**Analyzing the Similarity of U.S. Cities**

Adamm Hockman

IBM Data Science Certification

Coursera.Org

25 May 2020

# Table of Contents

[0 Table of Contents 1](#_Toc42098213)

[1 Introduction 2](#_Toc42098214)

[1.1 Background 2](#_Toc42098215)

[1.2 Problem 2](#_Toc42098216)

[1.3 Interest 2](#_Toc42098217)

[2 Data 3](#_Toc42098218)

[2.1 Data Acquisition 3](#_Toc42098219)

[2.1.1 Data Sources 3](#_Toc42098220)

[2.1.2 Wrangling Method 3](#_Toc42098221)

[2.2 Data Preparation 3](#_Toc42098222)

[2.3 Feature Selection 4](#_Toc42098223)

[3 Methodology 5](#_Toc42098224)

[3.1 Exploratory Data Analysis 5](#_Toc42098225)

[3.1.1 How are cities distributed? 5](#_Toc42098226)

[3.1.2 How are venues distributed in a typical city? 6](#_Toc42098227)

[3.1.3 How are alcoholic venues and population related? 8](#_Toc42098228)

[3.2 Modelling and Predicting 8](#_Toc42098229)

[3.2.1 K-means clustering of all us cities. 8](#_Toc42098230)

[3.2.2 Classifying parks in a city. 12](#_Toc42098231)

[3.3 Secondary and Tertiary Results 14](#_Toc42098232)

[3.3.1 Alcoholic Venues 14](#_Toc42098233)

[3.3.2 Parks in a City 14](#_Toc42098234)

[4 Results 12](#_Toc42098235)

[4.1 Observations 12](#_Toc42098236)

[4.2 Recommendations 15](#_Toc42098237)

[5 Conclusion 17](#_Toc42098238)

# Introduction

## Background

There are tens of thousands of cities across the U.S. with an array of characteristics, many of which are quantifiable. What exactly gives a city its character is a surprisingly complicated question. There are many stakeholders when it comes to this type of question. Suppose a travel agency must make a recommendation on which city to visit based only on a list of cities the client has indicated they enjoy the most. Imagine a new job requires you to work remotely and will pay moving expenses for any city in the U.S. as a perk. The applicant may want to move to a new city that is similar to cities they know they like. A restaurant may be looking at expanding into new cities and wants to make sure the new cities are similar to the cities where they have had success. In all of these situations, it is not clear how to define similarity between cities.

## Problem

Here we will utilize an unsupervised learning algorithm to cluster U.S. cities based primarily on the distribution of different venues in each city. We will also consider how additional factors such as population, politics of the state they reside in, as well as health statistics for the state influence this clustering. Finally, we will explore characteristics of venue distributions and determine whether they can be used to predict the presence of parks, and we will explore how population affects the number of alcohol-themed venues in a city.

## Interest

There is both personal interest as well as corporate interest in looking at this clustering of U.S. cities. Meaningful conclusions will improve the quality of life for individuals looking for a new city to call home. They will also help expanding businesses have more success in unknown cities across the U.S.

# Data

## Data Acquisition

### Data Sources

The U.S. Census Bureau provides population data from 2010 to 2019 for every census recognized area in the U.S. We use this to build our primary dataset based on individual cities in the U.S. After getting the names and populations, we will need coordinates for each city. Unfortunately, GeoPy does not allow heavy use of their library to return coordinates for a city. Hence, we will need to cut down the size of our dataset by removing cities with too small of a population.

As every city resides within a state, certain characteristics of the state will naturally influence the metrics for a city. Given this observation, we will use a combination of sources to build a dataset of all U.S. states with the following information: (a) population, (b) political affiliation, (c) health statistics, and (d) unemployment. Population data on the states also comes from the U.S. Census Bureau. For political affiliation, we scrape the Wikipedia page containing presidential election results for all U.S. states. Health statistics are scraped directly from the Centers for Disease Control. Finally, the unemployment rates come from the U.S. Bureau of Labor Statistics.

We will use Foursquare to obtain a list of the top 100 venues within a 500 ??? radius. This will give us a distribution of certain venue categories across each city. We are using the free developer account for API calls and may need to spread out the data collection over several days to not hit the limit of number of calls allowed per day.

