Boston House Prices: Executive Summary

Khushi Arya

I. Problem Statement

Perform exploratory data analysis to understand what variables are influential in predicting house prices for the city of Boston. Identify the better linear regression model of the two built by the predecessor at BCG based on prediction results and list the best predictions.

II. Analysis

1. The best predictor for predicting house prices based on the regression model was the **living area** of the house in square feet. It had a positive correlation of 0.64 with the assessed value of the house. This makes logical sense as bigger houses with a big living area would likely have a high assessed value.

2. As expected, the **region of the house** played a significant role in its value. Houses in Jamaica Plain had the highest median assessed value of $658,000. This was followed by Cambridge at $460,557.75, Roslindale at $433,500, Dorchester Center at $385,841.16, and Hyde Park at $338,700.

3. The **age of the property** was not a significant predictor of assessed value with a weak to no association of 0.12. This is interesting as you would expect newer homes to be valued more for their amenities and lack of the need for repairs. However, this was not the case in the dataset. Some old properties were more expensive most likely because of historical significance or "antique styles". There were also a couple of outliers that were built over 300 years ago.

4. Interestingly, the properties **not remodeled** had a $71,600 higher median assessed value of $487,000. However, it should be noted that there were only 509 remodeled properties. There are also a lot of outliers on the higher end of the assessed value for properties that were not remodeled. It is likely that the remodeled houses were already expensive and/or have a higher value for other reasons.

5. Houses **remodeled after 2015**, however, had a $76,877.42 higher median assessed value of $492,793.60 than that of "other" houses. The "other" category includes houses that were never remodeled, or the ones remodeled before 2015. There were, however, a lot of outliers on the higher end of the assessed value of the "other" category.

6. The regression model's 10 best predictions were off by just **$1.52 to $25.49**. The model is comprehensive in the sense that the top predictions did not have a pattern in variables like zip code, year built, median income, age, living area etc. On average, the predictions made by the model were off by +/- $23,476.36 from the actual assessed value.

III. Model Performance

- R-square

On analyzing the model, I found that the R-square was 0.948 meaning approximately 94.8% of the variability in the assessed value of properties can be explained by our model.

- Root Mean Squared Error

On average, the predictions were approximately $32,894.53 away from the assessed value.

- Mean Absolute Error

On average, the predictions made by the model are off by +/- $23,476.36 from the actual assessed value.

- The residual analysis showed no major signs of multicollinearity. However, in the Q-Q plot, points should ideally lie on the red line if they are perfectly normally distributed. The residuals are somewhat close to normal, though there are some deviations at the tails. This suggests presence of outliers and a potential non-linear relationship. There were some outliers in the dataset in terms of houses built over 300 years ago, a million-dollar house, and others.

IV. Recommendations

1. The age of the house does not matter but recent remodeling did have a positive relationship with assessed value. The city of Boston can increase tax revenue by increasing assessed values through remodeling incentives. For instance, a tax cut for remodeling would increase the taxes on the property and the city can recover the tax cut in a couple of years.

2. The region of the house clearly played an important role in its assessed value. Houses in Jamaica Plain, for instance, had a $319,300 higher median assessed value than that of houses in Hyde Park. To eliminate this bias and tax more efficiently, the city should build different prediction models for different regions that would give more accurate predictions.

3. The current model showed outlier bias in its residual Q-Q plot. This suggests a potential non-linear relationship. The model could be improved by adding a quadratic term and/or implementing a log transformation. In simpler words, adding a log term would improve the model's accuracy and likely decrease the mean absolute error of +/- $23,476.36.

4. Update the prediction model as needed since market trends change over time. The housing market is interconnected with the economy at large and the model should consider the supply and demand forces influencing the economy at different times.

5. Analyzing the outliers like the very old and expensive houses could be valuable in terms of understanding what's driving their value. They can be assessed differently for accuracy and fairness in taxes.

Appendix

I. Null Analysis

Only two variables contained nulls, namely: land_sf (parcel's land area in square feet) and yr_remod (year property was last remodeled).

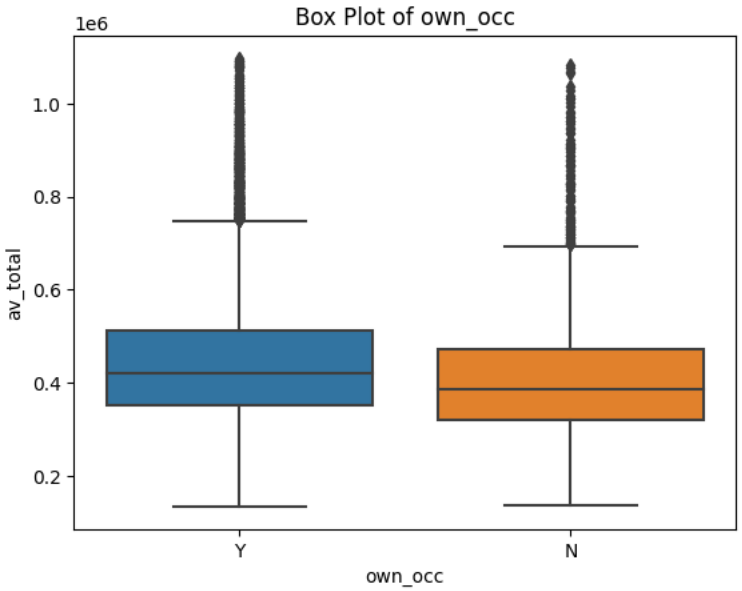
- *land_sf:* the 4 nulls appear to be recording errors, so they were dropped.
- yr_remod: Whether a house was remodeled is important in our analysis, so the nulls were not dropped. Instead, I created a binary variable 'is_remod'= 1 if the property was remodeled and 'is_remod'= 0 it wasn't remodeled.

