HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG BỘ MÔN LẬP TRÌNH PYTHON

BÁO CÁO BÀI TẬP LỚN PYTHON

Giảng viên hướng dẫn: Kim Ngọc Bách

Sinh viên thực hiện: Nguyễn Duy Khánh

Mã sinh viên: B22DCKH067

Hà Nội, ngày 3 tháng 11 năm 2024

Mục lục

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.

***Thư viện cần cài:** BeautifulSoup, requests, pandas.

➔ Cách cài: vào CMD gõ *"pip install BeautifulSoup requests pandas"* và nhấn ENTER.

***Ý TƯỞNG:**

- Sử dụng thư viện **requests** để gửi yêu cầu để truy cập vào **URL** của trang Web và lấy về nội dung HTML của trang đó. Sau đó, dùng thư viện **BeautifulSoup** để phân tích và truy xuất các thẻ <*table*> chứa dữ liệu cần sử dụng.

***TRONG CODE:**

- Đầu tiên lấy và xử lý HTML của trang Ngoại Hạng Anh 2023-2024 thông qua URL bằng thư viện **requests**, dùng thư viện **BeautifulSoup** để tìm thẻ <*table*> đầu tiên (bảng này chứa thông tin về các đội trong giải đó). **(Hình 1.1)**

```
if __name__ == "__main__":
    # URL to fetch
    url = 'https://fbref.com/en/comps/9/2023-2024/2023-2024-Premier-League-Stats'
    r = requests.get(url)
    soup = BeautifulSoup(r.content, 'html.parser')

    # Tìm bảng chứa thông tin các đội bóng trong mùa giải 2023-2024
    table = soup.find('table', {
        'class': 'stats_table sortable min_width force_mobilize',
        'id': 'results2023-202491_overall'
    })
```

Hình 1.1

- Sau đó, tạo một list ***team_data*** để chứa thông tin của các đội bóng, mỗi đội bóng sẽ có 2 thông tin là tên đội và URL của đội. Tiếp tục tìm các thẻ <*a*> trong thẻ <*table*> vừa tìm được và duyết lần lượt các thẻ <*a*> đó và chỉ xử lý các thẻ mà có chuỗi "squads" trong nội dung của *href* và lấy nội dung của thẻ đó ta được tên đội, lấy nội dung *href* ta được URL của đội và lưu vào list vừa tạo. **(Hình 1.2)**

  ⇨ **Đã có danh sách thông tin cần thiết về các đội bóng**

```python
# Danh sách chứa dữ liệu đội bóng và URL
team_data = []

if table:
    # Tìm thẻ <tbody> trong <table>
    tbody = table.find('tbody')
    if tbody:
        # Lấy tất cả các thẻ <a> có định dạng như yêu cầu trong <tbody>
        teams = tbody.find_all('a', href=True)

        for team in teams:
            if "squads" in team['href']:  # Kiểm tra nếu "squads" có trong href
                team_name = team.get_text(strip=True)
                team_url = "https://fbref.com" + team['href']
                team_data.append([team_name, team_url])

        print("+-+-+-+-+-+-+-+Danh sách các đội bóng đã được lấy thành công.+-+-+-+-+-+-+-+")
    else:
        print("Không tìm thấy thẻ <tbody>.")
else:
    print("Không tìm thấy thẻ <table>.")
```

Hình 1.2

- Tiếp theo, tạo một list ***players_data*** để chứa thông tin toàn bộ cầu thủ trong giải và gọi hàm **Crawl_Data_For_Each_Team** để xử lý yêu cầu đề bài. **(Hình 1.3)**

```python
# #  Danh sach chứa từng cầu thủ của đội bóng
players_data = []
players_data = Crawl_Data_For_Each_Team(players_data, team_data)
```

Hình 1.3

- Trong hàm **Crawl_Data_For_Each_Team**, sẽ duyệt từng thông tin các đội đã lưu ở ***team_data*** và sẽ lấy URL của đội đang xử lý để lấy HTML.

    - Sau khi lấy được HTML của đội, tạo một list ***player_data_tmp*** để lưu thông tin các cầu thủ của đội đang xử lý và một ***map*** tên ***mp*** với key là tên cầu thủ còn value sẽ chiếu đến list chứa thông tin chỉ số của cầu thủ đó.
    - Tiếp, xử lý 10 thẻ <*table*> mỗi thẻ này sẽ chứa các thông tin về cầu thủ và chỉ số của họ. 10 bảng này đều có nội dung class giống nhau nhưng nội dung id khác nhau nên có thể dễ dàng tìm kiếm.
    - Đầu tiên, xử lý bảng Standard Stats với id trong thẻ <*table*> là ***"stats_standard_9"***. Tìm thẻ <*tbody*> chứa dữ liệu cảu bảng sau đó tìm tất cả các thẻ <*tr*> chứa thông tin cầu thủ. Duyệt qua các thẻ <*tr*> vừa tìm được, tìm thẻ <*td*> với ***"data-stat = minutes"*** để lấy tổng thời gian thi đấu của cầu thủ đang xét, nếu nhỏ hơn 90 phút thì bỏ qua, còn không thì gọi một hàm con đặc biệt để xử lý thông tin (sẽ trả về 1 list). **(Hình 1.4)**

```
player_data_tmp = []
mp = {} # Map ánh xạ đến list chứa thông tin và chỉ số của cầu thủ thông qua key là tên cầu thủ

# Tìm bảng chứa thông tin các cầu thủ
player_table = soup_tmp.find('table', {
    'class': 'stats_table sortable min_width',
    'id': 'stats_standard_9'
})
if player_table:
    # Tìm thẻ <tbody> trong <table>
    tbody = player_table.find('tbody')
    if tbody:
        players = tbody.find_all('tr')
        for player in players:
            player_minutes_matches = player.find('td', {'data-stat': 'minutes'}).get_text(strip=True) if player.find('td', {'data-stat': 'minutes'}).get_text(strip=True) else "N/a"
            # Lọc ra những cầu thủ đã thi đấu ít nhất 90 phút
            if player_minutes_matches == "N/a" or int(player_minutes_matches.replace(',','')) < 90:
                continue
            player_data_tmp = Data_Processing_of_Footballer(player, team_name, player_data_tmp, mp)
    else:
        print(f"<Không tìm thấy thẻ <tbody> trong bảng cầu thủ>")
else:
    print(f"<Không tìm thấy thẻ <table> chứa cầu thủ trong trang của đội {team_name}.>")
```

(Hình 1.4)

