Análise do Desempenho de Atletas Olímpicos: Tendências Globais e Nacionais

Luciano Henrique Arendt Rodrigues
BCC 22 - Ciência de Dados - UNESP/Bauru

Junho de 2025

1 Introdução e Objetivo do Trabalho

O presente trabalho tem como objetivo aplicar técnicas de estatística descritiva e probabilidade na análise de dados reais, explorando o desempenho de atletas ao longo das edições dos Jogos Olímpicos. Utilizando o conjunto de dados "120 Years of Olympic History", disponível na plataforma Kaggle, são analisadas variáveis como idade, sexo, peso, altura, modalidade esportiva e conquistas de medalhas, buscando identificar padrões relevantes e tendências históricas no perfil e desempenho dos atletas olímpicos.

A análise é organizada em duas frentes complementares. A primeira consiste em uma abordagem exploratória e descritiva, que permite observar a distribuição de características físicas dos atletas, a evolução histórica da participação por gênero, os esportes com maior número de competidores e os países com mais medalhas. Entre os resultados observados, destaca-se que as estatísticas corporais dos atletas medalhistas são, em geral, muito semelhantes às dos demais participantes de cada modalidade, o que sugere que outros fatores além do perfil físico influenciam no sucesso olímpico.

Com base nessa constatação, a segunda parte do trabalho aplica uma abordagem preditiva, por meio de modelos probabilísticos e de aprendizado de máquina, com o intuito de estimar a probabilidade de um atleta conquistar uma medalha olímpica. Para isso, utiliza-se como entrada informações como gênero, idade, peso, altura e o esporte praticado, avaliando o desempenho de diferentes modelos na tarefa de classificação.

A escolha do tema "Desempenho de atletas" se justifica pela riqueza histórica e variedade de informações do conjunto de dados, o que permite uma análise ampla, contextualizada e estatisticamente relevante. Além disso, trata-se de um tema de interesse social, esportivo e científico, que possibilita compreender melhor a evolução do perfil dos atletas olímpicos e as variáveis associadas ao sucesso esportivo ao longo do tempo.

2. Descrição do Conjunto de Dados Utilizado

O conjunto de dados utilizado neste trabalho é intitulado *"120 Years of Olympic History: Athletes and Results"*, disponibilizado publicamente na plataforma Kaggle. Ele contém informações detalhadas sobre a participação de atletas nos Jogos Olímpicos, abrangendo

um período de 120 anos, desde a primeira edição da era moderna em **Atenas (1896)** até os Jogos do **Rio de Janeiro (2016)**.

O arquivo analisado, denominado `athlete_events.csv`, possui **271.116 linhas** e **15 colunas**. Cada linha representa a participação de um atleta individual em um evento específico de uma edição olímpica. Isso significa que um mesmo atleta pode aparecer diversas vezes, caso tenha competido em diferentes eventos ou edições dos Jogos.

A tabela a seguir apresenta as colunas presentes no conjunto de dados:

Coluna	Descrição
`ID`	Identificador único para cada atleta
`Name`	Nome do atleta
`Sex`	Sexo do atleta (M ou F)
`Age`	Idade do atleta (inteiro)
`Height`	Altura em centímetros
`Weight`	Peso em quilogramas
`Team`	Nome da equipe ou país representado
`NOC`	Código de três letras do Comitê Olímpico Nacional
`Games`	Combinação do ano e da estação (ex: "2008 Summer")
`Year`	Ano do evento (inteiro)
`Season`	Estação dos Jogos (Summer ou Winter)
`City`	Cidade sede da edição olímpica
`Sport`	Modalidade esportiva (ex: Athletics, Swimming)
`Event`	Evento específico dentro do esporte (ex: 100m Men)
`Medal`	Tipo de medalha conquistada (Gold, Silver, Bronze ou NA)

Tabela 1: Colunas do conjunto de dados `athlete_events.csv`

Esse conjunto de dados é especialmente adequado para análises de desempenho e perfil dos atletas olímpicos, pois combina variáveis demográficas e físicas com informações de desempenho esportivo. A amplitude temporal e a diversidade de países representados permitem análises históricas, comparativas e segmentadas por nacionalidade, esporte, sexo e outros critérios relevantes.

A fonte original do dataset é:
`https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-re`

3. Etapas de Limpeza e Tratamento dos Dados

Durante a construção deste trabalho, foram realizadas diferentes etapas de limpeza e transformação do conjunto de dados, de acordo com as necessidades específicas de cada abordagem. Nesta seção, os procedimentos são organizados em duas frentes: uma voltada para a análise exploratória e outra voltada para a aplicação de técnicas estatísticas preditivas.

3.1 Tratamento para Análise Exploratória

Antes de cada análise estatística e da construção dos respectivos gráficos, foram realizadas etapas de limpeza e preparação dos dados, com o objetivo de garantir a consistência e a relevância dos resultados obtidos.

Exclusão de valores nulos

Como o conjunto de dados possui entradas com valores ausentes em variáveis essenciais como `Age`, `Height`, `Weight` e `Medal`, optou-se por realizar a exclusão dessas linhas apenas nas análises que dependem diretamente dessas variáveis. Por exemplo:

- Para o cálculo da **média de altura**, apenas atletas com valor válido em `Height` foram considerados;
- Para a contagem de medalhas, foram removidas entradas com valor `NaN` na coluna `Medal`.

Conversão de tipos de dados

Foi adicionada uma nova coluna chamada `Medal Points`, a fim de facilitar a agregação de dados por tipo de medalha. Nessa coluna, o tipo de medalha foi convertido para valores numéricos:

- `Gold` $\rightarrow$ 3,
- `Silver` $\rightarrow$ 2,
- `Bronze` $\rightarrow$ 1,
- `NaN` $\rightarrow$ 0.

Essa transformação permitiu, por exemplo, calcular somatórios ou médias ponderadas de desempenho por país, esporte ou período.

Agrupamentos e agregações

Foram aplicados agrupamentos com base em múltiplas dimensões do conjunto de dados, com o objetivo de obter informações agregadas e comparativas. Alguns exemplos de operações realizadas:

- Contagem de atletas por sexo ao longo dos anos:

```
df_m.groupby(['Year', 'Sex'])['ID'].count().unstack()
```

- Cálculo da idade média dos atletas ao longo do tempo:

```
df_m.groupby(by='Year')['Age'].mean().plot()
```

- Cálculo do peso e altura médios ao longo dos anos:

```
df_m.groupby('Year')['Weight'].mean()
df_m.groupby('Year')['Height'].mean()
```

- Cálculo de médias por esporte:

```
df_m.groupby('Sport')['Age'].mean()
df_m.groupby('Sport')['Height'].mean()
df_m.groupby('Sport')['Weight'].mean()
```

- Identificação dos países com mais medalhas (ignorando valores nulos):

```
df_m[df_m['Medal'].notna()]['region'].value_counts().head(10)
```

Essas transformações foram fundamentais para permitir análises descritivas e visualizações claras e informativas ao longo da seção de Análise Exploratória.

3.2 Tratamento para Aplicação de Técnica Estatística Preditiva

Para a construção dos modelos preditivos apresentados neste trabalho, foi necessário preparar um subconjunto específico do conjunto de dados original, contendo apenas as variáveis relevantes para o problema de classificação. O objetivo era prever a probabilidade de um atleta olímpico conquistar uma medalha com base em suas características físicas e no esporte praticado.

