Does Remodeling Affect Housing Prices in the Greater Boston Area?

Andrew Clark

Department of Economics, Lehigh University

ECO 357- Econometrics

Muzhe Yang

December 4th, 2025

Abstract

This paper conducts an empirical study to examine the impact of remodeling post 2005 on the greater Boston Area housing assessment prices. We use information from the Boston.gov dataset called, “Property assessments FY2025”, with our final sample containing 68,169 observations across different structural and renovation variables. This study adds on to previously published pieces to understanding the relationships between different housing variables, including structural and renovation features. Within our model we control for neighborhood fixed effects (ZIP_CODE) and include structural and renovation to return a log housing assessment price.

Our model has a very high R-Squared at .8495. The model showed overall general property improvements had an 6.72% increase in log housing assessment prices. Targeted room remodeling also returns high increases in value, as remodeling a kitchen is associated with a 7.34% incremental increase in value and remodeling your bathroom is associated with a 5.86% incremental increase in value. A significance test confirmed that the value premium from a kitchen remodel is statistically different from a bathroom remodel.

1. Introduction

The housing market is essential to everyone's lives. With home equity representing a median of 45% of the Net Worth of U.S homeowners, residential property values are established as one of the most important assets in determining one's individual finances (Kochhar & Moslimani, 2023). It is important to understand what your home is worth and the factors that contribute to your home's value. We aimed to use a variety of factors to determine housing assessment values to inform individuals what to change in their home. These include both structural and renovation based features. This is because quality differences can arise in markets for durable goods, markets such as real estate, where consumers have a choice of new or used goods, some that may have deteriorated to a lower quality(Sweeney, 1974).

This paper aims to complete an investigation into factors that impact housing assessment values in the greater Boston Area. More specifically, how homeowners can utilize renovations to get a premium on their home. One of the most important factors in determining home prices by anyone in the real estate world is three special words. “Location, Location, Location”. There is a, notably higher incentive to renovate in attractive areas, where people want to live, since the expected gain from renovations exceeds the cost by a greater margin due to demand (Gyourko & Saiz, 2004). Hence why we made sure our model accounted for location via zipcodes.

This paper conducts a 3 level analysis, leading to the final research question of, “to what extent do specific structural features and targeted home renovations contribute to the assessed value of residential properties in the greater Boston Area, after controlling for location”. We utilized a hedonic pricing model, with the dependent variable being the natural log of total housing assessment value. We get our estimates through a linear regression with 68,169 observations. Overall, we are expecting property size and overall condition to be the strongest predictors of log housing assessment values.

2. Data collection, Cleaning and Usage

2.1 Data Collection

The data collected and utilized for this study was sourced from Boston.gov, this is a free, publicly accessible data source that allows Boston residents and visitors to access information and resources from a variety of city departments. The data is more specifically from the assessing department and is to "ensure fair assessment of Boston taxable and non-taxable property of all types and classifications" (Boston.gov, 2025). The data is also updated yearly and the code written for this project is to be reused year after year. The original data set contains 66 variables across 183,446 observations, which in this case represent all properties in Boston. With this many observations there are bound to be errors inside the data.

