U.S. CITY NEIGHBORHOOD ARCHETYPES

Team 44 - Kriti Agrawal, Vivek Prakash, Tanmayee Kolli, Yuan Jack Yao, Sindhu Panthangi, Nikolaos Kakonas

1. Introduction

The purpose of our project is to conduct a broad socioeconomic analysis of U.S. city neighborhoods. There is a need to deeply understand how socioeconomic and demographic patterns in U.S. neighborhoods are changing over time. Research shows that city planners and policymakers increasingly rely on advanced modeling tools to evaluate housing, zoning, and community development decisions. This creates an opportunity to use analytical frameworks to highlight long-term neighborhood shifts and to advocate for underserved and historically marginalized communities across the country.

2. Problem Definition

We propose to build a census analysis map of the USA to discover its neighborhood archetypes and visualize these dynamic clusters on a map by integrating socioeconomic and demographic datasets.

Given a dataset with rows identified by their census tract codes and year:

- identify which cluster they belong to based on the selected 21 features.
- forecast feature values from 2030 - 2070 for each census tract.
- cluster future neighborhoods and visualize how they evolve over the time.

3. Literature Survey

We build on prior research from Bell[1965], which uses traditional models to create similar archetypes, and Delmelle et al., [2017], which uses data from the 80s to classify and forecast neighborhood shifts by using a larger data set and modern machine learning methods.

Most existing studies focus on limited neighborhood features or specific cities. For example, Spielman et al., [2008] mapped New York neighborhoods using 79 indicators, but lacked clear cluster labels. Wo et al., [2025] linked feature relationships in Los Angeles but examined only two: crime and greenspaces. Lynge et al., [2022] grouped South African cities with demographic data using a fixed K-Means approach. Overall, data remains fragmented across themes and regions, and no study has integrated social, demographic, and infrastructural layers. We aim to unify these fragmented data, use multiple indicators and clustering methods to get a comprehensive understanding of the neighborhood patterns in the major USA cities.

Current existing literature uses clustering techniques such as DBSCAN to improve upon OpenStreetMap data aggregations, giving access to more precise data (Fang et al., [2025]). Other papers build upon existing ML models that predict gentrification, such as in Thackway et al., [2023] This paper predicts neighborhood changes in Sydney, Australia using a model explanation tool that isolates the influence of each variable on their model output. This approach easily highlights which features are significant in predicting neighborhood changes in the future.
Pretner et al., [2011] indicates that infrastructure planners with access to proper tools can build data-driven city overviews to guide infrastructure and public space decisions. In addition, city planners and policymakers can track neighborhood changes in class, ethnicity, and life stage to shape zoning and housing policies (Weden et al., [2011]).

Gilling et al., [2021] uses public data to train models predicting whether a census tract will be gentrifying, declining, inclusively growing, or unchanging. Its limitation is that most tracts are classified as “unchanging,” making prediction difficult. Another paper, Corrigan et al., [2021] combines census data with spatial analysis to identify income and property value hotspots and declining cold spots but overlooks social and environmental factors beyond economics, race, and age.

4. Proposed Method

4.1. Intuition

Our project addresses the limitations of existing urban analysis by applying hybrid methodologies of analysis. The core intuition is that stable neighborhood typologies and accurate future projections require moving beyond simple algorithms, which is a gap noted in prior work. We achieve this by integrating a Self-Organizing Map (SOM) with K-Means clustering to define neighborhood archetypes. SOM preserves topology and maps multidimensional data onto a 2D map. Closer nodes on the SOM map are similar to each other compared to the ones farther away. This yields more robust and interpretable groups than traditional clustering methods like the K-Means approach used in *[Lynge et al. 2022]*. Further, we use K-Means to cluster the weights produced by SOM into macro clusters. This helps in better interpretation of the final resulting archetypes.

