



A T M E
College of Engineering



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING - DATA SCIENCE

BUSINESS ANALYTICS

Module 2

*Descriptive Analytics I: Nature of Data, Big Data,
and Statistical Modeling*

Module 2: Descriptive Analytics I: Nature of Data, Big Data, and Statistical Modeling

1. The Nature of Data in Analytics

Data is defined as a collection of raw facts and figures that can be processed into meaningful insights. It is usually obtained through experiences, observations, or experiments.

1.1 What is Data?

- Data may consist of numbers, words, images, sounds, or any other raw input.
- Data is the lowest level of abstraction — information and knowledge are derived from data.
- Data by itself does not carry meaning; processing transforms it into actionable information.

💡 Note: The progression is: Data → Information → Knowledge → Wisdom (DIKW hierarchy).

1.2 Types of Data

- Structured Data: Organized in rows and columns (e.g., sales transactions in a database).
- Semi-structured Data: Tagged or marked format — partially organized (e.g., XML, JSON).
- Unstructured Data: Free-form data without a predefined format (e.g., social media posts, videos, emails).

1.3 The 4 V's of Data

The four fundamental characteristics that define and differentiate modern data:

V	Description	Example
Volume	The massive size/quantity of data generated	Petabytes of social media data per day
Variety	Different formats and types of data	Text, video, audio, structured files
Velocity	The speed at which data is created and processed	Real-time sensor and IoT data
Veracity	The quality, accuracy, and trustworthiness of data	Fake or misleading user reviews

💡 Note: Some providers add more V's: Variability (SAS) and Value (IBM). The most cited are Volume, Variety, and Velocity.

1.4 Data Sources

- Business Processes: Systems like ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), and SCM (Supply Chain Management) generate structured operational data.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING - DATA SCIENCE

- Internet and Social Media: Platforms such as Facebook, Twitter, YouTube, and Instagram produce vast volumes of unstructured data in the form of posts, comments, likes, and shares.
- Machines / Internet of Things (IoT): Physical devices, sensors, and machines continuously emit operational and usage data — e.g., temperature sensors, smart meters, GPS trackers.

1.5 Data Storage and Protection

- All collected data is stored in centralized or cloud-based storage systems.
- Data protection ensures privacy, security, and compliance with applicable regulations (e.g., GDPR, HIPAA).

1.6 Analytics Processing Layer

Once data is collected and stored, it goes through a three-phase analytics process:

- Build: Developing analytical models and systems.
- Test: Ensuring accuracy and reliability of models.
- Validate: Checking and refining results for correctness.

Cloud storage and computing provide the scalability and computational power needed to handle large data volumes. Processed insights are delivered through applications (dashboards, reports, mobile apps) to end users who use them for decision-making, trend analysis, and knowledge discovery.

2. Data Quality Metrics

The following characteristics define the readiness level of data for an analytics study (Delen, 2020; Kock, McQueen, & Corner, 1997). High-quality data is essential for accurate and trustworthy analytical outcomes.

Metric	Description
Data Source Reliability	Trustworthiness of the original source; reduces risks of misrepresentation during transfers.
Data Accuracy	Correctness and alignment with intended meaning (e.g., recorded details match what was provided).
Data Accessibility	Ease of retrieval, especially with diverse storage systems (databases, data lakes, Hadoop, etc.).
Data Security & Privacy	Controlled access, legal compliance (e.g., HIPAA), and protection against unauthorized use.
Data Richness	Inclusion of sufficient variables and dimensions for meaningful analytics.
Data Consistency	Proper integration of data from multiple sources to avoid errors (e.g., mixing patient records).
Data Currency / Timeliness	Data should be up-to-date and recorded close to the event to ensure reliability.
Data Granularity	Level of detail must be sufficient for analysis (e.g., medical results with decimals, detailed demographics).
Data Validity	Values must fall within defined acceptable ranges (e.g., gender = male/female/other).
Data Relevancy	Only variables pertinent to the study should be included to avoid misleading results.

