Geospatial Analysis for Choosing Suitable Location to Start Hotel in Sri Lanka Using Machine Learning

Sachith Yamannage

February 8, 2024

Geospatial Analysis for choosing suitable location to start hotel in Sri Lanka Using Machine Learning

Sachith Nimesh Yamannage

B.SC (hon) Financial Mathematics and Industrial Statistics

University of Ruhuna, Sri Lanka

Email: sachithnimesh999@gmail.com

Abstract

The post-pandemic recapture of Sri Lanka's tourist industry is anticipated, and this paper suggests a hotel that employs geospatial data science for strategic placement and operational excellence. The hotel will use geospatial analytics to choose the best site, customize its hospitality offerings to meet local demands, assign resources as efficiently as possible, build strong community relationships, and navigate the competitive marketplace. The hotel will succeed in the resurgent Sri Lankan hospitality sector cheers to its data-driven strategy.

Keywords: geospatial data science, hotel, Sri Lanka, Machine Learning

1.Introduction

As Sri Lanka gets ready for a post-pandemic shrinking, the country's booming hotel industry grants a huge opportunity for success and innovation. To exploit on this expanding potential, we propose to build a hotel that logically integrates the revolutionary power of geospatial data science. Owing to our pioneering approach, which will modify our market analysis, customer interface, and operational efficacy with geographic location, our hotel will be palpable in this crowded market.

Using a huge dataset, our company put on geospatial data science to determine the top spot for our hotel in Sri Lanka, making sure it meets market demand and customer prospects.

Through an alertness of the intricate relationships between local characteristics and customer likings, geospatial data science enables establishments to tailor services to match the wants of their target audience. Optimizing resource provision and improving visitor experiences are critical in supply chain and marketing. One Sri Lankan hotel is devoted to leveraging geospatial data science to select the ideal location, operate efficiently, and make a lasting brand in the rapidly expanding hospitality sector.

1.1 Research Questions

1. How are hotels spatially distributed across different districts in Sri Lanka?
2. How can the identified clusters inform strategic decisions in the hotel industry, tourism, and regional planning?
3. What insights can be gained by combining machine learning techniques with geospatial data analysis in the context of the hotel industry?

2.Literature Review

After the sweeping, Sri Lanka's hospitality sector is expected to grow significantly, and to stay viable, lodging establishments are increasingly utilizing geospatial data science. The best place for a hotel may be found using geospatial data science, which can also be used to achieve competition, improve resource apportionment, customize hospitality products to meet local requirements, and build strong communal links.

One study by Mariani et al. (2021) found that geospatial data science can be used to identify areas with high tourist petition and low hotel supply. This information can then be used to develop targeted marketing operations and attract more guests.[1]

Another study by Centobelli and Ndou (2019) found that geospatial data science can be used to optimize hotel operations. For example, geospatial data can be used to realize the movement of guests and staff, which can then be used to rally staffing levels and reduce wait times.[2]

Geospatial data science is a powerful tool that can be used to expand the performance of hotels in Sri Lanka. By leveraging geospatial data, hotels can gain treasured insights into their target market, optimize their operations, and build a sturdy brand.

3.Methodology:

The approach adopted in the study of the info on hotels in Sri Lanka incorporates several stages to discover, process, and interpret the data. The main techniques employed are charted in the following. The machine learning language secondhand in this work was Python.

Data Processing:

In order to preserve the integrity of the dataset, missing values were handled appropriately by means of the drop NA method. After that, a GeoDataFrame (gdf) was built to assurance accuracy for statistical and geographic analysis.

Visualization Techniques:

A heatmap was used in the research to find trends in the latitude, longitude, and number of rooms. Hidden patterns were revealed by exploiting k-means clustering technique. The distribution of hotel kinds and locations in Sri Lanka was revealed graphically using Pie Charts and Scatter Plots.

Spatial Statistical Models:

Sri Lankan districts were divided into five clusters consistent with latitude and longitude using k-means clustering. A thorough distance matrix was erected to clarify the geographical relationships and patterns crosswise districts.