There was one observation in the data frame with yr_built=0 which was presumably a recording error so that was dropped to only include houses built after 1700.

After making these changes, the final data frame consisted of 14,220 rows and 37 columns.

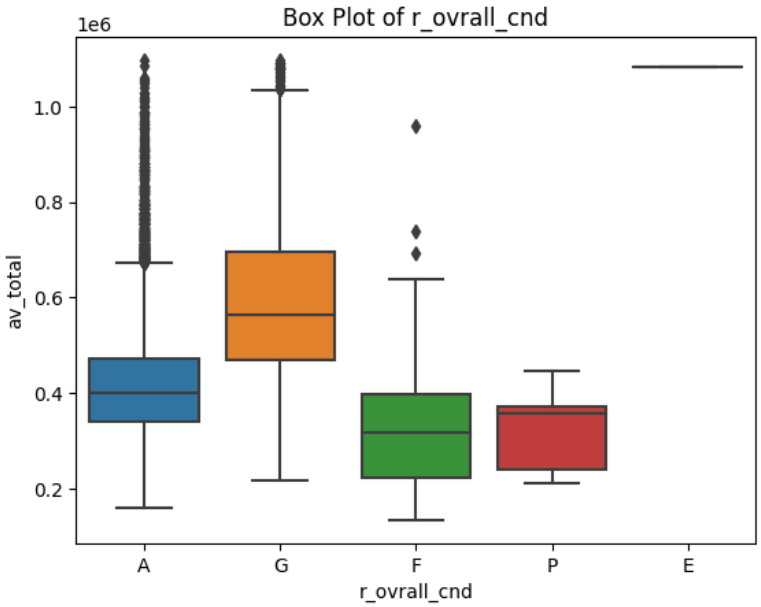
II. Categorical Analysis

1. Owner-occupied property (own_occ)



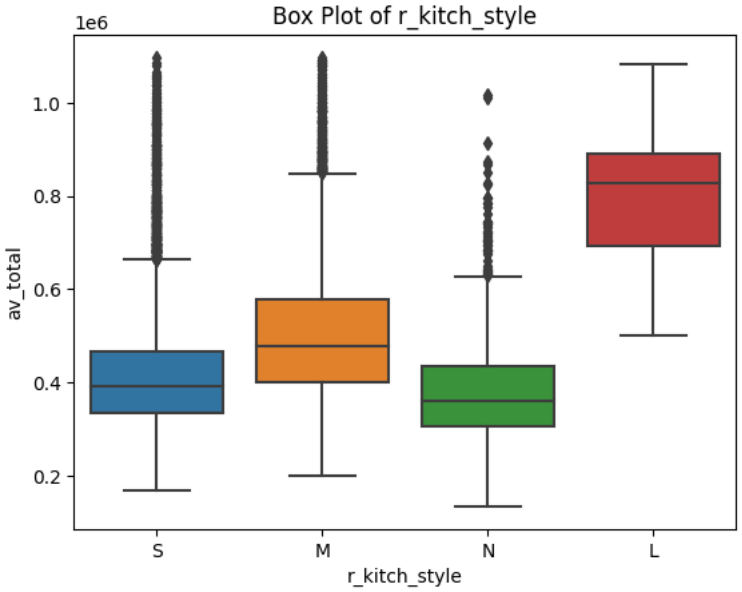
The graph above shows that owner occupied properties tend to have a higher assessed value compared to non-owner-occupied properties. The median assessed value for owner occupied properties is $420,968. 82 which is $33,731.82 higher than its counterpart. This seems to be an important determinant of the property's value. The City of Boston was correct in believing that owner occupied homes have a higher assessed value

2. Residential Overall Condition (r_ovrall_cnd)



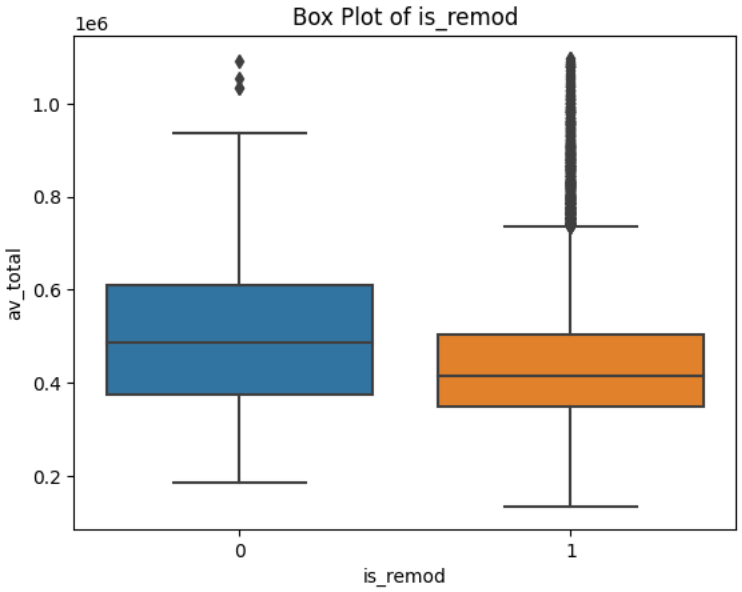
The properties were in the following conditions: Average, Good, Fair, Poor, and Excellent. As expected, the assessed value of properties in the "good" condition was the highest with a median of $564,055.62. This was followed by average, with a median of $400,300 assessed value. Most of the properties in the data frame were in these two categories with a total of 14,096 properties. It should be noted that there is 1 outlier property in excellent condition with a median assessed value of over a million dollars which would likely skew our data.

