Danmarks Tekniske Universitet

DTU

Machine learning & Data mining Report 1

Authors

James Alexander Cowie - s192911
Ghassen Lassoued - s196609
Chandykunju Alex - s200113

March 3, 2020

Contents

1 Introduction

The main objective of this project is to understand the standard data prepossessing methods and practise in most machine learning and data mining approaches.

This paper investigates a data set named "hprice2" which contains information on perceived relevant metrics related to generation of a hedonistic housing price model.
The aim is to explore some of the structures and relationships present in the data, with the goal of attempting future classification tasks. The techniques employed in the analysis are chosen with this in mind. Therefore, firstly the data set will get an introduction, followed by a description of the measured attributes and their properties. Lastly the relationship and quality of the data is explored.

2 Data

The data set HPRICE2 is taken from "Introductory Econometrics: A Modern Approach, 6e" by Jeffrey M. Wooldridge where it was used for hedonistic pricing modelling. Specifically predicting housing prices in the Boston Standard Metropolitan Area (SMSA) using data from 1970.

The data has no singular origin, but was collected from a plethora of sources. The first use of the data we could find is from the paper: "Hedonic Housing Prices and the Demand for Clean Air," published in Journal of Environmental Economics and Management 5, 81-102 by by D. Harrison and D.L. Rubinfeld.

Comparison of some basic attribute statistics serve as a simple quality check of the data. The computed values being compared with those presented in table V in Appendix B of the previously mentioned paper.

The attributes of this data set are a subset of data found in original paper. These values were used in what is referred to as the "basic housing value equation" (p. 98 of the paper) to quantify a measure of "willingness to pay for air quality improvements:

$$\begin{aligned} logMV = a_1 + a_2 RM^2 + a_3 AGE + a_4 logDIS + a_5 RAD + a_6 TAX \\ + a_7 PTRATIO + a_8 (B - 0.63) + a_9 logLSTAT + a_{10} CRIM \\ + a_{11} ZN + a_{12} INDUS + a_{13} CHAS + a_{14} NOX^p + \varepsilon \end{aligned}$$

Where MV , RM, DIS, RAD, TAX, PTRATIO, LSTAT, CRIM and NOX correspond to the attributes price, rooms, dist, radial, proptax, stratio, lowstat, crime and nox of HPRICE2.

a1, a2,.... , a14 are calculated using least squares regression and .p value is calculated using grid search on this model, They reached in the conclusion that p = 2 is the best fit, and resulting in a R2 of 0.81.

Because our data set is only a subset of the original we reduce the model to:

$$log(price) = a_1 + a_2 rooms2 + a_3 log(dist) + a_4 log(radial)$$
$$+a_5 proptax + a_6 stratio + a_7 log(lowstat) + a_8 crime + a_9 nox^2 + \varepsilon$$

Let's look at the possibilities of machine learning methods can be implemented in this data:

1. Classification: For doing classification crime rates would be a best choice of attributes. It would give the insight about high and low crime rates based on the grouping.

2.Regression: Since price is a time based attribute, it can be used for prediction.Especially when it comes to regression this would be a wise choice.

3.Clustering : Clustering can be perform many attributes The features like price, crime, nox, rooms, dist, radial, proptax, stratio and lowstat would be good choice for clustering techniques.

3 Attributes

At the first sight we would say that there are 12 attributes in the data set. However, the last three attributes lprice,lnox and lproptax are logarithmic transformations of price, nox and proptax respectively.

3.1 Attributes description

price: discrete since all values are integer and not real. It's ratio because zero means absence of price. However, in real life we don't expect the median price of a house to be zero dollars. But still, zero has a meaning.
crime: continuous since values are real. Having no crime has a meaning, so also zero here has a meaning. Therefore, it is ratio.
nox: for the same reasons mentioned above, it is continuous and ration.
rooms: continuous since values are real. A house with no rooms, i.e a house in which the average number of rooms is zero, has a meaning. So it is ratio.
dist: continuous and ratio for the same reasons.
radial: discrete since all values or integer because it is an index. It is also ordinal because the objects are ranked and can be sorted
proptax: continuous since values are real. A tax of zero has a meaning even if it is not realistic. Therefore it is ratio
strratio: continuous since values are real. It is ratio because the average students-teacher can be zero meaning that there is a teacher who has no students i.e absence of students that leads to absence of what is measured here.
lowstat: continuous since values are real. It is ratio because the percentage of lower status people can be zero and has a meaning
For the last three variables lprice,lnox and lproptax , the logarithm was applied. We can assume that zero can be calculated since 0=log(1). So zero here means absence for the loga-

rithm variables but not the absence of the original variables. More likely, these 3 logarithm attributes are ratio even though there is no physical meaning to zero.
The results are summarized in figure 1.

