

TRƯỜNG CAO ĐẲNG FPT POLYTECHNIC

BÁO CÁO DỰ ÁN 01



**FPT POLYTECHNIC**

**Chuyên ngành Xử Lý Dữ Liệu  
ĐỀ TÀI**

**Phân tích yếu tố ảnh hưởng đến  
thành tích của vận động viên**

Nhóm thực hiện : DP-05

Lớp : DP19302

GVHD : Chu Thị Ngân

1. Trần Tuấn Kiệt - PH52381
2. Đỗ Hồng Quân - PH53269
3. Chu Tiêu Quyết - PH53063
4. Phan Văn Trường - PH53187
5. Kiều Bình Quyền - PH53569

## DANH SÁCH THÀNH VIÊN

Tên	Email	Mã sinh viên
Trần Tuấn Kiệt	kiettph52381@gmail.com	PH52381
Đỗ Hồng Quân	quandhph53269@gmail.com	PH53269
Phan Văn Trường	truongpvph53187@gmail.com	PH53187
Chu Tiểu Quyết	quyetcph53063@gmail.com	PH53063
Kiều Bình Quyền	quyenkbph53569@gmail.com	PH53569

## MỤC LỤC

1 Giới thiệu dự án	6
1.1 Giới thiệu	6
1.2 Yêu cầu của công ty	7
1.3 Lập kế hoạch dự án	8
2 Phân tích yêu cầu khách hàng	10
2.1 Phân tích yêu cầu	12
2.2 Câu chuyện dữ liệu	14
2.2.1 Đặt vấn đề	14
2.2.2 Xác định câu chuyện	14
2.2.3 Xác định rõ đối tượng	15
2.2.4 Xác định câu chuyện chi tiết	16
2.2.5 Trình bày dữ liệu	17
2.2.6 Những điều cần lưu ý	18
2.3 Kiến trúc hệ thống	18
2.3.1 Kiến trúc	18
2.3.2 Giải thích	18
2.4 Giải thích về bộ dữ liệu khách hàng	20
2.4.1 Các khái niệm	20
3 Làm sạch và chuyển đổi dữ liệu	22
3.1 Chuẩn bị dữ liệu	22
3.1.1 Giải pháp lưu trữ dữ liệu	22
3.1.2 Giải pháp phân bố dữ liệu	23
3.2 Làm sạch dữ liệu	24
3.2.1 Các vấn đề ảnh hưởng tới dữ liệu	24
1. Thiếu dữ liệu (Data Missingness)	24
3.2.2 Các tiêu chí đánh giá chất lượng dữ liệu	24
3.2.3 Các bước làm sạch dữ liệu	25
3.3 Chuyển đổi dữ liệu	29
3.3.1 Các trường hợp cần chuyển đổi	29
3.3.2 Các kỹ thuật chuyển đổi	29
3.3.3 Trình bày các phép chuyển đổi trong dự án	31
4 Xử lý dữ liệu	32

4.1 Chuẩn hóa dữ liệu	32
4.1.1 Trình bày các bước chuẩn hóa trong dự án	32
4.2 Mô hình hóa dữ liệu	33
4.2.1 Các loại mô hình hóa	33
4.2.2 Các tiêu chí đánh giá mô hình dữ liệu	34
4.2.3 Trình bày các bước mô hình hóa	35
4.2.4 Trình bày các bước tạo bảng dữ liệu	42
4.3 Xử lý dữ liệu DAX	44
4.3.1 Measure	44
4.3.2 Calculated column	47
4.3.3 Tạo bảng	50
4.3.4 Filter	51
5 Trực quan hóa dữ liệu	52
5.1 Các kỹ thuật trực quan hóa	52
5.2 Các nguyên tắc trực quan hóa	52
- Hiểu rõ mục tiêu trực quan hóa	52
- Chọn loại biểu đồ phù hợp	52
- Đơn giản và dễ hiểu	52
- Sắp xếp và làm nổi bật thông tin quan trọng	52
- Sử dụng màu sắc hợp lý	53
- Cung cấp ngũ cảnh	53
- Duy trì tỷ lệ đúng	53
- Kiểm tra và đánh giá	53
- Tôn trọng tính trung thực của dữ liệu	53
5.3 Trình bày cách thêm visual mới	54
5.3.1 Tạo visual thống kê chi tiết	54
5.3.2 Tạo visual thống kê tổng thể	54
6 Xây dựng báo cáo	63
6.1 Dashboard và report	63
6.2 Xây dựng báo cáo	64
6.2.1 Dashboard vs Report	64
Tối ưu hiệu suất	64
6.2.2 Dashboard	66
6.2.3 Bookmark	69
7 KẾT LUẬN	69

---

7.1 Báo cáo	69
7.1.1 Các bước viết báo cáo	69
7.1.2 Tổng hợp	70
7.2 Khó khăn	70
7.3 Thuận lợi	70
7.4 Hướng phát triển	71

---

## MỤC LỤC BẢNG

---

1.3.1 Bảng lập kế hoạch dự án.....	27
2.2.6.1 Bảng trình bày dữ liệu.....	35
2.3.1 Kiến trúc hệ thống.....	36
2.4.1 Các khái niệm.....	37
2.4.1.1 Bảng các khái niệm.....	38
2.4.2.1 Bảng các trường dữ liệu.....	39
2.4.2.2 Bảng Performance.....	40
2.4.2.3 Bảng Event.....	40
2.4.2.4 Bảng Team.....	40
2.4.2.5 Bảng Athlete.....	41
2.4.2.6 Bảng Game.....	41
3.1.2.2.1 Trình bày cách phân bố dữ liệu.....	43

## MỤC LỤC ẢNH

3.2.3.2.1 Xóa các bản ghi trùng lặp.....	40
3.2.3.2.2 Dữ liệu đã được xóa.....	40
3.2.3.2.3 Cho cột tiêu đề lên dòng đầu tiên.....	40
3.2.3.2.4 Chuyển các giá trị NA của cột Age về giá trị trung vị.....	41
3.2.3.2.5 Chuyển các giá trị NA của cột Height về giá trị trung vị.....	41
3.3.3.1 Chuyển định dạng của cột ‘Age’ từ dạng text về dạng whole number...	45
3.3.3.2 Chuyển định dạng của cột ‘Weight’ từ dạng text về dạng whole number...	45
3.3.3.3 Chuyển định dạng của cột ‘Height’ từ dạng text về dạng whole number...	46
4.1.1.1 Mô hình dữ liệu.....	47
4.2.3.1 Đổi tên cột ID thành AthleteID.....	49
4.2.3.2 Group by cột Team.....	49
4.2.3.3 Giao diện group by của team.....	50
4.2.3.4 Tạo cột ID cho team.....	50
4.2.3.5 Group by cột Event.....	51
4.2.3.6 tạo cột EventID.....	51
4.2.3.7 Khôi phục các bảng chưa gộp.....	52
4.2.3.11 Tạo cột PerformanceID.....	54
4.2.4.1 Tạo bảng reference từ clean_data rename thành Performance.....	54
4.2.4.2 Tạo bảng reference từ clean_data rename thành Event.....	55
4.2.4.3 Tạo bảng reference từ clean_data rename thành Games.....	55
4.2.4.4 Tạo bảng reference từ clean_data rename thành Team.....	56
4.2.4.5 Tạo bảng reference từ clean_data rename thành Athlete.....	56
4.2.4.6 Chính các mối quan hệ của các bảng thành hai chiều.....	57
4.3.1.1.1 Tính tỉ lệ phần trăm vận động viên.....	57
4.3.1.2.1 Tổng số huy chương.....	58
4.3.1.3.1 Tổng số lượng huy chương bạc.....	58
4.3.1.4.1 Tổng số huy chương đồng.....	58
4.3.1.5.1 Tổng số huy chương vàng.....	58
4.3.1.6.1 Tổng số môn thể thao.....	59
4.3.1.7.1 Tuổi nhỏ nhất, và lớn nhất.....	59
4.3.1.8.1 Tổng số huy chương đồng.....	59

---

4.3.1.9.1 Tổng số huy chương vàng.....	59
4.3.1.10.1 Tổng số môn thể thao.....	59
4.2.1.11 Tuổi nhỏ nhất, và lớn nhất.....	60
4.3.1.7.1 Tuổi nhỏ và lớn nhất.....	60
4.2.1.12 Tổng số số nước tham gia.....	60
4.3.1.8.1 Tổng số môn thể thao và số nước tham gia.....	60
4.3.1.13.1 Chiều cao và cân nặng trung bình của vận viên nam.....	60
4.3.2.1 Chiều cao và cân nặng trung bình của vận động viên nữ.....	60
4.3.2.1 Tổng số huy chương ở các kỳ thế vận hội mùa đông.....	60
4.3.2.2.1 Tổng số huy chương ở các kỳ thế vận hội mùa hè.....	61
4.3.2.3.1 Tổng số vận động viên.....	61
4.2.3.3 Số vận động viên nữ.....	61
4.3.2.5.1 Số vận động viên nam.....	61
4.3.2.6.1 Tuổi trung bình của các vận động viên.....	61
4.3.2.7.1 Tuổi trung bình của vận động viên nữ.....	61
4.3.2.8.1 Tuổi trung bình của vận động viên nam.....	61
4.3.2.9.1 Tỉ lệ vận động viên nữ.....	61
4.3.3.1 Tỉ lệ vận động viên nam.....	61
4.2.4.1 Số vận động viên trong từng sự kiện.....	62
4.3.3.1.1 Số vận động viên trong từng sự kiện.....	62
4.3.3.2.1 Tổng số đội tuyển theo quốc gia.....	62
4.3.3.3.1 Số lượng sự kiện mà mỗi đội tuyển đã tham gia.....	62
4.3.4.2.1 Tạo cột continent.....	63
4.3.4.2.2 Cột Continent.....	63
4.3.4.3.1 Tạo cột nhóm chiều cao.....	64
4.3.4.4.1 Tạo cột nhóm cân nặng.....	64
4.3.4.5.1 Tạo cột BMI.....	65
4.3.4.6.1 Tạo cột Medal value.....	65
4.2.5.7 Tạo cột Age Group.....	66
4.3.4.8.1 Tạo cột Medal group.....	66
4.3.5.1.1 Tạo bảng Medal Count By Country.....	66
4.3.5.2.1 Tạo bảng Athlete Count By Country.....	67
4.3.6.1.1 Tạo filter Year.....	68
4.3.6.2.1 Tạo filter Medal để chọn loại huy chương.....	68
4.3.6.3.1 Tạo filter Season.....	68

---

---

4.3.6.4.1 Tạo filter Sex.....	68
4.4.6.5.1 Tạo filter Sport.....	68
5.3.2.1.1 Tạo visual filter lọc theo Sport.....	71
5.3.2.2.1 Tạo visual filter lọc theo Season.....	71
5.3.2.3.1 Tạo visual card thống kê số lượng huy chương mùa hè.....	71
5.3.2.4.1 Tạo visual card thống kê số lượng huy chương mùa đông.....	72
5.3.2.5.1 Tạo visual stacked column thống kê 5 môn thể thao tổ chức nhiều nhất	
72	
5.3.2.6.1 Tạo visual stacked bar thống kê 5 quốc gia có số lượng vận động viên tham gia nhiều nhất.....	73
5.3.2.7 Tạo visual donut thống kê tỷ lệ huy chương của các quốc gia.....	73
5.3.2.8 Tạo visual scatter mối tương quan giữa số lượng vận động viên và số huy chương của các quốc gia.....	74
5.3.2.8.1 Tạo visual scatter mối tương quan giữa số lượng vận động viên và số huy chương của các quốc gia.....	74
5.3.2.9 Tạo visual treemap số lượng huy chương của các quốc gia.....	74
5.3.3.1.1 Tạo visual filter lọc theo Medal.....	75
5.3.3.2.1 Tạo visual filter lọc theo Season.....	75
5.3.3.3.1 Tạo visual filter lọc theo Sex.....	75
5.3.3.4.1 Tạo visual filter lọc theo Year.....	75
5.3.3.5.1 Tạo visual card thống kê số vận động viên nữ.....	76
5.3.3.6.1 Tạo visual card thống kê số vận động viên nam.....	76
5.3.3.7.1 Tạo visual Pie Chart tỷ lệ giới tính của vận động viên.....	76
5.3.3.8.1 Tạo visual Stacked Bar Chart tổng số huy chương đạt được theo giới tính.....	77
5.3.3.9.1 Tạo visual Line Chart Biến động số lượng huy chương của nam và nữ theo thời gian.....	77
5.3.3.9.1.1 Tạo visual Clustered Column Chart tuổi trung bình của vận động viên qua các kỳ thi vận hội theo giới tính.....	78
5.3.3.9.2.1 Tạo visual Bar Chart các môn thể thao phổ biến nhất được phân chia theo giới tính.....	78
5.3.3.9.3.1 Tạo visual Stacked Bar Chart số huy chương theo môn thể thao và giới tính.....	79
5.3.3.9.4.1 Tạo visual Stacked Bar Chart tỷ lệ huy chương giữa các nhóm tuổi theo giới tính.....	80
5.3.3.9.5.1 Tạo visual Clustered Column Chart so sánh thành tích theo giới tính trong từng mùa thi đấu (Summer vs Winter).....	80

---

---

5.3.4.1.1 Chiều cao trung bình của vận động viên nữ.....	80
5.3.4.2.1 Cân nặng trung bình của vận động viên nữ.....	81
5.3.4.3.1 Chiều cao trung bình của vận động viên nam.....	81
5.3.4.4.1 Cân nặng trung bình của vận động viên nam.....	81
5.3.4.5.1 Mối tương quan giữa chiều cao và cân nặng và số huy chương.....	81
5.3.4.6.1 Mối tương quan giữa chiều cao và cân nặng và số huy chương.....	82
5.3.4.7.1 Chiều cao trung bình của vận động viên nữ.....	83
5.3.4.8.1 Số lượng huy chương theo nhóm chiều cao.....	83
5.3.4.9.1 So sánh số lượng huy chương giữa nam và nữ theo nhóm cân nặng ..	84
5.3.4.11.1 Chiều cao và cân nặng theo khu vực.....	85
5.3.4.12.1 Thống kê chiều cao và cân nặng theo môn thể thao và thành tích....	86
5.3.5.1 Tạo visual clustered column so sánh độ tuổi trung bình của vận động viên đạt huy chương so với không đạt.....	86
5.3.5.1.1 Tạo visual clustered column so sánh độ tuổi trung bình của vận động viên đạt huy chương so với không đạt.....	86
5.3.5.2 Tạo visual stacked column top 5 môn thể thao có số lượng người tham gia theo nhóm tuổi.....	87
5.3.5.2.1 Tạo visual stacked column top 5 môn thể thao có số lượng người tham gia theo nhóm tuổi.....	87
5.3.5.3 Tạo visual clustered bar so sánh sự khác biệt về số lượng người tham gia theo độ tuổi trong 2 mùa.....	87
5.3.5.3.1 Tạo visual clustered bar so sánh sự khác biệt về số lượng người tham gia theo độ tuổi trong 2 mùa.....	87
5.3.5.4 Tạo visual line xu hướng độ tuổi trung bình của vận động viên (phân theo giới tính) qua các năm.....	88
5.3.5.4.1 Tạo visual line xu hướng độ tuổi trung bình của vận động viên (phân theo giới tính) qua các năm.....	88
5.3.5.5 Tạo visual 100% stacked column tỉ lệ đạt huy chương ở mỗi nhóm tuổi... 88	
5.3.5.5.1 ảnh Tạo visual 100% stacked column tỉ lệ đạt huy chương ở mỗi nhóm tuổi.....	88
5.3.5.6.1 Tạo visual stacked column top 5 môn thể thao có số lượng huy chương được thu thập theo độ tuổi.....	89
5.3.5.7.1 So sánh số lượng huy chương giữa các nhóm tuổi theo giới tính.....	89
5.3.5.8.1 Top 5 môn thể thao có số lượng người tham gia theo nhóm tuổi.....	89
5.3.5.9.1 Tạo visual Matrix thống kê huy chương đạt được trong từng môn thể thao theo từng nhóm tuổi.....	90

---

---

5.3.6.1.1 Tạo visual filter lọc theo giới tính.....	90
5.3.6.2 Tạo visual filter lọc theo thời gian.....	90
5.3.6.3.1 Tạo visual filter lọc theo Season.....	91
5.3.6.4.1 Tạo visual thống kê chiều cao trung bình.....	91
5.3.6.5.1 Tạo visual thống kê cân nặng trung bình.....	91
5.3.6.6.1 Tạo visual thống kê độ tuổi trung bình.....	92
5.3.6.7.1 Tạo visual thống kê các nước tham gia.....	92
5.3.6.8.1 Tạo visual thống kê các môn thể thao.....	92
5.3.6.9.1 Tạo biểu đồ Clustered Column Chart phân tích số lượng vận động viên tham gia ở từng môn thể thao.....	93
5.3.6.10.1 Tạo biểu đồ Pie Chart phân tích Tỷ lệ huy chương theo môn thể thao..	93
5.3.6.11.1 Tạo biểu đồ Stacked Column Chart phân tích Các quốc gia đạt được nhiều thành tích nhất.....	94
5.3.6.12.1 Tạo biểu đồ Line Chart phân tích Chiều cao và cân nặng trung bình của vđv theo từng môn thể thao.....	94
5.3.6.13.1 Tạo biểu đồ Clustered Bar Chart phân tích Các môn thể thao có vận động viên đạt nhiều huy chương nhất.....	95
5.3.7.1.1 Tạo visual filter lọc theo giới tính.....	95
5.4.1.1 Tạo visual filter lọc theo môn thể thao.....	95
5.3.3.3.1 Tạo visual filter lọc theo thời gian(năm).....	96
5.3.3.4.1 Tạo visual filter lọc theo loại huy chương.....	96
5.3.7.5.1 Tạo visual filter lọc theo quốc gia.....	96
5.3.7.6.1 Tạo visual thống kê tổng số vận động viên.....	96
5.3.7.7.1 Tạo visual thống kê tổng số quốc gia.....	96
5.3.7.8.1 Tạo visual thống kê tổng môn thể thao.....	97
5.3.7.9.1 Tạo visual thống kê tổng kỳ thi đấu.....	97
5.3.8.1 Tạo visual thống kê tổng số huy chương vàng.....	97
5.3.8.1.1 Tạo visual thống kê tổng số huy chương bạc.....	97
5.3.8.2.1 Tạo visual thống kê tổng số huy chương đồng.....	98
5.3.8.3.1 Tạo visual tỷ lệ huy chương theo châu lục.....	98
5.3.8.4.1 Tạo visual thống kê những môn thể thao top đầu.....	99
5.3.8.5.1 Tạo visual thống kê tỷ lệ vận động viên theo nhóm tuổi.....	100
5.3.8.6.1 Tạo visual thống kê số lượng vận động viên tham gia theo khu vực.	101
5.3.8.7.1 Tạo visual thống kê thành tích của các vận động viên.....	102
5.4.2.8 Tạo visual thống kê tỷ lệ huy chương giữa nam và nữ.....	103

---

---

5.3.8.9.1 Tạo visual tỷ lệ giới tính.....	104
5.3.9.1 Tạo visual thống kê số lượng huy chương của các môn thể thao theo năm	
104	
6.2.2.1 Ảnh overview.....	108
6.2.2.2 Ảnh overview.....	108
6.2.2.3 Ảnh phân tích theo giới tính.....	110
6.2.2.4 Ảnh phân tích theo chiều cao và cân nặng.....	111
6.2.2.5 Ảnh phân tích chiều cao và cân nặng.....	112
6.2.2.5 Ảnh phân tích theo quốc gia và mù thi đấu.....	114
6.2.2.6 Ảnh phân tích theo môn thể thao.....	114
6.2.3.1 Ảnh công thức hồi quy tuyến tính.....	110
6.2.3.2 Ảnh kiểm tra dữ liệu đầu vào.....	110
6.2.3.3 Tính hệ số tương quan giữa các cột.....	111
6.2.3.4 Xây dựng mô hình hồi quy tuyến tính.....	112
6.2.3.5 Dự đoán giá trị và Đánh giá mô hình.....	112
6.2.3.6 Vẽ biểu đồ phân phối dữ lượng.....	113
6.2.3.7 Vẽ biểu đồ so sánh giá trị thực tế và giá trị dự đoán.....	114

# 1 Giới thiệu dự án

## 1.1 Giới thiệu

### Giới thiệu về hiện trạng:

Trong lĩnh vực thể thao chuyên nghiệp, áp lực về thành tích luôn là một yếu cầu lớn đối với vận động viên, huấn luyện viên và các nhà quản lý. Khi mà thành tích của một vận động viên không chỉ là kết quả của quá trình rèn luyện cá nhân mà còn phụ thuộc vào nhiều yếu tố khác nhau như thể chất, điều kiện thi đấu, và thậm chí cả đặc điểm tâm lý, việc phân tích các yếu tố ảnh hưởng đến thành tích của các vận động viên đóng một vai trò rất quan trọng.

Các yếu tố về nhân khẩu học (như độ tuổi, giới tính), thể chất (như chiều cao, cân nặng) và các yếu tố ngoại cảnh (như điều kiện thi đấu, khu vực) có ảnh hưởng đáng kể đến hiệu suất thi đấu. Ví dụ, ở một số môn thể thao, chiều cao và cân nặng của vận động viên có thể là lợi thế giúp họ đạt thành tích cao hơn, trong khi ở các môn khác, tốc độ hoặc sức bền lại đóng vai trò quan trọng. Thời tiết và địa điểm thi đấu cũng có thể ảnh hưởng trực tiếp đến kết quả thi đấu của vận động viên, đặc biệt là với các môn ngoài trời.

Ngoài ra, sự phát triển không đồng đều giữa các khu vực và quốc gia cũng tạo ra sự chênh lệch đáng kể trong thành tích của các vận động viên. Việc phân tích dữ liệu Olympic giúp làm rõ các yếu tố ảnh hưởng này và tạo cơ sở khoa học để các đội thể thao cải tiến quy trình đào tạo và huấn luyện. Hơn nữa, trong bối cảnh nhiều nước đã ứng dụng phân tích dữ liệu trong thể thao để tối ưu hóa thành tích, các kết quả từ dự án này sẽ giúp Việt Nam và các nước đang phát triển có cái nhìn mới, khoa học hơn về quá trình đào tạo và phát triển vận động viên.

### Thông tin bộ dữ liệu:

Bộ dữ liệu Olympic bao gồm dữ liệu lịch sử từ hơn 120 năm thi đấu (từ 1896 - 2016), với thông tin chi tiết về vận động viên (bao gồm giới tính, độ tuổi, chiều cao, cân nặng), môn thi đấu, thành tích đạt được, và cả thông tin về các kỳ Thế vận hội như địa điểm, mùa tổ chức. Bộ dữ liệu này là nguồn tài nguyên quý giá để thực hiện các phân tích chi tiết, từ đó đưa ra những kết luận và nhận định có giá trị cho ngành thể thao.

## 1.2 Yêu cầu của công ty

### Về mặt dữ liệu:

- Công ty yêu cầu thu thập, làm sạch, và chuẩn hóa dữ liệu Olympic để đảm bảo chất lượng dữ liệu phục vụ phân tích.
- Dữ liệu cần bao gồm các thông tin chi tiết về:
- Thông tin vận động viên: giới tính, tuổi, chiều cao, cân nặng.
- Thông tin môn thể thao: loại môn thể thao mà vận động viên tham gia.
- Thông tin thành tích: kết quả thi đấu của vận động viên.
- Thông tin kỳ thi đấu: địa điểm tổ chức, mùa thi đấu, năm thi đấu

### Quản lý và lưu trữ:

- Yêu cầu hệ thống lưu trữ và quản lý dữ liệu có tính hiệu quả và dễ mở rộng.
- Hệ thống cần đảm bảo tính sẵn sàng của dữ liệu cho các hoạt động phân tích.
- Hệ thống phải có các cơ chế bảo mật để bảo vệ dữ liệu khỏi các nguy cơ rò rỉ và mất mát.
- Cần xây dựng các cơ chế quản lý dữ liệu tiên tiến để duy trì tính toàn vẹn của dữ liệu trong suốt quá trình sử dụng.

### Mục tiêu:

- Xác định các yếu tố tác động mạnh mẽ nhất đến thành tích của vận động viên.
- Hướng tới xây dựng hệ thống phân tích dự đoán giúp tối ưu hóa kết quả thi đấu dựa trên các đặc điểm cá nhân và điều kiện thi đấu của vận động viên.
- Định hướng chiến lược huấn luyện của công ty dựa trên dữ liệu phân tích.
- Hỗ trợ tuyển chọn vận động viên bằng cách dựa vào những yếu tố khoa học và có cơ sở dữ liệu đáng tin cậy.

### Đánh giá tính khả thi

### Năng lực (kỹ năng hiện có):

- Xây dựng báo cáo và trực quan hóa liệu
- Biết sử dụng các ngôn ngữ lập trình Python và SQL, các công cụ như Power BI và Excel để quản lý và phân tích dữ liệu lớn.

### Năng lực (kỹ năng cần học thêm cho dự án):

- Kỹ thuật phân tích chuyên sâu trong lĩnh vực thể thao, bao gồm các phương pháp thống kê nâng cao, học máy và mô hình hóa dự đoán để xác định các yếu tố tác động chính.
- Các công cụ lưu trữ dữ liệu lớn
- Triển khai các mô hình phân tích trên các hệ thống có khả năng mở rộng để đáp ứng nhu cầu của dự án.

### 1.3 Lập kế hoạch dự án

TT	HẠNG MỤC	BẮT ĐẦU	KẾT THÚC	KẾT QUẢ	THÀNH VIÊN
1	Giới thiệu dự án	29/10/2024	30/10/2024	100%	Cả nhóm
n	Giới thiệu công ty	30/10/2024	30/10/2024	100%	Kiệt, Trường
1.2	Yêu cầu công ty	30/10/2024	31/10/2024	100%	Quyết, Quyền
1.3	Lập kế hoạch dự án	31/10/2024	31/10/2024	100%	Cả nhóm
2	Phân tích yêu cầu	31/10/2024	2/11/2024	100%	Quân, Trường
2.1	Phân tích yêu cầu KH	3/11/2024	6/11/2024	100%	Cả nhóm
2.2	Câu chuyện dữ liệu	6/11/2024	9/11/2024	100%	Kiệt, Quyền
2.3	Kiến trúc hệ thống	11/11/2024	13/11/2024	100%	Kiệt, Quyền
2.4	Giải thích về bộ dữ liệu khách hàng	14/11/2024	15/11/2024	100%	Quyền
3	Làm sạch và chuyển đổi dữ liệu	11/11/2024	16/11/2024	100%	Kiệt, Quyết
3.1	Chuẩn bị dữ liệu	13/11/2024	16/11/2024	100%	Cả nhóm
3.2	Làm sạch dữ liệu	11/11/2024	13/11/2024	100%	Kiệt, Quyết
3.3	Chuyển đổi dữ liệu	13/11/2024	15/11/2024	100%	Kiệt, Quyết, Quyền
4	Xử lý dữ liệu	15/11/2024	25/11/2024	100%	Quyết
4.1	Chuẩn hóa dữ liệu	15/11/2024	17/11/2024	100%	Kiệt, Quyết
4.2	Mô hình hóa dữ liệu	17/11/2024	19/11/2024	100%	Kiệt, Quyết

4.3	Xử lý dữ liệu DAX	19/11/2024	25/11/2024	100%	Cả nhóm
5	Trực quan hóa dữ liệu	25/11/2024	2/11/2024	100%	Cả nhóm
5.1	Các kỹ thuật trực quan hóa	25/11/2024	28/11/2024	100%	Quyết, Quyền
5.2	Các nguyên tắc trực quan hóa	25/11/2024	28/11/2024	100%	Kiệt, Trường
5.3	Trình bày cách thêm visual mới	29/11/2024	1/12/2024	100%	Cả nhóm
6	Xây dựng báo cáo	3/12/2024	3/12/2024	90%	Cả nhóm
6.1	Dashboard và Report	3/12/2024	3/12/2024	100%	Cả nhóm
6.2	Xây dựng báo cáo	3/12/2024	4/12/2024	90%	Cả nhóm
7	Kết luận	4/12/2024	5/12/2024	100%	Quyền, Quyết
7.1	Báo cáo	4/12/2024	5/12/2024	100%	Quân, Trường
7.2	Khó khăn	5/12/2024	6/12/2024	100%	Kiệt
7.3	Thuận lợi	5/12/2024	6/12/2024	100%	Kiệt
7.4	Hướng phát triển	6/12/2024	10/12/2024	100%	Kiệt

### 1.3.1 BÀNG LẬP KẾ HOẠCH DỰ ÁN

## 2 Phân tích yêu cầu khách hàng

### Phân tích tổng quan

- Top những môn thể thao dẫn đầu về thành tích: Xác định các môn thể thao có thành tích tốt nhất trong các kỳ thi đấu, dựa trên số lượng huy chương đạt được.
- Tỷ lệ giới tính giữa nam và nữ: Phân tích sự phân bố tỷ lệ vận động viên nam và nữ tham gia các kỳ thi đấu.
- Tỷ lệ vận động viên theo nhóm tuổi : Phân tích sự phân bố vận động viên theo các nhóm tuổi khác nhau.
- Số lượng vận động viên tham gia theo khu vực: Xác định số lượng vận động viên tham gia từ các khu vực khác nhau.
- Tỷ lệ huy chương giữa nam và nữ: Phân tích tỷ lệ huy chương đạt được giữa nam và nữ.
- Tỷ lệ huy chương theo châu lục: Phân tích thành tích của các châu lục
- Số lượng huy chương của các môn thể thao qua từng năm: Phân tích sự thay đổi số lượng huy chương của các môn thể thao qua các năm
- Thông kê thành tích của vận động viên: Cung cấp cái nhìn tổng quan về thành tích thi đấu của vận động viên.

### Phân tích theo giới tính

- Tỷ lệ giới tính của vận động viên: Cần có dữ liệu về giới tính của từng vận động viên để trực quan hóa tỷ lệ nam/nữ.
- Tổng số huy chương của các kỳ đạt được theo giới tính: Dữ liệu về giới tính,các kỳ thi đấu và số lượng huy chương để thể hiện số lượng huy chương của các kỳ theo từng giới tính.
- Biến động số lượng huy chương của nam và nữ theo thời gian: Dữ liệu về tổng số huy chương,giới tính và Year để xem sự khác biệt giữa nam và nữ ở các giai đoạn khác nhau.
- Tuổi trung bình của vận động viên theo giới tính qua các kỳ: Dữ liệu về tuổi, giới tính và năm tham gia để phân tích sự thay đổi độ tuổi trung bình của vận động viên theo thời gian.

- Các môn thể thao phổ biến nhất theo giới tính: Dữ liệu về môn thể thao và giới tính để xác định môn thể thao phổ biến nhất với từng giới tính.
- So sánh thành tích theo giới tính trong từng mùa thi đấu (Summer vs Winter) : Dữ liệu về mùa ,giới tính và số huy chương để đánh giá sự khác biệt về thành tích giữa Summer và Winter theo giới tính.
- Số huy chương theo môn thể thao và giới tính : Dữ liệu về môn thể thao,giới tính và số huy chương để so sánh trực tiếp số huy chương giữa nam và nữ trong cùng một môn thể thao.
- Tỷ lệ huy chương giữa các nhóm tuổi theo giới tính : Dữ liệu về nhóm tuổi,giới tính và tỉ lệ huy chương để xác định nhóm tuổi nào có thành tích nổi bật nhất theo từng giới tính.

### **Phân tích theo độ tuổi**

- Phân phối tuổi của vận động viên: Dữ liệu về tuổi của vận động viên để tạo biểu đồ phân phối.
- Tuổi trung bình của vận động viên qua từng năm: Dữ liệu về tuổi và năm tham gia của vận động viên để xác định xu hướng tuổi trung bình.
- Độ tuổi trung bình của vận động viên đạt huy chương so với không đạt : Cần thông tin về tuổi và trạng thái đạt huy chương của vận động viên để so sánh độ tuổi trung bình.
- Môn thể thao phổ biến nhất ở từng nhóm tuổi: Dữ liệu về môn thể thao và nhóm tuổi của vận động viên để xác định môn thể thao phổ biến trong mỗi nhóm tuổi.
- Phân phối độ tuổi của vận động viên nam và nữ: Dữ liệu về tuổi và giới tính để phân tích phân phối độ tuổi của vận động viên theo từng giới tính.

### **Phân tích theo chiều cao và cân nặng**

- Mối tương quan giữa chiều cao và cân nặng của vận động viên đạt huy chương: Phân tích mối quan hệ giữa chiều cao và cân nặng của các vận động viên đạt huy chương.
- Chiều cao và cân nặng trung bình theo môn thể thao: Tính toán chiều cao và cân nặng trung bình của vận động viên theo từng môn thể thao.

- Chiều cao và cân nặng trung bình của vận động viên qua các năm: Phân tích sự thay đổi chiều cao và cân nặng trung bình của vận động viên qua các năm.
- Phân tích đặc điểm thể chất và thành tích theo nhóm thể thao: phân tích mối quan hệ giữa đặc điểm thể chất theo từng nhóm môn thể thao
- So sánh số huy chương của nam và nữ theo nhóm cân nặng: Phân tích số lượng huy chương của vận động viên nam và nữ theo nhóm cân nặng.
- So sánh số huy chương của nam và nữ theo nhóm chiều cao: Phân tích số lượng huy chương của vận động viên nam và nữ theo nhóm chiều cao.
- So sánh chiều cao và cân nặng của từng loại huy chương: Phân tích chiều cao và cân nặng của vận động viên theo từng loại huy chương (vàng, bạc, đồng).
- phân tích thành tích theo nhóm cân nặng: Phân tích thành tích thi đấu của vận động viên theo các nhóm cân nặng.
- Phân tích thành tích theo nhóm chiều cao: Phân tích thành tích thi đấu của vận động viên theo các nhóm chiều cao.

### **Phân tích theo môn thể thao**

- Số lượng vận động viên tham gia ở từng môn thể thao: Dữ liệu về môn thể thao và vận động viên để thống kê số lượng tham gia.
- Tỷ lệ huy chương theo môn thể thao: Dữ liệu về số huy chương theo môn để tính toán tỷ lệ.
- Số lượng vận động viên nam và nữ ở mỗi môn: Dữ liệu về môn thể thao và giới tính để thể hiện sự tham gia của vận động viên nam/nữ.
- Môn thể thao phổ biến nhất qua các kỳ thi đấu: Dữ liệu môn thể thao và kỳ thi đấu để phân tích xu hướng phổ biến theo thời gian.
- Các nhóm môn thể thao nhiều thành tích nhất: Dữ liệu môn thể thao và huy chương để xác định nhóm môn thể thao đạt nhiều huy chương.

### **Phân tích theo thành phố và mùa thi đấu**

- Số lượng vận động viên nam tham gia nhiều nhất của 5 quốc gia: Dữ liệu về quốc gia, giới tính và số lượng tham gia để so sánh giữa các quốc gia.
- Sự phân bố các kỳ thi đấu giữa các thành phố: Dữ liệu về các kỳ thi đấu và thành phố để phân bố số kỳ theo từng thành phố.

- Các môn thể thao phổ biến nhất ở mỗi thành phố: Dữ liệu môn thể thao và thành phố để xác định môn phổ biến theo từng thành phố.
- Số lượng huy chương đạt được tại mỗi thành phố: Dữ liệu huy chương và thành phố để hiển thị tổng số huy chương tại mỗi địa điểm.
- Tổng số huy chương đạt được trong các mùa Summer và Winter: Dữ liệu huy chương và mùa thi đấu để tính tổng số huy chương theo từng mùa.
- Tỷ lệ huy chương của các quốc gia: Dữ liệu huy chương và quốc gia để biểu thị tỷ lệ huy chương giữa các quốc gia.

## 2.1 Phân tích yêu cầu

### 2.1.1 Dữ liệu:

**Loại dữ liệu:** Dữ liệu liên quan đến vận động viên bao gồm các đặc điểm như giới tính, độ tuổi, môn thể thao tham gia, chiều cao, cân nặng, số huy chương đạt được, quốc gia tham gia, và thông tin về các kỳ thi đấu. Ngoài ra, cần thu thập thông tin về các môn thể thao phổ biến theo giới tính, độ tuổi và các mùa thi đấu.

**Dữ liệu đầu vào:** Các dữ liệu này có thể được thu thập từ bảng xếp hạng, kết quả thi đấu, khảo sát về các môn thể thao, và các cơ sở dữ liệu quốc gia hoặc quốc tế về thể thao.

**Dữ liệu đầu ra:** Các kết quả phân tích như tỷ lệ huy chương theo giới tính, môn thể thao phổ biến theo độ tuổi, phân phối chiều cao, cân nặng, sự tương quan giữa chiều cao và huy chương, và số lượng vận động viên nam và nữ ở mỗi môn.

**Quản lý và lưu trữ:** Quản lý dữ liệu: Dữ liệu cần được tổ chức và phân loại rõ ràng, với các bảng lưu trữ thông tin về vận động viên, kết quả thi đấu, số huy chương, và các thông số liên quan khác. Các cơ sở dữ liệu này cần hỗ trợ tra cứu và phân tích hiệu quả.

**Lưu trữ dữ liệu:** Các dữ liệu này có thể được lưu trữ trong cơ sở dữ liệu quan hệ (SQL) hoặc trong các file dạng CSV/Excel tùy thuộc vào yêu cầu về tính mở rộng và tốc độ truy vấn. Dữ liệu cần được cập nhật định kỳ sau mỗi kỳ thi đấu.

**Công nghệ:** Phân tích dữ liệu với Python: Dùng để phân tích, tổng hợp và trích xuất các thông tin từ dữ liệu thu thập được. Sử dụng thư viện như Numpy để vẽ

đồ thị phân tích các chỉ số như tỷ lệ huy chương theo giới tính, chiều cao và cân nặng.

**Lý do chọn công nghệ:** Power Bi là công cụ mạnh mẽ để phân tích và trực quan hóa dữ liệu, giúp giải quyết các bài toán phân tích như tỷ lệ huy chương theo giới tính, phân tích độ tuổi và môn thể thao.

**Dữ liệu:** Cần thu thập và lưu trữ dữ liệu chi tiết về vận động viên, môn thể thao, kết quả thi đấu và các thông số thể chất (chiều cao, cân nặng) vào cơ sở dữ liệu.

**Quản lý và lưu trữ:** Hệ thống lưu trữ dữ liệu dựa trên Google Drive sẽ được sử dụng. Điều này có thể bao gồm việc sử dụng các driver cho phép truy cập trực tiếp vào các loại tệp khác nhau hoặc các giải pháp lưu trữ đám mây, tùy thuộc vào yêu cầu về khả năng mở rộng và tốc độ truy vấn của dự án. Dữ liệu cần được cập nhật định kỳ sau mỗi kỳ thi đấu.