As etapas de tratamento dos dados foram as seguintes:

- **Remoção de valores ausentes:** foram eliminadas todas as linhas com valores nulos nas colunas essenciais para a predição: `Sex`, `Age`, `Height`, `Weight` e `Sport`. Isso garantiu que apenas registros completos fossem utilizados nos modelos.
- **Conversão da variável alvo:** a coluna `Medal` foi convertida em uma variável binária, onde `1` representa atletas que conquistaram medalha (ouro, prata ou bronze) e `0` representa atletas que não conquistaram medalha. Essa transformação permitiu a formulação do problema como uma tarefa de classificação binária.
- **Codificação de variáveis categóricas:** as colunas `Sex` e `Sport`, originalmente categóricas, foram transformadas em valores numéricos por meio do uso do *LabelEncoder*, possibilitando seu uso direto nos algoritmos de aprendizado de máquina.
- **Seleção de colunas relevantes:** após o tratamento, foram mantidas apenas as colunas diretamente relacionadas à previsão: `Sex`, `Age`, `Height`, `Weight`, `Sport` e `Medal`.
- **Verificação e preenchimento residual de valores ausentes:** como etapa final de segurança, foi verificado se ainda havia valores ausentes nas colunas numéricas. Nos poucos casos encontrados, os valores foram preenchidos com a média da respectiva variável.

Essas transformações resultaram em um conjunto de dados limpo, estruturado e pronto para a aplicação de técnicas estatísticas preditivas, como os modelos de *Naive Bayes* e *Random Forest*, descritos na próxima seção.

Além da exclusão de valores ausentes, também foram realizadas codificações específicas para tratar variáveis categóricas e garantir compatibilidade com os algoritmos utilizados.

As etapas específicas de pré-processamento e a estrutura final dos dados utilizados nos modelos serão descritas na seção de aplicação preditiva.

4. Análise Exploratória (Estatística Descritiva e Gráficos)

Esta seção apresenta a análise exploratória dos dados com base em técnicas de estatística descritiva e visualizações gráficas. O objetivo é identificar padrões relevantes nos dados dos atletas olímpicos, como evolução temporal da participação, diferenças entre gêneros, variações físicas (idade, peso, altura) e distribuição de medalhas.

As análises incluem a distribuição dos atletas por gênero, a evolução da participação masculina e feminina ao longo do tempo, a quantidade de medalhas conquistadas por país, e a evolução física dos atletas (em termos de idade, peso e altura).

Além disso, são examinadas as modalidades esportivas com maior número de participantes e a relação entre características físicas e o tipo de esporte praticado. Os dados são representados por meio de gráficos e tabelas que auxiliam na visualização de tendências históricas e padrões relevantes no cenário olímpico mundial.

4.1 Número de participantes por ano olímpico

Uma das primeiras análises realizadas diz respeito ao número total de atletas participantes em cada edição dos Jogos Olímpicos. Esse indicador permite observar a evolução da escala do evento ao longo do tempo, refletindo tanto o crescimento do interesse mundial pelas Olimpíadas quanto a ampliação da participação feminina e de novos países ao longo das décadas.

É importante destacar que, a partir de 1994, o Comitê Olímpico Internacional decidiu **desmembrar a realização das Olimpíadas de Verão e de Inverno**, que até então ocorriam no mesmo ano. Desde então, os Jogos Olímpicos de Inverno passaram a ser realizados em anos pares alternados com os de Verão — coincidindo com os anos da *FIFA World Cup*. Essa mudança explica a presença de anos com número significativamente menor de participantes no gráfico abaixo, pois esses anos correspondem às edições de inverno, que tradicionalmente envolvem um número menor de atletas e modalidades.

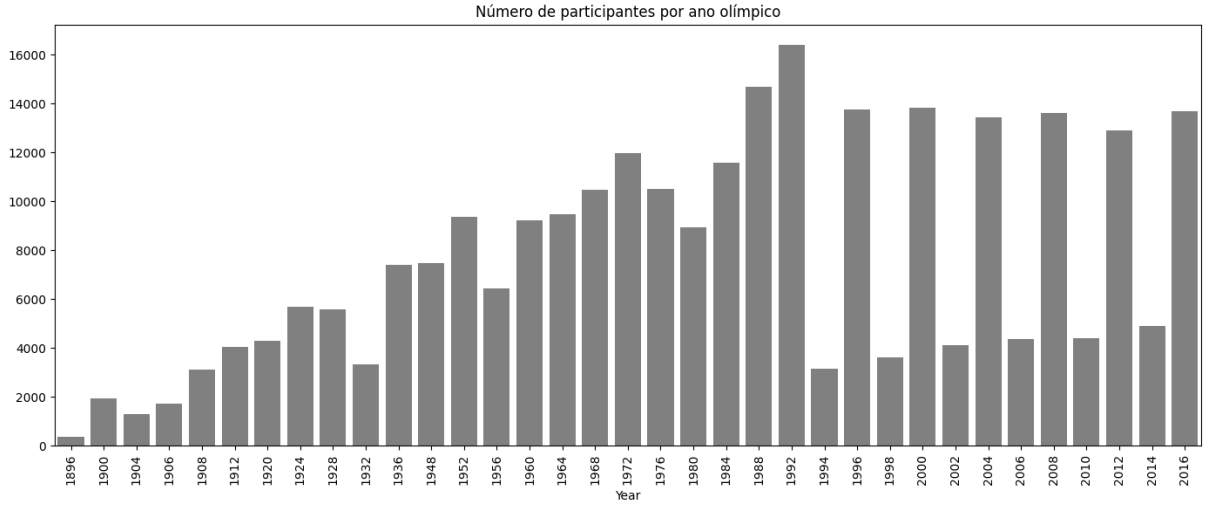

Figura 1: Número de atletas participantes por ano olímpico (Verão e Inverno)

4.2 Número de atletas por esporte

Esta análise busca identificar os esportes que historicamente tiveram maior número de atletas participantes nas Olimpíadas. Segundo o conjunto de dados analisado, os cinco esportes com maior volume de participação ao longo da história são: **Atletismo (Athletics)**, **Ginástica (Gymnastics)**, **Natação (Swimming)**, **Tiro Esportivo (Shooting)** e **Ciclismo (Cycling)**.

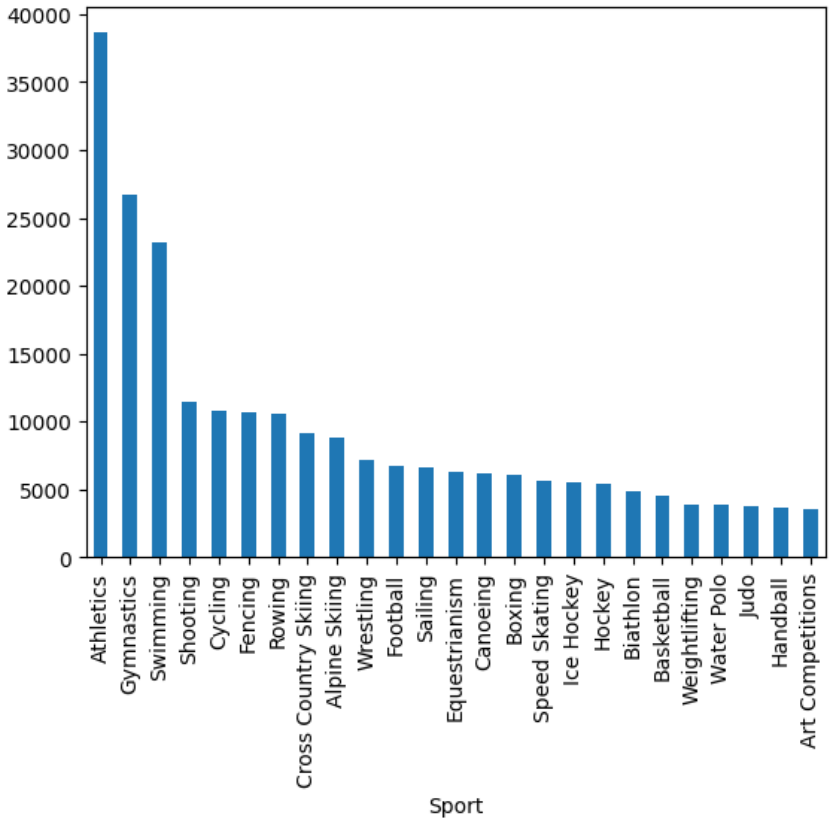

Figura 2: Número total de atletas por esporte ao longo das edições olímpicas

Observa-se que esses esportes, além de serem tradicionais, contam com um grande número de eventos diferentes dentro da mesma modalidade, o que contribui diretamente para o maior número de atletas registrados ao longo do tempo.

