Mark Valentino

30 November 2021

Cities Recommender

The goal of this project is to recommend cities and towns to individuals based on a city or town's "safety", and "economy". Many people often relocate to new cities that they think are better than the ones they already live in. Determining which city to move to is challenging. Cities that are both safe and economical are ideal, but finding an ideal city to move to is hard to determine. Someone trying to relocate might just chose a place off of word-of-mouth or a disorganized research of a handful of cities.

This AI can solve the issue by showing the best cities it finds via Pareto optimization, where cities shown are optimized between "safety" and "economy". First, to keep terminology simple, a "city" can also be a town, "safety" is referred to as a city's population divided by total numbers of crime committed, and "economy" is referred to as a city's median income divided by its cost of living. Higher values for safety and economy are better. Pareto optimization in this case involves making a virtual plot of cities, where cities with the lowest distance from the plot origin are the best cities. In order to determine where cities are plotted, the values of their "safety" and "economy" are ranked in descending order, with the value '0' being the top possible rank. Plotted cities have their distance from the origin calculated, which is referred to as "magnitude", and lower magnitudes are better. Magnitudes then indicate how good a city is overall with its safety and economy values. The lower the Magnitude, the more Pareto efficient a city is. Magnitudes can then be sorted in descending order, leaving the best cities at the top and the worst at the bottom.

Naturally as this sorted list is iterated through, many cities will start to increasingly skew more and more either to being more safe or more economical. Users most likely would have a preference of having a safer city or a more economical city, but users probably would not like to live in a city that is very safe with a bad economy or vice versa. This magnitude sorted list of cities can then be divided into two categories, one that skews safer, one that skews economical. These categories will still be sorted in terms of balance between safety and economy. Two cities from both categories can be shown to the user.

Datasets used are from the sources as follows: Cost of Living of Cities: https://advisorsmith.com/data/coli/ Median Income of Cities: https://www.census.gov/data/datasets.html Crime and Population of Cities: https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/tables/10tbl08.xls/view Cost of Living of States: https://meric.mo.gov/data/cost-living-data-series

In order to determine how good or bad cities are, data from the datasets above were first combined into one dataset (that is what the Java program does). To help with combining data, the Java program dealt with inconsistent naming practices, such as the state of Alabama being listed as both "AL" and "ALABAMA" in different datasets. Another example is that Honolulu was listed as both "Urban Honolulu" and just "Honolulu". The selection of cities in the census dataset were chosen as the main selection of cities since the census dataset was in the middle in terms of how many cities it covered. There were cities from the FBI crime dataset and the Advisorsmith dataset that were left over in that

they could not be applied to cities from the census dataset. The dataset of cities created by the Java program had data on nearly 4,000 cities. That dataset was then loaded into this notebook and was named as dfCities.

After the data was combined, analysis of the data took place. It was determined that a city's median income correlates to its cost of living. With this known, cost of living could be roughly estimated with a function generated by machine learning. That function was tested on cities that already had COL data, and was not accurate enough to use in recommending cities to move to. An experiment was then done to see if biasing the machine learning generated function with a city's respective state COL would improve accuracy. The results of that experiment resulted in an improvement in accuracy. For cities with a median income <= 62843, the formula of the machine learning function * 0.7 + 0.3 * state COL proved to be accurate within ± 4.78 points on average. For cities with a median income >= 62843, the formula of the machine learning function * 0.4 + 0.6 * state COL proved to be accurate within ± 6.67 points on average. It was determined that these improved predictions still were not good enough for recommending cities, but would be useful in allowing users to compare and contrast cities to see if relocating is worth the trouble. Cities with imputed COL could serve as cities a user is already from.

A result was then exported into CitiesRecomenderPart2, where the result only had cities with no missing data. Also, cities with an estimated COL were flagged. Values for safety, economy, "relocation number", and magnitude were all calculated in the result. "Relocation number" refers to how many users the AI will recommend a particular city to (with the assumption the user will move there). Having a limit on how many times a city is recommended will prevent people from overloading a city.

CitiesRecomenderPart2continued off of CitiesRecomenderPart1. Results were checked to see if they were sensible, which they were. Cities were divided up into two categories, one with cities skewing safer, and one skewing more economical, as mentioned in the last notebook. To make it easier for users to compare cities, radar charts were implemented. A user could compare his or her own city with the best city (that is not filled up with relocators) in the AI, and see if it is worth getting a recommendation for two cities. If the user decides to get a recommendation, two cities from both categories are also shown on a radar chart. To make the comparison more trustworthy, a disclaimer is added on the radar charts indicating potential ways they could be misleading.

Simulations of people using the AI were ran. One had every person willing to relocate to any city shown, while the other showed people choosing to relocate if the recommended cities were significantly better than the one they already lived in. To make the simulation fairer, people “from” worse cities were sorted to the top to give them more priority. This still is arguably not as fair as it could be since wealthy people from bad cities would get priority. Determining how wealthy a person is, is another problem in its own. Less people may want to use this AI system if they have to share more private financial details.