ISSN: 3066-9847
Review Article
Science Set Journal of Economics Research

Unraveling the Complexity: The Nexus Between Homelessness and Housing Prices in the San Francisco Bay Area and Throughout State of California (A Comprehensive Research Study)

Nathan Torento[1], Antony Scott[1], Samruddhi Mistry[1], and Bahman Zohuri[1,2*]

[1]*Graduate Students, Golden Gate University, Ageno School of Business, San Francisco, California, USA 94105*
[2]*Adjunct Professor and Project Advisor Professor, Golden Gate University, Ageno School of Business, San Francisco, California, USA 94105.*

***Corresponding author:** Bahman Zohuri, Adjunct Professor and Project Advisor Professor, Golden Gate University, Ageno School of Business, San Francisco, California, USA 94105

Submitted: 05 January 2024 **Accepted:** 09 January 2024 **Published:** 12 January 2024

https://doi.org/10.63620/MKSSJER.2024.1025

Citation: *Torento., N, Antony., Scott, Mistry., S, Zohuri, B. (2024) Unraveling the Complexity: The Nexus Between Homelessness and Housing Prices in the San Francisco Bay Area and Throughout State of California. Sci Set J of Economics Res 3(1), 01-11.*

Abstract

This research investigates the intricate relationship between homelessness and housing prices, unraveling the complexities prevalent in both the localized context of the San Francisco Bay Area and the broader scope of the entire state of California. Employing a multi-faceted approach, the study combines traditional research methodologies with advanced Artificial Intelligence (AI) and Machine Learning (ML) tools. Through comprehensive data analysis, the research explores the dynamics of homelessness and housing prices, identifying patterns and trends at both the regional and statewide levels. The study incorporates predictive modeling to discern future trends, spatial analytics to understand geographic variations, and natural language processing to gauge public sentiment.

By evaluating the impact of policies on homelessness rates, the research seeks to provide actionable insights for policymakers. Ethical considerations guide the implementation of AI, ensuring the responsible handling of sensitive data. Stakeholder collaboration and community engagement are integral components of this research, fostering a nuanced understanding of the challenges faced by diverse communities across the state. The research concludes with evidence-based policy recommendations, aiming to inform interventions that address the interconnected issues of homelessness and housing affordability in the San Francisco Bay Area and throughout the state of California.

Keywords: Homelessness, Housing Prices, San Francisco Bay Area, California, Affordable Housing, Socioeconomic Diversity, Artificial Intelligence, Machine Learning, Gentrification, Policy Interventions.

Introduction

This research delves into the intricate relationship between homelessness and housing prices in the state of California, employing cutting-edge Artificial Intelligence (AI) and Machine Learning (ML) tools for an exhaustive analysis. The study aggregates diverse datasets from regions across California, utilizing predictive modeling to discern statewide trends and variations. Through advanced spatial analytics and natural language processing techniques, the research aims to uncover nuanced patterns in homelessness clusters, migration dynamics, and public sentiments across communities. Evaluating the impact of statewide policies on homelessness rates through causal inference models, the study identifies vulnerabilities and populations at risk on a large scale.

Ethical considerations guide the implementation of AI, ensuring privacy and transparency. Stakeholder collaboration and community engagement play a pivotal role, contributing to a holistic understanding of challenges faced by different communities. The research concludes by translating findings into evidence-based policy recommendations at the state level, offering a roadmap for inclusive and equitable interventions to address the complex nexus of homelessness and housing affordability across California's diverse landscapes.

In the heart of the technological revolution and cultural diversity, the San Francisco Bay Area stands as a testament to the complexities that arise when rapid urban development collides with socioeconomic diversity. One of the most pressing issues gripping the region is the intricate relationship between homelessness and soaring housing prices. Public perception in San Francisco paints a vivid picture of a worsening homelessness crisis, exacerbating the challenges posed by an ever-expanding housing market.

The relationship between homelessness and housing prices in urban areas is a complex and multifaceted issue, particularly in rapidly changing and socioeconomically diverse regions like the San Francisco Bay Area. Public perception in San Francisco suggests a worsening homelessness crisis alongside soaring housing prices. On the contrary, there are views positing that rising homelessness negatively impacts housing prices. These conflicting opinions emphasize the need for a data-driven approach to understanding the true nature of this relationship.

The Overall Homelessness Crisis

The homelessness crisis in San Francisco is undeniable, with visible tent encampments dotting the cityscape and individuals seeking shelter in public spaces. Despite concerted efforts by local governments and nonprofits, the numbers continue to rise, sparking concerns about the adequacy of existing strategies.

One critical factor contributing to the homelessness crisis is the scarcity of affordable housing. As housing prices surge, the most vulnerable members of the community find themselves marginalized, unable to secure stable living arrangements. The mismatch between income levels and housing costs has left many individuals and families on the brink of homelessness, highlighting the urgency of addressing the affordability gap.

Soaring Housing Prices

The San Francisco Bay Area, often hailed as an economic powerhouse, has experienced a relentless surge in housing prices. The demand for housing, fueled by the tech industry's boom, has led to a shortage of affordable homes, pushing prices to unprecedented levels. The region's desirability has attracted a flood of high-income earners, driving up the cost of living and making it increasingly difficult for lower-income residents to find suitable accommodation.

The Gentrification Conundrum

Gentrification, a byproduct of rising housing prices, plays a pivotal role in reshaping urban landscapes. As affluent individuals move into once-affordable neighborhoods, property values rise, and long-time residents find themselves economically displaced. Gentrification not only contributes to the homelessness crisis but also underscores the need for comprehensive urban development policies that prioritize inclusivity and affordability.

Policy Challenges

Addressing the homelessness-housing price conundrum requires a multifaceted approach. Policymakers must grapple with the challenge of balancing economic growth with social equity. Strategies to increase affordable housing options, protect vulnerable populations, and mitigate the impacts of gentrification are essential.

Innovative Solutions

To combat homelessness and mitigate the effects of soaring housing prices, the San Francisco Bay Area must embrace innovative solutions. This includes the development of affordable housing projects, the expansion of supportive services, and the implementation of policies that protect vulnerable populations from economic displacement.

In conclusion, the intricate relationship between homelessness and housing prices in the San Francisco Bay Area highlights the urgent need for comprehensive and innovative solutions. As the region grapples with the challenges of rapid urban development and socioeconomic diversity, it is imperative that stakeholders collaborate to create a more equitable and sustainable future. By addressing the root causes of homelessness and implementing thoughtful policies, the San Francisco Bay Area can strive towards a more inclusive and resilient community.

