Introduction to Data Science Final Project

Nhóm 5:

Đặng Viết Khôi Nguyên – 1753022

Lê Minh Tiến - 18127229

Giới thiệu đề tài

- Có thể nói, sức hấp dẫn của bóng đá thu hút sự quan tâm của phần lớn dân số trên thế giới. Cùng sự đam mê và thích thú với môn thể thao vua này, nhóm cố gắng tìm hiểu và khám phá bằng góc nhìn của khoa học dữ liệu

Thu thập dữ liệu

- Ban đầu, để phù hợp với dự định, nhóm đã cố gắng tìm kiếm dữ liệu thật từ nhiều nguồn khác nhau. Kết quả, nhóm đã thu thập được danh sách các cầu thủ và danh sách các giải đấu từ nguồn **The Sport DB**.
- Tuy nhiên, nhận thấy các thông tin thu thập được từ nguồn trên không có nhiều tiềm năng, nhóm đã thu thập dữ liệu ảo từ nguồn game FIFA19 với các chỉ số cho các cầu thủ trong game

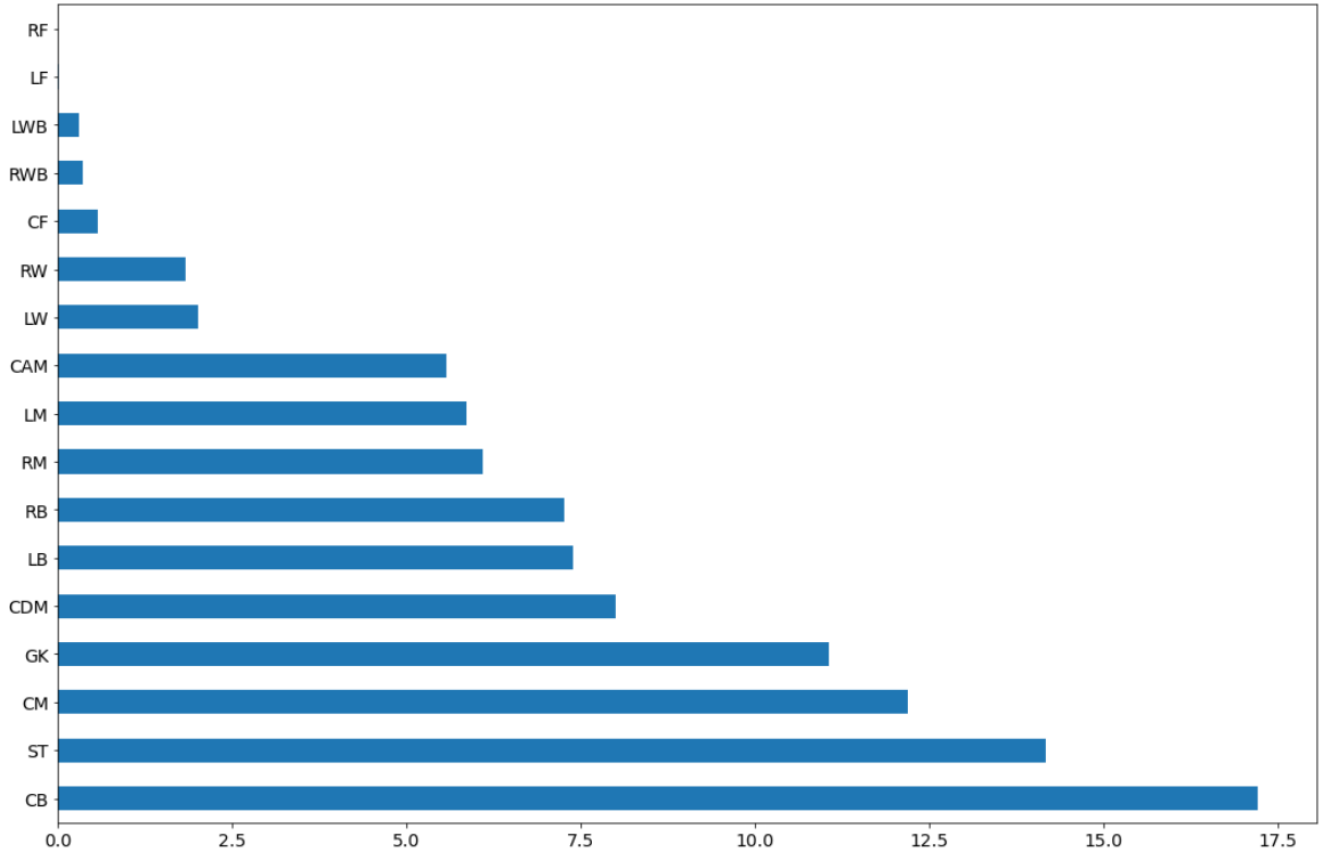
Khám phá dữ liệu

- Bộ dữ liệu gồm 15525 records, với mỗi records gồm 56 thuộc tính. Không có dữ liệu trống
- Các thuộc tính gồm tên, vị trí và các chỉ số thể hiện khả năng của cầu thủ

0	position	21	gkpositioning	41	volleys
1	composure	22	gkreflexes	42	weakFoot
2	height	23	headingaccuracy	43	traits
3	weight	24	interceptions	44	specialities
4	birthdate	25	jumping	45	atkWorkRate
5	age	26	longpassing	46	defWorkRate
6	acceleration	27	longshots	47	attributes
7	aggression	28	marking	48	name
8	agility	29	penalties	49	rarityId
9	balance	30	positioning	50	isIcon
10	ballcontrol	31	potential	51	quality
11	foot	32	reactions	52	isGK
12	skillMoves	33	shortpassing	53	positionFull
13	crossing	34	shotpower	54	id
14	curve	35	slidingtackle	55	baseId
15	dribbling	36	sprintspeed	56	rating
16	finishing	37	standingtackle		
17	freekickaccuracy	38	stamina		
18	gkdiving	39	strength		
19	gkhandling	40	vision		
20	gkkicking				

