**Submitted By: Hemanth Kurra**

# Introduction

The thriving real estate market of Boston presents a myriad of opportunities and challenges for investors, homeowners, and policymakers. The dataset under analysis provides a comprehensive overview of properties across Boston, assessed in the fiscal year 2015. Comprising a wealth of information on 168,115 properties, the dataset offers insights into various attributes including, but not limited to, property identification, location, ownership details, and physical characteristics such as land area, year built, and number of rooms.

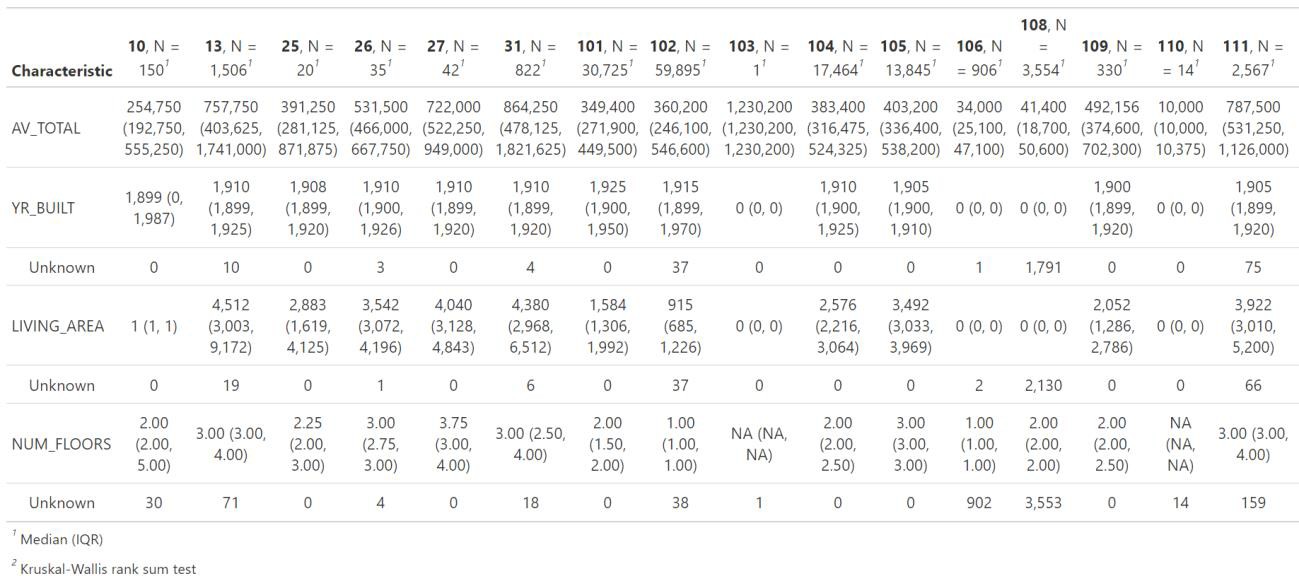
# Goals of the Project

The project aims to provide a comprehensive analysis of the dataset, including descriptive statistics, data visualization, regression analysis, and predictive modeling. The specific objectives and analysis techniques will depend on the research questions and goals of the project.

