Capstone: Battle of the Cities

Business Problem

Have you every wondered if you should move to a place for a specific job? Whether you'd be happy in a place if you were to move?

The USA has over nineteen-thousand cities, towns and villages. About 310 of them are considered at least medium size with populations above 100,000. It's easy to be overwhelmed when deciding to move, but it is especially true when you're not sure what to expect from the place you might be moving to.

Most people know in some way or another a city like New York, Boston, San Francisco, Chicago or Miami, but are otherwise unaware of the multitude of incredible cities in the US where they may be happier. People may be considering taking a job in a town they know little about, without knowing how similar it is to the cities they know they like or dissimilar from cities they know they dislike.

As an example, I love Boston and really like New York. I also really like Raleigh and Durham in North Carolina, but had I not attended Duke University, I would have not known. So now that I am graduating and considering job offers… where should I consider taking job in based on my preferences?

Our objective of this project is just that: clustering cities based on an amalgam of features and creating a recommendation system that, based on my input (or yours!), recommends cities based on these clusters and the city profiles of your rated cities.

Our main target audience are international students and people seeking and considering moving for professional purposes.

Data Requirements and Data Collection

To properly segment cities in the USA and create a competent recommender system, I need to generate detailed city profiles for each city in the study. This begs the following questions:

Which cities will I include?

The Wikipedia site 'List of United States cities by population' deemed appropriate for these purposes. Ideally, we would try to include a lot more cities, but the computational requirements to use thousands of cities instead of the largest 315 by population seems too high for the marginal benefit.

What data will be collected and how?

There is an enormous amount of data available on the web regarding cities, but a lot of it comes from untrustworthy sources, nor is it standardized. Lucky me, I found datausa.io, from which I can scrape a lot of relevant information from each city and generate good city profiles (Thank you Deloitte and Datawheel!). I will selectively scrape some metrics that I believe are important and combine them with information about the categories of the top 100 venues in each city from Foursquare's API to complete the city profiles.

Some of the metrics scraped and included are:

- Population and Population Change (Year to Year)
- Poverty Rate
- Median Age
- Median Household Income and Median Household Income Change (Year to Year)
- Number of Employees and Number of Employees Change (Year to Year)
- Median Property Value and Median Property Value Change (Year to Year)
- Average Male and Female Salary, and a ratio of Average Male to Female Salary
- Gini coefficient in 2017 and 2018, as well as it's change (Year to Year)
- Ratio of Patients to Clinicians (county-wise)
- Foreign-born population percentage
- Citizen population percentage
- Total degrees awarded in 2018 (higher education)
- Male to Female ratio of awarded degrees
- Number of degrees per capita
- Number of households in city
- Population per household (people per household)
- Homeownership Percentage (Rent vs Own)
- Average Commute Time (minutes)

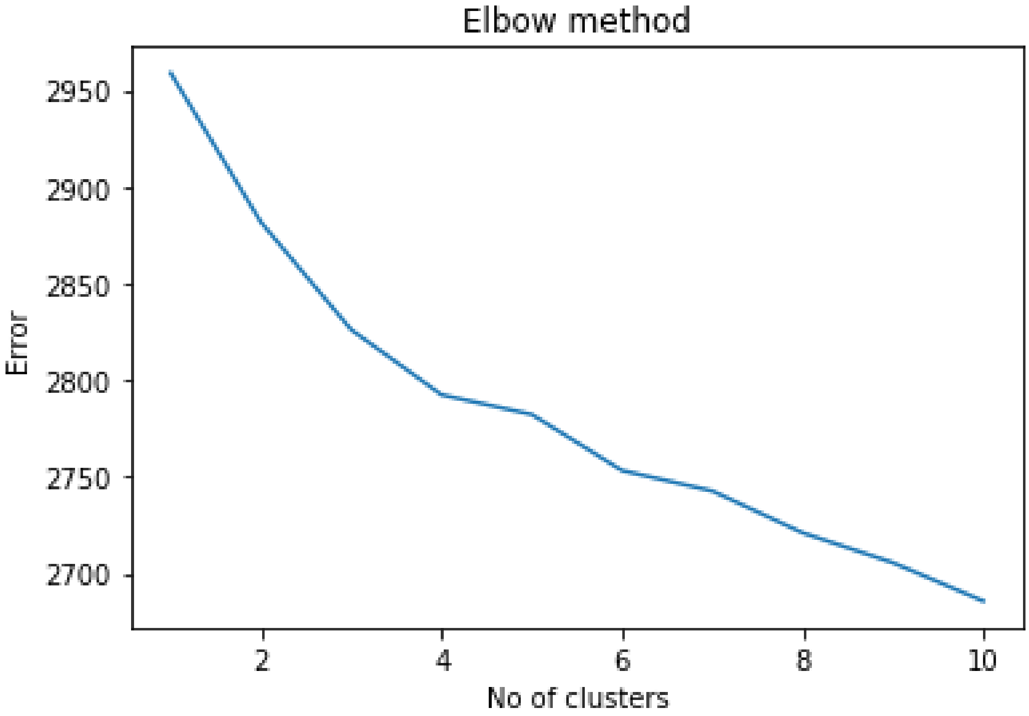
To see the final dataset, click here.