We have also used XGBoost Regressor with a First-Order Markov Process to predict features from 2030 to 2070. It achieves a R^2 value of 0.85, while other time-series methods like LSTM achieved a much lower R^2. We have used these predicted features to then predict neighborhoods using the same SOM and K-Means process as above. This is a significant step beyond papers like *[Delmelle et al. 2017]*, which classified shifts but did not predict future neighborhoods. This capability allows us to recursively forecast the long-term evolution of neighborhood features. Our approach offers a more accurate, stable, and actionable analytical tool than the current state of the art.

4.2. Description of Approaches

4.2.1. Data Exploration

Our primary data source is IPUMS NHGIS. It provided the vast majority of our required layers by offering source tables for our features of interest. We used Decennial Census data collected for the years of 1970 through 2000, and the modern ACS estimates for the years 2010, 2015, and 2022 for various census tracts, identified by the unique GISJOIN code. The LTDB and NHGIS crosswalks provided the links between historical GISJOINs and the 2010 GISJOIN standard, as well as the population weights used for interpolation. The raw dataset included over a million rows of data split across multiple files and is approximately 1.95 GB on the disk.

The features span several categories:

- **Core Demographics:** Total population, racial composition.
- **Housing Stock:** Total units, ownership/rental rates, and age/type.
- **Socioeconomic Status:** Estimated average household income, poverty, and education attainment.
- **Commute & Connectivity:** average commute time and transit use.

We also aimed to include environmental features using Air Quality Index (AQI) data. However, the attempt to integrate AQI was abandoned due to the high volume of missing data across the necessary historical timeline.

4.2.2 Data Cleaning and Harmonization

Rigorous cleaning and normalization were essential to unify the multi-source, multi-decade data. The process involved four main steps, executed using PySpark and Python libraries like Pandas, GeoPandas, etc:

- **Boundary Harmonization:** All data from 1970 to 2022 was interpolated and standardized to the 2010 census tract boundaries. This was achieved by using the crosswalk files to combine pieces of each historical tract into a single weighted 2010 aggregate.

- **Feature Mismatch Resolution:** The "Lowest Common Denominator" approach was used to resolve inconsistencies in classification bins across decades. For instance, narrower income bins from 1990 were summed to match the broader bins available in 1970. Estimated averages (e.g, average household income) were then calculated using midpoints.
- **Normalization:** Raw counts were converted into percentages (e.g., pct_white, pct_rented) to allow for effective comparison across the five decades of population growth and demographic shifts.

4.2.3 Feature Selection

To construct a consistent set of neighborhood indicators across all decades, we selected variables that were commonly used from the Census and ACS tables within each of our categories of interest, namely - demographics, housing stock, socioeconomic status, and family structure. This was done using the Popularity feature column on the IPUMS NHGIS datasets. The assumption was that tables frequently used in research and policy analysis are generally more stable, well-documented, and reliably measured over time. This approach allowed us to capture the necessary neighborhood characteristics without relying on inconsistently reported variables. The resulting features are Total Population, Average Household Income, Average Home Value, Total Housing Units, Percent Rented, Percent Vacant, Percent White, Percent Black, Percent Hispanic, Percent Foreign Born, Percent in Poverty, Percent with Bachelor's Degree or Higher, Percent Households with Children, Average Household Size, Average Commute Time, Percent Commute Work at Home, Percent Commute Public Transit, Percent Commute Drove Alone, Percent Built 1939 or Earlier, Percent Single-Family Detached, Percent 5+ Unit Buildings.

4.2.4 Clustering Model

The initial step in defining neighborhood archetypes involved a hybrid approach of Self-Organizing Map (SOM) and K-Means clustering applied to the fully harmonized dataset. This technique was used to overcome the non-linear, high-dimensional complexity of urban data and establish stable neighborhood typologies. The below steps were followed:

- **Scaling:** The data was filtered by year and for every year, all 21 socioeconomic features were standard scaled (mean of 0, standard deviation of 1) separately. This was done to ensure comparable feature magnitude.
- **Self Organizing Map (SOM):** Every census tract and year have been divided into separate rows. So if there are 7 different years for a census tract, there are 7 separate rows for the same tract. This is done so that all the years can be clustered differently based on their features and evolution of neighborhoods over the years can be studied. A 20 by 20 SOM grid (400 nodes) was trained over 5,000 iterations to perform non-linear dimensionality reduction. This step preserves the underlying relationships (topology) between similar neighborhood types by mapping them to nearby nodes. Similar neighborhoods would be closer together in the SOM space. In the top left corner of both of the images in Figure 1, we can see that while there is a low percentage of white populations, the percentage of black populations in the same corresponding areas are high. This shows a clear demarcation amongst the neighborhoods.

