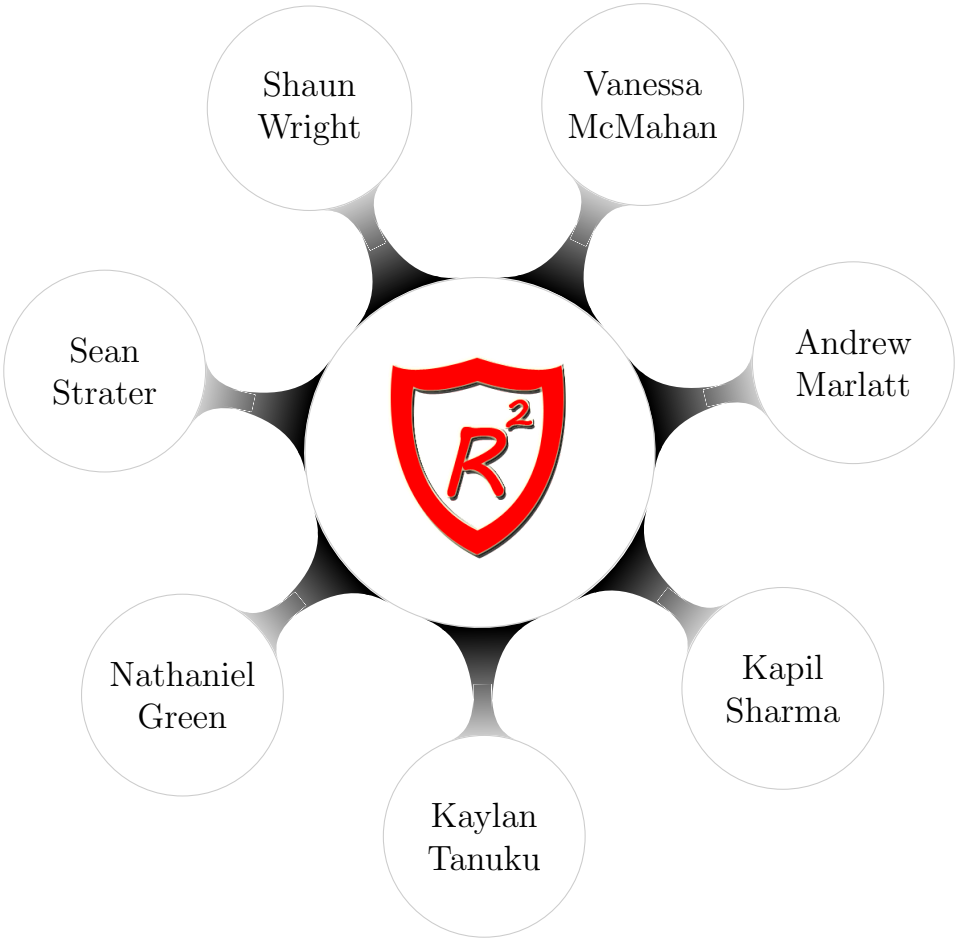
ANALYTICS
TEXAS A&M UNIVERSITY

REGRESSION ANALYSIS (STAT-608)
FALL 2017

TEAM PROJECT

Submitted By:



October 19, 2017

Executive Summary

Our team has created a multivariate regression model with the goal of predicting *Assessed Value* of a condominium using a combination of predictor variables for 8,230 properties in Ward 3 of the City of Boston. Based on the constant variance in the residuals, the Global Analysis of Variance test, and the Marginal Model Plots, along with other diagnostic tools at our disposal we have determined this model to be valid.

Our final model utilizes 23 predictor and one interaction variable in order to fit a model, of which all but four variables are statistically significant based on the protected t-tests. The model yields predictions which have percentage errors in the interquartile range of -6.2% to 9.5%.

Included in this report are the methodology and reasoning behind all transformations or conversions to dummy variables from the original data as well as supporting figures and references for all claims. In addition, it includes an interpretation of all coefficients represented in the final model, as well as analyses of over and under valued properties in the data, and an explanation of combinations of predictors which should produce the highest and lowest realistic value.

Our model does have weaknesses including 39 outliers, one leverage point and some high VIFs among the categorical variables. We provide explanation for the weaknesses where they are evident and discuss areas where we need to seek out further subject matter expertise.

Introduction

This report is a documentation of the multivariate regression model created to predict *Assessed Value* from predictors for real estate properties in Ward 3 of the City of Boston. The data set is public data provided by the City of Boston. This report includes the final model generated from the supplied data set, diagnosis of the model, identified outliers and leverage points, our interpretation of the model, and an explanation of weaknesses and shortcomings of the model.

- A Summary of Effect Tests for all predictor variables in the model is shown in Fig. 8 on page 11 in Appendix A.1.
- Fig. 9 on page 12 in Appendix A.2, shows a set of plots for studentized residuals plotted against all predictor variables in the model. Constant variance in all the plots proves the validity of this model.
- Fig. 10 in Appendix A.3 on page 13, shows the accuracy of the model prediction.

1. Model to Predict Assessed Value of Condominium Unit in Boston (Ward-3)

1.1 Model Construction Steps

- We applied a Log transformation to *Assessed Value* and *Living Area*, due to the data's shape across multiple orders of magnitude, and the non-constant error variance noticed in the residual plots.

