Projeto Fantasma

Consultores Responsáveis:
Antônio Olímpio

Requerente:
João Vitor Neves.

Brasília, 10 de novembro de 2024.

ESTAT
consultoria estatística

Sumário

1 Introdução

O seguinte projeto visa auxiliar no projeto que visa optimizar o desempenho dos atletas da academia House of Excellence, a partir de análises estatísticas utilizando dados das olimpíadas dos anos 2000 até o ano de 2016. Foi utilizado para esse estudo o banco de dados que contem a informação de todos os atletas, as medidas deles, se eles foram medalhistas e qual modalidade eles competiram. Este estudo faz uso de variáveis qualitativas nominais e ordinais e variáveis quantitativas discretas e contínuas para a realização das análises descritivas apresentadas. Todos os dados foram disponibilizados pelo cliente.

Primeiramente, serão apresentados os cinco países com a maior participação feminina no pódio entre todas as modalidades, o que apresenta onde o esporte feminino foi mais desenvolvido. Em seguida, buscou-se entender a diferença no IMC (Índice de Massa Corporal) entre algumas das modalidades.

A terceira análise apresenta os três maiores medalhistas da época, e também se existe alguma relação entre o atleta e as medalhas conquistadas. Por último, será compreendido se há alguma relação entre a altura dos atletas e o peso registrado.

A ferramenta utilizada para a produção do projeto foi o R versão 4.4.0 no software RStudio 2024.04.1+748 e o relatório foi produzido utilizando o software Quarto, utilizado para integrar códigos a relatórios de forma interativa, na versão 1.4.553.

2 Referencial Teórico

2.1 Análise Descritiva Univariada

2.1.1 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:

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

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

$$100 \times f_j$$

2.1.2 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:

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

2.1.3 Mediana

Sejam as n observações de um conjunto de dados $X = X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ 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:

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

2.1.4 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.

2.1.5 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.

2.1.5.1 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

2.1.5.2 Variância Amostral

Para uma amostra, a variância é dada por:

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

Com:

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

2.1.6 Desvio Padrão

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

2.1.6.1 Desvio Padrão Populacional

Para uma população, o desvio padrão é dado por:

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

Com:

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

2.1.6.2 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

2.1.7 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

2.1.8 Coeficiente de Assimetria

O coeficiente de assimetria quantifica a simetria dos dados. Um valor positivo indica que os dados estão concentrados à esquerda em sua função de distribuição, enquanto um valor negativo indica maior concentração à direita. A fórmula é:

$$C_{Assimetria} = \frac{1}{n} \times \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S}\right)^3$$

Com:

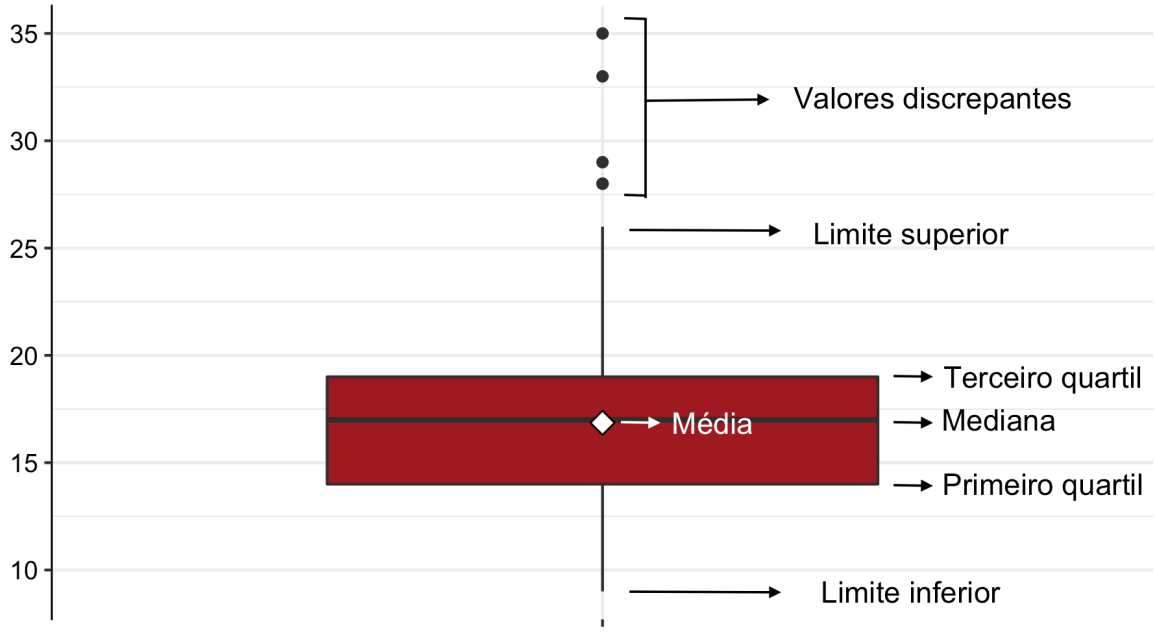
- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral

- $S =$ desvio padrão amostral
- $n =$ tamanho da amostra

2.1.9 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

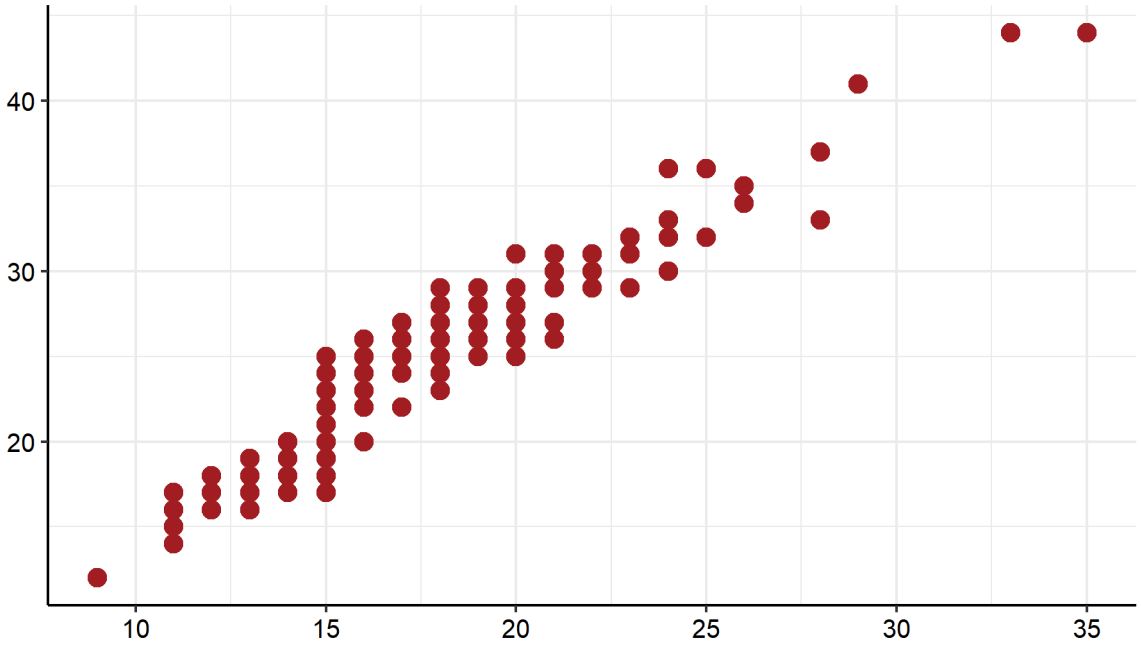


A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada 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.

2.1.10 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



2.1.11 Tipos de Variáveis

2.1.11.1 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)

2.1.11.2 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)

2.1.12 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 Análise de top 5 países com mais mulheres medalhistas

A primeira análise requisitada envolve buscar quais os países que obtiveram a maior quantia de medalhas na modalidade feminina dos jogos olímpicos. Esses dados podem ser utilizados para encontrar de forma mais eficiente o diferencial nos países com mais medalhas dos que apresentam uma menor quantidade.

Nesta análise, foram utilizadas, respectivamente, as variáveis qualitativas nominal e ordinal medalha, que indica o tipo de medalha que o time conquistou, e time, que indica o país de origem do atleta. Para essa análise foi necessário a utilização de gráficos de barras uni e bivariaveis, além de uma tabela.