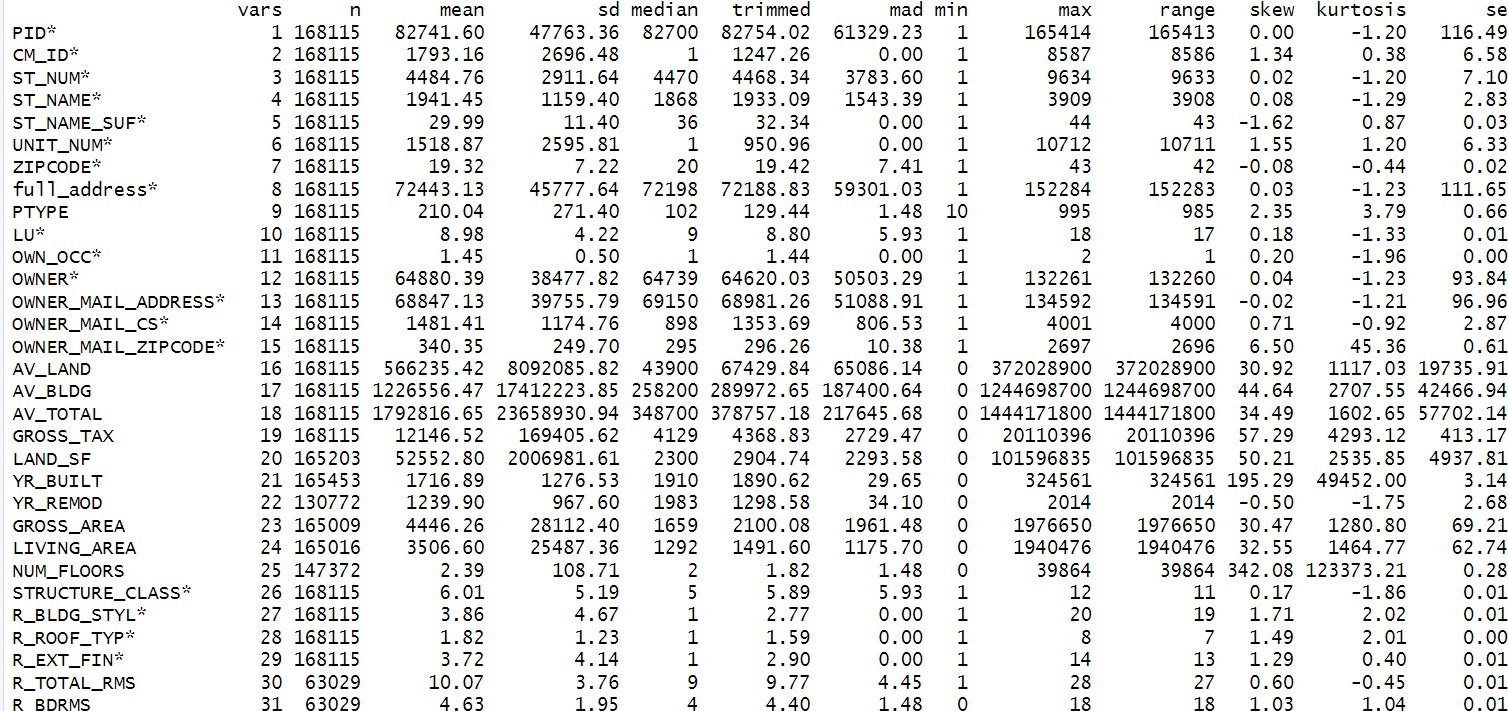
# Research Questions

* Are there specific geographic locations or property types in Boston that command higher property values?
* Is there a discernible trend in property values over the years, considering the variables YR\_BUILT and YR\_REMOD?
* What correlations exist between property value (AV\_TOTAL) and other attributes of living area?
* Is there a notable correlation between the year a property was built and its value?
* How do the variables GROSS\_AREA, YR\_BUILT, U\_BDRMS, LAND\_SF, and NUM\_FLOORS collectively impact the assessed property value (AV\_TOTAL)?