3. Residential Kitchen Style (r_kitchen_style)



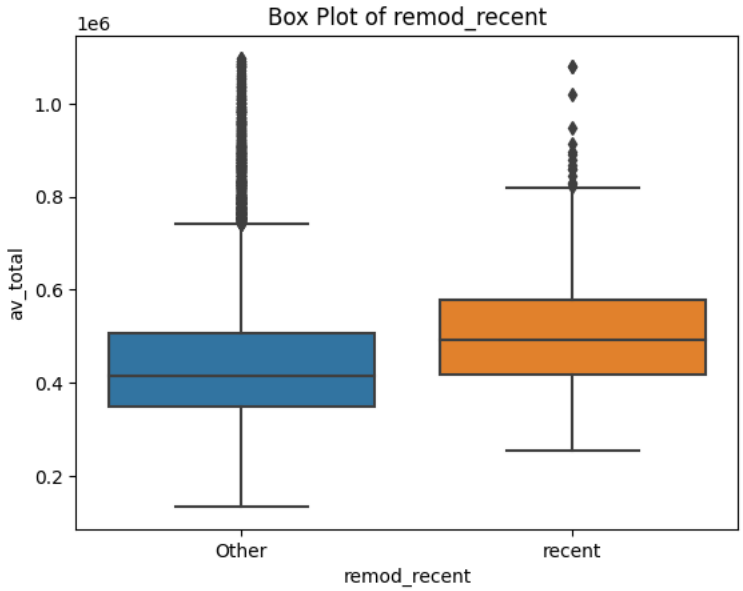
The graph above shows the assessed value of properties by their residential kitchen style. Of the four styles Semi-modern, Modern, No Remodeling, and Luxury, luxury style had the highest assessed value with a median of $826,200. This was followed by modern with a median of $477,800. It should be noted that there were a lot of outliers for all the kitchen styles except for luxury on the higher end of the assessed value. It could imply that there were other factors in those properties influencing their higher assessed value.

4. Remodeled or not (is_remod)



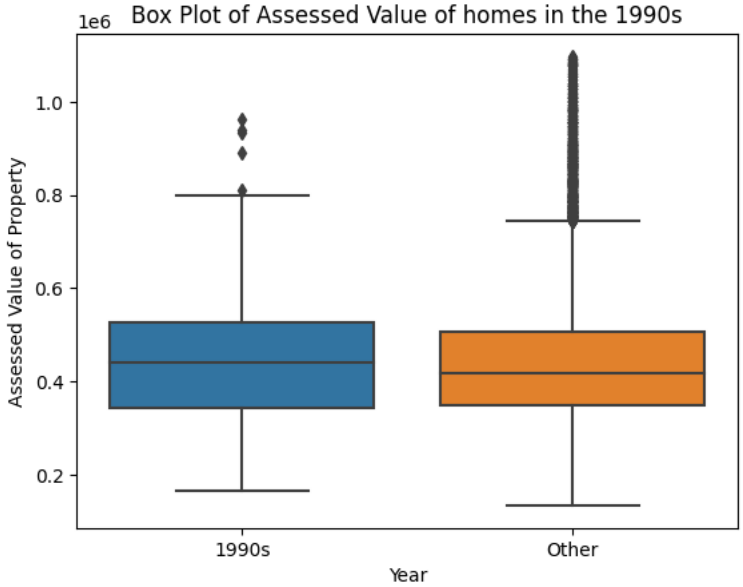
The graph above shows the assessed value of properties that were remodeled versus those that were not. Interestingly, the properties not remodeled had a $71,600 higher median assessed value of $487,000. However, it should be noted that there were only 509 remodeled properties. There are also a lot of outliers on the higher end of the assessed value for properties that were not remodeled. It is likely that the remodeled houses were already expensive and/or have a higher value for other reasons.

5. Remodeled recently



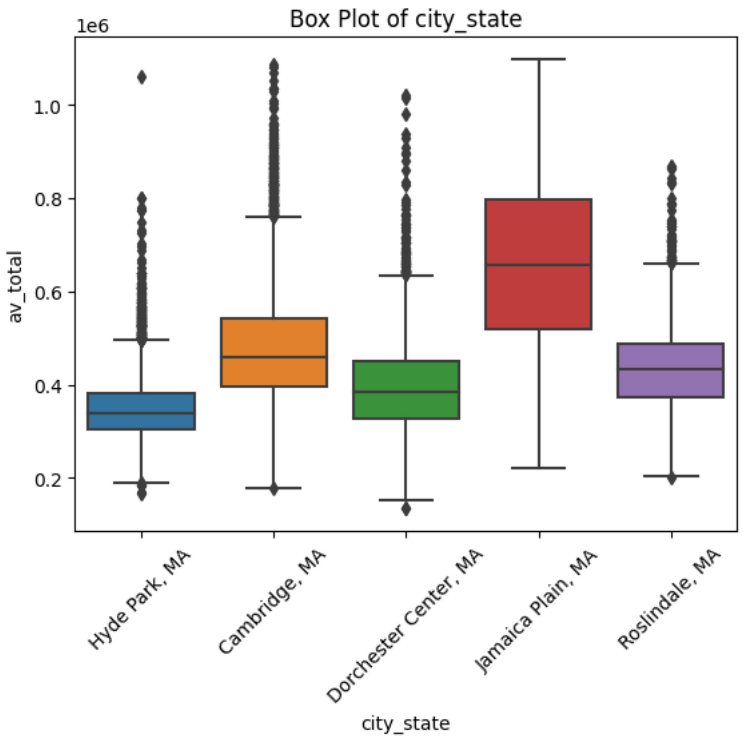
The graph above shows the assessed values of properties that were remodeled after 2015 compared to other properties remodeled earlier or not remodeled at all. Clearly, the recently remodeled houses had a median assessed value of $492,793.60 which was $76,877.42 higher than the median assessed value of the "other" houses. There were, however, a lot of outliers on the higher end of the assessed value of the "other" category. The City of Boston was correct in believing that recently remodeled houses have a higher assessed value.