Gráfico de colunas univariado da quantidade de medalhistas em cada equipe olímpica:

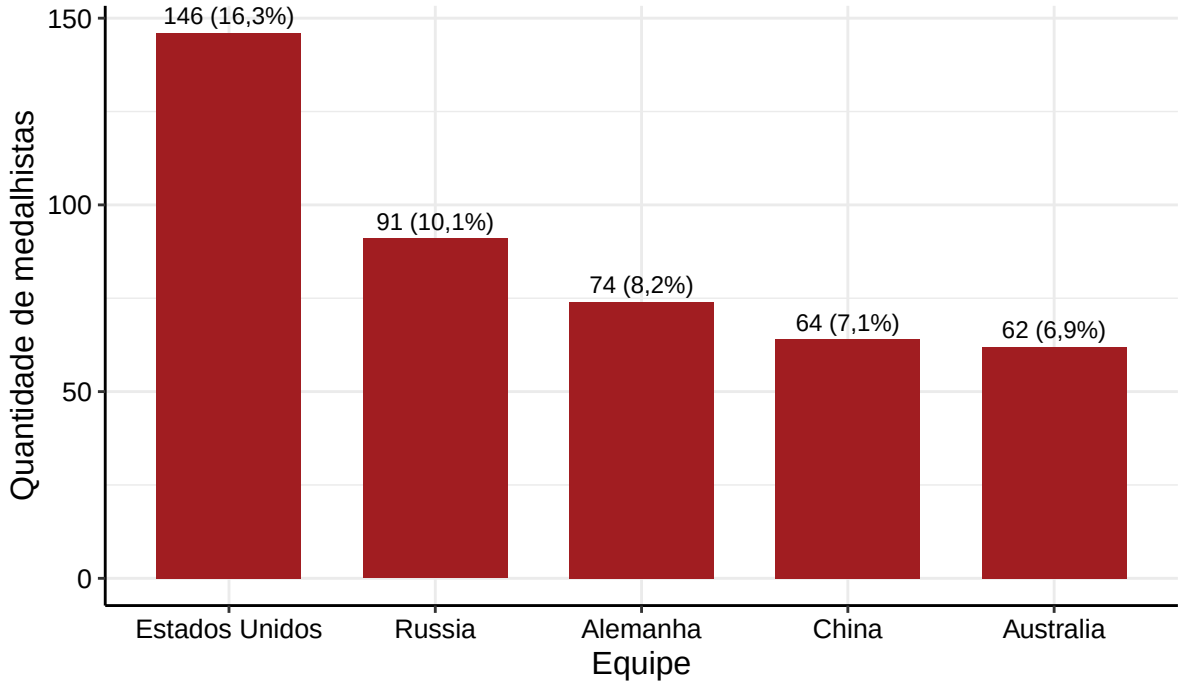


Gráfico de colunas bivariado da quantidade de cada tipo de medalha

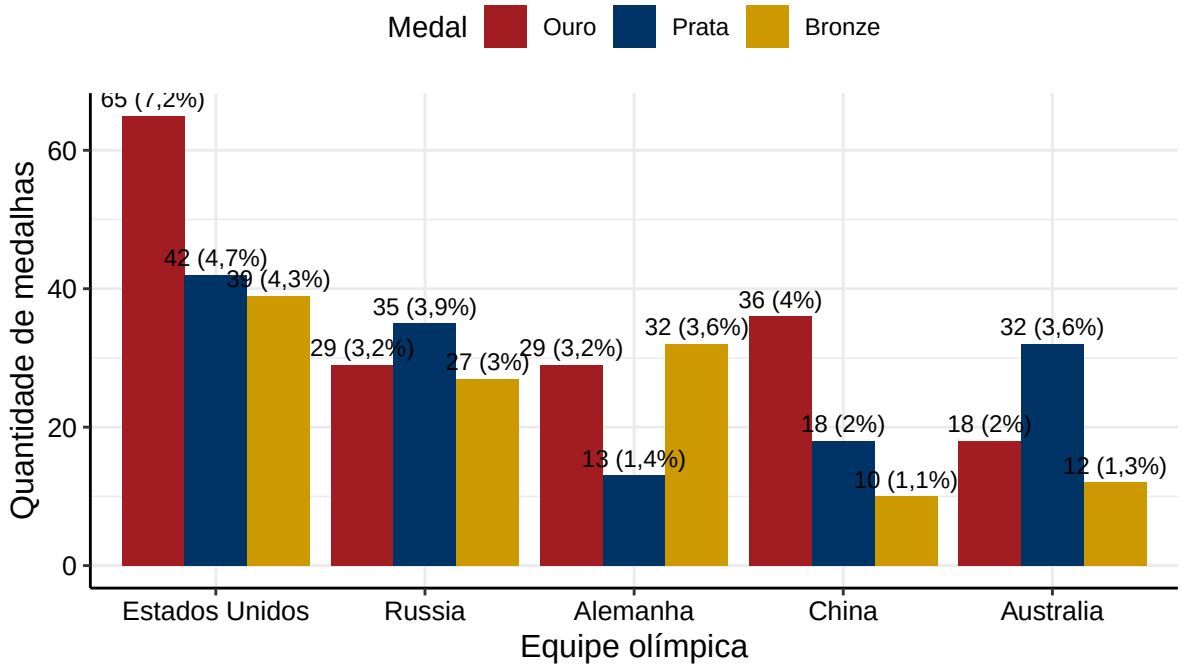


Tabela das medidas das medalhas gerais

Medidas	Estados Unidos	Rússia	Alemanha	China	Austrália	Somatorio	Total no banco
Medalhas Totais	146	91	74	64	62	437	898
Frequência Relativa	16,3%	10,1%	8,2%	7,1%	6,9%	48,6%	100%

O primeiro gráfico apresenta os 5 países com a maior quantia de medalhas dentro do banco de dados, esses sendo, em ordem: Estados Unidos, Rússia, Alemanha, China e Austrália. Após a análise dos gráficos, pode-se perceber que os Estados Unidos foram o país com a maior quantidade de mulheres medalhistas, enquanto, dentro do ranking, a Austrália foi a que obteve a menor quantidade, embora a China esteja muito próxima. No gráfico, é possível