

Báo Cáo Phân Tích Dữ Liệu Bóng Đá

Case study 12
Lớp: DHKL18A1

A. Giới thiệu

Vấn đề case 12:

Các câu lạc bộ bóng đá tham gia giải V.League (hoặc các giải đấu tương tự) thường được thống kê về danh sách đội bóng, cầu thủ, vị trí thi đấu, số bàn thắng và kết quả trận đấu. Case Study này giúp sinh viên thực hành làm sạch dữ liệu, truy vấn, groupBy, merge và pivot trong bối cảnh phân tích dữ liệu thể thao, cụ thể là bóng đá chuyên nghiệp tại Việt Nam.

Mục tiêu:

Mô tả **quy trình xử lý – phân tích dữ liệu**, trình bày **các kết quả chính**, đưa ra **nhận xét**, và **kết luận tổng hợp**. Báo cáo được trình bày theo logic: *Làm sạch dữ liệu → Truy vấn & thống kê → Phân tích nhóm → Hợp nhất dữ liệu → Phân tích tổng hợp bằng pivot*.

Các thành viên nhóm 9

	Thành viên	Mã sinh viên
1	Nguyễn Quang Trường *	24174600008
2	Cao Nguyên Ngọc	24174600052
3	Mai Thị Ngọc Huệ	24174600175
4	Nguyễn Hữu Hải Đăng	24174600038
5	Nguyễn Đức Nhật	24174600176

Phương pháp làm việc nhóm

Các thành viên cùng cùng làm 1 nhiệm vụ, sau đó trao đổi về cách làm và so sánh các kết quả để sửa lại lỗi. Nhóm trưởng có vai trò thống nhận xét các bài làm → thống nhất cấu trúc dự án → trình bày lại bài khoa học, hiệu quả hơn

B. Quy trình xử lý - phân tích dữ liệu

1. Đọc & làm sạch dữ liệu

Mục tiêu : Xử lý các dữ liệu thô thành dạng chuẩn phục vụ phân tích.

Dữ liệu đầu vào: các file CSV thô gồm thông tin đội bóng (team_info), cầu thủ (player_info) và kết quả trận đấu (match_results).

Các bước xử lý chính:

- Tạo hàm xử lý chung khoảng trắng chung cho các file.
- Xử lý giá trị thiếu bằng cách loại bỏ hoặc thay thế phù hợp.
- Chuẩn hóa đúng các dạng thông tin về cùng 1 kiểu và ép kiểu `astype()`

Kết quả: Dữ liệu sau làm sạch được lưu vào thư mục `data_clean/` để đảm bảo các bước phân tích sau sử dụng dữ liệu thống nhất và đáng tin cậy.

M001,T07,T07,3,3,2024-04-24	M001,T07,T07,3,3,2024-04-24
M002, T03,T07, 2,3,2024-01-08	M002,T03,T07,2,3,2024-01-08
M003,T07,T05,four,1,13/04/2024	M003,T07,T05,4,1,2024-04-13
M004, T02, T01, 2, 0,08/05/2024	M004,T02,T01,2,0,2024-05-08
M005,T05,T06 ,3 ,2,19/06/2024	M005,T05,T06,3,2,2024-06-19
M006, T03,T06 ,3 ,1 ,2024-01-28	M006,T03,T06,3,1,2024-01-28
M007, T03, T01,0,3,2024-02-18	M007,T03,T01,0,3,2024-02-18
M008, T01, T03,four,2,2024-06-17	M008,T01,T03,4,2,2024-06-17
M009, T02,T05,3 ,0,2024-06-21	M009,T02,T05,3,0,2024-06-21

Kết quả xử lý file `match_results.csv`

2. Truy vấn & thống kê

Mục tiêu : Tạo các bảng thống kê cơ bản.

Phương pháp:

- Sử dụng `groupby()` để đếm số đội theo thành phố (`home_city`)
- Thống kê số lượng cầu thủ theo mỗi đội (`team_id`)
- Kết hợp (merge) thông tin cầu thủ với tên đội để tạo bảng dễ đọc.

