Group 108: Predicting Homelessness using Previous Year Housing and Economic Indicators

Varshini Subhash, Jailyn Clark, Caitlin Lau, Jason Friedman

1) Background and History

According to a nationwide "point-in-time" count from January 2020 there were 580,466 people experiencing homelessness on a single night in the United States. This number represents a 2.2% year-over-year increase and is still likely to be an undercount of the true homeless population due to observation bias.

The recent nationwide uptick in homelessness may be attributed to a number of interrelated factors including but not limited to skyrocketing real-estate prices and a shortage of affordable housing. As a result of these economic trends, large urban centers (some of the most populated parts of the United States) are becoming increasingly unaffordable to low-income individuals. Consequently, some of the wealthiest, most-sought after cities and states also have the largest per-capita homeless population.

Homelessness is not a new phenomenon in the United States, having first occurred on a large scale following the Civil War. More recently, rates of homelessness have grown dramatically following the financial crisis of 2008 and the wave of service-sector layoffs sparked by the Covid-19 pandemic. We seek to model the relationship between homelessness and its main economic drivers in order to predict and explain future trends.

2) Description of the Data

We begin our study of homelessness and housing in the United States by exploring three datasets, each of which offers a unique lens on recent trends and determinants of homelessness.

The **first** dataset **housing.csv** contains data that was collected from approximately 500 Boston suburbs in 1978. It includes the following housing infrastructure features: median home value, average number of rooms in a house, percentage of low status population and pupil-teacher ratio by town. Although this localized dataset is not as recent as the homelessness data discussed below, we believe that it still offers valuable information on the determinants of housing affordability that generalize and inform broader national trends in homelessness.

The **second** dataset, **homelessness_2007_2016.csv,** contains data collected and published by the United States Department of Housing and Urban Development (HUD). It provides point-in-time estimates of the size and type of homeless individuals in cities and districts that received federal aid from 2006 to 2017.

Dataset 1: *housing.csv*

- **Source**: https://www.kaggle.com/schirmerchad/bostonhoustingmlnd
- **RM**: Average number of rooms per dwelling.
- **LSTAT**: Percentage of population considered lower status.
- **PTRATIO**: Pupil-teacher ratio by town.
- **MEDV**: Median value of owner-occupied homes.

Dataset 2: *homelessness_2007_2016.csv*

- https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/
- **Source**: https://www.kaggle.com/adamschroeder/homelessness-comparison-between-states/data
- This dataset captures a *state-level aggregation* of each category of homelessness between the years 2007 to 2016.
- The columns 'CoC Number' and 'CoC Name' describe the same information as the column 'State'.
- **State**: State under consideration.
- **Year**: Year under consideration.

- **Measures**: Category of homelessness.

Dataset 3: *state.csv*

- **Source**: https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html
- This dataset captures a *state-level aggregation* of each category of homelessness between the years 2007 to 2030.
- **State**: State under consideration.
- **Year**: Year under consideration.
- **Measures**: Category of homelessness.

Dataset 4: *Census Population 2000-2020*

- **Source**: https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html
- **State**: State under consideration.
- **Year**: Year under consideration.
- **Population**: State population

Dataset 5: *Census Housing Estimates 2000-2020*

- **Source**: https://www2.census.gov/programs-surveys/popest/tables/2000-2010/intercensal/
- **State**: State under consideration.
- **Year**: Year under consideration.
- **HU Estimate**: State population

Dataset 6: *Census Median Income 2000-2018*

- **Source**: https://www.census.gov/topics/income-poverty/income/data/tables.html
- **State**: State under consideration.
- **Year**: Year under consideration.
- **Median Income**: Annual median income by state

Dataset 7: *Median Home Value per city per year*

- **Source**: https://www.zillow.com/research/data/
- **State**: State under consideration.
- **Year**: Year under consideration.
- **Median Price**: Median annual home price by state

Dataset 8: *Homelessness Count by State over Years*

- **Source**: https://www.zillow.com/research/data/
- **State**: State under consideration.
- **Year**: Year under consideration.
- **Median Price**: Median annual home price by state

*Cleaned versions of Datasets 4-8 are in the google drive folder referenced in the submission comments. Each file is referenced and described in the notebook before being read in with the exception of a mapping file which converts State strings to abbreviations

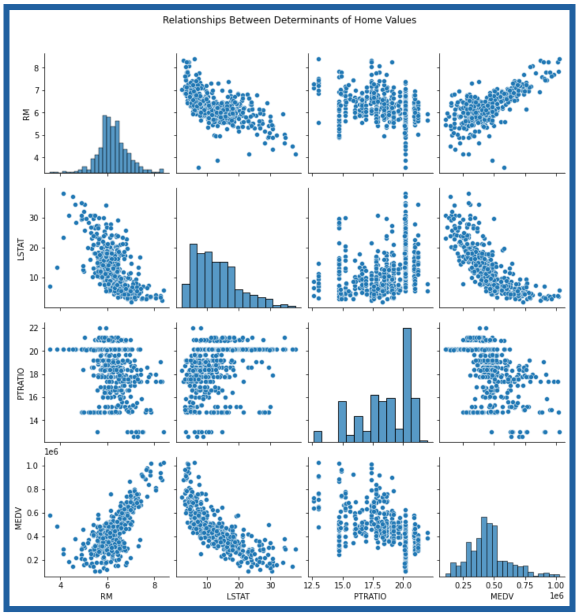
3) Visualizations and EDA

3.1 Provided DATASET 1:

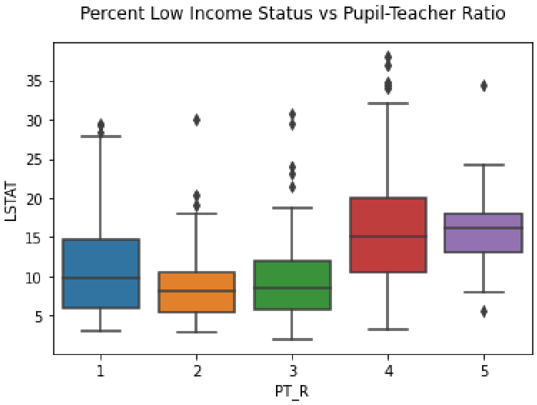
We perform a preliminary visualization of the dataset with the help of a pairplot using Seaborn, which plots all our columns pair-wise against each other. We note the following observations:

- The average number of rooms per dwelling (RM) appears to be positively correlated with the median value of the house (MEDV), which reflects the increase in cost of houses as the number of rooms increases.
- The percentage of low status population (LSTAT) appears to be negatively correlated with the number of rooms (RM) and the median value of the house (MEDV).
- We would expect the percentage of low status population (LSTAT) to have a relationship with the pupil-teacher ratio (PTRATIO) but the pairplot below does not seem to indicate any correlation. We would like to investigate this relationship further.

