UNIVERSITI MALAYSIA PAHANG AL-SULTAN ABDULLAH

BSD2434 DATA WAREHOUSING GROUP PROJECT

GROUP 02: THE SNACK ATTACKERS

TITLE:

INTEGRATED ANALYSIS OF FOOD SYSTEM: UNDERSTANDING HUNGER DYNAMIC (SDG 02: ZERO HUNGER)

PREPARED BY:

MATRIC ID	NAME	SECTION
SD22029	ALENA NG PEI YEEN (LEADER)	01G
SD22031	ANIS AQILAH BINTI MOHD ASRI	01G
SD22004	NUR AINUL NASUHA BINTI MOHD AZREEN	01G
SD21020	SITI NUR FARAHANA BT SAPUAN	01G

SD22034	MUHAMMAD AIMAN IRFAN BIN AHMAD MUHYE	01G

TABLE OF CONTENT

1.0 BACKGROUND

1.1 PROJECT DESCRIPTION

One of the Sustainable Development Goals (SDGs) of the United Nations, zero hunger (SDG 2) is the goal of ensuring that everyone has access to enough safe, nutrient-dense food to meet their dietary needs and preferences and they can obtain the food they need. Even though there is already more than enough food produced to feed everyone, up to 828 million people still suffer from chronic undernourishment. Malnutrition is having a severe negative impact on both developed and developing countries. In our project, we have chosen Integrated Analysis of Food System: Understanding Hunger Zero Dynamic as our title, which is closely connected with Zero Hunger of Sustainable Development Goals. This project analyzes the holistic project seeks to uncover the underlying causes of hunger, identify vulnerable populations, and propose sustainable solutions to alleviate food insecurity.

The main goal of this project is to discover ways to integrate multiple datasets for holistic analysis. To conduct an in-depth analysis of food security and nutrition outcomes, we have integrated several datasets. For instance, to find intricate relationships and factors causing food insecurity at the local, national, and international levels, we combined data concerning food prices with details regarding the global hunger index, climate change, malnutrition, risk factor, and crop production. Furthermore, we make use of these datasets that offer crucial data for efficient monitoring and optimization. The crop production dataset works as a first table that indicates the most important part for this project. Among all attributes, the values are designated as both primary key and measurements, with 'measure', 'subject', 'value' and 'year' categorized under attributes.

The crop production dataset contains location, subject (type of crop), time (year data was inserted), measure (metric used) and value (value according to metric used). The attributes help to analyze how many harvested productions per unit were made from all countries selected. It also shows the consumption patterns that have great incidence on agricultural commodities prices. From this dataset, we can detect how many products are made from the range and identify which type of products were produced. Therefore, we can detect what are the most demand products from each country. Also, we can address food availability and make wise decisions to improve food security based on the outcome from this study.

Besides that, we focused on the cause of death dataset, which describes the various reasons contributing to fatalities linked to zero hunger. We can't simply attribute these deaths to a single cause, as multiple factors like malnutrition and malaria play significant roles. We analyzed the relation of zero hunger from cause of death dataset by diseases like malaria, maternal disorders, meningitis, neonatal disorders, neoplasms, nutritional deficiencies, Parkinson's disease, poisonings and malnutrition. These diseases were chosen because they are closely connected to zero hunger. Determine zero hunger as a cause of death requires a mix of quantitative and qualitative techniques that consider broader health factors, such as food security, malnutrition, and socioeconomic circumstances.

Additionally, the information on climate change has crucial impacts on the agriculture industries. Changes in temperature and extreme weather contribute to reducing agriculture productivity. This dataset contains information about carbon dioxide emissions, country, date, humidity, location, precipitation, sea-level rise, temperature, and wind speed. By considering these variables, we obtained how weather conditions impact crop growth and ensure a consistent food supply for all. To address food insecurity, it is essential to comprehend how hunger is getting worse by climate change. Policymakers can set priorities for reducing hunger and ensuring sustainable food systems by examining these effects.

Additionally, consumer expenditure dataset contains parameters including code (short name from country), country, food as share of total consumer expenditure and year (2017-2021). The cost of food has a direct impact on people's capacity to obtain enough nutrition, which is essential for overcoming zero hunger. High prices can make it difficult for those who are poor to afford basic food items, which may increase food insecurity and stunt growth. Understanding this dataset allows us to anticipate problems and implement prevention measures, preventing long-term impacts on agricultural productivity and ensuring a stable food supply. Moreover, it enables us to monitor the ongoing process effectively.

On the other hand, the global hunger index dataset which contains code, country, global hunger index and year is important to generate accurate analysis. From this dataset, we can obtain the number of undernourishments from each country. Through this analysis, we aim to inform evidence-based policies and programs designed to achieve the goal of zero hunger and reduce food insecurity. In addition, our project will investigate the well-performing social safety nets, nutrition interventions, and food security initiatives currently in place work to address issues associated with hunger.

The disease of malnutrition continues to be the reason for effecting children much more vulnerable to sickness and death. This malnutrition dataset includes country, income classification, overweight, severe wasting, stunting, US population (‘000s), underweight and wasting. All this information is important to analyze malnutrition and places taken. We used the income attribute as we found out higher income produces less hunger issue since they have enough money to buy their own food and vice versa. From this, we can determine how malnutrition is related to zero hunger issue, and we developed ways to prevent it from keep going.

In conclusion, our project, "Integrated Analysis of the Food System: Understanding Hunger Dynamics," is designed to identify and analyze the primary causes of hunger. By examining multiple factors and applying various methodologies, we aim to optimize our approach and gain a deeper understanding of the underlying drivers of hunger.

1.2 PROBLEM TO BE SOLVED

Despite global efforts to eliminate hunger, millions of people around the globe still suffer from food insecurity and starvation. Understanding the complex dynamics of hunger requires an integrated analysis of food systems, encompassing several factors such as malnutrition, risk factor, global hunger index, food price, climate change and crop consumption. In this project, we aim to discover the integration of multiple datasets for holistic analysis towards achieving the sustainable development goal of zero hunger (SDG 2).

The problem to be solved:

1. What are the regions within the nation that have the highest levels of the Global Hunger Index, stunting, nutritional deficiencies, and protein-energy malnutrition fatalities?
2. How can nutritional deficiency and malnutrition of Malaysia countries as captured in the dataset, be analyzed to gain insights to promote balanced diets and address specific nutritional deficiencies?
3. How can crop production and food price datasets be used to understand its impact to enhance efficiency in food security systems?
4. What measure can be implemented, to leveraging insight from datasets on malnutrition, risk factor, global hunger index, food prices, climate change and crop consumption to improve overall food security performance to optimized rate of hunger in Malaysia?

These problems highlighted the need for data analysis and optimization strategies to improve various aspects of the food security infrastructure. The data warehouse findings give significant information to researchers, assisting in the evolution of food security knowledge. In addition, these insights assist with zero hunger decision-making processes, resulting in optimized rate of hunger, and lower total of death in Malaysia a year.

1.3 OBJECTIVE

1) To examine the highest level of the global hunger index, stunting, nutritional deficiencies, and protein-energy malnutrition fatalities across the nation.

2) To analyze the nutritional deficiency and malnutrition of countries from the dataset, to gain insight to promote balanced diets and address specific nutritional deficiencies in Malaysia

3) To study the impact of crop production and food prices to enhance the efficiency in food security system.

4) To improve overall food security performance by leveraging insights from datasets on malnutrition, risk factors, global hunger index, food prices, climate change and crop production to optimize the rate of hunger.

1.4 DATA SCHEMA

In the database world, the word schema refers to the organization and structure of the database. It contains the object of the schema, which can be a table, column, data type, view, stored procedures, relationships, primary key, foreign key, and so on. A database schema can be represented by a visual diagram, which describes the relationship between the database and the tables. The dataset of this project contains seven tables which are Cause of Death and Malnutrition acts as facts table and the other tables are Global Hunger Index, Consumer Expenditure on Food, Crop Production, and Climate Change. To prototype the data schema, we have used 4 library packages and python code in Jupyter Notebook as shown in Figure 1.4.1.

```
In [1]: import pandas as pd
        import psycopg2 as ps
        import pandas.io.sql as sqlio
        import missingno as msno
```

Figure 1.4.1: Libraries used to find data schema

```
In [3]: df_fact1 = pd.read_excel("CauseOfDeath.xlsx")
        df_fact1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 34 columns):
 #   Column                                  Non-Null Count  Dtype 
---  ------                                  --------------  ----- 
 0   Country/Territory                       30 non-null     object
 1   Code                                    30 non-null     object
 2   Year                                    30 non-null     int64 
 3   Meningitis                              30 non-null     int64 
 4   Alzheimer's Disease and Other Dementias  30 non-null     int64 
 5   Parkinson's Disease                     30 non-null     int64 
 6   Nutritional Deficiencies                30 non-null     int64 
 7   Malaria                                 30 non-null     int64 
 8   Drowning                                30 non-null     int64 
 9   Interpersonal Violence                  30 non-null     int64 
 10  Maternal Disorders                      30 non-null     int64 
 11  HIV/AIDS                                30 non-null     int64 
 12  Drug Use Disorders                      30 non-null     int64 
 13  Tuberculosis                            30 non-null     int64 
 14  Cardiovascular Diseases                 30 non-null     int64 
 15  Lower Respiratory Infections            30 non-null     int64 
 16  Neonatal Disorders                      30 non-null     int64 
 17  Alcohol Use Disorders                   30 non-null     int64 
 18  Self-harm                               30 non-null     int64 
 19  Exposure to Forces of Nature            30 non-null     int64 
 20  Diarrheal Diseases                      30 non-null     int64 
 21  Environmental Heat and Cold Exposure    30 non-null     int64 
 22  Neoplasms                               30 non-null     int64 
 23  Conflict and Terrorism                  30 non-null     int64 
 24  Diabetes Mellitus                       30 non-null     int64 
 25  Chronic Kidney Disease                  30 non-null     int64 
 26  Poisonings                              30 non-null     int64 
 27  Protein-Energy Malnutrition             30 non-null     int64 
 28  Road Injuries                           30 non-null     int64 
 29  Chronic Respiratory Diseases            30 non-null     int64 
 30  Cirrhosis and Other Chronic Liver Diseases 30 non-null  int64 
 31  Digestive Diseases                      30 non-null     int64 
 32  Fire, Heat, and Hot Substances          30 non-null     int64 
 33  Acute Hepatitis                         30 non-null     int64 
dtypes: int64(32), object(2)
memory usage: 8.1+ KB
```

Figure 1.4.2: Data Schema Causes of Death Tables

Figure 1.4.2 shows the global Causes of Deaths with various factors. This table contains 34 columns in total which are Country/Territory, Code, Year, Meningitis, Alzheimer's disease and other Dementias, Parkinson's disease, Nutritional Deficiencies, Malaria, Drowning, Interpersonal Violence, HIV/AIDS, Drug use Disorder, Tuberculosis, Cardiovascular Disease, Lower Respiratory Infections, Neonatal Disorders, Self-harm, Exposure to Forces of Nature, Diarrheal Disease, Environmental Heat and Cold Exposure, Neoplasms, Conflict and Terrorism, Diabetes Mellitus, Chronic Kidney Disease, Poisonings, Protein-energy Malnutrition, Road Injuries, Chronic Respiratory Diseases, Cirrhosis and other Chronic Liver Diseases, Digestive Diseases, Fire, Heat, and Hot Substance, and lastly Acute Hepatitis. There are 32 integer columns and 2 columns with object data type.

```
In [4]: df_fact2 = pd.read_excel("Malnutrition.xlsx")
        df_fact2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 152 entries, 0 to 151
Data columns (total 9 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Country                152 non-null    object
 1   Country.1              152 non-null    object
 2   Income Classification  152 non-null    int64
 3   Severe Wasting         140 non-null    float64
 4   Wasting                150 non-null    float64
 5   Overweight             149 non-null    float64
 6   Stunting               151 non-null    float64
 7   Underweight            150 non-null    float64
 8   U5 Population ('000s)  152 non-null    float64
dtypes: float64(6), int64(1), object(2)
memory usage: 10.8+ KB
```

Figure 1.4.3: Data Schema Malnutrition

Figure 1.4.3 shows global Malnutrition with various attributes. This table contains 9 columns in total which are Country, Country.1, Income Classification, Severe Wasting, Wasting, Overweight, Stunting, Underweight, and U5 Population ('000s). There are 3 data types in this table which are float, integer, and string. These data types are distributed to 6 float, 1 integer, and 2 string columns respectively.

```
In [5]: df_dim1 = pd.read_excel("CropProduction.xlsx")
        df_dim1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19437 entries, 0 to 19436
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Country  19437 non-null  object
 1   Subject  19437 non-null  object
 2   Measure  19437 non-null  object
 3   Year     19437 non-null  int64
 4   Value    19437 non-null  float64
dtypes: float64(1), int64(1), object(3)
memory usage: 759.4+ KB
```

Figure 1.4.4: Data Schema Crop Production

Figure 1.4.4 shows global crop production with various attributes. This table contains 5 columns in total which are Country, Subject, Measure, Year, and Value. There are 3 data types in this table which are float, integer, and object. These data types are distributed to 1 float, 1 integer, and 3 string columns respectively.

```
In [6]: df2 = pd.read_excel("ClimateChange.xlsx")
        df2.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            10000 non-null  datetime64[ns]
 1   Year            10000 non-null  int64
 2   Location        10000 non-null  object
 3   Country         10000 non-null  object
 4   Temperature     10000 non-null  float64
 5   CO2 Emissions   10000 non-null  float64
 6   Sea Level Rise  10000 non-null  float64
 7   Precipitation   10000 non-null  float64
 8   Humidity        10000 non-null  float64
 9   Wind Speed      10000 non-null  float64
dtypes: datetime64[ns](1), float64(6), int64(1), object(2)
memory usage: 781.4+ KB
```

Figure 1.4.5: Data Schema Climate Change

Figure 1.4.5 shows global climate change with various attributes. This table contains 10 columns in total which are Date, Year, Location, Country, Temperature, CO2 Emissions, Sea level Rise, Precipitation, Humidity and Wind Speed. There are 4 data types in this table which are datetime, float, integer, and string. These data types are distributed to 1 datetime, 6 float, 1 integer, and 2 string columns respectively.

```
In [7]: df_dim3 = pd.read_excel("GlobalHunger.xlsx")
        df_dim3.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 471 entries, 0 to 470
Data columns (total 5 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Country                      471 non-null    object
 1   Code                         471 non-null    object
 2   Year                         471 non-null    int64
 3   Global Hunger Index (2021)   471 non-null    float64
 4   411773-annotations           12 non-null     object
dtypes: float64(1), int64(1), object(3)
memory usage: 18.5+ KB
```

Figure 1.4.6: Data Schema Global Hunger

Figure 1.4.6 shows global hunger with various attributes. This table contains 5 columns in total which are Country, Code, Year, Global Hunger Index (2021). There are 3 data types in this table which are float, integer, and string. These data types are distributed to 1 float, 1 integer, and 3 string columns respectively.

```
In [8]: df_dim4 = pd.read_excel("ConsumerExpenditureOnFood.xlsx")
        df_dim4.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 4 columns):
 #   Column                                       Non-Null Count  Dtype
---  ------                                       --------------  -----
 0   Country                                      520 non-null    object
 1   Code                                         520 non-null    object
 2   Year                                         520 non-null    int64
 3   Food as share of total consumer expenditure  520 non-null    float64
dtypes: float64(1), int64(1), object(2)
memory usage: 16.4+ KB
```

Figure 1.4.6: Data Schema Consumer Expenditure on Food

Figure 1.4.6 shows the Consumer Expenditure on Food with various attributes. This table contains 4 columns in total which are Country, Code, Year and Food as share of total consumer expenditure. There are 3 data types in this table which are float, integer, and object. These data types are distributed to 1 float, 1 integer, and 2 string columns respectively.

1.5 DATA WAREHOUSE DESIGN

Inmon and Kimball are two well-known approaches that we discussed in depth during our group project discussion on data warehouse architecture for the integrated study of food systems. Each approach offers unique perspectives on how to structure data warehousing architecture, and we aimed to determine which would be most suitable for our project, "Integrated Analysis of Food System: Understanding Hunger Dynamics (SDG 02: Zero Hunger)."

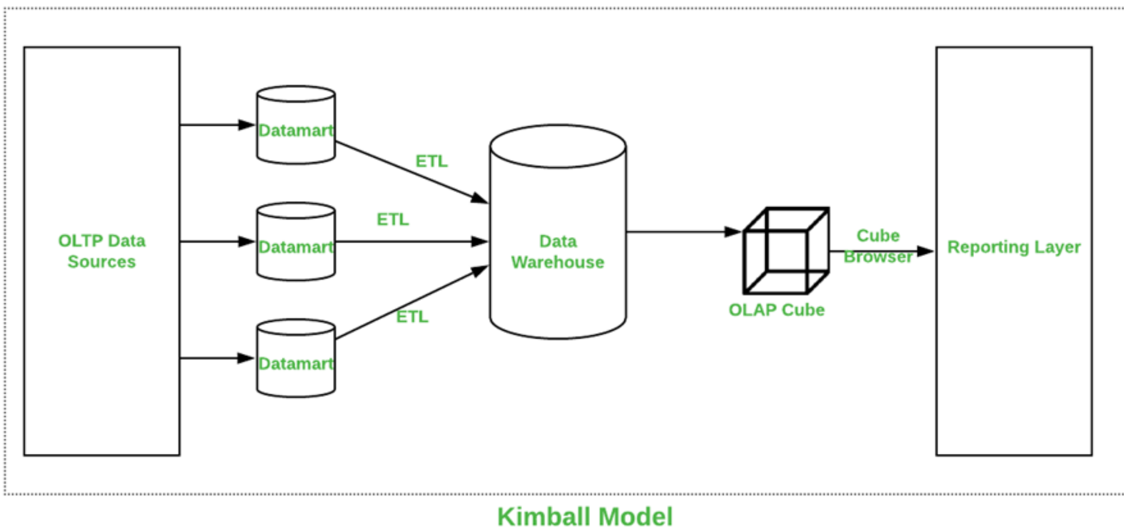


Figure 1.5.1: Kimball Model

Ralph Kimball presented the Kimball approach for creating a Data warehouse. The first step in this strategy is to identify the business process and queries that the data warehouse must address. These data sets are being thoroughly examined and then recorded. After gathering all the data from various data sources, or data marts, the Extract Transform Load (ETL) software loads it into a central location known as staging. This is subsequently converted into an OLAP cube.

