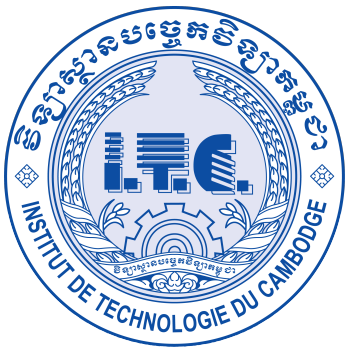

Optimal Cafe Location Analysis

Programming For Data Science

3rd year Engineer's Degree in Data Science
Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia

Name of Students	**ID of Students**	**Score**
KHUN Sithanut	*e*20211527	.....
PAV Limseng	*e*20211548	.....
PEANG Rattanak	*e*20210072	.....
PEL Bunkhloem	*e*20201314	.....
PEN Virak	*e*20211572	.....

Submission Date: 27, 06, 2024

Lecturer: Mr. OL Say (Course)

Lecturer: Mr. PHANN Raksmey (TP)

Academic year
2023-2024

Contents

1 Introduction

In the burgeoning cityscape of Phnom Penh, the strategic placement of business establishments, particularly cafes, plays a pivotal role in their success. As urban development and educational infrastructure grow, so does the opportunity to serve densely populated areas with significant foot traffic and customer engagement. This project leverages data science and machine learning, specifically k-means clustering, to pinpoint the optimal location for a new cafe that aims to cater primarily to a niche yet expansive market: the educational community within Phnom Penh.

The rationale behind targeting educational buildings as primary customers stems from their consistent population density and the daily presence of a diverse group of potential patrons, including students, faculty, and administrative staff. This demographic is known for its regular consumption of coffee and related products, primarily due to the long hours spent in educational pursuits and social engagement activities that are central to academic life.

This report begins by outlining the objectives of the project, which include the application of k-means clustering to identify clusters of educational buildings that are underserved by existing cafe offerings. By applying the elbow method, we aim to determine the most statistically appropriate number of clusters, thus identifying key areas in Phnom Penh where a new cafe could thrive.

Further, the importance of choosing the right location cannot be overstated, as it directly influences the cafe's accessibility, visibility, and profitability. A strategically located cafe not only ensures high customer retention but also enhances the overall viability of the business. The analysis presented in this report is based on a comprehensive dataset of educational buildings in Phnom Penh, which includes universities, colleges, and other academic institutions. This dataset provides a foundation for our clustering algorithm, which in turn informs our decision-making process for selecting an optimal cafe location.

In subsequent sections, we will delve into the methodology employed in collecting and preprocessing the data, the specifics of the k-means clustering process, and the application of the elbow method. The results will be thoroughly analyzed to offer insights into the potential locations and their respective characteristics. Finally, the discussion and conclusion will synthesize these findings and propose actionable recommendations for potential entrepreneurs looking to capitalize on this identified market opportunity.

By aligning the data-driven approach with strategic business planning, this report aims to bridge the gap between theoretical data science applications and practical business solutions, providing a blueprint for successful cafe placement that benefits both the business owner and the community it serves.

2.Methodology

2.1.Data Source

The primary source of data for this project is Google Maps. We utilized Google Maps to scrape data regarding the locations of various educational institutions and other places frequently visited by students in Phnom Penh. This data collection was facilitated by the use of the Google Maps API, which provided detailed information including coordinates (latitude and longitude), names of institutions, addresses, and types of places. The types of places we focused on include:

- Schools (kindergarten, primary school, secondary school)
- Universities
- Libraries
- Bookstores
- Museums
- Training centers

- Academies
- Colleges

These categories were selected based on their high likelihood of being frequented by students, our primary target customers for the café business.

1.1 Data Collection

To identify the optimal locations for opening a new café targeting students, we conducted a comprehensive data scraping operation using the Google Maps API. Our data collection efforts were concentrated in Phnom Penh, with a particular focus on a 10km radius around the central point of Stueng Mean Chey. This central location was chosen due to its strategic position within the city and its accessibility to various educational institutions.

Figure 1: Area for Data Collection on Google Map.

