Introduction

This project will utilize multiple data sources including the US Census, Google Geocoder and Foursquare to build out a list of grocery stores in the Greater Dallas-Forth Worth Area. Once we have collected all the data and shaped it into a useable dataset, we can begin to dig into the question on whether Denton, TX has cross a population threshold that would make a new Grocery Store a profitable investment.

Data

The Business Problem requires data on cities, their population, and what grocery stores are already available. To build a dataset that can provide the necessary information will require interacting with multiple locations.

The data will be source from the following APIs:

US Census

Googlemaps Geocoder

Foursquare Venue Search API

1) US Census Data

We need to find all of cities located in the 3 county area we are reviewing: Dallas, Denton, & Tarrant Counties. The US Census has quite a few different APIs, so we need to make sure and use a very specific API in order to have the most up-to-date information: 2017 Population Estimation API.

2) GeoCode the Cities

Googlemaps has provided a nice Python package that allows for geocoding cities into their center point Longitude/Latitude.

3) Foursquare

Now that we have City names and their Lon/Lat, we can take these information to the Foursquare API and begin querying for Grocery Stores relative to each city. In order to minimize overlap, we will be using a radius of "3218" which is 2 miles.

We are going to use a very specific "categoryID" to ensure the results are relevant to our end goal.

Methodology

We have to pull the down by county, which requires creating 3 different variables that eventually are concatenated into a single dataframe:

In [2]:
```
df_denton = pd.read_csv("https://api.census.gov/data/2017/pep/population?get=POP,GEONAME&for=place:*&in=state:48%20county:121
")
df_denton.rename(columns={'[["POP"':'Population', 'GEONAME':'City', 'place]':'Place'}, inplace=True)
df_denton['Population'] = df_denton['Population'].str[2:-1]
df_denton['City'] = df_denton['City'].str[:-22]
df_denton['Place'] = df_denton['Place'].str[:-1]
df_denton['City'] = df_denton['City'].str.strip('(pt.)')
df_denton['City'] = df_denton['City'].str.replace(' city', "")
df_denton['City'] = df_denton['City'].str.replace(' town', "")
df_denton.drop(columns=['state', 'Unnamed: 5', 'Place'], inplace=True)
df_denton = df_denton[:-1]
df_denton.head()
```

Out[2]:

	Population	City	county
0	4100	Argyle	121
1	3391	Aubrey	121
2	1732	Bartonville	121
3	79715	Carrollton	121
4	0	Celina	121

In [11]:
```
df_w = pd.concat([df_denton, df_worth, df_dal], ignore_index=True)
df_w.head()
```

Out[11]:

	Population	City	county	Latitude	Longitude
0	4100	Argyle	121	33.121232	-97.183347
1	3391	Aubrey	121	33.304283	-96.986118
2	1732	Bartonville	121	33.073177	-97.131679
3	79715	Carrollton	121	32.975642	-96.889964
4	0	Celina	121	33.366454	-96.764097

Feature Extraction:

We create 2 new columns that will be useful for this analysis: 'Store Count' (Number of Stores per City) and 'Popdense' (Population divided by Store County).

```
dfw_data = df_w
store_density = pd.DataFrame(dfw_groceryvenues['City'].value_counts().reset_index().rename(columns={'index': 'city', 'City':
'count'}))
store_density = store_density.sort_values(by='city', ascending=True)
store_density['Zip Code'] = dfw_groceryvenues['Zip Code']
dfw_data = dfw_data.join(store_density.set_index('city'), on='City')
dfw_data = dfw_data[['Population', 'City', 'Latitude', 'Longitude', 'count', 'county', 'Zip Code']]
dfw_data = dfw_data.rename(columns={'count' : 'Store Count'})
dfw_data = dfw_data.dropna()
dfw_data['Population'] = pd.to_numeric(dfw_data['Population'], errors='coerce')
dfw_data['Zip Code'] = pd.to_numeric(dfw_data['Zip Code'], errors='coerce')
dfw_data['Popdense'] = dfw_data['Population']/dfw_data['Store Count']
dfw_data = dfw_data.dropna()
dfw_data.head()
```

	Population	City	Latitude	Longitude	Store Count	county	Zip Code	Popdense
0	4100	Argyle	33.121232	-97.183347	2.0	121	76205.0	2050.000000
1	3391	Aubrey	33.304283	-96.986118	2.0	121	76205.0	1695.500000
2	1732	Bartonville	33.073177	-97.131679	5.0	121	76205.0	346.400000
3	79715	Carrollton	32.975642	-96.889964	60.0	121	76227.0	1328.583333
5	822	Coppell	32.954569	-97.015008	20.0	121	75019.0	41.100000

