Projeto Desempenho Esportivo

Consultores Responsáveis:
Renan de Andrade Marques

Requerente:
João Vitor Neves

Brasília, 12 de novembro de 2024.

ESTAT
consultoria estatística

Sumário

1 Introdução

Este projeto tem como objetivo compreender melhor os fatores que afetam a performance dos atletas. Para isso, foi feita uma análise estatistica descritiva acerca dos atletas que participaram das olimpíadas dos anos de 2000 até 2016. Foi estudado: O top 5 países com maior número de mulheres medalhistas, o comportamento do índice de massa corporal (IMC) nos esportes selecionados, o top 3 medalhistas gerais e a relação entre o peso e a altura dos atletas.

O banco de dados foi disponibilizado pelo cliente. Foi observado Nome, Sexo, Idade, País, Peso, Altura, Esporte, Modalidade e Medalha adquirida de 38366 atletas diferentes. Desses atletas, as análises desse relatório consideraram apenas os atletas medalhistas: 7294 atletas diferentes no total.

A manipulação e análise dos dados, além da confecção das figuras, foram feitas com auxilio do software estatístico RStudio versão 2024.04.2. O pdf do relatório foi gerado utilizando o Quarto.

2 Referencial Teórico

2.1 Frequência Relativa

A frequência relativa é utilizada para a comparação entre classes de uma variável categórica com c categorias, ou para comparar uma mesma categoria em diferentes estudos.

A frequência relativa da categoria j é dada por:

$$f_j = \frac{n_j}{n}$$

Com:

- $j = 1, ..., c$
- $n_j =$ número de observações da categoria j
- $n =$ número total de observações

Geralmente, a frequência relativa é utilizada em porcentagem, dada por:

$$100 \times f_j$$

2.2 Média

A média é a soma das observações dividida pelo número total delas, dada pela fórmula:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Com:

- $i = 1, 2, ..., n$
- $n =$ número total de observações

2.3 Mediana

Sejam as n observações de um conjunto de dados $X = X_{(1)}, X_{(2)}, \dots, X_{(n)}$ de determinada variável ordenadas de forma crescente. A mediana do conjunto de dados X é o valor que deixa metade das observações abaixo dela e metade dos dados acima.

Com isso, pode-se calcular a mediana da seguinte forma:

$$med(X) = \begin{cases} X_{\frac{n+1}{2}}, \text{para n ímpar} \\ \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}, \text{para n par} \end{cases}$$

2.4 Quartis

Os quartis são separatrizes que dividem o conjunto de dados em quatro partes iguais. O primeiro quartil (ou inferior) delimita os 25% menores valores, o segundo representa a mediana, e o terceiro delimita os 25% maiores valores. Inicialmente deve-se calcular a posição do quartil:

- Posição do primeiro quartil P_1:

$$P_1 = \frac{n+1}{4}$$

- Posição da mediana (segundo quartil) P_2:

$$P_2 = \frac{n+1}{2}$$

- Posição do terceiro quartil P_3:

$$P_3 = \frac{3 \times (n+1)}{4}$$

Com n sendo o tamanho da amostra. Dessa forma, $X_{(P_i)}$ é o valor do i-ésimo quartil, onde $X_{(j)}$ representa a j-ésima observação dos dados ordenados.

Se o cálculo da posição resultar em uma fração, deve-se fazer a média entre o valor que está na posição do inteiro anterior e do seguinte ao da posição.

2.5 Variância

A variância é uma medida que avalia o quanto os dados estão dispersos em relação à média, em uma escala ao quadrado da escala dos dados.

2.5.1 Variância Populacional

Para uma população, a variância é dada por:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

Com:

- $X_i = i$-ésima observação da população
- $\mu =$ média populacional
- $N =$ tamanho da população

2.5.2 Variância Amostral

Para uma amostra, a variância é dada por:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}$$

Com:

- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral
- $n =$ tamanho da amostra

2.6 Desvio Padrão

O desvio padrão é a raiz quadrada da variância. Ele avalia o quanto os dados estão dispersos em relação à média.

