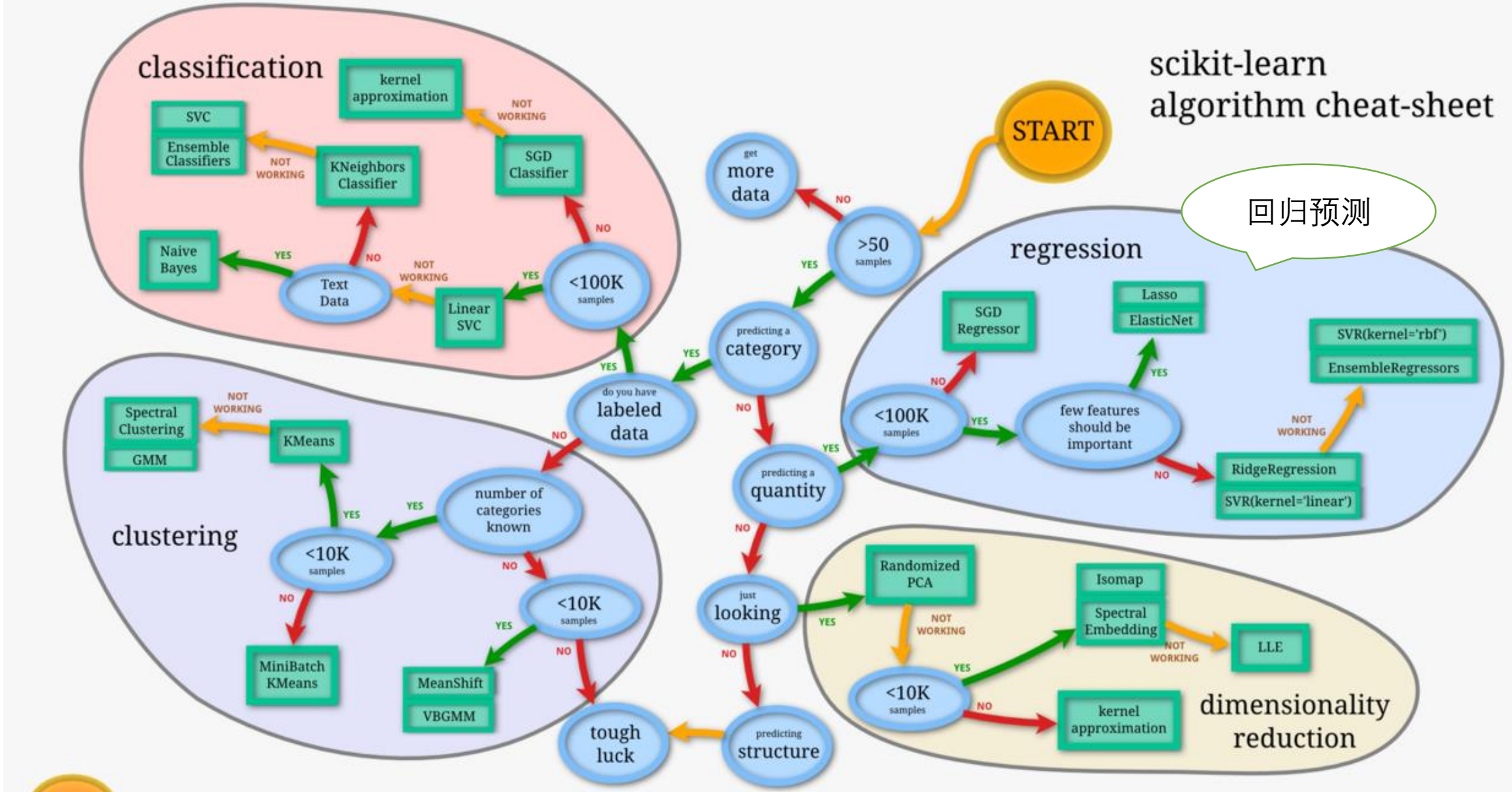
PREDICT BOSTON HOUSE PRICES

LINEAR REGRESSION


scikit-learn
algorithm cheat-sheet
START
回归预测
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
NO
RidgeRegression
SVR(kernel='linear')
YES
predicting a category
NO
do you have labeled data
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
MiniBatch KMeans
<10K samples
YES
NO
MeanShift
VBGMM
predicting a quantity
YES
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn


MASSACHVSETTS INSTITVTE OF TE

VE RI TAS

HARVARD
UNIVERSITY


FOR
SALE

“Numbers have an important story to tell.
They rely on you to give them a voice.”
— Stephen Few

The Boston Housing Dataset

A Dataset derived from information collected by the U.S. Census Service concerning housing in the area of Boston Mass.

Delve

This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. It was obtained from the StatLib archive (http://lib.stat.cmu.edu/datasets/boston), and has been used extensively throughout the literature to benchmark algorithms. However, these comparisons were primarily done outside of **Delve** and are thus somewhat suspect. The dataset is small in size with only 506 cases.

The data was originally published by Harrison, D. and Rubinfeld, D.L. `*Hedonic prices and the demand for clean air*', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Dataset Naming

The name for this dataset is simply **boston**. It has two prototasks: **nox**, in which the nitrous oxide level is to be predicted; and **price**, in which the median value of a home is to be predicted

Miscellaneous Details

- **Origin**
  The origin of the boston housing data is **Natural**.
- **Usage**
  This dataset may be used for **Assessment**.
- **Number of Cases**
  The dataset contains a total of **506** cases.
- **Order**
  The order of the cases is **mysterious**.
- **Variables**
  There are **14** attributes in each case of the dataset. They are:

  1. CRIM - per capita crime rate by town
  2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
  3. INDUS - proportion of non-retail business acres per town.
  4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
  5. NOX - nitric oxides concentration (parts per 10 million)
  6. RM - average number of rooms per dwelling
  7. AGE - proportion of owner-occupied units built prior to 1940
  8. DIS - weighted distances to five Boston employment centres
  9. RAD - index of accessibility to radial highways
  10. TAX - full-value property-tax rate per $10,000
  11. PTRATIO - pupil-teacher ratio by town
  12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
  13. LSTAT - % lower status of the population
  14. MEDV - Median value of owner-occupied homes in $1000's

- **Note**
  Variable #14 seems to be censored at 50.00 (corresponding to a median price of $50,000); Censoring is suggested by the fact that the highest median price of exactly $50,000 is reported in 16 cases, while 15 cases have prices between $40,000 and $50,000, with prices rounded to the nearest hundred. Harrison and Rubinfeld do not mention any censoring.

