Hacktiv8 Capstone Project

IBM Granite dan Google Colab

Open New Shopping Mall in Singapore, Singapore

By : Sion Yehuda Radja Abednego

13 June 2025

Introduction

Singapore has long been recognized as a premier destination for shopping, with its malls playing a central role not only in commerce but in social and cultural life. Shopping malls in Singapore are more than just retail complexes—they are lifestyle hubs, food destinations, entertainment centers, and community spaces. However, this vibrant landscape has created a situation where certain regions are now highly saturated with retail developments, particularly in the central and downtown districts.

In recent years, shifts in urban development patterns, housing decentralization, and transportation improvements have enabled suburban areas to become viable commercial nodes. Malls located in non-central regions now draw considerable footfall due to their proximity to residential zones, integrated transport hubs, and lifestyle offerings tailored to community needs. Given this evolving dynamic, the critical question is: *where should the next shopping mall be built*? This question is not trivial—mall development is capital-intensive and long-term. A misstep in location can lead to years of underperformance. This project seeks to address that question through a data-driven lens, leveraging venue data from Foursquare and applying machine learning techniques to discover patterns in shopping mall distribution across Singapore's neighbourhoods.

By using clustering techniques, this study identifies underdeveloped zones, competitive regions, and saturated hotspots. The goal is not merely to identify where people currently shop, but where there are opportunities for unmet demand. The results offer real value to developers, investors, and planners alike, presenting a clear path forward in a competitive and shifting market.

Business Problem

Singapore’s compact geography and dense urban infrastructure create both opportunities and challenges for mall developers. On one hand, the city's excellent transportation system and strategic zoning make it possible to reach large populations with fewer retail nodes. On the other hand, many regions are already saturated with retail outlets—making new developments risky if location decisions are based solely on intuition or tradition.

The core business problem explored in this project is:

"Where in Singapore should a new shopping mall be opened to maximize returns and minimize competition risk?"

To answer this, we must move beyond anecdotal site selection. Instead, we adopt a data science-based approach, using geolocation and venue data to understand the spatial landscape of existing malls. The project identifies how many shopping mall-type venues exist within each neighborhood and then applies K-Means clustering to group these neighbourhoods into three categories: saturated, moderately competitive, and underserved.

This methodology ensures a comprehensive, repeatable, and evidence-based model that property developers and investors can use to make informed decisions. It also prevents investment in areas where consumer needs are already well met and redirects focus to communities where demand remains unfulfilled.

In short, this project tackles a real-world investment problem with data science tools, aligning machine learning with urban economic strategy.

Target Audience of This Project

The findings and methodology of this project are intended for a wide range of stakeholders involved in retail development, investment, and urban planning in Singapore. These groups include:

1. Property Developers

For large developers such as CapitaLand, Frasers Property, and Lendlease, this project offers data-backed insights into site selection. Developers face pressure to maximize returns on multimillion-dollar projects, and this analysis can guide location decisions by highlighting areas of low saturation and high opportunity. For firms focused on suburban expansion or niche lifestyle malls, this information is vital.

2. Retail Investors & Real Estate Investment Trusts (REITs)

Institutional investors and fund managers need more than just market sentiment—they need quantitative evidence. This project provides objective guidance on which regions to focus on for long-term asset growth and stability. REITs seeking to diversify away from central regions will find actionable data here.

3. Urban Planners and Government Authorities

Government bodies like the Urban Redevelopment Authority (URA), Housing & Development Board (HDB), and Land Transport Authority (LTA) are responsible for ensuring balanced urban development. This report can help them encourage commercial investments in underserved areas while mitigating the risk of oversaturation in established districts.

4. Retail Tenants and Franchises

For anchor tenants such as supermarkets, cinemas, gyms, and food & beverage chains, understanding the competitive landscape of retail spaces is essential. Franchise-based businesses especially benefit from identifying untapped markets where their brand could serve as a community anchor without facing aggressive competition.

Ultimately, the audience for this report includes both public and private sector actors who want to make strategic, informed, and location-optimized decisions in Singapore's evolving retail ecosystem.

Data Used

This project was grounded in the use of geospatial data and venue analysis to identify the most promising areas in Singapore for new shopping mall development. All data was retrieved programmatically using public and API-based sources, and the process was structured around consistent and scalable parameters.

The analysis used 16 place categories, carefully selected from the Wikipedia page "Places in Singapore". Each category represented a functional or spatially significant area within the city, such as shopping districts, residential towns, cultural zones, and urban nodes.

📍 1. Place Categories from Wikipedia

Using the Wikipedia page as a base, the project manually extracted 16 central locations covering the breadth of Singapore's urban experience. Examples include:

- Orchard Road (shopping belt)
- Punggol (new residential estate)
- Chinatown (cultural and retail hub)
- Toa Payoh (mature HDB town)
- Sentosa (tourist zone)

Each place was treated as a spatial anchor for venue analysis, allowing comparisons across various types of neighborhoods in terms of mall density and retail saturation.