However, bear in mind that, research on homelessness and housing prices in California State particularly, in the San Francisco Bay Area demands innovative approaches to capture the complexity of the issue. Leveraging Artificial Intelligence (AI) and Machine Learning (ML) tools can enhance the depth and accuracy of one's research findings. As such, we utilized AI and ML tools as presented in Section 7.0 of this paper.

Furthermore, here is a guide on how to incorporate these technologies into your research:

1. **Define Research Objectives:** Clearly outline your research objectives. Identify specific questions or challenges within the realm of homelessness and housing prices that AI and ML can help address. For example, you might want to predict future homelessness trends based on housing market fluctuations or analyze the impact of specific policies on homelessness rates.
2. **Data Collection:** Acquire relevant datasets that encompass a wide range of variables, including housing prices, demographic information, social services availability, and historical homelessness trends. Publicly available datasets, government records, and nonprofit databases can be valuable sources. Ensure the data is cleaned and standardized for effective analysis.
3. **Implement Predictive Analytics:** Utilize machine learning algorithms for predictive analytics. Regression models can help identify patterns and relationships between housing prices and homelessness rates over time. Predictive models can forecast future trends, allowing for proactive policy measures.
4. **Spatial Analysis:** Leverage AI to perform spatial analysis. GIS (Geographic Information System) tools, combined with machine learning algorithms, can help identify geographic patterns in homelessness distribution and correlate them with housing price fluctuations. This can provide insights into localized factors contributing to homelessness.
5. **Natural Language Processing (NLP):** Implement NLP techniques to analyze qualitative data. Extract valuable information from news articles, public forums, or social media related to homelessness and housing prices. This can provide a real-time understanding of public sentiment and identify emerging issues.
6. **Identify Vulnerable Populations:** Use clustering algorithms to identify vulnerable populations at a higher risk of homelessness due to housing market dynamics. By analyzing demographic data, socioeconomic indicators, and housing trends, AI can help target interventions more effectively.
7. **Evaluate Policy Impact:** Assess the impact of existing policies or proposed interventions using causal inference models. Machine learning algorithms can help establish

causal relationships between policy changes and changes in homelessness rates, aiding policymakers in making informed decisions.

8. **Ethical Considerations:** Pay careful attention to ethical considerations, especially when dealing with sensitive data related to homelessness. Ensure data privacy and security and be transparent about the limitations and biases inherent in the AI and ML models.
9. **Collaborate with Stakeholders:** Collaborate with local authorities, nonprofits, and community organizations to validate findings and ensure the research aligns with real-world challenges. Engaging stakeholders can provide valuable context and help translate research findings into actionable policies.
10. **Communicate Results Effectively:** Clearly communicate your findings to both academic and non-academic audiences. Visualization tools, such as charts and maps generated through AI-driven analytics, can enhance the accessibility of your research.

By integrating AI and ML tools into your research on homelessness and housing prices in the San Francisco Bay Area, you can uncover nuanced insights and contribute to the development of more effective and targeted solutions to address this pressing societal issue.

As we extend our focus beyond the iconic San Francisco Bay Area, the intricate relationship between homelessness and housing prices becomes an even more pressing concern for the entire state of California. With diverse urban landscapes, economic disparities, and a housing market that resonates with the challenges of the tech-driven economy, leveraging Artificial Intelligence (AI) and Machine Learning (ML) tools becomes paramount for a comprehensive understanding of this complex issue.

1. **Statewide Data Aggregation:** Begin by aggregating comprehensive datasets from various regions across California. This should include housing price indices, demographic information, employment data, and existing homelessness statistics. AI-powered data aggregation tools can streamline this process, ensuring that the research captures the nuances of each distinct locality.
2. **Predictive Modeling for Statewide Trends:** Implement predictive modeling to discern statewide trends in homelessness and housing prices. Machine learning algorithms can analyze historical data to predict future patterns, offering invaluable insights for policymakers in planning region-specific interventions.
3. **Comparative Analysis Between Regions:** Leverage AI and ML tools to conduct a comparative analysis of different regions within California. Explore variations in housing market dynamics, demographic profiles, and socioeconomic factors to identify unique challenges and opportunities in addressing homelessness.
4. **Advanced Spatial Analytics:** Utilize advanced GIS tools and machine learning algorithms to conduct spatial analysis on a larger scale. This could involve mapping out clusters of homelessness, understanding migration patterns, and identifying correlations between housing price trends and vulnerable population distribution across the state.
5. **Sentiment Analysis Across Communities:** Extend natural language processing techniques to analyze sentiments across diverse communities in California. Extract insights from social media, community forums, and local news sources to understand public perceptions and concerns related to homelessness and housing prices.
6. **Policy Impact Evaluation at the State Level:** Employ AI-driven causal inference models to assess the impact of statewide policies on homelessness rates. Evaluate the effectiveness of initiatives such as housing affordability programs and rent control policies, providing a comprehensive overview of their success and potential areas for improvement.
7. **Identification of Statewide Vulnerabilities:** Apply clustering algorithms to identify statewide vulnerabilities and populations at higher risk of homelessness due to systemic factors. This can aid in the development of targeted interventions and support systems on a larger scale.
8. **Ethical AI Implementation:** Given the diversity and scale of California, prioritize ethical considerations in data handling, ensuring privacy and security. Transparently communicate the ethical practices employed in the AI and ML models to build trust among stakeholders.
9. **Stakeholder Collaboration and Community Engagement:** Collaborate with state-level agencies, nonprofits, and community organizations to ensure the research aligns with the broader Californian context. Engaging with stakeholders from diverse regions ensures a holistic understanding of the challenges faced by different communities.
10. **Policy Recommendations for Statewide Impact:** Translate research findings into actionable policy recommendations at the state level. AI-generated visualizations and analytical insights can contribute to evidence-based policymaking, helping shape interventions that address the interconnected issues of homelessness and housing prices statewide.

By extending our research to encompass the entire state of California, we can leverage AI and ML tools to unveil patterns, correlations, and opportunities for impactful interventions. The integration of technology into this research not only enhances its depth but also lays the foundation for a more inclusive and equitable approach to addressing homelessness and housing affordability challenges across the diverse landscapes of California.