Last Updated 10 October 1996
Comments and questions to: delve@cs.toronto.edu

波士顿房价数据集
http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

数据集特征:

实例数量:	506
属性数量:	13 数值型或类别型，帮助预测的属性

:中位数（第14个属性）经常是学习目标

属性信息 (按顺序):

- CRIM 城镇人均犯罪率
- ZN 占地面积超过2.5万平方英尺的住宅用地比例
- INDUS 城镇非零售业务地区的比例
- CHAS 查尔斯河虚拟变量 (= 1 如果土地在河边；否则是0)
- NOX 一氧化氮浓度（每1000万份）
- RM 平均每居民房数
- AGE 在1940年之前建成的所有者占用单位的比例
- DIS 与五个波士顿就业中心的加权距离
- RAD 辐射状公路的可达性指数
- TAX 每10,000美元的全额物业税率
- PTRATIO 城镇师生比例
- B 1000(Bk - 0.63)^2 其中 Bk 是城镇的黑人比例
- LSTAT 人口中地位较低人群的百分数
- MEDV 以1000美元计算的自有住房的中位数

缺失属性值: 无

创建者: Harrison, D. and Rubinfeld, D.L.

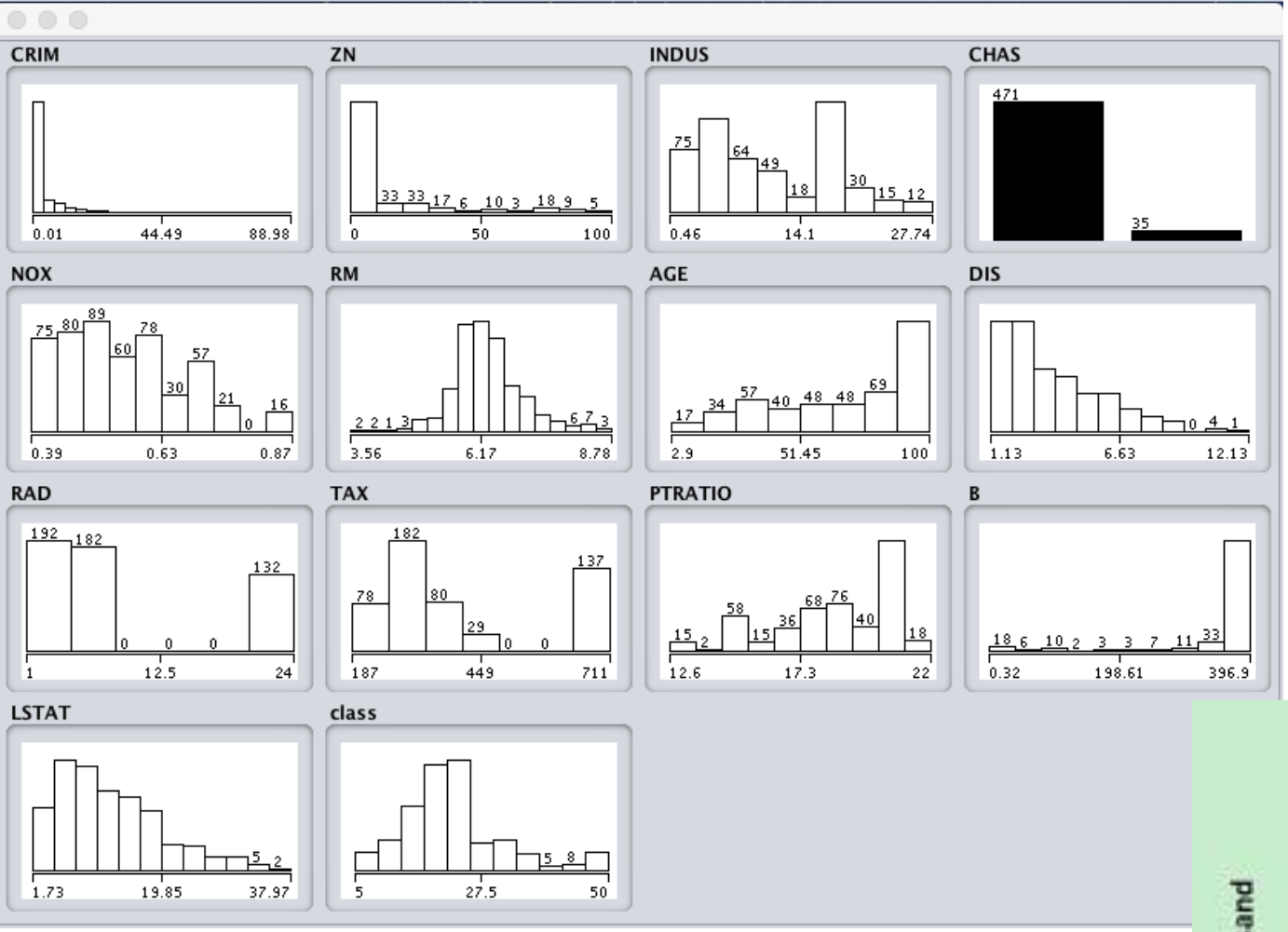
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	
2	0.00632	18.00	2.310	0	0.5380	6.5750	65.20	4.0900	1	296.0	15.30	396.90	4.98	24.00	
3	0.02731	0.00	7.070	0	0.4690	6.4210	78.90	4.9671	2	242.0	17.80	396.90	9.14	21.60	
4	0.02729	0.00	7.070	0	0.4690	7.1850	61.10	4.9671	2	242.0	17.80	392.83	4.03	34.70	
5	0.03237	0.00	2.180	0	0.4580	6.9980	45.80	6.0622	3	222.0	18.70	394.63	2.94	33.40	
6	0.06905	0.00	2.180	0	0.4580	7.1470	54.20	6.0622	3	222.0	18.70	396.90	5.33	36.20	
7	0.02985	0.00	2.180	0	0.4580	6.4300	58.70	6.0622	3	222.0	18.70	394.12	5.21	28.70	
8	0.08829	12.50	7.870	0	0.5240	6.0120	66.60	5.5605	5	311.0	15.20	395.60	12.43	22.90	
9	0.14455	12.50	7.870	0	0.5240	6.1720	96.10	5.9505	5	311.0	15.20	396.90	19.15	27.10	
10	0.21124	12.50	7.870	0	0.5240	5.6310	100.00	6.0821	5	311.0	15.20	386.63	29.93	16.50	
11	0.17004	12.50	7.870	0	0.5240	6.0040	85.90	6.5921	5	311.0	15.20	386.71	17.10	18.90	
12	0.22489	12.50	7.870	0	0.5240	6.3770	94.30	6.3467	5	311.0	15.20	392.52	20.45	15.00	
13	0.11747	12.50	7.870	0	0.5240	6.0090	82.90	6.2267	5	311.0	15.20	396.90	13.27	18.90	
14	0.09378	12.50	7.870	0	0.5240	5.8890	39.00	5.4509	5	311.0	15.20	390.50	15.71	21.70	
15	0.62976	0.00	8.140	0	0.5380	5.9490	61.80	4.7075	4	307.0	21.00	396.90	8.26	20.40	
16	0.63796	0.00	8.140	0	0.5380	6.0960	84.50	4.4619	4	307.0	21.00	380.02	10.26	18.20	
17	0.62739	0.00	8.140	0	0.5380	5.8340	56.50	4.4986	4	307.0	21.00	395.62	8.47	19.90	
18	1.05393	0.00	8.140	0	0.5380	5.9350	29.30	4.4986	4	307.0	21.00	386.85	6.58	23.10	
19	0.78420	0.00	8.140	0	0.5380	5.9900	81.70	4.2579	4	307.0	21.00	386.75	14.67	17.50	
20	0.80271	0.00	8.140	0	0.5380	5.4560	36.60	3.7965	4	307.0	21.00	288.99	11.69	20.20	
21	0.72580	0.00	8.140	0	0.5380	5.7270	69.50	3.7965	4	307.0	21.00	390.95	11.28	18.20	
22															

