# Data Formatting Demo

## Version 1.0

**Author: Shuiliang (Leon) Wu**

## Table of Contents

# Introduction

This demo provides a guide to format provided `final_cleaned_diabetes.csv` raw data (cleaned) to meet `Requirements of Data Format` so that date can be recognized by the source code for federated learning.

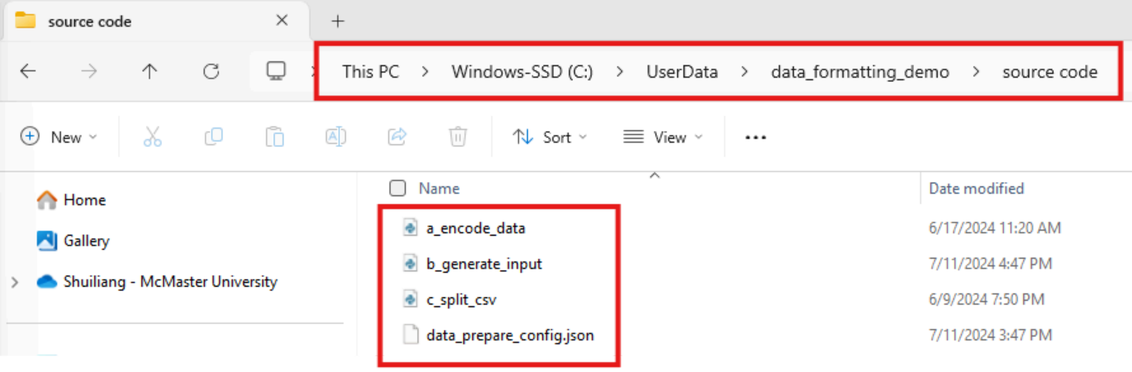
# What You Have Before Started

- Data Formatting Demo (this one)
- Source code in `data_formatting_demo` including four files:
  - `a_encode_data.py`
  - `b_generate_input.py`
  - `c_split_csv.py`
  - `data_prepare_config.json`
- Raw dataset `final_cleaned_diabetes.csv` (in `raw_data` folder)

# Getting Started

This Demo is demonstrated using Windows 11 (operations in macOS are very similar).

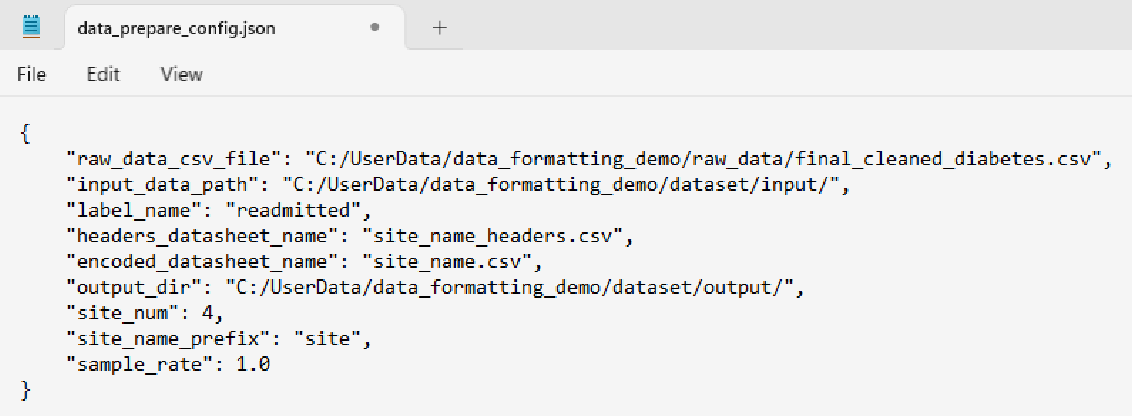
1. Open the `source code` folder in `data_formatting_demo`.



