Project Name:Analyzing the Disparity Between Cost of Living and Quality of Life
Team Member: Siyang Zhang
USC Email: szhang19@usc.edu
USC ID: 9033653228
GitHub Username: siyanz6717

Project Proposal

Research Question

People nowadays have more tendency and opportunity to work remotely, and therefore, people are increasingly interesting in seeking locations that offer high living standards at a reasonable cost. However, expensive cost does not always equal better life quality. This project aims to solve the problem of identifying undervalued cities where the quality of life exceeds what the cost of living would predict quantitatively. Specifically, we want to answer questions: which cities offer the highest return on capital spent for residents, is there a linear correlation between a city's safety index and its rent prices, and which regions currently offer the best trade-off between healthcare quality and cost in terms of city locations?

Data Collection

We will use Python (requests and BeautifulSoup) to scrape data from Numbeo.com, which is a global database of user-contributed data since it is available for scraping due to a lack of anti-scraping technology. We will focus on collecting data from the current year rankings to ensure the time relevance.

Data Collection Design

Source 1: Cost of Living Index Page
Variables: City Name, Cost of Living Index, Rent Index, Groceries Index, Restaurant Price Index.
Method: Scrape the main HTML table containing data for 285 cities.
Source 2: Quality of Life Index Page
Data Points: Safety Index, Health Care Index, Pollution Index, Traffic Commute Time Index.
Method: Scrape the corresponding Quality of Life tables, merge them with the Cost of Living data based on City Name. We will normalize City Name to avoid mismatches.

Analysis and Visualization

Analysis

We will create a value metric by creating a custom Quality to Cost Score calculated as Quality of Life Index / Cost of Living Index. We will compute the correlation coefficient between Rent Index and Safety Index to see if paying more rent guarantees safety. Identify cities that are statistical outliers such as the ones with very high cost index but low safety index or the opposite.

Visualization

We will use Matplotlib and Seaborn to create a scatter plot plotting Cost of Living vs. Quality of Life. We will add a regression line; cities significantly above the line will be highlighted as undervalued. We will also create a bar chart showing Top 10 Cities by Quality to Cost Score. We will also create a heatmap showing how different factors interact with each other using a correlation matrix.