Factors Influencing the Rate of Homelessness in the United States

Ihsan Alaeddin, Isak Dai, Xinran Li, Chaowei Wang

I. Introduction

“A record-high 653,104 people experienced homelessness on a single night in January 2023” (How Many Homeless People Are in the US? What Does the Data Miss?, 2024).

Homelessness is not a word we think of everyday, until it affects someone we know. The United States’ homelessness crisis has been exacerbated by a confluence of factors ranging from the personal to the systemic. Macroeconomic trends in inflation, prices, and unemployment can squeeze household budgets, and previous research has found a strong relationship between increases in the cost of living and the rate of homelessness (Heston, 2023). Rents and housing costs in particular are determiners of homelessness. As housing costs rise, competition for the lowest-quality housing becomes fierce and the costs of homelessness become preferable to the unbearable rents people on the brink of homelessness would need to pay to live in dilapidated housing (Quigley & Raphael, 2001).

Household debt is another potential cause of homelessness. Many people are currently in debt due to either student loans, the cost of living, cost of health care, divorce, etc. Debt is directly associated with homelessness. In severe cases, personal belongings can be taken to repay that debt, and thus we consider debt as a factor in this report. Another factor to consider is poverty. 2023 estimates show that more than 11% of Americans live below the poverty line (Shrider, 2024). Poverty can arise due to low minimum wage, debt, lack of education, and other. Poverty and homelessness go hand in hand as flagging incomes force difficult choices between health care, housing, food, and other necessities, which is why poverty was decided as a factor in this research.

Finally, personal factors like addiction are another potential contributing factor to homelessness. Drug addiction can refer to controlled substances as well as also legalized drugs such as alcohol and nicotine. Substance abuse and addiction has grown as an huge issue due to them becoming the norm for socializing, dealing with grief and depression, and the associated spike in mental health diagnoses in recent decades. Researchers debate the causality of substance abuse and homelessness i.e. whether substance abuse leads to homelessness as the disease may hurt someone's ability to keep a job or make financial decisions or whether substance abuse arises out of the stresses and conditions of being homeless. However, a meta-analysis of studies on substance abuse and homelessness shows consistent evidence that homeless populations have a much higher rate of substance abuse (Coombs et al., 2024).

Given the severity of the United States' homelessness crisis and its indiscriminate impact on all communities, we think it is pertinent and important to study which of these aforementioned factors contribute significantly to the rate of homelessness in different states. Our data science research question is as follows:

- *What are the most remarkable factors that affect homelessness?*

II. Datasets

Our homelessness data comes from Omdena, which aggregated counts of each state's homeless population collected by the Department of Housing and Urban Development from 2007 to 2022. The dataset was an Excel file and had sheets that were labelled yearly. Each sheet contained 58 rows, with each row labeled by state. We excluded columns which were irrelevant for this study, as many contained more detailed information about the demographics of each state's homeless population (e.g. gender, veteran status, age). In this study we are only interested in the overall homeless population (Overall Homeless) column and the state column. Using pandas, each sheet was extracted, with a new column that attached the year to each row, to create a new dataset. Then, we used pd.concat to combine each year's dataset. The state column in the dataset was the abbreviation of the states, so to get the full name of each state, the US library was used, and the state column was changed from abbreviations to the full name of the state. The rows that had missing values were dropped and the columns were made sure to be of type string for states and integers for population and year.

To measure household debt, we used a dataset from the Federal Reserve containing each state's aggregate household debt-to-income ratio. This dataset also had the state fips code; another column was added by searching the fips code and attaching the state name to a new column (State). Also, it had two columns with information about each state's debt-to-income ratio. The first was named, low, which refers to the lower bound of the debt-to-income ratio category. The second was named high, which refers to the upper bound of the debt-to-income ratio category. A new column was added named *average_debt*, which took the average of the low and high bounds of the debt-income ratio. Debt rate was also cleaned with the same methods.

We used a Census Bureau dataset to measure the rate of poverty. The format of the Excel file was unclean and could not be read by pandas, so from the year 2007-2023, and thus they were cleaned manually by deleting unnecessary rows and fixing the labelling. Using pandas, we then selected only the columns containing year, state, and the total population living in poverty.

Our drugs dataset was extracted out of a Kaggle dataset that collected data from "individual states as part of the NSDUH study." (mexwell , 2023). The dataset was very detailed in terms of demographics about the total population of drug users, ranging from ages, to drug types, to monthly usage, and yearly usage. Since this dataset had a lot of columns, the first step was to remove any column that had the word month or rate in it. The rest of the dataset was the divided population of multiple demographics of users of different drugs. Column indexes with the pandas library were used to sum these columns together into one, called the “'total_drug_population'.” After that a new dataset was created from the subset of the old one, containing only the year, state, and total_drug_population.

We also considered a set of economic variables that could influence the rate of homelessness. We obtained state-level unemployment data from the St. Louis Federal Reserve’s Federal Reserve Economic Data (FRED) site. The raw dataset contained the monthly unemployment rate for each state from January 1976 to October 2024. We averaged the monthly unemployment rate by year to create a dataset of each state’s yearly unemployment rate. Unemployment rates could influence homelessness as higher unemployment rates reflect a loose labor market that keeps unemployed people from quickly finding a new job that would allow them to keep their housing.

Inflation could also influence homelessness if rising prices impact the capacity to pay for housing, and we found a state-level inflation dataset from 1978-2018 constructed by Juan Herreño, Emi Nakamura, and Jón Steinsson out of Bureau of Labor Statistics data. This dataset contained quarterly inflation in the tradeable sector, non-tradeable sector, and overall for each state. We selected only the overall inflation and then averaged each year’s quarterly data to create an average rate of inflation for each year.

Finally, we also collected a dataset from the Department of Commerce’s Bureau of Economic Analysis containing regional price parities (RPP) for each state from 2008-2022. This dataset contains the relative prices in each state in four categories – goods, housing, utilities, and other – along with an overall RPP. National prices are set to an RPP value of 100, so if a state has an RPP of 110, this indicates that its prices are 10% higher than the national level. Previous studies have found that higher rents correlate with higher rates of homelessness (Quigley & Raphael 2001).

