Họ tên: **Nguyễn Hữu Hiệu**
MSSV: **20520506**

1. **Code book**

Thông tin	Nội dung
Tên bộ dữ liệu	Diabetes patient records
Nguồn thu thập	Từ 2 nguồn: 1. Máy đo tự động: có đồng hồ bấm giờ tự động, sẽ ghi lại chính xác thời gian tại lúc đo số liệu. 2. Thu thập bằng tay: Sử dụng bản ghi giấy (paper record), giờ được định sẵn vào các khung giờ: sáng (8:00), trưa (12:00), chiều (18:00) và tối (22:00).
Số thuộc tính	4
Số mẫu dữ liệu	29330
	(1) Date in MM-DD-YYYY format (2) Time in XX:YY format (3) Code 33 = Regular insulin dose 34 = NPH insulin dose 35 = UltraLente insulin dose 48 = Unspecified blood glucose measurement 57 = Unspecified blood glucose measurement 58 = Pre-breakfast blood glucose measurement 59 = Post-breakfast blood glucose measurement 60 = Pre-lunch blood glucose measurement 61 = Post-lunch blood glucose measurement

	62 = Pre-supper blood glucose measurement 63 = Post-supper blood glucose measurement 64 = Pre-snack blood glucose measurement 65 = Hypoglycemic symptoms 66 = Typical meal ingestion 67 = More-than-usual meal ingestion 68 = Less-than-usual meal ingestion 69 = Typical exercise activity 70 = More-than-usual exercise activity 71 = Less-than-usual exercise activity 72 = Unspecified special event (4) Value
Thông tin tác giả	kahn@informatics.WUSTL.EDU (Internet) or 70333,34 (CompuServe)

2. **Raw data**

Raw data gồm các file **data-01, data-02,… data-70**.

3. **Tidy data**

Tidy data được lưu lại thành file **diabetes.csv**.

4. **Instrucsiton list**

```
# Import libraries
import os
import time
import pandas as pd
```

```
# Insert the directory path in here
path = 'Diabetes-Data'
lst_df = []
# Extracting all the contents in the directory corresponding to path
l_files = os.listdir(path)

col_name = ["Date","Time","Code","Value"]

# Iterating over all the files
for file in l_files:
  file_path = os.path.join(path, file)

  if file.startswith('data-'):
    df = pd.read_csv(file_path, names=col_name, sep="\t")
    lst_df.append(df)
    # print(df.shape)

print('Task finished!')
print(df)
final_df = pd.concat(lst_df)

final_df.to_csv('diabetes.csv', index=False)
```