Assignment - Data Analysis

Francesco Muia - EMBA 20

Contents

1 Definition of the problem and dataset

The 'Boston House Prices Dataset,' available at a Kaggle repository[1], was gathered in 1978 and provides data for various Boston suburbs.[2] Our goal is to utilize a multivariate linear regression model to predict house prices based on a set of features representing Boston homes.

This dataset, devoid of NaN values, includes 506 rows, each representing a Boston suburb or town, and 15 columns that correspond to various dwelling features. Many of these features signify median values for the town or suburb. Specifically, the dataset columns comprise[3]

- *crim*: per capita crime rate by town;
- *zn*: proportion of residential land zoned for lots over 25000 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 (parts per 10 million);
- *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 $10000;
- *ptratio*: pupil-teacher ratio by town;

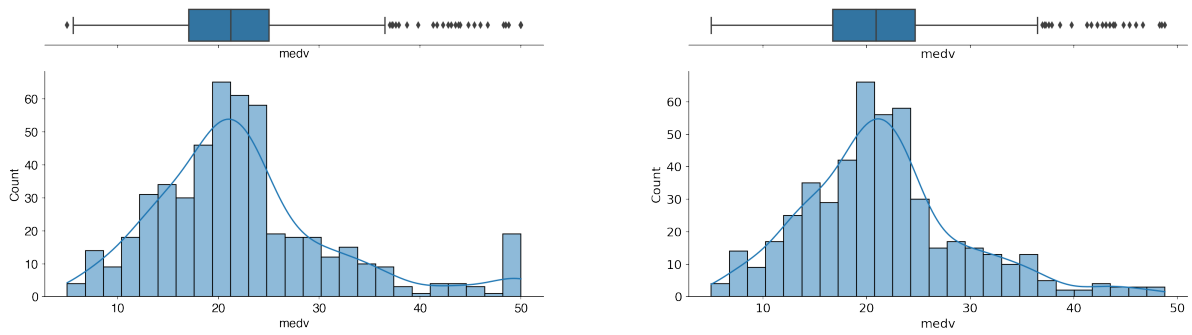
[1] Refer also to the original source and the first paper utilizing it for statistical analysis.
[2] Observe the intricacies and interesting facts shared at this link.
[3] The dataset also features a column named *Unnamed: 0*, which we have renamed to *idx* for comparison with JMP's indexing.

- *black*: $1000 \times (\text{Bk} - 0.63)^2$ where Bk is the proportion of black people by town;
- *lstat*: % lower status of the population;
- *medv*: median value of owner-occupied homes in $1000's.

The dependent variable, *medv*, is our target for prediction via multivariate linear regression. All other variables will serve as independent ones in our multivariate linear regression model (referred to simply as the *regression model* hereafter). Prior to the analysis, let us observe some general characteristics of the dataset:

- As seen in Fig. 1a, the dataset appears to be right-censored, i.e., all dwellings with *medv* exceeding 50 have been artificially designated a *medv* value of 50. Including these points in the analysis would skew the model parameter estimation, thus we eliminate them from the dataset beforehand. However, we must remember that our model may not extrapolate well to dwellings with *medv* values above 50. Upon removal of points artificially clustered in the last bin, the distribution assumes the form shown in Fig. 1b. This modified dataset, containing 490 rows, will be utilized in subsequent analyses.
- Another aspect worth noting is the *black* feature's behaviour as a function of the proportion of black people by town (Bk), as depicted in Fig. 2. The maximum value of the *black* feature corresponds to a minimal proportion of black residents, at $\text{Bk} \simeq 0$. For *black* values $\lesssim 137$, we cannot invert the function to find Bk from the value of *black*: for instance, $black \simeq 100$ corresponds to either $\text{Bk} \simeq 0.31$ or $\text{Bk} \simeq 0.96$.

To streamline the analysis, we transformed the *rad* variable, with unique values of $(1, 2, 3, 4, 5, 6, 7, 8, 24)$, into a categorical variable following the rule: $(1, 2, 3, 4) \leftrightarrow$ 'low', $(5, 6, 7, 8) \leftrightarrow$ 'medium', and $(24) \leftrightarrow$ 'high'. The other categorical variable, *chas*, only takes the values of 0 and 1.



(a) (b)

Figure 1: distribution of *medv* before and after the removal of the artificial points stacked on the last bin.

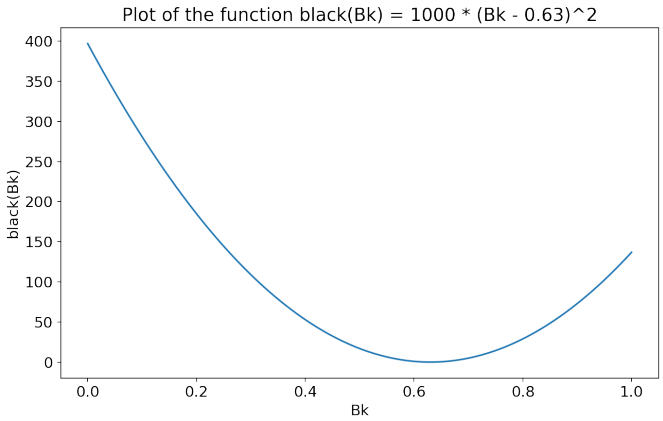
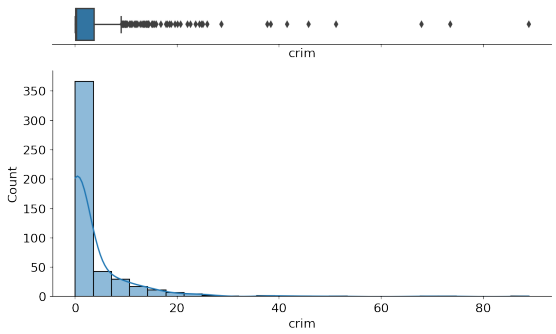
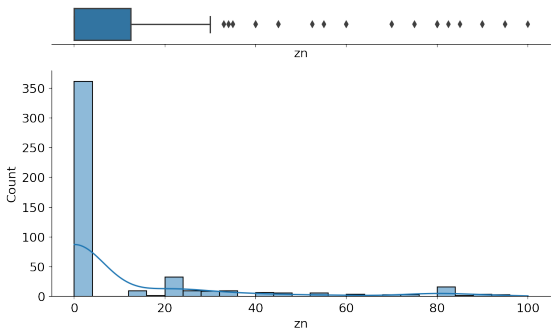


Figure 2: function describing the feature *black* as a function of the proportion of black people in town, Bk.

2 Exploratory analysis

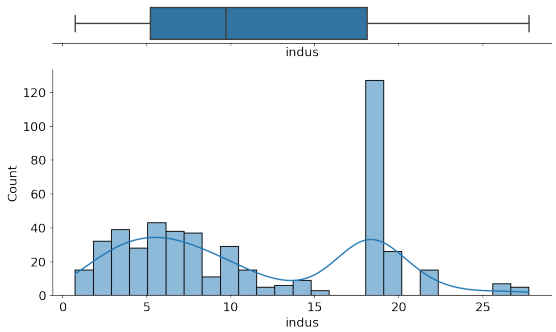


(a) Distribution of *crim*.

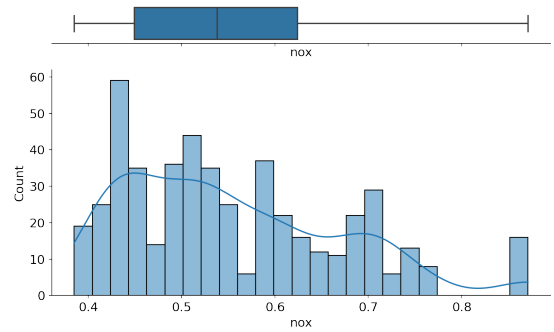


(b) Distribution of *zn*.

Figure 3

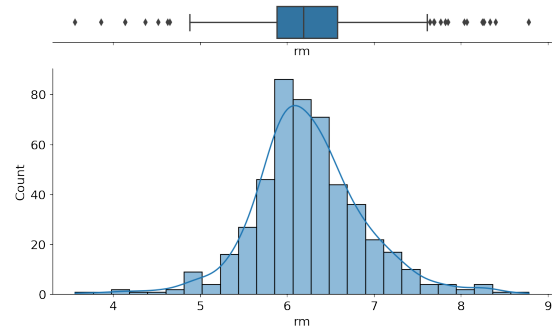


(a) Distribution of *indus*.

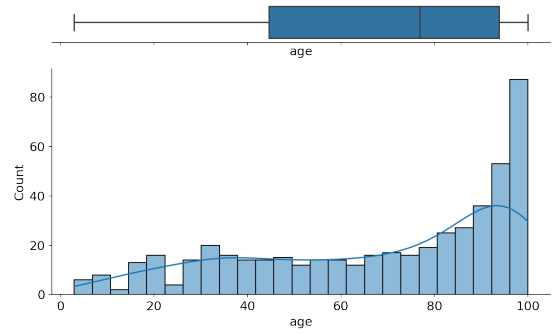


(b) Distribution of *nox*.

Figure 4



(a) Distribution of *rm*.

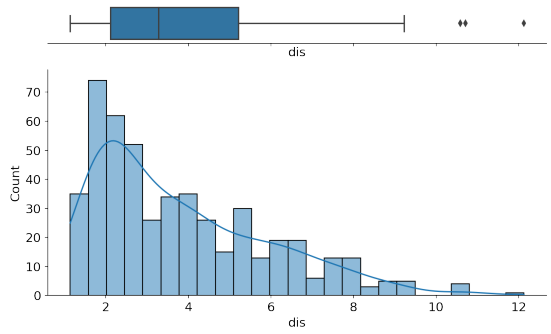


(b) Distribution of *age*.