We combined each of our datasets with an inner join and were left with a tidy database where each variable had its own column, each observation has a row (where an observation is a state in a given year), and each value has its own cell. This dataset spanned each year from 2008 to 2017. We also added a column containing each state's population in a given year pulled from the Census Bureau and used it to convert any count data (people living in poverty, people suffering from drug addiction, people experiencing homelessness) into rate data.

III. Exploratory Data Analysis

IIIa. Visualization of Trends in Variables

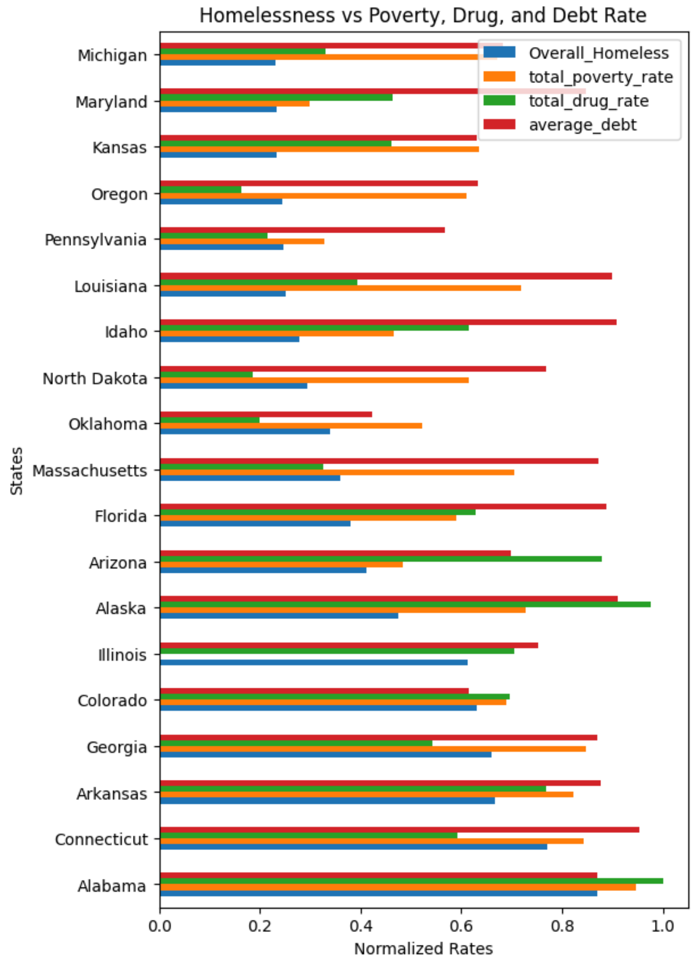
We began to explore our dataset by visualizing the distribution of different variables geographically. We grouped the data by state and then took the mean of each variable to find its mean in each state over the period from 2007 to 2018. We then plotted our data to create choropleths using a map of the United States, as it is both visually appealing and gives easily comprehensible analysis. Each factor was plotted in a separate US map, where the darker the state the higher the rate of the factor was. From this visualization, individuals can easily detect states that are the highest in each factor. These maps are found in the appendix for the sake of space. A heat map was also used for the homeless column, to show how big the proximity of a heat signal was for each separate state. For economic variables like inflation, unemployment, and RPP, we plotted choropleths using only a single year, the most recent one available for each dataset. We can also interpret the trends in these variables by looking at the five states with the highest values for each of these variables.

Table 1: Top Five States per Selected Variables

Homelessness	Poverty	Debt-Income Ratio	Drug Use
New York	Utah	Maryland	**Oregon**
Oregon	***Oregon***	Idaho	Colorado
California	***Nevada***	Virginia	Vermont
Washington	***California***	***California***	Rhode Island
Nevada	Idaho	Colorado	***Washington***

As we can see, states with the highest average rates of homelessness in this period were often among the states with the highest of poverty, date-to-income ratios, and drug use. Taking a more analytical approach, we created a line chart to show how certain variables changed as the rate of homelessness increased. For ease of visualization, we normalized the data as our selected variables had very different scales. For example, the debt-to-income ratio was on a much higher scale (centered around 1) than the homelessness rate. The data was then sorted based on the descending values of the homeless rate. The first visualization created based on the normalized data was a bar chart, where the top 20 states that had the highest homeless rate were visualized based on “*Overall_Homeless*", "*total_poverty_rate"*, "*total_drug_rate*", and "*average_debt*".

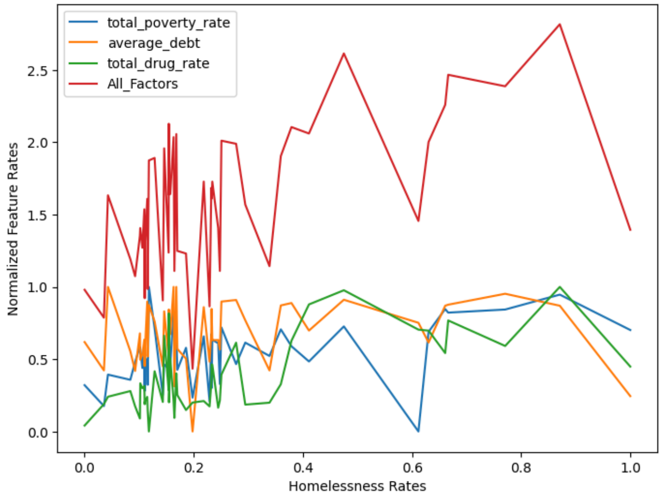
Plot 1: Bar Chart of Homelessness vs. Poverty, Drug, and Debt Ratio



From the bar chart, it is easily seen that the states that had the highest homeless rate also had high rates of the other different factors. Also, it is seen from the graph that the highest factor is usually the debt-to-income ratio, meaning that debt may have the strongest correlation with homelessness based on the small sample.

We also created a line graph to examine how other factors change with an increase in the homeless rate. The first line graph was chosen as it can easily show the relationship between the factors and the homeless rate. The second line graph is of the combined rates, in this case, the effects of the factors can be seen separately and joined together.

Plot 2: Line Chart of Homelessness vs. Poverty, Drug, and Debt Ratio



From the graph, it can be seen that the relationship between the different factors and homelessness is not directly linear. However, there appears to be a relationship as each factor increases when the homeless rate increases. All factors seem to have similar correlations with homelessness, as they similarly increase and decrease at the same points.

