

# Glossary

## Data Analytics

### Terms and Definitions

---



## A

**A/B testing:** The process of testing two variations of the same web page to determine which page is more successful at attracting user traffic and generating revenue

**Absolute reference:** A reference within a function that is locked so that rows and columns won't change if the function is copied

**Access control:** Features such as password protection, user permissions, and encryption that are used to protect a spreadsheet

**Accuracy:** The degree to which data conforms to the actual entity being measured or described

**Action-oriented question:** A question whose answers lead to change

**Administrative metadata:** Metadata that indicates the technical source of a digital asset

**Aesthetic (R):** A visual property of an object in a plot

**Agenda:** A list of scheduled appointments

**Aggregation:** The process of collecting or gathering many separate pieces into a whole

**Algorithm:** A process or set of rules followed for a specific task

**Aliasing:** Temporarily naming a table or column in a query to make it easier to read and write

**Alternative text:** Text that provides an alternative to non-text content, such as images and videos

**Analysis:** The process used to make sense of the data collected

**Analytical skills:** Qualities and characteristics associated with using facts to solve problems

**Analytical thinking:** The process of identifying and defining a problem, then solving it by using data in an organized, step-by-step manner

**Annotation:** Text that briefly explains data or helps focus the audience on a particular aspect of the data in a visualization

**Anscombe's quartet:** Four datasets that have nearly identical summary statistics but contain different plotted values

**Area chart:** A data visualization that uses individual data points for a changing variable connected by a continuous line with a filled in area underneath

**Argument (R):** Information needed by a function in R in order to run

**Arithmetic operator:** An operator used to perform basic math operations such as addition, subtraction, multiplication, and division

**Array:** A collection of values in spreadsheet cells

**Assignment operator:** An operator used to assign values to variables and vectors

**Attribute:** A characteristic or quality of data used to label a column in a table

**Audio file:** Digitized audio storage usually in an MP3, AAC, or other compressed format

**AVERAGE:** A spreadsheet function that returns an average of the values from a selected range

**AVERAGEIF:** A spreadsheet function that returns the average of all cell values from a given range that meet a specified condition

## B

**Bad data source:** A data source that is not reliable, original, comprehensive, current, and cited (ROCCC)

**Balance:** The design principle of creating aesthetic appeal and clarity in a data visualization by evenly distributing visual elements

**Bar graph:** A data visualization that uses size to contrast and compare two or more values

**Bias:** A conscious or subconscious preference in favor of or against a person, group of people, or thing

**Big data:** Large, complex datasets typically involving long periods of time, which enable data analysts to address far-reaching business problems

**Boolean data:** A data type with only two possible values, usually true or false

**Borders:** Lines that can be added around two or more cells on a spreadsheet

**Box plot:** A data visualization that displays the distribution of values along an x-axis

**Bubble chart:** A data visualization that displays individual data points as bubbles, comparing numeric values by their relative size

**Bullet graph:** A data visualization that displays data as a horizontal bar chart moving toward a desired value

**Business metric:** A standard of measurement used to solve a business task

**Business task:** The question or problem data analysis resolves for a business

## C

**C#:** An object-oriented programming language used to create games and mobile apps in the .NET open source developer platform

**C++:** An extension of the C programming language that is used to create console games, such as those for Xbox

**Calculated field:** A new field within a pivot table that carries out certain calculations based on the values of other fields

**Calculus:** A branch of mathematics that involves the study of rates of change and the changes between values that are related by a function

**CASE:** A SQL statement that returns records that meet conditions by including an if/then statement in a query

**Case study:** A common way for employers to assess job skills and gain insight into how a candidate approaches common data-related challenges

