

ĐỀ TÀI: DỰ ĐOÁN KẾT QUẢ CÁC TRẬN ĐẤU BÓNG ĐÁ (THẮNG/HÒA/THUA) TẠI NGOẠI HẠNG ANH

NHẬP MÔN KHOA HỌC DỮ LIỆU

Nhóm 4

Trương Tiến Đạt

Phạm Văn Hoàng Nam

Tống Trọng Tâm

Vũ Châu Minh Trí

Giáo viên hướng dẫn

Thầy Lê Nhựt Nam

GIỚI THIỆU

- Mục đích: Dự đoán kết quả các trận đấu bóng đá (thắng/hòa/thua) tại Ngoại hạng Anh
- Hiện tại: dùng Random Forest



OUTLINE

- 1. Thu thập dữ liệu và hiểu dữ liệu (trục quan hóa)**
- 2. Phân tích thống kê và tiền xử lý dữ liệu**
- 3. Xây dựng mô hình dự đoán và đánh giá, cải thiện**

1. Thu thập dữ liệu và trực quan hóa

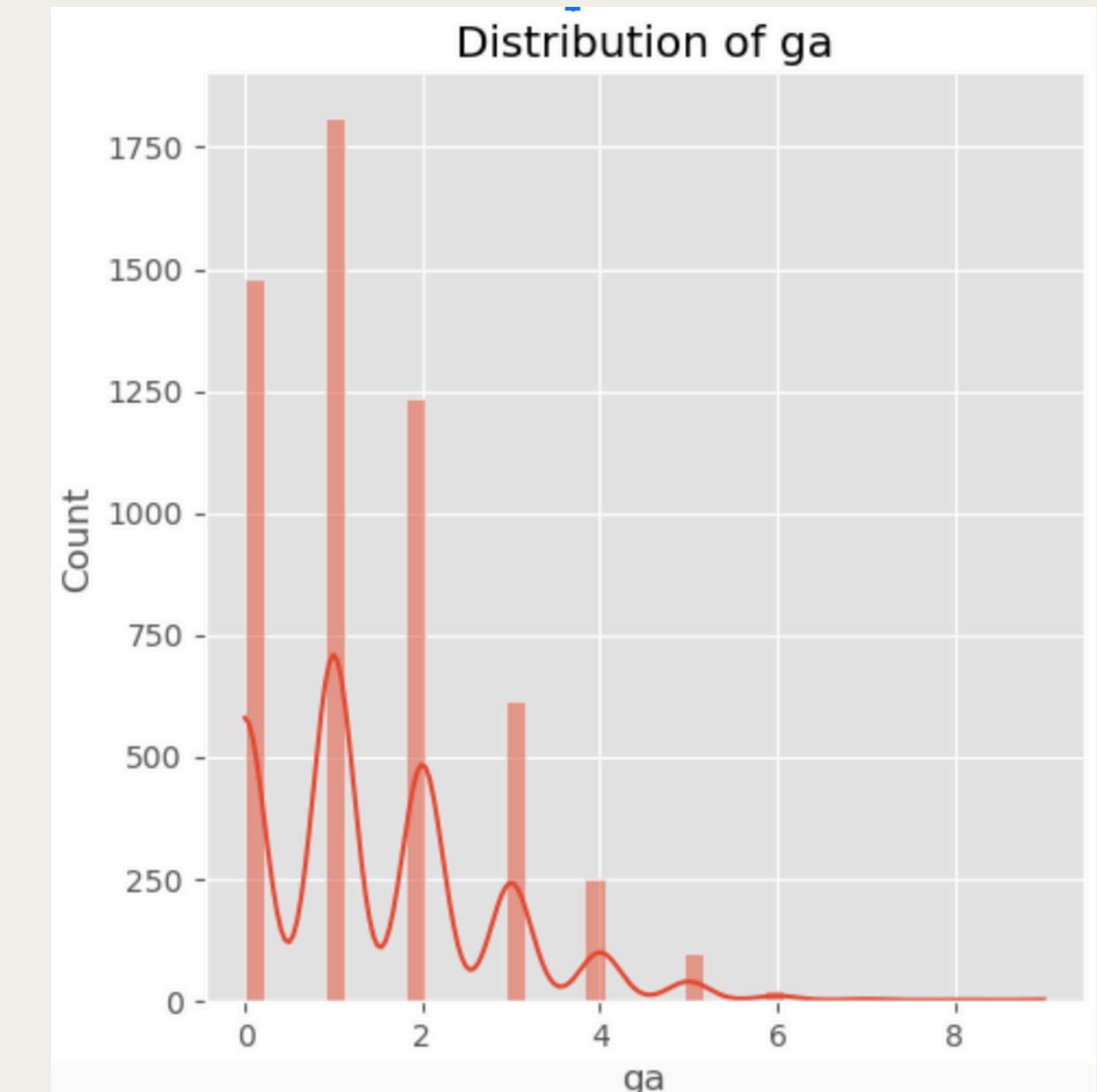
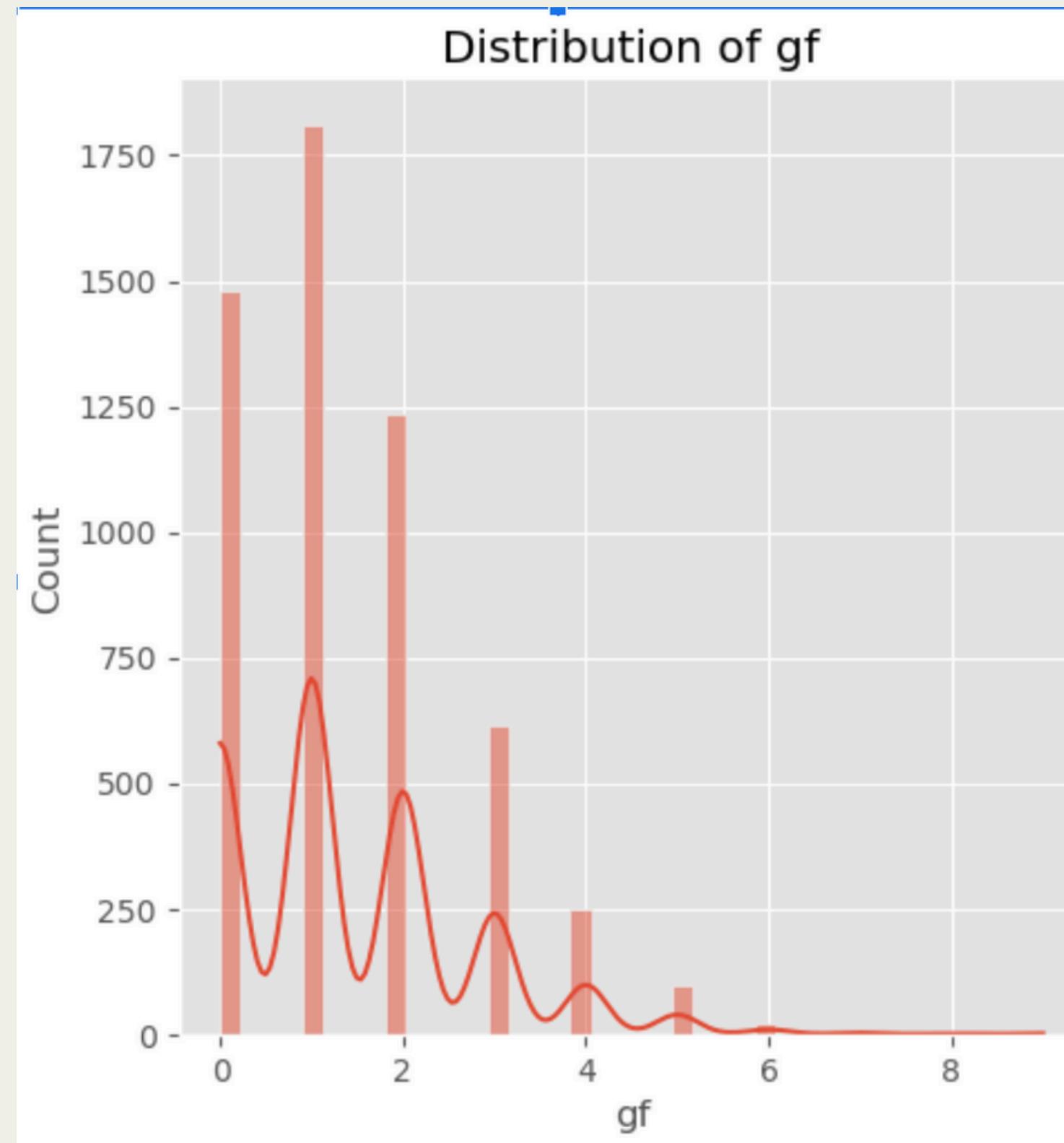
- Nguồn dữ liệu: Dữ liệu toàn bộ trận đấu tại Ngoại hạng Anh từ mùa giải 2017/2018 -> hiện tại (hết vòng 10 2024/2025) (5520 dòng)
- [HTTPS://FBREF.COM](https://fbref.com)
- [HTTPS://WWW.FOOTBALL-DATA.CO.UK](https://www.football-data.co.uk)



