# Ciencia de datos y nube

Todo el mundo «sabe» que agosto es el mes con más movimiento en las Illes Balears y que la mayoría de los turistas vienen de Alemania y del Reino Unido. Pero ¿cuántos, exactamente? ¿El Reino Unido va por delante del resto de España? ¿Cuánto más tranquilo es enero? Podrías adivinarlo, o podrías descargar las cifras oficiales y averiguarlo en veinte minutos con una hoja de cálculo.

Así es la ciencia de datos vista desde dentro: una pregunta clara, los datos adecuados, mucha limpieza, un gráfico y una frase honesta al final. En esta lección lo harás con datos reales del **Ibestat**, el instituto de estadística de las Illes Balears. Después verás dónde guardan las empresas sus datos en la nube y la regla con la que empieza todo plan de copias de seguridad.

**Qué aprenderás**

- Describir las etapas de la **ciencia de datos** y cómo cada una depende de la anterior
- Nombrar los principales **objetivos** de la ciencia de datos en una empresa
- Llevar un **conjunto de datos abiertos** real de una pregunta a un gráfico y una conclusión en Google Sheets
- Distinguir los principales tipos de **almacenamiento en la nube** y los procedimientos que protegen los datos que contienen
- Diseñar un plan de copias de seguridad con la **regla 3-2-1**
- Evitar dos trampas: saltarse la limpieza y creer que una carpeta sincronizada es una copia de seguridad

## Las etapas de la ciencia de datos

> La **ciencia de datos** combina estadística, programación y conocimiento del negocio para extraer información y conocimiento de los datos. Sigue un proceso.

| Etapa | Qué haces | Ejemplo del hotel |
|---|---|---|
| **1 · Plantear la pregunta** | Entender el problema de negocio y qué quieres saber o decidir | ¿En qué canal de venta deberíamos invertir? |
| **2 · Obtener los datos** | Encontrarlos y recogerlos: bases de datos, archivos, API, sensores, datos abiertos | Exportar las reservas del sistema del hotel |
| **3 · Limpiar y preparar** | Quitar duplicados, tratar valores vacíos o imposibles, unificar formatos | Borrar una reserva repetida; unificar «ES», «spain», «Spain» |
| **4 · Explorar y analizar** | Resumir, agrupar, hacer gráficos, buscar patrones | Reservas y valoración media por canal |
| **5 · Modelar** | Si hace falta, entrenar un modelo (ML) para predecir o clasificar | Un modelo que estima las cancelaciones |
| **6 · Comunicar y desplegar** | Explicar los resultados a quien decide y ponerlos en uso | Un cuadro de mando para el equipo directivo |

El proceso es **iterativo**: explorar los datos a menudo te hace volver atrás a buscar más datos o a reescribir la pregunta. Y las etapas están encadenadas: **una mala limpieza estropea todo**

**lo que viene después**. Uno de los modelos de referencia más conocidos, **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*, versión 1.0 publicada en 2000), también tiene seis fases y también vuelve atrás. No todos los proyectos llegan a la etapa 5: muchas preguntas se responden con un buen análisis descriptivo.

En el juego de construir ciudades, este es el momento en que dejas de admirar las capas y preguntas una cosa precisa: «¿por qué el distrito norte está siempre colapsado a las 8 de la mañana?». Entonces vas a buscar justo los datos que pueden responderla.

### Para qué usan las empresas la ciencia de datos

| Objetivo | Ejemplo |
|---|---|
| **Decidir con datos**, no por intuición | Abrir el segundo restaurante solo las semanas que lo necesitan |
| **Conocer al cliente** y personalizar la oferta | Correos distintos para familias y para parejas |
| **Optimizar procesos** y reducir costes | Planificar los turnos de limpieza a partir de la previsión de llegadas |
| **Prever** la demanda, las averías o la marcha de clientes | Predecir qué bomba fallará el mes que viene |
| **Detectar anomalías**: fraude, errores, intrusiones | Una tarjeta usada en dos países en diez minutos |
| **Crear nuevos productos y servicios** a partir de los datos | Vender estadísticas de movilidad anonimizadas a un ayuntamiento |

> **Cuidado: «la ciencia de datos es hacer modelos»**
> La mayor parte del tiempo se va en las etapas 2 y 3, obtener y limpiar los datos. La documentación de IBM sobre CRISP-DM calcula que la preparación de los datos suele ocupar entre el 50 % y el 70 % del tiempo y del esfuerzo de un proyecto. Las etapas 2, 3 y 6 son, en gran parte, trabajo de desarrollo: conectores, scripts de limpieza, cuadros de mando y API. Ese es tu futuro trabajo.

