engineering proceedings

MDPI

Proceeding Paper

Geopositional Data Analysis Using Clustering Techniques to Assist Occupants in a Specific City [†]

Sneha George, Jayakumar Keirolona Safana Seles, Duraipandi Brindha, Theena Jemima Jebaseeli * and Laya Vemulapalli

Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore 641114, India; snehageorge@karunya.edu (S.G.); keirolonasafana@karunya.edu (J.K.S.S.); brindha@karunya.edu (D.B.); layavemulapalli@karunya.edu.in (L.V.)

* Correspondence: jemima_jeba@karunya.edu

[†] Presented at the International Conference on Recent Advances on Science and Engineering, Dubai, United Arab Emirates, 4–5 October 2023.

Abstract: Geolocation and Geographic Information Systems (GIS) are becoming essential tools in several sectors. Clustering-based geopositional data analysis has enormous potential for helping the citizens of a given city. The insights gained from this kind of study can assist inhabitants and tourists in making better-educated decisions and improve overall quality of life by shedding light on numerous facets of the city's infrastructure, services, and facilities. Due to its capacity to combine databases and display geographic data, GIS has proven important in a variety of industries. City planners and other stakeholders may learn a lot about the requirements of the city's residents by clustering geopositional data. Making wise judgments based on this knowledge will raise the standard of living for everyone who lives, works, and visits the city. The purpose of this research is to use k-means clustering to identify the best houses to live in for immigrants according to their expectations, amenities, price, and proximity to the workplace or educational institution, and provide them with the best accommodation suggestions. After gathering the geolocational data of the city to which the immigrants have moved, the details will be cleaned and the data will be analyzed using different data pre-processing and data exploratory techniques. At last, the data will be clustered using the k-means clustering algorithm. It is computationally efficient and operates perfectly when clusters are spherical and comparable in size. It is essential to handle data privacy and security properly while working with geopositional data. The quality of life for those who live in cities can be improved by utilizing clustering algorithms to analyze geopositional data.

Keywords: Geographic Information Systems; immigrant accommodation; data analysis; K-Means; clustering; geolocation

check for updates

Citation: George, S.; Seles, J.K.S.; Brindha, D.; Jebaseeli, T.J.; Vemulapalli, L. Geopositional Data Analysis Using Clustering Techniques to Assist Occupants in a Specific City. *Eng. Proc.* **2023**, *59*, 8. https://doi.org/10.3390/engproc2023059008

Academic Editors: Nithesh Naik, Rajiv Selvam, Pavan Hiremath, Suhas Kowshik CS and Ritesh Ramakrishna Bhat

Published: 11 December 2023

1. Introduction

Information about places is stored, organized, and analyzed using a computerized system called a GIS. Making decisions and creating maps is certainly possible with this kind of information. Maps of various regions, such as towns, forests, or countries, are often made using GIS. Additionally, it is often used to analyze spatial data, including business location information. The allocation of land areas for residential, commercial, or industrial applications is made easier by GIS. Geolocational data collection is on the rise and presents great opportunities for technological advancements and the development of new information sources like social media and mobile devices. Analyzing geopositional data requires looking at and drawing conclusions from information that has spatial coordinates, such as latitude and longitude.

According to Kamel, Boulos, and Geraghty, the earliest map portraying the many geographic locations that affected people's health was created in 1694 as part of Italy's attempts to combat the plague pandemic [1]. According to the research, distinct geographical factors influence the location and transmission of illnesses. Maps have been used to explore and track the spread of cholera, influenza, and yellow fever since 1918 [2]. GIS is a powerful tool for combining geographic information from databases for better data analysis and map visualization [3,4]. Sarfo et al. discovered that 28.7% of the research in the GIS-related health field was specifically for tracking infectious illnesses [5]. Within a GIS, maps and other organized data are shown in concert via one user interface. An early examination of the convergence of numerous data layers, including statistical, genomic, geospatial, and epidemiologic, within a unique analytical tool that represents the COVID-19 problem using design and mapping methodologies was conducted [6]. GIS has been used in tourism research to look at the rise of tourism, identify tourist hotspots, and rate places, which indicates that GIS can be useful in promoting and operating tourist attractions. Data about transportation tend to be exclusive regarding transportation. Walking, biking, public transportation, railways, and navigable waterways are the six modes of transportation covered in Designing Geodatabases for Transportation [7].

