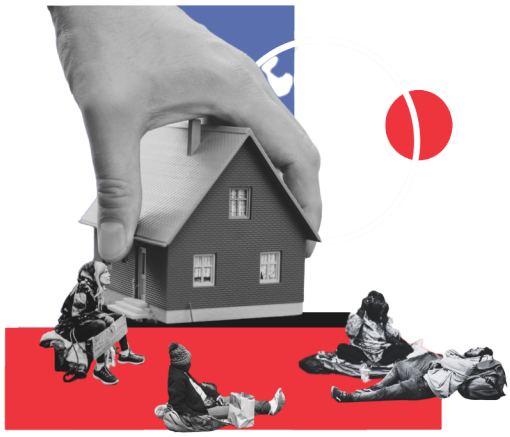
Predictive Model Based on Homelessness



Boston University
Faculty of Computing and Data Sciences
DS 701: Tools for Data Science

Instructor: Thomas Gardos

Team Members:

Syeda Aqeel	MSDS'25	(sharoo@bu.edu)
Shiheng Xu	MSDS'25	(timxu5@bu.edu)
Samritha Aadhi Ravikumar	MSDS'25	(samritha@bu.edu)
Kunshu Yang	MSDS'25	(luckykun@bu.edu)
Renjie Fan	MSDS'25	(renjief@bu.edu)

BOSTON UNIVERSITY

Table of Contents

Introduction

Project Goal & Overview

This project aims to develop a community-level predictive model for homelessness by analyzing data spanning 2010–2023. Unlike previous studies that primarily focused on predicting homelessness at the individual level, this project centers on identifying structural community-level factors, such as rent levels and economic conditions, that influence homelessness. Using data from approximately 400 Continuums of Care (CoC) funded by the U.S. Department of Housing and Urban Development (HUD), the project seeks to predict the number or rate of homelessness in each CoC, providing insights to guide policy decisions and resource allocation.

Big Impact

Homelessness is a complex societal challenge shaped by a range of factors including economic conditions, housing availability, healthcare access, and social services. This project aims to deepen the understanding of these dynamics, focusing on the community-level characteristics that drive homelessness. By analyzing these factors, the project provides valuable insights to help policymakers and community organizations allocate resources more effectively and enhance the impact of homelessness assistance programs. Furthermore, it offers a data-driven framework to guide long-term strategies for addressing homelessness. Through the prediction of homelessness rates and trends, the project seeks to identify high-risk communities, enabling timely interventions that can reduce homelessness on a larger scale.

Client Information:

Dr. Tom Byrne

Dr. Molly Richard

This project is guided by two leading experts in the field of homelessness research at Boston University. Dr. Tom Byrne, an Associate Professor at the School of Social Work, brings extensive expertise in studying the structural determinants of homelessness, while Dr. Molly Richard, a Postdoctoral Associate at the Center for Innovation in Social Science, specializes in community development and social services. Their combined insights and experience provide

crucial guidance, ensuring that the project remains focused on real-world challenges and produces findings that are both relevant and actionable for addressing homelessness effectively.

Data Description and Data Cleaning

The primary goal of this dataset is to facilitate predictive modeling of homelessness across Continuums of Care (CoCs) in the United States. The dataset combines multiple sources of data to analyze the relationships between homelessness and various socioeconomic, housing, and environmental factors, spanning from 2007 to 2023.

The dataset integrates information from the following sources:

1. **American Community Survey (ACS):**

   Among the largest annual surveys that the American Community Survey(ACS) carries out are those concerning quite granular-level information on demographic, social, economic, and housing characteristics. It is continuous rather than a count that occurs once in a decade, just like the decennial census. This shows that America's communities can provide a portrait on more regular intervals by compiling information on one- and five-year intervals, thus aiding high-quality, reliable estimates for small geographic areas.

   This project downloaded the direct data from the Census Bureau's Website for the years 2010 through 2022. In the period of time for the selected variables, it captures five-year estimates. The ACS is important in measuring the socioeconomic conditions underlying homelessness trends, such as rates of poverty, unemployment, housing vacancy, rent burden, and household composition.

- **GEO_ID**
- **Year**
- B01003_001E: Total Population
- B17001_002E: Population Below Poverty Level
- B25002_001E: Total Housing Units
- B25002_003E: Vacant Housing Units
- B25003_003E: Renter-Occupied Housing Units
- B25003_001E: Total Housing Units (for Tenure)
- B25106_001E: Total Households (for Housing Costs)
- B23025_003E: Civilian Labor Force Employed
- B23025_005E: Unemployed Individuals
- B25064_001E: Median Gross Rent
- B19013_001E: Median Household Income

- Unemployment rate: $\frac{Unemployed\ Individuals}{Civilian\ Labor\ Force\ Employed} = \frac{B23025\_005E}{B23025\_003E}$
- Vacancy Rate: $\frac{Vacant\ Housing\ Units}{Total\ Housing\ Units} = \frac{B25002\_003E}{B25002\_001E}$
- Renter Household Rate: $\frac{Renter-Occupied\ Housing\ Units}{Total\ Housing\ Units\ (for\ Tenure)} = \frac{B25003\_003E}{B25003\_001E}$
- Cost-Burdened Renter Rate: $\frac{Total\ Households\ (for\ Housing\ Costs)}{Renter-Occupied\ Housing\ Units} = \frac{B25106\_001E}{B25003\_003E}$
- Poverty Rate: $\frac{Population\ Below\ Poverty\ Level}{Renter-Occupied\ Housing\ Units} = \frac{B17001\_002E}{B25003\_003E}$

Since ACS data are collected over 5-year periods, the available data range is from 2010 to 2022. To extend this range, we duplicated the 2010 data for the years 2007 to 2009 and the 2022 data for 2023.

2. CoC Mapping dataset:

Used to align census tracts with CoC regions for geographic consistency.(Byrne, 2022) conducted a crosswalk from GEO_ID to CoC using two datasets: "tract_coc_match_2022" and "tract_coc_match_2019." The 2019 dataset was utilized for the years 2007 to 2019, while the 2022 dataset was applied for the years 2020 to 2023. Approximately 1% of GEO_ID records could not be matched to a CoC using the dataset.

