Coursera Capstone Project Presentation

Finding Similar City Analysis

Isaac Trussell

September 22, 2019

Problem Introduction

I currently live in Boise, ID. I have recently decided that I would like to make a change and move to another part of the US for both career and personal reasons. I don't really have a destination and am pretty open to a lot of different options. The thought of leaving Boise and starting brand new at a brand new place really scares me and seems like a big leap especially since I don't really have any problems with Boise itself other than that the job market for my current career field of mechanical engineering isn't exactly booming. This lead me perfectly to my idea for this project. I should cluster cities that have similar attributes to that of Boise. So basically my goal is to find a city that is super similar to mine and I can look for jobs in that area and it won't be that big of a change for me.

Objective

Find cities that have similar qualities to Boise, ID

Project Outline

The basic structure of the project should look as follows:

- Find data about population and location
- Find data about temperature and climate
- Gather information using the Foursquare API about popular venues in the area
- Cluster the cities based on similarities

Future steps for this analysis could include:

- Find job posting information for mechanical engineering jobs in the area
- Cost of living comparison
- Larger venue search radius

Data Sources

After a quick search for city data, it was pretty easy to find general information about different cities online. The first table that I came across was a wikipedia article that had the following information that I plan to use in my analysis:

- City
- 2018 Population Estimate
- 2010 Census Population
- 2016 Land Area
- 2016 Population Density
- Latitiude & Longitude

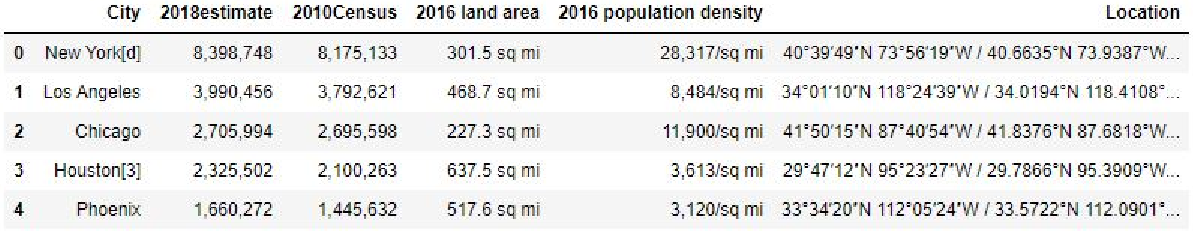
My next data source was found here. It has some interesting climate data for lots of different cities. Unfortunately it is smaller than that of the wikipedia data, so we will lose some cities that we can compare with Boise. The data I will be using from this website are the following:

- Average Winter Temp
- Average Spring Temp
- Average Summer Temp
- Average Fall Temp
- Average Yearly Precipitation
- Average Yearly Snowfall

Lastly, I plan to use the Foursquare API to pull popular venues in the area. I am going to categorize the venues based on their parent "category" as alot of the categories go into great depth. For example, I want to try and add up all the venues that are in the area that are simply resturants, rather than having 1 american food resturant and 1 chinese food resturant. For this high level analysis I don't think I need that in depth of information quite yet.

Methodology

I started by scraping the wikipedia article and making an initial DataFrame of the data that was there. That left me with this DataFrame:

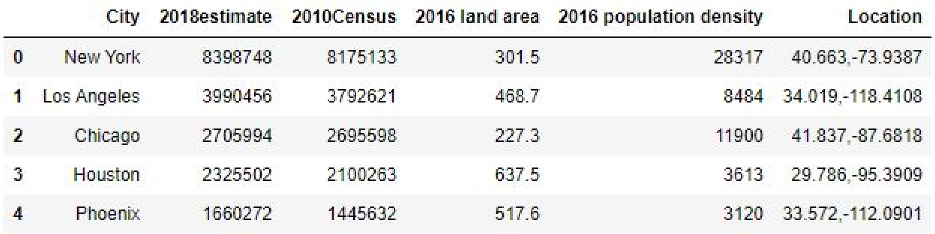


	City	2018estimate	2010Census	2016 land area	2016 population density	Location
0	New York[d]	8,398,748	8,175,133	301.5 sq mi	28,317/sq mi	40°39′49″N 73°56′19″W / 40.6635°N 73.9387°W...
1	Los Angeles	3,990,456	3,792,621	468.7 sq mi	8,484/sq mi	34°01′10″N 118°24′39″W / 34.0194°N 118.4108°...
2	Chicago	2,705,994	2,695,598	227.3 sq mi	11,900/sq mi	41°50′15″N 87°40′54″W / 41.8376°N 87.6818°W...
3	Houston[3]	2,325,502	2,100,263	637.5 sq mi	3,613/sq mi	29°47′12″N 95°23′27″W / 29.7866°N 95.3909°W...
4	Phoenix	1,660,272	1,445,632	517.6 sq mi	3,120/sq mi	33°34′20″N 112°05′24″W / 33.5722°N 112.0901°...

