TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - TIN

BÁO CÁO

Chủ đề: Sports Data Analysis

MÔN XỬ LÝ SỐ LIỆU THỐNG KÊ

Nhóm 19

22280045 – Đặng Lê Khiêm
22280047 – Nguyễn Lê Đăng Khoa
22280048 – Thái Anh Khoa
22280053 – Trần Đại Lộc
22280055 – Lê Thành Nam
22280103 – Nguyễn Hồ Tuyên

Giảng viên hướng dẫn

Ts.Tô Đức Khánh

Table of Contents

I.GIỚI THIỆU

Bài báo cáo này được viết để báo cáo về project cuối kì của môn Xử lý số liệu thống kê do nhóm 19 thực hiện

Chủ đề mà nhóm chọn là chủ đề 1: Sports Data Analysis

Mục tiêu của dự án này là phân tích đánh giá cầu thủ đựa trên các thông tin về tiền lương, quốc tịch, độ tuổi, câu lạc bộ mà họ hiện đang chơi và nhiều biện pháp đánh giá hiệu suất khác nhau. Việc phân tích đánh giá này sẽ giúp cho ban quản lý của câu lạc bộ đưa ra các quyết định mua sắm cầu thủ hợp lý dựa trên ngân sách của câu lạc bộ.

Tổng quan về bộ dữ liệu **Sports Data Analysis**

Dữ liệu gồm thông tin của 18207 cầu thủ, được tổng hợp trong 01 file dữ liệu fifa_eda_stats.csv, bao gồm 57 biến, chẳng hạn:

• ID - mã số của cầu thủ;

• Name - tên cầu thủ;

• Age - tuổi;

• Nationality - quốc tịch;

• Overall - điểm đánh giá tổng thể (tối đa 100);

• Potential - điểm đánh giá tiềm năng (tối đa 100);

• Club - tên câu lạc bộ đang chơi;

• Value - giá trị trên thị trường chuyển nhượng;

• Wage - tiền lương;

• Preferred.Foot - chân thuận;

• Release.Clause - chi phí giải phóng hợp đồng;

• Height - chiều cao;

• Weight - cân nặng;

• Position - vị trí thi đấu sở trường;

• và các biến khác đo các chỉ số đánh giá.

Chú ý có một số biến bị sai định dạng khi nhập vào file lưu trữ (số nhưng lưu ở dạng chữ), do đó cần hiệu chỉnh lại cho đúng trước khi xử lý chính.

II. NỘI DUNG

1. Các phương pháp đề xuất phân tích và xử lý số liệu

- Preprocessing xử lý trước bộ dữ liệu

- Khám phá bộ data dựa trên những phương pháp đã học như là Exploratory Data Analysis (EDA)

- A/B testing trên một số thuộc tính có ý nghĩa thống kê

- Xây dựng mô hình hồi quy chuẩn đoán một số biến quan trọng, có giá trị trong việc đưa ra quyết định lựa chọn cầu thủ

- Xây dựng mô hình phân loại chuẩn đoán một số biến quan trọng, có giá trị trong việc đưa ra quyết định lựa chọn cầu thủ

2. Preprocessing bộ dữ liệu

MỤC TIÊU

- Kiểm tra và xử lý định dạng của các cột dữ liệu định lượng
- Xử lý các cột dữ liệu định tính
- Xử lý dữ liệu bị thiếu

QUÁ TRÌNH

- **Làm sạch dữ liệu:**
    - Kiểm tra định dạng của dữ liệu ở các cột có đơn vị: Thay thế các chữ số bằng ký tự X, ta sẽ xem được định dạng của dữ liệu. Ta loại bỏ các ký hiệu đơn vị như đơn vị tiền tệ, đơn vị cân nặng…
    - Chuyển dữ liệu ở chung một cột về cùng một đơn vị
    - Kiểm tra các biến phân loại
    - Xử lý biến phân loại: Gộp nhóm các vị trí trong position thành 4 nhóm chính, phân tách cột work_rate thành 2 cột riêng, các biến phân loại đưa về dạng factor
- **Dữ liệu khuyết**
    - Đưa ra nhận xét về dữ liệu khuyết và lựa chọn phương pháp xử lý phù hợp

KẾT QUẢ VÀ ĐÁNH GIÁ

- Kết thúc quá trình ta thu được bộ dữ liệu ở dạng chuẩn, có thể sử dụng cho các bước phía sau và không chứa dữ liệu khuyết

3. Exploratory Data Analysis

Các mục tiêu phân tích cần đạt được

- Hiểu rõ hơn về đặc điểm của bộ dữ liệu FIFA (18,000 cầu thủ từ hơn 600 câu lạc bộ và hơn 160 quốc gia).
- Xác định các cầu thủ tốt nhất theo từng vị trí.
- Xác định các đặc trưng nổi bật theo vị trí, nhóm tuổi, hoặc tiềm năng của cầu thủ.
- Tìm kiếm các cầu thủ nổi bật, cầu thủ trẻ tiềm năng và phân tích các yếu tố liên quan đến giá trị thị trường và lương.

Các phương pháp và chiến lược (các bước) phân tích cho mỗi mục tiêu đã đề ra

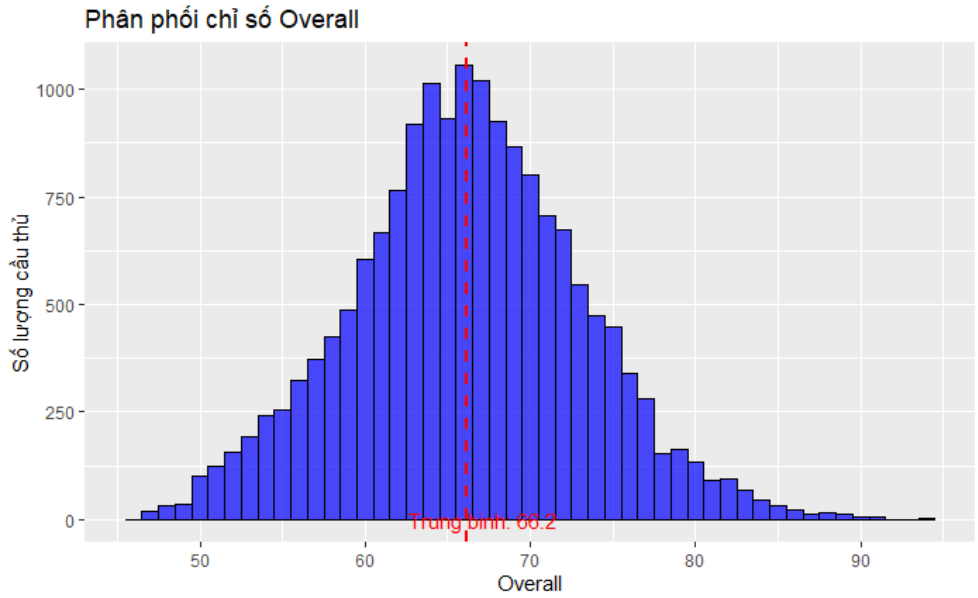
- Giá trị lớn nhất, nhỏ nhất, trung bình, trung vị, độ lệch chuẩn của overall, potential, value và wage
- Số các câu lạc bộ và quốc gia có trong bộ dữ liệu.
- Vẽ biểu đồ histogram cho các chỉ số quan trọng: overall, value, potential, age
- Mối quan hệ giữa overall và wage, value và release_clause bằng cách vẽ biểu đồ phân tán.
- Tìm các cầu thủ có overall, potential, value và wage cao nhất theo từng vị trí
- Giá trị trung bình của overall và potential của các cầu thủ theo từng quốc gia và sắp xếp theo top 10 quốc gia có số lượng cầu thủ nhiều nhất.
- Giá trị trung bình của overall và potential theo từng câu lạc bộ.
- Số lượng cầu thủ, overall trung bình của cầu thủ, lương trung bình và các chỉ số kĩ thuật theo từng vị trí.
- Phân nhóm tuổi để thấy sự khác biệt về overall, potential của 3 nhóm tuổi trẻ, đỉnh cao và xế chiều của sự nghiệp.
- Tìm ra các cầu thủ trẻ tiềm năng cao nhưng không thi đấu cho các câu lạc bộ lớn.

Mô tả và biểu diễn tổng hợp dữ liệu, phân tích kết quả, nhận xét

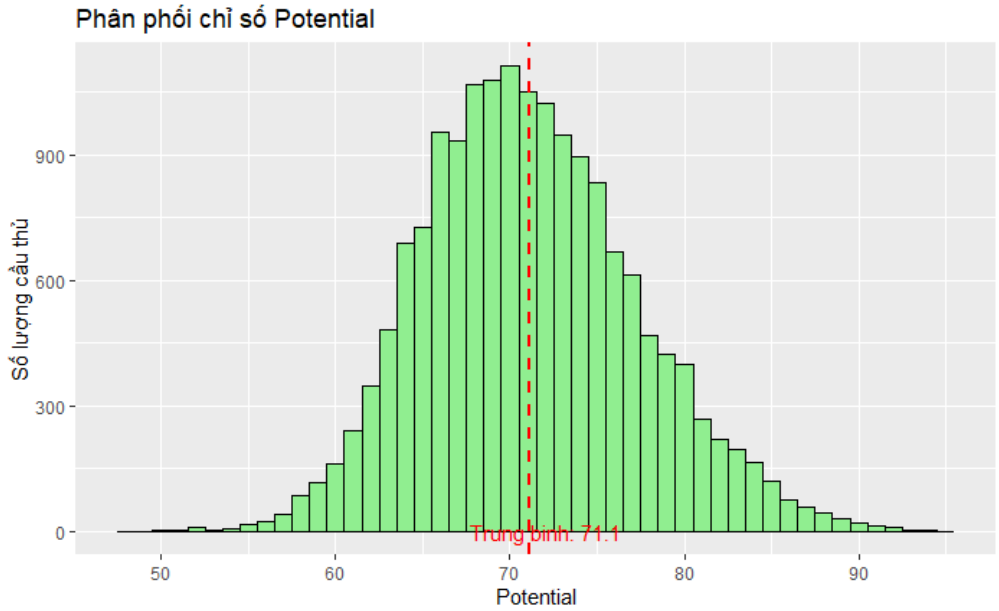
- Các giá trị thống kê cơ bản

bien	gtnn	gtln	tv	tb	dlc
overall	46.00	94.0	66.00	66.16	7.01
potential	48.00	95.0	71.00	71.14	6.15
value	0.01	118.5	0.68	2.44	5.72
wage	1.00	565.0	3.00	9.62	22.26

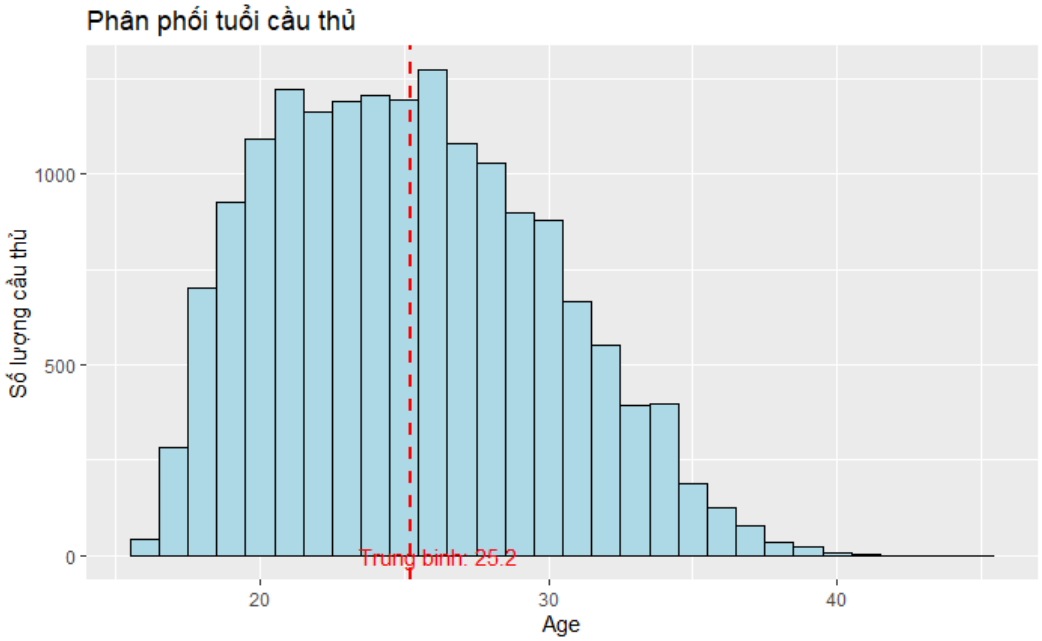
  Có 651 câu lạc bộ và 161 quốc gia trong bộ dữ liệu.
- Các biểu đồ histogram



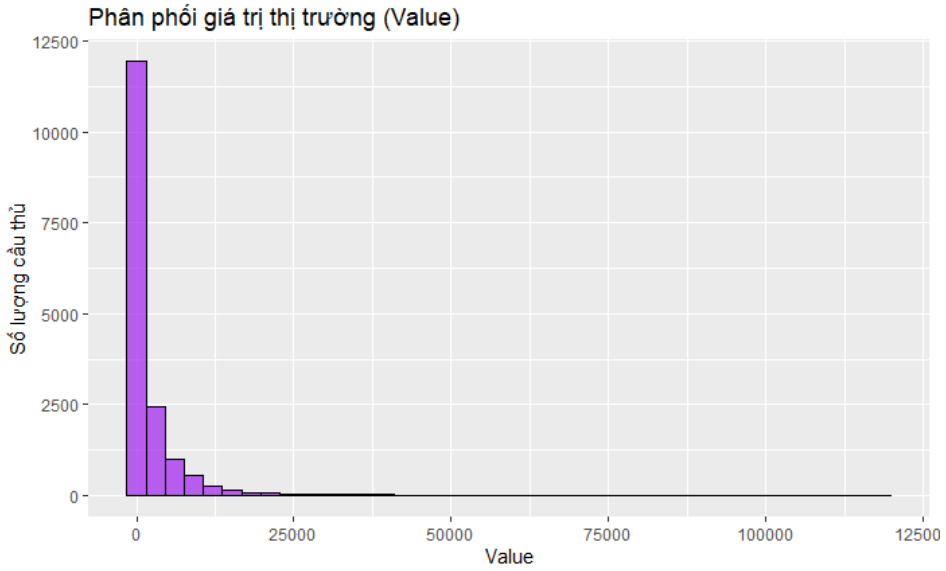
+ Biểu đồ cho thấy phân phối của chỉ số Overall có dạng hình chuông, tập trung chủ yếu xung quanh giá trị trung bình 66.2. Điều này cho thấy một phân phối gần giống với phân phối chuẩn.