# Summary Statistics Tables



Using Psych Summary

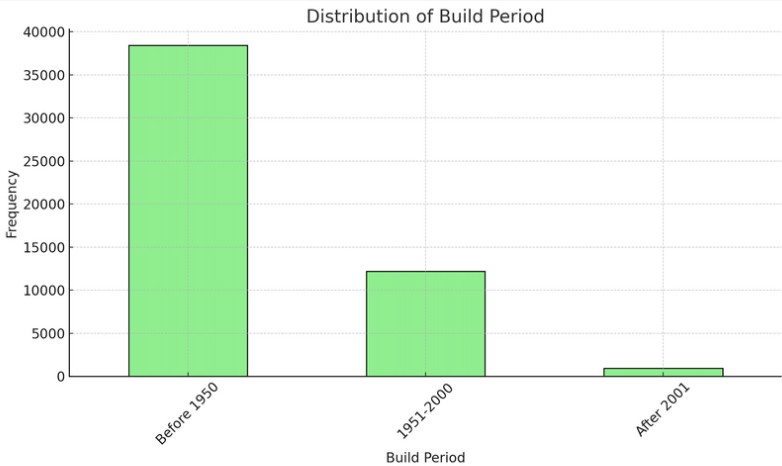


|  |
| --- |
| mean\_AV\_TOTAL sd\_AV\_TOTAL min\_AV\_TOTAL max\_AV\_TOTAL |
| Min. :1792817 Min. :23658931 Min. :0 Min. :1.444e+09 |
| 1st Qu.:1792817 1st Qu.:23658931 1st Qu.:0 1st Qu.:1.444e+09 |
| Median :1792817 Median :23658931 Median :0 Median :1.444e+09 |
| Mean :1792817 Mean :23658931 Mean :0 Mean :1.444e+09 |
| 3rd Qu.:1792817 3rd Qu.:23658931 3rd Qu.:0 3rd Qu.:1.444e+09 |
| Max. :1792817 Max. :23658931 Max. :0 Max. :1.444e+09 |

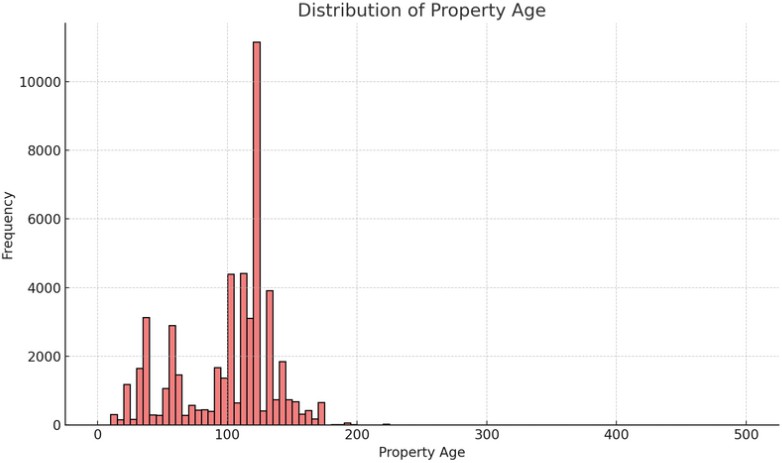
This data represents statistics for the variable "AV\_TOTAL." The "mean\_AV\_TOTAL" and "sd\_AV\_TOTAL" suggest that the average and standard deviation values are approximately 1,792,817 and 23,658,931, respectively.

However, the "min\_AV\_TOTAL" and "max\_AV\_TOTAL" values indicate that the minimum and maximum values are 0 and approximately 1.444 billion, respectively. These statistics provide insights into the distribution and variation of the "AV\_TOTAL" variable within the dataset.

