EXPLORACIÓ DE LES DADES

La ciència de dades ens permet descodificar la informació per tal d'adquirir més coneixements d'un sistema o del nostre entorn. Normalment, aquestes dades venen en forma de taula, conformada per dues entitats: files i columnes.

Les files, generalment, contenen els registres o observacions, i les columnes contenen les variables o atributs. Els registres, doncs, són els casos d'estudi que incorporen una informació determinada en cada atribut. Vegem-ne un exemple senzill:

	EDAT	PES	ALÇADA	CIUTAT	SEXE
JOAN	35	80	175	BARCELONA	HOME
ALBA	32	65	1,68	TARRAGONA	DONA
MIQUEL	47	74	1,78	GIRONA	HOME
MERITXELL	44	59	1,60	LLEIDA	DONA

- Les variables o atributs (columnes) són: edat, pes, alçada, ciutat i sexe.
- Les observacions o registres (files) són: Joan, Alba, Miquel i Meritxell.

Nota: d'ara en endavant sempre ens referirem a les files com observacions i a les columnes com a atributs.

És important saber que les taules poden venir amb dos tipus de dades: les **operacionals** i les **organitzatives**. Les primeres contenen informació directa d'observacions unitàries i puntuals com, per exemple, l'import i el producte de cada compra en un establiment, totes i cadascuna de les compres de bitllets d'avió d'una determinada companyia àrea, etc. Les segones, en canvi, recullen dades agrupades o tendències. Un exemple de dada organitzativa serien els casos d'una malaltia agrupats en diferents ciutats: aquí no estudiem cas per cas una observació individual, sinó una agrupació basada en certs criteris. Aquest segon grup es fa servir moltes vegades quan poden aparèixer conflictes relacionats amb la protecció de dades.

Recorda que el primer que necessitem establir abans de començar a treballar amb les dades són els **objectius**. Hem de saber respondre a la pregunta: quina informació rellevant volem obtenir amb aquestes dades? Una vegada ho tinguem clar, ens haurem de posar a treballar per aconseguir la resposta. Després del plantejament dels objectius, els següents passos són l'obtenció de la informació (que tindrem resolta, quan disposem d'una taula de dades) i, finalment, la preparació de les dades. Per fer aquest últim pas, hem de veure si tenim, o no, cel·les amb valors anormals o sense informació (*missing value*).

Nota: és important que ens familiaritzem amb expressions en anglès com *missing values*, ja que la indústria té com a referència aquest idioma i, tot i que treballis a Catalunya, de ben segur que les hauràs de fer servir.

Anem a veure, primer, un exemple de dades anormals:

	EDAT	PES	ALÇADA	CIUTAT	SEXE
JOAN	35	80	1,75	BARCELONA	HOME
ALBA	32	65	1,68	TARRAGONI	DONA
MIQUEL	47	74	1,78	TARRAGONA	HOME
MERITXELL	44	59	1,06	LLEIDA	DONA

En una exploració ràpida, podríem sospitar que l'alçada de la Meritxell pot tenir un valor anormal (probablement, degut a un error a l'hora de posar la dada). També veiem que el valor de ciutat de l'Alba és Tarragoni, i per tant, no correspon amb el valor real. Aquests errors, tard o d'hora, poden aparèixer i, si no els detectem, la informació que extraiem pot ser de baixa qualitat. A vegades, és pràcticament impossible detectar-los. Per això, és important treballar amb una base de dades àmplia, que tingui prou observacions per tal que aquests errors no ens alterin excessivament els nostres models predictius.

En els casos de cel·les sense valors, les causes poden ser molt diverses. Si, per exemple, registrem amb un termòmetre la temperatura cada cert període de temps i veiem cel·les sense valors, serà probablement degut a un mal funcionament de l'instrument de mesura. A continuació, mostrem un exemple d'un sistema format per cinc termòmetres, el qual ens serveix per mesurar la temperatura ambient cada 30 min en diverses zones d'una determinada ciutat.

L'objectiu d'aquesta taula podria ser, per exemple, veure si hi ha certes parts de la ciutat que són més caloroses que d'altres. L'estudi ens podria aportar coneixement important si podem relacionar les variacions de temperatura amb l'arquitectura del municipi o amb el trànsit. Això ens permetria proposar canvis urbanístics, si es considerés oportú.

