STRATEGIC LOCATIONS TO OPEN A PUBLIC EATERY IN LAHORE

THE BATTLE OF NEIGHBOURHOODS

Harris Irfan

Capstone Project

12/1/2019

Contents

Abstract

This report is a part of the final project of the online course 'Applied Data Science Capstone', which is the 9th and final course in the IBM Data Science Professional Certificate.

In the project, I have explored the famous areas of Lahore using the Foursquare API data, and debated upon the ideal area to open a public eatery in Lahore.

Introduction/Business Problem

Lahore, being the capital of the province of Punjab, and the second most populous city in Pakistan would be a great choice for a potential entrepreneur looking to start a food business. The city of Lahore is famous for its rich food culture and the availability of a variety of cuisines. The people of Lahore are hailed as food lovers throughout the country.

Lahore is also a tourist attraction owing to its cultural heritage and historical sites. The city is also a major centre of urbanization, attracting many people from nearby rural areas, because of having some of the country's top universities, colleges and employment opportunities. Owing to these reasons, Lahore is a favourable city for investment in the food sector.

The potential stakeholders that would be interested in this project would be entrepreneurs willing to start a public eatery, ranging from low cost public eateries to high cost. Stakeholders might include businessmen willing to open a juice or coffee shop, or a restaurant. Furthermore, this project will also help businessmen, having well established restaurants in other cities, who are willing to expand their operations and open a restaurant in Lahore. Established international food chains or franchises, who want to setup a restaurant in Pakistan can also use the insights drawn from this project.

Considering that the city is spanning on a total area of 1772 square kilometres, it might a difficult decision to choose an area to open an eatery. In this project, Foursquare location data is primarily used to determine the perfect spot for a public eatery. The notion behind the project is to explore the famous and populous towns within the city, and suggest an area with the least competitors in the food sector and hence more potential customers. Against every town, a list is prepared of the type of and number of public eateries present. The town with the least competitors might be a favourable choice, however, it should not be forgotten that it is not the only factor that determines the potential success of restaurant. There are a lot of other factors including the location of the town within the city. For example, a central town could attract more customers than a town at the suburbs of the city. Then a town having more universities/colleges or public offices nearby could also attract more customers. We would try to explore these factors as well, but with the foursquare location data, insights regarding competitor venues can be drawn much more easily than the other factors mentioned above.

An ideal place for a restaurant or public eatery would be a densely populated area, with less food joints, having more recreational sites nearby and more offices and educational institutes.

Data

First, the popular towns, from different parts of the city, were noted. The town/area names were extracted from Google Maps. Below is a list of the areas that we will be exploring within Lahore.

- Defence Housing Authority, Phase 3
- Phase 4, DHA
- Phase 5, Defence
- Cavalry Ground
- Cantt
- Paragon City
- Gulberg 2
- Gulberg 3
- Muslim Town
- Garden Town
- Faisal Town
- Model Town
- Allama Iqbal Town
- Johar Town
- Wapda Town
- Valencia Town
- Green Town
- Awan Town
- Kot Lakhpat
- Thokar Niaz Baig
- Ichhra
- Sant Nagar
- Garhi Shahu
- Mozang Chungi
- Gulshan-e-Ravi
- Samanabad
- Old Anarkali
- Gawalmandi
- Walled City
- Chah Miran
- China Scheme
- Ghaziabad
- Daroghawala
- Nabi Pura
- Aziz Bhatti Town
- Saddar Town
- Lake City
- Bahria Town

- Mughalpura
- Singhpura

In future, we can further extend this list and introduce more areas to explore.

For each area, the latitude and longitude coordinates are fetched from the Nominatim library of the Geopy python package. With the latitude and longitude, a call is made to the Foursquare API to explore the area and get nearby famous venues information, in a range of 2 kilometres. From the venues' information, venues having public eateries as their category can be extracted.

