2. Data acquisition and cleaning

2.1 Data sources

A list of cities within Los Angeles County can be found here on Wikipedia. This dataset does not provide geographical coordinates however, so that will be attained using Python's geocoder attribute.

The venue data used to locate existing cafés and visualize these competitors will be attained from Foursquare's database.

2.2 Data cleaning

The list of cities in Los Angeles County will be obtained from its Wikipedia site using the beautifulsoup package for Python. Once scraped, the data will then be merged with location information provided by Python's geocoder package.

Once we have the cities' geographical locations, we can then call on Foursquare's API to obtain café venues in the listed cities. From Foursquare's API, we will retrieve the following for each venue:

- **Name:** The name of the venue
- **Category:** The category type defined by the API
- **Latitude:** The latitude value of the venue
- **Longitude:** The longitude value of the venue

Once merged, we can then visualize and cluster the data to help prospective stakeholders decide which cities would be areas of interest for opening a café.