BOSTON REAL ESTATE RESIDENTIAL RENTAL TRENDS

Gregory Stoller

Applicability to My Business

In 1999, I started my current business (Massachusetts Corporation Division, 1999). It was initially founded as a commercial real estate holding company that bought, sold, and operated commercial properties and provided advisory and management services. I followed this model for seven years and then decided to expand. In 2008, I purchased my first residential building (Massachusetts Corporation Division, 2008) and later expanded to other residential and commercial properties. Currently, I have two residential apartment buildings that are located in downtown Boston. One is near a university campus, and the second is located one house away from one of Boston's more popular suburbs. For the past decade and a half, residential property prices and economic conditions have directly correlated to the rents I can charge. Well outside of my work in academia, I regularly read reports on the rental housing market (Harvard University, 2024), articles in the popular press involving real estate (Woods, 2024), and even lending sites such as Rocket Mortgage (2024).

Our doctoral class project is a perfect opportunity to expand my knowledge in the area of residential real estate by studying data analysis approaches using Python. After discussing the end-of-semester project in class, my first search for data involved nearly 30 different sites on Kaggle, GitHub, and other related online locations. I provided examples of ten of these in **Exhibit 1**. Unsurprisingly, some of these contained significantly older data, code that I needed help getting working, or only paid access options. The next step was to begin a new search from the perspective of the data itself in an attempt to work backward (see **Exhibit 2**). I posited that if I could find good data, I could write the code around it to complete the analysis. My efforts here resulted in moderate success. While I located some excellent data, most were at the national or state levels and had no applicability to Boston.

After several days of searching, I found some code that made much sense. However, it took me several weeks of reading to understand a lot of the nomenclature and concepts. I began my attempts to learn the basics with the book for our course (Delen, 2020), but at times, even that fell short in specifics. I found modest success in peer-reviewed articles but ultimately had to combine these with various websites and articles from the popular press.

Eventually, I became familiar enough with the theories that I could begin studying the data. Unfortunately, despite the comprehensive nature of the model, it only contained 506 entries and was based on a data set from the 1970s. Through successive research attempts, however, I added a contemporary aspect to the data through specific updates to Crime Rates, Nitrous Oxide levels, Real Estate Taxes, etc. The current number of observations is 6,073. Even though I expanded the data, many of the original analytical features remain intact:

A) There continues to be a high r-squared value, effectively explaining variations in the data.
B) Features that have the most significant impact on housing prices continue to be 1) Crime (coefficient: 8.384767e-17), 2) proximity to the Charles River

(coefficient: 3.138937e-16), 3) average number of rooms per dwelling (coefficient: -4.909781e-16) and 4) distance to and from large Boston employment centers (coefficient: 7.046419e-17).

Below, I have included some screenshots from the model (see **Figures** section on the next page) and detailed the specific areas I learned about, including the peer-reviewed and associated sources I consulted. These various coefficients summarize the applicability to my business: High crime areas tend to lower property values, while proximity to the Charles River increases them. The average number of rooms per dwelling dramatically increases property value, as does proximity to suburban locations (i.e., recent work-from-home trends, notwithstanding).

By extension, the higher the property sales prices, the more residential rents usually follow suit. This correlation was confirmed by multiple (and recent) articles in the Boston Business Journal (Welker, 2024), the Federal Reserve Bank of Boston (Loewenstein et al., 2023), rental agency Boston Pads (Salpoglou, 2024), and the Boston Globe (Fonseca, 2024). Moreover, even ignoring any research I have completed, it stands to reason that since Boston continues to be an attractive location due to our many healthcare options, universities, high-tech and start-up culture, sports teams, and a strong sense of community, high residential property prices and a robust rental market should continue for the foreseeable future.

