Table of Contents

Introduction 2
A. Background 2
B. Business Problem 2
C. Data 2
Methodology 2
D. Data Collection 2
a. Scraping Wikipedia 2
b. Gathering Coordinates 3
c. Identifying Businesses in Each Area 4
E. Exploratory Data Analysis 5
a. Clustering Areas 5
b. Selecting Areas 6
Results 7
F. Recommendations 7
Conclusion 7
Bibliography 8

Introduction

A. Background

The Iraqi invasion of Kuwait in August 1990 brought tremendous upheaval to a state that had already experienced unprecedented change in the previous three decades because of oil. Kuwait, unlike most of its neighbors, has a well-established national identity and a long history as a nation, dating back to the eighteenth century. The nation has transformed from a pre-oil to an oil economy. Its social structure and composition, including the country's tribal roots and key divisions (involving class, gender, immigrant labor, political tensions resulting from the nation's sudden wealth, as well as its relations with other countries in the Gulf and the Middle East) has created a unique culture.

We will be exploring the similarities populations share within their respective areas and the overall population, and how they differ between themselves and the overall population. Then using this information to identify key markets for business growth. As well as identify which businesses would be most successful in those markets.

B. Business Problem

The following exploratory data analysis could be helpful in business. For example, say a client would like to open a new business, however they do not know what kind of business they would like to open, let alone where to open it. In this article, we will determine the optimal areas in Kuwait to open a business, as well as the perfect business for that area.

C. Data

In order to answer the business question, data will need to be gathered data from a few sources. The first source will be Wikipedia. From Wikipedia, we will identify the areas of Kuwait. Next, using Google Maps' API, we will collect the approximate coordinates of each area. With the coordinates and area names collected, we will next be using the Foursquare API to collect businesses information for each area, within a designated radius. The venue data collected from Foursquare will then be used to determine the top businesses in each area.

Once the business information is gathered, the next step will be to cluster areas in Kuwait based on business categories. This information will allow us to cluster the population in each area into discrete groups which will help us understand sentiment. And then moving on to identify areas within the clusters which are prime candidates for a new business, as well as identifying which specific businesses would be the most lucrative.

Methodology

D. Data Collection

a. Scraping Wikipedia

The information needed to drive the entire project begins with collecting area names. Wikipedia is a good source because they have an article dedicated to the area names in Kuwait, as well as in which district they reside. Some cleaning needs to be done to properly populate our data frame with the correct information. After compiling the data into a table, we get a list of area names in Kuwait (see figure 1).

	Area
0	Kuwait City
1	Dasmān
2	Sharq
3	Mirgāb
4	Jibla
5	Dasma
6	Da'iya
7	Sawābir
8	Salhiya
9	Bneid il-Gār

Figure 1

b. Gathering Coordinates

With the area names identified and loaded into our data frame, the next objective is to collect the coordinates for each area. For this, we will be utilizing the Google Maps API. As the area names are transliterated into English, errors may occur when searching for certain areas. Therefore, it is important to thoroughly scan the data received for any misidentified areas and correct the errors before moving on. The most common error, beside transliterated spelling of area names, is wrongly identified areas by the Google Maps API. For example, a search for Granada would most likely return the coordinates for Granada, Spain. That is obviously undesirable. This highlights the need for careful analysis of the returned data before moving forward. If all is done correctly, we get the desired data (see figure 2a). We will then visualize the data using Folium maps (see figure 2b)

	Area	Latitude	Longitude
0	Kuwait City	29.375859	47.977405
1	Dasmān	29.387804	47.999790
2	Sharq	29.382323	47.988644
3	Mirgāb	29.366138	47.983692
4	Jibla	29.369934	47.968836
5	Dasma	29.366434	48.000698
6	Da'iya	29.360440	48.018371
7	Sawābir	29.376369	47.982453
8	Salhiya	29.363509	47.967598
9	Bneid il-Gār	29.373051	48.004744

