BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC GIA ĐỊNH**

**KHOA: ……………………..**



**BÁO CÁO BÀI TẬP**

**MÔN: KHAI PHÁ DỮ LIỆU**

**Ngành**: **CÔNG NGHỆ THÔNG TIN**

**Chuyên ngành: HỆ THỐNG THÔNG TIN**

**Giảng viên hướng dẫn: *Nguyễn Thị Phương Trang***

**Nhóm 8: Hồ Mỹ Quang Trung, Nguyễn Trường Duy**

**Lớp: K15HTT01**

TP. Hồ Chí Minh, tháng 11 năm 2023

**Khoa/Viện: …………………………**

**NHẬN XÉT VÀ CHẤM ĐIỂM CỦA GIẢNG VIÊN**

**TIỂU LUẬN MÔN: ………………………………………..**

1. **Họ và tên sinh viên: ……………………………………………………………………**
2. **Tên đề tài**:

1. **Nhận xét**:

***Những kết quả đạt được:***

***Những hạn chế:***

1. **Điểm đánh giá** *(theo thang điểm 10, làm tròn đến 0.5):*

Sinh viên:……………………………………………………………………………….

Điểm số: ……….…… Điểm chữ: ……………………………………………………..

|  |  |
| --- | --- |
|  | *TP. HCM, ngày … tháng … năm 20……*  **Giảng viên chấm thi**  *(Ký và ghi rõ họ tên)* |

Bài thực hành số 1

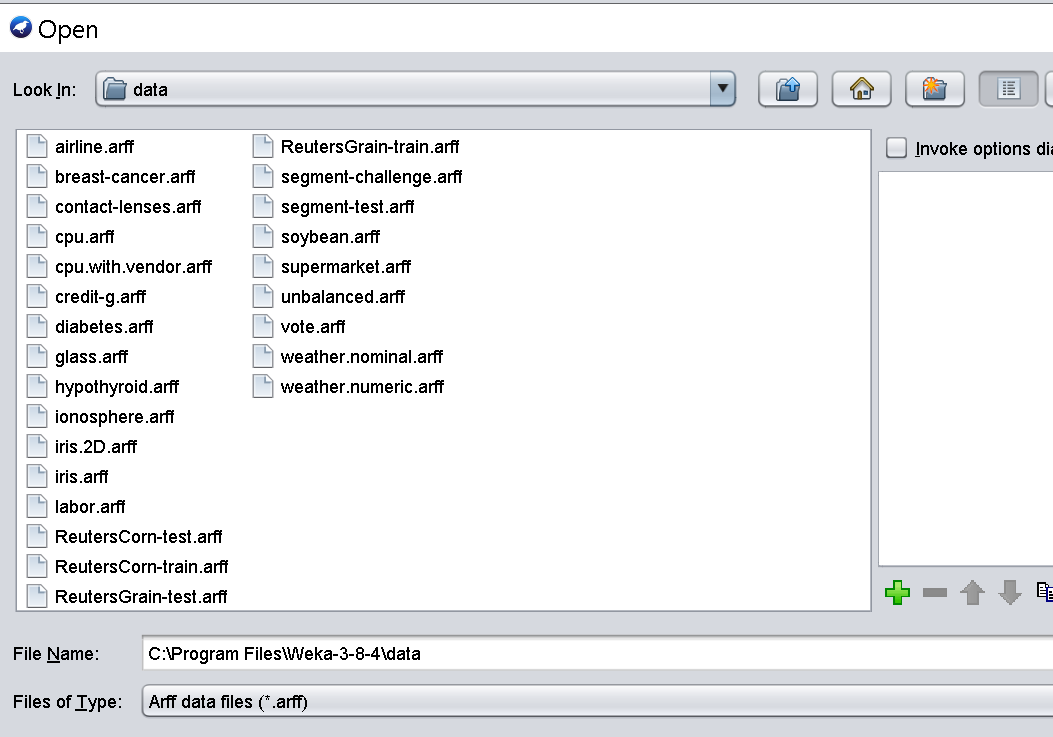
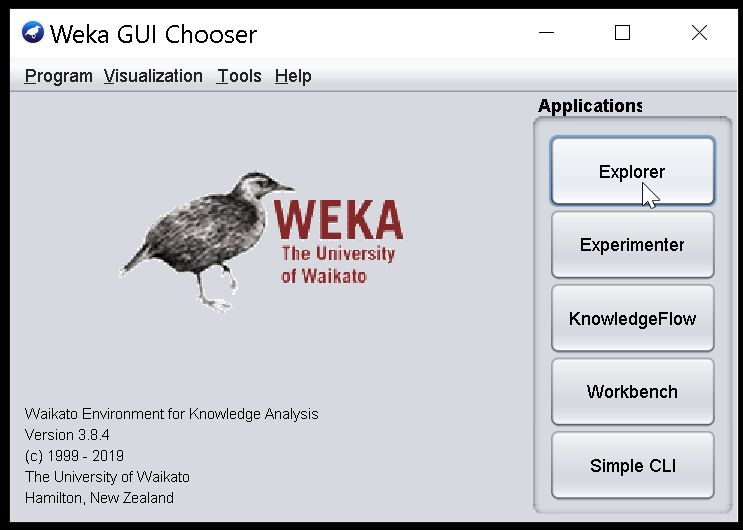
Thành lập nhóm, tối đa 2 sinh viên/nhóm.

Các câu phải làm: Câu 2. Câu 1,3 tìm hiểu thêm lấy điểm cộng.

Thời gian nộp: trước 11:30 am ngày 17/11/2023

**Câu 1:**

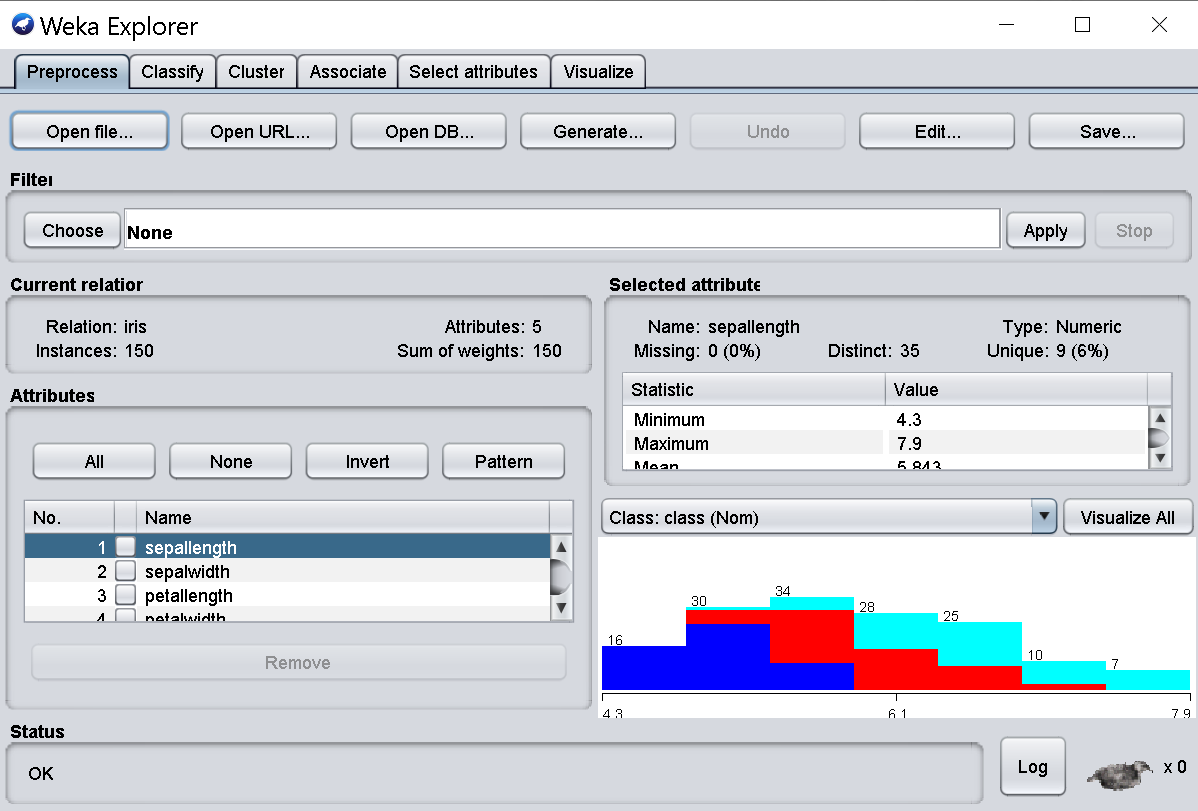
Sử dụng mạng Internet để tìm kiếm, tải về, cài đặt phần mềm WEKA của đại học Waikato New Zealand. Phần mềm này hoạt động dựa trên ngôn ngữ Java. Sau khi cài đặt xong, thử thực hiện Weka và dùng Explorer để thử nạp vào vài tập tin dữ liệu (do Weka cung cấp sẵn) để tìm hiểu các công cụ cho phép tìm hiểu các đại lượng trong các thuộc tính của bảng dữ liệu.



Lưu ý: Tập tin dữ liệu của Weka có phần mở rộng là arff. Có thể xem trực tiếp bằng các công cụ soạn thảo văn bản như NotePad, Note++, …

Ngoài ra Weka cũng cho phép người dùng nhập (import) dữ liệu từ các nguồn khác nhau (bảng tính, tập tin CSDL,…).

Nhóm cần tìm hiểu tất cả các thông tin, các tính năng trong thẻ tab Preprocess



**Câu 2:** Làm quen với việc tổng hợp dữ liệu và làm sạch dữ liệu.

Có 4 tập tin dạng bảng tính (Excel) lưu thông tin sinh viên gồm mã số, họ lót, tên, ngày sinh (ngày tháng và năm). Nhóm có nhiệm vụ tổng hợp 4 danh sách này thành 1 danh sách duy nhất (tổng hợp) sau đó làm sạch thông tin (phát hiện các cụm từ sai, các dữ liệu không hợp lệ,…) Kết quả làm việc sẽ có 2 tập tin: tập tin đầu tiên chứa danh sách có thông tin đầy đủ và chính xác, tập tin thứ 2: chứa các sinh viên có thông tin chưa đầy đủ hoặc không chính xác. Và 1 file word mô tả cách làm.

Nhóm có quyền sử dụng bất kỳ công cụ gì để thực hiện nhiệm vụ và bất kỳ ứng dụng gì để lưu lại kết quả làm việc (lưu ở dạng text, bảng tính, bảng dữ liệu Access, …).

**Câu 3:** Phần thu hoạch

Phân công thực hiện, thực hiện ở nhà, thời hạn 1 tuần và gửi vào chủ đề tronx1g diễn đàn.

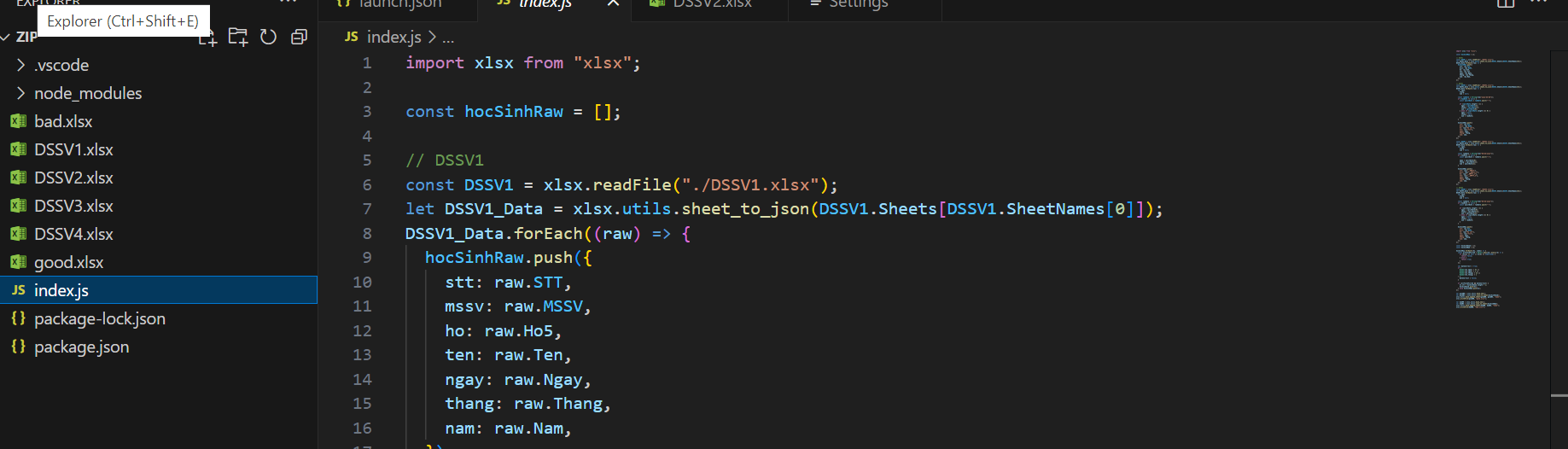
* Hướng dẫn cài đặt và làm quen WEKA.
* Hướng dẫn cài đặt và làm quen ngôn ngữ lập trình Python. Minh họa việc dùng Python để viết các chương trình đơn giản sau:

1. Nhập 4 số nguyên, in ra số lớn nhất và nhỏ nhất trong 4 số.
2. Nhập (hoặc gán trước) 1 mảng các số thực. Đếm xem có bao nhiêu số dương.
3. Nhập (hoặc gán trước) các số nguyên vào 1 mảng 2 chiều có số hàng số cột tùy ý chọn trước. Sau đó in ra các phần tử nằm trên hàng thứ k (với k do người sử dụng nhập vào).

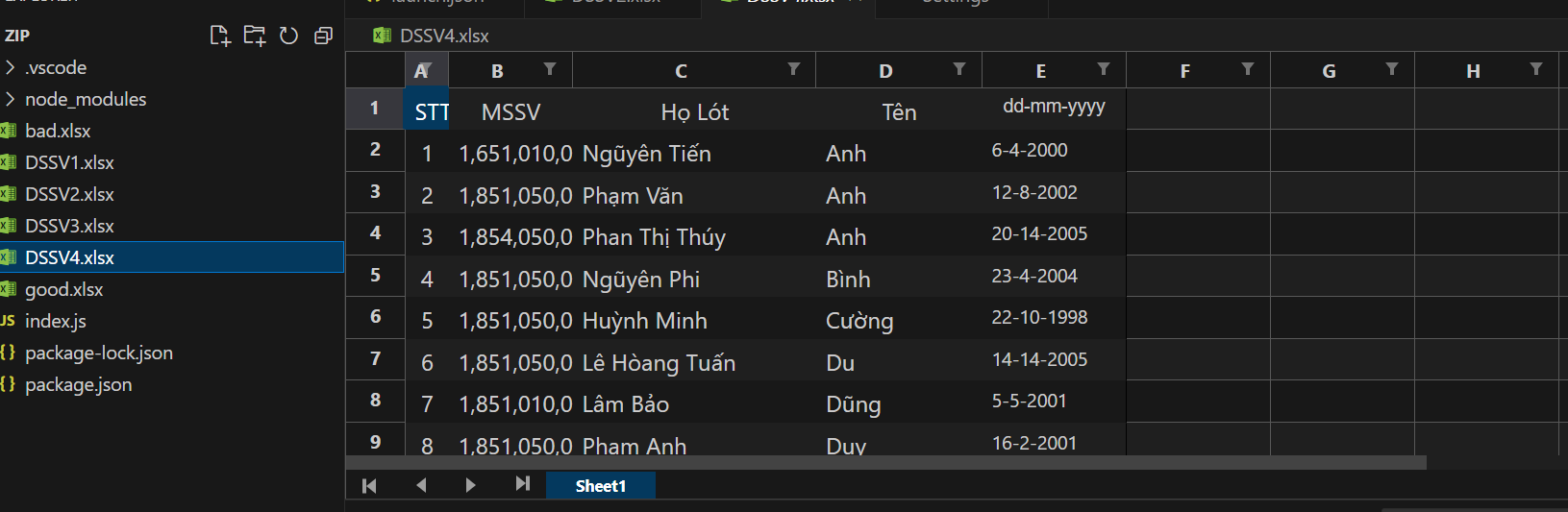
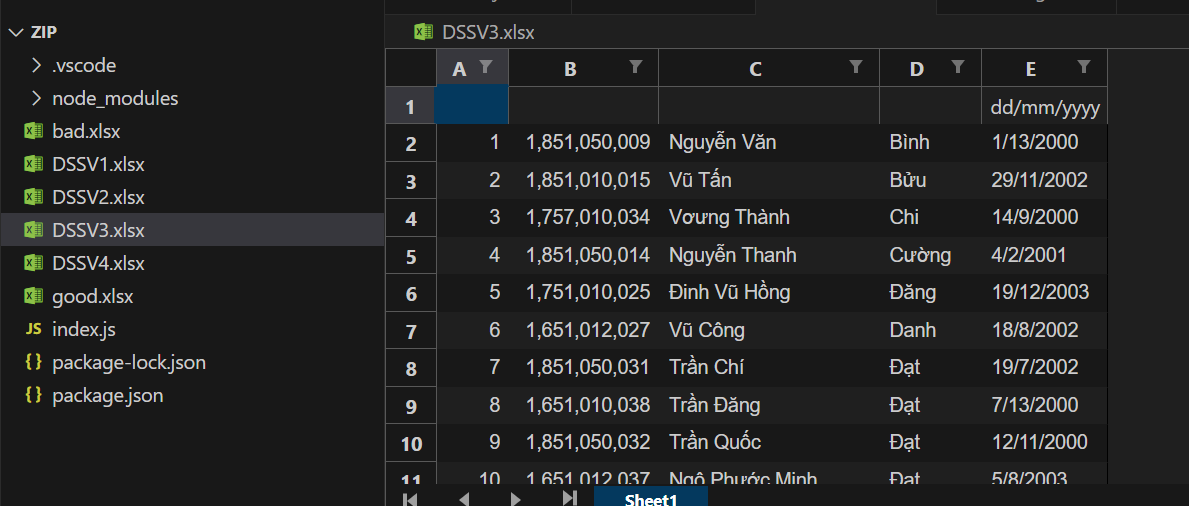
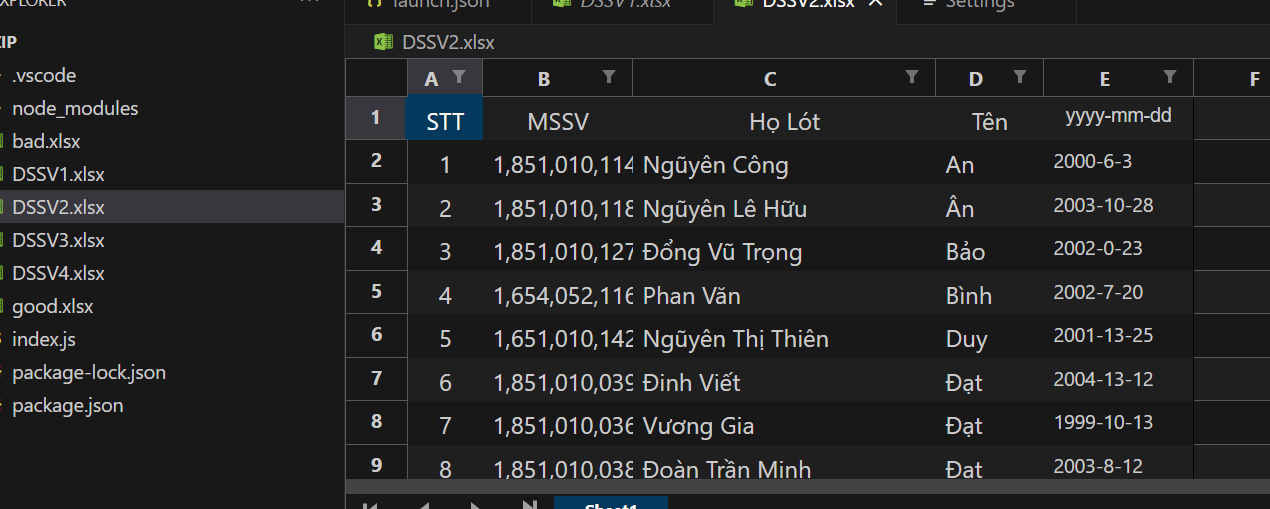
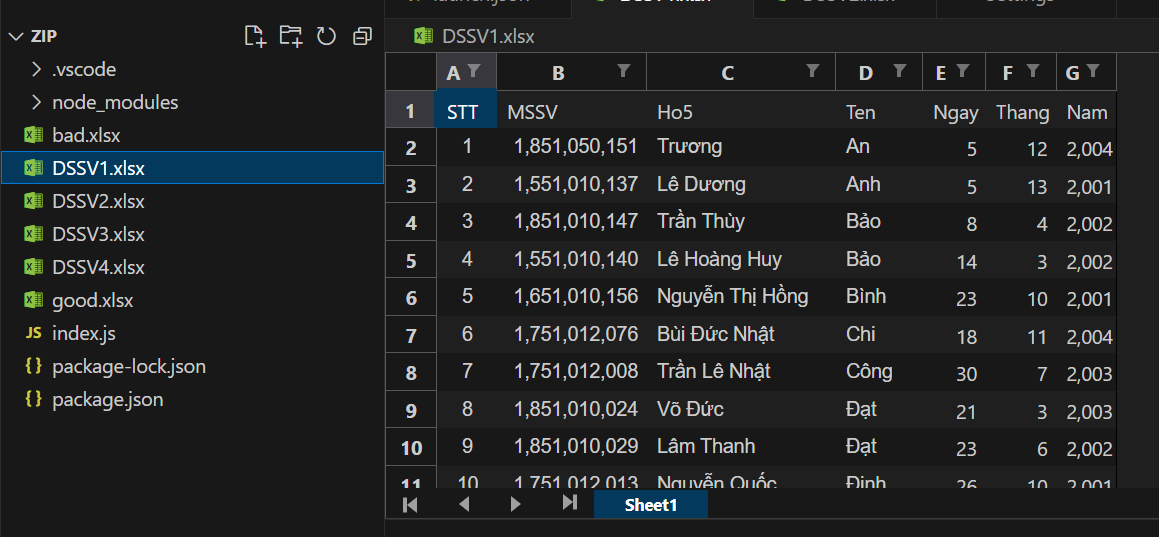
Lưu ý: Nhóm có thể dùng bất kỳ ứng dụng để thực hiện báo cáo (Power Point, Word, Clip/Video, …).

**Bài làm:**

Để giải quyết bài toán ở câu 2, nhóm em sẽ sử dụng ngôn ngữ javascript trên ứng dụng ViSualStudio Code

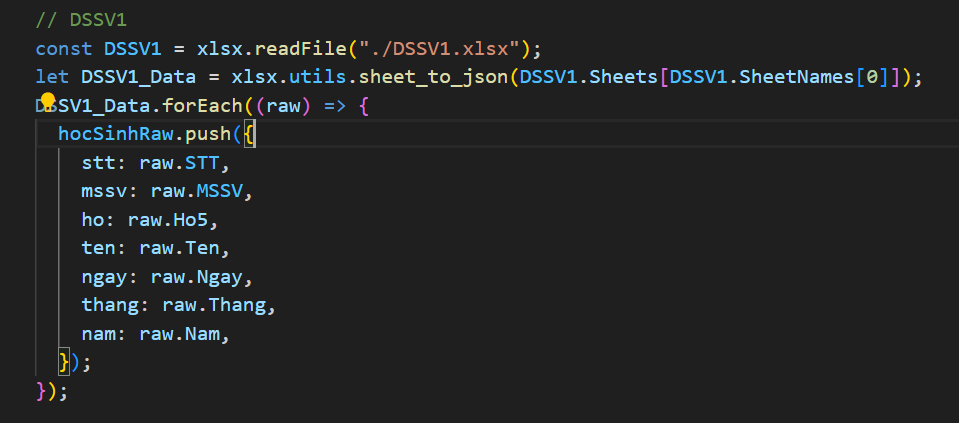


Việc đầu tiên là cài Excel viewer extension để có thể xem được file xlsx của excel trong ứng dụng, sau đó dùng câu lệnh import để import thư viện xlsx nhờ đó giúp ta đọc các file .xlsx DSSV 1-4.



Sau khi đọc các file xong, chúng ta bắt đầu viết code để lọc dữ liệu. Khai báo hằng số hocsinhraw là một mảng rỗng để lưu trữ các dữ liệu sau này.

Bắt đầu với bảng thứ 1:



Khai báo một hằng số (constant) có tên là DSSV1 và gán cho nó kết quả đọc file excel từ đường dẫn "./DSSV1.xlsx" bằng thư viện xlsx., tiếp theo khai báo một biến có tên là DSSV1\_Data và gán cho nó kết quả chuyển đổi nội dung của sheet trong file excel DSSV1 thành một mảng JSON sử dụng phương thức sheet\_to\_json từ thư viện xlsx. Sử dụng vòng lặp for-each để duyệt qua từng phần tử của mảng DSSV1\_Data. Trong mỗi lần lặp, đoạn mã thực hiện việc thêm một đối tượng vào mản hocSinhRaw. Đối tượng này chứa các key-value tương ứng với dữ liệu từ file excel, ví dụ như stt, mssv, ho, ten, ngay, thang, nam. Vì bảng thứ nhất chỉ có các hàng với các thông tin bị thiếu vậy nên việc lọc chỉ cần đơn giản là đảm bảo các hàng được điền đầy đủ thông tin thì mới được push vào mảng hocsinhraw.

Bảng thứ 2



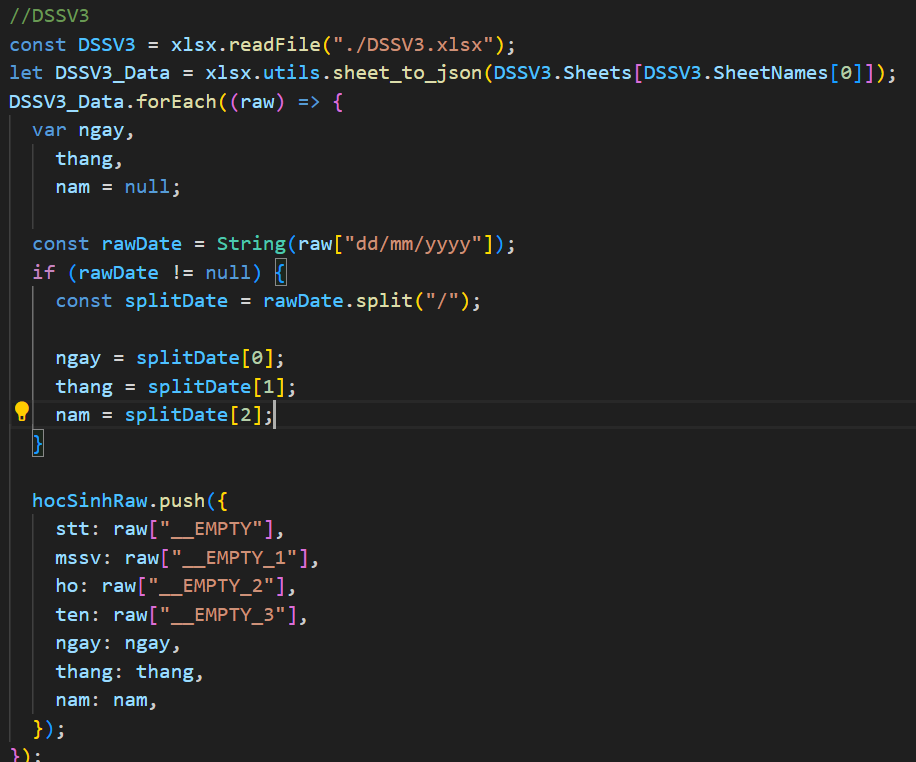
Tương tự với bảng DSSV1, ta import file DSSV2.xlsx và gán kết quả vào hằng số DSSV2. Tiếp đó khai báo biến DSSV2\_Data và gán cho nó kết quả chuyển đổi nội dung của sheet trong file excel DSSV2 thành một mảng JSON sử dụng phương thức sheet\_to\_json từ thư viện xlsx. Tiếp tục sử dụng vòng lặp For-Each với chức năng tương tự ở bản trước nhưng sẽ lọc dựa trên ngày, trong khối lệnh này, có một biến ngày (ngay), tháng (thang), năm (nam) được khởi tạo với giá trị mặc định là null.

Đoạn mã kiểm tra xem có dữ liệu ngày tháng năm (yyyy-mm-dd) trong đối tượng raw hay không. Nếu có, chuỗi ngày tháng năm được tách thành các phần bằng dấu "-" sử dụng phương thức split.v

Nếu chuỗi ngày tháng năm được tách thành nhiều phần tử (chiều dài splitDate > 2), thì ngày, tháng và năm được gán giá trị tương ứng từ splitDate.

Nếu chuỗi ngày tháng năm không thể tách thành các phần tử (chiều dài splitDate === 0), thì ngày, tháng và năm được gán giá trị null và giá trị của rawDate được gán cho năm.

Bảng thứ 3:



Vẫn tương tự với 2 bảng trước, import thư viện XLSX, khai báo biến DSSV3\_Data, gán cho nó kết quả chuyển đổi nội dung của sheet trong file excel DSSV3 thành một mảng JSON. Sử dụng vòng lặp forEach để duyệt qua từng phần tử của mảng DSSV3\_Data.

Trong khối lệnh này, có một biến ngày (ngay), tháng (thang), năm (nam) được khởi tạo với giá trị mặc định là null.

Đoạn mã chuyển đổi dữ liệu ngày tháng được lấy từ trường ["dd/mm/yyyy"] của đối tượng raw thành chuỗi bằng cách chuyển nó thành kiểu dữ liệu String.

Kiểm tra xem rawDate có giá trị khác null hay không.

Nếu giá trị của rawDate không phải là null, đoạn mã sẽ tách chuỗi rawDate thành các phần tử sử dụng dấu "/" thông qua phương thức split và gán giá trị tương ứng cho ngày, tháng, năm.

Cuối cùng, đoạn mã thêm một đối tượng vào mảng hocSinhRaw. Đối tượng này được khởi tạo với các thuộc tính tương ứng từ dữ liệu trong file Excel DSSV3.

Bảng số 4:



Cách làm tương tự với bảng số 2, sử dụng vòng lặp For-Each với chức năng tương tự ở bản trước nhưng sẽ lọc dựa trên ngày, trong khối lệnh này, có một biến ngày (ngay), tháng (thang), năm (nam) được khởi tạo với giá trị mặc định là null.

Đoạn mã kiểm tra xem có dữ liệu ngày tháng năm (yyyy-mm-dd) trong đối tượng raw hay không. Nếu có, chuỗi ngày tháng năm được tách thành các phần bằng dấu "-" sử dụng phương thức split.v

Nếu chuỗi ngày tháng năm được tách thành nhiều phần tử (chiều dài splitDate > 2), thì ngày, tháng và năm được gán giá trị tương ứng từ splitDate.

Nếu chuỗi ngày tháng năm không thể tách thành các phần tử (chiều dài splitDate === 0), thì ngày, tháng và năm được gán giá trị null và giá trị của rawDate được gán cho năm.

Tổng hợp dữ liệu đã được chọn lọc:



Khởi tạo hai mảng rỗng hocSinhFiltered và hocSinhUnqualifed để lưu trữ thông tin về học sinh.

Sử dụng vòng lặp forEach để duyệt qua từng phần tử của mảng hocSinhRaw.

Trong mỗi lần lặp, đoạn mã kiểm tra xem tất cả các trường dữ liệu của học sinh có được điền đầy đủ hay không. Sử dụng phương thức Object.keys(hs) để lấy danh sách các key (trường) của đối tượng hs và phương thức every để kiểm tra xem tất cả các trường đều có giá trị khác null hoặc undefined hay không.

Biến allFieldFilled sẽ được gán giá trị true nếu tất cả các trường đều được điền đầy đủ và false nếu có ít nhất một trường mà giá trị là null hoặc undefined.

Kiểm tra xem giá trị ngày và tháng trong đối tượng hs có hợp lệ hay không. Nếu giá trị ngày lớn hơn 31 hoặc nhỏ hơn 1, hoặc giá trị tháng lớn hơn 12 hoặc nhỏ hơn 1, biến dateCorrect sẽ được gán giá trị false.

Nếu tất cả các trường được điền đầy đủ và giá trị ngày và tháng hợp lệ, đoạn mã tiếp tục thêm đối tượng hs vào mảng hocSinhFiltered. Đồng thời, cập nhật giá trị trường "stt" của hs bằng giá trị tương ứng trong mảng hocSinhFiltered.

Nếu có ít nhất một trường không được điền đầy đủ hoặc giá trị ngày và tháng không hợp lệ, đối tượng hs sẽ được thêm vào mảng hocSinhUnqualifed.

Sau khi hoàn thành việc xử lý, đoạn mã tạo hai work book mới (FilteredWB và UnqualifedWB) bằng phương thức xlsx.utils.book\_new().

Tiếp theo, mỗi mảng hocSinhFiltered và hocSinhUnqualifed được chuyển thành một sheet trong work book tương ứng thông qua phương thức xlsx.utils.json\_to\_sheet.

Phương thức xlsx.utils.book\_append\_sheet được sử dụng để thêm sheet vào work book với tên "name".

Cuối cùng, đoạn mã ghi work book thành file Excel tương ứng bằng phương thức xlsx.writeFile.

Kết quả là đoạn mã tạo hai file Excel mới, "Filtered.xlsx" và "Unqualifed.xlsx", chứa thông tin về học sinh đã được lọc. File "Filtered.xlsx" chứa học sinh có tất cả các trường điền đầy đủ và giá trị ngày tháng hợp lệ, trong khi file "Unqualifed.xlsx" chứa học sinh có ít nhất một trường không điền đầy đủ hoặc giá trị ngày tháng không hợp lệ.