HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÁO CÁO
BÀI TẬP LỚN PYTHON

Họ và tên sinh viên : **Mạc Đức Duy**
Mã sinh viên : **B22DCCN129**
Lớp : **D22CQCN05-B**

Hà Nội – 2024

Mục Lục

I. Thu thập dữ liệu

***Yêu cầu**: Viết chương trình Python 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.*

***Code**: ThuThapDuLieu.py*

***Kết quả**: results.csv*

Ý tưởng :

- Sử dụng các thư viện chính:

  + **Selenium**: Tự động mở trang web và lấy nội dung.
  + **beautifulSoup**: Tìm và phân tích các bảng dữ liệu trong HTML.
  + **pandas**: Lưu dữ liệu thành bảng và xuất ra tệp CSV.

- Truy cập trang web thông qua link.

- Tìm bảng thống kê dữ liệu trong web thông qua các tag trong HTML, nếu bảng có tồn tại tiến hành trích xuất các dữ liệu có trong bảng. Có nhiều bảng dữ liệu nhỏ, cần tìm riêng từng bảng rồi gộp lại thành 1 bảng thống kê hoàn chỉnh.

- Sau khi đã thu thập được bảng dữ liệu, tiến hành tinh chỉnh lại theo yêu cầu đề bài (sắp xếp tên, điền các ô dữ liệu trống…) và xuất ra file csv.

=> Kết quả file ThuThapDuLieu.py thu được bảng dữ liệu có 172 trường dữ liệu với gần 500 cầu thủ. Dưới đây là một phần của file results.csv:

Name	Nation	Team	Position	Age	Playing time_Matches played	Playing time_Starts	Playing time_Minutes	Performance_non-Penalty Goals
Aaron Cresswell	ENG	West Ham	DFFW	33	11	4	436	0
Aaron Hickey	SCO	Brentford	DF	21	9	9	713	0
Aaron Ramsdale	ENG	Arsenal	GK	25	6	6	540	0
Aaron Ramsey	ENG	Burnley	MFFW	20	14	5	527	0
Aaron Wan-Bissaka	ENG	Manchester Utd	DF	25	22	20	1780	2
Abdoulaye DoucourÃ©	MLI	Everton	FWMF	30	32	32	2629	8
Adam Lallana	ENG	Brighton	MFFW	35	25	13	850	1
Adam Smith	ENG	Bournemouth	DF	32	28	25	2150	2
Adam Webster	ENG	Brighton	DF	28	15	13	1144	0
Adam Wharton	ENG	Crystal Palace	MF	19	16	15	1297	3
Adama TraorÃ©	ESP	Fulham	FWMF	27	17	1	377	5
Albert Sambi Lokonga	BEL	Luton Town	MF	23	17	16	1303	4
Alejandro Garnacho	ARG	Manchester Utd	FW	19	36	30	2565	11
Alex Iwobi	NGA	Everton	MF	27	2	2	140	0
Alex Iwobi	NGA	Fulham	FWMF	27	30	25	2192	7
Alex Scott	ENG	Bournemouth	MF	19	23	11	1014	2
Alexander Isak	SWE	Newcastle Utd	FW	23	30	27	2255	23
Alexis Mac Allister	ARG	Liverpool	MF	24	33	31	2599	10
Alfie Doughty	ENG	Luton Town	DF	23	37	34	2925	10

II. Truy vấn thông tin

***Yêu cầu 1**: Tìm top 3 cầu thủ có điểm cao nhất và thấp nhất ở mỗi chỉ số.*

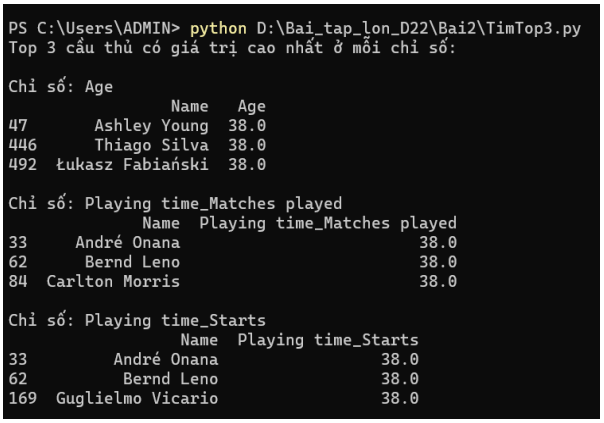
***Code**: TimTop3.py*

***Kết quả**: Terminal*

Ý tưởng:

- Sử dụng thư viện **pandas** để xử lý DataFrame.
- Truy cập vào results.csv ở phần I.
- Tiến hành lọc ra các bảng dữ liệu là số (số nguyên hoặc số thực).
- Sử dụng nlargest() và nsmallest() để tìm top của các chỉ số và in ra màn hình.

=> Kết quả file TimTop3.py in ra màn hình lần lượt là top 3 cao nhất các chỉ số, top 3 thấp nhất các chỉ số. Dưới đây là kết quả của 3 chỉ số đầu tiên:



```
PS C:\Users\ADMIN> python D:\Bai_tap_lon_D22\Bai2\TimTop3.py
Top 3 cầu thủ có giá trị cao nhất ở mỗi chỉ số:

Chỉ số: Age
                Name   Age
47       Ashley Young  38.0
446      Thiago Silva  38.0
492  Łukasz Fabiański  38.0

Chỉ số: Playing time_Matches played
             Name  Playing time_Matches played
33     André Onana                        38.0
62      Bernd Leno                        38.0
84  Carlton Morris                        38.0

Chỉ số: Playing time_Starts
                Name  Playing time_Starts
33         André Onana                38.0
62          Bernd Leno                38.0
169  Guglielmo Vicario                38.0
```

```
Top 3 cầu thủ có giá trị thấp nhất ở mỗi chỉ số:

Chỉ số: Age
              Name   Age
277  Leon Chiwome  17.0
284   Lewis Miley  17.0
120    David Ozoh  18.0

Chỉ số: Playing time_Matches played
              Name  Playing time_Matches played
13      Alex Iwobi                        2.0
188     Ionuț Radu                        2.0
320  Matheus Nunes                        2.0

Chỉ số: Playing time_Starts
                  Name  Playing time_Starts
120          David Ozoh                0.0
191       Ivan Perišić                0.0
226  Jesurun Rak Sakyi                0.0
```

***Yêu cầu 2**: Tìm trung vị (median), trung bình (mean) và độ lệch chuẩn (std) của mỗi chỉ số cho các cầu thủ trong toàn giải và của mỗi đội.*

***Code**: ChiSoTeam*

***Kết quả**: results2.csv*

Ý tưởng:

- Sử dụng thư viện **pandas** để xử lý DataFrame và xuất ra file csv.

- Truy cập vào results.csv ở phần I.

- Tiến hành lọc ra các bảng dữ liệu là số (số nguyên hoặc số thực).

- Tạo khung với cấu trúc: các cột chỉ số chia làm 3 cột nhỏ: Median of [chỉ số], Mean of [chỉ số], STD of [chỉ số]; các hàng tổ đội bao gồm All (toàn giải) và các đội.

- Sử dụng median(), mean(), std() để tính kết quả và xuất ra file csv.

=>Kết quả file ChiSoTeam.py thu được bao gồm 1 hàng all và 20 hàng các tổ đội. Dưới đây là một phần của file results2.csv, chứa 2 chỉ số đầu tiên:

Team	Median Of Age	Mean Of Age	Std Of Age	Median Of Playing time_matches played	Mean Of Playing time_matches played	Std Of Playing time_matches played
All	25	25.49899	4.127355	23	22.65720081	10.1369752
West Ham	27.5	28.27273	3.869069	23.5	23.36363636	10.82565495
Brentford	26	25.8	3.593976	26	22.96	10.34601373
Arsenal	24	24.7619	2.547641	27	26.80952381	10.1912661
Burnley	24	24.07143	3.838678	16	20.39285714	9.346575074
Manchester Utd	25.5	25.26923	4.414138	22	21.5	10.05286029
Everton	26	26.34783	4.858064	28	23.30434783	11.56182898
Brighton	23.5	24.78571	5.698324	20	20.92857143	8.751417119
Bournemouth	24.5	25.03846	3.538144	25.5	22.07692308	11.85216631
Crystal Palace	25.5	25.16667	4.280051	22.5	22.45833333	9.477566954
Fulham	27	27.90476	3.36013	29	27.23809524	7.993151831
Luton Town	26	26.32	3.051229	23	22.84	9.163696488
Newcastle Utd	25.5	26.125	4.87507	21	22.875	8.679373349
Liverpool	24	25.31818	3.822071	28	25.86363636	8.993624485
Chelsea	22	23	3.905125	23	21.88	9.404431579
Sheffield Utd	24	25.16667	4.25954	14.5	18.8	10.58430849
Nott'ham Forest	25.5	25.9	3.880544	20	19	9.955071485
Tottenham	25.5	25.125	3.53015	27.5	23.75	10.92762753
Manchester City	27	26	4.024922	29	24.95238095	9.351343168
Aston Villa	26	25.95652	3.548089	27	24.17391304	11.10958707
Wolves	24	24.68	4.422669	25	22.48	11.93077254

***Yêu cầu 3**: Vẽ historgram phân bố của mỗi chỉ số của các cầu thủ trong toàn giải và mỗi đội.*

***Code**: Historgram_all.py (đối với phạm vi toàn giải)*

Histogram_team.py (đối với phạm vi các đội)

***Kết quả**: Terminal*

Ý tưởng:

- Sử dụng thư viện

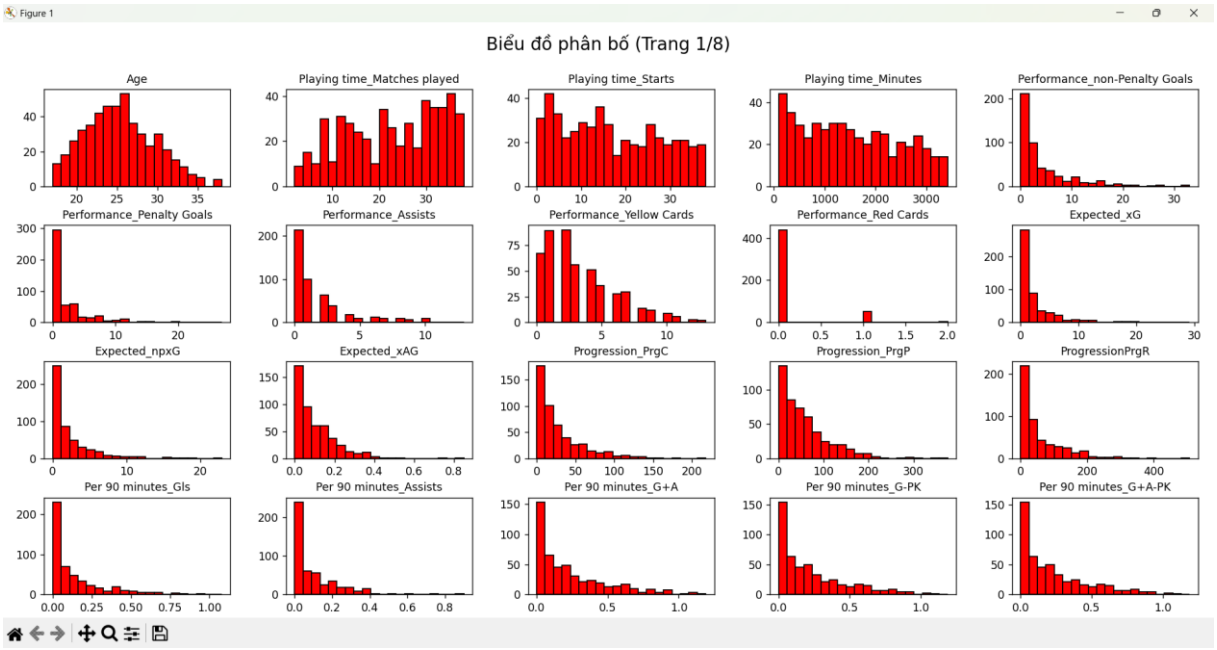
  + **Pandas:** Để xử lý DataFrame.
  + **matplotlib**: Tạo biểu đồ.