Ý nghĩa: Cung cấp cái nhìn tổng quan về phân bố đội bóng và nhân sự giữa các đội.

	team_id	player_count	team_name
0	T01	14	Clb Thanh Pho Hcm
1	T02	9	Clb Hai Phong
2	T03	11	Clb Ha Noi
3	T04	14	Clb Hai Phong
4	T05	9	Clb Thanh Pho Hcm
5	T06	8	Clb Ha Noi
6	T07	13	Clb Binh Duong
7	T08	13	Clb Binh Duong

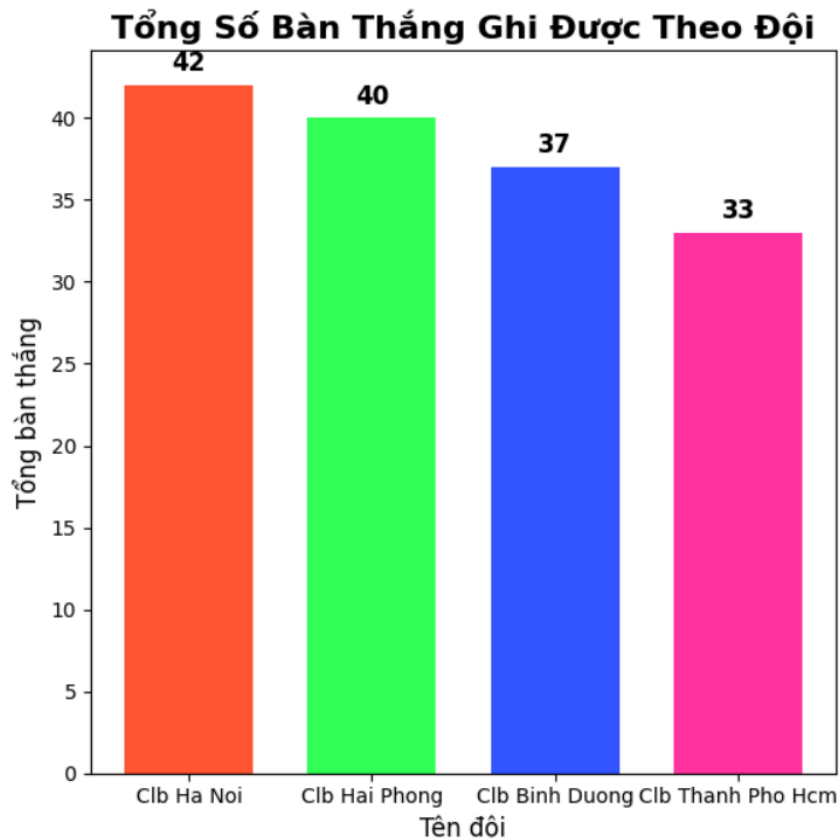
Bảng thông kê số lượng cầu thủ mỗi đội

3. Phân tích theo nhóm & tổng hợp

Mục tiêu : Phân tích dữ liệu về bàn thắng (tổng, trung bình) để đưa ra nhận xét chất lượng các đội.

Phương pháp:

- Nhóm các thông số về bàn thắng theo cầu thủ và đội bóng bằng `groupby()`
- Dùng `sort_values()` để xác định các đội và cầu thủ nổi bật.
- Tạo biểu đồ cơ bản để trực quan hóa các con số.



Biểu đồ cột thể hiện tổng bàn thắng của các đội trong các mùa

4. Phân tích bằng cách hợp nhất nhiều bảng dữ liệu

Mục tiêu : Kết nối các bảng dữ liệu rời rạc thành bộ dữ liệu hoàn chỉnh, đầy đủ, chi tiết.

Phương pháp:

- Merge thông tin cầu thủ với đội bóng (theo `team_id`).
- Merge dữ liệu trận đấu với thông tin đội nhà và đội khách.
- Kiểm tra sự trùng lặp hoặc bất thường dữ liệu `isna()` giữa các bảng.

