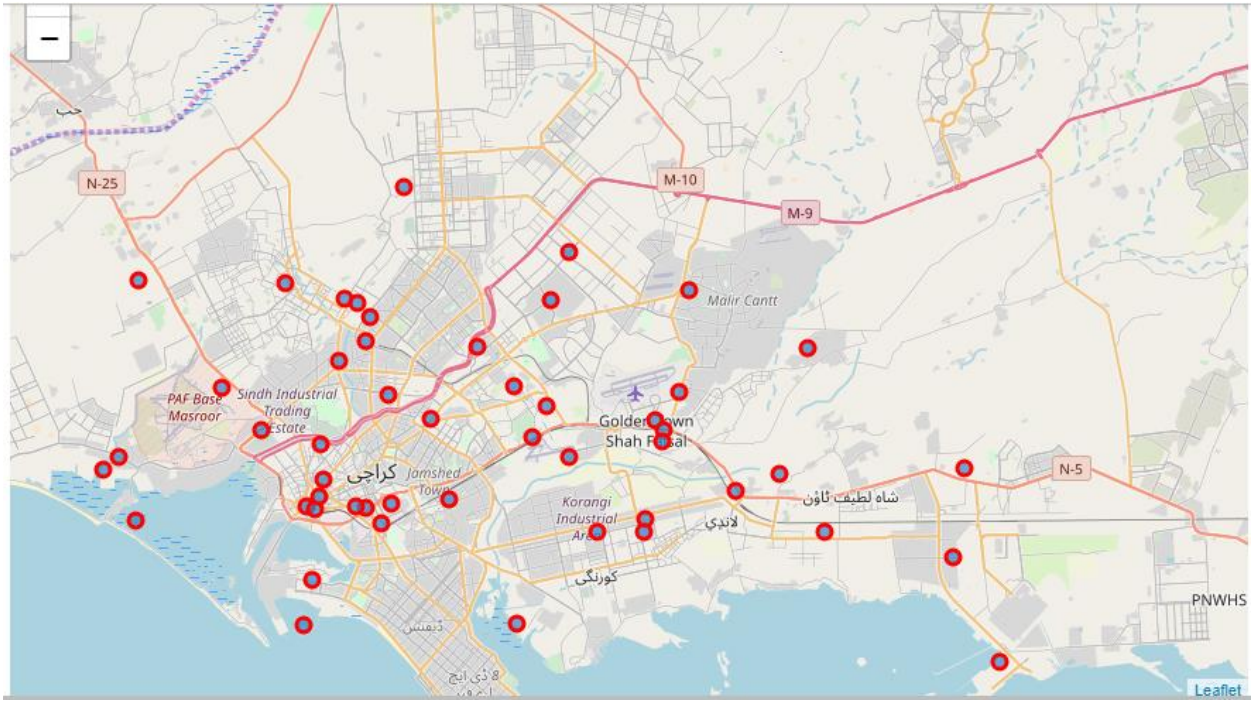
IBM Applied Data Science Capstone

ANALYZING TOWNS IN KARACHI AND IDENTIFYING POTENTIAL BUSINESS OPPORTUNITIES

Muhammad Saad Uddin

Jan 2020



INTRODUCTION

Karachi, despite being the biggest city and commercial hub of the country, very little data and statistics is available on its demographics. consumptions and trends. Several business and startups in the city have to research from scratch to find optimum locations or neighborhood for their product/services success with reasonable property valuation, So I've come up with a plan to analyze all the towns/Neighborhood of Karachi and find what type of venues each town is mostly consist of using towns/Neighborhood postal codes and leveraging Foursquare API.

Business Problem

Since there is very minimal data available for Karachi and what available is neither compact nor decision ready, existing and new both businesses of any type are having problem in finding a place which is both economically viable and suit their business model.

Target Audience

This project can be utilized by several groups, mainly business owners and new startups also tourist and citizens can use this project to find places of any category near where they are. Government can also utilize this project to track a certain town growth and what measure the decision makers can take to further improve particular towns conditions and attractions.

DATA

Following data has been used to work on this plan

- Postal Codes of all Neighborhood in Karachi
- Population Statistics of each Neighborhood
- Property estimated Value of each Neighborhood

Source of Data and method of extraction

The required data will be first extracted from KMC, Census, FBR Website and Pakistan Postal Codes website, then this data will be used to get coordinates of each town/neighborhood. Then leveraging API of Foursquare, I will call venues of each town and then perform necessary analysis on that data which will be useful for the targeted audience.

METHODOLOGY

The Analysis will start with scraping the webpage of Karachi metropolitan Corporation's for all areas of Karachi and their postal codes. Then I will scrape data of population, houses and commercial value per sq ft from FBR and Census Websites. Since this data would take extra effort for working in dataframe, I have cleaned it on excel sheet. Then for the purpose of analysis, I have merged both dataframes into one. After that I used Geocode for getting all the latitude and longitude of required areas.

This dataframe is cleaned and ready for analysis and before that I employed Foursquare API to get venues and category of venues by defining several functions so all these venues called via API will be stored in dataframe that can be analyzed. Foursquare output will return different number of venues each category wise with geo locations as well. This data is useful for advance analysis of each area or district, either venue or category wise. After acquiring this dataset, I will perform creative visualization using Folium, Waffle and WordCloud.