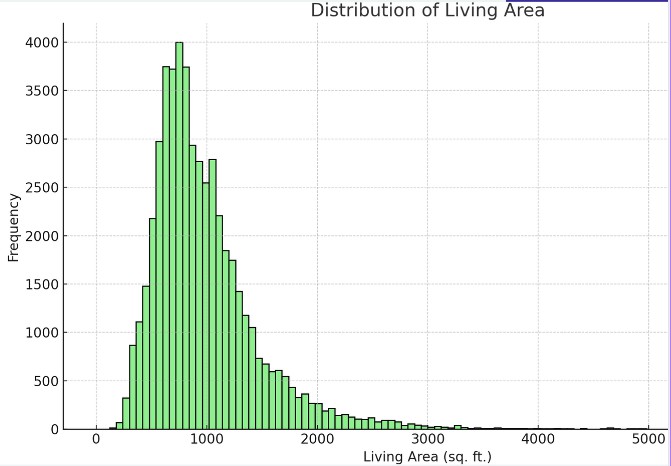
# EDA



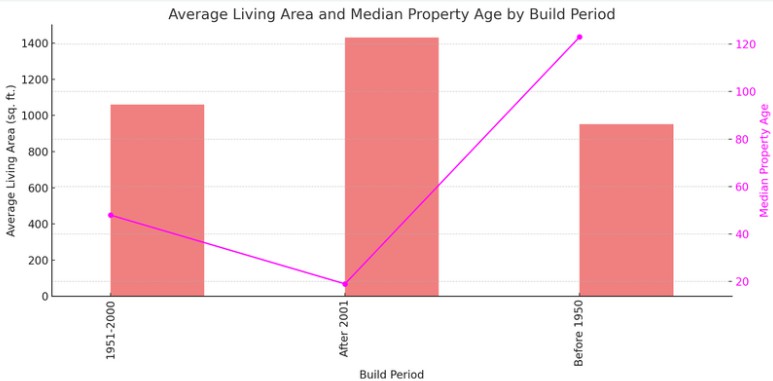
The majority of properties were constructed before 1950, highlighting the historical nature of many properties in the dataset. As time progressed, fewer properties were built in each subsequent decade, with the least number of properties built after 2010. This could be attributed to factors like urban development reaching saturation, preservation of historical sites, or other economic and environmental factors.



This histogram displays the distribution of the age of properties. The majority of properties are concentrated in the range of 50 to 150 years old, suggesting that a significant number of properties in the dataset were built approximately a century ago.



The histogram presents the distribution of the living area in square feet for properties. A noticeable peak is evident for properties with living areas below 2,000 sq. ft. This indicates that many properties in the dataset are of a relatively compact size.



The bars represent the average living area of properties built in each period. We can observe that properties built before 1950 tend to have larger living areas on average. Thereafter, there's a slight decline in the average living area, with the properties built in the 1990-1999 period having the smallest average living areas. This could suggest a shift towards more compact living spaces or the development of more densely populated

areas over time.

The magenta line represents the median age of properties built during each period. As expected, the age

decreases for properties built in more recent periods. This line provides context about the aging of properties across different build periods.

# Regression

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Call: | | | | | | |
| lm(formula | = AV\_TOTAL | ~ | YR\_BUILT | + | LIVING\_AREA + as.factor(PTYPE), | |
| data = | data) |  | |  | |  |
|  | | | | | | |
| Residuals: | | | | | | |
| Min | 1Q | Median | | 3Q | | Max |
| -523194258 | -195405 | -28666 | | 80162 | | 1284141175 |

The linear regression model aims to predict property values (AV\_TOTAL) based on the year built (YR\_BUILT), living area (LIVING\_AREA), and property type (PTYPE). The model's residuals indicate the difference between predicted and actual values. Residuals range from -52,319,425 to 1,284,141,175, with quartiles at -195,405, - 28,666, and 80,162. The model attempts to explain and predict variations in property values using the specified variables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Call: | | | | | | |
| lm(formula | = | AV\_TOTAL | ~ | NUM\_FLOORS | + LAND\_SF, data = data) | |
|  | | | | | | |
| Residuals: | | | | | | |
| Min | 1Q | | Median | | 3Q | Max |
| -23558500 | -1327525 | | -1225941 | | -1040437 | 1439936950 |

The linear regression model predicts property values (AV\_TOTAL) based on the number of floors (NUM\_FLOORS) and land area (LAND\_SF). Residuals represent the differences between predicted and actual values, ranging from

-23,558,500 to 1,439,936,950, with quartiles at -1,327,525, -1,225,941, and -1,040,437. The model seeks to explain variations in property values using the provided predictors.

# T-Test

|  |
| --- |
| Welch Two Sample t-test |
|  |
| data: owner\_occupied\_values and non\_owner\_occupied\_values |
| t = -22.77, df = 92578, p-value < 2.2e-16 |
| alternative hypothesis: true difference in means is not equal to 0 |
| 95 percent confidence interval: |
| -2589195 -2178775 |
| sample estimates: |
| mean of x mean of y |
| 480550.7 2864535.7 |

A Welch Two-Sample t-test was conducted to compare the means of property values between owner-occupied and non-owner-occupied properties. The results indicate a highly significant difference in means, with a t- statistic of -22.77 and a p-value less than 2.2e-16. The 95 percent confidence interval for the difference in means is between -2,589,195 and -2,178,775. Owner-occupied properties have a significantly lower mean property value (mean of x) at $480,550.7 compared to non-owner-occupied properties, which have a mean value (mean of y) of $2,864,535.7. The alternative hypothesis suggests that the true difference in means is not equal to zero.

