

1. **Introduction**

According to Kraak et al. (2018), The global population is growing at an incomprehensible rate and with it come complex environmental consequences that often result in social injustice. The United Nations has established a set of Sustainable Development Goals (SDGs) in an attempt to ameliorate inequality and promise safety for the masses. To reach these goals, a set of indicators have been identified and their associated data for each country are publicly available to measure how close each country is to each goal.

1. **Statement of the Problem/ Objectives**

B.1 **Problem Statement:**

The Philippines faces socio-economic challenges, particularly related to poverty, which is a critical concern in the context of the United Nations Sustainable Development Goal (SDG) 1 - "No Poverty." To address this issue effectively, there is a need to collect, store, and present socioeconomic data. Current data management approaches may lack the necessary structure and data-driven capabilities to support this goal.

B.2 **Objective:**

The primary objective of this project is to design and implement a comprehensive database schema tailored to the specific socioeconomic needs of the Philippine’s Region. This schema includes dimension (Dim) and fact (Fact) tables that closely align with the objectives of SDG 1, focusing on the eradication of poverty in all its forms. The proposed database pipeline aims to:

Collect and store data by creating a structured system capable of efficiently collecting, storing, and managing socio-economic data from various sources, with each dataset corresponding to a specific year and region.

Enables the systematic monitoring of poverty reduction progress, a fundamental aspect of SDG 1, by providing a framework for organizing and presenting data related to critical factors, including poverty rates, income, employment, and population.

In summary, the project's objective is to develop a robust database schema that supports the goals of Socio-Economic factor relating to SDG 1, by providing a structured and data-driven approach to address and monitor issues related to poverty within the Philippines.

C. **Solution**

Establishing a simple yet effective data pipeline for collecting and storing socioeconomic data is crucial, as illustrated in Figure 1 on page 3. We will divide the process into two phases. Phase I shown in Image 1 below, involves the extraction of data from the source webpage/site, followed by pre-processing, which will be done in Excel. Another transformation process to conforms to the structure of the Normalized Model will be carried out using Python, and finally, the transformed data will be loaded into the database using Python and SQL.

. For Phase II, as illustrated in Image 2 on page 3, the process will entail extracting data from the database, followed by transforming the schema based on the dimension and fact tables, then the data will be loaded into the data warehouse, and lastly the data warehouse will be connected to Excel via ODBC direct query connection for data visualization.

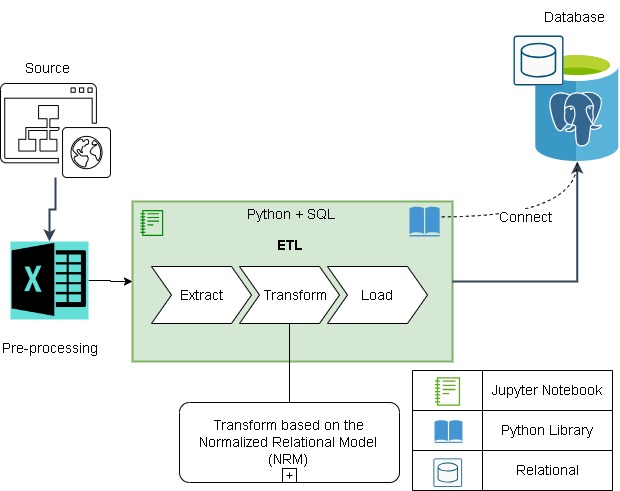
****

Image 1. PHASE I (Data pipeline)

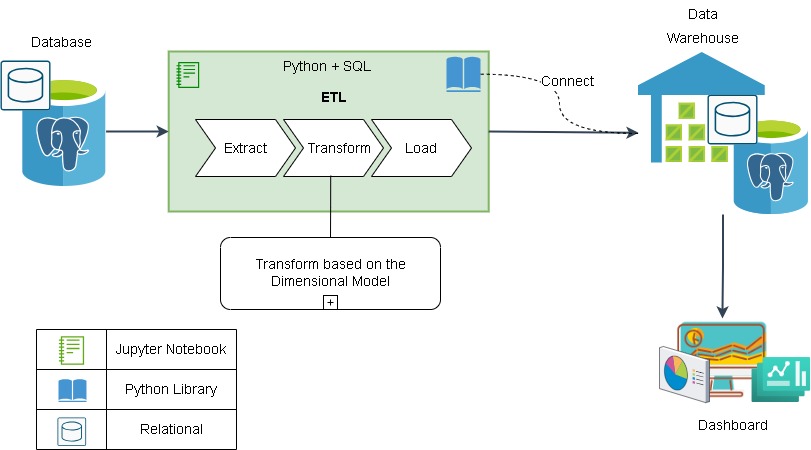
****

Image 2. PHASE II (Data pipeline)

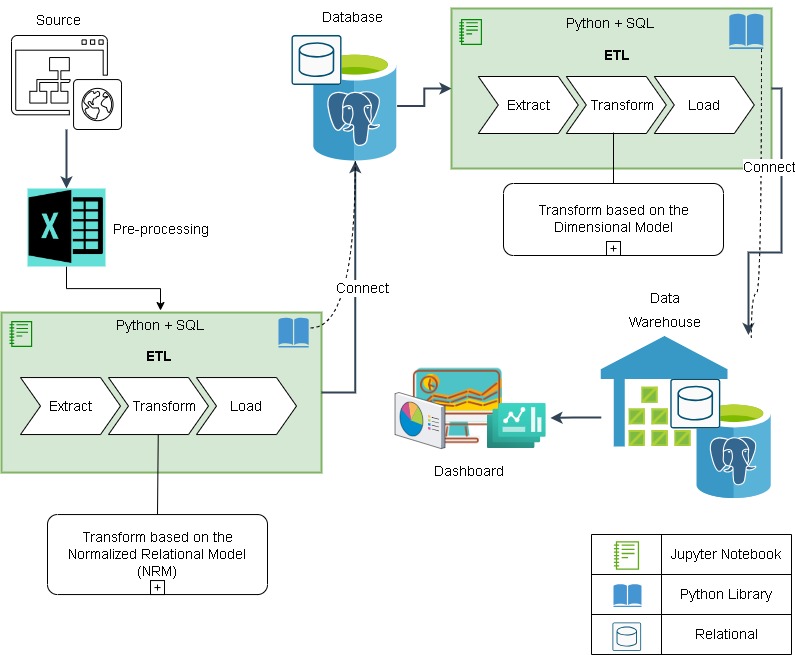


Figure 1. Data pipeline (Phase I & II)

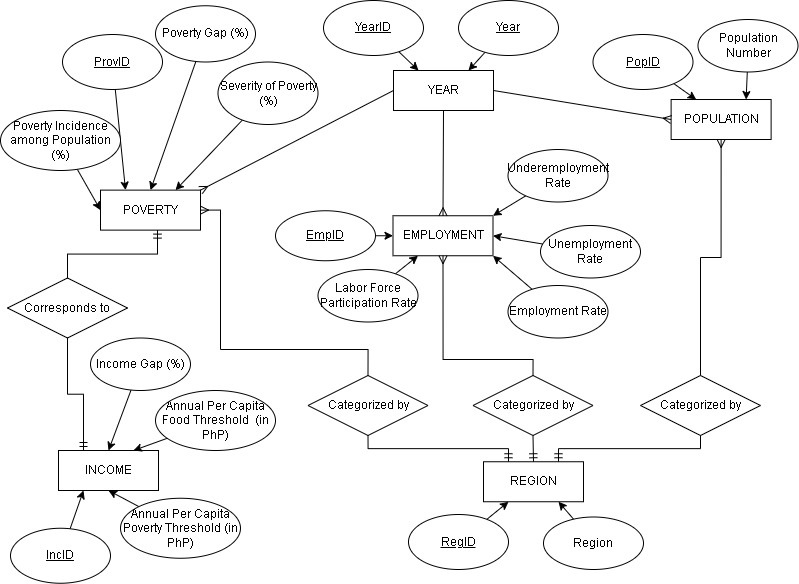


Figure 2. Entity-Relational Diagram

**Definition of Attributes**

**Population Number**

The number of persons in a specific region including all individuals, regardless of their age.

**Annual Per Capita Food Threshold**

Refers to a specific level or amount of expenditure or consumption of food that is considered the minimum necessary to meet basic nutritional needs for an individual or a household over the course of a year. "Annual Per Capita" means that this threshold is calculated on a per-person basis and is applicable for a year. When a person or household's food consumption or expenditure falls below this threshold, it indicates a risk of food insecurity or poverty.

