**COURSERA CAPSTONE: MACHINE LEARNING FOR SCIENTIFICALLY LOCATING WASTE MANAGEMENT PLANT FOR THE CITY OF KOCHI**

## Introduction

Waste disposal and management is one of the biggest challenges of Kochi city (also written as Cochin occasionally), the financial capital of the state of Kerala in southern India. Kerala is well known for its tourist destinations and is also heralded as “God’s Own Country” among travel buffs. Kochi city, is a port city with a trading history that dates back to at least 600 years with China, Europe and Arabia. Kochi city, although a centre of commerce with its own IT park christened as InfoPark, offers affordable residential spaces as well as a plethora of tourism spots and hence more troublesome is waste management. International container transshipment terminal, Cochin Stock Exchange - the only stock exchange of Kerala state, hometown of 4th largest private sector bank in India and Doing Business Group 2009 report [1] recognized Kochi as one of the major industrial cities of India among 17 cities they considered - these are only few of the instances that portray the economic importance of Kochi city. Kochi City generates 380 tonnes of waste a day, out of which 150 tonnes are biodegradable and 100 tonnes are plastic waste [2]. The present waste management plant located at Brahmapuram, only 5 kilometers away from InfoPark- a neighborhood of Kochi city is currently damaging the environment, which was previously a lush green area and borders with two rivers- which supply drinking water to nearby regions. The discomforting stench of the biodegradable garbage and rain water stagnation in disposable plastic carry bags are serious threats as these contribute to a huge swarm of houseflies and excessive mosquito nuisance, which points to the risk of an epidemic like malaria or dengue. Residents of Brahmapuram and school-going students are under these potential risks while some of the residents have already fled their homes[2]. Groundwater contamination is yet another disaster in the making. Due to a recent fire break out at Brahmapuram plant, that disrupted the whole city by the toxic gas emanating from plastic waste, the government of Kerala has presently stopped collecting household plastic garbage and now needs to bring up a solution to deal with the city waste management problem. The Kerala State Pollution Control Board (KSPCB) advised the government that the energy-from-waste option is not so cost-effective [3] . In addition to this, the amount of waste remaining to be treated is rising.

1.1 Business problem

This project aims at finding an ideal spot for waste disposal and management for the city of Kochi, as the growing city will generate more waste, an ideal plant should be able to accommodate the input it receives and only then comes the problem of optimal way of treatment. The city’s waste management should not interrupt the social or economic matters of the city and must be isolated well. The stakeholders of this project are the residents of Brahmapuram and the government of Kerala. The impact of this project will not only enhance the living standards of people of Kochi city but it will save a lot of government funding in long term, protect the environment meanwhile its benefits, both social and economic, will stimulate other districts of Kerala to join the bandwagon of this scientific way of locating waste management plant and thus propel the state towards prosperity.

## Data

This project will make use of Foursquare API for two data collection purposes. One, to find the venues surrounding the waste management plant of one of the world’s cleanest cities and two, to generate a dataset that comprises the neighborhoods and venues in the city of Kochi, Ernakulam district, southern India.

East Calgary Facility location will be obtained from the *geopy* package and Foursquare API will be then used to mine the nearby venues within 5000 metres radius. Then the East Calgary Facility datapoint will be created by representing the facility using the venue categories of each venue mined in the form of average value of one-hot encoded score. The East Calgary Facility datapoint is shown in Figure.2.1 (due to space constraint only few columns are depicted):



Figure.2.1: East Calgary Facility datapoint

Kochi city will be defined using the neighborhoods of Ernakulam district and the neighborhoods will be webscraped using BeautifulSoup from the website ‘<https://worldpostalcode.com/india/kerala/ernakulam/>’. The neighborhood names scraped will be then used to mine the geolocation data of each of these using the *geopy* package. A small excerpt from the scraped data is shown in Figure.2.2. The webscraped dataset will be then processed to eliminate any missing values and then fed to Foursquare API to obtain nearby venues for each neighborhood. Similar to the East Calgary Facility datapoint, neighborhoods of Ernakulam will also be processed and few part of it is represented in Figure.2.3.

