Data Validation and Data Verification Report

Author: Yangyang Cai

Date: 24/01/2022

Contents

[Purpose 3](#_Toc93970317)

[Method Used 3](#_Toc93970318)

[Data Validation File 3](#_Toc93970319)

[Data Assumptions 3](#_Toc93970320)

[Data Validation 4](#_Toc93970321)

[Data Verification 6](#_Toc93970322)

[Data Issues 7](#_Toc93970323)

[Transformation Suggestion 7](#_Toc93970324)

## Purpose

This report documents what data validation and data verification check have been done and what recommendations have been made for data transformation.

## Method Used

Python

## Data Validation File

* data\_validatoin\_and\_data\_verificaiton.ipynb
* data\_validatoin\_helper.py

## Data Assumptions

1. We only have 2 data sources for analysis:
   1. nmi\_info.csv, contains unique NMI information
   2. consumption data as csvs for each NMI, contains unique consumption information
2. NMI name is always formatted as NMI{ID}, which starts as NMI
3. Consumption data for each NMI is always saved as {nmi}.csv
4. Data Format Consumption:
   1. nmi\_info: NMI, STATE and INTERVAL as 3 columns
   2. consumption\_data: AESTTIME, QUANTITY and UNIT as 3 columns
5. Data Type Consumption:
   1. nmi\_info: NMI<String>, STATE<String> and INTERVAL<Integer>
   2. consumption\_data: AESTTIME<DateTime>, QUANTITY<Float>, and UNIT<String>
6. nmi\_info.csv is nmi master file, contains all NMI information
7. nmi\_info.csv is nmi master file with STATE as NSW, NT, QLD, SA, VIC, TAS, WA
8. nmi\_info.csv is nmi master file with INTERVAL as 15, 30
9. Consumption folder contains consumption data for all NMI
10. Consumption folder contains consumption data with UNIT as WH, KWH or MWH
11. Consumption folder contagions consumption data always has header
12. Both nmi\_info.csv and Consumption Data does not have missing data
13. The INTERVAL defined in mmi\_info.csv is consistent with consumption data
14. The UNIT in consumption data needs to be consistent for analysis
15. The quantity in consumption has reasonable data
16. Consumption data shows local date time

## Data Validation

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Row** | **ID** | **Category** | **File/Folder** | | **Check Item** | **Result** | **Data Consumption** | **Exception** |
| 1 | 1.1 | Format Check | | nmi\_info.csv | Valid csv file | Pass | 1.1 |  |
| 2 | 1.2 | Format Check | | Consumption Data Folder | All files are valid csv files | Pass | 1.2 | AESTTime has valid dataformat |
| 3 | 1.3 | Format Check | | Consumption Data Folder | All csv files have header | Pass | 11 |  |
| 4 | 2.1 | Source Check | | Consumption Data Folder | All files with csv as extension | Pass | 1.2 |  |
| 5 | 2.2 | Source Check | | Consumption Data Folder | All files with name start with NMI | Pass | 3.1 |  |
| 6 | 2.3 | Source Check | | Consumption Data Folder | No duplicate files | Pass | 1.2 |  |
| 7 | 3.1 | Missing Data Check | | nmi\_info.csv | Has related consumption data for each NMI | Not Pass | 1.1, 6 | NMIM2 is missing from nmi\_info.csv |
| 8 | 3.2 | Missing Data Check | | Consumption Data Folder | NMI does not have consumption data but in nmi\_info.csv | Not Pass | 1.1,9 | NMIK4 does not have consumption data |
| 9 | 3.3 | Missing Data Check | | nmi\_info.csv | NMI master file does not contain missing data | Pass | 12 |  |
| 10 | 3.4 | Missing Data Check | | Consumption Data Folder | Each consumption file does not have missing data | Not Pass | 12 | NMIA3.csv, NMIR2.csv,  NMIS3.csv have missing data |
| 11 | 4.1 | Schema Check | | nmi\_info.csv | Contains 3 columns | Pass | 5.1 |  |
| 12 | 4.2 | Schema Check | | nmi\_info.csv | NMI is string | Pass | 5.1 |  |
| 13 | 4.3 | Schema Check | | nmi\_info.csv | State is string | Pass | 5.1 |  |
| 14 | 4.4 | Schema Check | | nmi\_info.csv | Interval is integer | Not Pass | 5.1 | Interval defaults to be float when loaded as dataframe without any parameter setup |
| 15 | 4.5 | Schema Check | | Consumption Data Folder | Each consumption file contains 3 columns | Pass | 5.2 |  |
| 16 | 4.6 | Schema Check | | Consumption Data Folder | AESTTime is datetime with same format | Not Pass | 5.2 | AESTTime has different format and defaulted to be string when loaded into dataframe without any parameters setup |
| 17 | 4.7 | Schema Check | | Consumption Data Folder | Quantity is integer | Pass | 5.2 |  |
| 18 | 4.8 | Schema Check | | Consumption Data Folder | Unit is string | Pass | 5.2 |  |