3.2 Basic summary statistics of the attributes.

This summary allows us to better understand our variables and theirs observations. The results are summarized in the figure 1.
This summary shows the difference of scale in our attributes such as between price and nox. That's why we need a standardization.

price continuous

	price	crime	nox	rooms	dist	radial	proptax	stratio	lowstat	lprice	lnox	lproptax
Description	Median housing price, $	Crimes committed per capita	Nitrous oxide, parts per 100 mill.	Avg. number of rooms per house	Weighted dist. to 5 employment centers	Accessibiliy index to radial highways	Property tax per $1000	Average student-teacher ratio	% of people 'lower status'	log(price)	log(nox)	log(proptax)
Type	Discrete, Ratio	Continuous, Ratio	Continuous, Ratio	Continuous, Interval	Continuous, Ratio	Discrete, Ordinal	Continuous, Ratio	Continuous, Ratio	Continuous, Ratio	Continuous, Ratio	Continuous, Ratio	Continuous, Ratio
count	506	506	506	506	506	506	506	506	506	506	506	506
mean	22511.5	3.61154	5.54978	6.28405	3.79575	9.54941	40.8237	18.4593	12.7015	9.94106	1.69309	5.9314
std	9208.86	8.59025	1.1584	0.702594	2.10614	8.70726	16.8537	2.16582	7.23807	0.409255	0.20141	0.396367
min	5000	0.006	3.85	3.56	1.13	1	18.7	12.6	1.73	8.51719	1.34807	5.23111
25%	16850	0.082	4.49	5.8825	2.1	4	27.9	17.4	6.9225	9.73209	1.50185	5.63121
50%	21200	0.2565	5.38	6.21	3.21	5	33	19.1	11.36	9.96176	1.68269	5.79909
75%	24999	3.677	6.24	6.62	5.1875	24	66.6	20.2	17.0575	10.1266	1.83098	6.50129
max	50001	88.976	8.71	8.78	12.13	24	71.1	22	39.07	10.8198	2.16447	6.56667

Figure 1: Summary of attributes in HPRICE2

3.3 Data issues

There are no missing values in the data set. This data set is taken from a book named "Introductory Econometrics: A Modern Approach, 5th Edition". In the description in the site where we found it it was written that "Diego Garcia, a former Ph.D. student in economics at MIT, kindly provided these data, which he obtained from the book Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, by D.A. Belsey, E. Kuh, and R. Welsch, 1990. New York: Wiley. Data loads lazily. " This data set were used to teach. So we will assume there is no corrupted data.

4 Visualization

4.1 Distribution of attributes

To predict housing prices logic would tell us that the amount of rooms, crime and lowstat are highly correlated with housing price. Indeed upon inspection of the scatter plots of

all attributes (Figure A.1) in relation to one another there seems to be correlation, at least between rooms and lowstat. Lowstat in this case having a negative relation.