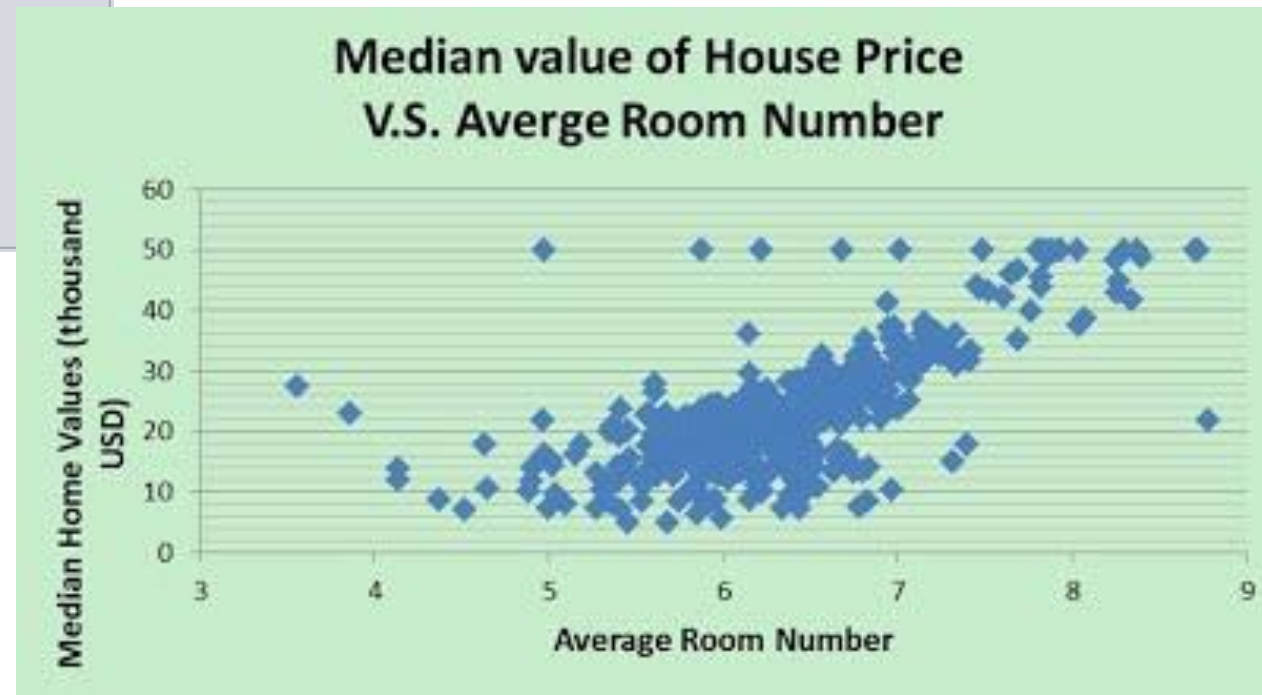
这是UCI ML（欧文加利福尼亚大学 机器学习库）房价数据集的副本。 http://archive.ics.uci.edu/ml/datasets/Housing

