A Study of Homelessness in the U.S.
(sponsored by Databricks)

Tanqiu Jiang [*]
University of Rochester
tjiang17@ur.rochester.edu
Max Wang [*]
University of Rochester
twang83@ur.rochester.edu
Ziyu Xiong [*]
University of Rochester
tjiang17@ur.rochester.edu
Yichi Qian [*]
University of Rochester
yqian17@ur.rochester.edu

1. Introduction

Homelessness has gradually become a national issue and has become a core problem in the field of municipal management. In a most recent report published by the U.S. Department of Housing and Urban Development (HUD), 580,466 people experience homelessness in the U.S. on a single night in 2020, which is a 2.2% increase compared to 2019. The studies with regards to this social phenomenon are still at a limited level compared to the other social-economic issues because the comprehensive state-level study did not start until 2007. Bucker et al. [2] mainly researched the impact of being homelessness on children. Somerville et al. [7] used a qualitative approach to seek the reasons behind homelessness.

Our study focuses on researching the factors that may affect the number of homelessness. Such study is very meaningful as it not only reveals the relationships between factors of interest and homelessness, but also provides insights that help legislators and governments to tackle such problems. In approaching this research, we collaborated with the teams from *Databricks* and used *Pyspark* as our data wrangling and engineering platform, and we appreciate their generous support and insightful feedback for this project.

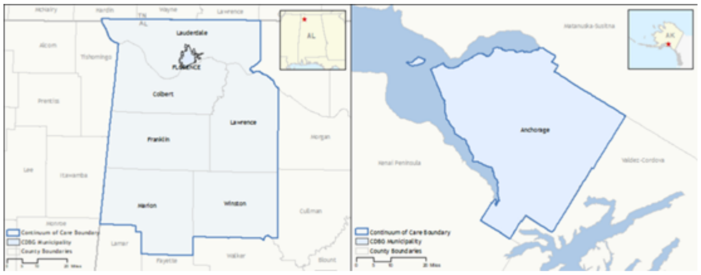


Figure 1. Examples of two different CoCs

2. Dataset Description

2.1. Main Dataset

Our main dataset is from our sponsor, Databricks, including the information of 252 COCs (Continuum of Care) and their relevant demographic information. There are 32 attributes from the original dataset. We did not utilize all of the variables for various reasons, and the most significant attributes from the dataset are: Total homeless count, sheltered/unsheltered homeless count, and city policy conservativeness rating. A number of attributes inside the dataset including population and GDP was unreliable, so we have acquired data from outside sources to supplement the original dataset. The collected data from outside sources are mostly at county level, and we manually aggregated them into CoC level. The attributes that we added to the main

[*] These authors contributed equally to this work

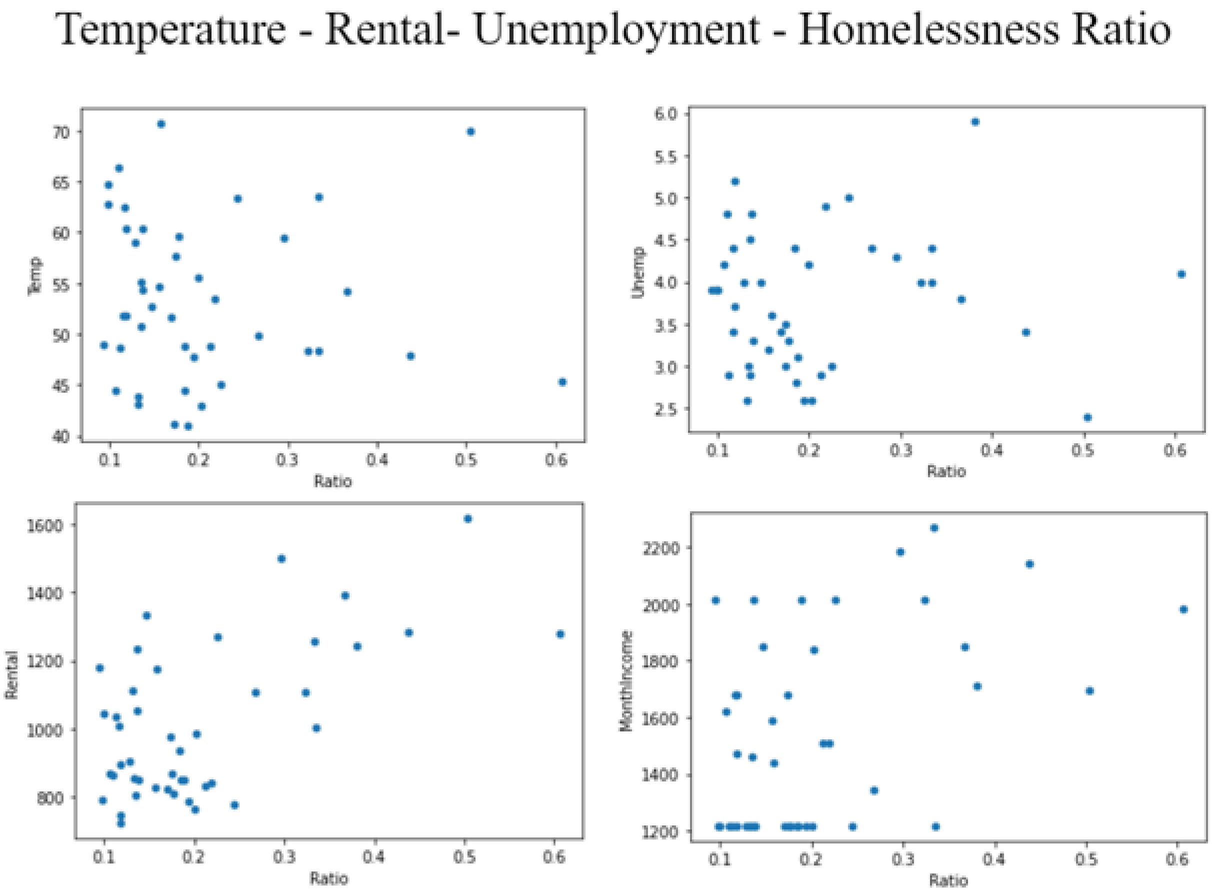


Figure 2. Dot Plot of Features of Interest against Homelessness Ratio

dataset are:

- Population: The population of each CoC [1]
- HomelessPer10000: Totalhomeless/Population (The homelessness count per 10k of population)
- GDPPERCAPITA: GDP per capita [1]
- WinterTemperature: The average temperature of each CoC during winter (Nov-Mar) [5]
- Urban: Proportion of urban population inside the CoC [1]
- Biden_share: Joe Biden's vote share in the presidential election 2020 [3]
- VoteType: Which party won the most vote count inside the CoC [3]

2.2. Base unit (CoC) and attribute aggregation

CoC continuum of care is the base unit to receive fundings from the federate government to support the homeless population. As we can see in Figure1, the sizes of CoC (continuum of care) units varies greatly across the country. The picture on the left is a CoC of six counties while the picture on the right is a CoC of the city Anchorage alone. To aggregate information such as population and GDP to each CoC, we manually checked all the CoC assignment files and then added the data base on the CoC arrangement.

2.3. Funding dataset

We scraped all the records of corresponding fundings provided to each CoC from the official website of the U.S. Department of Housing and Urban Development[1]. Our further analysis aggregated all the fundings by each CoC to investigate the relationship between homelessness and fundings.

[1] https://www.hud.gov/program_offices/public_affairs

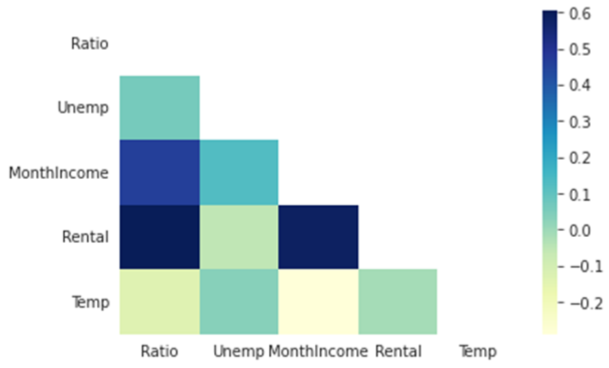


Figure 3. Heatmap Showing Correlations

3. Exploratory Data Analysis

3.1. Temperature, Rental Cost, Unemployment Rate and Minimum Wage

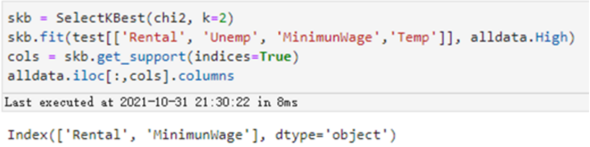
Since homeless people lack a fixed, regular and adequate nighttime residence, it is intuitive to consider high rental as a cause of homelessness. We expect a higher rental cost results in a higher homelessness ratio. By figure 2, from the plot in the left bottom corner, it indicates a positive correlation between Rental costs and Homelessness Ratio in each state. Also, we examine the correlation between Temperature and Homelessness Ratio. Since homeless people might lack enough clothes to resist the severe cold, they might be more willing to live in the south than the north. However, by figure 2, from the plot in the upper left corner, we cannot find such correlation between Temperature and Homeless Ratio. Then we check the correlations of Unemployment rate and Minimum wage in each state with Homelessness Ratio. From two plots on the right of figure 2, such correlations are not straight forward.

Then we compute the correlations of the previous 4 factors with Homelessness Ratio. We could find that Rental has the strongest correlation with Homelessness Ratio. Next is Minimum wage (MonthIncome). The other two factors, Temperature and Unemployment rate, have very weak correlations with Homelessness Ratio, since both absolute values are smaller than 0.2.

By previous analysis, as figure 3 shows, we find Rental cost and Minimum wage in each state are important factors to estimate Homelessness Ratio. To prove our idea, we perform a Chi-square test, as figure 4 shows.

3.2. Transportation Analysis

As a great leader, Deng, said: Building the road is the first step to becoming rich. This motivates us to analyze the influence of the transportation system. Since the most convenient way for homeless people to move across the city is public transportation. Also, they can move to places that



```
skb = SelectKBest(chi2, k=2)
skb.fit(test[['Rental', 'Unemp', 'MinimunWage','Temp']], alldata.High)
cols = skb.get_support(indices=True)
alldata.iloc[:,cols].columns
Last executed at 2021-10-31 21:30:22 in 8ms

Index(['Rental', 'MinimunWage'], dtype='object')
```

Figure 4. Cell of Code of Picking Strongest Pair

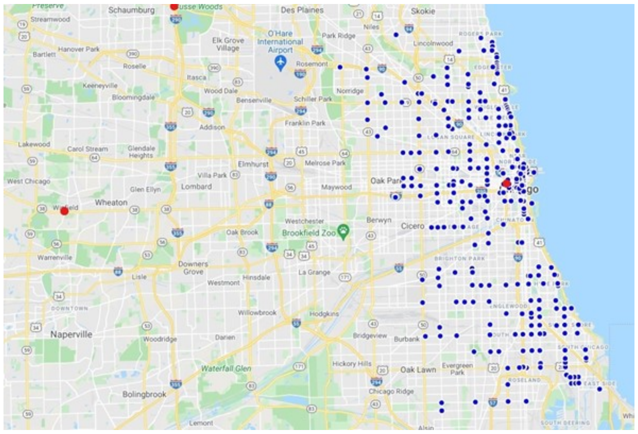


Figure 5. Distribution of Bus Stops (blue points) and CoCs (red points) in Chicago

