MTH208 – Data Science Project

Indian Institute of Technology, Kanpur

Analysis of Chennai Plot Prices for High-Potential Investments

An Interactive Analytics Platform for Chennai's Real Estate Ecosystem

Submitted by: Team 2

Members: Ananya Ghosh
Atreyee Kayal
Praval Venkata Shiv Sai Jiddu
Vivek Chandwani

Course Instructor: Dr. Akash Anand

Semester: 2025–26 Odd

Contents

1 Introduction

1.1 Overview

Real estate prices in any metropolitan city are influenced by a complex mix of factors such as environmental safety, accessibility, infrastructure, and proximity to amenities. In a rapidly expanding city like **Chennai**, these factors vary significantly across localities, making it crucial to analyze them comprehensively to understand property value trends and urban development dynamics.

The project **"Analysis of Chennai Plot Prices for High-Potential Investments"** aims to bridge the gap between spatial data and decision-making by performing a detailed analysis of how environmental risks, infrastructure quality, and amenity density collectively impact land prices across the city. Through interactive tools and visualizations, the project allows users to explore relationships between locality characteristics and plot prices in a transparent, data-driven manner.

The analysis is divided into four interactive modules—**Flood Risk Visualizer**, **Feature Influence Explorer**, **Investment Predictor**, and **Infrastructure Need Analyzer**—each focusing on a specific aspect of urban property analysis. Together, they provide a holistic view of Chennai's real estate ecosystem, enabling smarter planning, investment, and policy formulation.

1.2 Use Cases and Importance

- **For Investors and Home Buyers:** The system helps identify *high-growth potential localities* and avoid areas prone to flooding or infrastructure shortages.
- **For Urban Planners and Developers:** The insights assist in locating *underserved regions* where new infrastructure or amenities can generate the most impact. It supports evidence-based planning and prioritization of development efforts.
- **For Government Agencies and Policymakers:** The project provides an analytical foundation for urban policy formulation and disaster resilience planning. By identifying flood-prone zones, infrastructure gaps, and potential growth corridors, authorities can better allocate resources, improve zoning decisions, and design targeted interventions for sustainable city development.

Overall, this project serves as a **comprehensive urban analytics platform** that transforms spatial, demographic, and infrastructural data into actionable insights. It demonstrates how **exploratory data analysis** and **geospatial visualization** can empower governments, planners, and citizens to make informed decisions.

2 Data collection

2.1 Summary of Data Sources

Data Source	Link	Ethics Check and Justification
Area names and plot prices (Scraped)	MagicBricks Property Rates Chennai	Data is publicly available for informational purposes. Only aggregated price trends are used; no personal or identifiable information is collected.
Area coordinates (API)	Nominatim OpenStreetMap API	Open-source geocoding service under the OpenStreetMap license. Used only to retrieve public geographic coordinates; no user data involved.
Nearby amenities (API)	Overpass API (OpenStreetMap)	Publicly accessible OpenStreetMap data; all information pertains to public infrastructure, posing no privacy or ethical concerns.
Flooded streets (GeoJSON Dataset)	OSM-IN Flood Map Repository	Open-source spatial dataset derived from public records. Used for research on flood-prone areas; contains no personal or confidential information.
Airport, railway proximity, population, and density (Scraped)	GeoIQ Places Database	Publicly available demographic and geospatial information. Data is used in aggregate form for analysis, ensuring no violation of privacy or data ownership.

2.2 Data Collection Process

1. **Localities and their Minimum, Maximum, and Average Prices:** The list of localities was directly web-scraped from the MagicBricks website. Plot prices were collected manually. Localities without available plot price data were excluded from the dataset.

2. **Coordinates of Each Locality:** Coordinates were obtained using the Nominatim OpenStreetMap API, incorporating a delay of 5 seconds between consecutive API calls to comply with rate limits. The API was repeatedly called for localities where data retrieval initially failed until no missing values remained. Coordinates for 12 localities were manually entered, as they could not be extracted from the mentioned data source.

3. **Population, Population Density, Area, and Distance to the Nearest Airport and Railway Station:** Data for these attributes was obtained by performing a web search for the corresponding GeoIQ webpage of each locality using SerpAPI (as the websites did not follow a consistent format), and subsequently scraping the required information. Data for 3 localities was manually entered, as it was unavailable from the specified source.