The Centre for Systems Science and Engineering (CSSE) has created the initial geo-dashboard for COVID-19 instances for analyzing the data obtained [8]. It is challenging to provide data on the intended demographic that is accurate enough to be effective. Many technological and privacy constraints must be overcome in to collect data of excellent quality at a high level of spatial resolution. GIS offers a secure and comfortable living environment for people. Additionally, the location should be accessible, easy to reach, and alluring to potential customers, with all of the amenities. People would also be looking at the surrounding environment before considering the place [9,10]. The big data analytics technique to enhance strategic decision-making in tourist destination management using a design science research methodology is described by Miah et al. [11]. The introduction of social media platforms with geolocation functionality has ushered in a new era of data-driven study, allowing us to dive deeper into the complex connection between social contacts and activity preferences [12,13]. Open Street Map (OSM), which was created by a collaborative effort from volunteers all around the world, distinguishes amongst the existing datasets as a crowd-sourced geographical database [14,15].

Challenges

The use of geopositional data analysis to benefit city dwellers has several difficulties that must be overcome. These difficulties are listed as follows.

i. Data privacy: geopositional data may be utilized to follow people's movements and is frequently sensitive.
ii. Data quality: geopositional data might be unreliable and noisy.
iii. Scalability: the size and complexity of geopositional data can vary greatly.
iv. Interpretation: interpreting the outcomes of geopositional data analysis can be challenging. It is crucial to employ strategies to visualize and explain the findings.

2. Materials

The dataset used for the experimental setup of the proposed system consists of the details of people who cook, eat out, are employed, eat ethnic food, exercise, have an income, are on and off campus, their bhk, their rent, size, floor, area locality, city, tenant preference, bathroom, etc. [16]. The additional data are gathered from undergraduates and skilled employees, as the effort seeks to discover the best housing for immigrants.

3. Methodology

The primary objectives of the proposed study are to focus on the following.

i. To obtain information about rental homes that are currently available, their geographical location, cost, dimensions, and facilities.
ii. To use the geopositional data to identify a city's most popular neighborhoods for renters as a great way to improve the property search process.

iii. To use the geopositional data to determine the most well-liked spots for tenants in a city.

Figure 1 depicts the proposed system's design. To prepare the datasets for analysis, we cleaned them using different data-cleaning approaches.

