Battle of the Neighborhoods

Evaluation of Restaurant Locations

Joseph George Naduvathuserry

29 March 2021

IBM Applied Data Science Capstone

Contents

Tables

Figures

1 Business Problem

There are several factors to consider when entering the restaurant business, but location is without a doubt one of the most crucial. Whether a location is good or bad, however, depends on other aspects of the business – menu and cuisine, target customer segments, supply chain constraints, investment available, expansion plans and cost structure.

When the number of variables to consider is this large, an approach driven by data science makes sense. Even an experienced restaurant owner could benefit from the use of a similar approach, especially when expanding into new markets or new formats.

We can now define the generalized problem statement for this project as follows:

> Evaluate and quantify an area's suitability as a potential restaurant location, based on a selected set of weighted criteria.

The objective is to develop a framework that could be reused for similar problems, subject to the availability of relevant data. This project will focus on the city of Bangalore, located in Karnataka, India. Two example problems will be considered.

Table 1 Description of example cases

Parameter	Case 1	Case 2
Establishment type	Casual Dining	Restobar
Cuisines	Thai, Seafood	Alcoholic beverages, Burgers
Target demographic	Families, corporates, working professionals (30+ years)	College students, young working professionals (21-30 years)
Target pricing	₹ 600 per head	₹ 800 per head

Only the following attributes of a location will be considered in the scope of this project:

- **Population:** The most basic indicator of higher potential footfall.
- **Real estate prices:** Depending on the funding available, it might be prohibitively expensive to start a business in certain prime locations. Even if the investment is possible, profit margins will be affected.
- **Competition:** The restaurant business is highly competitive, and one has to compete to some extent with several different types and categories of establishment. Picking an already crowded area is a very high-risk move.
- **Complementary business:** Other non-restaurant establishments frequented by members of the target customer segment. Locations near these should experience higher footfall.

Detailed selection and weighting of these attributes will be on a case-by-case basis, and largely depends on the user's business plan, strategy and prior market or domain knowledge. The list of locations will be generated based on an evenly spaced grid spread across the selected area – for this project, the Bengaluru Metropolitan Area.

2 Data

The next step is to list the data that is required to meet the objectives set for this project. Potential sources were explored for the attributes listed below. Since the aim is to evaluate cells in a fairly fine grid across the city, localized data (neighborhood/administrative ward level) is required for this to be effective. Apart from this, other data quality factors considered included the recency of the data, and consistent availability across the Bengaluru area selected for this project.

- Geographic boundaries
- Population
- Real estate prices
- Restaurant data
- General venue data

2.1 Data Sources

After exploration and evaluation of several potential sources, the following were used in this project:

- **DataMeet GitHub Repository:** DataMeet is an Indian community of data science enthusiasts, and their repository hosts a selection of curated geospatial data, including a ward-level GeoJSON map of the Bengaluru metropolitan area (DataMeet, 2021). This file also includes the ward-level population and areas in the metadata. This information appears to originate from the 2011 Indian Census.
- **99Acres:** Cost per square foot for real estate in various localities in Bangalore (99Acres, 2021). The cost to buy per sq. ft. was the metric used in this project as it had the most comprehensive area coverage. Data was scraped from the webpage using the *BeautifulSoup* library for Python.
- **Foursquare – Places API:** Points of Interest (POI) data obtained using the search endpoint (FourSquare, 2021), with a selected list of category codes (FourSquare, 2021).
- **Geocoding (Forward & Reverse):** At various points in this project, a common requirement was the conversion of geographical coordinates into addresses and vice versa. This was achieved using two different services:
  - **Nominatim:** Free & open-source service (OpenStreetMap, 2021) with limited rates (1 request/second maximum).
  - **HERE Geocoding & Search:** Commercial service (HERE, 2021) with a freemium tier (250,000 free requests/month).

2.2 Limitations

Since this is not a commercial project, data sources were limited to free or open-sourced datasets and services. These represent potential areas for improvement in this project.

- **Income:** Data regarding median or mean income levels was not found at the level of localization required for this application (only city/district level data was available). Adding this would give us an additional important variable to rank our locations – depending on the restaurant target segment.

- **Demographics:** Again, the data is not localized enough to be useful in this case. This would also help in fine-tuning locations based on customer segmentation.
- **Detailed restaurant information:** The selected sources did not have the level of detail of sources such as Zomato (an Indian restaurant aggregator) or Google. At this time, Zomato does not appear to be allowing public access to their API, and Google's API rates are prohibitively high. With this data, the parameters for what restaurants are considered competition could be fine-tuned – with more detailed and consistent cuisines, costs, timings, ratings and even analysis of the menus if required

3 Methodology

This will be divided into three main sub-categories – data preparation, exploratory analysis and the final model.

3.1 Grid Generation

The first step in this analysis is to divide the area under consideration into an evenly spaced grid pattern. A hexagonal grid pattern was selected for this project for numerous reasons.

- **Area coverage:** Hexagons can form a perfect grid pattern, unlike circles.
- **Consistent neighbors:** Unlike a square grid, a hexagonal or honeycomb grid has consistent distance metrics even in diagonal distances, unlike a rectangular grid.
- **Location-based distortion:** A square or rectangular grid pattern is far more susceptible to distortion at different parts of the globe, due to inaccuracies in the estimation of the Earth's spherical surface as a flat 2-D plane.

Rather than develop a new system, his project makes extensive use of the H3 Geospatial Indexing system (Uber, 2021) developed by Uber, with a grid size of 8. The system divides the entire globe into a hierarchical hexagonal grid. Every unique set of geospatial coordinates can be mapped to a uniquely identifiable hexagon in several resolutions (Uber, 2021). These can also be converted to higher or lower resolutions as required by the project.

