TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN - ĐHQG TP. HỒ CHÍ MINH
KHOA TOÁN - TIN HỌC

—————oOo—————

Báo cáo môn Xử lý số liệu thống kê

GVHD: Tô Đức Khánh

Sinh viên thực hiện:

Họ và tên	*MSSV*
Dương Thị Ngọc Tuyền	**22110254**

TP. HỒ CHÍ MINH 2025

Mục lục

Project 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.

1. Bản đề xuất phân tích và xử lý số liệu

Nhóm quyết định chọn project [**Project 1- Sports Data Analysis**] để phân tích và xử lý số liệu. Trong quá trình thực hiện, nhóm sẽ áp dụng các phương pháp đã học trong học phần như:

- Làm sạch dữ liệu: loại bỏ giá trị bị thiếu, giá trị bất thường, giá trị không cần thiết.
- Mô tả và trực quan hóa dữ liệu bằng các biểu đồ.
- Sử dụng các thống kê để trả lời được các câu hỏi quan trọng của dữ liệu, trích xuất đặc trưng của dữ liệu để phục vụ cho việc xây dựng mô hình.
- Xây dựng và sử dụng mô hình dự đoán vị trí cầu thủ dựa trên các chỉ số.
- Đưa ra các lựa chọn tối ưu cho các huấn luyện viên, hỗ trợ huấn luyện viên trong việc xây dựng đội hình.

2. Các mục tiêu phân tích cần đạt được

- **Mục tiêu 1:** Nắm được các đặc trưng của dữ liệu, mối quan hệ giữa các thuộc tính, trực quan hóa được dữ liệu
- **Mục tiêu 2:** Xây dựng được mô hình có độ chính xác cao để giúp huấn luyện viên trong việc tìm ra vị trí tiềm năng của cầu thủ
- **Mục tiêu 3:** Xây dựng được mô hình hồi quy tuyến tính xác định các thông số ảnh hưởng đến giá trị cầu thủ
- **Mục tiêu 4:** Giúp huấn luyện viên xây dựng đội hình dựa trên ngân sách của đội

3. Quy trình thực hiện

3.1 Chuẩn bị và cấu hình môi trường làm việc

- Sử dụng RStudio và tạo file R Markdown mới với thiết lập đầu ra là HTML.
- Cài đặt các thư viện cần thiết

  ggplot2, janitor, tidyverse, leaps, readr, dplyr, stringr, ggcorrplot, VIM, maps , viridis, tidyr, caret

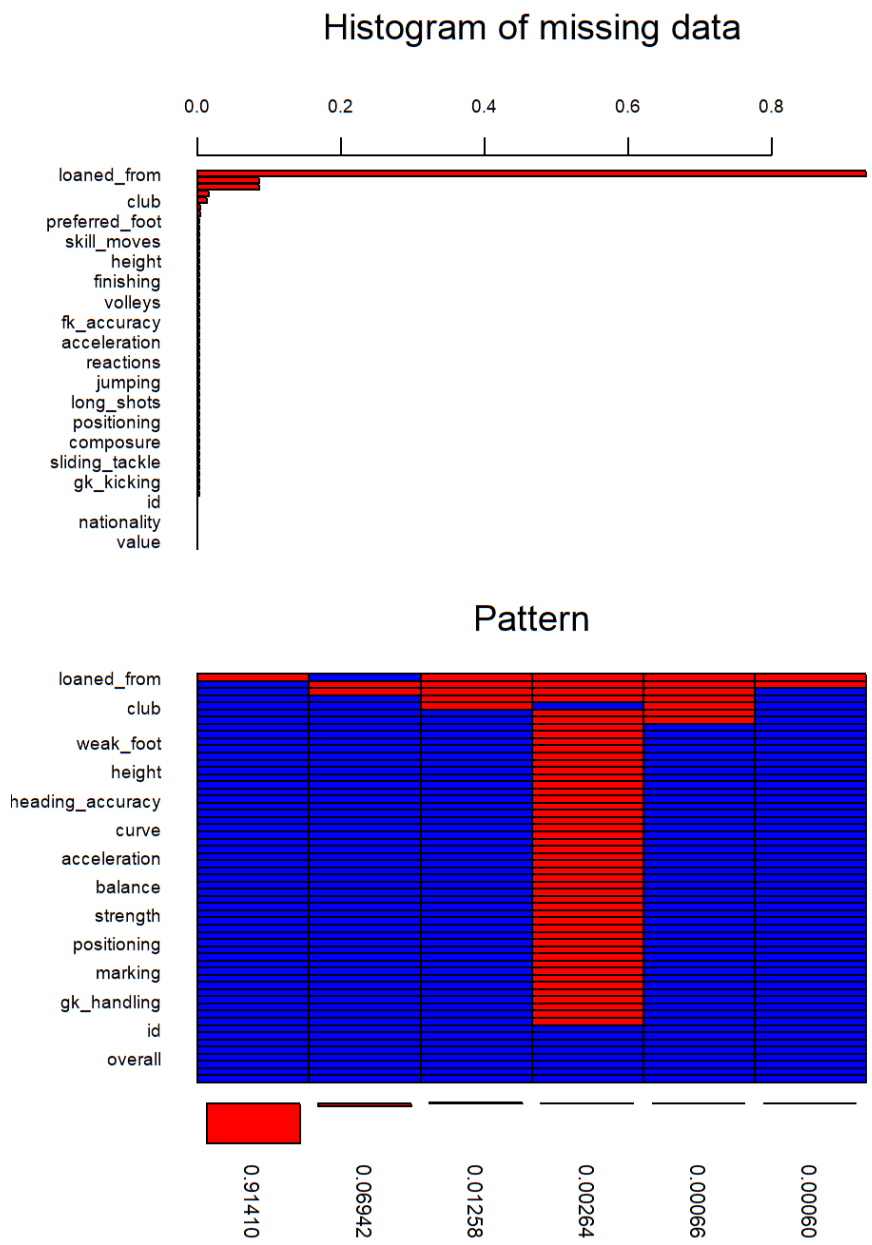
3.2 Xử lý dữ liệu ban đầu

1: Tải và Làm Sạch Dữ Liệu

Đầu tiên, dữ liệu được tải vào và kiểm tra các thông tin tổng quan về bộ dữ liệu. Các cột được làm sạch tên để dễ dàng xử lý trong quá trình phân tích.

2: Phân Tích Dữ Liệu Thiếu

Các giá trị thiếu trong bộ dữ liệu được kiểm tra bằng cách sử dụng các hàm thống kê. Một biểu đồ histogram được sử dụng để trực quan hóa phân phối các giá trị thiếu. Sau đó, chúng ta kiểm tra từng cột để xác định số lượng giá trị thiếu.



Hình 1: Missing data histogram

3: Xử Lý Dữ Liệu Thiếu

Các giá trị thiếu trong một số cột được xử lý bằng cách thay thế chúng bằng giá trị hợp lý hoặc loại bỏ các dòng chứa dữ liệu thiếu không thể khôi phục.

4: Chuyển Đổi Các Giá Trị Tiền Tệ

Các giá trị tiền tệ như lương, giá trị cầu thủ và phí giải phóng hợp đồng được chuyển đổi từ chuỗi ký tự (bao gồm các đơn vị như “K” và “M”) sang dạng số nguyên tương ứng. Cụ thể:

- Nếu giá trị có ký tự "K", giá trị được nhân với 1000.
- Nếu giá trị có ký tự "M", giá trị được nhân với 1,000,000.
- Nếu không có ký tự đặc biệt, giá trị sẽ được giữ nguyên dưới dạng số.

5: Xử Lý Các Cột Cân Nặng và Chiều Cao

- Cột ‘weight‘ được chuyển từ đơn vị pounds (lbs) sang kilograms (kg) bằng cách nhân với hệ số chuyển đổi 0.453592.
- Tương tự, chiều cao của cầu thủ được chuyển từ feet và inch sang cm.
- Các phép tính này được thực hiện bằng cách tách từng phần chiều cao, sau đó tính toán chiều cao tổng cộng bằng cách sử dụng công thức:

$$\text{Chiều cao (cm)} = \text{Feet} \times 30.48 + \text{Inch} \times 2.54$$

6: Xử Lý Các Cột Không Cần Thiết

Các cột không có ảnh hưởng lớn đến phân tích, như ‘$jersey\_number$‘ và ‘$loaned\_\ from$‘, được loại bỏ để giảm bớt sự phức tạp của bộ dữ liệu.