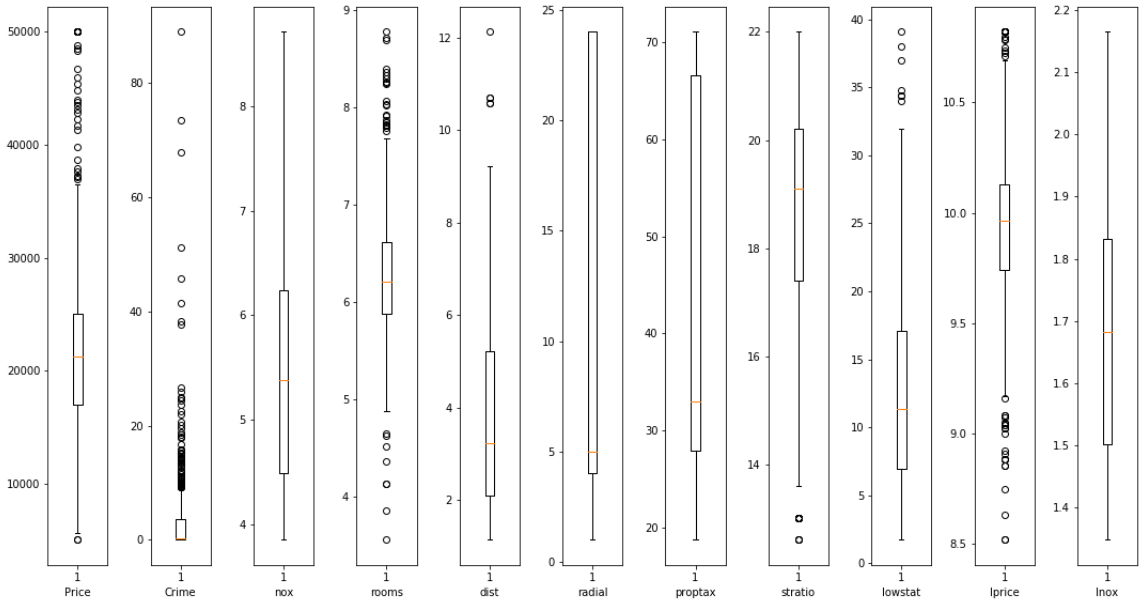

Figure 2: Boxplots of attribute distributions

In Figure 2 the distribution of the attribute values are plotted using boxplots. Observing the plots it appears the crime variable has an extremely sharp peak and long tail of values that lie outside the main distribution. We therefore define:

$$crime = log(crime) \tag{1}$$

The reasoning becomes apparent upon examination of the attributes histograms.

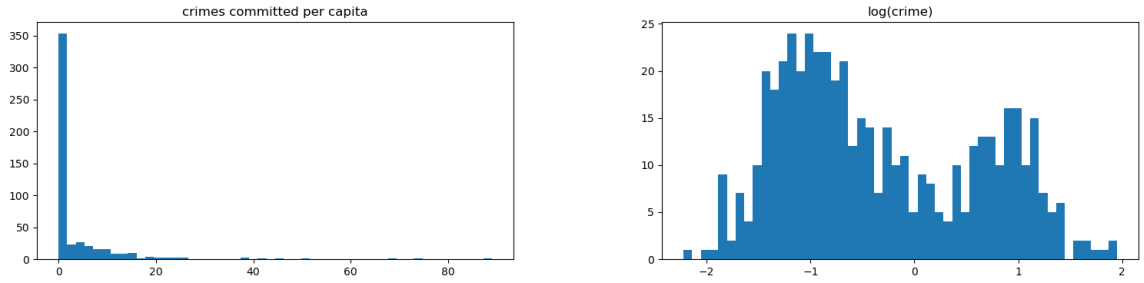

Figure 3: Histogram of crime attribute

It is near impossible to infer anything from the un-transformed distribution, however the logarithmic transformation indicates that there are two potential classes definable, of

high and low crime areas, separable around 0. Histogram plots of other attributes are also given below.