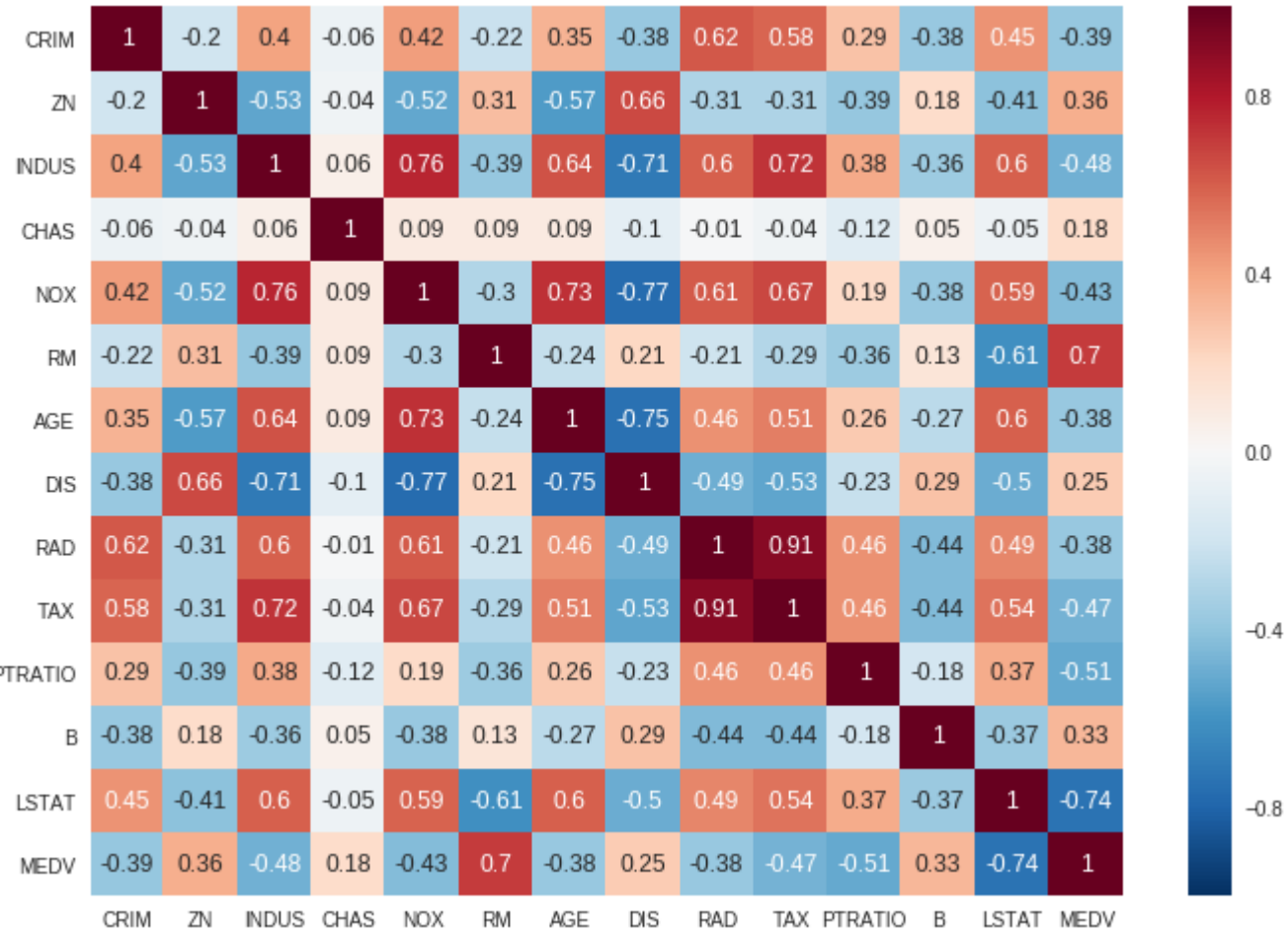
该数据集是从位于卡内基梅隆大学维护的StatLib图书馆取得的。

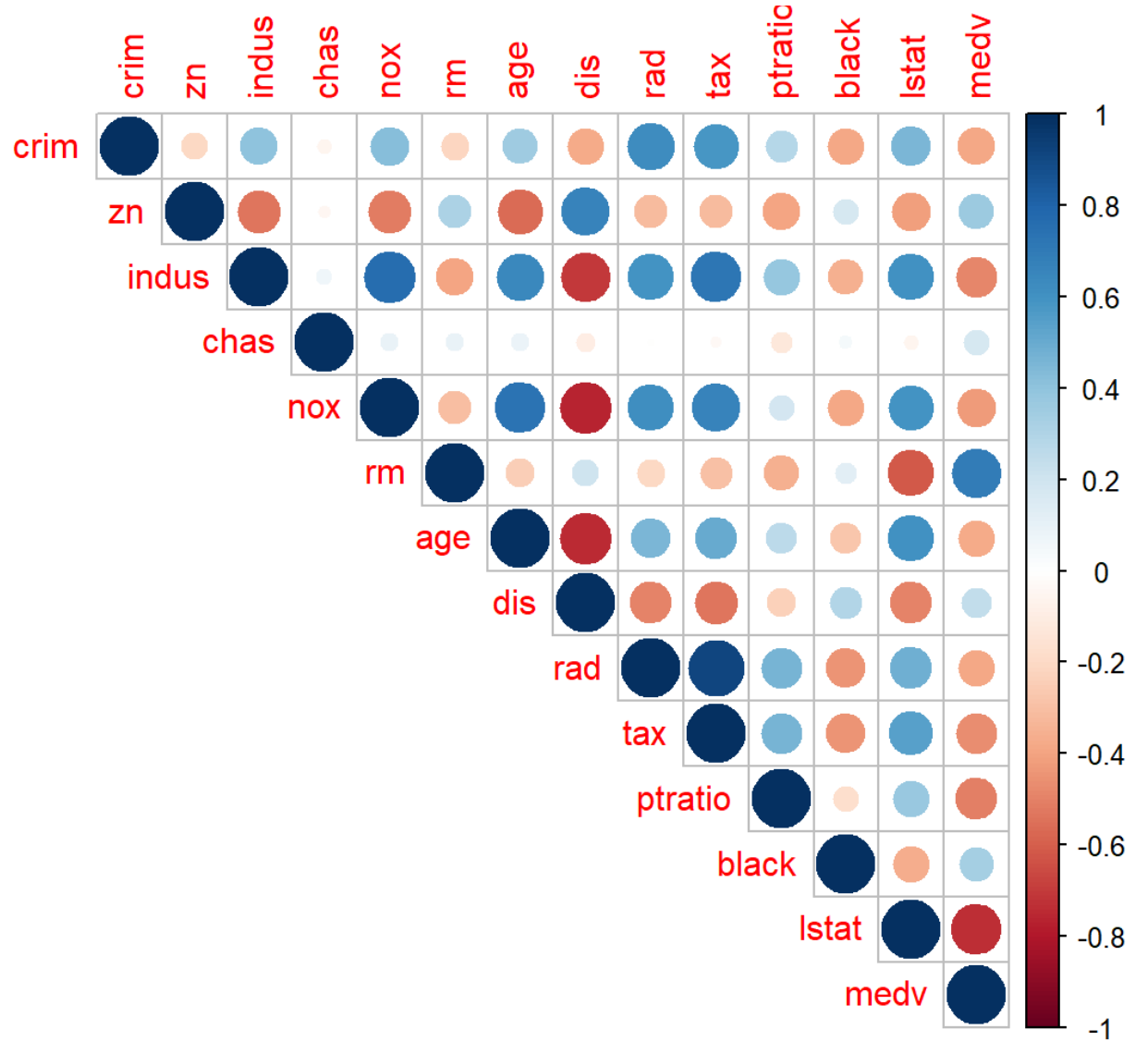
Harrison, D. 和 Rubinfeld, D.L. 的波士顿房价数据：'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978，也被使用在 Belsley, Kuh & Welsch 的 'Regression diagnostics ...', Wiley, 1980。 注释：许多变化已经被应用在后者第244-261页的表中。

https://archive.ics.uci.edu/ml/machine-learning-databases/housing/
https://www.kaggle.com/c/boston-housing


