Coursera Capstone Project : Applied Data Science

MANDADI.KARTHIK

OVERVIEW

- **Introduction**
- **Business Problem**
- **Data**
    1. Neighbourhoods
    2. Geocoding
    3. Venue Data
- **Methodology**
    1. Accuracy of the Geocoding API
    2. Folium
    3. Top 10 most common venues
    4. Optimal number of clusters
    5. K-means clustering
- **Results**
    1. Discussion
    2. Conclusion

INTRODUCTON

- Hyderabad, India, differentiates itself with **low living costs**. According to our city rankings, this is a good place to live with high ratings in **housing**, **startups** and **safety**.
- **Hyderabad** is the capital and largest city of the Indian state of Telangana and *de jure* capital of Andhra Pradesh.Hyderabad City has a population of about 6.9 million, with about 9.7 million in Hyderabad Metropolitan Region, making it the fourth-most populous city and sixth-most populous urban agglomeration in India. With an output of US$74 billion, Hyderabad is the fifth-largest contributor to India's overall gross domestic product.

BUSINESS PROBLEM

- XYZ is a company based in Vijayawada for organizing events, currently it works on a project to organize an event for 10 days for a group of studentss from all over the India.The company has to put a good program, including a hotel of residence, a hall for meetings, places of landscape to visit, stores for shopping, restaurants and cafes. So the company’s purpose is to make a list of places of landscape in Hyderabad, including the nearest restaurants, cafes, and shopping stores for each place

DATA

Neighbourhoods

The data of the neighbourhoods in Hyderabad can be extracted out by web scraping using BeautifulSoup library for Python. The neighbourhood data is scraped from a Wikipedia webpage.

Geocoding

The file contents from Hyderabad.csv is retrieved into a Pandas DataFrame. The latitude and longitude of the neighbourhoods are retrieved using Google Maps Geocoding API. The geometric location values are then stored into the intial dataframe

DATA

- Venue Data

  From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the FourSquare API, and creating another DataFrame to contain all the venue details along with the respective neighbourhoods.

Methodology

Accuracy of the Geocoding API

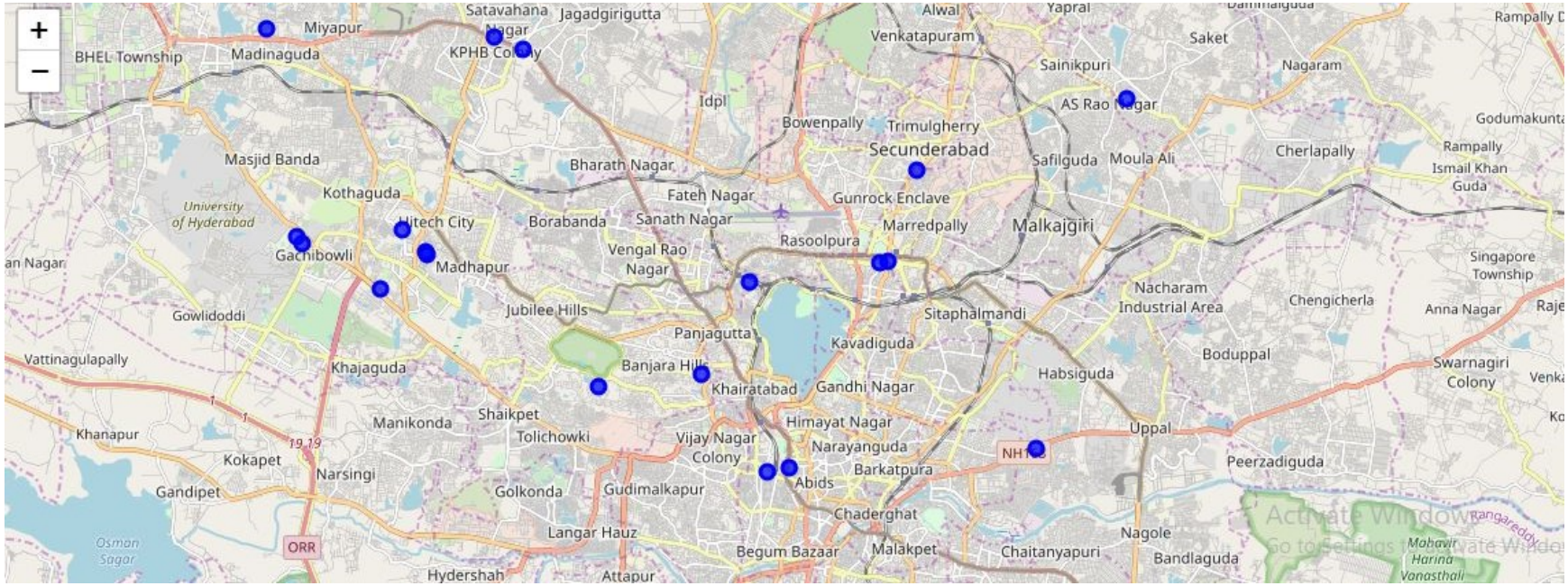
In the initial development phase with OpenCage Geocoder API, the number of erroneous results were of an appreciable amount, which led to the development of an algorithm to analyze the accuracy of the Geocoding API used.

