

Data Preprocessing

# Terminology

|  |  |
| --- | --- |
| Data Preprocessing | * وهى العمليات اللى بتم قبل ما ابدأ فى التحليل لان البيانات بتكون غير مهيئة للتحليل بسبب المشاكل اللى فى البيانات * Datasets are highly susceptible to noisy, missing, and inconsistent data. * Low-quality data will lead to low-quality mining results. * Data quality factors includes:   1. accuracy,   2. completeness,   3. consistency,   4. timeliness,   5. believability   6. interpretability. |
|  | * Inaccurate data   + Having incorrect attribute values (e.g., by choosing the default value “January 1” displayed for birthday) * Incomplete data   + Missing data (may not always be available or of interest) * Inconsistent data   + different assessments of the quality depends on the intended use of the data * Timeliness   + (e.g. month-end data are not updated in a timely fashion has a negative impact on the data quality. ) * Believability   + reflects how much the data are trusted by users * Interpretability   + reflects how easy the data are understood (e.g. sales codes) |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |