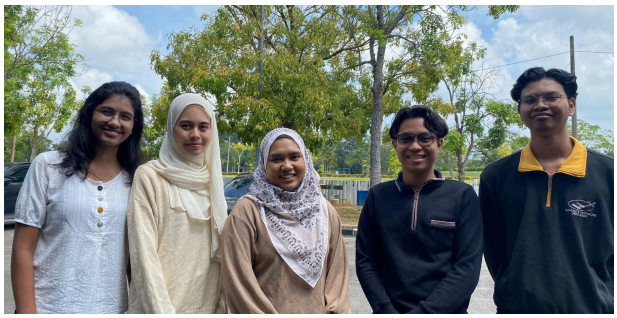
UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

BSD2343 DATA WAREHOUSING GROUP PROJECT

GROUP DATAOCEAN



TITLE: WATER AND HYGIENE IN SUDAN

SDG: CLEAN WATER & SANITATION (SDG6)

PREPARED FOR: DR NOR AZUANA BINTI RAMLI

STUDENT ID	NAME	SECTION
SD22002	ALMIRA DAMIA BINTI SYAHNIZAM	02G
SD22008	MUHAMMAD DANISH AIMAN HARISS BIN ROSLI	01G
SD22025	ABDUL HAZIQ AZIM BIN ABDUL MALIK	01G
SD22030	TUAN NURSHAFIEKA WAHIDA BINTI TUAN NADIN	02G
SD22065	VIENOSHA A/P SELVARAJU	02G

TABLE OF CONTENTS

1.0 BACKGROUND

1.1 Project Background

Clean water and good sanitation facilities should be highlighted and focused on, and that is why children in school have access to it. However, in countries such as those in Sudan for instance, many schools are unable to afford to meet even these basic needs. When children do not have access to clean water for drinking or available lavatory to use hygiene in schools it is unhealthy and learning too is affected. Various research findings have pointed out that this experience is not unique to students in Sudan schools but common in many other nations where students are subjected to inadequate resources to help them stay fit and healthy in class.

The aim of this project is to establish the reason behind schools in Sudan having scarce access to clean water and appropriate latrines. What we are doing now is trying to understand what is creating these problems and we are obtaining data from the likes of UNICEF as well as the World Bank. In our case, we want to know what stands in the way of schools being equipped with simple things like soap to use while washing hands or clean water to use for washing regardless of the country they are in. It is therefore through unveiling these challenges that we anticipate finding ways by which schools can enhance their facilities and provide a healthy environment to enhance children's learning environments. As much as this research focuses on Sudan, it is a small effort towards stopping people from suffering from these two factors which are water and sanitation in this world, and people should be given equal chances whether in Sudan or in any other part of the word. Addressing such problems in Sudan is all setting up solutions that help children globally and reducing the barrier for their health and education in school settings.

1.2 Problem to be solved

As we all know, water is important to all beings. Water is crucial for basic needs and daily use. In this dataset, we have used Water Hygiene in Sudanese Schools. In order to accomplish the Sustainable Development Goals 6 (SDG 6) we need to make sure that the WASH Facilities are sufficient in educational settings. Therefore, the problem that we are going to be solving is first the Water Contamination. Contaminated drinking water is one of the major problems in countries that are not that well developed. This can lead to sickness and even worse is death.The lack of proper water, sanitation, and hygiene facilities in schools makes this issue even worse, putting students and teachers at risk. Despite efforts to improve the situation, many children still do not have access to clean water and sanitation services at school. This highlights a significant gap in meeting hygiene standards in educational settings. It is important to address these issues to protect the health and education of children worldwide and work towards Sustainable Development Goals 3 and 6.

So, this is some of the crucial problems that need to be solved :

1. What is the area where the water is contaminated badly ?
2. What WASH Facilities that is not sufficient in the school ?
3. Is there a big gap in reaching Basic Hygiene Standards in schools ?
4. Do schools around the world provide the basic needs of WASH Services in schools ?

1.3 Objectives

1. To identify the Sudanese regions with the worst water contamination in school systems.
2. To assess the quality and accessibility of sanitation, drinking water, and hygiene services in Sudanese primary and secondary schools in both rural and urban regions.
3. To investigate the gaps in reaching basic hygiene standards in schools and determine the extent of these gaps.
4. To investigate the causes of the significant decreases in the observations for drinking water and hygiene services in rural areas and develop strategies to improve the coverage and quality of these services.

1.4 Data schema

In a data warehouse, a data schema is like a blueprint that organizes and structures the data stored within it. It sets the rules for how data is organized, connected, and used. Schemas are super important for keeping data in check and making sure queries run smoothly.In our project , we have 4 datasets which are the indicator, school, school hygiene and service level datasets.

So. we used two libraries in our jupyter notebook to display data schema which is pandas and numpy.

- Importing the library

Figure 1.5.1

Based on *Figure 1.5.1 ,* we used two libraries in our data to identify the data schema. In that figure we import pandas and numpy libraries.

- For indicator data schema

```
[4]: indicator.dtypes

[4]: country_name       object
     indicator_name     object
     indicatorvalue    float64
     yearindicator       int64
     dtype: object

[14]: indicator.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 2001 entries, 0 to 2000
     Data columns (total 4 columns):
      #   Column          Non-Null Count  Dtype
     ---  ------          --------------  -----
      0   country_name    2001 non-null   object
      1   indicator_name  2001 non-null   object
      2   indicatorvalue  2001 non-null   float64
      3   yearindicator   2001 non-null   int64
     dtypes: float64(1), int64(1), object(2)
     memory usage: 62.7+ KB
```

Figure 1.5.2

Based on *Figure 1.5.2 ,* we can see that the indicator consists of 4 columns which is country_name , indicator_name, indicatorvalue and yearindicator. For country_name and indicator_name has a string data type and for the indicatorvalue is a float data type.However for yearindicator is integer data type.

- For school data schema

```
[8]: school.dtypes

[8]: geographicarea          object
     indicatorproportion     object
     servicetype             object
     typeofschool            object
     residence               object
     timeperiod               int64
     observation            float64
     dtype: object

[15]: school.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 176 entries, 0 to 175
     Data columns (total 7 columns):
      #   Column               Non-Null Count  Dtype
     ---  ------               --------------  -----
      0   geographicarea       176 non-null    object
      1   indicatorproportion  176 non-null    object
      2   servicetype          176 non-null    object
      3   typeofschool         176 non-null    object
      4   residence            176 non-null    object
      5   timeperiod           176 non-null    int64
      6   observation          176 non-null    float64
     dtypes: float64(1), int64(1), object(5)
     memory usage: 9.8+ KB
```

Figure 1.5.3

Based on *Figure 1.5.3*, we can observe that the school data schema consists of 7 columns. Most of the columns are string data type except for timeperiod and observation where the data type is integer and float respectively.

- For school hygiene data schema

```
[11]: school_hygiene.dtypes

[11]: geographic_area         object
      indicatorproportion     object
      service_type            object
      residence               object
      time_period              int64
      observation            float64
      dtype: object

[16]: school_hygiene.info()

      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 139 entries, 0 to 138
      Data columns (total 6 columns):
       #   Column               Non-Null Count  Dtype
      ---  ------               --------------  -----
       0   geographic_area      139 non-null    object
       1   indicatorproportion  139 non-null    object
       2   service_type         139 non-null    object
       3   residence            139 non-null    object
       4   time_period          139 non-null    int64
       5   observation          139 non-null    float64
      dtypes: float64(1), int64(1), object(4)
      memory usage: 6.6+ KB
```

Figure 1.5.4

Based on *Figure 1.5.4* , this is the school hygiene data schema and it consists of 6 columns as well. The data type is mostly a string as well except for the time_period and observation as well which is an integer and float respectively as well.

- For service level data schema

```
[13]: service_level.dtypes

[13]: country           object
      residence_type    object
      service_type      object
      time_period        int64
      coverage         float64
      population       float64
      service_level     object
      dtype: object

[17]: service_level.info()

      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 103 entries, 0 to 102
      Data columns (total 7 columns):
       #   Column          Non-Null Count  Dtype
      ---  ------          --------------  -----
       0   country         103 non-null    object
       1   residence_type  103 non-null    object
       2   service_type    103 non-null    object
       3   time_period     103 non-null    int64
       4   coverage        103 non-null    float64
       5   population      103 non-null    float64
       6   service_level   103 non-null    object
      dtypes: float64(2), int64(1), object(4)
      memory usage: 5.8+ KB
```

Figure 1.5.5

Based on *Figure 1.5.5* , this is the service level data schema and it consists of 7 columns. The data type for country, residence_type, service_type and service_level are strings. However, time_period is an integer. The last two columns which are coverage and population are float data types

2.0 ARCHITECTURE AND ETL PIPELINE

2.1 Architecture Structure

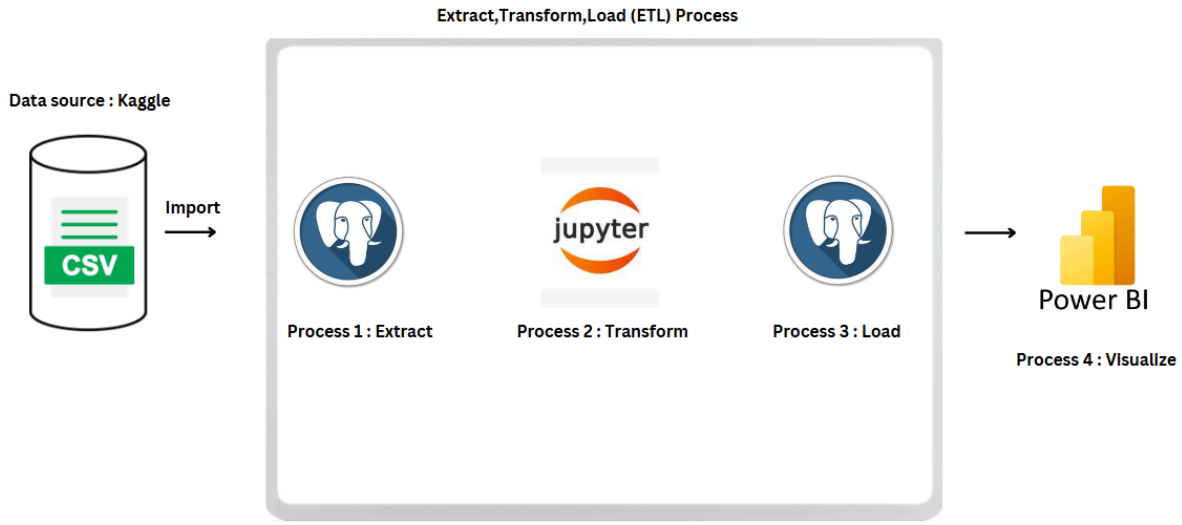


Figure 2.1

Based on Figure 2.1, the Water and Hygiene in Sudanese Schools data set was acquired in Kaggle.com. The data set consists of 4 tables which are indicator, school, school hygiene and service level.These tables were then imported into PostgreSQL. After successfully importing all the tables into PostgreSQL, the process continues as we connect to python in Jupyter Notebook for further operations. Some necessary libraries need to be installed first to extract data from PostgreSQL. After installing the libraries needed, the dataset was then cleaned and transformed where NULL values were identified and removed. After the cleaning and transformation process, the data integration process took place. Data integration process is where the tables are combined and provide a unified, single view of data. The OLAP operations like slicing, roll up and dicing were performed after the cleaned data load into PostgreSQL in order to provide a better view and make it easier for analysis steps that are going to take place. After OLAP operations are done, the results are then imported into powerBI for the visualization part. As in power BI there are various analyses that can be performed in order to observe the outcomes that we need.