Summary of our findings: The pairplot conveys the idea that wealthier neighborhoods which are more likely to have higher rooms per dwelling (RM) and a higher median value of the house (MEDV) would have a smaller percentage of low status population (LSTAT). This reinforces the idea that as a society becomes wealthier, we see a decrease in affordable housing. This in turn impacts populations belonging to a lower socioeconomic status (described by LSTAT and PTRATIO) and drives up homelessness.



Ex.1 *In the pairplot we observe that the number of rooms is **positively correlated** with median home value. The percentage of the population considered low status is **negatively correlated** with median home value. The percentage of the population considered low status **negatively correlated** with number of rooms. The percentage of the population considered low status data is **left skewed** and Pupil Teacher Ratio data is right skewed. We would assume prior to observing the data that there would exist a **stronger/ more visible relationship** between the pupil teacher ratio and the percentage of the population that is considered low status. This assumption is due to the fact that historically schools in lower income neighborhoods have higher student to teacher ratios. However our pair plot **does not support** this assumption in its current state. Our choice in order to determine whether a relationship exists is through binning the values of pupil teacher ratio and using a box plot to determine if there is any relationship.* **After reviewing the pair plot, feature descriptions and dataset origins we have determined that insights from this dataset will be useful for demonstrating the motivation behind the housing problem behind homelessness. This will be achieved by observing how housing prices (the median value of owner occupied homes) are impacted by the number of rooms in a home, the percentage of the population considered low status and the pupil teacher ratio.**



Ex.2 *In the box plot we observe that beginning at the bin for 17-18 students per teacher, each subsequent grouping has a higher median value for the percentage of the population considered low status. This indicates that the relationship that we originally inferred holds true for this dataset.*

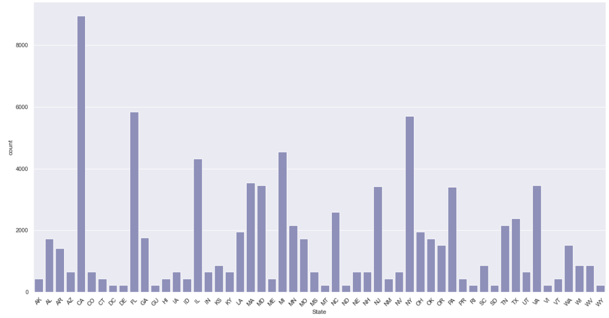
3.2 Provided DATASET 2:

We begin visualizing this dataset by plotting the count of homelessness for each state. We then plot the count of each category of homelessness and we note the following observations for both plots:

- In the plot ‘State versus Count’, we notice substantially higher numbers of homelessness in states such as California, Florida, Illinois, Michigan, New Jersey, New York, Pennsylvania and Virginia which might be due to the high presence of metropolitan cities in such states. We notice lower numbers of homelessness in relatively less populated states like Alaska, Washington DC, Delaware, Guam, Montana, North Dakota, Rhode Island, South Dakota, Virgin Islands, Wyoming.
- In the plot ‘Frequency of Types of Homelessness (Measures)’, we see the variation in the categories of homelessness and note the highest counts in chronically homeless populations.

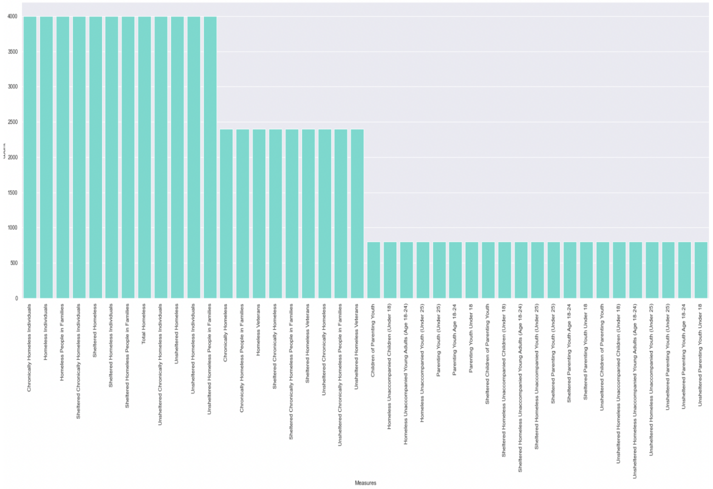
Summary of our findings: The plots convey the idea that states with more metropolitan cities seem to have higher counts of homelessness and the categories with highest counts are the ones for chronic homelessness. In general, we can conclude that chronic homelessness seems to be highly prevalent in states that have higher populations and a high presence of metropolitan cities.