The final step was researching City of Boston resources related to announced development plans. In 2014, then-Mayor Martin Walsh published a study entitled Housing a Changing City: Boston 2023 (Walsh, 2014), and the latest publication was last updated on September 24, 2024, under the Mayor Michelle Wu administration on the 2025 Boston Housing strategy (Boston Housing, 2024). I also checked the city's Planning Department website (Overview, City of Boston, 2024) to read neighborhood updates and, finally, reports from Zillow (Boston, MA Housing Market, 2024) and Redfin (Redfin, 2024).

Figures

Figure 1: Variable Overview

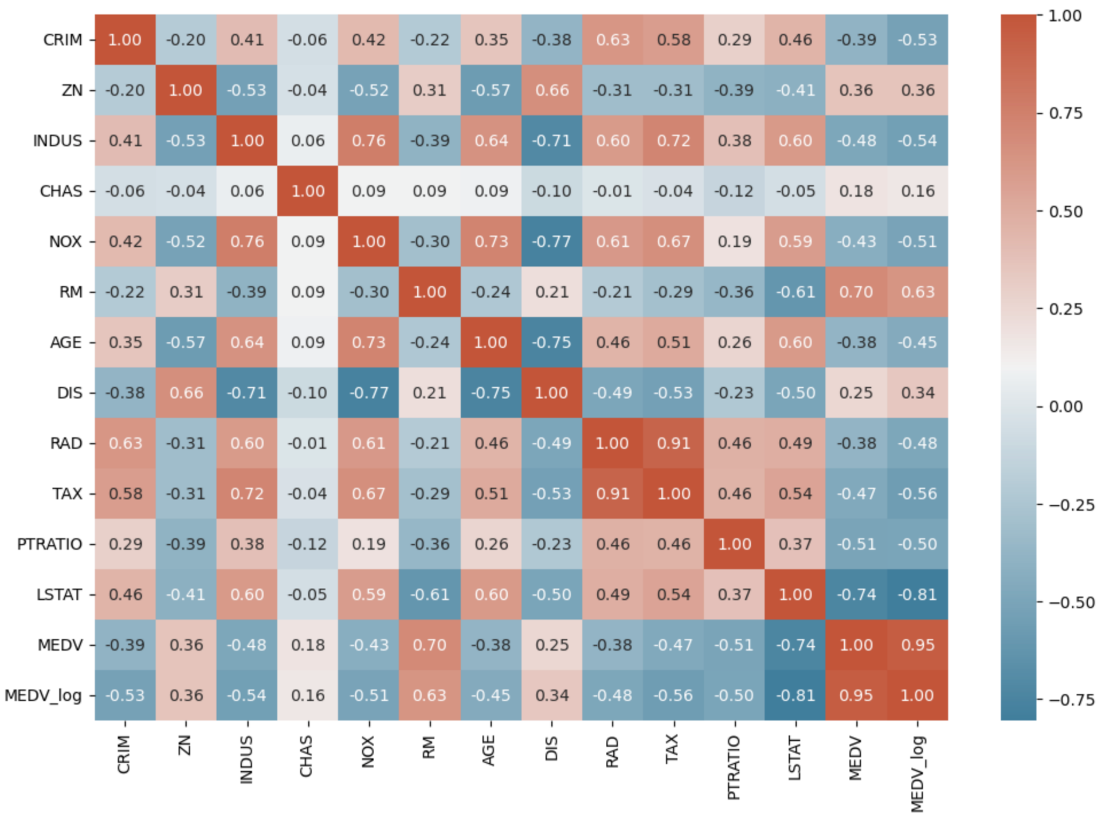
- **CRIM:** Per capita crime rate by town
  - Around 75% of the crime rate falls between ~0-4 with a max of 88 suggesting a possible **outlier**
- **ZN:** Proportion of residential land zoned for lots over 25,000 sq.ft.
  - Over 50% have 0% have residential land zoned for lots over 25,00sq.ft with the max 100%, suggesting this is **perhaps a rare commodity**.
- **INDUS:** Proportion of non-retail business acres per town
  - Ranges from 0.4-27% with an average of 11%, suggesting most towns have some industrial businesses.
- **CHAS:** Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
  - With a mean of 0.069 only ~7% of houses bound the Charles River.
- **NOX:** Nitric Oxide concentration (parts per 10 million)
  - Ranges from 0.38-0.87 with a average of 0.55. Distribution looks nominal.
- **RM:** The average number of rooms per dwelling
  - Ranges from 3.5-8.7 with an average of 6.2. Distribution looks nominal.
- **AGE:** Proportion of owner-occupied units built before 1940
  - Ranges from 2.9-100y with an averaga of 68y. Distribution looks nominal.
  - **Min age of 2.9y indicates that no houses in the database are newly built**
- **DIS:** Weighted distances to five Boston employment centers
  - Ranges form 1.1-12.1 with an average of 3.7. Distribution looks nominal.
- **RAD:** Index of accessibility to radial highways
  - Ranges from 1-24 with over 75% being the max 24.
  - There is a **large jump from the 50th percentile (5) and 75th percentile (24)**. Speculating that perhaps there are 2 cathegories of houses, those in rural areas and those more urban.
- **TAX:** Full-value property-tax rate per 10,000 dollars
  - Ranges from 187-711 with and average of 408. Distribution looks nominal.
  - **That range suggests these are mid to high income houses.**
- **PTRATIO:** Pupil-teacher ratio by town
  - Ranges from 12.6-22 with an avergage of 18.4. Distribution looks nominal.
- **LSTAT:** % lower status of the population
  - Ranges from 7-37.9% with an average of 12%. This indicates that most areas have little lower socio-economic class.
  - **The jump from 75th percentile (16.9%) to the max (37%) is indicative of a lower socio-economic area or less likely an outlier**
- **MEDV:** Median value of owner-occupied homes in 1000 dollars
  - Ranges from 5k-50k with an average of 22. Distribution looks nominal.