## Data Verification

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Row** | **ID** | **Category** | **File/Folder** | **Check Item** | **Result** | **Data Consumption** | **Exception** |
| 19 | 5.1 | Duplicate Check | nmi\_info.csv | No duplicate rows | Pass | 1.1 |  |
| 20 | 5.2 | Duplicate Check | Consumption Data Folder | No duplicate rows | Not Pass | 1.2 | NMIR2.csv have duplicate rows |
| 21 | 6.1 | Value Check | nmi\_info.csv | NMI name is string starts with NMI | Pass | 2 |  |
| 22 | 6.2 | Value Check | nmi\_info.csv | State code is correct | Pass | 7 |  |
| 23 | 6.3 | Value Check | nmi\_info.csv | Interval is correct (15,30) | Pass | 8 |  |
| 24 | 6.4 | Value Check | Consumption Data Folder | Unit is correct (WH, KWH, MWH) | Not Pass | 10 | Unit as Mwh, kWh initially |
| 25 | 7.1 | Consistent Check | nmi\_info.csv | Interval is consistent with AESTTime for consumption data | Not Pass | 13 | NMIA3.csv, NMIG2.csv,  NMIS3.csv are inconsistent |
| 26 | 7.2 | Consistent Check | Consumption Data Folder | Unit is consistent with single measurement with WH, KWH or MWH | Not Pass | 14 | NMIA1.csv has inconsistent values from ['Mwh', 'kWh'] |
| 27 | 7.3 | Consistent | Consumption Data Folder | AESTTime has consistent interval data | Not Pass |  | NMIG2.csv has inconsistent interval data |
| 28 | 8.1 | Outlier Check | Consumption Data Folder | Quality does not have outliers | Not Pass | 15 | NMIG1.csv, NMIM1.csv,  NMIR1.csv,  NMIR2.csv,  NMIS1.csv  NMIS2.csv, NMIS3.csv have outliers |

## Data Issues

* Nmi\_info.csv
  + NMI has consumption data but is missing from master file
  + NMI does not have consumption data but is in the master file
  + Columns names need to be standardized (i.e., all uppercase)
  + Column type needs to be standardized (i.e., INTERVAL as int)
  + Columns values need to be standardized (i.e., all uppercase)
  + Make sure no duplicate rows
  + Make sure no missing values
  + Interval needs further analysis which is not align with consumption data
* Consumption Data:
  + Column names need to be standardized (i.e., all uppercase)
  + Column type needs to be standardized (i.e., AESTIME same date format)
  + Columns values need to be standardized (i.e., UNIT all uppercase)
  + Unit measurement needs to be consistent format (i.e., all KWH)
  + Some files have duplicate rows
  + Some files have missing data
  + AESTTIME needs to be local time based on State
  + Need more date columns for further analysis
  + Need add common column (NMI) to join with nmi master dataset
  + Need mark outliers for further analysis

## Transformation Suggestion

### Transform NMI Master File

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Row** | **Data Issue** | **Data Consumption** | **Exception** | **Transform Method** |
| 1.1 | Has related consumption data for each NMI | 1.1, 6 | NMIM2 is missing from nmi\_info.csv | N/A (due to do not know the state info so could not update master file) |
| 1.2 | NMI does not have consumption data but in nmi\_info.csv | 1.1,9 | NMIK4 does not have consumption data | N/A (do inner join then do not need worry about this item) |
| 1.3 | Columns names need to be standardized (i.e., all uppercase) | 5.1 | Columns name is Nmi, State and Interval as default | Transformed to be all upper case. Because pandas are case-sensitive. |
| 1.4 | Column type needs to be standardized (i.e., INTERVAL as int) | 5.1 | Interval defaults to be float when loaded as dataframe without any parameter setup | Use pandas to define explicit date types |
| 1.5 | Columns values need to be standardized (i.e., all uppercase) | 7 |  | Transformed to be all upper case. Because pandas are case-sensitive. |
| 1.6 | Make sure no duplicate rows | 1.1 |  | Remove duplicate rows |
| 1.7 | Make sure no missing values (NMI & STATE) | 12 |  | Remove rows have missing NMI or STATE (because we used NMI for JOIN and STATE to transform consumption datetime to be local time |

