Battle of Neighborhoods – Bangalore City

Applied Data Science Capstone by IBM on Coursera

Author:

Saideep R Naik

Introduction:

This project deals with the major venue categories in the neighborhoods of Bangalore,India. This project would specifically help Business personal plan to start new Restaurants, Hotels, etc in Bangalore, Karnataka, India. The major Target Audience would be small scale business owners and stake holders planning to start their business at a location in Bangalore This project would help them find the optimal location based on the category of their business such as

1. What is the best location to start a new hotel in Bangalore with restaurants around?
2. Which area is best suitable for opening a Shopping Mall in Bangalore?

- The **Foursquare API** is used to access the venues in the neighborhoods. Since,it returns less venues in the neighborhoods, we would be analyzing areas for which countable number of venues are obtained.
- •Then they are clustered based on their venues using Data Science Techniques. Here the **k-means clustering algorithm** is used to achieve the task. The optimal number of clusters can be obtained using **silhouette score** metrics.
- **•Folium visualization library** can be used to visualize the clusters superimposed on the map of Bangalore city. These clusters can be analyzed to help small scale business owners select a suitable location for their need such as Hotels, Shopping Malls, Restaurants or even specifically Indian restaurants or Coffee shops.

Data:

- Bangalore has multiple neighborhoods. The Kaggle website has a dataset which has the list of locations in Bangalore along with their Latitude and Longitude in degree format. There is a total of 352 neighborhoods. We use following resources,

  1. https://www.kaggle.com/rmenon1998/bangalore-neighborhoods

  2. http://foursquare.com/

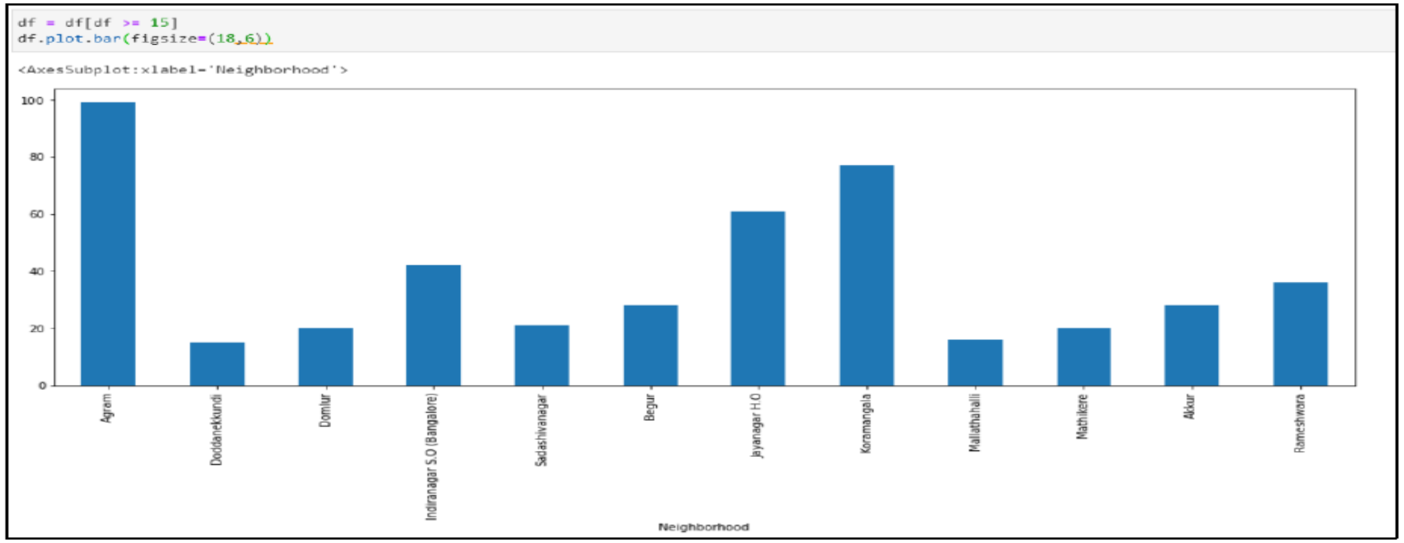
- A total of 776 venues data have been obtained from Foursquare. The resultant venues dataset is used for the analysis process.