Results

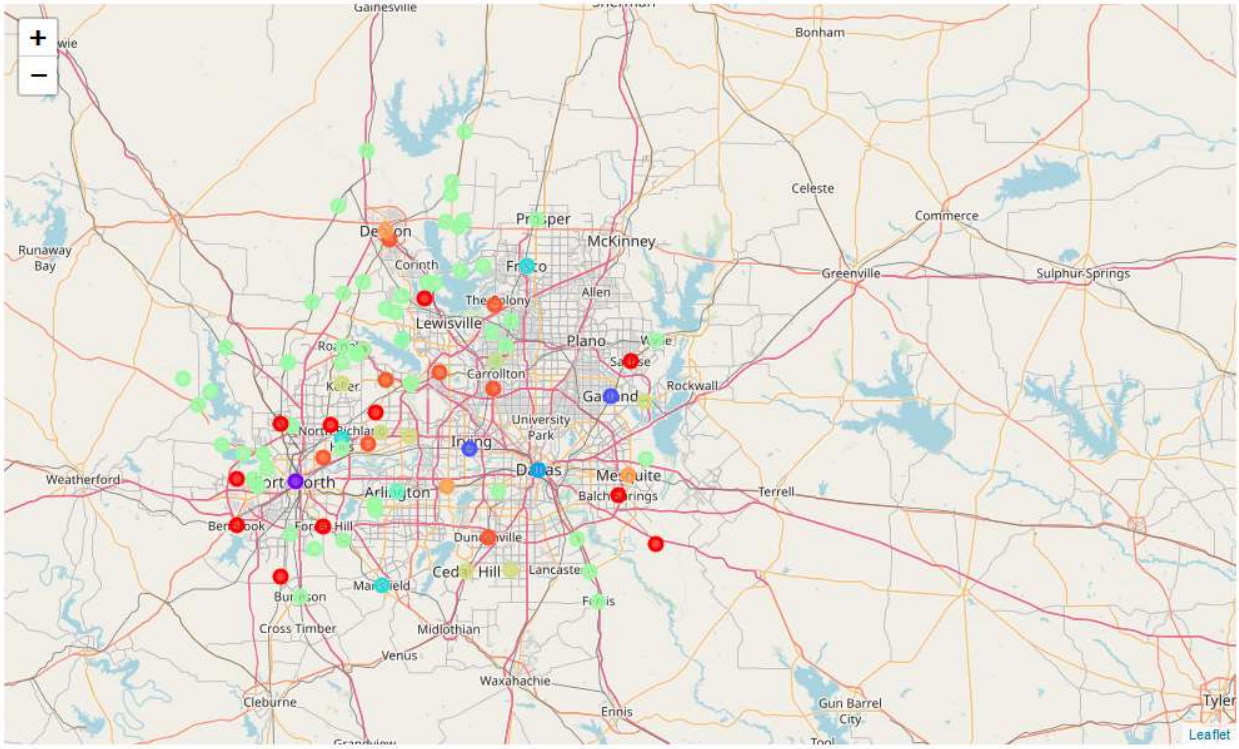
When looking at any data point, it is impossible to accurately draw a reasonable conclusion. As such, the goal from using clustering is for us to see clusters of similar data points. When we review the different clusters that were generated based off our dataset, you can begin to see exactly why these clusters were chosen.

The largest cluster in terms of cities was the very first cluster. Each city listed has very only 1 store with a tiny population, has a very small population , or has a low popdense. This cluster definitely highlights that there is no need to build a store in those cities.

The other clusters highlighted to varying degrees either that there was not enough population or popdense. There were a few clusters for outliers such as Dallas which has the largest population, most stores, and highest popdense.

There was a cluster that did stand out, as it highlighted 3 cities that had a large population (>100,000), high popdense (>6000), and moderate amount of stores available (20-40).

When mapped, the orange dots belong to the unique cluster.



Discussion and Conclusion

When we review all of the clusters there are a handful of observations worth sharing:

Popsense<1500 – No need for a new store

Popsense>13000 – Data is missing or incomplete

1501<Popsense<4000 – A new store might be acceptable but may require population>45000.

Denton was a member of the one cluster that stood out (pop >100,00, popsense > 6000,

20>store count>40). I strongly suspect that this cluster highlighted 3 cities that could very much benefit from a new grocery store. As such, I am confident that this analysis shows that the population of Denton, TX could be served well by another grocery store.