### 2.Transform Consumption Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Row** | **Data Issue** | **Data Consumption** | **Exception** | **Transform Method** |
| 2.1 | Column type needs to be standardized (i.e., AESTIME same date format) | 5.2 | NMIG2 has different date format | Use pandas functions to load data with correct data types |
| 2.2 | Column names need to be standardized (i.e., all uppercase) | 5.4 | Columns names are AESTTime, Quantity, Unit as default | Transformed to be all upper case. Because pandas are case-sensitive. |
| 2.3 | Columns values need to be standardized (i.e., UNIT all uppercase) | 10 | Unit as Mwh, kWh initially | Transformed to be all upper case. Because pandas are case-sensitive. |
| 2.4 | Unit measurement needs to be consistent format (i.e., all KWH) | 14 | NMIA1.csv has inconsistent values from ['Mwh', 'kWh'] | Convert Quality and Unit to be kWh |
| 2.4 | Some files have duplicate rows | 1.2 | NMIR2.csv have duplicate rows | Remove duplications |
| 2.5 | Some files have missing data | 12 | NMIA3.csv, NMIR2.csv,  NMIS3.csv have missing data | Missing data imputation – QUANTITY:  AESTTIME: |
| 2.6 | Unit measurement needs to be consistent format (i.e., all KWH) | 14 | NMIA1.csv has inconsistent values from ['Mwh', 'kWh'] | Convert Quality and Unit to be kWh |
| 2.7 | AESTTIME needs to be local time based on State | 16 | AESTTIME needs to be transformed based on STATE | Transform the datetime column to be local time |
| 2.8 | Need more date columns for further analysis |  | All files | Add more date features |
| 2.9 | Do not have common column to join with nmi master file |  | All files | Add NMI column when load consumption data |
| 2.10 | Quality does not have outliers | 15 | NMIG1.csv, NMIM1.csv,  NMIR1.csv,NMIR2.csv,  NMIS1.csvNMIS2.csv, NMIS3.csv have outliers | Mark outlier before further analysis |

## Further Discussion

### Missing Data Strategy

* NMI Master Data: NMI or STATE is missing - will drop all rows because do not have other datasets could be used to impute those two columns
* Consumption Data: QUANTITY data missing – used back fill method
* Consumption Data: AESTTime data missing – will drop all empty columns if AESTTime is missing because the interval in nmi\_info.csv is not always correct, which could not be used to impute missing datetime

### Remove Duplications Strategy

* All Datasets: Remove duplications with all columns
* NMI Master Data: Remove duplications with NMI because do not have other data sources to impute NMI
* Consumption Data: Remove duplications with AESTTime because could not use interval data to impute missing datatime

### Data Time Transform Strategy

* Reference Link: <https://en.wikipedia.org/wiki/Time_in_Australia>
* Reference Link: <https://en.wikipedia.org/wiki/List_of_tz_database_time_zones>

### Data Features Strategy

* The designed solution includes the following date features
  + DATE
  + YEAR
  + YEARDAY
  + MONTH
  + MONTHNAME
  + WEEK
  + DAY
  + DAYNAME
  + HOUR
  + MINUTE
  + WEEKDAY
  + WEEKEND
  + TIME
  + MONTHDAY
  + HOURMINUTE
  + SESSION:
    - Hour 0 – 4 AS 1 - Late Night
    - Hour 4 – 8 AS 2 - Early Morning
    - Hour 8 – 12 AS 3 – Morning
    - Hour 12 – 16 AS 4 – Noon
    - Hour 16 – 20 AS 5 – Eve
    - Hour 20 – 24 AS 6 - Night
  + SEASON
    - 3, 4, 5 AS 1 - Spring
    - 6, 7, 8 AS 2 - Summer
    - 9, 10, 11 AS 3 - Autumn
    - 12, 1, 2 AS 4 - Winter