Understanding Aging Populations and Home Value

Introduction

Our project seeks to understand how demographic changes in the state of California are impacting the housing market. Our group wanted to study the housing market due to its dual federal and local significance. The overall health of the housing market is seen as a fundamental indicator of national economic health as well as a key ingredient in establishing upwards social mobility and improving personal wellbeing. In 2020, spending allocated to residential construction, renovation, and brokerage fees alone accounted for 885 billion dollars in spending for roughly 4.2% of the national GDP. Likewise, this group is painfully aware of the housing problems facing our own communities; the neighborhoods of Santa Clara County and San Mateo County are both markets infamous for extremely high pricing and extremely low availability.

Understanding how the housing market evolves according to certain conditions is essential to crafting robust policies that both facilitate economic growth and are resilient to potential instability. One key area of instability is evolving demographic changes; specifically, an aging population. The U.S. Administration for Community Living projects by the year 2040, there will be over 80 million people in the U.S. over the age of 65, compared to 54.1 million people over the age of 65 in 2019. The demographic changes are largely out of the control of policy-makers and individual citizens. In order to craft policy and governmental responses to these changes, we require a better understanding of the consequences of demographic evolution in the housing market. Our research question is thus: **how does an evolving population age impact housing prices in the State of California?**

In this paper, we present two analyses of the role of population age in the overall housing market by specifically analyzing the impact of age on housing prices. We use Zillow's Home Value Index (ZHVI) as a real in-time measurement of housing price and our data set consists of this index and publicly available demographic information over time from the U.S. Census Bureau. The first analysis is an ordinary least squares regression of ZHVI against a county's median age for a given year. In the first analysis, we find a slight positive correlation between median age and home value in each year, which worsens over time. In 2019, three counties (Santa Clara County, San Mateo County, and San Francisco) were considerably above the regression line. The second analysis models home price as an order-1 quasi-autoregressive time series including percentage change in age demographics as exogenous inputs. We find that the AR model fits the given observational data well when including a time feature. We also find that the model predicts an increase in home value when younger population groups increase in size, but a considerably smaller increase in home value when older population groups increase in size.

Literature Review

Discerning the relationship between changes in age demographics and the housing market is a well studied subject. A study conducted by the Department of Housing and Urban Development analyzed to what extent elderly homeowners' homes appreciated at different rates relative to younger counterparts. The study was conducted in 2005 using data from the HUD's Health and Retirement Study, and it finds that homeowners aged 75 and older see their homes "appreciate in real terms 1.0 to 1.2 percentage points per year than houses of middle-aged (50 to 74 year old) owners." A key limitation of this study is age. Although the study confirms findings from previous studies (e.g., Davidoff 2004), the study was conducted in 2005, well before the Great Recession of 2008 radically upended mortgage lending and housing availability in global markets.

Firoozi et. al 2020 more directly touches on our research question; proposing a comparative analysis of percentage point changes in age demographics and home value in Singapore and the U.S. The model uses indicators of lending, interest rates, and housing availability as features to predict housing prices but not ZHVI data. In the U.S, it finds that percentage point increases in ages 45-54 and 55-59 are associated with increases in predicted home value, but that p.p. increases in ages 60-64 is associated with decreases in predicted home value: "the correlation flips." The AR model proposed by our group is significantly inspired by Firoozi et. al., with several key differences; mainly our usage of lagged ZHVI data (which makes our model autoregressive) instead of housing market health indicators and lending data; likewise, our model focuses solely on the State of California.

Data Sources

After clearly identifying the research question, viable data sources were compiled and evaluated. Upon evaluation, it was determined that the best data sources to reference could be sourced from Zillow and the United States Census Bureau for the state of California. In addition to compiling and listing houses on the market, Zillow also has open source datasets containing historical housing data such as home value forecasts, inventory data, list and sale prices, sale count and price cut, and home values (Zillow 2022).

From these available metrics provided by Zillow, the California home values dataset was identified to be the most relevant and useful for the project. This dataset listed the Zillow Home Value Index (ZHVI) over time. The ZHVI is defined as "a smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type. It reflects the typical value for homes in the 35th to 65th percentile range" (Zillow 2022) in USD.

In other words, this metric provided an accurate home value for houses sorted by county, state, and date, adjusting for seasonal and inflation changes. Given that the home value was optimal for comparing to demographics recorded in the US census, this was the chosen dataset. In order to analyze the appropriate demographics of the California counties, the other data source was acquired from the US Census Bureau through `tidycensus`.