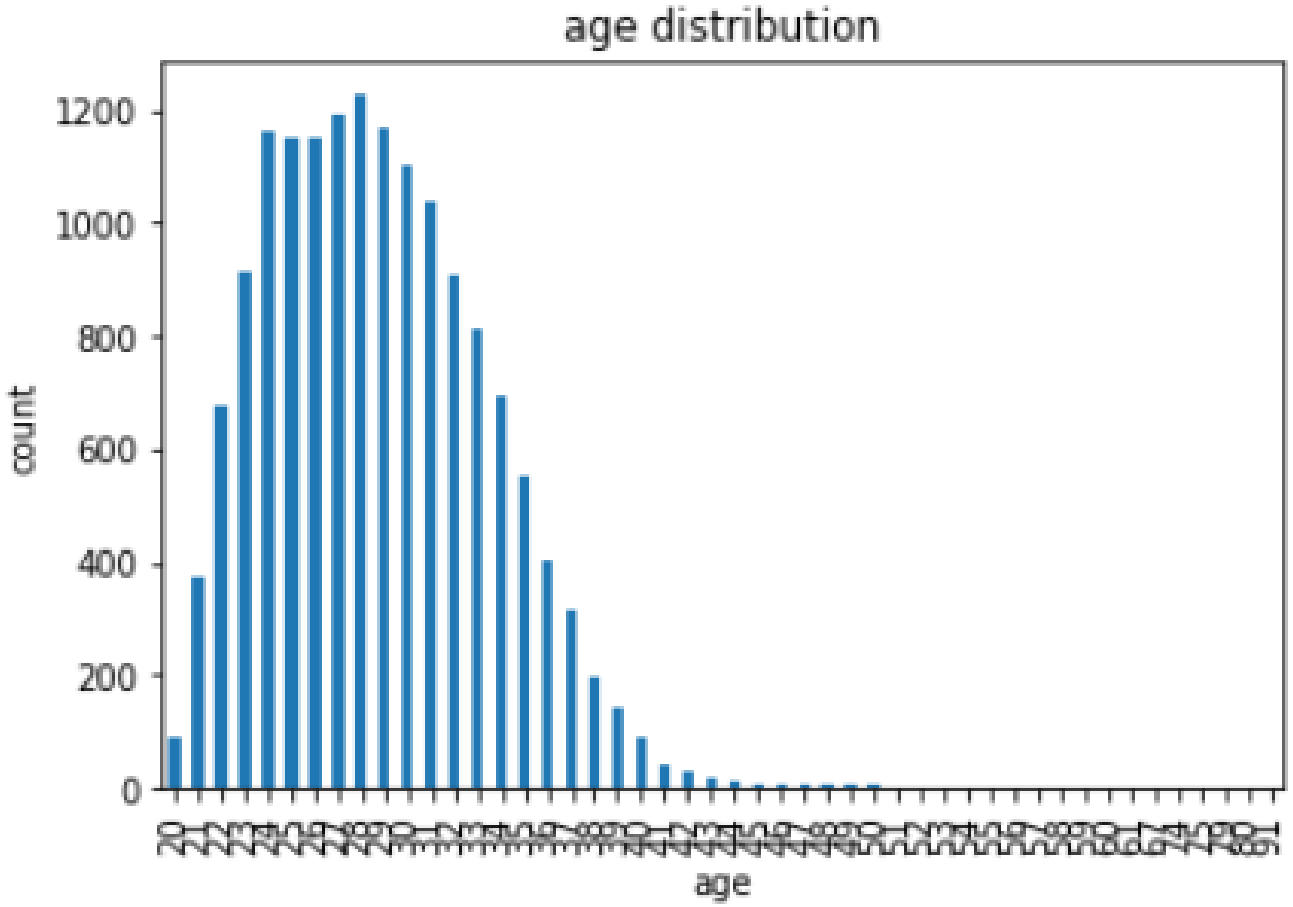
Khám phá dữ liệu

- Nhận thấy 4 vị trí ít phổ biến nhất là: RF, LF, LWB, RWB có số điểm dữ liệu ít hơn hẳn so với các vị trí khác.
- Nhóm nhận thấy các vị trí này đã không còn phổ biến trong đa số các chiến thuật ngày nay nên đã gộp các vị trí đấy vào các vị trí tương tự khác thay thế như sau:
  - RF -> RW
  - LF -> LW
  - LWB -> LB
  - RWB -> RB



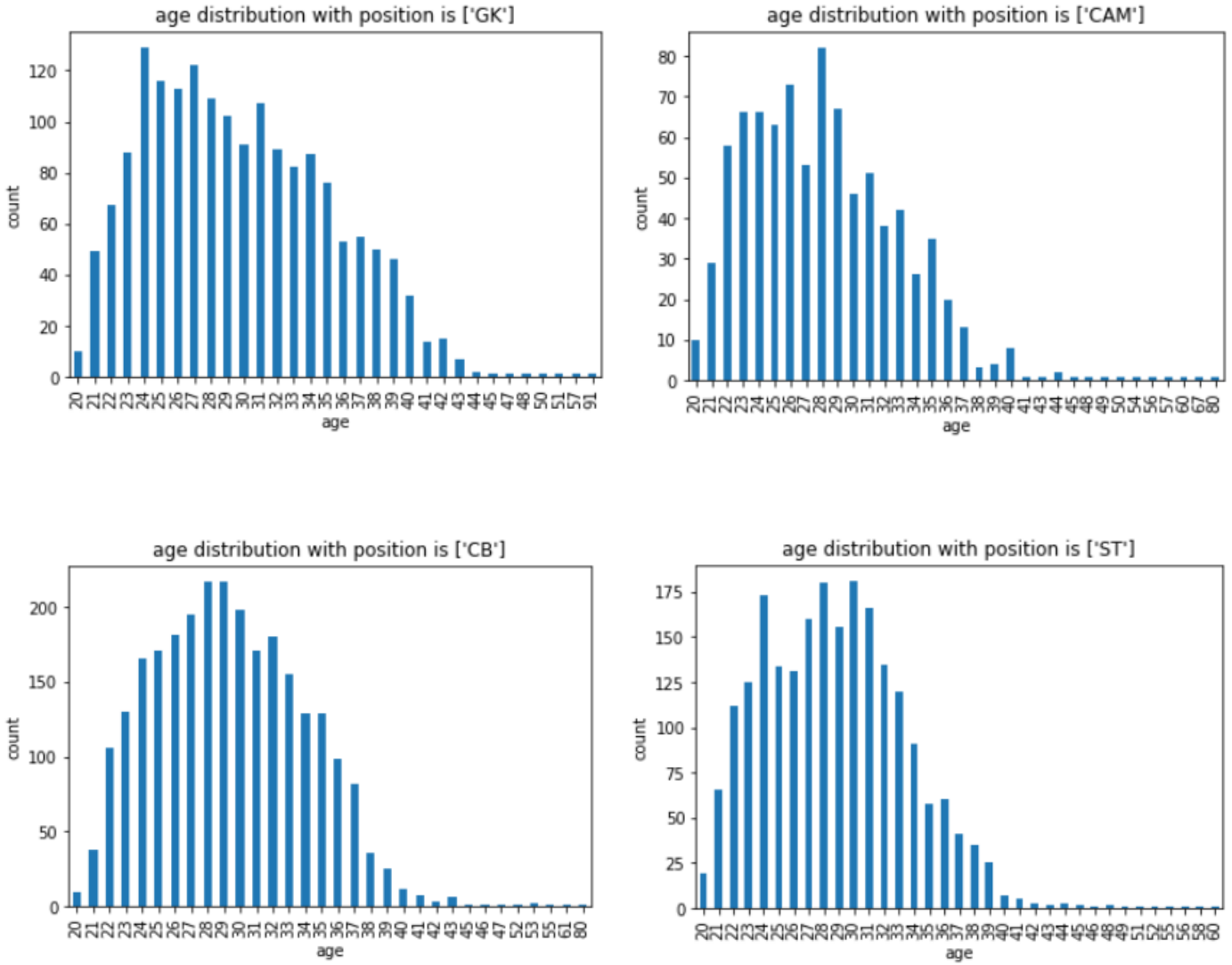
Trực quan hoá dữ liệu

- Đây là phân phối số lượng toàn bộ cầu thủ theo tuổi
- Phân phối số lượng cầu thủ theo tuổi tăng nhanh ở các tuổi từ 20 và đạt đỉnh tăng ở tầm 29, 30. Sau đó, số lượng cầu thủ giảm dần với tốc độ chậm hơn so với lúc tang.



