Capstone Project - The Battle of the Neighborhoods

Exploring neighborhoods in Pune, Maharashtra

By Nandhitha .V

I. Introduction

Pune also called Poona, is the second largest city in the Indian state of Maharashtra, after Mumbai. It is the ninth most populous city in the country with an estimated population of 6.4 million. Pune is ranked the number one city in India in the ease of living ranking index. But, for someone who is new to this state, it's puzzling to figure a neighborhood to settle.

I have recently got transferred to Pune for work. Which neighborhoods had got vegetarian restaurants? Which neighborhoods have shopping malls and theatres? Which neighborhoods are great for a coffee? Which neighborhoods are famous for its markets? Where are the ATMs? were the questions running on my mind as a new resident. So, in this capstone project, i am going to find a good neighborhood location in Pune to stay in by using data science methods and algorithms like clustering.

II. Data

The required data to do this project is as follows:

- Latitude and Longitude of neighborhood
- List of neighborhoods in Pune, Maharashtra
- Venue data(restaurants, ATM, theatres) of neighborhoods

Data sources

Following data sources will be needed to extract the required information:

- Web scraping data from Wikipedia to get information on Pune neighborhoods using Beautiful soup.

The link to the data is : 'https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Pune'

- Generating Longitude and Longitude coordinates of Pune as well the neighborhoods via geocoder , geopy package.

- Using Foursquare API to create a map of Pune and also to generate the venue data related to the neighborhood.

III. Methodology

The project is done following the seven Data Science methodology phases. The methodology will include:

- Firstly, packages such as numpy, pandas ,folium,kmeans,geopy etc which are required to do the project are imported.

- Generating the data required from Wikipedia 'https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Pune' by web scraping using Beautiful soup. Used functions like decompose for data cleaning which removes the unnecessary data from the website.

- The data collected includes the names of the neighborhoods in Pune.

- The geocodes (latitude and longitude) of the neighborhoods are generated using the Nominatim from geopy.geocoders package.

- A data-driven map of Pune is generated using folium package by adding location markers to it.

- The venues for each neighborhood is collected using the foursquare API by providing the credentials, latitude, longitude, version and url.

- A search query is written to find and explore the vegetarian restaurants in the neighborhood.

- The get_dummies method of pandas library is then used and the categories of venues is got.

- Then group_by is used to Group rows of data frame by location and the mean of the frequency of occurrence of each category in every location is found.

- Top five categories of each neighborhood is printed with the frequency.

- A function called return_most_common_venues is used to return ten most common venues in each neighborhood for easy-understanding.

- K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups called clusters. So, here k-means clustering is performed on the most common places of the neighborhood with k=2.Labels are assigned for each cluster to differentiate.

- Map of Pune before and after clustering is compared and visualized.
- Clusters are examined to derive cool insights from it.