	TERM 1	TERM 2	TERM 3	TERM 4	TERM 5
08:00	14,5	13,9	14,9	14,3	14,0
08:30	14,8	14,3	15,5	14,7	14,3
09:00	15,3	14,9	15,9		14,8
09:30	15,9	15,5	16,5		15,4
10:00	16,4	16,0	16,9		16,1
10:30	17,0	16,4	17,3	16,9	16,5
11:00	17,7	17,0	17,8	17,6	16,9

Nota: les columnes són els diferents termòmetres situats a diferents punts de la ciutat.

Les cel·les de color gris no contenen valors dins de la franja horària de 9:30 a 10:00. Per alguna raó desconeguda, el termòmetre no ha registrat la temperatura. En un cas real, hauríem d'analitzar si aquesta manca d'informació podria alterar significativament el nostre coneixement.

Descobreix tot el que Barcelona Activa pot fer per a tu

Acompanyament durant tot el procés de recerca de feina

barcelonactiva.cat/treball

Suport per posar en marxa la teva idea de negoci

barcelonactiva.cat/emprenedoria

Serveis a les empreses i iniciatives socioempresarials

barcelonactiva.cat/empreses

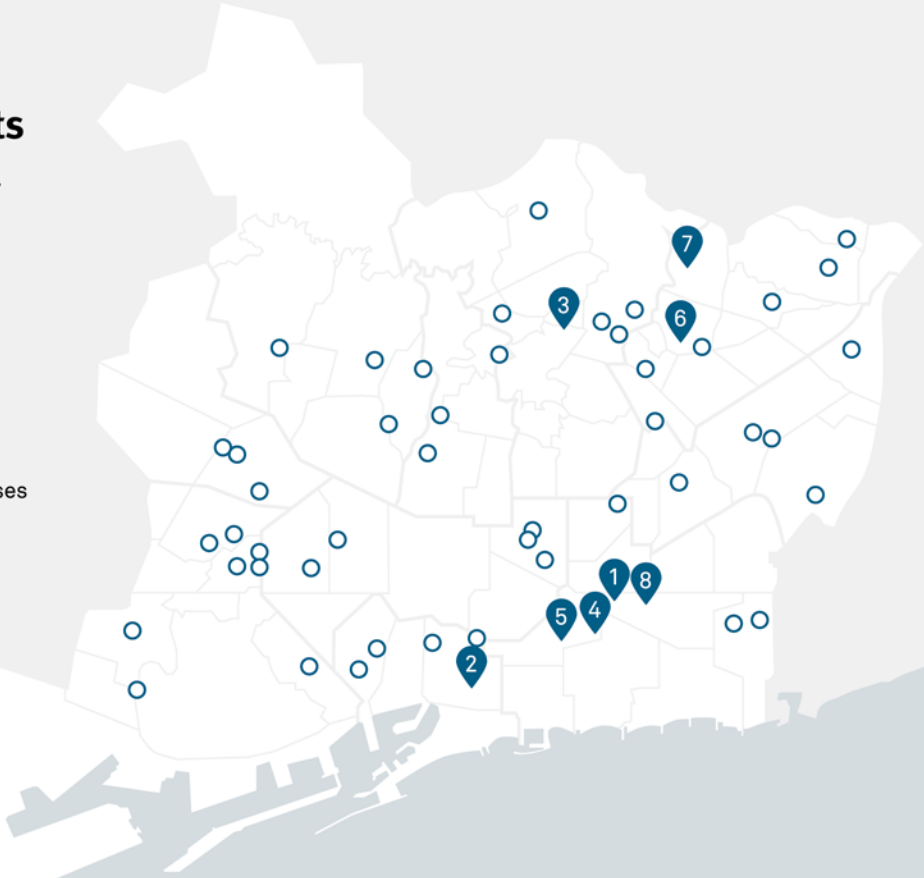
Formació tecnològica i gratuïta per a la ciutadania

barcelonactiva.cat/cibernarium

Xarxa d'equipaments de Barcelona Activa

1. Seu Central Barcelona Activa
   Porta 22
   Centre per a la Iniciativa Emprenedora Glòries
   Incubadora Glòries
2. Convent de Sant Agustí
3. Ca n'Andalet
4. Oficina d'Atenció a les Empreses
   Cibernàrium
   Incubadora MediaTIC
5. Incubadora Almogàvers
6. Parc Tecnològic
7. Nou Barris Activa
8. innoBA

○ Punts d'atenció a la ciutat



Barcelona Activa

Ajuntament de Barcelona

Darrera actualització 2019

Cofinançat per:

UNIÓ EUROPEA
Fons Europeu de Desenvolupament Regional

Segueix-nos a les xarxes socials:

barcelonactiva.cat/cibernarium
barcelonactiva
barcelonactiva
company/barcelona-activa