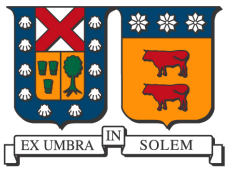

UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Boston Housing Problem: Una mirada a los factores que afectan el precio de una vivienda en Boston

Profesor:
- Felipe Osorio

Curso: MAT266

Nombre del alumno:
- Eric Zepeda

21 de julio de 2022

Índice

1. Introducción:

Boston Housing es una de las bases de datos más utilizadas para introducir al estudiante a la implementación de diferentes modelos de regresión cuando se estudia ciencias de datos. Sin embargo, para este caso el informe se limitará a analisar la información entregada para realizar un análisis estadístico.

Para esto, se utilizará Jupyter notebook. Este último, es un entorno de programación de lenguaje python, que otorga la facultad de separar el código en celdas e incorporar herramientas como latex para una mayor comprensión del trabajo que se esta haciendo. Además se contará con la ayuda de algunas librerias populares como las enunciadas a continuación:

- Numpy: Libreria que incorpora herramientas de Algebra lineal
- Pandas: Para la lectura y manipulación de la base de datos.
- Matplotlib: Gráficos para Python
- Sklearn: Incorpora herramientas de procesamiento de datos y modelos de Machine learning. Esto será útil para aplicar la regresión de forma directa.

También, cabe mencionar que la base de datos contempla 506 datos y las siguientes variables que serán de relevancia para el estudio a realizar:

- Cluster: Identificador de localidad. (Denotado como β_0)
- Crime: Índice de criminalidad per cápita por ciudad. (Denotado como β_1)
- Zone: Proporción de terrenos residenciales con zonificación para lotes de más de 25,000 pies cuadrados. (Denotado como β_2)
- indus: Proporción de acres comerciales no minoristas por ciudad. (Denotado como β_3)
- chas: Variable de carácter dummy relacionada con el río Charles (1 si el tramo limita con el río; 0 si no).(Denotado como β_4)
- noxsq: Concentración de óxido nítrico (partes por 10 millones). (Denotado como β_5)
- room: Número medio de habitaciones por vivienda. (Denotado como β_6)
- age: Proporción de unidades ocupadas por sus propietarios construidas antes de 1940.(Denotado como β_7)
- dist: Distancias ponderadas a cinco centros de empleo de Boston. (Denotado como β_8)
- rad: Índice de accesibilidad a carreteras.(Denotado como β_9)
- tax: Tipos de impuesto sobre bienes inmuebles por cada 10,000 dolares. (Denotado como β_{10})
- ptrat: Proporción de alumnos por profesor para cada ciudad.(Denotado como β_{11})
- black: $1000(Bk-0{,}63)^2$ donde Bk es la proporción de población de raza negra por ciudad. (Denotado como β_{12})
- lstat: Proporción de población de bajos ingresos. (Denotado como β_{13})

- medv: Valor mediano de las viviendas ocupadas por sus propietarios en miles de dólares. (Denotado como Y)

Con todo lo antes mencionado, se analisará el conjunto de datos de Boston Housing, que fue estudiado por Harrison y Rubinfeld en el año 1978. Donde buscaban una relación entre los precios de vivienda y la pureza del aire. Para esto se considerará un modelo de regresión de la forma:

$$Y = \beta X + \epsilon$$

Donde Y^T corresponde al vector generado por la columna medv, X la matriz de diseño a partir de la información entregada, β el vector de parámetros desconocidos, que se estimará a partir de las variables a considerar en el modelo enunciadas anteriormente y ϵ el intercepto.

2. Análisis Estadístico.

Antes de comenzar con en análisis de la información, es importante echar un vistazo a la muestra de la base de datos.

	Cluster	crime	zone	indus	chas	noxsq	room	age	dist	rad	tax	ptrat	black	lstat	medv
0	1	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	2	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	3	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

Figura 1: Datos de Boston Housing.

Claramente, cada variable es númerica, entero o flotante[1]. Usando este hecho, se utilizará la función describe para obtener un resumen de las variables.

