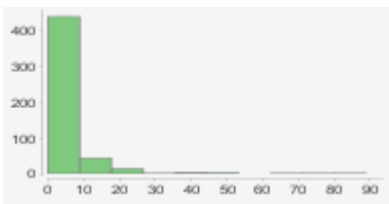
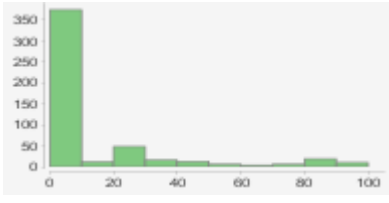
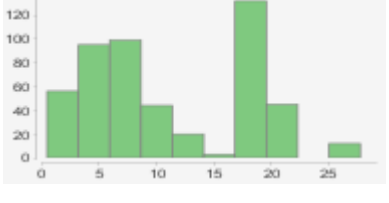

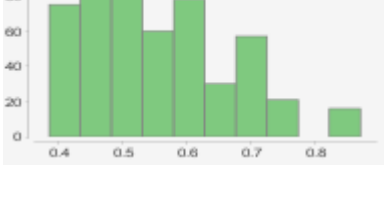
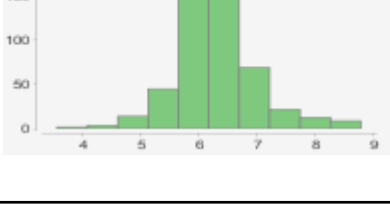


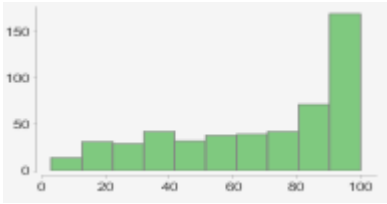
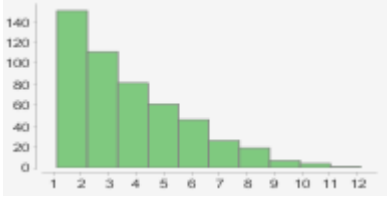
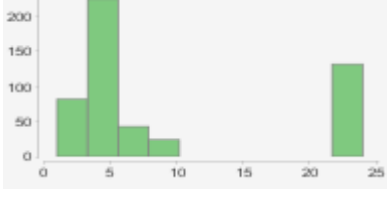
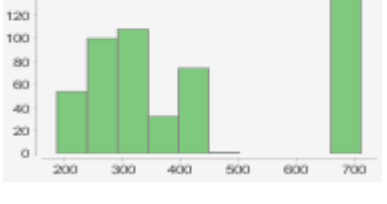
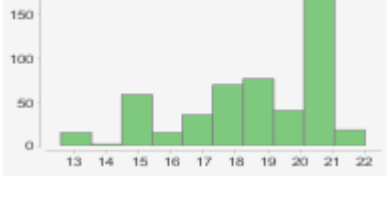
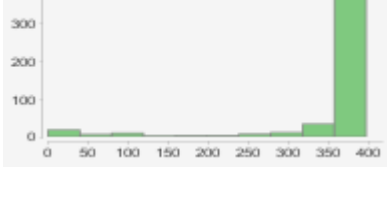
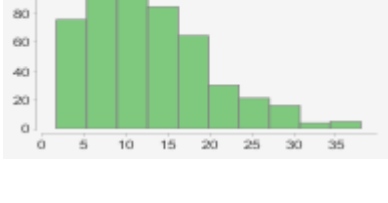
# Ejercicio 1

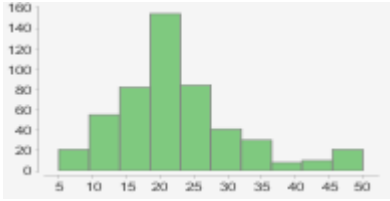
El dataset tratado tiene el propósito de entrenar un modelo de IA para predecir la mediana de precios de casas en ciertas ciudades. Los atributos con los que se cuentan para esto son los siguientes:

- **CRIM**: Tasa de crimen per cápita de la ciudad (números reales, de 0,006 a 88,976).
- **ZN**: Proporción de terreno residencial zonificado para lotes sobre 25000 pies cuadrados (números reales, de 0 a 100).
- **INDUS**: Proporción de acres de negocios minoristas en la ciudad (números reales, de 0,460 a 27,740).
- **CHAS**: Variable dummy Charles River (categórico, vale 1.0 si la ciudad limita con el río Charles River, 0.0 si no).
- **NOX**: Concentración de óxidos nítricos, en partes por 10 millones (números reales, de 0,385 a 0,871).
- **RM**: Número medio de habitantes por vivienda (números reales, de 3,561 a 8,780).
- **AGE**: Proporción de unidades ocupadas por sus propietarios, que fueron construidas antes de 1940 (números reales, de 2,900 a 100,000).
- **DIS**: Distancias ponderadas a cinco centros de empleo de Boston (números reales, 1,130 a 12,127).
- **RAD**: Índice de accesibilidad a carreteras radiales (números reales, de 1 a 24).
- **TAX**: Tasa de impuesto a la propiedad de valor total por \$10000 (números reales, de 187 a 711).
- **PTRATIO**: Proporción entre cantidad de alumnos y de maestros de la ciudad (números reales, de 12,6 a 22).
- **B**: Representa un atributo no presente directamente en el dataset llamado Bk, que es la proporción de gente negra sobre el total en la ciudad. Más concretamente, se calcula como  $1000 * (Bk - 0.63) ^ 2$  (números reales, de 0,320 a 396,900).
- **LSTAT**: Proporción de gente de estatus bajo sobre el total (números reales, de 1.730 a 37.970).
- **MEDV**: Variable de salida. Representa la mediana de valores de hogares ocupados por sus dueños, en miles de dólares (números reales, de 5 a 50).

Para que se tenga una mejor idea de los tipos de variables con los que se tratan, se presentan sus distribuciones y presencias de outliers.

Atributo	Imagen de la distribución	Distribución	Outliers
CRIM		Alta concentración en el rango 0-10, pocos ejemplos en el resto del rango	No se notan outliers
ZN		Alta concentración en el rango 0-10, pocos ejemplos en el resto del rango	No se notan outliers
INDUS		No se nota ninguna distribución particular	No se notan outliers
CHAS		Dato binomial	No se notan outliers
NOX		No se nota ninguna distribución particular, más allá de una tendencia decreciente	No se notan outliers
RM		Se asemeja a una distribución gaussiana, con media alrededor del 6 o 7	No se notan outliers

AGE		Distribución creciente, con una parte importante de los ejemplos en el rango 80-100	No se notan outliers
DIS		Distribución decreciente	No se notan outliers
RAD		Dos cúmulos de datos, uno en el rango 0-10 y otro en el rango 20-24	No se notan outliers
TAX		Dos cúmulos de datos, uno en el rango 187-500 y otro en el rango 650-700	No se notan outliers
PTRARIO		No se nota ninguna distribución particular, más allá de un pico en el rango 20-21	No se notan outliers
B		Alta concentración en el rango 350-400, pocos ejemplos en el resto del rango	No se notan outliers
LSTAT		No se nota ninguna distribución particular, más allá de una tendencia decreciente	No se notan outliers

MEDV	 <p>A histogram showing the frequency of MEDV values. The x-axis ranges from 5 to 50 in increments of 5. The y-axis ranges from 0 to 160 in increments of 20. The distribution is roughly bell-shaped, peaking at the 20-25 bin with a frequency of approximately 150. There is a slight increase in frequency at the 45-50 bin, reaching about 20.</p>	Distribución ligeramente similar a una distribución gaussiana, más un pico ligero en el rango 45-50	No se notan outliers
------	--	---	----------------------

## Ejercicio 2

Para este problema, decidimos utilizar un modelo de regresión lineal, el cual posee varios hiper-parámetros, entre los cuales se encuentran:

- *Feature selection*: Determina el algoritmo de selección de atributos, el cual filtra los atributos que participarán en la regresión lineal.
- *Eliminate collinear features*: Determina si, antes del entrenamiento, se deben eliminar atributos que tengan relación colineal entre sí.
- *Use bias*: Determina si se debería calcular el coeficiente independiente para el modelo, o si en su lugar se fija en 0.