Geovisualization:

Utilizing GeoPandas, the dataset was altered into a GeoDataFrame, facilitating the visualization of hotel distribution on a Sri Lankan map. Map overlaps incorporated hotel locations onto a global map, as long as additional context. Spatial analysis was hired

1

2

to visualize trends, patterns, and potential gaps in hotel distribution clusters.

Machine Learning for Geo-spatial Data Analysis:

By choosing related columns, one-hot encoding categorical variables, and building an extensive representation, the feature matrix was ready. StandardScaler allowed standardization, which guaranteed the feature matrix's consistency for precise machine learning analysis. Next, K-means clustering was used to identify regional hotel segmentation trends.

4.Findings

The collection contains hotel specifics such as name, address, number of rooms, grade, district, AGA division, and geographic coordinates. Examples of hotel types embrace boutique and secret (longitude and latitude). With information on associates with the Pradeshiya Sabha (PS), Municipal Council (MC), or Urban Council (UC), it seems to concentrate on hotels in various locales, most likely in Sri Lanka.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2130 entries, 0 to 2129
Data columns (total 10 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Type         2130 non-null   object
 1   Name         2130 non-null   object
 2   Address      2130 non-null   object
 3   Rooms        2130 non-null   int64
 4   Grade        1837 non-null   object
 5   District     2130 non-null   object
 6   AGA Division 2111 non-null   object
 7   PS/MC/UC     2127 non-null   object
 8   Logitiute    1368 non-null   float64
 9   Latitude     1370 non-null   float64
dtypes: float64(2), int64(1), object(7)
memory usage: 166.5+ KB
```

	Rooms	Logitiute	Latitude
count	2130.000000	1368.000000	1370.000000
mean	16.961972	80.430743	7.058470
std	36.657709	0.509183	2.119415
min	1.000000	79.705919	5.936771
25%	4.000000	79.967549	6.452453
50%	6.000000	80.396191	6.915005
75%	14.000000	80.729967	7.287510
max	541.000000	81.856883	80.791643

Figure 1

Heatmap shows clearly about correlations between numerical variables.

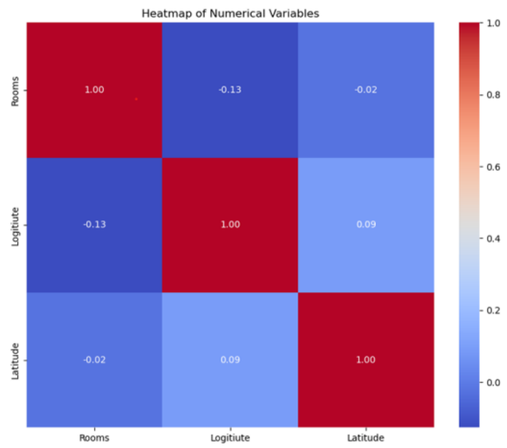


Figure 2

The dataset was grouped into discrete groups using clustering analysis; the cluster hubs at 9.89, 118.79, and 409.5 highlighted the characteristics of these groups for suitable categorization and analysis.

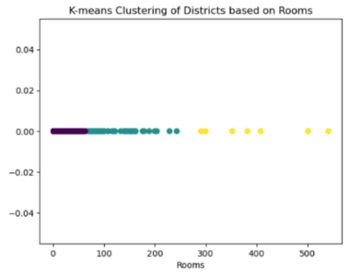


Figure 3

Matplotlib calm with Python Pandas were used to display the dataset. A pie chart with color coding professionally conveys the several percentages of different sorts of accommodations and provides a fleeting synopsis of the dataset's makeup. The scatter plot shows the distribution of hotels in Sri Lanka, highlighting major cities, outlying areas, and prospective areas for increase in the tourist industry.

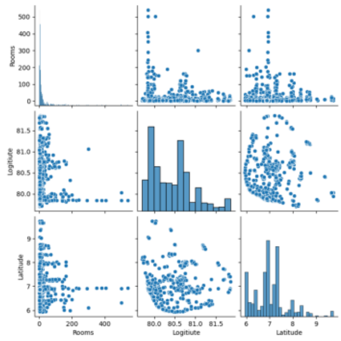


Figure 4

The dataset's longitude coordinates, rooms, and latitude were exposed using a boxplot created using Matplotlib and Seaborn. This figure makes usage of the "viridis" color scheme to obviously show significant differences while succinctly illustrating the distribution and highlighting outliers.

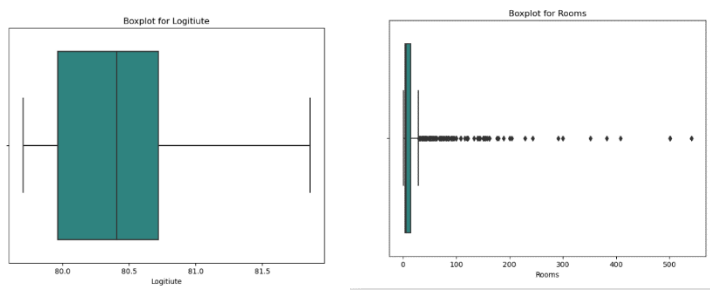
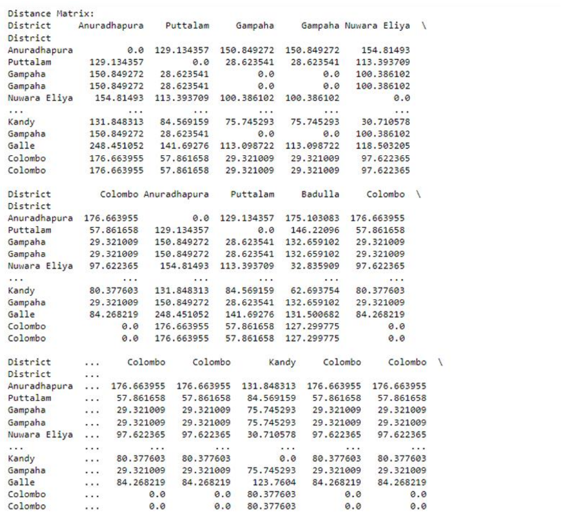


Figure 5

4.2 Quantitative analysis

k-means clustering recovers spatial statistical models, which are vital for identifying trends in geographically separate data. The blend of spatial models with k-means clustering, as this research using the PySAL, Pandas, and GeoPy libraries, expressions, greatly improves the accuracy of spatial data analysis, and benefits data scientists make well-informed verdicts. I thoroughly created a comprehensive distance matrix for each district in Sri Lanka, tightfitting the distances between them. With 1088 rows and columns, this matrix is crucial for spatial statistical modeling and offers information for regional planning and reserve allocation, among other uses.



