DATA CLEANING

MACQUARIE University
SYDNEY·AUSTRALIA

Data cleaning is imporant for graphical explorations for the variables can be helpful and numerical summaries is are easily understandable; counts (n), means and standard deviations for numerical variables). Without ensuring the data is valid, it is impossible to have insight into business operations and impossible to produce evidence to support decision making.

How to Prepare

a) AChoosing the variables for analysis
b) Checking the data for correctness
c) correcting errors, dealing with missing values, deciding what to do with extreme values
d) Deciding how to format the data (unit of measurement, data type, number of categories)
e) Creating new variables from the existing ones

Problems in data cleaning:

I**ncomplete:** variables have missing values
Noisy: data contains errors or outliers (Salary="-10")
Inconsistent: data contains discrepcodes or names (e.g., Age="42" Birthday="03/07/1997"

Missing Data

Missing at random:
The reason the data is missing is random.
Missing not at random:
The data is missing is systematically related to the unobserved data, that is, the missingness is related to events or factors which are not measured by the researcher.

Outlier

An outlier is an observation that lies a large (abnormal) distance from other values.

Outliers should only be removed if there is a good reason to believe they are faulty data

Keep an outlier If it is a reliable measurement