UNIT I — INTRODUCTION TO DATA ANALYTICS (DETAILED)

1.1 Data Analytics — Overview & Importance

Definition (formal)

Data analytics is the scientific process of inspecting, cleansing, transforming, modelling and interpreting data with the objective of discovering useful information, informing conclusions and supporting decision-making.

Core components (terminology)

- **Data ingestion / acquisition:** getting data from sources (sensors, databases, files, APIs).
- **Data wrangling / cleaning:** removing errors, imputing missing values, normalizing.
- **Exploratory Data Analysis (EDA):** summary statistics, visualizations to understand structure.
- **Feature engineering:** creating new variables (features) that make models work better.
- **Modeling / inference:** statistical or machine learning models to explain or predict.
- **Evaluation / validation:** accuracy, precision, recall, RMSE, R^2 etc.
- **Deployment / monitoring:** putting models into production and checking performance over time.

Importance (in data-analysis terms)

- **Evidence-based decisions:** replace heuristics with statistically supported choices.
- **Risk reduction:** quantify uncertainty (confidence intervals, p-values).
- **Optimization:** find optimal resource allocation (prescriptive analytics).

- **Automation:** automating repetitive decisions via predictive models.
- **Insights & discovery:** hypothesis generation using EDA and pattern detection.

Example (analyst view)

An operations analyst uses historical machine sensor logs, performs EDA to find patterns, builds a predictive maintenance model (time-to-failure regression), validates with cross-validation, and recommends preventative replacement schedules (prescriptive).

1.2 Types of Data Analytics (deep)

Analytics types are often presented as a chain — each adds value:

1. Descriptive Analytics

- **Question answered:** *What happened?*
- **Outputs:** aggregations (SUM, COUNT), group-by tables, time series plots, histograms.
- **Techniques/tools:** SQL, pandas, Excel, dashboards (Power BI/Tableau).
- **When to use:** reporting, KPI dashboards, baseline measurement.

Example outputs: monthly revenue, average downtime, counts by category.

---

2. Diagnostic Analytics

- **Question answered:** *Why did it happen?*
- **Outputs:** drill-downs, correlations, causal inference (where possible), root cause analyses.
- **Techniques:** correlation matrices, pivot tables, ANOVA, causal graphs, decision trees for rule extraction.
- **Key caution:** correlation ≠ causation — apply domain reasoning or experiments to infer causality.

Example: Use correlation and controlled comparisons to find that a price increase coincided with reduced sales in a region.

---

3. Predictive Analytics

- **Question answered:** *What is likely to happen?*
- **Outputs:** point forecasts, probability estimates, risk scores.
- **Techniques:** regression (linear, logistic), time-series models (ARIMA, exponential smoothing), tree models, ensemble (random forest, gradient boosting), neural networks.
- **Validation:** cross-validation, holdout sets, ROC/AUC for classification, RMSE/MAPE for regression.

Example: Predict monthly demand for a product, or probability a student will drop out.

---

4. Prescriptive Analytics

- **Question answered:** *What should we do?*
- **Outputs:** recommended actions, optimization plans, decision rules.
- **Techniques:** operations research (linear programming), simulation, multi-armed bandits, reinforcement learning, scenario analysis.
- **Metrics:** objective function value (cost, profit), constraints satisfaction.

Example: Given forecasted demand, compute optimal inventory levels that minimize holding + shortage costs.

---

5. Visual Analytics

- **Role:** make complex patterns understandable via visuals and interactive dashboards.
- **Visual types:** line charts, bar charts, histograms, boxplots, scatter plots, heatmaps, choropleth maps, network graphs.

- **Good practice:** match chart to task (compare → bar; trend → line; distribution → histogram/boxplot; relationship → scatter).
- **Tools:** Tableau, Power BI, matplotlib, ggplot, D3.js.

Dashboard design principles: clear title, one key message per chart, consistent axes, appropriate aggregation level, interactivity to filter and drill.

