Analysis of Correlates to Homelessness in the United States

Final Progress Report

by

L Srihari Boppana

Table of Contents

List of Figures

List of Tables

1. Introduction

Homelessness in the United States has escalated significantly over the past decade, prompting urgent concern among policymakers, researchers, and advocacy organizations. According to the U.S. Department of Housing and Urban Development's (HUD) 2024 Annual Homelessness Assessment Report, more than 770,000 individuals were experiencing homelessness on a single night in January 2024 marking an 18% increase from the previous year. This surge is the largest single-year increase on record, underscoring the growing challenges faced by economically vulnerable populations. Homelessness among families with children has been particularly impacted, rising by 39%, reflecting a worsening crisis in housing stability among low-income households [National Low Income Housing]. In contrast, homelessness among veterans has continued to decline, decreasing by 8%, due in part to targeted federal assistance programs, increased investment in permanent supportive housing, and VA-sponsored housing vouchers [Veteran].

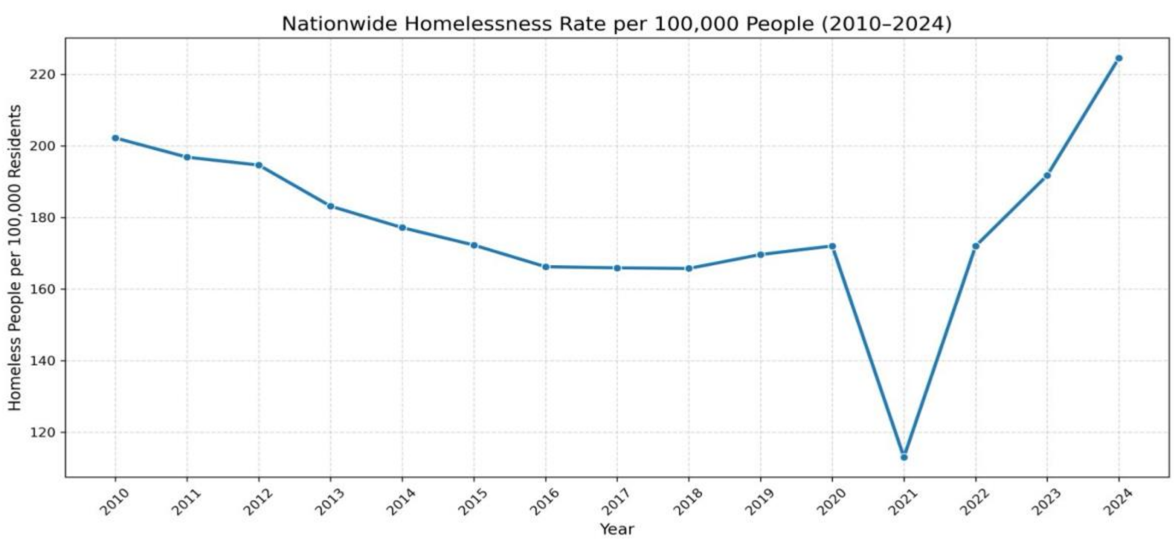

Figure 1.1: Trend of Overall Homelessness in USA(2010-2024)

As shown in Figure 1, the national homelessness rate followed a declining trend between 2010 and 2016, reflecting efforts made during the post-recession recovery period, including investments in housing-first policies and federal programs like the Emergency Solutions Grants (ESG) and Continuum of Care (CoC) programs (HUD). However, beginning in 2017, this progress began to stall due to factors such as rising rent burdens, limited affordable housing supply, and shifts in policy priorities (Harvard). The rate dropped sharply in 2020, which is largely attributed to incomplete data collection during the COVID-19 pandemic, particularly in counts of unsheltered homelessness (HUD). Many Point-in-Time (PIT) counts were either scaled back or canceled due to health and safety restrictions, leading to data gaps. However, despite this artificial decrease in reported homelessness, the pandemic exacerbated housing instability. By 2021, as data collection resumed, homelessness rates rose sharply, culminating in 2024 reaching the highest level in over a decade.

A growing body of research attributes this rise to the widening gap between housing costs and household incomes [Harvard]. In 2022, the median sale price for a single-family home in the U.S. was 5.6 times higher than the median household income—the highest ratio ever recorded. This housing affordability crisis disproportionately affects low-income renters, many of whom are spending more than 50% of their income on rent and are at high risk of eviction [Eviction]. By 2023, nearly half of all renter households—over 21 million—were spending more than 30% of their income on housing costs, classifying them as housing cost-burdened [Harvard]. Research suggests that even a $100 increase in median rent is associated with a 9% increase in homelessness rates, demonstrating how even modest housing cost fluctuations can have severe consequences for vulnerable populations [EPI].

While structural factors like housing affordability are critical, individual vulnerabilities, including mental health and substance use disorders, also play a significant role in contributing to homelessness. For example, studies have shown that homeless youth face barriers to employment, such as lack of education and skills, unstable housing, stigma in the workplace, mental health problems, substance use problems, and incarceration. [PMC] Studies indicate that up to 25% of the homeless population suffers from serious mental illness, while 38% experience substance use disorders (samhsa).

Given the multifactorial nature of homelessness, it is crucial to approach the problem through a comprehensive lens that considers both systemic and individual determinants. This project aims to analyze the drivers of homelessness in the United States using a data-driven approach, leveraging large-scale national datasets and advanced analytical techniques to uncover trends, assess key contributing factors, and support evidence-based policymaking. By integrating machine learning models, predictive analytics, and statistical modeling.

In addition to examining trends within the U.S., this study incorporates a comparative analysis with Ireland. Ireland presents a valuable case study due to its structured national homelessness reporting system and parallel challenges with rising housing costs, economic instability, and housing shortages (Housing). Ireland implements a centralized homelessness tracking system, Pathway Accommodation and Support System (PASS), which provides real-time data on emergency accommodations and service use. By comparing homelessness trends, population, crime statistics and economic indicators across both nations.

This research sets the foundation for building predictive models, evaluating regional differences, and generating actionable insights to support long-term solutions for homelessness in the United States and beyond.

2. Objectives

Understanding and addressing homelessness in the United States requires a data-driven approach to uncover key trends, analyze influencing factors, and develop predictive insights. One of the main challenges in studying homelessness is the complexity of its causes, which range from economic instability and rising housing costs to policy decisions and social factors. Furthermore, inconsistencies in data collection methods across states and localities make it difficult to develop a unified framework for analysis. By leveraging data science methodologies, this project aims to provide a clearer understanding of the drivers of homelessness and offer actionable insights for policymakers. To enhance the scope of the analysis, the project also incorporates a comparative study with Ireland, examining how homelessness in a similar high-income country is affected by policy, economic conditions, and social factors. This cross-national perspective will provide valuable context and highlight alternative intervention strategies.

This project will achieve several technical and research objectives. It will involve the collection, integration, and analysis of large-scale datasets from sources such as the U.S. Department of Housing and Urban Development (HUD), the American Community Survey, and other relevant databases. Advanced statistical and machine learning techniques will be applied to model homelessness trends, identify key contributing factors, and predict future shifts in homelessness rates. The study will also explore the relationship between economic indicators (such as unemployment rates, median income, and inflation) and homelessness, with a focus on identifying patterns that could inform more effective policies.

Additionally, the project will examine regional variations in homelessness by analyzing economic disparities, state-level policy interventions, and access to affordable housing. The study will also explore the role of public funding allocations and social services in mitigating homelessness, identifying best practices that could be scaled or adapted to different regions.

In terms of individual learning objectives, this project will provide hands-on experience in the full data science pipeline, including data ingestion, cleaning, exploratory data analysis (EDA), feature engineering, and predictive modeling. It will enhance proficiency in Python, R, and SQL for data analysis and machine learning. Moreover, the project will develop skills in effective data storytelling—translating technical findings into actionable insights that can be understood by nontechnical audiences. By engaging in this research, I aim to bridge the gap between data science and public policy, ensuring that evidence-based strategies are at the forefront of efforts to combat homelessness.

3. Data Data Sources Overview

In total, eight datasets were used for this project. The 2007-2024-PIT-Counts-by-State datasets were sourced from the U.S. Department of Housing and Urban Development (HUD), which tracks homelessness statistics over time through various data collection methods. These datasets provide a comprehensive view of trends in homelessness across different states, allowing for an in-depth analysis of patterns over nearly two decades.

HUD collects homelessness data primarily through the Point-in-Time (PIT) Count. The PIT Count is an annual survey conducted every January, providing a snapshot of homelessness by counting individuals experiencing sheltered and unsheltered homelessness on a single night. Local Continuums of Care (CoCs) are responsible for organizing and executing these counts, often relying on volunteers and outreach teams to gather data.

In 2020, homelessness was not recorded as accurately or comprehensively as in previous years due to disruptions caused by the COVID-19 pandemic. Many communities were unable to carry out the full Point-in-Time (PIT) Count, especially the unsheltered component, because of health and safety concerns for both volunteers and vulnerable populations. As a result, large portions of the country lacked consistent data on individuals experiencing homelessness.

For this project, I am using PIT homelessness data from 2010 to 2024, which allows for a longterm analysis of trends in both sheltered and unsheltered homelessness across the United States. While the 2020 data is included for continuity, it is important to acknowledge its limitations due to the exceptional circumstances of the pandemic.

This project utilizes data collected from the American Community Survey (ACS) and related public datasets spanning from 2010 to 2024. The ACS, conducted by the U.S. Census Bureau, provides annual, detailed information on housing, population, poverty, and socioeconomic indicators across the United States. It collects data through a combination of mail surveys, telephone interviews, and in-person visits, sampling over 3 million households each year. The ACS serves as a critical resource for understanding trends in rental affordability, household demographics, poverty levels, housing occupancy, and access to subsidized housing programs. This comprehensive data allows for in-depth analysis of housing insecurity and the structural factors contributing to homelessness over time.

While the dataset provides broad coverage, certain years contain missing or incomplete information. In 2020, data for variables such as poverty rate, total households, total occupied housing, and various categories of housing vacancy were unavailable or inconsistently reported, likely due to the disruptions caused by the COVID-19 pandemic. Similar gaps are observed in 2024 for several housing and income-related features, including median wages, rent data, and vacancy indicators. Housing vacancy data is also missing for 2018. These data gaps should be taken into account when conducting time-series analyses or drawing comparisons across years, as they may introduce limitations in interpreting certain trends.

