CIS 7030

GEOSPATIAL ANALYSIS

Table of Contents

How geospatial data science can be used for business.

Business plan - starting a new educational institute in optimal location in Sri Lanka

Introduction

In the modern world, where data is the king and technology is advancing at a rapid pace, companies are always on the lookout for new and creative ways to improve their operations, gain a competitive edge, and make informed decisions. Among the various game-changing instruments available, geospatial data science is particularly strong and useful, providing a vast toolbox that opens up new avenues for understanding consumer behavior, market dynamics, and operational efficiency. This proposal aims to explore the potential of geospatial data science to transform the retail industry by strategically integrating it into the market landscape, helping organizations navigate the complexities more effectively.

Geospatial data science is a powerful tool that can uncover patterns, trends, and correlations in geographical data such as location, demographics, and traffic patterns. This proposal recognizes the potential of geospatial research to generate actionable insights for informed decision-making and proposes its application to improve the performance of retail businesses. By analyzing geospatial data, businesses can gain a better understanding of their customers' preferences, manage their inventories more efficiently, position themselves strategically in the market, and increase customer engagement.

Our investigation is based on a large dataset that covers a wide range of topics related to schools in various parts of Sri Lanka. This dataset serves as a microcosm, exemplifying the application of geospatial data science in a dynamic and diverse setting. As we delve into the complex aspects of schools, their locations, and the corresponding demographic data, we draw comparisons to the retail industry. We cannot help but imagine the revolutionary effects that these insights may have on improving business tactics..

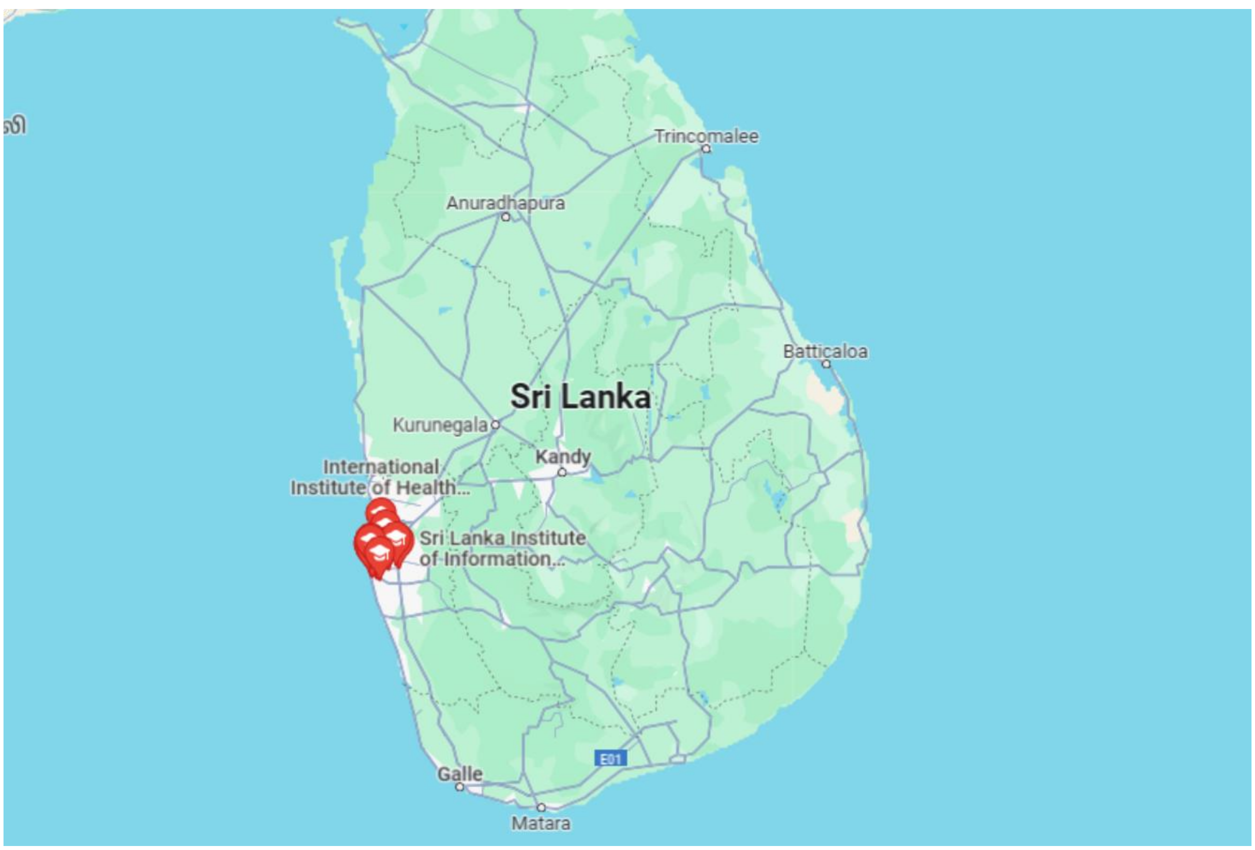
Join us on an exciting journey where we explore how the integration of retail operations with geographic data science can unlock new opportunities for productivity, creativity, and strategic growth. Our proposal aims to demonstrate how companies can leverage spatial data to overcome market challenges and thrive in the future.

Objectives

1: Strategic Location Identification

Geospatial analytics can be used to determine the most suitable location for a new school based on factors such as accessibility, population density, and proximity to potential students. The goal is to identify a site that maximizes ease of use and accessibility for the intended audience. In Sri Lanka, geospatial analytics can be employed to evaluate these factors and determine the optimal site for a new educational institution.:

- Density of population: The institute needs to be situated in a region with a large concentration of prospective pupils. This will guarantee that the demand for the institute's services is high enough.
- proximity to prospective students: The institute needs to be situated in a region that is easily accessible to prospective students. Students will find it simpler to get to and from the institute as a result.
- Accessibility: The institute should be situated in a place where public transit is readily available. Students who do not have their means of transportation will find it simpler to go to the institute thanks to this. (reference 1)



2: Demographic-Tailored Program Development

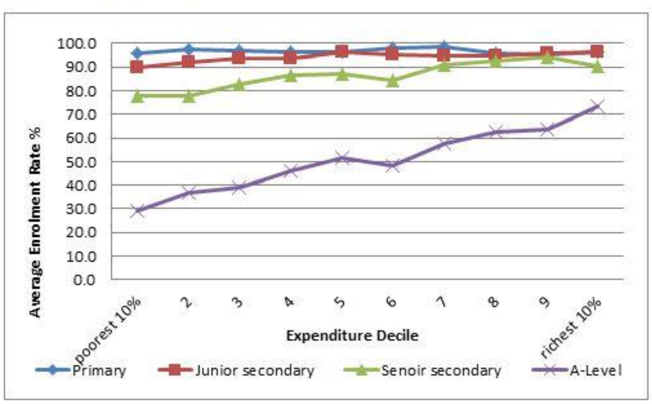
In order to provide relevant educational programs that cater to the unique demands and preferences of the local population, it is important to utilize insights gathered from geospatial data to examine the demographic makeup of the selected area. This will help educational establishments in Sri Lanka to attract a varied student body. Additionally, geospatial data insights can be used beyond education to improve various aspects of the community. Locate students in the selected area. Outreach and marketing activities can be targeted using this information.

- Monitor the enrollment and graduation rates of the selected student body.
- This data may be utilized to determine areas in need of development and assess how successful educational initiatives are. Make plans for upcoming development and growth.
- It is possible to find possible new locations for educational establishments using this information.

3: Resource Allocation Optimization

Maximizing the distribution of resources across an organization can be achieved through the use of geographic analytics. This includes efficiently allocating resources such as personnel, facilities, and classrooms based on the geographical dynamics of demand and student enrollment. By simplifying resource utilization, operational effectiveness can be improved, and wasteful spending can be reduced.. (reference 2)

Figure 2: Opportunity Curves for Access to Education, 2016

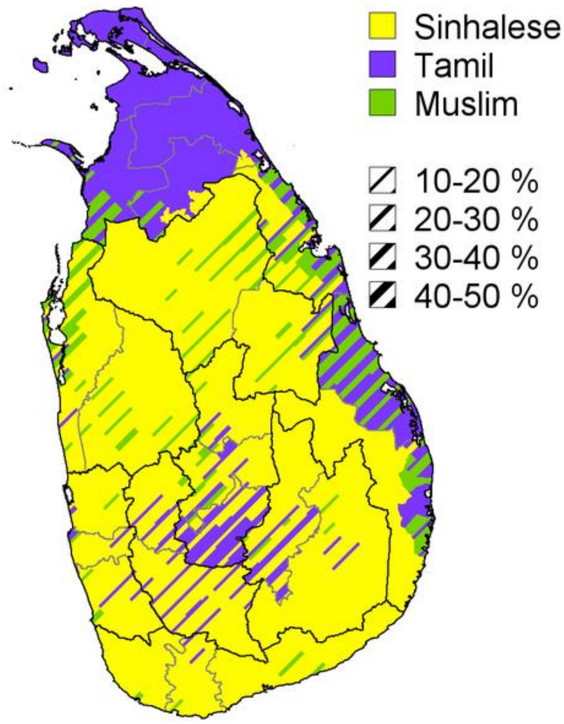


Source: Author's calculations using HIES 2016 data.

Note: Net enrolment rates measure enrollment of the official age group for a given level of education expressed as a percentage of the corresponding population

4: Community Engagement Strategy

To strengthen the links with the community and build a supportive environment for the educational institution, it is important to identify significant stakeholders and community hubs close to the institute. This can be achieved by using geospatial analytics to develop a community engagement plan. The plan will focus on building alliances with local companies, organizations, and schools, which will help to strengthen the engagement with community.



reference 3

According to latest data ,

The Sri Lankan community comprises three main ethnic groups: Sinhalese, Tamils, and Muslims. Sinhalese constitute the largest ethnic group at 74.9%, followed by Tamils at 18.6% and Muslims at 7.1%. The largest Sinhalese speakers are found in the Southern Province (95.8%), followed by the Uva Province (93.3%) and North Central Province (92.5%). Tamils are predominantly found in the Eastern and Northern Provinces, while Muslims are predominant in the districts of Puttalam and Ampara. The proportion of Muslims varies across provinces, with Ampara District having the highest percentage (73.4%).

5: Competitive Landscape Analysis

Perform a comprehensive analysis of the competitive landscape in the chosen area using geographic information. This includes understanding the distribution of existing educational institutions, the services they offer, and any gaps in the market. The objective is to position the new school strategically to differentiate it from competitors and capitalize on unfulfilled educational needs.

