Boston Housing Prices

(Data Mining)

Gustavo Jose Hernandez Sotres

Alberto Sandoval Castro

Rafael Juárez Badillo Chávez

Diego Pintor Ochoa

Repo: https://github.com/gustavoSo3/ml_project_2_boston_hosing

a) Download the Boston Housing Data from Kaggle at:

https://www.kaggle.com/altavish/boston-housing-dataset

a. How many rows are in this data set?

506

b. How many columns?

14

c. What do the rows and columns represent?

Las filas representan todas las entradas que fueron recopiladas

Y cada columna representa:

CRIM - per capita crime rate by town.

Que tanto crimen hay

ZN - proportion of residential land zoned for lots over 25,000 sq.ft.

Si hay más lotes grandes el número es mayor

INDUS - proportion of non-retail business acres per town.

Que tanta industria hay en la zona

CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)

Si está conectada al rio Charles

NOX - nitric oxides concentration (parts per 10 million)

La concentración de óxido nítrico

RM - average number of rooms per dwelling

El promedio de cuartos por edificio

AGE - proportion of owner-occupied units built prior to 1940

Proporción de dueños antes de 1940

DIS - weighted distances to five Boston employment centers

Distancia a los centros de trabajo

RAD - index of accessibility to radial highways

Que tan accesibles son las vías del tren

TAX - full-value property-tax rate per $10,000

El valor de propiedad en decenas de millares

PTRATIO - pupil-teacher ratio by town

La relación de estudiantes y maestros por miles

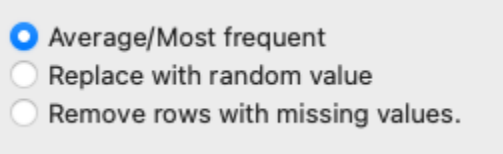
B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

LSTAT - % lower status of the population

MEDV - Median value of owner-occupied homes in $1000's

d. Resuelva los casos faltantes

En este caso los resolvimos poniendo el valor promedio



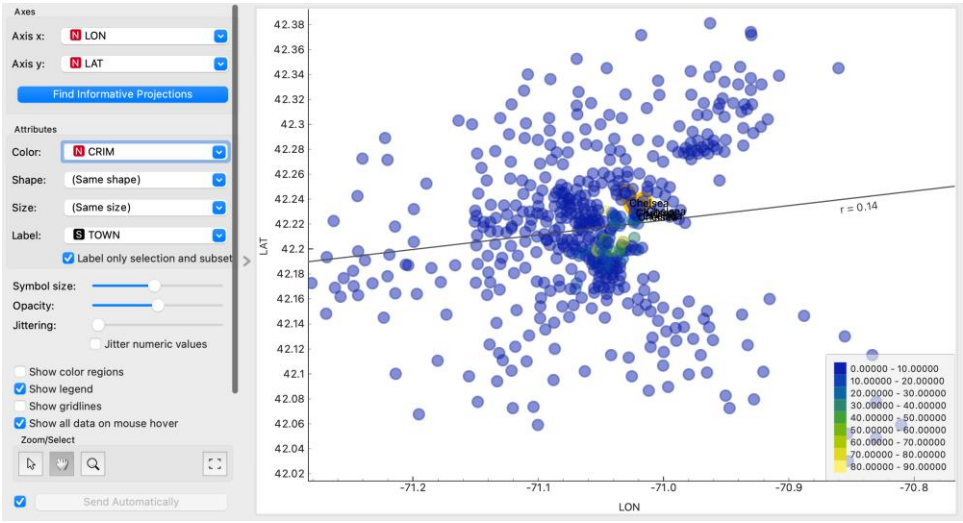
e. Descargue ahora los datos de

https://jakubnowosad.com/spData/reference/boston.html

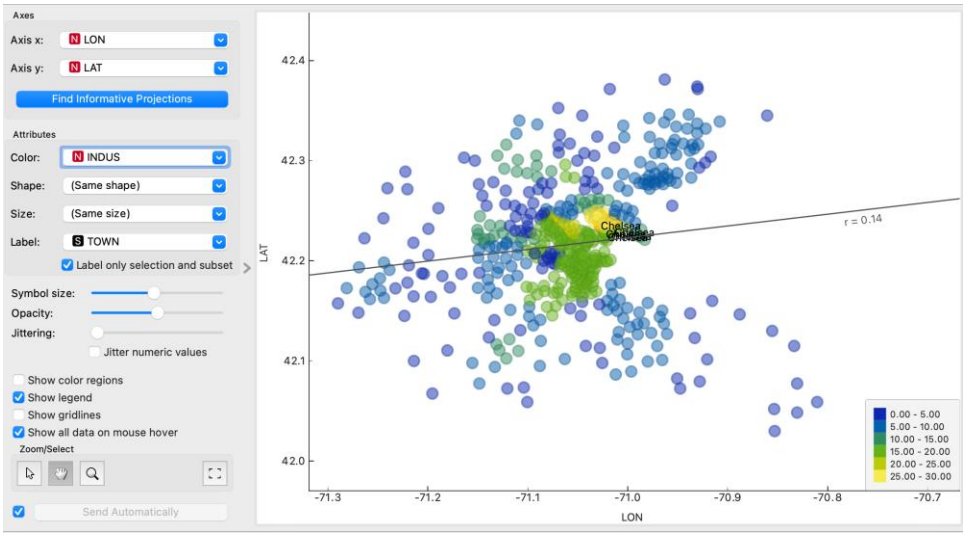
En las siguientes preguntas compare resultados entre los dos datasets.

