Brief for Level 5 Data Engineer - Formative Piece 1

Data Quality Assessment and Improvement Report

Prepared by: Paul Todd , November 2024

Introduction

Presented with an Excel file containing three tabs of Health facility data for a particular region.

Study objectives

Assess the data set for quality create a Schema using SQL as a DDL.

Data Quality

I took time to familiarise myself with the data before starting to look at the accuracy, completeness, consistency and uniqueness. Timeliness would need further discussions with the client/provider to determine how frequently the data is updated and how often any new analysis would be required.

TAB1: Mwanza\_r

Data is organized into a single table.

Extract Figure 1, analysis Figure 2

A screenshot of a computer

Description automatically generated

Figure 1Mwanza R table extract.

A screenshot of a computer

Description automatically generated

Figure 2 Mwanza R data analysis.

Tab 2: Pwani\_r

3 tables of data, assumed one table per Facility type.

Assumption based on the ‘name of the facilities’.

Table 3 values all contain ‘Hospital’.

Table 1, 83 entries contain ‘Dispensary.

Therefore, Table 2 would be ‘Health Centre’ as per the three types of facility described in Mwanza\_r HF TYPE.

Figure 3 shows an extract from Table 1, including formulas in columns D, E and F to determine duplicate values and the frequency of the word ‘Dispensary’ in the ‘NAME of FACILITY’.

A screenshot of a computer

Description automatically generated

Figure 3 Pwani r data analysis formula examples from table 1.

Figure 4 shows my analysis.

A screenshot of a computer

Description automatically generated

Figure 4 Pwani r data analysis summary

2nd Table extract in Figure 5, analysis Figure 6

A screenshot of a computer

Description automatically generated

Figure 5 Pwani r data table 2.

A screenshot of a spreadsheet

Description automatically generated

Figure 6 Pwani r table 2 analysis summary.

Figure 7 shows an extract of the third table, analysis figure 8.

A white paper with black text

Description automatically generated

Figure 7 Pwani R table 3 extract.

A white sheet with black text and numbers

Description automatically generated

Figure 8 Pwani r table 3 data summary.

Tab 3 DSM (iala)\_r

There are three tables, each split and displayed side by side (Figure 9). Each table is dedicated to a Facility type having 4 Headings.

A table with text on it

Description automatically generated

Figure 9 DSAM (Iala)\_r data table 1 extract.

I split out the data in the tables and concatenated into one non repeated table.

Table 1 Extract - Figure 10, Analysis Figure 11

A close-up of a sign

Description automatically generated

Figure 10 DSAM (Iala)\_r data concatenated into a linear table.

A white rectangular object with black text

Description automatically generated

Figure 11 DSAM (Iala)\_r data analysis summary.

Table 2 extract figure 12, analysis figure 13

A screenshot of a document

Description automatically generated

Figure 12 DSAM (Iala)\_r table 2 data extract.

A screenshot of a document

Description automatically generated

Figure 13 DSAM (Iala)\_r table 2 data analysis summary.

Table 3 extract figure 14, analysis figure 15

A screenshot of a computer

Description automatically generated

Figure 14 DSAM (Iala)\_r table 3 data extract.

A white sheet of paper with black text

Description automatically generated

Figure 15 DSAM (Iala)\_r table 3 data analysis summary.

Data Cleansing

Across all tabs, there are five common data points.

* Facility name– Text, max length 58
* Star Level – Integer - Assume that regions use the same rating system with integer values 0-4.
* Facility type- Text. 3 values Health Centre, Hospital and Dispensary
* Ownership (missing from tab 2) - Text, Private and Public. Missing values to be mapped as “Unknown”.
* Region – Text, derived from the tab name.

Using these 5 data points, I looked to standardise each table into a common format, cleansing the data of null lines, invalid data and merged columns.

Cleansing steps detailed in Appendix.

Cleansed Data from tab 1,2 and 3 in figures 16, 17 and 18 respectively

A table with a list of names

Description automatically generated

Figure 16 Cleansed data from Mwanza r tab.

A table with names and numbers

Description automatically generated

Figure 17 Cleansed data from Pwani r tab.

A table with a number of names

Description automatically generated

Figure 18 Cleansed data from DSM (Iala) r tab.

Schema Creation

The concatenated table can be used as a single ‘Fact table’ in a database.

Using the following code (Figure 19) to create the table schema:

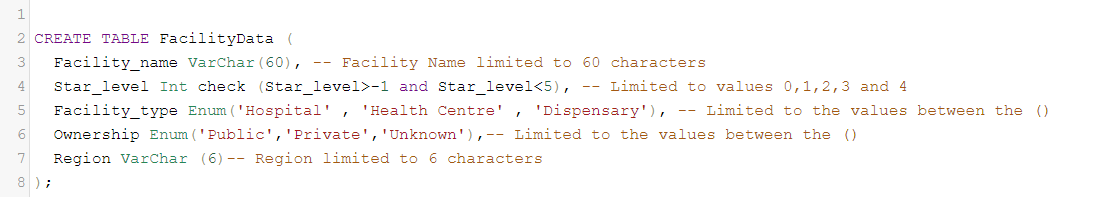


Figure 19 SQL code to create the table schema.

