Data Engineering Assessment - Take home

Power Outage Data Pipeline (2-3 hours)

Context

You're joining an energy policy team that monitors grid reliability. They've collected power outage incident reports but the data is messy and inconsistent. Your task is to build a foundation for analyzing this data.

The Data

Primary Dataset: DOE Form OE-417 Annual Summaries

- Download 2-3 recent years from: https://doe417.pnnl.gov/
- Excel files with ~300-400 incident records per year
- 11 columns: dates, times, location, NERC region, event type, impact metrics

What you'll encounter:

- Event types with inconsistent formatting ("- Vandalism" vs "Vandalism")
- NERC regions with multiple formats (e.g., "MRO/RF", "SERC,MRO", "SERC / RF")
- Long free-text "Alert Criteria" field (31-480 characters)
- ~40% of incidents have zero customers affected AND zero demand loss
- Date and time in separate columns with different types
- Area affected as semi-structured text ("California: Riverside County;")

Your Task

Build a data pipeline that:

1. Loads 2-3 years of Excel files
2. Cleans and normalizes the messy fields
3. Handles the data quality issues you find
4. Outputs structured data ready for analysis
5. Generates basic statistics/insights

What We're Looking For

- Do you explore the data before coding?
- How do you handle the inconsistencies?
- What assumptions do you make about the zeros?
- Do you normalize categories or keep them raw?

- Do you keep it simple or over-engineer?

Deliverables

Submit code + a brief write-up (1-2 pages) covering:

- What data quality issues did you find?
- What decisions did you make and why?
- What would you do differently with more time?
- Key statistics/insights from the data

Notes

- Use any language/tools you prefer
- Don't spend time on fancy visualizations
- We care more about your thinking than perfect code
- Document anything you find unclear or problematic

Hints

- Excel files have a title row - you'll need to handle headers correctly
- Event types have leading dashes and capitalization issues
- NERC regions use different delimiters (, / and spaces)
- Consider: what do the zero values mean? Bad data or valid incidents?
- Alert Criteria is long free text - do you parse it or leave it?
- How would this scale to 20+ years of data?