IBM Data Science Professional Certificate

Applied Data Science Capstone

The Battle of Neighborhoods

Breakfast Restaurant in Dubai

Final Report

Anthony Lapierre

+639178244277 | +971529086930 | anthonylapierre@hotmail.com | www.linkedin.com/in/anthonylapierre/

EXECUTIVE SUMMARY

This document serves as a Final Report for the Applied Data Science Capstone of the IBM Data Science Professional Certificate[1].

The final project involves the use of the venue database of Foursquare to resolve a business problem related to the selection of an ideal location for a new venue.

For this exercise, a fictional Breakfast Restaurant Chain from the US is interested in opening a branch in Dubai.

They are employing Data Scientists to identify the best location using the Foursquare API.

PLANNING PROCESS

The purpose of this section is to cover the planning process for the study.

BUSINESS PROBLEM

We were approached by a popular Breakfast Restaurant Chain to identify the best neighborhood to launch their first restaurant in Dubai.

The company is new to Dubai and want to better understand the neighborhoods to support their decision-making process.

Based on their experience in establishing a new branch into a new market, they have observed that neighborhoods with a lot of breakfast options are good areas to use as a benchmark.

We propose to use the Foursquare API combined with Machine Learning algorithms to group the various neighborhoods into clusters based on the similarity of venues within each neighborhood. In other words, neighborhoods with similar venues will be grouped into clusters.

An important aspect of this study is that Dubai, being one of the most cosmopolitan city in the world, have a variety of neighborhoods that are “unofficially” clustered through its residents. For example, some neighborhoods are mainly populated by Emiratis and Middle Eastern Expats, while other neighborhoods are populated by Western, Indian, Pakistani or Filipino Expats. The residents will have an impact on preferences and business to be launched in the neighborhood. Therefore, the clustering process will reveal interesting insights.

Once the clustering process is completed, we will overlay the Top 100 Breakfast venues on top of the clusters. We will identify the areas of the Clusters with the most Breakfast spots and recommend neighborhoods with limited breakfast spots as an ideal location.

[1] Link: https://www.coursera.org/specializations/ibm-data-science-professional-certificate

METHODOLOGY

We will be using the "IBM Foundational Methodology for Data Science" throughout this case study.

This diagram provides an overview of our strategy for all steps with the exception of "Deployment" and "Feedback" which will not be covered in this document.