Below is a sample from the data set that was fetched using the Foursquare API, and later transformed into a Pandas DataFrame:

Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Johnny & Jugnu	31.46613739	74.37926296	Burger Joint
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Cosa Nostra La Gelateria	31.47167664	74.37442413	Coffee Shop
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Tutti Frutti	31.47403806	74.37903722	Ice Cream Shop
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Esajee's	31.47229477	74.37719532	Food & Drink Shop
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Yum Chinese and Thai	31.47360381	74.37878442	Chinese Restaurant
Defence Housing Authority, Phase 3	31.4715878	74.3816205	McDonald's	31.47498499	74.37720688	Fast Food Restaurant
Defence Housing Authority, Phase 3	31.4715878	74.3816205	CC Park, DHA	31.46464303	74.3803715	Park
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Bundu Khan	31.47380254	74.37906897	BBQ Joint
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Hardees	31.47526556	74.37926676	Fast Food Restaurant
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Subway	31.47275785	74.3767199	Sandwich Place
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Espresso	31.471632	74.37665	Cafe
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Gloria Jean's Coffees	31.47397263	74.37898269	Coffee Shop
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Gourmet Grill	31.47444467	74.37871793	Pakistani Restaurant
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Bar.B.Q Tonight	31.47471411	74.37922062	Afghan Restaurant
Defence Housing Authority, Phase 3	31.4715878	74.3816205	English Tea House	31.47368438	74.37900092	Cafe
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Second Cup	31.47410521	74.37904999	Cafe
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Timmy's	31.474687	74.379013	Burger Joint
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Rina's Kitchenette	31.47200218	74.37424174	Cafe
Defence Housing Authority, Phase 3	31.4715878	74.3816205	SUBWAY	31.463472	74.384833	Sandwich Place
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Y Block Commercial Area	31.47267016	74.37690558	Market
Defence Housing Authority, Phase 3	31.4715878	74.3816205	Jalal Sons	31.47428583	74.37881167	Diner

Figure 1.0: Venues Data from Foursquare

The area/neighbourhood that is explored is listed in the first column, along with its latitude and longitude in the following two columns. Against the area, all famous nearby venues, fetched by the Foursquare API are listed along with their respective latitude and longitude values, and their category. The category column would be of most interest for us, as we can distinguish the venues that are public eateries from other venues. Mostly, the category information fetched is elaborative enough to even determine the cuisine of the restaurant or the type of public eatery. Using this information, against every neighbourhood, we can determine the neighbourhood with the least number of competitors in the fast food business for example.

Methodology

The count of the venues returned by the Foursquare API can also be one of the determining factors, as it gives an idea as to how popular the neighbourhood is, attracting potential customers.

Before diving further into the venues' details, let's first get a bird's eye view of the neighbourhoods that we explored in Lahore:

