Documentation: Cleaning HS 3542 UN Comtrade USA Exports Dataset

My Goal

The purpose of this cleaning step was to simplify the dataset so that it focuses on the variables that actually matter for analysis of U.S. semiconductor exports. I wanted to remove redundant or misleading fields, convert critical columns to numeric types, and produce a file that is lean, transparent, and ready for budgetary or trade-partner analysis.

Steps Taken

1. Reading the raw dataset

```
df = pd.read_csv('HS 3542 Un Comtrade USA Exports.csv', encoding='Latin1')
```

I loaded the original UN Comtrade export dataset into a pandas DataFrame. This file contained many fields, some of which were either redundant or consistently empty. I specified encoding='Latin1' to ensure proper handling of special characters.

2. Converting key columns to numeric

```
for col in ['partnerCode', 'cmdCode', 'qty', 'qtyUnitCode', 'altQtyUnitCode', 'altQty',
'fobvalue', 'primaryValue']:
    df[col] = pd.to_numeric(df[col], errors='coerce')
```

I explicitly converted important fields (partner codes, commodity codes, quantities, and values) into numeric types. This ensures that calculations and aggregations won’t break due to string formatting issues. Using errors='coerce' safely turns invalid entries into NaN, which is easier to handle than inconsistent text.

3. Dropping redundant or misleading columns

```
cols_to_drop = [
  'flowCode', 'partner2Code', 'partner2ISO', 'mosCode', 'motCode',
  'netWgt', 'grossWgt', 'legacyEstimationFlag', 'cifvalue',
  'primaryValue', 'altQtyUnitCode', 'altQtyUnitAbbr', 'altQty',
  'isAltQtyEstimated', 'motDesc', 'customsDesc', 'isGrossWgtEstimated',
  'isNetWgtEstimated', 'qty', 'qtyUnitCode', 'isAggregate',
  'cmdCode', 'partnerCode', 'partnerDesc', 'reporterISO'
]
df = df.drop(columns=cols_to_drop)
```

I removed columns that either:

- Added no analytical value (partner2Code, partner2ISO, mosCode).
- Were consistently empty or zero (netWgt, grossWgt).
- Were redundant (primaryValue duplicated fobvalue; altQty duplicated qty).
- Introduced unnecessary complexity (motCode, cifvalue, customsDesc).
- Were metadata fields not needed for analysis (reporterISO, partnerDesc).

This step was about clarity: keeping only the fields that contribute meaningfully to trade analysis.

4. Replacing placeholder values

```
df['reporterDesc'] = df['reporterDesc'].replace('X', 'United States')
```

In the raw dataset, the reporter country was sometimes marked as X. I standardized this by replacing X with **United States** to ensure consistency and avoid confusion in downstream analysis.

5. Saving the cleaned dataset

```
df.to_csv('Cleaned_HS_3542_Un_Comtrade_USA_Exports.csv', index=False)
```

Finally, I exported the cleaned dataset. The result is a leaner file that highlights the essentials: reporter, partner, commodity code, and FOB value.

Key Decisions and Rationale

- Weights dropped: Since netWgt and grossWgt were always zero, I removed them to avoid misleading interpretations.
- Logistics ignored: Transport mode (motCode, motDesc) was not detailed enough to support meaningful logistics analysis, so I focused instead on value and partner data.
- Redundancy reduced: Columns like primaryValue and altQty duplicated existing information, so I eliminated them for simplicity.
- Placeholder fixed: Replacing X with "United States" ensures clarity and avoids mislabeling in country-level analysis.
- Focus sharpened: By cleaning aggressively, I ensured the dataset is ready for budgetary analysis and partner comparisons, which are the most relevant insights here.

Data Source: UN Comtrade