Introduction

Bakeries are a popular type of food service establishment. The smell of freshly baked goods and fantastic coffee is what I call heaven.

My friend Mia loves to bake and just finished her baking school. Opening a bakery presents many unique challenges, soI wants to help her decide on a location where she can open a bakery with a low risk of competition.

Business problem

This project aims to analyze and select the best locations in Pune, India, to open a new bakery. This project mainly focuses on the geospatial analysis of Pune City to understand which would be the best place to open a new bakery. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Pune, if a person is looking to open a new bakery, where would you recommend that they open it?

Data

To solve the problem, we will need the following data:

• List of neighbourhoods in Pune: It defines this project's scope, which is confined to the city of Pune.

• Latitude and longitude coordinates of those neighbourhoods. This is required to plot the map and also to get the venue data

• Venue data, particularly data related to bakeries. We will use this data to perform clustering on the neighbourhoods.

Sources of Data and methods to extract the Data

This Wikipedia page is a list of neighbourhoods in Pune, with 200 neighbourhoods. I have used web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we can get the latitude and longitude coordinates of the neighbourhoods using Python Geocoder package. After that, I have used the Foursquare API to get the venue data for those neighbourhoods.

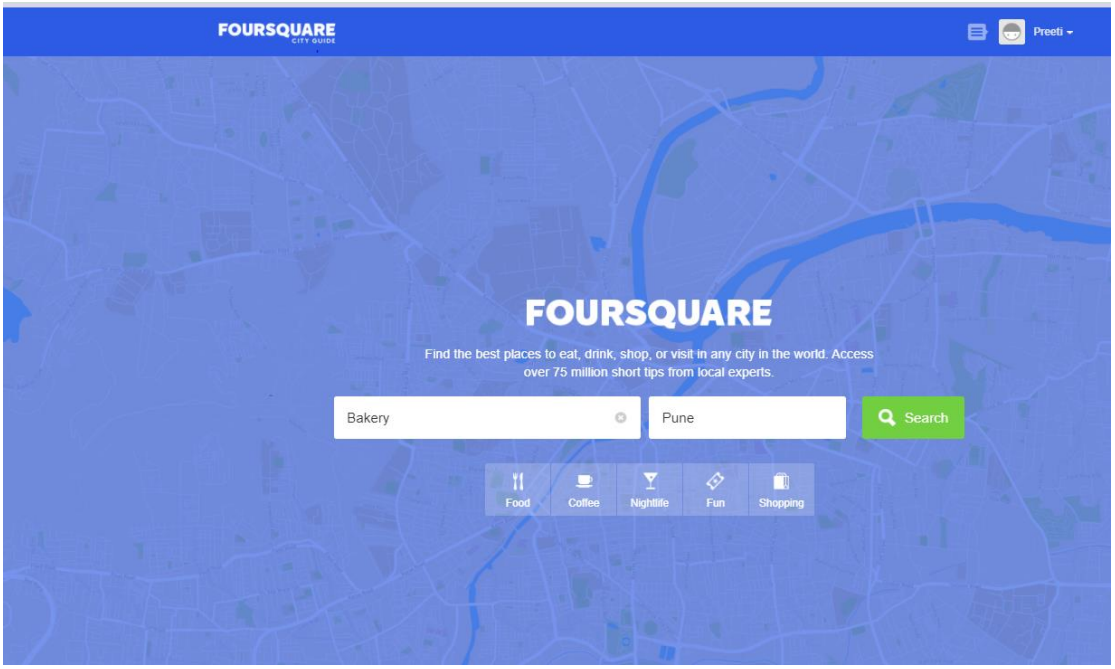
Foursquare API will provide many categories of the venue data, and we are particularly interested in the Bakery category in order to help us solve the business problem. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

