Week Eight: Capstone

Stephen Blamer

Colorado State University Global

MIS581 Capstone – Business Intelligence and Data Analytics

Dr. Lisa Bryan

10/10/2021

Table of Contents

Abstract 4

Introduction of Topic and Organization 5

Objectives 7

Overview of Study 8

Research Hypothesis 10

Literature Review 12

Research Design 13 Methods and Methodology 13

Limitations 15

Ethical, Privacy, and Security Considerations 15

Study to Results 20

Findings 21

Income 21

Housing Value 22

Annual Income to Monthly Housing Cost Ratio 24

Monthly Housing Costs 25

Monthly Housing Costs as a Percentage of Income, 30% or Greater 27

Owner-Occupied Housing as a Factor of US Adult Population 29

Conclusion 30

Recommendations 31

GitHub Link 32

References 33

Week Eight Capstone: US Housing

Abstract

This projects concerns how housing affordability has changed in the US from 2010 to 2019, following the 2008 housing crisis. It examines whether incomes, monthly housing costs, housing values, and the number of owner-occupied homes has increased. It also looks at how the ratio of monthly housing costs to annual income has changed in the study period, and whether owner-occupied housing has changed as a measure of the US adult population. The data was obtained from the US census bureau, reformatted for analysis, and analyzed using SAS, using paired t-tests and simple linear regression to examine changes.

Most of the null hypothesis, being no change, failed to be rejected, apart from the number of those paying =>30% of their income on housing. This value decreased significantly for all but the lowest income categories, showing that those who live in owner-occupied housing are overall paying less of their incomes on housing. The number of owner-occupied houses decreased by ~6%, with the US adult population increasing by ~6% over the same period, with a ratio loss of ~2%. This shows that fewer adults are owning and living in those homes in the US.

These findings were limited by a small N-value, due to the aggregated nature of the data. Other studies using individual years’ raw data, or incorporating rental, multi-family, and other housing types is suggested for more information. Overall, the study showed that while fewer adults are owning and living in their homes, those with them are more able to afford them, especially at higher income levels.

*Keywords*: housing, affordability, US Census, income

**INTRODUCTION OF TOPIC AND ORGANIZATION**

Home ownership has long been a staple of the “American Dream”. However, for many, it seems to be getting more and more difficult. This paper concerns itself with how the United States has recovered from the housing crisis of 2008. Specifically, it is examining how incomes have changed in the United States from 2010, the beginning of recovery, to modern times (2019). It approaches the subject by looking at incomes and housing prices for owner-occupied homes directly, as well as, how incomes relate to housing affordability, using several metrics. The data being examined comes from the United States Census Bureau, who collects information every ten years, as well as annually for many metrics.

The United States has conducted a census since 1790, and work was guided by the secretary of state with the work being done by Federal Marshals. This work consisted of gaining accurate counts of US populations and details of the citizens, land use, etc. The Census Bureau itself was formed into a permanent agency in 1902, and laws structuring statistics and other rules were added to the US code in 1954 (Census.gov, 2021). The US census is a mandate from the constitution and is required every year ending in ‘0’.

Obviously, tabulating the information from a census across a nation the size of the United States is a large undertaking, especially as the population was expanding rapidly. In the 1880s, Herman Hollerith created a tabulating machine to enter data from individuals using punch cards, based on a design of a weaving loom, then reading the cards and summarizing information and sorting the cards as well (Diamond, 2021). Hollerith’s inventions later were vital in creating the company now known as IBM. The most recent census, in 2020, was conducted by mail, phone, and electronically as much as possible, given the COVID-19 pandemic. Before this, interviews were a vital part of the process. Follow up work will be required in person to collect information from non-responsive members of the public. The modern question form consists of only nine questions, with additional information now gained through the American Community Survey, conducted annually, with roughly 3.5 million households asked these more detailed questions (Diamond, 2021).

The essential service of the Census Bureau is to maintain a count of every single household in the United States periodically, on a ten-year cycle. They also are responsible for the American Community Survey (ACS), the US Economic Census, and the Current Population Survey (US Census Bureau, 2021). They also conduct ongoing surveys about the people of the nation across a broad range of topics, including healthcare, housing, spending, crime, income, poverty, trade, and other programs. The US Census Bureau is not for profit, and does not collect fees for their services, and is funded from taxpayer dollars.

The census also provides an important role in allocation of federal funds, for things such as building schools, hospitals, transportation, and other public services. The data the Census Bureau collects also helps everyone from federal, tribal, and state officials to businesses and local communities, in making better informed decisions using their wide range of publicly available information. This data is also available for use by political parties, and for this reason, is at some substantial risk of political pressures.