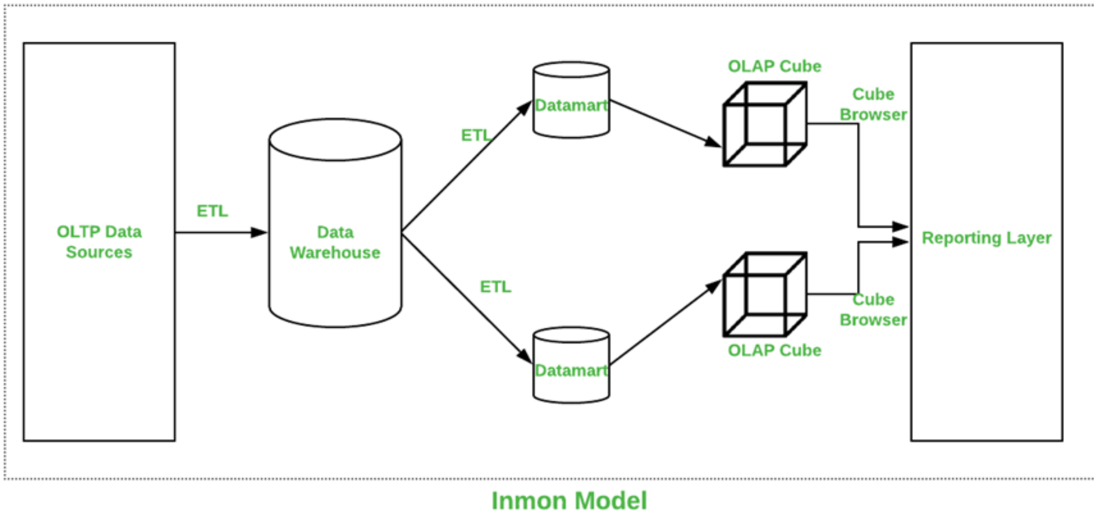


Figure 1.5.2: Inmon Model

Meanwhile, Inmon's approach in designing a Data warehouse was introduced by Bill Inmon. This model recognizes key areas and takes care of customers, products, and vendors. This model serves for the creation of a detailed logical model used for major operations. Details and models are then used to develop a physical model. This model is normalized and makes data redundancy less. This is a complex model that is difficult to use for business purposes for which data marts are created and each department can use it for their purposes.

We assessed the needs and goals in relation to our research on the dynamics of hunger within food systems to choose the best technique. Considering the detailed and diverse nature of our project, we have determined that the Inmon technique is more suitable for our requirements. Furthermore, the Inmon's methodology adaptability to the latest additions and modifications to the data model fits in well with the dynamic character of food systems and the developing knowledge of the dynamics of hunger. As new data sources become available or analytical techniques advance, we can easily incorporate these updates into the existing dimensional model and data marts without disrupting the entire data warehouse architecture.

In summary, the best strategy for building the data warehouse for our project is the Inmon approach as it provides a solid and flexible infrastructure. Furthermore, it focuses on creating a comprehensive and integrated data warehouse as the central repository of data, from which dimensional models and data marts can be derived, that makes us effectively analyse large, complicated datasets, get practical insights, and support Malaysia effort to achieve zero hunger.

2.0 ARCHETECTURE AND ETL PIPELINE

2.1 PIPELINE STRUCTURE

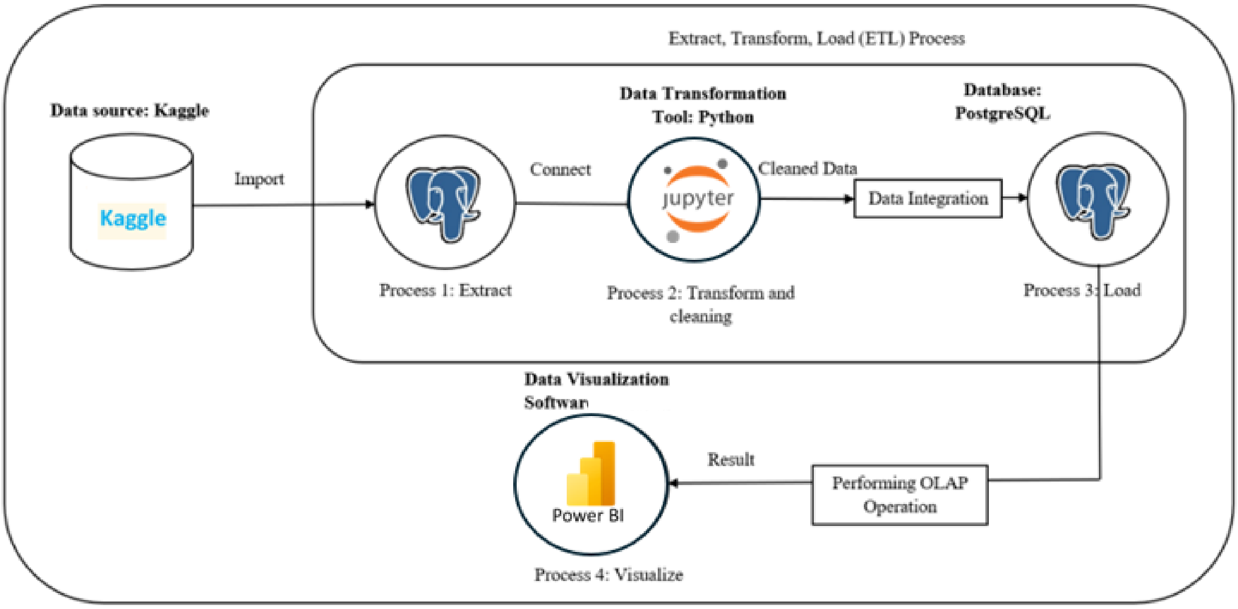


Figure 2.0 Pipeline structure

Based on the data collection and preparation, we utilize Kaggle as a platform to access relevant datasets related to malnutrition, risk factor, global hunger index, food price, climate change and crop consumption. Besides, we also utilize Department of Statistics Malaysia (DOSM) website to find datasets specifically in Malaysia since our group want to narrow down our insight to Malaysia country. Python was used to download and preprocess the datasets. Python provides libraries such as Pandas and Numpy that facilitate data extraction, cleaning, and transformation.

The dataset consists of 6 tables which are malnutrition, risk factor, global hunger index, food price, climate change and crop consumption. To start working with the dataset, we created a database and tables in PostgreSQL. After importing the tables, we used Jupyter for further data operations.

To transform the data, we start by installing the Pandas library for data cleansing, loading, and saving. We load multiple datasets, check the data types, identify missing values, and remove incomplete data. Tables are then combined to create a unified view, and the integrated data is saved as new clean CSV files. After importing the files back into PostgreSQL, we establish a connection to Python for further data preparation. By installing the necessary libraries and configuring the connection, we identify and remove NULL values from the data. This process ensures efficient data transformation and cleaning, providing reliable and prepared data for analysis.

Once the cleaning and transformation process is complete, the data is saved in a new CSV file and imported back into PostgreSQL. With the data imported, we can then perform OLAP operations for multidimensional analysis and proceed with the data visualization. This includes operations like roll-up, slicing, pivot and dicing, which provide deeper insights into the data and enable effective data exploration and visualization.

After completing the OLAP operations, the generated results are imported into Tableau for data visualization. Tableau provides a user-friendly platform for creating visualizations that allow us to observe and analyze the desired outcomes. With its interactive and customizable features, Tableau enables us to create insightful visual representations of the data, facilitating better understanding and decision-making based on the analysis results.

2.2 ETL Pipeline

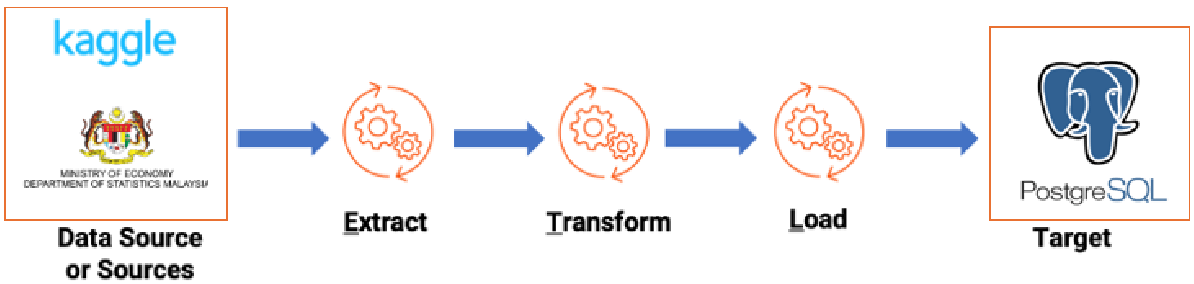
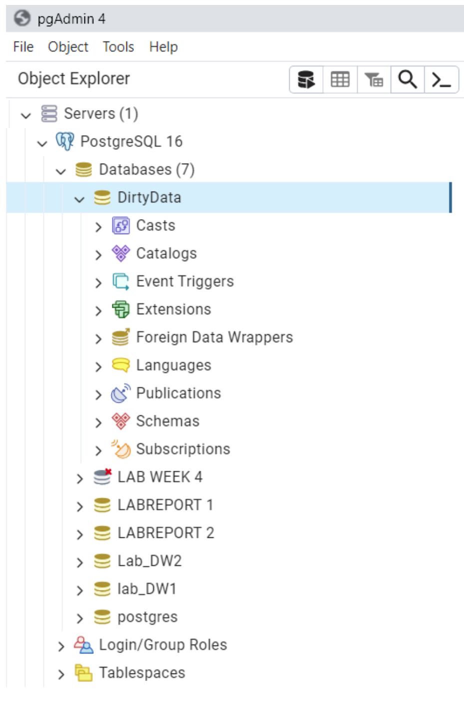