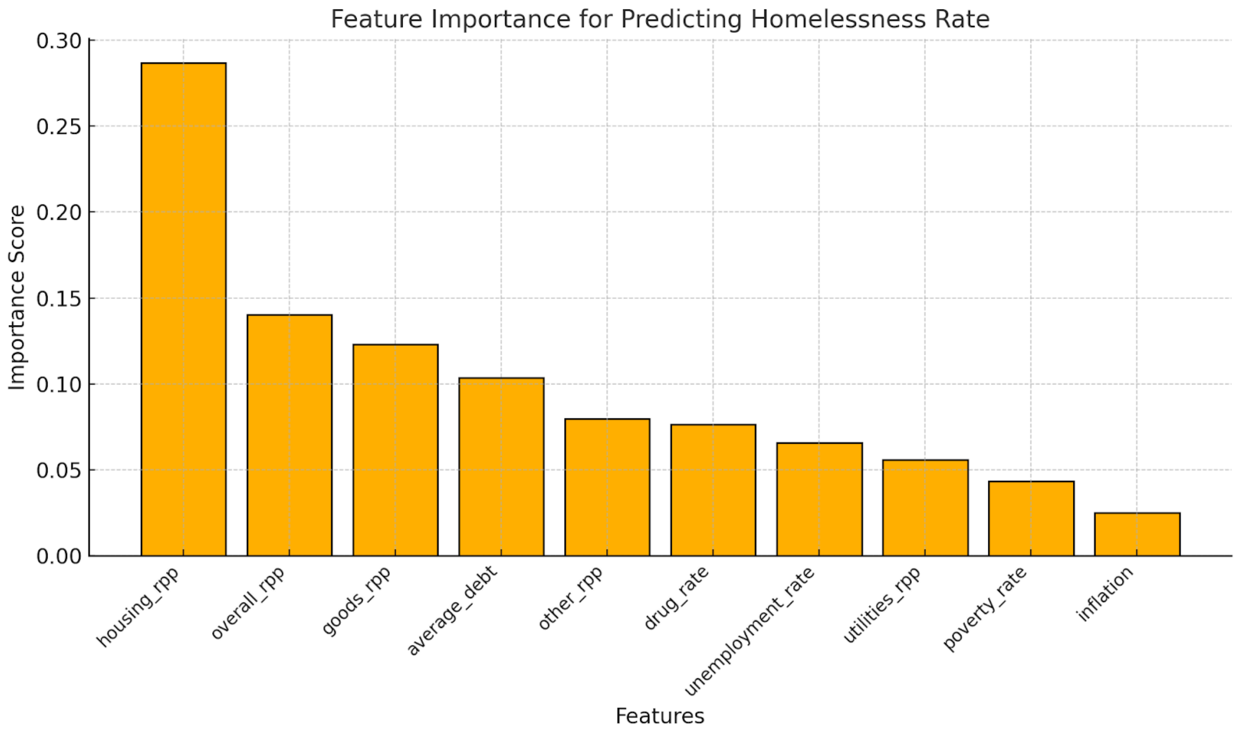
Since a relationship is evident between the factors and homelessness, a linear regression model for each variable could be used to examine the relationship more closely. A linear regression model was chosen as it is also capable of approximately the homelessness rate based on each factor and the strength of the relationship. This can be done by drawing a straight line following the path and slope of the line drawn by the model. For example, in the drug rate, if the drug rate reaches 0.00035, the homelessness rate is predicted to be 0.0028. The model does that by taking the data given (data from the past) to predict the trend of future data. However, a linear regression done in this manner would be inappropriate since it does not account for the impact of time on the correlation between data points. Since our observation is state in a given year, we must take time into consideration when conducting a linear regression and thus need to check if our residuals are autocorrelated.

IIIb. Variable Selection

To effectively reduce homelessness rates through data-driven policy analysis, it's practical to focus on a select few key variables rather than addressing all factors simultaneously. In this research, we concentrate on assessing the feasibility of lowering the homelessness rates of states that are currently above the upper quartile threshold when compared across all states.

To identify the most influential factors affecting homelessness rates over time, we apply a Random Forest algorithm to determine variable importance within our time series data. This method allowed us to assess which variables had the most significant impact on predicting homelessness rates while accounting for temporal dependencies.

Plot 3: Feature Importance in Predicting Homelessness Rate



In refining our selection, we chose to exclude *overall_rpp* from further consideration. The primary reason for this decision is that *overall_rpp* is a composite index that encompasses various economic factors, potentially masking the effects of individual components. For effective policymaking, especially when aiming for targeted interventions, it is crucial to focus on variables that offer local granularity and specific insights. By eliminating the composite *overall_rpp*, we ensure that our analysis centers on more actionable and precise variables. This approach enhances the relevance of our findings for policy development, as it allows policymakers to design strategies that directly address specific economic aspects influencing homelessness rates.

To ensure we focus on variables amenable to short-term policy interventions, we assess the variability of our candidate options using the Coefficient of Variation (CV) and Standard Deviation (SD). *goods_rpp* displayed low variability, indicating it is relatively stable and less likely to be influenced within a one-year timeframe. Consequently, we excluded *goods_rpp*, leaving Housing Regional Price Parity (housing_rpp) and Average Debt as our chosen variables.

To consolidate our decision on variable selection, we performed a lagged regression analysis with statistical testing to evaluate the influence of *housing_rpp* and *average_debt* on the homelessness rate. Using t-tests to assess the significance of their coefficients, we obtained the following results:

For *housing_rpp* (lagged), the regression yielded a coefficient of 0.0004 with a p-value of 0.031. This indicates that the lagged effect of *housing_rpp* on the homelessness rate is statistically significant ($p < 0.05$). Therefore, *housing_rpp* can be retained for further analysis as it significantly contributes to predicting homelessness rates.

In contrast, the regression for *average_debt* (lagged) resulted in a coefficient approximately equal to 0.0 and a p-value of 0.954. This high p-value suggests that the lagged effect of *average_debt* on the homelessness rate is not statistically significant ($p > 0.05$). As a result, *average_debt* should be excluded from further consideration since it does not have a meaningful impact on the homelessness rate.

