# Terms and definitions from Course 4

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

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

B

C

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

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

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

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

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

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

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

**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

**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

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

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

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

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

D

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

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

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

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

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

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

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

**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

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

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

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

E

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

F

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

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

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

G

H

**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

J

K

L

**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

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

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

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

N

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

O

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

P

Q

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

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

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

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

S

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

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

**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

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

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

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

T

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

**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

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

U

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

V

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

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

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

W

X

Y

Z