Data.census.gov is the home of the publicly available data the census bureau generates. The site not only offers data across a broad range of topics from various surveys and studies, but gives access to data going back to 2000, and historical data even farther. Some data is still being migrated over from legacy systems. This service gives anyone access to this data, while providing mapping and other tools. The service has the ability to generate ‘microdata’ using information from multiple data sources not found in their pre-created tables. Some data types include the ACS, American Housing Survey, Census data, business statistics, foreign trade, map data (and various generation tools), federal spending, etc.

The US Census Bureau headquarters is located in Suitland, Maryland, since 1942, with data processing facilities in Arizona, Indiana, and Maryland (US Census Bureau, 2021). The Maryland headquarters hosts ~4200 full-time employees, with this number expanding over a hundred-fold during census taking years. The 2010 census employed 635,000 temporary employees as enumerators (US Census Bureau, 2021).

The US Census Bureau was chosen as it provides information about income data and housing statistics across the entire US. They also should have little to no bias when it comes to this information, as opposed to say, Zillow or other housing information aggregates. The benefit of a full census of the US population provides detailed information as well. It should be noted that income and housing information comes from the more targeted/less encompassing ACS survey, but this is still a large representative portion of the US population, and the housing information is encompassing. Since the organization does not directly apply or use the data themselves, the information is simply collected for statistical purposes and used by other agencies, businesses and individuals. They collect the information to get an accurate account and leave the use of the information to third parties.

**OBJECTIVES**

The overall question of this project is “has housing affordability changed from 2010 to 2019?” This period represents the recovery from the 2008 housing crisis. The project is also concerned with incomes during this period, but mainly as they relate to housing affordability and owner-occupied housing costs as a factor of income. People paying more than thirty percent of their income to housing and shelter are said to be “rent burdened”, although this applies to owner-occupied housing, as well. This number rose from less than one-fifth of owner-occupied homes in 1980 to 25-percent in 2015 (Molloy, 2020). Income and housing value will be compared over the study period to see if a significant change has occurred. The total number of owner-occupied housing as it relates to the US populations in 2010 and 2019 will also be examined. This will tell if homeownership as a factor of population is rising, falling, or remaining stable. The objectives are to see how incomes and housing values for owner-occupied households has changed from 2010 to 2019, as well as, how affordability has changed, if so.

**OVERVIEW OF STUDY**

This project is examining how housing affordability has changed from 2010 to 2019 in the United States. This can be broken up into several smaller questions. Did annual household income rise from 2010 to 2019? Looking at the median income change will give us a quick answer to this question but examining the changes in the different income groups will give greater insight to this question. Did US housing value increase from 2010 to 2019? Again, looking at the median difference will give us an overview, but the differences in the particular housing levels can tell us how housing ranges changed over the time span. If the median price increases, but a larger percentage of higher value homes were sold, this paints a different picture than the inverse. Has the ratio of value to annual household income increased from 2010 to 2019? This is a vital metric, and has been the standard methodology of housing affordability (Stone, 2006). Have monthly housing costs increased from 2010 to 2019? The median values will be compared here, as well as, a detailed look at the increases within each monthly housing cost group. This figure is very important, as it relates to the affordability of housing directly, incorporating more than just mortgage value to housing value. Have monthly housing costs increased as a percentage of annual income from 2010 to 2019? This information will give some of the greatest insights into changes in housing affordability over the time span. Comparing the changes in values at the different income groups, with particular note of those spending thirty percent or more of their income on housing costs, will show income levels where housing costs are taking larger portions of income, or not. Has the percentage of owner-occupied homes as a factor of population in the US increased from 2010 to 2019? This will require population data, looking at number of adults in the US in the time span, as well as source data.

Housing affordability can be measured many ways, but one standard measure is the “rule-of-thumb” ratio standard (the ratio of monthly income to monthly housing costs) of 25 percent pre 1980, to 30 percent since (Stone, 2006). This will be a key analysis within this project, specifically, how (if) the ratios have changed from 2010 to 2019, and what this means for housing affordability. The actual mortgaged owner-occupied housing rates near the median income level for all analysis years can be used to see if these ratios still hold true for who actually owns and lives in their homes. After the economic crisis of 2008, a great deal of homes were foreclosed and repossessed, dropping the values of homes considerably. However, many large investors used this opportunity to purchase large numbers of homes, having the advantage of liquid capital and the ability to buy at scale (Christophers, 2021). Home cost was lowest at 2012, and large investors took advantage of this opportunity to buy thousands of properties (many foreclosed) (Brotman, 2021). This changed who owned homes in the US. Comparing home ownership percentages at median income levels between 2010 and 2019 may glean more information into how large of a problem this actually is. How housing values and incomes have changed between these periods will also be an important focus of analysis. Outside of the scope of these datasets, something of interest to measure would be housing value (cost) and intrinsic worth. Looking at home size, age, design, number of rooms and other quality factors relating to price over this timespan could give information about whether housing quality is changing over time, as well as affordability. Examining multifamily occupied housing would also give additional valuable information, however this data was not available in the dataset.