```
Distance Matrix:
District     Anuradhapura   Puttalam    Gampaha    Gampaha Nuwara Eliya  \
District
Anuradhapura          0.0 129.134357 150.849272 150.849272    154.81493
Puttalam       129.134357        0.0  28.623541  28.623541   113.393709
Gampaha        150.849272  28.623541        0.0        0.0   100.386102
Gampaha        150.849272  28.623541        0.0        0.0   100.386102
Nuwara Eliya    154.81493 113.393709 100.386102 100.386102          0.0
...                   ...        ...        ...        ...          ...
Kandy          131.848313  84.569159  75.745293  75.745293    30.710578
Gampaha        150.849272  28.623541        0.0        0.0   100.386102
Galle          248.451052  141.69276 113.098722 113.098722   118.503205
Colombo        176.663955  57.861658  29.321009  29.321009    97.622365
Colombo        176.663955  57.861658  29.321009  29.321009    97.622365

District        Colombo Anuradhapura   Puttalam    Badulla    Colombo  \
District
Anuradhapura 176.663955          0.0 129.134357 175.103083 176.663955
Puttalam      57.861658   129.134357        0.0  146.22096  57.861658
Gampaha       29.321009   150.849272  28.623541 132.659102  29.321009
Gampaha       29.321009   150.849272  28.623541 132.659102  29.321009
Nuwara Eliya  97.622365    154.81493 113.393709  32.835909  97.622365
...                 ...          ...        ...        ...        ...
Kandy         80.377603   131.848313  84.569159  62.693754  80.377603
Gampaha       29.321009   150.849272  28.623541 132.659102  29.321009
Galle         84.268219   248.451052  141.69276 131.500682  84.268219
Colombo             0.0   176.663955  57.861658 127.299775        0.0
Colombo             0.0   176.663955  57.861658 127.299775        0.0

