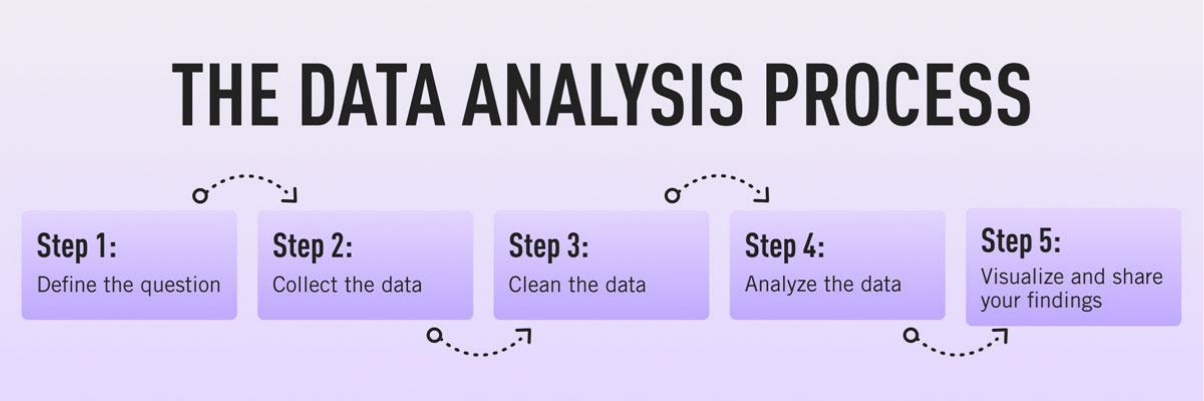
A Step-by-Step Guide to the Data Analysis Process

Like any scientific discipline, data analysis follows a rigorous step-by-step process. Each stage requires different skills and know-how. To get meaningful insights, though, it's important to understand the process as a whole. An underlying framework is invaluable for producing results that stand up to scrutiny.

In this post, we'll explore the main steps in the data analysis process. This will cover how to define your goal, collect data, and carry out an analysis. Where applicable, we'll also use examples and highlight a few tools to make the journey easier. When you're done, you'll have a much better understanding of the basics. This will help you tweak the process to fit your own needs.

Here are the steps we'll take you through:

On popular request, we've also developed a video based on this article. Scroll further along this article to watch that.



Ready? Let's get started with step one.

1. Step one: Defining the question

The first step in any data analysis process is to define your objective. In data analytics jargon, this is sometimes called the 'problem statement'.

Defining your objective means coming up with a hypothesis and figuring how to test it. Start by asking: What business problem am I trying to solve? While this might sound straightforward, it can be trickier than it seems. For instance, your organization's senior management might pose an issue, such as: "Why are we losing customers?" It's possible, though, that this doesn't get to the core of the problem. A data analyst's job is to understand the business and its goals in enough depth that they can frame the problem the right way.

Let's say you work for a fictional company called TopNotch Learning. TopNotch creates custom training software for its clients. While it is excellent at securing new clients, it has much lower repeat business. As such, your question might not be, "Why are we losing customers?" but, "Which factors are negatively impacting the customer experience?" or better yet: "How can we boost customer retention while minimizing costs?"

Now you've defined a problem, you need to determine which sources of data will best help you solve it. This is where your business acumen comes in again. For instance, perhaps you've noticed that the sales process for new clients is very slick, but that the production team is inefficient. Knowing this, you could hypothesize that the sales process wins lots of new clients, but the subsequent customer experience is lacking. Could this be why customers don't come back? Which sources of data will help you answer this question?

Tools to help define your objective

Defining your objective is mostly about soft skills, business knowledge, and lateral thinking. But you'll also need to keep track of business metrics and key performance indicators (KPIs). Monthly reports can allow you to track problem points in the business. Some KPI dashboards come with a fee, like Databox and DashThis. However, you'll also find open-source software like Grafana, Freeboard, and Dashbuilder. These are great for producing simple dashboards, both at the beginning and the end of the data analysis process.

2. Step two: Collecting the data

Once you've established your objective, you'll need to create a strategy for collecting and aggregating the appropriate data. A key part of this is determining which data you need. This might be quantitative (numeric) data, e.g. sales figures, or qualitative (descriptive) data, such as customer reviews. All data fit into one of three categories: first-party, second-party, and third-party data. Let's explore each one.

