Capstone Project 1 Milestone Report

For my first capstone project, I'm choosing to work with United States housing data taken from Zillow. I want to answer the question "How does the Bay Area housing market compare to the rest of the United States?". This problem relates to prospective homeowners looking to move to the Bay Area as it's a coastal region home to a lot of large tech companies and startups. They should know how expensive houses are and how fast or volatile home prices are growing at. Those same tech companies and politicians could use the results of this study to make decisions on what actions to take in the future since both are impacted by the housing market.

Zillow offers users to download data of the median housing price of single family homes by zip code from 1996 to April 2018 sampled by month. Once obtaining the data, the first step was to clean it up and remove any unnecessary data. A number of zip codes throughout the United States either wasn't being recorded a period time or was under development so data is missing for those zip codes so I replaced missing data points with 0's. The next step was to remove unnecessary columns that wouldn't pertain to this project such as 'Metro', 'RegionID', 'SizeRank' and 'City'. The following step required some research.

In the late 2000s, the United States was hit by a Great Recession. During this time, the housing market was greatly affected by a real estate bubble. After some research, I decided that January 2010 was the end of the Great Recession and a good starting point for my analysis. As for my data, I kept months from January 2010 until April 2018 and disregarded everything prior. What was left of the data set was every zip code of the United States ('RegionName'), county ('CountyName'), state ('State'), and housing data for every month from January 2010 to April 2018.

The next part of cleaning my data was to separate zip codes that corresponded to ones in the Bay Area. I decided to create a list of Bay Area counties by pulling it from a webpage, The California Metropolitan Transportation Commision. By doing a little web scraping and making a list comprehension,I was able to create the list of the 9 counties of the Bay Area. Next was to filter out my data set to only contain counties that matched with ones in the newly created list and delete any unnecessary columns that I wouldn't need for this analysis like 'State' since the Bay Area is constrained to be in California.

I now had two data set to analyze, one that contained housing data for all the zip codes in the Bay Area and another for the rest of the United States. The only difference between data frames was that the United States data frame still contained county names ('CountyName') and state ('State') since there are multiple counties with the same name but in different states.

Along with comparing overall house prices by zip code, I wanted to see how much they've grown since 2010 as a percent change. From the two data frames, I created two more with just January 2010 data and April 2018 data indexed by zip codes and keeping the county name (and state for the U.S. variant). I then used the following formula to add another column as the percent change.

$$percent\ change = \frac{(2018{-}04\ data) - (2010{-}01\ data)}{(2010{-}01\ data)}.$$

Now that I had the data for 2018 house prices and the percent change since 2010 indexed by zip code, I began my analysis.

The first method, as mentioned above, I looked at was the comparison between the most recent median house price per zip code and compare the Bay Area to the rest of the United States. Secondly, I calculated how much much house prices changed by zip code as a percentage and did the same comparison. Lastly, I noticed in my last report ('Data Story'), that

there seemed to be a division between Bay Area counties, mainly the 3 counties that inhabit most major businesses. In all my hypothesis testing, I used an alpha or significance level of 0.01.

2018 House Prices: Bay Area vs. United States

I started by testing the hypothesis that the mean price of a house in the Bay Area is the same for the United States. If statistically that isn't significant or reasonable, I'll reject that hypothesis and conclude that they aren't the same.