## En clase: de la pregunta al gráfico con datos abiertos reales

Trabaja **en parejas**, con un ordenador, en **Google Sheets**. **Sin herramientas de IA.**

**La pregunta:** *¿Cuáles son los tres lugares de residencia que enviaron más turistas a las Illes Balears en 2025, y cuánto cambia el número de turistas entre el mes con más movimiento y el más tranquilo?*

| Paso | Haz esto | Deberías ver |
|---|---|---|
| **1 · Obtener** | Ve a ibestat.es, busca *Turistes amb destinació principal les Illes Balears per lloc de residència. Illes Balears i mes* y ábrelo. Usa el botón **Descarregar** para descargarlo en CSV. Si la descarga falla, usa `ibestat-tourists-by-residence.csv` de Recursos, que es la misma tabla (de 2023 a agosto de 2026). | Un archivo CSV con periodos, lugares de residencia y números de turistas |
| **2 · Obtener** | En Google Sheets: *File → Import → Upload*, elige el archivo, *Replace spreadsheet*. Anota la fuente y la fecha de descarga en una celda. | Unas 660 filas y tres columnas |
| **3 · Limpiar** | Encuentra los problemas antes de contar nada. Busca: filas que son **totales anuales** mezcladas con las filas mensuales; filas que son **agregados** (*Món…*, *Altres…*) que se contarían dos veces; valores **vacíos**; el **orden** de las filas. Anota cada problema y qué has decidido. | Al menos cuatro problemas, cada uno con una decisión |

| Paso | Haz esto | Deberías ver |
|---|---|---|
| **4 · Analizar** | Quédate solo con los 12 meses de 2025. Haz una **tabla dinámica** (*Insert → Pivot table*): filas = lugar de residencia, valores = SUM de turistas. Ordénala de mayor a menor. | Alemania primero, después el Reino Unido y el resto de España |
| **5 · Analizar** | Para la fila *Món (exclosa Illes Balears)* (todos los turistas), lista los 12 meses de 2025 y encuentra el mes con más movimiento y el más tranquilo. | Un pico en verano y un valle en invierno |
| **6 · Representar** | Inserta un **gráfico de barras** de los primeros lugares y un **gráfico de líneas** de los 12 meses. Pon a cada uno un título y la fuente. | Dos gráficos |
| **7 · Comunicar** | Escribe **una frase** de conclusión para la dirección de un hotel y **una limitación** del análisis. | Dos frases |
| **8** | Entrega el enlace de la hoja de cálculo (compartida con el profesor) en Classroom. | |

La etapa 5, el modelo, se salta a propósito: esta pregunta no lo necesita. Deja vacías las celdas vacías: sustituir un valor que falta por 0 haría mentir a la media.

## Para tu proyecto

En la **sección B de la Entrega 5** repites este proceso por tu cuenta: una pregunta que importe a la empresa de tu proyecto, un **conjunto de datos abiertos real relacionado con su sector** (Ibestat, el catálogo de datos abiertos del Govern de las Illes Balears, datos.gob.es o el INE), las decisiones de limpieza, un gráfico y una conclusión con su limitación. La hoja de hoy es tu modelo.

## Dónde viven los datos: el almacenamiento en la nube

En la Unidad 3 viste la nube como modelo de servicio (IaaS, PaaS, SaaS): alquilar el agua y la electricidad a una compañía de suministro en lugar de construir tu propia central. Ahora mírala como el lugar donde se guardan los datos.

| Tipo | Cómo guarda los datos | Para qué se usa |
|---|---|---|
| **Almacenamiento de objetos** | Archivos enteros con metadatos dentro de contenedores (**buckets**), a los que se accede mediante una API | Imágenes, vídeos, copias de seguridad, registros, lagos de datos |
| **Almacenamiento de bloques** | Volúmenes que una máquina virtual monta como si fueran un disco | El disco del sistema y el disco de la base de datos de un servidor |
| **Almacenamiento de archivos** | Un sistema de carpetas compartidas por la red | Carpetas compartidas entre equipos o aplicaciones |
| **Base de datos gestionada** | Una base de datos SQL o NoSQL que administra el proveedor | Los datos operativos de una aplicación |
| **Almacén de datos (data warehouse)** | Datos estructurados ya preparados para el análisis | Informes de negocio y cuadros de mando |
| **Lago de datos (data lake)** | Datos en bruto en cualquier formato, a gran escala | Big data y entrenamiento de modelos |