b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

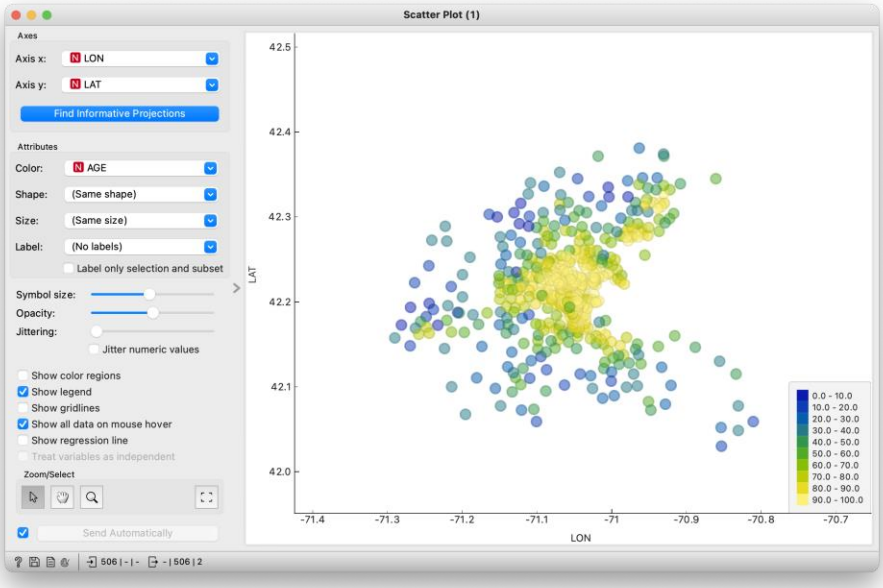
Lon & Lat, makes possible to kinda map the city, with CRIM, we're able to inference that crime is gathered at the center of the city

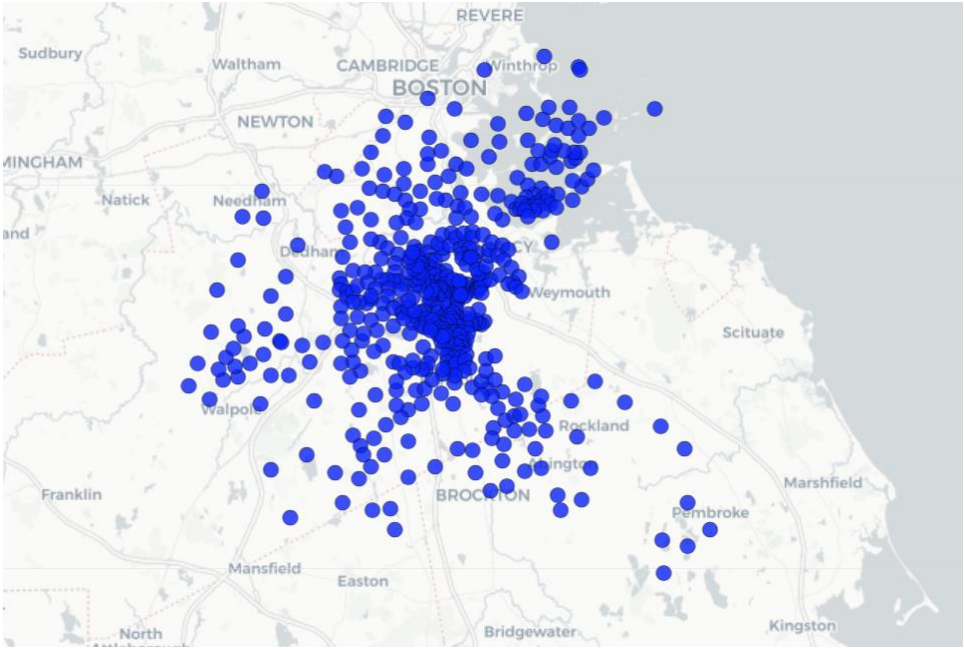


With Indus (numeric vector of proportions of non-retail business acres per town), and also with the image above, we found a correlation between non-retail business at the center of town and crime rate.

