**Predicting Restaurant and Construction business Locations in DFW, TX**

Yosef Woldeamanuel

November 14, 2019

1. **Introduction**
   1. **Background**

The Dallas–Fort Worth metroplex, officially designated the ***Dallas–Fort Worth–Arlington, TX*** Metropolitan Statistical Area by the U.S. Office of Management and Budget, is a metropolitan area in the U.S. state of Texas encompassing 12 counties and has a population of 7,539,711 according to the U.S. Census Bureau's 2018 population estimates, making it the most populous metropolitan area in both Texas and the South, the fourth largest in the U.S., and the tenth largest in the Americas.

It is the economic and cultural hub of North Texas. Residents of the area also refer to it as DFW, or the Metroplex. The region’s economy is primarily based on banking, commerce, telecommunications, technology, energy and healthcare. In 2019, DFW is home to 25 Fortune 500 companies, the third largest concentration in USA. In 2016, Dallas–Fort Worth ascended to the number one spot in the nation in year-over-year population growth. Due fast population growth knowing which area of the metropolitan will be suitable for different business will be advantages for new business owns during their business planning.

* 1. **Problem**

Data that might help to determining Restaurant Business type, location areas data for the ever-fast-growing Dallas-Fort Worth metroplex. This project aims to predict part of the metroplex where different restaurants are located.

* 1. **Interest**

Current restaurant owners who wish to expand their business and new restaurant owners in process of business planning would be very interested to understand which locations will be possible options for new restaurant.

1. **Data Acquisition and cleaning**
   1. **Data Sources**

List of cities, counties and zip codes for Texas can be found in [Zip\_TX](https://www.zip-codes.com/state/tx.asp) . Zip codes and their geographical location for the entire country can be found in [US Zip Code Latitude and Longitude](https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/). Population data for the metroplex can be found in [Dallas-Fort-Worth-Metroplex-Population](https://en.wikipedia.org/wiki/Dallas–Fort_Worth_metroplex)

* 1. **Data Cleaning**

All of these data contain information that are not relevant for this project and information for locations outside the metroplex. Data cleaning and extracting necessary information is required.

Texas Zip code data contains Zip code, City, county and Type for each Zip code in Texas. Type column is not important for this project and will be dropped.

Even if there are 12 counties in the metroplex, most of them are residential will not return much venue from foursquare. To clean up the zip code data, I used the top most populated counties in the metroplex.

The Zip code column contains additional information other than the number which I split the string and used only the zip code number.

Next the zip code latitude and longitude data needed cleaning since it contains the entire US zip code location data. Zip codes ranging from 75001 to 76670 are zip codes in the metroplex, which I used this range to create dataframe of metroplex zip code location from the bigger data.

After verifying the datatype of ZipCode columns on both data set are same, I used merge on zipcode to create the final metroplex data with zip code, city, county, latitude and longitude.

1. **Methodology**

K-means unsupervised clustering will be used to group cities in the metroplex based on certain similarities in the returned data from Foursquare. Even if K-means has disadvantage of predict K-value swipe at different K-value it will produce tighter clusters faster.

1. **Result** 
   1. **Population change**

Population change in the metroplex is as high as 23 % in Collin county with most of the counties showed population growth over 10%. The over all metroplex population change is 15.15%, which is higher than average growth for most cities in the country.

A screenshot of a cell phone

Description automatically generated

Comparison between population from 2010 census and 2017 estimated values for each counties and total population for the DFW metroplex.

A screenshot of a cell phone

Description automatically generated

* 1. **Clustering**

K-means clustering used to group the cities in to 6 different clusters. Based on the top 10 most common venues in each cluster, cluster 1, 5 and 6 are ideal places to open new restaurants due to currently available restaurants and other venues which drive people to go there either to attend sporting events or other activities the likes of banking, health/beauty and doctor's office visit.

A picture containing text, map

Description automatically generated

1. **Conclusions**

In this study, I analyzed restaurant locations in Dallas Fort Worth metroplex. I identified city of Irving, Grapevine and Arlington ideal places for restaurant owners to consider as their target market based on the currently available venues and population. I built bar graph and scatter plot to better understand population growth in the metroplex which to predict where future restaurant should be open to target the right market. Restaurant owner can use this data as firsthand indicator which are will be suitable to what kind of restaurant. However, due to fast population growth in the metroplex public available data might not be update.