File names:

01.1\_check\_4.py

01.2\_check\_3.py

02.1\_current\_4\_part1

02.2\_constant\_4\_part1

02.3\_current\_3\_part1

02.4\_constant\_3\_part1

03.1\_current\_4\_part2

03.2\_constant\_4\_part2

03.3\_current\_3\_part2

03.4\_constant\_3\_part2

01.1\_check\_4.py & 01.2\_check\_3.py

A pre-screen to check for ISICC 3 and ISICC 4 files,

Per country, year, and series\_code (and base year for constant files), is there more than one currency type('currency')?

Per country, year, and series\_code (and base year for constant files), is there more than one fiscal year type ('year\_type')?

All checked out.

02.1\_current\_4\_part1 (02.1\_constant\_4\_part1)

Initial steps

1. Import the original data.
2. Retrieve iso columns and country columns and create a iso mapping file. (no need for constant file)

Data frame transfer and analysis

1. Use pivot table method and create a data file with columns ‘country’, ‘year’, ‘sereis\_code’ (and ‘base’). And reset the index.
2. Add a new column aggregate footnotes. (columns\_to\_cal have different range)
3. Add a new column showing all missing sectors
4. Add a new column showing the number of sectors that are not showing NA.
5. Add new columns showing if column B is NA; if column D is NA; if column E is missing. Combining the information from number of sectors, add a new column showing if only sector B is missing, and another column showing if only D or E is missing.
6. Calculate the discard score with 0.005 threshold.
   1. In process, add a column as the sum of all sectors’ value.
   2. In process, calculate the discard score.
   3. Compare the score with 0.005 threshold and get True or False. If the score is over 0.005, then discard.
   4. In addition, if b.1g is empty, change the value in column ‘discard’ to False.
7. Take FISIM into consideration.
   1. Check footnotes. Need manual check to see if only b.1g is mentioning this footnote.
   2. If the footnote said FISIM is excluded from the b.1g.
      1. If the sum of sectors are smaller than b1.g, then there’s an issue, discard.
      2. If the sum of sectors are bigger than b1.g, then use the 0.075 as the new threshold.
   3. If the footnote said FISIM is included from the b.1g.
      1. If the sum of sectors are bigger than b1.g, then there’s an issue, discard.
      2. If the sum of sectors are smaller than b1.g, then use the 0.075 as the new threshold.
8. Save a copy of current discard column, call it ‘discard\_nc\_Nsector’
9. Check for the number of sector not showing NA and missing sectors. There should be in total 15 sectors, if anyone is missing, and that’s not D or E, then discard.

Final Data Clean and Export

1. Only keep the time range from 1990 to 2023.
2. Reset year and series into integer.
3. Merge iso3.
4. Export the current data file to ‘un4\_current\_Sectorvalues.csv’
5. Keep those no need to be discarded (discard = False) and export to ‘un4\_current\_intermediate.csv’
6. Put those need to be discarded and export to ‘un4\_current\_intermediate\_discard.csv’
7. Keep only columns ‘country’, ‘year’, ‘series\_code’, ‘iso3’. Export to ‘un4\_current\_4cols.csv’

03.1\_current\_4\_part2 (constant)

Initial Steps

1. Read the data file completed from part 1. The one ended with 4cols.
2. Data cleaning. Update index, etc.

Data Transforming and Analysis

1. Use pivot table function and get the data file with columns ‘country’, ‘year’, ‘base’, all series code columns (include, 100, 200, 1000, etc).
2. Reset index. Sort the data frame with 1)country, 2)base year, 3)year
3. Add a column ‘count’ that shows how many valid series are for each country and year combination.
4. Add a column to get the highest series for each country and year combination.
5. Add a column to show final series.
   1. If count =1, meaning only 1 valid series are there for the country & year combination, put the column name of the series into the final\_series cell.
   2. If count >1,
      1. First from bottom up (larger years are in the bottom), forward checking with next row, if same country and the next row has a series chosen that is also valid for this row, then use the same series of next row.
      2. Then from top down, backward checking with previous row, if same country and the previous row has a series chosen that is also valid for this row, then use the same series of the previous row.
   3. Check for the number of NAs in column ‘final\_series’. Check the countries that have NA in this column.
      1. If the number is limited, print it and check.
      2. If the number is big, export these countries with remaining data to a csv file and manually check.
6. For those still shows NA in final\_series, put the highest\_series as the final\_series for now. And there should no NA left.
7. Check for switch and overlap.
   1. Add a new column ‘n\_series’, showing how many different series are there for one country.
      1. Check for the distribution of switches. Getting a number of how many overlapping lines will need to be added in the future.
      2. Print the tail, instead of the head, of the data frame. Keep this number for future calculation of newly added rows.
   2. Add a new column ‘switch’, marking the rows for the switch of series within one country.
   3. Add a new column ‘overlap’. From bottom up, for the same country, if there’s a switch, check if there’s a row can serve as the overlap row. If so, add the new row of overlap row and mark them.
   4. Print the tail, instead of the head, for the data frame. For the calculation of how many lines are added.
      1. If the new lines added are the same as the overlap lines needed, then no need to worry about the overlap issue any more.
      2. If the new lines added < the overlap lines needed. It means there are some switch do not have overlap rows. Need to check for them.
         * If the lines need to be checked are limited, can print and do the manual check.
         * If the lines need to be checked are many, export to csv file then check.
8. Gap check. For some countries, there might be year gap although the series did not switch. Export the countries that has gap to a csv file for a manual check.

Final Data Clean and Export

1. Merge iso mapping table.
2. Export the data into csv files.