A dissertation submitted to the **University of Greenwich**
in partial fulfilment of the requirements for the Degree of

Master of Science
in
Big Data & Business Intelligence

Location Intelligence for Café Site Selection:

Predicting Success and Commercial Rent Prices in London

Name: Feruza Kachkinbayeva

Student ID: 001324705

Supervisor: Sanyaade Olufemi Adekoya
Submission Date: September 6, 2024
Word count: 16,361

LOCATION INTELLIGENCE FOR CAFÉ SITE SELECTION: PREDICTING SUCCESS AND COMMERCIAL RENT PRICES IN LONDON

Computing & Mathematical Sciences, University of Greenwich, 30 Park Row, Greenwich, UK.

Feruza Kachkinbayeva

(Submitted 6 September 2024)

ABSTRACT.

In London's vibrant yet competitive café industry, the success of new ventures is critically dependent on optimal site selection. This thesis explores the enhancement of café site selection processes through the integration of sophisticated data-driven frameworks that include machine learning techniques and the Analytic Hierarchy Process (AHP). The central aim is to employ these methodologies to forecast café success potential and estimate commercial rent prices across London's Lower Super Output Areas (LSOAs), thereby optimizing resource allocation during the site selection process.

The research commenced with the development of predictive models using Linear Regression (LR) and Random Forest Regressor (RFR) to accurately predict commercial rent prices. Initial explorations using clustering techniques such as K-Means and DBSCAN aimed to identify patterns in location characteristics but were ultimately not incorporated into the final model. Instead, the Analytic Hierarchy Process (AHP) was employed to systematically evaluate and prioritize key success factors such as income levels, public transport accessibility, and competitive density, which proved essential for effective site selection.

AHP's implementation allowed for a refined approach to comparing predicted rent prices, actual rental data, and café success potential for each LSOA. This comprehensive comparison enabled stakeholders to identify locations that balanced profitability with cost-effectiveness, significantly improving strategic decision-making in the café site selection process.

The findings confirm that areas with higher income levels and better transport connectivity are strongly correlated with higher café success potentials. The comparison between predicted and actual rents also helps stakeholders avoid overvalued areas, thus refining the overall site selection strategy. The study concludes that the use of AHP and machine learning techniques, supported by interactive geographical visualizations, offers a robust framework for enhancing the café site selection process in London.

Keywords: Café Success Prediction, Commercial Rent Estimation, Linear Regression, Random Forest Regressor, Analytic Hierarchy Process, Interactive Maps, Location Intelligence.

Acknowledgments

I extend my deepest gratitude to Dr. Sanyaade Olufemi Adekoya for his invaluable guidance and support throughout this MSc project. His expertise and consistent feedback have been instrumental in shaping this research. I am also grateful to Dr. Hai Huang for his insightful assessment during my thesis demonstration and for their flexibility in scheduling this critical event.

Special thanks to my family for their enduring support and encouragement, which has been crucial to my success. I appreciate all the module lecturers for their foundational teachings and to my classmates for their companionship and shared insights, which have greatly enriched my learning experience.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

AHP - Analytic Hierarchy Process

ANOVA - Analysis of Variance

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

EDA - Exploratory Data Analysis

GIS - Geographic Information Systems

K-Means - K-Means Clustering

LSOA - Lower Super Output Area

LR - Linear Regression

MCDM - Multi-Criteria Decision Making

MLP - Multilayer Perceptron

ONS - Office for National Statistics

PT - Public Transport

RFR - Random Forest Regressor

1. Introduction

1.1. Background

The cafe industry has seen significant growth in recent years, particularly in urban centers like London, known for its diverse and dynamic cafe culture. The UK coffee shop market has experienced significant growth in recent years, driven by evolving consumer preferences and an increasing demand for specialty coffee. As of 2023, the market value of coffee shops in the UK is estimated to be around £4.5 billion, reflecting a robust annual growth rate of approximately 6% since 2018 (Pujianto and Tannady, 2023). However, despite this expansion, the industry remains highly competitive, with studies suggesting that up to 74% of new cafes fail within their first five years of operation (Mishra, 2019).

One of the most critical determinants of a cafe's success is its location. Numerous studies have highlighted the importance of strategic location in the hospitality and retail sectors, as it significantly impacts customer foot traffic, accessibility, and visibility—all of which are vital for attracting and retaining customers. For example, Parsa et al. (2005) found that poor location choices were a leading cause of failure in the restaurant industry, accounting for a substantial percentage of business closures. Similarly, research by Adebayo et al. (2022)) identified location as the most crucial factor influencing the performance of retail outlets, emphasizing that accessibility and proximity to customer bases are key determinants of success. Additionally, rent costs, which often constitute a significant portion of a cafe's expenses, play a critical role in the financial sustainability of the business. If rent is disproportionately high relative to anticipated revenue, even a strategically located cafe may struggle to maintain profitability. Conversely, while lower rent might seem economically advantageous, it can lead to business failure if the location lacks sufficient amenities or does not attract the appropriate customer demographic (Parsa et al., 2005).

Traditionally, site selection has been driven by intuition, experience, or limited market research, which can result in suboptimal outcomes. However, as demonstrated by Aboulola (2018), the integration of big data and social media analytics into retail site selection has significantly enhanced the precision of data-driven decision-making, enabling more accurate identification of optimal store locations.

This thesis aims to explore the potential of integrating diverse data sources—including demographic data, competition density, and local amenities—into a comprehensive predictive model that can forecast the success of cafes in various locations across London. By leveraging these advanced analytical tools, this research seeks to provide actionable insights that can enhance strategic planning for cafe entrepreneurs.

In addition, this research will focus on predicting commercial property rent prices in various areas of London. Understanding rent dynamics is essential for cafe owners when evaluating the financial feasibility of potential sites. By forecasting both the success potential and the associated rent costs for different locations, this research aims to equip cafe owners with the tools needed to identify the most promising sites while ensuring alignment with financial constraints. Such an integrated approach ensures that strategic decisions made by cafe owners are informed by a thorough analysis of both profitability and cost-effectiveness, thereby improving their chances of achieving long-term success in a highly competitive market.

1.2. Problem Statement

In London's highly competitive cafe market, the ability to select the right location is critical to the success and longevity of a business. Despite the availability of various tools and models, many entrepreneurs still face significant challenges in making informed decisions due to the complexity of factors involved—ranging from local demographics and accessibility to rent costs and the competitive landscape (Wibisono and Marella, 2020). Current models either oversimplify these complexities or are inaccessible due to their reliance on data that is difficult for small business owners to obtain (Shaikh et al., 2020). This gap leaves cafe entrepreneurs with insufficient tools to accurately predict both the success potential and rent affordability of different locations, increasing the risk of business failure.

This thesis seeks to fill this gap by developing a comprehensive, data-driven framework that integrates multiple key factors into a single predictive model. The goal is to provide cafe owners with a practical tool that enhances their ability to make informed, strategic decisions about location, thereby improving their chances of long-term success in a competitive environment.

1.3. Objectives of the Research

The primary aim of this research is to create a robust, data-driven framework that enhances the accuracy of cafe site selection in London. This overarching goal is broken down into several specific, interrelated objectives:

1. To develop a comprehensive predictive model for assessing the success potential of cafes:

- **Objective**: This model will analyze key factors such as demographics, competition density, foot traffic, and the availability of local amenities. The goal is to systematically evaluate and prioritize these factors to provide entrepreneurs with actionable insights on the most promising locations for cafe success.
- **Results and Insights**: The model will produce a success score for each location, categorizing them into tiers such as low, medium, high, and very high success potential. This categorization will help cafe owners target areas that align with their business goals and customer demographics.
- **Visualization and Presentation**: The results will be presented using geospatial maps that visually highlight areas by their success potential. Additionally, bar charts and cluster diagrams will illustrate the distribution of success scores across different regions, providing a clear comparative view of the best locations.

2. To establish a robust framework for accurately estimating commercial rent prices:

- **Objective**: The framework will predict rent prices based on socio-economic and geographic data, ensuring that cafe owners can effectively budget for their chosen locations. Understanding rent dynamics is crucial for financial planning and long-term sustainability.
- **Results and Insights**: The framework will deliver detailed rent estimates for various locations, identifying areas with rent prices that are either above or below the predicted market value.

This insight will enable entrepreneurs to find affordable locations without compromising on success potential.

- **Visualization and Presentation**: Rent price predictions will be visualized through geospatial heat maps and scatter plots, displaying how rent prices vary across London. These visualizations will allow cafe owners to easily identify locations with favorable rent conditions.

3. To integrate success prediction with rent estimation for optimal cafe site selection:

- **Objective**: By integrating the success potential scores with rent price estimates, this approach will help entrepreneurs identify locations that strike the best balance between high success potential and financial viability. This integration is key to making strategic site selection decisions that maximize profitability.
- **Results and Insights**: The integrated model will generate a composite score or ranking for each location, combining success potential with rent affordability. This will result in a list of recommended sites that offer the best overall value, helping entrepreneurs choose locations that are both promising and cost-effective.
- **Visualization and Presentation**: The integration results will be showcased through multi-layered maps that combine success and rent data, along with bar charts that rank locations by their composite score. This will provide a clear visual representation of the most strategic locations for new cafes.

4. To validate and refine the predictive models through comparison with actual market data:

- **Objective**: This process involves comparing the model predictions with real-world data, including actual rent prices and observed success levels. The goal is to refine the models, ensuring they provide accurate and reliable predictions that can guide cafe site selection effectively.
- **Results and Insights**: The validation will highlight any discrepancies between predicted and actual outcomes, leading to refined models that better reflect market realities. This process will improve the reliability of the models, making them more practical for real-world application.
- **Visualization and Presentation**: Validation results will be presented through comparison charts and bar plots that show predicted versus actual data. The refined models will be

displayed alongside the initial versions, demonstrating the improvements made and providing confidence in their predictive power.

1.4. Significance of the Study

This research makes a valuable contribution to the growing field of location intelligence by introducing a robust, data-driven approach to cafe site selection. The findings of this study have the potential to significantly reduce the high failure rates observed in the cafe industry by equipping entrepreneurs with sophisticated tools and actionable insights that enable informed decision-making. By accurately predicting the success potential of a location, this research allows entrepreneurs to avoid costly mistakes and focus their resources on sites with the highest likelihood of success.

Moreover, the integration of success prediction with rent estimation adds a crucial layer of practicality to the decision-making process. Given that rent is a substantial and often unpredictable expense, it plays a critical role in determining the overall profitability of a cafe. This research enables the identification of locations where actual rent is lower than predicted market rent, particularly in areas with high success potential. As a result, entrepreneurs can secure properties that not only promise strong business potential but also align with their financial constraints, enhancing overall profitability (Tayman & Pol, 2011).

Additionally, the methodology developed in this thesis is versatile and has the potential to be adapted to other retail sectors and geographic regions, thus broadening its applicability. Similar predictive models could be employed to assess the success potential of restaurants, retail stores, or service-based businesses in diverse urban environments. By incorporating a variety of data sources, including real-time foot traffic and social media activity, this research sets a precedent for developing more dynamic and responsive decision-making tools in the retail industry. This adaptability underscores the broader relevance of the study, making it a valuable resource for both academia and industry practitioners.

2. Literature Review

In recent years, advancements in data analytics and machine learning have transformed how businesses approach location-based decisions, particularly in the cafe industry. With the availability of vast amounts of data and sophisticated modeling techniques, site selection has evolved into a more precise and data-driven process. This literature review explores the current methodologies and tools used in predicting cafe success and estimating commercial rent prices.

2.1 Predicting Cafe Success Potential

2.1.1 Clustering Techniques for Identifying Success Potential

The selection of optimal locations for cafes is a complex process, heavily influencing the success of such establishments. Clustering techniques have emerged as crucial tools in this domain, enabling the identification of areas with high potential for cafe success based on customer preferences, spatial dynamics, and socio-economic factors.

Likas, Vlassis, and Verbeek (2018) demonstrated the effectiveness of K-means clustering in segmenting areas based on factors like competition density and proximity to amenities, which helps in targeting locations with the highest profitability potential. Similarly, Susilo (2020) employed the stimulus–organism–response model to show how environmental factors, such as outdoor atmospherics, influence customer perceptions and intentions to visit cafes. This suggests that cafe owners can enhance customer satisfaction and loyalty by strategically leveraging environmental elements.

Prayag et al. (2012) investigated the clustering patterns of restaurant locations in Hamilton, New Zealand, using geo-coded data to identify land use patterns and potential areas for new development. Their study underscores the importance of spatial analysis in site selection, highlighting the benefits of historical data in identifying emerging trends in consumer behavior and location preferences.

Further, Kuhn et al. (2018) discuss the concept of "third places" and their role as social hubs, identifying key criteria for successful cafes, such as accessibility, digital engagement, and innovative business models. This research emphasizes that cafe success is not solely dependent on physical location but also on creating a welcoming environment for customers.

Wong et al. (2017) and Forsyth et al. (2012) explored the significance of location in relation to customer satisfaction, finding that convenience and accessibility are critical factors influencing customer decisions, reinforcing the importance of these elements in site selection. Finally, Ohri-Vachaspati et al. (2013) examined urban design's role in shaping cafe locations, highlighting how effective design can enhance visibility and accessibility, ultimately influencing cafe success.

2.1.2 Multi-Criteria Decision Methods (MCDM) Methods in Site Selection

Selecting an optimal site for a cafe involves evaluating multiple criteria to ensure the success and sustainability of the business. Multi-Criteria Decision-Making (MCDM) methods provide a structured approach to assess various factors influencing site selection, such as location, rent, competition, and customer demographics. This section explores the application of different MCDM techniques in cafe site selection, emphasizing the effectiveness of these methods in making informed decisions.

Importance of MCDM in Site Selection

MCDM techniques are crucial in making informed decisions when multiple conflicting criteria are involved. Ibrahim (2021) highlights the increasing use of Geographic Information Systems (GIS) combined with MCDM methods for site selection, emphasizing their effectiveness in spatial decision-making. This combination allows cafe operators to evaluate potential locations based on a range of factors, such as accessibility, visibility, and proximity to competitors, which are essential for business success (Ibrahim, 2021).

The Analytic Hierarchy Process (AHP)

The Analytic Hierarchy Process (AHP) is one of the most widely recognized MCDM methods for site selection. Günen (2021) demonstrates the application of AHP in evaluating suitable sites by ranking various criteria, such as land use and distance from amenities. AHP's structured approach allows cafe operators to systematically weigh factors like foot traffic, nearby

attractions, and rental costs, making it a particularly valuable tool for determining the best location for a cafe (Günen, 2021).

Fuzzy MCDM Approaches

Fuzzy MCDM approaches enhance decision-making by accommodating uncertainty and subjective judgments, which are common in site selection. Tavakkoli-Moghaddam et al. (2010) propose a decision support system that incorporates fuzzy quality function deployment for location selection. This approach allows for a more nuanced evaluation of both tangible and intangible factors, making it particularly relevant for cafes, where customer preferences and market trends can be difficult to quantify (Tavakkoli-Moghaddam et al., 2010).

PROMETHEE Method

The Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE) is another effective MCDM technique for site selection. Tavakkoli-Moghaddam et al. (2015**)** applied PROMETHEE in a fuzzy environment to assess facility location options, highlighting its capability to handle complex decision-making scenarios. Cafe operators can utilize PROMETHEE to rank potential sites based on multiple criteria, including expected revenue, operational costs, and market competition, thereby facilitating a more comprehensive evaluation process (Tavakkoli-Moghaddam et al., 2015).

TOPSIS Method

The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is a popular MCDM method that can be applied to cafe site selection**.** Asadzadeh et al. (2014) discuss the use of TOPSIS for evaluating site alternatives based on weighted criteria. By applying TOPSIS, cafe owners can compare potential sites against an ideal solution, thereby facilitating a more objective and structured selection process (Asadzadeh et al., 2014).

Analytic Hierarchy Process (AHP)

Among the various MCDM methods, AHP stands out due to its ability to handle complex, multi-criteria decisions that are often required in cafe site selection. AHP, developed by Saaty (1980), provides a structured method for organizing and analyzing complex decisions by breaking them down into a hierarchy of simpler sub-problems. This hierarchical structure allows for the

systematic evaluation of each criterion and sub-criterion independently, making it easier to manage the decision-making process.

Günen (2021) demonstrates the application of AHP in evaluating suitable sites by ranking various criteria, such as land use and distance from amenities. AHP's structured approach allows cafe operators to systematically weigh factors like foot traffic, nearby attractions, and rental costs, making it a particularly valuable tool for determining the best location for a cafe (Günen, 2021). This approach ensures that all relevant factors are considered and appropriately weighted based on their significance to the overall objective.

AHP is particularly advantageous in situations where subjective judgments and conflicting criteria are present, which is often the case in cafe site selection. For example, Fauzi, Indriyani, and Yanto (2021) applied AHP to coffee shop location selection, demonstrating its adaptability in optimizing site selection by systematically studying and prioritizing various factors that influence business success. Their study highlights the importance of a structured approach to decision-making, ensuring that all relevant factors are considered and appropriately weighted (Fauzi, Indriyani, & Yanto, 2021).

In broader contexts, AHP has proven versatile across different sectors. Chatterjee and Mukherjee (2013) applied AHP to hospital location selection in rural India, showcasing its ability to prioritize location criteria based on stakeholder input, a process crucial for cafe owners seeking to understand their target market. This adaptability of AHP is further supported by Croce et al. (2019), who compared AHP with other decision-making techniques, illustrating its strengths in site selection scenarios (Chatterjee & Mukherjee, 2013; Croce et al., 2019).

Moreover, the integration of AHP with Geographic Information Systems (GIS) has been shown to enhance the site selection process by providing spatial analysis capabilities. Zeng et al. (2020) utilized a GIS-supported AHP approach for business location selection, demonstrating how spatial data can inform decision-making. This integration is particularly relevant for cafe site selection, as it allows for the visualization of potential locations in relation to customer accessibility, competition, and other relevant factors, facilitating more informed and strategic decisions (Zeng et al., 2020).

In conclusion, while various MCDM methods offer valuable tools for site selection, AHP is particularly well-suited to the nuanced challenges of cafe site selection due to its ability to handle complex, multi-criteria decisions. Its structured approach, adaptability, and integration capabilities with GIS make it the preferred method for cafe operators aiming to optimize their location decisions.

2.1.3 Factors Influencing Cafe Success

Selecting an optimal location is crucial for the success of cafes, as it directly influences customer behavior, perceptions, and overall business viability. This section reviews key location-based factors that cafe owners should consider, including foot traffic, demographics, competition, crime rates, environmental factors, local amenities, and community engagement.

Foot traffic is essential for cafe success, as it increases visibility and the likelihood of spontaneous visits, leading to higher sales. Hsiao and Chen (2020) highlighted the importance of foot traffic and outdoor atmospherics in shaping customer perceptions and visit intentions. Additionally, Valentina (2023) emphasized that cafes in easily accessible locations, near public transportation and parking, tend to perform better due to increased convenience for customers.

The demographic profile of an area, including age, income level, and lifestyle, is critical for cafe success. Ton et al. (2022) found that service quality and cleanliness, which are often linked to demographic characteristics, significantly influence consumer preferences in cafes. Additionally, Hasyim et al. (2022) noted that cafes in densely populated areas benefit from a larger customer base, enhancing their long-term success prospects.

The presence of competitors can significantly impact a cafe's success. While some competition can indicate a healthy market, excessive saturation may limit a new cafe's potential. Papachristos et al. (2011) explored how the growth of coffee shops can signal economic development, suggesting that understanding the competitive landscape is crucial for differentiating a cafe’s offerings.

Crime rates can influence customer willingness to visit a cafe. Liu et al. (2021) found that cafes in safer neighborhoods attract more customers. Additionally, Liedka et al. (2016) noted that security measures, such as surveillance cameras, enhance perceived safety, further influencing customer decisions.

Environmental conditions, including air quality and noise levels, impact cafe success. Masjedi et al. (2019) examined the effects of air quality in cafes, finding that better environmental conditions positively affect customer health and satisfaction. Cafes near parks or scenic views often attract more customers due to the pleasant surroundings.

Proximity to local amenities, such as universities and shopping centers, can enhance a cafe's appeal. Hasyim et al. (2022) found that cafes near universities perform better due to the steady influx of students and young professionals. Being near complementary businesses can also create synergies that attract more foot traffic.

2.2 Estimating Commercial Rent Prices

2.2.1 Machine Learning for Rent Estimation

Machine learning techniques are highly effective in estimating commercial rent prices due to their ability to handle complex datasets and identify patterns that may not be apparent through traditional methods. Jung et al. (2022) demonstrated that ML models can reduce information asymmetry for non-local investors in commercial real estate, improving investment decisions through more accurate price predictions.

Various machine learning algorithms have been employed in rent estimation, with studies showing that models like Random Forest outperform traditional regression techniques. Čeh, Kovačić, and Šojat (2018) found that Random Forest models were more effective than multiple regression in predicting apartment prices, a methodology that can be applied to commercial real estate. Similarly, Hu and Tang (2023) proposed a Geographically Weighted Stacking Ensemble Learning model (GERPM) that enhances predictive accuracy by integrating multiple ML approaches, making it particularly useful for commercial properties where spatial variations are significant.

The selection of appropriate machine learning models and their parameters is crucial for accurate predictions. Zhang (2017) discussed the importance of model selection and optimization, presenting a neural network-based model specifically designed for commercial real estate price evaluation. This approach is well-suited for capturing the non-linear relationships that characterize rent pricing.

External factors such as economic conditions and regulatory changes also play a significant role in rent pricing. Stacy, Smith, and Jones (2023) used ML to analyze the impact of land-use reforms on housing costs, illustrating the broader economic context that influences rent prices. Additionally, Miao, Zhang, and Wang (2020) explored the cyclical nature of price-rent dynamics in commercial real estate, emphasizing the need for models that can adapt to economic fluctuations.