Figure 2a

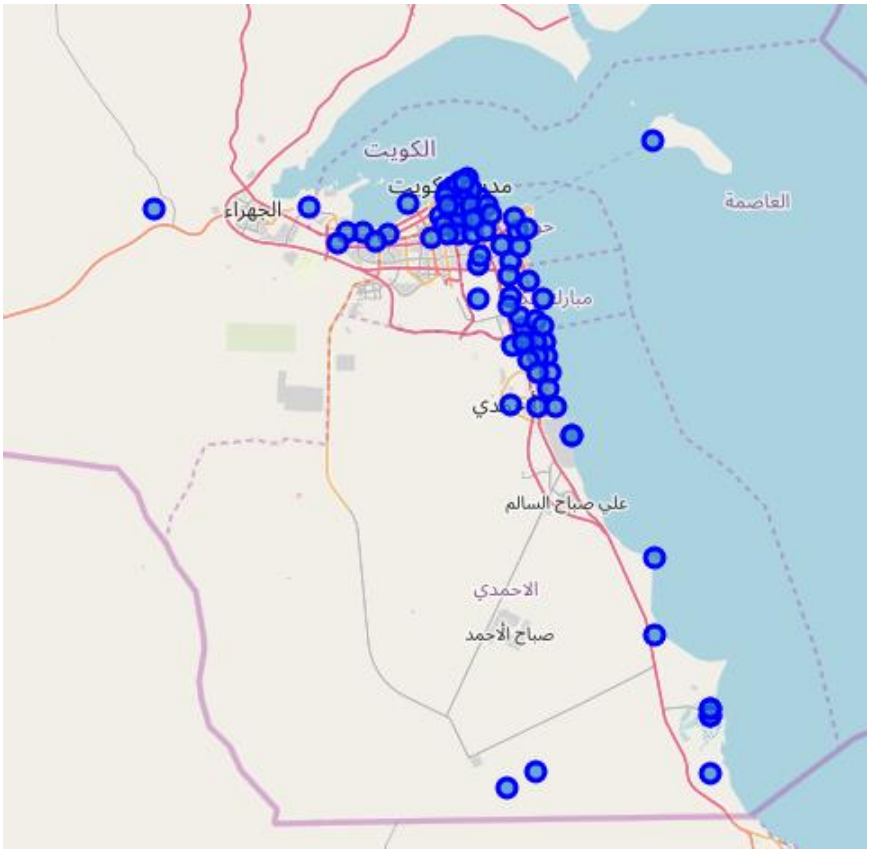

Figure 2b

c. Identifying Businesses in Each Area

Moving on, the next crucial step is to collect information about the available businesses in each area. To do this, we will utilize the Foursquare API. For anyone unfamiliar with Foursquare, it is an application that allows users to identify venues, check-in at places they are visiting, as well as rate the business and leave feedback. Using this crowdsourced data, we will populate our data frame with all businesses within a specified radius. For this study, a 700-meter radius was used to return enough results to perform our data analysis and clustering. Once the data is gathered, we populate the data frame (see figure 3a and 3b).

	Area	Area Lat	Area Long	Venue	Venue Lat	Venue Long	Category
0	Kuwait City	29.375859	47.977405	Concrete	29.375199	47.977508	Breakfast Spot
1	Kuwait City	29.375859	47.977405	Pick Yo (بك يو)	29.375684	47.977405	Frozen Yogurt Shop
2	Kuwait City	29.375859	47.977405	HABRA	29.376388	47.977048	Steakhouse
3	Kuwait City	29.375859	47.977405	Tobys Estate (توبيس إستيت)	29.375201	47.978188	Café
4	Kuwait City	29.375859	47.977405	CAF cafe	29.376773	47.978062	Café
5	Kuwait City	29.375859	47.977405	مطعم الف ليله وليله	29.376701	47.979291	Middle Eastern Restaurant
6	Kuwait City	29.375859	47.977405	Caffeine (كافين)	29.374840	47.976097	Coffee Shop
7	Kuwait City	29.375859	47.977405	% Arabica	29.377180	47.978590	Coffee Shop
8	Kuwait City	29.375859	47.977405	Little Ruby's (ليتل روبيز)	29.376590	47.977952	American Restaurant
9	Kuwait City	29.375859	47.977405	BON Cafe	29.375750	47.978382	Café

