Data Cleaning & Rationale

Comprehensive Data Cleaning & Exploratory Analysis of Job Market Trends

Anu Sharma, Cindy Guzman, Gavin Boss

# Data Cleaning Overview

In this section, we outline key decisions made during the data preprocessing phase, focusing on column relevance, code redundancy, and the impact on downstream analysis.

## Identifying Irrelevant or Redundant Columns

After reviewing the dataset, we identified several columns that were either:

* **Irrelevant** to our analysis goals (e.g., internal IDs, timestamps not tied to labor trends)
* **Redundant** due to duplication or overlapping information

### Examples of Columns Removed:

* record\_id, submission\_timestamp: Metadata not used in analysis
* naics\_code\_2017, naics\_code\_2022: Multiple versions of the same classification system
* soc\_code\_2010, soc\_code\_2018: Legacy codes that overlap with updated versions

## Why Remove Multiple Versions of NAICS/SOC Codes?

|  |
| --- |
| Note |
| **NAICS (North American Industry Classification System)** and **SOC (Standard Occupational Classification)** codes are updated periodically. Including multiple versions introduces: - Redundancy - Confusion in grouping industries or occupations - Risk of double-counting or misalignment in trend analysis |

We retained only the **most recent version** of each code to ensure consistency and relevance to 2024 labor market trends.

## How This Improves Analysis

Cleaning the dataset in this way improves our analysis by:

* **Reducing noise**: Fewer columns means clearer signals
* **Improving interpretability**: Analysts and readers can focus on current classifications
* **Enhancing visualizations**: Charts and tables are easier to read and more meaningful
* **Ensuring consistency**: Aligns with external sources like Lightcast and BLS data

## Next Steps

With a cleaner dataset, we’re now ready to explore key workforce themes—such as AI-driven job growth, salary disparities, and gender-based employment patterns—using reliable, streamlined data.