Our Research Objective

The main objective of this research is to understand the connection between homelessness levels and housing prices across the Bay Area and extend it throughout the State of California. To achieve this, two primary data sources were leveraged. 1) Government datasets tracking annual homelessness counts by various demographics in California from 2017 to 2022 (California Interagency Council on Homelessness, 2023), and 2) a dataset from the California Association of Realtors detailing median prices of existing single-family homes from 1990 to 2023 (California Association of Realtors, 2023). [1-3]

Step one was preparing the data for analysis. We began by aligning the datasets to the overlapping time frame of 2017 to 2022 and then integrating them based on the year and county, with a particular focus on "Continuum of Care" groupings, which we will elaborate on later. Our analysis was conducted at three levels: San Francisco, the broader Bay Area, and the entire state of California.

The Python programming language was used in each step of this study, starting with the initial exploratory data analysis to identify trends and patterns, followed by more targeted regression analysis to pinpoint demographic factors most strongly associated with homelessness and median housing prices. In the following section, we delve deeper into the methodology and findings of our study.

Data Collection and Cleaning

The following steps were taken for consideration of this class project, assigned to us by our instructor and advising professor.

Homelessness Dataset

In our study, we sourced our homelessness data from Data.gov, a platform managed by the U.S. government for public data dissemination. The dataset, titled "People Receiving Homeless Response Services by Age, Race, Ethnicity, and Gender," comprises four subsets corresponding to these demographic categories across California. Each subset annually records data from 2007 onwards, detailing unique individuals receiving homeless response services within different demographic subcategories.

Key to our analysis is the concept of the Continuum of Care (CoC), a statewide initiative established by the U.S. Department of Housing and Urban Development. The CoC program aims to facilitate rapid rehousing and long-term support for the homeless (U.S. Department of Housing and Urban Development, 2023). In this paper, "CoCs" refers both to the regional planning bodies and the geographic areas they cover. Notably, in California, some CoCs encompass multiple counties. [1-3]

Each dataset row presents the year, CoC, a demographic subcategory, and the number of individuals receiving services for that particular group. The age categories are '18-24', '25-34', '35-44', '45-54', '55-64', '65+', 'Under 18', and 'Unknown'. Race categories include 'American Indian, Alaska Native, or Indigenous', 'Asian or Asian American', 'Black, African American, or African', 'Multiple Races', 'Native Hawaiian or Pacific Islander', 'Unknown', and 'White'. Ethnicity is broken down into 'Hispanic/Latinx', 'Not Hispanic/Latinx', and 'Unknown'. Gender categories comprise 'Female', 'Male', 'Non-Singular Gender', 'Questioning Gender', 'Transgender', and 'Unknown'.

Table 1: Merged Homeless Demographic Data Sample

	Year	CoC	Total_Homeless	Age:18-24	Age:25-34	Age:35-44	Age:45-54	Age:55-64	Age:65+	Age:Under 18	Age:Unknown	Race:American Indian, Alaska Native, or Indigenous
31	2017	Tehama	395.0	30.0	79.0	71.0	62.0	78.0	22.0	53.0	0.0	19.0
95	2019	Santa Barbara	2801.0	160.0	452.0	476.0	508.0	579.0	223.0	417.0	39.0	120.0
132	2020	Santa Cruz	2396.0	151.0	384.0	421.0	376.0	362.0	218.0	471.0	48.0	97.0
171	2021	Tehama	286.0	22.0	39.0	50.0	28.0	33.0	18.0	74.0	23.0	13.0
173	2021	Yolo	1490.0	94.0	227.0	241.0	204.0	220.0	110.0	399.0	18.0	59.0
178	2022	Butte	2749.0	192.0	464.0	543.0	463.0	439.0	234.0	379.0	86.0	121.0
182	2022	El Dorado	434.0	45.0	41.0	51.0	41.0	62.0	38.0	152.0	0.0	15.0
190	2022	Orange	22967.0	1852.0	3453.0	3259.0	2715.0	2551.0	1075.0	4424.0	4042.0	499.0
196	2022	San Francisco	18214.0	1528.0	3578.0	3577.0	2700.0	2179.0	957.0	3937.0	303.0	1054.0
199	2022	San Mateo	5090.0	304.0	817.0	794.0	837.0	882.0	412.0	1090.0	83.0	252.0

Table 2: Housing Prices Data Sample

	Mon-Yr	CA	Alameda	Amador	Butte	Calaveras	Contra-Costa
0	1990-01-01	194952.0	226148.902299	NaN	102142.75	NaN	NaN
1	1990-02-01	196273.0	219306.000000	NaN	83333.00	NaN	NaN
2	1990-03-01	194856.0	225162.000000	NaN	100000.00	NaN	NaN
3	1990-04-01	196111.0	229333.000000	NaN	107999.60	NaN	NaN
4	1990-05-01	195281.0	232291.000000	NaN	100000.00	NaN	NaN

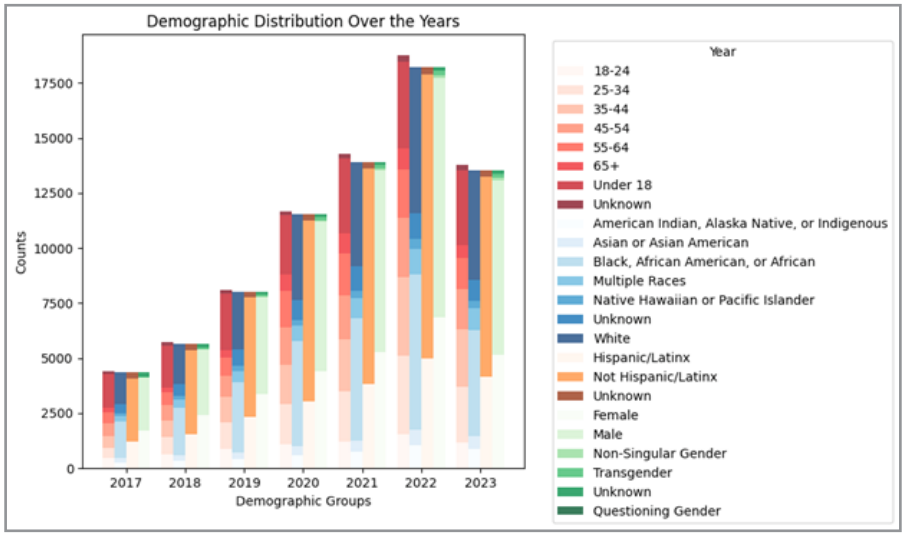

Figure 1: Distribution of Demographic Groups by Year for San Francisco from 2017 to 2022

