Data Science Capstone Project

Exploring the Cities in Texas

Introduction

Texas is one of the most diverse states in the US in term of both ethnic and economic diversity. This makes it a prospective location for restaurant business especially Thai cuisine that has captivated people all over the world with its aromatic ingredients, spicy taste and unique flavor.

The following data analysis is conducted to explore the potential of opening a Thai restaurant in Texas. The analysis will be focused on the five most populous cities in Texas which are Houston, San Antonio, Dallas, Austin, and Fort Worth. Several factors will be taken into consideration such as demographic and economic information, people's interests, current competition and other common venues in the cities.

Target audiences are potential investors who are looking to open or expand a restaurant in Texas, as well as current restaurant customers who might want to explore the cities further for other restaurants. The analysis should also be beneficial for current and future data scientists who want to use exploratory data analysis, data visualization, and machine learning to find insights from the data and further answer other questions.

Data

1. Geographical Coordinates of Each City: Geopy

The latitude and longitude of each city were obtained using Geopy and stored in a dataframe.

2. Demographic and Economic Information: Web Scraping

Statistical information such as population, race, income and housing for each city was obtained from the United States Census Bureau website (https://www.census.gov/quickfacts/fact/table/US/PST045219). The data tables were read from the website into a dataframe using pandas. The parameters of interest are %Asian, Median household income, and Median gross rent.

3. Restaurants Search and Venues Exploring: Foursquare location data

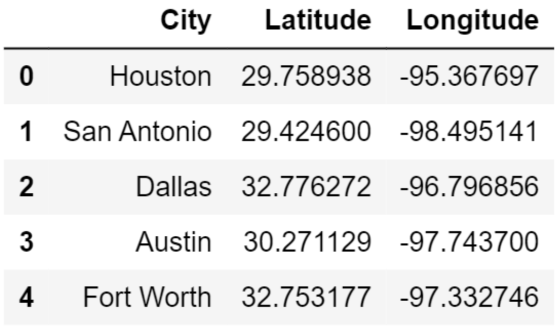
Foursquare API was used to search for restaurants to study the competition as well as explore common venues in the city to study the interests of people. The calls returned JSON files from which relevant data could be extracted. The limit for search is 50 places while the limit for explore is 100 venues. The radius of interest is within 10 km.

Methodology

1. Import and Explore Data

1.1 Obtain Geographical Coordinates

The focus of the analysis is on the five most populous cities in Texas which are Houston, San Antonio, Dallas, Austin, and Fort Worth. Since we are going to search and explore venues using the Foursquare API as well as plotting them on the map, we will first need the geographical coordinates of each city. The geographical coordinates of each city were obtained using Geopy, which transforms the address into latitude and longitude values. Their latitude and longitude values were then stored in a dataframe. The code is shown on the notebook on the Github repository and can be accessed through the attached link. The resulting dataframe is shown in Figure 1.

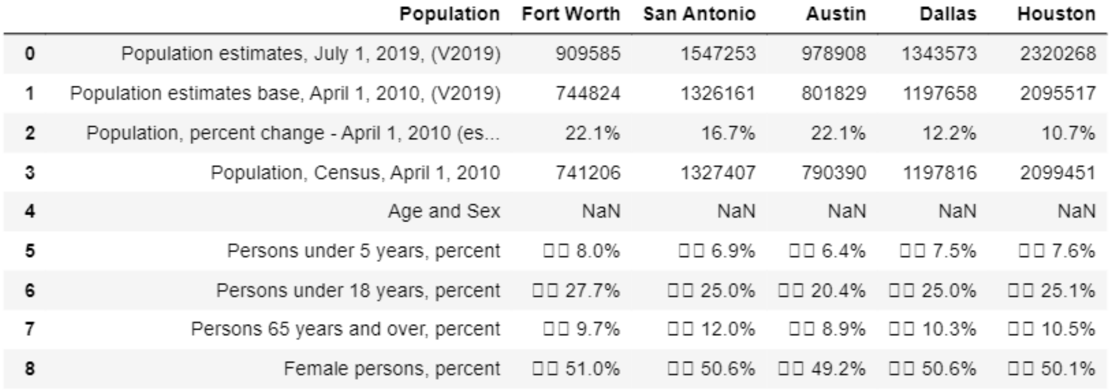


	City	Latitude	Longitude
0	Houston	29.758938	-95.367697
1	San Antonio	29.424600	-98.495141
2	Dallas	32.776272	-96.796856
3	Austin	30.271129	-97.743700
4	Fort Worth	32.753177	-97.332746

Figure 1: Geographical coordinates of cities obtained from Geopy

1.2 Import Demographic and Economic Information

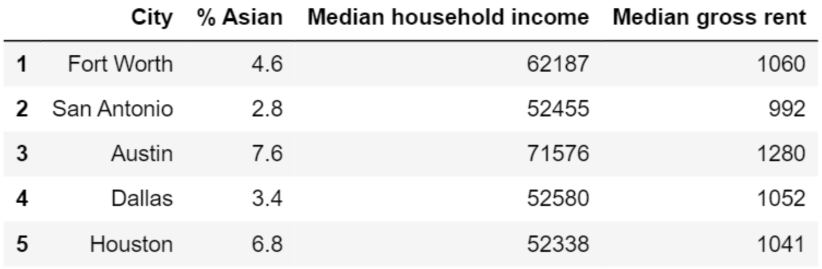
We can obtain some demographic and economic information of each city from web scraping. The United States Census Bureau website (https://www.census.gov/quickfacts/fact/table/US/PST045219) provides statistical information such as population, race, income and housing information for each city. The data tables were read from the website into a dataframe using pandas. Figure 2 shows the first few lines of the dataframe obtained as an example.