**Công nghệ:** Power bi sẽ giúp thực hiện phân tích, trực quan hóa dữ liệu về tỷ lệ huy chương theo giới tính, sự phân bố các môn thể thao theo độ tuổi, quốc gia và mùa thi đấu.

## 2.2 Câu chuyện dữ liệu

### 2.2.1 Đặt vấn đề

#### Mô tả thực trạng:

Trong thế giới thể thao cạnh tranh ngày càng gay gắt, việc hiểu các yếu tố ảnh hưởng đến thành tích của vận động viên trở nên quan trọng hơn bao giờ hết. Các yếu tố như giới tính, tuổi tác, chiều cao, cân nặng, loại hình thể thao và điều kiện tại thành phố tổ chức đều có thể ảnh hưởng đến hiệu suất thi đấu. Việc phân tích những dữ liệu này có thể giúp các liên đoàn thể thao, huấn luyện viên và vận động viên tối ưu hóa chiến lược, cải thiện thành tích và đạt được lợi thế cạnh tranh.

#### Dữ liệu liên quan:

Bộ Dữ liệu Olympic từ 1896 đến 2016 chứa thông tin chi tiết về các vận động viên và thành tích của họ trong các kỳ Thế vận hội. Bộ dữ liệu bao gồm các trường thông tin như: Tên, Giới tính, Độ tuổi, Quốc gia, Môn thể thao, Sự kiện thi đấu, Năm tham gia, và Huy chương đạt được

### Mục tiêu:

Xây dựng câu chuyện dữ liệu hấp dẫn, trực quan và dễ hiểu, giúp người đọc nắm bắt thông tin nhanh chóng và chính xác. Câu chuyện cần làm rõ vấn đề, chỉ ra các insight quan trọng và đề xuất giải pháp khả thi, góp phần tối ưu hóa chiến lược thi đấu của các huấn luyện viên cũng như phương pháp tập luyện của vận động viên.

### 2.2.2 Xác định câu chuyện

#### Thông điệp từ dữ liệu:

Câu chuyện này sẽ tập trung khai thác cách các yếu tố như giới tính, độ tuổi, thể trạng, địa điểm, và mùa giải ảnh hưởng đến thành tích thi đấu. Qua đó, chúng ta có cái nhìn sâu sắc về sự phát triển của thể thao toàn cầu và vai trò quan trọng của dữ liệu trong việc thúc đẩy thành công và bình đẳng trong thể thao.

#### Mục tiêu:

- Khám phá mối quan hệ giữa thể trạng, độ tuổi, và thành tích thi đấu để cải thiện hiệu quả luyện tập và thi đấu.
- Phân tích ảnh hưởng của các yếu tố môi trường như địa điểm và mùa giải đối với thành tích của vận động viên.

#### Giải pháp:

- Tối ưu hóa các chương trình huấn luyện, điều chỉnh theo thể trạng và độ tuổi phù hợp với từng môn thể thao.
- Tăng cường nghiên cứu về ảnh hưởng của thời tiết và văn hóa địa phương đến thành tích, từ đó hỗ trợ vận động viên chuẩn bị tốt hơn cho thi đấu.

#### Cách tiếp cận:

- Phân tích mối tương quan giữa các yếu tố như giới tính, độ tuổi, và thể trạng với thành tích huy chương.
- Xác định xu hướng phát triển thành tích thể thao qua các mùa Olympic.
- Rút ra so sánh giữa các địa điểm tổ chức và mùa giải để tìm hiểu tác động của môi trường thi đấu.

## 2.2.3 Xác định rõ đối tượng

### Người đọc:

Ban lãnh đạo và huấn luyện viên: Đối tượng chính là ban lãnh đạo và huấn luyện viên trong lĩnh vực thể thao, những người cần hiểu rõ các yếu tố ảnh hưởng đến thành tích của vận động viên để hỗ trợ lập kế hoạch chiến lược, đưa ra các quyết định về huấn luyện, và nâng cao hiệu quả thi đấu.

### Mức độ am hiểu:

Ban lãnh đạo: Họ cần thông tin tổng quan, dễ hiểu về xu hướng và các yếu tố chính tác động đến thành tích, cùng những khuyến nghị chiến lược có thể áp dụng. Ban lãnh đạo có thể không đi sâu vào các kỹ thuật phân tích dữ liệu nhưng cần thấy rõ kết quả, số liệu, và tác động thực tế của các yếu tố phân tích.

Huấn luyện viên: Họ có kiến thức sâu về kỹ thuật thể thao và cần chi tiết về cách từng yếu tố (giới tính, độ tuổi, thể trạng, địa điểm) ảnh hưởng đến vận động viên. Dữ liệu nên tập trung vào việc hỗ trợ họ tối ưu hóa kế hoạch huấn luyện, nắm bắt các yếu tố giúp cải thiện thành tích cụ thể của vận động viên.

## 2.2.4 Xác định câu chuyện chi tiết

### Bối cảnh:

Báo cáo này được xây dựng trên cơ sở bộ dữ liệu 120 năm lịch sử của Thế vận hội Olympic – sự kiện thể thao toàn cầu lớn nhất, quy tụ các vận động viên ưu tú từ khắp nơi trên thế giới tham gia tranh tài ở nhiều môn khác nhau. Mỗi kỳ Olympic không chỉ là dịp để các vận động viên khẳng định tài năng và mang vinh quang về cho quốc gia mà còn là cơ hội để khoa học thể thao thu thập và phân tích những yếu tố ảnh hưởng đến thành tích thi đấu. Các yếu tố đáng chú ý trong nghiên cứu này bao gồm tuổi tác, chiều cao, cân nặng, và giới tính, những yếu tố thể chất và nhân khẩu học có thể ảnh hưởng trực tiếp hoặc gián tiếp đến hiệu suất thi đấu của vận động viên. Ngoài ra, các yếu tố môi trường đặc trưng của từng kỳ Olympic (như thời tiết, độ cao, và đặc điểm địa lý) cùng những tiến bộ trong khoa học kỹ thuật, dinh dưỡng và y học thể thao cũng có thể ảnh hưởng đến thành tích thi đấu theo thời gian. Hiểu rõ tác động của những yếu tố này sẽ giúp huấn luyện viên và nhà quản lý thể thao tối ưu hóa chiến lược tập luyện, xác định tiềm năng phát triển của các vận động viên và hiệu quả hóa quá trình tuyển chọn tài năng. Đối với các liên đoàn thể thao quốc gia, những thông tin phân tích này là nền tảng quan trọng để đầu tư vào các vận động viên có tiềm năng đạt thành tích cao nhất. Tóm lại, báo cáo nhằm mang lại cái nhìn toàn diện

về những yếu tố ảnh hưởng đến thành tích thi đấu của vận động viên trong bối cảnh Olympic, đồng thời cung cấp các thông tin thực tiễn hỗ trợ chiến lược phát triển thể thao hiệu quả trên sân chơi quốc tế.

### 2.2.5 Trình bày dữ liệu

Câu chuyện chi tiết	Biểu đồ	Mô tả
Các môn thể thao phổ biến nhất theo giới tính	Biểu đồ clustered bar chart	Hiển thị môn thể thao nào phổ biến nhất với từng giới tính.
Tỷ lệ huy chương đạt được theo giới tính	Biểu đồ clustered column chart	So sánh số huy chương (Vàng, Bạc, Đồng) giữa nam và nữ.
Môn thể thao phổ biến nhất ở từng nhóm tuổi	Biểu đồ clustered bar chart	Môn thể thao phổ biến nhất trong mỗi nhóm tuổi.
Tỉ lệ độ tuổi của vận động viên nam và nữ	Biểu đồ Pie	So sánh độ tuổi của vận động viên nam và nữ.
Môn thể thao phổ biến nhất qua các kỳ thi đấu	Biểu đồ Line	Xem các môn thể thao có thay đổi về mức độ phổ biến qua thời gian.
Số lượng vận động viên nam và nữ ở mỗi môn	Biểu đồ clustered bar chart	Sự phân bổ giới tính ở từng bộ môn.
Phân phối chiều cao và cân nặng của vận động viên	Biểu đồ scatter	Thể hiện mối quan hệ giữa chiều cao và cân nặng của vận động viên
Tương quan giữa chiều cao và huy chương đạt được	Biểu đồ clustered column	Thể hiện chiều cao trung bình của từng nhóm huy chương.
Số lượng huy chương đạt được của mỗi quốc gia	Biểu đồ Treemap	So sánh số lượng huy chương đạt được theo từng quốc gia
Các môn thể thao phổ biến nhất của 2 mùa	Biểu đồ clustered column	So sánh môn thể thao phổ biến ở từng thành phố.

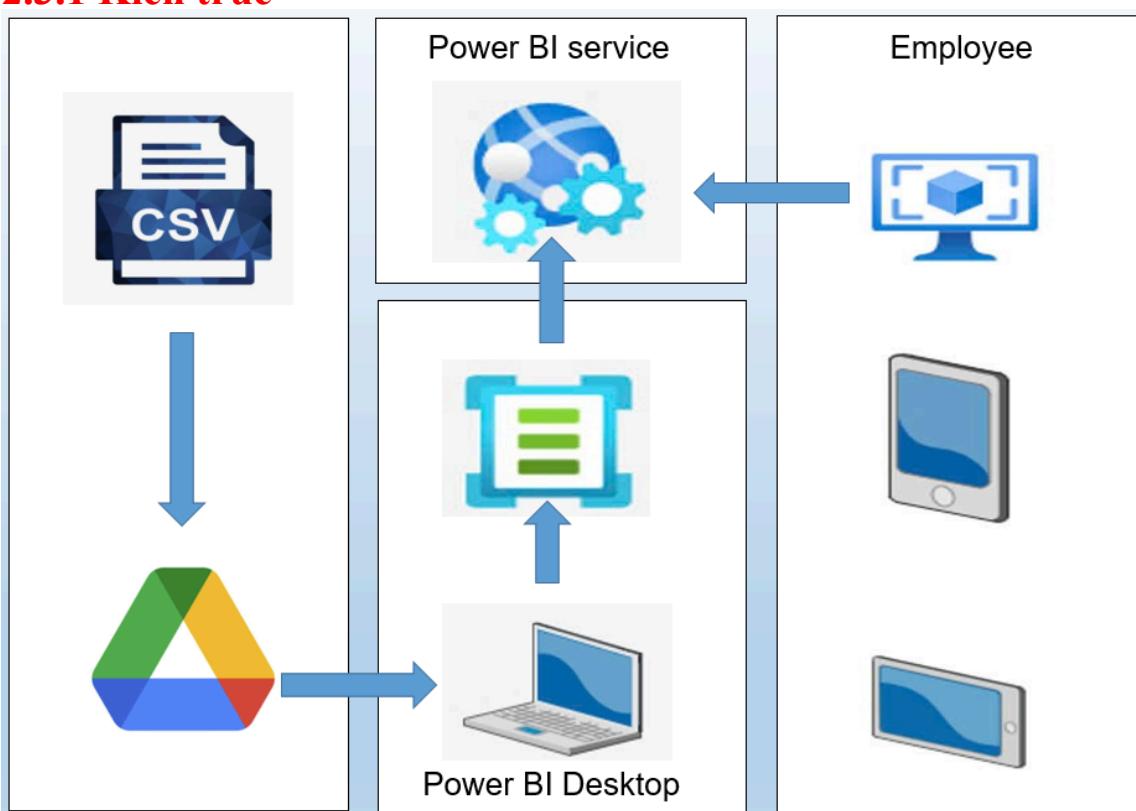
### 2.2.6.1 BẢNG TRÌNH BÀY DỮ LIỆU

## 2.2.6 Những điều cần lưu ý

- **Tính nhất quán:** Đảm bảo dữ liệu phù hợp và có nguồn gốc rõ ràng.
- **Rõ ràng và súc tích:** Chọn biểu đồ rõ ràng, dễ hiểu, tránh rối mắt.
- **Giải thích ý nghĩa:** Luôn giải thích các biểu đồ, không chỉ cung cấp dữ liệu thô mà còn cần truyền đạt ý nghĩa đằng sau các phát hiện.

## 2.3 Kiến trúc hệ thống

### 2.3.1 Kiến trúc



2.3.1 KIẾN TRÚC HỆ THỐNG

### 2.3.4 Giải thích

#### Nguồn dữ liệu (Tệp CSV) :

Tệp CSV chứa dữ liệu thô. Tệp này đại diện cho nguồn dữ liệu sẽ được phân tích và trực quan hóa.

#### Google Drive :

Tệp CSV được tải lên Google Drive, cho phép lưu trữ dễ dàng truy cập và tích hợp với các công cụ khác. Google Drive đóng vai trò là kho lưu trữ tệp, nơi dữ liệu có thể được lưu trữ, chia sẻ và truy cập khi cần.

### **Power BI Desktop :**

Từ Google Drive, dữ liệu CSV được nhập vào Power BI Desktop. Power BI Desktop là một công cụ phân tích và trực quan hóa dữ liệu mạnh mẽ, có thể được chuyển đổi, làm sạch và mô hình hóa, tạo báo cáo và bảng thông tin chi tiết dựa trên dữ liệu đã nhập.

### **Dịch vụ Power BI :**

Sau khi báo cáo được phát triển trong Power BI Desktop, báo cáo đó sẽ được xuất bản lên Power BI Service, một nền tảng đám mây để chia sẻ và cộng tác.

Dịch vụ Power BI cho phép phân phối báo cáo dễ dàng hơn, tự động làm mới dữ liệu và cập nhật bảng thông tin theo thời gian thực.

### **Quyền truy cập của nhân viên :**

Nhân viên có thể truy cập dữ liệu và thông tin chi tiết thông qua nhiều thiết bị khác nhau, bao gồm máy tính, điện thoại và máy tính bảng.

Power BI Service đảm bảo rằng nhân viên có thể truy cập từ xa và trên nhiều thiết bị vào báo cáo và bảng thông tin.

Thiết lập này cho phép nhân viên đưa ra quyết định dựa trên dữ liệu và xem thông tin mới nhất trên mọi thiết bị được hỗ trợ, giúp tăng cường tính linh hoạt và năng suất.

## **2.4 GIẢI THÍCH VỀ BỘ DỮ LIỆU KHÁCH HÀNG**

### **2.4.1 Các khái niệm**

Vận động viên (Athlete)	Người tham gia Thế vận hội, với các đặc điểm nhân khẩu học và thể chất có thể ảnh hưởng đến thành tích thi đấu.
Đội tuyển (Team)	Đội hoặc quốc gia mà vận động viên đại diện; quốc gia có thể ảnh hưởng đến cơ sở hạ tầng và cơ hội tập luyện.
Môn thể thao (Sport)	Môn thi đấu và sự kiện cụ thể mà vận động viên tham gia, vì từng môn thể thao có các yêu cầu khác nhau về thể chất và kỹ năng.
Mùa thế vận hội (Season)	Thế vận hội có hai mùa là mùa hè và mùa đông, mỗi

	mùa có các môn thể thao đặc trưng.
Huy chương (Medal)	Kết quả thi đấu của vận động viên (Vàng, Bạc, Đồng hoặc không có huy chương) là chỉ số thành tích trực tiếp để phân tích yếu tố tác động.

#### 2.4.1.1 BẢNG CÁC KHÁI NIỆM

#### 2.4.2 Các trường dữ liệu

STT	Tên trường	Kiểu dữ liệu	Mô tả
1	ID	Decimal number	Mã định danh của vận động viên, giúp xác định từng cá nhân trong dữ liệu.
2	Name	Text	Tên của vận động viên, để nhận diện cá nhân.
3	Sex	Text	Giới tính của vận động viên (M/F), có thể ảnh hưởng đến loại môn thể thao và thành tích.
4	Age	Text	Tuổi của vận động viên, ảnh hưởng đến sức mạnh thể chất và kinh nghiệm thi đấu (có thể thiếu dữ liệu cho một số vận động viên).
5	Height	Text	Chiều cao của vận động viên có thể ảnh hưởng đến khả năng thi đấu tùy theo từng môn thể thao (có thể thiếu dữ liệu).
6	Weight	Text	Cân nặng của vận động viên có thể ảnh hưởng đến khả năng thi đấu tùy theo từng môn thể thao (có thể thiếu dữ liệu).
7	Team	Text	Quốc gia hoặc đội tuyển của vận động viên.
8	NOC	Text	Mã quốc gia (mã viết tắt ba chữ

			cái của Ủy ban Olympic Quốc gia).
9	Games	Text	Tên kỳ Thế vận hội mà vận động viên tham gia, bao gồm năm và mùa.
10	Year	Decimal number	Thành phố tổ chức Thế vận hội, có thể ảnh hưởng đến thành tích do điều kiện thời tiết hoặc độ cao.
11	Season	Text	Mùa của Thế vận hội (Summer hoặc Winter).
12	City	Text	Thành phố tổ chức kỳ Thế vận hội.
13	Sport	Text	Môn thể thao mà vận động viên tham gia.
14	Event	Text	Tên sự kiện (cuộc thi cụ thể) mà vận động viên tham gia.
15	Medal	Text	Loại huy chương vận động viên giành được (Gold, Silver, Bronze) hoặc không có nếu không giành huy chương.

#### 2.4.2.1 BẢNG CÁC TRƯỜNG DỮ LIỆU

### Bảng Performance

STT	Tên trường	Kiểu dữ liệu
1	PerformanceID	Whole number
2	AthleteID	Whole number
3	EventID	Whole number
4	GameID	Whole number
5	TeamID	Whole number
6	Medal	Text

#### 2.4.2.2 Bảng Performance

### Bảng Event

STT	Tên trường	Kiểu dữ liệu
1	EventID	Whole number
2	Event	Text
3	Sport	Text

#### 2.4.2.3 Bảng Event

### Bảng Team

STT	Tên trường	Kiểu dữ liệu
1	TeamID	Whole number
2	Noc	Text
3	region	Text
4	notes	Text
5	Team	Text

#### 2.4.2.4 Bảng Team

### Bảng Athlete

STT	Tên trường	Kiểu dữ liệu
1	AthleteID	Whole number
2	Sex	Text
3	Age	Whole number
4	Weight	Whole number
5	Height	Whole number
6	Name	Text

#### 2.4.2.5 Bảng Athlete

### Bảng Game

STT	Tên trường	Kiểu dữ liệu
1	GameID	Whole number
2	City	Text
3	Game	Text
4	Year	whole number
5	Season	Text

#### 2.4.2.6 Bảng Game

### 3 Làm sạch và chuyển đổi dữ liệu

#### 3.1 Chuẩn bị dữ liệu

##### 3.1.1 Giải pháp lưu trữ dữ liệu

Google Drive được lựa chọn làm giải pháp lưu trữ dữ liệu. Google Drive là một nền tảng lưu trữ đám mây miễn phí, cung cấp một cách dễ dàng để lưu trữ, tổ chức và chia sẻ dữ liệu. Với dung lượng miễn phí lên đến 15GB và khả năng truy cập từ bất kỳ thiết bị nào, Google Drive là một lựa chọn lý tưởng cho các dự án nhỏ và vừa, đặc biệt khi làm việc nhóm.

**Lý do lựa chọn Google Drive:**

- Phù hợp với quy mô dự án:** Dự án phân tích các yếu tố ảnh hưởng đến thành tích của vận động viên có thể bắt đầu với quy mô nhỏ. Google Drive cung cấp một giải pháp đơn giản, linh hoạt và dễ sử dụng để lưu trữ dữ liệu mà không yêu cầu thiết lập hệ thống phức tạp.
- Dễ dàng sử dụng và chia sẻ:** Google Drive cho phép lưu trữ và chia sẻ các tệp dữ liệu với nhiều định dạng (CSV, Excel, hình ảnh, video). Các thành viên có thể dễ dàng tải lên, chỉnh sửa và làm việc trên dữ liệu mà không cần cài đặt phần mềm chuyên dụng.
- Khả năng tổ chức:** Dữ liệu được lưu trữ trong các thư mục riêng biệt theo cấu trúc logic (Raw Data, Processed Data, Reports) giúp dễ dàng quản lý và tìm kiếm.
- Tích hợp tốt với các công cụ khác:** Google Drive hỗ trợ tích hợp với các công cụ phân tích dữ liệu như Power BI hoặc Python, giúp dễ dàng tải dữ liệu lên và thực hiện các thao tác phân tích, trực quan hóa.
- Chi phí thấp:** Google Drive cung cấp dung lượng miễn phí, tiết kiệm chi phí đáng kể cho dự án, đặc biệt là trong giai đoạn đầu.

##### 3.1.2 Giải pháp phân bố dữ liệu

###### 3.1.2.1 Ý nghĩa việc phân bố dữ liệu

- Tối ưu hóa quản lý:** Phân bố dữ liệu rõ ràng giúp dễ dàng tìm kiếm, quản lý và tránh nhầm lẫn giữa các file.

- **Tăng hiệu quả làm việc nhóm:** Việc tổ chức và chia sẻ dữ liệu có cấu trúc hỗ trợ quá trình cộng tác, đảm bảo mọi người luôn sử dụng phiên bản dữ liệu mới nhất.
- **Linh hoạt trong xử lý:** Google Drive hỗ trợ chỉnh sửa và làm việc trực tiếp với các công cụ như Google Sheets, Power BI, hoặc Python, giúp tiết kiệm thời gian xử lý.
- **Đảm bảo an toàn:** Tính năng lưu trữ đám mây giúp bảo vệ dữ liệu trước nguy cơ mất mát do sự cố máy tính hoặc phần cứng.
- **Hỗ trợ phân tích và trình bày:** Dữ liệu có cấu trúc tốt giúp bạn nhanh chóng tạo các báo cáo phân tích và trình bày kết quả một cách rõ ràng, chuyên nghiệp.

### 3.1.2.2 Trình bày cách phân bố dữ liệu

Loại dữ liệu	Tên bảng	Trường dữ liệu
Olympic	Performance	PerformanceID, AthleteID, EventID, GameID, TeamID, Medal
Vận động viên	Athlete	AthleteID, Name, Sex, Age, Weight, Height
Môn thể thao	Game	GameID, City, Games, Season, Year
Sự kiện	Event	EventID, Event, Sport
Quốc gia	Team	TeamID, Team, Noc, Region

#### 3.1.2.2.1 Trình bày cách phân bố dữ liệu

## 3.2 Làm sạch dữ liệu

### 3.2.1 Các vấn đề ảnh hưởng tới dữ liệu

#### 1. Thiếu dữ liệu (Data Missingness)

Các hàng có giá trị NA cho các cột như chiều cao và cân nặng sẽ dẫn đến thiếu thông tin, gây ảnh hưởng đến việc phân tích và kết luận. Các mô hình học máy

và các phân tích thống kê có thể đưa ra kết quả sai lệch khi dữ liệu không đầy đủ.

## 2. Sự sai lệch (Bias)

Dữ liệu thiếu có thể làm cho mẫu không đại diện đầy đủ cho tổng thể, dẫn đến sai lệch trong các phân tích. Ví dụ, nếu các giá trị chiều cao và cân nặng chủ yếu thiếu ở các vận động viên từ một số khu vực hoặc một số môn thể thao, phân tích sẽ bị ảnh hưởng nghiêm trọng.

## 3. Độ chính xác của mô hình (Model Accuracy)

Các mô hình phân tích dữ liệu và học máy có thể hoạt động kém hiệu quả khi có nhiều giá trị NA, do thiếu các biến đầu vào cần thiết. Điều này làm giảm độ chính xác và tính ổn định của các dự đoán.

## 4. Khả năng tính toán và phân tích (Computational Challenges)

Các công cụ phân tích có thể không xử lý tốt dữ liệu thiếu, dẫn đến việc phải xử lý bổ sung, như việc loại bỏ dữ liệu, điền dữ liệu (imputation), hoặc phát triển các mô hình mạnh mẽ hơn để xử lý các trường hợp thiếu dữ liệu.

### 3.2.2 Các tiêu chí đánh giá chất lượng dữ liệu

#### Tính chuẩn xác (validity)

- Mandatory constraints (Bắt buộc về tính đầy đủ của dữ liệu): Một số trường thông tin trong bảng bắt buộc không được để trống.
- Data-type constraints (Thống nhất về định dạng dữ liệu): Các giá trị trong một cột phải thuộc một dạng dữ liệu nhất định.
- Range constraints (Giới hạn phạm vi dữ liệu): Giá trị tối thiểu và tối đa cho dạng dữ liệu số hoặc ngày tháng.
- Foreign-key constraints (Thống nhất về foreign-key): đảm bảo mỗi liên kết giữa các bảng với nhau, giá trị của cột foreign-key phải tương ứng với giá trị primary-key ở bảng khác.
- Unique constraints (Yêu cầu về tính duy nhất): Một số cột trong bảng phải là duy nhất trong bảng dữ liệu, ví dụ như cột customer\_ID trong bảng dim customer.

- Regular expression patterns (Ràng buộc về cách thức diễn đạt): Các trường dữ liệu có định dạng free-text cần có quy tắc xác thực theo cách này, để dữ liệu được lưu trữ một cách thống nhất.
- Cross-field validation (Xác thực chéo giữa các trường dữ liệu): Một số cột dữ liệu lặp lại giữa các bảng khác nhau cần được xác thực chéo để đảm bảo tính đúng đắn về các giá trị dữ liệu.

**Tính chính xác (accuracy):** Phản ánh mức độ trung thực của dữ liệu, các giá trị đo được trong bảng dữ liệu có gần với giá trị đích/ giá trị thực tế hay không.

**Tính đầy đủ (completeness):** Mức độ chi tiết và toàn diện của dữ liệu về một vấn đề, chủ đề.

**Tính nhất quán (consistency):** Mức độ thống nhất của dữ liệu khi đổi chiếu trên toàn bộ hệ thống dữ liệu và công cụ lưu trữ khác nhau.

**Tính đồng nhất (uniformity):** Đơn vị đo lường giống nhau được sử dụng trong tất cả các hệ thống dữ liệu.

**Tính minh bạch của nguồn dữ liệu (traceability):** Dữ liệu cần đảm bảo có thể truy tìm lại nguồn thu thập.

**Tính kịp thời (timeliness):** Dữ liệu phải được thu thập và cập nhật từ các nguồn mới nhất, “up-to-date” nhất.

### 3.2.3 Các bước làm sạch dữ liệu

#### 3.2.3.1 Trình bày các bước làm sạch

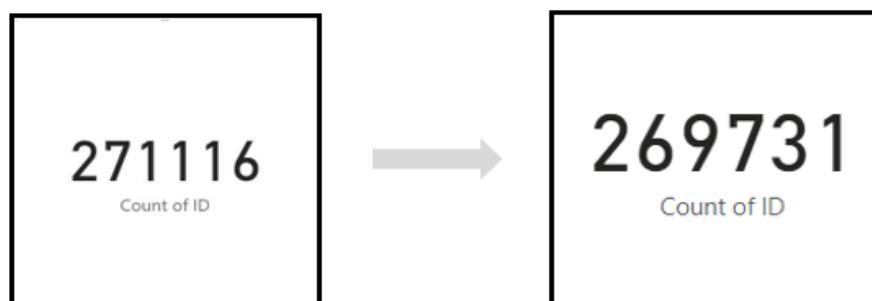
- Dữ liệu sai định dạng: chuyển về đúng định dạng
- Dữ liệu bị thiếu: Đánh dấu và loại bỏ hoặc thay thế các dữ liệu bị thiếu

#### 3.2.3.2 Trình bày các bước làm sạch trong phạm vi dự án

- Xóa các bản ghi trùng lặp

A <sub>C</sub> Event	A <sub>C</sub> Games	A <sub>C</sub> Team	A <sub>C</sub> Name	1.2 ID	A <sub>C</sub> Sex	A <sub>C</sub> Age	A <sub>C</sub> Height
1 Equestrianism Mixed Jumping, I...	1900 Summer	France	Henri Leclerc	67708	M	24	175
2 Equestrianism Mixed Jumping, I...	1900 Summer	France	Henri Leclerc	67708	M	24	175
3 Cycling Men's Sprint	1908 Summer	Netherlands	Gerard Dagobert Henri Bosch van Dra...	13725	M	20	175
4 Cycling Men's Sprint	1908 Summer	Netherlands	Gerard Dagobert Henri Bosch van Dra...	13725	M	20	175
5 Cycling Men's Sprint	1908 Summer	Netherlands	Antonie "Anton" Gerrits	39812	M	22	175
6 Cycling Men's Sprint	1908 Summer	Netherlands	Antonie "Anton" Gerrits	39812	M	22	175
7 Cycling Men's Sprint	1908 Summer	Netherlands	Dorotheus Magdalenus "Dorus" Nijland	86180	M	28	175
8 Cycling Men's Sprint	1908 Summer	Netherlands	Dorotheus Magdalenus "Dorus" Nijland	86180	M	28	175
9 Cycling Men's Sprint	1908 Summer	Netherlands	Johannes Jacobus van Spengen	124923	M	21	175
10 Cycling Men's Sprint	1908 Summer	Netherlands	Johannes Jacobus van Spengen	124923	M	21	175
11 Cycling Men's Sprint	1908 Summer	United States	George Guthrie Cameron	17624	M	26	175
12 Cycling Men's Sprint	1908 Summer	United States	George Guthrie Cameron	17624	M	26	175
13 Cycling Men's Sprint	1908 Summer	France	Andr Auffray	5935	M	23	175
14 Cycling Men's Sprint	1908 Summer	France	Andr Auffray	5935	M	23	175
15 Cycling Men's Sprint	1908 Summer	France	Marcel Carlos Paul Gaston Delaplane...	27195	M	26	175
16 Cycling Men's Sprint	1908 Summer	France	Marcel Carlos Paul Gaston Delaplane...	27195	M	26	175
17 Cycling Men's Sprint	1908 Summer	France	Gaston Dreyfus	30019	M	24	175
18 Cycling Men's Sprint	1908 Summer	France	Gaston Dreyfus	30019	M	24	175
19 Cycling Men's Sprint	1908 Summer	France	mile Marchal	74907	M	24	175
20 Cycling Men's Sprint	1908 Summer	France	mile Marchal	74907	M	24	175
21 Cycling Men's Sprint	1908 Summer	France	Georges Perrin	93561	M	16	175
22 Cycling Men's Sprint	1908 Summer	France	Georges Perrin	93561	M	16	175
23 Cycling Men's Sprint	1908 Summer	France	Andr Pouain	96534	M	24	175
24 Cycling Men's Sprint	1908 Summer	France	Andr Pouain	96534	M	24	175

### 3.2.3.2.1 Xóa các bản ghi trùng lặp



### 3.2.3.2.2 Dữ liệu đã được xóa

- Cho cột tiêu đề lên dòng đầu tiên

The screenshot shows the Power BI Data Editor interface. In the top navigation bar, 'Manage' is selected. In the 'Queries [2]' section, 'noc\_regions' is the active query. A context menu is open over the first row of the 'athlete\_events' table, with 'Copy Entire Table' highlighted. Other options in the menu include 'Use First Row as Headers', 'Add Custom Column...', 'Add Column From Examples...', and 'Invoke Custom Function...'. The table itself has three columns: 'Column1', 'Column2', and 'Column3', with some sample data visible.

### 3.2.3.2.3 Cho cột tiêu đề lên dòng đầu tiên

- Chuyển các giá trị NA của cột Age về giá trị trung vị



A <sup>B</sup> <sub>C</sub>	Age
NA	

I <sup>2</sup> <sub>3</sub>	Age
24	
24	
24	
24	
24	
24	
24	
24	
24	
24	

#### 3.2.3.2.4 Chuyển các giá trị NA của cột Age về giá trị trung vị

- Chuyển các giá trị NA của cột Height về giá trị trung vị

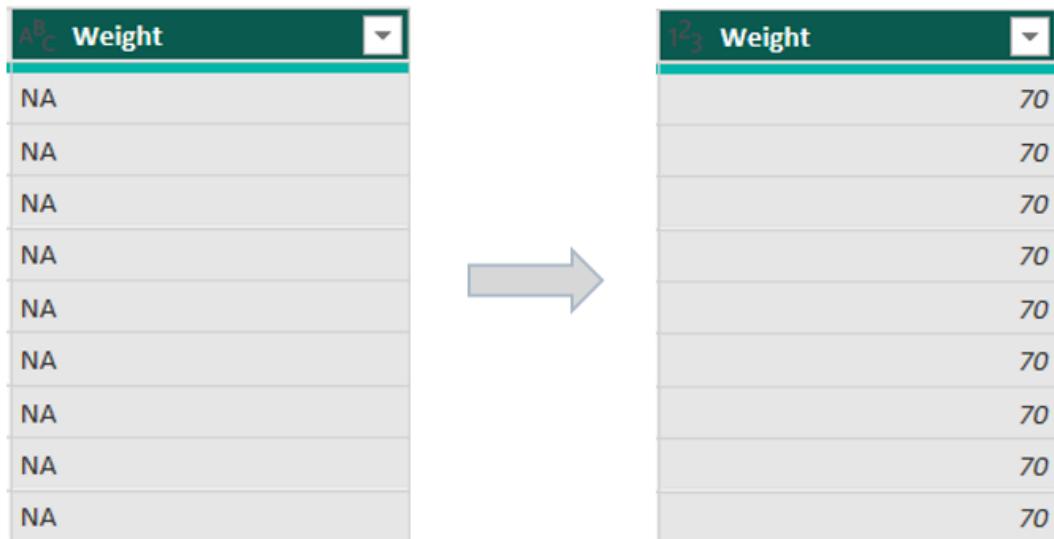


A <sup>B</sup> <sub>C</sub>	Height
180	
170	
NA	
NA	
185	
185	
185	
185	

I <sup>2</sup> <sub>3</sub>	Height
175	
175	
175	
175	
175	
175	
175	
175	

#### 3.2.3.2.5 Chuyển các giá trị NA của cột Height về giá trị trung vị

- Chuyển các giá trị NA của cột Weight về giá trị trung vị



The diagram illustrates a transformation process. On the left, there is a table titled "Weight" with columns A, B, and C. The "Weight" column contains nine rows, all of which are labeled "NA". On the right, there is another table with the same structure, also titled "Weight". This table shows the result of the transformation: the "Weight" column now contains nine rows, each labeled "70". A large grey arrow points from the left table to the right table, indicating the direction of the transformation.

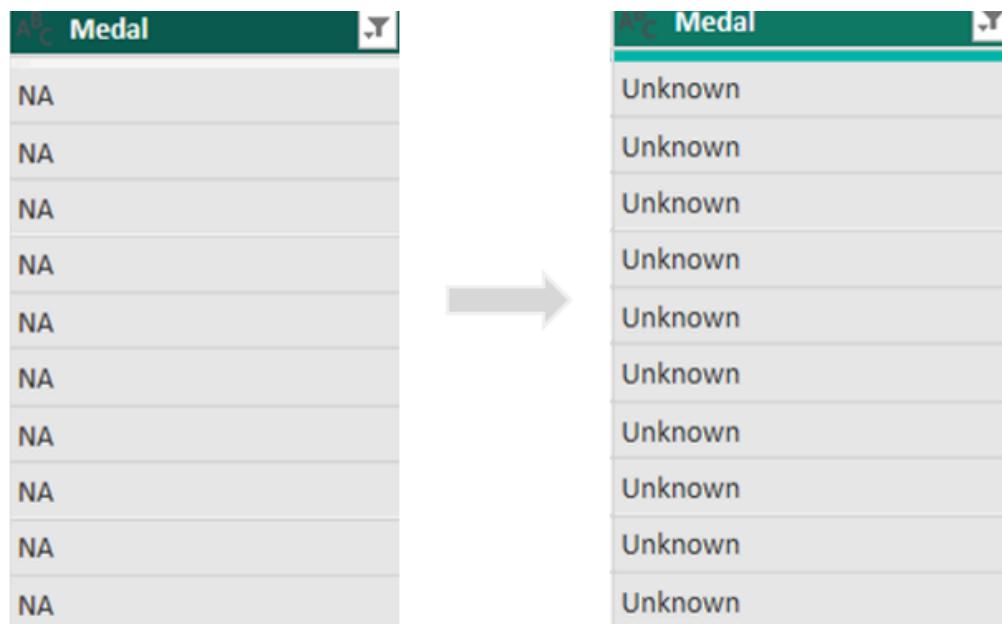
A	B	C	Weight
			NA

A	B	C	Weight
			70
			70
			70
			70
			70
			70
			70
			70
			70

### 3.2.3.2.6 Chuyển các giá trị NA của cột Weight về giá trị trung vị

- Chuyển các giá trị NA của cột Medal thành “Unknown”



The diagram illustrates a transformation process. On the left, there is a table titled "Medal" with columns A, B, and C. The "Medal" column contains nine rows, all of which are labeled "NA". On the right, there is another table with the same structure, also titled "Medal". This table shows the result of the transformation: the "Medal" column now contains nine rows, each labeled "Unknown". A large grey arrow points from the left table to the right table, indicating the direction of the transformation.

A	B	C	Medal
			NA

A	B	C	Medal
			Unknown

### 3.2.3.2.6 Chuyển các giá trị NA của cột Medal thành “Unknown”

### 3.3 Chuyển đổi dữ liệu

#### 3.3.1 Các trường hợp cần chuyển đổi

- Chuyển đổi kiểu dữ liệu của trường ‘Age’ từ dạng text về dạng whole number
- Chuyển đổi kiểu dữ liệu của trường ‘Weight’ từ dạng text về dạng whole number
- Chuyển đổi kiểu dữ liệu của trường ‘Height’ từ dạng text về dạng whole number

#### 3.3.2 Các kỹ thuật chuyển đổi

- **Sửa đổi (Revising)**: là một hình thức sửa đổi dữ liệu bằng cách giảm mô hình dữ liệu về dạng “bình thường” mà không có dư thừa hoặc một-nhiều giá trị trong một cột. Chuẩn hóa làm giảm nhu cầu lưu trữ và làm cho mô hình dữ liệu ngắn gọn hơn và dễ đọc hơn đối với các nhà phân tích.
- **Làm sạch (Data cleansing)**: dữ liệu chuyển đổi các giá trị dữ liệu để tương thích với định dạng.
- **Sửa đổi/chuyển đổi (format revision/conversion)**: định dạng thay thế các ký tự không tương thích, chuyển đổi đơn vị, chuyển đổi định dạng ngày tháng và thay đổi kiểu dữ liệu.
- **Tái cấu trúc khóa (key restructuring)**: tạo ra các số nhận dạng chung ngoài các giá trị có ý nghĩa tích hợp, vì vậy chúng có thể được sử dụng như các khóa cố định, duy nhất trên các bảng.
- **Lọc dữ liệu trùng lặp (Deduplication)**: có nghĩa là xác định và loại bỏ các bản ghi trùng lặp.
- **Xác thực dữ liệu (Data validation)**: đánh giá tính hợp lệ của một bản ghi bằng tính đầy đủ của dữ liệu, thường bằng cách loại trừ các bản ghi không đầy đủ.
- **Việc loại bỏ các cột không sử dụng và lặp lại (removing unused and repeated columns)**: cho phép bạn chọn các trường bạn muốn sử dụng làm tính

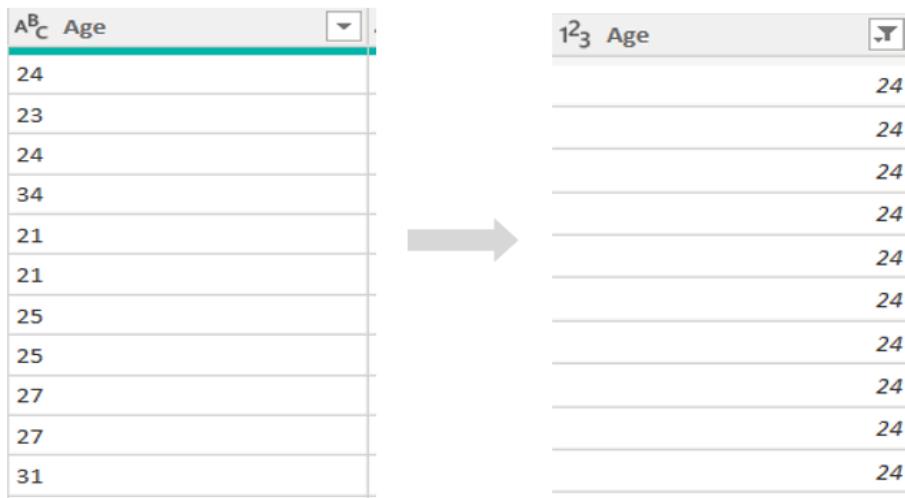
năng, tức là các biến đầu vào cho mô hình dự đoán. Nó cũng có thể cải thiện hiệu suất và tính dễ đọc tổng thể của một mô hình.