**Annual Per Capita Poverty Threshold (in PhP)**

This indicator represents the income or consumption level (measured in Philippine Pesos, PhP) below which a person or household is considered to be living in poverty. It takes into account not only the cost of food but also other basic necessities such as housing, clothing, education, healthcare, and transportation. Similar to the food threshold, it is calculated on a per-person basis, considering the size of the household.

**Poverty Incidence among Population (%)**

A statistical measure that represents the percentage of a population living in poverty within a specific region. This indicator is a fundamental measure for assessing the extent and prevalence of poverty in a given geographic area.

**Income Gap**

The disparity or difference in income levels among individuals or households within a population. It is often measured using metrics such as the Gini coefficient or other income distribution measures. A higher income gap indicates greater income inequality, while a lower gap suggests more equal distribution of income.

**Poverty Gap**

This will measure the depth or severity of poverty within a population. It represents the average income shortfall of people living in poverty compared to the poverty line or threshold. In other words, it quantifies how far below the poverty line the average poor person's income falls. A higher poverty gap indicates a larger income shortfall among those in poverty.

**Severity of Poverty**

Assesses the intensity or severity of deprivation experienced by people living in poverty. It considers not only the incidence (percentage of people in poverty) but also the depth of poverty. It provides insight into how poor individuals or households are relative to the poverty line. A higher severity of poverty indicates that people living in poverty are experiencing more significant deprivation.

**Labor Force Participation Rate**

Measures the proportion of the working-age population (typically individuals aged 15-64). by calculating the percentage of people who are either employed or actively looking for work relative to the total population within the specified age range. The labor force participation rate also provides insights into the willingness and ability of the working-age population to engage in the labor market.

**Employment Rate (Employment-to-Population Ratio)**

Measures the proportion of the working-age population that is employed/has a job. It calculates the percentage of people who are currently employed relative to the total working-age population. The employment rate is a measure of the extent to which the working-age population is participating in gainful employment.

**Unemployment Rate**

Calculates the percentage of unemployed individuals relative to the total labor work force. The unemployment rate provides insight into the level of joblessness within the labor force, it is also a key indicator of economic health and labor market conditions.

**Underemployment Rate (or Inadequate Employment Rate)**

The underemployment rate measures the proportion of employed individuals who are working in jobs that do not fully utilize their skills and qualifications or who are working part-time involuntarily (i.e., they want full-time work but can't find it). It calculates the percentage of underemployed individuals relative to the total employed population. The underemployment rate highlights the extent to which individuals are not fully employed or are in jobs that do not match their qualifications.

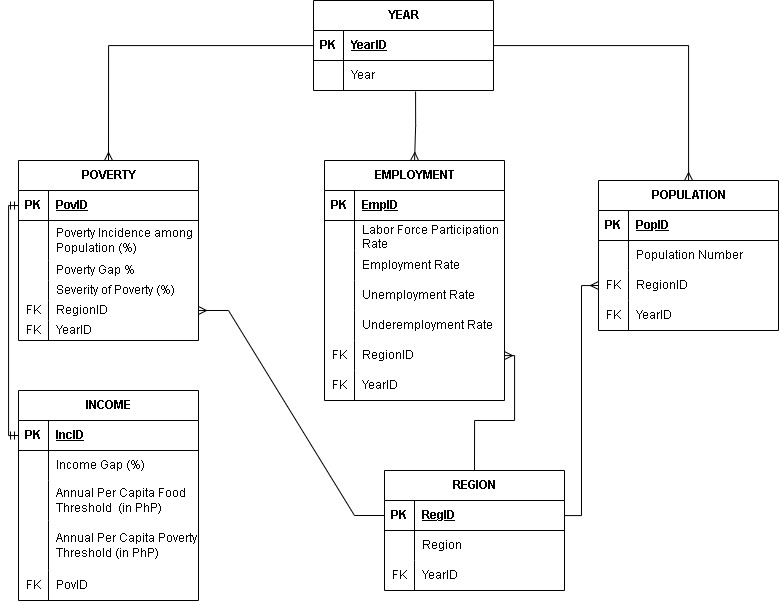


Figure 3. Normalized Model

Based on the ERD (Entity-Relational Diagram) in Figure 2 on page 4, we can now derive the NM (Normalized Model) for the structure of the database, named “povdb”. as show in Figure 3 on page 6.

Relationship:

One to One (1 : 1)

* Income to Poverty

One to Many (1 : N)

* Year to Poverty
* Year to Employment
* Year to Population
* Region to Poverty
* Region to Employment
* Region to Population

Example Table: (just to visualized the table format, not the actual data presented)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **Region** | **Poverty score** | **Employment rate** | **Population No. (M)** |
| 2015 | Region I | 12 | 92 | 5.1 |
| 2015 | Region II | 9 | 95 | 3.6 |
| 2015 | Region III | 10 | 95 | 12.4 |
| 2015 | Region IV | 13 | 93 | 16.1 |
| 2015 | Region V | 12 | 96 | 7.9 |
| 2015 | Region VI | 13 | 93 | 16.1 |
| 2019 | Region VII | 13 | 93 | 16.1 |
| 2019 | Region I | 13 | 93 | 16.1 |
| 2019 | Region II | 11 | 92 | 6 |
| 2019 | Region III | 12 | 96 | 7.9 |
| 2019 | Region IV | 9 | 95 | 3.6 |
| 2019 | Region V | 10 | 95 | 12.4 |
| 2019 | Region VI | 13 | 93 | 16.1 |
| 2019 | Region VII | 13 | 93 | 16.1 |
| 2021 | Region I | 13 | 93 | 16.1 |
| 2021 | Region II | 12 | 96 | 7.9 |
| 2021 | Region IV | 12 | 96 | 7.9 |
| 2021 | Region V | 9 | 95 | 3.6 |
| 2021 | Region VI | 10 | 95 | 12.4 |
| 2021 | Region VII | 13 | 93 | 16.1 |

The table above is an example of the merge data of year, region, poverty, employment, and population table. The Normalized Model Figure 3 in page 6, is designed in-order to eliminate redundancy and to achieved normalization.

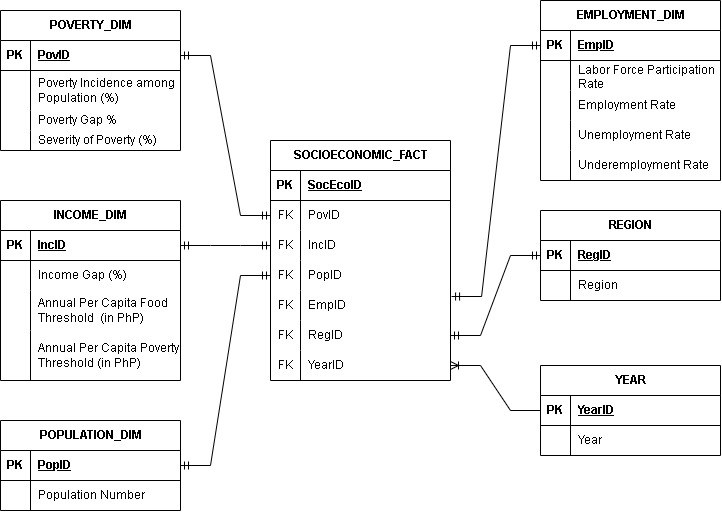
****

Figure 4. Dimension and Fact Table (Star schema)

The dimension and Fact table was derived from the normalized model. It will be use for the structure of the data warehouse, named "povdw," as depicted in Figure 4 above. In this schema, the fact table is referred to as "fact-less" because its primary purpose is to establish relationships between elements from different dimensions and it does not contain aggregated or calculated data.

**ABOUT DATA**

|  |  |  |
| --- | --- | --- |
| Dataset | Year | Frequency of Update (PSA) |
| Population | 2010, 2015, 2020 | Every 5 years |
| Poverty | 2015, 2018, 2021p | Every after 2 years |
| Income | 2015, 2018, 2021p | Every after 2 years |
| Employment | 2015, 2018, 2021p | Every after 2 years |

The frequency of update for the population dataset does not align with the poverty, income and employment data except for year 2015, therefore we can employ a linear interpolation formula on the population data.

Linear Interpolation:

y' = y1 + (x - x1) \* ((y2 - y1) / (x2 - x1))

The formula above will be utilized during transformation staged of population data to get the year 2018 and 2021.

After presenting the data pipeline and establishing the ERD, NM (Normalized Model), and Dimension & Fact Tables, will now proceed to create the database in PostgreSQL based on the Normalized Model structure and create the data warehouse based on the Dimension & Fact table (Star-schema) structure.