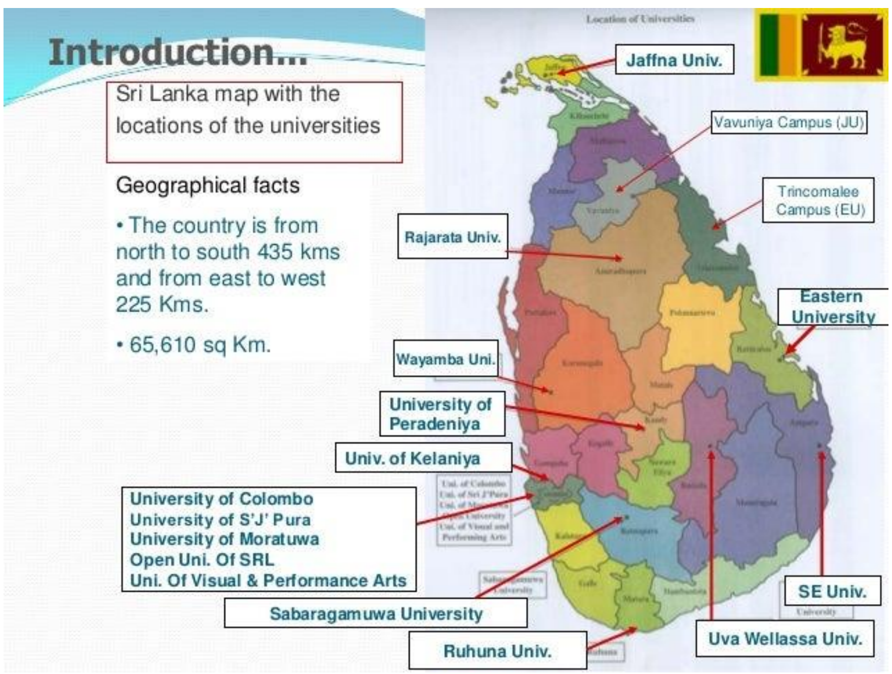
Visualization

Here are some appropriate, regionally specific infographics for the context of opening a new school in Sri Lanka:

1. **Proposed Institute Locations Map:**

It is required to create a comprehensive map that displays the proposed locations of educational institutions in every district of Sri Lanka. The district boundaries can be used to highlight the well-planned placement of each institute, providing a clear and visual representation of their distribution. To differentiate between rural and urban areas, use color-coded markers while considering accessibility and population density. (reference 4)

As an example, consider university geography.



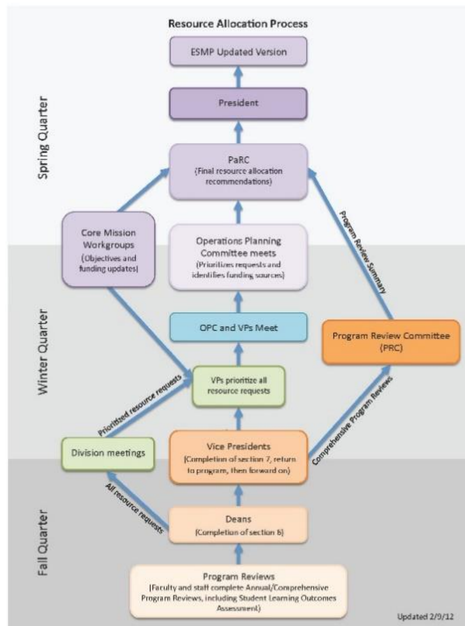
2. **Spatial Patterns of Student Enrollment Heatmap:**

Creating a heatmap that illustrates the preferred student enrollment patterns of each district is crucial. To identify clusters of students and highlight areas with higher demand for educational programs, geospatial analytics can be utilized. This data can be used to tailor educational offerings to better align with the interests and characteristics of students in different regions of Sri Lanka..

3. **Optimized Resource Allocation Flowchart:**

Please create a flowchart that illustrates the optimal allocation of resources for each institute. The flowchart should depict how personnel, infrastructure, and classrooms can be efficiently distributed in Sri Lanka, considering the unique geographical dynamics of student demand and enrollment. The focus of the flowchart should be on operational efficiency by showcasing the optimized use of resources.

The below figure shows an example:



4. **Community Engagement Network Map:**

Can you create a network map showcasing the links formed in various districts of Sri Lanka due to community participation techniques? In the map, edges should represent alliances and cooperative efforts, while nodes can signify local businesses, schools, and community groups. This map will demonstrate how the institute has been integrated into multiple communities across Sri Lanka.

5. **Competitive Landscape Radar Chart for Sri Lankan Districts:**

To better understand the competitive environment faced by educational institutions in Sri Lankan districts, a radar graphic needs to be created. This graphic should include important components such as community involvement, facilities, and program options. Each institute should be represented on the radar chart, allowing for a visual comparison of their strengths and shortcomings unique to the Sri Lankan setting. By utilizing this depiction, a new educational institution can be effectively positioned within the local competitive environment..

These visual aids provide a unique perspective for the business strategy of establishing a new educational institution, as they are tailored to the specific geographic and demographic characteristics of Sri Lanka. In the Sri Lankan education context, they tell a visual story that aligns with the strategic goals of selecting the best site, ensuring program relevance, improving operational efficiency, fostering community inclusion, and enhancing competitive positioning.

Conclusion

In conclusion, integrating geospatial data science in establishing a new educational institution in Sri Lanka could lead to groundbreaking outcomes. By utilizing spatial analytics, the proposed methods aim to enhance decision-making processes, improve the overall educational experience, and optimize operational efficiency within the intricate network of Sri Lanka's multiple districts..

A comprehensive strategy is developed through competitive landscape analysis, community involvement, optimal allocation of resources, creation of programs tailored to demographics, identification of strategic locations, and resource optimization. The strategy is designed to align with the unique cultural, geographic, and educational characteristics of Sri Lanka.

To ensure that the institute's footprint is in compliance with the requirements of nearby towns, suggested maps are provided for the best locations. These maps take into account the unique features of each area. When viewed within the context of Sri Lanka, the geographical patterns of student enrollment choices provide insights that can be used to customize curricula in accordance with the national culture.

In Sri Lanka, the education industry is highly competitive. Using geospatial analytics to optimize resource allocation has the potential to enhance operational efficiency, which is a crucial determinant for success. The geospatial network-mapped community engagement method is designed to foster significant connections with nearby schools, companies, and organizations, in line with the spirit of cooperation prevalent in Sri Lankan society.

Through its strategic location and the insights provided by geospatial analysis, the institute is well-equipped to overcome obstacles and make the most of the opportunities specific to Sri Lanka, as the competitive landscape radar graphic demonstrates. With this comprehensive geographic strategy, the new educational institution is guaranteed to be an active participant in the educational ecosystem rather than just a spectator, promoting long-term growth in Sri Lanka's competitive and dynamic education sector.

Descriptive explanations

Exploratory Spatial Data Analysis

The dataset appears to include several features about schools, pupils, and instructors in several districts of Sri Lanka. Let's examine each of your dataset's columns in detail:

1. **District:** The name of the district in Sri Lanka.

2. **Longitude and Latitude**: The geographical coordinates of the district.

3. **Schools_Feeleying, Schools_Nonfeeleying, Schools_Special education, Schools_Total:** The number of schools, categorized by fee type (fee-paying, non-fee-paying, special education), and the total number of schools in each district.

4. **Students_Male, Students_Female, Students_Total:** The number of students, categorized by gender, and the total number of students in each district.

5. **Teachers_Male, Teachers_Female, Teachers_Total:** The number of teachers, categorized by gender, and the total number of teachers in each district.

This dataset contains detailed information about education in various districts of Sri Lanka. It provides the geographic location of each district, making it useful for spatial analysis. The dataset includes information on the number of special education-focused schools, as well as fee-paying and non-fee-paying schools. It also provides data on the distribution of teachers and students by gender.

This dataset can be used to understand the distribution of teachers, the number of schools, and student demographics, among other things. It can also be used to identify trends and investigate potential relationships between variables. The dataset can be a starting point for conducting statistical and geographic analyses to gain valuable insights into the educational system of Sri Lanka.

first rows in our collection,

```
In [5]: gdf.head()
```

Out[5]:

	District	longitude	latitude	Schools_Feeleying	Schools_Nonfeeleying	Schools_Special education	School
0	Colombo	6.94	79.85	14	16	4	
1	Gampaha	7.09	79.99	5	6	3	
2	Kalutara	6.58	79.96	3	4	3	
3	Kandy	7.30	80.64	1	7	1	
4	Matale	7.47	80.62	1	0	1	

```
In [5]: gdf.head()
```

Out[5]:

l	Students_Male	Students_Female	Students_Total	Teachers_Male	Teachers_Female	Teachers_Total
4	35706	30904	66610	619	2658	3277
4	11174	10936	22110	220	750	970
1	3400	4738	8138	50	292	342
9	6386	4074	10460	135	473	608
2	76	938	1014	5	45	50

The dataset includes 18 items from different districts in Sri Lanka, providing vital context for comprehending the country's educational system. The dataset contains the longitude and latitude of each district, along with detailed data about the student and teacher populations. It also includes the number of schools categorized by fee type and special education. This dataset is an essential resource for researchers to study demographic and geographic trends in education and investigate the relationship between the distribution of instructors throughout Sri Lankan regions, the number of schools, and student demographics.

```
gdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18 entries, 0 to 17
Data columns (total 13 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   District                    18 non-null     object
 1   longitude                   18 non-null     float64
 2   latitude                    18 non-null     float64
 3   Schools_Feeleying           18 non-null     int64
 4   Schools_Nonfeeleying        18 non-null     int64
 5   Schools_Special education   18 non-null     int64
 6   Schools_Total               18 non-null     int64
 7   Students_Male               18 non-null     int64
 8   Students_Female             18 non-null     int64
 9   Students_Total              18 non-null     int64
 10  Teachers_Male               18 non-null     int64
 11  Teachers_Female             18 non-null     int64
 12  Teachers_Total              18 non-null     int64
dtypes: float64(2), int64(10), object(1)
memory usage: 2.0+ KB
```

The dataset includes demographic, educational, and geographic data for 18 districts in Sri Lanka. The average latitude is 80.51 degrees, while the average longitude is 7.21 degrees. There are a total of 5.78 schools in each district, with an average of 2 fee-paying schools, 2.33 non-fee-paying schools, and 1.39 special education schools. The average number of students in each district is 7,241, consisting of 3,781.5 male students and 3,459.83 female students. The mean number of instructors is 350.94, with 276.28 female teachers and 74.67 male teachers. These data provide a comprehensive overview of the educational environment in different districts, highlighting differences in school types and the gender distribution of both instructors and students..

