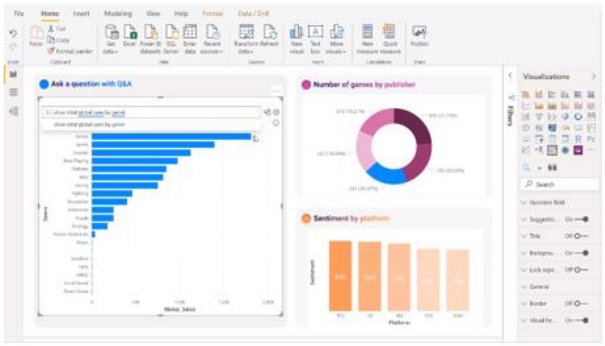
MODULE-2 Data Analyst

Data Analytics (Analysis)

Data analysis is the process of identifying, cleaning, transforming, and modelling data to discover meaningful and useful information. The data is then crafted into a story through reports for analysis to support the critical decision-making process.

As a data analyst, you are on a journey. Think about all the data that is being generated each day and that is available in an organization, from transactional data in a traditional database, telemetry data from services that you use, to signals that you get from different areas like social media.



For example, today's retail businesses collect and store massive amounts of data that track the items you browsed and purchased, the pages you've visited on their site, the aisles you purchase products from, your spending habits, and much more.

With data and information as the most strategic asset of a business, the underlying challenge that organizations have today is understanding and using their data to positively affect change within the business. Businesses continue to struggle to use their data in a meaningful and productive way, which impacts their ability to act.

A retail business should be able to use their vast amounts of data and information **in such a way that impacts the business, including:**

- **Tracking inventory**
- **Identifying purchase habits**
- **Detecting user trends and patterns**
- **Recommending purchases**
- **Determining price optimizations**
- **Identifying and stopping fraud**

Additionally, you might be looking for daily/monthly sale patterns. Common data segments that you might want to examine include day-over-day, week-over-week, and month-over-month so that you can compare how sales have been to where they were in the same week last year, for example.

The key to unlocking this data is being able to tell a story with it. In today's highly competitive and fast-paced business world, crafting reports that tell that story is what helps business leaders take action on the data. Business decision makers depend on an accurate story to drive better business decisions. The faster a business can make precise decisions, the more competitive they will be and the better advantage they will have. Without the story, it is difficult to understand what the data is trying to tell you.

However, having data alone is not enough. You need to be able to act on the data to affect change within the business. That action could involve reallocating resources within the business to accommodate a need, or it could be identifying a failing campaign and knowing when to change course. These situations are where telling a story with your data is important.

The underlying challenge that businesses face today is understanding and using their data in such a way that impacts their business and ultimately their bottom line. You need to be able to look at the data and facilitate trusted business decisions. Then, you need the ability to look at metrics and clearly understand the meaning behind those metrics.

This requirement might seem daunting, but it's a task that you can accomplish. Your first step is to partner with data experts within your organization, such as data engineers and data scientists, to help get the data that you need to tell that story. Ask these experts to participate in that data journey with you.

Your journey of telling a story with data also ties into building that data culture within your organization. While telling the story is important, *where* that story is told is also crucial, ensuring that the story is told to the right people. Also, make sure that people can discover the story, that they know where to find it, and that it is part of the regular interactions.

Data analysis exists to help overcome these challenges and pain points, ultimately assisting businesses in finding insights and uncovering hidden value in troves of data through storytelling. As you read on, you will learn how to use and apply analytical skills to go beyond a single report and help impact and influence your organization by telling stories with data and driving that data culture.

Overview of data analysis

Before data can be used to tell a story, it must be run through a process that makes it usable in the story. Data analysis is the process of identifying, cleaning, transforming, and modelling data to discover meaningful and useful information. The data is then crafted into a story through reports for analysis to support the critical decision-making process.

As the world becomes more data-driven, storytelling through data analysis is becoming a vital component and aspect of large and small businesses. It is the reason that organizations continue to hire data analysts.

Data-driven businesses make decisions based on the story that their data tells, and in today's data-driven world, data is not being used to its full potential, a challenge that most businesses face. Data analysis is, and should be, a critical aspect of all organizations to help determine the impact to their business,

including evaluating customer sentiment, performing market and product research, and identifying trends or other data insights.

While the process of data analysis focuses on the tasks of cleaning, modeling, and visualizing data, the concept of data analysis and its importance to business should not be understated.

To analyse data, core components of analytics are divided into the following categories:

- **Descriptive**
- **Diagnostic**
- **Predictive**
- **Prescriptive**
- **Cognitive**

Descriptive analytics

Descriptive analytics help answer questions about what has happened based on historical data. Descriptive analytics techniques summarize large datasets to describe outcomes to stakeholders.

By developing key performance indicators (KPIs), these strategies can help track the success or failure of key objectives. Metrics such as return on investment (ROI) are used in many industries, and specialized metrics are developed to track performance in specific industries.

An example of descriptive analytics is generating reports to provide a view of an organization's sales and financial data.

Diagnostic analytics

Diagnostic analytics help answer questions about why events happened. Diagnostic analytics techniques supplement basic descriptive analytics, and they use the findings from descriptive analytics to discover the cause of these events. Then, performance indicators are further investigated to discover why these events improved or became worse. Generally, this process occurs in three steps:

1. Identify anomalies in the data. These anomalies might be unexpected changes in a metric or a particular market.
2. Collect data that's related to these anomalies.
3. Use statistical techniques to discover relationships and trends that explain these anomalies.

Predictive analytics

Predictive analytics help answer questions about what will happen in the future. Predictive analytics techniques use historical data to identify trends and determine if they're likely to recur. Predictive analytical tools provide valuable insight into what might happen in the future. Techniques include a variety of statistical and machine learning techniques such as neural networks, decision trees, and regression.

Prescriptive analytics

Prescriptive analytics help answer questions about which actions should be taken to achieve a goal or target. By using insights from prescriptive analytics, organizations can make data-driven decisions. This technique allows businesses to make informed decisions in the face of uncertainty. Prescriptive analytics techniques rely on machine learning as one of the strategies to find patterns in large datasets. By analyzing past decisions and events, organizations can estimate the likelihood of different outcomes.

Cognitive analytics

Cognitive analytics attempt to draw inferences from existing data and patterns, derive conclusions based on existing knowledge bases, and then add these findings back into the knowledge base for future inferences, a self-learning feedback loop. Cognitive analytics help you learn what might happen if circumstances change and determine how you might handle these situations.

Inferences aren't structured queries based on a rules database; rather, they're unstructured hypotheses that are gathered from several sources and expressed with varying degrees of confidence. Effective cognitive analytics depend on machine learning algorithms, and will use several natural language processing concepts to make sense of previously untapped data sources, such as call center conversation logs and product reviews.

Example

By enabling reporting and data visualizations, a retail business uses descriptive analytics to look at patterns of purchases from previous years to determine what products might be popular next year. The company might also look at supporting data to understand why a particular product was popular and if that

trend is continuing, which will help them determine whether to continue stocking that product.

A business might determine that a certain product was popular over a specific timeframe. Then, they can use this analysis to determine whether certain marketing efforts or online social activities contributed to the sales increase.

An underlying facet of data analysis is that a business needs to trust its data. As a practice, the data analysis process will capture data from trusted sources and shape it into something that is consumable, meaningful, and easily understood to help with the decision-making process. Data analysis enables businesses to fully understand their data through data-driven processes and decisions, allowing them to be confident in their decisions.

As the amount of data grows, so does the need for data analysts. A data analyst knows how to organize information and distill it into something relevant and comprehensible. A data analyst knows how to gather the right data and what to do with it, in other words, making sense of the data in your data overload.

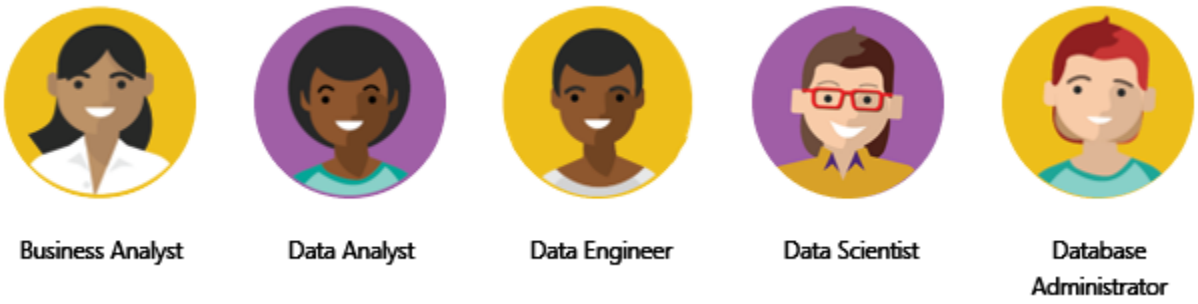
Roles in data

Telling a story with the data is a journey that usually doesn't start with you. The data must come from somewhere. Getting that data into a place that is usable by you takes effort that is likely out of your scope, especially in consideration of the enterprise.

Today's applications and projects can be large and intricate, often involving the use of skills and knowledge from numerous individuals. Each person brings a unique talent and expertise, sharing in the effort of working together and

coordinating tasks and responsibilities to see a project through from concept to production.

In the recent past, roles such as business analysts and business intelligence developers were the standard for data processing and understanding. However, excessive expansion of the size and different types of data has caused these roles to evolve into more specialized sets of skills that modernize and streamline the processes of data engineering and analysis.



The following sections highlight these different roles in data and the specific responsibility in the overall spectrum of data discovery and understanding:

- Business analyst
- Data analyst
- Data engineer
- Data scientist
- Database administrator

Business analyst

While some similarities exist between a data analyst and business analyst, the key differentiator between the two roles is what they do with data. A business analyst is closer to the business and is a specialist in interpreting the data that

comes from the visualization. Often, the roles of data analyst and business analyst could be the responsibility of a single person.

Data analyst

A data analyst enables businesses to maximize the value of their data assets through visualization and reporting tools such as Microsoft Power BI. Data analysts are responsible for **profiling, cleaning, and transforming data**. Their responsibilities also include **designing and building scalable and effective data models,** and enabling and implementing the advanced analytics capabilities into reports for analysis. A data analyst works with the pertinent stakeholders to identify appropriate and necessary data and reporting requirements, and then they are tasked with turning raw data into relevant and meaningful insights.

A data analyst is also responsible for the management of Power BI assets, including reports, dashboards, workspaces, and the underlying datasets that are used in the reports. They are tasked with implementing and configuring proper security procedures, in conjunction with stakeholder requirements, to ensure the safekeeping of all Power BI assets and their data.

Data analysts work with data engineers to determine and locate appropriate data sources that meet stakeholder requirements. Additionally, data analysts work with the data engineer and database administrator to ensure that the analyst has proper access to the needed data sources. The data analyst also works with the data engineer to identify new processes or improve existing processes for collecting data for analysis.

Data engineer

Data engineers' provision and set up data platform technologies that are on-premises and in the cloud. They manage and secure the flow of structured and unstructured data from multiple sources. The data platforms that they use can include relational databases, nonrelational databases, data streams, and file stores. Data engineers also ensure that data services securely and seamlessly integrate across data platforms.

Primary responsibilities of data engineers include the use of on-premises and cloud data services and tools to ingest, egress, and transform data from multiple sources. Data engineers collaborate with business stakeholders to identify and meet data requirements. They design and implement solutions.

While some alignment might exist in the tasks and responsibilities of a data engineer and a database administrator, a data engineer's scope of work goes well beyond looking after a database and the server where it's hosted and likely doesn't include the overall operational data management.

A data engineer adds tremendous value to business intelligence and data science projects. When the data engineer brings data together, often described as data wrangling, projects move faster because data scientists can focus on their own areas of work.

As a data analyst, you would work closely with a data engineer in making sure that you can access the variety of structured and unstructured data sources because they will support you in optimizing data models, which are typically served from a modern data warehouse or data lake.

Both database administrators and business intelligence professionals can transition to a data engineer role; they need to learn the tools and technology that are used to process large amounts of data.

Data scientist

Data scientists perform advanced analytics to extract value from data. Their work can vary from descriptive analytics to predictive analytics. Descriptive analytics evaluate data through a process known as exploratory data analysis (EDA). Predictive analytics are used in machine learning to apply modelling techniques that can detect anomalies or patterns. These analytics are important parts of forecast models.

Descriptive and predictive analytics are only partial aspects of data scientists' work. Some data scientists might work in the realm of deep learning, performing iterative experiments to solve a complex data problem by using customized algorithms.

Anecdotal evidence suggests that most of the work in a data science project is spent on data wrangling and feature engineering. Data scientists can speed up the experimentation process when data engineers use their skills to successfully wrangle data.

On the surface, it might seem that a data scientist and data analyst are far apart in the work that they do, but this conjecture is untrue. A data scientist looks at data to determine the questions that need answers and will often devise a hypothesis or an experiment and then turn to the data analyst to assist with the data visualization and reporting.

Database administrator

A database administrator implements and manages the operational aspects of cloud-native and hybrid data platform solutions that are built on Microsoft Azure data services and Microsoft SQL Server. A database administrator is responsible for the overall availability and consistent performance and optimizations of the database solutions. They work with stakeholders to identify and implement the policies, tools, and processes for data backup and recovery plans.

The role of a database administrator is different from the role of a data engineer. A database administrator monitors and manages the overall health of a database and the hardware that it resides on, whereas a data engineer is involved in the process of data wrangling, in other words, ingesting, transforming, validating, and cleaning data to meet business needs and requirements.

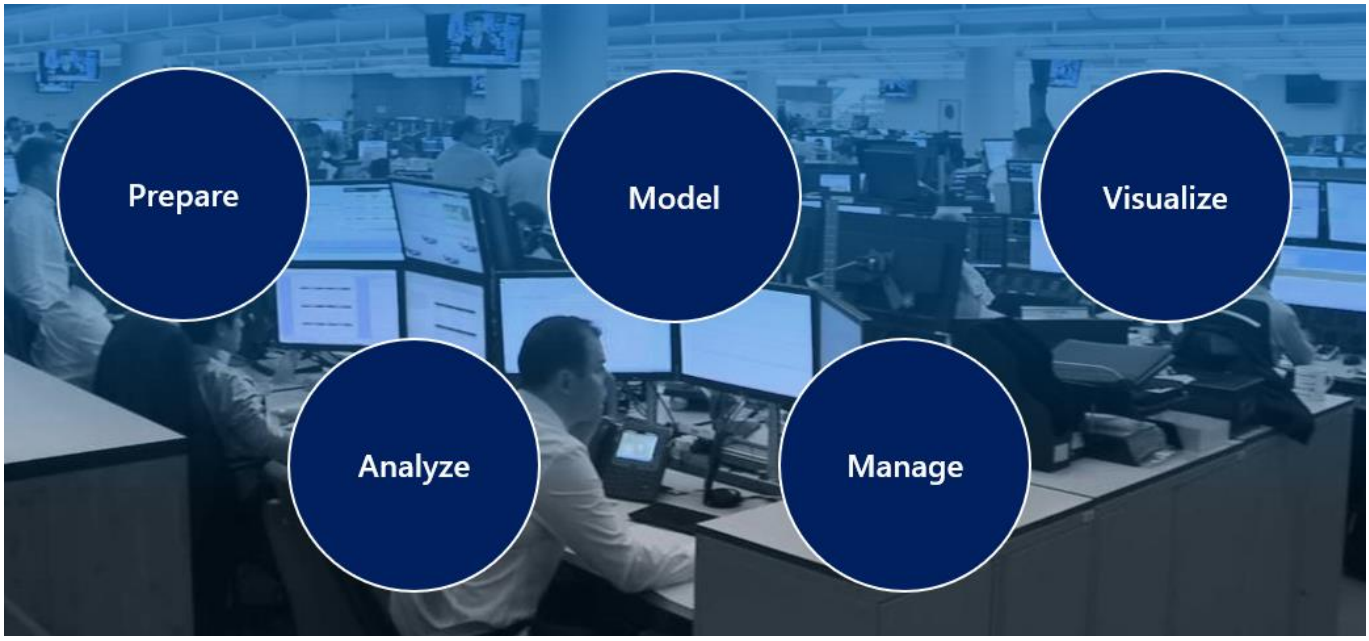
The database administrator is also responsible for managing the overall security of the data, granting and restricting user access and privileges to the data as determined by business needs and requirements.

Tasks of a data analyst

A data analyst is one of several critical roles in an organization, who help uncover and make sense of information to keep the company balanced and operating efficiently. Therefore, it's vital that a data analyst clearly understands their responsibilities and the tasks that are performed on a near-daily basis.

Data analysts are essential in helping organizations gain valuable insights into the expanse of data that they have, and they work closely with others in the organization to help reveal valuable information.

The following figure shows the five key areas that you'll engage in during the data analysis process.



Prepare

As a data analyst, you'll likely divide most of your time between the prepare and model tasks. Deficient or incorrect data can have a major impact that results in invalid reports, a loss of trust, and a negative effect on business decisions, which can lead to loss in revenue, a negative business impact, and more.

Before a report can be created, data must be prepared. Data preparation is the process of profiling, cleaning, and transforming your data to get it ready to model and visualize.

Data preparation is the process of taking raw data and turning it into information that is trusted and understandable. It involves, among other things, ensuring the integrity of the data, correcting wrong or inaccurate data, identifying missing data, converting data from one structure to another or from one type to another, or even a task as simple as making data more readable.

Data preparation also involves understanding *how* you're going to get and connect to the data and the performance implications of the decisions. When connecting to data, you need to make decisions to ensure that models and reports meet, and perform to, acknowledged requirements and expectations.

Privacy and security assurances are also important. These assurances can include anonymizing data to avoid oversharing or preventing people from seeing personally identifiable information when it isn't needed. Alternatively, helping to ensure privacy and security can involve removing that data completely if it doesn't fit in with the story that you're trying to shape.

Data preparation can often be a lengthy process. Data analysts follow a series of steps and methods to prepare data for placement into a proper context and state that eliminate poor data quality and allow it to be turned into valuable insights.

Model

When the data is in a proper state, it's ready to be modelled. **Data modelling is the process of determining how your tables are related to each other**. This process is done by defining and creating relationships between the tables. From that point, you can enhance the model by defining metrics and adding custom calculations to enrich your data.

Creating an effective and proper data model is a critical step in helping organizations understand and gain valuable insights into the data. An effective data model makes reports more accurate, allows the data to be explored faster and more efficient, decreases time for the report writing process, and simplifies future report maintenance.

The model is another critical component that has a direct effect on the performance of your report and overall data analysis. A poorly designed model can have a drastically negative impact on the general accuracy and performance of your report. Conversely, a well-designed model with well-prepared data will ensure a properly efficient and trusted report. This notion is more prevalent when you are working with data at scale.

From a Power BI perspective, if your report is performing slowly, or your refreshes are taking a long time, you will likely need to revisit the data preparation and modelling tasks to optimize your report.

The process of preparing data and modelling data is an iterative process. Data preparation is the first task in data analysis. Understanding and preparing your data before you model it will make the modelling step much easier.

Visualize

The visualization task is where you get to bring your data to life. The ultimate goal of the visualize task is to solve business problems. A well-designed report should tell a compelling story about that data, which will enable business decision makers to quickly gain needed insights. By using appropriate visualizations and interactions, you can provide an effective report that guides

the reader through the content quickly and efficiently, therefore allowing the reader to follow a narrative into the data.

The reports that are created during the visualization task help businesses and decision makers understand what that data means so that accurate and vital decisions can be made. Reports drive the overall actions, decisions, and behaviors of an organization that is trusting and relying on the information that is discovered in the data.

The business might communicate that they need all data points on a given report to help them make decisions. As a data analyst, you should take the time to fully understand the problem that the business is trying to solve. Determine whether all their data points are necessary because too much data can make detecting key points difficult. Having a small and concise data story can help find insights quickly.