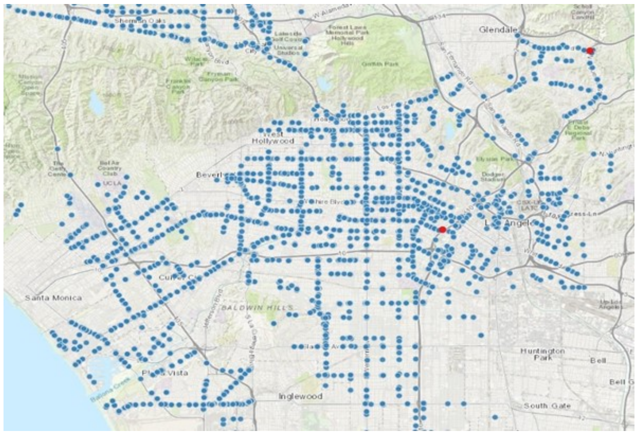


Figure 6. Distribution of Bus Stops (blue points) and CoCs (red points) in Los Angeles

provide food and necessities by public transportation. First, we look at available free long trips for homeless people. Free bus tickets are available from the Guardian that helps them move across states. By the figure 7, a larger circle means more homeless people move to that place.

Next, we check whether homeless people are more willing in a place with better public transportation. The figure 5 indicates the distribution of bus stops and CoCs in Chicago. Most bus stops are located in the downtown of Chicago, and there is one CoC here. There are two CoCs outside, one is located on the top, the other is located on the left, and no bus

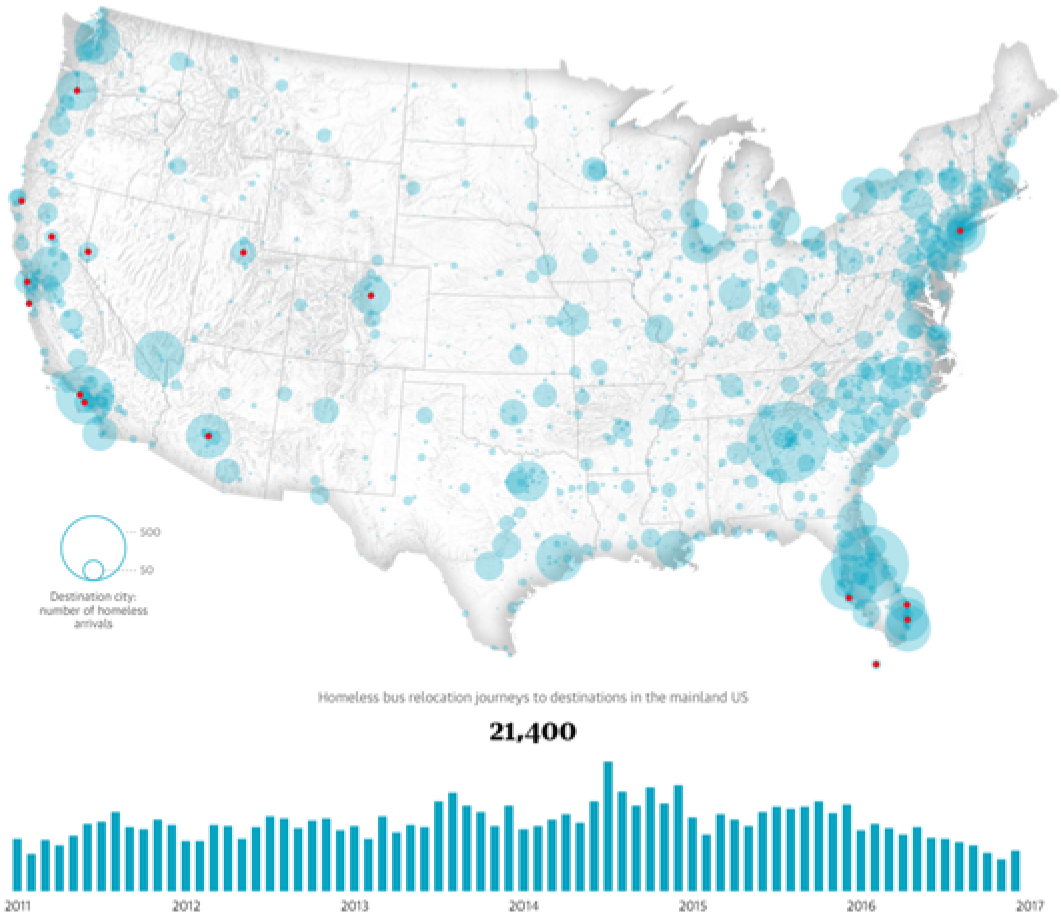

500
50
Destination city: number of homeless arrivals
Homeless bus relocation journeys to destinations in the mainland US
21,400
2011
2012
2013
2014
2015
2016
2017

Figure 7. Free bus tickets from the Guardian helps homeless people move across states

stops are near them. The figure 6 indicates the distribution of bus stops and CoCs in Los Angeles. There are 2 CoCs in LA, and both are connected by public transportation.

Then we compute the HR and SR for Chicago and Los Angeles. HR is: the homelessness population in the city over the homelessness population in the county. SR is: the size of the city over the size of the county. From the figure 8, we find that the HR of both cities are much higher than SR, especially for Chicago. There are about 90% homeless people in the county living in Chicago. This indicates that homeless people are willing to live in a place with better public transportation.

Even though we know that public transportation is attractive to homeless people, is it a key factor? We do a comparison between Los Angeles and San Diego. We do the same plot for bus stops and CoC in San Diego. We find the density of bus stops in SD is higher than LA. However, only about 9,000 homeless people live in SD, while about 44,000 homeless people live in LA. Although SD provides them with better public transportation, they still choose to live in LA. Hence, although public transportation is attractive to homeless people, it is not a decisive factor. This finding motivates us to analyze other factors, such as GDP and policies.