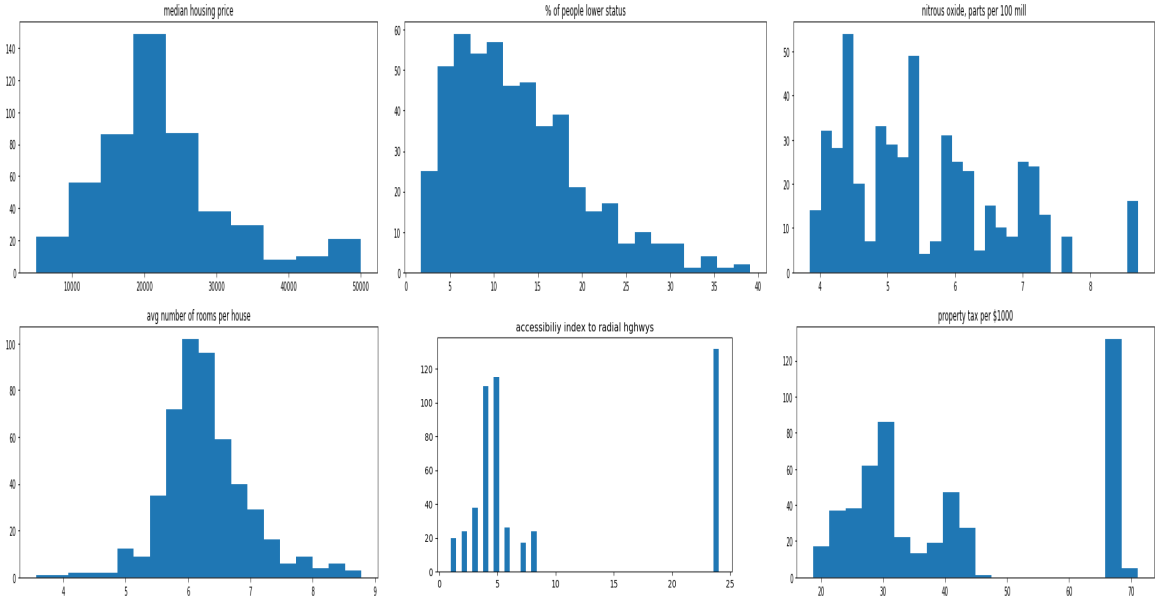

Figure 4: Histogram of some attributes

There are some distinguishable distributions present. With rooms appearing to have a super gaussian distribution, given the length of the "tails" evident from Figure 3.

Let's have a look at the scatter plots presented in "Appendix A".

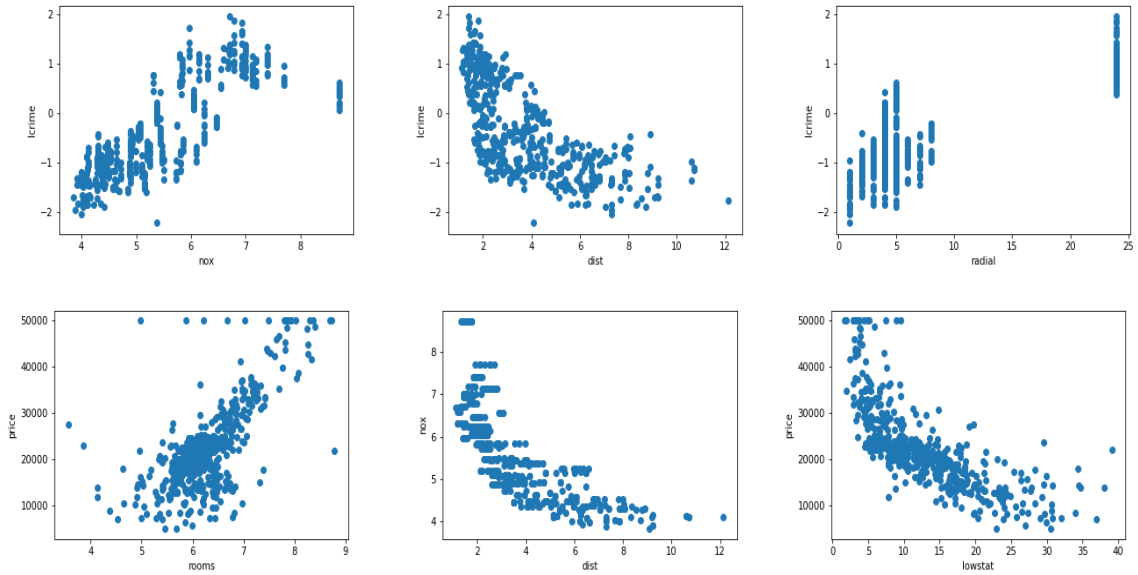

Figure 5: Scatter plots of some attributes

It gives the idea that how the attributes are correlated with each other. Moreover it gives the visual representation of how these attributes will fit in to the linear regression model and to help us determine the correlation between them. As we can see in the plots the housing price and average number of rooms are proportional, which comes as no surprise. Moreover the price is inversely proportional to the lower status metric, also to be expected. In addition to that the attributes nox and the distance are well correlated. That means the pollution is higher near to the city area.

Based on the study we have done it is concluded that price prediction using a linear regression model will be a nice choice to pursue.

