The Battle of Housing Affordability: A Comparison of Los Angeles and San Francisco

Anandita Juneja

Ajaiy Praveen Thiruchelvam

Lifu ‘Mario’ Ma

1. Executive Summary

The article explores the issue of housing affordability in two of California's largest cities, Los Angeles and San Francisco. To determine which city is more expensive to live in, the article considers various factors such as the type of house, number of bedrooms and bathrooms, location, and neighborhood population. In addition, the article defines affordability based on the expense of purchasing a home relative to the median household income in that area.

To obtain the data necessary for the analysis, the authors turned to Trulia, an online real estate marketplace widely used in the United States. The authors explain that they chose Trulia due to its comprehensive data on housing, including price, number of bedrooms and bathrooms, property type, and address. They also noted that Trulia is a popular website for people searching for homes, making it an excellent source for representative data.

To extract the data, the authors utilized the Beautiful Soup package in Python, a library that simplifies the process of scraping information from web pages. They also used the Position Stack API and Precisely API to obtain latitude and longitude data and determine neighborhoods based on that data.

2. Background and Domain Knowledge

Los Angeles and San Francisco are two of California's largest cities, with millions of people living and working in both. However, despite the countless opportunities for access to top-level jobs and global culture, the house prices in these two cities have become unaffordable over time. After the pandemic, the house price of these two cities also changed significantly. Now, which city is more expensive to live in, and which city is more affordable? These are the two questions we want to ask and answer through our dataset.

First, which city is more expensive to live in? There are many dimensions to measure the price of a house. To be more specific, usually, there can be some characteristics of a house itself, such as the type of the house, the number of bedrooms, the number of bathrooms, the area of the house etc. There can also be some external factors, such as the location of the house and the population of the neighborhood of the house. Therefore, to capture the real value of a house, we need to consider both its internal and external factors.

Second, which city is more affordable? Affordability is always an interesting topic discussed by scholars. By referring to the idea of Chris Herbert and Daniel McCue [6], to determine whether a house is affordable, we can calculate the expense we need to pay for the house over the median household income people earn in that district. If the calculation result is less than 30%, then we can say that the house is affordable for people who live there. Therefore, to measure affordability, we also need to capture the median household income information of people in the same district, calculate the annual house spending based on the house price, and compute the ratio of the spending on housing.

Now we've defined the problem and collected enough domain information. Let's start creating the dataset!

3. Introduction to Data

In order to perform our analysis of housing prices based on the city ,we have fetched the data from Trulia website.[1]Trulia is an American online real estate marketplace which is a subsidiary of Zillow which is also an American tech real-estate marketplace company . It facilitates buyers and renters to find homes and neighborhoods across the United States through recommendations, local insights, and map overlays that offer details on a commute, schools, churches and nearby business.[2]

3.1 Why did we choose Trulia for Web Scraping?

Our decision to choose Trulia for web scraping was based on several factors.

1. The data contains comprehensive information on houses, including price, number of bedrooms and bathrooms, property type, and address. This level of detail allows us to make informed decisions and derive valuable insights from the data.
2. It is one of the most widely used websites in the United States for searching homes, particularly among new students relocating from another city or country. This popularity ensures that we have a sizable and representative sample of data to work with, which is essential for achieving reliable and meaningful results.
3. Lastly, the absence of a bot is a significant advantage that we experienced while scraping data from Trulia, as opposed to other websites. The absence of a bot ensures that our data is accurate and not skewed by automated activity, which can be a problem with other websites.

3.2 Sources

1. To commence the process of web scraping, we obtained data from Trulia[1], extracting critical information such as price, number of bedrooms, number of bathrooms, total square footage, and address.

2. To extract the required data, we utilized the Beautiful Soup package in Python, a library that simplifies the process of scraping information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree. [3]

3. To obtain latitude and longitude data, we utilized the Position Stack API, a reliable and straightforward solution for forward and reverse geocoding. This API covers over 2 billion places and addresses worldwide, making it an excellent choice for our purposes.To use this API, users are required to have a unique API key.[4]

4. To enhance our analysis, we also utilized the Precisely API, which helped determine the neighborhood based on the provided latitude and longitude data. To use this API, users are required to have a unique API Key and Secret Key. Additionally, this API is limited to processing a maximum of 500 requests per user.[5]

5. We collected data on each neighborhood's demographics, which allowed us to calculate the median income for each one.[7]

3.3 Description of the Web-Scraping Routine

We started the web scraping process by utilizing the Beautiful Soup library in Python to extract data from Trulia web pages.Our focus was primary on San francisco and Los Angeles based out in california united states to fetch the dataset because these are among the most popular cities in the country for individuals looking to buy or rent a property.We attempted to include cities such as New York, Houston, and Chicago in our dataset. However, we encountered limitations in the amount of data we could scrape from Trulia at once, which prevented us from obtaining data from these cities.

Each page on Trulia contained approximately 40 houses. To obtain the desired dataset, we iteratively downloaded the contents of pages 1-18 for each selected city and saved them locally. We used a naming convention that included the city name and page number, resulting in file names such as "truliaSanFrancisco1.html." This allowed us to easily organize and access the downloaded data for further analysis.