Methodology:

- Now, we have the neighborhoods data of Bangalore 352 neighborhoods. We also have the most popular venues in each neighborhood obtained using Foursquare API A total of 776 venues have been obtained in the whole city and 179 unique categories But as seen we have multiple neighborhoods with less than 15 venues returned In order to create a good analysis let's consider only the neighborhoods with more than 15 venues.
- We can perform one hot encoding on the obtained data set and use it find the 10 most common venue category in each neighborhood Then clustering can be performed on the dataset Here K Nearest Neighbor clustering technique have been used To find the optimal number of clusters silhouette score metric technique is used.
- The clusters obtained can be analyzed to find the major type of venue categories in each cluster This data can be used to suggest businesses, suitable locations based on the category.

Analysis:

- Looking into the dataset we found that there were many neighborhoods with less than 15 venues which can be remove before performing the analysis to obtain better results The following plot shows only the neighborhoods from which 15 or more than 15 venues were obtained The resultant dataset consists of 12 neighborhoods.



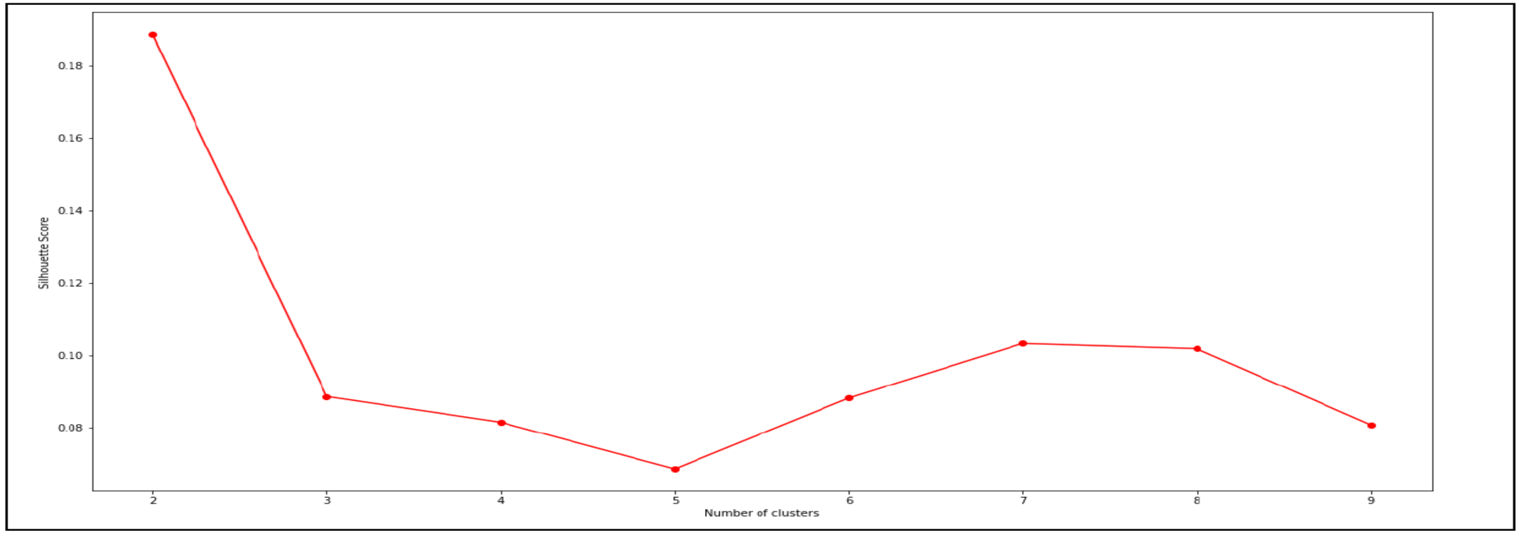
```
df = df[df >= 15]
df.plot.bar(figsize=(18,6))
```

```
<AxesSubplot:xlabel='Neighborhood'>
```

Figure representing Neighborhoods with more than 15 venues.