In conclusion, machine learning offers a promising approach to estimating commercial rent prices, with the ability to incorporate a wide range of influencing factors and improve predictive accuracy. These advanced techniques provide valuable insights for investors and policymakers in making informed decisions.

2.2.2 Factors Influencing Commercial Rent Prices

Commercial rent prices are shaped by various location-based factors, including proximity to amenities, transportation accessibility, urban development strategies, and neighborhood characteristics. Understanding these factors is essential for real estate stakeholders, as they directly influence investment decisions and pricing strategies.

Proximity to essential services, such as educational institutions and urban amenities, significantly impacts commercial rent prices. Tomal and Helbich (2022) demonstrated that closeness to educational facilities increases rental property desirability and prices in urban areas like Cracow, Poland. Similarly, Weldegebriel et al. (2021) found that urban redevelopment and access to amenities drive up land values and rental prices in Addis Ababa.

Transportation networks are critical in determining rent growth and demand. Adams and Verbrugge (2021) emphasized that proximity to efficient public transport systems increases location attractiveness, leading to higher rents. Zhang et al. (2019) used geographically weighted regression to show that transportation infrastructure significantly influences rent prices in Nanjing.

Urban development policies, particularly those involving land monetization, can significantly affect commercial rent prices. Garang et al. (2021) noted that such policies lead to higher commercial land prices in redevelopment areas. This is supported by Weldegebriel et al. (2022),

who found that land monetization in Addis Ababa often results in higher rent prices, especially in centrally located areas.

Neighborhood socio-economic status and demographic trends also impact rent prices. Bera and Uyar (2019) found that affluent neighborhoods in Istanbul command higher office rents due to increased demand. Similarly, Li and Xiao (2021) observed that access to leisure, shopping, and educational facilities contributes to higher price-rent ratios in Guangzhou.

In conclusion, factors such as proximity to amenities, transportation accessibility, urban development strategies, and neighborhood characteristics play crucial roles in determining commercial rent prices. Understanding these dynamics is vital for making informed investment and urban planning decisions.

2.3 Integration of Success Prediction and Rent Estimation

Integrating success prediction with rent estimation offers a comprehensive approach to cafe site selection, allowing entrepreneurs to assess both profitability and financial viability. Zhang, Li, and Hu (2020) demonstrated that combining revenue projections with cost estimates, like rent, enhances decision-making in retail site selection.

Financial metrics, such as the rent-to-revenue ratio, further refine site evaluation. Brotman (2021) emphasized the importance of income and rental ratios as key indicators for real estate investment decisions, aiding cafe operators in assessing the profitability of potential locations.

This integrated approach ensures that decisions are informed by both success potential and cost-effectiveness, thereby improving the likelihood of long-term success.

2.4 Gaps in the Existing Literature

Despite advancements in predictive modeling and site selection methodologies, notable gaps remain in applying these methods to the cafe industry, particularly in urban settings like London.

Many studies focus on broader retail or hospitality sectors, often overlooking the unique challenges faced by cafes, such as customer dwell time, social dynamics, and ambiance. This gap suggests that existing models may miss critical factors specific to cafe success (Zhao, Zong, & Wu, 2023).

Additionally, much of the research lacks a focus on the specific dynamics of London's cafe market, characterized by high competition, diverse demographics, and varying rent levels across neighborhoods. This raises concerns about the generalizability of findings from other regions to London (Wibisono & Marella, 2020; Mao et al., 2019). The current study directly addresses this by focusing on London's unique conditions.

Another significant gap is the lack of integrated models that combine success prediction with rent estimation. Most studies treat these elements separately, limiting the ability to make fully informed decisions. This research fills this gap by developing a comprehensive model that considers both profitability and financial viability (Dong, Ratti, & Zheng, 2019).

Furthermore, existing studies often underutilize granular demographic information at the neighborhood level. By incorporating these into predictive models, this study enhances the accuracy and responsiveness of cafe site selection (Ouyang et al., 2020).

This research aims to address these gaps by creating a data-driven framework specifically tailored to cafe site selection in London, integrating success prediction with rent estimation and utilizing advanced demographic data.

3. Theoretical Background

3.1. Clustering Techniques

3.1.1. K-Means Clustering

K-means clustering is an iterative unsupervised machine learning algorithm that partitions a dataset into K distinct, non-overlapping clusters, where each data point belongs to the cluster with the nearest mean (Terada, 2014). The algorithm works by iteratively assigning each data point to one of the K clusters, minimizing the within-cluster sum of squares (WCSS), which represents the squared distance between each data point and the centroid of its assigned cluster.
Mathematically, this can be expressed as:

$$\boldsymbol{WCSS} = \sum_{i=1}^{K} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$$

Where:

- K represents the number of clusters.
- C_i is the i-th cluster.
- x_j is a data point in cluster C_i.
- μ_i is the centroid of cluster C_i
- $|| \cdot ||^2$ is the squared Euclidean distance.

The algorithm proceeds by first initializing K centroids randomly, assigning each data point to the nearest centroid, recalculating the centroids, and repeating the process until the centroids stabilize.

Determining the Optimal Number of Clusters

To determine the optimal number of clusters in K-Means clustering, two commonly used methods are the Elbow Method and the Silhouette Score. These methods help ensure that the number of clusters chosen balances simplicity with the accuracy of data representation.

1. Elbow Method: The Elbow Method involves plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters (K). WCSS measures the sum of squared distances

between each point in a cluster and the centroid of that cluster. As the number of clusters increases, WCSS decreases. The optimal number of clusters is often identified at the "elbow point," where the rate of decrease sharply slows, indicating that adding more clusters does not significantly improve the model. This method is effective but may sometimes produce ambiguous results, making it challenging to identify the exact elbow point (Humaira & Rasyidah, 2020).

2. Silhouette Score: The Silhouette Score evaluates the quality of clustering by measuring how similar a data point is to its own cluster compared to other clusters (Shahapure & Nicholas, 2020). It ranges from -1 to 1, with higher values indicating better clustering quality. The formula for the Silhouette Score $S(i)$ for a single data point i is:

$$S(i) = \frac{b(i)-s(i)}{max(a(i),\ b(i))}$$

Where:

- $a(i)$ is the average distance between i and all other points in the same cluster.
- $b(i)$ is the average distance between i and all points in the nearest cluster.

A high Silhouette Score suggests that the data point is well matched to its own cluster and poorly matched to neighboring clusters. This method is particularly useful when determining the quality of the clustering structure (Shahapure & Nicholas, 2020).

3.1.2. DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering algorithm that identifies clusters based on the density of data points in a given space Zhang (2023). Unlike traditional clustering methods such as K-Means, which require the number of clusters to be specified in advance, DBSCAN can discover clusters of arbitrary shapes and sizes while effectively handling noise and outliers (Li, 2020). This flexibility makes DBSCAN particularly suitable for datasets with noise and varying densities.

DBSCAN operates by examining the local density of data points to form clusters. The algorithm classifies points into three categories:

- **Core Points:** A point is a core point if it has at least **MinPts** (a user-defined minimum number of points) within a radius ε (epsilon).
- **Border Points:** A point is a border point if it is within the ε radius of a core point but does not itself have enough points within its ε radius to be considered a core point.
- **Noise Points:** A point is considered noise if it is neither a core point nor a border point.

Clusters are formed by grouping together core points that are reachable from one another within the distance ε. Border points that are within ε of a core point are added to the nearest cluster, while noise points are discarded (Ester et al., 1996).

Mathematically, DBSCAN operates as follows:

1. For each point p in the dataset, calculate its ε -neighborhood:

$$N\epsilon\ (p) = \{q\ \epsilon\ D \mid dist(p,q)\ \leq \epsilon\}$$

where $D$is the dataset, and $dist(p,q)$ is the distance between points p and q.

2. A point p is a core point if:

$$|N\epsilon\ (p)|\ \geq\ MinPts$$

The algorithm proceeds by iteratively forming clusters by linking core points within ε distance and assigning border points to the nearest core point's cluster. Points that do not satisfy the core point condition and do not belong to any cluster are labeled as noise (Schubert et al., 2017).
DBSCAN's effectiveness in identifying clusters of arbitrary shape and its robustness to noise make it a powerful tool in clustering applications. However, the choice of $\boldsymbol{\varepsilon}$ and **MinPts** significantly impacts the results, often requiring empirical testing or domain knowledge for optimal parameter selection (Ester et al., 1996).

3.2. Analytic Hierarchy Process (AHP)

The Analytic Hierarchy Process (AHP) is a prominent Multi-Criteria Decision-Making (MCDM) technique developed by Thomas Saaty in the 1970s. AHP is widely used to aid decision-makers in evaluating complex problems that involve multiple criteria, often with conflicting objectives (Saaty, 1980). This method is especially valuable in scenarios where decisions need to be made by considering both qualitative and quantitative aspects.

AHP decomposes a complex decision problem into a hierarchy of more manageable sub-problems, each of which can be analyzed independently. The hierarchy generally consists of three levels:

1. **Goal**: The overall objective of the decision-making process.
2. **Criteria and Sub-Criteria**: The factors that influence the decision, often categorized into primary criteria and sub-criteria.
3. **Alternatives**: The possible choices or courses of action.

This hierarchical structure allows decision-makers to focus on smaller, more manageable parts of the problem, enabling a detailed analysis before synthesizing the results into an overall decision (Saaty, 1980).

Mathematical Foundations of AHP

1. Pairwise Comparison and the Pairwise Comparison Matrix:

In AHP, decision-makers compare the relative importance of pairs of criteria or alternatives using a scale of relative importance. This process generates a pairwise comparison matrix A, where each element α_{ij} represents the importance of criterion C_i relative to criterion C_j. The elements of the matrix are based on a standardized scale, such as 1 (equal importance), 3 (moderate importance), 5 (strong importance), 7 (very strong importance), and 9 (extreme importance) (Saaty, 1980). The matrix A is reciprocal, meaning $\alpha_{ij} = \frac{1}{\alpha_{ji}}$ for $i \neq j$.

The pairwise comparison matrix $\boldsymbol{A}$ is represented as:

$$A = \begin{bmatrix} 1 & \alpha_{12} & \cdots & \alpha_{1n} \\ \frac{1}{\alpha_{12}} & 1 & \cdots & \alpha_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ \frac{1}{\alpha_{1n}} & \frac{1}{\alpha_{2n}} & \cdots & 1 \end{bmatrix}$$

where n is the number of criteria (Saaty, 1980).

2. Calculation of the Priority Vector (Eigenvector):

The next step involves calculating the priority vector ω, which represents the relative weights of the criteria. This is done by deriving the principal eigenvector of the pairwise comparison matrix A. Mathematically, this is expressed as:

$$A \cdot \omega = \lambda_{max} \cdot \omega$$

where λ_{max} is the maximum eigenvalue of matrix A, and $\omega = [\omega_1, \omega_2, \ldots, \omega_n]^T$ is the eigenvector corresponding to λ_{max}. The elements of ω are normalized to provide the weights of each criterion (Saaty, 1980).

3. Consistency Check: Consistency Index (CI) and Consistency Ratio (CR):

A critical aspect of AHP is the consistency of the pairwise comparisons. Consistency is checked using the Consistency Index (CI) and the Consistency Ratio (CR). The Consistency Index is calculated as:

$$CI = \frac{\lambda_{max} - n}{n - 1}$$

where n is the number of criteria. The Consistency Ratio (CR) is then computed as:

$$CR = \frac{CI}{RI}$$

where RI is the Random Consistency Index, derived from random matrices. A CR of less than 0.1 is generally considered acceptable, indicating that the comparisons are consistent enough to be reliable (Subramanian & Ramanathan, 2012).

4. Synthesis of Results:

After calculating the priority vectors for all levels of the hierarchy, the overall priority for each alternative is determined by aggregating the weights across all criteria. This involves multiplying the local priorities of the alternatives by the global priorities of the criteria and summing these products to obtain the final scores for the alternatives. The alternative with the highest score is considered the most preferred choice according to the AHP analysis (Subramanian & Ramanathan, 2012).

3.3. Linear Regression

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It is widely applied in various fields, including economics, social sciences, and machine learning (Maulud & Abdulazeez, 2020). The primary goal of linear regression is to find the best-fitting linear relationship between the dependent variable (often denoted as y) and the independent variable(s) (denoted as x for a single variable or X for multiple variables).

The general formula for a simple linear regression model can be expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

- y is the dependent variable,
- x is the independent variable,
- β_0 is the y-intercept,
- β_1 is the slope of the line (indicating the relationship between x and y),
- ϵ is the error term, representing the difference between the observed and predicted values of y (Huang, 2020).

Linear regression is grounded in several key assumptions and theoretical principles that ensure the reliability and interpretability of the model. These assumptions include linearity, independence, homoscedasticity, and normality:

1. **Linearity**: The relationship between the independent and dependent variables is assumed to be linear, meaning that a change in the independent variable x leads to a proportional change in the dependent variable y. This linearity is crucial as it forms the basis of the regression model's predictive capabilities.
2. **Independence**: The observations in the dataset must be independent of each other. This means that the value of the dependent variable for one observation should not influence the value of the dependent variable for another observation.

3. **Homoscedasticity**: The variance of the error terms ϵ should be constant across all levels of the independent variables. If the variance changes, it can lead to inefficient estimates and affect the model's predictions.
4. **Normality**: The error terms ϵ are assumed to be normally distributed. This assumption is important for conducting hypothesis tests and constructing confidence intervals.

The mathematical formulation of linear regression is based on the method of **Ordinary Least Squares (OLS)**, which seeks to minimize the sum of the squared residuals (the differences between the observed and predicted values of y). The OLS estimator can be expressed in matrix form as:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

where:

- X is the matrix of independent variables,
- y is the vector of observed dependent variable values,
- $\hat{\beta}$ is the vector of estimated coefficients (Huang, 2020).

This estimator provides the best linear unbiased estimates of the coefficients under the Gauss-Markov theorem, assuming the aforementioned assumptions hold.
In practice, linear regression can be extended to handle multiple independent variables (Multiple Linear Regression) and polynomial relationships (Polynomial Regression), allowing for more complex models that can capture non-linear relationships between variables. The flexibility and simplicity of linear regression make it a powerful tool for data analysis and prediction in various domains (Su et al., 2012).

Multiple Linear Regression

Multiple Linear Regression extends the simple linear regression model by incorporating more than one independent variable. The general form of the Multiple Linear Regression model can be expressed as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

where:

- $X_1, X_2, \ldots, X_p$ are the independent variables,
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients corresponding to the independent variables.

This formula allows the model to consider the effect of multiple predictors on the dependent variable simultaneously, making it possible to analyze more complex real-world scenarios (Su et al., 2012).

Polynomial Regression

Polynomial Regression is a special case of linear regression where the relationship between the independent variable(s) and the dependent variable is modeled as an nth-degree polynomial. The general form for Polynomial Regression can be expressed as:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \epsilon$$

where $x^2, x^3, \ldots, x^n$ are the polynomial terms.

This model is particularly useful when the relationship between the independent and dependent variables is not linear, allowing for the modeling of more complex, non-linear relationships while still using linear regression techniques (Huang, 2020).

3.4. Random Forest Regressor

Random Forest Regression is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and control overfitting. Originally proposed by Leo Breiman in 2001, the Random Forest algorithm constructs a "forest" of decision trees during training, where each tree independently predicts the output. The final prediction is the average of all tree predictions for regression tasks (Breiman, 2001).

Theoretical Foundation

Random Forests operate on the principle of "bagging" (Bootstrap Aggregating), which involves training multiple models (trees) on different random subsets of the data. This approach reduces variance without significantly increasing bias, leading to a robust and accurate predictive model.

1. **Bootstrap Sampling**: Each tree in the Random Forest is trained on a bootstrap sample, which is a random sample drawn with replacement from the original dataset. This introduces variability among the trees, leading to decorrelated models and reducing the likelihood of overfitting (Scornet, Biau, & Vert, 2014).
2. **Random Feature Selection**: At each split in a tree, Random Forests select a random subset of features to consider for the split. This ensures that the trees in the forest are diverse, further reducing correlation between them. The combination of bootstrap sampling and random feature selection creates an ensemble of "weak learners" that, when aggregated, form a strong predictor (Biau & Scornet, 2015).
3. **Out-of-Bag (OOB) Error Estimation**: Since each tree is trained on a bootstrap sample, about one-third of the data is not used in the training of that tree. This out-of-bag data can be used to estimate the prediction error, providing a built-in validation mechanism without needing a separate validation set. This method offers an unbiased estimate of the model's performance (Zhe-xue, 2013).

Mathematical Foundation

The Random Forest algorithm can be mathematically described as follows:

- **Construction of Trees**: For each tree T_k in the forest, a bootstrap sample D_k of the training data D is drawn. The tree is then grown using this sample. At each node of the tree, a random subset of features is chosen, and the best split among these features is used to split the node.
- **Generalization Error**: The generalization error of a Random Forest decreases as the number of trees increases, provided the trees are not too correlated and have low bias. The generalization error can be expressed as:

- **Prediction**: For a new observation x, the prediction of the Random Forest is obtained by averaging the predictions of all the trees:

$$\hat{y}(x) = \frac{1}{M}\sum_{k=1}^{M} T_k(x)$$

where M is the number of trees in the forest, and $T_k(x)$ is the prediction of the k-th tree for the input x.

- **Generalization Error**: The generalization error of a Random Forest decreases as the number of trees increases, provided the trees are not too correlated and have low bias. The generalization error can be expressed as:

$$GE = \sigma^2(1 - \frac{\rho}{M})$$

where σ^2 is the variance of the error of an individual tree ρ is the average correlation between the trees, and M is the number of trees in the forest (Breiman, 2001).

Hyperparameters in Random Forest

Hyperparameters play a critical role in optimizing the performance of a Random Forest model. The most important hyperparameters include:

1. **Number of Trees (n_estimators)**: This determines how many trees the forest should contain. Increasing the number of trees generally improves performance by reducing variance, but it also increases computational cost.

2. **Maximum Depth of the Trees (max_depth)**: This parameter controls the maximum depth of each tree in the forest. Limiting the depth can prevent the model from overfitting, especially in datasets with a large number of features.

3. **Minimum Samples per Split (min_samples_split)**: This parameter specifies the minimum number of samples required to split an internal node. Higher values prevent the model from learning overly specific patterns, thus controlling overfitting.

4. **Minimum Samples per Leaf (min_samples_leaf)**: This parameter sets the minimum number of samples required to be at a leaf node. A larger value forces the tree to consider more data points in each prediction, smoothing the model.
5. **Maximum Features (max_features)**: This determines the number of features to consider when looking for the best split. Common choices include:
    - auto: Use all features (default for regression).
    - sqrt: Use the square root of the number of features.
    - log2: Use the logarithm (base 2) of the number of features.
6. **Bootstrap Sampling (bootstrap)**: This boolean parameter controls whether bootstrap samples are used when building trees. If set to True, each tree is trained on a bootstrap sample of the data; if False, the entire dataset is used.

Tuning these hyperparameters is crucial for maximizing the model's performance. Techniques such as Grid Search and Random Search can be used to systematically explore different combinations of hyperparameters to find the best configuration (Schonlau & Zou, 2020).

Advantages of Random Forest Regression

Random Forest Regression offers several advantages:

- **Robustness to Overfitting**: Due to the combination of multiple trees and randomness in feature selection, Random Forests are less prone to overfitting compared to individual decision trees (Biau, 2010).
- **Handling High Dimensional Data**: Random Forests perform well even when the number of features is much larger than the number of observations, making them suitable for high-dimensional data (Biau & Scornet, 2015).
- **Variable Importance**: The algorithm provides measures of feature importance, which can be used to identify the most significant predictors in the dataset.

4. Methodology

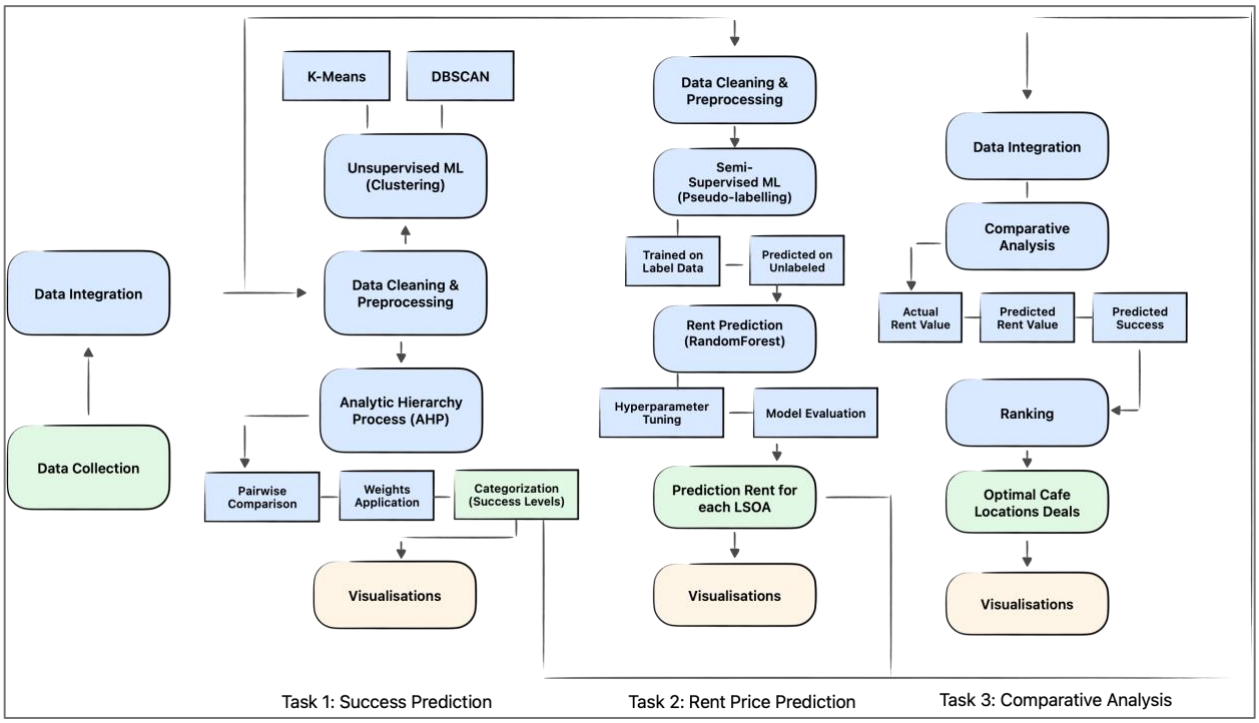
This research adopts a systematic approach to data collection and analysis using secondary data sources such as demographics, competition density, and rental prices. Data is sourced through web scraping with Apify and existing databases. The preprocessing stage includes imputation for missing values, MinMax normalization for data scaling, and various data conditioning procedures to prepare for thorough analysis. Execution of this analysis depends on six essential datasets, which underpin the modeling processes.