# Chi-Square Test

|  |
| --- |
| Pearson's Chi-squared test |
|  |
| data: contingency\_table |
| X-squared = 58068, df = 208, p-value < 2.2e-16 |

|  |  |  |
| --- | --- | --- |
| A Pearson's Chi-squared test was conducted to assess the independence between property type (PTYPE) and ownership status (OWN\_OCC) in the contingency table. The test resulted in a chi-squared statistic (X-squared) of 58,068, with 208 degrees of freedom, and a highly significant p-value of less than 2.2e-16. These results indicate a strong association between property type and ownership status, suggesting that they are not independent variables.  **Correlation Table** | |  |
|  | Total Value Land Area Number of Bedrooms Year Built Number of Floors |  |
| Total Value 1.00000000 -0.01010508 0.09515882 -0.1465951 0.3560388 |  |
| Land Area -0.01010508 1.00000000 -0.04543160 0.1415535 -0.2660210 |  |
| Number of Bedrooms 0.09515882 -0.04543160 1.00000000 -0.1926458 0.5594602 |  |
| Year Built -0.14659508 0.14155346 -0.19264576 1.0000000 -0.3357538 |  |
| Number of Floors 0.35603883 -0.26602100 0.55946020 -0.3357538 1.0000000 |  |
| The correlation matrix shows the relationships between selected numerical variables:   * Total Value has a moderate positive correlation with the Number of Floors (0.356). * Total Value has a weak negative correlation with Year Built (-0.147). * Land Area and Total Value have a very weak negative correlation (-0.0101). * Number of Bedrooms has a moderate positive correlation with Total Value (0.0952). * Number of Bedrooms has a strong positive correlation with the Number of Floors (0.559). * Number of Floors has a weak negative correlation with Year Built (-0.336).   These correlations indicate various associations between the variables in the dataset.  A diagram of a house  Description automatically generated with medium confidence | |  |