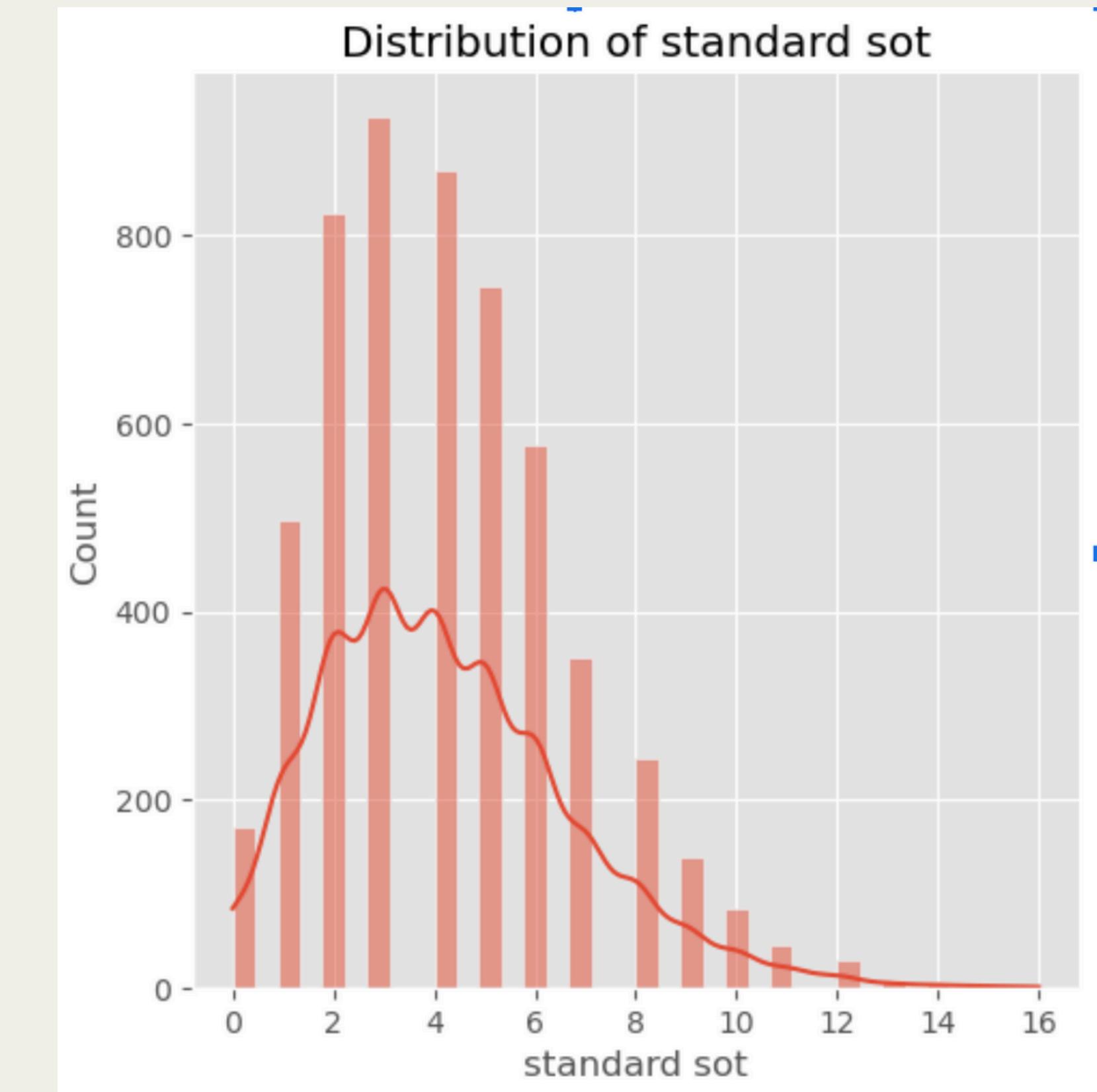
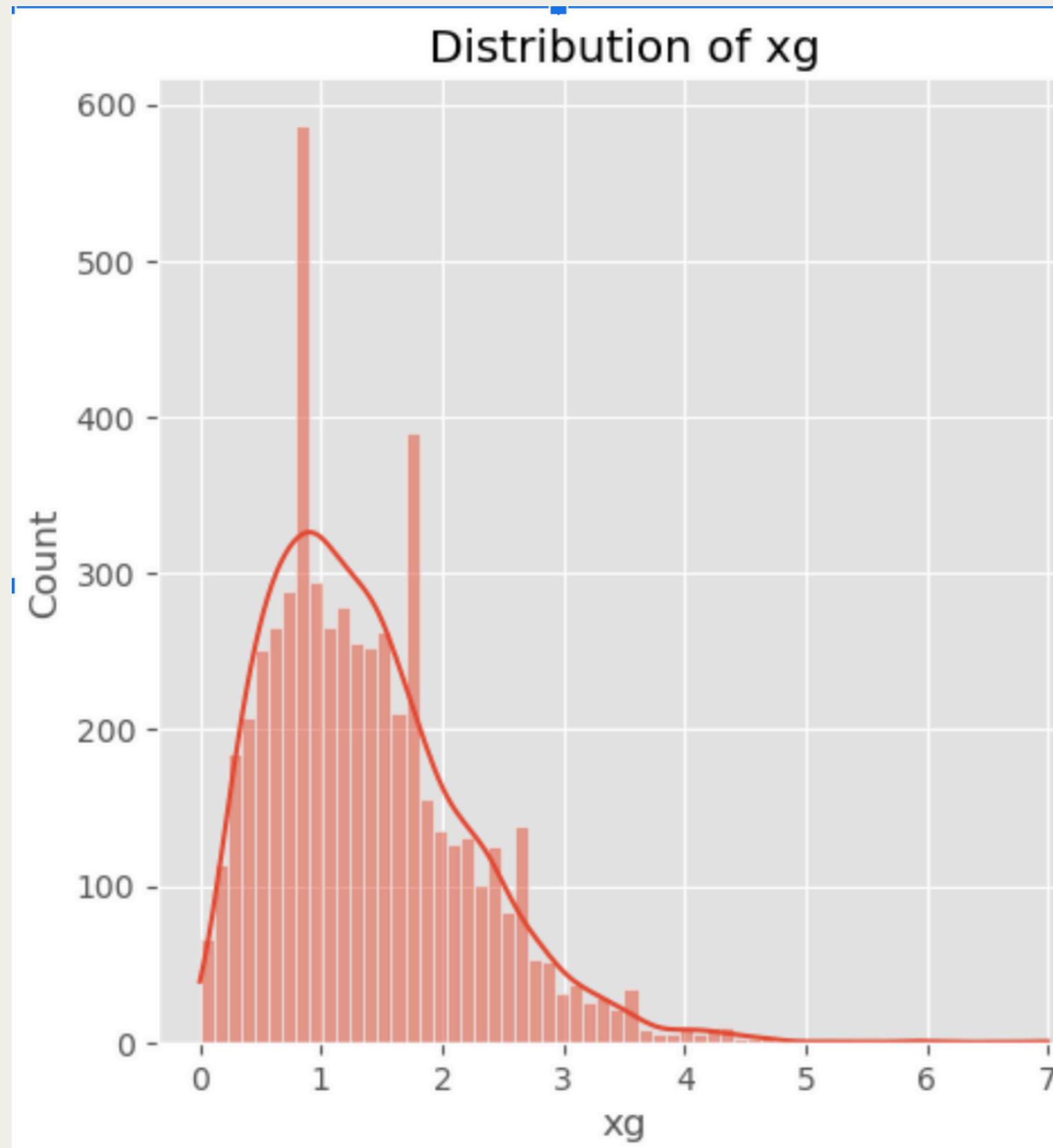
FBREF



