IN THIS CHAPTER

» Understanding what generative AI means for data work

» Seeing how large language models change data analysis

» Distinguishing hype from reality in AI capabilities

» Getting hands-on with your first AI-assisted data task

Chapter 1

Doing Data Science in an AI World

Late 2022 and early 2023 felt like a turning point for professionals across all industries. ChatGPT burst onto the scene, and suddenly everyone working in data science had the same questions: "Is my job obsolete? Do years of learning mean nothing now?" Since then, the headlines have been all over the place: "Will AI Replace Data Analysts?" "Does ChatGPT Make Programmers Redundant?" "Data Science Is a Dead Career." The sense of excitement and panic has been palpable.

People who are new to data science are right to question whether there's a place for them. While data science has provided a solid career path since 2012, something fundamental has shifted with the advent of generative AI. The gentle stream of AI development that had been flowing since the 1950s suddenly turned into rapids. What once required teams of programmers, statisticians, and data engineers can now happen in a conversation with an AI assistant. The specialized, code-heavy discipline that took years to master has collided with a new reality, one where common AI tools make sophisticated analysis accessible to anyone who can ask good questions.

If you're wondering whether data science is dying because of AI, let me be direct with you now: it's not dying, it's evolving. And that evolution is creating career opportunities for people who know how to work with AI tools rather than against them (or in fear of them). The natural questions remain, however: Should I still study data science? Is coding necessary? What separates AI capabilities from what data scientists do on a daily basis?

The answers to these questions are worth a deeper examination. In this chapter, I look at exactly how generative AI has changed the data science game. I start you off with your first AI-assisted analysis. By the end of this chapter, you'll understand what AI-augmented data science means and how to start doing real analytical work today, regardless of your coding background.

Taking a High-Level View of Data Science and Generative AI

In its truest form, *data science* represents the optimization of processes and resources. Data science produces *data insights* — actionable, data-informed conclusions or predictions that you can use to understand and improve any quantified occurrence in the natural world. Using data science insights is like being able to see in the dark. For any goal or pursuit you can imagine, you can find data science methods to help you predict the most direct route from where you are to where you want to be — and to anticipate every pothole in the road between both places.

Capturing rapid-fire change across the data science landscape

The data science field is rapidly shifting from a coding-first discipline to a thinking-first discipline, where workers are expected to use AI tools to execute analytical requirements. This is what I call *AI-augmented data science*. Sophisticated data analysis no longer requires writing code, but it still demands critical thinking and knowledge of when AI is right and when it's hallucinating nonsense (again).

The good news for newbies is that the barrier to entry into data science has almost disappeared, but with this development, the need for analytical thinking and domain expertise has increased.

Placing AI, GenAI, machine learning, and LLMs within the conversation

When people say "AI," they're often referring to vastly different things. Traditional AI includes everything from phone autocomplete to Netflix recommendations. *Machine learning* is AI that makes predictions by learning from data rather than following pre-programmed rules. But *generative AI* is more niche: It involves AI systems that create new content — text, images, code, analysis — based on patterns they've learned from massive bodies of training data.

For data science work, the most relevant type of generative AI is *large language models*, or LLMs. These models power ChatGPT, Claude, Google's Gemini, and other similar tools. They've been trained using enormous collections of text — books, articles, websites, code repositories, and datasets — to understand and generate human-like text.

Seeing how LLMs work

Here's the simple version of how these models work: LLMs predict the next most likely token, a subword unit that may be a whole word, part of a word, or punctuation, based on patterns in their training data. Think of it like a sophisticated autocomplete application that learned from reading millions of books, articles, and datasets.

For example, when you type, "The capital of France is," your phone might suggest "Paris" because it's seen that pattern many times. LLMs work on the same principle, but are vastly more sophisticated. They don't "understand" in the human sense, but they've learned patterns so well that they can generate intelligent-sounding text, write working code, and perform complex analysis tasks.

Here are some key terms you'll need to know:

» **Large Language Models (LLMs):** These are AI systems that are trained using vast amounts of text and that can "understand" and generate human-like text. Examples include OpenAI's GPT models (for example, ChatGPT), Anthropic's Claude Sonnet (or alternate models Haiku, Sonnet, Opus), and Gemini LLM models. These are reasoning and text-generation engines with which most people are already familiar.