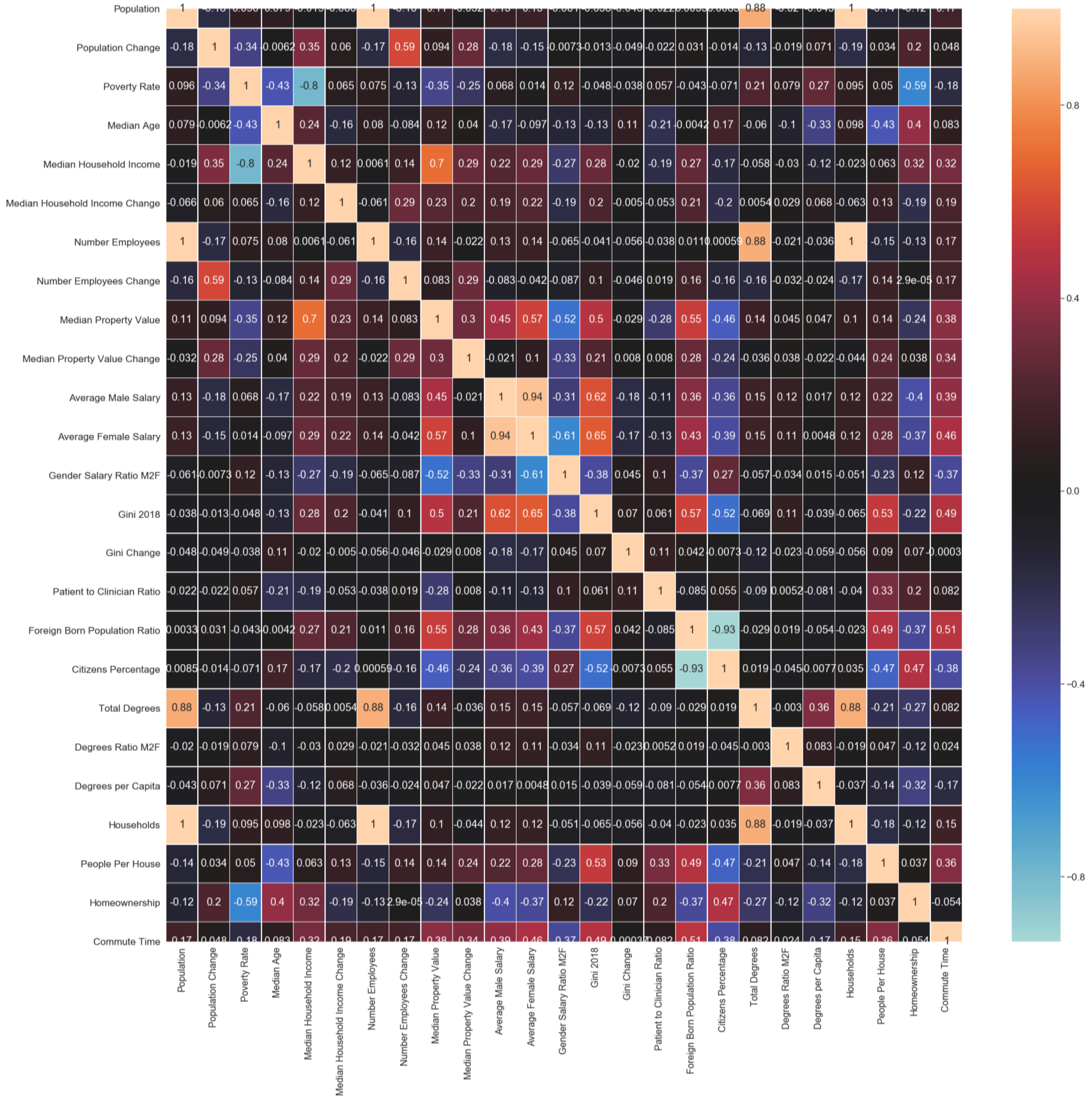
Methodology

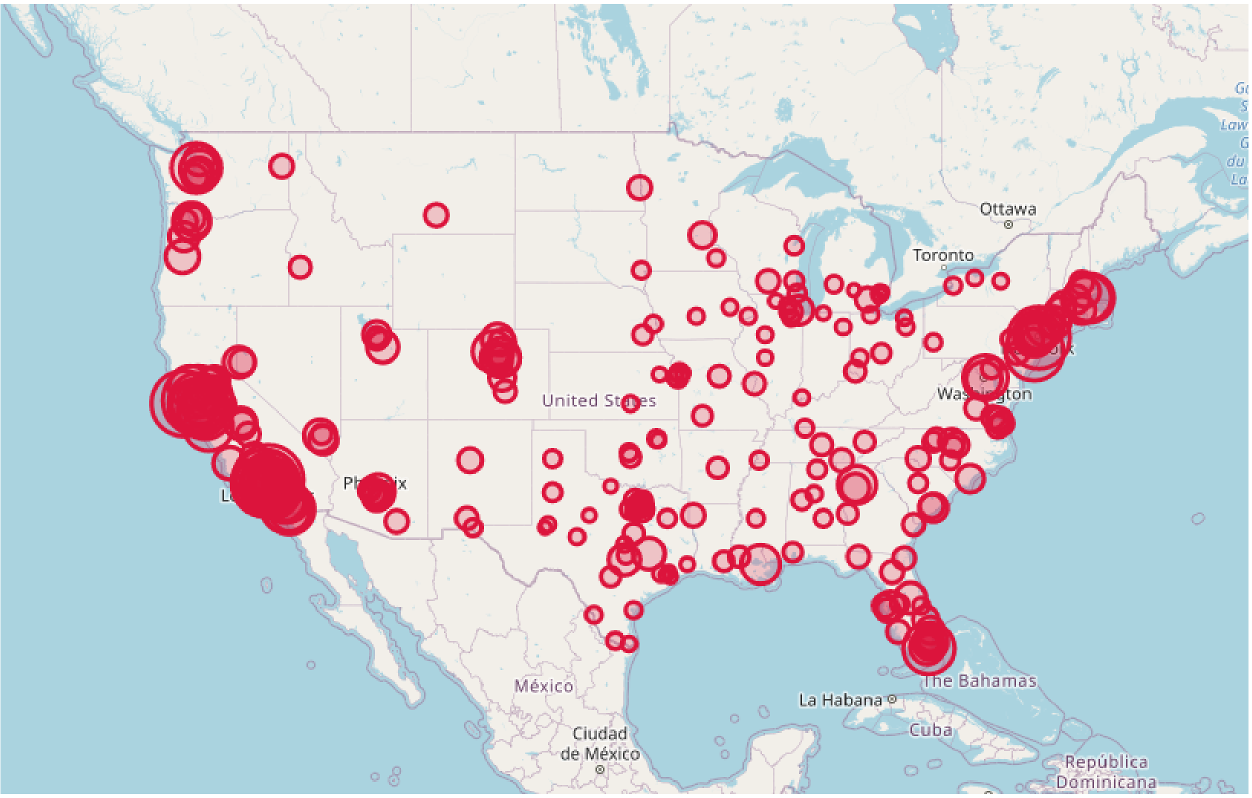
After the data was gathered, I briefly analyzed the data by summarizing basic statistics for each variable and plotting a correlation plot. While interesting insights lie within the correlation plot and variable statistical descriptions, feature selection was already done based on intuition during the data collection stage. Furthermore, the machine learning model that we applied does not require an in-depth analysis of each feature.

Before modeling, I preprocessed the data, normalizing every variable.

The model utilized is kMeans, to cluster the cities based on the scraped and collected metrics. Using the elbow method, I tried to determine the number of clusters to include. However, not matter how many clusters I included in the Elbow method, the error kept decreasing without a clear ‘elbow’. Hence, I opted to use the default: 8 clusters.