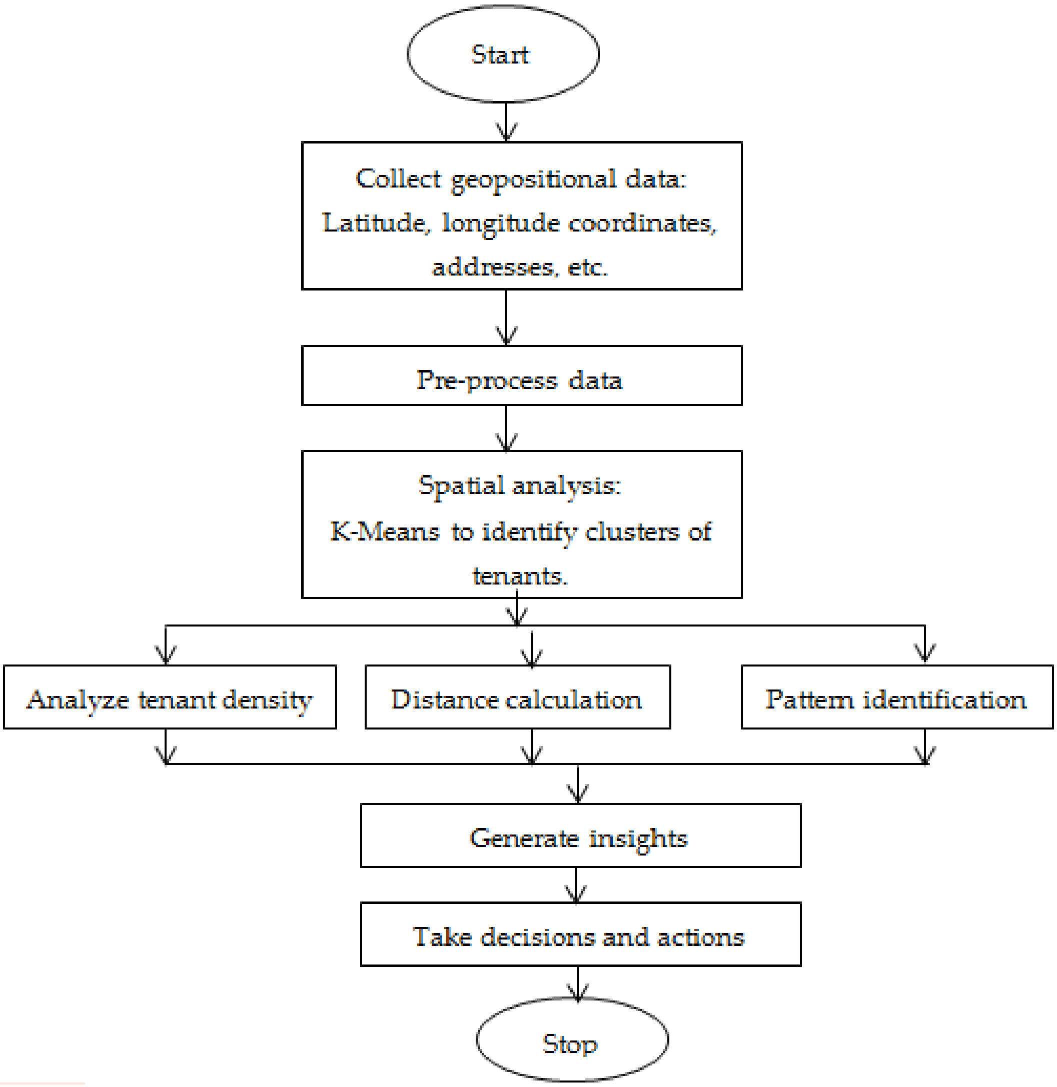

Start
Collect geopositional data: Latitude, longitude coordinates, addresses, etc.
Pre-process data
Spatial analysis: K-Means to identify clusters of tenants.
Analyze tenant density
Distance calculation
Pattern identification
Generate insights
Take decisions and actions
Stop

Figure 1. The proposed system's architecture for analyzing the geopositional data of tenants in a certain city.

Cross-validation approaches managing or imputing outliers can gauge the effect on model performance and confirm that the chosen procedures do not cause bias. ScikitLearn's K-Means clustering was used to generate the cluster location. Finally, we used Seaborn/Folium to explore the locations on the map. This step included removing null values, inconsistencies, merging multiple datasets, handling missing values, and encoding categorical variables.

Standard box plots, which show the distribution of a dataset, including the median, quartiles, and probable outliers, may be made using Seaborn and Matplotlib. The box plots in Figure 2 are the visual representation of differences, similarities, and potential outliers for every column in the dataset.