	Population	Fort Worth	San Antonio	Austin	Dallas	Houston
0	Population estimates, July 1, 2019, (V2019)	909585	1547253	978908	1343573	2320268
1	Population estimates base, April 1, 2010, (V2019)	744824	1326161	801829	1197658	2095517
2	Population, percent change - April 1, 2010 (es...	22.1%	16.7%	22.1%	12.2%	10.7%
3	Population, Census, April 1, 2010	741206	1327407	790390	1197816	2099451
4	Age and Sex	NaN	NaN	NaN	NaN	NaN
5	Persons under 5 years, percent	□□ 8.0%	□□ 6.9%	□□ 6.4%	□□ 7.5%	□□ 7.6%
6	Persons under 18 years, percent	□□ 27.7%	□□ 25.0%	□□ 20.4%	□□ 25.0%	□□ 25.1%
7	Persons 65 years and over, percent	□□ 9.7%	□□ 12.0%	□□ 8.9%	□□ 10.3%	□□ 10.5%
8	Female persons, percent	□□ 51.0%	□□ 50.6%	□□ 49.2%	□□ 50.6%	□□ 50.1%

Figure 2: Portion of data obtained from the United States Census Bureau website

In order to evaluate the demand, affordability, and economic conditions of the upcoming restaurant, some parameters such as %Asian, Median household income, and Median gross rent were taken into consideration. These parameters were selected and assigned into a new dataframe. We also removed special characters in the imported values and converted their data types into float and integer values for future plotting and analysis. The resulting dataframe is shown in Figure 3.

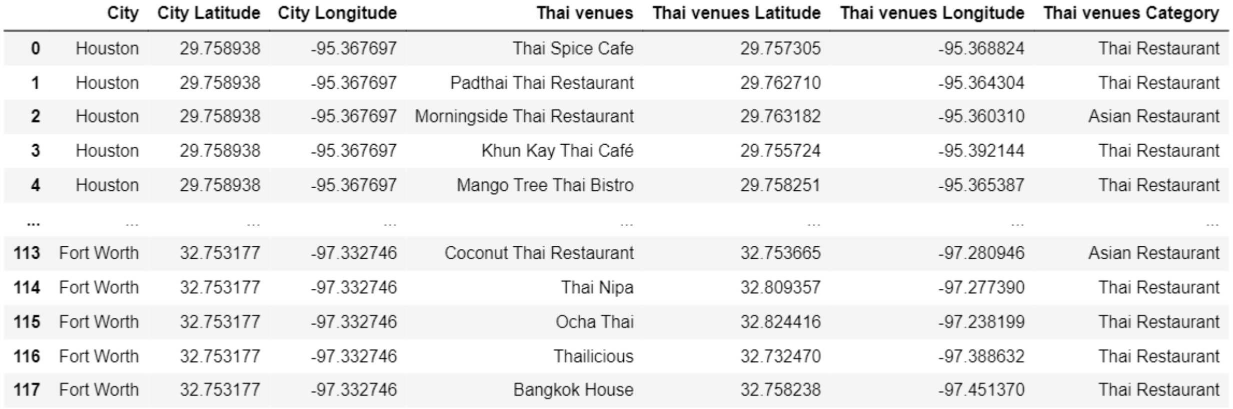


	City	% Asian	Median household income	Median gross rent
1	Fort Worth	4.6	62187	1060
2	San Antonio	2.8	52455	992
3	Austin	7.6	71576	1280
4	Dallas	3.4	52580	1052
5	Houston	6.8	52338	1041

Figure 3: Dataframe of selected demographic and economic information

1.3 Search Thai Restaurants

In order to study restaurant business in the area, we should start from our direct competitors which are other Thai restaurants. Foursquare API was used to search Thai restaurants in each city. The search keyword is 'Thai'. The call returned a JSON file from which we could extract the venues' names, coordinates and categories and store them in a new dataframe. The radius of interest is within 10 km and the search limit is 50 venues (less than explore limit which is 100 venues). Since the search results also contain other venue categories such as Thai offices and massage studios, we narrowed them down to only the venues with 'restaurant' categories. The results are displayed in Figure 4.



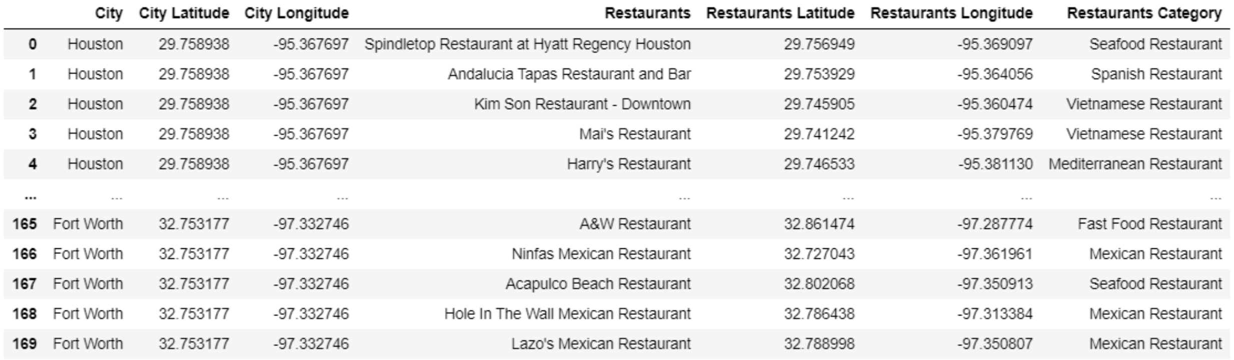
	City	City Latitude	City Longitude	Thai venues	Thai venues Latitude	Thai venues Longitude	Thai venues Category
0	Houston	29.758938	-95.367697	Thai Spice Cafe	29.757305	-95.368824	Thai Restaurant
1	Houston	29.758938	-95.367697	Padthai Thai Restaurant	29.762710	-95.364304	Thai Restaurant
2	Houston	29.758938	-95.367697	Morningside Thai Restaurant	29.763182	-95.360310	Asian Restaurant
3	Houston	29.758938	-95.367697	Khun Kay Thai Café	29.755724	-95.392144	Thai Restaurant
4	Houston	29.758938	-95.367697	Mango Tree Thai Bistro	29.758251	-95.365387	Thai Restaurant
...	...	...	...	...	...	...	...
113	Fort Worth	32.753177	-97.332746	Coconut Thai Restaurant	32.753665	-97.280946	Asian Restaurant
114	Fort Worth	32.753177	-97.332746	Thai Nipa	32.809357	-97.277390	Thai Restaurant
115	Fort Worth	32.753177	-97.332746	Ocha Thai	32.824416	-97.238199	Thai Restaurant
116	Fort Worth	32.753177	-97.332746	Thailicious	32.732470	-97.388632	Thai Restaurant
117	Fort Worth	32.753177	-97.332746	Bangkok House	32.758238	-97.451370	Thai Restaurant

Figure 4: Search results for Thai restaurants in each city using Foursquare API

1.4 Search All Restaurants

Other than Thai restaurants, we should also study other types of restaurants in the area to evaluate the market. Another search was conducted to search all restaurants in each city using Foursquare API. The search keyword is 'restaurant' and the radius of interest is within 10 km. Since the search results also contain other food places such as bars and hotels, we filtered them further to only the venues with 'restaurant' categories.

