**COURSE: 01**

The six steps of the data analysis process that you have been learning in this program are: **ask, prepare, process, analyze, share,** and **act**. These six steps apply to any data analysis.

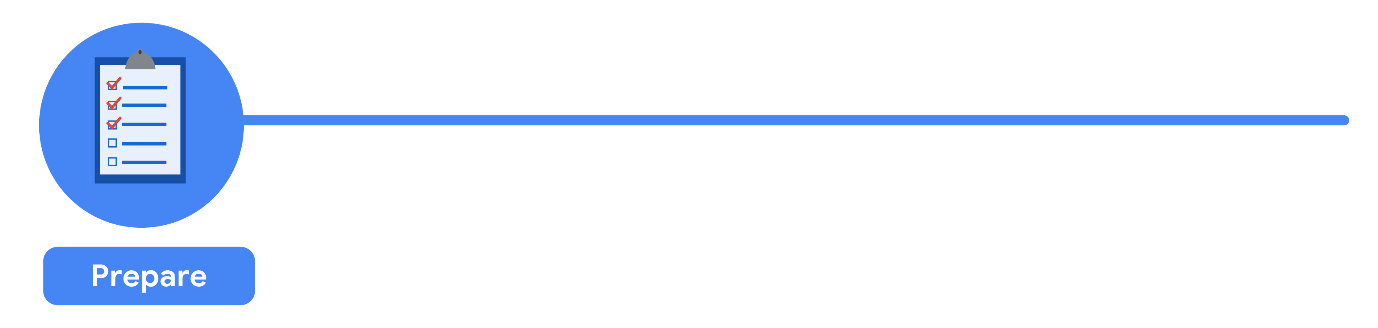
An organization was experiencing a high turnover rate among new hires. Many employees left the company before the end of their first year on the job. The analysts used the data analysis process to answer the following question: **how can the organization improve the retention rate for new employees?**

Here is a break down of what this team did, step by step.



First up, the analysts needed to define what the project would look like and what would qualify as a successful result. So, to determine these things, they **asked** effective questions and collaborated with leaders and managers who were interested in the outcome of their people analysis. These were the kinds of questions they asked:

* What do you think new employees need to learn to be successful in their first year on the job?
* Have you gathered data from new employees before? If so, may we have access to the historical data?
* Do you believe managers with higher retention rates offer new employees something extra or unique?
* What do you suspect is a leading cause of dissatisfaction among new employees?
* By what percentage would you like employee retention to increase in the next fiscal year?



It all started with solid **preparation**. The group built a timeline of three months and decided how they wanted to relay their progress to interested parties. Also during this step, the analysts identified what data they needed to achieve the successful result they identified in the previous step - in this case, the analysts chose to gather the data from an online survey of new employees. These were the things they did to prepare:

* They developed specific questions to ask about employee satisfaction with different business processes, such as hiring and onboarding, and their overall compensation.
* They established rules for who would have access to the data collected - in this case, anyone outside the group wouldn't have access to the raw data, but could view summarized or aggregated data. For example, an individual's compensation wouldn't be available, but salary ranges for groups of individuals would be viewable.
* They finalized what specific information would be gathered, and how best to present the data visually. The analysts brainstormed possible project- and data-related issues and how to avoid them.



The group sent the survey out. Great analysts know how to respect both their data and the people who provide it. Since employees provided the data, it was important to make sure all employees gave their consent to participate. The data analysts also made sure employees understood how their data would be **collected, stored, managed, and protected**. Collecting and using data ethically is one of the responsibilities of data analysts. In order to maintain confidentiality and protect and store the data effectively, these were the steps they took:

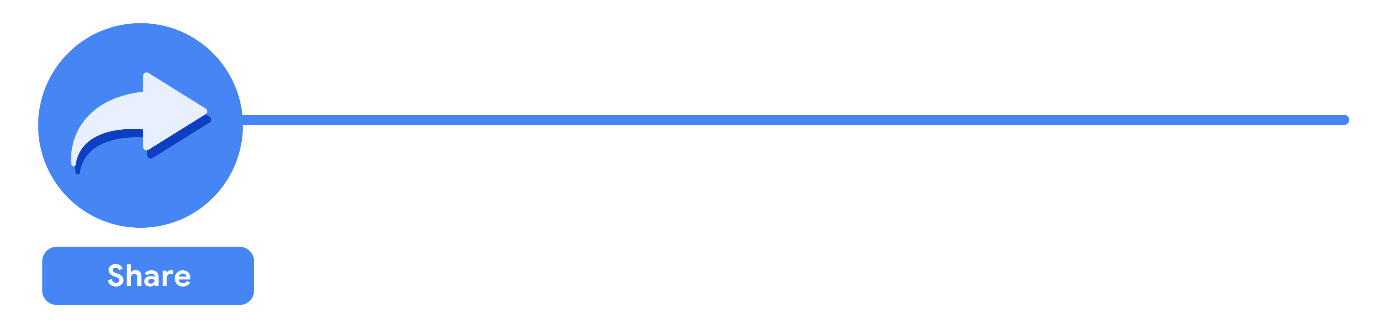
* They restricted access to the data to a limited number of analysts.
* They cleaned the data to make sure it was complete, correct, and relevant. Certain data was aggregated and summarized without revealing individual responses.
* They uploaded raw data to an internal data warehouse for an additional layer of security.



Then, the analysts did what they do best: analyze! From the completed surveys, the data analysts **discovered** that an employee’s experience with certain processes was a key indicator of overall job satisfaction. These were their findings:

* Employees who experienced a long and complicated hiring process were most likely to leave the company.
* Employees who experienced an efficient and transparent evaluation and feedback process were most likely to remain with the company.

The group knew it was important to **document** exactly what they found in the analysis, no matter what the results. To do otherwise would diminish trust in the survey process and reduce their ability to collect truthful data from employees in the future.



Just as they made sure the data was carefully protected, the analysts were also careful **sharing the report**. This is how they shared their findings:

* They shared the report with managers who met or exceeded the minimum number of direct reports with submitted responses to the survey.
* They presented the results to the managers to make sure they had the full picture.
* They asked the managers to personally deliver the results to their teams.

This process gave managers an opportunity to **communicate the results** with the right context. As a result, they could have productive team conversations about next steps to improve employee engagement.



The last stage of the process for the team of analysts was to work with leaders within their company and decide how best to **implement changes and take actions** based on the findings. These were their recommendations:

* Standardize the hiring and evaluation process for employees based on the most efficient and transparent practices.
* Conduct the same survey annually and compare results with those from the previous year.

A year later, the same survey was distributed to employees. Analysts anticipated that a comparison between the two sets of results would indicate that the action plan worked. Turns out, the changes improved the retention rate for new employees and the actions taken by leaders were successful!

**Is people analytics right for you?**

One of the many things that makes data analytics so exciting is that the problems are always different, the solutions need creativity, and the impact on others can be great — even life-changing or life-saving. As a data analyst, you can be part of these efforts. Maybe you’re even inspired to learn more about the field of people analytics. If so, consider learning more about this field and adding that research to your data analytics journal. You never know: One day soon, you could be helping a company create an amazing work environment for you and your colleagues!

Before you write your entry in your learning log, reflect on [the case study](https://www.coursera.org/learn/foundations-data/supplement/nhC19/case-study-new-data-perspectives) from earlier. The data analysts wanted to use data to improve employee retention. In order to do that, they had to break this larger project into manageable tasks. The analysts organized those tasks and activities around the six phases of the data analysis process:

1. Ask
2. Prepare
3. Process
4. Analyze
5. Share
6. Act

The analysts **asked** questions to define both the issue to be solved and what would equal a successful result. Next, they **prepared** by building a timeline and collecting data with employee surveys that were designed to be inclusive. They **processed** the data by cleaning it to make sure it was complete, correct, relevant, and free of errors and outliers. They **analyzed** the clean employee survey data. Then the analysts **shared** their findings and recommendations with team leaders. Afterward, leadership **acted** on the results and focused on improving key areas.

