**Introduction**

A few friends have recently started an Iranian restaurant in Frisco, TX, which is a suburb located outside of Dallas, TX. The restaurant has done very well and they have begun thinking about trying to expand into the Dallas area. They are confident that the restaurant can do equally as well if not better in the Dallas area. However, this type of branching out would be a very big leap for the company and is critical to their success. However, if the new restaurant fails, this could be detrimental to their original restaurant and possibly cause it to take severe losses or even close down.

Due to the magnitude of this decision, the group wants to do everything they can to make sure they are absolutely confident about opening up the new store. One of the most important tasks with opening up a new store is figuring out the location of this new store. They would like to have a quick, efficient, and overall smart way to find a good location to possibly place the new store. Traditional ways of trying to find a suitable area may take too long or may not be as accurate (contacting a real estate agent, checking local listings, etc). So the group decided to reach out to a data scientist to help them find the best location for the new store.

The way to find the best location for the store is to classify areas in Dallas as "clusters" based on similar traits. Features such as types of venues and most popular venues in these areas will help the group identify similar neighborhoods, which will then help the group decide which area is more suitable for the new restaurant. From there, they can decide on specific locations.

**Description of the data that will be used to solve the problem**

Like stated in the introduction, we will have to break Dallas up into different areas within the city. After breaking Dallas up into these areas, we can further analyze them for types of venues to determine their similarities to each other. How similar they are will determine which cluster (or "group") they get put into. Once you determine which cluster each area belongs to, then you can decide which are the most optimal to place the new restaurant based on the group's criteria.

Our first step is to find location data for the city of Dallas, including zip codes, latitudes, and longitudes. We were able to find a free downloadable online source containing location data for all zip codes in the US, the only stipulation being that we had to mention the source providing this data. You can download the same data from <https://simplemaps.com/data/us-zips> . While the file contains many attributes for each zip code, we will only be focusing on zip code, latitude, longitude, city, county, and state. Using this information, we will then use Foursquare API to see types of venues and specific venues within these zip codes to help determine the clusters. We will also be using K-means classification to determine the optimal number of clusters, as well as the accuracy for each cities relation to the cluster.

The analysis of this data will be similar to the analysis done in the Segmenting and Clustering Toronto neighborhoods lab. Gathering all of the location data for a city, clustering the neighborhoods within that city, leveraging Foursquare API to determine number of venues and types of venues within each neighborhood, and using that to determine which cluster each neighborhood belongs in is how we completed that lab and how we will complete this analysis.