Year	unmatched_count
2007	591
2008	591
2009	591
2010	591
2011	578
2012	570
2013	570
2014	569
2015	565
2016	565
2017	565
2018	565
2019	565
2020	1494
2021	1494
2022	616
2023	616

Then we take next step to aggregate to CoC level at following method:

Variable Name	Aggregation Method	Description
B01003_001E (Total Population)	Sum	Sum of total population in each CoC
B17001_002E (Poverty Population)	Sum	Sum of total population below the poverty line in each CoC
B25002_001E (Total Housing Units)	Sum	Sum of total housing units in each CoC
B25002_003E (Vacant Housing Units)	Sum	Sum of total vacant housing units in each Coc
B25003_003E (Renter-Occupied Units)	Sum	Sum of renter-occupied housing units in each CoC
B25003_001E (Total Occupied Units)	Sum	Sum of total occupied housing units in each CoC
B25106_001E (Cost-Burdened Households)	Sum	Sum of renter households spending more than 30% of income on housing
B23025_003E (Civilian Labor Force)	Sum	Sum of the total civilian labor force in each CoC
B23025_005E (Unemployed Individuals)	Sum	Sum of unemployed individuals in each CoC
B25064_001E (Median Gross Rent)	Population-Weighted Average	Weighted average of median gross rent using population as weights
B19013_001E (Median Household Income)	Population-Weighted Average	Weighted average of median household income using population as weights

3. Point-in-Time (PIT) Homeless Counts:

The Point-in-Time (PIT) dataset, provided by the U.S. Department of Housing and Urban Development (HUD), captures a snapshot of homelessness across the United States on a single night each year. It includes data on the number of homeless individuals, families, and the unsheltered population, collected by Continuums of Care (CoCs) across the country. The PIT dataset is used to assess the scope of homelessness, track trends over time, and inform

policy and funding decisions. It is a crucial resource for understanding the dynamics of homelessness, helping to guide interventions and allocate resources effectively to address the issue.

- **CoC_Number**
- **Year**
- Overall Homeless
- Overall Homeless Individuals
- Overall Homeless People in Families
- Unsheltered Homeless
- Sheltered Homeless

4. **Housing Inventory Counts (HIC):**

The Housing Inventory Count (HIC) dataset, provided by the U.S. Department of Housing and Urban Development (HUD), aggregates data on shelter and housing capacity across various Continuums of Care (CoCs) in the United States. It includes detailed information on the availability of year-round beds for different types of housing, such as emergency shelters (ES), transitional housing (TH), safe havens (SH), rapid re-housing (RRH), and permanent supportive housing (PSH). The HIC dataset helps track the supply of housing for individuals and families experiencing homelessness, providing crucial data for understanding shelter availability and informing resource allocation in homelessness response systems.

- **CoC_Number**
- **Year**
- Total Year-Round Beds (ES, TH, SH)
- Total Year-Round Beds (ES)
- Total Year-Round Beds (TH)
- Total Year-Round Beds (SH)
- Total Beds for Households with Children (ES, TH, SH)
- Total Beds for Households without Children (ES, TH, SH)
- Total Beds for Households with Children (ES)
- Total Beds for Households without Children (ES)
- Total Beds for Households with Children (TH)
- Total Beds for Households without Children (TH)
- Total Beds for Households with Children (SH)
- Total Beds for Households without Children (SH)
- Total Year-Round Beds (RRH)
- Total Beds for Households with Children (RRH)
- Total Beds for Households without Children (RRH)

- Total Year-Round PSH Beds
- Total Beds for Households with Children (PSH)
- Total Beds for Households without Children (PSH)

Housing Type	Description
ES	Emergency Shelter
TH	Transitional Housing
SH	Safe Haven
RRH	Rapid Re-Housing
PSH	Permanent Supportive Housing

The Housing Inventory Count (HIC) dataset by HUD aggregates data on shelter and permanent housing capacity over multiple years, with distinct conventions in column naming across time. To facilitate merging data from 2007 to 2023 into a unified, standardized format, column renaming was necessary.

The column naming conventions for data from 2014 to 2023 followed a consistent structure, making them an ideal standard for the entire dataset. However, column names from 2007 to 2013 varied significantly from this format. These inconsistencies posed challenges for merging and analyzing the data as a single entity.

The standardization process involved aligning all column names with the 2014–2023 naming convention. This ensured uniformity across all years, making it easier to analyze shelter and housing capacity indicators. Each renamed column was carefully mapped to reflect the original data while adopting the standardized naming convention. This approach ensured consistency, improved interpretability, and enhanced the usability of the dataset for further analysis.

Original Column Name	Renamed Column Name
Total Year-Round ES Beds	Total Year-Round Beds (ES)
Total ES Beds for Households with Children	Total Beds for Households with Children (ES)
Total ES Beds for Households without Children	Total Beds for Households without Children (ES)

Total Year-Round TH Beds	Total Year-Round Beds (TH)
Total TH Beds for Households with Children	Total Beds for Households with Children (TH)
Total TH Beds for Households without Children	Total Beds for Households without Children (TH)
Total Year-Round SH Beds	Total Year-Round Beds (SH)
Total Year-Round SH Beds without Children	Total Beds for Households without Children (SH)
Total Year-Round RRH Beds	Total Year-Round Beds (RRH)
Total RRH Beds for Households with Children	Total Beds for Households with Children (RRH)
Total RRH Beds for Households without Children	Total Beds for Households without Children (RRH)
Total Year-Round Beds (PSH)	Total Year-Round PSH Beds
Total PSH Beds for Households with Children	Total Beds for Households with Children (PSH)
Total PSH Beds for Households without Children	Total Beds for Households without Children (PSH)

5. Weather data:

The Weather dataset, retrieved from the NOAA Climate at a Glance, (National Oceanic & Atmospheric Administration) included average temperature data from 2007 to 2023 with 49 rows and 18 columns.

- Year
- Average Temperature for States of US

The data preparation process began by loading the ACS and Weather datasets. The ACS dataset included 6,524 rows and 42 columns, capturing detailed demographic and economic information. To integrate these datasets, state initials were mapped to the ACS data, ensuring that only the 50 U.S. states were included while excluding Union Territories. This mapping allowed the weather data to be merged as a new variable, "Average Temperature," into the ACS dataset.

During data cleaning, several issues were identified and addressed. For instance, the "Average Temperature" for Washington was reassigned to reflect the correct data for the District of Columbia (D.C.). Hawaii's average temperature data was entirely missing, affecting 34 rows. Additionally, South Dakota's weather data had been mistakenly swapped with Puerto Rico's, which was corrected by assigning accurate temperature data to South Dakota and marking Puerto Rico's missing values as NaN.

Further investigation revealed an additional 34 rows with missing values across various columns, including "Average Temperature." These rows were updated with NaN values where applicable. After all corrections, the final merged dataset expanded to 6,524 rows and 43 columns, ensuring no data loss during the merging process. However, 102 rows still had missing temperature values due to incomplete data for Hawaii, Puerto Rico, and unidentified rows. Additionally, 34 rows lacked corresponding state initials because of incomplete information, leaving some gaps in the dataset despite extensive cleaning efforts.