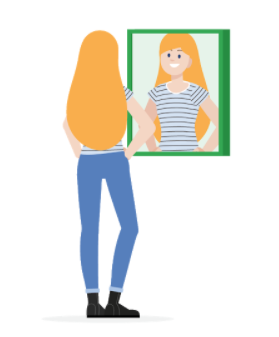
# Data and gut instinct

Detectives and data analysts have a lot in common. Both depend on facts and clues to make decisions. Both collect and look at the evidence. Both talk to people who know part of the story. And both might even follow some footprints to see where they lead. Whether you’re a detective or a data analyst, your job is all about following steps to collect and understand facts.

Analysts use data-driven decision-making and follow a step-by-step process. You have learned that there are six steps to this process:

1. **Ask** questions and define the problem.
2. **Prepare** data by collecting and storing the information.
3. **Process** data by cleaning and checking the information.
4. **Analyze** data to find patterns, relationships, and trends.
5. **Share** data with your audience.
6. **Act** on the data and use the analysis results.

But there are other factors that influence the decision-making process. You may have read mysteries where the detective used their gut instinct, and followed a hunch that helped them solve the case. **Gut instinct** is an intuitive understanding of something with little or no explanation. This isn’t always something conscious; we often pick up on signals without even realizing. You just have a “feeling” it’s right.



## Why gut instinct can be a problem

At the heart of data-driven decision making is data. Therefore, it's essential that data analysts focus on the data to ensure they make informed decisions. If you ignore data by preferring to make decisions based on your own experience, your decisions may be biased. But even worse, decisions based on gut instinct without any data to back them up can cause mistakes.

Consider an example of a restaurant entrepreneur, partnering with a well known chef to develop a new restaurant in a bustling part of the city’s central shopping district. The well known chef has several restaurants across the city. Banking on their reputation, the restaurant entrepreneur and chef followed gut instinct and created another uniquely themed restaurant. However, fundraising efforts fell short to fund the opening of the restaurant after months of planning and preparation. The property will go back on the market to be sold at a loss. Had the entrepreneur done more research, they would've found data showing prospective customers in this new restaurant location were very different from the chef's other restaurants.

The more you understand the data related to a project, the easier it will be to figure out what is required. These efforts will also help you identify errors and gaps in your data so you can communicate your findings more effectively. Sometimes past experience helps you make a connection that no one else would notice. For example, a detective might be able to crack open a case because they remember an old case just like the one they’re solving today. It's not just gut instinct.