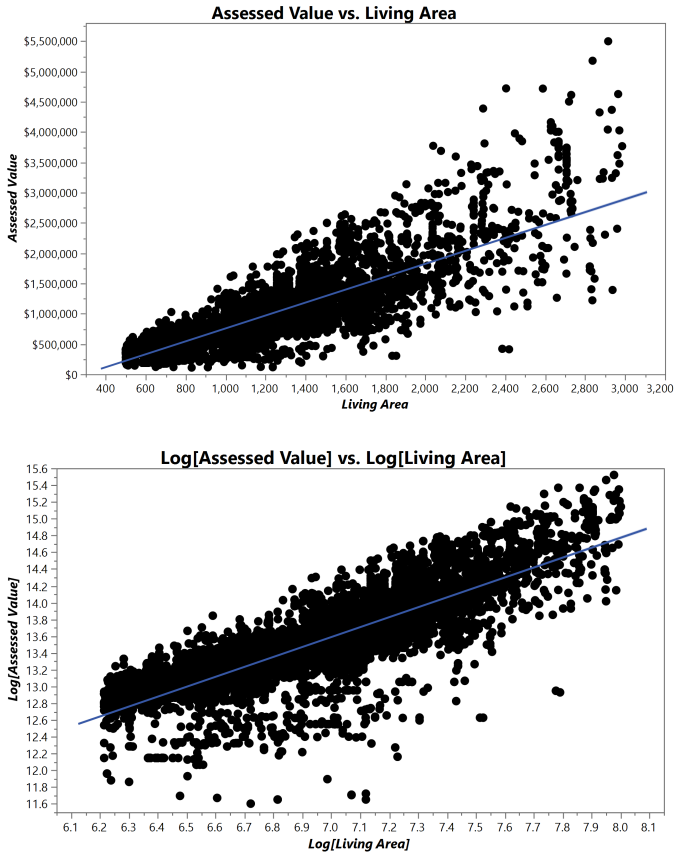
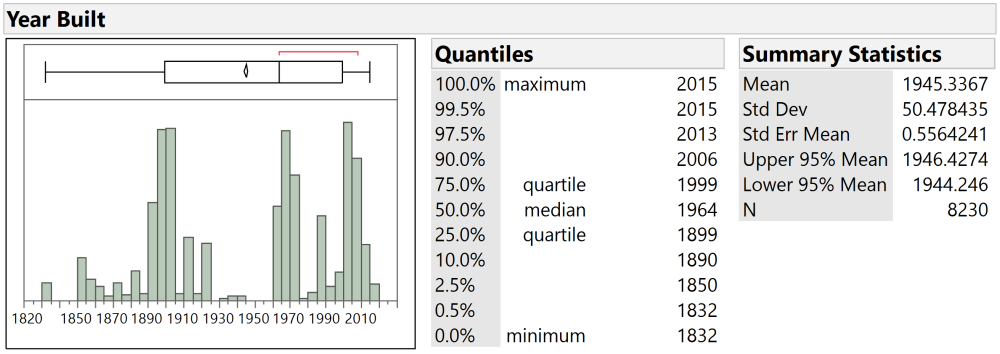


Fig. 1: Log transformation of *Assessed Value* and *Living Area*

- We identified a highly complex distribution to *Year Built*. From the distribution we can see that there were clearly impacts on the amount of housing construction from the Depression and from World War II. Due to its complexity, we chose to use the method of "binning", and chose the bin size of 25 years through trial and error. 25 years proved a small enough bin size to accurately represent short term building trends and practices, but not so large as to create too many categories in the variable.



Quantiles		
100.0%	maximum	2015
99.5%		2015
97.5%		2013
90.0%		2006
75.0%	quartile	1999
50.0%	median	1964
25.0%	quartile	1899
10.0%		1890
2.5%		1850
0.5%		1832
0.0%	minimum	1832

Summary Statistics	
Mean	1945.3367
Std Dev	50.478435
Std Err Mean	0.5564241
Upper 95% Mean	1946.4274
Lower 95% Mean	1944.246
N	8230

Fig. 2: Distribution of *Year Built*

- Similar to *Year Built*, we noticed a non-normal distribution to *Base Floor*, and decided the effect of *Base Floor* is better modelled as a binned variable than a continuous. Other transformation methods were ineffective; we couldn't apply a log transformation due to values of

zero being present, and a square root transformation created coefficients which were difficult to logically interpret. Trial and error led us to use a bin size of 2 floors.

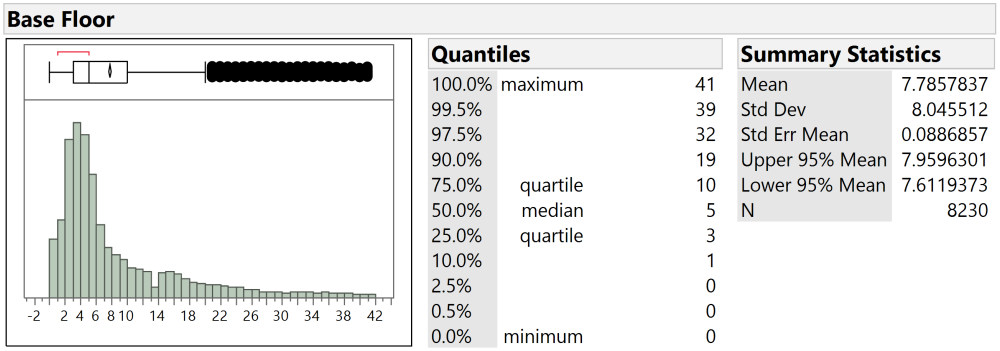


Fig. 3: Distribution of *Base Floor*

- We added an interaction term between *Zip Code* and *Log[Living Area]*. We chose this interaction because of the clear variations in slope between different *Zip Codes*, and the fact that both *Log[Living Area]* and *Zip Code* were among the most important predictors with the most logically intuitive relationship.

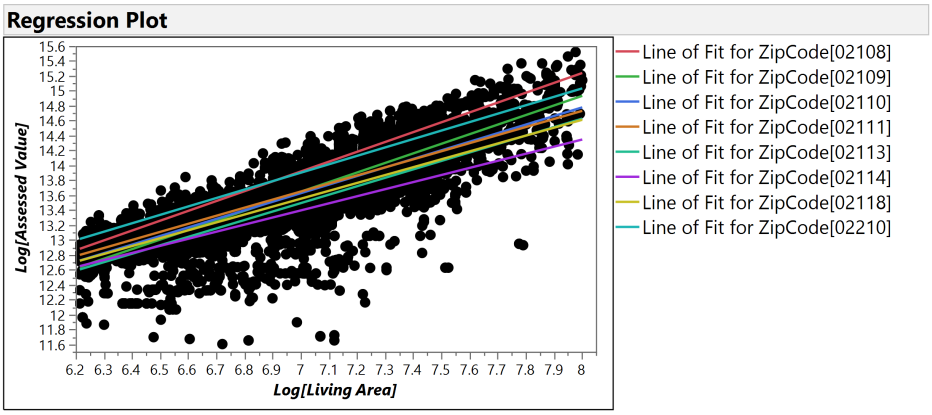


Fig. 4: Interaction between *Zip Code* & *Log[Living Area]*

- We created the predictor **Years Since Renovation**, which is the year 2017 - MAX(*Year Built* or *Year Renovated*). This takes into effect a continuous quantitative measure of when the last renovation (or none) occurred.
- To simplify the effects of *Air Conditioning* and *Half Baths*, we created dummy variables to represent the binary effect of these conditions if present or not.