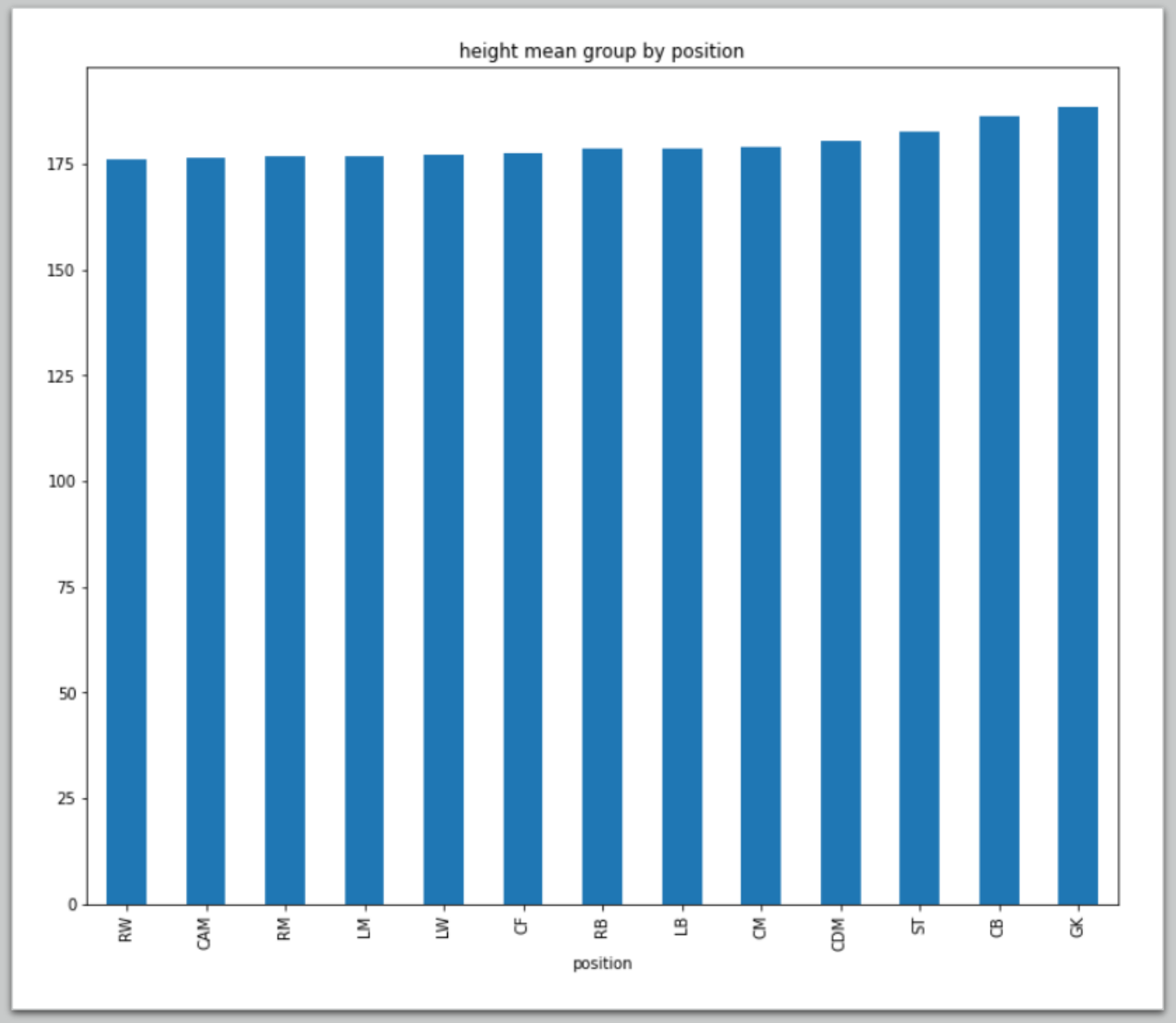
Trực quan hoá dữ liệu

- Và đa số các vị trị đều có phân phối tương tự như vậy (tăng nhanh đạt đỉnh ở tầm dưới 30 và giảm chậm dần vào các năm sau đó)
- Tuy nhiên, vị trí GK có phân phối tăng nhanh hơn và đạt đỉnh sớm hơn (tầm 24 tuổi) và giảm chậm hơn hẳn so với các vị trí khác.