4.2 Principle Component Analysis

The main aim of PCA is to find a lower dimensional representation of higher dimensional data. With the objective of emphasizing the variation in the data. This means we deemphesize the directions that don't contribute much to the variation, and project the data onto axis' defined by the directions that maximize variance.
The PCA algorithm generally goes as follows:

1. Subtract the mean
2. Divide by standard deviation (optional)
3. Compute the SVD: $\mathbf{USV^T} = \bar{\mathbf{X}}$
4. Project $\bar{\mathbf{X}}$ onto the subspace defined by the amount of principle components.

In Figure 6 the direction of the attribute projections are plotted along with a graph describing the data variance explained as a function of principle components.

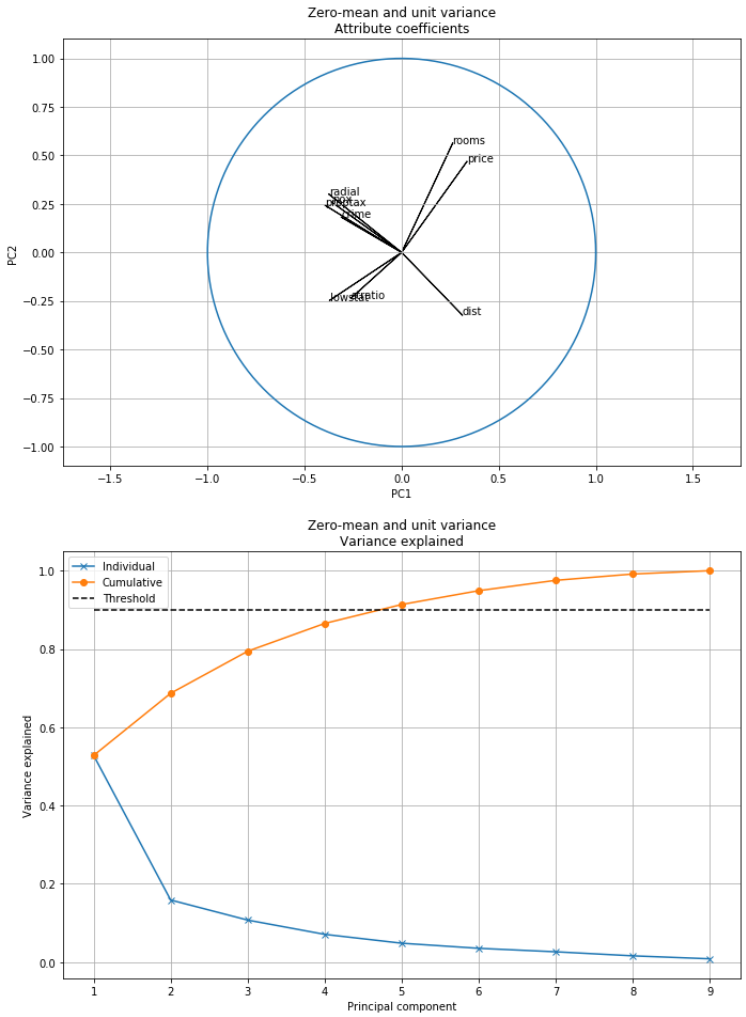

Figure 6: Attribute coefficients and variance explained

The result shows that 5 principle components are sufficient in order to explain 90% of the variance. Figure 7 shows a projection of the data onto the axis' defined by the first two principle components.