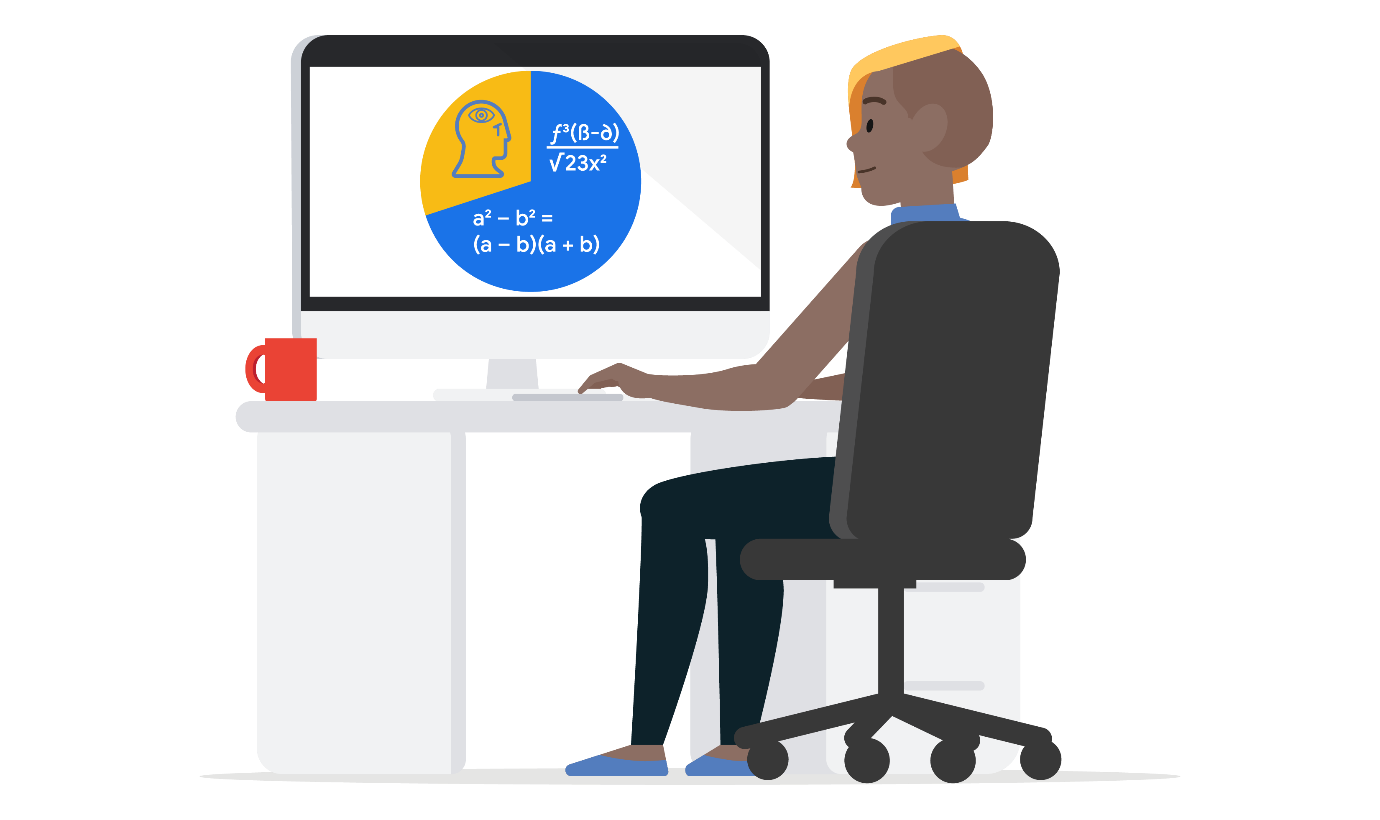
## Data + business knowledge = mystery solved

Blending data with business knowledge, plus maybe a touch of gut instinct, will be a common part of your process as a junior data analyst. The key is figuring out the exact mix for each particular project. A lot of times, it will depend on the goals of your analysis. That is why analysts often ask, “How do I define success for this project?”

In addition, try asking yourself these questions about a project to help find the perfect balance:

* What kind of results are needed?
* Who will be informed?
* Am I answering the question being asked?
* How quickly does a decision need to be made?

For instance, if you are working on a rush project, you might need to rely on your own knowledge and experience more than usual. There just isn’t enough time to thoroughly analyze all of the available data. But if you get a project that involves plenty of time and resources, then the best strategy is to be more data-driven. It’s up to you, the data analyst, to make the best possible choice. You will probably blend data and knowledge a million different ways over the course of your data analytics career. And the more you practice, the better you will get at finding that perfect blend.



You have already learned about the five essential aspects of analytical skills: **curiosity, understanding context, having a technical mindset, data design, and data strategy.** You have also discovered that you’re already practicing these skills. Now, you’ll complete an entry in your learning log exploring your own analytical strengths and weaknesses and your goals for the future. By the time you complete this activity, you will have a stronger understanding of your analytical skill set and how you can practice and improve them. These analytical skills are key to helping you solve problems and create insights using data analysis. Thinking about them now will help you grow as a data analyst!

