Capstone Final Report

Introduction

Failed Ideas

1. My hometown has a lot of grocery stores clustered in the middle of it and everyone as the city has grown has to drive in. I would like to identify where new grocery stores would be best put to reach a good number of people and be away from the others. That is to say to stop making people drive in so far to get a to a store.
2. Do the same as one but for gas stations which are in a similar situation.

Why They Failed

Why they failed is due to the fact that my hometown of about 60,000 people being the second largest in my state is too small to have enough data to do this on. I can get the gas stations and grocery store locations. but the smallest geographical boundaries I could obtain where zipcodes. The zipcodes as shown on the official USA zipcode map had one small inner city one and then there was a second one that covered most of the rest the city. Then the edges of the city barely reached into 4 additional zipcode areas. These would not have been good break downs for a distribution within my city for the above problems. So, I couldn't do anything smaller than a city level analysis in my hometown.

Other Failed Ideas

I also considered that since my state has lots of National Forests and campgrounds. I could identify different qualities of campgrounds and build a state camping map by campsite quality. This failed as Foursquare does not contain any information for campgrounds in National Forest Land. While I could get that data from the National Forest Service directly, there would be no use of Foursquare in the analysis which is required.

So, What did I do

I decided at this point that my hometown which I love is a bad idea for data science application. So, I decided to go with a simpler and in the end a more personal application as follows below.

The Problem

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

Recent college graduates and those just starting their own family often worry greatly about where to live. This is usually decided for them by a job as to what city, state or country their job is located in. However within a given city, state or country, there is a great diversity between states, counties, or even different neighborhoods. Thus keeping in trying to live somewhere similar to where they grew up but still having opportunity to have work with good wages.

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.

The 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

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 as large cities and metropolitan areas that cover multiple counties can be split into pieces and small places where several cities are in a county can be grouped together.

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. This removes a little more than the top 100 most populous counties in the US according to 2010 Census population counts as reported on the following Wikipedia page [Most Populous Counties in the US]

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

Here I imported all the useful libraries and functions as can be seen in the notebook in the code. 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. Thus, I could obtain the population center and not geographical center of the counties. This was very important as in most places the population center which is usually the county seat is very different from the geographical center that geocoders return.

Step 2 Wrangle the Data

This involves cleaning, formatting, and initial exploratory analysis. Most of the initial exploratory analysis was done in the researching of the topic just by looking through the data and thinking about the problem so only a few things were directly examined here.

Firstly, I selected only the relevant columns from the county table from Wikipedia. Then I filtered my list from the original 3000+ US counties and equivalents to those that fit my population parameter of 50,000 to 500,000. 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. If the rate limit wasn't a concern then I would gladly have done the full 870 or so that I had. I obtained the sample using pandas sample function.

Finally I remove the home county from the rest of the data so that I could track and examine it separately and avoid possible duplication somewhere. In the end, I figured out a way to where I didn't need to this and simply just re-appended it later on

Step 3 Get Venues

Here we use a function to obtain the venues for our entire county list. To keep things simple, I used a limit of 25 venues and 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

Now to get the venue data into a form that we can use it for more analysis. First, we onehot or dummy encode our venue categories. This puts them into a numeric form that we can actually use for further operations.

Next, we then group by the county by the average of each category type. This gives us a table of each county with its venue category types with each type assigned a score of how common it is to that county. To this I do the same to the home county and then use the append function to add it on making sure to enforce all NA values to zero using fillna. This way we have our full data.

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 like within a city or so where we can essentially say the earth is "flat" which it is not but in a small area is close enough, but not across a state or larger due to spherical corrections being necessary. A better method would be to use a centroid algorithm or convert to Cartesian vectors and then average there and reform back into lat and long. Since those methods are very complicated and somewhat hard to implement by hand, we will just use our approximation for now with no additional coding beyond the mean in the group by that is already used.