With the built-in AI capabilities in Power BI, data analysts can build powerful reports, without writing any code, that enable users to get insights and answers and find actionable objectives. The AI capabilities in Power BI, such as the built-in AI visuals, enable the discovering of data by asking questions, using the Quick Insights feature, or creating machine learning models directly within Power BI.

An important aspect of visualizing data is designing and creating reports for accessibility. As you build reports, it is important to think about people who will be accessing and reading the reports. Reports should be designed with accessibility in mind from the outset so that no special modifications are needed in the future.

Many components of your report will help with storytelling. From a color scheme that is complementary and accessible, to fonts and sizing, to picking the right visuals for what is being displayed, they all come together to tell that story.

Analyze

The analyze task is the important step of understanding and interpreting the information that is displayed on the report. In your role as a data analyst, you should understand the analytical capabilities of Power BI and use those capabilities to find insights, identify patterns and trends, predict outcomes, and then communicate those insights in a way that everyone can understand.

Advanced analytics enables businesses and organizations to ultimately drive better decisions throughout the business and create actionable insights and meaningful results. With advanced analytics, organizations can drill into the data to predict future patterns and trends, identify activities and behaviors, and enable businesses to ask the appropriate questions about their data.

Previously, analyzing data was a difficult and intricate process that was typically performed by data engineers or data scientists. Today, Power BI makes data analysis accessible, which simplifies the data analysis process. Users can quickly gain insights into their data by using visuals and metrics directly from their desktop and then publish those insights to dashboards so that others can find needed information.

This feature is another area where AI integrations within Power BI can take your analysis to the next level. Integrations with Azure machine learning, cognitive services, and built-in AI visuals will help to enrich your data and analysis.

Manage

Power BI consists of many components, including reports, dashboards, workspaces, datasets, and more. As a data analyst, you are responsible for the management of these Power BI assets, overseeing the sharing and distribution of items, such as reports and dashboards, and ensuring the security of Power BI assets.

Apps can be a valuable distribution method for your content and allow easier management for large audiences. This feature also allows you to have custom navigation experiences and link to other assets within your organization to complement your reports.

The management of your content helps to foster collaboration between teams and individuals. Sharing and discovery of your content is important for the right people to get the answers that they need. It is also important to help ensure that items are secure. You want to make sure that the right people have access and that you are not leaking data past the correct stakeholders.

Proper management can also help reduce data silos within your organization. Data duplication can make managing and introducing data latency difficult when resources are overused. Power BI helps reduce data silos with the use of shared datasets, and it allows you to reuse data that you have prepared and modeled. For key business data, endorsing a dataset as certified can help to ensure trust in that data.

The management of Power BI assets helps reduce the duplication of efforts and helps ensure security of the data.

Get data from files

Organizations often export and store data in files. One possible file format is a flat file. A flat file is a type of file that has only one data table and every row of data is in the same structure. The file does not contain hierarchies. Likely, you're familiar with the most common types of flat files, which are comma-separated values (.csv) files, delimited text (.txt) files, and fixed width files. Another type of file would be the output files from different applications, like Microsoft Excel workbooks (.xlsx).

Power BI Desktop allows you to get data from many types of files. You can find a list of the available options when you use the **Get data** feature in Power BI Desktop. The following sections explain how you can import data from an Excel file that is stored on a local computer.

Scenario

The Human Resources (HR) team at Tailwind Traders has prepared a flat file that contains some of your organization's employee data, such as employee name, hire date, position, and manager. They've requested that you build Power BI reports by using this data, and data that is located in several other data sources.

Flat file location

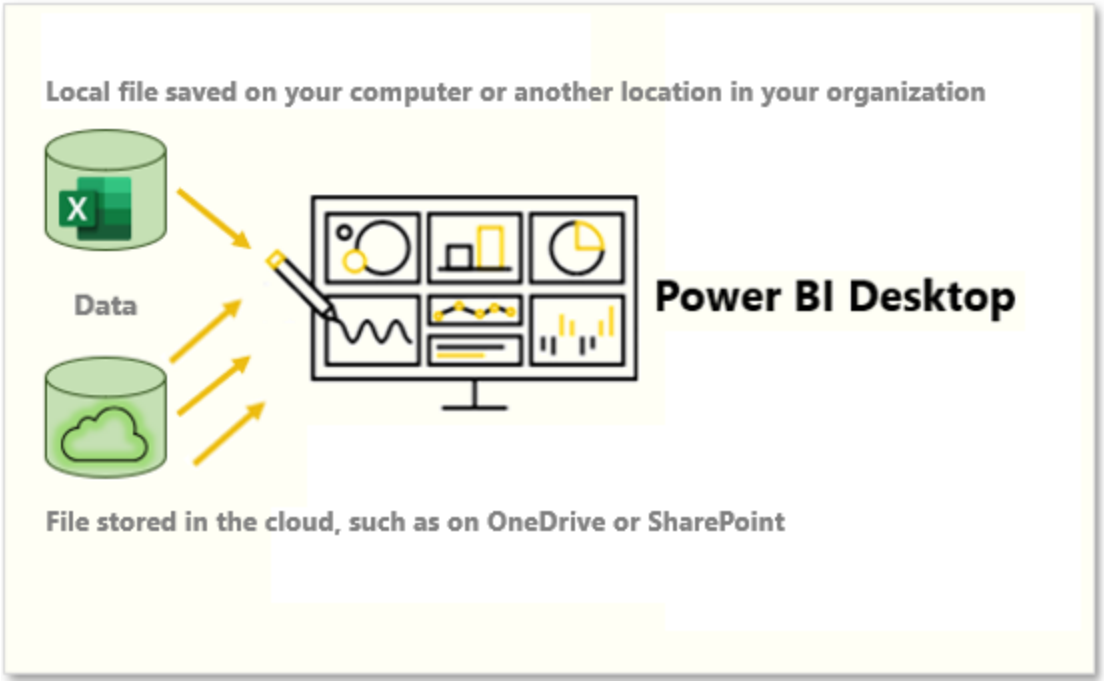
The first step is to determine which file location you want to use to export and store your data.

Your Excel files might exist in one of the following locations:

- **Local** - You can import data from a local file into Power BI. The file isn't moved into Power BI, and a link doesn't remain to it. Instead, a new dataset is created in Power BI, and data from the Excel file is loaded into it. Accordingly, changes to the original Excel file are not reflected in your Power BI dataset. You can use local data import for data that doesn't change.
- **OneDrive for Business** - You can pull data from OneDrive for Business into Power BI. This method is effective in keeping an Excel file and your dataset, reports, and dashboards in Power BI synchronized. Power BI connects regularly to your file on OneDrive. If any changes are found, your dataset, reports, and dashboards are automatically updated in Power BI.
- **OneDrive - Personal** - You can use data from files on a personal OneDrive account, and get many of the same benefits that you would with OneDrive for Business. However, you'll need to sign in with your

personal OneDrive account, and select the **Keep me signed in** option. Check with your system administrator to determine whether this type of connection is allowed in your organization.

- **SharePoint - Team Sites** - Saving your Power BI Desktop files to SharePoint Team Sites is similar to saving to OneDrive for Business. The main difference is how you connect to the file from Power BI. You can specify a URL or connect to the root folder.



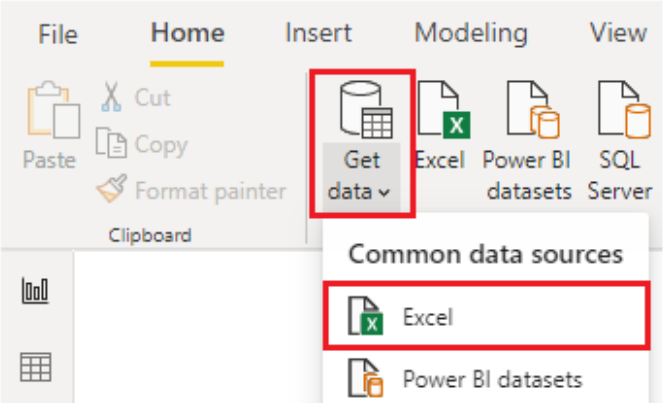
Using a cloud option such as OneDrive or SharePoint Team Sites is the most effective way to keep your file and your dataset, reports, and dashboards in Power BI in-sync. However, if your data does not change regularly, saving files on a local computer is a suitable option.

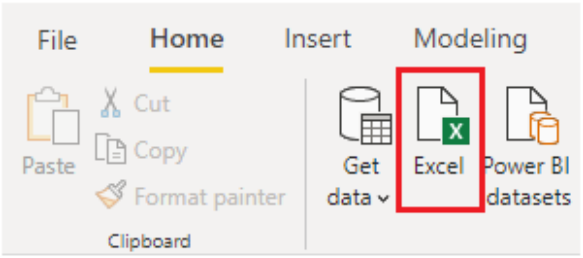
Connect to data in a file

In Power BI, on the **Home** tab, select **Get data**. In the list that displays, select the option that you require, such as **Text/CSV** or **XML**. For this example, you will select **Excel**.

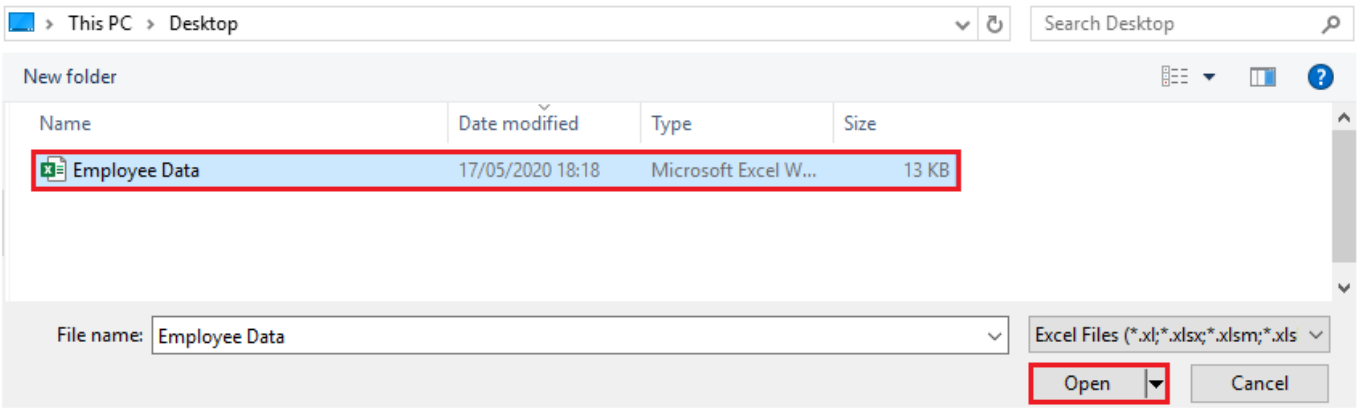
Tip

The **Home** tab contains quick access data source options, such as **Excel**, next to the **Get data** button.





Depending on your selection, you need to find and open your data source. You might be prompted to sign into a service, such as OneDrive, to authenticate your request. In this example, you will open the **Employee Data** Excel workbook that is stored on the Desktop (Remember, no files are provided for practice, these are hypothetical steps).

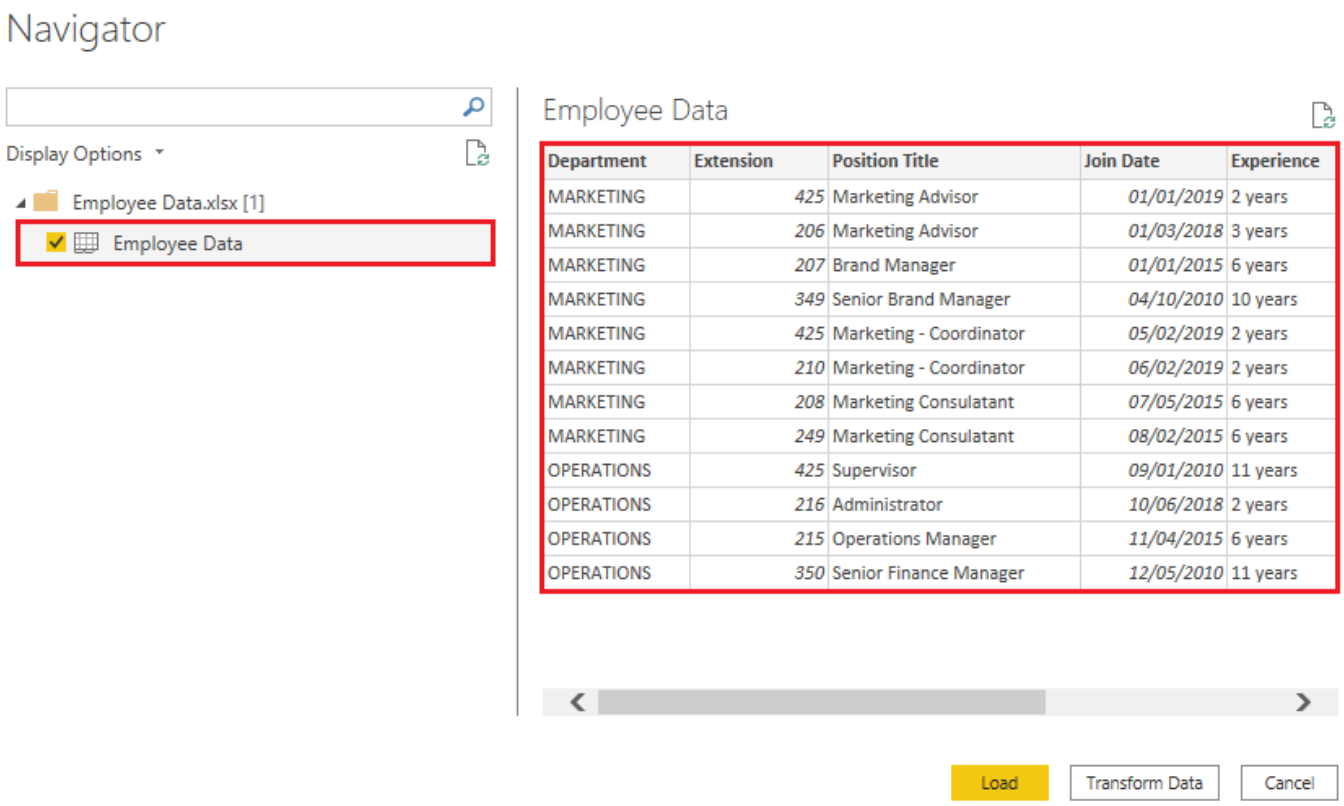


Select the file data to import

After the file has connected to Power BI Desktop, the **Navigator** window opens. This window shows you the data that is available in your data source (the Excel

file in this example). You can select a table or entity to preview its contents, to ensure that the correct data is loaded into the Power BI model.

Select the check box(es) of the table(s) that you want to bring in to Power BI. This selection activates the **Load** and **Transform Data** buttons as shown in the following image.



Department	Extension	Position Title	Join Date	Experience
MARKETING	425	Marketing Advisor	01/01/2019	2 years
MARKETING	206	Marketing Advisor	01/03/2018	3 years
MARKETING	207	Brand Manager	01/01/2015	6 years
MARKETING	349	Senior Brand Manager	04/10/2010	10 years
MARKETING	425	Marketing - Coordinator	05/02/2019	2 years
MARKETING	210	Marketing - Coordinator	06/02/2019	2 years
MARKETING	208	Marketing Consulatant	07/05/2015	6 years
MARKETING	249	Marketing Consulatant	08/02/2015	6 years
OPERATIONS	425	Supervisor	09/01/2010	11 years
OPERATIONS	216	Administrator	10/06/2018	2 years
OPERATIONS	215	Operations Manager	11/04/2015	6 years
OPERATIONS	350	Senior Finance Manager	12/05/2010	11 years

You now have the option to select the **Load** button to automatically load your data into the Power BI model or select the **Transform Data** button to launch the Power Query Editor, where you can review and clean your data before loading it into the Power BI model.

We often recommend that you transform data, but that process will be discussed later in this module. For this example, you can select **Load**.

Change the source file

You might have to change the location of a source file for a data source during development, or if a file storage location changes. To keep your reports up to date, you'll need to update your file connection paths in Power BI.

Power Query provides a number of ways for you to accomplish this task, so that you can make this type of change when needed.

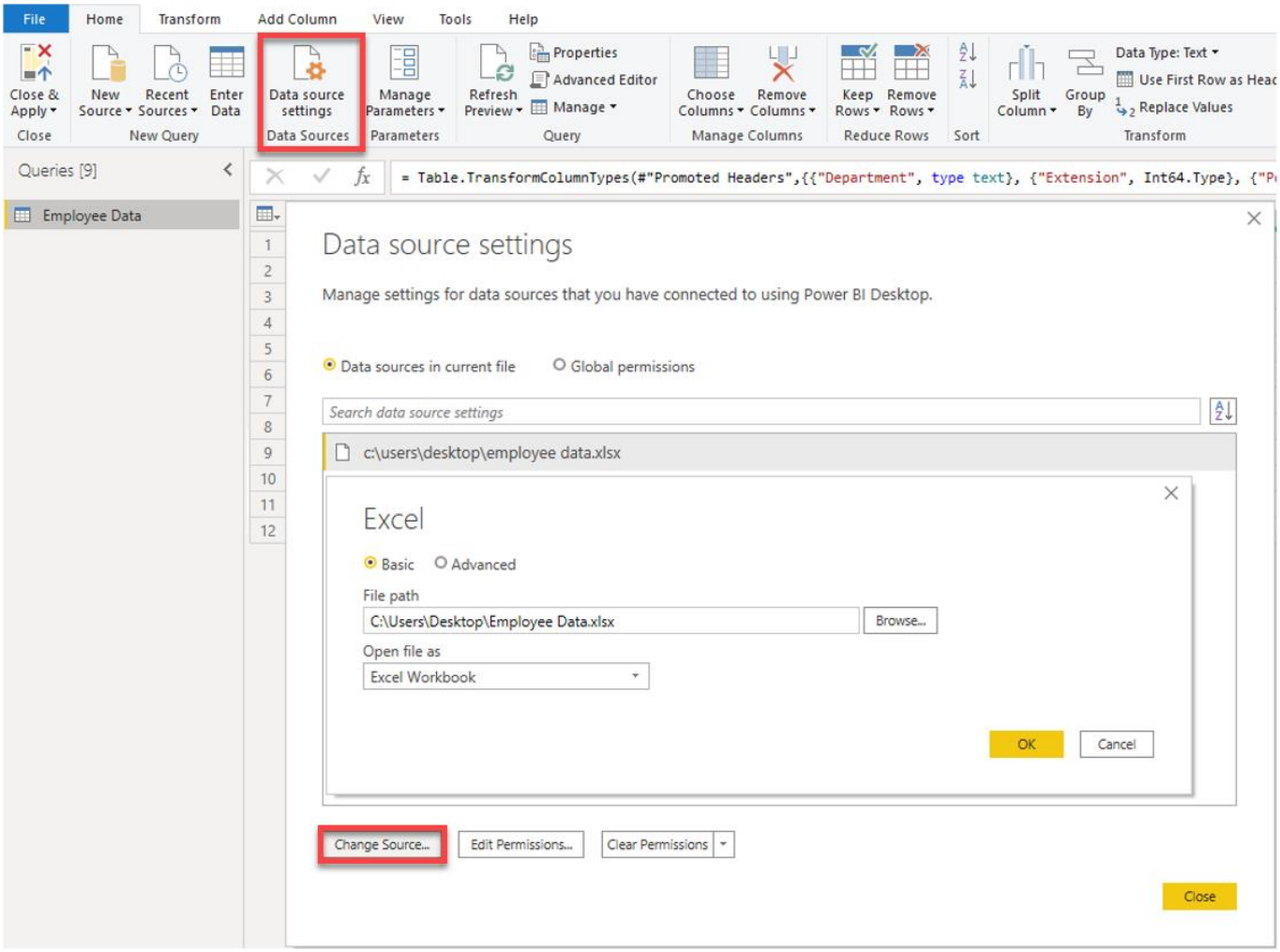
1. Data source settings
2. Query settings

3. Advanced Editor

Warning

If you are changing a file path, make sure that you reconnect to the same file with the same file structure. Any structural changes to a file, such as deleting or renaming columns in the source file, will break the reporting model.