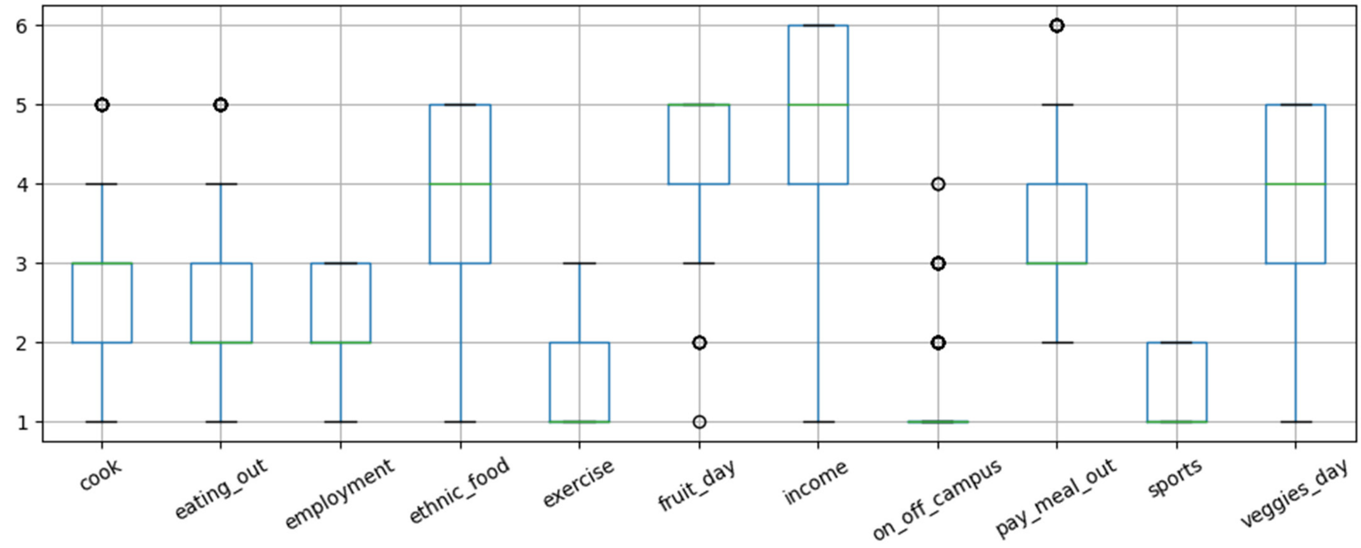


Figure 2. Box plot of the data frame.

The geolocational data was collected using REST (Representational State Transfer) APIs and Foursquare API. Access to location-based information, such as locations, venues, and user check-ins, was made possible using the Foursquare APIs. Finally, the server created a response providing the requested geolocational data, often in JSON or XML format.

3.1. Pseudocode to Obtain Geolocational Data from API

The geolocational data obtained from the Foursquare API can be used for various purposes, including location-based services, mapping applications, recommendations, and analyzing patterns and trends in user behavior. It provides a valuable resource for accessing up-to-date and comprehensive geospatial information that can enhance the functionality and user experience of applications that rely on location-based data. It is expressed as follows.

1. Define the API endpoint and necessary parameters:
   - apiEndpoint = https://api.example.com/geolocation (accessed on 11 July 2023)
   - apiKey = "YOUR_API_KEY"
   - location = "Chennai City"
   - request_date = "2023-7-11"
2. Construct the API request URL with parameters:
   - apiUrl = apiEndpoint + "?key=" + apiKey + "&location=" + location
3. Send an HTTP GET request to the API
   - response = HTTP_GET(apiUrl)
4. Check the status code:
   - if status_code == 200:
   - data = response.json() # Parse the JSON response
   - latitude = data.latitude
   - longitude = data.longitude
   - locationName = data.location_name