2.2 Data Cleaning

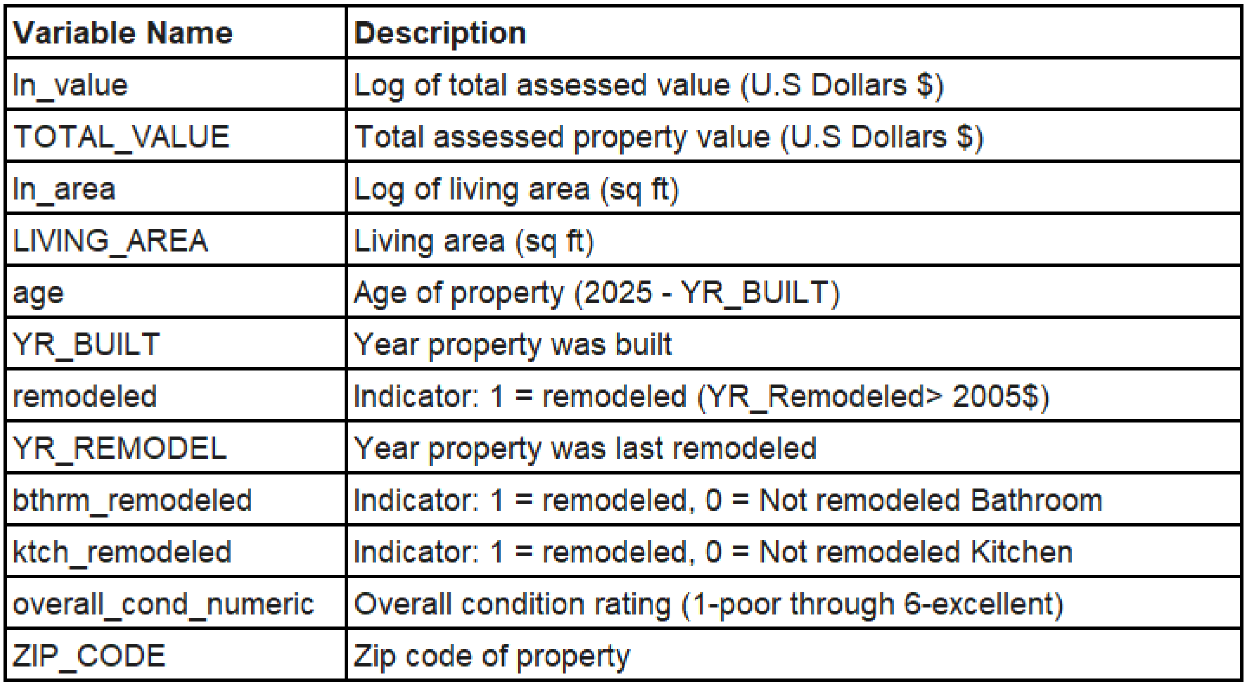
This section describes the technique used to clean the extremely large data set. It is important to note that the data was cleaned using programming language R. First we needed to filter out for residential only land codes, as the data set contained ALL Boston properties. Following this, some extra filtering was performed in order to make sure items including, YEAR_BUILT, TOTAL_VALUE and YEAR_REMODLED all make sense from a logical stand point. Refer to Appendix section 2.2.1. Next we created our own variables (through mutations) using the data already in the dataset. We created a renovation flag (so houses after 2005 are considered remodeled), log of both TOTAL_VALUE and LIVING_AREA, Property age and effective tax rate variables. Also created bathroom and kitchen renovation dummy variables from categorical data. Refer to Appendix section 2.2.2 for the code to create these variables.

Next we merged duplicated categorical variables to create consistent rating and converted useful factors into numerics through order level factoring for easier analysis. This way the data's importance does not change, only how it is labeled changes. Refer to Appendix section 2.2.3. We then got rid of some extreme outliers while maintaining the integrity of the dataset by dropping the top 1% and bottom 1% of log TOTAL_VALUE. Effectively keeping the middle 98% of values. The data is then prepared for export where all remaining NA's are dropped. We claim this was acceptable as it only removed a small % of the sample and allowed for all our observations to be complete and intact for regression analysis. Refer to Appendix section 2.2.4 for more details. Finally the data is written into an EXCEL to be used in stata to run regressions.

2.3 Data Usage

The following section shows the variables used with their description in the regressions. These variables were carefully selected from the metadata for analysis.

