Excel converter for Jihyun

GSA project

December 13, 2015

Contents

1 Introduction

The document records the implementation of the data converter from excel to csv for Jihyn.

2 Definition of the mapping from source to target format

Via the discussion in the first meeting, the mapping from source PM table (containing energy consumption information) to the target EUAS template is depicted in Table 1: As described

Table 1: Excel merging script field mapping table

Source sheet name	field name	target	Condition
Properties	State/Province	State	
Properties	Gross Floor Area	GSF	
Properties	Year Built	Year Built	
Properties	Postal Code	Postal Code (first 5 digit)	
Property Use	Self-Selected Primary Function	Use Type	
Meter Consumption Data	Property Name	Building ID	
Meter Consumption Data	End Date (month)	Month	
Meter Consumption Data	End Date (year)	Year	
Meter Consumption Data	(Usage/Quantity, Meter Type)	Elec Amt	if Meter Type = Electric - Grid
		Gas Amt	if Meter Type = Natural Gas
		Oil Amt	if Meter Type = Fuel Oil (No. 2)
		Water Amt	if Meter Type = Potable: Mixed Indoor/Outdoor
Meter Consumption Data	(Usage/Quantity, Meter Type)	Elec Cost	if Meter Type = Electric - Grid
		Gas Cost	if Meter Type = Natural Gas
		Oil Cost	if Meter Type = Fuel Oil (No. 2)
		Water Cost	if Meter Type = Potable: Mixed Indoor/Outdoor
Meter Consumption Data	Portfolio Manager ID	Portfolio Manager ID	
Meter Consumption Data	Portfolio Manager Meter ID	Portfolio Manager Meter ID	

in following sections, this template design of creating several columns (Usage/Quantity, Meter ID, unit, Cost ($))for each new energy usage type have several draw backs:

- The large number of different energy usage types (19 in total) will result in large number of columns
- The large variety in the number of energy record for each types of energy source will result in a large number of waste space in the table: for example, there are only 12 records for 'Electric - Wind' but 111065 records for 'Electric - Grid'. For the EUAS template with different columns for different energy source, there will be $111065 - 12 = 111053$ empty cells in the column of 'Electric - Wind'
- If in the future, the new type of energy usage is included, the structure of the table will change (the number of columns), this requires reprocessing the whole table again. If we maintain the way of keeping all energy consumption stacked on top of each other (the way in PM table), we can either append records of the new resource to the end or save it to another table without affecting the already tidied data.

Thus I propose to maintain the structure of the PM in recording energy usage information. Hence the mapping of fields from source to target table would be as the following:

Table 2: Excel merging script field mapping table

Source sheet name	field name	target
Properties	State/Province	State
Properties	Country	Country
Properties	Gross Floor Area	GSF
Properties	Year Built	Year Built
Properties	Postal Code	Postal Code (first 5 digit)
Property Use	Self-Selected Primary Function	Use Type
Meter Consumption Data	Property Name	Building ID
Meter Consumption Data	End Date (month)	Month
Meter Consumption Data	End Date (year)	Year
Meter Consumption Data	(Usage/Quantity, Meter Type)	Usage/Quantity
Meter Consumption Data	(Usage/Quantity, Meter Type)	Cost ($)
Meter Consumption Data	Portfolio Manager ID	Portfolio Manager ID
Meter Consumption Data	Portfolio Manager Meter ID	Portfolio Manager Meter ID

2.1 Input data quality check - Energy Information

2.1.1 Number of missing value

From the output of the script, one can see there are missing data in the following fields:

- End Date: 39 missing data.

- "Cost$": 12942 missing data.

