1 INTRODUCTION: BUSINESS PROBLEM

According to a statistical study by General Authority for Statistics in Saudi Arabia, about 38% of the population do not own a permanent residence (Aawsat, 2019). The decision of owning a home is a critical action to many families in Saudi Arabia in terms of effort, time, and most importantly money. One point that stands as a crucial element in this process is the location of your house. Such a decision is crucial to determine many of your future life activities including the quality of living in a specific destination within the city and proximity to various services. This project addresses this issue by performing an analysis of Khobar City, SA and clustering the promising neighborhoods according to the following factors:

1. Proximity to city center
2. Services and venues within the neighborhood including:
    a. Supermarkets
    b. Clothing stores
    c. Gyms
    d. Mosques
    e. Restaurants
    f. Coffee Shops
    g. Parks
    h. Schools
3. Distance from factories and industrial areas

Consequently, a Machine Learning (ML) clustering model is fitted to cluster the neighborhoods based on the venues nearby them. The segmentation of neighborhoods can cluster similar neighborhoods for an easy process of finding your future home's location. The ML does not rate the neighborhoods but rather it shows the similarity and dissimilarity between neighborhoods. For example, a neighborhood located within the edges of city with few services would be in a different cluster than a neighborhood close to the center with many services. Additionally, the model would place neighborhoods of dense commercial areas together which can help both customers and workers to choose a convenient location.

2 DATA UNDERSTANDING

2.1 Data Sources

The data required to cluster the neighborhoods are retrieved from different sources. Followingly, they are combined to carry on Exploratory Data Analysis (EDA), wrangling, and fitting the ML model. The data includes several sources as the following:

2.1.1 National Address API

National Address is a system that was issued by the Saudi Post Corporation as unified national post addressing system. It covers several levels varying from regions and cities to neighborhoods and buildings. On the other hand, National Address API is tool for providing agencies with data services. The API is used to address Khobar's neighborhoods' coordinates.

2.1.2 GeoJson.io

GeoJson.io is a website for creating and showing Geojson files either manually or by loading geo data files. The Khobar's Neighborhoods' coordinates and names are obtained through National Address API as mentioned earlier. Consequently, the Boundaries are obtained through Google Earth as a KML file which is converted to a Geojson file by GeoJson.io. By using Shapely library, the center of each polygon is identified and distances to several venues are calculated.

2.1.3 Foursquare API

Foursquare API is used to request results about venues within each neighborhood. The results include the Universal Transverse Mercator (UTM) coordinates along with the name and category of each venue. Based on a preliminary search, Foursquare showed that it is capable of giving a sufficiently accurate count of the venues within Khobar's neighborhood. Additionally, the venues are separated into different categories with a unique ID for each to group similar venues. Although the results of Foursquare depends largely on users' inputs, which could lead to inaccurate information, the large number of venues should overcome these discrepancies. The data are checked in terms of format, repetition, and misfit of categories through the EDA. All previous steps can give the safe assumption of data to be accurate for fitting the ML model.

2.2 Retrieving Data

2.2.1 Neighborhoods Data

Firstly, the neighborhoods' names and coordinates are retrieved from National Address API. The UTM coordinates are then converted into World Geodetic System (WGS 84). This step is important to find the Euclidean distance as UTM system relies on latitudes and longitudes, both in degress, rather than X and Y coordinates as in WGS 84. Table 1 shows the resulted data frame of 10 Khobar's neighborhoods.

Table 1 Sample of Khobar's neighborhoods data frame

	Neighborhood	Latitude	Longitude	X	Y	Distance
0	THUQBAH	26.274439	50.191364	419260.957823	2.906329e+06	1466.819086
1	SOUTH THUQBAH	26.273737	50.205405	420662.446460	2.906243e+06	106.211710
2	ISKAN	26.257636	50.208391	420949.699536	2.904457e+06	1709.884239
3	HAMRA	26.226879	50.204523	420542.507086	2.901053e+06	5100.947741
4	TAAWIN	26.224262	50.186858	418776.065391	2.900774e+06	5716.547726
5	KHUZAMA	26.208595	50.186968	418776.199502	2.899039e+06	7372.194020
6	DUGHEITHER VILLAGE	26.274842	50.213668	421488.268917	2.906360e+06	798.957141
7	SAHIL	26.255873	50.217583	421866.439607	2.904257e+06	2216.201050
8	NORTH KHOBAR	26.290838	50.213636	421495.859179	2.908132e+06	2127.867270
9	MADINAT UMAL	26.293403	50.203582	420493.793127	2.908422e+06	2281.309626

Later, Geojson.io was utilized to generate a geojson file that includes the boundaries of each neighborhood such that a visualization map could be shown to show analysis such as the a choropleth map of neighborhoods' distances from Khobar City center as shown in Figure 1.