2.6.1 Desvio Padrão Populacional

Para uma população, o desvio padrão é dado por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \mu\right)^2}{N}}$$

Com:

- $X_i =$ i-ésima observação da população
- $\mu =$ média populacional
- $N =$ tamanho da população

2.6.2 Desvio Padrão Amostral

Para uma amostra, o desvio padrão é dado por:

$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}}$$

Com:

- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral
- $n =$ tamanho da amostra

2.7 Coeficiente de Variação

O coeficiente de variação fornece a dispersão dos dados em relação à média. Quanto menor for o seu valor, mais homogêneos serão os dados. O coeficiente de variação é considerado baixo (apontando um conjunto de dados homogêneo) quando for menor ou igual a 25%. Ele é dado pela fórmula:

$$C_V = \frac{S}{\bar{X}} \times 100$$

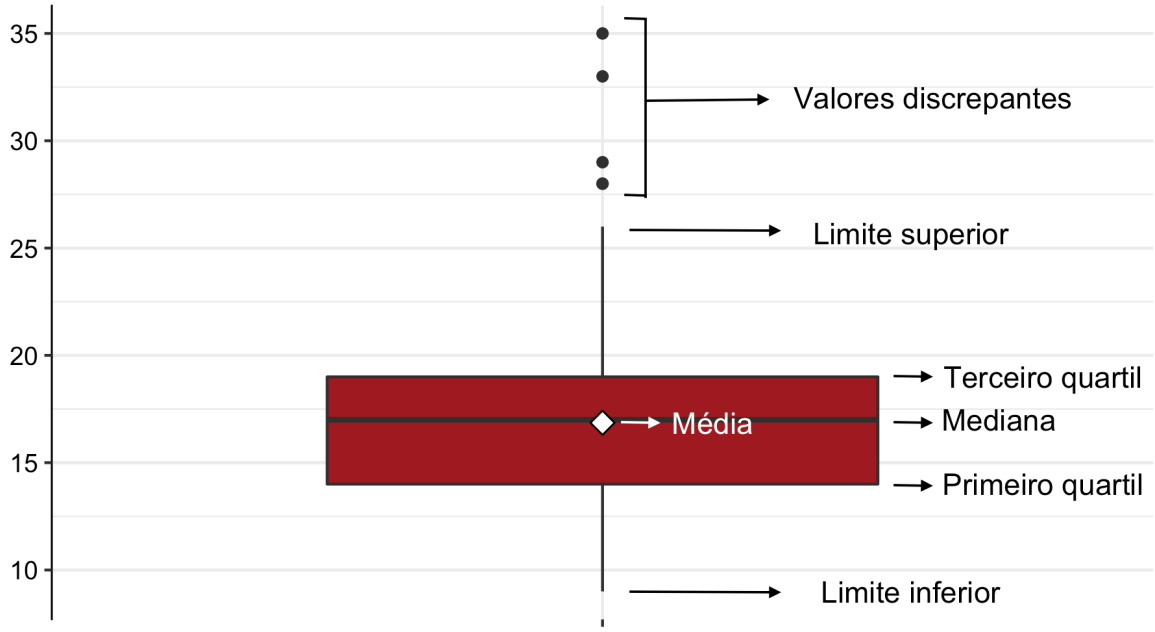
Com:

- $S =$ desvio padrão amostral
- $\bar{X} =$ média amostral

2.8 Boxplot

O boxplot é uma representação gráfica na qual se pode perceber de forma mais clara como os dados estão distribuídos. A figura abaixo ilustra um exemplo de boxplot.