### Wrangling Method

We will primarily rely on the webpage scraping tools provided by the pandas library. Using the Excel Reader and the HTML Reader, we can get all of our data besides the venue information. To get the venue distribution for each city we will make API calls to Foursquare using the free edition of a Foursquare developer account.

## Data Preparation

After obtaining the population figures from the U.S. Census, we need to get the latitude and longitude for each city. To achieve this, we rely on the free GeoPy package offered in Python. However, this service does not allow heavy use. We cannot make more than one call per second. The census data consists of 19,500 cities and would take approximately five and a half hours to complete. To reduce this overhead, we choose to remove those cities with a population of less than 20,000. This reduces our sample to 1,800 cities. In fact, since our venue data comes from Foursquare, this step is necessary to obtain venue data.

Next, we can add the state data to our data frame using an inner join on the state. After some cleaning (i.e. removing redundant columns, re-labelling, and sorting), we have each city as well as the auxiliary statistics associated with it.

Our final step is to obtain the venue data. Using the Foursquare API calls, we get the top 100 venues in each city, within a 5,000 meter radius of the coordinates. We then apply one-hot encoding to the venue category column, group the results by city, and take the sum over each group. Hence, we are left with the total number of venues of various categories (approximately 500) within each city.

If a city’s coordinates could not be found using GeoPy, or if a city returns no venues, then we will drop these rows. This does not affect our outcome since we will be using an unsupervised learning algorithm.

## Feature Selection

After processing the data, we end up with 1,737 total cities, with 558 features. The features we use to train our model include the following:

|  |  |
| --- | --- |
| Table 1. Feature selection for primary dataset (city-level). | |
| Column | Description |
| 'Location' | “City, State” (e.g. “St. Louis, Missouri”) |
| 'City Population 2019' | The cities estimated population for 2019, by the U.S. Census Bureau. |
| 'City Growth (% since 2010)' | The total population in 2019 divided by the total population in 2010. |
| 'City Latitude' | City latitude returned by geopy. |
| 'City Longitude' | City longitude returned by geopy. |
| 'Total Venues' | Total venues returned by Foursquare. |
| 'ATM' | Total number of ‘ATM’ for city. |
| 'Accessories Store' | Total number of 'Accessories Store' for city. |
| 'Adult Boutique' | Total number of 'Adult Boutique' for city. |
| … | … |
| "Women's Store" | Total number of " Women's Store" for city. |
| 'Yoga Studio' | Total number of 'Yoga Studio' for city. |

|  |  |
| --- | --- |
| Table 2. Feature selection for secondary dataset (state-level). | |
| Column | Description |
| 'State' | State name (e.g. ‘Missouri’) |
| 'State Population 2019' | Population estimated by the census. |
| 'State Growth (% since 2010)' | Population growth for state. |
| 'State Birth Rate' | Births per 100k |
| 'State Death Rate' | Deaths per 100k |
| 'State Election 2000' | One-hot encoded with \_D and \_R |
| 'State Election 2004' | One-hot encoded with \_D and \_R |
| 'State Election 2008' | One-hot encoded with \_D and \_R |
| 'State Election 2012' | One-hot encoded with \_D and \_R |
| 'State Election 2016' | One-hot encoded with \_D and \_R |
| 'State Unemployment' | Rate as of April 2020 by the BLS. |

The columns with the election results and the various venue categories have been one-hot encoded, with indicator variables in place of categorical variable values.

In the process of cleaning the data, we encountered several instances of features that contain redundant information. For example, total births is a less meaningful feature than birth rate because rates are independent of the how long the window was open to record totals, hence a more objective comparison. We also condensed the population totals for 2010-2018 into one feature—population growth across that period.

# Methodology

## Exploratory Data Analysis

### How are cities distributed?

We had to filter down the 19,500 cities returned from the U.S. Census Bureau. I wanted to make sure: (A) we do not remove too many cities from any one state, and (B) we remove cities for which it will be more difficult to obtain Foursquare data.

|  |  |
| --- | --- |
| Figure 1. Distribution of all cities. | Figure 2. Distribution of subset. |
| A screenshot of a cell phone  Description automatically generated | A picture containing drawing  Description automatically generated |

After exploring the distribution of cities across various factors, it was clear that we should remove cities that have a population below a set minimum. I landed on a minimum value of 20,000 to err on the side of keeping more cities. This reduced the number of cities to 1,800.