Table 2.3.2 Variables used in regression



Variable Name	Description
ln_value	Log of total assessed value (U.S Dollars $)
TOTAL_VALUE	Total assessed property value (U.S Dollars $)
ln_area	Log of living area (sq ft)
LIVING_AREA	Living area (sq ft)
age	Age of property (2025 - YR_BUILT)
YR_BUILT	Year property was built
remodeled	Indicator: 1 = remodeled (YR_Remodeled> 2005$)
YR_REMODEL	Year property was last remodeled
bthrm_remodeled	Indicator: 1 = remodeled, 0 = Not remodeled Bathroom
ktch_remodeled	Indicator: 1 = remodeled, 0 = Not remodeled Kitchen
overall_cond_numeric	Overall condition rating (1-poor through 6-excellent)
ZIP_CODE	Zip code of property

Note that some of the variables above were transformed into log variables. The rationale behind this was due to right skew on both TOTAL_VALUE and on LIVING_AREA shown by the statistics output below, resulting in the use of log to make the data more normalized.

Table 2.3.3 Descriptive summary of TOTAL_VALUE (U.S Dollars ($))

MIN	Q1	MEDIAN	MEAN	Q3	MAX	SD
238,400	551,600	749,800	902,219	1,040,100	4,721,000	582,577

Table 2.3.4 Descriptive summary of LIVING_AREA (Sq Ft)

MIN	Q1	MEDIAN	MEAN	Q3	MAX	SD
340	815	1272	1596	2133	5466	1008

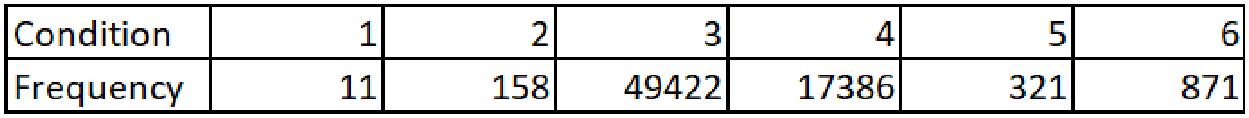
The descriptive statistics for age are as expected. No transformations were used on the variable because keeping the variable in linear form allows us to interpret the effect of age on property value as constant with each additional year of the property's life. Effectively capturing depreciation over the time for the asset. Using vce(robust) captures the rest of the error in our third level of analysis.

Table 2.3.5 Descriptive summary of age

MIN	Q1	MEDIAN	MEAN	Q3	MAX	SD
3	96	120	108	126	315	36

Next we have the distribution of overall_cond_numeric which is ordinal data from 1 (poor condition) to 6 (excellent condition). We will talk more about this in the limitation section.

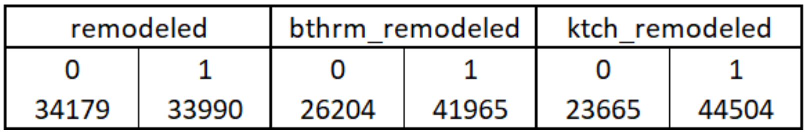
Table 2.3.6 Frequency distribution for overall_cond_numeric



Condition	1	2	3	4	5	6
Frequency	11	158	49422	17386	321	871

Then we have our remodeled, kitchen remodeled and bathroom remodeled statistics which are dummy variables and have the following results. The remodeled section is split to be about half the population, the bthrm_remodeled and ktch_remodel are a bit less but are still useful.

Table 2.3.7 Frequency distributions for remodeled, bthrm_remodeled and ktch_remodeled



remodeled		bthrm_remodeled		ktch_remodeled	
0	1	0	1	0	1
34179	33990	26204	41965	23665	44504

As far as ZIP_CODE goes refer to the frequency table in Appendix 2.3.8 for more details. ZIP_CODE is absorbed for analysis and denoted as u_z.

3. The 3 levels of analysis

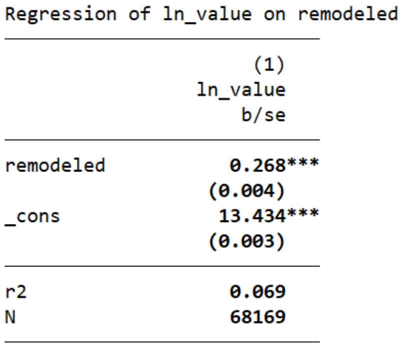
These sections hold the regression models and their results used to analyze the relationship between property characteristics, including renovation status, and the natural logarithm of a property TOTAL_VALUE (ln_value). From here on we switched from using R to using STATA. The reasoning being STATA is easier to use when working with only one data frame because of the simpler commands.

