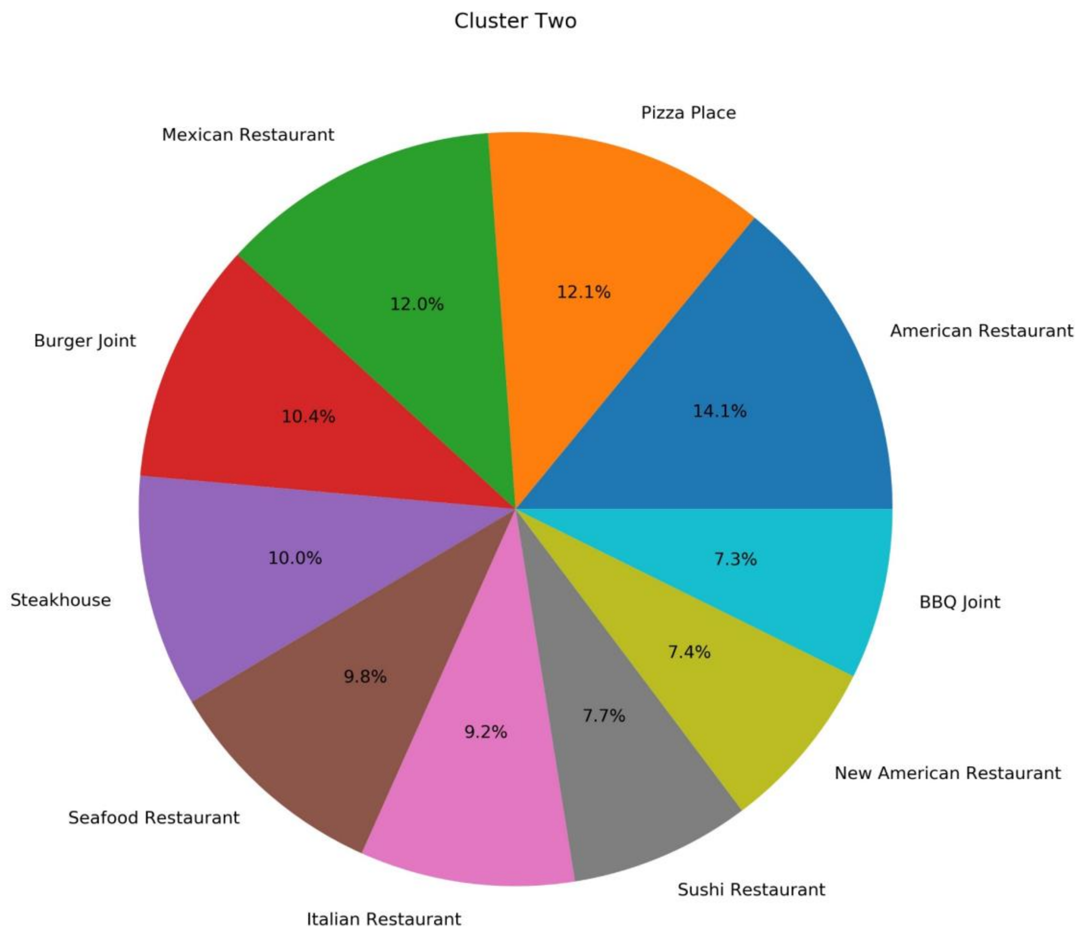
Figure 5: The Pie Chart Distribution of Restaurant Types for Cluster 1



Cluster 1 City List:

City	State	Population	Latitude	Longitude	City	State	Population	Latitude	Longitude
Tuscaloosa	Alabama	101,113	33.2065	-87.5346	Rockford	Illinois	146,526	42.2588	-89.0646
Montgomery	Alabama	198,218	32.3472	-86.2661	Fort Wayne	Indiana	267,633	41.0882	-85.1439
Huntsville	Alabama	197,318	34.699	-86.673	South Bend	Indiana	101,860	41.6769	-86.269
Anchorage	Alaska	291,538	61.1743	-149.2843	Evansville	Indiana	117,963	37.9877	-87.5347
Surprise	Arizona	138,161	33.6706	-112.4527	Wichita	Kansas	389,255	37.6907	-97.3459
Tucson	Arizona	545,975	32.1531	-110.8706	Topeka	Kansas	125,904	39.0347	-95.6962
Peoria	Arizona	172,259	33.7862	-112.308	Louisville	Kentucky	620,118	38.1654	-85.6474
Little Rock	Arkansas	197,881	34.7254	-92.3586	Grand Rapids	Michigan	200,217	42.9612	-85.6556
Vacaville	California	100,154	38.3539	-121.9728	Springfield	Missouri	168,122	37.1942	-93.2913
Roseville	California	139,117	38.769	-121.3189	Billings	Montana	109,550	45.7885	-108.5499
Fairfield	California	116,884	38.2593	-122.0321	Lincoln	Nebraska	287,401	40.8105	-96.6803
Ontario	California	181,107	34.0394	-117.6042	Reno	Nevada	250,998	39.5491	-119.8499
Victorville	California	122,312	34.5277	-117.3536	Albuquerque	New Mexico	560,218	35.1056	-106.6474
Clovis	California	112,022	36.8282	-119.6849	Fayetteville	North Carolina	209,468	35.0828	-78.9735
Elk Grove	California	172,886	38.4146	-121.385	Durham	North Carolina	274,291	35.9811	-78.9029
Corona	California	168,819	33.862	-117.5655	Greensboro	North Carolina	294,722	36.0951	-79.827
Lancaster	California	159,053	34.6936	-118.1753	Cary	North Carolina	168,160	35.7809	-78.8133
Visalia	California	133,800	36.3273	-119.3289	High Point	North Carolina	112,316	35.99	-79.9905
Palmdale	California	156,667	34.591	-118.1054	Fargo	North Dakota	124,844	46.8652	-96.829
Rancho Cucamonga	California	177,751	34.1233	-117.5642	Tulsa	Oklahoma	400,669	36.1279	-95.9023
Oxnard	California	209,877	34.2023	-119.2046	Norman	Oklahoma	123,471	35.2406	-97.3453
Moreno Valley	California	209,050	33.9233	-117.2057	Broken Arrow	Oklahoma	109,171	36.0365	-95.781
Modesto	California	215,030	37.6375	-121.003	Salem	Oregon	173,442	44.9237	-123.0232
Rialto	California	103,440	34.1118	-117.3883	Columbia	South Carolina	133,451	34.0291	-80.898
Fresno	California	530,093	36.7836	-119.7934	Sioux Falls	South Dakota	181,883	43.5383	-96.732
Sacramento	California	508,529	38.5666	-121.4686	Chattanooga	Tennessee	180,557	35.066	-85.2484
Fontana	California	213,739	34.109	-117.4629	Clarksville	Tennessee	156,794	36.5664	-87.3452
Riverside	California	330,063	33.9381	-117.3932	Killeen	Texas	149,103	31.0777	-97.732
Stockton	California	311,178	37.9763	-121.3133	El Paso	Texas	682,669	31.8484	-106.427
Jurupa Valley	California	108,393	34.0026	-117.4676	Fort Worth	Texas	895,008	32.7815	-97.3467
Bakersfield	California	383,579	35.3212	-119.0183	Tyler	Texas	105,729	32.3173	-95.3059
Ventura	California	111,128	34.2678	-119.2542	League City	Texas	106,244	29.4901	-95.1091
San Bernardino	California	215,941	34.1416	-117.2936	College Station	Texas	116,218	30.5852	-96.2964
Centennial	Colorado	110,831	39.5906	-104.8691	Beaumont	Texas	118,428	30.0849	-94.1453
Pueblo	Colorado	111,750	38.2699	-104.6123	Pasadena	Texas	153,219	29.6586	-95.1506
Greeley	Colorado	107,348	40.4153	-104.7697	Waco	Texas	138,183	31.5601	-97.186
Fort Collins	Colorado	167,830	40.5482	-105.0648	Denton	Texas	138,541	33.2166	-97.1414
Augusta	Georgia	196,939	33.3655	-82.0734	McAllen	Texas	143,433	26.2322	-98.2464
Macon	Georgia	153,095	32.8088	-83.6942	San Antonio	Texas	1,532,233	29.4724	-98.5251
Columbus	Georgia	194,160	32.5102	-84.8749	Provo	Utah	116,702	40.2453	-111.6448
Athens	Georgia	125,964	33.9496	-83.3701	Kenosha	Wisconsin	100,164	42.5822	-87.8456