### How are venues distributed in a typical city?

After pulling the Foursquare venue data, we effectively have a large sample of venue distributions. One interesting question is whether or not all cities share a common venue distribution. For example, we might consider whether for all cities, roughly 40% of the venues are restaurants.

First, we need to reduce the number of venue categories by selecting more encompassing groups, i.e. ‘Restaurants’ instead of ‘Korean Restaurants’, ‘American Restaurants’, etc. all considered individually. We ended up with six distinct groups with pre-defined terms related to the category: restaurants, entertainment, stores, cars and care, education, parks and rec. If the term was found in the column names (venue categories), it was aggregated into a total for that group. Since some cities did not return the full 100 venues, it is not right to compare total counts of each group. Therefore, we took the totals and divided them by the total venues returned. We are left with the relative frequency for each group for each city.

We then looked at the distribution of these values to determine if there is any regularity to how often certain venue categories appear in each group. We made the following boxplots to conclude that the distribution is far too heavily-skewed.

|  |
| --- |
| Figure 3. Boxplots of each venue grouping. |
| A screenshot of a video game  Description automatically generated |

### How are alcoholic venues and population related?

We tried modelling the number of venues with alcohol licenses (not including big box stores) to determine if there were some perfect conditions for obtaining high liquor licenses in an area (suppose your business requires a license). We found the following, with what appear to be roughly 4 clusters. There appear to be 3 or 4 planes onto which the data could be projected.

|  |  |
| --- | --- |
| Figure 4. Alcoholic venues and population. | |
| A close up of a map  Description automatically generated | A picture containing text  Description automatically generated |

## Modelling and Predicting

### K-means clustering of all us cities.

This was the main aim of our study. As described in the data section, here we used a collection of 1,800 cities across the U.S. with various statistics on population, health, state politics, as well as the data on roughly 100 venues per city. To get a better idea of how many cities we are comparing, consider the map below:

|  |
| --- |
| Figure 5. Cities in the dataset. |
| A close up of a map  Description automatically generated |

For each of the cities, we did use all of the columns as features for clustering—except ‘Latitude’ and ‘Longitude’ since those are not a descriptive feature for a city. Once the data was in all numeric form, we used the K-Means clustering algorithm implemented via the scikit-learn package in python. We then created a function to visualize the clustering using the folium package.

The most difficult part of modelling this was choosing the right value for . The following shows the clustering assignments for .

|  |  |
| --- | --- |
| Figure 6. K-Means clustering results. | |
| A close up of a map  Description automatically generated |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

There is another issue when choosing , and that is for some ‘middle-ground’ values, we see individual cities assigned to their own cluster. While interesting, this defeats the purpose of the study to identify ‘similar cities’ for all of the cities in our dataset. We decide on a value of by observing the distribution of cities per cluster for values up to 15. The histograms below count the size of each cluster for values of .

|  |
| --- |
| Figure 7. Choosing the value of based on cluster sizes. |
|  |

### Classifying parks in a city.

This inquiry concerned the presence of parks and whether we can use the distribution of venues in the city to classify whether or not a city has a park. I decided to use a logistic regression model with the one-vs-all classification method. Since we are dealing with strictly indicator variables as the inputs, some of the other models seemed inappropriate. Without normalizing the data (since we have zeros and ones), we split the data into 70% training, and 30% test.

# Results

## Observations

Here we primarily discuss the results from clustering the cities. From Figure 7, we can see what happens as the number of clusters increases. With this being the most subjective part of the study, it is important to remember the choice is arbitrary and could be improved depending on the context of a specific application.

Considering the map for , if we swap the colors then we essentially have a political map showing ‘liberal’ and ‘conservative’ leaning states in past presidential elections. This effectively illustrates the impact of choosing too few of clusters. Since political affiliation of cities’ home states was not the main goal of this study, we need to increase at least until the obvious political affiliations become blurred and even unrecognizable. We want the venues to be the primary predictors, with the additional city statistics as secondary predictors, and the state statistics as tertiary predictors.

