Data 230 Term Project

Can People in the US Afford Homes Today?

Athena Hung 015261246

San Jose State University

Professor Andrew Bond

May 23, 2021

Abstract

This project will discuss the issue of "if people in the US can afford homes today?" utilizing all the techniques and tools learned in DATA 230. First of all, necessary data will be collected online and will be cleansed as well as transformed into desired format. A story and an interactive dashboard will be created in Tableau with all detailed information.

Housing price in general in the US will be discussed first, and we focus on California for further analysis. Artheton in San Mateo county is the city that is the most expensive city to live in. Salary data is brought in next and we find out the fact that the Bay Area has much higher average salary than other counties in California. By utilizing table calculation, we initially find that people need about 9.4 years to buy a median price housing with an average salary. However, considered the fact that people are not able to save the entire salary and spend them on buying houses, we take the saving rate into consideration. In the end, we realize that people in California need about a hundred years in order to buy a house with average salary. Residents in Kings county would need 150 years to afford homes just by their wages which is impossible. Conclusions and some future works are also discussed at the end of this project report.

Introduction and Background

The housing price in the US has been increasing for a long time and people are always complaining that it is difficult or impossible to buy a home in the US. Without owning a home, people might need to move around frequently as increased rent or landlord issues. As a result of that, I would like to discuss more about this question for this project and allow the data to talk for itself. This project utilized all the techniques and tools taught in DATA 230. A final project

report, a presentation to the class with PowerPoint slides, a Tableau file, and all the dataset will be submitted for this term project as the project deliverable.

Project Details

First of all, I collect all the necessary data from Zillow (housing price data), Statista (personal saving rate data), and Bureau of Labor Statistics (BLS) from U.S. Department of Labor (salary data). I chose them as my data sources as they have higher accountability and credibility so data quality is high with higher accuracy and completeness during data collection. Zillow is an American online real estate marketplace company and it is one of the biggest in the real estate industry. Housing price data was collected from it and it contains the typical value of single family houses for the past 25 years. Only data in the 35^{th} to 65^{th} percentile range was selected to avoid outliers. Statista is a database company specializing in market and consumer data and all contents on the Statista platform need to pass a multi-stage peer-review process prior to publication. As a result of that, this source is reliable can be concluded. Personal saving rate data was collected from it and saved for future use. Salary data was collected from Bureau of Labor Statistics (BLS) from U.S. Department of Labor and there is no doubt that this data source is official and legit.

Some data pre-processing was performed after the data collection including data cleaning and data transformation in Excel and Tableau. Sample data of housing price data can be found in Figure 1. First of all, I removed some unnecessary columns. For example, StateName and State are basically the same so one of them is taken off to avoid confusion. RegionID does not mean too much for users like me so this column is also removed. Some of the columns were renamed

for better analysis, such as renaming RegionName to ZipCode and deleting RegionType so it is more straightforward. Pivot table was also utilized to convert the dates from columns to rows so Tableau can graph from it easily.

RegionID	SizeRank	RegionName	RegionType	StateName	State	City	Metro	CountyName	1/31/1996	2/29/1996	3/31/1996	4/30/1996	5/31/1996	6/30/1996	7/31/1996
84654	1	60657	Zip	IL	IL	Chicago	Chicago-Napervil	Cook County	381952	381188	380605	379933	378569	378236	377525
91982	3	77494	Zip	TX	TX	Katy	Houston-The Wo	Harris County	201351	201599	201402	200208	199065	198397	199164
84616	4	60614	Zip	IL	IL	Chicago	Chicago-Napervil	Cook County	588901	588436	587596	587531	585026	585045	584287
91940	5	77449	Zip	TX	TX	Katy	Houston-The Wo	Harris County	98062	98054	98012	98032	98027	97998	97954
91733	7	77084	Zip	TX	TX	Houston	Houston-The Wo	Harris County	98634	98658	98583	98575	98533	98513	98384
93144	8	79936	Zip	TX	TX	El Paso	El Paso	El Paso Count	82792	82747	82717	82679	82761	82869	83000
84640	9	60640	Zip	IL	IL	Chicago	Chicago-Napervil	Cook County	276783	275027	273606	270865	268413	267431	266840
62037	10	11226	Zip	NY	NY	New York	New York-Newarl	Kings County	216705	215555	214479	213132	212683	213406	214180
61807	11	10467	Zip	NY	NY	New York	New York-Newarl	Bronx County	232116	231409	231121	229903	229433	228807	229208
92593	12	78660	Zip	TX	TX	Pflugervill	Austin-Round Ro	Travis County	153861	153690	153583	153482	153517	153617	153638
97564	13	94109	Zip	CA	CA	San Franci	San Francisco-Oa	San Francisco	555467	560196	563528	567020	567530	562148	559092
74101	16	37013	Zip	TN	TN	Nashville	Nashville-Davids	Davidson Cou	109978	110425	110807	111598	112415	113254	114115
71831	17	32162	Zip	FL	FL	The Villag	The Villages	Sumter Count	108777	109683	109843	110560	110597	110527	110497
84646	18	60647	Zip	IL	IL	Chicago	Chicago-Napervil	Cook County	181477	181990	182515	183362	183428	183787	184563

Figure 1. Sample housing price data

A story with all the graphs and information as well as an interactive dashboard were made for this project. A Tableau file with all the graphs, story and dashboard will be attached along with this report. This report will go deeper into the story behind each graphs with all the details.

Figure 2 is the line graph of US housing price for the past 25 years. It has been increasing fairly quickly and this is also the main thing that interested ne to look into issue of "if people in the US still can afford homes today". Even though the price dropped during the financial crisis in 2008, it has increased from 120k to 300k which means today's housing price is 2.5 times more than the housing price 25 years ago.