**CAST:** A SQL function that converts data from one datatype to another

**Causation:** When an action directly leads to an outcome, such as a cause-effect relationship

**Cell reference:** A cell or a range of cells in a worksheet typically used in formulas and functions

**Changelog:** A file containing a chronologically ordered list of modifications made to a project

**Channel:** A visual aspect or variable that represents characteristics of the data in a visualization

**Chart:** A graphical representation of data from a worksheet

**Circle view:** A data visualization that shows comparative strength in data

**Clean data:** Data that is complete, correct, and relevant to the problem being solved

**Cloud:** A place to keep data online, rather than a computer hard drive

**Cluster:** A collection of data points on a data visualization with similar values

**COALESCE:** A SQL function that returns non-null values in a list

**Coding:** The process of writing instructions to a computer in the syntax of a specific programming language

**Column chart:** A data visualization that uses individual data points for a changing variable, represented as vertical columns

**Combo chart:** A data visualization that combines more than one visualization type

**Compatibility:** How well two or more datasets work together

**Completeness:** The degree to which data contains all desired components or measures

**Computer programming:** The process of giving instructions to a computer in order to perform an action or set of actions

**CONCAT:** A SQL function that adds strings together to create new text strings that can be used as unique keys

**CONCATENATE:** A spreadsheet function that joins together two or more text strings

**Conditional formatting:** A spreadsheet tool that changes how cells appear when values meet specific conditions

**Conditional statement:** A declaration that if a certain condition holds, then a certain event must take place

**Confidence interval:** A range of values that conveys how likely a statistical estimate reflects the population

**Confidence level:** The probability that a sample size accurately reflects the greater population

**Confirmation bias:** The tendency to search for or interpret information in a way that confirms pre-existing beliefs

**Consent:** The aspect of data ethics that presumes an individual's right to know how and why their personal data will be used before agreeing to provide it

**Consistency:** The degree to which data is repeatable from different points of entry or collection

**Context:** The condition in which something exists or happens

**Continuous data:** Data that is measured and can have almost any numeric value

**CONVERT:** A SQL function that changes the unit of measurement of a value in data

**Cookie:** A small file stored on a computer that contains information about its users

**Correlation:** The measure of the degree to which two variables change in relationship to each other

**COUNT:** A spreadsheet function that counts the number of cells within a range that meet a specified condition

**COUNTA:** A spreadsheet function that counts the total number of values within a specified range that meet specified criteria

**COUNTIF:** A spreadsheet function that returns the number of cells within a range that match a specified value

**COUNT DISTINCT:** A SQL function that only returns the distinct values in a specified range

**CRAN (Comprehensive R Archive Network) (R):** An online archive with R packages, source code, manuals, and documentation

**CREATE TABLE:** A SQL clause that adds a temporary table to a database that can be used by multiple people

**Cross-field validation:** A process that ensures certain conditions for multiple data fields are satisfied

**CSS (Cascading Style Sheets):** A programming language used for web page design that controls graphic elements and page presentation

**CSV (comma-separated values) file:** A delimited text file that uses a comma to separate values

**Currency:** The aspect of data ethics that presumes individuals should be aware of financial transactions resulting from the use of their personal data and the scale of those transactions

## D

**Dashboard:** A tool that monitors live, incoming data

**Dashboard filter:** A tool for showing only the data that meets a specific criteria while hiding the rest

**Data:** A collection of facts

**Data aggregation:** The process of gathering data from multiple sources and combining it into a single, summarized collection

**Data analysis:** The collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision-making

**Data analysis process:** The six phases of ask, prepare, process, analyze, share, and act whose purpose is to gain insights that drive informed decision-making

**Data analyst:** Someone who collects, transforms, and organizes data in order to draw conclusions, make predictions, and drive informed decision-making

**Data analytics:** The science of data

**Data anonymization:** The process of protecting people's private or sensitive data by eliminating identifying information

**Data bias:** When a preference in favor of or against a person, group of people, or thing systematically skews data analysis results in a certain direction

**Data blending:** A Tableau method that combines data from multiple data sources

**Data composition:** The process of combining the individual parts in a visualization and displaying them together as a whole