7: Tính Trung Bình Các Chỉ Số Cho Từng Vị Trí Cầu Thủ

Để hiểu rõ hơn về sự phân bổ các chỉ số trong từng vị trí, ta tính toán giá trị trung bình của các chỉ số cho từng nhóm vị trí (thủ môn, hậu vệ, tiền vệ, tiền đạo).

- Đối với mỗi vị trí, các cột dữ liệu liên quan (không bao gồm cột như cân nặng và chiều cao) được lấy trung bình.
- Sau khi tính toán trung bình cho mỗi nhóm, kết quả được lưu trữ vào một bảng tổng hợp với các vị trí như "Goalkeeper", "Defender", "Midfielder", "Forward".

8: Xử Lý Cột Work Rate

- Cột $work\_rate$ được tách thành hai cột riêng biệt: $work\_rate\_attack$ và $work\_rate\_defense$.

Hai cột này lần lượt biểu thị mức độ hoạt động của cầu thủ trong các tình huống tấn công và phòng thủ.

- Cụ thể, giá trị trong cột *work_rate* được phân tách dựa trên ký tự phân cách (thường là dấu cách hoặc dấu gạch). Sau khi tách, các giá trị này được chuyển đổi thành dạng tiêu chuẩn để dễ dàng phân tích.
- Sau khi tách, chúng ta thực hiện phân tích mối quan hệ giữa các giá trị trong *work_rate_attack* và *work_rate_defense* với từng nhóm vị trí cầu thủ (như thủ môn, hậu vệ, tiền vệ, tiền đạo) và có thể thấy biến *work_rate* thể hiện lối chơi riêng của từng cầu thủ chứ không hẳn ảnh hưởng đến vị trí. Do vậy, ta sẽ thống kê lối chơi của một số cầu thủ nổi bật.

```
## # A tibble: 10 x 4
##    name              overall work_rate_attack work_rate_defense
##    <chr>               <dbl> <chr>            <chr>
##  1 L. Messi               94 Medium           " Medium"
##  2 Cristiano Ronaldo      94 High             " Low"
##  3 Neymar Jr              92 High             " Medium"
##  4 De Gea                 91 Medium           " Medium"
##  5 K. De Bruyne           91 High             " High"
##  6 E. Hazard              91 High             " Medium"
##  7 L. Modric              91 High             " High"
##  8 L. Suárez              91 High             " Medium"
##  9 Sergio Ramos           91 High             " Medium"
## 10 J. Oblak               90 Medium           " Medium"
```

Hình 2: Một số cầu thủ nổi bật

9: Xử Lý Cột Release Clause Và Các Cột Liên Quan Đến Thời Gian

- Các cột như '*contract_valid_until*' và '*joined*' chứa thông tin về thời gian hợp đồng và thời gian gia nhập đội. Đối với cột '*contract_valid_until*', do dataset ở năm 2018, ta cũng không có một thời điểm cụ thể để làm mốc xác định các cầu thủ nào sẽ hết hạn hợp đồng trước khi ta mua, tức là ta không xác định được giá trị các cầu thủ do không biết có cần tính thêm *release_clause* (phí phá vỡ hợp đồng) hay không, do đó ta sẽ hiểu rằng các cầu thủ không có dữ liệu ở *release_clause*, tức là không có phí phá vỡ hợp đồng nghĩa là giá trị sẽ là 0 nên ta sẽ điền giá trị 0 vào các cầu thủ không có dữ liệu ở *release_clause*. Còn các cầu thủ có dữ liệu ở *release_clause*, ta sẽ giữ nguyên giá trị của cột này và tính vào giá trị để chuyển nhượng và đạt được 1 cầu thủ.

- Sau đó, do '*contract_valid_until*' và '*joined*' không ảnh hưởng lắm đến việc phân tích nữa nên ta sẽ bỏ 2 cột này.

10: Xử Lý Các Cầu Thủ Tự Do

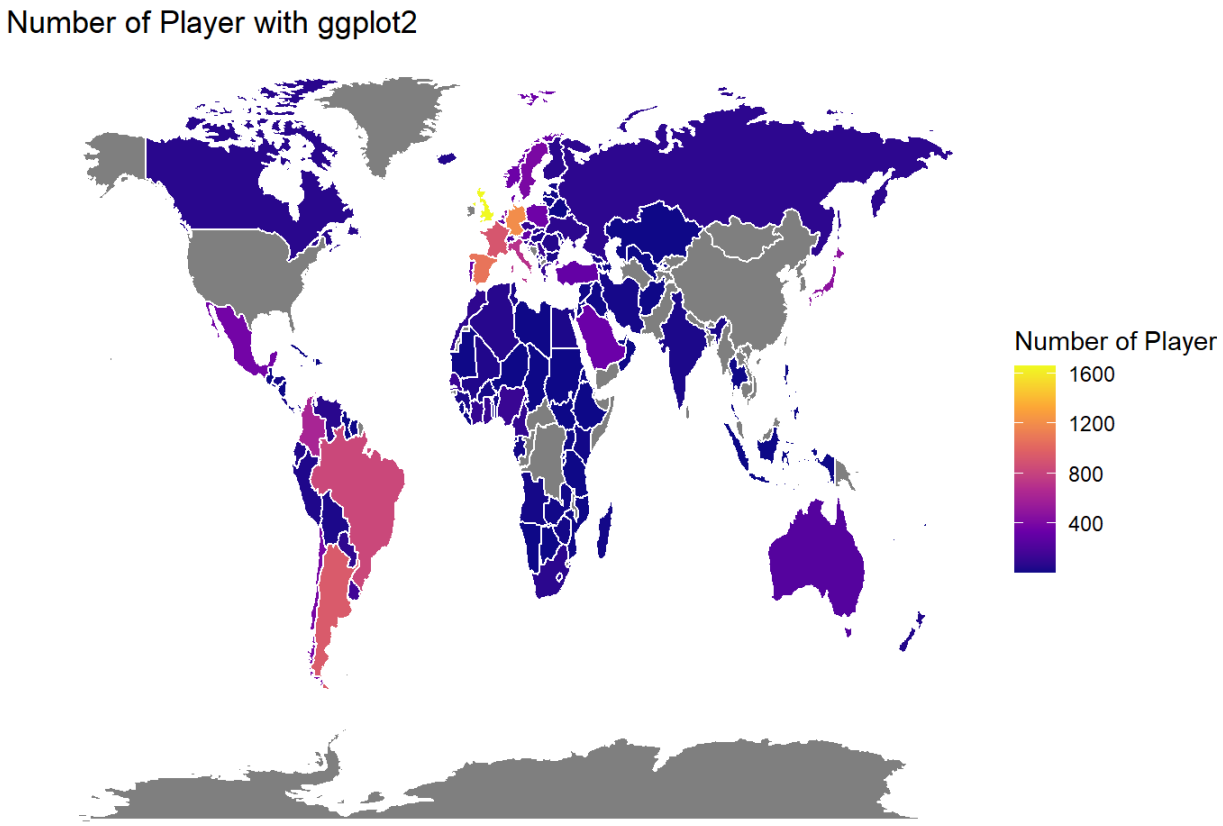
Các cầu thủ không có câu lạc bộ sẽ được gán giá trị "Free Agent" trong cột 'club'. Điều này giúp phân biệt các cầu thủ tự do với các cầu thủ thuộc câu lạc bộ cụ thể.

3.3 Phân tích dữ liệu và xây dựng mô hình

3.3.1 Thống kê và trực quan hóa dữ liệu

Bản đồ phân bố cầu thủ theo quốc gia

Cột `nationality` được sử dụng để thống kê số lượng cầu thủ từ mỗi quốc gia và vẽ bản đồ thế giới. Dữ liệu quốc gia `England` được gộp chung vào `UK`. Sử dụng thư viện `ggplot2`, số lượng cầu thủ được biểu diễn trên bản đồ thế giới theo thang màu `viridis`.