CRIM
0.01
44.49
88.98
ZN
0
50
100
INDUS
0.46
14.1
27.74
CHAS
471
35
NOX
0.39
0.63
0.87
RM
3.56
6.17
8.78
AGE
2.9
51.45
100
DIS
1.13
6.63
12.13
RAD
1
12.5
24
TAX
187
449
711
PTRATIO
12.6
17.3
22
B
0.32
198.61
396.9
LSTAT
1.73
19.85
37.97
class
5
27.5
50


Median value of House Price
V.S. Averge Room Number
Median Home Values (thousand USD)
Average Room Number
0
10
20
30
40
50
60
3
4
5
6
7
8
9


CRIM 1 -0.2 0.4 -0.06 0.42 -0.22 0.35 -0.38 0.62 0.58 0.29 -0.38 0.45 -0.39
ZN -0.2 1 -0.53 -0.04 -0.52 0.31 -0.57 0.66 -0.31 -0.31 -0.39 0.18 -0.41 0.36
INDUS 0.4 -0.53 1 0.06 0.76 -0.39 0.64 -0.71 0.6 0.72 0.38 -0.36 0.6 -0.48
CHAS -0.06 -0.04 0.06 1 0.09 0.09 0.09 -0.1 -0.01 -0.04 -0.12 0.05 -0.05 0.18
NOX 0.42 -0.52 0.76 0.09 1 -0.3 0.73 -0.77 0.61 0.67 0.19 -0.38 0.59 -0.43
RM -0.22 0.31 -0.39 0.09 -0.3 1 -0.24 0.21 -0.21 -0.29 -0.36 0.13 -0.61 0.7
AGE 0.35 -0.57 0.64 0.09 0.73 -0.24 1 -0.75 0.46 0.51 0.26 -0.27 0.6 -0.38
DIS -0.38 0.66 -0.71 -0.1 -0.77 0.21 -0.75 1 -0.49 -0.53 -0.23 0.29 -0.5 0.25
RAD 0.62 -0.31 0.6 -0.01 0.61 -0.21 0.46 -0.49 1 0.91 0.46 -0.44 0.49 -0.38
TAX 0.58 -0.31 0.72 -0.04 0.67 -0.29 0.51 -0.53 0.91 1 0.46 -0.44 0.54 -0.47
PTRATIO 0.29 -0.39 0.38 -0.12 0.19 -0.36 0.26 -0.23 0.46 0.46 1 -0.18 0.37 -0.51
B -0.38 0.18 -0.36 0.05 -0.38 0.13 -0.27 0.29 -0.44 -0.44 -0.18 1 -0.37 0.33
LSTAT 0.45 -0.41 0.6 -0.05 0.59 -0.61 0.6 -0.5 0.49 0.54 0.37 -0.37 1 -0.74
MEDV -0.39 0.36 -0.48 0.18 -0.43 0.7 -0.38 0.25 -0.38 -0.47 -0.51 0.33 -0.74 1
CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT MEDV
0.8
0.4
0.0
−0.4
−0.8


crim
zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
black
lstat
medv
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


MEDV
RM
ZN
B
DIS
CHAS
AGE
RAD
CRIM
NOX
TAX
INDUS
PTRATIO
LSTAT
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Data Cleaning/ Checking

Variable Selection
(Random Forest)