**Data constraints:** The criteria that determine whether a piece of a data is clean and valid

**Data design:** How information is organized

**Data-driven decision-making:** Using facts to guide business strategy

**Data ecosystem:** The various elements that interact with one another in order to produce, manage, store, organize, analyze, and share data

**Data element:** A piece of information in a dataset

**Data engineer:** A professional who transforms data into a useful format for analysis and gives it a reliable infrastructure

**Data ethics:** Well-founded standards of right and wrong that dictate how data is collected, shared, and used

**Data frame:** A collection of columns containing data, similar to a spreadsheet or SQL table

**Data governance:** A process for ensuring the formal management of a company's data assets

**Data-inspired decision-making:** Exploring different data sources to find out what they have in common

**Data integrity:** The accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle

**Data interoperability:** The ability to integrate data from multiple sources and a key factor leading to the successful use of open data among companies and governments

**Data life cycle:** The sequence of stages that data experiences, which include plan, capture,

manage, analyze, archive, and destroy

**Data manipulation:** The process of changing data to make it more organized and easier to read

**Data mapping:** The process of matching fields from one data source to another

**Data merging:** The process of combining two or more datasets into a single dataset

**Data model:** A tool for organizing data elements and how they relate to one another

**Data privacy:** Preserving a data subject's information any time a data transaction occurs

**Data range:** Numerical values that fall between predefined maximum and minimum values

**Data replication:** The process of storing data in multiple locations

**Data science:** A field of study that uses raw data to create new ways of modeling and understanding the unknown

**Data security:** Protecting data from unauthorized access or corruption by adopting safety measures

**Data storytelling:** Communicating the meaning of a dataset with visuals and a narrative that are customized for an audience

**Data strategy:** The management of the people, processes, and tools used in data analysis

**Data structure:** A format for organizing and storing data

**Data transfer:** The process of copying data from a storage device to computer memory or from one computer to another

**Data type:** An attribute that describes a piece of data based on its values, its programming language, or the operations it can perform

**Data validation:** A tool for checking the accuracy and quality of data

**Data validation process:** The process of checking and rechecking the quality of data so that it is complete, accurate, secure and consistent

**Data visualization:** The graphical representation of data

**Data warehousing specialist:** A professional who develops processes and procedures to effectively store and organize data

**Database:** A collection of data stored in a computer system

**Dataset:** A collection of data that can be manipulated or analyzed as one unit

**DATEDIF:** A spreadsheet function that calculates the number of days, months, or years between two dates

**Decision tree:** A tool that helps analysts make decisions about critical features of a visualization

**Delimiter:** A character that indicates the beginning or end of a data item

**Density map:** A data visualization that represents concentrations, with color representing the number or frequency of data points in a given area on a map

**Descriptive metadata:** Metadata that describes a piece of data and can be used to identify it at a later point in time

**Design thinking:** A process used to solve complex problems in a user-centric way

**Digital photo:** An electronic or computer-based image usually in BMP or JPG format

**Dirty data:** Data that is incomplete, incorrect, or irrelevant to the problem to be solved

**Discrete data:** Data that is counted and has a limited number of values

**DISTINCT:** A keyword that is added to a SQL SELECT statement to retrieve only non-duplicate entries

**Distribution graph:** A data visualization that displays the frequency of various outcomes in a sample

**Diverging color palette:** A color theme that displays two ranges of data values using two different hues, with color intensity representing the magnitude of the values

**Donut chart:** A data visualization where segments of a ring represent data values adding up to a whole

**dplyr (R):** An R package in Tidyverse that offers a consistent set of functions to complete common data-manipulation tasks

**DROP TABLE:** A SQL clause that removes a temporary table from a database

**Duplicate data:** Any record that inadvertently shares data with another record

**Dynamic visualizations:** Data visualizations that are interactive or change over time

## E

**Emphasis:** The design principle of arranging visual elements to focus the audience's attention on important information in a data visualization