3.1 First level of Analysis: Simple Regression

The goal of this basic analysis was to estimate the effect of different types of renovation on the property value. This is done with a log- base analysis with the dependent variable being ln_value to the independent variable of remodeled. This creates a simple t-test to make sure our hypothesis that remodeling has a positive effect on housing price.

Equation: $\ln\_value = \beta_0 + \beta_1 remodeled + u$

Results (Table 3.1.1)



```
Regression of ln_value on remodeled
----------------------------------------
                      (1)
                 ln_value
                     b/se
----------------------------------------
remodeled           0.268***
                  (0.004)
_cons              13.434***
                  (0.003)
----------------------------------------
r2                  0.069
N                   68169
----------------------------------------
```

Interpretation

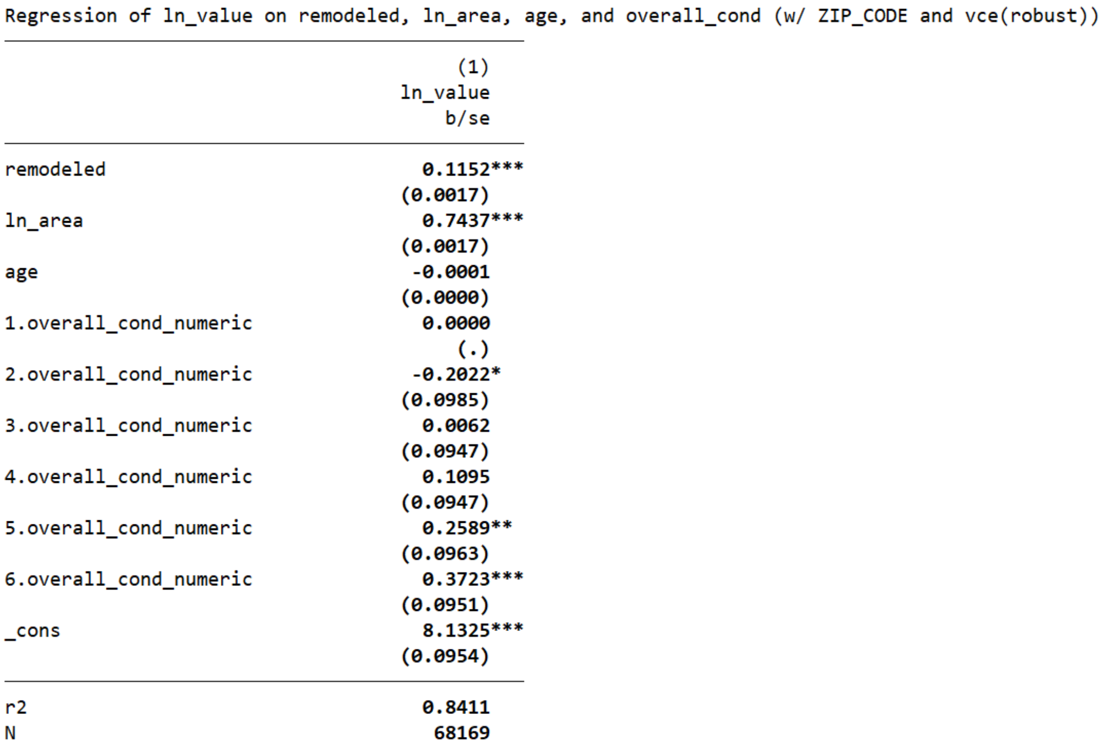
First we need to change β_1 to be an exact value because of the log - base equation. To find this, we take $100*(e^{.268}-1) = 30.73\%$. This model shows that a property that has been remodeled is predicted to have about a value that is 30.73% higher than a property that has not been remodeled after 2005. This effect is highly statistically significant as the p-value is .004 which is less than alpha = .01. Note that we are not including any other factors in this equation, it is the most simple model.

