AITEZAZ AKHTAR

Exploring Neighborhoods in Karachi, Pakistan

Report

1. Introduction

1.1 Background

Karachi is the largest and most populous metropolitan city in Pakistan. Karachi had an estimated population of more than 16 million people, which makes it the 7th largest urban agglomeration and the largest city in the Muslim world.

Karachi is the city of lights that never sleeps and it has always been a great attraction to the tourists for the variety of activities and tourist spots that it has to offer. Karachi is as renowned for its beaches and dining as it is for its architecture and atmosphere.

1.2 Problem Statement

Whenever a person is visiting a city they start looking for best places to visit during their stay. They primarily look for places based on the venue ratings across all venues and the average prices such that the locations fit in their budget.

Thus, our aim here is to identify places that someone can visit during their stay. Here, we'll explore the neighborhoods of different borough to find the 10 most common venues in each neighborhood and identify places that are fit for various individuals based on the data collected.

1.3 Interest

Expats who are considering visiting Karachi will be interested to explore its neighborhoods and common venues around each neighborhood.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

2.1.1 Using Foursquare Location Data

I have used the Foursquare API to retrieve information about the popular spots around the Neighborhood of Karachi. It returns venues from each Neighborhood within radius of 2km.

From Foursquare API (https://developers.zomato.com/api), I retrieved the following for each venue:

- Name: The name of the venue.
- Category: The category type as defined by the API.

- Latitude: The latitude value of the venue.
- Longitude: The longitude value of the venue.
- Price : Price range of the venue as Cheap, Moderate and Very Expensive
- Tips : Reviews received by the venue
- Ratings : Ratings giving by each customer in range of 1-10

2.1.2 Getting Coordinates of Major Districts : Geopy Client

Next objective is to get the coordinates of these boroughs using geocoder class of Geopy client.

2.1.3 Scraping Karachi District and Town from Wikipedia

We can get the data of different district and town through scraping different Wikipedia pages and create a dataframe using this information. For this, I've used requests and Beautifulsoup4 library to create a data-frame.

The data source is the list of Neighborhoods and Boroughs as found on a wikipedia page. This dataset is created from scratch using the list of neighborhood available on the site, the following are columns:

- Neighborhood: Name of the neighborhood in the Borough.
- Borough: Name of the Borough.

2.2 Data Cleaning

The data preparation for each of the three sources of data is done separately. First we extract Boroughs from Wikipedia page and then using these boroughs we extract these neighborhoods using web scraping. We drop the boroughs which do not has neighborhood data on Wikipedia page.

	Borough	Neighborhood
0	Bin Qasim Town	Abdullah Goth
1	Bin Qasim Town	Cattle Colony
2	Bin Qasim Town	Gaghar
3	Bin Qasim Town	Green Park City
4	Bin Qasim Town	Gulshan-e-Hadeed

The second data which we will extract is the coordinates (latitude and longitude) of each neighborhood from the extracted list. To do this we will using geocoder class of Geopy client. We drop the data which do not has coordinates in it.

	Borough	Neighborhood	Latitude	Longitude
0	Bin Qasim Town	Abdullah Goth	24.8665	67.2837
1	Bin Qasim Town	Cattle Colony	24.8793	67.1987
3	Bin Qasim Town	Green Park City	24.8577	67.2225
4	Bin Qasim Town	Gulshan-e-Hadeed	24.87	67.3601
5	Bin Qasim Town	Ibrahim Hyderi	24.7916	67.1409

The third data is the venues list from each neighborhood. First we explore the venues from each neighborhood, we get around 3000+ venues from it. Then we explore around 500 venues because of limitation of API calls. Than we narrow down our venues list by selecting only that venue which has rated, has received tips and have mentioned its price ranges, so in the end we left with 247 venues.

	Ratings	Tips	Price
0	7.1	13	Moderate
17	8.6	5	Cheap
18	7.5	2	Cheap
19	7.3	8	Cheap
20	7.6	3	Cheap

Finally, we merged our dataset. As a final dataset, we're left with 247 venues with 10 columns as described in below figure.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Ratings	Tips	Price
0	Cattle Colony	24.879311	67.198723	Anwar Baloch	24.869620	67.200499	BBQ Joint	8.7	17.0	Cheap
1	Cattle Colony	24.879311	67.198723	The Broast Restaurant	24.884536	67.182993	Fast Food Restaurant	8.6	5.0	Cheap
2	Cattle Colony	24.879311	67.198723	Student Biryani	24.884636	67.182260	Restaurant	8.6	5.0	Cheap
3	Cattle Colony	24.879311	67.198723	Taj Chaiye Ka Hotel	24.887062	67.183052	Tea Room	8.6	5.0	Cheap
4	Cattle Colony	24.879311	67.198723	Liaquat Market	24.887072	67.183061	Market	8.6	5.0	Cheap

