**Coursera Capstone Final Project**

**Data Requirements**

* List of cities in Los Angeles county and their geo coordinates
* Cities demographics including per capita income, populate and crime rate
* Venue data such as restaurants types, ratings, etc.

**Data Sources:**

* List of cities:

[List of cities in Los Angeles County, California - Wikipedia](https://en.wikipedia.org/wiki/List_of_cities_in_Los_Angeles_County,_California)

* Cities demographics: (Population and income)

<https://en.wikipedia.org/wiki/List_of_California_locations_by_income>

* Crime rate by city:

<https://en.wikipedia.org/wiki/California_locations_by_crime_rate>

* Venues data: Foursquare
* Geo coordinates: Python Geocoders library

**Methodology**

***Step 1:*** Data will be acquired from Wikipedia as HTML tables and saved into CSV files for further processing. This will be done using Pandas Read function.

***Step 2:*** The acquired datasets (data frames) consisting of Cities demographic data will be consolidated into one single data frame. The resulting consolidated data frame will contain city name, population, per capita income, and violent crime rate.

***Step 3:*** All rows with county other than Los Angeles will be removed from the data frame.

***Step 4:*** Add geospatial data to the data frame using the Nominatim function in the Conda-forge Geopy package.

***Step 5:*** Using the Folium package, create a map of cities in LA County.

***Step 6:*** Using Foursquare API, retrieve data for all restaurants within 1 mile of each city. This will include the restaurant categories, such as Chinese, Indian, Middle Easter, Mediterranean, etc.

***Step 7:*** Transform the data frame with restaurant venues/categories to one-hot encoding (0/1).

***Step 8:*** Use grouping to show the frequency of each restaurant category in each city.

***Step 9:*** Using the info from step 8 above, create a data frame with most common restaurant categories for each city.

***Step 10:*** Use silhouette score to identify the optimal number of parameter for the number of clusters.

**Step 11:** Using the silhouette score for number of clusters, run a ***k-means clustering algorithm*** (unsupervised learning) from the Scikit-learn package.

***Final Step:*** Review the clusters to pick the one with most venues and make recommendation to client.