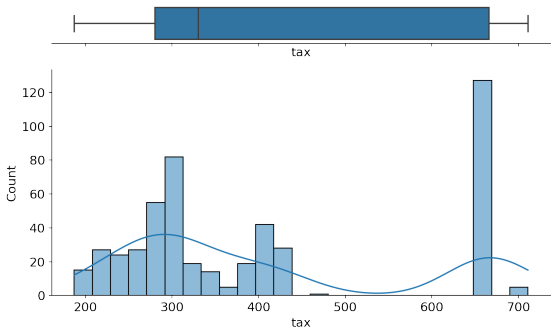
Figure 5

The univariate exploratory data analysis provides the following insights:

- *crim* (Fig. 3a): the distribution of per capita crime rate by town is highly skewed to the right, with most towns having a low crime rate. There are a few towns with exceptionally high crime rates. In particular the data points with value of *crim* larger than ~ 30 should be monitored as potential outliers;
- *zn* (Fig. 3b): a large fraction of the points (361 out of 490) has value of *zn* equal to 0, i.e. many suburbs

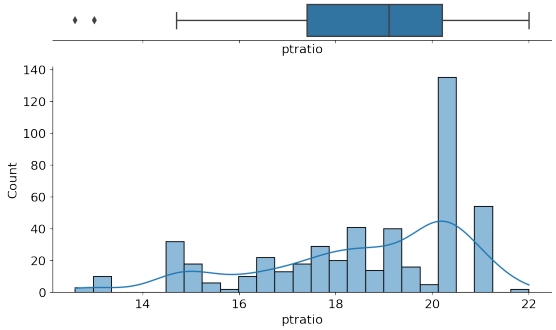


(a) Distribution of *dis*.

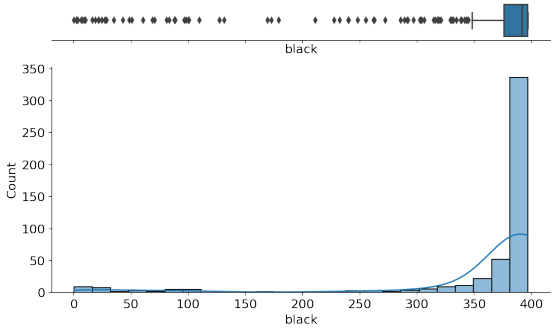


(b) Distribution of *tax*.

Figure 6

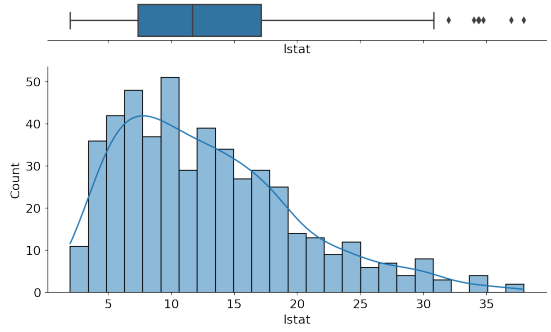


(a) Distribution of *ptratio*.

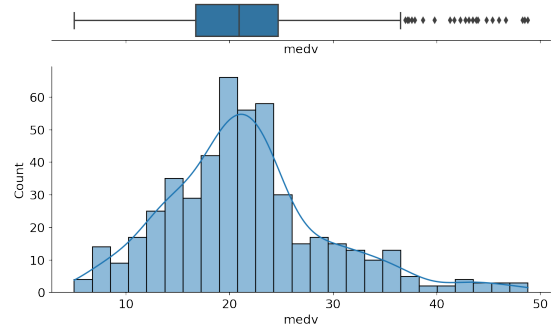


(b) Distribution of *black*.

Figure 7



(a) Distribution of *lstat*.

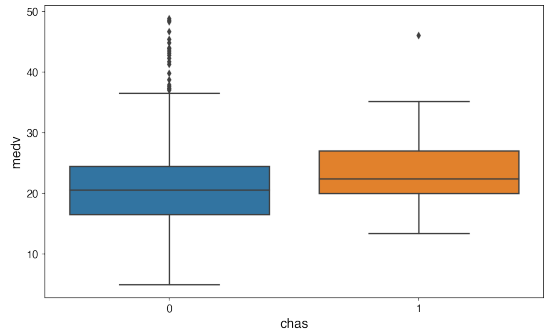


(b) Distribution of *medv*.

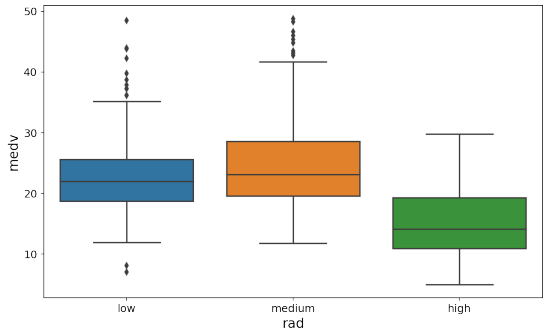
Figure 8

have no residential land zoned for lots over 25000 sq ft. We will see that these zones play an important role in the analysis. Data points with value of *zn* larger than ~ 50 should be monitored as potential outliers;

- *indus* (Fig. 4a): the proportion of non-retail business acres per town appears to be bimodal, with peaks around the low and high ends of the distribution. There is a particularly high bin around *indus* $\simeq 18$, corresponding to 127 points. We will see that these play an interesting role in the analysis;
- *nox* (Fig. 4b): the distribution of nitric oxide concentrations seems to be slightly skewed to the right. There is a gap between the largest value of *nox* (the last bin contains 16 points) and the bulk of the points which deserves further investigation: these points should be monitored as potential outliers;
- *rm* (Fig. 5a): the average number of rooms per dwelling follows an approximately normal distribution,



(a) Distribution of *chas*.



(b) Distribution of *rad*.

Figure 9

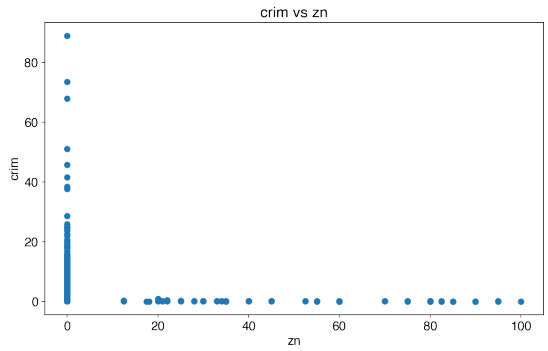
with a few potential outliers on both sides of the distribution;

- *age* (Fig. 5b): the proportion of owner-occupied units built prior to 1940 is skewed to the left. This indicates a higher number of older houses: there are 229 points with $age > 80$, corresponding to $\sim 47\%$ of the dataset;
- *dis* (Fig. 6a): the distribution of weighted distances to five Boston employment centers is right-skewed, indicating that most towns are relatively close to these centers. A few points (5 in total) at $dis > 10$ should be monitored as potential outliers;
- *tax* (Fig. 6b): the full-value property tax rate per \$10000 appears to be bimodal, with peaks around the lower end and the higher end. There is a big gap between the lower and higher ends of the distribution and a very large bin (corresponding to 127 points) close to the highest end. This should already ring a bell: as we saw before there is a large bin containing 127 points in the distribution of *indus*. We will investigate this further below;
- *ptratio* (Fig. 7a): the distribution of pupil-teacher ratios by town seems to be slightly skewed to the right. There is a very large bin at $ptratio \simeq 20$ and a few points at $ptratio < 14$ that should be monitored as potential outliers;
- *black* (Fig. 7b): the distribution of the *black* variable is heavily left-skewed. As most of the points lie at $black \simeq 400$, it means that most of the towns have a very low population of black people, see Fig. 2. We should monitor points with $black \lesssim 250$ as potential outliers;
- *lstat* (Fig. 8a): this variable is right-skewed, indicating that there are more towns with a low percentage of the population in lower status;
- *medv* (Fig. 8b): the median value of owner-occupied homes is slightly skewed to the right, indicating that most homes are in the lower to mid-range price bracket.

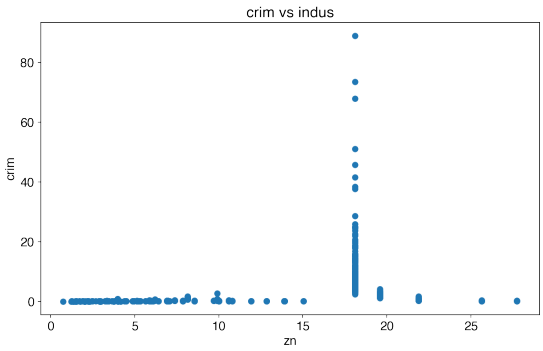
To preserve the business-oriented character of the assignment, we will perform the analysis without transforming the variables (e.g. using a log transformation), in order to prioritize the interpretability of the model over the rest. We will further comment on this in the following sections. Concerning the categorical variables:

- *chas* (Fig. 9a): there are many more points with $chas = 0$ (461) rather than with $chas = 1$ (29). This implies that most of the towns are not touched by the river. Note that there is one potential outlier for $chas = 1$, corresponding to $idx = 283$;
- *rad* (Fig. 9b): the number of data corresponding to rad = (low, medium, high) is $(188, 175, 127)$. Note that the median of *medv* is similar for *rad* equal to low and medium, while it is lower for *rad* equal to high. It seems that a higher accessibility to radial highways has a negative impact on dwellings' values. Likely, these correspond to more peripheral towns, where the industrialization is high as well as the criminality rate. Curiously, the number of points with *rad* = 'high' is 127, the same number of points contained in the anomalously large bins in the distributions of *indus* and *tax*.

2.1 Clustering

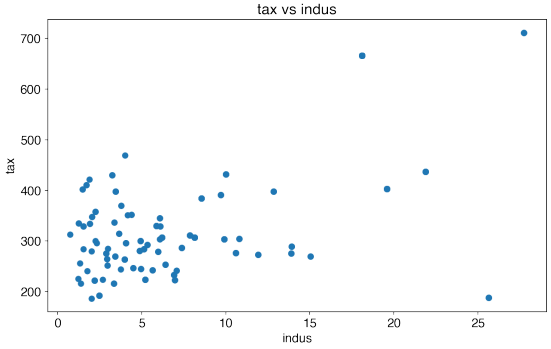


(a) Bivariate distribution in the plane *zn* vs *crim*.

