1. Introduction

1.1. Background

Los Angeles is well known for having diverse types of food. Often, different cuisine concentrated in different neighborhoods. For example, Sawtelle is known for having many Japanese restaurants, Koreatown has a high concentration of Korean restaurants, and North Hollywood often has lots of food trucks.

1.2. Problem

With so many different neighborhoods in Los Angeles, foodies need to know which neighborhoods are enriched in what kinds of cuisine and which neighborhoods are similar in their food options.

1.3 Target

Business looking to open new food venues would be interested in the kind of cuisine represented in the neighborhood. Foodies living near LA or visiting LA would also be interested in this information.

1.4 Approach

I am going to use FourSquare API to obtain information about food venues in the different neighborhoods in the downtown Los Angeles area. In order to do that, I need the neighborhood names, longitude, and latitude of the neighborhood. I will present the different

2. Data Acquisition and Cleaning

2.1 Data Requirement

For each of the neighborhood in the downtown area, I need the zip code, neighborhood name, longitude, and latitude in order to use the FourSquare API to obtain a list of food venues in that neighborhood. The main information I need from the list of food venues is the type of cuisine the venue belongs to.

2.2 Data Source

The zip code and name of the neighborhood in Los Angeles County can be obtained through LA almanac. The longitude and latitude of all areas in California can be obtained through OpenDataSoft.

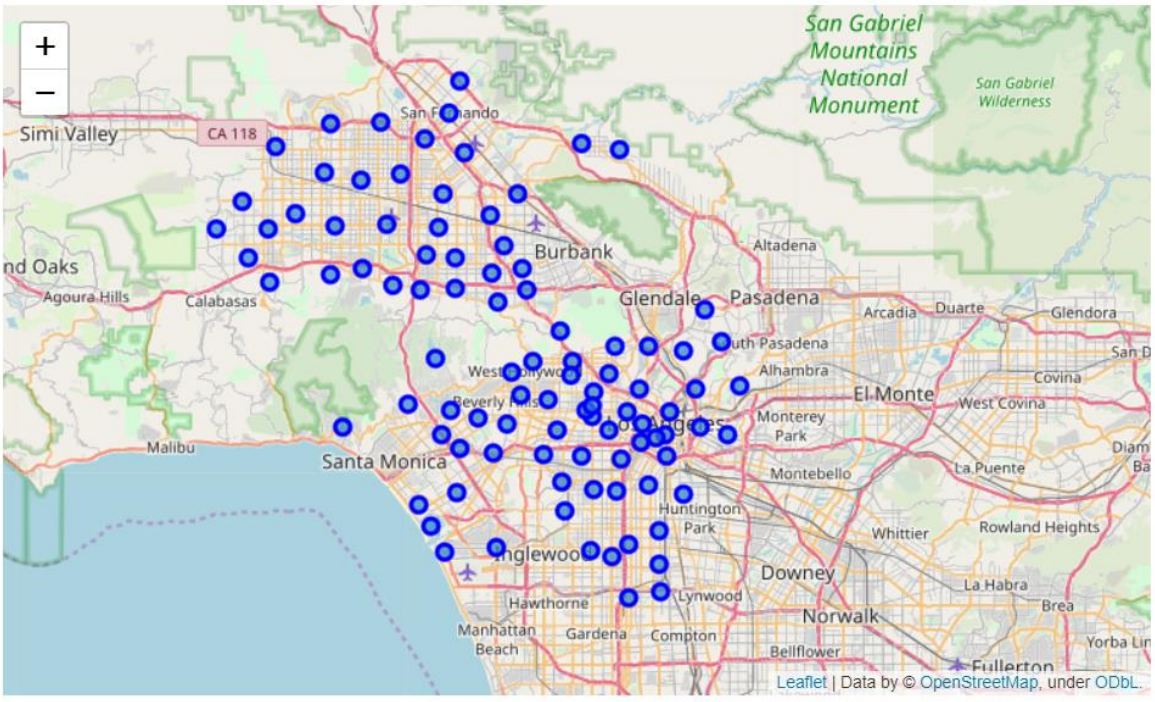
2.3 Data Cleaning

Data obtained from LA almanac is converted to a pandas dataframe with zipcode, name of neighborhood, and the median income. The median income column was dropped. The neighborhood column has information about whether the neighborhood is in downtown Los Angeles or elsewhere in the county. To focus only on downtown Los Angeles, all rows with neighborhoods outside of the Los Angeles downtown were dropped.

The OpenDataSoft data is converted into a pandas dataframe. We combined it with the neighborhood and zipcode dataframe by using inner join on zip code. The resulting dataframe has zipcode, neighborhood, city, latitude and the longitude.

	Zip Code	Neighborhood	City	Latitude	Longitude
0	90001	South Los Angeles	Los Angeles	33.972914	-118.24878
1	90002	Southeast Los Angeles, Watts	Los Angeles	33.948315	-118.24845
2	90003	South Los Angeles, Southeast Los Angeles	Los Angeles	33.962714	-118.27600
3	90004	Hancock Park, Rampart Village, Virgil Village,...	Los Angeles	34.077110	-118.30755
4	90005	Hancock Park, Koreatown, Wilshire Center, Wils...	Los Angeles	34.058911	-118.30848

From the table, the neighborhood name, latitude, and longitude are used to make markers on Folium map.



These information is passed through FourSquare API selecting only for food venues, with a limit of 100 per neighborhood and a radius of 500m. A list of up to 100 food venues in each neighborhood is received, containing information about the category of food each belongs to.

There are many categories that are very specific and can be classified into a broader category of cuisine, as the “General Category” column. For example, donburi restaurant, donburi bowl, ramen restaurant, and sushi restaurant all could be classify as Japanese restaurant. Additionally, fast food and similar ubiquitous stores such as donut shops were excluded from the dataset since they do not represent any cuisine.

Neighborhoods with less than 5 venues were dropped from the dataset, and general categories with less than 3 venues were also dropped. This resulted in 76 neighborhoods and 29 general categories of cuisine. The general category was one-hot encoded. Using groupby function, the percentage of each cuisine in the neighborhood was calculated and used as features in the K-mean clustering.