### **Báo cáo kết quả tiền xử lý, trực quan hóa và phân tích dữ liệu từ tập dữ liệu các giải đấu bóng đá Châu Âu (2008-2016)**

### 

### 

### **Báo cáo được thực hiện bởi**

### Nguyễn Văn Thức - SIC0241 - HL AI1

**Ngày 11 tháng 8 năm 2024**

**Mục lục**

1. Mô tả bài toán.........................................................3
2. Mô tả dữ liệu...........................................................4
3. Tiền xử lí dữ liệu.....................................................5
4. Trực quan hóa và phân tích dữ liệu...........................................................................7
5. Kết quả và thảo luận.............................................18
6. Những điểm hạn chế…………………………….19
7. Kết luận………………………………………….19
8. Tài liệu tham khảo………………………………19

### 

### 

### 

### 

### 

### 

### 

### 

### **1. Mô tả bài toán**

#### **1.1. Bài toán**

Đây là tập dữ liệu về các giải đấu hàng đầu Châu Âu từ mùa giải 2008-2009 đến 2015-2016. Dữ liệu này bao gồm cả những dữ liệu dự đoán trận đấu của hai nhà cái là bet365 và betwin. Bài toán đặt ra là làm thế nào để tiền xử lý dữ liệu, trực quan hóa và phân tích dữ liệu để rút ra những điều thú vị trong bóng đá cũng như tính toán xem tỷ lệ dự đoán chính xác của các nhà cái là bao nhiêu và có nên chơi cá cược bóng đá không?

**1.2. Bối cảnh và tầm quan trọng**

Bóng đá luôn là môn thể thao được nhiều người yêu quý nhất trên toàn cầu. Chúng ta đều biết rằng rất nhiều nước cũng như khu vực trên thế giới rất phát triển về bộ môn thể thao này, trong đó có Châu Âu. Các giải đấu hấp dẫn hàng đầu thế giới đều tập trung ở khu vực này. Phân tích dữ liệu từ các giải đấu này sẽ giúp chúng hiểu hiểu rõ hơn về xu hướng bóng đá cũng như các yếu tố ảnh hưởng đến kết quả của trận đấu, hiểu rõ hơn về việc phát triển kinh tế trong bóng đá. Ngoài ra, chúng ta cũng hiểu rõ hơn về tỷ lệ dự đoán chính xác của các nhà cái. Có cái nhìn sâu sắc hơn về cá độ bóng đá. Qua đó chúng ta rút ra được những bài học quý giá về tác hại của chơi cá cược bóng đá.

**1.3. Mục tiêu bài tập**

In this project, we will focus on the following aspects:

Trong bài tập này, chúng tôi sẽ tập trung vào những mặt:

1. Kiểm định giả thuyết: Kiểm định những giả thuyết về trung bình, xác suất của nhiều đối tượng như số bàn thắng ghi được hay tỉ lệ thắng của đội khách trên sân nhà.
2. Xây dựng khoảng tin cậy cho những biến quan trọng ảnh hưởng đến kết quả trận đấu.
3. Những phân tích khác như đánh giá tỷ lệ dự đoán đúng của các nhà cái hay phân tích từ những biểu đồ trực quan dữ liệu.

**1.4. Cấu trúc bài báo cáo**

**Bài báo cáo được chia thành các phần chính sau:**

1. **Mô tả bài toán**: Đưa ra bài toán, nếu bối cảnh và tầm quan trọng của phân tích dữ liệu, và nêu mục tiêu bài tập
2. **Mô tả dữ liệu**: mô tả thông tin dữ liệu cũng như chú thích của từng cột
3. Tiền xử lý dữ liệu: Sắp xếp lại dữ liệu và loại bỏ những thông tin không cần thiết.
4. **Trực quan hóa và phân tích dữ liệu**: Phân tích dữ liệu và đưa ra đồ thị minh họa
5. **Kết quả và thảo luận**: Tổng hợp lại những kết quả quan trọng và đưa ra nhận định chung.
6. **Những điểm hạn chế:** nêu những thiếu xót về dữ liệu, về phương pháp thống kê
7. **Kết luận**
8. **Tham khảo**

### **2. Mô tả dữ liệu**

- Đây là file dữ liệu thứ 11: FootballDataEurope.csv trong 20 tập dữ liệu đã được cung cấp.

- Dữ liệu bao gồm 18 cột và 25,979 dòng, mỗi dòng là thông tin về một trận đấu.

- Ý nghĩa của các cột trong dữ liệu:

1. id: mã số của trận đấu
2. country\_name: tên quốc gia
3. league\_name: tên giải đấu
4. season: mùa giải
5. stage: vòng đấu trong một giải
6. date: ngày diễn ra trận đấu
7. home\_team: đội chủ nhà
8. away\_team: đội khách
9. home\_team\_goal: số bàn thắng đội nhà
10. away\_team\_goal: số bàn thắng đội khách
11. B365H, BSH: tỉ lệ cược của đội nhà thắng
12. B365D, BSD: tỉ lệ cược hai đội hòa nhau
13. B365A, BSA:tỉ lệ cược đội khách thắng
14. diff goals: số bàn thắng chênh lệch của đội nhà và đội khách
15. target: số bàn thắng chênh lệch dự đoán

* Giải thích các cột B365H, B365D, B365A, BSH, BSD, BSA: Các cột này thể hiện tỉ lệ cược của nhà cái. Lấy ví dụ là nhà cái Bet365. Đội nhà thắng trong trường hợp B365H < B365A, Đội nhà thua trong trường hợp B365H > B365A. Và hai đội hòa nhau trong trường hợp B365D là nhỏ nhất trong 3 số. Tương tự các nhà cái khác cũng đánh giá như vậy.