For "End Date", we will discard these records with, for "Cost ($)", we'll first mark the missing data with "-1", and discard it when doing cost related analysis

```
checking numer of missing values for columns
## ----------------------------------------##
Portfolio Manager ID
non_Null    344509
dtype: int64
## ----------------------------------------##
Portfolio Manager Meter ID
non_Null    344509
dtype: int64
## ----------------------------------------##
Meter Type
non_Null    344509
dtype: int64
## ----------------------------------------##
End Date
non_Null    344470
Null            39
dtype: int64
## ----------------------------------------##
Usage/Quantity
non_Null    344509
dtype: int64
## ----------------------------------------##
Usage Units
non_Null    344509
dtype: int64
## ----------------------------------------##
Cost ($)
non_Null    331567
Null         12942
dtype: int64
```

2.1.2 Ranges

From the range checking, one can see there are missing values for 'End Date' (marked as 'inf') and illegal values for the 'Usage/Quantity' (negative values)

Portfolio Manager ID	600	4428021
Portfolio Manager Meter ID	519	15550834
End Date	inf	2015-09-01 00:00:00
Usage/Quantity	-1385200.0	513798464.0
Cost ($)	0.0	7858632.0

Note: when pandas read in date time, it converts missing datetime data to current date by default, which is why it shows up as inf (infinity)

2.1.3 Non-negativity

Checking the number of negative records for each group of energy consumption. From the result, we can see there are 108 records in District Chilled Water and 151 records in District Hot Water with negative energy consumption records, which is identified as illegal records that needs to be removed.

```
value  Meter Type
<0     District Chilled Water - Electric                    108
       District Hot Water                                   151
>=0    District Chilled Water - Absorption                  153
       District Chilled Water - Electric                    528
       District Chilled Water - Engine                       49
       District Chilled Water - Other                      5925
       District Hot Water                                   220
       District Steam                                     15784
       Electric - Grid                                   111065
       Electric - Solar                                    1900
       Electric - Wind                                       12
       Fuel Oil (No. 2)                                   20534
       Natural Gas                                        79492
       Other Indoor                                          14
       Other:                                               580
       Other: Mixed Indoor/Outdoor                          467
       Potable Indoor                                      1177
       Potable: Mixed Indoor/Outdoor                      98474
       Power Distribution Unit (PDU) Input Meter             16
       Power Distribution Unit (PDU) Output Meter           106
       Uninterruptible Power Supply (UPS) Output Meter     7754
dtype: int64
```

After removing 39 missing "End Date" and 259 negative "Usage/Quantity", there are 344211 legal records to be further processed.

2.2 Input data quality check - Static info

2.2.1 Missing data in PM file

There are no missing data for the static information in sheet-0 of the PM file

```
checking numer of missing values for columns
## ----------------------------------------##
Property Name
```

```
non_Null    850
dtype: int64
## ----------------------------------------##
Portfolio Manager ID
non_Null    850
dtype: int64
## ----------------------------------------##
State/Province
non_Null    850
dtype: int64
## ----------------------------------------##
Postal Code
non_Null    850
dtype: int64
## ----------------------------------------##
Year Built
non_Null    850
dtype: int64
## ----------------------------------------##
Gross Floor Area
non_Null    850
dtype: int64
```

2.2.2 Duplicate and missing record in EUAM template

First read in "Building ID" and "Region" columns from EUAM template table, for each ("Building ID", "Region")pair, there are 11 duplicate records.

```
# number of records
Building ID    11713 non-null object
Region         11713 non-null int64

# number of unique values
Building ID: 1065
Region: 11
```

Per Jiyhun's advice, I should look up the 'Region' field with building id in PM table from the EUAS template, since there are more buildings in EUAS (1065) than in PM(850). However, after reading both tables, I found there are only 120 common buildings in the two files, which means one cannot use EUAS table as a lookup table to retrieve region information for buildings in the PM file.

```
850 buildings in PM
1065 buildings in EUAS
120 common building records between PM and EUAS
```

Jiyhun pointed out the link to the definition of the GSA lookup, and GSA region map.

