# What Is Data Science?

Data is everywhere, and data science is the job of turning it into clear answers. What the job is, the four-step pipeline, and how a data scientist differs from a data analyst.

## Introduction

Look around you for one minute. Your phone knows how many steps you walked. A shop knows what it sold yesterday. A cricket app knows every ball bowled in a match. All of that is **data** — facts and numbers collected from real life. Data science is the work of taking that raw data and turning it into something a person can decide with.

In one line, the job is this: **find a question, get the data, clean it up, study it, and explain the answer in plain words.** A data scientist does not just make charts. Charts are only the last step. Most of the real work is earlier: finding the right data, fixing its mistakes, and asking the right question in the first place.

Here is a simple example. A school wants to know why marks fell in one class. The data — attendance, homework, test marks — already exists. A data scientist collects it, cleans it, looks for patterns, and then tells the story: "Marks fell most for students who missed more than five classes." That one sentence can change what the school does next.

Two things confuse beginners most, so let us clear them now. First, data science is not magic and it is not the same as artificial intelligence. Artificial intelligence is one tool a data scientist may use, but most data science is much simpler: counting, comparing, and explaining clearly. Second, a data scientist does not need huge data to start. A shop's one-month sales notebook is real data, and a good question about it is real data science. Start small, answer one question well, and the bigger tools will make sense later.

*(Roman Urdu: Data science ka matlab hai data se jawab nikalna — pehle data jama karo, phir usay saaf karo, phir us ka mutala karo, aur aakhir mein jawab saaf alfaaz mein batao.)*

## Key points

- **Data is recorded facts.** Numbers, text, dates, photos, clicks — anything that can be stored can be data.
- **The pipeline has four steps: collect** (gather the data), **clean** (fix mistakes and gaps), **analyse** (look for patterns and answers), and **tell the story** (explain the result so anyone can act on it).
- **A question comes first, data second.** "Why did sales fall?" is a good question. "Let us look at the data and see" usually wastes time, because data without a question has no direction.
- **Cleaning is the biggest part of the job.** Real data is messy: missing values, repeated rows, wrong spellings. Expect to spend more time cleaning than charting.
- **A data analyst and a data scientist overlap, but they differ.** An analyst usually answers questions about what already happened, with tables and reports. A data scientist also builds models that predict what may happen next, and works with larger, messier data.

- **You do not need fancy tools to start.** The thinking — question, clean, pattern, story — matters more than the tool. You can practise the whole pipeline with a notebook and a small table.
- **A good answer is honest about its limits.** "Students who missed classes scored lower" is a pattern. It does not prove missing class *caused* the low marks — other reasons may exist too.
- **A model learns patterns from past data.** Show it many past examples, and it guesses what comes next — a bit like learning from past exam papers. A model is only as good as the data it learned from.
- **Every step can be checked.** Someone else should be able to repeat your steps and get the same answer. If a step cannot be checked, it cannot be trusted.

## Examples

### Example 1 — The pipeline on a tiny shop problem

Question: why did a small shop sell fewer notebooks this week?

```
Step 1 – Collect:  sales list for this week and last week
Step 2 – Clean:    remove the double-entered sale, fix "Notbook" to "Notebook"
Step 3 – Analyse:  compare sales day by day, check the price change
Step 4 – Tell:     "Sales fell after the price rose from 50 to 80 rupees"
```

What this means: the four steps turn a pile of receipts into one decision — maybe lower the price again, or explain the rise to customers. Without the steps, the receipts are just paper.

### Example 2 — The same data, two different jobs

A school has one year of student marks.

An analyst's job:

```
Make a table: average marks per class, per subject, per term.
Report: "Class 9 average in maths fell from 72 to 64 in term 2."
```

A data scientist's job:

```
Use attendance, homework and past marks to predict
which students may struggle next term – so teachers
can help them BEFORE the exam, not after.
```

What this means: the analyst explains the past, the data scientist tries to help the future. Both use the same data. The difference is the question they are asked to answer.

### Example 3 — A bad question, a good question

Bad question:

```
What does the data say?
```

This is too wide. The data "says" thousands of things, and none of them helps you decide anything.

Good question:

```
Which subject had the biggest fall in average marks between term 1 and term 2, and for which
class?
```

This question has a clear answer you can compute, and a clear next step once you have it. Lesson DS.7 works a full project exactly like this, from question to findings.