**Engagement:** Capturing and holding someone's interest and attention during a data



presentation

**Equation:** A calculation that involves addition, subtraction, multiplication, or division (also called a math expression)

**Estimated response rate:** The average number of people who typically complete a survey

**Ethics:** Well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues

**Experimenter bias:** The tendency for different people to observe things differently (Refer to Observer bias)

**External data:** Data that lives, and is generated, outside of an organization

## F

**Facets (R):** A series of functions that splits data into subsets in a matrix of panels

**Factor (R):** An object that stores categorical data where the data values are limited and usually based on a finite group, such as country or year

**Fairness:** A quality of data analysis that does not create or reinforce bias

**Field:** A single piece of information from a row or column of a spreadsheet; in a data table, typically a column in the table

**Field length:** A tool for determining how many characters can be keyed into a spreadsheet field

**Fill handle:** A box in the lower-right-hand corner of a selected spreadsheet cell that can be dragged through neighboring cells in order to continue an instruction

**Filled map:** A data visualization that colors areas in a map based on measurements or dimensions

**Filtering:** The process of showing only the data that meets a specified criteria while hiding the rest

**Find and replace:** A tool that finds a specified search term and replaces it with something else

**First-party data:** Data collected by an individual or group using their own resources

**Float:** A number that contains a decimal

**Foreign key:** A field within a database table that is a primary key in another table (Refer to primary key)

**Formula:** A set of instructions used to perform a calculation using the data in a spreadsheet

**Framework:** The context a presentation needs to create logical connections that tie back to the business task and metrics

**FROM:** The section of a query that indicates from which table(s) to extract the data

**Function:** A preset command that automatically performs a specified process or task using the data in a spreadsheet

**Function (R):** A body of reusable code for performing specific tasks in R

**FWF (fixed-width file):** A text file with a specific format, which enables the saving of textual data in an organized fashion

## G

**GAM (generalized additive model) smoothing (R):** A process for smoothing plots with a large number of points

**Gantt chart:** A data visualization that displays the duration of events or activities on a timeline

**Gap analysis:** A method for examining and evaluating the current state of a process in order to identify opportunities for improvement in the future

**Gauge chart:** A data visualization that shows a single result within a progressive range of values

**General Data Protection Regulation of the European Union (GDPR):** Policy-making body in the European Union created to help protect people and their data

**Geolocation:** The geographical location of a person or device by means of digital information

**Geom (R):** The geometric object used to represent data

**ggplot2 (R):** An R package in Tidyverse that creates a variety of data visualizations by applying different visual properties to the data variables in R

**Good data source:** A data source that is reliable, original, comprehensive, current, and cited (ROCCC)

**GROUP BY:** A SQL clause that groups rows that have the same values from a table into summary rows

## H

**HAVING:** A SQL clause that adds a filter to a query instead of the underlying table that can only be used with aggregate functions

**head() (R):** An R function that returns a preview of the column names and the first few rows of a dataset

**Header:** The first row in a spreadsheet that labels the type of data in each column

**Headline:** Text at the top of a visualization that communicates the data being presented

**Heat map:** A data visualization that uses color contrast to compare categories in a dataset

**Highlight table:** A data visualization that uses conditional formatting and color on a table

**Histogram:** A data visualization that shows how often data values fall into certain ranges

**HTML5:** A programming language that provides structure for web pages and connects to hosting platforms

**Hypothesis:** A theory that one might try to prove or disprove with data

**Hypothesis testing:** A process to determine if a survey or experiment has meaningful results

## I

**IDE (Integrated Development Environment):** A software application that brings together all the tools a data analyst may want to use in a single place