Data Merging

The data merging process integrated four key datasets—HIC, PIT, ACS, and NOAA—by aligning them on the common fields of "Year" and "CoC_Number." The Housing Inventory Count (HIC) and Point-in-Time (PIT) datasets provided information on shelter capacity and homelessness counts, respectively, while the American Community Survey (ACS) contributed socioeconomic and housing characteristics at a granular level. The NOAA dataset added environmental context with average temperature data.

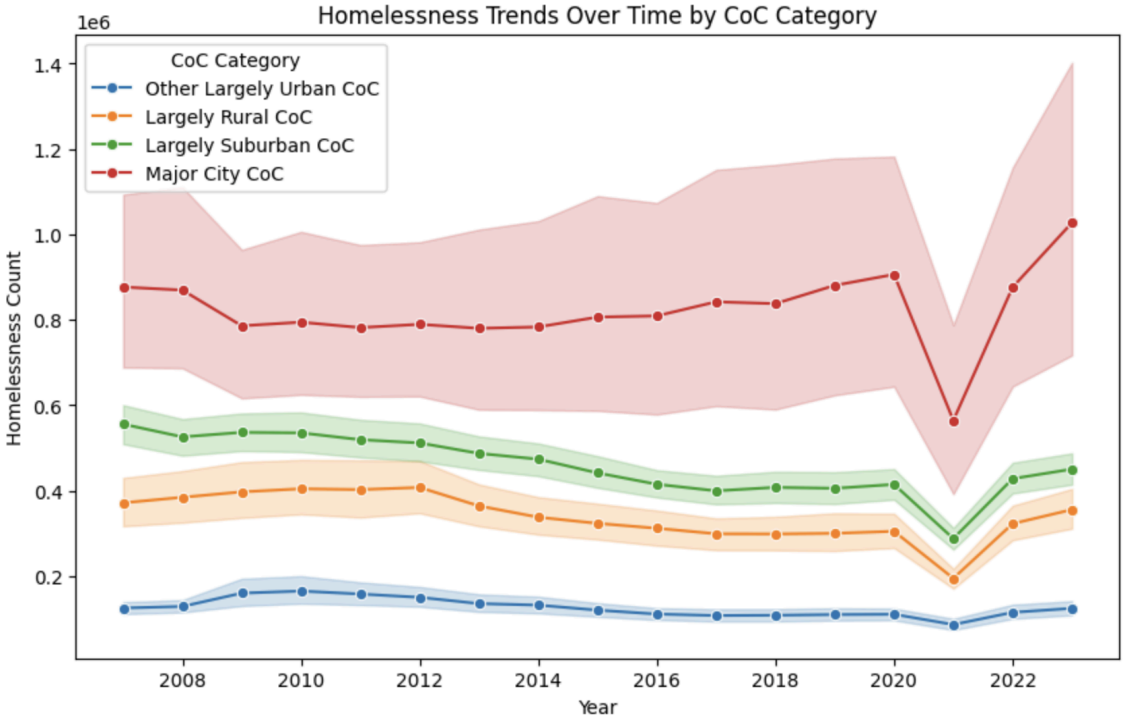
To ensure compatibility, preprocessing steps included mapping state initials, standardizing variable naming conventions, and aggregating data to the Continuum of Care (CoC) level. Aggregation involved summing population-related variables, calculating weighted averages for metrics like rent and income, and ensuring consistent definitions across datasets. For example, the HIC data from 2007 to 2013 required column renaming to match the 2014–2023 format, ensuring uniformity in analysis.

Data cleaning addressed discrepancies such as swapped state temperatures and missing values. For instance, Puerto Rico's temperature data was marked as NaN, and Hawaii's missing data was flagged. After corrections, the datasets were merged without data loss, expanding to 6,524 rows and 43 columns. This comprehensive integration created a robust foundation for predictive modeling, capturing the interplay of housing, economic, and environmental factors driving homelessness.

BOSTON UNIVERSITY

Data Visualization

1. Homelessness Trends by CoC Category

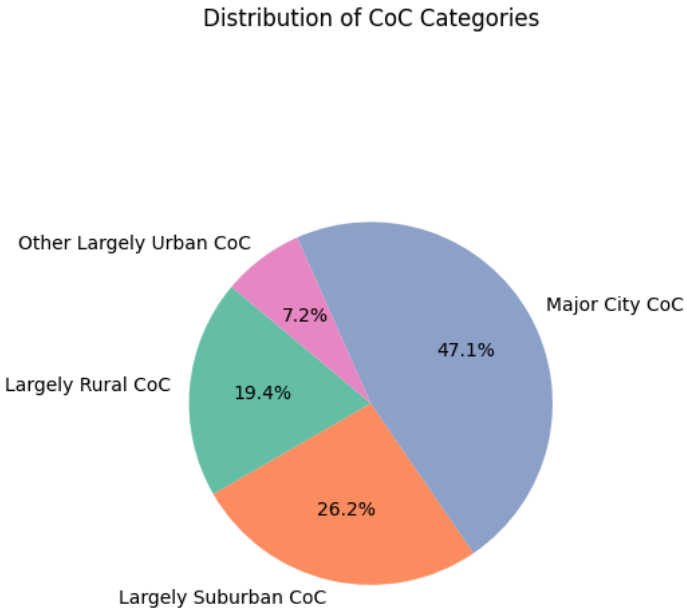


The visualization tracks homelessness trends from 2007 to 2023 across four types of Continuum of Care (CoC): Major City CoC, Largely Suburban CoC, Largely Rural CoC, and Other Largely Urban CoC. The y-axis represents the number of homeless individuals, while the x-axis shows the time period.

Key findings include the consistently high levels of homelessness in Major City CoCs, with annual counts approaching 1 million. A sharp drop in 2021, followed by a significant rebound in 2022, likely reflects the temporary effects of COVID-19 policies such as emergency housing programs or data reporting changes. Largely Suburban and Other Largely Urban CoCs show moderate levels of homelessness, with gradual declines over time and fewer fluctuations compared to urban areas. Largely Rural CoCs consistently report the lowest homelessness counts, generally under 200,000, with only a slight downward trend.

These findings highlight the disproportionate impact of homelessness in Major City CoCs, suggesting the need for targeted interventions in urban areas. The sharp 2021 decline reflects the temporary nature of pandemic-related measures, raising concerns about the long-term sustainability of such strategies.

