**The Tech Migration to Dallas–Fort Worth Metropolitan Area**

****

**Author – Roopam Chakrabarty**

**Date: July 2020**

* 1. **Introduction - Business Problem**

Dallas–Fort Worth Metropolitan area has quickly become a hub for tech companies, tech departments of Banking and Financial Services organizations and start ups, with companies moving from high cost and denser areas like Bay Area and NY Metro. The migration is not only driving growth and innovation; it is also impacting the real estate market. The objective of this project is to cluster neighborhoods in Dallas–Fort Worth Metropolitan, so that migrating tech workers moving from Bay Area or any other Tech Metro areas across US can make an informed decision on the neighborhood to choose for their future homes.

**1.2 Who would be interested in this project?**

Few of my colleagues moved from the NY Metro area to Dallas–Fort Worth Metropolitan in the recent past, and many more are planning to move to Dallas–Fort Worth Metropolitan in the future. This project will recommend neighborhoods/zip codes based on facilities, thus helping my colleagues to make a prudent decision while choosing a neighborhood/zip code for their future stay based on their lifestyle.

**2.0 Data**

I will explore, segment, and cluster the neighborhoods based on the Zip codes in the Dallas–Fort Worth Metropolitan area. The Wikipage [https://en.wikipedia.org/wiki/Dallas%E2%80%93Fort\_Worth\_metroplex#Dallas%E2%80%93Plano%E2%80%93Irving\_metropolitan\_division[26][27]](https://en.wikipedia.org/wiki/Dallas%E2%80%93Fort_Worth_metroplex%23Dallas%E2%80%93Plano%E2%80%93Irving_metropolitan_division%5b26%5d%5b27%5d%20) has all the information we need to explore and identify the major cities in the Dallas–Fort Worth Metropolitan area.

Based on the analysis the important cities in the Dallas–Fort Worth Metropolitan area are:

* Dallas
* Plano
* Irving
* Fort Worth
* Arlington
* Grapevine

After exploring numerous websites, I found the website <https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/> , for getting all relevant information related to zip codes, latitude and longitude coordinates for Dallas, Plano, Irving, Fort Worth, Arlington, and Grapevine. The data in this website was downloaded in the form of a CSV file. I uploaded the CSV file in this project and converted it to Pandas Data frame.

Dallas–Fort Worth metro area consists of 7 major cities with population more than 200,000. These 7 major cities are Arlington, Dallas Fort Worth, Grape Vine, Irving, Lake Dallas, and Plano. These 7 major cities consist of 121 unique zip codes with distinct latitude and longitude. The CSV file was converted into the data frame consist of Zip code, City, State, Latitude and Longitude.

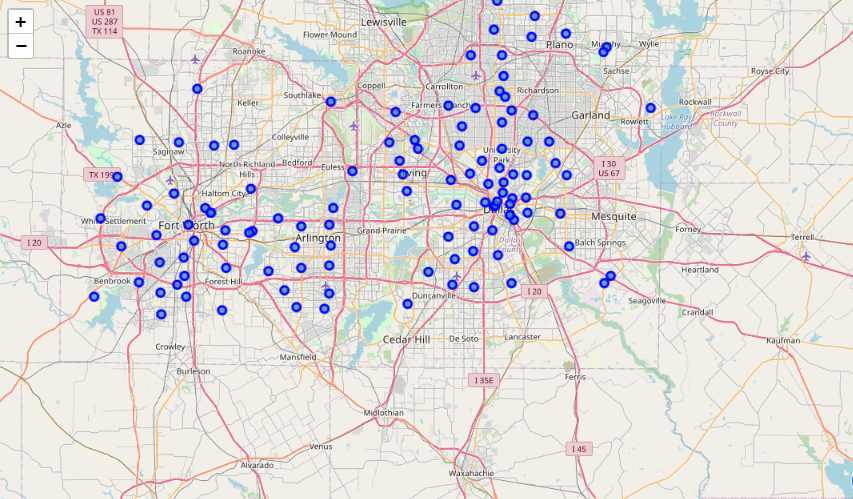


Also, I will use the Foursquare API to explore this neighborhoods/zip codes in Dallas–Fort Worth metro area. I will use the \*\*explore\*\* function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. I will use the \*k\*-means clustering algorithm to complete this task. Finally, I will use the Folium library to visualize the neighborhoods in Dallas–Fort Worth metro area and their emerging clusters.

**3.0** **Methodology**

After the data exploration and identifying the source of data, we will apply the K-Means machine learning technique for creating clusters of Zip codes representing similar facilities.

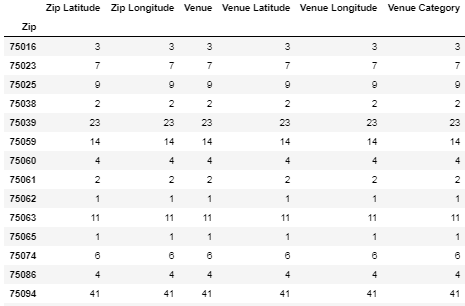
The converted data frame from the CSV file was loaded on a map leveraging Folium to explore if the zip codes were covering all the major cities that were part of Dallas–Fort Worth Metropolitan area.