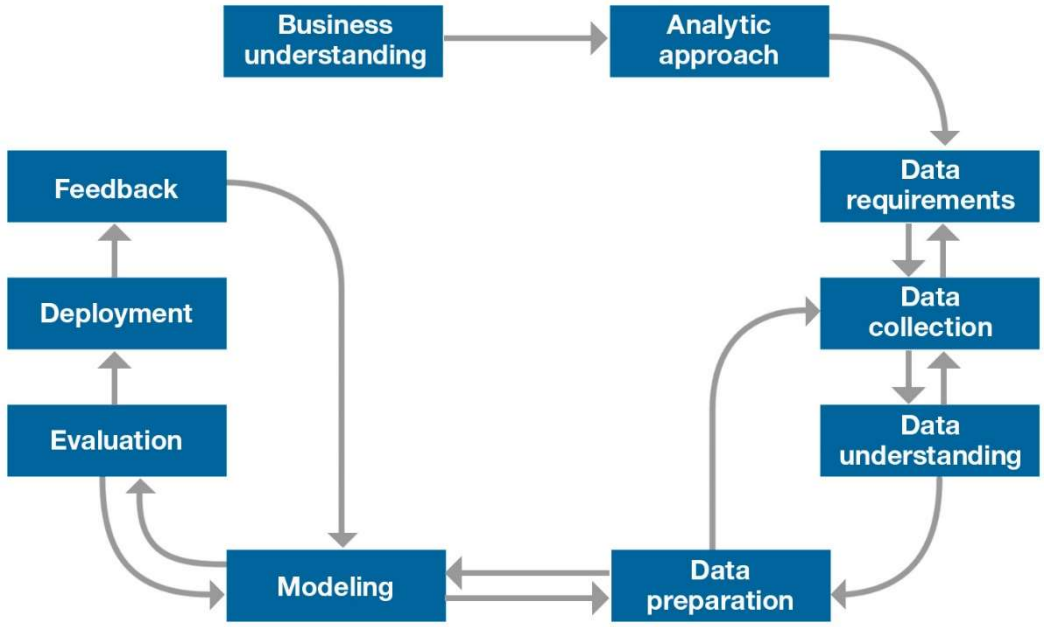


Figure 1 - IBM Foundational Methodology for Data Science[2]

BUSINESS UNDERSTANDING

Our Business Sponsor already started to study the Dubai Restaurant Market and provided us with the following context to develop our Business Understanding:

- It is a well-known fact that the choice of location for a Restaurant in Dubai is a key success factor: accessibility, parking space, visibility, high footfall, high urban commercial/resident population within a one-kilometer radius are all important factors.
- We will also assume that neighborhoods with a lot of breakfast options should be used as an example to evaluate our options for the new location.

ANALYTIC APPROACH

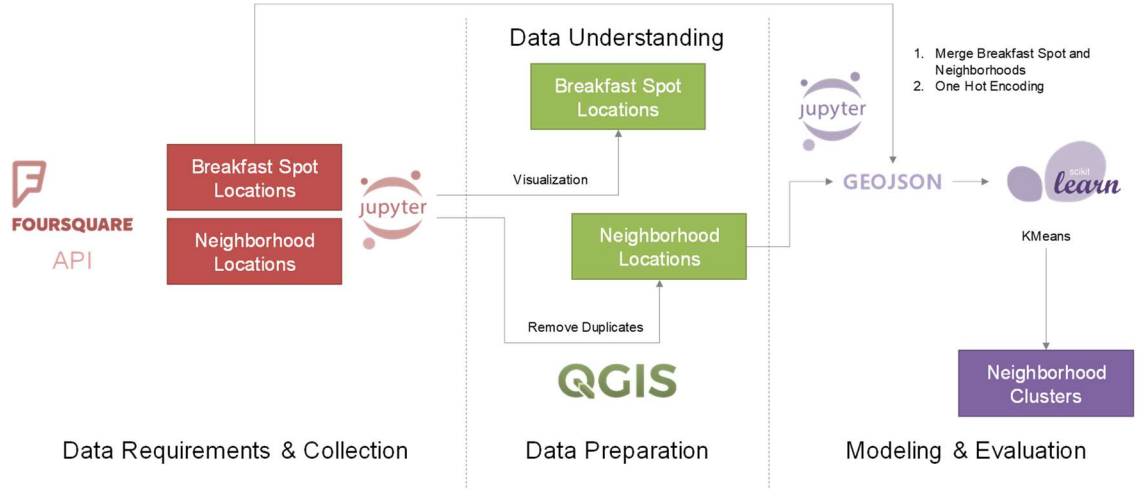
1. The analysis must be based on data from the Foursquare API Venue Database.
2. The business sponsor wants the restaurant to be profitable as soon as possible; we must avoid isolated areas or areas under construction, which will be removed from our list of neighborhoods.

[2] Link: https://www.slideshare.net/JohnBRollinsPhD/foundational-methodology-for-data-science/

FROM REQUIREMENTS TO EVALUATION

Below is the step-by-step process:

1. Export Neighborhoods and Breakfast Spots from Foursquare API using our Jupyter notebook.
2. There might be duplicates, inaccuracies or inconsistencies for neighborhood data. We will be using QGIS to make corrections to the data. We will also use QGIS to visualize the Breakfast Spots results and develop our understanding of the data. We will then export the neighborhood locations to GeoJSON format into our Jupyter notebook.
3. We will build our Data Model and use K Means to cluster the neighborhoods.