**SQL script for povdb based on NM (database)**

|  |
| --- |
|  |
|  |

**SQL script for povdw base on Dimension and Fact Table (data warehouse)**

|  |
| --- |
|  |
|  |
|  |

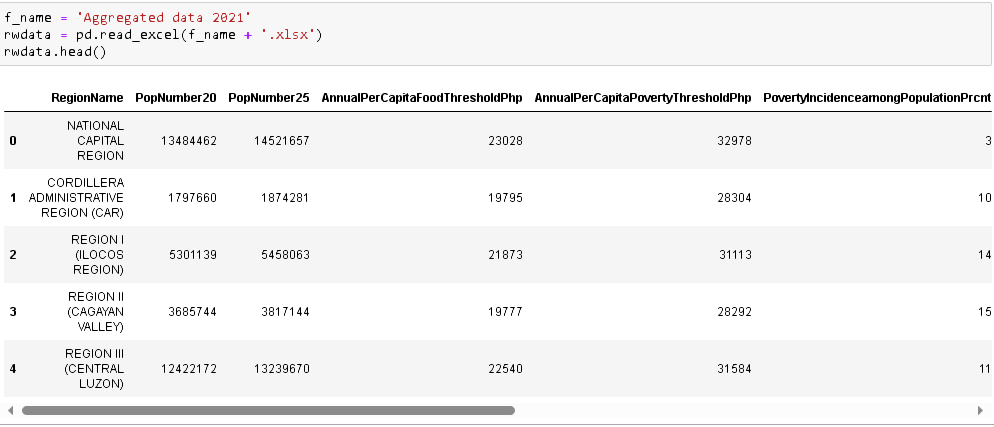
After the creation of the database (povdb) and data warehouse (povdw) in PostgreSQL, the next step is to perform the Phase I of the data pipeline.

**PHASE I**

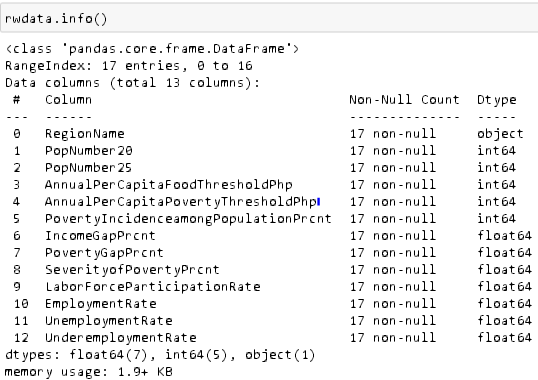
Note: The following screen cap shows the 2021 data undergoing Phase I and II processes.

**I. 1 EXTRACT**

Output of the pre-processing done by aggregating the data using excel.



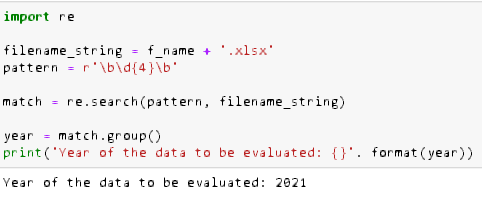
Information of all the columns in the aggregated file.



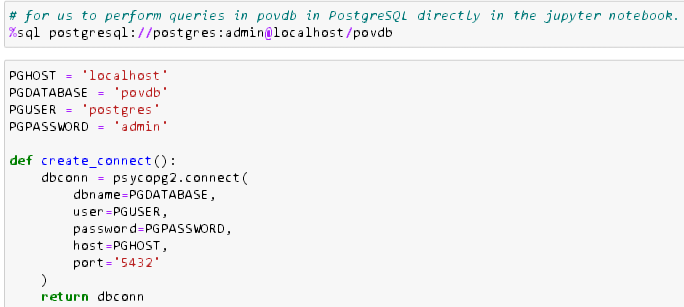
**I. 2 EVALUATION**

In this section of the ETL process, it will check if the data are already on the database or it is a new data to be added.

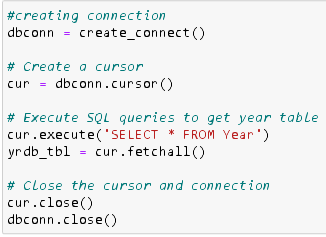
The code below will checked what year is the data to be evaluated.



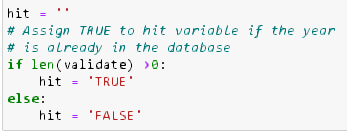
Then connect the jupyter notebook to database povdb in PostgreSQL.

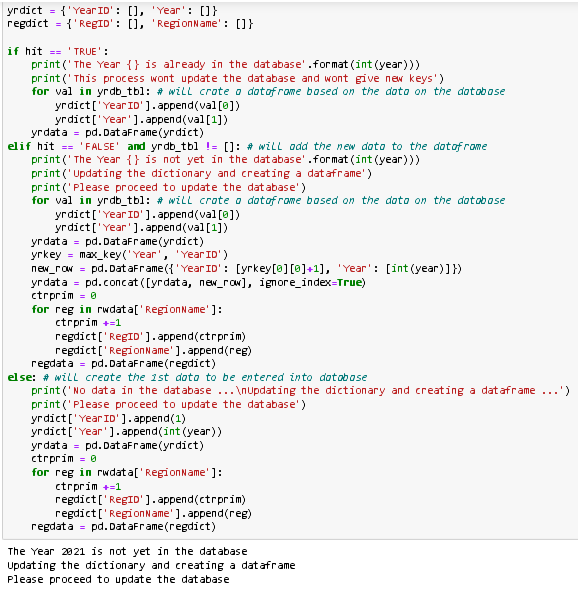


Get the year table in the database povdb to be evaluated



The code below with “hit:” variable will be assigned a “TRUE” value if the 2021 data are already in the database and “FALSE” if otherwise.



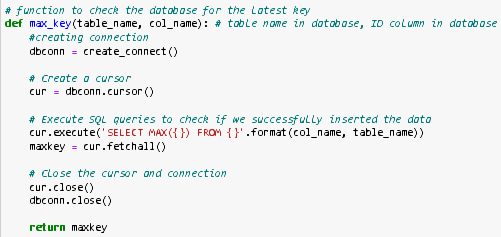


The code above is a simple if statement implemented to determine whether a newly imported file constitutes new data or if its already exists in the database. For instance, if the code has been executed multiple times, the output will indicate, “The Year 2021 is already in the database.” This mechanism serves the dual purpose of ensuring data integrity and preventing duplicate entries in the database by avoiding the assignment of new primary and foreign keys during the transformation process.

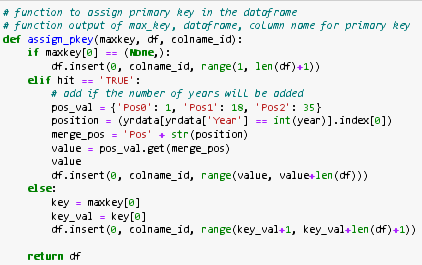
**I. 3 TRANSFORM**

During this stage, the data frame will be partitioned into its respective categories, and primary as well as foreign keys will be assigned. It's important to note that this assignment will only take place if the algorithm that was established during the evaluation phase returns a false outcome and hit variable will have the value of “FALSE”, signifying that the data is not yet present in the database.

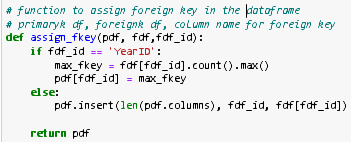
Check for the max key value in the database (function)



Assign pkey (function)



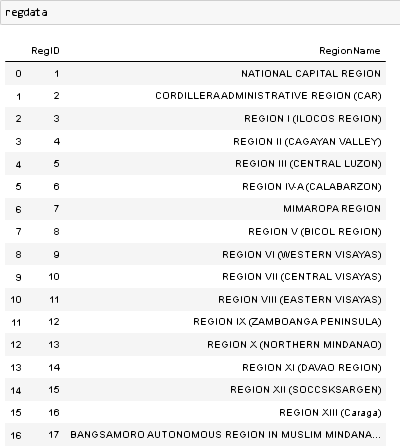
Assign fkey (function)



The max\_key function will determine the most recent primary key in the povdb database. For instance, when it examines the PovID column (the primary key) of the Poverty table for the first time, the max\_key function returns the value of (None,), indicating that no data has been loaded yet. This returned value will be used in the conditional statement within the assign\_pkey function.