- **Generative AI:** This is the broader category of AI that creates new content rather than just classifying or predicting. Generative AI applications can generate text, images, videos, code, and music, but text and code capabilities matter the most for data work, of course.
- **Training data:** This is the massive collection of text, code, and information that LLMs learn patterns from. Think of it as the AI's "reading material" during its education — except that it reads from a huge fraction of the public internet, millions of books, and vast code repositories.
- **Tokens:** These are the building blocks LLMs work with — in some cases, these blocks are roughly equivalent to words or word parts. When people talk about "token limits," they're referring to how much content the AI can process or generate at a given time.
- **Hallucinations:** This is when AI confidently states false information as fact. This is critical to understand because AI can sound authoritative even when it is completely wrong.

WARNING

Hallucinations are most likely to occur when a model is asked to generate information that requires precise factual recall, multi-step numerical reasoning, extrapolation beyond the provided data, or knowledge that may have changed over time.

You've probably heard rumors that AI is going to replace data analysts, but what's actually happening is far more complex. Companies are using generative AI to augment data teams, but not to directly replace them. Ride-sharing companies like Uber use AI tools to help analysts understand ride patterns more quickly. Retailers use AI to analyze customer behavior and generate initial insights, which human analysts validate and act upon. Financial institutions use AI to spot fraud patterns, with human experts needing to make the final decisions. In every case, AI handles technical grunt work while humans provide judgment, domain expertise, and strategic thinking.

Understanding why ChatGPT can read your data

At this point, you may be asking yourself, "Why can ChatGPT work with data?" The thing is, generative AI tools have been trained on code, datasets, and countless analytical explanations. Because of this, they "understand" statistical concepts, they can write code in Python, R, SQL, and other languages, and they can interpret data patterns. Generative AI models have essentially "read" millions of examples of how people analyze data and of what good analysis looks like. However, the central fact remains that these models don't have real domain expertise or judgment, per se.

Recognizing what GenAI can and cannot do with data

Before getting started using generative AI tools for analytical tasks, it's important to understand exactly what the tools can and cannot do with data. Examples of this are shown in Figure 1-1.

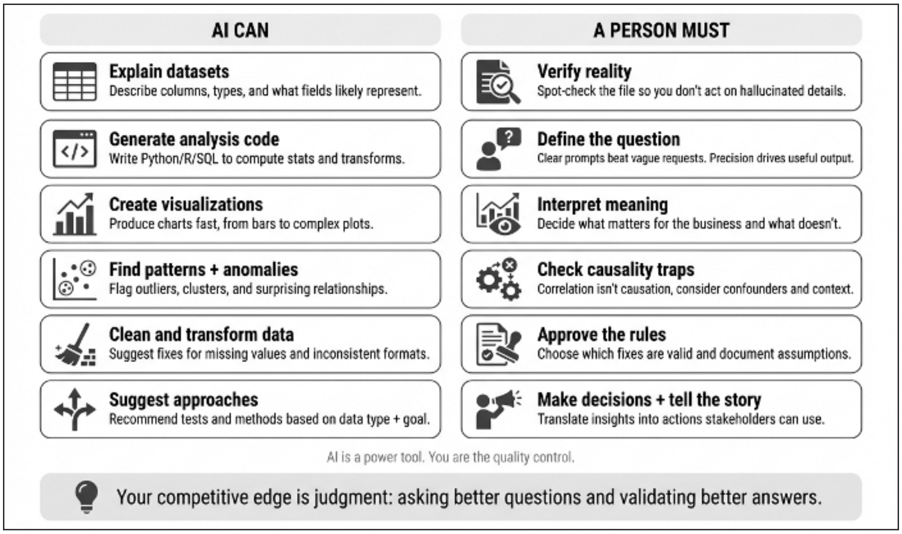

FIGURE 1-1: What GenAI can do versus what a person must do.