- Trong hàm **Data_Processing_of_Footballer**, sẽ lấy thông tin từng chỉ số thông qua các thẻ *<td>* với ***"data-stat"*** riêng biệt **(Hình 1.5)**. Trong hàm này sẽ trả về 1 list với thông tin các chỉ số và sẽ tạo một *key-value*

```
# Hàm xử lý dữ liệu cầu thủ
def Data_Processing_of_Footballer(tmp_tr, team_name, player_data_tmp, mp):
    # Lấy thông tin cầu thủ
    player_name = tmp_tr.find('th', {'data-stat': 'player'}).get_text(strip=True)
    player_national = tmp_tr.find('td', {'data-stat': 'nationality'}).find('a')['href'].split('/')[-1].replace('-Football', ' ') if tmp_tr.find('td', {'data-stat': 'nationality'}).find('a') else "N/a"
    player_position = tmp_tr.find('td', {'data-stat': 'position'}).get_text(strip=True)
    player_age = tmp_tr.find('td', {'data-stat': 'age'}).get_text(strip=True)
    #Playing time
    player_games = tmp_tr.find('td', {'data-stat': 'games'}).get_text(strip=True) if tmp_tr.find('td', {'data-stat': 'games'}).get_text(strip=True) else "N/a"
    player_games_starts = tmp_tr.find('td', {'data-stat': 'games_starts'}).get_text(strip=True) if tmp_tr.find('td', {'data-stat': 'games_starts'}).get_text(strip=True) else "N/a"
    player_minutes = tmp_tr.find('td', {'data-stat': 'minutes'}).get_text(strip=True) if tmp_tr.find('td', {'data-stat': 'minutes'}).get_text(strip=True) else "N/a"
```

(Hình 1.5)

```
# Thêm thông tin cầu thủ vào danh sách
tmp = [
    player_name, player_national, team_name, player_position, player_age, player_games, player_games_starts, player_minutes,
    player_goals_pens, player_pens_made, player_assists, player_cards_yellow, player_cards_red, player_xg,
    player_npxg, player_xg_assist, player_progressive_carries, player_progressive_passes,
    player_progressive_passes_received, player_goals_per90, player_assists_per90, player_goals_assists_per90,
    player_goals_pens_per90, player_goals_assists_pens_per90, player_xg_per90, player_xg_assist_per90, player_xg_xg_assist_per90,
    player_npxg_per90, player_npxg_xg_assist_per90
]
player_data_tmp.append(tmp)
mp[player_name] = player_data_tmp[-1]

return player_data_tmp
```

(Hình 1.6)

- Thứ hai, xử lý bảng **Goalkeeping** với id trong thẻ *<table>* là ***"stats_keeper_9".*** Tìm thẻ *<tbody>* và thẻ *<tr>* tương tự như ở bảng trên. Tạo một list tên ***list_tmp*** chứa tam thời tên các thủ môn. Duyệt lần lượt các thẻ *<tr>*, lấy tên thủ môn từ thẻ *<th>* và kiểm tra xem có tồn tại thủ môn này trong danh sách ***player_data_tmp*** không. Nếu có thì gọi hàm để xử lý thông tin thủ môn này (hàm này cũng trả về list). Để nối thông tin thủ môn vs thông tin thủ môn từ bảng trước, ta sẽ sử dụng nối list thông qua phép + và rồi thêm tên thủ môn vào ***list_tmp***. Với các cầu thủ khác, ta duyệt qua ***phayer_data_tmp*** và kiểm tra xem có tên cầu thủ đó trong ***list_tmp*** không, nếu không thì cộng vào list thông tin cầu thủ đó với list toàn chuỗi *"N/a"* **(Hình 1.7)**

```python
#Tìm bảng chứa thông tin các thủ môn
Goalkeeper_table = soup_tmp.find('table', {
    'class': 'stats_table sortable min_width',
    'id': 'stats_keeper_9'
})
if Goalkeeper_table:
    # Tìm thẻ <tbody> trong <table>
    tbody = Goalkeeper_table.find('tbody')
    if tbody:
        players = tbody.find_all('tr')
        # Danh sách lưu tạm thời tên các thủ môn
        list_tmp = []

        for player in players:
            player_name = player.find('th', {'data-stat': 'player'}).get_text(strip=True)
            if player_name in mp:
                mp[player_name] += Data_Processing_of_Goalkeeper(player)
                list_tmp.append(player_name)

        for player in player_data_tmp:
            if player[0] not in list_tmp:
                player += ["N/a"] * 15
    else:
        print(f"<Không tìm thấy thẻ <tbody> bảng thủ môn.>")
else:
    print(f"<Không tìm thấy thẻ <table> chứa thủ môn trong trang của đội {team_name}.>")
```

Hình 1.7

- 8 bảng còn lại sẽ xử lý giống với bảng **Goalkeeping**, mỗi bảng có một hàm xử lý thông tin riêng biệt (xem code để rõ ràng hơn).
- Sau khi xử lý xong 10 bảng, ta sẽ thêm list ***player_data_tmp*** vào ***players_data*** và sẽ tạm dừng khoảng 10 giây để xử lý đội tiếp theo (tránh bị web chặn). **(Hình 1.9)**

```python
# Thêm dữ liệu các cầu thủ của đội bóng vào danh sách chứa dữ liệu của tất cả các cầu thủ
players_data += player_data_tmp
print(f"<<<<<<<<Đã cào xong dữ liệu cầu thủ của đội {team_name}.>>>>>>>")

# Tạm nghỉ trước khi cào đội tiếp theo
time.sleep(10)
```

Hình 1.9

- Sau khi đã xử lý và lấy thông tin các cầu thủ của các đội, ta sẽ ***sort players_data*** theo thứ tự từ điển của *first name* và *tuổi giảm dần(nếu first name bằng nhau)*. Tiếp đó, dùng thư viện **pandas** để tạo một ***DataFrame*** từ ***players_data*** và thêm tên cho các cột chỉ số, thông tin rồi lưu vào file ***"results.csv"*** thông qua hàm của thư viện **pandas**. **(Hình 1.10)**

```python
# # Chuyển dữ liệu thành DataFrame và lưu thành file CSV
df_players = pd.DataFrame(players_data, columns=["Player Name", "Nation", "Team", "Position", "Age", "Matches Pla
                                                 "Goalkeeping_GA", "GGoalkeeping_GA90", "Goalkeeping_SoTA", "Goal
                                                 "Shooting_Gls", "Shooting_Sh", "Shooting_SoT", "Shooting_SoT%",
                                                 "Passing_Cmp", "Passing_Att", "Passing_Cmp%", "Passing_ToDist",
                                                 "Pass_Types_Live", "Pass_Types_Dead", "Pass_Types_FK", "Pass_Typ
                                                 "GSCreation_SCA", "GSCreation_SCA90", "GGSCreation_SCA_PassLive"
                                                 "DActions_Tkl", "DActions_TklW", "DActions_Def3rd", "DActions_Mi
                                                 "Possession_Touches", "Possession_Def Pen", "Possession_Def 3rd"
                                                 "PTime_Starts", "PTime_Mn/Start", "PTime_Compl", "PTime_Subs", "
                                                 "MStats_Fls", "MStats_Fld", "MStats_Off", "MStats_Crs", "MStats_
                                                 ])
df_players.to_csv("results.csv", index=False, encoding='utf-8-sig')
print("<----------------Đã lưu thông tin các cầu thủ vào file results.csv---------------->")
```

(Hình 1.10)

CÂU 2: Tìm top 3 cầu thủ có điểm cao nhất và thấp nhất ở mỗi chỉ số. 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. Vẽ historgram phân bố của mỗi chỉ số của các cầu thủ trong toàn giải và mỗi đội. 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.