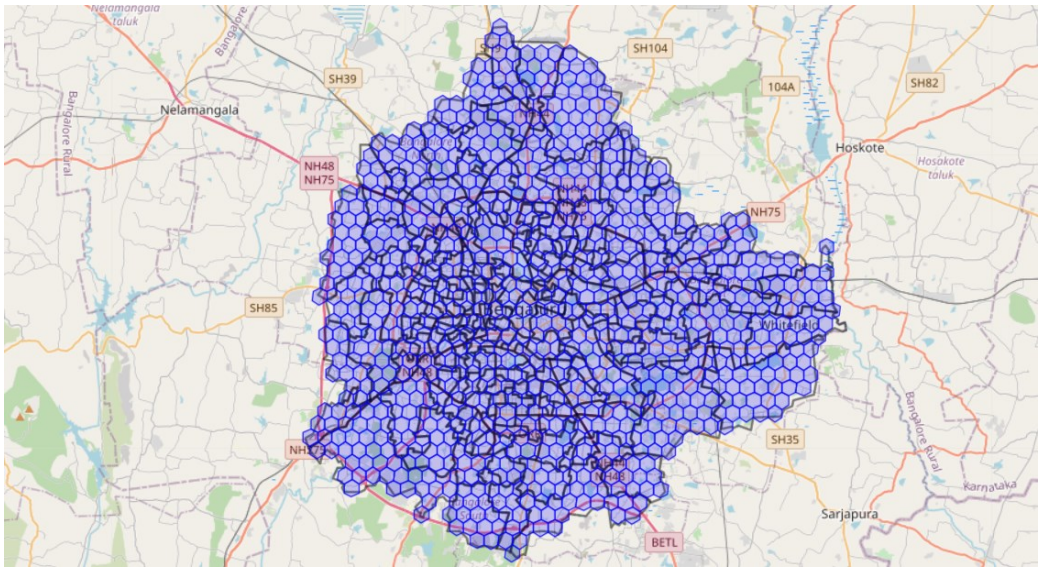

Figure 1 Generated H3 grid (resolution 8, 942 cells)

Table 2 Details of possible H3 resolutions

H3 Resolution	Average Hexagon Area (km^2)	Average Hexagon Edge Length (km)	Number of unique indexes
0	42,50,546.85	1,107.71	122
1	6,07,220.98	418.6760055	842
2	86,745.85	158.2446558	5,882
3	12,392.26	59.81085794	41,162
4	1,770.32	22.6063794	2,88,122
5	252.9033645	8.544408276	20,16,842
6	36.1290521	3.229482772	1,41,17,882
7	5.1612932	1.220629759	9,88,25,162
8	0.7373276	0.461354684	69,17,76,122
9	0.1053325	0.174375668	4,84,24,32,842
10	0.0150475	0.065907807	33,89,70,29,882
11	0.0021496	0.024910561	2,37,27,92,09,162
12	0.0003071	0.009415526	16,60,95,44,64,122
13	0.0000439	0.003559893	1,16,26,68,12,48,842
14	0.0000063	0.001348575	8,13,86,76,87,41,882
15	0.0000009	0.000509713	56,97,07,38,11,93,162

3.2 Geospatial Interpolation

Now that we have decided to map the evaluated locations to a hexagonal grid, we need to map our other data to the same grid pattern. Two different methods have been used in this project – areal interpolation for population and inverse distance weighting for real estate pricing. The difference is due to the data available in each case.

- **Inverse Distance Weighting:** In the standard Shepard's method (GIS Geography, 2020), each point is assigned a weight based on the inverse of the distance from a reference point. A modified version of the Shepard's method was used, only considering the nearest 5 neighbors for each point. The formula used is described below. The power factor *p* was set at 1 in this case.

$$u(\boldsymbol{x}) = \begin{cases} \dfrac{\sum_{i=1}^{N} w_i(\boldsymbol{x})\boldsymbol{u_i}}{\sum_{i=1}^{N} w_i(\boldsymbol{x})}, & if d(x, x_i) \neq 0 \; for \; all \; i \\ u_i, & if \; d(x, x_i) = 0 \; for \; some \; i \end{cases}$$

$$where \quad w_i(\boldsymbol{x}) = \frac{1}{d(x, x_i)^p}$$

  The results obtained show the expected distribution, with central locations noticeably more expensive than others.

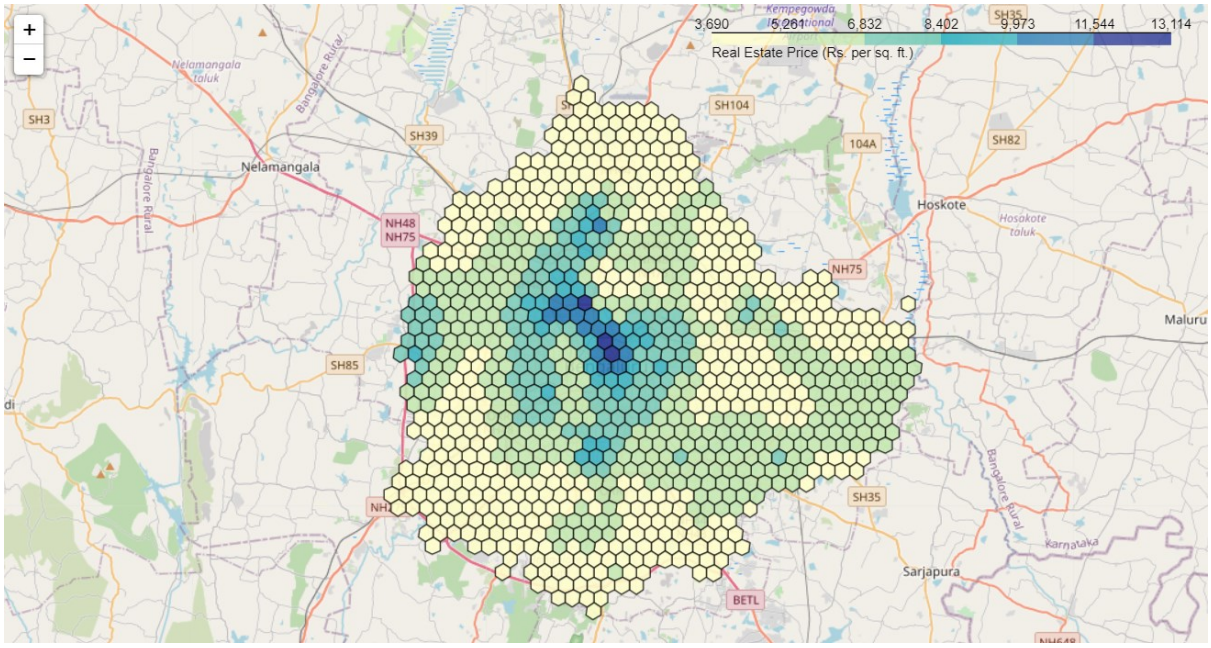

Figure 2 Real estate prices in Bangalore

- **Areal interpolation:** The wards are of different areas, some larger than and some smaller than our hexagonal grid size. This project requires the data to be mapped to the hexagonal cell level for further analysis. The *Tobler* Python package was used to calculate these interpolated values from the ward level data in the BBMP GeoJSON.

