Subject number : XXXXXXXXX

Sujet corrigé - 4 - Data Wrangling

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 **According to the ISO 25012 standard, which of the following is NOT a data quality dimension?**

1 Point - Only one correct choice

☐ **A.** Validity
☐ **B.** Accuracy
☐ **C.** Completeness
☒ **D.** Scalability

> Scalability is not a data quality dimension; it refers to the ability of a system to handle growing amounts of data, not the quality of the data itself.

2 **What is the main risk of imputing missing data with the mean?**

1 Point - Only one correct choice

☐ **A.** It increases the influence of outliers on the dataset
☒ **B.** It underestimates variability and weakens relationships among variables
☐ **C.** It has no measurable statistical bias if data are missing completely at random
☐ **D.** It increases the effective sample size without adding information

> Replacing missing values with a constant (the mean) reduces variance and weakens correlations.

3 **In the lab, why was the median used to impute missing values in the `Age` column?**

1 Point - Only one correct choice

☐ **A.** The median preserves the exact distribution of the original data.
☒ **B.** The median is robust to outliers and invalid data
☐ **C.** The median ensures that the imputed values match the most frequent age in the dataset.
☐ **D.** The median is computationally faster than other imputation methods.

> The median was used because it is robust to outliers and invalid values, such as -1.

4 **In the lab, how were the incorrect values in the `Oxygen Saturation` column corrected?**

1 Point - Only one correct choice

☐ **A.** Deleted
☐ **B.** Imputed with the mean
☒ **C.** Capped at 100%
☐ **D.** Left as is and flagged

> Incorrect values in the `Oxygen Saturation` column were capped at 100%.

5 **After we have assessed data quality, we handle missing values, outliers, and duplicates. What is the typical order?**

1 Point - Only one correct choice

☐ **A.** Handling missing values, then outliers, and finally duplicates
☐ **B.** Handling outliers, then duplicates, and finally missing values
☒ **C.** Handling duplicates, then missing values, then outliers
☐ **D.** Handling outliers, then missing values, and finally duplicates

6 **What is an outlier?**

1 Point - Only one correct choice

☐ **A.** A value that lies within one standard deviation of the mean.
☐ **B.** A value that is equal to the mode of the dataset
☒ **C.** A value that is significantly different from the others
☐ **D.** A value that is incorrect

> An outlier is a data point that is significantly different from the rest of the data, which may be due to variability, errors, or rare but valid occurrences.

7 **Assume that we have a price feature in a dataset with values [14, 17, 20, 23, 26]. What will the value 20 become after applying Z-score standardization?**

1 Point - Only one correct choice

☒ **A.** 0
☐ **B.** 0.5
☐ **C.** 1
☐ **D.** 20

> 20 is the mean of the value, and Z-score standardization transforms the values to a Gaussian with center (mean) 0 and width (standard deviation) 1.

8 **What is a potential drawback of one-hot encoding?**

1 Point - Only one correct choice

☐ **A.** It converts categorical data into numerical data
☒ **B.** It can significantly increase the dimensionality of the dataset
☐ **C.** It reduces the interpretability of categorical variables.
☐ **D.** It cannot handle missing values in categorical data.

> One-hot encoding can significantly increase the dimensionality of the dataset, especially when there are many unique categories, which can lead to computational inefficiency.

9 **Which reshaping operation corresponds to a SQL join?**

1 Point - Only one correct choice

☐ **A.** Pivoting
☒ **B.** Merging
☐ **C.** Melting
☐ **D.** Concatenation

> Merging combines datasets based on one or more keys (common columns), essentially corresponding to a SQL Join.

10 **What does data aggregation refer to in data wrangling?**

1 Point - Only one correct choice

☐ **A.** Combining related datasets to have all the features available at the same time
☒ **B.** Combining multiple data points to produce summary statistics
☐ **C.** Combining groups from a dataset that was split for detailed inspection
☐ **D.** Combining records without altering the underlying values

> Aggregation summarizes or consolidates data (e.g., sums, averages, counts) to make it easier to analyze or report.