What does this mean for you if you want to work with data? Well, for starters, you no longer need to be a programmer first. You can now be an analytical thinker who uses AI tools effectively. The technical barrier that kept many capable people out of data science has dropped dramatically.

For many common business problems, you can do sophisticated analysis without writing code. You don't need to memorize statistical formulas or spend months learning database query languages. You can focus on what matters most: asking the right questions, understanding your business or domain, and making good data-based decisions.

You do, however, need to develop some critical thinking skills that are actually harder to pick up than coding. You need to know when AI is right, when it's hallucinating, and when you need to dig deeper. You need to understand enough about data and analysis to spot when AI gives you technically correct but practically meaningless results. These judgment skills take time to develop, but they're more valuable and less automatable than coding skills.

Who can do AI-augmented data science? You can. This democratization opens professional opportunities across roles. I'm talking about the following:

- Business analysts independently exploring sales trends and customer patterns
- Marketers independently understanding campaign performance and customer behavior
- Product managers analyzing user data and feature adoption
- Executives independently making data-informed decisions about strategy and investment
- Researchers working with datasets in any field
- Anyone with questions about data

You don't need to understand the complex mathematics behind LLMs to use them effectively for data work. What you do need, though, are critical thinking skills and the ability to verify AI outputs. Think of it like driving a car: you don't need to understand internal combustion engines to be a good driver, but you do need to know the rules of the road and when something doesn't sound right.

Seeing Who Data Science Is for and Why It's Changing

Data science used to be for people who had computer science degrees and strong programming skills. That's changing fast. The walls that kept data science separated are rapidly coming down, and the field is opening up to more people who can bring domain expertise, business acumen, and analytical thinking, even if they lack sophisticated coding experience.

Defining data science in plain English

In plain English, *data science* is the practice of extracting insights and knowledge from data to answer questions and solve problems. At its core, it's the practice of asking good questions, gathering relevant data, finding patterns, and communicating what those patterns mean. What does this actually look like in real life? Take the following example. Your company's sales dropped 15 percent last quarter, and leadership wants to know why. The data science process looks like this:

1. **Gather relevant data.** (Sales figures, customer information, marketing budget, economic indicators, competitor activities.)
2. **Explore it for patterns.** (Did all products decline equally? Did certain regions perform worse? Did customer types change?)
3. **Analyze what changed compared to previous quarters and test hypotheses about why.** (Was it seasonal? Did a competitor launch something new? Did a price change affect behavior?)
4. **Communicate findings in a way that leads to action.**

What's changed with the advent of generative AI is that each step can now be AI-augmented. You can ask ChatGPT to analyze your sales data, identify patterns, test statistical relationships, and suggest what might be driving the decline. Twenty years ago, this analysis required a team of statisticians and programmers working for weeks. Ten years ago, it may have taken one skilled data scientist working for days. Today, it's fully possible to get initial insights in minutes — then you spend your time validating those insights, digging deeper into interesting findings, and figuring out what actions to take.

The traditional path required you to first master advanced math, statistics, programming, databases, and visualization tools. With this knowledge, you were then safe and clear to finally do data analysis. This sequential approach meant years of studying technical prerequisites before you could even ask your first analytical question. The AI-augmented path inverts this. Now you can start with questions, use AI tools for technical execution, focus on interpretation and critical thinking, and then grasp technical details as you need them. These days, you can do real data analysis on day one, as you're about to see.

Exploring how GenAI is reshaping data analysis

How does generative AI reshape data analysis? By handling the coding grunt work. *Data cleaning* — finding missing values, standardizing formats, fixing inconsistencies — has traditionally been estimated to consume 60 to 80 percent of a data scientist's time (CrowdFlower, 2016). Now you can ask AI to "check this dataset for data quality issues and recommend fixes," review its assessment, and implement solutions in minutes rather than hours. *Data transformation* — reshaping data, creating calculated fields, joining multiple datasets — used to require careful coding. Now you can describe what you need in plain English and AI generates and executes the code.

This transformation frees humans up for doing what they're uniquely good at: strategic thinking. What questions should you ask? Do these results make business sense? What confounding factors might you be missing? What actions should follow from these insights? AI can help you answer these questions, but it can't (and shouldn't) answer them alone.