Hình 3: Number of players

Thống kê số liệu cầu thủ

- **Top 10 quốc gia có số lượng cầu thủ cao nhất:** Thực hiện `group_by` theo quốc gia, sau đó sắp xếp giảm dần số lượng cầu thủ và chọn 10 quốc gia đứng đầu.

```
##          region Number of Player
## 1            UK              1657
## 2       Germany              1195
## 3         Spain              1071
## 4     Argentina               936
## 5        France               911
## 6        Brazil               825
## 7         Italy               699
## 8      Colombia               616
## 9         Japan               478
## 10 Netherlands               452
```

Hình 4: Top 10 quốc gia có số lượng cầu thủ cao nhất

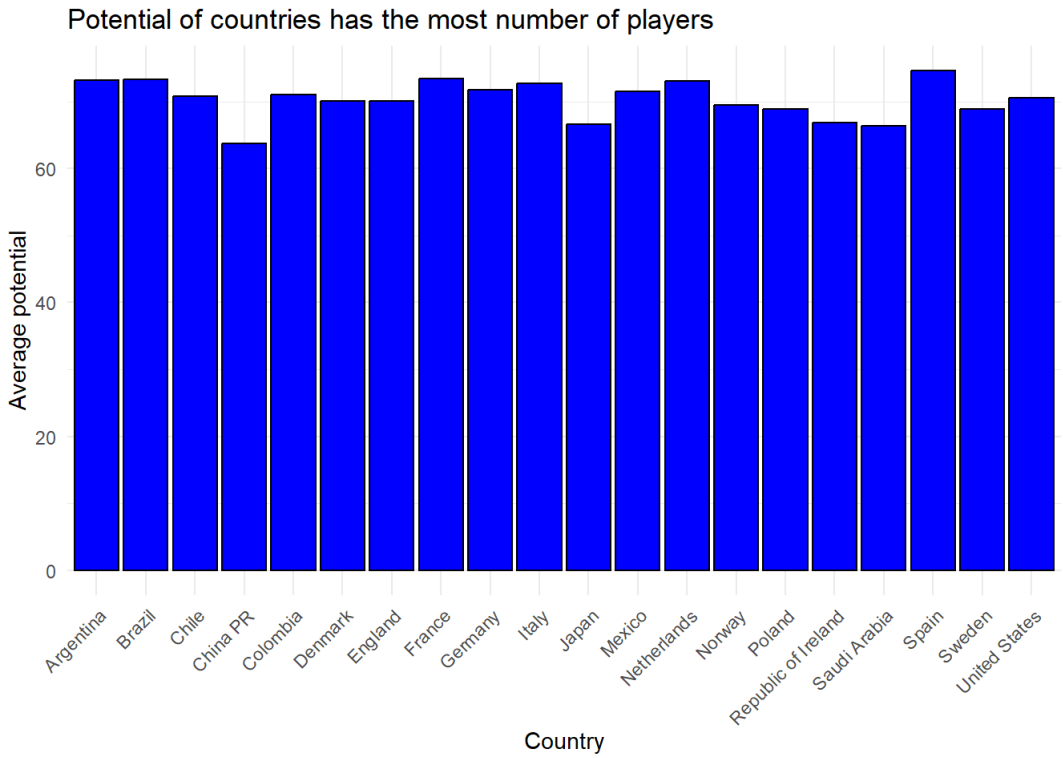
- **Top 20 quốc gia có tiềm năng trung bình cao nhất:** Tính tiềm năng trung bình (`avg_potential`) của cầu thủ từng quốc gia và chọn ra 20 quốc gia đầu bảng.

```
## # A tibble: 20 x 3
##    nationality          avg_potential numofplayers
##    <chr>                        <dbl>        <dbl>
##  1 Dominican Republic            80.5            2
##  2 Chad                          78              2
##  3 United Arab Emirates          78              1
##  4 Central African Rep.          76              3
##  5 Russia                        75.3           79
##  6 Portugal                      75.3          322
##  7 Equatorial Guinea             75.2            5
##  8 Ukraine                       75.1           73
##  9 Spain                         74.6         1071
## 10 Croatia                       74.5          126
## 11 Belgium                       74.4          259
## 12 Greece                        74.4          102
## 13 Ivory Coast                   74.1          100
## 14 Indonesia                     74              1
## 15 Tanzania                      74              3
## 16 Gambia                        73.9           15
## 17 Nigeria                       73.9          121
## 18 Georgia                       73.8           26
## 19 Zambia                        73.8            9
## 20 Mozambique                    73.8            4
```

Hình 5: Top 20 quốc gia có tiềm năng trung bình cao nhất

Dựa vào bảng trên, ta có thể thấy các quốc gia có trung bình tiềm năng cao nhất là do quốc gia có ít cầu thủ nên trung bình sẽ cao, do đó ta sẽ tính điểm tiềm năng trung bình cho 20 quốc gia có nhiều cầu thủ nhất

- **Potential trung bình của top 20 quốc gia có số lượng cầu thủ nhiều nhất:** Xếp hạng các quốc gia dựa trên số lượng cầu thủ, sau đó trực quan hóa tiềm năng trung bình (`avg_potential`) của họ qua biểu đồ cột.



Hình 6: Potential of countries has the most number of players

Thống kê cầu thủ và câu lạc bộ

- **Top 10 cầu thủ có chỉ số tổng thể (`overall`) cao nhất.**

```
## # A tibble: 10 x 2
##    name              overall
##    <chr>               <dbl>
##  1 L. Messi               94
##  2 Cristiano Ronaldo      94
##  3 Neymar Jr              92
##  4 De Gea                 91
##  5 K. De Bruyne           91
##  6 E. Hazard              91
##  7 L. Modric              91
##  8 L. Suárez              91
##  9 Sergio Ramos           91
## 10 J. Oblak               90
```

Hình 7: Top 10 cầu thủ có chỉ số tổng thể cao nhất

- **Top 10 cầu thủ có giá trị thị trường (`value`) cao nhất.**

```
## # A tibble: 10 x 2
##    name              overall
##    <chr>               <dbl>
##  1 L. Messi               94
##  2 Cristiano Ronaldo      94
##  3 Neymar Jr              92
##  4 De Gea                 91
##  5 K. De Bruyne           91
##  6 E. Hazard              91
##  7 L. Modric              91
##  8 L. Suárez              91
##  9 Sergio Ramos           91
## 10 J. Oblak               90
```

Hình 8: Top 10 cầu thủ có giá trị thị trường cao nhất

- **Top 10 câu lạc bộ có tổng giá trị thị trường cao nhất:**

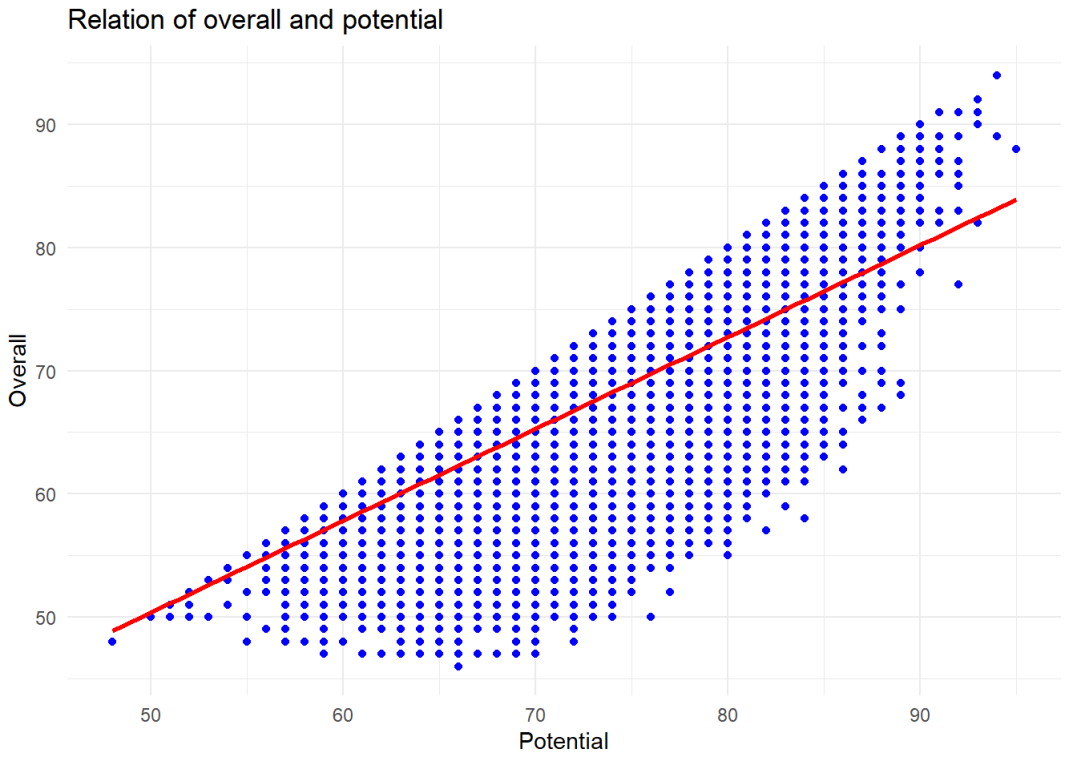
```
## # A tibble: 10 x 2
##    club                total_value
##    <chr>                     <dbl>
##  1 Real Madrid           874425000
##  2 FC Barcelona          852600000
##  3 Manchester City       786555000
##  4 Juventus              704475000
##  5 FC Bayern München     679025000
##  6 Atlético Madrid       644525000
##  7 Paris Saint-Germain   625325000
##  8 Tottenham Hotspur     618450000
##  9 Chelsea               606815000
## 10 Manchester United     588850000
```