- **Tính toán chéo (derivation):** bao gồm các phép tính cột chéo đơn giản.
- **Tóm tắt (summarization):** bao gồm việc sử dụng các hàm tổng hợp để tạo ra các giá trị tóm tắt.
- **Xoay vòng (pivoting):** biến các giá trị hàng thành cột và ngược lại.
- **Sắp xếp và lập chỉ mục (sorting, ordering and indexing):** tổ chức các bản ghi theo một số thứ tự để cải thiện hiệu suất tìm kiếm.
- **Chia tỷ lệ và chuẩn hóa (scaling, standardization and normalization):** đặt các con số trên một thang đo nhất quán, chẳng hạn như các phân số của độ lệch chuẩn trong chuẩn hóa điểm Z. Điều này cho phép các con số khác nhau được so sánh với nhau.
- **Vector (vectorization):** hóa chuyển đổi dữ liệu không phải số thành mảng số. Có rất nhiều ứng dụng học máy của những chuyển đổi này, chẳng hạn như để xử lý ngôn ngữ tự nhiên (NLP) và nhận dạng hình ảnh.
- **Tách biệt (splitting):** việc phân tách bao gồm việc phân chia các giá trị thành các phần câu thành của chúng. Các giá trị dữ liệu thường được kết hợp trong cùng một trường vì tính riêng trong thu thập dữ liệu, nhưng có thể cần được tách riêng để thực hiện phân tích chi tiết hơn.
- **Lọc loại trừ (filtering):** dữ liệu trên cơ sở các giá trị hàng hoặc cột nhất định.
- **Kết nối (joining):** là hành động liên kết dữ liệu giữa các bảng.
- **Hợp nhất (Merging):** còn được gọi là thêm hoặc kết hợp, kết hợp các bản ghi từ nhiều bảng, bằng cách kết hợp hai bảng sử dụng một cột chung
- Các kỹ thuật áp dụng cho dự án
  - + Sắp xếp và lập chỉ mục

+ Sửa lỗi/chuyển đổi

### 3.3.3 Trình bày các phép chuyển đổi trong dự án

- Chuyển định dạng của cột ‘Age’ từ dạng text về dạng whole number



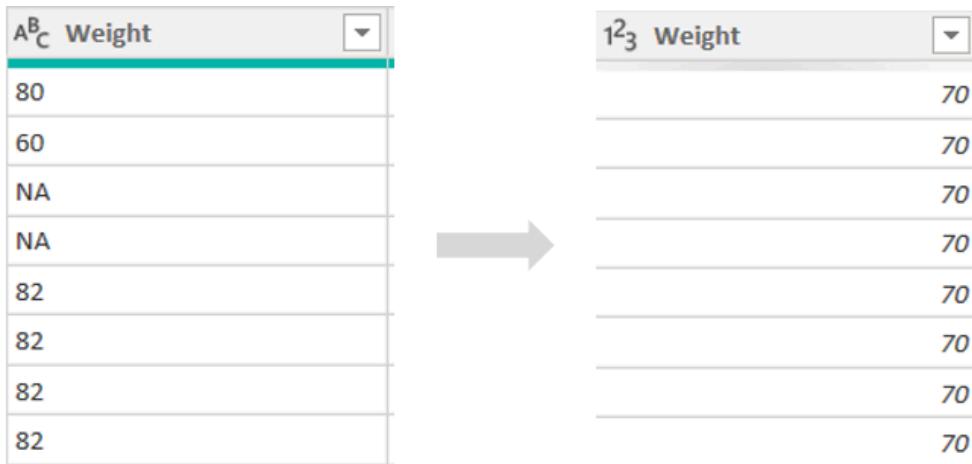
A <sup>B</sup> <sub>C</sub>	Age
	24
	23
	24
	34
	21
	21
	25
	25
	27
	27
	31

1 <sup>2</sup> <sub>3</sub>	Age
	24
	24
	24
	24
	24
	24
	24
	24
	24
	24

#### 3.3.3.1 Chuyển định dạng của cột ‘Age’ từ dạng text về dạng whole number

- Chuyển định dạng của cột ‘Weight’ từ dạng text về dạng whole number



A <sup>B</sup> <sub>C</sub>	Weight
	80
	60
	NA
	NA
	82
	82
	82
	82

1 <sup>2</sup> <sub>3</sub>	Weight
	70
	70
	70
	70
	70
	70
	70
	70

#### 3.3.3.2 Chuyển định dạng của cột ‘Weight’ từ dạng text về dạng whole number

- Chuyển định dạng của cột ‘Height’ từ dạng text về dạng whole number



A	B	C	Height
			180
			170
			NA
			NA
			185
			185
			185
			185

A	B	C	Height
			175
			175
			175
			175
			175
			175
			175
			175

### 3.3.3.3 Chuyển định dạng của cột ‘Height’ từ dạng text về dạng whole number

## 4 Xử lý dữ liệu

### 4.1 Chuẩn hóa dữ liệu

#### 4.1.1 Trình bày các bước chuẩn hóa trong dự án

Bước 1: Đảm bảo chuẩn 1NF (first normal form)

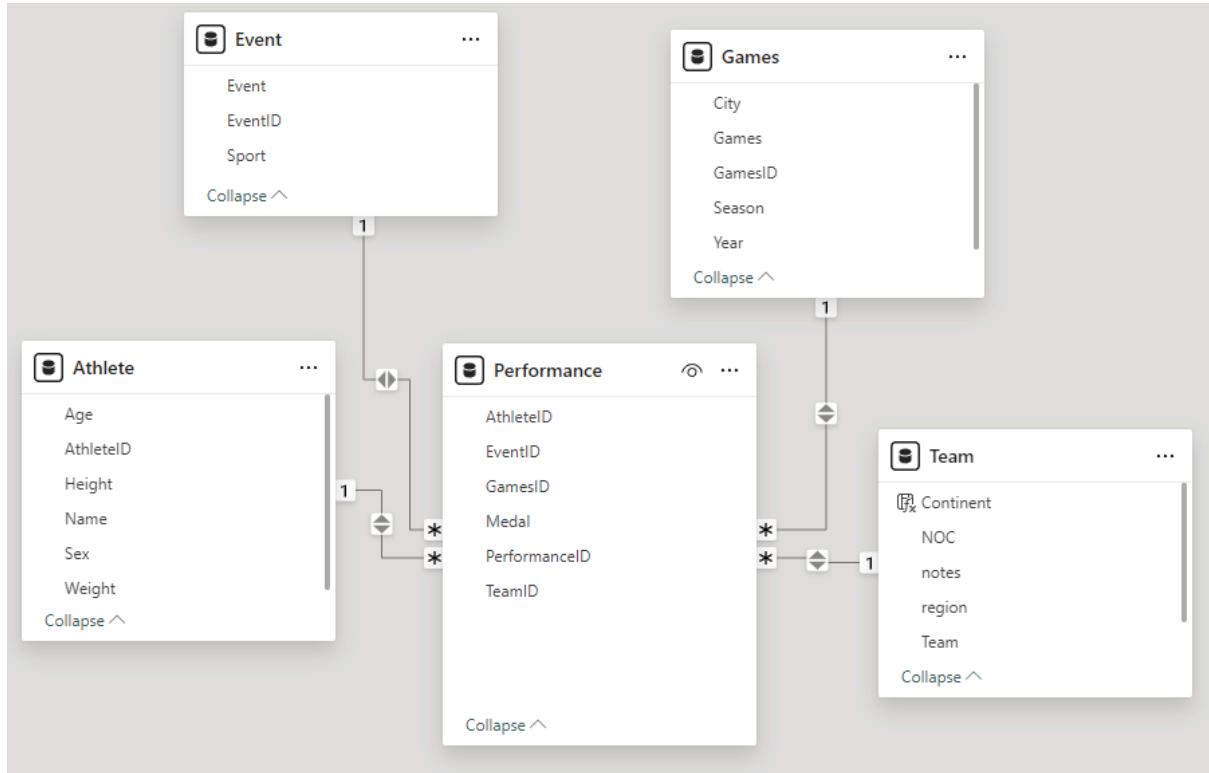
- Đảm bảo dữ liệu dạng nguyên tử: Mỗi ô trong bảng chỉ chứa một giá trị duy nhất. Các trường như name, sex, age, height, weight, team, noc, games, year, season, city, sport, event, medal đều là nguyên tử.

- Loại bỏ nhóm lặp lại: Đảm bảo không có các nhóm dữ liệu lặp lại trong cùng một bảng. Điều này yêu cầu dữ liệu của mỗi vận động viên được tách riêng.

Bước 2: Đảm bảo chuẩn 2NF (second normal form)

- Tạo bảng độc lập: Phân chia các bảng sao cho mỗi bảng chỉ chứa dữ liệu liên quan trực tiếp đến một thực thể, và không có thuộc tính không khóa nào phụ thuộc vào một phần khóa chính.

- Loại bỏ sự phụ thuộc từng phần: Đảm bảo tất cả các thuộc tính không khóa phải phụ thuộc hoàn toàn vào khóa chính. Ví dụ, các thuộc tính như medal chỉ phụ thuộc vào athlete, không phụ thuộc vào một phần khóa chính.



#### 4.1.1.1 Mô hình dữ liệu

### 4.1.2 Các loại mô hình hóa

Các loại mô hình:

- **Conceptual data models (Mô hình Dữ liệu Khái niệm)**: Mô hình này ghi lại các mối quan hệ cấp cao nhất giữa các thực thể khác nhau. Nó chứa các thực thể thiết yếu và mối quan hệ giữa chúng, nhưng không có thuộc tính hoặc khóa chính nào được chỉ định.
- **Logical Data Model (Mô hình Dữ liệu Logic)**: Mô hình này xác định cấu trúc của thông tin mà không quan tâm đến việc nó được lưu trữ vật lý như thế nào trong cơ sở dữ liệu. Mục tiêu chính của nó là ghi lại cấu trúc, quy trình, quy tắc và mối quan hệ của dữ liệu kinh doanh.
- **Physical data models (Mô hình Dữ liệu Vật lý)**: Mô hình này mô tả cách hệ thống sẽ được triển khai bằng cách sử dụng một hệ thống quản lý cơ sở

dữ liệu cụ thể. Nó thường được tạo bởi các chuyên gia quản trị dữ liệu và nhà phát triển, với mục đích triển khai cơ sở dữ liệu thực tế.

**Mô hình áp dụng cho dự án:** Physical data models (Mô hình Dữ liệu Vật lý):

### 4.1.3 Các tiêu chí đánh giá mô hình dữ liệu

#### 5 Tiêu chí đánh giá mô hình dữ liệu tốt:

**Clearness** (Tính rõ ràng):

- Sự dễ hiểu đối với những người sử dụng.
- Hầu hết thời gian developers đọc mã thay vì viết. Vì vậy chúng ta cần hiểu rõ ràng những gì chúng ta đang làm với dữ liệu của mình.

**Flexibility** (Tính linh hoạt):

- Khả năng phát triển của mô hình mà không cần phải tác động quá lớn đến các đoạn code.
- Công ty startup mà bạn làm việc đang phát triển, vì vậy các hệ thống sẽ thay đổi và các mô hình dữ liệu đãng sau sẽ cần phải phát triển theo thời gian.

**Performance** (Hiệu suất)

- Đây là một chủ đề rất rộng.
- Trong môn học này sẽ không nói về các nhà cung cấp cơ sở dữ liệu (database vendors) hoặc một số chỉnh sửa kỹ thuật để cải thiện tốc độ đọc và ghi dữ liệu.
- Cách thức thiết kế data model đúng đắn cũng đem lại lợi ích về hiệu suất.

**Productivity** (Năng suất)

- Dưới góc nhìn của lập trình viên (developer), chắc hẳn bạn sẽ muốn có một mô hình dữ liệu dễ làm việc mà không cần sử dụng nhiều thời gian.

**Traceability** (Khả năng truy xuất nguồn gốc)

- Cuối cùng, các công ty không chỉ muốn có dữ liệu liên quan đến người dùng của mình mà còn có dữ liệu liên quan đến chính hệ thống.

- Dữ liệu có thể cung cấp thông tin những gì đã xảy ra trong quá khứ, những giá trị công ty có tại một thời điểm nào đó.

#### 4.1.4 Trình bày các bước mô hình hóa

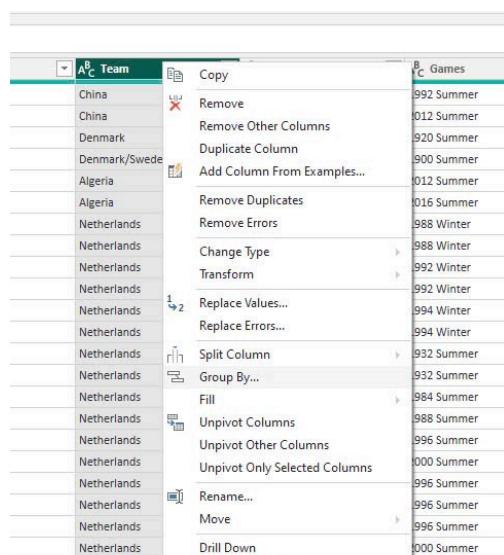
- Mô hình dữ liệu Athlete:
  - + Chọn cột ID đổi tên thành AthleteID để dễ nhận biết

The screenshot shows the Power BI Data Editor interface. On the left, there's a sidebar with 'Queries [3]' containing three items: 'noc\_regions', 'athlete\_events', and 'Performance'. The 'Performance' query is currently selected. The main area displays a table with three columns: 'AthleteID', 'Name', and 'Sex'. The 'AthleteID' column is highlighted with a green border. The table has 8 rows of data. The first row contains '1 A Dijiang'. The last row contains '5 Christine Jacoba Aaftink'.

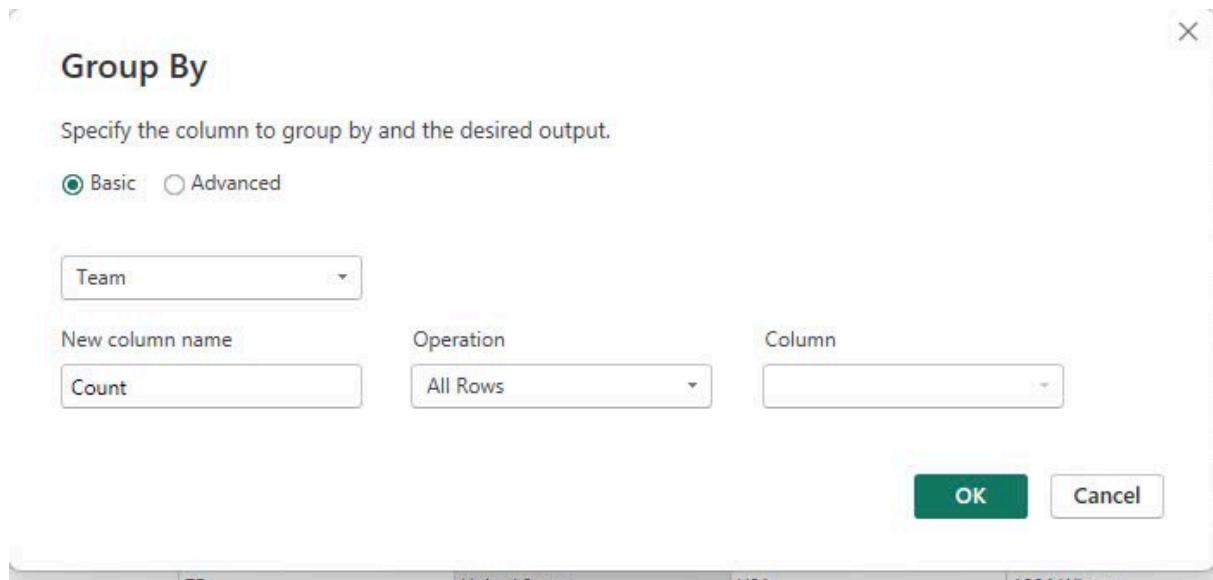
AthleteID	Name	Sex
1	A Dijiang	M
2	A Lamusi	M
3	Gunnar Nielsen Aaby	M
4	Edgar Lindenau Aabye	M
5	Christine Jacoba Aaftink	F
5	Christine Jacoba Aaftink	F
5	Christine Jacoba Aaftink	F
5	Christine Jacoba Aaftink	F

##### 4.2.3.1 Đổi tên cột ID thành AthleteID

- Mô hình dữ liệu Team:
  - + Group by cột Team và tạo cột TeamID



##### 4.2.3.2 Group by cột Team



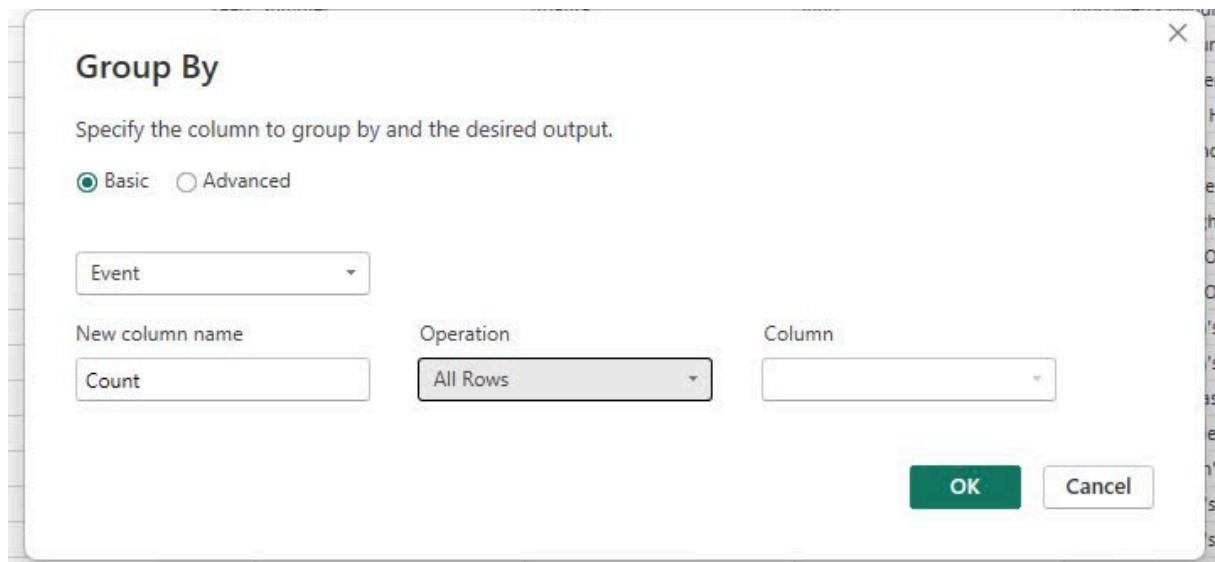
#### 4.2.3.3 Giao diện group by của team

	Team	Count	Index
1	China	Table	0
2	Denmark	Table	1
3	Denmark/Sweden	Table	2
4	Algeria	Table	3
5	Netherlands	Table	4
6	Argentina	Table	5
7	United States	Table	6
8	Azerbaijan	Table	7
9	Belarus	Table	8

#### 4.2.3.4 Tạo cột ID cho team

- Mô hình dữ liệu Event:

+ Group by cột Event và tạo cột EventID

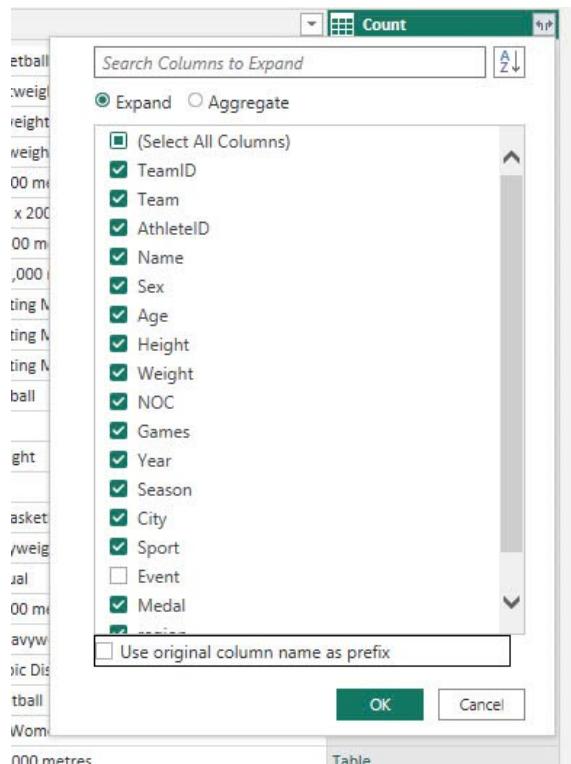


#### 4.2.3.5 Group by cột Event

The screenshot shows the Power BI desktop interface with a query named 'mohinh2'. The 'Event' column has been grouped by 'Event' and a new 'EventID' column has been created. The table contains 16 rows of event data, each with an 'EventID' and a corresponding 'Event' name. The 'Event' column is highlighted in red.

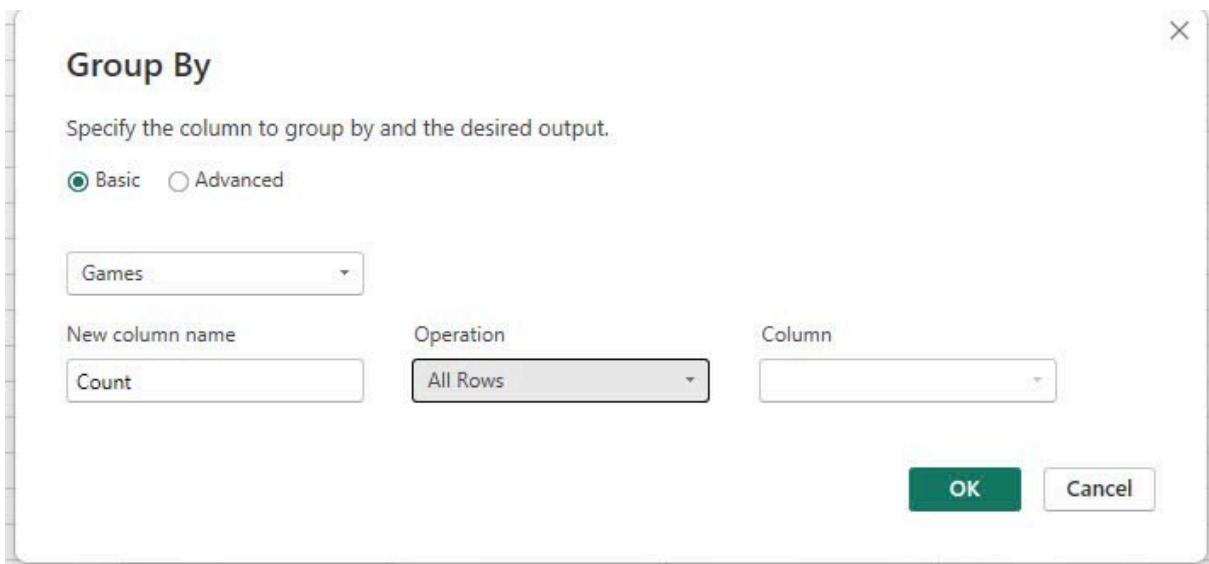
EventID	Event	Count
E1	Basketball Men's Basketball	Table
E2	Judo Men's Extra-Lightweight	Table
E3	Boxing Men's Middleweight	Table
E4	Wrestling Men's Lightweight, Greco-Roman	Table
E5	Swimming Women's 200 metres Freestyle	Table
E6	Swimming Women's 4 x 200 metres Freestyle Relay	Table
E7	Speed Skating Men's 500 metres	Table
E8	Speed Skating Men's 1,000 metres	Table
E9	Short Track Speed Skating Men's 500 metres	Table
E10	Short Track Speed Skating Men's 1,000 metres	Table
E11	Short Track Speed Skating Men's 5,000 metres Relay	Table
E12	Softball Women's Softball	Table
E13	Hockey Men's Hockey	Table
E14	Judo Men's Middleweight	Table
E15	Curling Men's Curling	Table
E16	Basketball Women's Basketball	Table

#### 4.2.3.6 tạo cột EventID



#### 4.2.3.7 Khôi phục các bảng chưa gộp

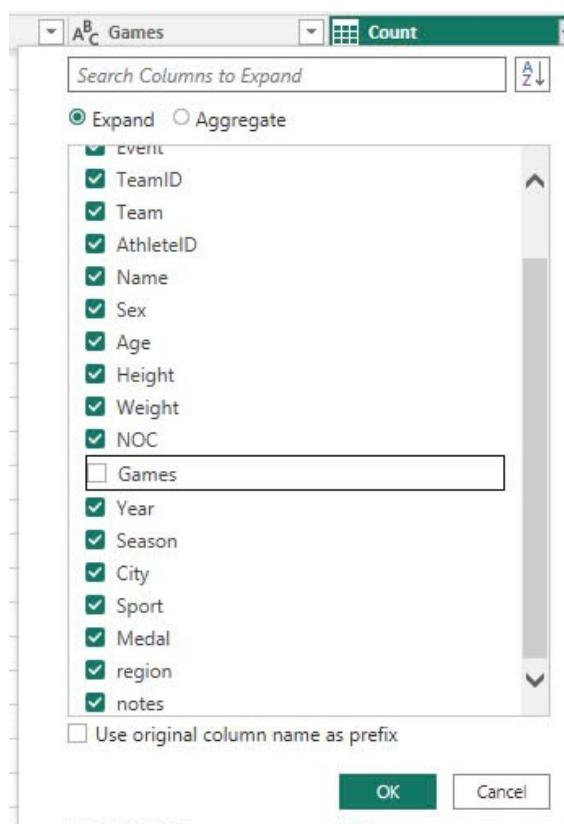
- Mô hình dữ liệu Games:
  - + Group by cột Games và tạo cột GamesID



#### 4.2.3.8 Group by cột Games

	A GamesID	B Games	C Count
1	G1	1992 Summer	Table
2	G2	2008 Summer	Table
3	G3	2012 Summer	Table
4	G4	2004 Summer	Table
5	G5	1948 Summer	Table
6	G6	2016 Summer	Table
7	G7	1936 Summer	Table
8	G8	1988 Summer	Table
9	G9	1996 Summer	Table
10	G10	2000 Summer	Table
11	G11	1984 Summer	Table
12	G12	1952 Summer	Table
13	G13	1976 Summer	Table
14	G14	1960 Summer	Table
15	G15	1972 Summer	Table
16	G16	1964 Summer	Table
17	G17	1968 Summer	Table
18	G18	1956 Summer	Table
19	G19	1980 Summer	Table
20	G20	1920 Summer	Table
...	...	...	...

#### 4.2.3.9 tạo cột GamesID



#### 4.2.3.10 Khôi phục các bảng chưa gộp của Game

- Mô hình dữ liệu Performance:
  - + Tạo cột PerformanceID

#### 4.2.3.11 Tạo cột PerformanceID

### 4.1.5 Trình bày các bước tạo bảng dữ liệu

- Tạo bảng reference từ clean\_data rename thành Performance giữ lại các cột "PerformanceID", "GamesID", "EventID", "TeamID", "AthleteID", "Medal"

	A <sub>B</sub> PerformanceID	A <sub>B</sub> GamesID	A <sub>B</sub> EventID	A <sub>B</sub> TeamID	L2 AthleteID	A <sub>B</sub> Medal
1	P1	G1	E1	T1		1 Unknown
2	P2	G1	E1	T1		41426 Unknown
3	P3	G1	E1	T1		50576 Unknown
4	P4	G1	E1	T1		69323 Unknown
5	P5	G1	E1	T1		72723 Unknown
6	P6	G1	E1	T1		109208 Unknown
7	P7	G1	E1	T1		113134 Unknown
8	P8	G1	E1	T1		116458 Unknown
9	P9	G1	E1	T1		116469 Unknown
10	P10	G1	E1	T1		128651 Unknown

#### 4.1.5.1 Tạo bảng reference từ clean\_data rename thành Performance

- Tạo bảng reference từ clean\_data rename thành Event giữ lại các "EventID", "Event", "Sport", remove duplicate cột EventID

Queries [8]

	A <sup>B</sup> <sub>C</sub> EventID	A <sup>B</sup> <sub>C</sub> Event	A <sup>B</sup> <sub>C</sub> Sport
1	E1	Basketball Men's Basketball	Basketball
2	E2	Judo Men's Extra-Lightweight	Judo
3	E3	Boxing Men's Middleweight	Boxing
4	E4	Wrestling Men's Lightweight, G...	Wrestling
5	E5	Swimming Women's 200 metre...	Swimming
6	E13	Hockey Men's Hockey	Hockey
7	E14	Judo Men's Middleweight	Judo
8	E16	Basketball Women's Basketball	Basketball
9	E17	Wrestling Men's Heavyweight, ...	Wrestling
10	E18	Archery Men's Individual	Archery
11	E19	Swimming Women's 200 metre...	Swimming
12	E20	Boxing Men's Light-Heavyweight	Boxing
13	E23	Rhythmic Gymnastics Women's	Rhythmic Gymnastics

#### 4.1.5.2 Tạo bảng reference từ clean\_data rename thành Event

- Tạo bảng reference từ clean\_data rename thành Games giữ lại các "GamesID", "Games", "Year", "Season", "City", remove duplicate cột GamesID

Queries [8]

	A <sup>B</sup> <sub>C</sub> GamesID	A <sup>B</sup> <sub>C</sub> Games	1.2 Year	A <sup>B</sup> <sub>C</sub> Season	A <sup>B</sup> <sub>C</sub> City
1	G1	1992 Summer		1992 Summer	Barcelona
2	G2	2008 Summer		2008 Summer	Beijing
3	G3	2012 Summer		2012 Summer	London
4	G4	2004 Summer		2004 Summer	Athina
5	G5	1948 Summer		1948 Summer	London
6	G6	2016 Summer		2016 Summer	Rio de Janeiro
7	G7	1936 Summer		1936 Summer	Berlin
8	G8	1988 Summer		1988 Summer	Seoul
9	G9	1996 Summer		1996 Summer	Atlanta
10	G10	2000 Summer		2000 Summer	Sydney
11	G11	1984 Summer		1984 Summer	Los Angeles
12	G12	1952 Summer		1952 Summer	Helsinki
13	G13	1976 Summer		1976 Summer	Montreal

#### 4.1.5.3 Tạo bảng reference từ clean\_data rename thành Games

- Tạo bảng reference từ clean\_data rename thành Team giữ lại các "TeamID", "Team", "NOC", "region", "notes", remove duplicate cột GamesID

Queries [8]

	A <sup>B</sup> TeamID	A <sup>B</sup> Team	A <sup>B</sup> NOC	A <sup>B</sup> region	A <sup>B</sup> notes
1	T1	China	CHN	China	
2	T7	United States	USA	USA	
3	T19	Spain	ESP	Spain	
4	T43	Germany	GER	Germany	
5	T47	Australia	AUS	Australia	
6	T73	Unified Team	EUN	Russia	
7	T78	Brazil	BRA	Brazil	
8	T107	Croatia	CRO	Croatia	
9	T114	Angola	ANG	Angola	
10	T115	Venezuela	VEN	Venezuela	
11	T120	Puerto Rico	PUR	Puerto Rico	
12	T152	Lithuania	LTU	Lithuania	
13	T6	Argentina	ARG	Argentina	
14	T11	Bulgaria	BUL	Bulgaria	
15	T18	Egypt	EGY	Egypt	
16	T21	France	FRA	France	
17	T94	Hungary	HUN	Hungary	

#### 4.1.5.4 Tạo bảng reference từ clean\_data rename thành Team

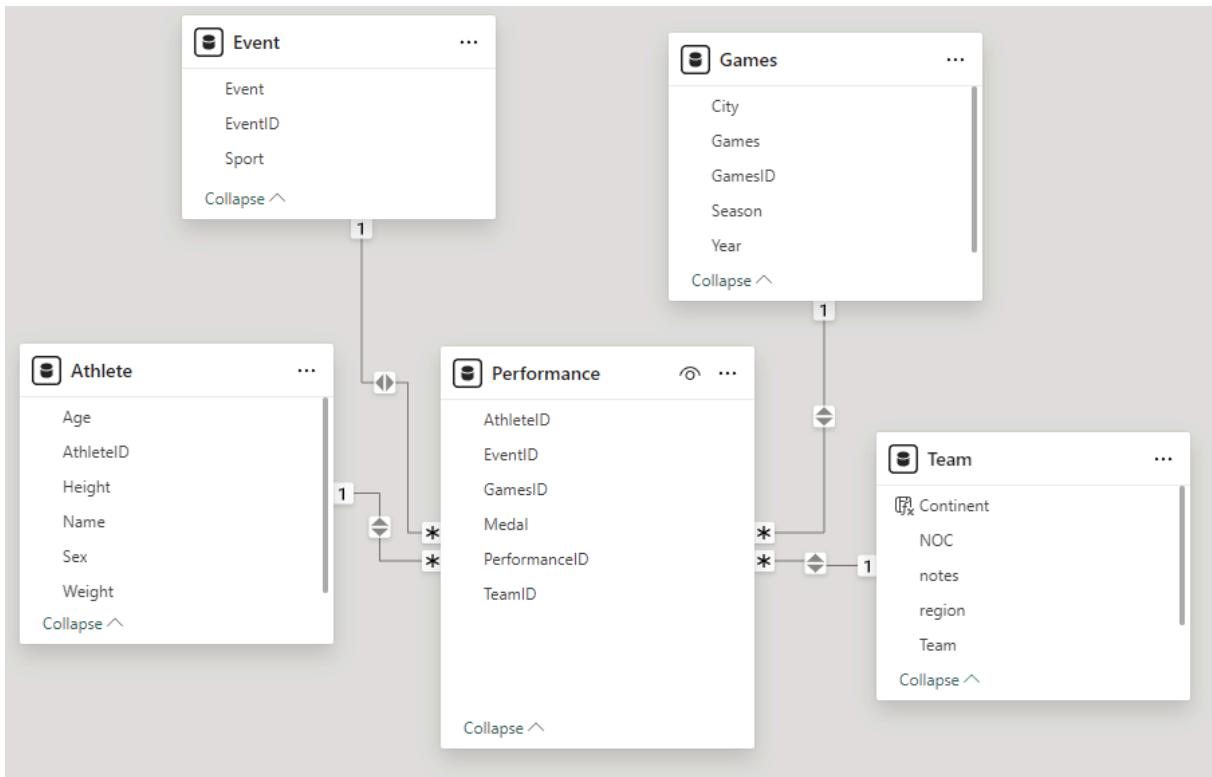
- Tạo bảng reference từ clean\_data rename thành Athlete giữ lại các "AthleteID", "Name", "Sex", "Age", "Height", "Weight", remove duplicate cột AthleteID

Queries [8]

	1.2 AthleteID	A <sup>B</sup> Name	A <sup>B</sup> Sex	1.2 Age	1.2 Height	1.2 Weight
1	1	A DJiang	M	24	180	80
2	41426	Gong Xiaobin	M	22	202	100
3	50576	Hu Weidong	M	21	198	87
4	69323	Li Chunjiang	M	29	190	85
5	72723	Ma Jian	M	22	200	90
6	105208	Shan Tao	M	22	215	120
7	113134	Song Ligang	M	25	200	95
8	116458	Sun Fengwu	M	30	190	85
9	116469	Sun Jun	M	22	197	100
10	128651	Wang Zhidan	M	21	214	110
11	131812	Wu Qinglong	M	27	190	87
12	134608	Zhang Yongjun	M	29	186	85
13	7901	Charles Wade Barkley	M	29	198	114
14	11668	Larry Joe Bird	M	35	205	100
15	30009	Clyde Austin Drexler	M	30	200	101
16	33553	Patrick Aloysius Ewing	M	29	213	109
17	55424	Earvin "Magic" Johnson, Jr.	M	32	205	100

#### 4.1.5.5 Tạo bảng reference từ clean\_data rename thành Athlete

- Chính các mối quan hệ của các bảng thành hai chiều



#### 4.1.5.6 Chính các mối quan hệ của các bảng thành hai chiều

## 4.2 Xử lý dữ liệu DAX

### 4.2.1 Measure

#### 4.2.1.1 Tỷ lệ phần trăm vận động viên

```

1 Percentage =
2 DIVIDE(
3     COUNTROWS('Athlete'),
4     CALCULATE(COUNTROWS('Athlete'), ALL('Athlete'))
5 )
    
```

##### 4.2.1.1.1 Tính tỉ lệ phần trăm vận động viên

#### 4.2.1.2 Tổng số lượng huy chương

```
Total Medals = SUM('Performance'[MedalValue])
```

4.2.1.2.1 Tổng số huy chương

#### 4.2.1.3 Tổng số lượng huy chương bạc

```
1 SilverCount =
2 CALCULATE(
3     COUNTROWS('Performance'),
4     'Performance'[Medal] = "Silver"
5 )
```

4.2.1.2.1 Tổng số lượng huy chương bạc

#### 4.2.1.4 Tổng số huy chương đồng

```
1 BronzeCount =
2 CALCULATE(
3     COUNTROWS('Performance'),
4     'Performance'[Medal] = "Bronze"
5 )
```

4.2.1.4.1 Tổng số huy chương đồng

#### 4.2.1.5 Tổng số huy chương vàng

```
1 GoldCount =
2 CALCULATE(
3     COUNTROWS('Performance'),
4     'Performance'[Medal] = "Gold"
5 )
```

