**Final Project Progress Report 1**

# **Project Description**

## *Research Questions and Evolution*

This project investigates the evolution of gentrification in Chicago from 2018 to 2023. Our primary research question is: How has gentrification in Chicago evolved from 2018 to 2023, and what socio-economic factors contributed to these changes? We particularly focus on housing market dynamics (especially price-to-rent ratios), demographic shifts, and infrastructural changes.

The methodological approach has evolved significantly from our initial proposal. Rather than attempting to replicate the Urban Displacement Project's expert-derived typology for 2023, we now employ machine learning to:

1. Learn the underlying patterns in the 2018 classifications
2. Apply these learned patterns to 2023 data
3. Identify neighborhoods where status changes occurred
4. Analyze the factors associated with these changes

This methodological shift occurred because we recognized that designing new categorization rules might not capture the nuanced expertise embedded in the original [UDP framework](https://www.urbandisplacement.org/maps/chicago-gentrification-and-displacement/). Machine learning allows us to identify both explicit and implicit patterns in neighborhood change while maintaining the rigor of the original classifications.

## *Social Science Relevance*

This research contributes to urban sociology and housing studies in several ways. First, it addresses a methodological gap in gentrification research by using machine learning to capture complex patterns of neighborhood change that might be overlooked by traditional typological frameworks. Second, it examines how housing market dynamics, particularly price-to-rent ratios, interact with demographic and infrastructural changes to drive gentrification. This approach aligns with recent scholarship emphasizing the role of housing markets in neighborhood transformation (e.g., studies by urban economists on housing affordability and displacement).

# **Key Concepts and Operationalization**

## *Gentrification Status*

Our study operationalizes gentrification status using the baseline classifications from the 2018 Urban Displacement Project typology. To capture these classifications through machine learning, we incorporate features from multiple temporal and categorical dimensions. The historical foundation comes from Census data spanning 1990 and 2000, providing long-term demographic and economic trajectories for each neighborhood. Contemporary indicators are drawn from the American Community Survey, encompassing population demographics including racial and ethnic composition, as well as educational attainment levels across communities.

## *Housing Bubble (Measured by Price-to-Rent Ratio)*

Housing market dynamics are measured through a comprehensive set of metrics scraped from Zillow databases. These include home values, rental prices, housing unit density measures, etc. Those data together capture the economic pressures and housing market changes that often characterize gentrifying areas. The temporal aspect of these housing metrics allows us to track both gradual and sudden shifts in neighborhood housing markets.

## *Neighborhood Context*

The neighborhood contextis captured through several geographic and infrastructural measures. We aim to find historical demographic shifts, infrastructural improvements and development projects to conduct qualitative analysis in understanding their potential influence on neighborhood change.

Each of these dimensions - contemporary population characteristics, housing market dynamics, and infrastructural context - serves as input features for our machine learning model. By training on the 2018 UDP classifications, the model learns to recognize the complex interactions between these various factors that correspond to different gentrification states. When applied to 2023 data, this same set of features allows us to identify how neighborhoods' statuses may have shifted over the five-year period.

# **Hypotheses**

1. Housing Market Hypothesis:

Neighborhoods with rapidly increasing price-to-rent ratios between 2018-2023 will show higher likelihood of transitioning to more advanced stages of gentrification.

1. Spatial Diffusion Hypothesis:

Gentrification status changes will show spatial clustering, with neighborhoods adjacent to previously gentrified areas more likely to experience status changes.

1. Infrastructure-Led Change Hypothesis:

Areas experiencing significant infrastructure improvements or increased transit accessibility between 2018-2023 will show higher probabilities of status change, particularly in neighborhoods closer to downtown Chicago.

1. Market Resistance Hypothesis:

Some neighborhoods with strong community organizations or protective housing policies will maintain stable classifications despite experiencing housing market pressures typical of gentrifying areas.