***Thư viện cần cài:** pandas, tabulate (vẽ bảng đẹp), matplotlib, seaborn.
→ Cách cài: vào CMD gõ *"pip install pandas tabulate matplotlib seaborn"* và ấn ENTER.

***Ý TƯỞNG:**

- Dùng **pandas**, dùng để xử lý và thao tác với dữ liệu dạng bảng thông qua dữ liệu lấy từ file csv ***"resulsts.csv"*** ở câu 1.
- Dùng **tabulate**, giúp định dạng dữ liệu thành bảng một cách dễ nhìn, tiện lợi khi hiển thị kết quả trong *terminal*.
- Dùng **matplotlib**, thư viện chính dùng để vẽ biểu đồ trong Python, còn **seaborn** được xây dựng trên **matplotlib**, cung cấp các hàm trực quan hóa dữ liệu tốt hơn và giúp biểu đồ dễ nhìn hơn.
- Dùng **Collections** để đếm tần xuất các giá trị trong tập dữ liệu ở ý cuối, thư viện **os** để làm việc với hệ thống tệp và thư mục, thư viện **time** để quản lý thời gian.

***TRONG CODE:**

- Đọc file csv ***"results.csv"*** và chỉ lấy từ cột chỉ số thứ 4 trở đi. Chuyển kiểu dữ liệu với các giá trị *"N/a"* thành *NaN* dễ dàng xử lý và các giá trị khác *"N/a"* thành số vì file csv tất cả giá trị ở dạng chuỗi. **(Hình 2.1)**

```
if __name__ == "__main__":
    df = pd.read_csv("results.csv")

    columns_to_analyze = df.columns[4:]  # Chỉ chọn các cột chỉ số từ cột "Age" trở đi

    # Chuyển đổi kiểu dữ liệu, với các giá trị không phải số (như 'N/a') chuyển thành NaN
    df[columns_to_analyze] = df[columns_to_analyze].apply(pd.to_numeric, errors='coerce')
```

Hình 2.1

- Cung cấp cho người dùng 5 chức năng, trong đó có 4 chức năng ứng với từng yêu cầu đề đề bài và khi không muốn sử dụng nữa thì có chức năng thoát. **(Hình 2.2)**

```
while True:
    print("Chọn chức năng muốn thực hiện: ")
    print("1. Tìm Top 3 người có chỉ số cao nhất và thấp nhất")
    print("2. Tính trung vị, trung bình và độ lệch chuẩn của các chỉ số của toàn giải và các đội")
    print("3. Vẽ biểu đồ histogram cho toàn giải và từng đội")
    print("4. Tìm đội có giá trị cao nhất ở từng chỉ số và tần suất của từng đội và đánh giá")
    print("5. Thoát chương trình")
    choice = int(input("Nhập lựa chọn của bạn: "))
    while choice < 1 or choice > 5:
        choice = int(input("Vui lòng nhập lại: "))
    if choice == 1:
        get_top3(df, columns_to_analyze)
    elif choice == 2:
        get_statistics(df, columns_to_analyze)
    elif choice == 3:
        print_historgram(df, columns_to_analyze)
    elif choice == 4:
        get_best_team(df, columns_to_analyze)
    else:
        print("Đã thoát chương trình!")
        break
```

Hình 2.2

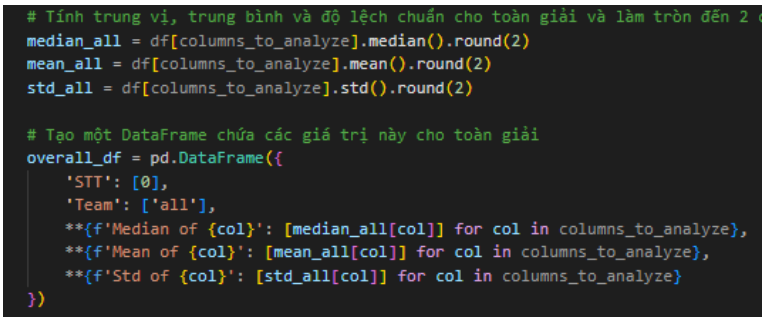
1. **Tìm top 3 cầu thủ có điểm cao nhất và thấp nhất ở mỗi chỉ số.**
   - Xử lý trong hàm **get_top3.**
   - Dùng ***nlargest(3, column)*** để tìm 3 cầu thủ có chỉ số cao nhất, đung thư viện tabulate để vẽ bảng dễ nhìn hơn **(Hình 2.3)** và ghi vào file Top3NguoiChiSoCaoNhat.txt. Tương tự với 3 cầu thủ có chỉ số thấp nhất. **(Hình 2.4)**

```
Top 3 cầu thủ cao nhất cho chỉ số 'Age':
╒═══════╤═══════════════════╤═══════════╤═══════╕
│       │ Player Name       │ Team      │   Age │
╞═══════╪═══════════════════╪═══════════╪═══════╡
│    47 │ Ashley Young      │ Everton   │    38 │
├───────┼───────────────────┼───────────┼───────┤
│   447 │ Thiago Silva      │ Chelsea   │    38 │
├───────┼───────────────────┼───────────┼───────┤
│   493 │ Łukasz Fabiański  │ West Ham  │    38 │
╘═══════╧═══════════════════╧═══════════╧═══════╛
```

