Coursera Capstone Project

IBM Applied Data Science Capstone

Opening a Debut Bakery for an International Chain in Karachi, Pakistan

By: Roha Farooq

May 2020

Table of Contents

Table of Figures

Introduction

Bakeries are becoming an essential part of a city's fabric. From daily requirements of bread and comfort foods to birthday cakes and sweets, it is a sound business investment especially in metropolitan cities. Everyone eats bread and thus we patronize select bakeries in our area. Consequently, loyal customers are developed who, by word of mouth, spread the news about our products. We will use this as basis of our new venture. The largest metropolitan city in Pakistan is Karachi (Wikipedia, n.d.). For this project we will be considering a hypothetical company by the name of "**Pan**". This company has a chain of bakeries in France and Spain and want to introduce themselves in a new market. They require data for towns in Karachi and already existing bakeries in those areas, so that they can strategize accordingly.

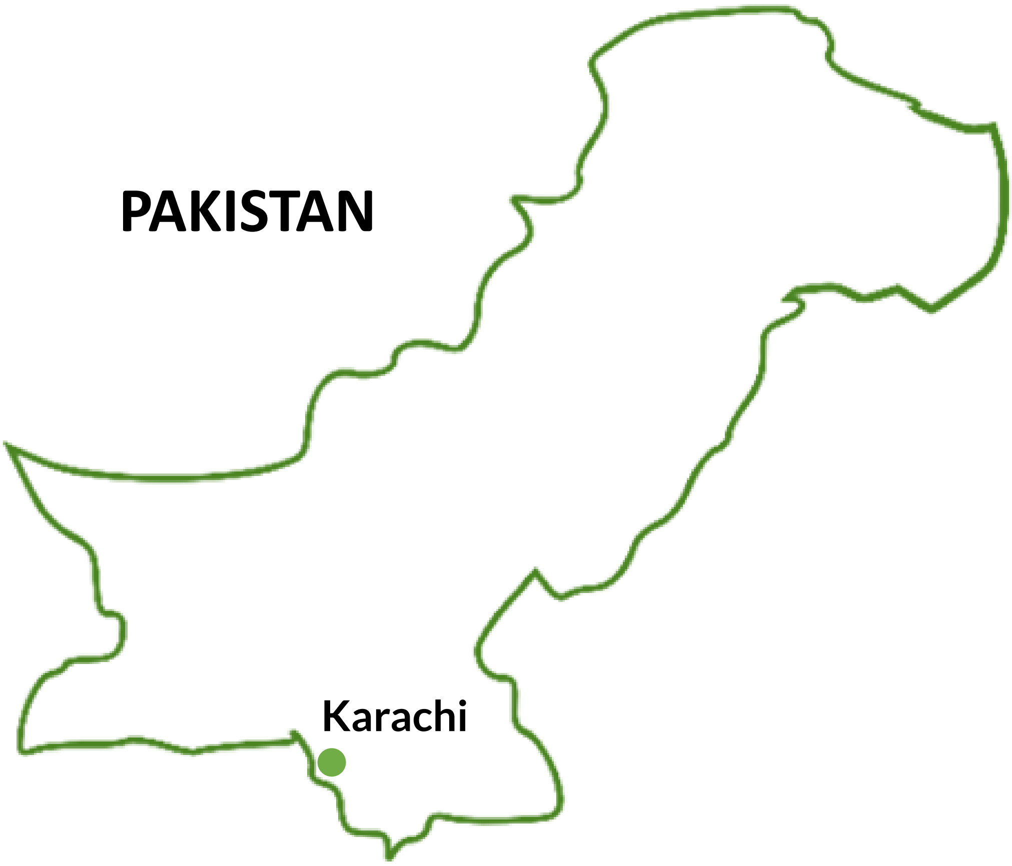


Figure 1: Map of Pakistan & Location of Karachi *(clipart-library, n.d.)*

Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of Karachi, Pakistan to open a new bakery. Using data science methodology and machine learning techniques like clustering, this project aims to provide solution to answer the business question: <u>In the city of Karachi, Pakistan, how can **Pan** establish itself by opening bakeries in a few key locations</u>?

In this project we will provide **Pan** with location data of 18 neighborhoods and all venues present within those areas. We will then proceed to analyze the location of all the bakeries in the city and identify potential segments for our debut bakery.