Validation of the analytical models incorporates silhouette scoring, consistency checks, and cross-validation to confirm their accuracy and reliability. The entire analytical procedure employs Python, hosted on Google Colab, with advanced data manipulation and geospatial visualization facilitated by Folium and GeoPandas. This methodological setup is crucial for making informed decisions on the optimal selection of café sites across London, analyzing a complex matrix of socio-economic variables to determine viability.

4.1. Methodological Framework

The methodological framework for this research is designed to address the complex challenge of selecting optimal cafe locations by integrating various data-driven techniques. The approach combines unsupervised learning, decision-making processes, and predictive modeling to provide a comprehensive solution that assesses both the success potential of locations and the associated rent costs. The framework, illustrated in Figure 4.1, outlines the sequential and interconnected steps taken to achieve the research objectives.

Figure 1: Methodological Framework



The framework is divided into several key phases:

1. **Data Collection and Integration**: The process begins with the collection of diverse datasets, including demographic information, socio-economic data, competition density, and rent prices across various locations in London. These datasets are then integrated to form a comprehensive database that serves as the foundation for subsequent analyses.

2. **Data Cleaning and Preprocessing**: Before analysis, the data undergoes rigorous cleaning and preprocessing to ensure accuracy and consistency. This step is crucial for removing outliers, handling missing values, and normalizing data, thus preparing it for effective clustering and modeling.

3. **Unsupervised Machine Learning (Clustering)**: Two clustering techniques, K-Means and DBSCAN, are employed to categorize locations based on their characteristics. These methods help in identifying patterns and grouping similar locations, which is essential for understanding the spatial distribution of potential cafe sites.

4. **Analytic Hierarchy Process (AHP)**: AHP is applied to incorporate expert judgment and prioritize the criteria that influence site selection. This step transforms qualitative assessments into quantifiable weights, which are then used to categorize locations into success levels.
5. **Rent Prediction using Semi-Supervised Machine Learning**: A semi-supervised machine learning approach, specifically pseudo-labeling, is utilized to predict rent prices for different locations. This step involves training a Random Forest model on labeled data and predicting rent values for unlabeled data, followed by model evaluation and hyperparameter tuning.
6. **Comparative Analysis**: The predicted rent values are compared with actual rent prices, and the predicted success levels are integrated to rank locations. This comparative analysis allows for the identification of optimal cafe locations that balance high success potential with favorable rent costs.
7. **Ranking and Visualization**: The final step involves ranking the locations based on their composite scores derived from the comparative analysis. Visualizations are generated to clearly present the results, aiding in decision-making and providing actionable insights for cafe entrepreneurs.

4.2. Legal, Social, Ethical, and Professional Issues

Legal Issues:

Data Privacy and Compliance: This research strictly adheres to GDPR guidelines by anonymizing and aggregating data across London's Lower Super Output Areas (LSOAs), ensuring compliance with UK data protection laws. The approach protects individual privacy while allowing for comprehensive demographic and geographic analysis.

Licensing and Data Usage: All data sources utilized, including those from the Office for National Statistics (ONS) and local administrative databases, were accessed under appropriate licenses that permit academic use. This ensures that no intellectual property rights are infringed upon and supports the ethical use of existing data resources.

Social Issues:

Community Impact: The research assesses the impact of café placements on community dynamics and local economies. It addresses potential displacement issues and explores strategies for integrating new businesses in a manner that supports local economic growth and community development.

Accessibility and Inclusion: Emphasis is placed on selecting café sites that are accessible to all individuals, including those with disabilities. This inclusive approach ensures that the benefits of new café locations extend to a diverse demographic, promoting greater social integration and community engagement.

Ethical Issues:

Bias and Fairness in Algorithms: The use of machine learning models, including Linear Regression and Random Forest, is scrutinized for bias. Measures are implemented to ensure these models do not perpetuate socioeconomic disparities, thereby fostering fairness and equity in predictive outcomes.

Transparency in Decision-Making: The research adheres to high ethical standards by maintaining transparency in data handling, model training, and parameter influence. This clarity in methodology supports trustworthiness and ethical integrity in the research findings.

Professional Issues:

Adherence to Professional Standards: This study conforms to rigorous academic and professional protocols in data science and urban planning. By employing recognized analytical techniques and methodologies, the research ensures the reliability and validity of its conclusions.

Continuous Improvement and Adaptation: The study recommends regular updates and revisions to the predictive models based on new data and evolving urban conditions. This commitment to continuous professional development helps maintain the applicability and accuracy of the research over time.

5. Data Collection and Integration

5.1. Data Collection

This research involved gathering data from various trusted sources to build an accurate model for predicting cafe success and rent prices across London's Lower Super Output Areas (LSOAs). The datasets used in this study were selected to provide a comprehensive view of the demographic, geographic, economic, and business-specific characteristics of each LSOA. The data collected formed the basis for the machine learning models and decision-making processes implemented in this research.

5.1.1. Rationale for Using LSOAs

LSOAs were chosen as the primary unit of analysis due to their granular representation of small geographic areas. Each LSOA is smaller than boroughs or wards, making them well-suited for neighborhood-level analysis. This granularity is particularly important in a city like London, where socio-economic, property, and business conditions can vary significantly within short distances. By using LSOAs, the research captures these variations and provides more detailed insights into the factors influencing rent prices and cafe success.

The LSOAs used in this research are based on the most recent geographic boundaries, post-2011 revisions, ensuring that the data reflects current geographic and administrative divisions. The unified dataset includes a total of 4,835 LSOAs, covering the entire Greater London area.

5.1.2. Data Extraction Methods

Data was collected using a combination of open data sources, web scraping techniques, and manual extraction, depending on the type and availability of the required information:

1. **Open Data Sources**: Government-backed open data platforms were the main sources for demographic, economic, and geographic information. Datasets from the Office for National Statistics (ONS) and the London Data Store provided essential information such as population estimates, income levels, crime rates, and public transport accessibility. These sources are recognized for their reliability and are frequently updated, ensuring that the data is accurate and relevant.
2. **Web Scraping**: To collect business-specific data such as the number of cafes, customer reviews, and amenities within each LSOA, web scraping tools like the Apify Google Maps

Extractor and TripAdvisor Scraper were employed. This method allowed for the collection of up-to-date information directly from websites that list and review businesses. Specific search strings were used to target relevant establishments, including cafes, coffee shops, bakeries, and other amenities like tourist attractions and transport hubs.

3. **Manual Extraction**: Data on rent prices was not readily available through open data platforms. Therefore, a manual extraction process was conducted using commercial property websites like OnTheMarket and PropertyLink. Two sets of rent data were collected: one with 347 records for use in the machine learning model and another smaller set of 20 records for validation purposes and comparison with predicted rents.

5.1.3. Data Sources

The data used in this study was gathered from various secondary sources, each providing key features necessary for building the predictive models. The table below outlines the key features collected for this study, along with the data type, extraction method, and source for each dataset:

Table 1. List of Data Sources

Feature	Year	Data Type	Extraction Method	Source	Description
Metadata: LSOA Code, LSOA Name, Postcode, District, District Code, London Zone, Latitude, Longitude	2024	Categorical	Open Data	London Postcodes Dataset (Doogal, 2024)	Geographical area identifiers for each LSOA for reference and mapping.
Average Income	2024	Numerical	Open Data	London Postcodes Dataset (Doogal, 2024)	Average household income in each LSOA.
Index of Multiple Deprivation	2024	Numerical	Open Data	London Postcodes Dataset (Doogal, 2024)	Deprivation index for each LSOA.
Distance to Station	2024	Numerical	Open Data	London Postcodes	Distance to the nearest public transport station.

				Dataset (Doogal, 2024)	
Population	2015	Numerical	Open Data	Office for National Statistics (ONS, 2015)	Population estimates for each LSOA.
Average Age	2015	Numerical	Open Data	Office for National Statistics (ONS, 2015)	Average age calculated based on population distribution.
Median Household Income	2015	Numerical	Open Data	Data.gov.uk (Data.gov.uk, 2015)	Median household income for each LSOA.
Public Transport (PT) Accessibility Levels	2014	Numerical	Open Data	LSOA Atlas, London Data Store (London Data Store, 2014)	Public Transport Accessibility Levels (PTAL) score for each LSOA.
Employment Rate	2014	Numerical	Open Data	LSOA Atlas, London Data Store (London Data Store, 2014)	Employment rate for each LSOA.
Median House Prices	2023	Numerical	Open Data	Office for National Statistics (ONS, 2023)	Median house prices for each LSOA.
Competitors	2024	Numerical	Web Scraping (Search Strings: Cafe, Coffee Shop, Bakery, Tea Room, Coffee Store, Bubble Tea Store)	Apify Google Maps Extractor (Apify, 2024)	Number of cafes and similar businesses within each LSOA.
Cafe Score	2024	Numerical	Web Scraping (Search Strings: Cafe, Coffee Shop, Bakery, Tea Room, Coffee Store, Bubble Tea Store)	Apify Google Maps Extractor (Apify, 2024)	Average score of cafes within each LSOA.
Reviews	2024	Numerical	Web Scraping (Search Strings: Cafe, Coffee Shop, Bakery, Tea Room, Coffee Store, Bubble Tea Store)	Apify Google Maps Extractor (Apify, 2024)	Total number of reviews for cafes within each LSOA.
Amenities	2024	Numerical	Web Scraping (Search Strings: Business Center, Shopping Center, Tourist Attraction,	Apify Google Maps Scraper & TripAdvisor	Number of amenities, including tourist attractions, hotels, and

			University, College, Museum, Gallery, Theatre, Cinema, Gym)	Scraper (Apify, 2024)	stations within each LSOA.
Crime Rate	2023-2024	Numerical	Open Data	Metropolitan Police Service (Metropolitan Police, 2024)	Crime data per 1,000 residents over the most recent 12 months.
Rent Prices (347 Records)	2024	Numerical	Manual Extraction	PropertyLink: Retail for Rent in London	Rent prices collected manually, consisting of 347 records for rent price prediction ML model.
Actual Rent Prices (20 Records)	2024	Numerical	Manual Extraction	OnTheMarket: Commercial Property to Rent in London	Actual rent prices, collected manually, for validation and comparison purposes with predicted rents.
Shapefile for Interactive Map Visualizations	2012	Geospatial	Open Data	Statistical GIS Boundary Files for London	Contains National Statistics data © Crown copyright and database right 2012. Used for detailed mapping and geographic analysis in interactive map visualizations.

5.1.4. Challenges in Data Collection

Despite the comprehensive data collection, several challenges were encountered during the process:

1. **Absence of Direct Foot Traffic Data**: One of the notable limitations during data collection was the unavailability of direct foot traffic data at the LSOA level. Foot traffic is a key predictor of business success, particularly for cafes, as high pedestrian activity often correlates with higher customer footfall. To compensate for the lack of this data, the number of amenities, such as tourist attractions, hotels, and transport hubs within each LSOA, was used as a proxy for foot traffic. It was assumed that areas with a greater number of amenities would experience higher pedestrian traffic, thus acting as a reasonable substitute for direct footfall data.

2. **Limited Rent Data**: Another challenge was the sparse availability of rent data, particularly data that included both property prices and corresponding postcodes. Rent data was not consistently available for every LSOA, which could potentially limit the accuracy of the predictive models. To address this, a semi-supervised learning approach was adopted, where

labeled rent data (i.e., data with both price and postcode information) was combined with unlabeled data to enhance the coverage and depth of the dataset. This approach allowed the research to generate a more complete view of rent patterns across London.

3. **Timeliness of Data**: All datasets used in this research were carefully chosen for their relevance to the study's objectives and credibility as reliable sources of accurate information. These datasets provided valuable insights into the socio-economic environment, geographic accessibility, and competitive landscape of each LSOA, ensuring that the subsequent analysis and machine learning models were grounded in robust, trustworthy data. However, it is important to note that some datasets date back to 2014 or earlier, as more recent data was unavailable for specific metrics at the LSOA level. For instance, Public Transport Accessibility Levels (PTAL) data was collected in 2014, and Employment Rate data was obtained from as early as 2011, with more current datasets being inaccessible. While efforts were made to use the most recent data available, these temporal limitations may affect the timeliness of some aspects of the analysis. Nonetheless, the general trends in the data are expected to remain relevant due to the gradual nature of changes in these indicators.
4. **Data Consistency Across Sources**: Merging data from multiple sources, especially those collected through different methods (e.g., web scraping, open data, manual extraction), required careful preprocessing to ensure consistency. Variations in formats, missing values, and inconsistent naming conventions were addressed through standardization and imputation techniques.

5.2. Data Integration

The integration of various datasets was a critical step in the research to ensure a unified and comprehensive dataset that could be effectively used for the analysis of both tasks: predicting cafe success and rent prices in London's LSOAs. The objective of the data integration process was to combine information from multiple disparate sources into a single dataset, referred to as the LSOA Statistics dataset. This unified dataset formed the basis for all subsequent machine learning and decision-making processes in this study.

5.2.1. Integration of Demographic and Economic Data

The process began with loading and merging key demographic and economic data for each LSOA. The primary dataset included LSOA codes, names, employment rates, and public transport accessibility levels. Population estimates and additional socioeconomic indicators, such as the median household income and average age, were added by merging the Office for National Statistics (ONS) data, ensuring that all demographic information was available for each LSOA. Each data source provided information with varying formats, necessitating consistent renaming and restructuring.

The ONS Population Data (2015) was merged with the base LSOA dataset. Using weighted formulas, the average age for each LSOA was calculated. This stage provided a complete view of each LSOA's demographic profile, including the population structure and employment figures.

5.2.2. Geographic Data and Postcode Mapping

Geographic data for LSOAs, including postcodes, district names, and transport zones, was integrated using the London Postcodes Dataset. This dataset was crucial for linking LSOAs with specific postcodes, which later facilitated the extraction of competitor and amenities data. Postcodes were concatenated and aligned with the LSOA codes, and missing or redundant data were systematically handled during this process.

To capture geographic accessibility, each LSOA was linked to public transport stations, and the Distance to the Nearest Station was included as a feature in the dataset. This was computed by merging the LSOA data with the distance information from the London Stations Dataset, ensuring geographic precision in the dataset.

5.2.3. Integration of Business and Amenities Data

The Apify Google Maps Extractor and TripAdvisor Scrapers actors were used to collect data on cafes, amenities, and tourist attractions within each LSOA. The postcodes in the business and amenities datasets were matched to the LSOA codes in the unified dataset. Metrics such as the number of competitors (e.g., cafes, coffee shops, and similar businesses), average cafe scores, and total reviews were aggregated by LSOA. This provided insights into the competitive landscape within each LSOA.

Amenities were used as a proxy for foot traffic, a critical missing variable in the dataset. Given the lack of direct footfall data, the number of amenities (hotels, tourist attractions, stations, universities,

business centers, etc.) was used to estimate potential pedestrian flow in each LSOA. This assumption was based on the logic that areas with more attractions and transport hubs are likely to experience higher foot traffic.

5.2.4. Crime Data Integration

Crime data was integrated from the Metropolitan Police Service, providing a detailed view of crime rates per 1,000 residents for the period from June 2023 to June 2024. This information was critical to understanding safety levels in each LSOA, which can influence both rent prices and business success. Crime rates were calculated by summing the total number of crimes across the selected period and normalizing them against the population figures to provide a standardized crime rate.

5.2.5. Final Unified Dataset

After merging the demographic, economic, geographic, business, amenities, and crime data, a comprehensive dataset was created for all LSOAs in London. The final LSOA Statistics dataset included 4,835 records, each representing an individual LSOA, and contained a wide range of features that were instrumental in predicting both cafe success and rent prices.

6. TASK 1: Café Success Prediction

The first task of this research is the Prediction of Cafe Success across London's Lower Super Output Areas (LSOAs). The primary objective is to create a robust predictive model that identifies areas where cafes are most likely to succeed based on a combination of demographic, economic, geographic, and business-specific factors. This task leverages the comprehensive LSOA Statistics dataset to assess potential locations and rank them according to their predicted success.

6.1. Exploratory Data Analysis (EDA)

The EDA was conducted to gain initial insights into the dataset and understand the distribution of key features. The dataset contains 4,835 LSOAs across 16 features, including demographic, geographic, economic, and business-specific attributes.

The key findings from this analysis are summarized as follows:

- **Numerical Features**: All columns except Median House Price 2023 were correctly recognized as numerical types, with no missing values in most of the columns except for Median House Price 2023 and Crime Rate per 1000.
- **Non-Numeric Values**: The Median House Price 2023 column was found to contain non-numeric values (indicated by the presence of ':'), necessitating the replacement of these values with NaNs and further imputation for consistency in analysis.
- **Missing Values**: There were missing values in two columns: Median House Price 2023 (386 missing values) and Crime Rate per 1000 (182 missing values). These were identified as important features for cafe success prediction, making it crucial to handle these missing values effectively during data cleaning.
- **Data Distribution**: Histograms (Figure 2) and boxplots (Figure 3) revealed positive skewness in features such as Population 2015, Median House Price 2023, and Competitors, indicating that the majority of LSOAs had smaller populations and fewer competitors, with a few outliers. Features such as Average Income and Employment Rate 2011 showed balanced distributions, while Amenities and Reviews were heavily skewed.

Figure 2. Histogram Subplots Distribution of Key Variables.

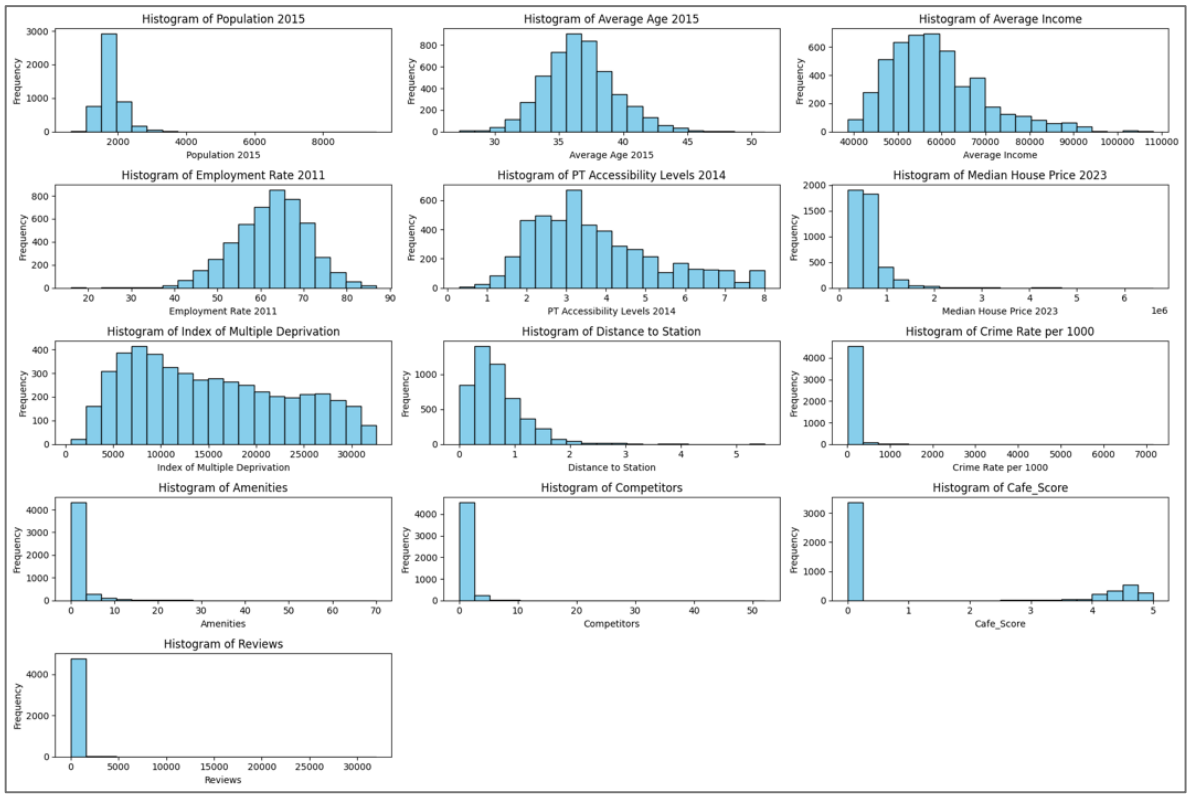
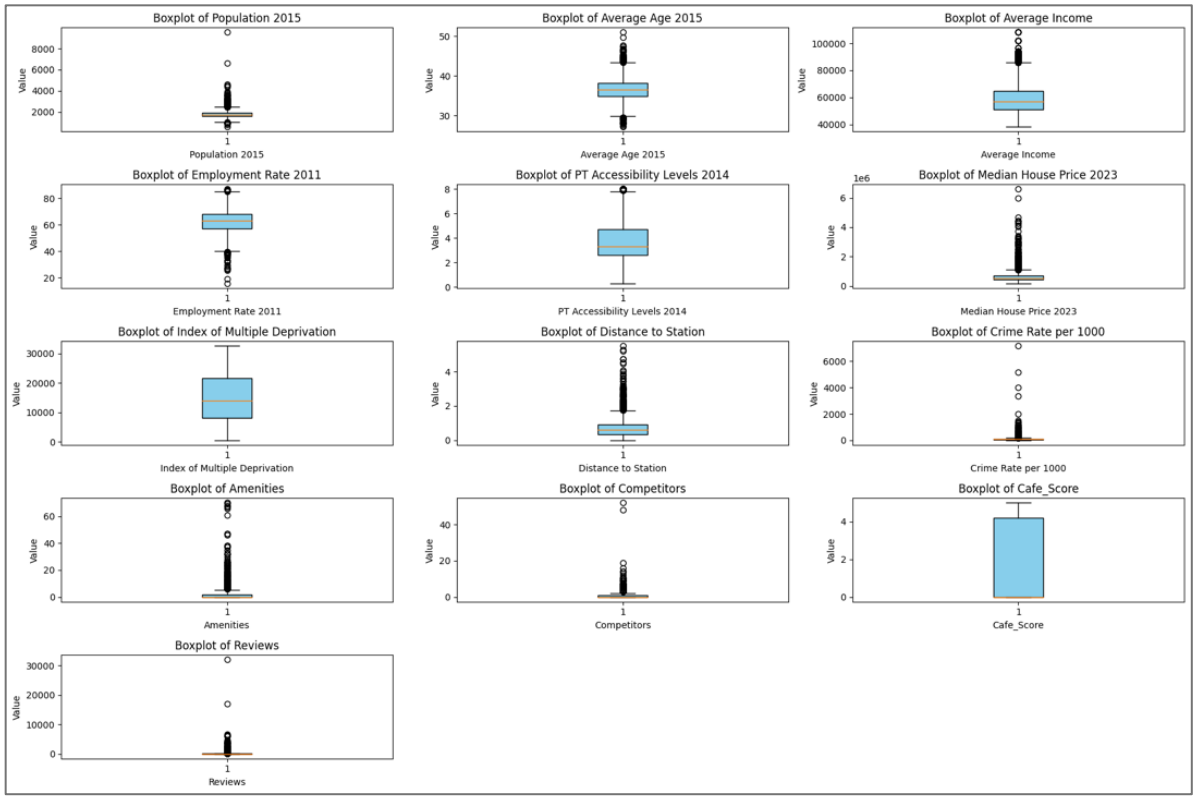


Figure 3. Boxplot Subplots Distribution of Key Variables.