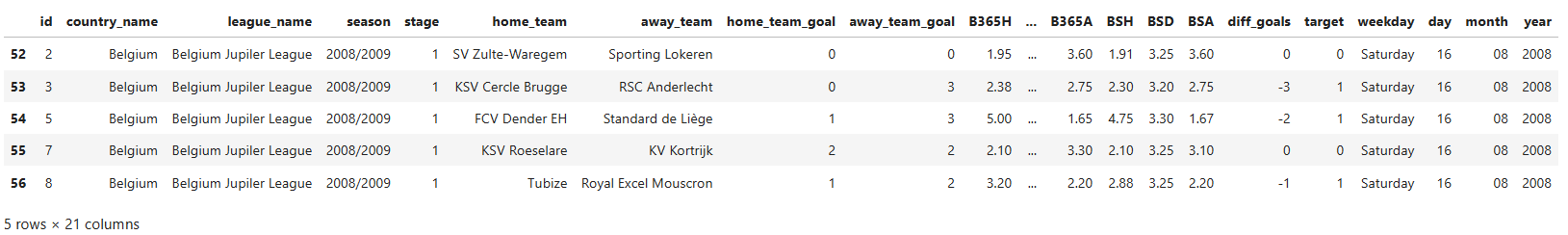
### **3. Tiền xử lý dữ liệu**

Ở đây, bài toán sẽ chia làm 2 bài toán nhỏ là phân tích dữ liệu khi không có dữ liệu của nhà cái và phân tích dữ liệu khi có dữ liệu của cả 2 nhà cái. Do đó, cách làm sạch cho mỗi bài toán nhỏ cũng sẽ khác nhau.

Cả hai bài toán đều có những bước chung sau:

* Sắp xếp dữ liệu theo từng quốc gia, giải đấu, mùa giải, vòng đấu và thời gian diễn ra của từng trận đấu.
* Kiểm tra các giá trị null và lặp lại.
* Định dạng lại cột date sao cho chỉ còn thời gian theo thứ tự ngày-tháng-năm.
* Chia cột date thành 3 cột riêng biệt: ngày, tháng, năm.
* Thêm một cột thể hiện thứ trong tuần tương ứng với ngày diễn ra trận đấu.
* Xóa cột date sau khi đã thực hiện xong việc tách cột .

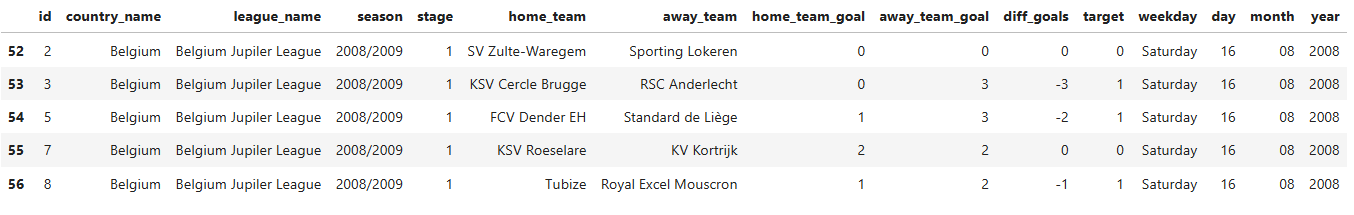
Sau khi thực hiện các bước chung trên, ta sẽ có dữ liệu với 5 dòng đầu như sau:



**Bài toán nhỏ 1: phân tích dữ liệu không dùng số liệu của nhà cái**

* Xóa 6 cột dữ liệu của nhà cái bao gồm B365H, B365D, B365A, BSH, BSD, BSA.

Sau khi xóa, ta sẽ có dữ liệu với 5 dòng đầu như sau:



**Bài toán nhỏ 2: Phân tích dữ liệu có dữ liệu của cả 2 nhà cái**

* Xóa các cột có giá trị null

Sau khi xóa hết các cột có giá trị null, ta sẽ được dữ liệu mới gồm 14160 dòng. Dữ liệu sẽ như sau:

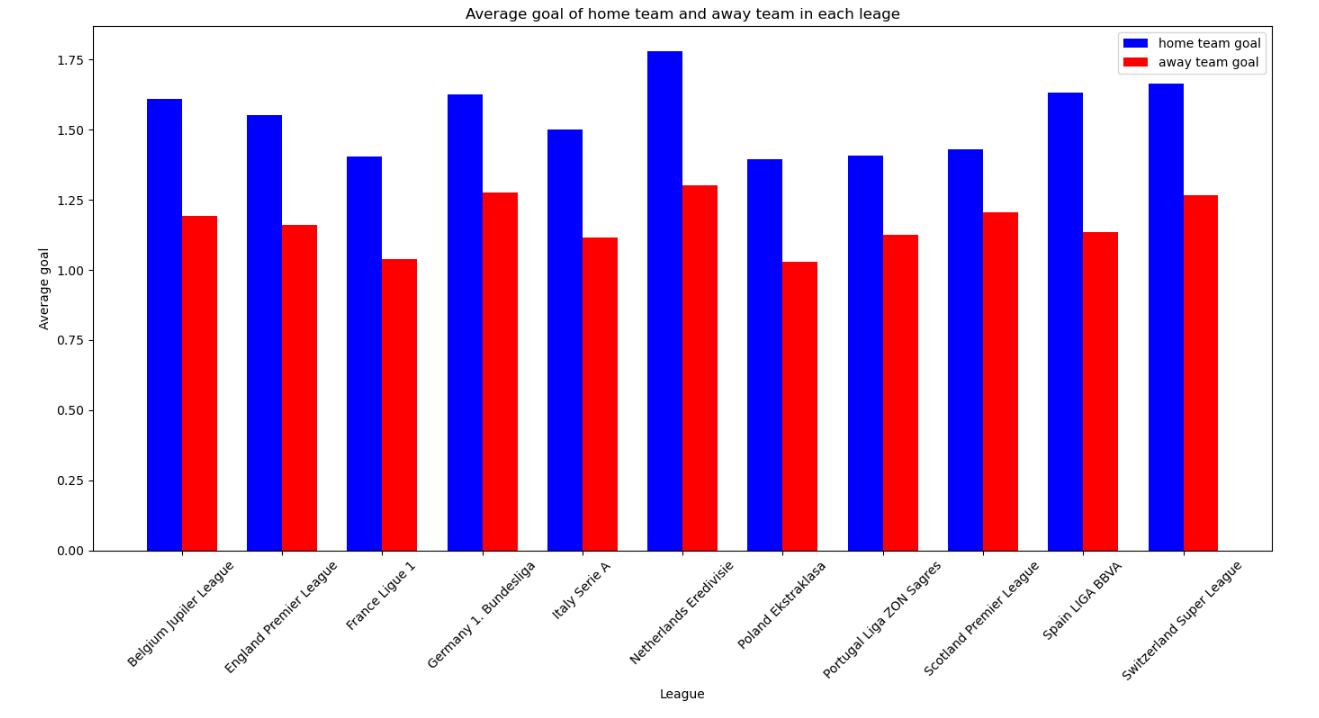


**4. Trực quan hóa và phân tích dữ liệu**

**Bài toán nhỏ 1**: Phân tích dữ liệu thực của trận đấu mà không có dữ liệu của nhà cái

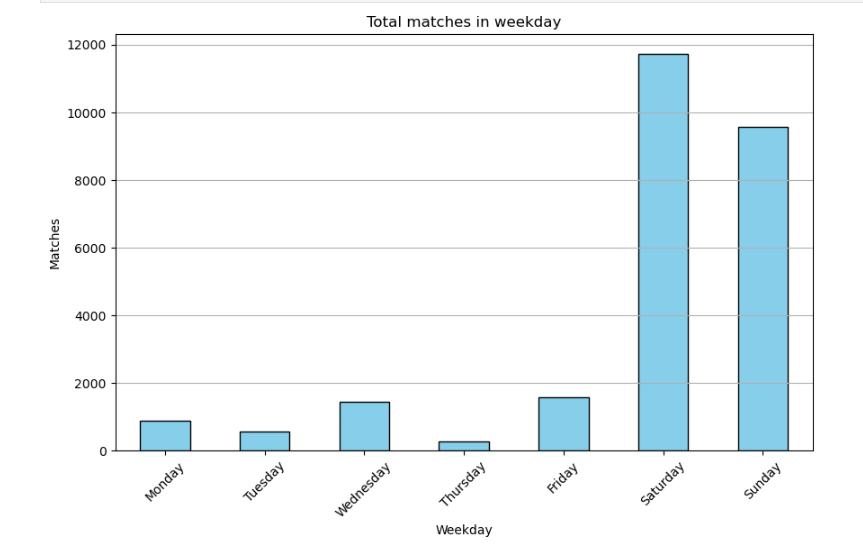
**4.1 Phân tích dữ liệu từ biểu đồ trực quan hóa**

a, Biểu đồ thể hiện số bàn thắng trung bình của đội nhà và đội khách theo từng giải đấu