Hình 9: Top 10 câu lạc bộ có tổng giá trị thị trường cao nhất

- **Tổng hợp giá trị cầu thủ (`value`) theo từng câu lạc bộ.**

Phân tích mối tương quan

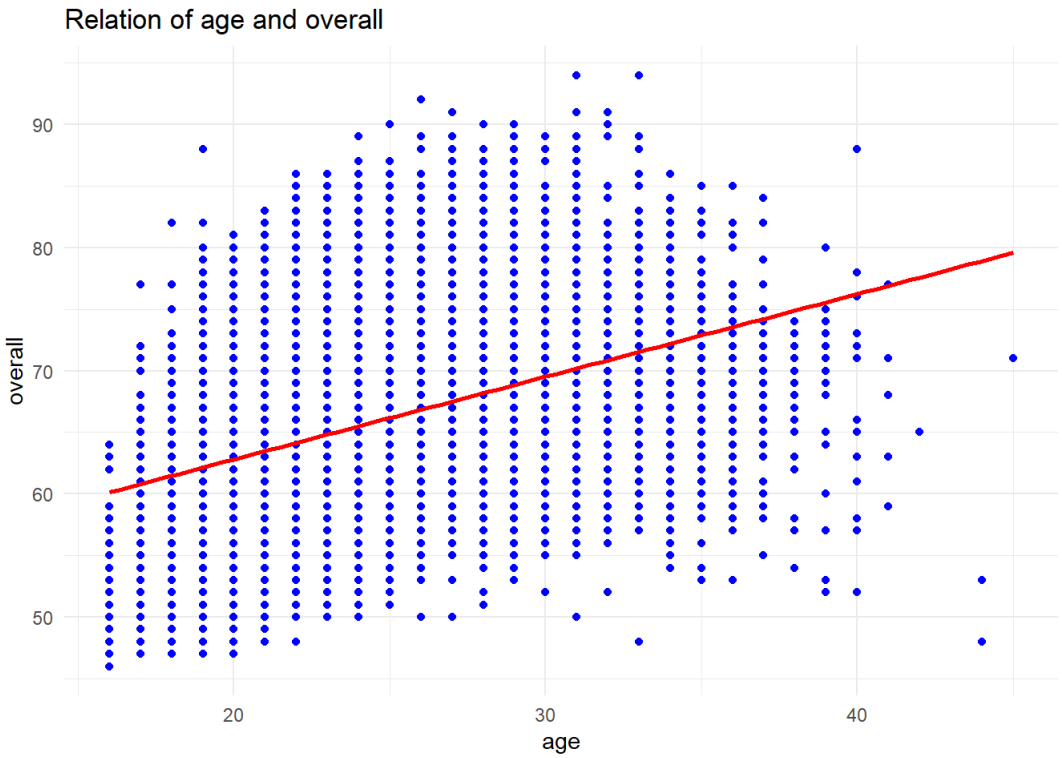
- **Tương quan giữa `potential` và `overall`:** Sử dụng biểu đồ phân tán và tính hệ số tương quan Pearson để phân tích.



Hình 10: Relation of overall and potential

⇒ Ta thấy rằng potential và overall có tương quan mạnh, có thể nói rằng potential cao thì overall cũng cao, nhưng không phải lúc nào cũng đúng, vì có thể cầu thủ có potential cao nhưng không được đánh giá cao về overall, có thể do cầu thủ đó chưa phát huy hết khả năng của mình.

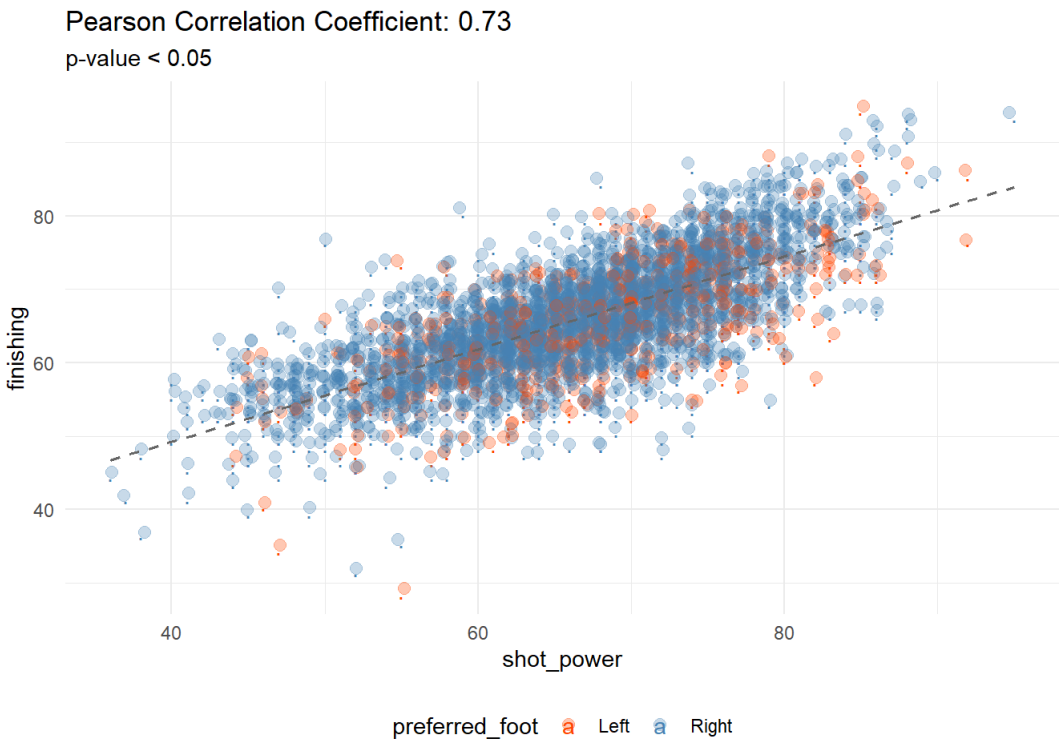
- **Tương quan giữa `age` và `overall`:** Kiểm tra liệu tuổi tác có ảnh hưởng đáng kể đến chỉ số tổng thể của cầu thủ.