  The results obtained show a cluster of very high-density areas near the city center, which is expected. However, it must be noted that this data is almost a decade old, and the population and its distribution will be significantly different at present.

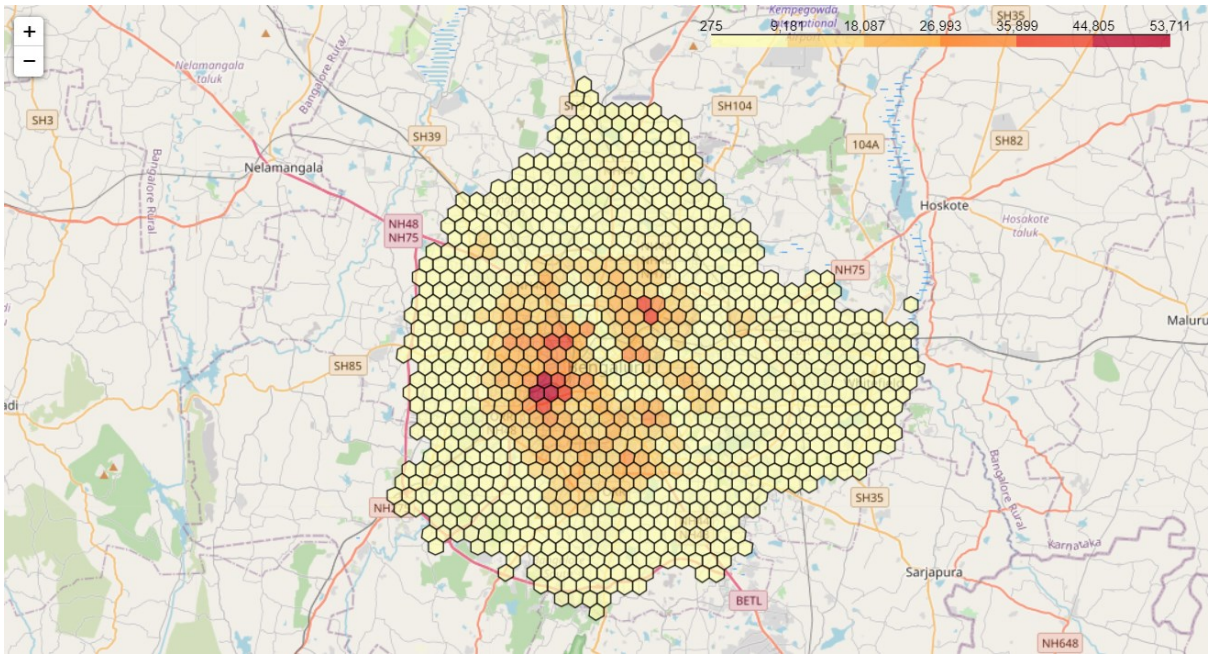

Figure 3 Population density in Bangalore

3.3 Clustering

Clustering is a very common technique in marketing where a large set of data points is grouped into a smaller number of distinct groups that can be targeted individually. As part of the exploratory data analysis, a K-Means unsupervised clustering model was used to cluster the locations into meaningful distinct categories.

The variables used for clustering the hexagons were:

- Estimated population of the hexagon
- Average real estate cost per square foot
- Count of Foursquare venues in each category (see Appendix Table 7)

The venue categories were one-hot encoded to separate variables, and then the entire set of variables was normalized using Scikit-Learn's MaxAbsScaler. This is an important step for K-Means clustering, since it is based on the Euclidean distance metric which requires similar variable scales to produce consistent results.

After evaluating several models using both the elbow method and silhouette scores (see Appendix Figure 8), the number of clusters was set at **4** for the final model. The clusters obtained using this model were color coded and plotted to a map. The properties of each cluster (Scaled frequencies of different venue categories) were also plotted in order to create a visual profile of each cluster.

3.4 Location Evaluation

Finally, a metric needed to be developed to provide a method of comparison between a given pair of locations. The metric would need to fulfil three criteria:

- Incorporate all the attributes required for the comparison into a single variable – population density, real estate costs, and complementary or competing businesses.
- Function well as an ordinal variable – the aim is to rank the set of possible locations and select the best suited, therefore the metric should provide sufficient differences to reliably create this ranked list.
- Be flexible enough to adjust for different types of establishment without affecting quality of measurement.

The implemented solution was to calculate a weighted average score on the three variables – population, real estate cost and frequency of a selected subset of venue categories (different for each use case). The weights used were negative in the case of detrimental factors (competition, costs) and the magnitude was based on the perceived relative importance for that particular scenario.

$$Hex\ score, S_h = w_p(P_h) + w_r(R_h) + \sum_{i=1}^{N} w_i(V_{ih})$$

$where$:
$w_p, w_r, w_i = weights\ for\ population, real\ estate\ cost, each\ indivudal\ category\ i$
$P_h, R_h = Normalized\ population\ and\ real\ estate\ cost\ per\ hex$
$V_{ih} = Normalized\ frequency\ for\ each\ of\ N\ venue\ categories\ in\ hex$

To also give some consideration to a slightly wider area around each hex, the scores of the immediate neighboring cells was also added to each hex score, with 20% weight for each neighbor. This aims to provide a more balanced scoring system.

Case 1

The weights selected for the first case (Thai casual dining restaurant - see Table 1) are listed below. The aim here was to locate high footfall areas with low presence of other restaurants (especially other Asian restaurants). In addition, areas close to residences and offices were preferred owing to the target customer base.

