Coursera Capstone Report

Daniel Wilcox

Introduction

This project will focus on comparing the top 200 most populous cities in the US based on the types of venues in the area. The cities will then be clustered using the k-means algorithm. The target audience for this project would be people who are looking to expand their businesses from one city to another. Cities in the same cluster may have like-minded citizens and similar business needs.

For example, if cities A and B are in the same cluster, a business owner in city A may have reason to believe that his or her business would also succeed in city B. People that are moving to a different city may also have interest in this project. They could use the clusters to determine which new cities are most like the one where they currently live.

Data

This project will utilize data from Foursquare to compile the most common types of venues in each of the cities. It will also use US city population data from the internet and location data accessed with geopy. For example, the 8th largest city in the US is San Diego, as shown on this web page https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population. This table could be scraped into a pandas dataframe using the BeautifulSoup package. The coordinates of San Diego are 32.8153, -117.1350, which could be added to the cities dataframe through the use of the geopy python package.

Using the location data in the dataframe, an API call could be made to Foursquare to retrieve the first 200 venues within a 1 mile radius of the given coordinates. The venues would then be classified using one-hot encoding of the venue types. This process would be repeated for the remaining 199 cities. Once this process is finished, the top 10 venue types in each city would be compiled and entered into the k-means algorithm to generate the clusters.

Methodology

Data on the top 200 most populous cities in the United States was initially pulled from a table on Wikipedia (https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population). The only necessary columns at this point were the rank, city, and state columns. All of the other columns were discarded as they contain information that is not necessary for this project. The city names were further cleaned up by removing any footnote tags that were present in the Wikipedia table. While these tags are helpful in a table on the web, they simply clutter up the city names in a pandas dataframe. Regular

expressions were used to iterate through the rows of the dataframe and remove any text containing the sequence "[…]" or similar.

	2019rank	City	State[c]	2019estimate	2010Census	Change	2016 land area	2016 land area.1	2016 population density	2016 population density.1	Location
0	1	New York[d]	New York	8336817	8175133	+1.98%	301.5 sq mi	780.9 km2	28,317/sq mi	10,933/km2	40°39′49″N 73°56′19″W / 40.6635°N 73.9387°W
1	2	Los Angeles	California	3979576	3792621	+4.93%	468.7 sq mi	1,213.9 km2	8,484/sq mi	3,276/km2	34°01′10″N 118°24′39″W / 34.0194°N 118.4108°W
2	3	Chicago	Illinois	2693976	2695598	−0.06%	227.3 sq mi	588.7 km2	11,900/sq mi	4,600/km2	41°50′15″N 87°40′54″W / 41.8376°N 87.6818°W
3	4	Houston[3]	Texas	2320268	2100263	+10.48%	637.5 sq mi	1,651.1 km2	3,613/sq mi	1,395/km2	29°47′12″N 95°23′27″W / 29.7866°N 95.3909°W
4	5	Phoenix	Arizona	1680992	1445632	+16.28%	517.6 sq mi	1,340.6 km2	3,120/sq mi	1,200/km2	33°34′20″N 112°05′24″W / 33.5722°N 112.0901°W

The initial dataframe

Location data was pulled from the geocoders.geopy.Nominatim() module. This module uses open-source geographical data from nominatim.org. The latitude and longitude for each city was retrieved and added to the existing cities dataframe.

	Rank	City	State	Latitude	Longitude
0	1	New York	New York	40.712728	-74.006015
1	2	Los Angeles	California	34.053691	-118.242767
2	3	Chicago	Illinois	41.875562	-87.624421
3	4	Houston	Texas	29.758938	-95.367697
4	5	Phoenix	Arizona	33.448437	-112.074142

The dataframe after cleaning and adding the geographical coordinates

Foursquare API calls were used to retrieve venue information for each city. The first 100 venues within a 1-mile radius of the given coordinates were used for this analysis. These venues were then grouped by city and counted to get an idea of how many venues were pulled for each city.