3.2 Second Level of Analysis: adding in complexity

Some factors that are important to assess value were not included in the first level of analysis including how big the home is (ln_area) as a bigger is associated with more and therefore a higher value. How old the home is (age), as older homes will slowly depreciate over time, as they require more and more work. The condition of the home (i.overall_cond_numeric) is a huge factor considering people want to live in nicer homes, as opposed to a home that is in rough shape. Last but not least, if the house is located in a more desired neighborhood (ZIP_CODE) the value of the home will increase due to demand. It is important to note that we are absorbing ZIP_CODE into variable u_z using the areg function from STATA. So u_z is capturing unobserved factors constant within ZIP_CODE. u captures the rest of the unobserved factors on home assessment value. In the simple model, remodeling captures the effects of all these other factors, leading to potential bias, or what is more commonly known as omitted variable bias. This new model attempts to isolate the effect of remodeling through the control of these new variables that can commonly be associated with housing value. Also used vce(robust) to account for heteroscedasticity.

Equation: ln_value = β_0 + β_1 remodeled + β_2 ln_area + β_3 age + β_4 2.overall_cond_numeric + β_5 3.overall_cond_numeric + β_6 4.overall_cond_numeric + β_7 5.overall_cond_numeric + β_8 6.overall_cond_numeric + u_z + u

Results (Table 3.2.1)



```
Regression of ln_value on remodeled, ln_area, age, and overall_cond (w/ ZIP_CODE and vce(robust))
----------------------------------------------
                                     (1)
                                ln_value
                                    b/se
----------------------------------------------
remodeled                          0.1152***
                                (0.0017)
ln_area                            0.7437***
                                (0.0017)
age                               -0.0001
                                (0.0000)
1.overall_cond_numeric             0.0000
                                     (.)
2.overall_cond_numeric            -0.2022*
                                (0.0985)
3.overall_cond_numeric             0.0062
                                (0.0947)
4.overall_cond_numeric             0.1095
                                (0.0947)
5.overall_cond_numeric             0.2589**
                                (0.0963)
6.overall_cond_numeric             0.3723***
                                (0.0951)
_cons                              8.1325***
                                (0.0954)
----------------------------------------------
r2                                 0.8411
N                                   68169
```

Interpretation

This model did a significantly better job at explaining the ln_value dependent variable. This new model has an r2 of 84.11% which is much better then the simple model that had a r2 of 6.9%. Starting with the remodeled variable, we have to change β_1 to be an exact value by doing 100*($e^{.1152}$-1) = 12.19%. A remodeled house after 2005 is associated with 12.19% higher value than a non-remodeled house holding other factors constant. The effect is highly statistically significant as the p-value is .0000 which is less than alpha =.01.

Looking at the ln_area variable we see that a 1% increase in the area of the house is associated with a .7437% increase in the house's value holding all other factors constant. This was analyzed through a log - log comparison. The effect is highly statistically significant as the p-value is .0000 which is less than alpha =.01.

The age coefficient can be interpreted as a one-year increase in the age causing a 0.01% decrease in housing assessment value when holding all other factors constant. No need to do an "e" transformation as the number is very close to zero. With a p-value equal to 0 this coefficient is statistically significant. So the model correctly incorporates depreciation, however its effect is essentially negligible.

Overall_cond_numeric had condition 1 be dropped, so we compare the rest of the results to the dropped one for categorical variables. Used a table to explain the data better.