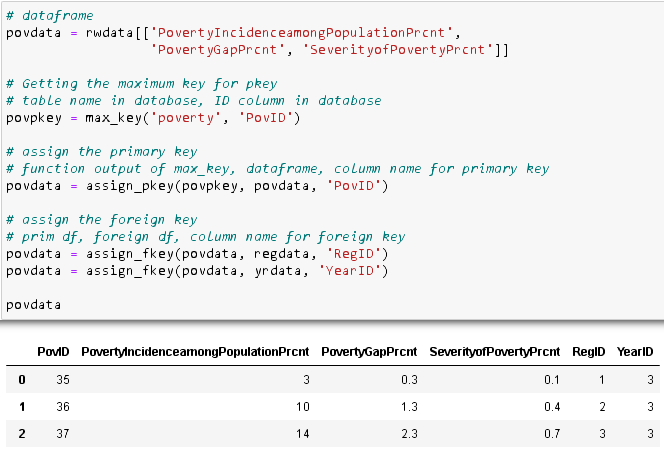
For example, when the max\_key value is 34, the assign\_pkey function proceeds to assign primary key values ranging from 35 to the 'n' rows of data in the 'Poverty' column. This process is also applicable to the assign\_fkey function. However, a minor adjustment is made in the conditional statement due to the relationship between the 'Year' and 'Region' tables versus the 'Poverty', 'Employment', 'Income', and 'Population' tables, as shown on page 7.

**Region**

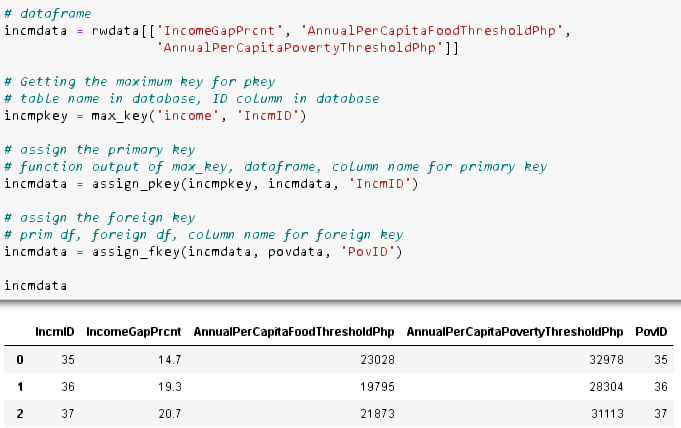


The Region table above contains static values and is classified as a lookup table since its primary purpose is to standardize and categorize data in other tables.

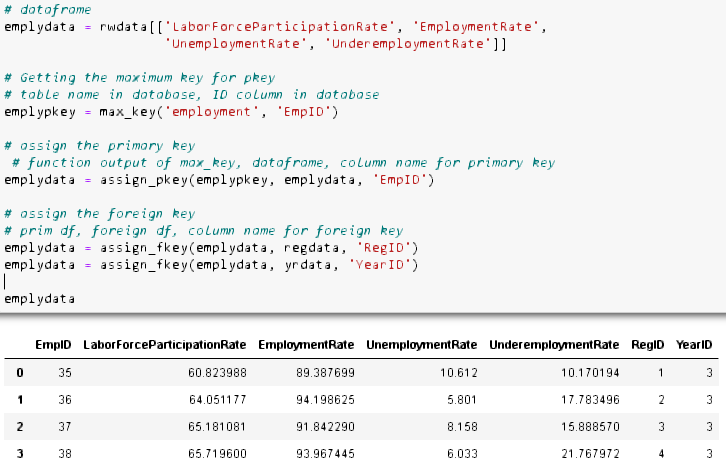
**Poverty**



**Income**

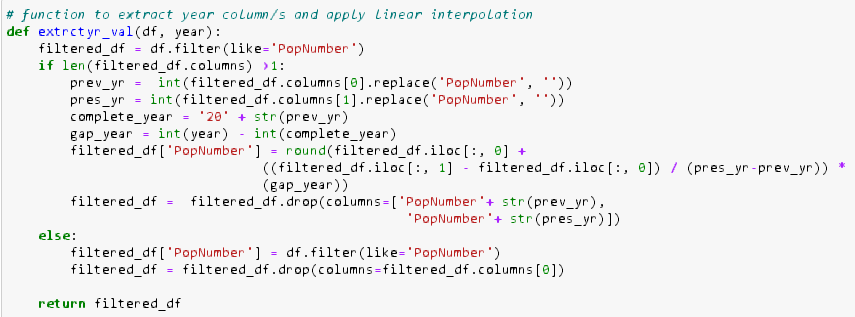


**Employment**



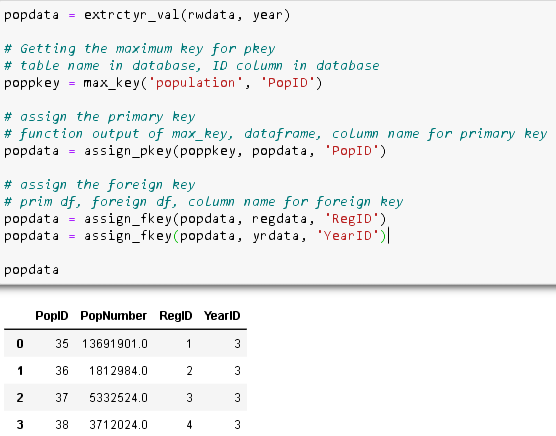
**Population**

This is where the Linear Interpolation formula will be used and place inside a function to align the population dataset with poverty, income, and employment data. The function below will also output non-interpolated data if there are no two columns consisting population data in the file, similar to the ‘Aggregated data 2015’.



For the year 2021 dataset, linear interpolation will be performed since the if statement will be true as there are two columns with population data in the file, namely 'PopNumber20' and 'PopNumber25 as shown in page 12.

**Population (Interpolated)**



**I.4 LOAD**

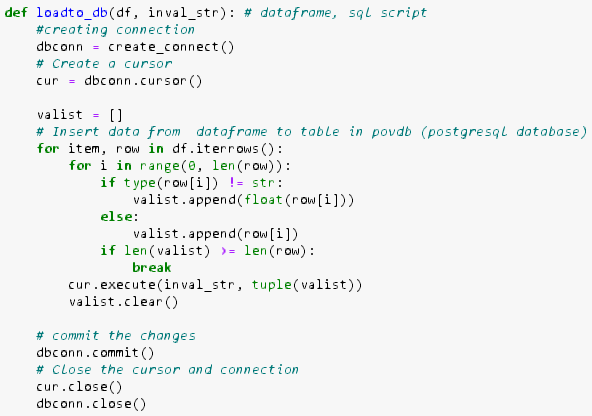
Load the dataset into povdb database after the transformation process.

Create SQL script for loading data (function)

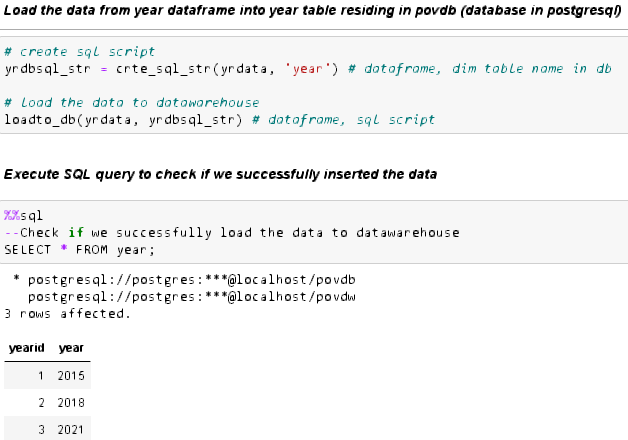


The crte\_sql\_str function above will output a string of SQL syntax, this syntax will be use in the loadto\_db function to load the data into povdb database.