Table 3 Weights assigned for Case 1

Category	Variable	Weight
General	Population	10
	Real estate cost (per sq. ft.)	-15
Competition	Asian Restaurant	-10
	Indian Restaurant	-7
	Restaurant	-7
	Vegetarian / Vegan Restaurant	-5
	Quick Bites	-3
	Fast Food	-3
Complements	Residence	15
	Office	12
	Shopping Mall	10
	Movie Theater	10

Case 2

The weights assigned for the second case (Restobar targeting college students – see Table 1) are described below. Here, the main goal was to identify areas with potentially high footfall, with a high proportion of students and young working professionals – so colleges and office locations were targeted.

Table 4 Weights assigned for Case 2

Category	Variable	Weight
General	Population	10
	Real estate cost (per sq. ft.)	-15
Competition	Nightlife Spot	-15
Complements	Residence	5
	Office	10
	Shopping Mall	10
	Movie Theater	10
	College & University	15
	Arts & Entertainment	7

4 Results

4.1 Clustering

The K-Means clustering model (see Section 3.3) produced 4 clusters of reasonable sizes (133, 439, 134 and 169) with similar clusters obtained in multiple test iterations. This indicates a relatively stable and consistent model, although K-Means will always have some random variation.

Below are the cluster profiles obtained – which show the relative frequencies of the top 10 venue categories in each cluster (Standardized Z-Scores used for consistent scaling). This gives us a good indication of the types of venue common in each cluster and allows us to make an informed judgement on the likely demographics and common features of the cluster.

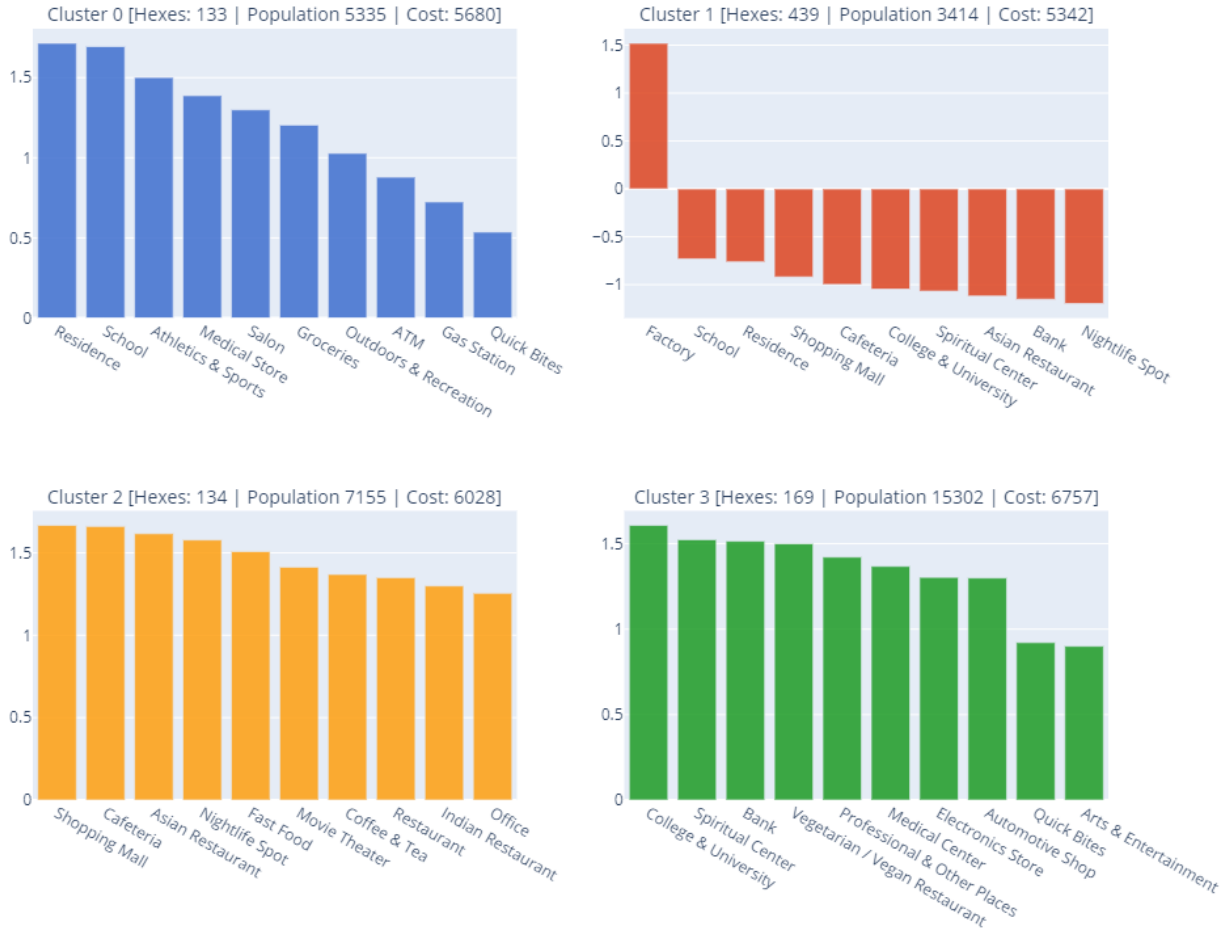

Figure 4 Profiles of the neighborhood clusters

These clusters were then plotted to the hexagonal grid map of Bangalore using Folium, to study the geographical distribution of these clusters.

