Data science and the cloud

Everybody "knows" that August is the busiest month in the Balearic Islands and that most tourists come from Germany and the United Kingdom. But how many, exactly? Is the UK ahead of the rest of Spain? How much quieter is January? You could guess, or you could download the official figures and find out in twenty minutes with a spreadsheet.

That is what data science looks like from the inside: a clear question, the right data, a lot of cleaning, a chart and one honest sentence at the end. In this lesson you will do it with real data from **Ibestat**, the Balearic statistics institute. Then you will look at where companies keep their data in the cloud, and at the one rule every backup plan starts with.

What you will learn

- Describe the stages of **data science** and how each one depends on the previous one
- Name the main **objectives** of data science in a company
- Take a real **open dataset** from a question to a chart and a conclusion in Google Sheets
- Tell the main kinds of **cloud storage** apart and the procedures that protect data in them
- Design a backup plan with the **3-2-1 rule**
- Avoid two traps: skipping the cleaning, and thinking a synced folder is a backup

The stages of data science

> **Data science** combines statistics, programming and knowledge of the business to extract information and knowledge from data. It follows a process.

Stage	What you do	Hotel example
1 · Ask the question	Understand the business problem and what you want to know or decide	Which sales channel should we invest in?
2 · Get the data	Find and collect them: databases, files, APIs, sensors, open data	Export the bookings from the hotel system
3 · Clean and prepare	Remove duplicates, handle empty or impossible values, unify formats	Delete a repeated booking; unify "ES", "spain", "Spain"
4 · Explore and analyse	Summarise, group, chart, look for patterns	Bookings and average rating per channel
5 · Model	If needed, train a model (ML) to predict or classify	A model that estimates cancellations
6 · Communicate and deploy	Explain the results to whoever decides and put them to use	A dashboard for the management team

The process is **iterative**: exploring the data often sends you back to look for more data or to rewrite the question. And the stages are chained: **bad cleaning ruins everything after it**. One of the best-known reference models, **CRISP-DM** (*Cross-Industry Standard Process for Data*

Mining, version 1.0 published in 2000), also has six phases and also loops back. Not every project reaches stage 5: many questions are answered with a good descriptive analysis.

In the city-builder game this is the moment you stop admiring the overlays and ask one precise thing: "why is the north district always jammed at 8 a.m.?" Then you go and get exactly the data that can answer it.

What companies use data science for

Objective	Example
Decide with data, not by intuition	Open the second restaurant only on the weeks that need it
Know the customer and personalise the offer	Different emails for families and for couples
Optimise processes and cut costs	Plan cleaning shifts from the arrivals forecast
Forecast demand, breakdowns or customers leaving	Predict which pump will fail next month
Detect anomalies: fraud, errors, intrusions	A card used in two countries in ten minutes
Create new products and services from data	Sell anonymised mobility statistics to a town council

> **Careful: "data science is building models"**
> Most of the time goes into stages 2 and 3, getting and cleaning the data. IBM's documentation of CRISP-DM estimates that data preparation usually takes between 50% and 70% of a project's time and effort. Stages 2, 3 and 6 are, to a large extent, development work: connectors, cleaning scripts, dashboards and APIs. That is your future job.

In class: from question to chart with real open data

Work **in pairs**, one computer, in **Google Sheets**. **No AI tools.**

The question: *Which three places of residence sent the most tourists to the Balearic Islands in 2025, and how much does the number of tourists change between the busiest and the quietest month?*