💡 Note: Collectively, these metrics are also referred to as 'data fitness criteria' — data that scores well on all these dimensions is said to be analytics-ready.

3. Simple Taxonomy of Data Types

Data can be classified into several types based on its nature and the kind of analysis it supports. Understanding data types is critical for choosing appropriate analytical techniques.

3.1 Categorical (Discrete) Data

Labels used to divide variables into groups (e.g., race, sex, age group). Values are finite and symbolic, not fractional.

Nominal Data

- Type of categorical data where codes are labels only, without any meaningful order.
- Example: Marital status (single, married, divorced); eye colour (blue, green, brown).
- Can be binomial (yes/no) or multinomial (multiple categories).

Note: Nominal data can be counted but not ordered or ranked.

Ordinal Data

- Categorical data that has a meaningful rank order.
- Example: Credit score (low, medium, high); customer satisfaction levels (poor, average, excellent).
- Order matters, but the exact numerical difference between categories is not precisely measurable.

3.2 Numerical (Continuous) Data

Variables with measurable numeric values (e.g., age, income, distance). Can be integer (whole numbers) or real (fractions/decimals). Also called continuous data.

Interval Data

- A numeric scale with meaningful, equal intervals but no true zero.
- Example: Temperature in Celsius or Fahrenheit — 0°C does not mean 'no temperature'.
- Differences between values are meaningful; ratios are not.

Ratio Data

- A numeric scale with a true, non-arbitrary zero point.
- Example: Length, mass, time, Kelvin temperature.
- Ratios are meaningful — 20 kg is twice as heavy as 10 kg.
- All mathematical operations (addition, subtraction, multiplication, division) are valid.

3.3 Other Data Types

- Text, spatial, image, video, and audio data must be converted into categorical or numeric form for analytics.
- Static Data: Does not change over time (e.g., historical census records).
- Dynamic (Temporal / Time-Series) Data: Changes over time (e.g., daily stock prices, monthly sales figures).

3.4 Data Type Conversion for Machine Learning

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING - DATA SCIENCE

- Many ML algorithms (e.g., neural networks, SVMs, logistic regression) require numeric input — categorical variables must be encoded.
- One-hot encoding is a common technique for converting nominal categories into numeric form.
- Care is needed: converting a categorical variable with many unique values creates too many additional columns.
- Some algorithms (e.g., ID3 decision trees, rough sets) natively require categorical inputs.
- Modern analytics tools increasingly automate data transformation.

3.5 Summary Table: Data Types

Data Type	Definition	Examples	Order?	True Zero?	Key Notes
Categorical (Discrete)	Groups/labels dividing variables into classes	Race, sex, age group	No	No	Finite, symbolic; not fractional
Nominal	Labels without order	Marital status, colours, yes/no	No	No	Used for naming only
Ordinal	Categories with meaningful rank order	Credit score, education level	Yes	No	Differences not measurable
Numeric (Continuous)	Measurable numbers (integer or real)	Age, income, distance	Yes	Depends	Allows fractional values
Interval	Equal intervals, no true zero	Celsius/Fahrenheit temperature	Yes	No	Differences meaningful
Ratio	Numeric scale with meaningful zero	Length, mass, time, Kelvin	Yes	Yes	Full mathematical use

4. The Art and Science of Data Preprocessing

Raw data is often incomplete, inconsistent, or noisy. Data preprocessing refers to the series of steps taken to clean, transform, and reduce raw data into a form suitable for analysis and machine learning.

💡 Note: Studies suggest that data scientists spend approximately 60–80% of their time on data preprocessing — it is one of the most critical and time-consuming phases in any analytics project.

Step 1: Data Consolidation (Blending)

- Relevant data is collected from identified sources and merged together.
- The integration or merging process is called data blending.
- Key challenges: schema matching (aligning different column names) and entity resolution (matching the same real-world entity across sources).

