**How the Other Half Chooses to Live: A Two-Level Analysis on Housing Prices Drivers in U.S. High-Poverty Neighborhoods**

**MACS30112 Final Project Progress Report**

**1. Project info**

* Team Name: UMai
* Team Members:
  + Kuang Sheng ([ksheng@uchicago.edu](mailto:ksheng@uchicago.edu))
  + Tian Lan ([lant1@uchicago.edu](mailto:lant1@uchicago.edu))
  + Xuewei Li ([xueweili@uchicago.edu](mailto:xueweili@uchicago.edu))
* GitHub Repository Link: <https://github.com/macs30112-winter24/final-project-umai>

**2. Project Description**

**A. Research questions & Relevance**

Affordable housing is a critical social determinant impacting health, education, and individuals’ social well-being. Research shows that both **household circumstances**, like income (Acolin, Goodman, & Wachter, 2018) and housing quality (Kain & Quigley, 1970), and **neighborhood attributes**, such as attached social resources (Kane, Riegg, & Staiger, 2006) and racial factors (Daniels, 1975) significantly influence housing prices. Yet, residents of high-poverty neighborhoods often face a unique challenge: disproportionate housing costs and severely limited housing options. Our project investigates which housing attributes and social resources are most valued within these disadvantaged neighborhoods. The results will help identify key housing affordability drivers and support evidence-based intervention policies to promote fairer and more equitable living communities.

We perform a two-level analysis, at the city level and the inter-city level, in high-poverty neighborhoods in New York City, City of Los Angeles, and Chicago. The three cities contain the highest number of persistently poor neighborhoods over the decades (Benzow & Fikri, 2022), which makes them the most appropriate cases for observing long-lasting patterns in how the U.S. low-income populations make housing decisions. Specifically, we ask: What are the key drivers of housing prices in U.S. high-poverty neighborhoods?

Our project speaks to our computational social scientist peers in two ways. First, affordable housing is a long-standing, complex issue in social science research. Our two-level analysis distinguishes the separate impact of house and neighborhood attributes on housing prices. This distinction offers policymakers clearer insights to inform solutions for the problem. Specifically, at the city level analysis, we utilize regression models to find out how housing attributes and some neighborhood attributes relate to individual property prices in high-poverty neighborhoods within the same city. At the inter-city level, we employ the same method to further study the role of neighborhood-level demographics (i.e. median age, dominant ethnic group, and population density) in this relationship across the three cities. Overall, four groups of regression models are used for our quantitative analysis. For the first three models, the level of analysis is each individual household. For the last model, it is each high-poverty neighborhood in one of the three cities.

Methodologically, we combine computational power with quantitative rigor and data visualization to analyze extensive housing data. We hope to inspire computational social scientists working with big data and map visualization to explore robust data processing and mapping tools for tackling large-scale research questions. Our data processing pipeline handles over 2,117 property listings across high-poverty neighborhoods, together with their neighborhood attributes extracted from official public sources.

**B. Concepts & Operationalization**

In Table 1 and Table 2, we introduce key concepts in the project and how we plan to operationalize them.

**Table 1. Main Concepts**

| **Concept** | **Definition** | **Source** |
| --- | --- | --- |
| Housing price | The market value of a listed property in the housing market. |  |
| High-poverty neighborhoods | Neighborhoods with poverty rate over 20%. | U.S. Department of Commerce, Bureau of the Census’s Official Poverty Measure[[1]](#footnote-0) |
| Housing attributes | Dwelling-unit quality; Interior quality; Exterior quality; Unit size | The Value of Housing Attributes (Kain & Quigley, 1970)[[2]](#footnote-1) |
| Neighborhood attributes | Socioeconomic characteristics of the neighborhood; Quality of some critical public services; Accessibility to the CBD |

