TIPOLOGIA I CICLE DE VIDA DE LES DADES

PAC2: Introducció a la neteja i anàlisi de dades

Marc Cervera Rosell

Semestre: febrer 2025 - juny 2025

Màster en ciència de dades

Índex

Exercici 1 1
Pregunta 1 1
Pregunta 2 1
Exercici 2 3
Pregunta 1 3
Pregunta 2 3
Exercici 3 3
Pregunta 1 3
Pregunta 2 3
Exercici 4 3
Pregunta 1 3
Pregunta 2 3
Exercici 5 3
Pregunta 1 3

Índex de figures

1 Exemple outlier diagrama de caixa 2
2 Exemple outlier IQR 2
3 Exemple outlier Z-scoring 3

Exercici 1

Pregunta 1

La reducció de la dimensionalitat consisteix a disminuir el nombre de paràmetres d'un *dataset* mantenint aquella informació més important. Aquests mètodes es poden dividir en paramètrics i no paramètrics. Els primers estimen les dades amb un model i els segons no treballen amb models sinó que treballen directament sobre les dades que hi ha disponibles. Un exemple de reducció de la dimensionalitat seria aplicar la tècnica PCA (*Principal Component Analysis*) en un sistema de reconeixement facial. Només es conservarien les característiques clau que permeten la identificació.
La reducció de la quantitat de dades consisteix a retallar la volumetria original de les dades per una volumetria menor, però que "expliqui"el mateix que les dades originals. És a dir, la reducció de quantitat de dades consisteix a agafar un subconjunt més petit de dades del *dataset* original. Un exemple de reducció de dades podria ser el resultat de la següent consulta SQL:

```
SELECT TOP 10000 * FROM [schema].[taula]
```

Aquesta consulta retorna les 10000 primeres files d'una taula que pot tenir milions de registres.

Pregunta 2

La primera tècnica de detecció de valors atípics és mitjançant el mètode del rang interquartílic (IQR). El rang interquartílic és la diferència entre el tercer quartil (Q3) i el primer (Q1). Per tant, tot valor fora del rang [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] serà considerat un *outlier*.
El segon mètode és mitjançant els diagrames de caixa. Els diagrames de caixa mostren una representació gràfica de la distribució de les dades. Els valors atípics en aquest cas, són tots aquells punts que queden fora dels bigotis de la caixa.
El tercer mètode és la puntuació Z. La puntuació Z és una mesura estadística que indica quantes desviacions estàndard es troba un valor respecte a la mitjana de les dades. Amb la puntuació Z de cada valor calculada (funció *zscore* en Python) es compara aquesta (en valor absolut) i si és major a 3 es considera un valor atípic.

Exemples:

Diagrama de caixa

IQR

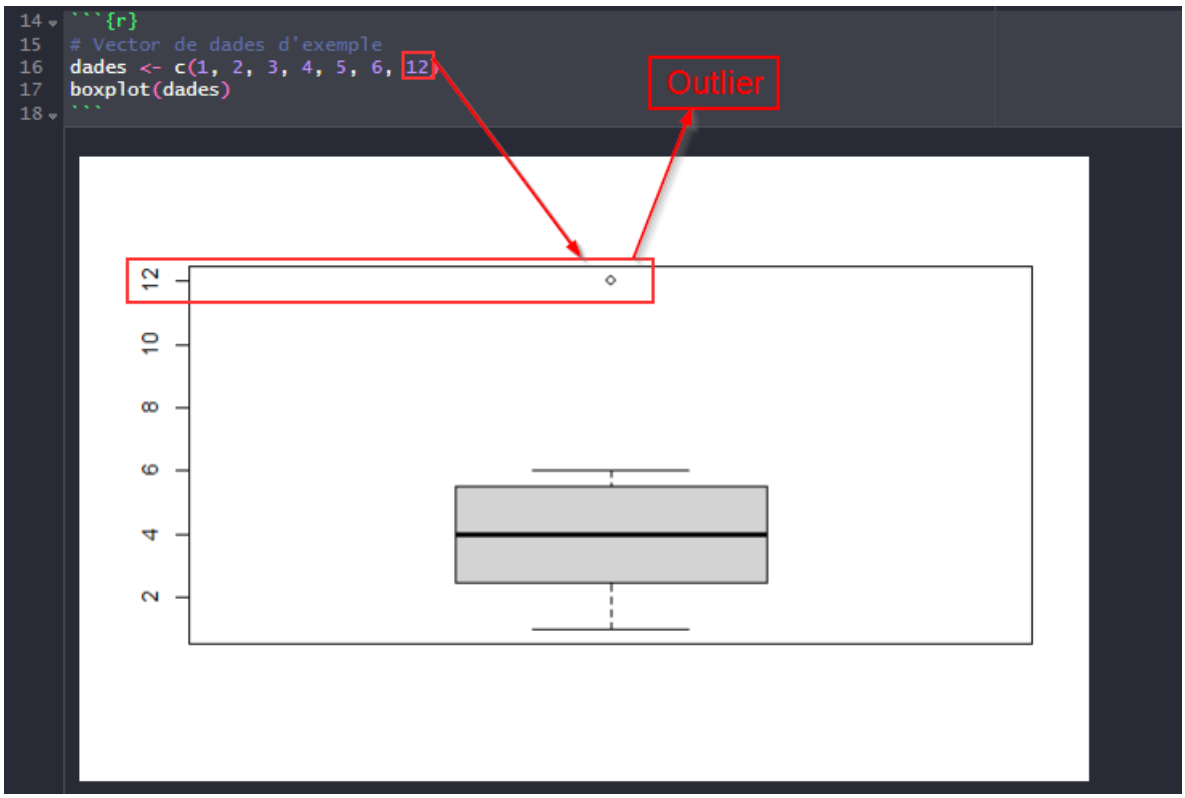

Figura 1: Exemple outlier diagrama de caixa

Puntuació Z

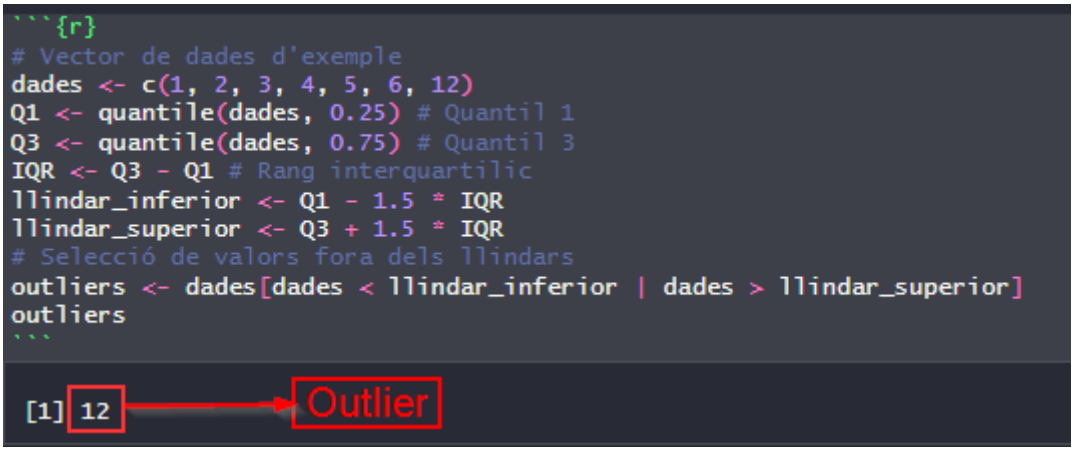

Figura 2: Exemple outlier IQR

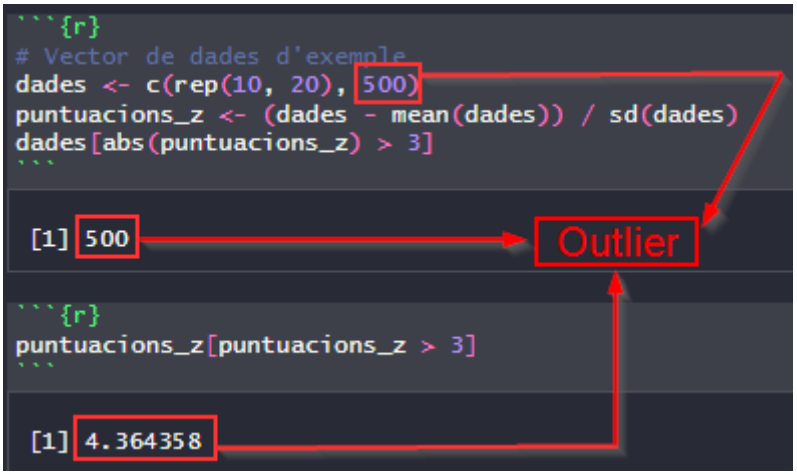

Figura 3: Exemple outlier Z-scoring

Exercici 2

Pregunta 1

Pregunta 2

Exercici 3

Pregunta 1

Pregunta 2

Exercici 4

Pregunta 1

Pregunta 2

Exercici 5

Pregunta 1