- Truy cập vào results.csv ở phần I.

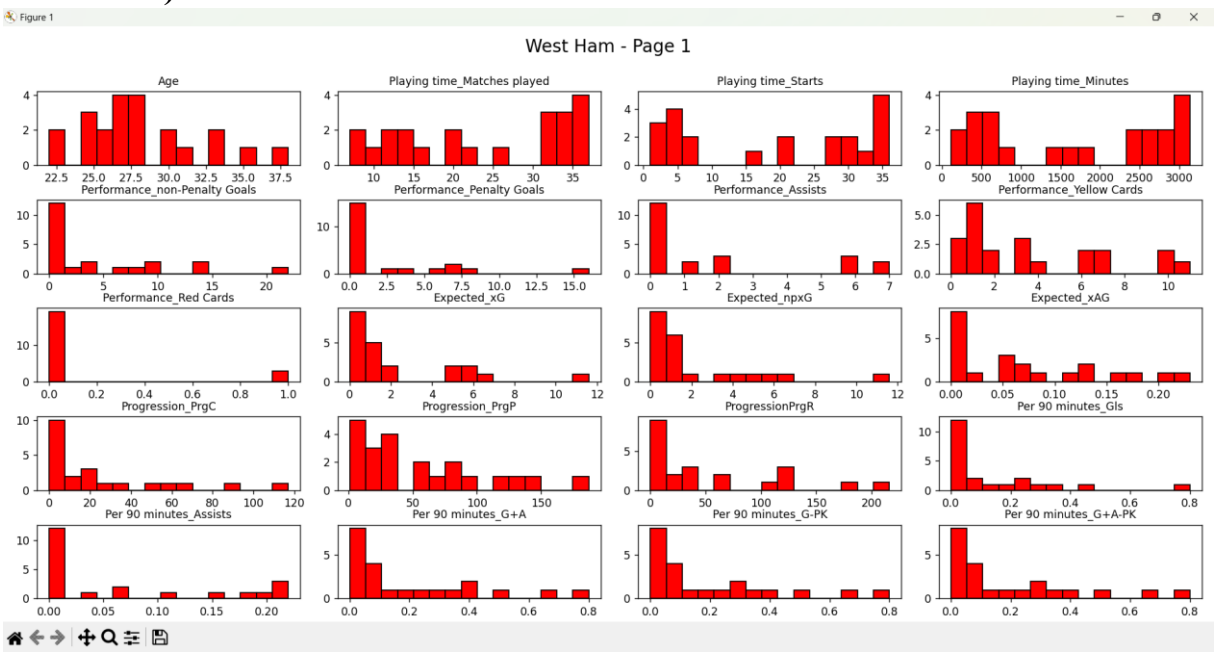
- Tiến hành lọc ra các bảng dữ liệu là số (số nguyên hoặc số thực).