1.2 Predictors Used For The Model

Some of the predictors from given data-set were modified/transformed for the model as shown in Table 1 on the next page. Other predictors provided for the model were used as shown here:

Categorical: *Interior Condition*, *Interior finish*, *Kitchen Style*, *Main Bathroom Style*,*Number of bedrooms*, *Number of Fireplaces*, *Number of Floors*, *Number of full baths*,*Number of Parking Spaces*, *Orientation of unit within the building*, *Type of Heating*,*Type of Kitchen* & *View*

Continuous: *Total Number of Rooms*

Table 1: Table showing predictors which were modified/transformed for the model

Type	Predictor	Notes
Categorical	Base Floor Binned (Size: 2)	***Base Floor*** predictor was binned with a size of 2
	Year Built Binned (Size: 25)	***Year Built*** predictor was binned with a size of 25
Indicator	Has 1 or more half baths?	Because the count of units for ***Number of half baths*** was distributed as: 6690 units with 0, 1529 units with 1 & 11 units with 2, this predictor was converted to an indicator variable showing if the unit has 1 or more half baths.
	Has Central A/C?	***Air conditioning*** predictor was converted to an indicator variable showing if the unit has "Central A/C".
	Is Owner Occupied?	***Owner Occupied*** predictor was converted to an indicator variable showing if the unit is owner occupied.
	Is Unit located in the corner of building?	***Unit located in the corner of the building*** predictor was converted to an indicator variable showing if the unit is located in the corner of the building.
Continuous	Log[Living Area]	***Living Area*** was Log Transformed (see Fig. 12 on page 14)
	Years Since Renovation	***Year Remodeled*** predictor was converted to *Years Since Renovation*, by subtracting year when the unit was remodeled from current year. In cases where the unit was "Unrenovated", year of build was used.
Interaction	Zip Code * Log[Living Area]	***Zip Code*** was crossed with ***Log[Living Area]*** to create an interaction variable

2. Marginal Model Plots

Marginal Model Plots were created using *Log[Assessed Value]* and *Predicted Log[Assessed Value]* for 3 continuous predictors in our model namely, *Log[Living Area]*, *Years Since Renovation* & *Total Number of Rooms*, using 0.5 smoothness for comparison:
Marginal Model Plots shown in Fig. 5 on the next page indicate that each of the continuous predictor variables is correctly specified in the model as demonstrated by the overlap between predicted and actual lines.

3. Variance Inflation Factors (VIFs)

For all three continuous predictors used in the model, VIFs are well below threshold of 10 as shown below in Fig. 6 on the following page, hence there are no issues with multicollinearity.

4. Studentized Residual & Cook's Distance

We have one point that is an influential point. Property 0303598012 has a Cook's Distance 216.15 as shown in Fig. 7 on page 6. Upon examining the data, we see that this is because this is the only property with an *Interior Condition* of "Poor". Since this is the only data point with a value in that category, it necessarily has a high influence of the prediction for that category. If "Poor" *Interior Condition* had a similar coefficient and p-value to another category in *Interior Condition*, we would try to combine it with that category. However, since it does not, we will leave it as is

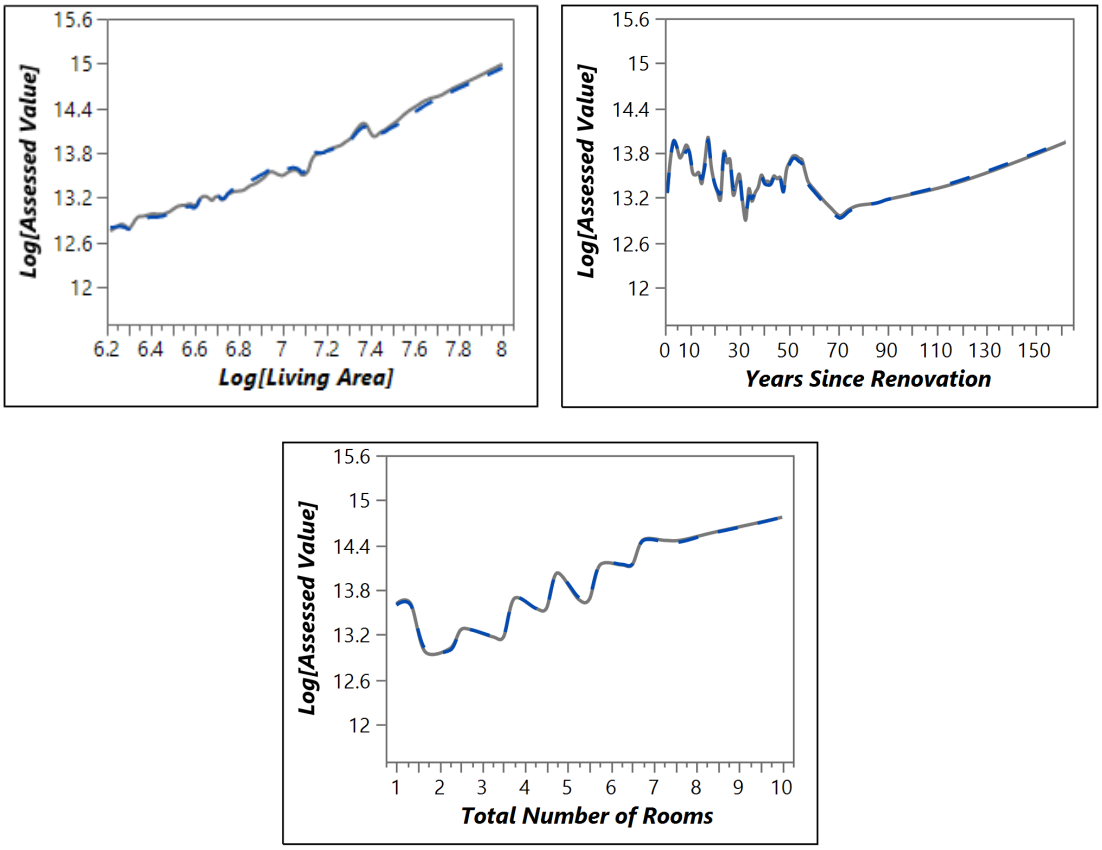


Fig. 5: Marginal Model Plots for continuous predictors

Parameter Estimates

Term	Estimate	Prob>|t|	VIF
Years Since Renovation	0.0004575	0.0006*	1.5335513
Total Number of Rooms	0.0118185	0.0009*	3.7735244
Log[Living Area]	0.7205773	<.0001*	5.9924999

Fig. 6: VIFs for continuous predictors

and use caution when predicting assessed value for future properties with an interior condition of "Poor".
Zooming in on the horizontal axis we see that there are no other influential points since they all have Cook's Distances of well below 1.

Looking closer at studentized residuals, we see that 39 out of the 8,230 records have values which are less than -5. These are the outliers for which we do not have a good explanation despite input from subject matter experts. This would be a good subject for potential further research.