This PC > Windows-SSD (C:) > UserData > data_formatting_demo > source code

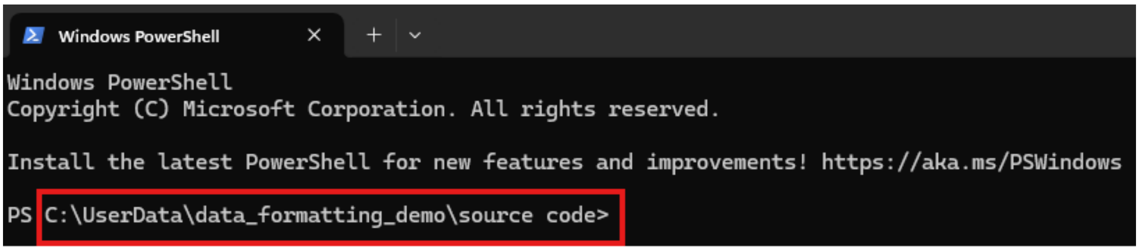
| Name | Date modified |
|---|---|
| a_encode_data | 6/17/2024 11:20 AM |
| b_generate_input | 7/11/2024 4:47 PM |
| c_split_csv | 6/9/2024 7:50 PM |
| data_prepare_config.json | 7/11/2024 3:47 PM |

2. Right-click `data_prepare_config.json`, select `Edit in Notepad` then update the parameters accordingly.



```
{
    "raw_data_csv_file": "C:/UserData/data_formatting_demo/raw_data/final_cleaned_diabetes.csv",
    "input_data_path": "C:/UserData/data_formatting_demo/dataset/input/",
    "label_name": "readmitted",
    "headers_datasheet_name": "site_name_headers.csv",
    "encoded_datasheet_name": "site_name.csv",
    "output_dir": "C:/UserData/data_formatting_demo/dataset/output/",
    "site_num": 4,
    "site_name_prefix": "site",
    "sample_rate": 1.0
}
```

3. Right-click at the blank space in `source code` folder, select `Open in Terminal`. The terminal shall be pop-up as below:



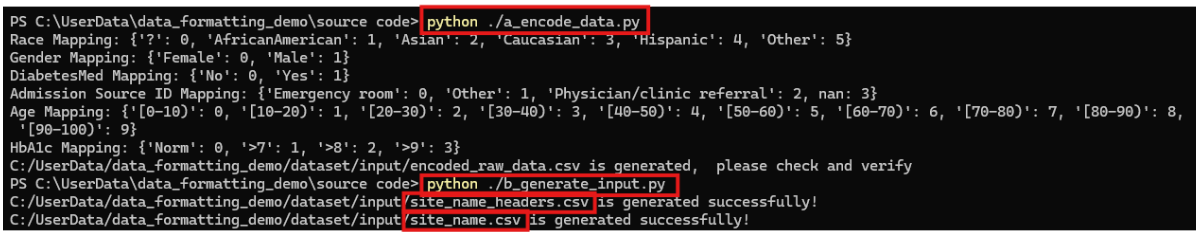
```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\UserData\data_formatting_demo\source code>
```