Table 3.2.2

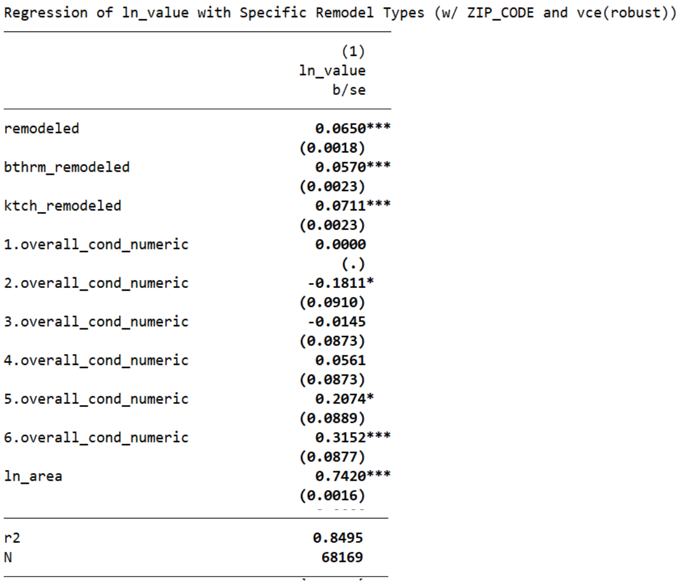
Condition Level	**Interpretation Vs Condition Level 1**	**Significant level ($\alpha = .05$)**
2 (β_4)	20.22% lower value	Not significant given α
3 (β_5)	0.62% higher value	Not significant given α
4 (β_6)	10.95% higher value	Not significant given α
5 (β_7)	25.89% higher value	Significant given α
6 (β_8)	37.23% higher value	Significant given α

3.3 Third Level of Analysis: Targeted remodeling

It is very important to understand what types of remodeling can provide greater value to a home. So we compared the effect of general renovations, to the effect of kitchen remodeling, to the effect of bathroom remodeling. This is all done while keeping the same controlling variables that we used before, and zipcode is still absorbed and denoted as u_z.

Equation: ln_value = β_0 + β_1 remodeled + β_2 bthrm_remodeled+ β_3 ktch_remodeled + β_4 2.overall_cond_numeric + β_5 3.overall_cond_numeric + β_6 4.overall_cond_numeric + β_7 5.overall_cond_numeric + β_8 6.overall_cond_numeric + β_9 ln_area + β_{10} age + u_z + u

Results (Table 3.3.1)



```
Regression of ln_value with Specific Remodel Types (w/ ZIP_CODE and vce(robust))
------------------------------------------------
                                    (1)
                               ln_value
                                   b/se
------------------------------------------------
remodeled                       0.0650***
                               (0.0018)
bthrm_remodeled                 0.0570***
                               (0.0023)
ktch_remodeled                  0.0711***
                               (0.0023)
1.overall_cond_numeric          0.0000
                                    (.)
2.overall_cond_numeric         -0.1811*
                               (0.0910)
3.overall_cond_numeric         -0.0145
                               (0.0873)
4.overall_cond_numeric          0.0561
                               (0.0873)
5.overall_cond_numeric          0.2074*
                               (0.0889)
6.overall_cond_numeric          0.3152***
                               (0.0877)
ln_area                         0.7420***
                               (0.0016)
------------------------------------------------
r2                              0.8495
N                                68169
------------------------------------------------
```

Interpretation

This third level of analysis focuses mostly on remodeling dummy variables. Looking at β_1, a property that has been remodeled in general, is associated with a 6.72% increase in housing assessment value from the equation 100*($e^{.0650}$-1). Looking at β_2, a property that has had a bathroom remodel is associated with a 5.86% increase in housing assessment value from the equation 100*($e^{.0569}$-1). Looking at β_3, a property that has had a kitchen remodel is associated with a 7.34% increase in housing assessment value from the equation 100*($e^{.0711}$-1). What we want to know is if remodeling your kitchen has a different affect on housing value then remodeling your bathroom. Running a quick t-test we get the following:

Figure 3.3.2

```
 ( 1)  bthrm_remodeled - ktch_remodeled = 0

       F(  1, 68127) =   12.00

            Prob > F =    0.0005
```

The results show that at alpha = .01 we can conclude that remodeling your kitchen is statistically different then remodeling your bathroom for housing values. Therefore, by these results, you should prioritize a kitchen remodel over a bathroom remodel.

