Analyzing the Boston Housing Market: A Study of Imputation Methods and Outlier Detection Techniques

## Team Members Ansh Gangdev - N00652175 Deepanshi Garg - N01510660 Kapil Sharma - N01501686 Kivan Ilangakoon - N01510394 Saransh Vaid - N01510833

Submitted in partial fulfillment of the Machine Learning - 1 (BIA-53020) course of the Business Insights & Analytics Certificate, Humber Institute of Technology and Advanced Learning

# Abstract

*The dataset created by Harrison Jr. David and Daniel L Rubinfield, entails a spreadsheet and an article summarizing the connection between clean air and pleasant living situations in the region of Boston, Massachusetts. The dataset contains a total of 11 different attributes and a total 167 records within the dataset which was retrieved from the US Census Bureau. Within the data set, NAN (not a number) was assigned to non numeric cells, and furthermore omissions and imputations were made to outliered data. The article investigates the issues of attempting to correlate housing market data with the concept of citizens of Boston wanting to pay for clean air based upon location. The Boston Housing dataset is a valuable resource for researchers and practitioners in the field of machine learning and statistics.*

# Attributes

The data set has been categorized by 11 attributes which are:

| CRIM | per capita crime rate by town |
| --- | --- |
| ZN | proportion of residential land zoned for lots over 25,000 sq.ft. |
| INDUS | proportion of non-retail business acres per town |
| CHAS | Charles River dummy variable (1 if tract bounds river; 0 otherwise) |
| NOX | nitrogen oxides concentration (parts per 10 million). |
| RM | average number of rooms per dwelling. |
| AGE | proportion of owner-occupied units built prior to 1940. |
| DIS | weighted mean of distances to five Boston employment centres. |
| RAD | index of accessibility to radial highways. |
| TAX | full-value property-tax rate per $10,000. |
| PTRATIO | pupil-teacher ratio by town. |

There are several techniques used within the article and analysis of the dataset that were completed by the authors. Firstly is to proxy the willingness of residents to pay by measuring additional cost to society in relation to increased air pollution particularly or vice versa wherein there is a reduced cost with better air quality. Secondly, is the utilization of the analysis on the housing market, with the idea being that residents and potential residents will opt to pay more for housing in areas with less air pollution versus paying less for more polluted areas. Lastly, is the creation of a four-step model to infer conclusions based on the two previous techniques.

1. Develop an equation to determine the value of a hedonic housing value with air pollution as a factor
2. A calculation of each households maximum ability to pay based on incremental changes in air pollution alongside the hedonic housing value equation
3. Estimation of the ability/willingness to pay for households combined with a demand curve for clean air. This can also be understood as the demand for clean air within the previous two calculations.
4. Finally step 4 combines the previous calculations with air pollution data pre and post air control solutions in order to identify the value of the dollar amount of benefits after a control strategy has been implemented.

As per the data on the Excel document, there are 167 records indicated among the 11 different attributes. Different outliers were observed among these records which will be identified and highlighted in the dataset. There are 3 possible causes of outliers identified in the data:

1. Non-numeric data entries
2. Errors in the placement of decimal points during data entries
3. Genuine outliers (entries out of the interquartile range)

| CRIM | float64 |
| --- | --- |
| ZN | float64 |
| INDUS | object |
| CHAS | int64 |
| NOX | object |
| RM | float64 |
| AGE | float64 |
| DIS | object |
| RAD | int64 |
| TAX | int64 |
| PTRATIO | object |

*Table 1: Attribute data type*

The presumption as mentioned in the article states: Individuals are willing to pay more for a unit located in an area with good air quality than for an otherwise identical unit located in an area with poor quality.

