Coursera Capstone Final Project: Weeks 1 & 2

Miguel A. Villarreal

February 2020

I'm a Texan, Where Should I Go When I Visit or Move to Other Texas Cities? A Clustering Approach to Texas Neighborhoods

Introduction & Business Problem

Texas is the second-most populous state in the United States at over 29 million residents and growing rapidly, adding an estimated 500,000 net new residents every year of the last decade.[1] Three of the largest cities, Houston, Dallas, and Austin are among the country's fastest growing cities. The 3 cities are located within relatively close driving distance (2-4) hours and frequently exchange visitors and residents between and among them.

Visitors and potential new residents from one city may need a guide when visiting or moving to a new city. In this context, a frame of reference relative to their origin city may prove beneficial to them. For example, if I am moving from Dallas to Houston, I may want to explore neighborhoods in Houston that are similar to my neighborhood in Dallas. This data analysis is designed to help supplement such a system by finding intra- and inter-city clusters of neighborhoods that have similar venue profiles.

While there are many other potential data points to use for this type of analysis (demographic information, real estate price analysis,) this analysis will focus solely on the distribution of retail venues for the purposes of brevity, and to fulfill the Coursera condition of using the Foursquare

[1] See U.S. Census Data, 2018

API. Likewise, the use of retail -oriented venue data is dual use in that it can be used to power applications for both potential visitors to a region, and to potential new residents. Possible next steps for adding more depth to the data analysis will be discussed at the conclusion of this project.

The practical application for this data analysis in the business context could be:

- Recommendations in travel/city guide websites;
- location-based mobile applications that suggest possible destinations based on user location, user origin, or user preferences; and
- Recommendations for potential retail business owners and/or venue operators looking for neighborhoods with certain profiles that would support their business

Proposed Approach using Data Science Tools & Methodology

I propose the following approach to preparing the report

- Obtaining a list of all neighborhoods/locations in the target cities (Dallas, Houston, Austin)
- Call the Foursquare API to obtain retail venue frequency locations
- Use K-Means Clustering to group the neighborhoods into categories, regardless of city location
- Use these clusters & groupings to derive meaningful correlations between and among different parts of each city.

1. Obtaining Datasets

For this step, I propose to obtain a list of neighborhoods with geographical coordinates in Texas via public data. The primary identifier will be postal/zip codes. I was able to obtain a table of zip codes and coordinates via: https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?refine.state=TX

From this site, one can filter the data down to the 3 desired metropolitan areas using the “city” field. The site makes the data exportable into .csv and other formats that will each export cleanly into a Pandas dataframe.

Later, though by no means essential, it may be optimal to add in neighborhood names rather than just zip codes, depending on the business requirements for the final product. For now though, since they are readily available, I will use zip codes as a proxy for the neighborhood and work through any issues that it generates (for example, zip codes may be coded as integers initially but I may need to convert them to strings).

2. Cleaning, Calling & Clustering

After obtaining the list of zip codes & locations and suitably cleaning the data using techniques like joining & merging tables and other Pandas commands, I will call the Foursquare API to get a categorical list of common venues in using the coordinates of each zip code in my dataframe.

After this step, I will use one hot encoding and other techniques to clean the data so that it may be clustered appropriately.

When the data is finally ready, I will use k-means clustering based on frequency of venue categories to create a list of possible categories.

3. Analysis & Next Steps

Ideally, the categories will break down into a broad geographic distribution and not be clustered entirely in a single city, for example:

Category 1 (Has high proximity to restaurants)
Houston Neighborhood A
Dallas Neighborhood B
Austin Neighborhood C

Category 2 (has high proximity to public parks)
Houston Neighborhood D
Dallas Neighborhoods A, D
Austin Neighborhood B

<continue until all categories defined>

Ideally, since this data set is not immense, it can be exported and further analyzed to determine if these types of groupings make sense. Additional data may be added to the data set at this point to supplement the findings. Furthermore I will spot check the data to see if the information being passed is valid or if the clustering model or data inputs need further adjustment.

There are a number of risks that I will need to watch out for, including:

Risk of Overfitting the Clusters - a bad result would be if the clusters are overfitted and all localized in particular cities. This would not be helpful for end users looking to explore or move to one city from another city. Likewise, uneven sized clusters may also be problematic.

Outliers Driving Categorization - another potentially bad outcome would be low frequency venues (Zoos, Stadiums, etc) creating their own clusters and creating superficial categories

Inadequate Categorization - this will be discussed more at the project's eventual conclusion but there is some risk involved in relying solely on venue categorization, as there may be considerable nuance within each category that is not captured by Foursquare Data.

