Humber Institute of Technology and Advanced Learning

Factors Affecting the Boston Housing Market

Machine learning Project

Submitted by:

First name	Last name	Student number
Vishva	Shah	N01580093

Introduction

Boston, being a significant U.S city, has low vacancy rates and has experienced a slowdown in multi-family housing production. This has pushed the rent making it one of the tightest cities in the nation. Much of the percentage of housing stock (approx. 57%) belonged to houses that were constructed before 1939. The housing production has since seen a decline (Paul L. McCann, November 1999, History of Boston's Economy - Growth and Transition: 1970 – 1998, 1).

The economic conditions, interest rates, etc. contributed to ever increasing prices of the housing units. This led to an increase in the wealth of families with owner-occupied homes, however, made it more difficult for younger families to buy a new home (Paul L. McCann, November 1999, History of Boston's Economy - Growth and Transition: 1970 – 1998, 1).

The housing market is affected by several factors. These include factors like inflation, economic growth, interest rates, number of households, and availability of mortgages. Factors like affordability of houses, speculative demand, number of sellers and construction of new homes could affect prices (Tejvan Pettinger, April 2022, Factors that affect the housing market, 2). These are all high-level factors that affect the housing market. However, there are elements at the individual level that could contribute to prices. Here, we look at these factors.

The data we are using here was published by Harrison Jr., David, and Daniel L. Rubinfeld. "Hedonic housing prices and the demand for clean air." Journal of environmental economics and

management 5.1 (1978): 81-102. Through its comprehensive array of features, the dataset offers a valuable insight into the complex relationship between socio-economic factors; like crime rate, proportion of non-residential businesses in the area, proximity to Charles River, nitric oxide concentration, age, tax, etc., and Boston housing prices.

Literature Review

In the past few decades, the housing market has accounted for a significant percentage of the overall GDP of the US economy, peaking at 19%. Most of the units in the market are owner occupied, a growth in the demand for housing promotes investment and employment opportunities (Introduction to U.S. Economy: Housing Market, January 2023, Congressional Research Service, 3). An economy is also affected by inflation which affects purchasing power, investment decisions and interest rates, among other things. While the un-adjusted annual rate of inflation was 3.4% for the year 2023 (Economic New Release – Consume Price Index, January 2024, 4), the 1970s had an elevated average annual inflation rate of 7.1%, topping out at 11% in 1979 (Lawrence Yun, May 2021, A Taste of the 1970s?, National Association of Realtors, 5). One of the major reasons for this high rate of inflation is excessive printing of money (Lawrence Yun, May 2021, A Taste of the 1970s?, National Association of Realtors, 5), that being said, we try to focus on the housing market. Even though there could be multiple reasons for the increase in housing prices, here, we try to understand the components that directly affected the prices in 1970s, at an individual level.

The data we're using contains information about the crime rate of the town, the proportion of land zoned for residential lots (over 25,000 sq ft.), the proportion of non-retail businesses (in acres) per town, nitric oxide concentration (per 10 million), age of the property, accessibility to the highways, the pupil-teacher ratio by town, the percentage of lower status of the population, the average number of rooms, weighted distance to the five Boston employment centers, and taxes. We also have the data of properties that bound the Charles River (it's a binary variable).

We clean the dataset acquired, adjust for outliers and aim to understand some basic characteristics of the dataset. Once the exploratory analysis is completed, we move on to applying a linear regression model. With the help of linear regression, we aim to find the leading features that impacted the median value of owner-occupied homes.

Data summary

With an emphasis on the Boston region, a thorough summary of housing characteristics in different neighborhoods is included in each record of the dataset. The dataset includes 506 entries, each being a unique town, with 14 attributes offering a multifaceted perspective on the housing landscape. A summary of the variables is provided below:

The acronym CRIM, or crime rate, stands for the rate of crime in each town and denotes the degree of public safety in that location.

The percentage of residential land zoned for lots larger than 25,000 square feet, or ZN- represents the amount of land set aside for big residential lots following land use and spatial planning regulations.

The INDUS, or percentage of land occupied by non-retail businesses, provides information on the distribution of commercial activity within each town and describes the level of industrialization within each town.

Characterizing whether a town's tract is bordered by the Charles River gives information about waterfront property. This binary variable is called CHAS (Charles River proximity).

Nitric oxide concentration, or NOX, is a measurement of the amount of nitric oxide present in parts per million, which indicates the degree of pollution and air quality.