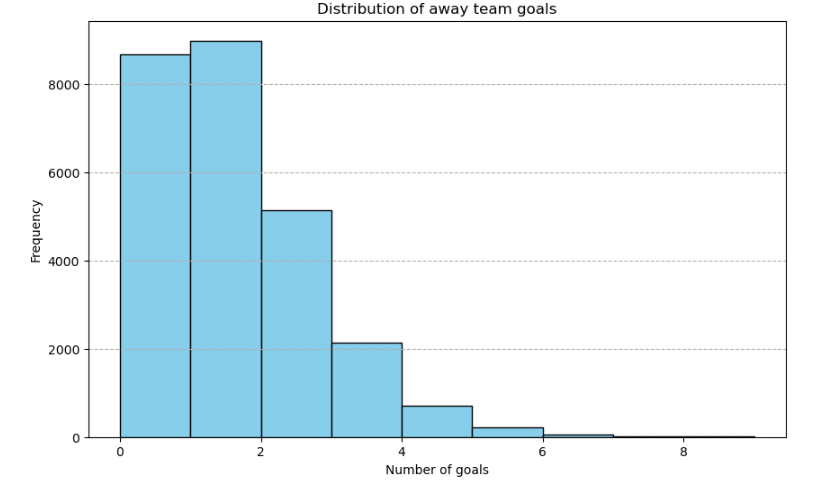
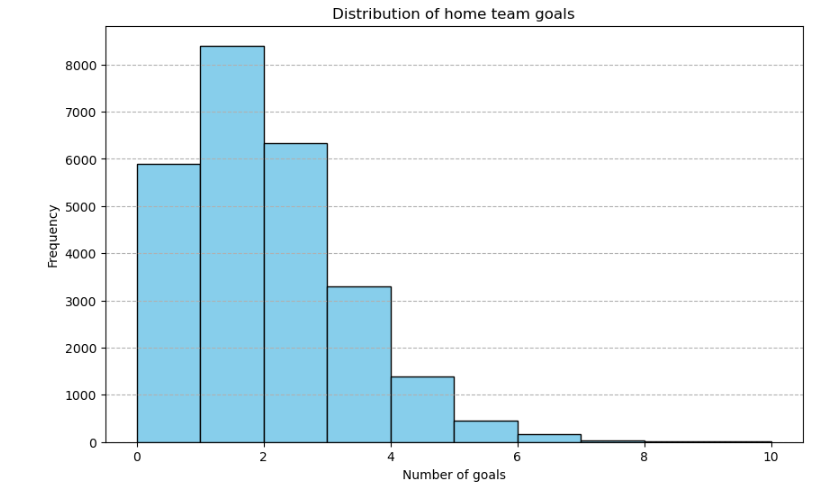
Nhìn vào biểu đồ ta có thể thấy rằng ở các giải đấu thì số lượng bàn thắng trung bình của đội nhà luôn cao hơn số lượng bàn thắng trung bình của đội khách. Điều đó có nghĩa rằng khi được chơi ở trên sân nhà, được sự cổ động viên nồng nhiệt của các fan hâm mộ, các cầu thủ sẽ tự tin hơn và màn trình diễn của họ sẽ trở nên tốt hơn dẫn đến số bàn thắng sẽ đến nhiều hơn. Ngược lại khi ở sân khách, các cầu thủ sẽ chịu một áp lực tâm lý rất lớn, dẫn đến việc thực hiện những chiến thuật đề ra bị cản trở khá nhiều cũng như số bàn thắng ghi được sẽ ít hơn so với khi chơi ở sân nhà.

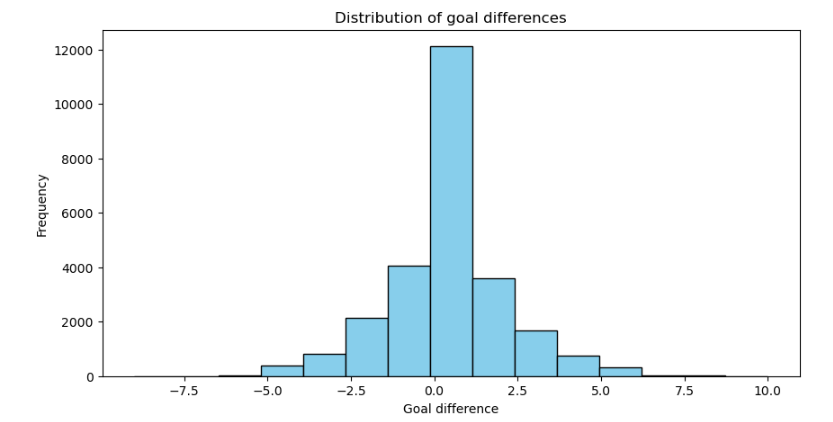
b, Biểu đồ thể hiện tổng số trận đấu theo các ngày trong tuần



Biểu đồ cho ta thấy được rằng số lượng trận đấu diễn ra vào cuối tuần là rất lớn, cao hơn gấp nhiều lần so với các ngày trong tuần. Đó là khoảng thời gian mà hầu hết mọi người không phải đi làm mà sẽ nghỉ ở nhà. Việc tổ chức các trận đấu và cuối tuần không chỉ giúp cho khán giả có thể theo dõi được nhiều trận đấu hơn mà nó còn mang lại lợi nhuận khổng lồ cho các đội bóng thông qua việc bán vé và các đồ lưu niệm tại sân vận động.

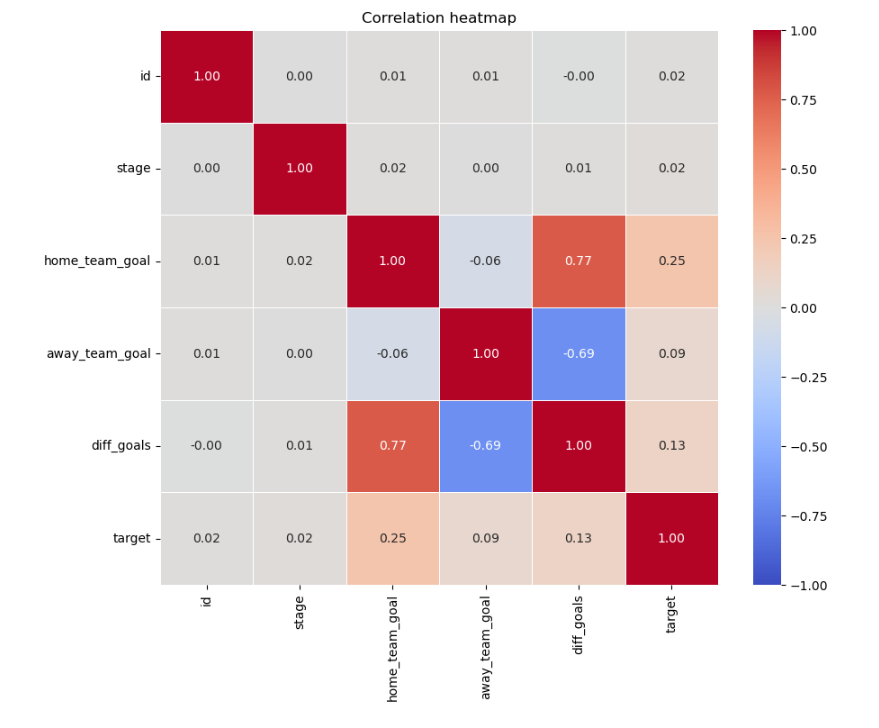
c, Biểu đồ thể hiện phân bố của số bàn thắng đội nhà và đội khách, và sự chênh lệch bàn thắng của hai đội