For example, try changing the data source file path in the data source settings. Select **Data source settings** in Power Query. In the **Data source settings** window, select your file and then select **Change Source**. Update the **File path** or use the **Browse** option to locate your file, select **OK**, and then select **Close**.



Get data from relational data sources

Completed100 XP

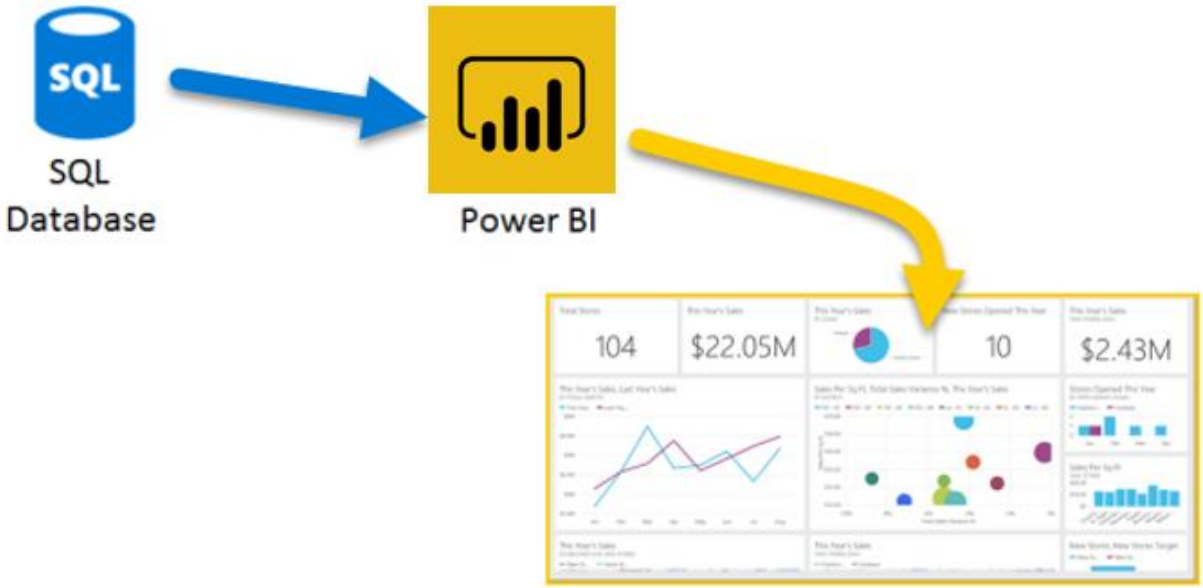
- 6 minutes

If your organization uses a relational database to record its sales transactions, you can use Power BI Desktop to establish a connection to your organization's relational database, rather than getting data from individual flat files.

Connecting Power BI to your database will help you to monitor the progress of your business and identify trends, so you can forecast sales figures, plan budgets and set performance indicators and targets. Power BI Desktop can connect to many relational databases that are either in the cloud or on-premises.

Scenario

The Sales team at Tailwind Traders have requested that you connect to the organization's on-premises SQL Server database and get the sales data into Power BI Desktop so you can build sales reports.

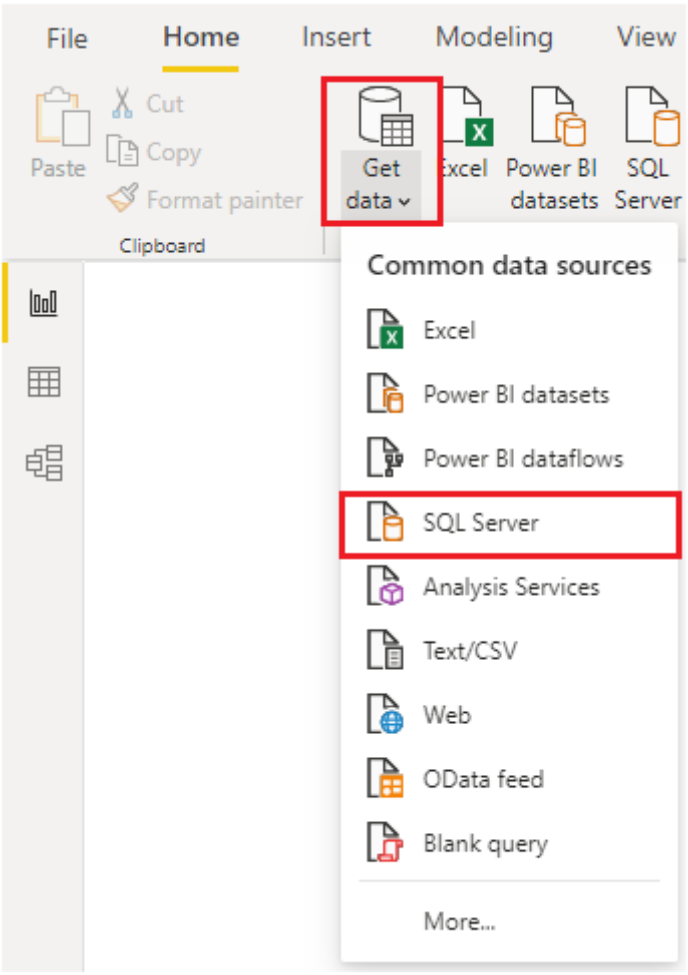


Connect to data in a relational database

You can use the **Get data** feature in Power BI Desktop and select the applicable option for your relational database. For this example, you would select the **SQL Server** option, as shown in the following screenshot.

Tip

Next to the **Get Data** button are quick access data source options, such as **SQL Server.**



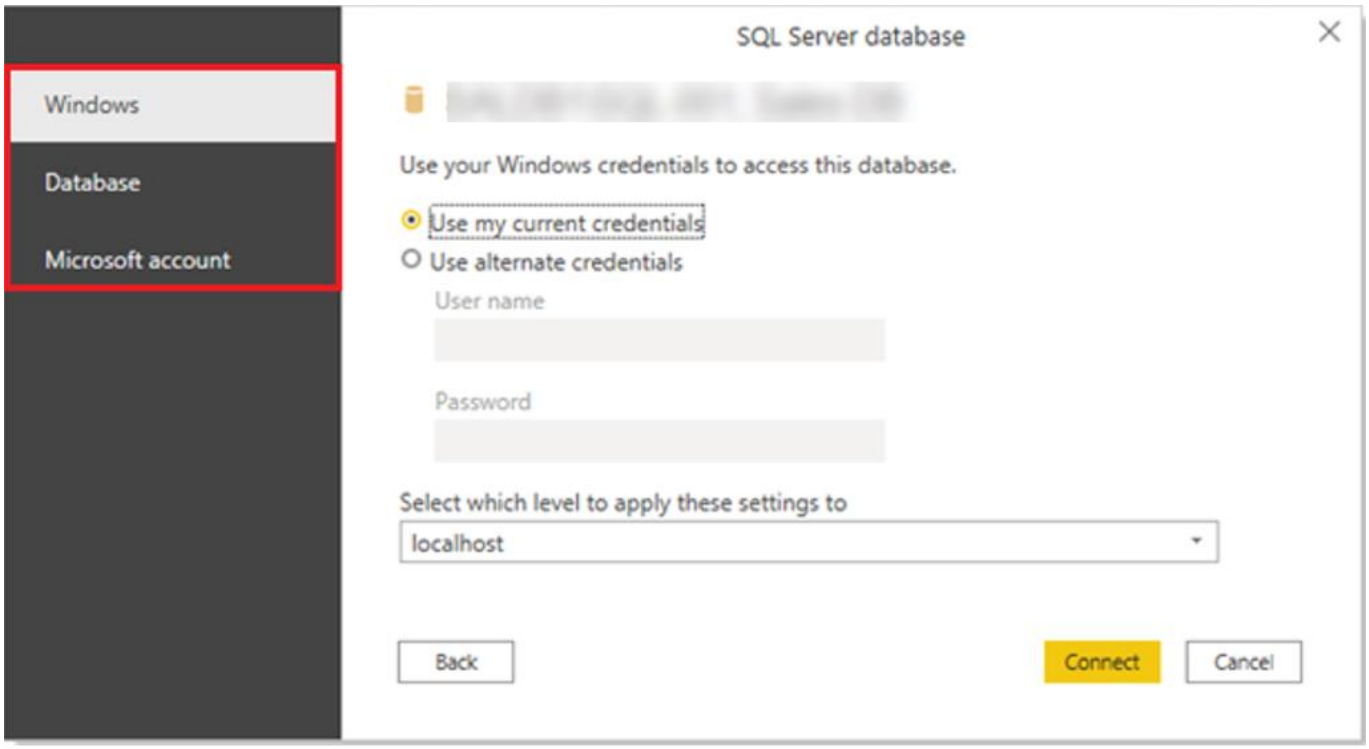
Your next step is to enter your database server name and a database name in the **SQL Server database** window. The two options in data connectivity mode are: **Import** (selected by default, recommended) and **DirectQuery**. Mostly, you select **Import.** Other advanced options are also available in the **SQL Server database** window, but you can ignore them for now.

After you have added your server and database names, you will be prompted to sign in with a username and password. You will have three sign-in options:

- **Windows** - Use your Windows account (Azure Active Directory credentials).

- **Database** - Use your database credentials. For instance, SQL Server has its own sign-in and authentication system that is sometimes used. If the database administrator gave you a unique sign-in to the database, you might need to enter those credentials on the **Database** tab.
- **Microsoft account** - Use your Microsoft account credentials. This option is often used for Azure services.

Select a sign-in option, enter your username and password, and then select **Connect**.

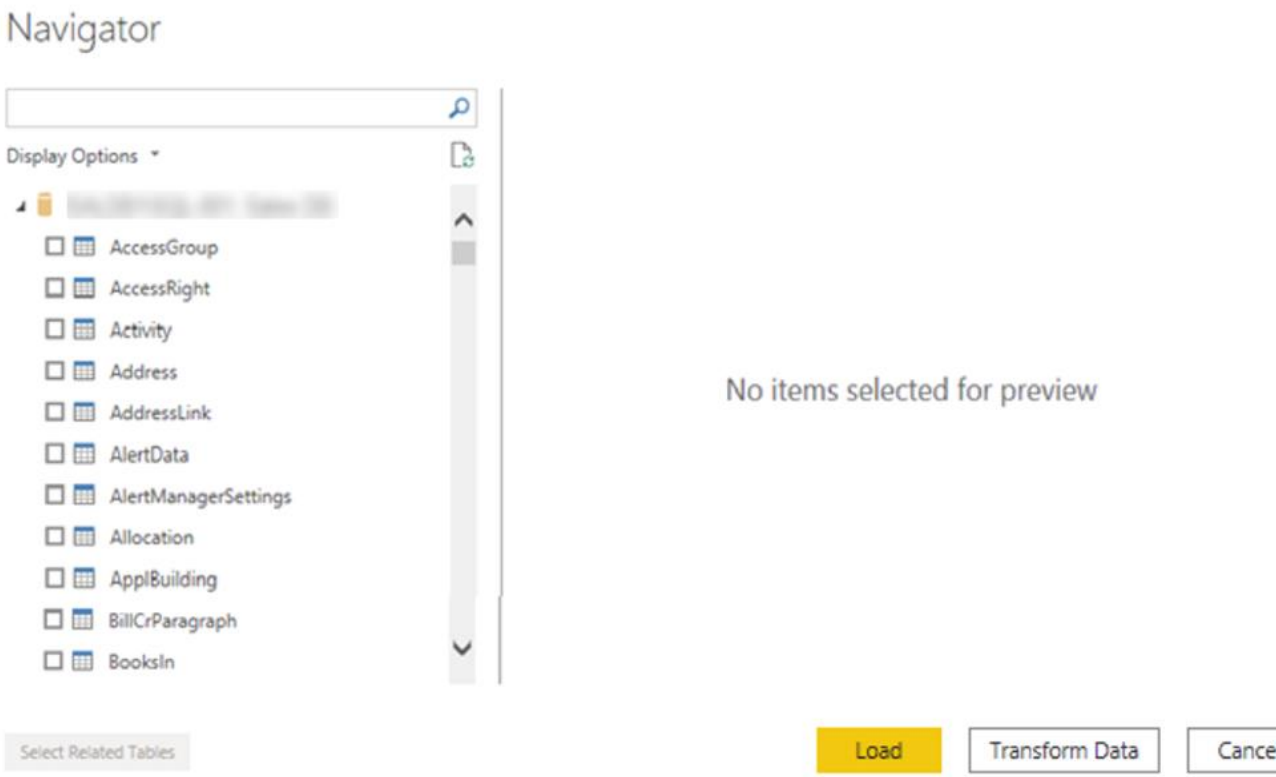


Select data to import

After the database has been connected to Power BI Desktop, the **Navigator** window displays the data that is available in your data source (the SQL database in this example). You can select a table or entity to preview its contents and make sure that the correct data will be loaded into the Power BI model.

Select the check box(es) of the table(s) that you want to bring in to Power BI Desktop, and then select either the **Load** or **Transform Data** option.

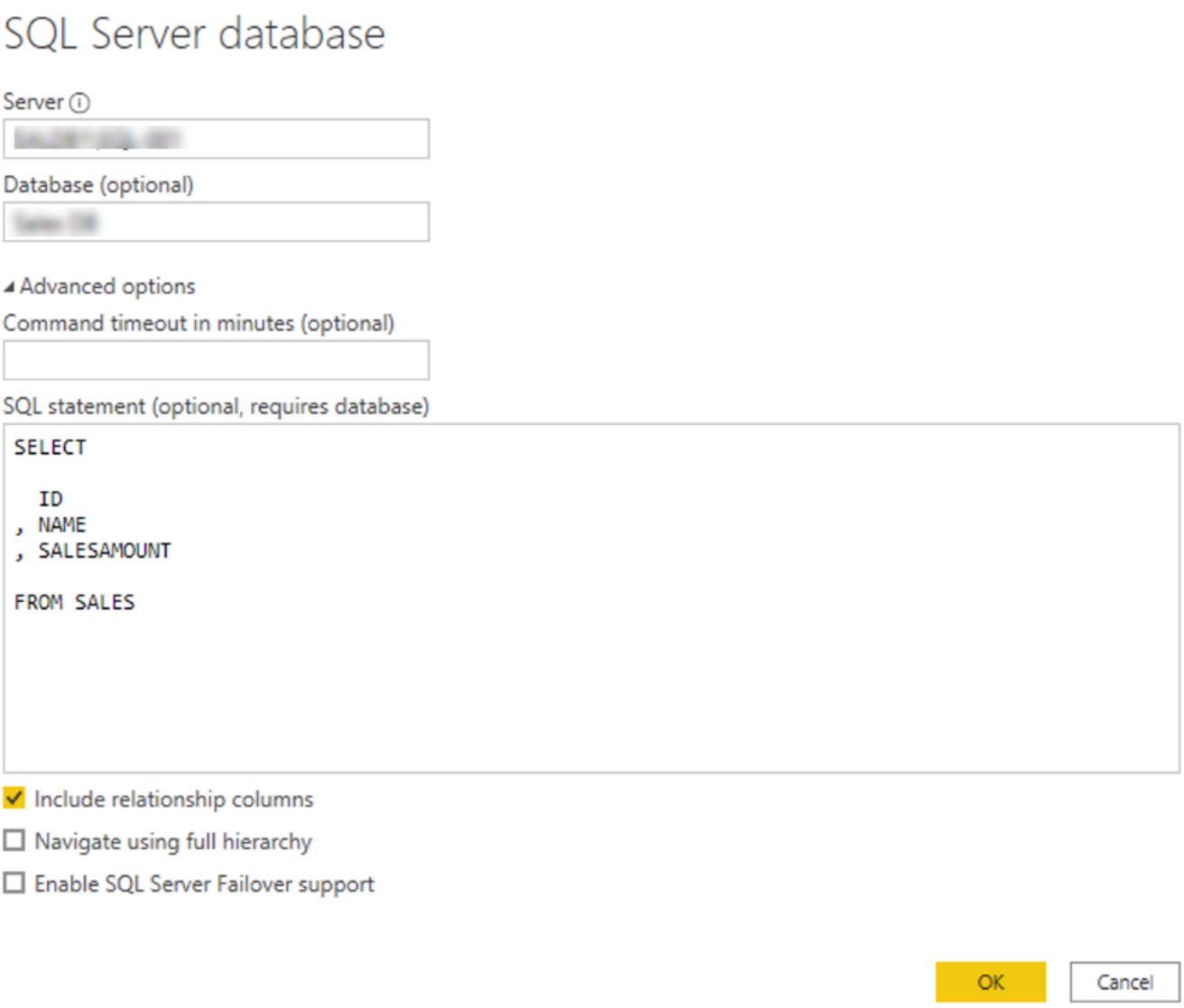
- **Load** - Automatically load your data into a Power BI model in its current state.
- **Transform Data** - Open your data in Microsoft Power Query, where you can perform actions such as deleting unnecessary rows or columns, grouping your data, removing errors, and many other data quality tasks.



Import data by writing an SQL query

Another way you can import data is to write an SQL query to specify only the tables and columns that you need.

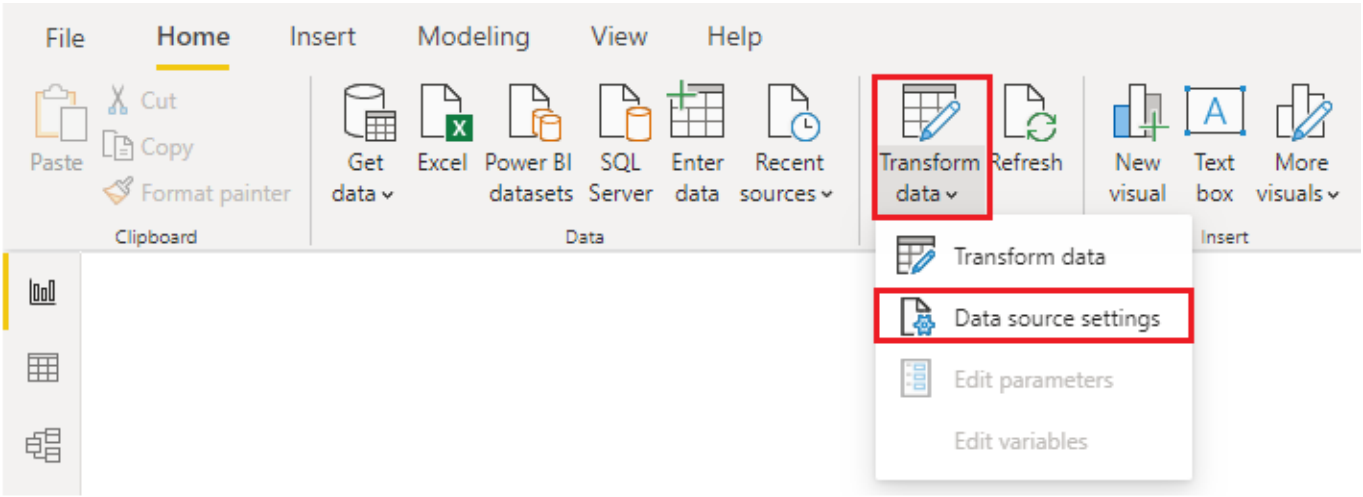
To write your SQL query, on the **SQL Server database** window, enter your server and database names, and then select the arrow next to **Advanced options** to expand this section and view your options. In the **SQL statement** box, write your query statement, and then select **OK**. In this example, you will use the **Select** SQL statement to load the ID, NAME and SALESAMOUNT columns **from** the SALES table.