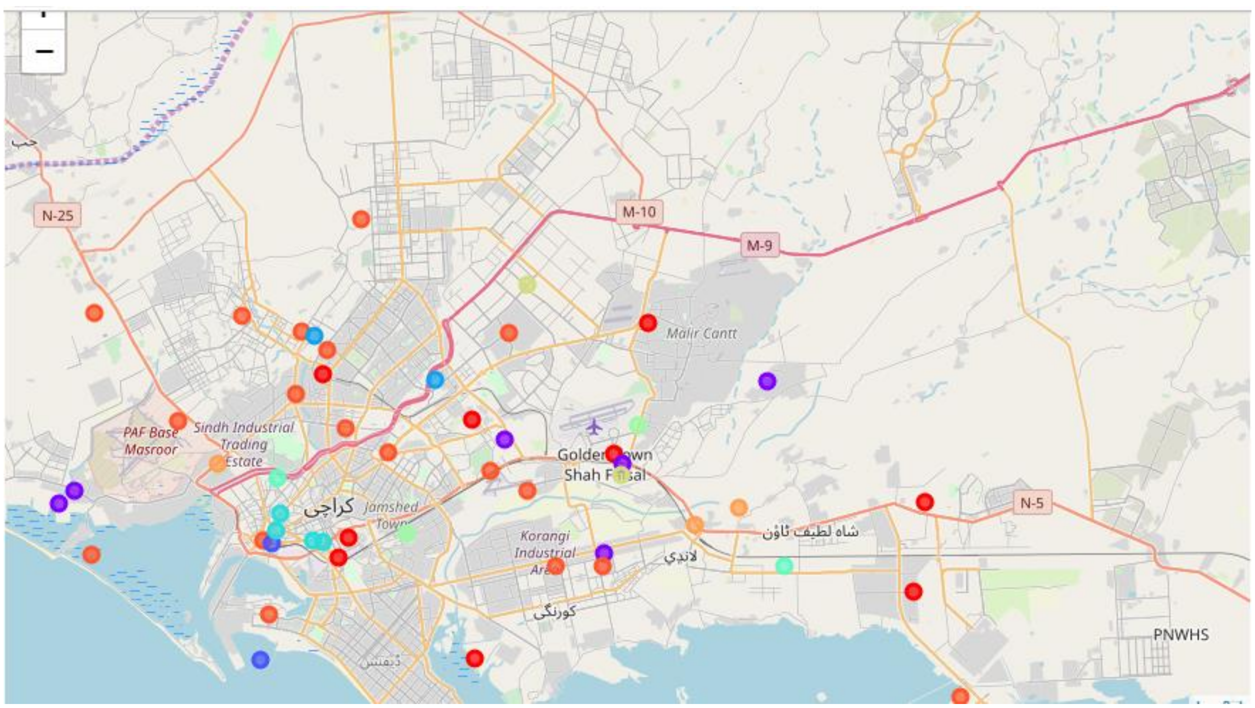
In the latter step, I will employ un-supervised machine learning algorithm: Density based Clustering – DBSCAN. To cluster area based on latitude, longitude, Population, Houses and Price. This will help me to understand which cluster with what frequency is right for which venue category and future potentials

RESULT

The most common Venue Category overall is Restaurant, Café and Market. Pizza and BBQ Joints are also spread across the city.

- District Central is dominated by Fast food and Bakery, having the third highest population and No. of houses.
- District Korangi is favored by Fast Food and Burger Joints with second highest domination in both population and No. of houses.
- District West has most beaches as venues and is the most populated district both in term of population and houses.
- District South has most market and Hotels and has opportunity for future developments for both houses and population.
- District East also dominated by Fast Food and Pizza and is moderately populated and has opportunity for future growth
- District Malir has most Fast Food and Farm and has opportunity for future developments for both houses and population.


Restaurant
Shop
Store
Place
Station
Airport
Park
BBQ
Café
Cream
Steakhouse
Sandwich
Train
Snack
Lounge
Auto
Dessert
Pizza
Coffee
Shopping
Donut
Persian
Venue
Multiplex
Middle
Sports
Chinese
Bakery
Soccer
Gym
Hotel
Social
Performing
Fast
Asian
Italian
Terminal
Juice
Mall
Afghan
Athletics
Home
Spot
Department
Bistro
Clothing
Diner
Electronics
Food
Bar
Ice
Indian
Club
Beach
Furniture
Market
Pharmacy
Bus
Eastern
Theater
Joint
Pakistani
Arts
Yoga
Burger
Japanese
Seafood
American
Fitness
Court
Center
Steakhouse
Event
Miscellaneous


N-25
M-10
M-9
N-5
Malir Cantt
PAF Base
Masroor
Sindh Industrial
Trading
Estate
Golden Town
Shah Faisal
کراچی
Jamshed
Town
Korangi
Industrial
Area
لانڈھی
شاہ لطیف ٹاؤن
کورنگی
PNWHS

Discussion

It is been evident that restaurant is the most common venue of the city. The result for each district and cluster wise provide insights that each district have different venue proportions and within each district there is marginally different proportions of opportunity and cluster distributions.

The limitations in this report is that forsquare has only provided data points of major chains and popular venues and real number of venues might be much greater, for further research google api can be used for much broader analysis. Moreover, the unavailability of data for the city is biggest limitation as most of the time consumed is to find and put together the data. This Project can be further enhanced by adding real time data collaborating several companies and government stake holders to check area wise real time monitoring of circulation of wealth, economics and potential business opportunity for any kind of business.

Conclusion

To conclude, Karachi's each district has its own essence in terms of venues and by combining their analysis with DBSCAN it is indicated that what will be the feasible opportunity based on current venues in a cluster.