Session 1 Handout: Prepare the Data

FAST-11 AI-Assisted Analytics and Business Intelligence · 7formations · Instructor: Wadood Chaudhary · All examples fictional

Why this session matters

Every analysis inherits the problems in its data. AI features make cleanup much faster: they spot duplicates, mixed date formats, and inconsistent names in seconds. They will also happily "fix" a blank by guessing or delete a real outlier. This session is about using AI to find problems while you make the fixes yourself, with repeatable, logged steps.

Key concepts

1. Data rules come first

Never in class tools	OK in class
Customer personal information (name plus SSN, license, account or card number; G.L. c. 93H)	Fictional course datasets
PHI, even with names removed	De-identified, fictional scheduling files
Confidential financials, bids, pricing	Public data (for example, Census tables)

At work: check the file's sensitivity label and your employer's approved tools before loading anything.

2. Cleaning with AI help

- Ask the AI to **find** problems: duplicates, mixed date formats, name variants, inconsistent capitalization, blanks, outliers.
- Ask for **formulas or query steps**, not a cleaned file.
- **Apply fixes yourself** in Excel, Power Query, or Tableau Prep.
- **Never guess** a blank or delete an outlier without checking the source.
- Spot-check rows before and after.

3. Document every transformation

Step	What changed	Rows	Reason	Method
2	Machine names standardized to VF-4	2	Variants split totals	Replace values

Power Query records each change as an **Applied Step** that reruns on refresh, which is a log you get for free. Keep your W1 log next to the dataset. **Test:** could a coworker repeat your cleanup from the log alone?

Worked example (fictional)

Bayside's 20-row scrap excerpt had a duplicate (same event in two date formats), machine variants ("VF4", "vf 4"), a blank quantity, and a quantity of 60 that was a typo for 6. Cleaned total: **$1,365.50**. The raw file would have reported **$2,430.50**, overstating scrap by $1,065. The typo alone added $993.60.

Checklist

- ☐ Data classified; approved tool confirmed.
- ☐ Raw file kept untouched.
- ☐ AI used to find problems; fixes applied with formulas or steps.
- ☐ Blanks and outliers checked at the source, not guessed.
- ☐ Every step logged with rows affected.
- ☐ Totals compared before and after.

Do and don't

Do	Don't
Ask for steps you can rerun	Accept a mystery "cleaned file"
Confirm duplicates by inspection	Delete every repeated part number
Flag blanks for the source owner	Fill blanks with averages
Log rows removed and values replaced	Clean silently