STOR 320: Introduction to Data Science

Submit to gradescope by 11:59 PM, Dec 09, 11:59 PM.

Final Paper Group 7

Group members: Ivy Nangalia, Ximing Sun, Kharan Joshi, Alisha Nazir

Introduction

Homelessness is a pervasive issue affecting communities in the U.S., yet its patterns and causes remain complex and multifaceted. By analyzing trends in the homeless population and identifying factors contributing to homelessness, we can better understand this pressing social challenge and inform strategies to mitigate it. Our project focuses on two critical questions:

1. What trends exist in the homeless population, and how does geography affect these trends?
2. Are race, age, and gender good predictors of homelessness in New York City?

These questions aim to uncover hidden patterns and provide actionable insights for policymakers, social service providers, and others working to reduce homelessness.

The first question explores the geographical dynamics of homelessness. Understanding how homelessness varies by region, urban density, and local economic conditions is essential for designing targeted interventions. For instance, homelessness in urban areas may be driven by high housing costs, while in rural regions, it may be linked to a lack of access to resources. Exploring these geographical nuances reveals where the problem is most severe and helps identify the unique factors driving homelessness in different areas. This insight is invaluable for organizations allocating resources and developing region-specific solutions.

The second question examines the role of demographic factors—race, age, and gender—in predicting homelessness. These characteristics intersect with

systemic inequalities, shaping individuals' vulnerability to housing instability. We can identify high-risk groups and address underlying disparities by investigating these predictors. For example, if specific demographics are disproportionately affected, understanding the reasons behind this can inform equitable policy reforms and foster more inclusive support systems.

These questions are both intellectually intriguing and socially significant. Uncovering these insights could guide evidence-based decision-making and fostering partnerships with stakeholders to build a better world. For the country at large, the answers hold promise for addressing one of the most visible and urgent manifestations of poverty and inequality. By answering these questions, our project aims to contribute to the collective effort to end homelessness and improve the lives of vulnerable populations worldwide.

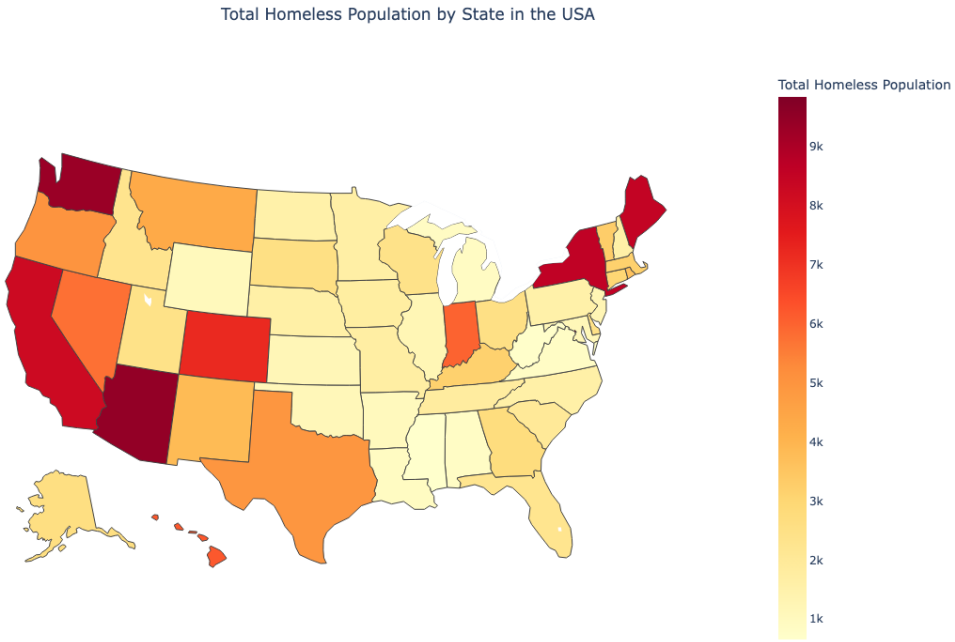
Data

The data we used to answer our research question, "What trends are there in the number of homeless people, and how do geographical factors affect these trends?" is from HUD, the U.S. Department of Housing and Urban Development. We used the Point-in-Time (PIT) Count and Housing Inventory Count (HIC) records, which show how many homeless people there are in the U.S. on one night in January every year. Surveys, observations by trained volunteers, and study in public databases are some of the ways the data is gathered. Local Continuums of Care (CoCs) keep an eye on all of it. Local governments, nonprofits, and housing providers are some of the groups that make sure the data is consistent and reliable so that it can be used to make comparisons over time and between areas. The structured and consistent data collection process from HUD allows us to confidently analyze trends across time and regions, making it easier to identify areas and populations most affected by homelessness.

