**U.S. Cities – Where to Live**

**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, diversity, 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.

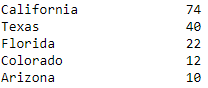
The next source of data is from the U.S. Census Bureau. The U.S. Census Bureau is a principal agency of the U.S. Federal Statistical System, responsible for producing data about the American people and economy. The census bureau provides city level data on the following categories: population characteristics, 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, diversity, 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

**Methodology**

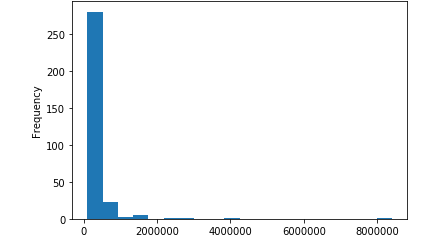
**Exploratory Data Analysis**

We perform exploratory data analysis on our list of cities with population greater than 100k. The states with the most cities with population greater than 100k are:

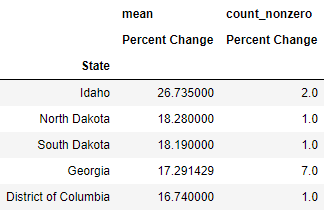


California, Texas, and Arizona also have some of the top 10 populated cities.

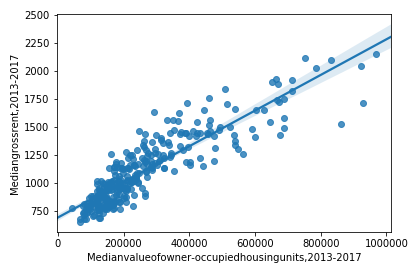
We see that the mean of the population is greater than the median of the population, which is as expected because we have less cities with high population (New York, Los Angeles, Chicago, etc.). Our distribution of populations are skewed to the right, with a long tail of high population cities pulling the mean up.



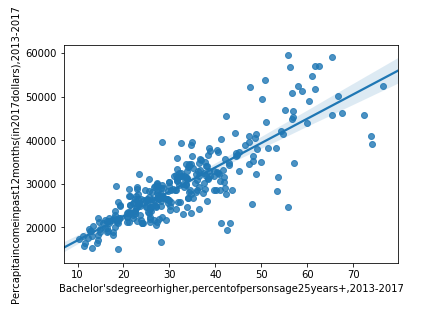
We examine if there is a relationship between population size and population change, but find that there is no relationship. We see that cities with 100k+ population in Idaho are experiencing the greatest population growth, followed by North Dakota, South Dakota, Georgia, and Washington D.C. It’s important to note that most of these states have a lower count of cities with 100k+ population.



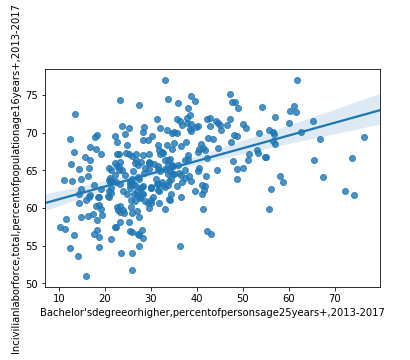
We then perform data analysis on our census bureau data. Reviewing the correlations between all statistics with population, we find only strong correlations between population related statistics (i.e. 2010 census population, number of households, number of firms, and total retail sales). We compute the number of firms per person and number of retail sales per person and test whether the strong positive correlation holds. As expected, we find that the relationship no longer holds. We then go on to explore other relationships in the data. We find that there is a strong correlation between housing values and rent (r = 0.890). We will therefore use only one of these measures, gross rent, in our model of city clusters.



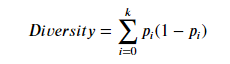
We check the correlation between college education and per capita income. We see that the correlation is strongly positive with r = 0.845. But in the graph below, we see some fanning out of the data as percent educated increases, so we will use both measures in our model of city clusters.



We then check the correlation between college education and the employment levels. We see that the correlation is weakly positive with r = 0.454. We’ll include employment levels in our model in order to capture the variation.

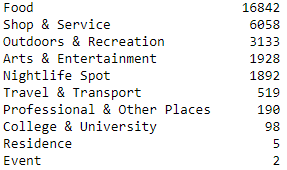


We are also interested in capturing the level of diversity in each city. In order to measure diversity, we will utilize Simpson's diversity index. For each city, we calculate the sum of the frequency of each population group Pi multiplied by the rareness of the group 1 - Pi.



This measure has the following appealing interpretation: pick two members of the population at random, then the diversity represents the probability that the two individuals are from different groups. After calculating the diversity index for each our cities, we looked at the top ten cities for diversity, finding that 7 of the top ten are in California. We see represented our most populous city of New York City in 8th. The most diverse city according to Simpson’s diversity index is Vallejo, California. The average diversity index for the US cities with population greater than 100k is .579. The states with the least diverse cities of 100k+ are Idaho, Montana, and North Dakota. We will include diversity in our model as this can be a consideration for people who are looking for a new city to move.

Next, we analyze the high-level venue category data per city. For the venues provided for our cities, the most common recommended venue was Food, followed by Shop & Service and Outdoors & Recreation:



Our assumption is that the cities with more Food and Shop&Service recommendations will have a smaller number of activities available for its residents. If Food and Shop & Service venues are more likely to be recommended, we assume that there are less options available for activity-based venues (Outdoors & Recreation, Arts & Entertainment, Nightlife Spot).

Next we transform our venue category data by doing one hot encoding to calculate the relative frequency of each of the high level categories per city and combine our census bureau data with our high-level venue category data. We review the top 10 cities with highest relative frequency of Outdoors & Recreation venues. With the top ten, we see major cities for tourism: New York City, San Francisco, Los, Angeles, Boston, Honolulu. We review the top 10 cities with highest relative frequency of Arts & Entertainment venues. It’s interesting to note that the top 2 cities are Anaheim (CA) Orlando (FL), as they have Disneyland and Disneyworld, respectively.