**The analytical skills table**



First, you’ll fill out an Analytical Skills Table in your learning log entry. The table will appear like this in the template:

Analytical skill column: -Curiosity -Context -Technical mindset -Data design -Data strategy

The table has a row for each essential aspect of analytical skills:

* **Curiosity:** a desire to know more about something, asking the right questions
* **Understanding context:** understanding where information fits into the “big picture”
* **Having a technical mindset:** breaking big things into smaller steps
* **Data design:** thinking about how to organize data and information
* **Data strategy:** thinking about the people, processes, and tools used in data analysis

You will put an X in the column that you think best describes your current level with each aspect. The three ratings are:

* **Strength:** This is an area you feel is one of your strengths
* **Developing:** You have some experience with this area, but there’s still significant room for growth
* **Emerging:** This is new to you, and will gain experience in this area from this course

Then update the Comments/Plans/Goals column with a quick note to yourself about why you chose those ratings.

# Variations of the data life cycle

You learned that there are six stages to the data life cycle. Here is a recap:

1. **Plan:** Decide what kind of data is needed, how it will be managed, and who will be responsible for it.
2. **Capture:** Collect or bring in data from a variety of different sources.
3. **Manage:** Care for and maintain the data. This includes determining how and where it is stored and the tools used to do so.
4. **Analyze:** Use the data to solve problems, make decisions, and support business goals.
5. **Archive:** Keep relevant data stored for long-term and future reference.
6. **Destroy:** Remove data from storage and delete any shared copies of the data.

**Warning:** Be careful not to mix up or confuse the six stages of the data life cycle (Plan, Capture, Manage, Analyze, Archive, and Destroy) with the six phases of the data analysis life cycle (Ask, Prepare, Process, Analyze, Share, and Act). They shouldn't be used or referred to interchangeably.

The data life cycle provides a generic or common framework for how data is managed. You may recall that variations of the data analysis life cycle were described in [Origins of the data analysis process](https://www.coursera.org/learn/foundations-data/supplement/WWlrt/origins-of-the-data-analysis-process). The same can be done for the data life cycle. The rest of this reading provides a glimpse of how government, finance, and education institutions can view data life cycles a little differently.

## U.S. Fish and Wildlife Service

The U.S. Fish and Wildlife Service uses the following data life cycle:

1. Plan
2. Acquire
3. Maintain
4. Access
5. Evaluate
6. Archive

For more information, refer to [U.S. Fish and Wildlife's Data Management Life Cycle](https://www.fws.gov/data/life-cycle) page.

## The U.S. Geological Survey (USGS)

The USGS uses the data life cycle below:

1. Plan
2. Acquire
3. Process
4. Analyze
5. Preserve
6. Publish/Share

Several cross-cutting or overarching activities are also performed during each stage of their life cycle:

* Describe (metadata and documentation)
* Manage Quality
* Backup and Secure

For more information, refer to the [USGS Data Lifecycle](https://www.usgs.gov/products/data-and-tools/data-management/data-lifecycle) page.

## Financial institutions

Financial institutions may take a slightly different approach to the data life cycle as described in [The Data Life Cycle](https://sfmagazine.com/post-entry/july-2018-the-data-life-cycle/), an article in Strategic Finance magazine:

1. Capture
2. Qualify
3. Transform
4. Utilize
5. Report
6. Archive
7. Purge

## Harvard Business School (HBS)

One final data life cycle informed by Harvard University research has eight stages:

1. Generation
2. Collection
3. Processing
4. Storage
5. Management
6. Analysis
7. Visualization
8. Interpretation

For more information, refer to [8 Steps in the Data Life Cycle](https://online.hbs.edu/blog/post/data-life-cycle).

# Key data analyst tools

As you are learning, the most common programs and solutions used by data analysts include spreadsheets, query languages, and visualization tools. In this reading, you will learn more about each one. You will cover when to use them, and why they are so important in data analytics.