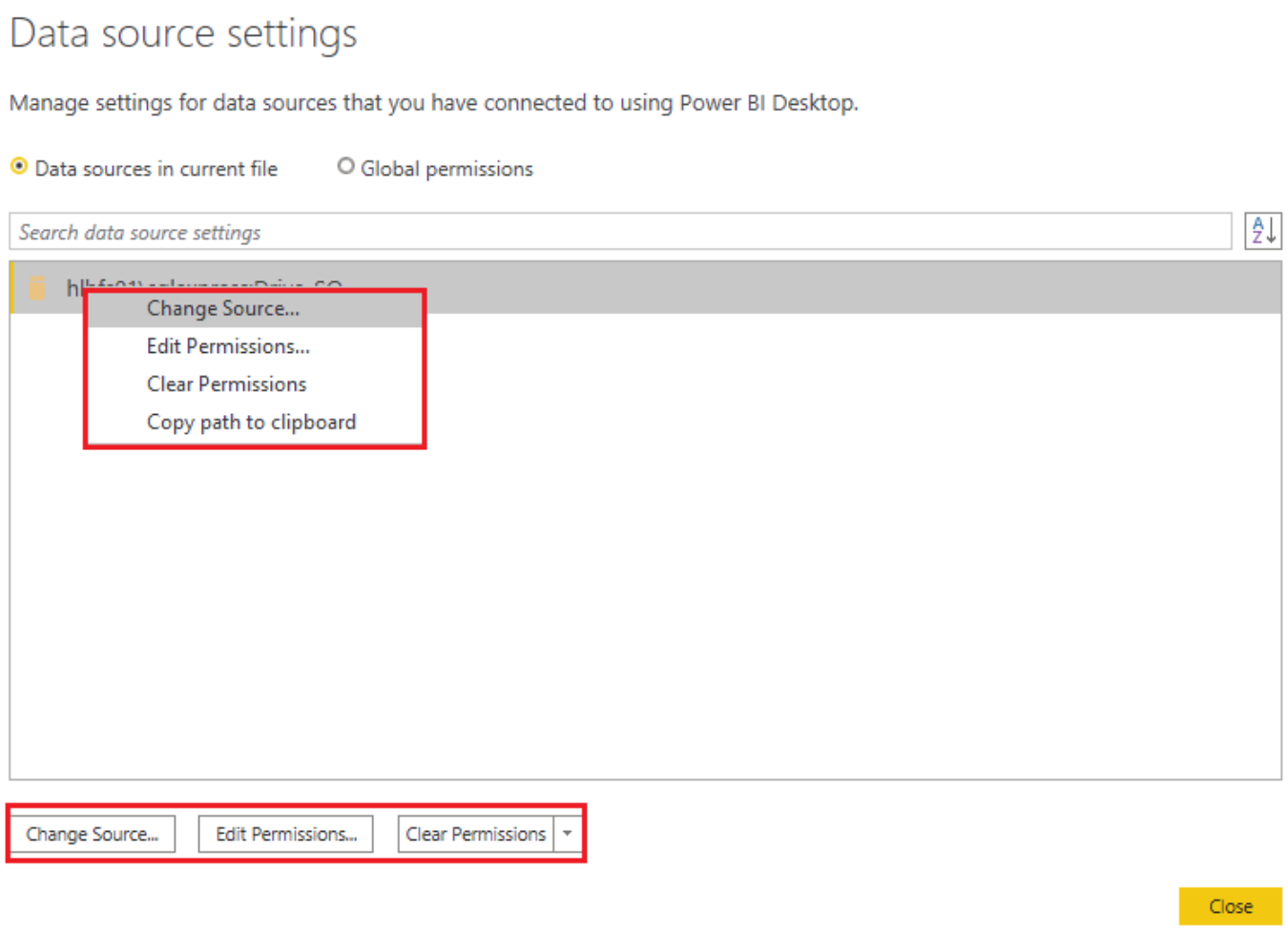
Change data source settings

After you create a data source connection and load data into Power BI Desktop, you can return and change your connection settings at any time. This action is often required due to a security policy within the organization, for example, when the password needs to be updated every 90 days. You can change the data source, edit permissions or clear permissions.

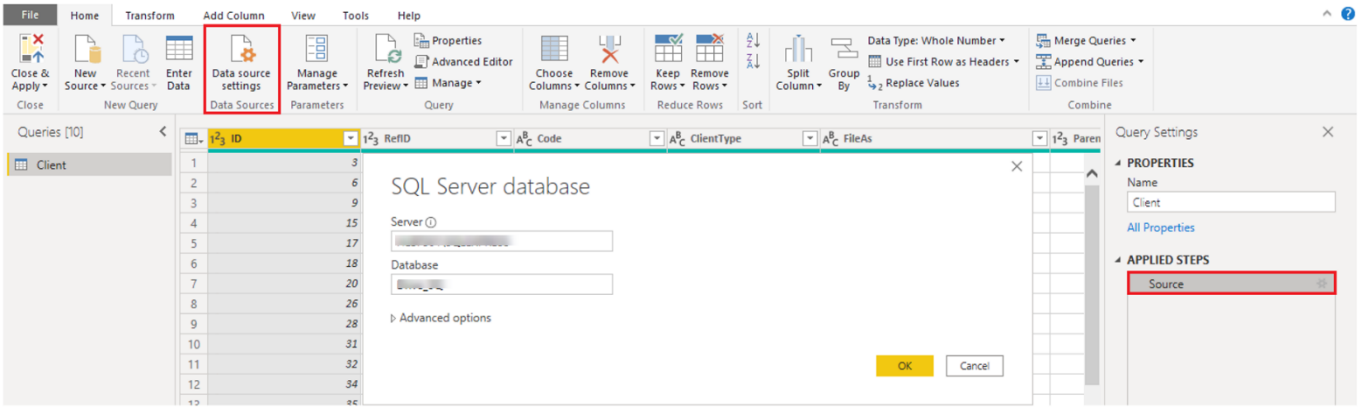
On the **Home** tab, select **Transform data,** and then select the **Data source settings** option.



From the list of data sources that displays, select the data source that you want to update. Then, you can right-click that data source to view the available update options or you can use the update option buttons on the lower left of the window. Select the update option that you need, change the settings as required, and then apply your changes.



You can also change your data source settings from within Power Query. Select the table, and then select the **Data source settings** option on the **Home** ribbon. Alternatively, you can go to the **Query Settings** panel on the right side of the screen and select the settings icon next to Source (or double Select Source). In the window that displays, update the server and database details, and then select **OK**.



After you have made the changes, select **Close and Apply** to apply those changes to your data source settings.

Write an SQL statement

As previously mentioned, you can import data into your Power BI model by using an SQL query. SQL stands for Structured Query Language and is a standardized programming language that is used to manage relational databases and perform various data management operations.

Consider the scenario where your database has a large table that is comprised of sales data over several years. Sales data from 2009 is not relevant to the report that you are creating. This situation is where SQL is beneficial because it allows you to load only the required set of data by specifying exact columns and rows in your SQL statement and then importing them into your data model. You can also join different tables, run specific calculations, create logical statements, and filter data in your SQL query.

The following example shows a simple query where the ID, NAME and SALESAMOUNT are selected from the SALES table.

The SQL query starts with a **Select** statement, which allows you to choose the specific fields that you want to pull from your database. In this example, you want to load the ID, NAME, and SALESAMOUNT columns.

SQLCopy

```
SELECT
ID
, NAME
, SALESAMOUNT
FROM
```

FROM specifies the name of the table that you want to pull the data from. In this case, it's the SALES table. The following example is the full SQL query:

SQLCopy

```
SELECT
ID
, NAME
, SALESAMOUNT
FROM
SALES
```

When using an SQL query to import data, try to avoid using the wildcard character (*) in your query. If you use the wildcard character (*) in your SELECT statement, you import all columns that you don't need from the specified table.

The following example shows the query using the wildcard character.

SQLCopy

```
SELECT *
FROM
SALES
```

The wildcard character (*) will import all columns within the **Sales** table. This method is not recommended because it will lead to redundant data in your data model, which will cause performance issues and require additional steps to normalize your data for reporting.

All queries should also have a **WHERE** clause. This clause will filter the rows to pick only filtered records that you want. In this example, if you want to get recent sales data after Jan 1, 2020, add a **WHERE** clause. The evolved query would look like the following example.

SQLCopy

```
SELECT
ID
, NAME
, SALESAMOUNT
FROM
SALES
WHERE
OrderDate >= ‘1/1/2020’
```

It is a best practice to avoid doing this directly in Power BI. Instead, consider writing a query like this in a view. A view is an object in a relational database, similar to a table. Views have rows and columns, and can contain almost every operator in the SQL language. If Power BI uses a view, when it retrieves data, it participates in query folding, a feature of Power Query. Query folding will be explained later, but in short, Power Query will optimize data retrieval according to how the data is being used later.

Model data in power BI

Introduction

Creating a great data model is one of the most important tasks that a data analyst can perform in Microsoft Power BI. By doing this job well, you help make it easier for people to understand your data, which will make building valuable Power BI reports easier for them and for you.

The pages in this module are instructional only, no data files are provided. You will have a chance to work with real data in the labs.

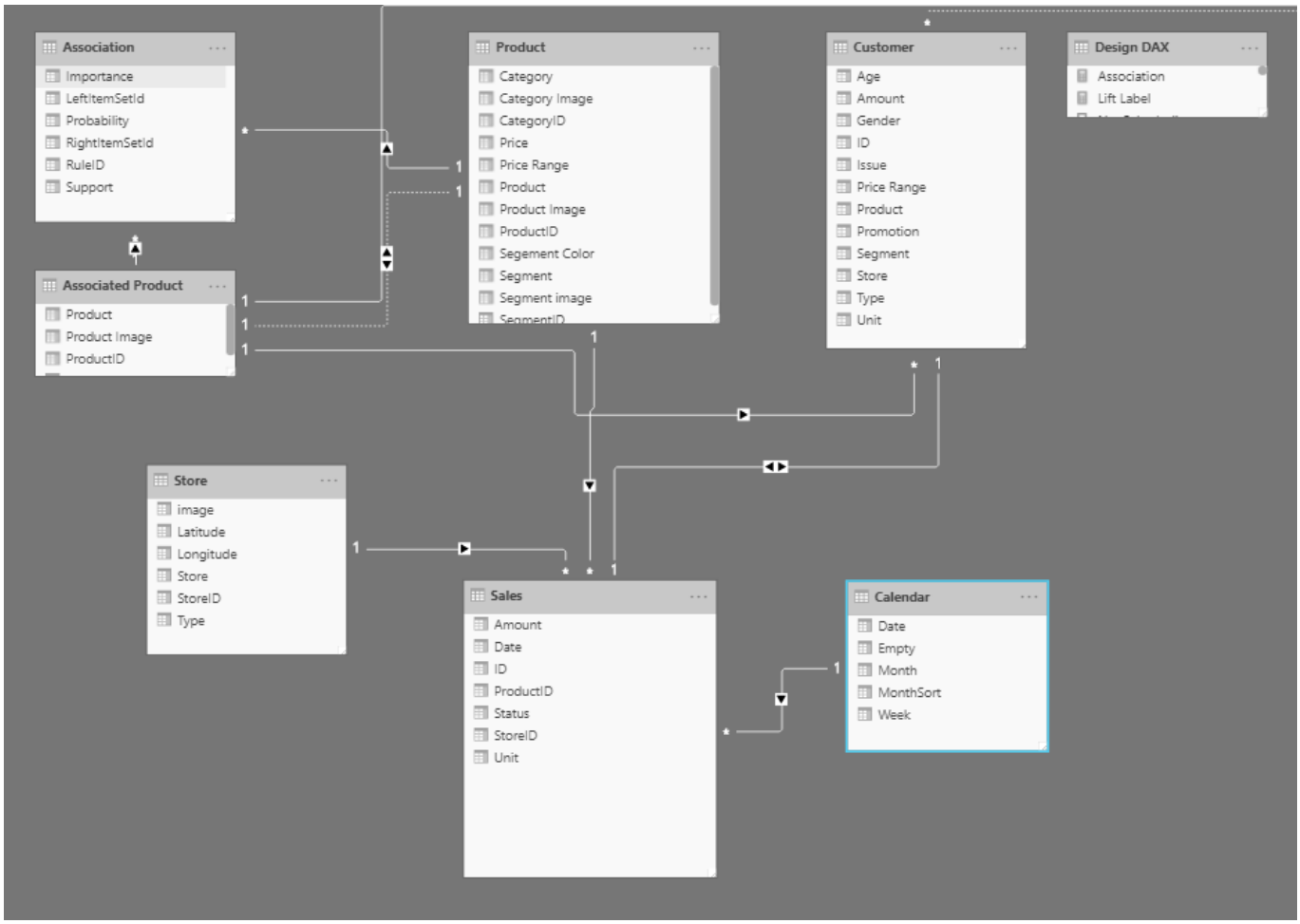
A good data model offers the following benefits:

- **Data exploration is faster.**
- **Aggregations are simpler to build.**
- **Reports are more accurate.**
- **Writing reports takes less time.**
- **Reports are easier to maintain in the future.**

Providing set rules for what makes a good data model is difficult because all data is different, and the usage of that data varies. Generally, a smaller data model is better because it will perform faster and will be simpler to use. However, defining what a smaller data model entails is equally as problematic because it's a heuristic and subjective concept.

Typically, a smaller data model is comprised of fewer tables and fewer columns in each table that the user can see. If you import all necessary tables from a sales database, but the total table count is 30 tables, the user will not find that intuitive. Collapsing those tables into five tables will make the data model more intuitive to the user, whereas if the user opens a table and finds 100 columns, they might find it overwhelming. Removing unneeded columns to provide a more manageable number will increase the likelihood that the user will read all column names. To summarize, you should aim for simplicity when designing your data models.

The following image is an example data model. The boxes contain tables of data, where each line item within the box is a column. The lines that connect the boxes represent relationships between the tables. These relationships can be complex, even in such a simplistic model. The data model can become easily disorganized, and the total table count in the model can gradually increase. Keeping your data model simple, comprehensive, and accurate requires constant effort.



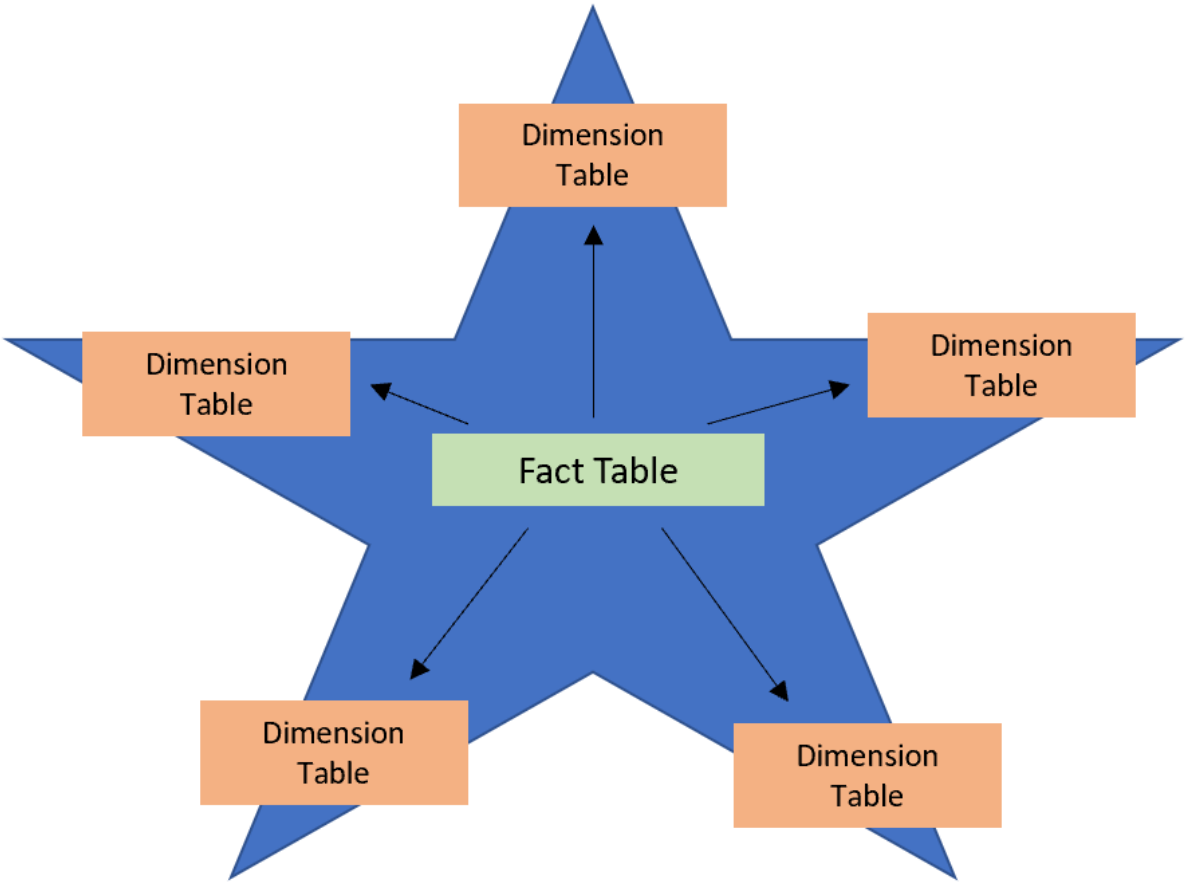
Relationships are defined between tables through primary and foreign keys. Primary keys are column(s) that identify each unique, non-null data row. For instance, if you have a Customers table, you could have an index that identifies each unique customer. The first row will have an ID of 1, the second row an ID of 2, and so on. Each row is assigned a unique value, which can be referred to by this simple value: the primary key. This process becomes important when you are referencing rows in a different table, which is what foreign keys do. Relationships between tables are formed when you have primary and foreign keys in common between different tables.

Power BI allows relationships to be built from tables with different data sources, a powerful function that enables you to pull one table from Microsoft Excel and another from a relational database. You would then create the relationship between those two tables and treat them as a unified dataset.

Now that you have learned about the relationships that make up the data schema, you'll be able to explore a specific type of schema design, the star schema, which is optimized for high performance and usability.

Star schemas

You can design a star schema to simplify your data. It's not the only way to simplify your data, but it is a popular method; therefore, every Power BI data analyst should understand it. In a star schema, each table within your dataset is defined as a dimension or a fact table, as shown in the following visual.



Fact tables contain observational or event data values: **sales orders, product counts, prices, transactional dates and times, and quantities.** Fact tables can contain several **repeated values**. For example, **one product can appear multiple times in multiple rows, for different customers on different dates**. These values can be aggregated to create visuals. For instance, a visual of the total sales orders is an aggregation of all sales orders in the fact table. With fact tables, it is common to see columns that are filled with numbers and dates. The numbers can be units of measurement, such as sale amount, or they can be keys, such as a customer ID. The dates represent time that is being recorded, like order date or shipped date.