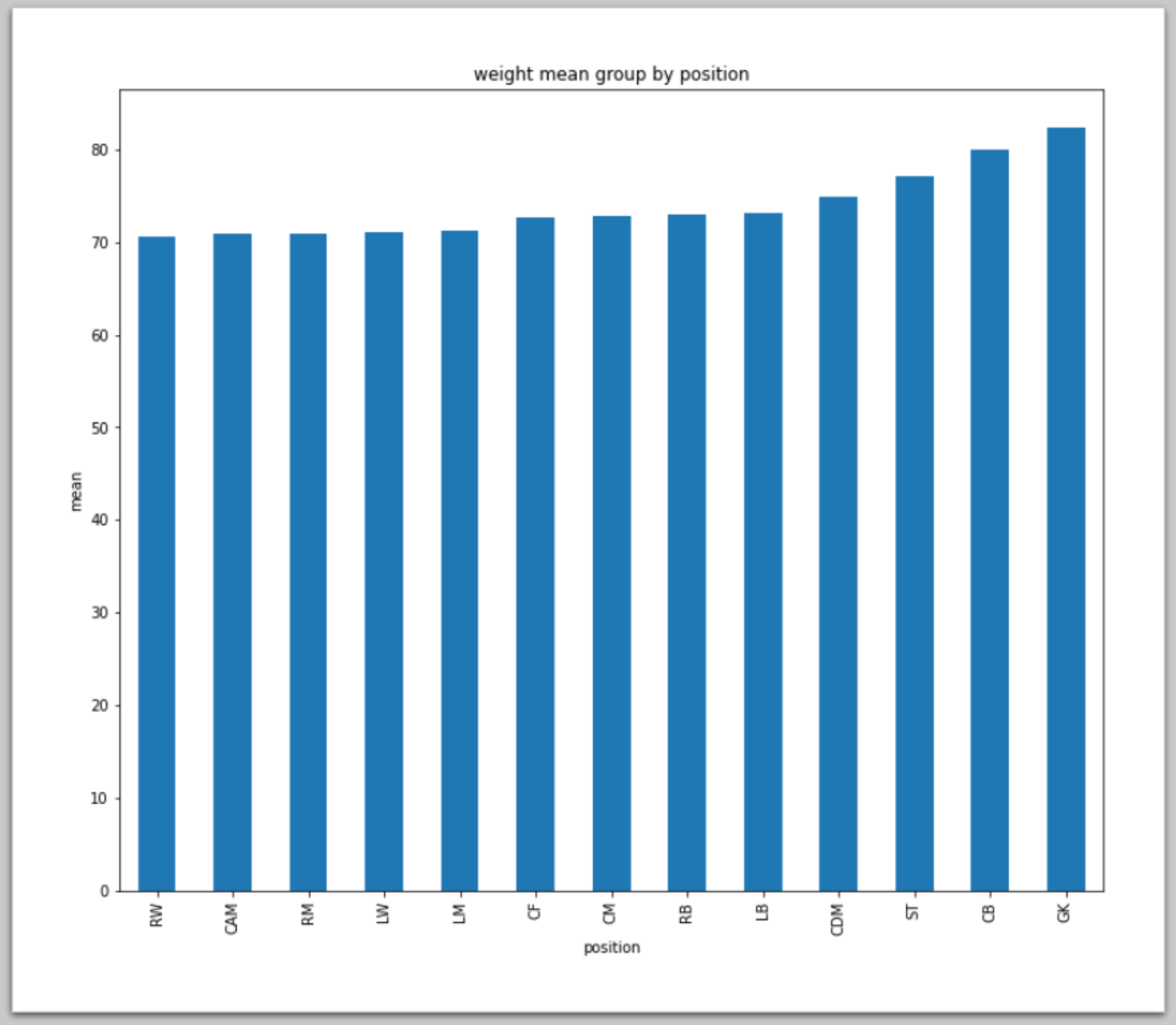
Trực quan hoá dữ liệu

- Bên đây là chiều cao trung bình của các cầu thủ theo từng vị trí



Trực quan hoá dữ liệu

- Bên đây là cân nặng trung bình của các cầu thủ theo từng vị trí

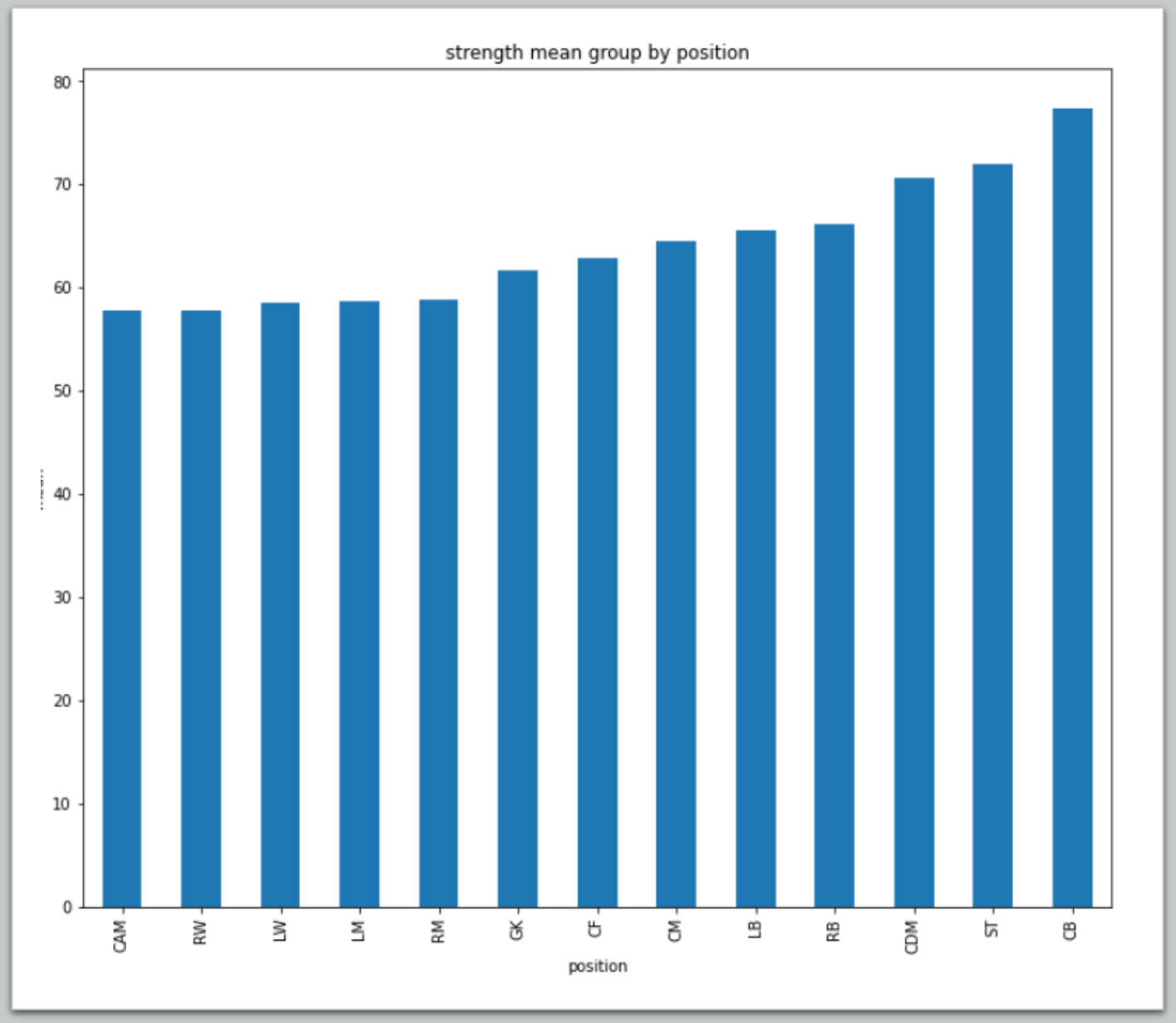


Trực quan hoá dữ liệu

- Dễ dàng thấy được, chiều cao và cân nặng trung bình của các cầu thủ ở vị trí trung tâm hoặc 2 cánh thấp hơn so với các cầu thủ đá ở trung tâm, và nhất là các vị trí gần với khung thành (ST, CB, GK). ST, CB và nhất là GK là 3 vị trí đòi hỏi khả năng tranh chấp bóng bổng, do đó, chiều cao và cân nặng của các cầu thử ở 3 vị trí đó tốt hơn các vị trí khác.
- Ngoài ra, vị trí CDM cũng là vị trí đòi hỏi tranh chấp khu vực giữa sân nên 2 chỉ số này cũng có phần cao hơn các vị trí khác.
- Ngược lại, các khu vực cho các vị trí xa khung thành, không yêu cầu tranh chấp bóng bổng nhiều như các cầu thủ chơi cánh hoặc tiền vệ công thì lại có 2 chỉ số này không tốt bằng các vị trí kể trên.

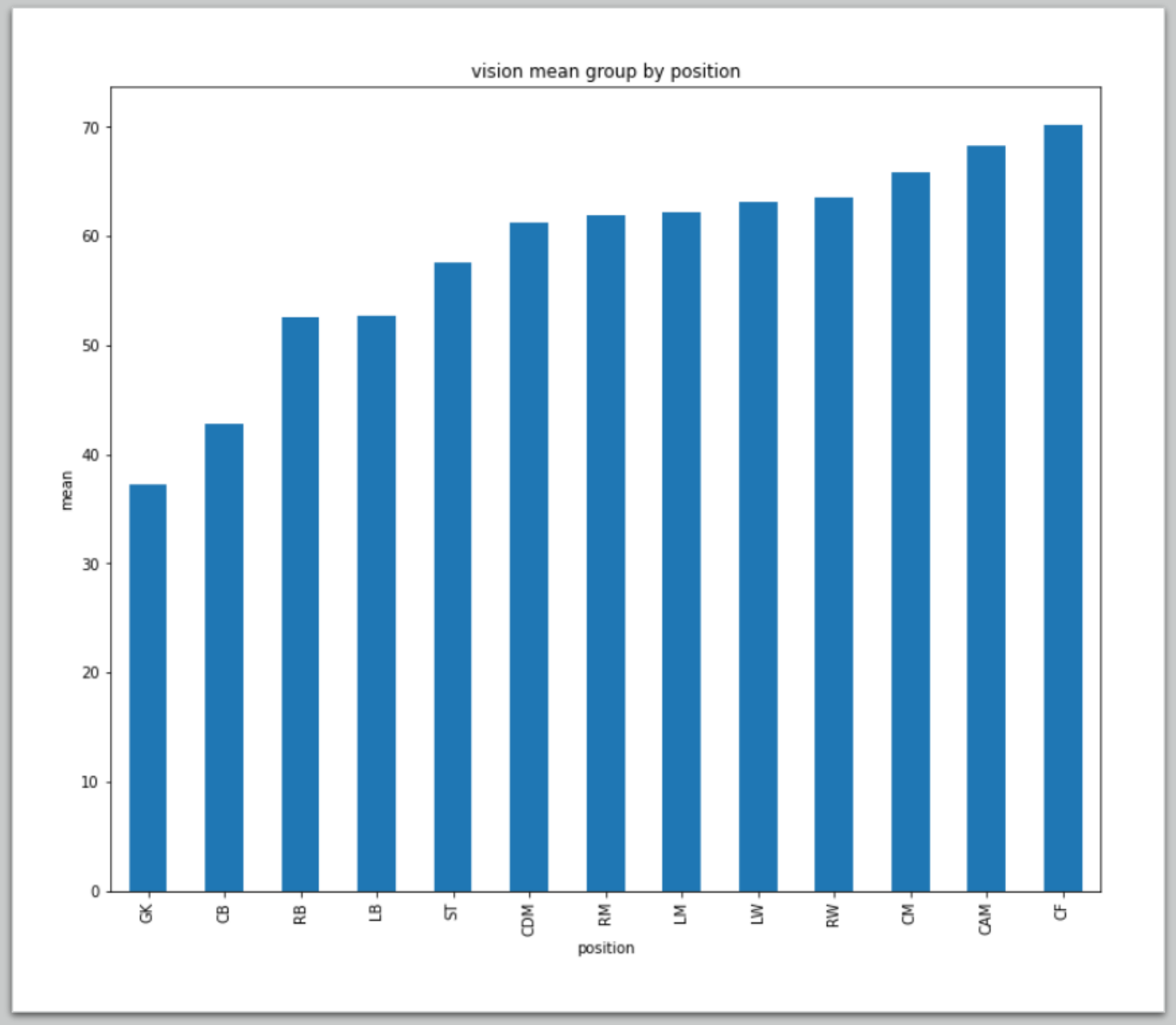
Trực quan hoá dữ liệu

- Bên đây là chỉ số sức lực trung bình của các cầu thủ theo từng vị trí
- Rõ ràng, 3 vị trí thường xuyên phải tranh chấp về sức mạnh như các vị trí hậu vệ (LB, RB, CB) cùng với ST và CDM là các vị trí có chỉ số strength tốt. Ngược lại, các tiền đạo cánh, các tiền vệ và thủ môn có chỉ số strength không ấn tượng.