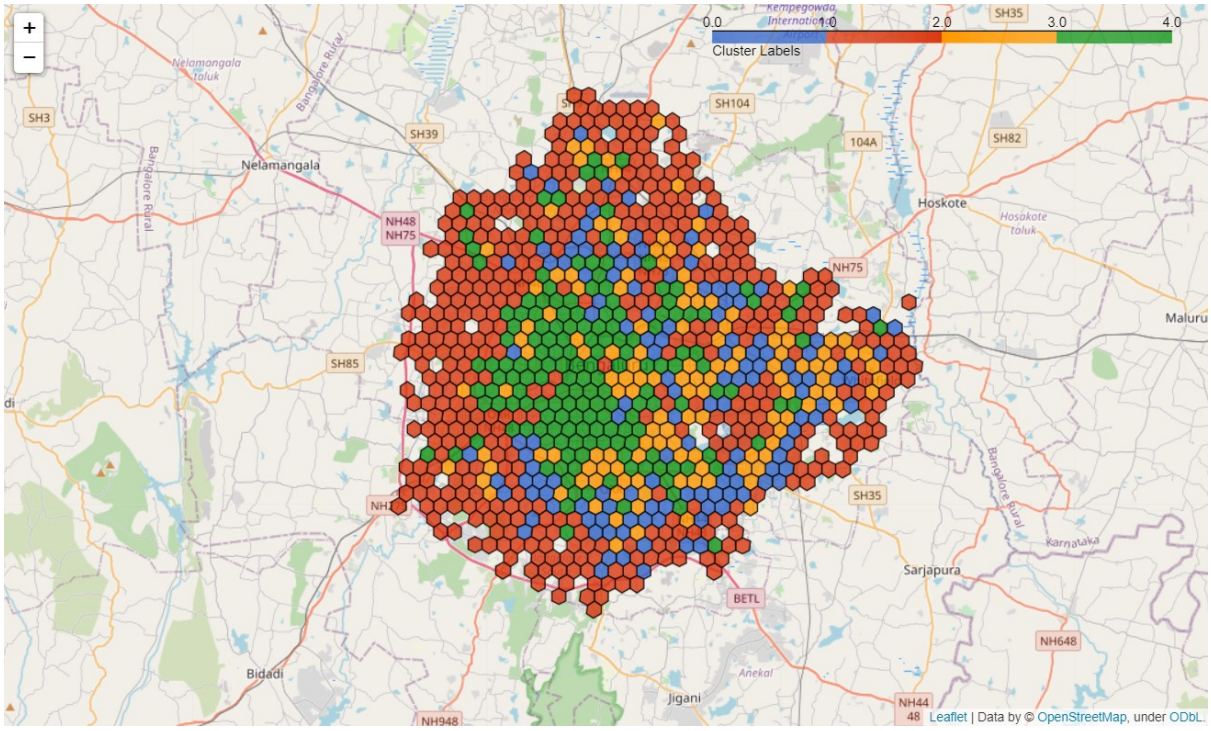

Figure 5 Clustered locations in Bangalore

Studying the results, we can describe the four clusters as follows:

- **Cluster 0: Residential.** Relatively less dense areas that mainly contain private residences, schools, parks and outdoor recreation, and similar venues.
- **Cluster 1: Industrial.** Mostly located in the outer parts of the city, these are less densely populated areas where most of the factories in the city are located.
- **Cluster 2: Shopping & Entertainment.** Expensive areas in the central and eastern part of the city containing shopping and entertainment venues.
- **Cluster 3: University/Commercial.** Mainly located near the city center, these are densely populated areas containing universities, banks, medical centers and places of worship.

4.2 Location Evaluation

For each case, the weighted scores for each location were calculated according to the previously described procedure (see Section 3.4). The locations were then ranked according to the scores, from highest to lowest, and this data was used to produce the following two outputs:

1. Choropleth map highlighting the areas with highest scores. The Choropleth color scale was tweaked manually – discrete steps were used rather than a continuous scale, and the steps were selected in order to highlight the best candidates (>99 and >95 percentile), and the secondary options for consideration (>85 percentile).
2. List of the top 10 potential locations for the particular scenario, along with addresses, H3 Hex ID and the location cluster obtained earlier (see Section 4.1)