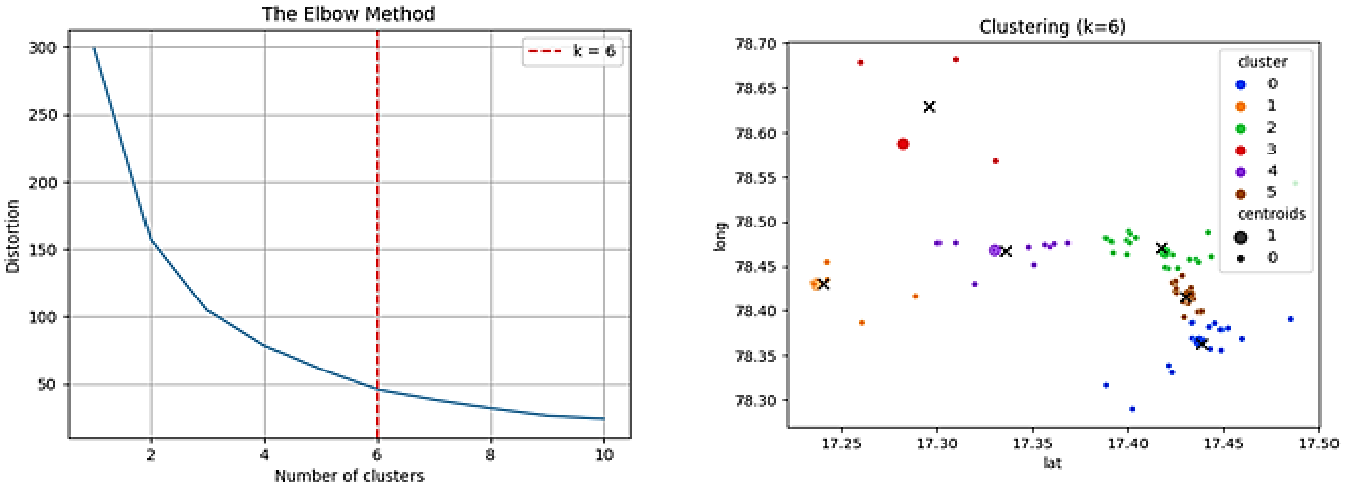
There are various clustering methods, and each has pros and cons. K-means is ideal for high-dimensional data since it is quite quick and effective. The process begins by selecting k data points at random. Centroids are the names for these points. The cluster with the nearest centroid is then given to each data point. After that, the centroids are modified, and the procedure is repeated until the clusters stop changing.

3.2. Mathematical Modeling for K-Means Clustering

A more advanced clustering method that may be used to discover clusters of various sizes and forms is DBSCAN. It operates by locating groups of dense points or core points. A hierarchy of clusters is created using the recursive clustering technique known as hierarchical clustering. At first, each data point is in its cluster. The next step is to combine the two nearest clusters, and so on, until there is just one cluster remaining. The method that could be used to locate clusters of various sizes and forms is hierarchical clustering. The k-means algorithm is described mathematically using the following expressions:

1. Initialization:
    - Choose the number of clusters, k.
    - Initialize the cluster centroids, denoted by μ1, μ2, . . ., μk.
2. Assignment Step:
    - Calculate distance to each centroid: d(xi, μk) = ||xi − μk||, where ||.|| represents a distance metric;
    - Cluster C(i) = argmin_k d(xi, μk), where C(i) denotes the cluster assignment of data point xi.
3. Update Step:
    - Update the centroids μk = (1/Nk) ∑(xi ∈ C(k)) xi
4. Repeat Steps 2 and 3:
    - Convergence can be defined based on various criteria, such as the centroids no longer changing significantly or a predefined threshold being met.
5. Output:
    - The final cluster for each data point;
    - The final cluster centroids μk.

The elbow method determines the optimum clusters in a dataset, as shown in Figure 3a. While computing the within-cluster sum of squares (WCSS) for each value of k and plotting the cure, where k rises, the WCSS declines. This is due to more clusters of data points being moved closer to their centroids. The elbow point is when the rate of decline in WCSS starts to noticeably slow down, providing the curve with an elbow form. When k = 6, the identified clusters and centroid formations are shown in Figure 3b. Also, "×" shows the centroid.



(**a**) (**b**)

Figure 3. (**a**) Elbow technique for counting the number of clusters; (**b**) finding the centroids for the cluster.

Table 1 shows the algorithm to generate a map with markers representing the locations from the dataset.

Table 1. Algorithm to generate a map with markers.

Steps	Process Details
1	Import the folium libraries.
2	Load the dataset containing the geographic coordinates. Assuming the Data Frame as data with columns "latitude" and "longitude": coordinates = data[['latitude', 'longitude']]
3	Create a map centered around mean coordinates.
4	Add markers for each location on the map.
5	Display the map with the plotted locations: map_locations.

