IBM Applied Data Science Capstone

Coursera Project

A Comparison of America's Large Cities In Terms of Access to Ethnic Foods

Fereshteh Bashiri

November 2019

1 Introduction

The United States of America has been one of the primary, if not the most prominent, destinations for many immigrants in the past 200 years. People from all over the world have come to the US and built a new home for themselves. Some might have come seeking freedom of speech or religion, and others might have come hoping for the American dream and greater economic opportunities. While the reason behind immigration might be different from time to time, and for one person to another, there is a collective experience for most immigrants: People come from different cultures and go through great hardship and emotional distress. Food is not just nutrients; it is a simple way to bring back memories and make people feel good. It is also a way to connect with local residents.

Over the past two hundred years, many immigrants have opted for large cities in the US, such as Chicago, NYC, Los Angeles, etc., resulting in abundant ethnic restaurants in major cities. In this project, we are going to explore some major US cities in terms of diversity and access to local cuisines. Also, we will pay attention to this question of whether one can trace the history of immigration to these cities. Can we find out whether a specific city had been a significant destination for people of a particular ethnicity? While such an exploration provides useful insights into the culture and common traditions in these areas, inherited from the first generations of immigrants, we will particularly look for tentative business opportunities. We want to find out what type of cuisine is not well-represented. Would it be economically justifiable to invest in opening a new ethnic restaurant?

1.1 Interested Audience

This project is in the interest of two groups. First, historians and social science research communities who are interested in the impact of immigration on American societies. The process of migration leads to mutual cultural exchange, one of which is the impact of immigrant culture on the target community. Second, business investors and entrepreneurs who are interested in the establishment of a new ethnic restaurant will be able to find out what type of cuisine is not well-represented. While investing in the food and restaurant industry in a large city can be fruitful, the kind of restaurant must be carefully selected to avoid losing out on competition with other city restaurants. The result of this project will provide useful insight into planning toward establishing a new ethnic restaurant.

2 Data

The focus of this project is on ethnic restaurants available in some major US cities. I use Foursquare API to obtain a list of restaurants within each city. In this project, which intends solely to practice data science in Python, I am using the non-premium version of APIs for developers. Therefore, Foursquare API only provides 100 restaurants per request, which I think is sufficient for our purpose. To preserve the flexibility of adding and removing cities of interest easily, I limit this project to cuisines within a specified radius of the center of each city. In such case, a list of neighborhoods in each city is not required.

The first information that we need is the geographical coordinates (latitude and longitude) of each city of interest, which can be obtained using the Nominatim geocoder. Then, exploring nearby venues of each location with a query of "restaurant" will give us a list of 100 restaurants. The hierarchy of venue categories by Foursquare is listed here. By looking at the hierarchy of categories, we realize that in order to filter ethnic cuisines from unrelated categories, such as coffee shops and bakeries, we need to keep those only that their venue category contains the term "Restaurant."

3 Method

3.1 Data Pre-processing

The input to this project is a list of cities of interest, i.e., New York, Boston, Philadelphia, Chicago, Houston, Denver, San Francisco, Seattle, Indianapolis, Madison, Albuquerque, Atlanta, Columbus, Phoenix, and Dallas. I used geographical coordinates obtained from Nominatim Geolocator with Foursquare API in order to request for a list of restaurants within 1.5 kilometer radius of each city. As I used free version of the Foursquare API, the number of venues returned after each request is limited to 100, which is enough for the purpose of this project. The next step is to put the data into a dataframe, and remove rows of data in which venue category is irrelevant to ethnic foods. The Foursquare API provides us 1500 venues within 15 cities of interest, which presents 93 unique categories. After the pre-processing step, we have 851 venues from 61 unique categories. It is important to note that Foursquare API returns a slightly