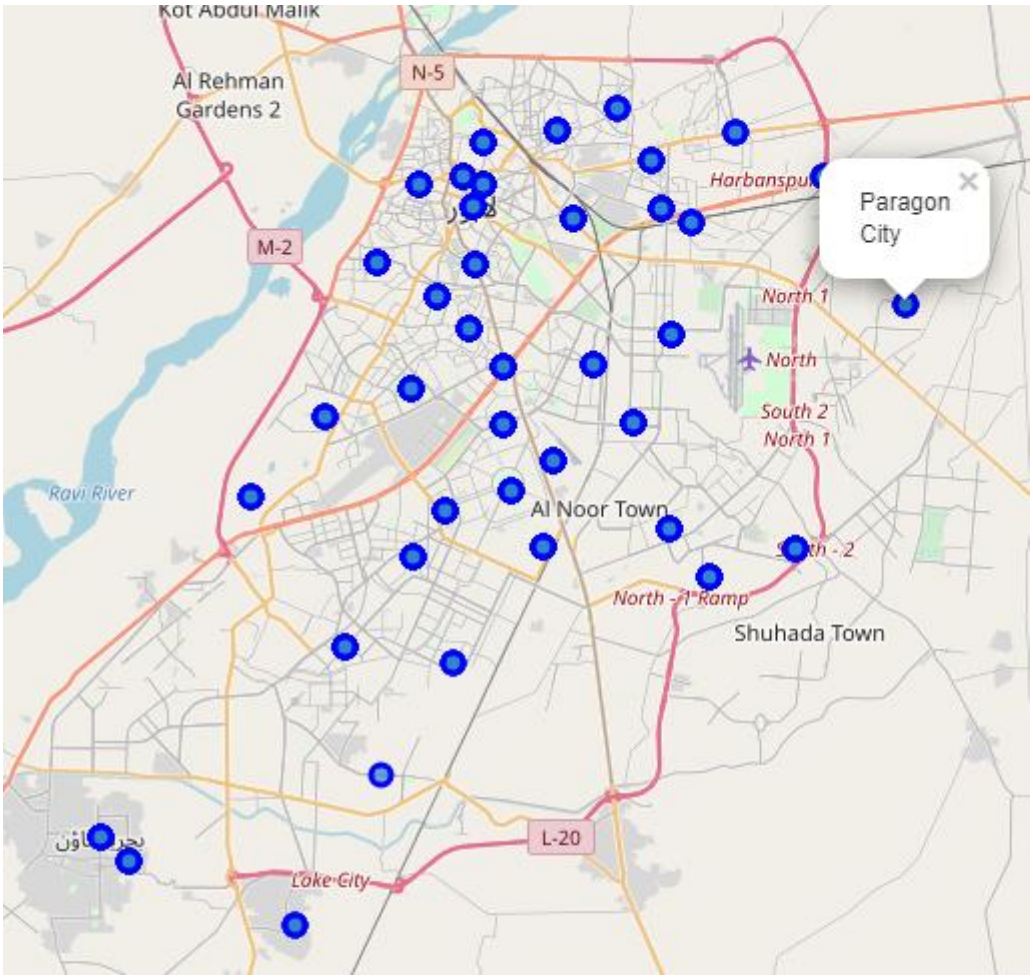

Figure 2.0: Map of Lahore

Below is a visual representation of the number of venues fetched against every neighbourhood.