## Spreadsheets

Data analysts rely on spreadsheets to collect and organize data. Two popular spreadsheet applications you will probably use a lot in your future role as a data analyst are Microsoft Excel and Google Sheets.

Spreadsheets structure data in a meaningful way by letting you

* Collect, store, organize, and sort information
* Identify patterns and piece the data together in a way that works for each specific data project
* Create excellent data visualizations, like graphs and charts.

## Databases and query languages

A database is a collection of structured data stored in a computer system. Some popular Structured Query Language (SQL) programs include MySQL, Microsoft SQL Server, and BigQuery.

Query languages

* Allow analysts to isolate specific information from a database(s)
* Make it easier for you to learn and understand the requests made to databases
* Allow analysts to select, create, add, or download data from a database for analysis

## Visualization tools

Data analysts use a number of visualization tools, like graphs, maps, tables, charts, and more. Two popular visualization tools are Tableau and Looker.

These tools

* Turn complex numbers into a story that people can understand
* Help stakeholders come up with conclusions that lead to informed decisions and effective business strategies
* Have multiple features

- **Tableau**'s simple drag-and-drop feature lets users create interactive graphs in dashboards and

worksheets

- **Looker** communicates directly with a database, allowing you to connect your data right to the visual

tool you choose

A career as a data analyst also involves using programming languages, like R and Python, which are used a lot for statistical analysis, visualization, and other data analysis.

# Choosing the right tool for the job

As a data analyst, you will usually have to decide which program or solution is right for the particular project you are working on. In this reading, you will learn more about how to choose which tool you need and when.

Depending on which phase of the data analysis process you’re in, you will need to use different tools. For example, if you are focusing on creating complex and eye-catching visualizations, then the visualization tools we discussed earlier are the best choice. But if you are focusing on organizing, cleaning, and analyzing data, then you will probably be choosing between spreadsheets and databases using queries. Spreadsheets and databases both offer ways to store, manage, and use data. The basic content for both tools are sets of values. Yet, there are some key differences, too:

| **Spreadsheets** | **Databases** |
| --- | --- |
| Software applications | Data stores - accessed using a query language (e.g. SQL) |
| Structure data in a row and column format | Structure data using rules and relationships |
| Organize information in cells | Organize information in complex collections |
| Provide access to a limited amount of data | Provide access to huge amounts of data |
| Manual data entry | Strict and consistent data entry |
| Generally one user at a time | Multiple users |
| Controlled by the user | Controlled by a database management system |

You don’t have to choose one or the other because each serves its own purpose. Generally, data analysts work with a combination of the two, as both tools are very useful in data analytics. For example, you can store data in a database, then export it to a spreadsheet for analysis. Or, if you are collecting information in a spreadsheet, and it becomes too much for that particular platform, you can import it into a database. And, later in this course, you will learn about programming languages like R that give you even greater control of your data, its analysis, and the visualizations you create.

# Planning a data visualization

Earlier, you learned that **data visualization** is the graphical representation of information. As a data analyst, you will want to create visualizations that make your data easy to understand and interesting to look at. Because of the importance of data visualization, most data analytics tools (such as spreadsheets and databases) have a built-in visualization component while others (such as Tableau) specialize in visualization as their primary value-add. In this reading, you will explore the steps involved in the data visualization process and a few of the most common data visualization tools available.



## Steps to plan a data visualization

Let’s go through an example of a real-life situation where a data analyst might need to create a data visualization to share with stakeholders. Imagine you’re a data analyst for a clothing distributor. The company helps small clothing stores manage their inventory, and sales are booming. One day, you learn that your company is getting ready to make a major update to its website. To guide decisions for the website update, you’re asked to analyze data from the existing website and sales records. Let’s go through the steps you might follow.

