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’](https://en.wikipedia.org/wiki/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](https://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](https://www.foursquare.com/)’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**](https://github.com/eliasmelul/Coursera_Capstone/blob/master/cities_profiles_freqVenues.csv)**.**