Documentation: Cleaning Occupational Employment and Wage Statistics (OEWS)

Purpose:

The purpose of this cleaning step was to simplify the OEWS dataset so that it focuses on the variables that matter for occupational wage and employment analysis. The original file contained over **6 million records**, including metadata fields such as footnote codes that added no analytical value. By removing unnecessary columns and ensuring uniqueness, the dataset is now leaner, transparent, and ready for large-scale labor market analysis.

Steps Taken

1. Reading the raw dataset

```
df = pd.read_csv('Occupational Employment and Wage Statistics (OEWS).txt', sep='\t')
```

I loaded the original OEWS dataset into a pandas DataFrame. The file was tab-delimited and contained millions of rows with multiple series identifiers, values, and metadata.

2. Inspecting the data

```
print(df.head())
```

I printed the first few rows to confirm the structure and ensure the dataset loaded correctly.

3. Dropping redundant columns

```
col_to_drop = 'footnote_codes'
df = df.drop(columns=col_to_drop)
```

I removed the **footnote_codes** column, which contained metadata flags but no analytical value for wage or employment comparisons.

4. Checking for duplicates

print(df.duplicated().sum())

I verified that the dataset contained **0 duplicate rows**, ensuring accuracy in aggregations and avoiding inflated totals.

5. Saving the cleaned dataset

df.to_csv('Clean Occupational Employment and Wage Statistics (OEWS).txt', sep='\t', index=False)

Finally, I exported the cleaned dataset in tab-delimited format. The result is a leaner file that highlights the essentials: occupational employment counts, wage estimates, percentiles, and statistical measures.

Key Decisions and Rationale

- Footnote codes dropped: These were metadata only and did not contribute to wage or employment analysis.
- Duplicates checked: Confirmed none existed, ensuring reliable totals and averages.
- Large-scale efficiency: Cleaning steps were designed to handle over 6 million records efficiently while maintaining auditability.
- Business focus sharpened: By keeping only occupational identifiers, employment counts, and wage measures, the dataset now directly supports labor market analysis and wage benchmarking.
- Transparency ensured: Every cleaning step was documented, making the dataset reproducible and audit-ready.

Data Source: Databases, Tables & Calculators by Subject