Step 2: Data Cleaning (Cleansing)

The process of identifying and correcting (or removing) errors and inconsistencies in data to improve its quality.

Key Tasks in Data Cleaning

- Handling Missing Values: Replace with mean/median/mode, use interpolation, or remove affected rows/columns.
- Removing Duplicates: Identify and delete repeated records that can skew analysis.
- Correcting Errors: Fix typos, invalid entries, inconsistent formats (e.g., date formats, currency symbols).

💡 Note: Missing data can be handled using imputation (filling in estimated values) or deletion. The choice depends on the extent and pattern of missingness.

Step 3: Data Transformation

Converting data into a suitable format or structure for analysis.

Common Transformation Techniques

- Normalization / Scaling: Rescaling values to a common range (e.g., 0–1) to prevent features with large ranges from dominating the model.
- Aggregation: Summarizing data (e.g., daily sales summed to monthly totals).
- Binning / Discretization: Converting continuous values into discrete bins (e.g., age 0–18, 19–35, 36–60, 60+).
- Date/Time Features: Extracting day, month, year, weekday, etc., from timestamp data.
- Encoding: Converting categorical variables to numeric form (label encoding, one-hot encoding).

Step 4: Data Reduction

Reducing the size of the dataset while preserving as much useful information as possible. This improves processing time and helps avoid overfitting.

Common Data Reduction Techniques

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING - DATA SCIENCE

- Sampling: Using a representative subset of the data instead of the entire dataset.
- Dimensionality Reduction: Techniques like PCA (Principal Component Analysis) reduce the number of features while retaining variance.
- Numerosity Reduction: Replacing the original data with a smaller representation (e.g., histograms, clusters, parametric models).

Summary Table: Preprocessing Steps

Step	Purpose	Key Techniques
Data Consolidation	Combine data from multiple sources	Schema matching, entity resolution
Data Cleaning	Improve quality by fixing errors and omissions	Imputation, deduplication, correction
Data Transformation	Convert data into suitable formats for analysis	Scaling, encoding, aggregation
Data Reduction	Reduce size while preserving insights	Sampling, PCA, feature selection

5. Definition and Fundamentals of Big Data

Big Data refers to datasets that are so large, fast-moving, or complex that traditional data processing tools are insufficient to handle them. The term has evolved and now encompasses multiple dimensions beyond just volume.

5.1 Definition

Traditionally, Big Data described the massive volumes of data analyzed by huge organizations like Google or research science projects at NASA. Today, it is commonly defined by three or more 'V's.

5.2 The V's of Big Data

- **Volume:** The massive amount of data generated every second. Example: Facebook generates over 4 petabytes of data per day.
- **Velocity:** The speed at which data is created, collected, and processed. Example: Stock market data or IoT sensor data is generated in real time.
- **Variety:** Different types of data (structured, semi-structured, unstructured). Example: Tweets (text), YouTube videos, and Excel files.
- **Veracity:** The quality, accuracy, and trustworthiness of data. Example: User reviews on e-commerce sites may contain fake or misleading information.
- **Value:** The usefulness of data in generating insights and decisions. Example: Netflix uses viewing data to recommend shows.
- **Variability:** Inconsistency in the meaning or format of data. Example: The word 'Apple' could mean the fruit or the tech company, depending on context.

💡 Note: IBM emphasizes Veracity; SAS emphasizes Variability. The core three V's recognized universally are Volume, Variety, and Velocity.

5.3 Why Big Data Matters

Big Data by itself, regardless of size, type, or speed, is worthless unless organizations extract value from it. Success in Big Data analytics depends on several key factors:

1. A Clear Business Need

- Initiatives should start with a specific, well-defined problem — not random data collection.
- Example: An e-commerce company wants to reduce shopping cart abandonment rates.

2. Strong, Committed Sponsorship

- Top management must actively support and fund analytics projects.
- Example: A hospital's leadership invests in predictive analytics to reduce patient readmissions.