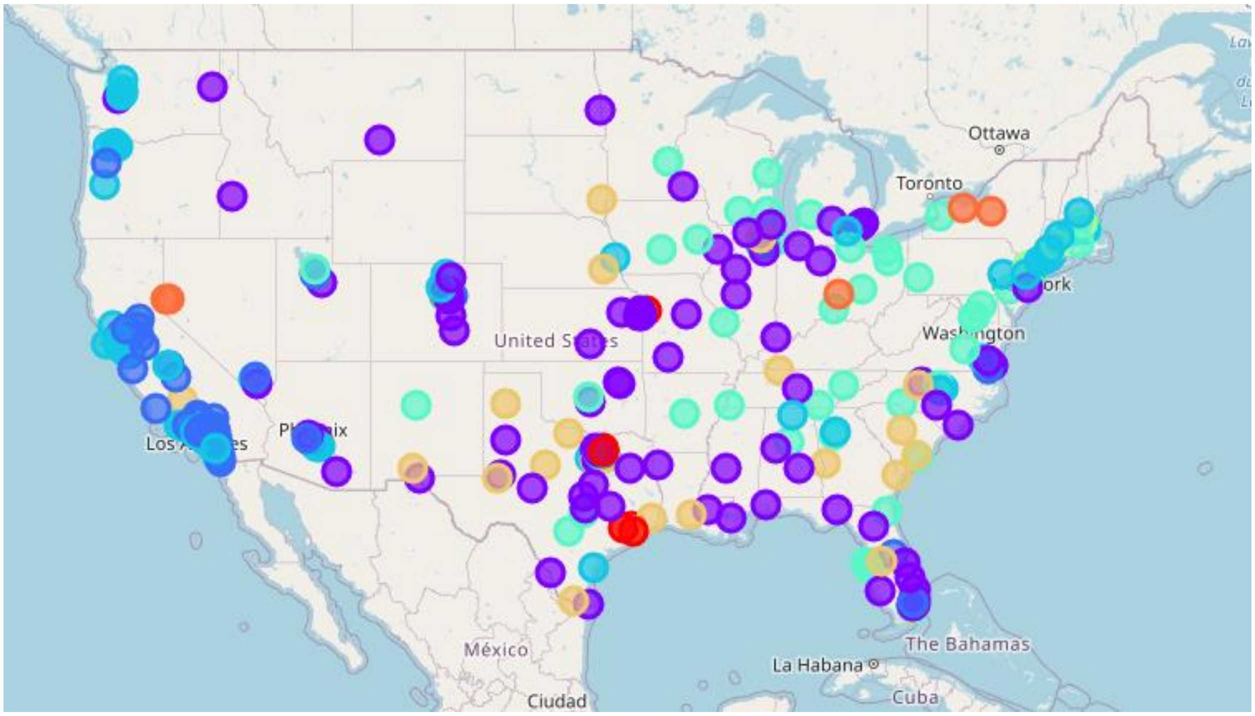

Elbow method
Error
2950
2900
2850
2800
2750
2700
2
4
6
8
10
No of clusters


Population 1 1 0.88 1
Population Change -0.18 1 -0.34 -0.0062 0.35 0.06 -0.17 0.59 0.094 0.28 -0.18 -0.15 -0.0073 -0.013 -0.049 -0.022 0.031 -0.014 -0.13 -0.019 0.071 -0.19 0.034 0.2 0.048
Poverty Rate 0.096 -0.34 1 -0.43 -0.8 0.065 0.075 -0.13 -0.35 -0.25 0.068 0.014 0.12 -0.048 -0.038 0.057 -0.043 -0.071 0.21 0.079 0.27 0.095 0.05 -0.59 -0.18
Median Age 0.079 -0.0062 -0.43 1 0.24 -0.16 0.08 -0.084 0.12 0.04 -0.17 -0.097 -0.13 -0.13 0.11 -0.21 -0.0042 0.17 -0.06 -0.1 -0.33 0.098 -0.43 0.4 0.083
Median Household Income -0.019 0.35 -0.8 0.24 1 0.12 0.0061 0.14 0.7 0.29 0.22 0.29 -0.27 0.28 -0.02 -0.19 0.27 -0.17 -0.058 -0.03 -0.12 -0.023 0.063 0.32 0.32
Median Household Income Change -0.066 0.06 0.065 -0.16 0.12 1 -0.061 0.29 0.23 0.2 0.19 0.22 -0.19 0.2 -0.005 -0.053 0.21 -0.2 0.0054 0.029 0.068 -0.063 0.13 -0.19 0.19
Number Employees 1 -0.17 0.075 0.08 0.0061 -0.061 1 -0.16 0.14 -0.022 0.13 0.14 -0.065 -0.041 -0.056 -0.038 0.011 0.00059 0.88 -0.021 -0.036 1 -0.15 -0.13 0.17
Number Employees Change -0.16 0.59 -0.13 -0.084 0.14 0.29 -0.16 1 0.083 0.29 -0.083 -0.042 -0.087 0.1 -0.046 0.019 0.16 -0.16 -0.16 -0.032 -0.024 -0.17 0.14 2.9e-05 0.17
Median Property Value 0.11 0.094 -0.35 0.12 0.7 0.23 0.14 0.083 1 0.3 0.45 0.57 -0.52 0.5 -0.029 -0.28 0.55 -0.46 0.14 0.045 0.047 0.1 0.14 -0.24 0.38
Median Property Value Change -0.032 0.28 -0.25 0.04 0.29 0.2 -0.022 0.29 0.3 1 -0.021 0.1 -0.33 0.21 0.008 0.008 0.28 -0.24 -0.036 0.038 -0.022 -0.044 0.24 0.038 0.34
Average Male Salary 0.13 -0.18 0.068 -0.17 0.22 0.19 0.13 -0.083 0.45 -0.021 1 0.94 -0.31 0.62 -0.18 -0.11 0.36 -0.36 0.15 0.12 0.017 0.12 0.22 -0.4 0.39
Average Female Salary 0.13 -0.15 0.014 -0.097 0.29 0.22 0.14 -0.042 0.57 0.1 0.94 1 -0.61 0.65 -0.17 -0.13 0.43 -0.39 0.15 0.11 0.0048 0.12 0.28 -0.37 0.46
Gender Salary Ratio M2F -0.061 -0.0073 0.12 -0.13 -0.27 -0.19 -0.065 -0.087 -0.52 -0.33 -0.31 -0.61 1 -0.38 0.045 0.1 -0.37 0.27 -0.057 -0.034 0.015 -0.051 -0.23 0.12 -0.37
Gini 2018 -0.038 -0.013 -0.048 -0.13 0.28 0.2 -0.041 0.1 0.5 0.21 0.62 0.65 -0.38 1 0.07 0.061 0.57 -0.52 -0.069 0.11 -0.039 -0.065 0.53 -0.22 0.49