Từ 3 biểu đồ trên, chúng ta có thể nhận thấy rằng hầu hết số bàn thắng của đội nhà và đội khách trong một trận đấu là 1-2 bàn thắng. Lý do cho việc ít bàn thắng này là bởi các nước Châu Âu đã chú trọng hơn trong chiến lược phát triển bóng đá hiện đại. Cơ sở vật chất cũng như sự đầu tư cho chuyên môn của các đội bóng đã góp phần làm thu hẹp khoảng cách về trình độ của các đội bóng lại. Từ đó, các trận bóng sẽ trở lên hấp dẫn hơn, làm hài lòng các fan hâm mộ cuồng nhiệt.

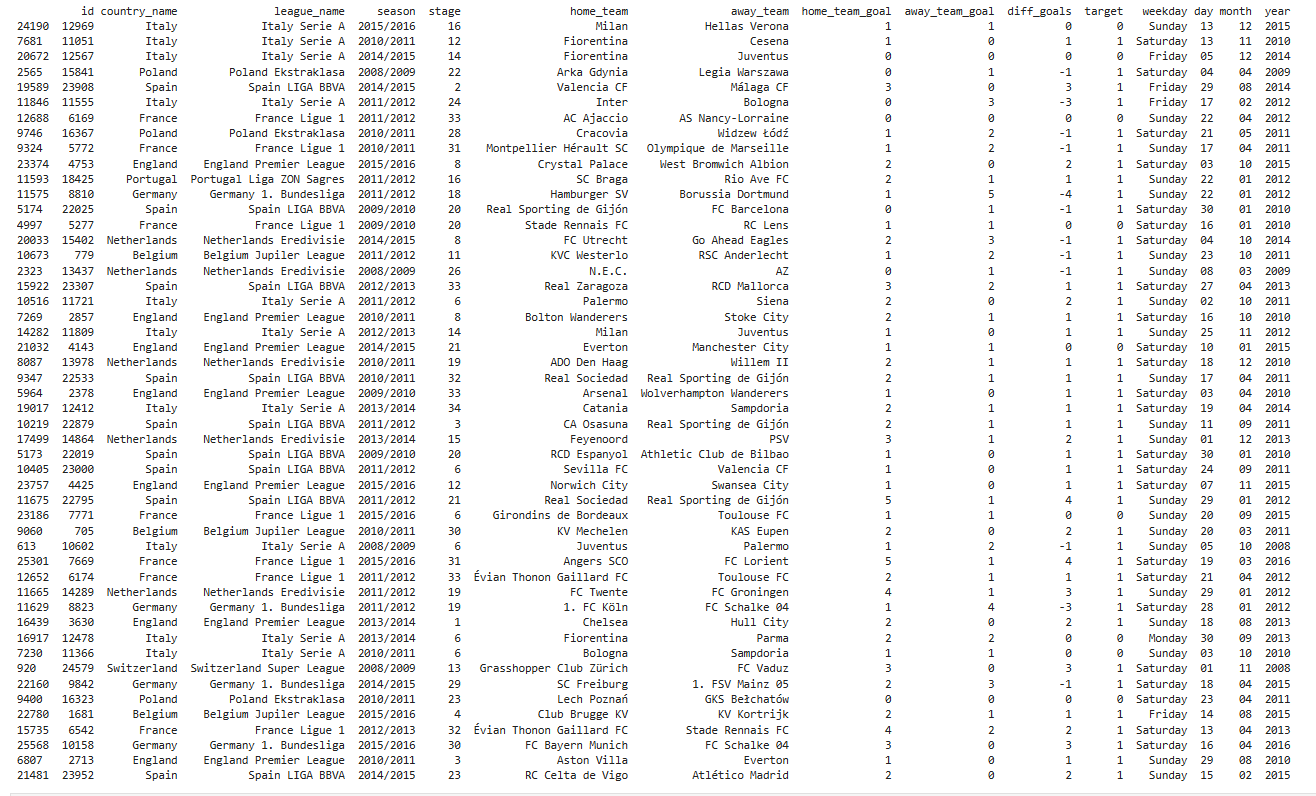
d, Biểu đồ nhiệt thể hiện mối tương quan giữa các cột trong dữ liệu



Biểu đồ cho thấy mối tương quan của từng cột trong bảng với nhau. Nhìn vào biểu đồ ta có thể dễ thấy rằng sự ảnh hưởng lẫn nhau của diff\_goals và home\_team\_goal là rất lớn với hệ số là 0.77. Ngược lại, hệ số tương quan giữa bàn thắng đội khách với hiệu số là hệ số âm (-0.69). Điều đó có nghĩa là số bàn thắng ghi được của đội nhà và hiệu số bàn thắng là tương quan mạnh. Số bàn thắng của đội nhà sẽ là yếu tố quyết định đến hiệu số bàn thắng của hai đội. Ở mục a đã chỉ ra số bàn thắng trung bình của đội nhà là lớn hơn so với đội khách, đây là nguyên nhân dẫn đến mối tương quan trên.