Also, the overall condition betas (β_4- β_8) did not change too much. We compare these betas the same way we did in the second level of analysis.

Table 3.3.3

Condition Level	**Interpretation Vs Condition Level 1**	**Significant level (α = .05)**
2 (β_4)	18.11% lower value	Not significant given α
3 (β_5)	1.45% lower value	Not significant given α
4 (β_6)	5.61% higher value	Not significant given α
5 (β_7)	20.74% higher value	Significant given α
6 (β_8)	31.52% higher value	Significant given α

Looking at the ln_area variable we see that a 1% increase in the area of the house is associated with a .7420% increase in the house's value holding all other factors constant. This was analyzed through a log - log comparison. The effect is highly statistically significant as the p-value is .0000 which is less than alpha =.01.

Finally look at the age coefficient we see 0 effect of age on housing value. All of the factors relating to age are counted for by other variables (overall_cond_numeric and remodeling dummies). So this is not a surprising result.

4.Limitations

The primary limitation comes from the overall_cond_numeric variable. Looking at the frequency table 2.3.6 we see that conditions 3 and 4 have the highest frequency. Which makes sense as most properties should be considered “Average” or “Good”. However, condition 1 contains only 11 observations. Since condition 1 is the omitted group in analysis it reduces the reliability for the interpretation of the coefficients. For example, some conditions may appear insignificant, just because the reference category (condition 1) has very little observations.

Another limitation is the measurement error in the assessment value of homes. These homes are assessed by a home assessor but not all by the same assessor. Different assessors may interpret different conditions for the same home, and therefore get different home values for the same home. So this can be a high source of error in the data set.

Another limitation is that the remodeling variables are calculated primarily from categorical fields. If a home was classified wrong this can lead to high error as it affects all of the results. It can change the effects of the coefficients, meaning the true value of renovation premiums could be higher (or lower) than reported in this paper.

One final note on remodeling. Remodeling is not random, but usually from the house deteriorating, or about to be sold. Might be other important unobserved factors affecting analysis.

5.Conclusion

Overall, this study researched how different home characteristics and renovation types affect the housing assessment prices of homes in the Greater Boston area. This analysis concludes that after controlling for the most important variable in real estate, location, and other key variables in home valuations, we find that remodeling is associated with a higher home assessment value. We also concluded through an F test, that renovating your kitchen will have a higher effect on housing assessment value than remodeling your bathroom. However, we should be careful with the results as specified in the limitations section.

References

Boston.gov. (2025). *Property assessment - dataset - analyze Boston*. Property Assessment. https://data.boston.gov/dataset/property-assessment

Glaeser, E. L., Kincaid, M. S., & Naik, N. (2018). *Computer vision and real estate: Do looks matter and do incentives determine looks*. National Bureau of Economics Research. https://www.nber.org/papers/w25174

Gyourko, J., & Saiz, A. (2004). *Reinvestment in the housing stock: The role of construction costs and the supply side.* Journal of Urban Economics, *55*(2), 238–256. https://doi.org/10.1016/j.jue.2003.09.004

Kochhar, R., & Moslimani, M. (2023). *4. the assets households own and the debts they carry*. Pew Research Center. https://www.pewresearch.org/2023/12/04/the-assets-households-own-and-the-debts-they-carry/

Mamre, M. O., & Sommervoll, D. E. (2022). *Coming of age: Renovation premiums in housing markets - the Journal of Real Estate Finance and Economics*. SpringerLink. https://link.springer.com/article/10.1007/s11146-022-09917-w

Sweeney, J. L. (1974). *Quality, commodity hierarchies, and housing markets.* Econometrica: Journal of the Econometric Society, 147–167. https://doi.org/10.2307/1913691