4.3 Distribuição de Medalhas e Países com Maior Desempenho

O conjunto de dados revela que houve uma distribuição levemente desigual entre os tipos de medalha. De forma curiosa, há um número ligeiramente maior de medalhas de **ouro** do que de **bronze**, e um pouco mais de medalhas de **bronze** do que de **prata**. Essa diferença pode ocorrer por motivos históricos, como empates em finais, mudanças de regras ao longo das décadas e número variável de eventos por edição.

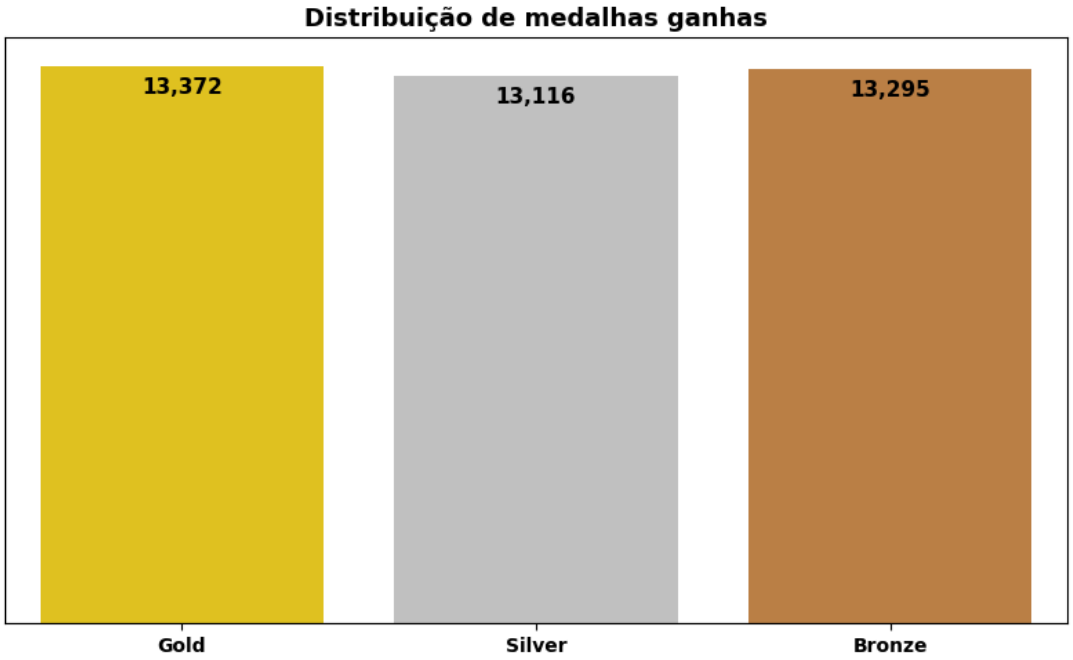

Figura 3: Distribuição total de medalhas por tipo (Ouro, Prata, Bronze)

No que diz respeito ao desempenho por país, a análise dos dez Comitês Olímpicos Nacionais com maior número de medalhas ao longo da história mostra clara dominância de alguns países. Os **Estados Unidos** (USA) lideram com ampla vantagem, seguidos por **Rússia** (considerando também os períodos como URSS) e **Alemanha**, que mantêm alta performance ao longo das décadas.

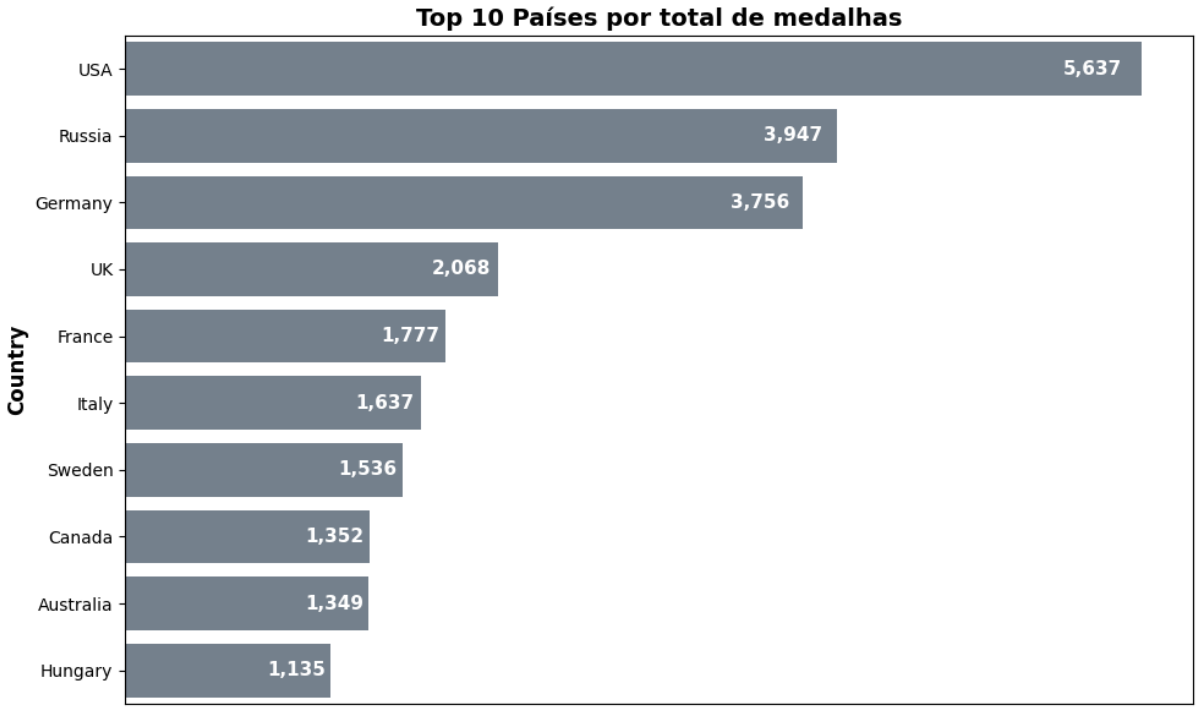

Figura 4: Top 10 países com maior número de medalhas olímpicas

Observa-se que os países no topo da lista são, em geral, aqueles com maior investimento histórico em esporte de alto rendimento e com maior participação nas edições dos Jogos. O desempenho nessas nações reflete também o número de atletas enviados, a variedade de modalidades em que competem e, em alguns casos, o contexto político (como o boicote a certas edições).

4.4 Distribuição de atletas por gênero

A análise da distribuição geral de atletas por gênero no conjunto de dados revela uma disparidade significativa: são **196.594 atletas do sexo masculino** contra **74.522 atletas do sexo feminino**. Essa diferença reflete, em parte, as restrições históricas à participação de mulheres nos Jogos Olímpicos, especialmente nas edições mais antigas.