City	State	City Latitude	City Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
Akron	**Ohio**	91	91	91	91	91	91
Albuquerque	**New Mexico**	100	100	100	100	100	100
Alexandria	**Virginia**	100	100	100	100	100	100
Amarillo	**Texas**	35	35	35	35	35	35
Anaheim	**California**	64	64	64	64	64	64
...	**...**	...	...	...	...	...	...
Washington	**District of Columbia**	100	100	100	100	100	100
Wichita	**Kansas**	100	100	100	100	100	100
Winston–Salem	**North Carolina**	92	92	92	92	92	92
Worcester	**Massachusetts**	100	100	100	100	100	100
Yonkers	**New York**	56	56	56	56	56	56

City venue counts

The categories for each venue were converted to categorical variables using one-hot encoding. The venues were once again grouped up by city and the mean values for each category were calculated to get an idea of the top venues in each city.

	City	State	ATM	Accessories Store	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Service	American Restaurant	...	Weight Loss Center	Whisky Bar	Wine Bar
0	Akron	Ohio	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.032967	...	0.0	0.010989	0.010989
1	Albuquerque	New Mexico	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.020000	...	0.0	0.000000	0.000000
2	Alexandria	Virginia	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.040000	...	0.0	0.000000	0.000000
3	Amarillo	Texas	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.057143	...	0.0	0.000000	0.028571
4	Anaheim	California	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.015625	...	0.0	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
194	Washington	District of Columbia	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.030000	...	0.0	0.000000	0.000000
195	Wichita	Kansas	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.050000	...	0.0	0.000000	0.000000

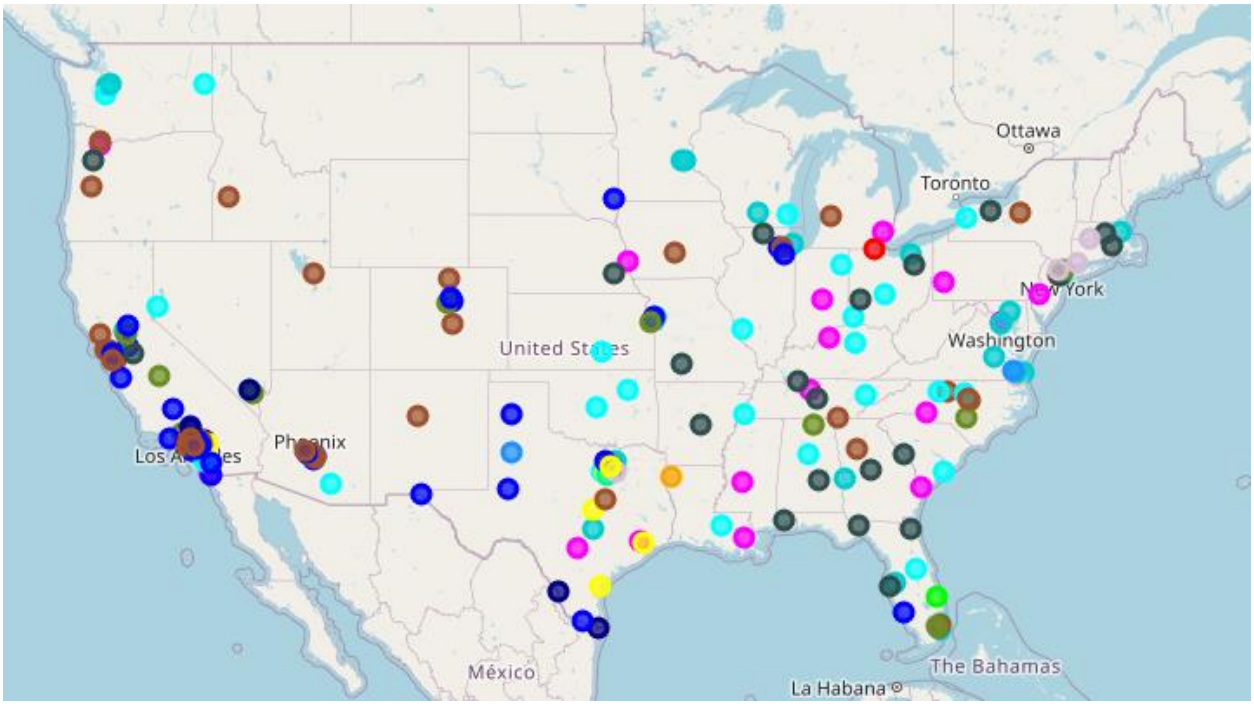
Grouped and average venue data for each city

These venue data were then sorted to get the top ten most common venue categories in each city.