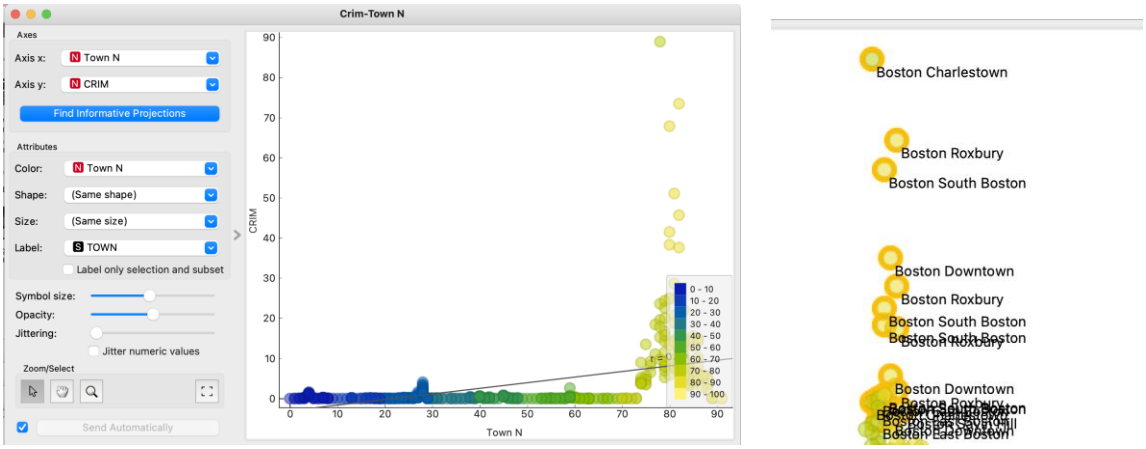


LON – LAT x AGE



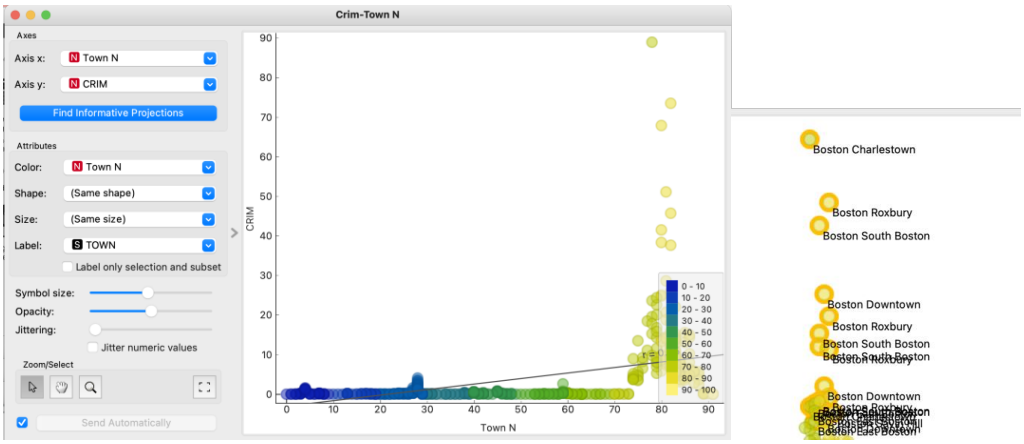


c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

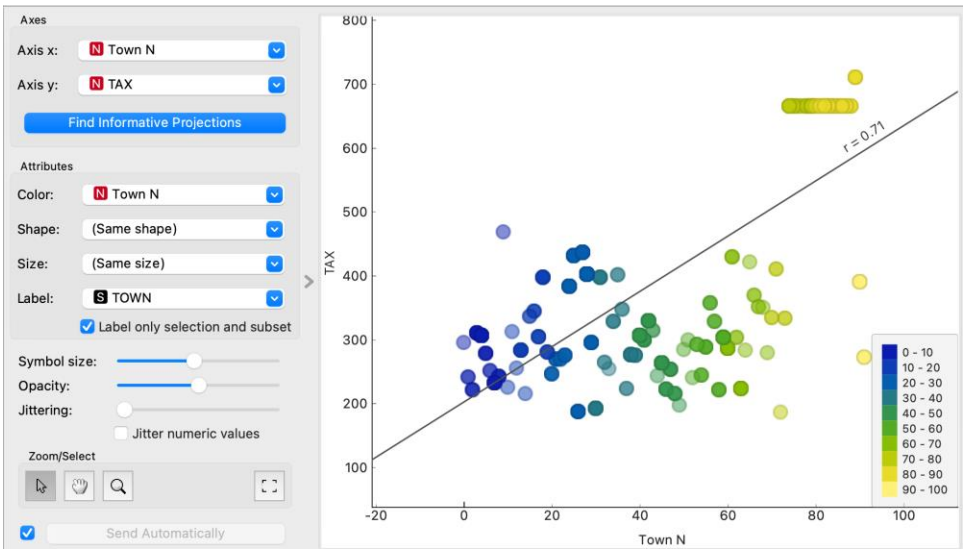


d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

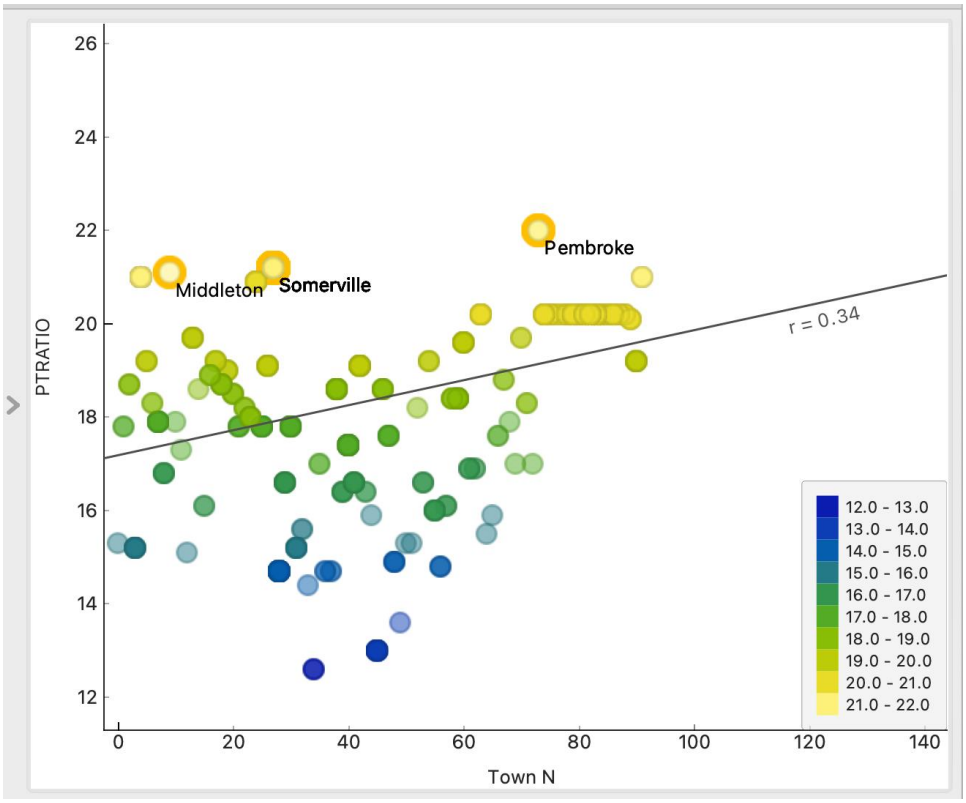
Boston CharlesTown, Boton Roxbury, Boston South Boston. 60-89 Crim