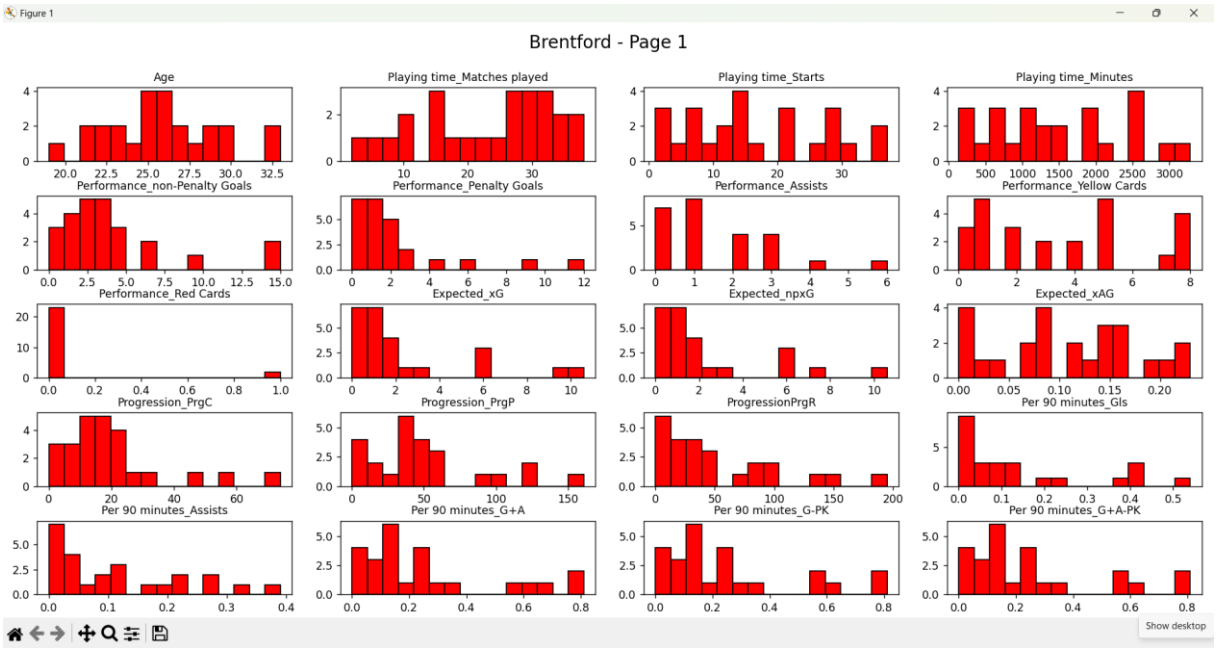
- Vì số lượng biểu đồ lớn nên sẽ hiển thị 20 biểu đồ trong 1 trang, cài đặt các thông số khác để trang hiển thị được hợp lý. Đối với phạm vi các đội thì sẽ lọc các chỉ số dựa theo Team.

=> Kết quả file Historgram_all.py thu được 8 trang chứa các biểu đồ. Dưới đây là trang đầu tiên:



=> Tương tự, kết quả file Histogram_team.py thu được 20 team, mỗi team có 8 trang biểu đồ. Dưới đây trang đầu của 2 đội đầu tiên (West Ham và Brentford):





***Yêu cầu 4**: Tìm đội bóng có chỉ số điểm số cao nhất ở mỗi chỉ số. Theo bạn đội nào có phong độ tốt nhất giải ngoại Hạng Anh mùa 2023-2024.*

***Code**: TopChiSo.py*

***Kết quả**: Terminal*

Ý tưởng:

- Sử dụng thư viện **pandas** để xử lý DataFrame.

- Truy cập vào results.csv ở phần I.

- Tiến hành lọc ra các bảng dữ liệu là số (số nguyên hoặc số thực).

- Sử dụng max() để tìm top 1.

=> Kết quả file TopChiSo.py lần lượt in ra top 1 các chỉ số dưới dạng Chỉ số -Team – Số điểm. Dưới đây là một phần kết quả với 5 chỉ số đầu tiên:

```
PS C:\Users\ADMIN> python D:\Bai_tap_lon_D22\Bai2\TopChiSo.py
Đội bóng có điểm số cao nhất ở mỗi chỉ số:
Age: Everton với điểm số 38.0
Playing time_Matches played: Manchester Utd với điểm số 38.0
Playing time_Starts: Manchester Utd với điểm số 38.0
Playing time_Minutes: Manchester Utd với điểm số 3420.0
Performance_non-Penalty Goals: Chelsea với điểm số 33.0
```

=> Dựa vào kết quả, đội phong độ nhất giải là <u>Manchester City</u> vì xuất hiện trên top các chỉ số nhiều nhất (trừ chỉ số Age).