4. **Count of Various Amenities within 2 km and 5 km:** Using the locality coordinates obtained earlier and the Overpass API, counts were retrieved for 17 different amenities-considering radii of both 2 km and 5 km-resulting in a total of 34 columns generated for each locality.

3 Methodology and Key Visualizations

3.1 Module 1: Flood Risk Visualizer

This module was developed to identify and visualize flooded streets and flood-prone localities across Chennai using spatial data and interactive mapping.

Flood data in GeoJSON format and locality information from the final dataset were read using the `sf` and `dplyr` packages. Each locality was represented as a spatial point based on its latitude and longitude coordinates. The module employed the `leaflet` library to create an interactive map centered on Chennai, displaying the following features:

- **Flooded Streets:** Represented as blue polylines derived from the flood dataset.
- **Localities:** Depicted as circular markers-colored red for flood-prone (unsafe) and green for safe localities. A locality was classified as flood-prone if a flooded street was located within a 500-meter radius.

The module also incorporated a **search feature**, allowing users to query any locality by name. Upon search, the map automatically zooms to the corresponding location and displays a popup with its safety status.

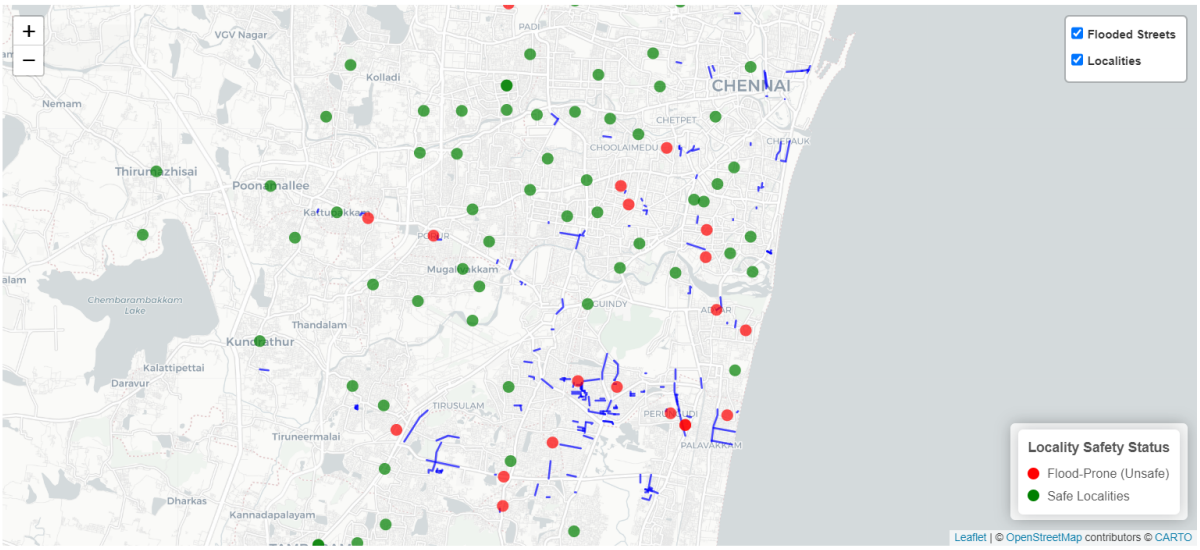

Figure 1: Interactive flood-risk visualization displaying safe and flood-prone localities in Chennai.

3.2 Module 2: Feature Influence Explorer

This module was developed to analyze and visualize the relative influence of various spatial, infrastructural, and demographic features on plot prices across Chennai using standardized linear regression analysis.

The `final_df.csv` dataset was first cleaned by removing redundant columns for linear regression such as `locality`, `min_price`, and `max_price`, `latitude`, `longitude`. Features representing similar amenities were group together producing consolidated metrics such as "Medical Facilities," "Schools & Kindergartens," "Food & Beverages," and "Colleges & Universities."

A **linear regression model** was then trained with standardized features to identify their impact on the dependent variable `avg_price`. The standardized coefficients were interpreted as influence scores—positive coefficients indicating features associated with higher prices and negative coefficients indicating those associated with lower prices.

The module provided **two visualization modes**:

- **Bar Chart View:** Displays features in descending order of absolute influence, with a diverging color scale—blue for positive and red for negative impacts.
- **Heatmap View:** Offers a compact overview where each feature's coefficient is represented as a color-coded cell with its numerical value annotated.

Users could **dynamically switch between 2 km and 5 km influence modes** and run the analysis interactively. The output also highlighted the **top five most influential features**, providing a quick interpretation of major drivers of land price variation.

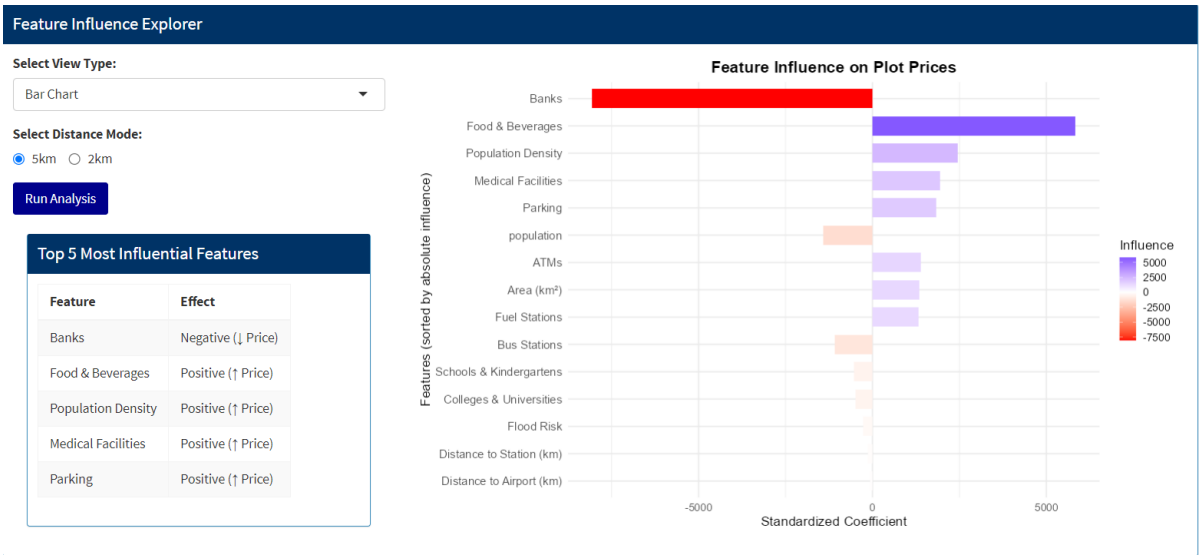

Figure 2: Visualization of feature influence on Chennai plot prices using standardized regression coefficients.

3.3 Module 3: Investment Predictor

This module was developed to identify high-growth and risky real-estate localities across Chennai by comparing the **predicted property prices** (based on model coefficients from Module 2) with their **actual average prices**. The analysis helps highlight undervalued areas with strong investment potential and overvalued areas that may depreciate.

Using the regression coefficients obtained from Module 2, predicted prices were computed for each locality. The difference between predicted and actual prices (`predicted_price - avg_price`) served as an indicator of investment potential:

- **High-Growth (Best Plots):** Localities where predicted prices exceeded actual prices, indicating undervaluation and future appreciation potential.
- **Depreciating (Risky Plots):** Localities where actual prices were higher than predicted, suggesting possible overvaluation.

An interactive interface allows users to select the type of analysis (high-growth or depreciating) and specify the number of top localities to display. The output visualization, created using `ggplot2` and `ggrepel`, presents a **slope chart** comparing actual and predicted prices for each locality. Upward blue slopes indicate undervalued areas with strong growth potential, while downward red slopes highlight risky, overvalued regions.

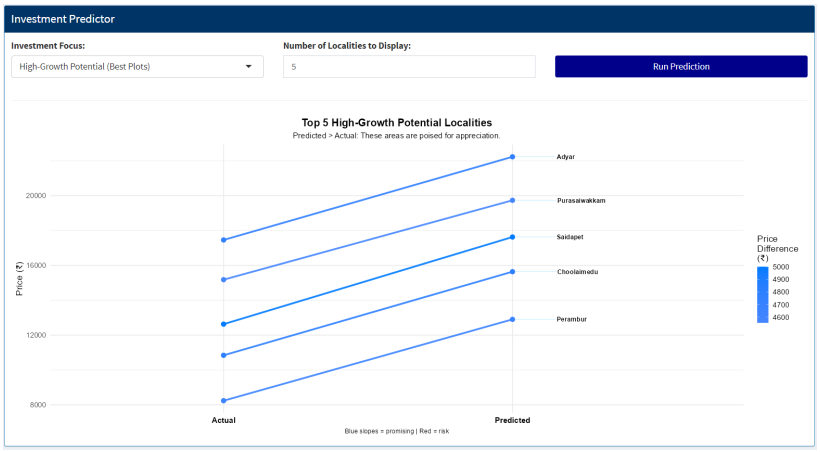

Figure 3: Top 5 high-growth or undervalued localities in Chennai.