3. Alignment Between Business and IT Strategy

- Business goals and IT capabilities must be aligned for projects to succeed.
- Example: A bank wants to improve fraud detection — the business team defines the goal; IT implements real-time monitoring.

4. A Fact-Based Decision-Making Culture

- Organizations must trust data-driven insights over gut feelings and intuition.
- Example: Netflix uses viewing data to recommend shows instead of relying solely on human intuition.

5. A Strong Data Infrastructure

- Requires reliable data storage, processing pipelines, and management systems.
- Example: Amazon uses AWS cloud infrastructure to process huge volumes of customer and transaction data.

6. The Right Analytics Tools

- Choosing appropriate tools and technologies for analysis is critical.
- Example: Hadoop and Spark for Big Data processing; Tableau/Power BI for visualization.

7. Personnel with Advanced Analytical Skills

- Skilled data scientists, engineers, and analysts are crucial to success.
- Example: Google hires top AI researchers to develop models like Google Translate.

5.4 Business Problems Addressed by Big Data Analytics

The top business problems addressed by Big Data analytics include process efficiency and cost reduction. Other key areas include:

- Process efficiency and cost reduction
- Brand management
- Revenue maximization, cross-selling, and up-selling
- Enhanced customer experience
- Churn identification and customer recruiting
- Improved customer service
- Identifying new products and market opportunities
- Risk management
- Regulatory compliance
- Enhanced security capabilities

6. Big Data Technologies

Three Big Data technologies are considered transformative for the business analytics and data management markets: Hadoop, MapReduce, and NoSQL.

6.1 Hadoop

Hadoop is an open-source framework for processing, storing, and analyzing massive amounts of distributed, unstructured data.

- Originally created by Doug Cutting at Yahoo!, inspired by Google's MapReduce model (early 2000s).
- Designed to process large data volumes across many machines without relying on a single server.
- Built to be scalable, fault-tolerant, and cost-effective.
- Uses inexpensive commodity computers (nodes) connected in a cluster rather than expensive high-end hardware.

How Hadoop Works

Data flows from sources into the Hadoop ecosystem through these components:

- HDFS (Hadoop Distributed File System): Splits data into smaller parts and distributes them across multiple nodes. Replicates parts for fault tolerance — if one node fails, another has a copy.
- NameNode: Keeps track of where data resides and manages node communication.
- Secondary Node: A backup to the NameNode; periodically replicates and stores data from the NameNode.
- JobTracker: Initiates and coordinates MapReduce jobs (the processing of the data).
- Worker Nodes: Store data and execute processing tasks as directed by the JobTracker.

Hadoop Ecosystem Components

- Hive: A Hadoop-based data warehousing framework (originally by Facebook). Allows SQL-like queries via HiveQL, which are converted to MapReduce.
- Pig: A Hadoop-based query language developed by Yahoo! — adept at very deep, long data pipelines.
- HBase: A non-relational (NoSQL) database that enables low-latency lookups in Hadoop. Adds transactional capabilities (updates, inserts, deletes). Used heavily by eBay and Facebook.
- Flume: A framework for populating Hadoop with data — deploys agents throughout IT infrastructure (web servers, mobile apps) to collect and integrate data.

6.2 MapReduce

MapReduce is a programming model popularized by Google for distributing the processing of very large, multi-structured data files across a large cluster of machines.

- High performance is achieved by breaking processing into small parallel units of work.
- MapReduce is a programming model, not a programming language.

How MapReduce Works (Colored Square Counter Example)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING - DATA SCIENCE

- Splitting: The input file is divided into multiple splits. Each split is processed independently and in parallel.
- Map Phase: Each map program processes a split and outputs key-value pairs (e.g., color: count).
- Shuffle & Sort: Intermediate results are merged and grouped by key (same color grouped together). Combiners and custom functions optimize this step.
- Reduce Phase: The reduce program sums the counts for each key and produces the final output (color: total count).