### Example 4 — The pipeline on a cricket team problem

Question: which bowler is best in the last 10 overs of a match?

```
Step 1 — Collect:  ball-by-ball records of the last 5 matches
Step 2 — Clean:    remove the two overs with missing records,
                   fix one bowler's name spelled two ways
Step 3 — Analyse:  runs given per over in overs 41-50, bowler by bowler
Step 4 — Tell:     "Hasan gives away 6.2 runs per over at the death,
                   while the team average is 8.1 — bowl him last."
```

What this means: the same four steps work far from shops and schools. The question is about cricket, but the method is identical — and the answer is a team decision, not just a number.

### Example 5 — A price change, measured with two averages

A shop sells notebooks. Daily sales for five days before a price cut:

```
120, 135, 110, 140, 125  ->  total 630, average 126 per day
```

After the price cut, five more days:

```
150, 165, 140, 170, 155  ->  total 780, average 156 per day
```

What this means: sales rose by 30 notebooks a day on average after the cut. That is a real before-and-after comparison — the simplest kind of analysis, and often the most useful. A careful analyst would still ask: did anything else change those weeks, like a school opening nearby? The numbers alone cannot answer that, which is why the question and the limits matter as much as the maths.

### Example 6 — Real-world case: how a ride app sets your waiting time

When you open a ride-hailing app, it tells you "car in 4 minutes". That number is data science at work. The company collects past trips: where cars were, what time it was, how long pickups really took. It cleans the records (a trip with no end time is useless), studies the pattern (Friday evenings near a market are slow), and

then predicts your wait from the same pattern. If the prediction is wrong, riders complain — so the company keeps checking its predictions against real waits, and keeps improving them. The pipeline you learned in this lesson is the same pipeline, just running millions of times a day.

*(Roman Urdu: Ride app ka "4 minute" wala andaza bhi isi pipeline se banta hai — purani trips ka data, safai, pattern, phir naya andaza.)*

## Quick recap

- Data is recorded facts from real life.
- Data science = question, collect, clean, analyse, tell the story.
- Cleaning takes the most time, and skipping it ruins the result.
- An analyst explains the past; a data scientist also predicts the future.
- A good result ends in a plain sentence someone can act on.

## Common mistakes

1. **Jumping to charts before cleaning the data** — a chart of dirty data looks fine and says something wrong. Fix: clean first, chart second, always.
2. **Collecting data before choosing a question** — you end up with a huge pile and no direction. Fix: write the question in one sentence first, then collect only the data that answers it.
3. **Confusing a pattern with a cause** — two things moving together does not mean one caused the other. Fix: say "students who missed classes scored lower", not "missing classes caused low marks", unless you have proof.
4. **Thinking data science is only about fancy tools or AI** — beginners often learn tool names before they can ask a good question. Fix: practise the four steps on paper first; tools only speed up steps you already understand.
5. **Trusting one number without asking how much data sits behind it** — "average 90" from 2 students is not the same as "average 90" from 200 students. Fix: always say how many rows your answer comes from.

## Practice

1. Write one data-science question about your own class, shop, or cricket team. Keep it to one sentence.
2. List three sources of data that could answer it — for example, an attendance register, a sales notebook, or a match scorecard.
3. For the same dataset, write one analyst question (about the past) and one data scientist question (about the future).

1. Take the question you wrote in Practice 1 and run it through the four steps on paper: what would you collect, what would need cleaning, what would you compare, and what would your one-sentence answer be?

2. Find one number in a news story or a YouTube video (for example, "sales rose 40%"). Write down two questions you would ask before trusting it: how many rows, and what else changed at the same time?
3. Write one analyst question and one data scientist question about your own phone usage — for example, screen time last week (past) versus screen time you expect next week (future).

## Practice quiz

**1. What are the four steps of a data science pipeline, in order?**

- Tell, clean, collect, analyse
- Collect, clean, analyse, tell the story
- Analyse, collect, tell, clean
- Clean, tell, collect, analyse

**Answer:** Collect, clean, analyse, tell the story

**2. What is the difference between a data analyst and a data scientist?**

- There is none
- An analyst mainly explains what happened; a scientist also predicts and builds models
- Analysts only use Excel
- Scientists never clean data

**Answer:** An analyst mainly explains what happened; a scientist also predicts and builds models

**3. Where does data come from in daily life?**

- Only from computers
- From phones, shops, schools, hospitals — almost everything we do
- Only from the internet
- Only from banks

**Answer:** From phones, shops, schools, hospitals — almost everything we do

---

**Watch the video lesson:** https://youtu.be/kUCbp5xqlFY

**Read online:** https://scopetreasure.com/notes/ds-what-is-data-science · More free tools and notes at https://scopetreasure.com