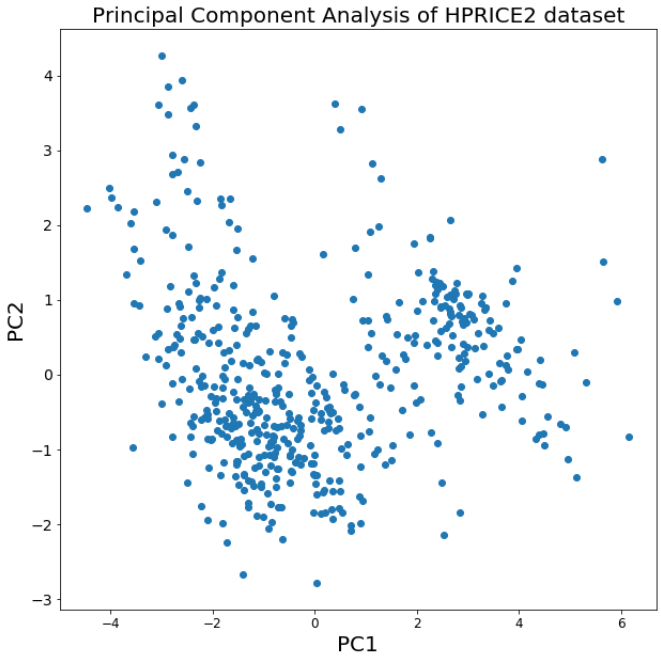

Figure 7: Data projected onto the first 2 principle components

There are some potentially classifiable clusters that are observable here. Note that a variation of 1 along the PC1 axis explains more of the variance in the data than a variation of 1 along the PC2 axis. We saw in Figure 6 that the first two principle components can explain roughly 70% of the variance. In Figure 8 the first three components are plotted

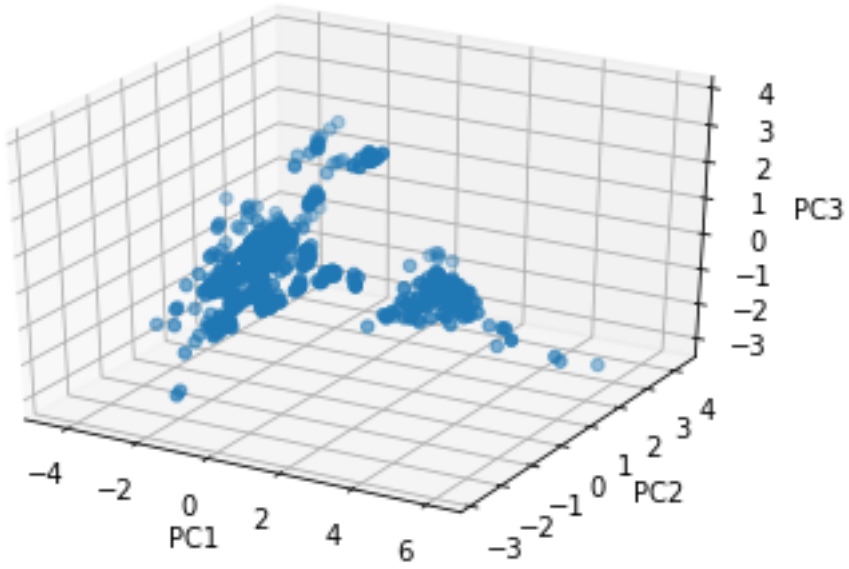

Figure 8: Data projected onto the first 3 principle components

While 3D plotting is not generally encouraged it serves to roughly illustrate the fact that including more principle components explains more of the data.

5 Discussion

While we examine the correlation between the attributes, it is concluded that this data will perform well with linear regression algorithm. The rooms and lowstat attributes being particularly important in the price prediction. Moreover it is difficult to find a relation between nox and price, potentially indicating a lack of concern regarding pollution.

The attribute nox is related to radial, lcrime and dist, this means there are connections between this atrributes in city but we couldn't find a valid relation between [rooms, price, lowstat] and [lcrime, nox, nox, dist].These two groupings of attributes could be treated as independent dimensions of houses. Referring PCA plot, Figure 6, the almost orthogonal relation between these two groupings can be obtained on the projection of the principle component. This analysis also gives the conclusion that the suitable attribute groups are [lowstat, stratio, rooms, price] and [nox, dist, lcrime, proptax, radial].

6 Conclusion

Finding relationship between attributes in the data HPRICE2 is the major task of this project. We found that every individual house has two major qualities. The first quality we term status: High price = large number of rooms, few low status people etc. And the second one is whether is is urban or not. The urban area is determined by quality of air, low distance to job, more crime and tax on property. From the principal component analysis we are able to visualize the independent nature of these two groups. From the study and analysis of the data we concluded that we need to take the first group of attributes (The attributes belongs to high class) should be given to the regression model.