State versus Count



Ex.3 *Above image is the count per state, where count is the number of homeless people identified as homeless from CoC programs.*

Frequency of Types of Homelessness (Measures)



Ex.4 *The above image is the count of the number of people identified within each measure of homelessness in the dataset obtained from CoC programs.*

3.3 Provided DATASET 3:

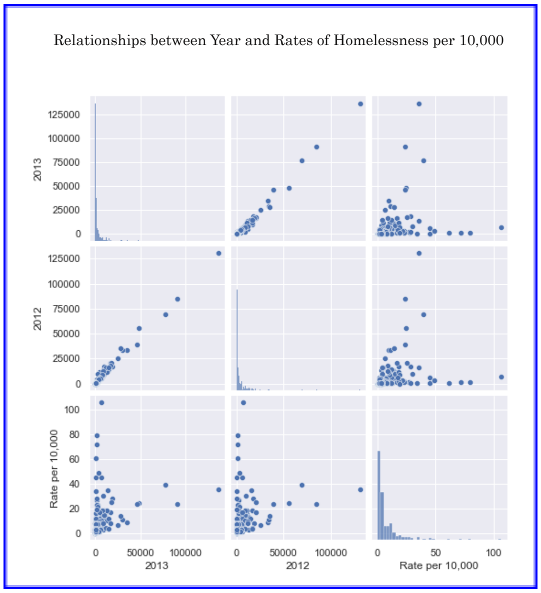
We repeat our preliminary pairplot visualization for the third dataset, which plots all our columns pair-wise against each other. We also plot the counts per category of homelessness across different states for both years - 2012 and 2013.

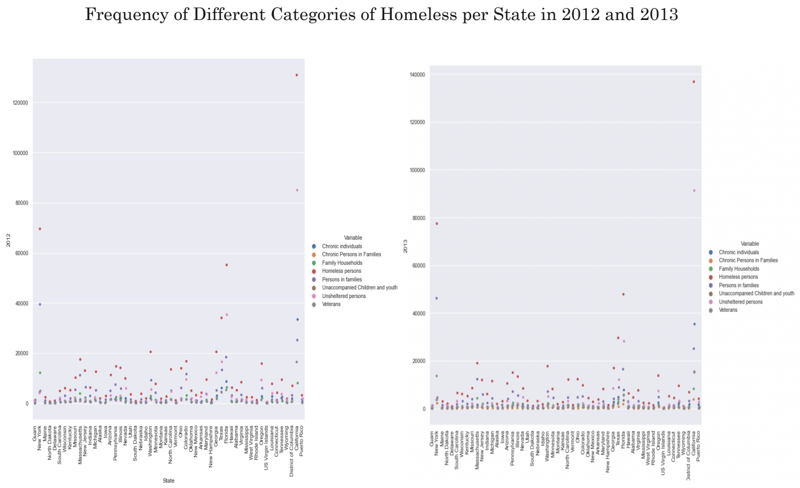
We note the following observations:

- We notice a clear positive correlation between the columns ‘2012’ and ‘2013’ which indicates that homelessness increased during the period 2012 to 2013.
- We see the category ‘Homeless Persons’ show the highest numbers across all states, for both years - 2012 and 2013. We see the lowest numbers in the category ‘Veterans’. California shows the highest counts for homelessness, which matches the conclusion we drew from Dataset 2.

- We notice that the category for homeless people has much higher numbers than chronic homelessness which is not what we observed in Dataset 2. This could possibly be due to different methods of categorization employed by the dataset sources. We also note that this could be a potentially interesting line of exploration, in the event that there is an underlying reason for this discrepancy.

Summary of our findings: The plots convey the idea that there was a definitive increase in homelessness from 2012 to 2013.The chronic homelessness category does not show the highest counts, unlike the previous dataset where we noticed clear dominance of chronic homelessness across all states. We note that this discrepancy could be due to the method of dataset collection or difference in categorization. This discrepancy could also be due to an underlying reason which we could explore further.





Ex.5 (left) *Initial pairplot of the data which shows positive correlation between 2012 counts and 2013 counts for homelessness. Also can see that for each numerical feature all data is left skewed.*

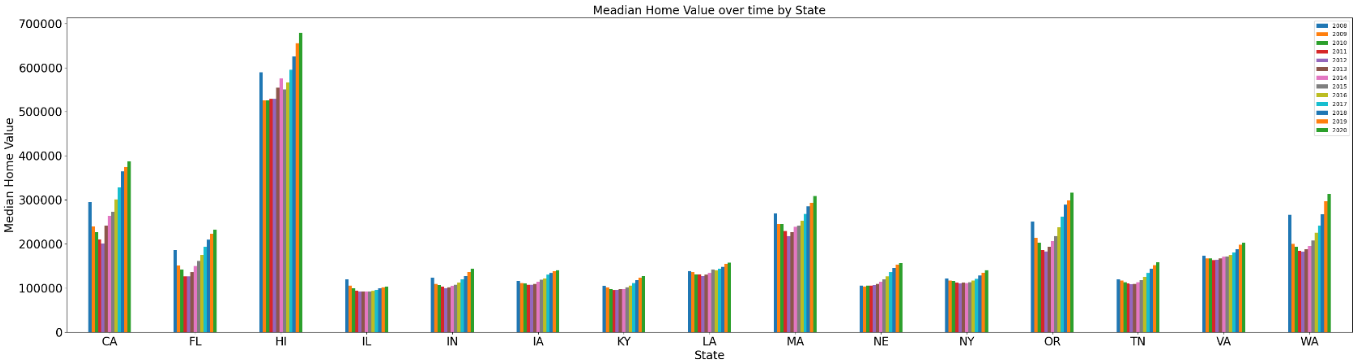
Ex.6 (right) States *with large Metropolitan cities have larger counts of all categories of homelessness. Consistently the category "Homeless persons" has the largest number of people.*

3.4 Combined Dataset:

Our combined dataset aggregates the 5 datasets representing: the population of each state over time (feature: POP), the count of homelessness within each state over time (feature: COUNT, perc_hml, and Share which was not used in the model), the median value of homes per state over time (feature: MEDV), the median income per state over time (feature: INCOME), and the estimated number of houses in each state over time (feature: HU or HUESTIMATE). We received all of our data from reputable sources including the United States Government for accurate numbers based on the Census and Department of Housing and Urban Development. When getting all the columns for the dataset we also calculated but chose not to use a variable we named share which was the percentage of homelessness each state represented within the country.