Figure 2.2.1 ETL Pipeline

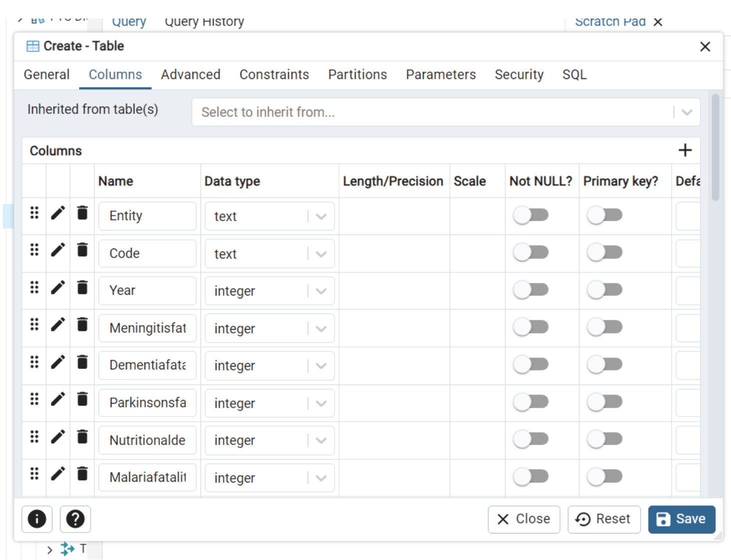
The dataset utilized in this project is earned from Kaggle and Department of Data Malaysia (DOSM). The dataset is composed of six tables which are malnutrition, risk factor, global hunger index, food price, climate change and crop consumption. According to the data processing pipeline, the datasets are initially imported into a PostgreSQL data warehouse. The process of data cleaning and data integrating is being done by using Jupyter Notebook. Jupyter Notebook is employed to extract the tables from the PostgreSQL data warehouse. This extraction process enables the subsequent steps to be executed on the extracted data. Subsequently, the data undergoes thorough cleaning and transformation, ensuring its quality and consistency. After the data has been cleaned and organized, it is transformed into a new clean CSV file format. This new clean CSV file is then uploaded into the PostgreSQL database, making it available for further analysis in OLAP operation to create visualization using Tableau.

2.2.1 Steps to load the raw data into PostgreSQL

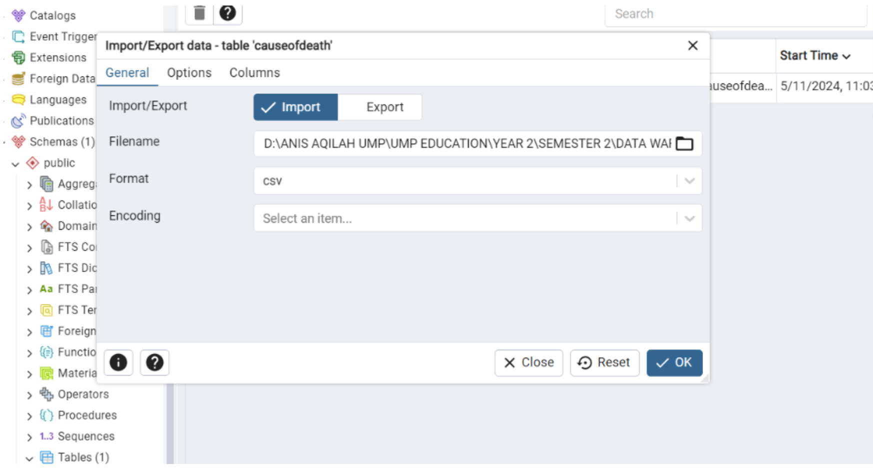
1. Database “DirtyData” is created in PostgreSQL.



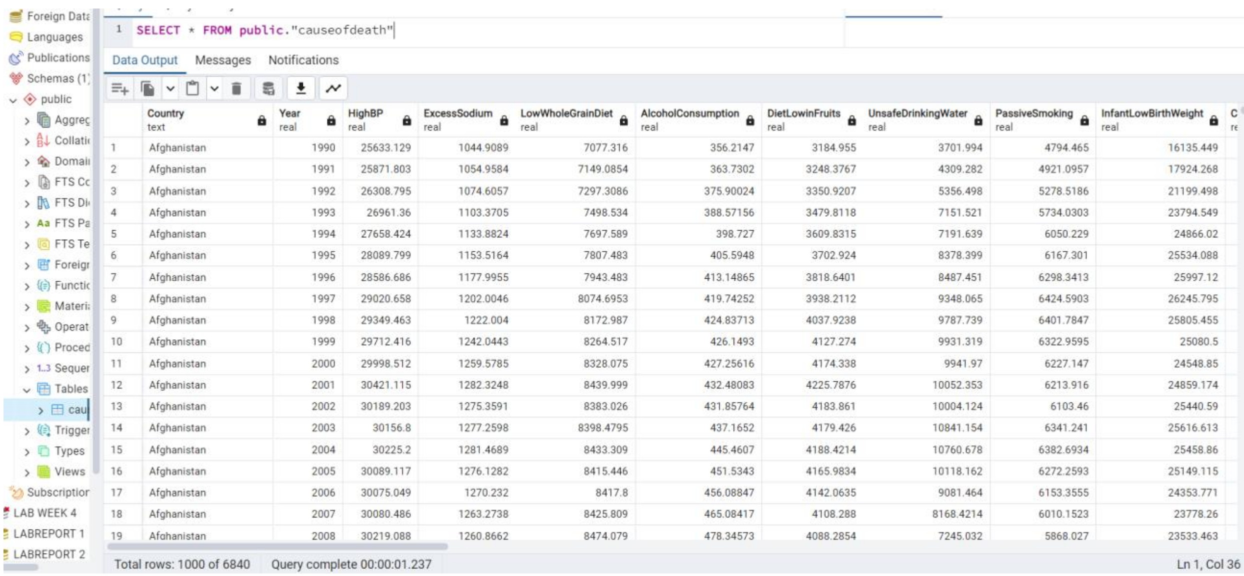
2. Causeofdeath table is created. There are 30 columns in causeofdeath table.



3.Import dataset “causeofdeath” into table causeofdeath.



1. SELECT * FROM public."causeofdeath"



	Country text	Year real	HighBP real	ExcessSodium real	LowWholeGrainDiet real	AlcoholConsumption real	DietLowinFruits real	UnsafeDrinkingWater real	PassiveSmoking real	InfantLowBirthWeight real
1	Afghanistan	1990	25633.129	1044.9089	7077.316	356.2147	3184.955	3701.994	4794.465	16135.449
2	Afghanistan	1991	25871.803	1054.9584	7149.0854	363.7302	3248.3767	4309.282	4921.0957	17924.268
3	Afghanistan	1992	26308.795	1074.6057	7297.3086	375.90024	3350.9207	5356.498	5278.5186	21199.498
4	Afghanistan	1993	26961.36	1103.3705	7498.534	388.57156	3479.8118	7151.521	5734.0303	23794.549
5	Afghanistan	1994	27658.424	1133.8824	7697.589	398.727	3609.8315	7191.639	6050.229	24866.02
6	Afghanistan	1995	28089.799	1153.5164	7807.483	405.5948	3702.924	8378.399	6167.301	25534.088
7	Afghanistan	1996	28586.686	1177.9955	7943.483	413.14865	3818.6401	8487.451	6298.3413	25997.12
8	Afghanistan	1997	29020.658	1202.0046	8074.6953	419.74252	3938.2112	9348.065	6424.5903	26245.795
9	Afghanistan	1998	29349.463	1222.004	8172.987	424.83713	4037.9238	9787.739	6401.7847	25805.455
10	Afghanistan	1999	29712.416	1242.0443	8264.517	426.1493	4127.274	9931.319	6322.9595	25080.5
11	Afghanistan	2000	29998.512	1259.5785	8328.075	427.25616	4174.338	9941.97	6227.147	24548.85
12	Afghanistan	2001	30421.115	1282.3248	8439.999	432.48083	4225.7876	10052.353	6213.916	24859.174
13	Afghanistan	2002	30189.203	1275.3591	8383.026	431.85764	4183.861	10004.124	6103.46	25440.59
14	Afghanistan	2003	30156.8	1277.2598	8398.4795	437.1652	4179.426	10841.154	6341.241	25616.613
15	Afghanistan	2004	30225.2	1281.4689	8433.309	445.4607	4188.4214	10760.678	6382.6934	25458.86
16	Afghanistan	2005	30089.117	1276.1282	8415.446	451.5343	4165.9834	10118.162	6272.2593	25149.115
17	Afghanistan	2006	30075.049	1270.232	8417.8	456.08847	4142.0635	9081.464	6153.3555	24353.771
18	Afghanistan	2007	30080.486	1263.2738	8425.809	465.08417	4108.288	8168.4214	6010.1523	23778.26
19	Afghanistan	2008	30219.088	1260.8662	8474.079	478.34573	4088.2854	7245.032	5868.027	23533.463

Total rows: 1000 of 6840 Query complete 00:00:01.237 Ln 1, Col 36

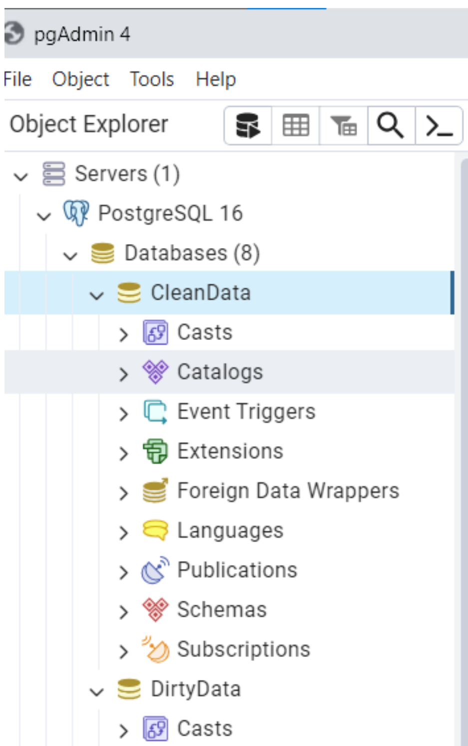
The rest progress for table climatechange, table cropproduction, table globalhungerindex, table malnutrition, and table consumerexpinditure can be check via this link: https://drive.google.com/drive/folders/1M9wtYIz64yjRX5ua0xqAwSOiVgalINFn?usp=sharing