The datasets for this project were obtained from data.census.gov at https://data.census.gov/cedsci/table?q=&t=Housing%20Value%20and%20Purchase%20Price%3AIncome%20%28Households,%20Families,%20Individuals%29&tid=ACSST5Y2010.S2506&hidePreview=true&moe=true. They comprise five-year ranges for 2010 and 2019. The datasets show information about mortgaged housing over these years, including number of homes, prices of homes, and mortgage status. They also include information about incomes of those owning the mortgages such as income ranges, percent of income dedicated to paying the monthly mortgage, and costs of monthly payments, among other information related to the housing or the individuals/families owning them. These datasets will help tell how housing has changed in the past decade, from how many houses are now owned (mortgaged), to how much those houses cost, both absolutely and as a factor of income. The year 2010 was chosen as the start year as this was shortly after/during the recovery from the 2008 housing crisis. This data is compared to the current market, where the housing values have been increasing rapidly in many areas of the United States.

**RESEARCH HYPOTHESIS**

For “did annual household income rise from 2010 to 2019?”, the alternative hypothesis will be: Household income increased from 2010 to 2019. The null hypothesis will be: household income did not increase from 2010 to 2019. The null hypothesis is inclusive of a decrease in annual household income. For “did housing value increase from 2010 to 2019?”, the alternative hypothesis will be: housing value increased from 2010 to 2019. The null hypothesis will be: housing value did not increase from 2010 to 2019. The null hypothesis is inclusive of a decrease in housing value.

For “has the ratio of value to annual household income increased from 2010 to 2019?”, the alternative hypothesis will be: the ratio of value to household income from 2010 to 2019 has increased. The null hypothesis will be: the ratio of value to household income from 2010 to 2019 has not increased. The null hypothesis is inclusive of a decrease in the ratios of value to household income from 2010 to 2019. For “have monthly housing costs increased from 2010 to 2019?”, the alternative hypothesis will be: monthly housing costs have increased from 2010 to 2019. The null hypothesis will be: monthly housing costs have not increased from 2010 to 2019. The null hypothesis is inclusive of a decrease in housing costs.

For “have monthly housing costs increased as a percentage of annual income from 2010 to 2019 for the >=30% level?”, the alternative hypothesis will be: monthly housing costs have increased as a percentage of income from 2010 to 2019. The null hypothesis will be: monthly housing costs as a percentage of income have not increased as a percentage of income. The null hypothesis is inclusive of a decrease in monthly housing costs as a percentage of income.

For “has the percentage of owner-occupied homes as a factor of population in the US increased from 2010 to 2019?”, the alternative hypothesis will be: the percentage of owner-occupied homes as a factor of population in the US has decreased from 2010 to 2019. The null hypothesis will be: the percentage of owner-occupied housing as a factor of population in the US has not decreased from 2010 to 2019. The null hypothesis is inclusive of an increase in the percentage of owner-occupied homes as a factor of population in the US from 2010 to 2019.

**LITERATURE REVIEW**

De Winter’s paper (De Winter, 2013) covers the use of the student’s t-test with regard to small sample sizes. Although the data in this paper is aggregated from a very high number of data points (the US census), it only covers five-year averages across two time points. It is used in review and consideration of statistical methodologies.

Molly’s paper (Molloy, 2020) discusses the effects of regulation on housing affordability. It gives examples of good and bad use of regulation. By examining the effects of regulation, or lack thereof, it looks at the influence of government at regulating housing affordability. It also gives specific statistics related to affordability in the United States. It is primarily used for reference, and statistical values.

Stone (Stone, 2006) discusses housing affordability in their paper. The paper consists of ways to measure affordability, such as the residual income approach, which will not be used in this discussion. He does discuss the standard approach of using monthly housing costs as a reflection of income on housing affordability. This approach is used within this paper. He also presents other methods of examining affordability and goes into the pros and cons of these measures, as well as how to obtain them.

O’Rourke’s paper (O'Rourke, 2005) is a ‘how-to’ guide on conducting statistical tests within SAS. This paper was used to consider different analytical techniques for the methodology section. It discusses the pros and cons of various statistical tests. It compares these tests in their requirements for use, their methodology, and their implementation. Specifically with regard within the SAS software. It also provides some methodology, “step-by-step”, which may or may not be used.