Hình 2.3

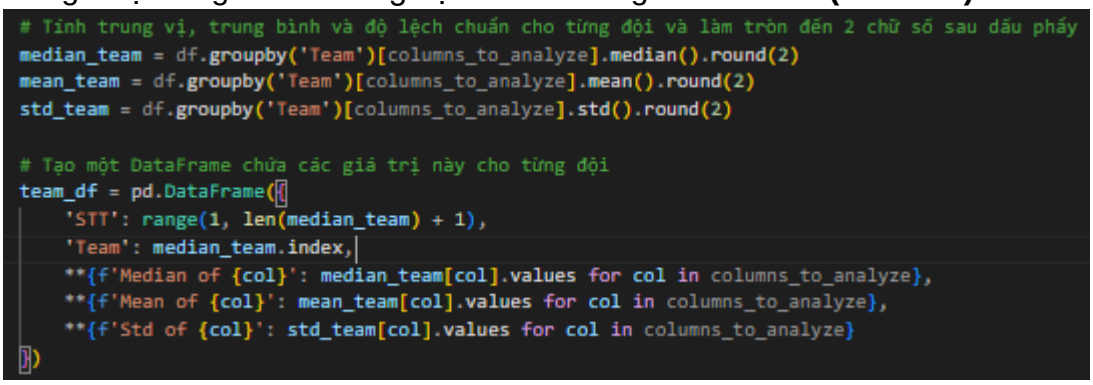
```python
def get_top3(df, column_to_analyze):
    # Ghi kết quả tìm kiếm Top3 cao nhất vào file Top3NguoiChiSoCaoNhat.txt
    with open("Top3NguoiChiSoCaoNhat.txt", "w", encoding="utf-8") as file:
        for column in columns_to_analyze:
            file.write(f"\nTop 3 cầu thủ cao nhất cho chỉ số '{column}':\n")
            top_highest = df.nlargest(3, column)[['Player Name', 'Team', column]]
            file.write(tabulate(top_highest, headers='keys', tablefmt='fancy_grid') + "\n")

        print("<<<<<<<Đã ghi kết quả tìm kiếm Top3 cao nhất vào file Top3NguoiChiSoCaoNhat.txt>>>>>>")

    # Ghi kết quả tìm kiếm Top3 thấp nhất vào file Top3NguoiChiSoThapNhat.txt
    with open("Top3NguoiChiSoThapNhat.txt", "w", encoding="utf-8") as file:
        for column in columns_to_analyze:
            file.write(f"\nTop 3 cầu thủ thấp nhất cho chỉ số '{column}':\n")
            top_lowest = df.nsmallest(3, column)[['Player Name', 'Team', column]]
            file.write(tabulate(top_lowest, headers='keys', tablefmt='fancy_grid') + "\n")

        print("<<<<<<<Đã ghi kết quả tìm kiếm Top3 thấp nhất vào file Top3NguoiChiSoThapNhat.txt>>>>>>")
```

2. **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.**
   - Xử lý trong hàm **get_statistics**.
   - Tính trung vị bằng gọi hàm ***median***, trung bình gọi hàm ***mean***, độ lệch chuẩn gọi hàm ***std***, làm tròn gọi hàm ***round***.
   - Tạo một ***DataFrame*** là ***overall_df*** để lưu trữ các giá trị thông kê của toàn giải sau khi đã dùng 4 hàm ở trên. **(Hình 2.5)**



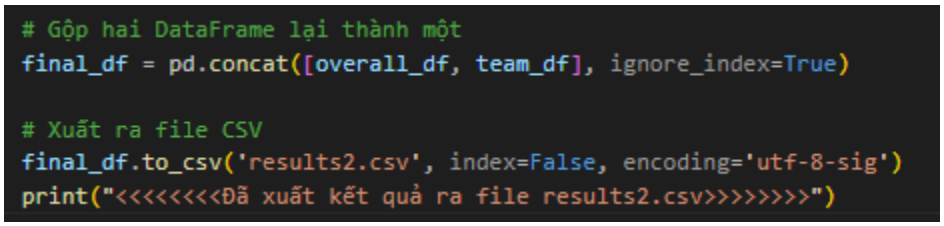
Hình 2.5

- Đối với các đội bóng thì cũng tạo một ***DataFrame*** là ***team_df*** để lưu trữ các giá trị thông kê của từng đội sau khi dùng 4 hàm trên. **(Hình 2.6)**



Hình 2.6

- Sau đó sẽ gộp hai ***DataFrame*** là ***overall_df*** và ***team_df*** thành ***final_df***. Tiếp đó xuất ra file csv ***"results2.csv"*****. (Hình 2.7)**



Hình 2.7

3. **Vẽ historgram phân bố của mỗi chỉ số của các cầu thủ trong toàn giải và mỗi đội.**
   - Xử lý trong hàm **print_historgram**.
   - Vì số lượng biểu đồ rất nhiều nên sẽ tạo một *folder* (nếu không tồn tại) để chứa các biểu đồ chỉ số của toàn giải, tạo một *folder* (nếu không tồn tại) để chứa các biểu đồ chỉ số của từng đội bóng (trong *folder* này sẽ chia thành nhiều *folder* con có tên là tên đội bóng để lưu các biểu đồ của đội bóng đó). Xem code để rõ hơn về cái này.
   - Với toàn giải, thì sử dụng thư viện **matplotlib** với **seaborn** để vẽ với các thông số biểu đồ đã được tạo trong code. **(Hình 2.8)**

```
# Vẽ histogram cho toàn giải
for col in columns_to_analyze:
    plt.figure(figsize=(8, 6))
    sns.histplot(df[col], bins=20, kde=True, color='blue')
    plt.title(f'Histogram of {col} - Toàn Giải')
    plt.xlabel(col)
    plt.ylabel('Số lượng cầu thủ (Người)')
    plt.grid(True, linestyle='--', alpha=0.5)
    # Lưu biểu đồ vào thư mục "histograms_all"
    plt.savefig(os.path.join(output_folder_1, f"{df.columns.get_loc(col)}.png"))
    plt.close()

print("Đã vẽ xong biểu đồ cho toàn giải")
```

Hình 2.8

- Với từng đội, vẫn dùng 2 thư viện như trên để vẽ. Trước tiên phải lấy danh sách các đội bóng khác nhau thông qua hàm ***unique()***. Duyệt lần lượt các đội và vẽ các biểu đồ. Dừng lại 3 giây sau đó mới vẽ tiếp đội khác. **(Hình 2.9)**

```
# Vẽ histogram cho từng đội
teams = df['Team'].unique()
for team in teams:
    # Tên thư mục của đội
    team_folder = os.path.join(output_folder_2, team)
    # Tạo thư mục nếu chưa tồn tại
    if not os.path.exists(team_folder):
        os.makedirs(team_folder)

    team_data = df[df['Team'] == team]

    for col in columns_to_analyze:
        plt.figure(figsize=(8, 6))
        sns.histplot(team_data[col], bins=20, kde=True, color='green')
        plt.title(f'Histogram of {col} - {team}')
        plt.xlabel(col)
        plt.ylabel('Số lượng cầu thủ (Người)')
        plt.grid(True, linestyle='--', alpha=0.5)
        # Lưu biểu đồ vào thư mục của đội
        plt.savefig(os.path.join(team_folder, f"{df.columns.get_loc(col)}.png"))
        plt.close()

    print(f"Đã vẽ xong biểu đồ cho đội {team}")
    time.sleep(3)

print("<<<<<<<<<<Đã vẽ xong biểu đồ cho toàn giải và từng đội>>>>>>>>>>")
```

