EDA_TREAT - Questions

Q1. When examining missingness in a multi-system SCM dataset, which scenario most strongly indicates MNAR (Missing Not At Random)?

A. Temperature_C missing evenly across all Product Categories
B. Random gaps in Weight_kg due to scanner outages
C. Notes missing specifically when shipments had no issues recorded
D. Missing Delivery_Time_hrs appearing after dtype coercion

Correct answer: C. Notes missing specifically when shipments had no issues recorded

Q2. Why must logical cleaning (negative or zero values) come before median imputation?

A. Because median computation requires all columns to be categorical
B. Logical errors must be turned into NaN before they distort the imputation value
C. Weight values cannot be imputed without fixing Notes first
D. Outlier treatment cannot run otherwise

Correct answer: B. Logical errors must be turned into NaN before they distort the imputation value

Q3. Converting Delivery_Time_hrs using pd.to_numeric(errors='coerce') is necessary because

A. It replaces all numeric values with strings
B. It forces pandas to guess missing categories
C. It detects corrupted text entries and turns them into NaN for proper cleaning
D. It automatically fixes outliers

Correct answer: C. It detects corrupted text entries and turns them into NaN for proper cleaning

Q4. After plotting Delivery_Time_hrs before TREAT, the distribution shows heavy right tail delays. What is the correct TREAT response?

A. Replace all values above the mean with the median
B. Delete all rows with high delivery times
C. Use IQR thresholds and apply .clip() to cap extreme values
D. Expand the scale to hide the outliers

Correct answer: C. Use IQR thresholds and apply .clip() to cap extreme values

Q5. Why is Weight_kg imputed by Product_Category instead of using the global median?

A. Different categories have different typical weights, so a global median would distort product-level

KPIs

B. Warehouse_ID determines the weight distribution

C. Temperature sensors influence weight

D. MAR cannot occur in SCM datasets

Correct answer: A. Different categories have different typical weights, so a global median would distort product-level KPIs