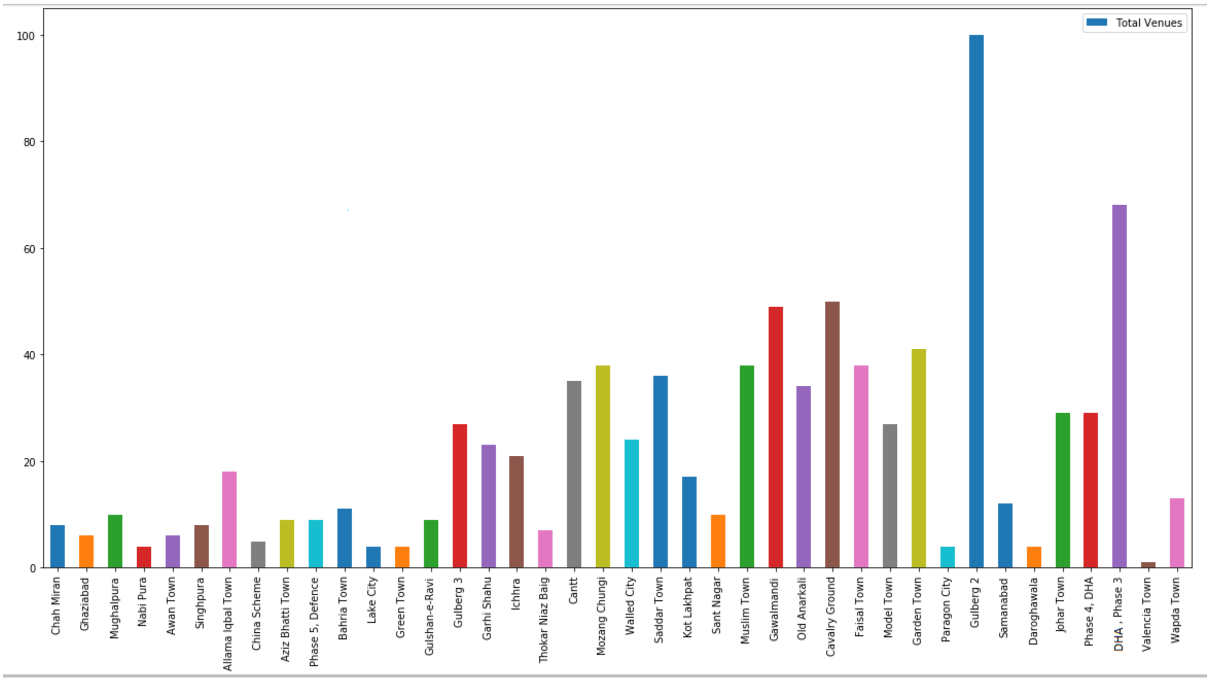

Figure 3.0: Total Number of Venues Against Neighbourhoods

As mentioned above, it is critical to filter food related venue from other venues in this project. It is manually noted from a sample data set that if the below mentioned key words are present within the category description of a particular venue, then the venue can be associated with a 'Food' Venue.

Food Related Categories – Key Words:

- Burger
- Coffee
- Cream
- Food
- Drink
- Restaurant
- Fast
- BBQ
- Sandwich
- Café
- Diner
- Pizza
- Dessert
- Bakery
- Donut
- Steak
- Steakhouse
- Breakfast
- Fish
- Chips

- Snack
- Tea
- Frozen
- Yogurt
- Deli
- Bodega
- Seafood
- Fried
- Chicken
- Juice
- Bar

All other venues would fall in the ‘Other Venues’ Category. Remember, a neighbourhood with the least Food Venues, and most Other Venues would be a favourable choice for our problem.

Count of food venues was also calculated for all neighbourhoods, and below is a visual representation of % of food venues against all neighbourhoods (food venues/total venues). The neighbourhoods were sorted in increasing % of food venues:

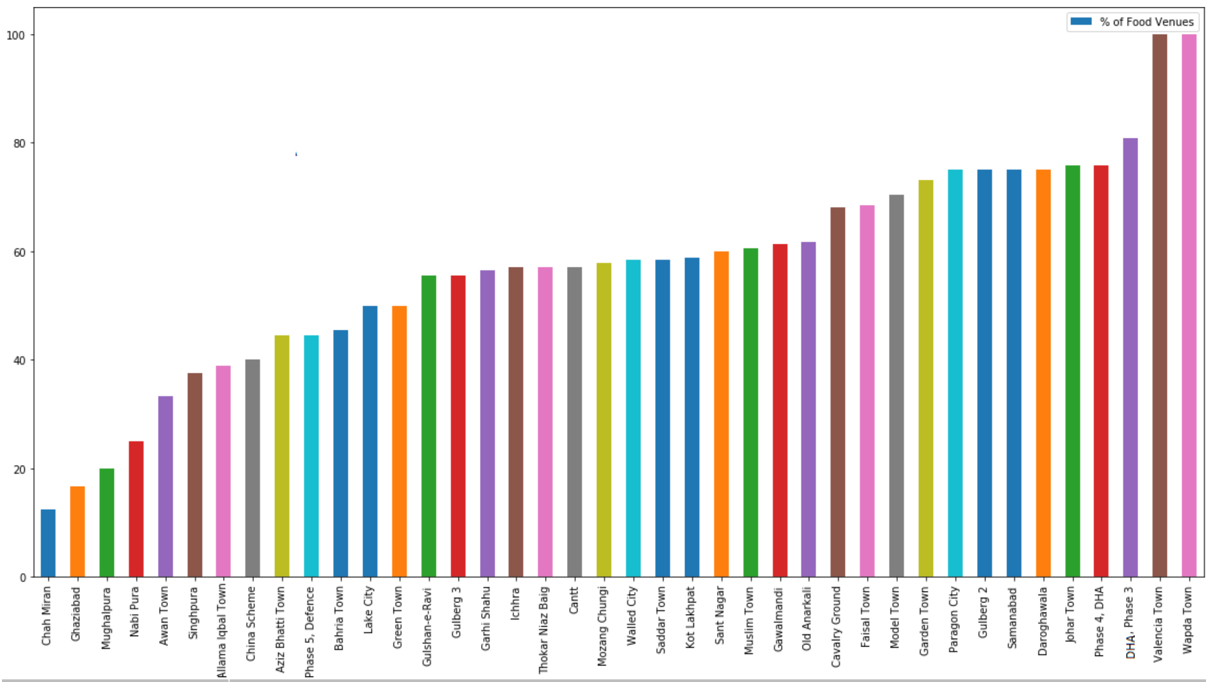

Figure 4.0: % of Food Venues in Neighbourhoods

All of this exploratory data analysis was done with the foursquare data, and manually identifying food-related key words as listed above.

Later, clustering technique was applied to cluster the neighbourhoods into three clusters on the basis of venues to draw insights. Clustering technique used was K Means clustering with k = 3.

The motivation behind using the clustering technique is that we have unlabelled data set. Therefore, we have to resort to unsupervised machine learning. We do have the venue information of all the neighbourhoods; however, target values i.e. the profit of each restaurant for example, are not present. Because of confidentiality, data related to the profit of each public eatery is not available. We are also trying to determine the potential success of a particular eatery in different neighbourhoods.

Clustering was applied after applying one-hot encoding to the venue categories shown in Figure 1.0.

There were 118 unique categories curated from the data.

```
area_venues_df_onehot = pd.get_dummies(area_venues_df[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
area_venues_df_onehot['Area'] = area_venues_df['Area']

# move neighborhood column to the first column
fixed_columns = [area_venues_df_onehot.columns[-1]] + list(area_venues_df_onehot.columns[:-1])
area_venues_df_onehot = area_venues_df_onehot[fixed_columns]

area_venues_df_onehot.head()
```

	Area	Accessories Store	Afghan Restaurant	Airport	Airport Terminal	American Restaurant	Asian Restaurant	BBQ Joint	Badminton Court	Bakery	...	Supermarket	Tea Room	Thai Restaurant	Theater	Theme Park	Train Station	Turkish Restaurant	Video Store	Wom St
0	Defence Housing Authority, Phase 3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
1	Defence Housing Authority, Phase 3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
2	Defence Housing Authority, Phase 3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
3	Defence Housing Authority, Phase 3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	

Figure 5.0: One-hot encoding of venue categories

area_venues_df is the DataFrame shown in Figure 1. As shown, against every area, the categories of the venues that were present in that area were marked as 1 while other categories were marked as 0.

Later this data was grouped by neighbourhood and by taking the mean of the frequency of occurrence of each category:

```
lahore_grouped = area_venues_df_onehot.groupby('Area').mean().reset_index()
lahore_grouped
```

	Area	Accessories Store	Afghan Restaurant	Airport	Airport Terminal	American Restaurant	Asian Restaurant	BBQ Joint	Badminton Court	Bakery	...	Supermarket	Tea Room	Thai Restaurant	Theater	Theme Park	Train Station	Turkish Restaurant
0	Allama Iqbal Town	0.055556	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.055556	...	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00
1	Awan Town	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.166667	0.00	0.000000	...	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00
2	Aziz Bhatti Town	0.000000	0.000000	0.000000	0.000000	0.000000	0.111111	0.000000	0.00	0.000000	...	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00
3	Bahria Town	0.000000	0.000000	0.000000	0.000000	0.000000	0.090909	0.000000	0.00	0.000000	...	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00
4	Cantt	0.000000	0.000000	0.000000	0.028571	0.000000	0.000000	0.028571	0.00	0.142857	...	0.028571	0.00	0.000000	0.000000	0.028571	0.000000	0.00
5	Cavalry Ground	0.000000	0.000000	0.000000	0.000000	0.040000	0.020000	0.020000	0.00	0.020000	...	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.02
6	Chah Miran	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	...	0.000000	0.00	0.000000	0.000000	0.125000	0.125000	0.00
7	China Scheme	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	...	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00
8	Daroghawala	0.000000	0.000000	0.000000	0.000000	0.000000	0.250000	0.000000	0.25	0.000000	...	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00
9	Defence Housing Authority, Phase 3	0.000000	0.029412	0.000000	0.000000	0.000000	0.000000	0.014706	0.00	0.014706	...	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00
10	Faisal Town	0.000000	0.000000	0.000000	0.000000	0.026316	0.052632	0.026316	0.00	0.131579	...	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00
11	Garden Town	0.000000	0.000000	0.000000	0.000000	0.024390	0.048780	0.073171	0.00	0.000000	...	0.000000	0.00	0.024390	0.024390	0.000000	0.000000	0.00