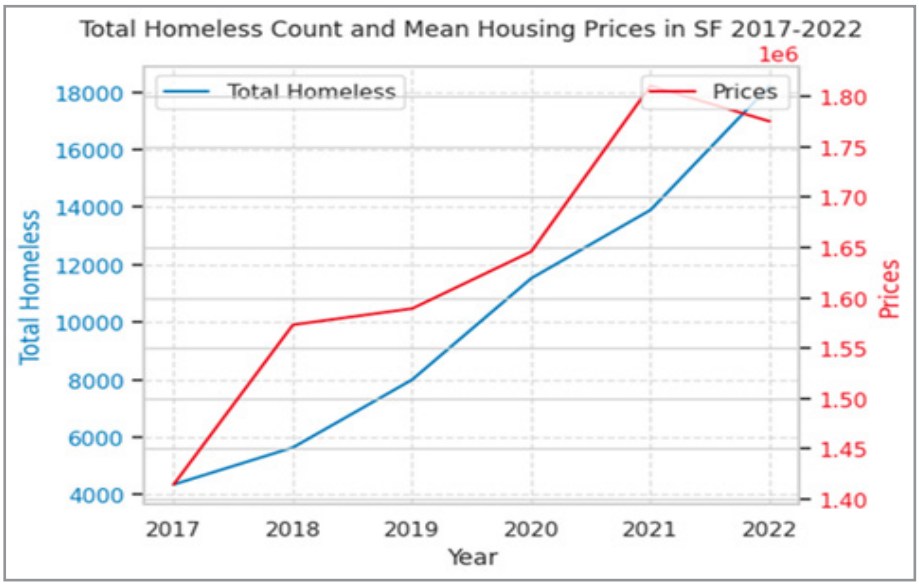

Figure 2: Total Homeless Count and Mean Housing Prices in San Francisco from 2017 to 2022

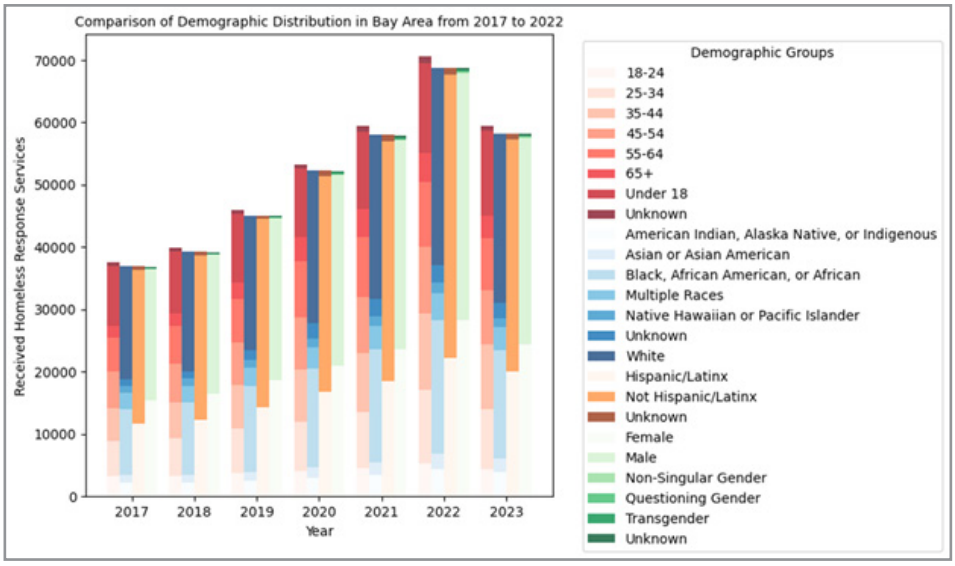

Figure 3: Distribution of Demographic Groups by Year for San Francisco from 2017 to 2022

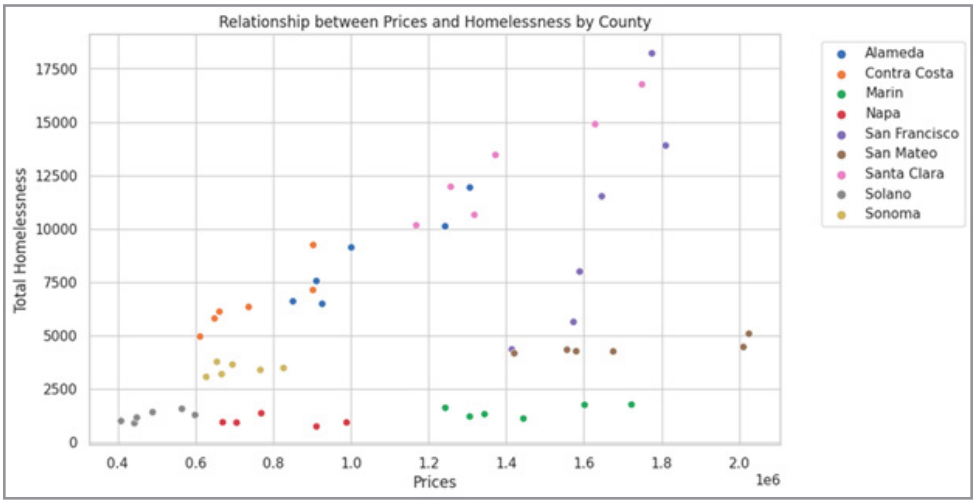

Figure 4: Relationship between Prices and Homeless by County

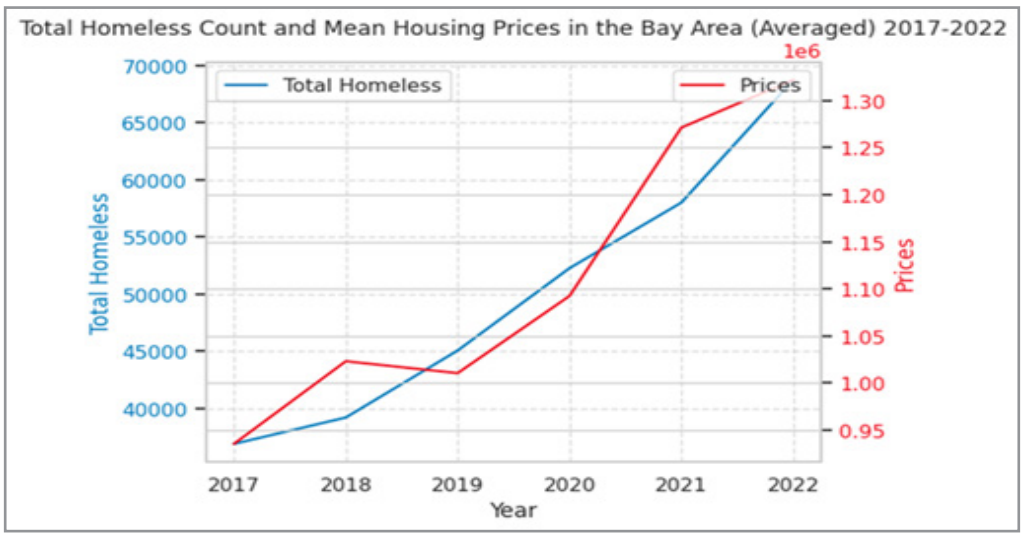

Figure 5: Total Homeless Count And Mean Housing Prices In California From 2017 To 2022