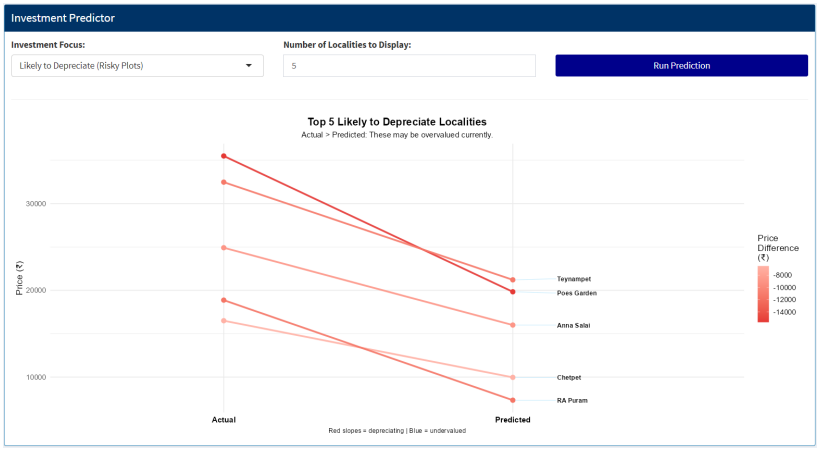

Figure 4: Top 5 depreciating or overvalued localities in Chennai.

3.4 Module 4: Infrastructure Need Analyzer

This module identifies clusters of localities with high demand for key facilities such as schools, hospitals, pharmacies, banks, ATMs, restaurants, cafes, and fuel stations. Users can dynamically select a facility type and visualization mode-either a spatial map or a comparative lollipop chart.

For the selected facility (e.g., `school`), a threshold equal to 25% of its average count is computed in 2km mode. Localities below this threshold are classified as underserved. These are then spatially clustered using the `DBSCAN` algorithm (3 km radius), grouping nearby underserved areas.

For each cluster, a **Need Score** defined as:

$$\text{Need Score} = \frac{\text{Total Population}}{\text{Average Facility Count} + 1}$$

is computed to quantify infrastructure demand intensity.

The results are visualized as:

- **Map View:** Color-coded clusters on a `leaflet` map, with centroids annotated by need scores and population.
- **Lollipop Chart View:** Cluster-wise comparison of need scores and population sizes.