2.2 ETL Pipeline

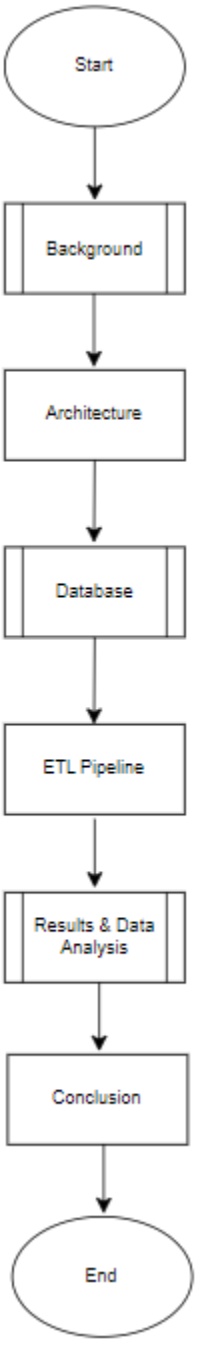


Figure 2.2 Flow of the project

Figure 2.2 shows the overall procedure of the project, which our project will complete in six stages.

2.3 ETL Process

2.3.1 Extract

The datasets must be saved in a PostgreSQL database before the ETL procedure can begin. To begin with, make a new database and utilize database connectors to pull pertinent information from every table.

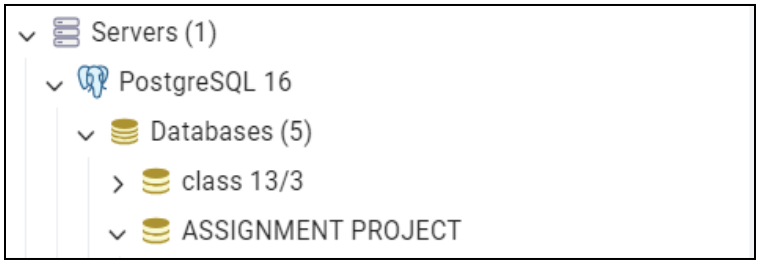


Figure 2.3.1.1 Database in PostgreSQL

Figure 2.3.1 shows that we created a database named 'ASSIGNMENT PROJECT' in PostgreSQL.

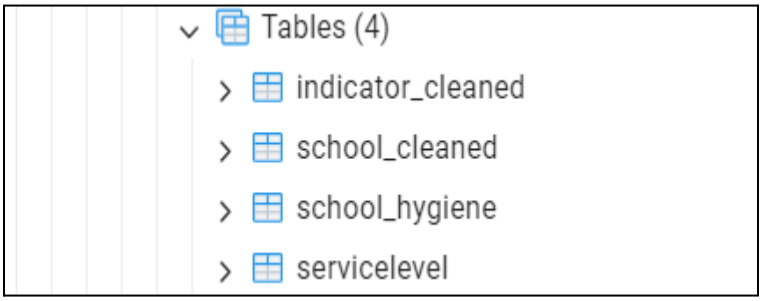


Figure 2.3.1.2 Tables created

Query to create tables:

a)
```
create table indicator_cleaned
(
country_name text,
indicator_name text,
indicatorvalue numeric,
yearindicator numeric
);
```

b)
```
create table servicelevel
(
country text,
residence_type text,
service_type text,
time_period numeric,
coverage numeric,
population numeric,
service_level text
);
```

c)
```
create table school_hygiene
(
geographic_area text,
indicatorproportion text,
service_type text,
residence text,
time_period numeric,
observation numeric
);
```

d)
```
create table school_cleaned
(
geographicarea text,
indicatorProportion text,
servicetype text,
typeofschool text,
residence text,
timeperiod numeric,
observation numeric
```

);

Query to copy the data from csv file into table:

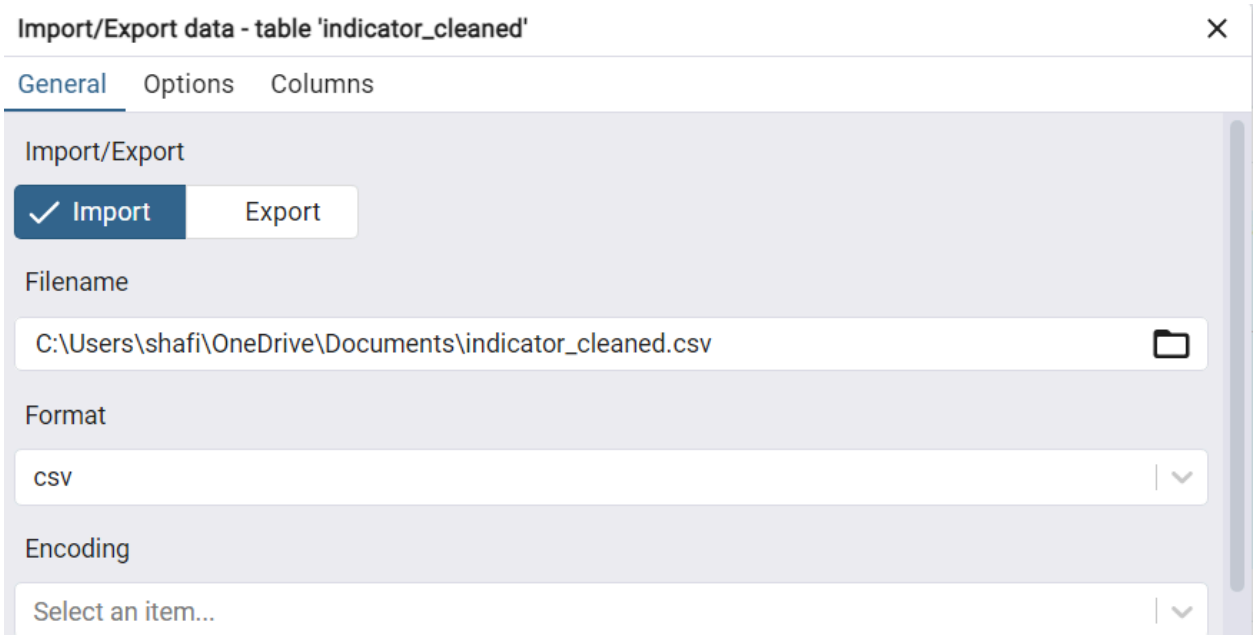


Figure 2.3.1.3 Import table indicator cleaned data from csv

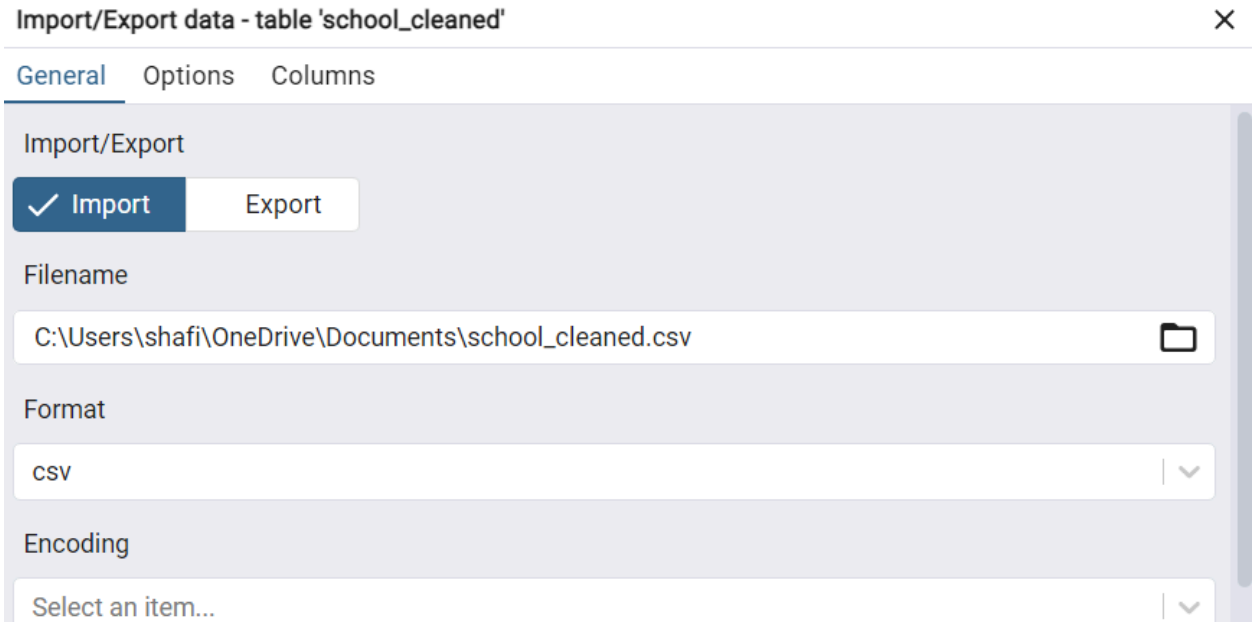


Figure 2.3.1.4 Import table school cleaned data from csv

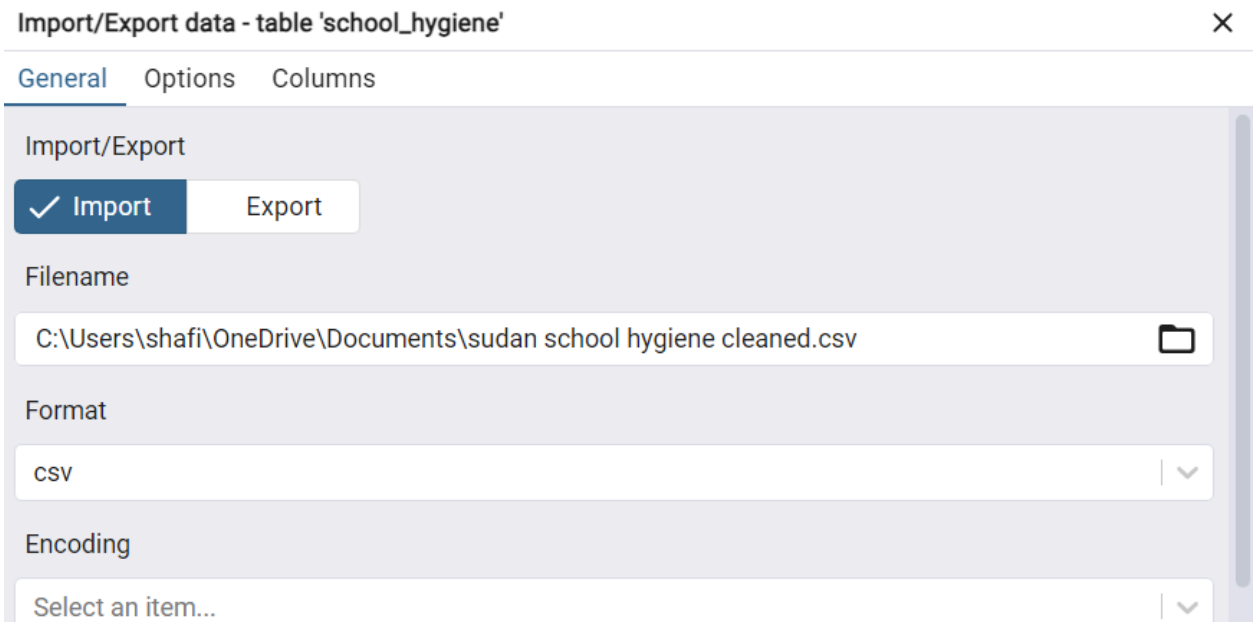


Figure 2.3.1.5 Import table school hygiene data from csv

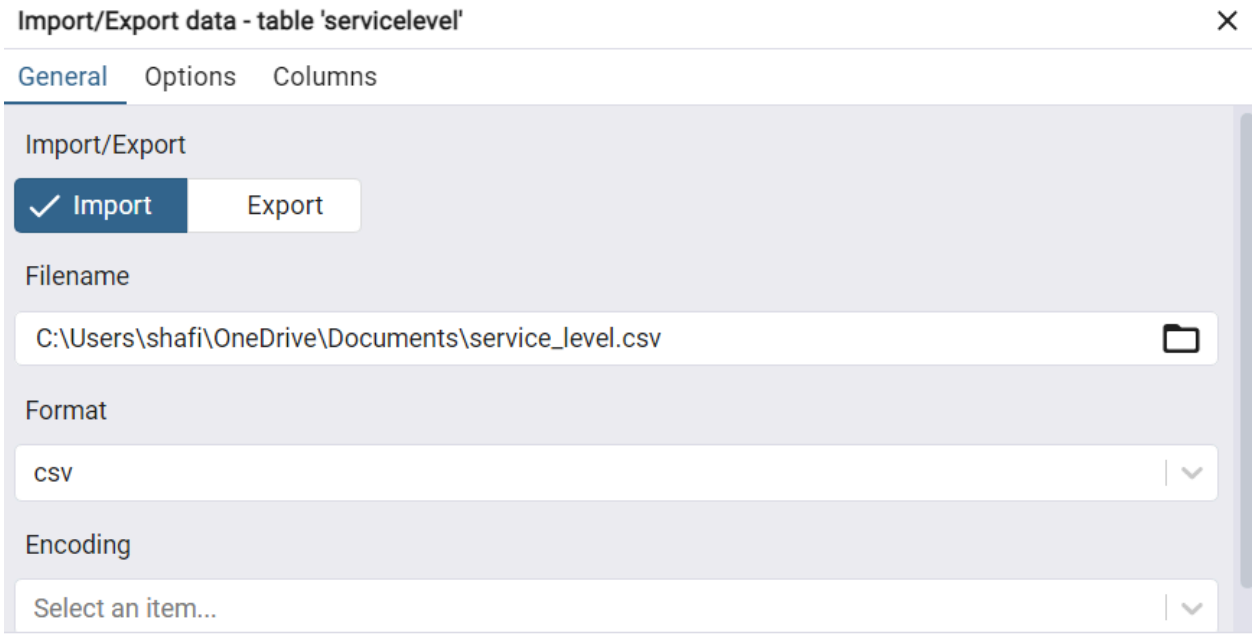
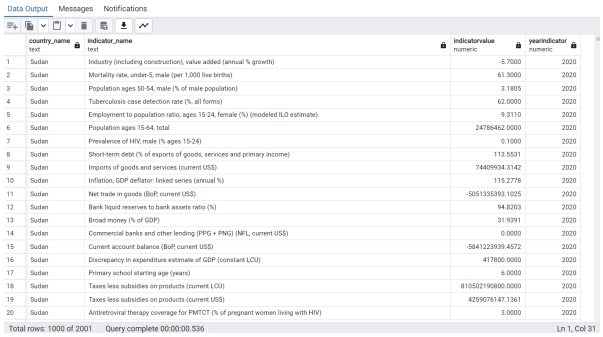
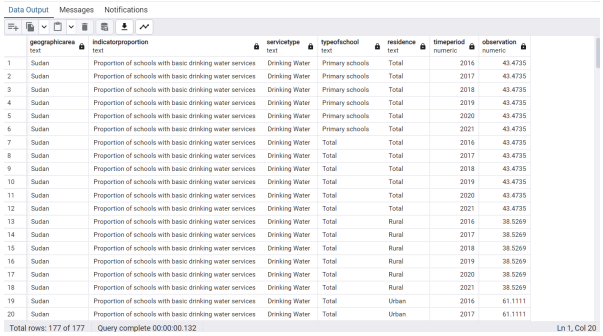
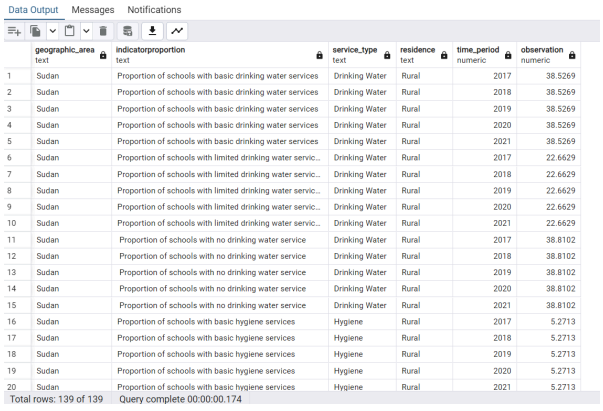


Figure 2.3.1.6 Import table service level data from csv

Run a query (Select * from {table name}) to view the data in the table:

QUERY	OUTPUT
SELECT *FROM indicator_cleaned	
SELECT *FROM school_cleaned	
SELECT *FROM school_hygiene	

SELECT *FROM servicelevel

Data Output | Messages | Notifications

	country text	residence_type text	service_type text	time_period numeric	coverage numeric	population numeric	service_level text
1	Sudan	total	Sanitation	2018	36.6727	15402187.7288	At least basic
2	Sudan	rural	Sanitation	2018	24.2341	6652198.0443	At least basic
3	Sudan	urban	Sanitation	2018	60.1402	8749989.9766	At least basic
4	Sudan	total	Drinking water	2018	60.4671	25395606.3873	At least basic
5	Sudan	rural	Drinking water	2018	54.4624	14949776.6469	At least basic
6	Sudan	urban	Drinking water	2018	71.7960	10445830.0651	At least basic
7	Sudan	total	Hygiene	2018	16.0087	6723507.1430	Basic service
8	Sudan	total	Hygiene	2018	26.3621	11071824.6133	Limited service
9	Sudan	total	Sanitation	2018	8.3345	3500420.6191	Limited service
10	Sudan	rural	Sanitation	2018	6.3995	1756646.2016	Limited service
11	Sudan	urban	Sanitation	2018	11.9853	1743774.4742	Limited service
12	Sudan	total	Drinking water	2018	26.8768	11287982.1623	Limited service
13	Sudan	rural	Drinking water	2018	28.1097	7716036.8380	Limited service
14	Sudan	urban	Drinking water	2018	24.5506	3571945.4265	Limited service
15	Sudan	total	Hygiene	2018	57.6292	24203726.5080	No handwashing facility
16	Sudan	total	Sanitation	2018	21.4077	8991009.4779	Open defecation
17	Sudan	rural	Sanitation	2018	28.8269	7912920.0451	Open defecation
18	Sudan	urban	Sanitation	2018	7.4099	1078089.4446	Open defecation
19	Sudan	total	Drinking water	2018	4.1331	1735860.6876	Surface water
20	Sudan	rural	Drinking water	2018	5.8447	1604363.8642	Surface water

Total rows: 103 of 103 Query complete 00:00:00.164

After the raw data has been extracted into pgAdmin, we need to connect our pgAdmin with the Jupyter Notebook to proceed to the next step which transforms the data.

```
[1]: ! pip install ipython-sql
     ! pip install sqlalchemy
     ! pip install psycopg2
     ! pip install python-sql
     ! pip install pandas-sql
     ! pip install sql-queries
```

After the packages have been installed, we need to load ipython-sql with the following command:

```
[2]: %reload_ext sql
```

Calling the create engine function:

```
[3]: from sqlalchemy import create_engine
```

Importing necessary libraries for ETL Process:

```
import pandas as pd
import psycopg2 as ps
import pandas.io.sql as sqlio
import missingno as msno
```

Extracting data from PgAdmin into Jupyter Notebook:

```
connectpg=ps.connect(dbname="ASSIGNMENT PROJECT",
                     user="postgres",password="1234",host="localhost",
                     port="5432")
```

2.3.2 Transform

After connecting the PgAdmin with the Jupyter Notebook the dataset needs to be cleaned as it is really important to do the data cleaning process. Some connectors are installed to ensure the data can be transferred from PostgreSQL to Python. The data needs to be stored in a data frame by the pandas library in order to ease the data cleaning process.

```
sql="""SELECT * FROM pg.catalog.pg_tables"""
```

	country	residence_type	service_type	time_period	coverage	population	service_level
0	Sudan	total	Sanitation	2018.0	36.6727	1.540219e+07	At least basic
1	Sudan	rural	Sanitation	2018.0	24.2341	6.652198e+06	At least basic
2	Sudan	urban	Sanitation	2018.0	60.1402	8.749990e+06	At least basic
3	Sudan	total	Drinking water	2018.0	60.4671	2.539561e+07	At least basic
4	Sudan	rural	Drinking water	2018.0	54.4624	1.494978e+07	At least basic
...	...	...	...	...	...	...	...
98	Sudan	rural	Drinking water	2022.0	3.2975	9.899004e+05	Surface water
99	Sudan	urban	Drinking water	2022.0	0.4423	7.454997e+04	Surface water
100	Sudan	total	Drinking water	2022.0	3.9482	1.850671e+06	Unimproved
101	Sudan	rural	Drinking water	2022.0	6.1648	1.850672e+06	Unimproved
102	Sudan	urban	Drinking water	2022.0	0.0000	0.000000e+00	Unimproved

Figure 2.3.2.1 shows the data stored in data frame

After the data is stored in a data frame, now the data cleaning process can take part.