+ Tương tự như Overall, biểu đồ cho thấy phân phối của Potential gần giống phân phối chuẩn với trung bình 71.1

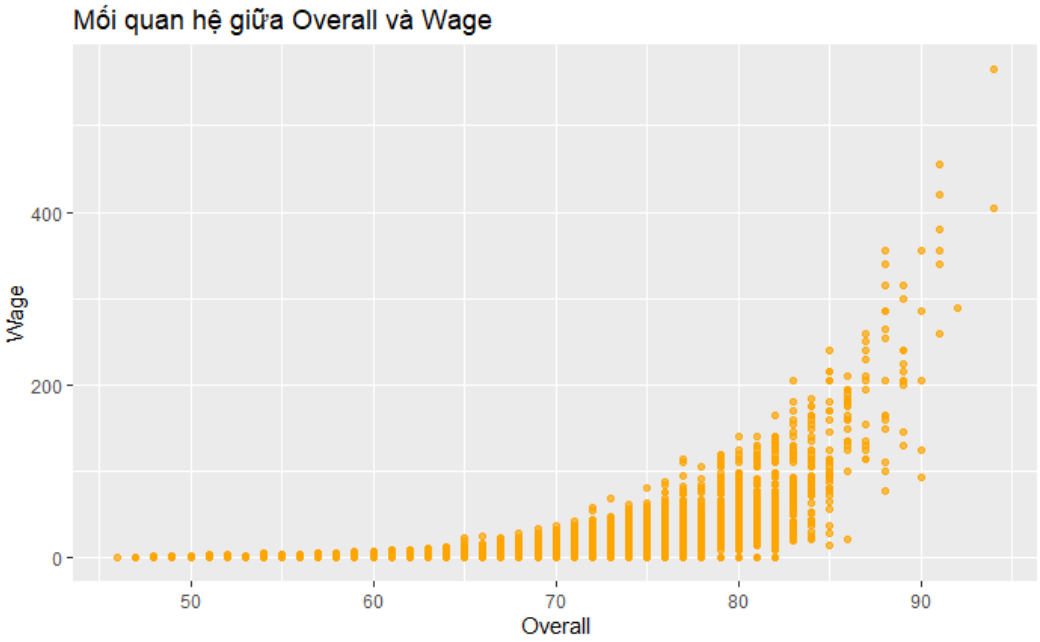


+ Biểu đồ có hình dạng gần giống với phân phối chuẩn, với trung bình 25.2 tuổi. Có một số ít cầu thủ có độ tuổi lớn hơn 35, tạo nên phần đuôi bên phải của biểu đồ.

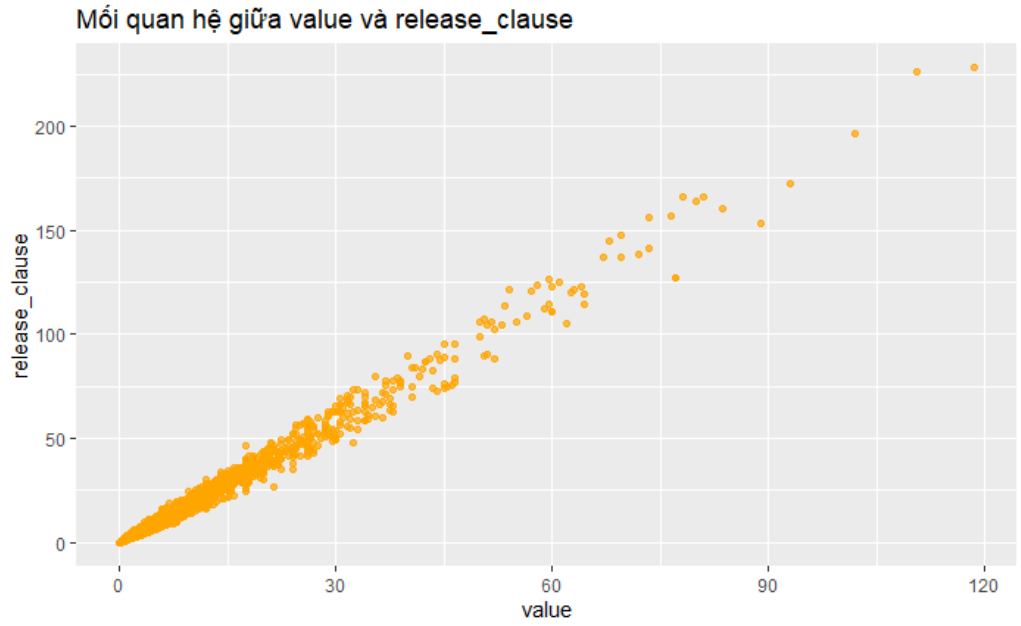


+ Biểu đồ có sự lệch phải, cho thấy có rất nhiều cầu thủ có giá trị dưới 1 triệu euro.

- Các dạng biểu đồ phân tán



+ Các điểm dữ liệu có xu hướng tăng dần từ trái sang phải, cho thấy mối quan hệ dương giữa Overall và Wage. Nói cách khác, khi chỉ số Overall tăng lên thì mức lương cũng có xu hướng tăng theo. Tuy nhiên mối quan hệ này không hoàn toàn tuyến tính, ban đầu, khi chỉ số Overall tăng, mức lương tăng khá chậm, đến khi chỉ số Overall đạt một ngưỡng nhất định, mức lương tăng nhanh hơn.



+ Value và release_clause có mối quan hệ gần như tuyến tính, ta cũng có thể thấy rằng release_clause thường cao hơn nhiều so với value. Điều này có thể hiểu rằng các câu lạc bộ muốn giữ chân cầu thủ của mình nên đưa ra chi phí giải phóng hợp đồng rất cao.

- Các cầu thủ có Overall, Potential, Wage, Value cao nhất theo từng vị trí

position	highest_overall	overall highest_potential	potential highest_wage	wage highest_value	value
Defender	Sergio Ramos	91 S. Umtiti	92 Sergio Ramos	380 S. Umtiti	57.0
Forward	L. Messi	94 L. Messi	94 L. Messi	565 Neymar Jr	118.5
Goalkeeper	De Gea	91 G. Donnarumma	93 De Gea	260 De Gea	72.0
Midfielder	K. De Bruyne	91 K. Mbappé	95 L. Modrić	420 K. De Bruyne	102.0

- Giá trị trung bình của Overall và Potential theo quốc gia

nationality	avg_overall	avg_potential	count
England	63.46576	69.82644	1475
Germany	66.08949	71.60035	1151
Spain	69.73306	74.47639	974
France	67.73857	73.32708	853
Argentina	68.52581	73.08043	833
Brazil	71.25127	73.19162	788
Italy	68.24352	72.23489	579
Colombia	65.22632	70.74211	570
Japan	62.46154	66.44396	455
Netherlands	67.57042	72.92488	426

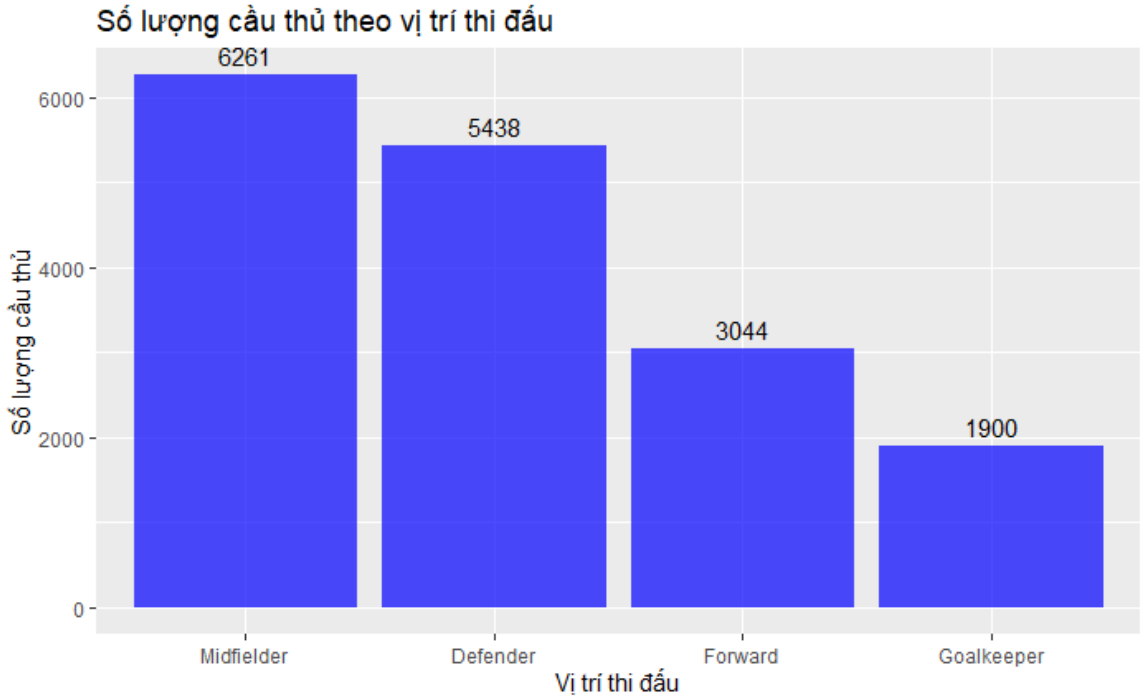
Nước Anh đóng góp nhiều cầu thủ nhất trong bộ dữ liệu với 1475 cầu thủ, kế đến là Đức, Tây Ban nha, Pháp và Argentina ở top 5.

- Giá trị trung bình của Overall và Potential theo câu lạc bộ

club	avg_overall	avg_potential	count
Juventus	82.28000	85.52000	25
Napoli	80.04167	83.62500	24
Inter	79.61905	81.38095	21
Real Madrid	78.24242	84.63636	33
FC Barcelona	78.03030	85.30303	33
Milan	77.54167	82.00000	24
Paris Saint-Germain	77.43333	83.56667	30
Roma	77.40000	82.04000	25
Manchester United	77.24242	82.66667	33
SL Benfica	77.07407	81.74074	27

Theo bảng tổng hợp, Juventus là câu lạc bộ có chỉ số trung bình cầu thủ cũng như tiềm năng là cao nhất, những cái tên còn lại trong top 10 lần lượt là Napoli, Inter, Real Madrid, FC Barcelona, Milan, Paris Saint-German, Roma, Manchester United và Benfica. Điều đặc biệt ta có thể nhận thấy là trong top 10 có đến 5 câu lạc bộ đến từ Ý.

- Các dạng biểu đồ cột
  + Số lượng cầu thủ theo vị trí



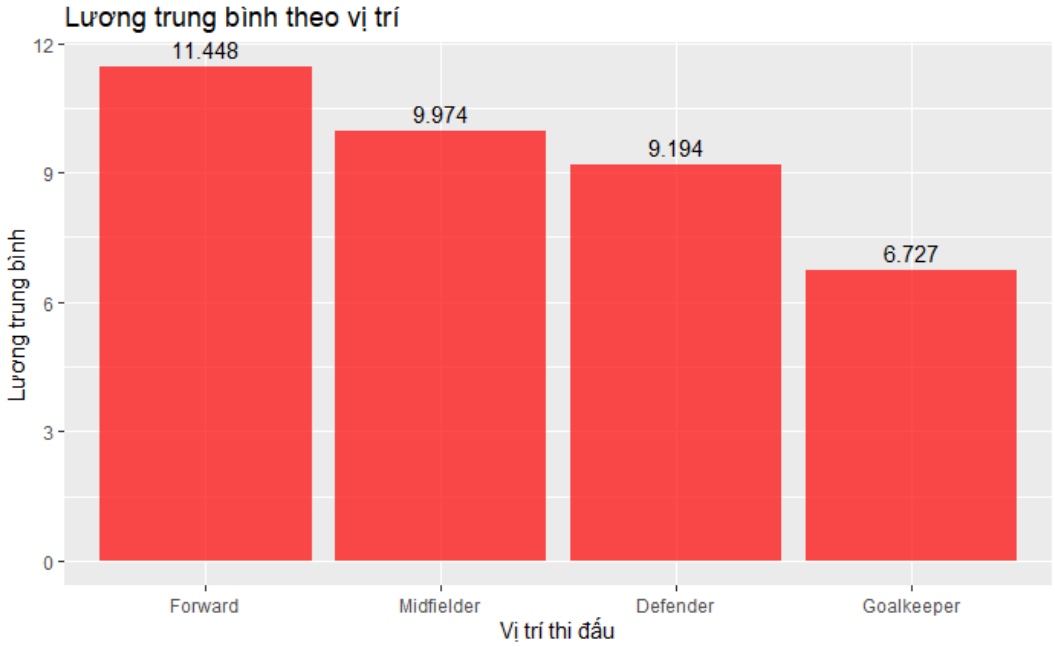
+ Số lượng tiền vệ là nhiều nhất, kế đến lần lượt là hậu vệ và tiền đạo, thủ môn là ít nhất.

- Overall trung bình theo vị trí

position	avg_overall
Midfielder	66.45248
Defender	66.37164
Forward	66.25460
Goalkeeper	64.46316

Nhìn chung, chỉ số overall trung bình của các vị trí khá tương đồng, chỉ có thủ môn là thấp hơn hẳn.

- Trung bình lương theo vị trí (đơn vị: nghìn euro/tuần)

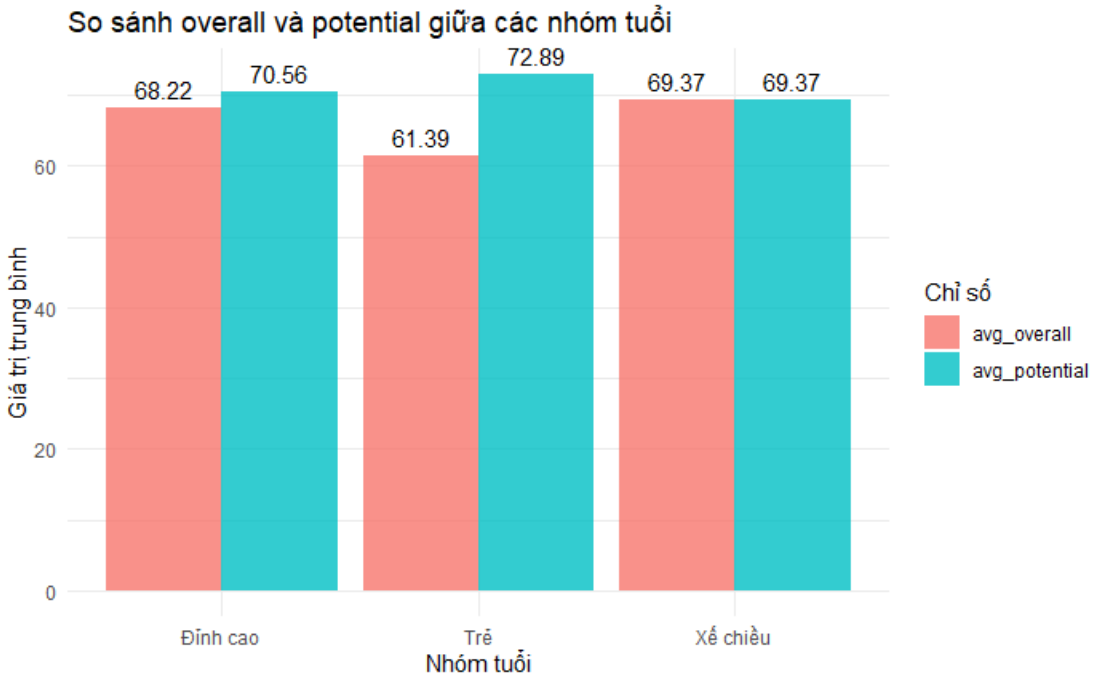


Tiền đạo có lương trung bình cao nhất, tiếp theo lần lượt là tiền vệ, hậu vệ và thủ môn.