4.2.1.5.1 Tổng số huy chương vàng

#### 4.2.1.6 Tổng số môn thể thao

```
Count Sport = DISTINCTCOUNT(Event[Sport])
```

4.2.1.6.1 Tổng số môn thể thao

#### 4.2.1.7 Tuổi nhỏ nhất, và lớn nhất

```
tuoimax = max('Athlete'[Age])
```

```
tuoimin = min('Athlete'[Age])
```

4.2.1.7.1 Tuổi nhỏ và lớn nhất

#### 4.2.1.8 Tổng số huy chương đồng

```
1 BronzeCount =
2 CALCULATE(
3     COUNTROWS('Performance'),
4     'Performance'[Medal] = "Bronze"
5 )
```

4.2.1.8.1 Tổng số huy chương đồng

#### 4.2.1.9 Tổng số huy chương vàng

```
1 GoldCount =
2 CALCULATE(
3     COUNTROWS('Performance'),
4     'Performance'[Medal] = "Gold"
5 )
```

4.2.1.9.1 Tổng số huy chương vàng

#### 4.2.1.10 Tổng số môn thể thao

```
Count Sport = DISTINCTCOUNT(Event[Sport])
```

4.2.1.10.1 Tổng số môn thể thao

#### 4.2.1.11 Tuổi nhỏ nhất, và lớn nhất

```
tuoimax = max('Athlete'[Age])
```

```
tuoimin = min('Athlete'[Age])
```

4.2.1.11.1 Tuổi nhỏ và lớn nhất

#### 4.2.1.12 Tổng số số nước tham gia

```
countcountry = distinctcount(Team[NOC])
```

4.2.1.12.1 Tổng số môn thể thao và số nước tham gia

#### 4.2.1.13 Chiều cao và cân nặng trung bình của vận viên nam

```
avgheightNam = CALCULATE(AVERAGE(Athlete[Height]), 'Athlete'[Sex] = "M")
```

```
avgweightNam = CALCULATE(AVERAGE(Athlete[Weight]), 'Athlete'[Sex] = "M")
```

4.2.1.13.1 Chiều cao và cân nặng trung bình của vận viên nam

#### 4.2.2 Chiều cao và cân nặng trung bình của vận động viên nữ

```
avgheightNu = CALCULATE(AVERAGE(Athlete[Height]), 'Athlete'[Sex] = "F")
```

```
avgweightNu = CALCULATE(AVERAGE(Athlete[Weight]), 'Athlete'[Sex] = "F")
```

4.2.2.1 Chiều cao và cân nặng trung bình của vận động viên nữ

#### 4.2.2.1 Tổng số huy chương ở các kỳ thế vận hội mùa đông

```
HUY_CHUONG_MUA_DONG = CALCULATE(COUNT(Performance[Medal]), 'Games'[Season] = "Winter")
```

4.2.2.1.1 Tổng số huy chương ở các kỳ thế vận hội mùa đông

#### 4.2.2.2 Tổng số huy chương ở các kỳ thế vận hội mùa hè

```
HUY_CHUONG_MUA_HE = CALCULATE(COUNT(Performance[Medal]), 'Games'[Season] = "SUMMER")
```

4.2.2.2.1 Tổng số huy chương ở các kỳ thế vận hội mùa hè

#### 4.2.2.3 Tổng số vận động viên

```
1 Tổng số Vận Động Viên = COUNTROWS(Athlete)
```

4.2.2.3.1 Tổng số vận động viên

#### 4.2.2.4 Số vận động viên nữ

```
1 Số vận động viên nữ = COUNTROWS(FILTER('Athlete', 'Athlete'[Sex] = "F"))
```

4.2.2.4.1 Số vận động viên nữ

#### 4.2.2.5 Số vận động viên nam

```
1 Số vận động viên nam = COUNTROWS(FILTER('Athlete', 'Athlete'[Sex] = "M"))
```

4.2.2.5.1 Số vận động viên nam

#### 4.2.2.6 Tuổi trung bình của các vận động viên

```
avg_age = AVERAGE(Athlete[Age])
```

4.2.2.6.1 Tuổi trung bình của các vận động viên

#### 4.2.2.7 Tuổi trung bình của vận động viên nữ

```
Tuổi trung bình của VĐV nữ = AVERAGEX(FILTER(Athlete, Athlete[Sex] = "F"), Athlete[Age])
```

4.2.2.7.1 Tuổi trung bình của vận động viên nữ

#### 4.2.2.8 Tuổi trung bình của vận động viên nam

```
Tuổi trung bình của VĐV nam = AVERAGEX(FILTER(Athlete, Athlete[Sex] = "M"), Athlete[Age])
```

4.2.2.8.1 Tuổi trung bình của vận động viên nam

#### 4.2.2.9 Tỉ lệ vận động viên nữ

Tỉ lệ vận động viên nữ = `DIVIDE([Số vận động viên nữ], [Tổng số Vận Động Viên], 0) * 100`

##### 4.2.2.9.1 Tỉ lệ vận động viên nữ

#### 4.2.3 Tỉ lệ vận động viên nam

Tỉ lệ vận động viên nam = `100 - 'Measure'[Số vận động viên nữ]`

##### 4.2.3.1 Tỉ lệ vận động viên nam

#### 4.2.3.1 Số vận động viên trong từng sự kiện

số vận động viên trong từng sự kiện = `DISTINCTCOUNT(Performance[AthleteID])`

##### 4.2.3.1.1 Số vận động viên trong từng sự kiện

#### 4.2.3.2 Tổng số đội tuyển theo quốc gia

Tổng số đội tuyển theo quốc gia = `DISTINCTCOUNT(Team[TeamID])`

##### 4.2.3.2.1 Tổng số đội tuyển theo quốc gia

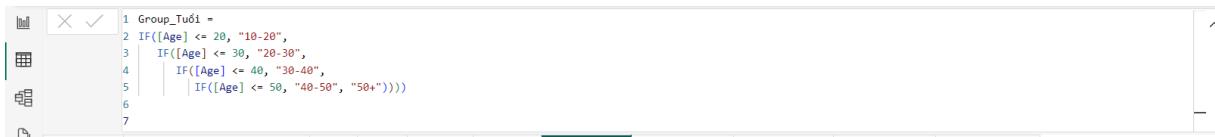
#### 4.2.3.3 Số lượng sự kiện mà mỗi đội tuyển đã tham gia

Số lượng sự kiện mà mỗi đội tuyển đã tham gia = `DISTINCTCOUNT(Performance[EventID])`

##### 4.2.3.3.1 Số lượng sự kiện mà mỗi đội tuyển đã tham gia

#### 4.2.4 Calculated column

##### 4.2.4.1 Tạo cột nhóm tuổi



		1 Group_Tuoi =
		2 IF([Age] <= 20, "10-20",
		3     IF([Age] <= 30, "20-30",
		4         IF([Age] <= 40, "30-40",
		5             IF([Age] <= 50, "40-50", "50+"))))
		6
		7

AthleteID	Name	Sex	Age	Height	Weight	Group_Tuoi	BMI	BMICategory	Height Group	Weight Group
A50364	Hsu Chaohsung	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A70271	Liu Yunchang	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A127873	Wang Yutseng	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A132495	Yu Chinghiao	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A99791	Kamal Riad Noseir	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A127213	Abdel Monein Wahibi Hassanein	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A90006	Sergio Paganella	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A44380	Jak Habib	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A13488	Vctor Hugo Borja Morca	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A40953	Emil Ging	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A78423	Irving "Toots" Meretsky	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A23974	Jacinto Cira Cruz	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A74892	Franco Marquicias	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A75381	Primitivo Martnez	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A75590	Jess Marzan	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A87289	Amador O. Obordo	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A89532	Bibiano Ouano	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A95335	Jean Pollet	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A26477	Cammino de Pilla	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A70880	Miguel Pedro Martinez Lopes	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A58049	Maksis Kazks	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A35649	Antonio Narciso Flecha Alvarez	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg

##### 4.2.4.1.1 Tạo cột nhóm tuổi

#### 4.2.4.2 Tạo cột continent

```

1 Continent =
2 SWITCH(
3     TRUE(),
4     'Team'[Region] IN {"Antigua", "Argentina", "Aruba", "Bahamas", "Barbados", "Belize", "Bermuda", "Bolivia", "Brazil", "Canada", "Cayman Islands", "Chile",
5     "Colombia", "Costa Rica", "Cuba", "Curacao", "Dominica", "Dominican Republic", "Ecuador", "El Salvador", "Grenada", "Guatemala", "Guyana", "Haiti", "Honduras",
6     "Jamaica", "Mexico", "Nicaragua", "Panama", "Paraguay", "Peru", "Puerto Rico", "Saint Kitts", "Saint Lucia", "Saint Vincent", "Suriname", "Trinidad", "USA",
7     "Uruguay", "Venezuela", "Virgin Islands", "British", "Virgin Islands, US"), "America",
8     'Team'[Region] IN {"Albania", "Andorra", "Austria", "Belarus", "Belgium", "Bosnia and Herzegovina", "Bulgaria", "Croatia", "Cyprus", "Czech Republic", "Denmark",
9     "Estonia", "Finland", "France", "Georgia", "Germany", "Greece", "Hungary", "Iceland", "Ireland", "Italy", "Kosovo", "Latvia", "Liechtenstein", "Lithuania",
10    "Luxembourg", "Macedonia", "Malta", "Moldova", "Monaco", "Montenegro", "Netherlands", "Norway", "Poland", "Portugal", "Romania", "Russia", "San Marino", "Serbia",
11    "Slovakia", "Slovenia", "Spain", "Sweden", "Switzerland", "Ukraine", "UK"), "Europe",
12    'Team'[Region] IN {"Afghanistan", "Armenia", "Azerbaijan", "Bahrain", "Bangladesh", "Bhutan", "Brunei", "Cambodia", "China", "India", "Indonesia", "Iran", "Iraq",
13    "Israel", "Japan", "Jordan", "Kazakhstan", "Kuwait", "Kyrgyzstan", "Lao", "Lebanon", "Malaysia", "Maldives", "Mongolia", "Myanmar", "Nepal", "North Korea",
14    "Oman", "Pakistan", "Palestine", "Philippines", "Qatar", "Saudi Arabia", "Singapore", "South Korea", "Sri Lanka", "Syria", "Taiwan", "Tajikistan", "Thailand",
15    "Timor-Leste", "Turkey", "Turkmenistan", "United Arab Emirates", "Uzbekistan", "Vietnam", "Yemen"), "Asia",
16    'Team'[Region] IN {"Algeria", "Angola", "Benin", "Botswana", "Burkina Faso", "Burundi", "Cape Verde", "Central African Republic", "Chad", "Comoros", "Democratic
17    Republic of the Congo", "Djibouti", "Egypt", "Equatorial Guinea", "Eritrea", "Eswatini", "Ethiopia", "Gabon", "Gambia", "Ghana", "Guinea", "Guinea-Bissau", "Ivory
18    Coast", "Kenya", "Lesotho", "Liberia", "Libya", "Madagascar", "Malawi", "Mauritania", "Mauritius", "Morocco", "Mozambique", "Namibia", "Niger", "Nigeria",
19    "Republic of Congo", "Rwanda", "Sao Tome and Principe", "Senegal", "Seychelles", "Sierra Leone", "Somalia", "South Africa", "South Sudan", "Sudan", "Tanzania",
20    "Togo", "Tunisia", "Uganda", "Zambia", "Zimbabwe"), "Africa",
21    'Team'[Region] IN {"American Samoa", "Australia", "Fiji", "Guam", "Kiribati", "Marshall Islands", "Micronesia", "Nauru", "New Zealand", "Papua New Guinea",
22    "Samoa", "Solomon Islands", "Tonga", "Tuvalu", "Vanuatu"), "Oceania",
23    "Other"
24 )

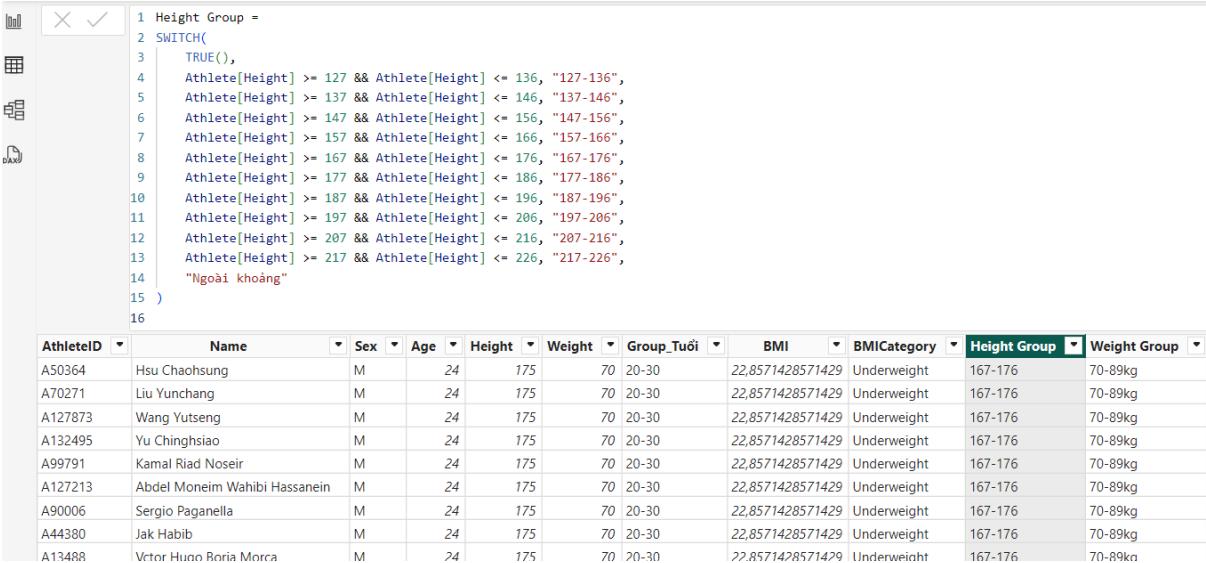
```

##### 4.2.4.2.1 Tạo cột continent

TeamID	Team	NOC	Region	Continent
T10	France	FRA	France	Europe
T427	Socit Nautique de Bayonne-2	FRA	France	Europe
T731	Socit Nautique de la Basse Seine-1	FRA	France	Europe
T280	Club Nautique de Dieppe-5	FRA	France	Europe
T440	Club Nautique de Franais-1	FRA	France	Europe
T499	Cercle de l'Aviron Roubaix-4	FRA	France	Europe
T643	Societ Nautique de la Marne-3	FRA	France	Europe
T1022	Club Nautique de Lyon-2	FRA	France	Europe
T222	Socit Nautique de Bayonne	FRA	France	Europe

##### 4.2.4.2.2 Cột Continent

#### 4.2.4.3 Tạo cột nhóm chiều cao



The screenshot shows the Microsoft Access query builder interface. On the left, there's a tree view of tables and queries. In the center, a query definition window displays the following SQL code:

```

1 Height Group =
2 SWITCH(
3   TRUE(),
4   Athlete[Height] >= 127 && Athlete[Height] <= 136, "127-136",
5   Athlete[Height] >= 137 && Athlete[Height] <= 146, "137-146",
6   Athlete[Height] >= 147 && Athlete[Height] <= 156, "147-156",
7   Athlete[Height] >= 157 && Athlete[Height] <= 166, "157-166",
8   Athlete[Height] >= 167 && Athlete[Height] <= 176, "167-176",
9   Athlete[Height] >= 177 && Athlete[Height] <= 186, "177-186",
10  Athlete[Height] >= 187 && Athlete[Height] <= 196, "187-196",
11  Athlete[Height] >= 197 && Athlete[Height] <= 206, "197-206",
12  Athlete[Height] >= 207 && Athlete[Height] <= 216, "207-216",
13  Athlete[Height] >= 217 && Athlete[Height] <= 226, "217-226",
14  "Ngoài khoảng"
15 )
16

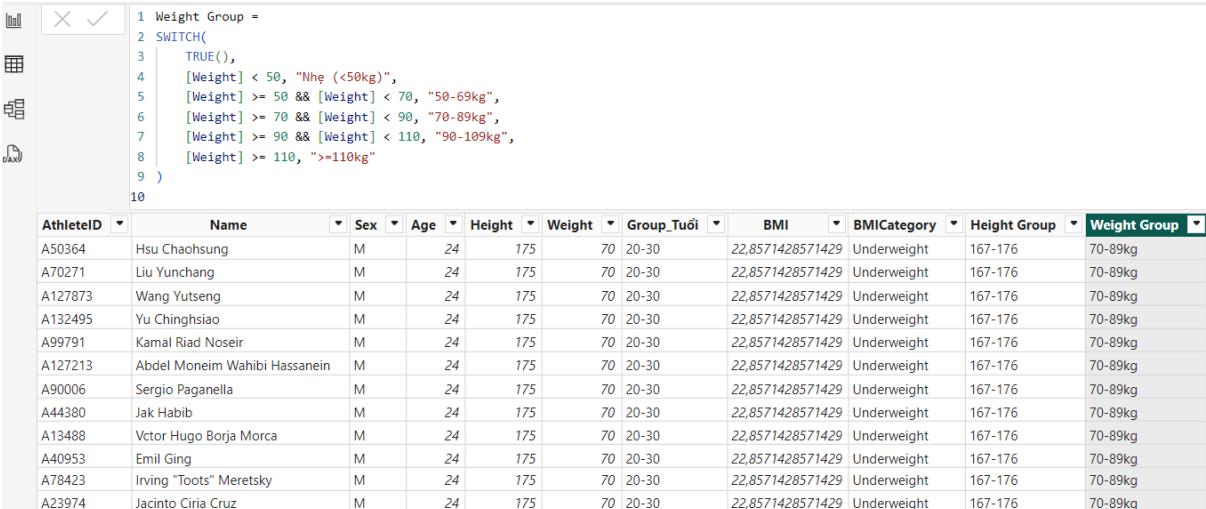
```

Below the code is a table view showing athlete data with additional columns for Height Group and Weight Group.

AthleteID	Name	Sex	Age	Height	Weight	Group_Tuổi	BMI	BMICategory	Height Group	Weight Group
A50364	Hsu Chaohsung	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A70271	Liu Yunchang	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A127873	Wang Yutseng	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A132495	Yu Chinghsiao	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A99791	Kamal Riad Noseir	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A127213	Abdel Moneim Wahib Hassanain	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A90006	Sergio Paganella	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A44380	Jak Habib	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A13488	Vctor Hugo Borja Morca	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg

##### 4.2.4.3.1 Tạo cột nhóm chiều cao

#### 4.2.4.4 Tạo cột nhóm cân nặng



The screenshot shows the Microsoft Access query builder interface. On the left, there's a tree view of tables and queries. In the center, a query definition window displays the following SQL code:

```

1 Weight Group =
2 SWITCH(
3   TRUE(),
4   [Weight] < 50, "Nhẹ (<50kg)",
5   [Weight] >= 50 && [Weight] < 70, "50-69kg",
6   [Weight] >= 70 && [Weight] < 90, "70-89kg",
7   [Weight] >= 90 && [Weight] < 110, "90-109kg",
8   [Weight] >= 110, ">=110kg"
9 )
10

```

Below the code is a table view showing athlete data with additional columns for Height Group and Weight Group.

AthleteID	Name	Sex	Age	Height	Weight	Group_Tuổi	BMI	BMICategory	Height Group	Weight Group
A50364	Hsu Chaohsung	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A70271	Liu Yunchang	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A127873	Wang Yutseng	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A132495	Yu Chinghsiao	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A99791	Kamal Riad Noseir	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A127213	Abdel Moneim Wahib Hassanain	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A90006	Sergio Paganella	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A44380	Jak Habib	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A13488	Vctor Hugo Borja Morca	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A40953	Emil Ging	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A78423	Irving "Toots" Meretsky	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg
A23974	Jacinto Ciria Cruz	M	24	175	70	20-30	22,8571428571429	Underweight	167-176	70-89kg

##### 4.2.4.4.1 Tạo cột nhóm cân nặng

#### 4.2.4.5 Tạo cột BMI



```

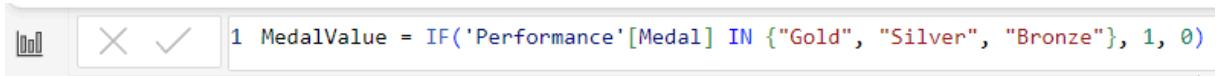
1 BMI =
2 'Athlete'[Weight] / ( 'Athlete'[Height] / 100 * 'Athlete'[Height] / 100 )
3

```

AthleteID	Name	Sex	Age	Height	Weight	Group_Tuổi	BMI
A50364	Hsu Chaohsung	M	24	175	70	20-30	22,8571428571429
A70271	Liu Yunchang	M	24	175	70	20-30	22,8571428571429
A127873	Wang Yutseng	M	24	175	70	20-30	22,8571428571429
A132495	Yu Chinghsiao	M	24	175	70	20-30	22,8571428571429
A99791	Kamal Riad Noseir	M	24	175	70	20-30	22,8571428571429
A127213	Abdel Moneim Wahibi Hassanein	M	24	175	70	20-30	22,8571428571429
A90006	Sergio Paganella	M	24	175	70	20-30	22,8571428571429
A44380	Jak Habib	M	24	175	70	20-30	22,8571428571429
A13488	Vctor Hugo Borja Morca	M	24	175	70	20-30	22,8571428571429
A40953	Emil Ging	M	24	175	70	20-30	22,8571428571429

##### 4.2.4.5.1 Tạo cột BMI

#### 4.2.4.6 Tạo cột Medal value



```

1 MedalValue = IF('Performance'[Medal] IN {"Gold", "Silver", "Bronze"}, 1, 0)

```

PerformanceID	GamesID	EventID	TeamID	AthleteID	Medal	MedalValue
P209618	G24	E578	T7	3669	Unknown	0
P209619	G24	E578	T7	5497	Unknown	0
P209620	G24	E578	T7	8190	Unknown	0
P209621	G24	E578	T7	8413	Unknown	0
P209622	G24	E578	T7	10377	Unknown	0
P209623	G24	E578	T7	10396	Unknown	0
P209624	G24	E578	T7	10939	Unknown	0
P209625	G24	E578	T7	11132	Unknown	0
P209626	G24	E578	T7	11817	Unknown	0
P209627	G24	E578	T7	12660	Unknown	0

##### 4.2.4.6.1 Tạo cột Medal value

#### 4.2.4.7 Tạo cột Age Group

```

1 Age Group =
2 SWITCH(
3   TRUE(),
4   Athlete[Age]< 20, "Dưới 20",
5   Athlete[Age] >= 20 && Athlete[Age] <= 30, "20-30",
6   Athlete[Age] > 30 && Athlete[Age] <= 40, "30-40",
7   Athlete[Age] > 40, "Trên 40",
8   "Không xác định" // Trường hợp nếu không có dữ liệu tuổi
9 )
10

```

##### 4.2.4.7.1 Tạo cột Age Group

#### 4.2.4.8 Tạo cột Medal group

```
1 Medal group = if(Performance[Medal] in {"Gold", "Silver", "Bronze"}, "Have medal", "Don't have medal")
```

##### 4.2.4.8.1 Tạo cột Medal group

#### 4.2.5 Tạo bảng

##### 4.2.5.1 Tạo bảng Medal Count By Country



The screenshot shows the Power BI Data View interface. On the left, there's a DAX editor window with the following code:

```
1 MedalCountByCountry =
2 SUMMARIZE(
3     FILTER('Performance', 'Performance'[Medal] <> "Unknown"),
4     'Team'[NOC],
5     "MedalCount", COUNT('Performance'[Medal])
6 )
7 )
```

On the right, there's a preview table titled "MedalCount" with columns "NOC" and "MedalCount". The data is as follows:

NOC	MedalCount
CHN	989
USA	5632
ESP	489
GER	2163
AUS	1320
EUN	279
BRA	475
CRO	149
VEN	15
PUR	9
LTU	61

##### 4.2.5.1.1 Tạo bảng Medal Count By Country

#### 4.2.5.2 Tạo bảng Athlete Count By Country



The screenshot shows the Power BI Data View interface. On the left, there are three icons: a grid for measures, a cube for tables, and a document for data flows. The main area displays a DAX query and a resulting table.

```
1 AthleteCountByCountry =  
2 SUMMARIZE(  
3     'Team',  
4     'Team'[NOC],  
5     "AthleteCount", COUNT('Athlete'[Name])  
6 )  
7
```

The table has two columns: NOC and AthleteCount. The data is as follows:

NOC	AthleteCount
FRA	6186
USA	9656
GBR	6272
CHN	2664
ESP	2640
GER	4872
AUS	3820
EUN	604
BRA	2053
CRO	428

##### 4.2.5.2.1 Tạo bảng Athlete Count By Country

## 4.2.6 Filter

### 4.2.6.1 Tạo filter Year để giới hạn dữ liệu theo thời gian.

A screenshot of a 'Year' filter interface. It features two input fields at the top with the values '1896' and '2016'. Below these is a horizontal slider with circular endpoints, spanning from approximately 1896 to 2016.

4.2.6.1.1 Tạo filter Year

### 4.2.6.2 Tạo filter Medal để chọn loại huy chương : Bronze ( Đồng ),Silver ( Bạc ),Gold( Vàng)

A screenshot of a 'Medal' filter interface. It shows a dropdown menu with the option 'All' selected, indicated by a small downward arrow icon.

4.2.6.2.1 Tạo filter Medal để chọn loại huy chương

### 4.2.6.3 Tạo filter Season để chọn mùa: Summer, Winter.

A screenshot of a 'Season' filter interface. It shows a dropdown menu with the option 'All' selected.

4.2.6.3.1 Tạo filter Season

### 4.2.6.4 Tạo filter Sex để chọn giới tính: M( Nam ),F( Nữ )

A screenshot of a 'Sex' filter interface. It shows a dropdown menu with the option 'All' selected.

4.2.6.4.1 Tạo filter Sex

### 4.2.6.5 Tạo filter Sport để chọn các môn thể thao.

A screenshot of a 'Sport' filter interface. It shows a dropdown menu with the option 'All' selected.

4.2.6.5.1 Tạo filter Sport

## 5 Trực quan hóa dữ liệu

### 5.1 Các kỹ thuật trực quan hóa

#### - Biểu đồ

- + **Biểu đồ cột:** So sánh dữ liệu giữa các danh mục.
- + **Biểu đồ đường:** Hiển thị sự thay đổi của dữ liệu theo thời gian.
- + **Biểu đồ tròn:** Hiển thị tỷ lệ phần trăm của các danh mục trong một tổng thể.
- + **Biểu đồ phân tán:** Hiển thị mối quan hệ giữa hai biến số.
- + **Biểu đồ nhiệt độ:** Hiển thị dữ liệu số bằng màu sắc.

#### - Bản đồ

- + **Bản đồ địa lý:** Hiển thị dữ liệu trên bản đồ địa lý.
- + **Bản đồ nhiệt:** Sử dụng màu sắc để thể hiện mật độ dữ liệu trên bản đồ.
- + **Bản đồ cây:** Hiển thị mối quan hệ giữa các dữ liệu theo cấu trúc cây.

### 5.2 Các nguyên tắc trực quan hóa

#### - Hiểu rõ mục tiêu trực quan hóa

- + **Xác định đối tượng người xem:** Biết rõ người xem là ai (nhà quản lý, nhà phân tích, khách hàng, v.v.) để chọn cách trình bày phù hợp.
- + **Đặt câu hỏi:** Biểu đồ này nhằm mục đích gì? Để so sánh, hiển thị xu hướng, hay trình bày sự phân phối?

#### - Chọn loại biểu đồ phù hợp

- + **Không dùng sai biểu đồ:**
  - Biểu đồ cột để so sánh.
  - Biểu đồ đường để thể hiện xu hướng.
  - Biểu đồ tròn chỉ dành cho các tỷ lệ tổng cộng là 100%.
- + **Tránh quá tải:** Không nên dùng quá nhiều loại biểu đồ trên cùng một giao diện

#### - Đơn giản và dễ hiểu

- + **Hạn chế thông tin thừa:** Tránh các yếu tố không cần thiết như màu sắc quá nhiều, hiệu ứng 3D gây nhiễu.
- + **Giảm thiểu chữ viết tắt:** Các nhãn và tiêu đề nên rõ ràng và dễ hiểu.
- + **Tránh phức tạp:** Một biểu đồ đơn giản nhưng rõ ràng luôn tốt hơn một biểu đồ phức tạp.

#### - Sắp xếp và làm nổi bật thông tin quan trọng

- + **Tập trung vào nội dung chính:** Làm nổi bật các dữ liệu quan trọng bằng màu sắc, kích thước hoặc vị trí.
- + **Sắp xếp logic:** Dữ liệu nên được sắp xếp theo thứ tự có ý nghĩa (ví dụ: giảm dần, tăng dần, hoặc theo thời gian).

#### - Sử dụng màu sắc hợp lý

- + **Màu sắc nhất quán:** Dùng cùng màu để biểu diễn cùng loại dữ liệu trong nhiều biểu đồ.
- + **Tránh lạm dụng:** Không dùng quá nhiều màu, và tránh các màu khó phân biệt (ví dụ: xanh lá và đỏ với người mù màu).
- + **Dùng màu để nhấn mạnh:** Sử dụng màu sắc sáng hoặc tương phản để làm nổi bật thông tin quan trọng.

#### - Cung cấp ngũ cảnh

- + **Tiêu đề rõ ràng:** Mỗi biểu đồ cần có tiêu đề cụ thể để người xem hiểu rõ nội dung.
- + **Chú thích (Legend):** Đảm bảo có chú thích nếu sử dụng nhiều màu hoặc ký hiệu.
- + **Trục được gắn nhãn:** Luôn gắn nhãn rõ ràng cho các trục (X, Y) và đảm bảo tỷ lệ chính xác.

#### - Duy trì tỷ lệ đúng

- + **Không làm méo dữ liệu:** Đảm bảo tỷ lệ giữa các trục phù hợp, tránh gây hiểu lầm.
- + **Bắt đầu từ 0 (khi cần thiết):** Với biểu đồ cột, trục Y nên bắt đầu từ 0 để tránh làm sai lệch nhận thức về sự khác biệt.

### - Kiểm tra và đánh giá

- + **Kiểm tra trước khi trình bày:** Xem biểu đồ có dễ hiểu không, có phù hợp với mục tiêu không.
- + **Nhận phản hồi:** Thu thập ý kiến từ người xem để cải thiện.

### - Tôn trọng tính trung thực của dữ liệu

- + **Không bóp méo dữ liệu:** Trực quan hóa phải phản ánh đúng bản chất của dữ liệu, không cố ý gây hiểu lầm.
- + **Tránh loại bỏ thông tin quan trọng:** Đảm bảo biểu đồ không bị đơn giản hóa quá mức.

## 5.3 Trình bày cách thêm visual mới

### 5.3.1 Tạo visual thông kê chi tiết

### 5.3.2 Phân tích theo quốc gia và mùa thi đấu

#### 5.3.2.1 Tạo visual filter lọc theo sport



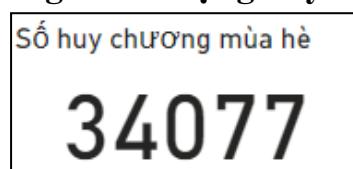
5.3.2.1.1 Tạo visual filter lọc theo Sport

#### 5.3.2.2 Tạo visual filter lọc theo season



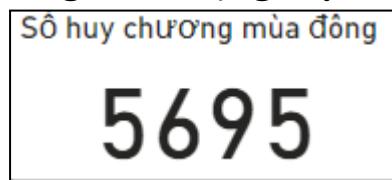
5.3.2.2.1 Tạo visual filter lọc theo Season

#### 5.3.2.3 Tạo visual card thống kê số lượng huy chương hè



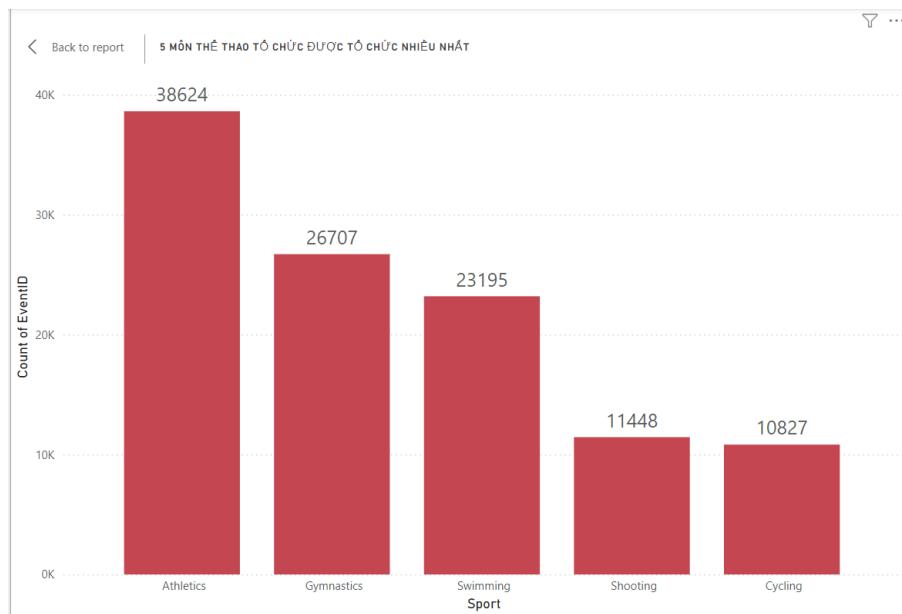
5.3.2.3.1 Tạo visual card thống kê số lượng huy chương hè

### 5.3.2.4 Tạo visual card thống kê số lượng huy chương mùa đông



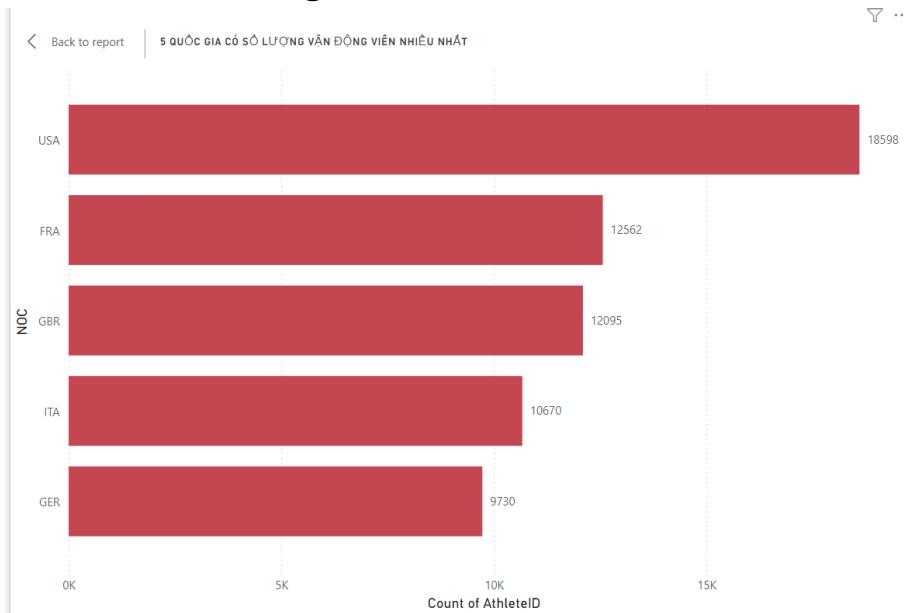
#### 5.3.2.4.1 Tạo visual card thống kê số lượng huy chương mùa đông

### 5.3.2.5 Tạo visual stacked column thống kê 5 môn thể thao tổ chức nhiều nhất



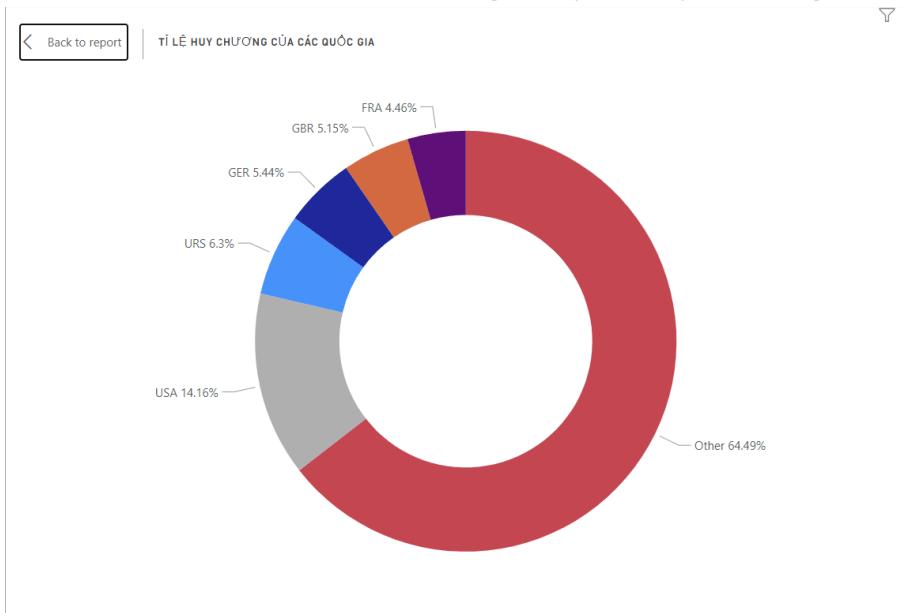
#### 5.3.2.5.1 Tạo visual stacked column thống kê 5 môn thể thao tổ chức nhiều nhất

### 5.3.2.6 Tạo visual stacked bar thống kê 5 quốc gia có số lượng vận động viên tham gia nhiều nhất



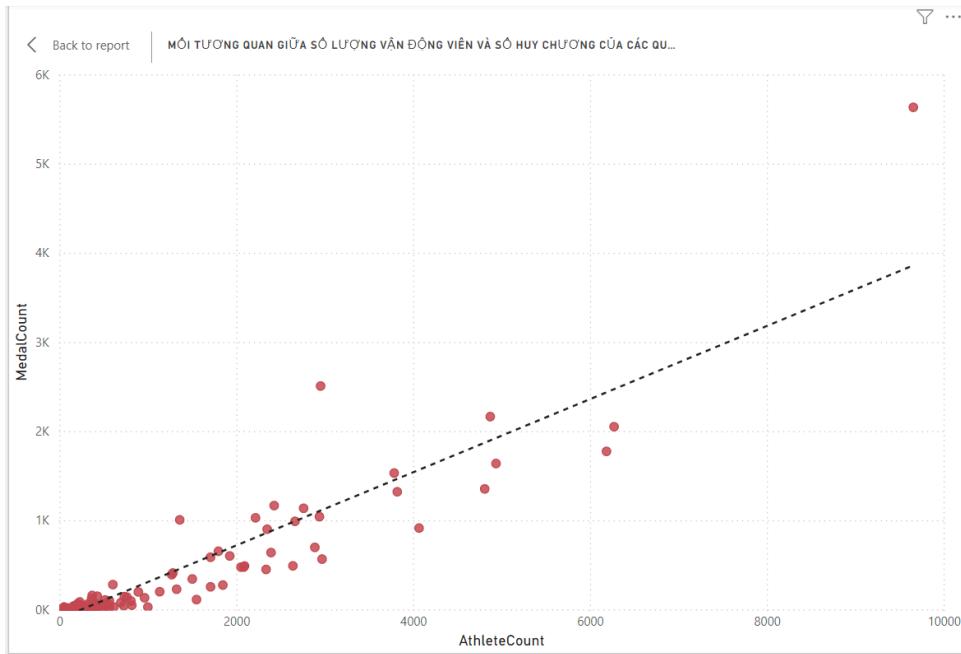
#### 5.3.2.6.1 Tạo visual stacked bar thống kê 5 quốc gia có số lượng vận động viên tham gia nhiều nhất

