**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO THỰC HÀNH**

**PHÂN TÍCH DỮ LIỆU**

**MÔN HỌC: PYTHON**

**Đề tài: Phân tích dữ liệu giải bóng đá**

**Premier League 2000 – 2023**

**Trần Công Hậu – N22DCCI010**

**Hồ Chí Minh**

[PHẦN 1: TỔNG QUAN DỰ ÁN 3](#_bookmark1)

MỤC LỤC

1. [Báo cáo công việc: 3](#_bookmark2)

1.1 [Giới thiệu: 3](#_bookmark2)

1.2 [Mục tiêu nghiên cứu: 3](#_bookmark2)

1.3 [Phạm vi nghiên cứu: 3](#_bookmark2)

1.4 [Phương pháp nghiên cứu: 3](#_bookmark2)

[PHẦN 2: PHÂN TÍCH DỮ LIỆU 4](#_bookmark5)

1. [Mô tả dữ liệu: 4](#_bookmark6)
2. [Các câu hỏi đặt ra cho dữ liệu: 4](#_bookmark7)
3. [Phân tích và trực quan hóa dữ liệu từ câu hỏi: 5](#_bookmark8)

3.1 [Đội bóng dành nhiều chúc vô địch nhất EPL: 5](#_bookmark9)

3.2 Các đội bóng vô địch liên tiếp nhiều lần nhất EPL.[: 6](#_bookmark9)

3.3  [Đội vô địch mà không thua trận đấu nào.: 7](#_bookmark9)

3.4  [Chuỗi thắng liên tiếp dài nhất trong một mùa giải.: 8](#_bookmark9)

3.5  [Chuỗi bất bại dài nhất trong một mùa: 9](#_bookmark9)

3.6  [Tính trung bình số lượng bàn thắng của mỗi CLB trong một mùa giải.: 10](#_bookmark9)

3.7  [Tìm đội có tỷ lệ thắng cao nhất trong mùa giải: 11](#_bookmark9)

3.8  [Tìm CLB có nhiều trận không để thủng lưới nhất trong một mùa giải: 12](#_bookmark9)

* 1. [Trong 5 mùa giải gần nhất đưa ra thống kê số trận thắng và hiệu suất ghi bàn: 13](#_bookmark9)

1. [Kết phân tích : 14](#_bookmark8)

[PHẦN 3:KẾT LUẬN 14](#_bookmark25)

1. [Đánh giá dữ liệu: 14](#_bookmark26)
2. [Kiến thức rút ra: 14](#_bookmark27)
3. [Các tài liệu dùng tham khảo: 14](#_bookmark27)

2

# **PHẦN 1: TỔNG QUAN VỀ DỰ ÁN**

1. **Báo Cáo công việc**
2. **Giới thiệu:**

* **Premier League ( Ngoại Hạng Anh )**: Giải bóng đá Ngoại hạng Anh, thường được biết đến với tên gọi English Premier League hoặc EPL, là hạng đấu cao nhất của hệ thống các giải bóng đá ở Anh. Gồm 20 câu lạc bộ, giải đấu sử dụng hệ thống thăng hạng và xuống hạng với English Football League (EFL).

Mùa giải kéo dài từ tháng 8 đến tháng 5 với mỗi đội chơi 38 trận đấu (đấu với 19 đội khác trên sân nhà và sân khách). Đa số các trận đấu được diễn ra vào chiều Thứ Bảy và Chủ Nhật.

* **Tên dữ liệu**: Dữ liệu giải bóng đá Premier League ( Ngoại Hạng Anh ) của mùa giải 1999/2000 – 2022/23.
* **Lý do lựa chọn đề tài**: Em là một người yêu thích bóng đá, đặc biệt là bóng đá ngoại hạng anh. Em luồn muốn tìm hiểu về thành tích và các thông số của các dội bóng ở Ngoại Hạng Anh cho nên đã lựa chọn đề tài này.