Gini Change -0.048 -0.049 -0.038 0.11 -0.02 -0.005 -0.056 -0.046 -0.029 0.008 -0.18 -0.17 0.045 0.07 1 0.11 0.042 -0.0073 -0.12 -0.023 -0.059 -0.056 0.09 0.07 -0.0003
Patient to Clinician Ratio -0.022 -0.022 0.057 -0.21 -0.19 -0.053 -0.038 0.019 -0.28 0.008 -0.11 -0.13 0.1 0.061 0.11 1 -0.085 0.055 -0.09 0.0052 -0.081 -0.04 0.33 0.2 0.082
Foreign Born Population Ratio 0.0033 0.031 -0.043 -0.0042 0.27 0.21 0.011 0.16 0.55 0.28 0.36 0.43 -0.37 0.57 0.042 -0.085 1 -0.93 -0.029 0.019 -0.054 -0.023 0.49 -0.37 0.51
Citizens Percentage 0.0085 -0.014 -0.071 0.17 -0.17 -0.2 0.00059 -0.16 -0.46 -0.24 -0.36 -0.39 0.27 -0.52 -0.0073 0.055 -0.93 1 0.019 -0.045 -0.0077 0.035 -0.47 0.47 -0.38
Total Degrees 0.88 -0.13 0.21 -0.06 -0.058 0.0054 0.88 -0.16 0.14 -0.036 0.15 0.15 -0.057 -0.069 -0.12 -0.09 -0.029 0.019 1 -0.003 0.36 0.88 -0.21 -0.27 0.082
Degrees Ratio M2F -0.02 -0.019 0.079 -0.1 -0.03 0.029 -0.021 -0.032 0.045 0.038 0.12 0.11 -0.034 0.11 -0.023 0.0052 0.019 -0.045 -0.003 1 0.083 -0.019 0.047 -0.12 0.024
Degrees per Capita -0.043 0.071 0.27 -0.33 -0.12 0.068 -0.036 -0.024 0.047 -0.022 0.017 0.0048 0.015 -0.039 -0.059 -0.081 -0.054 -0.0077 0.36 0.083 1 -0.037 -0.14 -0.32 -0.17
Households 1 -0.19 0.095 0.098 -0.023 -0.063 1 -0.17 0.1 -0.044 0.12 0.12 -0.051 -0.065 -0.056 -0.04 -0.023 0.035 0.88 -0.019 -0.037 1 -0.18 -0.12 0.15
People Per House -0.14 0.034 0.05 -0.43 0.063 0.13 -0.15 0.14 0.14 0.24 0.22 0.28 -0.23 0.53 0.09 0.33 0.49 -0.47 -0.21 0.047 -0.14 -0.18 1 0.037 0.36
Homeownership -0.12 0.2 -0.59 0.4 0.32 -0.19 -0.13 2.9e-05 -0.24 0.038 -0.4 -0.37 0.12 -0.22 0.07 0.2 -0.37 0.47 -0.27 -0.12 -0.32 -0.12 0.037 1 -0.054
Commute Time -0.17 0.048 -0.18 0.083 0.32 0.19 0.17 0.17 0.38 0.34 0.39 0.46 -0.37 0.49 0.00037 0.082 0.51 -0.38 0.082 0.024 -0.17 0.15 0.36 -0.054 1
Population
Population Change
Poverty Rate
Median Age
Median Household Income
Median Household Income Change
Number Employees
Number Employees Change
Median Property Value
Median Property Value Change
Average Male Salary
Average Female Salary
Gender Salary Ratio M2F
Gini 2018
Gini Change
Patient to Clinician Ratio
Foreign Born Population Ratio
Citizens Percentage
Total Degrees
Degrees Ratio M2F
Degrees per Capita
Households
People Per House
Homeownership
Commute Time
0.8
0.4
0.0
-0.4
-0.8