### 5.3.2.7 Tạo visual donut thống kê tỷ lệ huy chương của các quốc gia



#### 5.3.2.7.1 Tạo visual donut thống kê tỷ lệ huy chương của các quốc gia

### 5.3.2.8 Tạo visual scatter mối tương quan giữa số lượng vận động viên và số huy chương của các quốc gia



5.3.2.8.1 Tạo visual scatter mối tương quan giữa số lượng vận động viên và số huy chương của các quốc gia

### 5.3.2.9 Tạo visual treemap số lượng huy chương của các quốc gia



#### 5.3.2.9.1 Tao visual treemap số lượng huy chương của các quốc gia

### 5.3.3 Phân tích theo giới tính

#### 5.3.3.1 Tạo visual filter lọc theo Medal



5.3.3.1.1 Tạo visual filter lọc theo Medal

#### 5.3.3.2 Tạo visual filter lọc theo Season



5.3.3.2.1 Tạo visual filter lọc theo Season

#### 5.3.3.3 Tạo visual filter lọc theo Sex



5.3.3.3.1 Tạo visual filter lọc theo Sex

#### 5.3.3.4 Tạo visual filter lọc theo Year



5.3.3.4.1 Tạo visual filter lọc theo Year

### 5.3.3.5 Tạo visual card thống kê số vận động viên nữ



#### 5.3.3.5.1 Tạo visual card thống kê số vận động viên nữ

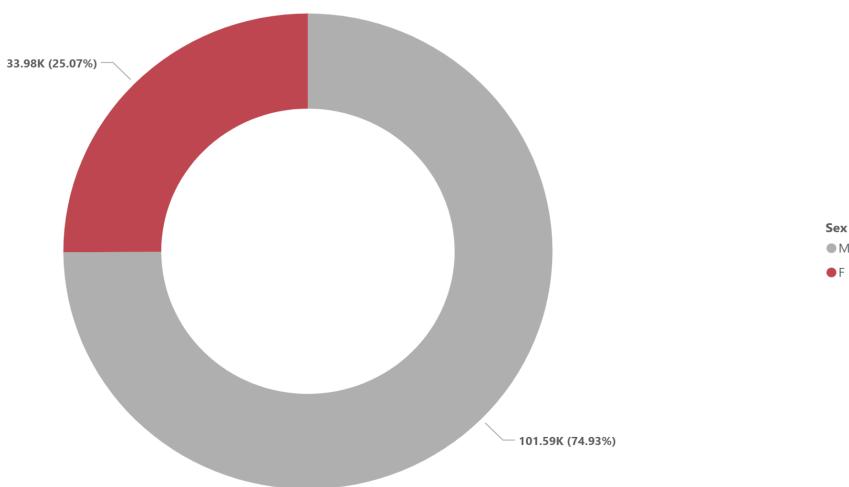
### 5.3.3.6 Tạo visual card thống kê số vận động viên nam



#### 5.3.3.6.1 Tạo visual card thống kê số vận động viên nam

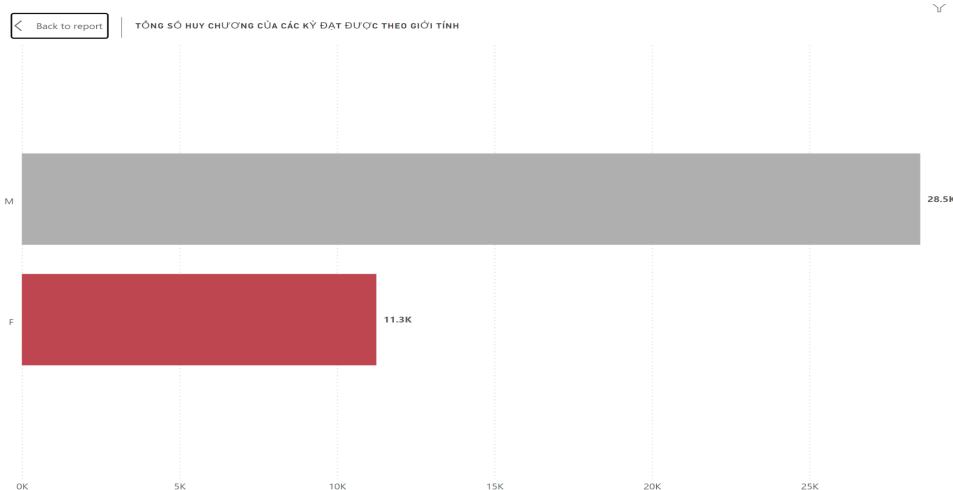
### 5.3.3.7 Tạo visual Pie Chart tỷ lệ giới tính của vận động viên

TỶ LỆ GIỚI TÍNH CỦA VẬN ĐỘNG VIÊN



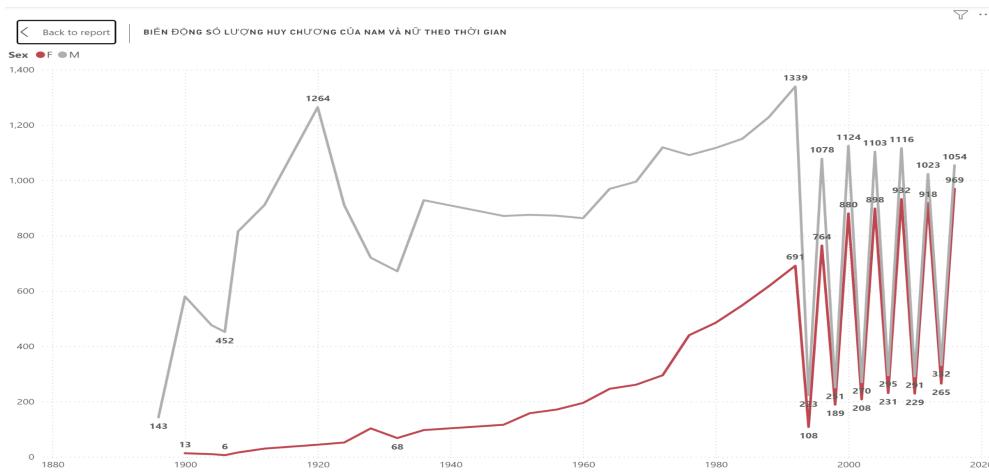
#### 5.3.3.7.1 Tạo visual Pie Chart tỷ lệ giới tính của vận động viên

### 5.3.3.8 Tạo visual Stacked Bar Chart tổng số huy chương đạt được theo giới tính



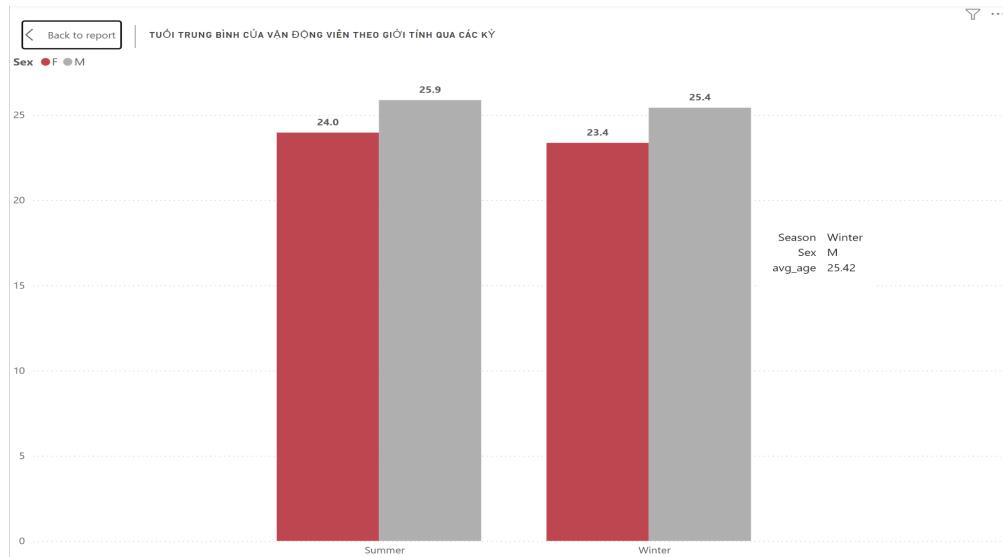
#### 5.3.3.8.1 Tạo visual Stacked Bar Chart tổng số huy chương đạt được theo giới tính

### 5.3.3.9 Tạo visual Line Chart Biến động số lượng huy chương của nam và nữ theo thời gian



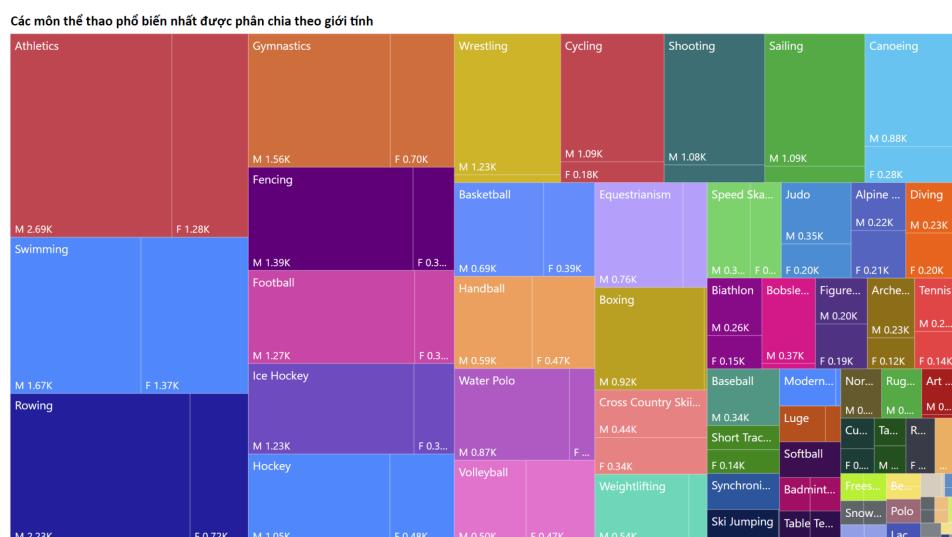
#### 5.3.3.9.1 Tạo visual Line Chart Biến động số lượng huy chương của nam và nữ theo thời gian

### 5.3.3.9.1 Tạo visual Clustered Column Chart tuổi trung bình của vận động viên qua các kỳ thi vận hội theo giới tính



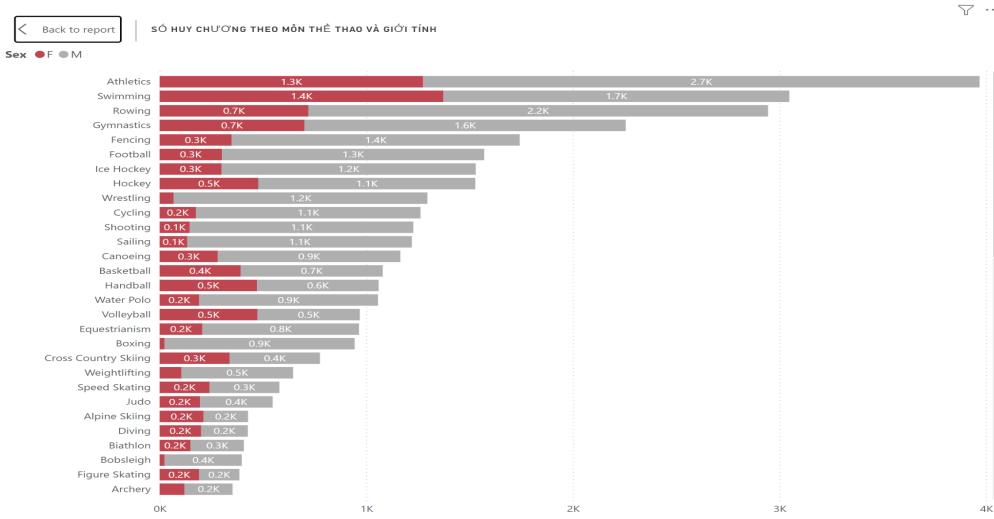
#### 5.3.3.9.1.1 Tạo visual Clustered Column Chart tuổi trung bình của vận động viên qua các kỳ thi vận hội theo giới tính

### 5.3.3.9.2 Tạo visual Bar Chart các môn thể thao phổ biến nhất được phân chia theo giới tính



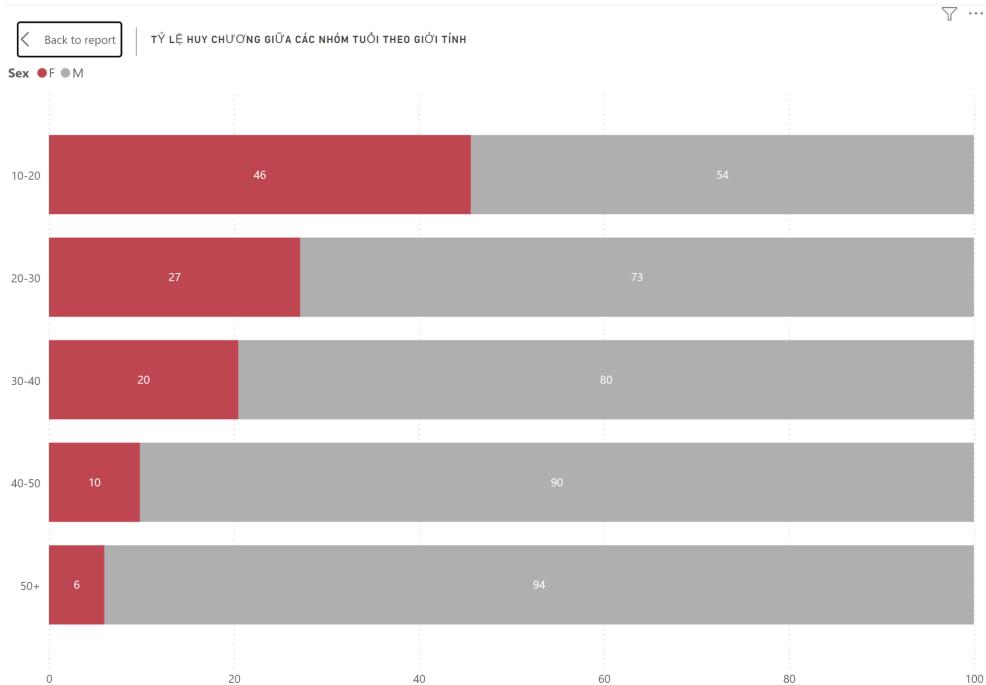
#### 5.3.3.9.2.1 Tạo visual Bar Chart các môn thể thao phổ biến nhất được phân chia theo giới tính

### 5.3.3.9.3 Tạo visual Stacked Bar Chart số huy chương theo môn thể thao và giới tính



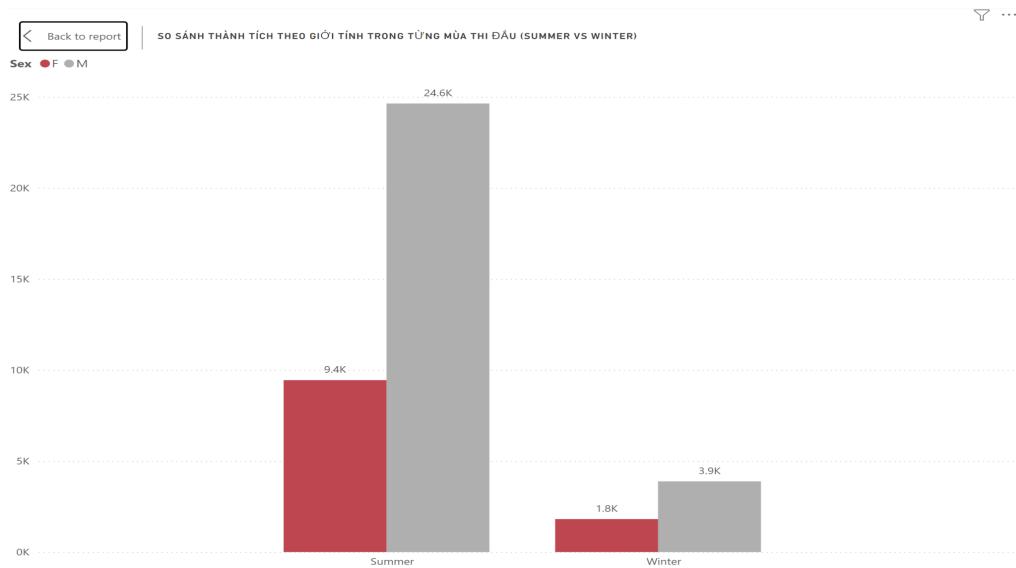
### 5.3.3.9.3.1 Tạo visual Stacked Bar Chart số huy chương theo môn thể thao và giới tính

### 5.3.3.9.4 Tạo visual Stacked Bar Chart tỷ lệ huy chương giữa các nhóm tuổi theo giới tính



### 5.3.3.9.4.1 Tạo visual Stacked Bar Chart tỷ lệ huy chương giữa các nhóm tuổi theo giới tính

### 5.3.3.9.5 Tạo visual Clustered Column Chart so sánh thành tích theo giới tính trong từng mùa thi đấu (Summer vs Winter)



### 5.3.3.9.5.1 Tạo visual Clustered Column Chart so sánh thành tích theo giới tính trong từng mùa thi đấu (Summer vs Winter)

## 5.3.4 Phân tích theo chiều cao và cân nặng

### 5.3.4.1 Chiều cao trung bình của vận động viên nữ

Chiều cao trung bình  
**169.60**  
 Nữ

5.3.4.1.1 Chiều cao trung bình của vận động viên nữ

### 5.3.4.2 Cân nặng trung bình của vận động viên nữ

Cân nặng trung bình  
**62.34**  
 Nữ

5.3.4.2.1 Cân nặng trung bình của vận động viên nữ

### 5.3.4.3 Chiều cao trung bình của vận động viên nam

Chiều cao trung bình
<b>178.12</b>
Nam

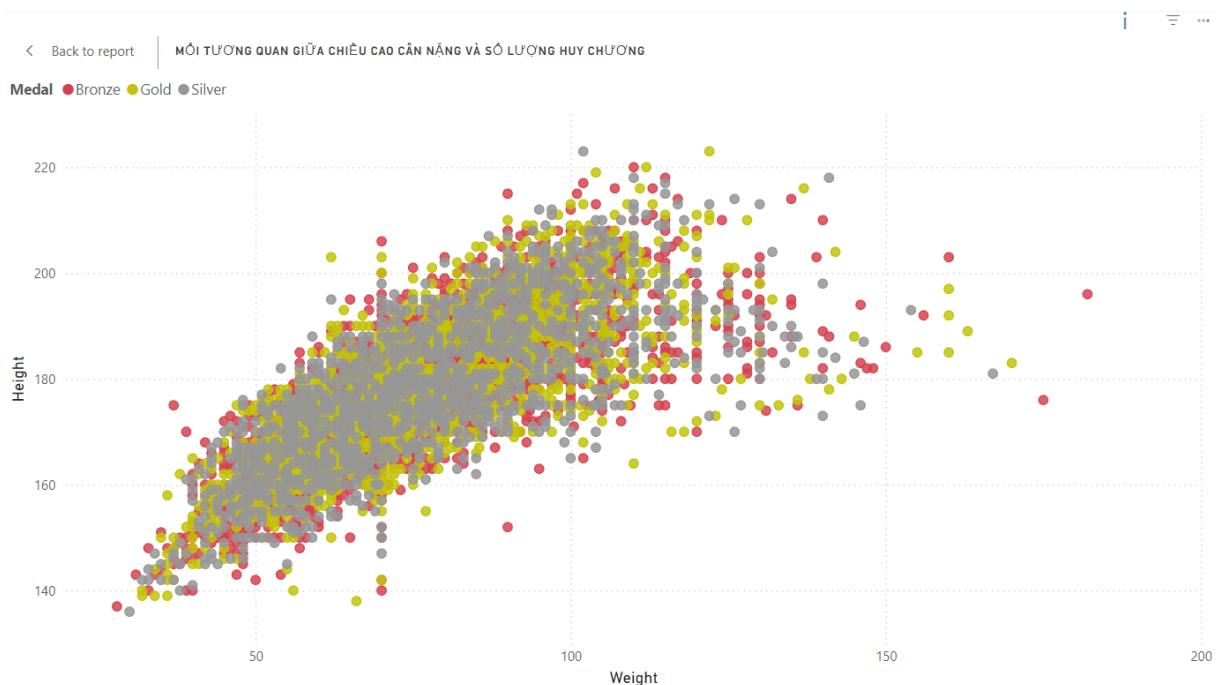
#### 5.3.4.3.1 Chiều cao trung bình của vận động viên nam

### 5.3.4.4 Cân nặng trung bình của vận động viên nam

Cân nặng trung bình
<b>74.51</b>
Nam

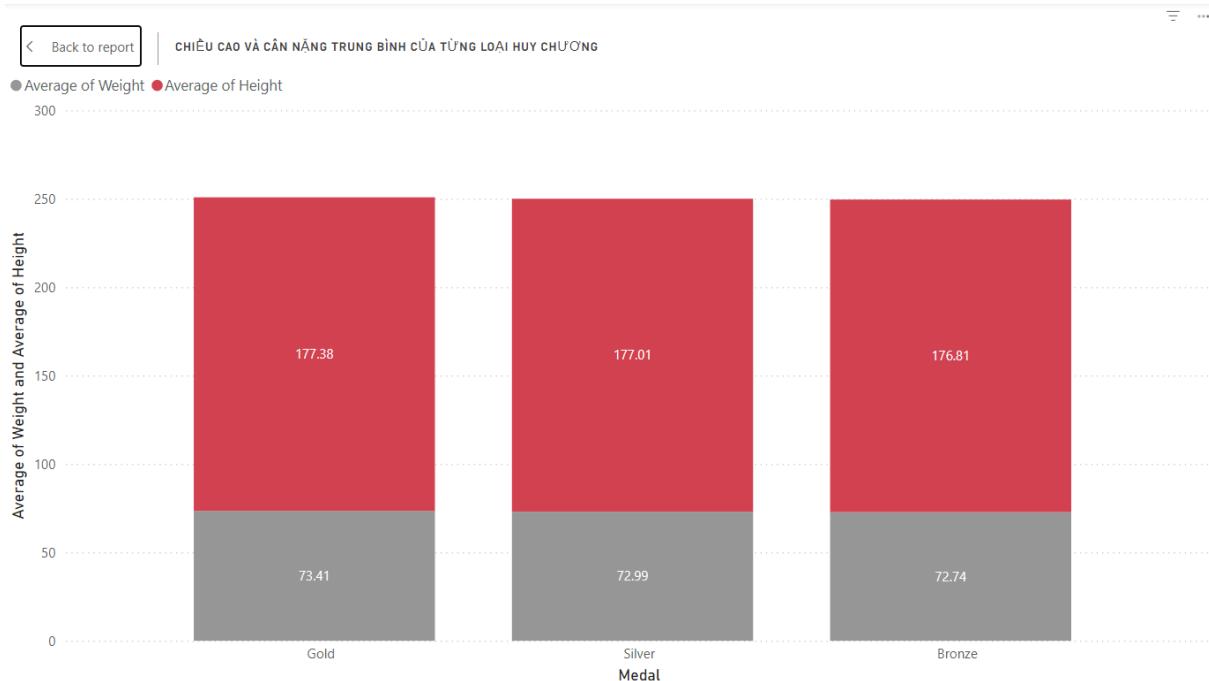
#### 5.3.4.4.1 Cân nặng trung bình của vận động viên nam

### 5.3.4.5 Mối tương quan giữa chiều cao và cân nặng và số huy chương



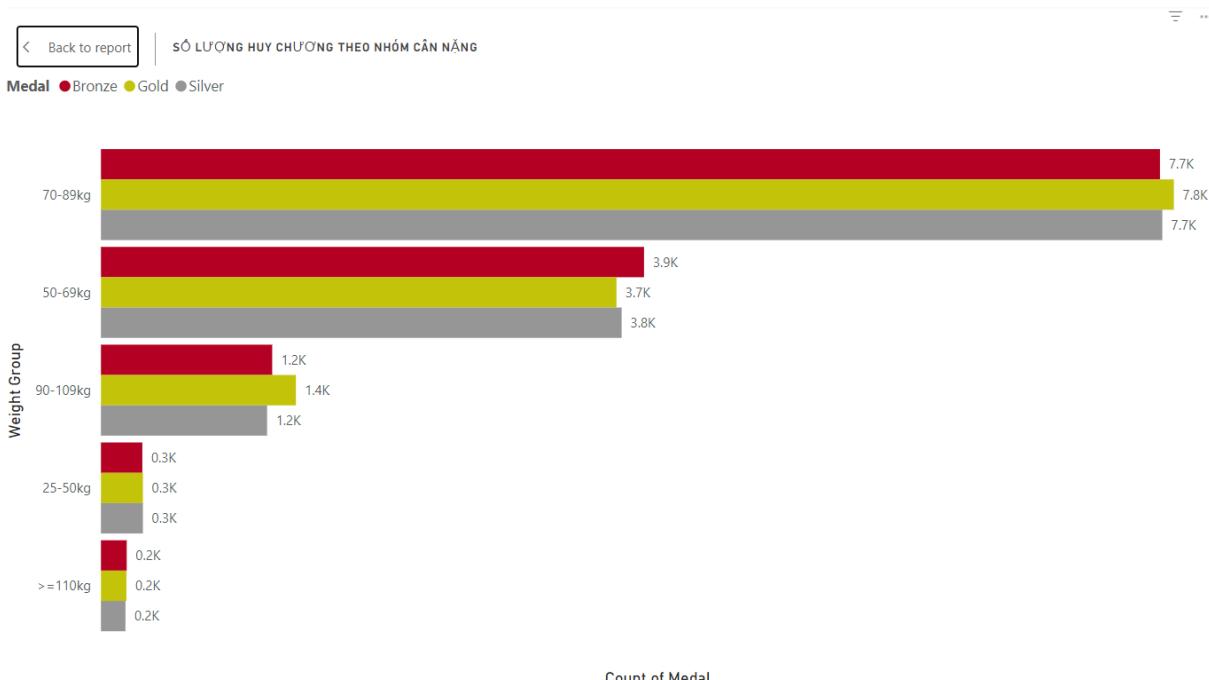
#### 5.3.4.5.1 Mối tương quan giữa chiều cao và cân nặng và số huy chương

### 5.3.4.6 So sánh chiều cao và cân nặng của từng loại huy chương



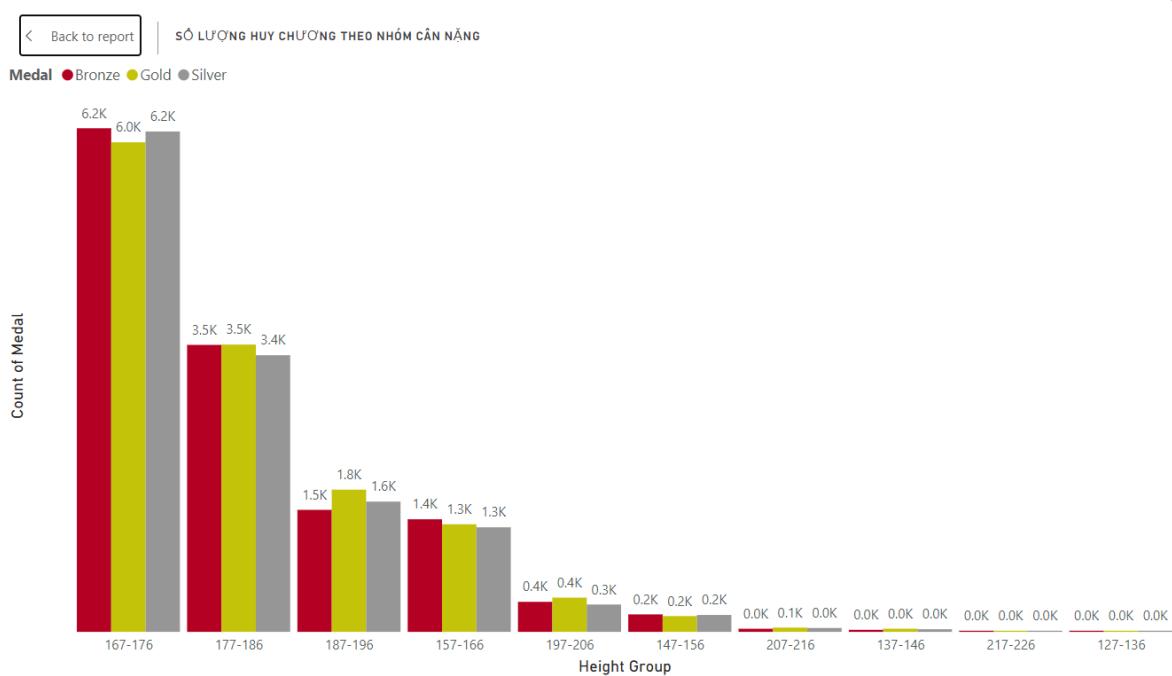
#### 5.3.4.6.1 Mối tương quan giữa chiều cao và cân nặng và số huy chương

#### 5.3.4.7 Số lượng huy chương theo nhóm cân nặng



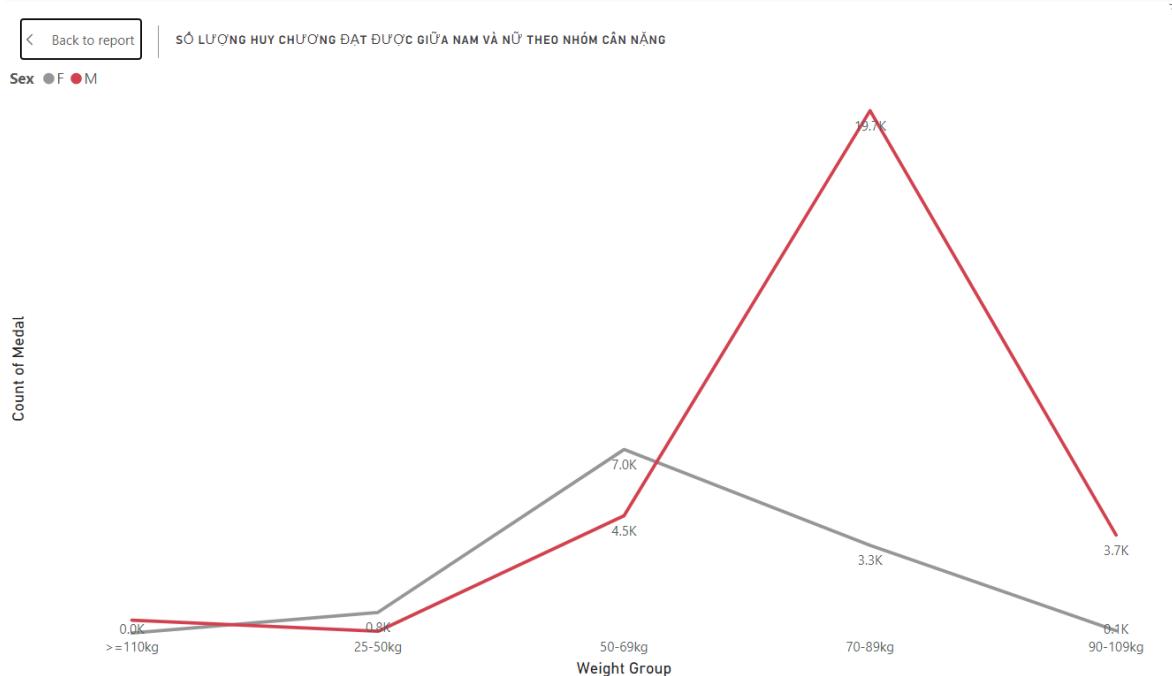
#### 5.3.4.7.1 Chiều cao trung bình của vận động viên nữ

### 5.3.4.8 Số lượng huy chương theo nhóm chiều cao



#### 5.3.4.8.1 Số lượng huy chương theo nhóm chiều cao

### 5.3.4.9 So sánh số lượng huy chương giữa nam và nữ theo nhóm cân nặng



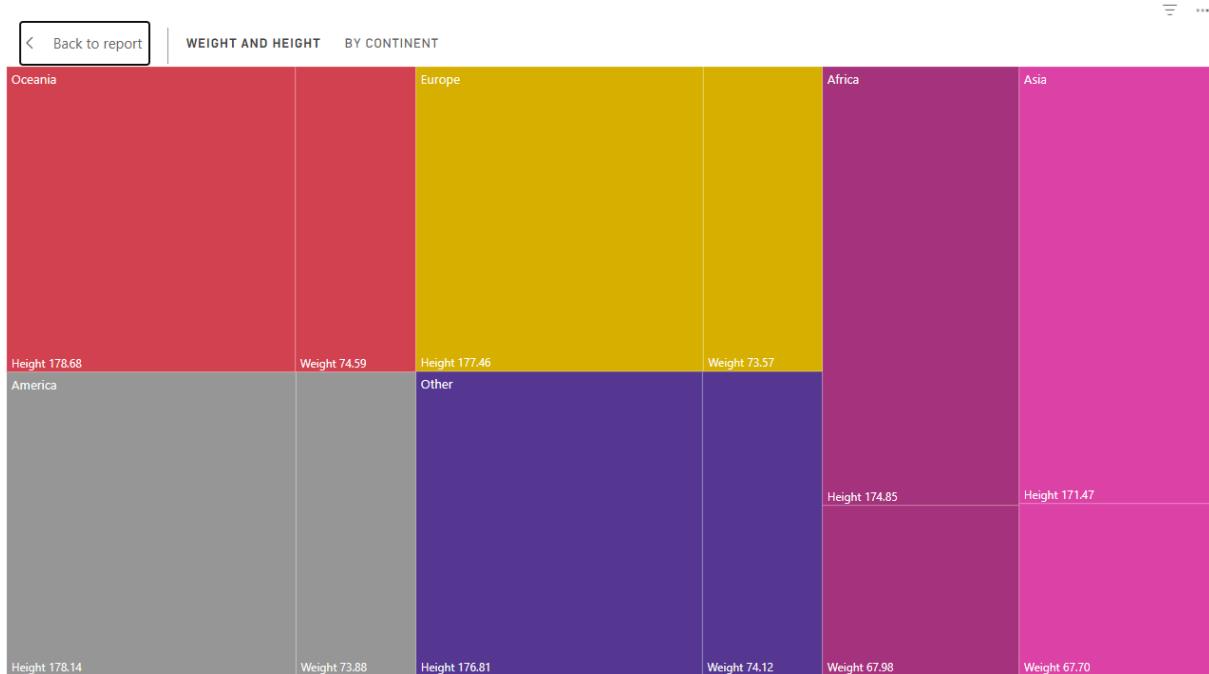
#### 5.3.4.9.1 So sánh số lượng huy chương giữa nam và nữ theo nhóm cân nặng

### 5.3.5 So sánh số lượng huy chương giữa nam và nữ theo nhóm cân nặng

Height Group	F	M
127-136	5	
137-146	88	7
147-156	451	166
157-166	2760	1243
167-176	5508	12906
177-186	2060	8431
187-196	351	4512
197-206	28	1101
207-216	2	135
217-226		18

5.3.4.10.1 So sánh số lượng huy chương giữa nam và nữ theo nhóm chiều cao

### 5.3.6 Chiều cao và cân nặng theo khu vực



#### 5.3.4.11.1 Chiều cao và cân nặng theo khu vực

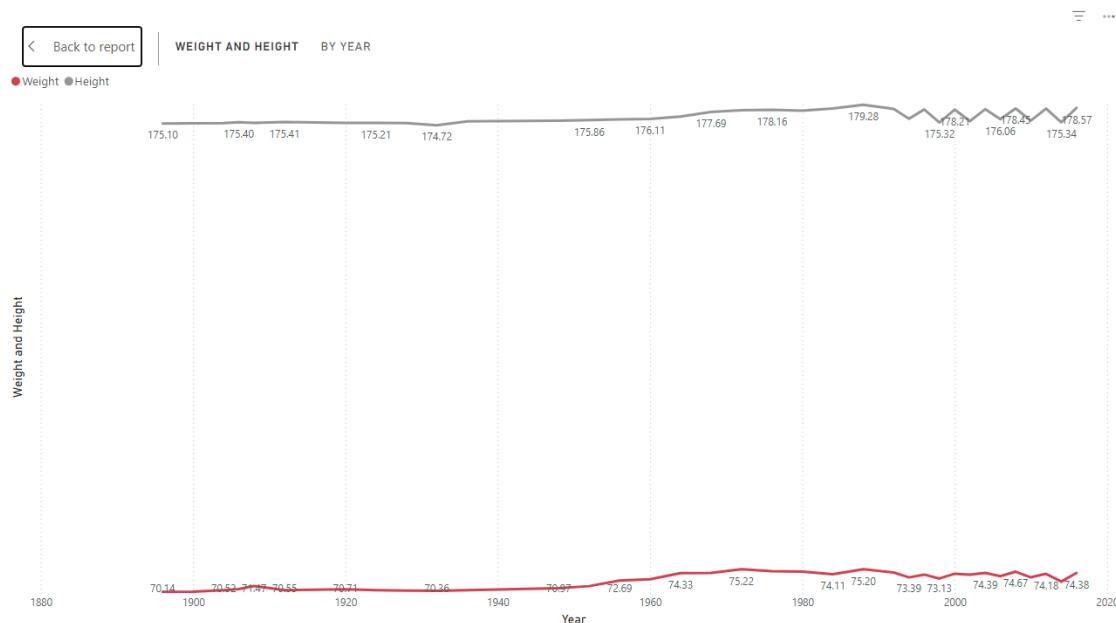
### 5.3.7 Thống kê chiều cao và cân nặng theo môn thể thao và thành tích

[Back to report](#)