**Table 2. Concept Operationalization**

| **Variable** | **Operationalization** | **Source(s)** |
| --- | --- | --- |
| **Dependent variable** | | |
| Housing price | The asking price of property listings in neighborhoods with over 20% poverty rate in New York City, City of Los Angeles, and Chicago on Redfin. | Redfin |
| **Independent variables** | | |
| Housing Attributes | | |
| Property size | The total land area in square feet is included in the designated parcel of the real estate listing. | Redfin |
| Property age | The number of years since the initial construction of the property. |
| Property type | A categorical variable reflecting the property structure and ownership arrangements attached to a property (eg. condominium, townhouse, etc.). |
| Neighborhood Attributes | | |
| A. Socioeconomic characteristics | | |
| Population density | The total number of residents living within a square mile of a designated neighborhood boundary. | City-Data |
| Median age | The median age of all residents within a neighborhood. |
| Dominant race | The most prevalent racial group within a neighborhood. |
| B. Quality of critical public services | | |
| Public transit (train) | Distance from a property to the nearest train or subway station. | Raw data from local government websites; further variable construction by team members |
| Public transit (bus stop) | The number of bus stops located within one kilometer of a property. |
| Hospitals | The distance from a property to the nearest hospital. |
| Park | The distance from a property to the closest publicly accessible park. |
| C. Accessibility to the CBD | | |
| Commercial areas | The number of food and beverage stores of a property within one kilometer. | Same as above |

\**Note*: We aggregate values of numeric variables and use their median values for neighborhood-level comparison. For example, the housing price assigned to a given neighborhood would be the median of all housing prices within that neighborhood. For categorical and ordinal variables, such as property type, we calculate the frequency of each value, determine the mode, and identify the most common property type.

**C. Hypotheses**

At the inter-city level, where the unit of analysis is each neighborhood in one of the three cities, we posit that neighborhood attributes, particularly the availability of social capital and community resources, are the primary drivers of housing prices in high-poverty neighborhoods.

We expect to see more unexpected findings from the complex interplay between variables at the city-level analyses, where the unit of analysis is each individual household.

**3. Data Sources**

1. [Redfin](https://www.redfin.com/) - A real estate company providing residential real estate brokerage and mortgage origination services. Since Redfin no longer permits web-scraping on their data, we found a feasible workaround. With a limitation of 350 recordsdownloaded at a time, we used housing price as a controlling factor to download separate CSV data files from the website. Then we ran a Python code and used Pandas Dataframe to combine CSV file contents into one single dataset. By doing this, we ensure that we collect all the listing properties in the city on the day of collection and avoid possible duplicate data. Due to the nature of Redfin, we can only collect the listings on the day of collection. No historical data can be retrieved.
2. [City of Chicago](https://data.cityofchicago.org/) - The city of Chicago provides a data portal for infrastructural spatial data access. We will download shapefiles from this portal. Shapefile is the most popular data type for GIS software and analysis. Shapefiles can also be transformed from and to CSV files for more flexible data analysis. Chicago’s data portal provides access to different layers updated at different time points.
3. [NYC Open Data](https://opendata.cityofnewyork.us/) - Open Data is free public data published by New York City agencies and other partners.
4. [City of Los Angeles GeoHub](https://geohub.lacity.org/) - The GeoHub is the City's public platform for exploring, visualizing, and downloading location-based Open Data.
5. [City-Data](https://www.city-data.com/) - Free and open-source information website presenting annually updated information pertaining to U.S. cities, including detailed demographic data by neighborhood in New York City, City of Los Angeles, and Chicago. We plan to scrape population density, dominant race, and median age of high-poverty neighborhoods from the website.
6. [NYU Furman Center](https://furmancenter.org/neighborhoods), [Poverty by city and community (arcgis)](https://www.arcgis.com/apps/dashboards/7846c3c37dff4728923609a9f55f849c), & [Illinois Policy](https://www.illinoispolicy.org/black-brown-chicago-neighborhoods-endure-highest-poverty-rates/) - Three federal/local-government-backed platforms with poverty rate in the three cities by neighborhood.

**4. Data cleaning/wrangling**

**A. Data cleaning**

Some information about housing attributes, such as the longitude and latitude of a property, are rightly omitted on Redfin due to safety concerns. We plan to remove all houses without complete house attributes information, which is essential to the project. By the same logic, we also plan to remove reported crimes with neighborhood/location information left blank.

**B. Data wrangling**

At the household level analysis, the neighborhoods of interest are high-poverty neighborhoods. To get data on the targeted neighborhoods, we plan to link the dataset containing housing attributes with the dataset containing neighborhood poverty rates by establishing a relational database. Specifically, we have a dataset with housing prices and housing attributes. We have another dataset with the names of neighborhoods in the three cities with their poverty rate. We then join the two datasets on common keys (i.e. the name of a neighborhood and a city) and find all neighborhoods with a poverty rate of over 20%. By doing so, we obtain the dataset we need to perform a housing price comparison at the household level.