Figura 1: Exemplo de boxplot

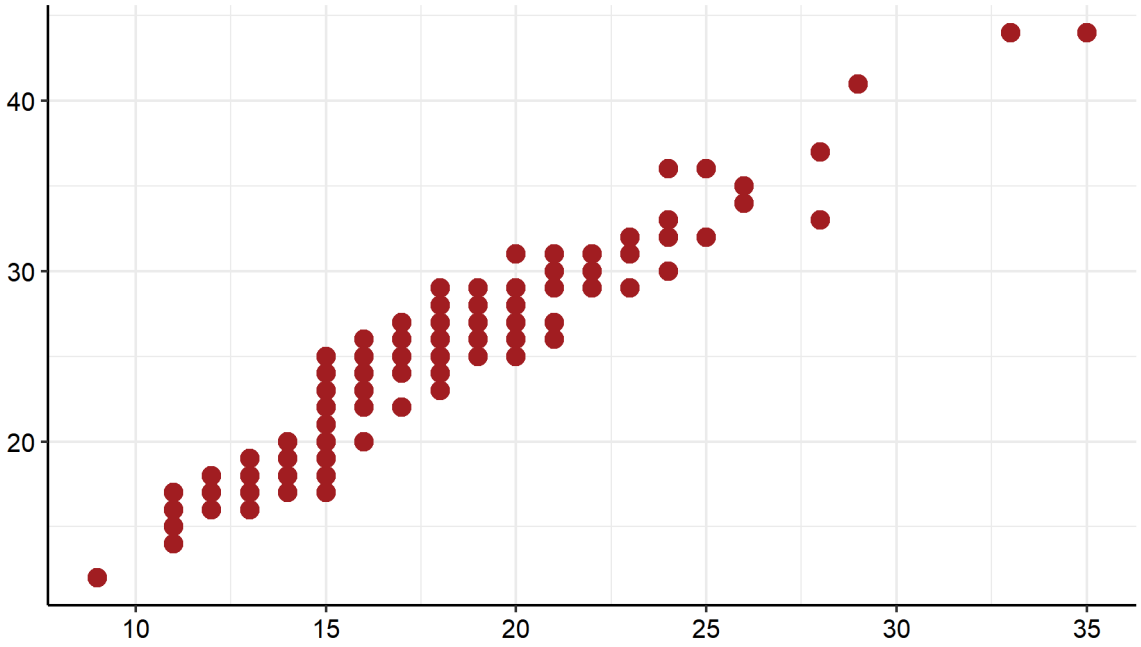


A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada pelo losango branco e os pontos são *outliers*. Os *outliers* são valores discrepantes da série de dados, ou seja, valores que não demonstram a realidade de um conjunto de dados.

2.9 Gráfico de Dispersão

O gráfico de dispersão é uma representação gráfica utilizada para ilustrar o comportamento conjunto de duas variáveis quantitativas. A figura abaixo ilustra um exemplo de gráfico de dispersão, onde cada ponto representa uma observação do banco de dados.

Figura 2: Exemplo de Gráfico de Dispersão



2.10 Tipos de Variáveis

2.10.1 Qualitativas

As variáveis qualitativas são as variáveis não numéricas, que representam categorias ou características da população. Estas subdividem-se em:

- **Nominais**: quando não existe uma ordem entre as categorias da variável (exemplos: sexo, cor dos olhos, fumante ou não, etc)
- **Ordinais**: quando existe uma ordem entre as categorias da variável (exemplos: nível de escolaridade, mês, estágio de doença, etc)

2.10.2 Quantitativas

As variáveis quantitativas são as variáveis numéricas, que representam características numéricas da população, ou seja, quantidades. Estas subdividem-se em:

- **Discretas**: quando os possíveis valores são enumeráveis (exemplos: número de filhos, número de cigarros fumados, etc)
- **Contínuas**: quando os possíveis valores são resultado de medições (exemplos: massa, altura, tempo, etc)

2.11 Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida que verifica o grau de relação linear entre duas variáveis quantitativas. Este coeficiente varia entre os valores -1 e 1. O valor zero significa que não há relação linear entre as variáveis. Quando o valor do coeficiente r é negativo, diz-se existir uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando r é positivo, diz-se que as duas variáveis são diretamente proporcionais.

O coeficiente de correlação de Pearson é normalmente representado pela letra r e a sua fórmula de cálculo é:

$$r_{Pearson} = \frac{\sum_{i=1}^{n} \left[(x_i - \bar{x}) \, (y_i - \bar{y}) \right]}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \times \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

Onde:

- $x_i =$ i-ésimo valor da variável X
- $y_i =$ i-ésimo valor da variável Y
- $\bar{x} =$ média dos valores da variável X
- $\bar{y} =$ média dos valores da variável Y