Figures, continued

Figure 2: Data Overview

	count	mean	std	min	25%	50%	75%	max
CRIM	6072.0	3.613524	8.593749	0.00632	0.08199	0.25651	3.67822	88.9762
ZN	6072.0	11.363636	23.301315	0.00000	0.00000	0.00000	12.50000	100.0000
INDUS	6072.0	11.136779	6.854135	0.46000	5.19000	9.69000	18.10000	27.7400
CHAS	6072.0	0.069170	0.253764	0.00000	0.00000	0.00000	0.00000	1.0000
NOX	6072.0	0.554695	0.115773	0.38500	0.44900	0.53800	0.62400	0.8710
RM	6072.0	6.284634	0.701980	3.56100	5.88500	6.20850	6.62500	8.7800
AGE	6072.0	68.574901	28.123348	2.90000	45.00000	77.50000	94.10000	100.0000
DIS	6072.0	3.795043	2.103802	1.12960	2.10000	3.20745	5.21190	12.1265
RAD	6072.0	9.549407	8.699367	1.00000	4.00000	5.00000	24.00000	24.0000
TAX	6072.0	408.237154	168.384361	187.00000	279.00000	330.00000	666.00000	711.0000
PTRATIO	6072.0	18.455534	2.162983	12.60000	17.40000	19.05000	20.20000	22.0000
LSTAT	6072.0	12.653063	7.134589	1.73000	6.93000	11.36000	16.96000	37.9700
MEDV	6072.0	22.532806	9.188768	5.00000	17.00000	21.20000	25.00000	50.0000

