CAPSTONE PROJECT: BATTLE OF THE CITIES

Elias Melul

Business Problem

Should you move for work? Would you be happy in your new city if you moved?
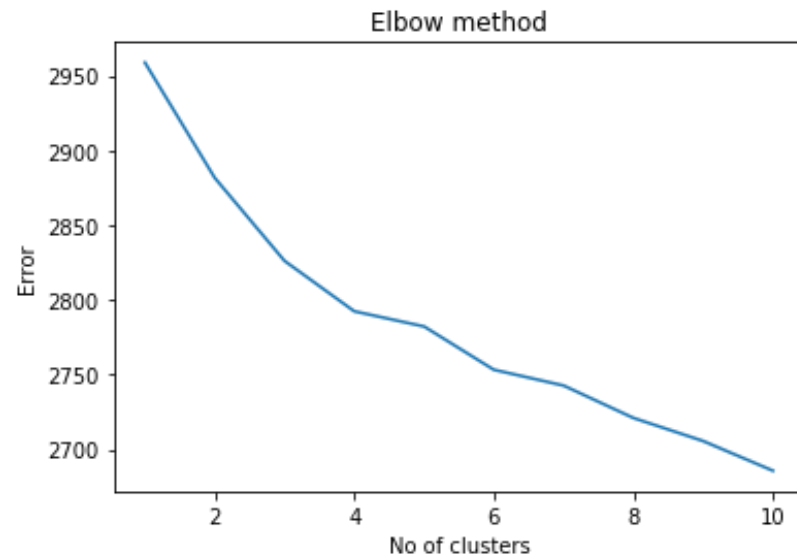
- Cluster cities based on metrics about each city
- Build recommender system to predict which cities are most recommended based on user's preferences

Data Requirements and Collection

○ Which cities will be included?

  ○ The Wikipedia site 'List of United States cities by population' deemed appropriate for these purposes. It includes the largest 315 cities by population.

○ What data will be collected and how?

  ○ Web scraping using BeautifulSoup on datausa.io

  ○ Foursquare's API

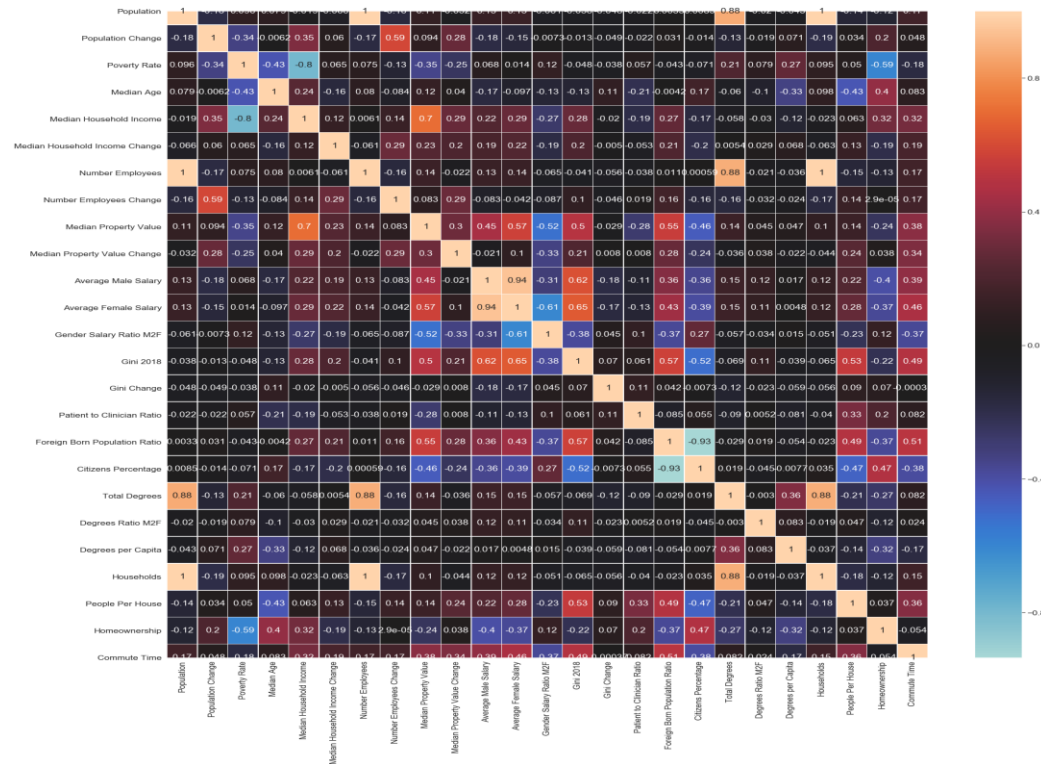○ To see the final dataset, click here.