Results

The clusters came out to be as follows, with clear distinctions between large metropolitan areas, richer areas and educated areas being very distinct.

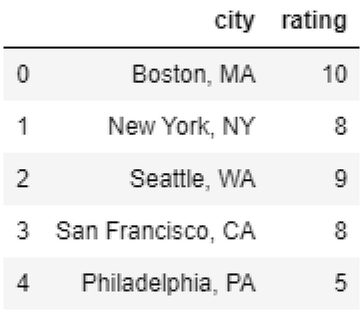


With these clusters, I can attempt to do a recommendation already. Say you like New York City. Hence, any other cities included in the cluster that NYC pertains to are likely to appeal to you as well. This cities include:

1. Sacramento, CA
2. Chicago, IL
3. Austin, TX
4. Seattle, WA
5. Durham, NC
6. Washington, DC
7. Boston, MA
8. Atlanta, GA

Recommendation System

Yet, I also built a content-based recommendation system. I rated some cities as follows:



	city	rating
0	Boston, MA	10
1	New York, NY	8
2	Seattle, WA	9
3	San Francisco, CA	8
4	Philadelphia, PA	5

According to such user profile, the cities that are recommended to the user are:

1. Jersey City, NJ
2. Los Angeles, CA
3. Cambridge, MA
4. Chicago, IL
5. Berkeley, CA
6. Pittsburgh, PA
7. Washington, DC
8. Atlanta, GA
9. Oakland, CA
10. Austin, TX

Discussion

Some of the clusters in the KMeans model are quite straightforward, including highly educated and populated metropolitan areas. However, for natives or people with more knowledge about the cities and towns in the US, this may be a good place to start if they want to predict the likelihood of liking a city or not.

The recommender model also seems to perform decently well, although the large number of variables seem to bias some of the results. Nonetheless, most of the results it reported are likely to be attractive for the specified user.

Conclusion

While the KMeans model and the recommender system seem to perform well, based on intuition, a lot more can be done to improve this project. First, a more comprehensive and concise list of metrics should be done, because the number of metrics representing a specific city's characteristics are not equally distributed and might be biasing the results. Second, it would be great if more than 300 cities could be included. Third, the characteristics of each city are equally weighted, and it may be the case that a user's preference is based on a specific characteristic of a city. Integrating a manual input parameter to give more importance to a specific metric could improve the recommender system.