Opioid overdose data in this project was sourced from the Kaiser Family Foundation (KFF) it collects opioid overdose statistics from official public health sources, including the Centers for Disease Control and Prevention (CDC) and state health departments. The data reflects reported overdose deaths across states, broken down by age group and year. In addition to raw opioid overdose counts, the dataset also includes a calculated feature, Drug overdose data, which measures overdose deaths relative to population size to allow for more accurate comparisons between states. However, there are inconsistencies in the availability of this data. Specifically, both opioid_overdose and drug_overdose are only avaiable till year 2022 and data for the years 2023 and 2024 is missing, and one value is missing for 2012. These inconsistencies likely stem from delays in state reporting, incomplete submissions, or differences in classification standards. While opioid-related data adds critical context to the intersection of public health and housing insecurity, these gaps should be acknowledged when interpreting long-term trends

Data Cleaning and Preprocessing

While the raw data was accessible from these sources, significant preprocessing and transformations were required before analysis. The following sections outline the specific modifications applied to ensure consistency and usability within the project.

After careful cleaning and transformation, a final preprocessed dataset was developed for analysis. The cleaned dataset consists of 783 rows and 29 columns, representing a structured integration of HUD PIT data, American Community Survey (ACS) indicators, and public health data. This final dataset includes both raw counts and derived features designed to allow meaningful cross-state and year-over-year comparisons.

To enable better comparisons across regions with different population sizes, a number of per capita variables were calculated. These per capita features provide a more standardized view of housing and public health indicators, allowing for better analysis across states with varying population sizes.

A comprehensive data quality assessment was conducted on the final dataset to ensure consistency and accuracy for analysis. The dataset was confirmed to have no duplicate rows, which validates the uniqueness of each state-year entry. However, missing values were identified across multiple features, primarily due to inconsistencies in data reporting from source agencies, as previously discussed. These gaps are most likely the result of delayed submissions, non-standardized reporting procedures across states, or disruptions in data collection during certain years, such as the COVID-19 pandemic period.

To maintain the integrity of the analysis, missing values were not imputed. This decision was made to avoid introducing artificial bias and to ensure that the results reflect only the observed data. Additionally, during preprocessing, several formatting inconsistencies were addressed—particularly in columns that were incorrectly interpreted as strings due to formatting artifacts (e.g., commas or special characters in numeric fields). These values were properly converted into the correct data types, such as converting strings to numeric or float formats, ensuring the dataset was clean and ready for accurate and meaningful analysis.

Feature Engineering

To make the data more meaningful and suitable for analysis, I created new features based on existing columns. These help us compare homelessness and housing conditions across states more fairly.

To make cross-state comparisons more meaningful, I created a per capita homelessness metric. This standardizes homelessness figures relative to the population size of each state, ensuring fairer comparisons regardless of how large or small a state is.
I calculated Overall Homelessness Per Capita using the following formula:

$$\text{Homelessness Per Capita} = \frac{\text{Total Homeless Population}}{\text{Total Population}}$$

Homelessness vs. Median Rent (2023)

The scatter plot below visualizes the relationship between median rent and homelessness per 100K people in different states.

Each blue dot represents a state. The red dashed line is the trend line, which shows the general pattern that homelessness increases as rent increases. The shaded area around the trend line represents the confidence interval, meaning there is some variation in the relationship. Most states with rent below $1,000 have lower homelessness rates. High-rent states like California, Hawaii, and New York have some of the highest homelessness rates. Some states, like Vermont and Alaska, have high homelessness rates despite not having the highest rent, meaning other factors could be involved.

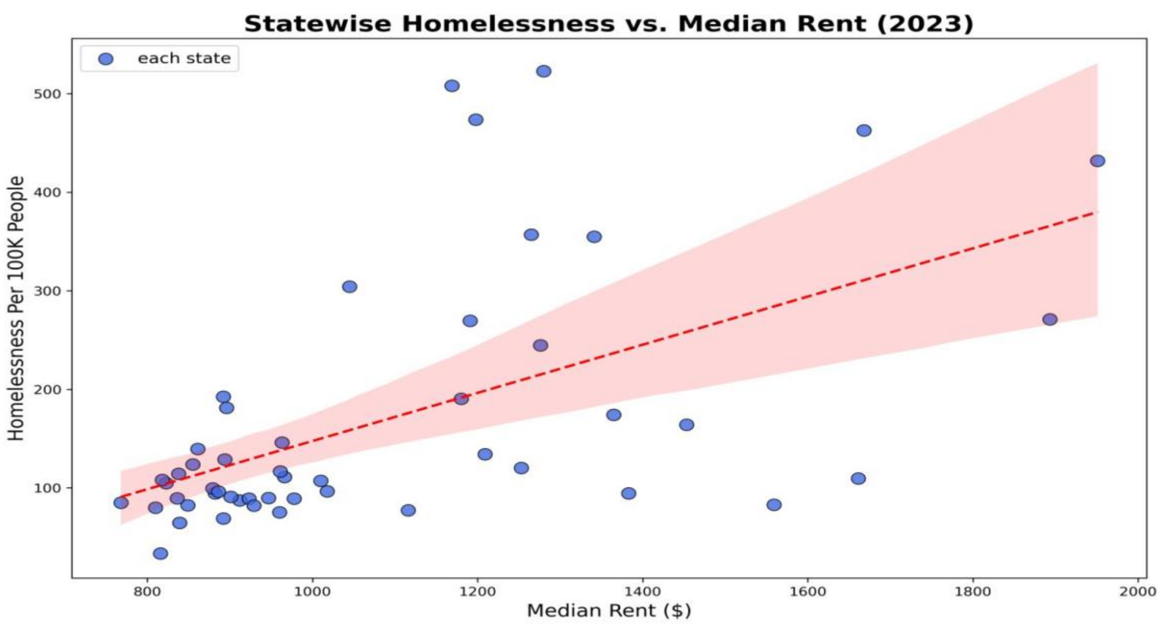

Figure 3.1: Statewise Homelessness vs. Median Rent 2023

The goal of this analysis is to understand the relationship between median rent and homelessness per capita in different U.S. states. Housing affordability is often linked to homelessness, so this analysis looks at whether states with higher rent also tend to have more people experiencing homelessness.

To determine if there is a relationship between Median Rent ($) and Homelessness Per 100K People, I calculated the Pearson Correlation Coefficient (r). Pearson correlation measures the linear relationship between two variables. It tells us:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \times \sqrt{\sum(Y_i - \bar{Y})^2}}$$

Where:

- X_i and Y_i are the individual values of Median Rent and Homelessness Per 100K, respectively.
- X' is the mean of Median Rent, and Y' is the mean of Homelessness Per 100K.
- The numerator calculates the covariance, measuring how the two variables change together.
- The denominator normalizes this by the standard deviation of both variables, ensuring that the correlation is always between -1 and 1.

r > 0: A positive correlation (when one variable increases, the other also tends to increase).
r < 0: A negative correlation (when one variable increases, the other tends to decrease). r = 0: No correlation (no relationship between the variables).
Pearson Correlation: 0.55

This means there is a moderate positive correlation between rent prices and homelessness rates. In simple terms, states with higher rent tend to have higher homelessness rates, but other factors also influence homelessness. According to a report by the Pew Charitable Trusts, increases in rent directly contribute to rising homelessness rates, with areas experiencing high rental costs also witnessing elevated levels of homelessness. This relationship underscores the significant impact of housing affordability on homelessness trends. Source: [pew]

State-Level Distribution of Homelessness Rates (2024)

To better understand the geographical distribution of homelessness across the United States, I visualize the overall homelessness rate per 100,000 people for each state in 2024. This choropleth map categorizes states into four tiers: low (<50), moderate (50–150), high (150–300), and very high (>300) rates of homelessness.

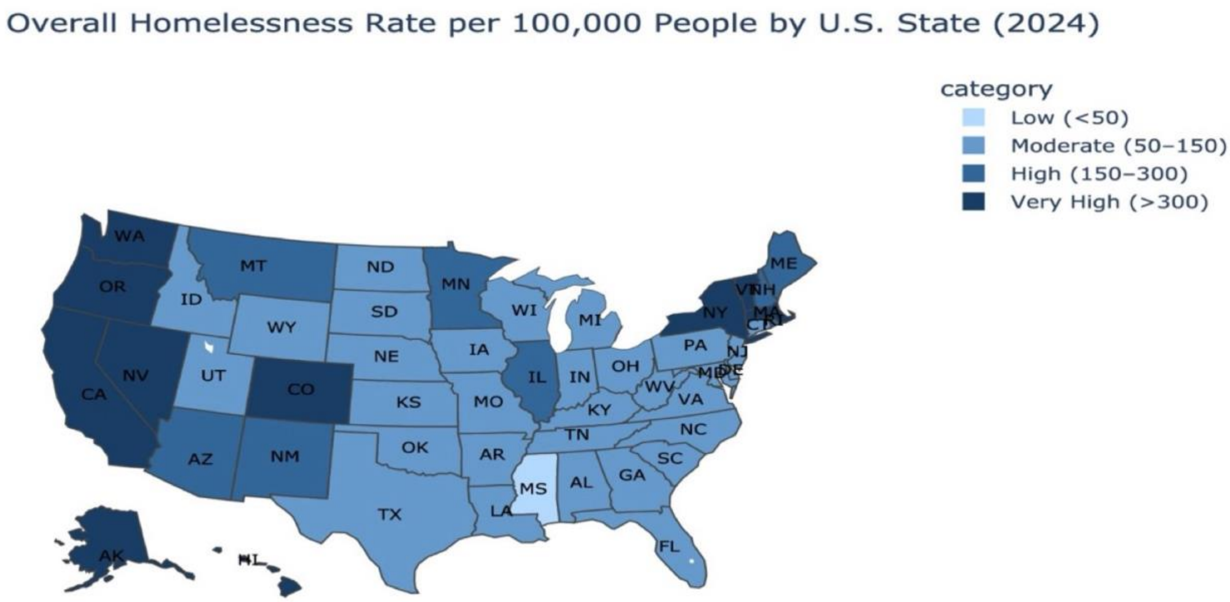

Figure 3.2: Overall Homelessness Rate per 100,000 People by U.S. State 2024

The map clearly illustrates significant regional disparities in homelessness. States along the West Coast—including California, Oregon, Washington, Alaska, and Hawaii—fall into the very high category, reflecting the impact of high housing costs, urban population densities, and a shortage of affordable housing units in these areas. Similarly, several Northeastern states such as New York, Massachusetts, Maine, and Vermont also exhibit very high homelessness rates.

In contrast, states in the South and Midwest, such as Mississippi, Louisiana, Texas, and Kentucky, fall into the low or moderate categories. These variations may reflect differences in housing costs, economic conditions, urbanization levels, and state-level policy interventions.

Homelessness Trend per State (2010 - 2024)

The map illustrates the homelessness trend across U.S. states between 2010 and 2024, categorized into two groups: increasing and decreasing trends. States marked in green represent those that experienced a decrease in homelessness per capita over the 15-year period, while states shown in red indicate an upward trend in homelessness rates.