Case 1

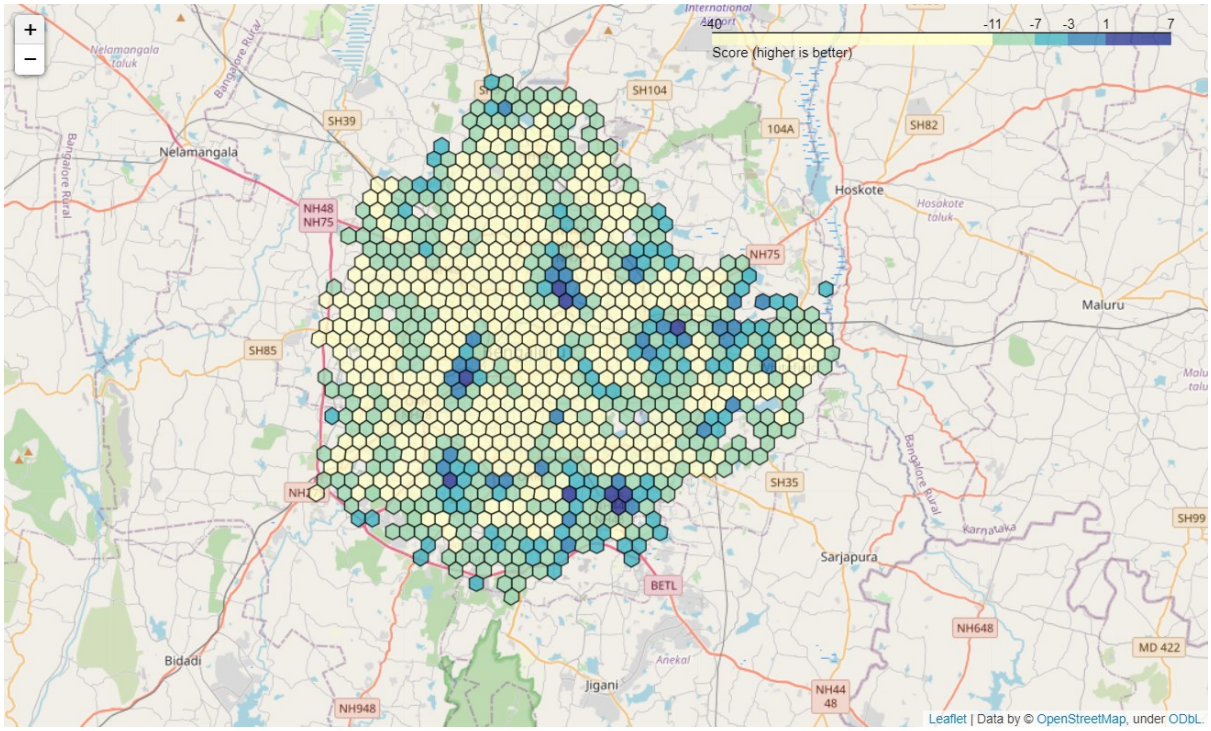

Figure 6 Location scores for Case 1

Table 5. Top 10 locations for Case 1

H3 ID	Cluster	Score	Address
8861892461fffff	0	7.27	Mangammanapalya, Bommanahalli Zone, Bengaluru - 560068
8861892463fffff	1	3.53	Mangammanapalya, Bommanahalli Zone, Bengaluru - 560068
8861892439fffff	0	2.96	Bommanahalli Ward, Bommanahalli Zone, Bengaluru - 560068
8861892eb1fffff	0	2.49	Richards Town, Sagayarapuram Ward, East Zone, Bengaluru - 560084
8861892eb7fffff	1	2.13	Muneshwara Nagar, East Zone, Bengaluru - 560084
8860145a2dfffff	1	1.44	K H Ranganath Colony, Rayapuram Ward, West Zone, Bengaluru - 560026
8861892e03fffff	0	1.34	Mapple Heights Apartments, Vijnana Nagar, Mahadevapura Zone, Bengaluru - 560093
886189246bfffff	0	0.97	AECS Layout, A block, Singasandra, Bommanahalli Zone, Bengaluru - 560068
88618924b3fffff	0	0.68	AGS Layout, Uttarahalli, Bommanahalli Zone, Bengaluru - 560061
8861892ebdfffff	3	0.64	Lingarajapuram, Lingarajapura Ward, East Zone, Bengaluru - 560043

Case 2

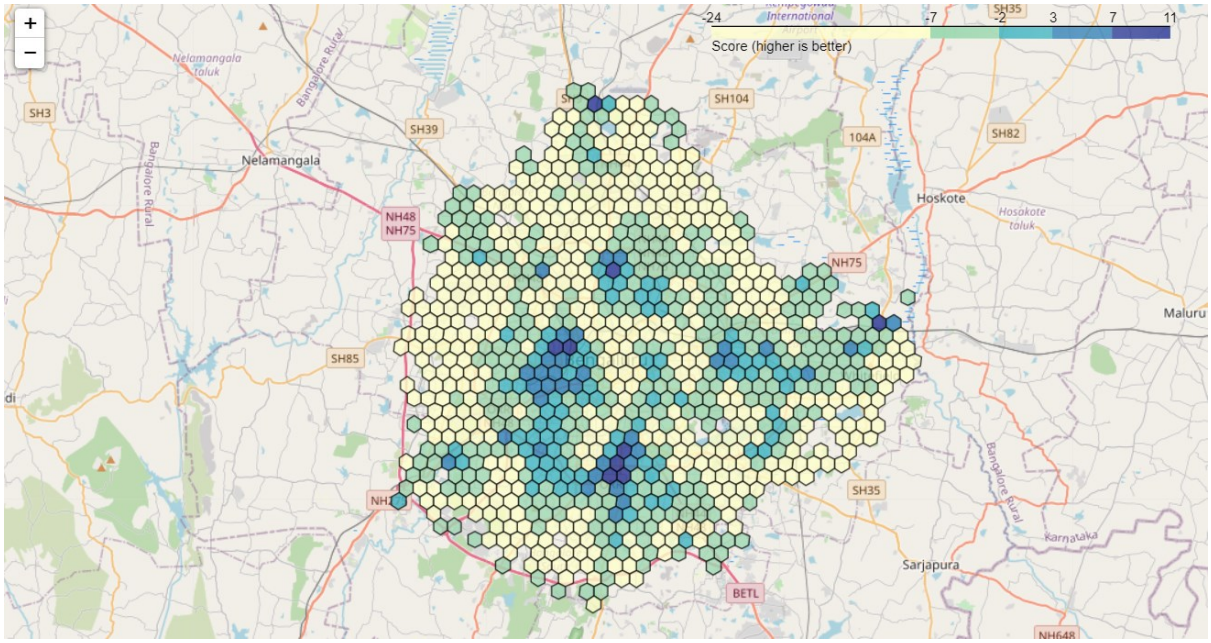

Figure 7 Location scores for Case 2

Table 6 Top 10 locations for Case 2

H3 ID	Cluster	Score	Address
8860145b1dfffff	3	11.47	Hospital-ayurvedic Homeo Clinic, 2nd Main Road, Prakash Nagar Ward, West Zone, Bengaluru - 560021
88618925d9fffff	2	8.63	MICO Layout, BTM Layout Ward, South Zone, Bengaluru - 560069
8861892185fffff	3	8.52	Kadugodi, Mahadevapura Zone, Sheegehalli, Bengaluru - 560067
8860145b0bfffff	3	8.24	Ramachandra Pura, Okalipuram Ward, West Zone, Bengaluru - 560020
88618925d5fffff	3	7.80	Bismillah Nagar, Gurappanapalya Ward, South Zone, Bengaluru - 560029
88618925dbfffff	3	7.69	JP Nagar 2nd Phase, South Zone, Bengaluru - 560069
8861892cd7fffff	0	7.54	4th Cross Road, Anandanagar, Hebbala Ward, East Zone, Bengaluru - 560032
8860169625fffff	1	7.21	Chowdeswari Ward, Yelahanka Zone, Bengaluru - 560064
88618925d1fffff	2	7.19	Sahakari Vidyakendra AHPS Jayanagar, East End B Main Road, NAL Layout, Jayanagar East Ward, Bengaluru - 560069
8860145b03fffff	3	7.08	7th Main Road, Sriramapura, Dayananda Nagar Ward, West Zone, Bengaluru - 560003

5 Discussion

The results obtained for each case display clear differences.The score values themselves are clearly not comparable – the weighting system used and the proportion of positive and negative weights will have a large impact on this. However, as an ordinal measure allowing us to rank different locations, it appears to meet the requirements.