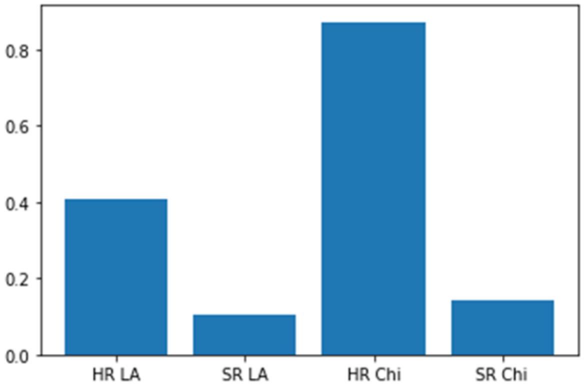


Figure 8. HR and SR of Los Angeles and Chicago

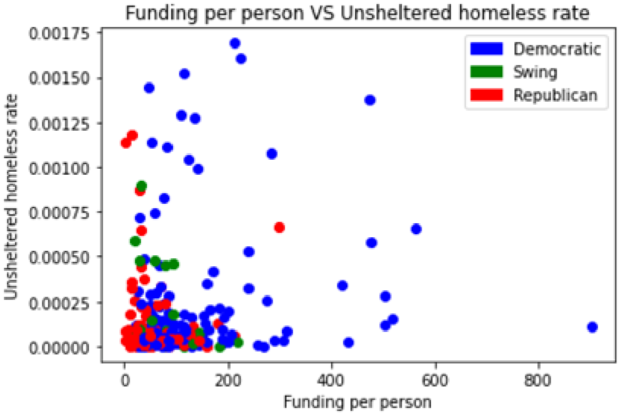


Figure 9. Dot plot of Supportive Funding against Homelessness Rate

3.3. CoC Funding and Unsheltered Homeless Analysis

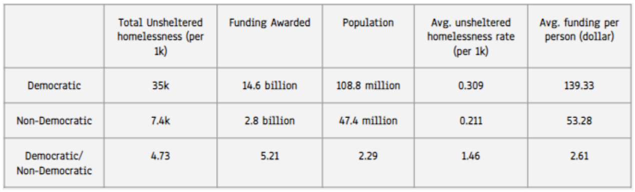
We have been discussing that policy has a significant impact on the homeless rate. We want to step further to explore how the government contributes to solving the homeless problem and reducing the unsheltered homeless rate. In this part of analysis, we introduced another two features called funding and unsheltered homeless. All these two features are normalized by county population.

3.3.1 CoC Funding VS Unsheltered Homeless

In the Funding Per Person vs Unsheltered Homelessness Rate scatter plot 9 labeled by democratic, swing and republican, most of the data points are distributed on the bottom left corner, which is showing a good pattern that government is trying to efficiently use the funding to reduce unsheltered the homeless rate and not overuse the money. In addition, we also notice that, for those counties in democratic states, the distribution is much more scattered and has significantly higher funding per person and unsheltered homelessness rate.

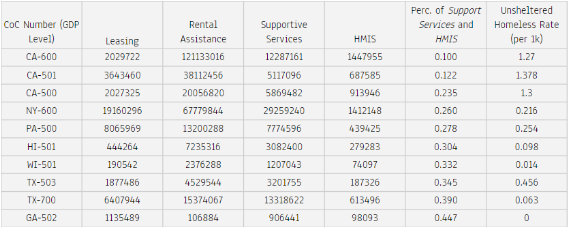
3.3.2 Statistics on Funding and Unsheltered Homeless

In figure 10, for these counties in democratic states, they have received more than 4 times of funding than those counties in non-democratic states. At the same time, counties in democratic states also have almost 4 times more unsheltered homeless people than counties in non-democratic states. As for the unsheltered homeless rate, counties in democratic states is 46 percent times higher than non-democratic.



	Total Unsheltered homelessness (per 1k)	Funding Awarded	Population	Avg. unsheltered homelessness rate (per 1k)	Avg. funding per person (dollar)
Democratic	35k	14.6 billion	108.8 million	0.309	139.33
Non-Democratic	7.4k	2.8 billion	47.4 million	0.211	53.28
Democratic/ Non-Democratic	4.73	5.21	2.29	1.46	2.61

Figure 10. Descriptive Statistics on CoC Funding and Unsheltered Homeless



CoC Number (GDP Level)	Leasing	Rental Assistance	Supportive Services	HMIS	Perc. of *Support Services* and *HMIS*	Unsheltered Homeless Rate (per 1k)
CA-600	2029722	121133016	12287161	1447955	0.100	1.27
CA-501	3643460	38112456	5117096	687585	0.122	1.378
CA-500	2027325	20056820	5869482	913946	0.235	1.3
NY-600	19160296	67779844	29259240	1412148	0.260	0.216
PA-500	8065969	13200288	7774596	439425	0.278	0.254
HI-501	444264	7235316	3082400	279283	0.304	0.098
WI-501	190542	2376288	1207043	74097	0.332	0.014
TX-503	1877486	4529544	3201755	187326	0.345	0.456
TX-700	6407944	15374067	13318622	613496	0.390	0.063
GA-502	1135489	106884	906441	98093	0.447	0

Figure 11. CoC Housing Budget Contribution

3.3.3 Higher GDP Counties Are Less Likely to Be Restricted by Policy

Based on our research, we notice that CoC funding is highly correlated to the economics of the county, with the correlation coefficient of 0.47. In figure12, we split original scatter plot9 into democratic plot and non-democratic plot. We labeled all the data as follow - 0 for GDP per capita lower than 44924, 1 for GDP per capita within the range (44924, 56554), 2 for GDP per capita higher than 56554.

