# The Battle of the Neighborhoods

## Coursera Capstone Project Report

### By Muhammad Musaddiq Sajjad

# Introduction

### Background

I live in the charming city of Al-Khobar, situated near the sea in the Eastern Province in the beautiful country of Saudi Arabia. It is one of the twin cities of Al-Khobar and Dammam, that reside side-by-side and have a lot in common.

Both cities are hustling and bustling with life, being the centers of the major activity in the Eastern Region. Everything from industrial zones, shopping malls, resorts, luxury apartments, lakes to restaurant chains and traditional markets can be found here.

Needless to say, both the cities, while being quite close and similar, are still a substantially varied mix of businesses and venues.

### Problem

For this report, I will be focusing on developing a system to recommend locations for businesses of different categories to open in one of these two cities. Any business that is entering the market can make use of this data to determine likely opening locations.

We aim to provide an idea of which district, location and cluster (from K-means clustering) the business should consider opening in.

### Interest

This report will be of interest to any business owner looking to open a new business in the Al-Khobar and Dammam area, as it will deliver, given an input business category, locations where the business is likely to succeed.

It can also be of use for anyone looking for recommendations of businesses, such as where to find a good Italian Restaurant.

# Data

### Data Sources

The venue data will be sourced from the Foursquare API, which will return location data and nearby venues, along with the venue category. This will help us in categorizing the different kinds of businesses to build our models.

Furthermore, we will be using publicly available district boundaries data and parsing that into a GeoJSON to establish our district boundaries.

Source: https://github.com/homaily/Saudi-Arabia-Regions-Cities-and-Districts

### Data Cleaning

Since the Foursquare API returns data that is already formatted, there is no data cleaning involved in that part.

However, the district boundary data sourced from the above linked Github repository does need to be prepared.

Specifically, it is JSON data, in the below format:

{'district\_id': 10100003002,

'city\_id': 3,

'region\_id': 1,

'name\_ar': 'حي النموذجية',

'name\_en': 'Al Namudhajiyah Dist.',

'boundaries': [[[24.65018372, 46.70227584],

[24.64939455, 46.7014039],

[24.64915715, 46.70115918],

[24.64892224, 46.70091159],

[24.64868987, 46.70066116],

[24.64857349, 46.70053129],

[24.64846099, 46.70039739],

[24.64835249, 46.70025959],

[24.6482481, 46.70011803],……

}

Note that it is using latitude,longitude coordinates system, so it will have to be reversed to be used as a GeoJSON. More information here, see "Position" section: <https://macwright.com/2015/03/23/geojson-second-bite.html>

This will have to be parsed into the standard GeoJSON format as below:

{

"type": "FeatureCollection",

"features":

[{

"type": "Feature",

"geometry":

{"type": "Polygon",

"coordinates": [[[50.13177952, 26.42026955], [50.13205218, 26.42399558]

…

}

}

}

# Methodology

### Exploratory Data Analysis

The cities of Al Khobar and Dammam constitute 44 and 83 districts respectively, so it is apparent that Dammam is the larger of the two cities by a factor of almost two.

In Al Khobar, the Foursqaure API returned 818 points of interest, and the most common venues were Coffee Shops, Cafes, and Food Trucks, with 107, 45 and 34 venues each.

In Dammam, the Foursqaure API returned 1093 points of interest, and the most common venues were Coffee Shops, Cafes, and Middle Eastern Restaurants, with 69, 53 and 50 venues each.

There were 160 unique venue categories in Al Khobar, whereas Dammam had 172 unique venue categories.

Combined, the two cities represented 221 unique venue categories.

From the above, it is clear that both of the cities have varied, vibrant venues, share a lot in common, and like to eat and dine out.

### K-means Clustering

After gathering the venues data from Foursquare API, one-hot encoding was performed on the venues. Then, these were arranged into the top 10 most common venues for each district.

After that, a K-means Clustering was performed on the Districts to categorize them into clusters. The chosen number of clusters was 5, as clustering with a higher or lower number of clusters resulted in ineffective clustering.