For simplicity of a single variable relationship with a limited time frame (10 years), we can use an OLS regression method. The OLS regression analysis revealed an R-squared value of 0.437, indicating that 43.7% of the variability in homelessness rates can be explained by variations in housing RPP. The intercept of -0.0009 (p = 0.087) represents the baseline homelessness rate when housing RPP is zero, though it is not statistically significant at the 5% level. The housing RPP coefficient of 0.0024 (p < 0.001) demonstrates a statistically significant positive relationship. The F-statistic of 23.26 (p < 0.001) confirms the overall statistical significance of the model, and the low standard error for the housing RPP coefficient (0.0005) underscores the precision of this estimate. These results highlight the potential impact of housing affordability policies on addressing homelessness. However, given the time series dependencies within our data, we should be wary of interpreting the results of our linear regression at face value without examining the autocorrelation of our residuals which could lead to underestimated standard errors. Due to the limited size of our data and the constraints of using single-variable prediction methodologies, we cannot precisely determine how increases in housing prices relative to the national level might influence homelessness rates. As a result, we use this analysis simply as a way of considering whether housing RPP might be an important factor in further statistical analyses.

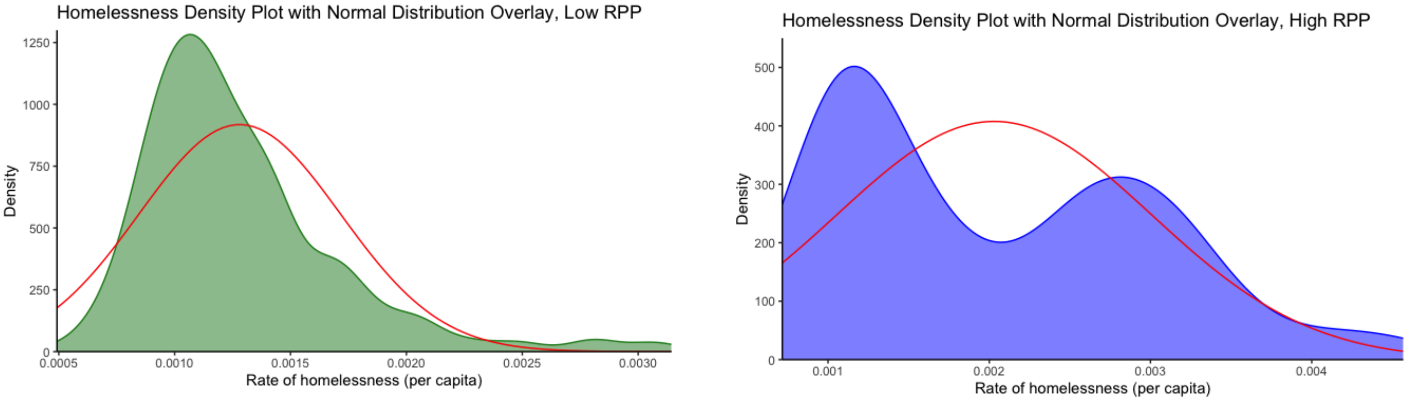
IV. Statistical Methods

IVa. Bootstrap difference in means test

Our regional price parity (RPP) data offers an intuitive way to split our observations into two groups. The RPP data, which includes information about overall prices as well as prices in specific sectors like housing, is indexed to national average prices for each year. For instance, the housing RPP for California in 2017 was 162.521 and in Arkansas, it was 57.64. Since the national RPP is always 100, the cost of housing in California was 62.5% higher than the national average housing prices. On the other hand, in Arkansas were only 57.6% of the national average. With this information in mind, we can split our observations into two groups, one including all observations where the housing RPP is greater than 100 and another where the housing RPP is less than or equal to 100. Since there are not an equal number of observations in each of our groups (n = 122 for above-national RPP, n = 313 for below-average RPP), if we want to conduct a t-test we cannot use a Student's t-test since this test is predicated on each group having an equal size if the two samples do not have the same variance. A Welch's t-test relaxes these assumptions and allows us to compare means between two independent approximately-normally distributed groups.

Initially, a Welch's t-test for difference in means seems best given our two groups. However, conducting a Welch's t-test would be inappropriate given the distribution of the homelessness rates for each group. Below are the density plots of the distributions of homelessness rates within each group with an overlay of the normal distribution with the same mean and standard deviation in red.

Plot 4: Distribution of Homelessness Rates within High and Low RPP Groups



As we can see, the distribution of homelessness within each group deviates significantly from normality. The distribution of homelessness rates for high-housing RPP observations is bimodal, and the distribution of homelessness rates for low-housing RPP observations has a long right tail. As a result, a Welch's t-test is inappropriate since it is a parametric test that assumes that both groups are normally distributed, even if they have different variances.

As a result, our best option is to conduct a non-parametric hypothesis test which makes no assumptions about the underlying distribution of our groups. One non-parametric test to consider is the permutation test. In a permutation test, we repeatedly relabel the data to test whether or not an observed difference in means is due simply to chance. After testing every permutation of labels (in this case, every permutation of our data that ensures our low-housing RPP group and high-housing RPP groups are 313 and 122 observations, respectively), we can then examine the distribution of differences in means to see whether our observed difference in means is statistically significant. However, the null hypothesis of a permutation test is that both groups come from the same distribution, meaning that our two groups need to have the same variance to fulfill the assumptions of the null hypothesis. A bootstrap distribution of the ratio between the variance of high-housing RPP and low-housing RPP groups has a 95% percentile interval of (3.675, 7.074), and none of our bootstrap sample ratios is less than or equal to 1. As a result, we can be confident that our two groups have different variances and thus cannot come from the same distribution as required by the null hypothesis of a permutation test. In this situation, it is best to use a more robust bootstrap difference in means test to account for the non-normality of our data and their different variances.

In a bootstrap difference in means test, we take each group and resample it without replacement to generate many new samples. We then calculate the mean of each of these samples, and per the Central Limit Theorem, the distribution of these sample means will be approximately normally distributed. Finally, we can calculate the difference between our large number of sample means to generate a bootstrap distribution of the difference in mean homelessness rates between our two groups. We can use this distribution as the basis for a hypothesis test, with our p-value represented by the proportion of bootstrap sample differences in means where the mean rate of homelessness is higher in low-RPP observations than high-RPP observations. An advantage of this test is that it is non-parametric, meaning it makes no assumptions about the data's distribution (such as normality) or any underlying structure (like both groups originating from the same population).