notar, também, que a diferença de medalhas entre os Estados Unidos e a Austrália é de 84 medalhas, enquanto dos Estados Unidos para a Rússia, segunda maior em medalhas, é de 55. Podem indicar uma tendência das atletas dos Estados Unidos performarem melhor. Além disso, pode-se perceber que os cinco países apresentados compõem 48,6% das medalhas do banco de dados, o que é quase metade, e dentro desse valor, os Estados Unidos compuseram 16,3% das medalhas, uma quantidade elevada.

O segundo gráfico apresenta a quantidade de cada medalha conquistada pelos países, e pode-se ver que embora a China esteja na quarta posição, ela tem a segunda maior quantia de medalhas de ouro, assim como pode ser visto que a Austrália tem

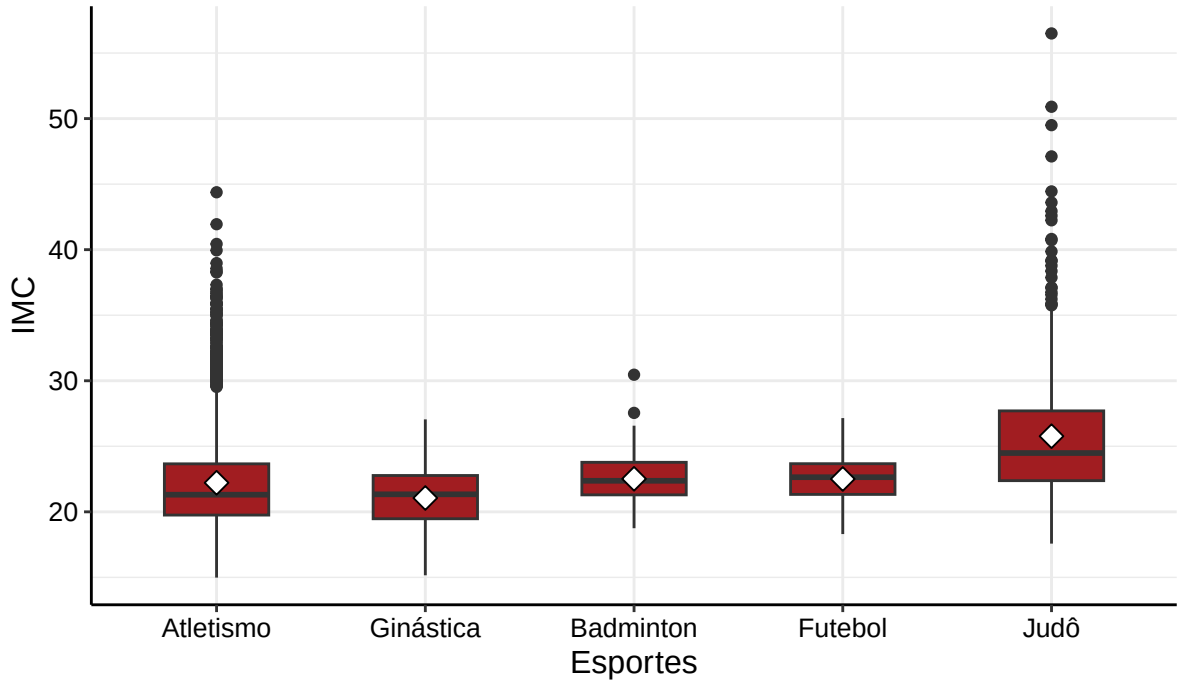
a terceira maior quantia de medalhas de prata, logo atrás da Rússia, que fica atrás apenas dos Estados Unidos e da Alemanha nas medalhas de bronze.

