Ciència de dades i núvol

Tothom "sap" que l'agost és el mes amb més feina a les Illes Balears i que la majoria de turistes vénen d'Alemanya i del Regne Unit. Però quants, exactament? El Regne Unit va per davant de la resta d'Espanya? Quant més tranquil és el gener? Ho podries endevinar, o podries descarregar les xifres oficials i descobrir-ho en vint minuts amb un full de càlcul.

Així és la ciència de dades vista per dins: una pregunta clara, les dades adequades, molta neteja, un gràfic i una frase honesta al final. En aquesta lliçó ho faràs amb dades reals de l'**Ibestat**, l'institut d'estadística de les Illes Balears. Després veuràs on guarden les empreses les seves dades al núvol, i la regla amb què comença qualsevol pla de còpies de seguretat.

Què aprendràs

- Descriure les etapes de la **ciència de dades** i com cada una depèn de l'anterior
- Anomenar els principals **objectius** de la ciència de dades en una empresa
- Dur un **conjunt de dades obertes** real d'una pregunta a un gràfic i una conclusió a Google Sheets
- Distingir els principals tipus d'**emmagatzematge al núvol** i els procediments que hi protegeixen les dades
- Dissenyar un pla de còpies de seguretat amb la **regla 3-2-1**
- Evitar dues trampes: saltar-se la neteja i pensar que una carpeta sincronitzada és una còpia de seguretat

Les etapes de la ciència de dades

> La **ciència de dades** combina estadística, programació i coneixement del negoci per extreure informació i coneixement de les dades. Segueix un procés.

Etapa	Què fas	Exemple a l'hotel
1 · Fer la pregunta	Entendre el problema del negoci i què vols saber o decidir	En quin canal de venda hauríem d'invertir?
2 · Obtenir les dades	Trobar-les i recollir-les: bases de dades, fitxers, API, sensors, dades obertes	Exportar les reserves del sistema de l'hotel
3 · Netejar i preparar	Treure duplicats, tractar valors buits o impossibles, unificar formats	Esborrar una reserva repetida; unificar "ES", "spain", "Spain"
4 · Explorar i analitzar	Resumir, agrupar, fer gràfics, cercar patrons	Reserves i valoració mitjana per canal
5 · Modelar	Si cal, entrenar un model (ML) per predir o classificar	Un model que estima les cancel·lacions
6 · Comunicar i desplegar	Explicar els resultats a qui decideix i posar-los en ús	Un quadre de comandament per a l'equip de direcció

El procés és **iteratiu**: explorar les dades sovint et fa tornar enrere per cercar-ne més o per reescriure la pregunta. I les etapes estan encadenades: **una mala neteja ho espatlla tot el que ve després**. Un dels models de referència més coneguts, **CRISP-DM** (*Cross-Industry Standard*

Process for Data Mining, versió 1.0 publicada el 2000), també té sis fases i també torna enrere. No tots els projectes arriben a l'etapa 5: moltes preguntes es responen amb una bona anàlisi descriptiva.

Al joc de construir ciutats, aquest és el moment en què deixes d'admirar les capes i preguntes una cosa precisa: "per què el districte nord està sempre col·lapsat a les 8 del matí?" Llavors vas a cercar exactament les dades que ho poden respondre.

Per a què fan servir les empreses la ciència de dades

Objectiu	Exemple
Decidir amb dades, no per intuïció	Obrir el segon restaurant només les setmanes que cal
Conèixer el client i personalitzar l'oferta	Correus diferents per a famílies i per a parelles
Optimitzar processos i reduir costos	Planificar els torns de neteja a partir de la previsió d'arribades
Preveure la demanda, les avaries o els clients que se'n van	Predir quina bomba fallarà el mes que ve
Detectar anomalies: frau, errors, intrusions	Una targeta usada en dos països en deu minuts
Crear productes i serveis nous a partir de les dades	Vendre estadístiques de mobilitat anonimitzades a un ajuntament

> **Compte: "la ciència de dades és fer models"**
> La major part del temps se'n va a les etapes 2 i 3, obtenir i netejar les dades. La documentació d'IBM sobre CRISP-DM estima que la preparació de les dades sol ocupar entre el 50% i el 70% del temps i l'esforç d'un projecte. Les etapes 2, 3 i 6 són, en bona part, feina de desenvolupament: connectors, scripts de neteja, quadres de comandament i API. Aquesta és la teva futura feina.