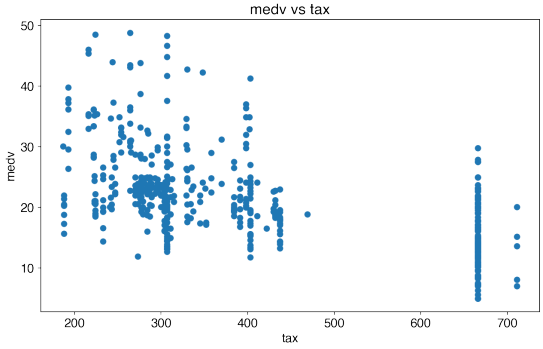


(b) Bivariate distribution in the plane *indus* vs *crim*.

Figure 10

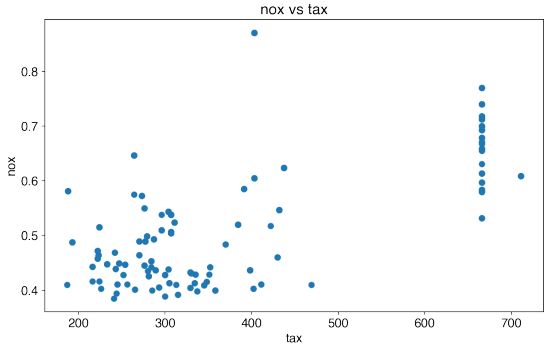


(a) Bivariate distribution in the plane *indus* vs *tax*.

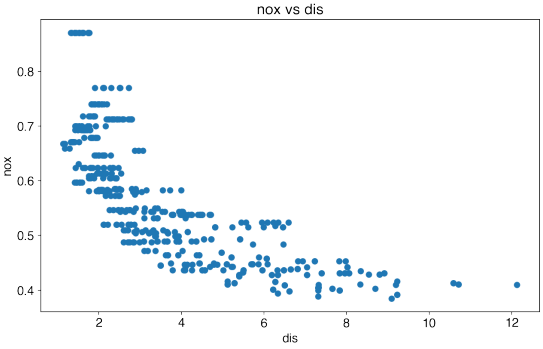


(b) Bivariate distribution in the plane *tax* versus *medv*.

Figure 11



(a) Bivariate distribution in the plane *tax* versus *nox*.

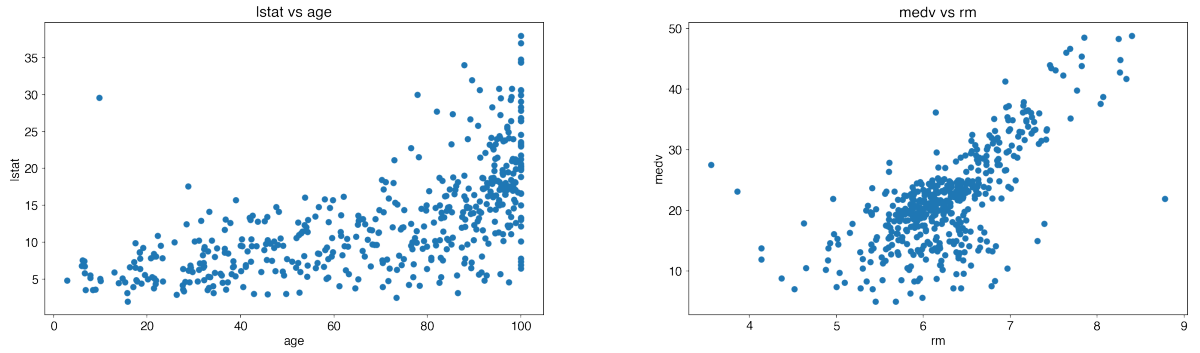


(b) Bivariate distribution in the plane *dis* versus *nox*.

Figure 12

In this subsection, we will propose a splitting of the dataset - based on the intuition gained in the analysis of Sec. 2 - into two clusters which contain more homogeneous populations. In particular:

- from the *crim* vs *zn* plot in Fig. 10a we infer that *crim* can only be non-zero for $zn = 0$. These points presumably correspond to suburbs where there are no residential lands zoned for lots over 25000 sq. ft., which can typically be associated with business parks, strip shopping centers, shopping malls;



(a) Bivariate distribution in the plane *age* versus *lstat*. (b) Bivariate distribution in the plane *rm* versus *medv*.

Figure 13

- from the *crim* vs *indus* plot in Fig. 10b we infer that high values of *crim* correspond to a single value of $indus = 18.1$. Along with the information from Fig. 10a, we infer that high a criminality rate (*crim*) corresponds to towns with no residential lands zoned for lost over 25000 sq. ft., which in turn correspond to a single value of $indus = 18.1$. These towns correspond to the 127 points forming the anomalous bins in the distributions of *indus* and *tax*, as seen in Sec. 2. Note that the data points at $indus \gtrsim 25$ are potential outliers, see Fig. 10b.

- from Fig. 11a we can observe that all the 127 houses with $indus = 18.1$ correspond to a single tax rate ($tax = 666$), suggesting that these towns belong to a region that enforces a single fiscal policy. This is a further hint toward the existence of two clusters, one of which is represented by the towns in this high criminality rate region. Note again that the data points at $indus \gtrsim 25$ are potential outliers.

- from Fig. 11b we recognize the 127 points corresponding to $indus = 18.1$ as the vertical series of points located at $tax = 666$. Furthermore, there are 5 points at $tax = 711$ that are potential outliers even after the clustering. These points correspond to the rightmost point in the *tax* vs *indus* plot in Fig. 11a;

- the *nox* vs *tax* plot in Fig. 12a, beyond confirming the previous considerations, highlights other 16 potential outliers, corresponding to the single point located at $nox = 0.87$. As they correspond to the same value of *tax*, these towns likely belong to the same region. The same points can also be seen at the top of the *nox* vs *dis* plot in Fig. 12b, where we can also note three points at $dis > 10$ that should be monitored as potential outliers;

- the *nox* vs *tax* plot in Fig. 13a features an isolated point that represents a possible outlier ($idx = 215$). At the same time, the *medv* vs *rm* plot in Fig. 13b shows a seemingly linear relation between these two variables, with potential outliers at $rm < 4$ and another isolated point at $rm \simeq 9$.

Based on the insights obtained from these bivariate plots, we are now prepared to segment the dataset into two distinct clusters using a specific rule:

$$tax = 666 \quad \leftrightarrow \quad \text{cluster } 1\,, \qquad\qquad tax \neq 666 \quad \leftrightarrow \quad \text{cluster } 2\,.$$

The rationale for this division is that each cluster aligns with different regions, such as towns or suburbs, possessing more uniform characteristics that influence the median home values. An alternate option could have been to incorporate the points with a tax value of 711 into cluster 1 (see Fig. 11b). However, we aim to utilize the relationship between points with a *tax* value of 666, which corresponds to an *indus* value of 18.1 and high crime rates, as depicted in Fig. 10a. Such a relation would not hold true for points with $tax = 711$, that correspond to $indus > 25$ (see Fig. 11a) for which $crim \simeq 0$ (see Fig. 10b).

Cluster 1

Let us examine the multivariate distribution for the points in cluster 1. First, it is important to note that not all variables remain necessary: beyond *tax* which is fixed at 666 for cluster 1, the other variables that are fixed

are $zn = 0$, $indus = 18.1$ and $ptratio = 20.2$. Also, there is no town or suburb with $rad \neq$ 'high' in cluster 1. Consequently, we have targeted towns or suburbs with substantial industrialization, high accessibility to radial highways, and no residential land zoned for lots over 25000 sq. ft., which we interpret as an absence of shopping centers. We hypothesize that cluster 1 comprises the most peripheral towns and suburbs included in the dataset.

After eliminating the redundant variables, the dataset structure becomes considerably simpler, enabling us to generate the scatter plot[4] in Fig. 20, from which we can deduce that:

- there are a few potential outliers identified with colored circles;
- there appears to be a positive linear relationship between *medv* and *dis*, as well as between *medv* and *rm*. This suggests that house values increase as one gets closer to residential areas, which are distanced from employment centers, and also with an increase in the median number of rooms. Additionally, there is a negative linear relationship between *medv* and *lstat*, which intuitively makes sense: as the percentage of lower status population increases, house values tend to decrease.

	idx	crim	chas	nox	rm	age	dis	rad	black	lstat	medv	Mahalanobis
365	366	4.56	0	0.72	3.56	87.90	1.61	high	354.70	7.12	27.50	5.55
380	381	88.98	0	0.67	6.97	91.90	1.42	high	396.90	17.21	10.40	6.67
418	419	73.53	0	0.68	5.96	100.00	1.80	high	16.45	20.62	8.80	5.30

Table 1: potential outliers according to the Mahalanobis distance method for cluster 1.

To pinpoint the outliers, we will initially identify possible outliers from the bivariate distributions, as detailed in Sec. 2 and Fig. 20. These will then be evaluated using the Mahalanobis distance method. We will pay special attention to the points marked as potential outliers by both methods, and only remove those from the dataset for which we can provide a business rationale for their anomalous nature.

We will proceed to identify potential outliers using the Mahalanobis distance method[5], setting a threshold at $\alpha = 1\%$. We uncover three potential outliers (2.36% of the data in cluster 1), as shown in Tab. 1. All these points were already included in the list of outliers identified in Sec. 2 and in the scatterplot in Fig. 20. In particular:

- The first point, labeled with $idx = 366$, matches the extreme point encircled in green in Fig. 20 as well as in Fig. 13b. This point does not appear to be anomalous, as it represents a town or suburb with relatively high property prices, situated in a residential area consisting mainly of older houses. The area exhibits a lower black population percentage, a small proportion of the population in the lower status, and a below-average number of rooms per dwelling. It is plausible that the area consists of many recently refurbished apartments. Consequently, we will not classify this point as anomalous.
- The final two points in Tab. 1 correspond to the two highest points in Fig. 10a and the two most extreme points encircled in red in Fig. 20. These points appear to be anomalous when considering the *crim* vs *lstat* plot shown in Fig. 14a. While there is a clear correlation between the percentage of the lower-status population and the crime rate, these two points are distinctly isolated, suggesting they may be considered anomalies. Further investigation is needed to understand the social/envirnomental/regulatory conditions causing this effect, but for the purposes of this assignment, these two points, labeled with $idx = (381, 419)$, will be identified as outliers and subsequently removed from the cluster 1 dataset in all following analyses.