	Cluster	crime	zone	indus	chas	noxsq	room	age	dist	rad	tax	ptrat	black	lstat	medv
count	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000
mean	48.530	3.614	11.364	11.137	0.069	0.555	6.285	68.575	3.696	9.549	408.237	18.456	356.674	12.653	22.533
std	27.569	8.602	23.322	6.860	0.254	0.116	0.703	28.149	2.000	8.707	168.537	2.165	91.295	7.141	9.197
min	1.000	0.006	0.000	0.460	0.000	0.385	3.561	2.900	0.586	1.000	187.000	12.600	0.320	1.730	5.000
25%	27.250	0.082	0.000	5.190	0.000	0.449	5.885	45.025	2.074	4.000	279.000	17.400	375.378	6.950	17.025
50%	43.000	0.257	0.000	9.690	0.000	0.538	6.208	77.500	3.107	5.000	330.000	19.050	391.440	11.360	21.200
75%	79.000	3.677	12.500	18.100	0.000	0.624	6.624	94.075	5.113	24.000	666.000	20.200	396.225	16.955	25.000
max	92.000	88.976	100.000	27.740	1.000	0.871	8.780	100.000	9.223	24.000	711.000	22.000	396.900	37.970	50.000

Figura 2: Resumen de los datos.

Este resumen consta de:

- Count: número de datos.

[1]Este hecho fácilitará mucho la implementación de un modelo más adelante.

- mean: media.
- std: Desviación Estándar.
- min: Valor minimo de dicha variable.
- $25\,\% - 50\,\% - 75\,\%$ Corresponde a los cuantiles respectivos.
- max: Valor máximo de dicha variable.

Para tener una visión más clara sobre los datos anómalos. Una herramienta que suele ser utilizada con mucha frecuencia es el gráfico de caja, donde este tipo de datos se puede observar con claridad como puntos fuera de caja.

A partir de esto, se realiza un gráfico de caja para cada variable a cosiderar.