7 Work Distributions

Each group member has contributed to every section in the report. That being said the primary contributions to the sections are as follows:

Intoduction: Alex
Dataset: Alex
Attributes: Ghassen
Visualization: James
Discussion: Alex, James, Ghassen
Conclusion: Alex, James, Ghassen

List of Figures

A Correlation

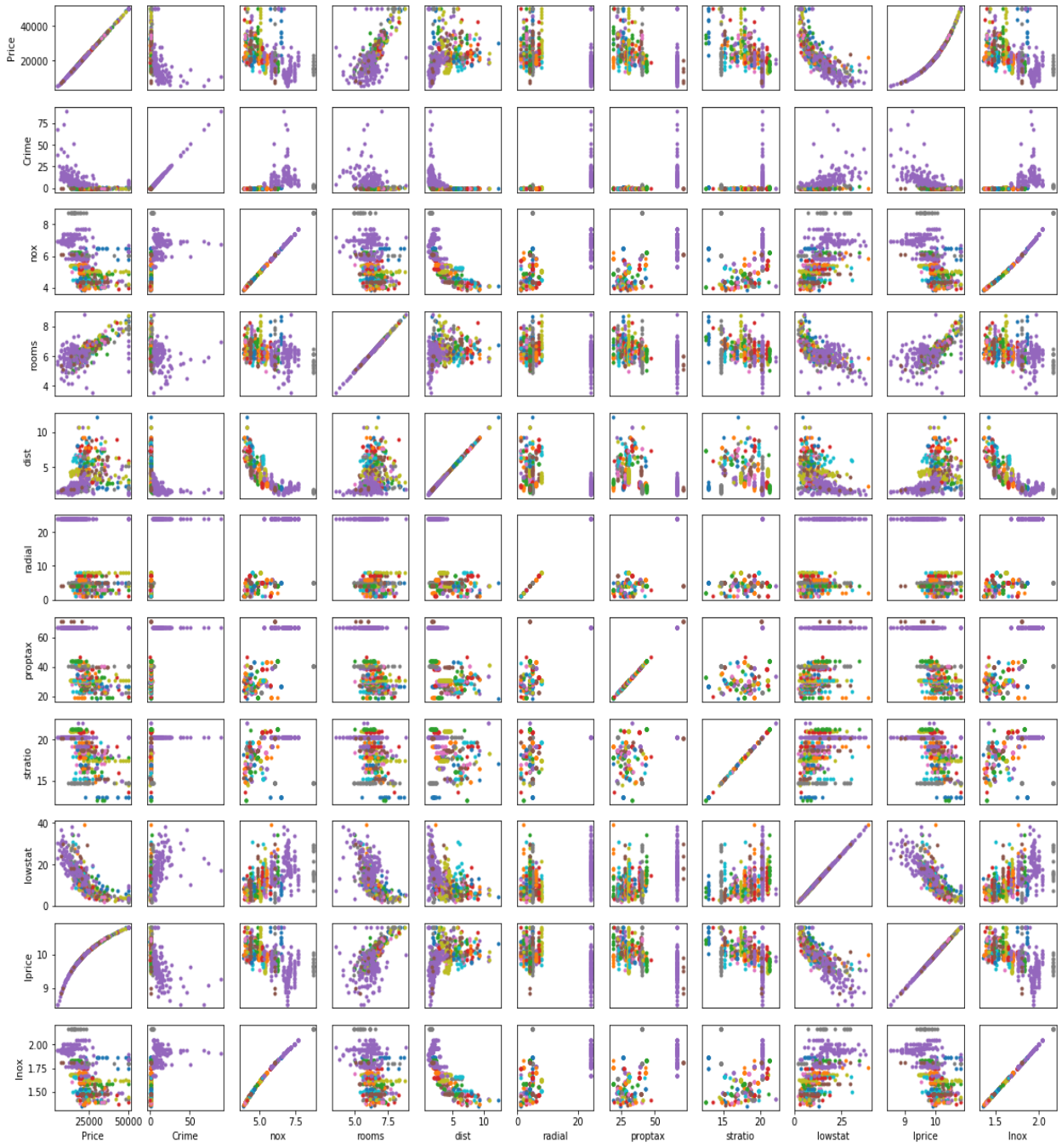

Figure A.1: Attribute correlations

B Attribute descriptions

TABLE IV

Variables used in the Housing Value Equations

Variable	Definition	Source
Dependent		
MV	Median value of owner-occupied homes.	1970 U. S. Census
Structural		
RM	Average number of rooms in owner units. *RM* represents spaciousness and, in a certain sense, quantity of housing. It should be positively related to housing value. The RM^2 form was found to provide a better fit than either the linear or logarithnic forms.	1970 U. S. Census
AGE	Proportion of owner units built prior to 1940. Unit age is generally related to structure quality.	1970 U. S. Census
Neighborhood		
B	Black proportion of population. At low to moderate levels of *B*, an increase in *B* should have a negative influence on housing value if Blacks are regarded as undesirable neighbors by Whites. However, market discrimination means that housing values are higher at very high levels of *B*. One expects, therefore, a parabolic relationship between proportion Black in a neighborhood and housing values.	1970 U. S. Census
LSTAT	Proportion of population that is lower status = ½ (proportion of adults without some high school education and proportion of male workers classified as laborers). The logarithmic specification implies that socioeconomic status distinctions mean more in the upper brackets of society than in the lower classes.	1970 U. S. Census
CRIM	Crime rate by town. Since *CRIM* gauges the threat to well-being that households perceive in various neighborhoods of the Boston metropolitan area (assuming that crime rates are generally proportional to people's perceptions of danger) it should have a negative effect on housing values.	FBI (1970)
ZN	Proportion of a town's residential land zoned for lots greater than 25,000 square feet. Since such zoning restricts construction of small lot houses, we expect *ZN* to be positively related to housing values. A positive coefficient may also arise because zoning proxies the exclusivity, social class, and outdoor amenities of a community.	Metropolitan Area Planning Commission (1972)
INDUS	Proportion nonretail business acres per town. *INDUS* serves as a proxy for the externalities associated with industry—noise, heavy traffic, and unpleasant visual effects, and thus should affect housing values negatively.	Vogt, Ivers, and Associates [33]