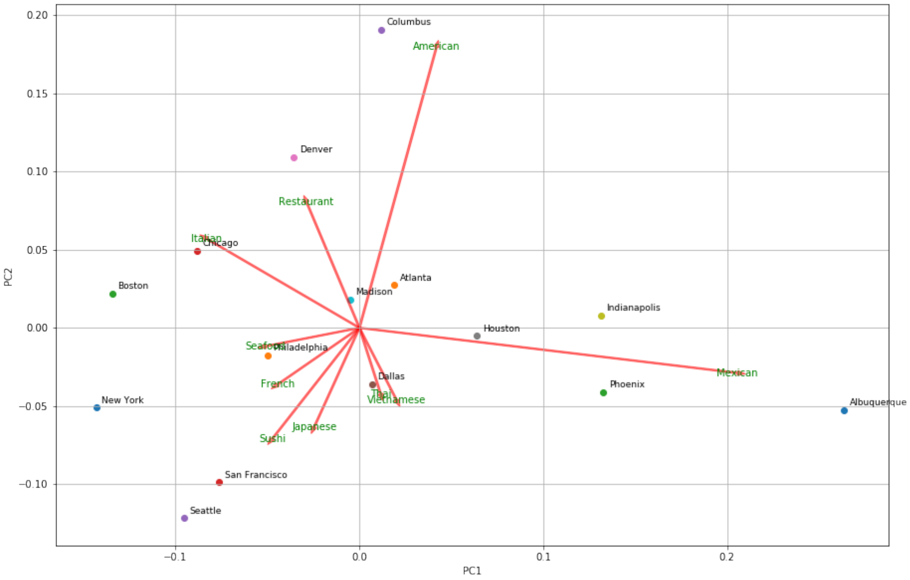

Figure 1: Visualization of the data on a 2D plane with PCA. The first two principal components and the 10 largest principal component loading vectors are illustrated on a biplot.

different set of restaurants every time we run the program, which may result in a different number of venues and unique categories.

3.2 Data Analysis

The first part of my analysis is to reorganize the data in a dataframe in which columns correspond to each unique venue categories, rows correspond to each city, and each element in the dataframe represents the average number of seeing a specific type of food in a city. From the information summarized at the end of the pre-processing step we already know that the reorganized data contains 15 rows and 61 columns. Also, we can expect that each row sums up to 1.0 and each value in the table is a float number between 0 and 1.

The next step in data analysis is to visualize the data and observe how cities are spread in the space. However, we are unable to visualize a 61-dimensional data. Therefore, I used Principal Component Analysis (PCA) – a linear method of dimensionality reduction – to project the data onto a 2D plane. The result is displayed in Figure 1 - a biplot in which principal component scores and the first 10 loading vectors with the largest magnitude are illustrated. Just as a brief observation, it is interesting how cities with a great number of Mexican residents are separated from other cities by spreading along Mexican loading vector, and San Francisco is in the direction of Sushi loading vector. I will discuss the figure in Section 4 in more details.

To obtain a better understanding of the above plot, I looked at the distribution of different cuisines at each city separately. In other words, I summarized the data in mini-tables (a table for each city) that each reports what are the most common types of restaurant in the city. For that purpose, I extracted top 5 most frequent cuisines and the average number of their appearance, as presented in Table 1.

	Albuquerque		Atlanta		Boston	
	Venue	Freq.	Venue	Freq.	Venue	Freq.
1	Mexican	0.36	Southern/Soul	0.15	Italian	0.22
2	American	0.11	American	0.15	Sea Food	0.19
3	Fast Food	0.07	Mexican	0.08	Restaurant	0.10
4	Vietnamese	0.06	Italian	0.06	American	0.08
5	Restaurant	0.06	Restaurant	0.06	French	0.07
	Chicago		Columbus		Dallas	
	Venue	Freq.	Venue	Freq.	Venue	Freq.
1	Restaurant	0.18	American	0.29	Sea Food	0.16
2	Italian	0.15	Restaurant	0.12	American	0.12
3	New American	0.15	Italian	0.10	Mexican	0.12
4	American	0.10	Mexican	0.08	Italian	0.07
5	Mediterranean	0.06	New American	0.04	Japanese	0.07
	Denver		Houston		Indianapolis	
	Venue	Freq.	Venue	Freq.	Venue	Freq.
1	American	0.22	Mexican	0.15	Mexican	0.22
2	Italian	0.16	American	0.15	New American	0.15
3	Restaurant	0.09	Vietnamese	0.10	American	0.15
4	New American	0.09	Italian	0.10	Southern/Soul	0.06
5	Mexican	0.07	New American	0.08	Italian	0.06
	Madison		New York		Philadelphia	
	Venue	Freq.	Venue	Freq.	Venue	Freq.
1	American	0.15	Italian	0.17	Italian	0.11
2	New American	0.12	Japanese	0.09	American	0.09
3	Italian	0.08	New American	0.09	Vegetarian/Vegan	0.09
4	Mexican	0.08	Restaurant	0.08	Mexican	0.07
5	Asian	0.08	Thai	0.06	Restaurant	0.07
	Phoenix		San Francisco		Seattle	
	Venue	Freq.	Venue	Freq.	Venue	Freq.
1	Mexican	0.22	Sushi	0.16	Vietnamese	0.10
2	American	0.10	New American	0.11	Sushi	0.10
3	New American	0.10	American	0.07	Italian	0.10
4	Thai	0.07	French	0.06	Sea Food	0.09
5	Restaurant	0.07	Japanese	0.06	French	0.09

Table 1: A summary of most common (top 5) restaurants in each city downtown and the average number of their appearance.

