Bài 1:

Mục tiêu chính của chương trình:

- Thu thập dữ liệu thống kê của các cầu thủ có số phút thi đấu trên 90 phút tại giải Ngoại hạng Anh.
- Sắp xếp dữ liệu theo tên (First Name) và giảm dần theo độ tuổi nếu trùng tên.
- Xuất dữ liệu ra file CSV với tên là results.csv, trong đó mỗi cột tương ứng với một chỉ số cụ thể.

Chỉ số thu thập: Chương trình sẽ thu thập nhiều chỉ số liên quan đến cầu thủ, bao gồm:

- Quốc gia, đội bóng, vị trí và tuổi của cầu thủ.
- Thời gian thi đấu, bao gồm số trận đã chơi, số trận xuất phát và tổng số phút thi đấu.
- Hiệu suất ghi bàn, với các chỉ số như số bàn thắng không phạt đền, số bàn thắng phạt đền, số kiến tạo, cũng như số thẻ vàng và thẻ đỏ.
- Các chỉ số dự đoán như xG (Expected Goals) và npxG (Non-Penalty Expected Goals).
- Thống kê về các hành động tấn công và phòng ngự.

Cấu trúc và logic của chương trình: Chương trình sẽ thực hiện các bước chính sau:

1. **Thu thập dữ liệu**: Sử dụng thư viện Selenium để tự động truy cập vào các trang thống kê của từng đội bóng và thu thập dữ liệu cầu thủ.
2. **Xử lý dữ liệu**: Dữ liệu thu thập được sẽ được chuyển đổi thành DataFrame bằng thư viện Pandas, giúp dễ dàng xử lý và sắp xếp.
3. **Lọc cầu thủ**: Chương trình sẽ lọc ra những cầu thủ có số phút thi đấu lớn hơn 90 phút.
4. **Sắp xếp và xử lý dữ liệu trống**: Dữ liệu sẽ được sắp xếp theo tên (First Name) và nếu có trùng tên, sẽ sắp xếp theo độ tuổi từ cao đến thấp. Các trường dữ liệu trống sẽ được xử lý và ghi nhận là "N/a".
5. **Lưu trữ kết quả**: Cuối cùng, dữ liệu đã được xử lý sẽ được xuất ra file results.csv, với cấu trúc mà mỗi cột tương ứng với một chỉ số, và thứ tự các cầu thủ đã được sắp xếp theo yêu cầu.

Bài 2:
Mục tiêu chính của đoạn mã này là xử lý và phân tích dữ liệu cầu thủ bóng đá từ file results.csv, bao gồm các nhiệm vụ sau:

1. Tìm ra 3 cầu thủ có điểm cao nhất và thấp nhất ở mỗi chỉ số.
2. Tính toán các thông số thống kê như trung vị, trung bình và độ lệch chuẩn cho từng chỉ số, cả theo toàn giải và theo từng đội.
3. Lưu kết quả vào các file results1.txt ,csv, png để đưa ra kết quả
4. Vẽ biểu đồ histogram để minh họa phân bố của các chỉ số.

Các bước chính trong chương trình

Đọc và chuẩn bị dữ liệu:

1. Đầu tiên, chương trình sử dụng thư viện Pandas để đọc file results.csv và kiểm tra danh sách các cột. Điều này giúp xác định cấu trúc dữ liệu và lựa chọn đúng các cột chỉ số cần phân tích.
2. Một số cột như Team, Nation, và Position không cần thiết cho phân tích sẽ được loại bỏ khỏi danh sách các chỉ số.

Xác định các chỉ số và tìm top/bottom 3 cầu thủ:

1. Sau khi xác định được các cột số cần phân tích, chương trình thực hiện việc tìm kiếm 3 cầu thủ có điểm cao nhất và 3 cầu thủ có điểm thấp nhất ở từng chỉ số. Kết quả này được lưu vào một dictionary có tên là top_bottom_players.
2. Mỗi kết quả sẽ được ghi vào file results1.txt theo định dạng dễ đọc, hiển thị rõ ràng top 3 và bottom 3 cho từng chỉ số.

Tính toán thống kê cho toàn giải và từng đội:

1. Chương trình tính trung vị, trung bình và độ lệch chuẩn cho toàn bộ giải đấu (được gọi là "All") và lưu kết quả vào dictionary stats.
2. Sau đó, dữ liệu sẽ được nhóm theo từng đội bóng để tính toán các chỉ số thống kê tương tự. Kết quả cuối cùng sẽ được lưu vào DataFrame stats_df và xuất ra file results2.csv. File này có cấu trúc bảng rõ ràng, dễ theo dõi, giúp người dùng so sánh thống kê giữa các đội.

Vẽ biểu đồ phân phối bằng histogram:

1. Đoạn mã có thể được mở rộng để sử dụng matplotlib.pyplot.hist vẽ biểu đồ histogram cho từng chỉ số. Việc này giúp người dùng dễ hình dung sự phân bố của các chỉ số, cả ở mức toàn giải và theo từng đội bóng. Biểu đồ histogram cung cấp cái nhìn trực quan hơn về cách các chỉ số phân bố giữa các cầu thủ.

Câu 3:

Lựa chọn phân loại:

Chương trình sử dụng thuật toán K-means để phân loại các cầu thủ thành các nhóm dựa trên các chỉ số giống nhau. Cầu thủ được phân loại thành bốn nhóm tương ứng với bốn vị trí khác nhau trên sân: thủ môn, hậu vệ, tiền vệ và tiền đạo. Việc phân loại này giúp hiểu rõ hơn về đặc điểm và phong cách chơi của từng nhóm cầu thủ.

Lý do lựa chọn phân loại

Kết quả phân loại cho thấy cầu thủ ở mỗi vị trí có những đặc điểm khác nhau về các chỉ số như số bàn thắng, kiến tạo, và các chỉ số phòng ngự. Sự phân nhóm này giúp

nhận diện điểm mạnh của cầu thủ trong mỗi vị trí và cung cấp cái nhìn tổng quan về năng lực trong các vị trí khác nhau.

Thuật toán PCA

Tiếp theo, thuật toán PCA (Phân tích thành phần chính) được áp dụng để giảm số chiều dữ liệu xuống còn hai chiều. Việc này giúp trực quan hóa các cụm dữ liệu trên mặt phẳng 2D, cho thấy rõ sự phân tách giữa các nhóm cầu thủ. Trực quan hóa giúp dễ dàng nhận diện các cụm và xác định mối liên hệ giữa các chỉ số.