2. Distribution of CoC Categories:

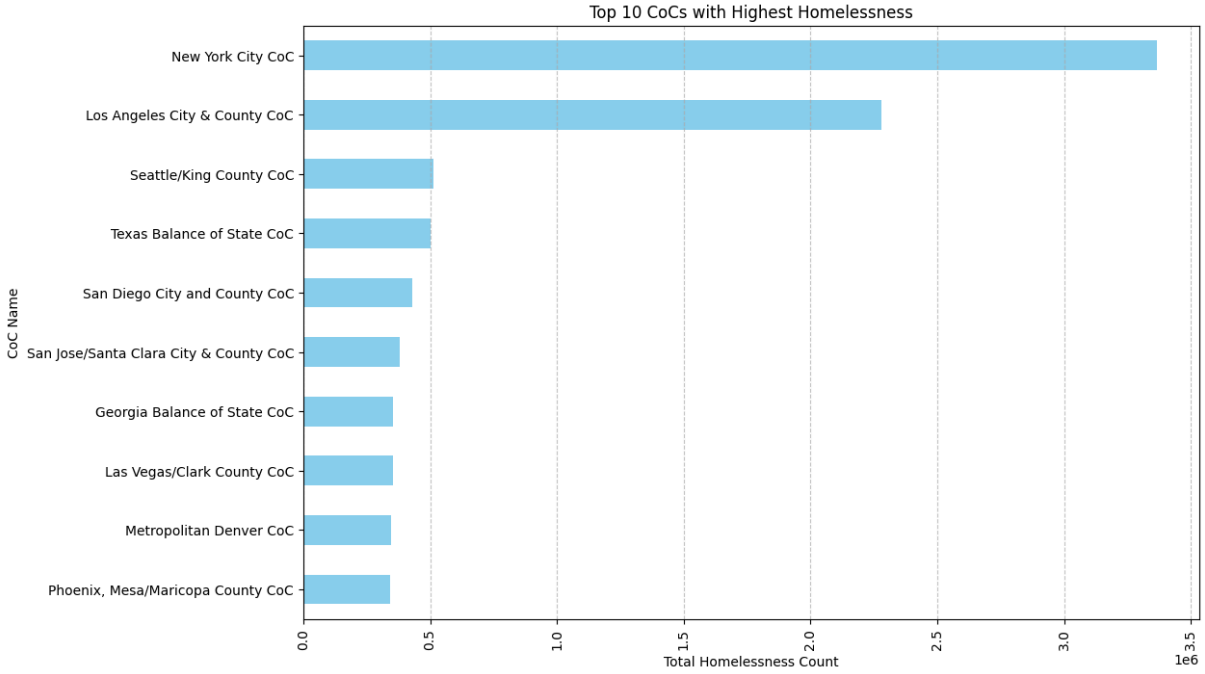


The pie chart highlights the distribution of homelessness across different Continuums of Care (CoC) categories. Major City CoCs account for 47.1% of the total, emphasizing the significant concentration of homelessness in urban areas. Largely Suburban CoCs make up 26.2%, while Largely Rural CoCs contribute 19.4%, reflecting a notable, though smaller, share compared to urban centers. Other Largely Urban CoCs represent the smallest portion at 7.2%, further illustrating the dominant role that major cities play in homelessness trends. This distribution underscores the need for targeted interventions that address the unique challenges faced by urban communities.

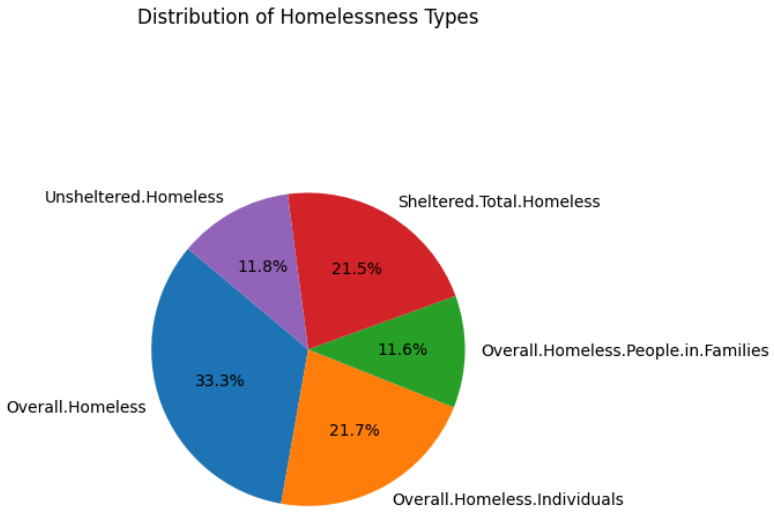
3. Top 10 CoCs with Highest Homelessness:

This bar chart highlights the top 10 Continuums of Care (CoCs) with the highest homelessness counts across the United States. CoCs represent geographic regions that receive federal assistance for homelessness programs. The chart reveals key insights into regional homelessness patterns. New York City stands out with the highest homelessness count, driven by its large population, high cost of living, and unique policies, such as the right-to-shelter law, which may influence reported figures. Los Angeles follows closely, facing significant housing affordability issues and a large unsheltered population. Other urban areas like Seattle/King County, San Diego, and San Jose/Santa Clara face similar challenges, particularly high housing costs. Additionally, some less urbanized areas, such as Texas and Georgia's Balance of State CoCs, also

rank highly, likely due to their broader geographic coverage and the aggregation of data from multiple smaller communities. These findings highlight significant regional disparities in homelessness, underscoring the need for tailored, localized strategies to address housing and homelessness crises.

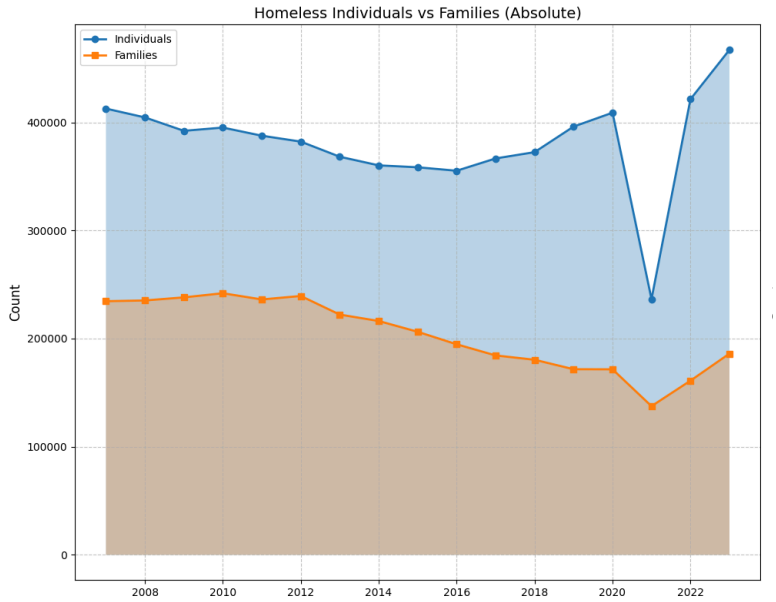


4. Distribution of Homelessness Types:



The pie chart illustrates the distribution of homelessness types, with Overall Homeless making up the largest proportion at 33.3%, representing all categories combined. Following this, Overall Homeless Individuals and Sheltered Homeless each contribute significant shares, at 21.7% and 21.5%, respectively, highlighting the prevalence of individuals and those in sheltered environments. In contrast, Unsheltered Homeless and Homeless Families account for smaller portions, with 11.8% and 11.6%, respectively, suggesting that both unsheltered individuals and family units represent a lesser proportion of the total homelessness population. This distribution underscores the dominance of individual and sheltered homelessness in the overall landscape.

