Data Cleaning and Preprocessing

a) Data Cleaning and Data Transformation

Below are the data cleaning method that have been performed on this dataset.

1) Remove columns from continents_2:
   - alpha_2
   - country-code
   - iso_3166-2
   - sub-region
   - intermediate-region
   - region-code
   - sub-region-code
   - intermediate-region-code

2) Rename columns name in First Merge table:
   - continents2.name → country_name
   - continents2.region → continent

3) Standardize the naming of countries in column continents2.name:
   - Congo (Democratic Republic Of The) → Democratic Republic Of The Congo
   - Micronesia (Federated States of) → Federated States Of Micronesia
   - Korea, Republic of → North Korea

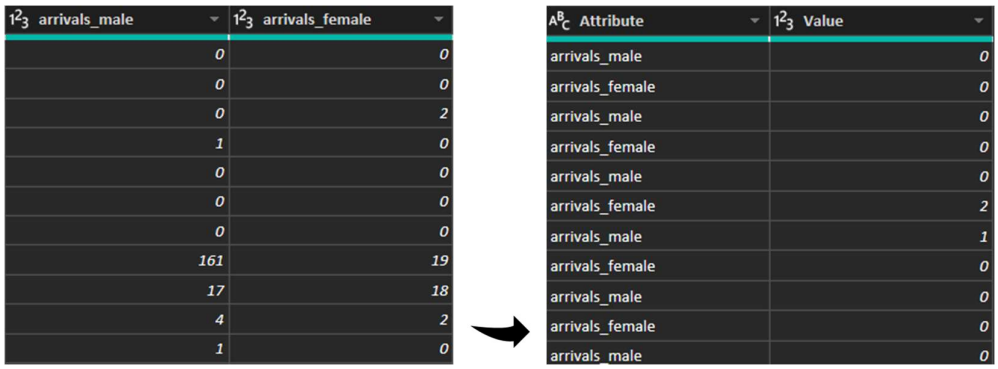
4) Populate the missing continent in continents2.region:

Missing Values	Populated Values
alpha-3	**region**
ATA	Antarctica

5) Populate the missing country in First Merge Table with these matching values with reference to the continents2 table.

Missing Values	Populated Values		
country	**country**	**country_name**	**continent**
ANT	ATG	Antigua and Barbuda	Americas
ZAR	COD	Congo (Democratic Republic Of The)	Africa
XXX	XXX	Unspecified	Unspecified

6) In Second Merge table, unpivot arrivals_male and arrivals_female to put into matrix format and create a new table named Finalize Table.



arrivals_male	arrivals_female
0	0
0	0
0	2
1	0
0	0
0	0
0	0
161	19
17	18
4	2
1	0

Attribute	Value
arrivals_male	0
arrivals_female	0
arrivals_male	0
arrivals_female	0
arrivals_male	0
arrivals_female	2
arrivals_male	1
arrivals_female	0
arrivals_male	0
arrivals_female	0
arrivals_male	0

7) Next, standardize the naming of gender in column Attribute at Finalize Table:
   - arrivals_male → Male
   - arrivals_female → Female

8) Rename the column name in Finalize Table:
   - Attribute → gender
   - Value → total_arrivals

9) Remove column arrivals in Finalize Table.

10) The Finalize Table is now clean and will be used for analysis and visualization. This table has a total of seven columns which are date, country, country_name, continent, soe, gender and total_arrivals with 185,348 rows.