Hình 2.9

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.**
    - Xử lý trong hàm **get_best_team.**
    - Chỉ lấy cột chỉ số từ ***Non – Penalty Goals*** trở đi để dễ dàng đánh giá. Và nhóm theo tên đội dùng ***groupby*** và tính trung bình của các chỉ số.**(Hình 2.10)**

```python
def get_best_team(df, columns_to_analyze):

    columns_to_analyze = df.columns[8:]  # Chỉ chọn các cột chỉ số từ cột "Non-Penalty Goals" trở đi
    df[columns_to_analyze] = df[columns_to_analyze].apply(pd.to_numeric, errors='coerce')

  # Nhóm theo 'Tên Đội' và tính trung bình các chỉ số
    team_summary = df.groupby('Team')[columns_to_analyze].mean()
```

Hình 2.10

- Tạo một list là ***results*** để chứa kết quả sau khi xử lý tìm đội có giá trị cao nhất ở các chỉ số **(Hình 2.11 và 2.12).** Sau đó, đếm tần suất đội đó có tên trong ***results***. Kết quả dự đoán sẽ dựa trên đội có tần suất cao nhất chứng tỏ đội đó có hiệu suất, phong độ tốt nhất.**(Hình 2.13 và 2.14)**

```python
# Tạo một danh sách để chứa kết quả
results = []

# Tìm đội có giá trị cao nhất ở từng chỉ số
for column in columns_to_analyze:
    best_team = team_summary[column].idxmax()
    max_value = team_summary[column].max()
    results.append([column, best_team, max_value])

# In kết quả dưới dạng bảng
headers = ["Chỉ số", "Team", "Giá trị"]
print(tabulate(results, headers=headers, tablefmt="grid"))
```

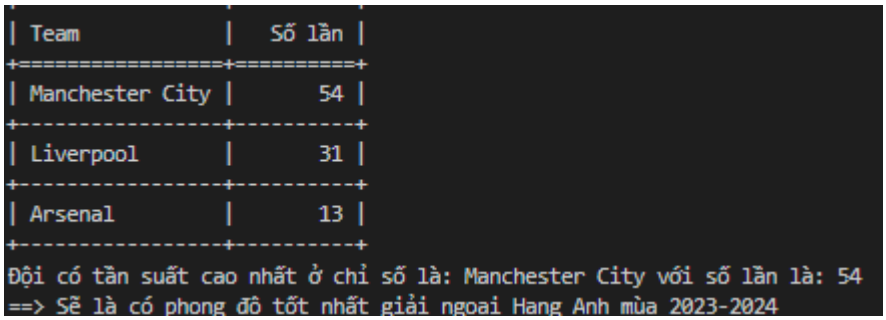
Hình 2.11

```
+--------------------------+-----------------+--------------+
| Chỉ số                   | Team            |     Giá trị  |
+==========================+=================+==============+
| Non-Penalty Goals        | Manchester City |     4.04762  |
+--------------------------+-----------------+--------------+
| Penalties Made           | Arsenal         |     0.47619  |
+--------------------------+-----------------+--------------+
| Assists                  | Manchester City |     3.2381   |
+--------------------------+-----------------+--------------+
```

Hình 2.12

```python
# Đếm tần suất của từng đội
team_counts = Counter([row[1] for row in results])

# Chuyển kết quả đếm tần suất thành dạng bảng
frequency_table = [[team, count] for team, count in team_counts.items()]
frequency_table.sort(key=lambda x: x[1], reverse=True)

# In bảng tần suất của từng đội
print("\nTần suất của từng đội bóng:")
print(tabulate(frequency_table, headers=["Team", "Số lần"], tablefmt="grid"))

print("Đội có tần suất cao nhất ở chỉ số là: " + str(frequency_table[0][0]) + " với số lần là: " + str(frequency_table[0][1]))
print("==> Số là có phong độ tốt nhất giải ngoại Hạng Anh mùa 2023-2024")
```

Hình 2.13



Hình 2.14

CÂU 3: Sử dụng thuật toán K-means để phân loại các cầu thủ thành các nhóm có chỉ số giống nhau. Theo bạn thì nên phân loại cầu thủ thành bao nhiêu nhóm? Vì sao? Bạn có Nhận xét gì về kết quả. Sử dụng thuật toán PCA, giảm số chiều dữ liệu xuống 2 chiều, vẽ hình phân cụm các điểm dữ liệu trên mặt 2D. Viết chương trình python vẽ biểu đồ rada (radar chart) so sánh cầu thủ.

***Thư viện cần cài:** pandas, sklearn, numpy, matplotlib.

→ Cách cài: vào CMD và gõ *"pip install pandas numpy sklearn matplotlib"* và ấn ENTER.