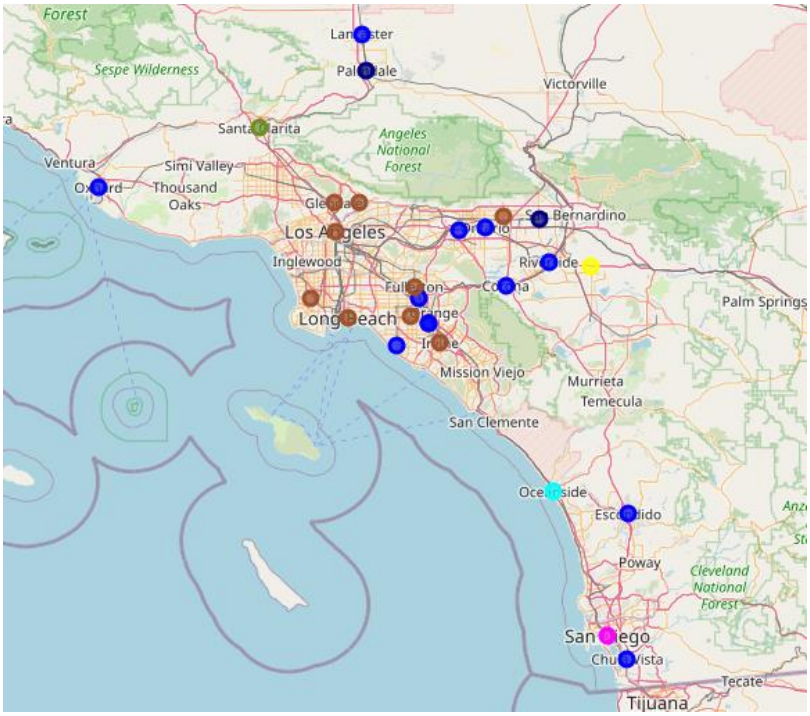
	City	State	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Akron	Ohio	Sandwich Place	Bar	Rental Car Location	Italian Restaurant	Coffee Shop	Music Venue	American Restaurant	Bank	Café	Fast Food Restaurant
1	Albuquerque	New Mexico	Coffee Shop	Pizza Place	Sandwich Place	Brewery	Hotel	Restaurant	Mexican Restaurant	Bar	New American Restaurant	Diner
2	Alexandria	Virginia	Park	Coffee Shop	Pizza Place	American Restaurant	Grocery Store	New American Restaurant	Seafood Restaurant	Spa	Hotel	Gastropub
3	Amarillo	Texas	Mexican Restaurant	Restaurant	Sandwich Place	Café	American Restaurant	Sushi Restaurant	Convenience Store	Performing Arts Venue	Furniture / Home Store	Fast Food Restaurant
4	Anaheim	California	Mexican Restaurant	Convenience Store	Taco Place	Coffee Shop	Indian Restaurant	Burger Joint	Brewery	Ice Cream Shop	Bakery	Steakhouse

Top 10 venue categories in each city

All of the categorical city venue data was used to train a k-means clustering algorithm with fifteen clusters. The clusters were visualized with a map using the folium package. For example, the blue dots on the map represent cluster 12.



Map of cities, different colors represent distinct clusters



Zoomed-in map of Southern California

Finally, a dictionary of individual cluster dataframes was created to better organize all of the data. This dictionary consists of 15 keys, one for each unique cluster, and contains the city, state, and top ten most common venues for each data point. As a final step of organization, another dictionary was created to hold the top venue categories in each cluster.