Appendix

The following code was written in program language called: R

Appendix 2.2.1

```
boston_clean <- boston %>%
  filter(
    TOTAL_VALUE > 0,
    YR_BUILT > 1700 & YR_BUILT <= 2025,
    LU %in% res_codes,
    YR_REMODEL <= 2025
)
```

Appendix 2.2.2

```
boston_clean_var <- boston_clean %>%
  mutate(
    # renovation flag
    remodeled = if_else(YR_REMODEL > 2005, 1, 0, missing = 0),
    # log values
    ln_value = log(TOTAL_VALUE),
    #Log Area
    ln_area = log(LIVING_AREA),
    # property age
    age = 2025 - YR_BUILT,
    # Tax Rates might be useful
    eff_tax_rate = GROSS_TAX/TOTAL_VALUE,
    #Bathroom remodeled or not
    bthrm_remodeled = ifelse(
      BTHRM_STYLE1 != "N - No Remodeling" & BTHRM_STYLE1 != "S - Semi-Modern" &
        BTHRM_STYLE1 != "" &
        !is.na(BTHRM_STYLE1), 1, 0),

    #Kitchen remodeled or not
    ktch_remodeled = ifelse(
      KITCHEN_STYLE1 != "N - No Remodeling" & KITCHEN_STYLE1 != "S -
Semi-Modern" &
        KITCHEN_STYLE1 != "" &
        !is.na(KITCHEN_STYLE1), 1, 0)
  )
```

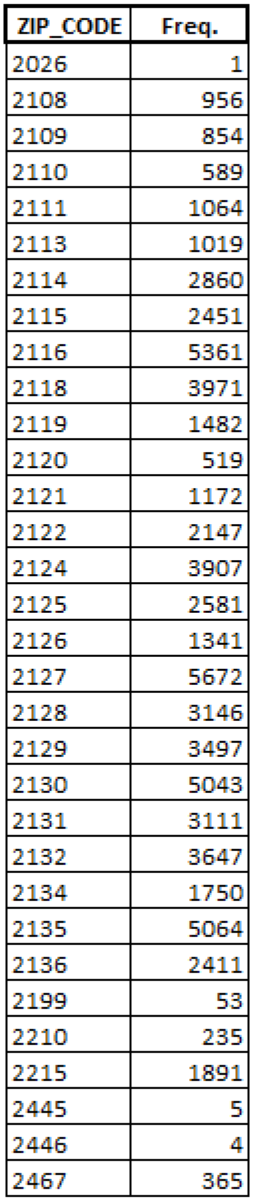
Appendix 2.2.3

```
boston_clean_var <- boston_clean_var %>%
  mutate(
    overall_cond_factor = factor(
      OVERALL_COND,
      levels = c("P - Poor", "F - Fair", "A - Average", "G - Good", "VG - Very Good", "E -
Excellent"),
      ordered = TRUE
    ),
    overall_cond_numeric = as.numeric(overall_cond_factor))
```

Appendix 2.2.4

```
boston_clean_var <- boston_clean_var %>%
  filter(between(ln_value, quantile(ln_value, 0.01, na.rm=TRUE),
                 quantile(ln_value, 0.99, na.rm=TRUE)),
         between(ln_area, quantile(ln_area, 0.01, na.rm=TRUE),
                 quantile(ln_area, 0.99, na.rm=TRUE)))
```

Appendix 2.3.8



ZIP_CODE	Freq.
2026	1
2108	956
2109	854
2110	589
2111	1064
2113	1019
2114	2860
2115	2451
2116	5361
2118	3971
2119	1482
2120	519
2121	1172
2122	2147
2124	3907
2125	2581
2126	1341
2127	5672
2128	3146
2129	3497
2130	5043
2131	3111
2132	3647
2134	1750
2135	5064
2136	2411
2199	53
2210	235
2215	1891
2445	5
2446	4
2467	365