3.2 Análise da diferença de IMC entre esportes

Essa análise busca entender a diferença de IMC (Índice de Massa Corporal), uma medida utilizada para identificar se o peso de algum indivíduo está acima ou abaixo do saudável, entre alguns esportes, sendo eles: Atletismo, Badmínton, Ginástica, Judô e Futebol. Essa análise pode ser utilizada para descobrir qual o IMC ideal, ou pelo menos o mais popular em cada esporte estudado.

Para essa análise, foram utilizadas as variáveis quantitativas contínuas peso e altura, que formaram a também quantitativa contínua IMC. Além disso, foi utilizada a variável qualitativa nominal esporte, essa que indica o esporte que o atleta competiu. Para a realização dessa análise, foi necessário ajustar o peso de libras para quilogramas e a altura de pés para metros, além de ajustar o banco de dados para que cada atleta fosse representado apenas uma vez, além de utilizar o gráfico boxplot e quadros de medidas.

Figura 3: Bloxplot Esportes



Quadro de medidas dos esportes

Medidas	Atletismo	Ginástica	Badminton	Futebol	Judô
Média	22,25	20,84	22,50	22,51	25,77

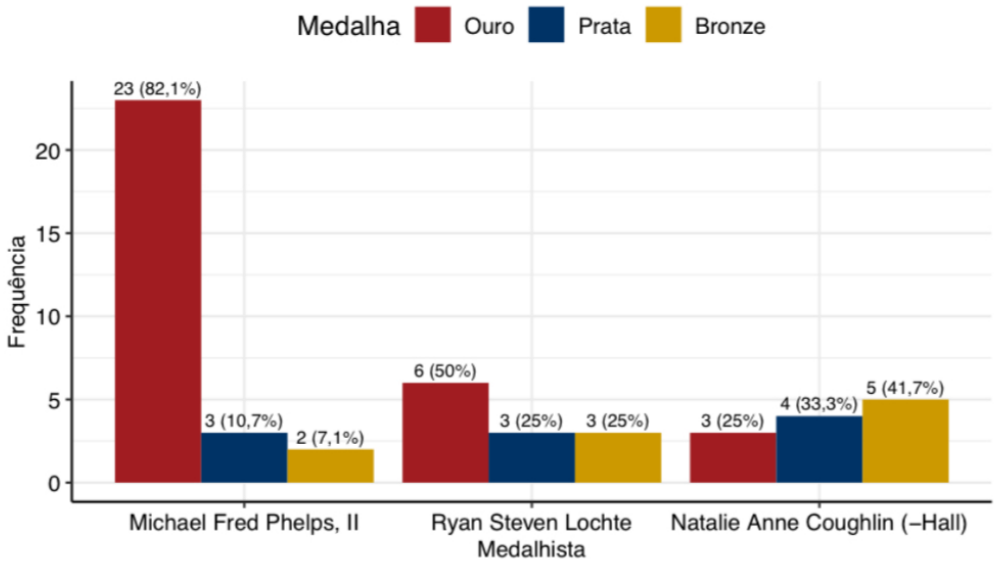
Medidas	Atletismo	Ginástica	Badminton	Futebol	Judô
Desvio Padrão	3,92	2,42	1,82	1,63	5.47
Variância	15,38	5,86	3,33	2,67	29,91
Mínimo	14,98	15.16	18,75	18,31	17,58
1° Quartil	18,70	18,87	21,22	21,33	22,38
Mediana	21,30	21,10	22,33	22,64	24,49
3° Quartil	23,67	22,65	23,77	23,67	23,67
Máximo	44,38	27,05	30.46	27,15	56,50