1. **Mục tiêu nghiên cứu:**

* **Dự đoán kết quả** của các trận đấu trong tương lai dựa vào số trận thắng của đội bóng chủ nhà với đội sân khách
* **Đội bóng có khả năng vô địch** cho mùa giải tiếp theo của giải đấu.
* **Đội bóng có hiệu số tốt nhất** qua từng mùa giải từ đó đánh giá sức mạnh của đội bóng đó với giải đấu đó.
* **Phân tích** các dữ liệu để đưa ra các kết luận về giải bóng đá được cho là hay nhất hiện tại này

1. **Phạm vi nghiên cứu:**

* Dự án này sẽ phân tích dữ liệu của EPl trong 23 mùa giải liên tiếp từ 1999/20 – 2022/23
* Dữ liệu được sư tầm trên github và các trang web bóng đá chính thống khác của EPL qua từng mùa giải.

1. **Phương pháp nghiên cứu:**

* Dự án sử dụng các phương pháp sau.
* **Thu thập dữ liệu**: Dữ liệu được thu thập từ các trang web github, sử dụng ngôn ngữ lập trình Python để đưa ra các dữ liệu sạch hơn so với dữ liệu gốc đến từ các trang web
* **Phân tích dữ liệu**: sử dụng ngôn ngữ Python và các thư viện như NumPy, PanDas, Maplotlib để đưa ra các phân tích về giải đấu

3

# **Phần 2: Phân tích dữ liệu**

1. **Mô tả dữ liệu**



**Season end**: Là ngày mùa giải kết thúc

**Round**: Là các vòng đấu của 1 mùa giải

**Date**: Ngày thi đấu

**Home**: là đội bóng chủ nhà **/** **HomeGoal**: là bàn thắng của đội chủ nhà

**Away**: là tên đội khách **/ AwayGoal**: là bàn thắng của đội khách

**FTR ( Full Time Result** ):

**H**: đội chủ nhà chiến thắng

**D**: hai đội có kết quả hòa

**A**: đội khách dành chiến thắng

*\* Dữ liệu vẫn có thể có những chỗ sai cho nên có thể dẫn đến việc kết quả của việc trả lời các câu hỏi có vài thông số không đúng, kết quả sẽ gần đúng so với kết quả thực tế*

*\* Bảng trên chỉ là một phần nhỏ của dữ liệu, nó chưa thể hiện được hết toàn bộ dữ liệu thực tế,*

1. **Các câu hỏi đặt ra cho dữ liệu**

**Câu 1**: Đội bóng dành nhiều chức vô địch nhất EPL.

**Câu 2**: Đội bóng vô địch liên tiếp nhiều lần nhất EPL.

**Câu 3**: Đội vô địch mà không thua trận đấu nào.

**Câu 4**: Chuỗi thắng liên tiếp dài nhất trong một mùa giải.

**Câu 5**: Chuỗi bất bại dài nhất trong một mùa

**Câu 6**: Tính trung bình số lượng bàn thắng và thua của mỗi CLB trong một mùa giải.

**Câu 7**: Tìm đội có tỷ lệ thắng cao nhất trong mùa giải.

**Câu 8**: Tìm CLB có nhiều trận không để thủng lưới nhất trong một mùa giải

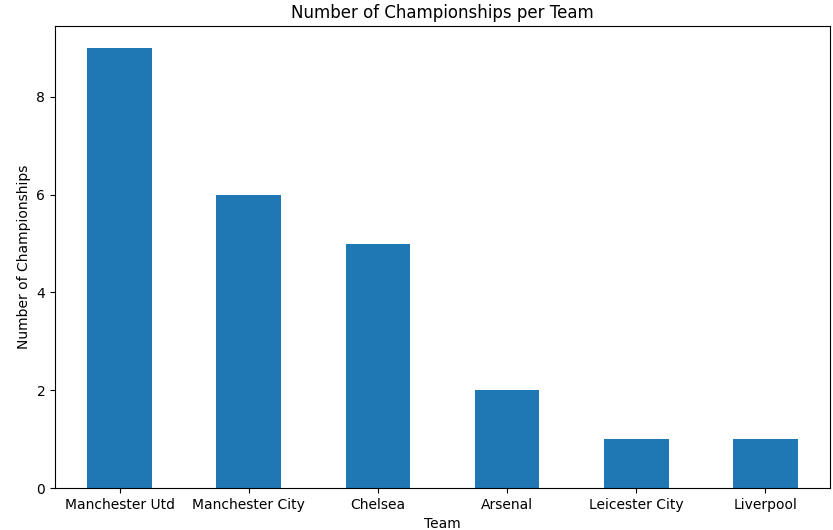
**Câu 9:** Trong 5 mùa giải gần nhất đưa ra thống kê số trận thắng và hiệu suất ghi bàn