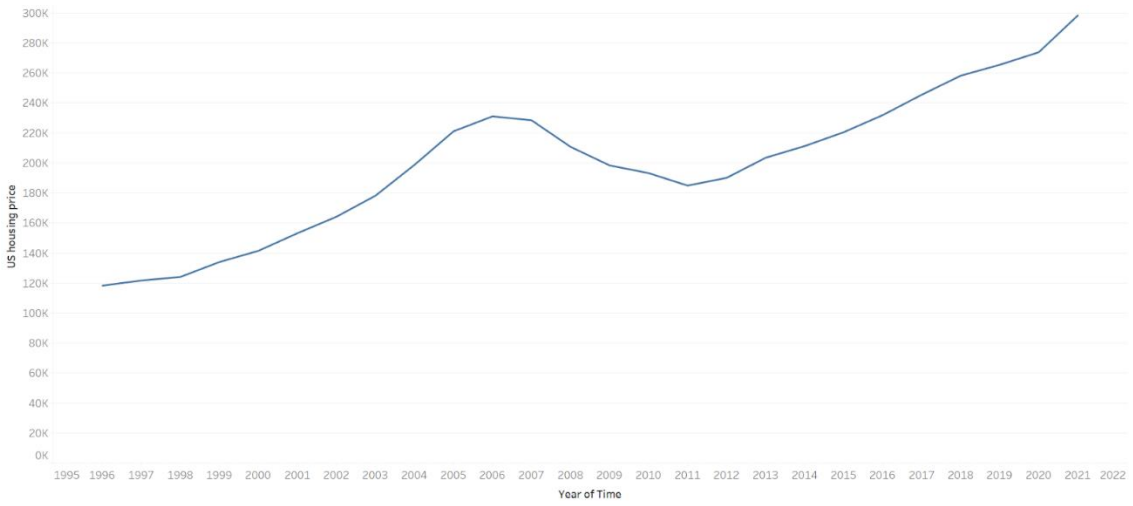

Figure 2. US housing

If we had to guess, there would be many people guess that California and New York have the most expensive housing. Figure 3 is the graph of total housing price for each state, and we can see that California and New York are the only two states that are not in the green zone. However, this is only a myth and not true. The reason why only these two states are red or yellow is just because they have larger population and more houses and it does not really mean that residents there are paying more than other states.

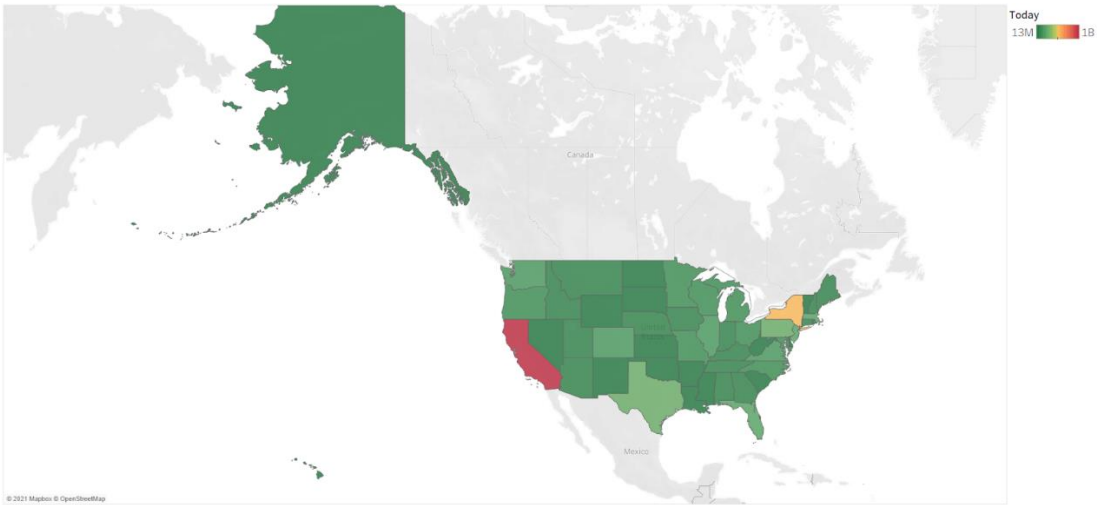

Figure 3. Total housing price for each state

Figure 4 shows median housing price for each state. Unlike the previous graph, now it seems like Hawaii has the highest housing price as it seems like it is the only state that has color of red. Other states are all in orange, yellow or green. The reason why houses in Hawaii are more expensive might be there are a lot of vacation houses that are more big and luxury like beach houses.

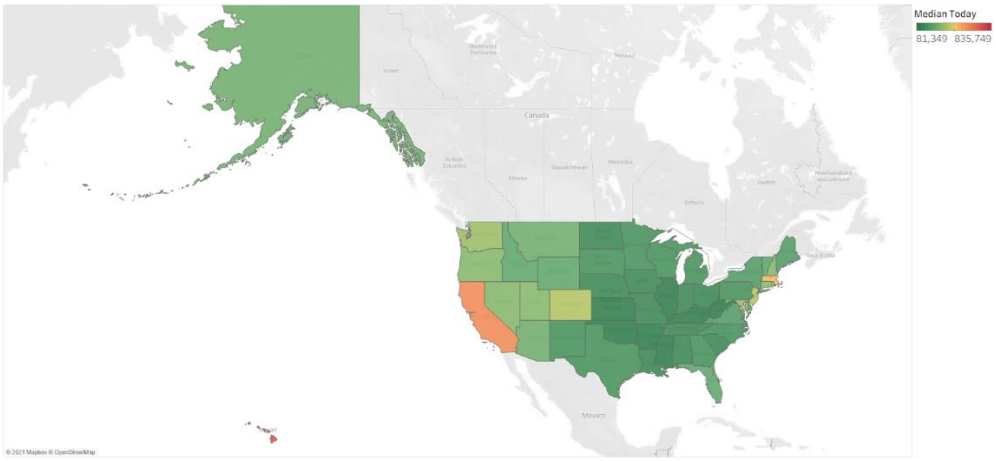

Figure 4. Median housing price for each state

We though Hawaii is the state with the most expensive housing, but in fact, Washington DC is the most expensive state for housing. We can see this from Figure 5 that it is four times more than the average of all of the states.