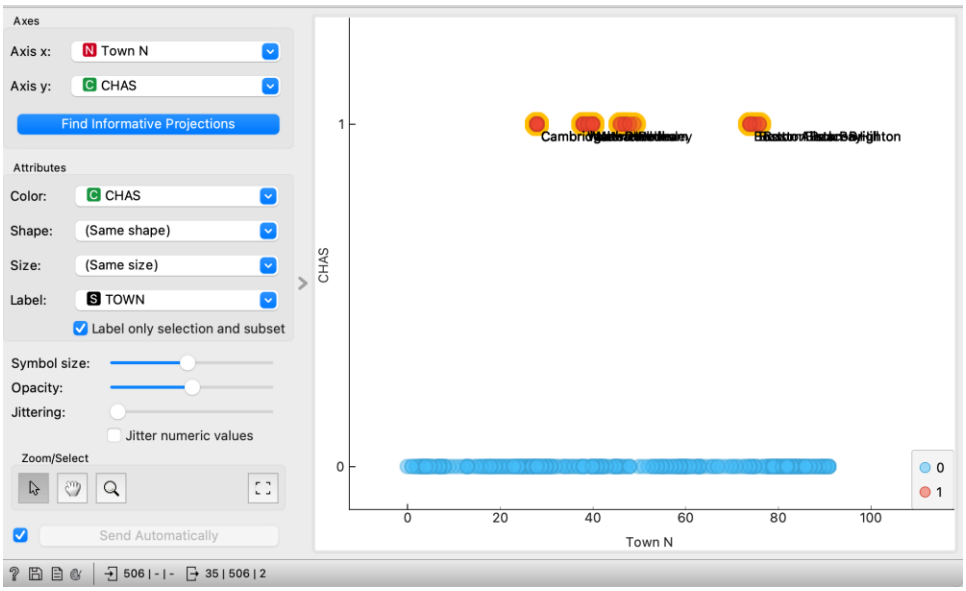
For full-value property-tax rate per USD 10,000 per town, Tax Rate, Chelsea



Pupil-teacher radio - Pembroke, Somerville, Middleston

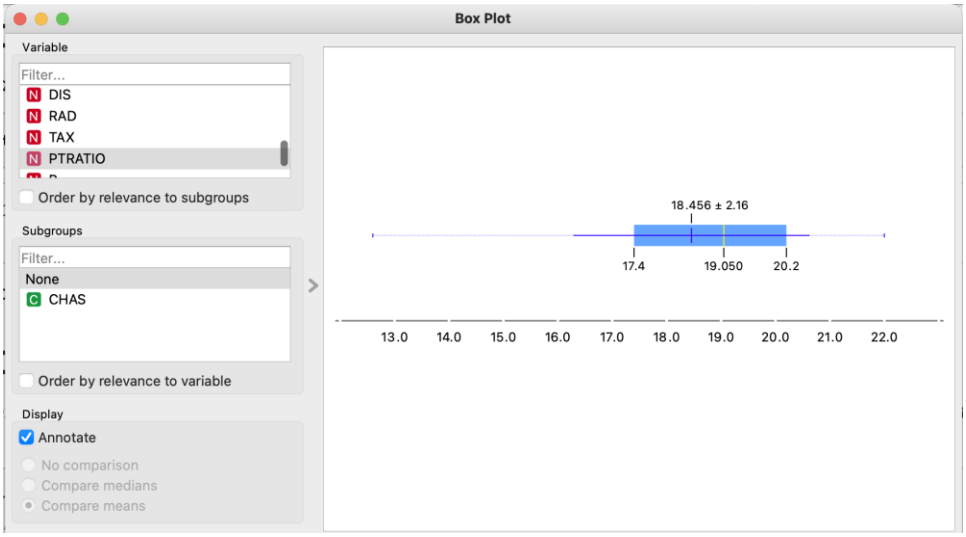


e) How many of the suburbs in this data set bound the Charles River?