Figure 6.0: Grouping Neighbourhood Data

Clustering was carried out on the grouped data using the cluster module from Python's scikit-learn Package:

```
from sklearn.cluster import KMeans

# set number of clusters
kclusters = 3

lahore_grouped_clustering = lahore_grouped.drop('Area', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(lahore_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[:10]
```

Figure 7.0: Clustering Using K-Means

For details regarding the result of the clustering technique applied on the data, and discussion regarding the exploratory data analysis, please refer to the Results and Discussion section below.

Results and Discussion

According to Figure 3.0, Gulberg 2 can be considered as one of the most popular towns within the city. It has 100 venues in the Foursquare Location Data, while other areas barely compete with Gulberg 2 in terms of the number of famous venues. However, it is pertinent to note here that Gulberg 2 has 75% food venues (Figure 4.0), meaning that it would offer a lot of competition to a new entrepreneur. However, if a stakeholder is opening a unique eatery or restaurant that is new for the city, for example a Japanese restaurant, then they can target this area as there won't be any competition with other Japanese Food eateries. It was drawn from the data that there are no Japanese Restaurants in the selected areas of the city according to Foursquare data.

Chah Miran, Ghaziabad and Mughalpura can be considered, as they have the least % of food venues. However, they also have very small number of other venues, and hence they probably won't attract much customers.

To go for a compromise, Allama Iqbal Town seems to be a better choice, having 18 total venues and 7 food venues in the vicinity. Gulberg 3, having 12 food venues and a total of 27 venues is also a good choice. Gawalmandi, Old Anarkali, Garhi Shahu and Ichhra are also favourable areas.

Having discussed the exploratory data analysis, let's move on to the results of clustering. The majority of the neighbourhoods were distributed to two clusters, while only one neighbourhood, Valencia Town, was placed in the third cluster.

The top five most common venues were listed against every neighbourhood that was assigned a cluster, to examine the nature of the neighbourhoods of a particular cluster. Below are the results:

Area	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Allama Iqbal Town	0	Fast Food Restaurant	Department Store	Park	Market	Accessories Store
Aziz Bhatti Town	0	Movie Theater	Department Store	Golf Course	Café	Fast Food Restaurant
Bahria Town	0	Movie Theater	Department Store	Restaurant	Café	Fast Food Restaurant
Cantt	0	Bakery	Pakistani Restaurant	Fast Food Restaurant	Shopping Mall	Movie Theater
Cavalry Ground	0	Café	Hotel	Pakistani Restaurant	Fast Food Restaurant	Coffee Shop
China Scheme	0	Historic Site	Pakistani Restaurant	Fast Food Restaurant	Business Service	Shopping Mall
Daroghawala	0	Pakistani Restaurant	Asian Restaurant	Badminton Court	Fish & Chips Shop	Zoo
Defence Housing Authority, Phase 3	0	Café	Burger Joint	Fast Food Restaurant	Pizza Place	Coffee Shop
Garden Town	0	Pakistani Restaurant	Café	BBQ Joint	Bus Line	Asian Restaurant
Garhi Shahu	0	Hotel	Chinese Restaurant	Pakistani Restaurant	Mediterranean Restaurant	Lounge
Gawalmandi	0	Pakistani Restaurant	Breakfast Spot	Hotel	Restaurant	Food Court
Green Town	0	Coworking Space	Pizza Place	Business Service	Food & Drink Shop	Zoo
Gulberg 2	0	Café	Chinese Restaurant	Pakistani Restaurant	Shopping Mall	Fast Food Restaurant
Gulberg 3	0	Café	Department Store	Coffee Shop	Bus Line	Breakfast Spot
Johar Town	0	Fast Food Restaurant	Pizza Place	Asian Restaurant	Ice Cream Shop	Café
Lake City	0	Diner	Farmers Market	Multiplex	Burger Joint	Zoo
Model Town	0	Pizza Place	Asian Restaurant	Park	Bakery	Coffee Shop