Hình 11: Relation of age and overal

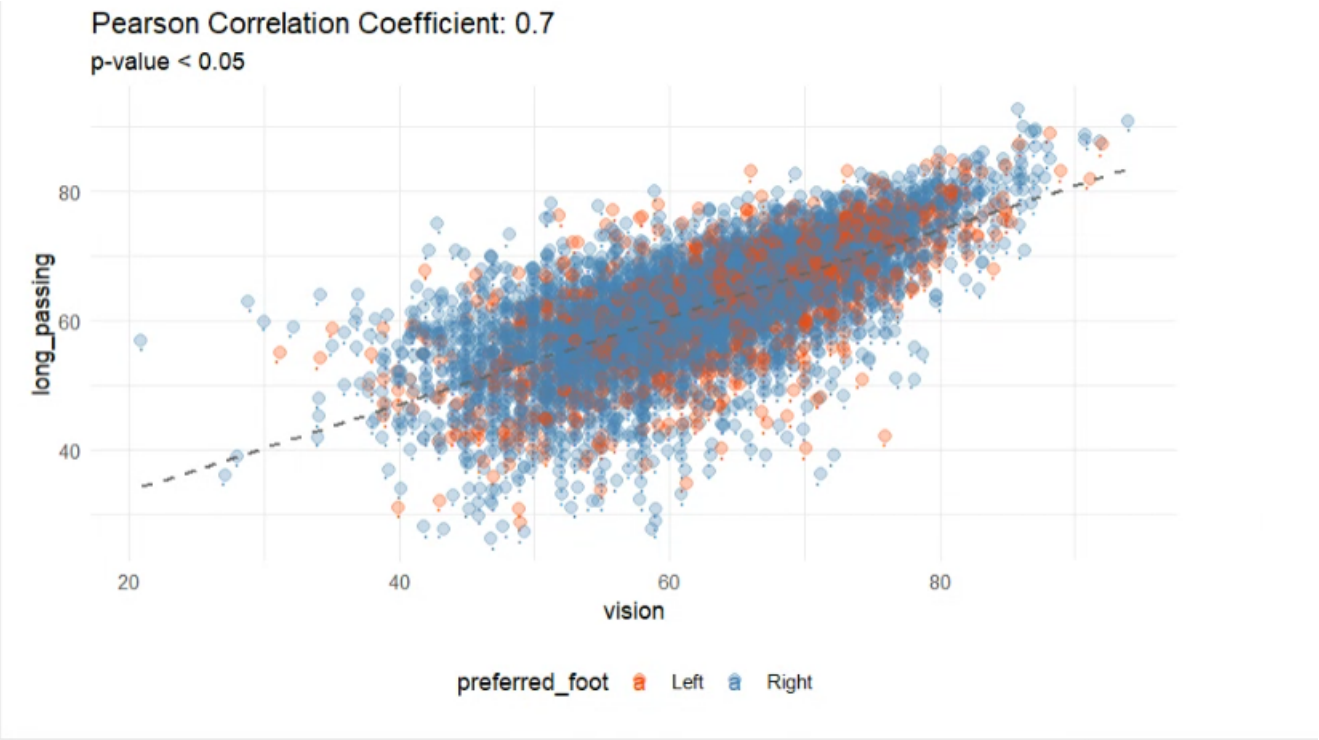
⇒ Có vẻ age không ảnh hưởng quá nhiều trực tiếp đến overall

- **Tương quan giữa `finishing` và `shot_power` của tiền đạo:** Do dữ liệu không tuân theo phân phối chuẩn, hệ số tương quan Kendall và Pearson được tính để so sánh.



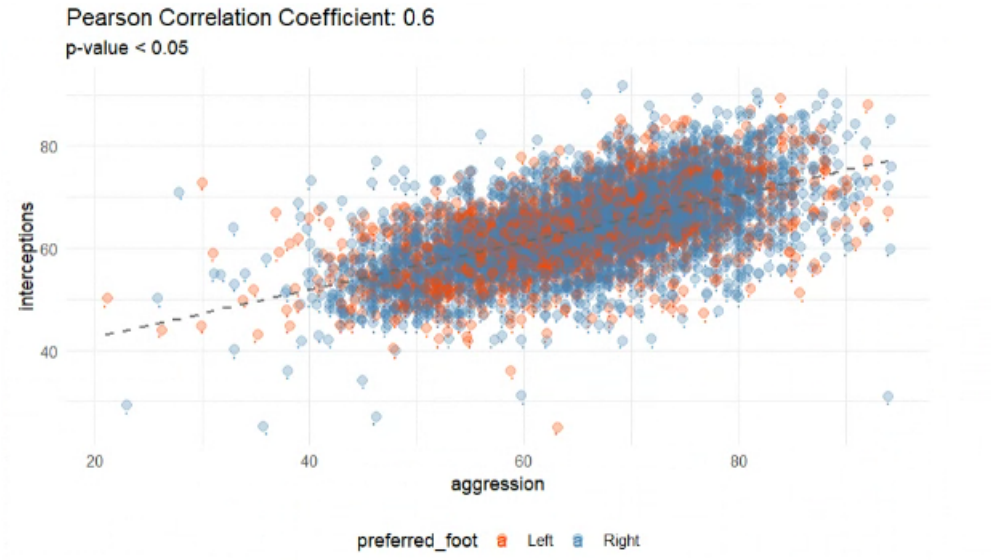
Hình 12: *shot_power* và finishing

- Ta thấy mối tương quan giữa *shot_power* và finishing khá cao đối với cả cầu thủ thuận chân trái và thuận chân phải
- Đa số các cầu thủ có khả năng dứt điểm thường sút mạnh



Hình 13: long pass và vision

- *long_pass* và vision là 2 chỉ số quan trọng của hàng tiền vệ
- Dựa vào data frame, ta có thể thống kê rằng, các tiền vệ xuất sắc đều có 2 chỉ số này ở mức cao vì nó có khả năng quyết định sức tấn công của đội bóng



Hình 14: aggression và interceptions

Ta thấy mối tương quan giữa aggression và interceptions của cầu thủ hậu vệ cũng khá cao, tuy nhiên vẫn có khá nhiều trường hợp ngoại lệ đối với các cầu thủ có chỉ số thấp hơn 60

3.3.2 Xây dựng mô hình dự đoán

Phân loại vị trí cầu thủ

- Ta sẽ xây dựng mô hình Random Forest để dự đoán vị trí của cầu thủ dựa vào các chỉ số của cầu thủ, sau đó sử dụng mô hình này để dự đoán vị trí của các cầu thủ không có dữ liệu ở cột position
- Dữ liệu `position` được phân loại thành các nhóm:
  - `GK (Goalkeeper)`: Thủ môn.
  - `DF (Defender)`: Hậu vệ, gồm các vị trí CB, LB, RB, LCB, RCB, LWB, RWB.
  - `MF (Midfielder)`: Tiền vệ, gồm các vị trí CDM, CM, CAM, LDM, RDM, LM, RM, LCM, RCM, LAM, RAM.
  - `FW (Forward)`: Tiền đạo, gồm các vị trí ST, CF, LW, RW, RS, LS
  - `NA (Not available)`
- Sử dụng mô hình `Random Forest` để dự đoán vị trí cầu thủ không có dữ liệu vị trí (`NA`).
- Đầu tiên chúng ta chia dữ liệu sau khi đã tổng quát hóa vị trí thành tập train để huấn luyện mô hình. Chúng ta chọn tập test là những cầu thủ chưa xác định vị trí. Phần còn lại được đưa vào tập train.
- Trong quá trình huấn luyện mô hình, ta chia dữ liệu thành 5 phần bằng nhau. Ta lấy 4 phần để huấn luyện mô hình, phần còn lại ta dùng để đánh giá mô hình, từ đó chọn ra mô hình tốt nhất
- Sau khi huấn luyện, ta đánh giá độ chính xác, in ra Confusion Matrix.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   DF   FW   GK   MF
##         DF 1112    3    0   97
##         FW    1  548    0  105
##         GK    0    0  405    0
##         MF   60  132    0 1165
##
## Overall Statistics
##
##               Accuracy : 0.8903
##                 95% CI : (0.8797, 0.9003)
##    No Information Rate : 0.3768
##    P-Value [Acc > NIR] : < 2.2e-16
##
##                  Kappa : 0.8443
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: DF Class: FW Class: GK Class: MF
## Sensitivity             0.9480    0.8023    1.0000    0.8522
## Specificity             0.9593    0.9640    1.0000    0.9151
## Pos Pred Value          0.9175    0.8379    1.0000    0.8585
## Neg Pred Value          0.9748    0.9546    1.0000    0.9111
## Prevalence              0.3233    0.1883    0.1116    0.3768
## Detection Rate          0.3065    0.1510    0.1116    0.3211
## Detection Prevalence    0.3341    0.1803    0.1116    0.3740
## Balanced Accuracy       0.9536    0.8832    1.0000    0.8837
```

Hình 15: Kết quả mô hình

- Ta sử dụng mô hình trên để dự đoán vị trí cầu thủ chưa xác định vị trí trong tập test, ta được kết quả như sau:

```
## # A tibble: 12 x 33
##    position crossing finishing heading_accuracy short_passing volleys dribbling
##    <fct>       <dbl>     <dbl>            <dbl>         <dbl>   <dbl>     <dbl>
##  1 DF             25        36               72            56      19        41
##  2 FW             64        73               65            64      52        67
##  3 DF             59        39               59            33      37        44
##  4 FW             52        70               54            57      63        74
##  5 DF             72        48               44            66      31        57
##  6 GK             15        20               15            23      17        14
##  7 DF             51        33               47            28      31        51
##  8 MF             53        47               39            57      56        57
##  9 FW             47        51               40            50      45        46
## 10 FW             35        56               49            38      38        53
## 11 FW             25        65               48            38      48        47
## 12 DF             22        21               48            36      20        38
## # i 26 more variables: curve <dbl>, fk_accuracy <dbl>, long_passing <dbl>,
## #   ball_control <dbl>, acceleration <dbl>, sprint_speed <dbl>, agility <dbl>,
## #   reactions <dbl>, balance <dbl>, shot_power <dbl>, jumping <dbl>,
## #   stamina <dbl>, strength <dbl>, long_shots <dbl>, aggression <dbl>,
## #   interceptions <dbl>, positioning <dbl>, vision <dbl>, penalties <dbl>,
## #   composure <dbl>, marking <dbl>, standing_tackle <dbl>,
## #   sliding_tackle <dbl>, gk_diving <dbl>, gk_handling <dbl>, ...
```

Hình 16: Đánh giá và kết quả

Hồi quy tuyến tính dự đoán giá trị cầu thủ (`value`)

- **Dữ liệu đầu vào:** `overall`, `wage`, `release_clause`, `potential`, và `international_reputation`.
- **Xây dựng mô hình:** Phương pháp hồi quy tuyến tính được áp dụng.

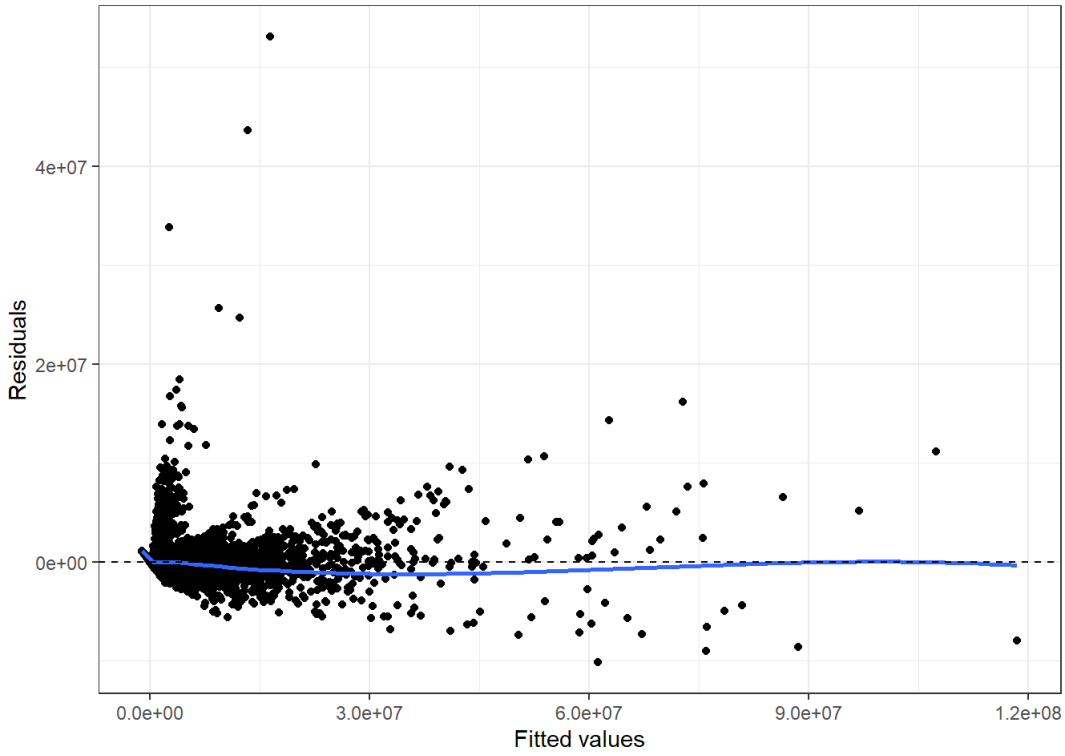
$$value \sim overall + wage + release\_clause + international\_reputation$$

