Documentation: Cleaning Quarterly Census of Employment and Wages

My Goal:

The purpose of this cleaning step was to simplify the dataset so that it focuses on the variables that matter for analysis of employment, wages, and contributions across industries and regions. I wanted to remove redundant metadata, disclosure codes, and technical identifiers that added no analytical value, while retaining establishment counts, employment levels, wage measures, and over-the-year changes. The result is a lean, transparent dataset that is ready for labor market and economic analysis.

Steps Taken

1. Reading the raw dataset

```
df = pd.read_csv('Quarterly Census of Employment and Wages.csv')
```

I loaded the original QCEW dataset into a pandas DataFrame. The file contained many fields, including aggregation codes, disclosure flags, and industry identifiers that were not essential for the intended analysis.

2. Inspecting the data

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

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

3. Dropping redundant or misleading columns

```
cols_to_drop = ['agglvl_code', 'size_code', 'disclosure_code',
        'lq_disclosure_code', 'oty_disclosure_code', 'industry_code']
df = df.drop(columns=cols_to_drop)
```

I removed columns that either:

- Added no analytical value (disclosure_code, lq_disclosure_code, oty_disclosure_code).
- Were metadata only (agglvl_code, size_code).
- Were redundant (industry_code was already represented by ownership and area codes).

This step sharpened the dataset to focus on establishments, employment, wages, contributions, and change metrics.

4. Checking data types

```
print(df.dtypes)
```

I verified the data types of each column to ensure numeric fields were properly recognized for calculations and aggregations.

5. Checking for duplicates

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

I confirmed that the dataset contained **0 duplicate rows**, which ensures accuracy in aggregations and avoids inflated totals.

6. Saving the cleaned dataset

```
df.to_csv('Cleaned Quarterly Census of Employment and Wages.csv', index=False)
```

Finally, I exported the cleaned dataset. The result is a leaner file that highlights the essentials: establishment counts, employment levels, wage totals, contributions, average weekly wages, and over-the-year changes.

Key Decisions and Rationale

- Disclosure codes dropped: These fields were metadata for suppression rules and did not add analytical value.

- Aggregation and size codes removed: These were technical identifiers that introduced unnecessary complexity.
- Industry code excluded: Since ownership and area codes already provide classification, this field was redundant.
- Duplicates checked: Confirmed that no duplicate rows existed, ensuring reliable totals and averages.
- Focus sharpened: By keeping establishments, employment, wages, contributions, and change metrics, the dataset now directly supports labor market and wage trend analysis.
- Transparency ensured: Every cleaning step was documented, making the dataset reproducible and audit-ready.

Data Source: Databases, Tables & Calculators by Subject