Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

18-02-2020

The more suitable place to launch a tuinsian restaurant.
Riyadh, Saudia Arbia.

Prepared by Hosni Mrizek

IBM Data Science Capstone Final Project

I. Introduction

As part of my final capstone project for IBM Data Science Professional Certificate in Coursera, I choose to explore neighborhoods for Riyadh city, where I'm living, using Foursquare location data. Riyadh is a metropolitan area with more than 6 million (macrotrends web site) people living there, the Tunisian people who immigrate to Saudi Arabia and specially to Riyadh are in exponential increase and the need to find and enjoy Tunisian cuisine is on the rise. Due to the limit number of Tunisian restaurant in Riyadh, the idea to find more suitable place to begin such project comes up.

II. Business Problem

The question to answer is **"Where is the more suitable place to launch such project?"** in Riyadh, Saudi Arabia. I will try to answer this question by using data science, machine learning method (clustering) and Foresquare Api to extract needed data.

III. Target Audience

For any entrepreneur or businessman who wants to find a location to launch such project.

IV. Data

The data needed for the project is:

- Riyad location (Latitude and Longitude)
- Neighborhood list in Riyadh city.
- Latitude and Longitude relative to these neighborhoods.
- Venue List of North African restaurant.

IV.1 Datasets

1- List of neighborhood

from https://en.wikipedia.org/wiki/Riyadh#City_districts

2- Foresquare Api to extract the venue list for each neighborhood

IV.2 Extraction the data

- Scrapping Riyadh Neighborhood via Wikipedia.
- Get latitude and longitude for each neighborhood extracted in previous section using geocoder.
- Extract venue list of each neighborhood using Foresquare Api.

IV.2.1 Scrapping Riyadh Neighborhood list

Get the list of district in Riyadh city from Wikipedia page and save it to pandas DataFrame

	District
0	Al-Deerah
1	Mikal
2	Manfuha
3	Manfuha Al-Jadidah
4	Al-Oud

IV.2.2. Get Riyadh Neighborhoods coordinate

We need to get the coordinate of Riyadh neighborhoods in order to get the list of venues

	District	latitude	longitude
0	Ad Difa	24.592784	46.833947
1	Al Iskan	21.400517	39.780900
2	Al Izdihar	24.780321	46.717530
3	Al Mansouriyah	24.625390	46.522381
4	Al-Arid	24.499165	47.000378

we see that we have 71 districts.

IV.1.6 Get venues list from Foresquare API

Now we will extract coordinate of districts extracted previously from Wikipedia page. Let's use Foursquare API to get info on restaurants in each neighborhood. We're interested in venues in restaurants category. So we will include in our list only venues that have 'restaurant' in category name.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Al Iskan	21.400517	39.7809	Dunkin' Donuts (دانكن دوناتس)	21.401011	39.780961	Donut Shop
1	Al Iskan	21.400517	39.7809	Baskin-Robbins	21.400857	39.781141	Ice Cream Shop
2	Al Iskan	21.400517	39.7809	مطعم الدوحان للكباب العيرو	21.401180	39.780820	Mediterranean Restaurant
3	Al Iskan	21.400517	39.7809	Boga Superfoods (بوقا سوبر فودس)	21.399679	39.785251	Sandwich Place
4	Al Iskan	21.400517	39.7809	Little Caesare (ليتل سيزرز)	21.400687	39.780630	Pizza Place

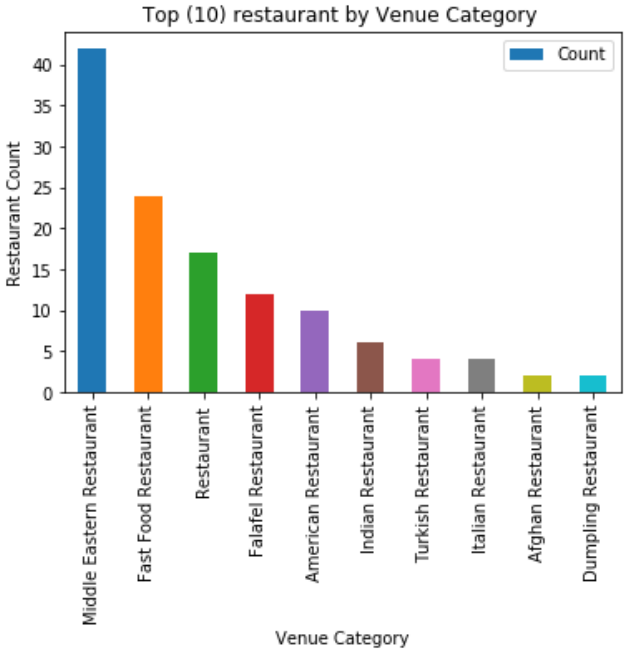
IV.1.6.1 Filter only restaurant category venues