**Incomplete data:** Data that is missing important fields

**Inconsistent data:** Data that uses different formats to represent the same thing

**Incorrect/inaccurate data:** Data that is complete but inaccurate

**INNER JOIN :** A SQL function that returns records with matching values in both tables

**Inner query:** A SQL subquery that is inside of another SQL statement

**Internal data:** Data that lives within a company's own systems

**Interpretation bias:** The tendency to interpret ambiguous situations in a positive or negative way

## J

**Java:** A programming language widely used to create enterprise web applications that can run on multiple clients

**JOIN:** A SQL function that is used to combine rows from two or more tables based on a related column

# K

# L

**Label:** Text in a visualization that identifies a value or describes a scale

**Labels and annotations (R):** A group of R functions used for customizing a plot

**Leading question:** A question that steers people toward a certain response

**LEFT:** A function that returns a set number of characters from the left side of a text string

**LEFT JOIN:** A SQL function that will return all the records from the left table and only the matching records from the right table

**Legend:** A tool that identifies the meaning of various elements in a data visualization

**LEN:** A function that returns the length of a text string by counting the number of characters it contains

**Length:** The number of characters in a text string

**Library:** A directory containing all of a data analyst's installed packages

**LIMIT:** A SQL clause that specifies the maximum number of records returned in a query

**Line graph:** A data visualization that uses one or more lines to display shifts or changes in data over time

**List:** A vector whose elements can be of any type

**Live data:** Data that is automatically updated

**Loess smoothing (R):** A process used for smoothing plots with fewer than 1,000 points

**Log file:** A computer-generated file that records events from operating systems and other software programs

**Logical operator:** An operator that returns a logical data type

**Long data:** A dataset in which each row is one time point per subject, so each subject has data in multiple rows

# M

**Mandatory:** A data value that cannot be left blank or empty

**Map:** A data visualization that organizes data geographically

**Mapping (R):** The process of matching up a specific variable in a dataset with a specific aesthetic

**Margin of error:** The maximum amount that sample results are expected to differ from those of the actual population

**Mark:** A visual object in a data visualization such as a point, line, or shape

**MATCH:** A spreadsheet function used to locate the position of a specific lookup value

**Math expression:** A calculation that involves addition, subtraction, multiplication, or division (also called an equation)

**Math function:** A function that is used as part of a mathematical formula

**Matrix:** A two-dimensional collection of data elements with rows and columns

**MAX:** A function that returns the largest numeric value from a range of cells

**MAXIFS:** A spreadsheet function that returns the maximum value from a given range that meets a specified condition

**McCandless Method:** A method for presenting data visualizations that moves from general to specific information

**Measurable question:** A question whose answers can be quantified and assessed

**Mental model:** A data analyst's thought process and approach to a problem

**Mentor:** Someone who shares knowledge, skills, and experience to help another grow both professionally and personally

**Merger:** An agreement that unites two organizations into a single new one

**Metadata:** Data about data

**Metadata repository:** A database created to store metadata

**Metric:** A single, quantifiable type of data that is used for measurement

**Metric goal:** A measurable goal set by a company and evaluated using metrics

**MID:** A function that returns a segment from the middle of a text string

**MIN:** A spreadsheet function that returns the smallest numeric value from a range of cells

**MINIFS:** A spreadsheet function that returns the minimum value from a given range that meets a specified condition

**Modulo:** An operator (%) that returns the remainder when one number is divided by another

**Movement:** The design principle of arranging visual elements to guide the audience's eyes from one part of a data visualization to another

**mutate() (R):** An R function that makes changes to a dataframe separating and merging columns or creating new variables

## N

**Naming conventions:** Consistent guidelines that describe the content, creation date, and version of a file in its name