What is first-party data?

First-party data are data that you, or your company, have directly collected from customers. It might come in the form of transactional tracking data or information from your company's customer relationship management (CRM) system. Whatever its source, first-party data is usually structured and organized in a clear, defined way. Other sources of first-party data might include customer satisfaction surveys, focus groups, interviews, or direct observation.

What is second-party data?

To enrich your analysis, you might want to secure a secondary data source. Second-party data is the first-party data of other organizations. This might be available directly from the company or through a private marketplace. The main benefit of second-party data is that they are usually structured, and although they will be less relevant than first-party data, they also tend to be quite reliable. Examples of second-party data include website, app or social media activity, like online purchase histories, or shipping data.

What is third-party data?

Third-party data is data that has been collected and aggregated from numerous sources by a third-party organization. Often (though not always) third-party data contains a vast amount of unstructured data points (big data). Many organizations collect big data to create industry reports or to conduct market research. The research and advisory firm Gartner is a good real-world example of an organization that collects big data and sells it on to other companies

Tools to help you collect data

Once you've devised a data strategy (i.e. you've identified which data you need, and how best to go about collecting them) there are many tools you can use to help you. One thing you'll need, regardless of industry or area of expertise, is a data management platform (DMP). A DMP is a piece of software that allows you to identify and aggregate data from numerous sources, before manipulating them, segmenting them, and so on. There are many DMPs available. Some well-known enterprise DMPs include **Salesforce DMP**, **SAS**, and the data integration platform, **Xplenty**. If you want to play around, you can also try some open-source platforms like **Pimcore** or **D:Swarm**.

Want to learn more about what data analytics is and the process a data analyst follows? We cover this topic (and more) in our free introductory short course for beginners.

3. Step three: Cleaning the data

Once you've collected your data, the next step is to get it ready for analysis. This means cleaning, or 'scrubbing' it, and is crucial in making sure that you're working with **high-quality data**. Key data cleaning tasks include:

- **Removing major errors, duplicates, and outliers**—all of which are inevitable problems when aggregating data from numerous sources.
- **Removing unwanted data points**—extracting irrelevant observations that have no bearing on your intended analysis.
- **Bringing structure to your data**—general 'housekeeping', i.e. fixing typos or layout issues, which will help you map and manipulate your data more easily.
- **Filling in major gaps**—as you're tidying up, you might notice that important data are missing. Once you've identified gaps, you can go about filling them.

A good data analyst will spend around 70-90% of their time cleaning their data. This might sound excessive. But focusing on the wrong data points (or analyzing erroneous data) will severely impact your results. It might even send you back to square one...so don't rush it

Carrying out an exploratory analysis

Another thing many data analysts do (alongside cleaning data) is to carry out an exploratory analysis. This helps identify initial trends and characteristics, and can even refine your hypothesis. Let's use our fictional learning company as an example again. Carrying out an exploratory analysis, perhaps you notice a correlation between how much TopNotch Learning's clients pay and how quickly they move on to new suppliers. This might suggest that a lowquality customer experience (the assumption in your initial hypothesis) is actually less of an issue than cost. You might, therefore, take this into account.

Tools to help you clean your data

Cleaning datasets manually—especially large ones—can be daunting. Luckily, there are many tools available to streamline the process. Open-source tools, such as **OpenRefine**, are excellent for basic data cleaning, as well as high-level exploration. However, free tools offer limited functionality for very large datasets. Python libraries (e.g. Pandas) and some R packages are better suited for heavy data scrubbing. You will, of course, need to be familiar with the languages. Alternatively, enterprise tools are also available. For example, **Data Ladder**, which is one of the highest-rated data-matching tools in the industry. There are many more. Why not see which free data cleaning tools you can find to play around with?

Curious about a career in Data Analytics?

4. Step four: Analyzing the data

Finally, you've cleaned your data. Now comes the fun bit—analyzing it! The type of data analysis you carry out largely depends on what your goal is. But there are many techniques available. Univariate or bivariate analysis, time-series analysis, and regression analysis are just a few you might have heard of. More important than the different types, though, is how you apply them. This depends on what insights you're hoping to gain. Broadly speaking, all types of data analysis fit into one of the following four categories.