At the conclusion of this step I will:

- Identify any red flags or warning signs that may necessitate further work on the data analysis **(**inadequate models, need better data sources, etc)
- Identify any preliminary conclusions and insights that can be drawn from the data that may be valuable to business stakeholders
- Identify next steps and recommendations that could hypothetically be taken to take this project to its logical conclusion

Week 2: Execution, Discussion & Conclusions:

Data Transformation & Analysis

A link to the Jupyter notebook which I used for this project is provided below:

https://github.com/fantasticmaverick/capstone/blob/master/Final%20Project%20copy%202.ipynb

A quick summary of the work performed is:

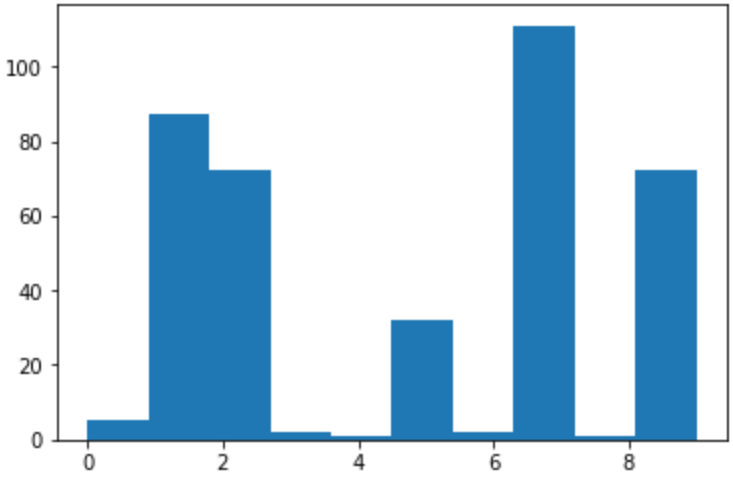
1. Obtain datasets on Houston, Austin, Dallas from public sources with lat/long information

2. Merge these datasets, explore, remove extraneous categories & prepare it for further analysis
3. Load Foursquare & define get requests to obtain nearest & most common venues to each coordinate set
4. Use One Hot encoding to obtain a sorted list of most common venues per neighborhood
5. Sort the table by most common venues
6. Apply k-means clustering to obtain groupings. I reran this using a number of different k-means numbers to see if any performed better or worse.
7. Export the data to csv for visual inspection/sanity check using Excel
8. Evaluate the data using matplotlib - in this case I used a histogram to examine the distribution into the various clusters.

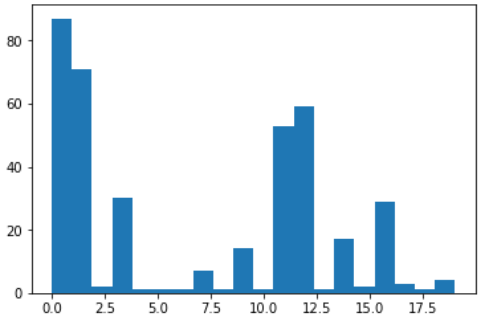
Results & Discussion

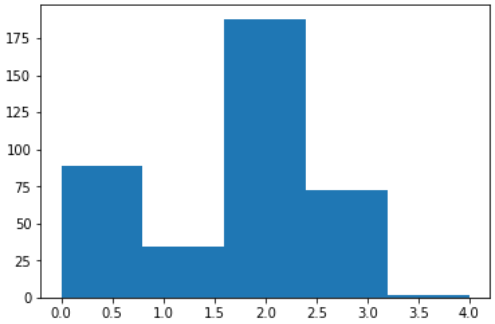
In terms of using clustering, the best performing model, initially, at least **appeared to be the use of 10 k-clusters** vs 5 or 20 k-clusters. This is based on two modes of inspection:

The 10 cluster grouping appeared to create the least-lumpy distribution of neighborhoods/zip codes into ~ 5 well defined bins, with ~5 thinly populated outlier groupings



This can be contrasted with the 20 and 5 cluster groupings, below. The groupings in the 20 cluster created a number of well defined broups but at least 10 or more outlier groups and the data, after further inspection was too thinly sliced to be meaningful. The groupings in the 5 cluster set were overbroad, with not enough meaningful correspondence between the various groups after further inspection.





As the data set is not enormously large, it can be exported locally to CSV and further analysis can be performed. Furthermore there is no “gold standard” for clustering neighborhoods among these three geographies to perform a python-based model evaluation so human analysis is necessary here (and manageable, given the small size of the data set.)

A link to the csv is here:
https://github.com/fantasticmaverick/capstone/blob/master/data%20(9).csv

Evaluation of Meaningful Clusters using 10 cluster model

Cluster 0

Cluster	Number of Zips	Cities
0	5	3
1	87* (3 coordinates)	2
2	72* (2 coordinates)	2
3	2	1
4	1	1
5	30* (3 coordinates)	2
6	2	1
7	111	3
8	1	1
9	72	3

*A data anomaly appeared in these categories where the initial coordinate data set appears to list a large number of city under the same lat & longitude values, I verified this with the initial

reference data, and determined this anomaly did not occur during processing; , this will have to be further studied to determine if it affects the validity of the model.

The ideal clusters will be ones that feature a number of cities, therefore I will analyze clusters below that yield results in multiple cities

Analysis of Meaningful Categories

Here I will analyze each multiple-city categorization to see if we can draw any conclusions about the clusters

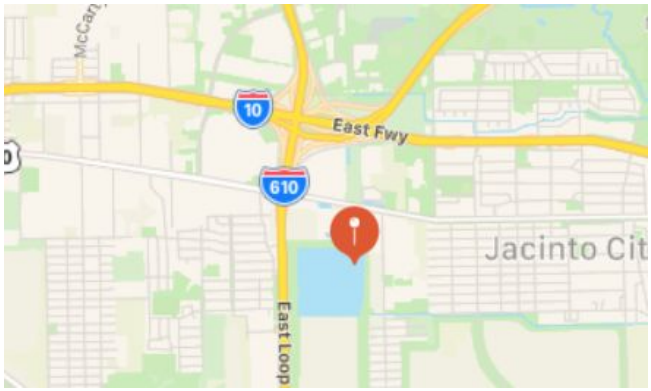
Cluster 0

The zips in cluster zero all appeared to have the common feature of having a Zoo, campground, and construction supply shop nearby. Ostensibly, the model performed adequately in clustering here.

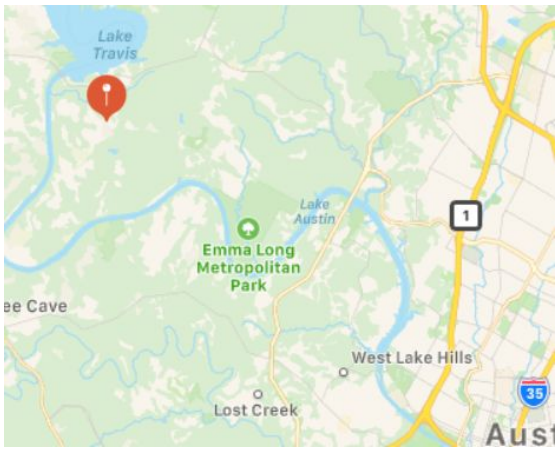
These locations all appear to be at or near the geographic fringes of their city so these may be meaningful categories, for example:

75236 (Dallas)

77029 (Houston)



78732 (Austin)



Cluster 1

Cluster 1 features 3 general neighborhoods in 2 cities

East Little York, Houston, a lower middle class neighborhood, Candlelight Plaza, Houston, an upper middle class neighborhood, and Fox Run Ridge, an expensive suburb of Austin.

THe model did not perform well here, as these neighborhoods are all fairly distinct demographically. It is interesting that they did not work under any clustering model. They appear to be outliers or just not good fits for this type of analysis.

Cluster 2

Cluster 2 features 2 locations in 2 cities, the neighborhoods correspond to:

South Boulevard, Dallas and Acres Homes, Houston. Both of these neighborhoods share the same demographic makeup in terms of being lower-income but gentrifying neighborhoods and a selection of venues to match.

In this instance, the model appears to have performed well.

Cluster 5

Cluster 5 features 3 locations in 2 cities:

Rowlett, a Dallas suburb, Balcones Park, an Austin suburb, and an area of Southwest Dallas suburbs.

The first two locations offer a similar

The model performed somewhat satisfactorily here, with some moderate correspondence between 2 of the 3 neighborhoods.

Cluster 7

Cluster 7 features a wide variety of locations in all 3 cities. Including dozens of locations in and around Houston, including Meyerland, Downtown Dallas, Downtown Houston. Downtown Austin and the University of Texas.

There appears to be a fair amount of correspondence with the locations in this cluster, however it may be overbroad owing to the large number of central downtown areas included and it may be difficult to disambiguate the correspondences here

THe model performed well, the results may not be helpful however to business stakeholders.

Cluster 9

Cluster 9 is similar to cluster 7 in that it features a wide variety of locations in all three cities. Like cluster 7, there is a fair amount of correspondence among the reason, however it may be difficult to disambiguate the regions.

Next Steps & Conclusions

A clustering model performed reasonably well here to correspond the 3 Texas city neighborhoods in certain instances. However there were notable deficiencies that need further study or clarifications:

Lumpy Distribution - the bulk of neighborhoods were fit into 1-2 categories, regardless of the number of clusters that were formed. My hypothesis is that the fact that these areas tended to

be central business districts is a function of radiuses in the Foursquare data call. Basically, for certain zip codes/neighborhoods, the radius of nearby venues **should be drawn smaller** to account for the greater proximity of venues. This will result in a more diverse array of results and potentially a better clustering model

Add more variables - the radius issue above and certain other factors may indicate that solely using a neighborhood-venue oriented approach is not adequate to cluster these locations. Other variables may be needed to be added to the model, such as

1. Demographic data, like income, real estate value, median age, density etc
2. A way to control for outliers - low frequency items like stadiums, zoos, etc may distort the results
3. The overall list of locations may have been too small. It's possilbe that if I include additional large cities in Texas such as San Antonio, Fort Worth, Arlington, El Paso the clustering will result in a more appropriate outcome.

Final Thoughts

In conclusion, I would say that a neighborhood clustering approach shows *some* promise in terms of analyzing similarity of Texas city neighborhoods, but that a somewhat higher level of data inputs and model refinement will be required to make it commercially meaningful.