Its not super pretty and all of the fields dtypes were objects. We need to clean all this up and make it usable. I made everything usable by using the .apply function and creating custom functions for each of the following columns.

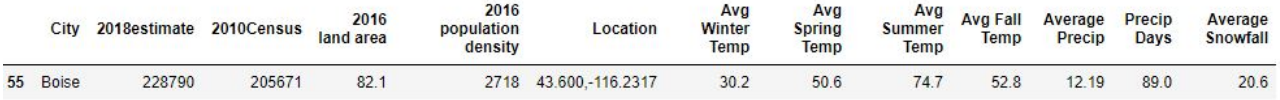
```
removeComma = lambda x: int(x.replace(",", ""))
land_clean = lambda x: float(x[:-6].replace(",", ""))
pop_clean = lambda x: x[:-6]
remove_hyper = lambda x: x[:-3] if x[-1] == "]" else x
def location_cleaning(x):
    x = x.split("/")[1]
    x = x.split(" ")[1:3]
    new = []
    for i in x:
        temp = []
        if i[-1] == "S" or i[-2] == "W":
            new.append("-" + i)
        else:
            new.append(i)
    return ",".join([a[:-3] for a in new])

wiki_data["2018estimate"] = wiki_data["2018estimate"].apply(removeComma)
wiki_data["2010Census"] = wiki_data["2010Census"].apply(removeComma)
wiki_data["2016 land area"] = wiki_data["2016 land area"].apply(land_clean)
wiki_data["2016 population density"] = wiki_data["2016 population density"]\
        .apply(pop_clean).apply(removeComma)
wiki_data["City"] = wiki_data["City"].apply(remove_hyper)
wiki_data["Location"] = wiki_data["Location"].apply(location_cleaning)
```

This gave us a cleaner looking DataFrame:



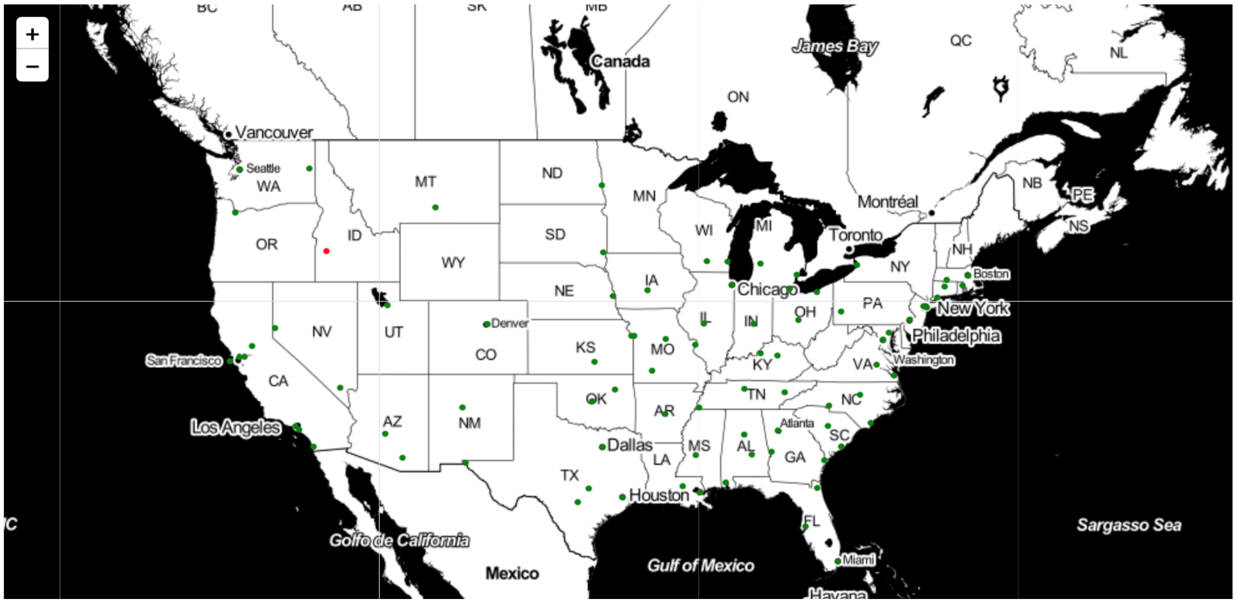
	City	2018estimate	2010Census	2016 land area	2016 population density	Location
0	New York	8398748	8175133	301.5	28317	40.663,-73.9387
1	Los Angeles	3990456	3792621	468.7	8484	34.019,-118.4108
2	Chicago	2705994	2695598	227.3	11900	41.837,-87.6818
3	Houston	2325502	2100263	637.5	3613	29.786,-95.3909
4	Phoenix	1660272	1445632	517.6	3120	33.572,-112.0901

Next, I did a similar process for the temperature data. This dataset had less cities and I had to drop a couple cities out of the analysis. After dropping cities we were left with 87 cities and our feature set looks like this:



	City	2018estimate	2010Census	2016 land area	2016 population density	Location	Avg Winter Temp	Avg Spring Temp	Avg Summer Temp	Avg Fall Temp	Average Precip	Precip Days	Average Snowfall
55	Boise	228790	205671	82.1	2718	43.600,-116.2317	30.2	50.6	74.7	52.8	12.19	89.0	20.6

Using the folium package we are able to see the distribution of these cities on a map of the US. All cities included in our analysis are shown in green, and Boise is shown in red.