After scrapping process, we got 183 locations of our main target educational buildings. To facilitate further analysis, we saved the collected data into an Excel file, providing a structured format for easier manipulation and examination.

	Name	Address	Latitude	Longitude	Type
0	CJCC (Cambodia-Japan Cooperation Center)	Rupp-CJCC, មហាវិថី សហាព័ន្ធរុស្ស៊ី (990), ភ្នំពេញ	11.568929	104.893694	school, point_of_interest, establishment
1	Angkor Computer Center	#95E0, Saint 164, Phnom Penh	11.563684	104.912819	school, point_of_interest, establishment
2	SAS Santhormuk - Stanford American School	#197, St.146, Teuk laok 2, Phnom Penh	11.564989	104.899767	school, primary_school, secondary_school, poin...
3	Sovannaphumi School, Tep Phan Campus	6A Oknha Tep Phan St. (182), Phnom Penh	11.563052	104.900276	school, point_of_interest, establishment
4	Aii Language Center (Aii), Mao Tse Tong (QLH B...	217 ABCD Mao Tse Tong Blvd, ភ្នំពេញ	11.546374	104.907986	school, point_of_interest, establishment

Figure 2: Dataset.

1.2 Data Preprocessing

The data preprocessing phase was essential to ensure the quality and usability of the collected data for subsequent analysis. We began by fetching data for the places of interest (educational institutions and other student-frequented locations) using the Google Maps API. This data, initially viewed in an HTML file to verify correct scraping, included 183 addresses with corresponding coordinates, names, and types of institutions. Utilizing the Folium library, we plotted all addresses on a map using their latitude and longitude coordinates, creating a visual representation of the locations in Phnom Penh. This initial mapping helped us understand the spatial distribution and identify any potential anomalies or outliers.

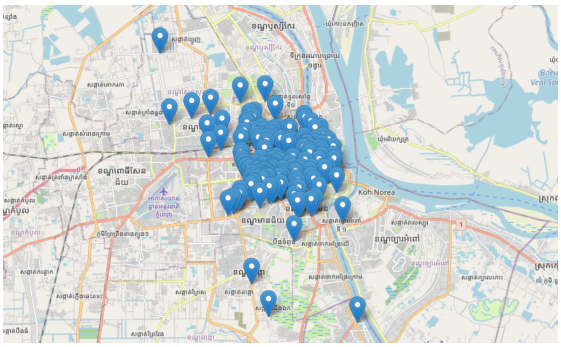

Figure 3: All Educational buildings Location.

Upon visual inspection, we noticed some addresses were significantly distant from the main cluster, indicating potential data inaccuracies or less relevant locations. We systematically examined the spatial distribution to identify these outliers, checking the distances of each locations from the central cluster. We identified 8 locations as outliers (figure 4) based on their significant distance from the main cluster and plotted them again using Folium to visually confirm their status.

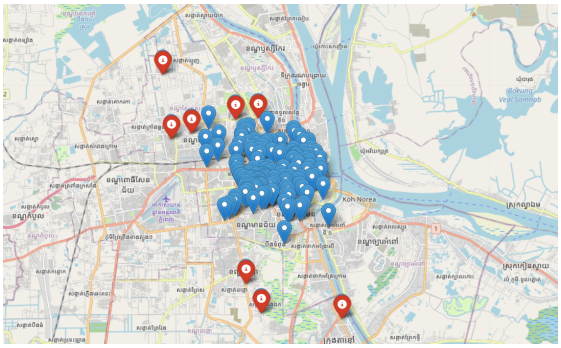

Figure 4: Outlier Locations(Red marker).

After confirming these outliers, we excluded them from our dataset to maintain the integrity and relevance of our analysis, focusing on core areas where our target demographic is concentrated. After removing the 8 outliers, we were left with a refined dataset of 175 addresses, which was then used for further analysis and the application of the K-means clustering algorithm. By meticulously preprocessing the data, we ensured our analysis was based on accurate and relevant information, crucial for effectively determining the optimal locations for opening new cafés targeting the student population in Phnom Penh.