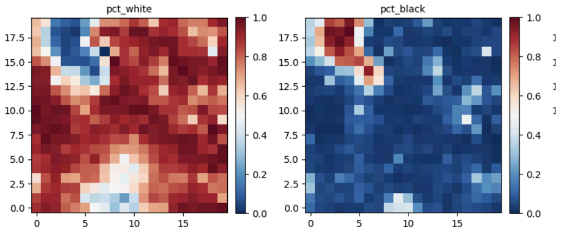

Figure 1: SOM clusters for percent white and black populations

- **K-Means for Macro-Clusters:** K-Means was applied directly to the 400 SOM node weight vectors, grouping them into larger and coherent "macro-clusters." Based on manually understanding the clusters, 9 macro-clusters were selected to represent the final, stable neighborhood archetypes.
- **Assignment**: Every record (GISJOIN-year) in the approx 478,000 row panel was assigned one of the 9 macro-cluster labels.

Analysis of the cluster centroids shows 9 distinct neighborhood profiles based on highly differentiating features. Features such as ethnicity, home value, no. of people, and kind of houses varied greatly over all the clusters.

4.2.4 Predictive Model

The second step is to forecast the evolution of neighborhood feature values up to 2070 using a supervised XGBoost model with a First-Order Markov Process, where the next state only depends on the immediately previous state. We used tabular time-indexed data where every unique GISJOIN is listed 7 times, once for each of the 7 time periods, along with values for all 21 features. XGBoost was chosen as it is optimized for vectorized processes and saves computational time; this means it is able to make thousands of predictions at once. The following steps were followed:

- **Data Preparation (Markov Process):** We sorted the data by GISJOIN and YEAR and created a lag of one year for all 21 features to allow XGBoost to learn temporal relationships between the data. A lag of 1 was the most optimal option given the size of our dataset (7 distinct time periods), as having more lags would have resulted in loss of too much information when training the XGBoost model.
- **Training (XGBoost):** The model trains 21 independent XGBoost Regressors, one per feature, data from time periods t and t-1.
- **Prediction:** After the model learns the relationship between historical data from 1970 to 2022, it then predicts values from 2030 to 2070. In this process, the prediction for a time period relies exclusively on the data of the prior period, where data is forecasted for every future year by batch. The input is the current time period’s data and lag 1 values, and the output are values for the next time period.

4.2.5 Visualization

The final output is an interactive Census Analysis Map tool created using D3 and designed to make complex data on neighborhood change accessible to end-users. Below are the parts of the UI:

- **Archetype Layer:** The clustering layer displays the assignment of the macro-clusters, allowing users to visualize the distribution of these archetypes. A slider moves across the years to show the changing neighborhoods with time. Cities with the most change from 1970 to 2022 have been selected and can be visualized using the dropdown.
- **Prediction Layer:** This layer visualizes all the individual predicted features. Movement of different features through the years can be seen for the 12 cities we selected.
- **Interactive Context:** When a user interacts with a specific census tract, a pop-up displays detailing the tract’s cluster information or prediction feature value based on the layer. Different years, cities and features can be selected for visualization.

5. Evaluation

5.1. Experimental Questions

Our evaluation was conducted using the dataset for the entirety of the United States, standardized to the 2010 census tract boundaries. The experiments were designed to answer the following core questions:

- Do the preprocessing, harmonization, and interpolation steps produce internally consistent historical data?

- Do the clusters correspond to recognized socioeconomic and spatial patterns in the real world?
- Based on the highest R^2 values, which of our predictive models (XGBoost, LightGBM, and LSTM) best forecasts future neighborhood features values across all variables?
- Is the visualization intuitive and responsive enough to support exploration across decades and cities?