III. Một số thuật toán Machine Leaning

***Yêu cầu 1**: Sử dụng thuật toán K-means để phân loại các cầu thủ thành các nhóm có chỉ số giống nhau. Nên phân loại cầu thủ thành bao nhiêu nhóm? Vì sao?*

***Code**: Kmeans.py*

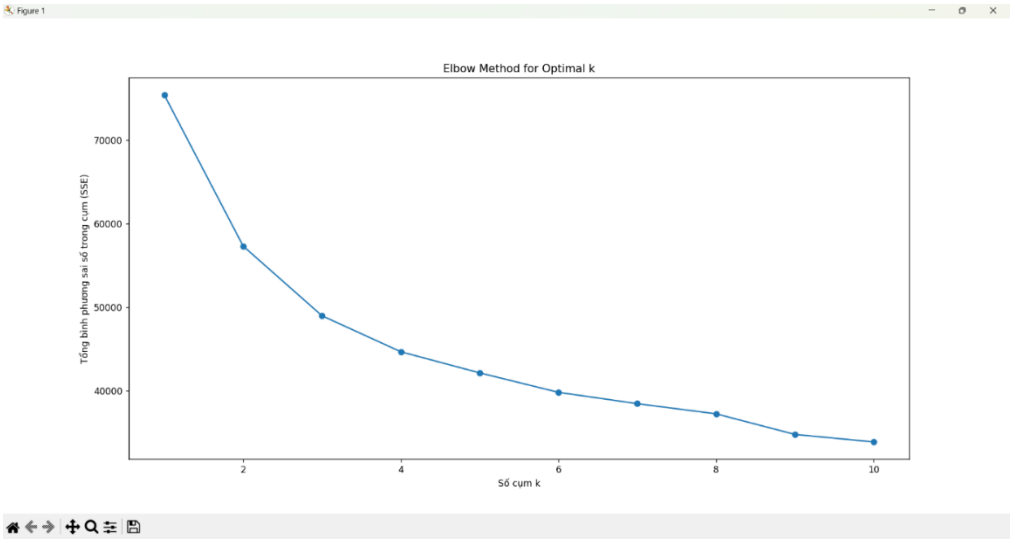
Elbow.py

***Kết quả**: Terminal*

<u>Ý tưởng:</u>

- Phân cụm bằng phương pháp Elbow: Dựa vào việc tính tổng bình phương sai số trong cụm, khi giá trị tổng này giảm không đáng kể nữa thì số cụm là hợp lý.

=> Kết quả file Elbow.py:



=> Nhận xét: Dựa vào biểu đồ, ta thấy điểm "elbow" nằm ở vị trí số 3. Trước vị trí này, ta thấy biểu đồ có độ dốc lớn. Sau vị trí này, biểu đồ có dốc thoải, cho thấy mức giảm có giảm dần nhưng không còn đột ngột, dù có thêm cụm cũng không cải thiện nhiều cho độ chính xác. <u>Vì vậy chọn số cụm bằng 3.</u>

- Thuật toán K-means phân loại cầu thủ thành các nhóm có chỉ số giống nhau. Dựa vào kết quả của thuật toán phân cụm trước, ta tiến hành viết thuật toán với số cụm là 3.

=> Kết quả file Kmeans.py với nhóm đầu tiên của thuật toán:

```
PS C:\Users\ADMIN> python D:\Bai_tap_lon_D22\Bai3\Kmeans.py
Phân nhóm các cầu thủ dựa trên các chỉ số:

Nhóm 1:
                  Name   Age  ...  Miscellaneous Stats_Aerial Duels_Lost  Miscellaneous Stats_Aerial Duels_Won%
4      Aaron Wan-Bissaka  25.0  ...                                 19.0                                   52.5
7            Adam Smith  32.0  ...                                 22.0                                   45.0
17   Alexis Mac Allister  24.0  ...                                 28.0                                   52.5
20               Alisson  30.0  ...                                  2.0                                   75.0
21       Alphonse Areola  30.0  ...                                  1.0                                   90.9
..                   ...   ...  ...                                  ...                                    ...
475      Wes Foderingham  32.0  ...                                  1.0                                   50.0
476          Will Hughes  28.0  ...                                 17.0                                   41.4
478       William Saliba  22.0  ...                                 49.0                                   59.5
486      Youri Tielemans  26.0  ...                                  9.0                                   59.1
488        Yves Bissouma  26.0  ...                                 15.0                                   48.3

[143 rows x 154 columns]
```