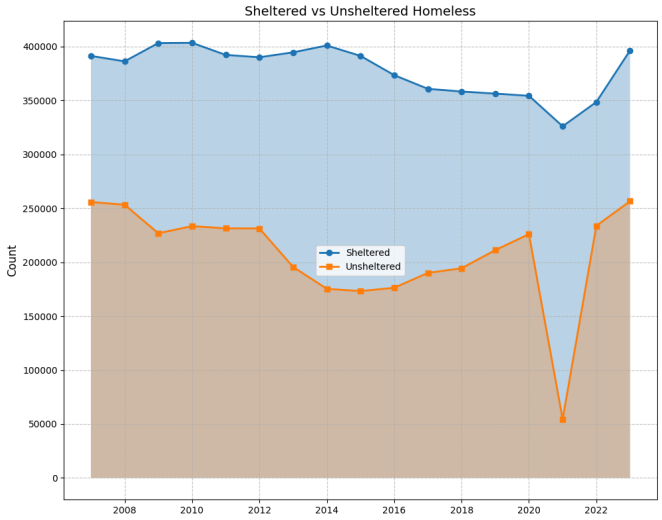
5. Trends in Homelessness: Individuals vs. Families:



Homelessness trends reveal that individuals consistently outnumber families, highlighting the predominance of individual homelessness in the overall count. This suggests that policy efforts should prioritize addressing homelessness among individuals. From 2007 to 2019, both individuals and families saw a gradual decline in homelessness, likely driven by interventions like Housing First programs. A sharp drop occurred in 2020-2021, primarily due to COVID-19-related policies such as eviction moratoriums and emergency shelter provisions, although data collection challenges during the pandemic may have contributed to this decrease. However, a sharp rebound in 2023 points to the lasting impact of economic pressures, such as rising rent and unemployment, coupled with reduced policy support. In contrast, family homelessness demonstrated a more consistent decline, possibly due to targeted programs like

federal child and family welfare initiatives, which have been more effective in addressing the needs of families compared to individuals.

5. Trends in Homelessness: Sheltered vs. Unsheltered Populations:



Sheltered homelessness has shown a relatively stable trend over time, indicating consistent utilization of shelter services. However, this stability may suggest that the shelter system is nearing its capacity, with limited room for expansion to accommodate growing needs. In contrast, unsheltered homelessness has exhibited more volatility, particularly with a significant decline in 2021, likely due to emergency measures such as expanded shelters and temporary housing programs during the pandemic. This underscores the effectiveness of short-term interventions in response to crises. However, the subsequent rebound in unsheltered homelessness in 2023 reveals underlying systemic issues, such as a lack of permanent housing solutions and barriers to shelter access. Additionally, fluctuations in unsheltered populations are influenced by regional weather conditions, with colder climates seeing higher shelter utilization and warmer regions, like California, experiencing larger unsheltered populations.

BOSTON UNIVERSITY

Data Modeling and Data Analysis

Client Question 1: To what extent community-level measures of rent, poverty, and other housing market conditions accurately predict the number of persons experiencing homelessness in a community?

The objective of data modeling is to determine how well community-level measures—such as rent, poverty, and housing market conditions—predict homelessness rates across various categories. By using a range of socioeconomic, housing, and environmental features, we sought to identify which factors most influence homelessness in different groups. This approach allowed for a comprehensive understanding of the relationship between community dynamics and homelessness, supporting the development of targeted interventions. The significance of this process lies in the identification of key predictors, which can inform policy decisions and improve resource allocation to reduce homelessness effectively.

Features

The following features were used in the analysis to predict homelessness rates across different categories. These features include a mix of socioeconomic, housing, and environmental factors.

1. Total Population
2. Median Gross Rent
3. Median Household Income
4. Poverty Rate
5. Vacancy Rate
6. Unemployment Rate
7. Cost Burdened Rate
8. Renter Household_Rate
9. Total Year-Round Beds (ES, TH, SH)
10. Average Temperature

Target variables:

The target variables for this analysis focus on different dimensions of homelessness, measured as rates per 1,000 individuals.

1. Overall Homelessness Rate Per 1000 Individuals
2. Overall Homelessness Individuals Rate Per 1000 Individuals
3. Overall Homelessness People in Families Rate Per 1000 Individuals
4. Unsheltered Homelessness Rate Per 1000 Individuals
5. Sheltered Homelessness Rate Per 1000 Individuals

Data Preprocessing Steps

Standard Scaling of Features:
Continuous predictors were standardized to ensure all features had a mean of 0 and a standard deviation of 1. This step was critical for models like linear regression and gradient boosting algorithms, which are sensitive to feature scaling.

Log Scaling of Target Variables:
Target variables representing homelessness rates were log-transformed to reduce skewness and stabilize variance. This transformation improved model performance, particularly for targets with wide value ranges.

Handling Missing Values:
Missing values in the target variables were imputed with zeros. This decision was based on the assumption that missing homelessness rate data may correspond to areas reporting no homelessness cases during the period of analysis.

Model Development and Evaluation

Linear Regression:

The Linear Regression model achieved moderate performance across all categories, with RMSE values ranging from 0.24 for Families to 0.41 for Unsheltered. Its R^2 scores were generally low, with the highest value of 0.44 observed for the Sheltered category, indicating that the model explained only a small proportion of the variance in the homelessness data.

Random Forest:

The Random Forest model displayed the highest overall performance, achieving an R^2 of 0.78 and an RMSE of 0.99. It performed exceptionally well for predicting Families (R^2: 0.82) and

Sheltered homelessness (R^2: 0.88). However, the RMSE values for Unsheltered and Individuals were relatively high, suggesting that the model may have experienced overfitting or variability in its predictions for these categories.