In non-democratic plot, the data points are "regulated" in the bottom left, high GDP counties trying to escape the magnitude (having relative higher funding per person and unsheltered homeless rate) but still within the "regulation" area. On the other hand, the high GDP counties in Democratic plot is "out of controll", many data points have significantly high funding per person or unsheltered homeless rate or even both. In this case, the policy cannot effectively regulated with higher GDP counties.

3.3.4 Natural Disaster is Causing More Unsheltered Homeless Rate

As aforementioned, counties with lower GDP level tend to be regulated by policy. However, we still notice some outliers in Non-Democratic plot. Myrte Beach/Sumter City County is from North Carolina, New Orleans/Jefferson Parish County is from Louisiana, and Jacksonville-Duvel Clay Counties is from Florida. All these three counties and states are suffering from hurricanes over years.

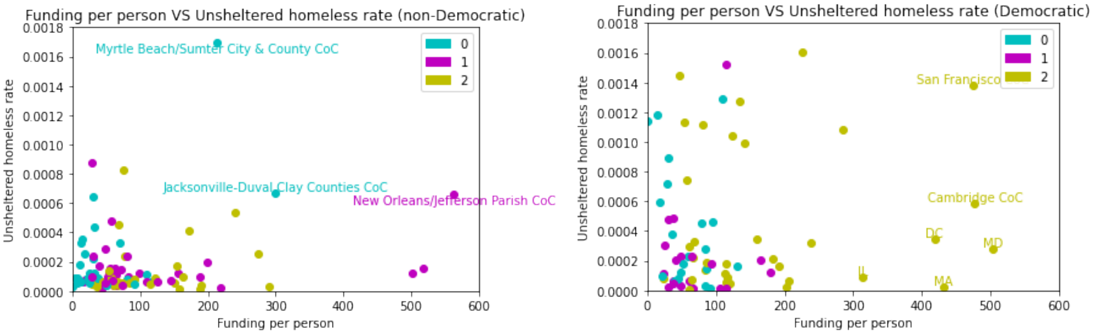


Figure 12. Dot plot of Supportive Funding against Homelessness Rate (Democratic and Non-Democratic)

3.3.5 High Living Expenses Needs to be Solved

In the democratic plot of Figure 9, these pointed out outliers are mostly in the counties with high living expenses, such as San Francisco, Cambridge, and Washington D.C. In addition, there is a large portion of funding budget used to pay for the homeless rent and leasing, especially in those counties with higher GDP level. According to figure **??**, the more percentage spent on Support Services and Homeless Management Information System (HMIS), the lower unsheltered homeless rate tends to be. High living expense is forcing the government to put more money help to pay the rent. In this scenario, if the government was still willing to pay a fair amount of money to these support services, the unsheltered homeless would be getting much better.

3.4. How does economy play a role?

To examine the relationship between economic factors and homelessness, we decided to use the GDP value as an indicator of the economic status. Since the areas CoCs serve are usually at the county level, we need to find the corresponding GDP sum of the areas each CoC is serving. In addition, we normalized the sum of GDP by dividing it by the total population of all serving areas of each county. In our study, we particularly focus on the GDP data between 2016 and 2019.

In figure 13 and 14, we show the total number of homelessness and sheltered homelessness for each year. It can be observed that the the total homelessness has decreased for a decade since 2007, but in recent years there is an increasing trend. For the total number of sheltered homelessness, it has maintained a decreasing trend. By comparing both trends, it can be concluded that the more homeless people remain unsheltered in recent years.

In figure 15, it can be observed that the more economically developed an area is, the higher homelessness ratio it will have. The possible reasons behind such phenomenon might be that richer community has better infrastructures, as these can better support the basic needs for homeless peo-