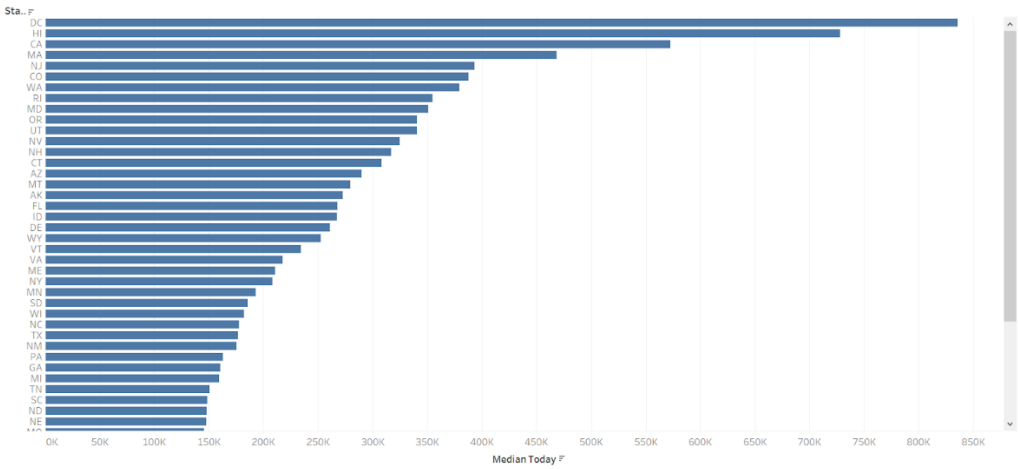

Figure 5. Median housing price histogram

We can see more clearly from Figure 6. We did not notice it because the red of Washington DC was too small and it blended in the greens around it. I think the reason of this is because Washington DC does not really have rural area unlike other states.

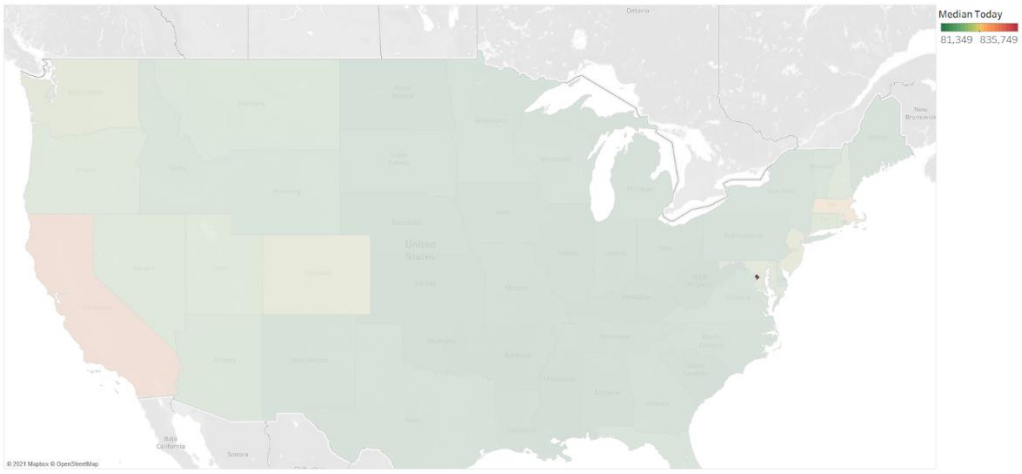

Figure 6. Highlight Washington DC

It is more obvious in Figure 7. The size of each square is the number of houses, i.e. California has the most number of houses, and the color is the median housing price. We can see clearly that Washington DC does not have a lot of houses and chances are that most of them are really expensive. Median price is 835k and average price is about 900k in Washington DC.

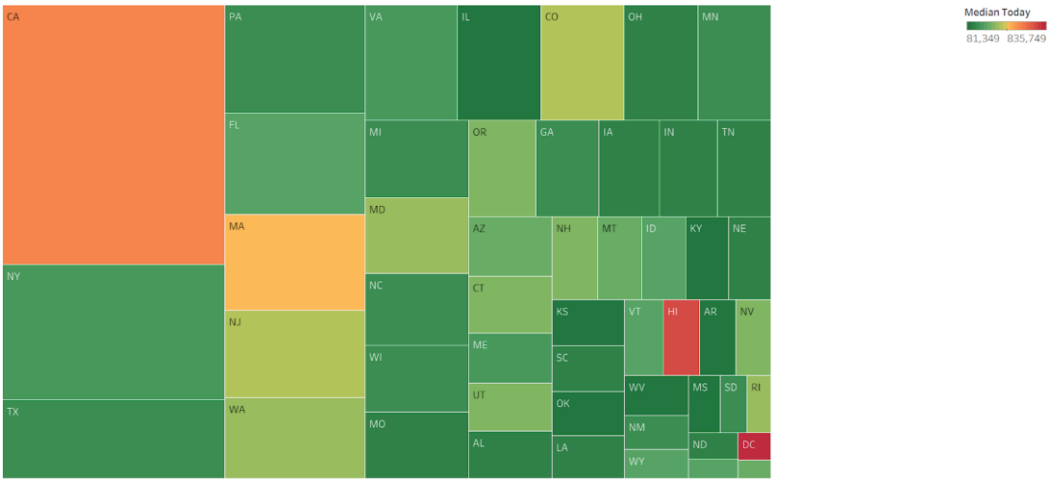

Figure 7. Median housing price vs Total number of houses

I wanted to focus more on California and look into more detail in it as it has the most housing numbers. Figure 8 is the median housing price by each zip code and cities in California. The sizes and colors of the circles are both median housing prices in that city. The bigger the circle is, the median housing price of that city is more expensive. It is also more expensive when the color of the circle is more towards to red than green.