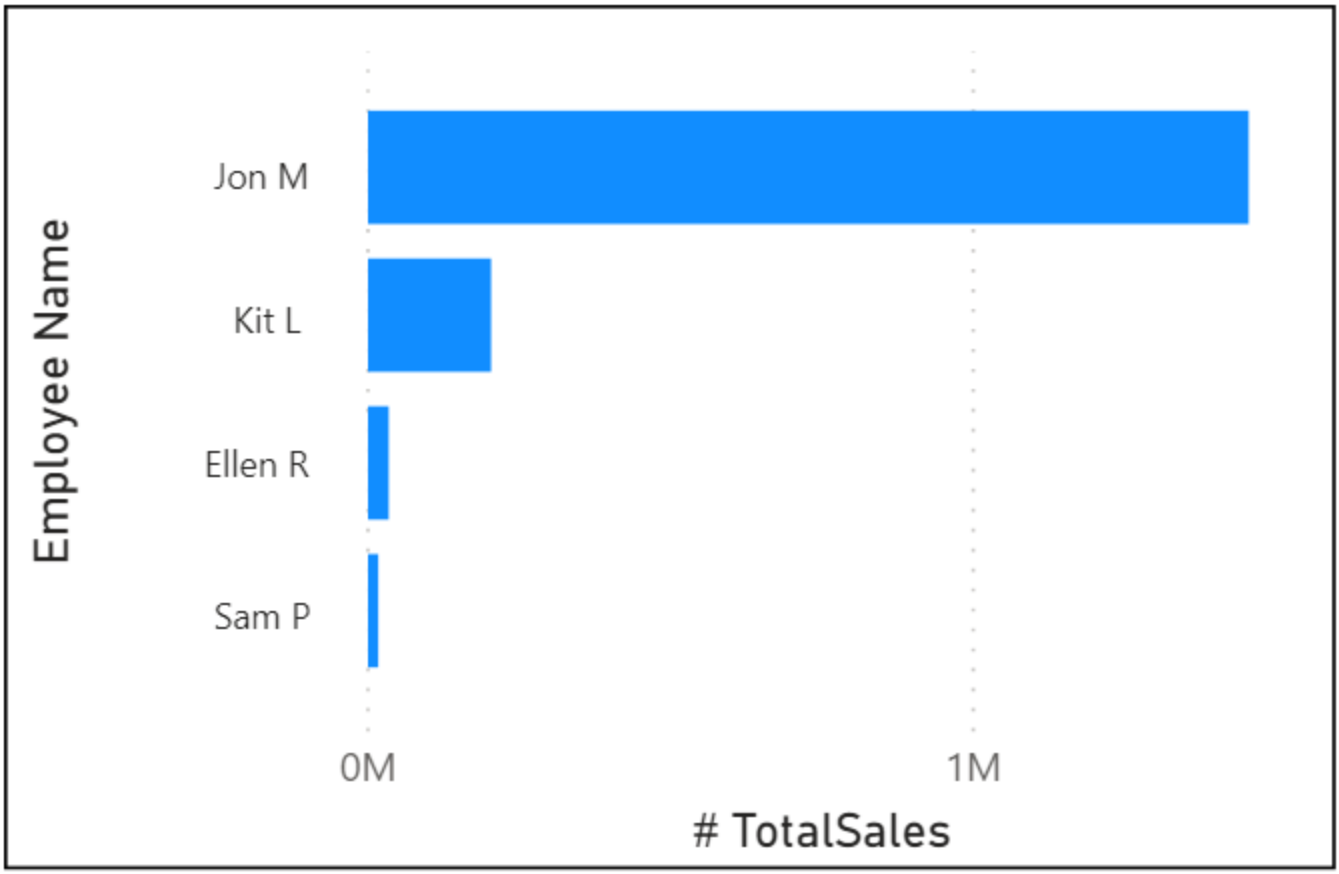
Dimension tables contain the details about the data in fact tables: **products, locations, employees, and order types.** These tables are connected to the fact table through key columns. Dimension tables are used to filter and group the data in fact tables. The dimension tables, by contrast, contain unique values, for instance, one row for each product in the Products table and one row for each customer in the Customer table. For the total sales orders visual, you could group the data so that you see total sales orders by product, in which product is data in the dimension table.

Fact tables are usually much larger than dimension tables because numerous events occur in fact tables, such as individual sales. Dimension tables are typically smaller because you are limited to the number of items that you can filter and group on. For instance, a year contains only so many months, and the United States is comprised of only a certain number of states.

Considering this information about fact tables and dimension tables, you might wonder how you can build this visual in Power BI.

The pertinent data resides in two tables, Employee and Sales, as shown in the following data model. Because the Sales table contains the sales order values, which can be aggregated, it is considered a fact table. The Employee table contains the specific employee's name, which filters the sales orders, so it would be a dimension table. The common column between the two tables, which is the primary key in the Employee table, is **EmployeeID**, so you can establish a relationship between the two tables based on this column.

When creating this relationship, you can build the visual according to the requirements, as shown in the following figure. If you did not establish this relationship, while keeping in mind the commonality between the two tables, you would have had more difficulty building your visual.

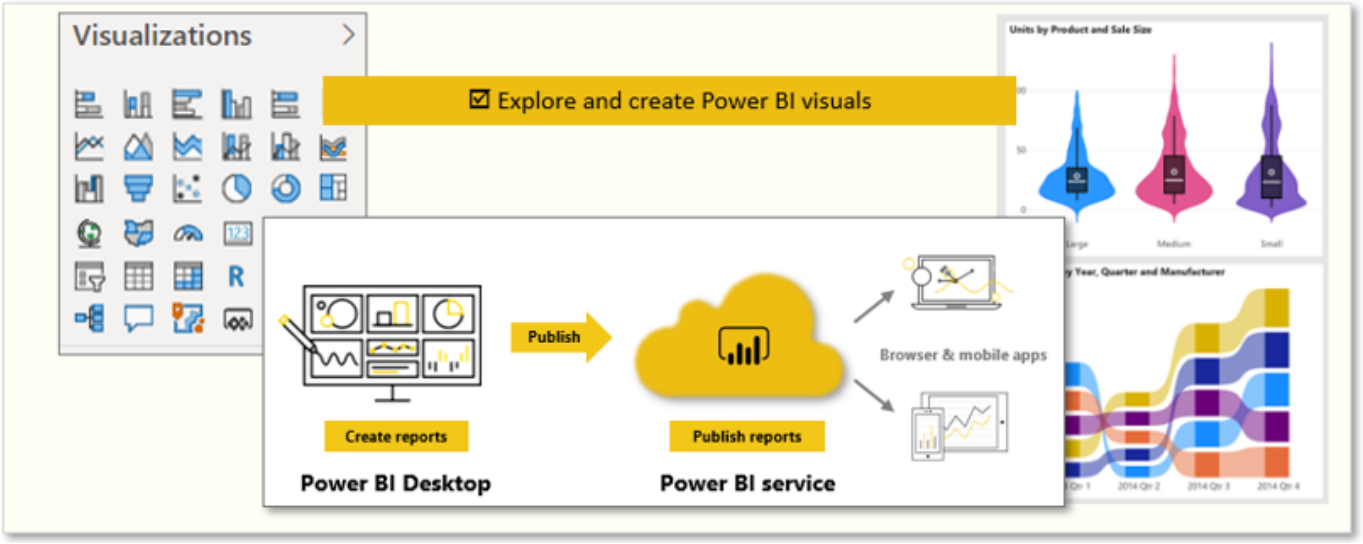


Star schemas and the underlying data model are the foundation of organized reports; the more time you spend creating these connections and design, the easier it will be to create and maintain reports.

Visualize data in power BI

Add visualization items to reports

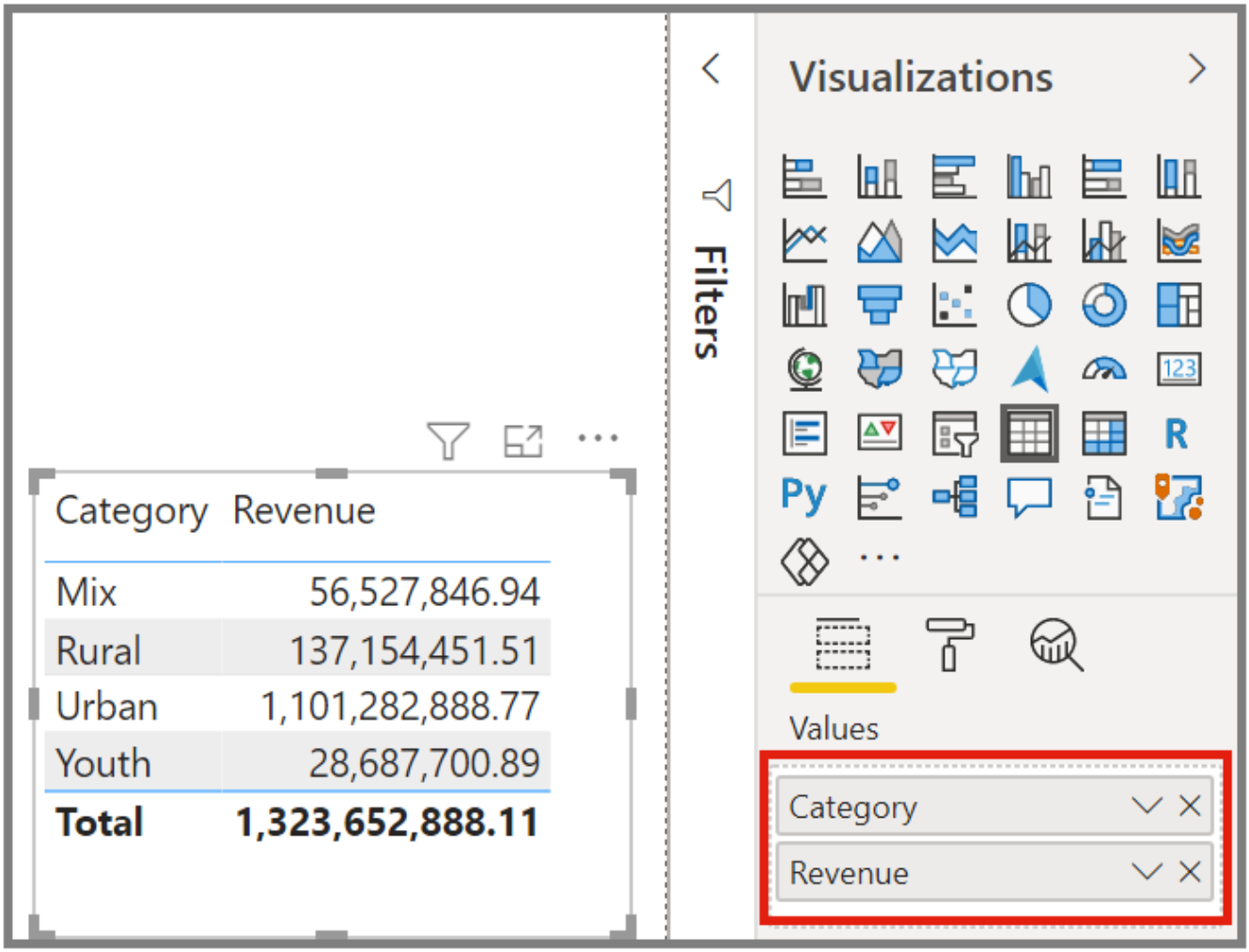
Power BI has a variety of visuals that you can use to report on the data in your data model. Visuals allow you to present the important information and insights that you discovered in the data in a compelling and insightful way. The report consumers rely on these visualizations as a gateway to the underlying data.



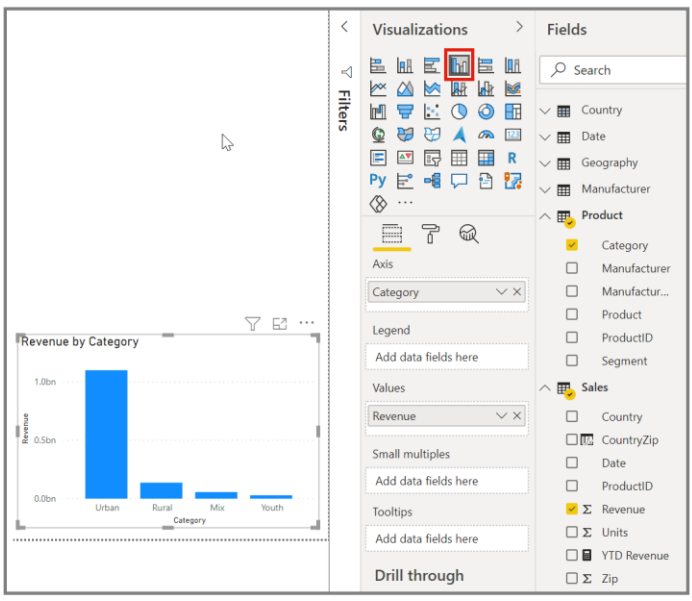
In Power BI Desktop, each visual is represented by an icon in the **Visualizations** pane. The types of visuals that are available include charts, maps, cards, a table, a matrix, and many more. You will learn how to select the correct visual later in this module.

This module is instructional, you will have an opportunity to practice in the lab near the end of the module.

In this example, you want to add a visualization to the report that displays sales data by category name. You start by selecting the **Category** and **Revenue** fields in the **Fields** pane. Power BI Desktop then automatically selects a visualization for you, depending on the data type of the fields that you selected. In this case, the default visualization type is a table.



While the visual is selected, you can change the visualization type by selecting a different visual from the **Visualizations** pane.



Choose an effective visualization

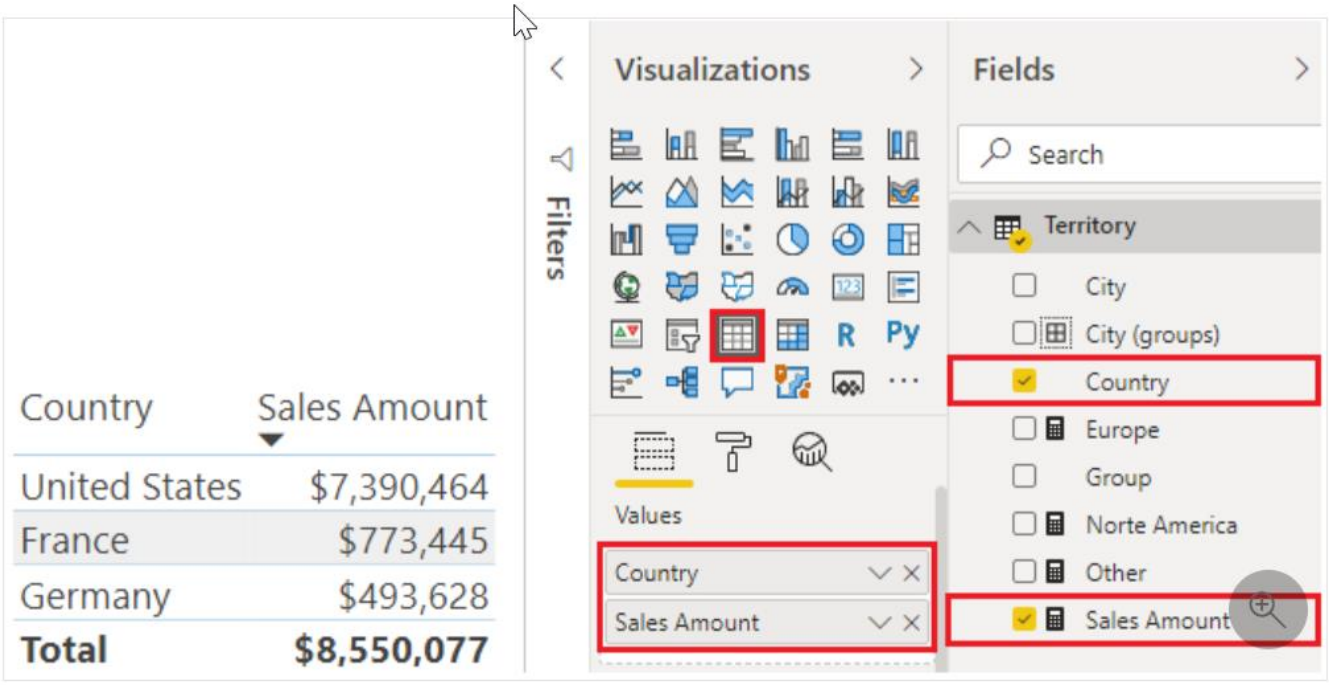
Power BI Desktop offers a range of out-of-the-box visualization options that are available directly from the **Visualizations** pane. When you select the fields that you want to display in a visualization, you can experiment with all the different visualization types to find the one that best suits your needs. If you can't find a visual that meets your needs, you can download other visuals from Microsoft AppSource or import your own custom visuals.

Depending on the type of data in your selected fields, one or more visualizations might not be suitable. For example, geographic data will not display well as a funnel chart or line chart visualization.

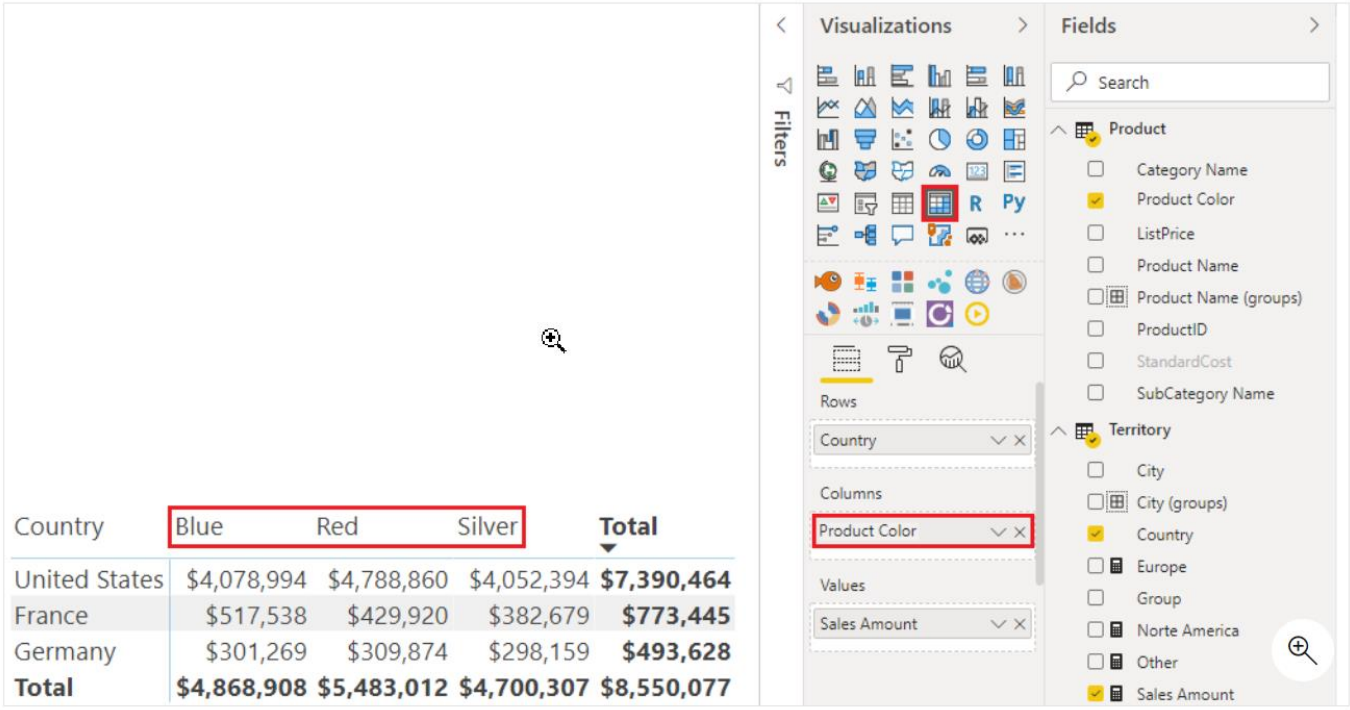
It is important that you choose an effective visualization to ensure that you display the data in the best way possible. The following sections outline the different types of visualizations that are available within Power BI Desktop, using the same data source for illustration purposes.

Table and Matrix visualizations

In the previous example, the **Table** visualization was selected by default. The table is a grid that contains related data in a logical series of rows and columns. The table supports two dimensions and it can also contain headers and a row for totals.