District     ...    Colombo    Colombo      Kandy    Colombo    Colombo  \
District     ...
Anuradhapura ... 176.663955 176.663955 131.848313 176.663955 176.663955
Puttalam     ...  57.861658  57.861658  84.569159  57.861658  57.861658
Gampaha      ...  29.321009  29.321009  75.745293  29.321009  29.321009
Gampaha      ...  29.321009  29.321009  75.745293  29.321009  29.321009
Nuwara Eliya ...  97.622365  97.622365  30.710578  97.622365  97.622365
...          ...        ...        ...        ...        ...        ...
Kandy        ...  80.377603  80.377603        0.0  80.377603  80.377603
Gampaha      ...  29.321009  29.321009  75.745293  29.321009  29.321009
Galle        ...  84.268219  84.268219   123.7604  84.268219  84.268219
Colombo      ...        0.0        0.0  80.377603        0.0        0.0
Colombo      ...        0.0        0.0  80.377603        0.0        0.0
```

Figure 6

The dataset was separated into five clusters using K-means clustering on Sri Lankan districts using latitude and longitude coordinates. These clusters were graphically showed by a plot that showed a distinct geographic separation. Clusters with diverse district makeup were numbered 0–4. Interestingly, Cluster 0 consisted of Puttalam, whereas Badulla, Ratnapura, and Kurunegala were part of Cluster 1. The average longitude was between 80.0 and 82.0 degrees east, while the average latitude was between 9.5 and 6.5 degrees north. By revealing distinct geographic patterns, this spatial clustering approach providing insightful information for regional planning, resource allocation, and realizing the fundamental structure of district data.

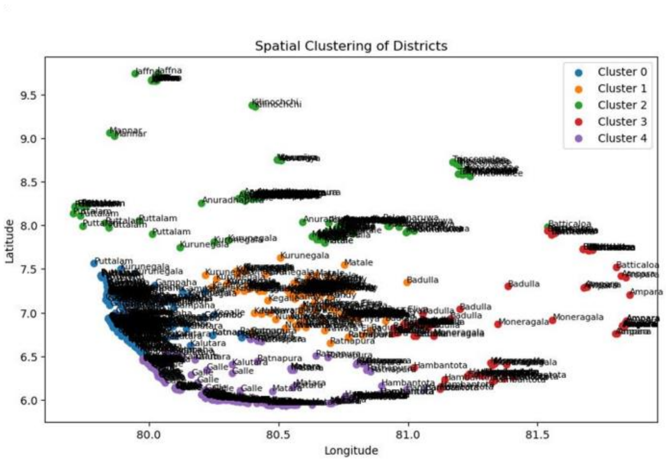


Figure 7

4.3 Geovisualization:

A key module of spatial data science is geovisualization, which is the graphic depiction of complex geographic patterns using graphs, charts, and maps. I used GeoPandas to produce a data frame for my inquiry out of a dataset that included room, district, latitude, and longitude information. Stakeholders may study more about the distribution of hotels by covering this geospatial data on a map of Sri Lanka that has red markers for hotels (scaled by number of rooms). This offers a comprehensive picture of hotel clusters, their capacity, and possible trends, which benefits with decision-making in the hotel industry, tourism, and district development. The command of geographical trends in the Boutique Hotel dataset is enhanced by this geovisualization.

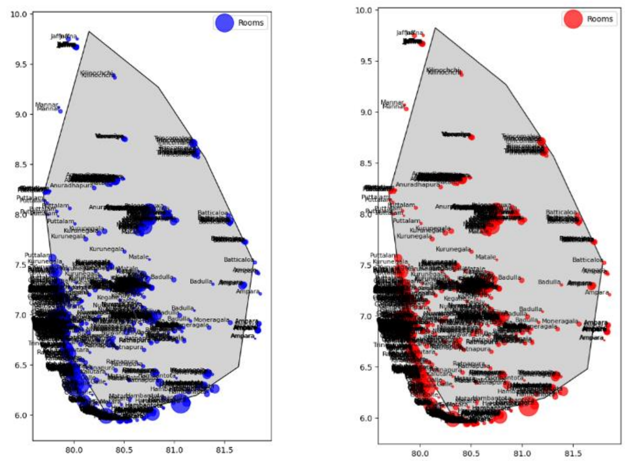
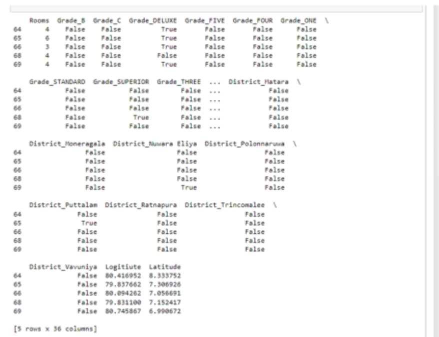


Figure 8

4.4 Machine learning for Geo-spatial data analysis:

Machine learning brings about a revolutionary change in geospatial data analysis by making pattern detection and trend forecasting possible. Relevant columns such as "District" and "Grade" in the Hotels Data Frame are one-hot encoded to generate an widespread feature matrix. Based on room capacity, district locations, and grades, this matrix shows the fundamental linkages between hotels. In the matrix, rows are calm of numerical properties such as "Rooms," "Longitude," and "Latitude." Encoded categories are epitomized by binary columns, such as "Rating DELUXE" or "District Colombo." This method makes trends obvious via machine learning algorithms. With the totaling of geographic coordinates, the feature matrix becomes much more insightful, revealing beforehand ignored features and architecture of Sri Lankan hotels.



