AN ANALYSIS OF UNITED STATES CITIES

Abstract

Using k-means clustering an analysis is used to determine different types of large cities in the United states. Venue category data is used as well as city size and grow rates to compare the largest 150 cities in the United States.

Nathan Jones

Introduction

1 Introduction

1.1 Background

> “America is so vast that almost everything said about it is likely to be true, and the opposite is probably equally true.”
>
> – James T. Farrell

The United States is a vast country spanning almost 3.8 million square miles with cities that are diverse in culture, lifestyle, and size. Even among large US cities there is variety among the cosmopolitan make up. The largest U.S. city, New York, New York boasts a population of about 84,000,000 people while Garden Grove, California, the smallest city in our analysis has a population of around 172,000. Determining similar cities based on culture, size, and change in growth is something that could benefit both individuals and businesses.

1.2 Problem

Factors such as the types of venues cities value can give us insight into the culture of that city. Using data about the top 300 venues in each city will help us establish distinguish cities that value a city that values coffee shops and bookstores over a city which prefers pizza and gastro pubs. Using this data such as this will help us categorize cities. In addition we can also look at the size of a city, population density, and growth rates to distinguish a quickly growing dense metropolis from smaller spread out cities.

1.3 Interest

An analysis such as this is beneficial to both individuals and businesses. An individual looking at moving to a different part of the country would be well served in finding a city that is similar to one that shares their personal preferences. A business looking to expand into new markets would benefit by understanding if a new city is similar to markets, they currently operate in hopes of repeating their success, or identifying new cities in which to diversify their customer base.

2. Data acquisition and cleaning

2.1 Data sources

The data identifying large US cities can be found on this Wikipedia page. Using web scraping methods, we can extract the data from the table found on the page. In addition foursquare provides and application programing interface (API) which we leverage to extract the venue data of 300 top venues for each city.

2.2 Cleaning the data

Using the Beautiful Soup package in Python we scraped the JSON data available from the Wikipedia page to create a DataFrame with the 150 largest US cities, their population, population density, change in size (from 2010 to 2018), and longitude and latitude values.

Several changes were then made to put the data set into a useable format. Using indicator symbols in the JSON data, information such as state capital status was stored with names. Using Regular Expressions, we removed these indicators. In the columns of population, population density, and change in size we cast these from strings into numeric values. Lastly we remove the extra data attached to the longitude and split it into two columns one for longitude and another for latitude containing the expected numeric values.

2.3 Obtaining Data from Foursquare

Using the Foursquare API we extracted the data venue category type from each venue returned. We did this for the top 250 venues (within 2000 meters of the city center) for all 150 cities and created a DataFrame. The city venues DataFrame contained the city name, city latitude, city longitude, venue name, venue latitude, venue longitude, and venue category. In the analysis we found 431 unique venue categories containing values such as BBQ Joint, Candy Store, Campground, and Gaming Café.

3. Methodology

3.1 Preparing the data for analysis

Using the city venue DataFrame we create a new data frame using the one hot encoding method. From this data frame we group the data by city and normalize the values. This gives us a DataFrame with a city column for each city and 431 columns for each venue type. The entry in each row contains a float with a value containing the proportion that venue makes up out of the top 250 venues in each city. That is the value of 0.040 in the row of Akron in the column of American Restaurant tells us 4% of the top 250 venues in Akron are American Restaurants.

We also want to consider the city size, population density, and change in size. Values in the hundreds of thousands would skew values of venues which are all between 0 and 1. Therefore we use the Min Max scaler method to scale each of these values between 0 and 1. For example New York's population size would be 1 (as it is the largest city in the US) while Garden Grove would get a value of 0. Using this scaling on all three columns allows us to use them with our venue data.

Lastly, we merge these DataFrames to create one that contains a column for city, a column for each venue type, and columns for our scaled population statistics.

3.2 Using k-means clustering

The first question we want to answer is how many clusters should we partition our US cities into? To answer this, we run the K-means algorithm with different values of clusters from 1 cluster to 9 cluster and check their distortions (or sum of distances to their cluster centers). Doing so results in the following graph (Fig 1).

