Governing Bengaluru (Bangalore)

1.Introduction/Business Problem

1.1. Background

Bengaluru (formerly and commonly known as Bangalore) is one of the fastest growing cities in the world. Many people move to Bengaluru to pursue career opportunities, a grand city life, and the fantastic weather. People often refer to it as the 'Silicon Valley of India' as it is a hotbed for Information Technology, Artificial Intelligence, and Data Science. As the infrastructure and population continue to grow, efficient governance has become a problem. Many argue over the split of the Bruhat Bengaluru Mahanagara Palike (BBMP) and whether splitting up Bengaluru would promote smoother administration.

1.2. Problem

Bengaluru is a very diverse place, containing different cultures and spanning over both urban and rural territory. Splitting up Bengaluru, so that similar policies/projects can be efficiently co-implemented is a difficult problem to pursue intuitively. Within a small radius, one can find small fish markets and some of the most modern malls. Thus, a data science based approach may prove useful to solve this issue.

1.3. Interest

The policy-makers in Bengaluru would be the interested party in such an analysis. Splitting up Bengaluru into similar neighborhoods would aid in smoother administration, and policies can be geared to solve problems that are likely similar within similar neighborhoods.

2. Data

2.1. Source

Kaggle is a community based environment for data scientists and machine learning enthusiasts. Google created this environment, and it is a fantastic place for people to share data and projects. The data for Bengaluru neighborhoods was obtained from the site, and the original data source would be a central Indian website 'data.gov.in'.

2.2. Data Cleaning

The data obtained look to have some obviously incorrect outliers. Places that are in Bengaluru cannot have such a wide range of latitudes and longitudes. So outliers are

removed and replaced by values found online. In addition, duplicate neighborhoods and excessive columns were removed to obtain a clean dataset.

2.3. Data Usage

The data contains latitudes and longitudes that can be used in foursquare to procure nearby venues. With this data, one can cluster similar neighborhoods together. This can be used to answer the question of how to split up Bengaluru into similar neighborhoods for effective governance. Table 1 is an example of how the foursquare data is received and input into a dataframe. Knowing the most common venues can predict which neighborhoods are similar to each other. A specific example would be places with many shopping malls and pubs might be a determining factor in grouping similar neighborhoods.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Achitnagar	Recreation Center	Bakery	Asian Restaurant	Restaurant	Dry Cleaner	Eastern European Restaurant	Electronics Store	Event Service	Event Space	Yoga Studio
1	Adugodi	Indian Restaurant	Café	Pizza Place	Coffee Shop	Gym	Lounge	Chinese Restaurant	Clothing Store	Tea Room	Dessert Shop
2	Agram	Indian Restaurant	Hotel	Restaurant	Ice Cream Shop	Asian Restaurant	Pub	Clothing Store	Bar	Pizza Place	Café
3	Akkur	Fast Food Restaurant	Bus Station	Yoga Studio	Duty-free Shop	Flea Market	Fishing Spot	Financial or Legal Service	Field	Farmers Market	Farm
4	Alahalli	Food & Drink Shop	Indie Movie Theater	Duty-free Shop	Food	Flea Market	Fishing Spot	Financial or Legal Service	Field	Fast Food Restaurant	Farmers Market

Table 1: Most Common Venues for Bengaluru Neighborhoods

3. Methodology

3.1. Exploratory Data Analysis

Bengaluru is a growing city that locates its international airport outside of the city center. It used to have its airport in the city, but demands for a much bigger and international airport forced the government's hand. Looking into the data and the history of the city it is expected that there will be a large number of neighborhoods in the center and the number of neighborhoods that surround will be of lower concentrations. Furthermore, with the airport located far from the city center, it is expected that there will be a cluster of neighborhoods that have accumulated near that location. Analyzing the data backs these claims with many neighborhood locations (latitude and longitude) clustered around the city center. We also see a higher than usual number of neighborhoods near the airport compared to locations equally distant from the city center.

3.2. Inferential statistical testing

The data definitely has some outliers that look like it was wrongly retrieved or misinput into the data set. No location in Bangalore can have a negative longitude when its city center has a value in the positive seventies. To weed out the outliers, the statistical tool z-score was used. The z-score represents how close or far an observation is from the overall mean. For a normal distribution as shown in the Figure 1 below, 99.7% of the data is within three standard deviations of the mean. Values outside this range are considered outliers. The z-score provides an impactful statistical tool to find outliers. Using python's scipy library, the z-scores of the locations were retrieved. After the outliers were highlighted, they could be replaced manually with correct latitudes and longitudes found online. This created an updated dataset with correct values, verified at the end with a map which had locations in the correct places.