Depicted below is an example of the first 5 rows of data from the data.

|  | **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.00632 | 18 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.09 | 1 | 296 | 15.3 |
| **1** | 0.02731 | 0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 |
| **2** | 0.02729 | 0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 |
| **3** | 0.03237 | 0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 |
| **4** | 0.06905 | 0 | 7.07 | 0 | 0.458 | 7.147 | 54.2 |  | 3 | 222 | 18.7 |

*Table 2: Top 5 rows of the dataset*

The table above illustrates a sample providing the first 5 rows for the Boston Housing data set. As seen it provides numerical values for all the 11 given attributes, and in one case there is even a null cell, which will be addressed as a NAN (not a number) entry.

| CRIM | 0 |
| --- | --- |
| ZN | 0 |
| INDUS | 28 |
| CHAS | 0 |
| NOX | 6 |
| RM | 0 |
| AGE | 0 |
| DIS | 4 |
| RAD | 0 |
| TAX | 0 |
| PTRATIO | 0 |

*Table 3: Number of outliers in all column except ‘PTRATIO’*

The table above shows a total count of all the outliers (except for PTRATIO) identified in the data set. The outliers here include non-numeric data entries, errors in the placement of decimal points, and entries that were outside the interquartile range.

|  | **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.00632 | 18 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.09 | 1 | 296 | 15.3 |
| **1** | 0.02731 | 0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 |
| **2** | 0.02729 | 0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 |
| **3** | 0.03237 | 0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 |
| **4** | 0.06905 | 0 | 7.07 | 0 | 0.458 | 7.147 | 54.2 | nan | 3 | 222 | 18.7 |
| **5** | 0.02985 | 0 | nan | 0 | 0.458 | 6.43 | 58.7 | 6.0622 | 3 | 222 | 137 |
| **6** | 0.08829 | 12.5 | 7.07 | 0 | 0.524 | 6.012 | 66.6 | 5.5605 | 5 | 311 | 15.2 |
| **7** | 0.14455 | 12.5 | nan | 0 | 0.524 | 6.172 | 96.1 | 5.9505 | 5 | 311 | 15.2 |
| **8** | 0.21124 | 12.5 | 7.87 | 0 | 0.524 | 5.631 | 100 | 6.0821 | 5 | 311 | 15.2 |
| **9** | 0.17004 | 12.5 | nan | 0 | 0.524 | 6.004 | 85.9 | 6.5921 | 5 | 311 | 15.2 |
| **10** | 0.22489 | 12.5 | 7.87 | 0 | 0.524 | 6.377 | 94.3 | 6.3467 | 5 | 311 | 15.2 |

*Table 4 : Sample of highlighted NAN values for all columns except ‘PTRATIO’*

Table 4 above gives an indication of the data that have been adjusted for the non numerical entries. As seen they have been identified and highlighted in red and labelled with ‘nan’.

| CRIM | 0 |
| --- | --- |
| ZN | 0 |
| INDUS | 28 |
| CHAS | 0 |
| NOX | 6 |
| RM | 0 |
| AGE | 0 |
| DIS | 4 |
| RAD | 0 |
| TAX | 0 |
| PTRATIO | 3 |

*Table 5 : Total invalid values in the data set including ‘PTRATIO’*

Table 5 above shows the total count of all the outliers, inclusive of PTRATIO identified in the data set. It is seen that a total of 3 outliers have been identified in the data set for PTRATIO.

