Subject number : XXXXXXXXX

Sujet corrigé - 2 - Data Lifecycle

This is the **question paper**. It is **NOT** the answer sheet.
Please check that the number on your question paper matches the number on your answer sheet. To complete the answer sheet correctly, you must:

- use a **black** ink pen
- shade in the boxes **completely without going over the edges**
- if you make a mistake, erase the box with a whiteout ("Tipp-Ex"), **but do not redraw it**
- every question has a **unique correct answer**.

Box correctly ticked

1 Ⓐ ● Ⓒ Ⓓ Ⓔ

Box incorrectly ticked

1 Ⓐ ☒ Ⓒ Ⓓ Ⓔ

1 Which step of the data lifecycle involves detecting outliers and correcting errors?

1 Point - Only one correct choice

□ **A.** Collect, because data issues are identified during acquisition
□ **B.** Store, because errors are fixed during data persistence
☒ **C.** Clean, because this stage focuses on addressing data quality issues
□ **D.** Communicate, because findings are validated during presentation

> While data issues can be noticed in other stages, cleaning is specifically dedicated to correcting errors and outliers.

2 What is the first step in the data lifecycle?

1 Point - Only one correct choice

□ **A.** Cleaning, because raw data is always messy
□ **B.** Analysis, because questions drive the entire process
☒ **C.** Collection, because data must be gathered before any other steps
□ **D.** Storage, because data must be saved immediately

> Collection logically precedes all other steps, as no analysis or cleaning can occur without data.

3 Which stage of the data lifecycle is often the most time-consuming?

1 Point - Only one correct choice

□ **A.** Analysis, because modeling and statistics require deep expertise
☒ **B.** Cleaning and preprocessing, because raw data often contains inconsistencies, missing values, and errors
□ **C.** Storage, because choosing the right format is complex
□ **D.** Collection, because gathering data from multiple sources is labor-intensive

> While all stages require effort, cleaning and preprocessing typically consume the most time due to data imperfections.

4 What is the purpose of the storage stage in the data lifecycle?

1 Point - Only one correct choice

□ **A.** To visualize data trends over time
☒ **B.** To save cleaned data in a structured format for future use
□ **C.** To collect additional data from new sources
□ **D.** To generate reports for stakeholders

> Storage ensures that cleaned data is preserved and accessible for further analysis or sharing.

5 In the Titanic dataset lab, how were missing values in the "age" column handled?

1 Point - Only one correct choice

□ **A.** By deleting the entire column to avoid bias
□ **B.** By dropping all rows with missing values to ensure data integrity
☒ **C.** By filling them with the average age to retain as much data as possible
□ **D.** By replacing them with zeros to simplify calculations

> Filling missing values with the mean/median is a common imputation strategy to avoid losing rows.

6 **In the analysis stage, what is the purpose of using "groupby" in pandas?**
1 Point - Only one correct choice

□ **A.** To remove irrelevant columns from the dataset
□ **B.** To merge multiple datasets into one
☒ **C.** To compute aggregated statistics for specific subsets of data
□ **D.** To generate visualizations of trends

> "groupby" enables aggregation and comparison across data subsets (e.g., by category).

7 **In the analysis stage, what does "df.describe()" provide?**
1 Point - Only one correct choice

□ **A.** A list of all missing values in the dataset
☒ **B.** Summary statistics for numeric columns
□ **C.** A visualization of data distributions
□ **D.** A cleaned and filtered version of the dataset

> "describe()" offers a statistical overview of numeric data, such as central tendency and spread.

8 **Why is it important to document the data lifecycle process?**
1 Point - Only one correct choice

□ **A.** Documentation is only necessary for raw data sources
☒ **B.** To ensure reproducibility, transparency, and collaboration
□ **C.** To delete intermediate files and reduce clutter
□ **D.** To skip the communication stage in future projects

> Documentation enables others (or your future self) to understand and replicate the process.

9 **For a smart city traffic management project with high-velocity, unstructured data (e.g., sensor logs, images), which storage and processing strategy is most appropriate?**
1 Point - Only one correct choice

□ **A.** Store data in a single CSV and process it hourly using Excel
□ **B.** Use a relational database and SQL queries for structured analysis
☒ **C.** Implement a distributed system with a NoSQL database for scalability
□ **D.** Preprocess data at the source (edge computing) and store summaries in CSV files

> Distributed systems and NoSQL databases handle high-volume, unstructured, real-time data effectively.

10 **When presenting findings to a non-technical audience (e.g., city officials), which communication strategy is most effective?**
1 Point - Only one correct choice

□ **A.** Share raw data tables and pandas outputs for transparency
□ **B.** Provide the Python code to demonstrate rigor
☒ **C.** Use interactive visualizations and focus on actionable insights
□ **D.** Include a technical appendix with p-values and confidence intervals

> Visual, intuitive presentations with clear takeaways resonate best with non-experts.