Data Analytics Approach: Boston Housing Market

Alexander Swanepoel[1]

[1]**University of Warwick, Computer Science MSc**

This paper seeks to undergo exploratory analysis to derive unforeseen relationships in the Boston Housing Data [6]. We employ machine learning functionalities yielding classification and regression descriptions of presupposed relationships found between crime rate per capita and pupil-teacher ratios. We also demonstrate remarkable data visualisation with respect to span.

Keywords: Machine Learning, Data Mining, Spatial Data Analysis,Python, R, WEKA, Exploratory Analysis, Data Visualisation

INTRODUCTION

The pioneering research with respect to this topic was conducted by Harrison, D. and Rubinfeld, D.L. [6], elucidating the methodological concerns of expressing the readiness to pay for clean air as a function of housing market data. The analysis derived a model describing households willingness to purchase, in which the conclusion could be drawn, that contrary to assumption, an increase in air pollution as described by nitric oxide concentration in flute gases i.e. free air, increase with the level of household income with respect to the Boston area.

Sequentially a follow-up study was conducted by Bivand, R. [5], displaying that by extension of data to include spatial domains, i.e. spatially lagged independent variables, the conclusion can be drawn that considering the willingness to pay, and by proxy median house values as a function of air pollution, contrarily does increase the willingness to pay by a factor of three. The research denotes the fundamental value of spatial consideration, showing observational units are of deep importance when considering the particular dataset.

DATASETS

I. Boston Housing Data

The pioneering study performed as mention above, whilst drawing valid and interesting conclusions, remains incomplete in exploratory analysis; we look to explore the data and conduct analysis to for potential underlying correlations revealed. The set contains information as below,

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	nitric oxides concentration 10ppm.
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted 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
B	$1000(AFAM - 0.63)^2$ where AFAM is the proportion of African-Americans by town
LSTAT	% lower status of the population
MEDV(Class)	Median value of owner-occupied homes in $1000's

II. Boston Housing Data Appended

By extension the appended dataset containing spatial dimensions (longitude and latitude), will also be included in exploratory analysis.

Software

Python

A strongly typed programming language, within Python, the following packages were implemented: NumPy, SciPy, Pandas, matplotlib.pyplot and seaborn.

R

A functional strongly typed programming language, used primarily for data analysis and predictive modelling. Within R the following packages were implemented: ggmap, ggplot2, mapproj, catools.

WEKA

WEKA is a GUI (graphical user interface) providing machine learning algorithms and modelling capabilities. Predominantly used for classification, regression, visualisation or clustering, the program offers a no-code solution to machine learning and data mining.

Anaconda Command Line Prompt

A command Line Tool useful for the usual command line functionalities as well as package installation and data analysis.

Microsoft Excel

A spreadsheet formatted cell based program primarily used for its analysis and accounting functionalities, as well as data conversion and storage facilities.

Data Cleaning

The data required cleaning via NumPy using the replace NaN function `.replace(np.NaN,'')` replacing all missing datapoints with a 0. Datasets also required labelling and conversions to CSV filetypes, and by proxy to ARFF filetypes for use in WEKA.

Hypothesis

It may be reasonable to assume that due to observational qualitative analysis, a sparse relationship suggesting the the per capita crime rate by town is a function of pupil-teacher ratio by town, due to restriction of available resources declining quality of education, by proxy reducing education of the subsequent population and reducing the earning potential of individuals.

ANALYSIS

We begin exploratory analysis(EDA, exploratory data analysis) in WEKA and python, by visualising the raw data to identify any prospective correlations, this will be the basis of the examination. To begin, correlation weightings were added to each prospective attribute contained within the dataset, for an approximate measure of the trends. The r^2 correlation coefficient [1][7] was computed such that for a set of n variables, $y(n) \in y_1 ... y_n$ such that, for a predicted value $f(n) \in f_1 ... f_n \exists\ e_i = y_i - f_i$ where e_i is the residual measurement of the predicted value $f(n)$ and the measured value $y(n)$, then the variance is given by $S_{Var} = \sum_i (y_i - \bar{y}_i)^2$ where $\bar{y}$ is the observed value and S_{Var} is the sum of the squares. From this, $S_{res} = \sum_i (y_i - f_i)^2$, the sum of the residuals, gives $r^2 = 1 - \frac{S_{res}}{S_{var}}$; which is plotted into the heatmap via seaborn as shown in Figure 1.