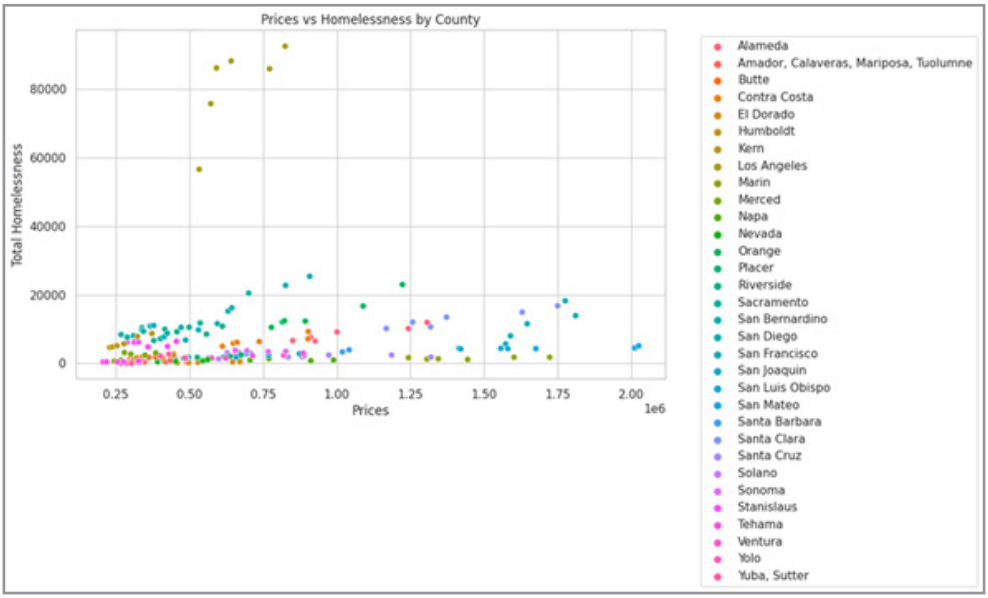

Figure 6: Scatterplot of Prices and Homelessness counts by County with Los Angeles

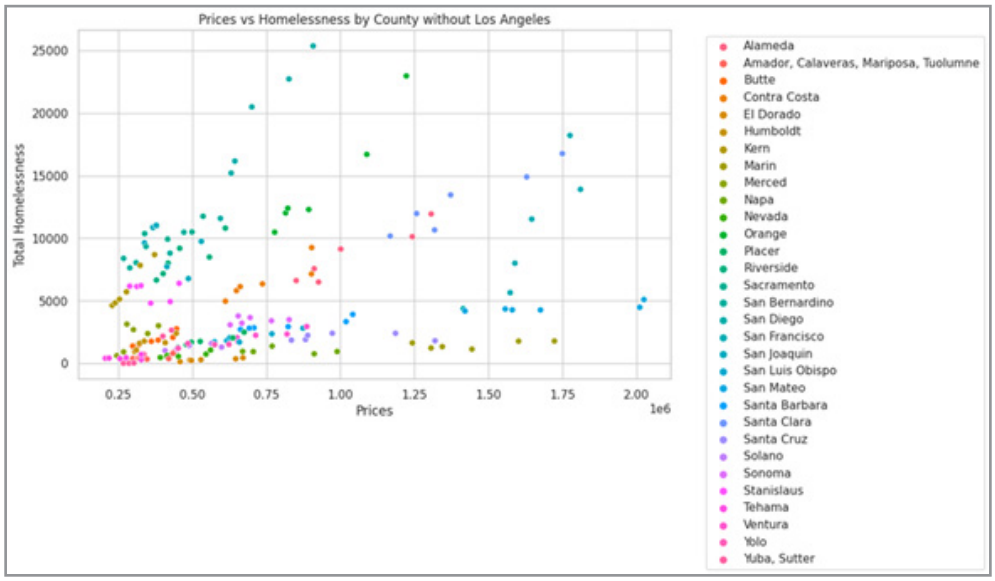

Figure 7: Scatterplot of Prices and Homelessness counts by County without Los Angeles

To streamline the data analysis process, we carefully developed custom functions within our code for efficient data cleaning and merging to create tables like the one above. These functions were crucial in consolidating and preparing the data from these multiple datasets into one unified format sufficient for our analysis. This would also inform the reliability and validity of our subsequent findings.

Limitations

It is crucial to emphasize that the homelessness dataset, which captures individuals who have accessed homeless response services within a Continuum of Care (CoC) for a specific year, has several inherent limitations that are crucial for understanding the broader context of our findings. Firstly, the dataset likely underrepresents the total homeless population, as it includes only those who have sought and received services. Many individuals who are homeless may not access these services due to barriers such as lack of awareness, mistrust of authorities, or logistical challenges. Additionally, there is a potential issue of data duplication, as individuals might move between different CoCs and be counted more than once, leading to inflated figures and a potentially inaccurate representation of the homelessness situation in a particular area. Furthermore, while the dataset provides demographic breakdowns, it may not fully capture the diversity and complexity of the homeless population. Groups with undocumented status or those not conforming to traditional gender categories might be underrepresented.

Another important thing to note is that the dataset's limitations include temporal constraints. Its annual nature might not effectively capture the dynamic nature of homelessness, which can fluctuate significantly within a year due to factors like seasonal employment, weather conditions, and policy changes. The variability in the availability and accessibility of homeless response services between different Continuum of Cares (CoCs) can also affect the accuracy of the data, influencing who gets counted. Additionally, different CoCs are likely to employ different methodologies for data collection on the ground, leading to inconsistencies in data gathering and reporting, which can affect the comparability of data across regions. Policy and funding changes at local, state, and federal levels can influence both the availability of services and individuals' willingness to seek them out, leading to year-to-year fluctuations in the data that may not solely reflect changes in the homeless population. Moreover, the dataset may not adequately reflect long-term trends in homelessness, focusing instead on the conditions of a specific year.