V. Results

Va. Bootstrap difference in means test

We conducted a bootstrap difference in means test to determine whether an observed difference in mean homelessness rate between high-housing RPP and low-housing RPP observations is statistically significant. Our null and alternative hypotheses are:

$H_0 : \mu_{\text{High RPP}} = \mu_{\text{Low RPP}}$

H_A: $\mu_{\text{High RPP}} > \mu_{\text{Low RPP}}$

Where $\mu_{\text{High RPP}}$ indicates the mean rate of homelessness among states with housing prices higher than the national average in a given year and $\mu_{\text{High RPP}}$ indicates the mean rate of homelessness among states with housing prices lower than the national average in a given year.

As previously mentioned, the bootstrap difference in means test is non-parametric, and thus we make no assumptions about the distribution or any parameters of the underlying data. Instead, we use our existing samples to determine the variance of the difference in means between our two groups to see whether the observed difference is statistically significant or simply due to chance.

Table 2: Summary Statistics of Above- and Below-National Housing RPP Groups

Group	Size	Mean Rate of Homelessness
Above National Housing RPP	122	0.002030577
Below National Housing RPP	313	0.001280915

From our table, we can see that states with above-national housing prices have a higher mean rate of homelessness than states with below-national housing prices (0.0007496614, or 0.750 per 1,000 people higher).

We generate 5,000 bootstrap samples by sampling with replacement within each group and then finding the difference in means between each of these 5,000 samples. Below is a density plot of our bootstrap distribution of the difference in means with a normal distribution with the same mean and standard deviation overlaid in red.

Plot 5: Bootstrap Distribution of the Difference in Means

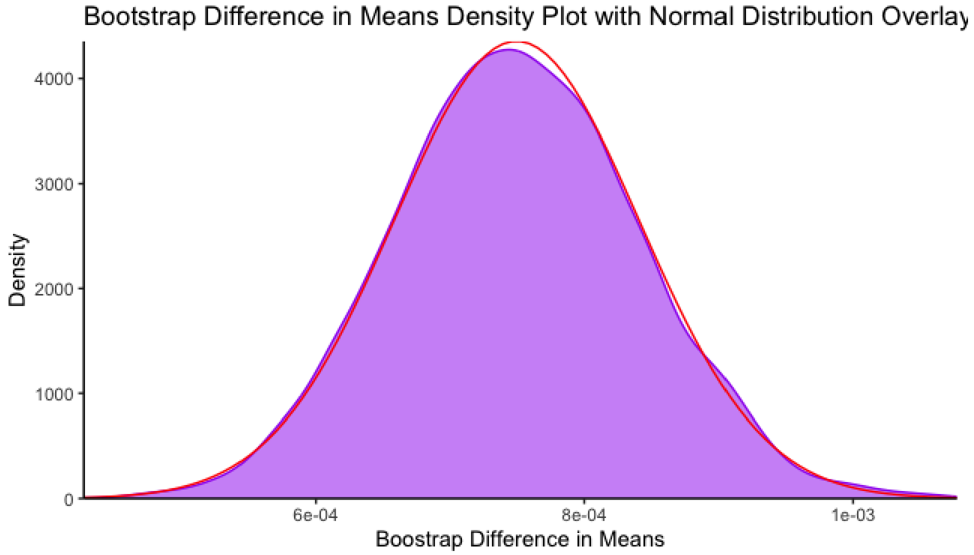


Table 3: Summary Statistics of Bootstrap Distribution of Difference in Means

Mean	**Median**	**Standard Deviation**	**2.5th Percentile**	**97.5th Percentile**
0.0007495503	0.0007486003	9.158168e-05	0.0.0005747357	0.0009299310

A 95% percentile interval of this bootstrap distribution is (0.0005747357, 0.0009299310), which means that 95% of our bootstrap differences in means are between 0.575 per 1,000 people and 0.930 per 1,000 people. Furthermore, none of our bootstrap sample differences are negative, which would indicate that the bootstrap sample mean homeless rate in above-national housing RPP states is lower than in below-national housing RPP states. As a result, our p-value is 0, and we can reject the null hypothesis and say there is enough evidence to suggest that the group of states with higher housing prices than the national average in a given year have a higher mean rate of homelessness. It is important to note that the interpretation of this result is not analogous to a linear regression between housing RPP and homelessness rate.

This test does not indicate whether or not there is a significant linear relationship between housing RPP and homelessness rates, but instead shows that the group of states with higher housing prices than the national average had a higher mean homelessness rate than the group of states with lower housing prices than the national average.

Vb. Linear Regression

We perform a linear regression analysis on the unemployment rate and poverty rate to evaluate whether there is a significant relationship between the two. The model parameters (such as slope, intercept, R^2, p-value, etc.) are output to assess the goodness of fit and significance of the model. We have the following null and alternate hypotheses:

$H_0 : \beta_{\text{Unemployment}} = 0$

$H_A : \beta_{\text{Unemployment}} \neq 0$

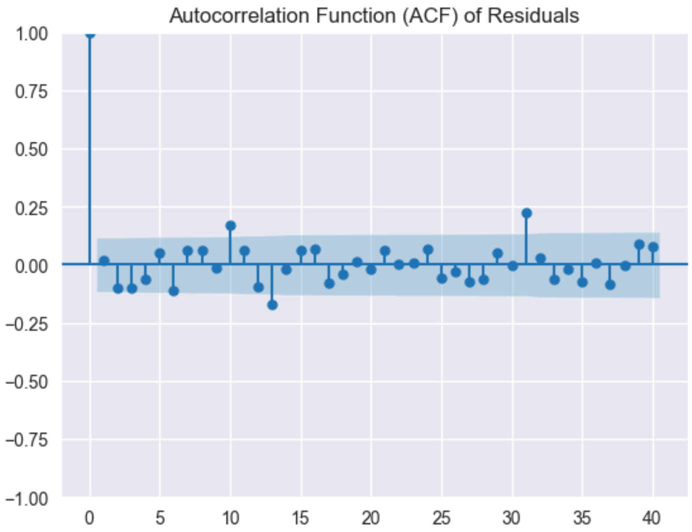
If the p-value is less than 0.05, it indicates that the linear relationship between the unemployment rate and poverty rate is significant. R^2 represents the proportion of variation in the poverty rate explained by the unemployment rate, with higher values indicating stronger explanatory power. By normalizing the data, we can more clearly compare different variables and make the subsequent charts more intuitive. Additionally, sorting by homelessness rate helps identify the areas (such as states or regions) with the most severe issues.