1.3 Life Cycle of Data Analytics; Data Quality & Quantity; Measurement

Life Cycle (end-to-end steps with outputs and QC)

1. **Problem formulation / requirement gathering**
    - Output: analytic question, KPIs, success criteria.
2. **Data collection / ingestion**
    - Sources: transactional DBs, CSV, logs, APIs, sensors. Output: raw datasets.
3. **Data cleaning & preprocessing**
    - Tasks: missing value handling, deduplication, type conversion, outlier detection. Output: clean dataset + data profiling report.
4. **Exploratory Data Analysis (EDA)**
    - Tasks: summary stats, visualizations, feature distributions, correlation checks. Output: EDA notebook, hypotheses.
5. **Feature engineering & selection**
    - Tasks: transforms (log, scaling), encoding categorical variables (one-hot, target encoding), dimensionality reduction (PCA). Output: modeling dataset.
6. **Model building**
    - Tasks: choose algorithm, train, tune hyperparameters. Output: trained model(s).
7. **Evaluation & validation**

- Tasks: metrics calculation, error analysis, cross-validation, checking bias/variance. Output: performance report.

8. **Deployment & monitoring**
    - Tasks: put model into production, monitor drift, retrain schedule. Output: production model + monitoring dashboards.
9. **Business action & feedback loop**
    - Tasks: decision implementation, A/B testing, measure impact. Output: business KPIs change; iterate.

Data Quality (dimensions & how to measure)

- **Accuracy:** values are correct. (Check by sampling against trusted source.)
- **Completeness:** no missing values for required fields. (Measure % missing.)
- **Consistency:** no conflicting values across sources. (Use cross-field rules.)
- **Uniqueness:** no duplicate records. (Dedup %.)
- **Timeliness:** data is up-to-date for the use case.
- **Validity:** conforms to schema/rules (range checks, regex).

Quality metrics examples: % missing, % invalid, number of duplicates, time lag in hours.

Data Quantity

- More data can reduce estimator variance and improve model generalization, but requires more compute and careful quality checks.
- **Curse of dimensionality:** adding many features without enough samples can harm performance — apply feature selection or regularization.

Measurement (scales & allowed statistics)

Four measurement scales determine permissible analyses:

1. **Nominal** — categories, no order (e.g., color). Allowed: mode, counts, chi-square tests.

2. **Ordinal** — ordered categories (e.g., rating A/B/C). Allowed: median, percentiles, nonparametric tests.
3. **Interval** — numeric with equal intervals, no true zero (e.g., temperature °C). Allowed: mean, SD, correlation.
4. **Ratio** — numeric with true zero (e.g., weight). Allowed: all arithmetic including ratios.

1.4 Data Types; Measures of Central Tendency; Measures of Dispersion

Data Types (structured → unstructured)

- **Structured:** tabular data — rows/columns (SQL tables, CSV).
- **Semi-structured:** XML/JSON — keys & nested structures.
- **Unstructured:** text, images, audio, video. Requires specialized processing (NLP, computer vision).

Also: **categorical** vs **numerical** (discrete/continuous) — choose stats accordingly.

Measures of Central Tendency (detailed)

Goal: summarize the centre of distribution.

Mean (arithmetic mean)

[
\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i
]

Worked numeric example (step-by-step): sample = [12, 15, 14, 10, 19], n = 5.

1. Sum step-by-step: 12 + 15 = 27. 27 + 14 = 41. 41 + 10 = 51. 51 + 19 = 70.
2. Mean = 70 ÷ 5 = 14.
   So (\bar{x} = 14).

Median

- Sort the data and pick middle element.
  Example sorted: [10, 12, 14, 15, 19] → median = 14.

Mode

- Most frequent value. Example: if data = [2,4,4,6,4,8] → mode = 4.
  If no repeat, there is no mode or distribution is multimodal.

When to use which: mean is sensitive to outliers; median is robust to outliers.

Measures of Dispersion (detailed)

Goal: describe spread around center.

Range

[
\text{Range} = \max(x) - \min(x)
]
Example: max 19 − min 10 = 9.