5. Effect of Living Area on Assessed Value across Zip Codes

When running the analysis, it was determined that *Log[Living Area]* predictor had a varying impact on *Assessed Value* across different *Zip Codes*, so we determined that it would be a good idea to cross *Zip Code* with *Log[Living Area]*. Table. 2 on the following page shows us coefficients of the predictors from this interaction. In this table, values under **Sum** column are sum of the coefficients

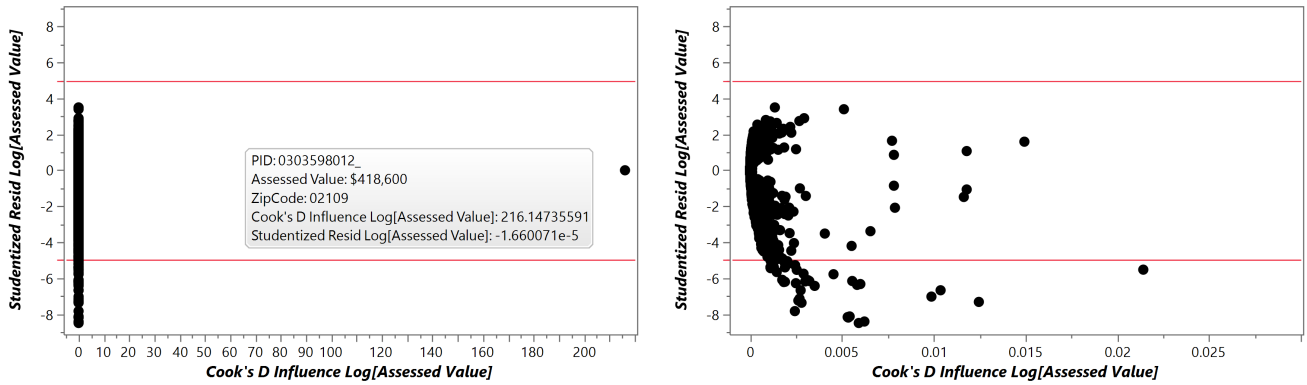


Fig. 7: Studentized Residual & Cook's Distance

for *Log[Living Area]* and *Log[Living Area] * Zip code*. **% Change in Assessed Value** column transforms the coefficient into a percentage change of *Assessed Value* for every 1% change in *Living Area*.

Table 2: Table showing effect of *Living Area* on *Assessed Value* in different *Zip Codes*

Term	**Estimate**	**P-Value**	**Statistically Significant**	**Sum**	**% Change in Assessed Value**
Log[Living Area]	0.7205773	<.0001	Yes	-	-
Log[Living Area]*ZipCode[02110]	0.1346888	<.0001	Yes	0.8552661	0.8546%
Log[Living Area]*ZipCode[02118]	0.0993966	<.0001	Yes	0.8199739	0.8192%
Log[Living Area]*ZipCode[02111]	-0.072204	<.0001	Yes	0.6483733	0.6472%
Log[Living Area]*ZipCode[02108]	-0.145758	<.0001	Yes	0.5748193	0.5736%
Log[Living Area]*ZipCode[02109]	0.0223803	0.1825	No	-	-
Log[Living Area]*ZipCode[02113]	0.0158735	0.4663	No	-	-
Log[Living Area]*ZipCode[02114]	0.0147767	0.428	No	-	-
Log[Living Area]*ZipCode[02210]	-0.069154	0.1754	No	-	-

Half of the interaction terms between Log[Living Area] and *Zip Codes* are statistically significant, these are the ones associated with *Zip Codes*: 02108, 02110, 02111 & 02118. Other *Zip Codes* do not have a statistically significant impact on the effect of *Log[Assessed Value]* due to their high P-Values.

For every 1% increase of the *Living Area* in the associated *Zip Code*, price of the home will increase by the value shown in **% Change in Assessed Value** column. For example, if a home is in *Zip Code* 02108, for every 1% increase in *Living Area*, *Assessed Value* of the home will increase by only 0.5736%. In 02110, for every 1% increase in *Living Area*, *Assessed Value* of the home will increase by 0.8546%. So in Boston's Ward 3, 02108 is the least valued *Zip Code* per square foot while 02110 is the most valued *Zip Code* per square foot.

6. Interpretation of Coefficients

P-values shown in the model's overall effect tests were used to first validate the coefficient's statistical significance. To decide if there was statistical significance in the predictors tested, a standard 95% confidence level and .05 alpha applied, so p-values of less than 0.05 were required for each global predictor. Four global predictors did not pass the p-value effect test requirement, which meant we

also ignored expanded predictor effects in categorical variables (Main Bathroom Style and Number of bedrooms). Variables with acceptable p-values and with categorical data were explored in the expanded estimates. Each subcategory predictor was subjected to the same p-value test and resulted in either an ignored or considered predictor based its statistical significance. Converting the effects of predictor variables into meaningful effects against the model's dependent variable (y = logValue), involved an algebraic transformation of the effect estimate. If the predictor was also logged, then a % change in X meant a % change in the dependent variable. Otherwise, the resulting effects for originally non transformed predictors meant a unit change in X resulted in a % change in the dependent variable. See description below for added details. See Table 3 for the interpretation of each predictor's effect in the model. Table 8 on page 15 in Appendix A.5 lists the coefficients for all statistically non-significant variables in the model.
Table 3 lists all statistically significant predictors in the model along with their estimates, P-Values and effect on *Assessed Value*. In this table, values in column **Change in X** should be interpreted as:

- **For Categorical Variables:** **1** ⇔ "Category present in the Condominium"
- **For Indicator Variables:** **1** ⇔ "Variable present in the Condominium"
- **For Continuous Variables:**
  - **Log[Living Area]:** **1%** ⇔ "1% Change in the *Log[Living Area]*"
  - **All Others:** **1** ⇔ "1 Unit Change in the Variable"
- **For Interaction Variables of *Log[Living Area]* & *Zip Code*:** **1%** ⇔ "1% change in *Living Area* for *Zip Code* of that interaction variable"

Table 3: Table showing interpretation of coefficients for statistically significant variables

Term	Estimate	P-Value	Change in X	Change in Assessed Value (%)
Log[Living Area]	0.7205773	<.0001*	1%	0.7196
Number of Parking Spaces[0]	-0.214392	<.0001*	1	-19.2968
Number of Parking Spaces[2]	0.0937193	<.0001*	1	9.8251
View[Special]	0.217196	<.0001*	1	24.2588
View[Excellent]	0.157452	<.0001*	1	17.0525
View[Average]	-0.048749	<.0001*	1	-4.7580
View[Fair]	-0.147367	<.0001*	1	-13.7023
View[Poor]	-0.19108	<.0001*	1	-17.3933
Orientation of unit within the building[End]	-0.107863	<.0001*	1	-10.2249
Orientation of unit within the building[Middle]	-0.054705	<.0001*	1	-5.3236
Orientation of unit within the building[Rear above]	0.027383	0.0003*	1	2.7761
Orientation of unit within the building[Through]	0.1068303	<.0001*	1	11.2745
Interior condition[Excellent]	0.1660354	<.0001*	1	18.0615
Year Built Binned (Size: 25)[1830 - 1855]	0.0753185	<.0001*	1	7.8228
Year Built Binned (Size: 25)[1880 - 1905]	-0.059307	<.0001*	1	-5.7583
Year Built Binned (Size: 25)[1955 - 1980]	-0.056394	<.0001*	1	-5.4833
Year Built Binned (Size: 25)[1980 - 2005]	-0.025009	0.0013*	1	-2.4699
Year Built Binned (Size: 25)[2005 - 2030]	0.0248999	0.0095*	1	2.5212
Log[Living Area]*ZipCode[02108]	-0.145758	<.0001*	1%	0.5736
Log[Living Area]*ZipCode[02110]	0.1346888	<.0001*	1%	0.8546
Log[Living Area]*ZipCode[02111]	-0.072204	<.0001*	1%	0.6472

