Assignment 11 - Bus Data Analysis using Pandas[1]

File `BusData.txt` chứa dữ liệu về quá trình di chuyển bằng xe bus giữa nhà và trường Đại học Khoa học Tự nhiên (HUS) của sinh viên X trong khoảng thời gian 2018-2020. Một số thông tin liên quan đến bộ dữ liệu này:

- Sinh viên X chỉ di chuyển trên một tuyến bus cố định
- Lộ trình chính thức của tuyến bus này là cố định.
- Dữ liệu được thu thập và ghi lại theo trình tự thời gian.

Cấu trúc của file dữ liệu:

- Một số dòng giải thích.
- Các cột mốc thời gian (ví dụ như `#16.04.2018`)
- Dữ liệu về những lần di chuyển. Thông tin của mỗi lần di chuyển được ghi trong một dòng.

Ngoài ra, ngày diễn ra một lần di chuyển nằm giữa hai cột mốc thời gian phía trên gần nhất và phía dưới gần nhất của lần di chuyển đó. Giải thích chi tiết về định dạng của dữ liệu có thể được tìm thấy trong file dữ liệu. Về cơ bản, mỗi dòng của file dữ liệu có thể chứa những thông tin sau:

- Địa điểm xuất phát: nhà (H) hoặc trường (U).
- Ngày trong tuần.
- Thời điểm lên xe.
- Thời gian di chuyển.
- Thời điểm xuống xe.

1. *(Data Cleaning)*
Trong phần đầu (khoảng 25%) của file, thông tin về thời điểm xuống xe không được thu thập, những trường thông tin khác có thể được thu thập hoặc không[2]. Trong phần còn lại của file, dữ liệu được trình bày theo một cách có hệ thống hơn. Tìm những sai sót (nếu có) trong file dữ liệu của X .
Bài tập này yêu cầu chuyển dữ liệu từ file thành một `pandas.DataFrame`. Trước tiên, hãy lưu chọn kiểu dữ liệu (data types) hợp lý cho từng trường thông tin của dữ liệu (data attributes). Giải thích về lựa chọn này.
Trình bày câu trả lời trong một file *Jupyter notebook* với tên được đặt theo mẫu

`Assignment11_FullName_DataCleaning.ipynb`

2. *(Basis Processing)*
Trong một số dòng của dữ liệu, thông tin về thời gian di chuyển không được ghi lại. Hãy tính toán thời gian di chuyển dựa theo thời điểm lên xe và thời điểm xuống xe của dòng dữ liệu đó. Hãy tìm hiểu, tính toán những thông tin cơ bản liên quan đến bộ dữ liệu này. Dưới đây là danh sách một số câu hỏi gợi ý:

[1] Ảnh/dữ liệu (ngoại trừ file dữ liệu đã cho) có thể được gửi kèm bài tập nếu cần thiết.
[2] hãy quen với việc dữ liệu là không hoàn hảo

(a) X cần trung bình bao nhiêu phút để di chuyển tới trường bằng xe bus? Thời gian đó có phụ thuộc vào thời điểm di chuyển trong ngày hay không?

(b) X thường đến trường muộn nhất vào ngày nào trong tuần?

(c) Số lần X đi xe bus từ nhà tới trường có bằng số lần X đi xe bus từ trường về nhà hay không?

Trình bày câu trả lời trong một file *Jupyter notebook* với tên được đặt theo mẫu

`Assignment11_FullName_BasisProcessing.ipynb`

3. *(Assumptions Proposing)*
Việc đặt ra giả sử đóng vai trò quan trọng trong việc đưa ra những kết luận sâu sắc khi phân tích một bộ dữ liệu. Những giả sử này giúp ta có thể thông tin để phân tích bộ dữ liệu. Tuy nhiên, cần tránh đưa ra những giả sử vô lý, khiến việc phân tích đi sai hướng.
Trong trường hợp của bài toán này, một giả sử hợp lý là X luôn đến trường sát giờ vào học, còn một giả sử không hợp lý là X lên/xuống xe bus tại những bến khác nhau trong những ngày khác nhau.
Trình bày những giả thiết hợp lý liên quan đến bộ dữ liệu trong một file *Jupyter notebook* với tên được đặt theo mẫu

`Assignment11_FullName_AssumptionsProposing.ipynb`

4. *(Bus Data Analysis)*
Từ những giả sử được đặt ra ở bài tập (3), hãy phân tích file dữ liệu theo mọi khía cạnh có thể và đưa ra những kết luận hợp lý từ những phân tích đó.
Ví dụ, với giả sử X luôn đến trường sát giờ vào học, ta có thể dựa vào thời điểm X xuống bến xe bus trong một ngày, kết hợp với thông tin về những thời điểm bắt đầu một tiết học tại HUS để suy ra thời điểm X bắt đầu tiết học ngày hôm đó.
Trình bày câu trả lời trong một file *Jupyter notebook* với tên được đặt theo mẫu

`Assignment11_FullName_BusDataAnalysis.ipynb`

Bonus: Hãy trả lời những câu hỏi sau[3].

(a) Trong thực tế, xe bus có thể di chuyển khác với lộ trình đã công bố hay không? Nếu câu trả lời là có thì sự thay đổi này xảy ra trong trường hợp nào?

(b) Chi phí X đã dành cho việc đi học bằng xe bus trong khoảng thời gian $2018 - 2020$ là bao nhiêu?

(c) Tuyến bus mà X sử dụng là tuyến bus số bao nhiêu?

(d) Biết rằng xuất phát từ nhà, X lên xe bus tại một bến cố định A và xuống xe bus tại một bến cố định B. Tương tự, xuất phát từ trường đại học, X lên xe bus tại một bến cố định C và xuống xe bus tại một bến cố định D. Hỏi các bến bus A, B, C, D ở đâu?

(e) Nhà của X ở khu vực nào?

[3] bằng cách tìm hiểu về hệ thống xe bus tại Hà Nội

Additional Problems

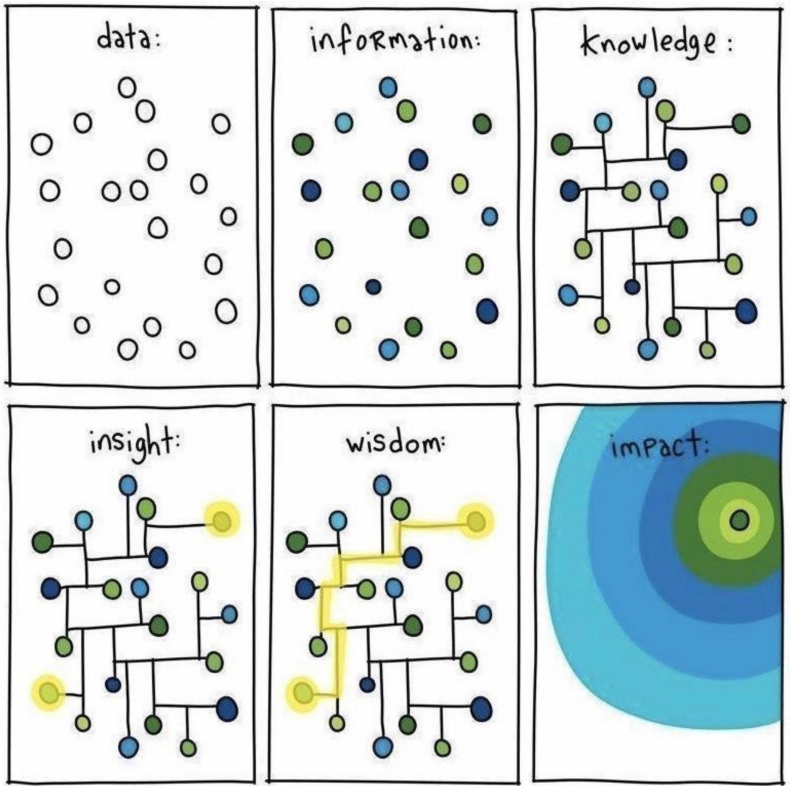
5. *(Schedule Analysis)*
Phân tích thời khóa biểu của trường Đại học Khoa học Tự nhiên để đưa ra những thông tin liên quan đến giảng viên, thời gian học, phòng học của trường.
Trình bày câu trả lời trong một file *Jupyter notebook* với tên được đặt theo mẫu

`Assignment11_FullName_ScheduleAnalysis.ipynb`

6. *(Do Anything You Prefer)*
Hãy đặt ra một câu hỏi và tìm câu trả lời cho câu hỏi đó thông qua một bộ dữ liệu nào đó. Hoặc chọn một bộ dữ liệu và trả lời những câu hỏi liên quan. Trình bày câu trả lời của bạn theo cách bạn muốn.
Bonus: làm thế nào để biết được ngày sinh nhật của một bạn cùng trường mà mình thầm để ý?

Related Topics

- Modin.pandas package.
- Interactive mode in Jupyter notebook.



Hình 1: How to make data powerful?