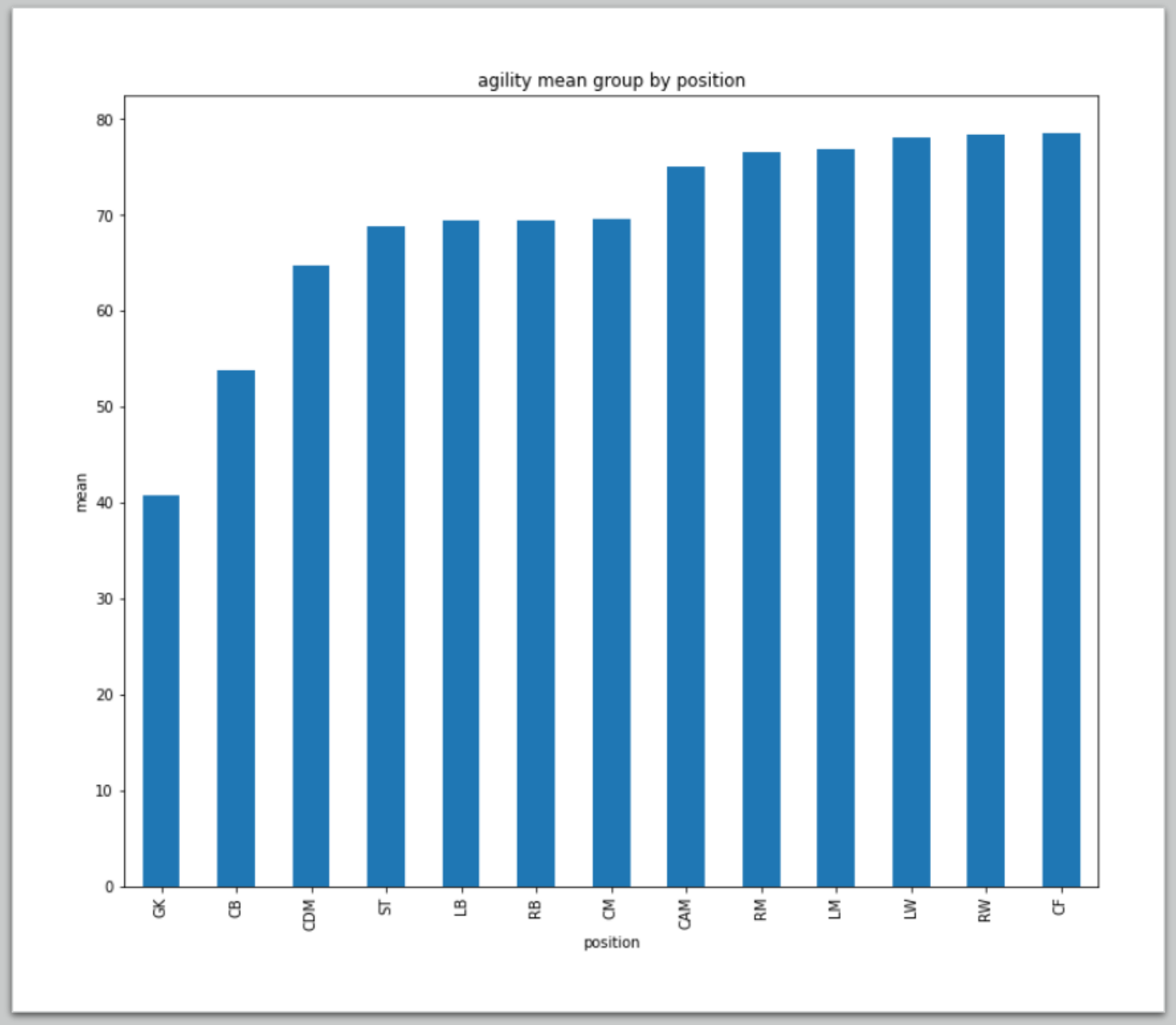
Trực quan hoá dữ liệu

- Bên đây là chỉ số tầm nhìn trung bình của các cầu thủ theo từng vị trí



Trực quan hoá dữ liệu

- Bên đây là chỉ số nhanh nhẹn trung bình của các cầu thủ theo từng vị trí

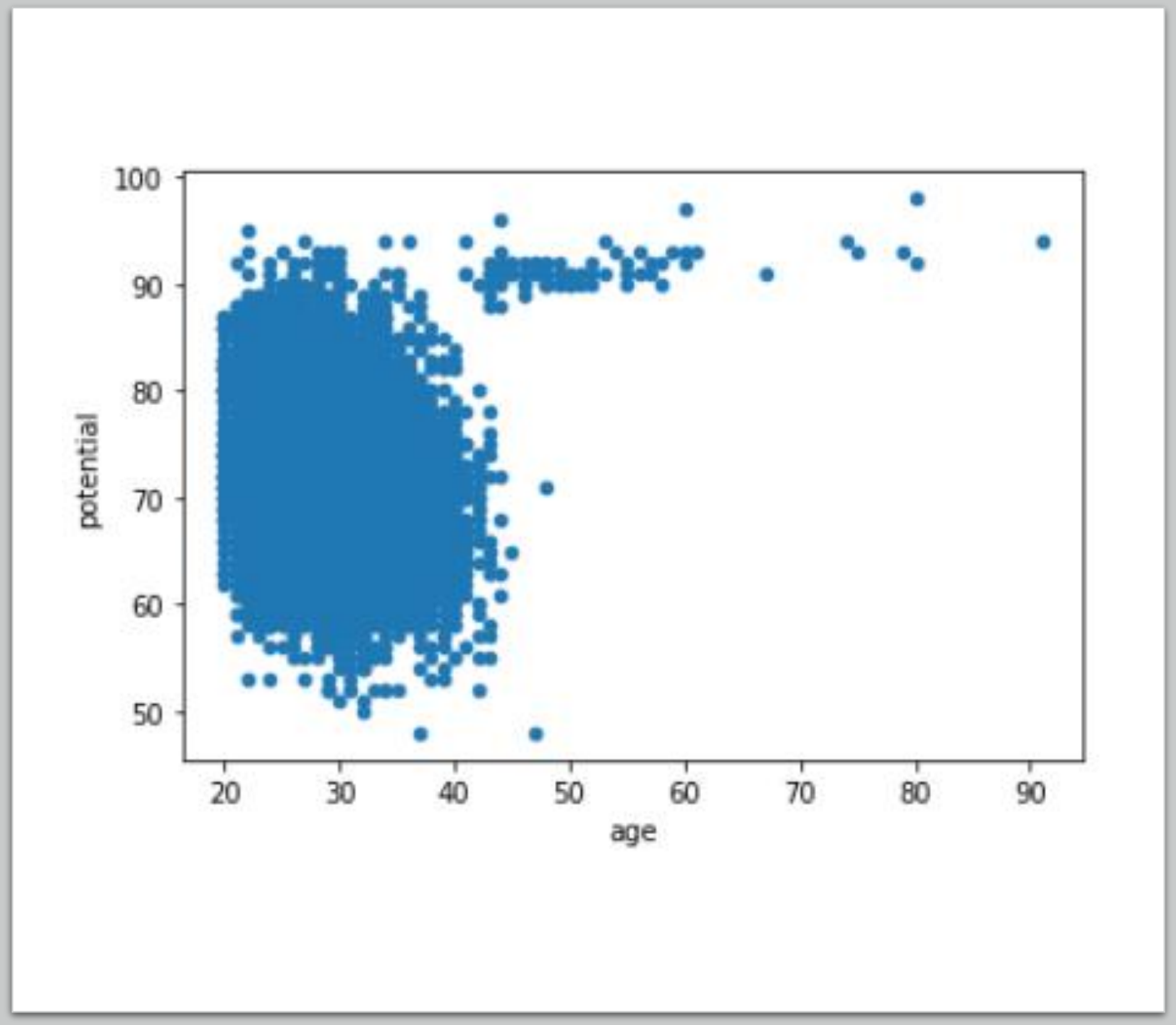


Trực quan hoá dữ liệu

- Thua thiệt về chiều cao, cân nặng và sức mạnh nhưng các tiền đạo cánh, tiền vệ và các hậu về cánh lại có chỉ số tầm nhìn và nhanh nhẹn tốt hơn.
- Đặc biệt, vị trí CF là vị trí có 2 chỉ số này tốt nhất. Còn lại, các vị trí cánh thì nhanh nhẹn hơn các vị trí trung tâm và ngược lại, các cầu thủ trung tâm có tầm nhìn tốt hơn các đồng đội ở cánh.

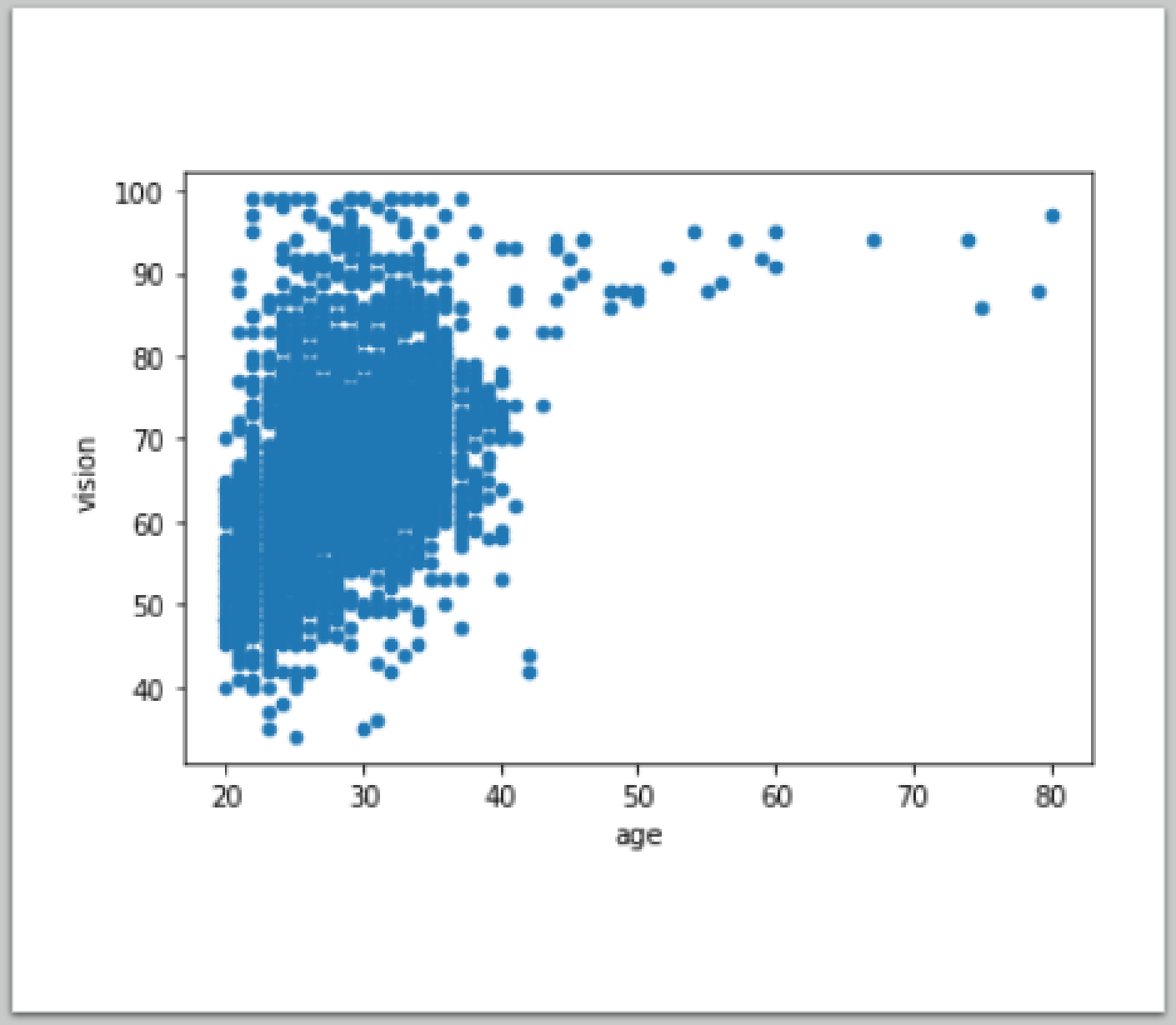
Trực quan hoá dữ liệu

- Bên đây là đồ thị thể hiện sự liên quan giữa tiềm năng của cầu thủ với độ tuổi.
- Dễ hiểu và dễ thấy, chỉ số tiềm năng tỉ lệ nghịch với số tuổi