The dataset has a lot of variables but we used a few main variables, like race, ethnicity, gender, and age. It also has information about where the people live, such as state-level data and whether the places are urban, suburban, or rural. It also tells the difference between homeless people who are housed and those who are not, which helps show differences in living conditions. This is a detailed picture of homelessness in the U.S. at a certain point in time, meaning each observation is a person or family that was counted during the PIT survey. By separating the data by location and population, we can see patterns. For example, the high number of homeless people on the West Coast is likely due to the high cost of living, and Black and Hispanic people are overrepresented in states like New York, Texas, and California. Our study of differences in

homelessness and how geography plays a part is directly linked to this knowledge. Clustering can happen in places with a lot of people or expensive homes, and it's common for people of certain races or ethnicities to be affected more than others. The collection also gives us a way to look into trends over time or more specific patterns in space. This makes it an important tool for learning about homelessness and what causes it.

This chloropleth illustrates the distribution of homeless individuals by racial categories across different U.S. states, as reported in the HUD Point-in-Time (PIT) Count. It highlights racial disparities, showing that states like California and New York have significantly higher homelessness counts, particularly among Black, White, and Hispanic populations.



The below table provides a detailed breakdown of the homeless population by race across various states, categorizing groups such as White, Black or African American, Asian, and others. This data shows the overrepresentation of certain racial groups, such as Black or African American individuals, in states like California, Texas, and Florida, offering a more granular view of demographic information.

The second research question we explored was if race, age, and gender are good predictors of homelessness in New York, using U.S. Census Bureau data, based on their every-ten-year surveys of New York City. This information is gathered by the Census Bureau through a mix of administrative records, home surveys, and door-to-door techniques to make sure that all population groups are fully

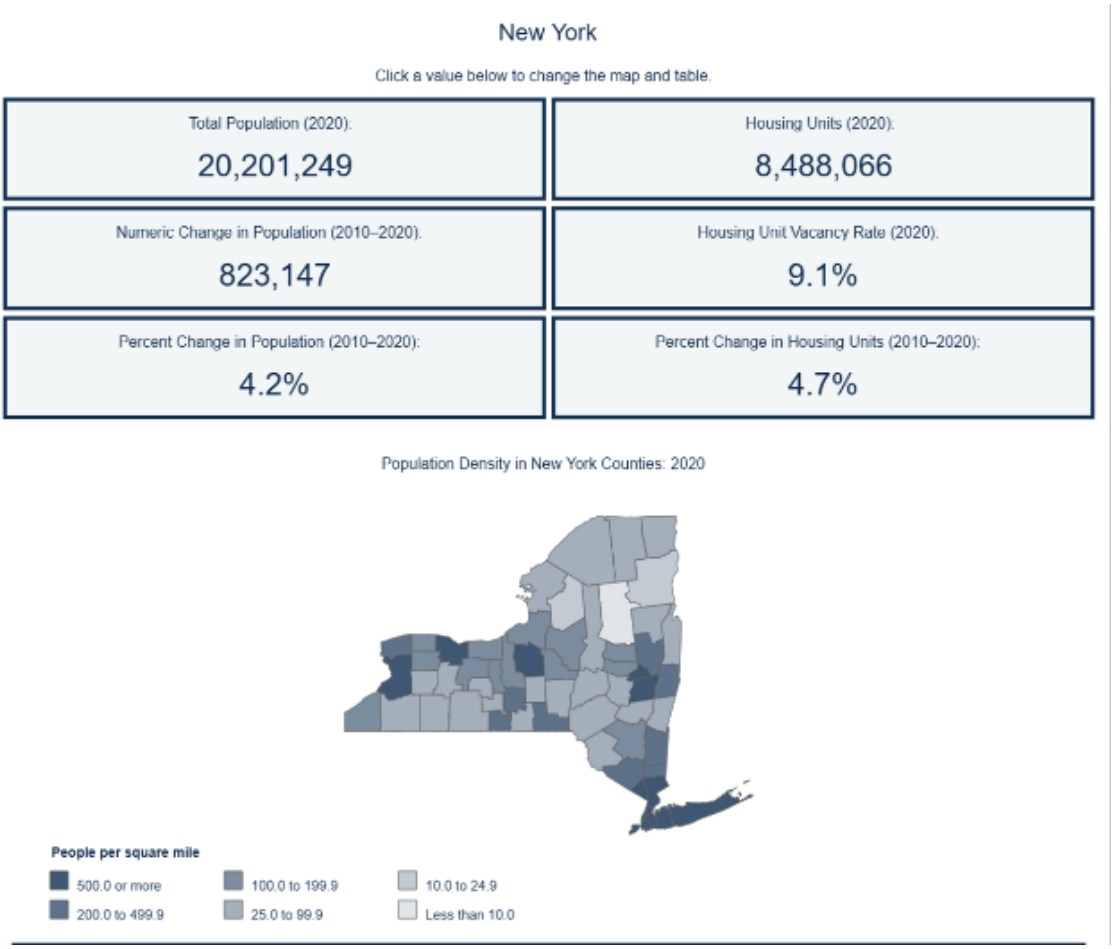
covered. Answers come directly from households through mailed surveys, online submissions, or interviews with trained census workers that happen in person. The process is carefully planned to reduce mistakes and increase precision, which makes the dataset a trustworthy source for figuring out housing and population trends. This information shows changes in the number of people living in cities versus those living in rural areas, the number of homes available, and the occupancy rates. These findings add to the HUD data by showing how changes in population and housing affect the trends in homelessness in New York.