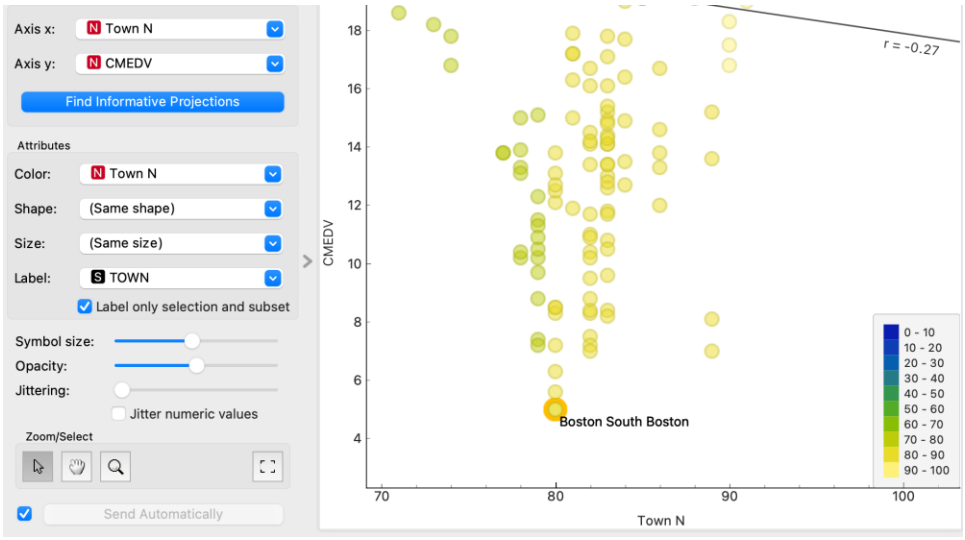
f) What is the median pupil-teacher ratio among the towns in this data set?

19.050



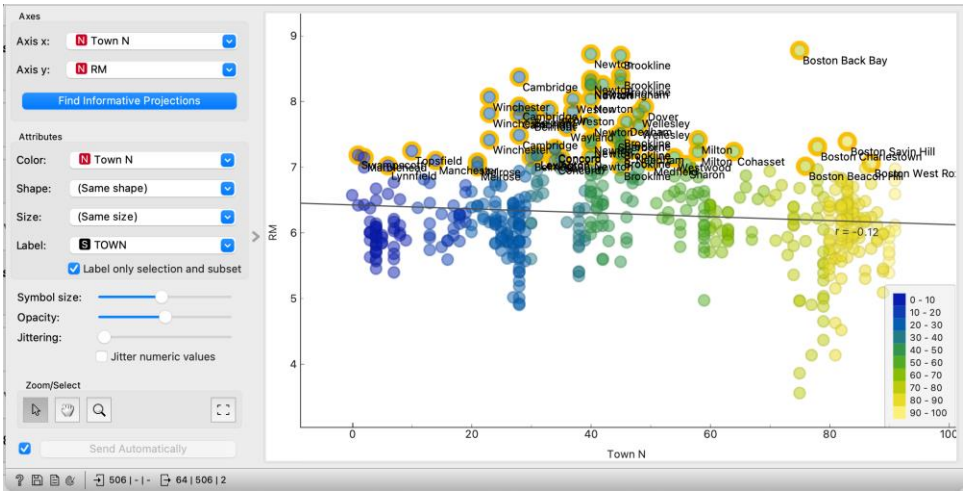
g) Which suburb of Boston has the lowest median value of owner-occupied homes?

   a. What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.



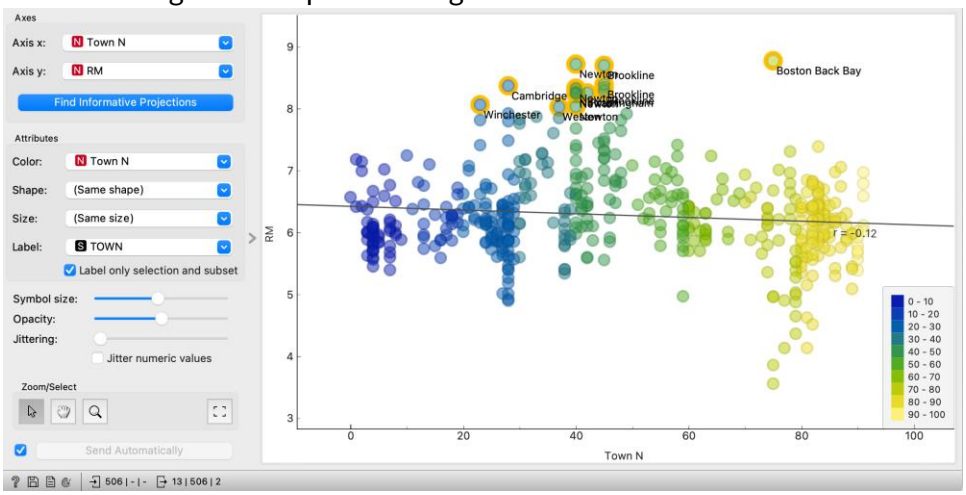
Boston South Boston, which also has a high crime-rate.

h) In this data set, how many of the suburbs average more than seven rooms per dwelling?