Log[Living Area]*ZipCode[02118]	0.0993966	<.0001*	1%	0.8192
ZipCode[02108]	0.095982	<.0001*	1	10.0739
ZipCode[02113]	-0.069985	<.0001*	1	-6.7592
ZipCode[02114]	-0.04711	<.0001*	1	-4.6018
ZipCode[02118]	0.0305215	<.0001*	1	3.0992
Number of fireplaces[0]	-0.104586	0.0024*	1	-9.9303
Type of heating[Electric]	0.0335975	0.0007*	1	3.4168
Type of heating[Forced Air]	-0.049535	<.0001*	1	-4.8328
Type of heating[Hot Water]	-0.043922	<.0001*	1	-4.2971
Number of full baths[1]	-0.150665	<.0001*	1	-13.9864
Number of full baths[2]	-0.102832	<.0001*	1	-9.7721
Number of full baths[4]	0.1561618	<.0001*	1	16.9015
Kitchen Style[Luxury]	0.111319	<.0001*	1	11.7751
Kitchen Style[Semi-modern]	-0.032903	0.0003*	1	-3.2368
Kitchen Style[No remodeling]	-0.078804	<.0001*	1	-7.5779
Interior finish[Elaborate]	0.1093288	<.0001*	1	11.5529
Interior finish[Substandard]	-0.094491	0.0022*	1	-9.0164
Type of kitchen[Full eat-in]	0.0434896	0.0002*	1	4.4449
Has 1 or more Half Baths?	0.0452187	<.0001*	1	4.6257
Base Floor Binned (Size: 2)[0 - 2]	-0.052347	<.0001*	1	-5.1000
Years Since Renovation	0.0004575	0.0006*	1	0.0458
Is Owner Occupied?	-0.016913	0.0002*	1	-1.6771
Total Number of Rooms (Continuous)	0.0118185	0.0009*	1	1.1889
Number of Floors[1]	-0.066439	0.0043*	1	-6.4280
Number of Parking Spaces[3]	0.1314191	0.0120*	1	14.0446
Year Built Binned (Size: 25)[1905 - 1930]	-0.021536	0.0118*	1	-2.1306
Base Floor Binned (Size: 2)[8 - 10]	-0.024915	0.0180*	1	-2.4607
Base Floor Binned (Size: 2)[10 - 12]	-0.024372	0.0311*	1	-2.4077
Base Floor Binned (Size: 2)[24 - 26]	0.0414907	0.0246*	1	4.2363
Number of Floors[2]	-0.055157	0.0164*	1	-5.3663

7. Over & Under Valued Properties

Table 4 lists top 10 over-priced condominiums in dollars based on prediction error ($). Although the studentized residuals for these properties can be considered reasonable for a large sample size, the error exceeds $1 million. Something to note on these properties is that at least 3 of them are called out as penthouses. In addition, the properties that are on Battery ST or Battery Wharf are in buildings that border the water. The model may benefit from additional variables that account for the address or proximity to the water.

Table 4: Top-10 over-priced condominiums based on Prediction Error ($)

CM_ID	Address	Assessed Value	Predicted Assessed Value	Prediction Error ($)	Prediction Error (%)
0303040000_	50 BATTERY ST, Unit: PH9, Zip: 02109	$4,720,250	$2,492,773	$2,227,477	47%
0304705000_	45 PROVINCE ST, Unit: PH-1B, Zip: 02108	$5,498,820	$3,717,468	$1,781,352	32%
0303041010_	2 5 BATTERY WHARF, Unit: 4611, Zip: 02109	$4,717,235	$2,936,645	$1,780,590	38%
0303038200_	27 UNION WH, Unit: 27, Zip: 02109	$3,811,654	$2,310,001	$1,501,653	39%
0302347000_	44 PRINCE ST, Unit: 100, Zip: 02113	$3,319,218	$1,920,814	$1,398,404	42%
0303041010_	2 5 BATTERY WHARF, Unit: 3311, Zip: 02109	$3,772,137	$2,419,248	$1,352,889	36%
0303038200_	21 UNION WH, Unit: 21, Zip: 02109	$4,387,838	$3,085,955	$1,301,883	30%
0303041010_	2 5 BATTERY WHARF, Unit: 3411, Zip: 02109	$3,688,459	$2,511,184	$1,177,275	32%
0304705000_	45 PROVINCE ST, Unit: PH-1A, Zip: 02108	$5,177,196	$4,008,715	$1,168,481	23%
0303028300_	63 COMMERCIAL WHARF EAST, Unit: 63-6-8, Zip: 02110	$2,879,384	$1,714,472	$1,164,912	40%

Table 5 lists top 10 under-priced condominiums based on prediction error ($). Since, the first three share an address there maybe a local affordable housing policy that may be impacting the assessed value for a certain amount of time. Remaining seven have very reasonable studentized residuals however prediction error associated with them is well over $900,000. Unit on 1 Avery St. is particularly interesting because it is a penthouse.

Table 5: Top-10 under-priced condominiums based on this model

CM_ID	Address	Assessed Value	Predicted Assessed Value	Prediction Error ($)	Prediction Error (%)
0306531000_	522 524 HARRISON AV, Unit: 4, Zip: 02118	$419,848	$1,687,011	($1,267,163)	-302%
0306531000_	522 524 HARRISON AV, Unit: 3, Zip: 02118	$419,848	$1,598,366	($1,178,518)	-281%
0306531000_	522 524 HARRISON AV, Unit: 1, Zip: 02118	$412,000	$1,463,749	($1,051,749)	-255%
0303474000_	120 FULTON ST, Unit: 5B, Zip: 02109	$1,684,084	$2,700,781	($1,016,697)	-60%
0306954000_	9 UPTON ST, Unit: 1, Zip: 02118	$2,083,162	$3,093,760	($1,010,598)	-49%
0303512000_	220 COMMERCIAL ST, Unit: 2, Zip: 02109	$1,661,550	$2,656,395	($994,845)	-60%
0303747000_	100 STATE ST, Unit: 5, Zip: 02109	$1,520,000	$2,484,708	($964,708)	-63%
0305729000_	21 DWIGHT ST, Unit: 1, Zip: 02118	$2,020,000	$2,979,425	($959,425)	-47%
0302953018_	500 ATLANTIC AV, Unit: 16N, Zip: 02210	$2,572,583	$3,480,112	($907,529)	-35%
0304832020_	1 AVERY ST, Unit: PH 2B, Zip: 02111	$3,478,000	$4,382,188	($904,188)	-26%

8. Combination of Predictors to Produce High & Low Assessed Values

Based on the coefficients which are statistically significant at both the predictor level and for individual categorical variables, following are the combinations of predictors which we can say should produce the highest and lowest assessed values respectively, based on our model. These combinations of predictors were determined by calculating a weighted measure of all statistically significant coefficients for each property. These represent examples of real world condominiums.