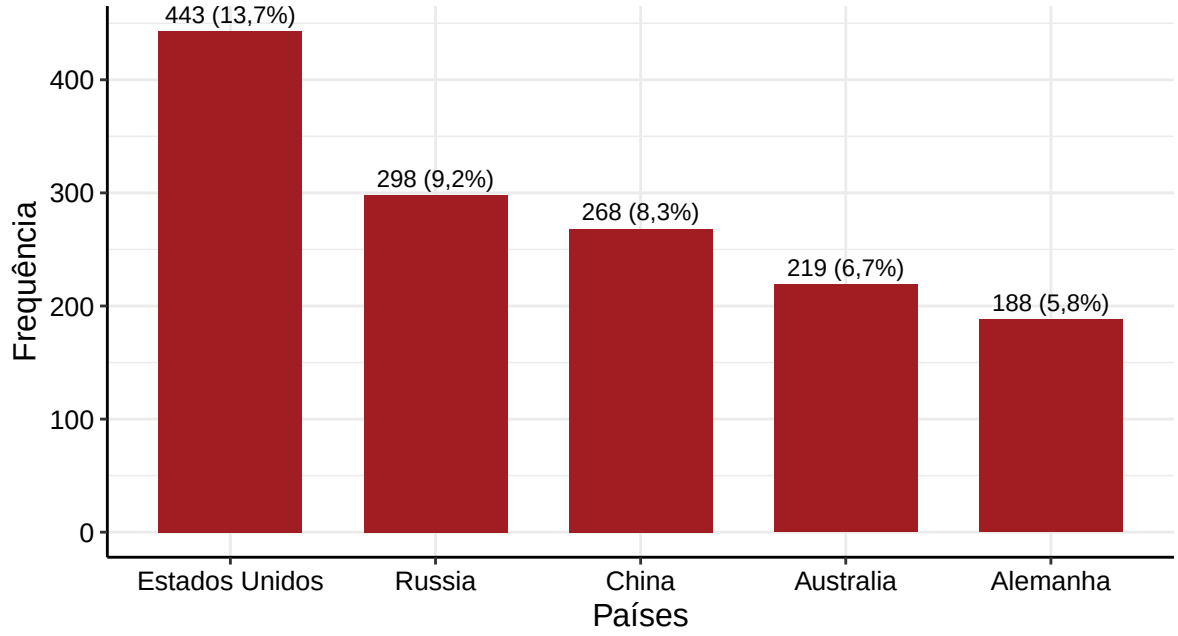
Vale ressaltar que o coeficiente de Pearson é paramétrico e, portanto, sensível quanto à normalidade (simetria) dos dados.

3 Análises

3.1 Top 5 países com maior número de mulheres medalistas

Está análise tem o intuito de identificar quais são os países com maior quantidade de mulheres medalhistas. Para isso foram utilizadas as variáveis Sexo e Medalhas, a primeira sendo qualitativa nominal e a segunda sendo qualitativa ordinal. Para ilustrar a análise sera utilizado um gráfico de barras.

