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

**KHOA CÔNG NGHỆ THÔNG TIN**

**-----🙠🕮🙢-----**

**BÁO CÁO BÀI TẬP LỚN**

**Lập trình Python**

**Sinh viên thực hiện:**

**Họ và tên : Lương Minh Quý**

**Mã SV: B22DCKH099**

**Hà Nội, ngày 3/11/2023**

* Câu 1: Thu thập dữ liệu thống kê [\*] của tất cả các cầu thủ có số phút thi đấu nhiều hơn 90 phút tại giải bóng đá ngoại hạng Anh mùa 2023-2024.
* Ý tưởng :
  + Sử dụng thư viện BeautifulSoup và requests để lấy dữ liệu từ trên trang web về
  + Cách cài đặt : dung lệnh “pip install BeautifulSoup requests” để cài đặt hai thư viện
  + Ta chia những chỉ số thành nhiều phần trong cùng 1 table để dễ dàng sử lý
  + Ta có thể chia các chỉ số thành 10 phần sau:

+ Base\_data

+ Goalkeep\_data

+ Shooting\_data

+ Passing\_data

+ Goalshot\_data

+ Defensive\_data

+ Posess\_data

+ Playtime\_data

+ Miscell\_data

* Đầu tiên , ta thu thập dữ liệu về các đội bóng tham gia ngoại hạng Anh bằng 2 thue viện nói trên (HÌnh 1.1)

A screen shot of a computer code

Description automatically generated

(Hình 1.1)

* Tiếp theo tạo 10 danh sách ứng vs 10 phần dữ liệu đã được chia ra (Hình 1.2)

A screen shot of a computer

Description automatically generated

(Hình 1.2)

* Từ danh sách các đội tìm được , tiến hành lấy thông tin từ các đội ở từng bảng và đưa vào danh sách , mỗi data sẽ tương tứng vs 1 hàm riêng , minh họa hàm tìm base\_data (HÌnh 1.3)

A screen shot of a computer program

Description automatically generated

(Hình 1.3)

* Sau khi tìm được tất cả các dữ liệu cần thiết , t chuyển tất cả các list tìm được thành dataframe và đưa hết vào 1 list chứa , minh họa cho vc chuyển list base\_data thành df và đưa vào list chưa, những hàm khác tương tự (HÌnh 1.4)

A screen shot of a computer code

Description automatically generated

(Hinh 1.4)

* Cuối cùng , ta lấy tổng hợp tất cả các df bằng outer join đòng thời loại bỏ trùng lặp và them điều kiện số phút thi đấu phải lớn hơn 90 (Hình 1.5)

A screen shot of a computer program

Description automatically generated

(Hình 1.5)

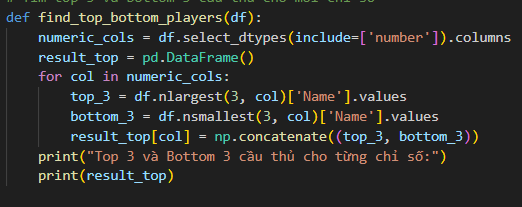
* Câu 2 : Từ những thông tin từ file results.cvs , thực hiện các yêu cầu
* TÌm top 3 cầu thủ có điểm cao nhất và thấp nhất ở mỗi chỉ số (Hình 2.1)

+ Ý tưởng :

\_ Đầu tiên ta lọc ra các column có kiểu số

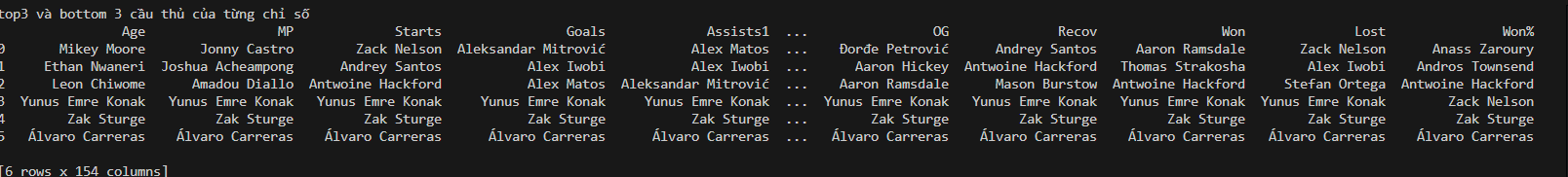
\_ TIếp đó tìm ra 3 cầu thủ có chỉ số cao nhất và thấp nhất ứng với tùng column và đưa vào list\_column