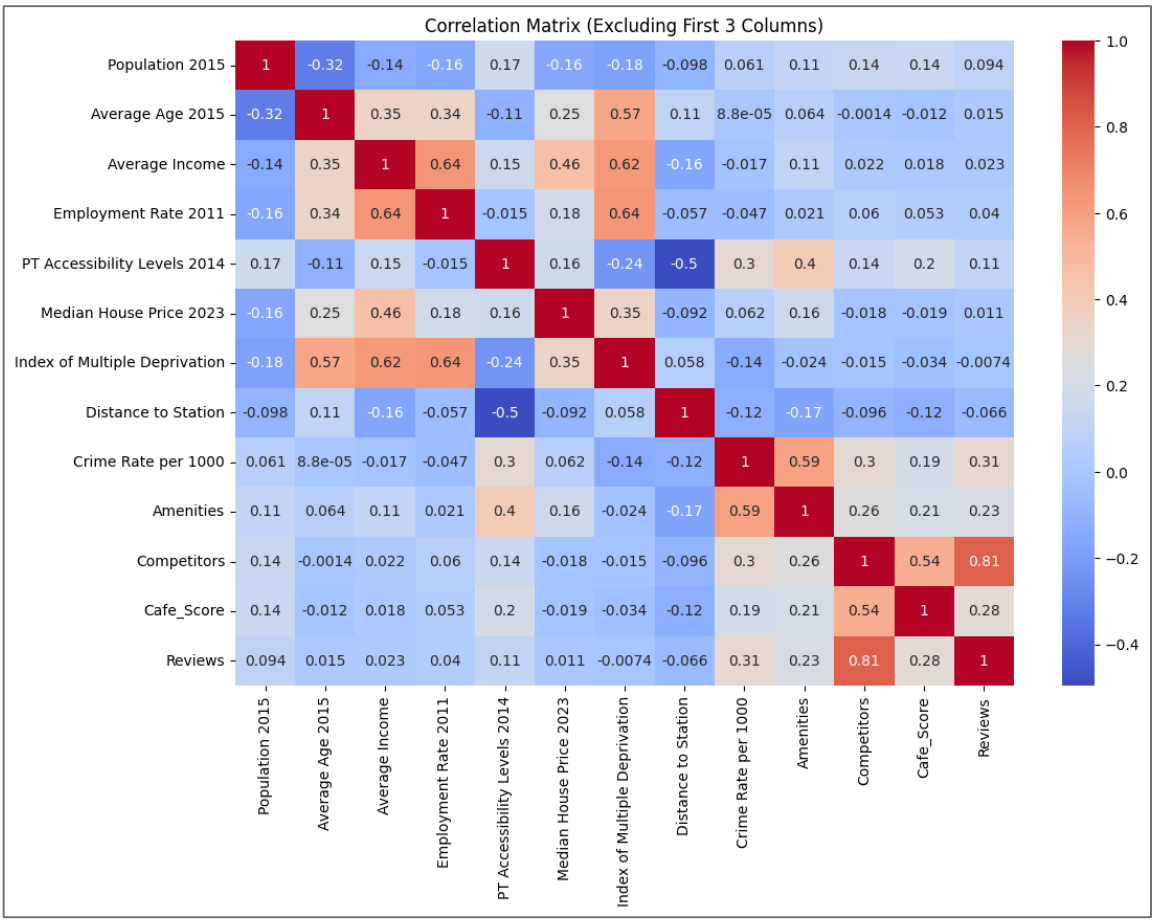


6.2. Data Cleaning

Data cleaning was an essential part of the preprocessing stage to address missing values and inconsistencies in the dataset:

- **Missing Value Imputation**: The missing values in Median House Price 2023 and Crime Rate per 1000 were imputed using the mean value for each column. This imputation strategy was chosen as it is a common approach when data is missing at random, ensuring the dataset remains complete for further analysis.
- **Data Type Correction**: The Median House Price 2023 column was initially stored as an object type due to non-numeric entries. These were identified and replaced with NaNs, and the column was converted to a numeric type to allow for proper analysis.
- **Removal of Redundant Features**: The Reviews feature, which was highly correlated with Competitors (correlation = 0.81) as seen from Figure 5, was dropped to prevent multicollinearity. Retaining Competitors provided more direct insights into the business landscape, which is critical for the success of cafes in each area.

Figure 4. Correlation Matrix



6.3. Data Preprocessing

Before applying machine learning models, data preprocessing steps were performed to standardize the data and prepare it for modeling:

Feature Normalization: MinMax scaling was employed to normalize the numeric features and bring them onto a common scale, ranging between 0 and 1. This is particularly important for distance-based models, such as clustering, where features of different magnitudes can disproportionately influence the results. According to Smolic (2024), MinMax scaling works by subtracting the minimum value of a feature and dividing by the range (the difference between the maximum and minimum values), as shown in the following formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

In this formula:

- X is the original value of the feature.
- X_{min} and X_{max} represent the minimum and maximum values of the feature, respectively.

This scaling technique preserves the relationships between data points while ensuring that all features contribute equally to distance-based calculations.

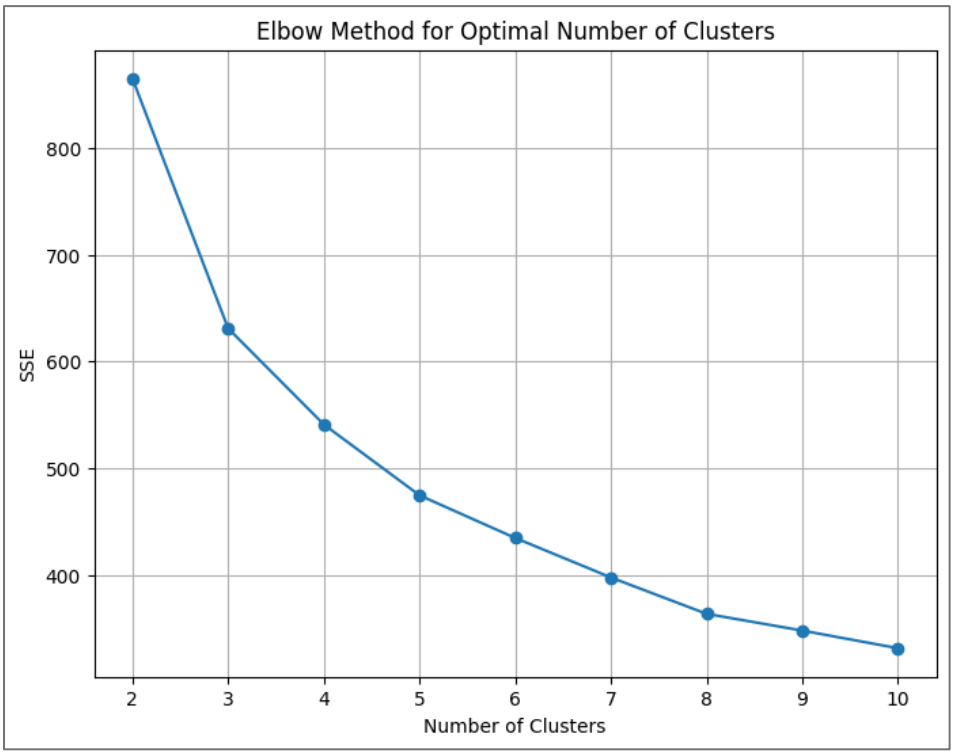
6.4. Unsupervised Machine Learning

In this section, two unsupervised machine learning methods, K-Means and DBSCAN, were explored to cluster London's Lower Super Output Areas (LSOAs) based on key features influencing cafe success. These methods provided initial insights into how different locations across London could be categorized, based on demographic, socio-economic, and business-related data, to inform future decisions about optimal cafe locations.

6.4.1. K-Means Clustering

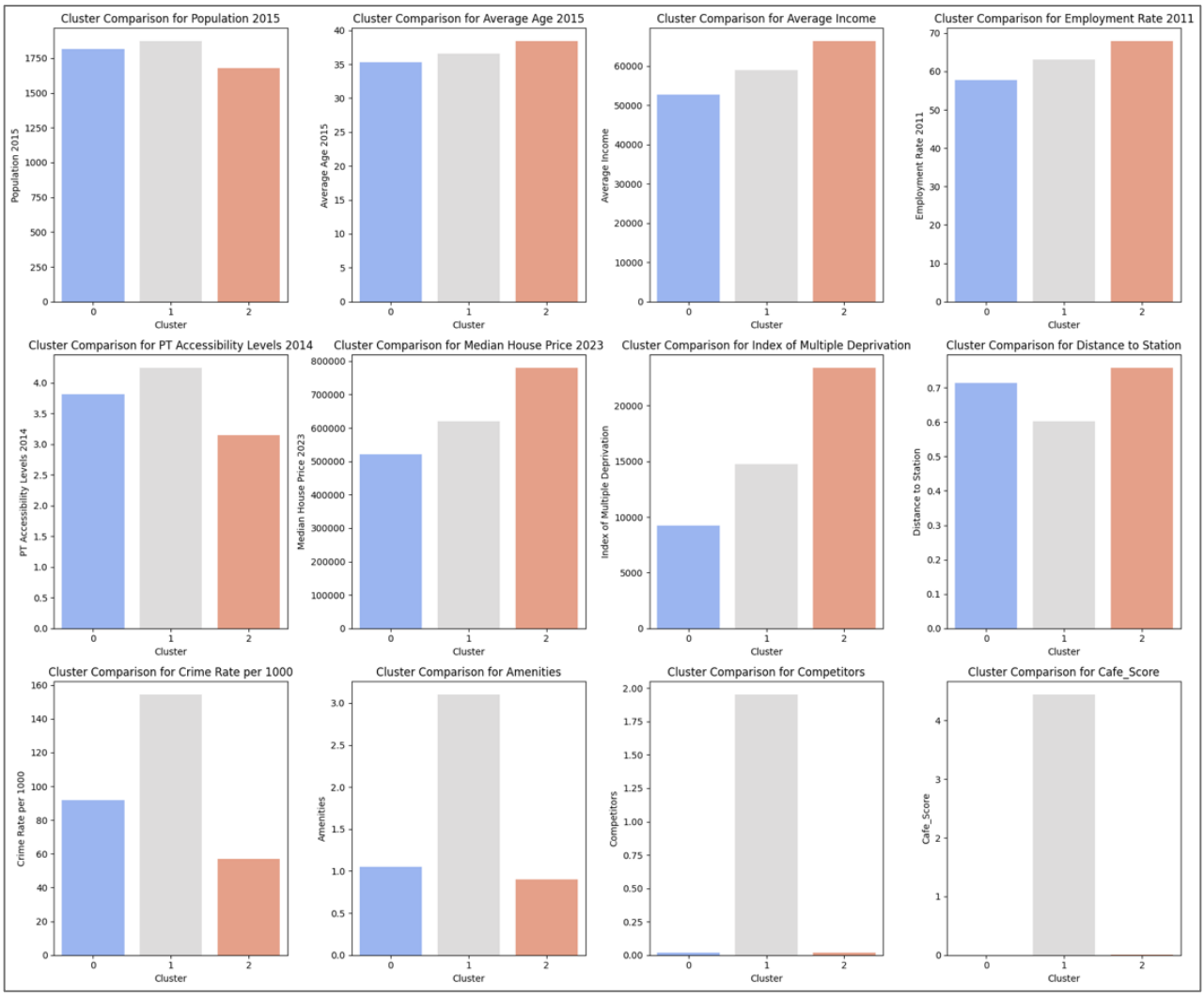
K-Means clustering was applied to the normalized dataset, with the goal of identifying clusters of LSOAs with similar characteristics. The optimal number of clusters was determined using the Elbow Method, which plots the Sum of Squared Errors (SSE) against the number of clusters. As shown in Figure 5 (Elbow Method for Optimal Number of Clusters), the SSE decreases steadily, with the "elbow" point appearing around 3 to 4 clusters, which indicates diminishing returns in adding more clusters.

Figure 5. Elbow Method for Optimal Number of Clusters



Using the Elbow Method, three clusters (K = 3) were selected for K-Means. After running the algorithm, the Silhouette Score—a measure of cluster cohesion and separation—was 0.37, indicating moderate overlap between clusters. Figure 6 (Cluster Comparison for Key Features Using K-Means) presents the comparison of features across the clusters, showing how the clusters differ in terms of population, income, crime rates, and other variables.

Figure 6. Cluster Comparison for Key Features Using K-Means



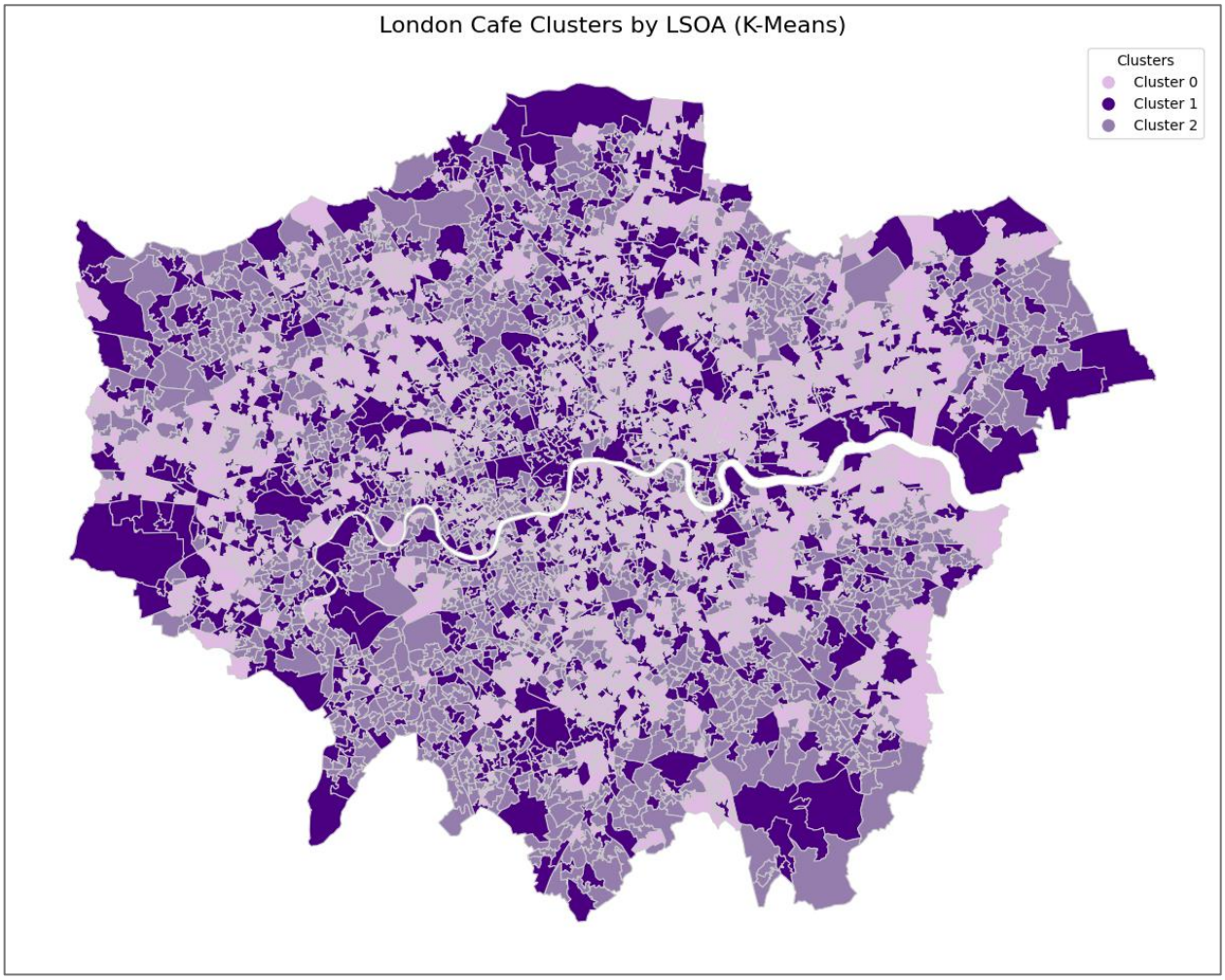
Key observations from the K-Means clustering:

- **Cluster 0**: This cluster had lower average income, lower crime rates, and fewer amenities.
- **Cluster 1**: Areas in this cluster exhibited moderate income and crime rates, but a higher number of amenities and competitors.
- **Cluster 2**: This cluster was characterized by the highest average income and moderate crime rates.

Map Visualisation

The clustering results were also visualized on a map, as shown in Figure 7 (K-Means Cluster Map). The map was generated using the GeoPandas and Matplotlib libraries in Python. First, the LSOA shapefile, containing the geographic boundaries of each LSOA, was loaded, and then the clustering results were merged with this geographic data based on the LSOA codes. Each LSOA area was then color-coded according to its assigned cluster. The use of GeoPandas allowed for seamless integration of geographic data with clustering results, while Matplotlib provided the framework for visualizing the map. The map provides a general spatial distribution of clusters, but the overlap between clusters suggests that K-Means may not have fully captured the underlying structure of the data, especially given the diverse socio-economic landscape of London

Figure 7. K-Means Cluster Map

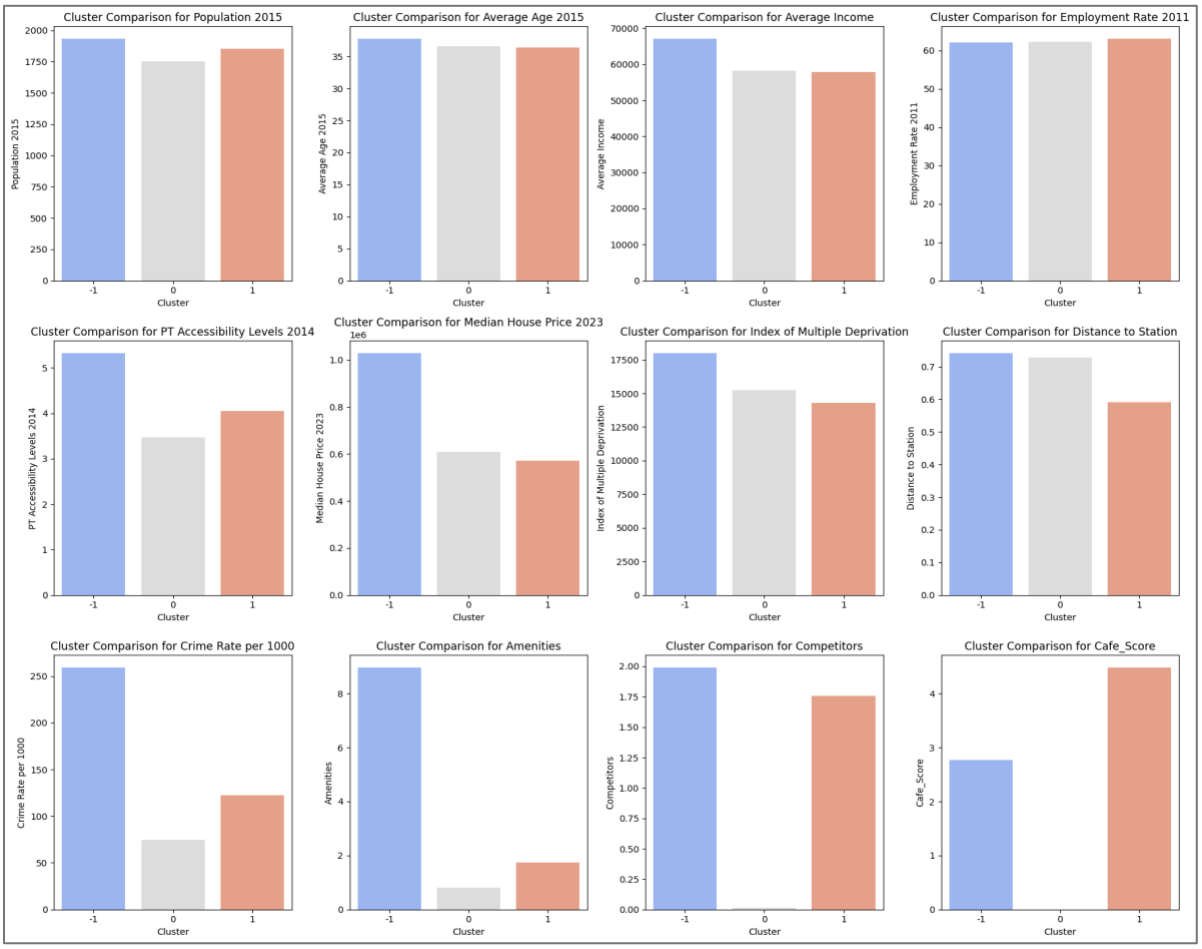


6.4.2. DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was applied next, offering an alternative clustering approach that does not require a predefined number of clusters. DBSCAN identifies clusters based on the density of data points and has the advantage of flagging noise points (outliers), which may be significant in this context given the diversity of London's neighborhoods.