And assessed by inserting 14 records (Figure 20),

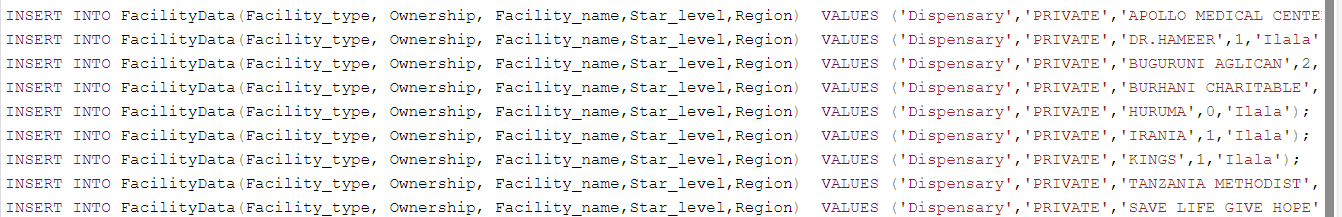


Figure 20 Testing the schema and data by inserting a sample data set.

Results in figure 21

A screenshot of a computer

Description automatically generated

Figure 21 Table result.

Reflections

For regular ingestion of data, I would recommend to the data owners:

* + - 1. Complete the Ownership details for Pwani\_r.
      2. Cleanse duplicates and erroneous values noted.
      3. Change the Star level from text to Integer, allowing for calculations such as average by type. Ask the data owners if the Star Rating is consistently calculated across the 3 regions and recommend a common method if not.
      4. Rather than mapping Tab 3’s Ownership prior to loading, a table could be created in the database and a relationship matching the original Ownership data to the Public/Private values in the mapping table Appendix (figure 24).
      5. Suggest to data owners that each facility be given a unique identifier that could be used to create or update records as and when data is refreshed. This could be used in the SQL tables as a primary key enhancing data integrity by allowing for 2 facilities with the same name but perhaps different Facility\_types.

Appendix

A screenshot of a computer

Description automatically generated

Figure 22 Dictionary definition of parastatal.

A close-up of a text

Description automatically generated

Figure 23 Google search result for FBO.

Parastatal refers to government owned (figure 22) and will be mapped to Public.

FBO refers to Faith Based organisation (Figure 23) and will be mapped to Private.

A table with text on it

Description automatically generated

Figure 24 Proposed ownership mapping.

Schema and Test SQL code

https://sqlfiddle.com/mysql/online-compiler?id=f85ca6c7-a14c-457d-af4d-3fe640f59536

CREATE TABLE FacilityData (

Facility\_name VarChar(100), -- Facility Name limited to 100 characters

Star\_level Int check (Star\_level>-1 and Star\_level<5), -- Limited to values 0,1,2,3 and 4

Facility\_type Enum('Hospital' , 'Health Centre' , 'Dispensary'), -- Limited to the values between the ()

Ownership Enum('Public','Private','Unknown'),-- Limited to the values between the ()

Region VarChar (20)-- Region limited to 20 characters

);

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','APOLLO MEDICAL CENTER',1,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','DR.HAMEER',1,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','BUGURUNI AGLICAN',2,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','BURHANI CHARITABLE',1,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','HURUMA',0,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','IRANIA',1,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','KINGS',1,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','TANZANIA METHODIST',0,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','SAVE LIFE GIVE HOPE',1,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','MADONA',2,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','PRINCE SAUDI',2,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','SUNSHINE MUSLIM VOLUNTEER SMV',2,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','TIKAYA',3,'Ilala');

INSERT INTO FacilityData(Facility\_type, Ownership, Facility\_name,Star\_level,Region) VALUES ('Dispensary','PRIVATE','EBRAHIM HAJI',2,'Ilala');

select \* FROM FacilityData;

Data Cleansing Steps:

1. Tab 1

Replace HC, DISP and HOSP with Health Centre, Dispensary and Hospital in HF Type

Remove duplicate lines from HF Name, remove line with value ??Bot in HFName

Remove ‘-Star’ from Star Rating

Replace 1 with Public and 2 with Private in Ownership column.

Remove columns SRA Code, HFR Code

1. Tab 2
   * 1. Table 1
        + 1. Add a column called Facility type. Enter Dispensary in each field.
          2. Add a column ‘Ownership’ and enter ‘Unknown’ in each field.
          3. Remove ‘-Star’ from Star Rating
          4. Remove duplicate lines in Name of Facility
          5. Remove 1st column with unique code (no header)
     2. Table 2
        + 1. Add a column called Facility type. Enter Health Centre in each field.
          2. Add a column ‘Ownership’ and enter ‘Unknown’ in each field.
          3. Remove ‘-Star’ from Star Rating
          4. Remove duplicate lines in Name of Facility
          5. Remove blank line.
          6. Remove SN column.
     3. Table 3
        + 1. Add a column called Facility type. Enter Hospital in each field.
          2. Add a column ‘Ownership’ and enter ‘Unknown’ in each field.
          3. Remove ‘-Star’ from Star Rating
          4. Remove blank lines.
          5. Remove SN column.
     4. Concatenate all three tables.
2. Tab 3
   1. Table 1
      1. Spilt side by side table and concatenate into one table of four columns.
      2. Add a column called Facility type. Enter Hospital in each field.
      3. Insert a column called Ownership Mapping. Map values into this column as Private or Public (See appendix figure 23 for mapping table)
      4. Remove SN column.
   2. Table 2
      1. Spilt side by side table and concatenate into one table of four columns.
      2. Add a column called Facility type. Enter Health Centre in each field.
      3. Insert a column called Ownership Mapping. Map values into this column as Private or Public (See appendix figure 23 for mapping table)
      4. Unmerge column Star Level
      5. Remove SN column.
   3. Table 3
      1. Spilt side by side table and concatenate into one table of four columns.
      2. Add a column called Facility type. Enter Dispensary in each field.
      3. Insert a column called Ownership Mapping. Map values into this column as Private or Public (See appendix figure 23 for mapping table)
      4. Remove SN column.
      5. Unmerge all columns.
      6. Concatenate all three tables.
      7. Delete Ownership Column
3. Concatenate all 3 Tab data into one data set, lining up the relevant columns.