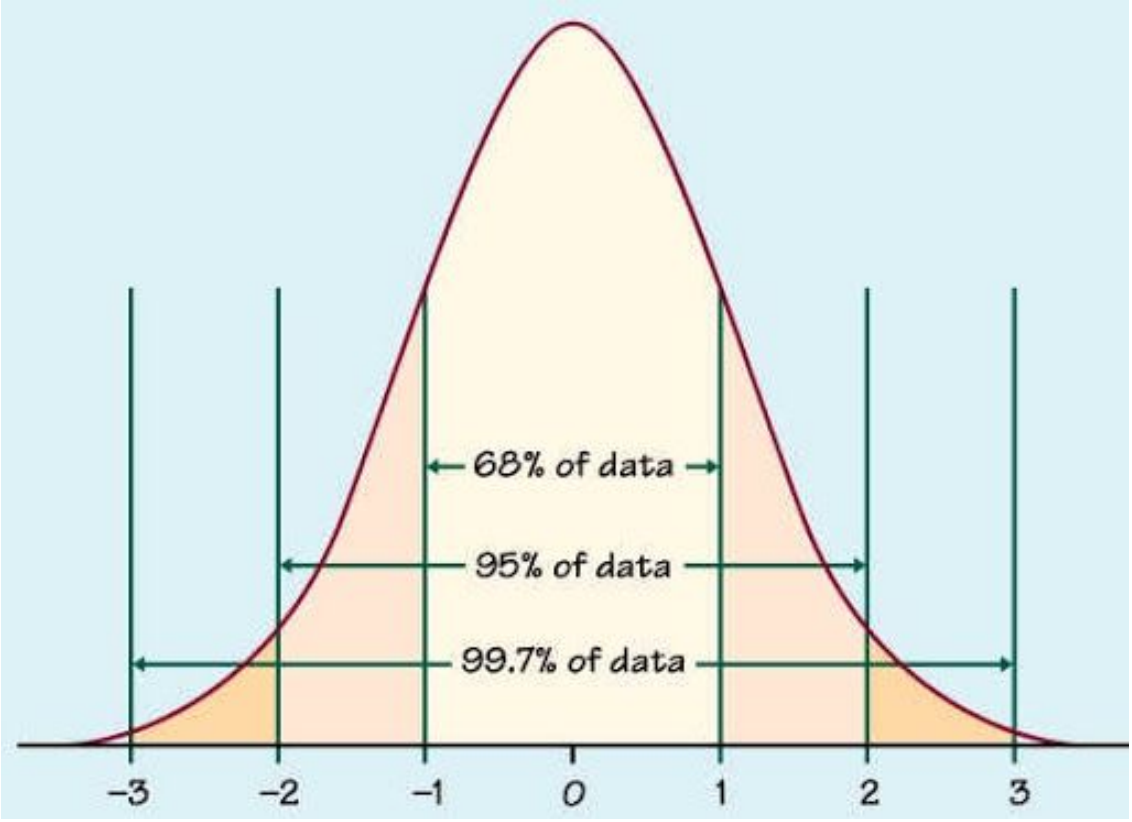

Figure 1: Normal Distribution Z-Scores
(Source:http://www.ltcconline.net/greenl/courses/201/probdist/zScore.htm)

3.3. Data Science and Machine Learning Techniques

First, foursquare data was used to collect venues nearby to each neighborhood. Foursquare is a great location to retrieve location data in a developer friendly manner. Using each latitude and longitude, calls were made to the website retrieving the nearby venues. Once received, the basis for the analysis is obtained. These categorical variables can be used to identify similar neighborhoods. For a numerical algorithm to be used, the categorical variables need to be converted to numerical values, which can be done by using dummy variables. Each venue type is put into a separate column, and if the location has the venue type it will be recorded as one, and if it isn't present it is recorded as zero. This provides a consistent basis for comparison. Table 2 below shows the resultant table ready for machine learning to be used.

	Neighborhood	ATM	Accessories Store	Afghan Restaurant	Airport	Airport Service	Airport Terminal	American Restaurant	Andhra Restaurant	Arcade	...	Turkish Coffeehouse	Udupi Restaurant	Vegetarian / Vegan Restaurant
0	Achitnagar	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.00	0.000000
1	Adugodi	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.00	0.000000
2	Agram	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.01	0.000000
3	Akkur	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.00	0.000000
4	Alahalli	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.00	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
208	Whitefield	0.0	0.0	0.0	0.000000	0.0	0.0	0.027027	0.000000	0.0	...	0.0	0.00	0.013514
209	Yadavanahalli	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.00	0.000000
210	Yelachenahalli	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.00	0.021505
211	Yelahanka	0.0	0.0	0.0	0.032258	0.0	0.0	0.032258	0.032258	0.0	...	0.0	0.00	0.032258
212	Yeliyur	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.00	0.000000

Table 2: One Hot Encoding

K-means clustering was the primary technique used to analyze the data. This technique is an unsupervised machine learning methodology that uses attempts to cluster similar data points together in a way that minimizes the intra-distance within a cluster and maximize the inter-distance between clusters. Due to the large number of neighborhoods and the diversity of Bengaluru, the number of clusters chosen was 20. This provides sufficient granularity to answer the questions posed in the introduction.

