BÀI THỰC HÀNH 5
THỰC HÀNH THAO TÁC DỮ LIỆU VỚI PANDAS

1. Một vài lệnh cơ bản thường sử dụng trong Pandas

Giả sử dataframe có tên sinhvien

Hiển thị n dòng đầu tiên của 1 dataframe

```
sinhvien.head(n)
```

Hiển thị n dòng cuối cùng của 1 dataframe

```
sinhvien.tail(n)
```

Xem giá trị của một thuộc tính

```
sinhvien['Col']
```

Hiển thị n dòng đầu tiên của Col trong dataframe sinhvien

```
sinhvien['Col'].head(n)
```

Lấy dữ liệu từ Col1 và Col2 của dataframe sinhvien

```
sinhvien['Col1','Col2']
```

Xem tên các thuộc tính (các cột) của bộ dữ liệu

```
sinhvien.columns
```

2. Một vài lệnh để xem thông tin của bộ dữ liệu

Hiển thị thông tin của bộ dữ liệu sinhvien

```
sinhvien.info()
```

Xem các thông số thống kê của bộ dữ liệu.

Lưu ý: chỉ có thuộc tính number mới được tính toán.

```
sinhvien.describe()
```

Xem số lượng dòng, cột của bộ dữ liệu.

```
sinhvien.shape()
```

Xem giá trị duy nhất của bộ dữ liệu.

sinhvien.value_counts(dropna=False)

Xem giá trị duy nhất của một thuộc tính

sinhvien[‘Col1’].value_counts()

Sắp xếp dữ liệu trong dataframe

sinhvien.sort_values(‘MSSV’,’HoTen’)

sinhvien.sort_values(‘MSSV’,’HoTen’, ascending=False)

3. Một số lệnh làm sạch dữ liệu

Lọc dữ liệu với điều kiện

sinhvien[sinhvien[‘GioiTinh’] == ‘Nam’]

Lọc dữ liệu với điều kiện, nhưng chỉ hiển thị 1 vài colums được chỉ định

sinhvien[sinhvien[‘GioiTinh’] == ‘Nam’,[‘HoTen’,’GioTinh’,’DiaChi’]]

Gom nhóm dữ liệu

sinhvien.groupby(‘MaLop’)

Một số hàm kết hợp với groupby: max, min, count, sum, median.

Kiểm tra giá trị null

sinhvien.isnull()

sinhvien[‘Diem’].isnull()

sinhvien[‘Diem’].isnull().sum() *#Xem có bao nhiêu giá trị null trong cột ‘Diem’*

Xóa các giá trị rỗng

sinhvien.dropna()

sinhvien.dropna(axis=1) *#Xóa cột có giá trị rỗng*

Điền vào giá trị null bằng 1 giá trị x cụ thể

```
sinhvien['Diem'].fillna(x)
sinhvien['Diem'].fillna(sinhvien['Diem'].mean())
```

Thay đổi giá trị F thành 'Nam' trong dataframe sinhvien

```
sinhvien.replace('F','Nam')
sinhvien.replace(['F','M'],['Nam','Nu'])
```

Thay đổi tên thuộc tính (tên cột) trong dataset

```
sinhvien.replace(columns={'Tên cũ':'Tên mới'})
```

Dữ liệu bị trùng

```
sinhvien.duplicated()           #Kiểm tra giá trị bị trùng lắp
sinhvien.duplicated().sum()     #Đếm xem có bao nhiêu dữ liệu trùng
sinhvien.drop_duplicate()       #Xóa dòng dữ liệu trùng
```

Xóa những thuộc tính không sử dụng trong bộ dữ liệu

```
#Loại bỏ 3 cột Điểm, Họ tên, Ngày sinh trong bộ dữ liệu sinh viên
col = ['Diem','Hoten','Ngaysinh']
sinhvien.drop(col, inplace=True, axis=1)
```