💡 Note: MapReduce transparently handles parallel computing — developers do not need to write parallel processing code themselves.

Why Use MapReduce?

- Ideal for: indexing and search, graph analysis, text analysis, machine learning, and data transformation.
- These applications are difficult to implement using standard SQL in relational DBMSs.
- Automatically handles distributed parallel execution — reduces developer burden.

6.3 Hadoop vs. Spark

Aspect	Hadoop	Spark
Processing	Disk-based, batch processing	In-memory; real-time, streaming, iterative
Performance	Slower (I/O bound)	Faster (up to 100x for in-memory workloads)
Cost	Cheaper (uses disk storage)	Costlier (requires more RAM)
Scalability	Uses HDFS for scaling	Uses HDFS for scaling
Security	More secure overall	Basic auth; integrates with Hadoop security
ML Support	Limited	MLlib (machine learning, regression, classification)
Best For	Very large data, batch jobs, low budget, archive analysis	Real-time streaming, iterative algorithms, ML applications

6.4 NoSQL Databases

NoSQL (Not Only SQL) databases are designed to handle large-scale, unstructured, and semi-structured data that traditional relational databases (RDBMS) cannot manage efficiently.

Characteristics

- Do not require fixed schemas — flexible structure adapts to changing data.
- Scale easily across distributed servers (horizontal scaling).
- Prioritize high performance, scalability, and low latency over strict ACID compliance.



When to Use NoSQL

- Real-time applications: Gaming leaderboards, stock trading, fraud detection.
- Scalable systems: Social media feeds, IoT sensor data, recommendation systems.
- Flexible schema requirements: Rapidly changing datasets (e.g., product catalogs in e-commerce).

7. Big Data and Stream Analytics

A stream is defined as a continuous sequence of data elements. Stream analytics processes data in real time, as it is created, enabling instant insights and faster decision-making.

The importance of stream analytics lies in the fact that time-to-action is critical in many domains. Modern technology now allows real-time capture and analysis of continuous data streams at scale.

7.1 Applications of Stream Analytics

Domain	Use Case
Energy / Smart Grids	Real-time power distribution, predicting demand peaks, optimizing renewable energy usage.
E-Commerce	Amazon/eBay analyze clicks, searches, and views instantly to recommend products and maximize sales.
Telecommunications	Analyze call detail records (CDR) to prevent churn, identify influencers, and improve campaigns.
Law Enforcement / Cybersecurity	Real-time crime detection, surveillance, fraud detection, and network intrusion prevention.
Financial Services	Real-time trading decisions, fraud detection, and monitoring for illegal activities.
Health Sciences	Medical devices stream patient data (ECG, blood pressure) for anomaly detection and interventions.
Government	Disaster management, environmental monitoring, traffic management using live sensor and GPS data.

💡 Note: Stream analytics is fundamentally different from batch analytics — instead of storing data and processing it periodically, stream analytics processes events as they arrive, enabling immediate responses.

8. Statistical Modeling for Business Analytics

Because of the increasing popularity of business analytics, traditional statistical methods are regaining their attractiveness as enabling tools to support evidence-based managerial decision-making.

8.1 Statistics in Descriptive Analytics

- Statistics (statistical methods and techniques) is traditionally considered part of descriptive analytics.
- Descriptive analytics has two main branches:
 - Statistics: Characterizes data either one variable at a time or multivariables together, using descriptive or inferential methods.
 - OLAP (Online Analytical Processing): Analyzing, characterizing, and summarizing structured data stored in organizational databases (often a data warehouse or data mart). Also called business intelligence.

Note: Some statistical methods such as discriminant analysis can also be considered part of predictive analytics.

8.2 Descriptive Statistics

Descriptive statistics summarizes and presents data through tables, graphs, and charts. It helps identify patterns, outliers, and unusual values.