Figura 1: Gráfico de colunas do número de mulheres medalistas

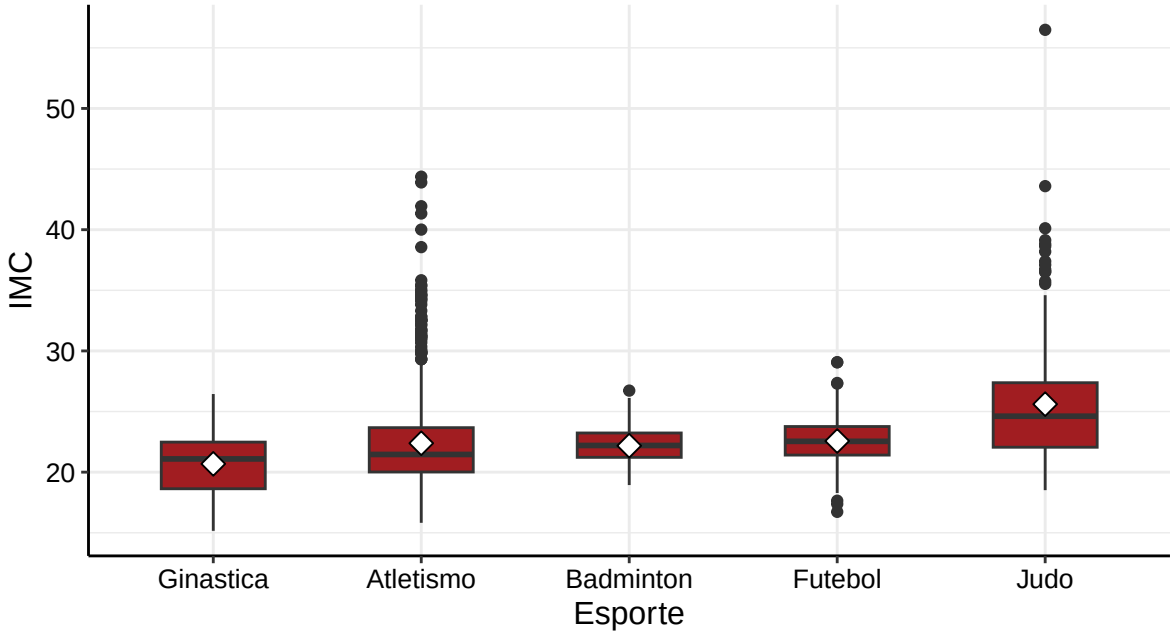


Como pode ser observado na figura 1, o ranque é formado por Estados unidos, Rússia, China, Australia e Alemanha respectivamente. Os Estados Unidos ocupam o primeiro lugar do ranque com mais de 100 mulheres medalistas de diferença do segundo colocado, a Rússia, e com mais que o dobro que a Australia e Alemanha, o quarto e o quinto colocado na devida ordem. Juntos esses 5 países sozinhos tem 43,7% das mulheres medalistas.

3.2 IMC por esportes

Essa análise tem como objetivo entender o comportamento do IMC nos esportes selecionados. Para tal, foram utilizadas as variáveis esporte e IMC, respectivamente qualitativa nominal e quantitativa continua. A variável IMC foi obtida dividindo o Peso do atleta pela Altura ao quadrado. O valor do índice de massa corporal (IMC) é um importante indicador da saúde de uma pessoa, o número representa o quanto a pessoa tem de massa muscular + massa de gordura + massa óssea. Para ilustrar as análises, serão utilizados um gráfico boxplot e um quadro de medidas resumo.

Figura 2: Boxplot do IMC pelo Esporte



Quadro 1: Medidas resumo do IMC

Estatística	Atletismo	Badminton	Futebol	Ginastica	Judo
Média	22.38	22.18	22.57	20.68	25.61
Desvio Padrão	4.01	1.59	1.77	2.42	5.05
Variância	16.10	2.52	3.12	5.86	25.50
Mínimo	15.82	18.93	16.72	15.16	18.51
1º Quartil	20.01	21.22	21.41	18.63	22.05
Mediana	21.46	22.20	22.54	21.09	24.61
3º Quartil	23.67	23.23	23.76	22.48	27.37
Máximo	44.37	26.73	29.06	26.44	56.49

Pode se observar pela figura 2 e o quadro 1 que o IMC segue um comportamento diferente para cada esporte. No judô há diversas categorias para pessoas com pesos diferentes e é um esporte que exige maior massa muscular, o que explica as medidas de variabilidade e as medidas de centralidade elevadas se comparadas com as dos outros esportes.