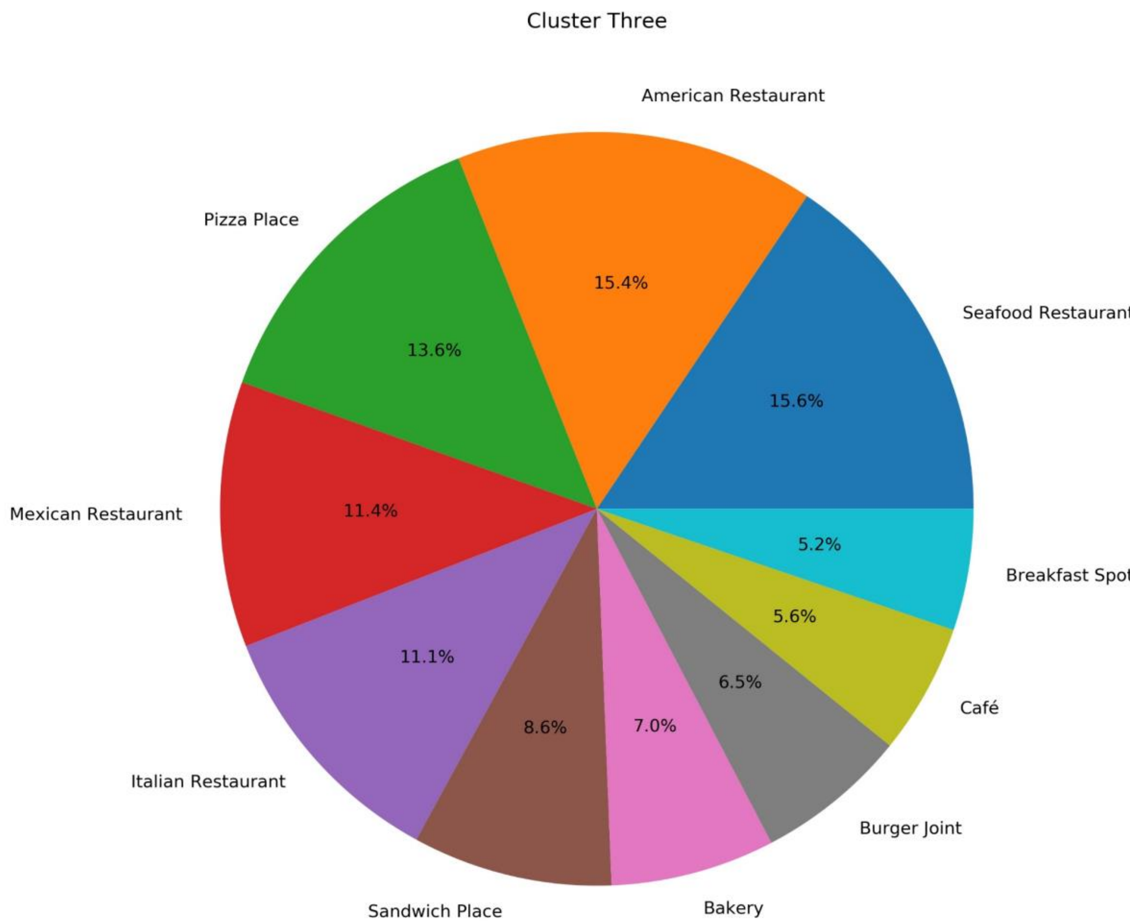
Figure 6: The Pie Chart Distribution of Restaurant Types for Cluster 2



Cluster 2 City List:

City	State	Population	Latitude	Longitude
Los Angeles	California	3,990,456	34.0194	-118.4108
Inglewood	California	109,419	33.9561	-118.3443
Torrance	California	145,182	33.835	-118.3414
Chicago	Illinois	2,705,994	41.8376	-87.6818
Lafayette	Louisiana	126,143	30.2074	-92.0285
Shreveport	Louisiana	188,987	32.4669	-93.7922
New Orleans	Louisiana	391,006	30.0534	-89.9345
Baton Rouge	Louisiana	221,599	30.4422	-91.1309
Jackson	Mississippi	164,422	32.3158	-90.2128
Las Vegas	Nevada	644,644	36.2292	-115.2601
Henderson	Nevada	310,390	36.0097	-115.0357
North Las Vegas	Nevada	245,949	36.2857	-115.0939
Raleigh	North Carolina	469,298	35.8306	-78.6418
Oklahoma City	Oklahoma	649,021	35.4671	-97.5137
Memphis	Tennessee	650,618	35.1028	-89.9774
Arlington	Texas	398,112	32.7007	-97.1247
Houston	Texas	2,325,502	29.7866	-95.3909
Sugar Land	Texas	118,600	29.5994	-95.6142
Richardson	Texas	120,981	32.9723	-96.7081
Pearland	Texas	122,149	29.5558	-95.3231
Dallas	Texas	1,345,047	32.7933	-96.7665
Round Rock	Texas	128,739	30.5252	-97.666
Carrollton	Texas	136,879	32.9884	-96.8998
Frisco	Texas	188,170	33.1554	-96.8226
McKinney	Texas	191,645	33.1985	-96.668
Grand Prairie	Texas	194,614	32.6869	-97.0211
Lewisville	Texas	106,586	33.0466	-96.9818
Irving	Texas	242,242	32.8577	-96.97
Garland	Texas	242,507	32.9098	-96.6303
Plano	Texas	288,061	33.0508	-96.7479
Mesquite	Texas	142,816	32.7629	-96.5888
Allen	Texas	103,383	33.0997	-96.6631