Final thing to note here now that we have our full data with location and venue information grouped by each county, we need to go find the index of our original home county. I found this just by scrolling through the list to where it was as the list sorts to alphabetically by default.

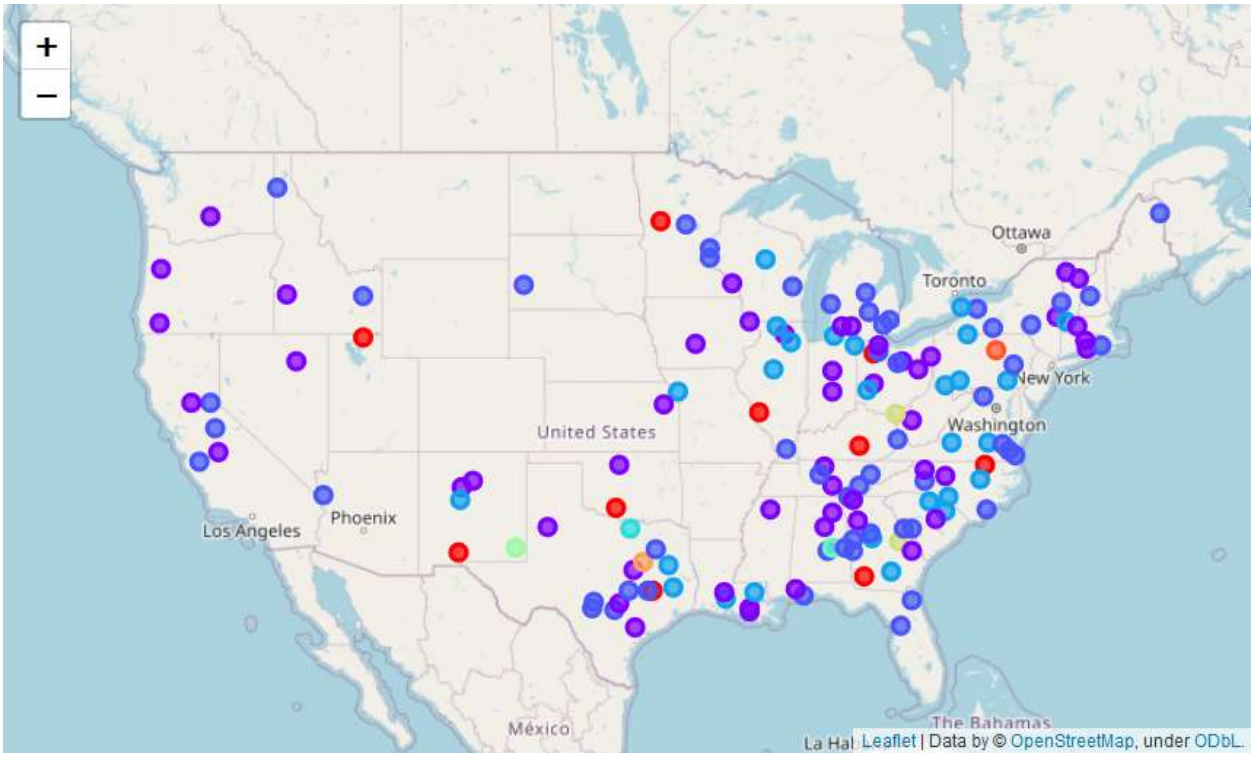
Home county of Bonneville for me was index 12 this will be used to identify its cluster later on after we run K-means.

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

Results

Map of all data minus Alaska so that the majority of data can be seen.



List of Places That Similar to Home

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
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
```

As is normal with K-means each time it is ran it comes out different. In my case I ran it with various k's a few times over to find a good k. 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.

Discussion

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. Interesting to note is that there was not a hidden geographical bias in the data. That means that my results in other words my target are not grouped into one area but instead are spread throughout the US fairly well matching with population centers. Another thing to note is that 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.

Conclusion

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.