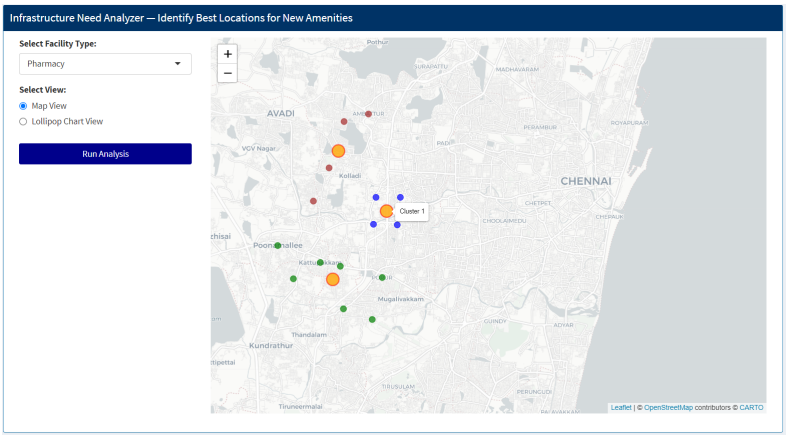

Figure 5: Map view of Pharmacy demand clusters highlighting underserved regions

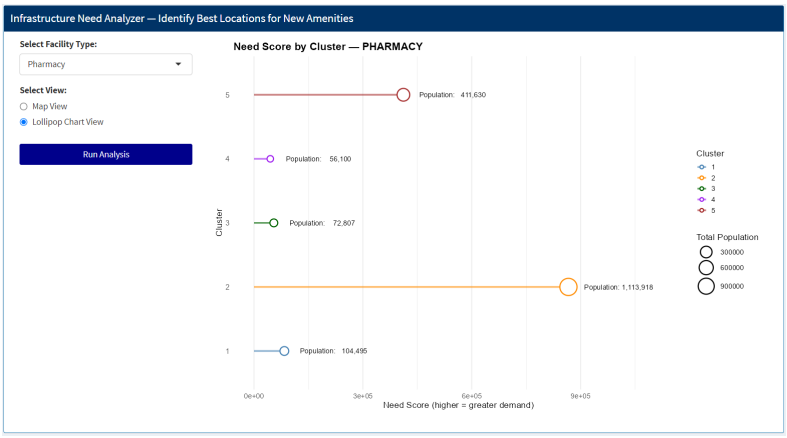

Figure 6: Lollipop Chart view of Pharmacy demand clusters highlighting underserved regions

4 Interesting Insights

1. **Population vs. Population Density Impact on Average Price:** The analysis reveals a positive correlation between average price and population density, but a negative correlation with total population-an observation that may initially appear counterintuitive. This can be explained by the spatial and demographic structure of the city. Densely populated areas, typically located near commercial or infrastructural centers, exhibit higher land values due to limited space and greater accessibility. In contrast, peripheral localities often cover larger geographic areas, resulting in higher total populations but lower densities. These regions generally have less developed infrastructure and lower market demand, contributing to the observed negative association between total population and average price.
2. **Impact of Closeness to Banks on Average Price:** Analysis in both 2 km and 5 km modes suggests that proximity to banks is highly negatively correlated with average plot price. This is because non-residential or high-traffic amenities cause noise, congestion, and land-use conflicts.
3. **High Growth-Potential Localities:** The analysis identifies Adyar, Purasaiwakkam, Saidapet, Choolaimedu, and Perambur as high growth-potential areas. Notably, Adyar emerges as the locality with the highest projected growth in Chennai, despite being classified as flood-prone, indicating that strong infrastructural development and market demand may outweigh environmental risks in influencing property appreciation.
4. **Likely to Depreciate Localities:** The analysis identifies Teynampet, Poes Garden, Anna Salai, Chetpet and RA Puram to be most likely to depreciate.

5 Instructions to Run the Code

5.1 Installing Requirements

1. Set the working directory as `MTH208_TEAM2`.
2. Run the following commands in R:

```
packages <- readLines("requirements.txt", warn = FALSE)
missing_pkgs <- setdiff(packages, rownames(installed.packages()))
if (length(missing_pkgs) > 0) {
  install.packages(missing_pkgs)
}
```

5.2 Running the Application

1. Set the working directory as the `MTH208_TEAM2` folder.
2. Run the following command in R:

```
shiny::runApp("app")
```

5.3 Replicating the Data Generation Process

1. Go to serpapi.com and obtain a SerpAPI key. (The free tier key is sufficient for our use.)
2. Alternatively, you may use one of the following API keys:
   - `a8d82d3ae00c72f83dc8d48d92797a2319d15a36b1d34a0a2f6d90d1553d1f5e`
   - `f7d8c1e03309a64c082032d11af1bf26d393468676fcb702818d7a1a07e9211e`
3. Set the working directory as the `MTH208_TEAM2` folder.
4. Run the following command in R:

   ```
   source("data/data_extraction_code.R")
   ```

5. When prompted, enter your SerpAPI key in the R console and press Enter.

After successful execution, a file named `created_df.csv` will be generated inside the `data` folder.

6 Limitations

1. The analysis assumes a linear relationship between average plot price and the selected features, which may not fully capture complex, nonlinear market dynamics.
2. The flooded streets data used in this study is from 2017. Current on-ground conditions may differ.
3. Minor discrepancies may exist in coordinate precision and calculated distances due to limitations in spatial data accuracy.
4. While identifying suitable locations for new infrastructure, the socio-economic conditions of each region were not explicitly considered. These could have been approximated using indicators such as the average plot price in the locality.

7 Ethical Considerations

All datasets used in this project were publicly available and applied only for academic and analytical purposes. No personal or sensitive information was collected or used at any stage. Care was taken to ensure that the models and visualizations were used responsibly - to explain patterns and support urban planning insights, not to predict or influence individual property decisions. The analyses reflect aggregate trends rather than exact real-world values, and results should be interpreted as informative indicators, not as precise forecasts. Efforts were made to avoid misleading interpretations and to maintain transparency in data sources, processing steps, and assumptions.