Figure 3a

	Area	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	Abdullah Port	Harbor / Marina	Dog Run	Donut Shop	Dry Cleaner	Egyptian Restaurant	Electronics Store	Event Service	Falafel Restaurant	Farm	Zoo Exhibit
1	Abdullah as-Salim	American Restaurant	Coffee Shop	Juice Bar	Cafeteria	Pizza Place	Grocery Store	Salon / Barbershop	Camera Store	Café	Dessert Shop
2	Abu 'Fteira	Plaza	Dessert Shop	Cafeteria	Restaurant	Fried Chicken Joint	Caribbean Restaurant	Supermarket	Middle Eastern Restaurant	Beach	Farmers Market
3	Abu Hassaniya	Café	Beach	Japanese Curry Restaurant	Middle Eastern Restaurant	Farm	Health Food Store	Hotel	Cupcake Shop	Coffee Shop	Boat or Ferry
4	Abu Hulaifa	Soccer Stadium	Dessert Shop	Soccer Field	Café	Arts & Crafts Store	Mexican Restaurant	Lounge	Burger Joint	Asian Restaurant	Grocery Store
5	Adiliya	Fast Food Restaurant	Pizza Place	Ice Cream Shop	Dessert Shop	Café	Burger Joint	Sandwich Place	Trail	Indian Restaurant	Middle Eastern Restaurant
6	Adān	Cosmetics Shop	Dive Spot	Bookstore	Food Truck	Shopping Mall	Bakery	Burger Joint	Burrito Place	Athletics & Sports	American Restaurant
7	Agricultural Wafra	Farm	Garden	Zoo Exhibit	Dry Cleaner	Egyptian Restaurant	Electronics Store	Event Service	Falafel Restaurant	Farmers Market	Doner Restaurant
8	Ahmadi	Garden	Fast Food Restaurant	Restaurant	Gym	Fried Chicken Joint	Soccer Field	Movie Theater	Middle Eastern Restaurant	Indian Restaurant	Arts & Crafts Store
9	Bayān	Fast Food Restaurant	Trail	Grocery Store	Coffee Shop	Intersection	Department Store	Garden	Cricket Ground	Supermarket	Ice Cream Shop

Figure 3b

E. Exploratory Data Analysis

a. Clustering Areas

With the data now gathered, we can begin clustering the data. To do this we will employ k-Means Clustering. k-Means Clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. (k-means clustering, n.d.) However, it should be noted that there exist drawbacks to using this approach for clustering. A key limitation of k-means is its cluster model.

The concept is based on spherical clusters that are separable so that the mean converges towards the cluster center. The clusters are expected to be of similar size, so that the assignment to the nearest cluster center is the correct assignment. Since data is split halfway between cluster means, this can lead to suboptimal splits. Therefore, to lower the risk of poor k-Means clustering, we will select the best k to fit our model. This is determined by trying all values of k within a predefined range (i.e. $k < 75$) that yields the highest Simplified Silhouette score. Simplified Silhouette is more suitable than the original Silhouette in the selection of the best result from k-means clustering with different k values. (Wang, Franco-Penya, Kelleher, Pugh, & Ross, 2017) By looping through various k's and calculating their simplified silhouette score, we find that the best k is $k = 8$, with a simplified silhouette score of 0.72053775. From here we set the k of our model to 8 and proceed to fit the model.

We assign the labels to each area and add the column labels to our data frame. To visualize the clustered areas, we generate a map of Kuwait with the areas color coded for each cluster (see figure 4). Using the labels, we now separate our data by cluster and determine the top three most popular businesses by cluster. This information is then used to explore intra-cluster differences.