Without diving into a detailed study of each location suggested by these results, we can use the cluster labels to verify that the results returned make logical sense. For the first case, six of the top 10 locations are residential areas – which agrees with the family restaurant parameter. The main shopping and entertainment centers are avoided, due to a combination of high costs and intense competition.

In the second case, six out of the top 10 locations are in Cluster 3, which was the cluster with the highest concentration of colleges and universities. Given the intended target audience, this also makes sense. A quick check also verifies that there are several university campuses, or in some cases office locations, located in and around these areas.

6 Conclusions

We can conclude that the procedure and metrics described here can function as a useful tool to evaluate and select locations for a new business venture. Quality of the results for different locations or businesses will be affected by a couple of factors:

- Availability and quality of data – geospatial, points of interest and demographics.
- Domain knowledge and careful selection of evaluation criteria and associated weights.

Provided these requirements are met, the method should be easily adapted to different use cases and scenarios.

7 Future Direction

There are several areas for improvement in this method, which were out of the scope of this project for various reasons.

1. Additional demographic data – obtaining age profiles, incomes and other demographic data at a finer geographical distribution level would allow for more evaluation parameters.
2. The venue categories used were aggregated into a list of 35 fairly high-level categories. More fine-tuned sub-categories could be used, provided the data source has this data consistently tagged and categorized.
3. Real estate listings online could be integrated into this, if sufficiently extensive sources can be obtained. Could be used to identify actual potential sites, with thresholds to filter out locations within the desired size or price range.
4. Additional venue information – especially restaurant information in this case. Data from a service such as Zomato is probably better suited for this purpose. They have an extensive coverage of restaurant data in India, and would provide additional parameters for evaluation such as restaurant prices, ratings etc.

8 References

99Acres. (2021, March 18). *Property Rates in Bangalore*. Retrieved from 99Acres: https://www.99acres.com/property-rates-and-price-trends-in-bangalore

Brath, R. (2015, October 15). *Equal Area Cartograms and Multivariate Labels*. Retrieved from Richard Brath: https://richardbrath.wordpress.com/2015/10/15/equal-area-cartograms-and-multivariate-labels/

DataMeet. (2021, March 16). Retrieved from DataMeet GitHub Repository: https://github.com/datameet/Municipal_Spatial_Data/blob/master/Bangalore/BBMP.GeoJSON

DeBelius, D. (2015, May 11). *Let's Tesselate: Hexagons For Tile Grid Maps*. Retrieved from NPR Visuals: https://blog.apps.npr.org/2015/05/11/hex-tile-maps.html

FourSquare. (2021, March 20). *FourSquare Places API*. Retrieved from FourSquare: https://developer.foursquare.com/docs/api-reference/venues/search/

FourSquare. (2021, March 20). *FourSquare Venue Category IDs*. Retrieved from FourSquare: https://developer.foursquare.com/docs/build-with-foursquare/categories/

GIS Geography. (2020, December 24). Retrieved from GIS Geography: https://gisgeography.com/inverse-distance-weighting-idw-interpolation/

HERE. (2021, March 18). *HERE Geocoding & Search*. Retrieved from HERE: https://www.here.com/platform/location-services/geocoding-and-search

OpenStreetMap. (2021, March 15). Retrieved from Nominatim: https://nominatim.org/

Uber. (2021, March 21). Retrieved from H3 Geospatial System: https://h3geo.org/

Uber. (2021, March 21). Retrieved from H3 Cell Dimensions: https://h3geo.org/docs/core-library/restable

Xu, T. (2020, June 20). *The Battle of the Neighborhoods — Open a Movie Theater in Montreal*. Retrieved from Towards Data Science: https://towardsdatascience.com/the-battle-of-the-neighborhoods-open-a-movie-theater-in-montreal-355cf5c679b8