1.3 Description of K-Means Clustering

In our analysis to determine optimal locations for opening new cafés targeting students in Phnom Penh, we employed the K-Means Clustering algorithm. K-Means is an unsupervised machine learning algorithm that partitions a dataset into K distinct, non-overlapping clusters based on feature similarity. Each cluster is defined by its centroid, which is the mean of all points in the cluster. The algorithm iteratively assigns each data point to the cluster with the nearest centroid and then recalculates the centroids until convergence is achieved, meaning the assignments no longer change.

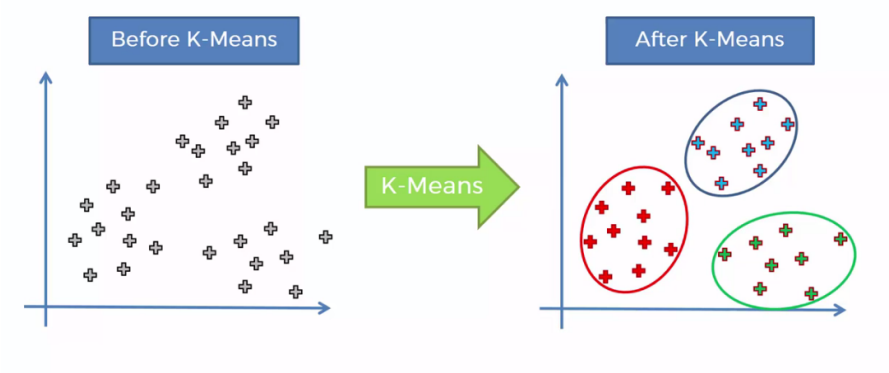

Figure 5: K-means clustering algorithm.

To determine the optimal number of clusters (K), we used the Elbow Method, a widely used heuristic for selecting the appropriate number of clusters in K-Means clustering.

2.4.Application of the Elbow Method

The Elbow Method involves plotting the total within-cluster sum of squares (WCSS) against the number of clusters. The WCSS is a measure of the variance within each cluster, summed over all clusters. As the number of clusters increases, the WCSS decreases, because points are closer to centroids. However, the rate of decrease in WCSS slows as more clusters are added. The point where the rate of decrease sharply shifts, resembling an "elbow," is considered the optimal number of clusters. In our analysis, we plotted the WCSS for different values of K and observed the "elbow" point. The plot indicated a distinct bend at K=4, suggesting that 4 clusters are optimal for our dataset.

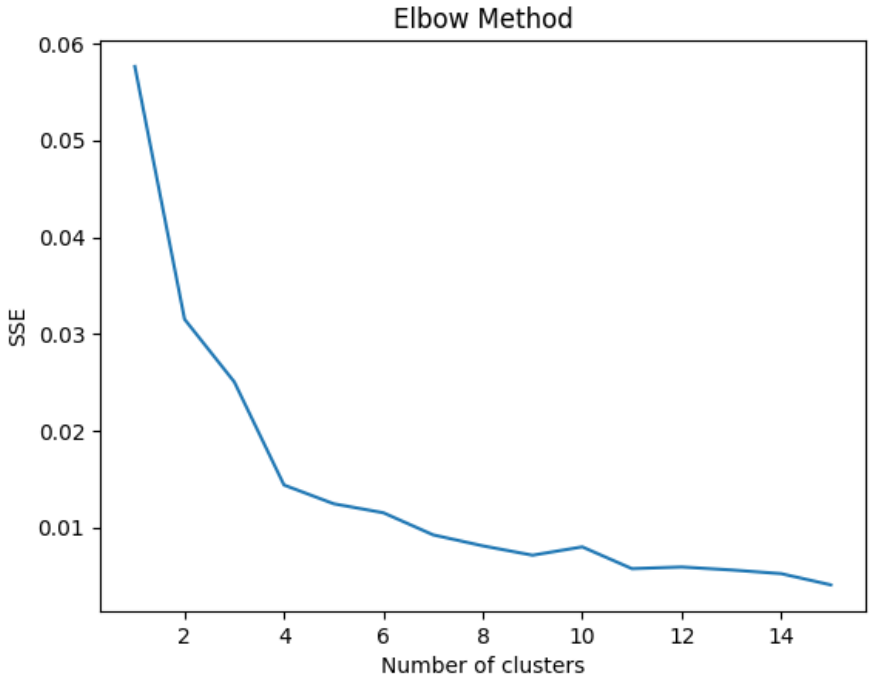

Figure 6: Number of Clusters.

This implies that dividing the locations into 4 clusters would balance the trade-off between

having a sufficient number of clusters to accurately represent the data and maintaining manageable complexity. Using K=4, we applied the K-Means algorithm to our refined dataset of 175 addresses, resulting in four distinct clusters representing potential optimal areas for opening new cafés. This clustering approach allows us to identify and target specific regions within Phnom Penh where student populations are concentrated, thereby maximizing the accessibility and appeal of our café business.

2 Result and Finding

2.1 Description of Clusters

The k-means clustering algorithm was applied to the dataset to identify optimal locations for opening a new cafe. The initial iteration of the clustering process is shown in Figure 7, and the final iteration after convergence is shown in Figure 8. In the initial iteration, the centroids are randomly placed and data points are assigned to the nearest centroid, resulting in the preliminary formation of clusters. The distribution of points in Figure 7 shows considerable overlap, indicating that further iterations are necessary for the algorithm to converge to an optimal solution.

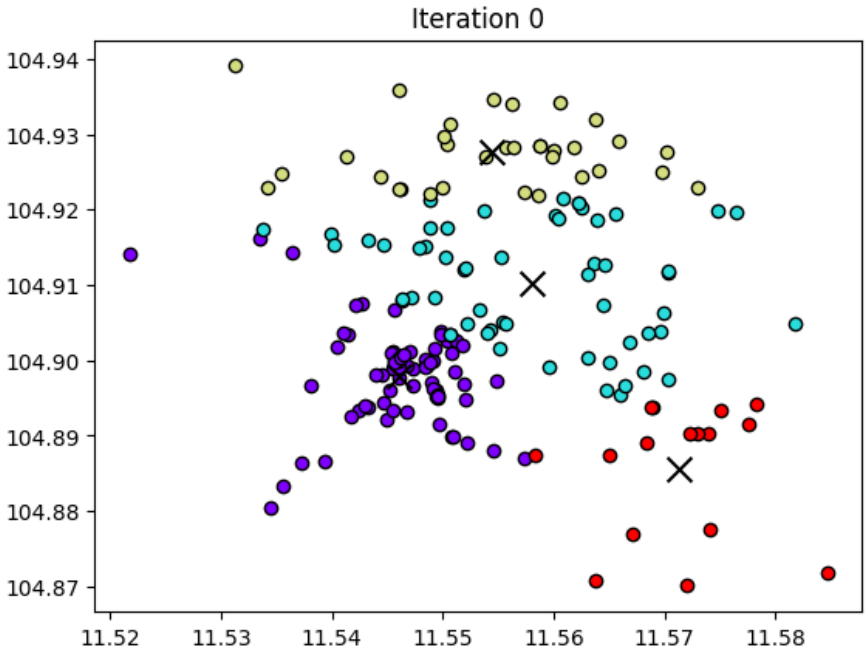

Figure 7: Initial iteration.

After 12 iterations, the algorithm converges, and the final cluster centroids are established. Figure 8 depicts the final state of the clusters with distinct separation between groups. Each color represents a different cluster, and the central black 'X' markers denote the centroid of each cluster.