## Ejercicio 3

El modelo resultante queda así (se decidió no utilizar *feature selection*):

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code ↑
INDUS	-0.814	2.362	-0.022	0.669	-0.345	0.731	
AGE	-0.197	1.685	-0.006	0.788	-0.117	0.907	
ZN	4.769	1.886	0.120	0.885	2.529	0.012	**
TAX	-6.254	2.693	-0.219	0.729	-2.323	0.021	**
CHAS	3.294	1.120	0.094	0.989	2.940	0.004	***
CRIM	-12.494	3.899	-0.131	0.863	-3.205	0.001	***
NOX	-7.589	2.514	-0.196	0.796	-3.019	0.003	***
B	4.155	1.510	0.097	0.909	2.751	0.006	***
RM	17.427	2.984	0.245	0.598	5.840	0.000	****
DIS	-16.498	2.981	-0.350	0.834	-5.535	0.000	****
RAD	8.760	2.054	0.359	0.763	4.266	0.000	****
PTRATIO	-7.626	1.699	-0.187	0.791	-4.488	0.000	****
LSTAT	-21.104	2.369	-0.461	0.491	-8.907	0	****
(Intercept)	27.109	3.239	?	?	8.370	0.000	****

## LinearRegression

```
3.294 * CHAS
- 12.494 * CRIM
+ 4.769 * ZN
- 0.814 * INDUS
- 7.589 * NOX
+ 17.427 * RM
- 0.197 * AGE
- 16.498 * DIS
+ 8.760 * RAD
- 6.254 * TAX
- 7.626 * PTRATIO
+ 4.155 * B
- 21.104 * LSTAT
+ 27.109
```

Como se puede ver, el modelo considera los atributos INDUS y AGE como poco significativos, por lo que al cambiar a una estrategia de *feature selection* de tipo *greedy*, resulta en el siguiente modelo:

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code ↑
ZN	4.835	1.870	0.121	0.886	2.585	0.010	**
CHAS	3.251	1.110	0.093	0.988	2.929	0.004	***
CRIM	-12.435	3.883	-0.131	0.862	-3.202	0.002	***
TAX	-6.654	2.424	-0.233	0.737	-2.745	0.006	***
B	4.134	1.499	0.097	0.909	2.757	0.006	***
NOX	-7.936	2.281	-0.205	0.808	-3.480	0.001	****
RM	17.496	2.938	0.246	0.603	5.955	0.000	****
DIS	-16.149	2.744	-0.343	0.820	-5.886	0.000	****
RAD	8.980	1.953	0.368	0.755	4.597	0.000	****
PTRATIO	-7.761	1.649	-0.190	0.798	-4.706	0.000	****
LSTAT	-21.238	2.261	-0.464	0.514	-9.394	0	****
(Intercept)	26.881	3.151	?	?	8.531	0.000	****

## LinearRegression

```

3.251 * CHAS
- 12.435 * CRIM
+ 4.835 * ZN
- 7.936 * NOX
+ 17.496 * RM
- 16.149 * DIS
+ 8.980 * RAD
- 6.654 * TAX
- 7.761 * PTRATIO
+ 4.134 * B
- 21.238 * LSTAT
+ 26.881

```

En lo que respecta a la performance de ambos modelos (con y sin *feature selection*), estos son los resultados:

Propiedad de la performance	Valor para el modelo sin <i>feature selection</i>	Valor para el modelo con <i>feature selection</i>
<i>Squared error</i>	17.991 +/- 35.884	17.878 +/- 35.496
<i>Correlation</i>	0.884	0.885
<i>Squared correlation</i>	0.781	0.783

## Ejercicio 4

Tras evaluar ambos modelos, se obtienen estadísticas acerca de los errores que tienen al predecir datos no vistos antes:

Modelo	Distribución del error	Notas
Sin <i>feature selection</i>		Rango: de -13.384 a 6.770
Con <i>feature selection</i>		Rango: de -13.353 a 6.763

Por un lado, no se nota una mejora clara en añadir *feature selection* al proceso. Ambos extremos del rango se acercaron un poco al 0, pero eso no implica que el modelo haya mejorado.

Por otro lado, se puede ver que ambos modelos fallan en, como máximo, 14000 dólares según la evaluación. Este es un nivel de error que, en un contexto donde la mayoría de casas cuestan, como mucho, 50000 dólares, puede considerarse importante.