2.2.2 Steps to extract and transform the dataset using Jupyter Notebook PostgreSQL

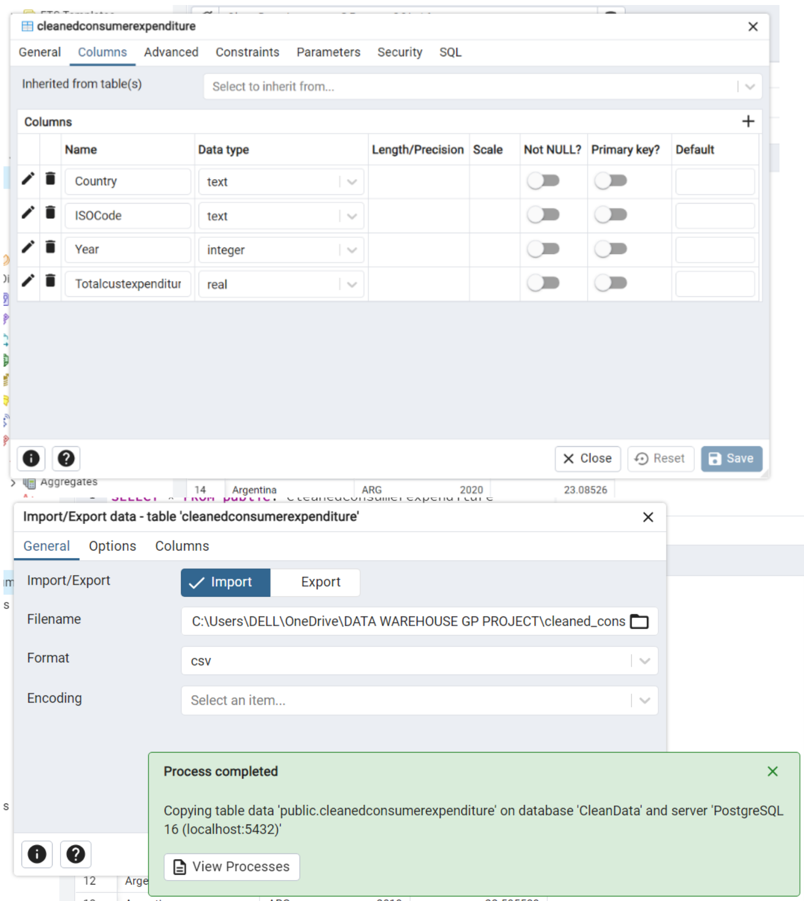
Refer to appendix

2.2.3 Steps to load the cleaned data into PostgreSQL

1. Database 'CleanData' is created in PostgreSQL.

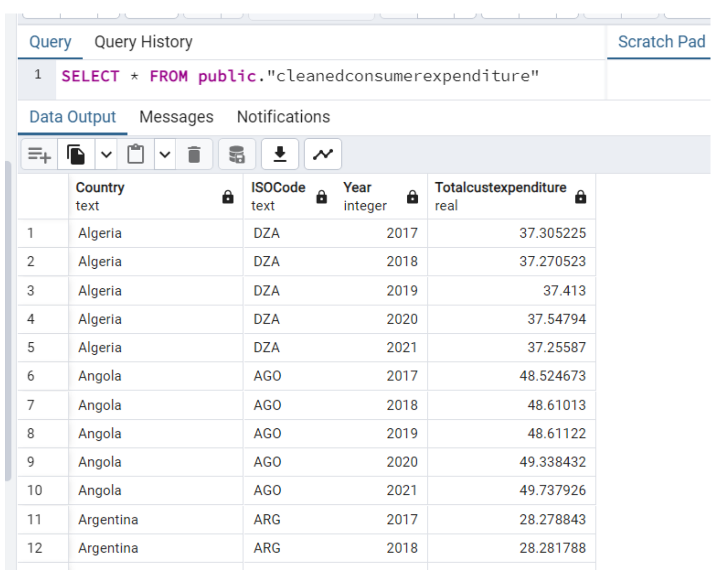


2. Table cleanedconsumerexpenditure is created. There are 4 columns in Table cleanedconsumerexpenditure



3. Import dataset ‘cleaned_consumerexpenditure into table ‘cleanedconsumerexpenditure

4. SELECT * FROM public." cleanedconsumerexpenditure "



The rest progress for table climatechange,table cropproduction, table globalhungerindex, table malnutrition, and table causeofdeath can be check via this link:
https://drive.google.com/drive/folders/1HI0l8js2jp8jailJJXUka_USKDdMYR-y?usp=drive_link

3.0 DATABASE

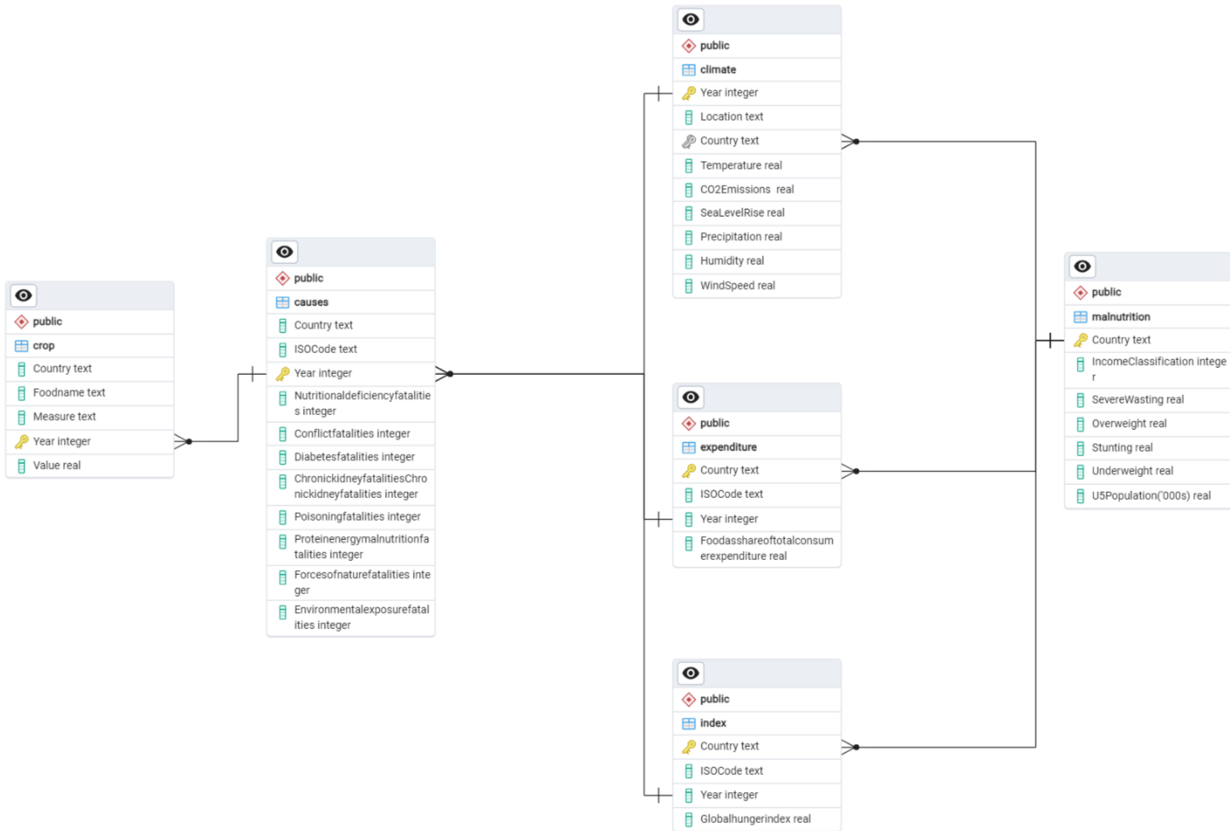


Figure 3.1: Data schema from PostgreSQL

Figure 3.1 shows the relational model of our database by inserting the 6 tables including expenditure, index, crop, climate, malnutrition, and causes. The Entity Relationship Diagram (ERD) clearly shows that our data schema is a Galaxy schema. Galaxy schema is the fundamental schema among the data mart, and it is the most flexible among other schemas.

Galaxy schema is a schema for representing multidimensional models. It is a collection of multiple fact tables that have some same dimension tables. The causes and malnutrition tables will be our fact tables as they are central to the galaxy schema and represent the primary focus of analysis. Index, climate, crop, and expenditure tables are the dimension tables as they support the fact tables by providing additional context.

Based on Figure 3.1, the fact tables in the Galaxy schema have one-to-many relationships with the dimension tables. The relationship between the causes (fact table) and index (dimension table) is one-to-many relationships with the foreign key year in the index table. Same goes to expenditure, climate, and crop are all one-to-many relationships with year is the foreign key. While for malnutrition (fact table) and climate (dimension table), expenditure (dimension table), and index (dimension table) are one-to-many relationship with country is the foreign key.