6. Homes built in the 1990s



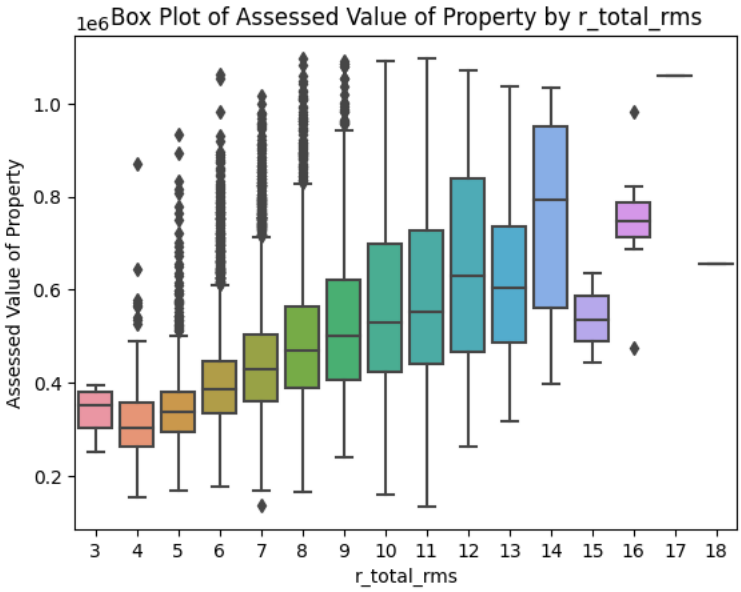
The graph above shows the assessed value of homes built in the 1990s vs other homes. The median assessed value of homes built in the 1990s at $440,246.75 is slightly ($23,466.6) higher than the median assessed value of other homes. It should be noted that there are a lot of outliers on the higher end of the median assessed value for “other” homes. Also, only 334 homes in the data set were built in the 1990s.

7. City-state



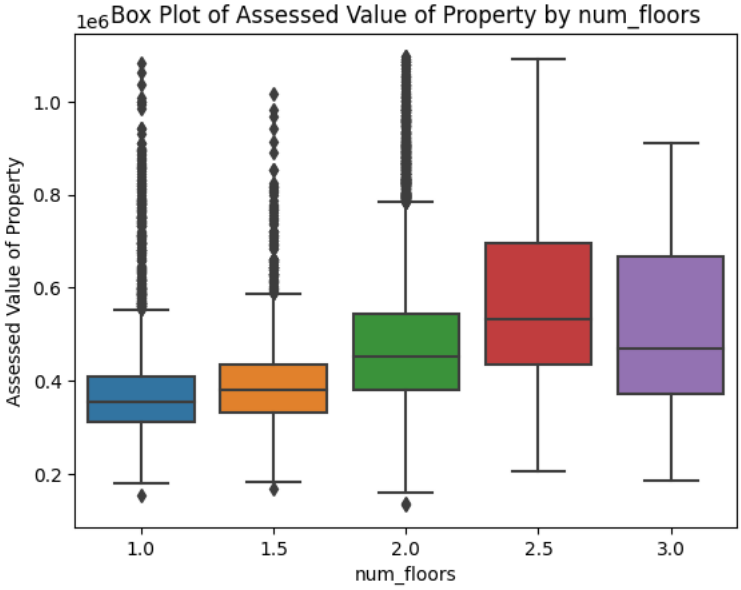
The graph above shows the assessed value of properties by different city-states. Properties in Jamaica Plain had the highest median assessed value of $658,000. This is followed by Cambridge at $460,557.75, Roslindale at $433,500, Dorchester Center at $385,841.16, and Hyde Park at $338,700. As expected, the city-sate is associated with the assessed value of houses.

8. Total number of rooms (r_total_rms)



The graph above shows the assessed value of properties by the number of the rooms. As expected, the properties with a bigger number of rooms have a greater assessed value. Most of the houses in the data have 6 to 8 rooms. It should be noted that the variability (spread) of assessed value also increases as the number of rooms increase. There are a couple of outliers with 17 and 18 rooms that have a very high assessed value of over a million dollars.

9. Number of levels in the structure (num_floors)

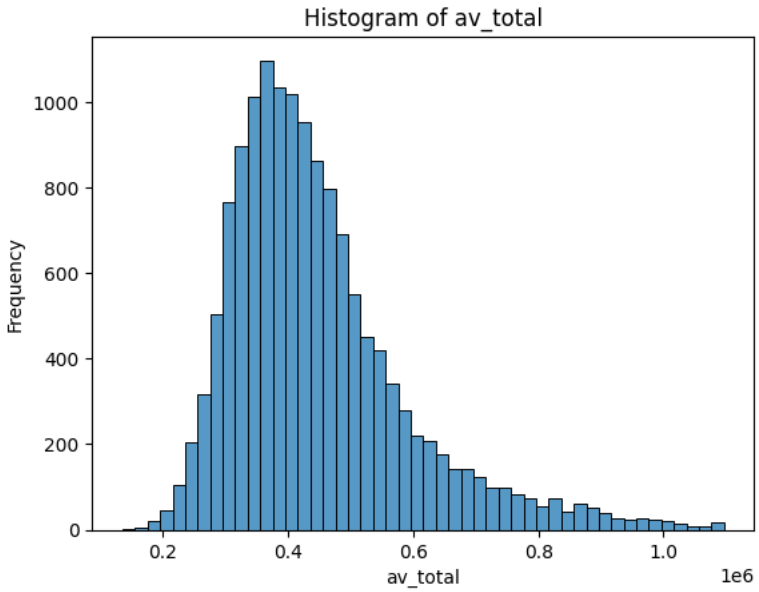


The graph above shows the assessed value of properties by the number of levels in the structure. The highest median assessed value was of properties with 2.5 floors at $205,802.92. This was

followed by properties with 3 floors at $185,300. It should be noted that there were a lot of outliers towards the highest end of the assessed value for properties with 2, 1, and 1.5 floors.

