Coursera Capstone Project

June 2020

**Building a Recommendation System for Finding Cities to Live in the United States of America**

Avnika Manaktala

# Introduction

*This project is part of the IBM Applied Data Science Capstone Project offered on Coursera.*

According to Wordometers the United States of America currently has a total population of 330.8 million people[[1]](#footnote-1). Although cities comprise of just 3.5% of the total land area in the country, 62.7% of the population chooses to live in these metropolitan centers[[2]](#footnote-2). These cities are most often not just a congregation of people, but culture and services.

The world today is very interconnected. With the development of technology, it is possible for people to continue to work and communicate remotely from all corners of the world. People today are also constantly in the search of finding a place to live where they are comfortable with sufficient facilities and services provided. It is therefore important for them to have a one stop solution where they can find a list of places where they are most comfortable living based on the services they require.

I am currently a graduate student studying in a small college town; however, I am graduating soon and I would like to know which major cities could offer me the best services. Like me, there are many people who are looking to find cities to live in that could offer the services they need. Therefore, this project would be suitable for people who are looking to move to a new city in the United States.

This project will ask people to rate their preferences on what services they prefer and recommend cities that would be the most suitable for them. To achieve this, cities will be clustered on how similar they are based on the services they offer as listed in the Foursquare API, and the entered data will be used to identify the cluster of choice that will be displayed.

# Data

## Data Sources

Based on the problem described above, the following are the factors that will influence the results of the search:

* The type of services that are offered in each city such as coffee shops, restaurants, parks and theaters
* The number of such services within each city which will provide an indication of what are the most popular services in each city

The data I will be using for this project is listed below:

1. Foursquare API: This API will be used to identify services, restaurants and shops available in each city and the frequency of each service that is present in each city, which will give an idea of its popularity.

2. U.S. Largest Cities (Source: Open data for largest cities in the US[[3]](#footnote-3)): This dataset holds data on the major cities in the U.S. and the location columns will be used for mapping the cities on the map. The cities listed in this database will be the candidates examined for this project.

## Data Cleaning

The cities dataset came as a downloadable csv file, which was convenient as no web scraping needed to be done. It contained the following columns listed with the description for each column:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| City | Name of the city |
| Rank | Rank (Population) |
| State | State it is located in |
| Growth from 2000 to 2013 | Population Growth from 2000-2013 |
| Population | Population of city (Numbers) |
| Coordinates | Latitude and Longitude |

*Table 1: List of columns and their descriptions of the cities dataset*

The coordinates column was split to produce separate latitude and longitude columns for ease of use for the rest of the analysis. Furthermore, the Rank, Growth from 2000 to 2013, Population and Coordinates columns were dropped as they were irrelevant for the analysis done in this project.

The data was then processed to check for NA values and duplicates. None were found, likely because this was an already processed and fairly popular dataset. The location of the cities was then visualized on a map as seen in Figure 1.

A close up of a map

Description automatically generated

*Figure 1: All cities considered for this analysis visualized on a map*

# Methodology

The cities dataset was used to parse for services in each city along with which category they fell under using the Foursquare API. The services were limited to a 25 Km radius along the central coordinates provided to account for large cities. Furthermore, the services were limited to a maximum of 2000 services as this acts as a large enough sample to understand the characteristics of each city. Table 2 shows the first 5 rows from the venues dataset that was generated using the API.

A screenshot of a cell phone

Description automatically generated

*Table 2: First 5 rows from the venues dataset created from the Foursquare API*

The services were then encoded for each city and the frequency of each service category was calculated to produce a dataset of the most common service categories for each city. This is shown in Table 3 that shows the first 5 rows of this dataset.

A screenshot of a cell phone

Description automatically generated

*Table 3: Most common services for each city*

This dataset was used to perform a K-means clustering analysis to group similar cities together. K- means clustering is a simple unsupervised machine learning technique where similar data points are put together in clusters[[4]](#footnote-4). For this analysis, I have chosen a k value of 10 as it provides a more specific list of cities for the user that can cater best to their interests. The average distance from each point to the cluster is 0.14, which is acceptable for this analysis.