Step	Do this	You should see
1 · Get	Go to ibestat.es, search for *Turistes amb destinació principal les Illes Balears per lloc de residència. Illes Balears i mes* and open it. Use the **Descarregar** button to download it as CSV. If the download fails, use `ibestat-tourists-by-residence.csv` from Resources, which is the same table (2023 to August 2026).	A CSV file with periods, places of residence and numbers of tourists
2 · Get	In Google Sheets: *File* → *Import* → *Upload*, choose the file, *Replace spreadsheet*. Write down the source and the download date in a cell.	About 660 rows and three columns
3 · Clean	Find the problems before you count anything. Look for: rows that are **yearly totals** mixed with monthly rows; rows that are **aggregates** (*Món…*, *Altres…*) that would be counted twice; **empty** values; the **order** of the rows. Write down every problem and what you decided.	At least four problems, each with a decision
4 · Analyse	Keep only the 12 months of 2025. Make a **pivot table** (*Insert* → *Pivot table*): rows = place of residence, values = SUM of tourists. Sort it from largest to smallest.	Germany first, then the United Kingdom and the rest of Spain
5 · Analyse	For the row *Món (exclosa Illes Balears)* (all tourists), list the 12 months of 2025 and find the busiest and the quietest month.	A peak in summer and a valley in winter

Step	Do this	You should see
6 · Represent	Insert a **bar chart** of the top places and a **line chart** of the 12 months. Give each a title and the source.	Two charts
7 · Communicate	Write **one sentence** of conclusion for a hotel manager and **one limitation** of the analysis.	Two sentences
8	Hand in the spreadsheet link (shared with the teacher) on Classroom.	

Stage 5, the model, is skipped on purpose: this question does not need one. Keep empty cells empty: replacing a missing value with 0 would make the average lie.

For your project

In **section B of Deliverable 5** you repeat this pipeline on your own: a question that matters to your project company, a **real open dataset related to its sector** (Ibestat, the Balearic Government's open data catalogue, datos.gob.es or the INE), the cleaning decisions, one chart and a conclusion with its limitation. Today's sheet is your model.

Where the data live: cloud storage

In Unit 3 you saw the cloud as a service model (IaaS, PaaS, SaaS): renting the water and electricity from a utility instead of building your own plant. Now look at it as the place where the data are kept.

Kind	How it stores data	Used for
Object storage	Whole files with metadata inside containers (**buckets**), reached through an API	Images, videos, backups, logs, data lakes
Block storage	Volumes that a virtual machine mounts as a disk	The system disk and database disk of a server
File storage	A shared folder system over the network	Folders shared between teams or applications
Managed database	A SQL or NoSQL database that the provider administers	The operational data of an application
Data warehouse	Structured data already prepared for analysis	Business reports and dashboards
Data lake	Raw data in any format, at large scale	Big data and training models

The usual procedures when data are stored in the cloud:

Procedure	What it protects against
Replication in several locations	A data centre going down (availability)
Backups and versioning	Deleting or overwriting by mistake, ransomware
Encryption at rest and in transit	Somebody reading the data who should not
Access control: identities, roles, least privilege, access logs	Too many people able to touch too much
Storage tiers: hot (frequent access, more expensive) and cold or archive (rare access, cheaper)	Paying "hot" prices for data you never read; it matches the archive stage of the lifecycle
Choice of region: where the data physically are	Legal problems with data leaving the EU (sovereignty, Unit 3)

Why the cloud matters for data: capacity on demand (scale up and down), pay per use without buying hardware, access from anywhere, high availability and analysis and AI services ready to use. Without the cloud, working with big data and AI would need an infrastructure that most small companies cannot afford.

> **Careful: the data are still yours**
> Moving data to the cloud does not move the responsibility. Under the **shared responsibility model**, the provider is responsible for the security **of** the cloud (buildings, hardware, its network); the company is still responsible for what it puts **in** the cloud: who has access, what is encrypted and how long it is kept.

The 3-2-1 backup rule

> **3-2-1 rule**: keep **3** copies of any important data (the original and two backups), on **2** different types of media, with **1** copy off-site.

It was popularised by the photographer Peter Krogh in his books on digital asset management, and the US government's computer emergency team (US-CERT, now part of CISA) recommended it in a 2012 paper, *Data Backup Options*. In the city-builder game it is keeping the city plans in the town hall, a copy on a different kind of storage in the fire station, and another copy in a vault in a different town.