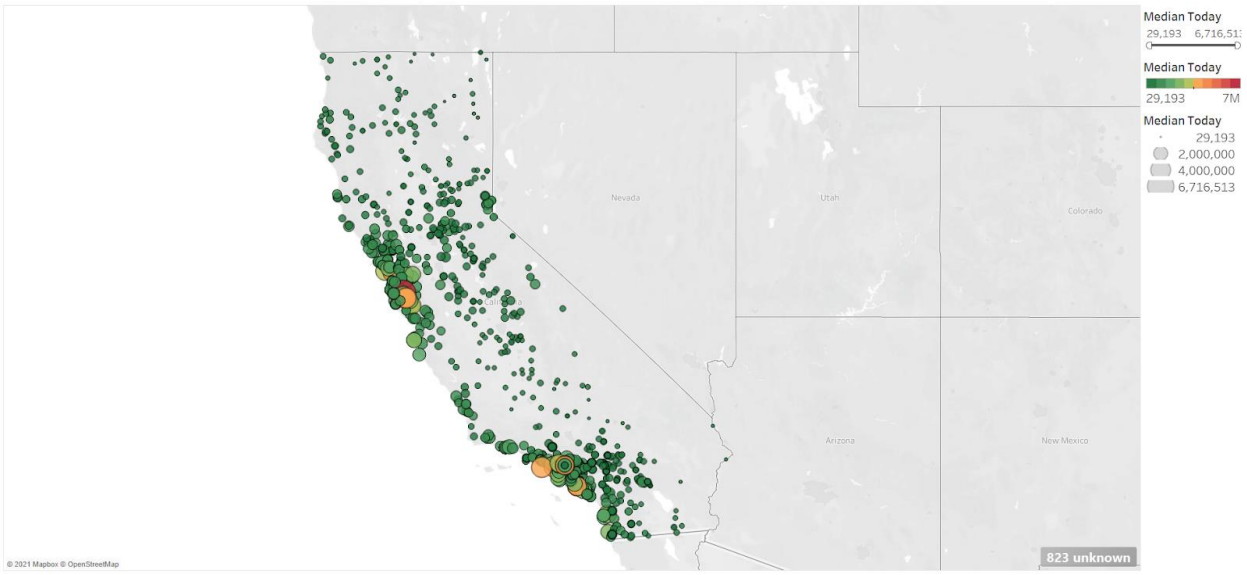

Figure 8. Median housing price by Zip code

Just like we mentioned before, housing price in California is also increasing in a fast pace for the past 25 years. The only difference is that the average housing price across the nation increased from 120k to 300k, California actually increases from 150k to 737k which is about five times more than 25 years ago.

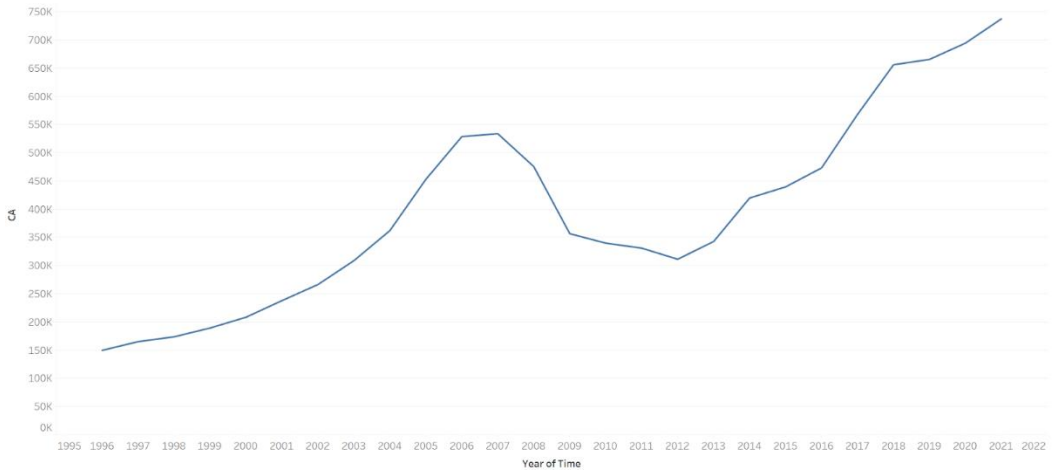

Figure 9. CA housing

These are the cities that median housing is higher than a million dollars in Figure 10. We can see that most of them are concentrated in the Bay Area and Greater Los Angeles. In Los Angeles, there are no red ones, and there are only oranges and yellows. As a result of that I will look into a little more detail into Bay Area.

Figure 10. Median housing price more than 1M

Figure 11 is houses cost more than a million dollars in the Bay Area. We can notice that there is actually only one red circle in this figure.