A classe: de la pregunta al gràfic amb dades obertes reals

Treballa **en parella**, amb un ordinador, a **Google Sheets**. **Cap eina d'IA.**

La pregunta: *Quins tres llocs de residència van enviar més turistes a les Illes Balears el 2025, i quant canvia el nombre de turistes entre el mes amb més feina i el més tranquil?*

Pas	Fes això	Hauries de veure
1 · Obtenir	Ves a ibestat.es, cerca *Turistes amb destinació principal les Illes Balears per lloc de residència. Illes Balears i mes* i obre-la. Fes servir el botó **Descarregar** per baixar-la en CSV. Si la descàrrega falla, fes servir `ibestat-tourists-by-residence.csv` de Recursos, que és la mateixa taula (del 2023 a l'agost del 2026).	Un fitxer CSV amb períodes, llocs de residència i nombres de turistes
2 · Obtenir	A Google Sheets: *File* → *Import* → *Upload* (*Fitxer* → *Importa* → *Penja*), tria el fitxer, *Replace spreadsheet* (*Substitueix el full de càlcul*). Escriu la font i la data de descàrrega en una cel·la.	Unes 660 files i tres columnes
3 · Netejar	Troba els problemes abans de comptar res. Cerca: files que són **totals anuals** barrejades amb les files mensuals; files que són **agregats** (*Món…*, *Altres…*) que es comptarien dues vegades; valors **buits**; l'**ordre** de les files. Apunta cada problema i què has decidit.	Almenys quatre problemes, cada un amb una decisió

Pas	Fes això	Hauries de veure
4 · Analitzar	Queda't només amb els 12 mesos del 2025. Fes una **taula dinàmica** (*Insert → Pivot table*, *Insereix → Taula dinàmica*): files = lloc de residència, valors = SUM dels turistes. Ordena-la de més gran a més petita.	Alemanya primer, després el Regne Unit i la resta d'Espanya
5 · Analitzar	Per a la fila *Món (exclosa Illes Balears)* (tots els turistes), fes la llista dels 12 mesos del 2025 i troba el mes amb més feina i el més tranquil.	Un pic a l'estiu i una vall a l'hivern
6 · Representar	Insereix un **gràfic de barres** dels llocs principals i un **gràfic de línies** dels 12 mesos. Posa a cada un un títol i la font.	Dos gràfics
7 · Comunicar	Escriu **una frase** de conclusió per al director d'un hotel i **una limitació** de l'anàlisi.	Dues frases
8	Lliura l'enllaç del full de càlcul (compartit amb el professor) a Classroom.	

L'etapa 5, el model, se salta a posta: aquesta pregunta no en necessita cap. Deixa buides les cel·les buides: substituir un valor que falta per un 0 faria mentir la mitjana.

Per al teu projecte

A l'**apartat B del Lliurament 5** repeteixes aquest circuit pel teu compte: una pregunta que importi a l'empresa del teu projecte, un **conjunt de dades obertes real relacionat amb el seu sector** (Ibestat, el catàleg de dades obertes del Govern de les Illes Balears, datos.gob.es o l'INE), les decisions de neteja, un gràfic i una conclusió amb la seva limitació. El full d'avui és el teu model.

On viuen les dades: emmagatzematge al núvol

A la Unitat 3 vas veure el núvol com un model de servei (IaaS, PaaS, SaaS): llogar l'aigua i l'electricitat a una companyia en lloc de construir la teva pròpia central. Ara mira'l com el lloc on es guarden les dades.

Tipus	Com emmagatzema les dades	S'usa per a
Emmagatzematge d'objectes	Fitxers sencers amb metadades dins contenidors (***buckets***), accessibles a través d'una API	Imatges, vídeos, còpies de seguretat, registres, llacs de dades
Emmagatzematge de blocs	Volums que una màquina virtual munta com un disc	El disc del sistema i el disc de la base de dades d'un servidor
Emmagatzematge de fitxers	Un sistema de carpetes compartides per la xarxa	Carpetes compartides entre equips o aplicacions
Base de dades gestionada	Una base de dades SQL o NoSQL que administra el proveïdor	Les dades operatives d'una aplicació
Magatzem de dades (*data warehouse*)	Dades estructurades ja preparades per a l'anàlisi	Informes de negoci i quadres de comandament
Llac de dades (*data lake*)	Dades en brut en qualsevol format, a gran escala	Big data i entrenament de models