3. Methodology and Exploratory Data Analysis

3.1 Methodology

As a first step, I retrieve the boroughs and using boroughs data extract its neighborhood. Second we would like to extract coordinates of each neighborhood which would be done by using Geopy Client. Third, I retrieve the venues in Karachi from Foursquare API up to a distance of 2 kilometers from each Neighborhood. Using this, I fetch the venue information including price and rating data. The final data will include the neighborhood, neighborhood latitude, neighborhood longitude, venue name, category, latitude, longitude, rating, price range. Using this dataset, I begin by analyzing the top venue types that exist in Karachi. I will then explore the venues on maps. This will allow us to better understand the location of various venues and the places where many venues co-exist and create place worth visiting. I'll also explore the venues based on the ratings and price range of various venues. The venues will be plot using proper color coding such that a simple glance at the map would reveal the location of the venues as well as give information about them. I aim to identify places which can be recommended to visitors based on their price and rating preferences. I'll also cluster the venues and see if we can draw meaningful information out of what kind of venues exist in Karachi. As a final step, I will analyze these plots and try to draw conclusions on what places can be recommended to visitors. I'll discuss my findings and any inferences I can draw.

3.2 Exploratory Data Analysis

3.2.1 Most Venues

We have different types of venues in each neighborhood. We will take a look at the venues and check which the majority venue neighborhood in the list.

Neighborhood	Venue
Clifton	100
Mahmudabad	100
Bath Island	100
Shah Rasool Colony	100
Civil Lines	91

From the above result we can clearly see that **Clifton, Mahmudabad and Bath Island** has most values in the list. So, if you are a tourist, you may find many places to enjoy there

3.2.2 Categories

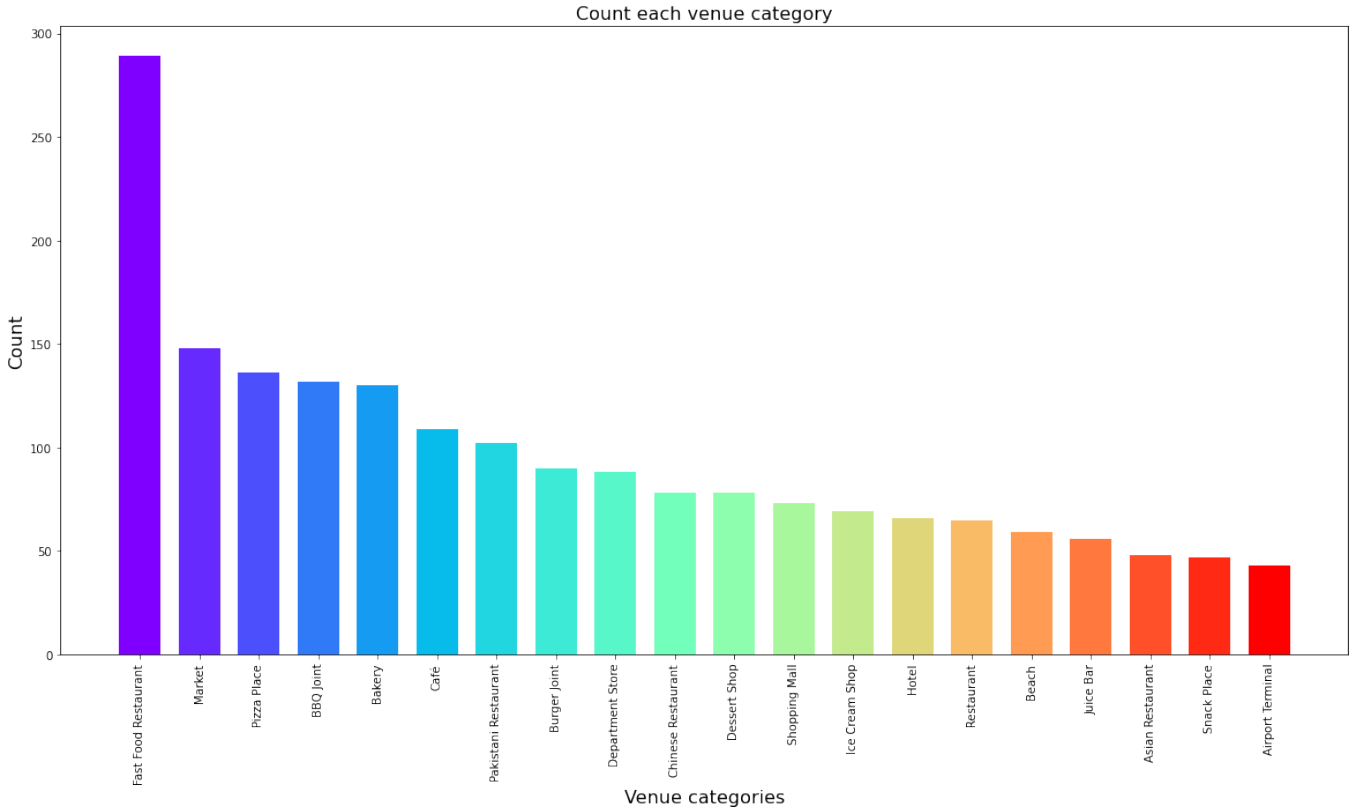
Now we will take a look at the venues and check how many unique venue categories are and which the majority venue categories in the list.

```
print('There are {} uniques venue categories visitor can find in Karachi'.format(len(khi_venues['Venue Category'].unique())))
```

```
There are 150 uniques venue categories visitor can find in Karachi
```

