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

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

Accuracy: The degree to which the 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

Agenda: A list of scheduled appointments

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

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

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

B

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

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

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

C

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

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

Compatibility: How well two or more datasets are able to work together

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

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

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

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

COUNT: A spreadsheet function that counts the number of cells in a range that meet a specified criteria

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

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

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

Data: A collection of facts

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

Data analysis process: Carrying out the six phases of ask, prepare, process, analyze, share, and act in order 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 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 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 strategy: The management of the people, processes, and tools used in data analysis

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 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

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

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

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

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

E

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

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

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

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

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

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

G

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

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

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

H

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

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

I

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

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

K

L

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

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

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

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

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

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

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

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 function that returns the smallest numeric value from a range of cells

N

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

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

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 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

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

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

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

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

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

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

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

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

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

Root cause: The reason why a problem occurs

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

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

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

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

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 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

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

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

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

Substring: A smaller subset of a text string

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

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

T

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

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

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

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

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

U

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

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

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

V

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

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

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

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

X

Y

Z