

**Reconocimiento de patrones**

**Tarea 3**

[Limpieza de datos I](https://github.com/rsotoc/pattern-recognition/blob/master/Limpieza%20de%20datos%20I.ipynb)l

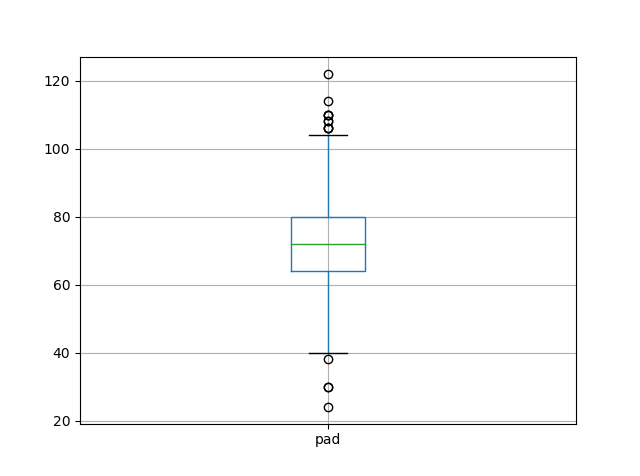
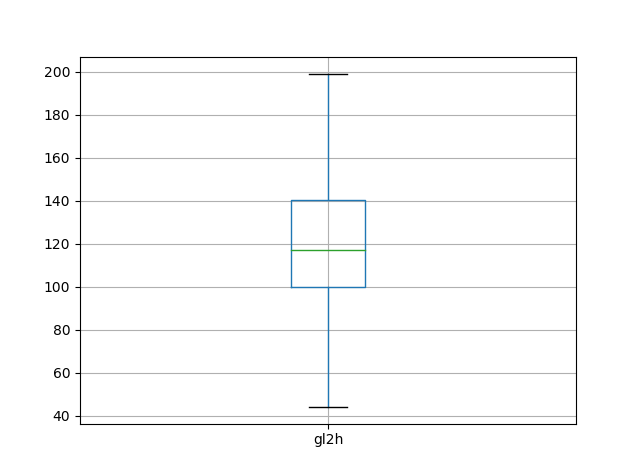
**Alumno:**

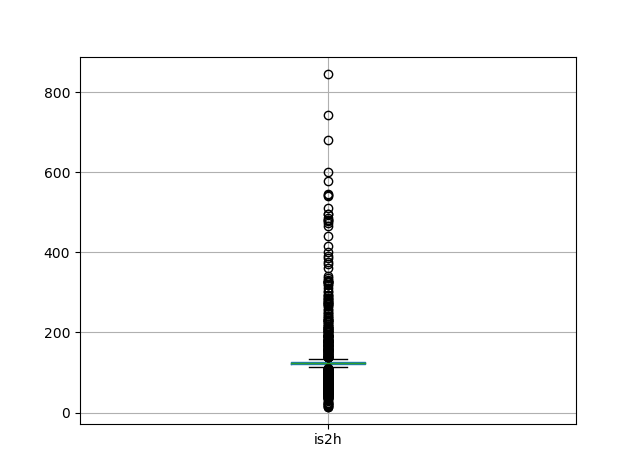
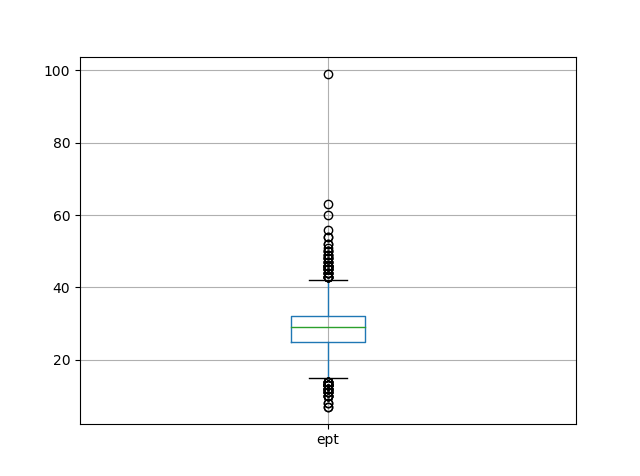
Pérez Rodríguez Raúl Francisco