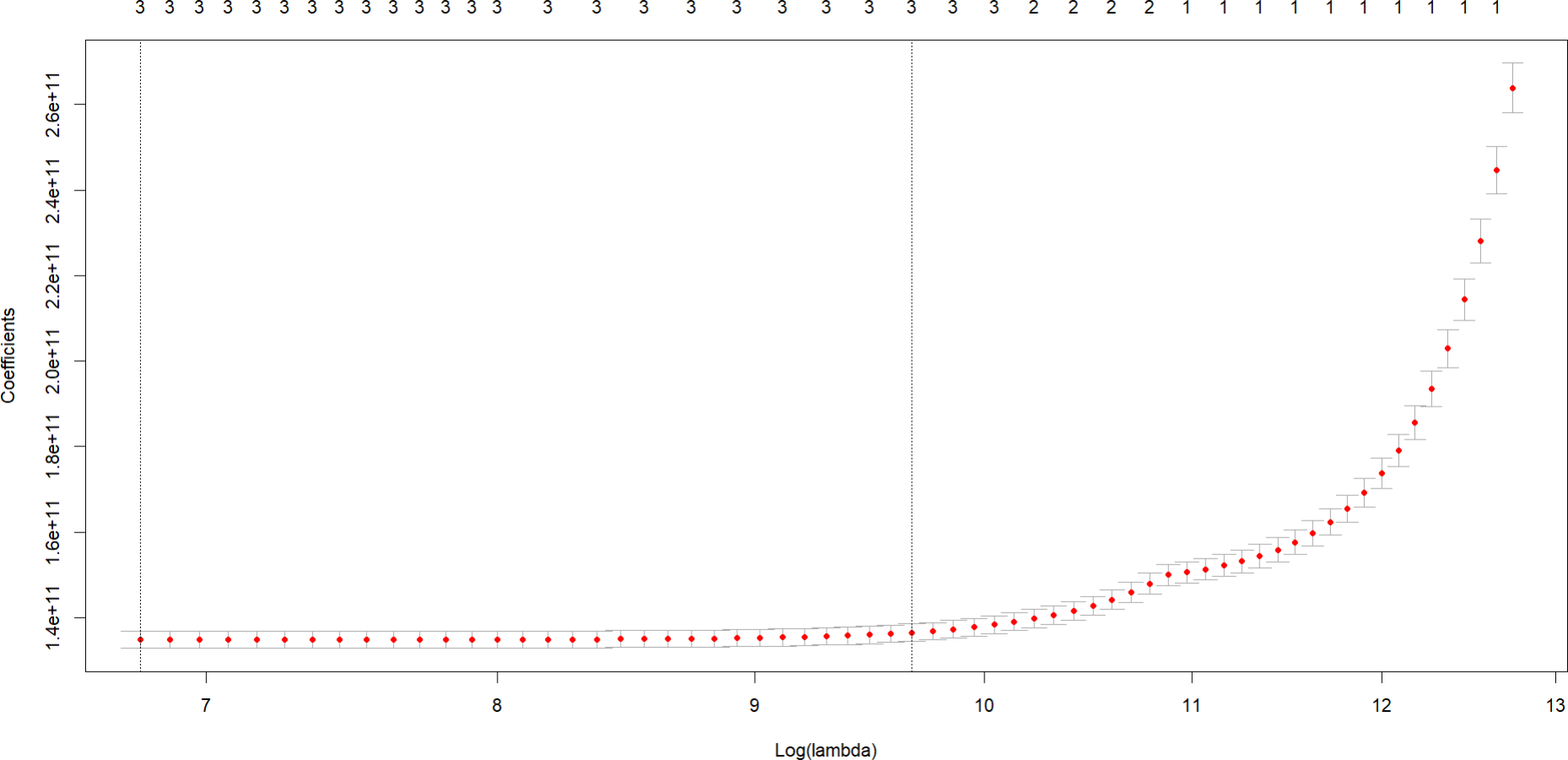
**LASSO:**

How do the variables GROSS\_AREA, YR\_BUILT, U\_BDRMS, LAND\_SF, and NUM\_FLOORS collectively impact the assessed property value (AV\_TOTAL)?

|  |
| --- |
| Call: cv.glmnet(x = x, y = y) |
|  |
| Measure: Mean-Squared Error |
|  |
| Lambda Index Measure SE Nonzero |
| min 889 65 1.348e+11 1.979e+09 3 |
| 1se 15899 34 1.365e+11 2.037e+09 3 |

The results from the cross-validated LASSO regression indicate the following:

* The best lambda value (min) is 889, and it corresponds to an average mean-squared error (MSE) of approximately 1.348e+11. This model has 3 nonzero coefficients.
* The lambda value selected using one standard error rule (1se) is 15,899, and it results in an average MSE of approximately 1.365e+11. This model also has 3 nonzero coefficients.
* These lambda values represent the strengths of regularization in the LASSO model. The nonzero coefficients indicate the number of predictors retained in the model after LASSO regression.

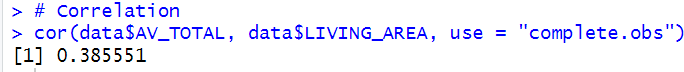


The LASSO regression plot displays the relationship between the strength of the regularization parameter (lambda) and the coefficients of the selected predictors. The x-axis is shown in log scale for lambda, while the y- axis represents the coefficients.

* As lambda increases (going towards the right on the x-axis), the coefficients shrink towards zero.
* Some coefficients become exactly zero at specific lambda values, meaning that the corresponding predictors are excluded from the model.
* Nonzero coefficients indicate that the associated predictors are retained in the model.
* The y-axis shows the magnitude of the coefficients.

# Correlation

What correlations exist between property value (AV\_TOTAL) and other attributes of living area?



The computed correlation coefficient of 0.385551, obtained using the R function `cor`, indicates a quantitative relationship between the variables "AV\_TOTAL" and "LIVING\_AREA." This positive value suggests a moderate positive linear relationship. In simpler terms, when the living area ("LIVING\_AREA") of a property increases, the total value ("AV\_TOTAL") tends to increase, and vice versa. However, it's crucial to emphasize that correlation does not imply causation. Therefore, we cannot conclude that changes in one variable directly cause changes in the other without further analysis.