Cluster Label	Count
0.0	3
1.0	21
2.0	11
3.0	40
4.0	5
5.0	1
6.0	22
7.0	1
8.0	7
9.0	28
10.0	1
11.0	1
12.0	35
13.0	6
14.0	17

Cluster city counts

Results and Discussion

Through analysis of venue categories and clustering of cities, the top 200 most populous cities in the United States have been grouped into fifteen different clusters. These clusters were created based on similiarites in the types of venues in a given area. Through visualizing all of the clusters on a map, one can see that these clusters also frequently share geographic similarities as well. Take for example cluster 12, which tends to be centered around the southwestern portion of the country. The cities within cluster 12 are displayed below:

	City	State
87	Glendale	Arizona
34	Mesa	Arizona
146	Peoria	Arizona
80	Chandler	Arizona
162	Lancaster	California
136	Ontario	California
163	Salinas	California
119	Huntington Beach	California
170	Pomona	California
109	Oxnard	California
172	Escondido	California
197	Orange	California
74	Chula Vista	California
188	Roseville	California
61	Stockton	California
57	Riverside	California
56	Santa Ana	California
54	Anaheim	California
51	Bakersfield	California
155	Corona	California
159	Hayward	California
189	Thornton	Colorado
53	Aurora	Colorado
131	Cape Coral	Florida
123	Aurora	Illinois
176	Joliet	Illinois
168	Kansas City	Kansas
138	Sioux Falls	South Dakota
117	Amarillo	Texas
93	Garland	Texas
178	Midland	Texas
184	McAllen	Texas
187	Denton	Texas
69	Plano	Texas
21	El Paso	Texas

The vast majority of the cities within this cluster are located within California, Texas, and Arizona. In addition to creating the fifteen clusters of cities, a dictionary was created that contains the most common venue types for each cluster. Continuing with the example of cluster 12, the most common venue types are as follows:

```
First most common venue types: ['Mexican Restaurant', 'Fast Food Restaurant', 'Burger Joint']
Second most common venue types: ['Convenience Store', 'Mexican Restaurant', 'Coffee Shop']
Third most common venue types: ['Fast Food Restaurant', 'Pizza Place', 'Sandwich Place']
```

We can see that some of the most common venue types are Mexican restaurants. This makes sense for the southwestern United States, as these states are close to the border with Mexico and have significant immigrant populations.

This project could be a powerful tool as a starting point for someone who is looking to open a business in a new city or move to a new city. This can be seen in the cluster 12 example. Someone who has lived in Chula Vista, California for a long time would likely want to move somewhere new that also has a high density of Mexican restaurants. Mesa, Arizona or El Paso, Texas could be good places to start a search. When moving to a new city, familiar venues could help someone to feel more at home.

One possible weakness of this project is it strongly biases larger cities. For example, no cities from Wyoming, Montana, or North Dakota are included in this analysis. Smaller cities would be part of distinct clusters as they likely have different venue distributions that large cities. This could be remedied by pulling additionaly city data from other webpages and training the model again. This would not

require too much effort, as it is as simple as changing the url in the request line or copying the cell and concatonating the new table with the existing one.

One other interesting feature of this analysis to note is that Foursquare did not retrieve any venue data for San Bernardino, California. This was remedied by using the pd.dropna() function to remove the missing data from the model but could be due to faulty location data from geopy. In future versions, location verification steps could be implemented to ensure that no cities are missed.

Conclusion

The goal of this project was to create a clustering model to group together the top 200 most populous cities in the United States based on the most common venue categories in each city. City data was firt pulled from Wikipedia through the use of the BeautifulSoup package. Geographical data for each city was then retrieved using the geopy package and concatonated to the existing cities dataframe. To get venue data, API calls were made to Foursquare using the previously obtained location data. In this analysis, the venue limit was set to 100 venues and the radius was set to one mile. The venue data was then transformed to categorial data using one-hot encoding and all of the venues were grouped together by city. To get an idea of the types of venues in each city, the mean values for each venue category were calculated. These mean values were then sorted to get the top ten venue categories for each city. After this, all of the city category data was put into a k-means clustering algorithm. This algorithm returned 15 distinct clusters of cities. These clusters are based off of most common venue types but also happen to share some geographical similarities, as seen on the folium map visualization. Finally, a dictionary was created containing the most common venue types for categories for each cluster.

This project is primarily useful as a starting point for people searching for new cities that are similar to the one in which they currently live. After getting an inital list of cities, they should still do more research into each of the cities, as the most common types of venues are not all that determines what it is like to live in a city or how well a business will do in a new city.