Recognizing these limitations is essential for a nuanced interpretation of the data. While the dataset provides valuable insights into the demographics and scale of homelessness as captured through service provision, it should be viewed as a starting point for analysis and our initial proxy for the count of people experiencing homelessness, rather than a definitive portrayal of the homelessness situation across California.

Housing Prices Dataset

For housing prices, we rely on the California Association of Realtors who collect and provide data on historical housing trends. In one page, they provide the Median Prices of Existing Single-Family Homes across California from 1990 to 2023. While this dataset offers simplified and potentially significant insights into the housing market, it's crucial to acknowledge the many limitations of selecting this dataset. See Table-2 in above.

Limitations

Firstly, the dataset focuses solely on single-family homes, excluding other housing types like apartments, condos, and multi-family units. This exclusion can skew our understanding of the overall housing market, especially in urban areas where diverse housing types are more common such as houses split into multiple rooms for rent in San Francisco. Secondly, the dataset presents median prices, which, while useful for indicating central tendencies, do not reflect the range and variability of housing prices within each region. This causes us to lose out on insights about high-end and low-end market dynamics.

Another limitation is the lack of granularity in geographical terms. The dataset does not differentiate between housing pric-

es at the neighborhood level, which can vary significantly even within the same city or county. This lack of specificity can lead to oversimplified interpretations of the housing market. Furthermore, the dataset does not account for the impact of external economic factors such as inflation, interest rates, and economic downturns, which can significantly influence housing prices. These factors are crucial for a comprehensive understanding of market dynamics but are also not directly captured in the dataset.

Finally, it is important to note that housing prices are influenced by a multitude of factors, including local policies, demographic changes, and shifts in housing demand and supply. Our analysis, therefore, requires careful consideration of these variables and an acknowledgment that the dataset serves as a limited representation of the complex housing market.

Despite these limitations, however, the dataset remains a valuable tool for our study, and it is the only one we've found that was publicly available and had comprehensive coverage of values for all counties in California. By being mindful of these constraints, we can use the data effectively as a proxy for understanding broader housing market trends in California.

Data Wrangling

Extensive data wrangling had to be performed before even beginning to explore the data (the accompanying .html or .ipynb file to this report goes is beyond the scope of this paper for publishing. If you are interested in obtaining a copy of it, please reach out to authors). Each of the four homelessness datasets had to have their rows for different demographic groups split into columns. Therefore, instead of having, say, "Year: 2017, CoC: Santa Clara, Age: 18-24, Experiencing Homeless Count: n" and "Year: 2017, CoC: Santa Clara, Age: 25-34, Experiencing Homeless Count: n2", we had consolidated rows to "Year: 2017, CoC: Santa Clara, Age: 18-24_Homeless: n, Age: 25-34_Homeless: n2".

This way, we could combine all the datasets and match the rows by Year and CoC. Then, for the median housing prices dataset, the data was actually split into months since 1990. We combined the data by getting the mean values per county per year. We then filtered for data only from 2017-2022, added the average values for counties in the Bay Area. Finally, we combined both datasets to have one final dataset containing the Year, CoC, the multiple columns for Homelessness Count by Demographic Group, and Prices.

Exploratory Data Analysis (EDA)

In this section, through all above figures from Figure-1 to Figure-7, we show what our python algorithm was plotting based on our collective data.

EDA: San Francisco

If we look at Figure-1 showing the stacked bar chart of demographic distribution for the homeless in San Francisco by year, are observed. Firstly, it is evident that there are always more values represented by age groups than other groups. This suggests that the completeness of data collection fluctuates over time, implying that there are challenges in consistently gathering full demographic information. In terms of age, the age groups '25-34' and '55-64' display notable representation, indicating that these age brackets are either more affected by the conditions being measured or are more likely to be included in the data collection process. In terms of gender, more male individuals access homeless response services, indicating a higher prevalence of homelessness among male-identifying individuals, or a higher propensity for men to seek out homeless response services. In terms of various racial groups, the majority is shared by the ''Black, African American, or African' group, followed by the 'White' group, while the rest of the categories present in lower numbers.

The trend from 2020 to 2023 shows a discernible increase in counts, which may reflect the consequences of recent socio-economic events or changes in the data collection methodology. There is a noticeable increase in the number of individuals receiving services from 2020 to 2022, which could be attributed to the economic impact of the COVID-19 pandemic. The groups identified as Non-Singular Gender and Transgender have lower visible counts, which could be indicative of their actual population numbers or point to larger systemic barriers in service access or data inclusion. The 'Unknown' category for age and race shows that there is a portion of the homeless population for which demographic data is not captured, reflecting inherent data collection challenges.

Figure 2 above for San Francisco from 2017 to 2022 shows a concurrent rise in both the total homeless count and mean housing prices, suggesting a potential correlation between the two. Notably, the data shows a parallel rise up to 2020, which coincides with the COVID-19 pandemic's start—a time of significant economic upheaval that probably had an impact on both the housing market and homelessness rates. Post-2020, while housing prices show signs of leveling off, the homeless count continues its upward trajectory, hinting at additional factors influencing homelessness beyond just housing costs.

The graph implies a strong relationship, with both the homeless count and housing prices following a similar growth pattern. However, as the scale in millions on the graph indicates, it is crucial to take into account San Francisco's high housing costs. Furthermore, despite the apparent correlation, causation cannot be established from this graph alone. Further analysis is required to account for confounding variables such as economic conditions, policy changes, and demographic shifts that could also be affecting these trends. This graph provides a clear starting point for discussing the dynamics between housing affordability and homelessness, but comprehensive statistical analysis is essential to fully understanding the complexity of these social issues.

EDA: San Francisco Bay Area

Figure-3 visualizes stacked bar charts average demographic distribution of individuals accessing homeless response services across the nine counties of the Bay Area from 2017 to 2022. When interpreting these averages, it's crucial to recognize the heterogeneity of the Bay Area; each county may have unique socio-economic conditions, housing markets, and resources for the homeless that the averaged data surely masks. Nonetheless, we can still glean some insight.

