Capstone Final

Finding Places to Live Similar to Home

Benjamin McDougal

Introduction

- ***The Audience***
  - My Audience is my Family and anyone else who wants to live in a quieter, less populous area, without all the fancy attractions but still has stuff to do similar to my Hometown of Idaho Falls, Idaho
- ***The Background***
  - Many young people are looking for new jobs and places to live. They including myself would like a place similar to where they grew up if at all possible. Directly with my comparison my city I grew up in is small enough that many would laugh at it being called a city at around 60,000 people and about 5 square miles. Jobs are more limited to a few big employers or owning/managing your own business. Thus, I and many other young adults are looking at moving to bigger cities even if that is not ideal for our lifestyle, habits, and upbringing.

Question

Are there other Counties similar to the one I grew up other places in the Nation where job opportunities and wages are better?

Limitations to Answering the Question

- As there are several thousand cities in America with great differences in size. I have decided instead of using cities to use counties which are a more equal distribution and manageable.
- Also, I am going to further restrict the counties to ones that are more similar in population as well to further balance the scale appropriately. I will work with counties between 50,000 and 500,000 thousand people.
- I am basing similarity of counties on venue types as I believe that reflects the culture and wealth of the area quite well. That alongside the population should allow me to find areas similar to where I grew up.

Data

- **Outside of Foursquare**
  - I will use the county name and population columns of the table of American county information as from Wikipedia's list here Wikiepdia American Counties
  - County names lets me group my data geographically. County Population helps to define similarity between counties.
- **Within Foursquare**
  - I also will use Foursquare data of venue types as a measure of similarity of culture and wealth of the county.

Methodology

- Step 1- Import libraries, functions and data
- Step 2 Wrangle the Data
- Step 3 Get Venues
- Step 4 Onehot and Combine with Latitude and Longitude Values
- Step 5 Run K-Means

Step 1
Import libraries, functions and data

- Import Libraries
- Create get venues function
- Import Data
  - With this I noted that the index for the table from the Wikipedia page on the data import is 3. That and in reading through the FourSquare API documentation. I learned FourSquare will automatically geocodes for me. I also found that using near instead of providing the latitude and longitude caused foursquare to automatically base the radius on density of venues.

Step 2
Wrangle the Data

- Firstly, I selected only the relevant columns
  - With that I dropped any NA values this gave me a list of about 870 counties total
- Then due to the FourSquare rate limit on a free account of about 950 non-premium requests a day I decided to only use a sample of 150 counties.

Step 3
Get Venues

- Here we use a function to obtain the venues for our entire county list.
- I used a limit of 25 venues
- As noted above I used near without a radius to allow FourSquare to choose a radius and venues according to its internal default of most densely located venues.
- With this we grab the venues category and location.

Step 4
Onehot and Combine with Latitude and Longitude Values

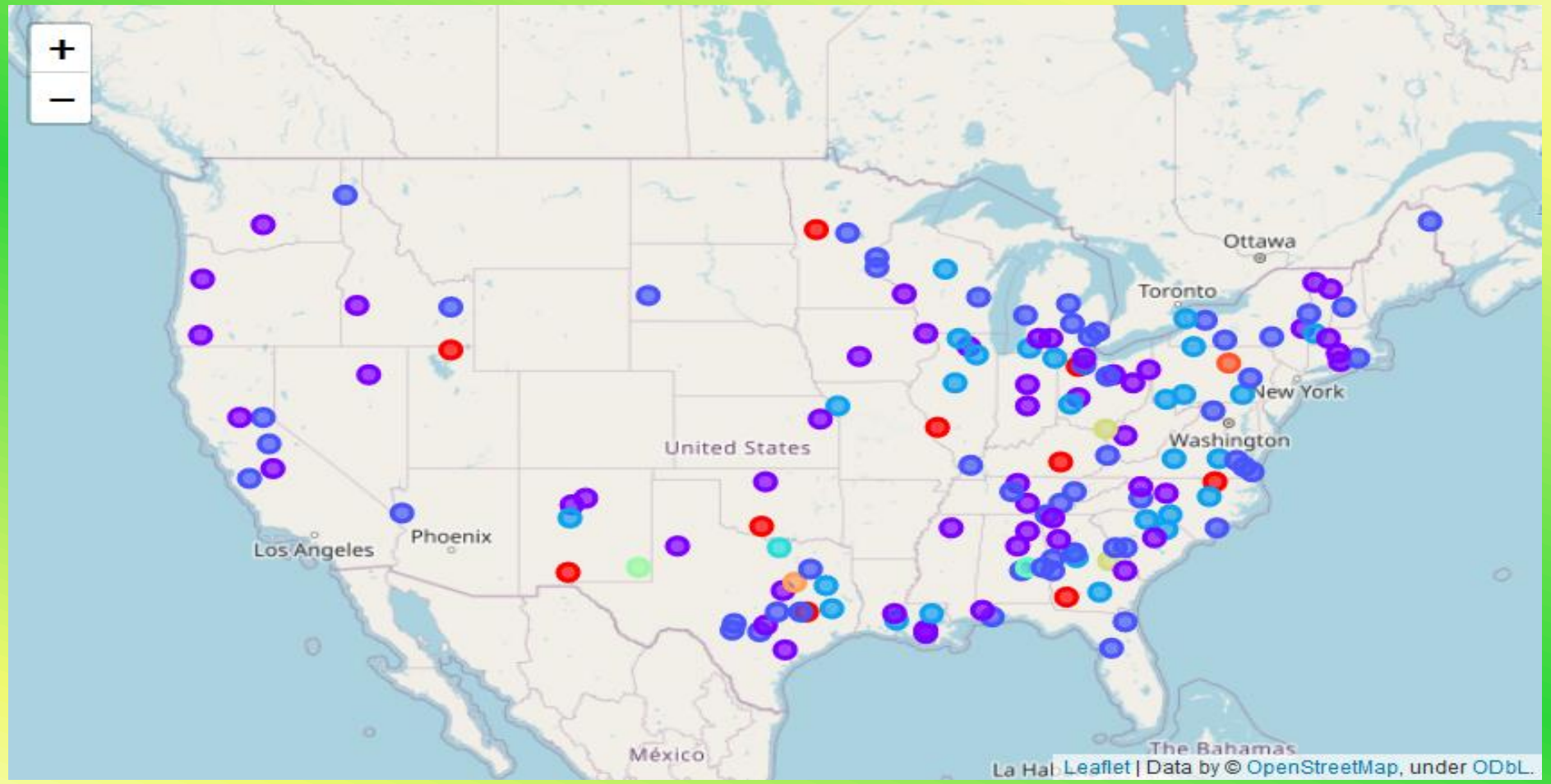
- Onehot or dummy encode our venue categories
- Then group by the county by the average of each category type.
- Above with the lat and long values in the group by we have to find the approximate center of each group of points to get one good lat and long location for each county.
  - . I am doing this by averaging the coordinates which will give alright results in small areas
- Find the index of our original home county
  - In my case this was 12