Ý nghĩa: Cho phép phân tích kết quả trận đấu gắn với thông tin đội bóng cụ thể.

	match_id	home_team_id	away_team_id	home_goals	away_goals	match_date	home_team_name	home_home_city	away_team_name	away_home_city
0	M001	T07	T07	3	3	2024-04-24	Clb Binh Duong	Ha Noi	Clb Binh Duong	Ha Noi
1	M002	T03	T07	2	3	2024-01-08	Clb Ha Noi	Ha Noi	Clb Binh Duong	Ha Noi
2	M003	T07	T05	4	1	2024-04-13	Clb Binh Duong	Ha Noi	Clb Thanh Pho Hcm	Ha Noi
3	M004	T02	T01	2	0	2024-05-08	Clb Hai Phong	Thanh pho Hcm	Clb Thanh Pho Hcm	Ha Noi
4	M005	T05	T06	3	2	2024-06-19	Clb Thanh Pho Hcm	Ha Noi	Clb Ha Noi	Da Nang

Bảng dữ liệu tổng hợp đầy đủ nhất về các kết quả trận đấu

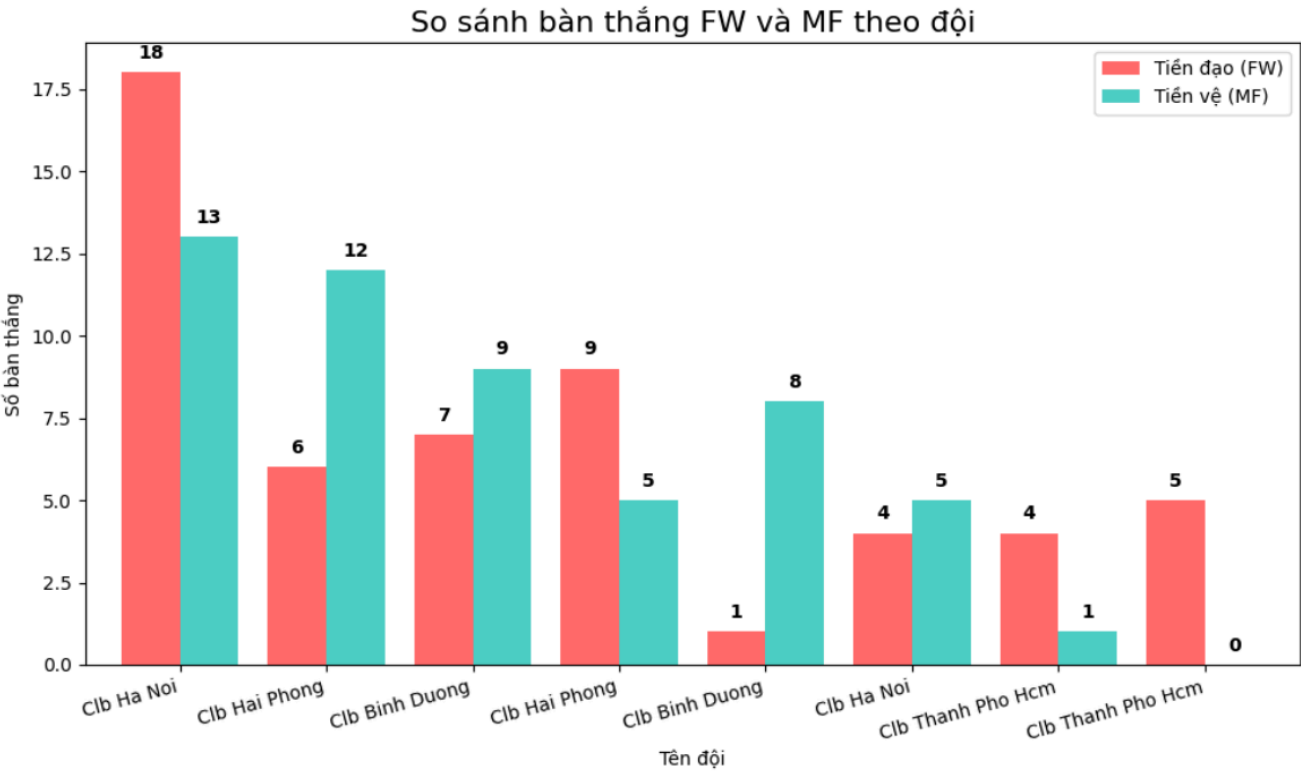
5. Phân tích bằng cách tạo Pivot Table

Mục tiêu : Phân tích với góc nhiều đa chiều hơn.