- Các chỉ số nổi bật nhất cho từng vị trí

position	metric	mean_value
Defender	strength	70.63
Defender	jumping	69.13
Defender	stamina	68.13
Defender	standing_tackle	66.66
Defender	aggression	65.51
Forward	sprint_speed	71.57
Forward	acceleration	71.15
Forward	agility	68.99
Forward	balance	66.77
Forward	jumping	66.38
Goalkeeper	gk_reflexes	65.96
Goalkeeper	gk_diving	65.20
Goalkeeper	gk_positioning	62.87
Goalkeeper	gk_handling	62.73
Goalkeeper	gk_kicking	61.32
Midfielder	balance	70.87
Midfielder	agility	70.35
Midfielder	acceleration	69.39
Midfielder	sprint_speed	68.47
Midfielder	stamina	68.16

+ Theo bảng tổng hợp, hậu vệ nổi bật ở các chỉ số như: sức mạnh, bật nhảy, bền bỉ, cướp bóng, quyết liệt. Tiền đạo nổi bật ở các chỉ số về: tốc độ dốc bóng, tăng tốc, nhanh nhẹn, thăng bằng, bật nhảy. Với tiền vệ, các chỉ số nổi bật là: thăng bằng, nhanh nhẹn, tăng tốc, tốc độ dốc bóng và bền bỉ, khá tương đồng với tiền đạo. Còn thủ môn thì dễ dàng thấy sự nổi bật ở các chỉ số: phản xạ, đổ người, chọn vị trí, dùng tay, phát bóng.



Chỉ số avg_potential thường cao hơn chỉ số avg_overall ở tất cả các nhóm tuổi, đặc biệt với nhóm tuổi trẻ thì cao hơn hẳn. Điều này gợi ý rằng tiềm năng của cầu thủ trẻ thường cao hơn so với hiệu suất thực tế của họ.
Sự chênh lệch giữa hai chỉ số này có xu hướng giảm dần khi tuổi tăng lên. Điều này có thể cho thấy rằng khi tuổi càng cao, khả năng khai thác hết tiềm năng của bản thân càng khó khăn hơn.

- Các cầu thủ trẻ tiềm năng không thuộc các CLB lớn

	nationality	id	name	age	age_group	overall	potential	club	value	wage
1	Argentina	231478	L. Martínez	20	Trẻ	79	86	Inter	18	
2	Argentina	236007	E. Barco	19	Trẻ	73	88	Atlanta United	8.5	
3	Belgium	216393	Y. Tielemans	21	Trẻ	78	87	AS Monaco	16	
4	Brazil	233299	Felipe Vizeu	21	Trẻ	75	87	Udinese	12	
5	Brazil	231943	Richarlison	21	Trẻ	79	86	Everton	18	
6	Brazil	238359	Wendel	20	Trẻ	76	86	Sporting CP	12.5	
7	Brazil	239970	Paulinho	17	Trẻ	71	86	Bayer 04 Leverkusen	4.7	

Bảng phía trên là danh sách các cầu thủ trẻ dưới 23 tuổi, có tiềm năng với chỉ số potential lớn hơn 85, quan trọng nhất là họ không thi đấu cho các câu lạc bộ lớn. Do đó các câu lạc bộ lớn có thể để mắt đến các cầu thủ này.

4. A/B Testing

Với bộ dữ liệu gồm 57 biến, nhóm chỉ ưu tiên chọn những biến cần thiết để đặt ra giả thuyết để khám phá bộ data một cách tối ưu nhất. Một số biến được đặt ra giả thuyết là

- position: Vị trí thi đấu

- body_type: Tỉ lệ cơ thể

- year_old: Được tạo thêm từ age, phân loại nhóm tuổi của cầu thủ

- region: Vùng của cầu thủ

A. Các mục tiêu phân tích cần đạt được

Mục tiêu là cần phải tìm được mức p.value, từ đó so sánh với độ tin cậy để xác định được các giả thuyết đặt ra là chấp nhận hay bác bỏ

Với thuộc tính position (vị trí thi đấu), các giả thuyết được đặt ra là

- ? Trung bình mức lương (wage) giữa các vị trí thi đấu là có như nhau, điểm toàn diện (overall) giữa các vị trí thi đấu có như nhau, điểm tiềm năng (potential) giữa các vị trí thi đấu có như nhau, giá trị cầu thủ (value) giữa các vị trí thi đấu có như nhau
- ? Vị trí thi đấu của cầu thủ có liên quan đến danh tiếng quốc tế (international_reputation), mức tấn công (attacking_work_rate), mức phòng thủ (deffensive_work_rate), số đo cơ thể (body_type)

Với thuộc tính body_type, các giả thuyết được đặt ra là

- ? Tỉ lệ cơ thể của cầu thủ có liên hệ đến mức tấn công (attacking_work_rate), mức phòng thủ (deffensive_work_rate)

Với thuộc tính year_old (độ tuổi), các giả thuyết được đặt ra là

- ? Trung bình mức lương (wage) hay điểm toàn diện (overall) hay điểm tiềm năng (potential) hay giá trị của cầu thủ (value) giữa 2 độ tuổi là như nhau (bên nào lớn hơn ?)
- ? Với các thuộc tính định lượng khác (ngoài 4 kể trên) trong bộ dữ liệu thì có khác nhau giữa 2 nhóm old và young

Với thuộc tính region, các giả thuyết được đặt ra là

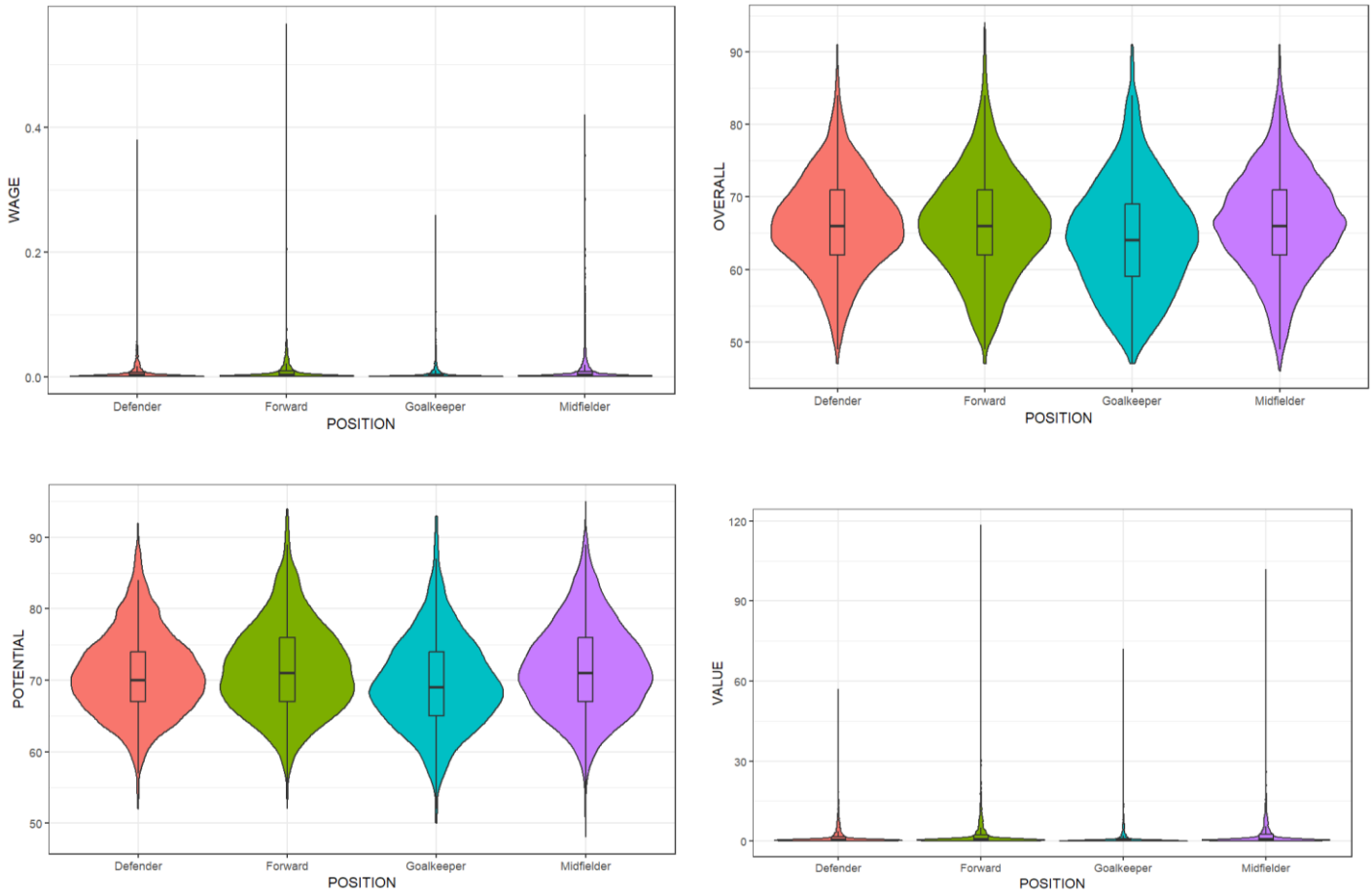
- ? Trung bình mức lương (wage) giữa các châu lục có như nhau, điểm toàn diện (overall) giữa các châu lục có như nhau, điểm tiềm năng (potential) giữa các châu lục có như nhau, giá trị cầu thủ (value) giữa các châu lục có như nhau
- ? South America nội trội hơn Asia về các thuộc tính định lượng nào

B. Các phương pháp và chiến lược (các bước) phân tích + Mô tả và biểu diễn tổng hợp dữ liệu

Thuộc tính position, 4 giả thuyết đầu tiên

? ***Trung bình mức lương (wage) giữa các vị trí thi đấu là có như nhau, điểm toàn diện (overall) giữa các vị trí thi đấu có như nhau, điểm tiềm năng (potential) giữa các vị trí thi đấu có như nhau, giá trị cầu thủ (value) giữa các vị trí thi đấu có như nhau***

Trước hết thì vẽ một biểu đồ box-plot để xác định thử mức phân phối của các biến này so với vị trí thi đấu



Dữ liệu về Wage và Value bị lệch phải rất nhiều vì thế mà rất khó để xác định nó có bằng nhau hay không chỉ bằng mắt thường

Dữ liệu về Potential và Overall nhìn thì có thể khá là bằng nhau giữa các cầu thủ tuy nhiên cần phải kiểm định mới biết rõ

- ✓ Áp dụng phương pháp **Permutation ANOVA** để xác định xem các chỉ số giữa các vị trí có trung bình bằng nhau hay không với mức ý nghĩa là 0.05
  H_0: Giá trị trung bình về mức lương (wage)/ điểm toàn diện (overall)/ điểm tiềm năng (potential)/ giá trị cầu thủ (value) giữa các vị trí thi đấu là như nhau
  H_1: Giá trị trung bình về mức lương (wage)/ điểm toàn diện (overall)/ điểm tiềm năng (potential)/ giá trị cầu thủ (value) giữa các vị trí thi đấu khác nhau

Thuộc tính position, 4 giả thuyết cuối

? ***Vị trí thi đấu của cầu thủ có liên quan đến danh tiếng quốc tế (international_reputation), mức tấn công (attacking_work_rate), mức phòng thủ (deffensive_work_rate), số đo cơ thể (body_type)***

✓ Các biến ở đây tất cả đều là biến định tính, giả thuyết liên quan đến sử dụng kiểm định độc lập với mức ý nghĩa là 0.05

H_0: Vị trí thi đấu của cầu thủ độc lập với danh tiếng quốc tế (international_reputation)/mức tấn công (attacking_work_rate)/mức phòng thủ (deffensive_work_rate)/ số đo cơ thể (body_type)

H_1: Vị trí thi đấu của cầu thủ có liên quan với danh tiếng quốc tế (international_reputation)/mức tấn công (attacking_work_rate)/mức phòng thủ (deffensive_work_rate)/ số đo cơ thể (body_type)

Tuy nhiên, trước hết ta nhìn vào bảng tần số kỳ vọng giữa position với các biến định tính này

Expected Frequencies: position and và international_reputation

	Defender	Forward	Goalkeeper	Midfielder
1	4947.230547	2769.284624	1728.5285105	5695.956318
2	377.062549	211.066274	131.7430752	434.128102
3	95.735985	53.589617	33.4494983	110.224899
4	16.010455	8.962086	5.5939434	18.433516
5	1.960464	1.097398	0.6849727	2.257165

Expected Frequencies: position and àttacking_work_rate

	Defender	Forward	Goalkeeper	Midfielder
High	1450.0898	811.7089	506.65145	1669.5498
Low	276.0987	154.5503	96.46698	317.8841
Medium	3711.8116	2077.7408	1296.88157	4273.5661

Expected Frequencies: position and deffensive_work_rate

	Defender	Forward	Goalkeeper	Midfielder
High	949.8447	531.6895	331.8693	1093.5965
Low	463.3230	259.3518	161.8819	533.4434
Medium	4024.8323	2252.9587	1406.2489	4633.9601

Expected Frequencies: position and body_type

	Defender	Forward	Goalkeeper	Midfielder
Lean	1914.0662	1071.4265	668.7616	2203.7456
Normal	3180.5259	1780.3459	1111.2540	3661.8743
Stocky	343.4079	192.2276	119.9844	395.3801

Ta thấy rõ ràng giữa biến position và international_reputation có tần số kỳ vọng <5 thế nên kết quả ở kiểm định này sẽ không được chính xác, do đó:

✓ Áp dụng phương pháp lấy lại mẫu (resampling method) tạo ra phân phối thực nghiệm hay vì phân phối lý thuyết cho riêng kiểm định này

<u>Thuộc tính body_type</u>, 2 giả thuyết

? ***Tỉ lệ cơ thể của cầu thủ có liên hệ đến mức tấn công (attacking_work_rate), mức phòng thủ (deffensive_work_rate)***

✓ Các biến ở đây tất cả đều là biến định tính, giả thuyết liên quan đến sử dụng kiểm định độc lập với mức ý nghĩa là 0.05

H_0: Tỉ lệ cơ thể (body_type) độc lập với mức tấn công (attacking_work_rate)/mức phòng thủ (deffensive_work_rate)

H_1: Tỉ lệ cơ thể (body_type) có liên với mức tấn công (attacking_work_rate)/mức phòng thủ (deffensive_work_rate)

Bảng tần số kỳ vọng giữa cho thấy không có tần số kỳ vòng nào nhỏ hơn 5 => Không cần hiệu chỉnh

Expected Frequencies: body_type and attacking_work_rate

	High	Low	Medium
Lean	1562.0864	297.42294	3998.4907
Normal	2595.6554	494.21559	6644.1291
Stocky	280.2582	53.36147	717.3803