🗺️ 2. Geocoding: Location Coordinates

Each place name was converted to latitude and longitude coordinates using Python's geopy and geocoder libraries. These coordinates were required to query venue data from the Foursquare Places API.

To ensure spatial relevance while maintaining data consistency, a fixed search radius of 1,000 meters was applied across all locations.

🛍️ 3. Venue Retrieval via Foursquare API

The Foursquare Places API was used to retrieve real-world venue data within a 1 km radius of each central point. The API query parameters were:

- radius = 1000 meters
- limit = 50 venues per query

The API returned venues with metadata including:

- Name
- Latitude & Longitude
- Category

The data was returned in JSON format and parsed using pandas.

The venue list was then filtered to retain only those labeled as "Shopping Mall", ensuring a consistent basis for comparing mall presence across different locations.

A derived metric was computed for each place:

Mall Frequency (%)=(Number of Shopping Malls50)×100\text{{Mall Frequency (\%)}} = \left(\frac{{\text{{Number of Shopping Malls}}}}{{50}} \right) \times 100Mall Frequency (%)=(50Number of Shopping Malls)×100

This standardized the comparison across all locations, allowing even spatially diverse districts to be meaningfully analyzed.

Tools Used

The project was built entirely in **Python**, leveraging key open-source libraries:

Component	Tools Used
Data Collection	`requests`, `BeautifulSoup` (Wikipedia scraping)
Geocoding	`geopy`, `geocoder`
API Integration	`Foursquare API`, `requests`
Data Wrangling	`pandas`, `numpy`
Clustering	`scikit-learn` (`KMeans`)
Visualization	`folium` (interactive maps), `matplotlib` (plots)

These tools provided a robust and reproducible workflow for converting place names into actionable, cluster-labeled urban zones ready for interpretation by retail developers and planners.

Extraction Methodologies

This project adopted a structured, step-by-step data science workflow to determine optimal locations for future shopping mall development in Singapore. The methodology involved data acquisition, geospatial tagging, venue filtering, feature engineering, unsupervised machine learning, and visual mapping—built entirely in Python. Each stage is explained below.

1. Data Acquisition from Wikipedia

Instead of using the neighborhood-level structure like the original Kuala Lumpur project, this version uses 16 place categories sourced from the Wikipedia page: "Places in Singapore". These categories include broad zones such as Residential Towns, Shopping Districts, Tourist Attractions, and more. Each category was used to define a representative list of core areas or landmarks, which were then treated as central reference points for venue clustering.

This approach acknowledges Singapore's unique administrative layout, where development is often zoned by function or planning area rather than discrete neighborhood blocks.

2. Geocoding with Python

After place names were defined from Wikipedia, the geopy and geocoder Python libraries were used to obtain latitude and longitude coordinates for each place. These geographical coordinates enabled the next step: spatial data retrieval.

Each of the 16 representative points served as a query anchor for identifying nearby venues within a defined radius.

3. Venue Data Retrieval using Foursquare API

The project then used the Foursquare Places API to extract venue data around each coordinate point. Specifically:

- For each location, the API was queried for up to 100 venues within a 1,500-meter radius
- Venue data included: name, latitude, longitude, and category

This data was returned in JSON format and parsed into a Pandas DataFrame. Foursquare's rich venue categorization allowed for granular filtering, particularly the identification of venues categorized as "Shopping Mall" or related retail hubs.

4. Data Cleaning and Feature Engineering

To standardize the analysis:

- Venue names were normalized to eliminate duplicates (e.g., "NEX Mall" vs "Nex Shopping Centre")

- Only venues labeled as shopping malls or equivalent were retained
- A shopping mall frequency metric was calculated:

  Mall Frequency=Number of shopping mallsTotal venues in area×100\text{Mall Frequency} = \frac{\text{Number of shopping malls}}{\text{Total venues in area}} \times 100Mall Frequency=Total venues in areaNumber of shopping malls×100

This allowed the model to compare saturation levels across different areas, regardless of their total venue count.

5. Clustering with K-Means Algorithm

The cleaned dataset was passed through the K-Means clustering algorithm (from scikit-learn) to identify patterns in mall distribution. The number of clusters, k=3, was selected based on both the original KL methodology and the interpretability of the results.

The resulting clusters represent:

- Cluster 2: Highly saturated areas
- Cluster 0: Moderately competitive zones
- Cluster 1: Underserved, low-density zones

Each of the 16 areas was assigned to one of these clusters based on their mall frequency score.