State	Overall Homeless - White	Overall Homeless - Black, African American, or African	Overall Homeless - Asian or Asian American	Overall Homeless - American Indian, Alaska Native, or Indigenous	Overall Homeless - Native Hawaiian or Other Pacific Islander	Overall Homeless - Multiple Races
AK	796	181	38	1,151	99	349
AL	1,349	1,787	12	37	11	108
AR	1,696	733	7	50	30	93
AS						
AZ	9,011	3,097	101	975	106	947
CA	96,385	53,369	7,012	8,589	3,413	12,631
CO	9,655	2,431	135	829	402	987
CT	1,653	1,117	14	38	7	186
DC	587	4,091	40	73	52	79
DE	385	773	4	3	2	78
FL	16,579	12,495	186	330	117	1,049
GA	4,603	7,165	43	58	20	405
GU	9	5	33	1	1,005	22
HI	1,295	246	603	50	2,336	1,693
IA	1,687	664	28	86	18	170
ID	1,902	50	8	160	21	157
IL	5,535	5,583	197	143	62	427
IN	3,684	1,962	27	39	36	289
KS	1,795	540	12	78	14	197
KY	3,386	1,131	16	36	11	186
LA	1,179	1,885	15	28	13	49
MA	9,223	8,659	181	126	134	818
MD	1,945	3,521	64	96	29	210
ME	2,044	2,013	17	34	5	145
MI	4,278	3,981	51	101	20	566
MN	2,999	3,152	193	961	45	1,043
MO	3,811	2,399	40	90	67	301
MP						
MS	429	505	7	8	5	28
MT	1,480	63	10	461	18	146

Age, gender, race, and housing-related factors like total housing units and occupancy rates are some of the most important variables used from this collection that we analyzed. The dataset also has classifications for urban and rural areas, which lets us look at how geography affects the stability of homes and the number of people living in an area. Each report in the dataset is a statistical summary for a certain New York county or city's population groups or housing units. This information gives us a general picture of the state's people and housing situations, which lets us look for links between these things and trends of homelessness. This information is very helpful for figuring out how differences and geographic factors affect homelessness in New York, which has a lot of people. For example, housing problems are more likely to happen in places