Table 4: Results of Regression Between Poverty and Unemployment

Variable	**Coefficient**	**t**	**p-value**
Constant	0.0010	401.773	0.0000
Unemployment	-5.618e-07	-1.562	0.119

Our R-squared (coefficient of determination) is 0.009, indicating that the model can only explain 0.9% of the variation in the poverty rate, which suggests that the explanatory power of the unemployment rate on the poverty rate is very weak. Adjusted R-squared is 0.005. After adjustment, the explanatory power remains very low, further indicating that the linear relationship between the unemployment rate and the poverty rate is not strong. The F-statistic and p-value are 2.440 and 0.119, respectively. The p-value is greater than 0.05, indicating that the model as a whole is not statistically significant, and the linear effect of the unemployment rate on the poverty rate is not significant. The regression coefficient is -5.618e-07. The coefficient for unemployment rate is negative but very close to zero, suggesting that changes in unemployment rate have a minimal impact on poverty rate.

Plot 6: Autocorrelation Function of Residuals of Poverty Regression



We examine whether the residuals of the linear regression model exhibit autocorrelation by using ACF and PACF plots for a visual assessment. Additionally, the Durbin-Watson statistic is used to detect the type of autocorrelation in the residuals (positive, negative, or no significant autocorrelation). If the Durbin-Watson value is close to 2, it indicates that the model residuals do not show significant autocorrelation. A bar chart is then created to display the top 20 states with the highest homelessness rates, while also comparing their poverty rates, drug abuse rates, and average debt rates. By visually comparing multiple states in terms of homelessness rates and influencing factors through the bar chart, it becomes easier to identify correlations or distribution patterns. Furthermore, this method provides preliminary observational data to support further in-depth analysis.

Based on the Durbin-Watson test, the statistic is 1.952, which is close to 2. This indicates that at the level of first-order lag, there is no significant autocorrelation among the residuals of the regression model, and the residuals exhibit a relatively random distribution. This aligns with one of the assumptions of the linear regression model, suggesting that the model meets the assumption regarding time-related independence.

In the Partial Autocorrelation Function (PACF) plot, the partial autocorrelation coefficients for most lags are close to 0 and fall within the 95% confidence interval, indicating that after controlling for the effects of other lags, the direct relationship between each lag and the current residual is not significant. This is consistent with the results of the ACF plot, further confirming that the model's residuals are random and lack temporal correlation.

The normality tests for residuals, including the Omnibus test and the Jarque-Bera test, show that the p-value for the Omnibus test is 0.000 and the p-value for the Jarque-Bera test is 1.07e-08, both indicating that the distribution of residuals deviates from normality.

From the skewness (Skew = -0.601) and kurtosis (Kurtosis = 4.303), it can be observed that the residual distribution exhibits some skewness and a higher kurtosis in the tails, which may suggest the presence of outliers or nonlinear characteristics.

As a result, we do not believe that a linear regression model is the best way to model the relationship between unemployment and poverty, as the relationship is insignificant and its non-normal residuals violate key assumptions of a linear regression model.

Vc. ANOVA

The one-way analysis of variance (ANOVA) was used to test whether the unemployment rate, when grouped into two categories (above and below the mean), has a significant impact on the variation in the poverty rate. The analysis aimed to determine whether there was a statistically significant difference in poverty rates between the two groups. Our null and alternate hypotheses are:

H_0: Mean poverty across all groups is the same

H_A: At least one group's mean poverty is different

Table 5: Results of ANOVA between Poverty and Unemployment

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**F**	**Pr> F**
Between Groups	3.263 e-09	235	0.897	0.69988
Error	6.809 e-08	44		

The F-statistic of 0.897 indicates the ratio of between-group variation to within-group variation. The relatively small F-value suggests that differences in group means do not explain a substantial portion of the total variation in poverty rates. The p-value of 0.6999 is much higher than the common significance level (e.g., α=0.05) indicating that grouping unemployment rates does not have a statistically significant effect on poverty rates.

Vd. Bootstrap T-test

The Bootstrap analysis was conducted by performing 1,000 iterations of resampling for poverty rates across the high-unemployment and low-unemployment groups to evaluate the statistical significance of group differences and the robustness of model results. The analysis showed that the mean t-statistic was -1.356, indicating a generally negative difference in poverty rates between the groups. The standard deviation of the t-statistic was 0.950, reflecting a moderate level of variability in the distribution. The 95% confidence interval ranged from [-3.126, 0.537], including 0, which suggests that the group difference may not be statistically significant. In terms of directional robustness, only 7.9% of iterations yielded positive effects, indicating that positive group differences are rare, and the robustness of the effect direction is classified as moderate.

Overall, the Bootstrap analysis serves to validate whether the poverty rate in the high-unemployment group is significantly higher than that in the low-unemployment group. By examining both the directionality and consistency of the t-statistic distribution, this method helps mitigate potential statistical biases caused by insufficient sample size and enhances the credibility of the results.

VI. Conclusion

As a whole, our analyses have shown that of the myriad factors that influence the rate of homelessness, it appears that Housing Regional Price Parity (housing RPP) is among the most important. While other variables like household debt, rates of addiction, and the poverty rate seem to have some correlation with the rate of homelessness, our random forest model found that housing RPP was the most influential factor in predicting the rate of homelessness. It is important to note that time series dependencies within our data make a direct linear regression between these two variables without accounting for autocorrelation tenuous. Therefore, due to the limited size of our data and the constraints of using single-variable prediction methodologies, we cannot precisely determine how increases in housing prices relative to the national level might influence homelessness rates.

Instead, the results of our bootstrap difference in means analysis gave statistically significant evidence that states that had housing prices above the national average in the period from 2008 to 2017 had a higher mean rate of homelessness. As discussed previously, this relationship can be due to multiple factors. First, as housing prices rise, they will force fiercer competition for the worst stock of housing. As this process continues, some people with the lowest ability to pay for housing will find themselves facing a choice between paying exorbitant costs for dangerous or dilapidated housing or simply not paying for housing and becoming homeless. The fact that housing prices were the factor that proved most influential may suggest that other factors that other research has shown to be correlated with homelessness like drug addiction may be more symptomatic of homelessness rather than a cause, although more research like a multivariate regression accounting for time series dependencies would be needed to test this hypothesis.