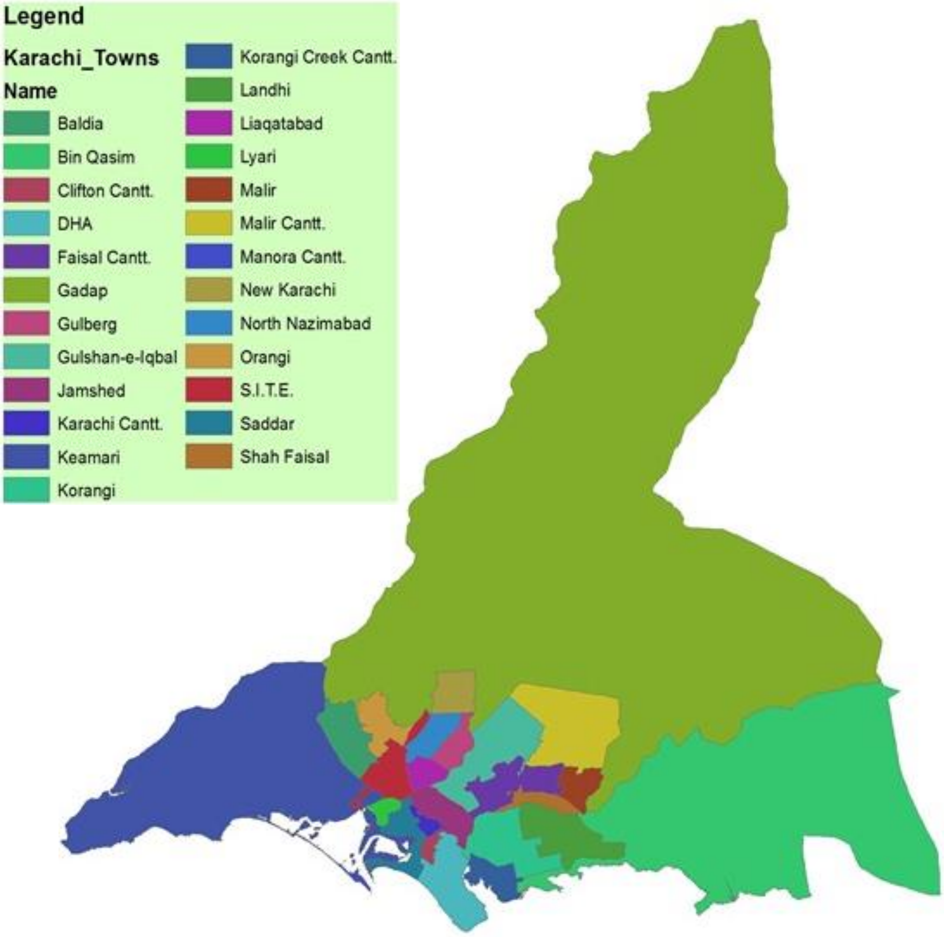


Figure 2: Map of Administrative Areas of Karachi

Data

To solve the problem, we will need the following data:

• **List of Towns (neighborhoods) in Karachi**: This defines the scope of this project which is confined to the city of Karachi, the former capital of the country Pakistan.

• **Coordinates (Latitude & Longitude)**: The latitude and longitude of these neighborhoods are required to plot them on map and get subsequent venue data.

• **Venue Data:** Venue data particularly related to bakeries. This data will be used to perform clustering on the neighborhoods to identify potential areas for new bakery opening.

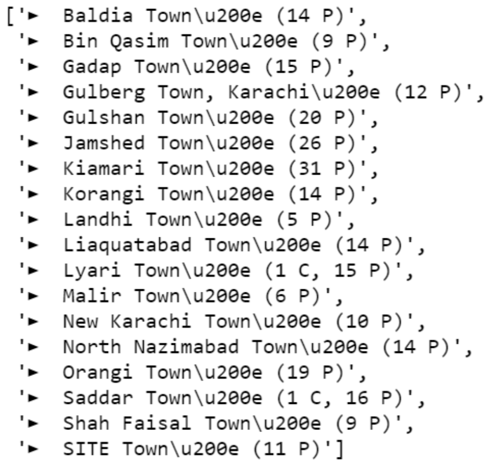
Sources of Data and Extraction Methods

1. Wikipedia page contains a list of neighborhoods in Karachi, with a total of 18 neighborhoods. (Wikipedia, n.d.)
    a. Data will be extracted from Wikipedia page, with the help of Python requests and BeautifulSoup packages. (Crummy, n.d.)
    b. Geographical coordinates of the neighborhoods will be obtained using Python OpenCageGeocode package which will give us the latitude and longitude coordinates of the respective neighborhoods. (Geocding API, n.d.)
2. Foursquare API to get the venue data for all neighborhoods (Foursquare, n.d.). Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the '**Bakery**' category in order to help us solve our business problem.