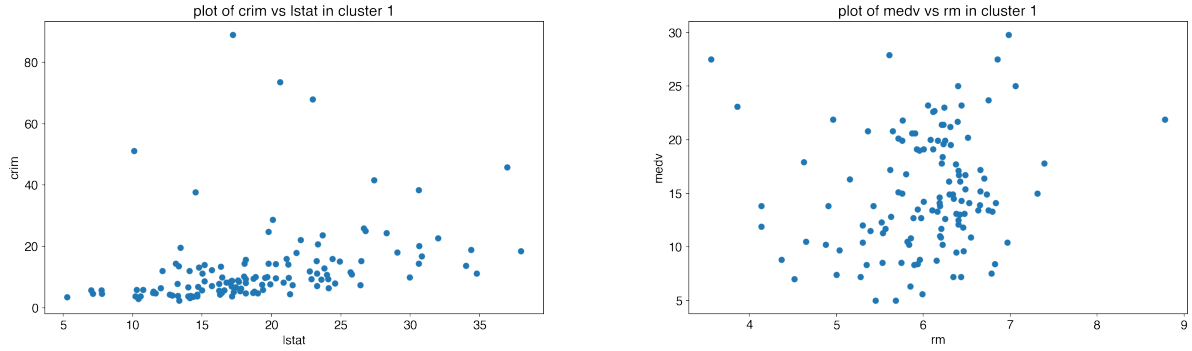
Some remarks regarding the other potential outliers identified in Sec. 2 are as follows:

- the potential outliers corresponding to large values of *nox*, *tax*, *dis* do not fall within cluster 1;
- the solitary point for $chas = 1$ in Fig. 9a is not part of cluster 1;

[4] We report the scatter plots in an Appendix for ease of visualization.

[5] Please note that the threshold we have identified differs slightly from the one proposed by JMP. Unfortunately, without access to the JMP source code, I cannot make a direct comparison regarding their computation methods. However, this does not impact our analysis. Our approach begins with the points having the largest Mahalanobis distances and among them, identifies a minimal number of outliers.

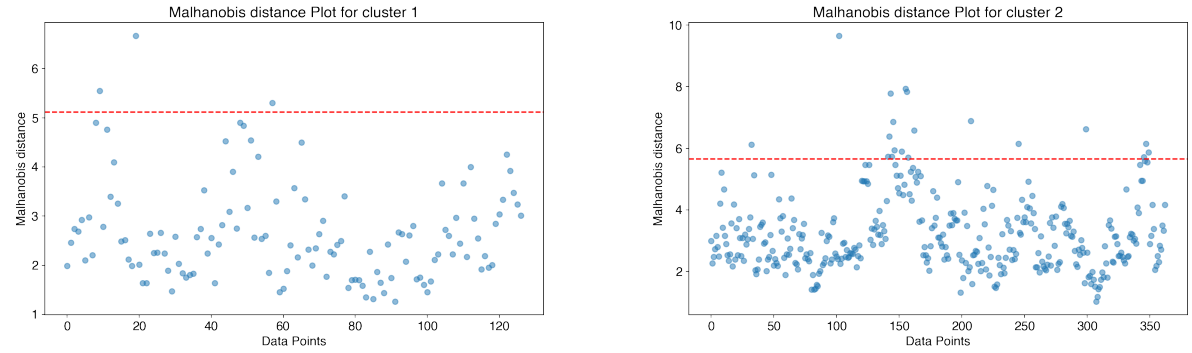
- we opt to retain the potential outlier corresponding to the isolated point at $rm \simeq 9$ in Fig. 14b (the point encircled in purple in Fig. 20) as we cannot justify its removal with a sound business argument.[6]



(a) Bivariate distribution in the plane *lstat* versus *crim*. (b) Bivariate distribution in the plane *rm* versus *medv*.

Figure 14

After eliminating the two outliers discussed above, we can also compute the correlation matrix, and visualize it with a heatmap in Fig. 16. This features a positive correlation of 0.55 between *medv* and *dis* and a stronger negative correlation (-0.7) betwen *medv* and *lstat*. Also, we identify a relatively strong multicollinearity (-0.72) between *age* and *dis*.



(a) Mahalanobis distance for cluster 1. (b) Mahalanobis distance for cluster 2.

Figure 15

Cluster 2

In this section we will follow the same strategy of Sec. 2.1 to analyse cluster 2. First, note that there is no house with *rad* = 'high' in cluster 2, implying that probably this cluster includes less peripheral towns and suburbs. From the scatterplot in Fig. 21 we can infer that

- a few potential outliers have been identified, noticeable as isolated points. (we do not highlight them with colored circles in this case for ease of visualization);
- there appears to be a positive linear relationship between *medv* and *rm*. This suggests that as the average number of rooms increases (likely indicating larger houses), the median value of the houses also increases. On the other hand, there is a negative relationship between *medv* and *lstat*. However, this relationship appears to be inverse (resembling a $1/x$ form) rather than linear.

Using the Mahalanobis distance method with a threshold of $\alpha = 1\%$ we identify 19 potential outliers (i.e. 5.23% of the points in cluster 2), see Tab. 2. More in detail:

[6] One might argue that the corresponding town features an above-average number of B&Bs, but this could also be true for data points with $rm \simeq 6-7$. Hence, we prefer to refrain from making too many additional assumptions and keep the point.

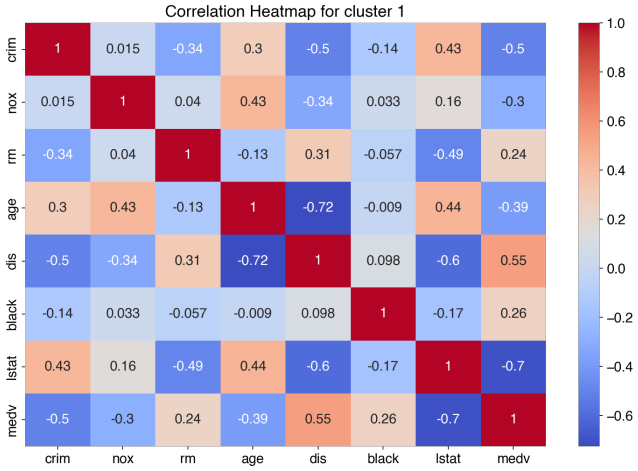


Figure 16: correlations in cluster 1.

idx	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv	Mahal.
33	1.39	0.00	8.14	0	0.54	5.95	82.00	3.99	l	307	21.00	232.60	27.71	13.20	6.12
103	0.23	0.00	8.56	0	0.52	6.41	85.40	2.71	m	384	20.90	70.80	10.63	18.60	9.66
142	1.63	0.00	21.89	0	0.62	5.02	100.00	1.44	l	437	21.20	396.90	34.41	14.40	5.73
143	3.32	0.00	19.58	1	0.87	5.40	100.00	1.32	m	403	14.70	396.90	26.82	13.40	6.39
144	4.10	0.00	19.58	0	0.87	5.47	100.00	1.41	m	403	14.70	396.90	26.42	15.60	7.78
145	2.78	0.00	19.58	0	0.87	4.90	97.80	1.35	m	403	14.70	396.90	29.29	11.80	5.73
146	2.38	0.00	19.58	0	0.87	6.13	100.00	1.42	m	403	14.70	172.91	27.80	13.80	6.86
147	2.16	0.00	19.58	0	0.87	5.63	100.00	1.52	m	403	14.70	169.27	16.65	15.60	5.94
153	1.13	0.00	19.58	1	0.87	5.01	88.00	1.61	m	403	14.70	343.28	12.12	15.30	5.90
156	3.54	0.00	19.58	1	0.87	6.15	82.60	1.75	m	403	14.70	88.01	15.02	15.60	7.94
157	2.45	0.00	19.58	0	0.87	5.27	94.00	1.74	m	403	14.70	88.63	16.14	13.10	7.84
158	1.22	0.00	19.58	0	0.60	6.94	97.40	1.88	m	403	14.70	363.43	4.59	41.30	5.70
166	2.92	0.00	19.58	0	0.60	6.10	93.00	2.28	m	403	14.70	240.16	9.81	25.00	6.58
215	0.29	0.00	10.59	0	0.49	5.41	9.80	3.59	l	277	18.60	348.93	29.55	23.70	6.88
254	0.37	22.00	5.86	0	0.43	8.26	8.40	8.91	m	330	19.10	396.90	3.54	42.80	6.15
311	2.64	0.00	9.90	0	0.54	4.97	37.80	2.52	l	304	18.40	350.45	12.64	16.10	6.62
489	0.15	0.00	27.74	0	0.61	5.45	92.70	1.82	l	711	20.10	395.09	18.06	15.20	5.72
491	0.21	0.00	27.74	0	0.61	5.09	98.00	1.82	l	711	20.10	318.43	29.68	8.10	6.15
493	0.11	0.00	27.74	0	0.61	5.98	83.50	2.11	l	711	20.10	396.90	13.35	20.10	5.86

Table 2: potential outliers according to the Mahalanobis eistance method for cluster 2.