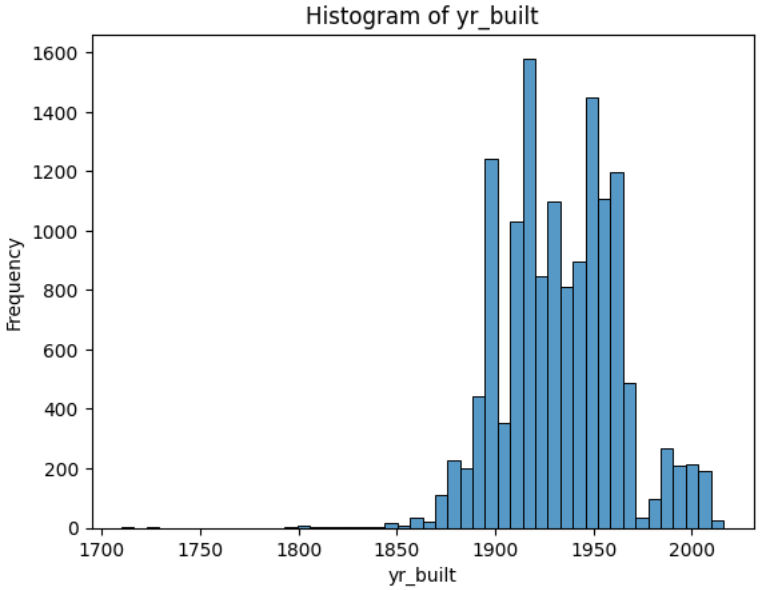
III. Numeric Analysis

1. Assessed Value (av_total)



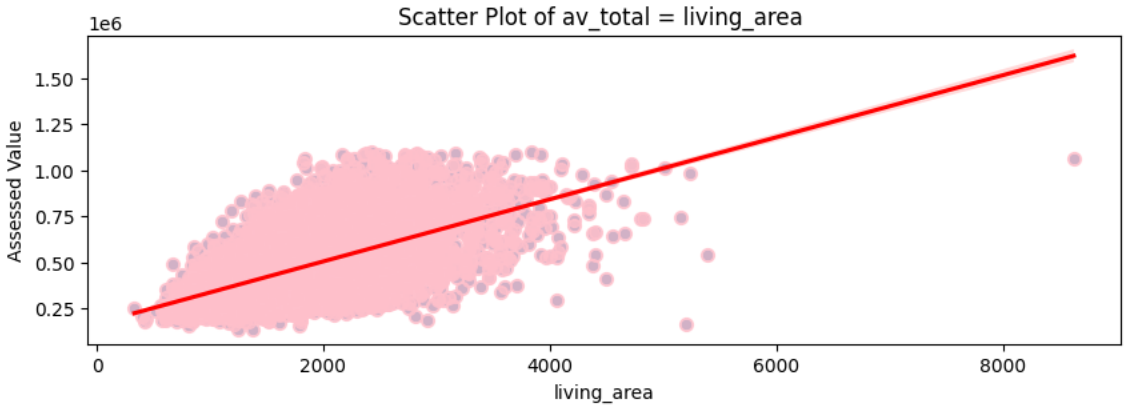
The histogram above shows that the assessed value of properties has a right distribution meaning most of the properties in the dataset are on the lower end of the assessed value. Most of the houses have an assessed value of $350,00 to $500,000.

2. Year built (yr_built)



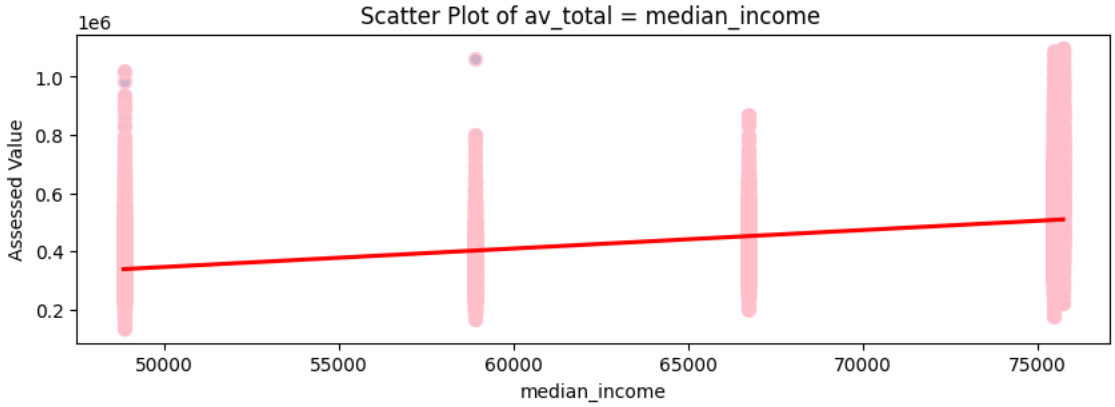
The histogram above shows the majority of properties were built between 1890 and 1960. There are outliers on both ends, a couple of properties were built in the 1700s and early 1800 and some were recently built in the 2010s.