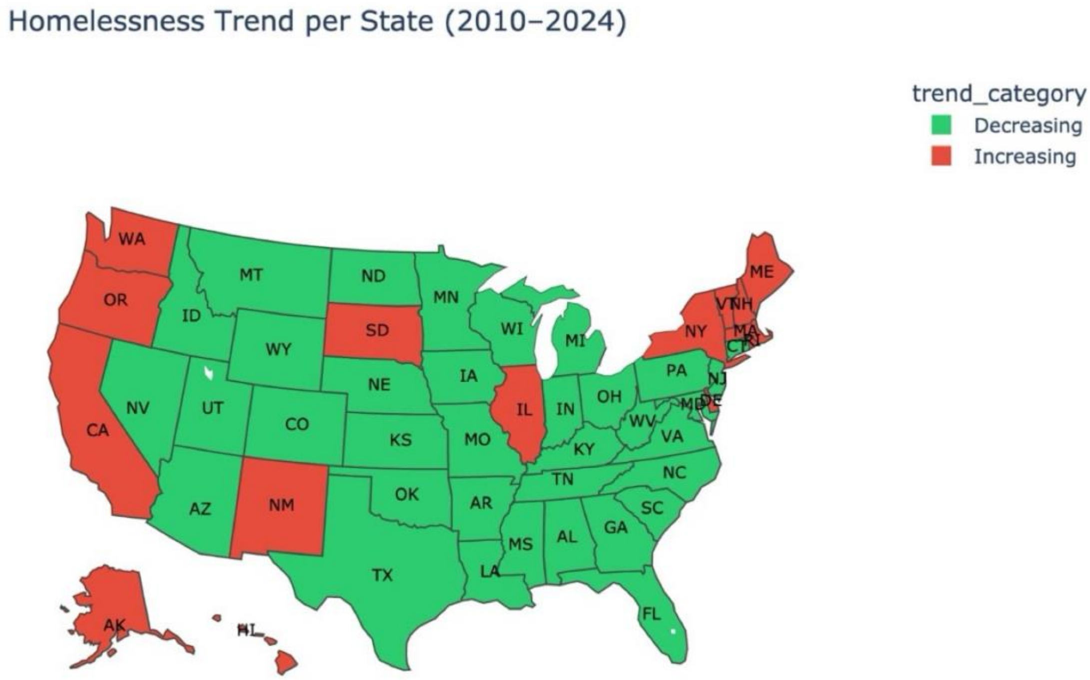

Figure 3.3: Homelessness Trend per State 2010 - 2024

A majority of states, particularly in the Midwest, South, and parts of the West, show a decreasing trend. In contrast, several states along the West Coast, Northeast, and a few in the central and southwestern regions display increasing trends. This visualization helps identify which states have made relative progress in reducing homelessness and which have seen growth in the issue over time. It provides a state-level overview of long-term changes in homelessness patterns.

State-Level Extremes in Homelessness Rates (2024)

To better understand regional disparities, the bar charts below illustrate the top 10 and bottom 10 U.S. states by homelessness per capita for the year 2024. These visualizations highlight the significant geographic variation in homelessness rates across the country.

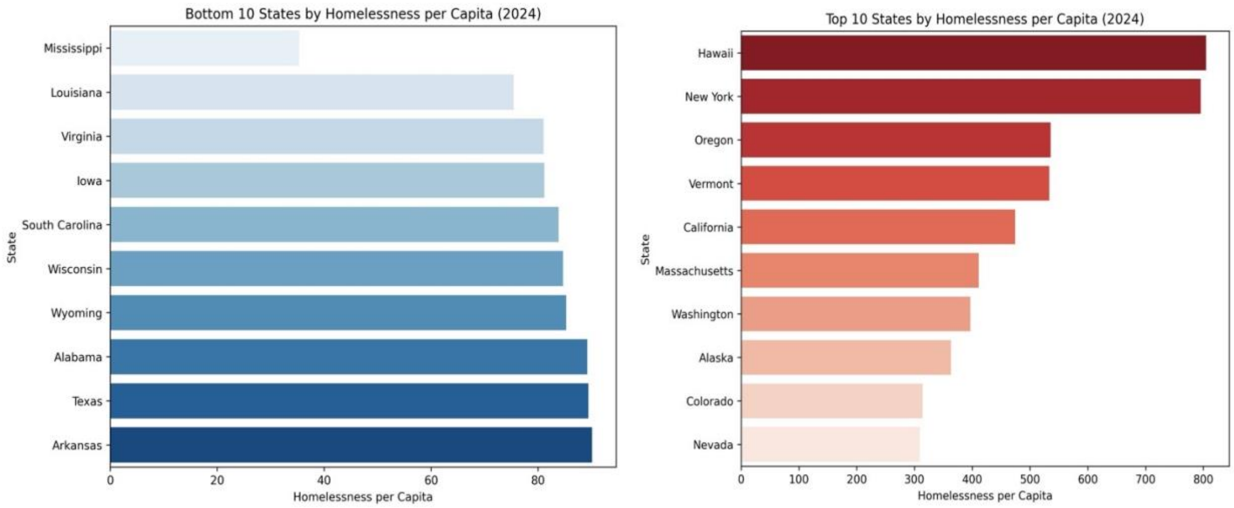

Figure 3.4: State-Level Extremes in Homelessness Rates (2024)

The top 10 states with the highest homelessness rates include Hawaii, New York, Oregon, and California states. Hawaii leads the nation with the highest per capita homelessness rate, followed closely by New York.

In contrast, the bottom 10 states with the lowest homelessness per capita such as Mississippi, Louisiana, Virginia, and Iowa. Mississippi recorded the lowest homelessness rate in 2024.

Correlation Analysis: Homelessness, Housing, and Substance Abuse (2010–2024)

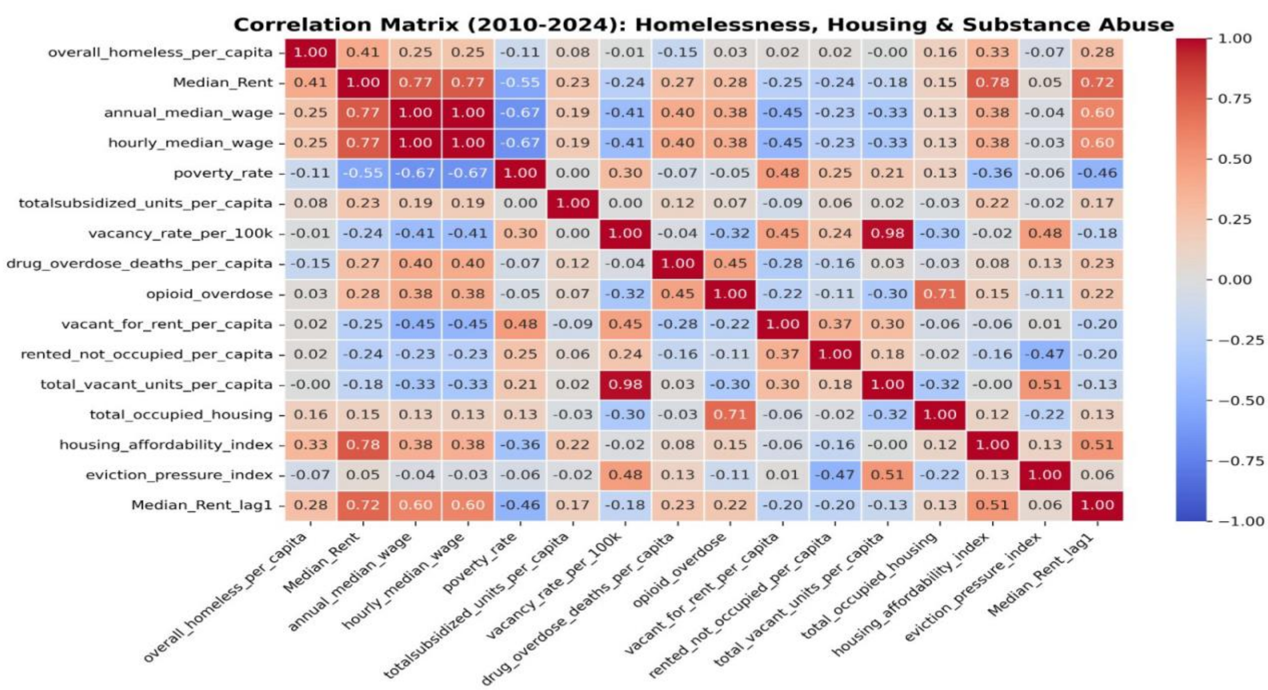

Figure 3.5: Correlation Matrix 2010 - 2024

A correlation analysis was also conducted, and the results are presented in the heatmap above. The target variable overall homelessness per capita—shows a moderate positive correlation with features related to housing cost and affordability, most notably Median Rent (0.41) and Lagged Median Rent (0.28). This supports the hypothesis that rising rental costs contribute to increased homelessness.

Interestingly, the housing affordability index demonstrates a moderate negative correlation (-0.33) with homelessness, suggesting that states with more affordable housing tend to experience lower rates of homelessness. Conversely, poverty rate shows a weak negative correlation (-0.11), indicating that poverty alone may not be a strong predictor of homelessness, a pattern also supported in prior research.

On the other hand, drug overdose deaths per capita and opioid overdose counts exhibit very weak correlations with homelessness (-0.15 and 0.03 respectively), suggesting limited direct linear association in this dataset.

For accuracy and consistency, only years in which all variables were available were included in this analysis. Any year with missing values for even one of the selected variables was excluded to avoid bias or skewed correlations caused by incomplete data.

Overall, the matrix highlights housing cost indicators as stronger correlates of homelessness compared to poverty and public health metrics. These insights guide variable selection for predictive modeling and emphasize the structural roots of homelessness.

Comparative Analysis with Ireland

To gain broader insight into homelessness as a global issue, this project extends its analysis by comparing homelessness trends in the United States with those in another developed nation, Ireland. The dataset compiled for this comparison covers the years 2014 to 2024 and includes key indicators such as total homeless population, national population, median gross household income (in euros), and annual crime counts for Ireland. These metrics are directly compared against similar variables for the United States over the same time period.

Ireland collects homelessness data primarily through the Department of Housing, Local Government and Heritage, which publishes monthly and annual reports on the number of individuals accessing state-funded emergency accommodation. These figures are gathered in collaboration with local housing authorities and service providers under the Pathway Accommodation and Support System (PASS). This system tracks individuals in real time, providing comprehensive data on homeless adults and families across various age groups and accommodation types.