Variance & Standard Deviation

- **Population variance:** (\sigma^2 = \frac{1}{N}\sum (x_i - \mu)^2).
- **Sample variance (unbiased):** (s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2). Use (n-1) to correct bias.

Using example sample [12,15,14,10,19], (\bar{x}=14):
Compute deviations and squared:

- $12 - 14 = -2 \rightarrow (-2)^2 = 4$
- $15 - 14 = 1 \rightarrow 1^2 = 1$
- $14 - 14 = 0 \rightarrow 0^2 = 0$
- $10 - 14 = -4 \rightarrow 16$
- $19 - 14 = 5 \rightarrow 25$

Sum of squared deviations = 4 + 1 + 0 + 16 + 25 = 46.

Sample variance (s^2 = 46/(n-1) = 46/4 = 11.5).

Standard deviation (s = \sqrt{11.5}).
Compute approximate: (\sqrt{11.5} \approx 3.391) (approx).

Interpretation: on average, data points are about 3.39 units away from the mean.

Interquartile Range (IQR)

IQR = Q3 – Q1; robust spread measure for skewed data.

Shape measures

- **Skewness:** direction and degree of asymmetry.
- **Kurtosis:** "tailedness" of distribution (higher kurtosis → heavier tails).

1.5 Sampling Funnel; Central Limit Theorem; Confidence Interval; Sampling Variation

Sampling Funnel — practical steps & methods

Why sample? Population may be too large/costly; sampling estimates population parameters.

Sampling steps:

1. Define population and target parameter (mean, proportion).
2. Choose sampling method.
3. Determine sample size (power / margin).
4. Collect sample and apply weighting if needed (survey weighting).
5. Analyze and adjust for non-response or sampling bias.

Common sampling methods (with pros/cons):

- **Simple Random Sampling (SRS):** every unit equal chance. (Pro: unbiased; Con: needs sampling frame.)
- **Systematic Sampling:** pick every k-th unit. (Pro: easy; Con: periodicity risk.)
- **Stratified Sampling:** divide population into strata and sample within each. (Pro: more precise estimates; Con: need strata info.)

- **Cluster Sampling:** sample clusters (e.g., schools), then sample within. (Pro: cheaper for spread populations; Con: higher variance.)
- **Convenience Sampling:** use what's easy. (Pro: cheap; Con: biased — avoid for inference.)

Sampling biases to watch:

- **Selection bias:** sample not representative.
- **Non-response bias:** certain groups don't respond.
- **Measurement bias:** instrument or question wording causes bias.
- **Survivorship bias:** only survivors included (e.g., only successful companies in dataset).

Central Limit Theorem (CLT) — formal statement & intuition

Formal: For a population with mean (\mu) and finite variance (\sigma^2), the sampling distribution of the sample mean (\bar{X}) from samples of size (n) approaches a normal distribution with mean (\mu) and variance (\sigma^2/n) as (n \to \infty):

[
\bar{X} \xrightarrow{d} N\left(\mu,\ \frac{\sigma^2}{n}\right)
]

Intuition: averaging many independent random variables reduces noise; sample means gather around the true mean and form a bell curve.

Practical rule: for many distributions, (n \ge 30) is often "large enough" for CLT to apply; for highly skewed data, larger n may be required.

Confidence Interval (CI) — derivation & example

Goal: provide a range that likely contains the true population parameter (mean or proportion) at a stated confidence level (e.g., 95%).

For population mean (σ known):
[
\text{CI} = \bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}

]
Where (Z_{\alpha/2}) is standard normal quantile (1.96 for 95%).

For population mean (σ unknown, small n): use **t-distribution**:
[
\text{CI} = \bar{X} \pm t_{\alpha/2,,n-1}\frac{s}{\sqrt{n}}
]

Worked example (step-by-step) — using previous sample [12,15,14,10,19]:

- (\bar{x} = 14) (calculated earlier).
- Sample standard deviation (s \approx 3.391).
- (n = 5). For 95% CI and small n use (t_{0.975, 4} = 2.776) (from t table).