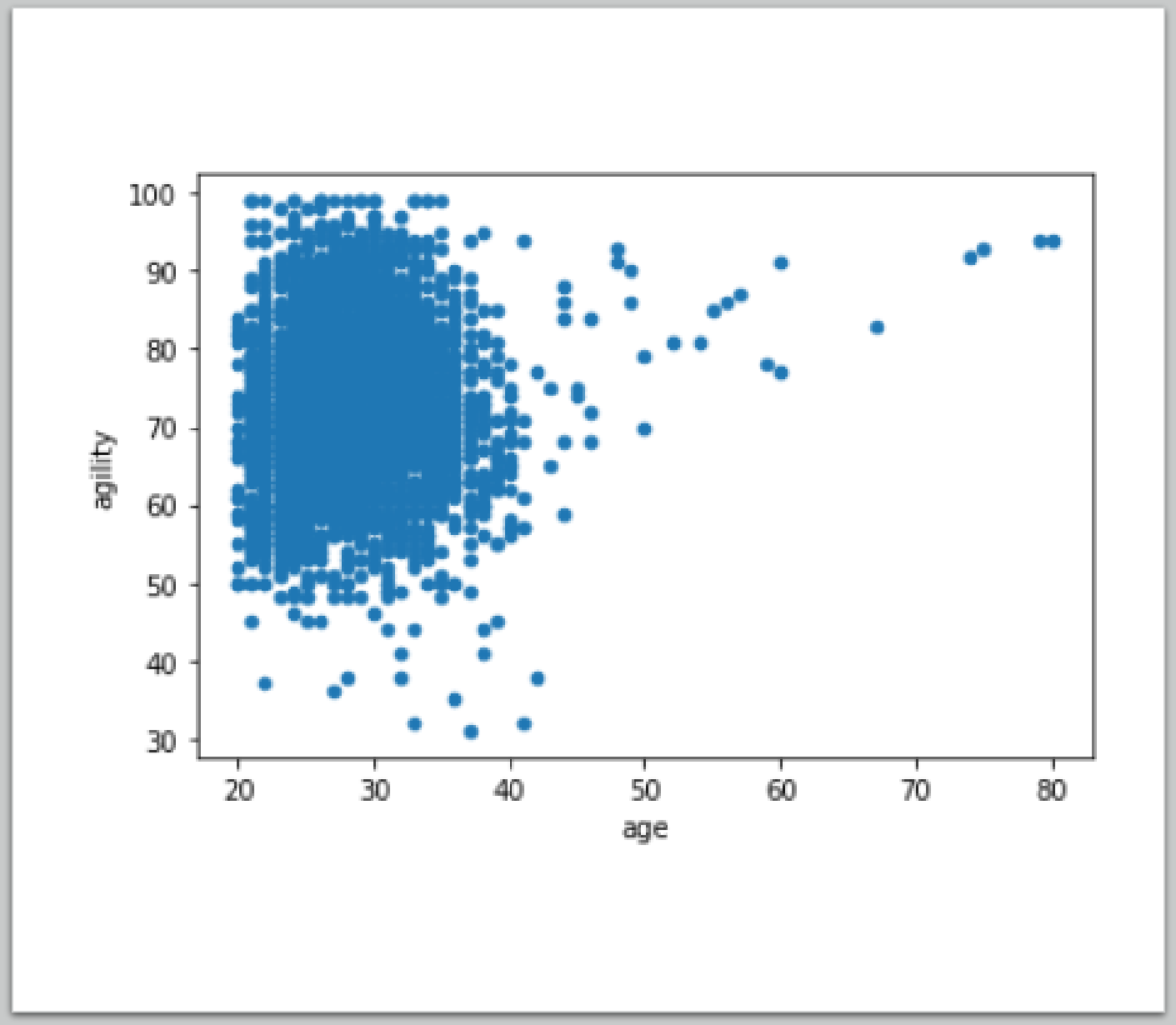
Trực quan hoá dữ liệu

- Bên đây là đồ thị thể hiện sự liên quan giữa tầm nhìn của cầu thủ tấn công (CF, CAM, CM, RW, LW, CM) với độ tuổi.
- Các vị trí tấn công này là các vị trí yêu cầu cao về tầm nhìn. Tại các vị trí này, những cầu thủ cao tuổi thường tốt hơn.



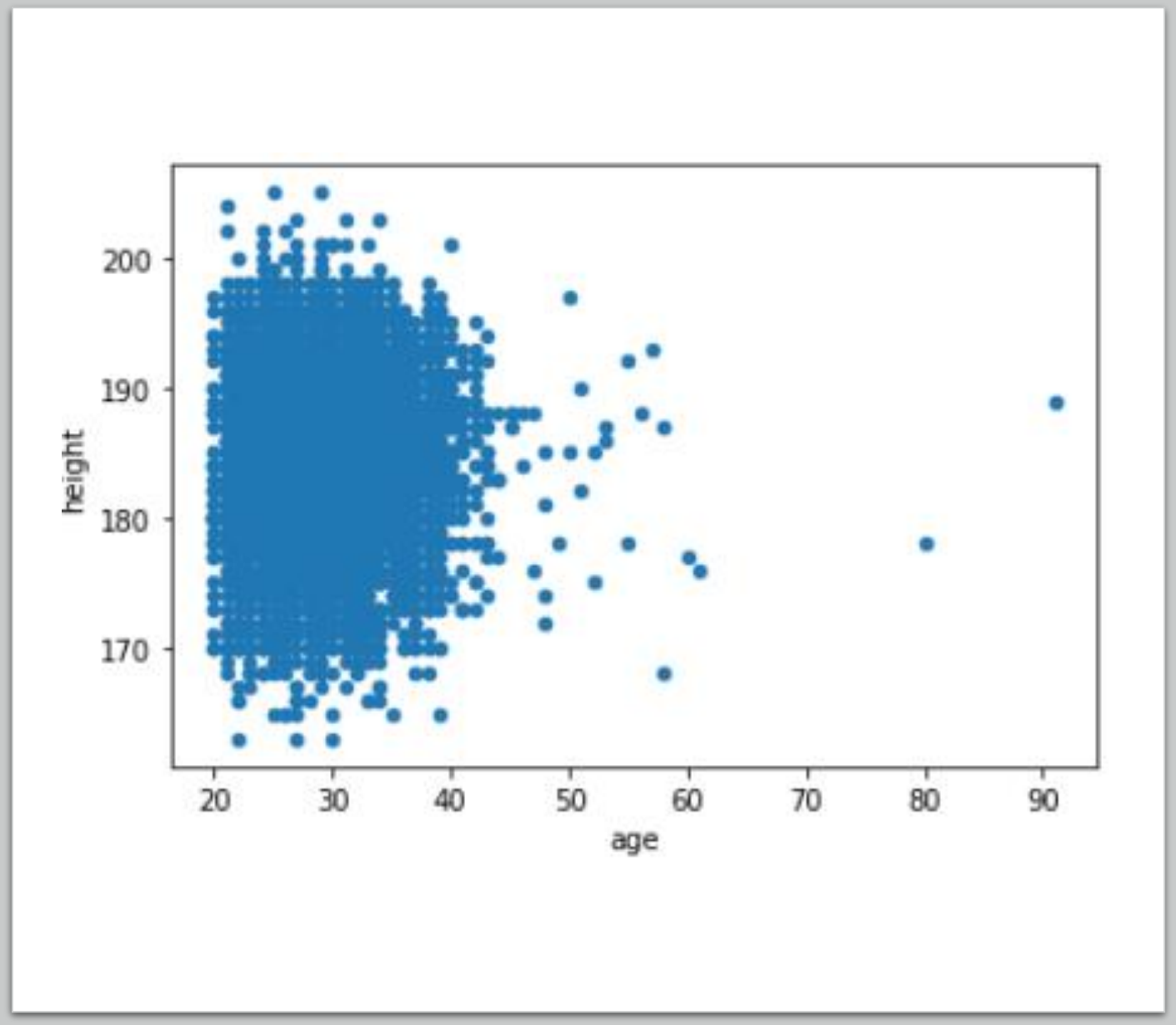
Trực quan hoá dữ liệu

- Bên đây là đồ thị thể hiện sự liên quan giữa tầm nhìn của cầu thủ tấn công (CF, CAM, CM, RW, LW, CM) với độ tuổi.
- Các vị trí tấn công cũng yêu cầu cao cho sự nhanh nhẹn. Tuy nhiên, sự nhanh nhẹn không phụ thuộc nhiều lắm vào tuổi tác.