Expected Frequencies: body_type and deffensive_work_rate

	High	Low	Medium
Lean	1023.2053	499.10737	4335.6873
Normal	1700.2186	829.34639	7204.4350
Stocky	183.5761	89.54624	777.8777

<u>Thuộc tính year_old</u>, 4 giả thuyết đầu tiên

? ***Trung bình mức lương (wage) hay điểm toàn diện (overall) hay điểm tiềm năng (potential) hay giá trị của cầu thủ (value) giữa 2 độ tuổi là như nhau (bên nào lớn hơn ?)***

Trước hết ta xét bảng thống kê giá trị trung bình các giá trị của year_old

year_old <chr>	n <int>	m_wage <dbl>	sd_wage <dbl>	m_overall <dbl>	sd_overall <dbl>	m_potential <dbl>	sd_potential <dbl>	m_value <dbl>	sd_value <dbl>
old	3352	0.012833234	0.02840069	69.45197	5.679395	69.45316	5.679247	2.400260	5.510696
young	13291	0.008807163	0.02034788	65.33323	7.067764	71.56655	6.192862	2.453362	5.772521

✓ Ta dùng kiểm định hoán vị cho để so sánh các giá trị giữa old và young với mức ý nghĩa 0.05

<u>+ Với potential và value</u>

H_0: Giá trị trung bình điểm tiềm năng (potential)/giá trị cầu thủ (value) giữa nhóm old và nhóm young bằng nhau

H_1: Giá trị trung bình điểm tiềm năng (potential)/giá trị cầu thủ (value) nhóm old thấp hơn nhóm young

<u>+ Với overall và wage</u>

H_0: Giá trị trung bình điểm toàn diện (overall)/ lương (wage) giữa nhóm old và nhóm young bằng nhau

H_1: Giá trị trung bình điểm toàn diện (overall)/ lương (wage) nhóm old cao hơn nhóm young

<u>Thuộc tính year_old</u>, các giả thuyết còn lại

? ***Với các thuộc tính định lượng khác (ngoài 4 kể trên) trong bộ dữ liệu thì có khác nhau giữa 2 nhóm old và young***

✓ Ta dùng kiểm định hoán vị cho để so sách các giá trị giữa old và young với mức ý nghĩa 0.05

H_0: Giá trị trung bình giữa các biến định lượng là giống nhau giữa young và old

H_1: Giá trị trung bình giữa các biến định lượng là khác nhau giữa young và old

Kết quả sẽ được ghi nhận thành bảng có dạng

Feature <chr>	p_value <lgl>	conclusion <lgl>
crossing	NA	NA
finishing	NA	NA
heading_accuracy	NA	NA
short_passing	NA	NA

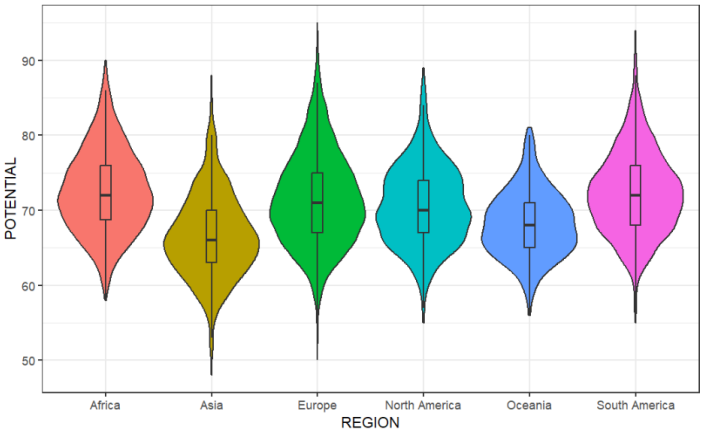
<u>Thuộc tính region</u>, 4 giả thuyết đầu

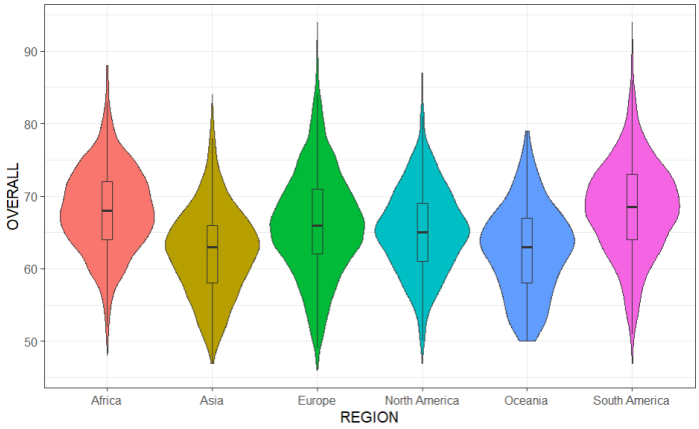
? ***Trung bình mức lương (wage) giữa các châu lục có như nhau, điểm toàn diện (overall) giữa các châu lục có như nhau, điểm tiềm năng (potential) giữa các châu lục có như nhau, giá trị cầu thủ (value) giữa các châu lục có như nhau***

Trước hết ta xét bảng dữ liệu

region <chr>	n <int>	m_wage <dbl>	sd_wage <dbl>	m_overall <dbl>	sd_overall <dbl>	m_potential <dbl>	sd_potential <dbl>	m_value <dbl>	sd_value <dbl>
Africa	1092	0.011170330	0.020628447	68.12363	5.859203	72.51923	5.545627	2.9636401	5.424251
Asia	1603	0.003835309	0.006531329	62.39301	6.065518	66.55770	5.556958	0.8252371	2.000144
Europe	9956	0.010382282	0.023516251	66.11852	7.190611	71.50462	6.212241	2.5952375	6.069388
North America	877	0.006369441	0.011823198	65.30787	5.838277	70.64994	5.298136	1.4905587	2.822079
Oceania	261	0.003597701	0.007333998	62.63602	6.055011	68.16858	4.376767	0.8073563	1.517991
South America	2854	0.011154870	0.026390855	68.26945	6.574542	72.34163	5.593419	3.0616766	6.606177

Về wage và value ta thấy có sự khác nhau rõ rệt giữa các vùng mà không cần dùng kiểm định, riêng potential và overall chưa rõ liệu sự chênh lệch có lớn hay không





- ✓ Tương tự với thuộc tính position khi so sánh 4 thuộc tính này với nhau, ta áp dụng permutation ANOVA với mức ý nghĩa là 0.05
  H_0: Giá trị trung bình về điểm toàn diện (overall)/ điểm tiềm năng (potential) giữa các vùng là như nhau
  H_1: Giá trị trung bình về điểm toàn diện (overall)/ điểm tiềm năng (potential) giữa các vùng khác nhau

Thuộc tính region, các giả thuyết cuối

- ? ***South America nội trội hơn Asia về các thuộc tính định lượng nào***
  Ta dùng kiểm định hoán vị cho để so sách các giá trị giữa old và young với mức ý nghĩa 0.05
  H_0: Giá trị trung bình giữa các biến định lượng là giống nhau giữa young và old
  H_1: Giá trị trung bình giữa các biến định lượng là khác nhau giữa young và old
  Kết quả sẽ được ghi nhận thành bảng có dạng

C. Phân tích kết quả

Thuộc tính position, 4 giả thuyết đầu tiên

- ? ***Trung bình mức lương (wage) giữa các vị trí thi đấu là có như nhau, điểm toàn diện (overall) giữa các vị trí thi đấu có như nhau, điểm tiềm năng (potential) giữa các vị trí thi đấu có như nhau, giá trị cầu thủ (value) giữa các vị trí thi đấu có như nhau***

Kết quả

```
summary(anova_po_wage)
```

```
## Component 1 :
##               Df R Sum Sq R Mean Sq Iter  Pr(Prob)
## position1      3    27848    9282.6 5000 < 2.2e-16 ***
## Residuals   16639  8220997     494.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_po_overall)
```

```
## Component 1 :
##               Df R Sum Sq R Mean Sq Iter  Pr(Prob)
## position1      3     6277   2092.31 5000 < 2.2e-16 ***
## Residuals   16639   811101     48.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_po_overall)
```

```
## Component 1 :
##               Df R Sum Sq R Mean Sq Iter  Pr(Prob)
## position1      3     6277   2092.31 5000 < 2.2e-16 ***
## Residuals   16639   811101     48.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_po_value)
```

```
## Component 1 :
##               Df R Sum Sq R Mean Sq Iter  Pr(Prob)
## position1      3     4114   1371.49 5000 < 2.2e-16 ***
## Residuals   16639   540505     32.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p.value của 4 lần chạy permutation ANOVA đều nhỏ hơn mức ý nghĩa 0.05, ta bác bỏ H0

=> Giá trị trung bình về mức lương (wage)/ điểm toàn diện (overall)/ điểm tiềm năng (potential)/ giá trị cầu thủ (value) giữa các vị trí thi đấu khác nhau

Thuộc tính position, 4 giả thuyết cuối

? ***Vị trí thi đấu của cầu thủ có liên quan đến danh tiếng quốc tế (international_reputation), mức tấn công (attacking_work_rate), mức phòng thủ (deffensive_work_rate), số đo cơ thể (body_type)***

Kết quả

```
## p-value: 0
## Observed p-value: 0
## Observed p-value: 0
## Observed p-value: 3.029882e-68
```

Tất cả các kiểm định có p.value gần bằng 0 nhỏ hơn mức ý nghĩa 0.05, bác bỏ H0

- Vị trí thi đấu của cầu thủ có liên quan với danh tiếng quốc tế (international_reputation)/mức tấn công (attacking_work_rate)/mức phòng thủ (deffensive_work_rate)/ số đo cơ thể (body_type)

Thuộc tính body_type, 2 giả thuyết

? ***Tỉ lệ cơ thể của cầu thủ có liên hệ đến mức tấn công (attacking_work_rate), mức phòng thủ (deffensive_work_rate)***

Kết quả

```
## Observed p-value: 5.462144e-13
## Observed p-value: 1.098452e-10
```

Tất cả các kiểm định có p.value gần bằng 0 nhỏ hơn mức ý nghĩa 0.05, bác bỏ H0

- Tỉ lệ cơ thể (body_type) có liên quan với mức tấn công (attacking_work_rate)/mức phòng thủ (deffensive_work_rate)

Thuộc tính year_old, 4 giả thuyết đầu tiên

? ***Trung bình mức lương (wage) hay điểm toàn diện (overall) hay điểm tiềm năng (potential) hay giá trị của cầu thủ (value) giữa 2 độ tuổi là như nhau (bên nào lớn hơn ?)***

+ Với potential và value

```
## P-value for potential: 0
```

p-value của potential < 0.05, bác bỏ H0 => Giá trị trung bình điểm tiềm năng (potential) nhóm old thấp hơn nhóm young

```
## P-value for value: 0.308
```

p-value của value > 0.05, chấp nhận H0 => Giá trị trung bình giá trị cầu thủ (value) nhóm old bằng (không thấp hơn) nhóm young

+ Với overall và wage

```
## P-value for overall: 0 ## P-value for wage: 0
```

p-value của overall và wage < 0.05, bác bỏ H0 => Giá trị trung bình điểm toàn diện (overall)/ lương (wage) nhóm old cao hơn nhóm young

Thuộc tính year_old, các giả thuyết còn lại

? ***Với các thuộc tính định lượng khác (ngoài 4 kể trên) trong bộ dữ liệu thì có khác nhau giữa 2 nhóm old và young***

Ta được bảng kết quả:

	Feature	p_value	conclusion
1	crossing	0.000	H1
2	finishing	0.008	H1
3	heading_accuracy	0.000	H1
4	short_passing	0.000	H1
5	volleys	0.000	H1
6	dribbling	0.000	H1
7	curve	0.000	H1
8	fk_accuracy	0.000	H1
9	long_passing	0.000	H1
10	ball_control	0.084	H0
11	acceleration	0.000	H1
12	sprint_speed	0.000	H1
13	agility	0.000	H1
14	reactions	0.000	H1
15	balance	0.000	H1
16	shot_power	0.000	H1
17	jumping	0.000	H1
18	stamina	0.000	H1
19	strength	0.000	H1
20	long_shots	0.000	H1
21	aggression	0.000	H1
22	interceptions	0.000	H1
23	positioning	0.005	H1
24	vision	0.000	H1
25	penalties	0.000	H1
26	composure	0.000	H1
27	marking	0.000	H1
28	standing_tackle	0.000	H1
29	sliding_tackle	0.000	H1
30	gk_diving	0.000	H1
31	gk_handling	0.000	H1
32	gk_kicking	0.000	H1
33	gk_positioning	0.000	H1
34	gk_reflexes	0.000	H1

- Chỉ có trung bình ball_control giữa hai nhóm old và young là bằng nhau, còn lại là khác nhau

Thuộc tính region, 4 giả thuyết đầu

? ***Trung bình mức lương (wage) giữa các châu lục có như nhau, điểm toàn diện (overall) giữa các châu lục có như nhau, điểm tiềm năng (potential) giữa các châu lục có như nhau, giá trị cầu thủ (value) giữa các châu lục có như nhau***

Kết quả

```
summary(anova_re_overall)
```

```
## Component 1 :
##               Df R Sum Sq R Mean Sq Iter  Pr(Prob)
## region1        5    43552    8710.4 5000 < 2.2e-16 ***
## Residuals   16637   773826      46.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_re_potential)
```

```
## Component 1 :
##               Df R Sum Sq R Mean Sq Iter  Pr(Prob)
## region1        5    43696    8739.2 5000 < 2.2e-16 ***
## Residuals   16637   586035      35.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Với p-value < 0.05, ta bác bỏ H0, cùng với giá trị wage và value đã xác định không bằng nhau từ trước

- Trung bình mức lương (wage) giữa các châu lục có như nhau, điểm toàn diện (overall) giữa các châu lục có như nhau, điểm tiềm năng (potential) giữa các châu lục có như nhau, giá trị cầu thủ (value) giữa các châu lục có sự khác nhau

Thuộc tính region, các giả thuyết cuối

? ***South America nội trội hơn Asia về các thuộc tính định lượng nào***

Ta thu được bảng kết quả

```
##              Feature p_value conclusion
## 1           crossing   0.000         H1
## 2         finishing   0.000         H1
## 3  heading_accuracy   0.000         H1
## 4     short_passing   0.000         H1
## 5           volleys   0.000         H1
## 6         dribbling   0.000         H1
## 7             curve   0.000         H1
## 8       fk_accuracy   0.000         H1
## 9      long_passing   0.000         H1
## 10     ball_control   0.000         H1
## 11     acceleration   0.000         H1
## 12     sprint_speed   0.000         H1
## 13          agility   0.000         H1
## 14        reactions   0.000         H1
## 15          balance   0.245         H0
## 16       shot_power   0.000         H1
```

```
## 17          jumping   0.000         H1
## 18          stamina   0.000         H1
## 19         strength   0.000         H1
## 20       long_shots   0.000         H1
## 21       aggression   0.000         H1
## 22    interceptions   0.000         H1
## 23      positioning   0.000         H1
## 24           vision   0.000         H1
## 25        penalties   0.000         H1
## 26        composure   0.000         H1
## 27          marking   0.000         H1
## 28  standing_tackle   0.000         H1
## 29   sliding_tackle   0.000         H1
## 30        gk_diving   0.999         H0
## 31      gk_handling   0.995         H0
## 32       gk_kicking   0.997         H0
## 33   gk_positioning   0.996         H0
## 34      gk_reflexes   0.991         H0
```