**4.2 Phân tích dữ liệu từ kiểm thử giả thuyết**

**4.2.1 Lấy mẫu ngẫu nhiên 50 trận đấu**

****

**4.2.2 Tính dữ liệu thống kê mẫu**

****

**4.2.3 Xây dựng khoảng tin cậy 95% và kiểm định giả thuyết (alpha = 0.05)**

**a, Giả thuyết 1:** Số lượng bàn thắng trung bình của đội nhà bằng số lượng bàn thắng của đội khách

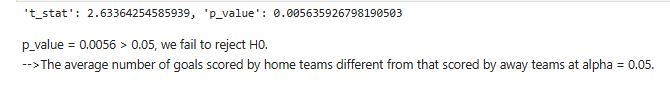
* Khoảng tin cậy 95% cho hiệu số bàn thắng của đội nhà và đội khách là từ 0.147 đến 1.093



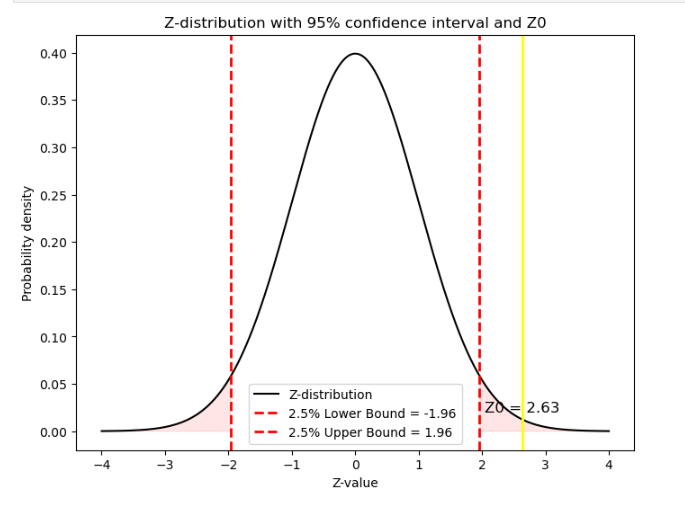
* Giả thuyết không (H0) và giả thuyết thay thế (H1):



* Kết quả kiểm định



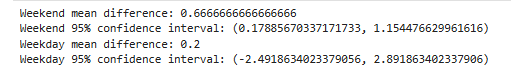
* Đồ thị trực quan:



#### Kiểm định cho thấy rằng số lượng bàn thắng trung bình của đội nhà là khác với số lượng bàn thắng trung bình của đội khách, cụ thể là lớn hơn do khoảng tin cậy luôn lớn hơn 0.

#### **b, Giả thuyết 2:** Hiệu số bàn thắng của đội nhà và đội khách là khác nhau giữa những trận diễn ra trong tuần và những trận diễn ra cuối tuần.

#### Khoảng tin cậy 95% cho hiệu số trong tuần và hiệu số cuối tuần của đội nhà và đội khách:



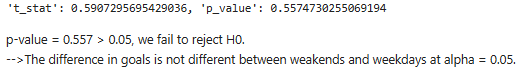
Trong tuần: từ 0.128 đến 1.154

Cuối tuần: từ -2.49 đến 2.89

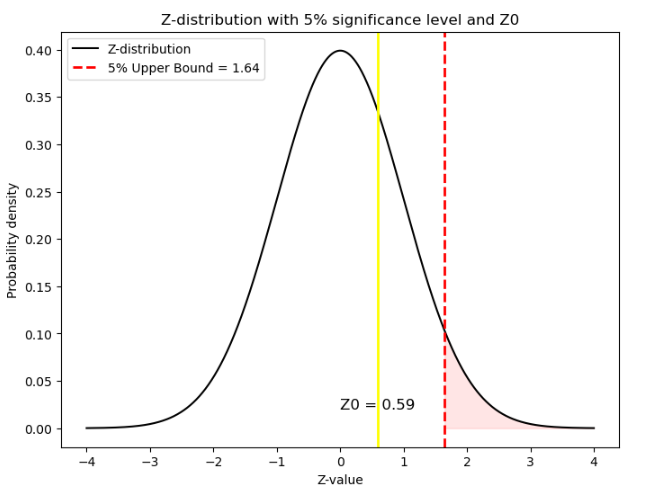
* Giả thuyết không (H0) và giả thuyết thay thế (H1):



* Kết quả kiểm định:



* Đồ thị trực quan:



#### Kết quả kiểm định cho thấy rằng hiệu số bàn thắng giữa đội nhà và đội khách ở trong tuần và ở cuối tuần là như nhau

**c, Giả thuyết 3:** Phần trăm tỉ lệ thắng của đội nhà ở những trận đấu cuối tuần cao hơn ở những trận đấu trong tuần

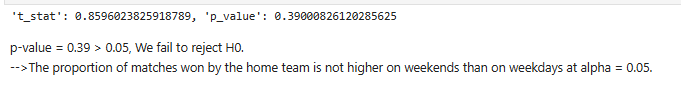
* Khoảng tin cậy 95% cho sự khác biệt về phần trăm tỉ lệ thắng của đội nhà trong tuần và cuối tuần là từ -0.373 đến 0.743