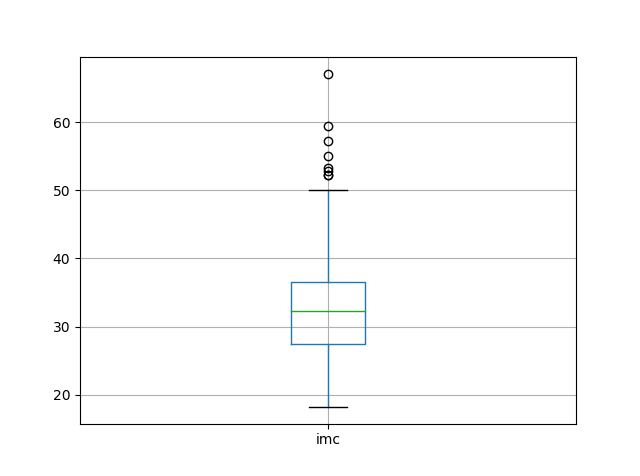
**Octubre 2017**

**Analice los problemas de valores atípicos en el conjunto de datos Pima Indians Diabetes completo.**

**Diagramas de caja**

****

****

****

Realizando una imputación con la mediana de las cinco variables que presentan valores faltantes y analizando el diagrama de caja de cada uno. Solo la variable 'gl2h' no presenta valores atipicos, 'imc' y 'pad' presentan pocos valores atípicos, pero 'ept' como 'is2h' presenta una gran cantidad de valores atípicos.

**Analice los problemas de valores faltantes y valores atípicos en los datos del ejemplo 2 (Rendimiento de combustible) y del ejemplo 3 (Taxonomía de flores).**

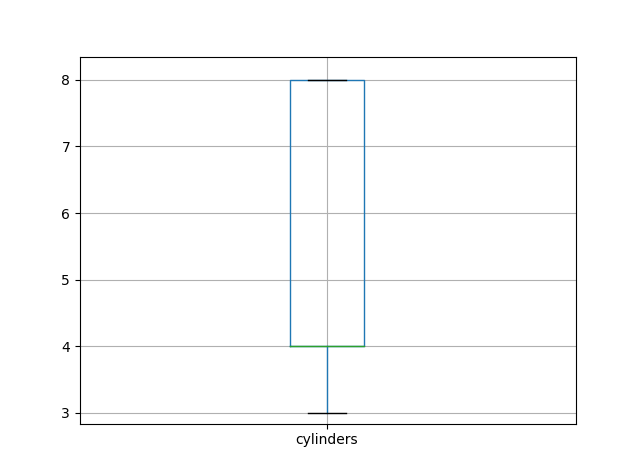
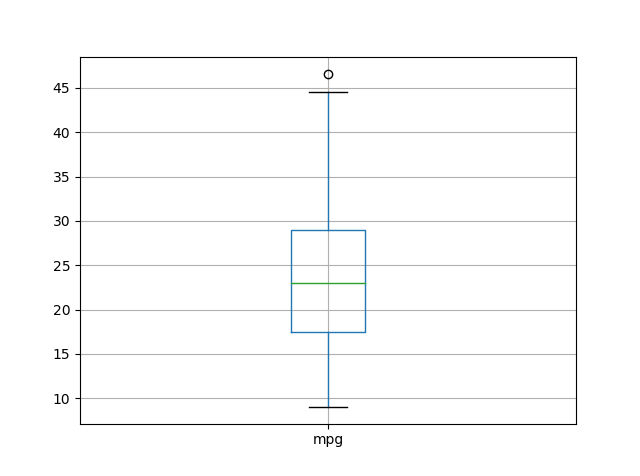
**Rendimiento de combustible**

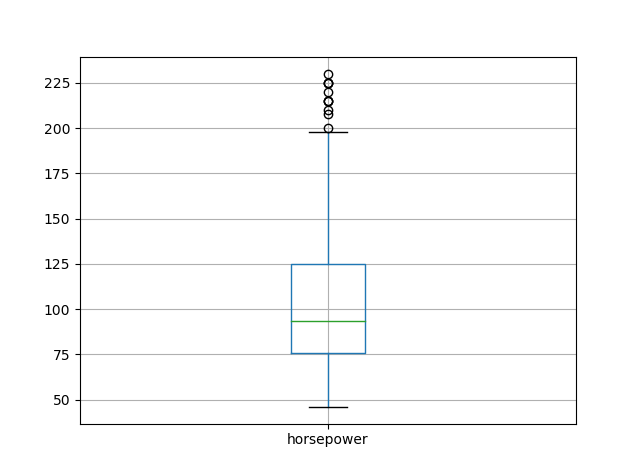
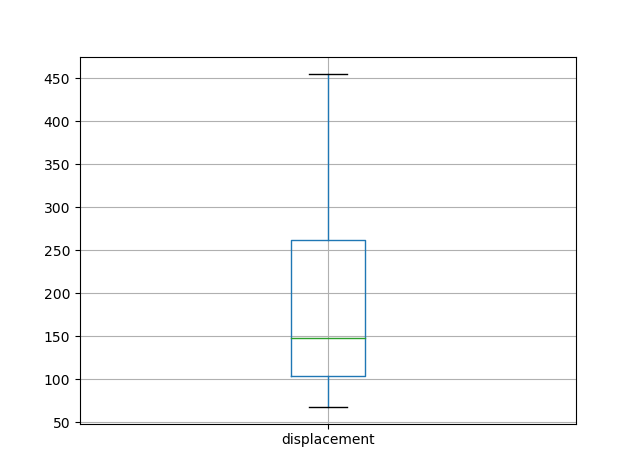
**Porcentaje de datos nulos por columna**