```
    Rooms  Grade_B  Grade_C  Grade_DELUXE  Grade_FIVE  Grade_FOUR  Grade_ONE  \
64      4    False    False          True       False       False      False
65      6    False    False          True       False       False      False
66      3    False    False          True       False       False      False
68      4    False    False         False       False       False      False
69      4    False    False          True       False       False      False

    Grade_STANDARD  Grade_SUPERIOR  Grade_THREE  ...  District_Matara  \
64           False           False        False  ...            False
65           False           False        False  ...            False
66           False           False        False  ...            False
68           False            True        False  ...            False
69           False           False        False  ...            False

    District_Moneragala  District_Nuwara Eliya  District_Polonnaruwa  \
64                False                  False                 False
65                False                  False                 False
66                False                  False                 False
68                False                  False                 False
69                False                   True                 False

    District_Puttalam  District_Ratnapura  District_Trincomalee  \
64              False               False                 False
65               True               False                 False
66              False               False                 False
68              False               False                 False
69              False               False                 False

    District_Vavuniya  Logitiute  Latitude
64              False  80.416952  8.333752
65              False  79.837662  7.306926
66              False  80.094262  7.056691
68              False  79.831100  7.152417
69              False  80.745867  6.990672

[5 rows x 36 columns]
```

Figure 9

By means of scikit-learns StandardScaler, the feature matrix was standardized to indorse equal contributions to clustering, instructive accuracy and reliability. This preprocessing stage makes pattern acknowledgment easier and is essential for K-means clustering. Three groups were found using K-means analysis in Sri Lankan hotels, portentous geographic division. These support besieged insights into joint traits, guiding strategic choices. Consuming geographic data with machine learning clustering advances the hotel industry's educated slant to result important trends.

```python
# Display the updated dataframe with cluster information
print(gdf[['District', 'Rooms', 'Grade', 'Logitiute', 'Latitude', 'Cluster']])
```

```
          District  Rooms     Grade  Logitiute  Latitude  Cluster
64     Anuradhapura      4    DELUXE  80.416952  8.333752        0
65         Puttalam      6    DELUXE  79.837662  7.306926        1
66          Gampaha      3    DELUXE  80.094262  7.056691        1
68          Gampaha      4  SUPERIOR  79.831100  7.152417        1
69     Nuwara Eliya      4    DELUXE  80.745867  6.990672        0
...             ...    ...       ...        ...       ...      ...
1895          Kandy      3  STANDARD  80.560632  7.159357        0
1897        Gampaha      1  STANDARD  79.875283  7.136658        1
1898          Galle      3  STANDARD  80.100697  6.139111        0
1899        Colombo      5  SUPERIOR  80.036651  6.850166        2
1900        Colombo      4  SUPERIOR  79.892414  6.872081        2