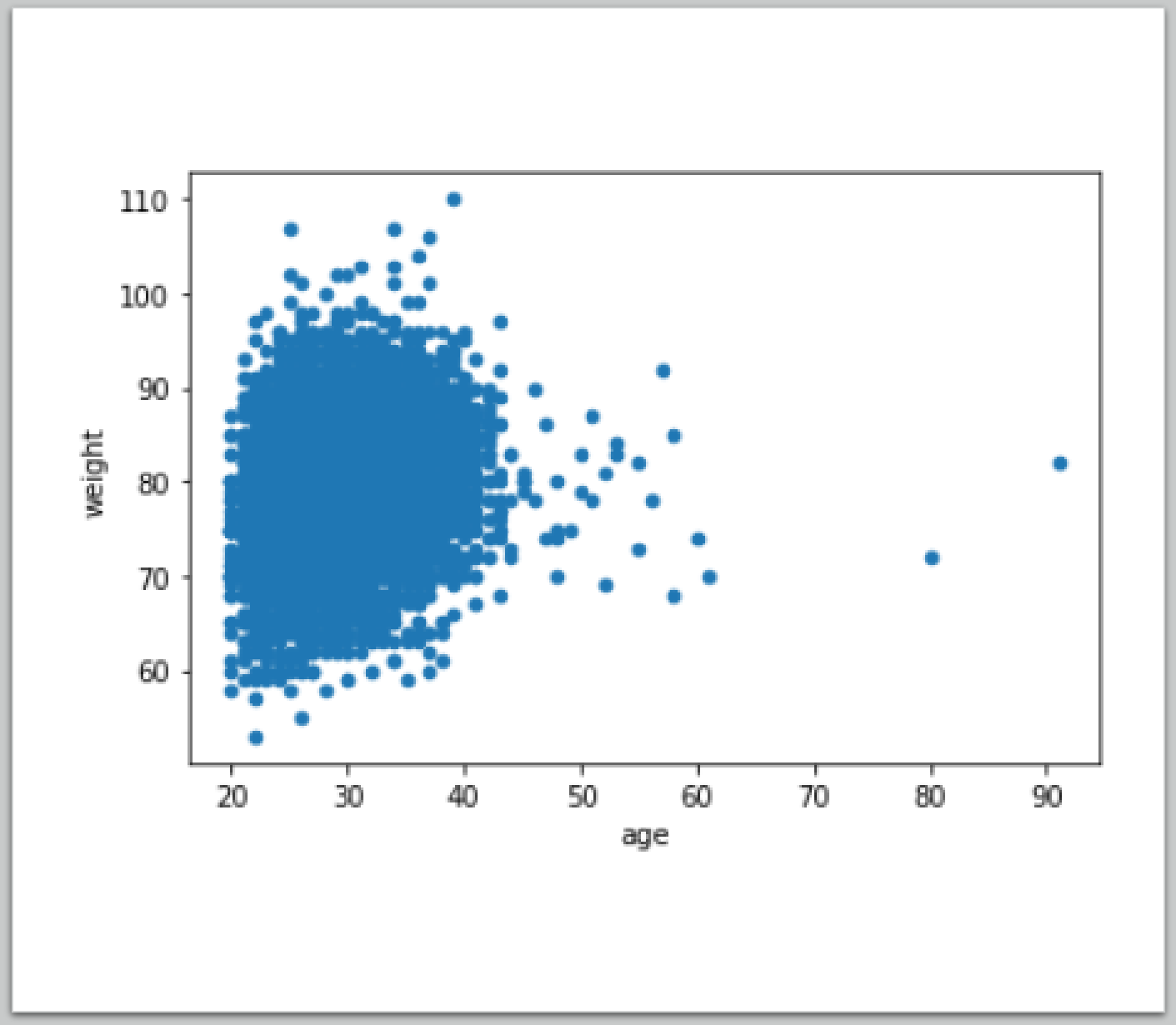
Trực quan hoá dữ liệu

- Bên đây là đồ thị thể hiện sự liên quan giữa chiều cao của các cầu thủ ST, CDM, CB, GK với độ tuổi.
- Đối với người trưởng thành, chiều cao là chỉ số khó có sự thay đổi rõ rệt theo thời gian



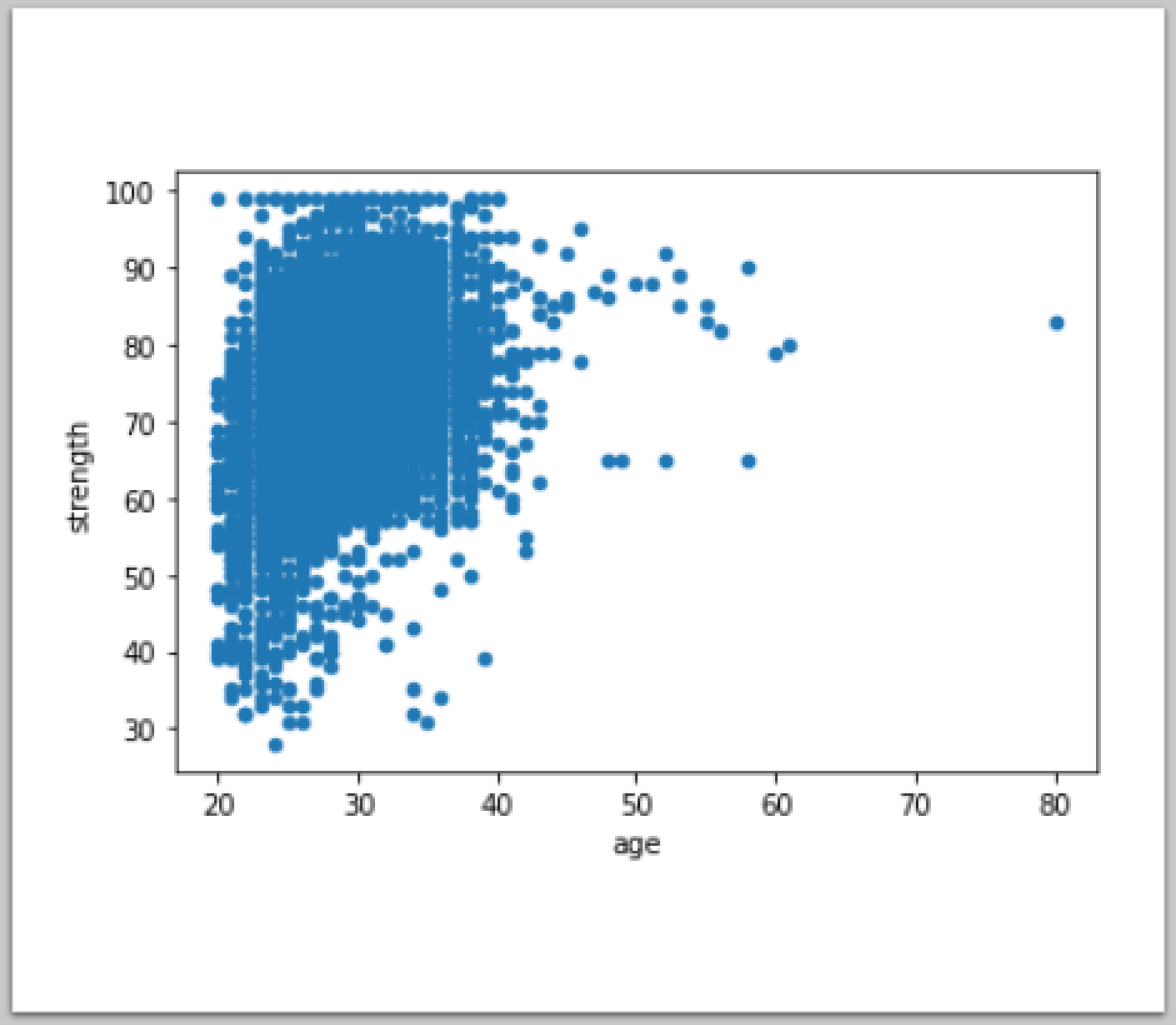
Trực quan hoá dữ liệu

- Bên đây là đồ thị thể hiện sự liên quan giữa cân nặng của các cầu thủ ST, CDM, CB, GK với độ tuổi.
- Theo đó, ta thấy, chỉ số cân nặng có xu hướng tỉ lệ thuận so với tuổi tác



Trực quan hoá dữ liệu

- Bên đây là đồ thị thể hiện sự liên quan giữa sức lực của các cầu thủ ST, CDM, CB với độ tuổi.
- Tương tự với cân nặng, chỉ số này cũng tỉ lệ thuận với số tuổi
- Từ đó, ta có thể hiểu được, vì là thi đấu ở các vị trí thường xuyên tranh chấp, 3 vị trí này phải tăng cường sức khỏe và cân nặng để có được lợi thế trong các tình huống tì đè, đối mặt.