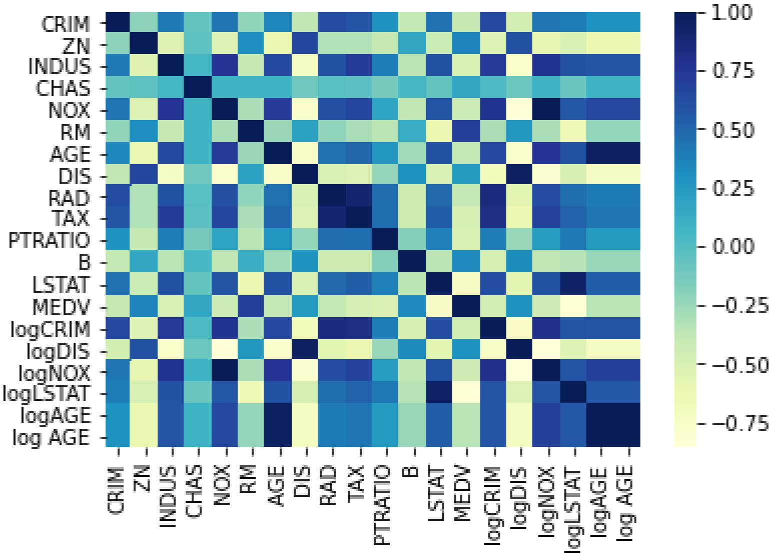


Figure 1. Correlation Of Determination, Heatmap Plot, Python

By observation we are able to see that there are some emergent correlations between NOX-INDUS, ZN-DIS, as well as RAD-TAX, as well as slight correlations in LSTAT-INDUS, and CRIM-PTRatio. Whilst correlations of NOX and INDUS are valid, this has already been previously explored. Moreover, we stipulate the relationship between ZN-DIS is uninteresting due to the potential colinearity; However, EDA is inconclusive and we seek to continue. The CRIM-PTRatio poses an interesting prospect, supported by a wide variety of literature, however we will stagnate on this particular pairwise analysis for now.

We are able to view the correlations graphically via WEKA, for a more intuitive view.