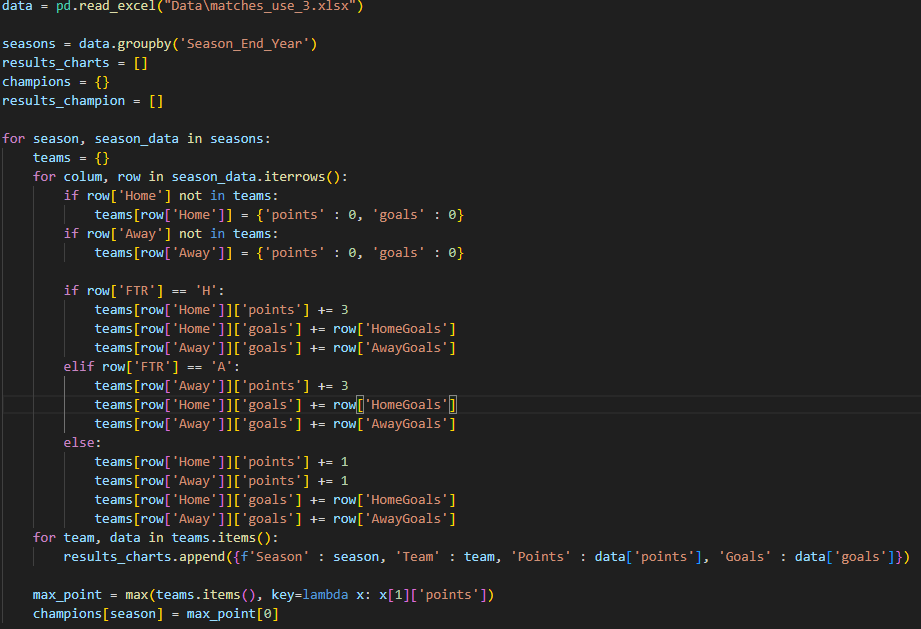
4

1. **Phân tích, Trực quan hóa dữ liệu từ các câu hỏi**

**Câu 1:** Đội bóng dành nhiều chức vô địch nhất EPL.

* Đội bóng vô địch là đội bóng có tổng số điểm cáo nhất tại mỗi mùa giải, trong trường hợp có đội bằng điểm nhau thì sẽ xét đến chỉ số phụ là hiệu số bàn thắng bại để phân hạng.
* Cách tính điểm như sau:
  + Thắng: 3 điểm
  + Hòa: 1 điểm
  + Thua: 0 điểm
* Ý tưởng bài toán:
  + Ta sẽ nhóm các dữ liệu theo các năm tạo ra list và dict để lưu thông tin
  + Duyệt qua dữ liệu và tính điểm theo trên
  + Cuối cùng ta sẽ tìm được dữ liệu mong muốn.
* Đoạn code làm sạch và vẽ sơ đồ dữ liệu