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Al Iskan	21.400517	39.78090	مطعم الدوحان للكباب العيرو	21.401180	39.780820	Mediterranean Restaurant
1	Al Iskan	21.400517	39.78090	Herfy (هرفي)	21.400527	39.780757	Fast Food Restaurant
2	Al Iskan	21.400517	39.78090	شاورمتك	21.400806	39.781047	Arepa Restaurant
3	Al Izdihar	24.780321	46.71753	فطائر تركية	24.779479	46.719074	Turkish Restaurant
4	Al Izdihar	24.780321	46.71753	فلافل ابو عدنان	24.778129	46.720055	Falafel Restaurant

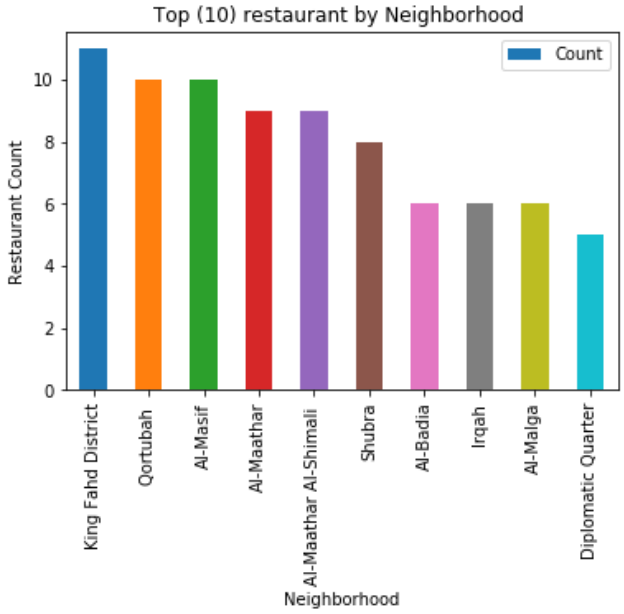
There are 26 unique categories.

IV.1.6.3 Exploring the Data

a. Plot the most frequent restaurant



b. Plot the most Neighborhood restaurant count



V. Methodology

Since there no Tunisian restaurant in Riyadh area, we will direct our efforts on detecting districts that have high restaurant density because, it will be more frequented.

In first step we have collected the number of restaurant by district.

Second step in our analysis will be calculation and exploration of '**restaurant number**' across different districts - we will use **bar chart** to identify a few promising areas of high number of restaurants and focus our attention on those areas.

In third and final step we will focus on most promising areas and within those create **clusters of locations**. We will present map of all such locations but also create clusters (using **k-means clustering**) of those locations to identify general zones / neighborhoods.

VI. Analysis

VI.1 Data preparation

First we encode the Venue Category in columns with 0 or 1 value.

	Neighborhoods	Afghan Restaurant	American Restaurant	Arepa Restaurant	Asian Restaurant	Australian Restaurant	Cantonese Restaurant	Chinese Restaurant	Dumpling Restaurant	Dutch Restaurant	...	Japanese Restaurant	Kebab Restaurant	Lebanese Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant
0	Al Iskan	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0
1	Al Iskan	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
2	Al Iskan	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0
3	Al Izdihar	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
4	Al Izdihar	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0