Examining the frequency of visits to various areas may utilize geopositional data to determine the most well-liked spots for tenants in a city. The tenant experience may be enhanced overall, and marketing activities can be targeted using these data. Many renters frequently go to the same coffee shop every day and discounts may be provided to tenants. The number of renters in a certain location and the distance they must travel to access vital services may utilize geopositional data to identify parts of the city that are underserved for tenants. Table 2 shows the algorithm to find the centroids for a cluster.

Table 2. Algorithm used to find the centroids for a cluster.

Steps	Process Details

1	Extract the data points belonging to the cluster of interest. Assuming the dataset as cluster_data with columns "latitude" and "longitude" and a cluster label column called "cluster":
	cluster_label = 0 # Specify the cluster label of interest
	cluster_points=cluster_data[cluster_data['cluster']==cluster_label][['latitude', 'longitude']]

2	Calculate the centroid of the cluster by taking the mean of the latitude, longitude values: centroid = cluster_points.mean().
3	The resulting centroid object contains the latitude and longitude coordinates of the cluster.
4	The centroid coordinates are used for further analysis, visualization, or any other relevant tasks related to the specific cluster.

4. Results and Discussion

The proposed system not only shows details on prospective flats, homes, and hostels, but it also provides ideas for various housing possibilities. It suggests hotels that are within your budget range for staying. There are multiple homes in the specified price range. Folium offers an interactive map that lets users examine the clusters and their distribution by zooming in and out, panning around, and interacting with the markers. Your geographic clustering results and their spatial patterns can be shown in a clear visual manner in this way. The proposed approach uses the kK-means method to avoid erroneous findings when two circular clusters with a comparable mean possess distinct radii. K-Means computes the cluster center using the median value instead of distinguishing between the two distinct clusters. Existing techniques fail when the resulting sets are not circular.

Folium provided the visual and interactive geographical patterns and insights discovered by analyzing the geopositional data of renters in a certain city; these are shown in Figure 4 for five clusters. It gives the visualization of rental details across different cities through the proposed approach based on the search criteria.