Methodology

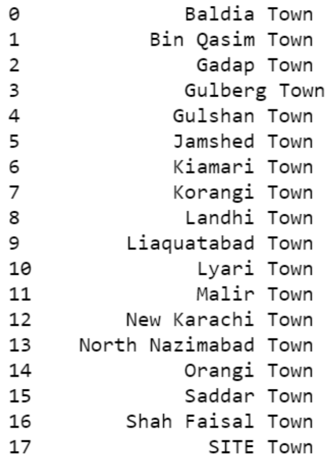
Data Acquiring & Cleansing

First we need to get the list of neighborhoods in the city of Karachi. The list is available on Wikipedia (Wikipedia, n.d.). We will scrap the webpage using python requests and BeautifulSoup packages to extract the list of neighborhoods (Crummy, n.d.). However, this is just a list of names. The data in this form cannot be used for analysis, and therefore has to be cleaned. All redundant characters and information, preceding and leading spaces are removed to get a clean list of names of the towns. These names are then stored in a dataframe and name of city (Karachi) and country (Pakistan) are added to the dataframe. This concludes the data acquiring and cleaning stage of our project.



```
['►  Baldia Town\u200e (14 P)',
 '►  Bin Qasim Town\u200e (9 P)',
 '►  Gadap Town\u200e (15 P)',
 '►  Gulberg Town, Karachi\u200e (12 P)',
 '►  Gulshan Town\u200e (20 P)',
 '►  Jamshed Town\u200e (26 P)',
 '►  Kiamari Town\u200e (31 P)',
 '►  Korangi Town\u200e (14 P)',
 '►  Landhi Town\u200e (5 P)',
 '►  Liaquatabad Town\u200e (14 P)',
 '►  Lyari Town\u200e (1 C, 15 P)',
 '►  Malir Town\u200e (6 P)',
 '►  New Karachi Town\u200e (10 P)',
 '►  North Nazimabad Town\u200e (14 P)',
 '►  Orangi Town\u200e (19 P)',
 '►  Saddar Town\u200e (1 C, 16 P)',
 '►  Shah Faisal Town\u200e (9 P)',
 '►  SITE Town\u200e (11 P)']
```

Figure 4: List of Neighborhoods before Cleaning

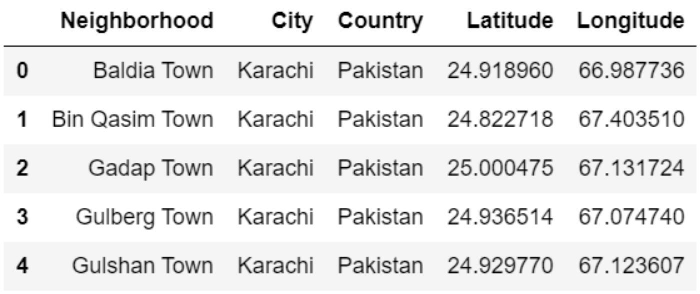


```
0            Baldia Town
1         Bin Qasim Town
2             Gadap Town
3            Gulberg Town
4           Gulshan Town
5           Jamshed Town
6           Kiamari Town
7           Korangi Town
8            Landhi Town
9       Liaquatabad Town
10            Lyari Town
11            Malir Town
12      New Karachi Town
13  North Nazimabad Town
14           Orangi Town
15           Saddar Town
16      Shah Faisal Town
17             SITE Town
```

Figure 3: List of Neighborhoods after Cleaning

Get Geographical Coordinates

We need to get geographical coordinates in the form of latitude and longitude in order use Foursquare API. For that, we will use OpenCageGeocode package that will allow us to convert address into geographical coordinates. After getting the latitude and longitude data for each location, we will populate the data into a pandas dataframe and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical co-ordinates data returned by OpenCageGeocode are correctly plotted within the city of Karachi.