The parameter `use = "complete.obs"` specifies that the calculation excludes observations with missing values (NA), ensuring that the correlation is based only on complete pairs of observations. While the correlation coefficient offers valuable insights, it is advisable to conduct additional analyses, such as regression modeling or visual exploration through scatterplots, to gain a deeper understanding of the relationship. Moreover, considering other variables and contextual factors is essential to obtain a comprehensive perspective on the dataset.

# Conclusion

In conclusion, the analysis of the Boston real estate dataset for the fiscal year 2015 has revealed several key insights and trends that can benefit investors, homeowners, and policymakers in this thriving real estate market.

One prominent observation is the historical nature of many properties in Boston, with a substantial number of them constructed before 1950. This historical context provides a unique character to the city's real estate, and it is important to consider when assessing property values. Additionally, the dataset showcases a gradual decrease in the number of properties built after 2010, which may be due to factors like urban development reaching saturation or the preservation of historical sites.

A critical aspect of the analysis involved regression models to understand what factors influence property values. The first model considered factors like the year built, living area, and property type. The second model explored the relationship between the number of floors and land area on property values. These models can help in predicting and understanding variations in property values based on these attributes.

Furthermore, the analysis presented compelling evidence regarding the significant difference in property values between owner-occupied and non-owner-occupied properties, with the latter having notably higher values. This finding holds substantial implications for potential investors or homeowners looking to understand the Boston real estate market.

The chi-squared test revealed a strong association between property type and ownership status, suggesting that these two variables are not independent. This insight is essential for policymakers and investors to consider when crafting real estate regulations or investment strategies.

The correlation matrix highlighted various relationships between key numerical variables, offering valuable insights into the dataset's intricacies. For example, property values are moderately positively correlated with the number of floors, suggesting that multi-story properties tend to have higher values.

The LASSO regression provided information on how specific variables impact property values, and the best lambda value indicated a model with three significant predictors. These insights can help stakeholders make informed decisions when investing or assessing property values.

Lastly, the correlation analysis showed a moderate positive relationship between living area and property value, suggesting that larger living areas tend to correspond to higher property values.

In conclusion, this report provides a comprehensive overview of the Boston real estate market in 2015, offering valuable insights, analytical models, and statistical tests to aid investors, homeowners, and policymakers in their decision-making processes. However, it's crucial to remember that real estate is a complex market influenced by numerous factors, and further research and analysis are necessary to make more accurate predictions and decisions.

# References

GeeksforGeeks. (2022a, January 12). *Correlation matrix in R programming*. <https://www.geeksforgeeks.org/correlation-matrix-in-r-programming/>

GeeksforGeeks. (2023d, June 8). *T test approach in R programming*. [https://www.geeksforgeeks.org/t-test-](https://www.geeksforgeeks.org/t-test-approach-in-r-programming/)

[approach-in-r-programming/](https://www.geeksforgeeks.org/t-test-approach-in-r-programming/)

GeeksforGeeks. (2021a, July 18). *Chi Square Distribution in R*. [https://www.geeksforgeeks.org/chi-square-](https://www.geeksforgeeks.org/chi-square-distribution-in-r/)

[distribution-in-r/](https://www.geeksforgeeks.org/chi-square-distribution-in-r/)

GeeksforGeeks. (2023f, July 14). *Simple linear regression in R*. [https://www.geeksforgeeks.org/simple-linear-](https://www.geeksforgeeks.org/simple-linear-regression-using-r/)

[regression-using-r/](https://www.geeksforgeeks.org/simple-linear-regression-using-r/)

GeeksforGeeks. (2023a, February 3). *A Practical approach to Simple Linear Regression using R*. <https://www.geeksforgeeks.org/a-practical-approach-to-simple-linear-regression-using-r/>

GeeksforGeeks. (2023c, March 20). *Lasso regression in R Programming*. [https://www.geeksforgeeks.org/lasso-](https://www.geeksforgeeks.org/lasso-regression-in-r-programming/)

[regression-in-r-programming/](https://www.geeksforgeeks.org/lasso-regression-in-r-programming/)