# **Data**

## *Data Source*

| Data Source | Collection Method | Link | Time Frame | Dataset Size | Reliability and validity Issues |
| --- | --- | --- | --- | --- | --- |
| Zillow | Scraping | https://github.com/yiningYvette/Dollar\_Hub/tree/main/zillow-listing-data | January 2025 | Rent data: ~ 4,700  Selling price data:  ~ 4,700 | Zillow data was obtained from a third-party API (<https://rapidapi.com/s.mahmoud97/api/zillow56>). Due to Zillow restrictions, we were unable to retrieve all lateset listings. However, through tuning our search parameters, we were able to obtain as much search results as possible. |
| Public Housing Buildings Data | Downloading | https://github.com/yiningYvette/Dollar\_Hub/tree/main/data |  | ~ 3,000 | This dataset is downloaded directly from <https://hudgis-hud.opendata.arcgis.com/datasets/public-housing-buildings>. We use the dataset to cross-reference scraped data from Zillow, ensuring the reliability of our estimation for the average regional housing price/rent. |
| Census | API | https://github.com/yiningYvette/Dollar\_Hub/tree/main/data | 2019-2024 | 1,901 | This data set is scrapped from the US census by using API. It is used to generate demographic data. |
| PUMS Data | downloading | https://github.com/yiningYvette/Dollar\_Hub/tree/main/data | 2019-2023 | ~8,5000 | This data is directly downloaded from <https://data2.nhgis.org/main>. It is used to generate population, housing, agriculture, and economic data along with a GIS file. |
| LIHTC Properties Data: | downloading | https://github.com/yiningYvette/Dollar\_Hub/tree/main/data | 2022-2023 | 1,076 | This data is download directly from <https://hudgis-hud.opendata.arcgis.com/>. Its full name is Low-Income Housing Tax Credit Properties |
| Transit Data | downloading | https://github.com/yiningYvette/Dollar\_Hub/tree/main/data |  | 444 | Directly download from <https://toddata.cnt.org/downloads.php>.  It contained all transport station in Chicago |
| ZIP Codes to Census Tracts Crosswalk | downloading | https://github.com/yiningYvette/Dollar\_Hub/tree/main/data | 2014 and 2022 | 1,049 | Directly download from <https://mcdc.missouri.edu/applications/geocorr2022.html>. It crosswalk zip code and tact code in year 2014 and 2022 |
| Census history | downloading | https://github.com/yiningYvette/Dollar\_Hub/tree/main/data | 1990 and 2000 | Both 1984 | Directly download from Github <https://github.com/urban-displacement/displacement-typologies/blob/main/data/outputs/downloads/Chicagocensus_00_2017.csv>. It is the historical demographic data from the US census. |
| Chicago Geospatial Data | downloading | https://github.com/yiningYvette/Dollar\_Hub/tree/main/zillow-listing-data | As of Jan 29, 2025 | 61 (by zipcode) | The data is directly retrieved from <https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-ZIP-Codes/unjd-c2ca/about_data>, which is a open data source released by Chicago Data Portal. The data is continuously updated by local authorities, which ensures its reliability. |

## 

## **Data Cleaning and Wrangling**

1. Census Data (2018-2023 ACS)

We aim to process ACS data (2018–2023) for Chicago census tracts by standardizing geographic identifiers, converting tract numbers to FIPS codes for consistency. After this, we will clean the dataset by addressing missing values and outliers. Finally, we will ensure uniform tract boundaries to maintain comparability across years.

1. Housing Market Data

To analyze Chicago’s housing market, we will first clean Zillow’s housing sale price and rent data by removing duplicates and empty values. Next, we will merge this data with Chicago geospatial datasets to generate heatmaps for exploratory analysis of housing price/rent.

1. Chicago Geospatial Data

For geospatial integration, we will combine Chicago’s spatial data with both housing market and census data, allowing us to create visual heat maps that effectively present our findings.

1. Infrastructure Data

Infrastructure data will also be processed to ensure consistency. We will clean CTA ridership data, standardize building permit records, and align location information with census geographies for accurate analysis.