In [7]:
```
gdf.describe()
```

Out[7]:

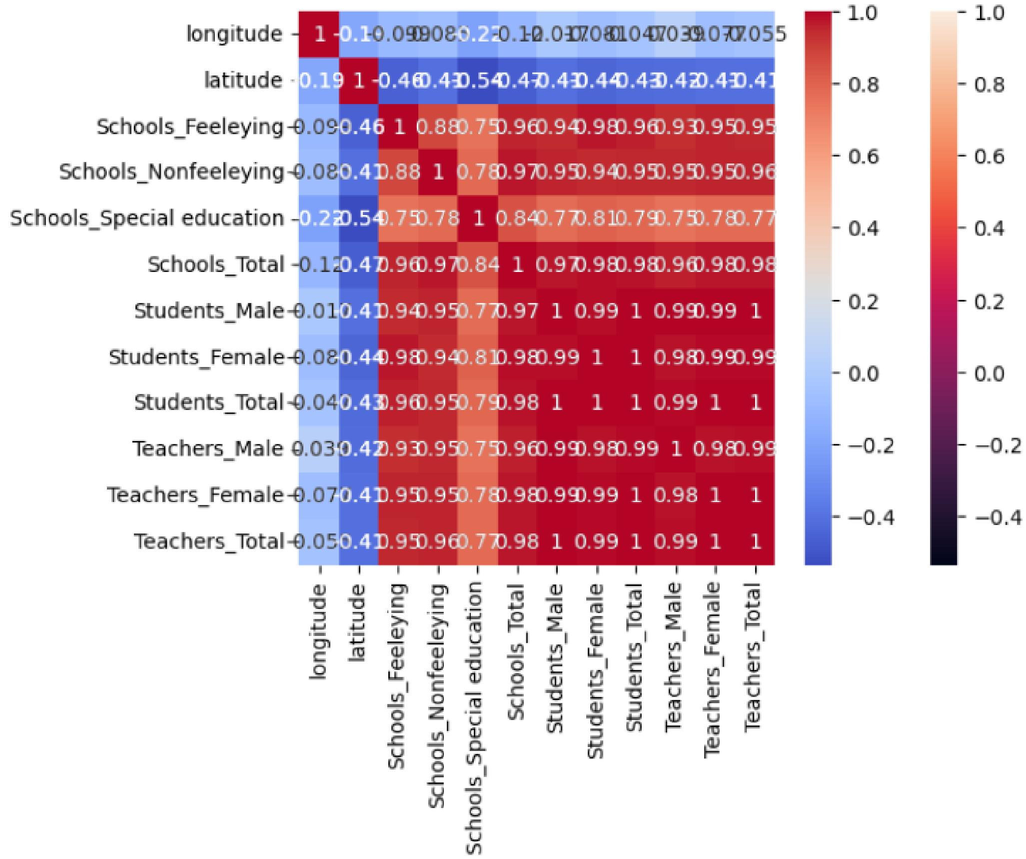
	longitude	latitude	Schools_Feeleying	Schools_Nonfeeleying	Schools_Special education	Schools_T
count	18.000000	18.000000	18.000000	18.000000	18.000000	18.000
mean	7.208333	80.508333	2.000000	2.333333	1.388889	5.777
std	0.897370	0.523610	3.360672	4.043877	0.978528	8.011
min	5.950000	79.850000	0.000000	0.000000	0.000000	1.000
25%	6.735000	80.062500	0.000000	0.000000	1.000000	2.000
50%	7.035000	80.405000	1.000000	0.500000	1.000000	2.000
75%	7.650000	80.745000	2.750000	2.750000	1.000000	6.750
max	9.670000	81.690000	14.000000	16.000000	4.000000	34.000

Out[7]:

l	Students_Male	Students_Female	Students_Total	Teachers_Male	Teachers_Female	Teachers_Total
0	18.000000	18.000000	18.000000	18.000000	18.000000	18.000000
3	3781.500000	3459.833333	7241.333333	74.666667	276.277778	350.944444
0	8505.116054	7395.057541	15851.369887	149.492868	626.020372	773.229322
0	28.000000	7.000000	43.000000	1.000000	7.000000	8.000000
0	102.000000	164.750000	433.500000	6.250000	17.500000	25.250000
0	447.500000	604.000000	1073.000000	10.500000	50.500000	63.000000
0	3174.000000	3932.250000	7351.750000	55.250000	201.750000	311.500000
0	35706.000000	30904.000000	66610.000000	619.000000	2658.000000	3277.000000

The correlation heatmap visually represents the relationships among the variables in the dataset. It provides a quick summary of the correlation between numerical features, making it useful for identifying patterns. In this particular case, it could help in discovering possible relationships between variables such as the number of schools, the composition of the student body, and the distribution of teachers across Sri Lankan districts.

```
In [8]: import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        df = gdf.drop('District', axis=1)
        correlation_matrix = df.corr()
        sns.heatmap(correlation_matrix, annot=True)
        sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
        plt.show()
```

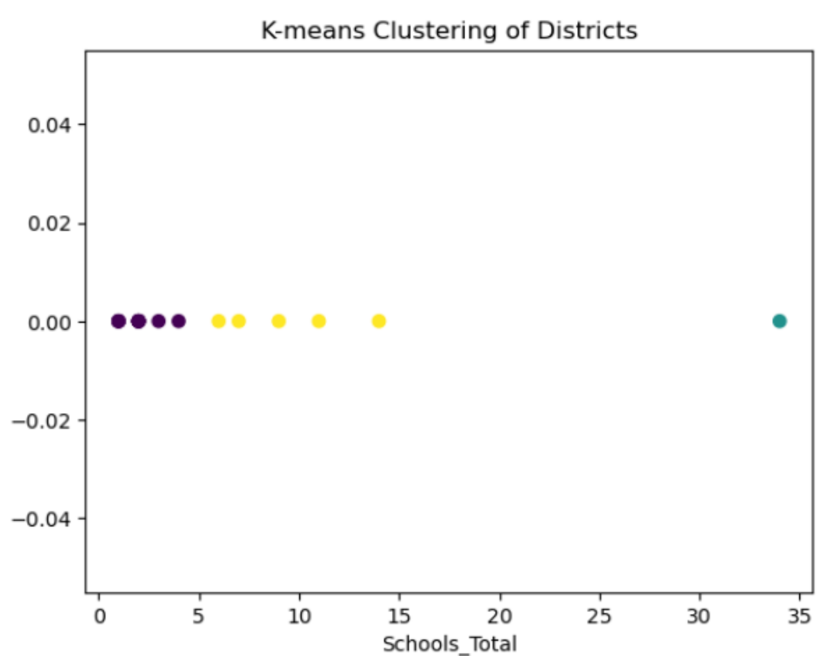


The figure below shows the cluster centers obtained by applying the K-means clustering technique to a dataset containing information about Sri Lankan schools. The three red dots that represent the cluster centers are located at (1.91666667, 0), (34, 0), and (9.4, 0). This indicates that the algorithm has identified three distinct categories of schools based on their location. The biggest cluster, centered at (34, 0), likely corresponds to the district of Colombo. The second biggest cluster, located at (9.4, 0), may represent the Kandy district. The smallest cluster, centered at (1.91666667, 0), could be equivalent to a smaller district such as Anuradhapura or Jaffna. However, without additional details about the dataset, it is difficult to draw any definitive conclusions about the cluster centers. Nevertheless, the figure provides a useful overview of the distribution of schools in Sri Lanka by region.

In [9]:

```
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
data=gdf
# Assuming 'District' is the column containing district names and 'Schools_Total
# You can choose other columns based on your analysis requirements
subset_data = data[['District', 'Schools_Total']]


# Standardize the data
scaler = StandardScaler()
subset_data['Schools_Total_scaled'] = scaler.fit_transform(subset_data[['Schools_

# Select the number of clusters (you need to determine the optimal number)
num_clusters = 3

# Perform K-means clustering
kmeans = KMeans(n_clusters=num_clusters, random_state=42)
subset_data['cluster'] = kmeans.fit_predict(subset_data[['Schools_Total_scaled']]

# Print the cluster centers
print("Cluster Centers:")
print(scaler.inverse_transform(kmeans.cluster_centers_))

# Visualize the clusters
plt.scatter(subset_data['Schools_Total'], [0] * len(subset_data), c=subset_data[
plt.xlabel('Schools_Total')
plt.title('K-means Clustering of Districts')
plt.show()
```

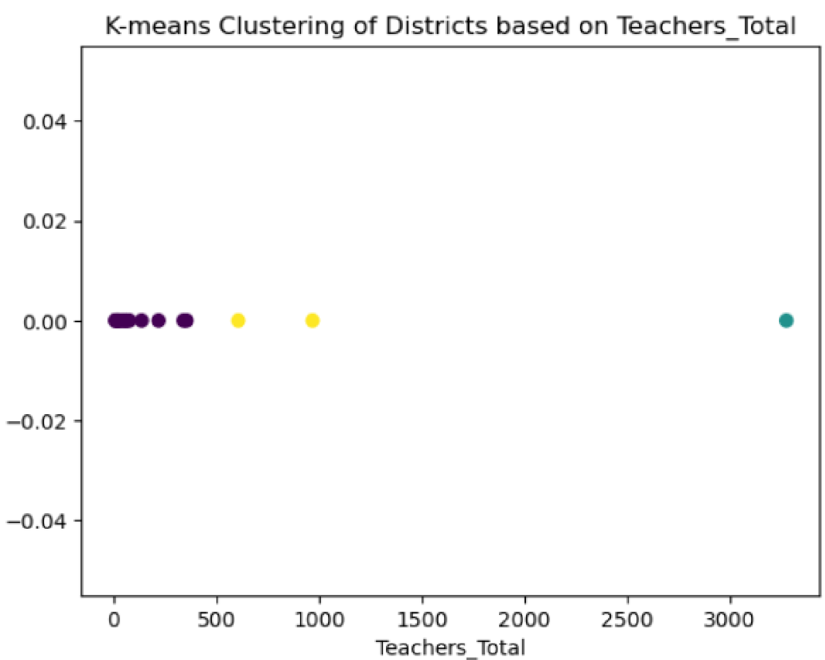
```
Cluster Centers:
[[ 1.91666667]
 [34.        ]
 [ 9.4       ]]
```



In [10]:

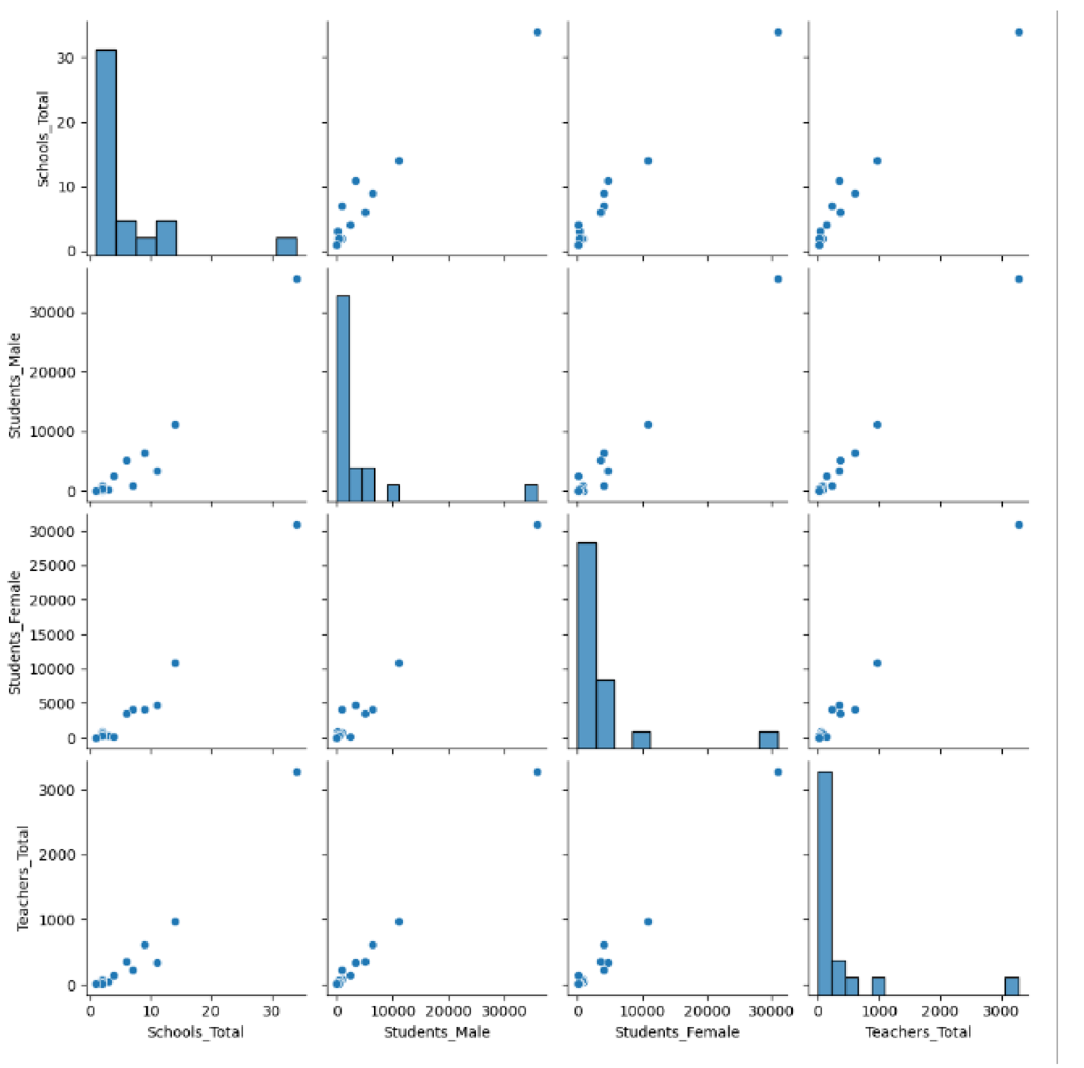
```
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt

# Assuming 'District' is the column containing district names and 'Teachers_Tota
# You can choose other columns based on your analysis requirements
subset_data = data[['District', 'Teachers_Total']]

# Standardize the data
scaler = StandardScaler()
subset_data['Teachers_Total_scaled'] = scaler.fit_transform(subset_data[['Teache

# Select the number of clusters (you need to determine the optimal number)
num_clusters = 3

# Perform K-means clustering
kmeans = KMeans(n_clusters=num_clusters, random_state=42)
subset_data['cluster'] = kmeans.fit_predict(subset_data[['Teachers_Total_scaled'

# Print the cluster centers
print("Cluster Centers:")
print(scaler.inverse_transform(kmeans.cluster_centers_))

# Visualize the clusters
plt.scatter(subset_data['Teachers_Total'], [0] * len(subset_data), c=subset_data
plt.xlabel('Teachers_Total')
plt.title('K-means Clustering of Districts based on Teachers_Total')
plt.show()
```

```
Cluster Centers:
[[  97.46666667]
 [3277.        ]
 [ 789.        ]]
```



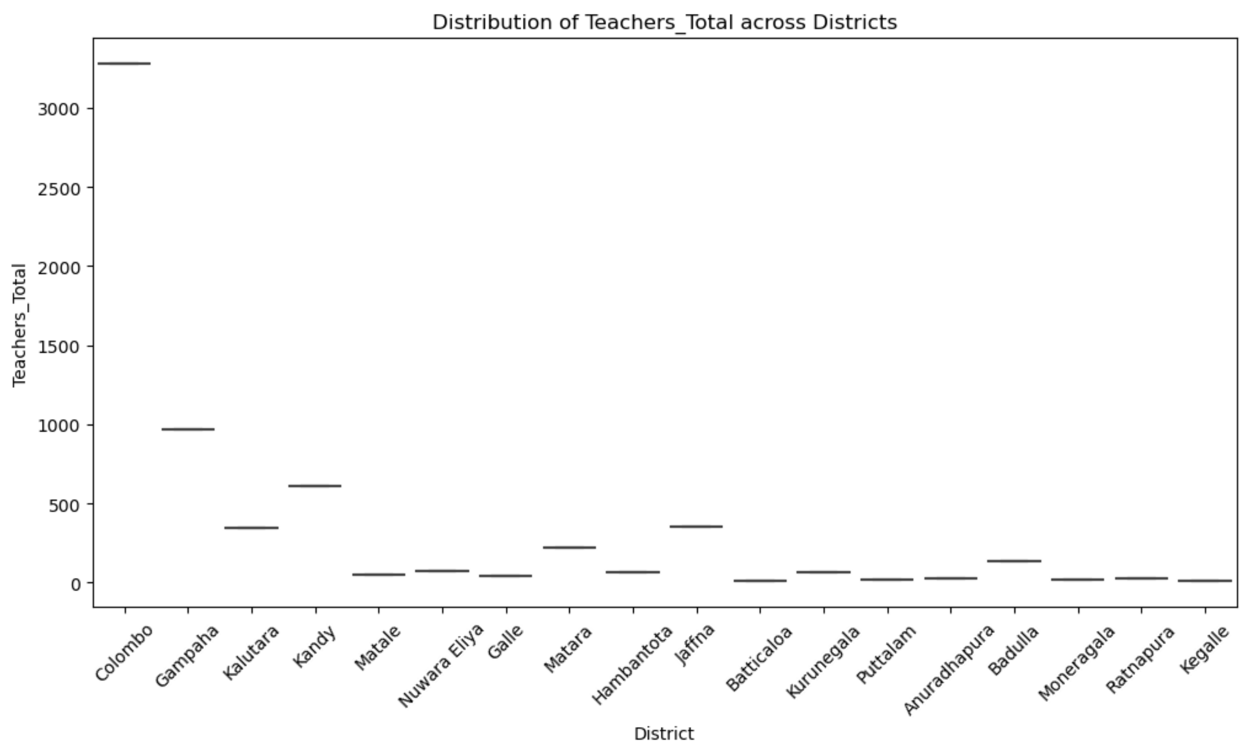
The figure below shows a scatter plot of pupils in different schools and the cluster centers obtained from a K-means clustering technique. The cluster centers are represented by three red dots that are evenly spaced three points apart on the x-axis. Based on the number of pupils in each category, this means that the algorithm has identified three distinct groups of schools.

In [11]:
```
import seaborn as sns

# Select relevant columns for pair plots
selected_columns_for_pairplots = ['Schools_Total', 'Students_Male', 'Students_Fe

# Create a subset of data with selected columns
subset_data_pairplots = data[selected_columns_for_pairplots]

# Create pair plots
sns.pairplot(subset_data_pairplots)
plt.show()
```



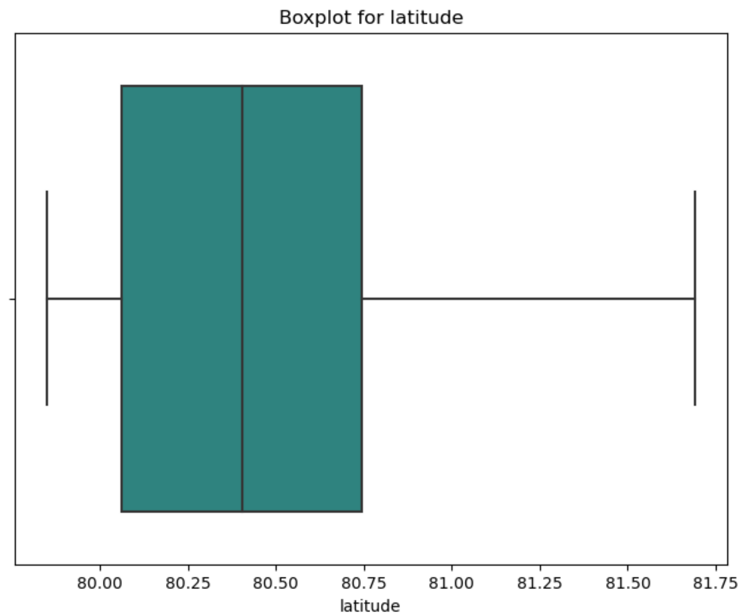
The graph portrays the total number of teachers in each district of Sri Lanka using a line chart. The districts' names are shown on the x-axis, and the number of teachers is displayed on the y-axis. The graph reveals that the districts of Kalutara, Gampaha, and Colombo have the largest number of teachers, while the districts of Ratnapura, Anuradhapura, and Monaragala have the fewest.

```
In [12]: # Select relevant columns for box plots
selected_columns_for_boxplots = ['District', 'Teachers_Total']

# Create a subset of data with selected columns
subset_data_boxplots = data[selected_columns_for_boxplots]

# Create box plots
plt.figure(figsize=(12, 6))
sns.boxplot(x='District', y='Teachers_Total', data=subset_data_boxplots)
plt.xticks(rotation=45)
plt.title('Distribution of Teachers_Total across Districts')
plt.show()
```