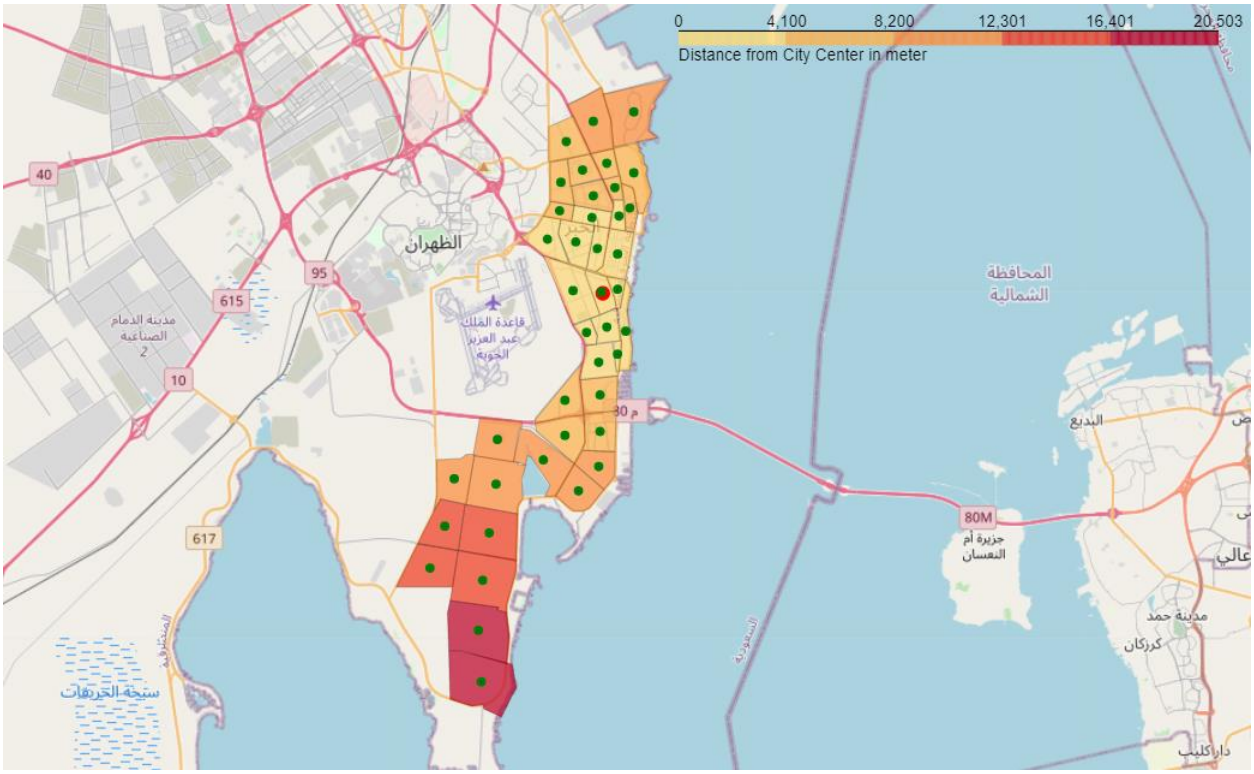

Figure 1 Choropleth map of neighborhoods' distances from Khobar City center

2.2.2 Foursquare Data

Based on the project objective, the data retrieved from Foursquare API must contain venues that are important or relatively important in the decision of the location. The selected venue categories along the category ID as per Foursquare API website are shown in Table 2. The table includes additionally factories which could influence the decision as these areas are usually avoidable and unsuitable due to their polluted environment and noises.

Table 2 The selected venue categories and their Foursquare API IDs

Category	ID	Category	ID
Supermarket	52f2ab2ebcbc57f1066b8b46	Elementary School	4f4533804b9074f6e4fb0105
Middle School	4f4533814b9074f6e4fb0106	High School	4bf58dd8d48988d13d941735
University	4bf58dd8d48988d1ae941735	Food	4d4b7105d754a06374d81259
Parks	4bf58dd8d48988d163941735	Mosque	4bf58dd8d48988d138941735
Gym & Fitness Center	4bf58dd8d48988d175941735	Art & Entertainment	4d4b7104d754a06370d81259
Beach	4bf58dd8d48988d1e2941735	Factory	4eb1bea83b7b6f98df247e06
Government Building	4bf58dd8d48988d126941735	Library	4bf58dd8d48988d12f941735
Medical Center	4bf58dd8d48988d104941735	Auto Workshop	56aa371be4b08b9a8d5734d3
Clothing Store	4bf58dd8d48988d103951735	Grocery Store	4bf58dd8d48988d118951735
Shopping Mall	4bf58dd8d48988d1fd941735	Shopping Plaza	5744ccdfe4b0c0459246b4dc

The algorithm to request calls about venues was to select a radius of 1,500m centered at each neighborhoods' center. Such method could sufficiently cover the area of each neighborhood with repeated venues from other nearby search areas being skipped. In each search area, Foursquare is requested to give 100 calls per each category. Figure 3 shows the coverage of each search area.