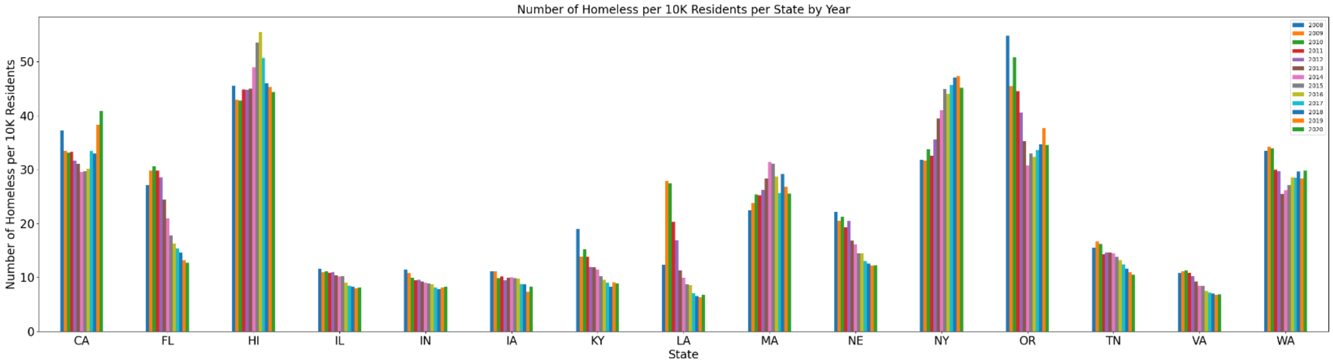
Upon completion of merging all data we explored the data set with various visualizations. These visualizations demonstrate trends in our dataset that could potentially allow us to better understand performance when modeling using this data. The following plots use a subset of states for the benefit of our reader's ability to view and interpret. Our first plot of Median Home Value over time by state demonstrates that for most states the value of homes decreases up until 2010 and then gradually increases. We recall that the United States suffered a subprime mortgage crisis between 2007 and 2010 which explains the decrease and the eventual increase. Our second visualization of Number of homeless per 10,000 residents shows that each state has varying trends over time. Interestingly, our plot for the person-to-home ratio appears to be nearly identical for each state. This allows us to reaffirm our initial conjecture that states with larger populations will have higher counts of homelessness. Our final grouped barplot of data features is the median state income per year. As aligns with assumptions, median income typically increases for each state each year.

Median Home Value Over Time by State



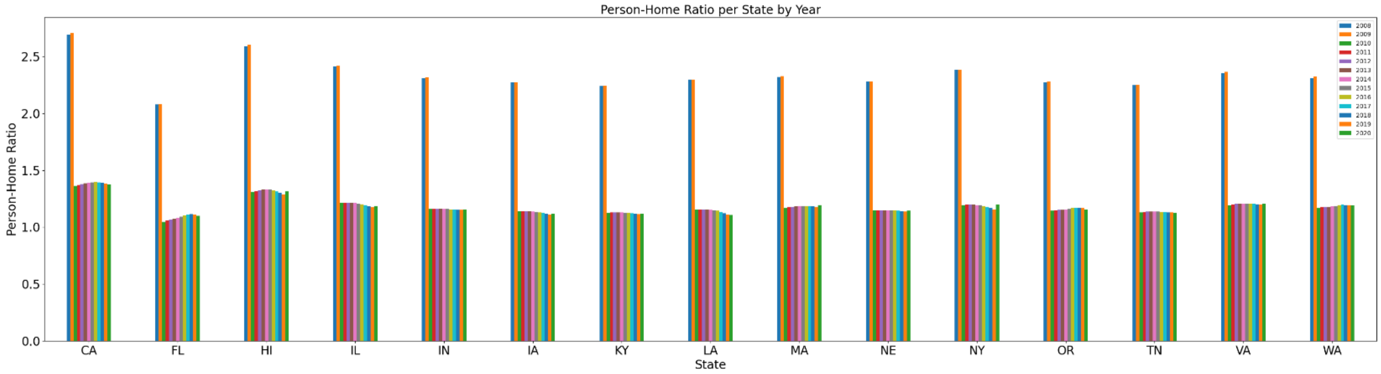
Ex.7: We plot the median home values for a few salient states, which seem to decrease initially, followed by a subsequent increase. Specifically, California, Florida, Hawaii, Massachusetts, Oregon and Washington show steep increases in home values, indicating a potential scarcity of homes and high demand.

Number of Homeless per 10K Residents per State by Year



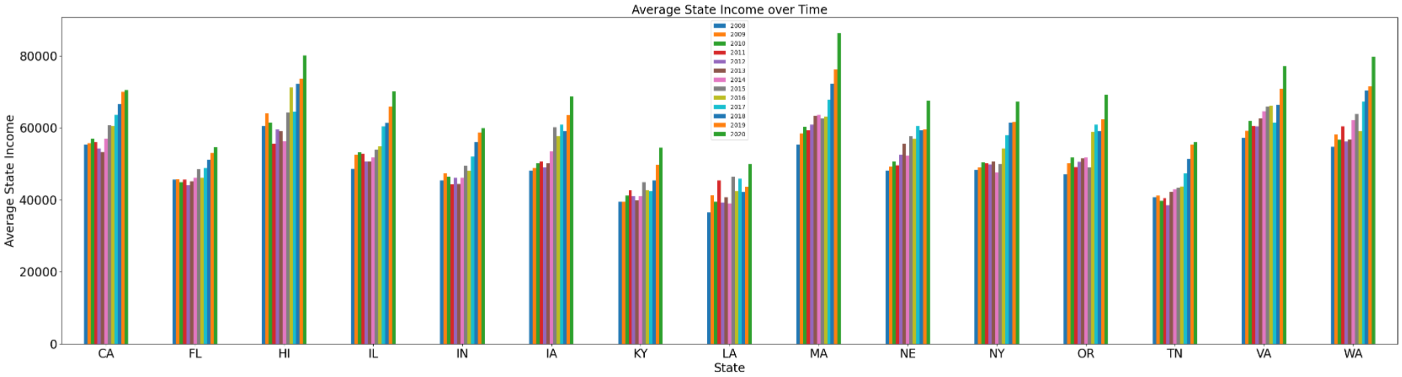
Ex.8: We plot the count of homeless people per 10000 residents per year which is a population normalized feature describing homelessness. We notice varying trends, where California, Massachusetts, New York and Washington show increasing trends while Florida, Illinois, Louisiana, Nebraska, Tennessee show prominent decreasing trends.