Selecting the first table servicelevel and checking if there are any null or missing values.

```
sql="""SELECT * FROM servicelevel"""
```

```
df1=sqlio.read_sql_query(sql,connectpg)
df1
```

```
df1.isnull().sum()
```

```
country           0
residence_type    0
service_type      0
time_period       0
coverage          0
population        0
service_level     0
dtype: int64
```

Selecting the second table school_hygiene and checking if there are any null or missing values.

```
sql="""SELECT * FROM school_hygiene"""
```

```
df2=sqlio.read_sql_query(sql,connectpg)
df2
```

```
df2.isnull().sum()
```

```
geographic_area         0
indicatorproportion     0
service_type            0
residence               0
time_period             0
observation             0
dtype: int64
```

Selecting the third table school_cleaned and checking if there are any null or missing values.

```
sql="""SELECT * FROM school_cleaned"""
```

```
df3=sqlio.read_sql_query(sql,connectpg)
df3
```

```
df3.isna().sum()
```

```
geographicarea          1
indicatorproportion     1
servicetype             1
typeofschool            1
residence               1
timeperiod              1
observation             1
dtype: int64
```

Dropping the null values in the third table.

```
df3_new=df3.dropna()
df3_new
```

```
df3_new.isna().sum()
```

```
geographicarea          0
indicatorproportion     0
servicetype             0
typeofschool            0
residence               0
timeperiod              0
observation             0
dtype: int64
```

Selecting the fourth table indicator_cleaned and checking if there are any null or missing values.

```
sql="""SELECT * FROM indicator_cleaned"""
```

```
df4=sqlio.read_sql_query(sql,connectpg)
df4
```

```
df4.isna().sum()
```

```
country_name      0
indicator_name    0
indicatorvalue    0
yearindicator     0
dtype: int64
```

2.3.3 Load

After cleaning the data in Jupyter, we have to load the data into PostgreSQL. By creating a table and database as shown below we can import the data from Jupyter Notebook into PostgreSQL.

```
import os

output_directory= r"C:\Users\shafi\OneDrive\Documents\PROJECT ASSIGNMENT DW"

os.makedirs(output_directory,exist_ok=True)

altered_tablenames= ['servicelevel','school_hygiene','school_cleaned','indicator_cleaned']

dfdict={
    'servicelevel':df1,
    'school_hygiene':df2,
    'school_cleaned':df3_new,
    'indicator_cleaned':df4
}

for table_name in altered_tablenames:
    csv_filepath=os.path.join(output_directory,f"{table_name}.csv")
    dfdict[table_name].to_csv(csv_filepath,index=False)
```

Figure 2.3.3.1 the data is being loaded to a new csv file

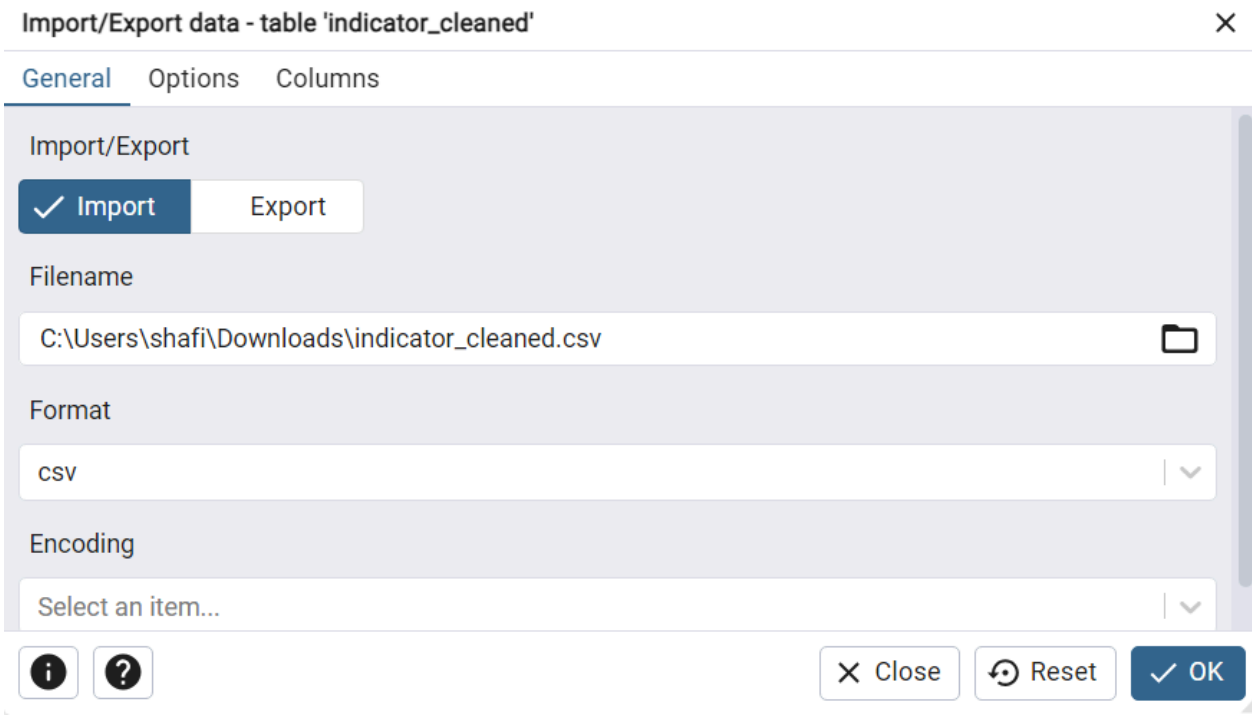
Query to create the cleaned tables from dataset:

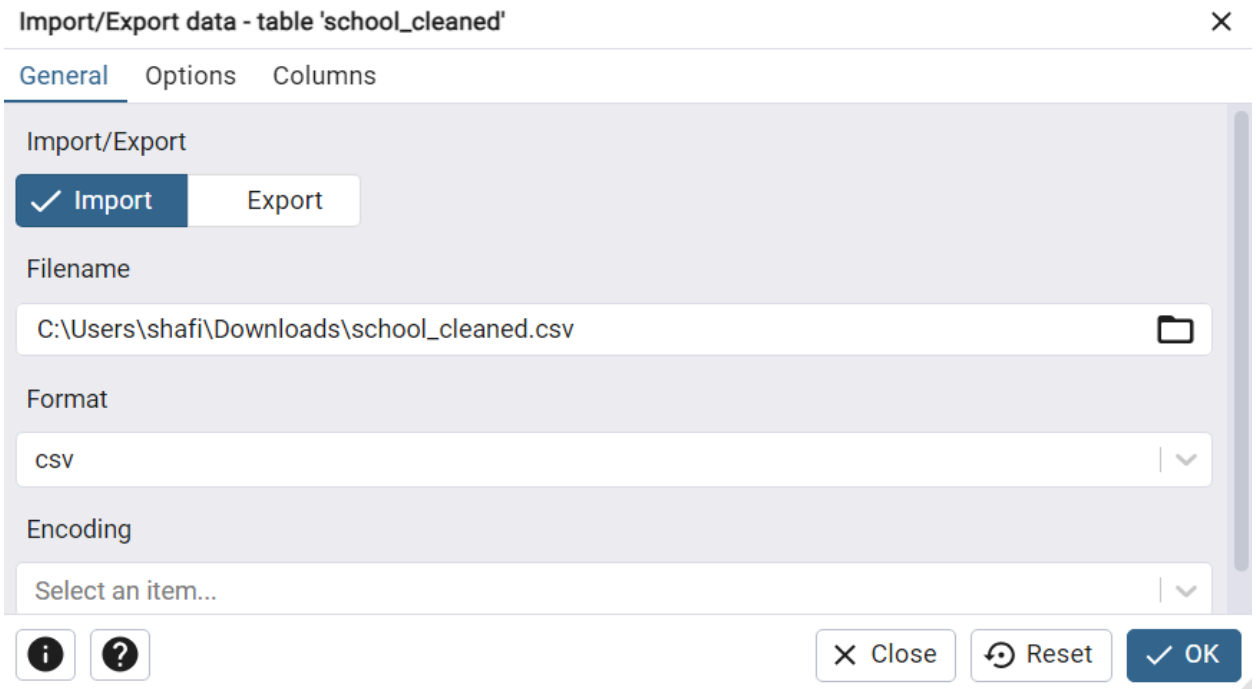
a)
create table indicator_cleaned
(
country_name text,
indicator_name text,
indicatorvalue numeric,
yearindicator numeric
);

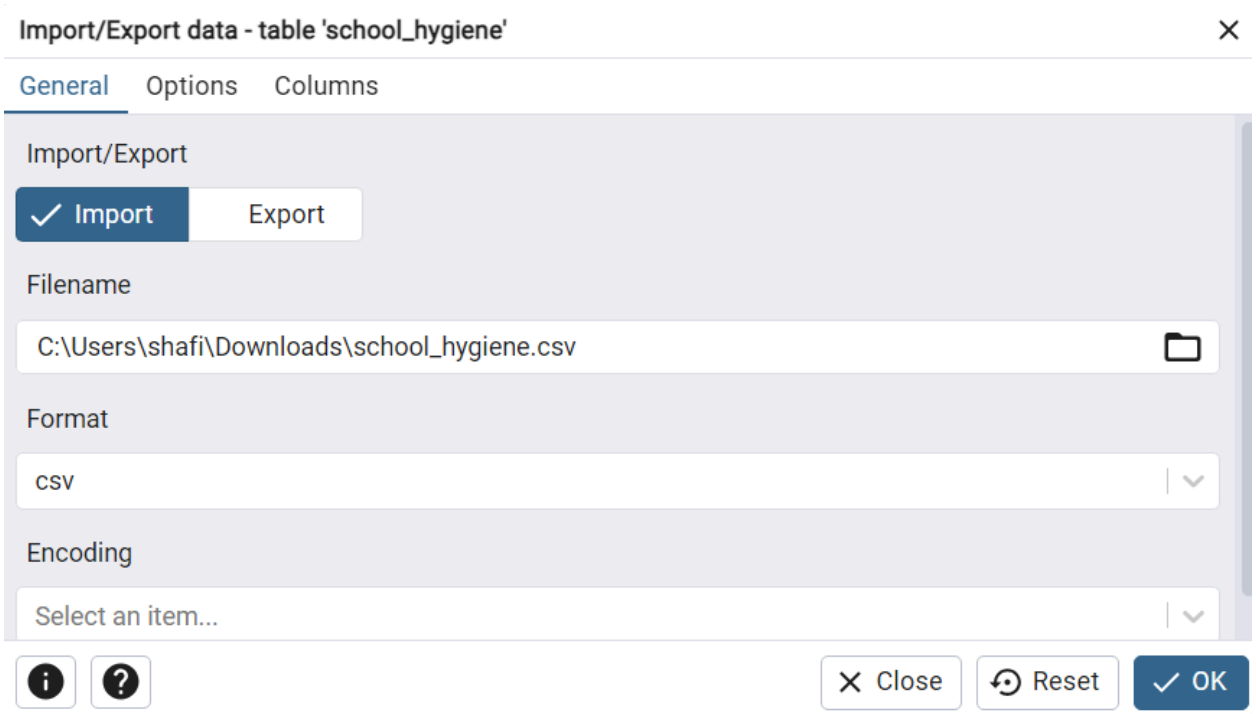
b)
create table servicelevel
(
country text,
residence_type text,
service_type text,
time_period numeric,
coverage numeric,