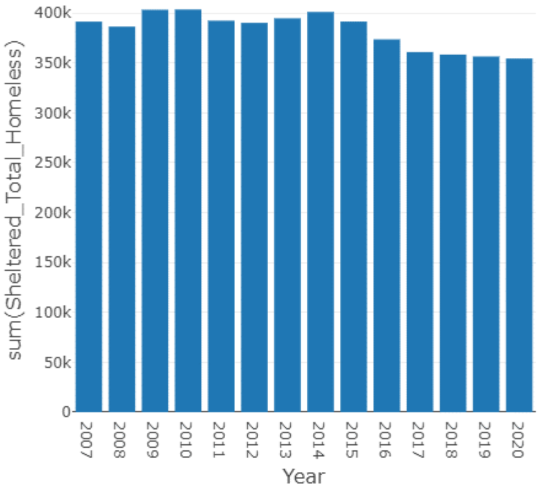


Figure 13. Total Number of Homelessness Overtime

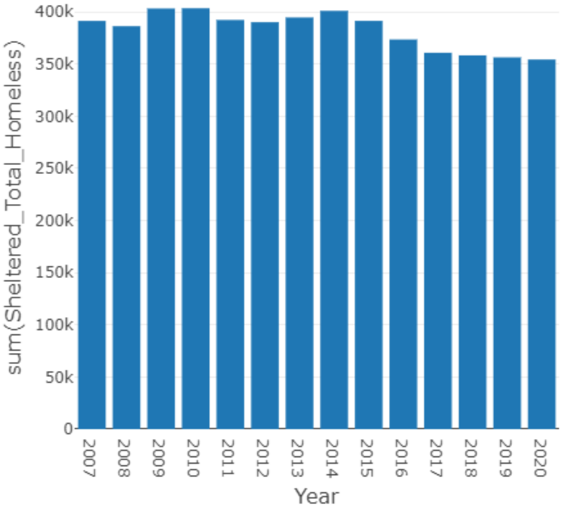
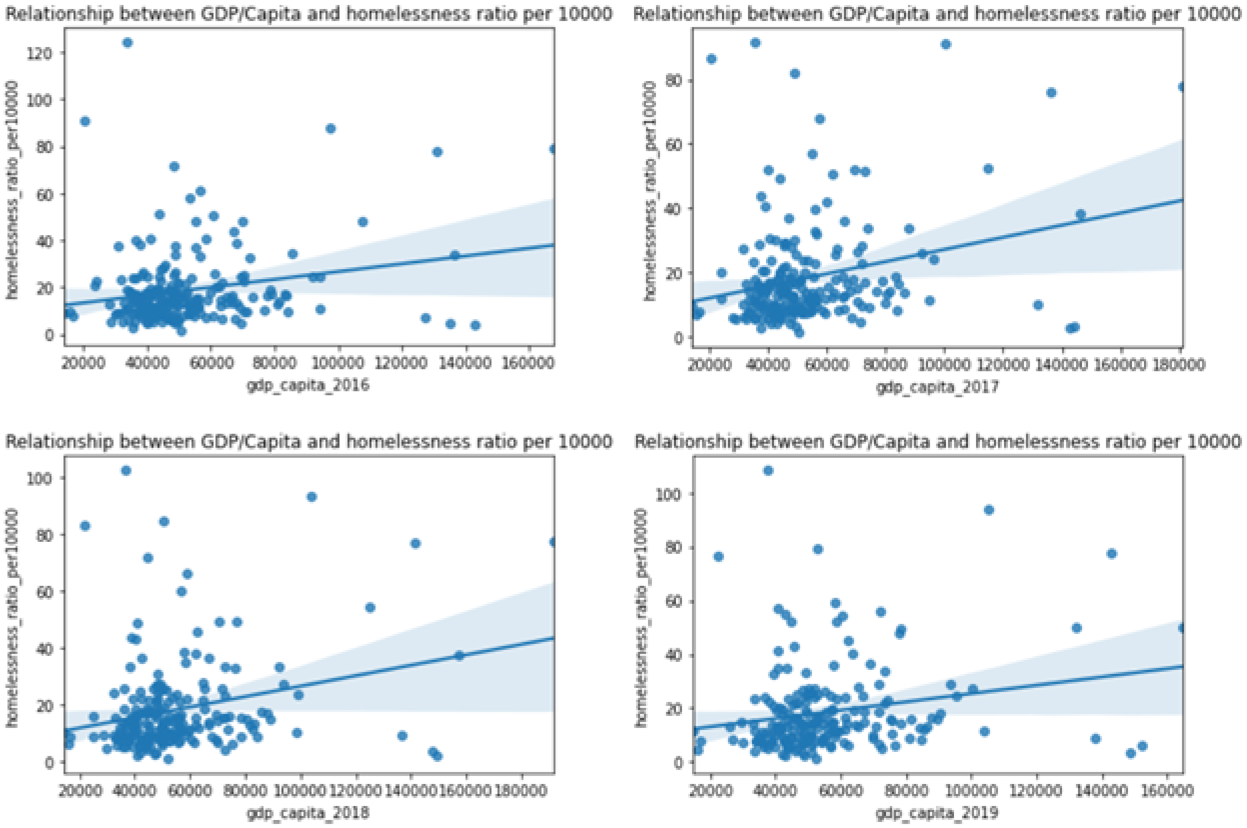


Figure 14. Total Number of Sheltered Homelessness Overtime



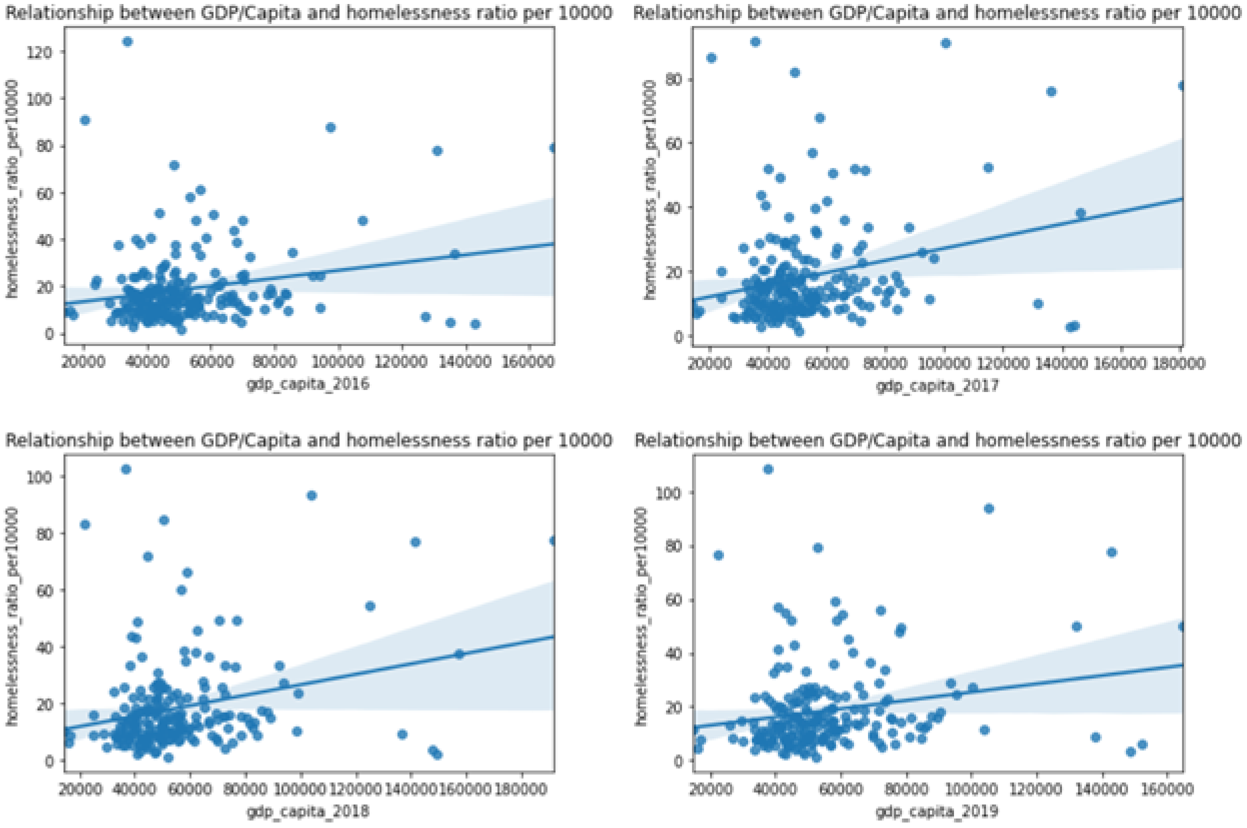
Relationship between GDP/Capita and homelessness ratio per 10000