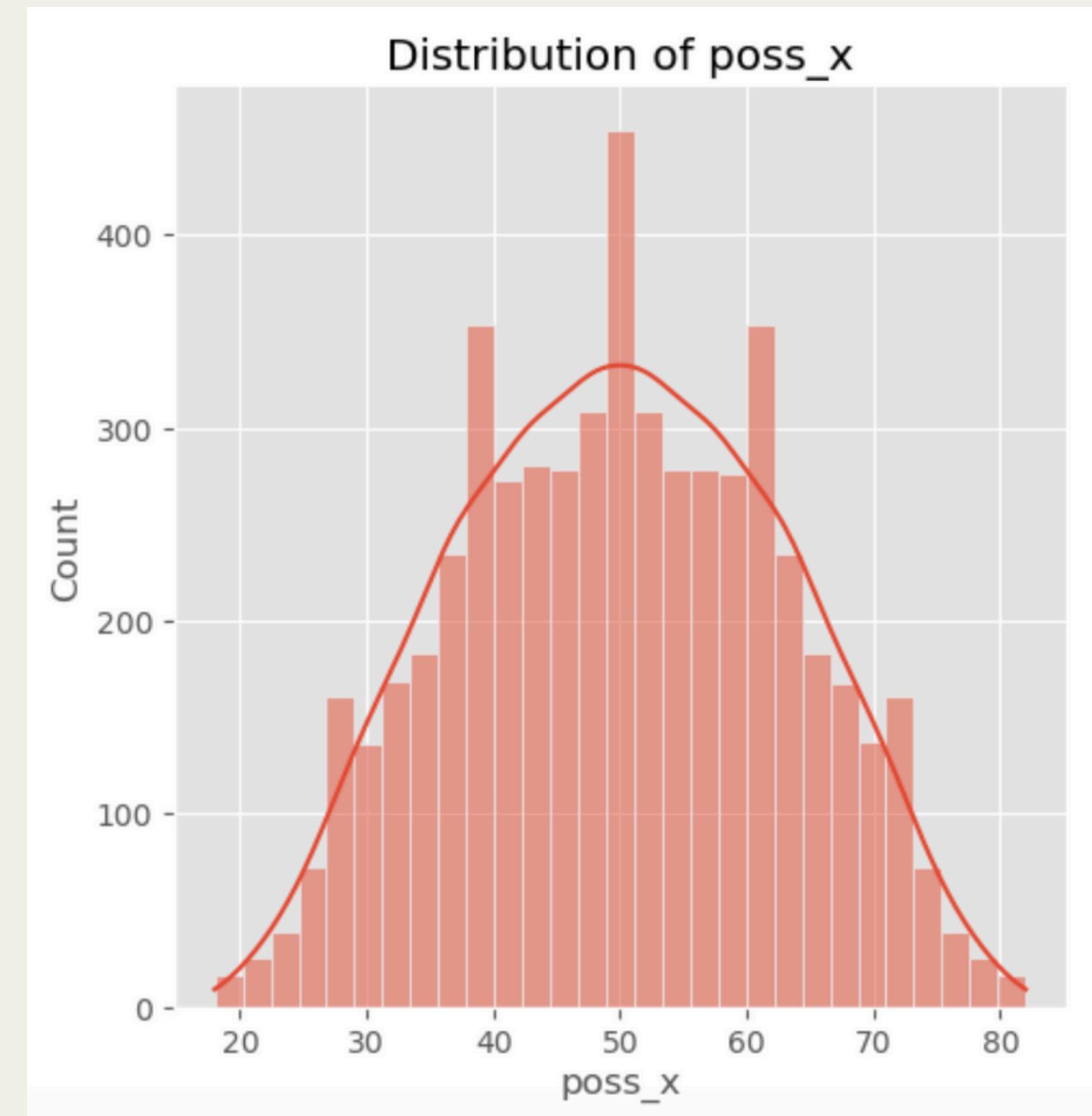
TRỰC QUAN HÓA DỮ LIỆU: gF & gA



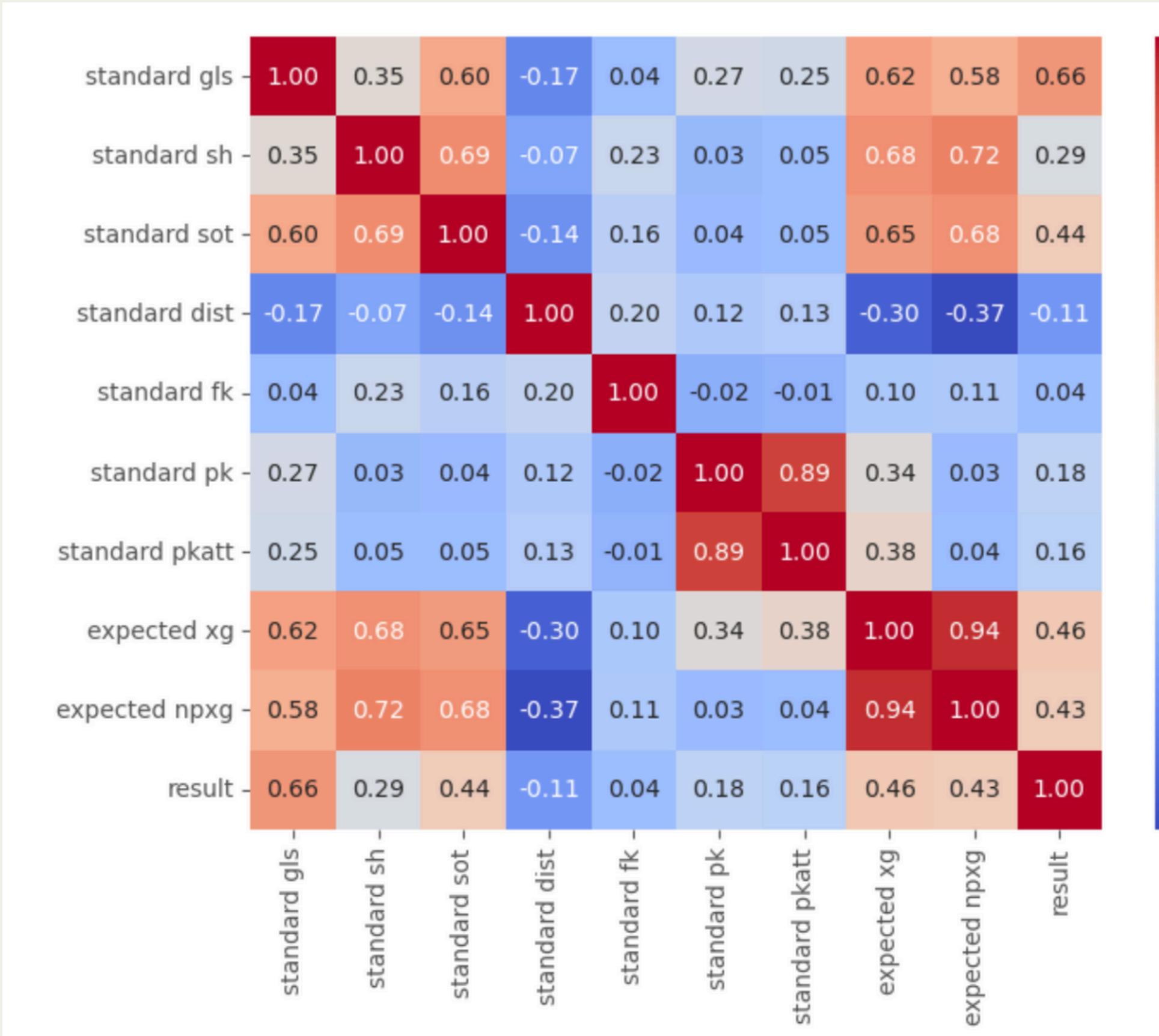
TRỰC QUAN HÓA DỮ LIỆU: xG & standard sot



TRỰC QUAN HÓA DỮ LIỆU: possession



CORRELATION MATRIX



MA TRẬN TƯƠNG QUAN GIỮA CÁC CỘT TRONG NHÓM SHOOTING VÀ RESULT:

Các chỉ số tương quan thuận như **số bàn thắng (standard gls)**, **số lần sút trúng đích (standard sot)** và **số bàn thắng kỳ vọng (xg)** có mối liên hệ chặt chẽ với kết quả, phản ánh rằng hiệu quả tấn công trực tiếp góp phần lớn vào kết quả.

Khoảng cách sút (standard dist) tương quan nghịch với kết quả => **sút xa khó có bàn thắng hơn.**

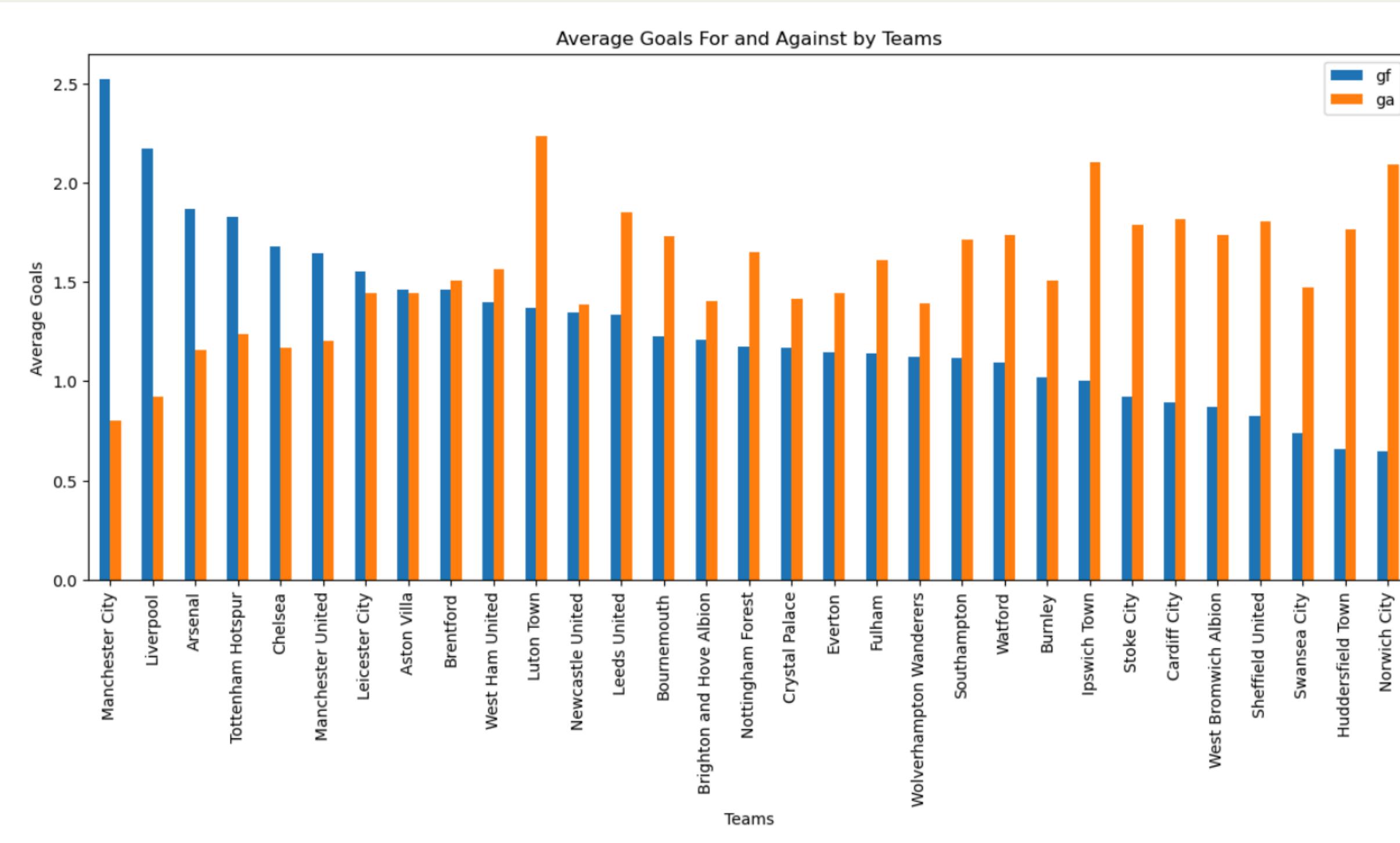
CORRELATION MATRIX

Ma trận tương quan giữa các cột trong nhóm passing và result:
 Để thấy đa phần đều có tương quan thuận với result => **chuyên bóng nhiều có thể dẫn đến kết quả tốt**