Please also note that since the limit for search is only 50 venues, not all the restaurants are captured in the results. We can obviously see that not as many Thai restaurants were returned compared to when we searched for Thai restaurants only. The same goes for other types of restaurants. The resulting dataframe is shown in Figure 5.

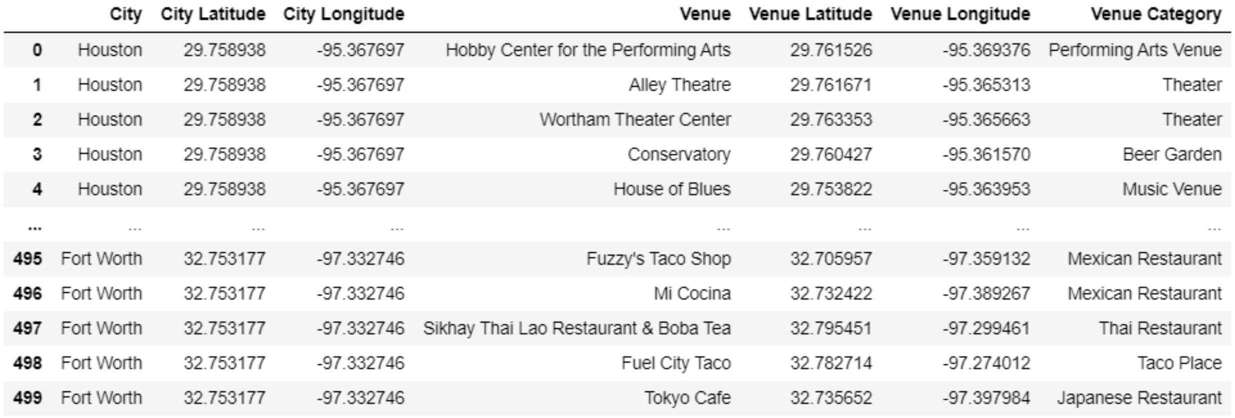


	City	City Latitude	City Longitude	Restaurants	Restaurants Latitude	Restaurants Longitude	Restaurants Category
0	Houston	29.758938	-95.367697	Spindletop Restaurant at Hyatt Regency Houston	29.756949	-95.369097	Seafood Restaurant
1	Houston	29.758938	-95.367697	Andalucia Tapas Restaurant and Bar	29.753929	-95.364056	Spanish Restaurant
2	Houston	29.758938	-95.367697	Kim Son Restaurant - Downtown	29.745905	-95.360474	Vietnamese Restaurant
3	Houston	29.758938	-95.367697	Mai's Restaurant	29.741242	-95.379769	Vietnamese Restaurant
4	Houston	29.758938	-95.367697	Harry's Restaurant	29.746533	-95.381130	Mediterranean Restaurant
...	...	...	...	...	...	...	...
165	Fort Worth	32.753177	-97.332746	A&W Restaurant	32.861474	-97.287774	Fast Food Restaurant
166	Fort Worth	32.753177	-97.332746	Ninfas Mexican Restaurant	32.727043	-97.361961	Mexican Restaurant
167	Fort Worth	32.753177	-97.332746	Acapulco Beach Restaurant	32.802068	-97.350913	Seafood Restaurant
168	Fort Worth	32.753177	-97.332746	Hole In The Wall Mexican Restaurant	32.786438	-97.313384	Mexican Restaurant
169	Fort Worth	32.753177	-97.332746	Lazo's Mexican Restaurant	32.788998	-97.350807	Mexican Restaurant

Figure 5: Search results for all restaurants in each city using Foursquare API

1.5 Explore the city

We should expand our analysis beyond the restaurant business and also study the popular spots in the city as well. This should help us learn the locals' interests and potentially use them to improve our strategy and marketing plans in the future. Therefore, we also used Foursquare API to explore each city. The radius of interest is within 10 km and the limit for explore is 100 venues. The explore results are shown in Figure 6.



	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Houston	29.758938	-95.367697	Hobby Center for the Performing Arts	29.761526	-95.369376	Performing Arts Venue
1	Houston	29.758938	-95.367697	Alley Theatre	29.761671	-95.365313	Theater
2	Houston	29.758938	-95.367697	Wortham Theater Center	29.763353	-95.365663	Theater
3	Houston	29.758938	-95.367697	Conservatory	29.760427	-95.361570	Beer Garden
4	Houston	29.758938	-95.367697	House of Blues	29.753822	-95.363953	Music Venue
...	...	...	...	...	...	...	...
495	Fort Worth	32.753177	-97.332746	Fuzzy's Taco Shop	32.705957	-97.359132	Mexican Restaurant
496	Fort Worth	32.753177	-97.332746	Mi Cocina	32.732422	-97.389267	Mexican Restaurant
497	Fort Worth	32.753177	-97.332746	Sikhay Thai Lao Restaurant & Boba Tea	32.795451	-97.299461	Thai Restaurant
498	Fort Worth	32.753177	-97.332746	Fuel City Taco	32.782714	-97.274012	Taco Place
499	Fort Worth	32.753177	-97.332746	Tokyo Cafe	32.735652	-97.397984	Japanese Restaurant

Figure 6: Explore results of each city using Foursquare API

2. Analyze and Visualize Data

Since we are planning to open a Thai restaurant, our target group of customers are primarily Asians and Americans with Asian descent. Thus, we started by visualizing the percentage of Asian population in each city using the data obtained from the United States Census Bureau website. A bar chart was created using pyplot and is shown under the Results section. In order to evaluate the economic condition of the city and affordability of the restaurant, we also looked at the median household income. Another bar chart was then created to show the median household income for each city. Also, the median gross rent was compared between different cities in the third chart.

In order to study the current competition in each city, we started by looking at our direct competitors which are other Thai restaurants. From the search for Thai restaurants using Foursquare API, the number of Thai restaurants in each city could be plotted for comparison. We were also able to create a leaflet map to visualize these Thai restaurants using Folium.