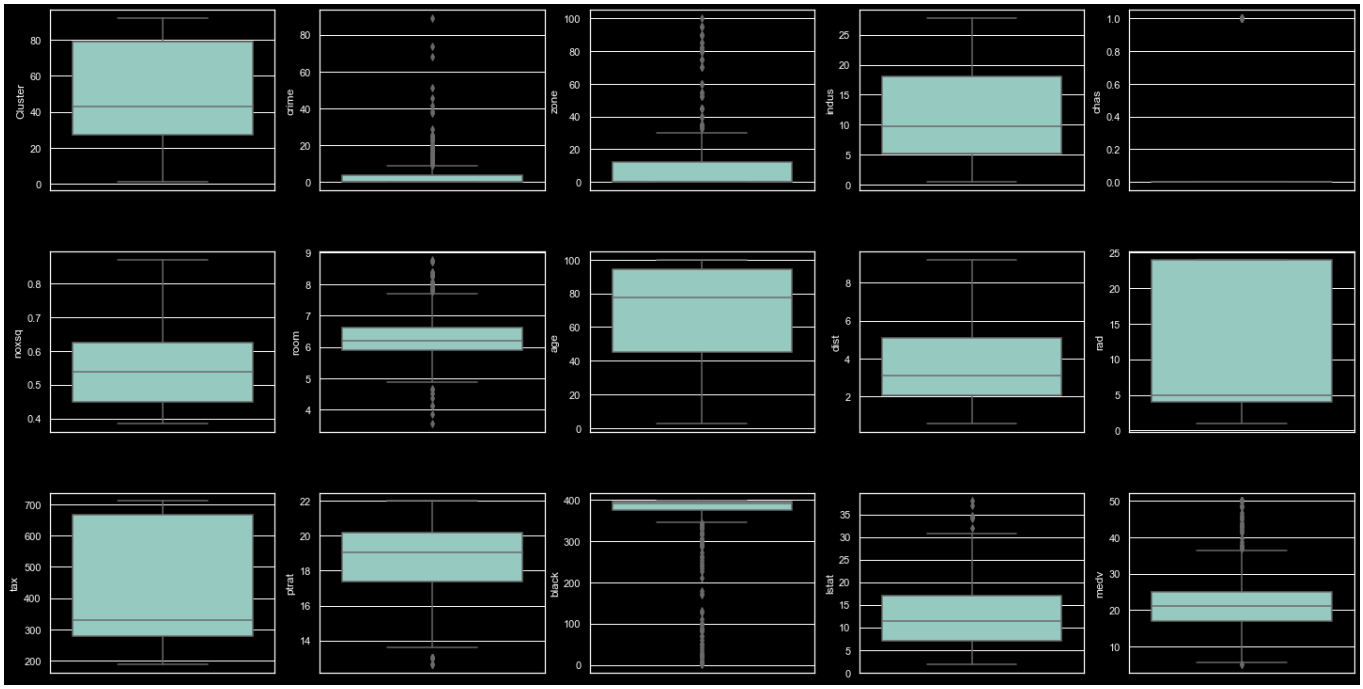

Figura 3: Gráficos de caja de cada variable.

Observe que las variables con más datos anomalos son: Crime, zone, room, black y chas. Específicamente el porcentage de outliers de cada variables es:

- Cluster:0,00 %
- Crime:13,04 %
- Zone: 13,44 %
- indus: 0,00 %
- chas: 100 %
- noxsq: 0,00 %
- room: 5,93 %
- age: 0,00 %

- rad: 0,00 %
- tax: 0,00 %
- ptrat: 2,96 %
- black: 15,22 %
- lstat: 1,38 %
- medv: 7,91 %

Con esta información, como existen variables con un alto porcentaje de datos anomalos dentro de la muestra. Es necesario implementar un mecanismo para que esa información no afecte negativamente a la eficacia del modelo a implementar. Es por ello que es importante realizar un análisis exploratorio de datos y limpieza de la información.

Existen multiples alternativas para mejorar los resultados de un modelo. Para contrastar esto, se considerará el modelo de regresión lineal y dos conjuntos de datos, que denominaremos raw y clean. Raw será la información sin limpiar y clean los datos con menos outliers.

Para esto existen varias alternativas para limpiar los datos. Desde eliminar todas las entradas con outliers reduciendo la cantidad de datos abruptamente, hasta sustituir los datos de las variables conflictivas con la media para no entorpecer el proceso. Se hará esto último con fines experimentales.

2.1. Resultados de datos Raw

Para esta parte, se consideró el dataframe Raw obteniendo la siguiente matriz de correlación:

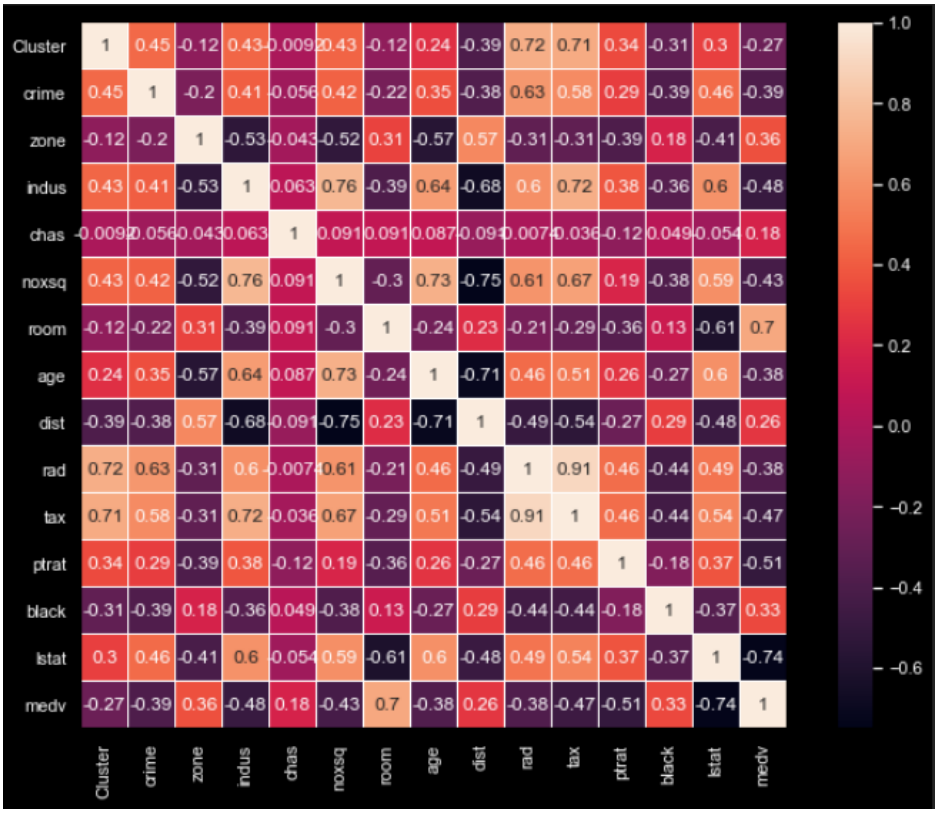

Figura 4: Matriz de correlación para datos Raw

De aquí se pueden reconocer 3 variables que tienen una alta influencia en medv por su alta correlación. Estos son: room, pstrat, lstat. Gráficamente, se relacionan con medv de la siguiente manera:

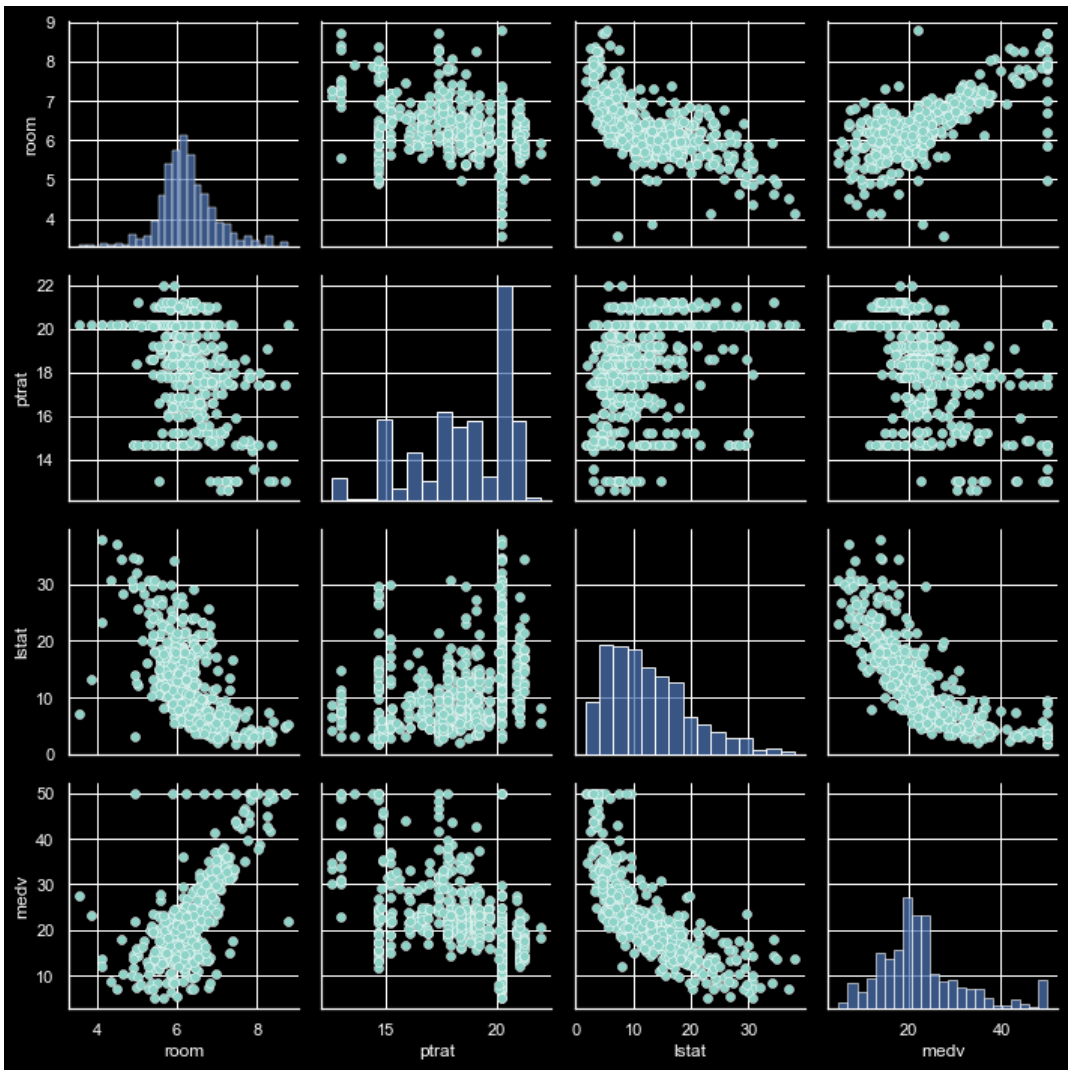

Figura 5: Diagrama de a pares para estudiar como se relacionan las variables altamente correlacionadas con medv.

Observe aquí que medv se distribuye normal a pesar de tener outliers. Además, existe una clara relación con medv para las variables room y lstat. Donde room se relaciona de manera directa con el precio de la vivienda medv, mientras que lstat de manera inversa.

Para finalizar, los datos obtenidos por el modelo se discutirán posteriormente.

2.2. Resultado de datos Clean

En esta parte, se cambian los datos anomalos por la media. Para esto, se manipularon los datos antes obtenidos con el fin de contrastar los resultados previamente obtenidos en la parte anterior, con el fin de ver qué tan sensibles pueden resultar este tipo de modelos dada una muestra inicial. Para comenzar, se adjunta la matriz de correlación obtenida:

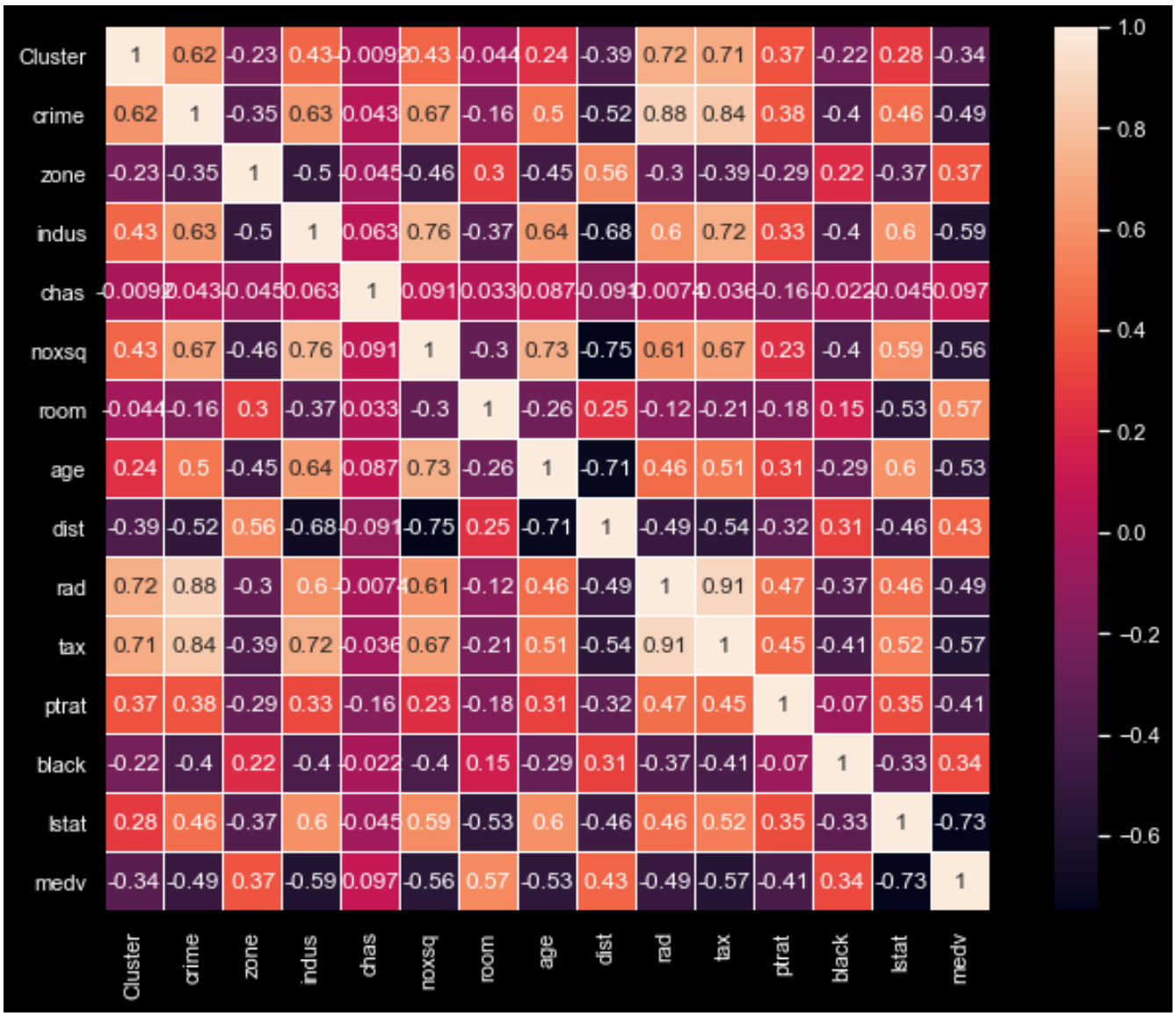

Figura 6: Correlación obtenida tras cambiar los datos anómalos por la media.

Observe que las correlaciones cambiaron bastante con respecto a la otra versión de los datos. La mayor parte de las correlaciones que se tienen con las variables y medv cambiaron en la medida que tenian mayor proporción de datos anómalos haciendo relevante nuevas variables como noxsq, age, tax y reduciendo otras, como room y chas. Esto claramente podría afectar al gráfico de datos pareados y en partícular a la regresión que se podría realizar en el problema. Es por esto que también se adjunta dicho gráfico con los datos limpiados.