Training/Test Set

Machine Learning

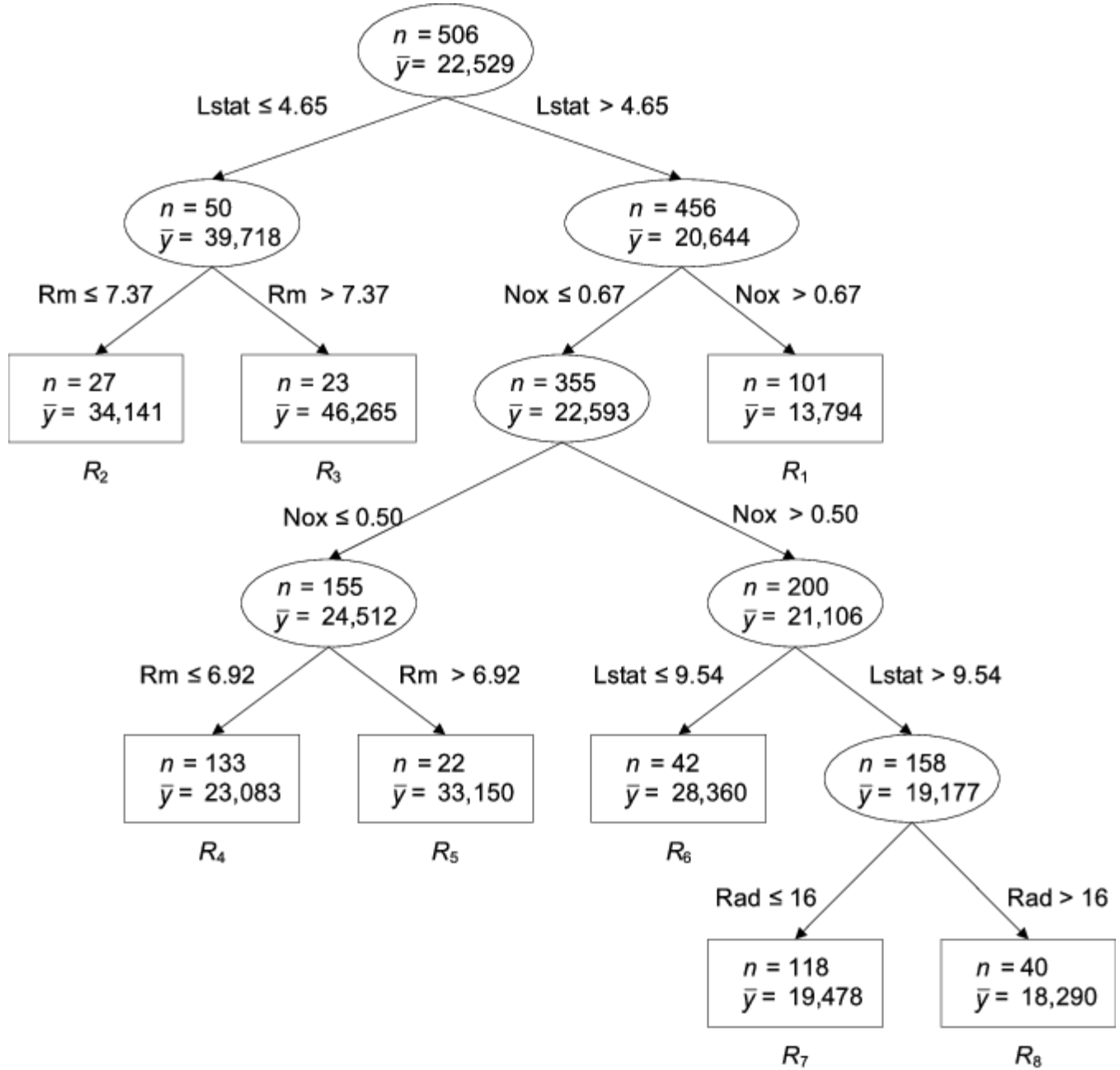
(1) Linear Regression

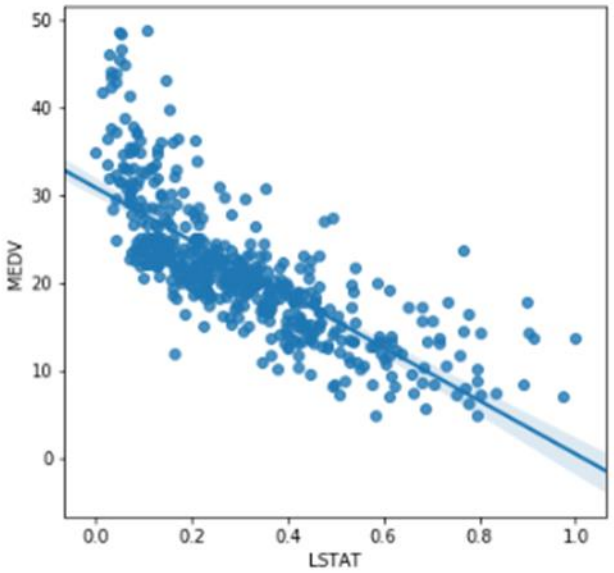
(2) Polynomial Regression

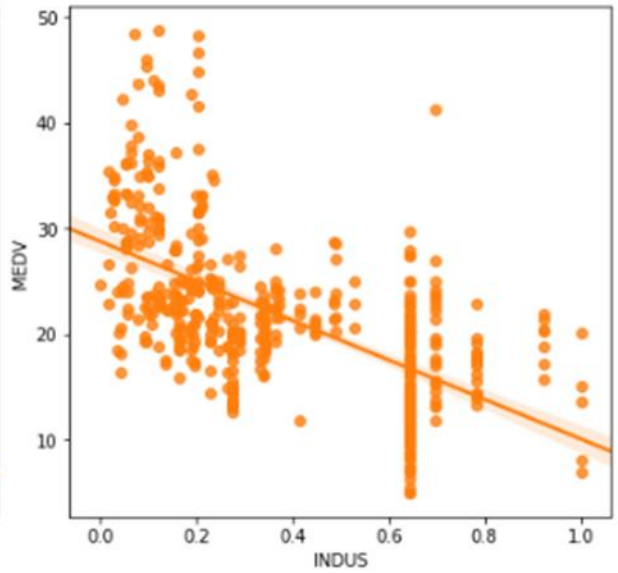
(3) Multivariate regression

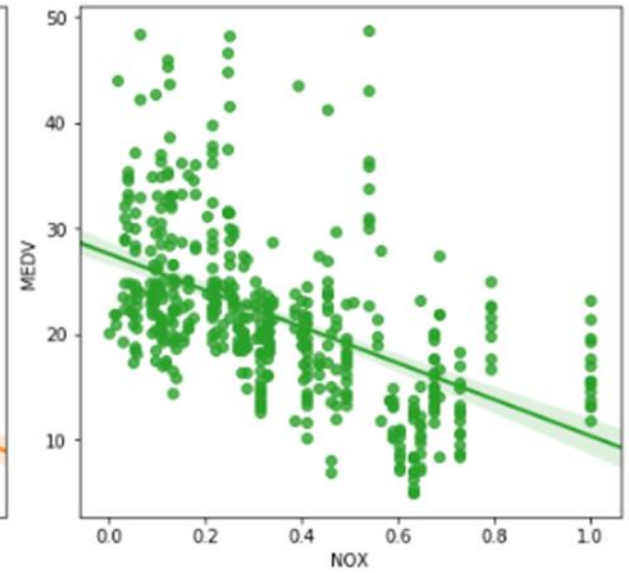
(4) Tree Regression

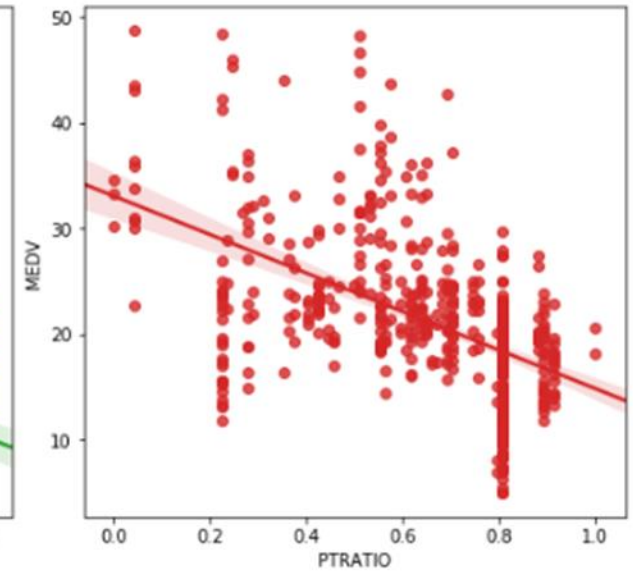
(5) Boosting/ Bagging