ast (số đường kiến tạo) và **xag** (số đường chuyền được kì vọng trở thành kiến tạo) tương quan mạnh với result => **các đường chuyền quyết định và khả năng tạo ra cơ hội ghi bàn có ảnh hưởng trực tiếp đến kết quả của trận đấu.**

	passes cmp	passes att	total passing dist	progressive passing dist	ast	xag	xa	kp	1/3	ppa	crspa	prgp	result
passes cmp	1.00	0.99	0.98	0.85	0.23	0.37	0.54	0.51	0.81	0.62	0.19	0.81	0.20
passes att	0.99	1.00	0.97	0.87	0.21	0.36	0.54	0.51	0.82	0.63	0.21	0.82	0.19
total passing dist	0.98	0.97	1.00	0.90	0.21	0.36	0.53	0.50	0.82	0.61	0.23	0.81	0.19
progressive passing dist	0.85	0.87	0.90	1.00	0.22	0.39	0.54	0.53	0.77	0.62	0.27	0.79	0.20
ast	0.23	0.21	0.21	0.22	1.00	0.57	0.40	0.36	0.15	0.23	0.01	0.17	0.58
xag	0.37	0.36	0.36	0.39	0.57	1.00	0.76	0.71	0.36	0.50	0.23	0.42	0.41
xa	0.54	0.54	0.53	0.54	0.40	0.76	1.00	0.69	0.56	0.70	0.34	0.62	0.30
kp	0.51	0.51	0.50	0.53	0.36	0.71	0.69	1.00	0.55	0.67	0.35	0.62	0.30
1/3	0.81	0.82	0.82	0.77	0.15	0.36	0.56	0.55	1.00	0.66	0.29	0.87	0.14
ppa	0.62	0.63	0.61	0.62	0.23	0.50	0.70	0.67	0.66	1.00	0.49	0.78	0.20
crspa	0.19	0.21	0.23	0.27	0.01	0.23	0.34	0.35	0.29	0.49	1.00	0.31	-0.07
prgp	0.81	0.82	0.81	0.79	0.17	0.42	0.62	0.62	0.87	0.78	0.31	1.00	0.16
result	0.20	0.19	0.19	0.20	0.58	0.41	0.30	0.30	0.14	0.20	-0.07	0.16	1.00

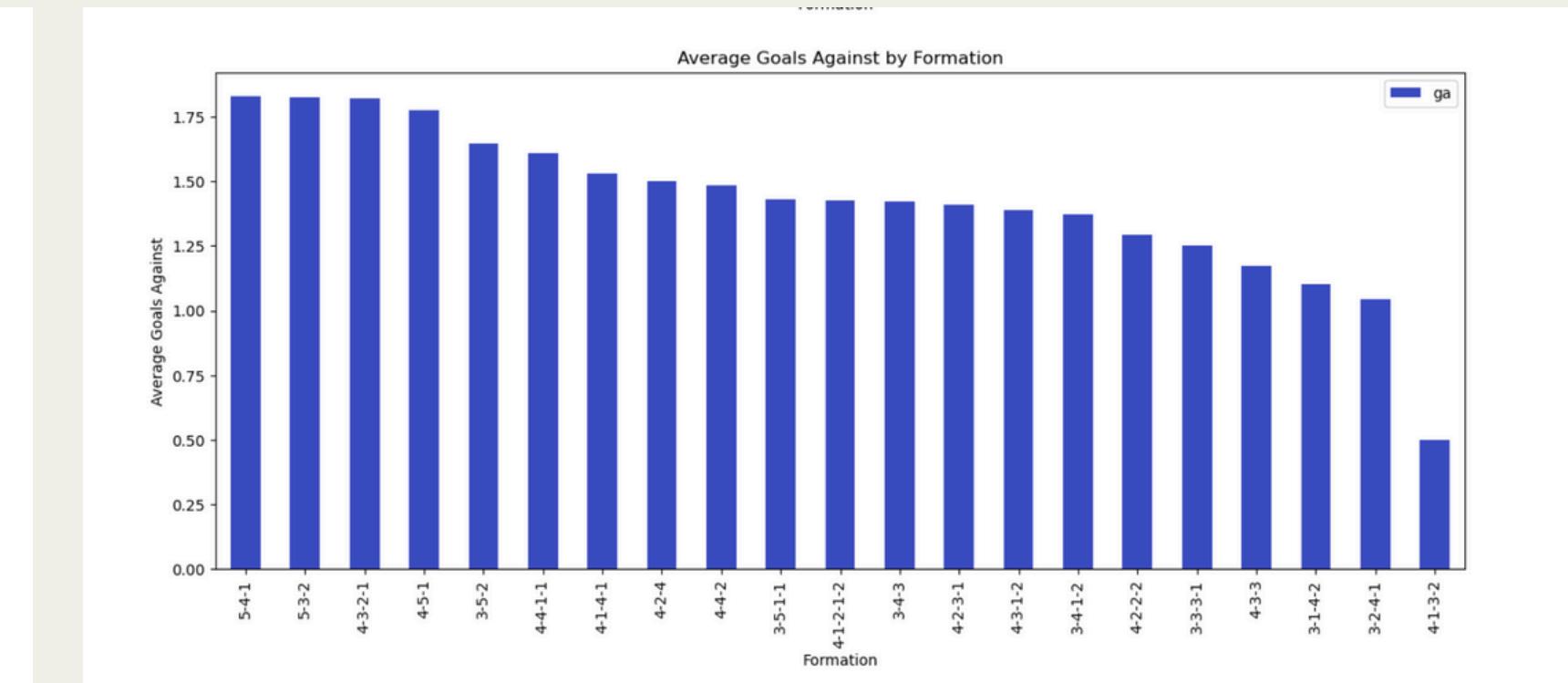
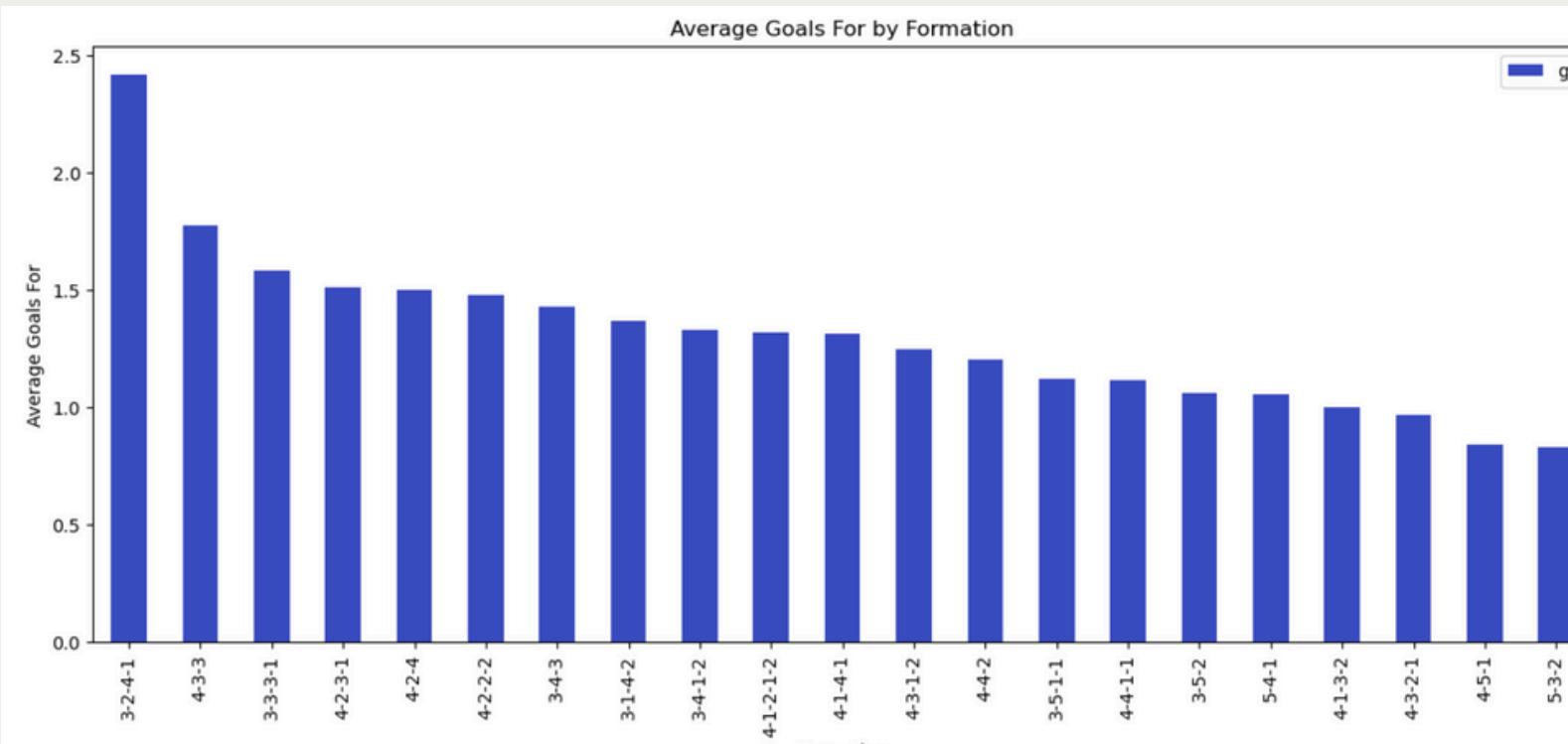
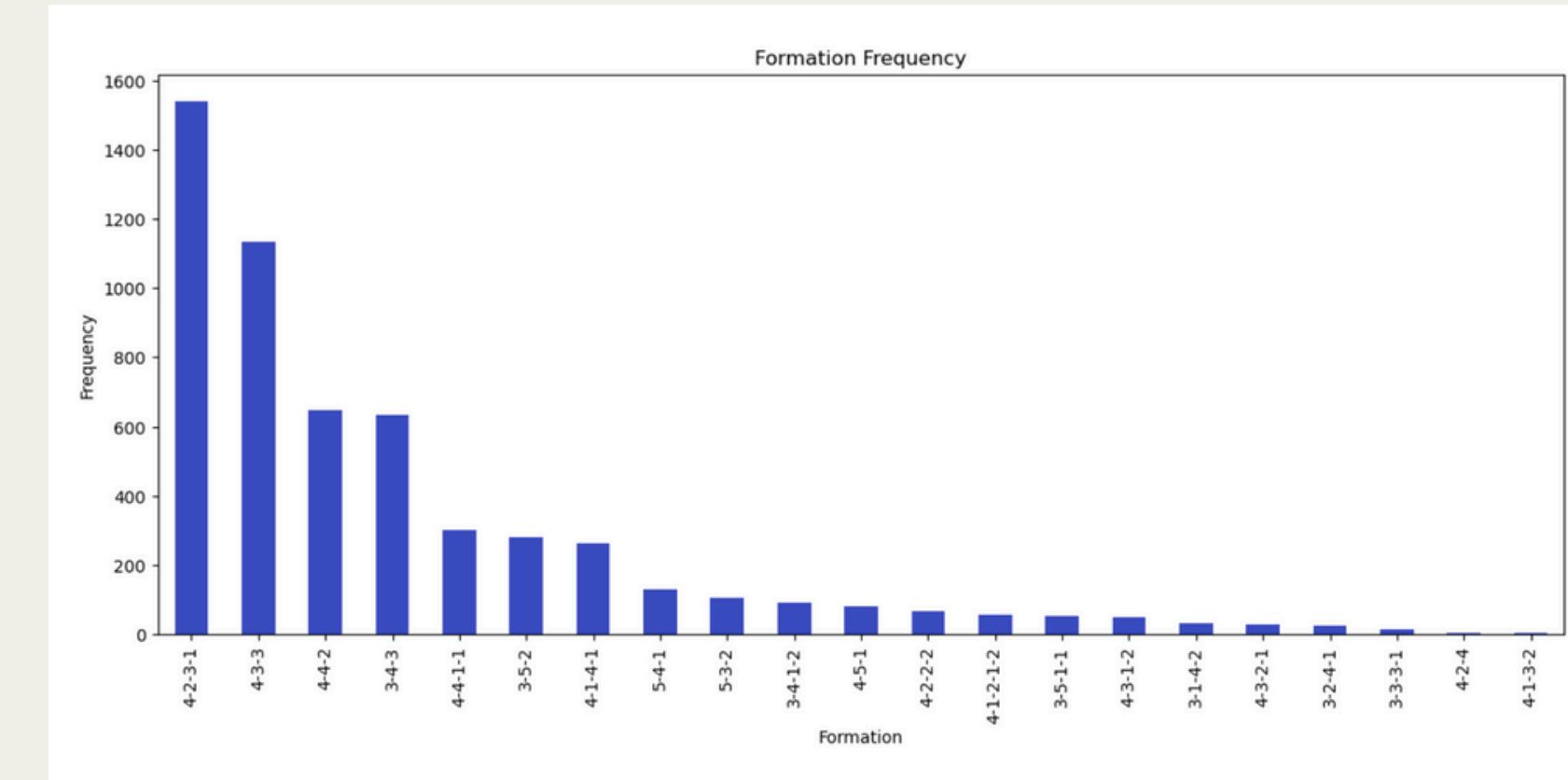
DỮ LIỆU TỪ ĐỘI CỤ THỂ