Sport	Height	Weight	Count	Medal
Baseball	182.02	84.39		894
Basketball	187.67	81.66		4536
Bobsleigh	179.13	81.56		3058
Handball	182.13	79.73		3665
Rugby Sevens	175.36	78.94		299
Water Polo	181.05	78.68		3846
Ice Hockey	178.04	78.29		5516
Beach Volleyball	185.52	78.21		564
Volleyball	186.06	78.08		3404
Weightlifting	169.68	77.68		3937
Tug-Of-War	176.17	77.15		170
Judo	173.94	76.75		3801
Rowing	181.12	76.56		10595

#### 5.3.4.12.1 Thống kê chiều cao và cân nặng theo môn thể thao và thành tích

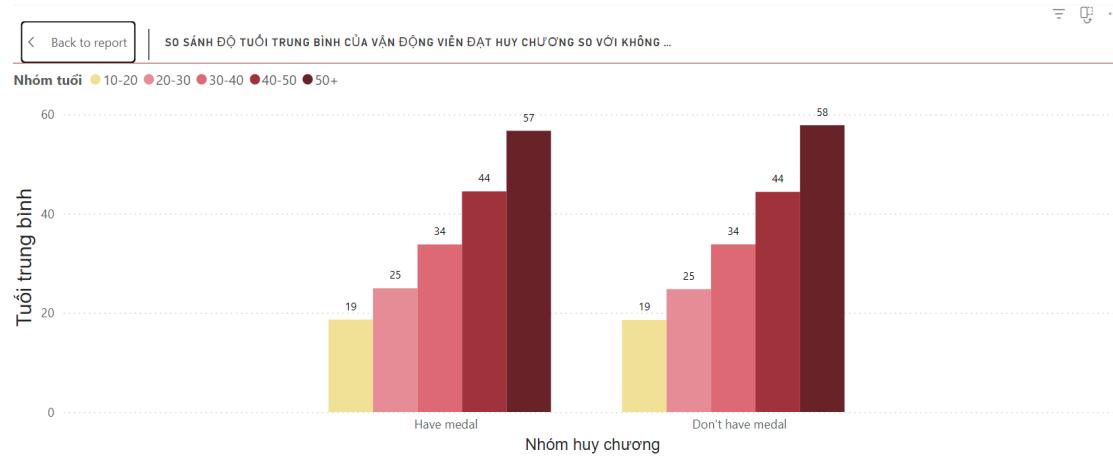
### 5.3.7.1 Xu hướng chiều cao và cân nặng theo từng năm



#### 5.3.4.13.1 Xu hướng chiều cao và cân nặng theo từng năm

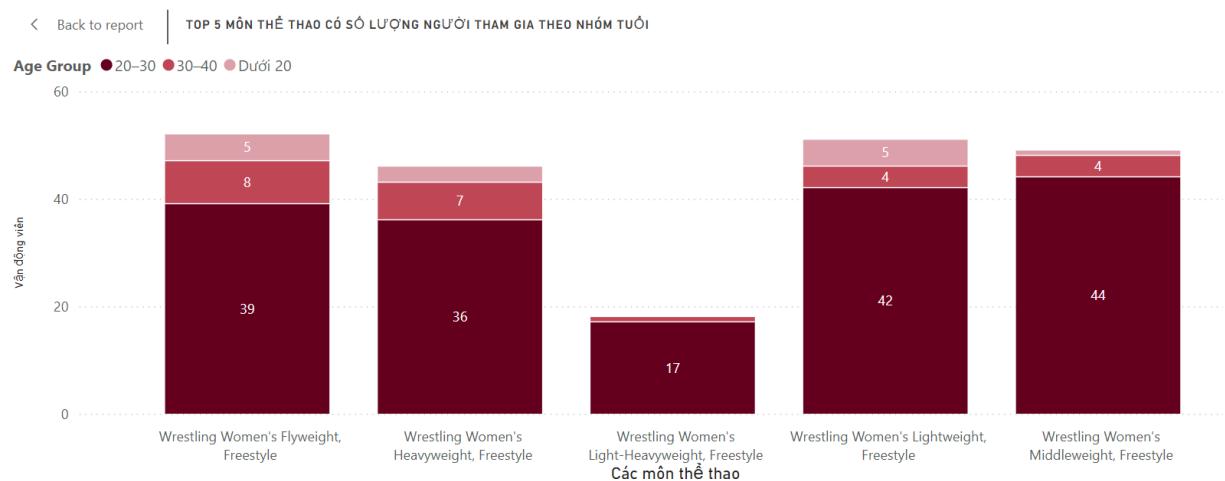
### 5.3.8 Phân tích theo tuổi

#### 5.3.5.1 Tạo visual clustered column so sánh độ tuổi trung bình của vận động viên đạt huy chương so với không đạt



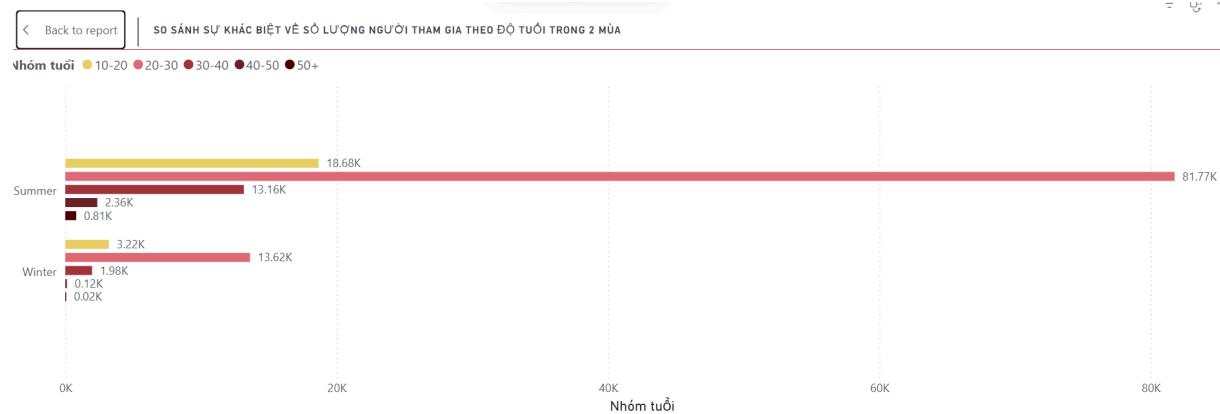
#### 5.3.5.1.1 Tạo visual clustered column so sánh độ tuổi trung bình của vận động viên đạt huy chương so với không đạt

#### 5.3.5.2 Tạo visual stacked column top 5 môn thể thao có số lượng người tham gia theo nhóm tuổi



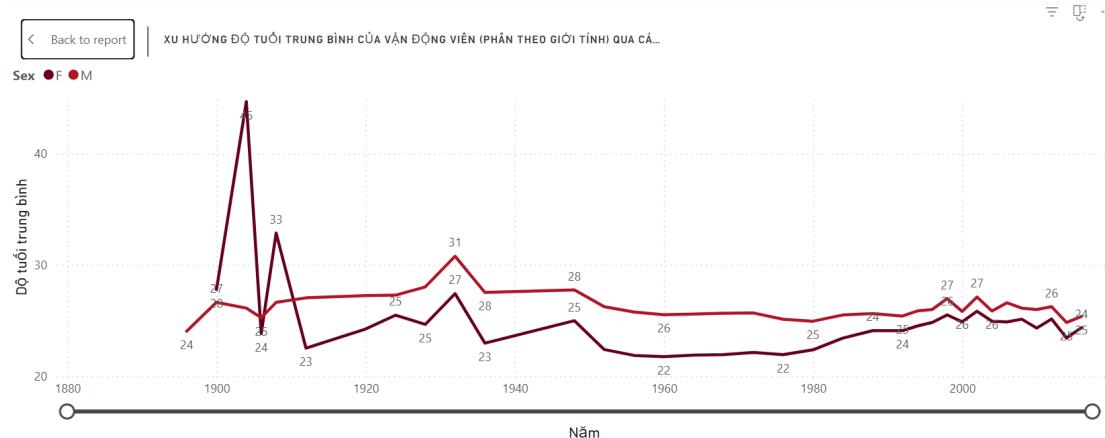
#### 5.3.5.2.1 Tạo visual stacked column top 5 môn thể thao có số lượng người tham gia theo nhóm tuổi

### 5.3.5.3 Tạo visual clustered bar so sánh sự khác biệt về số lượng người tham gia theo độ tuổi trong 2 mùa



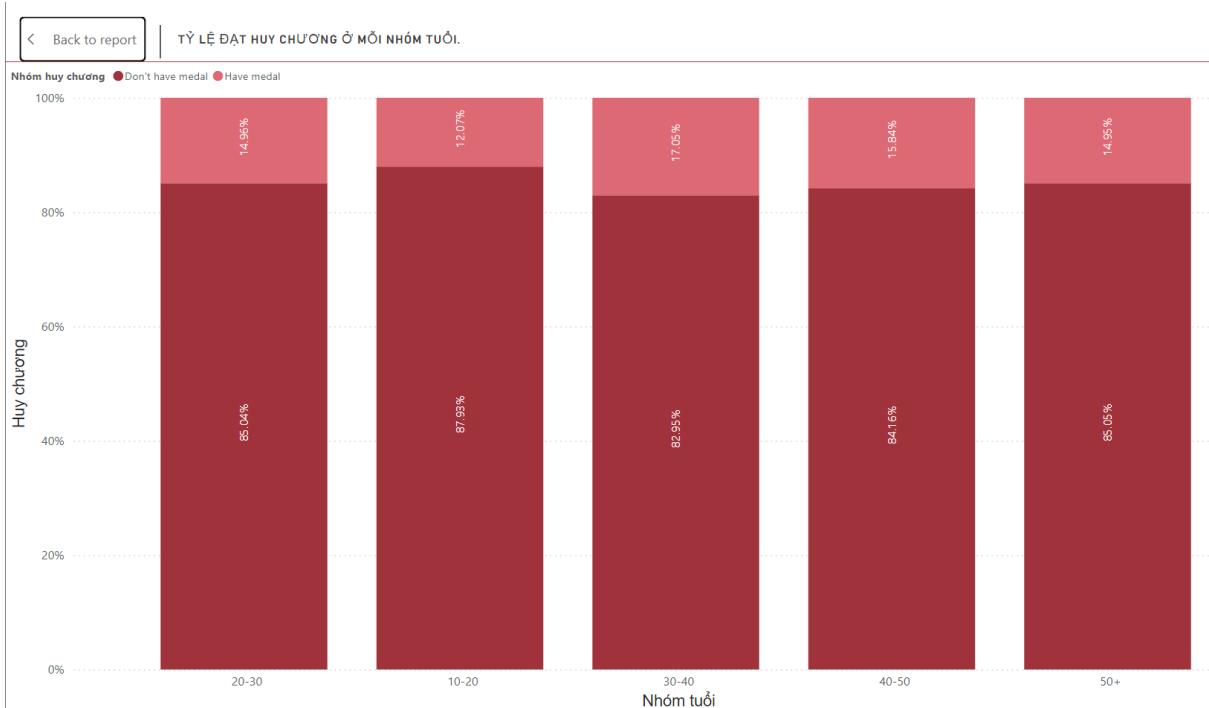
#### 5.3.5.3.1 Tạo visual clustered bar so sánh sự khác biệt về số lượng người tham gia theo độ tuổi trong 2 mùa

### 5.3.5.4 Tạo visual line xu hướng độ tuổi trung bình của vận động viên (phân theo giới tính) qua các năm



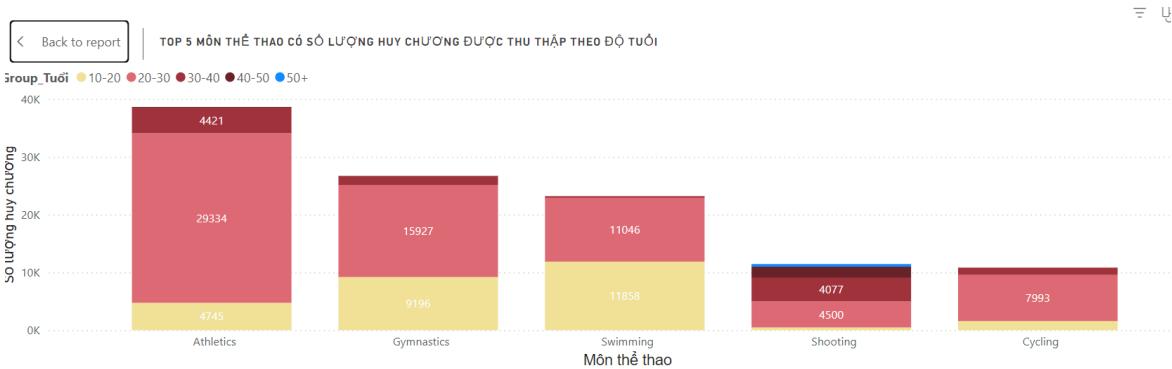
#### 5.3.5.4.1 Tạo visual line xu hướng độ tuổi trung bình của vận động viên (phân theo giới tính) qua các năm

### 5.3.5.5 Tạo visual 100% stacked column tỉ lệ đạt huy chương ở mỗi nhóm tuổi



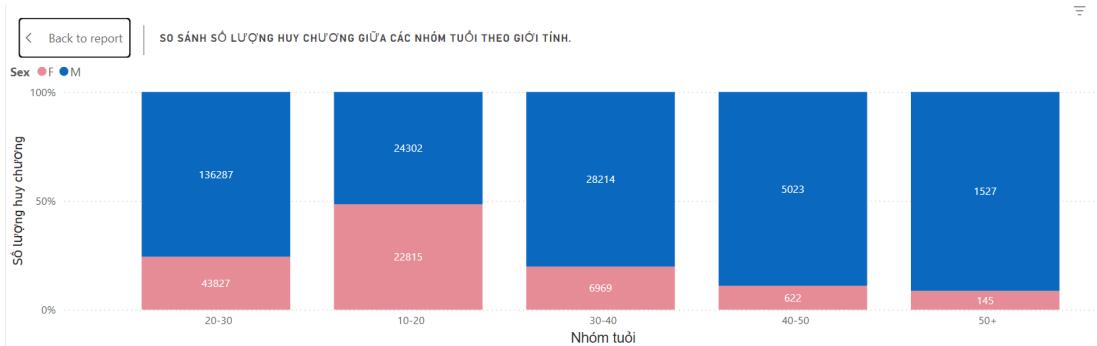
#### 5.3.5.5.1 ảnh Tạo visual 100% stacked column tỉ lệ đạt huy chương ở mỗi nhóm tuổi

### 5.3.5.6 Tạo visual stacked column top 5 môn thể thao có số lượng huy chương được thu thập theo độ tuổi



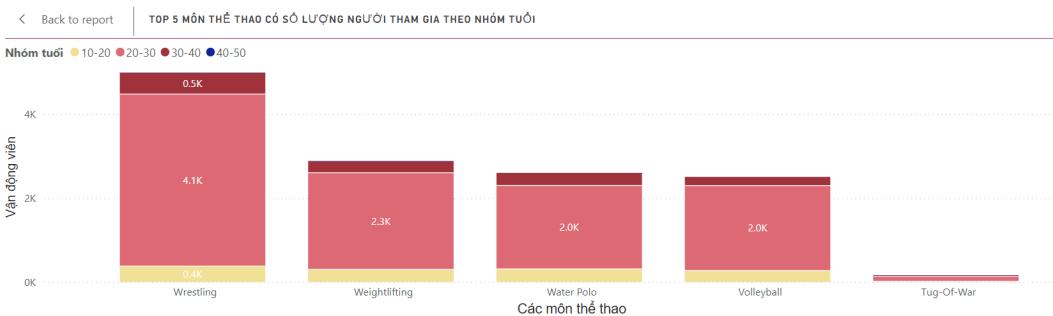
#### 5.3.5.6.1 Tạo visual stacked column top 5 môn thể thao có số lượng huy chương được thu thập theo độ tuổi

### 5.3.5.7 So sánh số lượng huy chương giữa các nhóm tuổi theo giới tính.



5.3.5.7.1 So sánh số lượng huy chương giữa các nhóm tuổi theo giới tính.

### 5.3.5.8 Top 5 môn thể thao có số lượng người tham gia theo nhóm tuổi



5.3.5.8.1

Top 5 môn thể thao có số lượng người tham gia theo nhóm tuổi

### 5.3.5.9 Tạo visual Matrix thống kê huy chương đạt được trong từng môn thể thao theo từng nhóm tuổi

Sport	THỐNG KÊ HUY CHƯƠNG ĐẠT ĐƯỢC TRONG TỪNG MÔN THỂ THAO THEO TỪNG NHÓM TUỔI				
	10-20	20-30	30-40	40-50	50+
Aeronautics		Gold			
Alpine Skiing	Bronze	Bronze	Bronze	Unknown	Unknown
Alpinism		Gold	Gold	Gold	Gold
Archery	Bronze	Bronze	Bronze	Bronze	Bronze
Art Competitions	Bronze	Bronze	Bronze	Bronze	Bronze
Athletics	Bronze	Bronze	Bronze	Bronze	Unknown
Badminton	Bronze	Bronze	Bronze		
Baseball	Bronze	Bronze	Bronze	Unknown	
Basketball	Bronze	Bronze	Bronze		
Basque Pelota		Gold			
Beach Volleyball	Bronze	Bronze	Bronze	Unknown	
Biathlon	Bronze	Bronze	Bronze	Unknown	
Bobsleigh	Bronze	Bronze	Bronze	Bronze	Unknown
Boxing	Bronze	Bronze	Bronze	Unknown	
Canoeing	Bronze	Bronze	Bronze	Bronze	

---

5.3.5.9.1 Tạo visual Matrix thống kê huy chương đạt được trong từng môn thể thao theo từng nhóm tuổi

### 5.3.9 Phân tích theo môn thể thao

#### 5.3.6.1 Tạo visual filter lọc theo giới tính



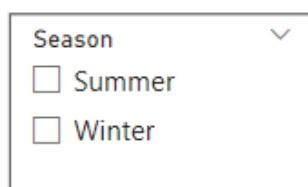
5.3.6.1.1 Tạo visual filter lọc theo giới tính

#### 5.3.6.2 Tạo visual filter lọc theo thời gian



5.3.6.2.1 Tạo visual filter lọc theo thời gian

#### 5.3.6.3 Tạo visual filter lọc theo Season



5.3.6.3.1 Tạo visual filter lọc theo Season

#### 5.3.6.4 Tạo visual thống kê chiều cao trung bình



---

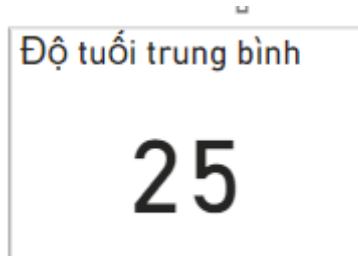
#### 5.3.6.4.1 Tạo visual thống kê chiều cao trung bình

#### 5.3.6.5 Tạo visual thống kê cân nặng trung bình



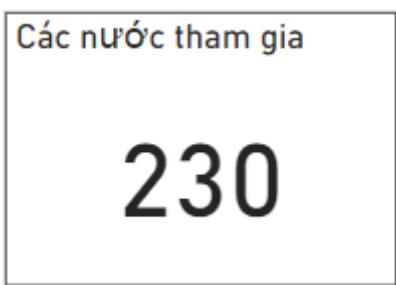
#### 5.3.6.5.1 Tạo visual thống kê cân nặng trung bình

#### 5.3.6.6 Tạo visual thống kê độ tuổi trung bình



#### 5.3.6.6.1 Tạo visual thống kê độ tuổi trung bình

#### 5.3.6.7 Tạo visual thống kê các nước tham gia



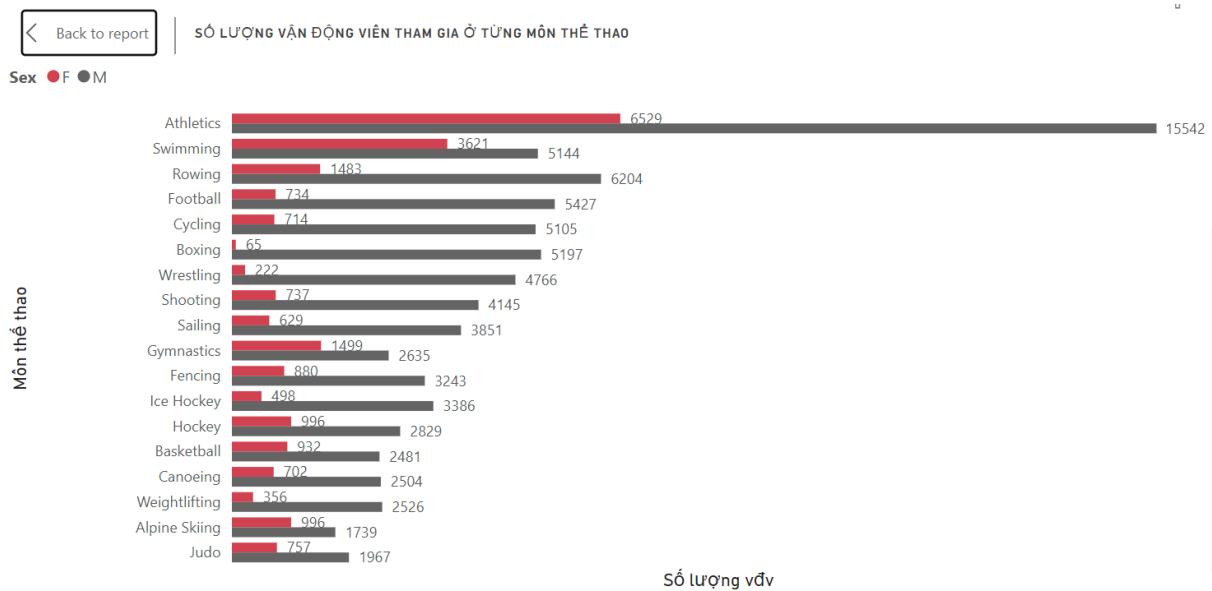
#### 5.3.6.7.1 Tạo visual thống kê các nước tham gia

### 5.3.6.8 Tạo visual thống kê các môn thể thao



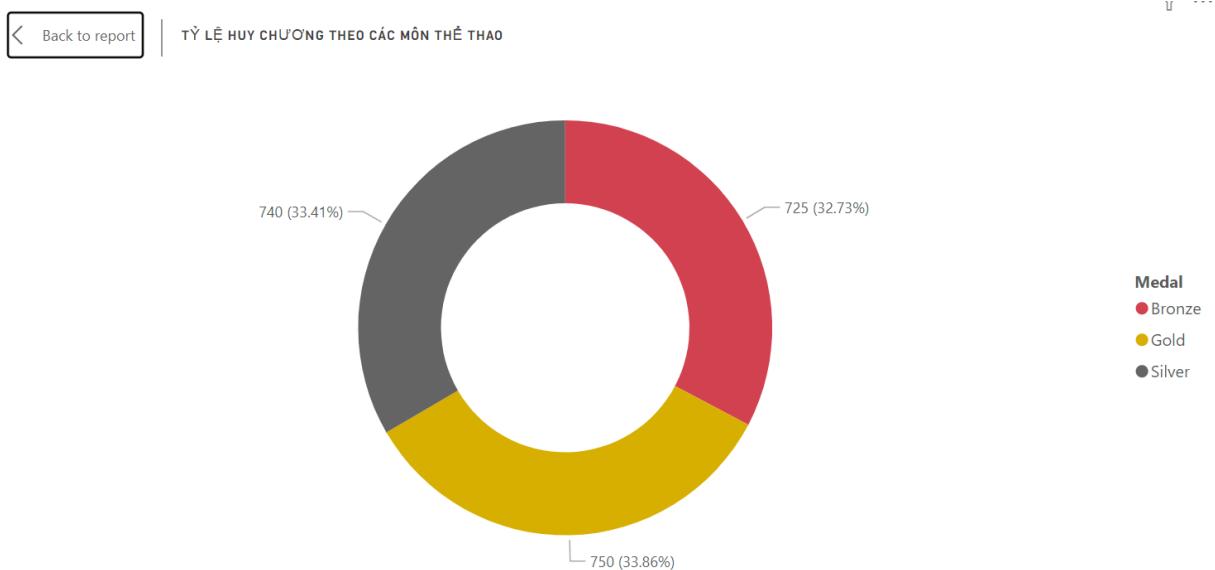
#### 5.3.6.8.1 Tạo visual thống kê các môn thể thao

### 5.3.6.9 Tạo biểu đồ Clustered Column Chart phân tích số lượng vận động viên tham gia ở từng môn thể thao



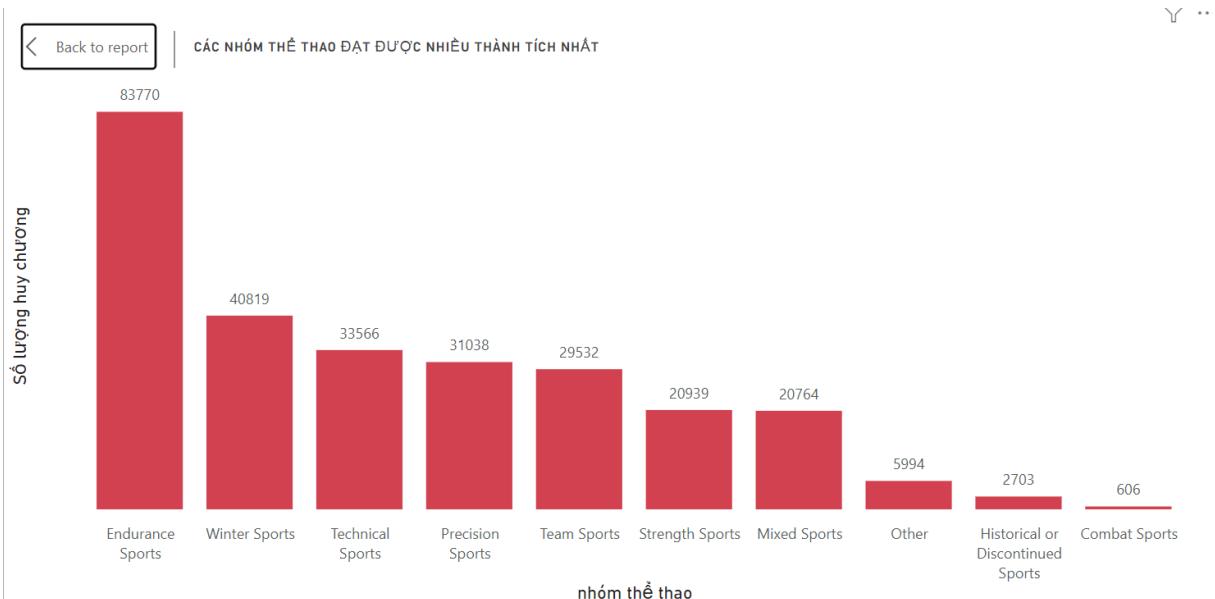
#### 5.3.6.9.1 Tạo biểu đồ Clustered Column Chart phân tích số lượng vận động viên tham gia ở từng môn thể thao

### 5.3.6.10 Tạo biểu đồ Pie Chart phân tích Tỷ lệ huy chương theo môn thể thao



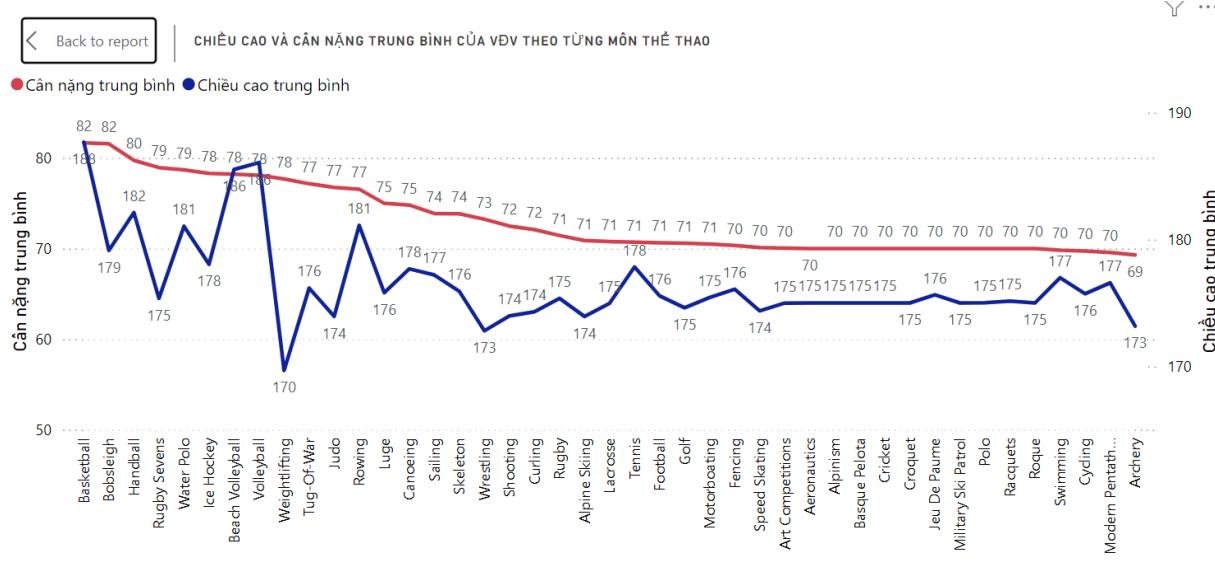
5.3.6.10.1 Tạo biểu đồ Pie Chart phân tích Tỷ lệ huy chương theo môn thể thao

### 5.3.6.11 Tạo biểu đồ Stacked Column Chart phân tích các nhóm thể thao đạt được nhiều thành tích nhất



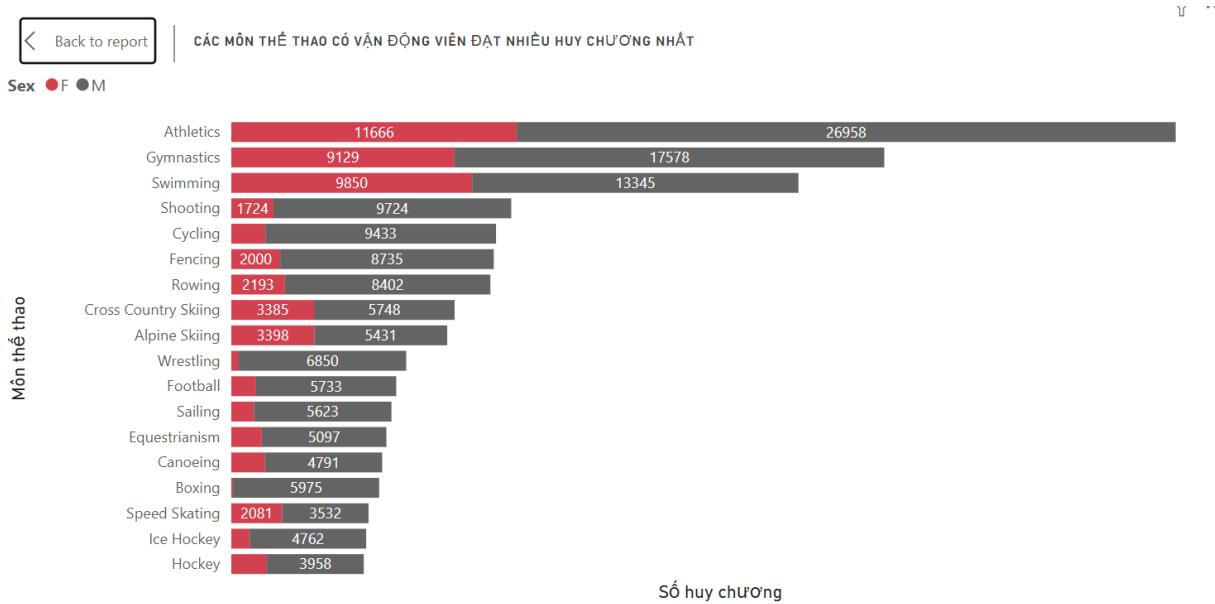
5.3.6.11.1 Tạo biểu đồ Stacked Column Chart phân tích nhóm thể thao đạt được nhiều thành tích nhất

### 5.3.6.12 Tạo biểu đồ Line Chart phân tích Chiều cao và cân nặng trung bình của vđv theo từng môn thể thao



### 5.3.6.12.1 Tạo biểu đồ Line Chart phân tích Chiều cao và cân nặng trung bình của vđv theo từng môn thể thao

### 5.3.6.13 Tạo biểu đồ Clustered Bar Chart phân tích Các môn thể thao có vận động viên đạt nhiều huy chương nhất



---

5.3.6.13.1 Tạo biểu đồ Clustered Bar Chart phân tích Các môn thể thao có vận động viên đạt nhiều huy chương nhất

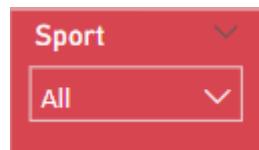
## 5.4 Tạo visual thống kê tổng thể

### 5.4.1 Tạo visual filter lọc theo giới tính



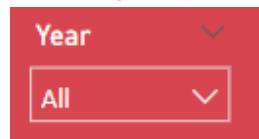
5.3.7.1.1 Tạo visual filter lọc theo giới tính

### 5.4.1.1 Tạo visual filter lọc theo môn thể thao



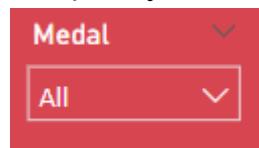
5.3.3.2.1 Tạo visual filter lọc theo môn thể thao

### 5.4.1.2 Tạo visual filter lọc theo thời gian(năm)



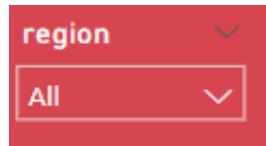
5.3.3.3.1 Tạo visual filter lọc theo thời gian(năm)

### 5.4.1.3 Tạo visual filter lọc theo loại huy chương



5.3.3.4.1 Tạo visual filter lọc theo loại huy chương

#### 5.4.1.4 Tạo visual filter lọc theo quốc gia



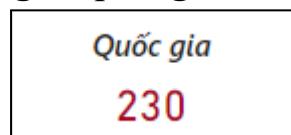
5.3.7.5.1 Tạo visual filter lọc theo quốc gia

#### 5.4.1.5 Tạo visual thống kê tổng số vận động viên



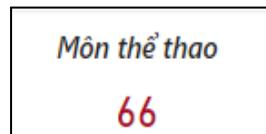
5.3.7.6.1 Tạo visual thống kê tổng số vận động viên

#### 5.4.1.6 Tạo visual thống kê tổng số quốc gia



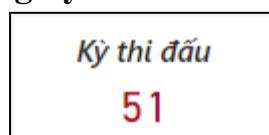
5.3.7.7.1 Tạo visual thống kê tổng số quốc gia

#### 5.4.1.7 Tạo visual thống kê tổng môn thể thao



5.3.7.8.1 Tạo visual thống kê tổng môn thể thao

#### 5.4.1.8 Tạo visual thống kê tổng kỳ thi đấu



5.3.7.9.1 Tạo visual thống kê tổng kỳ thi đấu

#### 5.4.2 Tạo visual thống kê tổng số huy chương vàng

Huy chương vàng

13.37K

5.3.8.1 Tạo visual thống kê tổng số huy chương vàng

#### 5.4.2.1 Tạo visual thống kê tổng số huy chương bạc

Huy chương bạc

13.30K

5.3.8.1.1 Tạo visual thống kê tổng số huy chương bạc

#### 5.4.2.2 Tạo visual thống kê tổng số huy chương đồng

Huy chương đồng

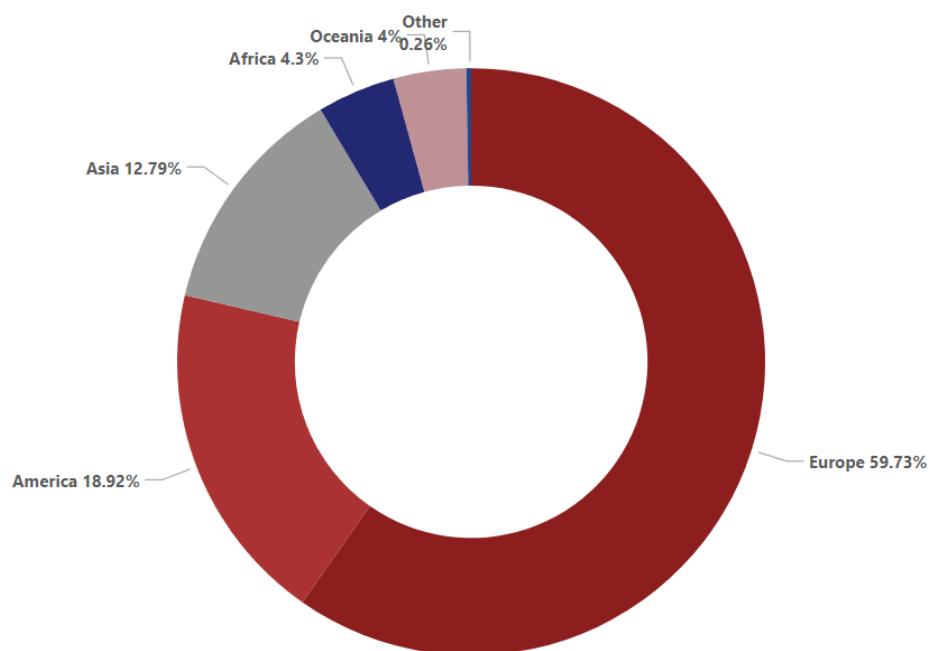
13.11K

5.3.8.2.1 Tạo visual thống kê tổng số huy chương đồng

### 5.4.2.3 Tạo visual tỷ lệ huy chương theo châu lục

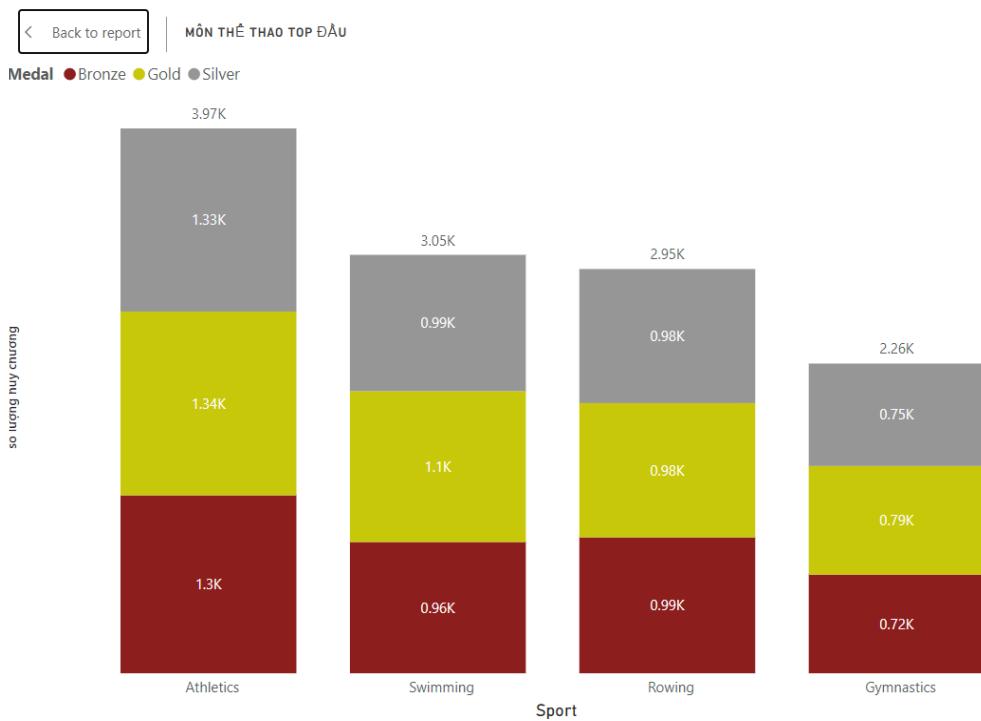
< Back to report

TỶ LỆ HUY CHƯƠNG THEO CHÂU LỤC



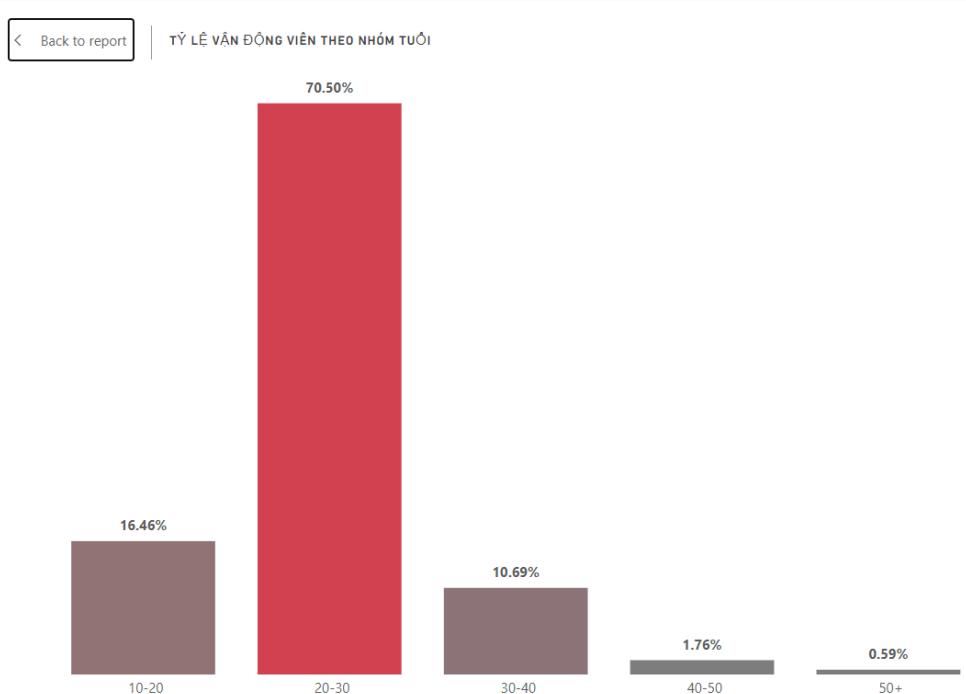
#### 5.3.8.3.1 Tạo visual tỷ lệ huy chương theo châu lục