The majority of the Ireland homelessness and housing data used in this project was obtained from the official government portal: https://www.gov.ie/en/collection/80ea4-homelessness-data/. Additional supporting data, such as Ireland's population, crime statistics, and median household income, were sourced from publicly available platforms including government websites and trusted aggregators like Google's public data sources. This combined approach enabled the construction of a well-rounded dataset for comparative analysis.

These datasets will be integrated and analyzed to identify patterns, develop predictive models, and generate data-driven policy recommendations to address homelessness effectively.

To visually represent this comparison, the graph below illustrates the homelessness rate per 100,000 population for both the United States and Ireland from 2014 to 2024. The red line represents the U.S. homelessness rate, while the blue line reflects Ireland's. As shown, the United States maintained a consistently higher rate of homelessness across the entire period. The U.S. trend remained relatively flat from 2014 to 2019, followed by a significant dip in 2021 due to pandemic-related disruptions in data collection. However, rates surged again from 2022 to 2024, reaching the highest level observed in over a decade.

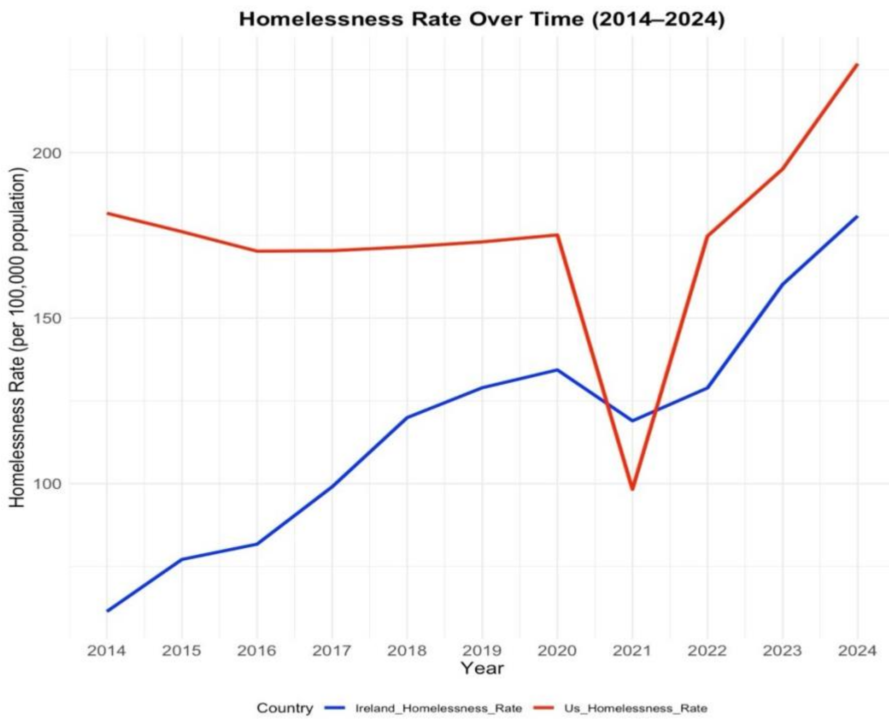

Figure 3.6: Homelessness Rate Over Time in Ireland and U.S. 2014 - 2024

In contrast, Ireland's homelessness rate steadily increased from 2014 to 2020 and A modest decline is visible in 2021, corresponding to the period when emergency eviction bans and rent freezes were enacted by the Irish government during the COVID-19 pandemic (Ireland) However, rates began climbing again after these temporary protections were lifted.

By 2024, both countries report their highest homelessness rates in over a decade, according to their respective datasets. This visual comparison provides context for understanding how different policy environments and national responses to crises such as COVID-19 can influence homelessness trends over time.

Homelessness Rate Vs Median Income (Ireland and U.S.)

This dual-axis line chart illustrates the trends in homelessness rates and median incomes for both the United States and Ireland between 2014 and 2024. The solid lines represent the homelessness rate (measured on the left y-axis as the number of homeless individuals per 100,000 population), while the dashed lines represent the median income levels (measured on the right y-axis, adjusted for purchasing power parity in USD).

From the chart, the U.S. exhibits a relatively stable homelessness rate from 2014 until 2020, followed by a sharp drop in 2021—likely due to pandemic-related reporting issues—and a significant surge from 2022 onward. Median income in the U.S. shows a steady increase throughout the time period.

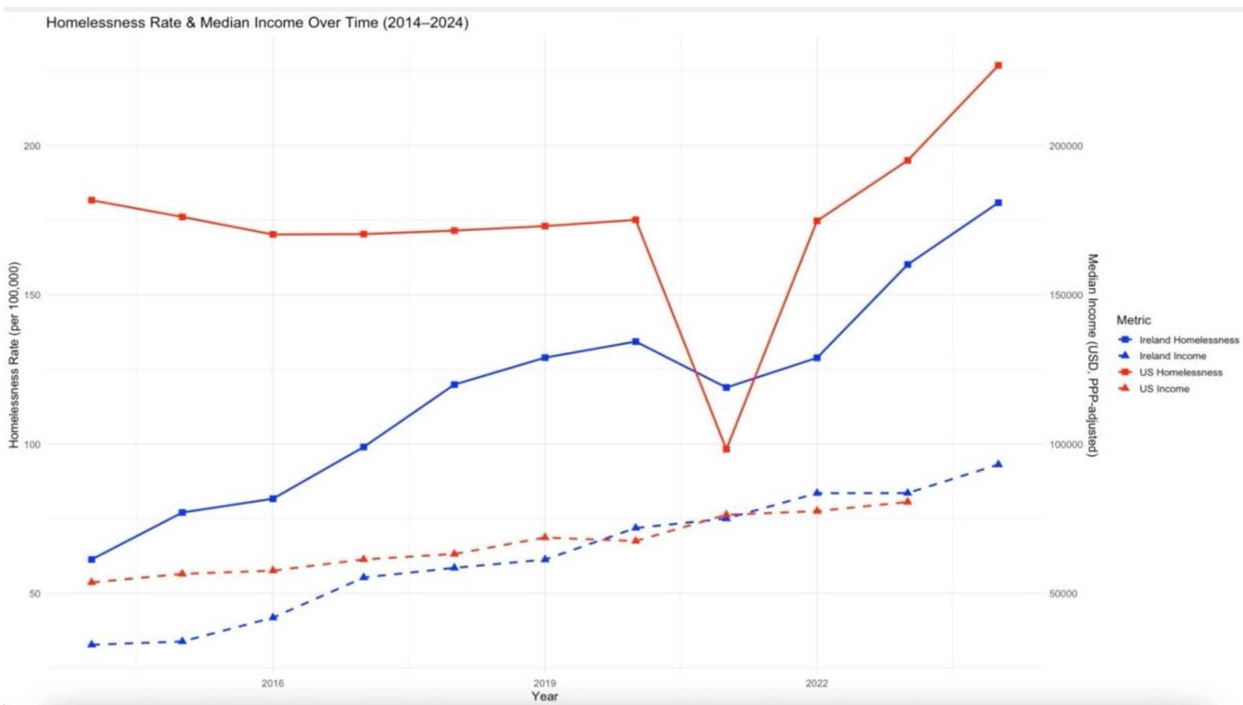

Figure 3.7: Homelessness Rate Vs Median Income (Ireland and U.S.)

In contrast, Ireland shows a consistent upward trend in both homelessness and income levels, with a brief decline in homelessness in 2021. Despite Ireland's lower median income levels compared to the U.S., its homelessness rate has steadily increased over the years. This visualization enables a clear year-over-year comparison between the two countries in terms of both economic and social metrics.

HomelessnessRate Vs Median Monthly Rent (in USD)

The plot examines the median monthly rent in US dollars between two nations, Ireland and the United States, from 2010 to 2023, showing varying patterns during the first few years of the recovery and then a significant divergence. The median rent fluctuated between the two countries from 2010 to 2015, indicating the complex nature of the housing market in the years following the global financial crisis. Due to the collapse of the property bubble and decreased demand for housing, Ireland experienced volatility in the early years of the decade, with "rents falling by 19% between 2008 and 2010 and stabilizing around 2011" due to the collapse of the property bubble and reduced housing demand [European Commission 2020].

However, a noticeable shift occurred after 2015, as Ireland began to experience a sharp increase in rental costs, which continued to accelerate through 2022. “As a result, rents nationwide were an average of 14.1% higher in the third quarter of 2022 than a year previously. That is the highest annual rate of inflation in market rents recorded in the Draft Report since it was launched.” [Rental Report Ireland Q3 2022]. This sudden growth in rent is attributed to multiple interlinked factors, including sustained population growth in urban areas, a persistent shortage of new housing construction, and a delayed recovery of the housing market following the financial crisis. In contrast, while rent in the United States also increased over this period, the pace of expansion was comparatively slower and periodically tempered by the broader economic shock, most notably the COVID-19 pandemic in 2020, which caused rental inflation in several regions to stabilize temporarily. According to the Federal Reserve's historical analysis, "rents began to recover in most markets by late 2010", following a temporary downturn in 2009 caused by weak labor markets and falling household incomes [Federal Reserve History, 2023].

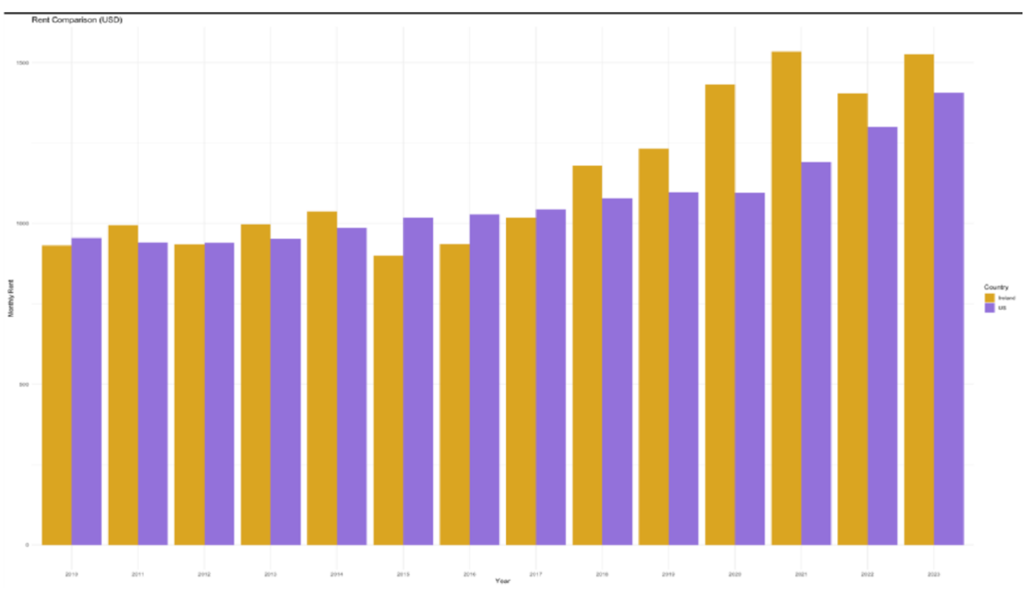

Figure 3.8: Median Monthly Rent Vs Year(Ireland and U.S.)