Figure 3: Correlation Heat Map



Figures, continued

Figure 4: Regression Results

OLS Regression Results

Dep. Variable:	MEDV_log	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	2.283e+31
Date:	Wed, 09 Oct 2024	**Prob (F-statistic):**	0.00
Time:	15:20:13	**Log-Likelihood:**	1.3856e+05
No. Observations:	4250	**AIC:**	-2.771e+05
Df Residuals:	4238	**BIC:**	-2.770e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-4.748e-15	1.05e-15	-4.521	0.000	-6.81e-15	-2.69e-15
CRIM	8.385e-17	4.35e-18	19.281	0.000	7.53e-17	9.24e-17
CHAS	3.139e-16	1.05e-16	2.983	0.003	1.08e-16	5.2e-16
NOX	1.762e-15	4.36e-16	4.040	0.000	9.07e-16	2.62e-15
RM	-4.91e-16	5.51e-17	-8.905	0.000	-5.99e-16	-3.83e-16
DIS	7.046e-17	2.04e-17	3.452	0.001	3.04e-17	1.1e-16
RAD	-2.134e-16	7.88e-18	-27.065	0.000	-2.29e-16	-1.98e-16
TAX	5.944e-18	4.11e-19	14.454	0.000	5.14e-18	6.75e-18
PTRATIO	1.498e-16	1.6e-17	9.391	0.000	1.19e-16	1.81e-16
LSTAT	-7.935e-17	7.08e-18	-11.200	0.000	-9.32e-17	-6.55e-17
MEDV	2.269e-17	1.17e-17	1.943	0.052	-2.01e-19	4.56e-17
MEDV_log	1.0000	2.95e-16	3.38e+15	0.000	1.000	1.000

Omnibus:	266.968	**Durbin-Watson:**	1.092
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	318.120
Skew:	0.664	**Prob(JB):**	8.34e-70
Kurtosis:	3.180	**Cond. No.**	1.88e+04

Figure 5: Correlation Coefficients

	Feature	Coefs
0	const	-4.747917e-15
1	CRIM	8.384767e-17
2	CHAS	3.138937e-16
3	NOX	1.762344e-15
4	RM	-4.909781e-16
5	DIS	7.046419e-17
6	RAD	-2.133581e-16
7	TAX	5.943928e-18
8	PTRATIO	1.498065e-16
9	LSTAT	-7.935020e-17
10	MEDV	2.269265e-17
11	MEDV_log	1.000000e+00

Areas I Learned About

a) **Univariate Analysis**: I had heard this term harkening back to my MBA statistics courses, but now I needed to appreciate the nuances. I read some peer-reviewed articles by Ivanova et al. (2021) for general knowledge and then mainly for real estate analysis (Whieldon, 2020).
b) **Log Transformation**: I was again familiar with the terminology but not its application. So, I found a very interesting set of articles by Bellégo et al. (2022) and Raymaekers et al. (2024) on interpreting the results.
c) **Correlations based on the dependent variable**. I liked the GitHub author's use of a heatmap to check the correlations. Agarwal et al. (2021) had a fascinating article about tracking a city's pulse through 3-D analysis. Li (2021) combined house price indices with machine learning, while Singh et al. (2020) discussed a two-way predictor approach for real estate.
d) **Removing Outliers from Dataset**: Based on college and graduate statistics classes, I understand what outliers are, but I did not realize it was possible to remove them, and certainly not with Python code. To learn more, I checked the peer-reviewed literature by reading articles from Kothandapani (2021), Yehia et al. (2022), and Rampini et al. (2022), especially those relating to real estate.
e) **Splitting the Dataset**: I was unaware that it is possible to split a data set into dependent and independent variables and then run a train and test set. It was also fascinating to see how multicollinearity was introduced. I read an interesting article by Birba (2020) outlining the general approach by Tran et al. (2022) relating to sci-kit-learn and then by Mulyanto et al. (2022) for detection and elimination in Python.
f) **Significance with Linear Regression Model**: I needed to read this section of the GitHub notes several times to understand it. Next, I had to study some articles. James et al. (2023) provided an excellent introductory overview, Capretto et al. (2020) discussed how to fit linear models, and finally, I read Khorsavi et al.'s piece from 2022 as it related to real estate applicability.
g) **Significance of Regression Coefficients**: The next step was to determine the significance of the regression coefficients. This time, I started with our Predictive Analytics book for the course (Delen, 2020) to establish a baseline and then read articles by Sgroi et al. (2021), Baak et al. (2020) and then a non-academic posting on StackOverflow (2015).
h) **Homoscedasticity**: I had no idea what this term meant when I began this project. It was not in Delen's (2020) book, so I again first resorted to peer-reviewed articles, such as those by Fernandes et al. (2023) and Hodeghatta et al. (2023). The most helpful article was from a site called GeeksforGeeks (2024).
i) **Cross-validation**: To learn about this term, I wanted to focus more on real estate as it was towards the end of the model. So, I found some excellent articles by Pai et al. (2020), Imran et al. (2021), and Munawar et al. (2020).
j) **Model Coefficients**: My learning crystallized when I finally understood how model coefficients can be used for predictability. This time, I focused on real estate articles and those relating to the Boston market. In this regard, I studied some good pieces by Chanasit et al. (2021), Lemeš et al. (2022), and Shahhosseini et al. (2020).