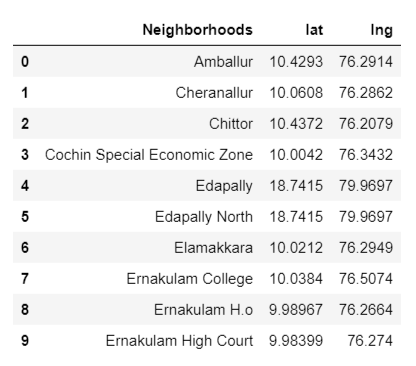


Figure.2.2: Webscraped data of Ernakulam’s neighborhoods

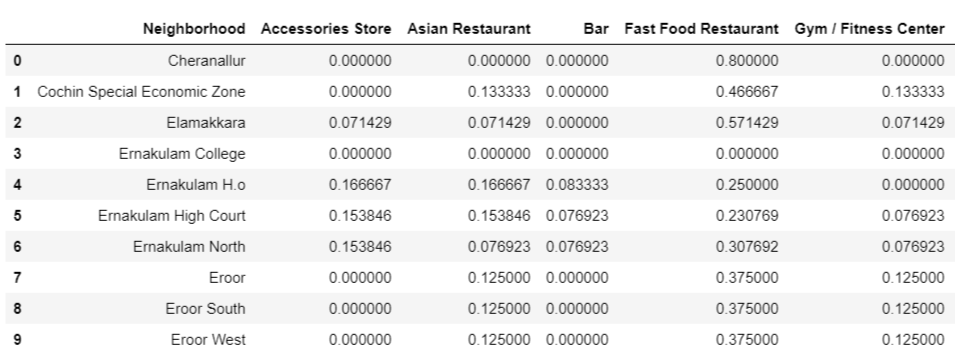


Figure.2.3: Ernakulam district neighborhood data

The East Calgary Facility datapoint will be concatenated with the Ernakulam district neighborhood dataframe and K-means clustering with 4 clusters will be performed to find the suitable location choices for Kochi city waste management.

## Methodology

The first step in this project is to find the best neighborhood that Kochi city would like to simulate. For that purpose, a simple Google search was done and it was found that the city of Calgary, Canada has consistently been ranked as one of the world's cleanest cities and it continues to be till 2020 [3]. It was found that Calgary city has two kinds of waste management facilities - one that deals with commercial waste and the other with industrial one. This study mainly focuses on commercial waste associated with Kochi city. The City of Calgary has two facilities that manage commercial waste - East Calgary and Spyhill. This study takes into consideration the East Calgary facility’s neighborhood based on simple convenience.

### 3.1. Tools used

1. Pandas - dataframe creation
2. Numpy - vectorized computation
3. Matplotlib - simple visualization, color palette selection
4. Requests - obtain webpages
5. Folium - Geospatial visualization
6. Sci-kit Learn - Machine learning
7. BeautifulSoup - Web-scraping
8. Geopy - mining geolocation data

### 3.2. Assumptions:

This study takes into consideration the following assumptions:

* East Calgary Facility location is an ideal one for waste management
* East Calgary Facility does not in any manner disrupt the surrounding economic or social activities
* This study assumes that geographical, political, climatic and community behavioural differences between the city of Calgary and Kochi are non-existent
* Population density and environmental hotspots of Kochi city are not considered

The entire project was then split into 6 steps:

1. **Generate East Calgary Facility datapoint** - Get East Calgary Facility location data and represent East Calgary landfill with respect to the surrounding venues using Foursquare explore option and generate a pandas dataframe of single row representing the East Calgary Facility. This is depicted in Figure.2.1
2. **Generate Ernakulam neighborhood dataset** - Firstly, the list of neighborhoods of Ernakulam is webscraped using *BeautifulSoup library* from the *bs4* package (Figure.2.2). Then using geopy, the location data for these neighborhoods are extracted. The neighborhood data is then fed to Foursquare API to mine 100 venues from each of the neighborhoods. This step will generate a pandas dataframe that represents Ernakulam district in terms of neighborhoods. This dataset is represented in Figure.2.3
3. **Prepare the clustering dataset** - Calgary and neighborhoods of Ernakulam will have differences in venue categories and hence the datasets must be pre-processed for clustering application. This preprocessing involves elimination of those neighborhoods with missing geolocation data and removal of non-overlapping venue categories to ensure the neighborhood of East Calgary and that of Ernakulam are adone on a common ground. After pre-processing, East Calgary Datapoint and Ernakulam neighborhood dataset are concatenated to develop the clustering dataset. Then clustering will be performed using Sci-kit learn’s implementation of K-Means API.
4. **Prepare the clustered dataset of Ernakulam** - The cluster labels generated in the previous step will be inserted into the clustering dataframe.
5. **Filter out the neighborhoods similar to East Calgary Facility** - the cluster label for East Calgary Facility will be determined first and using that label, neighborhoods of Ernakulam having the same label will be sliced from the dataframe.

## Result

Initially, Foursquare API explore endpoint option revealed 14 venues within 5000 metres of East Calgary Facility. The visualization of those 14 venues done using the *Folium* package is given in Figure 4.1.

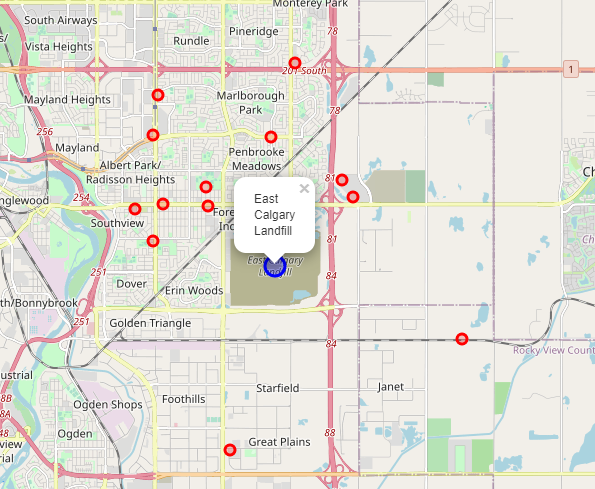


Figure.4.1: East Calgary nearby venues

Ernakulam neighborhood geolocation data extraction using *geopy* initially resulted in 52 venues but was reduced to 45 after elimination of some neighborhoods without geolocation data was ignored. Using these 45 neighborhoods of Ernakulam, Foursquare API was used to extract venues from each of them and yielded a pandas dataframe with 2730 venues in total. To generate the clustering dataset, both East Calgary datapoint as well Ernakulam neighborhood dataset had to be refined into having identical venue categories. This operation resulted in 326 venues in Ernakulam neighborhood dataset as there were only 10 common venues between the two places. The visualization of the refined Ernakulam neighborhood dataset is given in Figure.4.2.

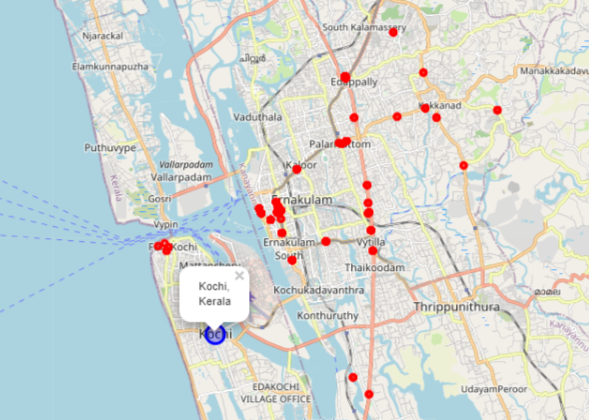


Figure.4.2. Some venues of Ernakulam neighborhood dataset

The two preprocessed datasets were combined to generate a clustering dataset. This then immediately followed an Elbow Curve analysis to obtain the optimal number of clusters to be considered. The plot for the analysis is given in Figure.4.3. As it became evident from the plot that cluster number beyond 4 does not cause significant reduction in segmentation loss, cluster count was fixed as 4.