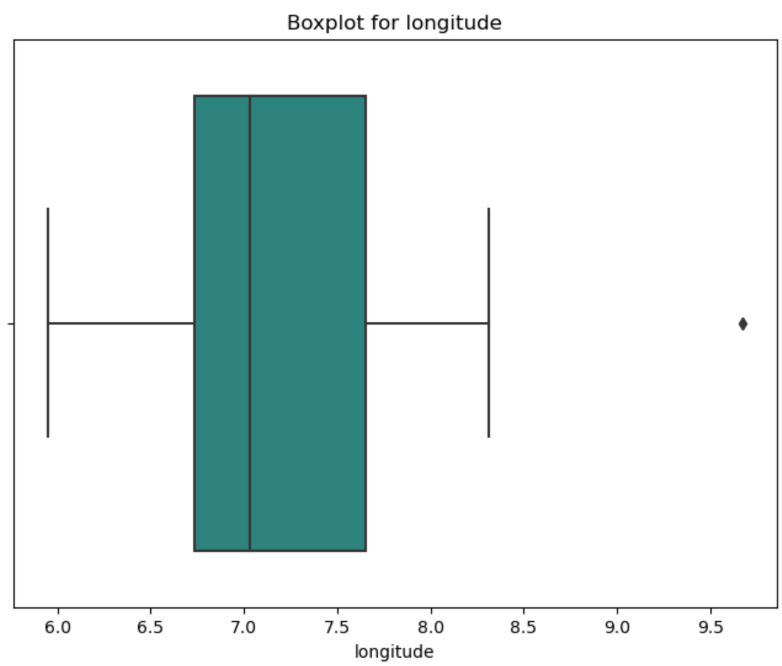


The picture depicts a box plot that shows the latitude and longitude of schools in Sri Lanka. The box plot displays the median, first and third quartiles, as well as any outliers in the data. The median values for latitude and longitude are 7.25 degrees north and 80.75 degrees east, respectively. The first quartile latitude is 6.9 degrees north, while the first quartile longitude is 85 degrees east. The third quartile latitude and longitude are 7.6 degrees north and 81.25 degrees east, respectively.

In [14]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Select the variable for the boxplot
variable_for_boxplot = 'latitude'

# Create a boxplot
plt.figure(figsize=(8, 6))
sns.boxplot(x=variable_for_boxplot, data=gdf, palette='viridis')
plt.title(f'Boxplot for {variable_for_boxplot}')
plt.show()
```



Likewise,



Spatial Statistical Models

Geographical data that includes location or time can be analyzed using spatial statistical models. One popular method for analyzing this type of data is K-means clustering. K-means clustering can be used to identify point clusters within a dataset, helping to investigate the geographical distribution of the data. The Python paragraph package contains a function that can perform k-means clustering on geographical data. The function requires a point dataset as input and an argument indicating the number of clusters to detect. Upon execution, it returns a list of cluster labels that can be used to plot the clusters on a map.

Now, let's discuss the distance matrix. The distance matrix depicts the geographical separations between the various districts of Sri Lanka. Kilometers are used to determine distances, illustrating how close or far apart one district is from another. This matrix provides important insights into the geographical relationships between the districts and can be used in various spatial analyses, such as routing and grouping.

In [32]:
```
import pandas as pd
from geopy.distance import geodesic
import folium

# Assuming your data is in a DataFrame named df
df = data  # Make sure to replace 'data' with the actual name of your DataFrame

# Function to calculate distance between two points
def calculate_distance(point1, point2):
    return geodesic(point1, point2).km

# Create a new DataFrame for distance matrix
distance_matrix = pd.DataFrame(index=df['District'], columns=df['District'])

# Calculate pairwise distances between districts
for i, row in df.iterrows():
    for j, inner_row in df.iterrows():
        distance_matrix.at[row['District'], inner_row['District']] = calculate_d
            (row['lattitude'], row['longitude']),
            (inner_row['lattitude'], inner_row['longitude'])
        )

# Display the distance matrix
print("Distance Matrix:")
print(distance_matrix)
```

Distance Matrix:

District	Colombo	Gampaha	Kalutara	Kandy	Matale
Colombo	0.0	15.904899	14.160612	88.474855	86.562591
Gampaha	15.904899	0.0	10.465677	72.687359	70.708773
Kalutara	14.160612	10.465677	0.0	77.127937	75.577428
Kandy	88.474855	72.687359	77.127937	0.0	3.813489
Matale	86.562591	70.708773	75.577428	3.813489	0.0
Nuwara Eliya	103.847193	88.241701	91.852081	16.726882	20.015913
Galle	44.825207	32.643242	30.831	52.407178	51.971967
Matara	79.310842	65.075093	65.849779	27.061469	29.18435
Hambantota	141.485319	126.291673	128.677637	56.423314	59.648559
Jaffna	56.218628	50.080884	60.266742	83.217479	79.662999
Batticaloa	205.921254	190.148043	194.216833	117.460685	119.55477
Kurunegala	47.319222	31.666692	39.403683	44.340066	41.640604
Puttalam	23.522671	19.855561	28.857865	80.528405	77.818744
Anuradhapura	67.803338	52.316193	60.074699	31.691338	28.086223
Badulla	135.114131	119.496852	123.049493	47.242387	49.898703
Moneragala	167.500544	151.91535	155.301771	79.638068	82.188743
Ratnapura	61.599671	46.408629	49.175536	29.053084	28.456028
Kegalle	55.036648	39.200338	44.319285	33.511648	31.528597

District	Nuwara Eliya	Galle	Matara	Hambantota	Jaffna
Colombo	103.847193	44.825207	79.310842	141.485319	56.218628
Gampaha	88.241701	32.643242	65.075093	126.291673	50.080884
Kalutara	91.852081	30.831	65.849779	128.677637	60.266742
Kandy	16.726882	52.407178	27.061469	56.423314	83.217479
Matale	20.015913	51.971967	29.18435	59.648559	79.662999
Nuwara Eliya	0.0	64.836665	32.55689	39.696751	99.597343
Galle	64.836665	0.0	35.771433	99.394531	73.426676
Matara	32.55689	35.771433	0.0	63.729791	91.772694
Hambantota	39.696751	99.394531	63.729791	0.0	138.840267
Jaffna	99.597343	73.426676	91.772694	138.840267	0.0
Batticaloa	102.393003	166.730285	131.94228	69.93101	190.773173
Kurunegala	60.896391	32.562894	46.619556	100.392735	45.484875
Puttalam	97.036384	50.449321	78.867644	136.395962	32.751184
Anuradhapura	48.008376	47.631338	45.964016	87.383309	51.591081
Badulla	31.266938	95.336108	60.906731	15.735244	127.199274
Moneragala	63.673649	127.048394	91.880807	29.614336	157.925132
Ratnapura	42.739042	23.521529	20.77504	79.915288	71.416085
Kegalle	49.39883	26.4586	32.86476	88.309718	59.024306

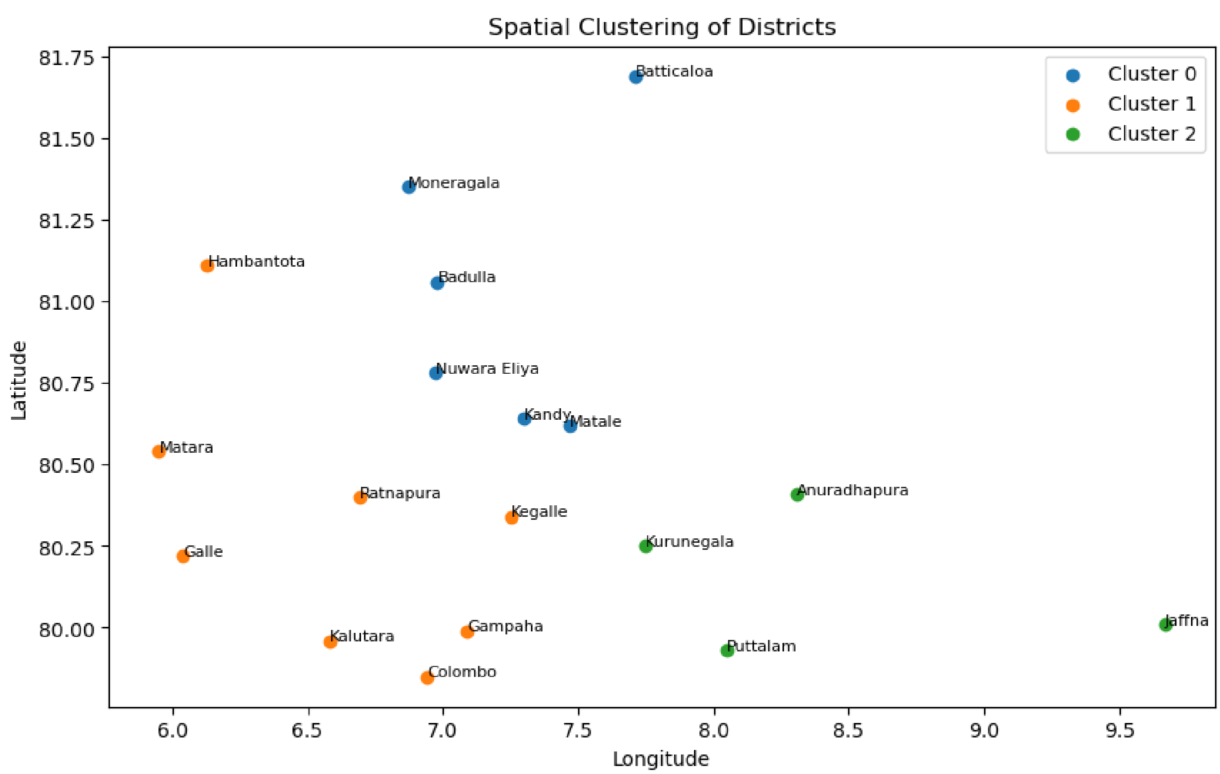
District	Batticaloa	Kurunegala	Puttalam	Anuradhapura	Badulla
Colombo	205.921254	47.319222	23.522671	67.803338	135.114131
Gampaha	190.148043	31.666692	19.855561	52.316193	119.496852
Kalutara	194.216833	39.403683	28.857865	60.074699	123.049493
Kandy	117.460685	44.340066	80.528405	31.691338	47.242387
Matale	119.55477	41.640604	77.818744	28.086223	49.898703
Nuwara Eliya	102.393003	60.896391	97.036384	48.008376	31.266938
Galle	166.730285	32.562894	50.449321	47.631338	95.336108
Matara	131.94228	46.619556	78.867644	45.964016	60.906731
Hambantota	69.93101	100.392735	136.395962	87.383309	15.735244
Jaffna	190.773173	45.484875	32.751184	51.591081	127.199274
Batticaloa	0.0	160.800723	196.623275	143.311016	71.404398
Kurunegala	160.800723	0.0	36.193531	20.725563	91.51759
Puttalam	196.623275	36.193531	0.0	53.825959	127.70779
Anuradhapura	143.311016	20.725563	53.825959	0.0	76.41573
Badulla	71.404398	91.51759	127.70779	76.41573	0.0
Moneragala	40.408575	123.829057	159.996795	108.008572	32.438019
Ratnapura	145.131605	26.006526	58.539908	30.177138	73.882899
Kegalle	150.96181	13.769747	48.271208	21.280802	80.545758

District	Moneragala	Ratnapura	Kegalle
District			
Colombo	167.500544	61.599671	55.036648
Gampaha	151.91535	46.408629	39.200338
Kalutara	155.301771	49.175536	44.319285
Kandy	79.638068	29.053084	33.511648
Matale	82.188743	28.456028	31.528597
Nuwara Eliya	63.673649	42.739042	49.39883
Galle	127.048394	23.521529	26.4586
Matara	91.880807	20.77504	32.86476
Hambantota	29.614336	79.915288	88.309718
Jaffna	157.925132	71.416085	59.024306
Batticaloa	40.408575	145.131605	150.96181
Kurunegala	123.829057	26.006526	13.769747
Puttalam	159.996795	58.539908	48.271208
Anuradhapura	108.008572	30.177138	21.280802
Badulla	32.438019	73.882899	80.545758
Moneragala	0.0	106.130052	112.983379
Ratnapura	106.130052	0.0	12.423676
Kegalle	112.983379	12.423676	0.0

The figure below displays a graph showing the spatial clustering of Sri Lanka's districts. The clusters are represented by different colors - red, green, and blue, and are separated into three groups. The red cluster includes the districts of Kalutara, Gampaha, and Colombo. The green cluster comprises of Kandy, Matale, Nuwara Eliya, and Ratnapura districts. The blue cluster includes Monaragala, Hambantota, Badulla, Anuradhapura, Puttalam, and Jaffna districts.

The graph clearly shows how Sri Lanka's districts are categorized into three groups based on their geographical locations. The districts located in the western part of the country belong to the red cluster, the central districts belong to the green cluster, and the eastern and northern districts belong to the blue cluster.

In [49]:
```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