Thêm một cột Xeploai vào dataframe với tất cả giá trị bằng 0

```
sinhvien['Xeploai'] = 0
```

4. Một số hàm thống kê

Thống kê dữ liệu cho các thuộc tính số

```
sinhvien.describe()
```

Tính độ tương quan giữa các thuộc tính

```
sinhvien.corr()
```

Đếm số lượng các giá trị không null của các thuộc tính

```
sinhvien.count()
sinhvien.isnull().count()
```

Tìm giá trị lớn nhất, nhỏ nhất của các thuộc tính trong bộ dữ liệu

```
sinhvien.min()      sinhvien.max()
```

Tính độ lệch chuẩn của mỗi thuộc tính số trong bộ dữ liệu

```
sinhvien[‘Diem’].std()
```

5. Một số hàm đọc, lưu dữ liệu

Đọc dữ liệu từ file .csv, .txt

```
pd.read_csv(file.csv’, encoding='utf-8')
pd.read_csv(‘filename.txt’,sep=’ ’)
```

Lưu dữ liệu dataframe New_SV thành tập tin Final.csv

```
New_SV.to_csv('Final.csv', encoding='utf-8', index=False, header=True)
```

Sử dụng bộ dữ liệu courses.csv, thực hiện các lệnh sau và cho biết kết quả. Mỗi lệnh có sự thay đổi như thế nào?

```
Cách viết 1: df = pd.read_csv('courses.csv')
             print(df)
Cách viết 2: df = pd.read_csv('courses.csv', index_col='Courses')
             print(df)
```

Cách viết 3: df = pd.read_csv('courses.csv', header=None, skiprows=2)

print(df)

Cách viết 4: df = pd.read_csv('courses.csv', usecols =['Courses','Fee','Discount'])

print(df)

Tạo một dataframe mới với 3 cột 'Courses', 'Fee' và 'Duration'. Tiền xử lý dữ liệu cho bộ dữ liệu vừa mới tạo. Đổi tên các thuộc tính lần lược là: 'Khoa Hoc', 'Le Phi' và 'Thoi Gian'. Sau đó lưu dữ liệu này thành file khoahoc.txt

6. Thực hành trên bộ dữ liệu student.csv

- Load bộ dữ liệu student.csv.

- Xem bộ dữ liệu có bao nhiêu quan sát và bao nhiêu thuộc tính.

- Hiển thị 10 dòng đầu và 10 dòng cuối của bộ dữ liệu này.

- Xem giá trị null của các thuộc tính. Thuộc tính nào có giá trị null nhiều nhất?

- Thay đổi giới tính 'F' thành 'Nam', và 'M' thành 'Nữ".

- Xử lý điểm cho các sinh viên bị thiếu bằng điểm trung bình của tất cả sinh viên.

- Kiểm tra số lượng dữ liệu bị trùng lắp.

- Kiểm tra thuộc tính Mark có bị giá trị outlier. Thay đổi các giá trị >10 bằng giá trị max của thuộc tính này.

- Loại bỏ những dòng dữ liệu bị trùng. Sau khi loại bỏ kiểm tra lại số dòng, số thuộc tính của bộ dữ liệu.

- Thay đổi tên các thuộc tính như sau: ID = MSSV, Name = Ho Ten, BOD = Ngay Sinh, Mark = Diem, Class = Lop.

- Tạo một danh sách sinh viên nam với thông tin là MSSV, Ho Ten và Lop. Sau đó lưu danh sách này thành file SV_nam.csv

- Load file SV_nam.csv để kiểm tra xem có thành công hay không.

7. Thực hành trên bộ dữ liệu wine.csv

- Thực hiện các công việc tương tự như trên.

- Thực hiện các công việc tiền xử lý dữ liệu: mising, null, duplicate, outlier…

- Xem độ tương quan giữa các thuộc tính.

- Loại bỏ những thuộc tính dư thừa. (tạo ra một dataframe mới không có những thuộc tính này)

- Lưu thành một file mới.