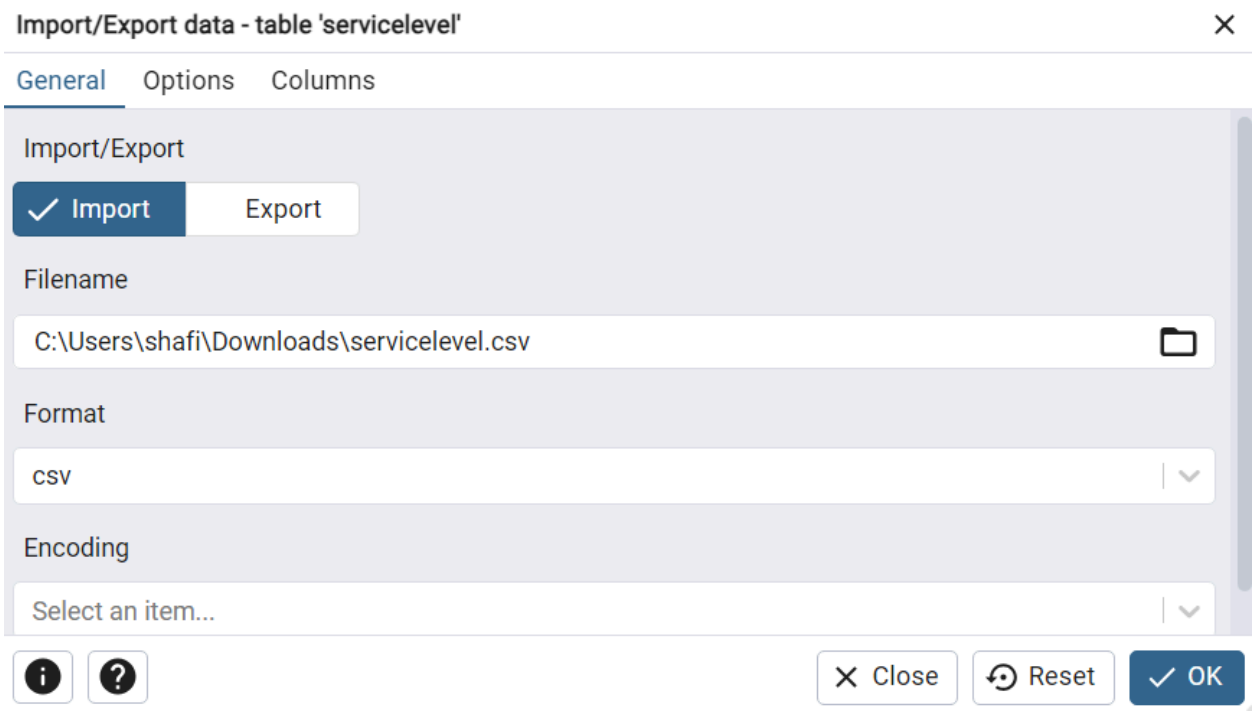
```
population numeric,
service_level text
);

c)
create table school_hygiene
(
geographic_area text,
indicatorproportion text,
service_type text,
residence text,
time_period numeric,
observation numeric
);

d)
create table school_cleaned
(
geographicarea text,
indicatorProportion text,
servicetype text,
typeofschool text,
residence text,
timeperiod numeric,
observation numeric

);
```

Query to copy the cleaned data from csv file into table:

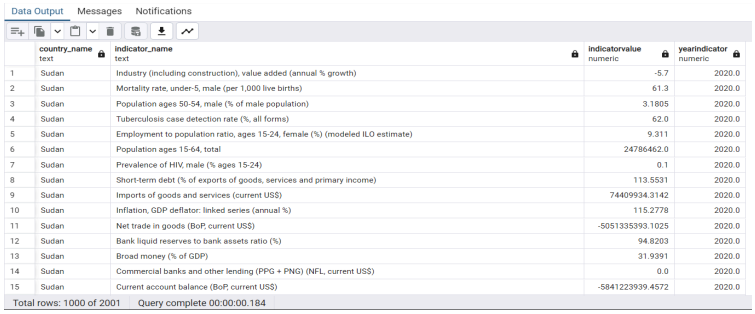
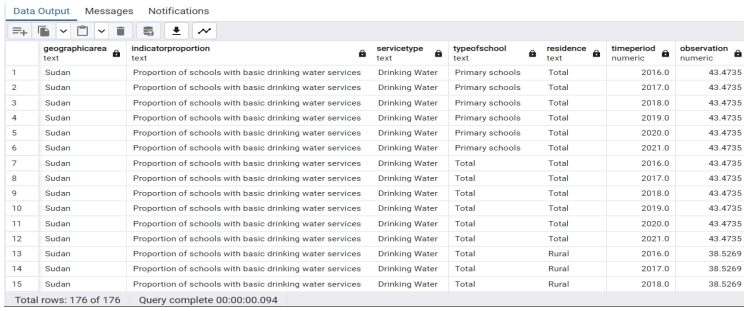
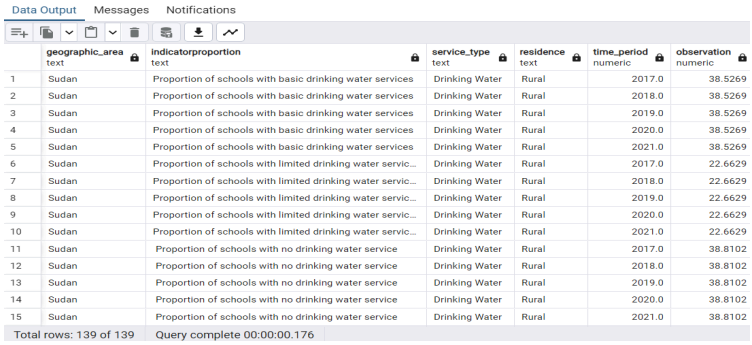
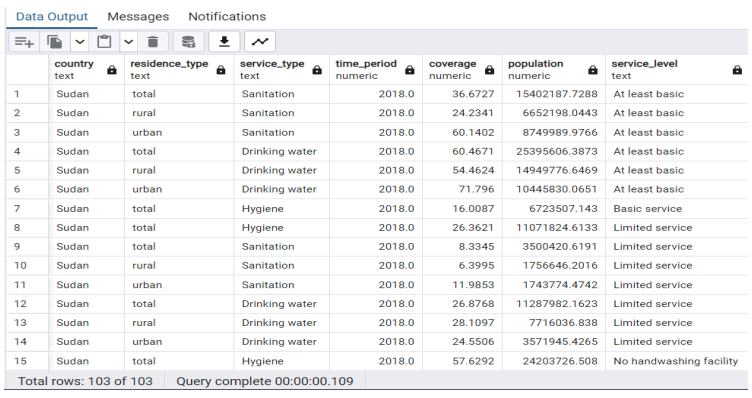





Import/Export data - table 'school_hygiene'
General
Options
Columns
Import/Export
Import
Export
Filename
C:\Users\shafi\Downloads\school_hygiene.csv
Format
csv
Encoding
Select an item...
Close
Reset
OK


Import/Export data - table 'servicelevel'
General
Options
Columns
Import/Export
Import
Export
Filename
C:\Users\shafi\Downloads\servicelevel.csv
Format
csv
Encoding
Select an item...
Close
Reset
OK

Run a query (Select * from {table name}) to view the cleaned data in the table:

QUERY	TABLE
SELECT *FROM indicator_cleaned	
SELECT *FROM school_cleaned	
SELECT *FROM school_hygiene	
SELECT *FROM servicelevel	

Data Output | Messages | Notifications

	country_name text	indicator_name text	indicatorvalue numeric	yearindicator numeric
1	Sudan	Industry (including construction), value added (annual % growth)	-5.7	2020.0
2	Sudan	Mortality rate, under-5, male (per 1,000 live births)	61.3	2020.0
3	Sudan	Population ages 50-54, male (% of male population)	3.1805	2020.0
4	Sudan	Tuberculosis case detection rate (%, all forms)	62.0	2020.0
5	Sudan	Employment to population ratio, ages 15-24, female (%) (modeled ILO estimate)	9.311	2020.0
6	Sudan	Population ages 15-64, total	24786462.0	2020.0
7	Sudan	Prevalence of HIV, male (% ages 15-24)	0.1	2020.0
8	Sudan	Short-term debt (% of exports of goods, services and primary income)	113.5531	2020.0
9	Sudan	Imports of goods and services (current US$)	74409934.3142	2020.0
10	Sudan	Inflation, GDP deflator: linked series (annual %)	115.2778	2020.0
11	Sudan	Net trade in goods (BoP, current US$)	-5051335393.1025	2020.0
12	Sudan	Bank liquid reserves to bank assets ratio (%)	94.8203	2020.0
13	Sudan	Broad money (% of GDP)	31.9391	2020.0
14	Sudan	Commercial banks and other lending (PPG + PNG) (NFL, current US$)	0.0	2020.0
15	Sudan	Current account balance (BoP, current US$)	-5841223939.4572	2020.0

Total rows: 1000 of 2001 Query complete 00:00:00.184

Data Output | Messages | Notifications

	geographicarea text	indicatorproportion text	servicetype text	typeofschool text	residence text	timeperiod numeric	observation numeric
1	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Primary schools	Total	2016.0	43.4735
2	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Primary schools	Total	2017.0	43.4735
3	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Primary schools	Total	2018.0	43.4735
4	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Primary schools	Total	2019.0	43.4735
5	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Primary schools	Total	2020.0	43.4735
6	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Primary schools	Total	2021.0	43.4735
7	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Total	Total	2016.0	43.4735
8	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Total	Total	2017.0	43.4735
9	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Total	Total	2018.0	43.4735
10	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Total	Total	2019.0	43.4735
11	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Total	Total	2020.0	43.4735
12	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Total	Total	2021.0	43.4735
13	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Total	Rural	2016.0	38.5269
14	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Total	Rural	2017.0	38.5269
15	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Total	Rural	2018.0	38.5269