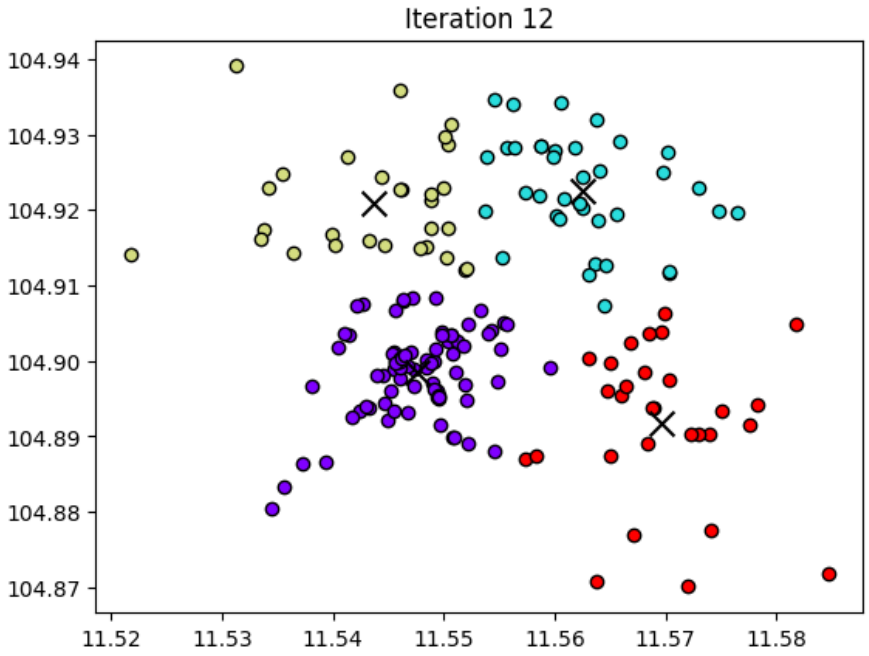

Figure 8: 12 iterations.

2.2 Analysis of Cluster Characteristics

Upon convergence, the data points have been grouped into four distinct clusters, each representing a potential area for opening a new cafe. The characteristics of each cluster can be analyzed based on the density and spread of the points:

- Cluster 1 (Yellow): This cluster has a higher density and smaller spread, indicating a concentrated area of potential customers.
- Cluster 2 (Cyan): Points in this cluster are more dispersed, suggesting a larger geographic area with potential customers spread out.
- Cluster 3 (Purple): This cluster shows moderate density and spread, providing a balance between customer concentration and area coverage.
- Cluster 4 (Red): Similar to Cluster 1, this cluster has a high density of points but in a different location.

2.3 Selection of Optimal Location

The selection of the optimal location for the new cafe involves considering both the density of potential customers and the strategic positioning relative to the centroids of the clusters. Based on the clustering results:

- Cluster 1 (Yellow) is identified as the optimal location due to its high density of data points, indicating a significant number of potential customers in a concentrated area.
- Cluster 3 (Purple) is also a strong candidate, offering a good balance between customer density and geographic coverage.

By focusing on these clusters, we can strategically position the new cafe to maximize reach and convenience for potential customers especially young students and college students.

3 Discussion

The k-means clustering model and elbow method were used to identify optimal locations for opening a cafe business in Phnom Penh, focusing on proximity to educational buildings.The results indicated that the optimal number of clusters (cafes) was determined to be 4.Each cluster represents a potential location where a new cafe could be established. The centroids of these clusters are strategically positioned to maximize accessibility and convenience for the target customer base—students and staff from nearby educational institutions. The clustering analysis revealed several key patterns:

- Clusters are densely populated around areas with a high concentration of educational buildings.
- Some clusters are located in regions that currently lack sufficient cafe services, indicating high potential for new businesses.
- The spread of clusters shows a balance between central and peripheral educational hubs, ensuring broad coverage.

These results suggest that the identified locations have significant potential for attracting a steady stream of customers from educational institutions, which aligns with the primary target demographic for the cafe business.

3.1 Comparison with Existing Cafe Locations

To validate the potential of the identified locations, a comparison was made with the existing cafes in Phnom Penh. The following observations were made:

- Several of the proposed locations coincide with areas that already have a few cafes, indicating that these are indeed high-demand areas.
- Some identified clusters are in regions currently underserved by existing cafes, highlighting new opportunities for market entry.
- Areas with a high density of educational buildings but fewer cafes suggest an untapped market with a high potential customer base.

The performance and popularity of existing cafes in these high-demand areas were also analyzed. Successful cafes in these regions reinforce the viability of the identified locations, while gaps in cafe presence point to lucrative opportunities for new businesses.