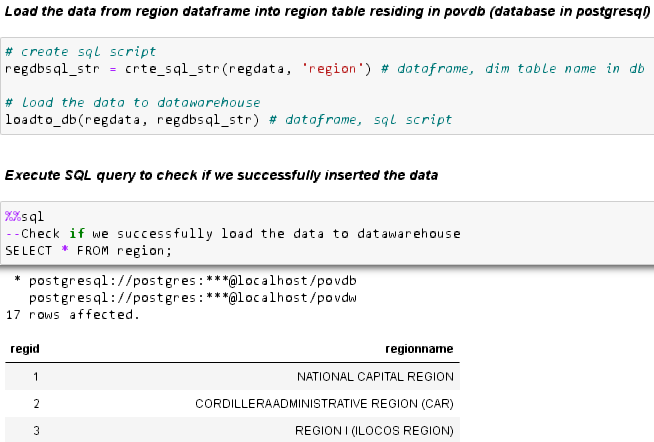
Load data to database (function)



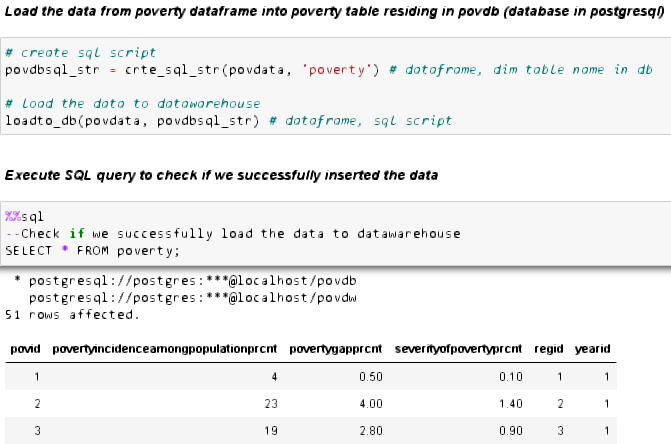
**Year**



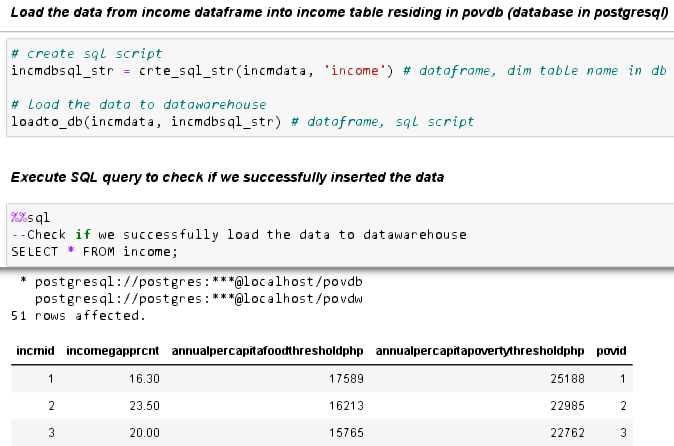
**Region**



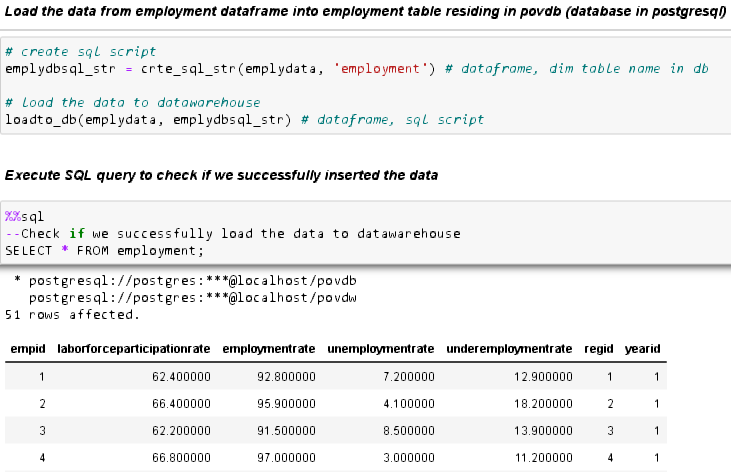
**Poverty**



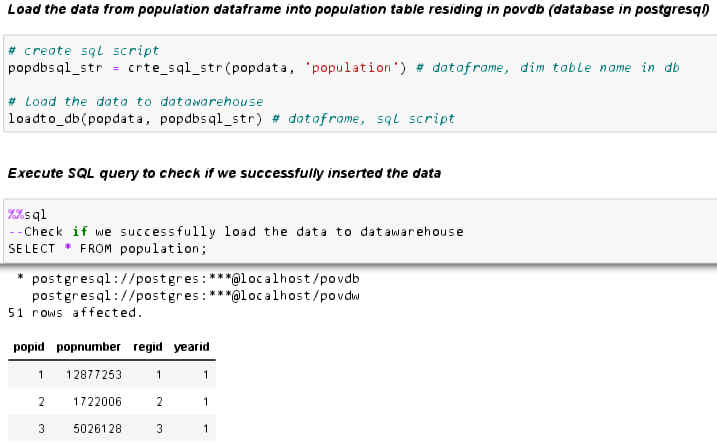
**Income**



**Employment**



**Population**



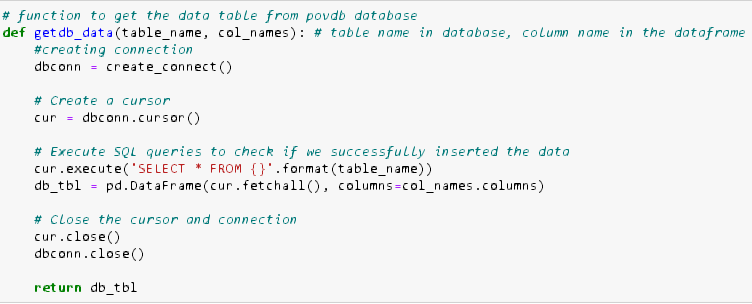
**PHASE II**

After the complete execution of Phase I pipeline, the Phase II will immediately be executed.

**II.1 EXTRACT**

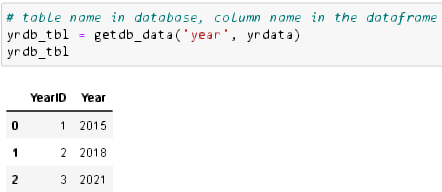
Extracting the data from povdb database.

Fetch the data from the database povdb (function)

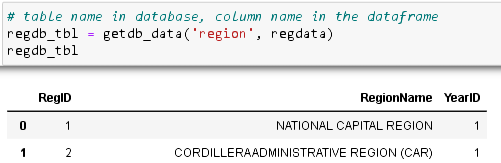


The function above will return a data frame after fetching the data from the 'povdb' database. The screenshot below will display only the first two rows of the entire data frame to conserve space in this documentation.

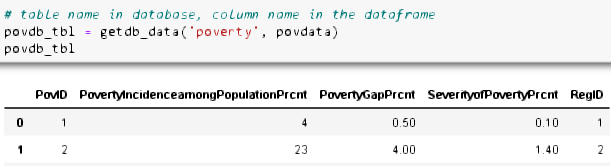
**Year**



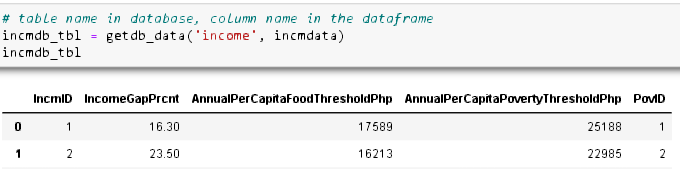
**Region**



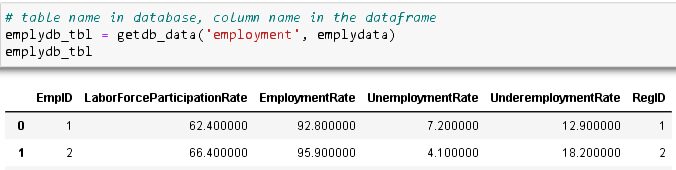
**Poverty**



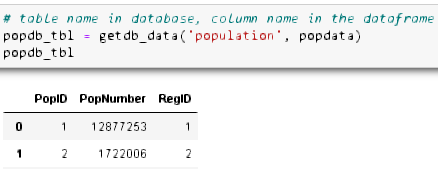
**Income**



**Employment**



**Population**



**II. 2 TRANSFORM**

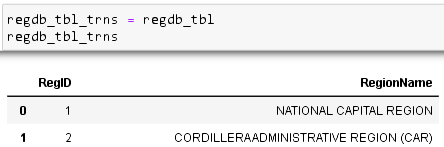
The data extracted from povdb database needs to be transformed to conform to the structure of dimension and fact table in data warehouse povdw except for the year and region table.

Removing the foreign key in the table (function)

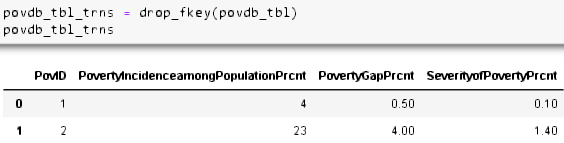


**Region**

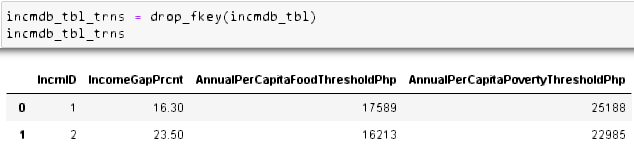
No transformation will be applied on region table as it is already conforming to the structure in data warehouse, however we need to assign the table into another variable for uniformity.