Person-Home Ratio per State by Year



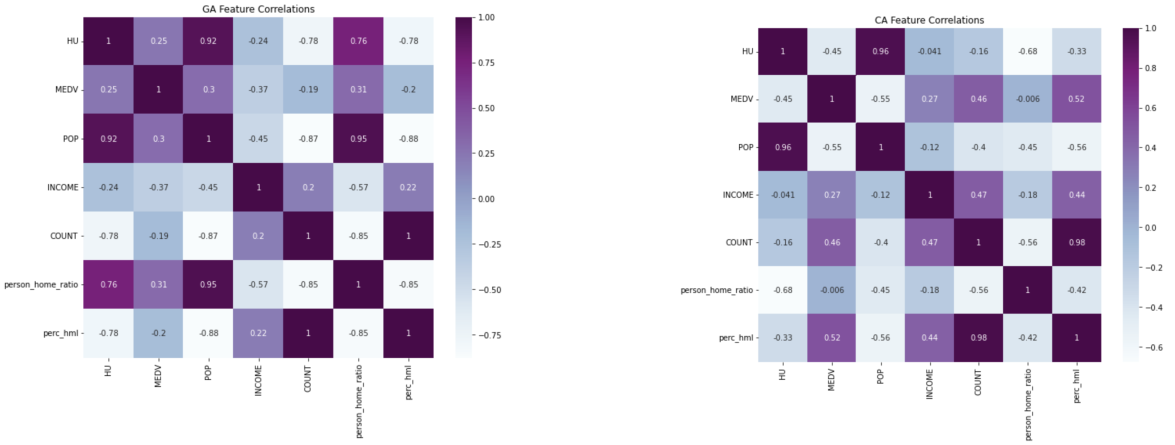
Ex. 9: We plot the person-home ratio which is the ratio of the population and the number of housing estimates in the state. This quantity indicates the relationship between the people living in the state and the available housing. We notice disproportionately high values during the years 2007 and 2008, which could potentially be due to the large number of individuals who lost their homes during the financial crisis of 2008. We see more stable numbers for the rest of the years across the states. Florida sees an increase which could indicate that while the population saw an increase, the available housing has been limited and not grown at the same rate. For states such as California, Hawaii, Louisiana and New York which show slightly decreasing trends, we could conclude that while the population seems to be increasing, the available housing might be increasing at an even higher rate. However, we are not certain about how this housing is distributed among the population. There is likely a large shortage of housing in densely populated, urban areas in these states but we would require further investigation in order to confirm this hypothesis.

Median State Income per State by Year



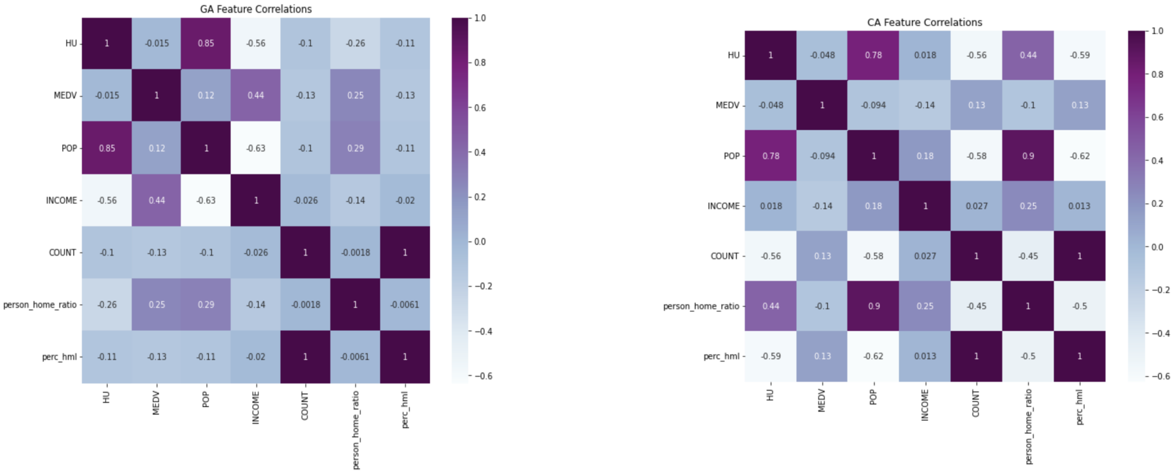
Ex.10: Here, we plot the average state income over time to decipher if the population in a given state can afford the available housing on average. We see an increase in average state income over time in all states but states such as California, Hawaii, Illinois, Indiana, Massachusetts, New York and Tennessee show sharper rates of increase in state income. This indicates that these states are generating jobs and more average income, which could be providing incentive for people to move into these cities and drive up the population. This could increase the homelessness count because we have a high demand for houses and a scarcity of the same.

Correlation plot of Nominal Values for Georgia and California



Ex.11: Here, we have correlation plots of each predictor in 2 different state datasets to show that as we know from examining the given first dataset and based on historical trends, our predictors are correlated. Count and percent of homeless in state are 100% correlated because percent of homeless in the state is the Count predictor normalized for state population.

Correlation plot of Year over Year Percent Change values for GA and CA



Ex.12: Here, we have correlation plots of the year of over year percent change of each predictor in 2 different state datasets to show that as we know from examining the given first dataset and based on historical trends, our predictors are correlated however using year over year percent change works to decorrelate some features.

Above we have plots of the correlations for each feature examined on a state level. What we observe is a confirmation of our assumptions from our initial model, that our predictors are correlated. We also show that since these plots only have predictors for one year that our predictors within the year we want to predict homelessness when nominal are correlated with percentage of homelessness in the state but when transformed using percent change they are not as useful in predicting homelessness. We mention this because we initially planned to use percent change to transform our entire dataset and ultimately chose not to citing this lack of correlation reasoning behind our decision.