5 rows × 27 columns

Then we extract the top 10 common venue by district

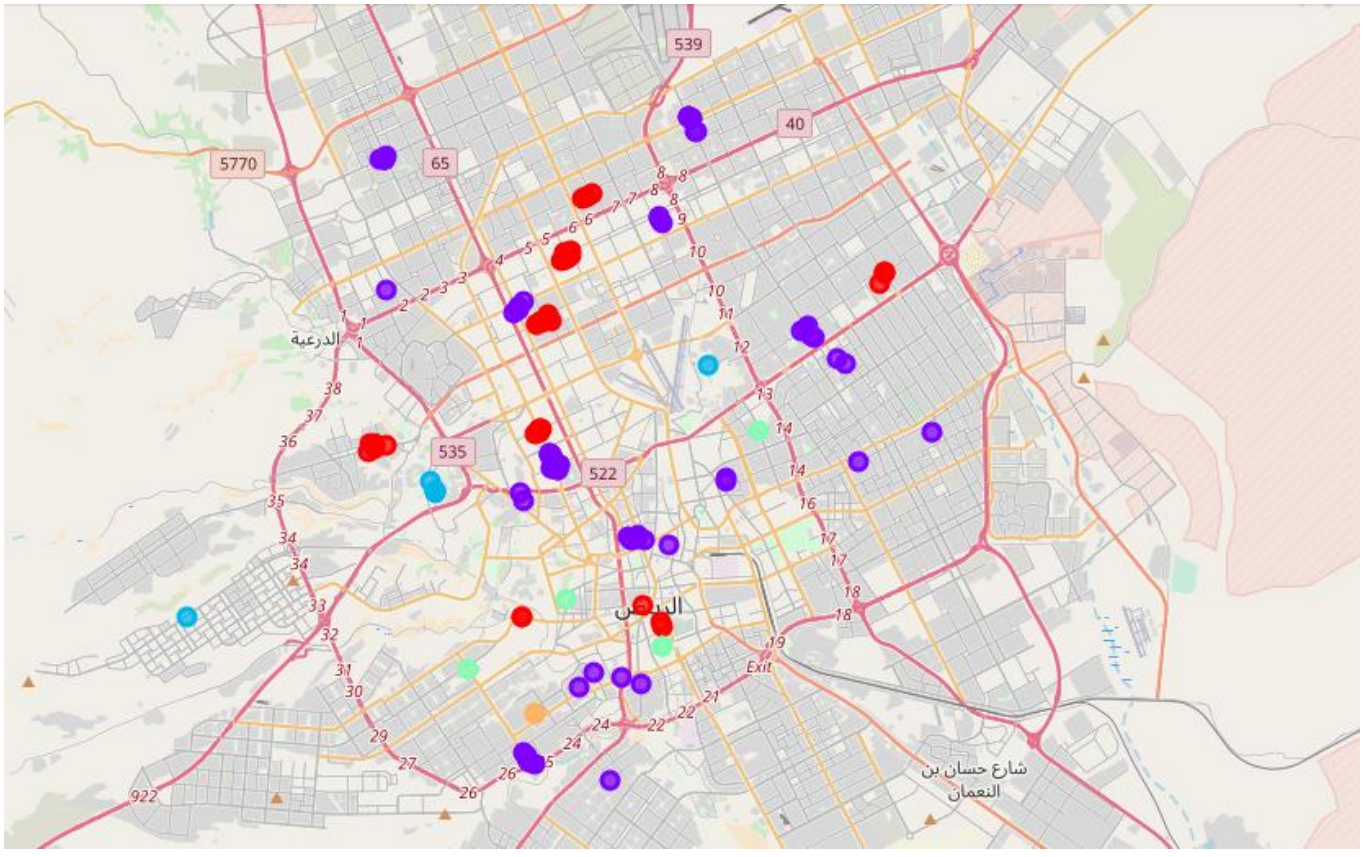
	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Al Iskan	Fast Food Restaurant	Arepa Restaurant	Mediterranean Restaurant	Falafel Restaurant	American Restaurant	Asian Restaurant	Australian Restaurant	Cantonese Restaurant	Chinese Restaurant	Dumpling Restaurant
1	Al Izdihar	Turkish Restaurant	Restaurant	Falafel Restaurant	Fast Food Restaurant	Peruvian Restaurant	Middle Eastern Restaurant	American Restaurant	Arepa Restaurant	Asian Restaurant	Australian Restaurant
2	Al Mansouriyah	Restaurant	Turkish Restaurant	Falafel Restaurant	American Restaurant	Arepa Restaurant	Asian Restaurant	Australian Restaurant	Cantonese Restaurant	Chinese Restaurant	Dumpling Restaurant
3	Al-Badia	Middle Eastern Restaurant	Restaurant	Chinese Restaurant	Turkish Restaurant	Falafel Restaurant	American Restaurant	Arepa Restaurant	Asian Restaurant	Australian Restaurant	Cantonese Restaurant
4	Al-Deerah	Middle Eastern Restaurant	Turkish Restaurant	Falafel Restaurant	American Restaurant	Arepa Restaurant	Asian Restaurant	Australian Restaurant	Cantonese Restaurant	Chinese Restaurant	Dumpling Restaurant