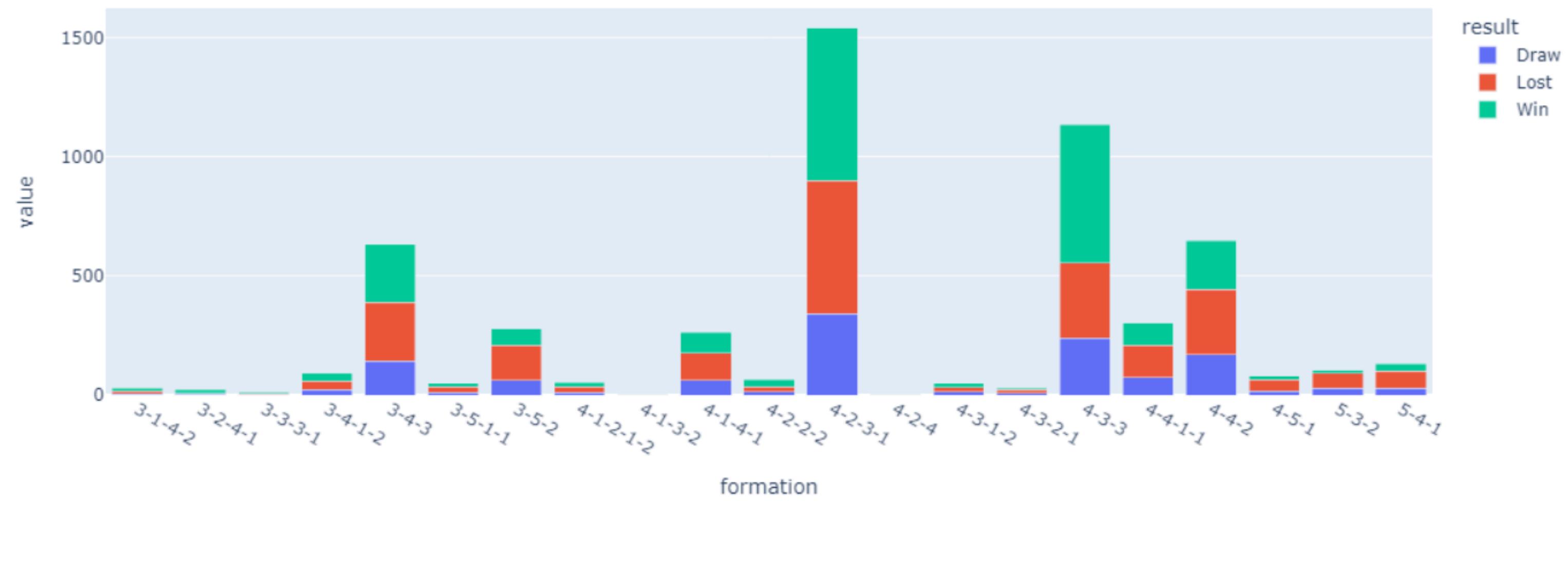
- **Nhóm đầu** (Manchester City, Liverpool, Arsenal,...) có số bàn thắng cao hơn hẳn số bàn thua.
- **Nhóm cuối** (Norwich City, Huddersfield Town, Sheffield United,...) có số bàn thua cao hơn hẳn số bàn thắng.

SỐ BÀN THẮNG CỦA DẠNG ĐỘI HÌNH

DỰA VÀO 3 BIỂU ĐỒ SAU (SỐ TRẬN
ĐƯỢC SỬ DỤNG CỦA MỖI SƠ ĐỒ, TRUNG
BÌNH SỐ BÀN THẮNG/ THUA MỖI TRẬN
KHI DÙNG MỖI SƠ ĐỒ) CÓ THỂ NHẬN
ĐỊNH 4-3-3 VÀ 4-2-3-1 ĐANG LÀ XU
HƯỚNG VÀ ĐƯỢC SỬ DỤNG BỞI NHIỀU
ĐỘI MẠNH