4) Discussion of our base model

When using only the first three datasets we created a base model. Our base model offers the motivation and intuition for our research question. This simple multi-linear regression model reveals the relative influence of the proportion of low socioeconomic inhabitants, size of available housing, and quality of education on the value of homes in the United states. Our model reveals the following:

- All else equal, if the number of rooms in a home increases by 1 the median value of the home increases by approximately $90,000
- All else equal, if the percentage of the population considered low status increases by 1 percentage point, the median home value decreases by approximately $10,000
- All else equal, if the student teacher ratio increases by one student, the median value of an owner occupied home decreases by approximately $20,000

These relationships imply that the quantity and quality of social support and housing infrastructure have a predictable association with the home values in a particular neighborhood at a certain point in time. As a neighborhood becomes more wealthy, this decreases the number of affordable houses in the area which displaces lower income populations, which in turn increases homelessness in the long run (albeit in different geographic locations). In our forthcoming analysis we plan to measure the relationship between average home value and the amount of homelessness in a community. We hypothesize that gentrification has a delayed effect on homelessness. That is, in the short term, we predict gentrification decreases local homelessness but in the long run it increases homelessness in surrounding communities and in the aggregate.

To test this hypothesis we will analyze trends over time in average home prices and the size of the homeless populations at the state level and city level (pending available data) to better understand the relationship between the two. We hope that our results will help inform more efficient allocation of resources aimed at improving the social safety net in America given its existing social and economic dynamics. Understanding the relationship between home price and homelessness can help us identify and perhaps even preempt unfavorable trends. The question we seek to answer is can the change in home values and other predictors over time serve as appropriate predictors of homelessness trends.

5) Model

5.1 Modeling Approach

Our data has a time-based order due to all of our features and response variables having data spanning the period of time from 2000 to 2020. The years that are common to our predictors and response variables are 2007 to 2018. In order to create models for our data over this 11 year time frame we had to reframe our time-series data for supervised learning. We were able to achieve this by implementing a shift transformation on our data set, then applying a time-series split (instead of a train-test split) in order to split the data into training and testing sets and still preserve the time-based ordering of the data.

In order to model our data robustly we used 5 different models. The first model we used was Multi-Linear Regression. Our initial exploratory data analysis on the given datasets demonstrated that a Linear regression model was able to measure the data and demonstrated that predictors such as median income, median home value, the number of homes, the person to home ratio, population can all predict homelessness, we also saw with our third given dataset that previous year homelessness is a good predictor of future homelessness. Our second model is a Multi-linear regression model with polynomial features of degree two. We observed in the pair plot from Dataset 1 of the given data that the feature that represented the percentage of the population that was categorized as 'lower status' was negatively correlated with the feature that represented the median home value. This negative correlation appeared to have a polynomial shape. Using polynomial features could be potentially beneficial when modeling for states that have steeper increases or declines in the percentage of homelessness populations when predicting with each feature. Our third model is a Random Forest Regressor. This model was useful on a state level as it helped us determine the features that were determined most important for each state and the splits are interpretable by humans, Random Forest also picks a random subset of features every time in order to de-correlate the trees. Our fourth model is a Lasso regularized linear model. We set our alpha level to be 0.01 for all states as a constant because that is what was determined to be most consistently optimal when testing randomly on various states. The use of a Lasso regularized linear model allowed us to tune hyper-parameters, which maximized the model performances without overfitting and reducing variance error. Our fifth and final model is a Boosting Regressor. The method of boosting fits several shallow Decision Tree Regressors, the first tree is fit to the model and the subsequent trees are fit to the residuals of each previous tree.We expect that boosting over several iterations will fit very well to the data if not overfitting on the training set.

For each model we used only the features that existed in the previous year, that is, if predicting homelessness in 2012 we used population, estimated number of houses, median income, median home value, person to home ratio, count of homelessness, percent of state that is categorized as homeless, and person to home ratio from 2011. We do not use anything from the year 2012. Our assumption in our project is that homelessness is something that is best predicted over time as the circumstances that cause homelessness might have a delayed reaction.

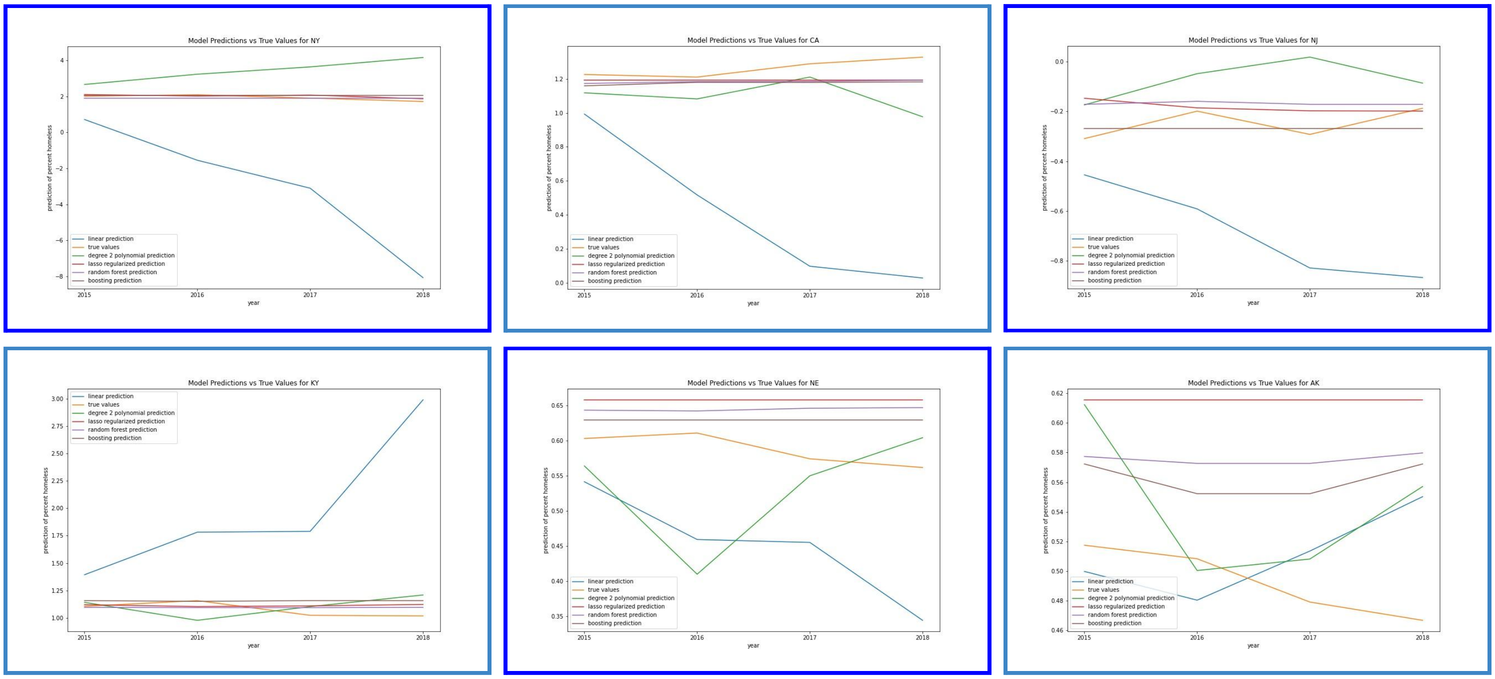
5.2 Model Performance:

For each state we recorded and stored the mean squared errors from each model in a list that was converted into a dataframe. For approximately one-third of the states, the mean squared errors for all models are fairly small. This means within those states the previous year homelessness predictors and values measuring homelessness are great predictors across all models. For other states, each model has varying performance. We were able to receive counts of how many states chose each model as the best model and a list of states that chose each model as the best model. The counts were given as:

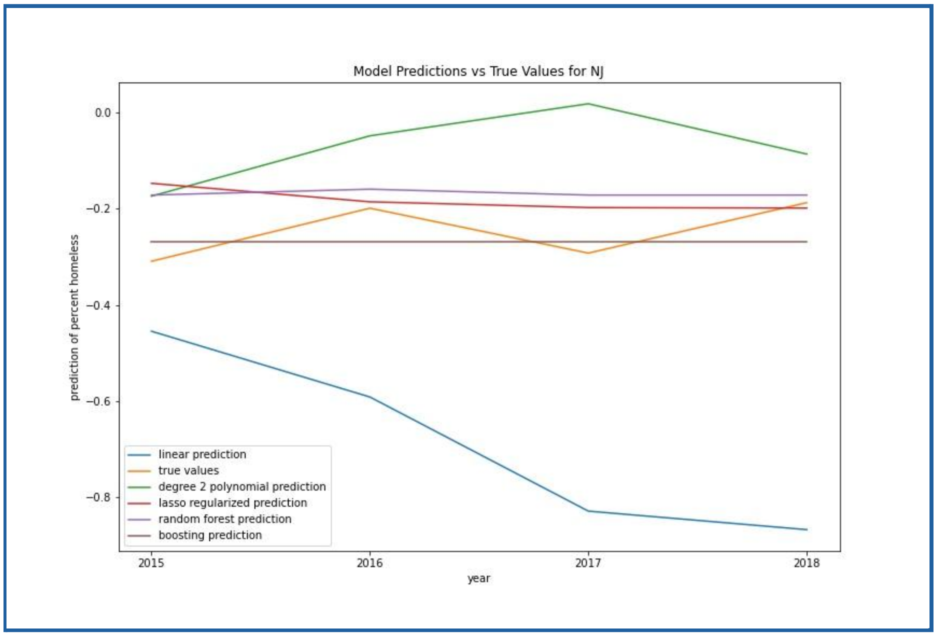
- The number of states that prefer the linear model is: 4.
- The number of states that prefer the polynomial of degree 2 model is: 4
- The number of states that prefer the random forest model is: 10
- The number of states that prefer the lasso linear model is: 9
- The number of states that prefer the boosting model is: 23

From this we are able to easily observe that the ensemble methods performed best overall, we know that this is due to underlying factors of homelessness and the fact that ensemble methods are able to capture the dynamics of population growth, changes in housing availability, and changes in housing price for states with higher populations and more aggressive changes year over year, while linear models will predict best for states where there is less dynamism. This is evidenced in our result visualizations. In our plot of the mean squared error we show visually that the mean squared errors for one-third of states are fairly small. This plot also provides visual representation of the fact that the linear regression model does not perform best for most states and most of the time is the worst performing model. In our other model visualizations we are able to observe visually the performance of each model over the final 4 years represented in our data set. When displaying the list of states (by their abbreviations and their preferences in terms of best model, we are able to observe which types of states prefer what model and what inferences we can make from that. In viewing the 4 states that prefer the simple multiple linear regression model, we observe that 3 of them are the least populous states in the country and thus due to lower population there is less change in the features over time that needs to be tuned in the model so the simplest model is best. We also observe that states that prefer the lasso linear model are states that are home to some of the largest US cities (CA, NY, FL). These states need a regularized linear model as the trends for homelessness are linear but have underlying factors best captured when a model is regularized. Finally as initially states random forest and boosting are ensemble methods that fit more to a particular states specific and unique trends, for states that are in between most populous and least populous, this fits our assumption that an ensemble method would be the best model to predict for future homelessness.

6) Model Visualizations and Results



Ex.13: Here we display the model predictions vs the true values for the 6 states :New York, California, Alaska, Nebraska, New Jersey, and Kentucky. We observe that the simple multi-linear regression line in blue tends to perform the worst for all states except Alaska where it performs the best. We also observe that for 5 out of the 6 states shown the lasso linear model, random forest model, and boosting model are all close in terms of predictions and in terms of distance to the true values



Ex.14: Here we examine New Jersey up close as it is one of the 23 states which had the lowest mean squared error on the testing set using a boosting model. Plots for other states which had the boosting model as its best model also are visually similar to the plot for New Jersey as well as they have the common trend of the linear regression model performing the worst.

