## INTRO

### Background

Imagine an individual is working for a large corporation with offices in cities across the United States and is 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... Their supervisor provides them with a list of cities to choose from to move to.

Or, similarly, imagine another individual has lived and worked in a city for all or the majority of their life but have decided that its time to make a change to a different city.

In both of these scenarios, what if the individual desired a new city that offers many of the same attractions, sites, foods, etc. that they are familiar with in your city now. They seek a new city but the same time desire a similar scene in terms of venues and offerings. They want different but not too different.

### Problem

How do you identify cities that are characteristically 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 top venues present in each city.

### Interest

As asserted in the background of this project, stakeholders of interest in this project are those who are trying to determine which city to move to. More broadly, stakeholders are those who are interested in evaluating the similarity of different cities across the United States.

## 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, their respective locations, 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 |