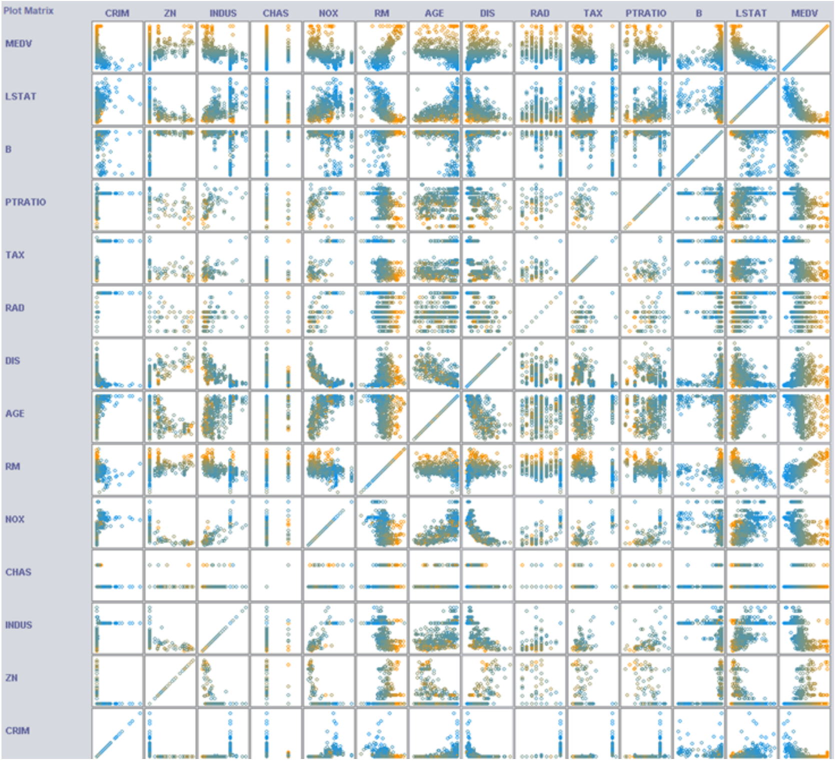


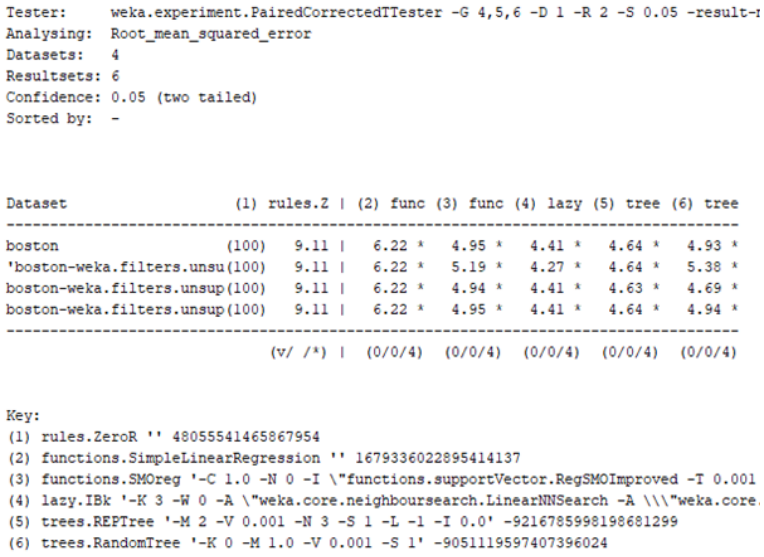
Figure 2. Correlation Visualisation, WEKA

By observation again, one may recognise the distinct graphical correlations between attributes such as DIS-NOX, AGE-NOX; moreover, between LSTAT-MEDV and RM-MEDV. These are noteworthy distinctions, and we seek to delve into the analysis for these relationships. The WEKA funcionality allows for a novel interface such that one is able to build an experiment with comparison of a multitude of datasets and machine learning (ML) algorithms. Within the functionality, we seek to normalise the data on a $\forall\, y(n) \in [0,1]$ basis, in order to smooth for ML algorithms; we also standardise the set, assigning attributes mean values set=0 ; Variance =1. Standardisation allows logistic regression and Naïve-Bayes algorithms that act on Gaussian distribution presuppositions to function smoothly.

Classification, Regression, and Predictive Modelling

We run analysis on multiple datasets$\{$numeric, auto-selected, normalise, standardised$\}$, as well as a plethora of ML algorithms, which we will take a brief digression to comment. We seek to implement the algorithms k-Nearest Neighbours (kNN) denoted in WEKA as IBk, simple linear regression, a C4.5 variant implemented in WEKA as REPTree, k-Random, implemented in WEKA as RandomTree and lastly a Support Vector Machine algorithm, known as SMOReg. This will form an adequate basis for analysis, where we will use the ZeroR(WEKA's r correlation classifier) as a benchmark for comparison.

In this instance, we will not adjust input parameters, and run the algorithms as default, on the training set. The results are as follows,



```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-r
Analysing:  Root_mean_squared_error
Datasets:   4
Resultsets: 6
Confidence: 0.05 (two tailed)
Sorted by:  -


Dataset                   (1) rules.Z | (2) func (3) func (4) lazy (5) tree (6) tree
--------------------------------------------------------------------------------------
boston                    (100)   9.11 |   6.22 *   4.95 *   4.41 *   4.64 *   4.93 *
'boston-weka.filters.unsu(100)   9.11 |   6.22 *   5.19 *   4.27 *   4.64 *   5.38 *
boston-weka.filters.unsup(100)   9.11 |   6.22 *   4.94 *   4.41 *   4.63 *   4.69 *
boston-weka.filters.unsup(100)   9.11 |   6.22 *   4.95 *   4.41 *   4.64 *   4.94 *
--------------------------------------------------------------------------------------
                              (v/ /*) |  (0/0/4)  (0/0/4)  (0/0/4)  (0/0/4)  (0/0/4)


Key:
(1) rules.ZeroR '' 48055541465867954
(2) functions.SimpleLinearRegression '' 1679336022895414137
(3) functions.SMOreg '-C 1.0 -N 0 -I \"functions.supportVector.RegSMOImproved -T 0.001
(4) lazy.IBk '-K 3 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.
(5) trees.REPTree '-M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0' -9216785998198681299
(6) trees.RandomTree '-K 0 -M 1.0 -V 0.001 -S 1' -9051119597407396024
```