### Step 1: Explore the data for patterns

First, you ask your manager or the data owner for access to the current sales records and website analytics reports. This includes information about how customers behave on the company’s existing website, basic information about who visited, who bought from the company, and how much they bought.

While reviewing the data you notice a pattern among those who visit the company’s website most frequently: geography and larger amounts spent on purchases. With further analysis, this information might explain why sales are so strong right now in the northeast—and help your company find ways to make them even stronger through the new website.

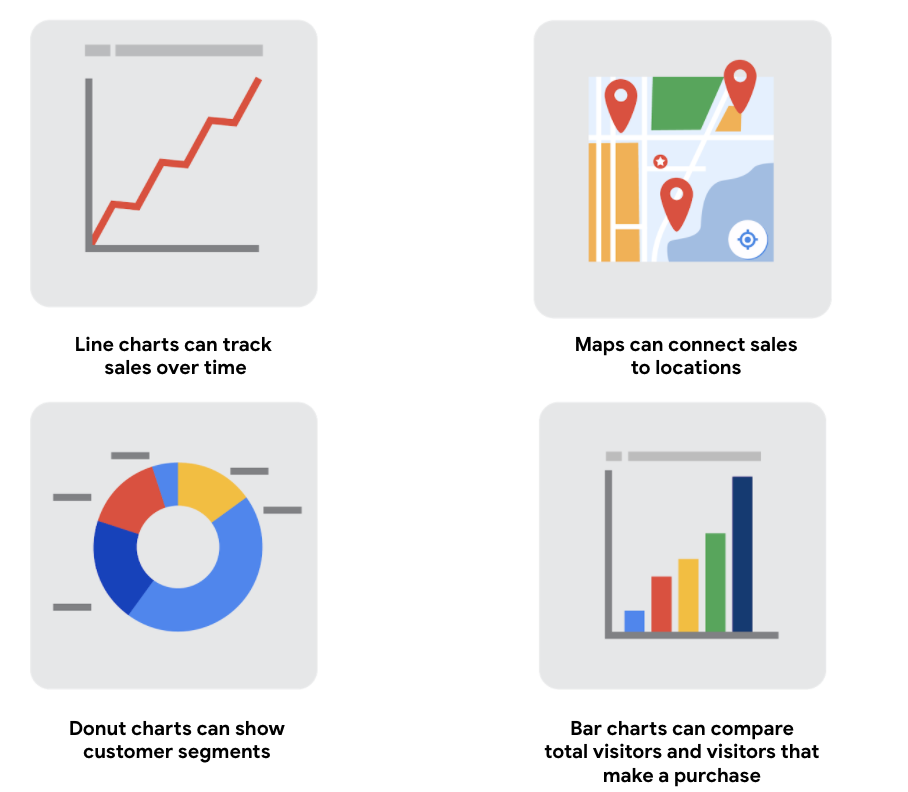
### Step 2: Plan your visuals

Next it is time to refine the data and present the results of your analysis. Right now, you have a lot of data spread across several different tables, which isn’t an ideal way to share your results with management and the marketing team. You will want to create a data visualization that explains your findings quickly and effectively to your target audience. Since you know your audience is sales oriented, you already know that the data visualization you use should:

* Show sales numbers over time
* Connect sales to location
* Show the relationship between sales and website use
* Show which customers fuel growth

### Step 3: Create your visuals

Now that you have decided what kind of information and insights you want to display, it is time to start creating the actual visualizations. Keep in mind that creating the right visualization for a presentation or to share with stakeholders is a process. It involves trying different visualization formats and making adjustments until you get what you are looking for. In this case, a mix of different visuals will best communicate your findings and turn your analysis into the most compelling story for stakeholders. So, you can use the built-in chart capabilities in your spreadsheets to organize the data and create your visuals.

1) line charts can track sales over time 2) maps can connect sales to locations 3) donut charts can show customer segments 4) bar charts can compare total visitors that make a purchase

## Build your data visualization toolkit

There are many different tools you can use for data visualization.