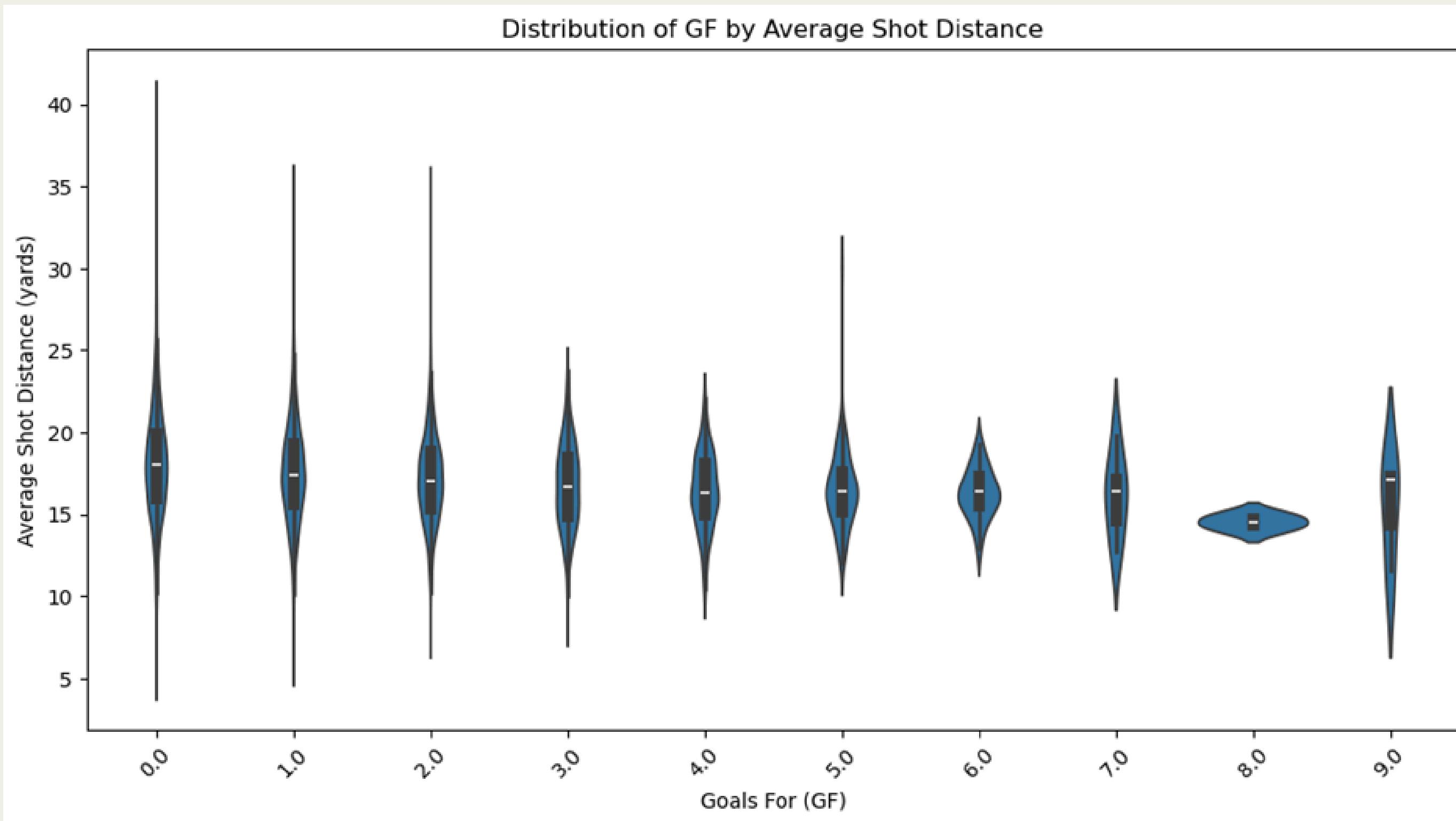


Win/Draw/Loss Distribution by Formation



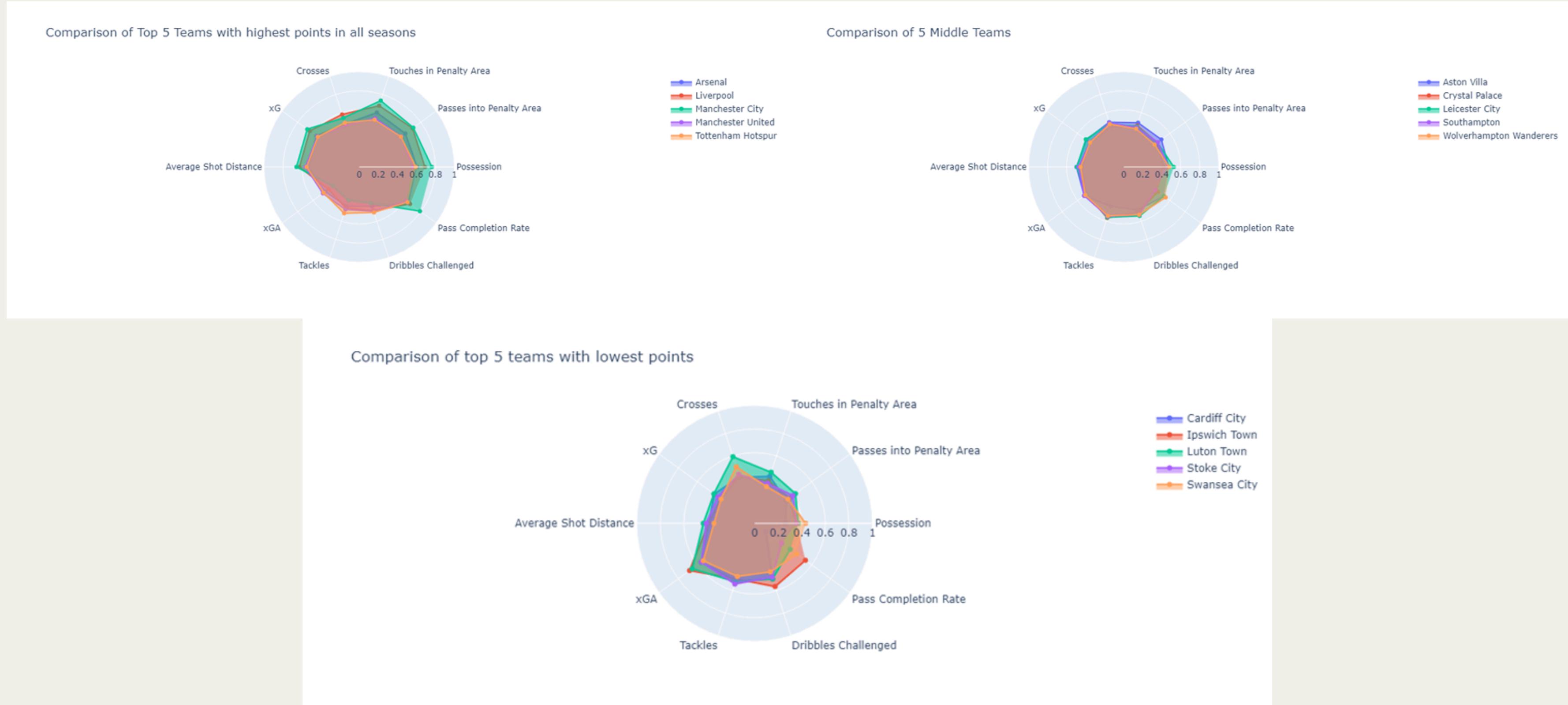
Biểu đồ cũng cỗ nhận định vừa rồi. Cũng dễ nhận thấy theo
ngay sau là **3 - 4 - 4** và **4 - 4 - 2**

Biểu đồ mô tả mối quan hệ giữa số bàn thắng ghi được và khoảng cách trung bình của cú sút



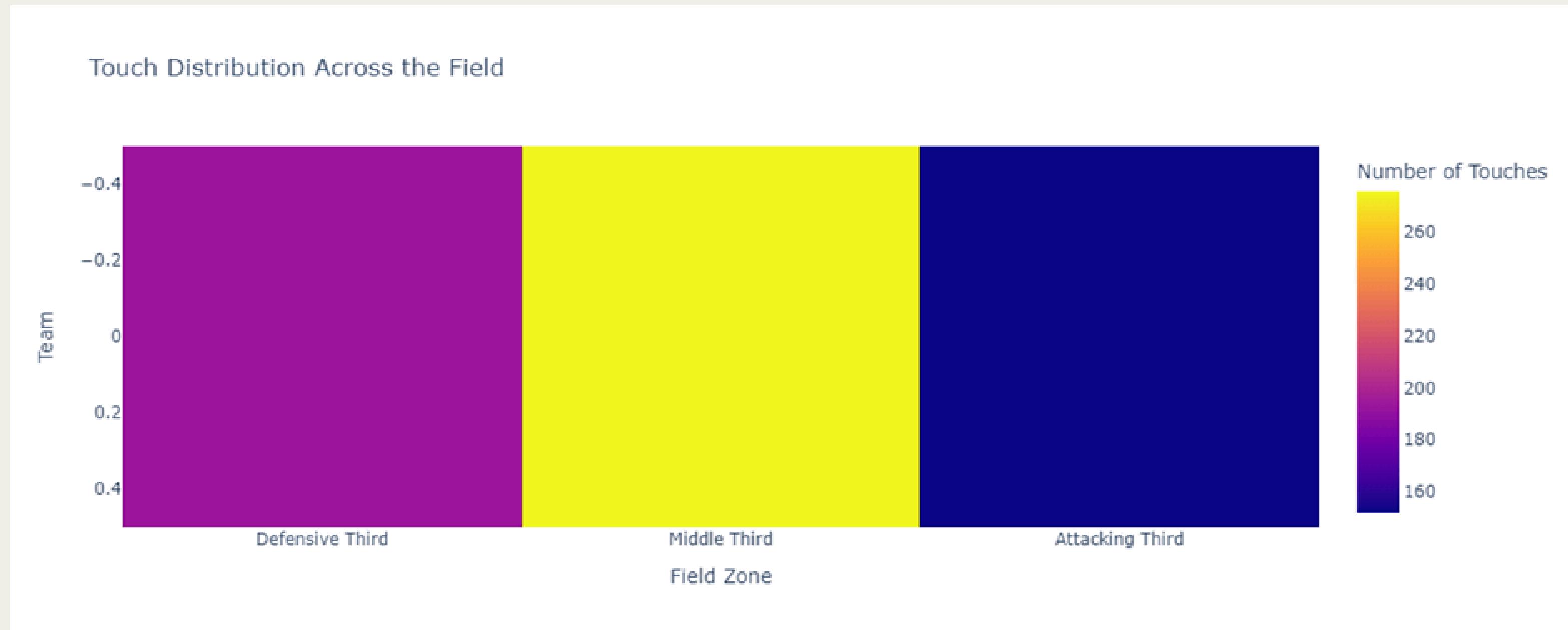
Từ biểu đồ, các cú sút thường tập trung ở khoảng cách 15-20 yards, khoảng cách có xu hướng giảm khi GF tăng