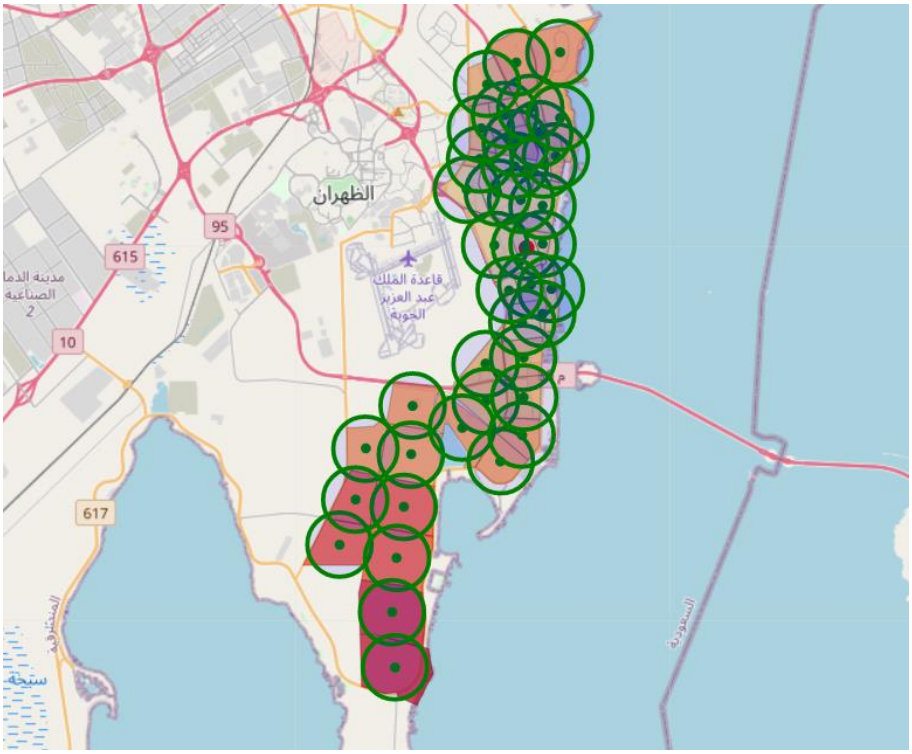

Figure 2 The search area coverage of each neighborhood

The figure shows almost a full coverage of Khobar's City along with some areas of Dhahran City, located directly west to Khobar's City. Although some of the resulted calls are outside of Khobar's city, they are, however, contributing services to the neighborhoods next to Dhahran City. The resulted calls are retrieved as json file which was filtered and read into a data frame as shown in Table 3. Since Foursquare is relatively unreliable in addressing the neighborhood of each venue, the Khobar's Geojson file was used to add an additional column about the neighborhood.

Table 3 Sample of filtered Foursquare API calls data frame

	ID	name	lats	lons	cat	sub	Neighborhood
0	577e08ba498ee027e6fb1727	تموينات القرني	26.273430	50.186268	Supermarket	Supermarket	THUQBAH
1	4f413350e4b0740e7b8c7f02	Farm Supermarket (أسواق المزرعة)	26.289101	50.192048	Supermarket	Supermarket	AQRABIYAH
2	571623c8498e0d3bcf4d6465	Meed (ميد)	26.264993	50.192548	Supermarket	Supermarket	THUQBAH
3	57706fb0cd100106d4872c5c	بقالة ماجد	26.264034	50.191147	Supermarket	Grocery Store	THUQBAH
4	5cc7ff1c65211f002c271991	اسواق الثلاجة العالمية	26.285603	50.193726	Supermarket	Supermarket	AQRABIYAH
...	...	...	...	...	...	...	...
2694	536b9b03498e1f325ba4336a	موحي	26.264544	50.191902	Food	Burrito Place	THUQBAH
2695	5cc8012416ef67002ca4cf44	مطعم رجل ألمع	26.267555	50.200341	Food	Asian Restaurant	THUQBAH
2696	52dae6a3498e8778e0e7a965	كافتيريا السندباد	26.267265	50.195027	Food	Burrito Place	THUQBAH
2697	5db153f0fe7ae00008189b6d	بوفية شواطئ الود	26.261724	50.190409	Food	Fast Food Restaurant	THUQBAH
2698	50a61760e4b0cd103aac5e2b	مطعم ومطبخ نور الحجاز	26.268066	50.197292	Food	Middle Eastern Restaurant	THUQBAH