Overview of database relationship:

Fact Tables:

1. causes: Country, Code, Year (primary key), Meningitis, Alzheimer's disease and other Dementias, Parkinson's disease, Nutritional Deficiencies, Malaria, Drowning, Interpersonal Violence, HIV/AIDS, Drug use Disorder, Tuberculosis, Cardiovascular Disease, Lower Respiratory Infections, Neonatal Disorders, Self-harm, Exposure to Forces of Nature, Diarrheal Disease, Environmental Heat and Cold Exposure, Neoplasms, Conflict and Terrorism, Diabetes Mellitus, Chronic Kidney Disease, Poisonings, Protein-energy Malnutrition, Road Injuries, Chronic Respiratory Diseases, Cirrhosis and other Chronic Liver Diseases, Digestive Diseases, Fire, Heat, and Hot Substance, and lastly Acute Hepatitis.
2. malnutrition: Country (primary key), Income Classification, Severe Wasting, Wasting, Overweight, Stunting, Underweight, and U5 Population ('000s).

Dimension Tables:

1. climate: Year (foreign key), Location, Country (foreign key), Temperature, CO2 Emissions, Sea level Rise, Precipitation, Humidity and Wind Speed.
2. expenditure: Country (foreign key), Code, Year (foreign key) and Food as share of total consumer expenditure.
3. crop: Country (foreign key), Subject, Measure, Year (foreign key), and Value.
4. index: Country (foreign key), Code, Year (foreign key), Global Hunger Index (2021)

4.0 RESULT AND DATA ANALYSIS

4.1 OLAP OPERATION

4.1.1 SLICING

```
SELECT DISTINCT "expenditure"."Country", "index"."Globalhungerindex"
FROM "expenditure"
JOIN "index" ON "expenditure"."ISOCode" = "index"."ISOCode"
WHERE "index"."Globalhungerindex" > 50
GROUP BY "expenditure"."Country", "index"."Globalhungerindex"
ORDER BY "expenditure"."Country" ASC;
```

OUTPUT:

	Country text	Globalhungerindex double precision
1	Angola	65
2	Ethiopia	53.5

Figure 4A: Output Slicing

The output of slicing operation is shown in Figure 4A. The Slice operations perform a selection on one dimension of the given cube, thus resulting in a subcube. Two tables involve in the operation which are expenditure and index. We perform slicing operation by selecting some attributes through join of two tables and conditions using <WHERE>, <GROUP BY>, and <ORDER BY> with 2 rows of result table. Through Slicing operation, we can filter the GlobalHungerIndex which exceed 50 from all the countries. In the operation, we can observe that Angola country and Ethiopia country have exceed 50 of global hunger index. The GHI more than 50 refers to extremely alarming, thus, the countries need to concentrate more on stunting cases among their citizens. Angola encountered rainfall shortages which lead in reducing agricultural production whereas Ethiopia faced extreme weather which is the main driver of food security. Malaysia in not in the list as it is not extremely severe on stunting cases, but actions need to be taken to achieve and maintain zero hunger.

4.1.2 DICING

```
SELECT DISTINCT "index"."Year", "index"."Country", "malnutrition"."Stunting", "index"."Globalhungerindex"

FROM public."malnutrition"

JOIN public."index" ON public."malnutrition"."Country" = public."index"."Country"

GROUP BY "index"."Year", "malnutrition"."Stunting", "index"."Globalhungerindex", "index"."Country"

HAVING "index"."Year" = 2021 AND "malnutrition"."Stunting" > 50

ORDER BY "malnutrition"."Stunting" ASC;
```

OUTPUT:

	Year integer	Country text	Stunting real	Globalhungerindex double precision
1	2021	Yemen	51.9	45.1
2	2021	Nepal	51.9125	19.1
3	2021	Madagascar	52.575	36.3
4	2021	Guatemala	54.1	19.6
5	2021	Bangladesh	54.229168	19.1
6	2021	Burundi	57.6	42.5

Figure 4B: Output Dicing

The output of dicing operation is shown in Figure 4B. This operation describes a subcube by operation a selection on two or more dimensions. Two tables are involved in the operation which are index and malnutrition. We perform dicing operation by selecting some attributes through joining two tables and conditions using <HAVING>, <GROUP BY>, and <ORDER BY> with 6 rows of result table. The main difference between Slicing and Dicing is our condition apply logical operator such as <AND> to filter more deeply information from various dimensions. Through Dicing operation, we filtered the Year which is equal to 2021 and Stunting is greater than 50. From the results, we can see that six countries have stunting more than 50 in 2021 which are Yemen, Nepal, Madagascar, Guatemala, Bangladesh, and Burundi. This happened due to various factors such as poor longer-term nutrition in early childhood, malnutrition, and unhealthy environment.

4.1.3 ROLL UP

SELECT

CASE

WHEN "Year" **IS NULL THEN** 0

ELSE "Year"

END AS "Year", "Country",

SUM("causes"."Nutritionaldeficiencyfatalities") **AS** "TotalNutritionaldeficiencyfatalities", **SUM**("causes"."Proteinenergymalnutritionfatalities") **AS** "TotalProteinenergymalnutritionfatalities"

FROM public."causes"

GROUP BY ROLLUP ("Country", "Year")

ORDER BY "Country", "Year";

OUTPUT:

	Year integer	Country text	TotalNutritionaldeficiencyfatalities bigint	TotalProteinenergymalnutritionfatalities bigint
1	0	Afghanistan	285812	538704
2	1990	Afghanistan	8348	14836
3	1991	Afghanistan	8612	14896
4	1992	Afghanistan	9764	15104
5	1993	Afghanistan	11348	15448
6	1994	Afghanistan	12324	15728
7	1995	Afghanistan	12524	15896
8	1996	Afghanistan	12700	16164
9	1997	Afghanistan	13000	16376
10	1998	Afghanistan	12772	16480
11	1999	Afghanistan	12460	16600
12	2000	Afghanistan	12240	16728
13	2001	Afghanistan	11892	17004
14	2002	Afghanistan	11180	16956
15	2003	Afghanistan	12156	17344
16	2004	Afghanistan	12132	17652
17	2005	Afghanistan	11516	17724
18	2006	Afghanistan	10908	17844
19	2007	Afghanistan	9952	17960
20	2008	Afghanistan	9108	18136
21	2009	Afghanistan	8160	18388
22	2010	Afghanistan	7896	18772

Total rows: 1000 of 7069 Query complete 00:00:00.096

Figure 4C: Output Roll-Up

The output of Roll-Up operation is shown in Figure 4C. The Roll-Up operations is used to provide abstract level details to the user. It performs further aggregation of data by reduction in dimension or by stepping up a concept hierarchy for a dimension. Only one table is involved in the operation which is ‘causes’. There are 7069 rows in the result of the operation. We perform Roll-Up operation by using SUM("causes"."Nutritionaldeficiencyfatalities") and SUM ("causes". "Proteinenergymalnutritionfatalities") to abstract level on these two variations. The ‘0’ in the results refers to the total of the two variations. As the year is in integer form, thus we represent ‘0’ and total. In the operation, we can observe the sum of nutritional deficiency fatalities and protein energy malnutrition fatalities of various countries.

4.2 VISUALIZATION

4.2.1 Sum of Nutritional Deficiency and Sum of Protein Energy Malnutrition by Year in Malaysia

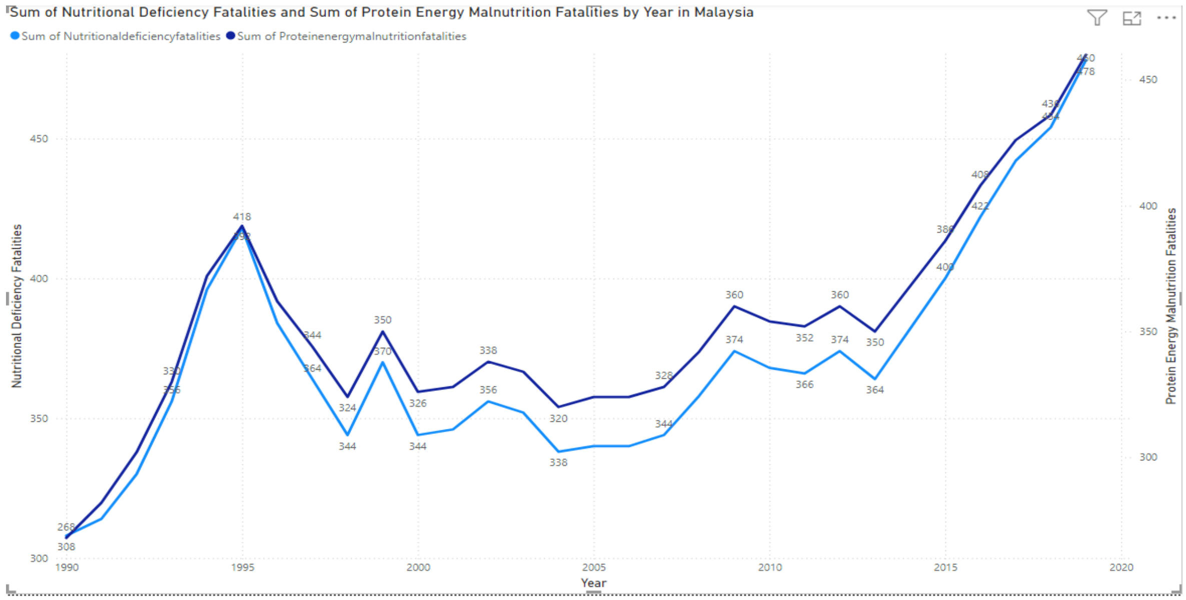


Figure 4.2.1: Sum of Nutritional Deficiency and Sum of Protein Energy Malnutrition by Year in Malaysia

The line on nutritional deficiency and protein energy malnutrition fatalities in Malaysia from 1990 to 2020 provides critical insights into the growing public health issue of malnutrition in the country. In the past 30 years, both types of deaths have increased. The increase started in 2015. By 2020, fatalities due to nutritional deficiencies reached a peak of 480, and those due to protein energy malnutrition peaked at 478, the highest levels recorded in the dataset. The increase in deaths may be due to poor nutrition or better ways of reporting deaths in the past few years. The strong connection between the two types of fatalities suggests that they likely share common risk factors, such as poverty, food insecurity, inadequate healthcare, and lack of access to diverse and nutritious foods.