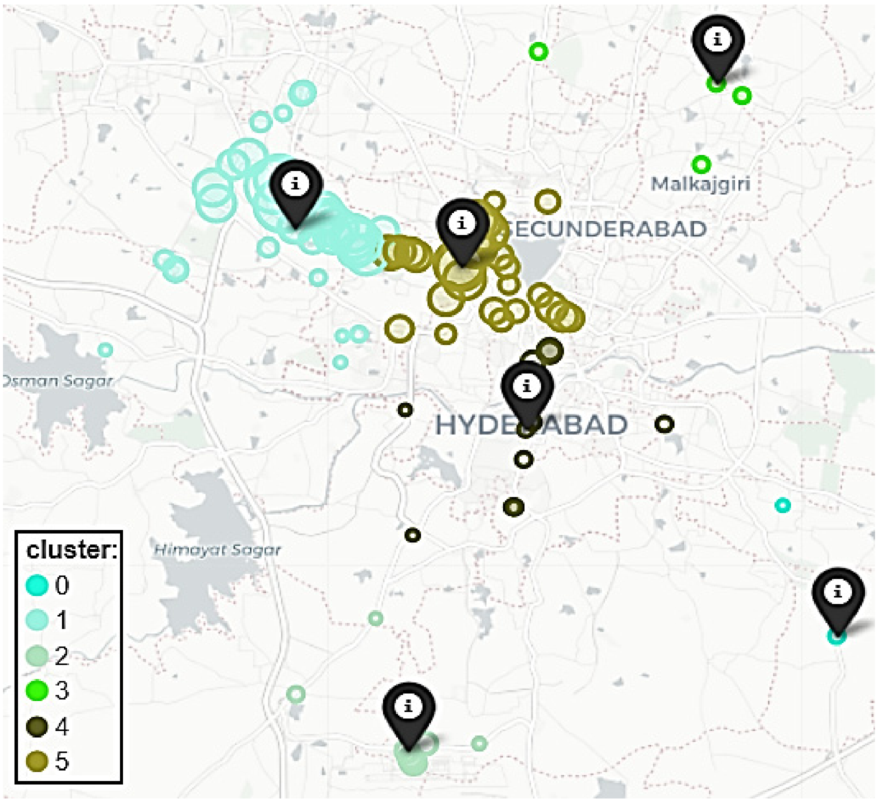


Figure 4. Final results after plotting the map using Folium.

The proposed method requires collecting the data before using k-means clustering to analyze the geopositional data of renters in a city. The latitude and longitude of the renters' residences, as well as the locations that they visit, may be included in these data. The next step would be to choose k clusters. Usually, the domain expert determines how many clusters to use based on his or her expertise. If k = 4, then one can be aware that there are six different categories of renters in the city. After deciding on k, the data will need to go through the k-means clustering method. Each data point will be repeatedly assigned by the algorithm to the cluster with the nearest mean. Once the clusters stop changing, the algorithm will stop iterating. The approach can have several clusters when the k-means clustering process is finished. Each cluster will represent a collection of renters that move similarly. The demands and preferences of renters in the city may then be gleaned from an analysis of these groupings [16]. One group of renters, for instance, may consist primarily of young professionals who reside in the city's center and work in the business district [17]. Couples with kids who travel to the city from the suburbs may make up another group of tenants. Understanding the various tenant groups will help us better target our advertising efforts and enhance the overall renter experience [18].

Data Privacy and Security Measures

Geolocational data contains details about a person's or a device's precise location that are extremely sensitive and have to be protected with strong data privacy and security safeguards [19]. Data are encrypted during network transmission using secure communication protocols, to make sure that only authorized individuals see or alter the data [20]. Role-Based Access Control (RBAC) is utilized to prevent unauthorized individuals from accessing, using, disclosing, altering, or destroying the data. To ensure sure the data privacy and security safeguards are active and effective, frequent security audits and assessments should be conducted.

5. Conclusions

The proposed system provided the value of geolocation in the hospitality sector as well GIS, and geolocations have been shown to be used to evaluate the amount of hotel competition in a well-known tourist area. The K-Means clustering technique may be used to analyze the geopositional data of renters in a particular city to provide important insights into their spatial preferences, behaviors, and patterns. K-Means is a popular clustering method that divides data points into clusters according to how close they are to a cluster centroid. Large datasets can be scaled for K-Means clustering. Because geopositional data may frequently be huge and complicated, this makes it a suitable option for analysis. In the future, this technique might be used in lodging competitiveness investigation to provide more detailed data on every area inside a single vacation spot to analyze profitability in the present tourist market.

Author Contributions: Conceptualization, S.G. and L.V.; methodology, L.V.; formal analysis, J.K.S.S.; investigation, D.B.; resources, T.J.J.; writing—original draft preparation, S.G., L.V. and T.J.J.; writing—review and editing, T.J.J. and D.B.; visualization, L.V.; supervision, S.G.; project administration, J.K.S.S.; funding acquisition, S.G., J.K.S.S., D.B., T.J.J. and L.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors would like to thank the Karunya Institute of Technology and Sciences for all of the support to complete this research.

Conflicts of Interest: The authors do not have any conflict of interest.