Exhibits

Exhibit 1: Non-selected Real Estate Predictive Models

a) Boston Housing: Data is too old (i.e., the model is seven years old).
b) Boston House Price predictions: I could not find an original data set.
c) BOSTON Housing (Regression Analysis w/ TensorFlow): This report covers TensorFlow well, but more predictive analysis is needed.
d) Boston Housing Prices - Evaluation & Validation: Interesting code, but the data is from 1978.
e) Predicting Boston Housing Prices: Excellent trial data, but it uses the same dataset from 1978.
f) House Price Prediction by ML & Deep Learning: This prediction uses deep learning algorithms well, but the data is not from Boston.
g) How to Build a Predictive Model in Python?: Paid site only; no trial possible.
h) Real Estate House Price Prediction Using Data Science: I love the interactive nature of this model, which allows you to input square feet, the number of bathrooms, and the number of bedrooms. However, the data is all from India.
i) House Price Prediction using Machine Learning in Python: This had some potential, but I could not get the code working without encountering copious errors.
j) Real Estate Price Prediction using Python: I used this site to better understand regressions re: predicted vs. actual prices and model performance.

Exhibit 2: Boston / Massachusetts Datasets

a) Housing Data from Zillow. The first link is the actual housing data, while the second is the Zillow website. This had much potential but was focused on Massachusetts rather than Boston.
b) USA Real Estate Dataset: This also had much potential but was only at the state level.
c) MassGIS Data: Property Tax Parcels: This had some excellent data, but it was only primarily in map form and needed to contain more information about property or rental prices.
d) Massachusetts Land Records: This is primarily from the Registry of Deeds only.
e) Massachusetts Housing Market: I like this because of the trend nature of the data, but it was only on median sales price and the number of homes sold.

Exhibit 3: Chosen Model and Data Updates

a) Boston House Price Prediction - Linear Regression: This was the code I ultimately chose.
b) Adjustments to data: The original data set was from the 1970s, with only 506 observations. I added a contemporary aspect to the data through dataset updates to crime rates, Nitrous Oxide, Real Estate taxes, etc. The current number of observations is 6,073.

References

Agarwal, S., Fan, Y., McMillen, D. P., & Sing, T. F. (2021). Tracking the pulse of a city—3D real estate price heat maps. *Journal of Regional Science*, *61*(3), 543–569.

Baak, M., Koopman, R., Snoek, H., & Klous, S. (2020). A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics & Data Analysis*, *152*, 107043.

Bellégo, C., Benatia, D., & Pape, L. (2022). Dealing with logs and zeros in regression models. *arXiv preprint arXiv:2203.11820*.

Birba, D. E. (2020). A Comparative study of data splitting algorithms for machine learning model selection.

Boston Housing Strategy 2025. Boston.gov. (2024). Accessed October 14, 2024. https://www.boston.gov/departments/housing/boston-housing-strategy-2025#:~:text=The%20Boston%20Housing%20Strategy%202025,ll%20use%20to%20assess%20progress

Boston, MA Housing Market. Zillow. (2014). Accessed October 14, 2024. https://www.zillow.com/home-values/44269/boston-ma/

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2020). Bambi: A simple interface for fitting Bayesian linear models in Python. *arXiv preprint arXiv:2012.10754*.

