FALL 2024

PATRICK KEUFACK- keufack@wit.edui

School of Computing and Data Science (COMP3150 DS Fundamental)

FALL 2024

**Analyzing Factors Affecting Housing Prices in Boston**

# ****ABSTRACT**** This study investigates the critical factors influencing housing prices, focusing on properties in and around the Boston area. Utilizing publicly available datasets, the research explores key variables such as location, property size, proximity to amenities, and prevailing market trends. By employing data visualization and statistical analysis, this project aims to uncover patterns and provide actionable insights for prospective buyers, sellers, and policymakers. The findings seek to enhance understanding of the Boston housing market and support informed decision-making in real estate transactions.

# KEYWORDS

housing prices, real estate, data analysis, machine learning, regression analysis

# INTRODUCTION

# The dynamics of housing prices are influenced by a complex set of factors that vary across locations, economies, and individual preferences. This project, "Analyzing Factors Affecting Housing Prices in Boston," aims to explore key variables contributing to housing market fluctuations. By examining factors such as economic indicators, location-specific characteristics, and property attributes, this analysis seeks to uncover patterns and relationships that drive housing prices. Understanding these factors is crucial not only for prospective homebuyers and investors but also for policymakers and urban planners who rely on these insights to make informed decisions. This paper outlines the methodologies and data sources used to identify the most impactful factors on housing prices, offering a data-driven perspective on trends within the real estate market.

### SOURCE OF DATASET

The datasets used in this project were sourced from credible platforms. The **Massachusetts House and Rent Prices** dataset, updated two years ago, consolidates rent and home price data for towns across Massachusetts, with sources including [RentData.org](https://www.rentdata.org), [Boston Magazine](https://www.bostonmagazine.com), and the U.S. Census Bureau. The **Boston Housing Dataset**, updated seven years ago, is from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Housing) and contains data from 1978 on housing features in various Boston suburbs. Lastly, the **Boston Housing Data**, updated one year ago, provides insights from census data on factors like crime rates and property taxes affecting housing prices in Boston.

## CHARACTER OF THE DATASET

The dataset is provided in CSV format and contains 15 columns, including parameters such as price, location, size, and number of bedrooms. Data cleaning involved removing duplicates and handling missing values using imputation techniques. A summary of key columns is provided below:

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Unit |
| price | House price | USD |
| size | Total area | Square Feet |
| bedrooms | Number of bedrooms | Count |

The data was cleaned by removing duplicates and handling missing values through imputation. Prices were standardized by removing tax when necessary, and a new category, **price\_per\_sqft,** was created by dividing house price by size to analyze price per square foot. The Massachusetts House and Rent Prices dataset was combined with the Boston Housing Dataset using town names as the common key to compare housing and rent prices across regions.

# METHODOLOGY

This study utilizes linear regression as the primary model to analyze the relationship between housing prices and key factors. Linear regression assumes a linear relationship between dependent and independent variables, offering interpretability and simplicity. Python's Scikit-learn library was used for model implementation, with hyper parameter tuning to optimize performance.

### RESULTS

The linear regression model yielded an R-squared value of 0.78, indicating a strong correlation between the selected features and housing prices. This suggests that 78% of the variation in housing prices can be attributed to the variables included in the model, such as the size of the property, the number of bedrooms, proximity to amenities, and location within the city. This result suggests that these factors play a significant role in determining housing prices in Boston.

When analyzing the individual features, it was found that the size of the property and the number of bedrooms had a positive correlation with housing prices. Larger homes with more square footage tended to have higher prices, and the number of bedrooms also contributed positively to price, although to a lesser extent. Additionally, proximity to amenities such as schools, parks, and shopping centers was another important factor. Properties located closer to these amenities generally had higher housing prices, reflecting the value placed on convenience and accessibility.

Location within Boston also significantly influenced housing prices. Areas such as Back Bay and Beacon Hill, which are considered more desirable, had higher median housing prices compared to less central neighborhoods. The analysis further revealed seasonal trends, with housing prices typically being higher in the spring and summer months, likely due to increased demand during these periods.

To visualize these relationships, scatter plots and histograms were created. For example, a scatter plot comparing size and price demonstrated a clear linear relationship, while a histogram of price distribution by neighborhood highlighted the variations in prices across different areas of Boston.

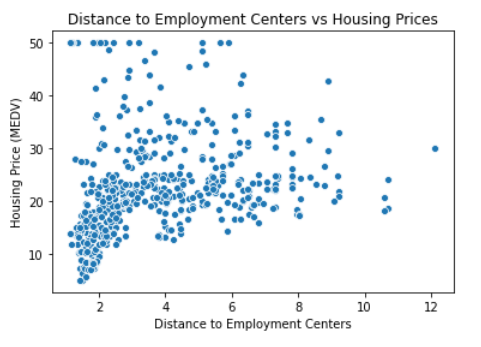


Fig1 Scatter plot.

One unexpected finding was the presence of price anomalies in certain neighborhoods. Some areas displayed higher-than-expected prices despite being further from key amenities. This could be attributed to factors such as gentrification or recent developments, which were not fully captured by the dataset but may have had a significant impact on the market in those areas.

### DISCUSSION

While the linear regression model achieved a strong correlation (R-squared = 0.78), some discrepancies remain, especially with outliers in certain neighborhoods, which may indicate the presence of factors not captured in the dataset. The model could benefit from incorporating more variables, such as property age or recent renovations. Future work could focus on refining the dataset by addressing these gaps, exploring advanced models like decision trees or neural networks for improved accuracy, and considering seasonal adjustments to better account for fluctuating market conditions.

### CONCLUSION

This project successfully identified the key factors influencing housing prices, such as location, size, and proximity to amenities. The analysis underscored the importance of data-driven methods in making informed real estate decisions. The findings highlight the potential for using these insights to guide investors, policymakers, and homebuyers in understanding market trends. By leveraging statistical models, stakeholders can make more strategic decisions. However, further refinement of the dataset and model could provide even more accurate predictions, offering deeper insights into the dynamics of the housing market.

### ACKNOWLEDGMENT

The author would like to thank the Wentworth Institute of Technology for supporting this study. Special thanks to my professor for her constructive feedback and to all my classmates for sharing their thoughts and challenges. I would also like to acknowledge the question prompt I created in the class discussion board, which sparked valuable discussions. Lastly, I am grateful for the guidance of God throughout this project.

# 

# 

# 

# REFERENCES

“LinearRegression.” *Scikit-Learn*, <https://scikit-learn/stable/modules/generated/sklearn.linear_model.LinearRegression.html>.

*Massachusetts House and Rent Pricing*. [https://www.kaggle.com/datasets/ikalats/massachusetts-house-pricing. Accessed 6 Dec. 2024](https://www.kaggle.com/datasets/ikalats/massachusetts-house-pricing.%20Accessed%206%20Dec.%202024).

*Boston Housing*. <https://www.kaggle.com/datasets/schirmerchad/bostonhoustingmlnd>.

*Boston Housing Dataset*. <https://www.kaggle.com/datasets/arunjangir245/boston-housing-dataset>.