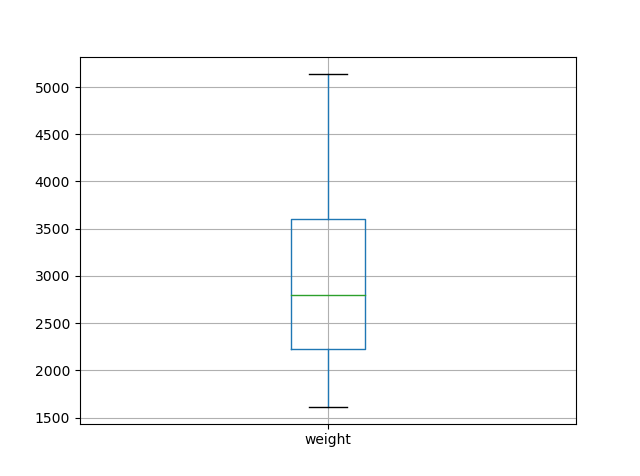
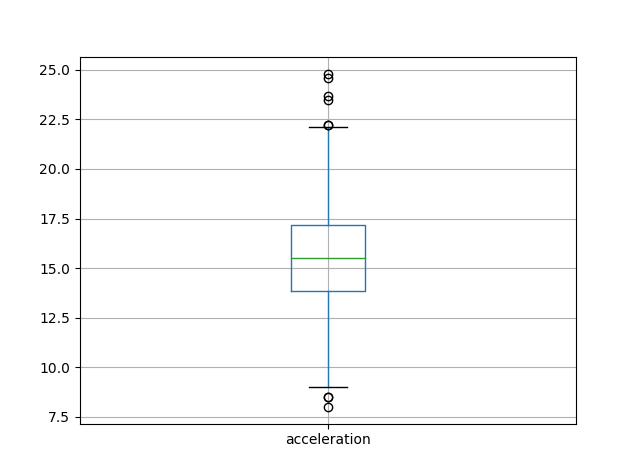
|  |  |
| --- | --- |
| **variable** | **porcentaje** |
| mpg | 0.00 % |
| cylinders | 0.00 % |
| displacement | 0.00 % |
| horsepower | 1.50 % |
| weight | 0.00 % |
| acceleration | 0.00 % |
| model-year | 0.00 % |
| origin | 0.00 % |
| car-name | 0.00 % |

Solo la variable “horsepower” presenta valores faltantes.

**Diagrama de caja**

****

****

****

En cuanto a los valores atípicos, las variables 'cylinders', 'displacement' y 'weight' no presentaron valores atípicos. 'acceleration', 'horsepower' y 'mpg' presentan valores atípicos, 'mpg' presenta un valor atípico, 'horsepower' y 'acceleration' presentan varios valores atípicos.