```
##
## Call:
## lm(formula = value ~ overall + wage + release_clause + potential +
##     international_reputation, data = model_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8169 -0.0537 -0.0050  0.0312  9.4732
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.508e-14  1.621e-03    0.00        1
## overall                   3.827e-02  2.388e-03   16.03   <2e-16 ***
## wage                      1.669e-01  3.061e-03   54.53   <2e-16 ***
## release_clause            7.654e-01  3.015e-03  253.86   <2e-16 ***
## potential                 3.742e-02  2.243e-03   16.68   <2e-16 ***
## international_reputation  3.538e-02  2.242e-03   15.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2184 on 18141 degrees of freedom
## Multiple R-squared:  0.9523, Adjusted R-squared:  0.9523
## F-statistic: 7.247e+04 on 5 and 18141 DF,  p-value: < 2.2e-16
```

Hình 17: Model

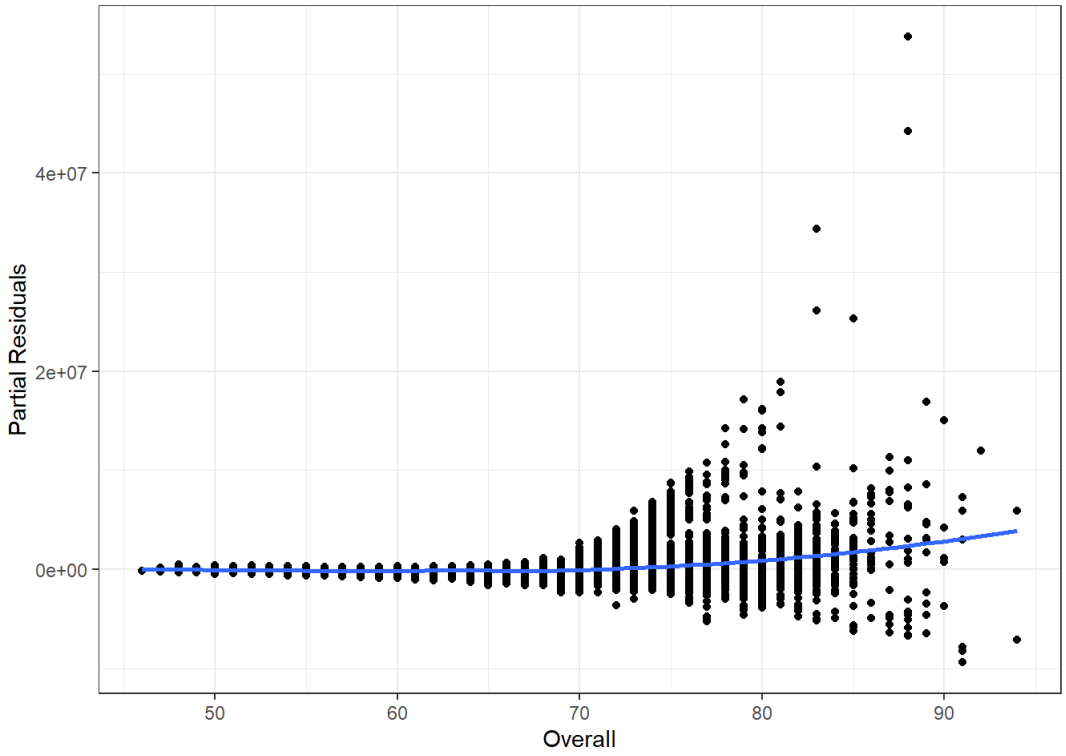
- Biến overall (hệ số 0.03827): Khi **overall** tăng 1 đơn vị, giá trị của *value* tăng trung bình 0.03827 đơn vị.
- Biến **wage** (hệ số 0.1669): Khi **wage** tăng 1 đơn vị, giá trị của *value* tăng trung bình 0.1669 đơn vị.
- Biến **release_clause** (hệ số 0.7654):
  - Khi **release_clause** tăng 1 đơn vị, giá trị của *value* tăng trung bình 0.7654 đơn vị.
  - Đây là biến có ảnh hưởng lớn nhất trong mô hình, vì hệ số của nó là cao nhất.
- Biến **potential** (hệ số 0.03742): Khi **potential** tăng 1 đơn vị, giá trị của *value* tăng trung bình 0.03742 đơn vị.
- Biến **international_reputation** (hệ số 0.03538): Khi **international_reputation** tăng 1 đơn vị, giá trị của *value* tăng trung bình 0.03538 đơn vị.

- **Kiểm định giả định:**
  - Kiểm tra tuyến tính hóa: Sử dụng biểu đồ phần dư (`Residuals vs Fitted`).



Hình 18: Đồ thị Residuals và Fitted values

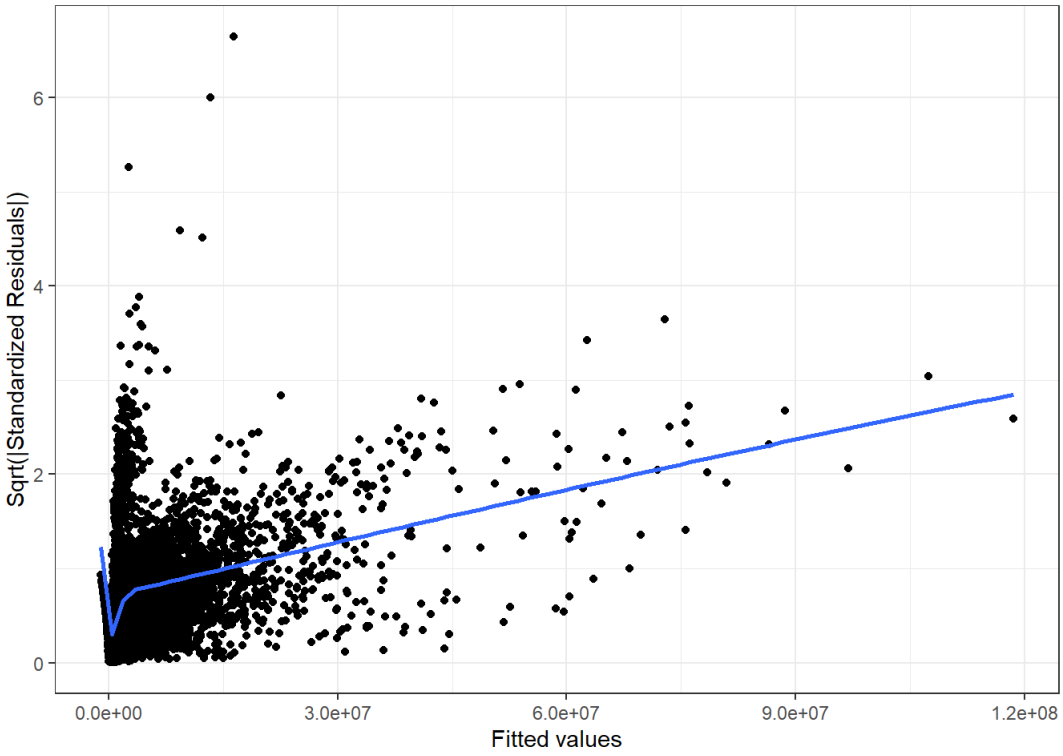
Từ hình vẽ, có thể thấy rằng các điểm thặng dư không biểu hiện bất kỳ xu hướng đường cong đáng kể nào. Điều này cho thấy rằng giả định về tính tuyến tính của mô hình là phù hợp



Hình 19: Partial Residual và Overall

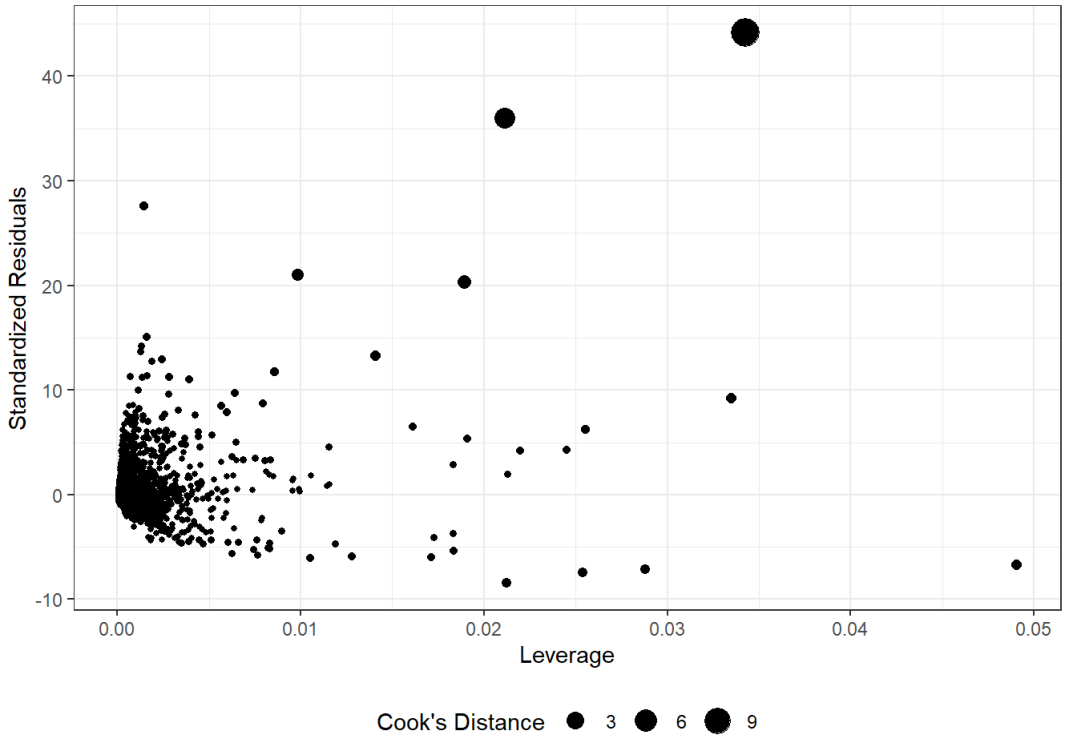
Hình 19 cho thấy đường tuyến tính (màu xanh dương) ước lượng tương đối khớp với dữ liệu.

– Kiểm tra tính đồng nhất phương sai: Dựa trên biểu đồ `Scale-Location`.



Hình 20: Kiểm tra tính đồng nhất phương sai

– Kiểm tra điểm ngoại lai: Sử dụng biểu đồ `Residuals vs Leverage`.



Hình 21: Kiểm tra điểm ngoại lai với biểu đồ Residuals và Leverage

Mô hình có một số điểm ngoại lai nhưng không đáng kể. Đa số tập trung ở ngưỡng 0.005

– Kiểm tra đa cộng tuyến: Tính chỉ số `VIF (Variance Inflation Factor)`.

```
##                overall                   wage         release_clause
##               2.169007               3.563890               3.459102
##              potential international_reputation
##               1.914175               1.912798
```

Hình 22: Kiểm tra đa cộng tuyến

⇒ Các giá trị đều bé hơn 5 nên không xảy ra hiện tượng đa cộng tuyến

- **Bootstrap:** Áp dụng phương pháp `Bootstrap` để ước lượng khoảng tin cậy 95

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = model_data2, statistic = fun_boot_md, R = 1000, formula = value ~
##     overall + wage + release_clause + potential + international_reputation)
##
##
## Bootstrap Statistics :
##          original        bias     std. error
## t1* -4.731655e+06  1.082098e+04 3.432112e+05
## t2*  3.101323e+04  7.580977e+01 2.657688e+03
## t3*  4.244700e+01 -3.476520e-01 7.220528e+00
## t4*  3.999610e-01  7.339032e-04 1.335688e-02
## t5*  3.419021e+04 -1.563443e+02 3.478123e+03
## t6*  5.029663e+05 -4.009418e+03 7.363221e+04
```

Hình 23: Bootstrap

3.3.3 Công cụ gợi ý cầu thủ cho huấn luyện viên theo số tiền và vị trí

Ta sẽ xây dựng một công cụ gợi ý cầu thủ cho huấn luyện viên dựa vào số tiền mà huấn luyện viên đó có và vị trí mà huấn luyện viên đó cần cầu thủ. Sau đó sẽ gợi ý cho huấn luyện viên top 20 cầu thủ tốt nhất cho vị trí đó dựa vào số tiền mà huấn luyện viên đó có.

Bước 1: Chuẩn bị dữ liệu

- Tạo dataframe `fifa_data3` từ `data`.
- Thêm cột `total_cost` vào `fifa_data3` để tính tổng chi phí mua cầu thủ:

$$\text{total\_cost} = \text{wage} \times 52 + \text{value} + \text{release\_clause}$$

Bước 2: Hàm `select_players_by_position`

1. **Lọc dữ liệu:**
   - Chỉ chọn những cầu thủ có tổng chi phí (`total_cost`) nhỏ hơn hoặc bằng ngân sách (`budget`).
   - Chỉ chọn những cầu thủ thuộc vị trí (`position`) mà huấn luyện viên yêu cầu.
2. **Sắp xếp:** Dữ liệu được sắp xếp giảm dần theo chỉ số `overall`.
3. **Xử lý thông tin chi tiết:**

- Phân nhóm vị trí: Xác định loại cầu thủ dựa trên vị trí (Thủ môn, Hậu vệ, Tiền vệ, Tiền đạo).
- Lựa chọn các cột thông tin liên quan, gồm thông tin cơ bản và chỉ số chi tiết (nếu cần).

