Data Ingestion and Integration

a) Data Source Identification

A total of two datasets are utilized for the purpose of this project. The first dataset is extracted in CSV format from Malaysia's Official Open Data Portal which can be accessible at data.gov.my. This dataset contain information on monthly foreigner arrivals to Malaysia from January 2020 until October 2024 provided by Immigration Department of Malaysia. It is important to note that approximately 0.01% of arrivals are diplomats, stateless individuals or refugees, thus they did not have specific nationality and labelled as unspecified with abbreviation 'XXX' in the dataset. Meanwhile, the second dataset is also in CSV format extracted from open-source platform, Kaggle. This dataset provides details on country ISO and region information which are useful to categorize the arrivals based on continents.

The two datasets are listed as below:

i) Monthly Arrivals by State of Entry, Nationality & Sex
ii) Country Mapping – ISO, Continent, Region

The details on the datasets are provided in Table 1 below:

Details	**Dataset 1** Monthly Arrivals by State of Entry, Nationality & Sex	**Dataset 2** Country Mapping – ISO, Continent, Region
Format	CSV	CSV
Source	Malaysia's Official Open Data Portal	Kaggle
No. of Rows	92,674	249
No. of Columns	6	11
Variables	i) date ii) country	i) name ii) alpha-2

	iii) soe iv) arrivals v) arrivals_male vi) arrivals_female	iii) alpha-3 iv) country-code v) iso_3166-2 vi) region vii) sub-region viii) intermediate region ix) region-code x) sub-region-code xi) intermediate-region-code

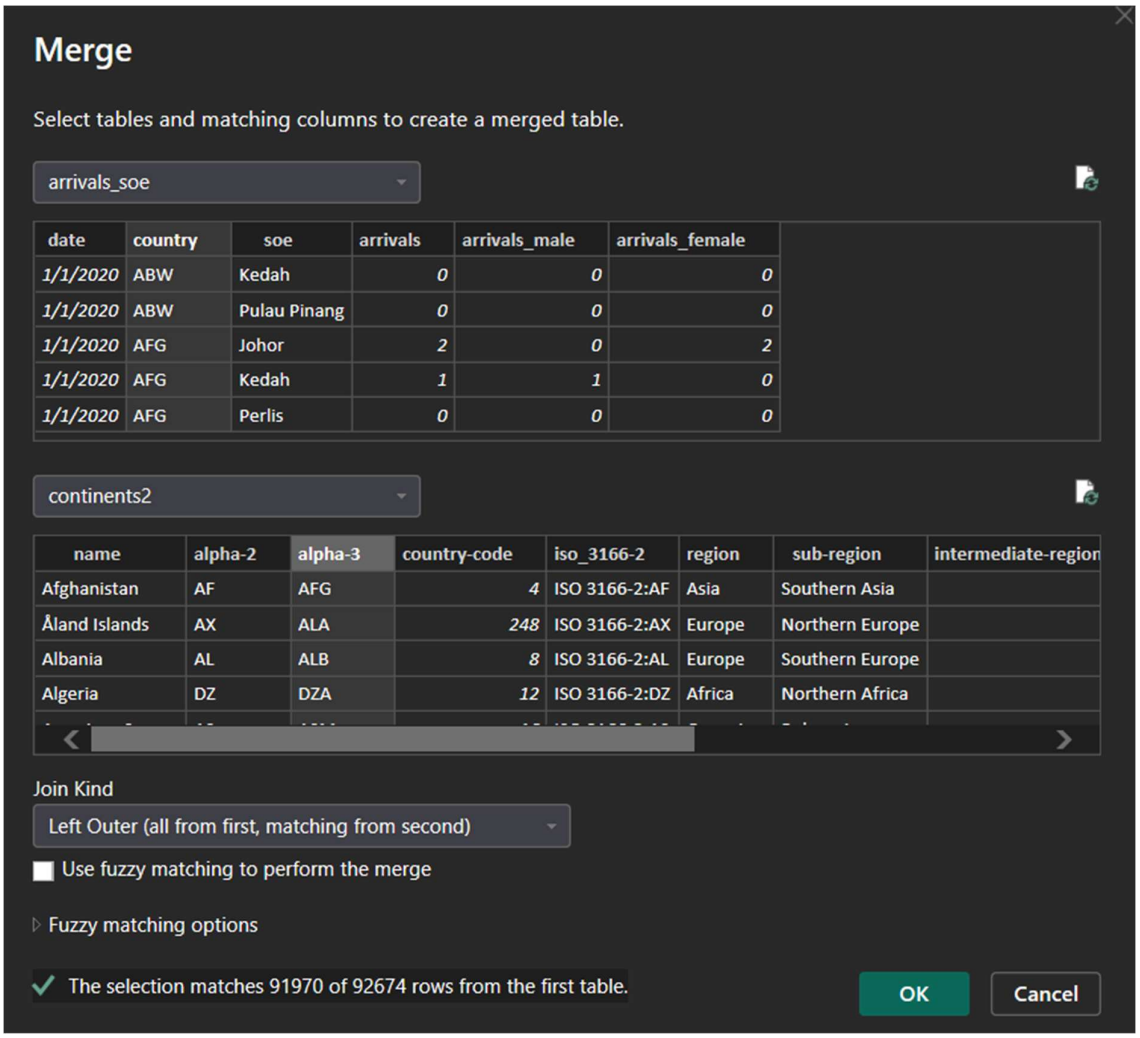
Table 2 below shows the data attributes and its purpose