These trends show how important public health initiatives are needed to prevent malnutrition in Malaysia. It is important to find and address the risk factors that are causing these increasing deaths. Providing affordable food, teaching people about healthy eating, and working together with local groups to address nutritional deficiencies are ways to make food more secure.

Also, hospitals can be improved to catch and treat malnutrition early. Malaysia could reduce the number of deaths linked to nutritional deficiencies and protein energy malnutrition by treating these underlying problems and encouraging balanced meals, eventually improving the general health of the country.

4.2.2 Income Classification, Overweight, Severe Wasting, Stunting, Underweight and Wasting in Malaysia.

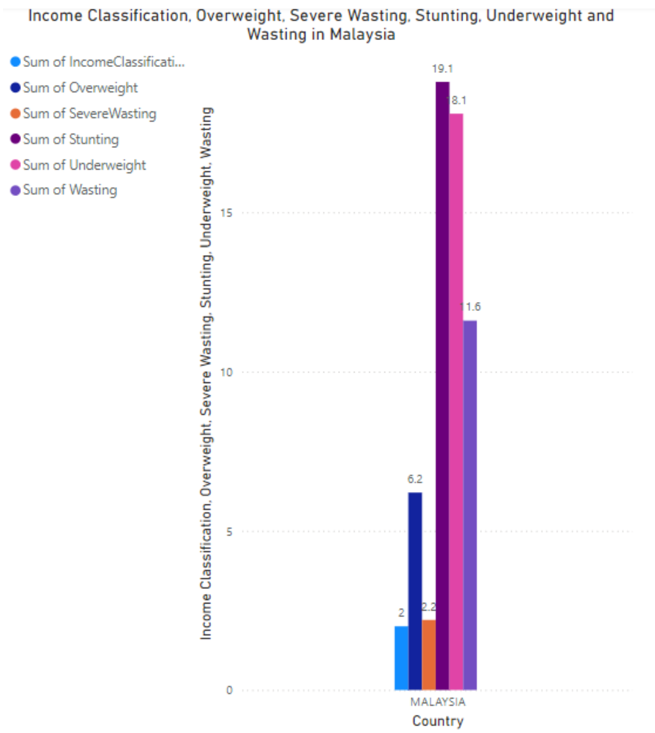


Figure 4.2.2: Income Classification, Overweight, Severe Wasting, Stunting, Underweight and Wasting in Malaysia.

This bar chart insight into Malaysia's income classification shows that malnutrition in the country is complicated and shows important nutritional problems. Along with severe wasting (1.6%) and wasting (2.2%), the large amount of stunting (19.1%) and underweight (8.1%) points to serious malnutrition and insufficient nutritional intake, which mostly affects young people. The proportion of overweight people (6.2%) indicates that overnutrition problems may coexist, suggesting that malnutrition may have a dual impact. This conflict highlights the importance of total diets that target both sides of the spectrum. Promoting balanced meals with important nutrients, raising public awareness of good eating practices, and making sure everyone has access to a variety of healthy meals are important strategies. To address the real causes of malnutrition and lower deaths caused by nutritional deficiencies in Malaysia, it is also important to improve food security and income gaps.

We need to make healthcare better at finding and treating people who are underweight. We should check them regularly and help them when they need it. Collaboration between government agencies, non-government organizations, and international partners can help fight malnutrition. By implementing these policies, Malaysia may try to lower the rate of undernutrition, improve health outcomes, and ensure a healthier future for the population.

4.2.3 Sum of Total Consumer Expenditure on Food in Malaysia by Year

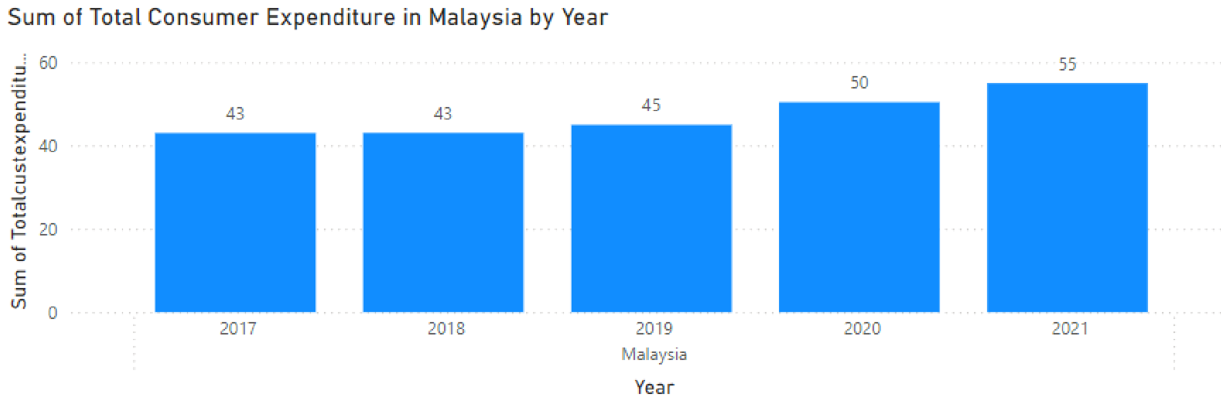


Figure 4.2.3: Sum of Total Consumer Expenditure on Food in Malaysia by Year

This visualisation is a bar chart that shows the sum of total consumer expenditures of food in Malaysia from 2017 to 2021. The x-axis of the graph shows the years while the y-axis shows the amount of consumer expenditure. There is a moderate increasement in consumer expenditure overr the years. In year 2017 to 2018, the total consumer expenditure remain constant at 43. This might be because of Malaysian economy that might experienced stability which indicate there are no significant growth or decline. Then from 2019 to 2021, the graph shows increasement from 45 to 55. This possible reasons for this situation are economic and population growth. Economic growth can happen due to many reason which are increased disposable income, more employment opprtunities and rising of the consumer confidence on both essential and non-essential item. Population growth can happen due to the increase demand for good and services such as food, clothing, and housing.

Increasement in the sum of total customer expenditure can lead to economic growth which will improved food security. The higher the economic growth, the higher the income of a household. So, they can afford to buy more nutritious food. This will reduced hunger and provide better food security among the Malaysians.

4.2.4 Sum of Value (Rice) in Malaysia by Year

Figure 4.2.4: Sum of Value (Rice) in Malaysia by Year

This visualization is an area chart that shows the sum of value of rice production in Malaysia from year 2000 to 2023. The x-axis of the graph shows the years while the y-axis shows the sum of value of rice production in Malaysia. The sum value of rice production shows a trend of increasing but with a decline phase around 2014 and 2015 followed by an increment after 2017. In 2013, it shows a significant increase from 2351.17 to the highest point which is 2873.30 in 2015. This is because of the increase in demand due to economic growth and higher crop production. Moreover, the export demand might be the reason why the crop production demand increases. In 2016 to 2017, there is an obvious decline of the values which might happen because of high demand of import rice in Malaysia and causes to lower the values of rice production. Other than that, there might be natural disaster happening that might cause the decreasing value of crop production. But it gradually increases back in 2017 until 2023 because of the balance of supply and demand.

It is important to underline this value of rice in Malaysia to understand the impacts of crop production trends on food security and agriculture sustainability in Malaysia. By understanding these trends, the responsible party can take action that can ensure the stable and enough nutrition for all population to achieve zero hunger in Malaysia.

5.0 CONCLUSION

In conclusion, the datasets chosen were all different but connected to each other and used to overcome zero hunger. We loaded the data into a database as the first step before performing further process in this project. ETL process is carried out to extract data, transform the data into suitable format and load it into data warehouse. Furthermore, the cleaning process was applied to avoid any duplicate, missing data and to standardize it into desired format so that it would be easier to perform further processes such as OLAP process. For the process, we utilized Python software, which is Jupyter and Power BI while in OLAP process, we utilized PgAdmin to implement all the operations like dicing and slicing and roll-up.

Data pertaining to consumer spending, rice production, nutritional deficiencies, and deaths resulting from hunger highlight a multifaceted issue. The rise in deaths attributed to malnutrition could indicate either an improvement in reporting or an ongoing decline in nutritional standards. There is a twice burden of malnutrition as evidenced from high incidence of stunting and underweight children living alongside overweight people. Due to technology and economic growth, consumer spending has been rising gradually, which could improve food security. The price of rice production has varied due to changes in demand and disasters. To address these issues and achieve zero hunger in Malaysia, balanced diets, increased food security, enhanced healthcare, and sustainable agriculture methods need to be implemented to guarantee better nutrition and health outcomes.

Each group member plays a big role whether it is from an idea's scratch until the end of this project. We faced various difficulties from each part, and we chose six datasets that are most related to our project to complete and obtain meaningful insights. We hope that this data will be beneficial for the government and public health institutions which leads to improved population health and overcome zero hunger.