3. Living area square footage (living_area)



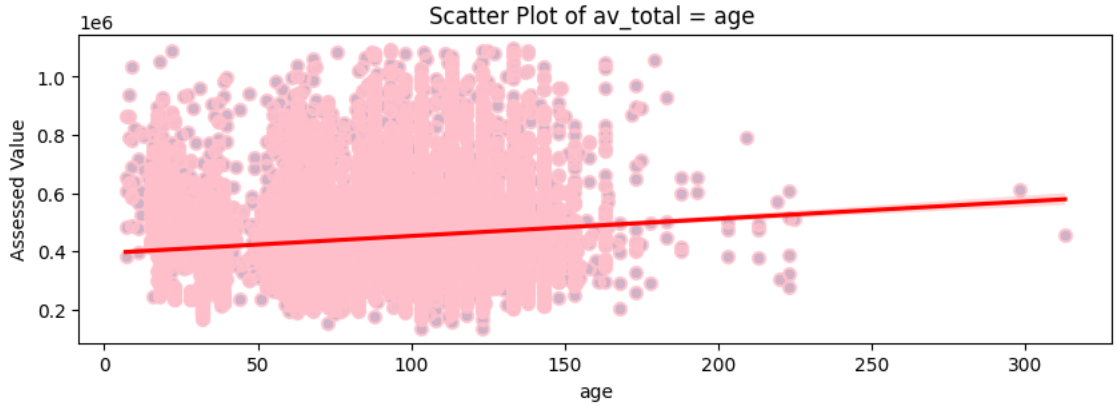
The scatter plot above shows the assessed value of properties by their living area square footage. We observe a positive linear relationship with a correlation of 0.64. This means that the living area is an important predictor of assessed value of a home.

4. Median income of residents of a particular zipcode (median_income)



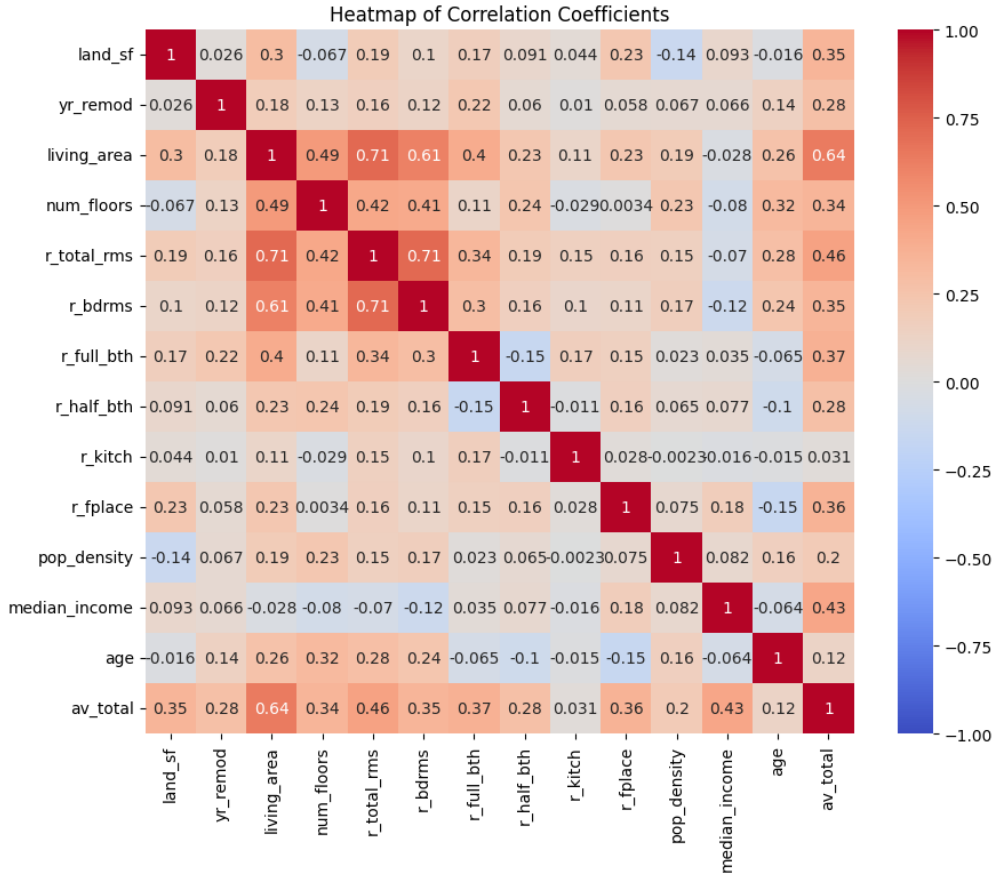
The scatterplot shows the assessed value of properties by the median income of residents of a particular zip code. We see a slight positive correlation of 0.42 between the two variables. Higher median income is associated with higher assessed values of homes, but the linear relationship is not very strong.

5. Age of properties (age)



The scatterplot shows the assessed value of properties by their age. Interestingly, the correlation is only 0.12 which is a weak and almost no linear relationship. This could be due to a variety of reasons; some old properties are more expensive because of historical significance or "antique styles". Newer homes tend to be valued more for their amenities and lack of the need for repairs. From this data, we can conclude that age is not a significant predictor of homes in Boston. There are a couple of outliers that are very old homes built over 300 years ago.

IV. Correlations



The correlation matrix shows the correlations between our main numeric variables. We are particularly focused on understanding the correlation of different explanatory variables with av_total (assessed value). Living_area has the strongest positive linear relationship of 0.64 with av_total. This implies that increasing living area square footage of the property is associated with increasing assessed value. This is followed by r_total_rms with a correlation of 0.45 with av_total. It should be noted that r_total_rms and living_area is highly correlated with a correlation of 0.71 which logically implies that large houses that have a lot of rooms also have a large living area. There are other correlated variables like total number of rooms and total total number of bedrooms (0.71) that should be considered while building the regression model since they would cause multicollinearity. None of the other correlations were significant but something else that stood out was that median income of the residents had a slight positive correlation of 0.42 with assessed value of properties. None of these variables had a negative correlation with the target variable av_total.