After tuning the hyperparameters (eps = 0.2, min_samples = 7), DBSCAN produced a Silhouette Score of 0.51, indicating better separation between clusters compared to K-Means. The results of the DBSCAN clustering are presented in **Figure 8** (Cluster Comparison for Key Features Using DBSCAN), where differences in population, crime rates, income, and cafe-related factors are clearly observed across clusters.

Figure 8. Cluster Comparison for Key Features Using DBSCAN



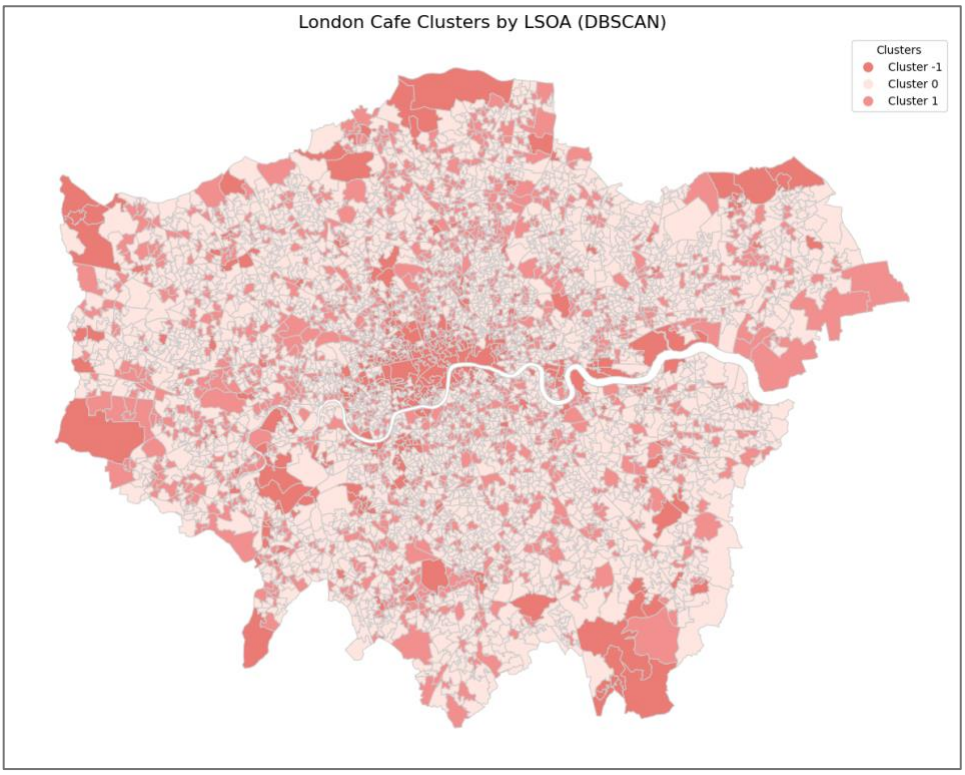
Key observations from the DBSCAN clusters:

- **Noise Cluster (-1)**: Representing high-income areas with elevated crime rates, a high number of amenities, and significant competition from other cafes.
- **Cluster 0**: Characterized by lower crime rates, fewer amenities, and minimal competitors.
- **Cluster 1**: Showcased moderate levels of crime, income, and amenities, with a higher Cafe_Score.

Map Visualisation

To visualize these clusters geographically, the DBSCAN results were mapped using the same approach as with K-Means. **Figure 9** (DBSCAN Cluster Map) shows the geographic distribution of the clusters across London. Similarly, the map was created by merging the clustering results with a shapefile of London LSOAs. DBSCAN identified distinct clusters, including areas flagged as noise (Cluster -1), which represent LSOAs that do not fit neatly into any of the identified clusters. The map reflects a more distinct separation of areas compared to K-Means, particularly highlighting how areas with high income, amenities, and crime rates cluster together.

Figure 9. DBSCAN Cluster Map



6.4.3. Evaluation and Insights

While DBSCAN provided better-defined clusters and a higher Silhouette Score compared to K-Means, it is important to note that the clustering results should be interpreted with caution. Both methods revealed useful initial groupings of LSOAs, yet there are still challenges in fully capturing the complexities of cafe success prediction across London. The geographic maps, Figures 5.3 and 5.5, provide useful spatial insights but highlight some limitations in reflecting the full socio-economic and business diversity of London.

In conclusion, the clustering techniques explored in this section provide an informative exploratory analysis, particularly in understanding how different neighborhoods in London may group based on key cafe success factors. However, further analysis is needed to incorporate more nuanced decision-making. Therefore, the next step in the research is to apply the Analytic Hierarchy Process (AHP), a multi-criteria decision-making tool, to enhance the accuracy and interpretability of cafe location predictions.

6.5. Analytic Hierarchy Process (AHP) Analysis

In this study, the Analytic Hierarchy Process (AHP) was employed to systematically determine the weights of various factors contributing to the success of cafes in London. AHP is a robust multi-criteria decision-making tool that ranks criteria based on their relative importance. While AHP traditionally relies on expert judgments for pairwise comparisons, this study used evidence from empirical studies and statistical correlations to determine the importance of each feature. This approach ensures that the weighting process remains grounded in objective data, rather than subjective expert opinion.

6.5.1 Pairwise Comparison Matrix

To construct the pairwise comparison matrix, each factor influencing cafe success was evaluated in relation to the others. The matrix quantifies the relative importance of the criteria by comparing them in pairs, with the resulting weights representing the priority of each factor. The factors considered were selected based on their relevance to cafe performance, as demonstrated in previous research, with particular attention to demographic, economic, and business-specific variables.

6.5.2 Calculating Weights

The pairwise comparison matrix was constructed based on these factors, with values assigned according to their relative importance. Using eigenvalue decomposition, the AHP method calculates the principal eigenvector, which represents the weights of each criterion. The calculated weights for each factor are listed below:

Table 2. Importance and Weight Assignment for AHP Criteria

Criterion	Importance	Weight (%)	Justification
Population Size (2015)	Moderate	1.82	Population size is essential for creating a potential customer base, but economic factors such as income and employment rate have a stronger influence on cafe success (An et al., 2013).
Average Age (2015)	Moderate	2.92	While age impacts cafe visit frequency, factors like income and employment carry more weight for profitability as disposable income plays a larger role (Sari et al., 2020).
Average Income	High	3.65	Income is directly related to spending power, making it a key factor for sustaining high-priced cafes, especially in affluent areas (Wibisono & Marella, 2020).
Employment Rate (2011)	Very High	5.03	Employment drives economic stability and disposable income, essential for increasing customer frequency and supporting cafes (Iraldo et al., 2017).
PT Accessibility Levels (2014)	Very High	16.92	Public transport accessibility significantly influences foot traffic, and ease of access is one of the highest-ranked factors for attracting customers (Carr et al., 2010).
Median House Price (2023)	High	13.32	House prices are a proxy for affluence, and areas with higher house prices can support premium cafes and upscale offerings (Wibisono & Marella, 2020).
Index of Multiple Deprivation	High (Inverse)	11.75	Deprivation negatively correlates with cafe success; affluent areas with lower deprivation tend to have better infrastructure and consumer spending (Iraldo et al., 2017).
Distance to Station	Moderate	10.17	Proximity to transport hubs enhances foot traffic but is secondary to more significant economic indicators such as income and employment (An et al., 2013).
Crime Rate per 1000	Moderate	8.80	Higher crime rates can deter customers, especially in areas with high foot traffic in the evening. However, this effect can be mitigated by strong economic indicators such as income and employment rates, which help sustain business despite safety concerns (Rosenthal & Ross, 2010).
Amenities	High	7.48	The presence of amenities drives foot traffic and increases the likelihood of customer visits, though it remains secondary to economic indicators (Carr et al., 2010).
Competitors	Moderate to High	7.56	Competitors indicate market demand, but too many competitors can saturate the market. A balance is needed between

			competition and market differentiation (Wibisono & Marella, 2020).
Cafe_Score	Very High	10.59	Customer satisfaction, as measured by the Cafe_Score, is a critical indicator of business success. High customer ratings and positive reviews not only attract new customers but also promote future customer loyalty and sustained business growth (Wang et al., 2016).

6.5.3 Consistency Check

After calculating the weights, a consistency check was performed to ensure the reliability of the pairwise comparison matrix. The Consistency Index (CI) and Consistency Ratio (CR) were calculated, with the CR found to be 0.0693, which is below the acceptable threshold of 0.10. This indicates that the pairwise comparisons were consistent and the results are valid for further analysis.

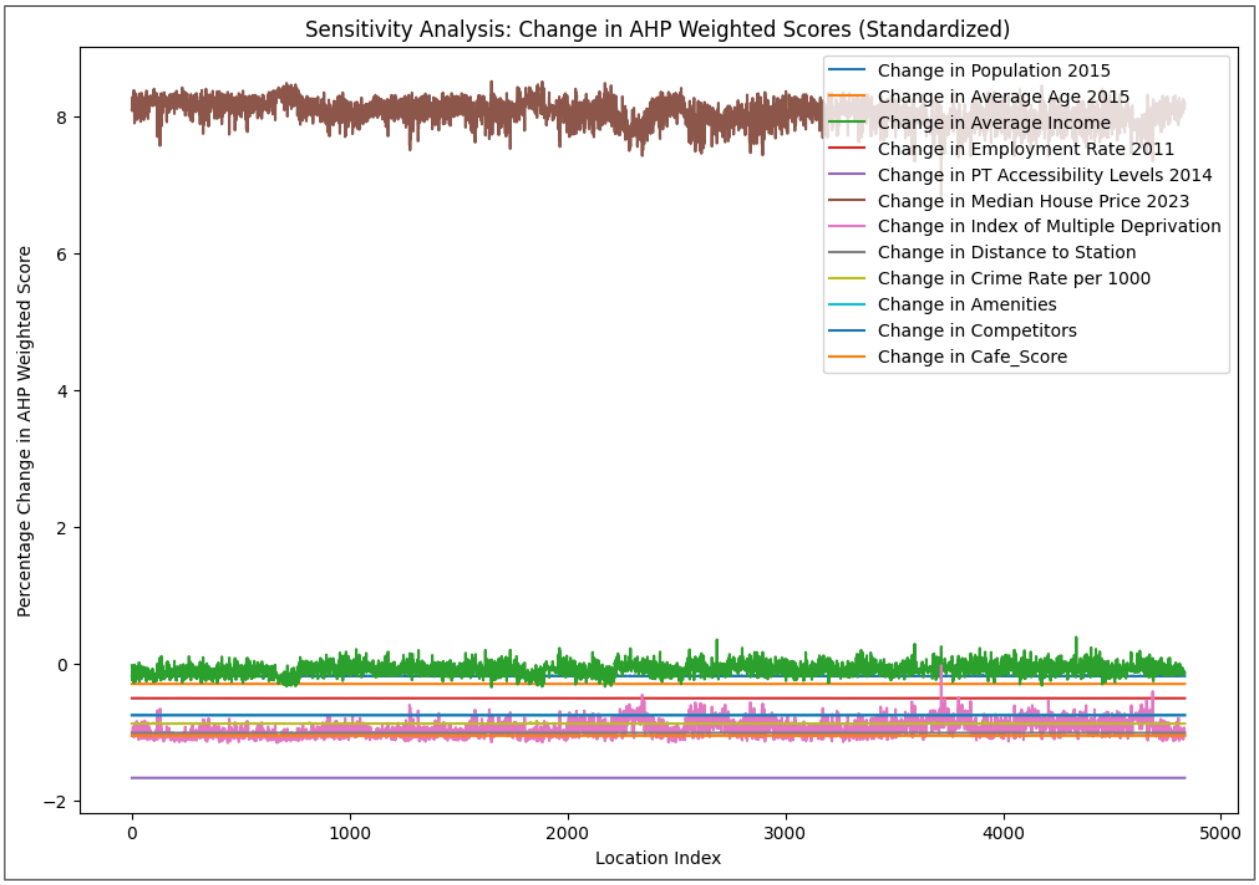
6.5.4 Applying AHP Weights

The weights were then applied to the dataset to compute the AHP Weighted Score for each location. This score represents the combined impact of all the factors for each Lower Super Output Area (LSOA) in London. The scores were used to rank the LSOAs according to their predicted success levels for cafes.

6.5.5 Sensitivity Analysis

Sensitivity analysis was conducted to assess the robustness of the AHP weights. This involved adjusting each criterion's weight by 10% and observing the resulting changes in the AHP Weighted Scores. The analysis showed that Public Transport Accessibility and Median House Prices were the most sensitive factors, significantly affecting the scores when their weights were altered.

Figure 10. AHP Sensitivity Analysis



6.5.6 Categorizing AHP Weighted Scores

The AHP Weighted Scores were divided into four categories—Low Success, Medium Success, High Success, and Very High Success—based on percentile thresholds. This binning strategy ensured a clear separation between different success levels, facilitating a better understanding of where cafes are likely to thrive in London.

6.5.7 Visualizing Results

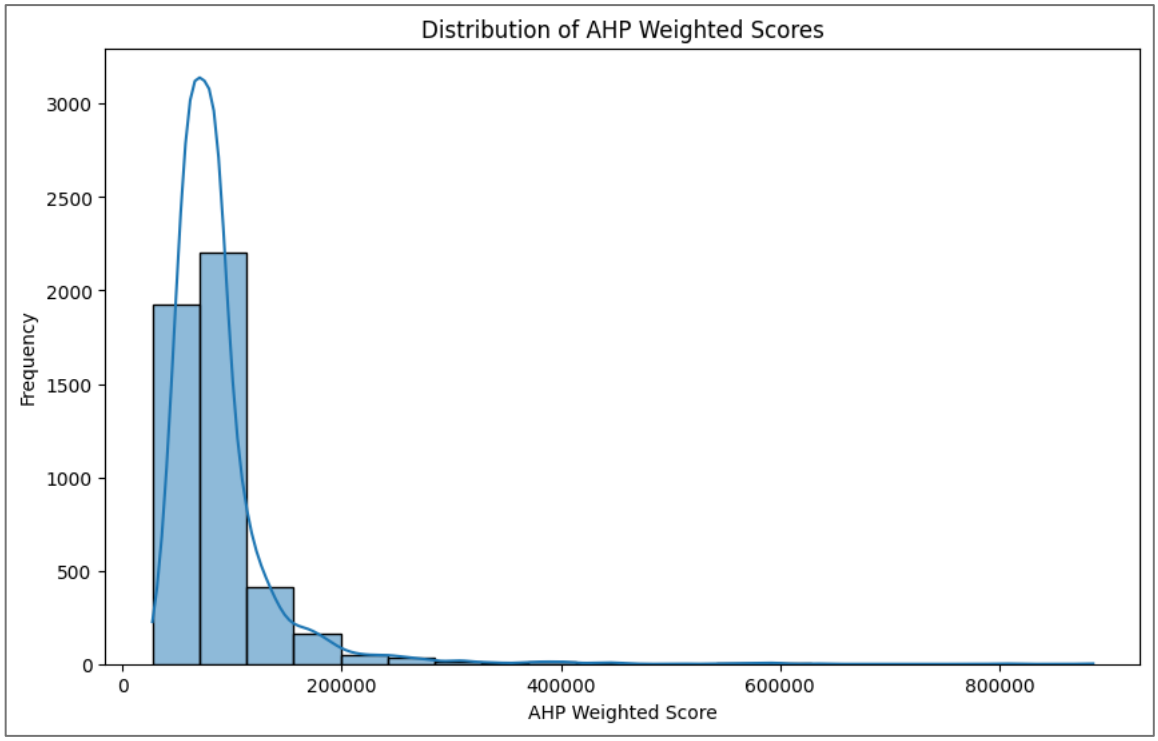
The results of the AHP analysis were visualized through multiple charts and maps to comprehensively interpret and convey the insights derived from the data:

1. **Histogram (Figure 11):**

   This figure presents a histogram that displays the distribution of AHP Weighted Scores across the 4,835 Lower Super Output Areas (LSOAs) in London. The histogram offers a clear overview of the frequency distribution of these scores, highlighting the concentration of

locations with varying degrees of cafe success potential. The bell-like shape suggests a fairly normal distribution, indicating that most LSOAs fall within a medium-to-high range of predicted success, with fewer areas at the extreme low or high ends. The histogram is particularly useful for understanding the general tendency of the scores and helps visualize how most LSOAs align in terms of potential cafe success.

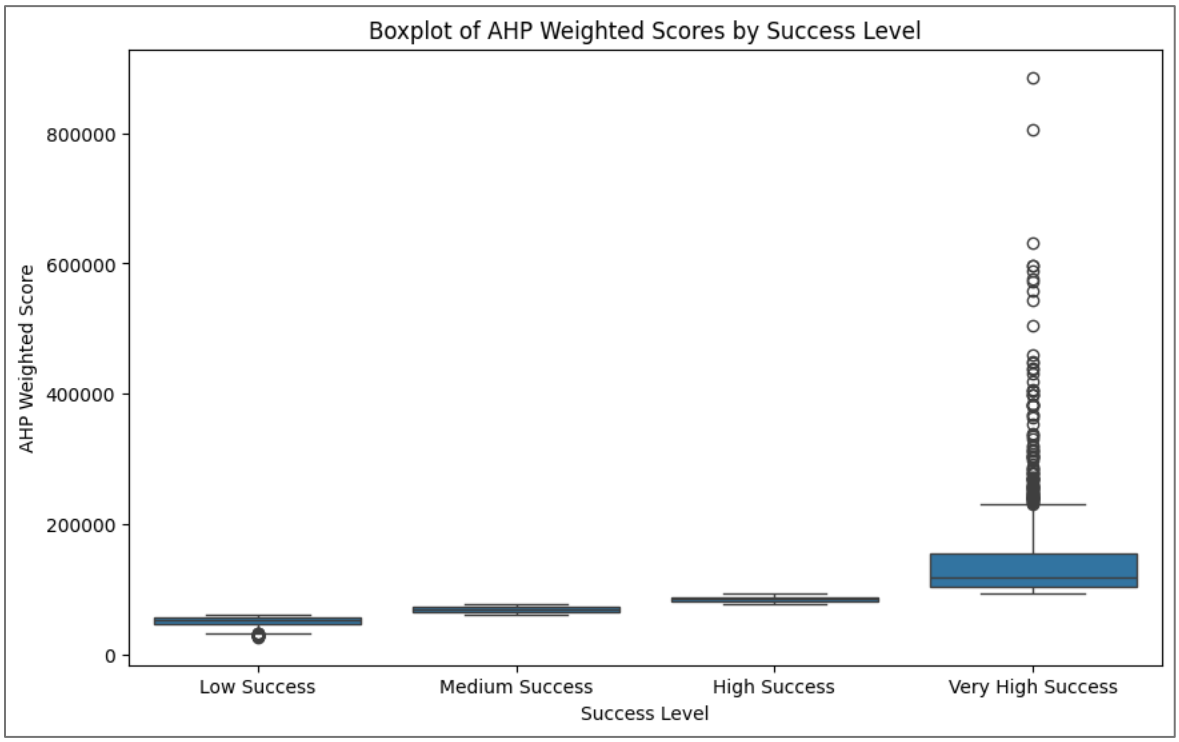
Figure 11. Histogram Distribution of AHP Weighted Scores



2. **Boxplot (Figure 12):**

In this figure, a boxplot compares the distribution of AHP Weighted Scores across each success level: Low Success, Medium Success, High Success, and Very High Success. The boxplot provides detailed insights into the spread, variance, and presence of outliers within each category. Notably, the "Very High Success" group exhibits the largest variance, indicating a wide range of success potentials among the top-performing LSOAs. There are also a few extreme outliers, suggesting that certain areas, while ranked very highly, may outperform the average LSOA within the same success category. This visual is critical for comparing the consistency and variability in success scores across the different success levels.