Relationship between GDP/Capita and homelessness ratio per 10000

Relationship between GDP/Capita and homelessness ratio per 10000

Relationship between GDP/Capita and homelessness ratio per 10000

Figure 15. Relationship between gdp per capita and homelessness ratio per 10000 people from 2016 to 2019



Relationship between GDP/Capita and homelessness ratio per 10000

Relationship between GDP/Capita and homelessness ratio per 10000

Relationship between GDP/Capita and homelessness ratio per 10000

Relationship between GDP/Capita and homelessness ratio per 10000

Figure 16. Relationship between GDP Per Capita and Sheltered Homelessness Ratio Per 10,000 People from 2016 to 2019

ple. Another possible explanation is that areas with higher GDP are usually populous metropolitan areas, which gives people better accessibility to job opportunities and related training programs.

Further, we examine the relationship between economy status and the sheltered ratio per 10,000 people. From Figure 16, it can be observed that the more economically developed an area is, the higher sheltered homelessness ratio it will have. One of the possible reasons behind such finding is that richer communities have more funds to support CoCs. The other possible reason is the better rotation rate provides higher capabilities. A higher rotation rate might be induced by the richer job opportunities in metropolitan areas and better supportive housing programs.

3.5. Does policy matter?

After the correlation analysis on different factors, we now focus on one of the biggest questions rose by our sponsor, namely: Does policy affect homelessness? To investigate this issue, we performed a sharp-null test and regression discontinuity design and then plotted against a few confounders to visualize the effect.

3.5.1 Sharp-null test

We used the ”City Policy Conserativeness” attribute from the original dataset to perform a sharp-null test. The City Policy Conservativeness attribute quantifies the policies of each local government into a spectrum from the most liberal(-1) to the most conservative(1). The sharp-null test shows an effect of +5.37 which means a CoC will receive an effect of having 5.37 more homeless people on average. The one-sided p-value of the sharp-null test is less than 0.01 which is very significant.

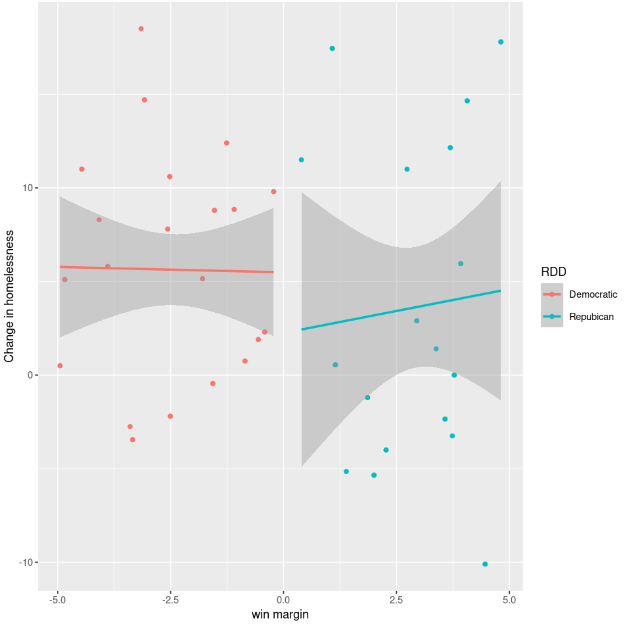


Figure 17. RDD plot of the effect of democratic(left) Vs. republican(right) policies on homelessness

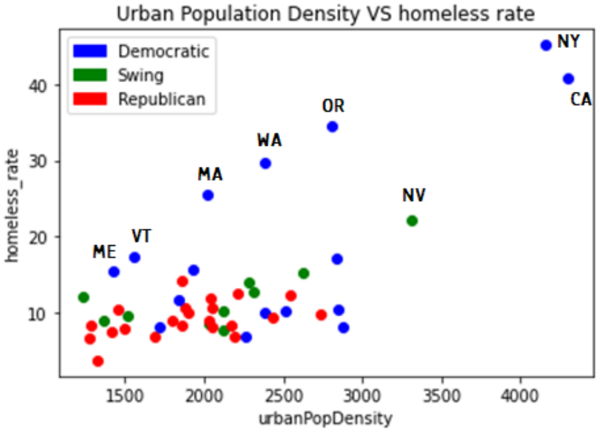


Figure 18. Urban population density (population per square mile Vs. Overall homeless rate

However, there is a very significant confounder ”urbanization” with conservativeness. Cities are overwhelmingly reinforcing liberal policy while as rural areas are mostly very conservative. In the meantime, most of the homeless people tend to gather in major cities for better facilities and supporting programs. Therefore, we need to find a way to remove all the other confounders and compare the effect of liberal versus conservative policies.

3.5.2 Regression Discontinuity Design

Regression Discontinuity Design (RDD) is one of the most widely used methods to study policies in the field of social and political studies. We wanted to use the Gubernatorial election data to perform RDD because it the governors make different policies on homelessness during their terms. And we want to measure the differences between the states where a Democratic candidate won by a small margin and those where a republican governor won by a small margin. The intuition is, theoretically, any small event can flip those elections so the actual election results must be randomized. If there is a big difference in homelessness during their terms, the reason is very likely because of the policy implemented by the elected governor.