Overall we aggregated states by

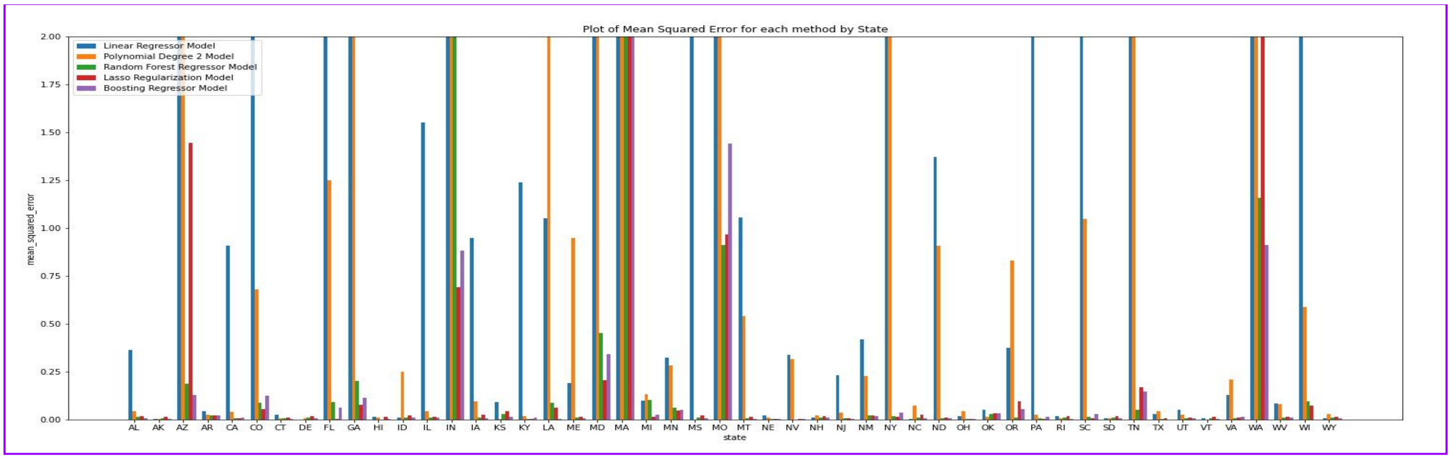
preferred model based on the model's mean squared error.

Below we share the states (in list form) grouped by model preference

- The states that prefer the linear model are: ['AK', 'DE', 'NC', 'WY'].
- The states that prefer the polynomial of degree 2 model are: ['KS', 'MS', 'OK', 'VT']
- The states that prefer the random forest model are: ['AR', 'HI', 'IL', 'KY', 'MA', 'MO', 'NV', 'OR', 'TN', 'UT', 'VA']
- The states that prefer the lasso linear model are: ['CA', 'CO', 'FL', 'GA', 'IN', 'MD', 'MI', 'MN', 'NY', 'PA', 'SC']
- The states that prefer the boosting model are: ['AL', 'AZ', 'CT', 'ID', 'IA', 'LA', 'ME', 'MT', 'NE', 'NH', 'NJ', 'NM', 'ND', 'OH', 'RI', 'SD', 'TX', 'WA', 'WV', 'WI']

As mentioned previously we use this to further interpret why each model performed the way it did and what characteristics of the state (population, population density, physical size, etc.) can explain the model performances.

Plot of Mean Squared Error for Each Model by State



Ex.15: Here we examine the mean squared errors as grouped bar plots per state to explore the performance across states and observe trends. What we observe is higher mean squared errors for the linear models across several states, and also the polynomial model. We also can see the states where boosting is not the best. Another insight is that the states that are not home to high population cities tend to have mean squared errors under 0.25 for all models.

7) Results and Implications

Our study confirms that homelessness is a predictable phenomenon that is driven by broader economic patterns. Specifically, the recent trends in decreasing housing affordability and availability have led to an increase in the homeless population. Specifically, we observe that desirable states with lots of economic opportunity and high cost of living also tend to have the highest rates of homelessness. Perhaps counter intuitively, it is the wealthiest and most liberal states like California, New York, and Massachusetts that have the largest per capita homeless population. We find random forests and boosting are the best approach for modeling these trends in the more densely populated and economically prosperous states as the models are most accurately able to capture the impact of these complex dynamics on homelessness. On the other hand, less populated and more rural states tend to exhibit more linearly predictable and less rapidly changing patterns that are better suited for a linear model. In conclusion, the heterogeneity of homelessness rates and differing economic dynamics among states imply the need for different modeling frameworks when it comes to predicting homelessness trends.

Some potential ethical implications that our models bring include the power to use this model to motivate institutional, economic, and governmental change in order to address the underlying root causes of homelessness such as building more affordable housing in the areas that are in greatest need of economic and infrastructural assistance for homeless people. However, potential misuse of our model that could exacerbate potential ethical and moral prejudices would be to make the cost of living so high as to functionally force homeless individuals to move to other states in order to survive. This can be seen when there are states that spend very little on infrastructure and governmental assistance for homeless individuals; however the states with the largest governmental spending to help homeless people like the Democratic states of California and Massachusetts have high levels of homelessness. This correlation could potentially be mistaken by bad-faith or malicious actors to pursue cuts in government funding to help Homeless people. Where in fact homeless people most likely relocate to safer havens for homeless people with more resources like Democratic States.

8) Conclusion and Potential Next Steps

Although our analysis reveals that the increase in homelessness that we have seen in recent years is predictable as a function of rising home prices and shrinking availability, it does not help us predict the impact to specific subpopulations at the local level. For example, homelessness may be going down in one city but going up in another as a result of income-families being displaced. Overall our model would be able to predict the macro level increase in homelessness but would not be able to pinpoint the causal mechanism. One future direction that we are interested in exploring is producing a similar model at the city level to better understand the impact of the economic factors in homelessness in another. It is one thing to understand the aggregate trends in homelessness, however these trends are the result of much smaller decisions and their consequences that add up to a much larger problem.

If we were to continue our work we would like to observe model performance with a year prediction shift of more than one year since our data shift was a one-year shift. We would also like to explore whether homelessness is correlated with legislative control in a state and see if homelessness could be a predictor of governor political party or majority party of the state using classification.