#### 5.4.2.4 Tạo visual thống kê những môn thể thao top đầu



##### 5.3.8.4.1 Tạo visual thống kê những môn thể thao top đầu

#### 5.4.2.5 Tạo visual thống kê tỷ lệ vận động viên theo nhóm tuổi



##### 5.3.8.5.1 Tạo visual thống kê tỷ lệ vận động viên theo nhóm tuổi

#### 5.4.2.6 Tạo visual thống kê số lượng vận động viên tham gia theo khu vực



##### 5.3.8.6.1 Tạo visual thống kê số lượng vận động viên tham gia theo khu vực

### 5.4.2.7 Tạo visual thống kê thành tích của các vận động viên

< Back to report

THÀNH TÍCH CỦA VẬN ĐỘNG VIÊN

Name	Sport	Region	Total Medals
Michael Fred Phelps, II	Swimming	USA	28
Larysa Semenivna Latynina (Diriy-)	Gymnastics	Russia	18
Nikolay Yefimovich Andrianov	Gymnastics	Russia	15
Borys Anfiyanovich Shakhlin	Gymnastics	Russia	13
Edoardo Mangiarotti	Fencing	Italy	13
Ole Einar Bjørndalen	Biathlon	Norway	13
Takashi Ono	Gymnastics	Japan	13
Aleksey Yuryevich Nemov	Gymnastics	Russia	12
Birgit Fischer-Schmidt	Canoeing	Germany	12
Dara Grace Torres (-Hoffman, -Minas)	Swimming	USA	12
Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)	Swimming	USA	12
Natalie Anne Coughlin (-Hall)	Swimming	USA	12
Paavo Johannes Nurmi	Athletics	Finland	12
Ryan Steven Lochte	Swimming	USA	12
Sawao Kato	Gymnastics	Japan	12
Carl Townsend Osburn	Shooting	USA	11
Mark Andrew Spitz	Swimming	USA	11
Matthew Nicholas "Matt" Biondi	Swimming	USA	11
Viktor Ivanovich Chukarin	Gymnastics	Russia	11
Vra slavsk (-Odlojov)	Gymnastics	Czech Republic	11
Akinori Nakayama	Gymnastics	Japan	10
Aladr Gerevich (-Gerei)	Fencing	Hungary	10
Aleksandr Nikolayevich Dityatin	Gymnastics	Russia	10
Franziska van Almsick	Swimming	Germany	10
Frederick Carlton "Carl" Lewis	Athletics	USA	10

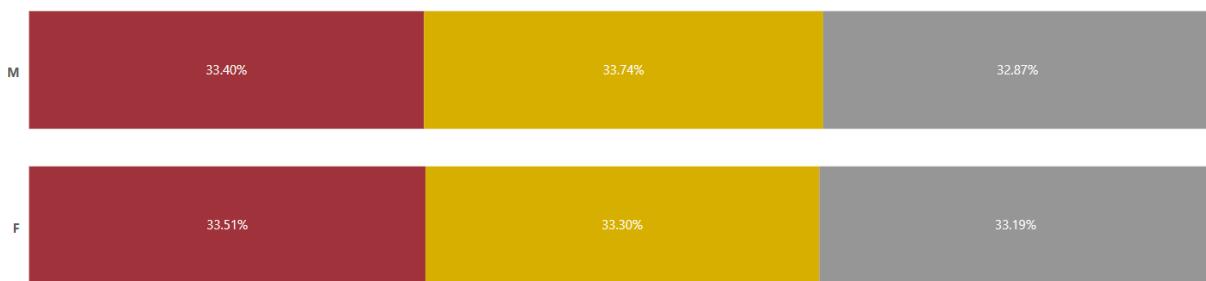
### 5.3.8.7.1 Tạo visual thống kê thành tích của các vận động viên

### 5.4.2.8 Tạo visual thống kê tỷ lệ huy chương giữa nam và nữ

< Back to report

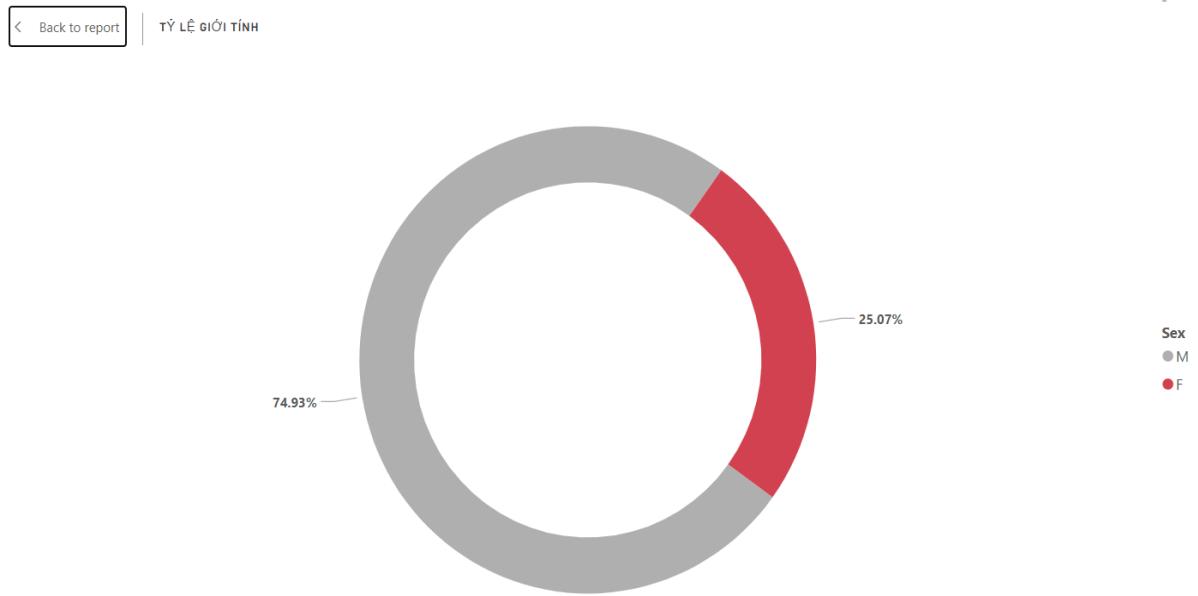
TỶ LỆ HUY CHƯƠNG GIỮA NAM VÀ NỮ

Medal ● Bronze ■ Gold ● Silver



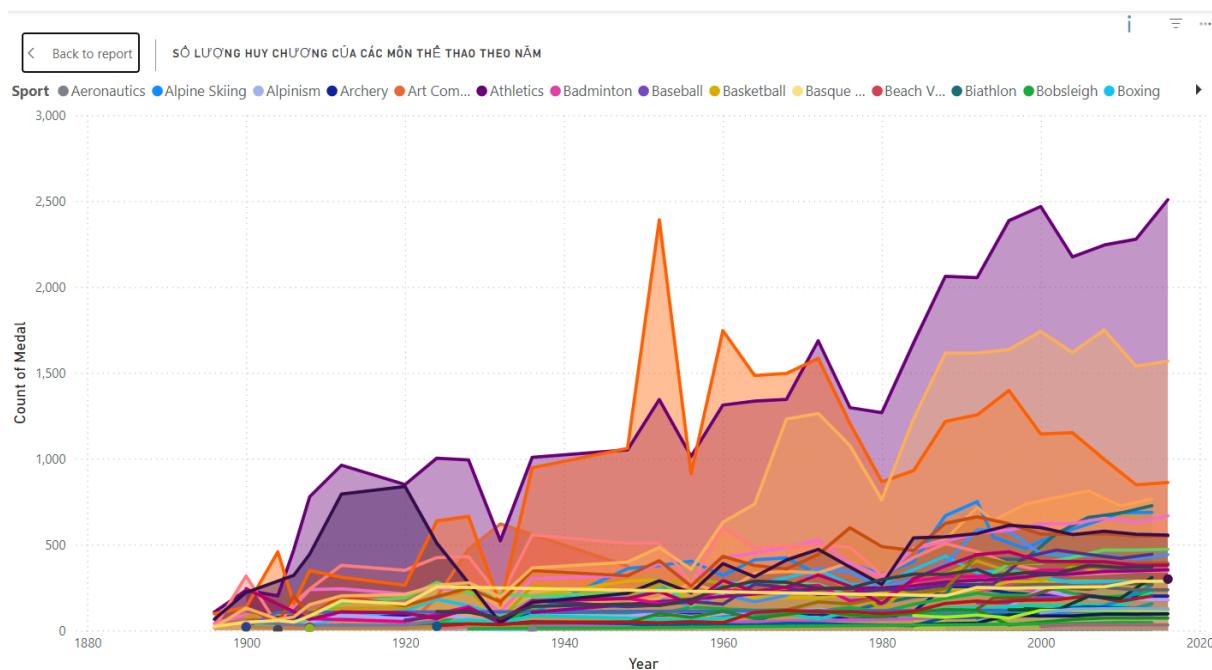
### 5.3.8.8.1 Tạo visual thống kê tỷ lệ huy chương giữa nam và nữ

### 5.4.2.9 Tạo visual thống kê tỷ lệ giới tính



#### 5.3.8.9.1 Tạo visual tỷ lệ giới tính

### 5.4.3 Tạo visual thống kê số lượng huy chương của các môn thể thao theo năm



#### 5.3.9.1 Tạo visual thống kê số lượng huy chương của các môn thể thao theo năm

## 6 Xây dựng báo cáo

### 6.1 Dashboard và report

#### Cách tối ưu hóa

##### Dashboard:

- **Sử dụng trực quan phù hợp:** Chọn biểu

đồ và đồ thị phù hợp để truyền đạt thông tin một cách hiệu quả. Ví dụ, sử dụng biểu đồ đường để theo dõi xu hướng theo thời gian, biểu đồ cột để so sánh các danh mục và biểu đồ tròn để hiển thị tỷ lệ phần trăm.

- **Sắp xếp hợp lý:** Sắp xếp các biểu đồ và đồ thị một cách logic và dễ hiểu. Đặt các thông tin quan trọng nhất ở vị trí nổi bật và sử dụng tiêu đề, chú thích rõ ràng để giải thích nội dung.

- **Thiết kế đơn giản:** Tránh sử dụng quá nhiều chi tiết và đồ họa không cần thiết. Thiết kế dashboard đơn giản, dễ nhìn và tập trung vào thông tin quan trọng.

##### Report:

### 6.2 Xây dựng báo cáo

#### 6.2.1 Dashboard vs Report

Tối ưu hóa dashboard trong Power BI là việc quan trọng để tăng hiệu quả sử dụng, cải thiện trải nghiệm người dùng và đảm bảo hiệu suất.

#### Tối ưu hiệu suất

-Loại bỏ các cột hoặc bảng không cần thiết.

-Sử dụng loại dữ liệu phù hợp (ví dụ: số nguyên thay vì số thập phân nếu không cần thiết).

-Tóm tắt dữ liệu trước khi đưa vào Power BI, chỉ lấy dữ liệu cần thiết.

### Sử dụng các mối quan hệ hiệu quả:

-Ưu tiên Star Schema thay vì Snowflake Schema để cải thiện hiệu suất.

-Tránh các mối quan hệ Many-to-Many khi không cần thiết.

### Tối ưu hóa Measures:

-Sử dụng các measure đơn giản và tránh tạo measure lồng nhau phức tạp.

-Tận dụng các hàm DAX như SUMX, CALCULATE một cách hiệu quả.

### Tổ chức trực quan:

-Sử dụng bố cục rõ ràng, không quá tải thông tin.

-Ưu tiên các thành phần quan trọng ở khu vực dễ nhìn (trên cùng hoặc bên trái).

### Sử dụng màu sắc hợp lý:

-Tránh dùng quá nhiều màu sắc, tập trung vào bảng màu thống nhất.

-Sử dụng màu sắc để làm nổi bật dữ liệu quan trọng.

### Tương tác thân thiện:

-Sử dụng Slicer, Button để tạo trải nghiệm tương tác mượt mà.

-Tận dụng Drillthrough và Tooltip để hiển thị thông tin chi tiết.

### Tối ưu kích thước dashboard:

-Thiết kế dashboard phù hợp với các thiết bị hiển thị khác nhau (desktop, tablet, điện thoại).

### Tạo Templates:

-Thiết kế một template chung để sử dụng lại cho nhiều dashboard khác nhau.

### Tái sử dụng Measures:

-Viết các measures chung có thể dùng cho nhiều mục đích thay vì lặp lại công thức.

#### Sử dụng Parameters:

-Tạo các tham số để dễ dàng thay đổi giá trị đầu vào cho báo cáo.

#### Giảm tải dữ liệu làm mới:

-Chỉ làm mới dữ liệu cần thiết thay vì toàn bộ dataset.

-Sử dụng chế độ Import thay vì DirectQuery nếu dataset không thay đổi thường xuyên.

#### Thiết lập thời gian làm mới hợp lý:

-Cấu hình tần suất làm mới phù hợp với nhu cầu người dùng.

#### Sử dụng Performance Analyzer:

-Công cụ trong Power BI giúp kiểm tra tốc độ tải và nhận diện các thành phần chậm.

#### Đánh giá phản hồi người dùng:

-Thu thập ý kiến từ người dùng để cải thiện giao diện và cách trình bày.

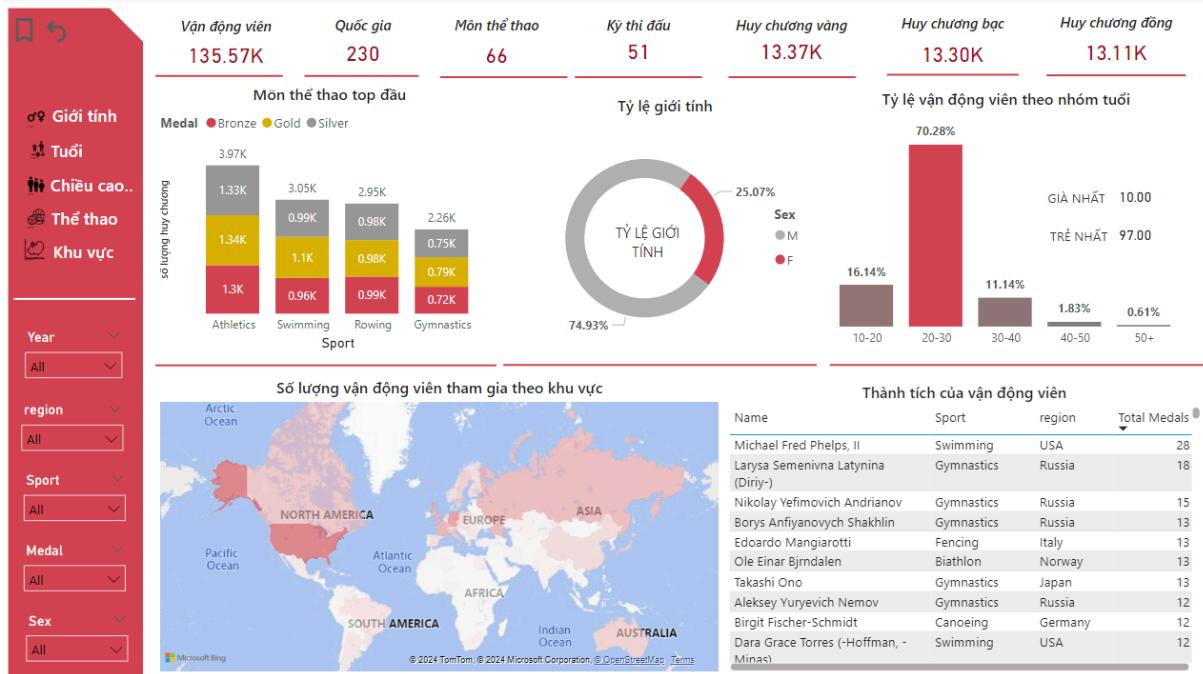
#### Giám sát qua Power BI Service:

-Theo dõi hiệu suất và thời gian làm mới qua Power BI Service.

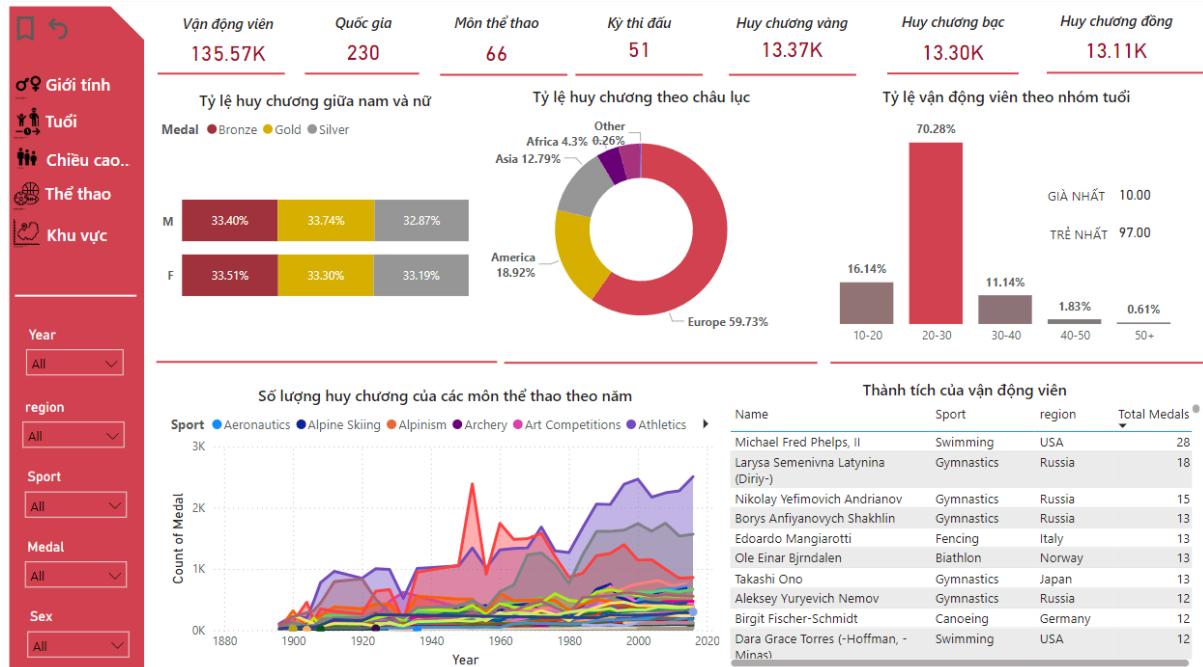
### 6.2.2 Dashboard

Dashboard được thiết kế nhằm trực quan hóa dữ liệu Olympic đa chiều, giúp phân tích các yếu tố ảnh hưởng đến hiệu suất như tuổi, giới tính, chiều cao, cân nặng, và khu vực. Nó hỗ trợ so sánh, đánh giá xu hướng, tối ưu hóa việc sử dụng dữ liệu lớn và kể câu chuyện dữ liệu hiệu quả, từ đó cung cấp thông tin chi tiết, hỗ trợ ra quyết định và phục vụ nhu cầu phân tích chuyên sâu của người dùng.

## Phân tích tổng quan



### 6.2.2.1 Ánh overview



### 6.2.2.2 Ánh overview

Tỷ lệ vận động viên theo nhóm tuổi - Clustered Column Chart:

**Mục đích:** Hiển thị tỷ lệ vận động viên theo từng nhóm tuổi. Nó giúp so sánh số lượng vận động viên ở các độ tuổi khác nhau, giúp nhận diện xu hướng và sự phân bố độ tuổi trong các kỳ Thế vận hội.

Tỷ lệ giành huy chương giữa các nhóm giới tính - Stacked Bar Chart:

**Mục đích:** Giúp so sánh tỷ lệ giành huy chương giữa nam và nữ trong các kỳ Thế vận hội. Nó thể hiện sự phân bổ huy chương (vàng, bạc, đồng) giữa các giới tính, từ đó có thể đánh giá sự tham gia và thành tích của nam và nữ trong các môn thể thao.

Tỷ lệ giới tính của các vận động viên - Donut Chart:

**Mục đích:** Cung cấp cái nhìn trực quan về tỷ lệ giới tính trong nhóm vận động viên tham gia Thế vận hội. Nó giúp hình dung tỷ lệ giữa nam và nữ trong tổng số vận động viên.

Số lượng vận động viên tham gia theo vùng địa lý - Filled Map:

**Mục đích:** Giúp hiển thị số lượng vận động viên tham gia từ các khu vực địa lý khác nhau. Nó cho phép so sánh sự tham gia giữa các khu vực như châu Á, châu Âu, châu Mỹ, v.v., giúp nhận ra các khu vực có sự tham gia đông đảo nhất.

So sánh thành tích thể thao giữa các môn theo thời gian - Line Chart:

**Mục đích:** Giúp so sánh sự thay đổi trong thành tích của các môn thể thao theo thời gian. Nó có thể giúp nhận ra xu hướng và sự thay đổi về thành tích trong các kỳ Thế vận hội, giúp phân tích sự tiến bộ của các môn thể thao qua các năm.

Top những môn thể thao giành nhiều huy chương nhất - Stacked Column Chart:

**Mục đích:** Giúp so sánh các môn thể thao giành nhiều huy chương nhất, thể hiện sự phân bổ huy chương (vàng, bạc, đồng) trong từng môn thể thao. Nó giúp xác định các môn thể thao nào thành công nhất tại các kỳ Thế vận hội.

Thống kê thành tích của các vận động viên - Table:

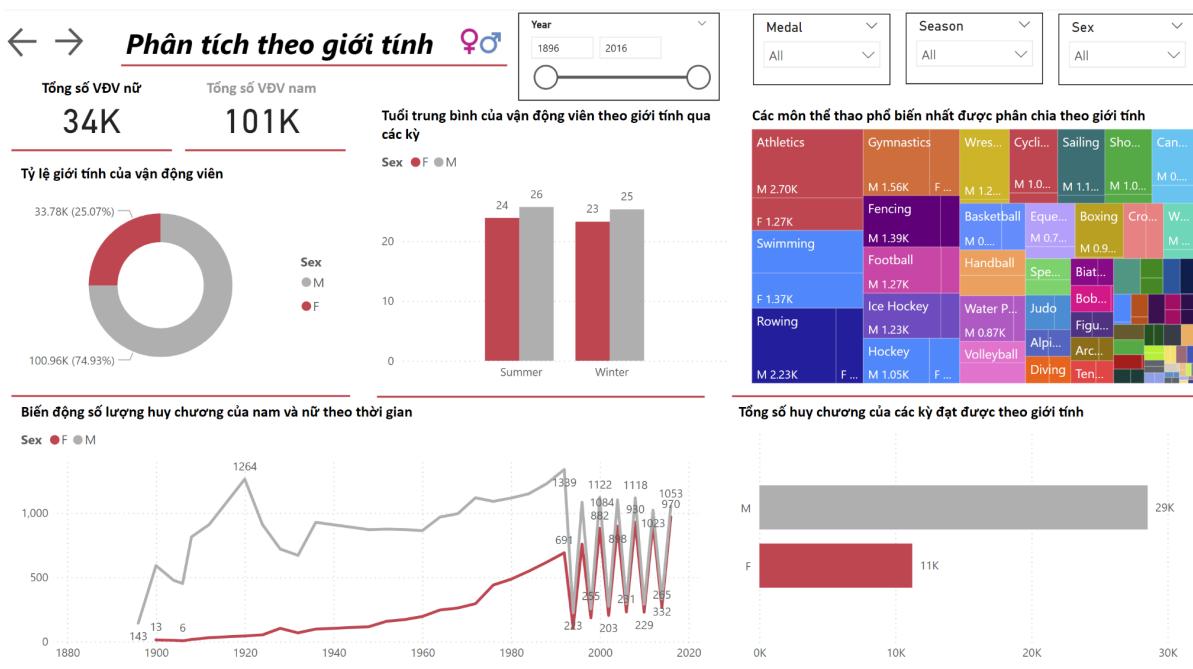
**Mục đích:** Cung cấp thông tin chi tiết về thành tích của các vận động viên, như số huy chương giành được, kết quả thi đấu, và các chỉ số quan trọng khác. Nó

giúp theo dõi và phân tích thành tích của từng vận động viên trong các kỳ Thế vận hội.

Tỷ lệ huy chương của các châu lục - Donut Chart:

**Mục đích:** Hiển thị tỷ lệ huy chương giành được bởi các châu lục trong Thế vận hội. Nó giúp nhận diện các châu lục có thành tích nổi bật nhất, từ đó đánh giá sự đóng góp của từng châu lục vào thành công chung của Thế vận hội.

### Phân tích theo giới tính

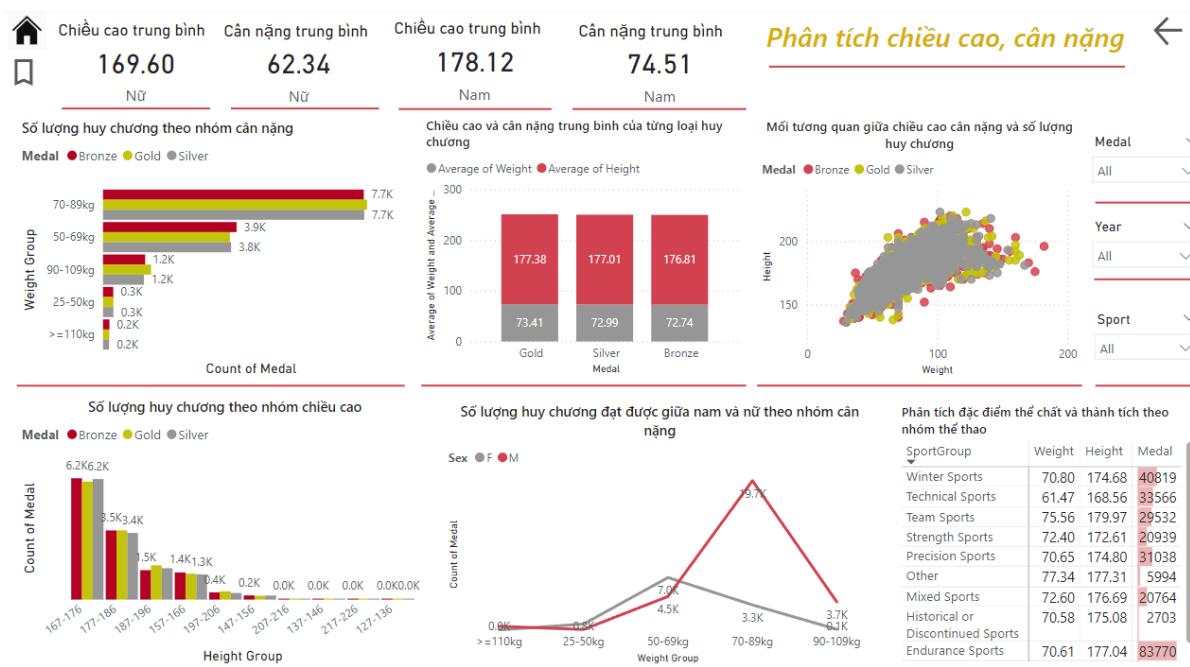


#### 6.2.2.3 Ảnh phân tích theo giới tính

- Tỷ lệ giới tính của vận động viên - Pie Chart
- Tổng số huy chương đạt được theo giới tính - Stacked Bar Chart
- Biến động số lượng huy chương của nam và nữ theo thời gian - Line Chart
- Tuổi trung bình của vận động viên qua các kỳ Thế vận hội theo giới tính - Clustered Column Chart
- Các môn thể thao phổ biến nhất được phân chia theo giới tính - Bar Chart

- So sánh thành tích theo giới tính trong từng mùa thi đấu (Summer vs Winter) - Clustered Column Chart
- Tỷ lệ huy chương giữa các nhóm tuổi theo giới tính - Stacked Bar Chart
- Số huy chương theo môn thể thao và giới tính - Stacked Bar Chart
- Số lượng huy chương theo nhóm tuổi: Clustered bar chart

### Phân tích theo chiều cao và cân nặng



#### 6.2.2.4 Ánh phân tích theo chiều cao và cân nặng

- Số lượng huy chương theo nhóm cân nặng: Clustered column chart

**Mục đích:** Hiển thị số huy chương mà các vận động viên giành được ở các nhóm cân nặng khác nhau, phân loại theo loại huy chương.

- Chiều cao và cân nặng trung bình của từng loại huy chương: Stacked column bar chart

**Mục đích:** So sánh chiều cao và cân nặng trung bình của các vận động viên đã giành huy chương vàng, bạc và đồng. Mỗi cột đại diện cho một loại huy chương,

trong đó các phần xếp chồng thể hiện chiều cao và cân nặng trung bình của vận động viên trong từng nhóm huy chương.

- Mối tương chiều cao cân nặng và số lượng huy chương: Scatter plot

**Mục đích:** Hiển thị mối tương quan giữa chiều cao, cân nặng của vận động viên và số huy chương của họ, trong đó các màu khác nhau đại diện cho các loại huy chương khác nhau.

- Số lượng huy chương theo nhóm chiều cao : Clustered column chart

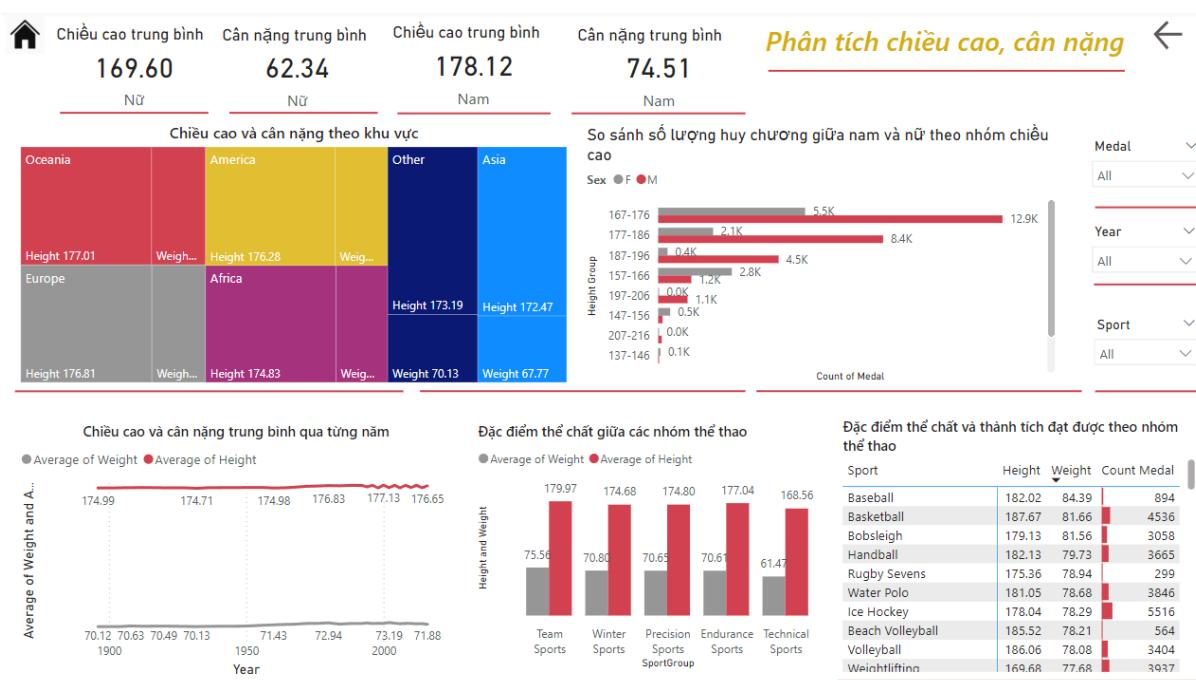
**Mục đích:** Hiển thị số huy chương mà các vận động viên giành được ở các nhóm chiều cao khác nhau, phân loại theo loại huy chương.

- Số lượng huy chương đạt được giữa theo nhóm cân nặng: Line

### Mục đích:

- Phân tích đặc điểm thể chất và thành tích theo nhóm thể thao: Table

**Mục đích:** Hiển thị thành tích của từng chiều cao và cân nặng trung bình của các vận động viên ở nhiều hạng mục thể thao khác nhau (ví dụ: Thể thao đồng đội, Thể thao mùa đông)



### 6.2.2.5 Ảnh phân tích chiều cao và cân nặng

- 
- Chiều cao và cân nặng theo khu vực: Tree map

**Mục đích:** Hiển thị sự phân bố chiều cao và cân nặng của các vận động viên trên khắp các khu vực khác nhau (Châu Đại Dương, Châu Âu, Châu Mỹ, Châu Phi, Châu Á). Kích thước của mỗi ô biểu thị tỷ lệ vận động viên và mã màu biểu thị số liệu chiều cao và cân nặng của từng khu vực.

- So sánh số lượng huy chương giữa nam và nữ: Clustered bar chart

**Mục đích:** giúp so sánh số lượng huy chương giành được giữa nam và nữ trong các môn thể thao. Nó thể hiện rõ ràng sự phân biệt về thành tích giữa hai giới, từ đó giúp đánh giá sự tham gia và thành tích của nam và nữ trong các kỳ Thế vận hội.

- Chiều cao và cân nặng trung bình qua từng năm: Line

**Mục đích:** Mục đích: Theo dõi sự thay đổi về chiều cao và cân nặng trung bình của các vận động viên Olympic qua nhiều năm.

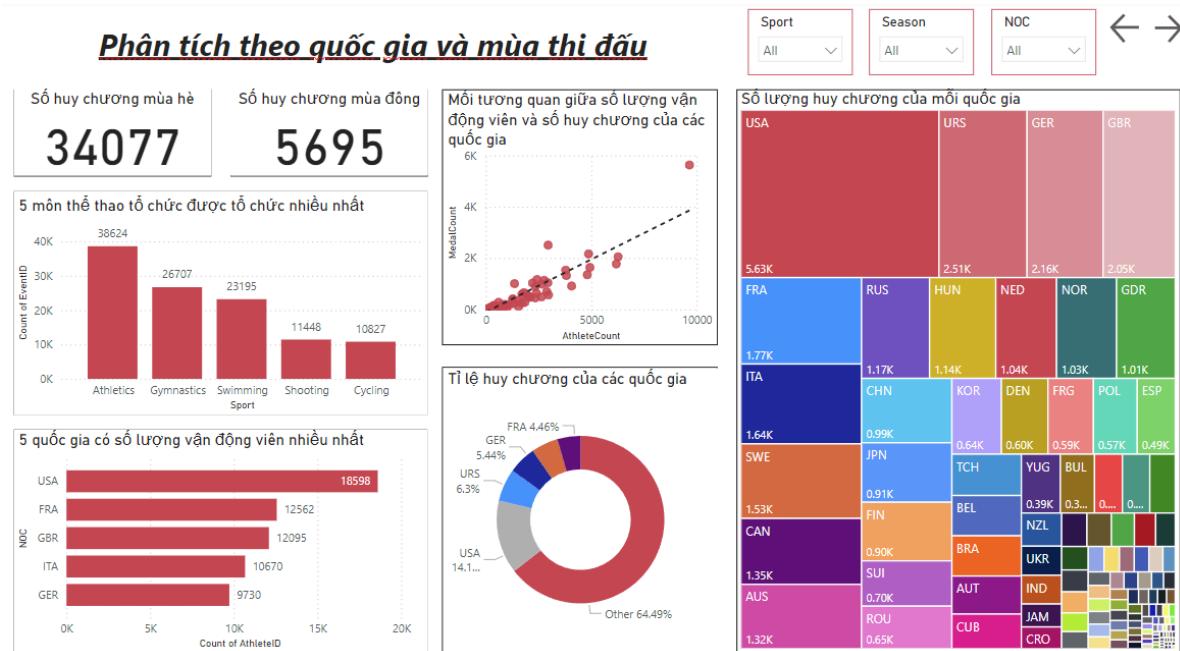
- Đặc điểm thể chất giữa các nhóm thể thao: Clustered column chart

**Mục đích:** giúp so sánh các đặc điểm thể chất (như chiều cao, cân nặng) của vận động viên giữa các nhóm thể thao khác nhau. Nó giúp nhận diện sự khác biệt về thể chất của các vận động viên tham gia các nhóm thể thao, từ đó hỗ trợ phân tích yếu tố thể chất ảnh hưởng đến thành tích thi đấu.