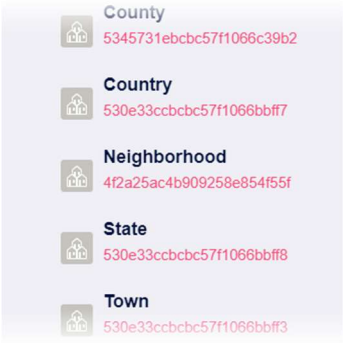


EXECUTION

The purpose of this section is to provide a detailed description of the execution of our study.

STEP 1 - DATA COLLECTION FROM FOURSQUARE

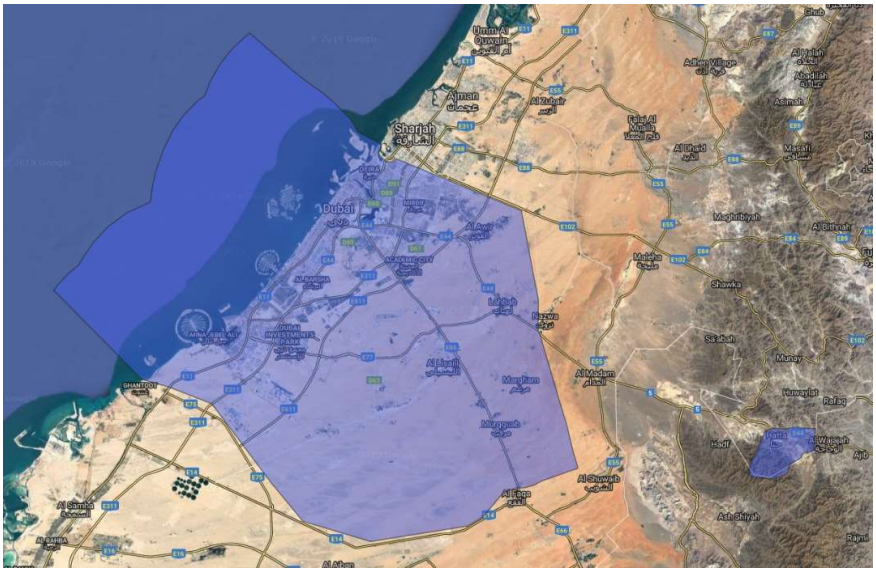
After studying the Foursquare API, we found out that there are neighborhoods in the database:



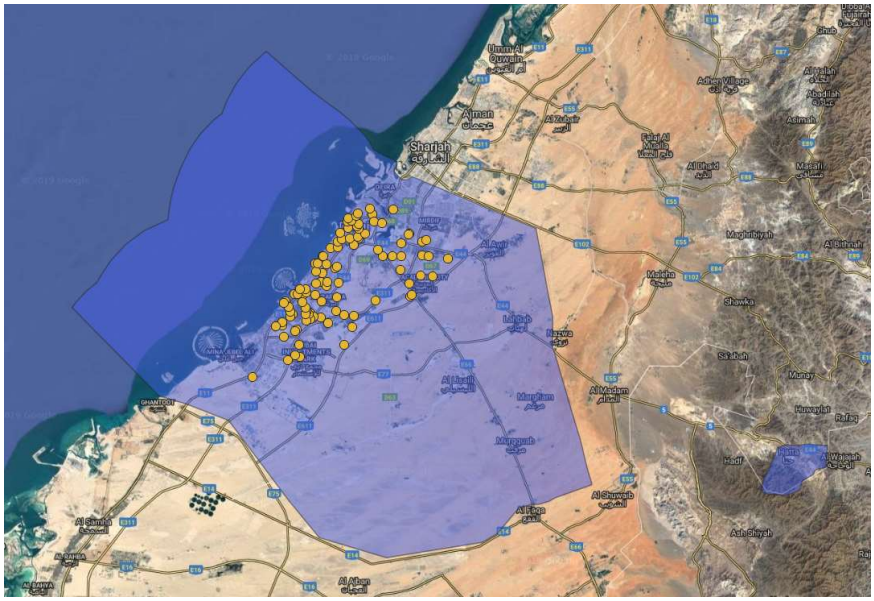
This is great news. As a result, we decided to extract those neighborhood locations from Foursquare and visualize them in QGIS. It is an opportunity for us to ensure that we focus our search on neighborhoods with enough information in Foursquare.