In the algorithm developed, Geocoding API from various providers were tested, and in the end, Google Maps Geocoder API turned out to have the least number of collisions (errors) in our analysis.

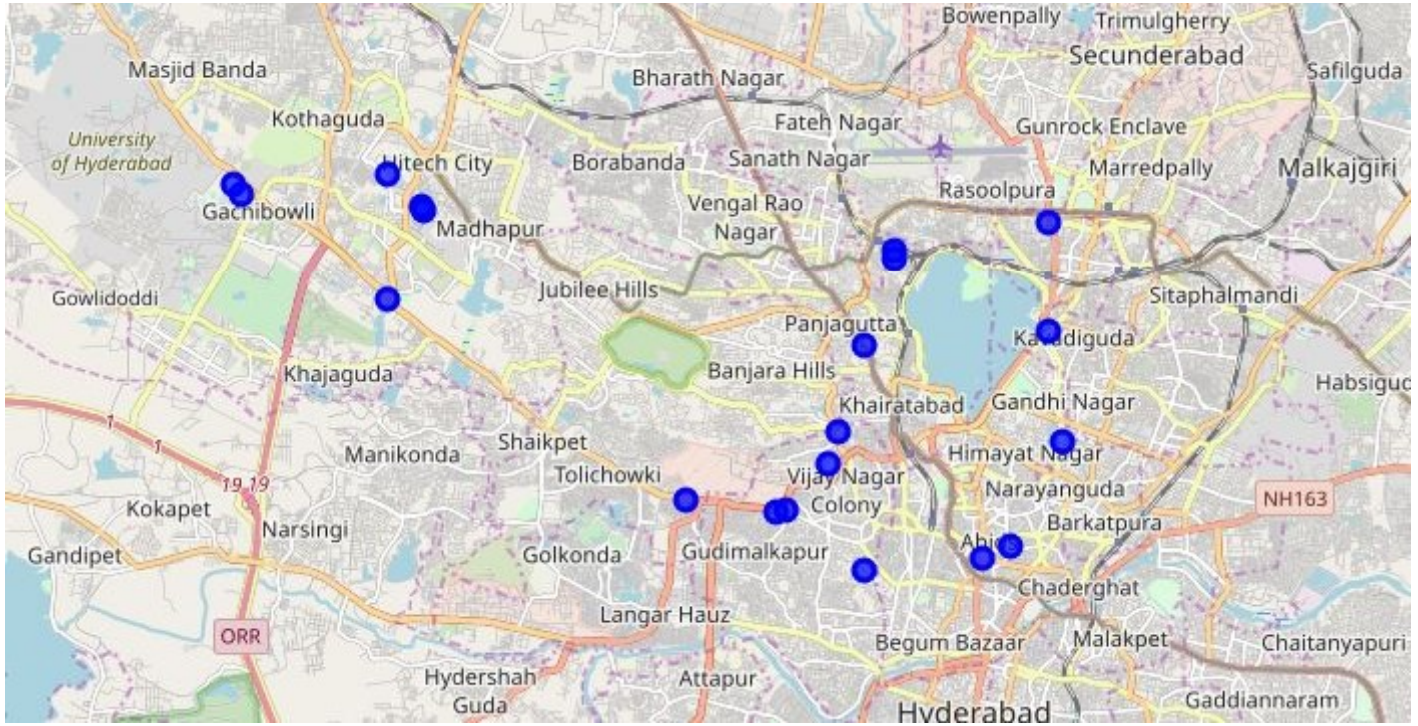
Folium

Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the leaflet.js library. All cluster visualization are done with help of Folium which in turn generates a Leaflet map made using OpenStreetMap technology.

map to visualize hotel neighbourhood including shopping stores and Cafeteria



map to visualize park neighbourhood including Restaurant and Cafeteria



Conclusion

- The data used in this project is provided by Foursquare location data. The data are grouped by landscape area, and each area included the information about this area and all information about restaurants, cafes, and stores which in this area.
- The company has made a list of places of landscape in Hyderabad, including the nearest restaurants, cafes, and shopping stores for each place. And it has to provide a map in case the Students can’t use his/here mobile application for any reason.