Measures of Central Tendency

- **Mean:** The sum of all data values divided by the total number of values. Represents the balance point of the dataset. Sensitive to outliers.
- **Median:** The middle value when data is arranged in order. Splits the dataset into two equal halves. Robust to outliers and skewed data. Suitable for ordinal, interval, and ratio data.
- **Mode:** The value that appears most frequently. Represents the most common observation. Useful for categorical (nominal) data. No standard formula — it is simply the most frequent value(s).

Measures of Dispersion (Spread)

These measures show how data values vary around the central point.

- Range: Maximum value minus minimum value. Simple but sensitive to outliers.
- Variance: The average of the squared deviations from the mean. Captures overall spread.
- Standard Deviation: Square root of variance. Most widely used measure of spread; in original data units.
- Mean Absolute Deviation (MAD): Average of the absolute differences from the mean.
- Quartiles and Interquartile Range (IQR): Focus on the middle 50% of data; robust to outliers.
- Box Plot: A visual tool (box-and-whisker diagram) for displaying centrality and spread simultaneously.

Shape of Distribution

- Normal Distribution: A symmetric, bell-shaped curve. Mean = Median = Mode.



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING - DATA SCIENCE

- Skewness: Measures the asymmetry of the distribution.
 - Positive skew: Long right tail ($\text{mean} > \text{median}$).
 - Negative skew: Long left tail ($\text{mean} < \text{median}$).
- Kurtosis: Measures the peakedness of the distribution. Value >3 indicates a peaked (leptokurtic) distribution; value <3 indicates a flat (platykurtic) distribution.

9. Regression Modeling for Inferential Statistics

Regression is a statistical method used to model the relationship between one or more independent (explanatory) variables and a dependent (response) variable. It helps in understanding, explaining, and predicting outcomes.

9.1 Purpose of Regression

- Hypothesis Testing: To find and test relationships between variables.
- Prediction / Forecasting: To estimate future values of the dependent variable.

9.2 Correlation vs. Regression

Concept	Definition	Key Difference
Correlation	Measures the degree of association/strength between variables	No cause-effect relationship implied
Regression	Determines dependency — how one variable affects another	Models cause-effect relationship

9.3 Types of Regression

- **Simple Linear Regression:** One independent variable and one dependent variable. Example: Predicting a person's weight from their height.
- **Multiple Linear Regression:** More than one independent variable predicts the dependent variable. Example: Predicting weight from height, BMI, and gender.
- **Nonlinear Regression:** Used when the relationship between variables follows a curve rather than a straight line.
- **Logistic Regression:** A classification technique used to predict categorical (binary) outcomes. Models probability using the logit function (log of odds).

9.4 Assumptions of Linear Regression

- Linearity: The relationship between variables is linear and additive.
- Independence: Errors (residuals) are independent of each other — no autocorrelation.
- Normality: Errors follow a normal distribution.
- Homoscedasticity (Constant Variance): Error variance remains the same across all values of the independent variable.
- No Multicollinearity: Independent variables are not highly correlated with each other.

Note: Violations of these assumptions can lead to biased or inefficient estimates. Always check residual plots to diagnose issues.

9.5 Logistic Regression

Logistic regression is a classification technique used to predict categorical outcomes (e.g., Yes/No, Pass/Fail, Spam/Not Spam) rather than continuous values.

- Models the probability of an event occurring using the logit function — the log of the odds ratio.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING - DATA SCIENCE

- Output is bounded between 0 and 1, representing probability.
- Types:
 - Binary Logistic Regression: Two outcome classes (e.g., accept/reject).
 - Multinomial Logistic Regression: More than two outcome classes (e.g., low/medium/high risk).
- Common applications: predictive analytics, medical research, social sciences, credit risk assessment.

9.6 Process Flow for Developing Regression Models

Step 1: Tabulated Data

- Starting point — the raw dataset organized in a table (rows as observations, columns as variables).
- Includes both dependent (target) and independent (predictor) variables.
- Example: A dataset with columns like Hours Studied (X) and Exam Score (Y).