1. Merge Operations

Finally, we will merge these datasets to create a comprehensive view of Chicago’s urban landscape. Census data will be joined with housing market data using ZIP codes, while infrastructure data will be integrated through spatial joins. The merged datasets will then be validated for completeness, ensuring duplicate or conflicting records are properly addressed.

## **Data Analysis and Visualization**

### *Variables*

Our analysis relies on a set of **features** that capture demographic shifts, housing market dynamics, and infrastructural developments to predict the **outcome variable**, which is the gentrification classification of each census tract.

#### ***Outcome Variable:*** *Gentrification Classification (Categorical, 9 classes):*

#### Based on the 2018 Urban Displacement Project (UDP) typology

#### Predicted for 2023 using machine learning models

#### ***Independent Variables (features)****:*

1. **Demographic Variables** (From ACS and Census data)

* Median household income (2018, 2023)
* Racial and ethnic composition
* Educational attainment levels
* Population change rate (2018-2023)

1. **Housing Market Indicators** (From Zillow, property records)

* Median (or mean) home value (2018, 2023), median (or mean) rent (2018, 2023), and rice-to-rent ratio
* Homeownership rates vs. rental occupancy
* Absolute loss of low-income households (2018-2023)
* Home affordability

1. **Infrastructural and Spatial Features** (From transit data, building permits, development projects) *(May not be used)*

* Distance to downtown
* Public transit accessibility (CTA ridership changes)
* New construction permits and development projects
* Proximity to major infrastructure improvements

1. **Neighborhood Change and Gentrification History**

* Prior gentrification status (1990-2000, 2000-2018)
* Rent gap (difference between local and regional median rent)
* Changes in community organization presence (proxy: LIHTC data, public housing buildings)

### **Data Analysis Methods**

We will use **descriptive statistics, machine learning models, and spatial analyses** to assess the evolution of gentrification in Chicago.

#### *1. Exploratory Data Analysis*

* Summarize key demographic and housing market trends across different neighborhoods.
* Visualize distributions of home values, rent increases, and population shifts.
* Identify outliers and trends in infrastructure investments.

#### *2. Machine Learning Classification Model*

We will train a supervised learning model to predict gentrification status for 2023 using 2018 data as the training set. The classification models under consideration include:

* **Random Forest Classifier** (interpretable, handles non-linearity well)
* **Gradient Boosting (XGBoost, LightGBM)** (high accuracy, feature importance insights)

#### 

#### *3. Spatial Analysis and Geographic Visualization*

* **Choropleth Maps**: Visualizing gentrification status changes across Chicago.
* **Heatmaps**: Identifying clusters of high housing price increases.
* **Spatial Clustering**: Using Moran’s I to detect spatial dependencies in gentrification patterns.

#### *4. Hypothesis Testing and Regression Analysis*

To validate our hypotheses, we will employ:

* **Machine Learning Models**: We will use supervised learning techniques to predict neighborhood transitions in gentrification stages based on economic, housing, and transit-related features.
* **Spatial Clustering and Spatial Lag Models**: These methods will help assess whether gentrification status changes exhibit spatial dependence.

### **Expected Visualizations**

* **Gentrification Change Map (2018 vs. 2023)**: Census tracts color-coded by classification shifts.
* **Price-to-Rent Ratio Change (2018-2023)**: Heatmap indicating high-increase areas.
* **Transit and Development Overlay Maps**: Examining infrastructure’s impact on gentrification trends.
* **Income and Rent Scatterplots**: Identifying correlations between income growth and rent increases.
* **Feature Importance Chart (Machine Learning Model)**: Highlighting the most influential variables in predicting gentrification.

By employing these analytical techniques and visualization methods, we aim to identify communities in Chicago that experienced significant changes in terms of gentrification/de-gentrification. Furthermore, the study will attempt to uncover reasons behind such changes, through literature review and focusing on key socio-economic drivers of gentrification in Chicago between 2018 and 2023.