ASSIGNMENT TITLE
R Data Analysis Assignment: Snakebite Clinical Dataset

Dataset provided: Snake_bite_dataset.csv
Observations: 100 patients
Variables: 27 (demographics, comorbidities, labs, severity, outcomes)

Purpose
This assignment is designed to objectively assess the candidate's proficiency in R for clinical and epidemiological data analysis. The task evaluates data handling, statistical analysis, visualization, and interpretation skills using a real-world observational dataset.

Candidates are expected to demonstrate not only correct R syntax, but also appropriate statistical reasoning and professional reporting standards.

DATASET

A CSV file containing anonymized data from an observational clinical study of adult patients admitted with snakebite is provided. The dataset includes demographic variables, clinical parameters at presentation, laboratory investigations, hospital course, and outcomes.

For the purpose of this assignment, variables are classified as follows:

1) Vital parameters are physiological measurements assessed at presentation that reflect immediate cardiovascular and respiratory status.

The vital parameters in this dataset include:

- Pulse rate (beats per minute)
- Respiratory rate (breaths per minute)
- Systolic blood pressure (mmHg)
- Diastolic blood pressure (mmHg)
- Oxygen saturation, SpO_2 (%)

Derived vital-status variable:

- Hypotension at presentation (Yes/No), defined as systolic blood pressure < 90 mmHg

2) Laboratory parameters are biochemical and hematological measurements obtained from blood samples at admission, reflecting inflammation, tissue injury, and physiological reserve.

The laboratory parameters in this dataset include:

- Total leukocyte count (cells/mm^3)
- Serum C-reactive protein (CRP) (mg/L)

- Serum lactate dehydrogenase (LDH) (U/L)
- Serum albumin (g/dL)

Derived laboratory abnormality variables:

- Leukocytosis (Yes/No)
- Elevated CRP (Yes/No)
- Elevated LDH (Yes/No)
- Hypoalbuminemia (Yes/No)

Candidates are expected to correctly classify variables according to these definitions before proceeding with analysis.

INSTRUCTIONS TO CANDIDATES

- Use R (version ≥ 4.0).
- All analyses must be reproducible.
- Clearly comment your code.
- Do not modify or delete observations.
- Interpret results objectively without causal claims.

SECTION A: DATA IMPORT AND STRUCTURE (FOUNDATIONAL SKILL)

Import the CSV dataset into R.
Report:

- Number of observations and variables
- Variable classes
- Convert appropriate variables to factors with meaningful labels.

Assessment focus:
Data import, structure inspection, factor handling

SECTION B: DESCRIPTIVE STATISTICS (CORE ANALYTICAL SKILL)

Continuous variables
Compute and present:

- Mean and standard deviation
- Median and interquartile range (Q1, Q3)
- Minimum and maximum

For:

- Age

- Vital parameters
- Laboratory parameters
- Duration of hospital stay
- ICU stay duration

Categorical variables
Generate frequency and percentage tables for:
- Age groups
- Sex
- Hypotension at presentation
- Comorbidities
- ICU admission
- Overall complications

Assessment focus:
Numerical summaries, categorical summaries, clarity of output

SECTION C: DATA VISUALIZATION (APPLIED R SKILL)

Create publication-quality plots:
- Boxplot for age distribution
- Bar chart for sex distribution
- Boxplots for laboratory parameters (any 2)
- Boxplot for hospital stay

Plots must include:
- Clear titles
- Axis labels
- Appropriate themes

Assessment focus:
ggplot2 usage, aesthetic clarity, correct variable mapping

SECTION D: ASSOCIATION WITH COMPLICATIONS (STATISTICAL REASONING)

Compare laboratory parameters between:
- Patients who developed complications
- Patients who did not develop complications

Tasks:
- Compare continuous variables using appropriate statistical tests

- Compare categorical laboratory abnormalities using chi-square or Fisher's exact test where appropriate

Report:

- Test used
- p-values
- Direction of association

Assessment focus:
Test selection, interpretation, correct execution in R

SECTION E: ASSOCIATION WITH ICU ADMISSION

Compare laboratory parameters between:

- ICU-admitted patients
- Non-ICU patients

Tasks:

- Continuous variable comparisons
- Categorical abnormality comparisons
- Assessment focus:
  Group-wise analysis, outcome-based comparison

SECTION F: CORRELATION ANALYSIS (ADVANCED COMPETENCY)

Perform correlation analysis between laboratory parameters and duration of hospital stay:

- Total leukocyte count
- Serum CRP
- Serum LDH
- Serum albumin

Use Spearman's rank correlation and justify its use.
Report

- Correlation coefficients
- p-values
- At least one visual representation (scatter plot)

Assessment focus:
Correlation analysis, method selection, interpretation

SECTION G: INTERPRETATION AND PROFESSIONAL JUDGEMENT

Answer the following briefly:

- Why is correlation analysis appropriate in this context?

- Why should regression modelling be used?
- Why should causality not be inferred?

Assessment focus:
Conceptual understanding, professional judgement

SECTION H: CODE QUALITY AND REPRODUCIBILITY

Code will be assessed for:
- Readability
- Logical organization
- Adequate commenting
- Reproducibility

Deliverables:
- One R script OR R Markdown file
- All tables and plots generated via code

This assignment reflects a real clinical data analysis workflow. Candidates are evaluated not only on whether they can run R commands, but on whether they can apply correct statistical methods and communicate findings responsibly.