Finally using Folium, the map with color coded clusters is obtained. This visually shows the clustering for Bengaluru neighborhoods. Tables are made to further analyze the data and see what is similar between neighborhoods using the ten most common venues amongst the locations. From this conclusions can be drawn on how to split up Bengaluru in order to effectively govern.

4. Results

For each neighborhood, cluster labels are assigned based on which neighborhoods are most similar or dissimilar. Many locations have Indian restaurants as the most common venue. Table 3 shows the ten most common venues for each location, along with its cluster label. This will be used for plotting the map that depicts the neighborhood clusters and provides a basis for understanding why certain locations were put in certain clusters.

Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Agram	12.958000	77.630800	13.0	Indian Restaurant	Hotel	Restaurant	Ice Cream Shop	Asian Restaurant	Pub	Clothing Store	Bar	Pizza Place	Café
Amruthahalli	13.066513	77.596624	1.0	Indian Restaurant	Café	Ice Cream Shop	Fast Food Restaurant	Garden	Resort	Pizza Place	Lake	Light Rail Station	South Indian Restaurant
Banaswadi	13.014162	77.651854	1.0	Indian Restaurant	Café	Ice Cream Shop	Department Store	Pizza Place	Korean Restaurant	BBQ Joint	South Indian Restaurant	Kerala Restaurant	Steakhouse
Bellandur	12.930400	77.678400	13.0	Café	Indian Restaurant	Pizza Place	Coffee Shop	Fast Food Restaurant	Hotel	Ice Cream Shop	Lounge	Gym	Sports Bar
Bhattarahalli	13.025800	77.714279	13.0	Hotel	Café	Bar	Indian Restaurant	Korean Restaurant	Convenience Store	Construction & Landscaping	Pizza Place	Vegetarian / Vegan Restaurant	Forest

Table 3: Head of Most Common Venues and Cluster Labels

Table 4 below highlights three key clusters - 1,6,13. Cluster 1 seems to be citizen hubs. People who lived in Bengaluru for long tend to be in these hubs, which is why Indian restaurants are the most common venue. Additionally we find many local types of venues such as shopping malls, bakeries, and yoga studios. Cluster 13 is the travel heavy neighborhoods. The cluster of neighborhoods surrounding the airport fall in this category. The ones that are closer to the city are heavy on the hotels, lounges and international cuisine restaurants. Finally, cluster 6 has the train station as the primary venue. The daily wage workers who ply their trade in flea markets, fishing spots and fields encompass these neighborhoods.

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
14	Fraser Town	12.997000	77.614400	1.0	Indian Restaurant	Café	Tea Room	Middle Eastern Restaurant	Bakery	Pub	Ice Cream Shop	Shopping Mall	Coffee Shop	Fast Food Restaurant
54	Chandapura	12.801700	77.711600	1.0	Indian Restaurant	Asian Restaurant	Coffee Shop	Train Station	Yoga Studio	Farm	Electronics Store	Event Service	Event Space	Falafel Restaurant
165	Doddajala	13.176735	77.652050	13.0	Fast Food Restaurant	Hotel	Airport Terminal	Bike Shop	Farm	Toll Booth	Food Truck	Airport	Event Service	Food & Drink Shop
241	Korati	12.971600	77.594600	13.0	Indian Restaurant	Hotel	Lounge	Brewery	Pub	Café	Ice Cream Shop	Italian Restaurant	Sushi Restaurant	Japanese Restaurant
264	Malur	13.006034	77.938284	6.0	Train Station	ATM	Men's Store	Duty-free Shop	Flea Market	Fishing Spot	Financial or Legal Service	Field	Fast Food Restaurant	Farmers Market
274	Marasandra	12.980402	77.873983	6.0	Train Station	Yoga Studio	Duty-free Shop	Food	Flea Market	Fishing Spot	Financial or Legal Service	Field	Fast Food Restaurant	Farmers Market

Table 4: Three Important Clusters

Figure 2 below shows the clustering of neighborhoods on the Bengaluru map. As expected, many neighborhoods are in the city center, and as one goes further from it, the number of neighborhoods drop. One notable exception is near the airport. Two primary clusters formed depicted in purple and green. They seem to be spread both in the city center and in the outskirts. Other clusters seem to be separated and only in the outskirts, shown by different shades of orange and red. Finally, it looks as though there is significant neighborhood clustering around the highways, highlighting their importance to any level of government.