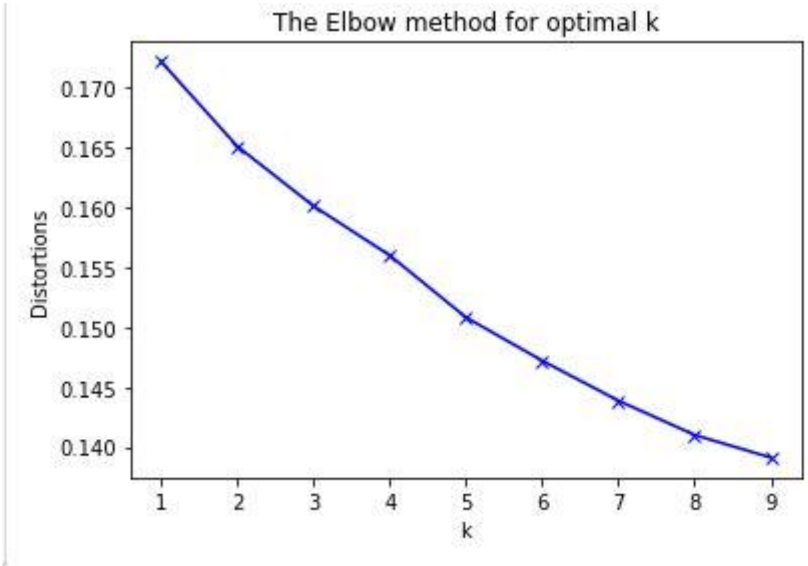

Figure 1: Distortions for various cluster values k.

Using the elbow method, we determine there are diminishing returns for clusters after 6 clusters.

From this method we run the k-means clustering algorithm on the data set using 6 clusters and assign a cluster label to each city.

4 Results

As a result of the analysis we were able to use the Folium package to make an interactive map showing cities grouped by color coding. A static image of the continental US is shown in Fig 2.

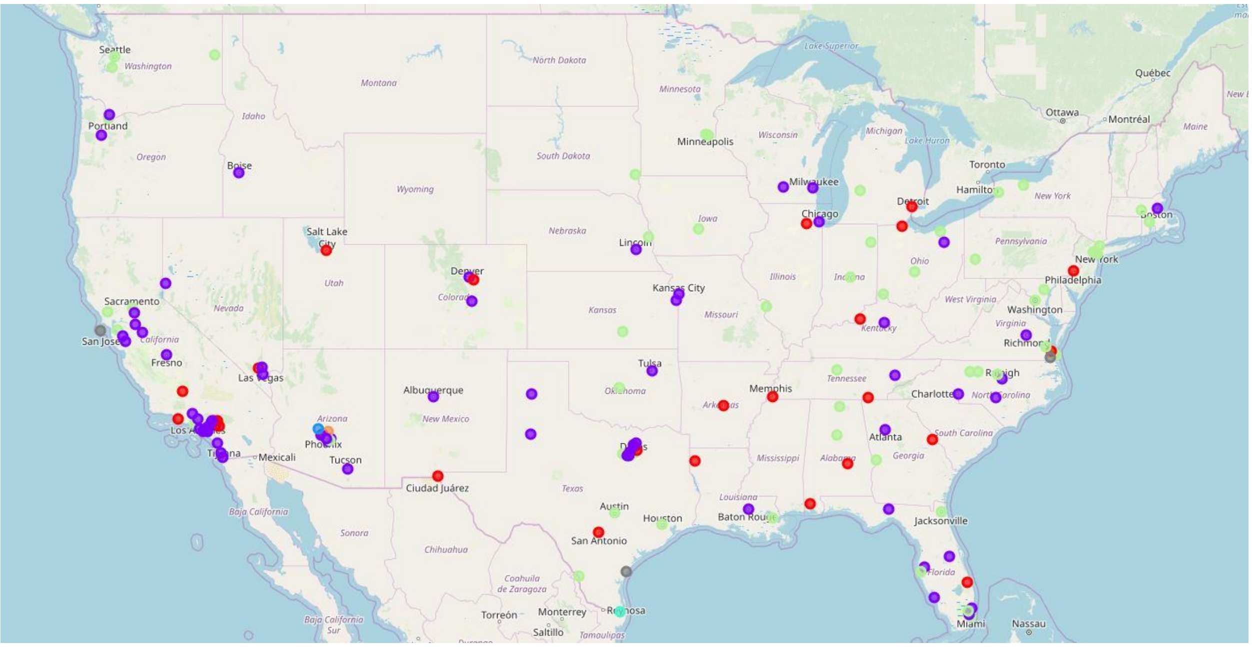

Figure 2: Clustering of cities in US

A dynamic version of this map can be found here.

5 Discussion

Our results indicate that there are six distinct city types in the United States. It can also be seen from the map that city type is not relegated to geographic location. We see that city types of all categories can be found distributed across the United States. Using this data an individual should be able to find a city of any type in any region of the US they desire. Companies likewise should be able to expand to any region into a similar cultural market as one they are currently operation in.

6 Conclusion

Our discussion concludes by noting that a similar analysis could be run using more cities, a larger radius for venues or scaled to use more venue values. The basic frame work for expanding this analysis would be very similar to the analysis already run.