Histogram-Gradient Boosting:

The Histogram-Gradient Boosting model delivered balanced performance with lower RMSE values compared to Random Forest, such as 0.35 for Individuals. Its R^2 scores were moderate, with the best performance observed for Individuals and Unsheltered homelessness (both with an R^2 of 0.59), indicating a reasonable fit but less explanatory power than Random Forest.

Extreme Gradient Boosting (XGBoost):

The Extreme Gradient Boosting model had the lowest R^2 values among the models, suggesting weaker explanatory power overall. Its RMSE values ranged from 0.38 for Families to 0.43 for Individuals, indicating that while the model provided stable predictions, it was less accurate compared to the Random Forest model.

MODELS	HOMELESSNESS TYPE	RMSE	R^2
Linear Regression	Overall	0.38	0.39
	Individuals	0.37	0.36
	Families	0.24	0.29
	Unsheltered	0.41	0.30
	Sheltered	0.28	0.44
Random Forest	Overall	0.99	0.78
	Individuals	0.78	0.77
	Families	0.37	0.82
	Unsheltered	0.78	0.69
	Sheltered	0.48	0.88

MODELS	HOMELESSNESS TYPE	RMSE	R^2
Histogram-Gradient Boosting	Overall	0.38	0.58
	Individuals	0.35	0.59
	Families	0.31	0.41
	Unsheltered	0.36	0.58
	Sheltered	0.33	0.52
Extreme Gradient Boosting (XGBOOST)	Overall	0.43	0.42
	Individuals	0.43	0.18
	Families	0.38	0.18
	Unsheltered	0.42	0.45
	Sheltered	0.41	0.28

Feature Importance

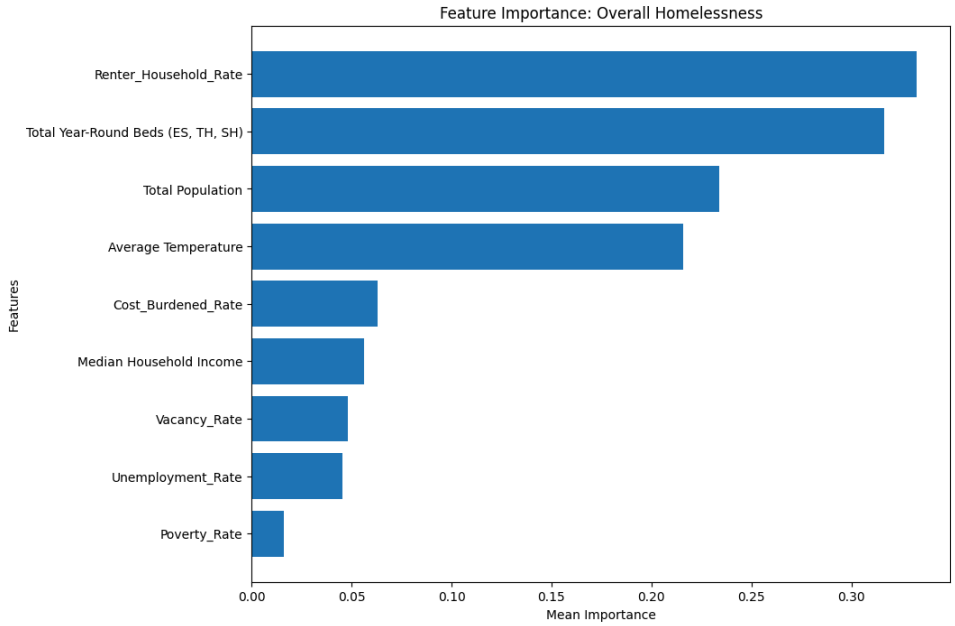
Client Question 2: What community-level measures are most important in predicting the number of persons experiencing homelessness in a community?

Since the Random Forest model performed the best in predicting homelessness across all models under consideration, we focus on its feature importance to understand the key drivers of homelessness. The feature importance analysis reveals which community-level factors, such as Median Gross Rent and Poverty Rate, most strongly influence predictions. This insight can guide targeted interventions to address the root causes of homelessness, particularly in areas with high rent burdens and economic vulnerability.

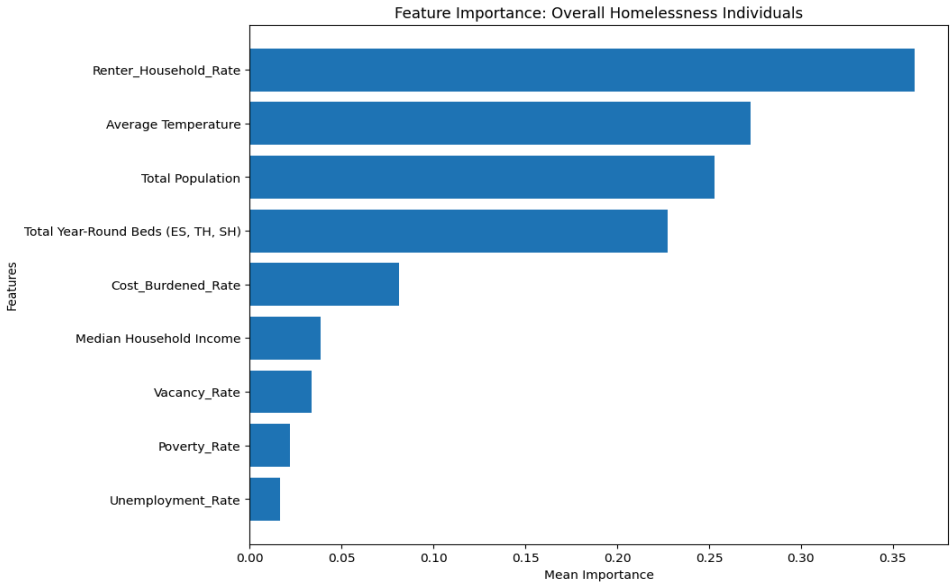
BOSTON UNIVERSITY

1. Top community-level determinants of Overall Homelessness

This plot highlights the most critical features for predicting the total number of people experiencing homelessness in a community. Median Gross Rent emerges as one of the highest-scoring features, underscoring the strong association between higher rent costs and increased homelessness rates. The Poverty Rate is another significant predictor, emphasizing the economic vulnerability of communities as a major driver of homelessness. Additionally, Total Year-Round Beds reflects the impact of shelter availability on homelessness counts, suggesting that communities with more shelter options may experience lower overall homelessness rates. These findings imply that policies aimed at controlling rent costs, expanding shelter capacity, or supporting low-income households could be effective strategies for reducing homelessness on a broader scale.

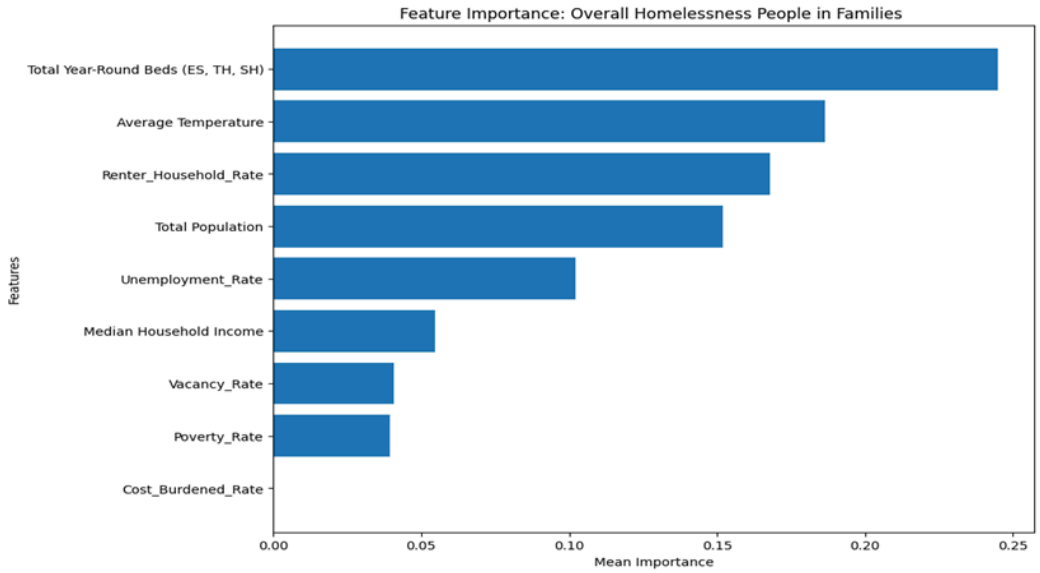


2. Top community-level determinants of Homelessness in Individuals



This plot highlights the most influential factors for predicting homelessness among individuals, distinct from families or other subgroups. A key determinant is the Unemployment Rate, which underscores the strong connection between job availability and individual homelessness rates. Similarly, Median Household Income emerges as an important feature, indicating that communities with lower income levels tend to experience higher rates of individual homelessness. Additionally, Average Temperature may play a role by influencing seasonal variations, as milder climates are often associated with higher rates of unsheltered homelessness. These findings suggest that addressing unemployment and fostering economic stability are critical strategies for reducing homelessness among individuals.

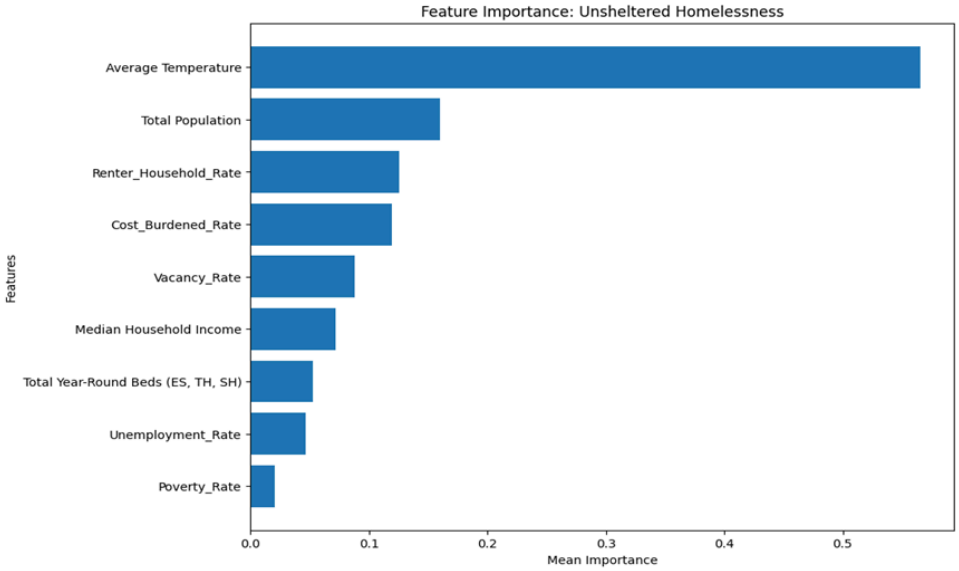
3. Top community-level determinants of Homelessness in Families



This plot highlights the key community-level measures that predict homelessness among family units, offering valuable insights into the factors that drive this issue. One of the most critical predictors is the Cost-Burdened Rate, which underscores the heightened risk of homelessness for families spending more than 30% of their income on housing. Additionally, the Renter Household Rate suggests that communities with a higher proportion of renter households may face increased rates of family homelessness, likely due to the instability associated with renting. Another significant factor is the Total Year-Round Beds, which reflects the importance of housing accommodations in mitigating family homelessness; greater availability of such resources is associated with lower rates of homelessness among families. These findings highlight the importance of enhancing affordable housing programs and rental assistance initiatives as effective strategies to prevent family homelessness and promote housing stability for vulnerable households.