By displaying the average number of rooms in houses, the RM (Average number of rooms per residence) gives a sense of the size and layout of housing.

A measure of owner-occupied housing stock's age distribution that provides insights into historical development patterns is called AGE (Percentage of owner-occupied units built prior to 1940).

The weighted distances to employment centers in Boston, or DIS, indicate how close each town is to the main employment hubs in the city, which has an impact on work prospects and commute habits.
RAD (Accessibility to Radial Highways) is an index reflecting transport infrastructure and connectivity that shows how accessible each municipality is to radial highways.

The property tax rate, or TAX (property tax rate), indicates the fiscal climate and tax burden in each location by providing the rate per $10,000 of assessed value.

PTRATIO, or the pupil-teacher ratio, shows the proportion of students to teachers in classrooms and shows the caliber of instruction and class sizes.

The percentage of the population with a lower socioeconomic status is represented by LSTAT (Percentage of Lower Status Population), which provides information on the demographics and economic disparities of the community.

The median value of owner-occupied homes in each town, expressed in thousands of dollars, is provided by MEDV (Median value of owner-occupied homes), a crucial target variable.

To aid in determining if a tract's median home value is greater than $30,000 (1) or not (0), a categorical variable is known as CAT.MEDV (Categorical Median Value Indicator) is used.

Below are some basic statistics that we calculated for each column :

	mean	median	min	max	range	Std. dev	length	miss_val
CRIM	3.708250	0.24522	0.00632	88.9762	88.96988	8.714712	491	0
ZN	10.733198	0.00000	0.00000	100.0000	100.00000	23.011313	491	0
INDUS	11.370692	9.90000	0.46000	27.7400	27.28000	6.828344	491	0
CHAS	0.071283	0.00000	0.00000	1.0000	1.00000	0.257560	491	0
NOX	0.553653	0.53200	0.38500	0.8710	0.48600	0.116288	491	0
RM	6.251493	6.18500	3.56100	8.7800	5.21900	0.675457	491	0
AGE	68.405703	77.00000	2.90000	100.0000	97.10000	28.277211	491	0
DIS	3.814045	3.27210	1.12960	12.1265	10.99690	2.103519	491	0
RAD	9.706721	5.00000	1.00000	24.0000	23.00000	8.789577	491	0
TAX	412.246436	337.00000	187.00000	711.0000	524.00000	169.440997	491	0
PTRATIO	18.624644	19.10000	13.60000	22.0000	8.40000	1.965453	491	0
LSTAT	12.801181	11.65000	1.73000	37.9700	36.24000	7.181111	491	0
MEDV	22.091853	20.90000	5.00000	50.0000	45.00000	8.868464	491	0
CAT_MEDV	0.142566	0.00000	0.00000	1.0000	1.00000	0.349986	491	0

Figure 1. Summary of all the statistics for each column

After thorough analysis and coding, it has been confirmed that there are zero missing data entries across all variables in the dataset.

The pie chart comparison of categorical median values ("CAT_MEDV") across various age categories reveals insightful patterns in the dataset. For areas with a lower percentage of owneroccupied units built before 1940 (0-25%), a substantial 71.4% show a categorical median value below $30,000. Similarly, in regions with a moderate percentage of older units (26-50%), the majority, accounting for 73.5%, also have a categorical median value below $30,000. As the percentage of older units increases, this trend persists: 84.5% in the 51-75% category and 88.9% in the 76-100% category show a prevalence of lower categorical median values. This suggests a potential correlation between a higher percentage of older units and lower median property values.