TAX	Full value property tax rate ($/$10,000). Measures the cost of public services in each community. Nominal tax rates were corrected by local assessment ratios to yield the full value tax rate for each town. Intra-town differences in the assessment ratio were difficult to obtain and thus not used. The coefficient of this variable should be negative.	Massachusetts Taxpayers Foundation (1970)

Continued

Figure B.1: Attribute descriptions ["Hedonic Housing Prices and the Demand for Clean Air," published in Journal of Environmental Economics and Management 5, 81-102 by by D. Harrison and D.L. Rubinfeld]

TABLE IV—*Continued*

Variable	Definition	Source
PTRATIO	Pupil–teacher ratio by town school district. Measures public sector benefits in each town. The relation of the pupil–teacher ratio to school quality is not entirely clear, although a low ratio should imply each student receives more individual attention. We expect the sign on *PTRATIO* to be negative.	Massachusetts Dept. of Education (1971–1972)
CHAS	Charles River dummy: =1 if tract bounds the Charles River; =0 if otherwise. *CHAS* captures the amenities of a riverside location and thus the coefficient should be positive.	1970 U. S. Census Tract maps
Accessibility		
DIS	Weighted distances to five employment centers in the Boston region. According to traditional theories of urban land rent gradients, housing values should be higher near employment centers. *DIS* is entered in logarithm form; the expected sign is negative.	Schnare [29]
RAD	Index of accessibility to radial highways. The highway access index was calculated on a town basis. Good road access variables are needed so that auto pollution variables do not capture the locational advantages of roadways. *RAD* captures other sorts of locational advantages besides nearness to workplace. It is entered in logarithmic form; the expected sign is positive.	MIT Boston Project
Air Pollution		
NOX	Nitrogen oxide concentrations in pphm (annual average concentration in parts per hundred million).	TASSIM
PART	Particulate concentrations in mg/hcm³ (annual average concentration in milligrams per hundred cubic meters).	TASSIM

TABLE V

Summary Statistics for Housing Value Equation Variables

Variable	Mean	SD
MV	22,532	9,197
RM	6.28	0.70
AGE	68.6	28.1
B	0.06	0.18
LSTAT	0.13	0.07
CRIM	3.61	8.60
ZN	11.36	23.32
INDUS	11.13	6.86
TAX	408.2	168.5
PTRATIO	18.5	2.16
DIS	3.79	2.10
RAD	9.55	8.70
NOX	5.55	1.16
PART	6.31	1.50

Figure B.2: Attribute descriptions continued ["Hedonic Housing Prices and the Demand for Clean Air," published in Journal of Environmental Economics and Management 5, 81-102 by by D. Harrison and D.L. Rubinfeld]