6.0 REFERENCES

SDG 2 : ZERO HUNGER. (n.d.). https://sdg-for-malaysian-states-sdsn.hub.arcgis.com/pages/sdg-2-zero-hunger

Martin. (2023, October 19). *Goal 2: Zero hunger - United Nations Sustainable Development*. United Nations Sustainable Development. https://www.un.org/sustainabledevelopment/hunger/

Topic: Health in Malaysia. (2023, December 21). Statista. https://www.statista.com/topics/5858/health-in-malaysia/#topicOverview

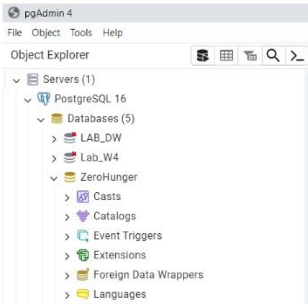
Goal 2 - Zero Hunger - dataset by sdgvizproject. (2023, June 2). data.world. https://data.world/sdgvizproject/goal-2-zero-hunger

APPENDIXES

Steps to extract and transform the dataset using Jupyter Notebook PostgreSQL

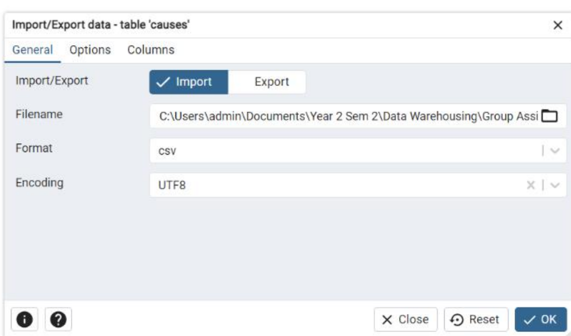
Steps to load clean data into database (PostgreSQL) to do OLAP Operation

1. Created databased 'ZeroHunger' PostgresSQL



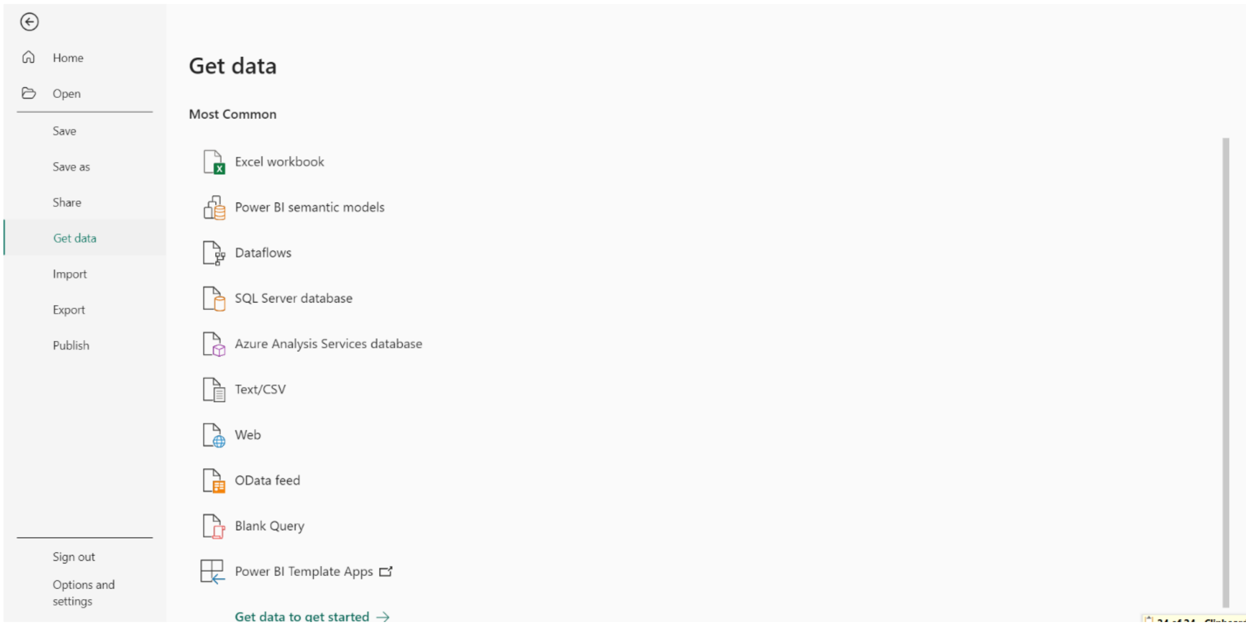
2. Created all of the six tables 'causes', 'climate', 'crop', 'expenditure', 'index', and 'malnutrition'.

3. Import dataset 'causes', 'climate', 'crop', 'expenditure', 'index', and 'malnutrition'.

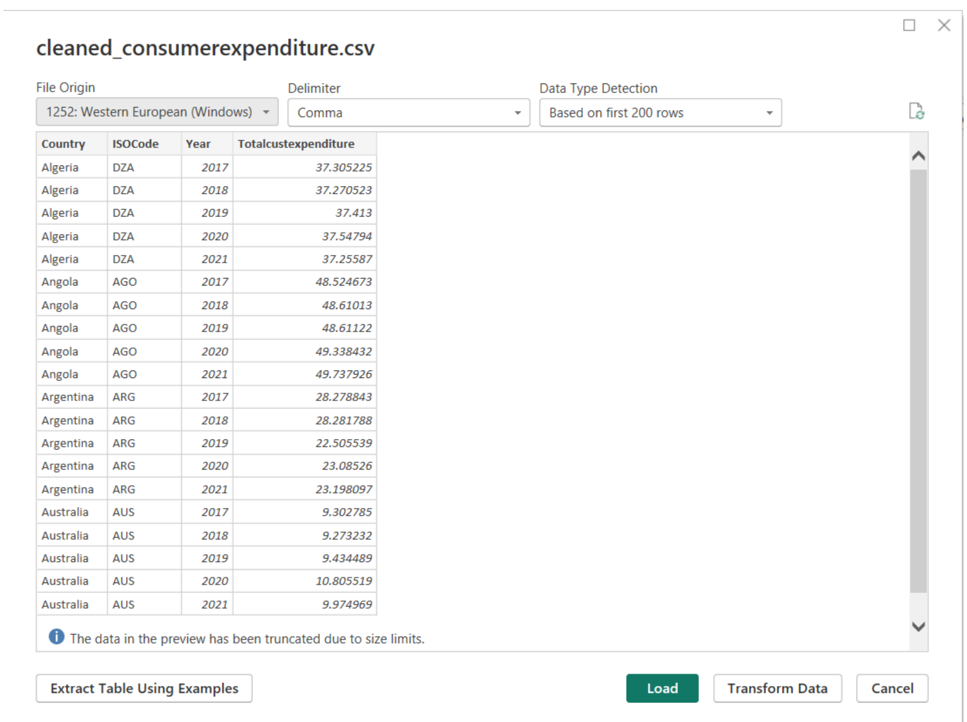


Steps to load dataset into Power BI

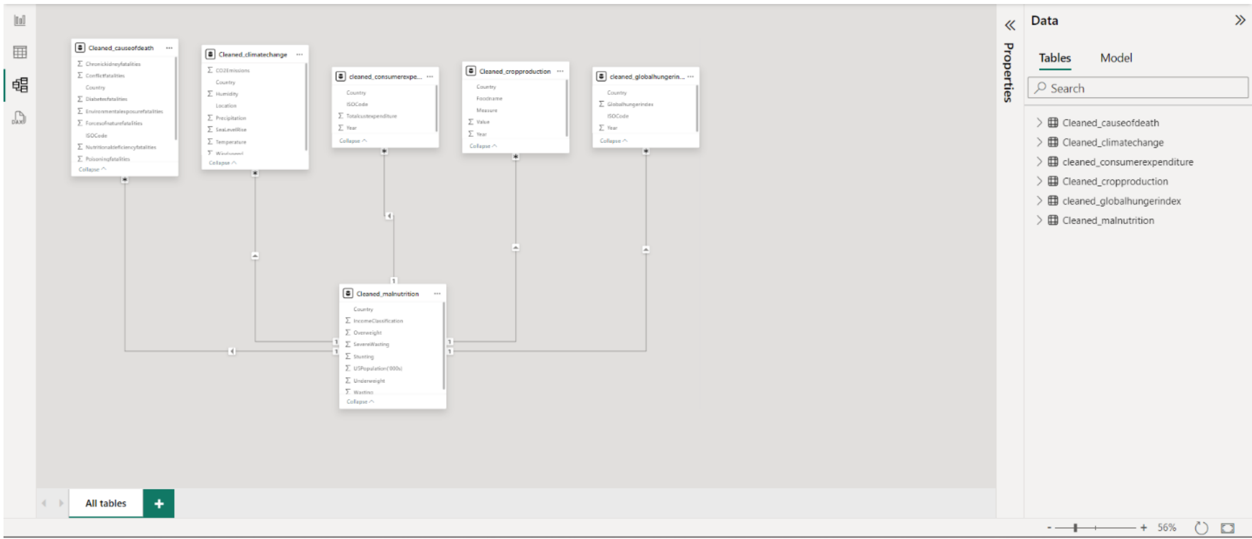
1) Open the Power BI and import the cleaned data.



2) After we choose the dataset, click the load.



3) After we load the data, we can continue to do the visualization.


Cleaned_causeofdeath
Cleaned_climatechange
Cleaned_cropproduction
Cleaned_malnutrition
Properties
Data
Tables
Model
Search
Cleaned_causeofdeath
Cleaned_climatechange
cleaned_consumerexpenditure
Cleaned_cropproduction
cleaned_globalhungerindex
Cleaned_malnutrition
All tables
56%