As mentioned previously, we also did another search for all restaurants in each city to evaluate the restaurant business beyond just the Thai restaurants. The radius for the search was within 10 km and the results were further filtered to only those with restaurant categories. We could then create horizontal bar charts that show the number of restaurants in each city by types.

We also used the Foursquare API to explore each city in order to study popular spots and learn the locals' interests for future marketing plans. The resulting dataframe can be used to analyze common venues in each city. There are a total of 130 unique venue categories. We started by using one hot encoding to assign 0 or 1 value to the venue categories. Then, we grouped the rows by city and took the mean of the frequency of occurrence of each venue category. Next, we created a function to sort the venue categories in descending order of occurrence frequency and created a new dataframe to display the top 10 venues for each city.

3. Cluster the Data

We can cluster these 5 cities based on their venues using K-Means clustering. The cities with similar venue categories are clustered together. We then added cluster labels to the venues dataframe and combine them with demographic and economic information as well as coordinates data for mapping.

Results

The bar chart showing % Asian in each city sorted from largest to smallest values is shown in Figure 7. We can see that Austin consists of the highest percentage of Asian population at 7.6% and San Antonio has the lowest % Asian at 2.8%.

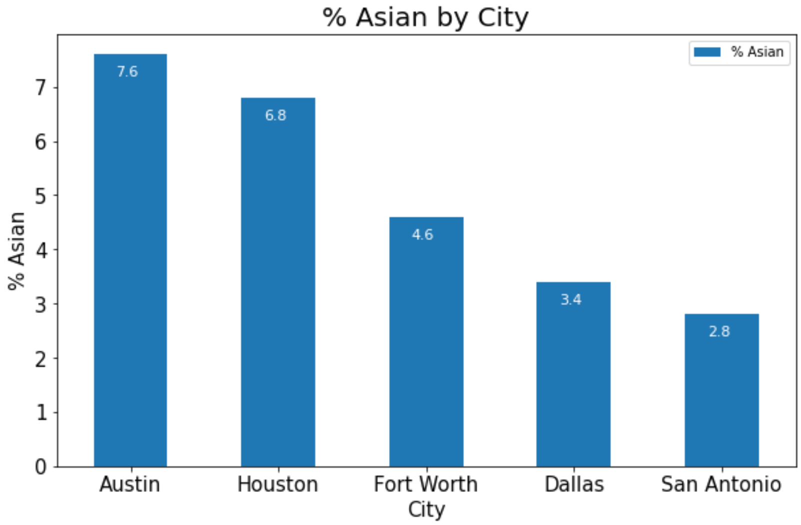


Figure 7: % Asian by city according to the United States Census Bureau website

The bar chart that shows the Median household income for each city is displayed in Figure 8. It is ranked from the city with highest to lowest values. As displayed in the chart, Austin has the highest median household income of $71,576 while Houston has the lowest median household income of $52,338.

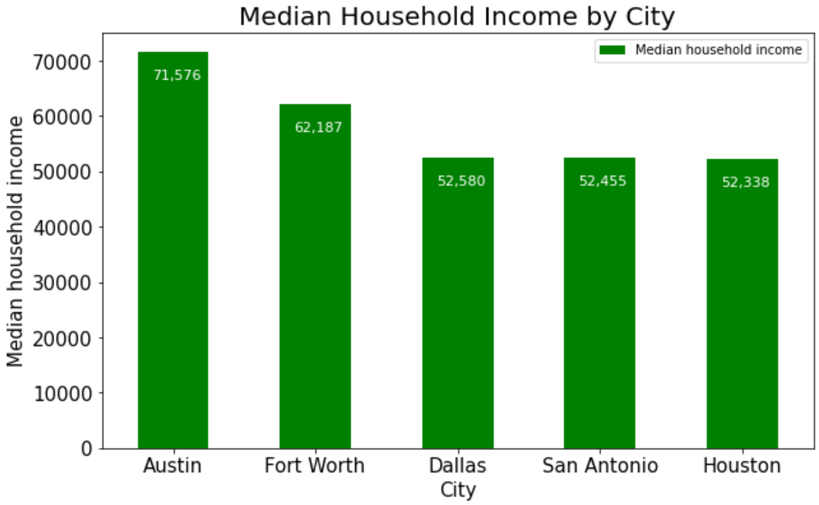


Figure 8: Median household income by city according to the United States Census Bureau website

The bar chart showing the Median gross rent for each city is displayed in Figure 9. It is also ranked from the city with highest to lowest values. The actual amount for restaurant lease will vary but the trend should still hold for cost comparison between cities. The higher the amount will result in the higher fixed costs for operating the restaurant in the future. We can see that Austin has the highest median gross rent while San Antonio has the lowest one.

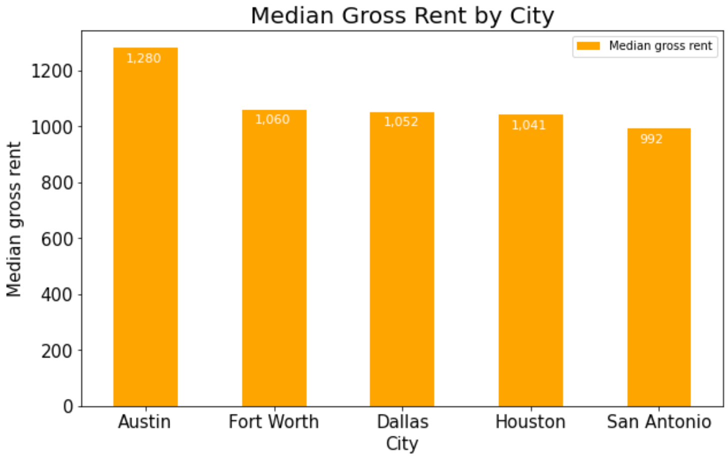


Figure 9: Median gross rent by city according to the United States Census Bureau website

From the search for Thai restaurants using Foursquare API, the number of Thai restaurants in each city can be plotted as shown in Figure 10. We can observe that the highest competition is in Austin and Houston with 28 and 27 Thai restaurants within 10 km radius, while the lowest competition is in Fort Worth with 16 Thai restaurants.