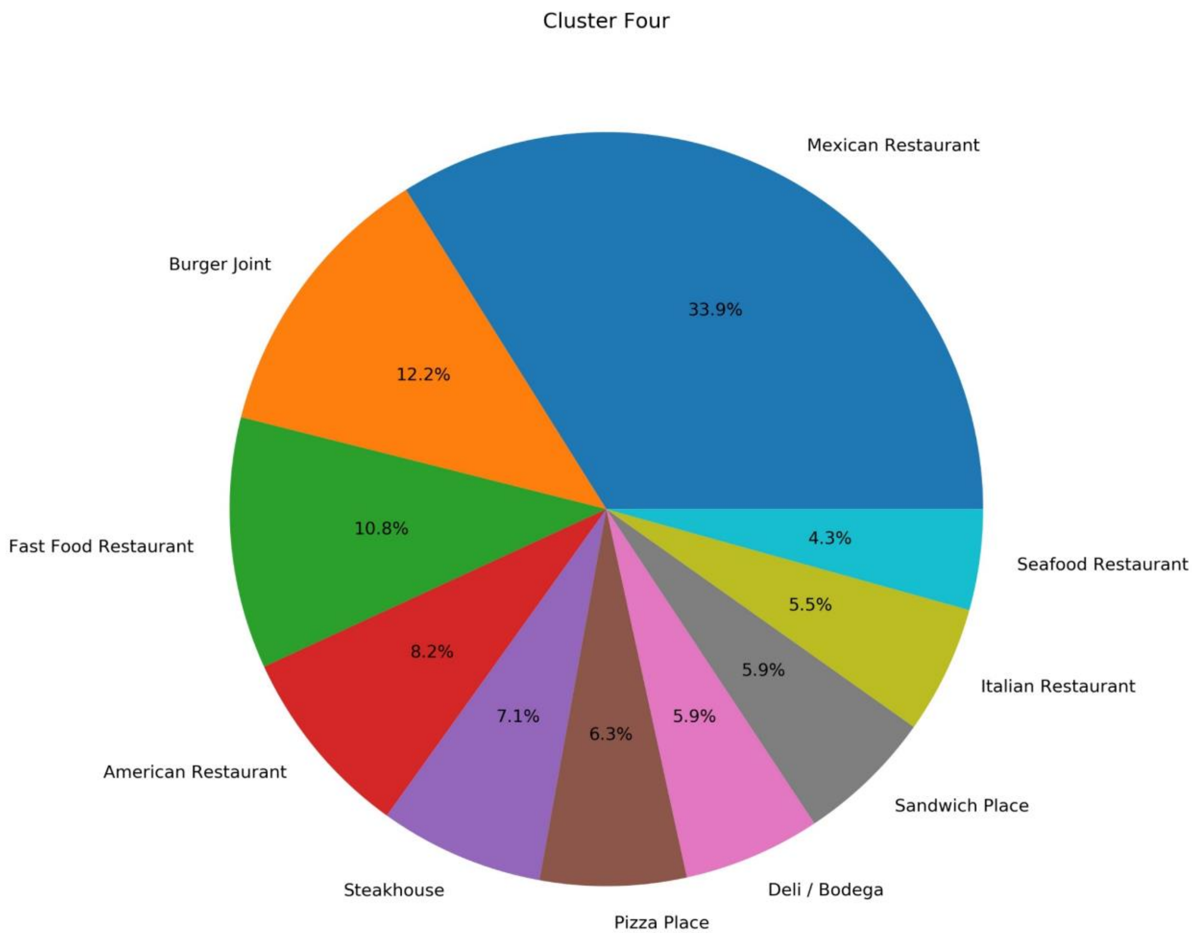
Figure 7: The Pie Chart Distribution of Restaurant Types for Cluster 3



Cluster 3 City List:

City	State	Population	Latitude	Longitude
Mobile	Alabama	189,572	30.6684	-88.1002
San Diego	California	1,425,976	32.8153	-117.135
Costa Mesa	California	113,615	33.6659	-117.9123
Carlsbad	California	115,877	33.1239	-117.2828
Simi Valley	California	125,851	34.2669	-118.7485
Thousand Oaks	California	127,690	34.1933	-118.8742
Orange	California	139,484	33.787	-117.8613
Escondido	California	152,213	33.1331	-117.074
Salinas	California	156,259	36.6902	-121.6337
El Cajon	California	103,241	32.8017	-116.9604
Garden Grove	California	172,646	33.7788	-117.9605
Huntington Beach	California	200,641	33.6906	-118.0093
Vista	California	101,224	33.1895	-117.2386
Chula Vista	California	271,651	32.6277	-117.0152
San Francisco	California	883,305	37.7272	-123.0322
Irvine	California	282,572	33.6784	-117.7713
Santa Ana	California	332,725	33.7363	-117.883
Hialeah	Florida	238,942	25.8699	-80.3029
Davie	Florida	106,558	26.0791	-80.285
Pompano Beach	Florida	111,954	26.2416	-80.1339
Miami Gardens	Florida	113,069	25.9489	-80.2436
Palm Bay	Florida	114,194	27.9856	-80.6626
Clearwater	Florida	116,478	27.9789	-82.7666
Miramar	Florida	140,823	25.977	-80.3358
Miami	Florida	470,914	25.7752	-80.2086
Hollywood	Florida	154,823	26.031	-80.1646
Port St. Lucie	Florida	195,248	27.2806	-80.3883
Cape Coral	Florida	189,343	26.6432	-81.9974
Fort Lauderdale	Florida	182,595	26.1412	-80.1467
Pembroke Pines	Florida	172,374	26.021	-80.3404
Savannah	Georgia	145,862	32.0025	-81.1536
Atlanta	Georgia	498,044	33.7629	-84.4227
Baltimore	Maryland	602,495	39.3	-76.6105
Boston	Massachusetts	694,583	42.332	-71.0202
Lakewood	New Jersey	104,157	40.0771	-74.2004
Wilmington	North Carolina	122,607	34.2092	-77.8858
Charleston	South Carolina	136,208	32.8179	-79.959
Corpus Christi	Texas	326,554	27.7543	-97.1734
Newport News	Virginia	178,626	37.0762	-76.522
Richmond	Virginia	228,783	37.5314	-77.476
Virginia Beach	Virginia	450,189	36.78	-76.0252
Alexandria	Virginia	160,530	38.8201	-77.0841
Everett	Washington	111,262	47.9566	-122.1914