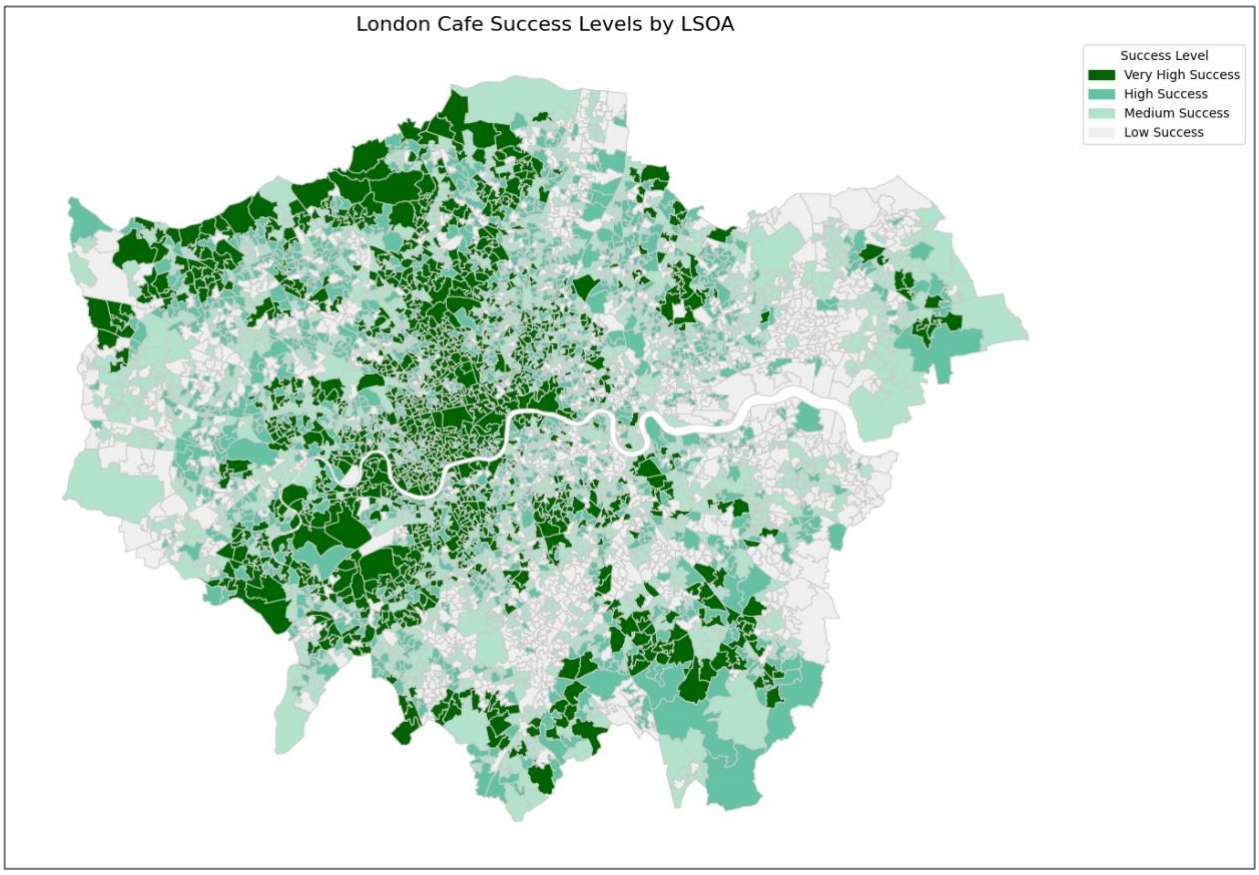
Figure 12. Boxplot Distribution of AHP Weighted Scores



3. **Map Visualization (Figure 13):** This map spatially represents the success levels of cafes across London, using color-coded regions (from light green for low success to dark green for very high success). The visualization merges AHP scores with geographical data, displaying the predicted success levels across London's LSOAs. Darker green areas (Very High

Success) are more affluent or central locations, while lighter areas (Low Success) tend to be in outer, less economically vibrant regions.

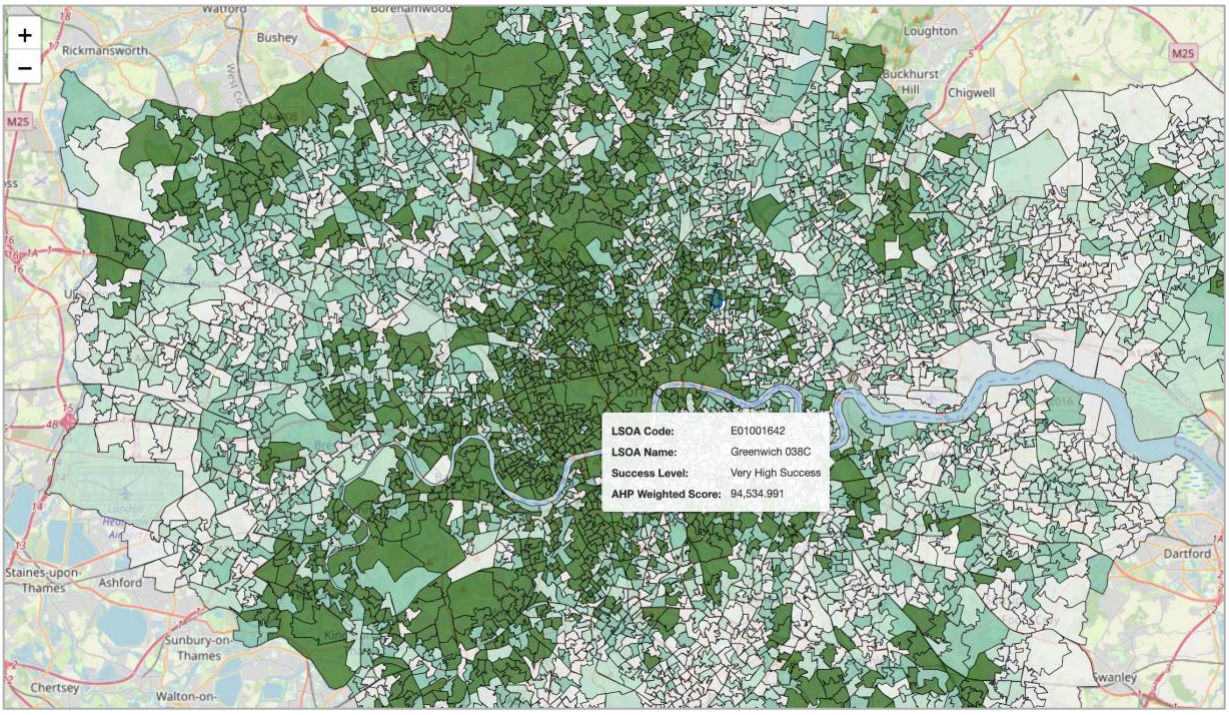
Figure 13. AHP Success Levels Map Visualisation



Explanation: The map was created using GeoPandas for spatial data handling and Matplotlib for visualization. It helps to visualize how geographic and socio-economic factors correlate with cafe success. Central London, with its affluent neighborhoods and high foot traffic, predominantly shows "Very High Success" levels, reflecting real-world market conditions where cafes thrive due to better amenities, transport accessibility, and a wealthier customer base.

4. **Interactive Map (Figure A):** An interactive version of the AHP results map was generated using **Folium**. This map enables users to zoom in and click on individual LSOAs to view detailed success levels and AHP Weighted Scores. For example, selecting Tower Hamlets displays its very high success level with an AHP Weighted Score of 139,481.173.

Figure 14: AHP Success Levels Interactive Map Visualization Using Folium
(Click the image to view the interactive map)



Explanation: The interactive map was developed to allow more user engagement and precise exploration of specific regions. Users can examine how each LSOA fares in terms of cafe success, facilitating a more granular understanding of potential cafe placement decisions. The tooltip provides detailed information such as the LSOA code, name, success level, and AHP score, making the map an invaluable tool for stakeholders analyzing geographic business opportunities.

6.6. Conclusion

Task 1 successfully predicted cafe success across London's Lower Super Output Areas (LSOAs) by analyzing key demographic, geographic, and economic factors. Initial clustering techniques like K-Means and DBSCAN provided insights but struggled to fully capture the complexity of cafe success due to overlapping clusters and limited interpretability.

The Analytic Hierarchy Process (AHP) was ultimately selected as the most effective model. AHP offered a structured approach to multi-criteria decision-making, allowing for the integration of both qualitative and quantitative factors. By assigning weighted importance to features such as public transport accessibility, income, and deprivation levels, AHP provided a clearer ranking of

locations based on cafe success potential. This method's transparency and alignment with existing research, such as Wang et al. (2016), supported its applicability in urban location planning.

In summary, AHP was chosen for its ability to prioritize key success factors and provide actionable, interpretable results, making it an optimal tool for cafe location decisions in London.

7. TASK 2: Commercial Property Rent Price Prediction

Predicting commercial property rent prices is crucial for enabling entrepreneurs and investors to make informed decisions. This section details the methodology for predicting rent prices through the utilization of two machine learning models: RandomForestRegressor and Linear Regression. It covers addressing missing rent values through semi-supervised learning, optimizing the models through hyperparameter tuning, and evaluating model performance using cross-validation and learning curve analysis.

7.1 Rent Data Integration

The rent prediction task begins with the integration of rent data into a comprehensive dataset that contains key variables from the Lower Super Output Areas (LSOA) in London. The rent data, which contains values for rent prices per square foot across various postcodes in London, was combined with the LSOA data to enrich the dataset with additional features, such as demographic information, accessibility, and neighborhood amenities.

The integration process involved matching rent prices to their corresponding LSOA regions using postcodes as the common key. The steps included:

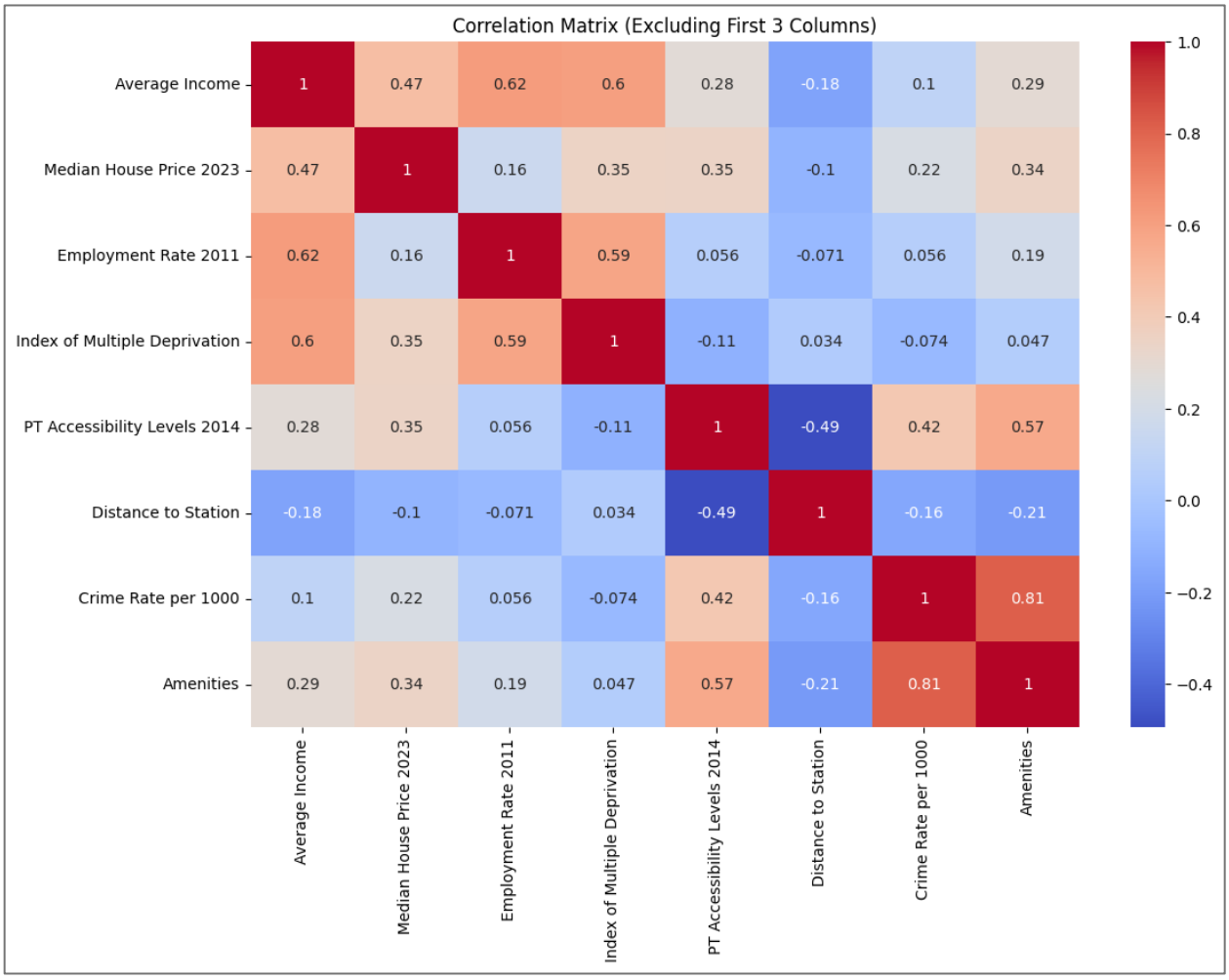
- **Dataset Loading:** Rent price data and LSOA statistics were loaded from their respective sources.
- **Postcode Expansion:** Postcodes in the LSOA data, which often contained multiple postcodes, were split to ensure a precise match with the rent data.
- **Data Merging:** Rent prices were merged with the LSOA statistics based on the postcodes to create a unified dataset. This dataset includes variables such as **average income**, **median house price**, **crime rate**, **distance to the nearest station**, and **public transport accessibility levels**. This expanded dataset provides a comprehensive view of the factors influencing rent prices in each LSOA.

7.2 Exploratory Data Analysis (EDA)

To gain insights into the relationships between rent prices and other variables, an Exploratory Data Analysis (EDA) was conducted on the merged dataset. Key findings from the EDA include:

- **Distribution of Rent Prices:** Rent prices exhibited significant variation across different regions, with values ranging from 15.18 to 62.59 GBP per square foot. This variation reflects the differing socio-economic and geographic conditions across London.
- **Feature Correlations:** A correlation matrix (as shown in the provided heatmap) was generated to assess the relationships between the various socio-economic and geographic features. Some key correlations observed were:
    - **Average Income and Employment Rate** showed a strong positive correlation (0.62), indicating that areas with higher incomes tend to have higher employment rates.
    - **Crime Rate and Amenities** showed a strong positive correlation (0.81), suggesting that areas with more amenities may also experience higher crime rates.
    - **Public Transport Accessibility and Crime Rate** showed a moderate positive correlation (0.42), implying that more accessible areas may have higher crime rates.

Figure 15: Correlation Matrix



- **Insights on Rent Predictors:** The correlation matrix provides valuable insights into which factors may influence rent prices. For instance, public transport accessibility, crime rate, and amenities are expected to be key predictors of rent prices due to their relatively high correlations with other variables.

7.3 Data Cleaning & Preprocessing

Data cleaning and preprocessing were critical steps to prepare the dataset for machine learning. The following steps were performed:

- **Handling Missing Values:** Several variables, including median house price and crime rate, contained missing or invalid values. These were addressed through imputation using the mean of each variable. Imputation ensures that the dataset remains complete and consistent, which is crucial for accurate predictive modeling.
- **Feature Engineering:** The postcode variable, which initially contained multiple entries, was expanded into individual postcodes. Additionally, categorical variables such as LSOA Name were excluded from the analysis since they do not provide numerical insight for regression models.
- **Conversion of Data Types:** Variables such as Median House Price were originally stored as objects due to the presence of non-numeric charact (colons). These characters were removed, and the columns were converted to numeric data types to ensure compatibility with machine learning algorithms.
- **Normalization:** Similarly to task 1, to ensure that all features contribute equally to the model, MinMax scaling was applied to normalize the numeric features. This transformation was essential for distance-based algorithms and helps improve the performance of machine learning models by bringing all features onto the same scale.
- **Correlation Analysis:** The correlation matrix, as shown in the heatmap, helped identify multicollinearity and the strength of relationships between the features. The results will guide the feature selection process for the rent prediction model, ensuring that only relevant predictors are included.

7.4. Predicting Missing Rent Values

The substantial number of missing rent values, constituting 7,653 out of 8,000 records, necessitated accurate predictions to prepare the dataset for further analysis. The RandomForestRegressor was selected for its robust ability to handle complex, non-linear relationships between features and the target variable.

Model Justification:

1. **Complex Interaction Handling:** The RandomForestRegressor is capable of modeling intricate interactions without the necessity for explicit feature engineering, ideal for datasets where relationships between variables are not straightforward.
2. **Robustness Against Overfitting:** This model mitigates the risk of overfitting through its ensemble approach, which uses multiple decision trees to ensure that the model remains generalizable.
3. **Effectiveness with Missing Data:** Suitable for scenarios with incomplete data, making it ideal for semi-supervised learning setups where both labeled and unlabeled data are utilized to enhance model learning accuracy.

Process of Rent Value Prediction:

1. **Data Imputation:** Utilizing SimpleImputer with a mean strategy ensured that missing feature values were consistently addressed across the dataset.
2. **Model Training:** The model was initially trained on a subset containing labeled rent values and was then utilized to predict missing rent values across the dataset.
3. **Performance Evaluation:** The effectiveness of the predictions was assessed based on:
    - **R^2 Score:** 0.6895, indicating a moderate fit to the data.
    - **Mean Absolute Error (MAE):** 12.08 GBP per square foot, providing a quantifiable measure of the average prediction error.

7.5. Hyperparameter Tuning for the Unlabeled Rent Data

Hyperparameter tuning was undertaken to enhance the accuracy and reliability of the RandomForestRegressor, which is critical in semi-supervised learning to ensure precise predictions for the unlabeled data.

Justification for Hyperparameter Tuning:

1. **Enhances Model Performance:** Fine-tuning the model's parameters improves predictive accuracy and ensures better generalization across unseen data.
2. **Optimizes Model Configuration:** Adjusting parameters such as the number of trees and their depth can significantly enhance the model's ability to learn from and make predictions about the data.

Process:

1. **GridSearchCV Implementation:** An exhaustive search was conducted over a predefined parameter grid to find the optimal settings for the model.
2. **Parameter Grid:**

Table 3. Random Forest Regressor Parameter Grid

Parameter	Values
n_estimators	100, 200, 300, 400, 500
max_depth	10, 20, 30, 40, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
max_features	auto', 'sqrt', 'log2'
bootstrap	True, False

3. **Optimized Parameters Achieved:**
    - **Enhanced Performance:** The tuning process yielded an R^2 Score of 0.7731 and MAE of 2.17 GBP per square foot, indicating a substantial improvement in the model's predictive capabilities.

7.6. Training the Final Models

Evaluation of two models was conducted to identify which would most accurately predict rent prices:

Linear Regression with Polynomial Features

Rationale: Real estate pricing often involves complex interactions that linear models typically cannot capture without transformation. Polynomial features allow these interactions to be modeled effectively.

Process:

1. **Feature Engineering:** Polynomial features of degree 2 were generated.
2. **Normalization:** StandardScaler was applied to ensure uniform scaling.
3. **Model Training and Evaluation:**
    - **R^2 Score (Cross-Validation):** 0.6194
    - **MAE:** 9.56 GBP per square foot

RandomForestRegressor

1. **Model Training and Hyperparameter Optimization:** Parameters optimized through GridSearchCV were applied.
2. **Performance Evaluation (Cross-Validation):**
    - **R^2 Score:** 0.9531
    - **MAE:** 1.08 GBP per square foot

7.6.1 Cross-Validation Results

A 5-fold cross-validation was conducted to rigorously assess the performance of the RandomForestRegressor and Linear Regression models, ensuring comprehensive utilization of the dataset for both training and validation. The RandomForestRegressor demonstrated exceptional performance, achieving an average R^2 score of 0.9531 and a Mean Absolute Error (MAE) of 1.08 GBP, with a standard deviation of 0.02 in the R^2 scores, indicating consistent and stable performance across various data subsets. In contrast, the Linear Regression model showed more variability with an average R^2 score of 0.6194, a MAE of 9.56 GBP, and a standard deviation of 0.05 in the R^2 scores, reflecting less stability.

The table below summarizes these results, clearly illustrating the RandomForestRegressor's superior accuracy and reliability, thus confirming its suitability for predicting rent prices.