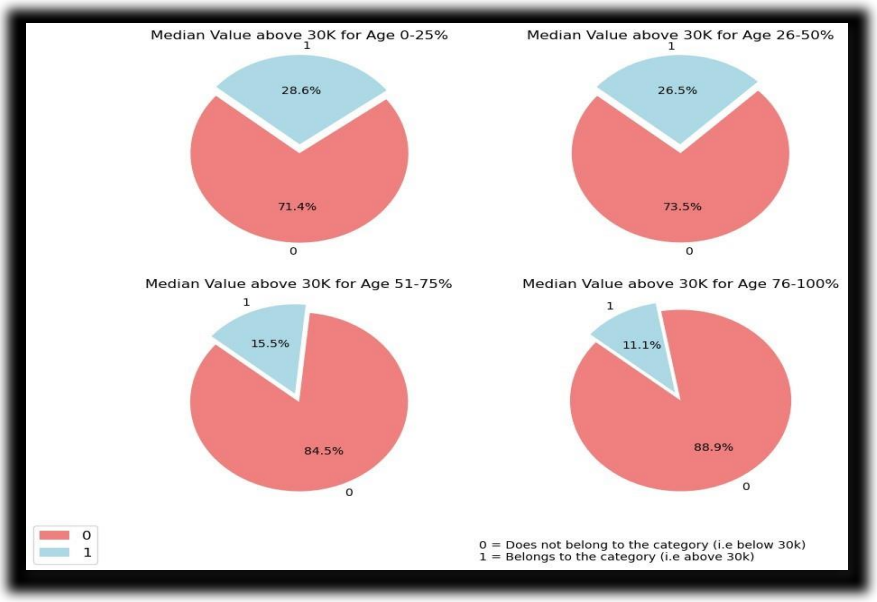


Figure 2. Age vs. Median Value: Pie Chart

Data preprocessing

During the data preprocessing step, we applied a proper procedure to address missing values. As a result, our inspection showed that the dataset has no missing values, with a sum of zero missing values across all variables. This absence assures that the dataset is comprehensive and reliable, laying a solid platform for further analysis and model development. With this achieved, we can confidently move forward with our inquiry, equipped with a clean dataset for essential exploration and interpretation.

We detected 15 outliers in the dataset for column ‘PTRATIO’ irrespective of the CAT_MEDV outcome for each. These outliers account for approximately 2.9% of the dataset. To maintain the integrity of our analysis and ensure the robustness of our findings, we made the decision to omit these outliers. Consequently, the new dataset comprises 491 cases (rows) and 14 attributes (columns). This reduction in the dataset size aligns with our aim of focusing on a more representative and refined subset of the data for subsequent analyses. The exclusion of outliers is a standard practice to enhance the reliability and interpretability of statistical analyses and model building, contributing to the overall quality of our investigation.

Exploratory data analysis

With the help of the following analysis, we aim to gain better understanding of the dataset. These visual patterns help us draw conclusions about each variable's variability, skewness, and presence of extreme values. These charts help in finding trends and understanding how the numbers are dispersed, where they tend to cluster, and whether any unusual values exist or not.

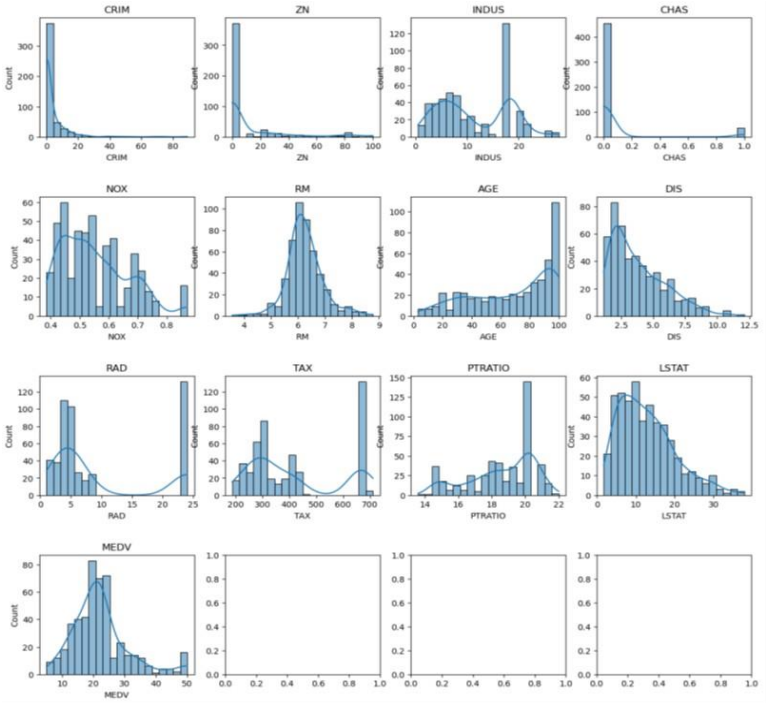


Figure 2. Histograms of the 14 variables

The feature with the largest variability is the one that shows a broad range of values on the histogram. After looking at the basic statistics and the histograms associated with each variable, we conclude that the TAX variable shows the largest variability in this case.

Skewness is determined from the shape of the histogram; whether the skewness is positive or negative is dependent on the direction of the tail in the histogram. A single side of skewed distributions has a larger tail. We search for variables whose distribution does not follow a symmetrical bell shape. Hence, skewness is seen in the variables "CRIM," "ZN," "CHAS," "AGE," "DIS," and "LSTAT." These variables show a trend towards greater or lower values, deviating from a normal distribution.

