

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

Final Report

Short Description

Today, since there are more and more job opportunities offering individuals the right to choose between working from home and in office, digital nomads are a new group of people who may concern about where they should stay at home to work and to have a comparably higher quality of life. The primary goal of this project is to identify undervalued cities where the quality of life significantly exceeds what the cost of living would quantitatively predict. By treating a city's cost as an investment and its quality of life as a return, this research calculates a Geo-Arbitrage Score to answer which global cities offer the highest return on capital for residents and to analyze how this trade-off varies across different geographic regions.

Data

The data for this project was collected by scraping Numbeo.com, a global database of user-contributed data regarding consumer prices and living conditions. Specifically, the project utilized Python scripts to scrape two distinct datasets which are correspondingly the Cost of Living Index table and the Quality of Life Index table for the current year. The data collection process involved iterating through the main ranking tables to extract indices for rent, groceries, safety, healthcare, and pollution. The initial scraping targeted the top 500 cities listed on the platform. After cleaning, merging, and filtering for complete records, meaning dropping city subjects with none values, the final dataset comprised 285 unique cities with full metric coverage.

Data Cleaning, Analysis, and Visualization

The data processing pipeline began with cleaning the raw scraped data. The primary challenge involved merging the Cost of Living dataset with the Quality of Life dataset. Since both datasets contained a City column and a Cost of Living Index column, it leads to potential duplication and naming conflicts. The cleaning script addressed this by normalizing city names to lowercase and stripping whitespace to ensure accurate matching, while we also drops redundant columns and empty Rank identifiers after the merge through column selections. After the final dataframe was preliminarily completed, a custom metric (variable/column) named the Geo-Arbitrage Score was calculated by dividing the Quality of Life Index by the Cost of Living Index and added to the final dataframe.

The analysis phase focused on three key areas including correlation, regional segmentation, and outlier detection. First, a correlation analysis was performed between the Rent Index and the Safety Index to test the premise that higher housing costs guarantee safety. The results yielded a correlation coefficient of approximately 0.0489, indicating a weak positive relationship. In other words, while higher rent often accompanies higher safety, the relationship is too weak to imply any causation or even strong connection, and it also suggests that expensive cities are not

automatically safer. Second, a regional analysis was conducted by grouping cities into continents. The findings revealed that Asia and Africa offer the highest average Geo-Arbitrage Scores (4.41 and 3.77 respectively), while Europe and North America lag behind with significantly lower average scores. The potential reason for this result might be that in the formula of the Geo-Arbitrage Score, the simple division might offer indices with inadequate weight while the calculation of indices themselves in the website might have distortions due to index normalization. Finally, outlier analysis using regression residuals identified specific cities that deviate from the market trend. Hyderabad, Muscat, Valencia, The Hague, and Plzen were identified as the top undervalued cities, that offer exceptional quality of life relative to their cost. Conversely, cities such as New York, Manila, Colombo, Caracas, and Beirut were identified as overvalued that present low quality of life despite their respective costs.

Visualization was executed using the Matplotlib and Seaborn libraries to communicate these findings effectively. A scatter plot illustrating Cost of Living vs. Quality of Life was generated. It features a red regression line to demarcate the average market trend. Cities situated significantly above this line represent high-value locations. Additionally, a bar chart was created to display the top 10 cities ranked by their Geo-Arbitrage Score, visually highlighting the dominance of Asian cities in the value rankings. Lastly, a correlation heatmap was produced to visualize the interdependencies between variables. It again shows that Safety index does not have correlation with either the Cost of Living index or Rent index. Surprisingly, the Geo-Arbitrage Score we generated shows different levels of correlations to the Cost of Living index and the Quality of Life index, which implies that we should tune the formula to reflect the weight of these two indices fairly.

Changes from Original Proposal

The project largely adhered to the original proposal, but specific technical adjustments were required during the implementation phase. The most significant deviation involved the data merging strategy. The original plan underestimated the difficulty of aligning city names between two separate web pages due to minor formatting differences and duplicate column names provided by the source website. To address this challenge, the cleaning pipeline was enhanced to strictly enforce string normalization and explicitly remove duplicate indices prior to merging, ensuring data integrity. Furthermore, while the original proposal focused heavily on city-level analysis, the final implementation expanded the scope to include a regional analysis. This addition was necessary to provide a more macroeconomic context to the findings and allow for broader conclusions about geographic desirability.

Future Work

Given more time and resources, this project could be significantly expanded by incorporating a temporal dimension. Since currently, the analysis is a snapshot of the present year, which is limited in temporal perspective, collecting historical data over the past decade would allow for the identification of trends, such as which cities are improving their value proposition over time. This project could also segment regions more accurately by eliminating the Other values. Also. Since we gather only user-contributed data, the accuracy of the study is questionable, and we could find

better sources to scrape and fetch data that can reflect the reality better. Besides, we could look into the dubious outliers more carefully to handle them appropriately case by case.

Additionally, the Cost of Living index, the Quality of Life index, and Geo-Arbitrage Score could be refined by redesigning formulas since the current results show that the Cost of Living and the Quality of Life indices are distorted due to normalization. For example, cities with bad safety index and low living cost can be normalized to a decent city while calculating the Quality of Life, and they can easily receive a high Geo-Arbitrage Score due to the low Cost of Living Index. The indices can also be improved by integrating external data, such as average internet speeds or weather consistency, which are critical factors for the target demographic of digital nomads. Analyzing these additional variables would create a more holistic index for remote work feasibility.