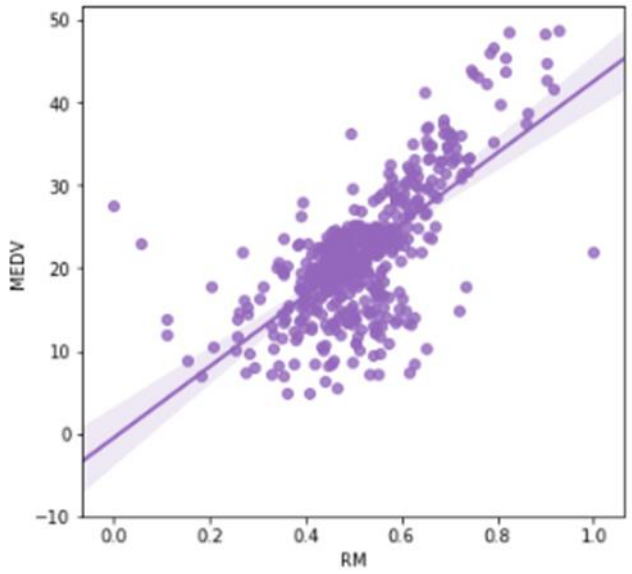


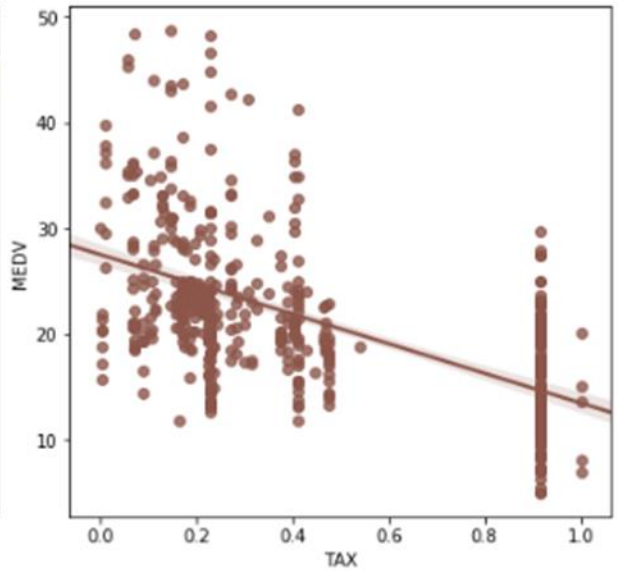

MEDV
LSTAT

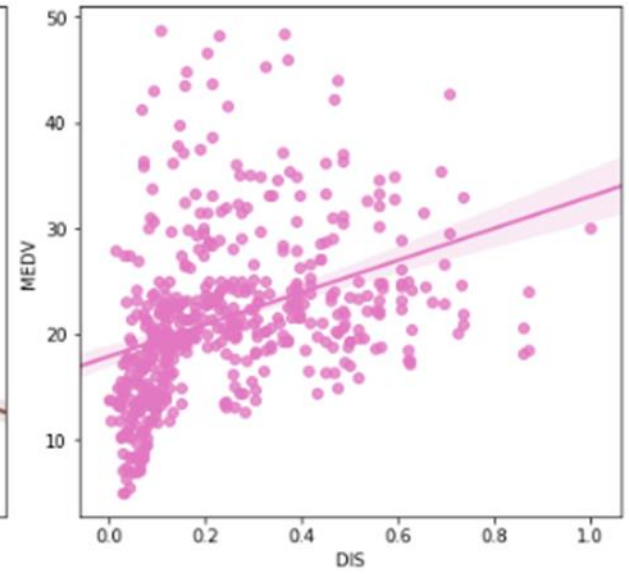

MEDV
INDUS

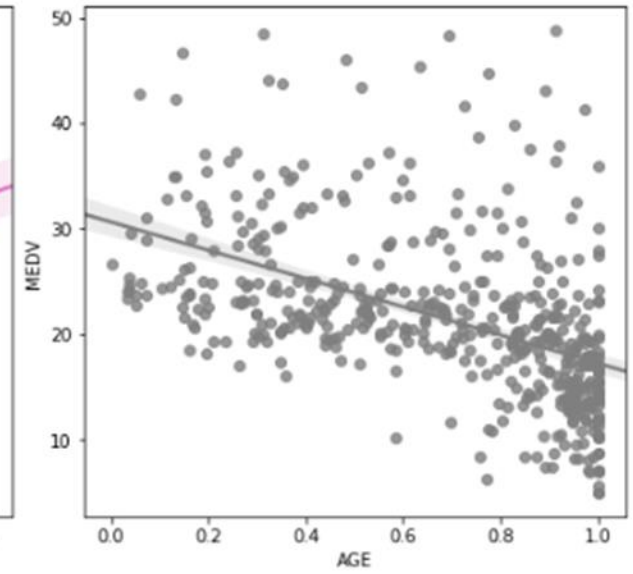

MEDV
NOX

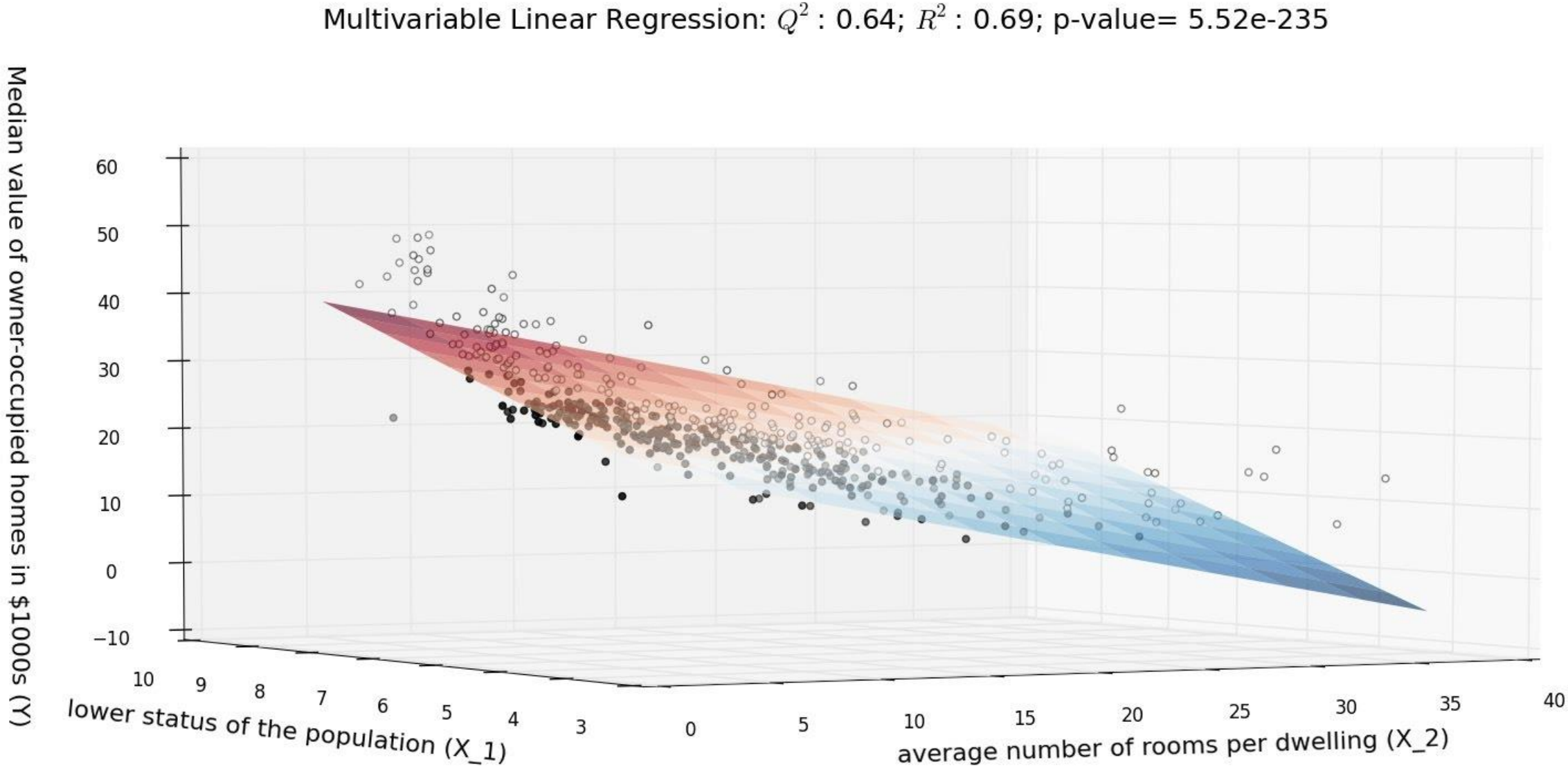

MEDV
PTRATIO

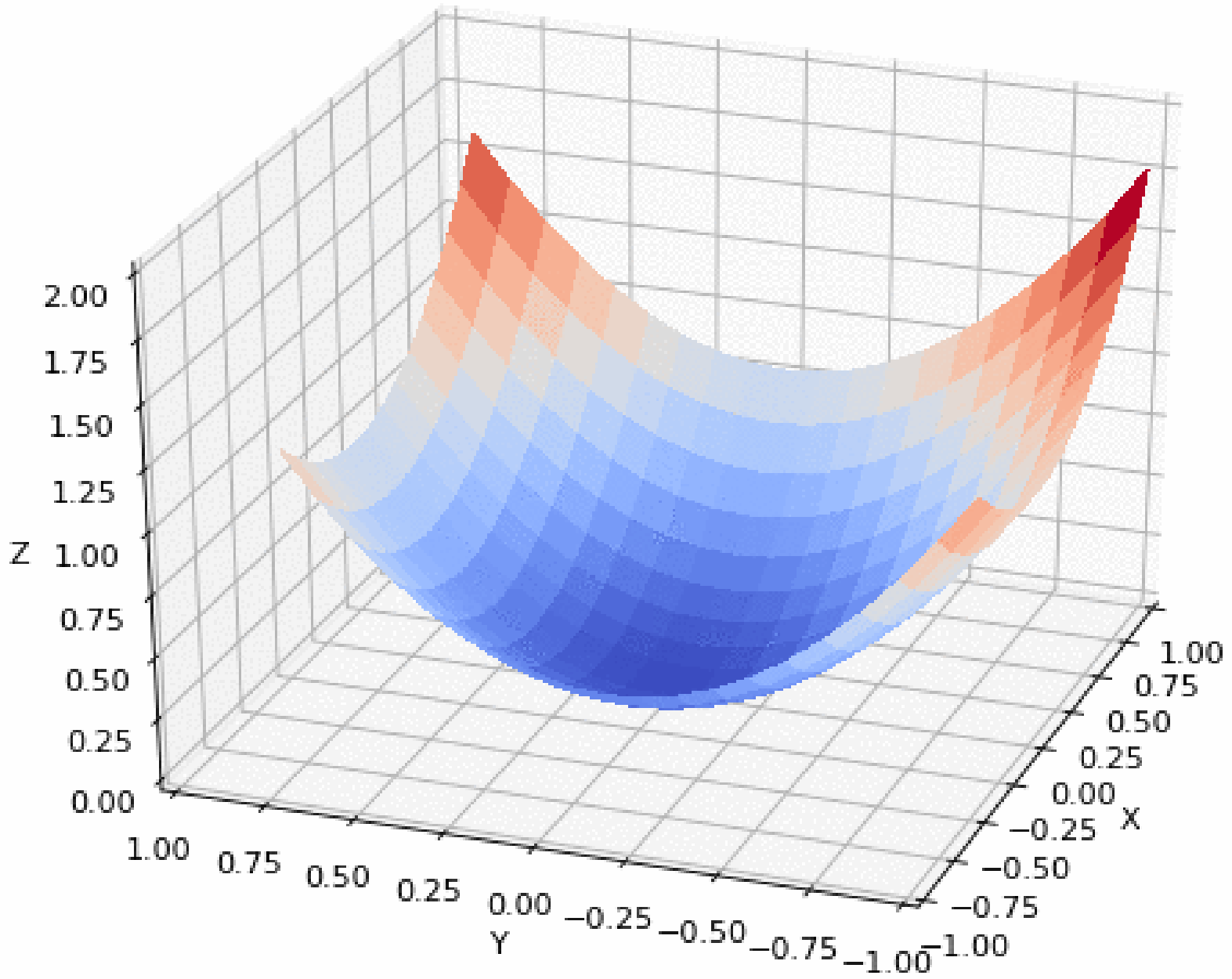

MEDV
RM


MEDV
TAX


MEDV
DIS


MEDV
AGE


Multivariable Linear Regression: Q^2 : 0.64; R^2 : 0.69; p-value= 5.52e-235
Median value of owner-occupied homes in $1000s (Y)
60
50
40
30
20
10
0
−10
10
9
8
7
6
5
4
3
lower status of the population (X_1)
0
5
10
15
20
25
30
35
40
average number of rooms per dwelling (X_2)


2.00
1.75
1.50
1.25
Z 1.00
0.75
0.50
0.25
0.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Y
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
X