Glossary

Data Analytics

Terms and Definitions

A

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) (Refer to Good data source)

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

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

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

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

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: 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 design: How information is organized

Data-driven decision-making: The process of 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 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: The process of exploring different data sources to find out what they have in common

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

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

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

Data visualization: The graphical representation of 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

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

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

E

Equation: A calculation that involves addition, subtraction, multiplication, or division (Refer to Math expression)

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

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 where the selected data comes from

Function: A preset command that automatically performs a 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) (Refer to Bad data source)

H

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

I

Inbox: Electronic storage where emails received by an individual are held

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

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

M

Math expression: A calculation that involves addition, subtraction, multiplication, or division (Refer to 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

Metadata: Data about data; in database management, it helps data analysts interpret the contents of the data within a database

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

MIN: A spreadsheet 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

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

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 (Refer to 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

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

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

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

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

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 the subset of a dataset

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

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

String data type: A sequence of characters and punctuation that contains textual information (Refer to 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

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

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 (Refer to String data type)

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

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

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

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)

W

WHERE: The section of a query that specifies criteria that the extracted 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