3.2 Potential Business Impacts of Chosen Location

The chosen locations are expected to have several positive business impacts:

- Increased Foot Traffic: Being close to educational institutions ensures a constant flow of potential customers, particularly during peak hours such as lunch breaks and after-school periods.
- Market Demand: The high concentration of students and faculty members in these areas suggests a steady demand for cafe services, including food, beverages, and a place to study or socialize.
- Competitive Advantage: Establishing cafes in underserved areas provides a first-mover advantage, allowing the business to capture market share before competitors enter the scene.
- Brand Visibility: Locations near educational buildings offer high visibility and accessibility, which are critical for brand recognition and customer acquisition.

- Revenue Potential: The strategic placement of cafes in high-traffic areas is likely to result in higher sales and revenue, driven by the consistent patronage of the target demographic.

In conclusion, the clustering analysis and elbow method have successfully identified four optimal locations for opening a cafe business in Phnom Penh. These locations not only align with the target customer base but also offer significant business potential due to their strategic positioning and market demand. By leveraging these insights, the cafe business can achieve sustainable growth and profitability.

4 Conclusion

4.1 Summary of Findings

The primary objective of this project was to identify the optimal location for opening a new cafe in Phnom Penh with a specific focus on proximity to educational buildings. Utilizing K-means clustering, we analyzed the spatial distribution of educational institutions and other relevant factors across the city. Our data collection process involved gathering comprehensive geographic information, including the locations of schools, universities, and other educational establishments.

Through the application of the Elbow Method, we determined that the optimal number of clusters was four. Each cluster represented a distinct area with varying densities of educational buildings. Cluster analysis revealed that certain regions of Phnom Penh, particularly in the central area around the Stung MeanChey bridge, exhibited a higher concentration of educational institutions. These areas emerged as prime candidates for the new cafe location due to the potential customer base comprising students, faculty, and staff.

Our analysis further highlighted that these clusters also have varying levels of commercial activity and accessibility factors, which are crucial for the success of a cafe. By evaluating these clusters, we were able to identify specific neighborhoods that not only host a significant number of educational buildings but also demonstrate robust foot traffic and commercial viability.

4.2 Implications for Future Cafe Openings

The findings of this project have several implications for future cafe openings in Phnom Penh and similar urban settings. Firstly, the methodology of using K-means clustering to analyze potential locations based on proximity to educational institutions can be applied to other types of businesses that rely on foot traffic from such demographics. The clustering approach allows for a data-driven decision-making process, reducing the risk associated with new business ventures.

Secondly, the identified clusters provide a strategic blueprint for expansion. Entrepreneurs can prioritize areas within these clusters for subsequent cafe openings, ensuring that new locations are established in regions with proven potential. This systematic approach not only maximizes the likelihood of success for individual cafes but also facilitates the creation of a cohesive brand presence within targeted neighborhoods.

Moreover, the insights gained from this project emphasize the importance of integrating various data sources such as demographic information and commercial activity indicators into location analysis. This comprehensive approach can help businesses anticipate and adapt to market dynamics more effectively.

4.3 Recommendations for Entrepreneurs

Based on the results and implications of this study, several recommendations can be made for entrepreneurs considering opening a cafe in Phnom Penh or similar urban environments:

- **Data-Driven Location Selection:** Utilize clustering algorithms like K-means to analyze potential locations based on relevant factors such as proximity to educational buildings, commercial activity, and accessibility. This will enable a more informed and strategic decision-making process.

- **Focus on High-Density Areas:** Prioritize locations within identified high-density clusters of educational institutions. These areas are likely to provide a steady stream of potential customers, including students, faculty, and staff.
- **Leverage Additional Data:** Incorporate additional data points such as pedestrian traffic, existing competition, and local demographics to refine location selection further. This will help in identifying micro-locations within clusters that offer the highest potential for success.
- **Strategic Expansion:** Use the clustering results to plan future expansions methodically. Establishing cafes within identified clusters can create a strong brand presence and operational synergy, enhancing overall business performance.
- **Continuous Monitoring and Adaptation:** Regularly update the analysis with new data to monitor changes in the urban landscape and educational building distributions. This proactive approach will help in adapting to shifts in market dynamics and maintaining competitive advantage.