Phương pháp:

- Tạo bảng 2 chiều tính dữ liệu bàn thắng cùng lúc giữa cầu thủ, đội bóng hoặc thời gian.
- Sử dụng **stack/unstack** thành thạo hơn khi làm việc trên multi index.
- Đánh giá, so sánh các giá trị cùng lúc trên 2 tiêu chí.

Ý nghĩa: Giúp nhận diện xu hướng theo thời gian và sự khác biệt giữa các đội.



Biểu đồ dạng nhóm cột so sánh hàng công các đội

C. Nhận xét dựa trên kết quả phân tích

- Một số đội bóng tập trung nhiều cầu thủ hơn, cho thấy lợi thế về chiều sâu đội hình.
- Hiệu suất ghi bàn không phân bố đồng đều giữa các đội; một số đội/cầu thủ nổi bật đóng góp phần lớn số bàn thắng.
- Phân tích theo tháng cho thấy sự biến động về số bàn thắng, có thể liên quan đến lịch thi đấu hoặc phong độ theo thời gian.

D. Đánh giá chất lượng công việc các thành viên

- Nhìn chung các thành viên đã đạt được hầu hết các nhiệm vụ.
- Các nhiệm vụ về sau tương đối khó, kết quả các bài cho ra khác nhau, xuất hiện nhiều lỗi, nhưng vẫn có cá nhân tìm được giải pháp đúng.
- Ngọc Huệ đóng góp nhiều về ý tưởng, chú thích chi tiết các bước làm giúp thành viên khác dễ hiểu, giải được các bài khó, chất lượng hoàn thiện chưa cao đã cải thiện qua các phần sau.
- Nguyễn Ngọc hoàn thành nhiều phần việc rất nhanh, tỉ lệ lỗi thấp, trình bày cần tối ưu thêm.
- Hải Đăng làm tốt các phần nhiệm vụ được giao, có sáng tạo các phương pháp mới cho các cách tính, còn gặp nhiều lỗi khi xử lý dữ liệu lỗi.
- Đức Nhật có khả năng học hỏi nhanh, xử lý tốt các lỗi gặp phải, một số phần khó chưa tìm được hướng đi đúng.
- Quang Trường (nhóm trưởng) nhận xét và sửa lỗi bài cho các thành viên, hỗ trợ làm các nhiệm vụ khó, trực quan hóa dữ liệu bằng biểu đồ, hoàn chỉnh các bài toán, làm báo cáo.

E. Kết luận

- Việc chuẩn hóa và làm sạch dữ liệu ở giai đoạn đầu đóng vai trò then chốt, giúp đảm bảo độ chính xác và độ tin cậy cho toàn bộ các kết quả phân tích phía sau.
- Các phương pháp phân tích như **groupby**, **merge** và **pivot table** đã cho phép khai thác dữ liệu ở nhiều cấp độ khác nhau: từ tổng quan (phân bố đội bóng, số lượng cầu thủ) đến chi tiết (hiệu suất ghi bàn theo cầu thủ, đội bóng và theo thời gian).
- Kết quả cho thấy hiệu suất thi đấu không đồng đều giữa các đội, trong đó một số đội và cầu thủ có đóng góp vượt trội về số bàn thắng, phản ánh sự khác biệt về chiến thuật, nhân sự và phong độ.
- Tổng kết, báo cáo không chỉ dừng ở việc xử lý và thống kê dữ liệu, mà còn minh họa rõ **giá trị của phân tích dữ liệu trong việc hỗ trợ ra quyết định**. Trong các nghiên cứu tiếp theo, hệ thống phân tích này có thể được mở rộng bằng trực quan hóa dữ liệu, so sánh nhiều mùa giải, hoặc áp dụng các mô hình dự đoán để nâng cao khả năng phân tích và dự báo kết quả thi đấu.