VI.2 Clustering the Districts

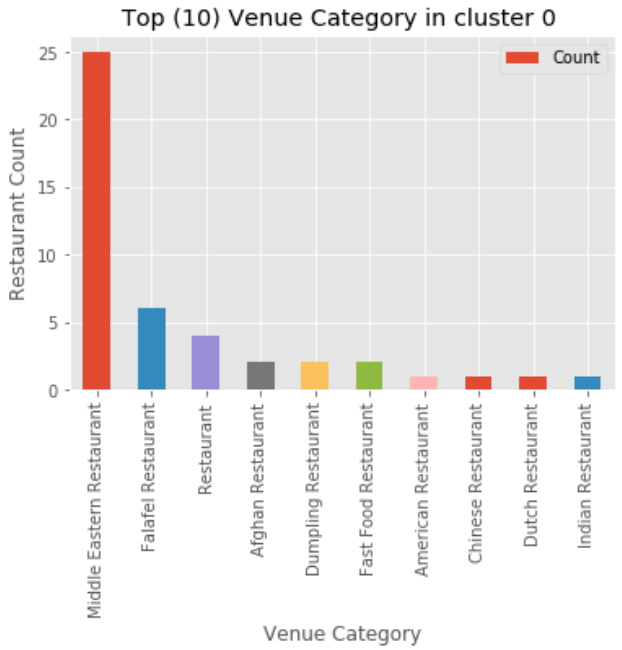
Next, we try to cluster districts based on the venue categories and use K-Means clustering. So our expectation would be based on the similarities of venue categories, these districts will be clustered.

