SAS Data Quality Monitoring

workshop exercises

Paul Van Mol

Contents

[1. Read Breakfast\_Items and Manufacturers with an import 3](#_Toc184766365)

[2. Add 2 discovery Agents as an SAS Viya Administrator for SASDM Compute and Public Caslib. 4](#_Toc184766366)

[3. Go to Discover Information Assets as Alex and search for the Breakfast\_Items table. 6](#_Toc184766367)

[4. Create Ruleset to check Validity of UPC (Uniform Product Code) in Breakfast\_Items: 8](#_Toc184766368)

[5. Create Ruleset to check Accuracy of UOM (Unit of Measure) 11](#_Toc184766369)

[6. Create a Monitoring Task: Mon\_Breakfast\_items 14](#_Toc184766370)

[7. Appendix 16](#_Toc184766371)

[Prepare Demo on Training Virtuallab 16](#_Toc184766372)

# Read Breakfast\_Items and Manufacturers with an import

Navigate to the Folder: /gelcontent/dqdashboard/data/

Add the BREAKFAST\_ITEMS.xlsx, then add an import step.

Analyze the columns and store the result in a table SASDM.BREAKFAST\_ITEMS

Add the MANUFACTUERS.xlsx, then add an import step.

Analyze the columns and store the result in a table SASDM.MANUFACTURERS

A screenshot of a computer

Description automatically generated

In order to check the values in BREAKFAST\_ITEMS: UPC, UOM and BRAND, add a One-Way Frequencies Step to the Breakfast\_Items table:

A screenshot of a computer

Description automatically generated

In options unselect include cumulative frequencies and percentages:

In order to check the values in MANUFACTURERS fields: COUNTRY, STATE\_PROV, CONTACT\_CNTRY, CONTACT\_STATE\_PROV add a One-Way Frequencies Step to the MANUFACTURERS table:

A screenshot of a computer

Description automatically generated

Finally, Standardize the COUNTRY and STATE\_PROV fields in the MANUFACTURERS table:

Using a Clean Data step:

A screenshot of a computer

Description automatically generated

# Add 2 discovery Agents as an SAS Viya Administrator for SASDM Compute and Public Caslib.

Login to SAS Viya as sasadm & lnxsas.

Go to Discover Information Assets

Select Discovery Agents

A screenshot of a search box

Description automatically generated

Select new Discovery Agent and choose the compute library SASDM.

A screenshot of a computer

Description automatically generated

No changes are required in the configuration:

A screenshot of a computer

Description automatically generated

Select Run Now. This will save the discovery agent and run the Discovery Agent.

Same can be done for the PUBLIC Caslib.

# Go to Discover Information Assets as Alex and search for the Breakfast\_Items table.

Search for the Breakfast\_Items table:

A screenshot of a computer

Description automatically generated

Select the Table and look at the Overview page:

A screenshot of a computer

Description automatically generated

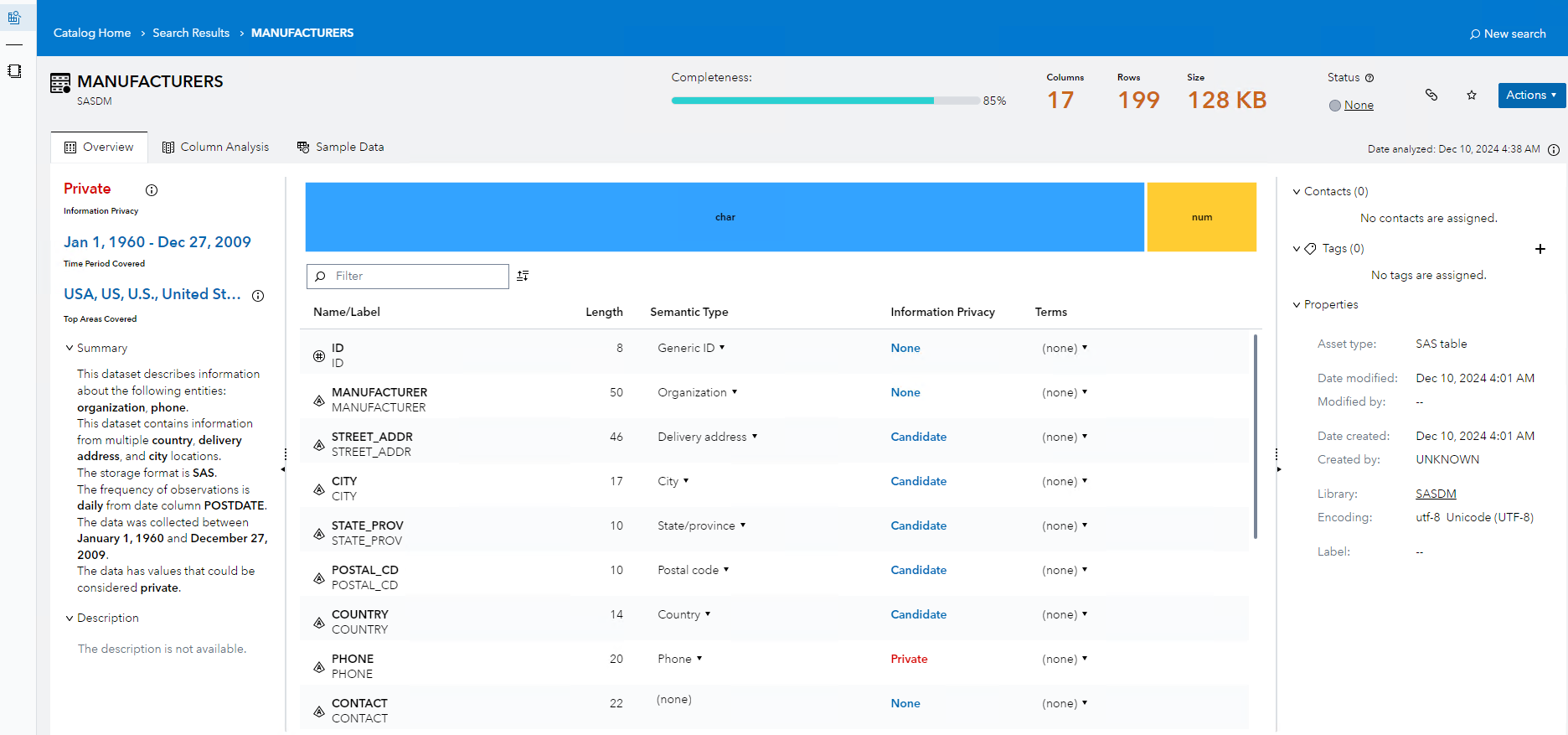
Select column analysis:

Investigate the columns UPC, UOM and BRAND.

* What is the completeness of each field?
* For UPC and UOM what is the most common Pattern?

Close the Breakfast\_items table

Now search for the Manufacturers table:



In the column analysis, investigate the completeness and Frequency Distribution of different fields:

What is the completeness of Street\_addr, STATE\_PROV, COUNTRY

# Create Ruleset to check Validity of UPC (Uniform Product Code) in Breakfast\_Items:

1. create Validity\_UOM Ruleset:

Duplicate the Ruleset Validity\_Phone and save as Validity\_UOM

* Rename Variables isAccuracy\_UOM (Boolean), Value\_UOM (string)

A screenshot of a computer

Description automatically generated

On the Rule Set tab:

* Rule: Validity\_UPC
* Dimension: Validity
* Rule\_Description: Checking type and length of UPC
* Expression:

Assign: isValidity\_UPC True

Change the custom expression to the expression below. It uses a Regex to check if not numbers are found in the Value\_UPC and if the length is not 14 long.

Value\_UPC= STRIP(Value\_UPC)

re= prxparse('/^\d+$/');

if ^prxmatch(re,Value\_UPC) and length(Value\_UPC) ^= 14 then do;

isValidity\_UPC= 0;

end;

A screenshot of a computer

Description automatically generated

1. Create a Decision Field\_UPC

Select Decisions: select Field\_Phone, check the checkbox, then top left menu: duplicate. Rename the Decision to Field\_UPC

A screenshot of a computer

Description automatically generated

Select the variables tab, rename all variables and initial values from Phone to UPC

Remark: sometimes a variable rn also contains an initial value that needs to be renamed:

A screenshot of a computer

Description automatically generated

After the renaming, the variables tab should like like following screen:

A screenshot of a computer

Description automatically generated

On the Decision Flow, remove the Validity\_Phone Ruleset and Add to the No Branch after the Completeness check, the Validity\_UPC Ruleset

A screenshot of a computer

Description automatically generated

Then add to the No branch the Validity\_UPC Ruleset:

A screenshot of a computer

Description automatically generated

Check the mappings of the Input and Output variables. The Value\_UPC in the decision should be mapped to the Value\_UPC in the Ruleset. Same for the isValidity\_UPC.

# Create Ruleset to check Accuracy of UOM (Unit of Measure)

1. Create Lookup Table UOMLookup:

CT=Carat, LB=Pound, OZ=Ounce, PK=Peck

Store the lookup table and Activate it

1. create Accuracy\_UOM Ruleset:

Duplicate the Ruleset Accuracy\_County and save as Accuracy\_UOM

* Rename Variables isAccuracy\_UOM (Boolean), Value\_UOM (string)
* Rule: UOM Check
* Rule\_Description: Checking the values for UOM
* Assign isAccuracy\_UOM False
* Add Rule: if Value\_UOM Lookup UOMLookup THEN then isAccuracy\_UOM=true

A screenshot of a computer

Description automatically generated

1. Create a Decision Field\_UOM

Select Decisions: select Field\_County, check the checkbox, then top left menu: duplicate

Rename the Decision to Field\_UOM.