Davis’ work here (Davis, 2012) is a short textbook about the ethics of big data usage. It covers case studies of ethical use as well as privacy and security concerns. It offers brief case studies as well as examples of how to examine ethics, security, and privacy. It’s example of the four requirements of ethical consideration for data analytics projects is used as a format for consideration in the ethics section of this paper, as well as one of its case studies.

**RESEARCH DESIGN**

**Methods and Methodology**

Excel will be used to restructure the datasets. This will be done both to combine the datasets for direct comparison, as well as to reorganize the data for cleaner importation into other programs (R, SAS, etc.). The data in its current form (US Census Bureau, 2021) would be very difficult to work with in R and SAS, in its current tabular format. For many of the research questions, the data is separated into discrete groups, so comparing information in histograms or pie charts would be useful for initial analysis. This is simply done in either R or SAS. I intend to primarily use SAS/SAS Enterprise and R to analyze this data. SAS and Tableau will also be used for visualization purposes. R’s ability to filter data and create new tables more focused on specific aspects of the dataset should be useful, although may not be necessary with some work with the dataset in Excel, especially in ‘de-tiering’ the variables which are separated into categories. For example, monthly housing costs as a percentage of household income is set up in two tiers with the data mixed among direct lines. I would like to look at this data from the datasets, as well as computing this information directly from income and housing costs, converting annual income to monthly income.

Heat maps and perhaps bubble plots could be great tools to show some of the data. Especially for information related to grouping the income levels against the housing data due to the multi-tiered effects (income vs housing value with bubbles as percentages of owner-occupied housing populations might show a lot of information, for example). These can be generated in SAS or R, although Tableau may be used if the figures ‘tell a better story’ or are easier to understand there. Pie charts and histograms will also be useful ways to demonstrate the data overviews. Multiple line graphs showing the changes in data values from 2010 to 2019 (for example for monthly housing cost changes) may also prove useful visually, if the values are distinct enough.

For the majority of the variables, direct comparison between the 2010 and 2019 sets is what is being tested. Although the data is collected from massive figures of housing and income data, in this study often the n-count will only be two. Student’s t-test would make a simple comparison of these values, however with the N value so small, even though the data is collected from such a high sample size, this test has difficultly showing the significance of variation (De Winter, 2013). The paired samples t-test will be much more applicable in this case. One-way ANOVA or simple multivariate MLR may be more appropriate for the more complex grouped items, such as income split into levels, or monthly housing costs as a percentage of annual income, however the minimal points still make the paired t-test ideal (Kim, 2015). The major downside to the paired t-test with this number of datapoints is the need for a much higher t-value to attain statistical significance (O’Rourke, 2005).

Standard deviation was required to be calculated by hand using sample size data, z scores, and margins of error available in the data set, as well as from other documentation available on the US Census Bureau website (US Census Bureau, 2021). This was necessary to find the significance of the differences between values. The general formula used was MOE=((z\*std.dev)/√n). The standard deviation for most measures in 2010 was 0.842, and in 2019 it was 0.872.

**Limitations**

This study may not give significant results with respect to the hypothesis testing. This is entirely possible, and has meaning; however, the limited number of data points (two five-year aggregated points) may play into this. Regression of the non-aggregated data may be necessary in this case. With regard to measuring number of owner-occupied homes over the study period to the total US population, there is a caveat that it is possible that the population is choosing to live in other conditions such as renting, for various reasons, and this could contribute to a dip in owner-occupied housing as a measure of total US population. This question is outside of the scope of this study, however.

**Ethical, Privacy, and Security Considerations**

For any data project, there are concerns about data privacy, ethical use, as well as security. Responsible use of big data can give incredible insights into business problems resulting in increased profits, new markets, among many other benefits. Irresponsible use can result in data breaches, loss of information, or even legal or other punitive impacts. This can be caused by ethical misuses, such as collecting private information without transparency, such as Apple regularly saving the position of iPhones to a hidden file (Davis, 2012). This resulted in some serious negative press, especially as the data was not being stored in a secure fashion, so there were security as well as privacy and ethical concerns. Davis identifies four common elements for big data ethics: identity, privacy, ownership, and reputation (Davis, 2012). These are defined as our offline and online identity relationships, who controls access to data, who owns/gives access to data, and the trustworthiness of data. For this project, the security, privacy, and ethical concerns relate to how the US Census Department interacts with, stores, and gathers their data, as well as how the data is used and presented within the project and presentations.