4. **Trả về kết quả:** Gợi ý `top 20` cầu thủ phù hợp nhất.

Ví dụ sử dụng:

- Chọn 20 cầu thủ tốt nhất với kinh phí 10 triệu bảng, vị trí hậu vệ trung tâm (CB), không cần thông tin chi tiết:

```
select_players_by_position(fifa_data3, 10000000, "CB", 20, detail = FALSE)
```

```
## # A tibble: 20 x 14
##    name   overall potential  value  wage release_clause heading_accuracy jumping
##    <chr>    <dbl>     <dbl>  <dbl> <dbl>          <dbl>            <dbl>   <dbl>
##  1 Hilton      78        78 0      18000              0               79      79
##  2 L. Ló~      77        77 7   e6 13000              0               79      87
##  3 Luís ~      77        77 0          0              0               80      79
##  4 M. Ha~      77        77 2.70e6 31000        4600000               66      74
##  5 P. Ja~      77        77 2.10e6 53000        4000000               77      81
##  6 K. Mb~      77        77 7   e6 21000              0               81      68
##  7 Danilo      77        77 2.70e6 26000              0               72      76
##  8 J. Mo~      76        76 2.3 e6 48000        4400000               74      83
##  9 D. Bo~      75        75 9   e5 18000        1900000               74      74
## 10 N. Sp~      75        75 1.6 e6 11000        2700000               78      68
## 11 K. Wi~      75        80 7.5 e6 35000              0               74      64
## 12 A. Ra~      75        75 2   e6 48000        3800000               75      69
## 13 M. Le~      75        80 7.5 e6  9000              0               72      79
## 14 M. An~      74        74 3.20e6 25000        5500000               76      85
## 15 C. Be~      74        74 2.3 e6  8000        4000000               76      80
## 16 T. My~      74        74 1.7 e6  1000        3600000               70      70
## 17 S. An~      74        74 3.20e6  5000        5000000               70      76
## 18 K. Dj~      74        76 5.50e6 19000              0               70      72
## 19 A. Se~      74        75 0          0              0               63      78
## 20 Bigas       74        75 5.50e6 10000              0               71      76
## # i 6 more variables: strength <dbl>, aggression <dbl>, interceptions <dbl>,
## #   marking <dbl>, standing_tackle <dbl>, sliding_tackle <dbl>
```

Hình 24: Chọn 20 cầu thủ tốt nhất với kinh phí 10 triệu bảng, vị trí hậu vệ trung tâm (CB), không cần thông tin chi tiết

- Chọn 20 cầu thủ tốt nhất với kinh phí 50 triệu bảng, vị trí tiền đạo cắm (ST), cần thông tin chi tiết:

```
select_players_by_position(fifa_data3, 50000000, "ST", 20, detail = TRUE)
```

```
## # A tibble: 20 x 54
##        id name             age nationality overall potential club   value   wage
##     <dbl> <chr>          <dbl> <chr>         <dbl>     <dbl> <chr>  <dbl>  <dbl>
##  1 113422 David Villa       36 Spain            82        82 New ~ 8   e6  12000
##  2 106231 Aduriz            37 Spain            82        82 Athl~ 8   e6  29000
##  3 159145 B. Gomis          32 France           81        81 Al H~ 1.4 e7  56000
##  4 204529 M. Batshuayi      24 Belgium          80        84 Vale~ 1.95e7 105000
##  5 189805 L. de Jong        27 Netherlands      80        80 PSV   1.6 e7  23000
##  6 186537 C. Stuani         31 Uruguay          80        80 Giro~ 1.35e7  41000
##  7 176600 K. Gameiro        31 France           80        80 Vale~ 1.35e7  45000
##  8 153244 A. Gignac         32 France           80        80 Tigr~ 1.20e7  70000
##  9  54050 W. Rooney         32 England          80        80 DC U~ 1.20e7  13000
## 10 230498 Luimo Boas Sa~    30 Brazil           79        79 Sant~ 1.20e7  26000
## 11 228941 André Silva       22 Portugal         79        87 Sevi~ 1.9 e7  84000
## 12 216497 M. Philipp        24 Germany          79        84 Boru~ 1.7 e7  53000
## 13 215353 L. Alario         25 Argentina        79        82 Baye~ 1.6 e7  67000
## 14 215061 D. Benedetto      28 Argentina        79        79 Boca~ 1.3 e7  32000
## 15 203757 Zé Luís           27 Cape Verde       79        80 Spar~ 1.45e7   1000
## 16 198715 Sergio León       29 Spain            79        79 Real~ 1.25e7  32000
## 17 192991 C. Tosun          27 Turkey           79        80 Ever~ 1.45e7 105000
## 18 192678 Kike García       28 Spain            79        79 SD E~ 1.3 e7  32000
## 19 192598 Elkeson           28 Brazil           79        79 Shan~ 1.3 e7  25000
## 20 185431 N. Kalinic        30 Croatia          79        79 Atlé~ 1.20e7  60000
## # i 45 more variables: preferred_foot <chr>, international_reputation <dbl>,
## #   weak_foot <dbl>, skill_moves <dbl>, position <chr>, jersey_number <dbl>,
## #   loaned_from <chr>, height_cm <dbl>, weight_kg <dbl>, crossing <dbl>,
## #   finishing <dbl>, heading_accuracy <dbl>, short_passing <dbl>,
## #   volleys <dbl>, dribbling <dbl>, curve <dbl>, fk_accuracy <dbl>,
## #   long_passing <dbl>, ball_control <dbl>, acceleration <dbl>,
## #   sprint_speed <dbl>, agility <dbl>, reactions <dbl>, balance <dbl>, ...
```

Hình 25: Chọn 20 cầu thủ tốt nhất với kinh phí 50 triệu bảng, vị trí tiền đạo cắm (ST), cần thông tin chi tiết

4. Kết luận

Mô hình đã cung cấp thông tin giá trị dự đoán khả quan, giúp xác định các yếu tố ảnh hưởng đến giá trị cầu thủ và dự đoán vị trí cầu thủ còn thiếu. Các giả định của mô hình hồi quy đều được kiểm định và đạt kết quả tốt.

5. Đề xuất

1. **Đàm phán, chuyển nhượng hợp lý:**
    - Sử dụng công cụ lựa chọn cầu thủ phù hợp với ngân sách để thực hiện chuyển nhượng tốt hơn. Dựa vào đó và các kết quả thống kê dữ liệu khác, xây dựng đội hình phù hợp

với ngân sách, chiến lược và mục tiêu của đội bóng.

- Khi đàm phán hợp đồng, cần đảm bảo điều khoản giải phóng phản ánh đúng giá trị thực tế của cầu thủ. Điều này giúp:
  - Bảo vệ lợi ích tài chính của câu lạc bộ nếu cầu thủ bị chuyển nhượng.
  - Tránh việc điều khoản giải phóng quá thấp khiến câu lạc bộ mất đi những cầu thủ quan trọng với giá không tương xứng.

2. **Ưu tiên gia hạn hợp đồng với cầu thủ tiềm năng:**
   - Thông qua mô hình dự đoán vị trí cầu thủ để huấn luyện hợp lý và khai thác điểm mạnh và sở trường của các cầu thủ trẻ
   - Đối với các cầu thủ có *potential* cao (≥ 70) nhưng giá trị hiện tại thấp, câu lạc bộ có thể đầu tư vào việc gia hạn hợp đồng dài hạn và đặt mức điều khoản giải phóng cao hơn để tối ưu lợi ích trong tương lai.
3. **Sử dụng công cụ dự đoán để định giá cầu thủ:**
   Dựa trên mô hình hồi quy, câu lạc bộ có thể định giá cầu thủ dựa trên các chỉ số như *overall*, *potential*, và *wage* để đưa ra chiến lược ký kết và bán cầu thủ hợp lý, tránh trả lương vượt mức giá trị thực tế.

Cảm ơn thầy đã theo dõi báo cáo của nhóm.