Exploratory Data Analysis

Question: How does the Bay Area housing market compare to the rest of the United States?

The way I approached this problem was to look at the median house price by zip code throughout the United States and how they've changed since the Great Recession ended. I used the median housing price in January 2010 to the most recent prices as of April 2018. I used those numbers to do my analysis on. The first method I looked at was to compare 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.

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.