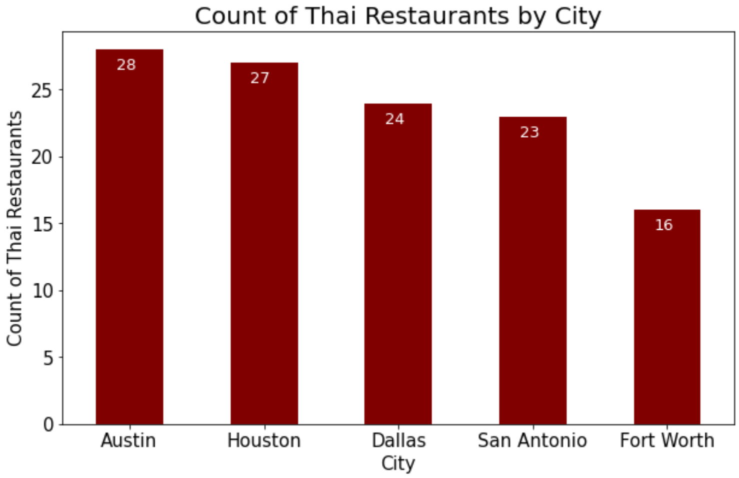


Figure 10: Count of Thai restaurants by city according to Foursquare API

The leaflet map showing the location of Thai restaurants in each city is displayed in Figure 11. The initial zoom is for Texas as a whole with all the Thai restaurants from the search shown in different colors according to the city. We can zoom in to look at the location of Thai restaurants in each city and click on the markers to display the name and city of the restaurants as shown in Figure 12.

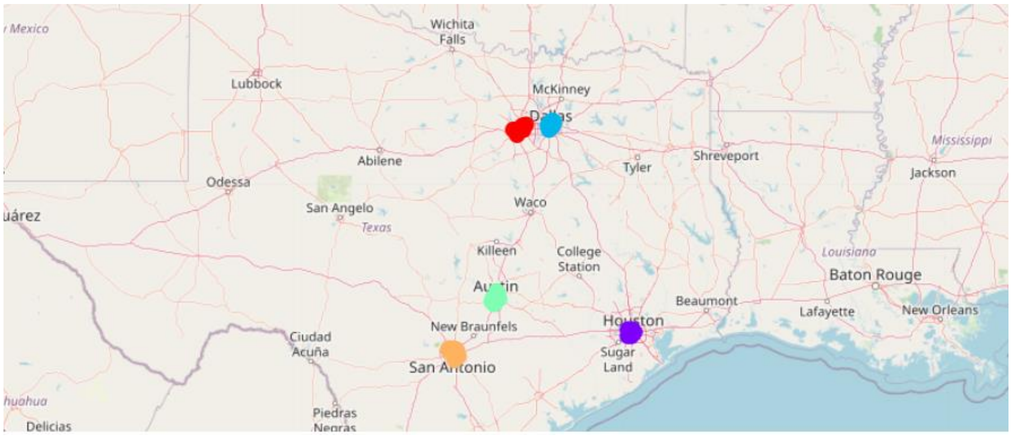


Figure 11: Leaflet map of Thai restaurants in Texas (initial zoom)

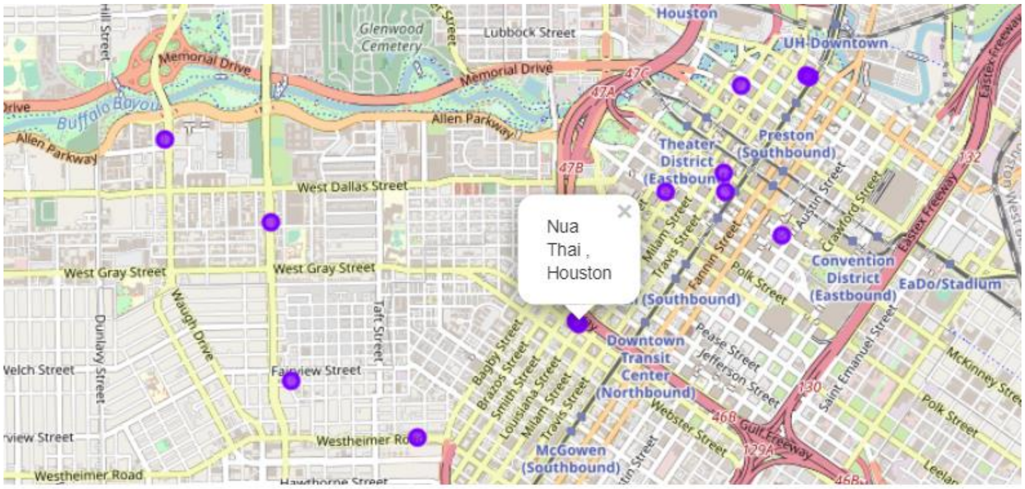


Figure 12: Zoomed-in version of leaflet map of Thai restaurants in Houston

According to the search for all restaurants using Foursquare API, a horizontal bar chart that shows the number of restaurants in each city by types can be created. Please note that since the limit for search is only 50 venues, not all the restaurants are captured in the results and the counts will therefore not be the most representative. However, we should get a good idea of the availability and proportion of different types of restaurants in each city.

The first bar chart is for restaurants in Houston. We can see that Mexican and American restaurants are the most prevalent, followed by Vietnamese restaurants. Vietnamese restaurants seem to be the most popular among Asian restaurants and there are also a moderate number of Chinese and Thai restaurants in the city.