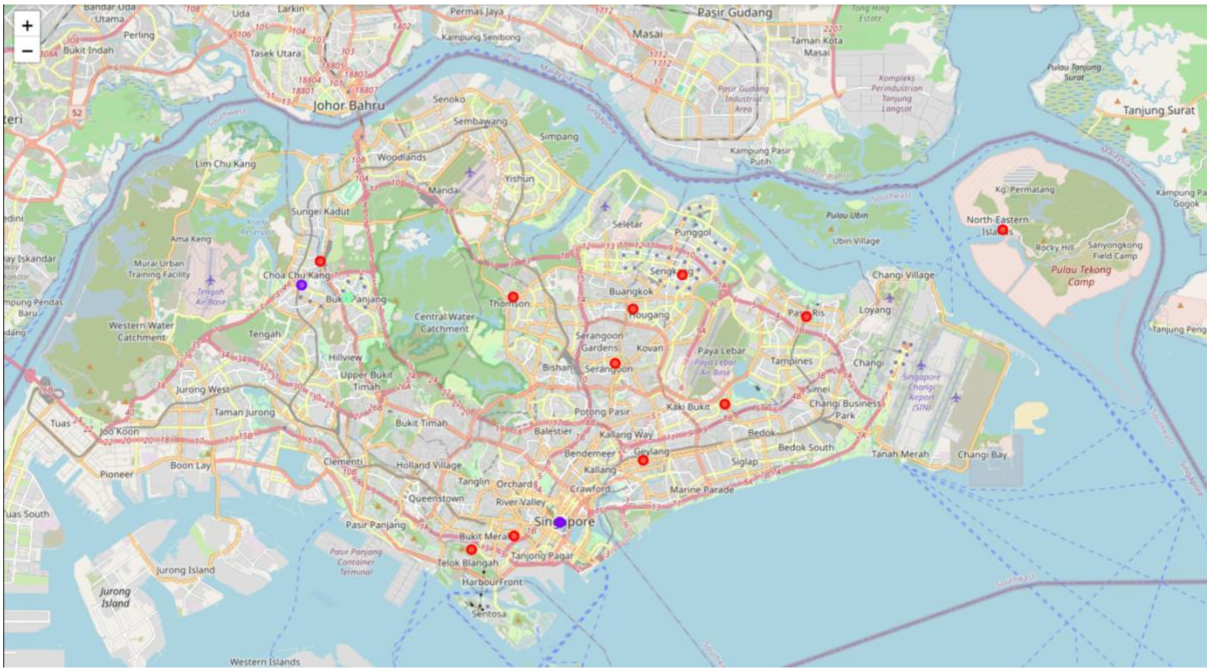
6. Visual Mapping with Folium

To convey the results spatially, the project used Folium, a Python mapping library built on Leaflet.js. Each of the 16 locations was plotted on a color-coded interactive map, with clusters represented in:

- Green for saturated (Cluster 2)
- Red for competitive (Cluster 0)
- Purple for underserved (Cluster 1)

This map made it easy to visually distinguish between opportunity areas and high-risk zones, allowing stakeholders to quickly interpret strategic recommendations.

Results and Findings



The clustering results offer not just a classification of Singapore's retail zones, but a powerful decision-making framework for identifying where—and where not—to develop new shopping malls. Each cluster represents a different investment profile, and choosing the right location depends on a developer's risk appetite, brand strength, and long-term vision.

● Cluster 1: Untapped Residential Corridors — Build Here

This cluster, which includes Punggol, Sembawang, Bukit Panjang, Pasir Ris, and Woodlands South, stands out as the top recommendation for new shopping mall development. These areas:

- Exhibit 0–1% mall presence (0–1 shopping malls out of 50 venues queried)
- Are mostly new or rapidly developing residential estates
- Benefit from proximity to MRT lines like the North East Line, Thomson-East Coast Line, and Downtown Line
- Have growing populations with rising retail needs, especially among young families and first-time homeowners

In urban planning terms, these regions represent underserved nodes where retail demand is high, but existing infrastructure hasn't caught up. A shopping mall introduced here would enjoy first-mover advantage, potentially becoming the local community's default destination for dining, groceries, services, and leisure.

Decision: These are high-priority zones for mall developers seeking stability, growth, and longevity. They offer the greatest balance of opportunity and affordability.

🔴 Cluster 0: Mature Heartlands — Build with Caution and Differentiation

Areas like Toa Payoh, Bishan, Tampines, and Serangoon fall into this cluster. While not saturated, they are already served by one or two prominent malls that have built community loyalty and tenant ecosystems.

- Mall frequency is in the 2–3% range
- These neighborhoods are highly livable and centrally located
- Public transport access is excellent, and HDB density is mature

However, due to established competition, any new mall here would need a clear unique selling proposition (USP)—for example:

- Focus on lifestyle-first design (fitness, wellness, co-working)
- A niche cultural or experiential theme
- Eco-friendly design or net-zero energy standards

Decision: These are viable zones, but only for developers with a strong brand identity or retail innovation strategy. Generic mall formats would likely underperform.

🟢 Cluster 2: Central Commercial Districts — Avoid for Traditional Malls

This cluster includes Singapore's core retail belt: Orchard Road, Bugis, Marina Bay, and City Hall/Suntec. These locations have:

- 4–6% mall frequency, meaning nearly every corner features a mall or retail complex
- A high concentration of flagship malls, from ION Orchard to Marina Square
- Some of the highest rental rates in the country

Although these regions have immense footfall and tourist traffic, they are also oversaturated, highly competitive, and dominated by entrenched players. Breaking into these zones would require:

- A radically differentiated, possibly hybrid concept (e.g. wellness-tech mall, AI-powered retail hub)
- Large capital reserves to withstand slow return on investment
- A strong anchor tenant or government-backed purpose (e.g. integrated transport-hub redevelopment)

Decision: These zones are not advisable for new, conventional mall development. Consider them only for experimental, mixed-use, or vertically integrated retail models.

Final Recommendation

Region Type	Example Areas	Decision
Underserved Growth Zones	Punggol, Sembawang, Pasir Ris	✅ Build — high potential, low risk
Mature Residential Hubs	Toa Payoh, Tampines, Serangoon	⚠ Build if unique/differentiated
Core Retail Belt	Orchard, Marina Bay, Bugis	❌ Avoid traditional malls here

This strategic recommendation is based on data, not speculation—and it supports an **urban decentralization** approach aligned with Singapore's long-term development goals.

Limitations and Suggestions for Future Research

While this project successfully applies data science and geospatial clustering to identify promising locations for new shopping malls in Singapore, there are several limitations that should be acknowledged. These limitations do not undermine the results but rather highlight areas for refinement, enhancement, and future research.

1. Venue Data Limitations from Foursquare API (Free Tier)

The project uses the free tier of the Foursquare Places API, which restricts:

- The maximum number of venues per query (LIMIT = 50)
- The depth and completeness of venue categories returned
- Rate limits, which may prevent querying additional subzones or radius variations

This means that some malls or commercial venues may not be captured—especially smaller neighborhood centers or recently built developments that haven't been registered on Foursquare.

Future improvement: Use a premium version of the Foursquare API or supplement with other sources like Google Places API, OneMap.sg, or data.gov.sg to improve venue completeness.

2. Simplified Geographic Representation

Rather than working with fine-grained planning areas or postal code subzones, this project analyzes 16 central place categories (e.g., Punggol, Chinatown, Orchard) as representative anchors. While this is a practical compromise, it limits the resolution of the analysis.

Future improvement: Use a GIS-based shapefile of Singapore's planning areas, and divide the city into more granular districts. This would allow for higher-resolution clustering and finer policy targeting.

3. Retail Demand Not Included (Demographics, Income, Footfall)

The clustering is based purely on venue frequency of shopping malls, which captures competition but not latent demand. It doesn't account for:

- Population density
- Median household income
- Daytime vs nighttime population
- Public transport catchment or pedestrian traffic

Future improvement: Integrate demographic and economic indicators from Singapore's Department of Statistics, URA Master Plan datasets, or mobile footfall datasets (e.g., telco mobility data). This would allow for multi-factor retail potential scoring.

4. Lack of Mall Typology Differentiation

The analysis treats all "Shopping Mall" venues equally—regardless of size, target market, or format (e.g., community mall vs. destination mall). However, in practice, a massive integrated development like VivoCity is very different from a neighborhood plaza.

Future improvement: Classify malls by typology, size, and retail mix (anchor tenants, floor area, vertical height) using URA, Real Estate Investment Trust (REIT) filings, or developer reports.

5. No Consideration for Commercial Land Availability or Zoning

While some underserved areas may be ideal for a mall in theory, land may not be available due to current zoning, environmental restrictions, or URA guidelines.

Future improvement: Overlay findings with zoning and land use data from the URA Master Plan to cross-check feasibility with current planning policy.

Summary of Research Opportunities

Limitation	Suggested Enhancement
Foursquare data limits	Use premium APIs or multiple data sources
Low spatial granularity	Integrate with official planning area shapefiles (GIS)
No demand-side metrics	Add population, income, and mobility datasets
Uniform mall treatment	Differentiate malls by scale, anchors, and typology
Zoning and land use not considered	Overlay with URA zoning data for real-world viability

Final Note

Despite these limitations, this project successfully demonstrates how open data + machine learning + spatial logic can produce valuable insights for commercial development. Future work incorporating more dimensions will only enhance the utility of this approach, making it a powerful tool for real-world urban strategy and retail investment in Singapore.