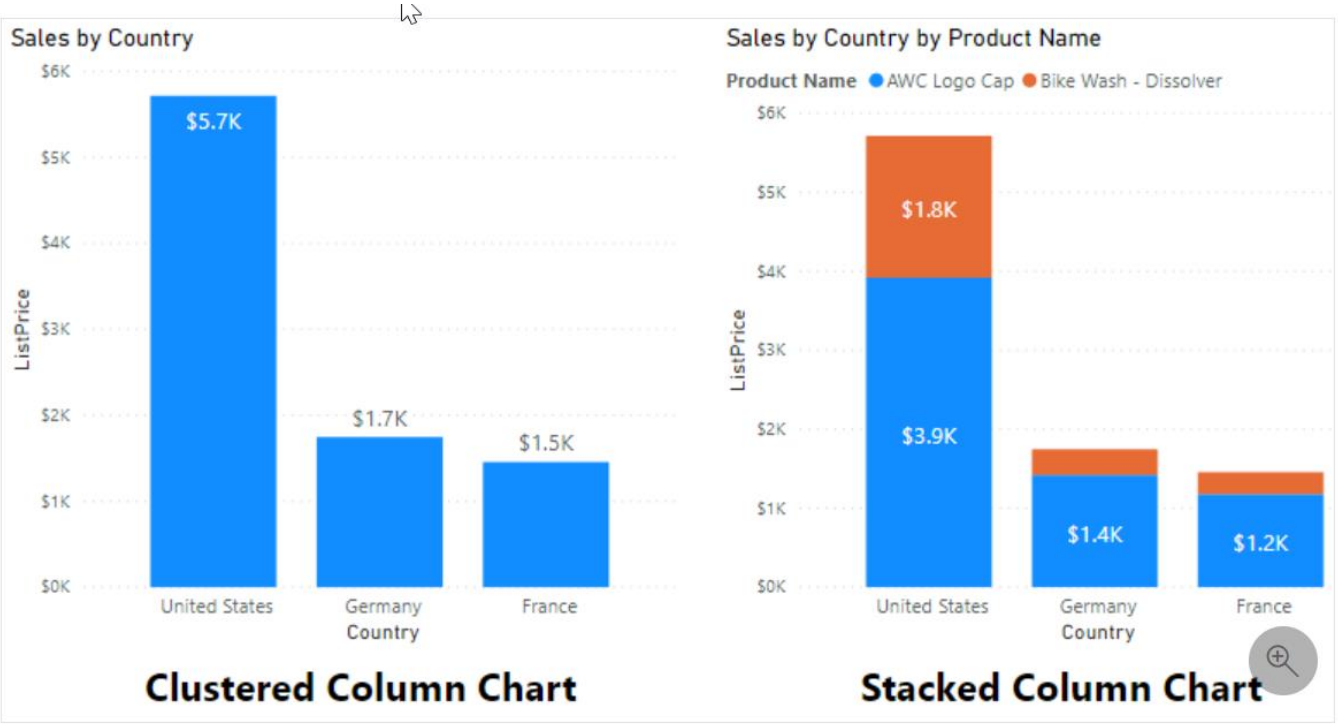
The **Matrix** visualization looks similar to the table visualization; however**, it allows you to select one or more elements (rows, columns, values) in the matrix** to cross-highlight other visuals on the report page. In the following image, notice that a new field called **Product Color** was added to the columns, and the available colors are now spanning across the table, with the categories listed in rows.



Bar and column charts

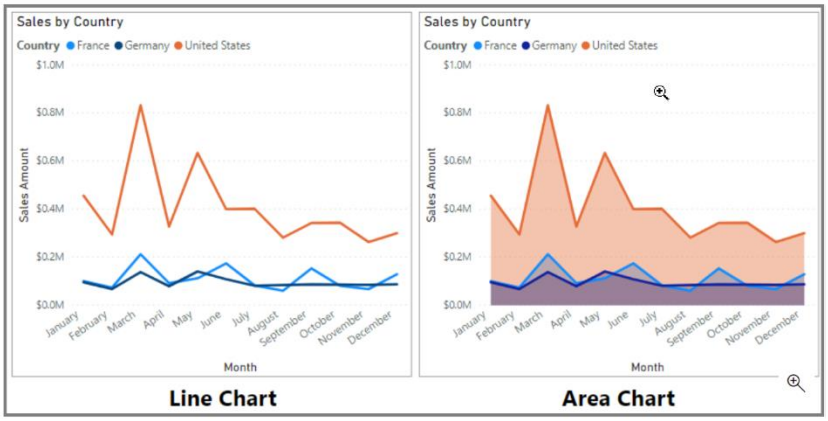
Power BI Desktop has a variety of bar and column chart visualizations that present specific data across different categories in a stacked or clustered format. The stacked format will stack the information items on top of each other.

For example, the following clustered column chart shows a single column with total sales for each country, whereas the stacked column chart shows data for sales by country, by product name. All sales data is stacked into one column to show you the total sales by country, broken down by how much each product contributed to the overall total sales.



Line and area charts

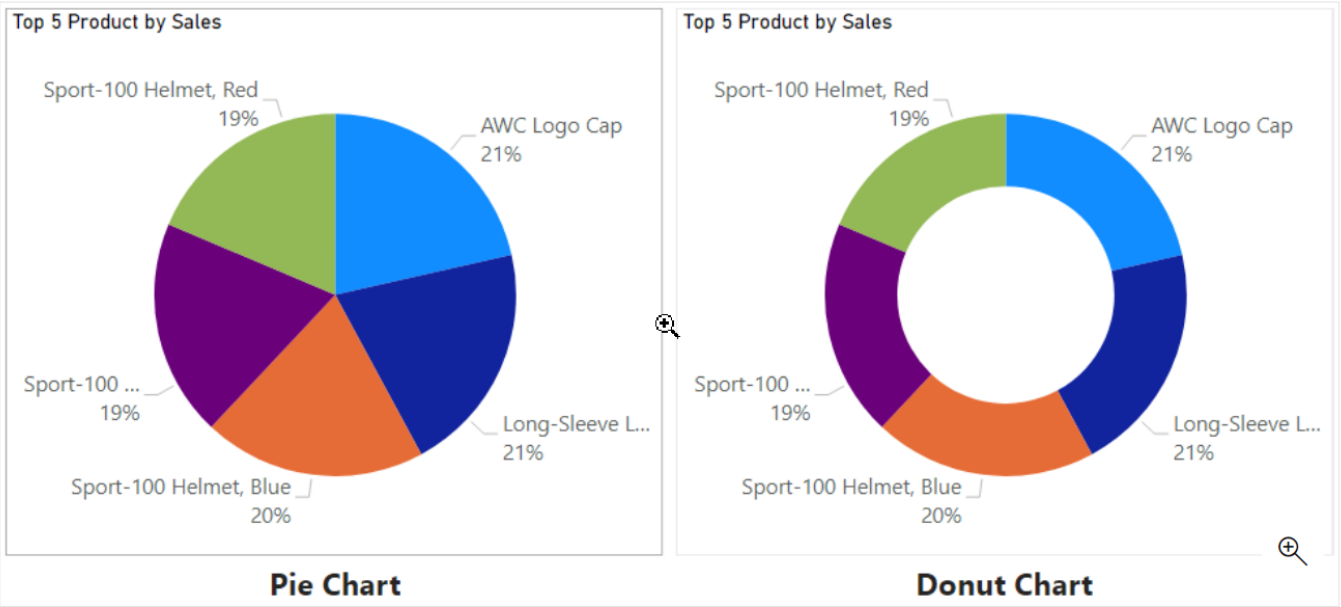
The **line chart** and **area chart** visualizations are beneficial in helping you present trends over time. The basic area chart is based on the line chart, with the area between axis and line filled in. The main difference between these two chart types is that the area chart highlights the magnitude of change over tim



Pie chart, donut chart, and Treemaps

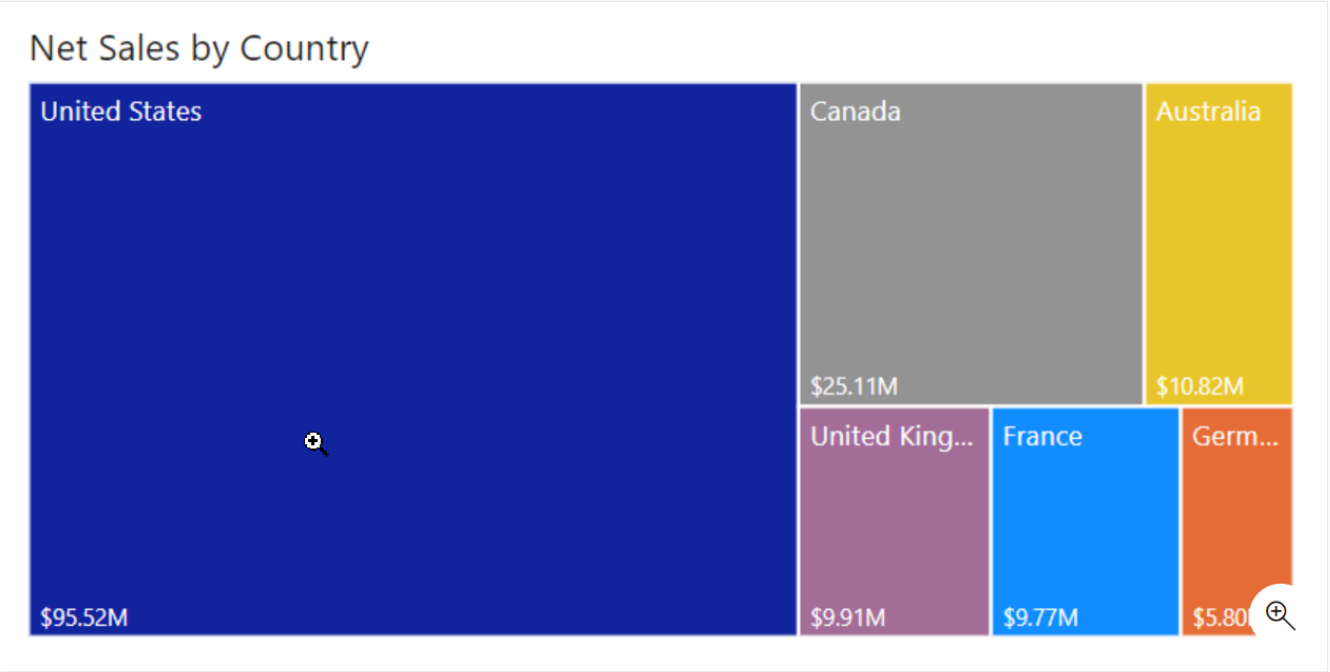
The **pie chart**, **donut chart**, and **Treemap** visualizations show you the relationship of parts to the whole by dividing the data into segments. From a data analysis perspective, these charts are not useful because interpreting the data that they present can be difficult. However, these charts are often used for aesthetic reasons due to the colorful segments that they display. These charts are best suited for illustrating percentages, such as the top five sales by product or country, or any other available categories.

The pie chart is a solid circle, whereas the donut chart has a center that is blank and allows space for a label or icon.



When using pie charts, donut charts, and **Treemaps**, try to avoid presenting too many categories because it results in thin slices (or rectangles) that provide no added value to the user. If you do need to present all categories in your dataset, it's better to use another type of visual, such as a column chart.

Pie charts and donut charts present data by dividing it into slices, while the **Treemap** visualization displays data as a set of nested rectangles. Each level of the hierarchy is represented by a colored rectangle (branch) containing smaller rectangles (leaves). The space inside each rectangle is allocated based on the value that is being measured. The rectangles are arranged in size from top left (largest) to bottom right (smallest).



A **Treemap** is ideal to visualize:

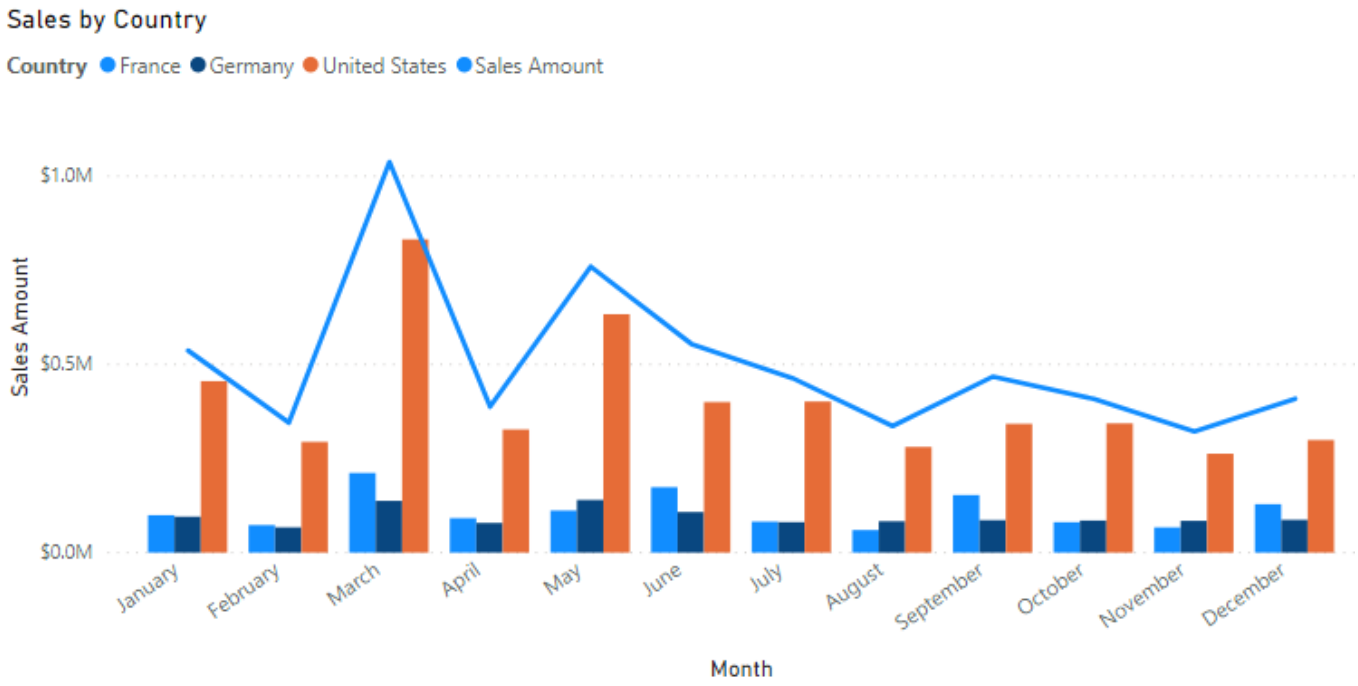
- Large amounts of hierarchical data when a bar chart can't effectively handle the large number of values.
- Proportions between each part and the whole.
- The distribution pattern of the measure across each level of categories in the hierarchy.
- Attributes, by using size and color coding.
- Spot patterns, outliers, most-important contributors, and exceptions.

Combo charts

The **combo** chart visualization is a combination of a column chart and a line chart that can have one or two Y axes. The combination of the two charts into one lets you:

- Compare multiple measures with different value ranges.
- Illustrate the correlation between two measures in one visual.

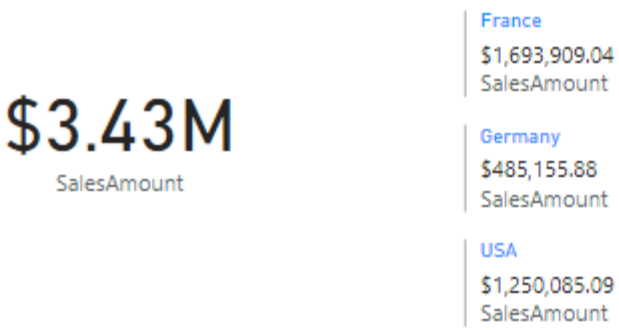
- Identify whether one measure meets the target that is defined by another measure.
- Conserve space on your report page.



Card visualization

The **card** visualization displays a single value: a single data point. This type of visualization is ideal for visualizing important statistics that you want to track on your Power BI dashboard or report, such as total value, YTD sales, or year-over-year change.

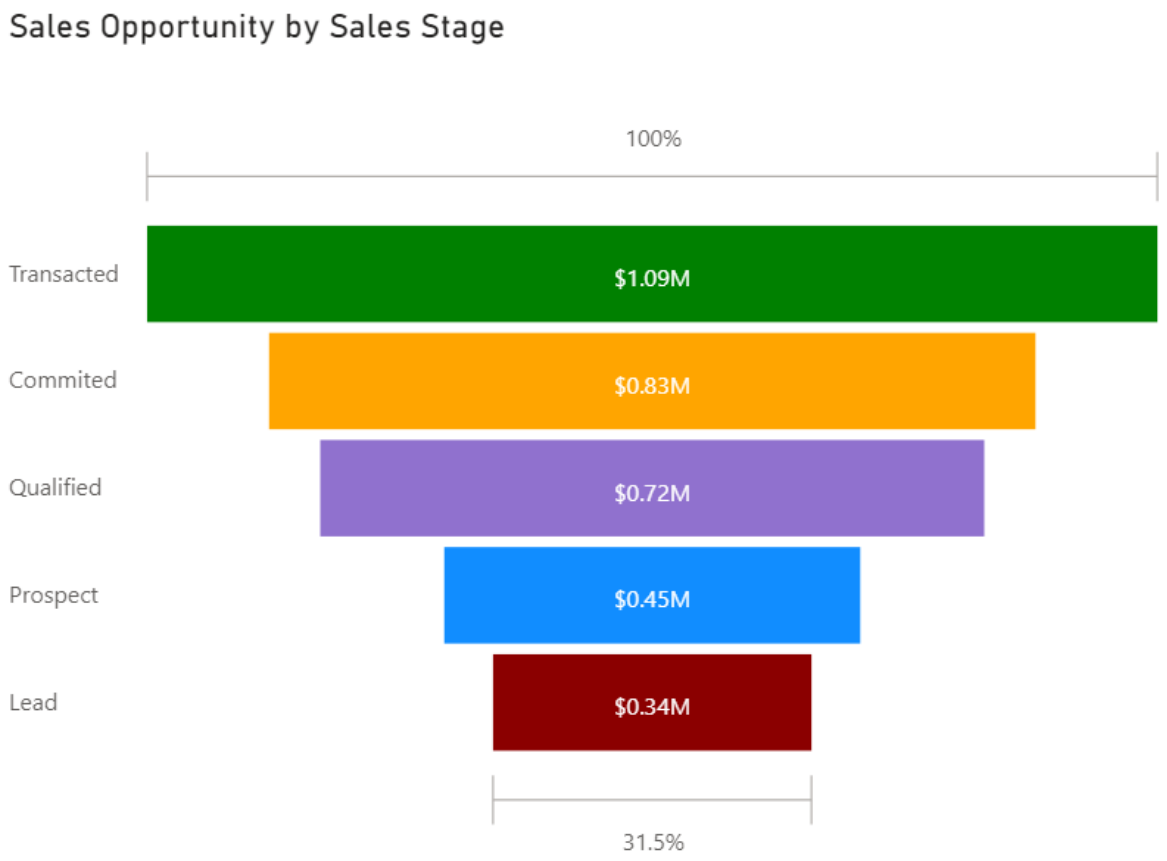
The **multi-row** card visualization displays one or more data points, with one data point for each row.



Funnel visualization

The **funnel** visualization displays a linear process that has sequential connected stages, where items flow sequentially from one stage to the next.

Funnel charts are most often seen in business or sales contexts. For example, they are useful for representing a workflow, such as moving from a sales lead to a prospect, through to a proposal and sale.



Funnel charts are great options in the following contexts:

- When the data is sequential and moves through at least four stages.
- When the number of items in the first stage is expected to be greater than the number of items in the final stage.