Ireland's median rent levels started to rise in 2016 and remained higher than the median rent levels of the United States for the following years. This is because of Ireland's growing economy, which has limited housing supply in major cities like Dublin. According to the European Commission, "post-2014 rent increases were driven by rapid economic recovery and worsening supply-demand imbalances" [European Commission 2020]. Meanwhile, median rent increased gradually in the United States, thanks to its broader geographic spread and larger housing inventory.

Unemployment Rate Vs Year

The line graph that contrasts the unemployment rates in the US and Ireland from 2010 and 2023 demonstrates how the two nations' post-2008 financial crisis recovery paths diverged. After reaching a peak of almost 15% in 2012, Ireland's unemployment rate gradually decreased to about 4.3% by 2023. The adoption of active labor market policies, increased labor market participation, and revitalized economic growth were the main drivers of this downward trend. According to the OECD, "by November 2023, Ireland's unemployment rate stood at 4.8%, just below its pre-pandemic level of 4.9% in September 2019" (OECD 2024).

In contrast, the United States saw unemployment fall from nearly 10% in 2010 to 3.5% in 2019, reflecting a prolonged period of labor market expansion. However, the onset of the COVID-19 pandemic led to a sudden and severe disruption. "Unemployment reached 14.8% in April 2020 the highest level since data collection began in 1948", following widespread job losses across sectors ([Federal Reserve History]). The U.S. job market began to recover through stimulus programs and labor reallocation efforts, with unemployment falling back to around 4.1% by December 2024 [TradingEconomics].

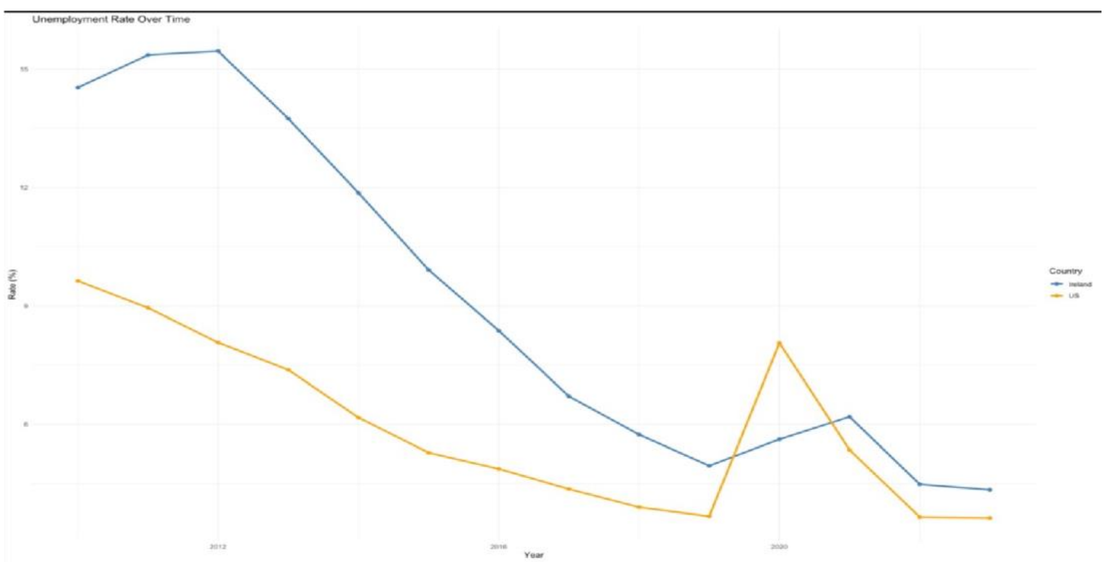

Figure 3.9: Unemployment Rate Vs Year(Ireland and U.S.)

Even though various economic systems and policy reactions influenced them, these trends highlight how resilient both labor markets are. While the United States experience demonstrates how vigorous stimulus and flexible job markets helped mitigate the worst effects of economic shocks, Ireland's slow recovery demonstrates the advantages of innovative employment measures and budgetary stability. When taken as a whole, these patterns provide insightful background information for understanding socioeconomic outcomes like homelessness and household instability.

Crime Rate Vs Year

National crime levels differ significantly and consistently between the United States and Ireland, as seen by the line graph that compares crime rates per capita between the two countries from 2010 to 2023. Per capita crime rates in the U.S. were continuously higher than those in Ireland for the whole period, with a particularly noticeable difference after 2015. A mixed but higher trend in overall crime levels in the United States was shown by the FBI's Uniform Crime Reporting Program, which reports that "in 2022, violent crime decreased 1.7% compared to 2021, while property crime increased by 6.7%" (FBI 2023). There are many reasons for the persistently higher prevalence in the United States, including social inequality, gun accessibility, and wider geographic variance.

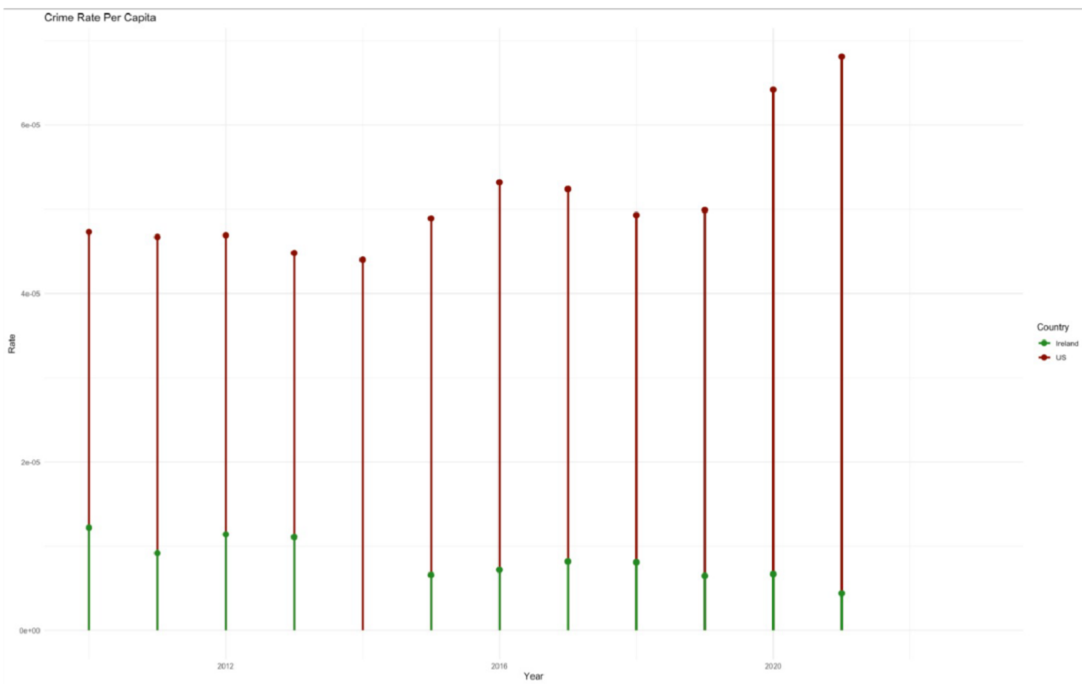

Figure 3.10: Crime Rate Vs Year(Ireland and U.S.)

On the other hand, Ireland maintained consistent and comparatively low crime rates per capita over the same period. Although overall crime rates were lower than in many industrialized countries, the Central Statistics Office of Ireland reported that "theft crime rose by 41% to 65,986 incidents in 2022, which was the highest rate of increase in any crime category."[Recorded Crime Ireland]. Contributing factors include lower population density, effective community policing, and fewer-firearm-related-incidents.

Also, In Ireland, a study by Focus Ireland found that "homelessness did not inevitably lead to criminal behaviour... for some, being homeless led to crime which in turn led to imprisonment. For others, it was being released from prison that led directly to homelessness" [Focus Ireland].

Similarly, in the United States, research has shown that “homeless individuals are 17 times more likely to have been the target of violent crime than those with permanent homes”, reinforcing the notion that homelessness both stems from and increases vulnerability to crime [Wiley]. These findings emphasize the need for integrated housing and criminal justice policies to reduce homelessness.

Poverty Rate Vs Year

The graph compares the poverty rates in the US and Ireland between 2010 and 2023. The at-riskof-poverty rate fluctuated throughout Ireland, but in 2022, it increased noticeably**.** "The at-riskof-poverty rate in 2022 was 13.1%, a 1.5 percentage point increase on the 2021 estimate of 11.6%," (CSO) reports**.** The increasing cost of living and the discontinuation of financial assistance related to COVID-19 are contributing to this issue**.**

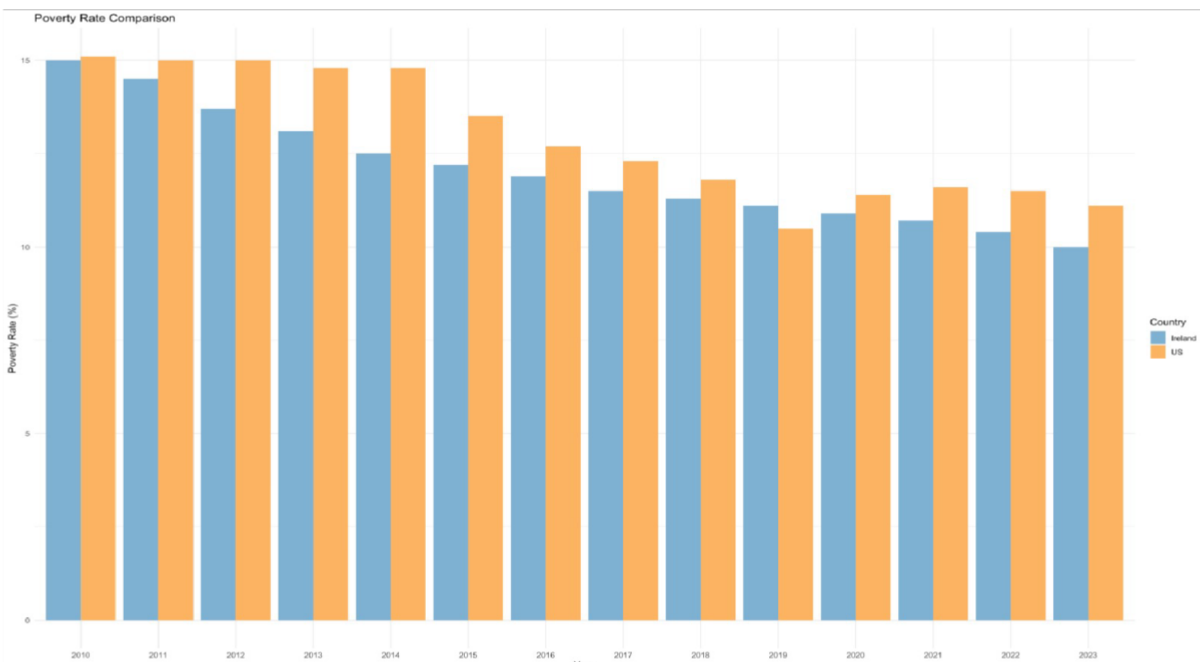

Figure 3.11: Poverty Rate Vs Year(Ireland and U.S.)