Clustering was then performed and its outcome is depicted in Figure.4.4. A standard Folium marker is used to pinpoint the current waste management facility at Brahmapuram. Finding out to which cluster did East Calgary Facility fall was the next step and it was found to be the second cluster (labeled as 1). There were 16 neighborhoods of Ernakulam that showed similarity to that of East Calgary Facility.

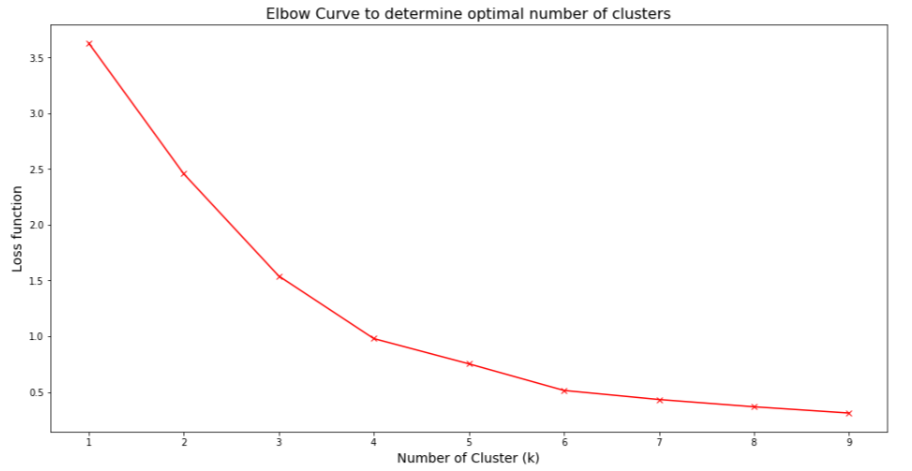


Figure.4.3: Elbow Curve plot

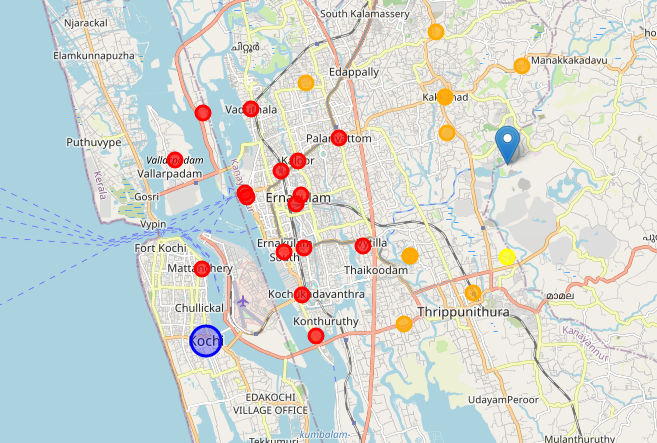


Figure.4.4: Clustered Ernakulam neighborhood dataset

A second level of filtering was done on the obtained number of neighborhoods of Ernakulam, by considering the public concerns of environmental damage, especially the contamination of water bodies [2]. The final set of choices are Palarivattom, Kaloor, Ernakulam North, Thammanam, Kadavanthara, Panampilly Nagar and Vyttila and these are depicted in Figure.4.5.

