datacamp

COURSE GLOSSARY

Introduction to Data Literacy

API (Application Programming Interface): A set of rules and endpoints that allows software systems to request and exchange specific data or functionality programmatically

Data bias: Systematic distortions in data or its collection process that lead to unrepresentative samples or misleading conclusions

Data cleaning: The process of detecting and correcting or removing errors, inconsistencies, and inaccuracies in a dataset to improve its quality for analysis

Data lake: A storage system that holds raw, unprocessed data in its native formats for future processing or analysis

Data literacy: The ability to read, work with, analyze, and communicate insights from data so you can make informed decisions and evaluate data-driven claims

Data source: Any origin of data such as public datasets, APIs, internal systems, sensors, surveys, or third-party providers that supply the raw inputs for analysis

Data warehouse: A centralized repository that stores processed and organized data optimized for querying and analysis rather than raw storage

Function: A named block of reusable code that performs a specific task and can be called with parentheses, such as built-in functions like type() and print()

k-Nearest Neighbors (KNN): An instance-based algorithm that predicts the label or value of a query point by aggregating the labels/values of its k closest training observations according to a distance metric

Descriptive analytics: Analytics that summarize past or current data using statistics and visualizations to answer the question “what is happening?”

Diagnostic analytics: Analytics aimed at identifying possible causes and explanations for observed events or patterns, often using drill-downs, correlations, and hypothesis tests

ETL (Extract, Transform, Load): A common data pipeline framework that extracts data from sources, transforms it into a usable format, and loads it into a target system like a warehouse

Exploratory Data Analysis (EDA): An open-ended approach to examining datasets with summary statistics and visualizations to discover patterns, anomalies, and hypotheses for further analysis

Machine learning: A set of computational techniques where algorithms learn patterns from data to make predictions or decisions on new, unseen inputs without being explicitly programmed for each task

Missing data: Instances where expected values are absent from a dataset, which can bias results if not handled by techniques like deletion or imputation

Pipeline: An automated sequence of steps that moves and processes data from one system or storage location to another to ensure data is available and up-to-date

Predictive analytics: Techniques that use historical data and models to estimate the likelihood of future outcomes or unknown present conditions, typically with associated uncertainty

Prescriptive analytics: Methods that recommend actions or decisions based on predicted outcomes and trade-offs, often using rule-based systems, simulations, or optimization

Qualitative data: Categorical or descriptive data that characterizes attributes or qualities and is often analyzed by grouping or interpretation rather than numeric calculations

Quantitative data: Numeric data that can be measured or counted and subjected to mathematical or statistical analysis

Relational database: A type of structured database that organizes data into tables with rows and columns and uses keys to define relationships between tables

Storytelling (with data): The practice of combining data, narrative structure, and visualizations to present insights in a memorable and persuasive way that motivates an audience to act

Structured data: Data organized in a predefined format such as rows and columns, making it easy to query and analyze with traditional tools

Unstructured data: Data that lacks a predefined schema or tabular format, such as text, images, audio, or video, which typically requires more preprocessing to analyze