From the data, it seems that the age groups '25-34' and '55-64' consistently show higher counts across the years. This trend suggests that these age groups are notably affected by homelessness

in the Bay Area, a pattern that mirrors observations made within San Francisco. While this could indicate a regional issue, it may also be influenced by specific counties with higher counts in these demographics. The racial demographics represented in the chart—particularly the 'White', 'Black, African American, or African', and 'Hispanic/Latinx' groups—are the most prominent in terms of accessing homeless services. This prominence, however, does not necessarily equate to the overall demographic makeup of the homeless population but rather reflects those who are utilizing services.

A notable increase in the counts is seen post-2020, a trend that might correlate with the socio-economic impacts of the COVID-19 pandemic. However, the chart does not reveal the differential impact on individual counties, which could vary based on local responses and the severity of the pandemic's effects. Gender distribution shows a higher count for males accessing services. This observation aligns with broader trends but also raises questions about how service utilization might differ by gender across counties. Furthermore, the relatively lower counts for 'Non-Singular Gender', 'Questioning Gender', and 'Transgender' individuals suggest potential barriers to service access or data collection challenges for these groups.

The scatter plot in Figure-4 illustrates the relationship between housing prices and homelessness counts in the 9 various counties of the Bay Area. Each point represents a county's median housing price plotted against its total homelessness count. The graph shows a range of housing prices from approximately $400,000 to over $2,000,000 (denoted as 0.4 to 2.0 on the x-axis that is scaled by 1e6 for millions of dollars) and homelessness counts up to 17,500.

From this visualization, we can observe that counties with higher housing prices, such as San Francisco and Santa Clara, also have higher counts of homelessness. Conversely, counties like Solano and Sonoma, with relatively lower housing prices, show lower homelessness counts. This pattern suggests a potential correlation where counties with more expensive housing markets might experience higher rates of homelessness. Notably, counties like Marin, San Mateo, and San Francisco appear as outliers, deviating from the more linear trend observed among the other counties. These outliers have higher homelessness counts relative to their housing prices compared to counties such as Santa Clara and Alameda, which align more closely with the linear trend formed by the 6 other counties.

The deviation of Marin, San Mateo, and San Francisco from the linear pattern could imply that factors other than housing prices are significantly influencing homelessness rates in these counties. For instance, these areas might have higher rental market pressures, a lack of affordable housing inventory, or more pronounced income inequality, which can exacerbate homelessness irrespective of the average housing price. Additionally, the presence of robust homeless services in these counties might lead to higher reported counts due to more comprehensive data collection. At these times, it's important to remember that this is a count of unique individuals who've received homeless response services.

EDA: California

The line graph illustrated in Figure-5 displays the total homeless count and mean housing prices in California from 2017 to 2022, revealing a simultaneous increase in both metrics over the years. The trend lines suggest a correlation between rising housing prices and an increase in homelessness across the state. However, while the trend lines run parallel for the initial years, there is a noticeable convergence starting in 2020, potentially indicating that the rate of increase in homelessness is outpacing that of housing prices or that other factors intensified during this period, such as the economic impact of the COVID-19 pandemic. This observation suggests that while housing price escalation may be a contributing factor to rising homelessness, it is likely not the sole driver, and other economic or social factors may also be influencing this trend.

Furthermore, Figure-6 presents a Scatterplot of Prices and Homelessness counts by County with Los Angeles.

The scatter plot in Figure-6 above displays the relationship between housing prices and homelessness counts by county in California. Immediately, we can see that Los Angeles is a significant outlier, so we've also remade the plot without Los Angeles. Focusing on the second graph, which excludes Los Angeles, allows for a clearer examination of patterns across other counties, including those in the Bay Area. Figure-7

With Los Angeles removed, we observe that Bay Area counties like San Francisco, Alameda, and San Mateo do not exhibit a simple linear relationship between housing prices and homelessness. While these counties have relatively high housing prices, their homelessness counts vary, with some not following the expected trend of higher prices correlating with higher homelessness counts. This suggests that in the Bay Area, other factors are influencing homelessness beyond housing costs alone. For instance, San Francisco, despite its high housing prices, does not have the highest homelessness count as one might expect if housing costs were the sole factor.

Other counties with lower housing prices, such as those in the Central Valley and more rural areas of the state, also show a wide range of homelessness counts, indicating that lower housing costs do not automatically equate to lower rates of homelessness. This further suggests that additional variables like employment rates, availability of social services, and local economic conditions play significant roles in influencing homelessness. Furthermore, the spread of homelessness counts among counties with similar housing prices on the second graph indicates a complex interaction of factors at the county level. For instance, counties with similar housing prices have vastly different homelessness counts, suggesting that each county's approach to housing, support services, and economic opportunities can greatly affect the number of homeless individuals.

Data Analysis

As illustrated in the following tables our data has been analyzed utilizing the Regression Analysis Methodology, for San Francisco and Bay Area Counties as well throughout State of California.

Regression Analysis: San Francisco

The analysis conducted in Table-3 is driven by three variations of Ordinary Least Squares (OLS) regression models which were applied to a dataset representing San Francisco's housing market and demographic variables. The first model, the original OLS regression, and the second, a pruned version removing less significant variables, both achieved perfect R-squared and adjusted R-squared values of 1.0. Typically, such scores would indicate that the model explains 100% of the variability in the target variable (housing prices). However, in this context, these perfect scores are highly indicative of overfitting. This suspicion of overfitting is primarily due to the extremely small sample size of the dataset, which consisted of only 6 rows. In such cases, the model tends to learn the noise and specific details of the training data to an extent that it perfectly predicts the outcome but fails to generalize to new, unseen data.

Table 3: San Francisco Regression Analysis

Model	R-squared	Adjusted R-squared
Original	1.0	1.0
Pruned (p<0.05)	1.0	1.0
PCA	0.0	0.0

On the other hand, the third model, which incorporated Principal Component Analysis (PCA) before regression, yielded an R-squared and adjusted R-squared of 0.0. This drastic shift suggests that the PCA transformation, in this case, removed or altered the features' predictive power. This outcome could be due to the dimensionality reduction process in PCA, which can sometimes lead to the loss of significant information, particularly in a small dataset.

The stark contrast between the perfect scores in the first two models and the complete lack of explanatory power in the PCA-based model underscores the challenges of working with very small datasets. It highlights the importance of having a sufficiently large and representative sample to build robust predictive models. The results from these models are more reflective of the limitations of the dataset, and rather require us to rely on a larger dataset, perhaps one that collects more data in intervals between the years. Consequently, further exploration of the results in the Bay Area counties and greater California area are necessary.

