



STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES
MASTER OF SCIENCE IN DATA SCIENCE & ANALYTICS
DSA 8203: PRINCIPLES OF DATA SCIENCE
END OF SEMESTER EXAM

Date: 9th February 2026

Time: 3 Hours

Instructions:

This Examination consists of **FIVE** questions

Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

Question One [20 Marks]

- a. A hospital is experiencing long delays in outpatient appointments and wants to understand the causes and reduce waiting times. You are hired as a data scientist to investigate.
 - i. Identify any 4 types of data the hospital should collect to analyze the causes of outpatient appointment delays. Provide brief justification for each data type **[4 Marks]**.
- b. A Kenyan government agency overseeing a cash transfer programme analyses an administrative database of beneficiaries. The analysis reveals that some individuals are recorded more than once in the dataset.
 - i. Explain three ways in which duplicate records can arise in administrative datasets of this nature **[3 Marks]**
 - ii. Describe how to clean duplicate records, clearly stating how potential duplicates are identified and how decisions are made on which records to retain in python **[4 Marks]**
- c. A bank reports how long customer complaints take to resolve.

```
df['resolution_time_days'].plot(kind='hist', bins=40)
```

- i. Explain what information this plot conveys about service performance **[3 marks]**
- ii. Explain how bin selection could distort interpretation **[4 marks]**
- iii. After presenting the histogram of complaint resolution times to senior executives, you are asked to include this visualization in a public-facing annual report. Explain a key risk associated with reusing the same histogram designed for internal analysis in a public report **[2 Marks]**

Use the Case Study below when responding to questions Two and Three

You are a data scientist working for a public health research institute. The institute has collected a dataset containing health and lifestyle information of 15,000 adults from multiple regions over a 3-year period.

The dataset includes the following variables: Age (years), Gender (Male, Female), Region (Urban, Semi-Urban, Rural), BMI (Body Mass Index), Physical_Activity_Level (Low, Moderate, High), Daily_Calorie_Intake (kcal), Smoking_Status (Non-smoker, Former, Current), Alcohol_Consumption (units per week), Systolic_BP (mmHg), Diastolic_BP (mmHg), Cholesterol_Level (mg/dL), Year (2021, 2022, 2023)

The dataset has been merged from multiple sources and no prior data quality checks have been performed.

Question Two [15 Marks]

- a. Identify two key data quality issues that are likely to arise when merging health data from multiple sources over multiple years **[2 Marks]**
- b. The variable `Cholesterol_Level` has 12% missing values.
 - i. Describe how you would determine whether the data is Missing Completely at Random (MCAR), Missing at Random (MAR) or Missing Not at Random (MNAR) **[3 Marks]**
 - ii. Propose two appropriate strategies for handling the missing values and justify when and why each strategy would be suitable in this context **[4 Marks]**
- c. You observe several extreme values for `Daily_Calorie_Intake` (e.g., >10,000 kcal/day).
 - i. Describe the steps you would take to determine whether these extreme values represent real values **[3 Marks]**
 - ii. Explain how you would decide whether to retain, transform, or remove these extreme values, including the criteria used to support your decision **[3 Marks]**

Question Three [15 Marks]

- a. Describe how you would examine the relationship between *Physical_Activity_Level* and *BMI*
 - i. Justify the statistical summaries you would compute **[2 Marks]**
 - ii. The visual tools you would use to support your analysis **[2 Marks]**
- b. During exploratory analysis, you observe that *Alcohol_Consumption* is right-skewed
 - i. Explain how you would confirm the presence of skewness **[3 Marks]**
 - ii. Describe possible reasons for this distributional pattern **[3 Marks]**
- c. Consider the Python/pandas code snippet:

```
df.groupby('Physical_Activity_Level')['BMI'].median()
```

- i. Explain step-by-step what this code does and what the output represents **[3 Marks]**

- ii. Give two reasons why the median might be preferred over the mean when comparing BMI across activity groups **[2 Marks]**

Question Four [15 Marks]

A university uses a data-driven model to decide which applicants are admitted into its MSc Data Science program. The dataset *admissions.csv* contains:

applicant_id, age, gender, degree_class, GPA, years_experience, admission_decision.

- a. For each attribute, identify its data type by stating:
 - i. Whether it is Numeric or Categorical
 - ii. If Numeric: whether it is Continuous or Discrete and if Categorical: whether it is Nominal or Ordinal

Provide your answers in a clear list or table **[6 Marks]**

- b. Describe three univariate analyses you would perform to understand applicant characteristics. For each analysis, clearly state: the variable being analyzed, the summary statistics you would compute and the insight the analysis provides about the applicant population **[6 Marks]**.
- c. Write Python (Pandas) code to: **[3 Marks]**
 - i. Remove rows where age < 15 or age > 80.
 - ii. Convert all values in degree_class that are exactly '1st' to 'First'.

Question Five [15 Marks]

A FinTech startup offering digital wallets wants to build an executive management dashboard. You are given several CSV files with the columns: *user_id, signup_date, region, current_balance, transactions*

where transactions may contain multiple rows per user with transaction amounts and timestamps.

- a. For each of the following business questions, propose one primary visualization (and optionally a secondary one), and explain why it is suitable: **[9 Marks]**
 - i. How many new users join each week?
 - ii. What is the distribution of wallet balances?
 - iii. Which regions have the highest transaction volumes?
- b. Describe three principles of effective data storytelling for this executive dashboard. For each principle, provide a short practical example of how it would appear in the dashboard. **[6 Marks]**