Table 4. Comparative Performance of Linear Regression and RandomForestRegressor

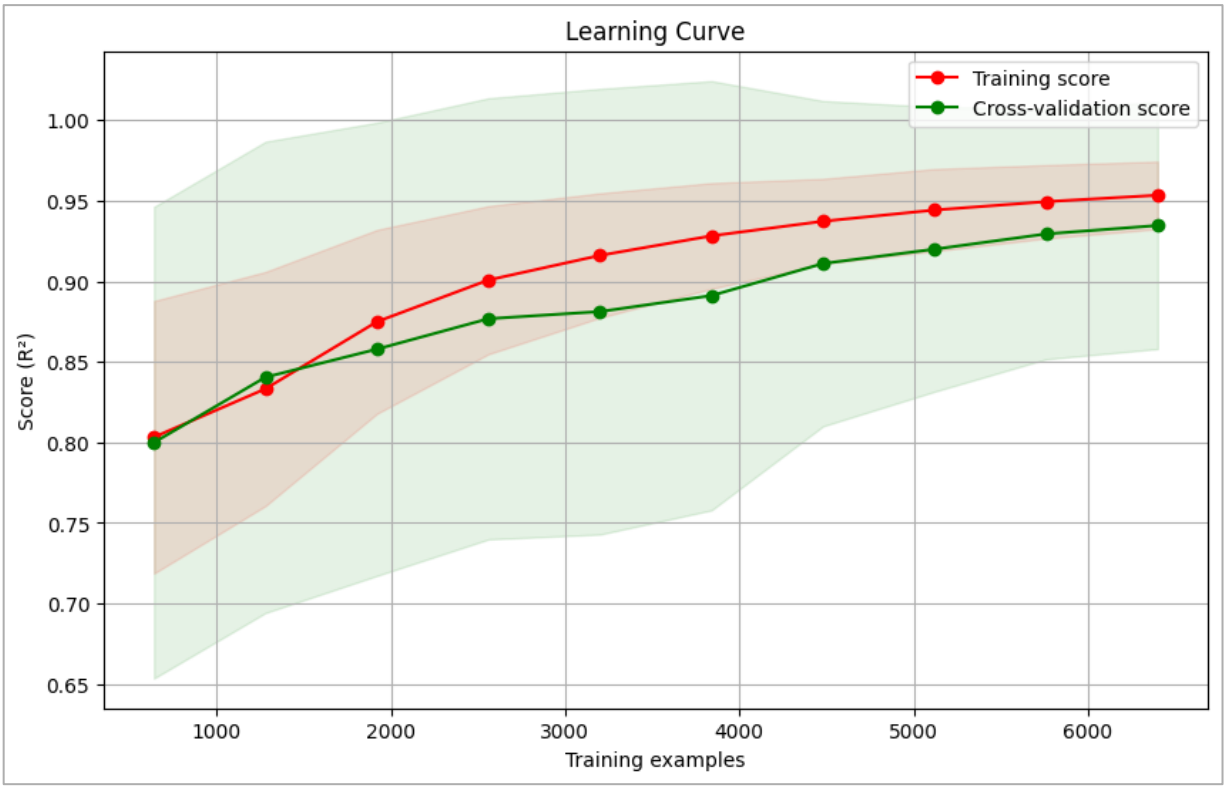
Model	R^2 Score (Cross-Validation)	Mean Absolute Error (Cross-Validation)	Standard Deviation (R^2 Scores)
Linear Regression	0.6194	9.56 GBP	0.05
Random Forest Regressor	0.9531	1.08 GBP	0.02

7.6.2. Learning Curve Analysis

The learning curve analysis for the RandomForestRegressor was conducted to visualize how the model's performance metrics evolve with increasing amounts of training data. As depicted in Figure 16, the analysis evaluates the model's performance in relation to the quantity of training data.

- **Training Score (Red Line):** This score demonstrates how well the model fits the training data, revealing an incremental improvement as more data points are utilized. This upward trend suggests that the model is effectively capturing the underlying patterns within the training set, enhancing its predictive accuracy as it learns from an increasing volume of data.
- **Cross-Validation Score (Green Line):** This score assesses the model's ability to generalize to unseen data. The upward trajectory that gradually converges toward the training score indicates the model's improving generalization capabilities. The convergence of these two scores, with the addition of more training examples, implies that the model is not only learning effectively but is also becoming more stable in its predictions across different data samples. This reduction in the variance between training and validation outcomes signifies a robust model that balances well between fitting the training data and generalizing to new data.

Figure 16. Learning Curve for RandomForestRegressor



7.6.3. ANOVA Analysis

To statistically assess the significance of the performance differences between the Linear Regression and Random Forest Regressor models, an Analysis of Variance (ANOVA) was performed. This analysis focused on comparing the variability of the R^2 scores obtained from the cross-validation process for both models, which helps in determining if the observed differences in model performance are statistically significant.

ANOVA Results: The ANOVA test yielded a p-value significantly less than the conventional alpha level of 0.05, confirming that the differences in R^2 scores between the two models are statistically significant. This outcome suggests that the superior performance metrics of the RandomForestRegressor are not due to random variations in the data but are attributable to the model's robustness and its ability to generalize effectively across diverse data sets.

Interpretation: These statistical results validate the comparative performance data presented earlier, offering robust evidence that supports the selection of the RandomForestRegressor as the more effective model for rent price prediction. The ANOVA thus provides a rigorous statistical foundation for the preference of one model over another, ensuring that the model selection process is both transparent and grounded in quantitative analysis.

Table 5. ANOVA Results

Source of Variation	Sum of Squares	Degrees of Freedom	F-Statistic	p-Value
Model Difference	0.334	1	5.27	0.022
Residual	2.76	18		

7.7. Predicting Rental Prices by LSOA

This analysis employed a sophisticated machine learning approach to forecast rental prices per square foot across London's Lower Super Output Areas (LSOAs). The process began with the acquisition of the 'LSOA Statistics.csv' dataset, which includes critical socio-economic and geographical indicators pertinent to rental market dynamics.

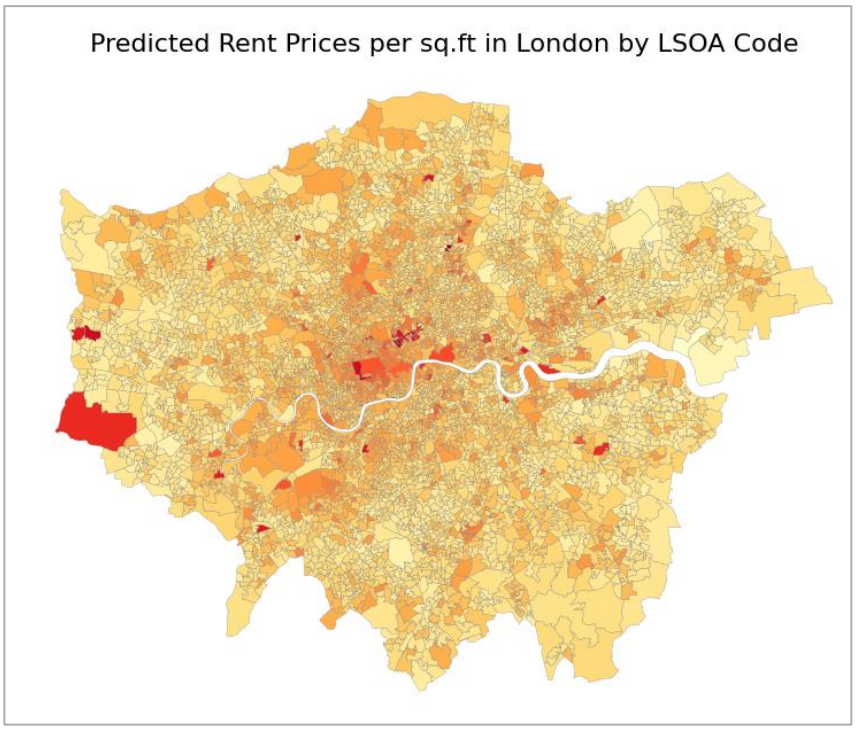
Data Preparation and Processing: The prepared dataset underwent a series of preprocessing steps, including imputation, polynomial feature transformation, and scaling, aligning it with the preprocessing protocols used during the training phase of the model. The RandomForestRegressor, which was previously optimized and rigorously validated, was then applied to predict rental prices. This model was chosen for its robustness and ability to handle varied data features effectively.

The application of this model facilitated the prediction of rental prices, which were then systematically mapped back to their corresponding LSOAs, providing a granular view of the predicted rental landscape across London.

7.8. Rent Prediction Map Visualizations

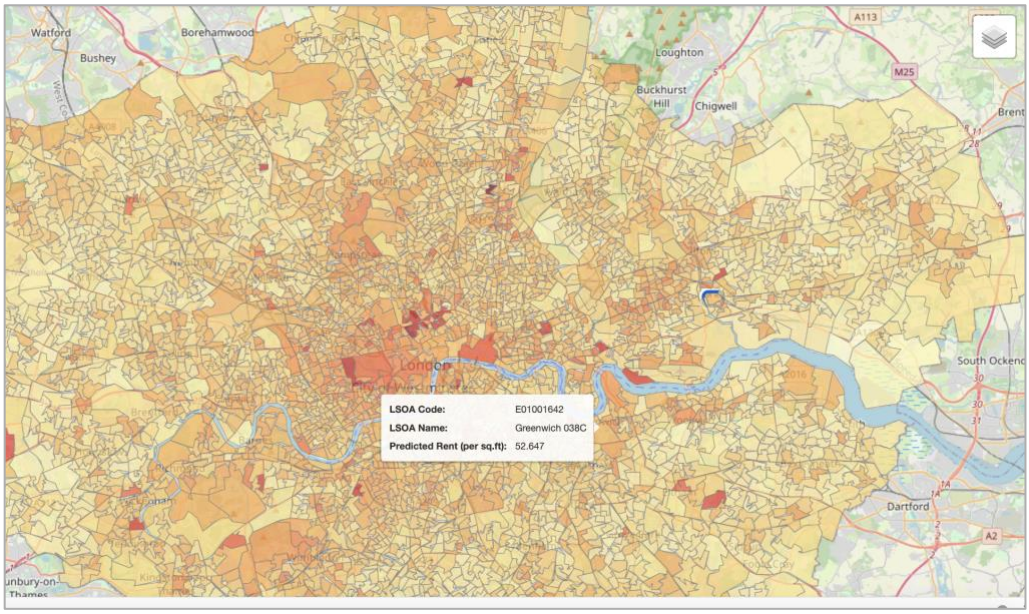
1. Following the prediction of rent prices, geospatial visualizations were created to depict the distribution of predicted rents across London. The predictive results were merged with LSOA boundary data using GeoPandas, facilitating detailed mapping of rent prices.

Figure 17. Commercial Rent Prediction Map Visualisation



2. Using Folium, an interactive map was produced, color-coded to reflect varying rent levels across the city. This map enhances the visualization by allowing users to interact with the data, offering tooltips that display specific rent prices and LSOA details for each area.

Figure 18. Commercial Rent Prediction Interactive Map Visualisation

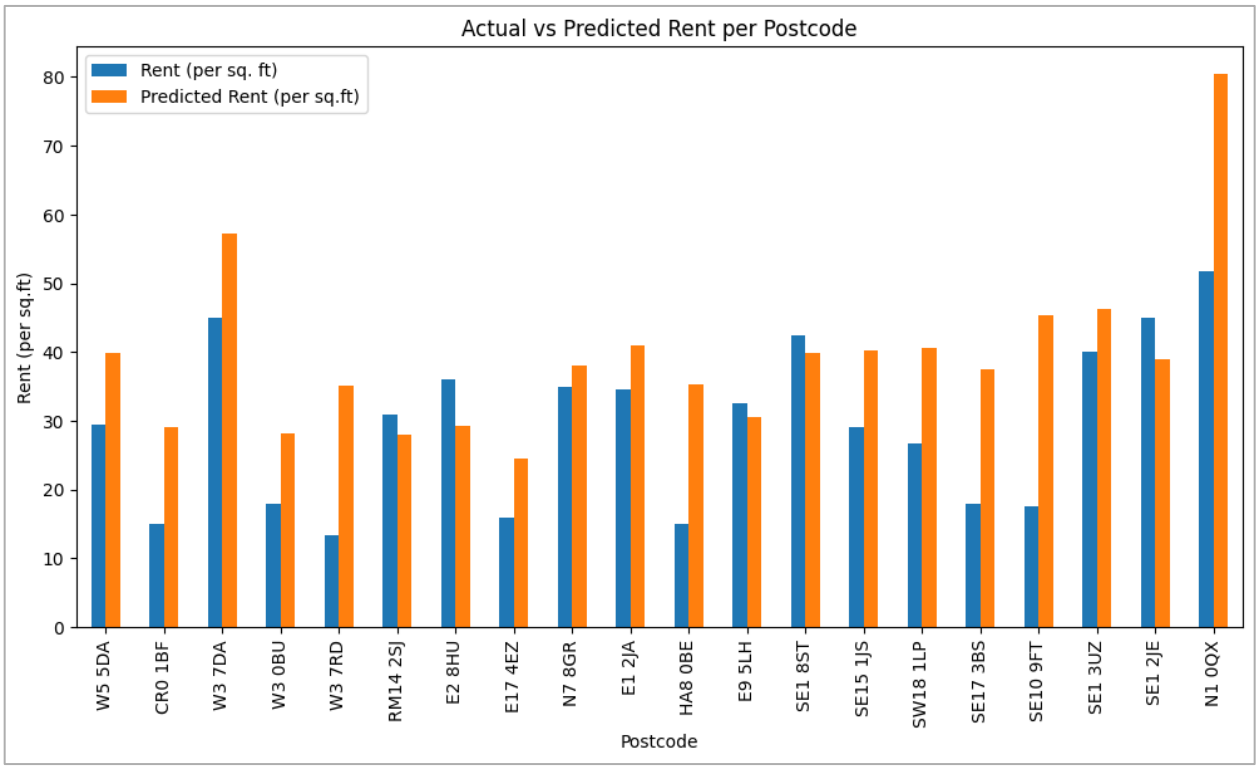


8. TASK 3: Comparative Analysis

8.1. Predicted vs. Actual Rent Evaluation

This stage integrates predicted rental prices derived from the machine learning model with actual market data to evaluate prediction accuracy. The analysis leverages a sample of 20 records, focusing on a direct comparison between predicted and actual rents. These records, merged based on 'LSOA Code', synchronize predicted rental values with actual figures obtained from market sources, creating a dataset that highlights both discrepancies and alignments between expected and actual rental values.

Figure 19. Actual vs Predicted Rent Prices per Postcode

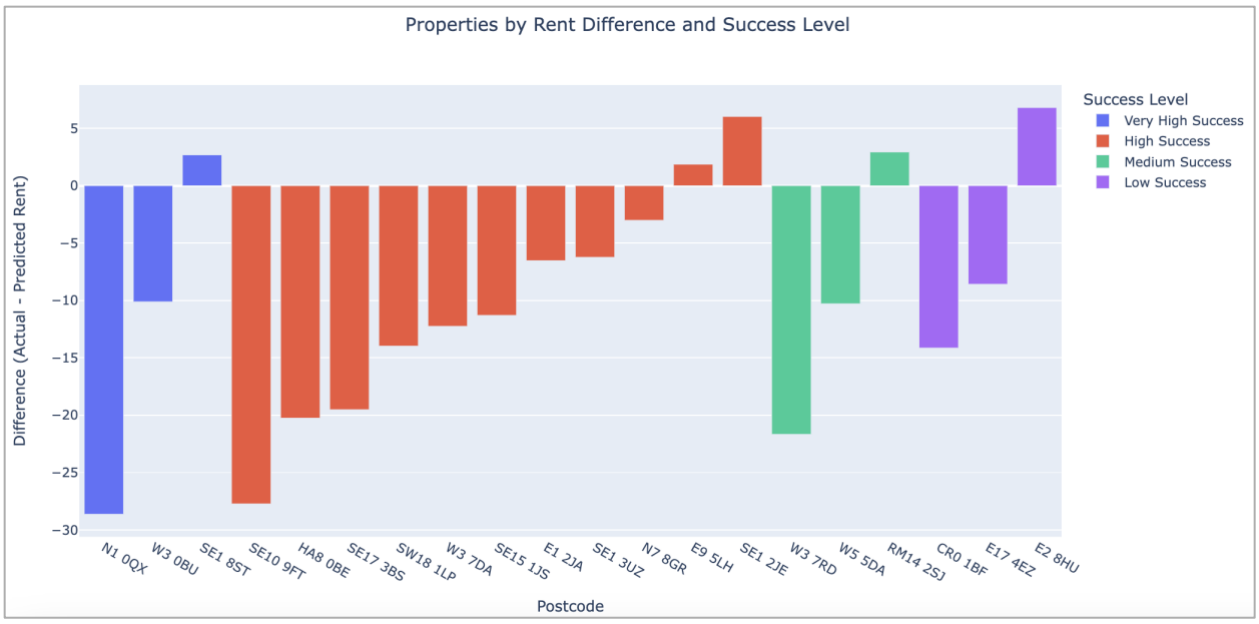


Explanation: This bar chart provides a side-by-side comparison of actual rents (market data) versus rents predicted by our machine-learning model across selected postcodes. Each pair of bars, one blue for actual rent and one orange for predicted rent per square foot, illustrates the model's accuracy in mirroring real market conditions, highlighting any consistent patterns of overestimation or underestimation across specific areas.

8.2. Visual Representation and Success Level Analysis

Differences between predicted and actual rents are quantified and classified into success levels—Very High, High, Medium, Low—based on their deviation from actual figures. This classification aids in identifying locations where predictions substantially diverge from market rents.

Figure 20. Properties by Rent Price Difference and Success Level



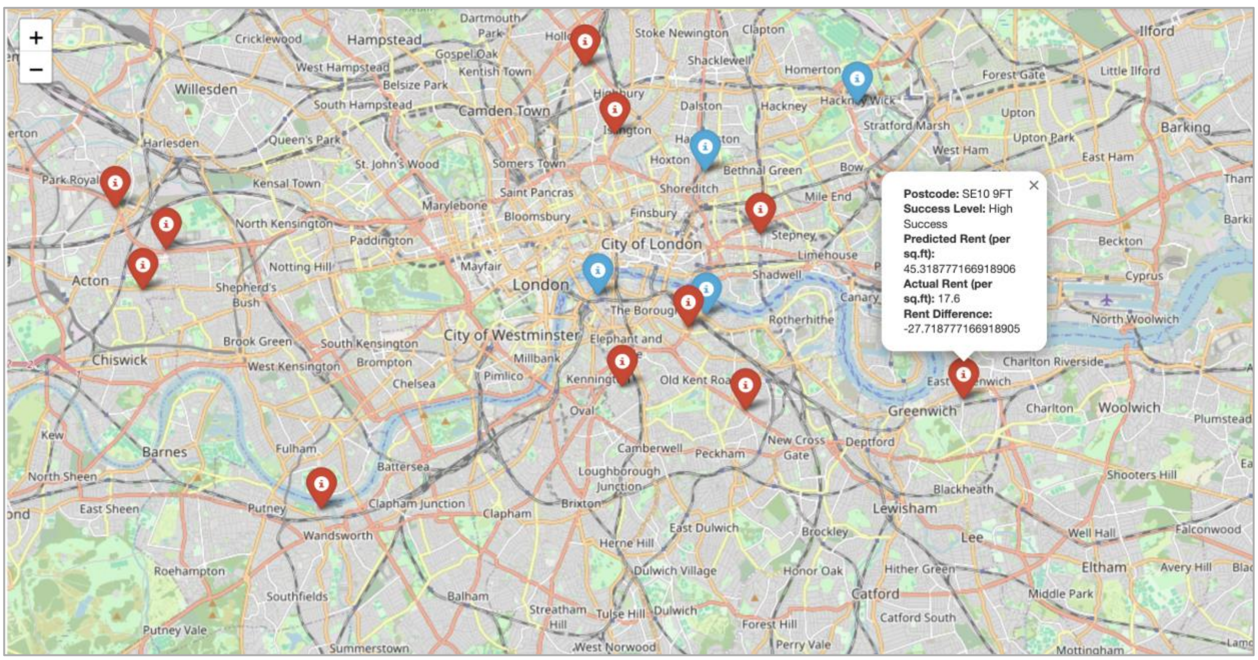
Explanation: This bar chart visualizes the "Rent Difference" across various postcodes, with each bar color-coded to indicate success levels: very high (blue), high (red), medium (green), and low (purple). Differences are based on the size and direction of rent discrepancies, where smaller or positive differences imply higher success, and larger negative differences suggest lower success. This visualization is crucial for assessing the rental prediction model's accuracy and pinpointing locations with significant discrepancies between predicted and actual rents.

- **Very High Success (Blue):** For example, the N1 0QX postcode shows a minimal negative difference, indicating a slight overestimation of rent but still categorized as very high success due to its minor deviation and strong alignment with market rates.
- **High Success (Red):** In postcodes like W3 7DA and SE1 2JA, the predictions were nearly accurate, showing modest deviations but close alignment with actual rents, providing reliable estimates.

- **Medium Success (Green):** The SE15 1JS postcode exhibited a slight positive difference, suggesting that actual rents exceeded predictions, reflecting a moderate success in mirroring market trends.
- **Low Success (Purple):** The E1 2JA postcode displayed a significant negative difference, highlighting a major overestimation of rents that could mislead stakeholders regarding investment viability.

8.3. Interactive Map Visualization

Figure 21. Interactive Map of Rent Prediction Success by Postcode



Explanation: This interactive map offers a geographic representation of rental prediction success across London, enhancing the spatial understanding of model performance. It features markers color-coded by success level and interactive capabilities, allowing users to explore details such as postcode, predicted and actual rents, and the difference in rents. This visualization not only spotlights areas where the model excels but also identifies regions needing model refinement or data adjustments.

9. Conclusions and Future Directions

9.1. Summary of Findings and Implications for Stakeholders

This research successfully addresses the complex challenge of selecting optimal cafe locations in London through a data-driven approach. By combining unsupervised learning, decision-making processes, and predictive modeling, the study has developed a comprehensive framework that balances the success potential of locations with associated rental costs. Key findings from this research have significant implications for several groups of stakeholders:

- **For Cafe Owners and Entrepreneurs:** The predictive models enable more informed decision-making, reducing the risk of business failure by identifying locations with high success potential. The success predictions, combined with the rent analysis, allow for strategic site selection that aligns with operational budgets and market positioning.
- **For Investors:** The model provides a systematic and data-informed approach for assessing investment opportunities in the cafe sector. By identifying locations where success potential is high but rent costs remain reasonable, investors can maximize return on investment while minimizing financial risk.
- **For Property Managers and Real Estate Developers:** Insights into commercial rent dynamics, derived from the predictive model, allow property managers to price properties more competitively, attracting sustainable and profitable businesses. The integration of success levels into the rent analysis offers additional metrics for adjusting rental pricing strategies.