V. Model Performance

1. Prediction model 1

- R-square

On analyzing the model, I found that the R-square was 0.439 meaning that approximately 43.9% of the variability in the assessed value of properties can be explained by our model.

- Root Mean Squared Error

On average, the predictions were approximately $107,810.50 away from the assessed value.

- Mean Absolute Error

On average, the predictions made by the model are off by +/- $77,494.68 from the actual assessed value.

2. Prediction model 2

- R-square

On analyzing the model, I found that the R-square was 0.948 meaning that approximately 94.8% of the variability in the assessed value of properties can be explained by our model.

- Root Mean Squared Error

On average, the predictions were approximately $32,894.53 away from the assessed value.

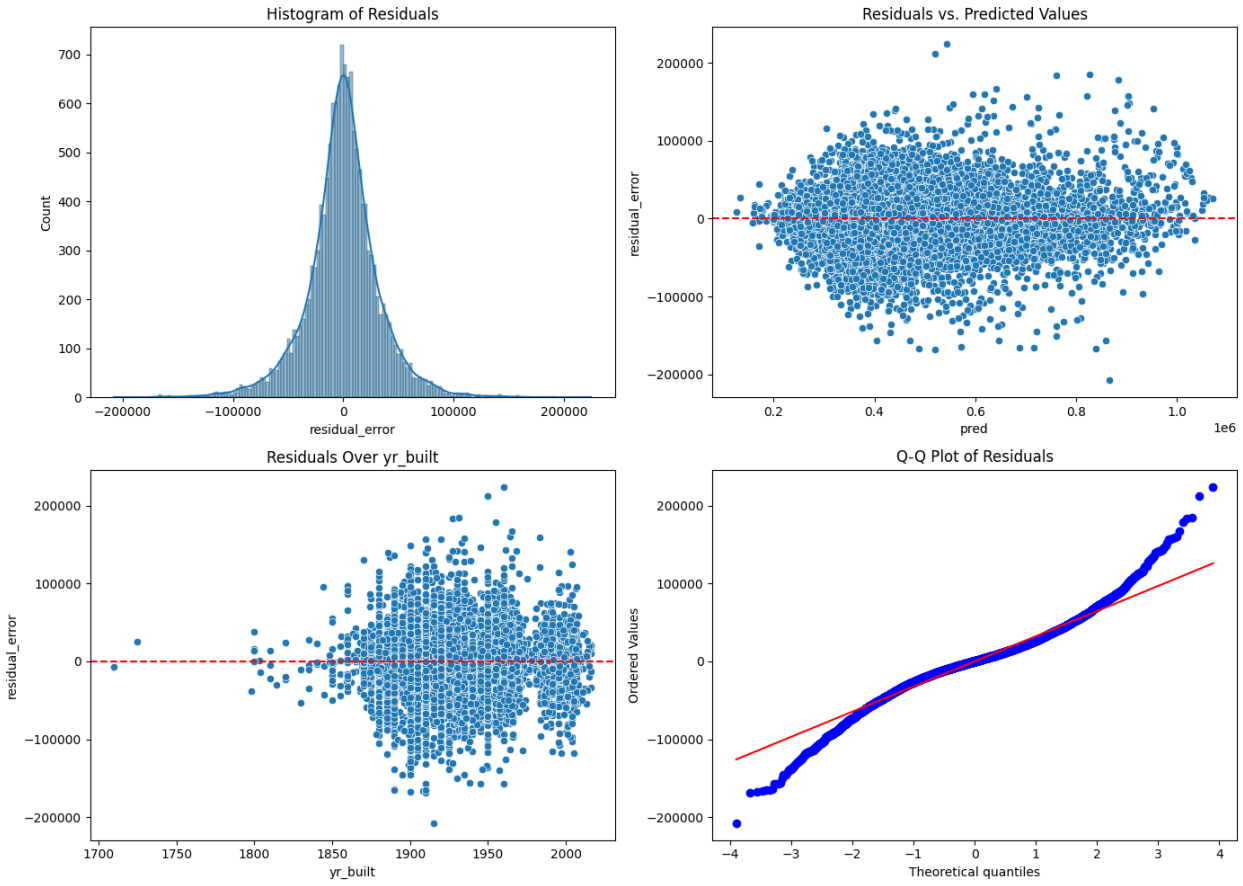
- Mean Absolute Error

On average, the predictions made by the model are off by +/- $23,476.36 from the actual assessed value.

3. Why model 2 is better:

Model 2 is clearly a better model because it explains 94.8% of the variability in the assessed value of properties (50.9% more than model 1). Its predictions were also significantly more accurate than model 1 being only $32,894.53 away from the actual assessed value in comparison to being $107,810.50 away. Finally, the mean absolute error of +/- $23,476.36 is $54,018.32 lower than model 1's mean absolute error of +/- $77,494.68.

4. Prediction model 2 residuals:



- **Histogram of residuals:** The residuals have a normal distribution which is ideal.
- **Residuals vs Predicted Values:** Residuals seem to have almost constant variance (homoscedasticity), although there are some outliers. Heteroskedasticity isn't a problem.
- **Residuals Over yr_built:** Picking one of the x variables, there doesn't seem to be too much of a pattern in residuals.
- **Q-Q Plot**: The Q-Q plot shows how the residuals compare to a normal distribution. Points should ideally lie on the red line if they are perfectly normally distributed. From the plot, the residuals are somewhat close to normal, though there are some deviations at the tails. This suggests the presence of outliers and a potential non-linear relationship.
- Overall, the residual analysis shows that this is a good model.