- One hot encoding is performed on the filtered data to obtain the venue categories in each neighborhood. Then group the data by neighborhood and take the mean value of the frequency of occurrence of each category.
- This is used to obtain the top 10 most common venues in each neighborhood i.e the 10 venues with the highest mean of frequency of occurrence.
- The resultant dataset can be used for the clustering algorithm. Here, the K Nearest Neighbor (clustering algorithm) is used. It is an unsupervised machine learning technique that clusters the given data into K number of clusters.
- For optimal result we need to select the best value for K. Here, the silhouette score is used to find the best value for K.

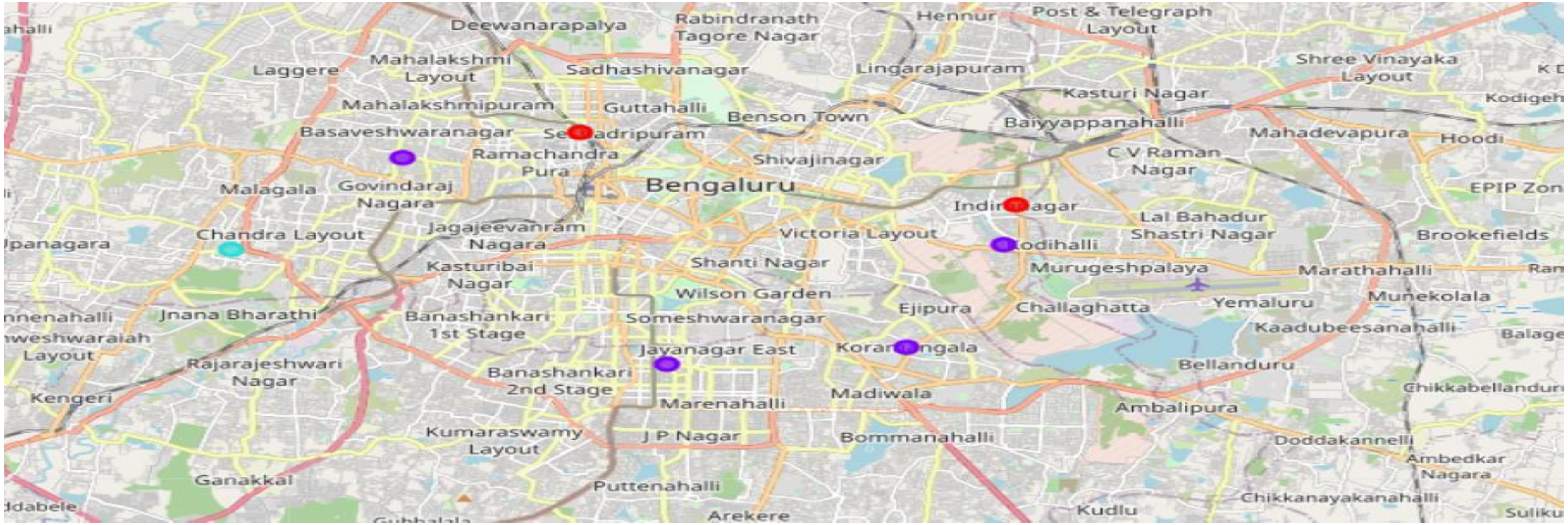
- A range of values from 2 to 10 was considered, KNN clustering was performed on the dataset and the silhouette score was calculated and plotted on a line plot. From the plot we can see that a K value of 7 provides the best score. This K value is used for the K Means Clustering Technique. The K Means labels obtained were included in the top neighborhoods dataset for examining the characteristics of each cluster.



Silhouette Score for different Number of Clusters

Results:

- From the below image we can see that the result shows the suitable areas for Restaurants and Hotels in the city of Bangalore.



Conclusion:

- Purpose of this project was to analyze the neighborhoods of Bangalore and create a clustering model to suggest places to start a new business based on the category.
- The neighborhoods data was obtained from an online source and the Foursquare API was used to find the major venues in each neighborhood. The best number of clusters i.e 7 was obtained using the silhouette score Each cluster was examined to find the most venue categories present, that defines the characteristics for that particular cluster.
- A few examples for the applications that the clusters can be used for have also been discussed A map showing the clusters have been provided Both these can be used by stakeholders to decide the location for the particular type of business.

Thank you!