* You can use the visualizations tools in your spreadsheet to create simple visualizations such as line and bar charts.
* You can use more advanced tools such as Tableau that allow you to integrate data into dashboard-style visualizations.
* If you’re working with the programming language R you can use the visualization tools in RStudio.

Your choice of visualization will be driven by a variety of drivers including the size of your data, the process you used for analyzing your data (spreadsheet, or databases/queries, or programming languages). For now, just consider the basics.

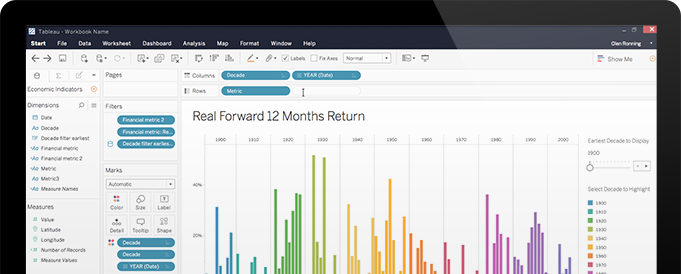
## Spreadsheets (Microsoft Excel or Google Sheets)

In our example, the built-in charts and graphs in spreadsheets made the process of creating visuals quick and easy. Spreadsheets are great for creating simple visualizations like bar graphs and pie charts, and even provide some advanced visualizations like maps, and waterfall and funnel diagrams (shown in the following figures).

But sometimes you need a more powerful tool to truly bring your data to life. Tableau and RStudio are two examples of widely used platforms that can help you plan, create, and present effective and compelling data visualizations.

## Visualization software (Tableau)

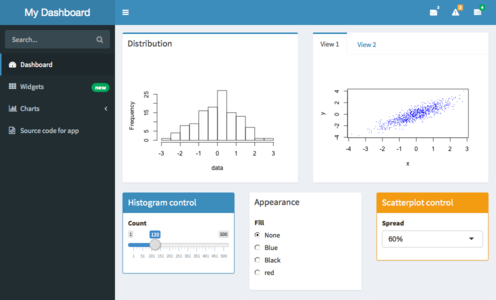
Tableau is a popular data visualization tool that lets you pull data from nearly any system and turn it into compelling visuals or actionable insights. The platform offers built-in visual best practices, which makes analyzing and sharing data fast, easy, and (most importantly) useful. Tableau works well with a wide variety of data and includes an interactive dashboard that lets you and your stakeholders click to explore the data interactively.



You can start exploring Tableau from the [How-to Video](https://public.tableau.com/en-us/s/resources) resources. Tableau Public is free, easy to use, and full of helpful information. The Resources page is a one-stop-shop for how-to videos, examples, and datasets for you to practice with. To explore what other data analysts are sharing on Tableau, visit the [Viz of the Day](https://public.tableau.com/en-us/gallery/?tab=viz-of-the-day&type=viz-of-the-day) page where you will find beautiful visuals ranging from the [Hunt for (Habitable) Planets](https://public.tableau.com/en-us/gallery/hunt-habitable-planets?tab=viz-of-the-day&type=viz-of-the-day) to [Who’s Talking in Popular Films](https://public.tableau.com/en-us/gallery/whos-talking-popular-films?tab=viz-of-the-day&type=viz-of-the-day).

## Programming language (R with RStudio)

A lot of data analysts work with a programming language called R. Most people who work with R end up also using RStudio, an integrated developer environment (IDE), for their data visualization needs. As with Tableau, you can create dashboard-style data visualizations using RStudio.



Check out their website to learn more about [RStudio](https://rstudio.com/).

You could easily spend days exploring all the resources provided at RStudio.com, but the [RStudio Cheatsheets](https://rstudio.com/resources/cheatsheets/) and the [RStudio Visualize Data Primer](https://rstudio.cloud/learn/primers/3) are great places to start. When you have more time, check out the webinars and videos which offer advice and helpful perspectives for both beginners and advanced users.