IBM Capstone Project Report

Introduction

To determine the cities with the best quality of life, cities are often compared based on general metrics, such as population density, public transport system, green spaces, etc. For an individual, other factors might be more important when it comes to choosing a place to live. Here, I compared the 20 largest cities in Germany with respect to my specific interests, using the Foursquare venue location data. Specifically, I chose some types of restaurants I enjoy and venues related to my hobbies, which are dancing and theatre. In the notebook, those venue categories can be specified by the user, such that for a different individual, different important factors can be chosen. In addition, I took into account that the closer those venues of interest are, the more convenient it will be to travel within the city. Based on those features, I developed a score to rank the cities with a potentially good quality of life for the individual interests of a person.

Data

The 20 largest cities of Germany, their population size and location, were gathered from this Wikipedia page https://en.wikipedia.org/wiki/List_of_cities_in_Germany_by_population with the BeautifulSoup library.

The venue categories for the Foursquare API were chosen from this webpage: https://developer.foursquare.com/docs/build-with-foursquare/categories

Finally, the venues within these categories in the specified cities were explored using the Foursquare API.

Methodology

First, I used the beautifulsoup package to extract the location of the 20 largest German cities and their population. Using pandas and folium, I cleaned the data and visualized the cities.

Next, I chose specific venue categories. The whole list of venue categories in the Foursquare API can be found on https://developer.foursquare.com/docs/build-with-foursquare/categories. In my case, those venue categories were: Jazz Club, Theatre, Dance studio, Costume Shop, Malay Restaurant, Bubble Tea Shop, Molecular Gastronomy Restaurant, Night club. For each city, I defined a circular radius based on the city's area. I then explored all venues within the radius with the chosen categories. All resulting venues in all categories and all cities were stored in a pandas dataframe.

I then defined several features that were important to rank the cities: The overall number of venues of interest (VOI), The diversity of venues of interest (diversity factor, DF), The median distance of these venues from the city center (dist), the population of the city (pop). I normalized those features and calculated a score that ranked the cities based on those factors. The score was:

Score = (VOI + DF + dist)/pop

Based on this score, I ranked the cities from the most to the least livable.

Results

Discussion

Conclusion