9 Appendix

K-Means Clustering

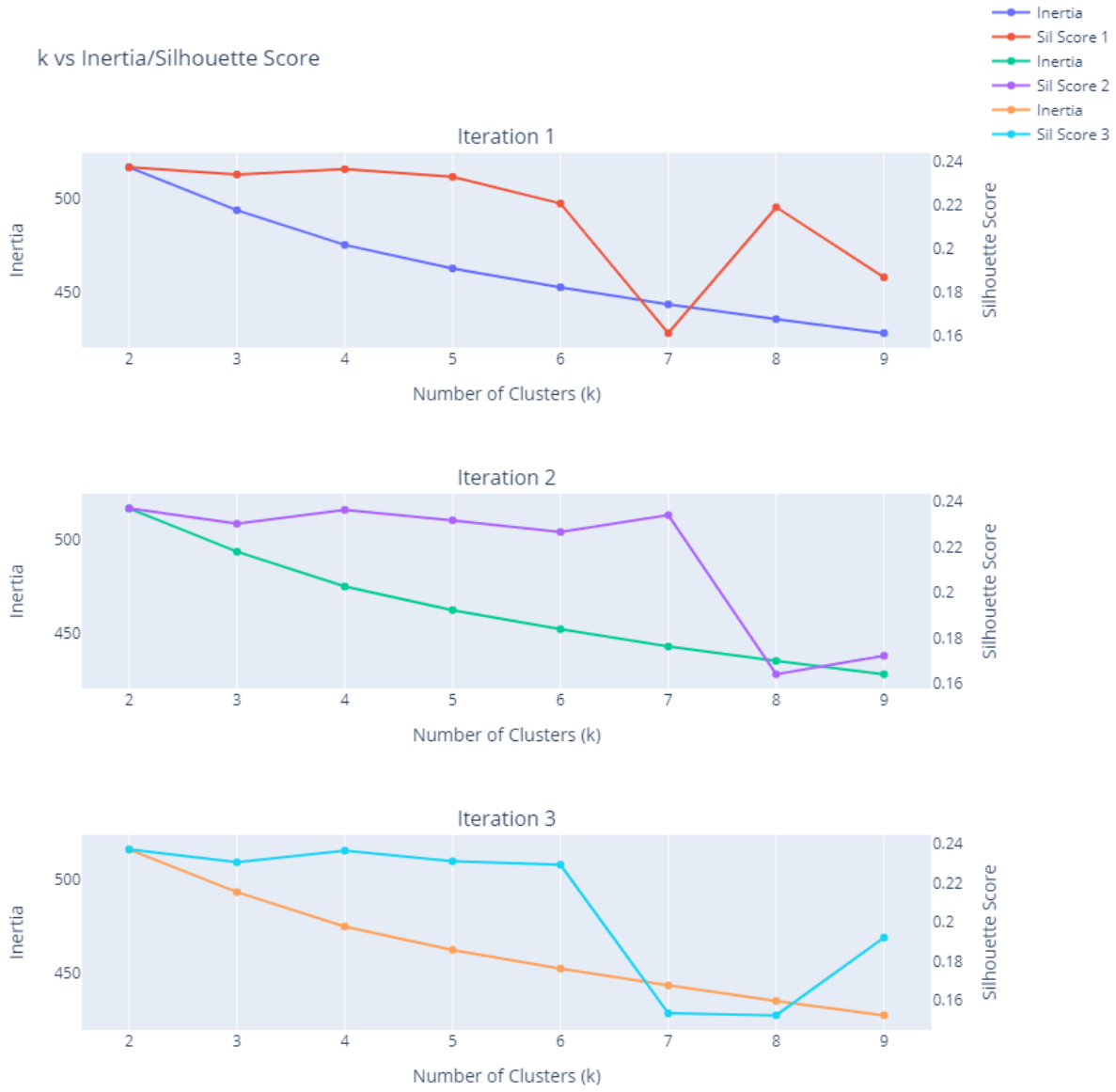

Figure 8 Inertia and silhouette score plots vs k, across three separate iterations

Foursquare Data

Table 7 Mapping of Foursquare category IDs to simplified category list

Category	Foursquare Categories
Movie Theater	Movie Theater
Asian Restaurant	Asian Restaurant
Indian Restaurant	Indian Restaurant
Vegetarian / Vegan Restaurant	Vegetarian / Vegan Restaurant
Bakery & Dessert	Bakery, Dessert Shop
Cafeteria	Cafeteria, Food Court
Fast Food	Fast Food Restaurant, Fried Chicken Joint, Burger Joint
Quick Bites	Food Truck, Snack Place
Coffee & Tea	Café, Coffee Shop, Tea Room, Juice Bar
Athletics & Sports	Athletics & Sports, Pool
Medical Center	Medical Center
School	School
Spiritual Center	Spiritual Center
Factory	Factory
Office	Office
ATM	ATM
Automotive Shop	Auto Dealership, Automotive Shop, Bike Shop, Car Wash, Motorcycle Shop
Groceries	Convenience Store, Department Store, Smoke Shop, Food & Drink Shop, Market
Bank	Bank
Electronics Store	Electronics Store, Mobile Phone Shop
Medical Store	Pharmacy, Optical Shop
Salon	Salon / Barbershop, Spa
Shopping Mall	Shopping Mall
Clothing & Jewelry	Clothing Store, Jewelry Store
Gas Station	Gas Station

Notes:

- For each of the Foursquare categories, all associated sub-categories were also included.
- In addition to this list, all the top-level categories provided by Foursquare were also included as is (without further sub-divisions)
- For the full list, see the complete Venue Categories list. (FourSquare, 2021)

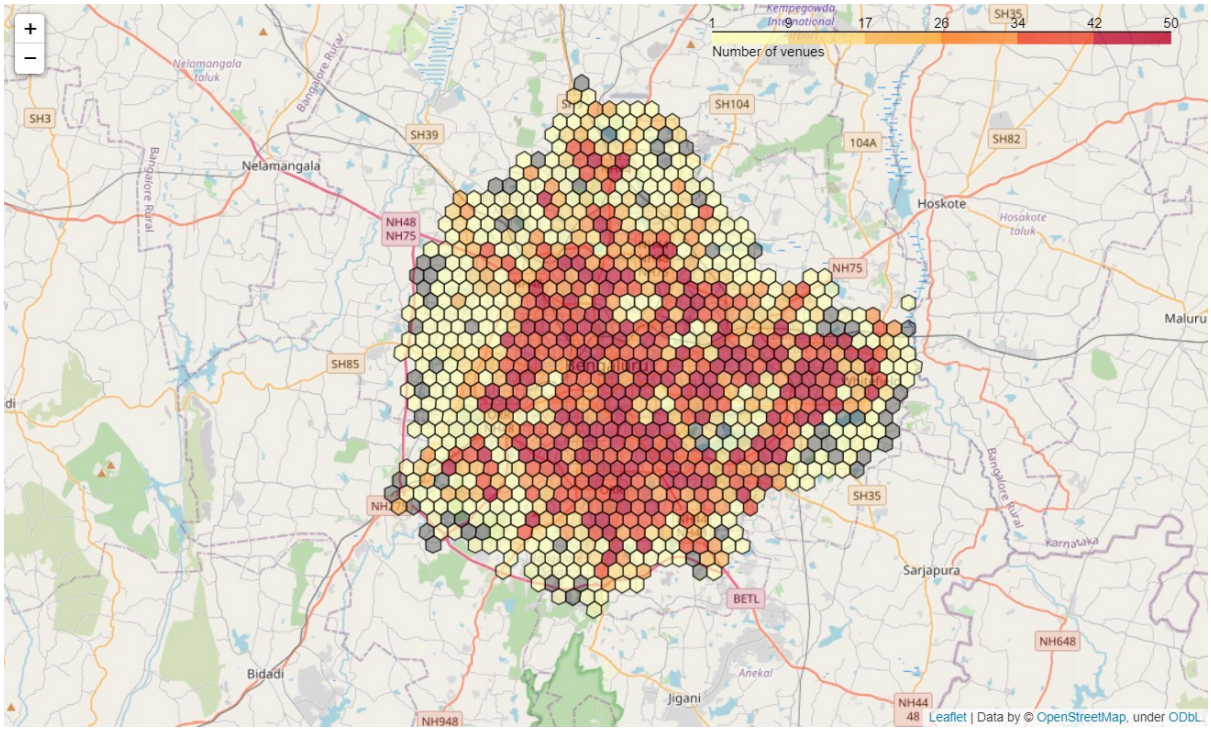

Figure 9 Distribution of Foursquare venues (all categories)