- $H_0 : \mu_{ba2018} = \mu_{us2018}$
- $H_1 : \mu_{ba2018} \neq \mu_{us2018}$

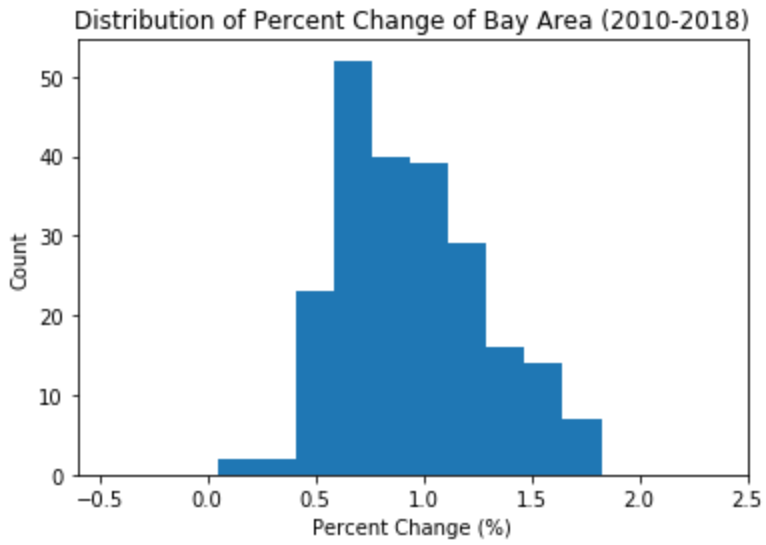
Given that I have the data for the entire population of United States homes and my sample sizes are above 30, I decided to use a z-test to compare the means. I manually calculated the z-score to be 17.65 which resulted in a p-value of 5.22×10^{-70}, well below my alpha. This means that there's statistically significant difference in the mean housing price in the Bay Area and the United States.

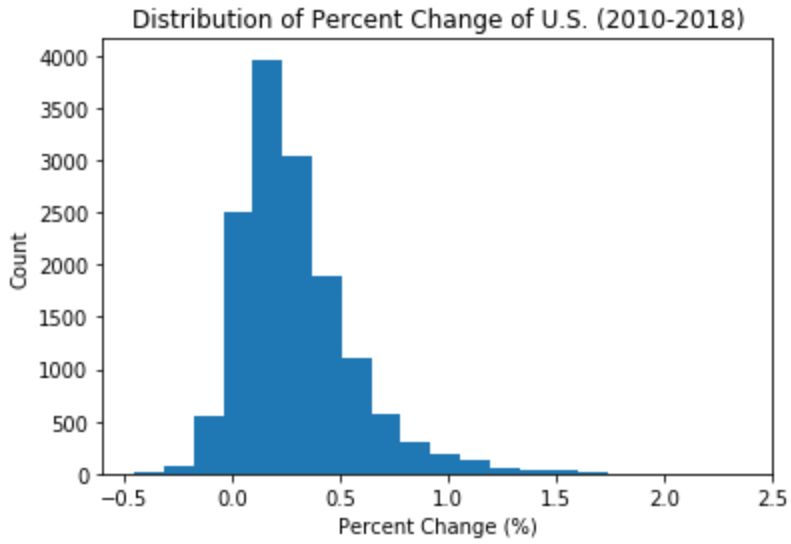
2010-2018 House Price Percent Changes: Bay Area vs. United States

Next was to compare how, percentage-wise, how zip codes across the U.S. have changed between January 2010 and April 2018. The following formula was used to calculate the percent change.

$$percent\ change = \frac{(2018\text{–}04\ data) - (2010\text{–}01\ data)}{(2010\text{–}01\ data)}$$

Given that equation, if data wasn't present, or 0, in 2010, the percent change rises to *infinity*. That distorts the data and created problems during my analysis but since those instances only accounted for 3% and 4% of my Bay Area and U.S. data respectively, I decided to ignore those data points and analyze the remaining zip codes. Below shows the distribution of percent change for the Bay Area and U.S.





Once again, since we have the population statistics and the data is sufficiently large, the z-test was used again to calculate the p-value that the mean percent change is the same for the Bay Area and the United States. With this test, I got a z-score of 29.39 and a p-value of 3.84×10^{-190} which again is significant enough to reject the hypothesis that the percent changes are the same for the Bay Area and United States.

2018 House Prices: Bay Area Counties

In my 'Data Story' report, I noticed a few Bay Area counties (San Francisco, San Mateo, Santa Clara, and Marin) had a mean house price that seemed to be clustered together, separate from the other 5 counties. I wanted to investigate if their prices had a statistically significant difference. I ran t-test between each county (36 total comparisons) and calculated the p-value of their significance assuming the null hypothesis that their means were the same. The following conclusions could be made from this analysis.