STEP 2 - DATA PREPARATION IN QGIS 3

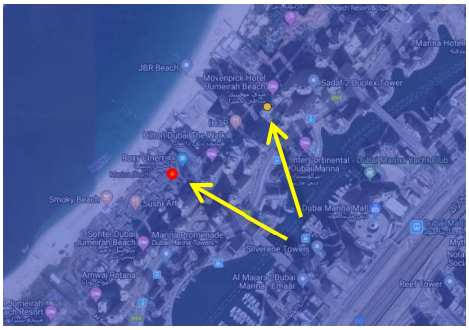
We will be using QGIS 3.6.0-Noosa as a platform to prepare our location data. Using the QuickOSM plug-in, we can extract the Dubai Municipality boundary vector from OpenStreetMap:



We will then import the CSV file of the neighborhood locations extracted from Foursquare. We can then have a better idea of the territory covered by Foursquare data. As a result, we will focus on those areas and ignore the desert areas. The Hatta exclave was not part of the radius that we established initially. Therefore, it will be ignored in this study.



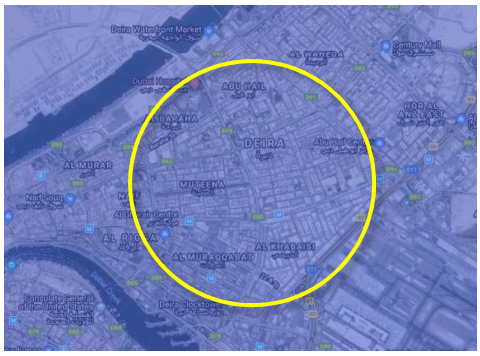
After analyzing the neighborhood locations, we observed a lot of duplication, inaccuracies and missing neighborhoods:



Duplication

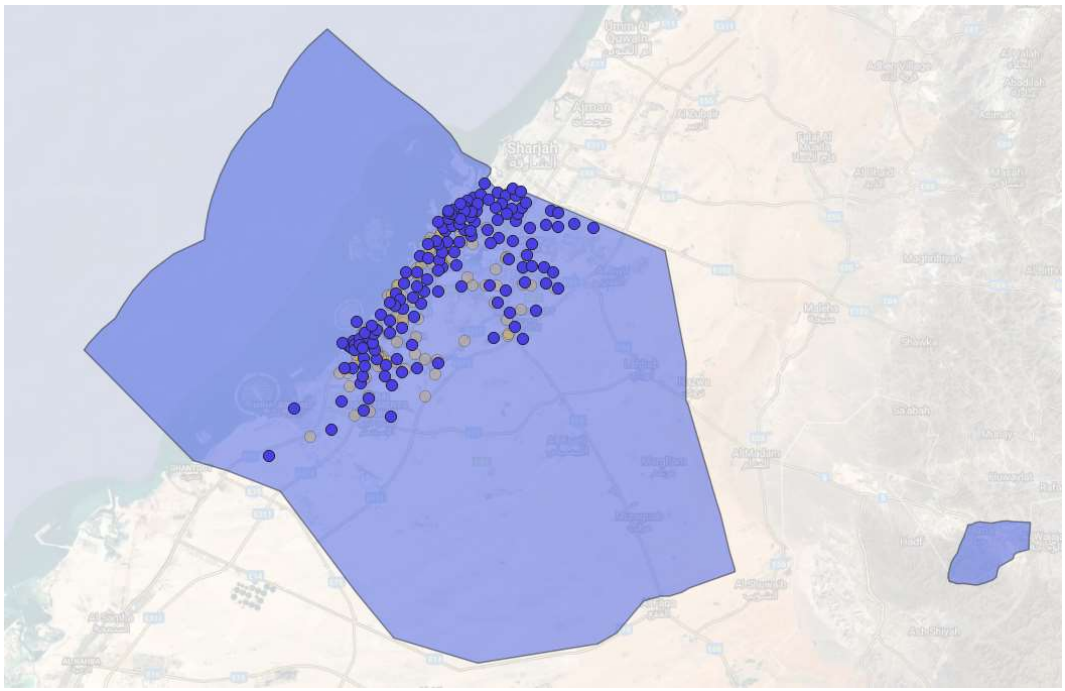
Neighborhoods not centered

Palm Jumeirah is missing



A major area of Dubai is missing

We ended up adding the neighborhoods manually (in blue), using Google Map as a background. We tried to stay close to the Foursquare neighborhoods (in yellow) and stay within the boundary of Dubai:

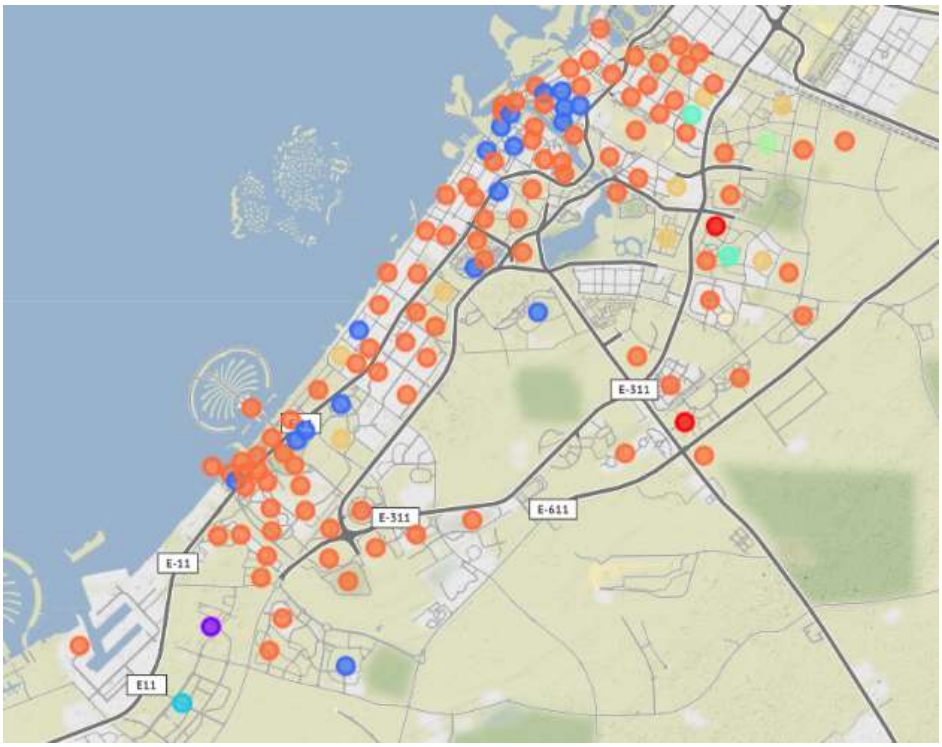


STEP 3 - CLUSTERING IN JUPYTER NOTEBOOK

As a first step, we created the neighborhood map using Folium:

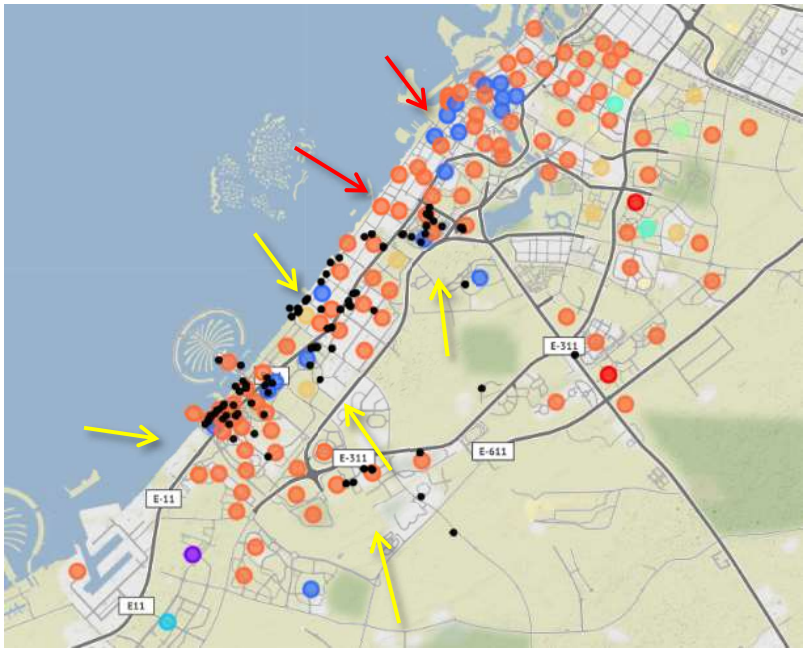
We then used the Foursquare API to get the nearby venues for each neighborhood, grouped them to get the counts and then performed one hot encoding on the results.

We then performed a K-Means Analysis using 8 clusters. After trying 4 to 7, we realized that 8 is an ideal number:



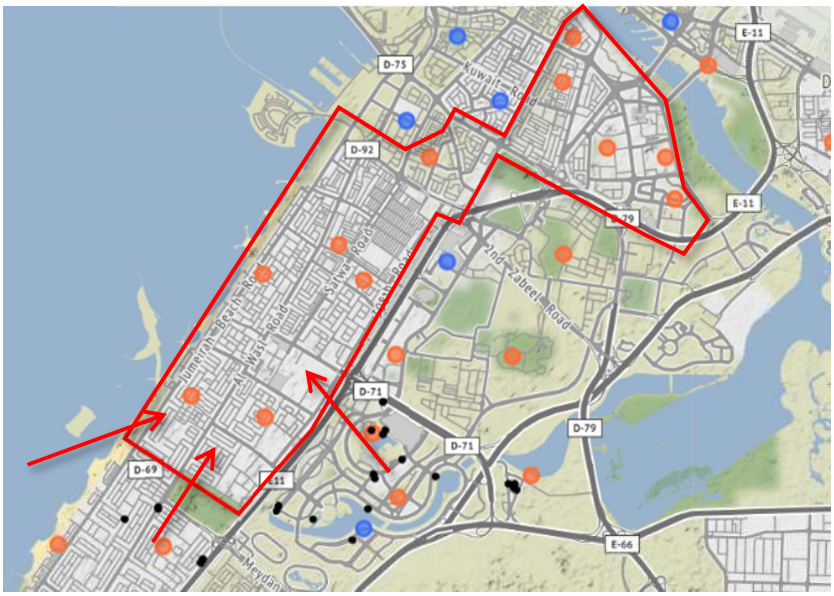
STEP 4 - EVALUATION

As we can see below, the Orange Cluster appears to be the “Typical Neighborhood” and it is also containing a great concentration of Breakfast Spots (See yellow arrows). We recommend focusing on the areas identified by the red arrows:



STEP 5 - RECOMMENDATION

Here is a zoomed-in view of the area. It is a residential area, close to other areas with great concentration of Breakfast Spots. The area contains the “Typical Neighborhood” cluster. We assume that the residents in these areas have similar tastes as residents in the typical cluster with breakfast options. Residents of neighborhood with breakfast options are likely to explore this new area, given that a new restaurant is launched (See red arrows):



FINAL THOUGHTS

This case study provides a short overview of some of the IBM Data Science Methodology and the tools that can be used for Location Analytics.

Here are a few important points to consider:

Since we are using a free version of the Foursquare API, we are limited to a maximum of 100 values when extracting the Breakfast Spots. Additionally, the free version does not allow us to get the ratings of a venue. With a premium account we could get the total number of venues, sorted by rating, which would be a way more accurate representation of successful breakfast restaurants.

We saw a great example of a technique for clustering the neighborhoods. However, the evaluation and recommendation requires the support of expert in the field and/or feedback from the business sponsor.