5.2 Data Preprocessing Evaluation

Because the project depends on a heavily engineered dataset, validating the correctness of preprocessing was a central component of evaluation. We applied the following strategies:

- **Consistency checks:** We examined raw counts and percentages for each variable over time to ensure that there were no artificial spikes caused by changing boundaries.
- **Distribution comparisons:** For each feature, we compared means, medians, and standard deviations to confirm no anomalous values were introduced due to data merges or codebook differences.
- **Replication of known patterns:** A critical validation step was examining whether the clusters created on our processed data reproduced documented urban phenomena. We discuss this further in the Clustering Evaluation section below.

5.3 Clustering Evaluation

The clustering patterns align with known demographic and economic shifts across the USA.

For instance over many years, **Las Vegas** experienced a decline in black populations with an increase in Hispanic populations. The population of lower-diversity White suburbs shrunk and affluent White areas remained stable, consistent with the city's broad immigration trends. [12]

Detroit reflects clear white flight from 1970s onwards, with affluent White residents moving to the suburbs and Black populations increasing in the urban core.[13]

Around **Central Park in NYC**, increasing diversity aligns with decades of gentrification and reinvestment.[14]

In **Atlanta**, especially around Midtown, clusters show rising gentrification, while many outlying areas display sustained or increasing Black population over time.[15]

Unlike KMeans or PCA, SOM preserves the structure of the data on a 2D grid. Points that are similar end up closer together on the map. Additionally, it doesn't force hard cluster boundaries. It naturally shows gradual transitions between neighborhoods or demographic profiles.

5.4 Predictive Model Evaluation

While demographic forecasting inherently carries a lot of uncertainty, we aimed for plausibility rather than exact accuracy. We ran the below evaluations for our predictive models to determine the most suitable approach:

- **Alignment with contemporary research:** Income and home value forecasts indicate a widening disparity between affluent and low-income neighborhoods. Some cities also exhibit signs of the "white flight" phenomenon, where rising poverty and increasing Black population shares in certain urban cores coincide with projected decline in White population shares. These patterns emerged without being explicitly encoded in the model, indicating that the learned relationships in the data appear realistic.

- **Model comparison:** We attempted to determine the best-performing forecasting model by comparing R^2 scores across all the features. The information is contained in the table below.

Model	Mean R^2	Median R^2	Std Dev R^2
XGBoost	0.770	0.852	0.177
LightGBM	0.224	0.550	1.020
LSTM	-25.137	-3.707	82.524

Overall, XGBoost emerged as the most robust and accurate model for neighborhood forecasting, while the poor performance of LSTM highlights the challenges of applying Deep Learning to sparse temporal Census data.

5.5 Visualization Evaluation

We tested our D3 user interface for:

- **Responsiveness:** While we do experience some performance issues during the initial loading of the clustering and prediction data, the interface is consistently responsive once all of the data is cached. Zooming, hovering, and switching views perform smoothly thereafter.
- **Clarity of tooltips:** The tooltips are informative and relate information associated to cluster labels or predicted feature values, giving the user a contextual understanding about the geographic unit, from the census tract level to the national scale.
- **Ease of use/exploration:** The buttons, dropdowns and sliders on the UI are self-explanatory and intuitive, allowing users to seamlessly navigate between time periods and cities.

D3 is better for visualization because it gives pixel level control making it easier to zoom in on a particular census tract. It is specifically built for the web, so hover, click and animations are configurable.

Overall, while there is room for technical improvement, the user interface effectively communicates neighborhood clusters and prediction data, enabling users to deeply explore urban phenomena.

6. Conclusion

Our approach combines clustering with time-series data to model and predict how neighborhoods evolve over time. By integrating datasets into one framework, we reveal dynamic links between urban, social, and demographic factors. Interactive visualizations show how neighborhoods shift between clusters and the factors driving these changes.