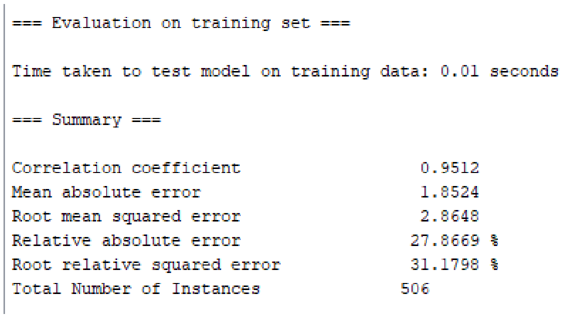
Figure 3. Algorithm Comparison on 4 Datasets Defined Above

All error values for their respective models are recorded and tabulated above; one can conclude, all algorithms outperform the R correlation classifier substantially. Note that in WEKA, statistically significant results are denoted by '*' providing clarity on the deviations. We also identify that there is negligible difference between the normalisation and standardisation filtered datasets, and whilst the auto-selection filter provides small improvement, we choose to return to the numeric corrected values of our initial dataset, as the improvement is also negligible.

Let us conclude from the result, k-Nearest Neighbours[4][2] under default parameters performs the most efficiently. Due to this deduction, we will adjust the parameters of k-Nearest Neighbours accordingly, for completeness. Post analysis it was identified that the values of $k \in k\{1,3,5,9,12,15\}$ such that $k = \{1,3\}$ produced the lowest errors when adjusted for Manhattan distances. The kNN algorithm considers pairwise vectors (X_1, Y_1) where Y is the class variable, projected onto some $\mathscr{L}_n$ vector norm where the norm is given in the form $||\bullet|| \in \mathbb{R}^d$. Then, if $\sum_{i=1}^{n} w_{ni} = 1$ is the weighting for the i'th nearest neighbour, k nearest neighbours can be mathematically described as assigning a weighting $1/k$ to the i'th node of the network, and 0 to all others. It follows that if subject to the correct conditions and post-asymptotic expansion, one may derive the weighting equation,

$$w^*_{ni} = \frac{1}{k*}\left[1+\frac{d}{2}-\frac{d}{2k^{k*2d}}\{i^{1+2/d}-(i-1)^{1+2/d}\}\right]$$

$\forall\ i = 1,2,...,k^*\ \forall\ w^*_{ni}\ \in\ i = k^*+1,...,n$ where $k^* = [B_n \frac{4}{d+4}]$ where B is some constant to be calculated, and d is the distance dimensionality of the norm. Then, we are able to see the reasoning behind the substantial efficacy of the Manhattan distance relative to the Euclidean distance. The set we for a given model with a high dimensionality, d, it is preferable to use smaller values of a function, p[3]. The implication being that the L1 distance metric is the most desirable for our specific high dimensionality dataset, due to the dimensionality of the Euclidean distance being given by $||x_2|| := \sqrt{x_1^2+...+x_n^2} = \sqrt{\mathbf{x}\cdot\mathbf{x}}$; $\forall n$ and the Manhattan distance given by $||x_1|| = \sum_i |x_i|$. Figure 4 allows us to deduce all instances were correctly classified, with a correlation coefficient of 0.95, deeming it statistically significant. With respect to our initial question, this suggests that the combination of attributes considered has consequences on the Median value of owner-occupied homes.



```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correlation coefficient                  0.9512
Mean absolute error                      1.8524
Root mean squared error                  2.8648
Relative absolute error                 27.8669 %
Root relative squared error             31.1798 %
Total Number of Instances              506
```

Figure 4. k-Nearest Neighbours Classification for $k = 3$

Visualisation

Resuming the previous proposition of exploring the relationships between CRIM-PTRatio, and the class variable itself. We seek to apply some R analysis, to aid visualisation of the variables. Implementation of the `get_map()` function in the ggmaps package allows us to retrieve jpeg images of the input longitude and latitude shown in figure 5.

