Zip & Link: Exploring the Relationship between Housing Prices and Accessibility to Essential Services

By: Nelu Wijegunasekera, Pragya Khanal, Vyshnavi Voleti

Project Overview

Zip & Link analyzes the relationship between essential services and housing affordability across neighborhoods in Chicago. The project explores how accessibility to key services—such as healthcare, public education, public transport, grocery stores, and parks—affects median property prices and other economic indicators. The project involves web scraping, bulk data downloads, data cleaning, and an analysis framework to generate an Accessibility Index for each ZIP code. This index is meant to help users understand the impact of essential services on housing affordability and highlight areas that require better urban planning. Our initial hypothesis was that the housing prices were positively correlated with the Accessibility Index. In order to test this and better visualize the relationship between the Accessibility Index and the economic indicators of interest, we developed 3 key visualizations on Dash: a choropleth map, a scatterplot and horizontal bar plots that help compare 2 specific zip codes, and found that the converse of our hypothesis was true.

Methodology and Data Documentation

To achieve the goal of our project, we first needed to obtain the housing prices and other zip-code specific economic indicators across Chicago. Secondly, we needed to obtain data on the 5 key variables that were used to compute the Accessibility Index - healthcare services, public education, public transport, grocery stores, and parks. Lastly, we also needed population data to calculate the Accessibility Index which we will explain in detail later on. Please refer to the table below to understand how we obtained all the relevant data.

Variables	Data Source	Challenges/Gaps
Median Housing Prices Median Housing Costs Owner Occupied Housing Costs Renter Occupied Housing Costs Housing Cost as % of Income Unemployment Rate Poverty Levels	Web Scraped ZipAtlas ● https://zipatlas.com/us/il/chicago/zip-code-comparison/lowest-property-prices.htm ● https://zipatlas.com/us/il/chicago/zip-code-comparison/lowest-housing-costs.htm ● https://zipatlas.com/us/il/chicago/zip-code-comparison/highest-owner-occupied-housing-costs.htm ● https://zipatlas.com/us/il/chicago/zip-code-comparison/highest-renter-occupied-housing-costs.htm ● https://zipatlas.com/us/il/chicago/zip-code-comparison/highest-housing-cost-as-percentage-of-income.htm ● https://zipatlas.com/us/il/chicago/zip-code-comparison/highest-unemployment-rate.htm ● https://zipatlas.com/us/il/chicago/zip-code-comparison/highest-poverty.htm	For all the data that was scraped from ZipAtlas, there were only 54 Chicago City Zip Codes that had data on all the key variables we were looking at. Solution: We had to drop around 6 zip codes for the sake of our analysis. Output: 54 Zip Codes in the City of Chicago and 6 economic indicators

Hospitals	• Web Scraped https://cookcountysheriffil.gov/departments/c-c-s-p-d/cemeteries/hospitals-cook-county/	Could not find data with all the health centers and hospitals together. Thus, I had to separate the hospital data source and health center data source.
Community Health Centers	1. Web Scraped City of Chicago Health Facilities (pdf) 2. Exported bulk data from HRSA Find a Health Center	Challenges for Data 1: Had to convert PDF into Excel and preprocess the data (Used Tabula) Challenges combining 1 & 2: Had to account for duplicates across datasets and find ways to eliminate them → Record Linkage using jaro_winkler_similarity
Public Schools	• Scraped CPS for the list of schools in Chicago using an API.	None
Parks and Grocery Stores	• Exported CSV from City of Chicago open data repository for the count of parks and grocery stores.	None
Public Transit	• Found a pdf with the number of public transit stops aggregated by zip code in the Chicago area that was used at the Institute of Social Research, University of Michigan.	None
Population	• Exported data from https://www.illinois-demographics.com/zip_codes_by_population	The dataset originally contained values for zip codes across the entire state of Illinois. It was necessary to filter out the relevant zip codes specific to Chicago

Calculation of Accessibility Index

Once all of these datasets were obtained, we preprocessed them individually to get zip code and count of each essential service. These datasets were then joined with each other on Zip Code and combined into one comprehensive dataset with Zip Code, all housing and economic indicators, population, park_count, grocery_store_count, num_public_transit_stops, school_count and total_healthcare_services (given by the sum of hospitals and community health centers). These last 5 columns were then normalized (scaled from 0 to 1), summed up, and divided by the population of the zip code to get our Accessibility Index. This Accessibility Index was once again normalized to ensure that each zip code was given a score of 0-1. With this, we proceeded to visualize any relationships between this Accessibility Index and other indicators by building a choropleth map, scatter plot and bar charts on Dash.

Project Structure & Flow

Our project repository contains 4 main folders:

1. zip_link/cleaning analysis: encompasses the scraping scripts written to extract data from the sources, clean the raw data sitting in zip_link/data/raw, merge the datasets into one file and determine the Accessibility Index.

    - **zipatlas_data.py** is the main script that connects all of these functions and produces a final cleaned dataset
2. zip_link/data folder: divided into raw and preprocessed folders
    - data/raw: contains the raw downloaded files and web scrape outputs
    - data/preprocessed: contains the zip code and count for each zip code of the individual data sources. In this folder is also the final merged data file (**zipatlas_bulk_merge.py**) that was subsequently used to build the accessibility index and in the visualizations
3. zip_link/visualization folder contains the python script for all three visualizations (the choropleth map, scatter plot and the bar charts), the necessary shape files along with the boundaries (Boundaries_-_ZIP_Codes_20250222.csv) that was obtained from the City of Chicago open data repository. **merge_visualization.py** is the key file to run here
4. zip_link/tests folder contains all the relevant tests for our data collection, data reconciliation, and visualizations

Team Responsibilities

1. Data Collection and Cleaning:
    - Vyshnavi - ZipAtlas scraping and data processing, Bulk data (Parks, Public Transit and Grocery Store Data) processing, Community Health Center data scraping and processing
    - Pragya: Hospital data scraping and processing, Population data and processing
    - Nelu: Schools data API and processing
2. Data Reconciliation: Vyshnavi, Pragya, Nelu
3. Developing Accessibility Index: Pragya
4. Data Visualization:
    - Nelu: Choropleth Map (Key part of our visualization)
    - Vyshnavi: Scatterplot comparing economic indicators' and essential services' distribution against Accessibility Index
    - Pragya: Horizontal Bar Charts for zip code comparison

Final Thoughts

As mentioned at the start, we were anticipating a positive correlation between accessibility of essential services and median house prices. However, our visualizations (particularly the choropleth map) showed us that it was in fact areas of lower median house prices, higher unemployment and higher poverty levels that the accessibility index was higher. While we found this rather surprising, we realized that this could be because these neighbourhoods are already receiving enough public investments, whereas the more affluent neighbourhoods have less "essential" services such as **public** schools (as most of the children there may be sent to private schools instead), and **public** transit connectivity (as more people have their own private transport). Interestingly, we also learnt that there are a greater number of healthcare services downtown and in the south of Chicago, as compared to the North. Hence, helping explain these differences in accessibility.

However, there are still improvements that can be made for the accessibility index to be more holistic. For example, we can add a 'lifestyle' accessibility index that would consider the number of restaurants/fine-dining options available in the neighbourhoods as well as recreational facilities such as publicly owned tennis courts, ice rinks, football fields etc. Another can be used to map **affordable** housing with accessibility to essential services. Lastly, accessibility by means of travel/commute time can be another variation that can be examined by neighbourhood. Perhaps, these approaches would result in more intuitive findings.