Select the variable tab, rename all variables and initial values from County to UOM

A screenshot of a computer

Description automatically generated

On the Decision Flow, remove the Accuracy\_County Ruleset and Add to the No Branch after the Completeness check, the Accuracy\_UOM Ruleset

Check the mappings of the Input and Output variables. The Value\_UOM in the decision should be mapped to the Value\_UOM in the Ruleset. Same for the isAccuracy\_UOM.

A screenshot of a computer

Description automatically generated

# Create a Monitoring Task: Mon\_Breakfast\_items

Start with a Duplicate of Mon\_Person Decision.

Delete all the Field\_ decisions from the flow, then delete the variables.

Add Variables from the Field\_UOM and Field\_UPC decisions from /Public/DataManagement/DQDashboard/Monitor Data Quality

A screenshot of a computer

Description automatically generated

On the Decision Flow, add the Field\_UOM and Field\_UPC Decisions from the folder

A screenshot of a computer

Description automatically generated

The resulting Monitor Task Decision flow should look like:

A screenshot of a computer

Description automatically generated

Select Scoring, add a new Test:

Select Breakfast\_items and map the columns UPC and UOM to Value\_UPC and Value\_UOM.

Run the test, check if the test is successful.

A screenshot of a computer

Description automatically generated

# Appendix

## Prepare Demo on Training Virtuallab

To ensure that DQ Dashboard demo works as desired follow the steps below to load the data into memory and to run the Studio Flows to load the DQ Dashboard tables.

1. Fork the following repository: <https://github.com/paulvanmol/dqdashboard.git> to your personal git account
2. Clone the Repository from SAS Studio

Create a new empty folder dqdashboard:

Create a new git profile:

A screenshot of a computer

Description automatically generated

1. Download the required files from the dqdashboard folder:
   1. Packages/DQDashboard.json package from SAS Studio to your Downloads folder.
   2. Data/ADDRESS.xlsx, PERSON.xlsx, DQ\_SCORE\_HIST\_RAW.xlsx
2. Import DQDashboard.json package in SAS Environment Manager
   1. Login as sasadm + lnxsas to have administrative privileges:
   2. Select Manage Environment:
   3. Select Import tab
   4. Import the DQDashboardPublic.json
   5. Map the Target Server to cas-shared-default, Map the Target Library PUBLIC, WORK CAS Library to the Public CAS Library:

A screenshot of a computer

Description automatically generated

* 1. Other tables should map to Compute, Public and WORK libraries:

A screenshot of a computer

Description automatically generated

* 1. Check if Import is successful:

A screenshot of a computer

Description automatically generated

Upload the data to SAS Studio Explorer:

Upload the data of the dqdashboard repository to the home directory of the student:

* Either by doing a Git Clone in SAS Studio to the /gelcontent/dqdashboard folder
* Or by using the upload Button in the Explorer:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Go to SAS Studio (Develop Code and Flows)

1.1. Go to location: SAS Content/GTPPub/Data Management/DQ Dashboard/SAS Programs

1.2. Open and run SAS job: Load tables into memory.sas

cas casauto;

%let outcaslib=PUBLIC;

%let incaslib=GTPPUB;

%let path=/gelcontent/dqdashboard;

proc cas;

session casauto;

table.dropcaslib /caslib="&incaslib" quiet=true;

table.addCaslib /

caslib="&incaslib"

description="Monitor data"

dataSource={srctype="path"}

path="&path/data";

table.dropTable /

caslib= "&outcaslib",

name= "ADDRESS",

quiet= True;

run;

table.loadTable /

caslib= "&incaslib",

path="ADDRESS.xlsx",

casout={

caslib="&outcaslib",

promote= True

};

run;

table.dropTable /

caslib= "&outcaslib",

name= "PERSON",

quiet= True;

run;

table.loadTable /

caslib= "&incaslib",

path="PERSON.xlsx",

casout={

caslib="&outcaslib",

promote= True

};

run;

table.dropTable /

caslib= "&outcaslib"

name= "DQ\_SCORE\_HIST\_RAW",

quiet= True;

run;

table.loadTable /

caslib= "&incaslib",

path="DQ\_SCORE\_HIST\_RAW.xlsx",

casout={

caslib="&outcaslib",

promote= True

};

run;

quit;

1.3. Go to location: SAS Content/Public/Data Management/DQ Dashboard/Dashboard Flows

1.4. Open and run flows:

* Monitor\_Person.flw
* Monitor\_Address.flw
* Write\_Dashboard\_Tables.flw

2. Go to SAS Drive (Share and Collaborate)

2.1. Open Dashboard: DQ Dashboard in location: SAS Content/Public/Data Management/DQ Dashboard/Dashboard to ensure the Dashboard data got generated correctly.