	Neighborhood	City	Country	Latitude	Longitude
0	Baldia Town	Karachi	Pakistan	24.918960	66.987736
1	Bin Qasim Town	Karachi	Pakistan	24.822718	67.403510
2	Gadap Town	Karachi	Pakistan	25.000475	67.131724
3	Gulberg Town	Karachi	Pakistan	24.936514	67.074740
4	Gulshan Town	Karachi	Pakistan	24.929770	67.123607

Figure 5: Latitude & Longitudes of Neighborhoods

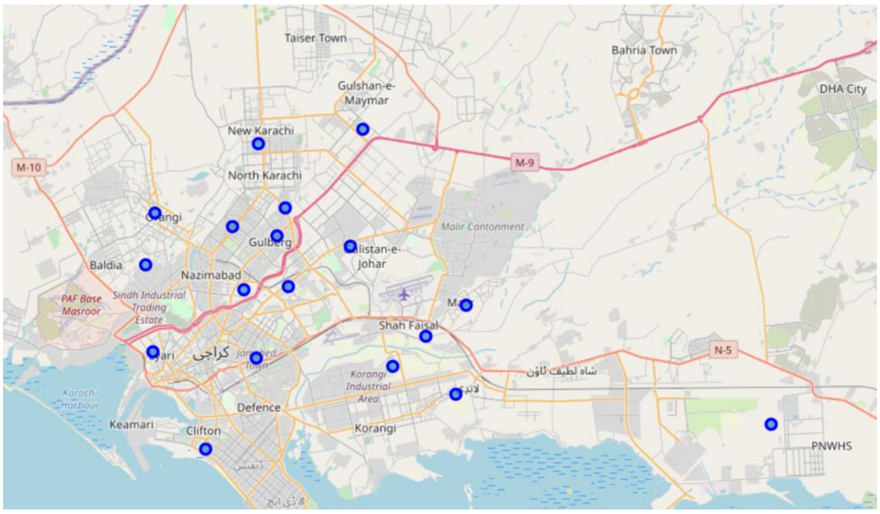


Figure 6: Neighborhoods Plotted on a Map of Karachi

Get Venue Data from Foursquare API

We will use Foursquare API to get the top 200 venues that are within a radius of 10000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. API calls are then made to Foursquare, by giving the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, category, latitude and longitude from it. With the data, we can check how many venues are returned for each neighborhood and examine how many unique categories can be curated. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data to be used in clustering. Since we are analyzing the “Bakery” data, we will filter the “Bakery” as venue category for the neighborhoods.

Neighborhood Clustering

Lastly, we will make clusters of the neighborhoods by using k-means clustering. K-means clustering algorithm identifies ‘k’ number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 4 clusters based on their frequency of occurrence for the venue category “Bakery”. The results will allow us to identify which neighborhoods have higher concentration of bakeries while which neighborhoods have fewer number of bakeries. Based on the occurrence of bakeries in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open our debut bakery.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 4 clusters based on the frequency of occurrence for "Bakery" (shown in figure 7)

Highest Concentration of Bakeries:

- Cluster 0: (**Red**) includes the following areas:
    - Shah Faisal Town
    - Site Town
    - North Nazimabad Town
    - Gulberg Town
    - Kiamari Town

Due to higher concentration of bakeries the area is already saturated and it would not be possible for a new company to establish itself.

High to Moderate Concentration of Bakeries:

- Cluster 3: (**Yellow**) includes the following areas:
    - Jamshed Town
    - Liaquatabad Town
    - Gulshan Town
    - Malir Town
    - Korangi Town
    - Baldia Town

In this clusters a lot of bakeries exist. It will be relatively less competitive but still can pose difficulties for a new chain.

Moderate to Low Concentration of Bakeries

- Cluster 2: (**Cyan**) includes the following areas:

    - New Karachi Town
    - Gadap Town
    - Orangi Town

In this cluster a moderate number of franchises exist. It would be more suitable for second phase of expansion of our retail chain.

Lowest Concentration of Bakeries

- Cluster 1: (**Blue**) includes the following areas:
    - Lyari Town
    - Saddar Town
    - Landhi Town
    - Bin Qasim Town

This cluster has the lowest number of bakeries. This would prove to be the best area for our debut bakery as competition will be less.

This clustering is purely based on the existing locations of the bakeries, however we can also consider many other factors, such as lifestyle and preferences of people living in those areas. Marketing can analyse that data and combine these 2 models to formulate a comprehensive paln.