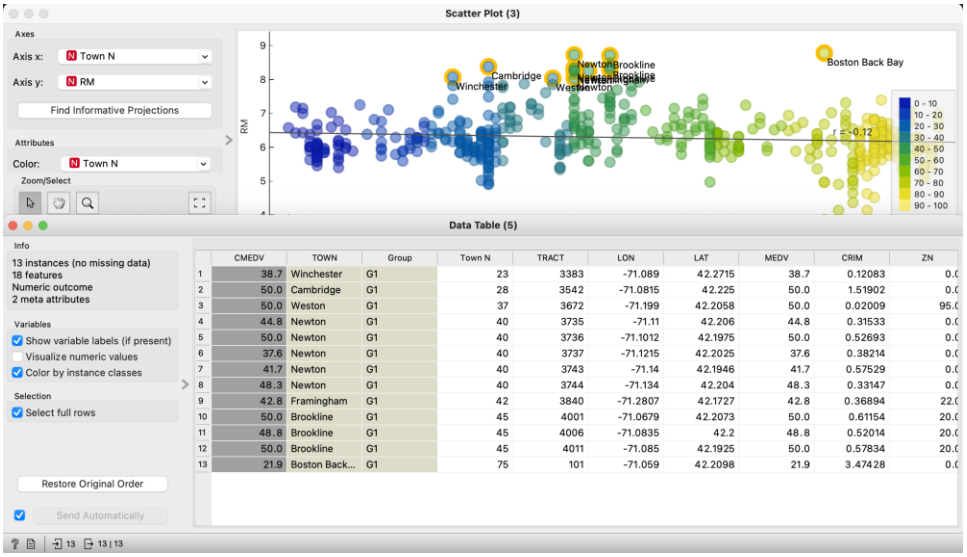


64

a. More than eight rooms per dwelling? 13



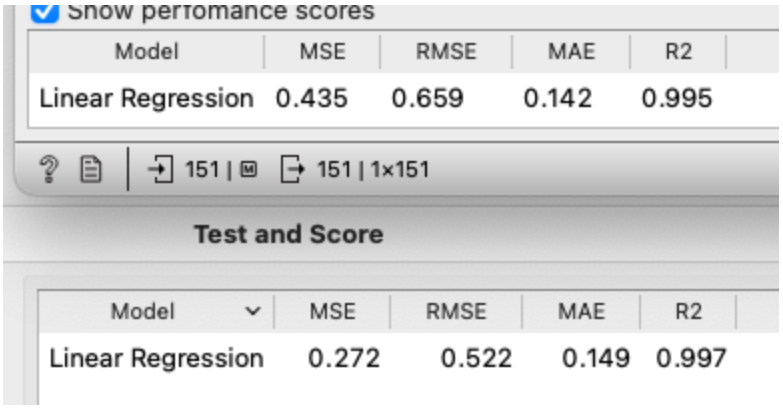
b. Comment on the suburbs that average more than eight rooms per dwelling.



	CMEDV	TOWN	Group	Town N	TRACT	LON	LAT	MEDV	CRIM	ZN
1	38.7	Winchester	G1	23	3383	-71.089	42.2715	38.7	0.12083	0.0
2	50.0	Cambridge	G1	28	3542	-71.0815	42.225	50.0	1.51902	0.0
3	50.0	Weston	G1	37	3672	-71.199	42.2058	50.0	0.02009	95.0
4	44.8	Newton	G1	40	3735	-71.11	42.206	44.8	0.31533	0.0
5	50.0	Newton	G1	40	3736	-71.1012	42.1975	50.0	0.52693	0.0
6	37.6	Newton	G1	40	3737	-71.1215	42.2025	37.6	0.38214	0.0
7	41.7	Newton	G1	40	3743	-71.14	42.1946	41.7	0.57529	0.0
8	48.3	Newton	G1	40	3744	-71.134	42.204	48.3	0.33147	0.0
9	42.8	Framingham	G1	42	3840	-71.2807	42.1727	42.8	0.36894	22.0
10	50.0	Brookline	G1	45	4001	-71.0679	42.2073	50.0	0.61154	20.0
11	48.8	Brookline	G1	45	4006	-71.0835	42.2	48.8	0.52014	20.0
12	50.0	Brookline	G1	45	4011	-71.085	42.1925	50.0	0.57834	20.0
13	21.9	Boston Back...	G1	75	101	-71.059	42.2098	21.9	3.47428	0.0

2 are near the river while the other ones are not, they don't have that much crime rate, 7 of them doesn't have proportions of residential land zone.

i) Split the data in training (70%) and test (30%) data, use the linear regression model with the training data, and test it to predict House value.

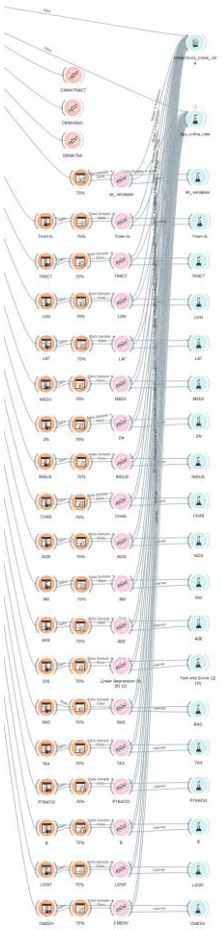


a. How well did the model perform?

Bien ya que podemos ver que tenemos un valor cercano a 1 para R2

b. How can you improve these results?

j) We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors. For each predictor, individually, fit a simple linear regression model to predict the response.



Model	MSE	RMSE	MAE	R2
all_variables	40.593	6.371	3.082	0.450
RAD	45.165	6.721	2.789	0.388
TAX	48.940	6.996	3.238	0.337
TRACT	51.978	7.210	3.610	0.296
LSTAT	58.650	7.658	3.780	0.206
Town N	59.286	7.700	4.054	0.197
NOX	60.893	7.803	3.723	0.175
INDUS	61.737	7.857	3.741	0.164
CMEDV	62.856	7.928	4.621	0.149
MEDV	62.948	7.934	4.630	0.148
B	63.060	7.941	4.200	0.146
DIS	63.318	7.957	4.412	0.142
AGE	64.878	8.055	4.436	0.121
PTRATIO	67.829	8.236	4.654	0.081
RM	70.378	8.389	4.772	0.047
ZN	70.992	8.426	4.775	0.039
LAT	73.428	8.569	4.851	0.006
LON	73.669	8.583	4.918	0.002
LON	73.669	8.583	4.918	0.002
CHAS	73.873	8.595	5.018	-0.000

a. Describe your results.