Figure 8.0: Cluster 1

Area	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Awan Town	1	Ice Cream Shop	Clothing Store	Housing Development	Department Store	BBQ Joint
Chah Miran	1	Historic Site	Pizza Place	Garden	Bus Station	Train Station
Faisal Town	1	Shopping Mall	Bakery	Pizza Place	Pakistani Restaurant	Department Store
Ghaziabad	1	Department Store	Clothing Store	Market	Bakery	Flea Market
Gulshan-e-Ravi	1	Bus Station	Bakery	Department Store	Mexican Restaurant	Pakistani Restaurant
Ichhra	1	Bakery	Market	Bus Station	Ice Cream Shop	Fried Chicken Joint
Kot Lakhpat	1	Shopping Mall	Ice Cream Shop	Furniture / Home Store	Pizza Place	Convenience Store
Mozang Chungi	1	Fast Food Restaurant	Market	Asian Restaurant	Bakery	Breakfast Spot
Mughalpura	1	Bakery	Market	Department Store	Flea Market	Multiplex
Nabi Pura	1	Badminton Court	Metro Station	Brazilian Restaurant	Bus Station	Electronics Store
Samanabad	1	Bakery	Mexican Restaurant	Fast Food Restaurant	Snack Place	Pizza Place
Singhpura	1	Clothing Store	Bakery	Department Store	Fast Food Restaurant	Market

Figure 9.0: Cluster 2

Figure 8.0 does not show all the areas in Cluster 1; however, it is enough to draw some insights. Figure 9.0 on the other hand shows all the areas in Cluster 2.

According to the venues' information of areas belonging to Cluster 1, it can be estimated that the majority of these area have a higher percentage of food venues, like cafes or restaurants. Areas in Cluster 2 happen to have more venues other than food venues, like recreational sites, Bus Stations, Markets, etc. Hence, areas in cluster 2 could be of more interest to our target audience for this project.

Conclusion

In the previous sections, we discussed that areas with less food venues and more other venues would be favourable choices to open a public eatery. There were some areas that were short listed as well, like Allama Iqbal Town, Garhi Shahu, Old Anarkali, Gawalmandi, etc. However, it is pertinent to note that the factors that we have explored, like the number of total venues or % of food venues in a neighbourhood are only few factors among many others which choosing a place to open a public eatery. They would vary depending on the type of eatery that the stakeholder is willing to launch. For example, high-cost eateries would have different factors as compared to low-cost ones.

The per-person capita data of people living in a neighbourhood is essential if one plans launch an expensive eatery. However, this data was not available. Then data related to the population of the neighbourhoods is also equally important, as densely populated neighbourhoods would attract more customers. However, the Lahore demographic data is not present at neighbourhood or union level. It is only present at tehsil (sub-division of a district) level, which is not much of use to us.

Another very important factor to consider would be the rent prices of commercial areas in the targeted neighbourhoods, or sale prices if the stakeholder is willing to buy a place for their eatery. Such a data set was also not available. It would take a lot of web scraping to get some data.

In conclusion, the inferences drawn in this project are reasonable for stakeholders, but only in the scope of a particular factor to determine the success of an eatery in Lahore. There are much more factors that stakeholders need to consider. They can however, use the inferences from this project to their benefit to some extent.

For details, you can refer to Python Notebook link present in the Appendix.

References

[1] Foursquare API

[2] Google Map

Appendix

Notebook URL:

https://github.com/harris4188/Capstone-Project---The-Battle-of-Neighbourhoods/blob/master/The%20Battle%20of%20Neighbourhoods.ipynb