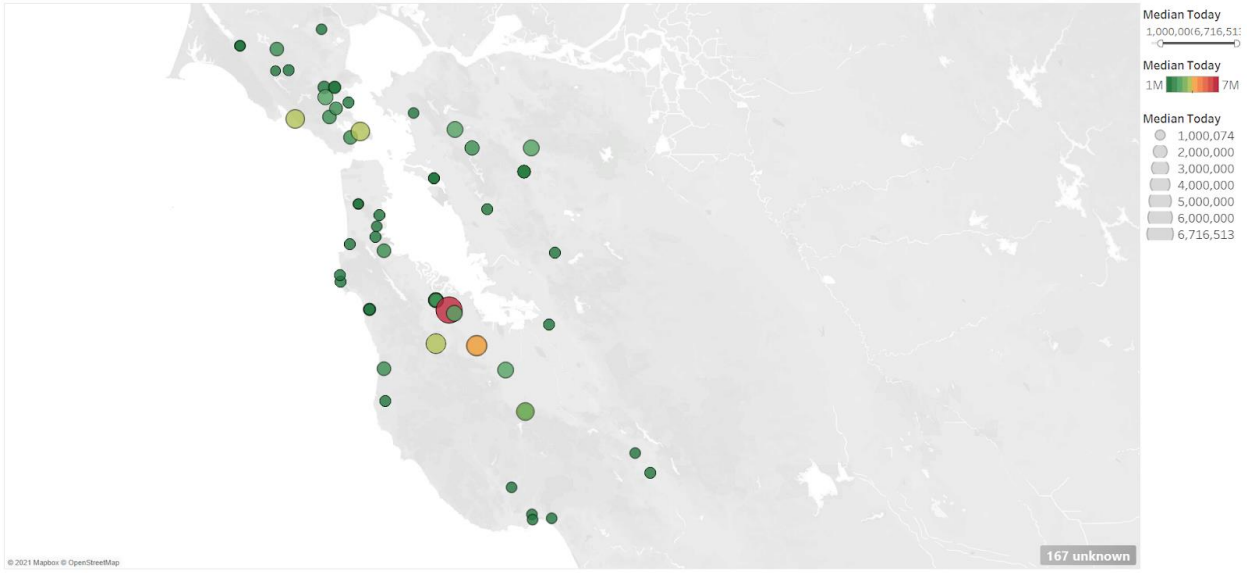

Figure 11. Median housing price more than 1M in the Bay Area

The red circle we noticed from the previous graph represents Artheton in San Mateo county, and it is also the most expensive city for housing in the entire California. the median price of Artheton is 6.7 million dollars. So here comes the question, are we able to afford it? Can people still afford homes today?

Figure 12. Artheton is the most expensive city for housing in CA

Figure 13 is the average weekly wage from BLS in every county in California. For average weekly wage from high to low is set to red to green so it follows the same pattern as the previous graphs to avoid confusion. We can see that the counties in the bay area, Santa Clara, San Mateo, and San Francisco are the ones that have the highest salary among all the counties in California, and all other counties are all in the green zone which means the average weekly wages there are lower.

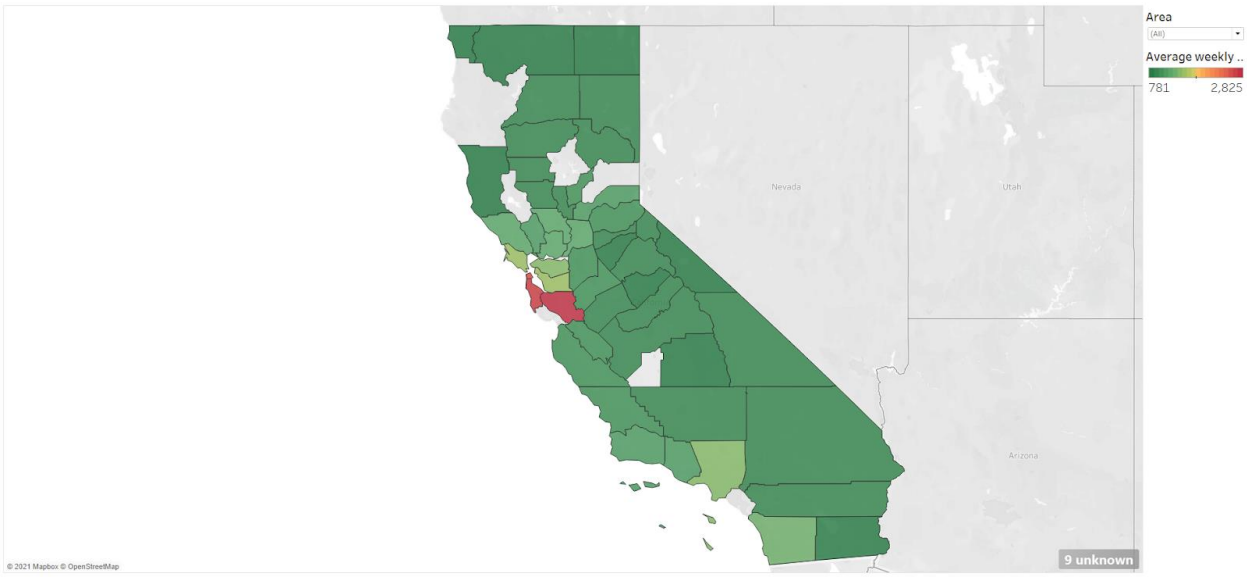

Figure 13. Average weekly wage in CA

Another histogram of average weekly wage was made in Figure 14 just to make sure we did not get tricked again like the housing price in DC. The top 5 highest salary counties in California are all in the Bay Area, and that most expensive city, Artheton, is also in the Bay Area. In this case, does that mean people should be able to afford the houses then?