Using the downloaded pages, we iteratively extracted details for each house using the Beautiful Soup 4 (bs4) object. This process involved parsing the HTML content and selecting the relevant elements containing the required information, such as the price, number of bedrooms and bathrooms, total square footage, and address.

In order to obtain the pricing information for each house, we employed a web scraping technique. Specifically, we targeted the main “div” element on the webpage, identified by the “data-testid” as "home-card-sale", which corresponded to each property listed. From there, we were able to extract the necessary details such as the price, number of bedrooms, number of bathrooms, and property address. To do this, we fetched a “div” element with the attribute “data-testid” set to "property_feature", which allowed us to retrieve the relevant data for each house in a systematic and efficient manner. To clean and manipulate the data, we employed regex (regular expressions) to remove the dollar sign from the price and extract only the numeric values for the number of bathrooms, bedrooms, and square footage. Additionally, we addressed any empty elements on the page by assigning them a null value, ensuring consistency in the dataset.

We obtained the latitude and longitude coordinates for each property by passing its address to the Position Stack API. The API requires the address to be formatted with the corresponding region information. Subsequently, we utilized these coordinates to retrieve the corresponding neighborhood information for each property using the Precisely API.

The data obtained from the web scraping process was stored in both Excel format and in MongoDB. The data was appended to a list before being saved to both storage locations, ensuring that all relevant information was captured and preserved for further analysis. Additionally, we collected data on the median income for each neighborhood in our target cities by utilizing neighborhood-level demographic data. With this information, we were able to calculate the estimated one-year house payment for each zip code and determine the ratio of house payment over median income. We then added this data to MongoDB using the update command.

For our analysis aimed at determining which city is most affordable, we utilized a comprehensive dataset that provided a wealth of information. This dataset included information on median incomes for each neighborhood, as well as estimated one-year house payments and the ratio of house payment over median income for each zip code. By utilizing this comprehensive dataset, we were able to conduct a thorough analysis and arrive at a well-informed conclusion.

Variable	**Data Type**	**Example**
Address	String	1100 Sacramento St #402, San Francisco, CA 94108
City	String	San Francisco
Latitude	Double	37.778008
Longitude	Double	-122.431272
Price	Integer	5200000
Bed Room	Integer	3
Bath Room	Integer	4
sqft	Integer	2426
Neighbourhood	String	Western Addition
Population	Integer	44391
Number of Households	Integer	17953
Median Income	Integer	98803

Average Income	Integer	130769
avg_price_per_neighborhood	Double	1284830.737
house_payment_1yr	Double	56362.56

Fig 1: Sample Dataset

3.4 Database Design Choices

We opted to store our data in MongoDB, a NoSQL database that utilizes a document-based data model, with data stored in JSON-like documents. The flexibility, scalability, and high-performance features of MongoDB were the key drivers for our choice. With MongoDB, we have the ability to effortlessly incorporate additional data by adding new variables to the document, eliminating the need to modify the database structure. This approach contrasts with traditional relational databases, where restructuring the database would be required to add new data.

4. Business Values

We can apply our dataset in the following situations. Overall, a house pricing dataset locality-wise with the median income of the population can provide valuable insights for a variety of businesses and industries involved in the real estate market.

4.1 Real Estate Investment

A house pricing dataset locality wise with the average income of the population can be used by real estate investors to identify potentially profitable investment opportunities. For example, if a locality has relatively low house prices but high average incomes, this may indicate that the area is ripe for real estate development and could provide high returns on investment.

4.2 Targeted Marketing

Real estate agents can use a house pricing dataset locality wise with the average income of the population to target potential homebuyers who are likely to be interested in properties in specific areas. By understanding the income level and other demographic factors of potential homebuyers, agents can more effectively market properties and increase the likelihood of a successful sale.

4.3 Economic Analysis

Economic analysts can use a house pricing dataset locality wise with the average income of the population to analyze the housing market in different localities and identify trends and patterns. By understanding the relationship between housing prices and income levels in different areas, analysts can make informed predictions about future economic growth and development.

4.4 Urban Planning

City planners use a variety of data sources to make decisions about urban development, including housing affordability and economic growth. A house pricing dataset locality-wise with the average income of the population can help planners identify areas where affordable housing and economic growth could be encouraged, leading to more sustainable and equitable urban development.

5. Summary & Conclusion

Comprehensive data was scraped from Trulia regarding house details for two major cities in California, namely San Francisco and Los Angeles using Beautiful Soup. The geolocation was fetched by passing the address through an API and in return the geolocation was again passed into an API to fetch the neighborhood of the house. One limitation of the API was that with one key a maximum of 550 requests could be made. All the parameters were then inserted into a collection in MongoDB. This data can be used to estimate which city is cheaper or more affordable to live in.

Appendix

[1] https://www.trulia.com/

[2] Wikipedia

[3]https://pypi.org/project/beautifulsoup4/#:~:text=Beautiful%20Soup%20is%20a%20library,and%20modifying%20the%20parse%20tree.

[4] https://positionstack.com/documentation

[5] https://www.precisely.com/product/precisely-addresses/precisely-addresses

[6] IS THERE A BETTER WAY TO MEASURE HOUSING AFFORDABILITY?

https://www.jchs.harvard.edu/blog/is-there-a-better-way-to-measure-housing-affordability

[7] https://www.point2homes.com/US/Neighborhood/CA/San-Francisco-Demographics.html