=> Ngoại trừ balance, gk_diving, gk_handling, gk_kicking, gk_positioning, gk_reflexes các thuộc tính còn lại South America đều có trung bình cao hơn Asia

D. Nhận xét, kết luận

Với position:

- Ta rút ra được ở các đặc tính wage, value, potential, overall ở mỗi vị trí sẽ khác nhau, do vai trò ở mỗi vị trí khác nhau nên giá trị mà cầu thủ đem lại cũng khác nên mới có sự khác nhau này
- Position có phụ thuộc vào các giá trị international_reputation, attacking_work_rate /deffensive_work_rate/body_type. Mỗi một vị trí sẽ xét đến các tiêu chí này để xem có phù hợp không

Với body_type:

- Body_type có liên quan với mức tấn công (attacking_work_rate)/mức phòng thủ (deffensive_work_rate). Khi chọn cầu thủ mạnh về attack hay defend thì cần xét tới body_type của cầu thủ này

Với year_old

- Các giá trị về wage, value, overall của nhóm old đa phần vượt trội hơn nhóm young. Nếu muốn đầu tư lâu dài và tiết kiệm chi phí có thể ưu tiên chọn những cầu thủ nhóm young, các cầu thủ nhóm old nên chọn trong trường hợp cần ngay vào luôn để làm mạnh đội hình
- Các đặc tính định lượng của nhóm old và young khác nhau (chỉ giống ở ball_control), khi chiêu mộ cầu thủ cần chú ý đến các đặc tính này nếu cần

Với region

- Ta rút ra được ở các đặc tính wage, value, potential, overall ở mỗi vùng sẽ khác nhau do có vùng tập trung vào bóng đá nhiều hơn các vùng còn lại, một phần cũng là do nền kinh tế định giá khác nhau ở mọi nước. Nếu muốn chọn cầu thủ tốt thì cần phải chú ý đến các vùng nội trội ở các đặc điểm này
- Ở vùng Bắc Mỹ có các chỉ số đều nội trội hơn vùng châu Á, nếu chiêu mộ cầu thủ thì nên ưu tiên ở Bắc Mỹ hơn nếu ngân sách cho phép. Từ đây cũng rút ra được châu Á đang bị kém ở những đặc điểm nào để cải thiện tốt hơn

5. Mô hình hồi quy

TỔNG QUAN

- Chỉ số **'overall'**, **'potential', 'wage','value'** là các biến quan trọng hỗ trợ huấn luyện viên trong việc đưa ra quyết định chiêu mộ cầu thủ. Mô hình hồi quy tuyến tính sẽ giúp đánh giá tiềm năng và năng lực hiện tại của một cầu thủ mới, ngay cả khi chưa có đầy đủ dữ liệu thực tế về họ.
- Mô hình dự đoán này sẽ đóng vai trò như một công cụ hỗ trợ đắc lực, giúp huấn luyện viên tối ưu hóa chiến lược chuyển nhượng dựa trên cả nhu cầu ngắn hạn và dài hạn của câu lạc bộ.

Mô hình hồi quy cho biến Value

Mục tiêu hồi quy

- Xây dựng mô hình hồi quy cho biến dự đoán *value* dựa trên các biến giải thích có trong bộ dữ liệu với độ chính xác cao hơn so với mô hình hồi tuyến tính đơn giản dựa trên tất cả các biến giải thích.

- Chọn lọc các biến có tác động lớn đối với biến *value* trong mô hình hồi quy
- Giảm độ phức tạp của mô hình cuối cùng so với mô hình ban đầu

Phương pháp, chiến lược phân tích + Mô tả và biểu diễn tổng hợp dữ liệu + Kết quả

- **Bước 1:** Xây dựng mô hình hồi quy cho biến dự đoán *value* dựa trên tất cả các biến phân loại và đánh giá mô hình hiện tại.
  Đối với mô hình này, khi đánh giá bằng phương pháp cross-validation, ta thu được kết quả sau: RMSE: 0.573021, Rsquared: 0.9901128, MAE: 0.2563513. Đây là kết quả rất tốt, cho thấy sự giải thích cao của các biến trong mô hình. Tuy nhiên, mô hình này rất phức tạp vì sử dụng tổng cộng khoảng 58 biến giải thích (bao gồm các biến dummy được tạo ra từ các biến định tính). Do đó, trong các bước tiếp theo, ta sẽ đơn giản hóa mô hình nhưng mong muốn vẫn đảm bảo hiệu suất của mô hình.
- **Bước 2:** Sử dụng phương pháp Hồi quy từng bước để đánh giá và Hồi quy Lasso nhằm giảm số lượng các biến giải thích
  Sau khi sử dụng phương pháp Hồi quy từng bước, ta xác nhận được rằng khi thêm biến giải thích vào mô hình hồi quy, độ chính xác của mô hình được cải thiện rất nhiều. Đây là dấu hiệu xác nhận sự phù hợp của mô hình hồi quy. Số lượng biến giải thích trong mô hình hồi quy còn 12 biến.
  Ta sử dụng phương pháp hồi quy Lasso với mong muốn tìm một mô hình hồi quy đơn giản hơn. Kết quả thu được chỉ còn 6 biến giải thích: *overall*, *wage*, *international_reputation*, *volleys*, *release_clause*, và *stamina*. Mô hình này có độ chính xác là Average RMSE: 0.5781514 Average R-squared: 0.9897874 Average R-squared adjusted: 0.9897874 khi sử dụng phương pháp cross-validation.
- **Bước 3:** Chuẩn đoán mô hình
  Bước này nhằm kiểm tra các giả thuyết cho mô hình hồi quy từ 6 biến giải thích trên.
- **Bước 4:** Mở rộng mô hình
  Bằng cách mở rộng mô hình hồi quy bằng mô hình Hồi quy Tổng quát và thêm các biến tương tác vào mô hình, ta thu được mô hình với độ chính xác Average RMSE: 0.5166711, Average R-squared: 0.9917792, , Average R-squared adjusted: 0.9917792

Tổng quan và hiệu quả, nhận xét

- Kết thúc quá trình ta xây dựng được mô hình hồi quy cho biến dự đoán value dựa vào 7 biến giải thích *overall*, *wage*, *international_reputation*, *volleys*, *release_clause*, và *stamina* với độ chính xác cao hơn mô hình ban đầu (RMSE giảm từ 0.573021 xuống 0.5166711)

Mô hình hồi quy cho biến Wage

Mục tiêu hồi quy

- Xây dựng mô hình hồi quy cho biến dự đoán *wage* dựa trên các biến giải thích có trong bộ dữ liệu với độ chính xác cao hơn so với mô hình hồi tuyến tính đơn giản dựa trên tất cả các biến giải thích.
- Chọn lọc các biến có tác động lớn đối với biến *wage* trong mô hình hồi quy
- Giảm độ phức tạp của mô hình cuối cùng so với mô hình ban đầu

Phương pháp, chiến lược phân tích + Mô tả và biểu diễn tổng hợp dữ liệu + Kết quả

- **Bước 1:** Xây dựng mô hình hồi quy cho biến dự đoán *wage* dựa trên tất cả các biến phân loại và đánh giá mô hình hiện tại.
  Đối với mô hình này, khi đánh giá bằng phương pháp cross-validation, ta thu được kết quả sau: RMSE: 10.67, Rsquared: 0.7711, MAE: 0.7703. Mô hình có khả năng giải thích thấp và độ phức tạp của mô hình cao vì sử dụng tổng cộng khoảng 58 biến giải thích (bao gồm các biến dummy được tạo ra từ các biến định tính). Do đó, trong các bước tiếp theo, ta sẽ tìm cách cải thiện độ chính xác của và đơn giản hóa mô hình.
- **Bước 2:** Sử dụng phương pháp Hồi quy từng bước để đánh giá và Hồi quy Lasso nhằm giảm số lượng các biến giải thích
  Sau khi sử dụng phương pháp Hồi quy từng bước, ta nhận thấy rằng khi tăng số lượng biến giải thích mô hình độ chính xác của mô hình giảm đi đánh kể. Dấu hiện này cho thấy việc sử dụng mô hình hồi quy cho biến dự đoán *wage* dựa trên các biến giải thích trong bộ dữ liệu là không phù hợp.
  Ta sử dụng phương pháp hồi quy Lasso vào chỉ giảm được 6 biến giải thích, trong đó các biến *international_reputation, position, value, deffensive_work_rate, skill_moves* là những biến có hệ số lớn hơn so với các biến còn lại. Để kiểm tra điều này, ta thử xây dựng mô hình hồi quy Lasso dựa trên các biến trên và cả hai mô hình trên không có sự khác biệt lớn. Có thể điều này do sự không phù hợp của mô hình hồi quy. Ta sẽ chọn mô hình đơn giản hơn cho các bước tiếp theo
- **Bước 3:** Chuẩn đoán mô hình
  Bước này nhằm kiểm tra các giả thuyết cho mô hình hồi quy từ 5 biến giải thích trên.
- **Bước 4:** Mở rộng mô hình
  Bằng cách mở rộng mô hình hồi quy bằng mô hình Hồi quy Tổng quát và thêm các biến tương tác giữa *release_clause* và *value* vào mô hình, ta thu được mô hình với độ chính xác trên toàn bộ tập dữ liệu là RMSE: 10.21275, , R-squared adjusted: 0.789

Tổng quan và hiệu quả, nhận xét

- Kết thúc quá trình ta xây dựng được mô hình hồi quy cho biến dự đoán *wage* dựa vào 5 biến giải thích *international_reputation, position, value, deffensive_work_rate, skill_moves* với độ chính xác cao hơn mô hình ban đầu (RMSE giảm từ 10.67 xuống 10.21275).
- Tuy nhiên, độ chính xác của mô hình rất thấp, do đó, việc xây dựng mô hình hồi quy cho biến wage dựa vào các bến giải thích trong bộ dữ liệu là không phù hợp

Mô hình hồi quy cho biến overall

Mục tiêu hồi quy

- **Chỉ số 'overall'**: Phản ánh năng lực hiện tại của cầu thủ, phù hợp để đáp ứng nhu cầu cấp thiết của đội bóng. Các cầu thủ có chỉ số 'overall' cao thường được lựa chọn để nâng cao chất lượng đội hình ngay lập tức.

Phương pháp, chiến lược phân tích + Mô tả và biểu diễn tổng hợp dữ liệu + Kết quả

- Bước 1: Chúng ta sẽ xây dựng baseline model bằng cách dung hồi quy biến được chọn với tất cả các biến còn lại.

	Metric	Train	Test
1	MSE	3.105704	3.1407562
2	MAE	1.376685	1.3953978
3	Adjusted R2	0.936822	0.9334985

- Bước 2: Ta sẽ thực hiện Lựa chọn mô hình
  + Cách 1: Sử dựng hồi quy từng bước nó sẽ giúp ta chọn được bộ biến có giá trị mean square error tốt nhất. Nó sẽ được đánh giá bằng Cross Validation

	Metric	Train	Test
1	MSE	3.2069351	3.2131129
2	MAE	1.3995615	1.4078213
3	Adjusted R2	0.9348659	0.9324005

  + Cách 2: Sử dụng Lasso để co hệ số, phương pháp này sẽ đẩy hệ số của các biến không quan trọng về 0 sau đó ta sẽ dụng Cross Validation tương tự trên.

	Metric	Train	Test
1	MSE	3.1328745	3.178213
2	MAE	1.3833992	1.402746
3	Adjusted R2	0.9363221	0.932931

- Phía trên là thông số từ các mô hình theo từng cách của bước 2, các thông số không tốt hơn tuy nhiên bằng Lasso ta có thể giảm độ phức tạp của thuật toán vì thế ta sẽ dung nó để thực hiện các bước tiếp theo
- Bước 3: Chuẩn đoán mô hình cho mô hình hồi quy
  + Kiểm tra tính tuyến tính mô hình
  + Kiểm tra tuyến tính từng phần
  + Kiểm tra đồng nhất phương sai
  + Kiểm tra điểm ngoại lai
  + Kiểm tra đa cộng tuyến
- Bước 4: Mở rộng mô hình
  + Ta thêm bậc 4 cho biến ‘age’, bậc 2 cho biến ‘value’ và ‘potential’ sau đó ta được mô hình.

	Metric	Train	Test
1	MSE	1.5574229	2.6839806
2	MAE	0.9411552	1.0527663
3	Adjusted R2	0.9663734	0.9434984

  Ta thấy kết quả thu được khá tốt
- Bước 5: Xây dựng khoảng tin cậy
  Ta dùng 1 điểm bất kì, có thông số ‘overall’ là 86 thì thu được:
  + Khoảng tin cậy cho trung bình

```
          2.5%    97.5%
82.71068 85.24096
```

+ Khoảng tin cậy cho biến dự đoán

```
          2.5%    97.5%
81.41045 86.92959
```

Tổng quan và hiệu quả, nhận xét

Bằng việc xây dựng mô hình hồi quy đa thức ta đã tạo được một mô hình có chỉ số ổn định từ ban đầu trên tập test, chỉ số MSE là 3.140707 đến cuối cùng ta thu được MSE trên tập test chỉ còn 2.68.

Mô hình đã dự đoán có kết quả chính xác cao và có thể ứng dụng được

Mô hình hồi quy cho biến potential

Mục tiêu hồi quy

Chỉ số 'potential': Đây là yếu tố ưu tiên đối với các cầu thủ trẻ, đại diện cho tiềm năng phát triển dài hạn. Dựa trên chỉ số này, câu lạc bộ có thể đầu tư vào những tài năng triển vọng, nhằm xây dựng đội hình bền vững trong tương lai.

Phương pháp, chiến lược phân tích + Mô tả và biểu diễn tổng hợp dữ liệu + Kết quả

- Bước 1: Chúng ta sẽ xây dựng baseline model bằng cách dung hồi quy biến được chọn với tất cả các biến còn lại.

```
        Metric     Train      Test
1          MSE 5.9263871 5.8796982
2          MAE 1.9126669 1.9026382
3 Adjusted R2 0.8443281 0.8349253
```

- Bước 2: Ta sẽ thực hiện Lựa chọn mô hình
  + Cách 1: Sử dựng Hồi quy từng bước nó sẽ giúp ta chọn được bộ biến có giá trị mean square error tốt nhất. Nó sẽ được đánh giá bằng Cross Validation

```
        Metric     Train      Test
1          MSE 5.9605624 5.9274720
2          MAE 1.9178459 1.9110744
3 Adjusted R2 0.8436898 0.8346958
```

  + Cách 2: Sử dụng Lasso để co hệ số, phương pháp này sẽ đẩy hệ số của các biến không quan trọng về 0 sau đó ta sẽ dụng Cross Validation tương tự trên.

