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 1.1 Description of sample 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.

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.

- **Areal interpolation:** The wards are of different areas, both larger than and less than our hexagonal grid size. Ward-level population data is mapped to hexagon level using the *Tobler* package for Python.
- **Inverse Distance Weighting:** This is a fairly simple weighted averaging process, where each point is weighted based on the inverse of the distance from a reference point. This was used for the real estate costs, since we had data at several unique points (geocoded) which needed to be generalized over a larger area.