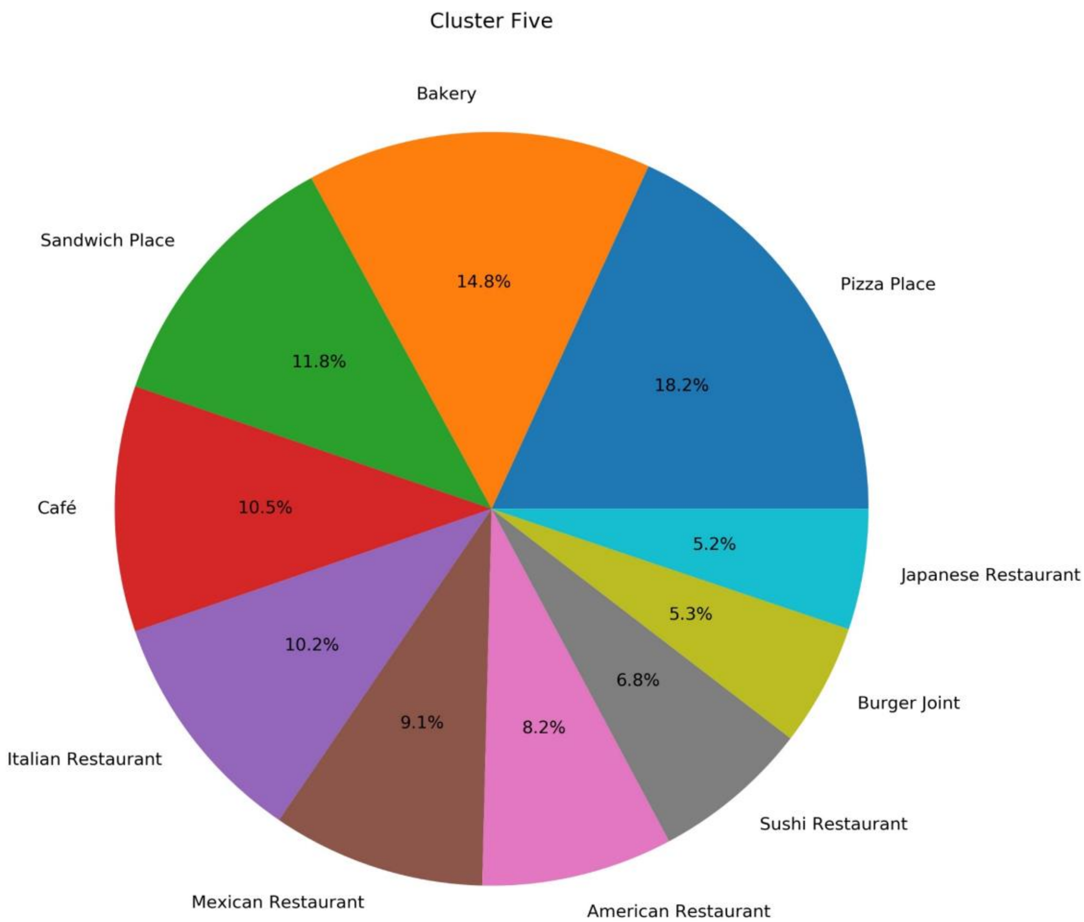
Figure 8: The Pie Chart Distribution of Restaurant Types for Cluster 4



Cluster 4 City List:

City	State	Population	Latitude	Longitude
Las Cruces	New Mexico	102,926	32.3264	-106.7897
Laredo	Texas	261,639	27.5604	-99.4892
Lubbock	Texas	255,885	33.5656	-101.8867
Amarillo	Texas	199,924	35.1999	-101.8302
Brownsville	Texas	183,392	25.9991	-97.455
Midland	Texas	142,344	32.0246	-102.1135
Odessa	Texas	120,568	31.8838	-102.3411
Wichita Falls	Texas	104,576	33.9067	-98.5259
San Angelo	Texas	100,215	31.4411	-100.4505