Histograms which show significant values near the tails are considered extreme values. After considering the basic statistics as well as the histogram for the variables we conclude that variables “RAD”, “TAX”, “INDUS” appeared to be extreme.

We look at the boxplots for RAD and DIS variables for each category in CAT_MEDV variable. We are trying to understand here if there are any values outside of the interquartile range for variables which calculate the distance between the employment centers (DIS), the index of accessibility to highways (RAD) for houses prices above and below 30,000 dollars mark (CAT_MEDV = 0 or 1).

The box plot for the variable 'RAD' is shown on the left subplot. The box in the plot represents where most of the data is in the middle which is the interquartile range. The line inside the box is

the middle value. Lines outside the box show the range of most values, with some exceptions (denoted by circles/dots). Similarly, the boxplot for DIS is constructed on the right side.

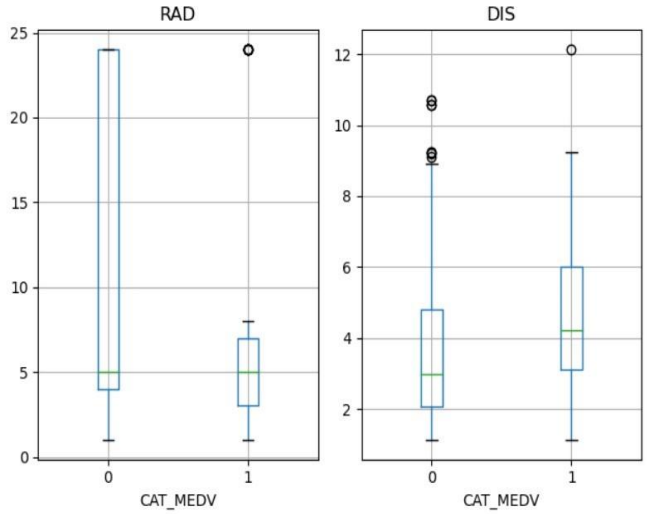


Figure 3. Boxplot comparison between Rad/CAT_MEDV and DIS/CAT_MEDV

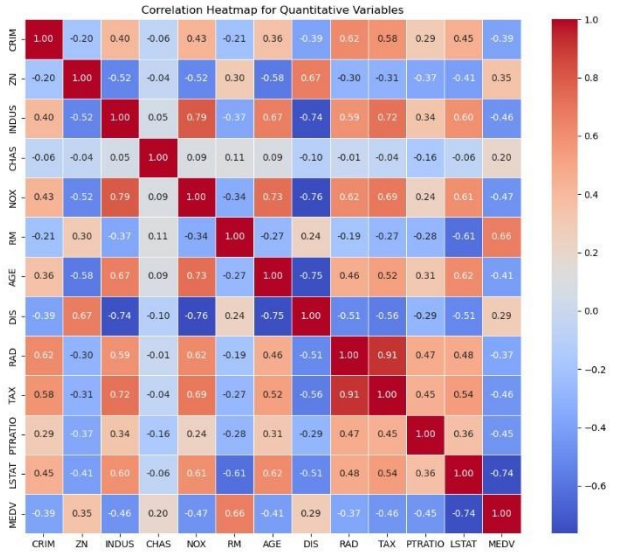
The boxplot between RAD and CAT_MEDV variables do not show a lot of outliers. The 75th percentile (Q3) is the same as the upper bound for CAT_MEDV = 0. This indicates that there are no values after the third quartile for this *category* in CAT_MEDV. However, we see outliers for CAT_MEDV = 1 category. This could be a genuine case of outlier.

The boxplot between DIS and CAT_MEDV shows several outliers for both the categories in CAT_MEDV. These outliers also seem like genuine cases of outliers. However, an in-depth analysis of the data source could help understand the outliers better.

A correlation analysis on the data was performed. First, we computed the data's correlation matrix. This matrix provides insights of the linear relationship between pairs of variables. Some variables show strong positive correlations such as TAX and RAD (0.910676), which indicates a high degree of linear association. On the contrary, some variables show negative correlations like NOX and DIS (-0.763542) which is an inverse relationship.

We can reduce the number of variables based on correlations by using techniques such as aggregation and statistics. Besides this, reducing the number of categories by combining close or similar categories works too. Principal component Analysis can be applied to transform the original variables into a smaller set of uncorrelated variables.

