Applied Data Science Capstone Project

Battle of the Neighborhoods

By: Susan Terrillion, July 2020

Purchasing rental properties in the Dallas-Fort Worth Metroplex

Business Problem

- The DFW area is spread out over 30 miles which makes it hard for a potential investor to truly get a feel for the different cities where rental investment opportunities may exist.

- Our objective here is to cluster and analyze data from both Dallas and Fort Worth.   We're looking to answer the question: **"Which of two cities, Dallas or Fort Worth, offers the greatest number of amenities for its residents?".**

- Investors interested in purchasing in the DFW area real estate market have a hard time isolating where to purchase.   Developers interested in converting existing structures into new apartments will find this information relevant also.

Data Sources

- For each of the cities, the zip codes and longitude/latitude values were gathered using Opendatasoft’s public data and fetched as a .CSV file.

- The FourSquare API was used to fetch all of the venues available in each defined segment in JSON. This data was summarized by venue category to support the investment opportunity analysis provided here.

These data sources will allow us to segment, cluster and map similar zip codes in the two cities.

Data Cleansing

- Data was loaded from both the Opendatasoft's Dallas and Fort Worth .CSV files separately
- The data was scrubbed to only include ZIP CODE, CITY, LONGITUDE, and LATITUDE.
- Duplicate rows were dropped from the data where multiple rows contained the same longitude and latitude values.
- The Dallas and Fort-Worth data sets were merged to be used by the Foursquare API to load venues for clustering.
- The Foursquare data was sorted by category with counts.
- Onehot encoding was used to convert the category items into numerical values to apply k-means clustering to the sample.
- Finally, each row was grouped by zip and categories given a frequency.

Methodology

- The data cleansing process prepared the Foursquare data as input to the k-means clustering tool
- It was determined that 8 clusters were appropriate for clustering the zips represented in the data set.
- After clustering, the cluster labels were merged into the DFW set with all values present.
- Geopy was used to ascertain the coordinates for each city for generating a map of the city and it’s clusters.
- A map for each city was then generated to visualize the clusters of zip codes.

Results

- The k-means clustering results showed that:
  - The largest cluster included 36 similar zip codes -- 19 clusters in Dallas and 17 clusters in Fort Worth.
  - The second largest cluster included 29 zip codes, with Dallas showing 24 of those and Fort Worth with only 5 zips.
  - The next largest cluster included 13 zip codes with 6 zip codes in Dallas and 7 in Fort Worth.
  - The five additional clusters contained only one zip code, making those clusters dissimilar to the other 3 clusters. Of these 5, Dallas was shown to have 3 locations and Fort Worth only 2.

Conclusion

- The analysis done here showed that *Dallas* is a more attractive investment opportunity than *Fort Worth* based on the number of venues identified for each city.
- A future analysis may include the introduction of purchase vs. rental value for the densest clusters to help with ROI analysis by investors