Step 5 Run K-Means

The final step is to run K-means to create groups of similar counties based on venue information so we need to drop the county name and location columns.

Then we do some trial and error to find a good k. In then end, I went with 10 for k. The final thing we do here is to then merge the cluster labels onto the original full data set

Map


+
−
Ottawa
Toronto
New York
Washington
United States
Phoenix
Los Angeles
México
The Bahamas
Leaflet | Data by © OpenStreetMap, under ODbL.

List of Places

1	Aiken, South Carolina, USA
4	Aroostook, Maine, USA
6	Autauga, Alabama, USA
7	Baldwin, Alabama, USA
12	Bonneville, Idaho, USA
14	Brazos, Texas, USA
18	Calumet, Wisconsin, USA
20	Catawba, North Carolina, USA
21	Catoosa, Georgia, USA
23	Chisago, Minnesota, USA
26	Clay, Florida, USA
27	Clinton, Michigan, USA
29	Columbia, Georgia, USA
32	Crow Wing, Minnesota, USA
40	Fairbanks North Star Borough, Alaska, USA
43	Franklin, New York, USA
45	Genesee, New York, USA
46	Grafton, New Hampshire, USA
47	Guadalupe, Texas, USA
51	Hampton, City of, Virginia, USA
54	Hernando, Florida, USA
57	Hunt, Texas, USA
58	Jackson, Georgia, USA
59	James City, Virginia, USA
60	Jefferson, West Virginia, USA
66	Knox, Tennessee, USA
67	Kootenai, Idaho, USA
74	Lee, Alabama, USA

List of Places cont.

75 Lenawee, Michigan, USA
83 McCracken, Kentucky, USA
85 McMinn, Tennessee, USA
86 Medina, Texas, USA
88 Midland, Michigan, USA
90 Mohave, Arizona, USA
91 Muscogee, Georgia, USA
92 Muskegon, Michigan, USA
95 Newport, Rhode Island, USA
100 Pender, North Carolina, USA
101 Pennington, South Dakota, USA
102 Pike, Kentucky, USA
103 Placer, California, USA
107 Portage, Ohio, USA
110 Richland, Ohio, USA
113 San Benito, California, USA
116 Schuylkill, Pennsylvania, USA
118 Shiawassee, Michigan, USA
122 Steuben, New York, USA
126 Taylor, Texas, USA
129 Troup, Georgia, USA
130 Tuolumne, California, USA
133 Van Zandt, Texas, USA
135 Virginia Beach, City of, Virginia, USA
136 Washington, Minnesota, USA
137 Washington, New York, USA
141 Williamson, Tennessee, USA

Discussion

- In the end I went with 10 for k as that gave me a result list of around 50 to 60 places out of the 150 sample I started with. I think the high k was necessary due to the large number of counties and would probably have to be higher with the full list of counties I had initially with ~870 counties. Interestingly, each time I ran k-means with the same k I got very similar results and my target group seemed to stay fairly consistent. This in my opinion shows that the analysis is valid.
- From the map we can easily find the home county of Bonneville, Idaho if you know Idaho geography (it is in the southeast part of the state). Also, it turns out that there are many places that are similar. None of them are right in big cities but near them. This is good as it means that there are multiple places to live similar to my home county that will due to proximity to bigger cities likely have better jobs and wages.

Conclusion and Final Thoughts

- With the final results, it would be beneficial to do a secondary analysis of those areas with additional demographic, housing, price, and job data. However, that is outside the scope of this analysis which was simply to find similar areas.
- Based on the results, I would suggest living where work can be found. As with either a reasonable commute or some effort in the job hunting, one can find a place like home in more places than they would initially think.