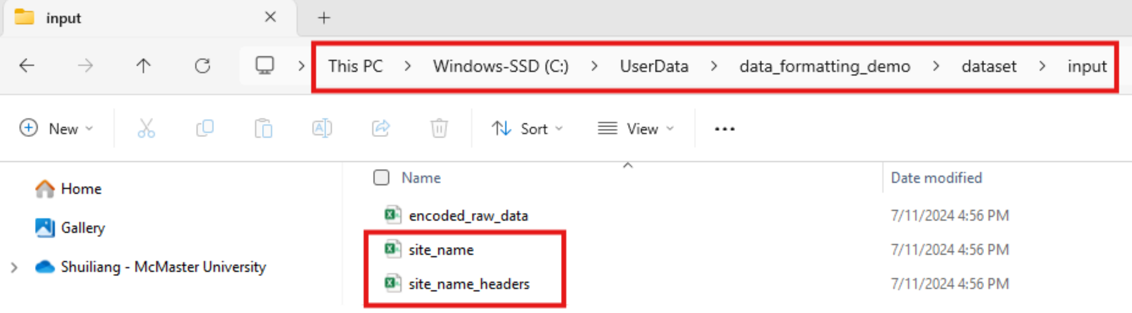
4. Formatting the data by running below in terminal one-by-one:

```
python ./a_encode_data.py
python ./b_generate_input.py
```



```
PS C:\UserData\data_formatting_demo\source code> python ./a_encode_data.py
Race Mapping: {'?': 0, 'AfricanAmerican': 1, 'Asian': 2, 'Caucasian': 3, 'Hispanic': 4, 'Other': 5}
Gender Mapping: {'Female': 0, 'Male': 1}
DiabetesMed Mapping: {'No': 0, 'Yes': 1}
Admission Source ID Mapping: {'Emergency room': 0, 'Other': 1, 'Physician/clinic referral': 2, nan: 3}
Age Mapping: {'[0-10)': 0, '[10-20)': 1, '[20-30)': 2, '[30-40)': 3, '[40-50)': 4, '[50-60)': 5, '[60-70)': 6, '[70-80)': 7, '[80-90)': 8,
 '[90-100)': 9}
HbA1c Mapping: {'Norm': 0, '>7': 1, '>8': 2, '>9': 3}
C:/UserData/data_formatting_demo/dataset/input/encoded_raw_data.csv is generated,  please check and verify
PS C:\UserData\data_formatting_demo\source code> python ./b_generate_input.py
C:/UserData/data_formatting_demo/dataset/input/site_name_headers.csv is generated successfully!
C:/UserData/data_formatting_demo/dataset/input/site_name.csv is generated successfully!
```

5. There shall be `site_name.csv` and `site_name_headers.csv` generated in `input` folder under `dataset` folder.  Replace `site_name` with your actual site name, and they are ready to be used.

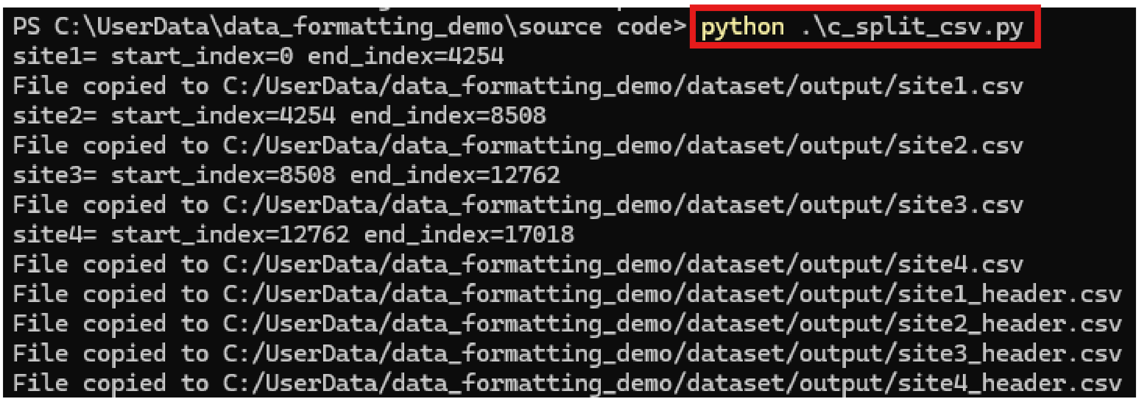


This PC > Windows-SSD (C:) > UserData > data_formatting_demo > dataset > input

| Name | Date modified |
|---|---|
| encoded_raw_data | 7/11/2024 4:56 PM |
| site_name | 7/11/2024 4:56 PM |
| site_name_headers | 7/11/2024 4:56 PM |