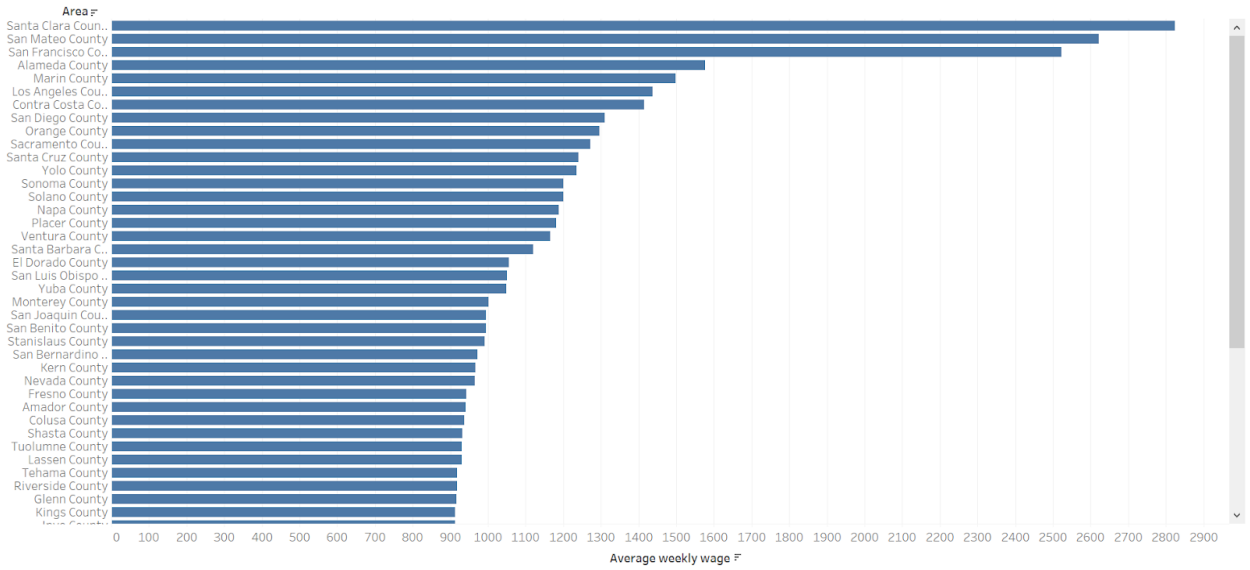

Figure 14. Histogram of average weekly wage

Figure 15 shows the correlation between salary and housing price. We can see that most counties are not stray too far from the trend line. We can also consider this line as a misery index. In other words, counties above this line should be able to buy houses more easily but it would be harder for counties below this line to afford a home just by their salaries.

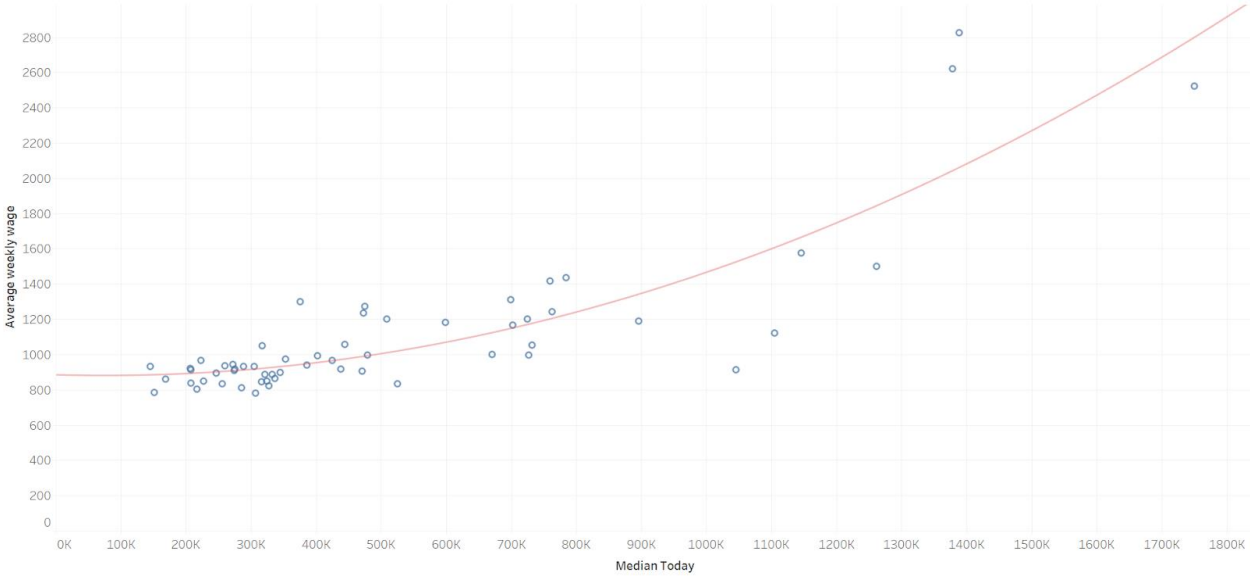

Figure 15. Scatter plot of salary and housing price

A table calculation was made in Tableau in order to figure out how many years the residents would need in order to buy a house for their own just by their salaries. With the average salary for each county, average of 9.4 years would be needed for people having average salary to buy a median price house. However, this also includes premises such as the house prices stay at what they are today, and people do not need to pay rents, pay taxes, or pay for any cost of living. If people do not pay for any of those and save all the salaries, we need about 10 years to afford a median price house in California with average salaries.