The poverty levels in the United States kept gradually declining from 2010 to 2023. In contrast, pandemic-related relief efforts resulted in a notable decline in poverty rates in the United States in 2021. In 2022, however, " The official poverty rate in 2022 was 11.5%, with 37.9 million people in poverty. Neither the rate nor the number in poverty was significantly different from 2021," according to the [U.S. Census Bureau]. This rebound demonstrates the short-term character of the assistance programs and the continuous difficulties in combating poverty.

One of the main causes of homelessness is high poverty rates, which restrict access to secure housing and raise financial stress. Because low-income households experience more housing instability, the National Alliance to End Homelessness claims that "poverty is the most consistent and powerful predictor of homelessness" [(NAEH)]. According to the [ESRI], "housing affordability challenges are particularly acute among low-income households in Ireland," increasing the likelihood that they may become homeless. These observations emphasize the necessity of housing programs that target poverty in both nations.

4. Methodology

4.1 Techniques

This project focuses on developing predictive models to estimate homelessness per capita using economic and housing-related data across the United States and Ireland. Homelessness per capita is defined as the ratio of the total number of homeless individuals in a region to the total population of that region. This formulation allows for meaningful comparisons across geographies with varying population sizes.

Given that our target variable is continuous, this constitutes a regression problem. A variety of machine learning models were utilized, including both linear regression techniques and non-linear ensemble models, to balance interpretability, complexity, and performance. These models were chosen based on their widespread applicability to tabular data and their strengths in managing collinearity, feature importance extraction, and non-linear relationships.

Three types of models were developed:

- A U.S.-only model using state-wise data from 2010 to 2023.
- Comparative models for the U.S. and Ireland, trained with the same feature set and structured year-wise to identify differences in influential variables across the two countries.

Each model went through a structured development process involving data preparation, feature selection, normalization, model training, evaluation, and interpretation.

Data Preprocessing

The dataset used consists of multiple features derived from economic, housing, public health, and infrastructure metrics. Prior to model training, the following preprocessing steps were applied:

- Removal of Non-Predictive Columns: Columns such as state and year were removed from the feature set as they are identifiers rather than predictive attributes.
- Handling Missing Values: An Iterative Imputer from scikit-learn was used to fill in missing values using multivariate feature correlations, enabling better preservation of structural patterns than mean/median imputation.
- Feature Scaling: All features were standardized using z-score normalization (StandardScaler) to ensure fair contribution across models, particularly for algorithms sensitive to distance or regularization.
- Multicollinearity Detection using VIF: To detect multicollinearity among input features, I computed Variance Inflation Factor (VIF) scores. A VIF score above 5 is typically indicative of high multicollinearity. However, in our dataset, all features showed acceptable VIF values (ranging from ~1.08 to ~2.6), so no features were removed.

VIF was calculated for each variable to identify redundancy:

Variance Inflation Factor (VIF):

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

Feature	**VIF**
annual_median_wage	3.167650
poverty_rate	2.282133
Median_Rent	1.558289
drug_overdose_deaths_per_capita	1.543492
opioid_overdose	1.490040
total_vacant_units_per_capita	1.352056
totalsubsidized_units_per_capita	1.080734

Table 4.1: Features VIF Score

4.2 Model Development:

Linear Models:

1. Linear Regression (Ordinary Least Squares)

Linear Regression, also known as Ordinary Least Squares (OLS), serves as the baseline model in our study. This method assumes that the target variable is a linear combination of the input features, plus an error term. The objective of OLS is to find the set of coefficients that minimizes the sum of squared differences between the actual and predicted values.

$$y = \beta 0 + \beta 1 x 1 + \beta 2 x 2 + \cdots + \beta p x p + \epsilon$$

Where:

- y is the homelessness rate per capita,
- xi are the predictor variables (e.g., median rent, unemployment), • βi are the coefficients I want to learn, • ϵ is the error term (residual).

2. Ridge Regression (L2 Regularization):

Many economic indicators in our dataset are correlated (e.g., rent and income). This multicollinearity can destabilize ordinary linear regression. Ridge regression addresses this by introducing L2 regularization, which shrinks the coefficients and improves generalization.

Ridge Regression is a regularized form of linear regression that adds an L2 penalty to the cost function. This penalty shrinks the size of the coefficients but retains all variables in the model. The main advantage of Ridge is its ability to handle multicollinearity**,** which is the presence of highly correlated predictors. In housing and economic datasets, features such as median rent**,** housing vacancy, and income often correlate strongly, making Ridge a valuable tool.

$$\text{Minimize: } [(y_{,} - y\backslash_{.})^{-} + \alpha\ [\ \beta^{-}{}_{/}$$

$Where$:

$y_{,}$ are the actual values, $y_{.}$ are the predicted values,b $\beta_{/}$ are the model coefficients, and α is the regularization strength that controls the amount of shrinkage.

3. Lasso Regression (L1 Regularization):

Lasso Regression (Least Absolute Shrinkage and Selection Operator) is another regularized variant of linear regression, but unlike Ridge, it uses an L1 penalty. The distinguishing feature of Lasso is its ability to shrink some coefficients to exactly zero, effectively eliminating unimportant features from the model.

The L1 penalty in Lasso makes it a powerful tool for feature selection, especially when dealing with datasets that contain many variables — not all of which may be equally informative. As α\alphaα increases, more coefficients are driven to zero, and the model becomes simpler, using only the most predictive features.

In our context, Lasso was instrumental in helping us narrow down the most influential socioeconomic variables associated with homelessness. By tuning α\alphaα carefully, I was able to observe which features consistently contributed to predictions.

$$Minimize: [(y_{\mathrm{i}} - \hat{y}_{\mathrm{i}})^{-} + \alpha\ [|\beta_{j}|$$

$Where$:

$y_{,}$ and $y_{.}$ are actual and predicted values,b $\beta_{/}$ are the model coefficients, and α controls the strength of the L1 penalty that can shrink coefficients to zero.

Nonlinear Models:

1. Random Forest Regressor (Ensemble Learning with Bagging)

Random Forest is an ensemble-based learning algorithm used for regression and classification tasks. In our study, it was selected for its ability to handle non-linear relationships, reduce overfitting, and automatically capture feature interactions , a necessity when modeling complex socio-economic and housing-related factors that influence homelessness.

Rather than building a single predictive model, Random Forest creates an ensemble of decision trees and aggregates their predictions to arrive at a more accurate and stable result.

Random Forest predicts by averaging outputs from K decision trees:

$$yn = \frac{1}{K} [_{023}^{1} \, T_0(x)$$

Where:

yn is the final prediction,

$T_0(x)$ is the prediction from the k-th decision tree, and K is the total number of trees in the ensemble.

2. XGBoost Regressor (Gradient Boosting Framework)

XGBoost (Extreme Gradient Boosting) is a powerful ensemble algorithm based on gradientboosted decision trees. Unlike Random Forest, which builds trees in parallel, XGBoost constructs trees sequentially, where each new tree attempts to correct the residual errors made by the ensemble of previously built trees.

In our implementation, a grid search was used to tune:

- Number of trees (n_estimators)
- Tree depth (max_depth)

- Learning rate (learning_rate)
- Subsample ratio (subsample)

In our study, XGBoost was tuned to learn complex non-linear trends while still controlling for bias and variance. It was particularly helpful in capturing interactions and layered effects among variables like rent, overdose, and subsidized_units.

XGBoost minimizes a regularized loss function:

$$\mathrm{Obj} = \overset{4}{\underset{,23}{[}}\, L(y_{,}, y\backslash_{.}) + \Omega(f)$$

Where:

$L(y_{,}, y\backslash_{.})$ is a loss function (e.g., squared error),

$\Omega(f)$ is a complexity penalty on the model (e.g., number of leaves).

3. K-Nearest Neighbors Regressor (Instance-Based Learning)

KNN is a non-parametric algorithm that makes predictions based on the local neighborhood around a new data point. It does not learn a global function but instead uses distance-based voting (or averaging, in regression) from the most similar observations in the training set.

Hyperparameters tuned via grid search included:

- Number of neighbors (n_neighbors)
- Weighting strategy (weights: uniform vs. distance)

KNN predicts based on the average of k nearest neighbors:

$$yn = \frac{1}{k} \sum_{i \in N_k(x)} y_{,}$$

Where:
yn is the predicted homelessness rate, $N_0(x)$ is the set of k nearest neighbors of x, $y_{,}$ are the corresponding target values.

KNN achieved the highest performance in our study ($R^2 \approx 0.69$ on test set), indicating that local patterns in homelessness data such as similarity between economically distressed regions are more informative than global patterns. It is especially useful when the data distribution is not uniform, and each region behaves differently.

Evaluation Metrics

Three primary evaluation metrics were used to compare model effectiveness:

- R^2 Score (Coefficient of Determination): Indicates the proportion of variance in the target variable that is predictable from the input features. A higher R^2 indicates a better fit to the data.
- Root Mean Squared Error (RMSE): Measures the average magnitude of prediction errors, with higher penalties for larger errors. It is sensitive to outliers and is reported in the same units as the target variable.

- Mean Absolute Percentage Error (MAPE): Represents the average absolute percent difference between actual and predicted values. MAPE is scale-independent and interpretable as a percentage error.
- Mean Absolute Error(MAE): The average absolute difference between predicted and actual values. MAE treats all errors equally and is more robust to outliers than RMSE. Including MAE allows us to compare our models fairly with those used in prior homelessness studies.

4.3 Test Set Evaluation

Each model was evaluated on a reserved 30% hold-out test set, which was never seen during training or cross-validation. This provided an unbiased measure of real-world predictive performance.