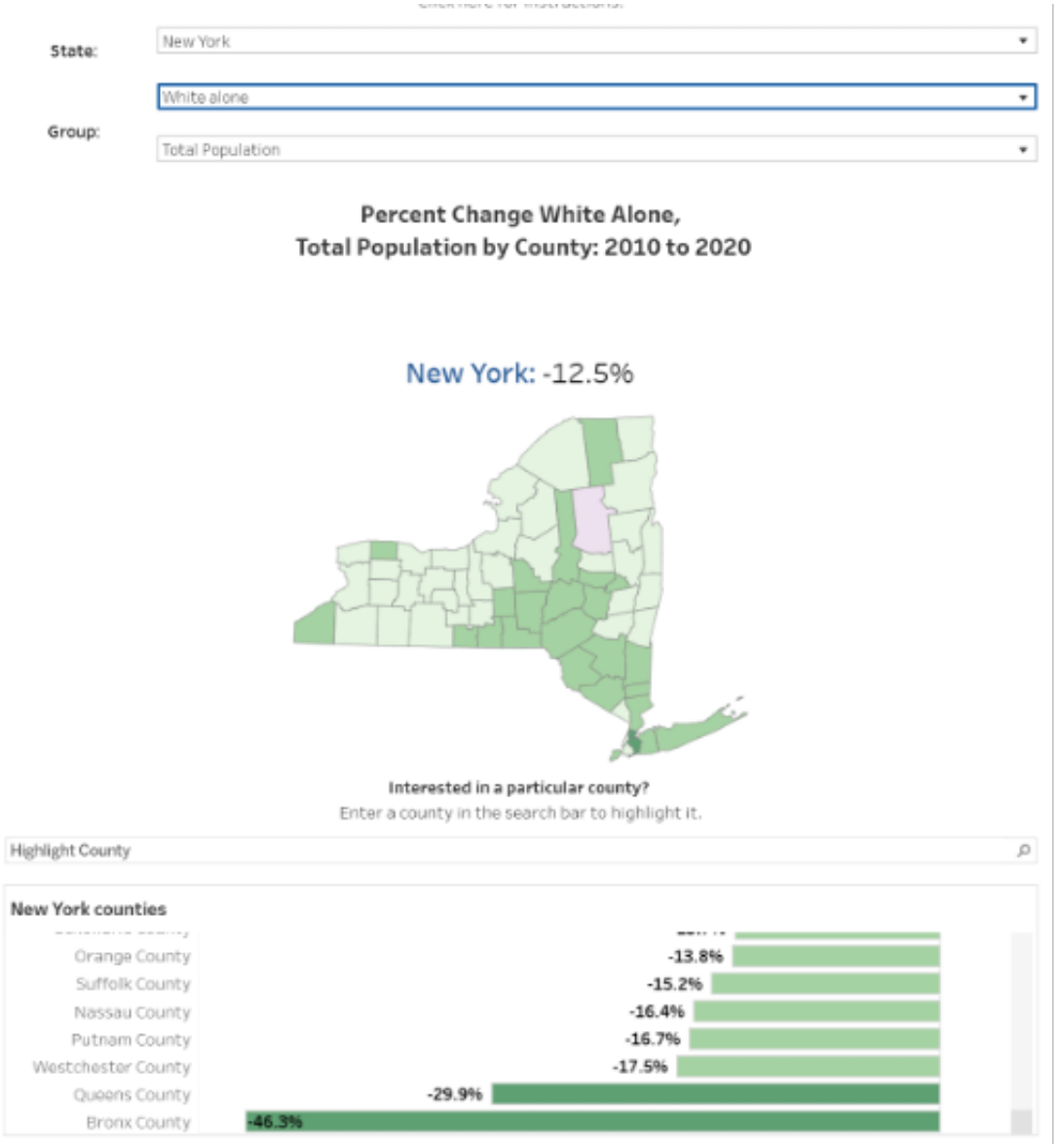
with high population densities. Including factors for race and gender also helps find differences in population, and dividing areas into urban and rural areas shows how trends change over time. The HUD data and this dataset together make it possible to look at both the direct effects and root causes of homelessness in New York.

This map visualizes population density across New York's counties, categorizing areas by the number of people per square mile. Urban centers like New York City and its surrounding counties are shown to have significantly higher densities compared to rural upstate areas. This data can be linked to homelessness analysis as higher population densities often correlate with a lack of housing affordability.



The second map illustrates the percentage change in the White population across New York counties over the past decade. This data includes an option to explore population changes for different racial groups, as highlighted in the screenshot below, which shows racial and ethnic categories like Black or African American, Asian, and Native Hawaiian. Observing these changes reveals patterns of demographic shifts, such as increases in some counties like Bronx County and decreases in others like Queens County. These trends may affect homelessness

by altering the dynamics and resource distribution in different areas. By putting these insights together with HUD data, we can look into how changes in population and more people living in cities affect trends in homelessness. For example, counties with fewer people and fewer housing needs may not have as much pressure on housing, while areas with more people may have trouble finding housing and making it affordable, which can hurt some race and ethnic groups more than others.



This table shows changes in race and ethnicity among the homeless people who were living in shelters over two time periods. It is related to our research on racial differences in homelessness. For example, the rise in the number of Hispanic people (+4.5%) and the falls in the numbers of Black (-2.1%) and White (-3.3%) people show changes in how different racial groups experience sheltered homelessness. This trend is in line with our analysis of the geographic and systemic factors that cause these differences. This lets us look into how

structural inequalities and regional trends might affect racial and ethnic groups in different ways over time. This is another piece of the data. We saw the same tables and changes for things like age.