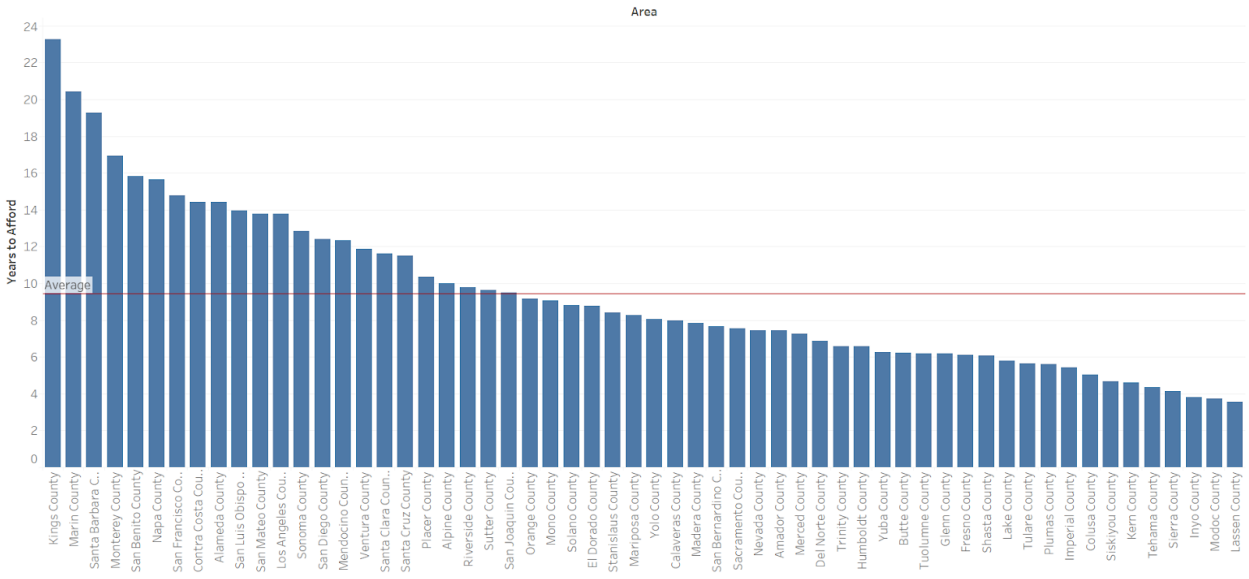

Figure 16. Years to buy a house

We all know that the premises above can never happen, so a line graph of personal saving rate in the US for about the past five years was made in Figure 17. The average of saving rate is only about ten percent for the past five years. We can also notice that people actually save a lot more during the pandemic, mostly like because there is no place to spend our money. Before COVID-19 outbreak, the personal saving rate was only right percent. Average of the whole five-year period was used to adjust the table calculation in Tableau earlier.

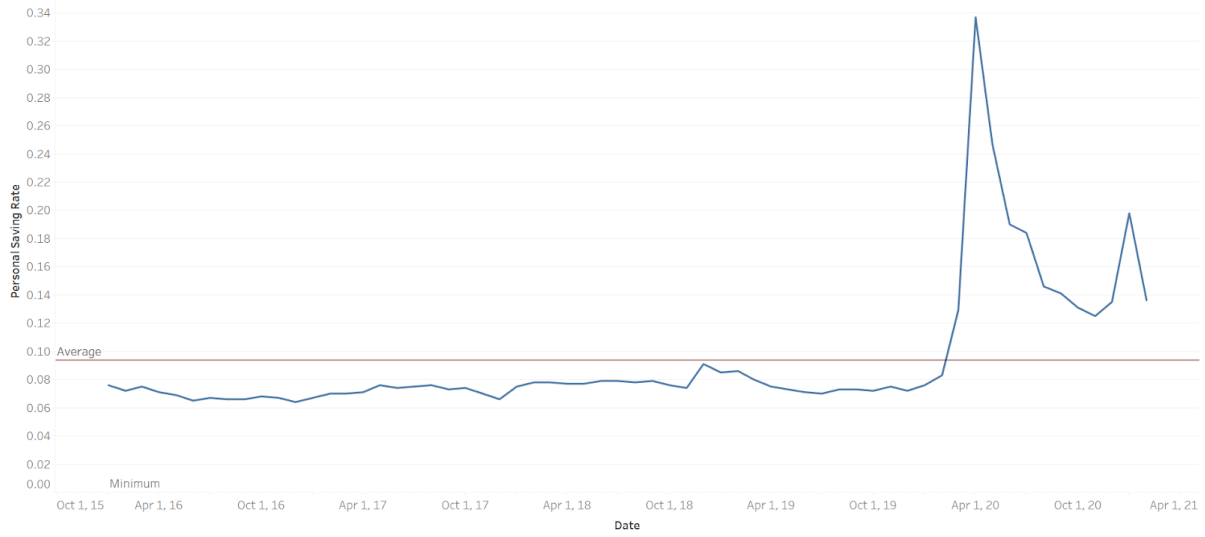

Figure 17. Personal saving rate

A new histogram was made with adjusted table calculation in Figure 18. Now we need to work and save money for a hundred years in order to afford a house. Residents in Kings county need about 250 years to afford a median price house in Kings. According to this figure, it is clear that people in the US generally are not really able to afford a home just by working and getting paid with salary.