We used the Gubernatorial election data [3] to plot the data. As we can see in Figure 17, There is a slight rise in the total homeless count for those states where the liberal governor won by a small margin. The difference is not very significant because the governor does not change policies easily. A policy can be enforced most effectively in trifecta states, where one party controls the majority in the state house, state senate, and the governor [6]. In the next step, we want to investigate the effect of being trifecta states while eliminating the confounder ”urbanization” at the same time.

3.5.3 Visualization on policy

To remove the confounder, we plotted the total homeless rate versus urbanization ratio and colored the dots based on the 2020 election results. As we can see in Figure 18, the blue dots marked are overwhelmingly democratic trifecta states. The trifecta states distinguishably jump out from the rest of the states and they showed a clear upward trend as urban population density goes up as well. The visualization demonstrates that liberal policies can significantly increase the overall homeless rate as they might be much more welcoming to homeless individuals. On the other hand, when we repeated the same tests on the sheltered ratio, there wasn’t a significant difference between policies and parties.

Conclusively, our analysis shows that liberal policies seem to attract and keep more homeless people while all states are providing a fair amount of shelter to the local homeless population.

4. Model Development

We aim to construct a model to predict Homelessness Per 10,000 for different CoCs. We choose Lasso, Neural Network, and Bayesian linear regression. Then we also construct two tree-based models, Random Forest and XgBoost. We found that tree-based models outperform other models since we have a limited amount of data and the features do not share much homogeneity.

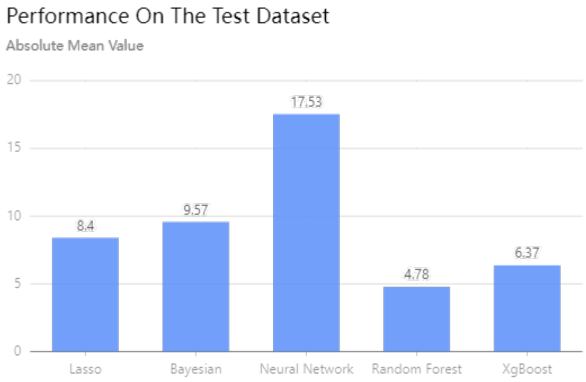


Figure 19. Performance of models on the test dataset with absolute mean error as standard (lower is better)

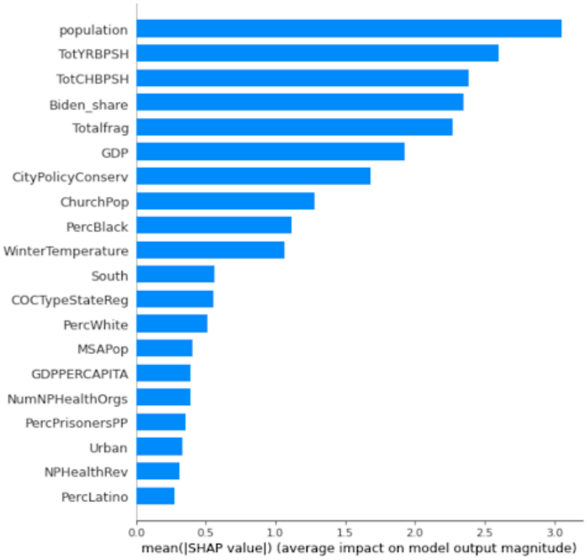


Figure 20. Mean SHAP value from XGBoost

5. Performance and Results

During training, we perform 5-fold and leave 10% data out as test data. We use absolute mean error as the cost function. Also, we fill None value with the corresponding state level mean value. By the figure 19, we find that the Neural Network is the worst in our case. Lasso and Bayesian perform similarly to each other, which outperform the Neural Network by about 50%. As we expect, Random Forest and XGBoost are the best 2 models that outperform the Neural Network by 73% and 64%. Tree-based models perform better under the lack of enough training data and homogeneity, they can still grow faster to fit data.

Based on the figure 20, we can easily find out which factors are more important. Population is the most important factor, since homelessness population is strongly correlated with the total population. Also, political factors play a major role, which confirms our previous analysis, and supporting programs, such as housing programs, are effective to lower the homelessness ratio.

6. Conclusion and Next Steps

In this project, we examined the homelessness issue from the perspective of policy, public transportation, economy, urbanization, temperature, and fundings for CoC programs. During the exploration process, we created several visualizations and conducted a couple of statistical experiments to further understand the relations. Lastly, we built a model that predicts the homelessness ratio per 10,000 people, and used “SHAP” [4] to investigate the extent of influence for each feature.

For our possible future steps regarding this project, we want to compose a blog with Databricks as it is a reflection of how this platform could be utilized to conduct social science research for social good. In addition, by no means is this study a comprehensive one that studies all possible factors that are related to homelessness. We intend to include more factors in our future study so that this project can be more inclusive. Lastly, based on our finding, we can propose possible solutions to, more or less, address this social issue.

References

[1] Decennial census of population and housing by decades. U.S. Census Bureau, 2021.

[2] John C Buckner. Understanding the impact of homelessness on children: Challenges and future research directions. *American Behavioral Scientist*, 51(6):721–736, 2008.

[3] Dave Leip. Dave leip's atlas of u.s. elections. 2021.

[4] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[5] National Oceanic and Atmospheric Administration National Center for Environmental Information. Climate at a glance. NOAA, 2021.

[6] Jennifer L. Pomeranz, Arjumand Siddiqi, Gabriella J. Bolanos, Jeremy A. Shor, and Rita Hamad. Consolidated state political party control and the enactment of obesity-related policies in the united states. *Preventive Medicine*, 105:397–403, 2017.

[7] Peter Somerville. Understanding homelessness. *Housing, theory and society*, 30(4):384–415, 2013.