Predicting Home Values in Los Angeles' South Bay

Springboard | Statistical Inference

By: Lauren Broussard

Date: 05/21/2020

As we looked to see if we could predict home prices in the South Bay, we used Bootstrap Inference to try to better understand the relationship between price and a few features in our dataset. Corresponding code can be found in the Jupyter notebook "Capstone1_Apply Statistics.ipynb."

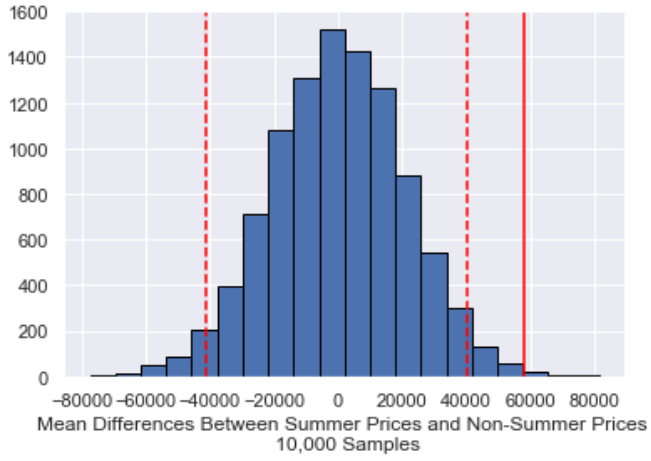
We asked the following questions about two features in particular, "Sold Date" & "Year Built":

- **Is there a difference in average home price for homes that sell in the summer vs. non summer months?**
- **Is there a difference in average home price between "newer" homes vs "older" homes?**

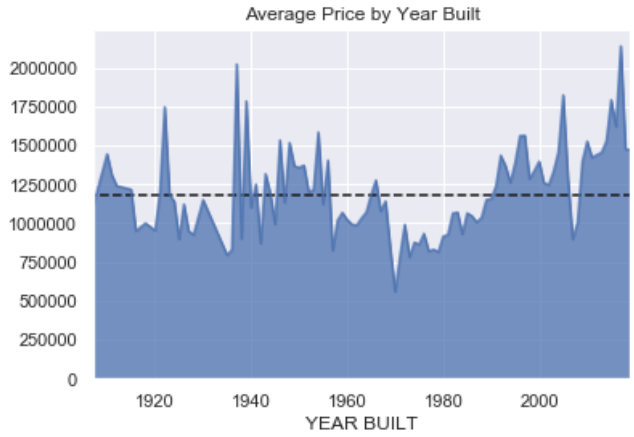
Summer vs. Non Summer Sales: Is there a "hot time" to buy a home? From initial data, it looked like the *number* of home sales in the summer months was greater than non summer months. Could the increase in sales be due to a difference in price in the summer months vs. other months? For the purpose of this test, we defined "summer" months as June, July, and August, and "non summer" months as the other 9 months in the year. We then got values for the average price in each group, as well as the difference in means.

Mean Price, Summer: $1,034,713.47
Mean Price, Not Summer: $976,924.60
Difference in Mean Prices: $57,788.87

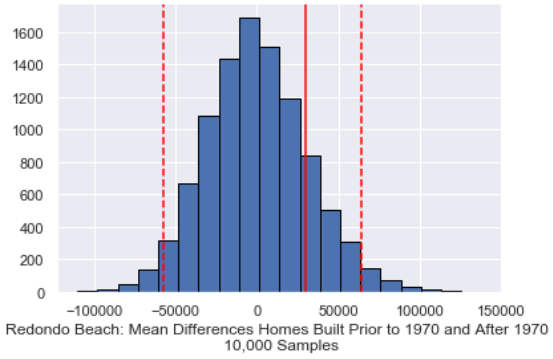
If the prices of the two groups were the same, we would expect to see that the difference in mean would be closer to 0. For this experiment, we hypothesized that there was no difference in the two groups. We chose a significance level of 0.05. To run the test, we used Bootstrap Inference to simulate resampling our data. We shifted the means of the groups, so that the average prices were equal in both groups, and ran our experiment 10,000 times to see how likely it would be to get the mean difference we observed above. Our p-value from this test was 0.0019, so we would reject the null hypothesis. There may then be some difference in price between summer and non-summer prices.



New Home vs. Older Home: Are newer homes really all the rage? To control for the fluctuation in prices between neighborhoods, we chose one neighborhood to look at - Redondo Beach. We split the homes then into those that were built prior to 1970 and those built after 1970. We chose this year as it seemed that this was about when prices began to steadily increase.



We again chose a Bootstrap test like the one above to compare the differences between these two groups. Running this test 10,000 times, we ended up with a p-value of 0.1571, for a significance level of 0.05. With this information, we would accept the null hypothesis that there is no difference in average price between homes built before 1970 and those built after in this particular neighborhood of Redondo Beach.



Other Considerations: With both of these tests, there are other factors that could be at play to explain the difference in pricing data. For instance, the recession in 2008 could have helped to bring down the average prices in the group. It may make sense to look at homes built pre-recession and post recession, which may have yielded a different result.

Additionally, in our test of home prices vs sale date -- home prices tend to go up over time in general, so the difference could be due to prices simply rising over time. In future tests, we will need to control for some other factors.