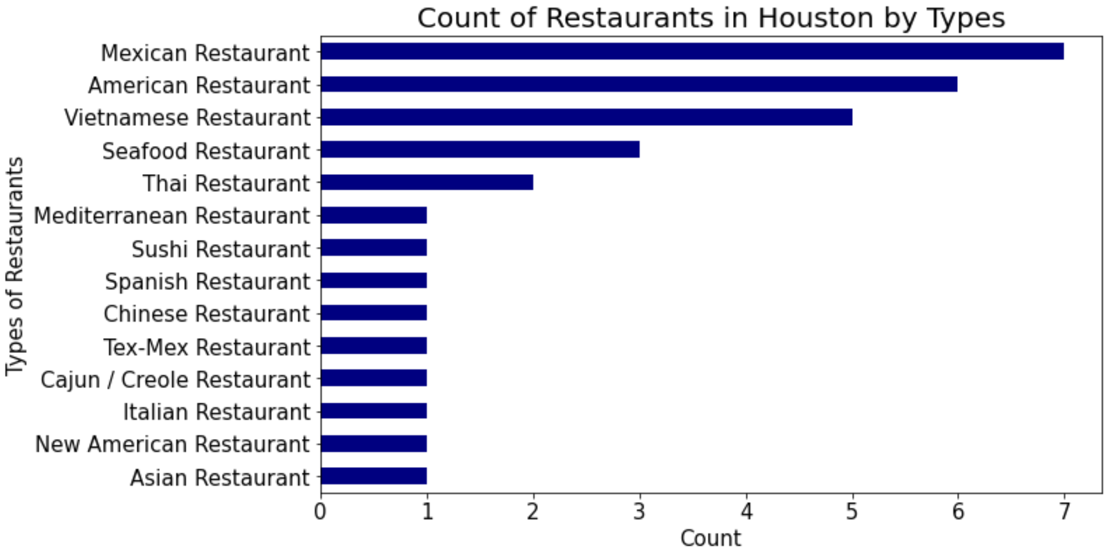


Figure 13: Count of restaurants in Houston by types according to Foursquare API

The second bar chart is for restaurants in Austin separated by types. Mexican restaurants seem to be the most popular restaurants in Austin, and the most common Asian restaurants here are Chinese restaurants.

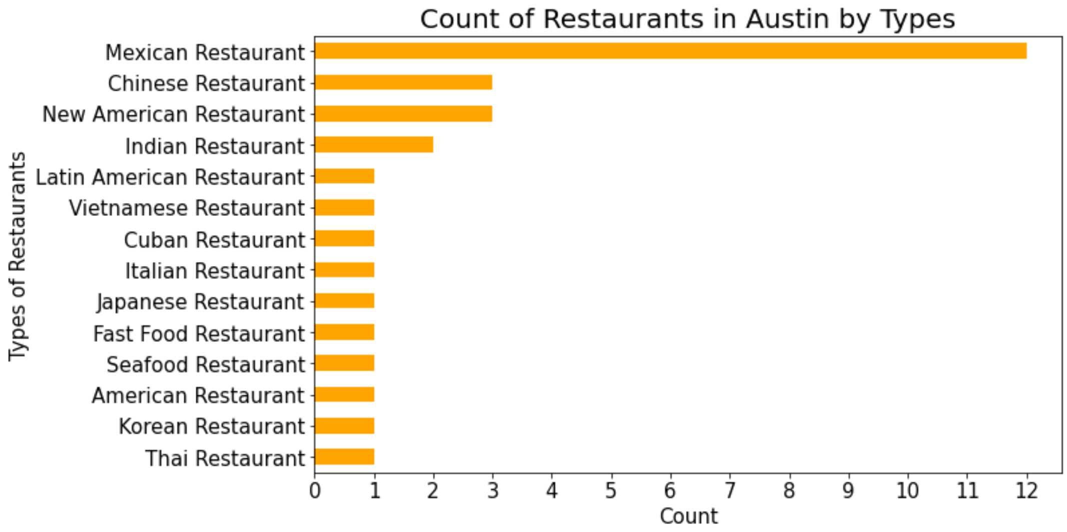


Figure 14: Count of restaurants in Austin by types according to Foursquare API

Next is the bar chart for restaurants in Dallas. Mexican and American restaurants are the most common restaurants. There are also a decent number of Italian restaurants. The more popular Asian restaurants here are Chinese and Vietnamese.

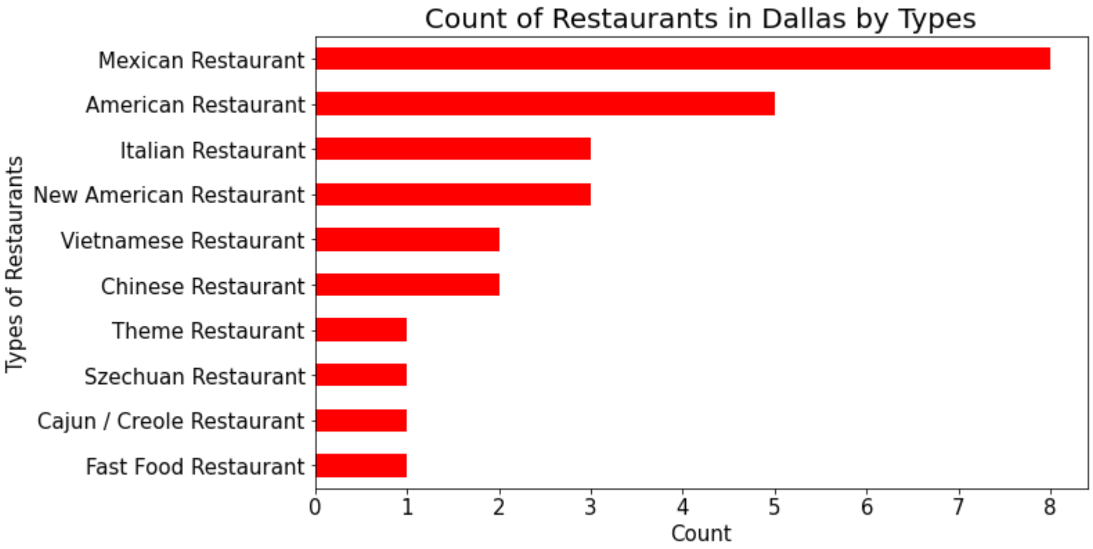


Figure 15: Count of restaurants in Dallas by types according to Foursquare API

The bar chart for restaurants in San Antonio is shown below. We can see that Mexican restaurants are significantly more popular than other types of restaurants. This is expected considering the city's Hispanic and Latino percentage according to the United States Census Bureau website. There also seem to be a fair share of international restaurants in the city.