Model	RMSE	R^2	MAPE
Linear Regression	0.0011	0.1517	46.63%
Ridge Regression (CV)	0.0011	0.1483	46.43%
Lasso Regression (CV)	0.0011	0.1439	45.91%
Random Forest (Tuned)	0.0008	0.5454	28.68%
XGBoost (Tuned)	0.0008	0.4984	32.86%
KNN Regression (Tuned)	0.0006	0.6938	20.69%

Table 4.2: Test Set Model Evaluation

As evident from the table, KNN Regression outperformed all other models on the test set, achieving the highest R^2 and lowest RMSE and MAPE. Random Forest and XGBoost followed, while linear models underperformed in comparison, likely due to their inability to capture nonlinear relationships.

Cross-Validation Performance

To ensure model robustness, I also conducted 5-fold cross-validation on the training data. This technique splits the training set into five subsets, training the model on four and validating on the fifth, rotating through each subset.

Model	CV R^2	CV RMSE	CV MAPE
Linear Regression	0.1732	0.0008	45.51%
Ridge Regression (CV)	0.1720	0.0008	45.24%
Lasso Regression (CV)	0.1752	0.0008	44.77%
Random Forest (Tuned)	0.3913	0.0007	30.32%

Table 4.3: Cross–Validation Performance

Cross-validation confirmed the generalizability of KNN and Random Forest models, with KNN again leading in all metrics.

The results suggest that non-parametric and ensemble methods significantly outperform linear models in predicting homelessness rates. In particular, KNN Regression captured local data structure and non-linear dependencies that linear models failed to model. These results support the use of KNN as the most reliable model for this task, both in accuracy and stability.

Comparative Modeling Framework: U.S. and Ireland

To evaluate the generalizability of our modeling framework across geopolitical contexts, I applied the same methodology to two distinct datasets: one from the United States and one from Ireland. Both datasets were structured with identical feature sets and aligned by year, allowing for a consistent cross-country comparison. The target variable in both cases was homelessness rate per capita, defined as the number of homeless individuals relative to the total population in a given year.

The same data preprocessing pipeline was used for both datasets, including imputation for missing values, feature scaling (z-score normalization). I then applied three regression models Ridge, Lasso, and Random Forest using Leave-One-Out Cross-Validation (LOOCV) to ensure unbiased performance evaluation. Model performance was assessed using RMSE, R^2, MAE, and MAPE, allowing for direct comparisons between the U.S. and Ireland in terms of model accuracy and predictive drivers.

This comparative approach not only highlights the versatility of our framework but also provides insight into regional differences in the socio-economic factors influencing homelessness.

5. Results

Model 1: U.S. (State + Year-Wise Model) — Main Analysis

Six regression models were evaluated using a reserved 30% test set and validated through 5-fold cross-validation to assess their predictive performance and generalizability. Metrics used included RMSE, R^2, and MAPE.

Model	**Test RMSE**	**Test R^2**	**Test MAPE**	**CV R^2**	**CV RMSE**	**CV MAPE**
Linear Regression	0.0011	0.1517	46.63%	0.1732	0.0008	45.51%
Ridge Regression	0.0011	0.1483	46.43%	0.1720	0.0008	45.24%
Lasso Regression	0.0011	0.1439	45.91%	0.1752	0.0008	44.77%
Random Forest	0.0008	0.5454	28.68%	0.3913	0.0007	30.32%
XGBoost	0.0008	0.4984	32.86%	0.2891	0.0007	36.00%
KNN Regression	0.0006	0.6938	20.69%	0.4945	0.0006	23.76%

Table 5.1: Model Performance Comparison for U.S. (State + Year-Wise Model)

On the test set, KNN Regression achieved the best performance overall, with the lowest RMSE (0.0006), lowest MAPE (20.69%), and the highest R^2 (0.6938), indicating that it captured the underlying patterns in the data most effectively. Random Forest and XGBoost also demonstrated strong performance, with R^2 values of 0.5454 and 0.4984, respectively, and lower error metrics compared to the linear models.

In contrast, the linear models, Linear, Ridge, and Lasso Regression showed relatively lower R^2 scores (all around 0.14–0.15) and higher MAPE values (above 45%), suggesting that they were less effective at modeling the non-linear relationships present in the data.

Cross-validation results were consistent with the test set findings. KNN Regression again stood out with the highest mean CV R^2 (0.4945 ± 0.1421), followed by Random Forest (0.3913 ± 0.1006). The linear models showed lower and more stable CV R^2 values around 0.17, reinforcing their limited ability to capture complex patterns in the dataset. Additionally, KNN and Random Forest had lower cross-validated RMSE and MAPE values, further confirming their reliability and predictive strength.

Overall, the results indicate that non-linear and ensemble-based models significantly outperformed linear approaches in predicting homelessness per capita in the U.S. dataset.

To better understand the factors influencing homelessness predictions, permutation importance was used to identify which features had the greatest impact on model performance. The figure below presents the top 15 predictors identified by the KNN model.

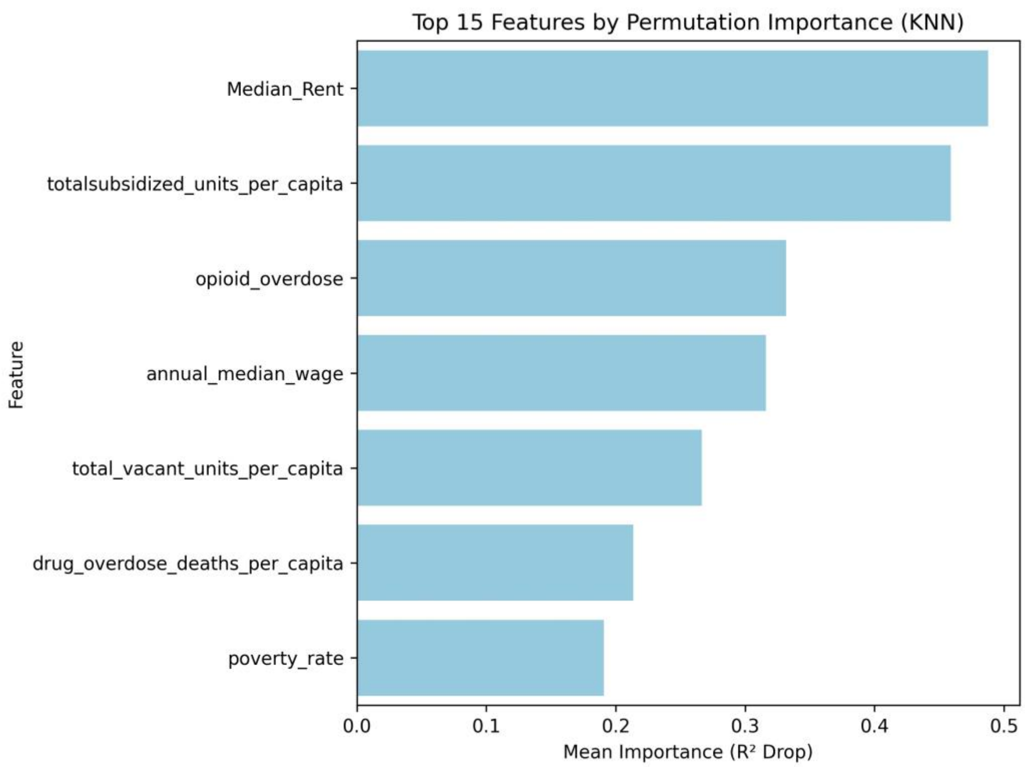

Figure 5.1: Top 15 Features by Permutation Importance in KNN Regression.

The most influential predictors identified in the KNN model were related to housing affordability and access, specifically Median Rent and subsidized housing units per capita.

The importance of Median Rent aligns with findings from the Urban Institute, which highlight that "increases in median rent are significantly associated with increases in homelessness rates, especially in metropolitan areas" [Urban Institute]. Even modest rent increases can result in a measurable rise in homelessness, particularly in low-income communities where individuals are already rent-burdened [Fischer & Sard, 2017].

Similarly, the presence of subsidized housing units per capita as a top predictor mirrors national data showing that investments in public housing and rental subsidies reduce homelessness.

According to a randomized evaluation by the U.S. Department of Housing and Urban Development (HUD*)*, long-term housing vouchers significantly reduce both homelessness and housing instability [HUD, 2016]. Furthermore, cities with higher rates of subsidized housing often report lower rates of unsheltered homelessness, controlling for other factors [Homelessness].

Beyond housing, other highly ranked features such as opioid overdose rates, drug-related mortality, poverty rate, and median wage reflect well-established connections between homelessness and broader public health and economic conditions. The CDC has documented that individuals experiencing homelessness face dramatically elevated risks of opioid overdose and other substance use-related mortality [CDC]. Likewise, research from the National Low Income Housing Coalition (NLIHC) shows that poverty and insufficient wages are among the strongest economic predictors of housing insecurity and homelessness [NLIHC].

These findings reinforce the idea that homelessness is a multi-dimensional crisis, but one that is fundamentally anchored in housing access and affordability, consistent with a growing body of literature **[Colburn & Aldern, 2022].**

Policy Implications (Based on Model Findings & Existing Evidence)

The model highlights housing-related variables (Median Rent and subsidized housing) as top predictors of homelessness. This supports existing research and policies that emphasize:

- Expanding affordable housing programs, including rental subsidies and vouchers, which have been shown to reduce homelessness in multiple large-scale studies [HUD Family Options Study].
- Providing rental assistance or regulating sharp rent increases, since rising rent burdens are linked to increased homelessness risk, particularly for low-income populations [Urban Institute].
- Integrating housing efforts with mental health and substance use services, because opioid overdose and drug-related deaths were also important predictors, consistent with studies showing better outcomes when housing and health services are coordinated [SAMHSA, 2021].

Comparative Models :

To enable cross-country comparison, models for the United States and Ireland were trained on year-wise data using a common set of features: rent, poverty rate, unemployment rate, mental health/substance use deaths per capita, crime rate, and wage.

Model Evaluation Metrics

United States

Model	**RMSE**	**R^2**	**MAE**
Ridge (LOOCV)	0.971	0.0798	0.710
Lasso (LOOCV)	0.610	0.656	0.474

Random Forest	0.669	0.616	0.535

Table 5.2: Model Evaluation Metrics (U.S. Comparision Model)

In the comparative U.S. model, Lasso Regression outperformed other methods, maintaining strong predictive accuracy ($R^2 = 0.656$) even with fewer input variables.