	Metric	Train	Test
1	MSE	5.9694848	5.9365298
2	MAE	1.9179766	1.9108106
3	Adjusted R2	0.8434086	0.8342419

Đây là thông số từ các mô hình theo từng cách ta thấy không có sự chênh lệch rõ ràng giữa cả hai cách vì thế ta chọn ngẫu nhiên 1 bộ data, trong trường hợp này là bộ data của hồi quy từng bước

- Bước 3: Chuẩn đoán mô hình cho mô hình hồi quy
  + Kiểm tra tính tuyến tính mô hình
  + Kiểm tra tuyến tính từng phần
  + Kiểm tra đồng nhất phương sai
  + Kiểm tra điểm ngoại lai
  + Kiểm tra đa cộng tuyến
- Bước 4: Mở rộng mô hình:
  Ta thêm bậc 4 cho biến 'age' thì thu được mô hình:

	Metric	Train	Test
1	MSE	2.1285713	3.2472175
2	MAE	1.0426589	1.1800429
3	Adjusted R2	0.9410803	0.9096891

Ta thấy kết quả thu được khá tốt từ Ajdjusted R2 trên tập train ban đầu chỉ khoảng 0.84 thì mô hình này đã tang lên khoảng 0.94.

- Bước 4: Xây dựng khoảng tin cậy
  Ta dùng 1 điểm bất kì, có thông số 'poential là 88 thì thu được:
  + Khoảng tin cậy cho trung bình

2.5%	97.5%
88.19318	88.50122

  + Khoảng tin cậy cho biến dự đoán

2.5%	97.5%
85.18643	92.16681

Tổng quan và hiệu quả, nhận xét

Bằng việc xây dựng mô hình hồi quy đa thức ta đã tạo được một mô hình có chỉ số ổn định từ ban đầu trên tập test, chỉ số MSE là 5.879 đến cuối cùng ta thu được MSE trên tập test chỉ còn 3.24

Mô hình đã dự đoán có kết quả chính xác cao và có thể ứng dụng được

6. Mô hình phân loại

A. Mục tiêu phân loại

- Thực hiện xây dựng mô hình phân loại để giúp ban quản lý của câu lạc bộ đưa ra các quyết định mua sắm cầu thủ hợp lý dựa trên ngân sách của câu lạc bộ.
- Phân loại đa nhóm:
  + Phân loại theo nhóm vị trí thi đấu
  + Phân loại cầu thủ dựa trên khả năng phù hợp với các chiến thuật của đội bóng
  + Phân loại dựa trên xu hướng di chuyển tấn công khi không có bóng.
  + Phân loại dựa trên xu hướng di chuyển phòng thủ khi không có bóng.
  + Phân loại dựa trên danh tiếng quốc tế.

- Phân loại hai nhóm:
  + Phân loại theo chân thuận của cầu thủ
  + Phân loại dựa trên danh tiếng quốc tế

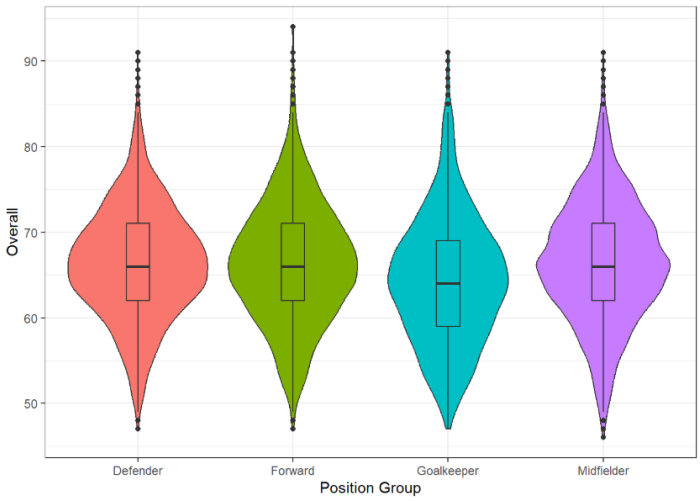
B. Phương pháp và chiến lược phân tích:

- Chia data thành tập train (70%) và tập test (30%)
- A/B testing để kiểm định cho sự khác nhau của các chỉ số overall,value của các nhóm phân loại.
- Kiểm tra sự cân bằng cân bằng của dữ liệu bằng phương pháp SMOTE, under-sampling,...
- Sử dụng các mô hình phân loại: Multinominal logistic cho phân loại đa nhóm và logistic, Naive Bayes,... cho phân loại hai nhóm.(các chỉ số được chọn phù hợp với từng loại nhóm khác nhau)
- Sử dụng tập test để đánh giá mô hình phân loại bằng các chỉ số Precision, Recall, Accuracy, Kappa,Macro_F1 cho các phân loại đa nhóm.
- Sử dụng ước lượng AUC, ROC curve để thể hiện sensitivity và specificity, và các chỉ số đánh giá hiệu suất như mô hình phân loại đa nhóm để đánh giá mô hình phân loại hai nhóm.
- Sử dụng Youden index và Closest top left để tìm ngưỡng threshold hợp lý giúp nâng cao hiệu suất của mô hình dự đoán phân loại hai nhóm.

C. Mô tả và biểu diễn tổng hợp dữ liệu (bảng tổng hợp, biểu đồ) + Kết quả

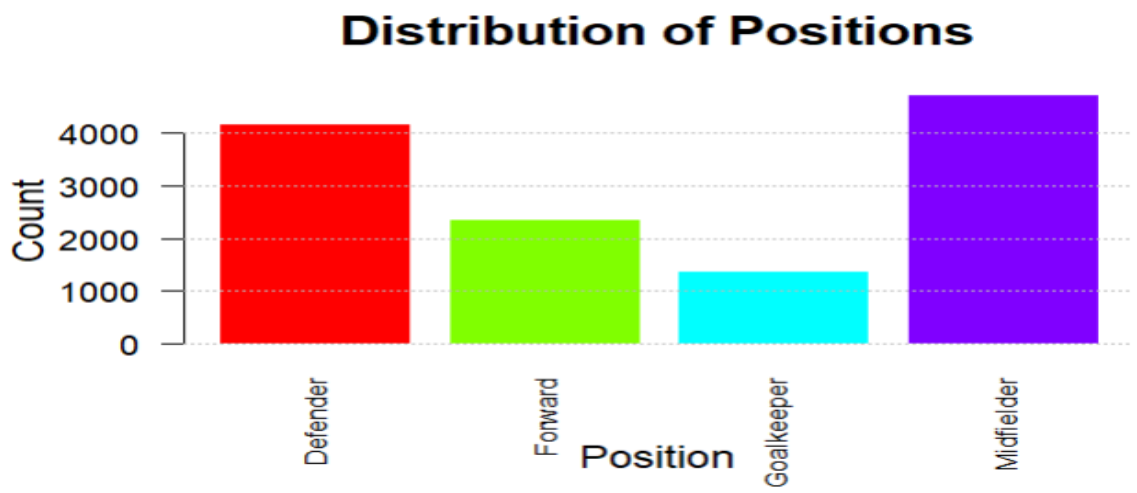
I. Phân loại theo nhóm vị trí thi đấu: (Phân loại đa nhóm)

- Bước 1: Kiểm định sự khác nhau của trung bình overall của các nhóm phân loại bằng A/B testing



Giả thuyết H0: Các nhóm vị trí cầu thủ có trung bình overall như nhau
Đối thuyết H1: Ít nhất có một nhóm có trung bình overall khác với những nhóm còn lại.
Kết quả kiểm định: p_value<2.2e-16<0.05 => Đủ cơ sở để bác bỏ H0 nên ta kết luận có ít nhất 1 nhóm có trung bình overall khác với các nhóm còn lại.

- Bước 2: Kiểm tra sự cân bằng dữ liệu
  Dữ liệu ban đầu không cân bằng nên ta sử dụng SMOTE để cân bằng dữ liệu

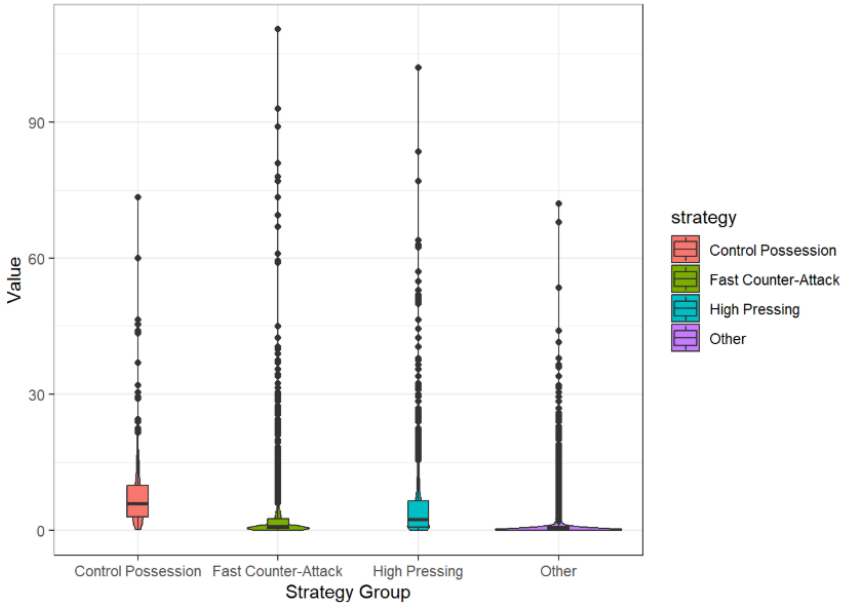


- Bước 3: Sử dụng mô hình phân loại Multinominal logistic ta được bảng kết quả đánh giá mô hình phân loại

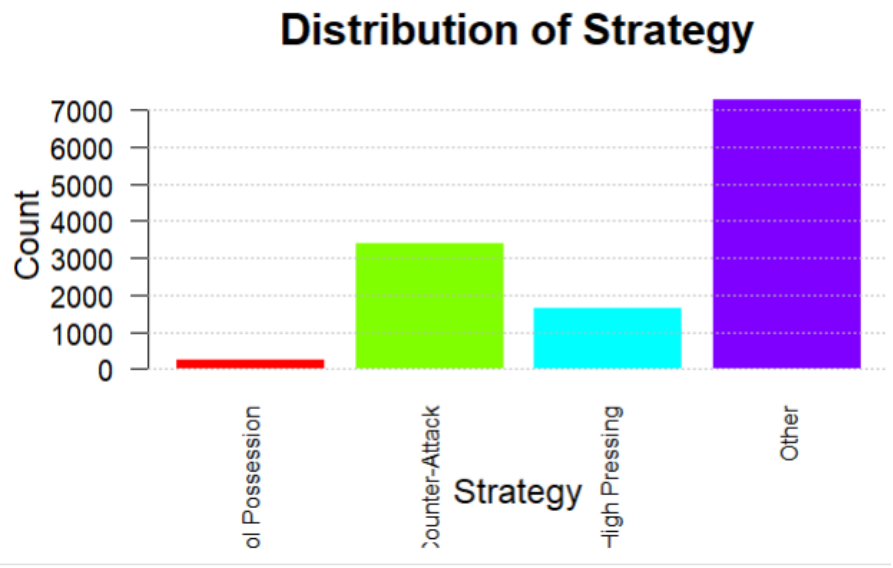
```
## $Precision
##   Defender    Forward Goalkeeper Midfielder
##  0.9128269  0.7554502  1.0000000  0.8506682
##
## $Recall
##   Defender    Forward Goalkeeper Midfielder
##  0.9156777  0.8597627  1.0000000  0.7896440
##
## $Accuracy
## [1] 0.8688163
##
## $Kappa
## [1] 0.8164421
##
## $Macro_F1
## [1] 0.6942795
```

II. Phân loại cầu thủ dựa trên khả năng phù hợp với các chiến thuật của đội bóng (Phân loại đa nhóm)

- Bước 1: Kiểm định sự khác nhau của trung bình value của các nhóm phân loại bằng A/B testing
  Kết quả kiểm định p_value<2.2e-16<0.05 =>Đủ cơ sở để bác bỏ H0 nên ta kết luận có ít nhất 1 nhóm có trung bình value khác với các nhóm còn lại



- Bước 2: Kiểm tra sự cân bằng dữ liệu
  Dữ liệu ban đầu không cân bằng nên ta sử dụng SMOTE để cân bằng dữ liệu

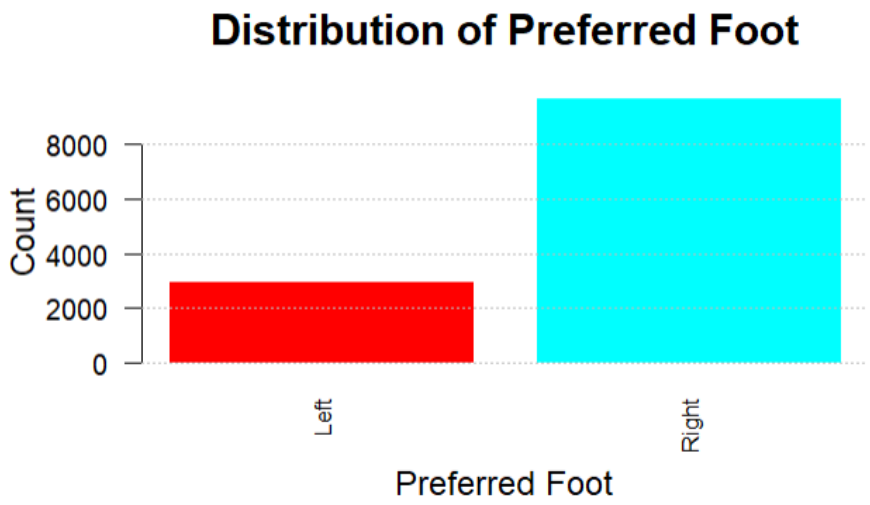


- Bước 3: Sử dụng mô hình phân loại Multinominal logistic ta được bảng kết quả đánh giá mô hình phân loại

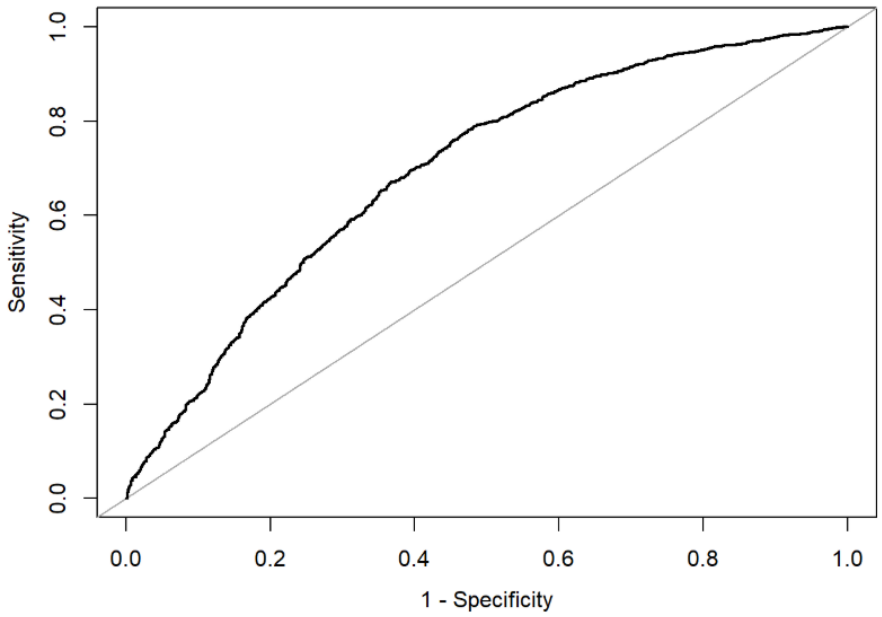
```
## $Precision
##  Control Possession Fast Counter-Attack       High Pressing               Other
##           0.4630631           0.8866964           0.7334711           0.9785607
##
## $Recall
##  Control Possession Fast Counter-Attack       High Pressing               Other
##           0.9483395           0.9149425           0.8781694           0.8765818
##
## $Accuracy
## [1] 0.8884979
##
## $Kappa
## [1] 0.816472
##
## $Macro_F1
## [1] 0.5740917
```