Identity of our online selves is a topic of great controversy with data use. Companies aggregate large amounts of data about people, including shopping habits, spending, and more sensitive things such as location, demographics, and financials. Users often do not know how their data is being used, and often this use benefits corporations to the detriment of users. Recent changes encourage better transparency and disclosures of data use, but these documents are often legally dense and difficult to parse for the average consumer, not to mention inconvenient. For example, new changes in the US require permission to collect “cookies”, but disabling these for every site you visit is tedious, to say the least. Ownership of data today is rarely tied to the user whose data is being collected. In most situations, user data is collected and not only used by the first party, but often sold to advertisers and other parties. The user does not receive a portion of the sale of this data, and in many cases is not able to opt out without being unable to use the software or service. One concern is the benefit of practices to the consumer relative to the benefits of the controller. Unfair practices were relabeled from “deceptive” to “unfair” by the FTC, and these are described as injury to customers which is not reasonably avoidable by consumers or outweighed by benefits to consumers (Khanan, 2019). Reputation is tied to the parties using the data. Using data poorly reflects badly on the company, such as the Apple case in the above paragraph. Conversely, using data responsibly attracts customers and builds trust.

For the analysis of the data in this dataset, the ethical considerations mainly point to treating the data without bias. False interpretation of the results would give an untrue representation of the status of the US housing markets, and how they relate to incomes and housing affordability. The data needs to be treated without bias and presented in a way that is not misleading. This applied to analysis methodology, as well as to how the results are presented. It is easy to manipulate a figure to tell a false or half-story, for example. The results need to be encompassing, as well as truthful.

Today, massive amounts of information is collected every moment for the purposes of big data collection and use. Big data can be massively beneficial to businesses, governments, and society at large with the use of analytic tools and techniques. This data is in a wide variety of formats, from text, video, picture, location, etc. This data is sourced from smart phones, computers, social networks, IoT (Internet of Things), space-based sensors, etc. The collection of this data has a large number of security and privacy issues and challenges. Interestingly, in Europe, data processing is often prohibited by law unless a legitimate legal basis is shown (Khanan, 2019). This is inverse of US common practices. Security of data includes physical and digital security, data integrity, and usage. Data access should be limited to few users with passwords or multi-factor digital security, as well as physical security to degrees able. Data integrity needs to be clean, (error and gap-free), without bias, and consistent and precise (Kitchin, 2016). Physical security includes things like physically protected servers, limited access to computers, etc. Digital security encompasses access limits, firewalls, digital security teams, and clean coding that is not open to exploitation.

For the analysis of the data in this dataset, relevant security concerns include protecting the data integrity, limiting access to the data, and presenting the project and results without PII (personally identifiable information). Protecting the data integrity will be important, as the dataset will need to be reorganized for analysis, and ensuring no data is lost/overwritten, etc. is vital. The data will be compared manually via excel to ensure no data loss. The data will only be accessed by myself, although the data is publicly available via the US Census Bureau (US Census Bureau, 2021). PII and other identifiable information has already been removed from the dataset by the US Census Bureau before data aggregation.

Privacy in data encompasses who we are, both in reality and online, what information about us is viewable and shareable, and how much control over that data exists. It is also related to security, especially when data can be PII but more importantly, how secure personal data is stored, protected, and accessed. One of the great features (and occasionally a problem) of the internet is the expectation of anonymity. This is of high importance in not only being secure in your online actions, but for personal safety at times. For example, being able to criticize your government without being identifiable is vital to having the voices of the people heard, especially in societies where criticism can be met with punishment.

There must exist a balance between privacy risks and the benefits of big data. These balances include the benefits of using patient information for medical research versus privacy and ethical concerns using that information, policing and protecting the internet and public at large versus personal freedoms, etc. Medical use has stringent standards for use of private information, while government surveillance of the internet is a constantly changing landscape of legislation and debate. Large concerns of the loss of privacy with regard to big data include profiling, discrimination, criminalization, and security (Polonetsky, 2013). Specific challenges include transparency of private information use, security of PII and other private information, cleaning datasets of reverse-identifiable information as well as directly identifiable information, and when to collect personal information. These concerns are especially important for data such as name, location, financial information, or demographics.

For the analysis of the data in this dataset, many of the privacy concerns have already been addressed. The data is aggregated from a vast information store, which does include private information such as name, address, and financial information such as income and wealth. However, this PII was removed from the dataset before it was released online, and would be nearly impossible to reverse engineer from the dataset itself, nor from the analysis within. The data is all publicly available and collected ethically as per the regulations and legislation applied to the US Census Bureau.