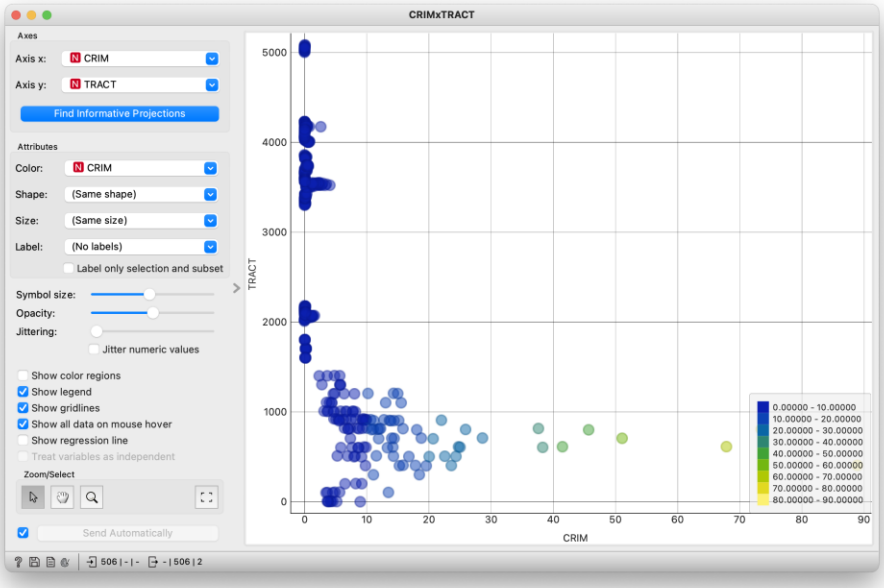
Vemos como el crimen se ve relacionado con el acceso que se tiene a las vías del tren. También como dependiendo de que vía está cerca es lo que influye el porcentaje de crimen

b. In which of the models is there a statistically significant association between the predictor and the response?

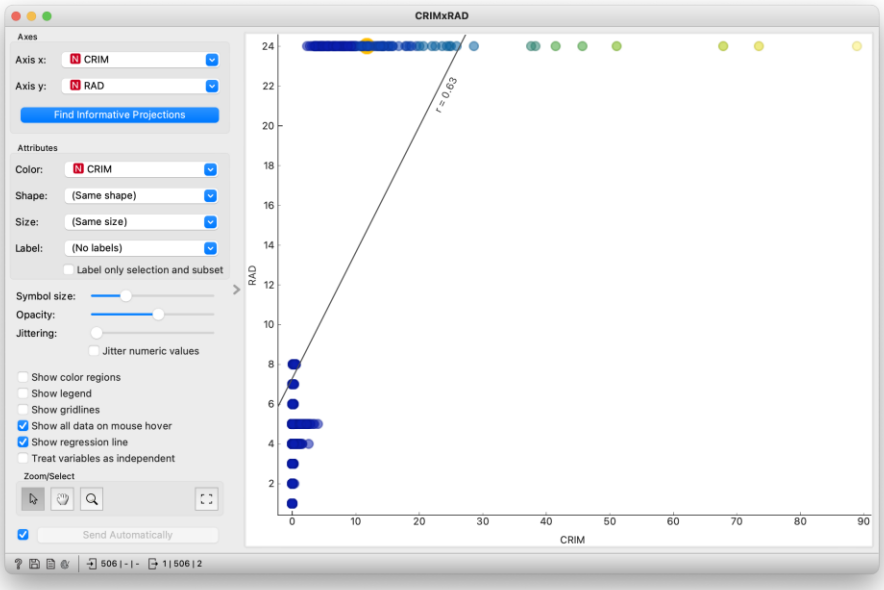
RAD que representa que tan accesibles son las vías del tren

TRACT que indica el identificador de la via de tren, lo que significa que cada vía se le puede asociar con un % de crimen en la zona

c. Create some plots to back up your assertions.