**Narrative:** (Refer to story)

**Nested:** Code that performs a particular function and is contained within code that performs a broader function

**Nested function:** A function that is completely contained within another function

**Networking:** Building relationships by meeting people both in person and online

**Nominal data:** A type of qualitative data that is categorized without a set order

**Normalized database:** A database in which only related data is stored in each table

**Notebook:** An interactive, editable programming environment for creating data reports and showcasing data skills

**Null:** An indication that a value does not exist in a dataset

## O

**Observation:** The attributes that describe a piece of data contained in a row of a table

**Observer bias:** The tendency for different people to observe things differently (also called experimenter bias)

**Open data:** Data that is available to the public

**Open-source:** Code that is freely available and may be modified and shared by the people who use it

**Openness:** The aspect of data ethics that promotes the free access, usage, and sharing of data

**Operator:** A symbol that names the operation or calculation to be performed

**ORDER BY:** A SQL clause that sorts results returned in a query

**Order of operations:** Using parentheses to group together spreadsheet values in order to clarify the order in which operations should be performed

**Ordinal data:** Qualitative data with a set order or scale

**Outdated data:** Any data that has been superseded by newer and more accurate information

**OUTER JOIN:** A SQL function that combines RIGHT and LEFT JOIN to return all matching records in both tables

**Outer query:** A SQL statement containing a subquery

**Ownership:** The aspect of data ethics that presumes individuals own the raw data they provide and have primary control over its usage, processing, and sharing

## P

**Package (R):** A unit of reproducible R code

**Packed bubble chart:** A data visualization that displays data in clustered circles

**Pattern:** The design principle of using similar visual elements to demonstrate trends and relationships in a data visualization

**PHP (Hypertext Preprocessor):** A programming language for web application development

**Pie chart:** A data visualization that uses segments of a circle to represent the proportions of each data category compared to the whole

**Pipe (R):** A tool in R for expressing a sequence of multiple operations, represented with “%>%”

**Pivot chart:** A chart created from the fields in a pivot table

**Pivot table:** A data summarization tool used to sort, reorganize, group, count, total, or average data

**Pixel:** In digital imaging, a small area of illumination on a display screen that, when combined with other adjacent areas, forms a digital image

**Population:** In data analytics, all possible data values in a dataset

**Portfolio:** A collection of materials that can be shared with potential employers

**Pre-attentive attributes:** The elements of a data visualization that an audience recognizes automatically without conscious effort

**Primary key:** An identifier in a database that references a column in which each value is unique (Refer to foreign key)

**Problem domain:** The area of analysis that encompasses every activity affecting or affected by a problem

**Problem types:** The various problems that data analysts encounter, including categorizing things, discovering connections, finding patterns, identifying themes, making predictions, and spotting something unusual

**Profit margin:** A percentage that indicates how many cents of profit has been generated for each dollar of sale

**Programming language:** A system of words and symbols used to write instructions that computers follow

**Proportion:** The design principle of using the relative size and arrangement of visual elements to demonstrate information in a data visualization

**Python:** A general-purpose programming language

## Q

**Qualitative data:** A subjective and explanatory measure of a quality or characteristic

**Quantitative data:** A specific and objective measure, such as a number, quantity, or range

**Query:** A request for data or information from a database

**Query language:** A computer programming language used to communicate with a database

## R

**R:** A programming language used for statistical analysis, visualization, and other data analysis

**Random sampling:** A way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen

**Range:** A collection of two or more cells in a spreadsheet

**Ranking:** A system to position values of a dataset within a scale of achievement or status

**readr (R):** An R package in Tidyverse used for importing data



**Record:** A collection of related data in a data table, usually synonymous with row

**Redundancy:** When the same piece of data is stored in two or more places

**Reframing:** The process of restating a problem or challenge, then redirecting it toward a potential resolution

**Regular expression (RegEx):** A rule that says the values in a table must match a prescribed pattern

**Relational database:** A database that contains a series of tables that can be connected to form relationships

**Relational operator:** An operator used to compare values (also called a comparator)

**Relativity:** The process of considering observations in relation or proportion to something else

**Relevant question:** A question that has significance to the problem to be solved

**Remove duplicates:** A spreadsheet tool that automatically searches for and eliminates duplicate entries from a spreadsheet