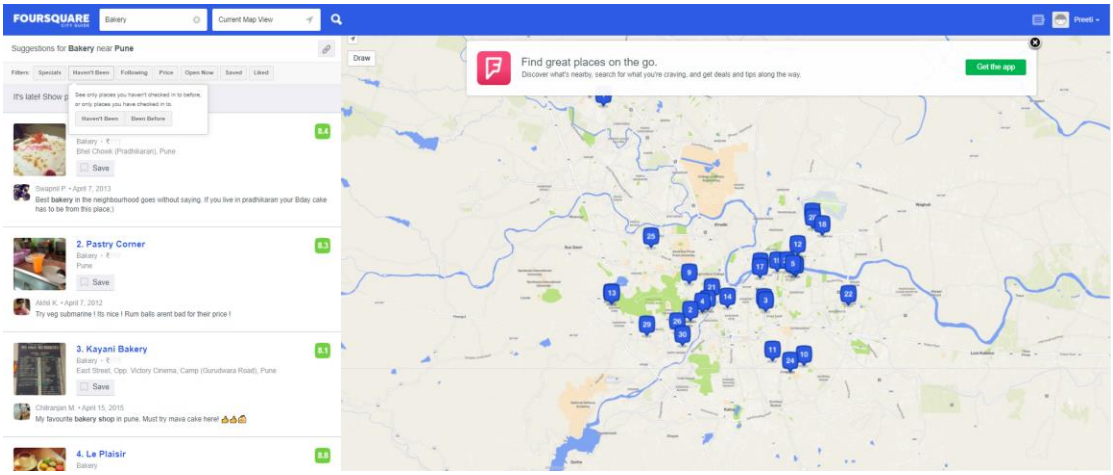
The Foursquare API allows application developers to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which you can send requests, so there’s really nothing to download

onto your server.