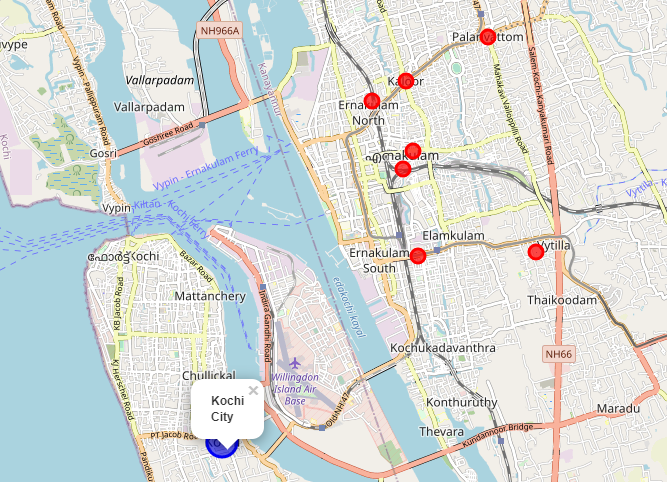


Figure.4.5: Final set of waste management locations

## Discussion

Kochi city is the financial capital of Kerala, southern India. Kochi city is facing a serious crisis in the area of city waste management and the current facility is being criticized by domain experts, environment as well as the residents of Brahmapuram for the unscientific location of the plant, the waste management practices and the catastrophes awaiting the state like ground-water contamination, climate change and damage to ecosystem. This study recognized the requirement of the city and implemented machine learning to extract the best possible locations to set up a new waste management plant- one that mitigates all the aforementioned problems. In short, this project demonstrates a scientific way to easily identify suitable locations for meeting city waste management needs using Foursquare API and simulating the neighborhood of one of the world’s cleanest city’s waste management plants, East Calgary Facility of the city of Calgary, Canada. This method thus minimizes the time required for city managers to narrow down possible locations by doing individual site visits which in turn, reduces initial plant set up costs and saves considerable time spent on making policies that govern the location and vicinity of waste management plants - the city managers can easily borrow the policies from one of the cleanest cities and tune it somewhat based on the requirements. And this solution can be replicated easily and is the greatest strength of this method- this means that this method can be implemented anywhere regardless of districts, states or countries. Hence, this solution will do justice to society, government and to nature. Now with this scientific method of locating the waste management plant, the authorities can focus all the efforts and resources onto identifying the best possible way of managing city waste - one step less towards making the world a better place.

The ML approach used in this project relies on the assumption that the geographical, political and social climate of both Calgary city and Kochi city are same and thus the regulations which operate the plant at Calgary city is expected to run the Brahmapuram facility of Kochi city with the same efficiency. This may not be that realistic. In addition to this, the population density and vegetation patterns of two cities vary and the final recommendation can change based on those considerations and the pro-environment regulations enforced at Kochi city. Another caveat is the use of Foursquare API. If the rules dictating erection of commercial or residential ventures nearby the waste management facility are not strict, Foursquare venues will be a poor indicator of any neighborhood. In short, the results of this study must be subjected to further filtering to arrive at an optimal solution.

## Conclusion

This study will help city managers and city planners to make the best decision regarding the first step towards city waste management - optimal location of the plant that does not interfere with social or economic aspects of the city. This endeavor will allow them to focus onto the more pressing issue - identify or research the best possible way to treat the city waste and hence mitigate the collateral damage done to nature and social life. The present study demonstrates the possibilities of Machine Learning based problem solving for better city administration.

6.1. Future Extensions

This study implemented K-Means clustering approach while Density Based Spatial Clustering for Applications with Noise (DBSCAN) methodology can be utilized which might yield different and improved results from this study. Future extensions can include a study conducted with different location data vendors or inclusion of population density or local government regulations or local geography. Another interesting aspect of this project is that this solution can be transformed into a solution for other domains, where users can decide their use-case and the application can suggest suitable localities which are served to users via a web-based or smartphone app based big data platform that can deliver even better results.

## References

1. Doing Business Group Report

<https://www.doingbusiness.org/en/reports/subnational-reports/india>

1. The problem of Brahmapuram-

<https://www.thenewsminute.com/article/walk-through-ghost-village-brahmapuram-deserted-thanks-kochis-garbage-41040>

1. Waste-to-energy not cost effective:

<https://www.thehindu.com/news/cities/Kochi/waste-to-energy-plant-may-not-be-most-cost-effective-option/article27190210.ece>

1. Top 10 cleanest cities in the world 2020:

<https://kenyaprime.com/top-10-cleanest-cities-in-the-world-2020-latest-ranking-things-you-need-to-know-about-the-cities-2/>