- 8 data points correspond to instances where $nox = 0.87$, see Fig. 12. We have not found a compelling business justification to exclude these points. Consequently, we will retain them in the dataset designated for cluster 2;
- the pont $idx = 103$ is characterized by a value of $black = 70.80$, which is the smallest in cluster 2. Since it is also the point with the largest Mahalanobis distance, for the purpose of this assignment, we will assume that there exists an underlying social/environmental/regulatory cause making this point anomalous and qualifying it as an outlier.[7]
- the point $idx = 254$, with $zn = 22$ and characterized by a very high $rm = 8.91$. We will keep this point in the analysis and the rationale for this decision aligns with the one used for retaining data in cluster 1;
- 6 points with $indus > 19$, which we will keep in the dataset as we do not have a good business argument to discard them;
- the point with $idx = 215$ which corresponds to the isolated point in Fig. 13a. For the purpose of this assignment we will consider this point an outlier, as it is characterized by a very anomalous relation

[7]Note that $black = 70.80$ corresponds to a proportion of black people equal to either 36% or 90%, due to the quadratic nature of the function in Fig. 2.

between *lstat* and *age*, implying that it corresponds to a quite recently built town or suburbs where the percentage of population in the lower status is extremely high;

- two more points with $idx = (33, 311)$ which do not seem anomalous and will be kept in the dataset;
- we analysed all the other points that we identified as potential outliers both in the univariate and in the bivariate distributions and we did not find any sound business justification to remove them.

From the correlation heatmap in Fig. 17 that there is a strong correlation between *medv* and *rm*, where this last feature is probably related to the size of the house.

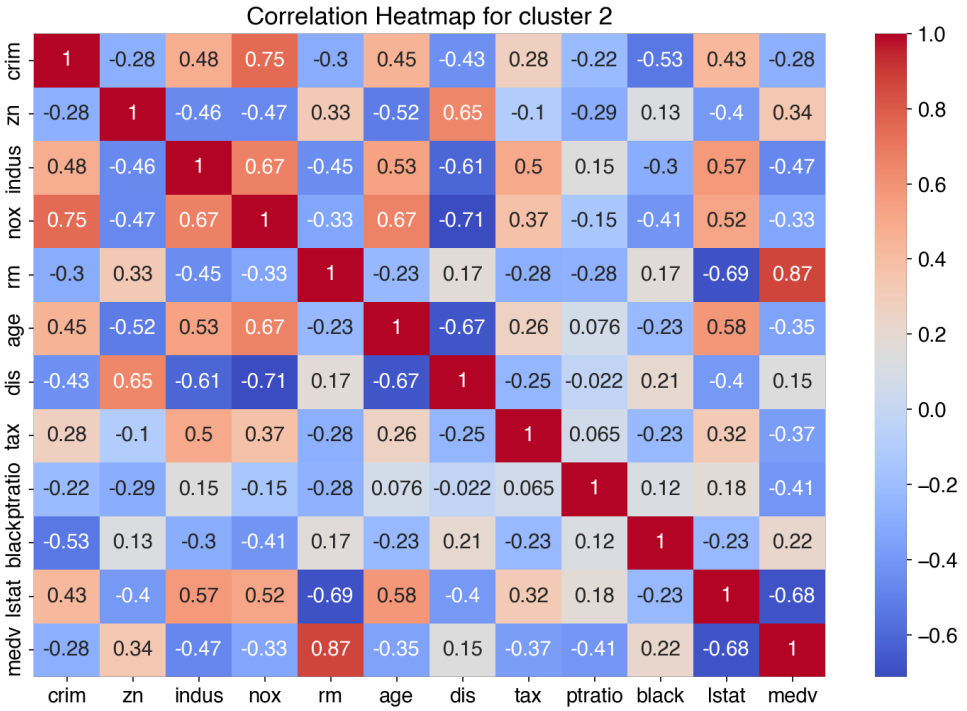


Figure 17: correlations in cluster 2.

3 Regression models

The regression models in this section are evaluated using the python *statsmodel* library.[8] All the results have been double-checked in JMP, whose screenshots are reported in Sec. B.

3.1 Cluster 1

3.1.1 Model 1.1

The base model for cluster 1 includes all the meaningful variables: *crim*, *nox*, *rm*, *age*, *dis*, *black*, *lstat*. The categorical variable *chas* has been transformed into a pair of variables $chas_0$ and $chas_1$, of which we keep only the second in the model to avoid the dummy variable trap.

The results of the regression are reported in Tab. 3. A few observations are in order:

- R-squared and Adj. R-squared are 0.64 and 0.61 respectively, meaning that more than 60% of the variance is explained by the model.
- The F-statistic equal to 26.07 tells us that definitely not all the coefficients of the model vanish.

[8] The Jupyter notebook used for the analysis is provided in the submission folder.

- Setting a p-value threshold at 5%, the variables *age*, *dis*, *black*, $chas_1$ are not significant for the model. We will build the second model for cluster 1 removing the non-significant variables.

R-squared	0.643
Adj. R-squared	0.618
F-statistic	26.07
Prob (F-statistic)	1.18e-22
Log-Likelihood	-323.52
No. Observations	125
Df Residuals	116
Df model	8

	coef	std err	t	P> |t|	[0.025	0.975]
const	43.0029	6.193	6.943	0.000	30.736	55.270
crim	-0.1420	0.036	-3.924	0.000	-0.214	-0.070
nox	-19.7833	5.705	-3.468	0.001	-31.082	-8.484
rm	-1.3890	0.483	-2.874	0.005	-2.346	-0.432
age	0.0298	0.036	0.828	0.409	-0.041	0.101
dis	0.5230	0.855	0.612	0.542	-1.170	2.216
black	0.0036	0.002	1.698	0.092	-0.001	0.008
lstat	-0.4918	0.062	-7.974	0.000	-0.614	-0.370
chas_1	3.1421	1.668	1.884	0.062	-0.162	6.446

Table 3: regression output for Model 1.1.

3.1.2 Model 1.2

Statistic	**Value**
R-squared	0.618
Adj. R-squared	0.606
F-statistic	48.60
Prob (F-statistic)	3.03e-24
Log-Likelihood	-327.63
No. Observations	125
Df Residuals	120
Df Model	4

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
const	45.6698	4.480	10.194	0.000	36.800	54.540
crim	-0.1523	0.034	-4.461	0.000	-0.220	-0.085
nox	-15.1777	4.981	-3.047	0.003	-25.040	-5.316
rm	-1.3916	0.472	-2.946	0.004	-2.327	-0.456
lstat	-0.5285	0.055	-9.539	0.000	-0.638	-0.419

Table 4: regression output for Model 1.2.

We remove the non-significant variables *age*, *dis*, *black*, $chas_1$ and evaluate again the regression, obtaining the output in Tab. 4. The fact that *lstat* belongs to the final model is consistent with what we observed in the correlation heatmap in Fig. 16. R-squared (and its adjusted version) is $\simeq 0.62$, meaning that more than 60% of the variance is explained by the model. An F-statistic of 48.6 suggests that there is a statistically significant relationship between at least one of the predictors and the dependent variable. The model is significantly better than a model with no predictors, and reads

$$y = \text{const} + \beta_{\text{crim}} \times \text{crim} + \beta_{\text{nox}} \times \text{nox} + \beta_{\text{rm}} \times \text{rm} + \beta_{\text{lstat}} \times \text{lstat}\,, \tag{1}$$

where the β's are the coefficients in the 'coef' column of Tab. 4. As all the variables are now significant, we can perform a VIF analysis to check for multicollinearity and compute the standard coefficients, using that

$$\beta_{\text{standard,i}} = \beta_{\text{non-standard,i}} \times \frac{\sigma_{X,i}}{\sigma_y} \qquad \forall i\,, \tag{2}$$

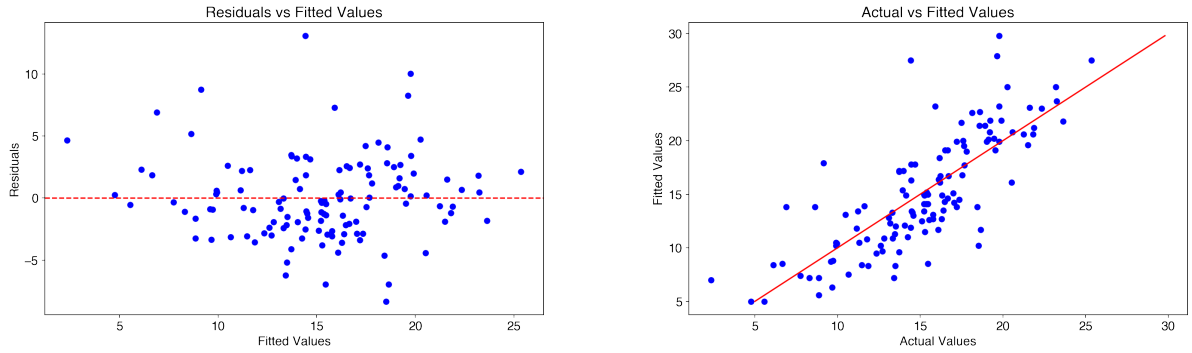
where σ_X is the standard deviation of the regressor variables, while σ_y is the standard deviation of *medv*. Using that $\sigma_y \simeq 5.41$, we obtain

Variable	$\beta_{\text{non-standard}}$	σ_X	β_{standard}	VIF
crim	-0.1523	10.04	-0.28	1.26
nox	-15.1777	0.06	-0.17	1.03
rm	-1.3916	0.72	-0.19	1.24
lstat	-0.5285	6.56	-0.64	1.42

Table 5

from which we infer that the most important driver for the house value is the variable *lstat*, followed by *crim*. Furthermore, as VIF < 5 for all the variables, the model apparently does not feature multicollinearity problems.

We can observe the output of the model visually as depicted in Fig. 18. Despite the residuals are reasonably distributed above and below 0, it is clear that the potential outliers, which we opted to include in our analysis, are contributing to the deviation of residuals from normality. We will analyse the limitations of the model in more detail in Sec. 3.3.



(a) Residuals of the regression model. (b) Actual versus fitted values for the regression model.

Figure 18: Regression plots for Model 1.2.

3.2 Cluster 2

3.2.1 Model 2.1

We proceed here as in Sec. 3.1. The first model for cluster 2 includes all the variables. Before performing the analysis, we obtain dummy variables for the categorical features *chas* and *rad*. Since *rad* = 'high' does not appear in cluster 2, and to avoid the dummy variable trap, eventually we will only have rad_{medium} in the model. The results of the regression are reported in Tab. 6.