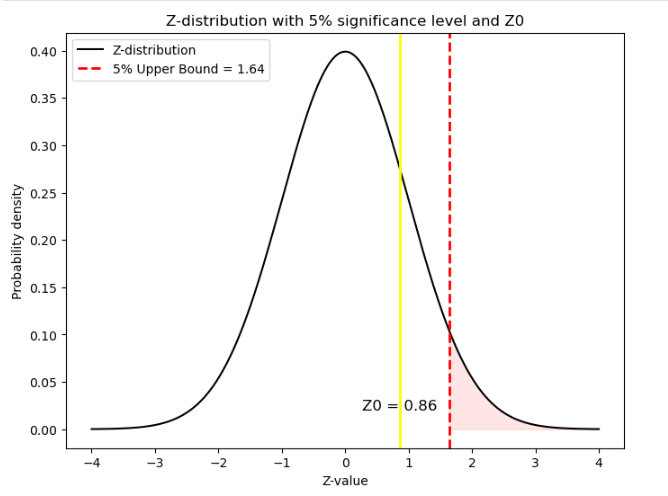
* Giả thuyết không (H0) và giả thuyết thay thế (H1):



* Kết quả kiểm định:



* Đồ thị trực quan:



#### Kết quả kiểm định cho thấy rằng tỉ lệ thắng của đội nhà ở cuối tuần là không cao hơn so với tỉ lệ thắng của đội nhà ở trong tuần.

**d, Giả thuyết 4:** Số bàn thắng trung bình ghi được của các đội bóng là khác nhau giữa Pháp (League 1) và Anh (Ngoại hạng Anh)

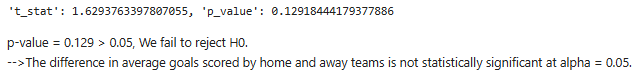
* Khoảng tin cậy 95% cho sự khác biệt giữa số bàn thắng trung bình của các đội bóng ở Pháp và ở Anh là từ -0.290 đến 3.147



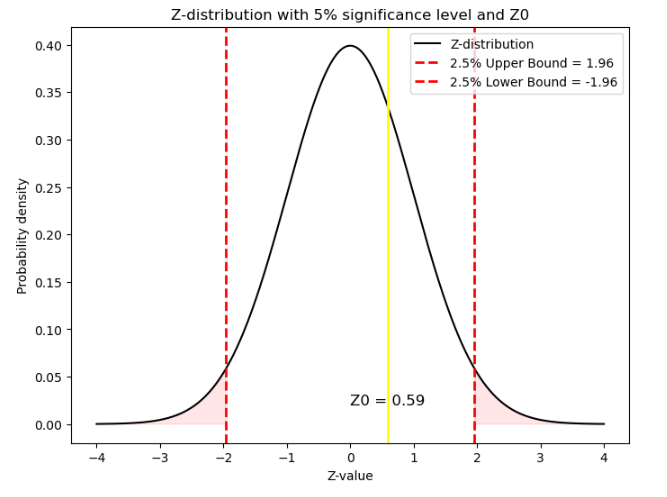
* Giả thuyết không (H0) và giả thuyết thay thế (H1):



* Kết quả kiểm định:



* Đồ thị trực quan:

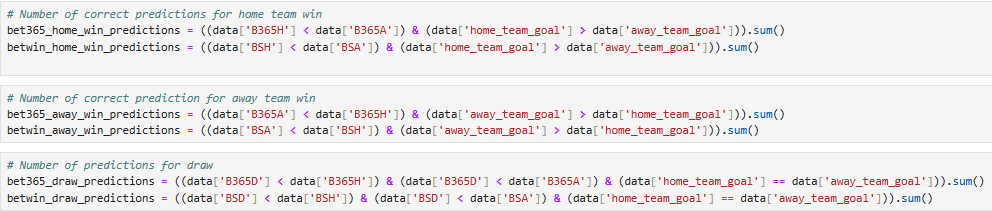


Kết quả kiểm định cho thấy rằng không có sự khác biệt quá lớn trong số bàn thắng trung bình của các đội bóng ở Pháp và Anh.

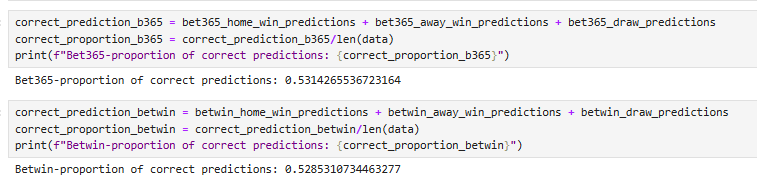
#### **Bài toán nhỏ 2: Phân tích tỷ lệ dự đoán của những trận đấu có dữ liệu của 2 nhà cái.**

**1, Tỷ lệ dự đoán chính xác thắng, hòa, thua của 2 nhà cái**

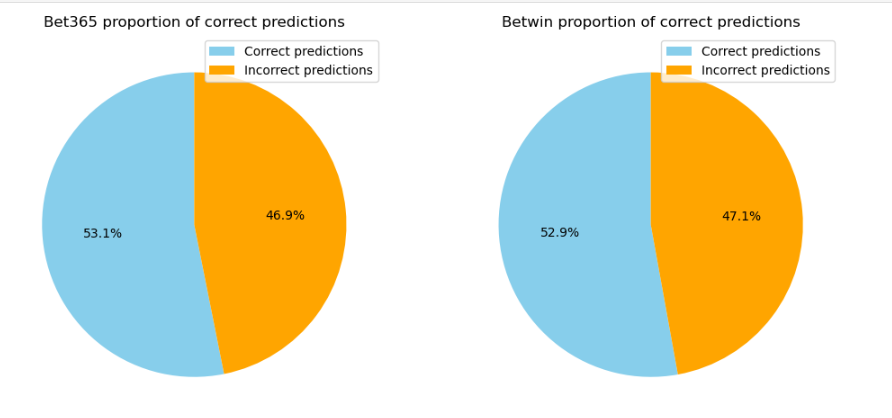
#### Đầu tiên, ta sẽ tính số trận dự đoán chính xác của 2 nhà cái cho 3 trường hợp đội nhà thắng, thua và hòa