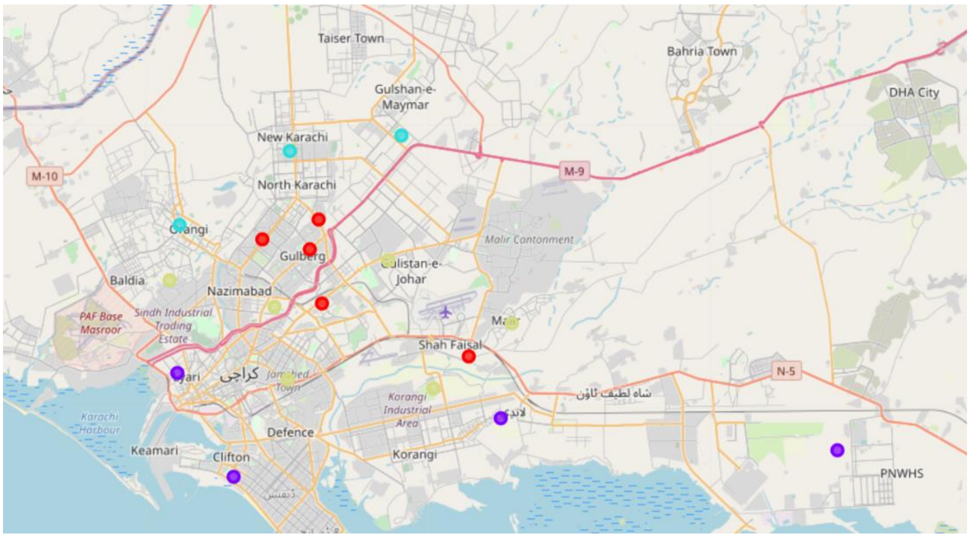


Figure 7: Clustering of Neighborhood based on presence of Bakeries

Discussion

As observations noted from the map in the Results section, most of the bakeries are concentrated in the central area of Karachi city, with the highest number in cluster 0 and a high to moderate number in cluster 3. On the other hand, cluster 1 has very low number to no bakery in the neighborhoods. This represents a great opportunity and high potential areas to open new bakeries as there is very little competition.

Location data suggests that neighborhoods on the out skirts of the city have less concentration of bakeries. They could provide good potential locations for our debut bakery, but for a decision of this magnitude, location cannot be the only deciding factor.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of bakeries, there are other factors such as population, lifestyle, habits and income of residents that could influence the decision of finding a location for the debut bakery. However, to the best available knowledge of this researcher such data is not publically available at the neighborhood level. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a bakery. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned as well as free version of OpenCageGeocode. Future research could make use of paid accounts to bypass these limitations and obtain more accurate and descriptive results.

Conclusion

In this project, we have gone through the processes of:

- Identifying a business problem and specific needs of the target market and the customer.
- Identifying data required for the project.
- Extracting, cleaning and preparing data for further analysis.
- Using machine learning for clustering the data into 4 clusters based on their similarities (presence of bakeries)

- Providing recommendations to the relevant stakeholders i.e. Pan.
- Identify potential areas in which to open the debut bakery for the chain.

To answer the business question that was raised in the introduction section, the answer proposed by this project is as follows: The neighborhoods in cluster 1 are the most preferred locations to open a new bakery as there is a low concentration of bakeries in that region and therefore competition will be less for the new company in the beginning. It will give them an opportunity to establish themselves and make a strong footing in the area before expansion of their operations. The findings of this project will help **Pan** to capitalize on the opportunities of high potential locations while avoiding areas which already have a higher concentration of bakeries.

References

Alamy. (n.d.). *Image: Bakery Items*. Retrieved from https://www.alamy.com/panorama-of-fresh-bread-products-isolated-on-white-background-image207438201.html

City Pulse. (n.d.). *Image: Karachi Town Administration Boundaries*. Retrieved from https://citypulse.com.pk/pakistangis/tag/karachi-map/

clipart-library. (n.d.). *Image - Pakistan Map*. Retrieved from http://clipart-library.com/pakistan-map-outline.html

Crummy. (n.d.). *BeautifulSoup Documentation*. Retrieved from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Foursquare. (n.d.). *Foursquare Develepors Portal*. Retrieved from https://foursquare.com/developers/apps

Geocding API. (n.d.). *OpenCage Geocoding API Documentation*. Retrieved from https://opencagedata.com/api

Shutterstock. (n.d.). *Image: Karachi Skyline*. Retrieved from https://www.shutterstock.com/es/search/pakistan+skyline

Wikipedia. (n.d.). *Category: Towns in Karachi*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Category:Towns_in_Karachi

Wikipedia. (n.d.). *Largest Cities*. Retrieved from https://en.wikipedia.org/wiki/List_of_largest_cities_in_Pakistan