Furthermore, the fact that the states with higher housing prices were shown to have a higher mean homelessness rate may also be indicative of a more complex causality behind homelessness. As shown by our choropleth maps, states with high housing costs also tended to have higher rates of unemployment, higher inflation, and higher rates of poverty and addiction. As such, it may be that high housing costs are simply positively correlated with these or other variables that cause homelessness, and thus states with high housing costs will also have high rates of homelessness. Again, a more rigorous multivariate regression accounting for time series dependencies would be needed to determine which of these variables is actually significant and which are confounding.

However, we are left with the impression that states with high housing costs also have high rates of homelessness. Whether or not the relationship is causal or a mere correlation, states should take immediate action to lower the costs of housing in an attempt to alleviate the crisis of homelessness. These actions could be supply-side – subsidizing homebuilding, relaxing zoning laws to allow for more density, or updating the stock of dilapidated housing – or demand-side – rent assistance, temporary free stabilizing housing, or incentivizing job growth in areas with ample housing supply – but action is needed now to reduce the rate of homelessness across the country.

VII: References

1. Coombs, T., Abdelkader, A., Ginige, T., Van Calster, P., Harper, M., Al-Jumeily, D., & Assi, S. (2024). Understanding drug use patterns among the homeless population: A systematic review of quantitative studies. *Emerging Trends in Drugs, Addictions, and Health*, *4*, 100059. https://doi.org/10.1016/j.etdah.2023.100059
2. Federal Reserve. (2024). *State Debt-to-Income Ratio, 1999—2024* [Dataset].
3. Herreno, J., Nakamura, E., & Steinsson, J. (n.d.). *State Consumer Price Index* [Dataset]. https://sites.google.com/view/jadhazell/state-consumer-price-index
4. Heston, T. F. (2023). The Cost of Living Index as a Primary Driver of Homelessness in the United States: A Cross-State Analysis. *Cureus*, *15*(10), e46975. https://doi.org/10.7759/cureus.46975
5. *How many homeless people are in the US? What does the data miss?* (2024, March 28). USAFacts. https://usafacts.org/articles/how-many-homeless-people-are-in-the-us-what-does-the-data-miss/
6. mexwell. (n.d.). *US Drug Abuse* [Dataset]. Kaggle. https://www.kaggle.com/datasets/mexwell/us-drug-abuse
7. Omdena. (n.d.). *2007-2022 Homeless Populations by State (USA)* [Dataset].
8. Quigley, J. M., & Raphael, S. (2001). The Economics of Homelessness: The Evidence from North America. *European Journal of Housing Policy*, *1*(3), 323–336. https://doi.org/10.1080/14616710110091525
9. Shrider, E. A. (2024, September 10). *Poverty in the United States: 2023*. United States Census Bureau. https://www.census.gov/library/publications/2024/demo/p60-283.html

10. U.S. Bureau of Economic Analysis. (n.d.). *SARPP Regional price parities by state* [Dataset]. https://apps.bea.gov/itable/?ReqID=70&step=1&_gl=1*6yae1v*_ga*MTk2NTAwMTU1Mi4xNzMyMjQ3Mzc0*_ga_J4698JNNFT*MTczMjI0NzM3My4xLjEuMTczMjI0NzgxNC4xNy4wLjA.#eyJhcHBpZCI6NzAsInN0ZXBzIjpbMSwyOSwyNSwzMSwyNiwyNl0sImRhdGEiOltbIlRhYmxlSWQiLCIxMDEiXSxbIk1ham9yX0FyZWEiLCIwIl0sWyJTdGF0ZSIsWyIwIl1dLFsiQXJlYSIsWyIwMDAwMCJdXSxbIlN0YXRpc3RpYyIsWyItMSJdXSxbIlVuaXRfb2ZfbWVhc3VyZSIsIkxldmVscyJdXX0=

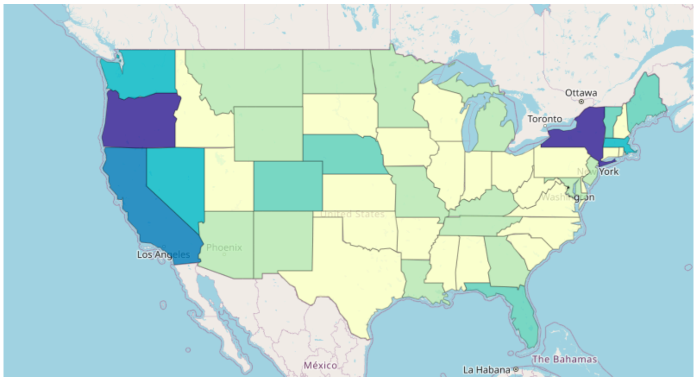
11. U.S. Bureau of Labor Statistics. (n.d.). *State Employment and Unemployment* [Dataset]. https://fred.stlouisfed.org/release?rid=112

12. U.S. Census Bureau. (n.d.). *Historical Poverty Tables: People and Families—1959 to 2023* [Dataset]. https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html

Appendix

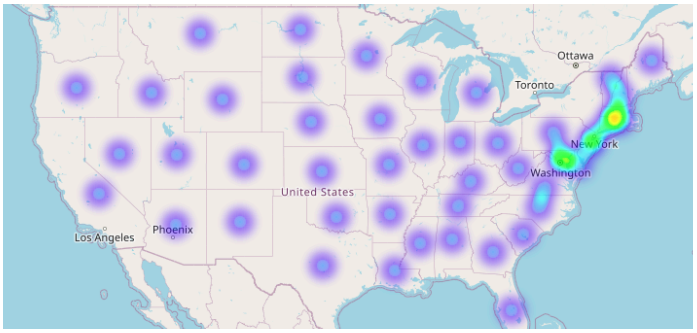
1. Homelessness Rate Integrated on USA Map.

    - Darker states mean that the rate of homelessness is higher.



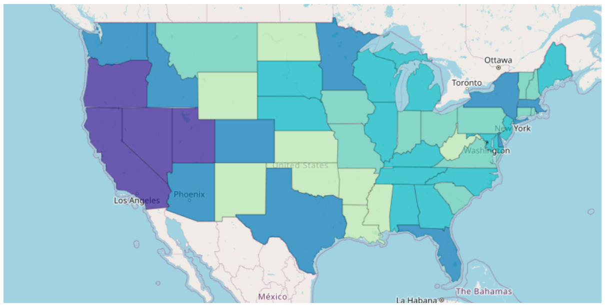
2. Homelessness Rate Integrated on Heat Map of the USA.