When deciding on a value of that keeps the clusters fairly robust (see Figure 8), something peculiar happened starting at clusters. For , the K-Means model consistently returned five main clusters, with the remaining clusters populated by an extremely small number of cities (i.e. less than 10). This suggests that there is something inherent to the data set that naturally forms five clusters.

Unfortunately, using does not actually identify all five clusters. Below is a visualization of the clustering assignment.

|  |
| --- |
| Figure 8. Using k=5 clusters. |
| A close up of a map  Description automatically generated |

In fact, the map above still appears to be too heavily influenced by the state election results. Therefore, we selected one of the models for and then removed the cities assigned to relatively small clusters.

For the final clustering, we chose clusters, then removed cities in a cluster with size less than 10. We obtained the following distribution among the clusters:

|  |
| --- |
| Figure 9. Distribution of final clusters. |
| A screenshot of a cell phone  Description automatically generated |

After removing the clusters of negligible size, we visualized our final clustering below. We can see an interesting partitioning of cities that seems more reliant on features outside political affiliation.

|  |
| --- |
| Figure 10. Final clustering. |
| A picture containing text, map  Description automatically generated |

## Secondary Results

### Alcoholic Venues

There does not appear to be any inherent clustering upon analysis of the DBSCAN. Perhaps with additional tuning of parameters for the model we could identify clusters, but we may also be seeing the clusters do the issues of scale and point-of-view.

|  |
| --- |
| Figure 11. DBSCAN clusters. |
| A close up of a piece of paper  Description automatically generated |

### Parks in a City

After training the logistic classifier, we tested the model on our data, and the results were not strong. The Jaccard similarity score for the model was 0.4698, and the confusion matrix is shown below. It does not appear that the distribution of venues in a city is a good predictor of whether the city is invested in parks.

|  |
| --- |
| Figure 12. Confusion matrix for parks classifier. |
| A screenshot of a cell phone  Description automatically generated |

# Discussion

## Observations

We attempted to decipher why these cities were clustered in such a way. First, we aggregated all the data into the five clusters, including the total venues. Next, we looked at the relative frequency of each venue category. We use the same groupings from our exploration of venue distribution earlier, with the addition of ‘Fitness’. We aggregate the columns under these groupings and consider the bar plots showing the distributions below. Note that as before we divide by the total number of venues for each cluster. This is to use relative frequency versus totals as a better comparison tool.

|  |  |
| --- | --- |
| Figure 13. Comparing clusters by venue categories. | |
| *Cluster 0* | A screenshot of a social media post  Description automatically generated |
| *Cluster 1* | A screenshot of a cell phone  Description automatically generated |
| *Cluster 2* | A screenshot of a cell phone  Description automatically generated |
| *Cluster 3* | A screenshot of a social media post  Description automatically generated |
| *Cluster 4* | A screenshot of a social media post  Description automatically generated |

From here it is a question of whether these distributions are significantly different. This could be an area for further study. By inspection, cluster 0 has a far higher percent of their venues accounted for by restaurants (66%). Cluster 4 appears to have a higher percentage of venues in entertainment (26%). Clusters 1 and 2 both have more shopping than the other clusters with 26% and 25%, respectively. Cluster 3 does not appear to have any distinguishing characteristics based on this assignment of terms.

## Recommendations

There are many directions one might go to further this study. A deeper exploration of what features are most relevant in clustering is certainly warranted. It is also worth doing significant feature analysis on the full data set. The number of cities and the number of clusters should both be significantly increased (with additional computing power) for a clearer understanding of what these clusters represent. We might also recommend a DBSCAN and use something like relative frequency for each city. This will give more input variation for logistic regression in the later predictions. In any case, given the right tweaking this type of analysis may be vital to a vested party’s success.

It is also necessary to state that the lists used for filtering the venue categories into broader groups was arbitrary and the more fine-tuned that stage, the higher quality analysis you can perform on the variety of clusters, using the same methodology presented here.

# Conclusion

This study was more of a beginning to a larger body of work, rather than a conclusion in its own right. It is clear that more data is needed to really appreciate the insights here. However, it was still curious to observe an entertainment cluster, a shopping cluster, and a restaurant cluster. That might be useful information to the right people.