BOSTON UNIVERSITY

4. Top community-level determinants of Unsheltered Homelessness

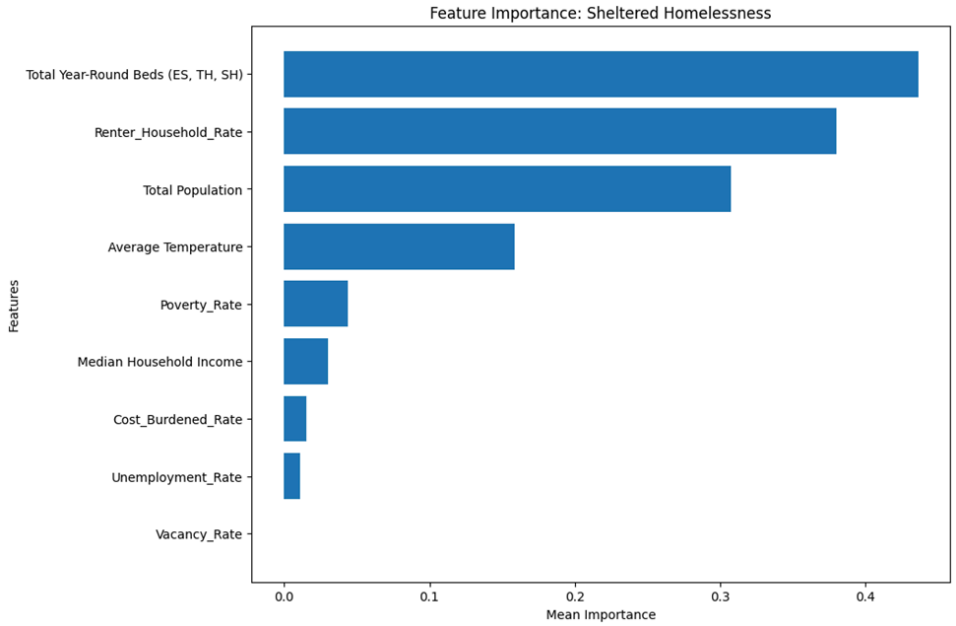


This plot highlights the key factors that most strongly predict unsheltered homelessness, offering insights into the structural and environmental conditions contributing to this issue. Vacancy Rate emerges as a notable predictor, suggesting that higher vacancy rates, while indicative of available housing, may also signal barriers to accessing these options, such as affordability or eligibility requirements. Unemployment Rate further underscores the role of economic factors, as limited job opportunities increase the likelihood of individuals living without shelter. Additionally, Weather Conditions, particularly average temperature, reveal that warmer climates may be associated with higher rates of unsheltered homelessness, as individuals in these regions might be less reliant on shelters during mild weather. These findings imply that targeted strategies, such as expanding job training programs to reduce unemployment and implementing policies to improve access to vacant housing, could play a critical role in mitigating unsheltered homelessness.

5. Top community-level determinants of Sheltered Homelessness

This plot illustrates the features that are most critical for predicting homelessness among individuals staying in shelters. One of the top predictors is Median Gross Rent, which suggests that higher rent burdens in a community can drive more people to seek shelter services due to housing unaffordability. Another significant factor is the Total Year-Round Beds (ES, TH, SH), indicating that the availability of shelter beds directly impacts the number of individuals experiencing sheltered homelessness, as limited capacity can

constrain access to shelter. Additionally, the Renter Household Rate emerges as a key predictor, highlighting that communities with a higher proportion of renter households may experience increased sheltered homelessness, potentially due to the instability often associated with renting. These insights emphasize the importance of expanding shelter capacity and implementing rental assistance programs to support individuals at risk of becoming sheltered homeless.



Conclusion

This project aimed to predict homelessness at the community level, utilizing data spanning from 2007 to 2023 and drawing from a variety of sources. By integrating extensive datasets such as U.S. Department of Housing and Urban Development's Point-in-Time counts, Housing Inventory Count, American Community Survey, and weather data from National Oceanic & Atmospheric Administration, we created a high-quality, multi-dimensional dataset. This dataset was aggregated to the Continuum of Care level, addressing challenges like missing values and geographical mismatches. As a result, it provided a solid foundation for training predictive models, allowing for a more accurate understanding of the factors contributing to homelessness in different communities.

The project led to the development of strong predictive models, with the Random Forest model emerging as the best performer. This model achieved superior results, particularly in predicting family homelessness rates, with an RMSE of 0.37 and an R^2 of 0.82. It demonstrated

comprehensive coverage, effectively predicting multiple dimensions of homelessness, including overall homelessness, family homelessness, and the distinction between sheltered and unsheltered homelessness. This adaptability and reliability underline the model's potential for informing policy decisions.

Through feature engineering and data analysis, the project identified several key community-level factors influencing homelessness. Renter household rates and total year-round beds were found to be the strongest predictors, emphasizing the importance of housing supply and demand in shaping homelessness trends. Population size and average temperature were additional influential factors, particularly for understanding regional and seasonal patterns in homelessness. Median household income, unemployment rates, and poverty rates emerged as secondary predictors, offering insights into the socioeconomic dynamics driving homelessness.

The project uncovered important trends and patterns related to homelessness. Regional variations highlighted that communities with higher rent burdens and unemployment rates tended to have significantly higher homelessness rates, particularly in densely populated urban areas like New York and Los Angeles. Furthermore, policy impacts were observed, with emergency shelter policies effectively mitigating short-term increases in homelessness during crises. However, long-term challenges remain linked to housing supply shortages, underscoring the need for sustained efforts to address the root causes of homelessness. These findings underscore the importance of integrating diverse data sources and applying advanced machine learning models like Random Forest to enhance homelessness prediction and inform data-driven policy-making.

Policy and Practical Recommendations

Based on our findings, we propose the following recommendations to address homelessness more effectively:

1. **Policy Interventions:**

   First, it is crucial to prioritize high-risk communities by allocating additional resources to areas with high rent burdens and unemployment rates. These communities should receive tailored housing and economic support programs to mitigate the risk of homelessness. Second, implementing rent stabilization measures, such as rent control and subsidies, can help alleviate housing cost pressures on vulnerable populations, making housing more affordable. Finally, enhancing shelter capacities is vital, particularly in high-density areas, by expanding year-round bed availability and accounting for weather conditions to better address seasonal demands for shelter.

2. **Model Optimization and Applications:**

To optimize the predictive models and improve their effectiveness, several strategies can be employed. First, incorporating dynamic variables, such as real-time weather conditions and economic fluctuations, could enhance the model's timeliness and accuracy by reflecting current trends that influence homelessness. Additionally, ensemble approaches should be explored, where the strengths of different models like Histogram-Gradient Boosting and Random Forest or Support Vector Regression are combined, increasing the model's robustness and overall predictive power. Finally, multi-objective optimization could be pursued by developing tailored models that not only predict overall homelessness rates but also address subgroup-specific outcomes, such as family homelessness or unsheltered populations, to ensure more precise and actionable predictions. These strategies can significantly improve model performance and provide more targeted insights for policy and intervention planning.

3. **Data-Driven Long-Term Strategies:**

Data-driven long-term strategies are essential for addressing homelessness effectively. Efficient resource allocation can be achieved by using predictive models to identify high-risk communities, ensuring that resources are distributed in a way that addresses the most pressing needs, both spatially and temporally. Additionally, policy impact monitoring plays a crucial role in refining intervention strategies. By comparing the predictions made by the models with actual outcomes, policymakers can evaluate the effectiveness of their actions, make necessary adjustments, and close the "data-policy-feedback" loop. This continuous process of evaluation and refinement helps to ensure that homelessness interventions are both timely and impactful.

This project advances the understanding of homelessness by focusing on structural determinants at the community level, bridging a critical gap in existing individual-level predictive models. By uncovering the drivers of homelessness, it equips policymakers and social service organizations with precise tools and actionable insights. The continued refinement and application of these models have the potential to significantly impact homelessness reduction efforts, ensuring a more stable future for vulnerable populations.