Total rows: 176 of 176 Query complete 00:00:00.094

Data Output | Messages | Notifications

	geographic_area text	indicatorproportion text	service_type text	residence text	time_period numeric	observation numeric
1	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Rural	2017.0	38.5269
2	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Rural	2018.0	38.5269
3	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Rural	2019.0	38.5269
4	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Rural	2020.0	38.5269
5	Sudan	Proportion of schools with basic drinking water services	Drinking Water	Rural	2021.0	38.5269
6	Sudan	Proportion of schools with limited drinking water servic...	Drinking Water	Rural	2017.0	22.6629
7	Sudan	Proportion of schools with limited drinking water servic...	Drinking Water	Rural	2018.0	22.6629
8	Sudan	Proportion of schools with limited drinking water servic...	Drinking Water	Rural	2019.0	22.6629
9	Sudan	Proportion of schools with limited drinking water servic...	Drinking Water	Rural	2020.0	22.6629
10	Sudan	Proportion of schools with limited drinking water servic...	Drinking Water	Rural	2021.0	22.6629
11	Sudan	Proportion of schools with no drinking water service	Drinking Water	Rural	2017.0	38.8102
12	Sudan	Proportion of schools with no drinking water service	Drinking Water	Rural	2018.0	38.8102
13	Sudan	Proportion of schools with no drinking water service	Drinking Water	Rural	2019.0	38.8102
14	Sudan	Proportion of schools with no drinking water service	Drinking Water	Rural	2020.0	38.8102
15	Sudan	Proportion of schools with no drinking water service	Drinking Water	Rural	2021.0	38.8102

Total rows: 139 of 139 Query complete 00:00:00.176

Data Output | Messages | Notifications

	country text	residence_type text	service_type text	time_period numeric	coverage numeric	population numeric	service_level text
1	Sudan	total	Sanitation	2018.0	36.6727	15402187.7288	At least basic
2	Sudan	rural	Sanitation	2018.0	24.2341	6652198.0443	At least basic
3	Sudan	urban	Sanitation	2018.0	60.1402	8749989.9766	At least basic
4	Sudan	total	Drinking water	2018.0	60.4671	25395606.3873	At least basic
5	Sudan	rural	Drinking water	2018.0	54.4624	14949776.6469	At least basic
6	Sudan	urban	Drinking water	2018.0	71.796	10445830.0651	At least basic
7	Sudan	total	Hygiene	2018.0	16.0087	6723507.143	Basic service
8	Sudan	total	Hygiene	2018.0	26.3621	11071824.6133	Limited service
9	Sudan	total	Sanitation	2018.0	8.3345	3500420.6191	Limited service
10	Sudan	rural	Sanitation	2018.0	6.3995	1756646.2016	Limited service
11	Sudan	urban	Sanitation	2018.0	11.9853	1743774.4742	Limited service
12	Sudan	total	Drinking water	2018.0	26.8768	11287982.1623	Limited service
13	Sudan	rural	Drinking water	2018.0	28.1097	7716036.838	Limited service
14	Sudan	urban	Drinking water	2018.0	24.5506	3571945.4265	Limited service
15	Sudan	total	Hygiene	2018.0	57.6292	24203726.508	No handwashing facility

Total rows: 103 of 103 Query complete 00:00:00.109

3.0 DATABASE

3.1 Relational Model and Relationship between Data

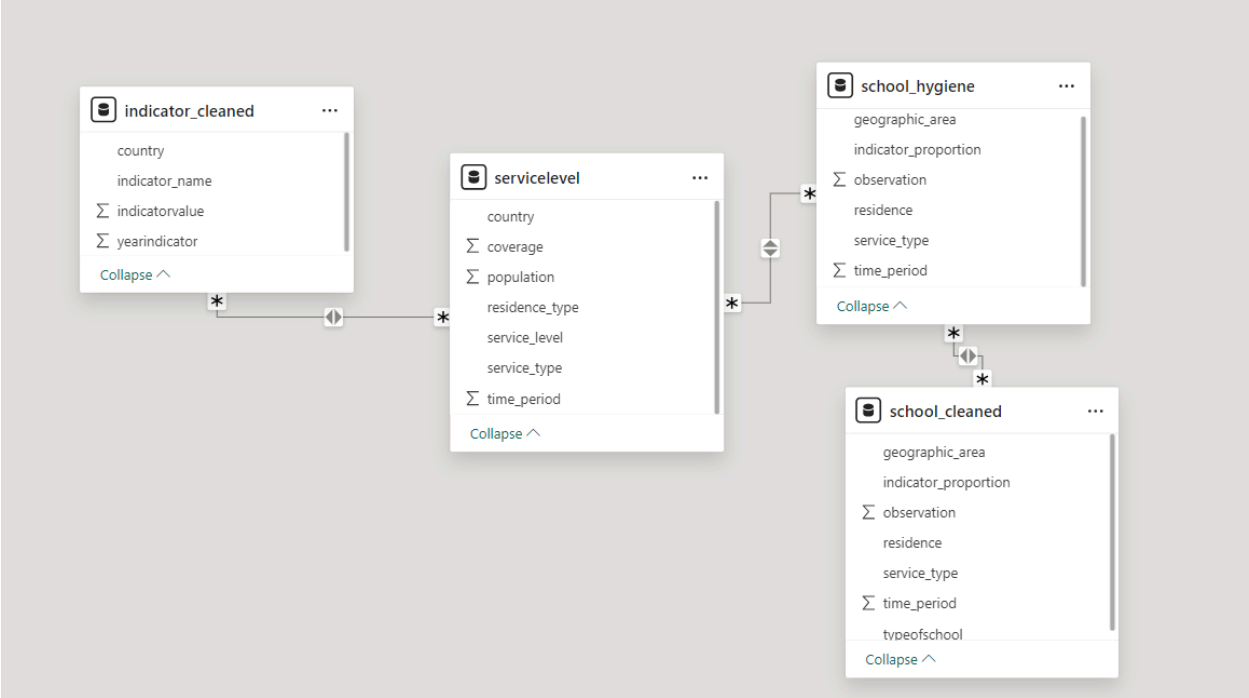


Figure 3.1 Relational Model using Power BI

3.2 Relationship between Data

DATA	RELATIONSHIP
Indicator_cleaned -> servicelevel	Many to many
Servicelevel -> school_hygiene	Many to many
School_hygiene -> school_cleaned	Many to many

3.3 Identification of Data Warehouse Schema

The data warehouse schema for these datasets is snowflake schema, as seen in Figure 3.1 above. A snowflake schema avoids redundancy but increases schema complexity by normalizing dimensions into several linked tables. There is a many-to-many relationship between the Servicelevel table and the Indicator_cleaned table. In a similar way, Servicelevel has many-to-many relationships with School_hygiene and School_hygiene with School_cleaned. The structure that is produced by this normalisation has dimension tables connected to the main fact table and to each other, like a complex web.

4.0 RESULTS AND DATA ANALYSIS

4.1 OLAP Coding

Slicing Operator

```
SELECT residence, servicetype, timeperiod, MAX(observation) AS max_observation,
typeofschool

FROM school_cleaned

WHERE timeperiod = 2019 AND observation > 75

GROUP BY residence, servicetype, timeperiod, typeofschool;
```

Output:

	residence text	servicetype text	timeperiod numeric	max_observation numeric	typeofschool text
1	Rural	Hygiene	2019.0	78.7597	Total

Interpretation:

By using slicing operators, the output shows there is only one data that has an observation above 75 which is 78.7595 from rural areas in 2019 where the service type is hygiene. The observation value indicates that the proportion of schools with limited hygiene services amounts to 78.7597. This suggests that there has been stagnant progress and a huge gap to keep up with the maximum observation. This also suggests that development and improvement must be uniform so it can be accessed to all.

Dicing Operator

SELECT geographic_area, service_type, residence, time_period, MAX(observation) as max_observation

FROM school_hygiene

WHERE service_type = 'Drinking Water'

AND residence = 'Rural'

GROUP BY geographic_area, service_type, residence, time_period

ORDER BY time_period;

Output:

	geographic_area text	service_type text	residence text	time_period numeric	max_observation numeric
1	Sudan	Drinking Water	Rural	2017.0	38.8102
2	Sudan	Drinking Water	Rural	2018.0	38.8102
3	Sudan	Drinking Water	Rural	2019.0	38.8102
4	Sudan	Drinking Water	Rural	2020.0	38.8102
5	Sudan	Drinking Water	Rural	2021.0	38.8102

Intepretation:

By using the dicing operator, the output shows the constant value observation of 38.8102 throughout 5 years ranging from 2017 to 2021 for drinking water service in the rural area. It shows that the rural area only received a constant value of drinking water service without an increase in value despite the possibility that the population kept on growing every year. This suggests that there has not been any effort or progress to improve the drinking water service.

Roll Up Operator

SELECT residence_type, service_type, time_period, AVG(coverage) as avg_coverage, AVG(population) as avg_population

FROM servicelevel

GROUP BY ROLLUP (residence_type, service_type, time_period)

ORDER BY avg_coverage desc;

Output:

	residence_type text	service_type text	time_period numeric	avg_coverage numeric	avg_population numeric
1	total	Hygiene	2018.0	33.3333333333333333	13999686.088100000000
2	rural	Sanitation	2019.0	25.0000250000000000	7032132.500000000000
3	total	Drinking water	2019.0	25.0000250000000000	10808023.250025000000
4	total	Drinking water	2022.0	25.0000250000000000	11718551.000000000000
5	rural	Drinking water	2020.0	25.0000250000000000	7193470.500000000000

Interpretation:

From the roll up operator, the output shows that the highest average coverage is 33.33 from an average population of 13.99 million. In 2019, rural areas had a sanitarian service coverage of 25.00 with a population average of 7.03 million. In the same year, the drinking water service had a similar coverage of 25.00 on the total residence, with an average population of 10.81 million. By 2020 and 2022, the coverage of drinking water will stay the same at 25.00, but the average population will be 11.72 million and 7.19 million, respectively.

Group by Cube Operator

SELECT servicetype, typeofschool, timeperiod, AVG(observation) as avg_observation

FROM school_cleaned

WHERE typeofschool = 'Primary schools'

GROUP BY CUBE (servicetype, typeofschool, timeperiod)

HAVING servicetype IS NOT NULL AND typeofschool IS NOT NULL AND timeperiod IS NOT NULL

ORDER BY timeperiod desc;

Output:

	servicetype text	typeofschool text	timeperiod numeric	avg_observation numeric
1	Drinking Water	Primary schools	2021.0	33.3333333333333333
2	Hygiene	Primary schools	2021.0	33.3333333333333333
3	Sanitation	Primary schools	2021.0	28.1000000000000000
4	Hygiene	Primary schools	2020.0	33.3333333333333333
5	Drinking Water	Primary schools	2020.0	33.3333333333333333
6	Sanitation	Primary schools	2020.0	28.1000000000000000
7	Drinking Water	Primary schools	2019.0	33.3333666666666667
8	Hygiene	Primary schools	2019.0	33.3333333333333333
9	Sanitation	Primary schools	2019.0	28.1000000000000000
10	Sanitation	Primary schools	2018.0	28.1000000000000000
11	Hygiene	Primary schools	2018.0	33.3333333333333333
12	Drinking Water	Primary schools	2018.0	33.3333666666666667

Intepretation:

From the group by cube operator, we can see that across the years, the observation of drinking water and hygiene service is consistently at 33.33. On the other hand, sanitation service is consistently at 28.19 across the years. This indicates that the drinking water and hygiene service have a stable coverage in primary school, while sanitation service lag behind. This suggests that there is a need for an improvement for all services, especially sanitation services to ensure better coverage of essential services for children in primary school.