Radar chart: so sánh các đội ở một số thông số để rút ra phong cách chơi

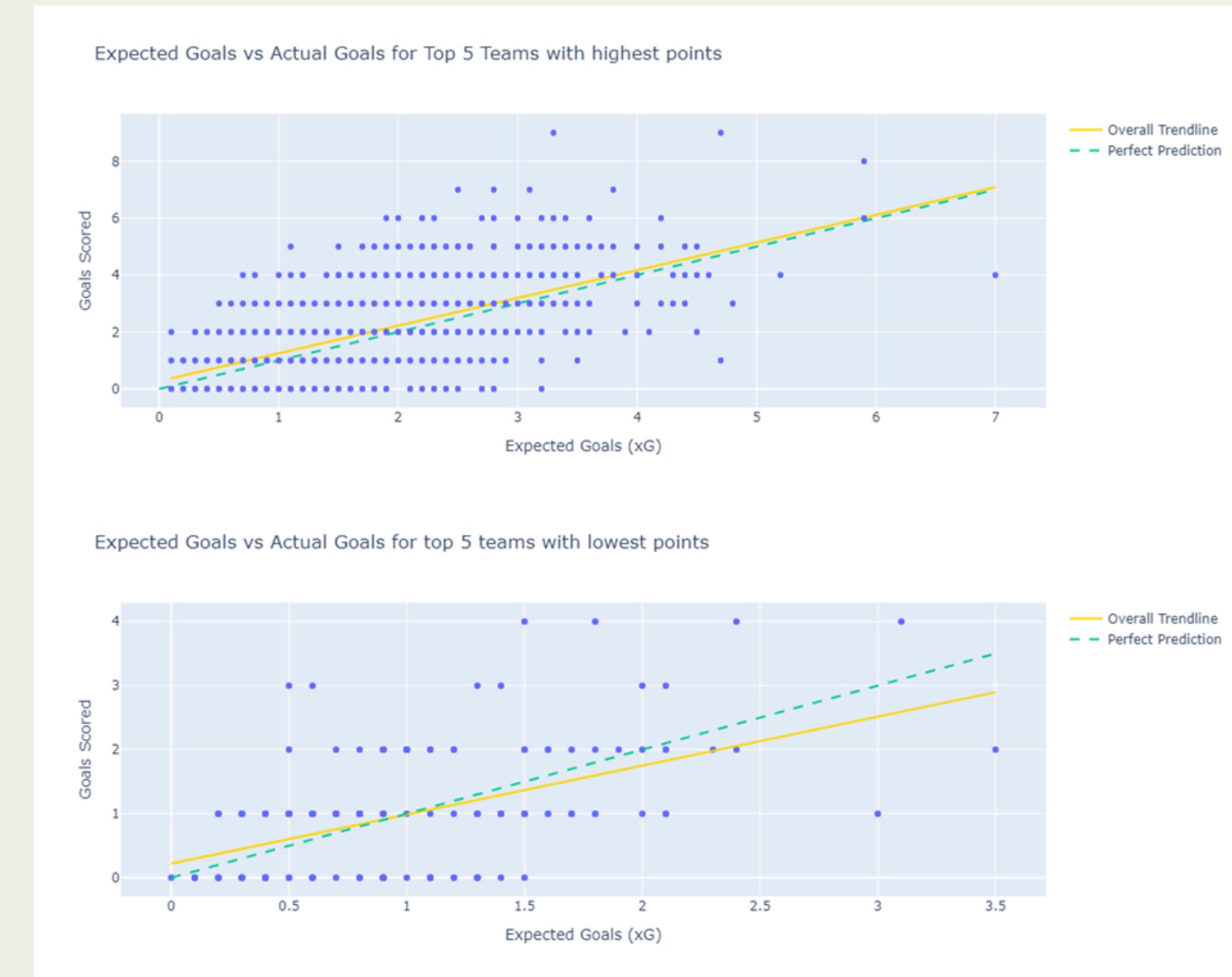


Ở các đội càng mạnh, các chỉ số tấn công và kiểm soát bóng (xG , Possession, Pass Completion Rate, Touches in Penalty Area, Passes into Penalty Area). Ngược lại ở các đội càng yếu, các chỉ số về phòng ngự (Tackles, Dribbles Challenged càng tăng) và xGA (số bàn thua kì vọng) cũng tăng.

Biểu đồ thể hiện phân bố số lần chạm bóng của các đội ở mỗi khu vực trên sân => bóng chủ yếu được luân chuyển ở khu vực giữa sân



Biểu đồ thể hiện mối quan hệ giữa bàn thắng kì vọng và bàn thắng thực tế ở 2 nhóm (top 5 team mạnh nhất và yếu nhất)



=> Các đội mạnh thường tận dụng cơ hội ghi bàn tốt hơn (đường trendline nằm trên đường perfect prediction)

2 biểu đồ thể hiện khả năng cầm bóng và qua người của 2 nhóm (top 5 team mạnh/yếu nhất)



Ở các đội mạnh, cầu thủ cầm bóng nhiều hơn và thực hiện qua người nhiều hơn. Từ đó có thể thấy các cầu thủ này có kỹ thuật tốt hơn

2. Tiền xử lý dữ liệu



Feature Encoding

Encode các category features như team, formation, captain, referee (đây cũng là các feature sẽ có trước khi trận đấu diễn ra)

Feature Aggregation

Tạo ‘form_<feature>’ cho các numeric features (các thông số chỉ có được sau khi trận đấu đã diễn ra) bằng cách tính trung bình các feature đó ở 4 trận gần nhất của đội.

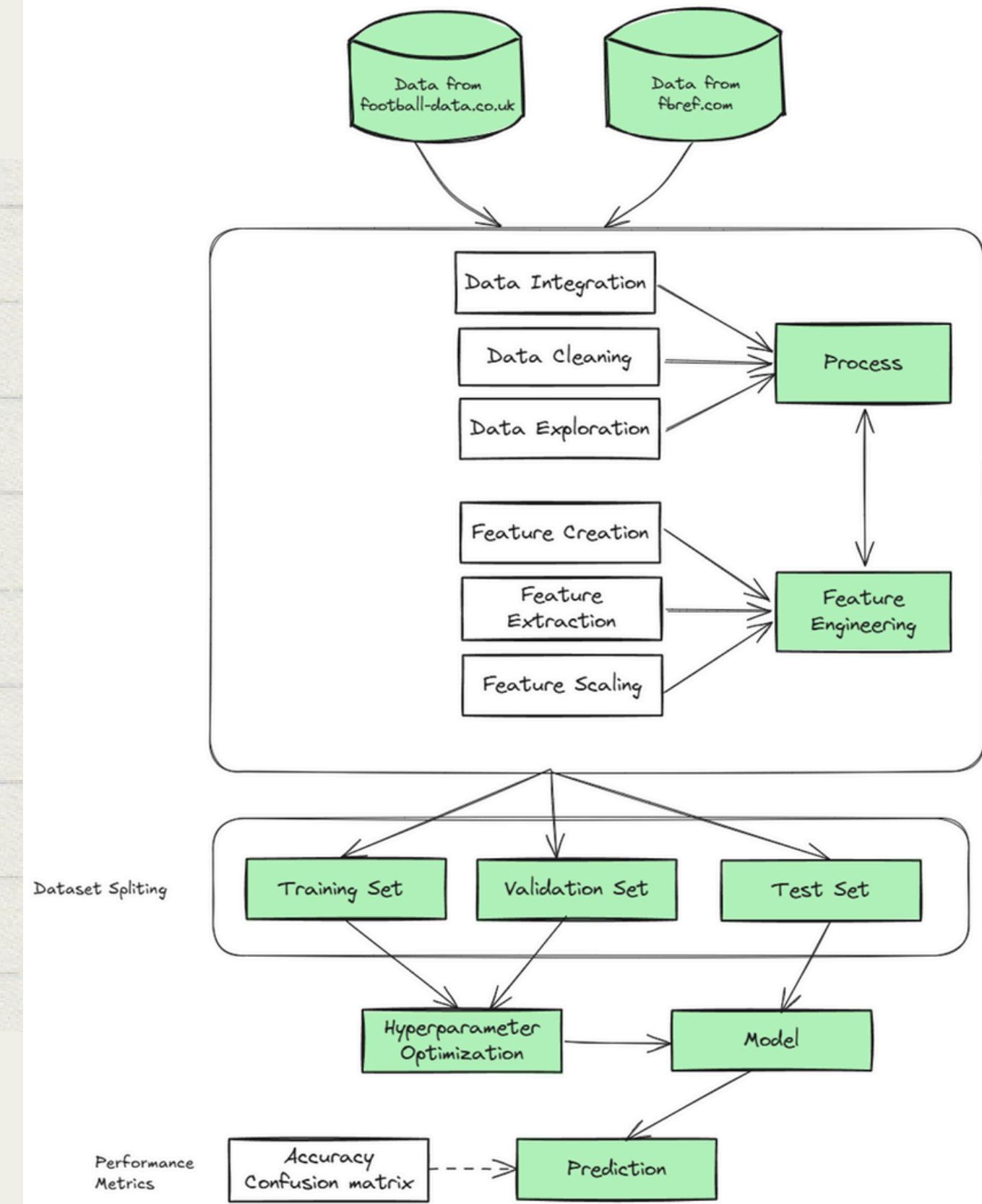
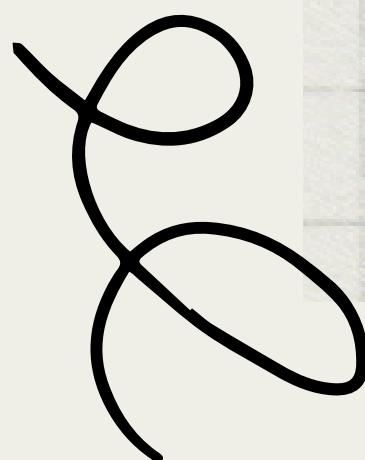
Dataset Splitting