Analisando o gráfico e o quadro de medidas, pode-se notar que as médias são muito próximas uma da outra, principalmente as de badmínton, atletismo e futebol, que se encontram ao redor de 22. Além disso, pode-se ver que atletismo e judô possuem a maior quantidade de outliers, valores extremos, o que pode ser justificado pela alta quantia de categorias dentro de cada esporte, um fator externo à análise, cada uma necessitando de um valor maior de IMC. Isso pode justificar o alto desvio padrão e variância desses esportes. Sobre IMC, ginastica, futebol e badmínton estão, em sua maioria, dentro da área considerada “normal”(18,5-24,9) de IMC, com alguns valores passando para “sobrepeso”(25-29,9), enquanto atletismo e judô possuem diversos valores chegando em “obesidade – grau II”(35-39,9) e até mesmo alguns chegando em “obesidade – grau III”(40+), principalmente em judô, que existe uma categoria apenas para atletas no lado mais extremo de IMC.

3.3 Análise dos 3 maiores medalhistas de 2000-2016

Essa análise visa observar os três atletas com a maior quantia de medalhas no período do ano 2000 ao ano de 2016 para uma possível investigação acerca dos treinos do atleta e talvez descobrir uma tendência entre o atleta e as medalhas conquistadas por ele. Foram usadas duas variáveis qualitativas para essa análise, o nome do atleta, que é qualitativo, e a medalha ganha por dito atleta, que é ordinal. Além disso, foi utilizado um gráfico de barras bivariado.