R-squared	0.848
Adj. R-squared	0.843
F-statistic	149.3
Prob (F-statistic)	2.37e-133
Log-Likelihood	-888.73
No. Observations	361
Df Residuals	347
Df model	13

	coef	std err	t	P> |t|	[0.025	0.975]
const	-11.0186	4.600	-2.395	0.017	-20.066	-1.971
crim	0.5199	0.422	1.232	0.219	-0.310	1.350
zn	0.0167	0.009	1.855	0.064	-0.001	0.034
indus	-0.0194	0.041	-0.477	0.633	-0.100	0.061
nox	-5.3774	3.401	-1.581	0.115	-12.067	1.312
rm	8.3230	0.389	21.414	0.000	7.559	9.087
age	-0.0504	0.009	-5.452	0.000	-0.069	-0.032
dis	-0.8658	0.134	-6.482	0.000	-1.128	-0.603
tax	-0.0101	0.002	-4.391	0.000	-0.015	-0.006
ptratio	-0.5598	0.090	-6.223	0.000	-0.737	-0.383
black	0.0143	0.005	2.736	0.007	0.004	0.025
lstat	-0.0723	0.047	-1.532	0.126	-0.165	0.021
$chas_1$	0.6549	0.629	1.041	0.299	-0.582	1.892
rad_{medium}	0.4794	0.354	1.355	0.176	-0.216	1.175

Table 6: regression output for Model 2.1.

A few observations are in order:

- R-squared and Adj. R-squared are 0.85 and 0.84 respectively, meaning that more than 80% of the variance is explained by the model.
- The F-statistic equal to 149.3 tells us that definitely not all the coefficients of the model vanish.
- Setting a threshold at 5%, the variables *crim*, *zn*, *indus*, *nox*, *lstat*, $chas_1$ and $chas_{medium}$ are not significant for the model. We will build the second model for cluster 2 removing the non-significant variables.

R-squared	0.843
Adj. R-squared	0.841
F-statistic	317.7
Prob (F-statistic)	3.56e-139
Log-Likelihood	-894.60
No. Observations	361
Df Residuals	354
Df model	6

	coef	std err	t	P> |t|	[0.025	0.975]
const	-16.6358	3.060	-5.436	0.000	-22.654	-10.617
rm	8.9476	0.281	31.894	0.000	8.396	9.499
age	-0.0620	0.008	-8.178	0.000	-0.077	-0.047
dis	-0.6674	0.099	-6.708	0.000	-0.863	-0.472
tax	-0.0105	0.002	-5.155	0.000	-0.015	-0.006
ptratio	-0.6379	0.076	-8.381	0.000	-0.788	-0.488
black	0.0144	0.005	3.164	0.002	0.005	0.023

Table 7: regression output for Model 2.2.

3.2.2 Model 2.2

We remove the non-significant variables *crim*, *zn*, *indus*, *nox*, *lstat*, $chas_1$ and rad_{medium} and evaluate again the regression, obtaining the output in Tab. 7. R-squared (and its adjusted version) is $\simeq 0.84$, meaning that more than 84% of the variance is explained by the model. As the F-statistic is 317.7, we can be sure that not all the coefficients of the model vanish, hence the final model, namely

$$y = \text{const} + \beta_{\text{rm}} \times \text{rm} + \beta_{\text{age}} \times \text{age} + \beta_{\text{dis}} \times \text{dis} + \beta_{\text{tax}} \times \text{tax} + \beta_{\text{ptratio}} \times \text{ptratio} + \beta_{\text{tax}} \times \text{tax}\,, \tag{3}$$

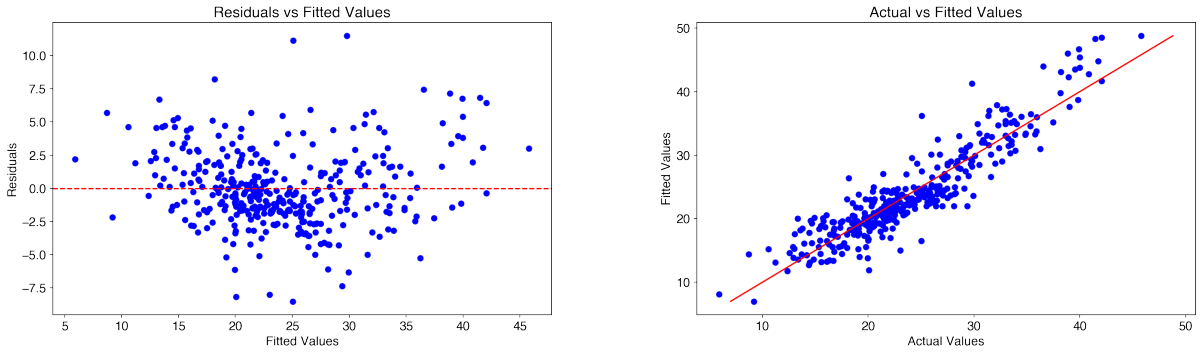
is statistically significant. As all the variables are now significant, we can perform a VIF analysis to check for multicollinearity and compute the standard coefficients, obtaining (for $\sigma_y \simeq 5.41$), we obtain

Variable	$\beta_{\text{non-standard}}$	σ_X	β_{standard}	VIF
rm	8.95	0.61	0.75	1.23
age	-0.06	-0.24	-0.17	1.91
dis	-0.67	2.11	-0.19	1.84
tax	-0.01	81.57	-0.12	1.16
ptratio	-0.64	2.16	-0.19	1.12
black	0.01	36.65	-0.07	1.11

Table 8

from which we infer that the most important driver for the house value in cluster 2 is the variable *rm*, followed by *dis* and *ptratio*. Since all the VIF values are below 5, apparently there is no multicollinearity problem.

We can have a visual representation of the model output as in Fig. 19: the residuals seem reasonably distributed around 0. However, as for Model 1.2, there is a normality issue that will be discussed more in Sec. 3.3.



(a) Residuals of the regression model for cluster 2. (b) Actual versus fitted values for the regression model.

Figure 19

3.3 Limitations and drawbacks of the models

Linear regression assumes that the relationship between the dependent variable and the independent variables is linear and that the errors are normally distributed, independent, and have a constant variance (homoscedastic-

ity). These assumptions, if met, assure the best linear unbiased estimates for the model parameters. However, real-world data often deviate from these assumptions. Therefore, statistical tests become critical in evaluating the adequacy of these assumptions and the overall validity of the model.

The most common statistical tests that provide information about the validity of the regression assumptions for both Model 1.2 and Model 2.2 are reported in Tab. 9.

	Model 1.2	**Model 2.2**
Omnibus	19.818	23.698
Prob(Omnibus)	0.000	0.000
Skew	0.774	0.488
Kurtosis	4.854	4.132
Durbin-Watson	1.337	1.225
Jarque-Bera (JB)	30.395	33.610
Prob(JB)	2.51e-07	5.03e-08
Cond. No.	503	1.01e+04

Table 9: Statistic tests for Model 1.2 and Model 2.2.

- Omnibus: the Omnibus test is used for testing the skewness and kurtosis. The Omnibus test for both models is quite high (19.818 for Model 1.2 and 23.698 for Model 2.2), suggesting that the residuals may not be normally distributed.
- Durbin-Watson: both models have a Durbin-Watson statistic slightly below 2, indicating there might be a slight positive autocorrelation.
- Skew: both models have a skew value below 1 which means the models have moderate skewness.
- Kurtosis: Model 1.2 (4.854) shows a higher kurtosis than Model 2.2 (4.132), indicating a more serious outlier problem in Model 1.2.
- Jarque-Bera (JB): both models have fairly high JB values, indicating the distribution may not be normal.
- Condition Number (Cond. No.): for both models, the condition number is high, especially for Model 2.2 (1.01e+04), indicating potential multicollinearity issues despite the VIF values are all smaller than 5.

The most apparent risk in both models comes from the potentially high multicollinearity (as indicated by the high Condition Number), which can affect the stability and interpretability of the coefficient estimates. To address the potential multicollinearity problem, given that we have already removed all the variables with high VIF value, we could consider using techniques such as ridge regression or principal component analysis, which goes beyond the scope of this assignment.

If the residuals are not normally distributed or if there are more outliers (as suggested by the high Kurtosis, Omnibus and Jarque-Bera tests), this could affect the validity of the model assumptions and potentially lead to unreliable predictions. The most straightforward initial solutions to consider include: re-evaluating the data for outliers that could impact the residuals' normality, applying transformations to the variables (such as log transformations) which would however affect the interpretability of the model, or utilizing a different model that does not necessitate this assumption. A thorough analysis of this problem would go beyond the scope of this assignment.

Note however that improving a model is not just about getting better statistics: it is about making it more representative of the underlying reality. As such, any changes should also make sense in the context of the data at hand. Even if a model does not pass all statistical tests, it may still hold value in a business context by providing a general idea of trends and serving as a foundation for deeper analysis. However, it becomes absolutely essential to disclose the model's limitations and potential drawbacks when presenting the results. This ensures that the users of the model comprehend the associated risks and are not misled by the outcomes.

4 Recommendations

In this section, we will take the perspective of a real estate agency. For such an agency, understanding the median value of homes is pivotal for a variety of reasons. These include the development of pricing strategies, offering informed investment advice to clients, and crafting effective marketing campaigns.

The analysis conducted in this assignment can benefit a real estate business in several ways. The first key insight is that it is possible to form two distinct clusters, each with a unique set of features that can predict the median house value. Interestingly, these two clusters are nearly 'orthogonal'. The only common predictive variable for median house prices across both clusters is *rm*. All other variables serve as predictors in at most one of the cluster models. We will now delve into specific recommendations catered to each of these clusters.