1. Napa county is only statistically different from San Francisco and Solano county. That may be due to Napa county's small sample size (7 zip codes). The confidence intervals were calculated and it's highly improbable that Napa county could have a higher mean than San Francisco. The confidence interval for the Napa county mean is between ($430,669 and $1,224,073) and San Francisco is between ($1,431,117 and $2,908,924).
2. Solano county, the cheapest county in the Bay Area on average, has statistically significant differences in mean price with every other Bay Area county. That means Solano county could be considered separate from other "lower" tier counties.
3. The "upper" and "lower" tier county house prices seem to be statistically different from each other. When comparing an "upper" tier county to a "lower" tier county, I was able to reject the null hypothesis on all instances except a few comparisons with Napa county. Again, that could be due to the small sample size of Napa county. When grouping "upper" tier counties and "lower" tier counties and running another t-test, the resulting p-value was 6.48×10^{-15}. That justifies rejecting the null hypothesis that the "upper" tier mean is the same as "lower" tier mean.

2010-2018 house Price Percent Changes: Bay Area Counties

As when comparing percent change between 2010 and 2018 Bay Area prices with the United States, I did the same when comparing Bay Area counties. T-tests were ran again

between the 9 Bay Area counties due to small sample sizes. The following conclusions could be drawn from the t-tests.

1. Again Napa county suffers from a small sample size which makes drawing conclusive analysis difficult.
2. The county with the highest percent change since 2010, Santa Clara county, was statistically different from all other Bay Area counties.
3. Three North Bay counties (Marin, Napa, and Sonoma) have all seem the three smallest percent change since 2010 and are each statistically different from the other six counties.
4. The only other significant test was between San Mateo county and San Francisco county. Of the remaining 5 counties (Alameda, Contra Costa, San Francisco, San Mateo, and Solano), those two had seen the highest and lowest percent change respectively with San Mateo seeing a 105% percent change and San Francisco seeing a 87% percent change.

Conclusions

From my analysis, it seems that the mean price of Bay Area housing is different from the mean price of housing for the rest of the United States. The also is true when comparing mean percent change of Bay Area zip codes and the United States. In both cases, the resulting p-values were significantly less than the alpha of 0.01. The mean price per zip code for the Bay Area in 2018 is $1,301,955 while the rest of the U.S. is $267,970. Not only is it statistically significant, but practically significant as well with Bay Area houses almost five times as expensive. As for percent change, the Bay Area has seen a 95.1% change since 2010 which is practically significant with Bay Area houses almost doubling since 2010 while the rest of the U.S. has seen a much more modest percent change of 28.8%.

The following table shows the average housing price per Bay Area county and the confidence interval of each with a significance level of 0.99. From this table, there seems to be a three divisions in housing prices within the Bay Area. Solano county appears the be the cheapest county compared to the rest. The middle group seems to be comprised of Alameda, Contra Costa, Napa, and Sonoma counties. Lastly, the most expensive counties are San Francisco, Marin, San Mateo, and Santa Clara. This is particularly interesting because these counties are home to big company headquarters like Google and Apple. Practically speaking, regardless of county, the lower bound of the mean for all of them are greater than the U.S. average house price.

	center	lower bound	upper bound
Alameda	1023660	878731	1168589
Contra Costa	849878	655199	1044557
Marin	1483475	955832	2011117
Napa	827371	430669	1224073
San Francisco	2170021	1431117	2908924

San Mateo	1856480	1162647	2550312
Santa Clara	1711873	1361649	2062096
Solano	441063	356363	525763
Sonoma	728903	616305	841502

The next thing to compare was the mean percent change by Bay Area county which is displayed in the following table. What stands out is that three North Bay counties (Napa, Marin, and Sonoma) have seen the least amount of change since 2010 while the other six have almost doubled in price with percent changes hovering around 1 (or 100%). In practice, the three North Bay counties with the least amount of change not not be significant. However the remaining six are what contribute to the Bay Area having mean percent change of 95.1% which is well above the United States mean percent change. The rest of the U.S. has barely changed while those six counties alone have almost doubled in price since 2010.

	change
Alameda	0.958301
Contra Costa	0.885270
Marin	0.594183
Napa	0.622233
San Francisco	0.976023
San Mateo	1.041695
Santa Clara	1.255284
Solano	0.930338
Sonoma	0.610309