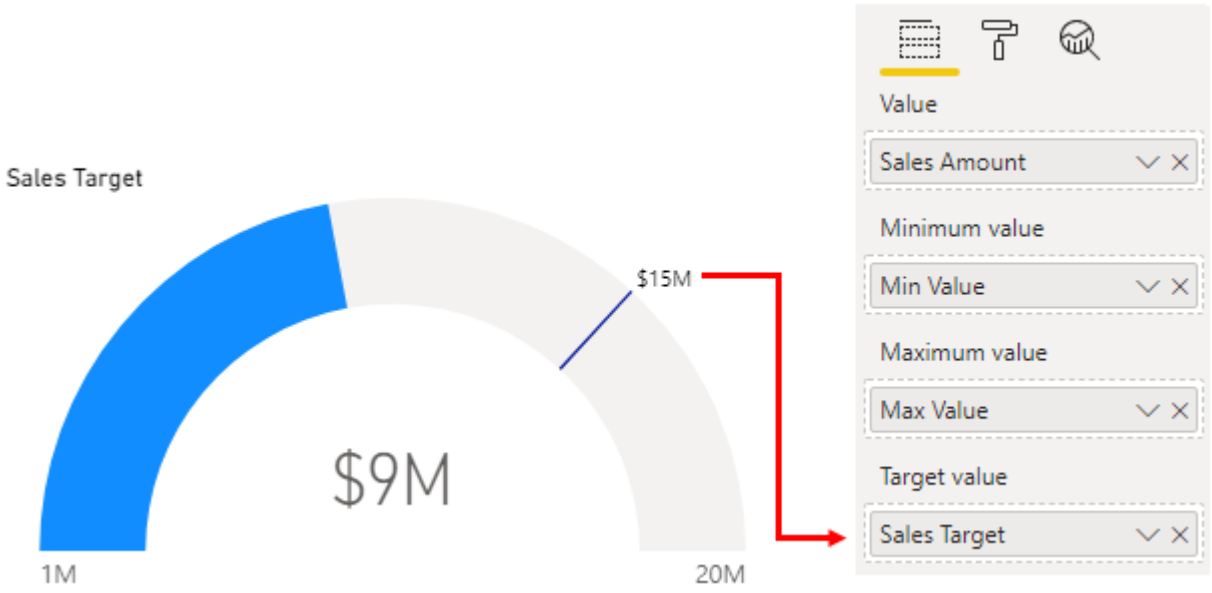
- To calculate a potential outcome (revenue, sales, deals, and so on) by stages.
- To calculate and track conversion and retention rates.
- To reveal bottlenecks in a linear process.

Gauge chart

A radial gauge chart has a circular arc and displays a single value that measures progress toward a goal or target.

The value at the end of the arc represents the defaulted maximum value, which will always be double the actual value. To create a realistic visual, you should always specify each of the values. You can accomplish this task by dropping the correct field that contains an amount into the **Target value**, **Minimum value**, and **Maximum value** fields on the **Visualization** pane.

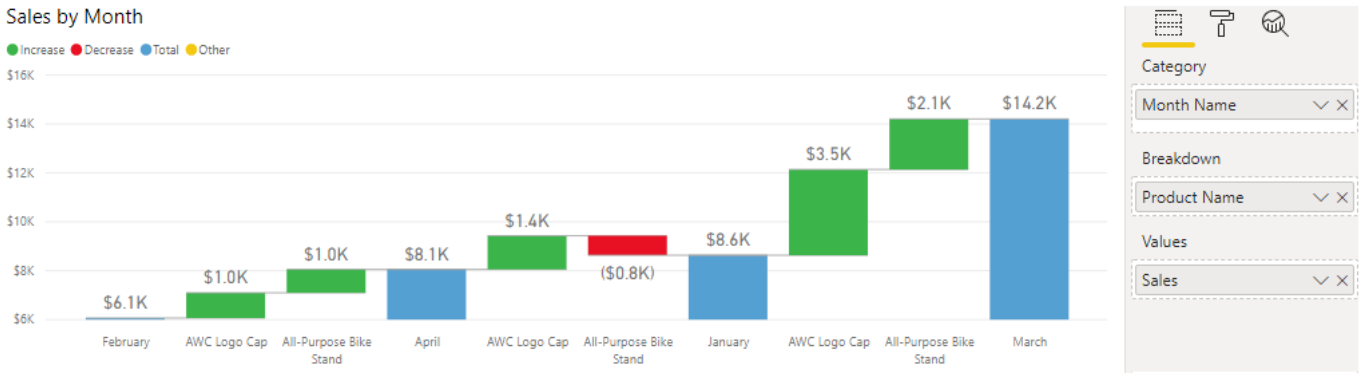
The shading in the arc represents the progress toward that target. The value inside the arc represents the progress value. Power BI spreads all possible values evenly along the arc, from the minimum (left-most value) to the maximum (right-most value).



Radial gauges can be used to show the progress that is being made toward a goal or target, or they can show the health of a single measure. However, radial gauges do take up a lot of space in comparison to the insights that they provide. It is more effective to use a pair of gauges with a spark line so users can see the trend and know what to do about it.

Waterfall visualization

The **waterfall** visualization (also known as a **bridge chart**) shows a running total as values are added or subtracted, which is useful in displaying a series of positive and negative changes. The chart consists of color-coded columns, so you can quickly identify increases and decreases. The initial and the final value columns often start on the horizontal axis, while the intermediate values are floating columns.



Waterfall charts can be used to:

- Visualize changes over time or across different categories.
- Audit the major changes that contribute to the total value.
- Plot your organization's annual profit by showing various sources of revenue to help determine the total profit (or loss).
- Illustrate the beginning and ending headcount for your organization in a year.
- Visualize how much money you earn and spend each month and the running balance for your account.

Scatter chart

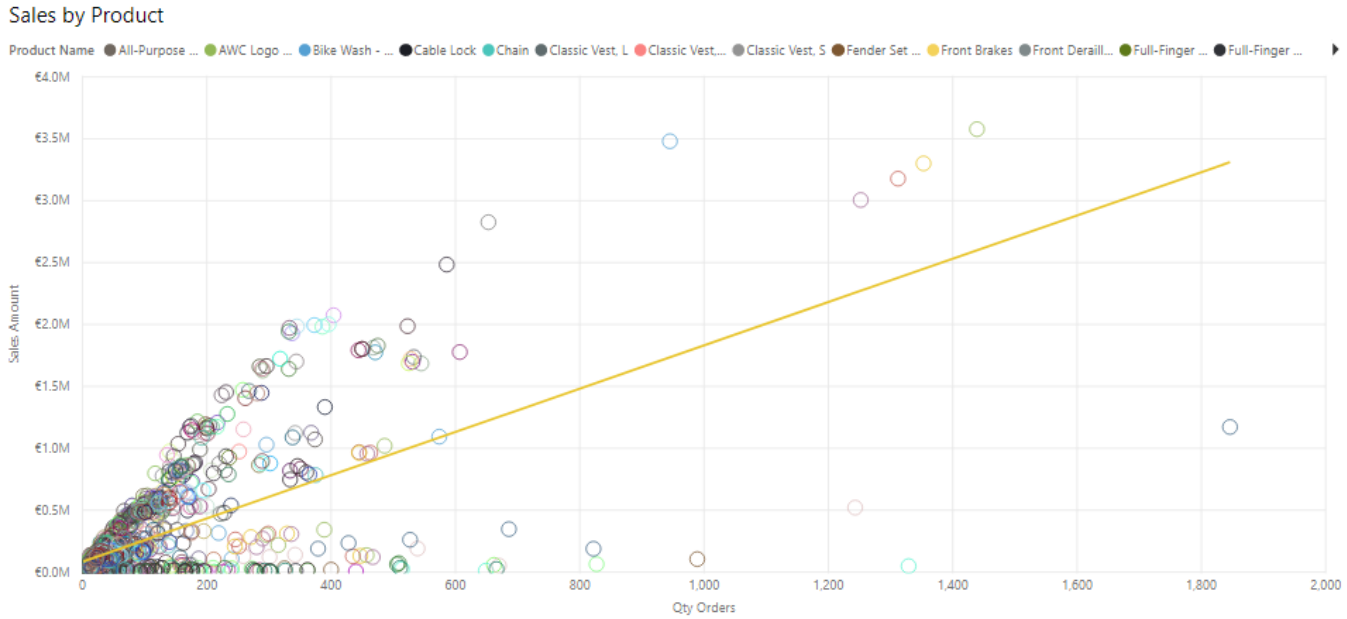
The **scatter** chart visualization is effective when you are comparing large numbers of data points without regard to time. **The scatter chart has two value axes to show: one set of numerical data along a horizontal axis** and **another set of numerical values along a vertical axis**. The chart displays **points at the intersection of an X and Y numerical value, combining these values into single data points**. These data points might be distributed evenly or unevenly across the horizontal axis, depending on the data. You can set the number of data points, up to a maximum of 10,000.

You might want to use a scatter chart instead of a line chart because it allows you to change the scale of the horizontal axis. **Scatter charts also allow you to:**

- **Show relationships between two numerical values.**
- **Plot two groups of numbers as one series of x and y coordinates.**
- **Turn the horizontal axis into a logarithmic scale.**
- **Display worksheet data that includes pairs or grouped sets of values.**
- **Show patterns in large sets of data, for example, by showing linear or non-linear trends, clusters, and outliers.**
- **Compare large numbers of data points without regard to time. The more data that you include in a scatter chart, the better the comparisons that you can make.**

The following example shows a scatter chart that displays outliers (anomalies) with a trendline going up. The chart clearly shows that most products were sold at the same quantity, and only some products were sold in larger quantities. By identifying those outliers, you can run further analysis and break them down by country and

region, which can help to improve logistics, decrease costs, and increase customer satisfaction.



Maps

Power BI integrates with Bing Maps to provide default map coordinates (a process called geocoding), so you can create maps. Together, they use algorithms to identify the correct location; however, sometimes, it's a best guess.

A *basic* map (*bubble* map) is used to associate categorical and quantitative information with spatial locations. This type of map visual display's precise geographical locations of data points on a map, as illustrated in the following image.

A *fill* map uses shading, tinting, or patterns to display how a value differs in proportion across a geographical region. Similarly, *shape* maps use colors to display relative comparisons of geographical regions. You can also use an ArcGIS map to display graphical information in a more interactive way.

Slicer visualization

The **slicer** visualization is a standalone chart that can be used to filter the other visuals on the page. Slicers provide a more advanced and customized way of filtering, in comparison to the **Filters** pane, which is suited to more basic filtering operations. You can learn more about these two filtering options in another module.

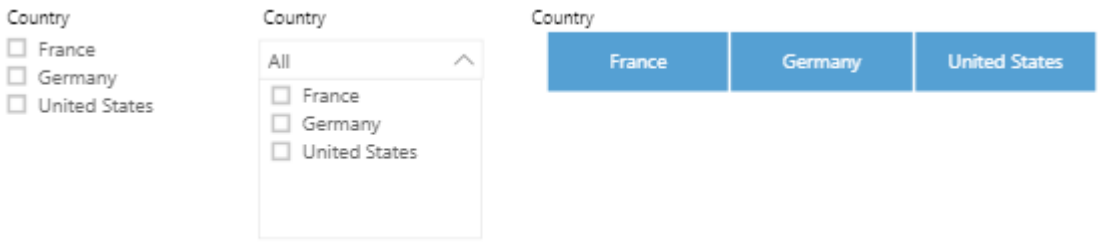
Slicers come in many different formats, including list, drop-down, and buttons, and they can be formatted to allow the selection of only one, many, or all available values.

Slicers are ideal to:

- Visualize commonly used or important filters on the report canvas for easier access.
- Simplify your ability to see the current filtered state without having to open a drop-down list.
- Filter by columns that are unneeded and hidden in the data tables.
- Create more focused reports by putting slicers next to important visuals.

Tip

Using a slicer that is set to a drop-down format will defer the queries that are being sent to the dataset and can help improve performance.

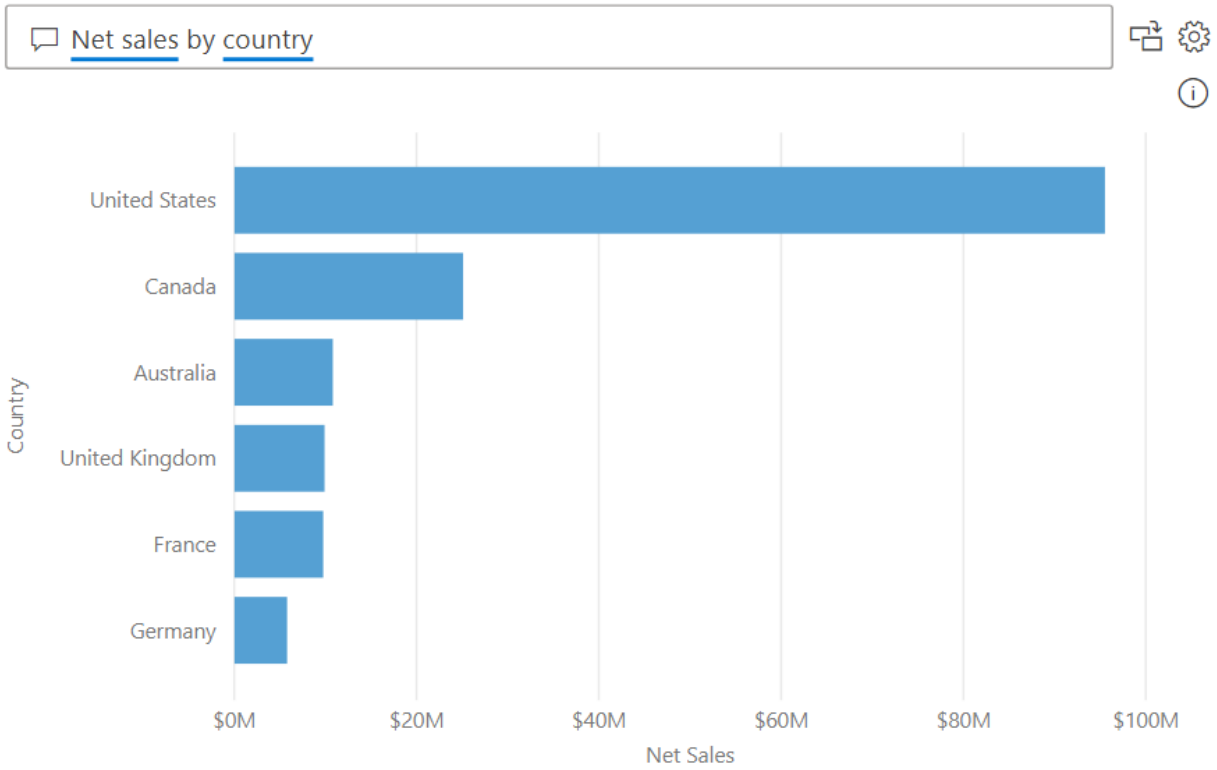


Q&A visualization

The **Q&A** visualization allows you to ask natural language questions and get answers in the form of a visual. This ability to ask questions is valuable to consumers and to you, the report author. This visualization type can help you create visuals in the report, and it can also be used as a tool for consumers to get answers quickly.

The Q&A visualization consists of the following four core components:

- The question box, where users enter their question and are shown suggestions to help them complete the question.
- A pre-populated list of suggested questions.
- An icon that users can select to convert the Q&A visual into a standard visual.
- An icon that users can select to open Q&A tooling, which allows designers to configure the underlying natural language engine. When entering natural language queries with Power BI Q&A, you can specify the visual type in your query. The following example illustrates how to implement **Net sales by country**.



Data Analysis in power BI

Explore statistical summary

Data is often intertwined with statistics because statistics are one way in which you can explore your data. Statistics show you the distribution of your data and help you to identify key takeaways and trends and determine whether outliers exist.

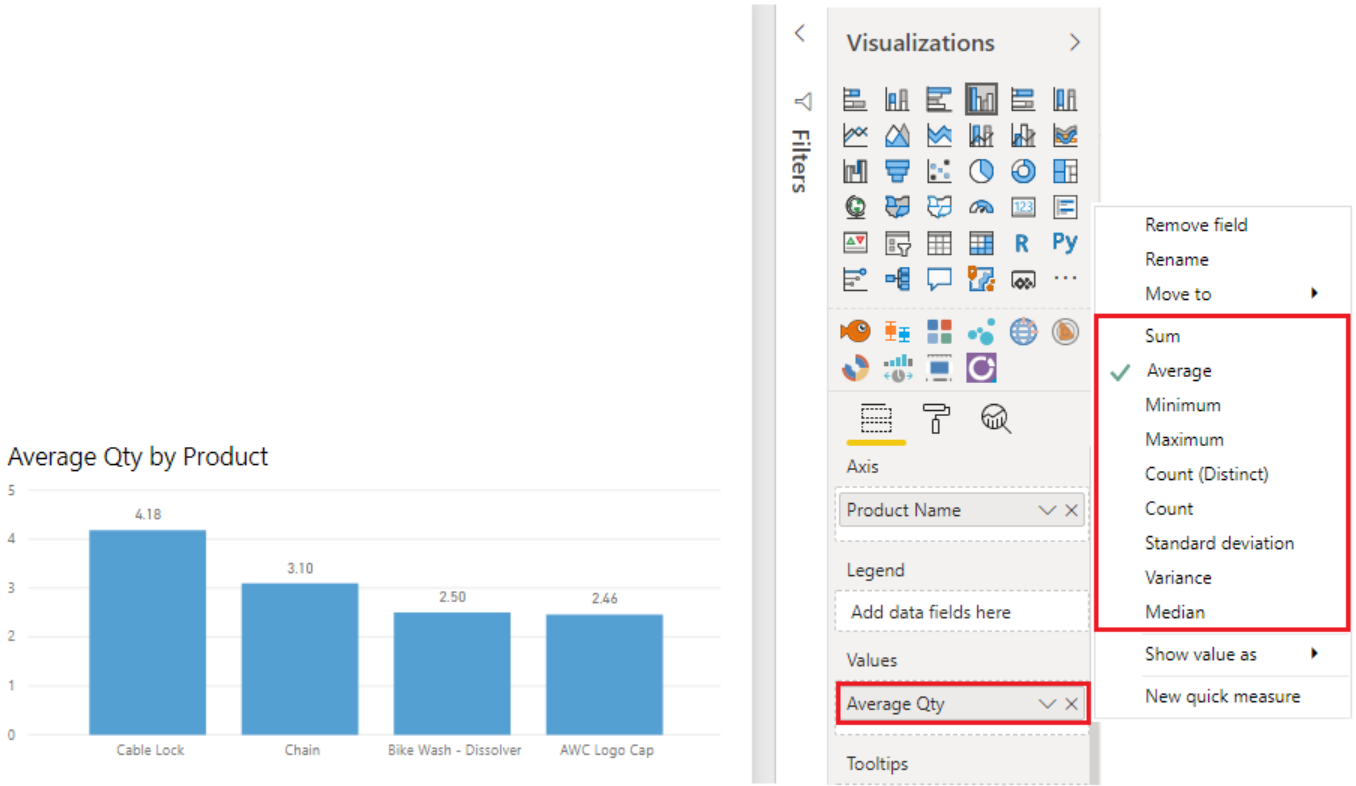
The statistical summary is the information that provides a quick and simple description of your data. Power BI has many functions that help you to conduct a statistical analysis, such as Data Analysis Expressions (DAX) functions, visuals such as histograms and bell curves, advanced analytics visuals, and statistical programming languages such as Python and R.

Exploring the statistical summary gives the user a high-level view of the available data, where they can see clusters, patterns on behavioral data, data averages, and more. They can gain insights about their data that will help drive business decisions.

For example, the Supply Chain team asks you to create a report that shows the frequency of orders for certain products and what the top 10 products are in terms of sales.

Statistical functions

Power BI Desktop has a number of DAX functions that you can use to get quick statistics based on your data. You can access these quick functions by right-clicking the **Values** field in the **Visualizations** pane, as illustrated in the following image.