Table 3: Demographic Characteristics of the Sheltered Population Experiencing Homelessness Over Time

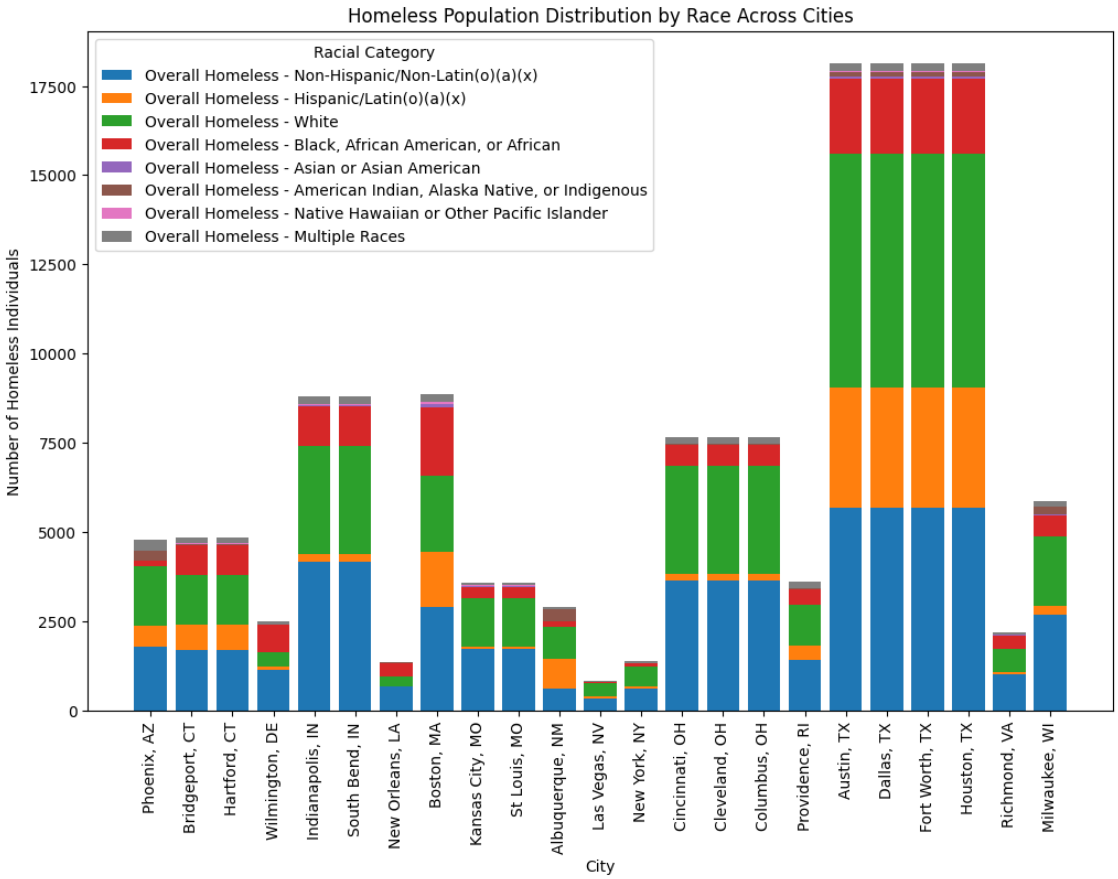
	2013-2017		2018-2022		Difference	
	Est.	S.E.	Est.	S.E.	Est.	S.E.
Under age 18	10.5	0.3	8.2	0.3	*-2.3	0.5
Ages 18 to 64	84.3	0.4	83.5	0.5	-0.8	0.6
Ages 65 and over	5.2	0.2	8.3	0.3	*3.1	0.4
Male	63.6	0.5	59.8	0.7	*-3.8	0.8
Female	36.4	0.5	40.2	0.7	*3.8	0.8
White, non-Hispanic	35.7	0.5	32.4	0.4	*-3.3	0.7
Black, non-Hispanic	38.9	0.5	36.8	0.5	*-2.1	0.7
Asian, non-Hispanic	1.2	0.1	1.3	0.1	0.1	0.2
Hispanic	18.1	0.4	22.6	0.4	*4.5	0.6
Native-born	90.1	0.3	88.2	0.4	*-2.0	0.5
Naturalized citizen	2.6	0.1	2.8	0.2	0.3	0.2
Not a citizen	7.3	0.2	9.0	0.3	*1.7	0.4
With a disability	35.7	0.5	34.6	0.5	-1.1	0.7
With no disability	64.3	0.5	65.4	0.5	1.1	0.7

*Estimate is different from zero at the 90 percent confidence level.
Note: Some percentages do not add up to 100 due to rounding.
Source: U.S. Census Bureau, 2013-2017 and 2018-2022 American Community Survey, 5-year estimates.