Chia dữ liệu thành 3 tập:

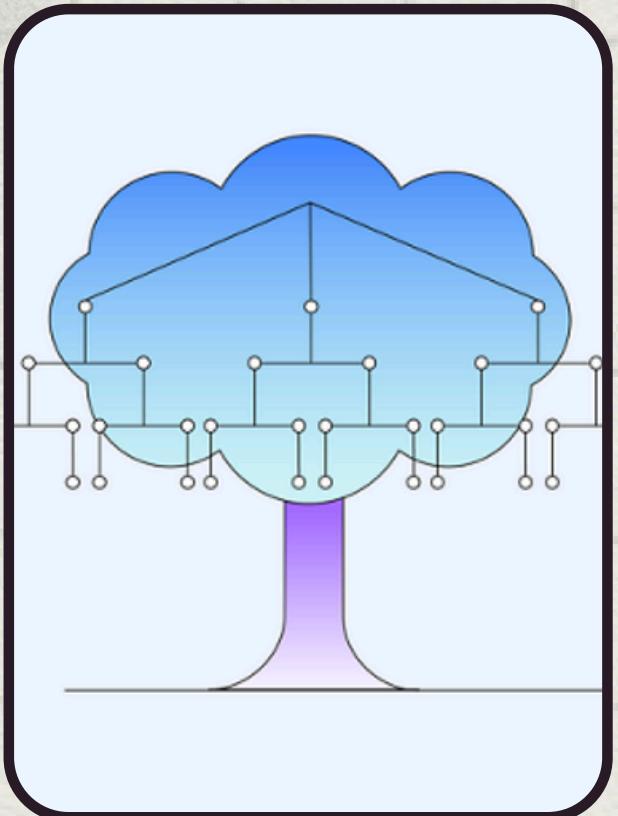
- train_set là các trận đấu từ đầu mùa 2017/2018 -> hết mùa 2020/2021
- valid_set là các trận đấu mùa 2021/2022
- test_set là các trận đấu từ mùa 2022/2023 đến hiện nay

3. Mô hình dự đoán

Nhóm chúng em sử dụng
Random Forest để thử
nghiệm và kiểm tra kết quả



RANDOM FOREST



Feature Selection

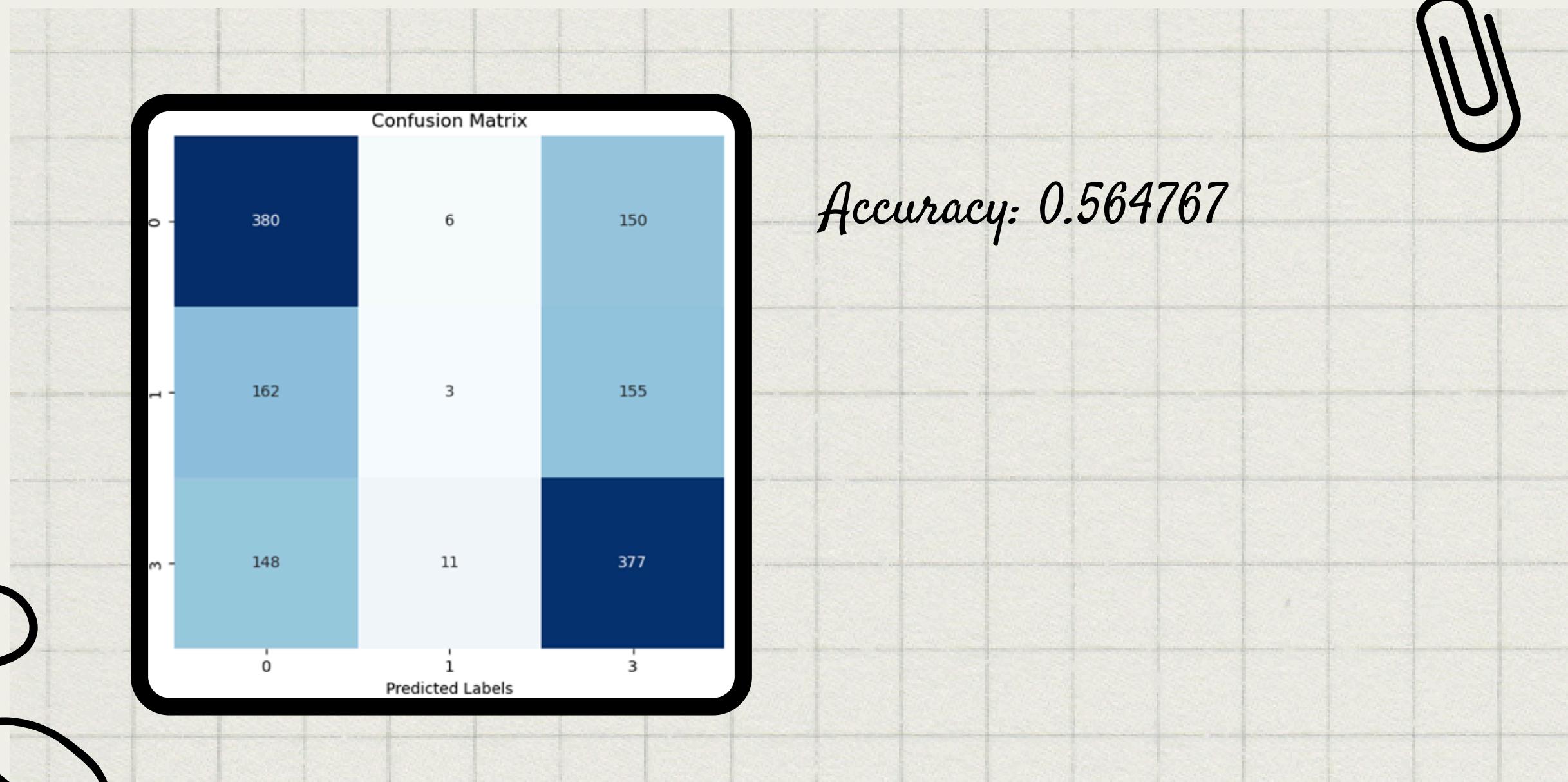
Kết hợp các categorical features (team, formation, captain, referee) và các form_features (trung bình của 4 trận gần nhất).



Hyperparameter

- Number of Trees (n_estimators): Tối ưu số lượng cây trong rừng.
- Maximum Tree Depth (max_depth): Điều chỉnh độ sâu tối đa của mỗi cây.

KẾT QUẢ



CẢI THIỆN

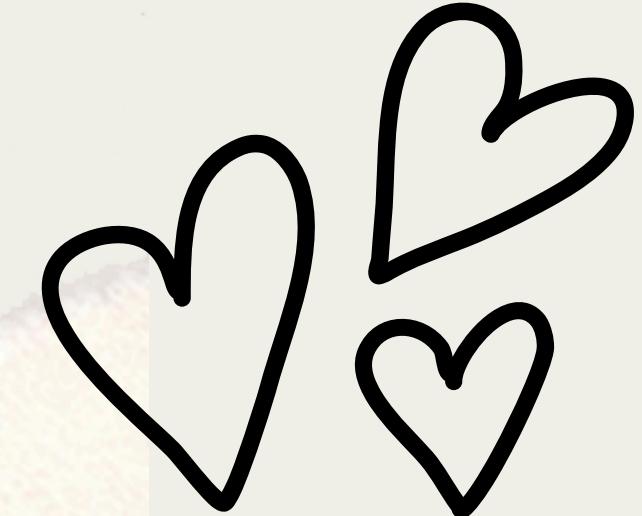


Model Optimization

Cải tiến mô hình bằng cách tối ưu các siêu tham số (hyperparameter tuning), áp dụng các kỹ thuật như Bootstrap Sampling và Boosting để tăng hiệu suất. Đồng thời, nhóm cũng sẽ thử nghiệm với các mô hình khác như RNN và GRU nhằm đánh giá khả năng dự đoán trong các phương pháp tiếp cận khác nhau.

Data Enhancement

Để cải thiện chất lượng mô hình, chúng em dự định thu thập thêm dữ liệu từ các nguồn khác, ví dụ như thông tin phong độ cá nhân của cầu thủ. Ngoài ra, dữ liệu mới sẽ được phân tích kỹ lưỡng, bao gồm kiểm tra sự tương quan giữa các đặc trưng để tối ưu hóa đầu vào.



KẾ HOẠCH THỰC HIỆN