As stated above, the visual analysis and presentation of the data is important for ethical considerations. The security concerns stem from using the data responsibly and not sharing the data with outside parties, but again, the data is freely accessible via the US Census Bureau. Integrity of the data as it relates to the visual analysis and presentation is important. Data should not be excluded because it does not tell the story the author desires, and should not be missed for precision and accuracy reasons. The privacy of the data here does not carry considerations related to visual analytics. The data is available freely, and all personally identifiable information has been scrubbed from the data. With regard to audience, this paper will likely only be seen by instructors and students of Colorado State University Global. If used in other locations, ensuring that the data and results are easy to parse is important for other audiences. However, this should be done within the scope of the project regardless.

**STUDY TO RESULTS**

Using this methodology and techniques, this project should give us insight into how income levels, housing values, and housing affordability has changed from 2010 to 2019, and whether the US is making a recovery in terms of owner-occupied home ownership and housing affordability. If it is not, considerations for what to do about the problem should definitely be a focus of the current and next administrations. It is possible that there is no change to housing affordability. In such a case, where incomes are rising with home value and affordability is remaining static, or improving, this is wonderful news. Comparing total number of owner-occupied homes to the change in the US population will also give us insight into whether the housing market is keeping up with supply, with the limitations expressed in the limitations section above. Hopefully the knowledge gained from this study can give us insight into the study questions. With this knowledge, we can decide whether changes in policy or development should be considered.

**FINDINGS**

**Income**

Figure 1

*Income Paired t-test, SAS*

**Chart, line chart, histogram

Description automatically generated**

Figure 2

*Income Levels, 2010 and 2019, SAS*

Chart, histogram

Description automatically generated *Chart, bar chart, histogram

Description automatically generated*

From our p value of 0.9940, we fail to reject the null hypothesis. We cannot prove that income levels grew significantly from 2010 to 2019. The median income did grow by ~$17,000, but most of that growth was driven by growth in the very highest income levels, as seen in figure two, with the very large growth in the $150,000 or more category. One difficulty in the data was testing for significance with the aggregated data. If, instead of percentages, the data was simply the raw incomes, more accurate testing for significance could have been done. We can still see here that incomes in the lowest levels shrank somewhat, with an upward trend, however, we cannot prove that this result is significant with the data as it is. The paired t-test in figure one is testing whether there was a significant enough change in the distribution of the income levels.

**Housing Value**

Figure 3

*Housing Value Paired t-test, SAS*

Chart, histogram

Description automatically generated

Figure 4

*Housing Values, 2010 and 2019, SAS*

*Chart

Description automatically generated* *Chart

Description automatically generated*

From our p value of 0.9864, we fail to reject the null hypothesis. We cannot prove that housing values grew significantly from 2010 to 2019. The median income did grow by ~$30,000, from ~$210k to ~$239k, however. Looking at figure four, small increases in mid-range housing occurred, as well as a larger increase in the number of homes at the highest end of the scale. The difficulty in testing for significance with the aggregated data exists here as well. The overall changes in housing value appear to be mostly related to the large increase in the highest levels of housing. Still, we cannot prove that this result is significant with the data as it is. The paired t-test in figure three is testing whether there was a significant enough change in the distribution of the housing values.

**Annual Income to Monthly Housing Cost Ratio**

Figure 5

*Annual Income to Monthly Housing Cost Ratio, SAS*

*Chart, histogram

Description automatically generated*

Figure 6

*Annual Income to Housing Cost Ratios, SAS*

*Chart, pie chart

Description automatically generated* *Chart, pie chart

Description automatically generated*

From our p value of 1.000, we fail to reject the null hypothesis. This is a very high p-value. We cannot prove that the ratio of annual income to monthly housing costs significantly increased from 2010 to 2019. In fact, it appeared to remain quite stable. Looking at figure six, the changes that did occur were a decrease in those paying a ratio of four or more, and an increase in those paying less than a ratio of two. This is a positive sign, and shows that the US population as a whole is generally paying less of their income for housing, with respect to that these figures only represent owner-occupied homes, and not renters, etc. The paired t-test in figure five is testing whether there was a significant enough change in the distribution of the ratio of annual income to monthly housing costs in the US. If anything, the ratios of housing costs to income appear to have decreased somewhat, at least at the highest end.