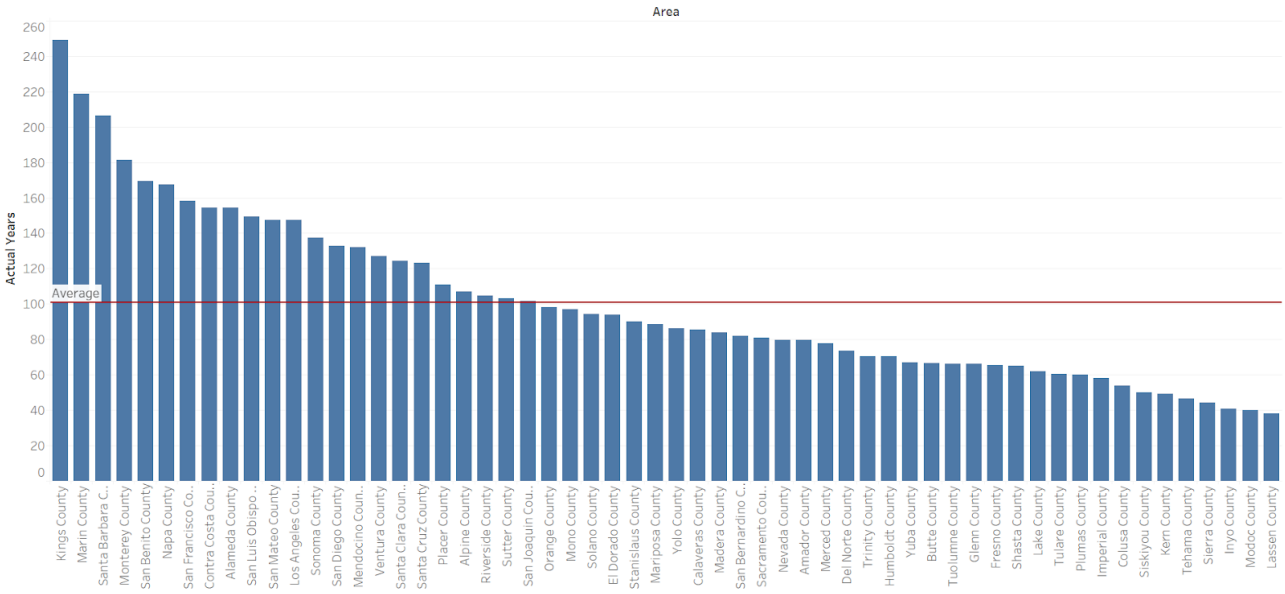

Figure 18. Actual years to buy a house

An interactive dashboard was also made in Tableau with a drop-down list that if a county was selected, it would highlight that county in both graphs for easy comparison.

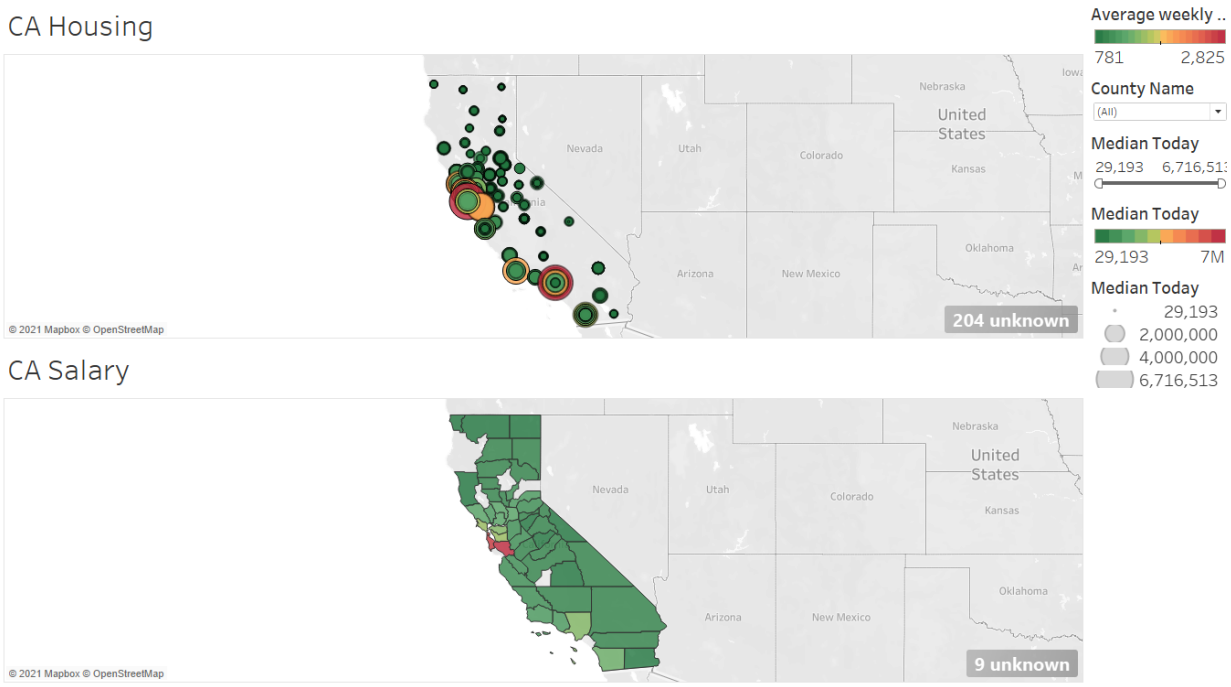

Figure 19. Interactive Dashboard

Conclusion and Future Works

There are also future works that I would like to dive deep into in the future. From what we had above, we can see that it is hard for people in the US nowadays to afford a median price house with median salary. I would like to consider more factors that affect the abilities to afford a home. For example, people with good understanding of investment or stock might have higher chance to be able to afford a house. Mortgage or mortgage rate could also be an important factor as well as housing price changing rate. If possible, I would also like to take different saving rate in different counties into consideration. There are also different tax rates as well as different cost of living in different counties. Chances of winning lotteries could also be another factor that affect the ability of people buy a home. Furthermore, I would like to expand the analysis to all the states in the US, not just California, or even to the world. I would love know figure out if citizens of a country are more easily to afford a home than another.

This term project utilized everything we learned from DATA 230 including data collection and transformation, table and graph design, visualization techniques for geospatial data, and interactive visualization. Different visualization tools were also tried for this project including Splunk and Elasticsearch, Logstash, and Kibana (ELK), but only Tableau was selected for this project in the end. I tried to include as many as visual variable in the project as possible, and I also tried all different kinds of visualization charts while working on the project to see if one fits better than another.

Resource

Bowers, M. (2021, March 25). *Housing Data*. Zillow Research. https://www.zillow.com/research/data/

County Employment and Wages in California - Fourth Quarter 2019: Western Information Office : U.S. Bureau of Labor Statistics. (2020, July 21). U.S. Bureau of Labor Statistics. https://www.bls.gov/regions/west/news-release/countyemploymentandwages_california.htm

Statista. (2021b, March 31). *Monthly personal saving rate in the U.S. 2015–2021*. https://www.statista.com/statistics/246268/personal-savings-rate-in-the-united-states-by-month/