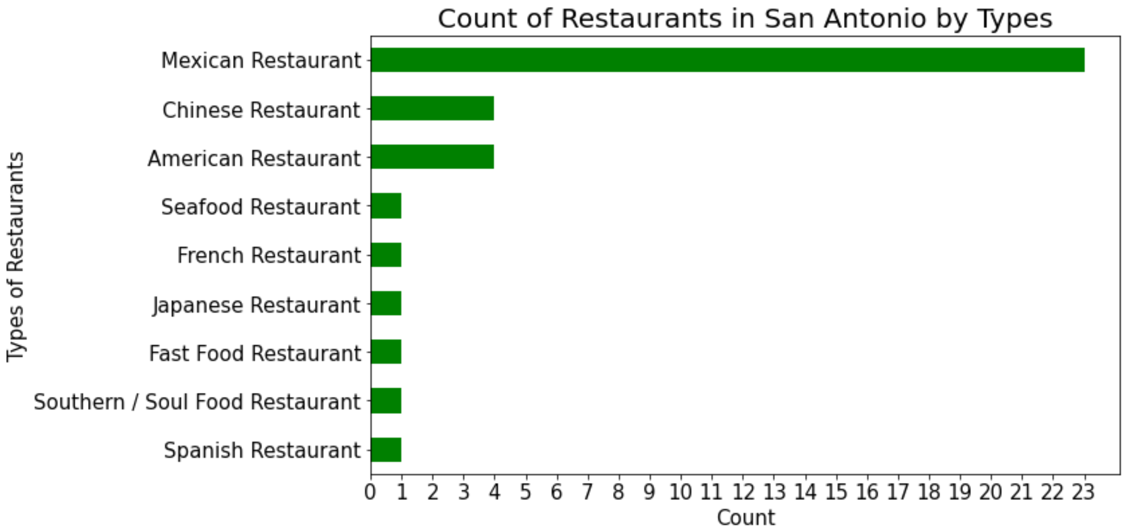


Figure 16: Count of restaurants in San Antonio by types according to Foursquare API

As for the bar chart of restaurants in Fort Worth shown below, Mexican restaurants appear to be the most common ones. Thai restaurants also seem to be the most popular Asian restaurants in the city.

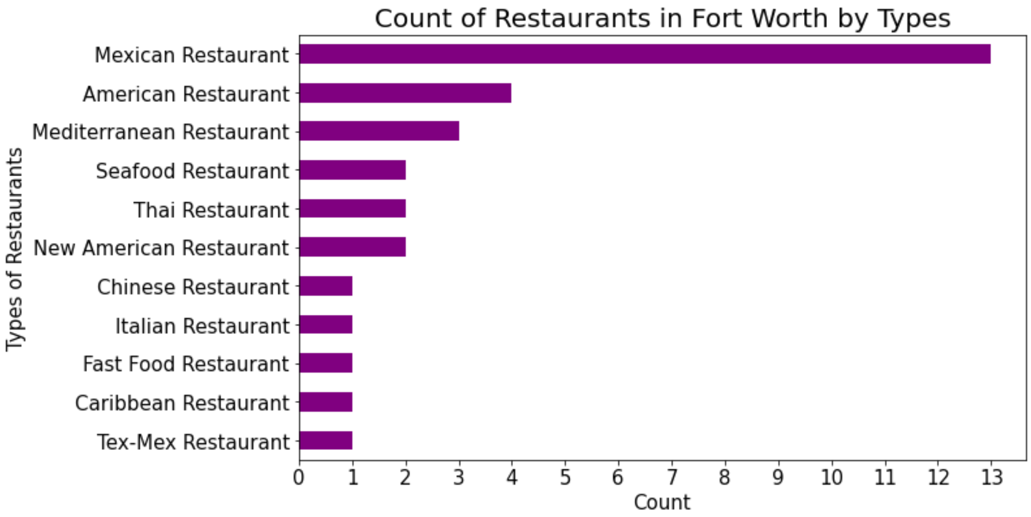
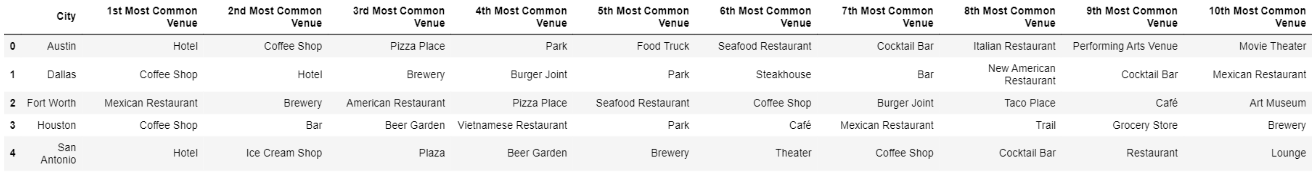


Figure 17: Count of restaurants in Fort Worth by types according to Foursquare API

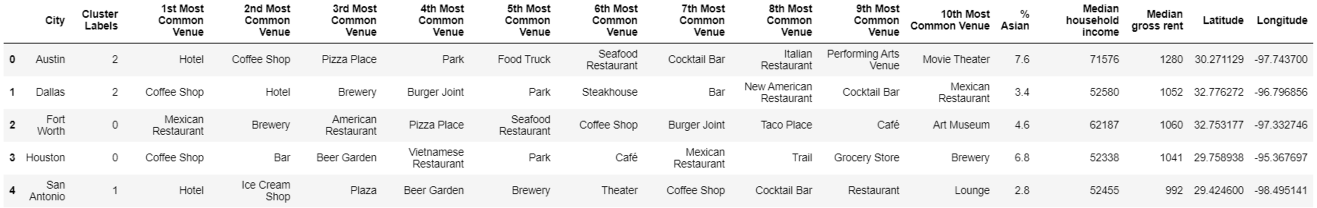
As a result of using Foursquare API to explore each city, we could analyze their common venues and create a dataframe to display the top 10 venues for each city as shown in Figure 18.



	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Austin	Hotel	Coffee Shop	Pizza Place	Park	Food Truck	Seafood Restaurant	Cocktail Bar	Italian Restaurant	Performing Arts Venue	Movie Theater
1	Dallas	Coffee Shop	Hotel	Brewery	Burger Joint	Park	Steakhouse	Bar	New American Restaurant	Cocktail Bar	Mexican Restaurant
2	Fort Worth	Mexican Restaurant	Brewery	American Restaurant	Pizza Place	Seafood Restaurant	Coffee Shop	Burger Joint	Taco Place	Café	Art Museum
3	Houston	Coffee Shop	Bar	Beer Garden	Vietnamese Restaurant	Park	Café	Mexican Restaurant	Trail	Grocery Store	Brewery
4	San Antonio	Hotel	Ice Cream Shop	Plaza	Beer Garden	Brewery	Theater	Coffee Shop	Cocktail Bar	Restaurant	Lounge

Figure 18: Top 10 venues for each city obtained from Foursquare data

K-Means clustering was used to cluster the five cities based on their venues. Cluster labels were added to other data as shown in Figure 19. The cities with similar venue categories are clustered together.