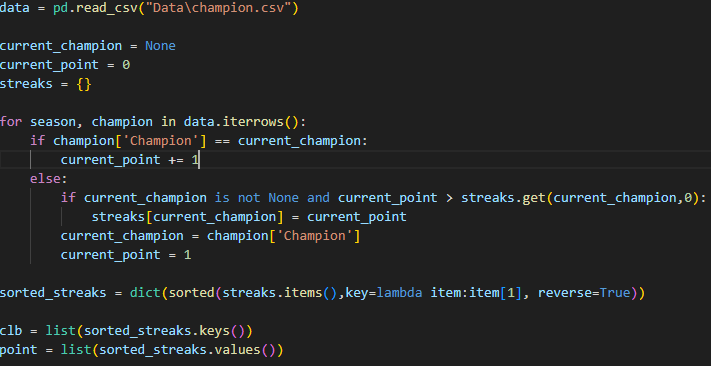


5

**Câu 2**: Các đội bóng vô địch liên tiếp nhiều lần nhất EPL.

* Để xác định đội bóng vô địch nhiều lần liên tiếp ta phải xác định được năm đầu tiên và năm kết thúc của chuỗi vô địch đó mới được tính là vô địch liên tiếp.
* Tiếp tục từ câu 1 ta xác định đội bóng vô địch liên tiếp nhiều nhất.
* Ý tưởng của bài toán:
  + Chúng ta sẽ sử dụng dữ liệu của câu 1 để không phải thức hiện truy xuất dữ liệu lại
  + Duyệt qua toàn bộ dữ liệu
  + Thêm đội hiện tại của vòng lặp vào nơi lưu trữ.
    - Nếu vòng lặp tiếp theo đối tượng vẫn giống trước thì tang biến đêm lên 1
    - Nếu đối tưởng khác đối tượng trước đó thì biến đếm giảm về 0
  + Kết thúc vong lặp ta tìm được danh sách các đội như dưới đây.
* Ta có được bảng thống kê và đoạn code của bài toán

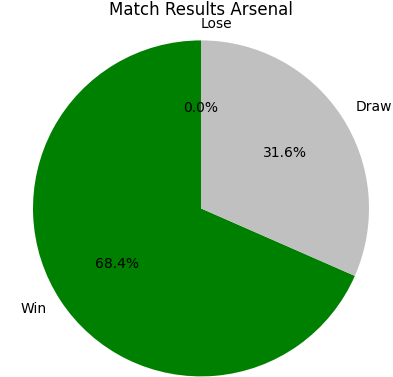




6

**Câu 3**: Đội bóng vô địch khi không thua trận đấu nào theo từng mùa giải.

* Để tìm được đội bóng theo yêu cầu của đề bài ta phải thỏa mãn hai điều kiện sau:
  + Là đội vô địch
  + Không thua trận đấu nào ( Có thể hòa )
* Ý tưởng cho câu hỏi:
  + Tìm kiếm các đội vô địch của các mùa giải
  + Kiểm tra các trận đấu nếu trong số 38 trận đấu ( bởi vì có 38 vòng đấu ) không có trận thua và đội đó vô địch thì thỏa mãn.
* Từ ý tưởng trên ta có biểu đồ dạng tròn và đoạn code như sau.

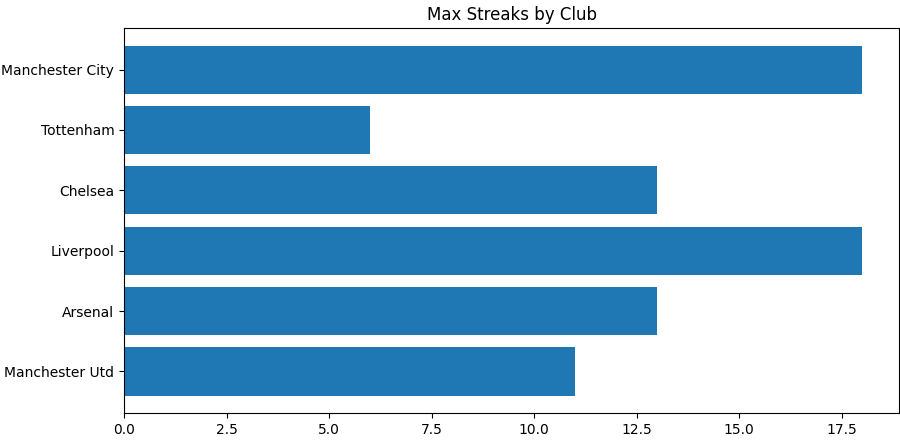




7

**Câu 4**: Các đội bóng có chuỗi thắng liên tiếp dài nhất trong một mùa giải.

* Chuỗi thắng là các trận đấu liên tiếp nhau chỉ thắng trong một mùa giải đội đó có thể không có chức vô địch và trận hòa sẽ không được chấp nhận
* Ý tưởng cho bài toán như sau:
  + Duyệt qua từng trận đấu.
  + Tìm kiếm và lưu trữ chuỗi thắng dài nhất và chuỗi thắng hiện tại
  + Cuối cùng duyệt qua chuỗi thắng dài nhất sẽ tìm được kết quả
* Đây là kết quả sau khi làm sạch dữ liệu và vẽ biểu đồ dạng cột.