III. Phân loại theo chân thuận (Phân loại hai nhóm)

- Bước 1: Kiểm tra sự cân bằng dữ liệu
  Dữ liệu ban đầu không cân bằng ta dùng under_sampling để cân bằng dữ liệu



- Bước 2: Xây dựng mô hình phân loại logistic
  Kết quả đánh giá mô hình:
  + Đường cong ROC nằm trên bên trái đường tham chiếu nhưng không quá cao nhưng mô hình vẫn có khả năng dự đoán phân loại có tính chính xác chấp nhận được
  + Giá trị của AUC ước lượng của mô hình là 0.6957 ,95% CI: 0.6776-0.7138 , do đó mô hình Logistic có độ chính xác chấp nhận được để dự đoán chân thuận của cầu thủ
  Đồ thị đường cong ROC:



- Bước 3:Xét ngưỡng threshold mặc định là 0.5
  Ta được bảng kết quả:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  750 1483
         1  375 2385

               Accuracy : 0.6279
                 95% CI : (0.6143, 0.6413)
    No Information Rate : 0.7747
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2099

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6667
            Specificity : 0.6166
         Pos Pred Value : 0.3359
         Neg Pred Value : 0.8641
             Prevalence : 0.2253
         Detection Rate : 0.1502
   Detection Prevalence : 0.4472
      Balanced Accuracy : 0.6416

       'Positive' Class : 0
```

- Bước 4: Tối ưu threshold
  a. Phương pháp Youden index (tối ưu hóa sự kết hợp giữa sensitivity và specificity)
  Threshold=0.4304604

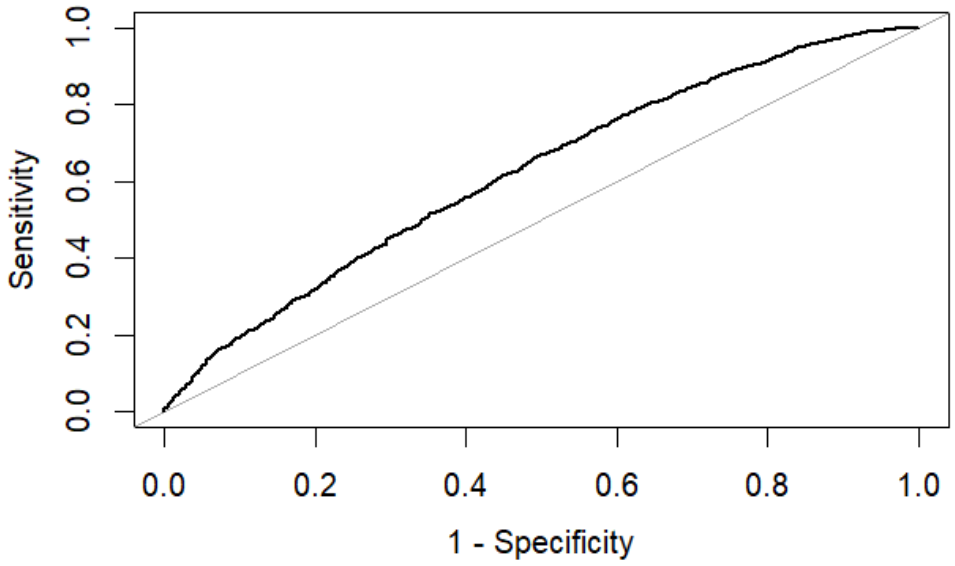
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  652  941
         1  568 3288

               Accuracy : 0.7231
                 95% CI : (0.711, 0.7349)
    No Information Rate : 0.7761
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2813

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5344
            Specificity : 0.7775
         Pos Pred Value : 0.4093
         Neg Pred Value : 0.8527
             Prevalence : 0.2239
         Detection Rate : 0.1197
   Detection Prevalence : 0.2923
      Balanced Accuracy : 0.6560

       'Positive' Class : 0
```

b. Phương pháp Closest top left (tìm ngưỡng sao cho điểm trên ROC curve gần nhất với điểm lý tưởng (1, 1) - là điểm có sensitivity và specificity tối đa)
Threshold=0.4792016

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  713 1276
         1  412 2592

               Accuracy : 0.6619
                 95% CI : (0.6486, 0.6751)
    No Information Rate : 0.7747
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2388

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6338
            Specificity : 0.6701
         Pos Pred Value : 0.3585
         Neg Pred Value : 0.8628
             Prevalence : 0.2253
         Detection Rate : 0.1428
   Detection Prevalence : 0.3984
      Balanced Accuracy : 0.6519

       'Positive' Class : 0
```

Ngưỡng threshold của phương pháp Youden index đem lại Accuracy tốt hơn của Closest top left

- Bước 5: So sánh với mô hình phân loại Naive Bayes



Xét ngưỡng threshold mặc định là 0.5

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  982 2911
         1  238 1318

               Accuracy : 0.4221
                 95% CI : (0.4089, 0.4353)
    No Information Rate : 0.7761
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0655

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8049
            Specificity : 0.3117
         Pos Pred Value : 0.2522
         Neg Pred Value : 0.8470
             Prevalence : 0.2239
         Detection Rate : 0.1802
   Detection Prevalence : 0.7144
      Balanced Accuracy : 0.5583

       'Positive' Class : 0
```

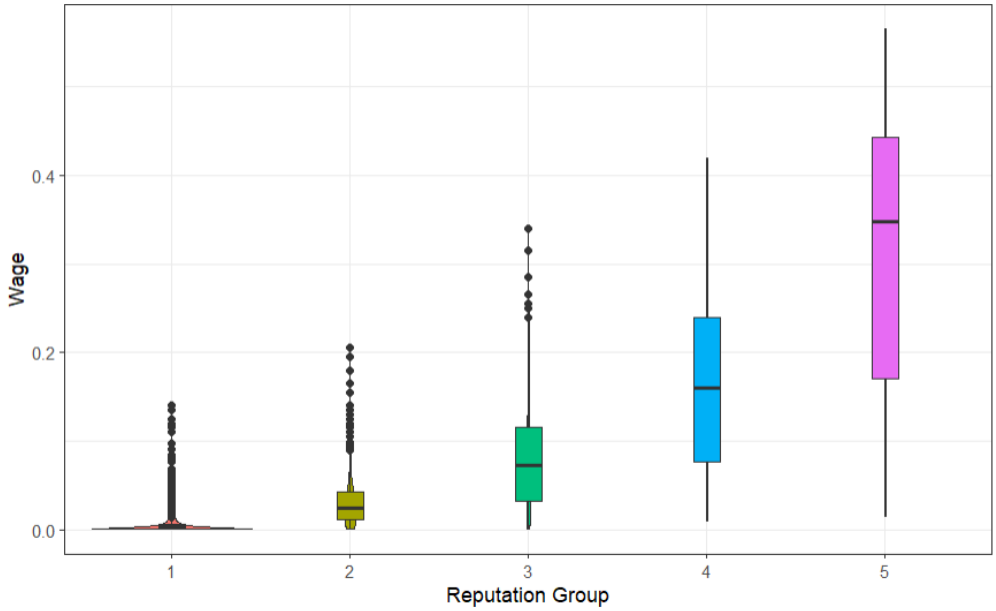
Đường cong ROC nằm trên bên trái đường tham chiếu nhưng không cao bằng mô hình phân loại logistic nhưng mô hình vẫn có khả năng dự đoán phân loại có tính chính xác chấp nhận được.

Giá trị của AUC ước lượng của mô hình là 0.6267 , 95% CI: 0.6079-0.6454, do đó mô hình Naive Bayes có độ chính xác chấp nhận được để dự đoán chân thuận của cầu thủ nhưng kém hơn là mô hình logistic.

IV. Phân loại dựa trên danh tiếng quốc tế (đa nhóm)

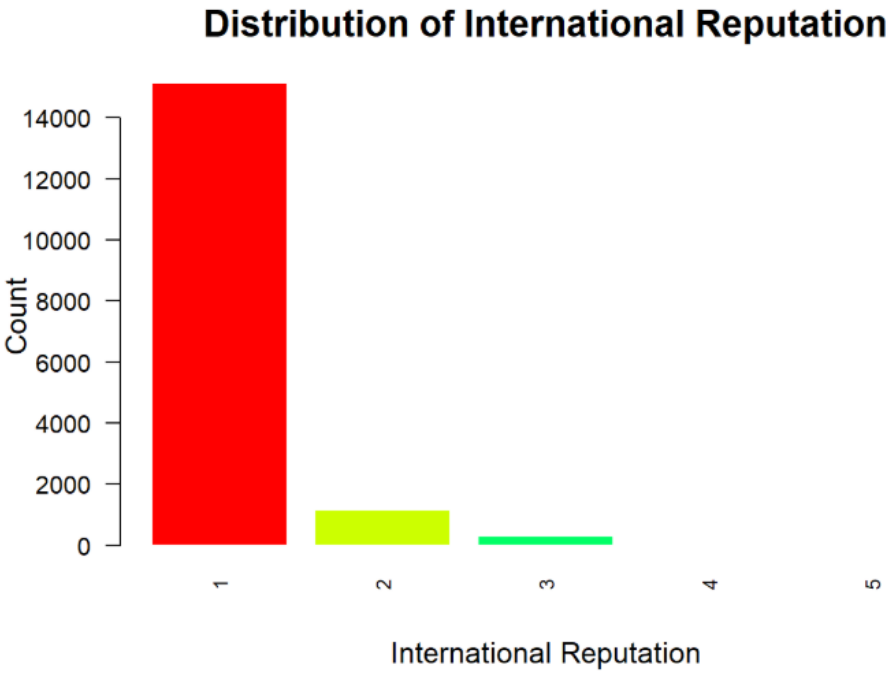
- Bước 1: Kiểm định sự khác nhau của trung bình wage của các nhóm phân loại bằng A/B testing
  Giả thuyết H0:các nhóm cầu thủ có độ nổi tiếng khác nhau thì có trung bình wage như nhau.

  Đối thuyết H1: Ít nhất có một nhóm có trung bình wage khác với những nhóm còn lại.



Kết quả kiểm định: p_value<2.2e-16<0.05 =>Đủ cơ sở để bác bỏ H0 nên ta kết luận có ít nhất 1 nhóm có trung bình wage khác với các nhóm còn lại.

- Bước 2: Kiểm tra sự cân bằng dữ liệu
  + Dữ liệu ban đầu không cân bằng nên ta sử dụng SMOTE để cân bằng dữ liệu.



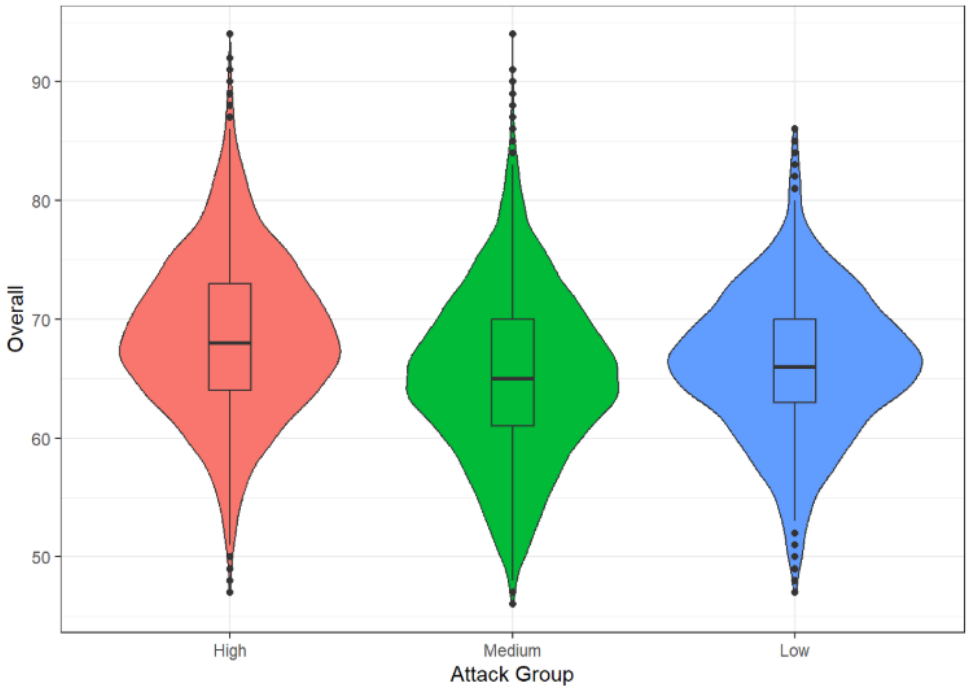
- Bước 3:Sử dụng mô hình phân loại Multinominal logistic ta được bảng kết quả đánh giá mô hình phân loại

```
## $Precision
##         1         2         3         4         5
## 0.9242637 0.7644986 0.7185629 0.8533204 1.0000000
##
## $Recall
##         1         2         3         4         5
## 0.8857527 0.7739215 0.7164179 0.8802793 1.0000000
##
## $Accuracy
## [1] 0.8512648
##
## $Kappa
## [1] 0.8140668
##
## $Macro_F1
## [1] 0.6178205
```

- Nhận xét
  + Mô hình có khả năng phân loại tốt các nhóm 1,4,5,Nhưng không hoạt động tốt với nhóm 2 và 3.
  + Các chỉ số như Kappa khá tốt nhưng Macro_F1 vẫn khá thấp.