\_tạo 1 dataframe được tạo nên từ các list column



(Hình 2.1)

+ Kết quả sau khi chạy code :



* Tìm trung vị của mỗi chỉ số. Tìm trung bình và độ lệch chuẩn của mỗi chỉ số cho các cầu thủ trong toàn giải và của mỗi đội

+ Ý tưởng :

\_ Đối với toàn bộ các cầu thủ thì ta có 3 bảng tìm về median,mean,std đối với mỗi chỉ số, sau khi tìm được thì t chỉ cần nối column của chúng lại tạo thành 1 dataframe1 mới (Hình 2.2)

A computer screen shot of text

Description automatically generated

(Hình 2.2)

\_ Còn đối vs đội cũng vậy , ta cũng có 3 bảng về median,mean,std đối vs mỗi chỉ số được group theo “team” , từ đó ra nối column chúng lại tạo thành 1 dataframe2 mới (Hình 2.3)

A computer code on a black background

Description automatically generated

(Hình 2.3)

\_Cuối cùng ta nối df1 và df2 theo hang bằng hàm concat và them dữ liệu vào file result2.csv (Hình 2.4)

A black background with white text

Description automatically generated

(Hình 2.4)

* Vẽ historgram phân bố của mỗi chỉ số của các cầu thủ trong toàn giải và mỗi đội (Hình 2.5)

+ Dầu tiên ta vẽ biểu đồ cho toàn giải (Hình 2.5) , sau đó ta tạo them thư mục để lưu trữ từng biểu đồ một (HÌnh 2.5)

A computer screen shot of text

Description automatically generated

(Hình 2.5)

* Tìm các đội bóng có chỉ số cao nhất ở mỗi chỉ số

+ Ý tưởng t duyêt qua từng chỉ số của df1 (được tạo ra từ vc grouby ”team” với phương thức mean) và lưu lại vào trong list tên đội , chỉ số, giá trị max của chỉ số đó

A computer screen with colorful text

Description automatically generated

(Hình 2.6)

A screenshot of a computer screen

Description automatically generated

(HÌnh 2.7 : bảng df\_results)

* Từ hình 2.6 , ta thấy rằng việc tiếp theo là tìm tần xuất điểm của các đội bằng cách dung hàm couter để đếm tần xuất và sort để tìm giá trị lớn nhất

+ Kết quả của thu được sau khi chạy code : 

* Manchester City là đội có phong độ tốt nhất trong giải ngoại hạng Anh
* Trong hàm main , ta vt lại các hàm (Hinh 2.8)

A screen shot of a computer code

Description automatically generated

(Hình 2.8)

* Câu 3:
* Sử dụng thuật toán K-means để phân loại các cầu thủ thành các nhóm có chỉ số giống nhau.
* Sử dụng thuật toán PCA, giảm số chiều dữ liệu xuống 2 chiều, vẽ hình phân cụm các điểm dữ liệu trên mặt 2D.

+ Hàm vẽ biểu đồ (Hinh 3.1)

A computer screen with text

Description automatically generated

(HÌnh 3.1.1)

+ Hàm chuẩn bị dữ liệu (HÌnh 3.1.2)

A screen shot of a computer code

Description automatically generated

(Hình 3.1.2)

+ Hàm chạy K-means tùy chỉnh (HÌnh 3.1.3)

A computer screen shot of text

Description automatically generated

(Hình 3.1.4)

+ Trong hàm main (Hình 3.1.4)

A computer screen with text and numbers

Description automatically generated

(Hình 3.1.4)

+ Sau khi chạy trương trình (Hình 3.1.5)

A diagram of colored dots

Description automatically generated

(Hình 3.1.5)

* Viết chương trình python vẽ biểu đồ rada (radar chart) so sánh cầu thủ :

+ hàm vẽ biểu đồ (Hình 3.2.1)

A screen shot of a computer program

Description automatically generated

(Hình 3.2.1)

+ Hàm main (Hình 3.2.2)