We next examine the correlations between the venue frequencies and population size. Of all the venue categories, we see the strongest correlations between population and Outdoors & Recreation (r = .551), Food (r = -.504), and Arts & Entertainment (r = .407). The negative correlation between population size and food requires more comments: the lower the population, the higher the percentage of food venues recommended, suggesting that smaller cities have less activity-based venues than larger cities.

**Machine Learning**

We decided to use k-means clustering machine learning algorithm to divide our cities into different groups based on the features included in the model. We selected k-means, as it is a model that is considered one of the simplest models for unsupervised learning, since our data is unlabeled (i.e. without defined groups). We use this model to quickly discover insights from our data, as the model we develop is meant to be a starting point in the evaluation of which cities a stakeholder would like to start their search.

**Results**

We decided to cluster the cities into 4 groups. We summarize the cluster types below:

Cluster 0 - Medium Population; Some Activities; High growth; High rent; High education; High income

Cluster 1 - Medium Population; Some Activities; Lower rent; Less diversity

Cluster 2 - Lower Population; Less Activities; Less educated; Higher rent with lower income

Cluster 3 - High Population; Lots of Activities; Greater diversity; reasonable rent; larger land area

The value counts of each cluster are represented below:



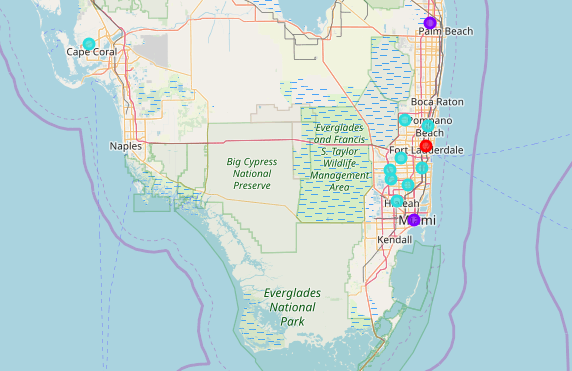
**Discussion**

In our analysis we were able to identify cities that were most likely the least attractive cities to move to. It appears that Cluster 2 cities are the least attractive, with less activities, less educated population, and higher rent with lower income. Cluster 2 cities had the highest value counts among the clusters. The states with the highest proportion of Cluster 2 cities are California, Virginia, and Florida, where California cities with 100k+ having 62% of these cities in Cluster 2. For California, we see the Cluster 2 cities in the following areas: East Bay (Northern California), Central California, and the greater Los Angeles Area (especially east of LA proper). For Florida, the concentration of Cluster 2 cities is in the South and for Virginia, it’s near the East Coast.

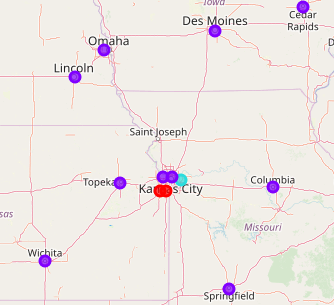
The next cluster that is likely more attractive than Cluster 2 cities is Cluster 1. Cluster 1 has a lower rent with similar income levels as Cluster 2 cities, and on average there are more activities to do in these cities. Cluster 1 cities are predominantly in the Midwest and Southern states, where rent is generally lower than the West and East coast cities. Some of the higher populated cities of Cluster 1 include Detroit, Baltimore, and Milwaukee.

The final two clusters of our model both have attractive features. We recommend that the stakeholder use the cities from these two clusters as a starting point in diving deeper into U.S. cities with population greater than 100k, to decide which city to move. We start first with discussing Cluster 0 cities. Cluster 0 is very attractive in that these cities are on average experiencing greater population growth, with higher per capita incomes and a more educated population. But with these, Cluster 0 is facing higher rents -> higher cost of living. It is interesting to note the states with Cluster 0 cities which are found in areas of the U.S. that have a high density of Cluster 2 and Cluster 1 cities:

1. Fort Lauderdale, Florida



1. Olathe and Overland Park, Kansas



Of the Cluster 0 cities, the highest rents of the country are happening in California, especially in Northern California in Silicon Valley. On the other hand, the highest per capita income paints a slightly different picture, where we see cities in Washington, Virginia, and Arizona, among others, experiencing higher incomes.

Cluster 3 cities are notable for their population size. Cluster 3 contains 9 of the top 10 most populous cities in the U.S. These large urban cities offer the most activities for their residents and have more diversity. The rent is more reasonable compared to Cluster 0 cities, but the per capita incomes are lower as well.

**Conclusion**

The purpose of this project was to group U.S. cities with population greater than 100,000 people into different clusters, with the intention of highlighting potential cities for stakeholders to move and live. We took into consideration some factors that contribute to the decision, including cost of living, employment opportunities, education levels, available activities, diversity, and city size. We were able to leverage different data sources to build our model, including U.S. Census Bureau data on city statistics and Foursquare venue data to approximate the number of activities available in each city. We explored the data through analysis of relationships between the different city level statistics, and even defined a custom feature for diversity using a common method for calculating city diversity found in the literature. With this data we were able to run a k-means clustering machine learning algorithm to cluster the cities into 4 different groups.

The final decision on best city to move to is a very personal and complex one, and stakeholders will end up considering many additional factors not modeled in this project. Other factors that could be considered are state tax landscape, crime rates, proximity to family, climate, education, healthcare, and industry specific employment opportunities, among others. We hope that the analysis performed thus far will provide stakeholders a place to start in exploring urban U.S. cities.