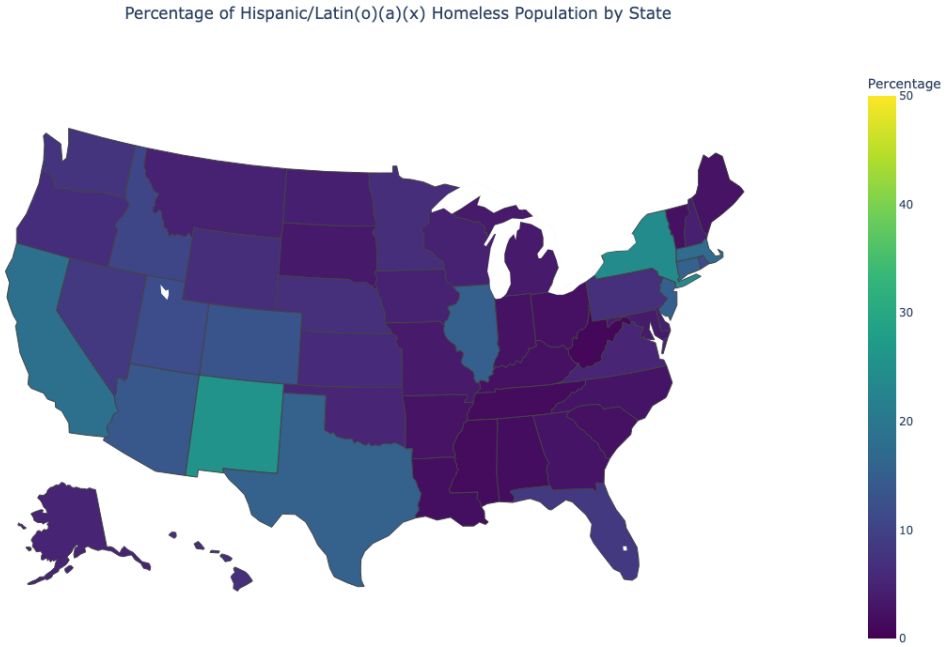
Results

Research Question 1: What trends exist in the homeless population, and how does geography affect these trends?

To answer this research question, we utilized the PIT datasets, picked the year 2023 & sliced the dataframe into the variables of interest: age, race, and location. We started our analysis by taking a random sample (n = 40) of the 385 CoC's and created a new dataframe with the sampled data. This data table provides an overview of homelessness statistics from various CoCs in the United States for 2023, focusing on demographics such as age, race, and gender, along with total homeless counts. The results highlight substantial variation in homelessness across CoCs. For example, Chicago, a main city CoC, reports an overall homeless count of 6,139, while smaller CoCs such as Dearborn Heights in Michigan report only 199 – likely due to differences in population and cost of living. There also appears to be a large variability of Hispanic homeless individuals between states.



We used plotly to create a choropleth of the percentage of hispanic homeless people. There tends to be a higher percentage of Hispanic homeless populations around the mexican-american border, likely because the Hispanic population is higher there. New Mexico has a high hispanic-american population due to its history of spanish colonization & mixing with indigenous populations, so the high hispanic homeless population makes sense. New York has about a 19% hispanic population due to a large wave of puerto-rican immigration during the 20th century, when Puerto Ricans were given American citizenship. However, since the homeless population is closer to 30%, there are likely some systemic factors that disproportionately affect hispanic and latino populations and their housing security.