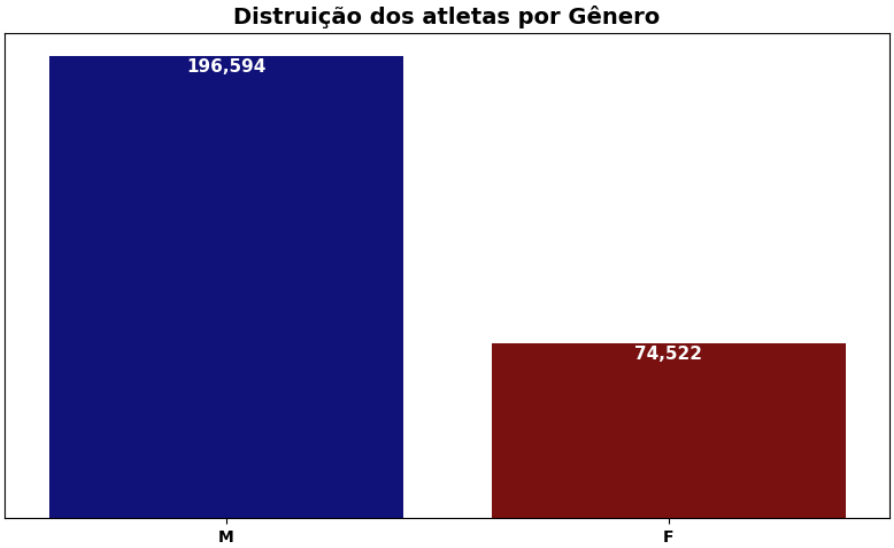

Figura 5: Distribuição geral de atletas por gênero

Quando observamos a evolução da participação por gênero ao longo do tempo, nota-se uma tendência clara de redução da desigualdade. A partir da década de 1990, especialmente após a reorganização do calendário olímpico em 1994, houve um aumento mais expressivo na participação feminina. Ainda assim, a quantidade de atletas do sexo masculino continua superior à de atletas do sexo feminino em praticamente todas as edições dos Jogos.

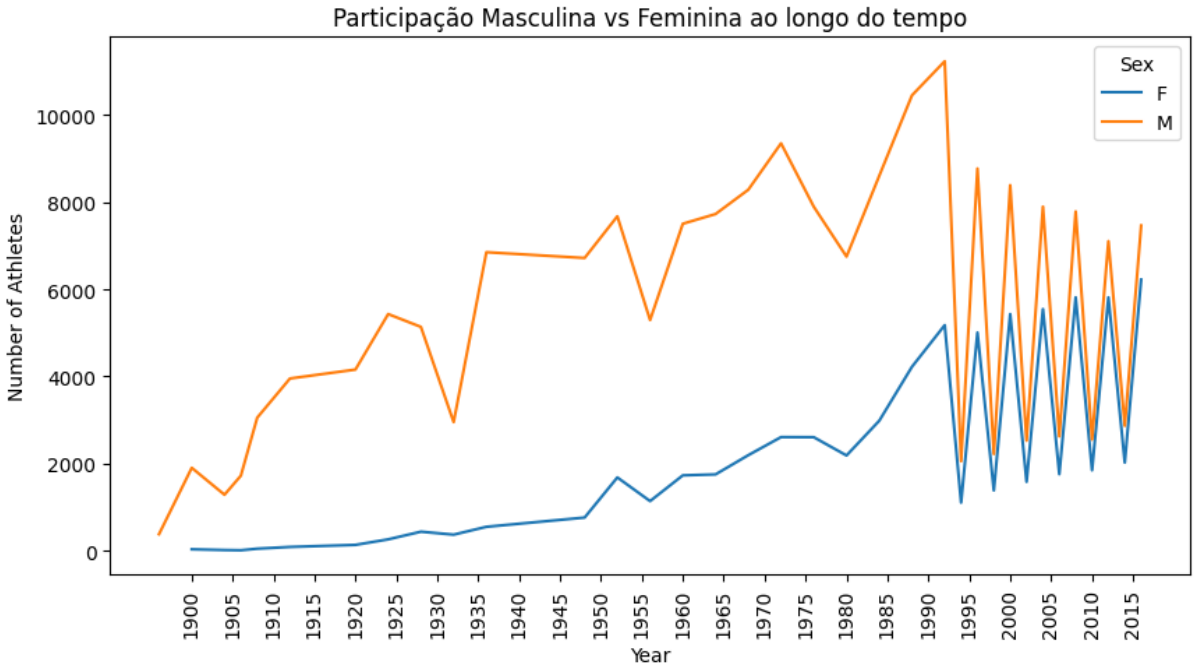

Figura 6: Distribuição de atletas por gênero ao longo do tempo

Esse aumento gradual da participação feminina está relacionado a mudanças institucionais promovidas pelo Comitê Olímpico Internacional, bem como à inclusão de novas

modalidades e eventos com participação mista ou exclusivamente feminina.

4.5 Distribuição da idade dos atletas

A seguir são apresentadas as estatísticas descritivas básicas da variável `Age`, com base nos 261.642 atletas para os quais essa informação está disponível:

Estatística	**Valor**
Total de registros válidos	261.642
Média	25,56 anos
Desvio padrão	6,39 anos
Idade mínima	10 anos
1º Quartil (25%)	21 anos
Mediana (50%)	24 anos
3º Quartil (75%)	28 anos
Idade máxima	97 anos

Tabela 2: Estatísticas descritivas da idade dos atletas

A figura a seguir apresenta um boxplot com a distribuição geral das idades:

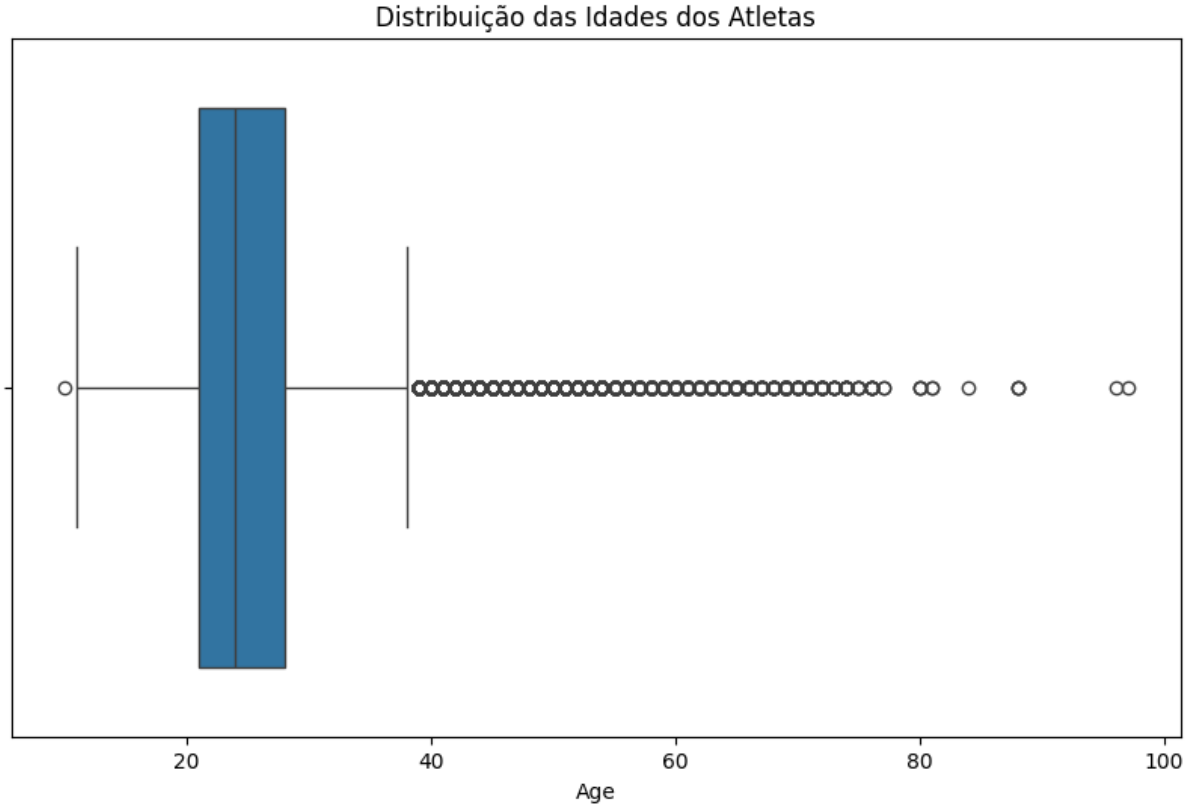

Figura 7: Boxplot da idade dos atletas olímpicos

Em seguida, observa-se a distribuição de frequência das idades dos atletas:

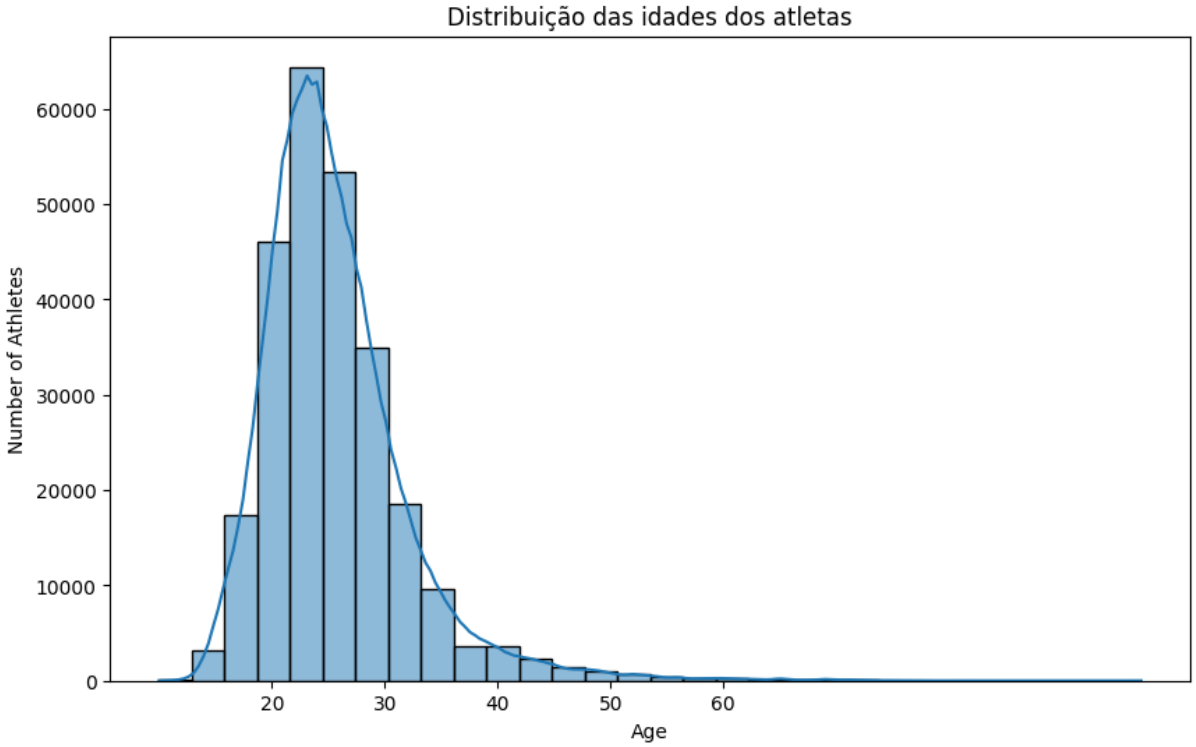

Figura 8: Distribuição da idade dos atletas olímpicos

A forma da distribuição sugere um comportamento aproximadamente **normal** ou possivelmente compatível com uma **distribuição gama**, com leve assimetria à direita. Isso é coerente com o fato de que a maioria dos atletas está na faixa dos 20 a 30 anos, enquanto casos de atletas com idades muito altas são mais raros.

No gráfico a seguir, observa-se a variação da idade média dos atletas ao longo dos anos:

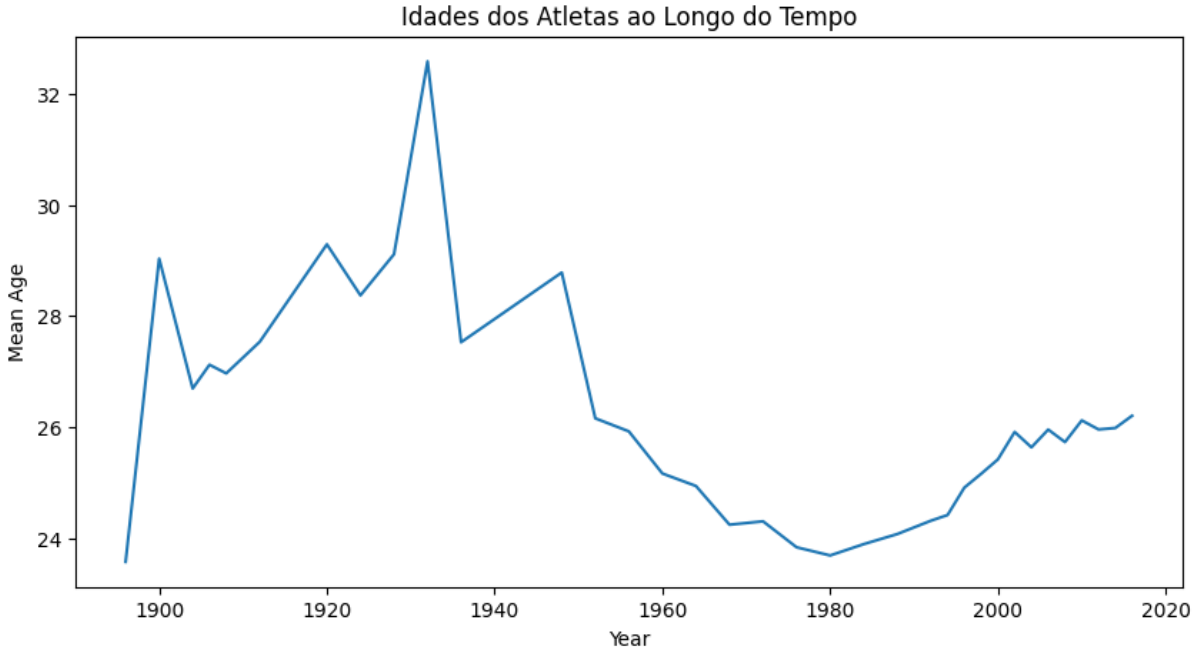

Figura 9: Idade média dos atletas ao longo do tempo

É possível notar que nas primeiras edições dos Jogos Olímpicos (final do século XIX e início do século XX), a média de idade dos atletas era ligeiramente mais alta, o que pode estar relacionado ao perfil mais elitizado e restrito da competição nessas épocas.

Por fim, a figura abaixo apresenta a idade média por modalidade esportiva:

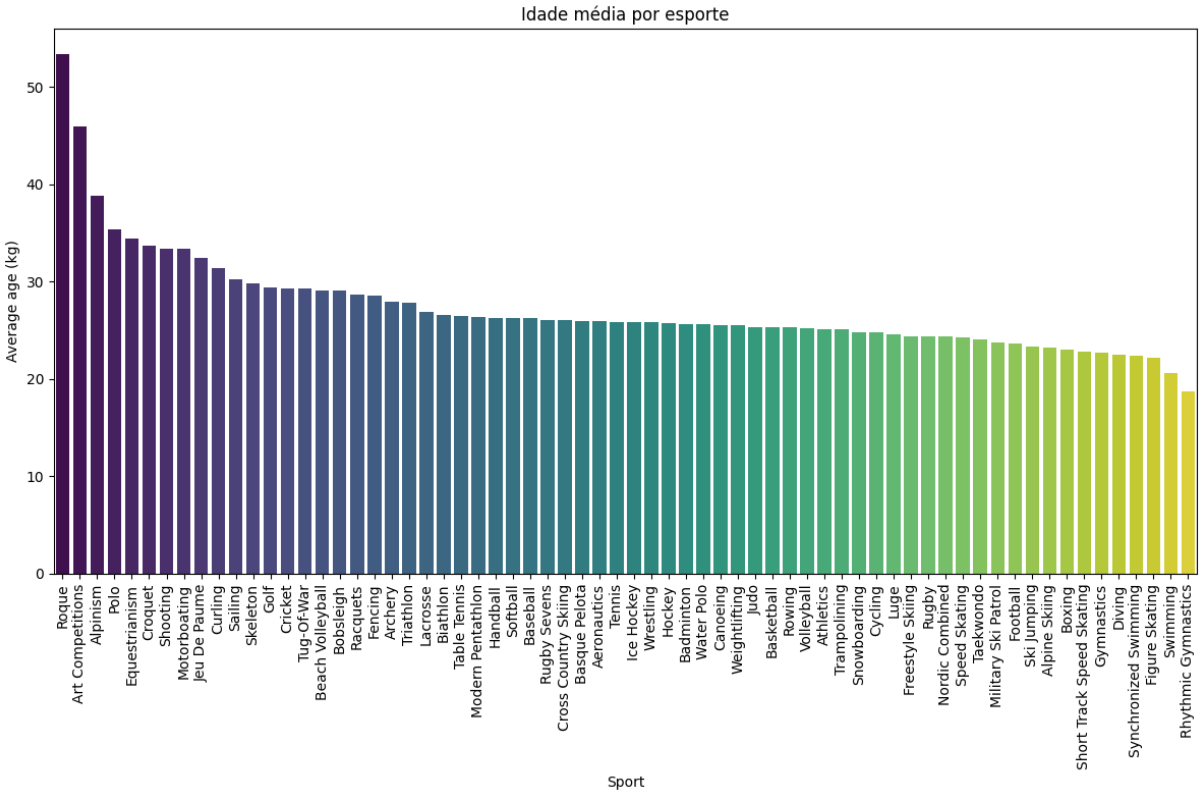

Figura 10: Idade média dos atletas por esporte

Destaca-se que o esporte com maior idade média é o **Roque**, com média superior a 50 anos, o que se justifica pelo fato de ter sido uma modalidade pontual, com baixa exigência física e disputada em um único ano (1904). Já o esporte com menor idade média é a **Ginástica Rítmica (Rhythmic Gymnastics)**, cuja exigência de flexibilidade e desempenho físico extremo favorece a participação de atletas mais jovens, frequentemente com menos de 20 anos.

Além da idade média geral por esporte, também foi analisada a **idade média dos atletas medalhistas por modalidade**, a fim de verificar se a conquista de medalhas está associada a faixas etárias específicas em cada esporte. O resultado, ilustrado na figura a seguir, mostra que as idades dos medalhistas por esporte seguem um padrão muito semelhante ao da média geral de atletas por esporte, o que indica que, em geral, os atletas vencedores pertencem às faixas etárias mais representativas de suas respectivas modalidades.

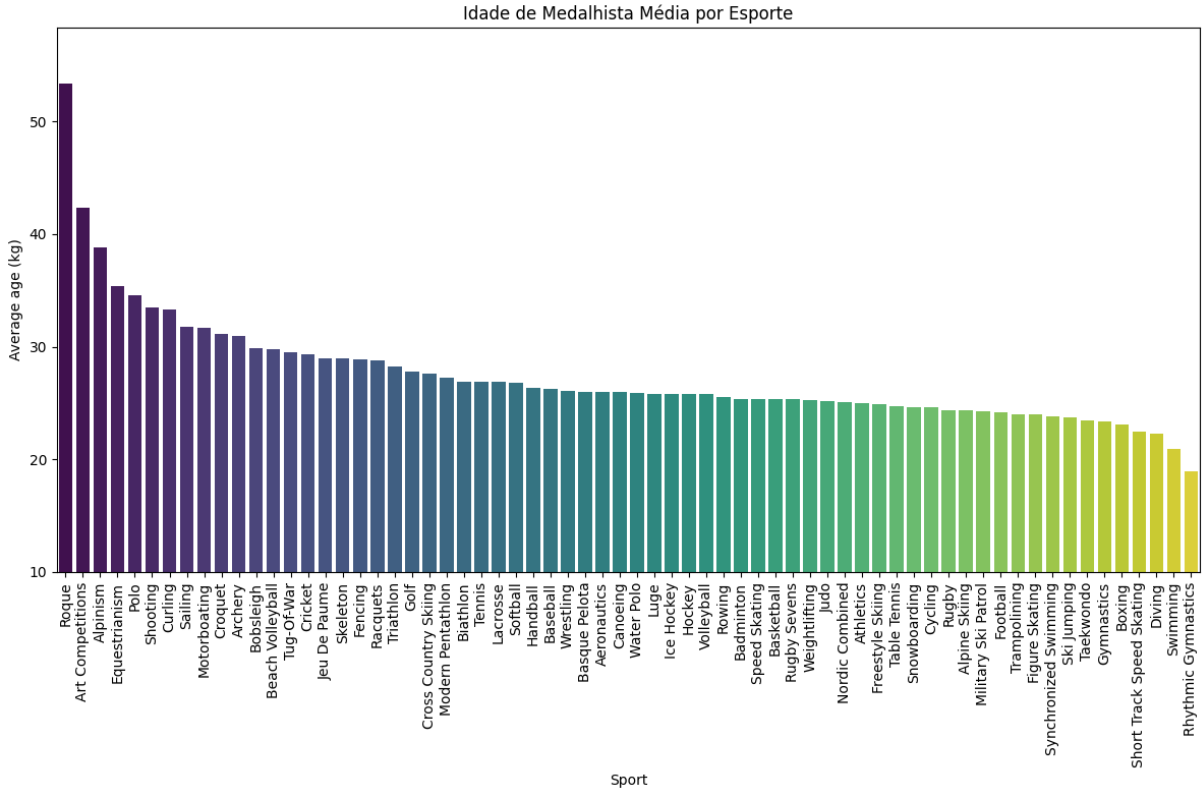

Figura 11: Idade média de atletas medalhistas por esporte

4.6 Distribuição do peso dos atletas

A seguir são apresentadas as estatísticas descritivas da variável `Weight`, considerando os 208.241 atletas para os quais essa informação está disponível:

Estatística	**Valor**
Total de registros válidos	208.241
Média	70,70 kg
Desvio padrão	14,35 kg
Peso mínimo	25 kg
1º Quartil (25%)	60 kg
Mediana (50%)	70 kg
3º Quartil (75%)	79 kg
Peso máximo	214 kg

Tabela 3: Estatísticas descritivas do peso dos atletas

A figura a seguir apresenta um boxplot da distribuição dos pesos registrados:

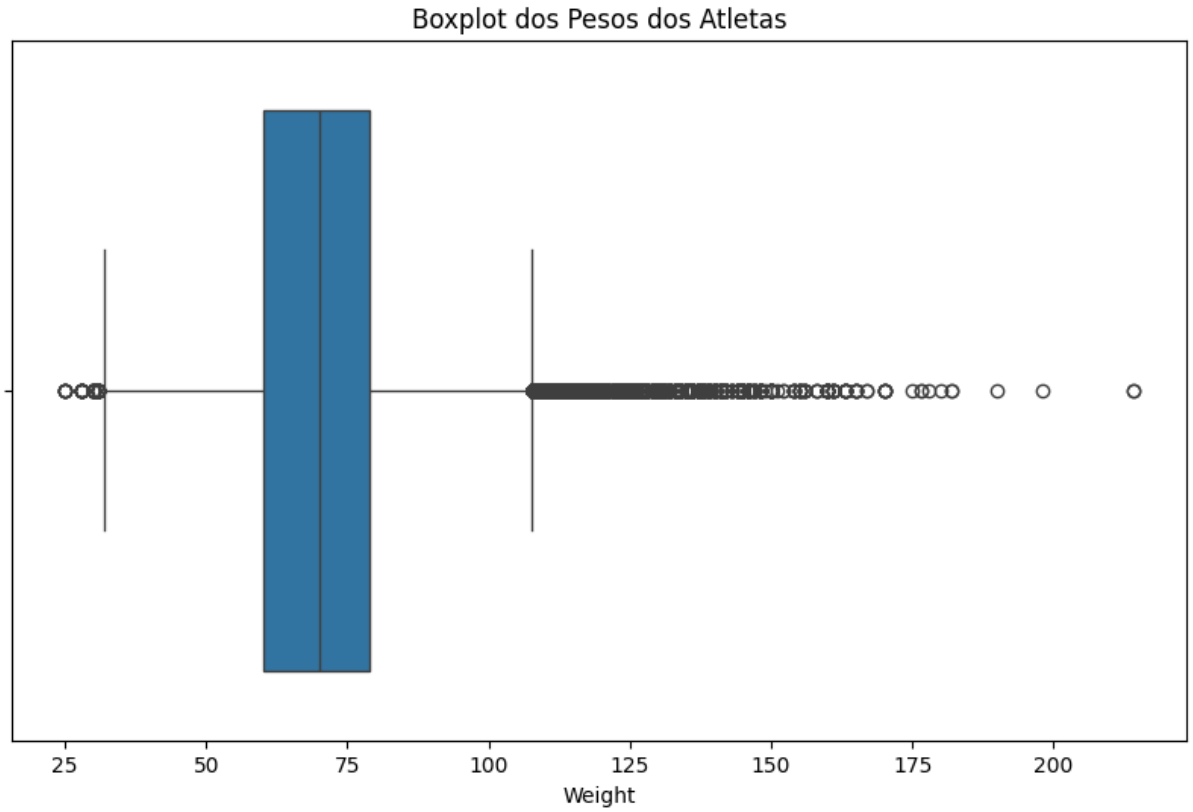

Figura 12: Boxplot do peso dos atletas olímpicos

A seguir, observa-se a distribuição da frequência dos pesos dos atletas:

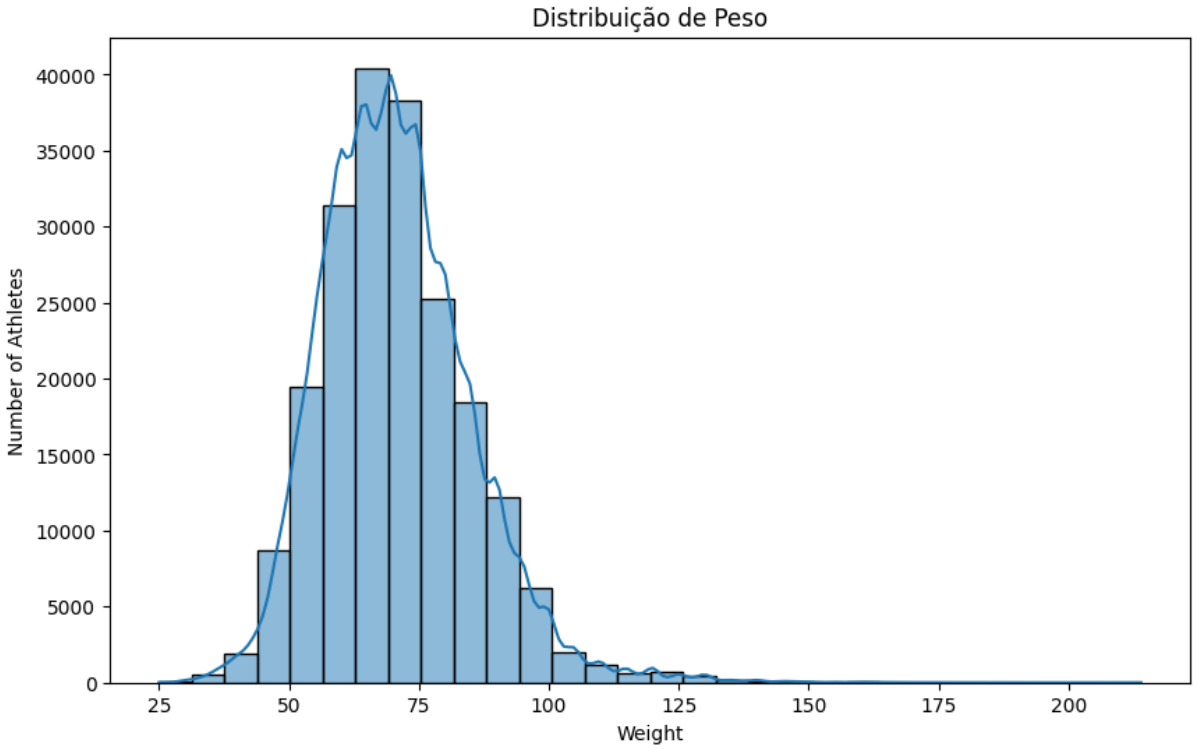

Figura 13: Distribuição do peso dos atletas olímpicos

Assim como na variável idade, a distribuição do peso apresenta formato aproximadamente **normal** ou com características próximas de uma **distribuição gama**, com leve assimetria à direita — sugerindo uma maior concentração em torno da média e poucos casos de atletas com pesos muito elevados.

O gráfico a seguir mostra a evolução do peso médio ao longo das edições dos Jogos Olímpicos:

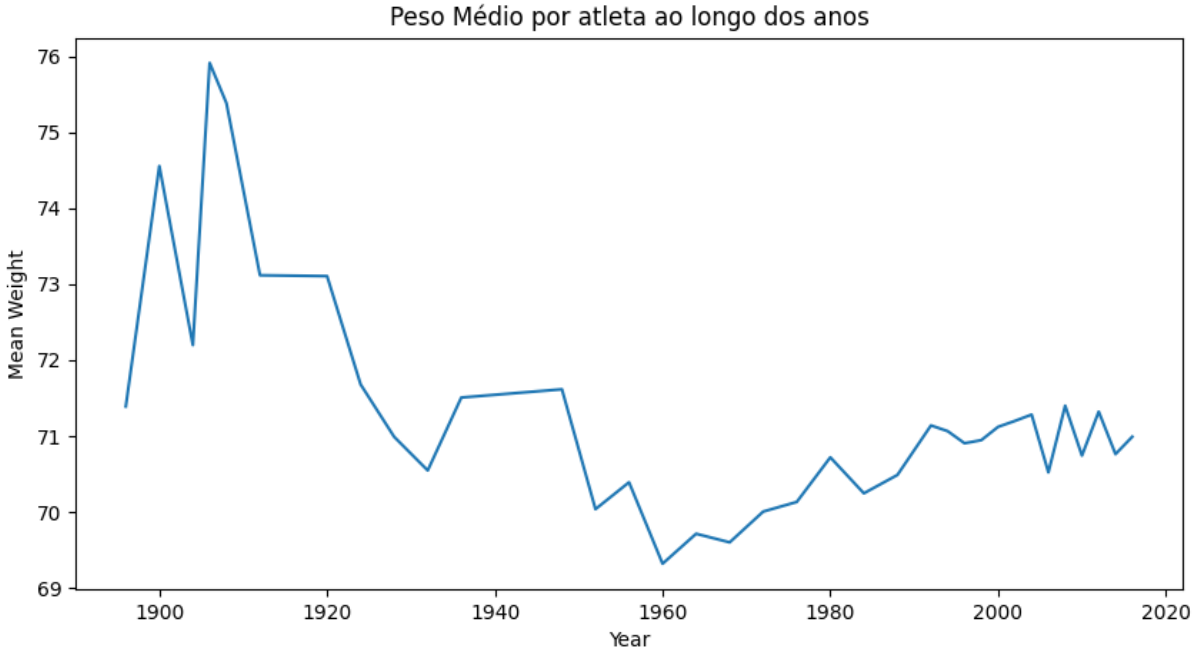

Figura 14: Peso médio dos atletas ao longo do tempo

Nas primeiras edições dos Jogos, nota-se que o peso médio era ligeiramente mais elevado, o que pode estar relacionado ao perfil esportivo das modalidades mais comuns na época (como remo, luta e levantamento de peso), bem como à ausência de algumas categorias femininas mais leves.