**Monthly Housing Costs**

Figure 7

*Monthly Housing Cost Paired t-test, SAS*

*Chart, histogram

Description automatically generated*Figure 8

*Monthly Housing Costs, SAS*

*Chart, pie chart

Description automatically generated* *Chart, pie chart

Description automatically generated*

From our p value of 1.000, we fail to reject the null hypothesis. This is another very high p-value. We cannot prove that monthly housing costs significantly increased from 2010 to 2019. Looking at figure eight, most categories remained stable. A note on figure eight, the ‘other’ category is inclusive of all monthly housing costs below $600. This, and the $600-$799 category were the only categories with significant change. Both decreased over the study period. This could indicate that monthly housing costs are rising for lower priced housing, or that there are fewer low-priced homes available, increasing mortgage costs. The paired t-test in figure seven is testing whether there was a significant enough change in the distribution of monthly housing costs in the US. With the exception of the lowest monthly housing cost categories, monthly housing costs appear fairly stable for owner-occupied homes. More granular data would be helpful to understand what is fully occurring here. Looking back at figure four, the percentage of homes below $50,000 decreased; this could definitely contribute to the decrease in the percentage of the lowest monthly housing costs.

**Monthly Housing Costs as a Percentage of Income, 30% or Greater**

Figure 9

*Monthly Housing Costs as a Percentage of Income, 30% or Greater Paired t-test, SAS*

*Chart, histogram

Description automatically generated*

Figure 10

*Monthly Housing Costs as a Percentage of Income, 30% or Greater, SAS*

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

From our p value of 0.0003 in figure nine, we can reject the null hypothesis. This is a very low p-value. The paired t-test in figure nine is testing whether there was a significant enough change in the distribution of monthly housing costs as a percentage of income in the US, with specific regard to the >=30% level. The >=30% level was looked at specifically due to its use in determining when housing begins to become unaffordable, with small crisis potentially leading to failure to pay a mortgage (Stone, 2006). Comparing the 2010 and 2019 graphs in figure ten, we can clearly see the decrease in those paying thirty percent or more of their incomes on their monthly housing costs in every single income category. This is very positive news for housing affordability in the US. With fewer people near or at the thirty percent limit, fewer failures to pay mortgages should occur. This could be due to differences in income levels, lending practice changes that occurred after the housing crisis, or other factors. This does come with a caveat however, specifically if this change is due to fewer people owning and occupying their own homes, which will be discussed in the next section. The one category that did not change significantly was the <$20k level.

**Owner-Occupied Housing as a Factor of US Adult Population**

Figure 11

*Owner Occupied Homes in the US as a Factor of Adult Population, 2010-2019, SAS*

*Chart, line chart

Description automatically generatedGraphical user interface, text

Description automatically generated* This section examines the changes in US population, the total number of owner-occupied homes in the US, and the relation of these two values over the time period. Statistical tests for the relation were a simple regression, with an R value of 1.0, which is highly significant. However, due to it being a simple comparison of total population, and owner-occupied homes in the US, with two data points, this cannot be used to reject the null hypothesis. To be tested more accurately, breaking down the five-year aggregations into individual years for more data points would make a simple linear regression easily possible. That being said, there are concrete differences in the data points between these two time frames.

Owner-occupied homes decreased by 3,280,214. At the same time, the US adult population increased by 19,493,985 people, giving shifts in the percentage of owner-occupied homes in the US as a function of adult population from 16.7% in 2010, to 14.8% in 2019, or a total loss of approximately two percent. This is a highly concerning finding. The change in population was an increase of about six percent, conversely the change in owner-occupied homes was a decline of six percent. If the population is increasing, we might expect the number of owner-occupied homes to follow the same trend. The fact that it tracked inversely could point to an increased difficulty in being able to afford or obtain a mortgage, lack of availability of houses for sale at affordable rates, a very large increase in rental population, or other factors. The US adult population data was found in the US population characteristics files from the US Census Bureau (US Census Bureau, 2021).

**CONCLUSION**

The majority of null hypothesis were unable to be rejected, with the exception of “have monthly housing costs increased as a percentage of annual income from 2010 to 2019 for the >=30% level?”. As stated, this percentage of income being spent on housing is a common dividing line to distinguish when housing begins to become unaffordable. Since this decreased for every category but the lowest incomes (<$20k), we can say that owner-occupied homes are becoming more affordable, at least for those who live in them. It does not answer the question of whether US citizens are more capable of owning a home, however. That question would ideally be answered with information about income levels, housing values, and housing market prices combined with geographical information. This would make for an enlightening study, however a complex one. Simply being able to afford a house does not mean that more people will purchase one. Complicated factors such as location, availability of careers or schools near housing, age, etc. can all influence housing purchasing decisions, and this is outside of the scope of the study.

Some factors that led to the majority of null hypothesis being failed to be rejected include the limitations of the dataset. The data for individual years is available using the same source, and, in hindsight, should have been used instead of the aggregated data. With additional measures at ten timepoints, the data would have been much easier to work with, and would enable the use of a number of statistical methods that could not be used with the dataset as it is. This is unfortunate, as visually, many of the hypothesis seem correct in the data, but the data was not robust enough to prove significance for most of the hypothesis.