Table 6: Predictors for Highest Value (Example, CM_ID: 0304705242)

Term	Value
Living Area	2872
Number of Parking Spaces	2
View	Excellent
Orientation of unit within the building	Through
Interior Condition	Excellent
Year Built	2005-2030
Zip Code	02108
Kitchen Style	Luxury
Interior finish	Elaborate
Has 1 or more Half Baths?	Yes (1)
Base Floor	24th or 25th
Years Since Renovation	10
Is Owner Occupied?	No (0)
Total Number of Rooms	5

Table 7: Predictors for Lowest Value (Example, CM_ID: 0302263002)

Term	Value
Living Area	516
Number of Parking Spaces	0
View	Average
Year Built	1880 - 1905
Zip Code	02113
Number of Fireplaces	0
Type of Heating	Hot Water
Number of full baths	1
Kitchen Style	No Remodeling
Interior finish	Substandard
Has 1 or more Half Baths?	No (0)
Base Floor	0 or 1st
Years Since Renovation	30
Is Owner Occupied?	Yes (1)
Total Number of Rooms	3

9. Weakness in Model

The model has displayed some weaknesses, specifically:

- High VIFs in some of the categorical values. The highest of which occur in some of the Interior Condition categories, Number of fireplaces, and Number of full baths.
- There are 39 outliers with a standard deviation below -5, so the error percentage on these was high. Ideally more variables or guided knowledge of the existing variables would be applied to reduce the number of outliers. We spoke with a subject matter expert who lives in this ward, and she informed us that the city mandates that a certain number of "affordable housing" units are available to lower income residents. The Boston city website, which is the source of this data, has guidelines for defining a maximum list price for affordable housing units, which is a function of both number of bedrooms and resident household income. Since our database does not include resident income, we cannot fully incorporate this factor into our model.
- There is a point with a High Cook's Distance as discussed earlier in Section 4. on page 4 it is a consequence of only one of our properties having a "Poor" interior condition. However, the Y-behavior is consistent with the other observations so this is not a bad leverage and would not be expected to impact our models coefficients.
- The coefficients for different categories in *Type of heating* that are statistically significant are not intuitive. There are coefficients that are unexpectedly negative, such as *Type of heating[Hot Water]*. In addition, the model shows the Space Heater category as statistically not significant. Intuitively, one would expect that if the heat source was only a Space Heater in Boston, it would be statistically significant and negative.
- The coefficient for *Years Since Renovation* is statistically significant and positive, which implies that an increase in time since renovation adds value to the property. This is counterintuitive to our logical understanding of the real estate model. This might be because so many other variables also indicate renovated status, such as *Kitchen Style*, *Interior finish*, *Interior condition*, etc.

A Appendix

A.1 Summary of Effect Tests

Response Log[Assessed Value]

Summary of Fit

RSquare	0.879894
RSquare Adj	0.878447
Root Mean Square Error	0.192064
Mean of Response	13.47015
Observations (or Sum Wgts)	8230

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	98	2197.3780	22.4222	607.8347
Error	8131	299.9419	0.0369	Prob > F
C. Total	8229	2497.3199		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Log[Living Area]	1	1	109.07138	2956.771	<.0001*
Number of Parking Spaces	3	3	38.62676	349.0389	<.0001*
View	5	5	20.91429	113.3914	<.0001*
Orientation of unit within the building	6	6	13.50951	61.0373	<.0001*
Interior condition	4	4	10.20080	69.1323	<.0001*
Year Built Binned (Size: 25)	7	7	6.42928	24.8984	<.0001*
Log[Living Area]*ZipCode	7	7	5.27526	20.4293	<.0001*
ZipCode	7	7	5.16162	19.9892	<.0001*
Number of fireplaces	5	5	4.51668	24.4881	<.0001*
Type of heating	4	4	4.18161	28.3394	<.0001*
Number of full baths	4	4	3.88781	26.3483	<.0001*
Kitchen Style	3	3	3.22887	29.1767	<.0001*
Interior finish	2	2	2.99389	40.5801	<.0001*
Type of kitchen	3	3	2.37785	21.4868	<.0001*
Has 1 or more Half Baths?	1	1	1.35582	36.7544	<.0001*
Base Floor Binned (Size: 2)	20	20	2.95645	4.0073	<.0001*
Years Since Renovation	1	1	0.43598	11.8187	0.0006*
Is Owner Occupied?	1	1	0.52808	14.3155	0.0002*
Total Number of Rooms	1	1	0.40914	11.0912	0.0009*
Number of Floors	3	3	0.42193	3.8126	0.0096*
Main Bathroom Style	3	3	0.23525	2.1258	0.0947
Number of bedrooms	5	5	0.31748	1.7213	0.1260
Has Central A/C?	1	1	0.00286	0.0775	0.7807
Is Unit located in the corner of the building?	1	1	0.00081	0.0218	0.8825

Fig. 8: Model to estimate *Assessed Value* of condominium units in Boston (Ward-3)

