Introduction / Business Problem

Imagine you work for a large corporation with offices in cities across the United States and are suddenly asked to relocate. It could be for a number of reasons: to help found a new office in a different city, to re-allocate human resources for efficiency purposes, or another... Your supervisor provides you with a list of cities you can choose from to move to. Or imagine you have lived and worked in a city for all or the majority of your life but have decided that its time to make a change to a new city. In both of these scenarios, what if you want a new city that offers many of the same attractions, sites, foods, etc. that you are familiar with in your city now. You want this new city to be different but at the same time have a similar scene in terms of venues and offerings.

How do you identify cities that are similar to the one you live in now?

This project will venture to answer this question via clustering of various cities in the United States by similarity in terms of venues present in each city.

Data

To complete this project, the following data for every city in the United States is necessary: city name, state name, longitude, latitude, population, and land area. This data is available from Kaggle.com in two separate datasets in csv format: "World City Database"—which contains latitude, longitude, and population data for every city in the world—and "US City Population Densities"—which contains population density, population, and land area data for every city in the United States. Merging these two datasets on US cities yields the necessary dataset of information.

An example entry in said merged dataset is as follows:

City	State	Area (sq mi)	City Radius (km)	lat	long	population
New York	New York	303	15.80	40.6943	-73.9249	19354922
Chicago	Illinois	228	13.71	41.8373	-87.6862	8675982

Next, venue data for each city is necessary. Using Foursquare API in combination with the latitudes, longitudes, and radius information of every city in the above-described dataset, each city is explored to yield its top 100 venues and their associated category in json format. This information will be cleaned and structured into a *pandas* dataframe for easier processing and manipulation.

An example entry in the resulting dataframe of city and venue data is as follows:

City	City Lat	City Long	Venue	Venue Lat	Venue Long	Venue Category
New York	40.6943	-73.9249	Carmenta's	40.70132	-73.92678	Italian Restaurant
New York	40.6943	-73.9249	Henry's Wine and Spirit	40.70105	-73.93025	Wine Shop