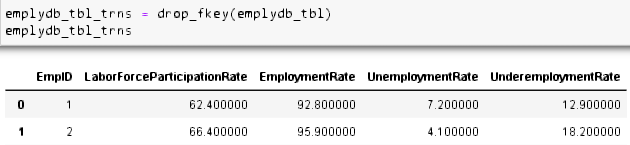
**Poverty**



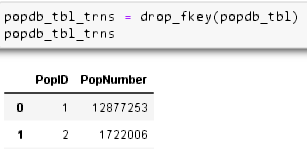
**Income**



**Employment**

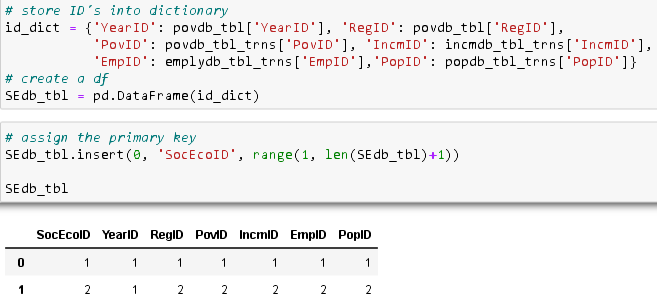


**Population**



**Socio Economic**

The Socio-Economic table will be loaded in the fact table. As mentioned previously on page 8, this fact table is called fact-less as it does not contain any calculated or aggregated data from the dimension tables, and this table will only serve as a bridge or connection.



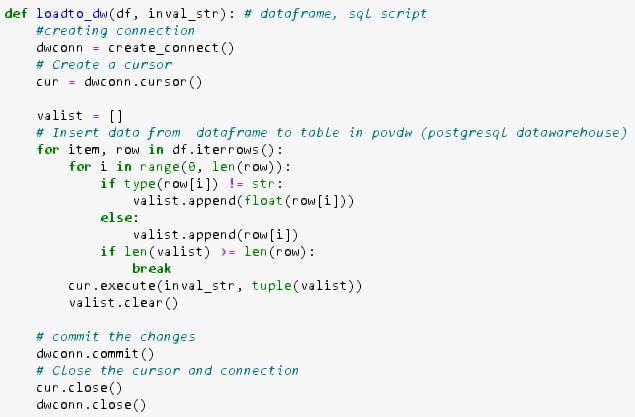
**II. 3 LOAD**

After the transformation process, will now load the data into the povdw data warehouse.

Creating connection to data warehouse.

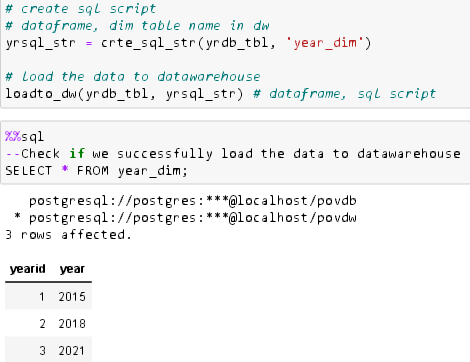


Load data to data warehouse (function)

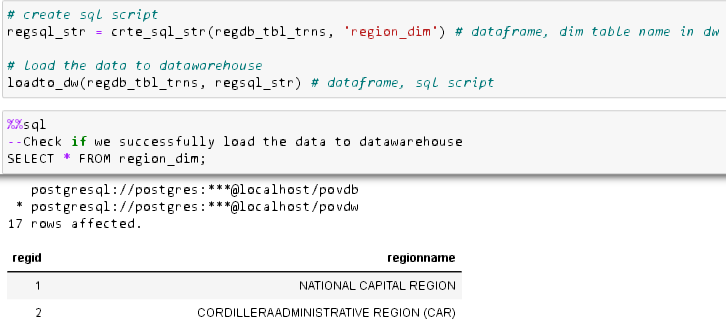


The function above will load the previously transformed data, but before using this function, will need to call again the previously created function named “crte\_sql\_string” to generate SQL string syntax.

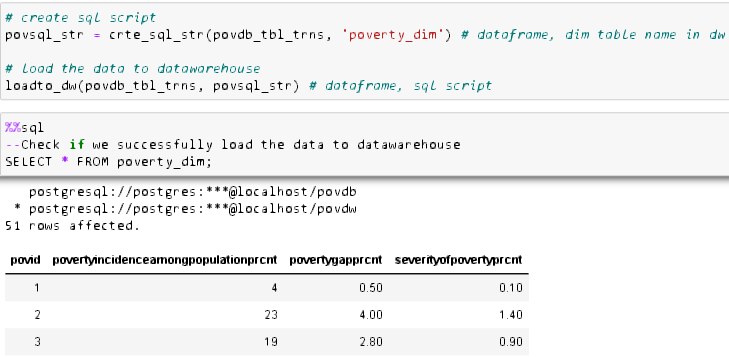
**Year\_dim**



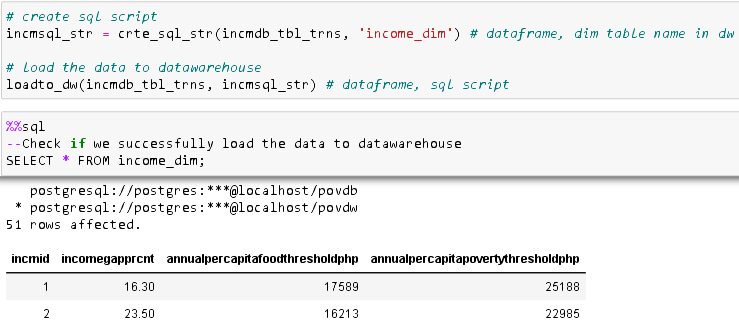
**Region\_dim**



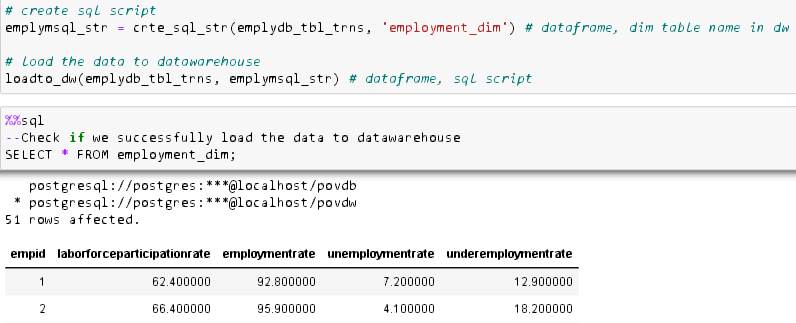
**Poverty\_dim**



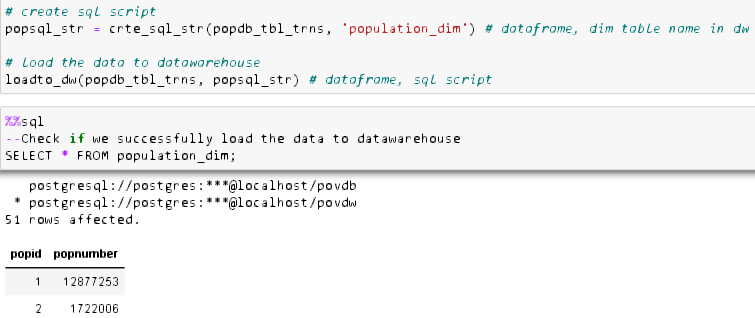
**Income\_dim**



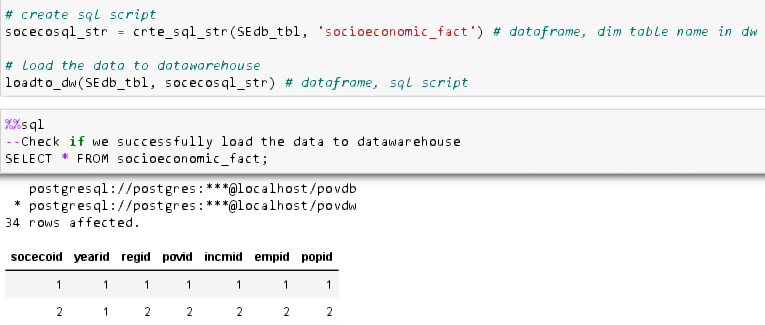
**Employment\_dim**



**Population\_dim**



**Socio-Economic\_fact**



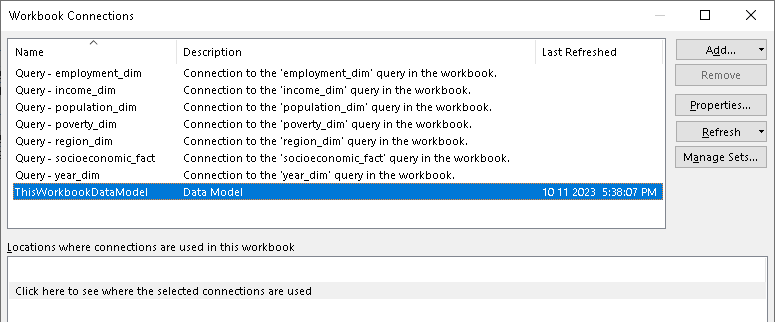
**Summary table**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Regions | Number of rows | Primary Keys |
| *2015* | NCR, CAR, Reg I – XIII, BARMM | 17 | 1 - 17 |
| *2018* | NCR, CAR, Reg I – XIII, BARMM | 17 | 18 - 34 |
| *2021p* | NCR, CAR, Reg I – XIII, BARMM | 17 | 35 - 51 |