Cluster 1 Cluster 1 comprises towns and suburbs in regions with a tax rate (*tax*) of 666. These regions likely encompass peripheral locations marked by high levels of industrialization and easy accessibility to radial highways. The key differentiating characteristic of this cluster is the non-zero crime rate, in contrast to cluster 2 where $crim \simeq 0$. Of the predictors given, *lstat* has the most significant effect on the price (given the highest absolute value of the standard β coefficient, -0.64), followed by *crim*, *rm*, and *nox*, as presented in Tab. 5. It is worth noting that all coefficients are negative. While the negative correlation for *crim*, *nox*, and *lstat* is intuitive, a negative β for *rm* contradicts expectations — typically, we would expect a house with more rooms (hence larger) to have a higher price. This unexpected relationship could have various explanations; for instance, the model may be capturing the relationship between *rm* and *dis*: large values of *dis* correspond to upper band of the variable *rm*. Although *dis* was discarded in Model 1.2, it had a correlation of 0.55 with *medv* (see Fig. 16).

A real estate agency could leverage the insights from Model 1.2 in several ways. For instance:

- Profit Maximization: the agency could focus on buying and selling properties in areas characterized by lower crime rates, lower nitric oxide concentrations, and a smaller proportion of lower-status populations to maximize profits.
- Sales Volume Increase: properties in areas with higher proportions of lower-status populations, elevated crime rates, and higher nitric oxide concentrations might be more affordable, potentially making them easier to sell in high volumes.
- Rehabilitation Projects: the agency could use this model to identify areas with lower house prices that are prime candidates for investment and improvement.
- Investor Services: the agency could provide clients with advice on the potential risks and rewards of investing in various areas.

Cluster 2 Cluster 2 incorporates locations characterized by an extremely low criminality rate. The most impactful of the predictors in this cluster these is the average number of rooms *rm*, with a standard β coefficient of 0.75, encoding the intuitive expectation of higher home prices with an increased number of rooms. This is followed by *age*, *dis*, and *tax*, which all exhibit negative beta coefficients, indicating a decrease in home prices with an increase in these predictor values. Interestingly, the pupil-teacher ratio by town *ptratio* and *black* also exhibit negative standard beta coefficients. While the former is slightly counterintuitive, the latter roughly suggests that housing prices grow when the percentage of black people decreases, see Fig. 2.

A real estate agency could apply these insights from Model 2.2 in several strategic ways:

- Profit Maximization: by focusing on properties with a higher number of rooms (*rm*) and lower *age*, *dis*, and *tax* values, the agency could potentially maximize its profits.
- Investor Services: the agency could advise investors on properties with low *ptratio* and *black* values, providing them with an understanding of potential investment opportunities and risks.
- Targeted Marketing: understanding the impact of these variables on housing prices could allow the agency to tailor their marketing strategies, focusing on properties that meet specific buyer profiles.
- Strategic Property Acquisition: the agency could target properties for acquisition based on these predictors, focusing on acquiring assets that align with these insights to ensure maximum return on investment.

A Scatterplots

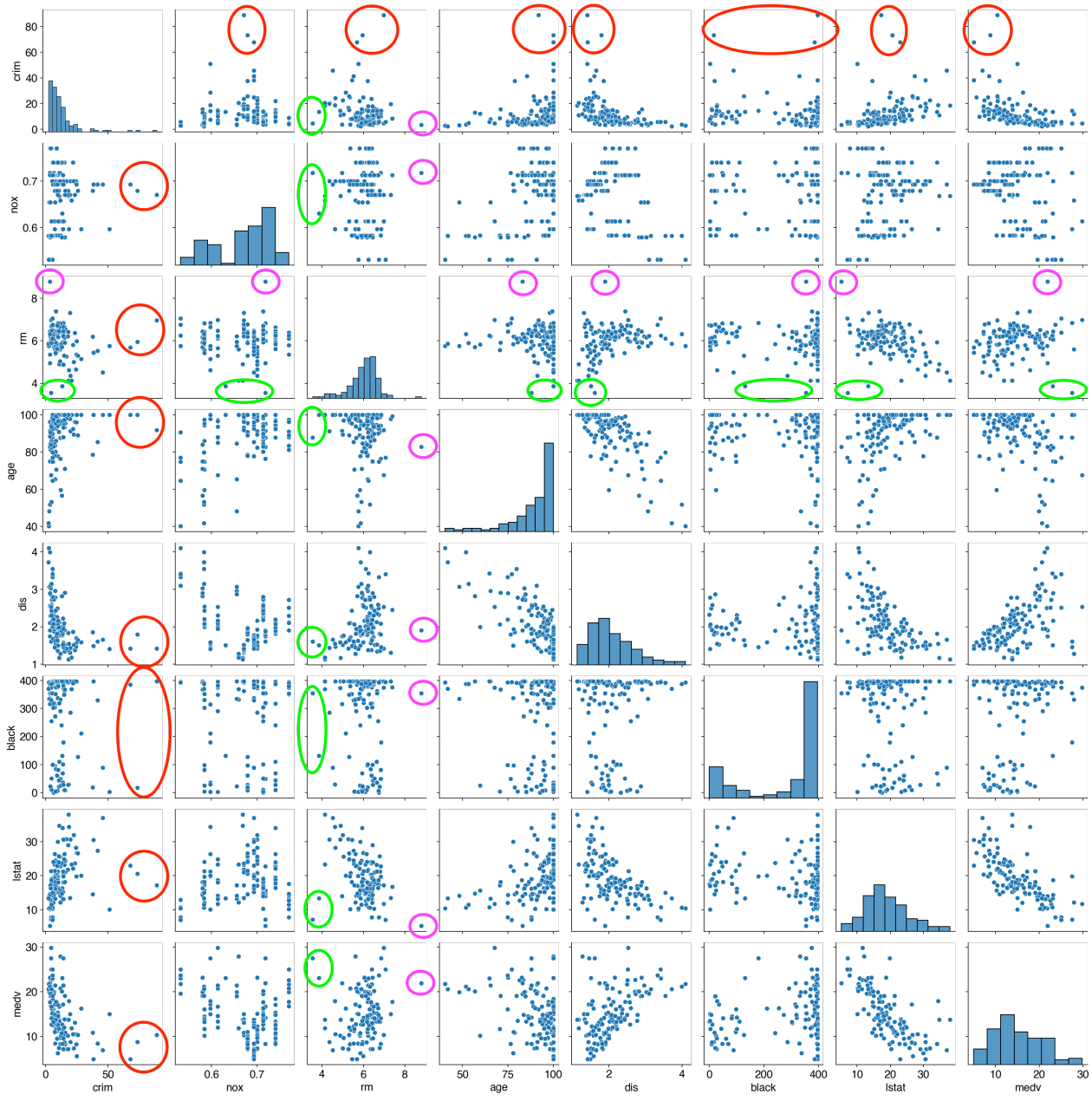


Figure 20: scatterplot for cluster 1.

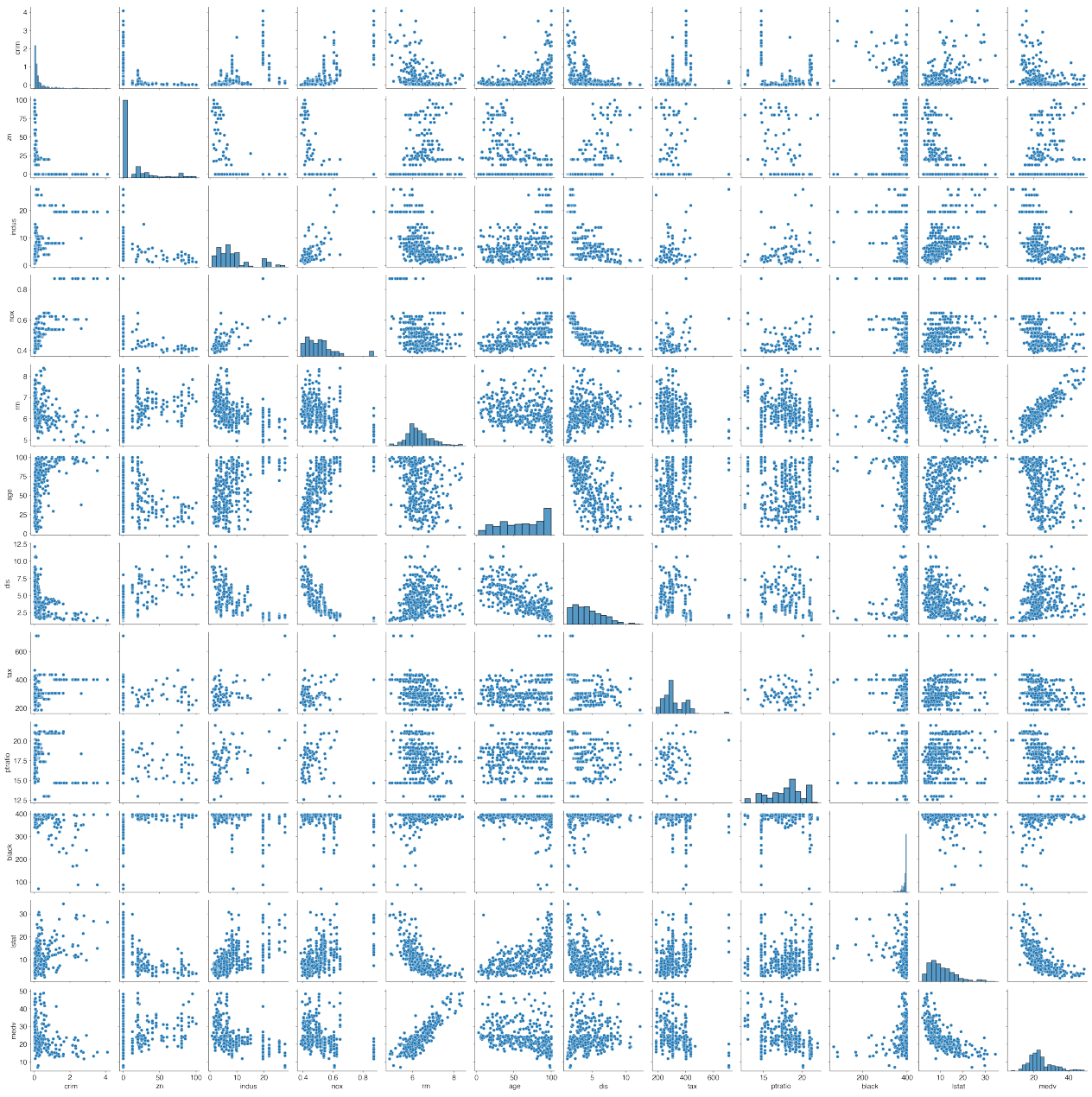


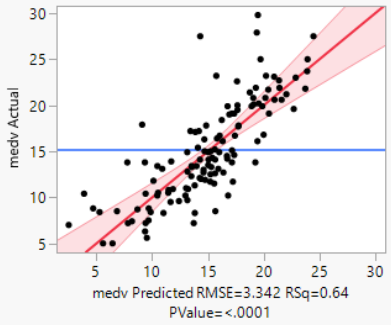
Figure 21: scatterplot for cluster 2.