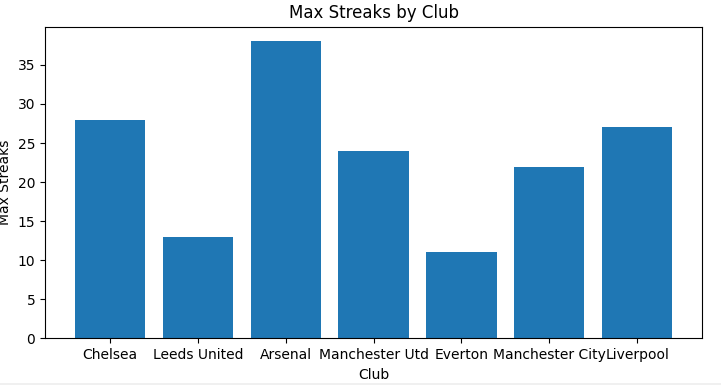
****

*\*\* Lưu ý: Biểu đồ trên sẽ so sánh đội có chuỗi thắng cao nhất, ví dụ nếu MU có 2 chuỗi thắng là 11 và 8 thì chúng ta sẽ chọn 11 để đem ra so sánh với các CLB khác.*

8

**Câu 5**: Các đội bóng có chuỗi bất bại dài nhất trong một mùa.

* Chuỗi bất bại khác với chuỗi thắng là những trận đấu hòa vẫn sẽ được tính vào kết quả
* Ý tưởng cho bài toán như sau:
  + Duyệt qua từng trận đấu trong dữ liệu
  + Tạo các biến để lưu trữ về trận thắng hoặc hòa dài nhất. nếu có trận bại chuỗi kết thúc.
  + Cuối cùng duyệt qua tất cả thành phần trong chuỗi của một mùa giải sẽ tìm được đội có chuỗi cao nhất trong một mùa giải
* Trong các mùa giải, vì có nhiều đội cho nên chúng ta sẽ chỉ chọn ra 6 đội được coi là 6 đội bóng lớn ở EPL.
* Ta có đoạn làm sạch dữ liệu và vẽ biểu đồ cột như sau.



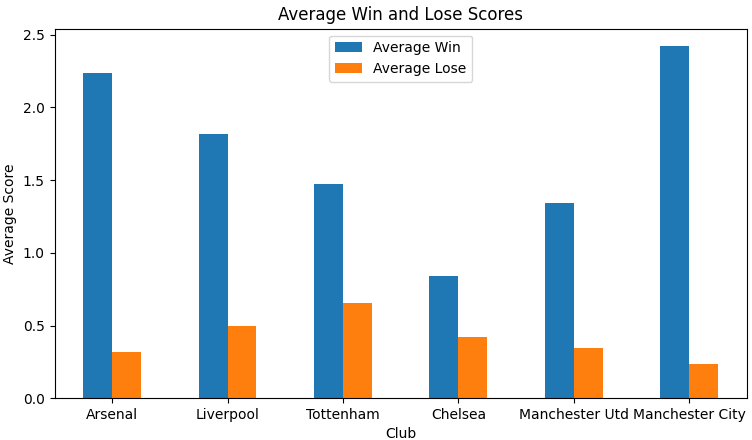
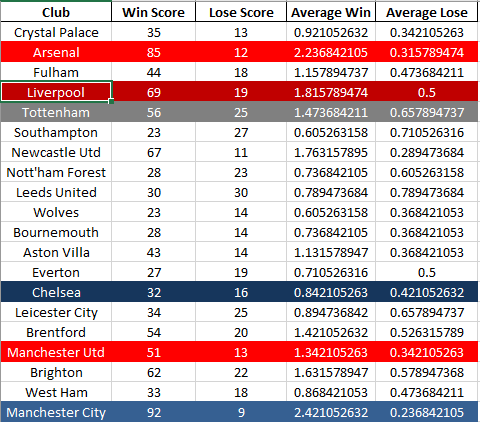