We build on the above dataset for between-neighborhood level analysis. We plan to aggregate values of numeric variables for each property listing and use their median values. For example, the housing price assigned to a given neighborhood would be the median of all housing prices within that neighborhood. For categorical and ordinal variables, such as property type, we calculate the frequency of each value, determine the mode, and identify the most common property type. After completing the calculation process, we will join a dataset containing socioeconomic attributes of qualified neighborhoods to the processed dataset on the same keys - the name of a neighborhood and a city. By doing so, we obtain the dataset crucial for performing a housing price comparison at the neighborhood level.

**5. Data analysis and visualization**

Our two-level analysis will focus on both household level and neighborhood level. The construction of one group of variables, the Quality of Critical Public Services, deserves special notice. We plan to incorporate QGIS software and Python coding by utilizing the function calculating the distance between two points in PA2 to construct this group of variables. For other groups of variables, please refer to Table 1 and Table 2 for detailed operationalization.

For both levels of analysis, we plan to conduct hedonic regression models, with housing prices for an individual household or the median of housing prices within a neighborhood as the dependent variable, and housing as well as neighborhood attributes as independent variables.

For the final result, we plan to utilize the mapping and data visualization capabilities empowered by QGIS to visualize the linear relation between different independent factors and local housing prices at both levels. Color-coded polygons of neighborhoods with high poverty rates can visualize the price distribution within and across neighborhoods. At the same time, with an overlay of different factors shapefiles on the same map, our final deliverable can present insightful findings with the locational proximity and factors that influence housing prices the most.

**6. Responsibilities & Timeline**

| **Progress** | **In charge** | **Timeline (Feb & March)** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | 11 | 18 | **21** | **22** | **23** | **24** | **25** | **26** | **27** | **28** | **29** | **3/1** |
| Data Collection | All | ccc | ccc |  |  |  |  |  |  |  |  |  |  |
| Data Cleaning & Matching | All | ccc | ccc |  |  |  |  |  |  |  |  |  |  |
| Data Analysis | Tian Lan |  |  | ccc | ccc | ccc | ccc |  |  |  |  |  |  |
| Data Visualization | Kuang Sheng |  |  |  |  | ccc | ccc | ccc | ccc |  |  |  |  |
| Presentation | All |  |  |  |  |  |  | ccc | ccc | ccc | ccc |  |  |
| Final Report | Xuewei Li |  |  |  |  |  |  |  |  |  |  | ccc | ccc |
| Short video | All |  |  |  |  |  |  |  |  |  |  | ccc | ccc |

**References**

Acolin, A., Goodman, L., & Wachter, S. M. (2018). Accessing homeownership with credit constraints. *Housing Policy Debate*, *29*(1), 108–125. https://doi.org/10.1080/10511482.2018.1452042

Daniels, C. B. (1975). The influence of racial segregation on housing prices. *Journal of Urban Economics*, *2*(2), 105–122. https://doi.org/10.1016/0094-1190(75)90042-x

Kain, J. F., & Quigley, J. M. (1970). Measuring the value of housing quality. *Journal of the American Statistical Association*, *65*(330), 532–548. https://doi.org/10.1080/01621459.1970.10481102

Kain, J. F., & Quigley, J. M. (n.d.). The value of housing attributes. Retrieved February 10, 2024, from NBER website: https://www.nber.org/books-and-chapters/housing-markets-and-racial-discrimination-microeconomic-analysis/value-housing-attributes

Kane, T. J. (2006). School quality, neighborhoods, and housing prices. *American Law and Economics Review*, *8*(2), 183–212. https://doi.org/10.1093/aler/ahl007

Provenzano, J. (2022, March 23). Neighborhood poverty project - The persistence. Retrieved February 10, 2024, from Economic Innovation Group website: https://eig.org/neighborhood-poverty-project/neighborhood-turnaround/

USDA ERS. (n.d.). Retrieved February 10, 2024, from Background and Uses website: https://www.ers.usda.gov/data-products/poverty-area-measures/background-and-uses/

1. For full definition of high-poverty neighborhoods, see: <https://www.ers.usda.gov/data-products/poverty-area-measures/background-and-uses/> [↑](#footnote-ref-0)
2. For a full list of housing attributes and respective explanations, see: <https://www.nber.org/books-and-chapters/housing-markets-and-racial-discrimination-microeconomic-analysis/value-housing-attributes> [↑](#footnote-ref-1)