**Taxonomía de flores**

**Porcentaje de datos nulos por columna**

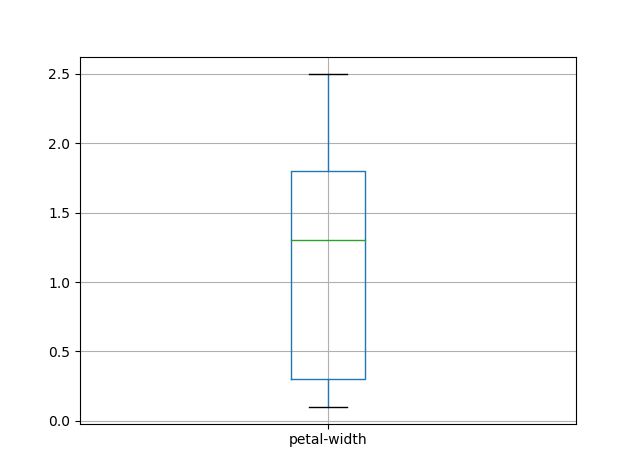
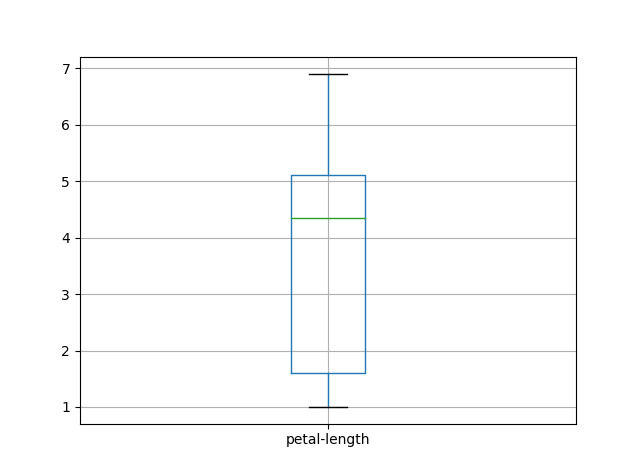
|  |  |
| --- | --- |
| **variable** | **porcentaje** |
| sepal-length | 0.00 % |
| sepal-width | 0.00 % |
| petal-length | 0.00 % |
| petal-width | 0.00 % |
| class 0.0 | 0.00 % |

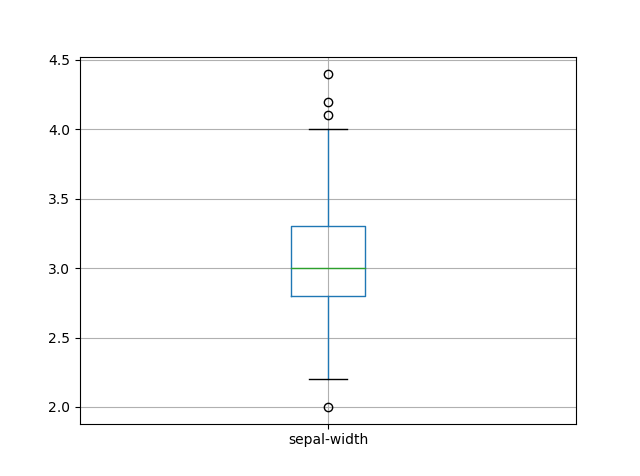
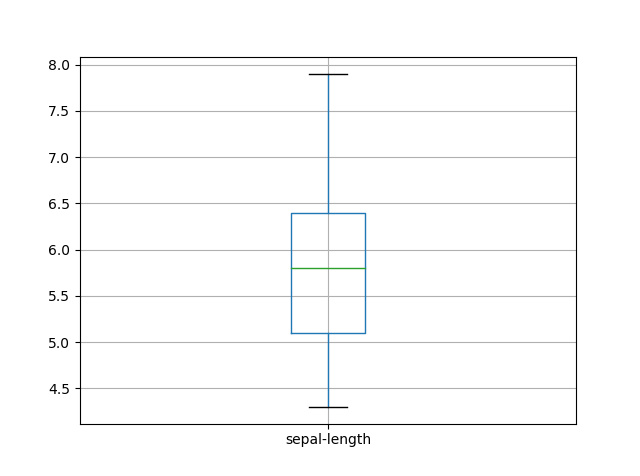
**Descripción de los datos**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **type** | **sepal-length** | **sepal-width** | **petal-length** | **petal-width** |
| **count** | 150.000000 | 150.000000 | 150.000000 | 150.000000 |
| **mean** | 5.843333 | 3.054000 | 3.758667 | 1.198667 |
| **std** | 0.828066 | 0.433594 | 1.764420 | 0.763161 |
| **min** | 4.300000 | 2.000000 | 1.000000 | 0.100000 |
| **25%** | 5.100000 | 2.800000 | 1.600000 | 0.300000 |
| **50%** | 5.800000 | 3.000000 | 4.350000 | 1.300000 |
| **75%** | 6.400000 | 3.300000 | 5.100000 | 1.800000 |
| **max** | 7.900000 | 4.400000 | 6.900000 | 2.500000 |

No hay variables que presenten valores faltantes.

**Diagramas de caja**

****



Únicamente la variable 'sepal-width' tiene valores atípicos con cuatro en total.