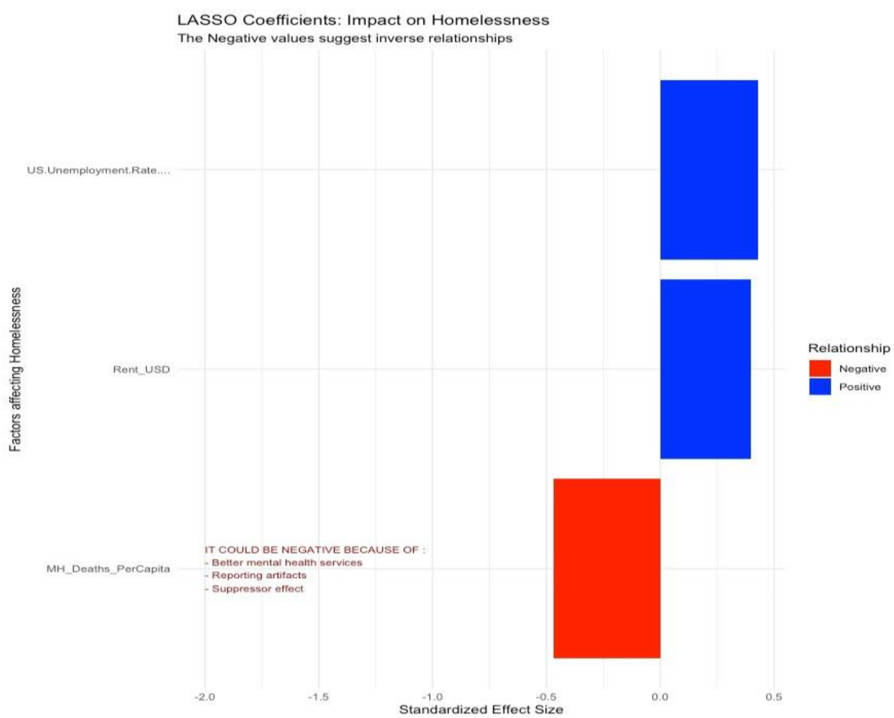

Figure 5.2: U.S. – Standardized Lasso Coefficients (Common Features)

The key features selected were:

- Rent (USD) → positive standardized effect
- Unemployment Rate (%) → positive effect
- Mental Health/Substance Use Deaths Per Capita → negative effect

These results are consistent with the main U.S. model, where rent emerged as the strongest driver, reinforcing the relationship between housing affordability and homelessness. The inclusion of unemployment aligns with well-established research showing job loss increases housing precarity **[Colburn & Aldern, 2022]** [NLIHC, 2022].

Ireland

Model	RMSE	R^2	MAE
Ridge (LOOCV)	0.000217	0.8951	0.000172
Lasso (LOOCV)	0.000229	0.8833	0.000186
Random Forest	0.000198	0.9132	0.000165

Table 5.3: Model Evaluation Metrics (Ireland Model)

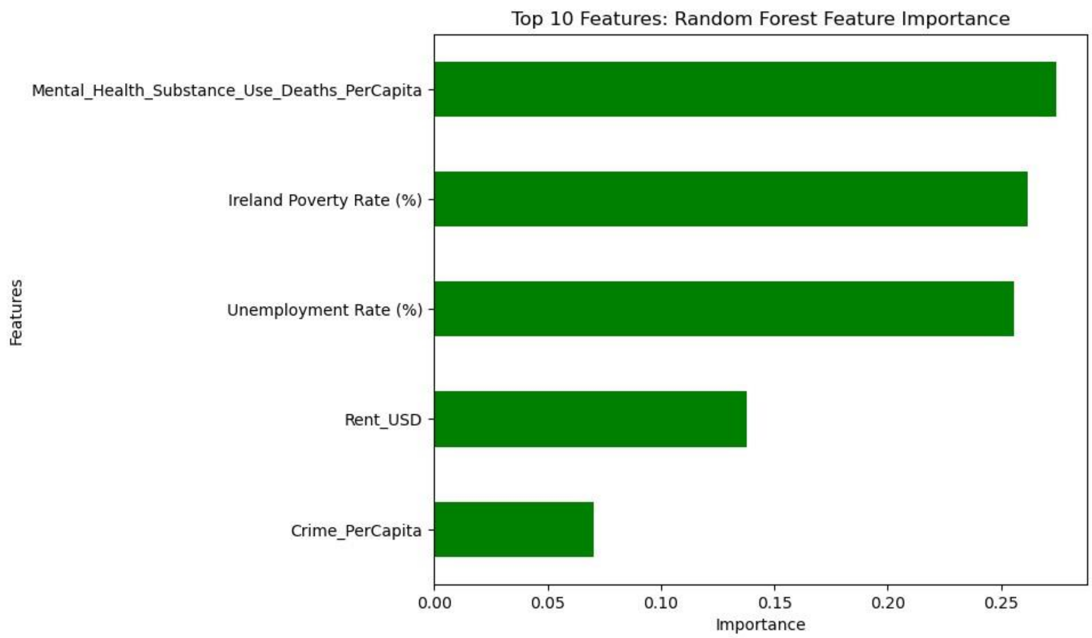

Figure 5.3: Ireland – Feature Importance from Random Forest (Common Features)

Ireland's best-performing model was Random Forest, with high predictive accuracy ($R^2 = 0.9132$). The most influential features were:

1. Mental Health/Substance Use Deaths Per Capita
2. Poverty Rate (%)
3. Unemployment Rate (%)

This reflects structural differences in how homelessness operates in Ireland:

- According to Focus Ireland, austerity-era cuts to mental health and welfare supports were directly followed by spikes in homelessness [Focus Ireland]
- Poverty and unemployment have long been shown to correlate with housing instability in Irish national data [ESRI 2018]
- Ireland's centralized housing system, with more rent regulation and public housing access, may reduce the direct impact of market rent on homelessness outcomes [OECD 2021]

Interpretation & Policy Link

- In the U.S., rent remains a primary structural driver, consistent with the full model and national housing literature. This reinforces the need for affordable housing expansion, rent caps, and unemployment support **[Colburn & Aldern, 2022]**.

- In Ireland, mental health and poverty emerged as dominant factors. This supports evidence from Irish policy studies calling for investment in public health, housing-first approaches, and poverty alleviation [Focus Ireland, 2020] [ESRI, 2018].
- The comparative models, despite using fewer features, support the main findings by validating the role of rent in the U.S. and highlighting structural differences in how homelessness is shaped in different policy environments.

6. Deliverables

The deliverables or outcomes of this project include a robust and accurate analysis of homelessness trends in the United States and Ireland, achieved through a comprehensive data collection, exploratory analysis, and predictive modeling. This project uses state-of-the-art machine learning algorithms, such as K-Nearest Neighbors (KNN), Random Forest, XGBoost, Ridge, and Lasso regression, along with feature engineering techniques and hyperparameter tuning. The model's performance will be evaluated using various metrics, including RMSE, R^2, MAPE, and crossvalidation results, ensuring the models provide reliable predictions for homelessness rates.

In the context of the problem objectives, the outcomes of this project can significantly impact both research and public policy. By developing predictive models that identify the key drivers of homelessness—such as housing costs, unemployment, and public health factors—this project can provide valuable insights to inform policymaking and guide future interventions. These models can help policymakers at local, state, and national levels better understand the socio-economic conditions that lead to homelessness and develop data-driven strategies to combat the crisis.

Moreover, the project's comparative analysis of homelessness trends in the U.S. and Ireland will offer a unique perspective on the different national approaches to the issue. By examining government initiatives such as the U.S. Federal Strategic Plan – All In and Ireland's Housing for All Plan, the project will highlight the role of policy interventions in shaping homelessness outcomes. This cross-national approach can provide a deeper understanding of the effectiveness of various strategies and contribute to ongoing efforts to reduce homelessness globally.

In summary, the project's deliverables and outcomes include not only an accurate and robust predictive model for understanding homelessness but also actionable insights for policymakers. By integrating large-scale datasets and advanced analytical techniques, this project contributes to the growing body of research on homelessness and provides valuable resources to support efforts to address this critical issue. The findings will also offer practical guidance for enhancing the effectiveness of national and regional policies aimed at eradicating homelessness, ultimately improving housing stability and socio-economic conditions for vulnerable populations.

7. References

- [1] U.S. Department of Housing and Urban Development (HUD). “2024 AHAR: Part 1 - PIT Estimates of Homelessness in the U.S.” Available at: https://www.huduser.gov/portal/datasets/ahar/2024-ahar-part-1-pit-estimates-ofhomelessness-in-the-us.html
- [2] U.S. Census Bureau. “American Community Survey (ACS) – Housing, Poverty, and Population Topics.” Available at: https://www.census.gov/topics/housing.html
- [3] U.S. Department of Housing and Urban Development. “Continuum of Care (CoC) Housing Inventory Count (HIC) Reports.” Available at: https://www.hudexchange.info/programs/coc/coc-housing-inventory-count-reports
- [4] Government of Ireland. “Homelessness Data and Statistics.” Available at: https://data.gov.ie/dataset?theme=Housing
- [5] U.S. Bureau of Labor Statistics. “Occupational Employment and Wage Statistics (OEWS).” Available at: https://www.bls.gov/oes/tables.htm
- [6] Kaiser Family Foundation (KFF). “Opioid Overdose Deaths by Age Group.” Available at: https://www.kff.org/other/state-indicator/opioid-overdose-deaths-byagegroup/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22s ort%22:%22asc%22%7D
- [7] National Low Income Housing Coalition. “Research and Data on Housing Affordability.” Available at: https://nlihc.org/
- [8] U.S. Department of Veterans Affairs. “Homeless Veterans Programs.” Available at: https://www.va.gov/homeless/
- [9] Joint Center for Housing Studies of Harvard University. “Research and Reports on Housing in America.” Available at: https://www.jchs.harvard.edu/
- [10] Eviction Lab. “National Database on Evictions and Housing Instability.” Available at: https://evictionlab.org/
- [11] Economic Policy Institute (EPI). “Research on Economic Inequality and Living Standards.” Available at: https://www.epi.org/
- [12] PubMed Central. “Barriers to employment among homeless youth.” Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC2856116/
- [13] Substance Abuse and Mental Health Services Administration (SAMHSA). “Behavioral Health Data and Programs.” Available at: https://www.samhsa.gov/
- [14] Housing Agency Ireland. “Housing Market Insights and Policy Resources.” Available at: https://www.housingagency.ie/
- [15] Pew Charitable Trusts. “How Housing Costs Drive Levels of Homelessness.” Available at: https://www.pewtrusts.org/en/research-and-analysis/articles/2023/08/22/how-housing-costs-drive-levels-of-homelessness
- [16] OpenAI. "ChatGPT: Language Models for Dialogue." Available at: https://openai.com/chatgpt
- [17] Grammarly. "Grammarly: Writing Assistance." Available at: https://www.grammarly.com/