Introduction/Business Problem

According to U.S. census data, since 2010 roughly 35.5 million Americans on average moved each year from one address to another. If that move includes moving from one city to another city, there are thousands of cities to choose from in the U.S. Determining which city to live is a complex decision. There are many factors that can contribute to the decision of which city to move, a few among them are cost of living, employment opportunities, education levels, available activities, and city size. Let's assume you are living in the U.S. and would like to move to a city which has a population of 100,000 people or greater. There are over 300 cities in the U.S. which have a population greater than 100,000 people. How would you begin to decide which city to choose from? We can use machine learning to cluster these 300 cities into different groups, each group with a different set of attributes. This will help you narrow down the search for a potential city to choose.

Data

We will be using different sets of data in our model of U.S. cities. The first source of data is from a Wikipedia page which contains a list of the largest U.S. cities by population for cities with 100,000 people or more. This list contains 314 cities based on 2018 population data.

The next source of data is from the U.S. Census Bureau. The U.S. Census Bureau is a principal agency of the U.S. Federal Statistical System, responsible for producing data about the American people and economy. The census bureau provides city level data on the following categories: population characteristics, housing, education, income, and geography. One method to access this data is through a feature called QuickFacts, which provides statistics for all states and counties, and for cities and towns with a population of 5,000 or more. The user must search for a specific city to obtain the statistics for that city. Using the Python libraries requests and Beautiful Soup, we can search for each of the 300+ cities with a population of 100,000 people or more and get the statistics for each city. The following statistics were used for the project: population, population percent change, median gross rent, percent of population age 25+ with bachelor’s degree or higher, percent employed, per capita income, land area in square miles.

An important consideration for choosing a city is the number of activities available for its residents, for example the number of Arts and Entertainment venues in a city. To retrieve this data, we use the Foursquare API called Places. The Places API offers real-time access to Foursquare’s global database of rich venue data. Ideally, we would extract all venues from the Foursquare database to understand the number of venues available in a city, but we are only limited to 100 venues per API call. Extracting all venues from Foursquare would require searching each possible latitude and longitude in a specific city. As an approximation to extracting all venues, we use the Foursquare Explore endpoint to get the top 100 recommended venues for each city in our list. If Arts and Entertainment venues are more likely to be recommended in one city versus another, then we assume that the city with the higher recommendations has relatively more Arts and Entertainment venues then the other city. To classify these venues at a higher level, we use the Foursquare Categories endpoint, which gives us the mapping from venue category to high-level venue category: Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Resident, Shop & Service, Travel & Transport. For example, the venue category of Park maps to the high-level venue category of Outdoors & Recreation