After the datasets were chosen, they were uploaded to R where the data sets were cleaned so that the appropriate and relevant information would remain for easy manipulation. In order to fully answer the research question, the problem was approached with different aspects in mind, such as conducting a basic regression between median age and ZHVI, analyzing outliers, mapping correlations between ZHVI and age over time, as well as exploring autoregression.

Results

Model 1: Linear Regression of ZHVI Against Median County Age

In this analysis, we are working with the following model:

$$ZHVI_{c,t} \approx \beta_0 + \beta_1 X_{c,t} + w_{c,t}$$

Here, $ZHVI_{c,t}$ refers to the Zillow Home Value Index of country c at time t, $X_{c,t}$ refers to the median age of country c at time t, and $w_{c,t}$ is some normally distributed noise. We fit the coefficients via ordinary least squares (`lm` in R). Note that each regression model is prepared on a year-to-year basis.

Statistical Hypothesis
Our research question is seeking to understand the contribution of aging populations to home values in California. Using median age, our claim is that if older counties are associated with higher home values, then as a county ages, we should expect to see its associated home value increase.
As such, we propose the following null hypothesis:

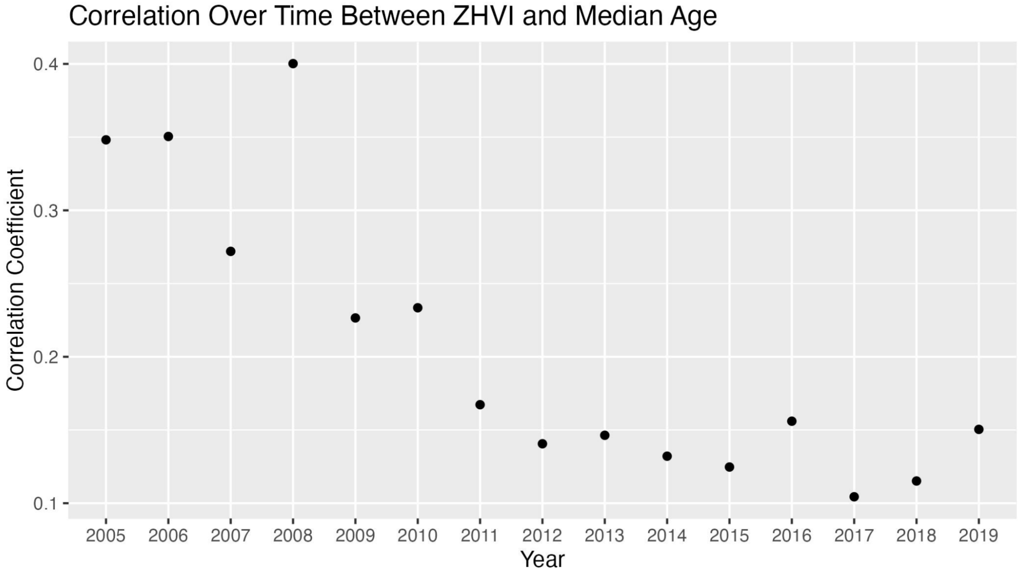
(NH) There is no relationship between median age and home value

Likewise, we also propose the following alternative hypotheses:

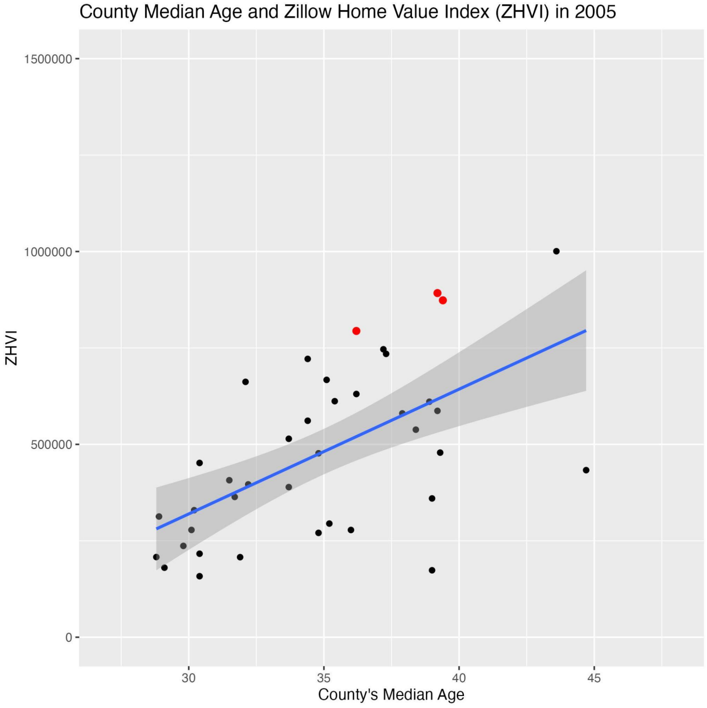
(AH 1) Counties with higher median ages have higher home values
(AH 2) Counties with higher median ages have lower home values