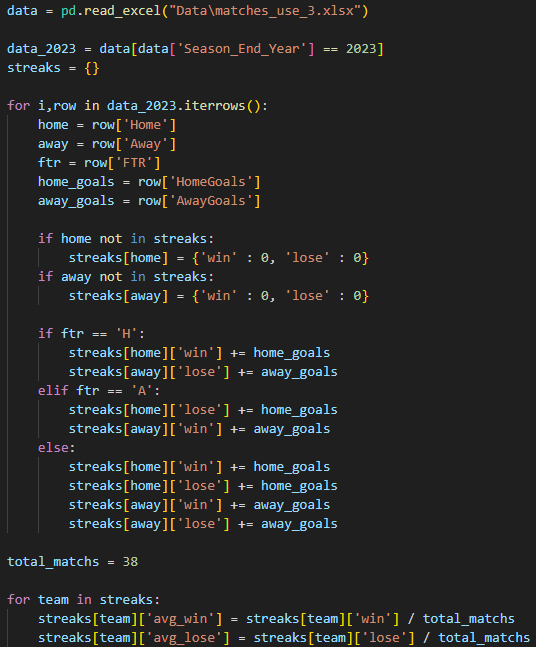
*\*\* Lưu ý: Biểu đồ trên sẽ so sánh đội có chuỗi thắng cao nhất, ví dụ nếu ARS có 2 chuỗi bất bại là 38 và 21 thì chúng ta sẽ chọn 38 để đem ra so sánh với các CLB khác.*

9

**Câu 6**: Tính trung bình số lượng bàn thắng và thua của mỗi CLB trong một mùa giải.

* Ý tưởng cho bài toán.
  + Tính tổng số tất cả các bàn thắng và bàn thua trong một mùa giải
  + Chúng ta sẽ tính được trung bình bàn thắng và bàn thua của đội và xây dựng biểu đồ so sánh giữa các bàn thắng và thua
  + Vì có nhiều mùa giải cho nên chúng ta sẽ chọn mùa giải gần nhất và xây dựng biểu đồ của 6 đội bóng lớn ở EPL
* Ta có biểu đồ và dữ liệu được làm sạch

