



SRI SATHYA SAI INSTITUTE OF HIGHER LEARNING

(Deemed to be University)

Department of Mathematics and Computer Science

ASSIGNMENTS

CIE

UDSC-402: Data Analysis and Visualization	Instructor: Dr. D Bhanu Prakash	Date: 20-08-2026
---	--	------------------

SET - 1

1. Write a one page note on set-based profiling.
2. A retail analytics team performs the following six operations on a raw point-of-sale feed: (i) splitting `TXN\_DATE` "03102015" into month, day and year columns; (ii) appending a third-party household-income table keyed on postal code; (iii) replacing NULL discount values with the column median; (iv) aggregating transactions into weekly totals per store; (v) inserting the source filename and load timestamp into every record; (vi) mapping "Californa" and "Californai" to "California". Classify each operation as structuring, enriching or cleansing, and justify the classification of (i), (ii) and (v) using the source-of-values test. **[BTL-3 — Apply]**
3. Define data wrangling and locate it precisely within the workflow framework. How does it relate to ETL?
4. A marketing department has 13 employees. Twelve salaries range from \$25,000 to \$165,000; the thirteenth is the Vice-President at \$475,000. The mean is \$100,000 and the median is \$62,000. (a) Which average would you report to a job applicant asking about a typical salary, and which to the CFO comparing departmental impact on expenses? (b) Justify each choice statistically. **[BTL-3]**

SET - 2

1. Write a one page note on Transformation: Structuring.
2. **[BTL-4 — Analyse]** A quarterly revenue report suddenly shows total revenue up 38% while unit volume is flat. Investigation reveals that a new `CUSTOMER\_MASTER` table was inner-joined to `TRANSACTIONS` on `customer\_id` last month. Diagnose the most probable cause, explain the mechanism by which it inflates revenue, and describe the profiling you would have run at the "assess data utility" stage to catch it before publication. Contrast this failure mode with its opposite.
3. Distinguish direct from indirect value and explain why one is a prerequisite for the other.

- Two warehouses each shipped 12 orders. Both have a sum of 51 days, a mean of 4.25 days and a median of 4.5 days, yet customers complain only about Warehouse B. Warehouse A shipped in 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 4 days; Warehouse B in 1, 1, 1, 3, 3, 4, 5, 5, 5, 6, 7, 10 days. Analyse why the summary statistics hid the problem, and determine which statistics would have exposed it. [BTL-4]

SET - 3

- Write a one page note on semantic profiling.
- You are handed a machine-generated e-commerce log with four columns: `ORDER\_TS` ("11072016", fixed width), `CUSTOMER` ("LLEWELLYN, CHARLES"), `NOTE` ("REFUND ISSUED (\$124.50 PENDING)") and `CART` ("{"gift_wrapped":"Y","items":["SKU19","SKU42","SKU07"]}"). Analytics must be run in a tool that accepts only rectangular input. Select the appropriate extraction technique for each column, justify each selection, and state what must be done to `CART` before analysis.
- Name the three data stages and their primary objectives. Why do most projects end in the refined stage?
- For each of the following five reporting requirements, decide whether a table or a graph should be used and justify your decision with the specific rule that applies: (i) a payroll clerk must find one employee's exact provident-fund balance; (ii) a director must see whether monthly sales are trending up or down over three years; (iii) a report must show units sold, revenue in rupees and attainment against forecast as a percentage; (iv) an analyst must find months whose values are abnormal; (v) a manager needs both monthly detail and the annual total on one page. [BTL-3]

SET - 4

- Write a one page note on syntactic profiling.
- [BTL-4 — Analyse] An analyst wrangling a 4-billion-row global transaction table works on a 1% simple random sample to keep the transform→profile loop interactive. The script converts every amount to USD using currency-specific rules. All profiling passes; in production the totals are badly wrong. Analyse why the sampling strategy caused the defect to escape, specify the sample design that would have caught it, and explain the cost that design imposes.
- Explain the data value funnel and derive from it the argument for data democratization.
- A colleague presents quarter-to-date regional sales as a three-dimensional, brightly coloured vertical bar graph with a textured background, a legend on the right, a heading "QTD Sales", and no other text. Analyse the design failures of this display and propose a redesign, justifying every change. [BTL-4]

SET - 5

1. Write a one page note on Transformations: Cleaning.
2. **[BTL-3 — Apply]** Your firm purchases a marketing-leads database of 2 million records to support a churn model. Before spending money on it, apply the five characteristics of a dataset to assess it. For each characteristic, state one assessment question you would ask and one refined-stage remedy you would apply.
3. Name the two primary bottlenecks in the traditional IT/line-of-business split and give the remedy for each.

SET - 6

1. Write a one page note on the five metadata characteristics.
2. **[BTL-4 — Analyse]** In a drug trial dataset the granularity is “one record per patient.” Records for patients who died during the trial, and for those who began showing abnormal biometrics, were removed before the file was shared. Analyse the effect of this on the validity of any efficacy analysis. Distinguish random from systematic missingness in your answer, and explain why the same 8% record loss can be benign in one case and fatal in the other.
3. Contrast schema-on-read with schema-on-write; give an infrastructure example of each.
4. Classify the categorical relationship in each of the following, then state whether a unidirectional or a bidirectional table would present it best: (i) sales by North, East, South and West; (ii) number of orders falling into order-size bands of ₹0–999, ₹1,000–1,999, ₹2,000–2,999; (iii) expenses by Division → Department → Group; (iv) sales by salesperson and by month; (v) sales, returns and net sales per salesperson. **[BTL-3]**

SET - 7

1. Write a one page note on Transformation: Enriching.
2. **[BTL-3 — Apply]** A customer table contains an `AGE` field in which unreported ages are stored as `-1`, and a `HOMETOWN` field containing values such as “San Francisco, CA USA”, “Sn Francisco, CA, USA” and “Moscow”. Apply the general recipe for profiling semantic constraints to both fields. State the summary statistics you would report, and identify which value is genuinely semantically invalid and why.
3. Distinguish generic from custom metadata using a single example (e.g. the CPG trading-partner fields).
4. Six market-share values are shown once as a pie chart and once as a sorted horizontal bar graph. Readers can rank the six values instantly from the bars but not from the pie. Analyse the perceptual reasons for this difference and state the conditions, if any, under which a pie chart is defensible. **[BTL-4]**