|  | **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **100** | 0.14866 | 0 | 8.56 | 0 | 0.52 | 6.727 | 79.9 | 2.7778 | 5 | 384 | nan |
| **101** | 0.11432 | 0 | 8.56 | 0 | 0.52 | 6.781 | 71.3 | 2.8561 | 5 | 384 | 20.9 |
| **102** | 0.22876 | 0 | nan | 0 | 0.52 | 6.405 | 85.4 | 2.7147 | 5 | 384 | 20.9 |
| **103** | 0.21161 | 0 | 8.56 | 0 | 0.52 | 6.137 | 87.4 | 2.7147 | 5 | 384 | 20.9 |
| **104** | 0.1396 | 0 | 8.56 | 0 | 0.52 | 6.167 | 90 | 2.421 | 5 | 384 | 20.9 |
| **105** | 0.13262 | 0 | 8.56 | 0 | 0.52 | 5.851 | 96.7 | 2.1069 | 5 | 384 | 20.9 |
| **106** | 0.1712 | 0 | 8.56 | 0 | 0.52 | 5.836 | 91.9 | 2.211 | 5 | 384 | 20.9 |
| **107** | 0.13117 | 0 | 8.56 | 0 | 0.52 | 6.127 | 85.2 | 2.1224 | 5 | 384 | 20.9 |
| **108** | 0.12802 | 0 | 8.56 | 0 | 0.52 | 6.474 | 97.1 | 2.4329 | 5 | 384 | 20.9 |
| **109** | 0.26363 | 0 | 8.56 | 0 | 0.52 | 6.229 | 91.2 | 2.5451 | 5 | 384 | 20.9 |
| **110** | 0.10793 | 0 | nan | 0 | 0.52 | 6.195 | 54.4 | 2.7778 | 5 | 384 | nan |
| **111** | 0.10084 | 0 | 10.01 | 0 | nan | 6.715 | 81.6 | 2.6775 | 6 | 432 | 17.8 |
| **112** | 0.12329 | 0 | 10.01 | 0 | 0.547 | 5.913 | 92.9 | 2.3534 | 6 | 432 | 17.8 |
| **113** | 0.22212 | 0 | 10.01 | 0 | 0.547 | 6.092 | 95.4 | 2.548 | 6 | 432 | 17.8 |
| **114** | 0.14231 | 0 | 10.01 | 0 | 0.547 | 6.254 | 84.2 | 2.2565 | 6 | 432 | 17.8 |
| **115** | 0.17134 | 0 | 10.01 | 0 | 0.547 | 5.928 | 88.2 | 2.4631 | 6 | 432 | 17.8 |
| **116** | 0.13158 | 0 | 10.01 | 0 | 0.547 | 6.176 | 72.5 | 2.7301 | 6 | 432 | 17.8 |
| **117** | 0.15098 | 0 | 10.01 | 0 | 0.547 | 6.021 | 82.6 | 2.7474 | 6 | 432 | nan |

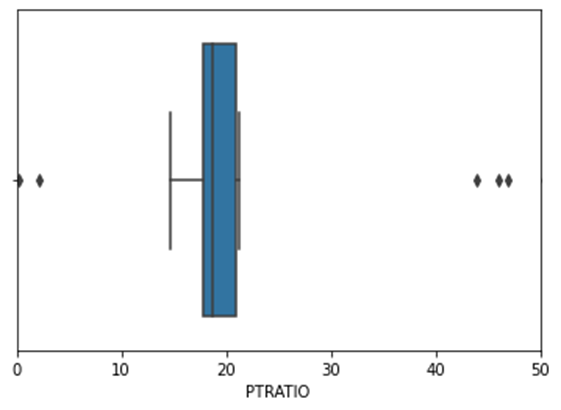
*Table 6: ‘PTRATIO’ non-numeric values highlighted*

In table 6, we can see the 3 outliers identified in the PTRATIO columns, and they have been respectfully highlighted and labelled as ‘nan’.

| CRIM | -0.166208 |
| --- | --- |
| ZN | -18.75 |
| INDUS | -4.89 |
| CHAS | 0 |
| NOX | 0.313 |
| RM | 5.03375 |
| AGE | -27.575 |
| DIS | -1.348225 |
| RAD | 0 |
| TAX | 81.5 |
| PTRATIO | 13.15 |
| dtype: | float64 |
| CRIM | 0.457333 |
| ZN | 31.25 |
| INDUS | 18.95 |
| CHAS | 0 |
| NOX | 0.673 |
| RM | 7.30775 |
| AGE | 163.825 |
| DIS | 9.450375 |
| RAD | 8 |
| TAX | 565.5 |
| PTRATIO | 25.55 |
| dtype: | float64 |

*Table 7: Interquartile ranges for box plot*

In the table above we can see an illustration of the interquartile ranges identified for all the attributes. We have considered just the interquartile ranges for PTRATIO. The first quartile of the PTRATIO is 13.15 and the third quartile of the PTRATIO is 25.55. This can now be used to construct the whiskers on the box plot, as illustrated on the next diagram.



