Texas Sized Data

An Exploration of Clustering Texas City Neighborhoods for Practical Applications

Business Problem

Texas is one of the fastest growing states in the United States, adding nearly 5 million residents in the last 10 years.

A substantial proportion of that population is located in Texas largest cities, all of which are located within a few hundred mile radius in the central, central-northeast, and southeast portions of the state.

With millions of new residents and visitors every year, as well as hundreds of new and diverse neighborhoods to explore there may be a benefit to using data analytics to cluster neighborhoods with an aim of surfacing them to Texas visitors and residents as they travel throughout the state.

Business Problem (cont'd)

The practical application for a well-clustered list of Texas city neighborhoods includes:

Recommendations for Travel & Relocation/Moving Guide websites

Location-based mobile applications that suggest recommended destinations to users based on user location or user's home location, when visiting a target location

Descriptive Analytics for Business Owners to be used in analyzing neighborhoods for potential new venue locations

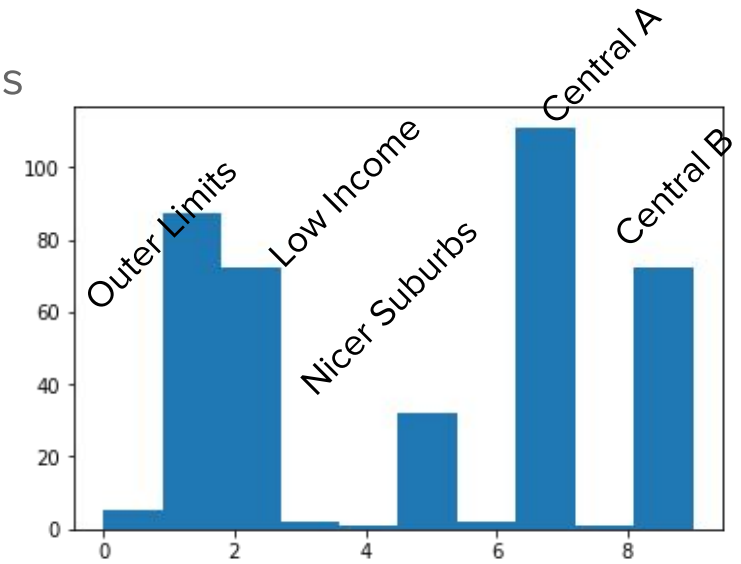
Methodology

- Choose 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 cluster to derive meaningful correlations between and among different parts of each city.

Results

Using K-Means Clustering, the results resulted in a histogram that identified approximately 4 meaningful groupings of neighborhoods:

1. Outer Limits Neighborhoods
2. Lower-Income Gentrifying Neighborhoods
3. Nicer Suburbs
4. Central City Group A
5. Central City Group B



Practical Applications

From the Data Analysis generated, a number of clusters can be generated to drive relationships between neighborhoods.

The sample chart below indicates some of the possible recommendations that might be made from the data analysis. Visitors from one neighborhood have several affinity neighborhoods in the other cities:

Visitor origin: Dallas	Affinity Neighborhood:Houston	Affinity Neighborhood: Austin
Oak Lawn/Turtle Creek (75219)	Rice Military (77007)	Old West Austin/Tarrytown (78703)
Uptown/Cityplace (75204)	East Downtown (77703)	West Campus (78705)

Note this is **sample data only** - the full dataset can be used to derive hundreds of these affinity groupings.

Deficiencies in the Modeling & Remediation

The primary deficiencies in the modeling was the sorting of the most dense, popular neighborhoods predominantly into 2 large buckets.

Next steps in terms of remediating this modeling should/will be, if the project is carried further:

- Application of clustering **within** the large clusters to get a better result set
- Development of a way to account for central business district venue proximty vs. outer suburb venue proximity, which distorts the model, potentially using two different calls to Foursquare
- Inclusion of additional data (demographics) of each neighborhood to develop a better multifactor clustering model.