Chanasit, K., Chuangsuwanich, E., Suchato, A., & Punyabukkana, P. (2021). A real estate valuation model using boosted feature selection. *IEEE Access*, *9*, 86938-86953.

Delen, D. (2020). *Predictive analytics: Data mining, machine learning and data science for practitioners*. FT Press.

Fernandes, A. P., Fonseca, A. R., Pacheco, F. A. L., & Fernandes, L. S. (2023). Water quality predictions through linear regression-A brute force algorithm approach. *MethodsX*, *10*, 102153.

Fonseca, C. (2024, September 1). Rents are high, vacancies are low. Yep, it's another September in Boston. *The Boston Globe*. Accessed October 10, 2024. https://www.bostonglobe.com/2024/09/01/business/boston-rents-september-apartments/

GeeksforGeeks (2024). Homoscedasticity in Regression, accessed October 10, 2024. https://www.geeksforgeeks.org/homoscedasticity-in-regression/

Harvard University (2024). America's Rental Housing. Accessed October 10, 2024. https://www.jchs.harvard.edu/americas-rental-housing-2024

Hodeghatta, U. R., & Nayak, U. (2023). Simple Linear Regression. In *Practical Business Analytics Using R and Python: Solve Business Problems Using a Data-driven Approach* (pp. 183–225). Berkeley, CA: Apress.

Housing a changing city: Boston 2030 plan. (2014, October). Accessed October 14, 2024. https://www.boston.gov/sites/default/files/embed/h/housing_a_changing_city-boston_2030_full_plan_1.pdf

Imran, I., Zaman, U., Waqar, M., & Zaman, A. (2021). Using machine learning algorithms for housing price prediction: the case of Islamabad housing data. *Soft Computing and Machine Intelligence*, *1*(1), 11-23.

Ivanova, M. V., Herron, T. J., Dronkers, N. F., & Baldo, J. V. (2021). An empirical comparison of univariate versus multivariate methods for the analysis of brain–behavior mapping. *Human Brain Mapping*, *42*(4), 1070-1101.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python* (pp. 69–134). Cham: Springer International Publishing.

Khosravi, M., Arif, S. B., Ghaseminejad, A., Tohidi, H., & Shabanian, H. (2022). Performance evaluation of machine learning regressors for estimating real estate house prices.

Kothandapani, H. P. (2021). A benchmarking and comparative analysis of python libraries for data cleaning: Evaluating accuracy, processing efficiency, and usability across diverse datasets. *Eigenpub Review of Science and Technology*, *5*(1), 16-33.

Lemeš, L., & Akagic, A. (2022, June). Prediction of Real Estate Market Prices with Regression Algorithms. In *International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies* (pp. 401–411). Cham: Springer International Publishing.

Li, Z. (2021, January). Prediction of house price index based on machine learning methods. In *2021 2nd International Conference on Computing and Data Science (CDS)* (pp. 472–476). IEEE.

Lowenstein L. & Willen, P. (2023). House Prices and Rents in the 21st Century. *Federal Reserve Bank of Boston*. Accessed October 10, 2024. https://www.bostonfed.org/publications/research-department-working-paper/2023/house-prices-and-rents-in-the-21st-century.aspx

Massachusetts Corporations Division (1999), Certificate of Organization. Accessed October 10, 2024. https://corp.sec.state.ma.us/CorpWeb/CorpSearch/CorpSearchFormList.aspx?sysvalue=udpRRwFe838Z76AxHjag4w--

Massachusetts Corporations Division (2008), Certificate of Organization. Accessed October 10, 2024. https://corp.sec.state.ma.us/CorpWeb/CorpSearch/CorpSearchFormList.aspx?sysvalue=udpRRwFe838Z76AxHjag4w--

Mulyanto, A. D. (2022). mVIF Package: A Tool for Detecting Multicollinearity without Dependent Variables. *MATICS: Jurnal Ilmu Komputer dan Teknologi Informasi (Journal of Computer Science and Information Technology)*, *14*(2), 70-73.