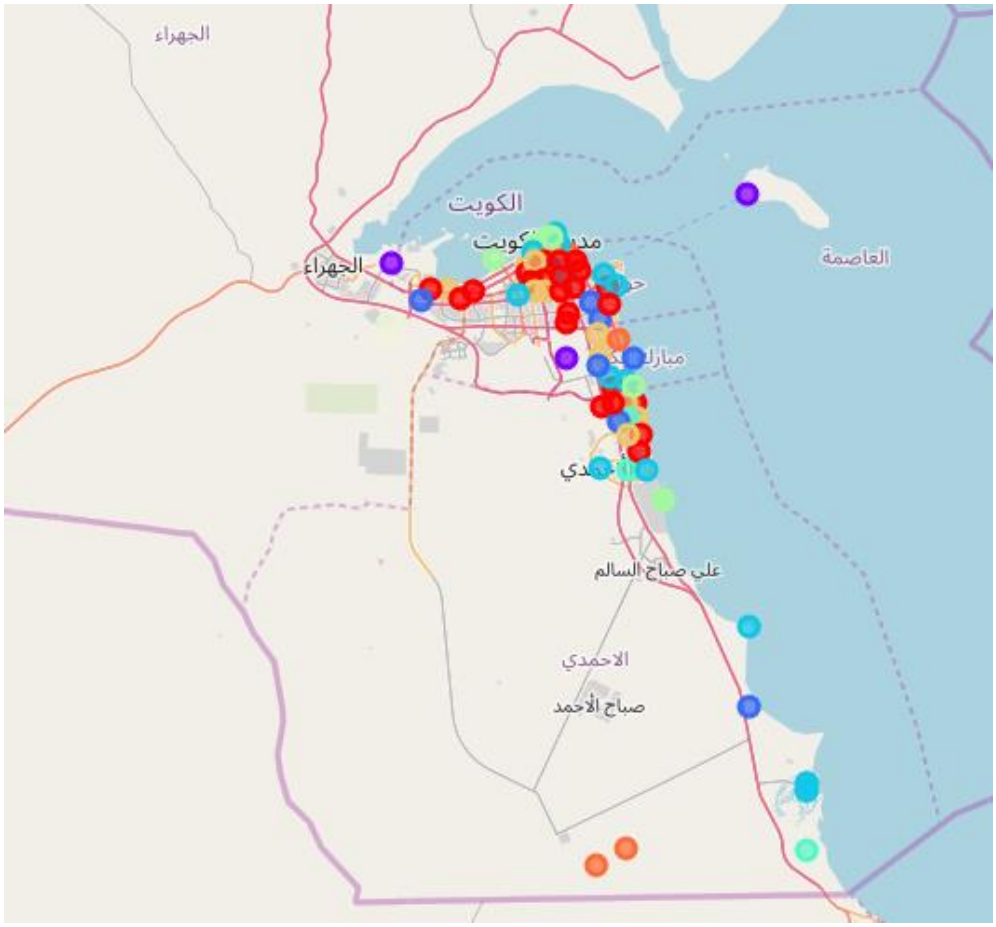

Figure 4

b. Selecting Areas

To do this we find the intra-cluster distance for each area and sort the areas in each cluster based on distance to the centroid. The areas with the highest distance are selected. These are the areas we explore for starting a business. The reason for this is that these areas have high inter-cluster distance, meaning that they are very different from the areas within other clusters. And high intra-cluster distance, meaning they are more like the areas within their own cluster. However, when compared to the intra-cluster areas, these areas are the weakest (furthest apart). So, we choose them to bring these samples closer to the centroid.

By comparing the businesses available in each of the selected areas with the top three businesses across the cluster we can determine what businesses to open and in which areas to open them.

Results

F. Recommendations

For Cluster 1, the most popular businesses among the cluster is coffee shop. Of the members of Cluster 0, Abdullah Al-Salem, Rumaithiya, and Surra areas are prime candidates for new coffee shop as they do not have any, or the available ones are not very popular. This suggests that if interested in opening a new coffee shop, we should focus on Cluster 0, as it is the cluster where coffee shops will thrive.

For Cluster 2, the most popular businesses happen to be outdoor in nature. This includes fishing piers, farms, and farmers markets. If we you are interested in opening any businesses that cater to outdoor activities, you should focus on Cluster 1. For this cluster, a good suggestion would be to start a Farmers Market. Areas in Cluster 1 include Jaber Al Ahmed. However, I suggest starting a Farmers Market in Subhan.

For Cluster 7, the most popular businesses are coffee shops, bakeries, and delis. For anyone interested in starting a deli, or bakery, I would strongly suggest areas in this cluster. The main area being 'Adan' as this population loves these kinds of businesses, and suffers from a lack of them. For a Deli or Bakery choose Adan.

Conclusion

In summation, by utilizing a k-Means clustering approach to classifying cities and towns we can determine areas that are prime candidates for new startups or business expansion, as well as weak areas. In addition, we can also identify which businesses are most likely to succeed in the selected areas through sentiment analysis of the top ten businesses in those areas. However, there are disadvantages to using this approach which largely arise from not enough data. Certain markets might not be right for certain businesses due to local factors that are not obvious when exploring top ten trends across segments. Therefore, it is important to conduct further analysis of each area to confirm the predictions of the model.

Bibliography

Frieze, A., Kannan, R., Vempala, S., Vinay, V., & Drineas, P. (2004). Clustering large graphs via the singular value decomposition. *Machine Learning, 56*(1–3), 9–33. Retrieved 5 1, 2019, from http://www.cc.gatech.edu/~vempala/papers/dfkvv.pdf

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient ''k''-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 881–892. Retrieved 5 1, 2019, from http://www.cs.umd.edu/~mount/Papers/pami02.pdf

k-means clustering. (n.d.). Retrieved 5 1, 2019, from Wikipedia: The Free Encyclopedia: http://en.wikipedia.org/wiki/K-means_clustering

MacKay, D. (2003). *Information Theory, Inference and Learning Algorithms.* Cambridge University Press. Retrieved 5 1, 2019, from http://www.inference.phy.cam.ac.uk/mackay/itila/book.html

Sarma, T. H., Viswanath, P., & Reddy, B. E. (2011). *A fast approximate kernel k-means clustering method for large data sets*. Retrieved 5 1, 2019, from http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.ieee-000006069372

Wang, F., Franco-Penya, H.-H., Kelleher, J. D., Pugh, J., & Ross, R. J. (2017). *An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity*. Retrieved 5 1, 2019, from https://link.springer.com/content/pdf/10.1007/978-3-319-62416-7_21.pdf