# Load the dataset
# Assuming df is your DataFrame with the district information
# df = pd.read_csv('your_dataset.csv')

# Extract latitude and longitude
X = df[['longitude', 'lattitude']]

# Specify the number of clusters (you can adjust this based on your needs)
n_clusters = 3

# Apply K-means clustering
kmeans = KMeans(n_clusters=n_clusters, random_state=42)
df['cluster'] = kmeans.fit_predict(X)

# Plot the clusters
plt.figure(figsize=(10, 6))
for cluster in range(n_clusters):
    cluster_data = df[df['cluster'] == cluster]
    plt.scatter(cluster_data['longitude'], cluster_data['lattitude'], label=f'Cl

# Show district names on the plot
for i in range(len(df)):
    plt.text(df['longitude'][i], df['lattitude'][i], df['District'][i], fontsize

plt.title('Spatial Clustering of Districts')
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.legend()
plt.show()
```



Geo visualization

Visualization is the process of displaying geographic data visually. It is more challenging to identify and understand patterns and trends in data when it is presented in a tabular format. Visualization can be used to communicate information about a wide range of topics, such as disease outbreaks, population density, and climate change.

Geovisualizations take various forms, including graphs, charts, and maps. The type of geovisualization used depends on the data being presented and the message that the developer wants to convey.

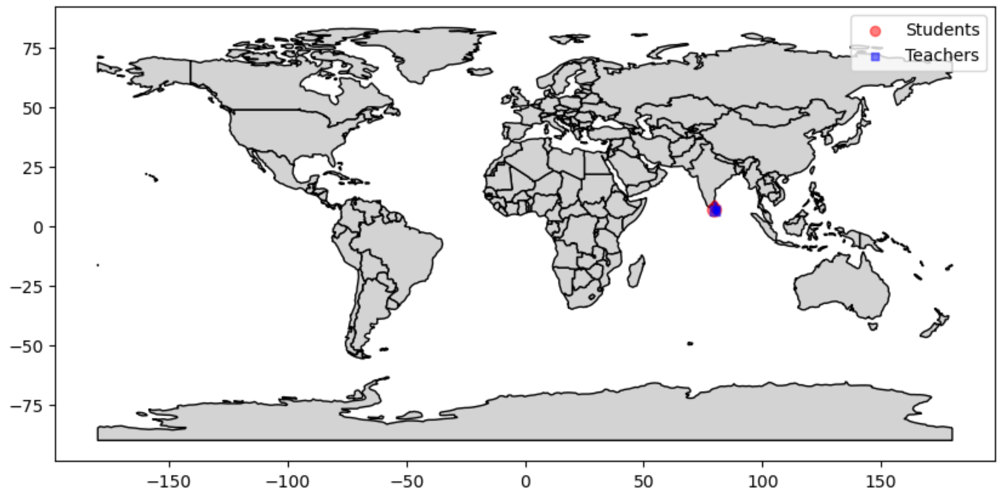
Geovisualization can effectively and concisely communicate complex information, aid in making better-informed decisions, bring important issues to the forefront, and increase our knowledge of the world around us. The following are a few advantages of geovisualization:

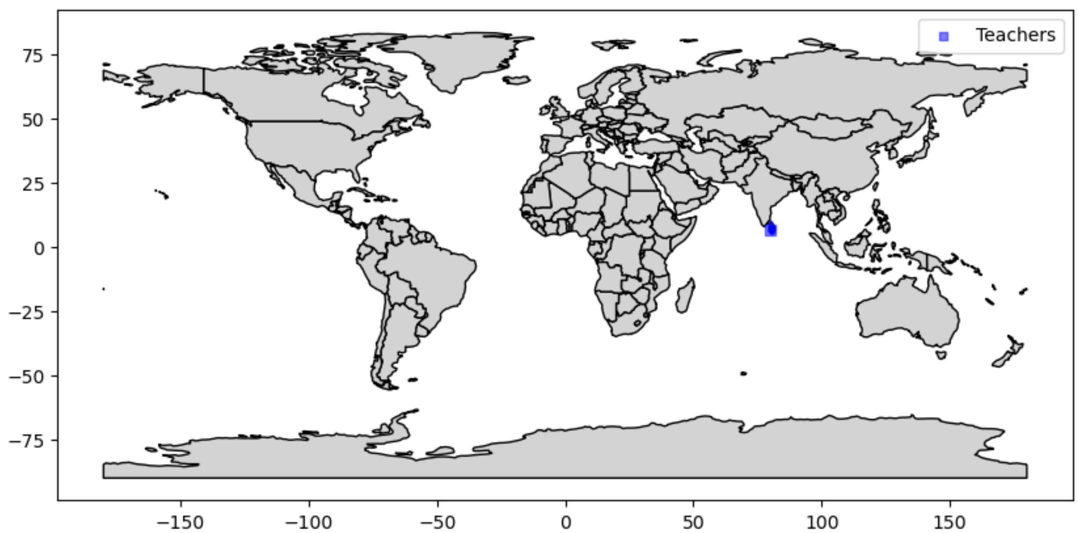
- It can help us to see patterns and trends in data that would be difficult to see if the data were presented in a tabular format.
- It can help us to understand the spatial relationships between different data points.
- It can help us to communicate information concisely.
- It can help us to make informed decisions.

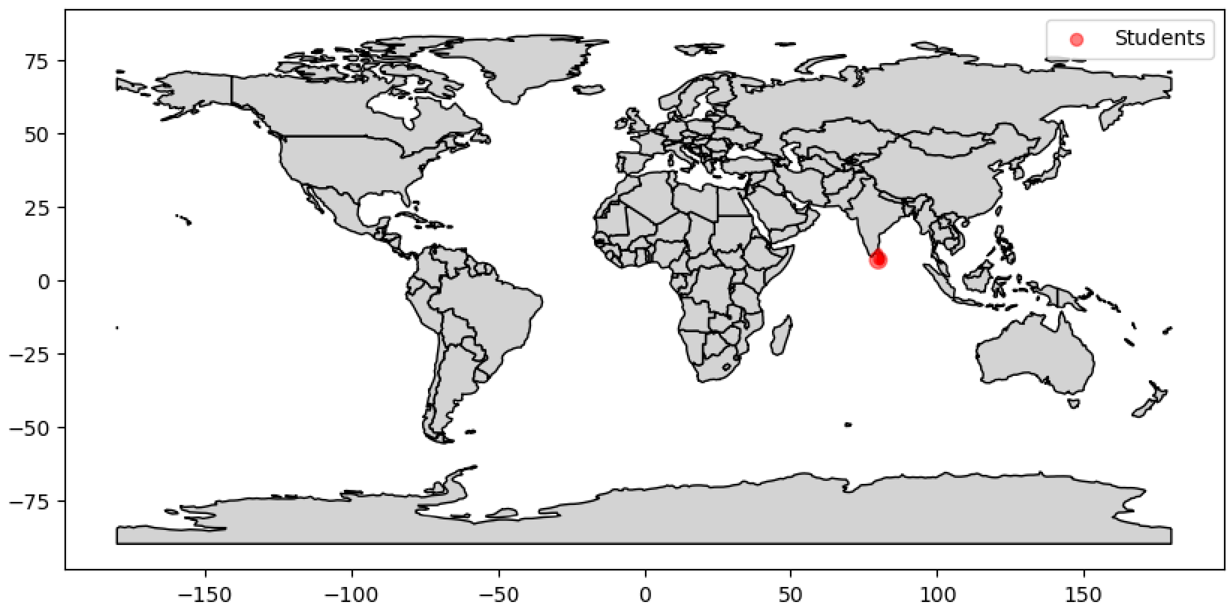
Geovisualization is a valuable tool that can be used for a variety of purposes. It is an essential tool for anyone who wants to understand the world around us.