Los procedimientos habituales cuando los datos se guardan en la nube:

| Procedimiento | Contra qué protege |
|---|---|
| **Replicación** en varias ubicaciones | La caída de un centro de datos (disponibilidad) |

| Procedimiento | Contra qué protege |
|---|---|
| **Copias de seguridad y versionado** | Borrar o sobrescribir por error, ransomware |
| **Cifrado** en reposo y en tránsito | Que lea los datos alguien que no debería |
| **Control de acceso**: identidades, roles, mínimo privilegio, registros de acceso | Demasiada gente que puede tocar demasiadas cosas |
| **Niveles de almacenamiento**: caliente (acceso frecuente, más caro) y frío o de archivo (acceso poco frecuente, más barato) | Pagar precios «calientes» por datos que nunca lees; encaja con la etapa de archivo del ciclo de vida |
| **Elección de la región**: dónde están físicamente los datos | Problemas legales si los datos salen de la UE (soberanía, Unidad 3) |

**Por qué la nube importa para los datos:** capacidad bajo demanda (escalar hacia arriba y hacia abajo), pago por uso sin comprar hardware, acceso desde cualquier lugar, alta disponibilidad y servicios de análisis y de IA listos para usar. Sin la nube, trabajar con big data e IA necesitaría una infraestructura que la mayoría de las empresas pequeñas no se pueden permitir.

> **Cuidado: los datos siguen siendo tuyos**
> Llevar los datos a la nube no traslada la responsabilidad. Según el **modelo de responsabilidad compartida**, el proveedor es responsable de la seguridad **de** la nube (edificios, hardware, su red); la empresa sigue siendo responsable de lo que pone **en** la nube: quién tiene acceso, qué está cifrado y cuánto tiempo se guarda.

## La regla 3-2-1 de copias de seguridad

> **Regla 3-2-1**: guarda **3** copias de cualquier dato importante (el original y dos copias de seguridad), en **2** tipos de soporte distintos, con **1** copia fuera de las instalaciones.

La popularizó el fotógrafo Peter Krogh en sus libros sobre gestión de archivos digitales, y el equipo de respuesta a emergencias informáticas del Gobierno de EE. UU. (US-CERT, hoy parte de CISA) la recomendó en un documento de 2012, *Data Backup Options*. En el juego de construir ciudades es guardar los planos de la ciudad en el ayuntamiento, una copia en otro tipo de soporte en el parque de bomberos y otra copia en una cámara acorazada de otra ciudad.

| Copia | Ejemplo para una pequeña empresa hotelera |
|---|---|
| **1 · Original** | La base de datos del sistema de gestión del hotel en su región de la nube |
| **2 · Segunda copia, otro soporte** | Una copia de seguridad cada noche en almacenamiento de objetos, o en un NAS (un disco de red) en la oficina |
| **3 · Copia fuera de las instalaciones** | Una copia en otra región o en otro proveedor, idealmente una que no se pueda modificar durante 30 días (inmutable) |

> **Cuidado: una carpeta sincronizada no es una copia de seguridad**
> Si un archivo se borra o lo cifra un ransomware en una carpeta sincronizada con Google Drive o OneDrive, la sincronización copia fielmente el daño a todas partes. Una copia de seguridad es una copia **separada** a la que puedes volver, y solo sabes que funciona si has **probado una restauración**. El desastre clásico de la pequeña empresa, en el

> que la única copia de seguridad estaba en el mismo servidor que acabó cifrado, incumple la parte «1 fuera de las instalaciones» de la regla.

## Ideas clave

- La **ciencia de datos** sigue etapas encadenadas: pregunta, datos, limpieza, análisis, modelo (solo si hace falta) y comunicación; es **iterativa**.
- La limpieza ocupa la mayor parte del tiempo, y una mala limpieza estropea todos los resultados que vienen después.
- Las empresas usan la ciencia de datos para decidir con datos, conocer a los clientes, optimizar, prever, detectar anomalías y crear nuevos servicios.
- Los principales tipos de **almacenamiento en la nube** son el de objetos, el de bloques, el de archivos, las bases de datos gestionadas, los **almacenes de datos (data warehouses)** y los **lagos de datos (data lakes)**, protegidos con replicación, copias de seguridad, cifrado, control de acceso, niveles de almacenamiento y la elección de la región.
- La **regla 3-2-1**: tres copias, dos tipos de soporte, una fuera de las instalaciones. Una carpeta sincronizada no es una copia de seguridad; una copia de seguridad que nunca has restaurado es una esperanza.