Munawar, H. S., Qayyum, S., Ullah, F., & Sepasgozar, S. (2020). Big data and its applications in smart real estate and the disaster management life cycle: A systematic analysis. *Big Data and Cognitive Computing*, *4*(2), 4.

Overview. City of Boston Planning Department. (2024). Accessed October 14, 2024. http://www.bostonplans.org/real-estate/overview

Pai, P. F., & Wang, W. C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences*, *10*(17), 5832.

Rampini, L., & Re Cecconi, F. (2022). Artificial intelligence algorithms to predict Italian real estate market prices. *Journal of Property Investment & Finance*, *40*(6), 588-611.

Raymaekers, J., & Rousseeuw, P. J. (2024). Transforming variables to central normality. *Machine Learning*, *113*(8), 4953-4975.

Redfin. (2014). *Boston Housing Market: House Prices & Trends*. Accessed October 14, 2024. https://www.redfin.com/city/1826/MA/Boston/housing-market

Rocket Mortgage (2024, September). Boston Housing Market Report. Accessed October 10, 2024. https://www.rockethomes.com/real-estate-trends/ma/boston

Salpoglou, D. (2024), 2024 Boston Apartment Rental Market Report, *Boston Pads*. Accessed October 10, 2024. https://bostonpads.com/boston-rental-market/2024-boston-apartment-rental-market-report/

Sgroi, G., Parasiliti Palumbo, G. A., Di Salvatore, V., Russo, G., & Pappalardo, F. (2021). Peak: A clever python tool for exploratory, regression, and classification data. A case study for covid-19. In *Bioengineering and Biomedical Signal and Image Processing: First International Conference, BIOMESIP 2021, Meloneras, Gran Canaria, Spain, July 19-21, 2021, Proceedings 1* (pp. 361-370). Springer International Publishing.

Shahhosseini, M., Hu, G., & Pham, H. (2020). Optimizing ensemble weights for machine learning models: a case study for housing price prediction. In *Smart Service Systems, Operations Management, and Analytics: Proceedings of the 2019 INFORMS International Conference on Service Science*(pp. 87-97). Springer International Publishing.

Singh, A., Sharma, A., & Dubey, G. (2020). Big data analytics predicting real estate prices. *International Journal of System Assurance Engineering and Management*, *11*, 208–219.

StackOverflow (2015), Find p-value (significance) in sci-kit-learn LinearRegression. Accessed October 10, 2024. https://stackoverflow.com/questions/27928275/find-p-value-significance-in-scikit-learn-linearregression

Tran, M. K., Panchal, S., Chauhan, V., Brahmbhatt, N., Mevawalla, A., Fraser, R., & Fowler, M. (2022). Python-based scikit-learn machine learning models for thermal and electrical performance prediction of high-capacity lithium-ion battery. *International Journal of Energy Research*, *46*(2), 786–794.

Welker, G. (2024). Renters in Boston make 39% less than they should to afford their housing costs, *Boston Business Journal*. Accessed October 10, 2024. https://www.bizjournals.com/boston/news/2024/07/05/boston-income-gap-for-renters-housing.html

Whieldon, L., & Ashqar, H. (2020). Predicting residential property value in Catonsville, Maryland: A comparison of multiple regression techniques. *arXiv preprint arXiv:2101.01531*.

Woods, E. (2024). An in-depth look at the Mass. rental market (Feb.), *boston.com*. Accessed October 10, 2024. https://www.boston.com/real-estate/renting/2024/02/08/in-depth-look-at-the-mass-rental-market-feb/

Yehia, T., Khattab, H., Tantawy, M., & Mahgoub, I. (2022). Removing the outlier from the production data for the decline curve analysis of shale gas reservoirs: a comparative study using machine learning. *ACS omega*, *7*(36), 32046-32061.