There seems to be a large amount of cities on the east coast compared to the west coast. The only cities in the northwest, which I consider similar to Boise, are Spokane, Portland, and Seattle. Continuing on, I want to look at popular venue types in all of the cities. I want to group venues by their parent category and add them up for up to 75 venues in 3 mile radius. The parent categories that I drew from were as follow:

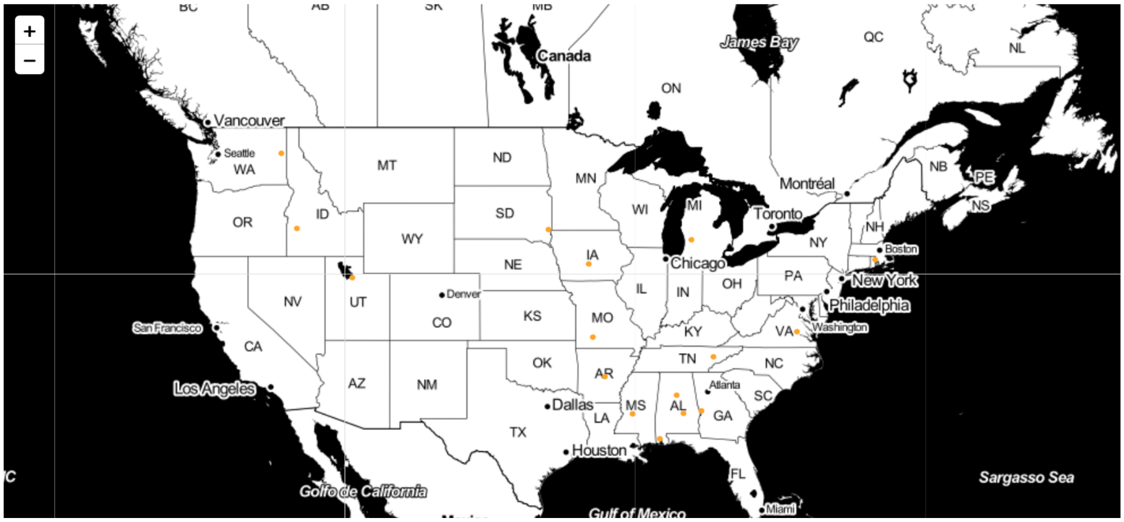
- Arts & Entertainment
- College & University
- Event
- Food
- Nightlife Spot
- Outdoors & Recreation
- Professional Places
- Residence
- Shop & Service
- Travel & Transport

After collecting all the items I normalized each of the rows by getting the average that each venue occurred in each city. Adding this last feature set left us with 23 features for each city.

$$\frac{\text{\# of occurrences in the city in the category}}{\text{Total \# of occurrences in the city}}$$

Finally, it was time to run create a model and get some results. I used KMeans with a number of clusters (*n*) of 7. I played around with a couple different *n* values, but was the most happy with the levels of separation that 7 had. This model output 36 cities with the same label as Boise. I was looking for a smaller set of cities to look into, so I decided to run the KMeans algorithm again on only the dataset that had the same label as Boise and see if I was able to get even more distinction in the cities. I ran it again, this time with a *n* value of 3. This almost cut the cities with the same label as Boise in half and I found my final results.

Results



The final 15 cities that shared the same label as Boise were:

Richmond	Spokane	Des Moines	Birmingham	Salt Lake City
Grand Rapid	Montgomery	Little Rock	Colombus	Mobile
Knoxville	Sioux Falls	Providence	Spingfield	Jackson

There was at least one city in every major region of the US in the final cluster as Boise. Looking at the other labeled data, a lot of it seemed to make sense to me, especially in population. LA and New York both were the only cities in their label. Large metropolitan areas like Austin, San Fransisco, Boston, Miami, and Detroit all fell into the same label.

Discussion

I would like to see some results on how impactful every feature was on the final clusters. It seems like population was a very heavy influence on the clustering of cities. The difference from largest to smallest population was only about 30,000, which is pretty small given the range of populations that were in the dataset. I was surprised that the cities that were in the cluster were all over the US, my expectation was that the majority of similar cities would be in the northwest.

Conclusion

I think that this analysis needs move venue heavy analysis rather than purely city statistics. I think that the model was too heavily influenced by population data rather than the other important things that make cities similar. I wished that my categorized venues were more similar in the final dataset.