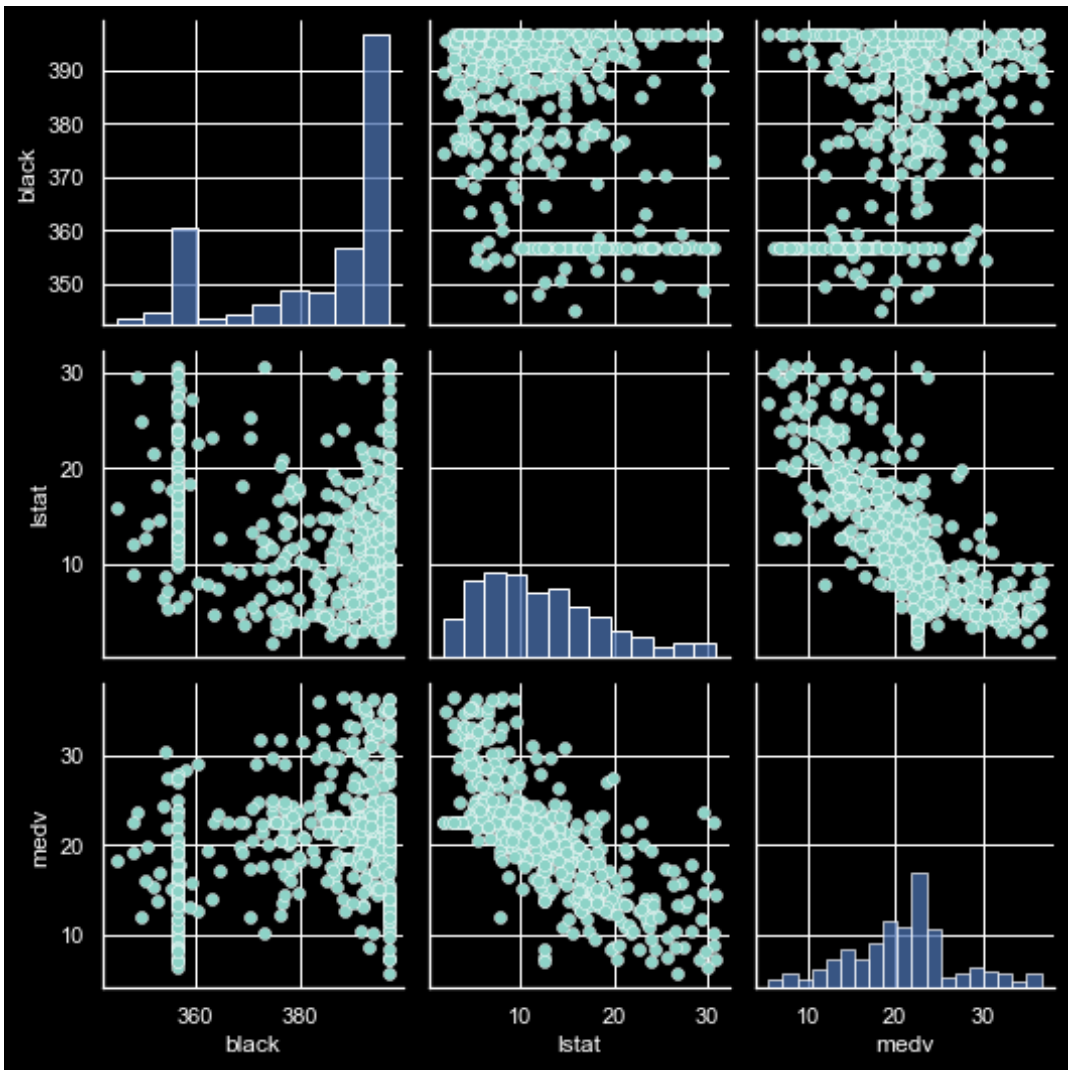

Figura 7: Datos pareados de las variables consideradas

Los datos en este gráfico se volvieron más dispersos con respecto a medv. No obstante, esto sumo peso a las otras variables por lo que el modelo puede recopilar más información al respecto. Además, cabe mencionar que medv sigue teniendo una forma similiar al de la distribución normal con una curtosis un poco más orientada hacia la derecha, pero puede explicarse porque la media supera la mediana en una unidad aproximadamente.

2.3. Resultados del modelo.

Finalmente, se hace un contraste de los resultados obtenidos al utilizar un modelo de regresión lineal para ver si la limpieza de datos tiene alguna repercusión en la precisión del modelo.

Datos usados	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	ϵ	Test Accuracy
Raw	−0,03	−0,13	0,02	0,07	3,13	−14,8	4,36	−0,01	−1,2	0,29	−0,01	−0,98	0,01	−0,54	29.98	0.69
Clean	−0,01	−0,08	0,02	−0,04	13,41	−0,13	3,31	−0,03	−0,32	0,02	−0,01	−0,31	0,02	−0,37	10.28	0,61

A partir de esto, es claro que para cada caso, se tiene una baja precisión. Esto debido a la baja correlación y el hecho de que la información cuenta con una cantidad de datos anomalos por sobre el 5 % en promedio. Es por esto que el valor de una vivienda no se puede correlacionar con la muestra recibida a la calidad del aire.

3. Conclusión

Los resultados parecieran ser pesimistas para el caso de los datos que han sido limpiados. No obstante, parece ser muestra de lo sensible que pueden ser estos modelos y que no existe una relación clara entre el precio de una vivienda y la calidad del aire por la baja correlación que hay entre estas.

4. Bibliografia

Referencias

[1] Yalin Başţanlar y Mustafa Özuysal. "Introduction to machine learning". En: *miRNomics: MicroRNA biology and computational analysis* (2014), págs. 105-128.

[2] David Harrison Jr y Daniel L Rubinfeld. "Hedonic housing prices and the demand for clean air". En: *Journal of environmental economics and management* 5.1 (1978), págs. 81-102.

[3] Felipe Osorio. "Notas de Clase: Análisis de Regresión". En: ().

[4] Jake VanderPlas. *Python data science handbook: Essential tools for working with data*. .ºReilly Media, Inc.", 2016.