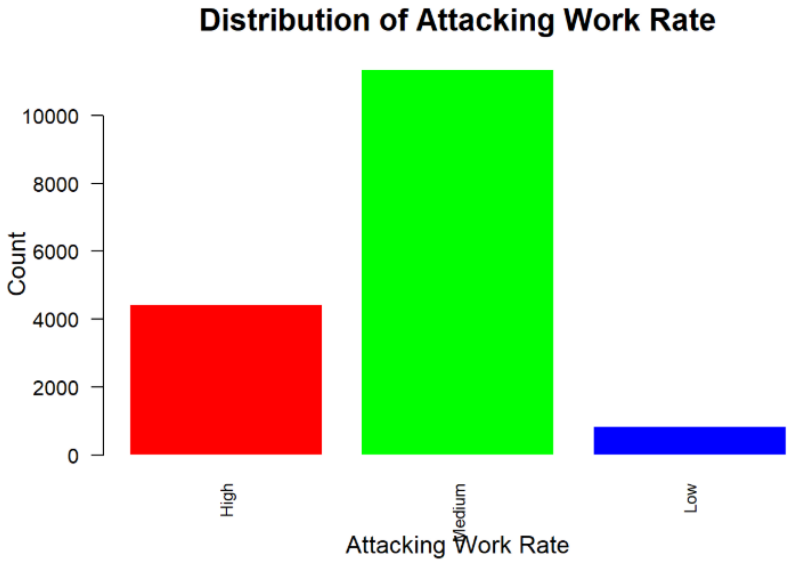
V. Phân loại dựa trên xu hướng di chuyển tấn công khi không có bóng

- Bước 1:Kiểm định sự khác nhau của trung bình overall của các nhóm phân loại bằng A/B testing.
  Giả thuyết H0:các nhóm cầu thủ có cùng xu hướng di chuyển thì có trung bình overall như nhau.
  Đối thuyết H1: Ít nhất có một nhóm có trung bình overall khác với những nhóm còn lại.



p-value<0.05 nên ta bác bỏ giả thuyết H0 và chấp nhận giả thuyết H1. Có ít nhất một nhóm có trung bình overall khác với các nhóm còn lại.

- Bước 2:Kiểm tra sự cân bằng dữ liệu:
  + Dữ liệu ban đầu không cân bằng nên ta sử dụng SMOTE để cân bằng dữ liệu.



- Bước 3:Sử dụng mô hình phân loại Multinominal logistic ta được bảng kết quả đánh giá mô hình phân loại

```
## $Precision
##      High    Medium       Low
## 0.7023718 0.6553525 0.7468165
##
## $Recall
##      High    Medium       Low
## 0.8134692 0.4397722 0.8760984
##
## $Accuracy
## [1] 0.7092136
##
## $Kappa
## [1] 0.5639848
##
## $Macro_F1
## [1] 0.3513438
```

Kết quả không thực sự tốt. Áp dụng một vài mô hình khác

- LDA

```
## $Precision
##      High    Medium       Low
## 0.8099247 0.4581691 0.8119508
##
## $Recall
##      High    Medium       Low
## 0.6905931 0.6025346 0.7595561
##
## $Accuracy
## [1] 0.6927817
##
## $Kappa
## [1] 0.5393343
##
## $Macro_F1
## [1] 0.3267524
```

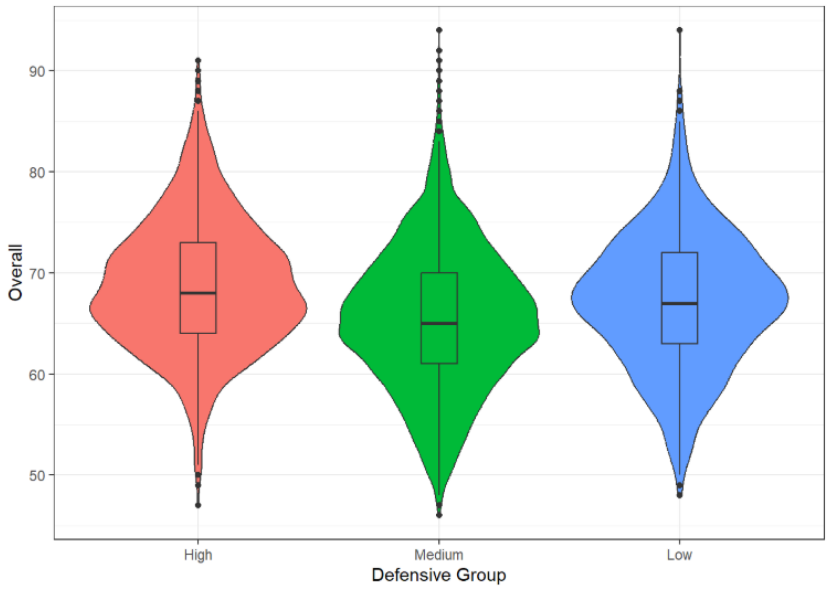
- QDA

```
## $Precision
##      High    Medium       Low
## 0.7682765 0.4007884 0.8985062
##
## $Recall
##      High    Medium       Low
## 0.6845638 0.6931818 0.6901789
##
## $Accuracy
## [1] 0.6886737
##
## $Kappa
## [1] 0.5331857
##
## $Macro_F1
## [1] 0.3274379
```

Kết quả của 3 mô hình có thể xem là tương đương. Đều không hoạt động tốt. Biến dự đoán attacking_work_rate không thể phân loại được dựa trên các biến giải thích

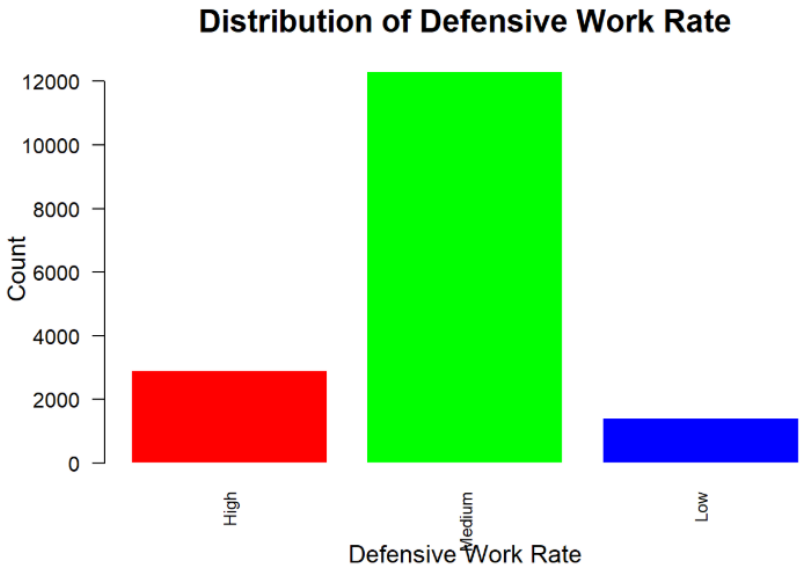
VI. Phân loại dựa trên xu hướng di chuyển phòng thủ khi không có bóng

- Bước 1: Kiểm định sự khác nhau của trung bình overall của các nhóm phân loại bằng A/B testing. Giả thuyết H0: Các nhóm cầu thủ có cùng xu hướng di chuyển thì có trung bình overall như nhau. Đối thuyết H1: Ít nhất có một nhóm có trung bình overall khác với những nhóm còn lại.



p-value<0.05 nên ta bác bỏ giả thuyết H0 và chấp nhận giả thuyết H1. Có ít nhất một nhóm có trung bình overall khác với các nhóm còn lại.

- Bước 2: Kiểm tra sự cân bằng dữ liệu:
  Dữ liệu ban đầu không cân bằng nên ta sử dụng SMOTE để cân bằng dữ liệu.



- Bước 3:Sử dụng mô hình phân loại Multinominal logistic ta được bảng kết quả đánh giá mô hình phân loại

```
## $Precision
##      High    Medium       Low
## 0.7035992 0.6401674 0.7187933
##
## $Recall
##      High    Medium       Low
## 0.7856580 0.4355429 0.8659148
##
## $Accuracy
## [1] 0.6951698
##
## $Kappa
## [1] 0.5427304
##
## $Macro_F1
## [1] 0.3307945
```

Kết quả không thật sự tốt. Áp dụng các mô hình khác.

- LDA

```
## $Precision
##      High    Medium       Low
## 0.7955083 0.4030094 0.8734336
##
## $Recall
##      High    Medium       Low
## 0.6923868 0.6494102 0.7090539
##
## $Accuracy
## [1] 0.6901637
##
## $Kappa
## [1] 0.5351921
##
## $Macro_F1
## [1] 0.324415
```

- QDA

```
## $Precision
##      High    Medium       Low
## 0.7895981 0.4005693 0.8947368
##
## $Recall
##      High    Medium       Low
## 0.6818646 0.7307122 0.6900773
##
## $Accuracy
## [1] 0.6942227
##
## $Kappa
## [1] 0.5413853
##
## $Macro_F1
## [1] 0.3399486
```

Kết quả của 3 mô hình có thể xem là tương đương. Đều không hoạt động tốt. Biến dự đoán defensive_work_rate không thể phân loại được dựa trên các biến giải thích.

VII. Phân loại dựa trên danh tiếng quốc tế (nhị phân)

- Biến international_reputation được đo lường thông qua 5 level. Nhưng số lượng dữ liệu được phân loại ở mức 1 và 2 là rất nhiều.Nên ta thử áp dụng mô hình phân loại cho 2 biến đó.

```
##
##     1     2
## 15141  1154
```

- Bước 1: Có sự chênh lệch giữa số lượng phần tử của 2 nhóm. Áp dúng SMOTE
- Bước 2: Áp dụng mô hình hồi quy logictis.
  Áp dụng với ngưỡng 0.5

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    2
##          1 2702  193
##          2  322 2840
##
##                Accuracy : 0.915
##                  95% CI : (0.9077, 0.9219)
##     No Information Rate : 0.5007
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8299
```

```
##  Mcnemar's Test P-Value : 1.697e-08
##
##             Sensitivity : 0.8935
##             Specificity : 0.9364
##          Pos Pred Value : 0.9333
##          Neg Pred Value : 0.8982
##              Prevalence : 0.4993
##          Detection Rate : 0.4461
##    Detection Prevalence : 0.4780
##       Balanced Accuracy : 0.9149
##
##        'Positive' Class : 1
##
```

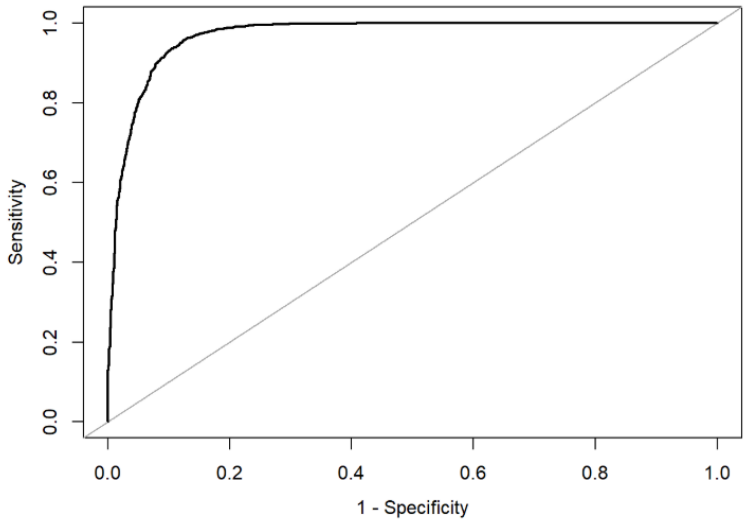
- Bước 3: Chúng ta sẽ sử dụng phân tích đường cong ROC và diện tích dưới đường cong (AUC)
  Tính toán AUC:

```
## Area under the curve: 0.9686
```

Khoảng tin cậy 95% cho AUC:

```
## 95% CI: 0.9645-0.9726 (DeLong)
```

Đường cong ROC ước lượng:



- Cải thiện mô hình: Tìm ra ngưỡng tối ưu.
  + Phương pháp Youden index và phương pháp closest top left

```
out_youd <- coords(out_roc, "best", ret = c("threshold", "specificity", "sensitivity"),
best.method = "youden")
print(out_youd)
```

```
##   threshold specificity sensitivity
## 1 0.4024343   0.8716931   0.9610946
```

```
out_clost <- coords(out_roc, "best", ret = c("threshold", "specificity", "sensitivity"),
best.method = "closest.topleft")
print(out_clost)
```

```
##   threshold specificity sensitivity
## 1  0.528847    0.901455   0.9287834
```

+ Sử dụng threshold của phương pháp closest top left.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    2
##          1 2726  216
##          2  298 2817
##
##                Accuracy : 0.9151
##                  95% CI : (0.9078, 0.922)
##     No Information Rate : 0.5007
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8303
##
##  Mcnemar's Test P-Value : 0.0003532
##
##             Sensitivity : 0.9015
##             Specificity : 0.9288
##          Pos Pred Value : 0.9266
##          Neg Pred Value : 0.9043
##              Prevalence : 0.4993
##          Detection Rate : 0.4501
##    Detection Prevalence : 0.4857
##       Balanced Accuracy : 0.9151
##
##        'Positive' Class : 1
##
```

Nhận xét: Kết quả của mô hình này khá tốt chỉ số Accuracy 0,9151/1 là ổn.
Kappa đang ở mức 0.8303.

D. Tổng quan và hiệu quả phân loại, nhận xét

- Các mô hình phân loại, trong đó mô hình phân loại Logistic đã hoạt động tốt trong việc phân loại hai nhóm cầu thủ theo chân thuận, danh tiếng quốc tế mang lại kết quả chấp nhận được và hợp lý, tuy nhiên cần cải thiện thêm bằng cách xem xét kĩ lưỡng lại các chỉ số và tạo ra nhiều chỉ số mới có độ quan trọng lớn hơn, đạt được độ chính xác cao hơn khi tối ưu hóa với các phương pháp kiểm tra như A/B Testing, cân bằng dữ liệu và tìm threshold hợp lý.
- Các mô hình phân loại đa nhóm cho ra kết quả tốt đến rất tốt phù hợp để sử dụng cho việc dự đoán phân loại.
- Việc áp dụng các chỉ số như Precision, Recall, AUC, Macro_F1,... giúp mô hình có khả năng đo lường hiệu suất một cách chi tiết và chính xác, giúp đánh giá mô hình trên nhiều phương diện khác nhau.
- Kết quả của việc phân loại dựa trên xu hướng di chuyển tấn công và phòng thủ khi không có bóng không tốt đối với các mô hình phân loại đã áp dụng.

III. Tổng kết

Bộ dữ liệu sports_data_analysis cho một cái nhìn đa dạng về bóng đá trên thế giới, với các phân tích dữ liệu phía trên mang lại nhiều khám phá thú vị

- EDA tổng hợp lại những thông tin hữu ích về bộ dữ liệu, cho thấy các số liệu quan trọng, tìm ra những vấn đề trong tập dữ liệu (imbalance data, outliers,…)
- A/B testing rút ra những giả thuyết cần thiết cho những huấn luyện viên, người chiêu mộ cầu thủ có thêm cập nhập về tình trạng hiện nay của các cầu thủ qua các đặc điểm về tuổi, vị trí thi đấu, vùng,…
- Mô hình hồi quy giúp dự đoán tiền lương, năng lực, tiềm năng, giá trị của một cầu thủ, giúp các nhà tuyển dụng dự đoán và chọn được các cầu thủ phù hợp
- Mô hình phân loại giúp dự đoán được vị trí thi đấu phù hợp, khả năng phù hợp, chân thuận,…giúp các huấn luyện viên đưa ra những chiến lược phù hợp ở các trận đấu khác nhau

Qua các phân tích, đánh giá xây dựng ở trên sẽ giúp cho ban quản lý của câu lạc bộ đưa ra các quyết định mua sắm cầu thủ hợp lý dựa trên ngân sách của câu lạc bộ.

Hết.