B JMP checks

Response medv

Whole Model

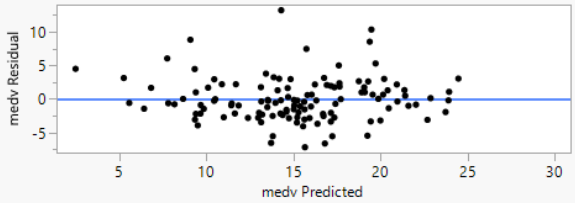
Actual by Predicted Plot



Effect Summary

Source	Logworth	PValue
lstat	11.921	0.00000
crim	3.830	0.00015
nox	3.133	0.00074
rm	2.317	0.00482
chas	1.207	0.06211
black	1.036	0.09214
age	0.388	0.40934
dis	0.266	0.54185

Residual by Predicted Plot



Summary of Fit

RSquare	0.642633
RSquare Adj	0.617987
Root Mean Square Error	3.34201
Mean of Response	15.1688
Observations (or Sum Wgts)	125

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	2329.8208	291.228	26.0746
Error	116	1295.6075	11.169	Prob > F
C. Total	124	3625.4283		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	43.002864	6.19338	6.94	<.0001*
crim	-0.141977	0.03618	-3.92	0.0001*
chas	3.1421267	1.66807	1.88	0.0621
nox	-19.78328	5.704803	-3.47	0.0007*
rm	-1.389045	0.483314	-2.87	0.0048*
age	0.0298006	0.035989	0.83	0.4093
dis	0.5230272	0.854868	0.61	0.5419
black	0.0036384	0.002142	1.70	0.0921
lstat	-0.491757	0.061673	-7.97	<.0001*

Durbin-Watson

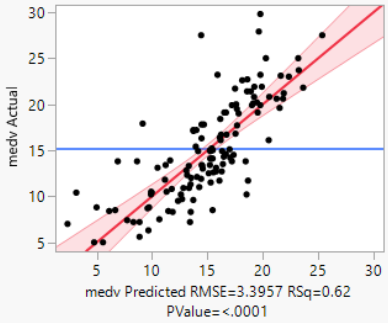
Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.4405219	125	0.2794	0.0002*

(a) JMP results for Model 1.1.

Response medv

Whole Model

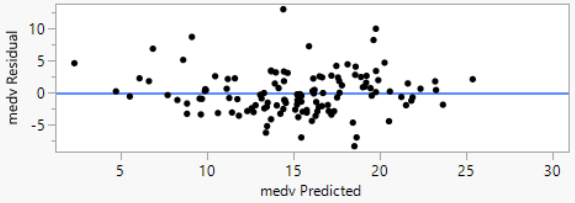
Actual by Predicted Plot



Effect Summary

Source	Logworth	PValue
lstat	15.665	0.00000
crim	4.732	0.00002
nox	2.546	0.00284
rm	2.413	0.00387

Residual by Predicted Plot



Summary of Fit

RSquare	0.618344	AICc	667.9687
RSquare Adj	0.605622	BIC	684.2267
Root Mean Square Error	3.395666		
Mean of Response	15.1688		
Observations (or Sum Wgts)	125		

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	2241.7629	560.441	48.6049
Error	120	1383.6655	11.531	Prob > F
C. Total	124	3625.4283		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	45.669823	4.479978	10.19	<.0001*	36.799779	54.539866
crim	-0.152264	0.034134	-4.46	<.0001*	-0.219847	-0.08468
nox	-15.17773	4.980984	-3.05	0.0028*	-25.03973	-5.315732
rm	-1.391625	0.472359	-2.95	0.0039*	-2.326863	-0.456386
lstat	-0.528456	0.055401	-9.54	<.0001*	-0.638147	-0.418765

Durbin-Watson

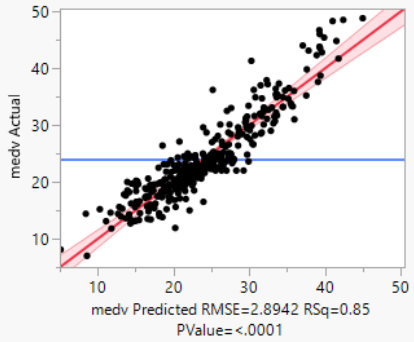
Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.3369803	125	0.3274	<.0001*

(b) JMP results for Model 1.2.

Figure 22

Response medv

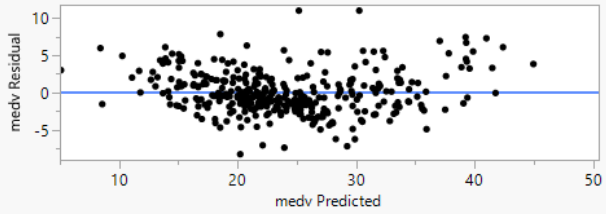
Actual by Predicted Plot



Effect Summary

Source	Logworth	PValue
rm	64.709	0.00000
dis	9.508	0.00000
ptratio	8.852	0.00000
age	7.023	0.00000
tax	4.823	0.00002
black	2.185	0.00654
zn	1.191	0.06445
nox	0.940	0.11479
lstat	0.898	0.12643
rad_medium	0.754	0.17621
crim	0.660	0.21864
chas	0.525	0.29855
indus	0.198	0.63340

Residual by Predicted Plot



Summary of Fit

RSquare	0.848373	AICc	1808.853
RSquare Adj	0.842692	BIC	1865.795
Root Mean Square Error	2.894164		
Mean of Response	23.9446		
Observations (or Sum Wgts)	361		

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	13	16262.455	1250.96	149.3469
Error	347	2906.537	8.38	Prob > F
C. Total	360	19168.992		<.0001*

Parameter Estimates

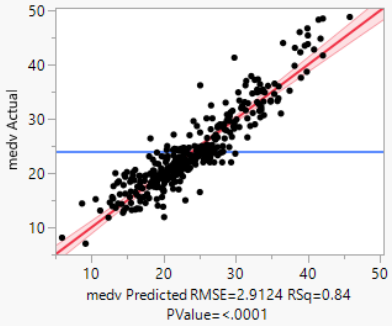
Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-11.01858	4.600124	-2.40	0.0171*	-20.06622	-1.970949
crim	0.5199065	0.421871	1.23	0.2186	-0.309839	1.3496517
zn	0.0167152	0.009011	1.85	0.0645	-0.001008	0.0344387
indus	-0.019445	0.040734	-0.48	0.6334	-0.099561	0.0606707
chas	0.6549009	0.629037	1.04	0.2985	-0.582304	1.8921061
nox	-5.377422	3.401294	-1.58	0.1148	-12.06717	1.3123247
rm	8.3230271	0.388669	21.41	<.0001*	7.5585827	9.0874715
tax	-0.010143	0.00231	-4.39	<.0001*	-0.014686	-0.005599
rad_medium	0.4793937	0.353726	1.36	0.1762	-0.216324	1.1751111
ptratio	-0.559837	0.089962	-6.22	<.0001*	-0.736775	-0.382898
black	0.0142987	0.005226	2.74	0.0065*	0.00402	0.0245775
lstat	-0.072331	0.047213	-1.53	0.1264	-0.165192	0.0205285
age	-0.050442	0.009252	-5.45	<.0001*	-0.068639	-0.032244
dis	-0.865777	0.13356	-6.48	<.0001*	-1.128467	-0.603087

(a) JMP results for Model 2.1.

Response medv

Whole Model

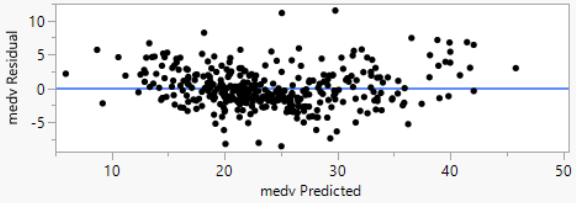
Actual by Predicted Plot



Effect Summary

Source	Logworth	PValue
rm	105.402	0.00000
ptratio	14.901	0.00000
age	14.282	0.00000
dis	10.106	0.00000
tax	6.374	0.00000
black	2.772	0.00169

Residual by Predicted Plot



Summary of Fit

RSquare	0.843364	AICc	1805.604
RSquare Adj	0.840709	BIC	1836.306
Root Mean Square Error	2.912354		
Mean of Response	23.9446		
Observations (or Sum Wgts)	361		

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	16166.433	2694.41	317.6689
Error	354	3002.559	8.48	Prob > F
C. Total	360	19168.992		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-16.63576	3.060128	-5.44	<.0001*	-22.65408	-10.61745
rm	8.9475608	0.280542	31.89	<.0001*	8.3958214	9.4993002
age	-0.062015	0.007583	-8.18	<.0001*	-0.076929	-0.047101
dis	-0.667433	0.099497	-6.71	<.0001*	-0.863113	-0.471753
tax	-0.010506	0.002038	-5.16	<.0001*	-0.014514	-0.006498
ptratio	-0.637857	0.076111	-8.38	<.0001*	-0.787543	-0.48817
black	0.0143594	0.004538	3.16	0.0017*	0.0054342	0.0232845

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.2250273	361	0.3734	<.0001*

(b) JMP results for Model 2.2.

Figure 23