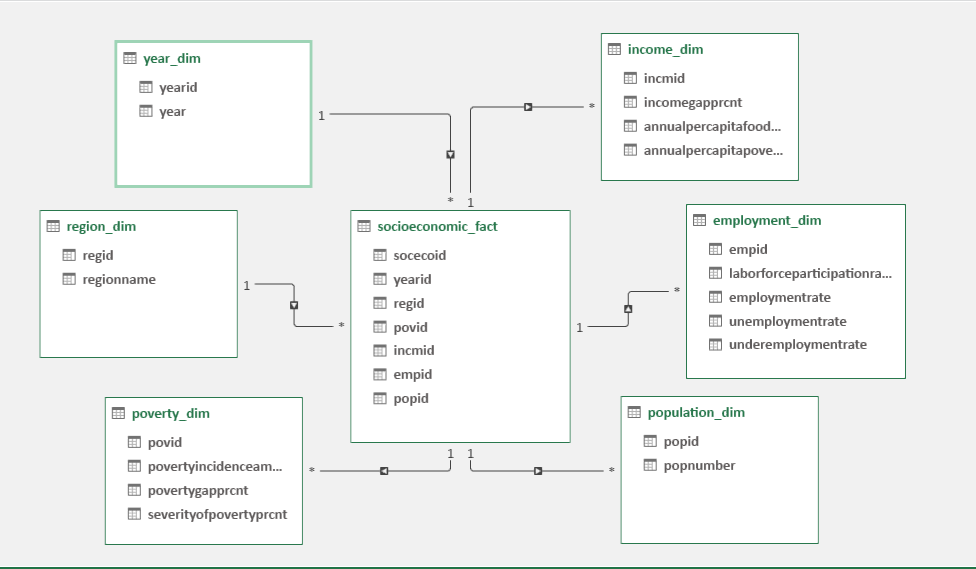
At the time of creating those screen captures during both Phase I and Phase II ETL processes, the displayed data is limited to two – three columns only, if you wish to view all the rows of data, please refer to the Jupyter notebook.

**Visualization (connecting PostgreSQL to Excel via ODBC)**

The image below is the established connection after connecting PostgreSQL povdw data warehouse to Excel via ODBC direct connection.

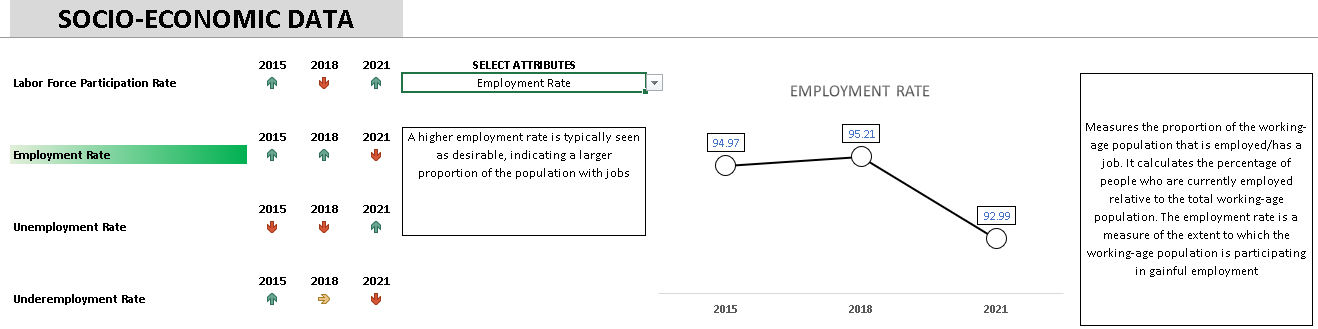


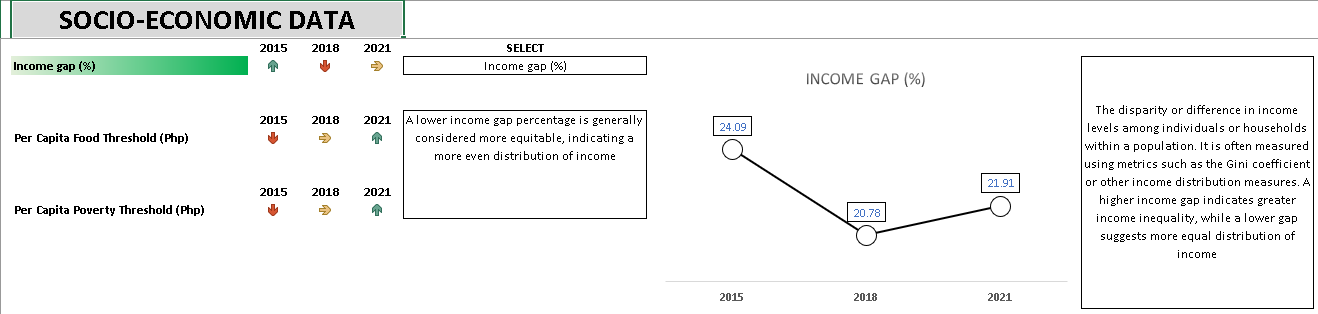
Pover Pivot Schema (Excel)

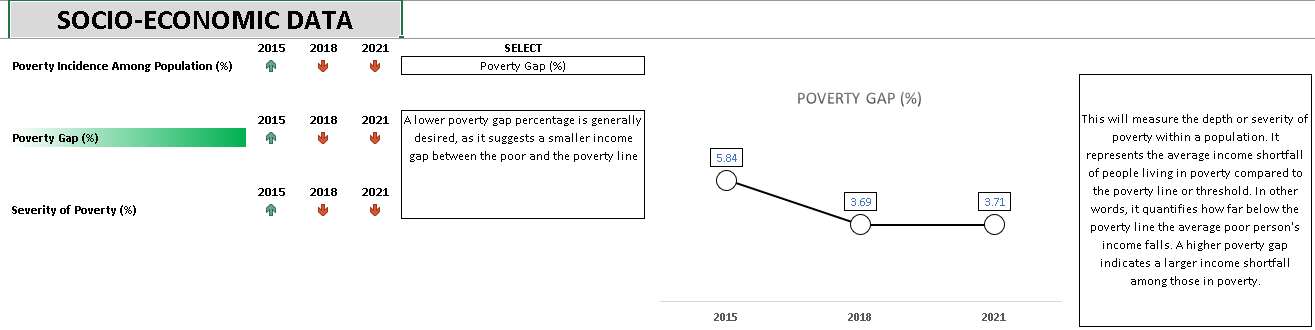


Creating a simple dynamic dashboard in excel to show the Socio-Economic data/status. **(see excel file for clearer view and more in-dept explanation)**