Els procediments habituals quan les dades s'emmagatzemen al núvol:

Procediment	Contra què protegeix
Replicació en diverses ubicacions	Que caigui un centre de dades (disponibilitat)

Procediment	Contra què protegeix
Còpies de seguretat i versions	Esborrar o sobreescriure per error, programari de segrest (*ransomware*)
Xifratge en repòs i en trànsit	Que algú que no hauria llegeixi les dades
Control d'accés: identitats, rols, privilegi mínim, registres d'accés	Massa gent que pot tocar massa coses
Nivells d'emmagatzematge: calent (accés freqüent, més car) i fred o d'arxiu (accés rar, més barat)	Pagar preus de "calent" per dades que no llegeixes mai; correspon a l'etapa d'arxiu del cicle de vida
Elecció de la regió: on són físicament les dades	Problemes legals per dades que surten de la UE (sobirania, Unitat 3)

Per què el núvol és important per a les dades: capacitat a demanda (créixer i decréixer), pagament per ús sense comprar maquinari, accés des de qualsevol lloc, alta disponibilitat i serveis d'anàlisi i d'IA a punt per fer servir. Sense el núvol, treballar amb big data i IA requeriria una infraestructura que la majoria d'empreses petites no es poden permetre.

> **Compte: les dades continuen sent teves**
> Passar les dades al núvol no hi passa la responsabilitat. Segons el **model de responsabilitat compartida**, el proveïdor és responsable de la seguretat **del** núvol (edificis, maquinari, la seva xarxa); l'empresa continua sent responsable del que posa **al** núvol: qui hi té accés, què està xifrat i quant de temps es conserva.

La regla 3-2-1 de les còpies de seguretat

> **Regla 3-2-1**: guarda **3** còpies de qualsevol dada important (l'original i dues còpies de seguretat), en **2** tipus de suport diferents, amb **1** còpia fora de les instal·lacions.

La va popularitzar el fotògraf Peter Krogh als seus llibres sobre gestió d'actius digitals, i l'equip d'emergències informàtiques del govern dels Estats Units (US-CERT, avui part de CISA) la va recomanar el 2012 en un document, *Data Backup Options*. Al joc de construir ciutats és guardar els plànols de la ciutat a l'ajuntament, una còpia en un altre tipus de suport al parc de bombers i una altra còpia en una cambra cuirassada d'una altra ciutat.

Còpia	Exemple per a una petita empresa hotelera
1 · Original	La base de dades del sistema de gestió de l'hotel a la seva regió del núvol
2 · Segona còpia, suport diferent	Una còpia de seguretat cada nit a l'emmagatzematge d'objectes, o a un NAS (un disc de xarxa) a l'oficina
3 · Còpia fora de les instal·lacions	Una còpia en una altra regió o en un altre proveïdor, idealment una que no es pugui modificar durant 30 dies (immutable)

> **Compte: una carpeta sincronitzada no és una còpia de seguretat**
> Si un fitxer s'esborra o el xifra un programari de segrest en una carpeta sincronitzada amb Google Drive o OneDrive, la sincronització copia fidelment el dany a tot arreu. Una còpia de seguretat és una còpia **separada** a la qual pots tornar, i només saps que funciona si n'has **provat una restauració**. El desastre clàssic de petita empresa, en què l'única còpia de seguretat era al mateix servidor que va quedar xifrat, incompleix la part "1 fora de les instal·lacions" de la regla.

Idees clau

- La **ciència de dades** segueix etapes encadenades: pregunta, dades, neteja, anàlisi, model (només si cal) i comunicació; és **iterativa**.
- La neteja ocupa la major part del temps, i una mala neteja espatlla tots els resultats que vénen després.
- Les empreses fan servir la ciència de dades per decidir amb dades, conèixer els clients, optimitzar, preveure, detectar anomalies i crear serveis nous.
- Els principals tipus d'**emmagatzematge al núvol** són el d'objectes, el de blocs, el de fitxers, les bases de dades gestionades, els **magatzems de dades** i els **llacs de dades**, protegits amb replicació, còpies de seguretat, xifratge, control d'accés, nivells d'emmagatzematge i l'elecció de la regió.
- La **regla 3-2-1**: tres còpies, dos tipus de suport, una fora de les instal·lacions. Una carpeta sincronitzada no és una còpia de seguretat; una còpia que no has restaurat mai és una esperança.