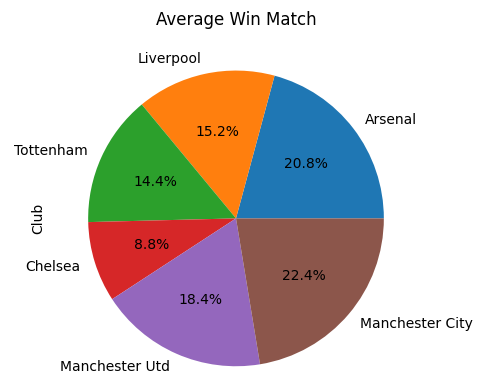
 

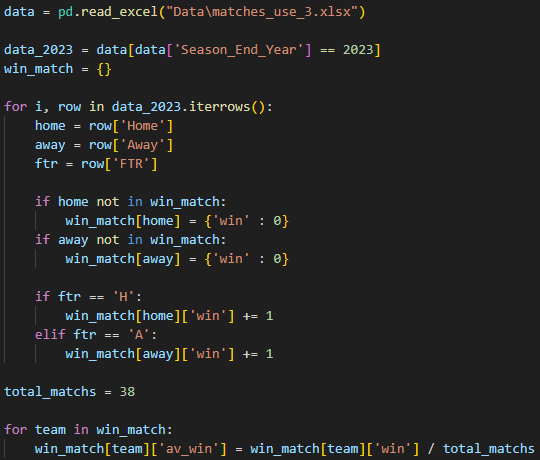


10

**Câu 7**: Tìm các đội có tỷ lệ thắng cao nhất trong mùa giải.

* Ý tưởng bài toán
  + Xác định số trận dành chiến thắng của các đội
  + Bởi vì có 38 vòng đấu cho nên số trận các đội phải thi đấu mặc định là 38 trận từ đó ta tính được tỷ lệ thắng cao nhất trong mùa giải
  + Giống câu trên ta sẽ tính tỉ lệ của năm 2023 với 6 đội bóng lớn ở EPL.
* Ta có dữ liệu sau khi làm sạch và biểu đồ so sánh sau

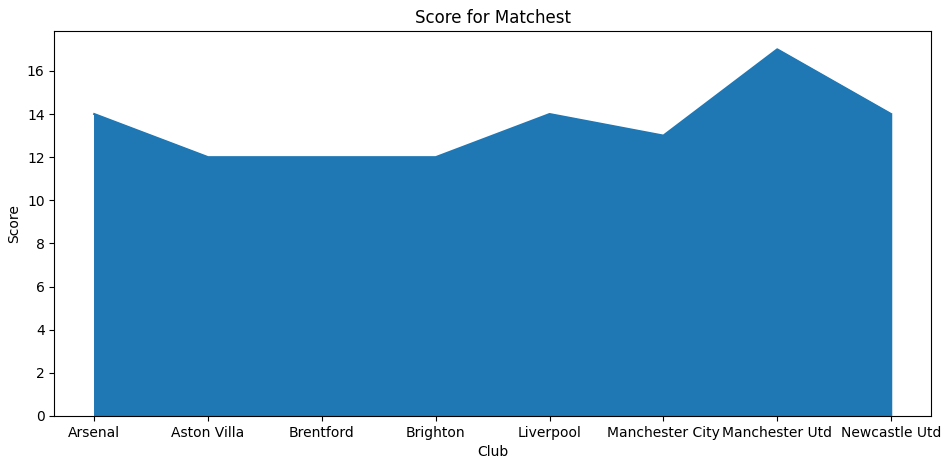
 

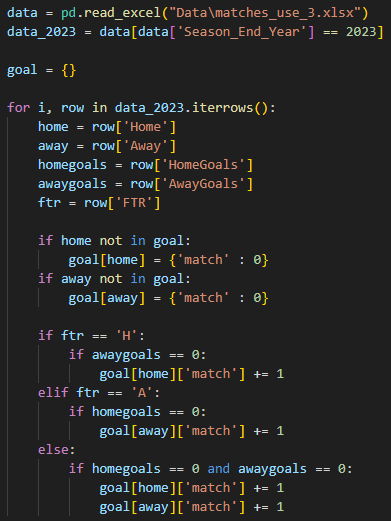


11

**Câu 8**: Tìm CLB có nhiều trận không để thủng lưới nhất trong một mùa giải

* Ý tưởng bài toán:
  + Duyệt qua các trận đấu trong mùa giải, tìm kiếm các trận thắng hoặc hòa của từng đội mà bàn thắng của đội bạn = 0, sẽ thỏa mãn yêu cầu.
  + Chúng ta vẫn dùng dữ liệu của năm 2023.
* Ta có dữ liệu được làm sạch và bảng thống kê các đội bóng cao nhất

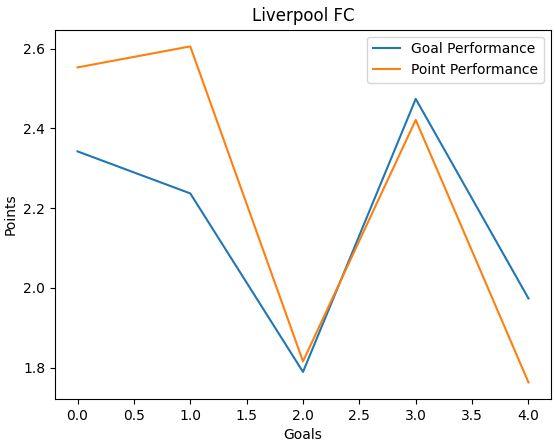
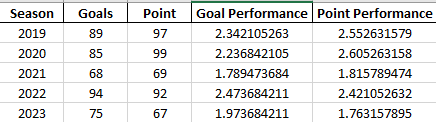
 