Copy	Example for a small hotel company
1 · Original	The hotel management system's database in its cloud region
2 · Second copy, different medium	A nightly backup to object storage, or a NAS (a network disk) in the office
3 · Off-site copy	A copy in another region or another provider, ideally one that cannot be changed for 30 days (immutable)

> **Careful: a synced folder is not a backup**
> If a file is deleted or encrypted by ransomware in a folder synced with Google Drive or OneDrive, the sync faithfully copies the damage everywhere. A backup is a **separate** copy you can go back to, and you only know it works if you have **tested a restore**. The classic small-company disaster, where the only backup sat on the same server that got encrypted, breaks the "1 off-site" part of the rule.

Key takeaways

- **Data science** follows chained stages: question, data, cleaning, analysis, model (only if needed) and communication; it is **iterative**.
- Cleaning takes most of the time, and bad cleaning ruins every result after it.
- Companies use data science to decide with data, know customers, optimise, forecast, detect anomalies and create new services.
- The main kinds of **cloud storage** are object, block, file, managed databases, **data warehouses** and **data lakes**, protected by replication, backups, encryption, access control, storage tiers and the choice of region.

- The **3-2-1 rule**: three copies, two types of media, one off-site. A synced folder is not a backup; a backup you have never restored is a hope.

Quick check

1 What is the first stage of a data-science project?

- **A** Making a chart
- **B** Asking the question: understanding the business problem
- **C** Training a model
- **D** Buying more storage

2 In the Ibestat table, yearly totals ("2025") are mixed with monthly rows. What happens if you add them all up for 2025?

- **A** Nothing: the result is the same
- **B** You count every tourist twice, because the yearly row already adds up the months
- **C** The spreadsheet refuses to add them
- **D** You get the average instead of the total

3 A value is missing in the data. What is the best decision?

- **A** Replace it with 0 so the formulas work
- **B** Leave it empty and note it as a limitation, so it is not counted as a real zero
- **C** Delete the whole column
- **D** Invent a reasonable number

4 Which kind of cloud storage keeps whole files in buckets, reachable through an API?

- **A** Block storage
- **B** Object storage
- **C** A data warehouse
- **D** File storage

5 What does the 3-2-1 rule ask for?

- **A** 3 passwords, 2 users, 1 administrator
- **B** 3 copies of the data, on 2 types of media, with 1 copy off-site
- **C** 3 cloud providers, 2 regions, 1 data centre
- **D** Backups every 3 days, kept 2 weeks, tested once a year

6 A company keeps all its documents in a folder synced with a cloud drive and says it is backed up. Ransomware encrypts the laptop. Pick the two true statements.

- **A** The sync can copy the encrypted files to the cloud too, so it is not a real backup
- **B** The cloud provider is responsible for recovering the files under the shared responsibility model
- **C** A separate copy that cannot be changed, and a tested restore, would have protected it
- **D** It is safe because the cloud is always encrypted

7 A water company wants to know which pumps are likely to fail next month. Which stage of data science might it need that today's tourism question did not?

- **A** Asking the question
- **B** Cleaning
- **C** Modelling: training a model that predicts failures
- **D** Communicating

Your data now have a home and a backup plan. The last step of the unit is protecting them from people who should not get in, and knowing which laws require it: Cybersecurity and regulation.

Solutions

1. B · Everything else depends on the **question**: it decides which data to get, how to clean them and which chart answers it.

2. B · The yearly row is an **aggregate** of the monthly ones. Counting both doubles the total. Spotting aggregates is part of **cleaning**, stage 3.

3. B · A 0 is a real value and would pull totals and averages down. Leaving it **empty** and saying so is honest cleaning.

4. B · **Object storage** keeps files with their metadata in **buckets**. It is the usual place for images, logs and backups.

5. B · Three copies (the original and two backups), two different media, and one copy **off-site**, so that one fire, theft or ransomware attack cannot destroy all of them.

6. A, C · A synced folder mirrors **damage** as well as changes. Under the **shared responsibility model** the data are still the company's responsibility. A separate, ideally immutable, copy and a tested restore are what make a backup.

7. C · Today's question was **descriptive**, so no model was needed. Predicting failures is a **predictive** question, and that is where stage 5, the model, usually comes in.