***Yêu cầu 2**: Sử dụng thuật toán PCA, giảm số chiều dữ liệu xuống 2 chiều, vẽ hình phân cụm các điểm dữ liệu trên mặt 2D.*

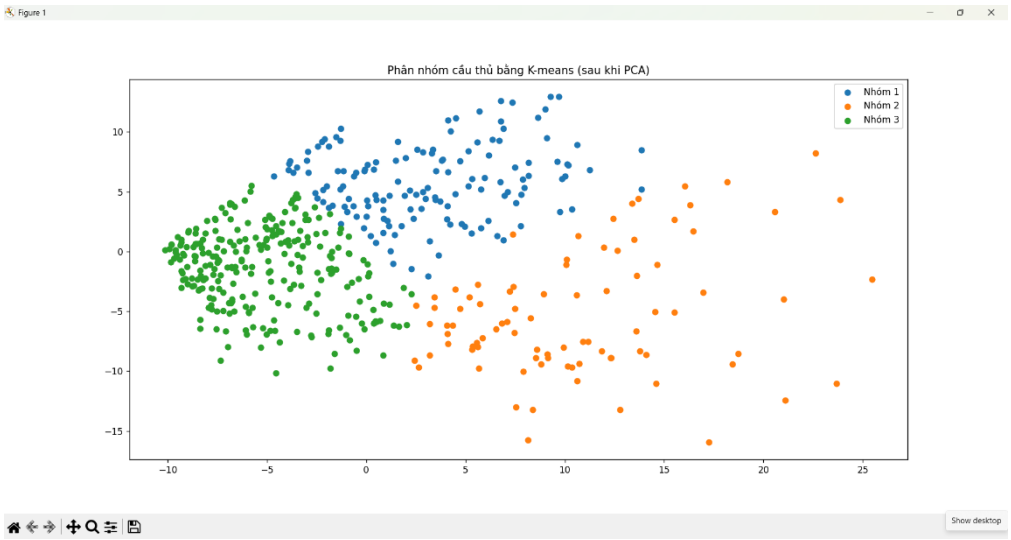
***Code**: PCA.py*

***Kết quả**: Terminal*

Ý tưởng:

- Trong các bài toán thực tế có thể có số chiều rất lớn, tới vài nghìn. PCA giảm chiều xuống 2, bỏ qua các chiều phương sai nhỏ, giữ lại những chiều quan trọng:

=> Kết quả PCA.py với số cụm là 3:



***Yêu cầu 3**: Vẽ radar chart so sánh 2 cầu thủ với tệp chỉ số nhập từ bàn phím.*

***Code**: radarChartPlot.py*

***Kết quả**: Terminal*

Giả sử đầu vào là:

python D:\Bai_tap_lon_D22\Bai3\radarChartPlot.py --p1 "Kevin De Bruyne" --p2 "Virgil van Dijk" --Attribute "Playing time_Matches played,Progression_PrgC,Shooting_Standard_SoT%,Passing_Long_Cmp,Pass Types_Pass Types_Dead"

Với:

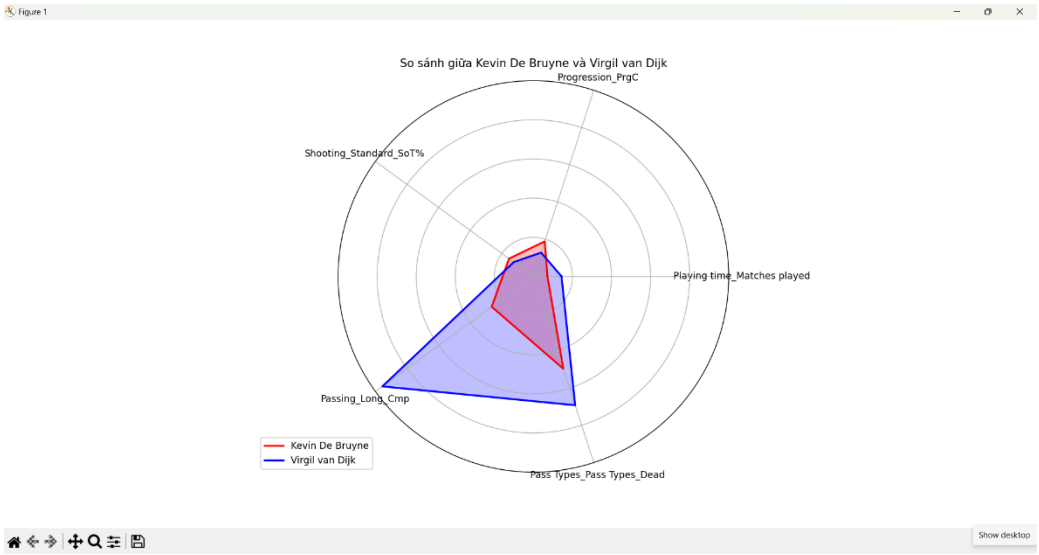
- Cầu thủ p1 là **Kevin De Bruyne**

- Cầu thủ p2 là **Virgil van Dijk**

- Atribute bao gồm 5 chỉ số:

  + **Playing time_Matches played**
  + **Progression_PrgC**
  + **Shooting_Standard_SoT%**
  + **Passing_Long_Cmp**
  + **Pass Types_Pass Types_Dead**

=> Kết quả file radarChartPlot.py:



Có thể thay thế 2 cầu thủ khác và các chỉ số (số lượng tùy thích) theo mong muốn, với điều kiện tên và các chỉ số trùng với bảng results.csv

IV. Chuyển nhượng cầu thủ

Đề xuất phương pháp định giá cầu thủ:

- Một số yếu tố quan trọng trong việc định giá:

+ **Tuổi**: Cầu thủ trẻ có tiềm năng phát triển thường có giá trị cao hơn. Khi cầu thủ lớn tuổi thì có thể lực và khả năng duy trì phong độ sẽ giảm.
+ **Vị trí thi đấu**: Một số vị trí có giá trị cao hơn trên thị trường. Ví dụ, tiền đạo và trung vệ có thể có giá trị cao do nhu cầu cao từ các câu lạc bộ.
+ **Số trận ra sân**: Cầu thủ có số trận thi đấu nhiều thường cho thấy độ bền bỉ và sức bền.
+ **Thành tích cá nhân**: Các chỉ số đều ảnh hưởng trực tiếp đến giá trị của cầu thủ (có thể dựa vào chỉ số đã thu thập trong file results.csv).
+ **Kinh nghiệm thi đấu quốc tế**: Các cầu thủ đã từng tham gia các giải đấu quốc tế lớn (như Champions League, World Cup) thường có giá trị cao hơn.
+ **Tình trạng thể chất và tiền sử chấn thương**: Một cầu thủ có tiền sử chấn thương nhiều thường bị đánh giá thấp hơn vì rủi ro cao.
+ **Thời gian còn lại trong hợp đồng**: Một cầu thủ có thời gian hợp đồng dài với câu lạc bộ hiện tại thường có giá trị chuyển nhượng cao hơn, vì câu lạc bộ có thể yêu cầu phí chuyển nhượng cao để bù đắp.
+ **Độ nổi tiếng và thương hiệu cá nhân**: Mặc dù đây không phải là một yếu tố kỹ thuật, độ nổi tiếng và thương hiệu cá nhân có thể ảnh hưởng đến giá trị cầu thủ. Một cầu thủ có sức hút lớn sẽ mang lại nhiều lợi ích thương mại cho câu lạc bộ.

- Việc định giá cầu thủ vẫn nên do những người có chuyên môn đánh giá và xem xét, khó có thể tạo ra một mô hình định giá tiêu chuẩn mà máy móc có thể tự động định giá.