Throughout the pages of this book, you'll be honing the skills that matter most in today's market:

- **Asking good analytical questions:** Know what to ask. This matters more than knowing how to calculate answers.
- **Prompt engineering and context engineering:** With prompt engineering you tell AI what you need clearly and specifically. This is the new "coding" literacy. With context engineering, you provide the relevant background information, constraints, examples, and reference materials the AI needs to generate high-quality outputs on a consistent basis.
- **Critical evaluation:** Spot when AI gives you technically correct but practically wrong answers.
- **Domain knowledge and business context:** Understand your industry, organization, and problem space.
- **Communication and storytelling with data:** Translate insights into narratives that drive action.
- **Verification and quality checks:** Validate AI outputs before trusting them.

It's still imperative that you understand basic statistics so that you can spot when results don't make sense. For instance, you should know that if AI reports that a p-value of 0.8 is "highly significant," that's backwards. Knowing your business domain helps you ask better questions and recognize when insights are novel versus when they're obvious. Communication skills matter more than ever because you need to translate AI-generated insights for non-technical stakeholders in ways that drive action and advance your career.

Comparing the old way versus the AI-assisted way

The difference between the old and new approaches becomes clear when you look at specific tasks. For data cleaning, the old way involved writing Python code to identify missing values, writing more code to handle each missing data type, debugging when your code broke on unexpected data types, and documenting the process. This could require 2 to 4 hours for a moderately complex dataset. The AI-assisted way is simply to ask AI to "check this dataset for data quality issues and suggest solutions." Then you review its assessment in 2 minutes. Ask it to implement the fixes you agree with. Verify the results in a few minutes, and get it all done in a total time of 15 to 30 minutes. Those saved hours can go to thinking about what the clean data reveals.

For exploratory analysis, the old way required writing code to calculate summary statistics, creating multiple visualization scripts, and iterating through dozens of charts trying to find patterns. You'd spend hours coding before seeing any insights. The AI-assisted way: Ask AI to "explore this dataset and highlight interesting patterns." You get summary statistics, visualizations of key relationships, and flags for unusual patterns in seconds. Then you can ask follow-up questions about the findings and request specific visualizations to dig deeper. You spend minutes on technical execution and hours on interpretation and strategy.

The key difference here is that the old way required you to think in code first, and analysis second. Your mental energy went to syntax, debugging, and technical details before you could think about meaning. The AI-assisted way lets you think in questions and natural language. Your focus stays on what matters — the analytical logic, business implications, strategic decisions — while AI handles the technical translation.

WARNING

You still need to understand what good analysis looks like in order to evaluate whether an AI output is valid. Without knowing that correlation doesn't imply causation, you'll draw wrong conclusions from AI's correlation analysis. Without recognizing misleading scales, you'll misinterpret visualizations. AI handles execution, but you provide quality control.

When working with AI tools for data analysis, always start with a clear question. While AI is terrific for execution, it can't define good questions for you. The clearer your question — "What factors predict customer churn?" versus "Tell me about customers" — the better your AI analysis will be.

Getting Your Hands Dirty with Your First GenAI Data Task

Enough theory, let's actually do something with AI and data. This section shows you how easy it is to start analyzing data using ChatGPT with advanced data analysis capabilities. No coding is needed here, though I will show you what happens behind the scenes. Be excited because this isn't just practice — this is real data quality assessment, which is a fundamental first step in any data science project.

Uploading a sample CSV to ChatGPT Advanced Data Analysis

For this walk-through, I'm using the "human_capital_project.csv" dataset that contains human capital indicators as published by the United Nations.

ON THE WEB

You can download the `human_capital_project.csv` dataset from `businessgrowth.ai`. The principles demonstrated here apply to any dataset. You can also download sample datasets from `www.kaggle.com/datasets`. Just browse categories like Movies and TV Shows, Clothing and Accessories, or Economics and download anything that looks interesting. Open data from sites like `data.gov` also offers hundreds of thousands of datasets on crime, commerce, and public health. The cool thing is that these are real datasets that analysts and researchers actually use.