The decrease in those spending thirty percent or more of their incomes on housing who live in the homes that they own is a positive sign for housing affordability in the US. Again, we cannot know from this study whether people are unable to afford homes and are stuck renting. The population increase and converse decrease in the number of owner-occupied homes seen in figure eleven may point to this being the case. Whatever the cause, a decrease in owner-occupied homes combined with an increase in population is cause for great concern.

**RECOMMENDATIONS**

Repeating this study using data from individual years from 2010 to 2019 instead of the five-year aggregated data would be prudent. This would likely give more insight into whether all hypotheses were truly able to be rejected or not. This would also allow for a number of statistical tests for significance that were not able to be performed on this data and examining changes to the variables over time. Hopefully, this would conclusively prove or disprove hypothesis without the need for additional information.

Instead of using summarized data, the raw Census data could also be used in a similar manner. The aggregated data is robust and trustworthy, but simply does not have enough datapoints for many types of statistical analysis. The raw data would include millions of points of information, such as actual income levels, housing values, etc. that would enable the examination of variable as continuous ones, instead of the discrete categories used in this study. This has the advantage of flexibility of analysis, and can tell many stories that we cannot see with this aggregated information. That is not to say that the information gleaned here is not valid or without value and potential insights, however.

That being said, a study of housing affordability inclusive of renters, owner-occupied housing, shared housing, and other housing types would be recommended to truly understand housing affordability in the US. The large decrease in owner-occupied homes coupled with the increase in the US adult population signifies that, for some reason, people are choosing to, or are not able to own and live in their own homes, a classic tenant of the ‘American Dream’.

**GITHUB LINK**

Documents, Excel spreadsheets, csv files, SAS files and code found here under the code tab:

https://github.com/slblamer/Capstone-581-2/projects

**REFERENCES**

Anna Diamond. (2021). *The Enumerated Story of the Census*. Smithsonian.

https://www.smithsonianmag.com/history/enumerated-story-census-180974648/

Census.gov. (2021). *Legislation.*

https://www.census.gov/history/www/reference/legislation/

Davis, K, & Patterson, D. (2012). *Ethics of Big Data.* O’Reilly Media, Inc.

De Winter, J. C. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation.* *18*(1), 10. <https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1307&context=pare>

Khanan, A., Abdullah, S., Mohamed, A. H. H., Mehmood, A., & Ariffin, K. A. Z. (2019). Big data security and privacy concerns: A review. *smart technologies and innovation for a sustainable future*, 55-61. http://cyberlaw.stanford.edu/files/publication/files/PolonetskyTene.pdf

Kitchin, R. (2016). Big data. *International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology*, 1-3. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118786352.wbieg0145

Molloy, R. (2020). The effect of housing supply regulation on housing affordability: A review. *Regional science and urban economics, 80*(C). https://www.aei.org/wp-content/uploads/2017/04/Overview-Talk-Panel-5.pdf

O'Rourke, N., Hatcher, L., & Stepanski, E. J. (2005). *A step-by-step approach to using SAS for univariate & multivariate statistics*. SAS institute. https://books.google.com/books?hl=en&lr=&id=pfUfzykTZ1AC&oi=fnd&pg=PP1&dq=multiple+paired+t+tests+at+same+time+sas&ots=8CVKvT9Lsc&sig=lh-R5-vQXgQHwWvOxFHd5IUONlc#v=onepage&q=paired%20t%20tests&f=false

Polonetsky, J., & Tene, O. (2013). Privacy and big data: making ends meet. *Stan. L. Rev. Online*, *66*, 25. http://cyberlaw.stanford.edu/files/publication/files/PolonetskyTene.pdf

Stone, M. E. (2006). What is housing affordability? The case for the residual income approach. *Housing policy debate*, *17*(1), 151-184. https://works.bepress.com/michael\_stone/5/

US Census Bureau. (2021). *Financial Characteristics for Housing Units with a Mortgage, 2010 and 2019 5 year estimates* https://data.census.gov/cedsci/table?q=&t=Housing%20Value%20and%20Purchase%20Price%3AIncome%20%28Households,%20Families,%20Individuals%29&tid=ACSST5Y2010.S2506&hidePreview=true&moe=true

US Census Bureau. (2021). *Sample Size.* https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/sample-size/index.php

US Census Bureau. (2021). *National Population by Characteristics.* <https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html>

US Census Bureau. (2021). *List of all Surveys and Programs.*

<https://www.census.gov/programs-surveys/surveys-programs.html>

US Census Bureau. (2021). *Location*

https://www.census.gov/about/census-careers/opportunities/location.html