No.	Variables	Data Type	Data Description
Dataset 1			
1	date	Ordinal	The first day of the month with date format dd/mm/yyyy
2	country	Nominal	The origin country of a foreigner with the country names using Alpha-3 code as per International Standards Organization (ISO)
3	soe	Nominal	State of Entry in which the foreigners entering from with the state names using the full spellings.
4	arrivals	Ratio	Total number of foreigners arrived regardless of gender.
5	arrivals_male	Ratio	Total number of males foreigners arrived.
6	arrivals_female	Ratio	Total number of females foreigners arrived.
Dataset 2			
7	name	Nominal	The name of a country.

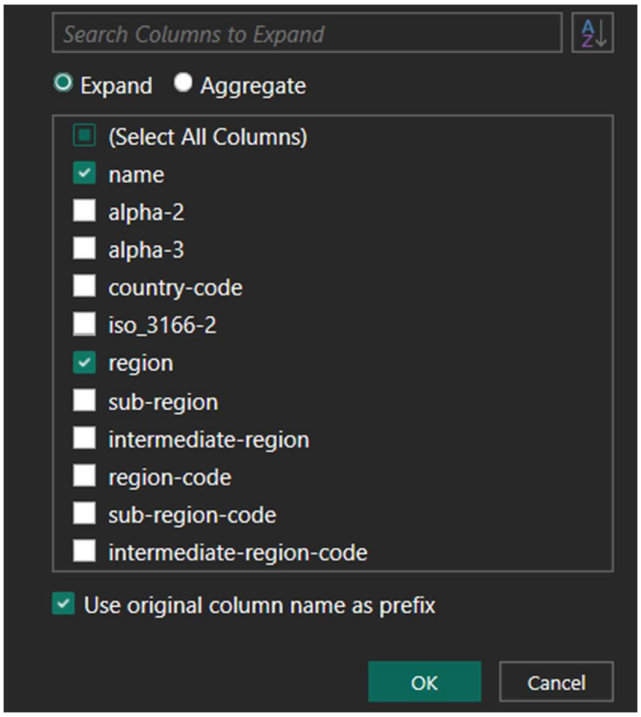
8	alpha-2	Nominal	A two-letter ISO code for each country.
9	alpha-3	Nominal	A three-letter ISO code for each country.
10	country-code	Nominal	A unique number assign to each country.
11	iso_3166-2	Nominal	The entry code for a country in ISO 3166.
12	region	Nominal	The continent of a country.
13	sub-region	Nominal	The sub-continent of a country.
14	intermediate region	Nominal	The smaller fraction under sub-region.
15	region-code	Nominal	The unique code for region.
16	sub-region-code	Nominal	The unique code for sub-region.
17	intermediate-region-code	Nominal	The unique code for intermediate region.

b) Data Integration

During the process of data integration, this project combines both datasets using merge function in Power Query. It merges using Left Outer Join on country column in arrivals_soe dataset and alpha-3 in continents2 dataset.



The new merged table was created and named as First Merge. The First Merge table has a new column at the end of query which contains the contents of the continents2 table that was merged with the arrivals_soe table. This project chooses to expand the columns and only include name and region column in the merged table.



Next, rename the columns. The continents2.name was renamed as country name, while continents2.region was renamed as continent.

continents2.name	continents2.region
Aruba	Americas
Aruba	Americas
Afghanistan	Asia

country name	continent
Aruba	Americas
Aruba	Americas
Afghanistan	Asia

The challenge encountered during data integration process is there are country codes that have no matching country name and continent.

country	country name	continent
XXX	*null*	*null*
ZAR	*null*	*null*
ANT	*null*	*null*

To solve this issue, this project creates a new table named Missing Country to populate the missing values. Then, Missing Country table is merged again with the First Merge table and creates a new table named Second Merge. The Second Merge table is now free from missing values.

	country	updated country	continents2.name	continents2.region
1	ANT	ATG	Antigua and Barbuda	Americas
2	ZAR	COD	Congo (Democratic Republic Of The)	Africa
3	XXX	XXX	Unspecified	Unspecified