[1088 rows x 6 columns]
```

Figure 11

4.5 Predictive analytics for geospatial application:

Key principles including room capacity, grade, and geographic coordinates are engaged into account in the study of Sri Lanka's hotel information with geospatial technology, namely K-means clustering. This innovative strategy recovers the hotel's profitability and long-term feasibility. The model predicts cluster preps by using K-means clustering on a fresh set of variables, which contain normalized attributes like "Rooms," "Type," and "Grade." Choosing a suitable neighborhood for a deliberate boutique hotel might be aided by utilizing the most frequent AGA Division in the projected cluster. By commending a neighborhood parallel to hotels in the same cluster, this model-driven proposal—which is based on patterns educated from the current dataset—improves decision-making and offers a data-driven foundation for tactically placing the new hotel in harmony with current spatial trends in Sri Lanka's hotel industry.

5.Discussion

The systematic examination of Sri Lanka's hotel dataset has provided important new information about the geographical subtleties of the sector. There are evident relationships between numerical variables and different types within the dataset, as demonstrated by the heatmap and clustering analysis. The several percentages of lodging categories and the geographic distribution of hotels are clearly interconnected by the visualizations, which include pie charts and scatter plots. The spatial statistical models expressly the one with k-means clustering offer important insights into the patterns seen in data that is dispersed spatially. Creating a detailed distance matrix for each district in Sri Lanka enlarges the possible customs for regional planning and resource distribution.

A real technique that gives stakeholders an innate considerate of hotel distribution, clusters, and possible trends is geovisualization. A prophetic element is added by the machine learning method, notably K-means clustering, which acclaims potential sites for new hotel based on patterns discovered from the offered information.

6.Conclusion

Of conclusion, a inclusive knowledge of the spatial undercurrents in Sri Lanka's hotel business is made possible by the grouping of exploratory data analysis, machine learning tactics, and spatial statistical models. The results afford industry stakeholders with practical insights to help them style well-informed conclusions on resource provision, regional planning, and calculated hotel placement, among other topics. Unconventional analytics combined with the compliance of geospatial technology equips the segment for innovation and long-term success. This research adds to our thoughtful of the hotel environment and arranges the groundwork for additional studies and claims in the rapidly developing subject of spatial data science in the hospitality business.

7.References

[1]. Mariani, M. M., Baggio, R., & Perri, S. (2021). Big data and analytics in hospitality and tourism: A systematic literature review. International Journal of Contemporary Hospitality Management, 33(11), 3642-3677

[2]. Centobelli, P., & Ndou, E. N. (2019). The role of geospatial data science in the tourism industry. Journal of Hospitality and Tourism Management, 42, 298-305.

8.Appendix

- Metadata - [Accommodation Information for Tourists | Open Data Portal - Sri Lanka]

Type	Name	Address
Boutique Hotels	THE THEVA RESIDENCY	11/B5/10-1 06TH LANE,
Boutique Hotels	HIGHLAND VILLA	351, ABIMANGAMA RO
Boutique Hotels	ULAGALLA WALAWWA RESORT	THIRAPPANE, ANURADH
Boutique Hotels	GALLE FORT HOTEL	NO.28, CHURCH STREET
Boutique Hotels	THE ELEPHANT CORRIDOR	POTHANA, KIBISSA, SIGI

Rooms	Grade	District	AGA Division
10		Kandy	Kandy Divisional Secretariat
10		Matara	Weligama Divisional Secretariat
21		Anuradhapura	Anuradhapura East
14		Galle	Galle Divisional Secretariat
21		Matale	N/A

PS/MC/UC	Logitiute	Latitude
Kandy	80.63541	7.276036
Weligama Pradeshiya Sabha	80.40997	5.960334
Anuradhapura	80.54506	8.205927
Galle	80.21756	6.026649
Matale	80.71074	7.943525