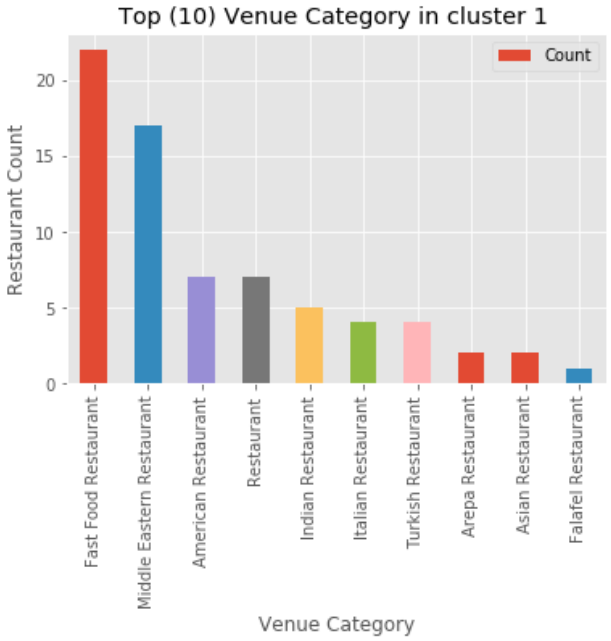
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	Al Iskan	21.400517	39.780900	مطعم الدومان للكباب العميرو	21.401180	39.780820	Mediterranean Restaurant	1	Fast Food Restaurant	Arepa Restaurant	Mediterranean Restaurant	Falafel Restaurant	American Restaurant	Asian Restaurant	Australian Restaurant	Cantonese Restaurant
1	Al Iskan	21.400517	39.780900	Herfy (هرفي)	21.400527	39.780757	Fast Food Restaurant	1	Fast Food Restaurant	Arepa Restaurant	Mediterranean Restaurant	Falafel Restaurant	American Restaurant	Asian Restaurant	Australian Restaurant	Cantonese Restaurant
2	Al Iskan	21.400517	39.780900	شاورمتك	21.400806	39.781047	Arepa Restaurant	1	Fast Food Restaurant	Arepa Restaurant	Mediterranean Restaurant	Falafel Restaurant	American Restaurant	Asian Restaurant	Australian Restaurant	Cantonese Restaurant
3	Al Izdihar	24.780321	46.717530	فطائر تركية	24.779479	46.719074	Turkish Restaurant	1	Turkish Restaurant	Restaurant	Falafel Restaurant	Fast Food Restaurant	Peruvian Restaurant	Middle Eastern Restaurant	American Restaurant	Arepa Restaurant
4	Al Izdihar	24.780321	46.717530	فلافل ابو عدنان	24.778129	46.720055	Falafel Restaurant	1	Turkish Restaurant	Restaurant	Falafel Restaurant	Fast Food Restaurant	Peruvian Restaurant	Middle Eastern Restaurant	American Restaurant	Arepa Restaurant
5	Al Izdihar	24.780321	46.717530	فروج الراية	24.779839	46.719020	Restaurant	1	Turkish Restaurant	Restaurant	Falafel Restaurant	Fast Food Restaurant	Peruvian Restaurant	Middle Eastern Restaurant	American Restaurant	Arepa Restaurant
6	Al Izdihar	24.780321	46.717530	تسالي صحوح	24.777475	46.720629	Fast Food Restaurant	1	Turkish Restaurant	Restaurant	Falafel Restaurant	Fast Food Restaurant	Peruvian Restaurant	Middle Eastern Restaurant	American Restaurant	Arepa Restaurant

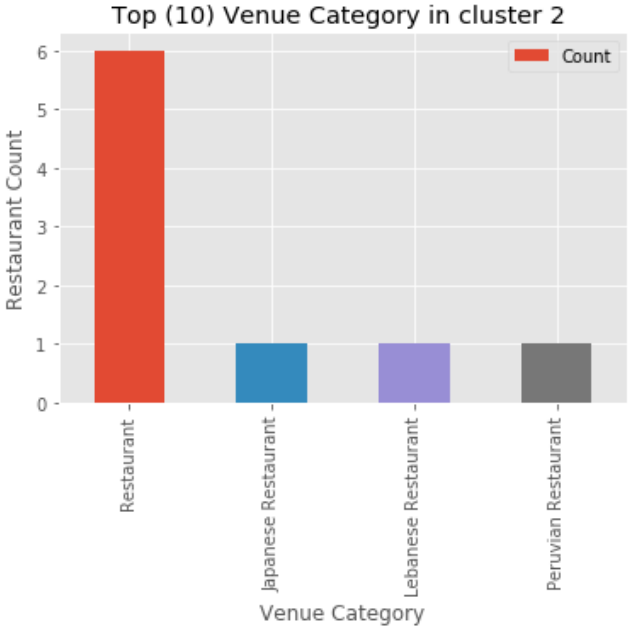
We can represent these clusters in a leaflet map using Folium library



We plot the top 10 venues count by venue category, we interest to the 3 first one.