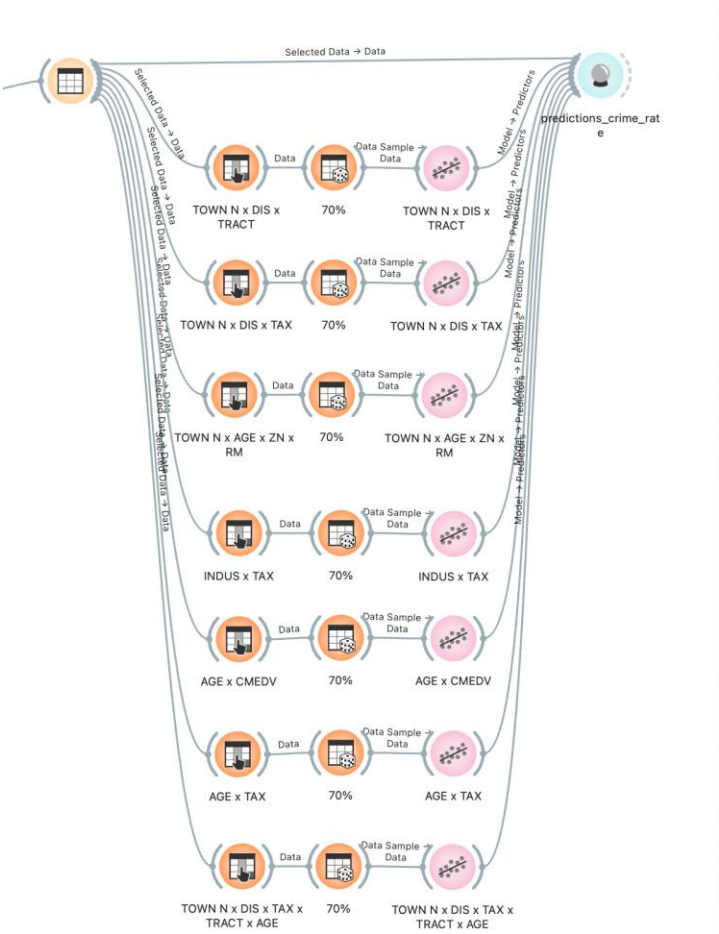


Vemos como dependiendo de que vía está cerca se tiene más porcentaje de crimen



Y vemos como si se tienen acceso a las vías igual el crimen sube

k) Fit a multiple regression model to predict the response using all the predictors.
   a. Describe your results.



Show perfomance scores

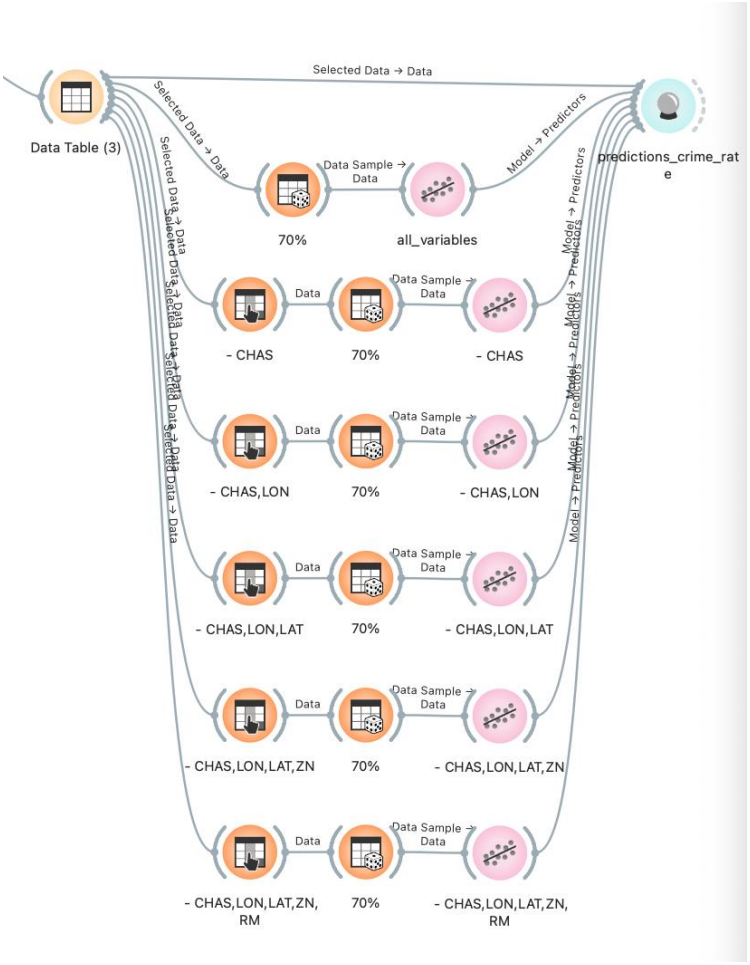
Model	MSE	RMSE	MAE	R2
TOWN N x DIS x TAX x TRACT x AGE	46.624	6.828	2.866	0.369
TOWN N x DIS x TRACT	47.598	6.899	2.891	0.355
TOWN N x DIS x TAX	48.241	6.946	3.236	0.347
AGE x TAX	48.605	6.972	3.271	0.342
INDUS x TAX	48.921	6.994	3.251	0.337
TOWN N x AGE x ZN x RM	53.447	7.311	3.811	0.276
AGE x CMEDV	59.214	7.695	4.339	0.198

b. For which predictors can we get the best results?

TOWN N x DIS x TAX x TRACT x AGE

Iniciamos por las de menos variables ya que vimos que AGE y TAX daban el mejor probamos con los resultados de cada predictor individual (TOWN N x DIS x TRACT) y tome los 3 primeros y para obtener el que mejor R2 score solo unimos esos 5 predictores

c. Which predictors can be eliminated?



Model	MSE	RMSE	MAE	R2
- CHAS	40.595	6.371	3.088	0.450
all_variables	40.597	6.372	3.086	0.450
- CHAS,LON	40.680	6.378	3.090	0.449
- CHAS,LON,LAT	40.683	6.378	3.091	0.449
- CHAS,LON,LAT,ZN	41.089	6.410	3.069	0.444
- CHAS,LON,LAT,ZN,RM	41.147	6.415	3.072	0.443

Para este nos basamos en los que menos influían de manera individual en el punto j ya que vemos no baja en gran cantidad, es solo cuando las combinamos cuando se nota que no da tan buenos resultados

l) How do your results from (j) compare to your results from (k)?

En que utilizamos la información de J para obtener mejores resultados en K. Ya que gracias a J supimos que predictors eran los que más y menos influían para saber que predictors se podían

quitar o cuales podíamos usar que influyeran más para que no variara tanto los resultados comparados con los modelos de más o menos predictores