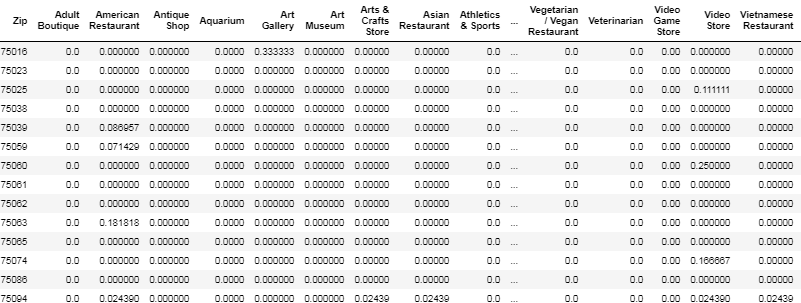
I defined a function to get all the neighborhood data using Foursquare API. The radius for the API call was restricted to 500 m and the number of response for venue was limited to 100. A new data frame was created for all the Zip codes along with the Venue information including the category.

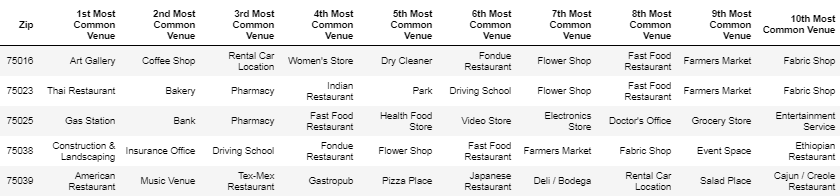


The number of unique venues for zip code was validated. 216 unique venue categories were found.



I analyzed zip code based on the unique categories of venue. The rows were group by Zip code and by taking the mean of the frequency of occurrence of each category.

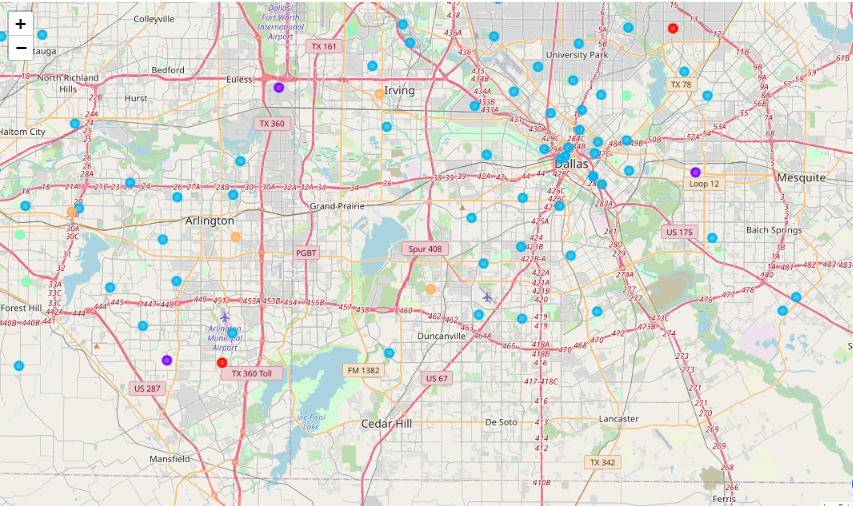


A new data frame was created to display the top 10 venues for each of the zip code.

The data was clustered using k-means. A new data frame was created that had the Cluster information along with the top 10 venues and zip codes.



Finally the clusters were visualized using folium map.



**4.1** **Result**

By looking at the cluster data, we can conclude that Cluster 3 is having most of the zip codes and the best place for settling down due to the wide variety venue categories required for a family to sustain a decent life.

The first cluster (Cluster label 0) is predominantly in an area with Athletics and Sports venue, so will not be suitable for living, but might be good for families into sports.

The second cluster (Cluster label 1) is having Mexican restaurant as the most common venue, so the zip codes part of this cluster might be liked by people loving Mexican food.

The third cluster (Cluster label 2) covers wide variety of categories with malls, convenience stores and supermarket. This is one of the clusters we can certainly recommend for living as some of the zip codes in this cluster are having condominium and apartment communities.

The fourth cluster (Cluster label 3) is having Construction and Landscaping as the most common venue. The zip codes in this cluster might not be suitable for living.

The fifth cluster (Cluster label 4) is also having variety of venue categories for living; also these zip codes are having parks, which can be liked by families and individuals believing in outdoor living.

**4.2** **Recommendation**

Based on the data Cluster 3 (Cluster label 2) and Cluster 5 (Cluster label 4) can be recommended for anyone planning to move to Dallas–Fort Worth Metropolitan area.

**5.0 Conclusion**

The project recommended neighborhoods/zip codes based on venue categories retrieved using Four Square API and k-mean clustering. The Clusters selected were based on the categories which were most common and anyone can go thru this detailed report and Notebook to select a zip code based on their personal preference. The analysis and project covered all the aspects of Data Science and Python Libraries such as Pandas, Scikit, Folium to name a few.