**Repetition:** The design principle of repeating visual elements to demonstrate meaning in a data visualization

**Report:** A static collection of data periodically given to stakeholders

**Return on investment (ROI):** A formula that uses the metrics of investment and profit to evaluate the success of an investment

**Revenue:** The total amount of income generated by the sale of goods or services

**Rhythm:** The design principle of creating movement and flow in a data visualization to engage an audience

**RIGHT:** A function that returns a set number of characters from the right side of a text string

**RIGHT JOIN:** A SQL function that will return all records from the right table and only the matching records from the left

**Root cause:** The reason why a problem occurs

**ROUND:** A SQL function that returns a number rounded to a certain number of decimal places.

**Ruby:** An object-oriented programming language for web application development

S

**Sample:** In data analytics, a segment of a population that is representative of the entire population

**Sampling bias:** Overrepresenting or underrepresenting certain members of a population as a result of working with a sample that is not representative of the population as a whole

**Scatterplot:** A data visualization that represents relationships between different variables with individual data points without a connecting line

**Schema:** A way of describing how something, such as data, is organized

**Scope of work (SOW):** An agreed-upon outline of the tasks to be performed during a project

**Second-party data:** Data collected by a group directly from its audience and then sold

**SELECT:** The section of a query that indicates from which column(s) to extract the data

**SELECT INTO:** A SQL clause that copies data from one table into a temporary table without adding the new table to the database

**Small data:** Small, specific data points typically involving a short period of time, which are useful for making day-to-day decisions

**SMART methodology:** A tool for determining a question's effectiveness based on whether it is specific, measurable, action-oriented, relevant, and time-bound

**Smoothing (R):** A process used to make data visualizations in R clearer and more readable

**Smoothing line (R):** A line on a data visualization that uses smoothing to represent a trend

**Social media:** Websites and applications through which users create and share content or participate in social networking

**Soft skills:** Nontechnical traits and behaviors that relate to how people work

**Sort range:** A spreadsheet menu function that sorts a specified range and preserves the cells outside the range

**Sort sheet:** A spreadsheet menu function that sorts all data by the ranking of a specific sorted column and keeps data together across rows

**Sorting:** The process of arranging data into a meaningful order to make it easier to understand, analyze, and visualize

**Specific question:** A question that is simple, significant, and focused on a single topic or a few closely related ideas

**SPLIT:** A spreadsheet function that divides text around a specified character and puts each fragment into a new, separate cell

**Sponsor:** A professional advocate who is committed to moving forward the career of another

**Spotlighting:** Scanning through data to quickly identify the most important insights

**Spreadsheet:** A digital worksheet

**SQL:** (Refer to Structured Query Language)

**Stakeholders:** People who invest time and resources into a project and are interested in its outcome

**Static data:** Data that doesn't change once it has been recorded

**Static visualization:** A data visualization that does not change over time unless it is edited

**Statistical power:** The probability that a test of significance will recognize an effect that is present

**Statistical significance:** The probability that sample results are not due to random chance

**Statistics:** The study of how to collect, analyze, summarize, and present data

**Story:** The narrative of a data presentation that makes it meaningful and interesting

**String data type:** A sequence of characters and punctuation that contains textual information (also called text data type)

**Structural metadata:** Metadata that indicates how a piece of data is organized and whether it is part of one or more than one data collection

**Structured data:** Data organized in a certain format such as rows and columns

**Structured Query Language:** A computer programming language used to communicate with a database

**Structured thinking:** The process of recognizing the current problem or situation, organizing available information, revealing gaps and opportunities, and identifying options