9.2. Strategic Recommendations

The findings from this research provide several strategic recommendations that stakeholders can implement to optimize cafe site selection and enhance investment opportunities:

Real-Time Data Integration: To improve the predictive accuracy of the models, future studies should incorporate real-time data streams such as current foot traffic, real estate market trends, and social media activity related to consumer behavior. These dynamic data sources can help refine predictions by accounting for short-term fluctuations in demand and market conditions.

Cross-City Application: While the model is tailored for London's market, its methodology can be extended to other metropolitan areas. Testing and adapting the model in different cities would

validate its scalability and provide global stakeholders with insights on the viability of similar cafe markets in various urban settings.

Collaboration with Urban Planners: Cafe owners, investors, and property managers should consider collaborating with urban planners and city officials to align their strategies with ongoing urban development projects. These collaborations could lead to the identification of up-and-coming neighborhoods that offer high success potential at lower costs, benefiting both entrepreneurs and urban developers.

Focus on Emerging Neighborhoods: The model's ability to identify success potential in lower-rent areas suggests opportunities for business expansion into emerging or redeveloping neighborhoods. Entrepreneurs could capitalize on these findings by strategically targeting these areas before rent prices escalate.

9.3 Limitations and Opportunities for Future Research

While this research has generated valuable insights, certain limitations are acknowledged, providing avenues for further investigation and improvement:

Data Availability and Granularity: The study's reliance on aggregated socio-economic and demographic data limits the granularity of predictions. Incorporating more detailed data, such as consumer spending patterns, neighborhood foot traffic, and cafe-specific reviews, could enhance model precision.

Model Generalizability: The predictive models are calibrated specifically for London’s unique urban landscape and market conditions. Future research should focus on applying and adjusting the model for different geographical and economic contexts to test its robustness and adaptability. Extending the methodology to other sectors, such as retail or hospitality, could provide broader applications of the model.

Handling of Data Scarcity: In this study, semi-supervised learning techniques like pseudo-labeling were used to compensate for the lack of labeled data. While effective, this method could benefit from further refinement, particularly by exploring other semi-supervised techniques like self-training or active learning to enhance predictive accuracy for unlabeled data.

9.4 Personal Reflection and Contribution to the Field

Reflecting on the research process, this study has demonstrated the growing importance of interdisciplinary, data-driven approaches in urban planning and business strategy. The integration of machine learning, decision theory, and real estate economics has not only advanced academic understanding but also offers tangible, real-world applications for cafe owners, investors, and property managers.

The methodologies developed and employed in this research provide a scalable, replicable framework for addressing similar challenges in other sectors. As the cafe industry, like many other retail sectors, becomes increasingly data-centric, the application of predictive analytics will be essential for strategic decision-making. By adopting data-driven models such as those developed in this study, stakeholders can navigate the complexities of urban business environments with greater confidence and foresight.

Moreover, the research journey has underscored the significance of continuous adaptation and learning in response to emerging data and technologies. As machine learning techniques evolve, so too must their application to urban planning and commercial development.

9.5 Final Remarks

In conclusion, this study provides a comprehensive, data-centric framework that addresses the practical challenges of selecting optimal cafe locations in a dynamic urban environment like London. Through a combination of machine learning models, decision-making techniques, and socio-economic analysis, the study offers a robust solution that balances both business success potential and rent affordability.

The findings of this research hold the potential to transform how stakeholders approach cafe site selection and broader urban commercial investments. The predictive model's capacity to integrate various data sources and provide actionable insights makes it a valuable tool not only for the cafe industry but also for real estate development and urban planning sectors. Going forward, continuous refinement of these models, alongside the inclusion of real-time data, will further enhance their predictive power, ultimately driving more informed, data-backed decision-making.

References

Aboulola, O. (2018). GIS Spatial Analysis: A New Approach to Site Selection and Decision Making for Small Retail Facilities. [online] ResearchGate. Available at: https://www.researchgate.net/publication/325536724_GIS_Spatial_Analysis_A_New_Approach_to_Site_Selection_and_Decision_Making_for_Small_Retail_Facilities.

Adams, B. and Verbrugge, R.J. (2021). Location, Location, Structure Type: Rent Divergence within Neighborhoods. *Working paper*. doi:https://doi.org/10.26509/frbc-wp-202103.

Adebayo, A.A., Greenhalgh, P., Muldoon-Smith, K. and Oyedokun, T. (2022). Towards Attaining Sustainable Retail Property Locations: The Relationships between Supply, Demand, and Accessibility of Retail Spaces. *Sustainability*, 14(7), p.3846. doi:https://doi.org/10.3390/su14073846.

An, H.-Y., Paik, J.-K. and Hong, W.-S. (2013). Importance and Performance Analysis of Customers' Selection Attributes for Social Enterprises Type Cafe. *Korean Journal of Food and Cookery Science*, 29(6), pp.637–645. doi:https://doi.org/10.9724/kfcs.2013.29.6.637.

Apify. (2024). Google Maps Extractor. Available at: https://apify.com/compass/google-maps-extractor.

Apify. (2024). Google Places Scraper. Available at: https://apify.com/compass/crawler-google-places.

Apify. (2024). TripAdvisor Scraper. Available at: https://apify.com/maxcopell/tripadvisor.

Asadzadeh, A., Sikder, S., Mahmoudi, F. and Kötter, T. (2014). Environmental Management and Sustainable Development ISSN. *Environmental Management and Sustainable Development*, 3(1). doi:https://doi.org/10.5296/emsd.v3i1.4874.

Bera, A.K. and Kangalli Uyar, S.G. (2019). Local and global determinants of office rents in Istanbul. *Journal of European Real Estate Research*, 12(2), pp.227–249. doi:https://doi.org/10.1108/jerer-12-2018-0052.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13, pp.1063–1095. doi:https://doi.org/10.5555/2503308.2343682.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), pp.197–227. doi:https://doi.org/10.1007/s11749-016-0481-7.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.

Brotman, B.A. (2022). Property investor decisions using income and rental ratio signals. *Journal of Property Investment & Finance*, 40(1), pp.2–13. doi:https://doi.org/10.1108/jpif-03-2020-0031.

Carr, L.J., Dunsiger, S.I. and Marcus, B.H. (2010). Validation of Walk Score for estimating access to walkable amenities. *British Journal of Sports Medicine*, 45(14), pp.1144–1148. doi:https://doi.org/10.1136/bjsm.2009.069609.

Čeh, M., Kilibarda, M., Lisec, A. and Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS International Journal of Geo-Information*, 7(5), p.168. doi:https://doi.org/10.3390/ijgi7050168.

Chatterjee, D. and Mukherjee, B. (2013). Potential Hospital Location Selection using AHP: A Study in Rural India. *International Journal of Computer Applications*, 71(17), pp.1–7. doi:https://doi.org/10.5120/12447-9144.

Croce, P.R., Azevedo, L.D.M., Hora, H.R.M. da and Morais, A.S.C. (2020). Three-stage location decision model for a retail point: a multicriteria AHP approach. *Revista Mundi Engenharia, Tecnologia e Gestão*, 5(2). doi:https://doi.org/10.21575/25254782rmetg2020vol5n21139.

Data.gov.uk. (2015). Median Household Income 2015. Available at: https://data.world/datagov-uk/5c4a083f-a8c6-42d8-ad40-36a9719a634c.

Doogal. (2024). London Postcodes Dataset. Available at: https://www.doogal.co.uk/london_postcodes.

Doogal. (2024). London Stations Dataset. Available at: https://www.doogal.co.uk/london_stations#google_vignette.

Dong, L., Ratti, C. and Zheng, S. (2019). Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proceedings of the National Academy of Sciences*, 116(31), pp.15447–15452. doi:https://doi.org/10.1073/pnas.1903064116.

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, pp.226–231. doi:https://dl.acm.org/doi/10.5555/3001460.3001507.

Fauzi, A., Indriyani, N. and Bayu Hasta Yanto, A. (2021). Selection of Coffee Shop Business Locations Using the Analytical Hierarchy Process Method. *Jurnal Teknologi Dan Open Source*, 4(2), pp.133–140. doi:https://doi.org/10.36378/jtos.v4i2.1771.

Forsyth, A., Wall, M., Larson, N., Story, M. and Neumark-Sztainer, D. (2012). Do adolescents who live or go to school near fast-food restaurants eat more frequently from fast-food restaurants? *Health & Place*, 18(6), pp.1261–1269. doi:https://doi.org/10.1016/j.healthplace.2012.09.005.

Garang, Z., Wu, C., Li, G., Zhuo, Y. and Xu, Z. (2021). Spatio-Temporal Non-Stationarity and Its Influencing Factors of Commercial Land Price: A Case Study of Hangzhou, China. *Land*, 10(3), p.317. doi:https://doi.org/10.3390/land10030317.

Greater London Authority. (2012). Statistical GIS Boundary Files for London. Contains National Statistics data © Crown copyright and database right. Available at: https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london.

Greater London Authority. (2014). LSOA Atlas: Demographics. Available at: https://data.london.gov.uk/dataset/lsoa-atlas.

Günen, M.A. (2021). Evaluation of GIS based Ranking and AHP methods in selecting the most suitable site: A case study in Kayseri, Turkey. *Research Square*. doi:https://doi.org/10.21203/rs.3.rs-239049/v1.

Hasyim, A., Kurniawan, E. and Purnamasari, W. (2022). Cafe market share using satellite image data and Google Database in Malang City. *Civil and Environmental Science*, 005(01), pp.055–063. doi:https://doi.org/10.21776/ub.civense.2022.00501.6.

Hsiao, Y.-H. and Chen, G.-T. (2020). Listening to Customer Kansei for Restaurant Location Evaluation. *Journal of Hospitality & Tourism Research*, 44(4), pp.666–693. doi:https://doi.org/10.1177/1096348020919024.

Hu, G. and Tang, Y. (2023). GERPM: A Geographically Weighted Stacking Ensemble Learning-Based Urban Residential Rents Prediction Model. *Mathematics*, 11(14), pp.3160–3160. doi:https://doi.org/10.3390/math11143160.

Huang, M. (2020). Theory and Implementation of linear regression. *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*. doi:https://doi.org/10.1109/cvidl51233.2020.00-99.

Huang Zhe-xue. (2013). A Brief Theoretical Overview of Random Forests. *Journal of Integration Technology*.

Humaira, H. and Rasyidah, R. (2020). Determining The Appropriate Cluster Number Using Elbow Method for K-Means Algorithm. *Proceedings of the 2nd Workshop on Multidisciplinary and*

Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia. doi:https://doi.org/10.4108/eai.24-1-2018.2292388.

Ibrahim, G.R.F. (2021). Multi Criteria Decision Analysis Technique for Solar Power Sites Selection in Duhok Governorate – Iraq. *Research Square*. doi:https://doi.org/10.21203/rs.3.rs-818565/v1.

Iraldo, F., Testa, F., Lanzini, P. and Battaglia, M. (2017). Greening competitiveness for hotels and restaurants. *Journal of Small Business and Enterprise Development*, 24(3), pp.607–628. doi:https://doi.org/10.1108/jsbed-12-2016-0211.

Jin Seung Jung, Kim, J. and Jin, C. (2022). Does Machine Learning Prediction Dampen the Information Asymmetry for Non-Local Investors? *International Journal of Strategic Property Management*, 26(5), pp.345–361. doi:https://doi.org/10.3846/ijspm.2022.17590.

Kuhn, V.R., Benetti, A.C., Anjos, S.J.G. dos and Limberger, P.F. (2018). Food services and customer loyalty in the hospitality industry. *Tourism & Management Studies*, 14(2), pp.26–35. doi:https://doi.org/10.18089/tms.2018.14203.

Li, S. (2020). An Improved DBSCAN Algorithm Based on the Neighbor Similarity and Fast Nearest Neighbor Query. *IEEE Access*, 8, pp.1–1. doi:https://doi.org/10.1109/access.2020.2972034.

Li, W. and Xiao, D. (2021). The Sustainable-based Impacts of Built Environmental Influencing Factors on Price-rent Ratio: A Case Study in Guangzhou. *Proceedings of the 57th ISOCARP World Planning Congress*, 13(1), pp.1–15. doi:https://doi.org/10.47472/txlfqttz.

Likas, A., Vlassis, N. and Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36(2), pp.451–461. doi:https://doi.org/10.1016/s0031-3203(02)00060-2.

Liu, Y. (2017). A commercial real estate price evaluation model based on GT-BCPSO-BP neural network. *International Journal of Applied Decision Sciences*, 10(4), p.335. doi:https://doi.org/10.1504/ijads.2017.087177.

Liu, Y., Li, W., Liu, G., Yang, X., Guo, Y. and Zhang, K. (2021). Influence of Places of Resident Activities on Spatial Distribution of Drug-Related Crimes. *Social Sciences*, 10(3), p.101. doi:https://doi.org/10.11648/j.ss.20211003.14.

Metropolitan Police Service. (2023). Recorded Crime Summary by LSOA. Available at: https://data.london.gov.uk/dataset/recorded_crime_summary.

Mao, X., Zhao, X., Lin, J. and Herrera-Viedma, E. (2019). Utilizing multi-source data in popularity prediction for shop-type recommendation. *Knowledge-Based Systems*, 165, pp.253–267. doi:https://doi.org/10.1016/j.knosys.2018.11.033.

Masjedi, M.R., Taghizadeh, F., Hamzehali, S., Ghaffari, S., Fazlzadeh, M., Jafari, A.J., Niazi, S., Mehrizi, E.A., Moradi, M., Pasalari, H. and Arfaeinia, H. (2019). Air pollutants associated with smoking in indoor/outdoor of waterpipe cafés in Tehran, Iran: Concentrations, affecting factors and health risk assessment. *Scientific Reports*, 9(1). doi:https://doi.org/10.1038/s41598-019-39684-3.

Maulud, D. and Abdulazeez, A.M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, 1(4), pp.140–147. doi:https://doi.org/10.38094/jastt1457.

Miao, J., Wang, P. and Zha, T. (2020). Discount Shock, Price-Rent Dynamics, and the Business Cycle. *Federal Reserve Bank of Atlanta, Working Papers*. doi:https://doi.org/10.29338/wp2020-07.

Mishra, S. (2019). Why do 74% of independent coffee shops fail in the first five years. [online] Medium. Available at: https://fabled.medium.com/why-do-74-of-independent-coffee-shops-fail-in-the-first-five-years-8fe19fc20f70.

Office for National Statistics. (2023). Median House Prices by LSOA. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/medianpricepaidbylowerlayersuperoutputareahpssadataset46.

OnTheMarket. (2024). Commercial Property to Rent in London. Available at: https://www.onthemarket.com/to-rent/commercial/property/london/.

Ouyang, J., Fan, H., Wang, L., Yang, M. and Ma, Y. (2020). Site Selection Improvement of Retailers Based on Spatial Competition Strategy and a Double-Channel Convolutional Neural Network. *ISPRS International Journal of Geo-Information*, 9(6), p.357. doi:https://doi.org/10.3390/ijgi9060357.

Papachristos, A.V., Smith, C.M., Scherer, M.L. and Fugiero, M.A. (2011). More Coffee, Less Crime? The Relationship between Gentrification and Neighborhood Crime Rates in Chicago, 1991 to 2005. *City & Community*, 10(3), pp.215–240.

Parsa, H.G., Self, J.T., Njite, D. and King, T. (2005). Why Restaurants Fail. *Cornell Hotel and Restaurant Administration Quarterly*, 46(3), pp.304–322. doi:https://doi.org/10.1177/0010880405275598.

Prayag, G., Landré, M. and Ryan, C. (2012). Restaurant location in Hamilton, New Zealand: clustering patterns from 1996 to 2008. *International Journal of Contemporary Hospitality Management*, 24(3), pp.430–450. doi:https://doi.org/10.1108/09596111211217897.

PropertyLink. (2024). Retail Properties for Rent in London. Available at: https://propertylink.estatesgazette.com/retail-for-rent/london.

Rosenthal, S.S. and Ross, A. (2010). Violent crime, entrepreneurship, and cities. *Journal of Urban Economics*, 67(1), pp.135–149. doi:https://doi.org/10.1016/j.jue.2009.09.001.

Sari, W.M., Fitria, F.L., Suharto, E., Ikhwani, Y., Wagino, W., Alamsyah, N. and Windarto, A.P. (2020). Improving the Quality of Management with the Concept of Decision Support Systems in Determining Factors for Choosing a Cafe based on Consumers. *Journal of Physics*, 1471(1), pp.012009–012009. doi:https://doi.org/10.1088/1742-6596/1471/1/012009.

Schubert, E., Sander, J., Ester, M., Kriegel, H.P. and Xu, X. (2017). DBSCAN Revisited, Revisited. *ACM Transactions on Database Systems*, 42(3), pp.1–21. doi:https://doi.org/10.1145/3068335.

Scornet, E., Biau, G. and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4), pp.1716–1741. doi:https://doi.org/10.1214/15-aos1321.

Shaikh, S.A., Memon, M.A., Prokop, M. and Kim, K. (2020). An AHP/TOPSIS-Based Approach for an Optimal Site Selection of a Commercial Opening Utilizing GeoSpatial Data. *International Conference on Big Data and Smart Computing*. doi:https://doi.org/10.1109/bigcomp48618.2020.00-58.

Smolic, H. (2024). The Importance of Normalization in Machine Learning. [online] Graphite Note. Available at: https://graphite-note.com/the-importance-of-normalization-in-machine-learning/#:~:text=Min%2DMax%20normalization%20rescales%20the.

Stacy, C., Davis, C., Freemark, Y.S., Lo, L., MacDonald, G., Zheng, V. and Pendall, R. (2023). Land-use reforms and housing costs: Does allowing for increased density lead to greater affordability? *Urban Studies*, 60(14), p.004209802311595. doi:https://doi.org/10.1177/00420980231159500.

Su, X., Yan, X. and Tsai, C.-L. (2019). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3), pp.275–294. doi:https://doi.org/10.1002/wics.1198.

Subramanian, N. and Ramanathan, R. (2012). A review of applications of Analytic Hierarchy Process in operations management. *International Journal of Production Economics*, 138(2), pp.215–241. doi:https://doi.org/10.1016/j.ijpe.2012.03.036.

Susilo, A. (2020). Identifying Factors that Affect Consumer Satisfaction of Parklatz Café in Ponorogo City, East Java, Indonesia: An Application of Exploratory Factor Analysis. *Falah: Jurnal Ekonomi Syariah*, 5(1). doi:https://doi.org/10.22219/jes.v5i1.11399.

Tavakkoli-Moghaddam, R., Hassanzadeh, S. and Zhang, G. (2010). A Proposed Decision Support System for Location Selection Using Fuzzy Quality Function Deployment. *InTech*.

Tavakkoli-Moghaddam, R., Sotoudeh-Anvari, A. and Siadat, A. (2015). A Multi-criteria Group Decision-Making Approach for Facility Location Selection Using PROMETHEE under a Fuzzy Environment. In: *Lecture Notes in Business Information Processing*. Springer, Cham, pp.145–156. Available at: https://doi.org/10.1007/978-3-319-19515-5_12.

Tayman, J. and Pol, L. (2011). Retail Site Selection And Geographic Information Systems. *Journal of Applied Business Research (JABR)*, 11(2), p.46. doi:https://doi.org/10.19030/jabr.v11i2.5874.

Terada, Y. (2014). Strong consistency of factorial K-means clustering. *Annals of the Institute of Statistical Mathematics*, 67(2), pp.335–357. doi:https://doi.org/10.1007/s10463-014-0454-0.

Tomal, M. and Helbich, M. (2022). The private rental housing market before and during the COVID-19 pandemic: A submarket analysis in Cracow, Poland. *Environment and Planning B: Urban Analytics and City Science*, 49(5), p.57. doi:https://doi.org/10.1177/23998083211062907.

Valentina, F. and Arini, E. (2023). Influence Of Service Quality, Location and Facilities For The Decision To Visit Back At The Promise Soul Cafe, Bengkulu City. *Jurnal Ekonomi Manajemen Akuntansi dan Keuangan*, 4(3). doi:https://doi.org/10.53697/emak.v4i3.1284.

Wibisono, Y.Y. and Marella, S. (2020). A decision making model for selection of café location: An ANP approach. *Journal of Physics: Conference Series*, 1477(1), p.052030. doi:https://doi.org/10.1088/1742-6596/1477/5/052030.

Wong, M.S., Peyton, J., Shields, T.W., Curriero, F.C. and Gudzune, K.A. (2017). Comparing the accuracy of food outlet datasets in an urban environment. *Geospatial Health*, 12(1). doi:https://doi.org/10.4081/gh.2017.546.

Zhang, S., Wang, L. and Lu, F. (2019). Exploring Housing Rent by Mixed Geographically Weighted Regression: A Case Study in Nanjing. *ISPRS International Journal of Geo-Information*, 8(10), p.431. doi:https://doi.org/10.3390/ijgi8100431.

Zhang, X. and Zhou, S. (2023). WOA-DBSCAN: Application of Whale Optimization Algorithm in DBSCAN Parameter Adaption. *IEEE Access*, 11, pp.91861–91878. doi:https://doi.org/10.1109/access.2023.3307412.

Zhao, J., Zong, B. and Wu, L. (2023). Site Selection Prediction for Coffee Shops Based on Multi-Source Space Data Using Machine Learning Techniques. *ISPRS International Journal of Geo-Information*, 12(8), p.329. doi:https://doi.org/10.3390/ijgi12080329.

Zeng, Q., Zhong, M., Zhu, Y. and Li, J. (2020). Business Location Selection Based on Geo-Social Networks. *Database Systems for Advanced Applications*, 12114, pp.36–52. doi:https://doi.org/10.1007/978-3-030-59419-0_3.