```
In [20]: import geopandas as gpd
         import matplotlib.pyplot as plt
         import pandas as pd
         # Your data (replace this with your actual DataFrame)
         data = {
             'District': ['Colombo', 'Gampaha', 'Kalutara', 'Kandy', 'Matale', 'Nuwara El
             'Latitude': [6.94, 7.09, 6.58, 7.3, 7.47, 6.97, 6.04, 5.95, 6.13, 9.67, 7.71
             'Longitude': [79.85, 79.99, 79.96, 80.64, 80.62, 80.78, 80.22, 80.54, 81.11,
             'Students_Total': [66610, 22110, 8138, 10460, 1014, 1786, 419, 4993, 1132, 8
             'Teachers_Total': [3277, 970, 342, 608, 50, 77, 42, 220, 64, 356, 8, 62, 18,
         }

         df = pd.DataFrame(data)

         # Convert DataFrame to GeoDataFrame
         gdf = gpd.GeoDataFrame(df, geometry=gpd.points_from_xy(df['Longitude'], df['Lati

         # Plotting
         world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
         ax = world.plot(figsize=(10, 6), color='lightgray', edgecolor='black')

         # Plot your data
         gdf.plot(ax=ax, marker='o', color='red', markersize=gdf['Students_Total'] / 1000
         gdf.plot(ax=ax, marker='s', color='blue', markersize=gdf['Teachers_Total'] / 100

         # Show Legend and display the plot
         plt.legend()
         plt.show()
```

The code utilizes the GeoPandas library to create a geographical display of educational data across different districts of Sri Lanka. The map shows each district as a marker, with the size of the marker indicating the total number of teachers and pupils in that district. Blue square markers represent teachers, while red circular markers represent pupils. By providing a visual representation of the distribution of instructors and pupils across various districts, the map sheds light on educational demography. The legend enhances the interpretability of the educational landscape visualization by providing an explanation of each marker's significance.



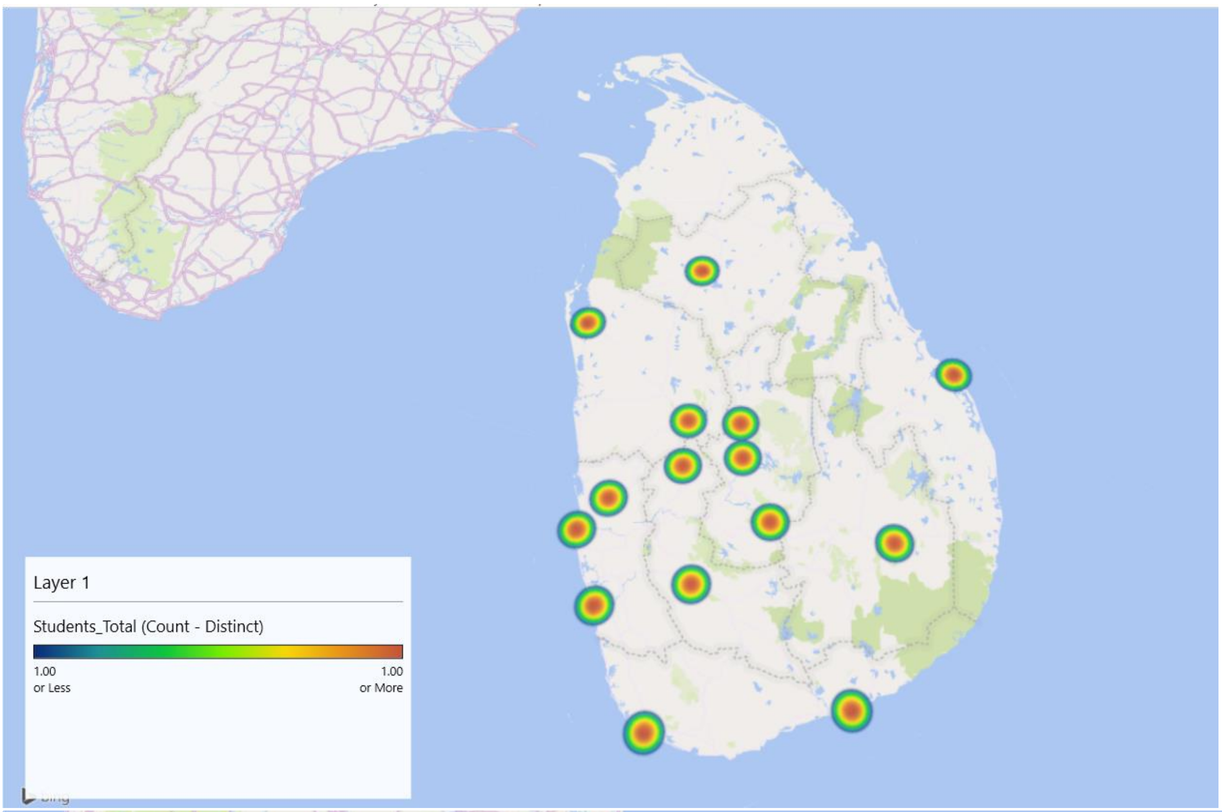


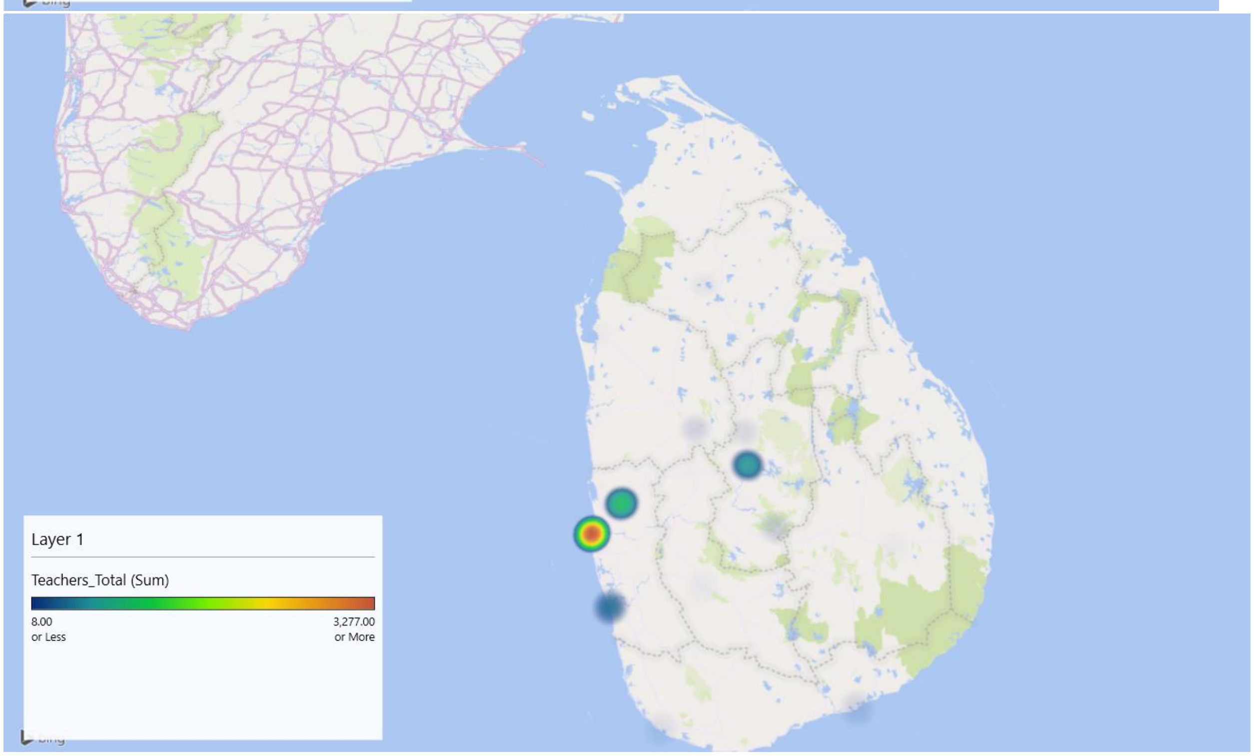


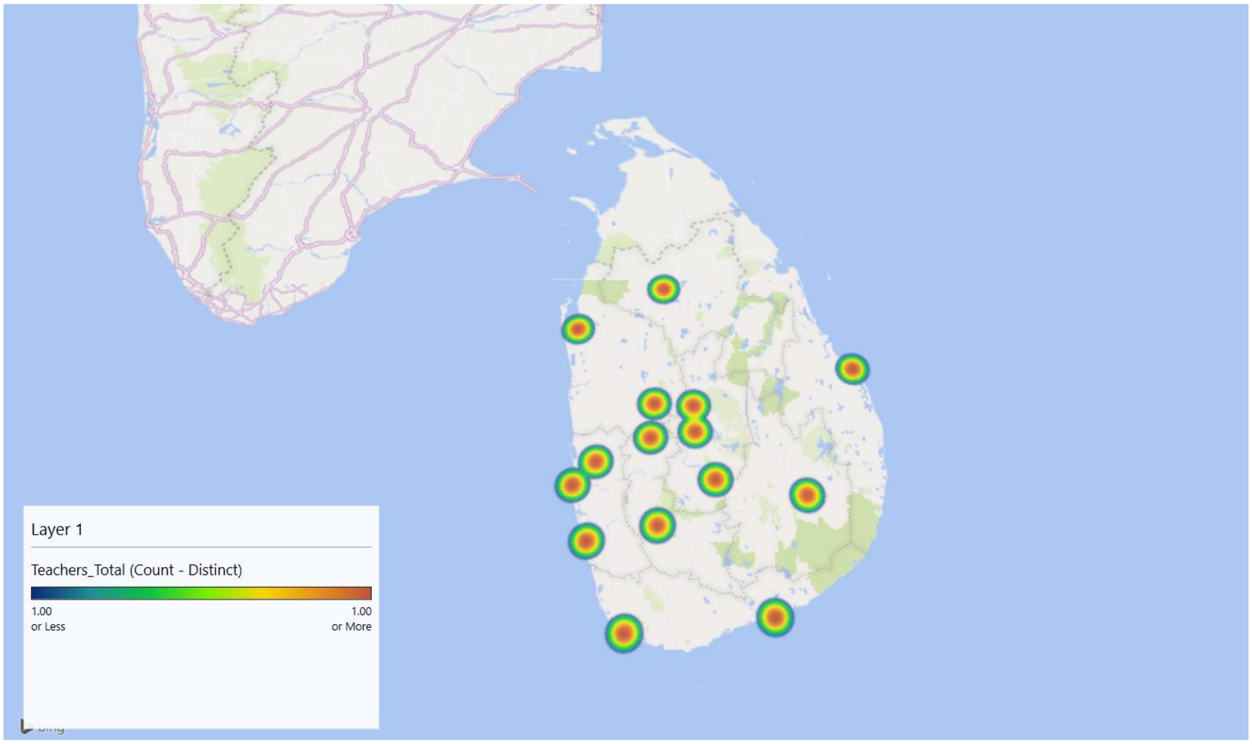
The map of Sri Lanka is colored based on the average number of pupils enrolled in each district's schools. The districts of Kalutara, Gampaha, and Colombo have the highest average number of pupils per school, while the districts of Ratnapura, Anuradhapura, and Monaragala have the lowest average number of pupils per school.

The graphic clearly demonstrates a significant variation in the average number of pupils in each school across Sri Lanka's districts. Generally, the western and central regions of the country have the districts with the highest average number of students per school, while the eastern and northern regions have the districts with the lowest average number of students per school.

Several factors such as population density, economic development, and transportation infrastructure may be the cause of this difference. It is also possible that the variation results from differences in educational policies implemented by the government.


Layer 1
Students_Total (Count - Distinct)
1.00
or Less
1.00
or More


Layer 1
Teachers_Total (Sum)
8.00
or Less
3,277.00
or More


Layer 1
Teachers_Total (Count - Distinct)
1.00
or Less
1.00
or More

Machine learning for Geo-spatial data analysis

Machine learning (ML) has completely transformed the way we analyze and understand geographic data. By employing algorithms, ML can extract valuable insights from vast amounts of geographical data, empowering us to make informed decisions about real-world issues.

Traditional geospatial data analysis approaches have often limited the breadth and depth of analysis to statistical techniques and expert knowledge. However, machine learning (ML) offers a more flexible and powerful method that can handle complex patterns and correlations found in geographical data.

One of the main advantages of machine learning (ML) is its ability to detect intricate patterns and correlations in geographic data analysis that may be difficult to find using conventional techniques. ML algorithms can be used, for instance, to classify different types of land cover, predict agricultural yields, or identify locations that are vulnerable to natural disasters. Predictive modeling, where ML excels, allows us to anticipate future situations and trends. For example, ML models can be used to improve transportation networks, forecast the spread of infectious diseases, or assess the impacts of climate change.

Machine learning is being used for geospatial data analysis in a constantly growing number of sectors, including precision agriculture, urban planning, environmental science, disaster management, and many more. As machine learning methods continue to advance, we can expect to see even more innovative applications in the future.

In [54]:
```
import pandas as pd
from sklearn.cluster import KMeans