Comprovació ràpida

1 Quina és la primera etapa d'un projecte de ciència de dades?

A Fer un gràfic
B Fer la pregunta: entendre el problema del negoci
C Entrenar un model
D Comprar més emmagatzematge

2 A la taula de l'Ibestat, els totals anuals ("2025") estan barrejats amb les files mensuals. Què passa si ho sumes tot per al 2025?

A Res: el resultat és el mateix
B Comptes cada turista dues vegades, perquè la fila anual ja suma els mesos
C El full de càlcul no les vol sumar
D Obtens la mitjana en lloc del total

3 A les dades hi falta un valor. Quina és la millor decisió?

A Substituir-lo per 0 perquè les fórmules funcionin
B Deixar-lo buit i apuntar-lo com a limitació, perquè no es compti com un zero real
C Esborrar tota la columna
D Inventar-se un número raonable

4 Quin tipus d'emmagatzematge al núvol guarda fitxers sencers en *buckets*, accessibles a través d'una API?

A Emmagatzematge de blocs
B Emmagatzematge d'objectes
C Un magatzem de dades
D Emmagatzematge de fitxers

5 Què demana la regla 3-2-1?

A 3 contrasenyes, 2 usuaris, 1 administrador
B 3 còpies de les dades, en 2 tipus de suport, amb 1 còpia fora de les instal·lacions
C 3 proveïdors de núvol, 2 regions, 1 centre de dades
D Còpies de seguretat cada 3 dies, conservades 2 setmanes, provades un cop l'any

6 Una empresa guarda tots els documents en una carpeta sincronitzada amb un disc al núvol i diu que en té còpia de seguretat. Un programari de segrest xifra el portàtil. Tria les dues afirmacions certes.

- **A** La sincronització pot copiar també al núvol els fitxers xifrats, així que no és una còpia de seguretat de veritat
- **B** El proveïdor del núvol és responsable de recuperar els fitxers segons el model de responsabilitat compartida
- **C** Una còpia separada que no es pugui modificar, i una restauració provada, l'haurien protegida
- **D** És segur perquè el núvol sempre està xifrat

7 Una empresa d'aigües vol saber quines bombes és probable que fallin el mes que ve. Quina etapa de la ciència de dades podria necessitar que la pregunta d'avui sobre turisme no necessitava?

- **A** Fer la pregunta
- **B** La neteja
- **C** Modelar: entrenar un model que predigui les avaries
- **D** Comunicar

Les teves dades ja tenen una casa i un pla de còpies de seguretat. L'últim pas de la unitat és protegir-les de qui no hi hauria d'entrar, i saber quines lleis ho exigeixen: Ciberseguretat i regulació.

Solucions

1. B · Tota la resta depèn de la **pregunta**: decideix quines dades obtenir, com netejar-les i quin gràfic la respon.

2. B · La fila anual és un **agregat** de les mensuals. Comptar-les totes dues duplica el total. Detectar els agregats forma part de la **neteja**, l'etapa 3.

3. B · Un 0 és un valor real i faria baixar els totals i les mitjanes. Deixar-lo **buit** i dir-ho és una neteja honesta.

4. B · L'**emmagatzematge d'objectes** guarda fitxers amb les seves metadades en ***buckets***. És el lloc habitual per a imatges, registres i còpies de seguretat.

5. B · Tres còpies (l'original i dues còpies de seguretat), dos suports diferents i una còpia **fora de les instal·lacions**, perquè un sol incendi, robatori o atac de programari de segrest no les pugui destruir totes.

6. A, C · Una carpeta sincronitzada replica tant els canvis com els **danys**. Segons el **model de responsabilitat compartida**, les dades continuen sent responsabilitat de l'empresa. Una còpia separada, idealment immutable, i una restauració provada són el que fa que una còpia de seguretat ho sigui.

7. C · La pregunta d'avui era **descriptiva**, així que no calia cap model. Predir avaries és una pregunta **predictiva**, i és aquí on sol entrar l'etapa 5, el model.