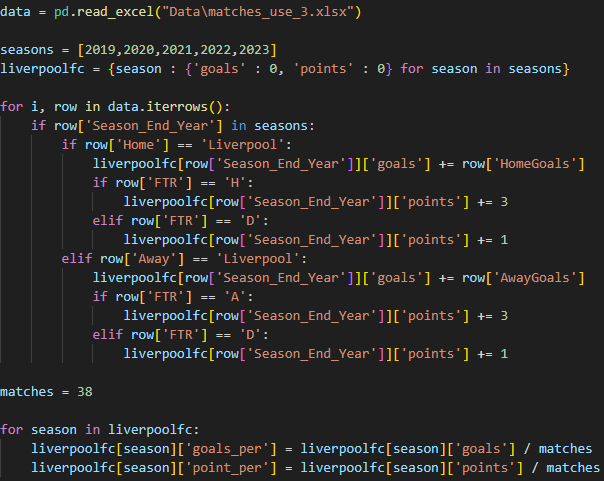


12

**Câu 9:** Trong 5 mùa giải gần nhất đưa ra thống kê số trận thắng và hiệu suất ghi bàn của CLB Liverpool.

* Ý tưởng bài toán
  + Duyệt qua các trận đấu của Liverpool trong 5 mùa giải
  + Chúng ta sẽ biết được điểm và số bàn thắng mà họ ghi trong 5 mùa giải, từ đó tìm được hiệu suất bàn thắng theo từng mùa
* Ta có đồ thị và dữ liệu sau.



13

1. **Kết quả phân tích.**

* Dựa vào các kết quả ở trên ta phân tích được các dữ liệu về toàn bộ các CLB ở giải và các thông số họ mang lại qua từng mùa
* Và ta biết được sự thống trị của 6 đội bóng lớn ở EPL thay nhau vô địch qua từng mùa giải, chỉ một vài mùa giải thì các đội mới có thể chen chân vào sự tranh đâu của 6 đội bóng đó.
* Kết quả phân tích dữ liệu được sử dụng hoàn toàn bằng ngôn ngữ lập trình Python với các thư viện Pandas ( để truy xuất và lọc các dữ liệu cần thiết cho câu hỏi ) và Matplotlib ( để vẽ các đồ thị thể hiện các thông số đó ). Không có một ngôn ngữ lập trình khác nào ngoài Python có ở trong đây

# **PHẦN 4: KẾT LUẬN**

1. **Đánh giá dữ liệu**
   * Dữ liệu trên là một dang dữ liệu đưa ra các thông số về các trận đấu của các CLB ở EPL, có một vài khung dữ liệu bị thiếu so với khung dữ liệu ngoài đời, cho nên dẫn đến việc một vài chỗ ở trong các câu hỏi có thể dẫn đến sai sót về kết quả của các câu hỏi.
2. **Kiến thức rút ra được từ bài tập**
   * Python là một ngôn ngữ tuyệt vời nó giúp em học được cách truy xuất các dữ liệu theo các điều kiện cụ thể hoặc theo kiểu các nhóm dữ liệu bằng cách sử dụng thư viện của Pandas.
   * Em cũng học được cách trực quan hóa các dữ liệu đó ra dạng đồ thị ( nhiểu kiểu đồ thị ) bằng cách sự dụng thư viện Matplotlib của Python, dung nó để trực quan hóa các dữ liệu số đưa ra một kiểu thống kê dễ nhìn hơn so với dữ liệu số
3. **Các nguồn tài liệu tham khảo về môn học.**
   * Em có tham khảo một số tài liệu về các thư viện em đã sử dụng để biết thêm nhiều hơn và dễ dàng sử dụng các thư viện ấy, các trang web gồm.
     + W3schools.com
     + Stackoverflow.com
     + Github.com
     + Python.org

14