*Illustration 1: Box plot of data set*

A box plot diagram has been illustrated to show the distribution of the data, and help identify the outliers in the data set. As seen in the diagram below, the illustration of the box plot looks at the PTRATIO and shows the outliers that are indicated by the diamond shapes on the ends of the box plot diagram. The interquartile range has been indicated by the whiskers of the box plot. On one end is the first quartile at 13.15, and the third quartile at 25.55.

|  | **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.00632 | 18 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.09 | 1 | 296 | 15.3 |
| **1** | 0.02731 | 0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 |
| **2** | 0.02729 | 0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 |
| **3** | 0.03237 | 0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 |
| **6** | 0.08829 | 12.5 | 7.07 | 0 | 0.524 | 6.012 | 66.6 | 5.5605 | 5 | 311 | 15.2 |
| **8** | 0.21124 | 12.5 | 7.87 | 0 | 0.524 | 5.631 | 100 | 6.0821 | 5 | 311 | 15.2 |
| **10** | 0.22489 | 12.5 | 7.87 | 0 | 0.524 | 6.377 | 94.3 | 6.3467 | 5 | 311 | 15.2 |

*Table 8: Omitted the rows with NAN values*

Captured in table 8 is the result after the omissions of the highlighted cells. These cells are the fields that were non numeric fields, and therefore captured as ‘nan’. As seen, omissions of rows have been done with such fields.

|  | **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.00632 | 18 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.09 | 1 | 296 | 15.3 |
| **1** | 0.02731 | 0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 |
| **2** | 0.02729 | 0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 |
| **3** | 0.03237 | 0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 |
| **4** | 0.06905 | 0 | 7.07 | 0 | 0.458 | 7.147 | 54.2 | 4.169953 | 3 | 222 | 18.7 |
| **5** | 0.02985 | 0 | 9.122878 | 0 | 0.458 | 6.43 | 58.7 | 6.0622 | 3 | 222 | 137 |
| **6** | 0.08829 | 12.5 | 7.07 | 0 | 0.524 | 6.012 | 66.6 | 5.5605 | 5 | 311 | 15.2 |
| **7** | 0.14455 | 12.5 | 9.122878 | 0 | 0.524 | 6.172 | 96.1 | 5.9505 | 5 | 311 | 15.2 |
| **8** | 0.21124 | 12.5 | 7.87 | 0 | 0.524 | 5.631 | 100 | 6.0821 | 5 | 311 | 15.2 |
| **9** | 0.17004 | 12.5 | 9.122878 | 0 | 0.524 | 6.004 | 85.9 | 6.5921 | 5 | 311 | 15.2 |
| **10** | 0.22489 | 12.5 | 7.87 | 0 | 0.524 | 6.377 | 94.3 | 6.3467 | 5 | 311 | 15.2 |

*Table 9: Imputed NAN values with Mean values*

Table 9 illustrates the imputations made on the data set. These are outlier figures that have been replaced by the average value of their respective attribute. Unlike omissions, these rows remain a part of the data set, except that the values are replaced with the average of all the valid values in the specific attribute.

Mentioned within the article, the procedural model is established when considering the following hypothesis:   
  
**Hypothesis:** **Households consider the level of air pollution as well as the quantity and quality of housing and other neighbourhood characteristics in making their household choices.**

Equation: Harrison Jr. David and Daniel L Rubinfield touches on the theoretical model by creating the following equation and variables to measure the utility function as is depicted.Text

Description automatically generated

# Hedonic Findings

According to Harrison and Rubinfeld, the results displayed a fairly sensitive attachment to the details of the Hedonic housing price equation and strategy mentioned earlier in the methodology presented, in comparison to that of the air quality demand equation.

This means that among the researchers present in the article, none found a direct correlation/ influence within the value set in terms of improvements in air pollution concentration with the level of air pollution and is exclusive of household income and popular choices as a result.They also claimed that the procedure used in the study identified that marginal air pollution damages increased alongside the levels of air pollution when connected with higher household incomes.

Finally, it can be surmised that the aggregate benefit estimate was directly influenced within the findings of the constant marginal valuations as well as being unbiased towards the willingness to pay function mentioned in the model.