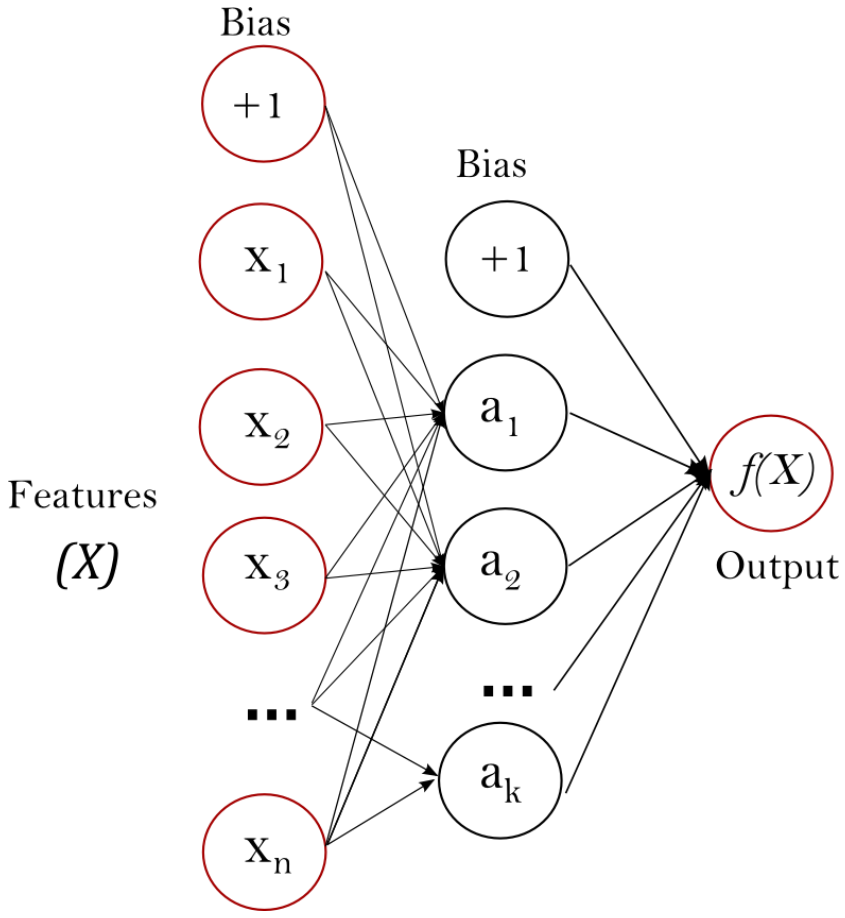
Mô hình hóa dữ liệu

Bài toán phân loại vị trí của cầu thủ: Với nguồn dữ liệu đáp ứng được việc mô tả đầy đủ các chỉ số của cầu thủ, việc phân loại vị trí của các cầu thủ là một nhiệm vụ vô cùng thú vị khi tìm được mối quan hệ giữa các chỉ số và vị trí của cầu thủ, góp phần tạo nên cái nhìn toàn diện hơn về sức ảnh hưởng của các chỉ số đến vị trí tiềm năng của một cầu thủ là như thế nào.

Mô hình hóa dữ liệu

- Các mô hình được sử dụng: Multi-layers Perceptron Classifier

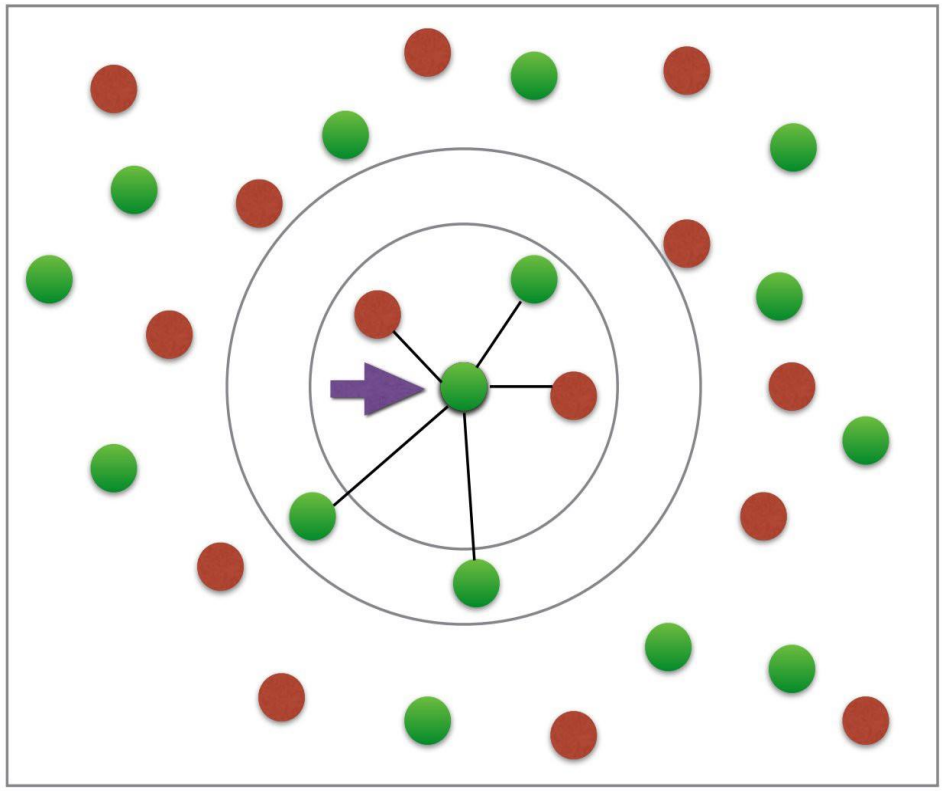
Mô hình MLPClassifier là một bộ phân lớp sử dụng mạng neuron nhân tạo để thực hiện việc các nhiệm vụ phân loại. Với việc sử dụng mạng neuron nhân tạo, mô hình MLPClassifier mang tính nhạy bén, linh hoạt cao và dễ dàng tìm kiếm mối quan hệ giữa đầu vào và đầu ra của mô hình



Mô hình hóa dữ liệu

- Các mô hình được sử dụng: k-Nearest Neighbors

K-Nearest Neighbors là một thuật toán học có giám sát trong lĩnh vực Học Máy và là một trong những mô hình mạnh mẽ trong nhiệm vụ phân loại. Thay vì tìm mối quan hệ giữa đầu vào và đầu ra, thuật toán kNN sẽ xác định đầu ra dựa vào tính tương đồng về thuộc tính của các đối tượng trong tập dữ liệu.



Lời cuối

Đồ án lần này đã cho chúng em khá nhiều niềm vui trong việc khám phá dữ liệu, đặc biệt là trong lĩnh vực yêu thích của cả hai thành viên là game và bóng đá. Không những thế, việc thực hiện đồ án giúp chúng em có thêm nhiều kiến thức mới trong lĩnh vực Data Science. Mặc dù đã cố gắng thực hiện đồ án, song chúng em cũng không thể tránh khỏi những khó khăn và thiếu sót trong quá trình làm việc, khó khăn lớn nhất là tìm kiếm data API để có thể thực hiện đồ án. Vì vậy, nếu có thêm thời gian nhóm chúng em sẽ cố gắng tìm kiếm những nguồn dữ liệu chất lượng tốt hơn, nghiên cứu thêm nhiều cách xử lý dữ liệu hơn, cũng như tìm hiểu và cài đặt nhiều mô hình hơn nữa.

Thank you!