## Comprobación rápida

**1** ¿Cuál es la primera etapa de un proyecto de ciencia de datos?

- **A** Hacer un gráfico
- **B** Plantear la pregunta: entender el problema de negocio
- **C** Entrenar un modelo
- **D** Comprar más almacenamiento

**2** En la tabla del Ibestat, los totales anuales («2025») están mezclados con las filas mensuales. ¿Qué pasa si los sumas todos para 2025?

- **A** Nada: el resultado es el mismo
- **B** Cuentas cada turista dos veces, porque la fila anual ya suma los meses
- **C** La hoja de cálculo se niega a sumarlos
- **D** Obtienes la media en lugar del total

**3** Falta un valor en los datos. ¿Cuál es la mejor decisión?

- **A** Sustituirlo por 0 para que funcionen las fórmulas
- **B** Dejarlo vacío y anotarlo como limitación, para que no cuente como un cero real
- **C** Borrar toda la columna
- **D** Inventarse un número razonable

**4** ¿Qué tipo de almacenamiento en la nube guarda archivos enteros en buckets, accesibles mediante una API?

- **A** Almacenamiento de bloques
- **B** Almacenamiento de objetos
- **C** Un almacén de datos (data warehouse)
- **D** Almacenamiento de archivos

**5** ¿Qué pide la regla 3-2-1?

- **A** 3 contraseñas, 2 usuarios, 1 administrador
- **B** 3 copias de los datos, en 2 tipos de soporte, con 1 copia fuera de las instalaciones
- **C** 3 proveedores de nube, 2 regiones, 1 centro de datos
- **D** Copias de seguridad cada 3 días, guardadas 2 semanas, probadas una vez al año

**6** Una empresa guarda todos sus documentos en una carpeta sincronizada con una unidad en la nube y dice que tiene copia de seguridad. Un ransomware cifra el portátil. Elige las dos afirmaciones verdaderas.

- **A** La sincronización puede copiar también los archivos cifrados a la nube, así que no es una copia de seguridad de verdad
- **B** El proveedor de nube es responsable de recuperar los archivos según el modelo de responsabilidad compartida
- **C** Una copia separada que no se pueda modificar, y una restauración probada, la habrían protegido
- **D** Está a salvo porque la nube siempre está cifrada

**7** Una empresa de agua quiere saber qué bombas tienen más probabilidades de fallar el mes que viene. ¿Qué etapa de la ciencia de datos podría necesitar que la pregunta de turismo de hoy no necesitaba?

- **A** Plantear la pregunta
- **B** Limpiar
- **C** Modelar: entrenar un modelo que prediga las averías
- **D** Comunicar

Tus datos ya tienen casa y plan de copias de seguridad. El último paso de la unidad es protegerlos de quien no debería entrar, y saber qué leyes lo exigen: Ciberseguridad y regulación.

## Soluciones

**1. B** · Todo lo demás depende de la **pregunta**: decide qué datos obtener, cómo limpiarlos y qué gráfico la responde.

**2. B** · La fila anual es un **agregado** de las mensuales. Contar las dos duplica el total. Detectar los agregados forma parte de la **limpieza**, la etapa 3.

**3. B** · Un 0 es un valor real y haría bajar los totales y las medias. Dejarlo **vacío** y decirlo es una limpieza honesta.

**4. B** · El **almacenamiento de objetos** guarda archivos con sus metadatos en **buckets**. Es el lugar habitual de imágenes, registros y copias de seguridad.

**5. B** · Tres copias (el original y dos copias de seguridad), dos soportes distintos y una copia **fuera de las instalaciones**, para que un solo incendio, robo o ataque de ransomware no pueda destruirlas todas.

**6. A, C** · Una carpeta sincronizada replica el **daño** igual que los cambios. Según el **modelo de responsabilidad compartida**, los datos siguen siendo responsabilidad de la empresa. Una copia separada, idealmente inmutable, y una restauración probada son lo que hace que algo sea una copia de seguridad.

**7. C** · La pregunta de hoy era **descriptiva**, así que no hacía falta ningún modelo. Predecir averías es una pregunta **predictiva**, y es ahí donde suele entrar la etapa 5, el modelo.