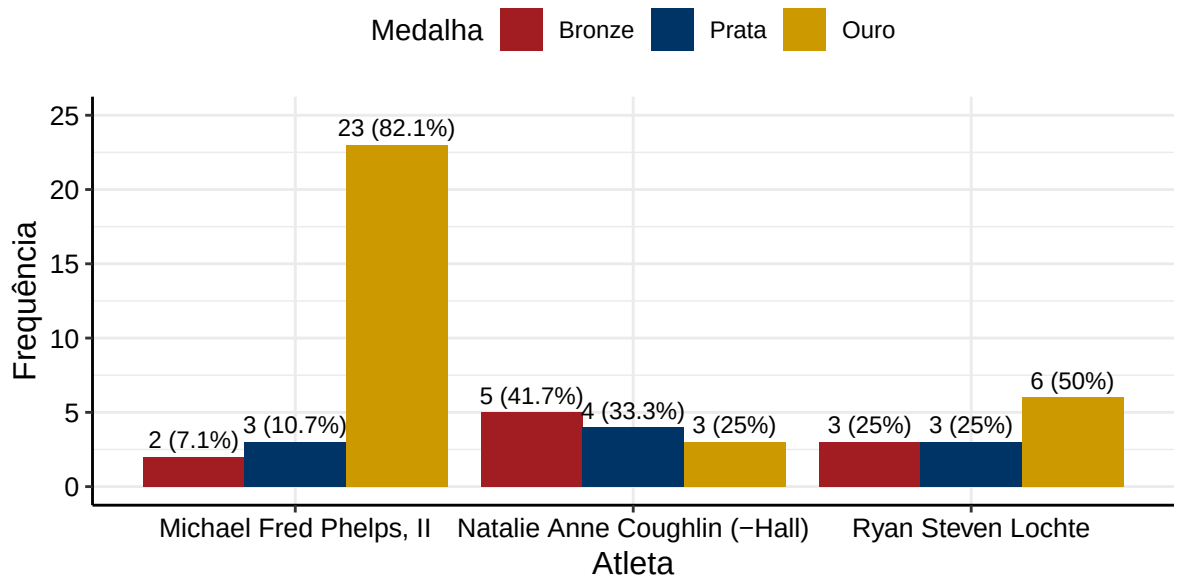
A média, medida de centralidade, no Badminton, na Ginastica, no futebol e no Atletismo são semelhantes: 21.97, 21.51, 22.72 e 22.44, respectivamente, entretanto cada um tem uma configuração única. No atletismo há uma assimetria positiva enquanto na ginastica há uma assimetria negativa, ou seja, no atletismo há uma maior concentração na parte inferior da amostra e na ginastica há o contrário. Além disso, o atletismo tem o segundo maior desvio padrão (4.09), medida de variabilidade, que se deve ao fato dele possuir modalidades com diferentes exigências de massa muscular, desde salto com vara até arremeso de peso.

No badminton e no futebol as medianas e as médias, medidas de centralidade, juntas e próximas do centro da caixa além dos baixos valores dos desvios padrões (1.34 e 1.69 respectivamente), medida variabilidade, mostram que esses esportes tem uma distribuição simétrica e centralizada. Em geral, quanto mais um esporte ou modalidade exige massa muscular, maior o IMC.

3.3 Top 3 medalhistas gerais

Esta análise tem como objetivo observar quais são os 3 maiores medalhistas e verificar se há relação entre o medalhista e o tipo de medalha conquistada. Para isso, foram utilizadas as variáveis Nome e Medalha, a primeira sendo qualitativa nominal e a segunda sendo qualitativa ordinal. Para ilustrar a análise, foi utilizado um gráfico de barras.