Inside ChatGPT, if you've got a paid plan, then you have built-in access to *Advanced Data Analysis* (ADA) — the feature that lets ChatGPT run calculations, clean data, and analyze the files that you upload (like CSVs and spreadsheets). Look for the file upload option located near the bottom of the chat interface (it's generally represented by a "+" icon). Click it to upload your file. ChatGPT accepts CSV, Excel (XLSX), JSON, and several other formats. Upload the file and wait for confirmation. The AI will typically acknowledge the upload and may give you a preview of what it sees.

WARNING

As with every tool, OpenAI's interface is subject to change. Be sure to check OpenAI's documentation if what you're seeing on your screen does not match that description above.

TIP

ADA works with multiple models, but tool access can vary by plan and workspace settings. In case you're wondering whether you have ADA access inside of ChatGPT, the quickest way to verify this is to start a new chat and look for the upload icon. If you can upload a CSV/XLSX file and ChatGPT can read it, you're most of the way there. Next, upload a small spreadsheet and enter a test prompt like "Calculate summary stats." If ADA is available, you'll see it do real analysis (tables, calculations, charts), and you'll usually see a "View analysis" option showing the work behind the answer. If you can upload files but then the model only replies in generic text with no calculations, no charts, and no analysis view, then you don't have ADA access and you should switch models and try again.

Asking AI to describe columns and assess data quality

With generative AI, what you ask matters a lot. Copy and paste this exact prompt into the chat:

"Please analyze this dataset. Tell me: 1) What columns are included and what they appear to represent, 2) Any data quality issues you notice such as missing values or inconsistencies, 3) Basic statistics about the data including row count and key summary statistics. Present your findings with a concise summary at the top, and then go into more detailed findings."

After allowing the AI to run, you should get an output similar to the one shown in Figure 1-2.

Concise summary

- **What this is:** A **monthly panel dataset** (long/tidy format) of **consumer price metrics** by geography, spanning **Jan 2000 → Jun 2025**.
- **Size: 190,332 rows × 8 columns**, covering **210 geographies** and **3 indicators.**
- **Biggest data quality note:** `Value` **has 4,767 missing values (2.50%)**, heavily concentrated in **"Food price inflation"** (5.07% missing), including a **structural-looking gap for all months in 2000** for most geographies (likely because inflation needs a prior-year baseline).
- **Distribution note:** `Value` is **extremely right-skewed** due to **hyperinflation outliers** (e.g., Venezuela CPI index values up to ~10^{12}), so **medians/percentiles are more informative than means.**

FIGURE 1-2: The results of a simple data quality assessment by ChatGPT.

As you can see, the AI produced a concise summary of the dataset and its quality issues, just like I requested. Not pictured in Figure 1-2 is the more detailed analysis that I requested, which was indeed returned by ChatGPT.

So, why did this prompt work well when others might fail? Well, it's specific about what you want, it covers the key aspects of data quality assessment, and it sets clear expectations for how the response should be formatted. Vague prompts like "tell me about this data" will yield vague, unfocused answers. Structured prompts like mine will guide AI to produce structured, useful output.

Interpreting the results AI gives you

Now here's where your work begins. You need to verify that this assessment isn't hallucinated. To do that, open the data file in Google Sheets or Excel and spot-check a few things. Check your file's row count (in Excel, you can use Ctrl+Down arrow to jump to the last row). Does the dataset actually have 8 columns and 190,832 rows? Does the summary that was produced by ChatGPT actually make sense? Are there lots of missing values when it comes to food price inflation? Do the summary statistics pass the "smell test"? If AI reports that Switzerland's economic data are extremely right-skewed because of hyperinflation, that should be an immediate red flag. Switzerland has one of the world's most stable economies and has not experienced hyperinflation in modern times. This suggests there may be issues in the underlying data or analysis that the model failed to identify. Are there red flags AI may have missed? Look at some actual data samples to see if there are issues AI didn't mention.

WARNING

This verification step is non-negotiable. AI tools can confidently describe datasets that don't actually match the files you uploaded, especially if the file format is ambiguous or corrupted. Always spot-check AI's descriptions against your actual data. This takes two minutes and can save you hours of working with wrong information.