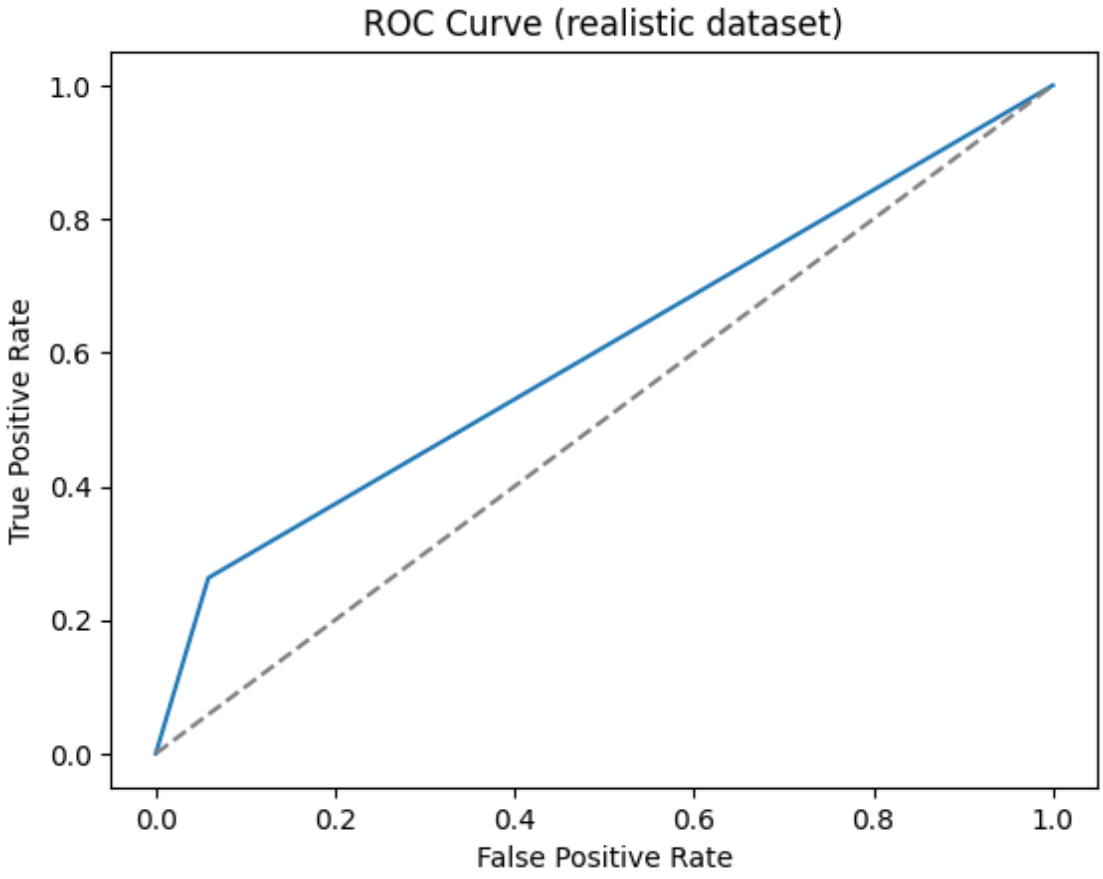
Research Question 2: Are race, age, and gender good predictors of homelessness?

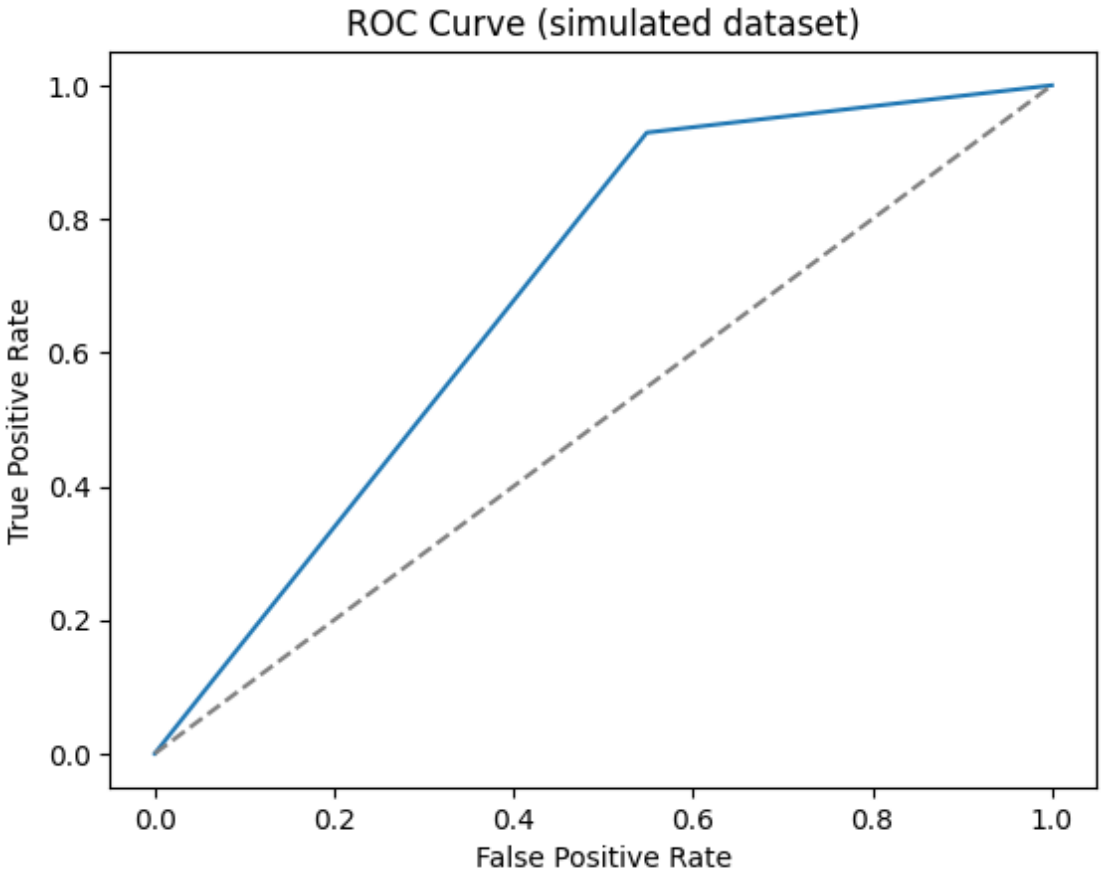
To answer the question of whether race, age, and gender are good predictors of homelessness, two datasets were constructed: a “simulated” NYC homelessness dataset and a “realistic” NYC homelessness dataset derived from census and homelessness records. Both datasets were used to evaluate the predictive power of these demographic factors. By employing complementary statistical and machine learning techniques, we aimed to uncover patterns and validate whether these variables hold consistent predictive power across both datasets. The realistic simulation dataset contained 88,025 rows for homeless people and 8,800,000 rows for non-homeless individuals, in line with the estimated NYC population by the Census and the estimated NYC homeless population contained in the PIT-CoC dataset. The dataset contained simulated data drawn from the demographics provided from both the Census and the PIT-CoC datasets in order to depict the inequalities in homelessness accurately. The “realistic” NYC dataset was inherently imbalanced, reflecting the true distribution of homelessness in the population.