Figura 3: Gráfico de colunas da quantidade de medalhas pelo tipo da medalha

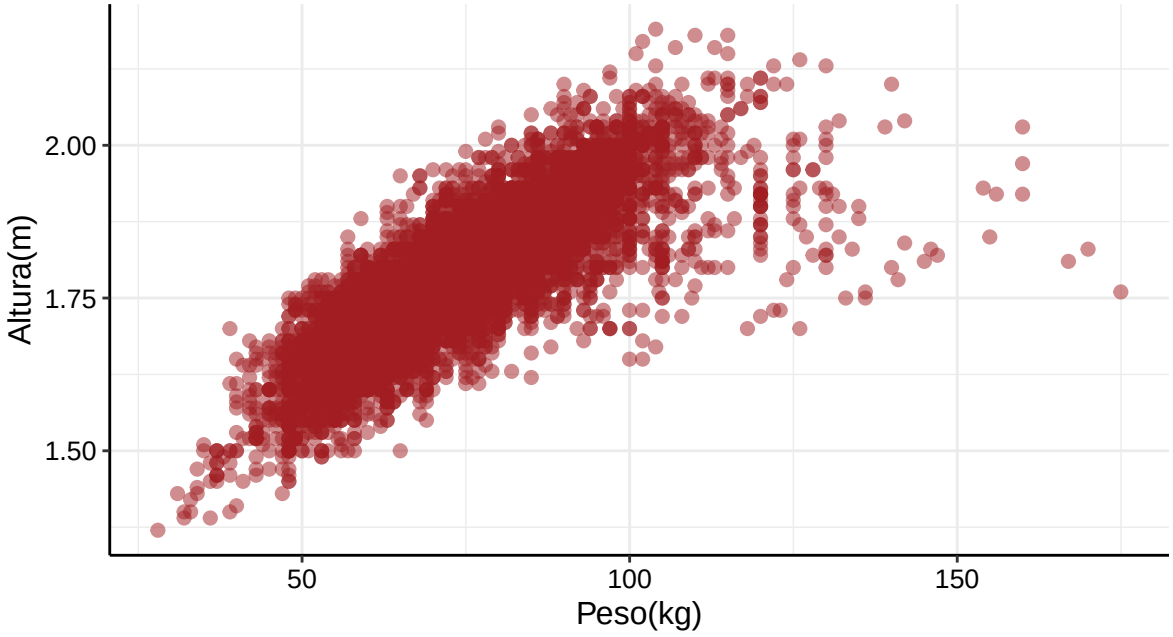


Observa-se pelo gráfico a relação entre o atleta e o tipo de medalha. Os três atletas que conquistaram mais medalhas nessas 5 edições dos jogos olímpicos foram: Michael Fred Phelps com 28 medalhas, Natalie Anne Coughlin e Ryan Steven Lochte ambos com 12 medalhas. Os três são estadunidenses e tem como esporte a natação, que possui diversas modalidades e , consequentemente, várias oportunidades de se obter uma medalha. Conforme o valor da medalha aumentou, Natalie conquistou menos medalhas, enquanto para Michael e Ryan o comportamento foi o contrário. Também é possivel observar que a grande maioria das medalhas que Michael Fred Phelps conquistou foram medalhas de ouro, demostrando grande desempenho esportivo do atleta.

3.4 Relação peso x altura

Esta análise tem o intuito de compreender a relação entre o peso e a altura dos atletas. Para isso, foram utilizadas as variáveis Peso(Kg) e Altura(m), ambas são quantitativas contínuas. O comportamento conjunto das variáveis está ilustrado pelo gráfico de dispersão a seguir.

Figura 4: Gráfico de dispersão do peso pela altura do atleta



Quadro 2: Medidas resumo da variável peso e altura do atleta

Estatística	Peso	Altura
Média	74.14	1.78
Desvio Padrão	16.25	0.12
Variância	264.05	0.01
Mínimo	28	1.37
1º Quartil	62.99	1.7
Mediana	71.99	1.78
3º Quartil	83.99	1.86
Máximo	174.97	2.19

Observa-se pela figura 4 e pelo quadro 2 o comportamento das duas variáveis. Os atletas têm, em geral, peso entre 63 e 84 quilos e altura entre 1.70 e 1.86 metros. Além disso o peso dos atletas varia mais que a altura (coeficiente de variação 21% e 6% respectivamente). O coeficiente de Pearson, que mostra a força e o sentido da associação de duas variaveis quantitativas e varia de -1 a 1, assumiu o valor 0.79, ou seja, observa se pelo gráfico e pelo coeficiente uma relação forte e positiva entre as variáveis, conforme a altura aumenta o peso tende a aumentar.

4 Conclusões

Os resultados das análises evidenciam aspectos importantes sobre os atletas. Observou-se que Estados unidos, Rússia, China, Australia e Alemanha são os países com maior número de mulheres medalhistas, com os Estados Unidos bem a frente dos demais. Em paralelo, constatou-se que os três atletas que conquistaram mais medalhas nessas 5 edições dos jogos olímpicos, Michael Fred Phelps, Natalie Anne Coughlin e Ryan Steven Lochte são estadunidenses e tem como esporte a natação, que possui diversas modalidades e , consequentemente, várias oportunidades de se obter uma medalha.. Dentre os três, destaca-se Michael Phelps com 28 medalhas das quais 23 são de ouro, quase a mesma quantidade de todas as medalhas de Natalie e Ryan somadas.

Além disso, pode se observar que, em média, os atletas que praticam ginástica, futebol, judô, atletismo e badminton têm IMC igual a 22, que indica eutrofia. As análises mostraram que o IMC dos atletas tem grande relação com o esporte que eles praticam. Em geral, quanto mais um esporte ou modalidade exige massa muscular, maior é o IMC. Os atletas têm, em geral, peso entre 63 e 84 quilos e altura entre 1.70 e 1.86 metros. Também se constatou que o peso e a altura dos atletas tem uma relação forte e positiva, ou seja, conforme a altura aumenta o peso tende a aumentar.