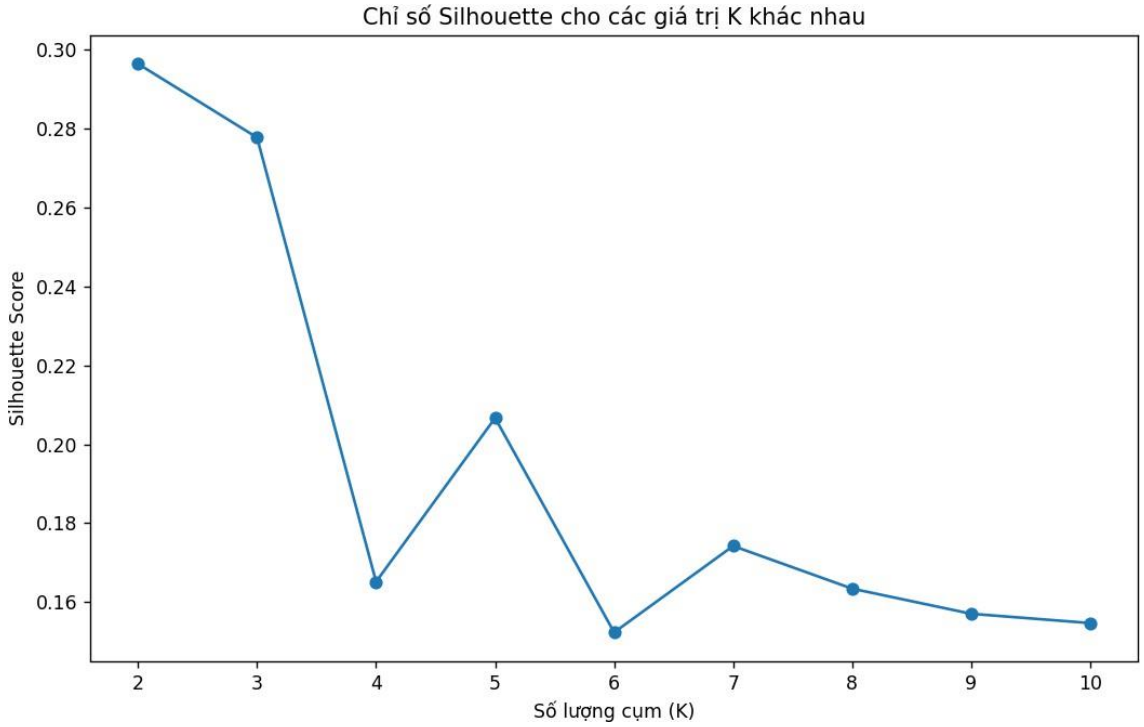
***Ý TƯỞNG:**

- Dùng **pandas** để đọc file csv chứa dữ liệu và xử lý dữ liệu trước khi dùng thuật toán
- Dùng **StandardScaler** từ **sklearn.preprocessing** để chuẩn hoá dữ liệu để mỗi đặc trưng có phương sai bằng 0 và độ lệch bằng 1, giúp cải thiện hiệu suất của *K-means*.
- Dùng **PCA** từ **sklearn.decomposition**: Giảm số chiều của dữ liệu để dễ trực quan hóa.
- Dùng **numpy** để tính toán số học và các thao tác mảng để thực hiện các bước của *K- means*, như tính khoảng cách *Euclidean* và trung bình của các điểm trong cụm.
- Dùng **matplotlib** để vẽ biểu đồ.
- Để phân loại cầu thủ thành số nhóm phù hợp thì dùng **phương pháp Silhouette Score** .
- Dùng thư viện **argparse** để phân tích các đối số dòng lệnh *(command-line arguments),* cho phép chương trình nhận các tham số từ người dùng khi chạy từ dòng lệnh (ý so sánh hai cầu thủ).

***TRONG CODE:**

1. Sử dụng thuật toán K-means để phân loại các cầu thủ thành các nhóm có chỉ số giống nhau. Theo bạn thì nên phân loại cầu thủ thành bao nhiêu nhóm? Vì sao? Bạn có Nhận xét gì về kết quả. Sử dụng thuật toán PCA, giảm số chiều dữ liệu xuống 2 chiều, vẽ hình phân cụm các điểm dữ liệu trên mặt 2D.

- Lựa chọn số lượng nhóm phù hợp thì dùng **phương pháp Silhouette Score** (code chạy trong file **Silhouette Score.py**) ta sẽ thu được biểu đồ. **(Hình 3.1)**



Hình 3.1

- Dựa vào biểu đồ **Silhouette Score** trên, ta thấy:
  - **K = 2** có **Silhouette Score** cao nhất (khoảng 0.30), cho thấy đây là số cụm có chất lượng phân cụm tốt nhất vì các điểm dữ liệu được phân cách rõ ràng giữa các cụm.
  - **K = 3** có **Silhouette Score** thấp hơn một chút (khoảng 0.28), nhưng vẫn tương đối cao, cho thấy cũng là lựa chọn khả thi.
  - Từ **K = 4** trở lên, **Silhouette Score** giảm đáng kể, chỉ dao động từ khoảng 0.16 đến 0.22, cho thấy chất lượng phân cụm không còn tốt nữa và các cụm không được phân biệt rõ ràng.

⇨ Có thể chọn ***K = 2 hoặc K = 3*** vì phù hợp với nội dung của phương pháp.

- Tiếp theo, sử dung **pandas** để đọc file csv chứa dữ liệu và xử lý các dữ liệu trước khi đưa vào chuẩn hoá. **(Hình 3.2)**

```
if __name__ == "__main__":
    # Đọc file csv
    data = pd.read_csv('results.csv')

    # Loại bỏ các cột  ở dạng chuỗi
    data = data.select_dtypes(exclude=['object'])

    # Điền các ô N/a bằng trung bình của cột đó
    data = data.fillna(data.mean())
```

Hình 3.2

- Chuẩn hoá dữ liệu bằng **StandardScaler** rồi dùng **PCA** giảm số chiều xuống 2. **(Hình 3.3)**

```
# Chuẩn hóa dữ liệu
scaler_standard = StandardScaler() # Khởi tạo
data = pd.DataFrame(scaler_standard.fit_transform(data), columns=data.columns)

# Áp dụng PCA giảm số chiều xuống 2
pca = PCA(n_components=2)
data = pca.fit_transform(data)
data = pd.DataFrame(data, columns=['PC1', 'PC2'])
```