	City	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	% Asian	Median household income	Median gross rent	Latitude	Longitude
0	Austin	2	Hotel	Coffee Shop	Pizza Place	Park	Food Truck	Seafood Restaurant	Cocktail Bar	Italian Restaurant	Performing Arts Venue	Movie Theater	7.6	71576	1280	30.271129	-97.743700
1	Dallas	2	Coffee Shop	Hotel	Brewery	Burger Joint	Park	Steakhouse	Bar	New American Restaurant	Cocktail Bar	Mexican Restaurant	3.4	52580	1052	32.776272	-96.796856
2	Fort Worth	0	Mexican Restaurant	Brewery	American Restaurant	Pizza Place	Seafood Restaurant	Coffee Shop	Burger Joint	Taco Place	Café	Art Museum	4.6	62187	1060	32.753177	-97.332746
3	Houston	0	Coffee Shop	Bar	Beer Garden	Vietnamese Restaurant	Park	Café	Mexican Restaurant	Trail	Grocery Store	Brewery	6.8	52338	1041	29.758938	-95.367697
4	San Antonio	1	Hotel	Ice Cream Shop	Plaza	Beer Garden	Brewery	Theater	Coffee Shop	Cocktail Bar	Restaurant	Lounge	2.8	52455	992	29.424600	-98.495141

Figure 19: Dataframe showing cluster label for each city along with other data

The visualization of these clusters in Folium map is illustrated in Figure 20. The clusters are separated by colors and the size of markers corresponds with the number of Thai restaurants in each city.

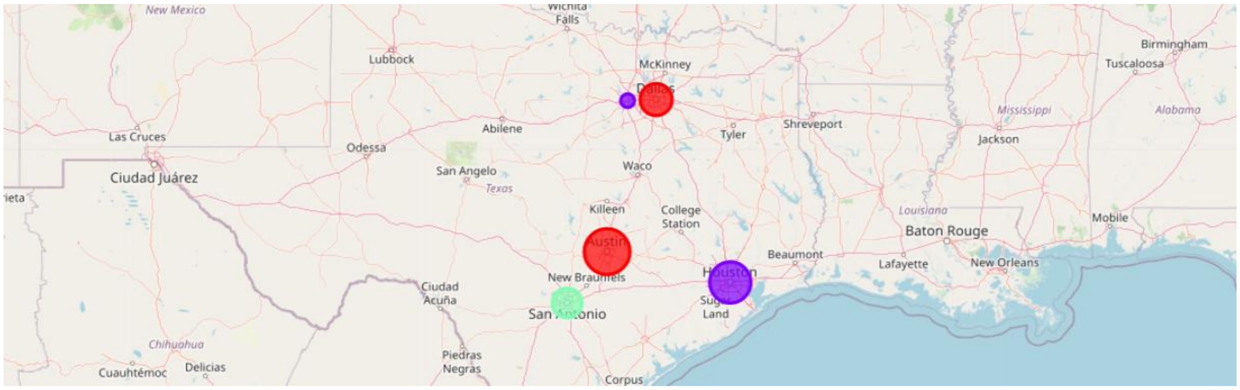


Figure 20: Map of five cities divided into three clusters based on their venues. Colors distinguish clusters and the size of marker represents the number of Thai restaurants in that city.

Discussion

Several general observations can be noted from the exploratory data analysis performed. First, Austin consists of the highest percentage of Asian population while San Antonio has the lowest Asian percentage. This should give us a good idea about the demand since our primary group of customers are Asians and Americans with Asian descent. In order to evaluate the economic condition of the city and affordability of the restaurant, we also analyzed the median household income and median gross rent. Austin appears to have the highest median household income while Houston has the lowest income. Nevertheless, Austin has the highest median gross rent and San Antonio has the lowest one. The higher the leasing amount will result in the higher fixed costs for operating the restaurant in the future. According to the search for Thai restaurants using Foursquare API, Austin and Houston have the highest number of Thai restaurants within the specified radius while Fort Worth has the lowest number of Thai restaurants. A leaflet map of all the Thai restaurants was also created so we can use it to explore the Thai restaurants in the cities further. When all types of restaurants were considered, Mexican restaurants seem to be the most common and popular food options in all the five cities. K-Means clustering was also performed to cluster the cities based on their venues. Austin and Dallas are clustered together with their top venues being hotels and coffee shops. Fort Worth and Houston are in another cluster with restaurants and bars being their common venues. Meanwhile, San Antonio is separated in a different cluster with a lot of hotels, ice cream shops and plazas.

An analysis was conducted primarily to evaluate the potential of opening a Thai restaurant in the five most populous cities in Texas. Fort Worth seems to be the city with lowest competition. It has the lowest number of Thai restaurants within the specified radius while having Thai restaurants as the most popular Asian restaurants in the city. Also, it has the second highest median household income and a decent percentage of Asian population. While Fort Worth appears to be the prospective city for a Thai restaurant, we should still drill down further in each city to study their neighborhoods and consider other factors such as distance from offices, proximity to public transportation, and distance from attractions. Also, we should be able to use the insights we gained from studying top venues in the cities to help in making future marketing plans. For instance, since coffee shops and ice cream shops seem to be the popular venues in these cities, we might plan to include them in the menus or open a mini dessert bar to attract more customers in the future. We should also be able to use Foursquare API to explore trending venues with highest foot traffic at a particular time to help in determining specific locations for the restaurant or where to promote the restaurant.

That being said, there are several limitations to the analysis. As mentioned previously, the limit for search using Foursquare API is only 50 venues, so not all of the restaurants were captured. This resulted in a slight bias when evaluating the overall restaurant business in the cities. Also, the clustering results might be different if other clustering techniques such as Hierarchical or Density-based clustering were used. However, it should give us some preliminary ideas about opening a restaurant in the five major cities of Texas.

Conclusion

The analysis was conducted to explore the potential of opening a Thai restaurant in the five most populous cities in Texas. The geographical coordinates, demographic and economic information, and venues data were obtained from Geopy, web scraping, and Foursquare API. The data were used to evaluate the demand, the competition, the economic condition and the affordability of the restaurant. Popular spots in the cities were also explored to study the locals' interests for future marketing plans. Several visualizations were created for the analysis such as charts and leaflet maps. One of the machine learning techniques, Clustering was used to group the cities into different clusters based on their venues. Several observations from the study were noted along with further recommendations. This project should provide a good example of using fundamental data science techniques to solve a real-world business problem.