The above mini-tables reveal interesting information about each city. For example, we can see that more than 30% of restaurants in downtown Albuquerque are Mexican, and one who does not like sea food in general, may not enjoy living in Seattle!

To facilitate comparing cities, I then summarized the above table into a table of top 3 venue categories for each city, reported in Table 2.

	rank 1	**rank 2**	**rank 3**
Albuquerque	Mexican	American	Fast Food
Atlanta	Southern / Soul	American	Mexican
Boston	Italian	Seafood	Restaurant
Chicago	Restaurant	Italian	New American
Columbus	American	Restaurant	Italian
Dallas	Seafood	American	Mexican
Denver	American	Italian	Restaurant
Houston	Mexican	American	Vietnamese
Indianapolis	Mexican	New American	American
Madison	American	New American	Italian
New York	Italian	Japanese	New American
Philadelphia	Italian	American	Vegetarian / Vegan
Phoenix	Mexican	American	New American
San Francisco	Sushi	New American	American
Seattle	Vietnamese	Sushi	Italian

Table 2: A summary of most common (top 3) restaurants in each city.

After I looked at each city separately and observed principal components of the dataset, we investigated how cities will cluster into a number of groups. For that purpose, I used K-Means clustering method from the scikit-learn package. K-Means clustering is an unsupervised learning method that measures the similarity of samples, and clusters them into a pre-defined number of groups. I considered the default Euclidean similarity metric. The first step in k-Means clustering is to define the number of desired clusters and instantiate a k-Means objects. Next, the data was fitted into the object. The result is reported in Table 3 below.

In the end, to have a better understanding of the result of the clustering combined with where each city is located on the map of the US, I created a Figure 2, presented below.

Cluster 0	Cluster 1	Cluster 2	Cluster 3
Columbus	Atlanta	Boston	Albuquerque
Denver	Dallas	Chicago	Houston
-	Madison	New York	Indianapolis
-	Philadelphia	San Francisco	Phoenix
-	-	Seattle	-

Table 3: The report of clustering cities of interest into 4 clusters using K-Means clustering.

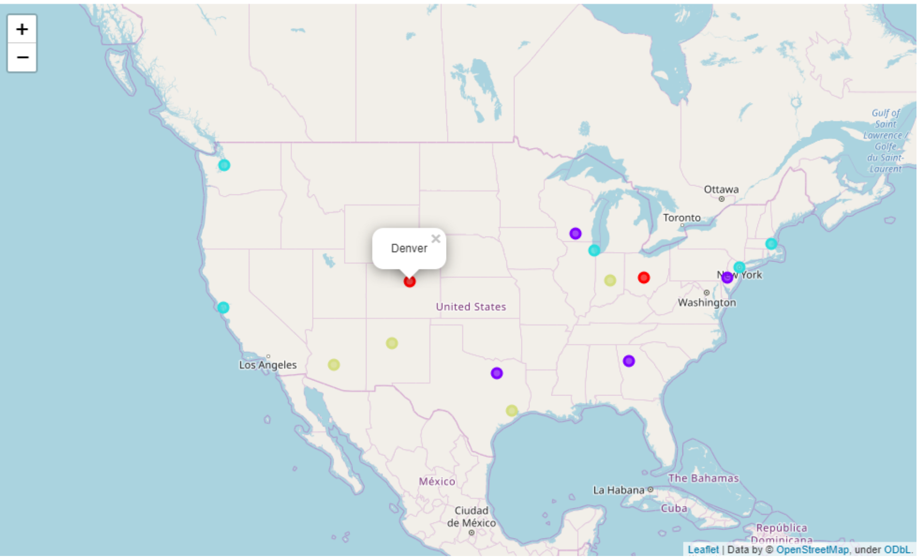

Figure 2: Visualization of the result of the clustering on a map of the US.

4 Results and Discussion

In previous sections, a list of US cities of interest was given for which I collected geographical latitude and longitude using Nominatim geolocator. I used the geographical location with Foursquare API to obtain a list of restaurants located in downtown area of each city. Then, I analyzed the collected data. In this section, I am going to discuss the results obtained in the above analysis.

The unfiltered collected data contains 1500 restaurants within 15 cities. The limited size of the data is the result of using the free version of Foursquare API, which for the purpose of this course is acceptable. Then, to remove venues that are not representative of ethnic foods, e.g., coffee shops, bakeries, I only kept venues in which the term ”restaurant” exists. The procedure leaves us with 851 venues within 61 categories spread in all cities. Next step in pre-processing the data is to put the data into a dataframe that illustrates the average number of each venue category serving in each city of interest. That is being done by creating dummy variables from all ethnic categories in the dataset. The final pre-processed data is a 15x61 dataframe, containing information on 61 ethnic categories within downtown area of 15 cities.