* Sau đó ta sẽ cộng 3 trường hợp trên lại rồi chia cho số trận đấu có sự dự đoán của cả hai nhà cái (14160 trận đấu) để tính tỷ lệ dự đoán chính xác.
* Kết quả:



* Đồ thị trực quan:



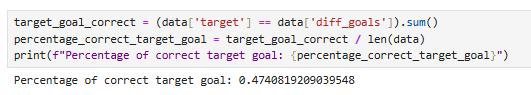
#### 

Tỷ lệ dự đoán chính xác của nhà cái Bet365 là khoảng 53.14%

Tỷ lệ dự đoán chính xác của nhà cái Betwin là khoảng 52.85%

#### **2, Tỷ lệ dự đoán chính xác hiệu số bàn thắng của 2 nhà cái**

* Đầu tiên ta sẽ đếm những trận đấu mà hiệu số bàn thắng dự đoán bằng với thực tế. Sau đó ta sẽ chia cho tổng số trận đấu có sự dự đoán của nhà cái (14160 trận)
* Kết quả:



* Đồ thị trực quan:

#### 

Kết quả tỷ lệ dự đoán chính xác hiệu số bàn thắng của 2 nhà cái là khoảng 47.41%

#### **5 Kết quả và thảo luận**

Sau khi phân tích và trực quan hóa dữ liệu về bóng đá Châu Âu từ mùa giải 2008-2009 đến 2015-2016, một số điểm chính được nhấn mạnh như sau:

1. Sự chênh lệch về tỉ số: Số bàn thắng ghi được của đội nhà là cao hơn của đội khách, tuy nhiên sự khác biệt này không phải là quá lớn
2. Thời gian diễn ra các trận đấu: Các trận đấu diễn ra vào cuối tuần nhiều hơn so với trong tuần. Những đội bóng có thứ hạng cao và có nhiều fan hâm mộ sẽ được ưu tiên đá vào cuối tuần.
3. Tỷ lệ chiến thắng khi đá ở sân nhà: Tỷ lệ chiến thắng khi đá trên sân nhà của các đội bóng giữa các ngày trong tuần và cuối tuần là như nhau. Điều đó phụ thuộc vào nhiều yếu tố khác nhau như đối thủ, thể trạng các cầu thủ và điều kiện thời tiết
4. Tỷ lệ dự đoán chính xác thắng, hòa, thua của 2 nhà cái là khoảng 53% và tỷ lệ dự đoán chính xác hiệu số trận đấu là vào khoảng 47%. Dựa vào kết quả này, chúng ta có thể thấy được rằng tỷ lệ dự đoán chính xác của nhà cái là khá thấp. Do đó chúng ta không nên chơi cá cược bóng đá để tránh phải rơi vào bẫy của các nhà cái dẫn đến mất tiền cũng như ảnh hưởng đến cuộc sống của chúng ta.

Nhìn chung, những kết quả trên thể hiện rằng những giải đấu ở Châu Âu diễn ra rất hấp dẫn với cân bằng về mặt chuyên môn của các đội bóng. Ngoài ra chúng ta cũng hiểu thêm được về tỷ lệ dự đoán chính xác của các nhà cái và hiểu được những rủi ro khi chơi cá cược bóng đá.

#### **6 Những điểm hạn chế**

* Hạn chế về dữ liệu: Dữ liệu có quá ít thông tin quan trọng khác về trận đấu dẫn đến việc phân tích chưa được đầy đủ về bóng đá Châu Âu.
* Hạn chế về phương pháp kiểm định: Phương pháp kiểm định còn đơn giản nên chưa kiểm định được nhiều giả thiết có ý nghĩa khác.

### **7. Kết luận**

Bài tập đã giúp chúng ta được dữ liệu các giải đấu hàng đầu Châu Âu từ năm 2008 đến năm 2016 để đưa ra được những nhận định có ý nghĩa bằng việc sử dụng trực quan hóa và các phương pháp thống kê. Những điểm ý nghĩa về bóng đá rút ra từ bài tập là về số bàn thắng đội nhà nhiều hơn đội khách, trận đấu thường diễn ra vào cuối tuần, không có sự khác biệt về tỷ lệ chiến thắng của đội nhà là như nhau ở trong tuần và cuối tuần. Thêm vào đó, tỷ lệ dự đoán chính xác trận đấu của các nhà cái là khá thấp (khoảng 50%). Những phân tích này sẽ giúp chúng ta hiểu rõ hơn phần nào về môn thể thao vua tại lục địa già cũng như rút ra được lời khuyên cho chính bản thân mình của việc chơi cá độ bóng đá.

### **8. Nguồn tham khảo**

<https://industri.fatek.unpatti.ac.id/wp-content/uploads/2019/03/088-Applied-Statistics-and-Probability-for-Engineers-Douglas-C.-Montgomery-George-C.-Runger-Edisi-5-2011.pdf>

<https://www.uefa.com/>

<https://www.quora.com/Why-do-European-football-leagues-especially-English-ones-start-games-at-lunch-time-in-the-weekend>

<https://www.researchgate.net/figure/The-percentage-of-goals-predicted-correctly_tbl2_342442670>