

Critical Skills and Practice

Lab C: Exploring Data

In this lab you will act as a member of a company's data analysis team. In groups of 4, you will be handed a paper dataset of employee salaries that is messy. Your task is to clean the dataset so it can be used to analyse pay disparities by gender, race, and education level, controlling for industry, job title, location, and years of experience. As you make changes, you must log every modification using pen on a physical version control document, so that the next team can pick up your dataset and continue the work without needing to ask you what you did.

Version control is a critical skill regardless of what career path you follow. Many workplaces use tools like GitHub, but others use proprietary or sector-specific systems or apply version control principles to documents and datasets rather than code. You will work in a team throughout, and hand off your dataset and documentation to the next lab group at the end of the session.

Dataset: A collection of data usually organised into rows and columns (like a spreadsheet). Each row is typically one "record" (e.g., one employee), and each column is one piece of information about that record (e.g., salary, job title).

Version control: A system for tracking changes made to a file or document over time, so you can see who changed what, when, and why. It allows multiple people to collaborate without overwriting each other's work or losing track of the dataset or document's history. In software, this is often done with tools like Git/GitHub. In this lab, you'll practice the same principle by hand.

1. Receiving and reviewing your materials (20 minutes)

You will be given:

- A paper copy of a messy salary dataset
- A blank Version Control Document and Log (provided separately)

As a team, read through the dataset and the version control log from previous teams. Discuss what "clean" needs to mean for your analysis (e.g. consistent formatting, no duplicates, plausible values, complete records) and which problems in the dataset are most urgent to fix first

2. Cleaning and logging (60 minutes)

Work through the dataset making changes directly on the paper with your pens. For every single change you make, you must add an entry to the Version Control Log. Each log entry should include:

- **ID:** A unique number for the change
- **Team:**
- **Message:** A clear description of what was changed and why
- **Type:** e.g. correction, deletion, standardisation
- **Location:** Where in the dataset the change was made (ID in column 1))
- **Old → New:** The original content and what you changed it to

3. Version Control (15 minutes)

You must also complete the top section of the Version Control Document:

- **Outgoing Team**
- **Date**
- **Description of Dataset:** What the dataset contains and its intended purpose
- **Total Changes Logged:** The final count once you finish
- **Summary of Changes:** A brief overview of the main types of cleaning carried out
- **Known Limitations:** Anything you were not able to fix or decisions the next team should double-check

Be as clear and specific as possible. The team receiving your dataset will only know what you tell them. If your log is vague or incomplete, they may repeat work, miss errors, or misunderstand your decisions.

4. Handover (5 minutes)

Pass your dataset and completed Version Control Document and Log to the demonstrator.