Coursera Capstone Final Project, Week 1

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 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

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

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