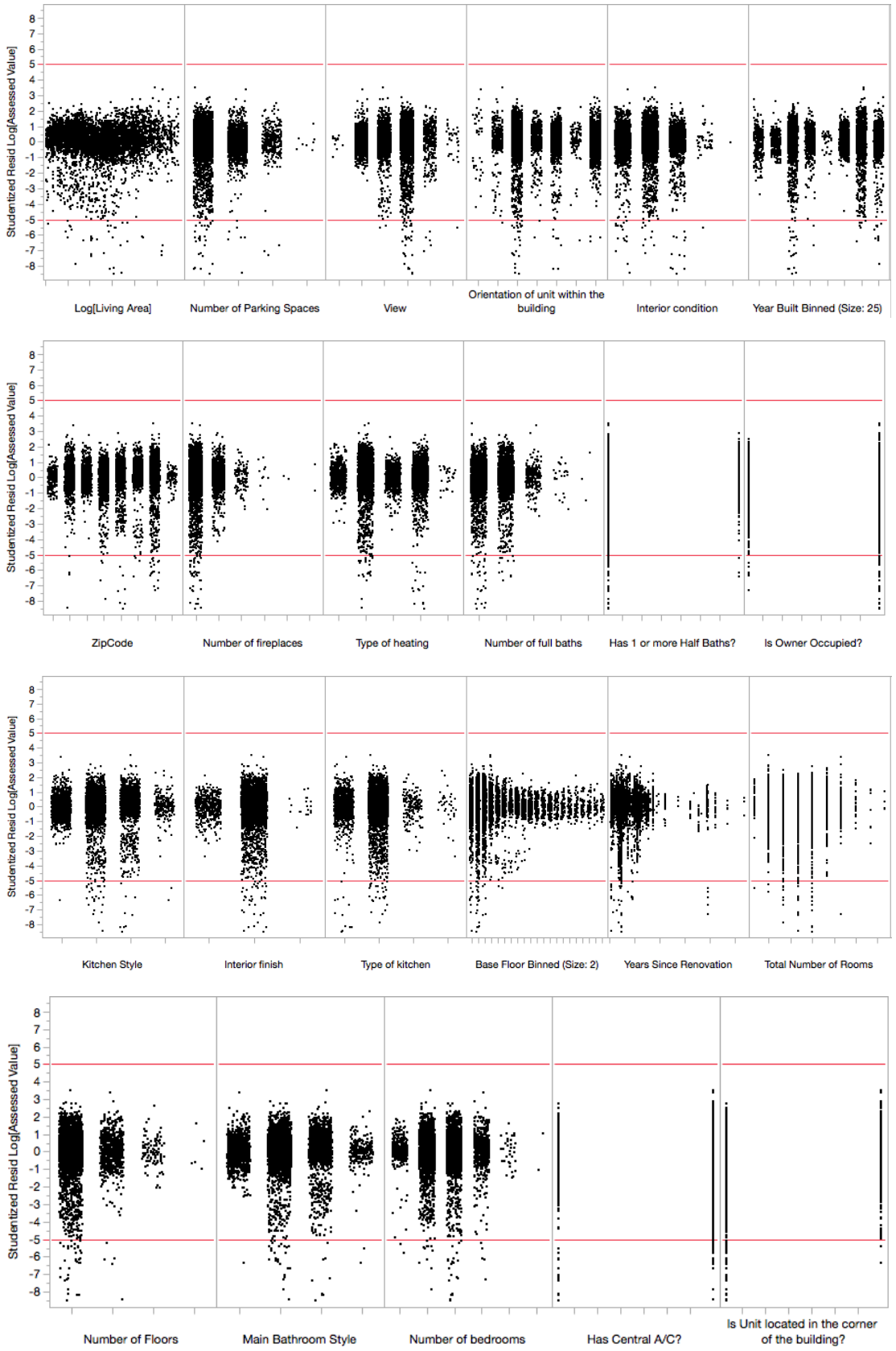


Fig. 9: Scatter plots showing Studentized Residuals against all predictors

A.2 Validity Of Model (Studentized Residual Plot against Predictors)

A.3 Comparison of Actual and Predicted Values - Accuracy of the Model

Prediction Error in Assessed Value ($)

Quantiles		
100.0%	maximum	$2,227,477
75.0%	quartile	$68,342
50.0%	median	$7,077
25.0%	quartile	($41,560)
0.0%	minimum	($1,267,163)

Prediction Error in Assessed Value (%)

Quantiles		
100.0%	maximum	48.7%
75.0%	quartile	9.5%
50.0%	median	1.2%
25.0%	quartile	-6.2%
0.0%	minimum	-408.1%

Fig. 10: Prediction error of this model

A.4 Reason for using Log Transform

Log Transform for Assessed Value: *Assessed Value* was log transformed because of skewness in the data. Fig. 11 shows the effect of applying log transform.

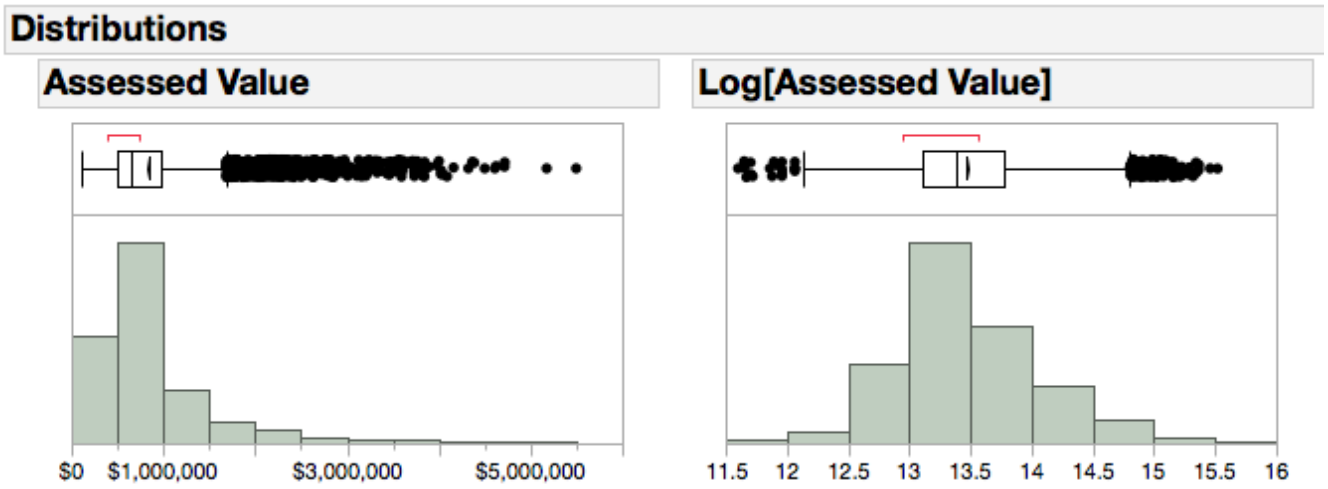


Fig. 11: Effect of applying log transform to ***Assessed Value***

Log Transform for Living Area: *Living Area* was log transformed because of skewness in the data. Fig. 12 on the following page shows the effect of applying log transform.