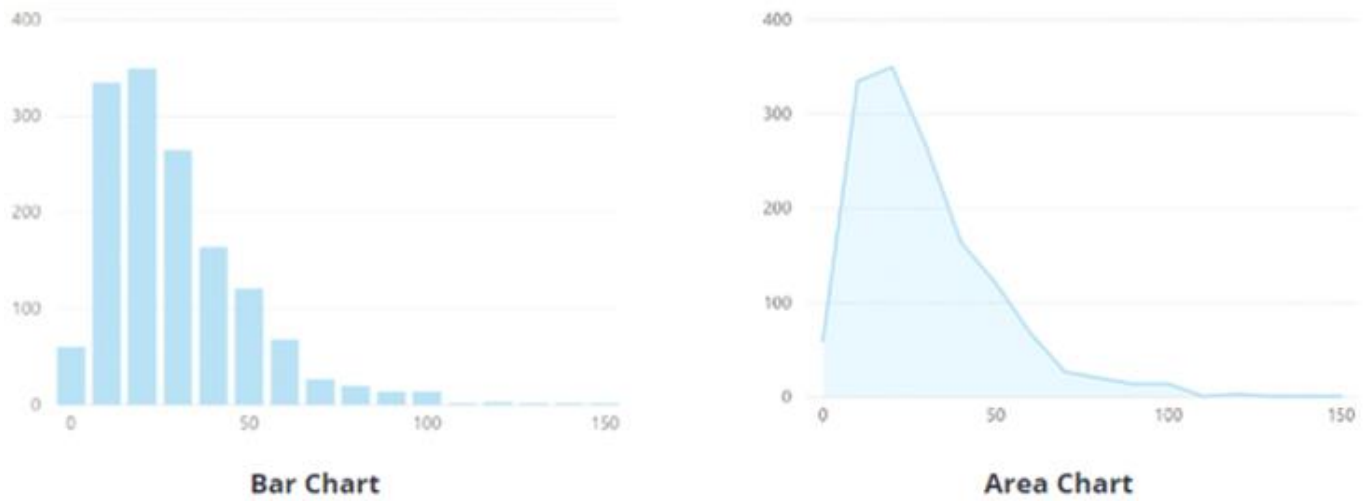
However, to avoid performance issues, it's better to create the statistical measures yourself by using DAX functions to calculate average, sum, min, max, and so on. For example, to analyze the inventory data to find the average order quantity for each product, you could use the following formula:

DAXCopy

```
Average Qty =
AVERAGE ( Sales[Order Qty] )
```

Histogram

Histograms and bell curves are the most common way to display statistics about your datasets. In Power BI terms, you can represent a histogram with one of the bar or column chart visuals and represent a bell curve with an area chart visual, as illustrated in the following image. You can also use the Q&A visual to ask a direct question about the top or bottom items in a list.



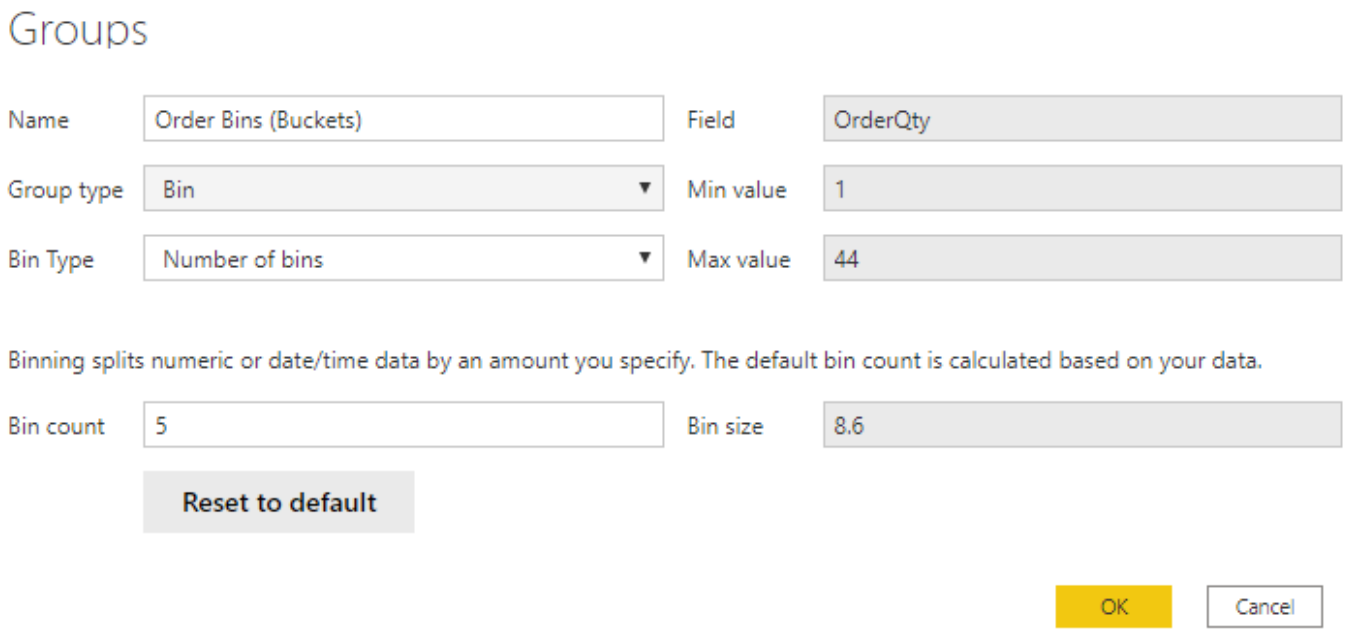
A typical bar or column chart visual in Power BI relates two data points: a measure and a dimension. A histogram differs slightly from a standard bar chart in that it only visualizes a single data point.

In this example, you use the clustered column chart visual to present a histogram that determines the order quantities by order sizes.

You start by selecting the clustered column chart icon on the **Visualization** pane. Next, create a new grouping for the x-axis. You will learn more about grouping and binning later in this module, but they are useful in this context also.

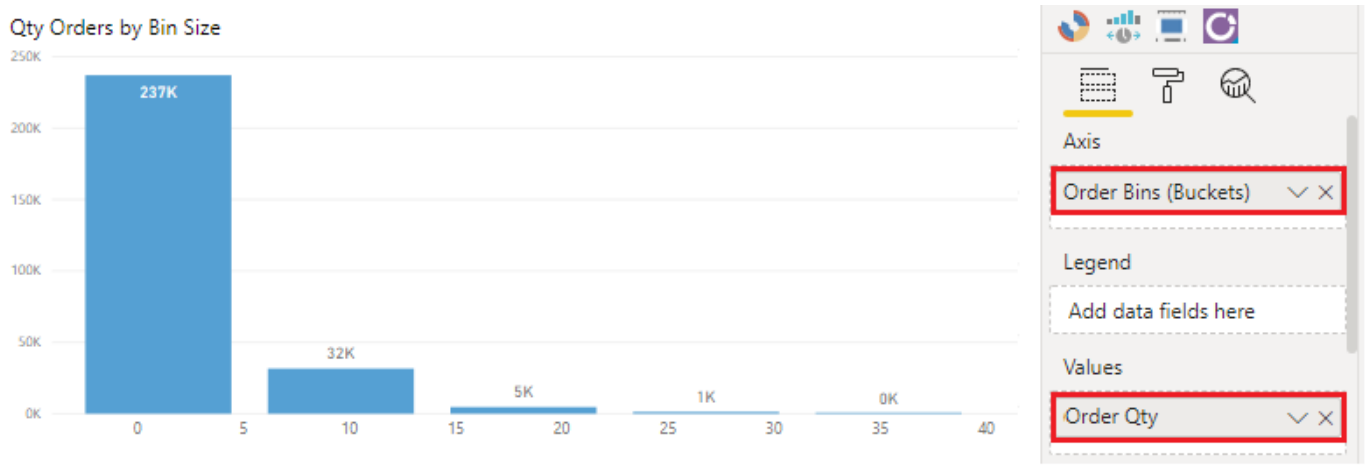
To create the group, in the **Fields** pane, right-click the data field that you want to analyze and then select **New Group**. In this case, you use the **OrderQty** field. In the **Groups** window that displays, set up the bin group as follows:

1. Rename the group as **Order Bins (Buckets)**.
2. Set the **Group type** option to **Bin** and the **Bin Type** option to **Number of bins**.
3. Enter **5** as the **Bin count**, **1** as the **Min value**, and **44** as the **Max value**.



Next, populate the visual as follows:

1. Drag and drop the **OrderQty** field from the **Fields** pane into the **Value** field on the **Visualizations** pane.
2. Drag and drop the **Order Bins (Buckets)** group from the **Fields** pane into the **Axis** field on the **Visualizations** pane.



The visual now shows that the data is grouped into buckets on the x-axis, with the order quantities of that variable on the y-axis.

You have now produced a histogram that displays the order quantity (**OrderQty** field) by order size buckets for the Supply Chain team.

Top N analysis

The TOPN DAX function returns the top N rows of a specified table. The Top N analysis is a great way to present data that might be important, **such as the top 10 selling products, top 10 performers in an organization, or top 10 customers.** Alternatively, you can look at it from the other perspective and present the bottom 10 items in a list, in other words, the worst performers. Depending on the requirements, you might want to use one or both of these options.

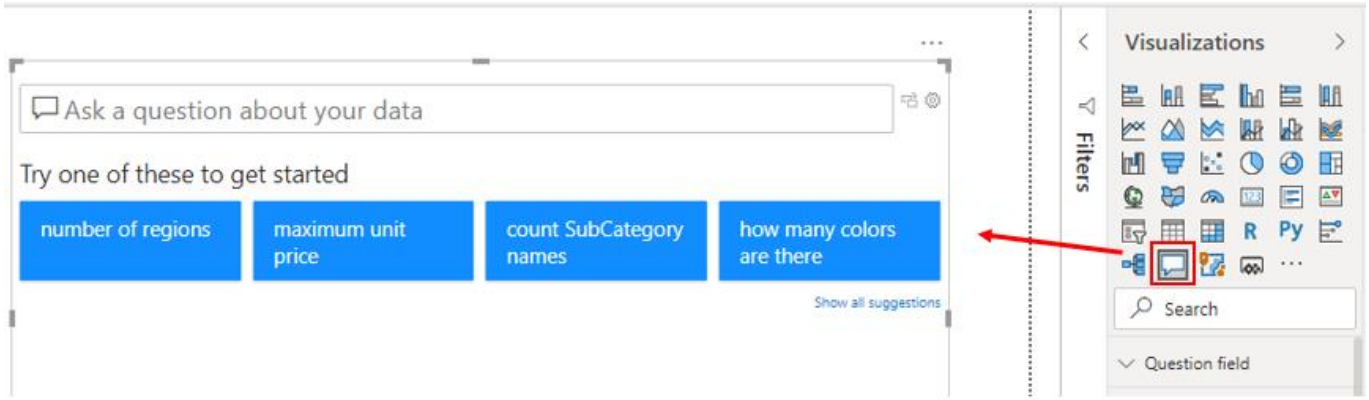
In this example, the Supply Chain team wants to know what the top 10 selling products are. You accomplish this task in one of three ways: by using a Q&A visual, using a Top N filter, or writing a DAX formula.

Use the Q&A visual to find the top N

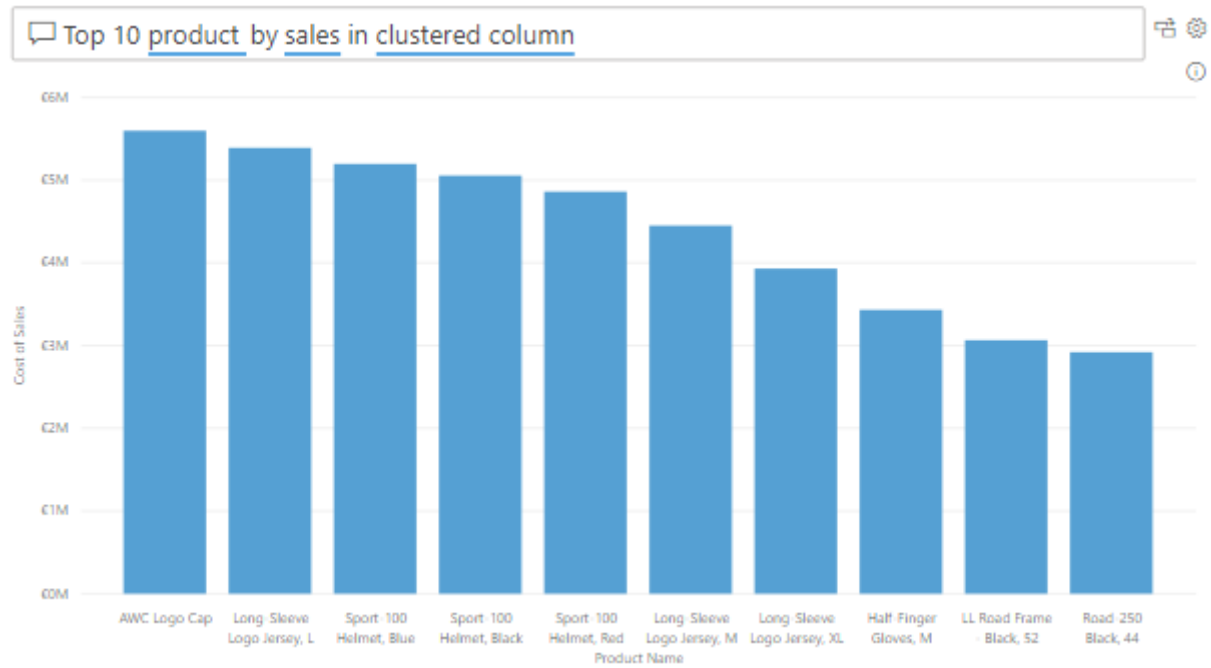
You've created a report for the Supply Chain team, and now the team members have questions about various other views or insights that they are interested in.

Power BI has a built-in Q&A visual that allows users to ask their own questions and get answers so you don't have to address each individual question. The Q&A visual is an effective tool because it allows users to quickly get answers about the data independently, which saves time for everyone involved. The Q&A visual is unique in that it does not require knowledge of Power BI to use the visual; users can ask their question and they, too, can create insightful visuals.

Add the **Q&A** visualization to your report, and then reposition the visual and customize its formatting, as required.

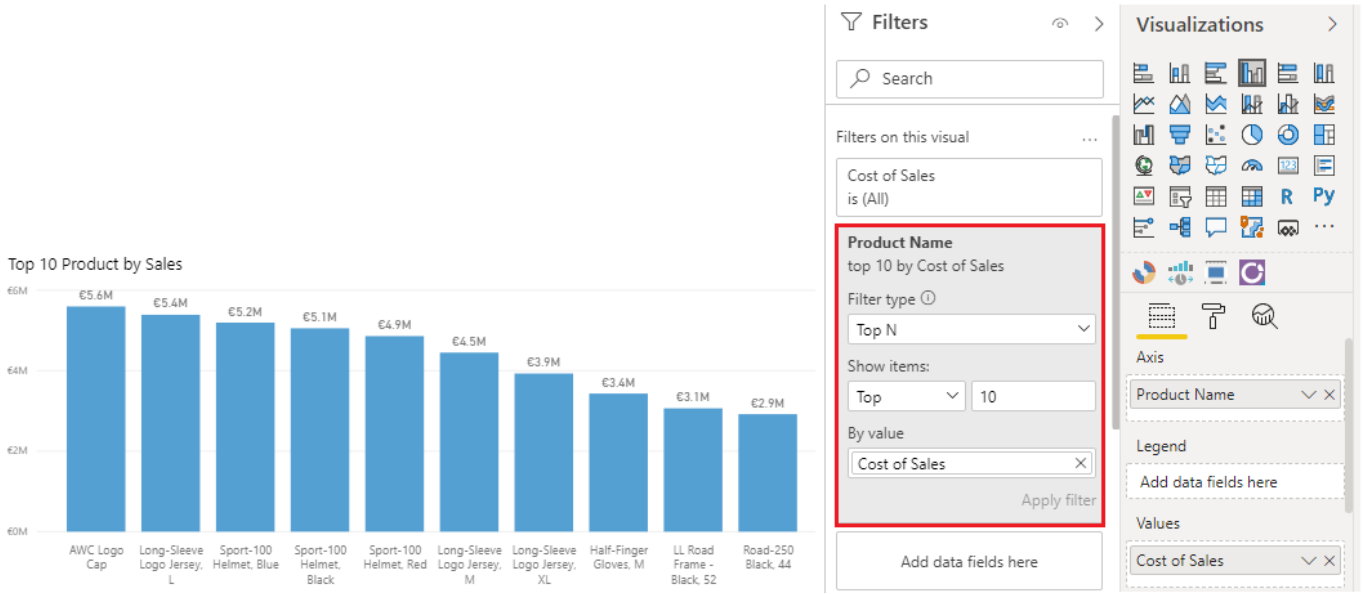


Now, you can use the visual to get answers. In this case, you want to know what the top 10 selling products are, so you enter a question such as, "What are my top 10 products by sales?" Power BI will automatically display those results for you.



Use a Top N filter type

Top N is a filtering option that is available on the **Filters** pane. Select the field that you want to analyze on your report page (in this example, it's the **Product Name** field). In the **Filters** pane, expand the **Filter type** list and select **Top N**. In the **Show items** settings, select **Top** and **10**. Then, select **Cost of Sales** as the value that you want to filter the field by. The visual updates accordingly.



Use a TOPN DAX function

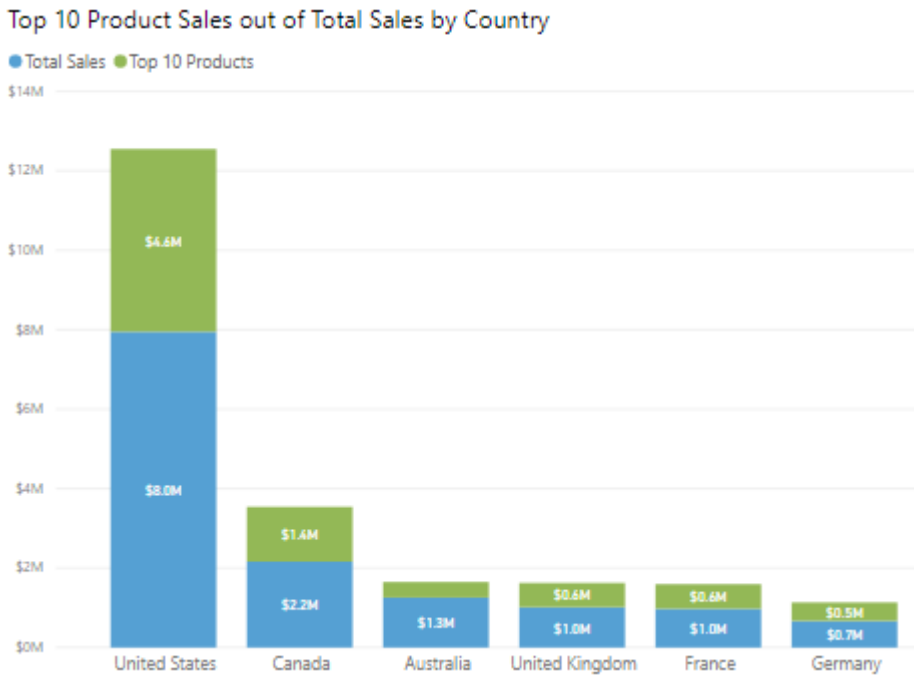
You can also calculate your top 10 products in DAX by using the TOPN function. This option could be useful if you want to present the top 10 in a different context, such as how much of the top 10 best-selling products contributed toward the overall total sales.

Start by creating a new measure called **Top 10 Products**. Then, use the TOPN function, along with the SUMX function, to calculate your top 10 products by total sales, as follows:

DAXCopy

```
Top 10 Products =
SUMX ( TOPN ( 10, Product, Product[Total Sales]), [Total Sales] )
```

The following image shows the top 10 products versus total sales for comparison.



You can adjust the DAX formula to present the same result in percentages.

For more information about the statistical capabilities of Power BI, see Statistical Functions - DAX.