Compute standard error (SE = s/\sqrt{n}).

- (\sqrt{n} = \sqrt{5} \approx 2.23607).
- (SE = 3.391 / 2.23607 \approx 1.517).

Margin of error (= t \times SE = 2.776 \times 1.517 \approx 4.212).

So 95% CI = (14 \pm 4.212) → lower = (14 - 4.212 = 9.788), upper = (14 + 4.212 = 18.212).
Interpretation: with 95% confidence the true population mean is between 9.788 and 18.212.

Note: For large n and unknown σ, you may use Z approximation.

Sampling variation

- **Definition:** different random samples will produce different estimates.
- **Quantified by:** standard error (SE) of estimator. Example: SE of mean = (\sigma/\sqrt{n}) (or (s/\sqrt{n}) if σ unknown).
- **Reducing variation:** increase n, reduce measurement error, use stratified sampling.

EXTRA: Statistical Testing & Practical Concepts (short but crucial)

Hypothesis testing (basics)

- Null hypothesis (H_0) vs alternative (H_A).
- Compute test statistic, compare to critical value or compute p-value.
- Common tests: t-test, chi-square test, ANOVA, Mann-Whitney for nonparametric.

Effect size & practical significance

- Statistical significance ($p < 0.05$) ≠ practical importance. Report effect sizes (Cohen's d, odds ratio).

Model validation & overfitting

- Split data: train / validation / test or use cross-validation (k-fold).
- Overfitting: model captures noise → poor generalization. Use regularization (L1/L2), pruning, simpler model.

VISUAL ANALYTICS — Charts & when to use

Task	Visualization
Compare categories	Bar chart
Trend over time	Line chart / area chart
Distribution	Histogram / boxplot
Relationship between two numeric variables	Scatter plot (add regression line)
Composition	Stacked bar / pie (use sparingly)
Correlation matrix	Heatmap
Outliers	Boxplot / scatter with labels

Best practices: label axes, show units, avoid deceptive scales, provide context.

Diplomawallah.in

PRACTICAL EXAM / VIVA PREP — Common short answers & formulas

Key formulas

- Mean: (\bar{x} = \dfrac{1}{n}\sum x_i)
- Sample variance: (s^2 = \dfrac{1}{n-1}\sum (x_i - \bar{x})^2)
- SD: (s = \sqrt{s^2})
- SE(mean): (SE = s/\sqrt{n})
- CI for mean: (\bar{x} \pm t_{\alpha/2, n-1}\cdot SE)
- Sample size (proportion): (n = \dfrac{Z^2 p(1-p)}{E^2})
- Sample size (mean): (n = \dfrac{Z^2 \sigma^2}{E^2})

Short viva answers (bullet style)

- **What is CLT?** — Distribution of sample mean tends to Normal(μ,σ²/n) as n increases.
- **Why use stratified sampling?** — reduces variance when strata are homogeneous.
- **Difference mean vs median?** — mean uses all values; median is robust to outliers.
- **Why use t-test instead of Z?** — when population σ unknown and sample size small.

Diplomawallah.in

CHECKLIST: What to include in your notes for each subtopic (recommended layout)

1. Title + one line definition.
2. Expanded explanation with terms and steps.
3. Small numeric example with calculation (where possible).
4. Diagrams or textual diagram descriptions (e.g., sampling funnel: Population → Sampling frame → Sample → Analysis).
5. Key formulas boxed.
6. Applications & example from engineering / JUT context.
7. 4–6 quick revision bullets (key takeaways).

8. 2–3 likely exam questions (theory + numerical).

SAMPLE QUESTIONS (practice)

1. Compute the mean, median, variance and standard deviation for [12, 15, 14, 10, 19]. (Show steps.)
2. Explain the Central Limit Theorem and its importance in sampling.
3. Given sample mean 14, s = 3.391, n = 5, compute 95% CI for the population mean.
4. Describe 5 types of analytics and give one algorithm example for each.
5. Explain measurement scales and give examples of permissible statistics for each.

Diplomawallah.in