



B.Com Honours

Semester II

Calicut University

Data Analytics for Business

Course Code: COM2MN109 • Module 2 Notes

1. Data Collection, Structures, and Pre-processing

An analytical model is only as good as the data fed into it ("garbage in, garbage out"). Data preparation is the most time-consuming phase in the data analytics lifecycle. This module covers the four scales of measurement, data structures, quantitative vs. qualitative data, and steps for data cleaning, pre-processing, and quality assurance.

The Four Scales of Measurement

Statistical data is classified into four measurement levels, which determine the type of analysis possible:

Scale of Measurement	Mathematical Properties	Business Example
Nominal Scale	Categorical data only. No mathematical order or rank.	Customer gender, payment method (UPI, Card, Cash).
Ordinal Scale	Categories with a clear, logical order or rank. Intervals are unequal.	Customer satisfaction ratings (Poor, Average, Excellent).
Interval Scale	Ordered categories with equal intervals. No true zero point.	Temperature (Celsius), credit scores (300-900).
Ratio Scale	Ordered categories, equal intervals, and a true zero point.	Monthly sales revenue, transaction count, height, weight.

Data Structures: Structured, Unstructured, Semi-Structured

Structured Data

Highly organized data fitting neatly into relational databases (tables with rows and columns). E.g., SQL tables, Excel sheets containing transaction details.

Unstructured & Semi-Structured

- **Unstructured:** Data that has no pre-defined organizational model (e.g., video files, emails, PDF reports, customer audio calls).
- **Semi-Structured:** Data that contains tags or markers to separate elements but doesn't fit a relational model (e.g., JSON, XML files).

Methods for Data Cleaning and Pre-processing

Before starting analysis, raw datasets must be cleaned and pre-processed:

1. **Handling Missing Values:** Deleting rows with missing elements, or using ****Imputation**** (filling missing gaps with the column mean, median, or mode).
2. **Deduplication:** Removing duplicate entries to prevent over-counting.
3. **Data Normalization:** Scaling numeric columns to a standard range (e.g., 0 to 1) so that columns with large scales don't distort model weights.
4. **Data Validation:** Running consistency checks (e.g., ensuring ages are positive numbers and email addresses contain "@").

Acing your exams is just a click away!

Visit www.degreeelive.in to download the next module for free.

DegreeLive