E1:
Crawl_Data - BeautifulSoup: sys,pandas,requests,time,csv

Khó khăn:
+ Tìm hiểu và sử dụng cấu trúc HTML
+ Khi sử dụng các thư viện liên quan (cài đặt thư viện, các bug khi code khó khăn khi fix,...)
+ Bị lỗi đường dẫn, chặn truy cập
+ Khi phân tích HTML với lượng thông tin lớn, mã hóa dữ liệu
+ Dữ liệu có thể chứa các ký tự đặc biệt hoặc mã hóa không đúng, gây lỗi khi xử lý.
+ Xử lí dữ liệu bị thiếu

Phân tích các ý chính bài toán:

+ Lấy danh sách cầu thủ
+ Truy cập trang chi tiết từng cầu thủ và thu thập chỉ số
+ Xử lý các chỉ số không có hoặc không áp dụng
+ Xếp và lưu kết quả vào file

E2:

Khó khăn:
+ Thiếu dữ liệu (các trường hợp của việc gán dữ liệu N/A)
+ Số lượng chỉ số lớn khi tính toán, đòi hỏi mã phải tối ưu và mất nhiều thời gian trong khi tính toán
+ Xử lí dữ liệu thiếu

+ Cần phải phân tích nhiều nhóm
+ Dữ liệu không đồng nhất so với thông tin tương ứng
+ Xử lí dữ liệu lớn

+ Việc định dạng dữ liệu (đúng kiểu số, dữ liệu không bị thừa)
+ Quản lí số lượng biểu đồ
+ Sắp xếp các trục của histogram

+ Tổng hợp nhiều chỉ số cần phải đánh giá và xem xét
+ Tìm tiêu chí hợp lí

E3:

K-means:
Khó khăn:
+ Xác định số lượng nhóm
+ Dữ liệu có nhiều chỉ số, việc phân nhóm phức tạp
+ Xử lí dữ liệu không trùng lặp (khác nhau về vị trí thi đấu)

--> Phân tích:
+ Số nhóm tối ưu phản ánh cách phân nhóm cho cầu thủ như độ liên kết theo cụm, vị trí, phong cách chơi, thành tích cá nhân
+ Kết quả phân nhóm cho thấy các nhóm cầu thủ có phong cách thi đấu

Khó khăn:

+ Việc số chiều giảm xuống 2 làm mất mát thông tin, gây ra thông tin có thể không chính xác
+ Trùng lặp nhóm
+ Lựa chọn chỉ số phù hợp