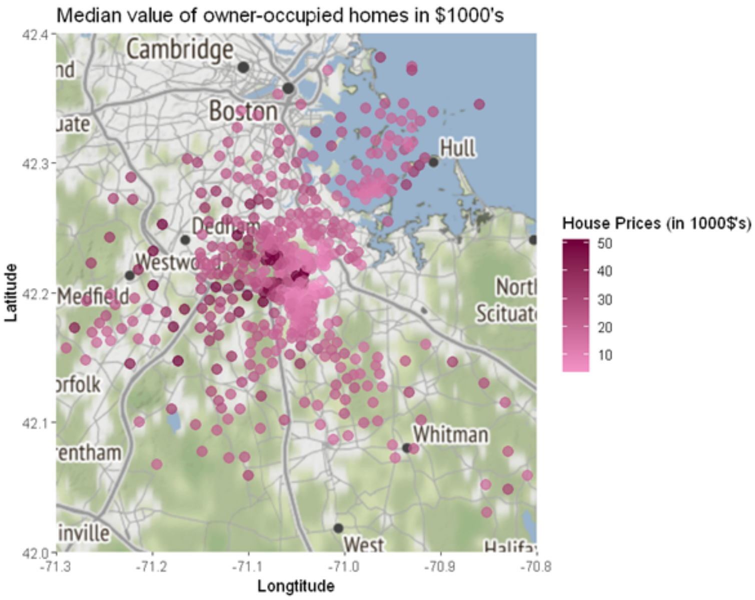


Figure 5. MEDV Class Variable Visualisation

The visualisation brings elements of reality to surface, allowing for tangible representation of the data. We can discern that the concentration of high valued owner-occupied homes are located in clusters around the town of Dedham, and on the border of Dedham and Milton; within the visual one can note that the high median-valued owner-occupied homes are scattered towards not only the left hand side of the map, however also in between the lowest quartile end of the median-value. Let us further delve into the analysis, by assigning quantile transformations to the measurements assigned to the class variable in the domain [0,1] implemented by the `cutpoints <- quantile(params,'')` in R.

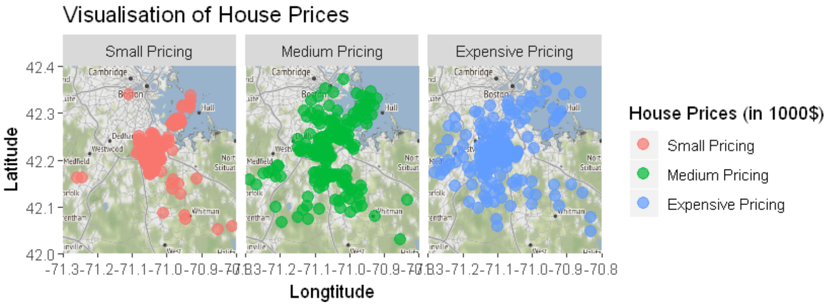


Figure 6. MEDV Class Variable Quantile Visualisation

Let us remark on the distribution of the smallest quantile of median-valued owner occupied homes, the area is concentrated around the bayside and and central Boston location. Upon further inspection, the location specified for the density of low-quantile median value housing is within a reasonable radius of the University of Massachusetts Boston, by speculation, this may be responsible for the high concentration, as students may wish to purchase their own homes if they so choose to remain in the city of study, and are within the junior years of their career; however this remains inconclusive as data remains sparse with respect to categories of home ownership. We wish to repeat subsequent analysis to aid visualisation for the other attributes.