Gráfico de barras bivariado dos 3 maiores medalhistas

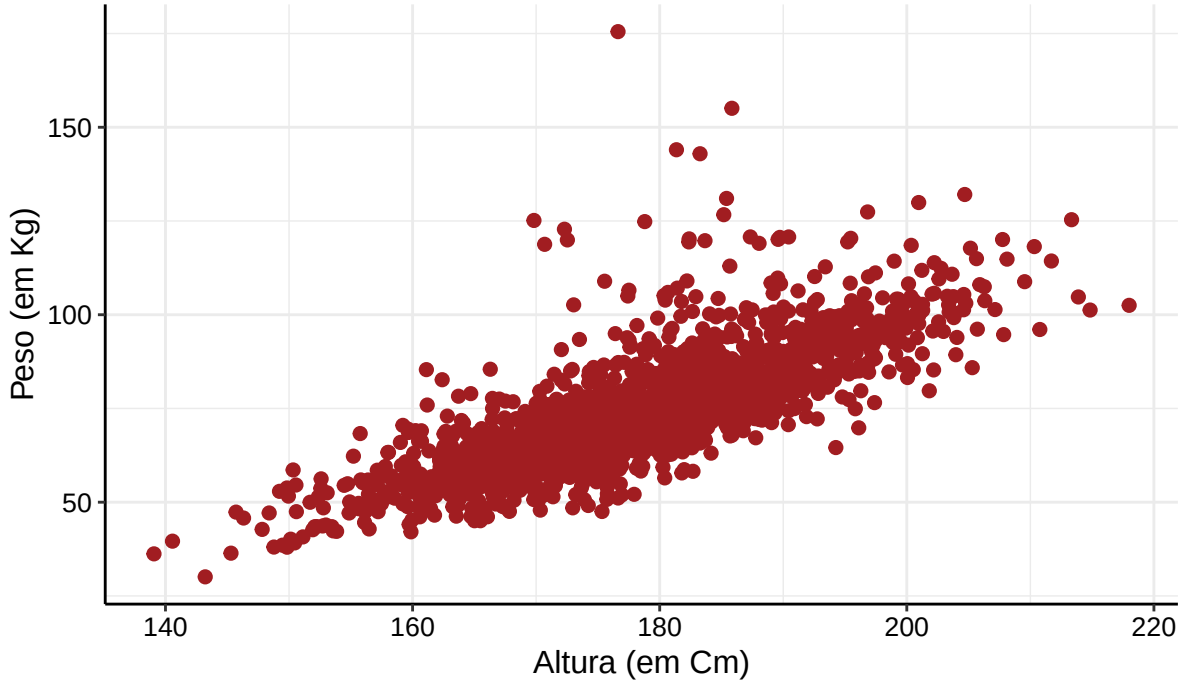


Analisando o gráfico, pode-se perceber que o Michael Phelps foi o maior medalhista do período, dando a ele uma média de 1,75 medalhas por ano, e 5,6 medalhas por olimpíada. Em comparação com o segundo colocado, existe uma diferença de 16 medalhas, mostrando uma superioridade para o atleta, mesmo que as modalidades possam variar.

3.4 Análise da correlação entre o peso e a altura

Essa análise foca em procurar se há alguma correlação entre as variáveis quantitativas contínuas Peso e Altura. Para a clareza no gráfico, o peso está em quilogramas e a altura está em centímetros. Nessa análise, o uso do coeficiente de correlação linear de Pearson é fundamental para o entendimento da correlação entre as variáveis apresentadas. Para essa análise, foram utilizados o gráfico de dispersão e dois quadros de medidas.

Gráfico de dispersão de Altura por Peso



Quadro de medida para Altura

Medidas	Altura
Média	178.29
Desvio Padrão	11.70
Variância	136,85
Mínimo	139.00
1°Quartil	170.00
Mediana	178.00
3°Quartil	186.00
Máximo	217.00
Amplitude	78.00

Quadro de medida para Peso

Medidas	Peso
Média	74.40
Desvio Padrão	16.16
Variância	261.08
Mínimo	31.00
1°Quartil	63.00
Mediana	73.00
3°Quartil	84.00

Medidas	Peso
Máximo	175.00
Amplitude	144.00

Após a análise do gráfico e das tabelas, é possível notar que existe uma tendência de quanto maior a altura, maior o peso, o que pode ser comprovado pelo coeficiente de correlação linear de Pearson que teve um valor de 0.79, mostrando alta correlação, mesmo que existam alguns outliers. Além disso, pode-se perceber que, por mais que a amplitude do peso seja quase o dobro da amplitude da altura, o desvio padrão apresenta valores relativamente próximos, apresentando mais uma semelhança entre as variáveis. Pode ser visto também, considerando os valores dos quartis das tabelas, uma concentração de peso entre 63 e 84 quilogramas e na altura entre 170 e 186 centímetros.

4 Conclusões

Na primeira análise, pôde ser observado que os Estados Unidos têm a superioridade na modalidade feminina. Além disso, por mais que a Alemanha tenha mais medalhas que a China no geral, a China tem medalhas mais bem sucedidas, então, para uma busca de técnicas para conquistar medalhas de ouro, os Estados Unidos e a China são os mais indicados por obterem mais sucesso.

Na segunda análise, é difícil analisar o atletismo e o judô em geral por ter muitas categorias diferentes, entretanto, o futebol, ginástica e badmínton tem valores que são mais fáceis de perceber uma tendência, levando a crer que o valor de IMC indicado para cada esporte se encontra no intervalo de seu respectivo boxplot.

Na terceira análise, ao analisar as 3 maiores quantidades de medalhas conquistadas, deu para perceber uma tendência de que todos os presentes no ranking têm pelo menos uma medalha de ouro, o que é um reflexo da habilidade do atleta. Ou seja, quanto mais medalhas algum atleta conquista, maior a chance de conquistar uma medalha de ouro.

Na quarta análise, é possível notar que existe, nas olimpíadas, uma tendência do peso aumentar à medida que a altura aumenta, provavelmente pelo desejo de manter os atletas em um IMC (Índice de Massa Corporal) saudável para a realização de atividades físicas, o que pode ser comprovado na análise de IMC de alguns esportes, que embora não apresente todos os esportes, ainda mostra uma tendência de média entre eles.

Após todas as análises, pode-se perceber diversas informações que podem ser utilizadas para o aprimoramento do treinamento dos atletas da House of Excellence, embora algumas das análises sejam necessários mais estudos em cima das informações para se obter o melhor resultado possível.