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)
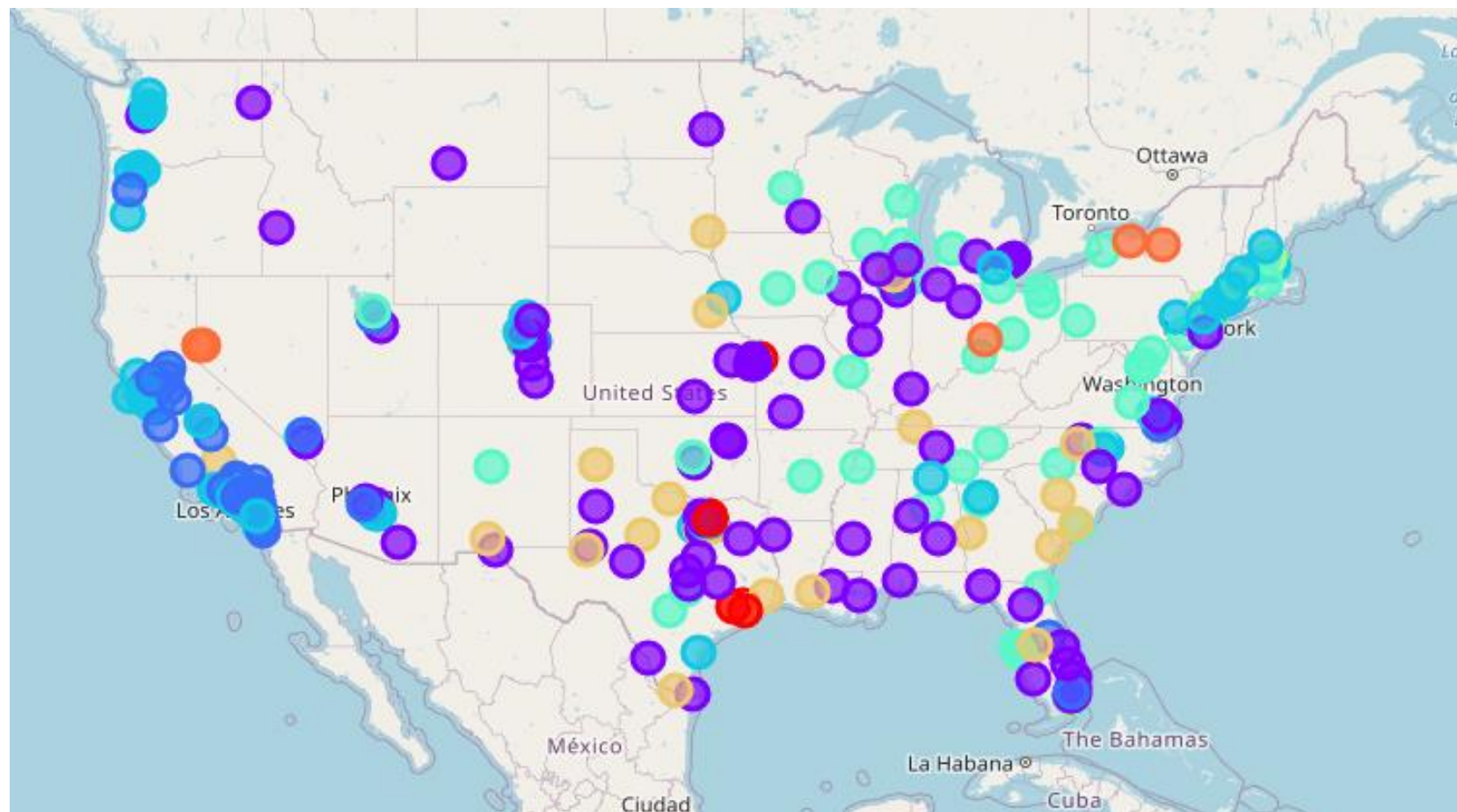- Average Commute Time (minutes)

Methodology

- Kmeans
  - 8 clusters due to lack of 'elbow'
- Content-based recommendation system

Results

◦ 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:

  ◦ Sacramento, CA

  ◦ Chicago, IL

  ◦ Austin, TX

  ◦ Seattle, WA

  ◦ Durham, NC

  ◦ Washington, DC

  ◦ Boston, MA

  ◦ Atlanta, GA

Results Continued

◦ User Profile (image)

◦ According to such user profile, the cities that are recommended to the user are:
- ◦ Jersey City, NJ
- ◦ Los Angeles, CA
- ◦ Cambridge, MA
- ◦ Chicago, IL
- ◦ Berkeley, CA
- ◦ Pittsburgh, PA
- ◦ Washington, DC
- ◦ Atlanta, GA
- ◦ Oakland, CA
- ◦ Austin, TX

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

Next Steps and Conclusion

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.

It would be great if more than 300 cities could be included

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.