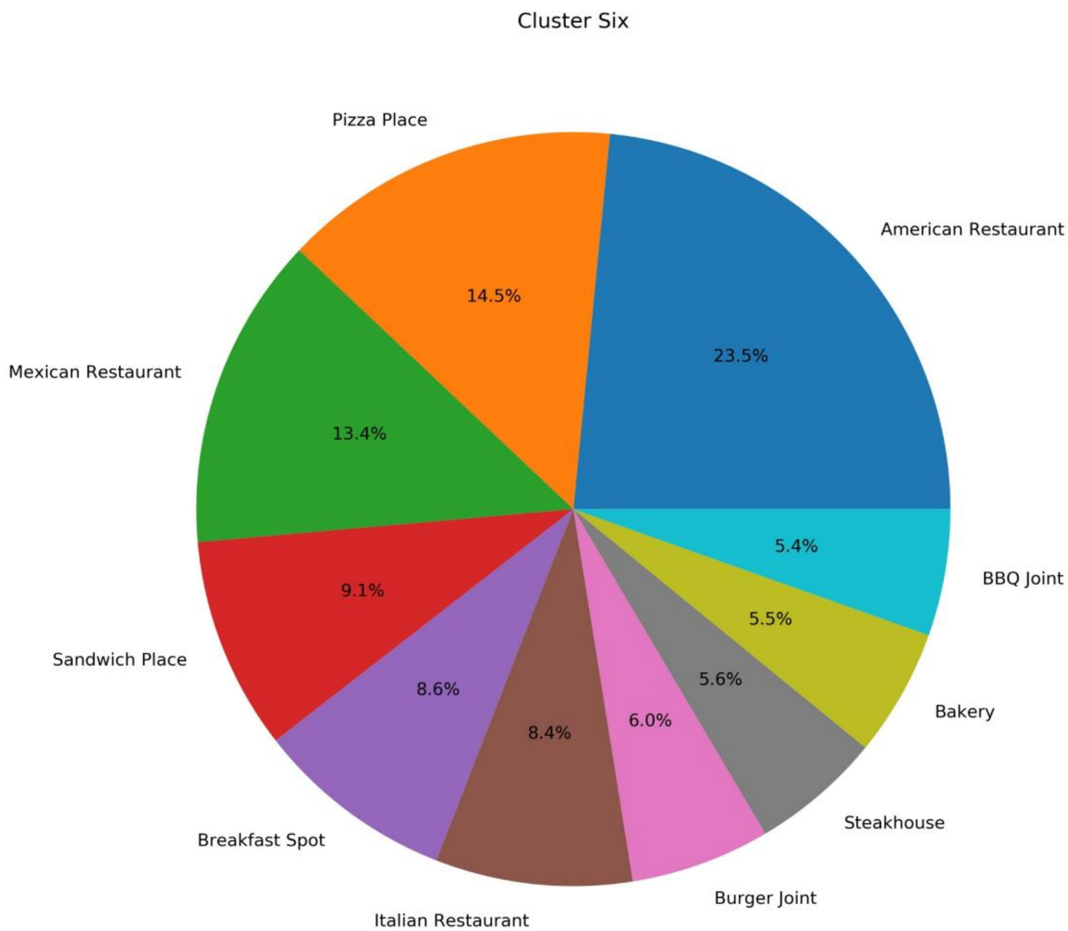
Figure 9: The Pie Chart Distribution of Restaurant Types for Cluster 5



Cluster 5 City List:

City	State	Population	Latitude	Longitude	City	State	Population	Latitude	Longitude
Santa Clarita	California	210,089	34.403	-118.5042	Honolulu	Hawaii	347,397	21.3243	-157.8476
Pasadena	California	141,371	34.1606	-118.1396	Lowell	Massachusetts	111,670	42.639	-71.3211
Santa Clara	California	129,488	37.3646	-121.9679	Cambridge	Massachusetts	118,977	42.376	-71.1187
Pomona	California	152,361	34.0585	-117.7611	Elizabeth	New Jersey	128,885	40.6664	-74.1935
Sunnyvale	California	153,185	37.3858	-122.0263	Paterson	New Jersey	145,627	40.9148	-74.1628
Hayward	California	159,620	37.6287	-122.1024	Woodbridge	New Jersey	100,450	40.5607	-74.2927
Vallejo	California	121,913	38.1079	-122.264	Edison	New Jersey	100,693	40.504	-74.3494
Berkeley	California	121,643	37.867	-122.2991	Newark	New Jersey	282,090	40.7242	-74.1726
El Monte	California	115,586	34.0746	-118.0291	Jersey City	New Jersey	265,549	40.7114	-74.0648
Downey	California	112,269	33.9382	-118.1309	New York City	New York	8,398,748	40.6635	-73.9387
Glendale	California	201,361	34.1814	-118.2458	Rochester	New York	206,284	43.1699	-77.6169
Fullerton	California	139,640	33.8857	-117.928	Yonkers	New York	199,663	40.9459	-73.8674
Antioch	California	111,535	37.9791	-121.7962	Cleveland	Ohio	383,793	41.4785	-81.6794
Concord	California	129,688	37.9722	-122.0016	Hillsboro	Oregon	108,389	45.528	-122.9357
Richmond	California	110,146	37.9523	-122.3606	Gresham	Oregon	110,158	45.5023	-122.4416
Daly City	California	107,008	37.7009	-122.465	Portland	Oregon	653,115	45.537	-122.65
Anaheim	California	352,005	33.8555	-117.7601	Eugene	Oregon	171,245	44.0567	-123.1162
West Covina	California	106,311	34.0559	-117.9099	Providence	Rhode Island	179,335	41.8231	-71.4188
Oakland	California	429,082	37.7698	-122.2257	Austin	Texas	964,254	30.3039	-97.7544
Long Beach	California	467,354	33.8092	-118.1553	Renton	Washington	102,153	47.4761	-122.192
Norwalk	California	105,120	33.9076	-118.0835	Kent	Washington	129,618	47.388	-122.2127
San Mateo	California	105,025	37.5603	-122.3106	Vancouver	Washington	183,012	45.6349	-122.5957
Burbank	California	103,695	34.1901	-118.3264	Bellevue	Washington	147,599	47.5979	-122.1565
San Jose	California	1,030,119	37.2967	-121.8189	Seattle	Washington	744,955	47.6205	-122.3509
Fremont	California	237,807	37.4945	-121.9412	Tacoma	Washington	216,279	47.2522	-122.4598
West Valley City	Utah	136,401	40.6885	-112.0118					
West Jordan	Utah	116,046	40.6024	-112.0008					
Salt Lake City	Utah	200,591	40.7769	-111.931					

Figure 10: The Pie Chart Distribution of Restaurant Types for Cluster 6



Cluster 6 City List:

City	State	Population	Latitude	Longitude
Santa Maria	California	107,408	34.9332	-120.4438
Temecula	California	114,742	33.4931	-117.1317
Murrieta	California	114,985	33.5721	-117.1904
Santa Rosa	California	177,586	38.4468	-122.7061
Thornton	Colorado	139,436	39.9194	-104.9428
Denver	Colorado	716,492	39.7619	-104.8811
Aurora	Colorado	374,114	39.688	-104.6897
Westminster	Colorado	113,479	39.8822	-105.0644
Arvada	Colorado	120,492	39.8337	-105.1503
Lakewood	Colorado	156,798	39.6989	-105.1176
Hartford	Connecticut	122,587	41.7659	-72.6816
West Palm Beach	Florida	111,398	26.7464	-80.1251
Gainesville	Florida	133,857	29.6788	-82.3461
Jacksonville	Florida	903,889	30.3369	-81.6616
Naperville	Illinois	148,304	41.7492	-88.162
Joliet	Illinois	148,099	41.5177	-88.1488
Peoria	Illinois	111,388	40.7515	-89.6174
Elgin	Illinois	111,683	42.0396	-88.3217
Springfield	Illinois	114,694	39.7911	-89.6446
Aurora	Illinois	199,602	41.7635	-88.2901
Davenport	Iowa	102,085	41.5541	-90.604
Cedar Rapids	Iowa	133,174	41.967	-91.6778
Kansas City	Kansas	152,958	39.1225	-94.7418
Olathe	Kansas	139,605	38.8843	-94.8195
Overland Park	Kansas	192,536	38.889	-94.6906
Lexington	Kentucky	323,780	38.0407	-84.4583
Worcester	Massachusetts	185,877	42.2695	-71.8078
Rochester	Minnesota	116,961	44.0154	-92.4772
Independence	Missouri	116,925	39.0855	-94.3521
Kansas City	Missouri	491,918	39.1251	-94.551
Sparks	Nevada	104,246	39.5544	-119.7356
Manchester	New Hampshire	112,525	42.9849	-71.4441
Winstonâ€“Salem	North Carolina	246,328	36.1027	-80.261
Akron	Ohio	198,006	41.0805	-81.5214
Murfreesboro	Tennessee	141,344	35.8522	-86.416
Knoxville	Tennessee	187,500	35.9707	-83.9493
Abilene	Texas	122,999	32.4545	-99.7381
Chesapeake	Virginia	242,634	36.6794	-76.3018
Norfolk	Virginia	244,076	36.923	-76.2446
Hampton	Virginia	134,313	37.048	-76.2971
Green Bay	Wisconsin	104,879	44.5207	-87.9842

Discussion:

Assumptions, Observations and Considerations:

There are a few assumptions, consideration and limitations that must be disclosed for consideration and as part of this analysis.

1. There was a hard limit of 100 restaurants returned for the Foursquare API used. Larger cities could have a different distribution of restaurant types. The sample size of 100 taken could be too small and may not be representative of the total restaurant types.