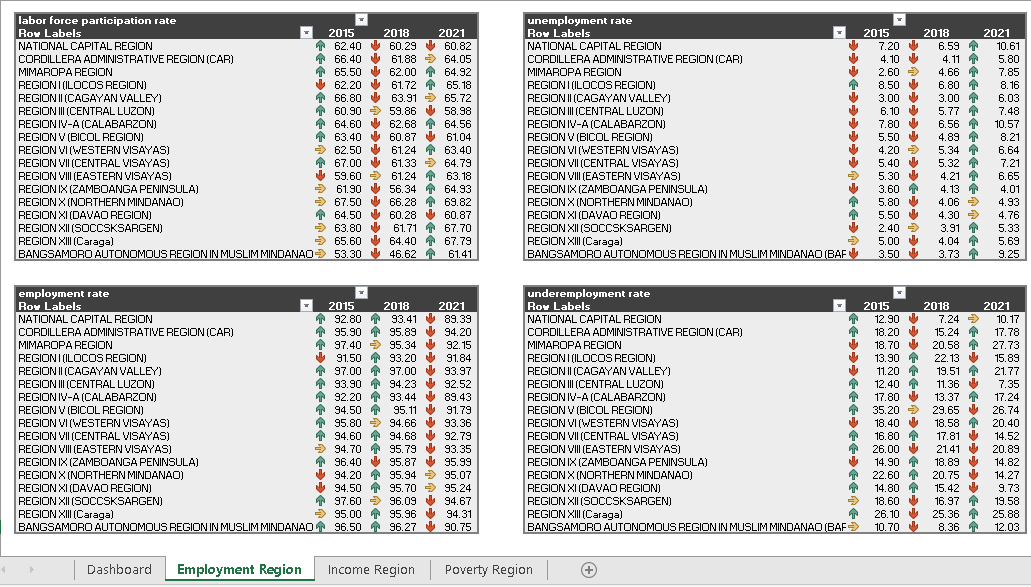
Overall Socio-Economic Status



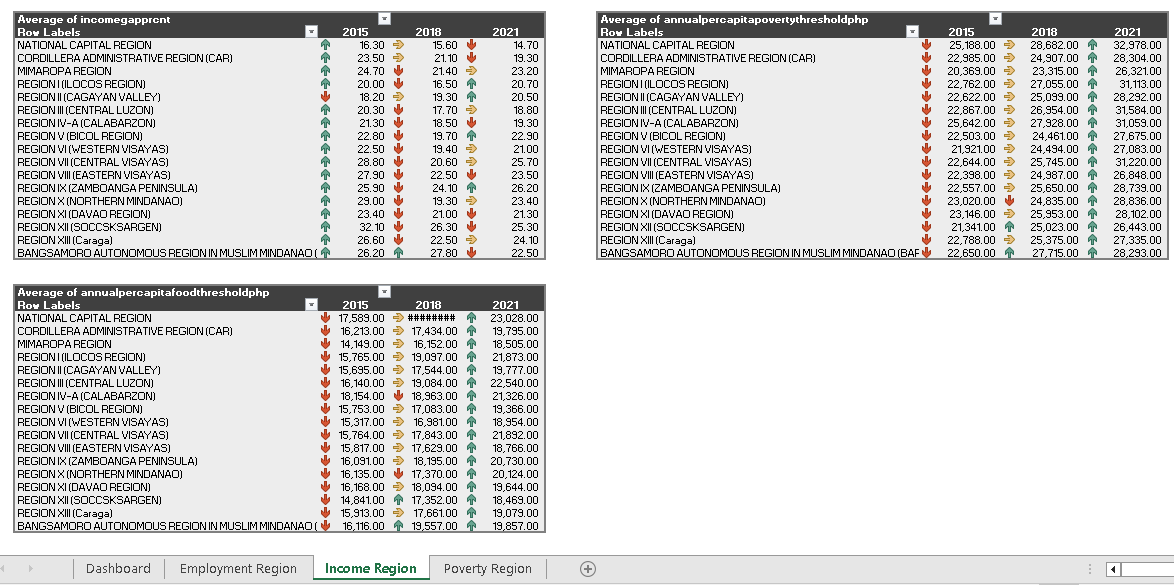




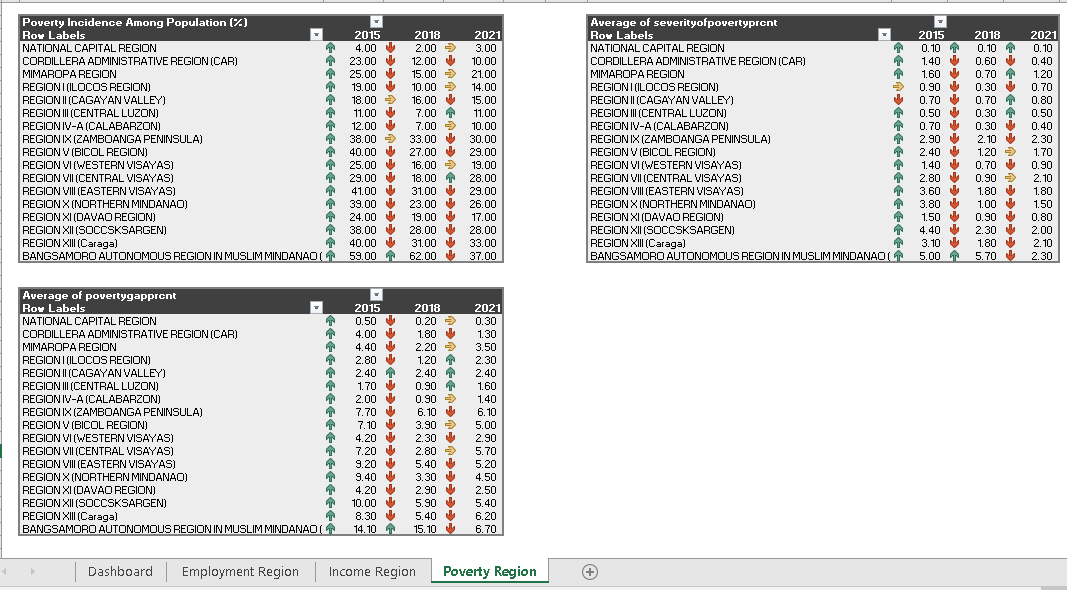
Employment by Region



Income by Region



Poverty by Region



D. **Conclusion**

Based on the table below, the government should focus on employment and, more importantly, on income attributes, namely 'per capita food threshold' and 'per capita poverty threshold' solutions, as they have been in a negative state since 2018.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Indicator** | **Attributes** | **2015** | **2018** | **2021** | **Conclusion for recent year** |
| Employment | Labor Force Participation Rate | + | - | + | Maintain |
| Employment Rate | + | + | - | The Government should focus on |
| Unemployment Rate | + | + | - | The Government should focus on |
| Employment Rate | - | +s | + | Maintain |
| Income | Income gap (%) | - | + | +s | For Improvement |
| Per Capita Food Threshold (Php) | + | -s | - | The Government should focus on |
| Per Capita Poverty Threshold (Php) | + | -s | - | The Government should focus on |
| Poverty | Poverty Incidence Among Population (%) | **-** | + | + | Maintain |
| Poverty Gap (%) | **-** | **+** | **+** | Maintain |
| Severity of Poverty (%) | **-** | **+** | **+** | Maintain |

(+) in positive state, the government should maintain this.

(-) in negative state, the government should take some intervention or action to solve the problem.

(S) slightly towards negative (-s) or positive (+s) state.

E. **References**

Kraak, M., Ricker, B., & Engelhardt, Y. (2018). Abstract. Challenges of Mapping Sustainable Development Goals Indicators Data. ISPRS International Journal of Geo-Information, 7(12), 482. https://doi.org/10.3390/ijgi7120482

Annual Labor and Employment Status<https://www.psa.gov.ph/content/2018-annual-labor-and-employment-status>

Annual Per Capita Food Threshold and Subsistence Incidence Among Population<https://openstat.psa.gov.ph/PXWeb/pxweb/en/DB/DB__1E__FY/0071E3DF040.px/?rxid=ca0a2b00-1d5a-412d-8a08-be01e009d80d>.

Annual Per Capita Poverty Threshold and Poverty Incidence Among Population<https://openstat.psa.gov.ph/PXWeb/pxweb/en/DB/DB__1E__FY/0031E3DF020.px/?rxid=aa90cec7-9eb7-43b8-b32c-0a563c706645>

Updated Projected Mid-Year Population for the Philippines Based on the 2015 POPCEN Results: 2020-2025 <https://psa.gov.ph/statistics/census/projected-population>

2021 Annual Labor Market Statistics (Preliminary Results) <https://psa.gov.ph/content/2021-annual-labor-market-statistics-preliminary-results>

2015 Annual Labor Survey <https://psa.gov.ph/sites/default/files/iesd/2022-12/TABLE%25204%2520Total%2520Population%252015%2520Years%2520Old%2520and%2520Over%2520and%2520Rates%2520of%2520Labor%2520Force%2520Participation%252C%2520Employment%2520Unemployment%2520%2520and%2520Underemployment%252C%2520by%2520Region%2520October%25202015.pdf>

Highlights of the Philippine Population 2020 Census of Population and Housing (2020 CPH) <https://psa.gov.ph/content/highlights-philippine-population-2020-census-population-and-housing-2020-cph>