A heatmap was generated to visually represent the correlation matrix. With heatmaps, we can easily identify the patterns and relationships between variables. Darker shades represent stronger correlations whereas lighter shades represent weaker correlations.



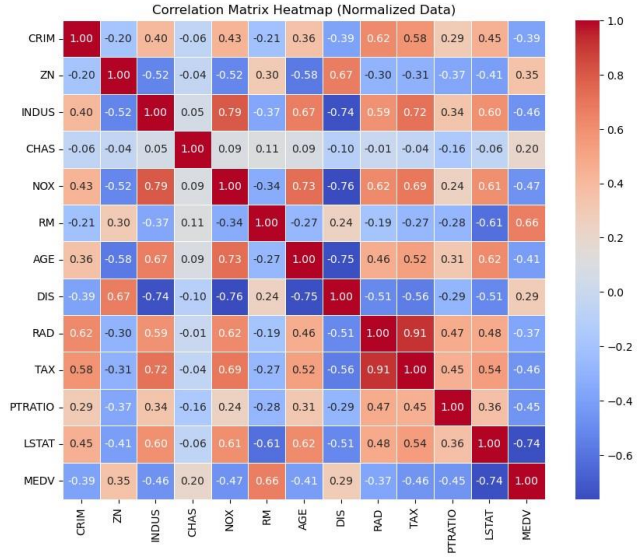


Figure 4. Correlation Heatmap for Quantitative Variables Figure 5. Correlation Matrix Heatmap (Normalized Data)

After that, data was normalized using Z-score. Normalization is done to standardize the scale of variables to ensure they contribute equally to the analysis. Despite normalization the correlation values remain unchanged.

Overall, this provided valuable insights into the interdependencies among the data attributes. To build a model and select features it is important to identify strong correlated variables. Besides this, normalization provides a comparison between variables by bringing them to a common scale.

Linear Regression

We use linear regression model to understand the most key features that change the prices of the housing units in Botson. To begin with our analysis, we use the cleaned dataset and divide it into training and test dataset at a split of 80-20.

Once we fit out model, we predict the y values and compare the forecasted values to the actual values. To understand the error rate, we use the mean squared error and compare this error to the mean squared error of the baseline model. The baseline model we consider is where the predicted values as the mean values.

The mean squared error for the baseline model is 55.93 and for the linear regression model its 12.10. This suggests the linear regression model is doing well. Hence, we can move on to understanding the most prominent features of the dataset that change the prices.

We get the following coefficients data for each variable:

	Feature	Coefficient
0	CRIM	-0.122795
1	ZN	-0.005215
2	INDUS	0.137449
3	CHAS	2.790639
4	NOX	-15.558904
5	RM	0.542332
6	AGE	0.000753
7	DIS	-0.607606
8	RAD	0.152872
9	TAX	-0.006184
10	PTRATIO	-0.551640
11	LSTAT	-0.533303
12	CAT_MEDV	12.696295

The NOX variable has the highest negative coefficient, this suggests that NOX variable decreases the prices by approximately 15.55 units. We can ignore the high coefficient for CAT_MEDV variable as this variable is derived from the MEDV variable.

The second most important variable is CHAS, this indicates that CHAS = 1 increases the MEDV values by 2.79 units.

Conclusion

After analyzing Boston Housing markets, we came to the following conclusions.

The dataset consisted of zero missing values and a few outliers of PTRATIO. Since outliers accounted for a very small percentage of the dataset, we chose to omit them. With the help of

histograms and box plots, we identified trends, skewness, and extreme values in a variety of variables. Correlation matrix and Heat maps helped us to understand the interdependencies between variables. After analyzing the matrix, we concluded that the data is on a similar scale. Linear Regression helped us understand the most important feature of the dataset that impacted the housing prices. The coefficient data provided insight into the effects of several components, with CHAS having a positive price impact and NOX (nitrous oxide content) having a negative price impact standing out as key drivers to changes in housing prices.

References

1. https://www.bostonplans.org/getattachment/15ca7a2f-56d1-4770-ba7f-8c1ce73d25b8
2. https://www.economicshelp.org/blog/377/housing/factors-that-affect-the-housing-market/
3. https://www.everycrsreport.com/files/2023-01-03_IF11327_5627e0cd68baa358be959508c171593733914f71.pdf
4. https://www.bls.gov/news.release/cpi.htm
5. https://www.nar.realtor/magazine/real-estate-news/economy/a-taste-of-the-1970s