Descriptive analysis

Descriptive analysis identifies what has already happened. It is a common first step that companies carry out before proceeding with deeper explorations. As an example, let's refer back to our fictional learning provider once more. TopNotch Learning might use descriptive analytics to analyze course completion rates for their customers. Or they might identify how many users access their products during a particular period. Perhaps they'll use it to measure sales figures over the last five years. While the company might not draw firm conclusions from any of these insights, summarizing and describing the data will help them to determine how to proceed.

Diagnostic analysis

Diagnostic analytics focuses on understanding why something has happened. It is literally the diagnosis of a problem, just as a doctor uses a patient's symptoms to diagnose a disease. Remember TopNotch Learning's business problem? 'Which factors are negatively impacting the customer experience?' A diagnostic analysis would help answer this. For instance, it could help the company draw correlations between the issue (struggling to gain repeat business) and factors that might be causing it (e.g. project costs, speed of delivery, customer sector, etc.) Let's imagine that, using diagnostic analytics, TopNotch realizes its clients in the retail sector are departing at a faster rate than other clients. This might suggest that they're losing customers because they lack expertise in this sector. And that's a useful insight!

Predictive analysis

Predictive analysis allows you to identify future trends based on historical data. In business, predictive analysis is commonly used to forecast future growth, for example. But it doesn't stop there. Predictive analysis has grown increasingly sophisticated in recent years. The speedy evolution of machine learning allows organizations to make surprisingly accurate forecasts. Take the insurance industry. Insurance providers commonly use past data to predict which customer groups are more likely to get into accidents. As a result, they'll hike up customer insurance premiums for those groups. Likewise, the retail industry often uses transaction data to predict where future trends lie, or to determine seasonal buying habits to inform their strategies. These are just a few simple examples, but the untapped potential of predictive analysis is pretty compelling.

Prescriptive analysis

Prescriptive analysis allows you to make recommendations for the future. This is the final step in the analytics part of the process. It's also the most complex. This is because it incorporates aspects of all the other analyses we've described. A great example of prescriptive analytics is the algorithms that guide Google's self-driving cars. Every second, these algorithms make countless decisions based on past and present data, ensuring a smooth, safe ride. Prescriptive analytics also helps companies decide on new products or areas of business to invest in.

5. Step five: Sharing your results

You've finished carrying out your analyses. You have your insights. The final step of the data analytics process is to share these insights with the wider world (or at least with your organization's stakeholders!) This is more complex than simply sharing the raw results of your work—it involves interpreting the outcomes, and presenting them in a manner that's digestible for all types of audiences. Since you'll often present information to decision-makers, it's very important that the insights you present are 100% clear and unambiguous. For this reason, data analysts commonly use reports, dashboards, and interactive visualizations to support their findings.

How you interpret and present results will often influence the direction of a business. Depending on what you share, your organization might decide to restructure, to launch a high-risk product, or even to close an entire division. That's why it's very important to provide all the evidence that you've gathered, and not to cherry-pick data. Ensuring that you cover everything in a clear, concise way will prove that your conclusions are scientifically sound and based on the facts. On the flip side, it's important to highlight any gaps in the data or to flag any insights that might be open to interpretation. Honest communication is the most important part of the process. It will help the business, while also helping you to excel at your job!

Tools for interpreting and sharing your findings

There are tons of **data visualization tools** available, suited to different experience levels. Popular tools requiring little or no coding skills include **Google Charts**, **Tableau**, **Datawrapper**, and **Infogram**. If you're familiar with Python and R, there are also many data visualization libraries and packages available. For instance, check out the Python libraries **Plotly**, **Seaborn**, and **Matplotlib**. Whichever data visualization tools you use, make sure you polish up your presentation skills, too. Remember: Visualization is great, but communication is key!

7. Summary

In this post, we've covered the main steps of the data analytics process. These core steps can be amended, re-ordered and re-used as you deem fit, but they underpin every data analyst's work:

- **Define the question**—What business problem are you trying to solve? Frame it as a question to help you focus on finding a clear answer.
- **Collect data**—Create a strategy for collecting data. Which data sources are most likely to help you solve your business problem?
- **Clean the data**—Explore, scrub, tidy, de-dupe, and structure your data as needed. Do whatever you have to! But don't rush…take your time!
- **Analyze the data**—Carry out various analyses to obtain insights. Focus on the four types of data analysis: descriptive, diagnostic, predictive, and prescriptive.
- **Share your results**—How best can you share your insights and recommendations? A combination of visualization tools and communication is key.