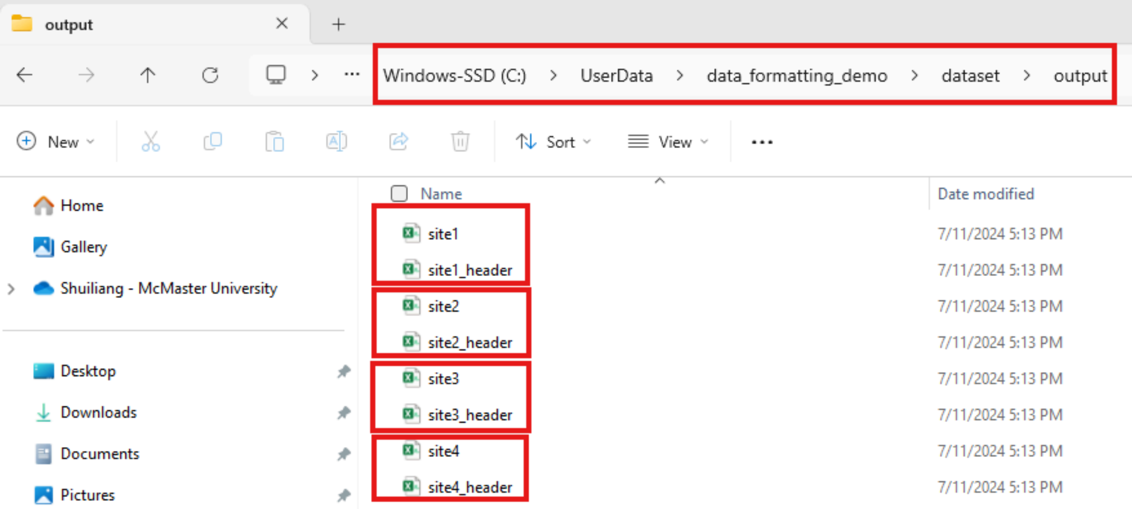
6. Since there will be four sites in total in the demo of `NVIDIA FLARE User Guide for Project Manager` and `NVIDIA FLARE User Guide for Site Admin`, the formatted dataset is split into four datasets with equal amount of data in each datasets by running:

```
python ./c_split_csv.py
```



```
PS C:\UserData\data_formatting_demo\source code> python .\c_split_csv.py
site1= start_index=0 end_index=4254
File copied to C:/UserData/data_formatting_demo/dataset/output/site1.csv
site2= start_index=4254 end_index=8508
File copied to C:/UserData/data_formatting_demo/dataset/output/site2.csv
site3= start_index=8508 end_index=12762
File copied to C:/UserData/data_formatting_demo/dataset/output/site3.csv
site4= start_index=12762 end_index=17018
File copied to C:/UserData/data_formatting_demo/dataset/output/site4.csv
File copied to C:/UserData/data_formatting_demo/dataset/output/site1_header.csv
File copied to C:/UserData/data_formatting_demo/dataset/output/site2_header.csv
File copied to C:/UserData/data_formatting_demo/dataset/output/site3_header.csv
File copied to C:/UserData/data_formatting_demo/dataset/output/site4_header.csv
```

7. There shall be four datasets generated in `output` folder under `dataset` folder. Replace `site#` with the actual site name, and they are ready to be used for demo.



Windows-SSD (C:) > UserData > data_formatting_demo > dataset > output

| Name | Date modified |
|---|---|
| site1 | 7/11/2024 5:13 PM |
| site1_header | 7/11/2024 5:13 PM |
| site2 | 7/11/2024 5:13 PM |
| site2_header | 7/11/2024 5:13 PM |
| site3 | 7/11/2024 5:13 PM |
| site3_header | 7/11/2024 5:13 PM |
| site4 | 7/11/2024 5:13 PM |
| site4_header | 7/11/2024 5:13 PM |