There are 150 uniques venue categories visitor can find in Karachi

We have various types of venues in the final dataset. We will take a look at the venues and check which the majority venue categories in the list.



We see that the majority venues are actually Fast Food Restaurant. This is followed by Markets and Pizza Place.

3.2.3 Ratings

Rating of a venue is an important factor on which a visitor decides whether it is worth it to visit the place. We'll first identify the various rating values and plot them as a bar plot with their counts to see the most common rating.

	Venue	Venue Latitude	Venue Longitude	Venue Category	Ratings	Tips	Price
0	Anwar Baloch	24.869620	67.200499	BBQ Joint	8.7	17.0	Cheap
1	The Broast Restaurant	24.884536	67.182993	Fast Food Restaurant	8.6	5.0	Cheap
2	Student Biryani	24.884636	67.182260	Restaurant	8.6	5.0	Cheap
3	Taj Chaiye Ka Hotel	24.887062	67.183052	Tea Room	8.6	5.0	Cheap
4	Liaquat Market	24.887072	67.183061	Market	8.6	5.0	Cheap

We can see that **Anwar Baloch** has highest rating so this might be the place you have to visit

We can also analyze which neighborhood has highest rating restaurant

Neighborhood	Ratings
Cattle Colony	8.477778
Green Park City	8.200000
Gulshan-e-Hadeed	8.200000
Landhi Colony	7.900000
Ibrahim Hyderi	7.900000

Here we can see that **Cattle Colony** has the highest rating restaurant

3.2.4 Price

We will now take a look the venues based on the price ranges. We have three price features for our venues, Cheap, Moderate and Very Expensive according to Foursquare API

Now we map the cheap restaurant in map

Users might also be interested in going to a place that fits in their budget. We will use the **Price** column to plot the venues on a map. We'll represent the venues with lower price in green and move towards red as the price increases.

		Price
Neighborhood	Price	
Mujahid Colony	Cheap	29
Gulshan-e-Iqbal I	Cheap	29
Shanti Nagar	Cheap	27
Gulshan-e-Iqbal II	Cheap	18
Gulistan-e-Johar	Cheap	17