2. Restaurant Type is assumed to represent Restaurant Cuisine. This could be a bad assumption. For example, a Pizza Place could be considered as an Italian Restaurant. A Burger Joint could be considered a Sandwich Place. Both of those types could be considered as American Cuisine.

3. There could be some cultural or subcultural biases introduced into the models based upon city locations. The cluster colors visualized on the Map show these potential biases. An example of this could be with Cluster 4. According to the data, 33% of the restaurant types is of Mexican cuisine. All of the cities in this cluster are also close to the American/Mexican border. However, this could be viewed as advantageous though. Perhaps the client wants to set up several chains of a given cuisine and leverage the close proximity to gain a more collective localized response from customers before expanding into other regions of the country.

4. There was no information on the ratings of the restaurants and the average prices. One could assume that a steakhouse would be more expensive than a fast food place. However, the demographics and sub-city (neighborhood) demographics of the areas around the cities could have a huge impact on right size pricing and the quality of the cuisine the Client's restaurants would potentially provide

Suggestions and Observations:

These are subjective suggestions and are based upon observing the results more closely.

1. Avoid Cities in Cluster 4. The number of cities belonging to this cluster is small. Additionally, although the population size was not a factor considered in this analysis, the combined population for this cluster is rather small in comparison to the other clusters.

2. Combine Italian and Pizza Cuisines.

3. Combine Burger and Sandwich Cuisines

4. Combine Bakery, Breakfast, and Café Cuisines

5. Japanese and Sushi pair well together

6. Seafood and Steakhouses typically pair well together.

7. All other suggestions are under the assumption that the client does not want to immediately have a lot of competition but some awareness and acceptance of that particular cuisine.

8. For Cluster 0, a Sushi Restaurant could be successful. Sushi is in the top 10 most frequent restaurants in this cluster representing about 5% of the sample taken. There are many cities belonging to this cluster and they are somewhat randomly spread out across with a bias towards the Eastern part of the country. The states of Arizona, Michigan, and Florida could be 3 localized environments to try opening up a few Sushi restaurants to test the performance locally before spreading out more.

9. For Cluster 1, a Steakhouse could be successful. Steakhouse type represents about 6% of the sample types taken for this cluster. The cities are well spread out across the country demonstrating the acceptance of a Steakhouse restaurants but also localized clusters of cities in certain states. For example, in California, North Carolina, and Texas (where beef is really popular) could be great localized starts for opening up a new Steakhouse chain.

10. For Cluster 2, the bakery, café, and breakfast types were not listed in the top 10most frequent types. The number of cities belonging to this cluster is somewhat small in comparison to other clusters. However, there are several of the top largest cities in this cluster such as Dallas, Houston, Chicago, and Los Angeles. The potential population pool is quite significant. There is also a localization effect based upon city member locations. Texas would be the perfect spot to take the risk and open up several breakfast/bakery/café style restaurants. If the restaurant is successful in Dallas, Houston, and some of the surrounding suburbs then the client could quickly set up restaurants in LA and Chicago.

11. For Cluster 3, seafood is the top cuisine listed. Given the previous suggestion of a steakhouse pairs well with seafood, the suggestion would be to open up a steakhouse for the cities in this cluster. Again, steakhouses are not in the top 10 just as before with the breakfast in cluster 2, but the great pairing of seafood and steak leads towards this suggestion. Examining the city makeup for this cluster, it makes sense that seafood would be the top cuisine. Most of these cities are coastal cities on the east or west sides of the country.

12. For Cluster 5, there is not a strong recommendation for this cluster. Given that sushi and Japanese cuisine were in the top 10 types, the client could introduce additional seafood restaurants to the cities in this cluster. Most of the cities in this cluster are in California. The largest city based upon population size (New York City) is also a member of this cluster.

13. For Cluster 6, the suggestion to the client would be to open up a café chain of restaurants. Perhaps they could combine it with a breakfast and bakery types as well. Bakeries and Breakfast types represent approximately 13% of the total sample size taken for this cluster. There are a couple of localized clusters that the client could experiment in to see if they could gain momentum in a certain region of the country. The states of Illinois or Colorado would be great places to try. The advantage of Illinois could be the cities in this cluster are very close to Chicago.

Conclusion:

This analysis was performed in order to provide objective information and analysis on where a restaurant investor could open up new restaurants and what cuisine they should sever to the customers. In this analysis, the top 300+ American cities based upon population sizes were observed. The data collected was 100 (if available) restaurants and the respective types. From this dataset, clustering was performed to demonstrate relationships between cities and provide the most common types of restaurants. The results of this analysis recommends the following strategy for the investor.

It is recommended that the restaurant investor focuses on a breakfast/café/bakery style restaurant opening up stores in the Dallas and surrounding areas, Houston and the surrounding areas, or Chicago and the surrounding areas. The idea behind this approach is to first to gain localized credibility before expanding into other areas in the country.