There is a Canada state from the PM file, the following output shows the number of buildings in each state/Country in the PM file:

Canada	British Columbia	1
United States	Alabama	12
	Alaska	10
	Arizona	16
	Arkansas	9
	California	44
	Colorado	43
	Connecticut	6
	Delaware	1
	District of Columbia (D.C.)	44
	Florida	25
	Georgia	27
	Hawaii	3
	Idaho	4
	Illinois	23
	Indiana	10
	Iowa	6
	Kansas	4
	Kentucky	11
	Louisiana	14
	Maine	24
	Maryland	35
	Massachusetts	16
	Michigan	19
	Minnesota	14
	Mississippi	9
	Missouri	11
	Montana	10
	Nebraska	5
	Nevada	6
	New Hampshire	5
	New Jersey	11
	New Mexico	14
	New York	48
	North Carolina	14
	North Dakota	15
	Ohio	21
	Oklahoma	7
	Oregon	11
	Pennsylvania	15
	Puerto Rico	4
	Rhode Island	2
	South Carolina	14
	South Dakota	5
	Tennessee	12

```
Texas                          80
Utah                            9
Vermont                        24
Virgin Islands of the U.S.      2
Virginia                       17
Washington                     34
West Virginia                  12
Wisconsin                       6
Wyoming                         6
```

The non-U.S. state should be dropped in the analysis, because it is not in the definition of GSA region.

2.3 Clean up summary

- PM sheet-0, static information
  - Turn 'Property Name' from 'XXXXXXXX - XXXXXXXXXX' to just the string before '-'
  - Keep only 5 digit for zip code
  - Drop non-U.S. state
- PM sheet-5, energy information
  - Drop missing data for 'End Date'
  - Mark missing cost data as '-1'
  - Drop negative energy consumption record
  - Drop non-U.S. state

  The cleaned up energy information

```
Checking non-negativity after initial clean
>=0    344211
dtype: int64
is_nn  Meter Type
>=0    District Chilled Water - Absorption                  153
       District Chilled Water - Electric                    528
       District Chilled Water - Engine                       49
       District Chilled Water - Other                      5925
       District Hot Water                                   220
       District Steam                                     15784
       Electric - Grid                                   111065
       Electric - Solar                                    1900
       Electric - Wind                                       12
       Fuel Oil (No. 2)                                   20495
       Natural Gas                                        79492
       Other Indoor                                          14
       Other:                                               580
       Other: Mixed Indoor/Outdoor                          467
       Potable Indoor                                      1177
       Potable: Mixed Indoor/Outdoor                      98474
       Power Distribution Unit (PDU) Input Meter             16
       Power Distribution Unit (PDU) Output Meter           106
       Uninterruptible Power Supply (UPS) Output Meter     7754

# Ranges of columns
        Portfolio Manager ID                      600                  4428021
```

Portfolio Manager Meter ID	519	15550834
End Date	1998-06-30 00:00:00	2015-09-01 00:00:00
Usage/Quantity	0.0	513798464.0
Cost ($)	-1.0	7858632.0

2.3.1 Retrieving region information summary

I digitized the GSA region in into a table from GSA lookup and combined it with the data frame with the static information. State abbreviation table is retrieved from `http://www.stateabbreviations.us/`

Attention should be paid for U.S. owned island, since they don't appear in GSA region map.

3 Cleaned up and combined file

After cleaning the static and the energy information, retrieving region information and merging these three tables: static, energy and region, there are 344118 records in the cleaned up table.

```
Data columns (total 16 columns):
Portfolio Manager ID          344118 non-null int64
Portfolio Manager Meter ID    344118 non-null int64
Meter Type                    344118 non-null object
End Date                      344118 non-null object
Usage/Quantity                344118 non-null float64
Usage Units                   344118 non-null object
Cost ($)                      344118 non-null float64
Year                          344118 non-null int64
Month                         344118 non-null int64
Building ID                   344118 non-null object
State                         344118 non-null object
Postal Code                   344118 non-null int64
Country                       344118 non-null object
Year Built                    344118 non-null int64
GSF                           344118 non-null int64
Region                        344118 non-null int64
```