Regression Analysis: Bay Area Counties

Table-4 below is the summary of results for an OLS Regression conducted across all the Bay Area counties and pruned three times to filter for variables with high p values until we landed on these variables.

The model's constant term has an estimated value of approximately 830,900, with a highly significant p-value, establishing a substantial baseline for housing prices in the absence of other variables. This points to other influential factors, beyond those captured in the model, contributing to this baseline value in the Bay Area's housing market. The age demographic '35-44' shows a negative coefficient of -621, implying that an increase in the homeless count within this age group is associated with a slight decrease in housing prices.

However, the p-value of 0.055 is just above the conventional threshold for statistical significance, which suggests that this result should be interpreted with caution.

Table 4: Summary Table for Pruned OLS Model of Demographic Groups vs Prices for Bay Area

Statistic	Value	
R-squared	0.717	
Adjusted R-squared	0.694	
Demographic Group	**Coef.**	**P>|t|**
const	8.309e+05	0.000
Age:35-44	-621.0004	0.055
Age: Unknown	4428.4561	0.000
Race: Multiple Races	-3898.9566	0.000
Gender: Male	584.8694	0.000
PCA	0.0	0.0

On the other hand, the 'Unknown' age category shows a strong positive correlation with housing prices, denoted by a coefficient of 4428.4561 and a p-value indicating high statistical significance. The 'Multiple Races' demographic exhibits a significant negative correlation, with a coefficient of -3898.9566. The 'Male' gender group is positively associated with housing prices, with a coefficient of 584.8694, again confirmed by a significant p-value. The model's R-squared value at 0.717 suggests a substantial explanatory power, accounting for over 71% of the variation in housing prices with the variables considered. The Adjusted R-squared, slightly lower at 0.694, accounts for the number of predictors, indicating that the model is not unduly complicated by unnecessary variables.

Regression Analysis: California

Table-5, is illustrates a summary table for Pruned OLS Model of Demographics Groups vs. Prices for California

Table 5: Summary Table for Pruned OLS Model of Demographic Groups vs Prices for California

Statistic	Value	
R-squared	0.497	
Adjusted R-squared	0.489	
Demographic Group	**Coef.**	**P>|t|**
Variable	**Estimate**	**p-value**
const	501600.0000	0.000
Age:45-54	-386.9682	0.000
Age:65+	744.3669	0.000
Race: Native Hawaiian or Pacific Islander	3227.2835	0.000
PCA	0.0	0.0

The regression model employed to explore the relationship between various demographic groups and homelessness rates in California offers insightful findings. The model's R-squared value stands at 0.497, indicating that about 49.7% of the variability in homelessness rates is explainable by the demographic variables included. The adjusted R-squared, at 0.489, further

underscores the model's robustness, adjusting for the number of predictors and affirming a good fit considering the degrees of freedom.

In terms of the model's coefficients and their significance, each demographic group's coefficient is accompanied by a p-value, providing a measure of statistical significance. The constant term, or the intercept, is set at 501,600, representing the estimated homelessness rate when all other variable values are zero. This intercept is statistically significant, as evidenced by a p-value of less than 0.001, suggesting a meaningful baseline for the model.

For the age group '45-54', the model reveals a coefficient of -386.9682, pointing to an association with a decrease in homelessness rates. This negative coefficient is supported by a highly significant p-value, indicating a robust relationship. In contrast, for the '65+' age group, the coefficient is 744.3669, suggesting an increase in homelessness rates associated with this demographic. This positive association is also statistically significant, as reflected in its p-value. Another notable finding is in the racial category 'Native Hawaiian or Pacific Islander', where the coefficient is a substantial 3,227.2835. This figure indicates a significant increase in homelessness rates for individuals in this demographic group, and the p-value being highly significant further validates this association.

Conclusion

This study set out to explore the relationship between homelessness levels and housing prices across the Bay Area. We leveraged government datasets on homelessness and realtor-provided median single-home family housing prices to delve into the socio-economic fabric of the Bay Area, San Francisco, and California at large. Our findings suggest a nuanced relationship between demographic variables and housing prices that was, on average, positively correlated.

In San Francisco, the perfect R-squared and adjusted R-squared values obtained from our OLS regression models pointed towards overfitting, likely due to the limited size of the dataset. This limitation was further highlighted in the PCA-based model, which showed no explanatory power, underscoring the challenges of working with small datasets and the potential loss of crucial information in dimensionality reduction processes.

As such, we knew we'd find more insight if we expanded our analysis to the Bay Area. Across the 9 counties, we found that certain demographic groups, such as '18-24' and 'Native Hawaiian or Pacific Islander', showed a positive correlation with housing prices, potentially indicative of gentrification or demographic shifts in specific neighborhoods. Conversely, groups like 'Asian or Asian American' and 'American Indian, Alaska Native, or Indigenous' were negatively correlated with housing prices, hinting at socio-economic challenges or regions with more affordable housing.

Expanding further to the California state level, the model's R-squared value of 0.717 provided a robust explanation for a substantial portion of the variance in housing prices. However, the adjusted R-squared of 0.694 indicated that there are still significant factors at play that are not captured by our model. The statewide analysis painted a diverse demographic landscape, with different age and race groups showing both negative and positive associations with housing prices, underscoring the multifaceted nature of homelessness and housing economics.

The study, while extensive in its exploration, is ultimately greatly limited due to the nature of the datasets and the inherent challenges in capturing the full spectrum of factors influencing both homelessness and housing markets. It is clear that a variety of interrelated factors, including but not limited to economic conditions, demographic changes, and local policies, have an impact on housing prices and homelessness levels.

In closing, our findings accentuate the intricacy of the relationship between the housing crisis and market dynamics. They underscore the critical need for nuanced and data-informed approaches to effectively address and understand the multifaceted issues faced by urban areas such as the San Francisco Bay Area and the broader California region. Our paper demonstrates that while there is a discernible correlation between certain demographic groups and housing prices, the true nature of this correlation is complex and demands thoughtful analysis to guide the development of effective and equitable housing policies.

References

1. California Association of Realtors (2023) Historical Housing Data. Median Prices of Existing Single-Family Homes.
2. California Interagency Council on Homelessness (2023) People receiving homeless response services by age, race, ethnicity, and gender. People Receiving Homeless Response Services by Age, Race, Ethnicity, and Gender (Datasets).
3. U.S. Department of Housing and Urban Development. (2023) Continuum of Care Program. CONTINUUM OF CARE PROGRAM.

Copyright: *©2024 Bahman Zohuri, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*