   - The bigger the heat wave the bigger the rate of homelessness.



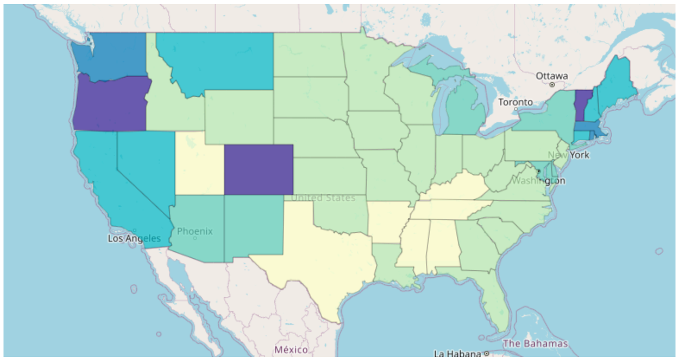
3. Poverty Rate Integrated on the USA Map.

   - Darker states mean that the rate of homelessness is higher.



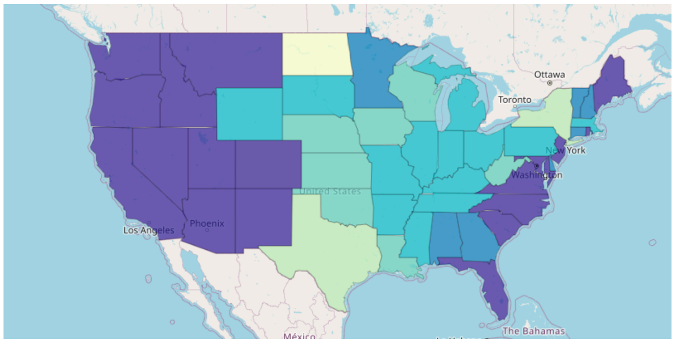
4. Drug Consumption Rate Integrated on the USA Map.

   - Darker states mean that the rate of homelessness is higher.



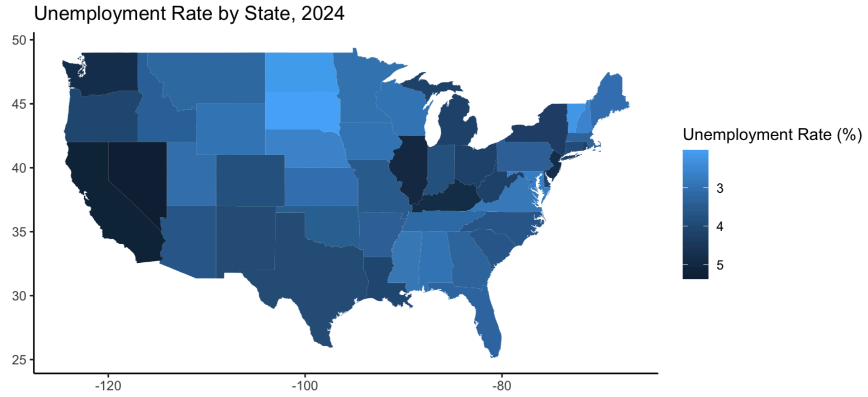
5. Debt vs Income Rate Integrated on the USA Map.

   - Darker states mean that the rate of homelessness is higher.



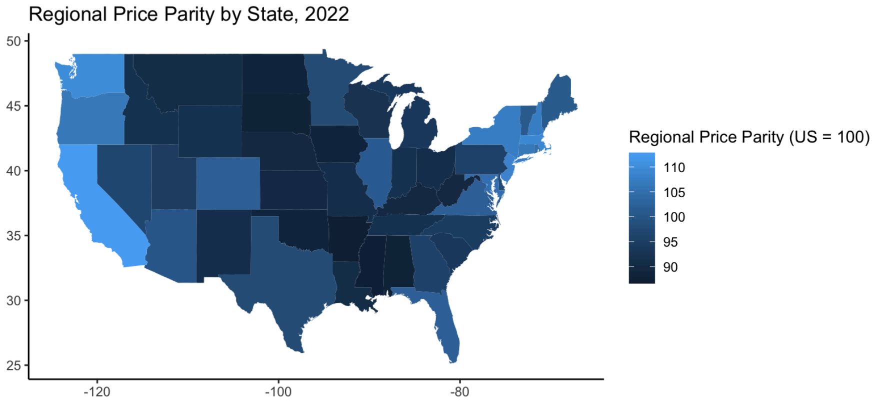
6. Unemployment Rate Integrated on the USA Map.

   - Darker states mean that the rate of homelessness is higher.



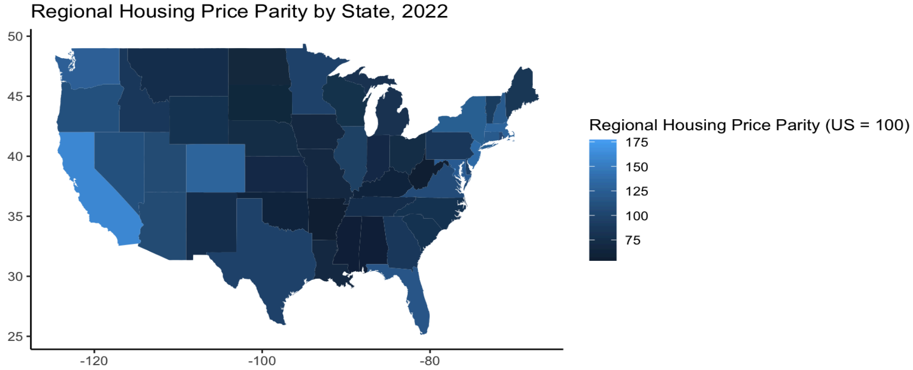
7. Regional Price Parity Integrated on the USA Map.

   - Darker states mean that the rate of homelessness is higher.


Regional Price Parity by State, 2022
50
45
40
35
30
25
-120
-100
-80
Regional Price Parity (US = 100)
110
105
100
95
90

8. Regional Housing Price Integrated on the USA Map.

    - Darker states mean that the rate of homelessness is higher.


Regional Housing Price Parity by State, 2022
50
45
40
35
30
25
-120
-100
-80
Regional Housing Price Parity (US = 100)
175
150
125
100
75