References

1. Kamel Boulos, M.N.; Geraghty, E.M. Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: How 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. *Int. J. Health Geogr.* **2020**, *19*, 8. [PubMed]
2. Koch, T. The map that launched the idea of global health. *Int. J. Epidemiol.* **2014**, *43*, 1014–1020. [CrossRef]
3. Sridharan, K.; Chitra, D.M. Web Based Agent And Assertion Passive Grading for Information Retervial. *ARPN J. Eng. Appl. Sci.* **2015**, *10*, 7043–7048.
4. Sumithra, M.; Malathi, D.S. Segmentation of Different Modalities using Fuzzy KMeans and Wavelet ROI. *Int. J. Sci. Technol. Res.* **2019**, *8*, 996–1002.
5. Sarfo, A.K.; Karuppannan, S. Application of Geospatial Technologies in the COVID-19 Fight of Ghana. *Trans. Indian Natl. Acad. Eng.* **2020**, *5*, 193–204. [CrossRef]
6. Franch-Pardo, I.; Napoletano, B.M.; Rosete-Verges, F.; Billa, L. Spatial analysis and GIS in the study of COVID-19. A review. *Sci. Total Environ.* **2020**, *739*, 140033. [CrossRef] [PubMed]
7. Butler, J.A. *Designing Geodatabases for Transportation*; ESRI, Inc.: Redlands, CA, USA, 2008.
8. Dong, E.; Ratcliff, J.; Goyea, T.D.; Katz, A.; Lau, R.; Ng, T.K.; Garcia, B.; Bolt, E.; Prata, S.; Zhang, D.; et al. The Johns Hopkins University Center for Systems Science and Engineering COVID-19 Dashboard: Data collection process, challenges faced, and lessons learned. *Lancet Infect. Dis.* **2022**, *22*, e370–e376. [CrossRef] [PubMed]
9. Benbrahim Ansari, O. Geo-marketing segmentation with deep learning. *Businesses* **2021**, *1*, 51–71. [CrossRef]
10. Yu, D.; Fang, C. Urban Remote Sensing with Spatial Big Data: A Review and Renewed Perspective of Urban Studies in Recent Decades. *Remote Sens.* **2023**, *15*, 1307. [CrossRef]
11. Miah, S.J.; Vu, H.Q.; Gammack, J.; McGrath, M. A big data analytics method for tourist behaviour analysis. *Inf. Manag.* **2017**, *54*, 771–785. [CrossRef]
12. Hasan, S.; Ukkusuri, S.V.; Zhan, X. Understanding social influence in activity location choice and lifestyle patterns using geolocation data from social media. *Front. ICT* **2016**, *3*, 10. [CrossRef]
13. Peixoto, J.P.; Costa, D.G.; da Franca Rocha, W.D.; Portugal, P.; Vasques, F. On the positioning of emergencies detection units based on geospatial data of urban response centres. *Sustain. Cities Soc.* **2023**, *97*, 104713. [CrossRef]
14. Xie, Y.; Gupta, J.; Li, Y.; Shekhar, S. Transforming smart cities with spatial computing. In Proceedings of the 2018 IEEE International Smart Cities Conference, Kansas City, MO, USA, 16–19 September 2018; pp. 1–9.

15. Vargas-Munoz, J.E.; Srivastava, S.; Tuia, D.; Falcao, A.X. Falcao. OpenStreetMap: Challenges and opportunities in machine learning and remote sensing. *IEEE Geosci. Remote Sens. Mag.* **2020**, *9*, 184–199. [CrossRef]
16. Dataset. Available online: https://github.com/She-Codes-Now/Intro-to-Data-Science-with-R/blob/master/food_coded.csv (accessed on 31 October 2023).
17. De Bruin, R. Informational Privacy and Trust in Autonomous Intelligent Systems. *Towards Trust. Artif. Intell. Syst.* **2022**, *102*, 47–60.
18. Lazar, E. *Privacy and Security of Health Data—What's at Stake? Regulating Cyber Technologies: Privacy Vs Security*; World Scientific Publishing Company: Singapore, 2023.
19. Kayhan, H. Ensuring trust in pharmaceutical supply chains by data protection by design approach to blockchains. *Blockchain Healthc. Today* **2022**, *5*, 232–249. [CrossRef] [PubMed]
20. Zum Felde, H.M.; Reding, J.L.; Lux, M. D-GATE: Decentralized Geolocation and Time Enforcement for Usage Control. In Proceedings of the 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Delft, The Netherlands, 3–7 July 2023; pp. 386–395.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.