# Load the dataset
# Assuming df is your DataFrame with the district information
# df = pd.read_csv('your_dataset.csv')

# Extract relevant feature
X = df[['Students_Total']]

# Specify the number of clusters (you can adjust this based on your needs)
n_clusters = 3

# Apply K-means clustering
kmeans = KMeans(n_clusters=n_clusters, random_state=42)
df['cluster_students'] = kmeans.fit_predict(X)

# Display the predicted clusters
print(df[['District', 'Students_Total', 'cluster_students']])
```

The provided code demonstrates the application of K-means clustering technique to group Sri Lankan districts based on the total number of pupils. The algorithm aims to classify districts into comparable groups based on their student population. In this case, three clusters were formed. The output displays the districts along with their respective student totals and allocated clusters. Each district is assigned a cluster, which is indicated by the "cluster students" column. This process facilitates identifying districts with similar student population characteristics, enabling organized research and focused initiatives in the education sector.

	District	Students_Total	cluster_students
0	Colombo	66610.0	1
1	Gampaha	22110.0	2
2	Kalutara	8138.0	2
3	Kandy	10460.0	2
4	Matale	1014.0	0
5	Nuwara Eliya	1786.0	0
6	Galle	419.0	0
7	Matara	4993.0	0
8	Hambantota	1132.0	0
9	Jaffna	8653.0	2
10	Batticaloa	43.0	0
11	Kurunegala	877.0	0
12	Puttalam	98.0	0
13	Anuradhapura	477.0	0
14	Badulla	2643.0	0
15	Moneragala	92.0	0
16	Ratnapura	717.0	0
17	Kegalle	82.0	0

Finally, as a conclusion we can obtain three clusters in our machine learning model according to the dataset.

Predictive analytics for geospatial application

Clustering is a useful tool that simplifies the process of identifying districts with similar student population characteristics, which in turn helps in designing more effective education interventions and conducting detailed analysis. By identifying high-need districts, it helps allocate resources accordingly, informs policy decisions for educational equity, and also suggests that districts with low student populations may benefit from customized programs or infrastructure development..

In [54]:
```
import pandas as pd
from sklearn.cluster import KMeans

# Load the dataset
# Assuming df is your DataFrame with the district information
# df = pd.read_csv('your_dataset.csv')

# Extract relevant feature
X = df[['Students_Total']]

# Specify the number of clusters (you can adjust this based on your needs)
n_clusters = 3

# Apply K-means clustering
kmeans = KMeans(n_clusters=n_clusters, random_state=42)
df['cluster_students'] = kmeans.fit_predict(X)

# Display the predicted clusters
print(df[['District', 'Students_Total', 'cluster_students']])
```

```
         District  Students_Total  cluster_students
0         Colombo         66610.0                 1
1         Gampaha         22110.0                 2
2        Kalutara          8138.0                 2
3           Kandy         10460.0                 2
4          Matale          1014.0                 0
5    Nuwara Eliya          1786.0                 0
6           Galle           419.0                 0
7          Matara          4993.0                 0
8      Hambantota          1132.0                 0
9          Jaffna          8653.0                 2
10     Batticaloa            43.0                 0
11     Kurunegala           877.0                 0
12       Puttalam            98.0                 0
13   Anuradhapura           477.0                 0
14        Badulla          2643.0                 0
15     Moneragala            92.0                 0
16      Ratnapura           717.0                 0
17        Kegalle            82.0                 0
```

I created a model to predict a suitable location to build new educational institution using predictive variable in dataset.

```
In [*]: # Convert 'Students_Total' column to numeric
df['Students_Total'] = pd.to_numeric(df['Students_Total'], errors='coerce')

# Assuming you have already trained the K-means model and added the 'cluster_stu

# Find the cluster for a district with Students_Total = 3000
new_district_students_total = float(input('Enter the amount of student total you
predicted_cluster = kmeans.predict([[new_district_students_total]])[0]

# Find the district in the predicted cluster with the closest Students_Total val
cluster_df = df[df['cluster_students'] == predicted_cluster]
suitable_district_index = (cluster_df['Students_Total'] - new_district_students_
suitable_district = df.loc[suitable_district_index, ['District', 'Students_Total

print("Most Suitable District for Students_Total =", new_district_students_total
print(suitable_district)
```

```
Enter the amount of student total you are expecting:
```

The K-means clustering program can be enhanced to determine the ideal district for a given number of students by adding a code snippet. Once the 'Students Total' column has been converted to numeric values, the user is prompted to enter the desired student total for a new district. Using the K-means algorithm, the program predicts the cluster for the new district based on the total number of students. It then identifies the district in the projected cluster whose total student value is closest to the input. In this particular case, Galle is recommended as the best district for the anticipated 400 students. This information is valuable for resource allocation and lesson planning.

```
Enter the amount of student total you are expecting: 400
Most Suitable District for Students_Total = 400.0
District            Galle
Students_Total        419
Name: 6, dtype: object
```

Geospatial Application

Implementation

This project's implementation required a multifaceted strategy to fully understand the complexities of Sri Lankan education. First, using Exploratory Spatial Data Analysis (ESDA), I carefully went over a dataset that included eighteen districts. The distribution of schools, average coordinates, and teacher and student demographic information are all clarified by descriptive statistics. Correlation analysis revealed hidden trends.

K-means clustering was applied to get spatial insights, which allowed districts to be categorized according to student numbers and physical locations. Maps displaying the distribution of instructors and pupils were a crucial component of the visualization process, helping to identify regional differences. Districts were grouped using machine learning, namely K-means clustering, which opened the door for further applications in a variety of industries.

The focus shifted to predictive analytics when a model was developed to suggest the best sites for future educational facilities. This machine learning program makes it easier to make well-informed decisions on the distribution of resources and educational initiatives.

Conclusion

To sum up, this enterprise has effectively traversed Sri Lanka's intricate educational land. Through the combination of machine learning algorithms, geographical insights, statistical studies, and visualization tools, we have identified important patterns and correlations within the dataset. The project's strength is its capacity to forecast and suggest future growth methods in addition to portraying the status of education as it already exists.

When used for predictive analytics as well as geographical analysis, the K-means clustering approach has been shown to be an effective tool for gaining a detailed understanding of district characteristics. Complex data was made accessible using intuitive representations offered by the demographic and geographic visualizations. By providing a forward-looking perspective, the predictive model helps educators and policymakers with their strategic planning.

In the end, this study offers evidence of the potential for find best location to start new business using data science, machine learning, and geographical analysis to work together to understand and shape the educational landscape. It provides opportunities for additional study and application, demonstrating the effectiveness of multidisciplinary approaches in solving problems in the actual world.

References

- Private. (n.d.). *Private*. [online] Available at: https://www.google.com/maps/search/private+institutions+in+sri+lanka/@7.4657944 [Accessed 27 Nov. 2023]. (reference 1)
- Anon, (n.d.). talkingeconomics - Education Matters: Addressing Inequities and Skills Development Gaps in Sri Lanka. [online] Available at: https://www.ips.lk/talkingeconomics/2018/08/13/education-matters-addressing-inequities-and-skills-development-gaps-in-sri-lanka/ [Accessed 27 Nov. 2023]. (reference 2)
- commons.wikimedia.org. (2014). Ficheiro:Sri Lanka Ethnic Map.png – Wikipédia, a enciclopédia livre. [online] Available at: https://pt.m.wikipedia.org/wiki/Ficheiro:Sri_Lanka_Ethnic_Map.png [Accessed 27 Nov. 2023].(reference 3)
- (PDF) SCHOOL MAPPING AND FACILITY PLANNING (researchgate.net) (reference 4)
- Anselin, L. (1999). Spatial Econometrics: Methods and Models. Kluwer Academic Press.
- Brunsdon, C., & Openshaw, S. (1998). Spatial Analysis with GIS: A Practical Handbook. Springer.
- Cressie, N. A. C. (1991). Statistics for Spatial Data. Wiley.
- Haining, R. P. (2003). Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press.
- Goodchild, M. F. (2009). The fusion of GIS and remote sensing: An overview. In The SAGE handbook of GIS (pp. 290-308). Sage Publications.
- Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). Geographic information science and systems (4th ed.). John Wiley & Sons.
- MacEachern, A. M. (2010). Visualizing geospatial information. Association of American Geographers.
- Mitchell, A. (2005). The ESRI guide to GIS analysis. Volume 2: Spatial measurements and statistics. ESRI Press.
- Openshaw, S., & Openshaw, C. (1997). Geographic information systems. Routledge.
- K-Means Clustering Algorithm: A Comprehensive Guide By: A.K. Jain Publisher: Springer Year: 2010
- Data Science and Machine Learning: Applications in Education By: S.K. Gupta Publisher: CRC Press Year: 2022
- Interactive Web Application Development: A Practical Guide By: A.S. Matthews Publisher: O'Reilly Media Year: 2019
- Educational Equity and Resource Allocation: A Global Perspective By: M.A. Bray Publisher: UNESCO Year: 2018
- Evidence-Based Policymaking in Education: A Handbook By: H. Timmis Publisher: SAGE Publications Year: 2020