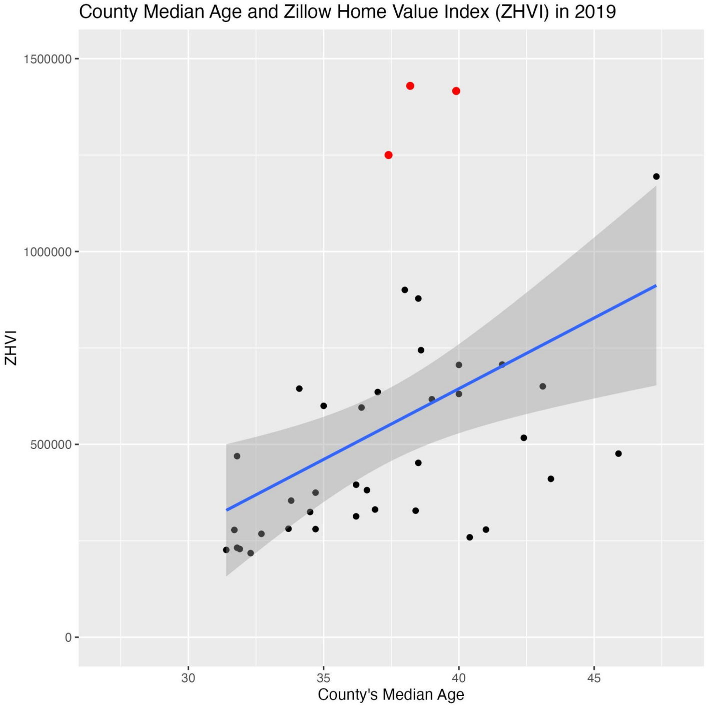
Results

After calculating correlation coefficients for every year from 2005 to 2019, we see an overall trend downwards in correlation:



This is perhaps best explained by increases in the housing value of three outliers: San Francisco, San Mateo, and Santa Clara County. When calculating correlations that *exclude* these outliers, we see correlation vastly improve for years with the worst disparities (notably, 2018 and 2019). This is expected, given how far these three counties move away from other California counties over time. To best visualize these changes, see our gif animation of the correlations; however, for purposes of concision, the 2005 and 2019 visualizations below capture roughly the same idea:

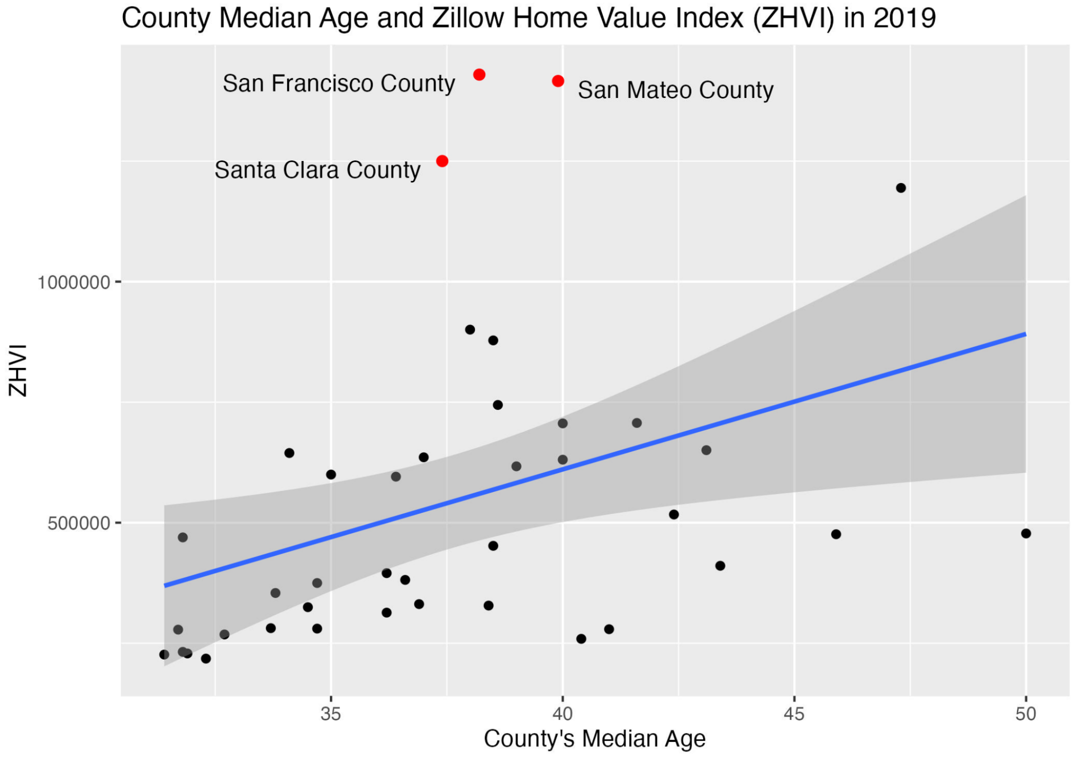




From simple linear regression alone, we were able to capture much of our statistical question. As counties age, it seems that housing value roughly increases. Though not a strong correlation, positive trends exist in every year from 2005 to 2019, and even the weak correlations can be explained. Our correlation coefficient only ever reaches 0.40 at maximum; however, from a purely non-statistical perspective, there is no limit on the number of variables that can inform housing value. Racial and income demographics, assets and generational wealth, quality of education and level of education completed, unemployment and job growth, political involvement and leanings - all of these factors and countless others are intertwined with housing pricing.

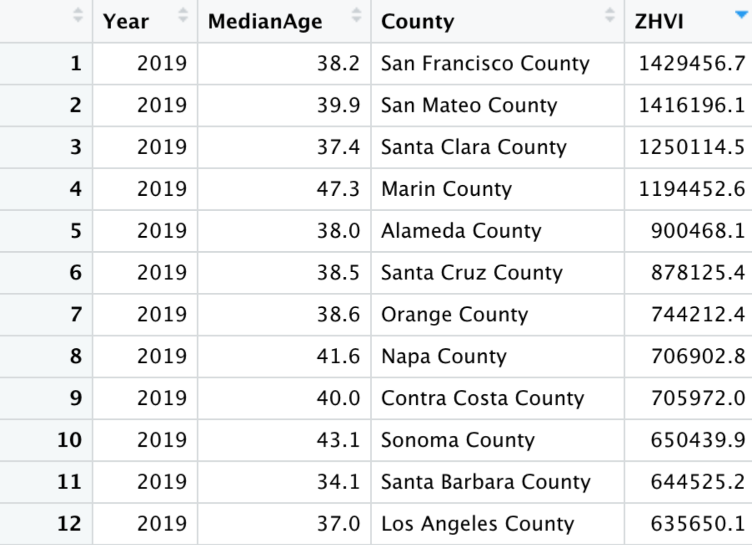
Why, then, visualize the relationship at all? For one, age is out of the control of policy-makers and individual citizens but changes in ways we can anticipate. For example, in our 2005 and 2019 visualizations, the regression lines shift right, capturing the aging of California homeowners. Secondly, though the actual impact of age may be minimal, it's still there, even as one of (seemingly) infinite factors. Perhaps *knowing* that a myriad of factors are at play, the fact that we even see a correlation of 0.40 (in 2008) is more noteworthy. Finally, what might not be captured in just one associative relationship is certainly seen over time. As we alluded to earlier, there are three jarring outliers that almost single-handedly tank the correlation coefficients in later years. These outliers are, in order of ZHVI value, San Francisco County, San Mateo County, and Santa Clara County. We now take a brief detour by expanding more thoroughly into characterizing the housing markets of these three counties.

Who are the Outliers?



From our analysis, we observed that San Francisco, San Mateo, and Santa Clara were outliers with the three highest Zillow Home Value Indices (ZHVIs) relative to low median ages under 40. Their ZHVIs were respectively $1,429,456.7; $1,416,196.1; and $1,250,114.5 in 2019. Although these counties occupy the three highest ZHVIs, it is notable that the top 6 highest counties ranked by ZHVI are *all* located in the Bay Area.