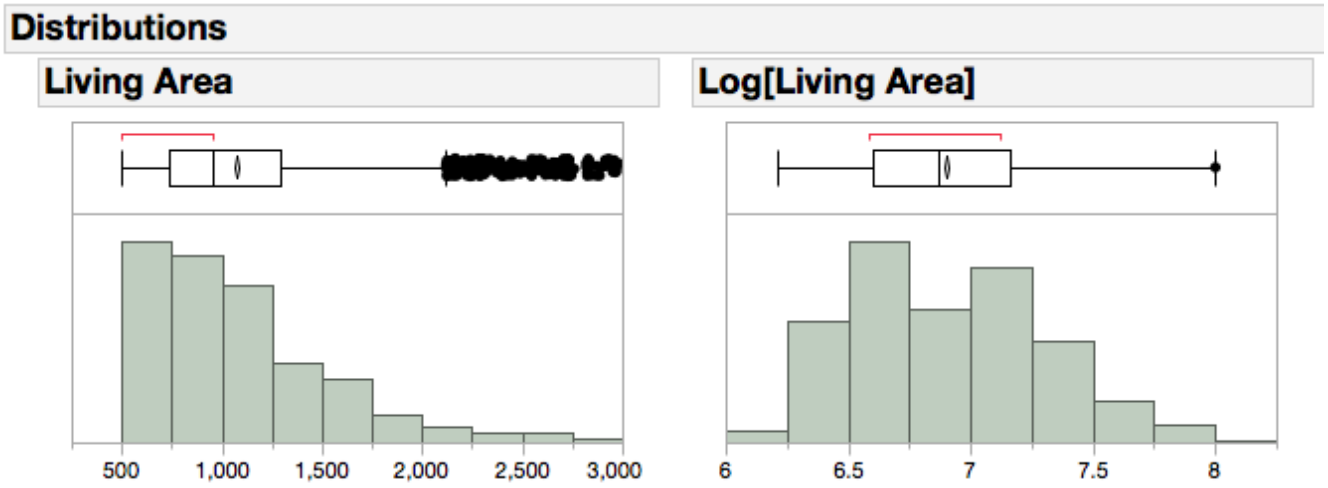


Fig. 12: Effect of applying log transform to ***Living Area***

A.5 Statistically non-significant variables in the model

Table 8: Table showing coefficients of statistically non-significant variables

Term	Estimate	P-Value	Change in X	Change in Assessed Value (%)
Number of Parking Spaces[1]	-0.010746	0.5513	-	-
View[Good]	0.0125483	0.3088	-	-
Orientation of unit within the building[Face courtyard]	0.0007349	0.9403	-	-
Orientation of unit within the building[Front/Street]	-0.004068	0.5328	-	-
Orientation of unit within the building[Rear below]	0.0316877	0.0605	-	-
Interior condition[Good]	0.0587551	0.1362	-	-
Interior condition[Average]	0.0110731	0.7797	-	-
Interior condition[Fair]	-0.017703	0.7034	-	-
Interior condition[Poor]	-0.218161	0.1592	-	-
Year Built Binned (Size: 25)[1855 - 1880]	0.0156279	0.1855	-	-
Year Built Binned (Size: 25)[1930 - 1955]	0.0464002	0.0864	-	-
Log[Living Area]*ZipCode[02109]	0.0223803	0.1825	-	-
Log[Living Area]*ZipCode[02113]	0.0158735	0.4663	-	-
Log[Living Area]*ZipCode[02114]	0.0147767	0.428	-	-
Log[Living Area]*ZipCode[02210]	-0.069154	0.1754	-	-
ZipCode[02109]	-0.006438	0.4262	-	-
ZipCode[02110]	0.0068122	0.4622	-	-
ZipCode[02111]	-0.010683	0.1199	-	-
ZipCode[02210]	0.0009004	0.976	-	-
Number of fireplaces[1]	-0.028029	0.4164	-	-
Number of fireplaces[2]	0.0125807	0.7364	-	-
Number of fireplaces[3]	0.0680097	0.2473	-	-
Number of fireplaces[4]	0.0570087	0.6243	-	-
Number of fireplaces[5]	-0.004984	0.9659	-	-
Type of heating[Heat Pump]	0.015637	0.1603	-	-
Type of heating[Space Heater]	0.0442217	0.1401	-	-
Number of full baths[3]	0.0079957	0.763	-	-
Number of full baths[5]	0.0893403	0.329	-	-
Kitchen Style[Modern]	0.0003878	0.9621	-	-
Interior finish[Normal]	-0.014838	0.3537	-	-

Type of kitchen[One person]	-0.009573	0.3846	-	-
Type of kitchen[Pull/alcove]	-0.017782	0.2293	-	-
Type of kitchen[None]	-0.016135	0.5918	-	-
Base Floor Binned (Size: 2)[2 - 4]	-0.010169	0.3057	-	-
Base Floor Binned (Size: 2)[4 - 6]	0.0130525	0.1726	-	-
Base Floor Binned (Size: 2)[6 - 8]	0.0149182	0.127	-	-
Base Floor Binned (Size: 2)[12 - 14]	-0.018549	0.177	-	-
Base Floor Binned (Size: 2)[14 - 16]	-0.003103	0.7894	-	-
Base Floor Binned (Size: 2)[16 - 18]	-0.013749	0.2738	-	-
Base Floor Binned (Size: 2)[18 - 20]	0.0028478	0.8483	-	-
Base Floor Binned (Size: 2)[20 - 22]	0.013881	0.3805	-	-
Base Floor Binned (Size: 2)[22 - 24]	0.0085082	0.6138	-	-
Base Floor Binned (Size: 2)[26 - 28]	0.0214661	0.3232	-	-
Base Floor Binned (Size: 2)[28 - 30]	0.0046223	0.8512	-	-
Base Floor Binned (Size: 2)[30 - 32]	0.0019571	0.9394	-	-
Base Floor Binned (Size: 2)[32 - 34]	0.0285569	0.2654	-	-
Base Floor Binned (Size: 2)[34 - 36]	-0.003291	0.8996	-	-
Base Floor Binned (Size: 2)[36 - 38]	0.0019522	0.9425	-	-
Base Floor Binned (Size: 2)[38 - 40]	-0.011925	0.7023	-	-
Base Floor Binned (Size: 2)[40 - 42]	0.009168	0.7826	-	-
Number of Floors[3]	-0.006133	0.8206	-	-
Number of Floors[4]	0.1277294	0.0509	-	-
Main Bathroom Style[Luxury]	0.0158756	0.1998	-	-
Main Bathroom Style[Modern]	-0.012053	0.1325	-	-
Main Bathroom Style[Semi-modern]	-0.019522	0.0302	-	-
Main Bathroom Style[No remodeling]	0.0156992	0.3926	-	-
Number of bedrooms[0]	-0.019371	0.4824	-	-
Number of bedrooms[1]	-0.039866	0.1095	-	-
Number of bedrooms[2]	-0.031063	0.1966	-	-
Number of bedrooms[3]	-0.034385	0.1648	-	-
Number of bedrooms[4]	-0.064868	0.0837	-	-
Number of bedrooms[5]	0.1895536	0.0989	-	-
Has Central A/C?	0.0022343	0.7807	-	-
Is Unit located in the corner of the building?	0.0009331	0.8825	-	-