A computer screen with text on it

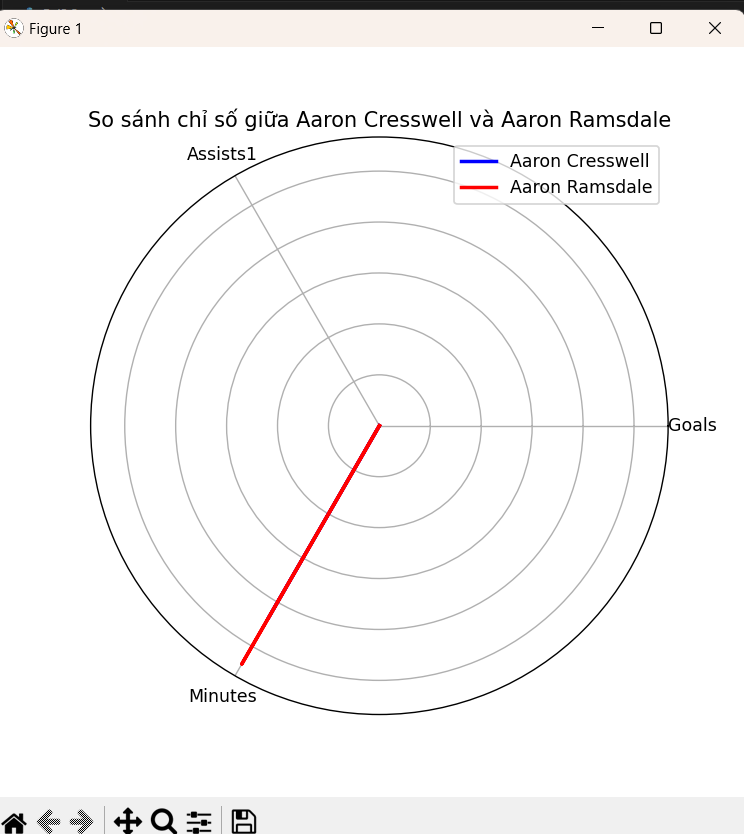
Description automatically generated

(Hình 3.2.2)

* Sau khi chạy lệnh ta dung lệnh <Đường dần file> <tên file> --p1 "Tên Cầu Thủ 1" --p2 "Tên Cầu Thủ 2" --Attribute "ChỉSố1,ChỉSố2,...,ChỉSốN"

\_ Cụ thể sau khi dung lệnh :

C:/Users/luong/AppData/Local/Programs/Python/Python312/python.exe Bai3,2.py --p1 "Aaron Cresswell" --p2 "Aaron Ramsdale" --Attribute "Goals, Assists1, Minutes" (đã được bôi xanh dòng cuối cùng) trên terminal (Hình 3.3.3)



(Hình 3.3.3)

* Bài 4 : Thu thập giá chuyển nhượng của các cầu thủ trong mùa 2023-2024 từ trang web
* Ý tưởng :
  + Cũng như bài 1 , ta dung hai thự viện requests và BeautifulSoup để thu thập và làm sạch dữ liệu trên web
  + Đầu tiên ta lấy dữ liệu của các team
  + TIếp theo , từ từng team ta lấy dữ liệu của các cầu thủ trong team đấy (GIá chuyển nhượng)
* **fetch\_teams(url)** (Hình 4.1)

+ Lấy URL của giải đấu và tải nội dung HTML về.

+ Tìm bảng chứa danh sách các đội và lấy tên đội cùng liên kết đến trang chi tiết của mỗi đội.

A screen shot of a computer program

Description automatically generated

(Hình 4.1)

* **fetch\_players(team\_name, team\_url)** (HÌnh 4.2)

+ Với mỗi đội, hàm này tải về trang chi tiết của đội đó, tìm bảng chứa danh sách cầu thủ và giá chuyển nhượng.

+ Tạo một danh sách các cầu thủ, trong đó mỗi phần tử chứa tên cầu thủ, tên đội, và chi phí chuyển nhượng.

A computer screen shot of text

Description automatically generated

(Hình 4.2)

* **main()** (Hình 4.3)

+ Gọi fetch\_teams() để lấy danh sách đội bóng và các URL tương ứng.

+ Duyệt qua từng đội và gọi fetch\_players() để lấy dữ liệu về cầu thủ.

+ Kết hợp tất cả dữ liệu vào all\_players\_data và tạo DataFrame.

+ In DataFrame ra màn hình hoặc lưu vào file CSV nếu cần.



(Hình 4.3)