Step 2: Data Assessment

- Explore and analyze data to understand relationships among variables.
- Scatter Plot: Visualize how one variable changes with another (linear or non-linear).
- Correlations: Measure the strength and direction of relationships between variables.
- Goal: Check whether a linear relationship exists and assess data quality.

Step 3: Model Fitting

- Create the linear regression equation and estimate parameters (coefficients).
- Transform Data: Apply transformations (e.g., log, square root) if the relationship is non-linear.
- Estimate Parameters: Use the least squares method to find the best-fit line that minimizes prediction errors.

Step 4: Model Assessment

- Evaluate the model to ensure it fits data well and satisfies regression assumptions.
- Test Assumptions: Check for linearity, independence, normality, and homoscedasticity of residuals.
- Assess Model Fit: Use performance metrics such as R^2 (coefficient of determination), RMSE (Root Mean Squared Error), and residual plots.
- If issues are found, loop back to Data Assessment or Model Fitting for refinement.

Step 5: Deployment

- Once the model performs well, it is deployed for practical use.
- One-time use: Apply to a single dataset for a specific analysis or prediction.
- Recurrent use: Implement in a production system for ongoing or automated predictions (e.g., sales forecasting, demand prediction).

9.7 Comparison: Linear vs. Logistic Regression

Aspect	Linear Regression	Logistic Regression
Output	Continuous numeric value (e.g., sales, price)	Categorical outcome / probability (0 to 1)
Use Case	Forecasting and prediction	Classification tasks
Equation	$Y = a + bX$	$\log(p/1-p) = a + bX$
Examples	House price prediction, exam score prediction	Spam detection, disease diagnosis, credit approval

💡 Note: Both regression types help organizations analyze trends, test hypotheses, and make data-driven decisions. The choice between them depends on the type of outcome variable.

10. Quick Revision: Key Terms Glossary

Term	Definition
Data	Raw facts and figures that are processed into meaningful insights.
Structured Data	Organized in rows and columns (e.g., SQL tables, Excel sheets).
Unstructured Data	Free-form data with no predefined format (e.g., social media posts, videos).
4 V's of Big Data	Volume, Variety, Velocity, Veracity — the defining characteristics of Big Data.
Data Preprocessing	The process of cleaning, transforming, and reducing raw data for analysis.
Imputation	Filling in missing values with estimated substitutes (mean, median, etc.).
Hadoop	Open-source framework for distributed storage and processing of large datasets.
HDFS	Hadoop Distributed File System — stores and replicates data across cluster nodes.
MapReduce	Google's programming model for parallel processing of large datasets across clusters.
NoSQL	Non-relational databases designed for scalability and flexible schemas.
Stream Analytics	Processing data continuously in real time as it is generated.
Descriptive Statistics	Summarizes and describes the main features of a dataset.
Inferential Statistics	Draws conclusions and makes predictions about a population from sample data.
OLAP	Online Analytical Processing — analyzes structured data in data warehouses.
Regression	Statistical method to model the relationship between dependent and independent variables.
R^2 (R-Squared)	Coefficient of determination — measures the proportion of variance explained by the model.
Logistic Regression	Classification algorithm predicting categorical outcomes using probability.
Homoscedasticity	Constant variance of residuals across all predictor values — an assumption of linear regression.
One-Hot Encoding	Technique to convert categorical variables into binary numeric columns for ML models.
PCA	Principal Component Analysis — dimensionality reduction technique preserving maximum variance.

Common Exam Questions

- What is Big Data? Explain its characteristics with examples.
- What are the keys to success in Big Data analytics? Explain with examples.
- Explain the MapReduce process flow with the colored square counter example.
- What are the steps in the data preprocessing pipeline?
- Compare descriptive and inferential statistics.
- Explain the process flow for developing a regression model.
- What is NoSQL? When should you use it?
- Differentiate between linear and logistic regression.