Once the cities were clustered in one of 10 clusters based on the frequency of the services they offer, these clusters were visualized on a map as shown in Figure 2.

A close up of a map

Description automatically generated

*Figure 2: Cities analyzed for this analysis and the different clusters they fall in(color)*

Using this above analysis, the program then asks the user for the services they perceive are the most important in a city. This is taken in as raw input and a separate row containing the preferences of the user is added into a separate dataset which is a clone of the most common services dataset shown in Table 3. This user preference dataset is shown in Table 4.

A screenshot of a cell phone

Description automatically generated

*Table 4: Cities with their most common venues along with user preference*

The K- means clustering code is then run to display which cluster the user defined preferences are closest to. In my case, this has given a result of cluster 6. A table containing the cities in cluster 6 along with their most common venues are then displayed as shown in Table 5.

A screenshot of a cell phone

Description automatically generated

*Table 5: Cities that are in cluster number 6 along with their most common venues*

In addition to the table the user may also visualize these cities on a map as shown in Figure 3 below.

A close up of a map

Description automatically generated

*Figure 3: Cities in Cluster 6 as visualized on a map.*

# Results and Discussion

The results show the different cities that are suitable for users with different interests. The number of clusters chosen for this project was 10, which means it can only recommend that the user's city would be in clusters 0-9. The following list shows what venues seem to be the most popular in each cluster:

* Cluster 0: Fast Food Restaurants, American Restaurants and Mexican Restaurants are the most important venues
* Cluster 1: American Restaurants and Parks seem to be the most important venues
* Cluster 2: Fast Food Restaurants, Theme Parks and Grocery Stores are the most important venues
* Cluster 3: Coffee Shops and Parks seem to be the most important venues
* Cluster 4: Beaches, Seafood Restaurants and Parks seem to be the most popular
* Cluster 5: Coffee Shops, Grocery Stores and Pizza Places seem to be the most popular venues
* Cluster 6: Grocery stores, Hot Dog Joints, Bakeries and Breweries are the most important venues
* Cluster 7: Supermarkets and Coffee Shops seem to be the most important venues
* Cluster 8: Fast Food Restaurants, Grocery stores and Coffee Shops are the most important venues
* Cluster 9: Parks, Bakeries and Pizza Places are the most important venues

While this project is able to classify the cities into 10 main groups, it could be helpful if an estimation can be done on the error that the clustering has, and this could be taken into account when dividing the cities into the number of clusters. A deeper analysis on the most effective number of clusters could be done to improve the model.

# Conclusion

While this project gives the suggested cities users may look into based on their requirements for the most popular services, there are other considerations, such as the size of the city, crime rate and housing prices, that the users should look into before deciding to move. This tool can give a suggestion on the cities as a starting point for users to look into and research about deeper.

Future work into this project could include an analysis on crime, size and housing cost as part of the analysis and ask the user specifically for their preferences on these statistics. This will give them a well-rounded analysis as it would be a one-stop solution for understanding the cities that are best suited for their needs. Furthermore, a deeper analysis on the best k value would prove to be beneficial to providing the user with a more accurate set of cities.

This project, therefore, uses a starting point of the largest cities in the U.S., uses the Fourquare API to identify the most important services in each city and clusters similar cities together. The code then asks the user to add the services most important to them in order and finds a cluster of cities that are most suitable for them.

1. <https://www.worldometers.info/world-population/us-population/> [↑](#footnote-ref-1)
2. <https://www.census.gov/newsroom/press-releases/2015/cb15-33.html#:~:text=U.S.%20Cities%20are%20Home%20to%2062.7%20Percent%20of%20the%20U.S.,3.5%20Percent%20of%20Land%20Area&text=A%20majority%20of%20the%20U.S.,by%20the%20U.S.%20Census%20Bureau.> [↑](#footnote-ref-2)
3. https://public.opendatasoft.com/explore/dataset/1000-largest-us-cities-by-population-with-geographic-coordinates/table/?sort=-rank [↑](#footnote-ref-3)
4. <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1> [↑](#footnote-ref-4)