Now it's time to congratulate yourself! You performed a data quality assessment, which is an important first step in any data science project. Professional data scientists do exactly this before they dive into deeper analysis. You've validated that your data is usable, identified issues that need fixing, and gained an initial understanding of what you're working with. That's real preparation for data science.

Taking a Peek at the Traditional Path

If AI can do all this, why would anyone study how to code? It's a valid question. A deeper understanding of the traditional path helps you evaluate when AI is doing good work, what its limitations are, and whether learning to code might be worth the time investment for your specific career goals.

Some people love coding and want to understand exactly what's happening under the hood. Others prefer the AI-assisted route and never want to touch code. Both paths are legitimate ways to do data science. There's no hierarchy here. People who use AI tools aren't "less than," and people who code aren't "more than." Different approaches fit different people and different situations. Understanding what code-focused data scientists do helps you make informed choices about your own professional development path.

Seeing what coders do

When Python users analyze data, they typically use a library called *pandas* — it's essentially a Python toolkit that makes it easy to work with data tables. Here's what simple data exploration code looks like:

```
df.info()  # Shows column names, data types, and non-null counts
df.describe()  # Provides summary statistics for numeric columns
df.isnull().sum()  # Counts missing values in each column
```

What does this code do? The first line displays information about your dataset's structure. The second line calculates summary statistics like mean, median, standard deviation, and quartiles for each numeric column. The third line counts how many values are missing in each column.

This is exactly what I just asked ChatGPT to do. The AI wrote and ran code like this behind the scenes. When you ask for a data quality assessment, AI generates similar pandas code, executes it, and then translates the results into plain English for you. The resulting numerical outputs should be identical; the difference is that you are thinking in questions while a coder would be thinking in syntax.

Identifying when you need to study coding

Coding skills offer tangible benefits. Here's my honest assessment of when learning to code actually matters for your career:

- Code is very beneficial when you're doing the same analysis repeatedly and you want to deploy with automation (think monthly reports, weekly data dumps, and routine quality checks).
- Coding matters when you need fine control that AI doesn't provide. This often shows up in the form of specific statistical techniques, custom algorithms, or precise outlier handling.
- Coding is valuable when you're working with huge datasets that AI tools can't handle — millions of rows that exceed token limits or cloud processing constraints.
- Coding is sometimes required when your organization mandates code for documentation and reproducibility, because some companies do require code-based workflows for auditability.
- Coding skills open doors when you want to advance to senior technical positions. The truth is that data science leadership roles still demand coding proficiency even if you're using AI for execution.

Conversely, AI-assisted approaches are usually enough when you're doing exploratory, one-off analyses and answering unique questions that won't come up repeatedly.

- They work fine when your questions change frequently.
- They're sufficient when you're working with moderate-sized datasets.
- They're perfectly acceptable when your organization accepts AI-assisted workflows.
- They're often better when your role is more strategic than technical.

My recommendation is to start with AI-assisted approaches. As you continue working with data, you'll naturally discover whether coding will help you. If you find yourself repeatedly explaining the same analysis to AI, that's a signal that automation through code may be valuable. If you're frustrated by AI's token limits or processing constraints, that's a sign you need to handle the bigger data yourself. If your organization requires code documentation, that makes learning it necessary for professional advancement.

REMEMBER

You can discover specific coding skills as you need them rather than trying to master programming before doing data analysis. Want to automate a weekly report? Master just enough Python and pandas to write that specific script (maybe 20 to 30 hours of focused learning). Need to work with databases? Master SQL basics for your specific use cases (another 20 to 30 hours). This targeted approach gets you productive fast while building technical skills where they matter most for your career development.

When choosing between AI-assisted data science or traditional coding, the goal should be to start doing data science in whatever way works best for you. AI tools lower the barrier to entry and allow you to focus on analytical thinking from day one. You can always add coding skills later if your work demands it. The important point is that you're asking good questions and that you're thinking critically about answers, whether those answers come from AI or from code you wrote yourself. That judgment is what makes you valuable in data science roles, not the specific tools you use to generate answers.