Coursera Data Science Capstone

Introduction

Summary This report seeks to analyze data from various cities on the East Coast of the United States. Namely, I am seeking to use information from FourSquare in order to ascertain which major cities have the smallest number of Chiropractors. One could leverage this data in determining where to open a new practice if a certain area is under-served. Alternatively, one may look to move to an area with many practices so one could start a career there as an employee rather than proprietor

Target Audience My target audience is composed of potential professionals in the field of Chiropractic. This may be interesting information to have when planning to visit conferences, start a practice, attend university, etc. Though individuals outside of the Chiropractic community may find the information useful in finding cities with a potentially larger footprint of alternative medicine providers

Primary Questions Which large cities in the Eastern United States have the smallest number of Chiropractors?
Which large cities in the Eastern United States have the largest number of Chiropractors?

Data

Data Sources Data came from two primary sources. First, the FourSquare API was used to collect data about various cities in the US and get the number and location of Chiropractors. Two, Wikipedia was used for the cities in question to acquire demographic and population size data for the cities being analyzed

One way I used the FourSquare data is the "CategoryID" subfield under "category". This field helped determine if a venue is a Chiropractor since a venue may have a name with "Chiropractor", "Chiropractic", etc.. Additionally, I used the "lat" and "lng" subfields of "location" to determine if a venue is within a city's limits

I leveraged Wikipedia to get the population totals for major cities within the US. This article (https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population) was of use, specifically

Nominatim was used to find each city's latitude and longitude values

Data Manipulation The majority of the data cleaning and manipulation involved stripping off extraneous characters from the Wikipedia dataset. The table on the site had many footnotes and other characters that would not allow me to systemically find proper latitude and longitude values for each city. Additionally, the output would not look particularly clean with these extra characters.

Rather than use a fixed radius for my queries to FourSquare, I had to use a bit of math to convert the square mileage of each city into meters and dynamically assign a radius to the API call for each city

Methodology

The methodology was relatively straight-forward. I started by pulling HTML data from Wikipedia to make my base dataframe of large US cities. The results were then filtered to include only the 50 cities with the largest population. The dataset was then further subdivided by only including cities east of St. Louis, MO

I iterated through the list to fetch each city's latitude and longitude values from Nominatim. These values were stored in the dataframe with each city.

Next, I passed those latitude and longitude values to FourSquare in order to get a total venue count for each city. The radius of the query was determined individually for each city by calculating a metric value based on a city's square mileage.

Lastly, I plotted each of the cities in the dataframe with their population and venue count (in this case, total number of Chiropractors supplied by FourSquare)

Results

Summary of Findings

Somewhat unsurprisingly, the two largest cities, New York and Chicago, had the largest number of Chiropractic firms. In fact, it would seem as if there may be a limitation in the results passed from the API as both cities returned only 50 chiropractic firms.

Miami, FL had the fewest Chiropractic firms. This may be related to the fact that Miami also had the smallest land area of any of the cities that were analyzed.

Resulting Dataframe

The following dataframe includes the results from the Wikipedia page and the FourSquare info:

In [18]:
```
Eastern_Cities
```

Out[18]:

	City	State	Population	Latitude	Longitude	Venue Count	2016 land area	2016 population density
0	new york	New York	8398748	40.712728	-74.006015	50	301.0	28,317/sq mi
1	chicago	Illinois	2705994	41.875562	-87.624421	50	227.0	11,900/sq mi
2	philadelphia	Pennsylvania	1584138	39.952724	-75.163526	38	134.0	11,683/sq mi
3	jacksonville	Florida	903889	30.332184	-81.655651	45	747.0	1,178/sq mi
4	columbus	Ohio	892533	39.962260	-83.000707	41	218.0	3,936/sq mi
5	charlotte	North Carolina	872498	35.227087	-80.843127	44	305.0	2,757/sq mi
6	indianapolis	Indiana	867125	39.768333	-86.158350	40	361.0	2,366/sq mi
7	washington	District of Columbia	702455	38.894893	-77.036553	44	61.0	11,148/sq mi
8	boston	Massachusetts	694583	42.360253	-71.058291	45	48.0	13,938/sq mi
9	detroit	Michigan	672662	42.331551	-83.046640	32	138.0	4,847/sq mi
10	nashville	Tennessee	669053	36.162230	-86.774353	43	475.0	1,388/sq mi
11	memphis	Tennessee	650618	35.149022	-90.051629	26	317.0	2,056/sq mi
12	louisville	Kentucky	620118	38.254238	-85.759407	36	263.0	2,339/sq mi
13	baltimore	Maryland	602495	39.290882	-76.610759	23	80.0	7,598/sq mi
14	milwaukee	Wisconsin	592025	43.034993	-87.922497	42	96.0	6,186/sq mi
15	atlanta	Georgia	498044	33.749099	-84.390185	46	133.0	3,539/sq mi
16	miami	Florida	470914	25.774266	-80.193659	19	36.0	12,599/sq mi
17	raleigh	North Carolina	469298	35.780398	-78.639099	41	145.0	3,163/sq mi
18	virginia beach	Virginia	450189	36.852984	-75.977418	43	244.0	1,850/sq mi
19	tampa	Florida	392890	27.947760	-82.458444	38	113.0	3,326/sq mi
20	new orleans	Louisiana	391006	29.949932	-90.070116	38	169.0	2,311/sq mi

Discussion

One item I would like to investigate is the size of New York and Chicago's datasets being returned by the API. This would be a pretty big limitation if 50 results were the maximum I could expect using the free API. In particular, if I were to investigate venues that were more numerous (say, restaurants), this limitation would make the data much less useful.

Another item to explore in the future is to pull in more variables from other data sources--average rent costs, median income of city residents, etc. With more data, it would be helpful to make more informed decisions about starting a business or career in the cities being analyzed.

Conclusion

What I hoped to provide was a high-level look at a venue type, in this case Chiropractic firms, and which cities have large numbers of those firms already. The hope would be that density of venues could be helpful for prospective employees and employers about whether to dig into the city further to see if it would be appealing to move there. In the future, I hope to include more data in the tables and analysis to help people make a more informed decision.