- Đặc điểm thể chất và thành tích đạt được theo từng môn thể thao: Table

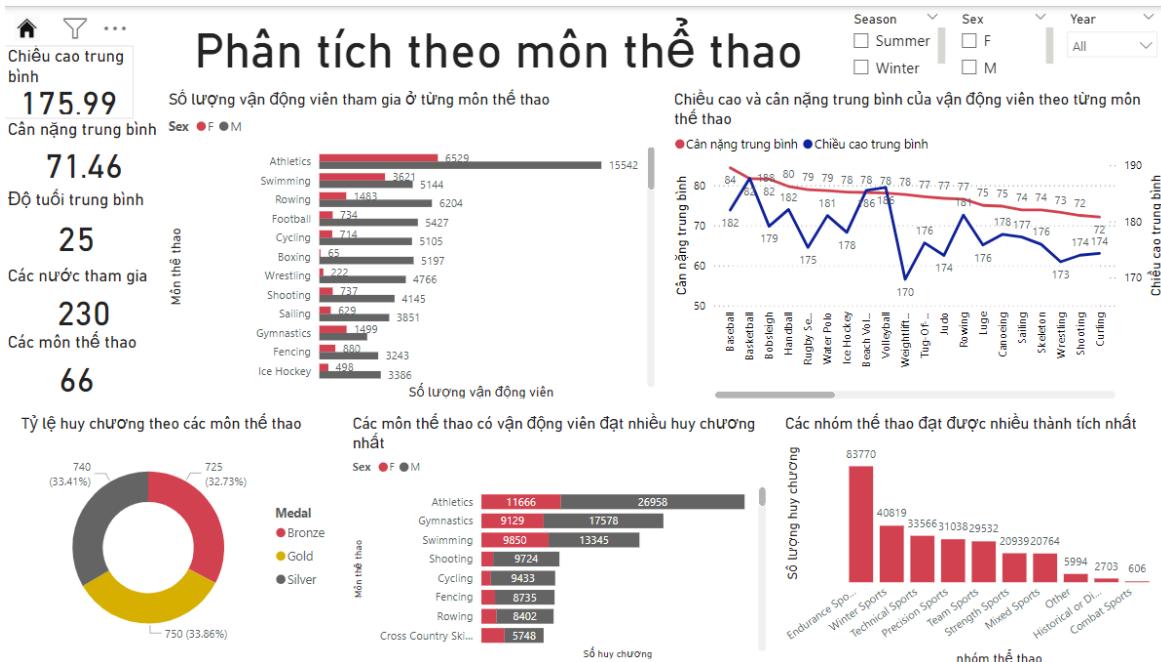
**Mục đích:** hiển thị các đặc điểm thể chất (như chiều cao, cân nặng) và thành tích (huy chương, kết quả thi đấu) của các nhóm thể thao. Nó giúp người xem dễ dàng so sánh thông tin chi tiết về thể chất và thành tích của vận động viên trong từng nhóm thể thao, hỗ trợ việc phân tích sự khác biệt giữa các nhóm thể thao.

## Phân tích theo quốc gia và mùa thi đấu



### 6.2.2.5 Ánh phân tích theo quốc gia và mùa thi đấu

## Phân tích theo môn thể thao



### 6.2.2.6 Ánh phân tích theo môn thể thao

- Số lượng vận động viên tham gia ở từng môn thể thao - Clustered column chart

**Mục đích:** So sánh số lượng vận động viên nam và nữ ở các môn thể thao.

Dễ dàng nhận biết các môn thể thao thu hút nhiều vận động viên nhất

- Tỷ lệ huy chương theo môn thể thao - Pie chart

**Mục đích:** Minh họa tỷ lệ phân bổ bổ sung của 3 loại huy chương (vàng, bạc, đồng) giúp người xem xác định mức độ cân bằng trong công việc phân phối huy động giữa các môn thể thao.

- Các quốc gia đạt được nhiều thành tích nhất - Stacked column chart

**Mục đích:** Phân tích thành tích huy chương theo nhóm thể thao lớn

Giúp thấy rõ nhóm nào chiếm ưu thế, với thể thao sức bền dẫn đầu.

- Chiều cao và cân nặng trung bình của vận động viên theo từng môn thể thao - Line chart

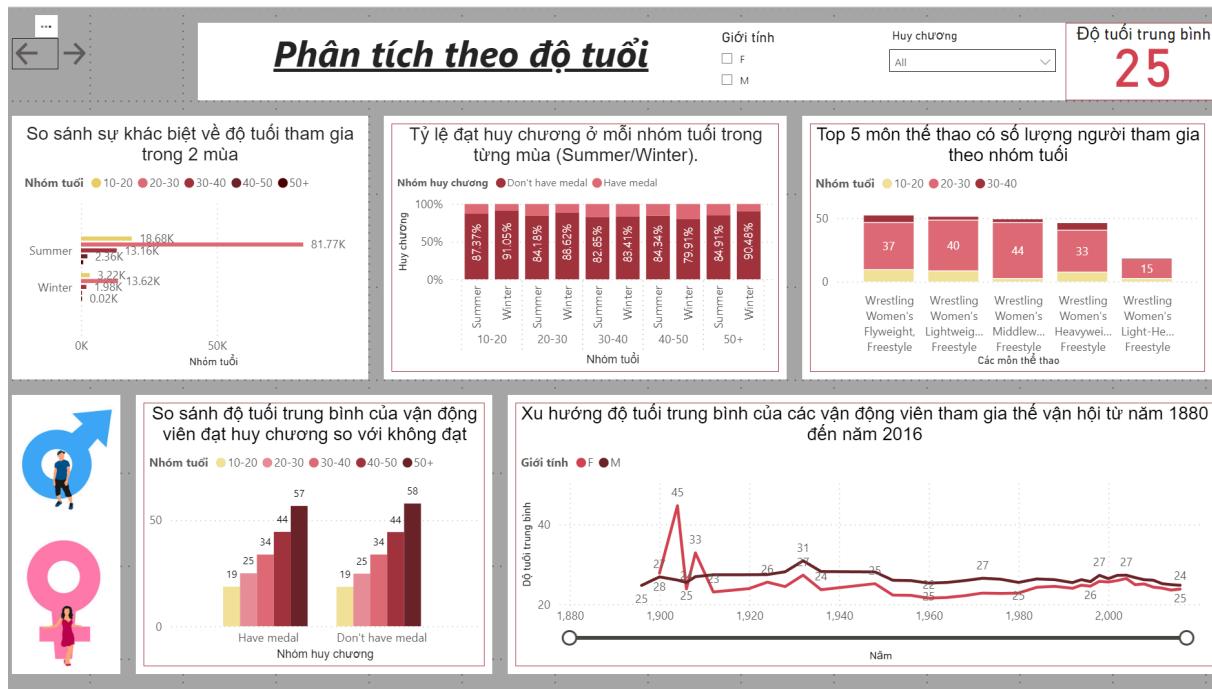
**Mục đích:** So sánh chiều cao và cân nặng trung bình của vận động viên giữa các môn thể thao.

Giúp xác định đặc điểm chất lý tưởng của vận động viên theo môn

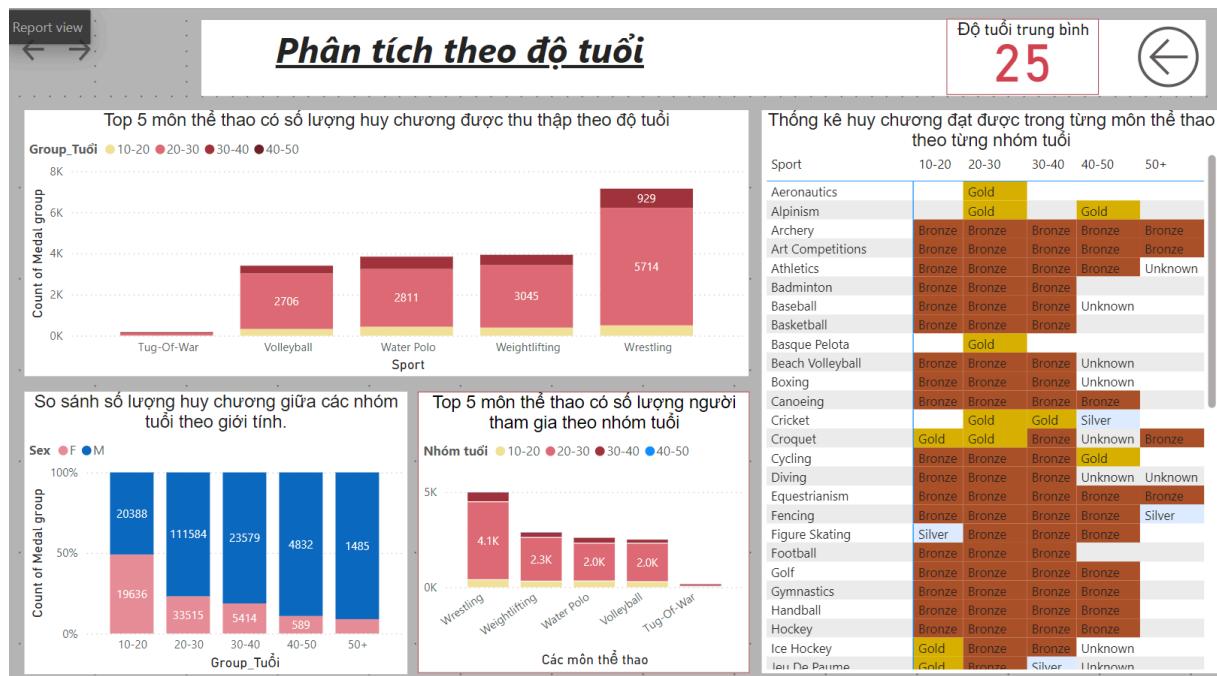
- Các nhóm thể thao đạt được nhiều thành tích nhất - Clustered bar chart

**Mục đích:** Xác định các môn thể thao đạt được nhiều huy chương nhất. Điền kinh (điền kinh) và thể dục dụng cụ (thể dục dụng cụ) là hai môn có thành tích huy chương vượt trội.

## Phân tích theo độ tuổi



6.2.2.7 Ảnh phân tích theo độ tuổi trang 1



6.2.2.7 Ảnh Phân tích theo độ tuổi trang 2

- So sánh sự khác biệt về độ tuổi tham gia trong 2 mùa - Clustered Bar Chart  
Mục đích: Hiểu sự phân bố độ tuổi của vận động viên ở hai mùa giải (Summer và Winter), từ đó xác định mùa nào có sự tham gia đông đảo hơn ở từng nhóm tuổi.

- Xu hướng độ tuổi trung bình của các vận động viên tham gia thể vận hội từ năm 1880 đến năm 2016 - Line Chart

Mục đích: Xem xét sự thay đổi độ tuổi trung bình của vận động viên nam và nữ theo thời gian, để nhận ra các xu hướng lịch sử trong thể thao Olympic.

- Tỷ lệ đạt huy chương ở mỗi nhóm tuổi trong từng mùa (Summer/Winter). - 100% stacked Column Chart

Mục đích: Tìm ra tỉ lệ đạt huy chương theo từng nhóm tuổi của từng mùa summer và winter

- Top 5 môn thể thao có số lượng người tham gia theo nhóm tuổi - Stacked Bar Chart

Mục đích: Xác định các môn thể thao phổ biến nhất theo từng nhóm tuổi, giúp nhận biết xu hướng lựa chọn môn thể thao dựa trên độ tuổi.

- So sánh độ tuổi trung bình của vận động viên đạt huy chương so với không đạt- Clustered Column Chart

Mục đích: Phân tích xem độ tuổi có ảnh hưởng như thế nào đến khả năng đoạt huy chương, và so sánh các nhóm tuổi giữa vận động viên có huy chương và không có huy chương.

- So sánh số lượng huy chương giữa các nhóm tuổi theo giới tính - 100% stacked Column Chart

- Top 5 môn thể thao có số lượng huy chương được thu thập theo độ tuổi - Stacked Column Chart

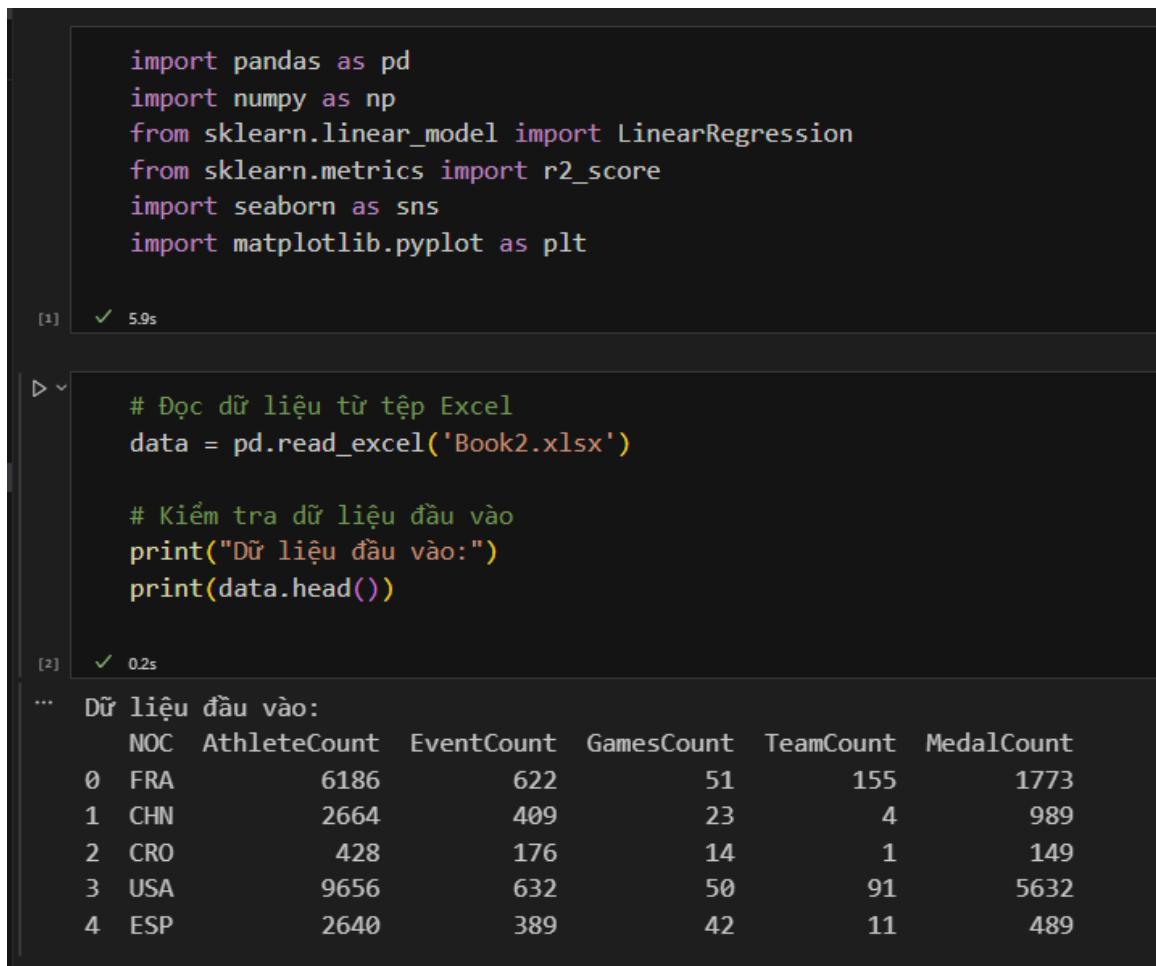
- Top 5 môn thể thao có số lượng người tham gia theo nhóm tuổi - Stacked Column Chart

### 6.2.3 Xây dựng mô hình dự báo

- Mô hình hồi quy tuyến tính:

$$\mathbf{y} = w_0 + \sum_{i=1}^m w_i * \mathbf{x}_i + \epsilon$$

#### 6.2.3.1 Ảnh công thức hồi quy tuyến tính



```

import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import seaborn as sns
import matplotlib.pyplot as plt

[1]: ✓ 5.9s

# Đọc dữ liệu từ tệp Excel
data = pd.read_excel('Book2.xlsx')

# Kiểm tra dữ liệu đầu vào
print("Dữ liệu đầu vào:")
print(data.head())

```

[2]: ✓ 0.2s

Dữ liệu đầu vào:

	NOC	AthleteCount	EventCount	GamesCount	TeamCount	MedalCount
0	FRA	6186	622	51	155	1773
1	CHN	2664	409	23	4	989
2	CRO	428	176	14	1	149
3	USA	9656	632	50	91	5632
4	ESP	2640	389	42	11	489

#### 6.2.3.2 Ảnh kiểm tra dữ liệu đầu vào

```
# Tính hệ số tương quan giữa các cột
correlation_matrix = data[['AthleteCount', 'EventCount', 'GamesCount', 'TeamCount', 'MedalCount']].corr()
print("Hệ số tương quan:")
print(correlation_matrix)

# Vẽ heatmap biểu diễn hệ số tương quan
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Ma trận tương quan")
plt.show()

[3]: ✓ 0.2s
```

... Hệ số tương quan:

	AthleteCount	EventCount	GamesCount	TeamCount	MedalCount
AthleteCount	1.000000	0.896596	0.749959	0.815560	0.921625
EventCount	0.896596	1.000000	0.838224	0.671013	0.738264
GamesCount	0.749959	0.838224	1.000000	0.584561	0.567468
TeamCount	0.815560	0.671013	0.584561	1.000000	0.744491
MedalCount	0.921625	0.738264	0.567468	0.744491	1.000000

... Ma trận tương quan

### 6.2.3.3 Tính hệ số tương quan giữa các cột

```
# Xây dựng mô hình hồi quy tuyến tính
X = data[['AthleteCount', 'EventCount', 'GamesCount', 'TeamCount']] # Các biến độc lập
y = data['MedalCount'] # Biến phụ thuộc

# Khởi tạo mô hình
model = LinearRegression()

# Huấn luyện mô hình
model.fit(X, y)

[5] ✓ 0.0s
...
    ▾ LinearRegression ⓘ ⓘ
LinearRegression()
```

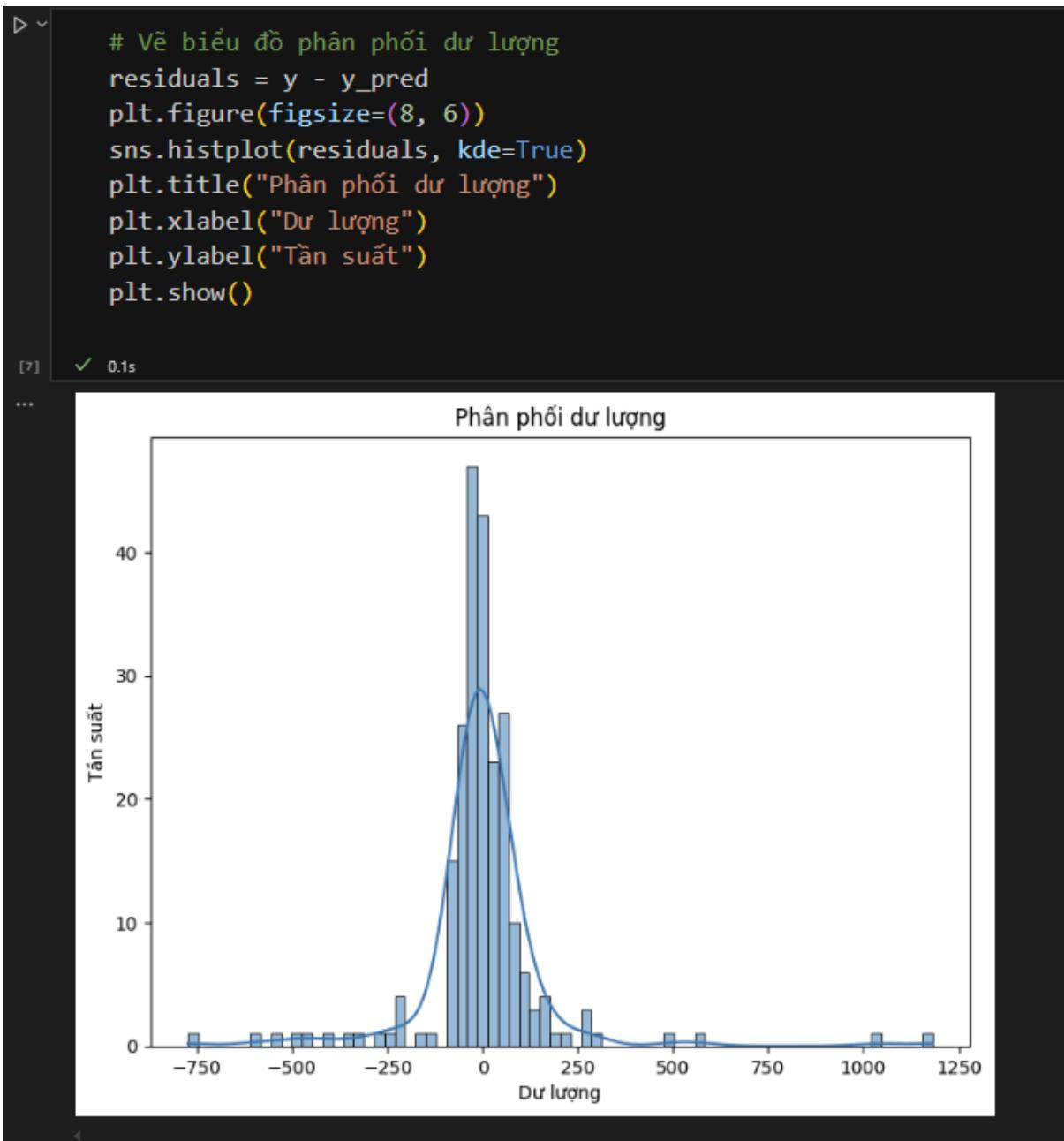
#### 6.2.3.4 Xây dựng mô hình hồi quy tuyến tính

```
# Dự đoán giá trị
y_pred = model.predict(X)

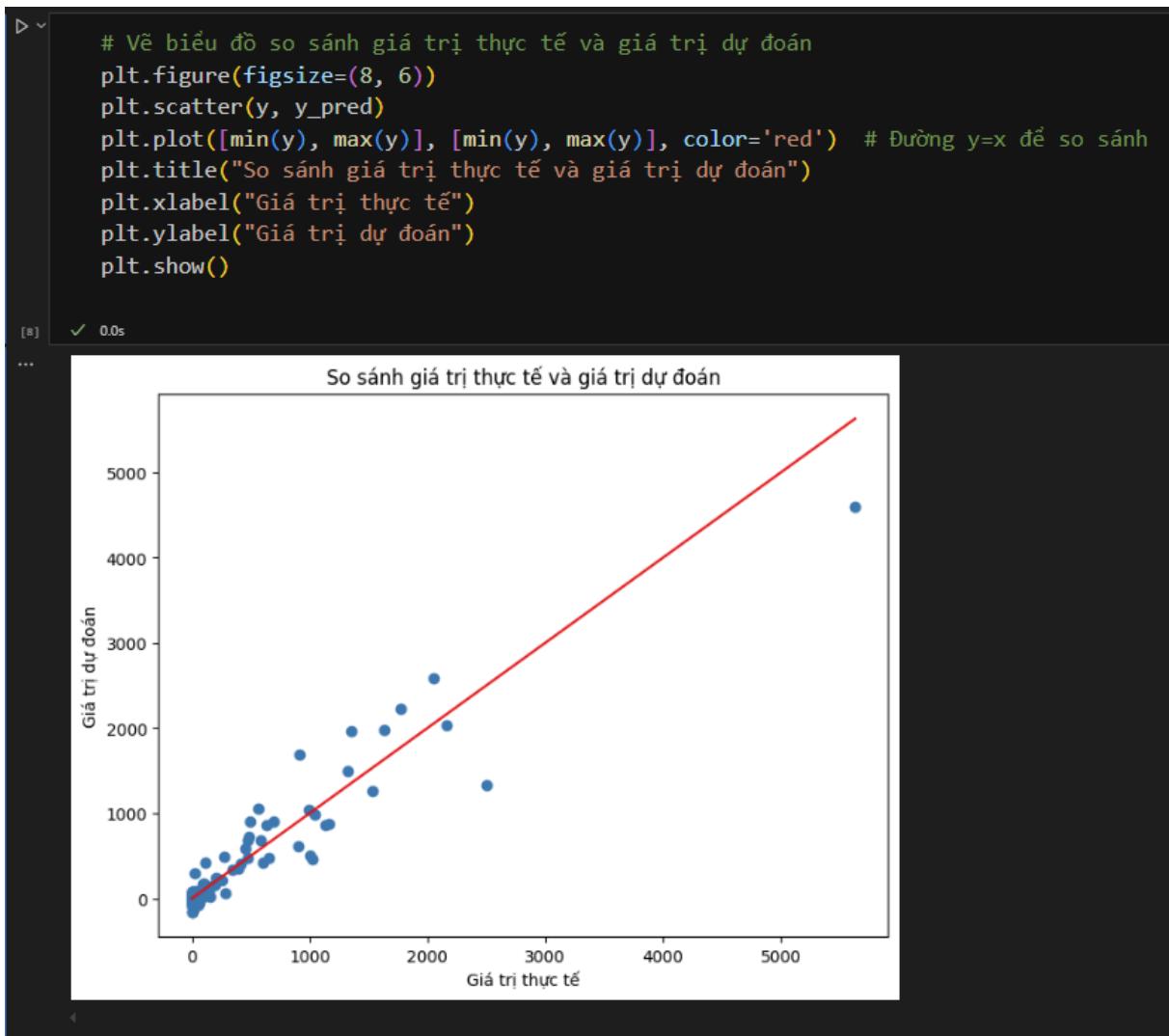
# Đánh giá mô hình
r2 = r2_score(y, y_pred)
print(f"Hệ số xác định (R^2): {r2}")
print("Hệ số hồi quy (Coefficients):", model.coef_)
print("Hệ số chặn (Intercept):", model.intercept_)

[6] ✓ 0.0s
...
    ▾ LinearRegression ⓘ ⓘ
LinearRegression()
```

#### 6.2.3.5 Dự đoán giá trị và Đánh giá mô hình



#### 6.2.3.6 Vẽ biểu đồ phân phối dư lượng



6.2.3.7 Vẽ biểu đồ so sánh giá trị thực tế và giá trị dự đoán

## 6.2.4 Bookmark tổng quan

### 6.2.4.1 Tạo bookmark hiển thị về giới tính



Mục đích: Chuyển sang mục phân tích chi tiết: phân tích theo giới tính

#### 6.2.4.2 Tạo bookmark hiển thị về tuổi



Mục đích : Chuyển sang mục phân tích chi tiết: phân tích theo tuổi

#### 6.2.4.3 Tạo bookmark hiển thị về chiều cao cân nặng



Mục đích: Chuyển sang mục phân tích chi tiết: phân tích theo chiều cao và cân nặng

#### 6.2.4.4 Tạo bookmark hiển thi về môn thể thao



Mục đích: Chuyển sang mục phân tích chi tiết: phân tích theo môn thể thao

#### 6.2.4.5 Tạo bookmark hiển thị về khu vực



Mục đích: chuyển sang mục phân tích chi tiết: phân tích theo khu vực

#### 6.2.4.6 Tạo bookmark xem các biểu đồ khác



Mục đích : Xem các biểu đồ khác:

- Tỷ lệ huy chương giữa nam và nữ
- Tỷ lệ huy chương theo châu lục
- Số lượng huy chương của các môn thể thao theo từng năm

#### 6.2.4.7 Tạo bookmark quay lại giao diện chính



Mục đích: Quay trở về giao diện ban đầu

### 6.2.5 Bookmark phân tích chi tiết

#### 6.2.6 Tạo bookmark quay lại giao diện chính của các mục phân tích chi tiết



Mục đích: Quay trở về giao diện tổng quan

#### 6.2.7 Tạo bookmark xem thêm các phân tích khác



Mục đích: Xem các phân tích khác

### 6.2.8 Tạo bookmark để trở về mục phân tích chi tiết ban đầu



Mục đích: Trở về mục phân tích chi tiết ban đầu

## 7 Kết luận

### 7.1 Báo cáo

#### 7.1.1 Các bước viết báo cáo

**Xác định mục tiêu:** Xác định rõ thông tin cần truyền đạt và đối tượng người đọc.

**Thu thập và phân tích:** Thu thập dữ liệu, sử dụng công cụ phân tích để tìm ra thông tin chi tiết.

**Cấu trúc báo cáo:** Bố cục rõ ràng, logic, thường gồm tóm tắt, giới thiệu, phương pháp, kết quả, thảo luận và kết luận.

**Viết nội dung:** Ngôn ngữ đơn giản, sử dụng biểu đồ, bảng biểu để minh họa.

**Kiểm tra và chỉnh sửa:** Đọc lại, kiểm tra lỗi và đảm bảo tính chính xác.

**Trình bày:** Trình bày gọn gàng, sử dụng công cụ hỗ trợ nếu cần.

#### 7.1.2 Tổng hợp

##### Loại Dữ Liệu và Dữ Liệu Đầu Vào

- Dữ liệu liên quan đến các đặc điểm của vận động viên: giới tính, độ tuổi, môn thể thao, chiều cao, cân nặng, số huy chương, quốc gia tham gia.

- 
- Dữ liệu đầu vào được thu thập từ bảng xếp hạng thi đấu, kết quả thi đấu, khảo sát thể thao và cơ sở dữ liệu quốc tế.

## **Dữ Liệu Đầu Ra và Mục Tiêu Phân Tích**

- Kết quả phân tích bao gồm tỷ lệ huy chương theo giới tính, môn thể thao phổ biến theo độ tuổi, phân phối chiều cao và cân nặng, mối quan hệ giữa chiều cao và huy chương.
- Phân tích số lượng vận động viên theo giới tính và môn thể thao tham gia.

## **Quản Lý và Lưu Trữ Dữ Liệu**

- Quản lý dữ liệu: tổ chức và phân loại rõ ràng thông qua Power bi quan hệ với các tệp CSV
- Lưu trữ dữ liệu: sử dụng Google drive hoặc các giải pháp lưu trữ đám mây để hỗ trợ truy xuất nhanh chóng và hiệu quả.
- Cập nhật định kỳ dữ liệu sau mỗi kỳ thi đấu.

## **Công Nghệ và Lý Do Lựa Chọn**

- Excel : sử dụng để tính toán và làm sạch
- Power BI: công cụ mạnh mẽ giúp trực quan hóa kết quả phân tích, hiển thị tỷ lệ huy chương, phân tích độ tuổi, giới tính, môn thể thao.

## **Kết Quả Dự Kiến**

- Phát hiện mối quan hệ giữa các yếu tố như chiều cao, cân nặng, độ tuổi, giới tính, môn thể thao và thành tích thi đấu.
- Cung cấp cái nhìn sâu sắc để xây dựng chiến lược huấn luyện hiệu quả cho các vận động viên và đội tuyển thể thao.

## 7.2 Khó khăn

**Dữ liệu không đầy đủ và không nhất quán:** Bộ dữ liệu ban đầu có một số trường bị thiếu giá trị hoặc có định dạng không nhất quán, gây khó khăn cho quá trình làm sạch và chuyển đổi dữ liệu.

**Phân chia công việc chưa hợp lý:** Khối lượng công việc lớn nhưng chưa có phân tách rõ ràng khiến cho nhóm đôi lúc bị chậm tiến độ

**Không có chuyên môn về lĩnh vực:** chưa am hiểu sâu về thể thao và các yếu tố liên quan đến thành tích của vận động viên, gây khó khăn trong công việc phân tích dữ liệu và đưa ra các kết luận chính xác.

## 7.3 Thuận lợi

**Dữ liệu có sẵn từ nguồn uy tín :** Bộ dữ liệu Olympic từ Kaggle có tính chất đầy đủ tương đối và được công nhận rộng rãi, cung cấp cơ sở vững chắc cho phân tích.

**Hỗ trợ công cụ và công nghệ :** Sử dụng các công cụ hiện đại như Excel, power bi để giúp tăng hiệu quả trong quá trình xử lý và trực quan hóa dữ liệu hóa học.

**Đã có Kỹ năng và kinh nghiệm làm việc với dữ liệu :** Nhóm đã từng thực hiện về phân tích dữ liệu

**Khả năng học hỏi và thích nghi nhanh :**

- Sẵn sàng học hỏi thêm các kiến thức mới liên quan đến lĩnh vực thể thao để bổ sung cho dự án.
- Tinh thần làm việc linh hoạt, có khả năng xử lý vấn đề hiệu quả khi gặp khó khăn.

## 7.4 Hướng phát triển

**Mở rộng nguồn dữ liệu:**

- Thu thập thêm các dữ liệu chi tiết từ nhiều quốc gia và các kỳ thi đấu lớn khác như Paralympic, Asian Games, World Championships để làm phong phú thêm phân tích.

- Kết hợp với dữ liệu từ các công ty thể thao, các tổ chức thể thao quốc tế để nâng cao độ chính xác và phạm vi phân tích.

### **Ứng dụng AI và Machine Learning:**

- Áp dụng machine learning để dự đoán thành tích của vận động viên dựa trên các yếu tố như tuổi, cân nặng, chiều cao, và môn thể thao.
- Sử dụng các thuật toán học máy để phân tích sự ảnh hưởng của các yếu tố tâm lý, chế độ luyện tập, và dịch bệnh tới thành tích của vận động viên.

### **Tạo ra các mô hình phân tích nâng cao:**

- Phát triển các mô hình phân tích phức tạp để nghiên cứu các yếu tố ảnh hưởng lâu dài như chế độ dinh dưỡng, phương pháp huấn luyện, và môi trường thi đấu đối với hiệu suất vận động viên.
- Tích hợp dữ liệu từ các thiết bị đeo (wearables) để theo dõi và phân tích sức khỏe, tim mạch, và các thông số cơ thể khác của vận động viên.

### **Xây dựng hệ thống dự báo và khuyến nghị:**

- Phát triển các công cụ dự báo có thể chỉ ra những thay đổi tiềm năng trong thành tích của vận động viên dựa trên các yếu tố bên ngoài như thời tiết, lịch thi đấu, hoặc các chấn thương.
- Cung cấp các khuyến nghị cho vận động viên về cách tối ưu hóa huấn luyện và dinh dưỡng dựa trên phân tích dữ liệu thực tế.

### **Phát triển dashboard và báo cáo tự động:**

- Tiếp tục cải tiến và phát triển các dashboard và báo cáo tự động để giúp huấn luyện viên và các nhà phân tích dễ dàng theo dõi hiệu suất của vận động viên qua từng giai đoạn.

- Tích hợp thêm các tính năng tương tác cho phép người dùng tùy chỉnh và tinh chỉnh báo cáo theo nhu cầu.

### **Ứng dụng phân tích đa chiều (Big Data):**

- Áp dụng phân tích dữ liệu lớn để kết hợp nhiều nguồn dữ liệu và tìm kiếm những mối quan hệ tiềm ẩn giữa các yếu tố ảnh hưởng đến thành tích của vận động viên.
- Phát triển nền tảng phân tích dữ liệu lớn có thể xử lý hàng triệu dữ liệu từ các kỳ thi đấu và vận động viên.

## 7.5 Tổng kết

- Quá trình làm dự án: 29/12/2024 - 10/12/2024
- Đã hoàn thành 95% Requirement đề ra
- Dự án đã phân tích và đánh giá các yếu tố quan trọng liên quan đến thành tích của vận động viên, từ đó cung cấp cái nhìn tổng quan và dữ liệu hữu ích để hỗ trợ các chiến lược nâng cao hiệu suất thi đấu. Các kết quả chính từ dự án bao gồm:

### Xác định các yếu tố quan trọng:

- Chiều cao và cân nặng: Là các yếu tố có tác động đáng kể đến thành tích, đặc biệt khi so sánh giữa các môn thể thao khác nhau.
- Giới tính: Tỷ lệ nam và nữ trong các môn thể thao và thành tích đạt được cho thấy sự chênh lệch nhất định.
- Nhóm tuổi: Hiệu suất thi đấu có sự khác biệt theo nhóm tuổi, phản ánh sự thay đổi thể chất và kinh nghiệm thi đấu.
- Quốc gia và khu vực: Sự phân bố vận động viên và thành tích có liên quan đến đặc trưng văn hóa, điều kiện địa lý, và tiềm lực thể thao của từng quốc gia.
- Môn thể thao: Môn thể thao cũng ảnh hưởng khá lớn đến thành tích vì thể thao có rất nhiều thể loại, và mỗi loại đều có sự phổ biến khác nhau

### Phân tích dữ liệu theo yếu tố thể chất và thành tích:

- Mỗi quan hệ giữa chiều cao và cân nặng với các loại huy chương (vàng, bạc, đồng) cho thấy sự khác biệt về yêu cầu thể chất ở từng môn thể thao.
- Các môn như Endurance Sports và Winter Sports có yêu cầu đặc thù về thể trạng và chiếm tỷ lệ huy chương cao nhất.
- Tỷ lệ vận động viên theo nhóm tuổi: Trẻ hơn không đồng nghĩa với thành tích cao hơn, nhưng kinh nghiệm đóng vai trò quan trọng với các vận động viên lớn tuổi.

### Đánh giá tác động:

- Các yếu tố như giới tính, thể chất, và môi trường thi đấu tác động mạnh mẽ đến cơ hội đạt huy chương, cung cấp dữ liệu hữu ích cho huấn luyện viên và nhà phân tích.
- Sự khác biệt về thành tích giữa các khu vực giúp nhận diện những quốc gia có tiềm năng phát triển môn thể thao cụ thể.
- Việc quản lý và phân tích dữ liệu đã hỗ trợ xây dựng chiến lược tối ưu hóa hiệu suất thi đấu cho vận động viên.

### **Khuyến nghị giải pháp:**

- Cải thiện chiến lược huấn luyện: Cá nhân hóa chương trình tập luyện dựa trên đặc điểm chiều cao, cân nặng, và môn thể thao.
- Khuyến khích nghiên cứu chuyên sâu: Tiếp tục thu thập dữ liệu từ các kỳ thi đấu quốc tế, kết hợp với dữ liệu từ thiết bị theo dõi vận động viên.
- Đầu tư vào phát triển thể thao địa phương: Dựa trên các phân tích về khu vực và nhóm môn thể thao, tập trung vào thế mạnh từng quốc gia để phát triển lâu dài.

**Trello**



**Slide**



**Dashboard**