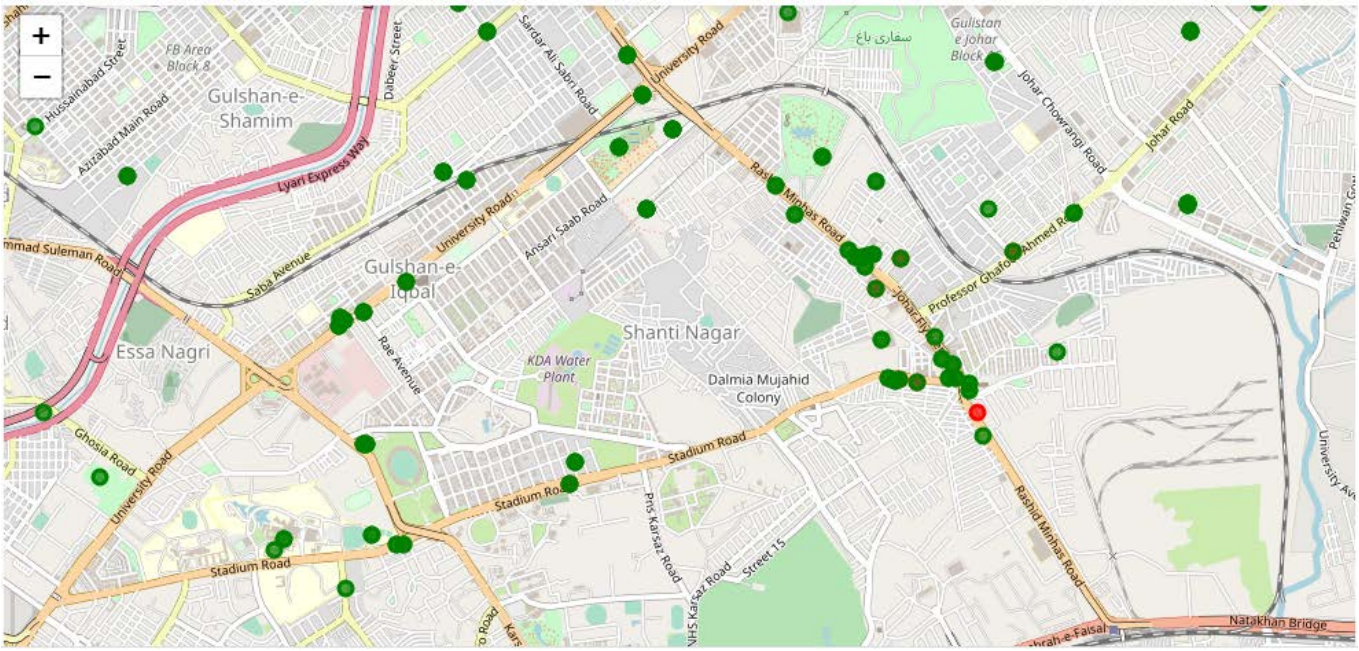
We can see that most of the cheap restaurants are found in **Gulshan-e-Iqbal and Mujahid Colony.**

3.2.5 Clustering

Finally, I cluster all the venues based on their price range, location and more to identify similar venues and the relationship amongst them. I used KMeans clustering and decided to cluster the venues into two separate groups.

We see the two clusters in below figure:

1. The first cluster (green) is spread across the whole city and includes the majority venues. These venues have rating spread around 6.72.
2. The second cluster (red) is very sparsely spread and has very limited venues. These venues have rating spread around7.59.



4. Results and Discussion

Based on our analysis above, we can draw a number of conclusions that will be useful to aid any visitor visiting the City of Lights Karachi, Pakistan. After collecting data from the Foursquare API, we got a list of 3113 different venues. However, we narrow it down to only 500 venues due to limitation of Premium API Calls.

From the above result we can identify that majority of venues are in Clifton, Mehmodabad and Bath Island. Also from list of venues majority them were Fast Food Restaurant, Markets and BBQ joint. A visitor who loves Fast Food would surely benefit from coming to Karachi. While the complete range of ratings range from 1 to 10, the majority venues have ratings close to 9. This means that most restaurants provide good quality food which is liked by the people of the city, thus indicating the high rating. From the data we also identify that most high rated restaurant are in Cattle Colony. When we take a look at the price ranges of each venue, we explore that many majority venues have lower prices. We also explore that majority of lower price range restaurants are in Gulshan-e-Iqbal.

Finally, through clusters we identified that there are many venues which are relatively lower priced but have an average rating of 3.57. On the other hand, there are few venues which are high priced and have average rating of 4.03. If you're looking for cheap places with relatively

high rating, you should check Gulshan-e-Iqbal. If you're looking for the best places, with the highest rating but might also carry a high price tag, you should visit Clifton.

A company can use this information to build up an online website/mobile application, to provide users with up to date information about various venues in the city based on the search criteria (name, rating and price).

5. Conclusion

The purpose of this project was to explore the places that a person visiting Karachi could visit. The venues have been identified using Foursquare API and have been plotted on the map. The map reveals that there are three major areas a person can visit: Cattle Colony, Clifton & Gulshan-e-Iqbal. Based on the visitor's venue rating and price requirements, he/she can choose amongst the three places.