6.2 Limitations

One of the main limitations with the project was the limited amount of data. We heavily relied on NHGIS data because other historical datasets were unavailable or inconsistent. The data from 1970 is incomplete, which shows up as a greyed area in the visualization. The D3 visualizations are a little laggy and required significant RAM, which affected smooth interaction with the dashboard. Limited forecasting accuracy reduces the reliability of clusters built on future predictions.

6.3 Potential Future Extensions

Future work could include looking at other methods for time series predictions. Another possible extension could be combining NHGIS data with other sources such as data from the EPA, regional Department of Transportation, and crime data, to get a richer understanding of the clusters. After SOM, we have used K-Means to create macro clusters. Other clustering algorithms may also be experimented with.

The final list of finished tasks are on the Gantt Chart, where all team members have contributed a similar amount of effort.

Key References

1. Bell, W. (1965). Urban neighborhoods and individual behavior. In M. Sherif & C. W. Sherif (Eds.), Problems of youth (pp. 235–264)

2. Delmelle, E. C. (2017). Differentiating pathways of neighborhood change in 50 U.S. metropolitan areas. Environment and Planning A, 49(10), 2402–2424. https://doi.org/10.1177/0308518x17722564
3. Spielman, S. E., & Thill, J.-C. (2008). Social area analysis, data mining, and GIS. Computers, Environment and Urban Systems, 32(2), 110–122. https://doi.org/10.1016/j.compenvurbsys.2007.11.004
4. Wo, J. C., Kim, Y.-A., & Berg, M. T. (2025). Exploring crime through physical and social neighborhood factors: Greenspace and social disconnection in Los Angeles. Journal of Criminal Justice, 98, 102410. https://doi.org/10.1016/j.jcrimjus.2025.102410
5. Lynge, H., Visagie, J., Scheba, A., Turok, I., Everatt, D., & Abrahams, C. (2022). Developing neighborhood typologies and understanding urban inequality: a data-driven approach. Regional Studies, Regional Science, 9(1), 618–640. https://doi.org/10.1080/21681376.2022.2132180
6. Fang, C., Zhou, L., Gu, X., Liu, X., & Werner, M. (2025). A data driven approach to urban area delineation using multi source geospatial data. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-93366-x
7. Thackway, W., Ng, M., Lee, C.-L., & Pettit, C. (2023). Building a predictive machine learning model of gentrification in Sydney. Cities, 134, 104192. https://doi.org/10.1016/j.cities.2023.104192
8. Pretner, D. G., & Gstach, D. (2011, January 1). Creating Added Value in Working Landscapes–The Development of the Atlanta BeltLine.
9. Weden, M. M., Bird, C. E., Escarce, J. J., & Lurie, N. (2011). Neighborhood archetypes for population health research: Is there no place like home? Health & Place, 17(1), 289–299. https://doi.org/10.1016/j.healthplace.2010.11.002
10. Gilling, G., Mishra, V., Gibli, J., & Hernandez, D. (2021). Predicting Neighborhood Change Using Publicly Available Data and Machine Learning. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3911354
11. Corrigan, A. E., Curriero, F. C., & Linton, S. L. (2021). Characterizing clusters of gentrification in metro Atlanta, 2000 to 2016. Applied Geography, 137, N.PAG–N.PAG. https://doi.org/10.1016/j.apgeog.2021.102597
12. Immigration. (n.d.). Cdclv.unlv.edu. https://cdclv.unlv.edu/healthnv/immigration.html
13. Gorrepati, R. (2024, May 7). Segregation and White Flight in Detroit. ArcGIS StoryMaps; Esri. https://storymaps.arcgis.com/stories/12b84ab10e734e72b59beac8df64c903
14. Smale, L. (2022, December 7). Urban Sustainability and Environmental Gentrification. ArcGIS StoryMaps; Esri. https://storymaps.arcgis.com/stories/421bf3f9933d48528a9396e7eddc575a

15. Pandey, L., & Sjoquist, D. (2022). An Exploration of Racial Residential Segregation Trends in Atlanta (pp. 1970–2020). https://cslf.gsu.edu/files/2022/04/cslf2201-4.pdf