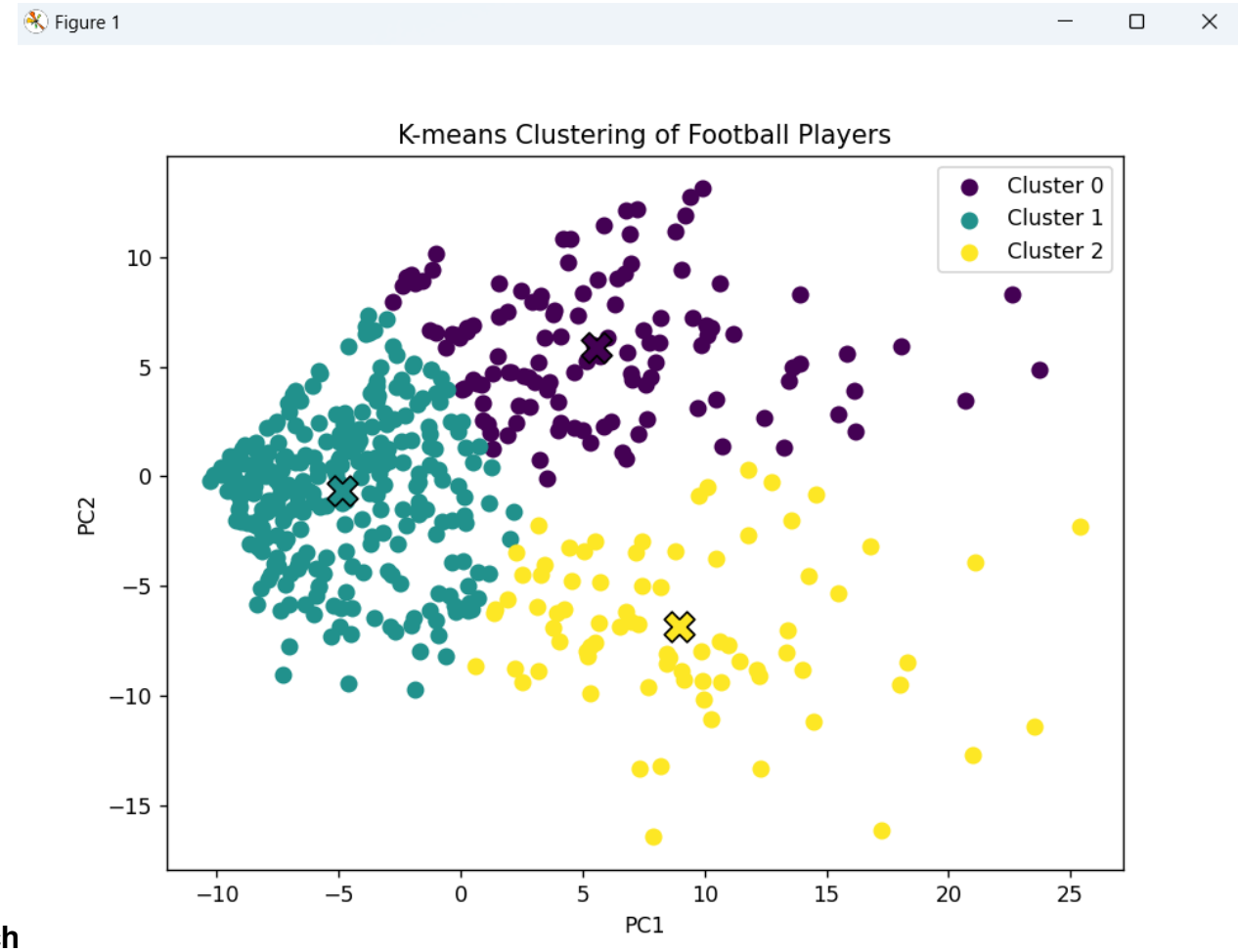
Hình 3.3

- Sau đó dùng ***K-means*** để phân loại cầu thủ với số lượng cụm (nhóm) ***K = 3*** đã suy ra từ **phương pháp Silhouette Score** và sẽ tạo ngẫu nhiên *K* tâm cụm. Nội dung đoạn code ***K-means*** là sẽ tính khoảng các từ các điểm đến *K* tâm cụm nếu gần với tâm cụm nào nhất thì sẽ đánh theo màu tâm cụm đó, sau đó lấy trung bình toạ độ của các điểm cùng màu để cập nhập tâm cụm mới và cứ tiếp tục lặp lại như thế đến khi nào tâm cụm sau khi cập nhập không thay đổi với trước khi cập nhập thì dừng và gọi hàm vẽ biểu đồ (xem code của hàm trong file **Cau3-1.py**). **(Hình 3.4)**

```
# K-means
# Số lượng cụm
k = 3

# Khởi tạo ngẫu nhiên các tâm cụm
centroids = data.sample(n=k).values

# Khởi tạo nhãn cho các điểm dữ liệu
clusters = np.zeros(data.shape[0])

epochs = 100
for step in range(epochs):  # Giới hạn số bước lặp
    # Bước 1: Gán nhãn dựa trên khoảng cách đến các tâm cụm
    for i in range(len(data)):
        distances = np.linalg.norm(data.values[i] - centroids, axis=1)
        clusters[i] = np.argmin(distances)

    # Bước 2: Cập nhật các tâm cụm
    new_centroids = np.array([data.values[clusters == j].mean(axis=0) for j in range(k)])

    # Kiểm tra nếu các tâm cụm không thay đổi thì kết thúc
    if np.all(centroids == new_centroids):
        # Vẽ biểu đồ
        plot_kmeans(data.values, centroids, clusters, step)
        break

    centroids = new_centroids
```

Hình 3.4

- **Kết quả:**



2. **Viết ch**
như sau:

+ python radarChartPlot.py --p1 <player Name 1> --p2 <player Name 2> --Attribute <att1,att2,...,att_n>
+ --p1: là tên cầu thủ thứ nhất
+ --p2: là tên cầu thủ thứ hai
+ --Attribute: là danh sách các chỉ số cần so sánh

- Khởi tạo parper để lấy thông số đầu vào. **(Hình 3.5)**

```
def main():
    # Khởi tạo parser để lấy thông số đầu vào
    parser = argparse.ArgumentParser(description='Compare two players using radar chart.')
    parser.add_argument('--p1', type=str, required=True, help='Player 1 name')
    parser.add_argument('--p2', type=str, required=True, help='Player 2 name')
    parser.add_argument('--Attribute', type=str, required=True, help='List of attributes to compare, separate
```

Hình 3.5

- Đọc dữ liệu từ file csv bằng **pandas** và xử lý các dữ liệu về dạng số (Vì dữ liệu đang ở dạng chuỗi). Gọi hàm để vẽ *rada (**plot_radar_chart**).* **(Hình 3.6)**

```
# Đọc dữ liệu từ file CSV
data = pd.read_csv('results.csv')

player1 = args.p1
player2 = args.p2
attributes = args.Attribute.split(',')

# Chuyển các cột dữ liệu về dạng số
for attr in attributes:
    data[attr] = pd.to_numeric(data[attr], errors='coerce')

# Vẽ biểu đồ radar
plot_radar_chart(data, player1, player2, attributes)
```

Hình 3.6

- Ở hàm **plot_radar_chart**, trích xuất dữ liệu của hai cầu thủ từ data dựa trên tên đã cung cấp *(player1 và player2)* và các thuộc tính *(attributes).* Xác định số lượng thuộc tính *(num_vars)* để tính toán các góc cho biểu đồ *radar*. **(Hình 3.7)**

```
def plot_radar_chart(data, player1, player2, attributes):
    # Lấy dữ liệu của hai cầu thủ
    p1_data = data[data['Player Name'] == player1].iloc[0][attributes].values.astype(float)
    p2_data = data[data['Player Name'] == player2].iloc[0][attributes].values.astype(float)

    # Số lượng thuộc tính
    num_vars = len(attributes)

    # Tạo các góc cho biểu đồ radar
    angles = np.linspace(0, 2 * np.pi, num_vars, endpoint=False).tolist()
```