# Results

### Clustering Results – Al Khobar

Cluster 0 – was mostly located nears the outskirts of the city, featuring a multitude of Hookah Bars, Wellness Centers, Gyms, Farms and Soccer Fields. It is safe to assume this cluster represents open areas with many leisure and healthy activity spots. As a resident of this city, I can say this is an accurate cluster, as these venues tend to be away from the hustle and bustle of the city and are usually found in relaxed, quiet areas on the outskirts of the city.

Cluster 1 – represents the biggest cluster, featuring many different venue categories that at first glance makes it seems like there isn’t a correlation. However, it is likely that this cluster represents the central city locations, where market activity and footfall tends to be both frequent and varied. A common term would be “downtown”.

Cluster 2 – has just one district, which features a resort and a yoga studio.

Cluster 3 – has two districts only, and both have Farms and Waterfronts.

Cluster 4 – My home, and the only cluster that has a Lake in it.

### Clustering Results – Dammam

Cluster 0 – is by far the most populated cluster, and includes everything from Auto Workshops, Furniture Stores and Restaurants to IT Service Providers and Electronics Stores

Cluster 1 – has a lot of Campgrounds, Trails and Lounges so it is a cluster for outdoor activity venues.

Clusters 2, 3 and 4 – have only one district, and without many distinctive features, which may point to the clustering model being inaccurate.

### Recommending Districts

A district will be recommended if similar venues are numerous and thriving in the same district. Since Foursquare by default returns only popular venues, we can safely assume that the list of venues represents the popular venues for that area.

Hence, we can simply sort the districts by the highest number of venues of the same category and present the results as a top 10 list.

We can then output a map showing the recommended districts and their boundaries.

### Recommending Locations

Obviously, if there are other venues in the same area, there is a higher competition. But sometimes we want to open near to other similar businesses to benefit from the footfall. So to recommend possible locations, we will do two things:

1) Show the user a heatmap of similar businesses.

2) Future Improvement: Take an input Coordinate, a Competition Importance Factor (CIF) and a Footfall Importance Factor (FIF), and return a Recommendation Factor (RF).

### Recommending Clusters for Businesses

Finally, we'd like to get a recommendation for the cluster where the business should open.

Since there isn't enough data to build the correlation between clusters and business success, we'll recommend the venue if it is in the top 10 venues for that cluster. We'll also assign a recommendation point, which will be the frequency of the venue occurring in the top 10 venues for every district.

# Discussion

### Problems with Data

The first major issue with the data collection turned out to be the quality of data available for this region of the world. A major challenge was getting the district boundaries, but with some effort it was resolved.

However, one problem that couldn’t be resolved was getting district population statistics, or a way to measure footfall of each venue. For example, a shawerma shop may see 100x the daily customers than a resort, but this isn’t apparent by any data given by Foursquare.

That leads to some corner cutting especially when trying to recommend the locations.

### Problems with Clustering

It was apparent looking at the sizes of each cluster and the distribution of venues in the cluster members that something was amiss. For example, many clusters have just one member. It doesn’t mean that the number of members should necessarily be evenly spread, it just rings a bell when there are multiple clusters with just one member.

Possible reasons may be problems in the algorithm implementation (I couldn’t find any), using the wrong number of clusters, or perhaps the data not having a large enough spread to identify different clusters effectively.

### Problems with Recommendation

When building a process to recommend an opening location for a business, a few factors have to be considered. Ideally, one would like their business to be located close enough to where similar businesses are thriving, but also not have too much competition. For some businesses, for example an ATM machine, it is best to spread out as much as possible as people don’t have a strong preference and will generally pick the spot closest to them.

### Recommending Districts for Businesses

A district will be recommended if similar venues are numerous and thriving in the same district. Since Foursquare by default returns only popular venues, we can safely assume that the list of venues represents the popular venues for that area.

Hence, we can simply sort the districts by the highest number of venues of the same category and present the results as a top 10 list.

We can then output a map showing the recommended districts and their boundaries.

### Recommending Locations for Businesses

Obviously, if there are other venues in the same area, there is a higher competition. But sometimes we want to open near to other similar businesses to benefit from the footfall. So to recommend possible locations, we will do two things:

1) Show the user a heatmap of similar businesses.

2) Future Improvement: Take an input Coordinate, a Competition Importance Factor (CIF) and a Footfall Importance Factor (FIF), and return a Recommendation Factor (RF).

# Conclusion