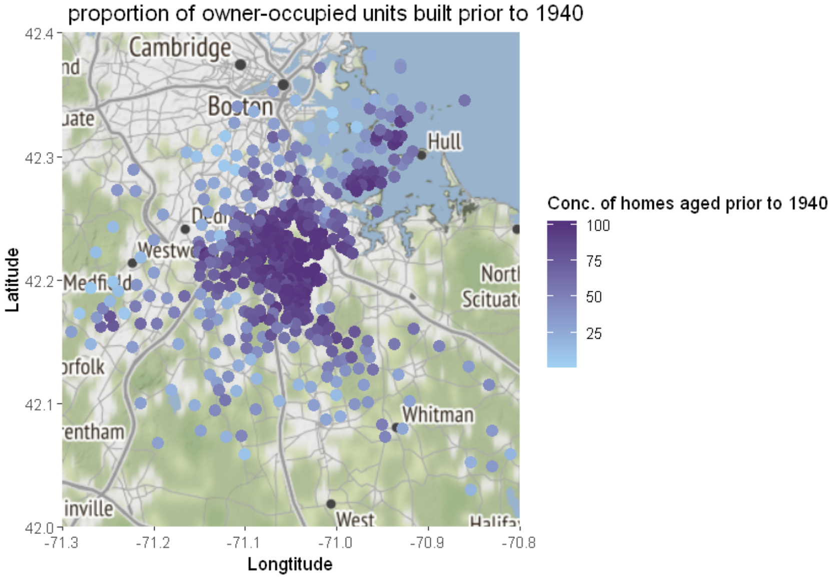


Figure 7. AGE Variable Visualisation

Also remark, the age distribution of the proportion of owner-occupied units built prior to 1940; we see a remarkable clustering similar to that of the MEDV class, with densities highly concentrated in the upper quantile of the span. However, we refer to a previous heatmap, for distribution of correlations 1. As shown above, the bimodal overlap may be coincidental, however correlation is not equivalent to causation. Commencing with our previous proposition, we see there is some sparse overlapping

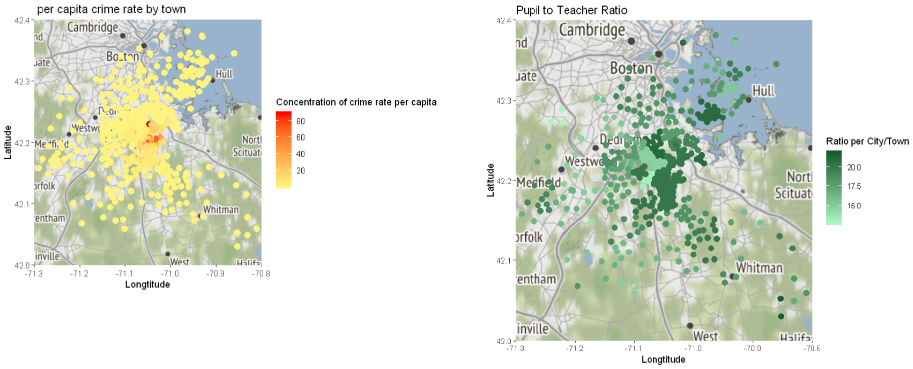
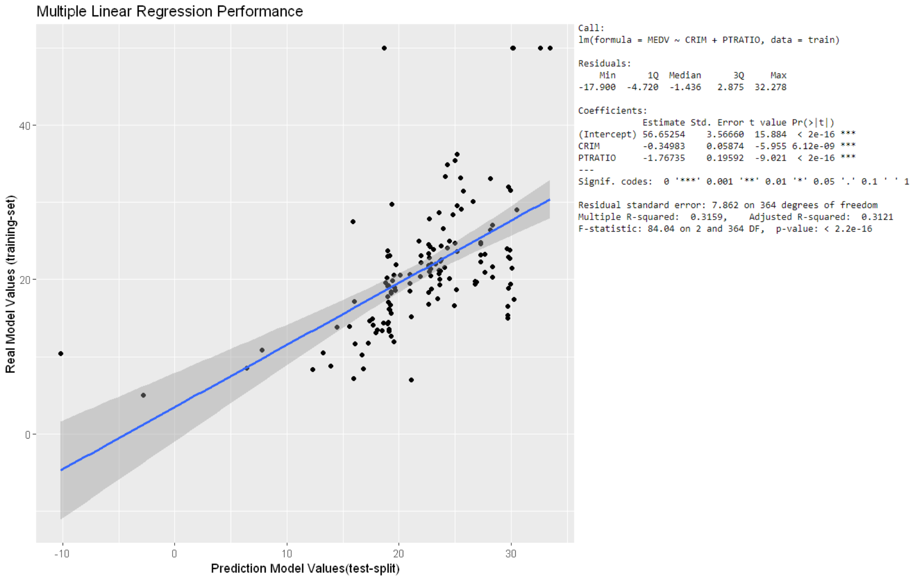


Figure 8. AGE Variable Visualisation

between the PTRatio (Pupil-Teacher Ratio) and CRIM (crime rate per capita), the sparse overlap may be statistically insignificant, however we are unable to draw conclusion as the visualisation remains ambigious, whilst all visualisation is qualitative.