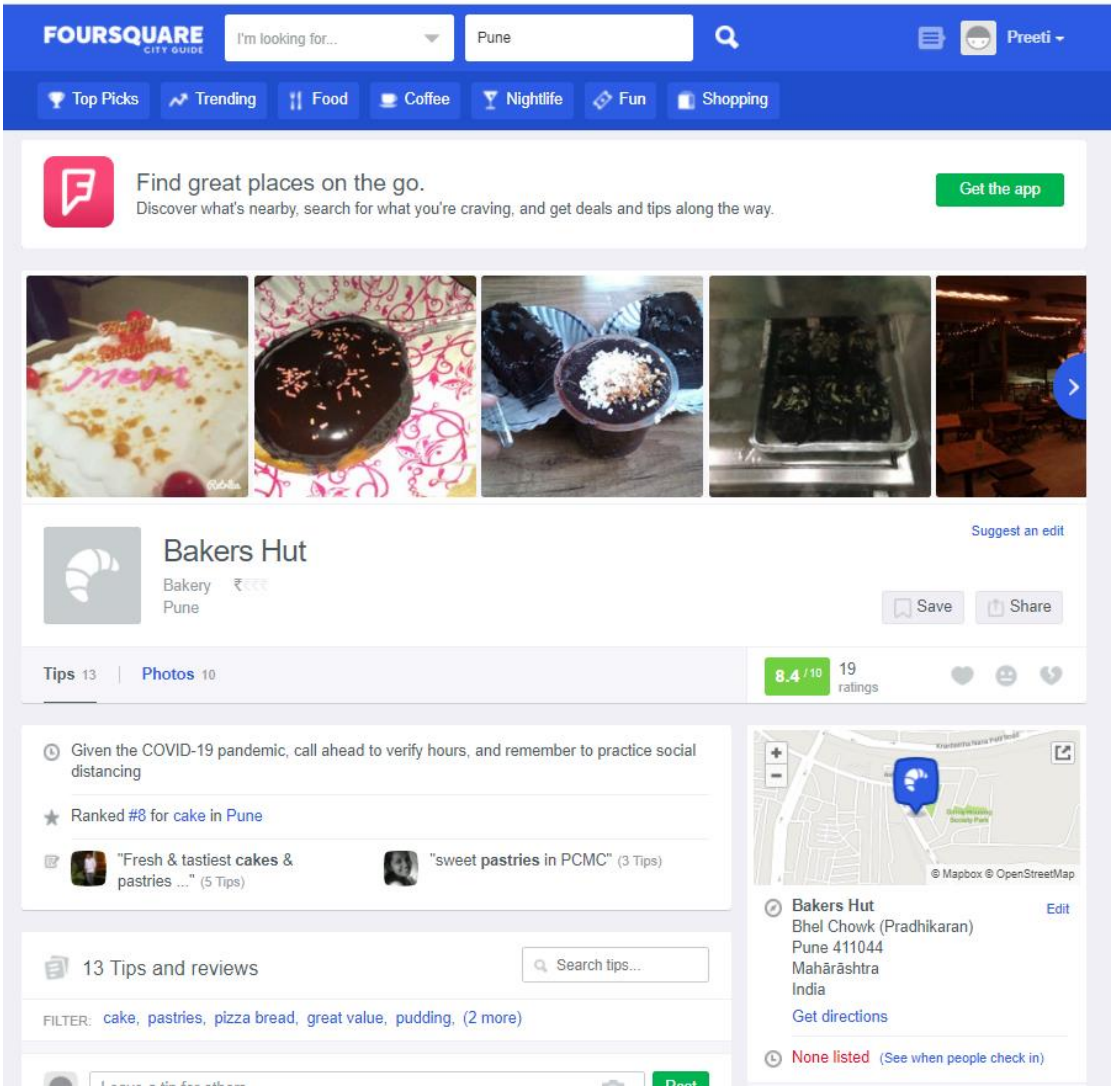
Search for Bakeries in Pune



Bakeries returned by Foursquare



Exploring venue:



If you click on the first one which is Bakers Hut, then you are redirected to this page where you see all the information in the Foursquare dataset about Bakers Hut. This includes its name, full address, working hours, tips and images that users have posted about the cafe. So similarly you can explore the bakeries in Pune city.

Web Scraping

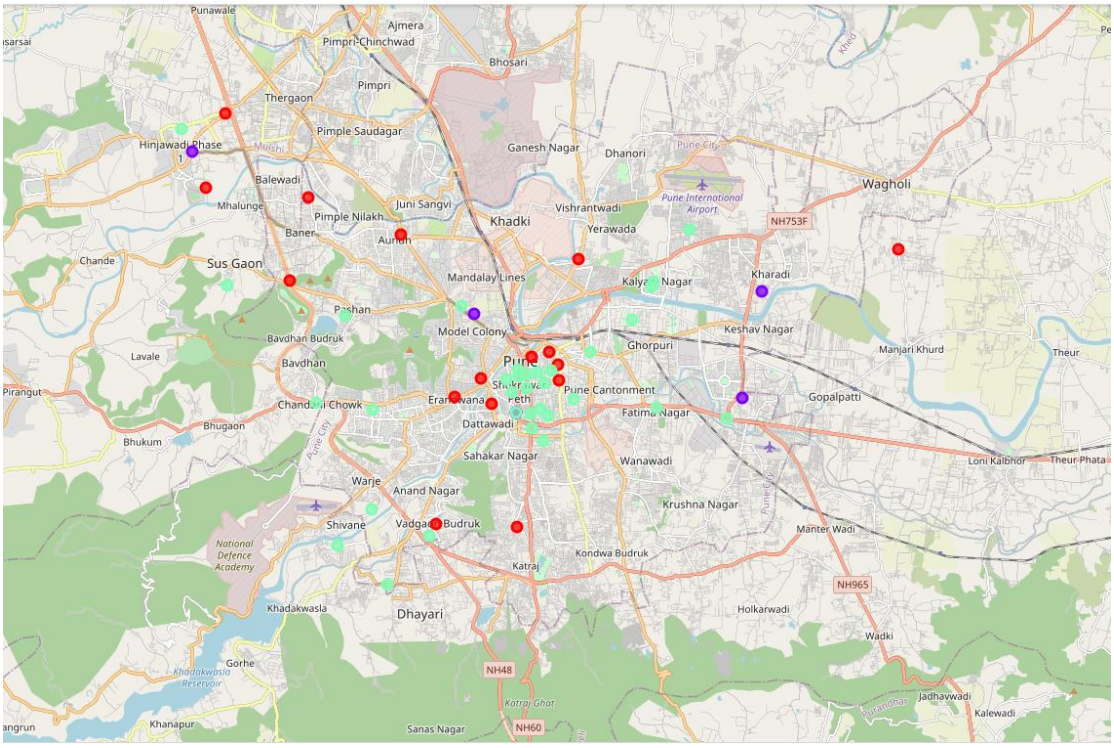
Web scraping is data scraping used for extracting data from websites. The web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. We performed scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data.

Results

The results from the K-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "Bakery":

- Cluster 0: Neighbourhoods with no Bakeries
- Cluster 1: Neighbourhoods with a high number of Bakery
- Cluster 2: Neighbourhoods with a moderate number of Bakery

We visualize the results of the clustering in the map with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Conclusion

Purpose of this project was to identify Pune areas with low number of bakeries in order to find a location for Mia to open a bakery by narrowing down the search for optimal location for a new bakery. By calculating bakery density distribution from Foursquare data we have first identified that there are 40 bakeries in the 56 Neighborhood. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decission on optimal bakery location will be made by the soon to be owner(Mia) based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration the number of bakeries already present in the neighborhood