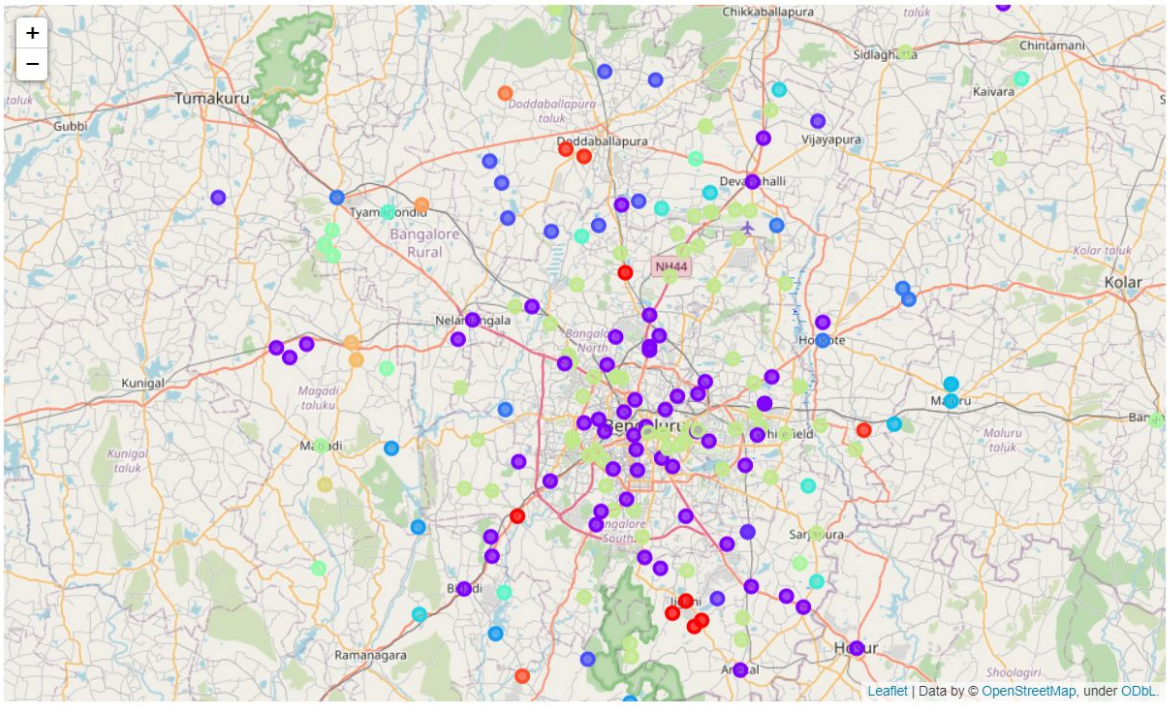

Figure 2: Map of Neighborhood Clusters

5.Discussion

The primary focus of the government should be to provide accurate and impactful funding in locations of need. The clusters provide a strong framework to do so. Splitting Bengaluru up into smaller sections by location would result in severe inefficiencies as neighborhoods near each other aren't necessarily similar. Splitting by type allows the government to focus limited resources in projects that can have the most impact. Furthermore, places that need the same kind of help can be assisted simultaneously. Cluster 1 would benefit most from housing projects. These are places with shopping malls, bakeries and yoga studios for the people who have settled down in Bengaluru. As the population grows, housing prices will spike. A proactive approach could aid future generations. Cluster 13 would benefit from infrastructure rebuilding. This cluster is for people who live intermittently in Bengaluru. One would expect neighborhoods in this cluster to have hotels as the centerpiece. Thus connecting the metro from these neighborhoods to the airport could be enormously beneficial. Cluster 6 would benefit from cheap transportation options as they are daily wage earners working in the flea market and in fishing. A greater number of buses and subsidized ticket prices would improve the standard of living immensely. Agricultural government is key in these areas. The map shows that neighborhoods grow out of Bengaluru through its highways. Petrol bunks would be necessary investments in such locations.

6. Conclusion

To govern Bengaluru effectively, the spread of neighborhoods must be understood. First, data on Bengaluru neighborhoods were cleaned and outliers were corrected. Second, using calls to foursquare, the most common venues surrounding a neighborhood were obtained. Third, using a k-means clustering approach, the neighborhoods were divided into clusters. This provided insight into how nearby neighborhoods may need different approaches. This could help the government allocate funding in the most impactful manner. Furthermore, this kind of data could help in determining where the metro in the city should be connected to the airport. Although intuitive thinking may suggest to put it right at the High Court, the data would suggest to put it nearer to Domlur for smoother travel. This kind of data- driven approach can aid in making accurate investment to neighborhoods rather than a one size fits all approach. Governing a city like Bengaluru is a difficult task and data science may prove useful in a successful outcome.