4.2 Data Visualisation

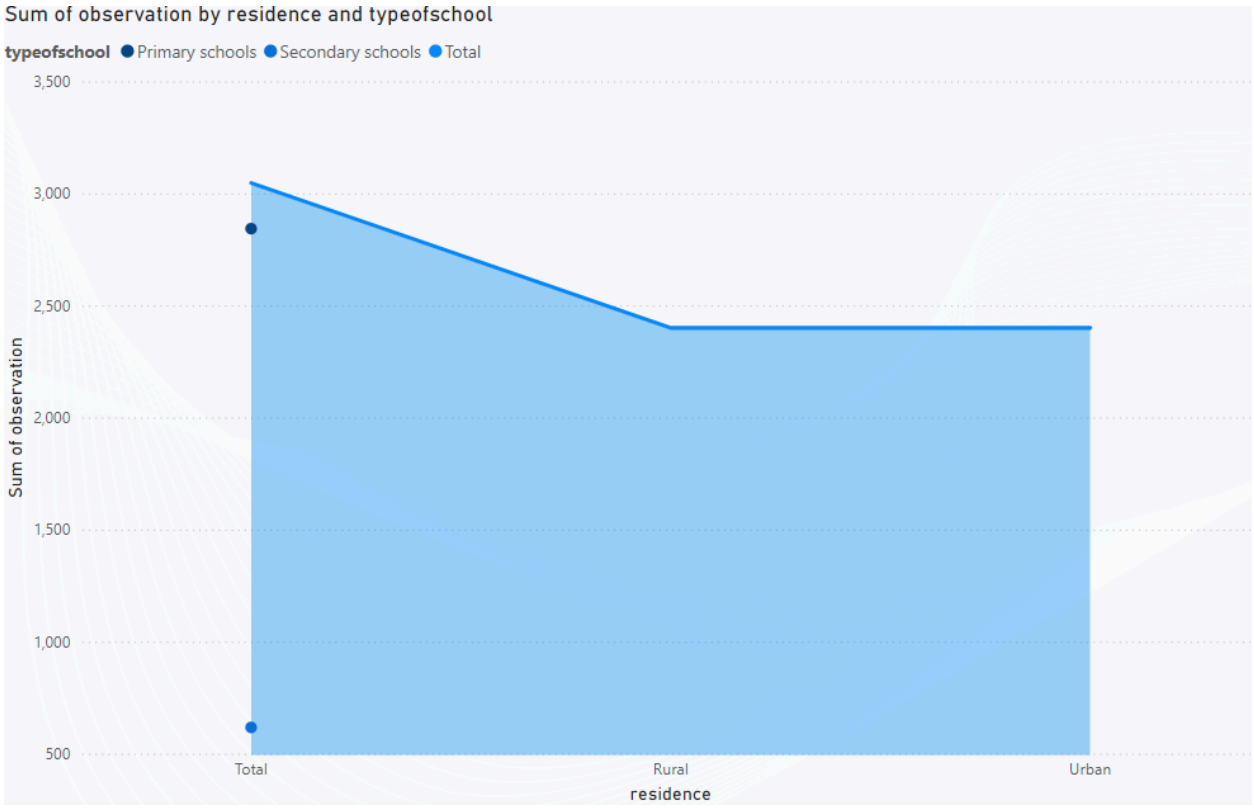


Figure 4.1.1 Sum of Observation by Residence and typeschool Area Chart

The region chart shows that primary schools with 2842.65 observations make up the majority of observations, far more than secondary schools with 617.20 observations, suggesting that primary education is given more priority. With 2400.00 observations in each, the distribution of observations between rural and urban locations implies that both regions are receiving equal attention. The sum for both urban and rural areas combined which is 4800.00 is more than the total number of observations of 3046.69, suggesting a possible contradiction in the data that requires further analysis. The prioritization of early education is evident in both rural and urban settings, as seen by the equal regional distribution and significant emphasis on elementary education. However, additional examination is necessary due to the inconsistent statistics.

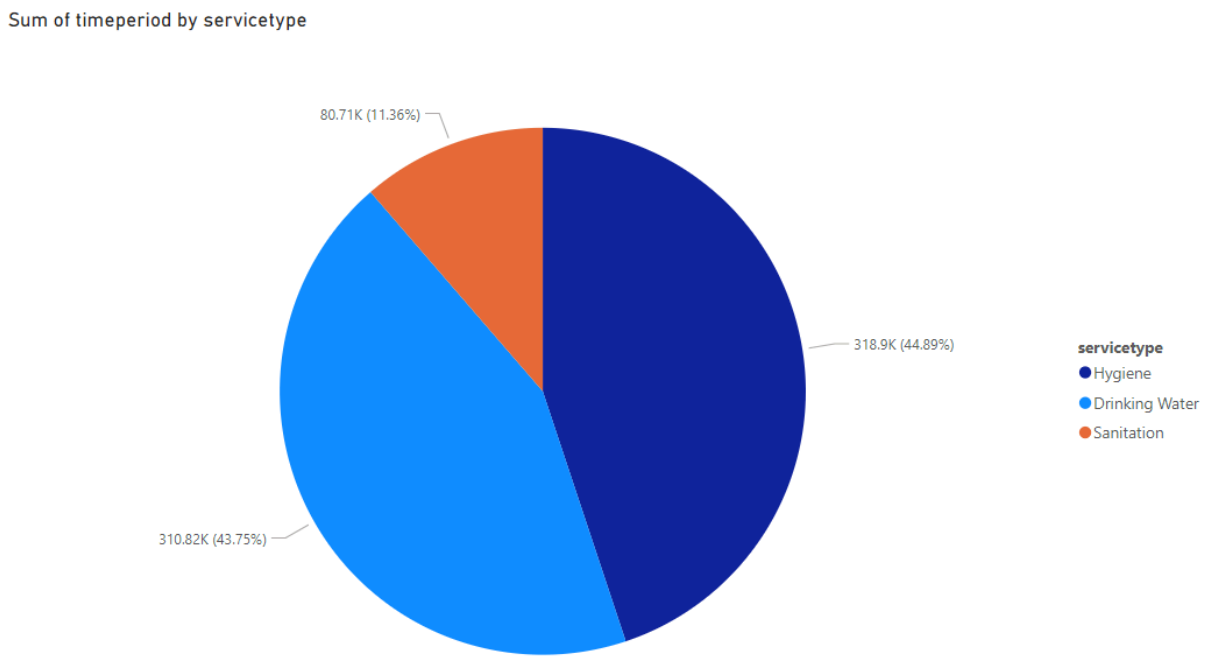


Figure 4.1.2 Sum of Timeperiod by Servicetype Pie Chart

The pie chart illustrates how the time period is allocated, highlighting the importance of drinking water services and hygiene. Hygiene services account for the majority, 318,904 units, or 44.89% of the total time period. In close succession, 310,820 units, or 43.75% of the total time, are assigned to drinking water services. These figures highlight the need of maintaining sanitation and ensuring a sufficient supply of clean drinking water.

Conversely, the allocation for sanitation services is insufficient, amounting to 80,714 units, or 11.36% of the total time. This notable disparity implies that sanitation services receive far less funding and attention than those related to drinking water and hygiene. The distribution suggests that people place a high value on clean, readily available water, but a low value on sanitation. This discrepancy highlights the need for more equitable resource allocation in order to properly address all aspects of public health and sanitation.

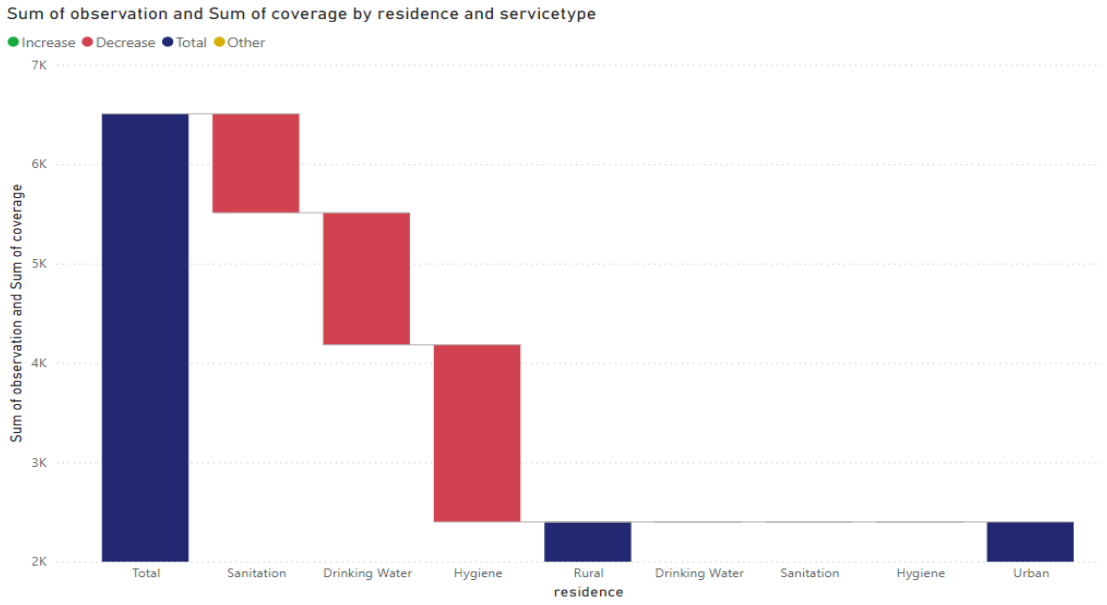


Figure 4.1.3 Sum of Observation and Sum of Coverage by Residence and Servicetype Waterfall Chart

Significant differences between urban and rural locations may be seen in the data from the waterfall chart, which shows the distribution and variations in observations for sanitation, drinking water, and hygiene services among different dwellings. There are 6506.54 observations in total, and the coverage sum is 2548.30. There is a significant general decline in sanitation services, as evidenced by a change in observations of -995.00, which suggests a decline in their availability or quality. While there has been no change in the coverage of drinking water services, there has been a notable loss of -1329.14 observations, or a 52.55% decrease, in rural areas. Overall, these observations stand at 2529.14. With 2982.40 observations overall, hygiene services likewise exhibit no change in coverage; but, in rural regions, there is a significant reduction of -1782.40 observations, or 59.76%.

The substantial decreases in rural observations for both drinking water and hygiene services despite an equal distribution of observations and coverage between urban and rural areas (both with an observation sum of 2400.00 and a coverage sum of 2548.30) highlight differences in service availability and quality. While services are significantly reduced in rural areas, they are maintained more steadily in urban areas. This emphasizes how critical it is to address service delivery disparities and guarantee that rural communities receive sufficient funding for basic

sanitation, drinking water, and hygiene services. The fact that sanitation services are generally declining highlights how critical it is to improve these services in order to improve living conditions and public health.