Table 1. California Counties Ranked by 2019 Highest Zillow Home Value Index



	Year	MedianAge	County	ZHVI
1	2019	38.2	San Francisco County	1429456.7
2	2019	39.9	San Mateo County	1416196.1
3	2019	37.4	Santa Clara County	1250114.5
4	2019	47.3	Marin County	1194452.6
5	2019	38.0	Alameda County	900468.1
6	2019	38.5	Santa Cruz County	878125.4
7	2019	38.6	Orange County	744212.4
8	2019	41.6	Napa County	706902.8
9	2019	40.0	Contra Costa County	705972.0
10	2019	43.1	Sonoma County	650439.9
11	2019	34.1	Santa Barbara County	644525.2
12	2019	37.0	Los Angeles County	635650.1

We sought to better understand the reason behind the presence of these outliers, as well as how they move away from other California counties. The outlier counties observe as large as a $500,000 increase in the difference between the median home value in comparison to other counties from 2014 - 2019. We suspect this is largely due to the major changes occurring in the Bay Area and the massive growth in Silicon Valley, of which there is already existing literature.

Model 2: Quasi-Autoregressive Time Series

In response to the impact of outliers on the above model, we sought a more sophisticated model to capture a better understanding of demographic changes *over time*. The model is as follows:

$$\Delta ZHVI_t \approx \beta_0 + \beta_1 \Delta ZHVI_{t-1} + \sum_{i \in age\ groups} \gamma_i \, \Delta X_{i,t} + \sum_{j \in year} \kappa_j \, I(t = j) + w_t$$

Above, $\Delta ZHVI_t$ refers to the percentage change in ZHVI from $t - 1$ to t, $\Delta X_{i,t}$ refers to the percentage change in age group i from time $t - 1$ to t and w_t is some normally distributed noise. We work with percentage change in ZHVI and age group because our initial observations

revealed that the ZHVI time-series is not stationary; home prices appreciate over time. Our age groups buckets are: <25 years old, 25-39 years old, 40-54 years old, 55-69 years old, and >69 years old. The relevant years are 2007 until 2019. Years 2005 and 2006 are present in the dataset but are lost when computing percentage changes and the order-1 lag for ZHVI data.
We fit the intercept (β_0), autoregressive coefficient (β_1), age group coefficients (γ_i), and factorized year coefficients (κ_j) using ordinary least squares. Our data consisted of a total of 601 observations, 102 of those were excluded due to missing ZHVI data. The R-Squared value is 0.8815. Selected coefficients and their standard errors are detailed in the following table:

Feature	**Coefficient Value** (% increase)	**Standard Error**
Autoregression	0.387	0.040
P.P. increase in age < 25	0.132	0.101
P.P. increase in age 25-39	0.058	0.045
P.P. increase in age 40-54	0.136	0.061
P.P. increase in age 55-69	0.0685	0.065
P.P. increase in age >69	-0.005	0.043

The coefficients in the table answer the question "given a sudden shock in the population of a particular age group, how much is the ZHVI for the current year expected to change?" To illustrate this with an example from the table, we see that holding all other factors equal (e.g., considering a fixed year and previous ZHVI percentage difference), a 1 percentage point increase in the population aged less than 25 is associated with a 0.132 percentage point increase in the year-over-year change in ZHVI (meaning the ZHVI is expected to increase *additionally* by 0.132% in response to the shock).

The corresponding standard errors of the coefficient are unfortunately quite large, which makes it difficult to draw conclusions by interpreting coefficient values. If we do analyze point estimates, however, we see that an increase in the population aged 69 and higher results in a smaller (possibly even negative) p.p increase in ZHVI. The fact that the correlation turns negative for a particular age group aligns with the findings of Firoozi et. al 2020 as described in the Literature Review section.

Overall, we also note that an increase in middle-aged citizens (people aged 40-54, ostensibly capturing the aging of older-millennial and younger-Gen-X classes) is associated with a positive pp increase in ZHVI, which gives more detail to the findings from Model 1. In particular, the AR model predicts that not only do home values increase as the population ages, but also that the greatest impact to home values is due to increases in the middle-aged population.

Limitations

The limitations on our research are mainly the availability of data and the inability to quantify how much age accounts for our statistical relationship. Zillow housing data is extremely granular and can provide housing values down to the month and neighborhood. To see more general trends and to more easily join data, we chose to analyze county by county, but home values often vary *within* counties, sometimes dramatically so. By generalizing to counties, we unfortunately gloss over these intracounty trends. Throughout our analyses, we also alluded to the complex nature of our core question. Given that we used a simple linear regression to see a *general* relationship between housing value and age, we were not able to capture the same level of complexity that innately informs the question. For the Quasi-Autoregressive model, its autoregressive quality is, frankly, somewhat poorly justified. Our time-series consists of just 13 years due to limitations in Zillow data and housing prices overall have experienced a mostly linear increase in the recent decade.