We hope to achieve a conclusion under quantitative foundation, therefore we will effectuate a multiple linear regression in R, whilst undergoing statistical analysis for verification. We choose to split the data into a training and test set, where Data={Training =0.7, Test =0.3}. Our implementation of multilinear regression requires a tangible graphical representation to supplicate our understanding.



```
Call:
lm(formula = MEDV ~ CRIM + PTRATIO, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-17.900  -4.720  -1.436   2.875  32.278

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.65254    3.56660  15.884  < 2e-16 ***
CRIM        -0.34983    0.05874  -5.955 6.12e-09 ***
PTRATIO     -1.76735    0.19592  -9.021  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.862 on 364 degrees of freedom
Multiple R-squared:  0.3159,	Adjusted R-squared:  0.3121
F-statistic: 84.04 on 2 and 364 DF,  p-value: < 2.2e-16
```

Figure 9. MultiLinear Regression Predicted vs. Real Plot, With Statistics

We commence dissection of statistics remarking the F-statistic; as shown in the tabular values, the degrees of freedom $Var_1 = 2$; $Var_2 = 364$ at a $P_r = 0.05$ the F-critical value is surpassed by the F-statistic hence we are able to reject the null hypothesis; the claim proceeding is there is a relationship between the Pupil-Teacher Ratio and the Crime rate per capita. When inquiring the graphical rendition we note the shaded area equivalent to statistically significant datapoints within a 0.95 confidence interval. Whilst there may be a statistically significant relationship, the dataset provided clusters outside of the significant span, as a perfectly fit model would have all points idealised on the line of the graph.

CONCLUSION

We underwent exploratory data analysis to shown emergent properties within the data yielding a basis of digression for quantitative analysis. We denoted an interesting sparse correlation between NOX-

INDUS,ZN-DIS, as well as RAD-TAX, as well as slight correlations in LSTAT-INDUS, and CRIM-PTRatio. Whilst valid, we elected to explore the unforeseen correlations in the dataset by previous studies, answering and posing new potential research questions. Upon progression, we examined a classification model to depict the optimal dataset for the problem, and conclusively showed that the dataset should be numeric and have k-nearest neighbours as a classifier, under Manhattan distance parameterisation, whilst forming a pedagogical approach for understanding. We commenced to procure a set of visualisation graphs for selected attributes of interest stated previously in the paper, that provided intuition to a difficult to conceptualise dataset.

From this, we concluded that the null hypothesis rejected stabilising a statistically significant relationship between the crime rate per capita per town, and the pupil-teacher ratio.

We conclude, whilst objective satisfaction occured in unveiling previously unforeseen trends in the data, the set itself is sparse and open to wild variance, therefore previous analysis may have been unable to reveal hidden trends with statistical surety. We require that further data may be collected to provide a larger breadth of instances and datapoints for a more thorough conclusive analysis.

Extensions

Possible extensions include previously identified yet unexplored trends in the exploratory data analysis stage of the proposal, namely; trends between NOX-INDUS, ZN-DIS, as well as RAD-TAX. Potential analysis could also be conducted on the possible relationship between AGE-MEDV as denoted in the visualisation section of the paper; the attributes appear to have a bimodal overlapping, allowing for interesting stipulations. A natural extension of the research would be the implementation of the procedure and techniques in the paper extended to other geographical domains.

REFERENCES

[1] Coefficient of determination, R^2. https://en.wikipedia.org/wiki/Coefficient_of_determination.

[2] k-Nearest Neighbours Algorithm. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

[3] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In J. Van den Bussche and V. Vianu, editors, *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[5] R. Bivand. Revisiting the boston data set - changing the units of observation affects estimated willingness to pay for clean air. *REGION*, 4(1):109–127, May 2017.

[6] D. Harrison and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81 – 102, 1978.

[7] O. Heinisch. Steel, r. g. d., and j. h. torrie: Principles and procedures of statistics. (with special reference to the biological sciences.) mcgraw-hill book company, new york, toronto, london 1960, 481 s., 15 abb.; 81 s 6 d. *Biometrische Zeitschrift*, 4(3):207–208, 1962.