To address this, we applied resampling techniques (undersampling the majority class) to ensure the models received equal representation of homeless and non-homeless groups. Using the same demographics from the other dataset, we created another “simulated” dataset, containing a balanced subset of 88,025 rows for homeless people and 88,025 for non-homeless individuals. This ensured

that our analysis would not be biased by differences in the sample sizes. In both datasets, categorical variables like race and gender were one-hot encoded, while the continuous variable age was normalized for consistency across the models.

For the “realistic” dataset, based on the Logistic Regression, we can see that race was the most significant predictor, with minority groups (Hispanic or Latino) having higher odds of homelessness. Gender had a moderate effect, with non-binary and transgender people being more likely to experience homelessness. Age demonstrated little effect, with only older (>60) individuals being more likely to experience homelessness. The model only achieved an accuracy of 4% and an ROC-AUC of 0.803. For the simulated dataset, the model replicated similar trends, with race and gender emerging as statistically significant predictors. Age effects were slightly weaker but remained consistent. However, the accuracy was 63%, which was much better than the previous one, with an ROC-AUC of 0.803.





In summary, both datasets revealed consistent trends, underscoring the importance of race and gender in predicting homelessness. Age effects, while present, were less pronounced compared to the other two variables. Based on the employing of logistic regression, the model with real dataset showed worse predictive ability and the model with simulated dataset showed moderate predictive ability. This final result suggested that while race, age, and gender are meaningful predictors of homelessness, additional factors might be necessary to research in order to improve the model's predicting accuracy.

Conclusion

Our investigation of the trend of homelessness and the predictive power of demographic factors has brought to light the complex interaction of geographic, economic, and systemic factors that shape homelessness in the U.S. The results of the first research question indicate a clear geographic clustering, with highly prevalent homelessness along the West Coast and border states. These regional trends are thoroughly intertwined with housing affordability, urban density, and systemic inequity. The disproportionate representation of Black/African American populations in states such as New York and Texas, along with Hispanic populations in states including New Mexico and California, underlines how

geography intersects with historical and structural marginalization. Such disparities call for region-specific, equity-driven solutions that take into consideration the local economic and demographic landscape.

The second research question revealed that race and gender are significant predictors of homelessness, whereas age has a less sharp role. Throughout both the "realistic" and "simulated" NYC datasets we selected, minority racial groups and people who are non-binary or transgender were more likely to be experiencing homelessness, reflecting the broader societal inequities faced by these groups. With only moderate predictive abilities found from the models for the simulated dataset, the realistic data accuracy being lower underlines challenges when trying to model such a complex issue with demographic factors. This could suggest other needed variables for enhancing predictive accuracy: economic status, mental health, and access to social services are other factors that may come in handy.

Taken together, these findings indicate the nature of homelessness as a social problem, influenced by geographic, economic, and demographic factors. Homelessness is not an issue that can be addressed exclusively through broad policies because it requires region-specific approaches that consider racial and gender-based systemic inequities. Additionally, the predictive models can be improved with more data to allow the fine-tuning of interventions and increasing resource allocation among those most vulnerable. We recommend that governments use the raw data collected in order to train and test these models, and that they use the models to inform policymaking and interventions. These will result in more effective and much fairer solutions while providing a better understanding of homelessness in the United States in general.