With more time, we would have created a multiple linear regression to see, presumably with much statistical uncertainty, *how much* age, on average, accounts for changes in home value, as compared to other covariates such as race and income demographics. This is a slightly different question than what the autoregression answers to. We would have liked to analyze trends in home value and ownership based on race and income, as well as explore *affordability* for Californians, as opposed to how homes are priced on the market.

Conclusion

Home ownership, often seen as a fixture of American life, is a rapidly disintegrating ideal, shadowed by wealth gaps and a housing market ever out of reach for middle class Americans. As a group of soon-to-be graduates witnessing exorbitant housing prices in real time, we hoped to explore the relationship between demographic changes and the housing market.

Regarding linear regression, we used median age as our demographic variable and the Zillow Housing Value Index (ZHVI) as our measure of housing value, ultimately finding a weakly positive relationship between age and housing price. Thus, we rejected the null hypothesis that there exists no relationship between the two variables. A simple linear regression also effectively pictured associations over time and was especially important in identifying outliers. The correlation coefficients over time steadily decrease with the “help” of San Francisco, San Mateo, and Santa Clara County. Though the distancing of these Bay Area counties is attributable to numerous factors, we find that the general consensus among researchers has to do with surges in income inequality and the expanding domain of tech companies as employers and property owners in Silicon Valley.

Next, we utilize a model typically used for prediction, a quasi-autoregressive time series, to further understand the statistical question. From this analysis, we were able to distill how “shocks” (or percentage point increases) to different age groups would affect housing value. Most notably, middle-aged populations (people aged 40-54) were responsible for the largest increases in housing value. By approaching our project first with a simple linear model, we were able to identify an interesting group of outlying Bay Area counties with exorbitant ZHVIs. Afterwards, we were able to get at group-by-group impacts for populations of different ages.

Our simple linear regression and quasi-autoregressive models did not provide substantial evidence of the existence of a trend; the correlations are simply too noisy and the standard errors too large. However, it is clear to our group that our analysis revealed the existence of some positive trend between aging and home value, and to the extent that the model is valid, we believe this is a topic warranting further statistical research.

Works Cited

Davidoff, Thomas and Welke, Gerd, Selection and Moral Hazard in the Reverse Mortgage Market (2004, October 21). Available at SSRN: https://ssrn.com/abstract=608666 or http://dx.doi.org/10.2139/ssrn.608666

Firoozi, Fathali, Jalilvand, Abolhassan, Lien, Donald and Oliver, Mikiko, (2020), The Impact of Population Aging on Housing Prices: A Comparative Study of Singapore and the U.S , *International Real Estate Review*, **23**, issue 4, p. 467-482.

Rodda, David and Patrabansh, Satyenda, (2005, December), The Relationship Between Homeowner Age and House Price Appreciation. Prepared for the U.S. Department of Housing and Urban Development Office of Policy Development and Research, Retrieved May 31, 2022 from https://www.huduser.gov/publications/pdf/HouseAppreciation_and_age_relationship.pdf.

Schuetz, J. (2020, December 9). Rethinking homeownership incentives to improve household financial security and shrink the racial wealth gap. *Brookings*. https://www.brookings.edu/research/rethinking-homeownership-incentives-to-improve-household-financial-security-and-shrink-the-racial-wealth-gap/

Weinstock, L. R. (2021, May 21). *Introduction to U.S. Economy: Housing Market*. Congressional Research Service. 3. Retrieved May 31, 2022 from https://crsreports.congress.gov/product/pdf/IF/IF11327

_, Causes and Effects of High House Prices in San Francisco. (2020, September 13). *The Cannon Yarder*. https://cannonyarder.com/2020/09/13/causes-and-effects-of-high-house-prices-in-san-francisco/

_, Housing Data. (n.d.). *Zillow Research*. Retrieved May 31, 2022, from https://www.zillow.com/research/data/

_, *Projected Future Growth of Older Population | ACL Administration for Community Living*. (n.d.). Retrieved May 31, 2022, from https://acl.gov/aging-and-disability-in-america/data-and-research/projected-future-growth-older-population