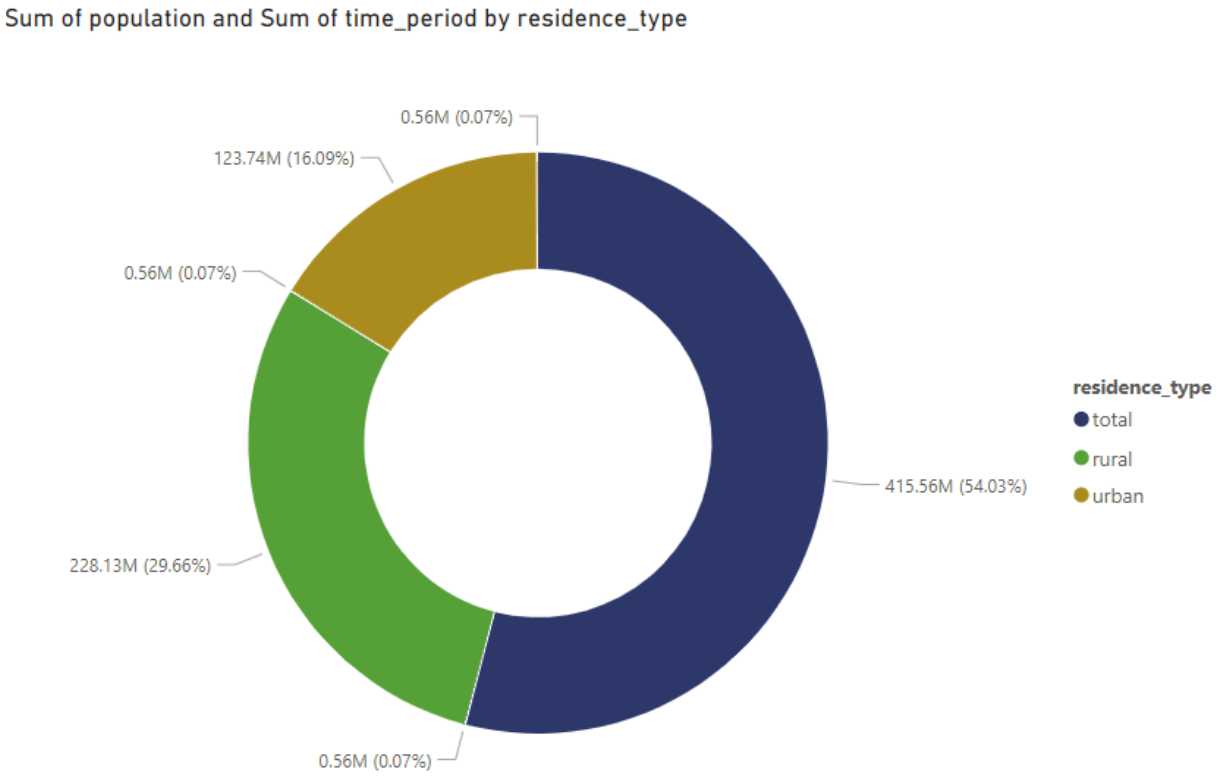


Figure 4.1.4 Sum of Population and Sum of Timeperiod by Residencetype Donut Chart

The donut chart's figures show that 228,130,864 people, or 29.66% of the population, live in rural areas, while 123,743,824 people, or 16.09% of the population, live in urban areas. According to this data, 415,557,483.66 people, or a significant share of the population which is 54.03%, reside in clearly defined urban and rural areas. The data indicates a higher concentration of individuals living outside of urban centres, with rural areas appearing to have a considerably larger population than metropolitan areas.

Nonetheless, the aggregate populace of urban and rural regions scarcely surpasses 50% of the overall populace, suggesting that a substantial portion of the populace is not categorized as either rural or urban. This may include those who are transitory, live in non-traditional housing arrangements, or reside in peri-urban locations. In order to ensure that all demographic groupings are taken into account in policy and development strategies, the data highlights the necessity of taking these unclassified segments into consideration when planning and allocating resources.

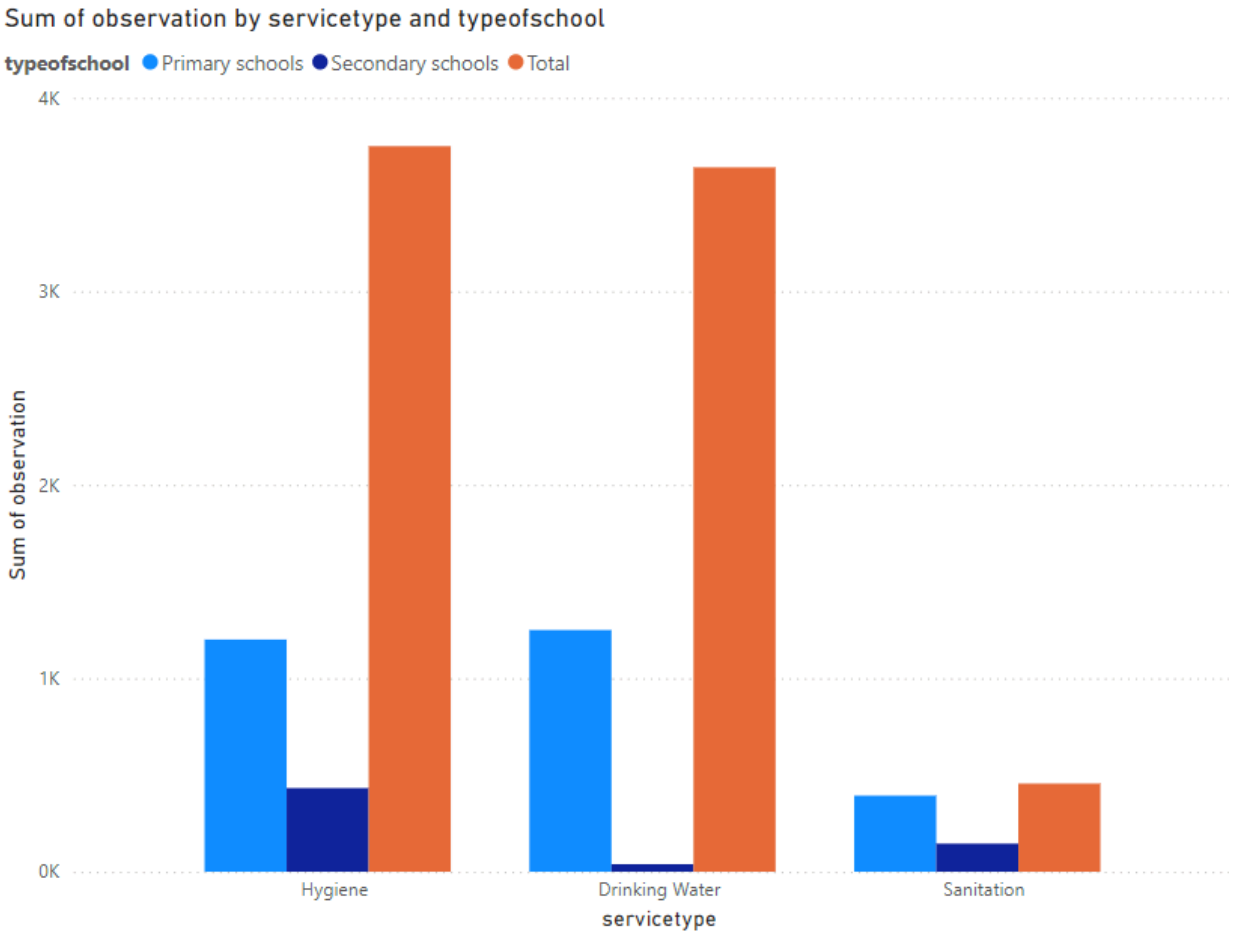


Figure 4.1.5 Sum of Observation by Servicetype and Typeofschool

Hygiene services had the most observations overall which is 3750.04, with primary schools receiving the most attention of 1200.00 observations and senior schools having 432.36 observations, according to the clustered column chart. There are also a lot of observations at 3640.65 for drinking water services, mainly from primary schools with 1249.25 observations,

while secondary schools only have a minimum of 39.25 observations. According to this statistics, drinking water and sanitary facilities are given top priority, particularly in primary schools where younger pupils have greater access to these vital resources.

In sharp contrast, elementary schools account for 393.40 observations and secondary schools for 145.60 observations, meaning that sanitation services have the fewest total observations of 456.00. This discrepancy suggests that compared to drinking water and hygiene services, sanitation services receive significantly less attention. The emphasis on drinking water and hygiene services in primary schools highlights the value of these resources for younger students, but the noticeably less attention paid to sanitation services points to a crucial area that has to be improved to guarantee students' overall health and wellbeing in educational settings.

5.0 CONCLUSION

According to Sudan's water and hygiene data, primary schools are given precedence when it comes to receiving clean water and hygiene services; secondary schools and sanitation services receive noticeably less observations. This implies that younger students have more access to these basic utilities, while services and sanitation for older students are conspicuously neglected. While the importance of primary education and basic services for younger children is acknowledged, there is an urgent need to improve sanitary services at all educational levels.

Rural areas had far fewer observations of drinking water and hygiene facilities, even when the population distribution in each is the same. This disparity in the provision of services indicates an inequity that needs to be addressed right away in order to guarantee fair and consistent access. Improving rural service delivery and sanitation facilities are crucial first steps towards resolving these discrepancies. Providing enough drinking water, sanitation, and hygiene assistance to every area, especially rural ones is essential for the general health and welfare of the populace.

6.0 References

Richman, J. (2024, January 11). What is a database schema? Types, use cases, & examples. *Estuary*. https://estuary.dev/database-schema/

THE 17 GOALS | Sustainable Development. (n.d.). https://sdgs.un.org/goals

UNESCO, WHO. (2018). *Drinking water, sanitation and hygiene in schools : global baseline report 2018*. http://repositorio.minedu.gob.pe/handle/20.500.12799/5941

Taylor, D. (2023, December 30). *What is OLAP? Cube, Analytical Operations in Data Warehouse*. Guru99. https://www.guru99.com/online-analytical-processing.html

GeeksforGeeks. (2024a, February 19). *Power BI Tutorial Learn Power BI*. GeeksforGeeks. https://www.geeksforgeeks.org/power-bi-tutorial/

GeeksforGeeks. (2023a, February 2). *ETL process in data warehouse*. GeeksforGeeks. https://www.geeksforgeeks.org/etl-process-in-data-warehouse/

Babic, T. (2020, December 3). *What Is a Many-to-Many Relationship in a Database? An Explanation with Three Examples*. Vertabelo Data Modeler. https://vertabelo.com/blog/many-to-many-relationship/

GeeksforGeeks. (2024d, April 2). *Relationships in SQL One-to-One, One-to-Many, Many-to-Many.* GeeksforGeeks. https://www.geeksforgeeks.org/relationships-in-sql-one-to-one-one-to-many-many-to-many/

Informatica. (n.d.). Informatica. https://www.informatica.com/resources/articles/what-is-etl-pipeline.html