**Subquery:** A SQL query that is nested inside a larger query

**SUBSTR:** A SQL function that extracts a substring from a string variable

**Substring:** A subset of a text string

**Subtitle:** Text that supports a headline by adding context and description

**SUM:** A spreadsheet function that adds the values of a selected range of cells

**SUMIF:** A spreadsheet function that adds numeric data based on one condition

**Summary table:** A table used to summarize statistical information about data

**SUMPRODUCT:** A function that multiplies arrays and returns the sum of those products

**Swift:** A programming language for macOS, iOS, watchOS, and tvOS

**Symbol map:** A data visualization that displays a mark over a given longitude and latitude

**Syntax:** The predetermined structure of a language that includes all required words, symbols, and punctuation, as well as their proper placement

## T

**Tableau:** A business intelligence and analytics platform that helps people visualize, understand, and make decisions with data

**Technical mindset:** The ability to break things down into smaller steps or pieces and work with them in an orderly and logical way

**Temporary table:** A database table that is created and exists temporarily on a database server

**Text data type:** A sequence of characters and punctuation that contains textual information (also called string data type)

**Text string:** A group of characters within a cell, most often composed of letters

**Third-party data:** Data provided from outside sources who didn't collect it directly

**Tibble (R):** A streamlined variation of data frames

**Tidy data (R):** A way of standardizing the organization of data within R

**tidyr (R):** An R package in Tidyverse used for data cleaning to make tidy data

**Tidyverse (R):** A system of packages in R with a common design philosophy for data manipulation, exploration, and visualization

**Time-bound question:** A question that specifies a timeframe to be studied

**Transaction transparency:** The aspect of data ethics that presumes all data-processing activities and algorithms should be explainable and understood by the individual who provides the data

**Transferable skills:** Skills and qualities that can transfer from one job or industry to another

**TRIM:** A function that removes leading, trailing, and repeated spaces in data

**TSV (Tab-separated values file):** A text file that stores a data table by separating columns of data with tabs

**Turnover rate:** The rate at which employees voluntarily leave a company

**Typecasting:** Converting data from one type to another

## U

**Unbiased sampling:** When the sample of the population being measured is representative of the population as a whole

**Underscores:** Lines used to underline words and connect text characters

**Unfair question:** A question that makes assumptions or is difficult to answer honestly

**Unique:** A value that can't have a duplicate

**United States Census Bureau:** An agency in the U.S. Department of Commerce that serves as the nation's leading provider of quality data about its people and economy

**Unity:** The design principle of using visual elements that complement each other to create aesthetic appeal and clarity in a data visualization

**Unstructured data:** Data that is not organized in any easily identifiable manner

## V

**Validity:** The degree to which data conforms to constraints when it is input, collected, or created

**VALUE:** A spreadsheet function that converts a text string that represents a number to a numeric value

**Variable (R):** A representation of a value in R that can be stored for later use

**Variety:** The design principle of using different kinds of visual elements in a data visualization to engage an audience

**Vector (R):** A group of data elements of the same type stored in a one-dimensional sequence in R

**Verification:** A process to confirm that a data-cleaning effort was well executed and the resulting data is accurate and reliable

**Video file:** A collection of images, audio files, and other data usually encoded in a compressed format such as MP4, MV4, MOV, AVI, or FLV

**Vignette (R):** Documentation for an R package that describes the problem the package is designed to solve, explains how its functions can be used, and lists any dependencies on other packages

**Visual form:** The appearance of a data visualization that gives it structure and aesthetic appeal

**Visualization:** (Refer to Data visualization)

**VLOOKUP:** A spreadsheet function that vertically searches for a certain value in a column to return a corresponding piece of information

## W

**WHERE:** The section of a query that specifies criteria that the requested data must meet

**Wide data:** A dataset in which every data subject has a single row with multiple columns to hold the values of various attributes of the subject

**WITH:** A SQL clause that creates a temporary table that can be queried multiple times

**World Health Organization:** An organization whose primary role is to direct and coordinate international health within the United Nations system

## X

**X-axis:** The horizontal line of a graph usually placed at the bottom, which is often used to represent time scales and discrete categories

## Y

**Y-axis:** The vertical line of a graph usually placed to the left, which is often used to represent frequencies and other numerical variables

## Z