In the first part of the analysis, I used PCA to project the 61-dimensional data into a two-dimensional plane for visualization. The PCA finds orthogonal directions along which data has maximum variance. In a biplot figure, principal component scores and the top 10 loading vectors with the largest magnitude were plotted. In that figure, we see that the first loading vector is aligned with Mexican restaurant, which means that it is highly correlated with Mexican restaurant feature and corresponds to the average number of Mexican restaurant in town. We also see that the second loading vector places a high weight on American restaurants. Also, we observe that cities in which a specific ethnic restaurant is more common are mapped close to each other on the plot. For example, Seattle and San Francisco are mapped close along with the direction of Japanese and Sushi restaurants, while Houston, Indianapolis, Phoenix and Albuquerque are spread along the first component which corresponds to Mexican restaurant.

Next, I looked at the distribution of different cuisines in each city separately. In other words, for each city, I obtained most common cuisines by ranking them in decreasing order. The results (category of cuisine and it's average number) were printed out for observation in the program. Also, I created a DataFrame of top most common foods in each city. Looking at these tables provide an interesting insight into each city. We can see that Mexican restaurants are very popular in many cities, especially in southern-central part of the US. For example, more than 30% of restaurants in downtown Albuquerque are categorized as Mexican food. Also, Italian cuisines are very popular on the east coast including Boston, New York, Philadelphia, and Chicago. Such an observation is accordant with the fact that many Italians were established in major industrial cities between 1880 and 1920, during a period of industrialization and urbanization on early 20-th century. According to Wikipedia of Italian-Americans, New York and New Jersey and some smaller cities on the northeast of the US have the highest population of Italian-Americans. Another observation is that while sea food and east Asian cuisines are popular in cities on the west coast (e.g., San Francisco and Seattle), American foods are more common in the central parts of the country (e.g., Denver and Columbus).

For the sake of simplicity of analysis, I created a dataframe from the above information, summarizing top most common cuisines within each city. Looking at this dataframe provides an insight into the 2D PCA projection. We can see that first rank cuisines, which have significant higher rate of appearance in downtown area, play a major role in clustering cities on the projected plot.

To further analyze the data and cluster cities of interest into a number of groups based on their similarity, I used K-means clustering technique. I chose to cluster the data into 4 groups using the Euclidean similarity metric. The result that is printed out in the program. Here is the explanation of the clustering, considering the PCA projection and the dataframe of top most common cuisines: cluster 0 contains cities in which American food is very common, cluster 1 contains cities in which American food is moderately common followed closely by another type of food, cluster 3 contains cities in which Italian and sea food (including sushi and Japanese) are very common, and finally, cluster 3 contains cities in which Mexican food is very popular. I also, for visualizing the results, plotted a map of the US marking cities of interest with different colors according to the result of city clustering.

The above analysis discussed about what we can observe in the data, such as trace of immigration from other countries to the US and similarity of cities with respect to variaty of ethnic foods. A business opportunity is about what we cannot see in the data. From the perspective of an investor, the right question is: what is missed in the data? What type of food is not well-represented, which could be an opportunity for investment? My answer to such a question is "Indian food". Indian food is not among top 10 ethnic restaurants within the cities of interest (please refer to the red loading vectors plotted on the PCA projection). However, according to the Wikipedia page of Indian-Americans, the Asian-Indian is one of the well-represented ethnics in the US. According to the United States Census, it is one of the fastest growing ethnics in the US, and in 2010 they made up 0.9% of the US population. Even though little India have emerged in some large cities including New York, Chicago, Philadelphia, and Houston, Foursquare API returned only 11 Indian restaurants (1.29% of the filtered data) within downtown of 15 cities. Considering the population of Indian people, which is also growing caused by the technology boom, and the American people's interest in ethnic foods, I believe investing on an Indian chain restaurant in these large cities, especially those with higher number of Indian-American residents, is a wise bet.

5 Conclusion

The purpose of this project was to practice Python and Machine Learning techniques to dig into data, and solve real-world problems. For that purpose, we were introduced to developer APIs, got familiarized with multiple packages including Numpy, Pandas, Matplotlib, Folium, and Scikit-learn. While conducting this project, I investigated the diversity of ethinc foods in multiple large cities of the US. Most popular cuisines were recognized. Cities were compared regarding the variety and size of ethnic foods in their downtown area. We observed traces of immigration into the US through the perspective of food. Last, but not least, an investing opportunity in the food industry was identified.