5. Top and Bottom 10 Record Predictions

1. 10 Best Predictions of the Model

g_area	num_floors	structure_class	r_bldg_styl	...	median_income	city_state	av_total	age	is_remod	remod_recent	yr_category	pred	residual_error	abs_residual
1182	2.0000	R	CL	...	58890	Hyde Park, MA	332042.1316	112	1	Other	Other	332043.6562	-1.5246	1.5250
1068	1.0000	R	BW	...	66735	Roslindale, MA	359000.0000	103	1	Other	Other	359010.5312	-10.5312	10.5310
972	1.0000	R	RN	...	75446	Cambridge, MA	344800.0000	63	1	Other	Other	344813.3750	-13.3750	13.3750
1478	2.0000	R	CL	...	48841	Dorchester Center, MA	285886.7055	133	1	Other	Other	285902.5938	-15.8883	15.8880
2161	2.0000	R	VT	...	48841	Dorchester Center, MA	658200.0000	143	1	Other	Other	658181.5000	18.5000	18.5000
1450	2.0000	R	CL	...	75446	Cambridge, MA	454388.7767	93	1	Other	Other	454407.9062	-19.1296	19.1300
1799	2.0000	R	CL	...	66735	Roslindale, MA	451563.1848	143	1	Other	Other	451585.1250	-21.9402	21.9400
1344	1.5000	R	CL	...	58890	Hyde Park, MA	314484.7140	139	1	Other	Other	314506.9688	-22.2548	22.2550
2118	2.0000	R	CL	...	75446	Cambridge, MA	553200.0000	15	0	Other	Other	553175.8750	24.1250	24.1250
1527	2.0000	R	CL	...	66735	Roslindale, MA	382567.6008	105	0	Other	Other	382593.0938	-25.4929	25.4930

We observe from the table above that the model's 10 best predictions were off by just $1.52 to $25.49. The model is comprehensive in the sense that the top predictions did not have a pattern in variables like zip code, year built, median income, age, living area etc.

2. Top 10 Overestimates of the Model

structure_class	r_bldg_styl	...	pop_density	median_income	city_state	av_total	age	is_remod	remod_recent	yr_category	pred	residual_error
R	CL	...	10618	75730	Jamaica Plain, MA	657900.0000	108	1	Other	Other	865683.0625	-207783.0625
R	CL	...	11505	66735	Roslindale, MA	351800.0000	113	1	Other	Other	520393.0625	-168593.0625
R	CL	...	10618	75730	Jamaica Plain, MA	671200.0000	123	1	Other	Other	838272.0000	-167072.0000
R	CL	...	11505	66735	Roslindale, MA	322100.0000	115	1	Other	Other	488623.1562	-166523.1562
R	CL	...	10618	75730	Jamaica Plain, MA	549800.0000	133	1	Other	Other	715284.3750	-165484.3750
R	CL	...	10618	75730	Jamaica Plain, MA	521800.0000	113	1	Other	Other	687193.4375	-165393.4375
R	CL	...	11505	66735	Roslindale, MA	407800.0000	133	1	Other	Other	572399.4375	-164599.4375
R	CL	...	10618	75730	Jamaica Plain, MA	313100.0000	78	1	Other	Other	470196.7812	-157096.7812
R	CL	...	15913	48841	Dorchester Center, MA	247800.0000	63	1	Other	Other	404759.9688	-156959.9688
R	CL	...	10618	75730	Jamaica Plain, MA	701400.0000	113	1	Other	Other	858199.5000	-156799.5000

We observe from the table above that the top 10 overestimates of the model ranged from -$207,783.06 to -$156,799.50. This is important because overestimating assessed value means taxing residents on a value more than the value calculated based on their actual home value. This helps us understand the worst outcomes of the prediction for residents.

3. Top 10 Underestimates of the Model

num_floors	structure_class	r_bldg_styl	...	pop_density	median_income	city_state	av_total	age	is_remod	remod_recent	yr_category	pred	residual_error
2.0000	R	CL	...	10618	75730	Jamaica Plain, MA	767500.0000	63	1	Other	Other	542955.3125	224544.6875
2.0000	R	CL	...	10618	75730	Jamaica Plain, MA	732300.0000	73	1	Other	Other	519895.8750	212404.1250
2.0000	R	CL	...	10618	75730	Jamaica Plain, MA	1011700.0000	92	1	Other	Other	826599.8750	185100.1250
2.0000	R	CL	...	10618	75730	Jamaica Plain, MA	944600.0000	96	1	Other	Other	760980.6875	183619.3125
1.0000	R	CP	...	10618	75730	Jamaica Plain, MA	1062620.0000	68	1	Other	Other	883951.8125	178668.1875
1.5000	R	CP	...	10618	75730	Jamaica Plain, MA	809300.0000	58	1	Other	Other	642110.1875	167189.8125
2.0000	R	CL	...	10618	75730	Jamaica Plain, MA	755500.0000	59	1	Other	Other	595222.0000	160278.0000
2.0000	R	SD	...	10618	75730	Jamaica Plain, MA	777500.0000	40	1	Other	Other	617846.3750	159653.6250
2.0000	R	CL	...	10618	75730	Jamaica Plain, MA	1060314.0000	88	1	Other	Other	902510.0625	157803.9375
2.0000	R	CL	...	10618	75730	Jamaica Plain, MA	978400.0000	113	1	Other	Other	821302.3125	157097.6875

We observe from the table above that the top 10 underestimates of the model ranged from $224,544.68 to $157,097.68. This is important because underestimating assessed value means taxing residents on a value less than the value calculated based on their actual home value. This helps us understand the worst outcomes of the city of Boston as it loses out on tax money by underestimating the assessed value.