Anthony Ayala

8/20/23

Executive Summary

We joined Boston Consulting Group recently and we were tasked to clean up the mess left by a previous consultant for the City of Boston. We were specifically tasked to look at Boston Home Pricing and tasked to perform analysis on what variables impact assessed value for properties. In other words, were tasked to help with predicting home value prices and figure out whether Model 1 or Model 2 had the best predictions. Our challenges are to find the correlation between the numeric variables and the target variable, choose relevant categorical variables that relate to the target variable, provide evidence on selection with the use of plots and tables, and then compare the linear regression results for the model.

In our analysis, we found several key findings for understanding assessed value for properties that range ranking cities in terms of home value, having descriptive statistics on assessed value for property, the combination of categorical variables on home value, the influence on price based on decade a house was built, and the correlation between explanatory variables. First, on page 7 of our appendix that has a histogram, boxplot, and table of our target variable ("assessed value for property"), we have a right skewed distribution that is unimodal. Hence, we reference the median as our preferred measure of center where the median price is $415,400.00 and see that the interquartile range of assessed value for property ranges from a low of $348,184.92 to $503,502.88. With this being known, we can add another layer of analysis by looking at the city states columns and see if these neighborhoods fit under this IQR or if they're an outlier. On the bottom of page 10 in our appendix, we have a box plot of "av_total = city state" and have a descriptive statistics table on the right, and we find that Jamaica Plain comes in 1st at $659,900.00 for highest home value, Cambridge in 2nd at $458,079.45, Roslindale in 3rd at $432,396.82, Dorchester in 4th at $387,089.37, and Hyde Park in 5th at $337,088.47. Jamaica Plain would be the one city state that explains the right skew in the distribution of "av_total". Another interesting piece of categorical analysis is on page 11 of the appendix with the big table, we start to slowly answer our client's beliefs on owner occupied homes and remodeling of homes causing higher assessed values. For example, we note significant differences in price like $79,309.05 when performing a comparison for the average home in Jamaica Plains that is owner occupied ($718,124.13) and has been remodeled to the average home in Jamaica Plains that is not owner occupied and has not been remodeled ($638,814.98). With our next important finding, we get to assess our client's beliefs on page 12 to 14 in our appendix with the boxplots and tables provided. We find that homes built in 1990s come in at the second highest for home values "the Mean Home Value $602,448.74 (2000s) > $454,867.33 (1980s) > $427,557.64 (1990s) and the 50% of Home values $585,800.00 (2000s) > $435,650.00 (1980s) > $433,600.00 (1990s)" (Page 13). Thus, we see that homes built in the 2000s come at the highest home value. With this finding, we see that this plays out for homes that were recently remodeled where the "2000s has highest home values than 1980s and 1990s: Mean $519,026.55 (2000s) > $480,845.13 (1990) > $440,722.12 (1980s) and Median $593,300.00 (2000s) > $438,400.00 (1990s) > $407,200.00 (1980s)" (Page 14). Finally, the last insight highlights the possibility of multicollinearity with the correlation matrix on page 16 in the appendix that shows the correlation between explanatory variables. For example, we can see how “s ome variables that are easily understood that have high correlations between each other are number of total rooms and living area, number of bedrooms and living area, and number of floors and living area. It’s easily understood by factoring in that the living area space a home has the more likely the home is to have more space for bedroom, an additional floor, and rooms. These are examples of positive relationships” (Page 16).

For our model performance, it was no issue at all for selecting the better model as Model 2 had the highest R-squared value of 94.7% (Method 1 of handling nulls) and 94.8% (Method 2 of handling nulls). While Model 1 had a R-Squared value of 43.7% (Method 1 of handling nulls) and 43.6% (Method 2 of handling nulls), these results come from page 2 and 3 in our appendix. Model 2 has 94.7% of variability in assessed value for property that can be explained by the model, it also true that Model 2 has low errors for Root Mean Squared Error and Mean Absolute Error. For example, when making another comparison between the models, Model 2 and Model 1 have drastic differences for both RMSE and MSE where “The Root Mean Squared Difference is −74,406.42 (Model 2 MSE - Model 1 MSE) and Absolute Mean Difference is −53,475.22 (Model 2 MAE - Model 1 MAE), so this serves as greater evidence in why we chose Model 2 to be the better and more accurate Model”(Page 4).

To recap everything, we find that Model 2 is the best at predicting home values, we have three very important critical categorical variables that impact home value, homes built in the 2000s or remodeled in the 2000s have the highest home value, and there might be a slight concern of multicollinearity between the explanatory variables. Now, let’s share some recommendations for our client, the city of Boston. We will first suggest that we should do a model based on location (“city_state”) for homes as we want to see what Jamaica Plains and Cambridge very expensive. Some questions that may arise could be the different sized homes, is there a college campus nearby, do a lot rich people live in these areas, and so forth. We also think that the owner/renters and collecting data on them about considering remodeling in the next few years, money saved up for remodeling, whether they have home insurance, year they bought/rented the home, and so forth. This recommendation will help address the insights that we gained into homes being remodeled or recently built that they do tend to have a higher home value, and seeing if that still plays true or what else influences remodeling and newly built homes. One final recommendation is that for potential new homeowners in Boston, it might be best to not live in Jamica Plains as it consistently has the most expensive homes.