VII. Results

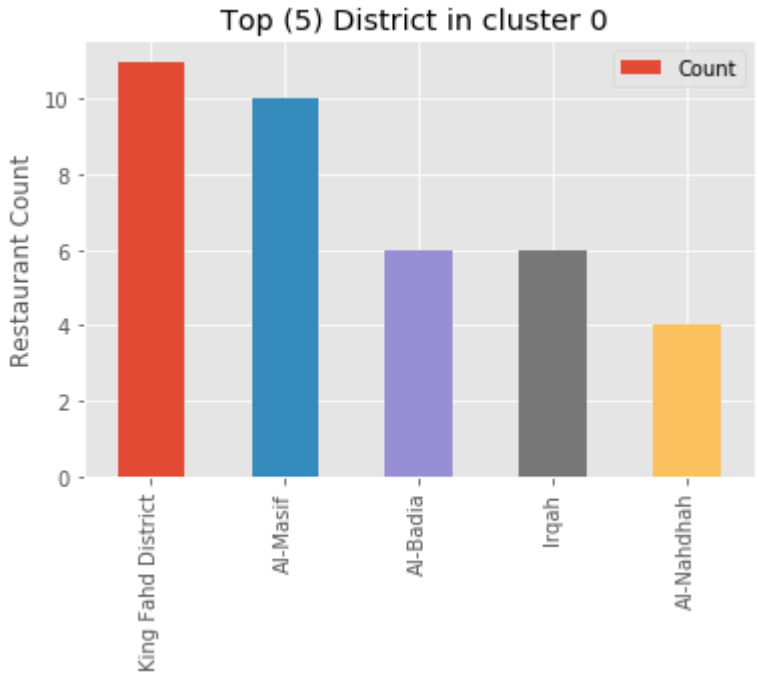
We decided earlier to choose the district we most frequent restaurants, we can after clustering that cluster 0 and cluster 1 have the biggest number of restaurant.

Cluster 0: "Middle Eastern Restaurant" is the most frequent restaurant.

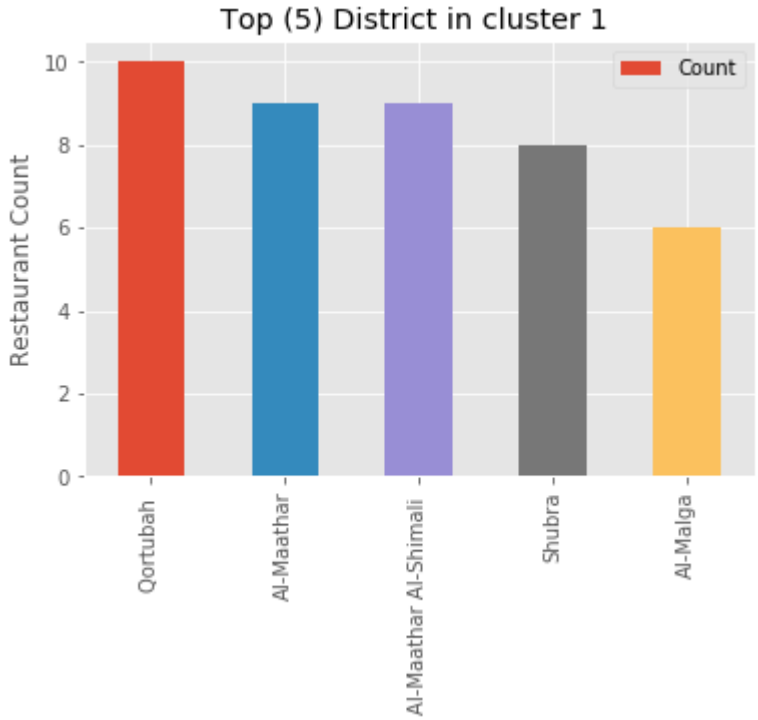
Cluster 1: "International Restaurant".

Let’s see in each cluster may District most suitable for each category of restaurant.

VII.1 Top 5 District in Cluster 0



VII.1 Top 5 District in Cluster 1



VIII. Discussion

According to this analysis, "King Fahd District" area will provide most suitable for an upcoming popular restaurant while "Qurtubah" could potentially be a target for starting quality restaurants.

The clustering is completely based on the most frequent number of restaurants venues obtained from Foursquare data. However, it definitely gives us some information on possibilities of opening restaurants around the districts of Riyadh.

IX. Conclusion

Because of lake of data, we only choose to analyze by the most frequent restaurant number. However, district distance from the Riyadh center, land price, district population density, may be very important information and gives us more homogeneous clusters, and more accurate estimation. Finally some of the drawbacks and chance for improvements to represent even more realistic pictures are mentioned. I hope it is help anyone who wants to get a preliminary idea about the best location to launch a Tunisian Restaurant.