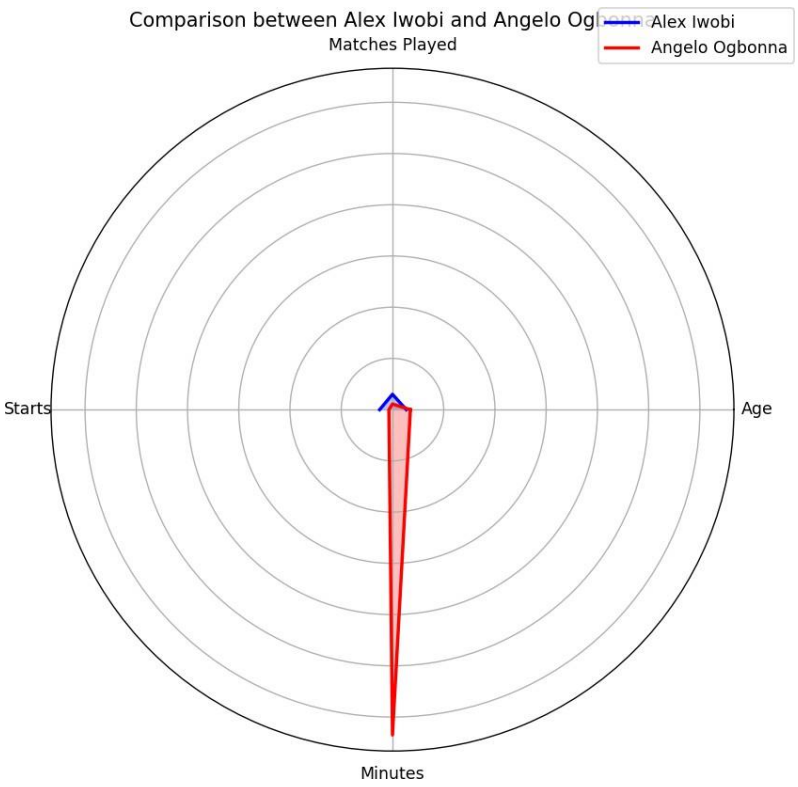
Hình 3.7

- Thiết lập chỉ số, thuộc tính để vẽ biểu đồ rada (xem chi tiết hơn trong file **Cau3- 2.py**).

➔ **Cách chạy file Cau3-2.py để so sánh 2 cầu thủ:** save file trước -> vào terminal gõ theo format này: *python Cau3-2.py –p1 “tên cầu thủ 1” –p2 “tên cầu thủ 2” –*
Attribute “các thuộc tính cần so sánh (sử dụng dấu phẩy để phân cách)” -> Nhấn ENTER để chạy

- **Ví dụ:** *python Cau3-2.py --p1 "Alex Iwobi" --p2 "Angelo Ogbonna" --Attribute "Age,Matches Played,Starts,Minutes"*

⇨ **Biểu đồ rada sẽ được như sau: (Hình 3.8)**



Hình 3.8

CÂU 4: Thu thập giá chuyển nhượng của các cầu thủ trong mùa 2023-2024 từ trang web https://www.footballtransfers.com. Đề xuất phương pháp định giá cầu thủ.

***Thư viện cần cài:** pandas, BeautifulSoup, requests.

→ Cách cài: vào CMD gõ *"pip install pandas BeautifulSoup requests"* và ấn ENTER.

***Ý TƯỞNG:**

- Dùng **requests** để lấy HTML từ URL của trang web. Sau đó, dùng **BeautifulSoup** phân tích và xử lý thẻ *<table>.*
- Dùng **pandas** để lưu dữ liệu xử lý được vào file csv.

***TRONG CODE:**

1. **Thu thập giá chuyển nhượng của các cầu thủ trong mùa 2023-2024 từ trang web17 https://www.footballtransfers.com.**

- Lấy HTML từ trang web thông qua URL và dùng **BeautifulSoup** để tìm thẻ *<table>* đầu tiên là nơi chứa thông tin các đội bóng. Sau đó, tạo list ***teams_data*** lưu thông tin các đội bóng và tìm thẻ *<tbody>* trong thẻ *<table>* vừa tìm được và tiếp tục tìm các thẻ *<a>* trong thẻ *<tbody>.* Duyệt lần lượt các thẻ *<a>* và lấy nội dung của thẻ *<a>* và nội dung của *href* trong

thẻ <*a*> đang xét này chính là tên đội bóng và URL của đội đó => lưu vào trong ***teams_data.*** **(Hình 4.1)**

```
if __name__ == "__main__":
    url = 'https://www.footballtransfers.com/us/leagues-cups/national/uk/premier-league/2023-2024'

    # Cào dữ liệu từ trang web
    response = requests.get(url)
    soup = BeautifulSoup(response.text, 'html.parser')

    table = soup.find('table', {
        'class': 'table table-striped table-hover leaguetable mvp-table ranking-table mb-0'
    })

    teams_data = []

    if table:
        tbody = table.find('tbody')
        if tbody:
            teams = tbody.find_all('a', href=True)

            for team in teams:
                teams_data.append([team.text.strip(), team['href']])

            print ("Hoàn thành lấy dữ liệu các team")
        else:
            print('Không tìm thấy tbody')
    else:
        print('Không tìm thấy bảng dữ liệu')
```

Hình 4.1

- Duyệt từng team trong ***teams_data***, lấy HTML của đội từ URL của đội. Tiếp là tìm thẻ <*table*> đầu tiên có *"**class** : table table-striped-rowspan ft-table mb-0"* và tìm thẻ <*tbody*> từ thẻ <*table*>. Tiếp tục, tìm tất cả thẻ <*tr*> trong thẻ <*tbody*>, duyệt lần lượt các thẻ <*tr*> và chỉ xử lý các thẻ có nội dung *class* là *"odd"* hoặc *"**even**"* và từ đó lấy tên cầu thủ và giá chuyển nhượng cầu thủ đó => lưu vào list ***players_data*** và lưu vào file csv. **(Hình 4.2)**

```
players_data = []

for team in teams_data:
    team_name = team[0]
    team_url = team[1]
    print(team_name, team_url)

    r_tmp = requests.get(team_url)
    soup_tmp = BeautifulSoup(r_tmp.text, 'html.parser')

    table_tmp = soup_tmp.find('table', {
        'class': 'table table-striped-rowspan ft-table mb-0'
    })

    if  table_tmp:
        tbody_tmp = table_tmp.find('tbody')
        if tbody_tmp:
            players = tbody_tmp.find_all('tr')

            for player in players:
                if "odd" in player['class'] or "even" in player['class']:
                    player_name = player.find('th').find('span').get_text(strip = True)
                    player_cost = player.find_all('td')[-1].get_text(strip = True)
                    players_data.append([player_name, team_name,  player_cost])

            print ("<------------Hoàn thành lấy giá các cầu thủ của team: ", team_name, "---------------->")
        else:
            print('Không tìm thấy tbody cầu thủ')
    else:
        print('Không tìm thấy bảng dữ liệu cầu thủ')

    time.sleep(3)

df = pd.DataFrame(players_data, columns=['Player', 'Team', 'Cost'])
df.to_csv("results4.csv", index=False, encoding='utf-8-sig')
print("<----------------Đã lưu thông tin  giá các cầu thủ vào file results4.csv-------------------->")
```

Hình 4.2