|  |
| --- |
| Tài liệu hướng dẫn |

## Thông tin môn học

|  |  |
| --- | --- |
| Tên môn học | Khoa học dữ liệu |
| Giáo viên | Nguyễn Văn Huy |

## Thông tin sinh viên

|  |  |
| --- | --- |
| Họ và tên | Hà Huy Nam |
| Mã sinh viên | K215480106063 |
| Lớp | K57KMT |

## Nội dung bài tập

|  |  |  |
| --- | --- | --- |
| Đề bài | Bài tập 4: Phân tích dữ liệu thời tiết  Yêu cầu bài tập:  1.Đọc dữ liệu và kiểm tra thông tin cột thời gian  2.Lọc dữ liệu: chỉ dữ lại dữ liệu từ 2020 trở đi  3.Trích xuất thông tin thời gian: Thêm cột year, month, day\_of\_week,hour  4.Lấy mẫu lại dữ liệu: Chuyển từ dữ liệu theo giờ thành trung bình nhiệt độ theo ngày  5.Nội suy dữ liệu bị thiếu để điền nhiệt độ bị khuyết  6.Tính toán xu hướng thời gian: Tính nhiệt độ trung bình theo tháng và vẽ biểu đồ | |
| Github bài làm | https://github.com/Hhuynam/KhoaHocDuLieu | |
| Link dữ liệu(Kaggle) | [Historical Hourly Weather Data 2012-2017](https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data)  https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data |  |

## Nội dung hướng dẫn

### **1. Đọc dữ liệu và kiểm tra thông tin cột thời gian**

Trong Vscode, cài đặt extension “EditCSV” giúp xem dữ liệu dạng bảng có giao diện

A screenshot of a computer

AI-generated content may be incorrect.

Ví dụ xem bảng “humidity.csv”

A screenshot of a computer screen

AI-generated content may be incorrect.

Quá trình phân tích dữ liệu như sau:

Đọc dữ liệu (5 dòng đầu) file csv: dùng head()

A screenshot of a computer

AI-generated content may be incorrect.

Kiểm tra kiểu dữ liệu của các cột: dùng info()

A screenshot of a computer

AI-generated content may be incorrect.

Cột datatime trong này đang có kiểu object-chuỗi văn bản chứ không phải kiểu datetime64 mà hay dùng để xử lý thời gian. Để chuyển sang datetime64 dùng to\_datetime()

Kiểm tra lại bằng print(data.dtypes), xem 5 dòng đầu tiên cột datetime bằng head()

A screenshot of a computer

AI-generated content may be incorrect.

### **2. Lọc dữ liệu**

Lọc dữ liệu: so sánh giá trị các trường year trong dataframe, in ra nếu >= 2015 bằng dùng head(). In ra số lượng dòng đã lọc bằng shape()

A screenshot of a computer

AI-generated content may be incorrect.

### **3.Trích xuất thông tin thời gian**

Trích xuất theo thông tin mong muốn ví dụ theo năm/tháng/... (year/month/day/hour): dùng dt.year,dt.datetime,...

A screenshot of a computer

AI-generated content may be incorrect.

### **4.Lấy mẫu lại dữ liệu: Chuyển từ dữ liệu theo giờ thành trung bình nhiệt độ theo ngày**

Mở đọc file về nhiệt độ “tempurature.csv”: tương tự như đã làm

A screenshot of a computer

AI-generated content may be incorrect.

Cần chuyển sang dạng datetime64 nếu chưa: cách tương tự đã làm

A screenshot of a computer

AI-generated content may be incorrect.

Sử dụng phương thức .resample() để lấy mẫu lại dữ liệu từ mức độ theo giờ sang theo ngày và tính trung bình nhiệt độ

A screenshot of a computer

AI-generated content may be incorrect.

### **5.Nội suy dữ liệu bị thiếu để điền nhiệt độ bị khuyết**

Sử dụng các hàm như isnull() hoặc isna() trong Pandas để kiểm tra giá trị bị thiếu (NaN)

A screen shot of a computer

AI-generated content may be incorrect.

Phương pháp nội suy: sử dụng giá trị trước và sau để dự đoán giá trị bị thiếu

method='linear': Nội suy tuyến tính, sử dụng các giá trị trước và sau, có thể sử dụng các phương pháp khác như method='time' A screenshot of a computer

AI-generated content may be incorrect.

### **6.Tính toán xu hướng thời gian: Tính nhiệt độ trung bình theo tháng và vẽ biểu đồ**

Vẽ xu hướng nhiệt độ trung bình theo ngày: dùng figure() và tham số figsize để chỉ định kích thước

Vẽ nhiệt độ trung bình cho từng thành phố (thay cột phù hợp từ tập dữ liệu)

A screen shot of a graph

AI-generated content may be incorrect.