A figura a seguir apresenta a média de peso dos atletas por esporte:

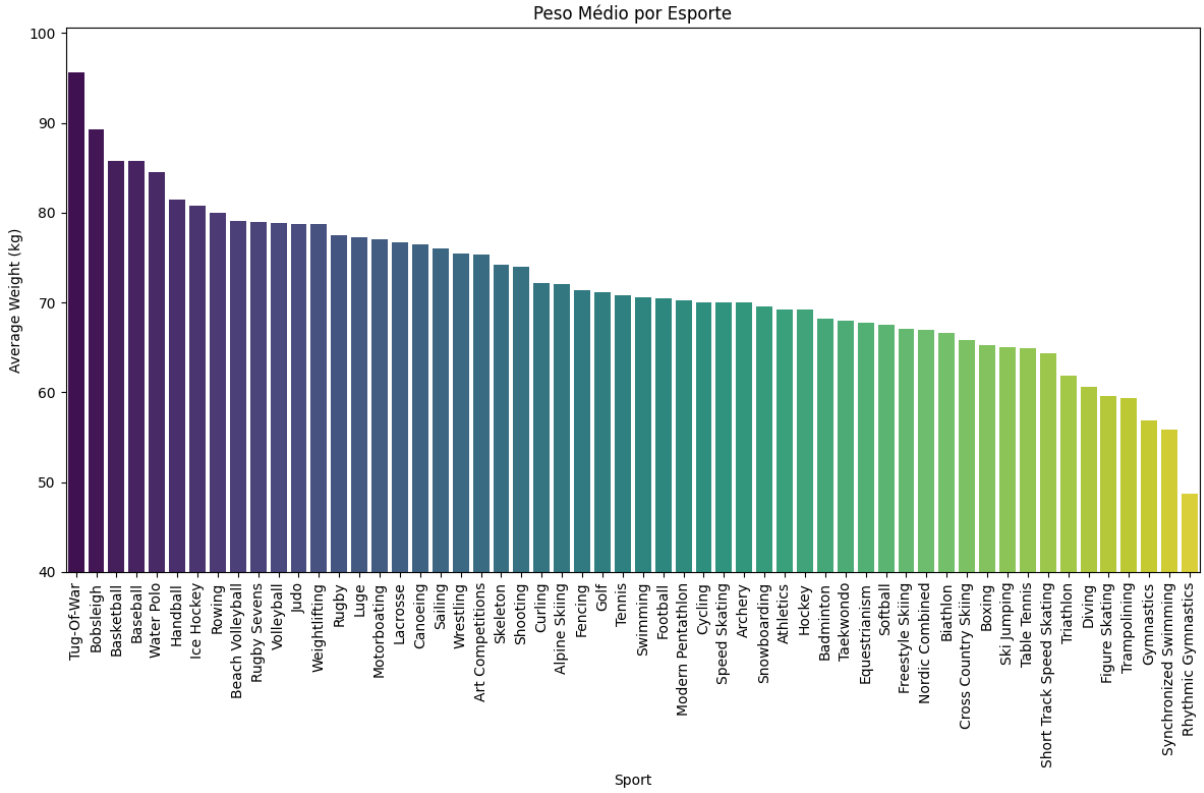

Figura 15: Peso médio dos atletas por esporte

O esporte com maior peso médio é o **Cabo de Guerra (Tug-of-War)**, com média aproximada de 96 kg. Essa modalidade exigia força bruta e foi disputada entre 1900 e 1920. Por outro lado, o esporte com menor peso médio é a **Ginástica Rítmica (Rhythmic Gymnastics)**, com atletas pesando, em média, 48 kg — característica esperada de uma modalidade que exige leveza, flexibilidade e controle corporal extremo.

Também foi analisado o **peso médio dos atletas medalhistas por esporte**, com o objetivo de verificar se há alguma diferença significativa em relação ao perfil geral de cada modalidade. Conforme apresentado na figura a seguir, observa-se que os pesos dos medalhistas se mantêm bastante próximos às médias gerais de suas respectivas modalidades. Isso sugere que, em esportes olímpicos, o perfil físico ideal — em termos de peso — está geralmente alinhado com aquele que historicamente alcança o pódio.

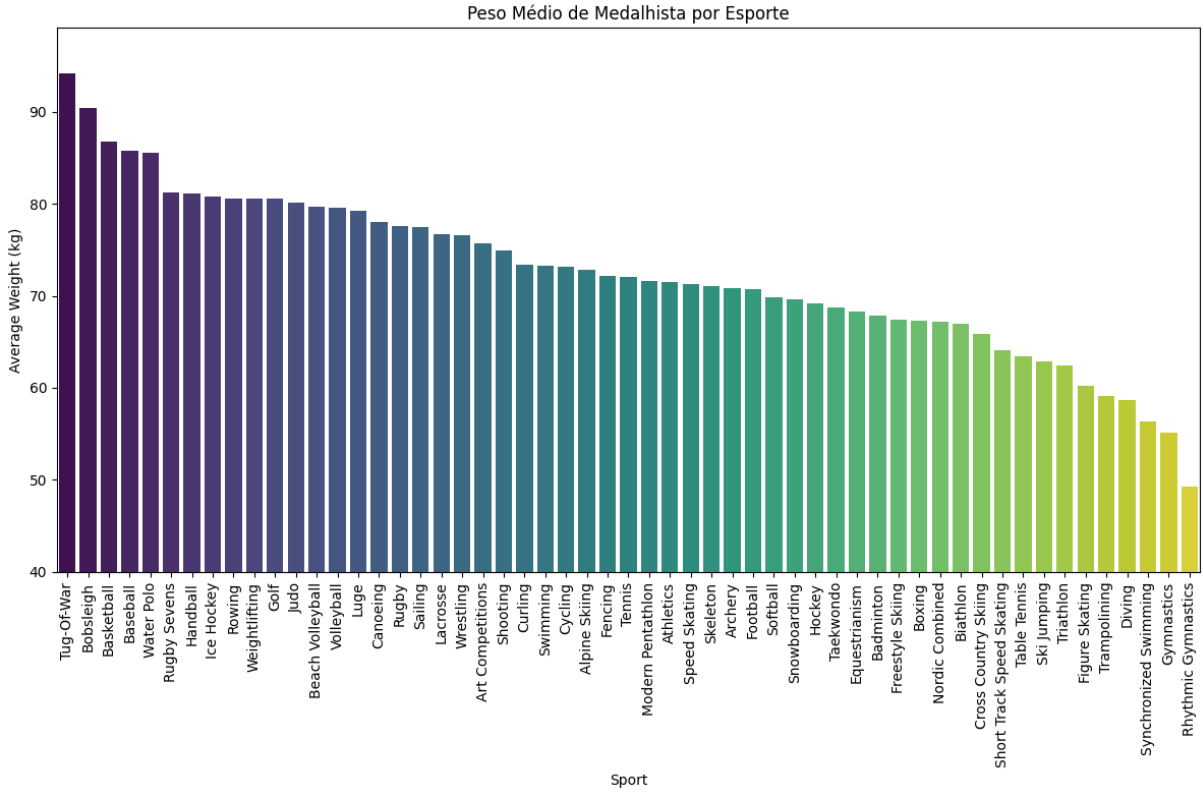

Figura 16: Peso médio de atletas medalhistas por esporte

4.7 Distribuição da altura dos atletas

A seguir são apresentadas as estatísticas descritivas da variável `Height`, considerando os 210.945 atletas para os quais essa informação está disponível:

Estatística	**Valor**
Total de registros válidos	210.945
Média	175,34 cm
Desvio padrão	10,52 cm
Altura mínima	127 cm
1º Quartil (25%)	168 cm
Mediana (50%)	175 cm
3º Quartil (75%)	183 cm
Altura máxima	226 cm

Tabela 4: Estatísticas descritivas da altura dos atletas

A figura a seguir apresenta o boxplot da distribuição geral das alturas dos atletas:

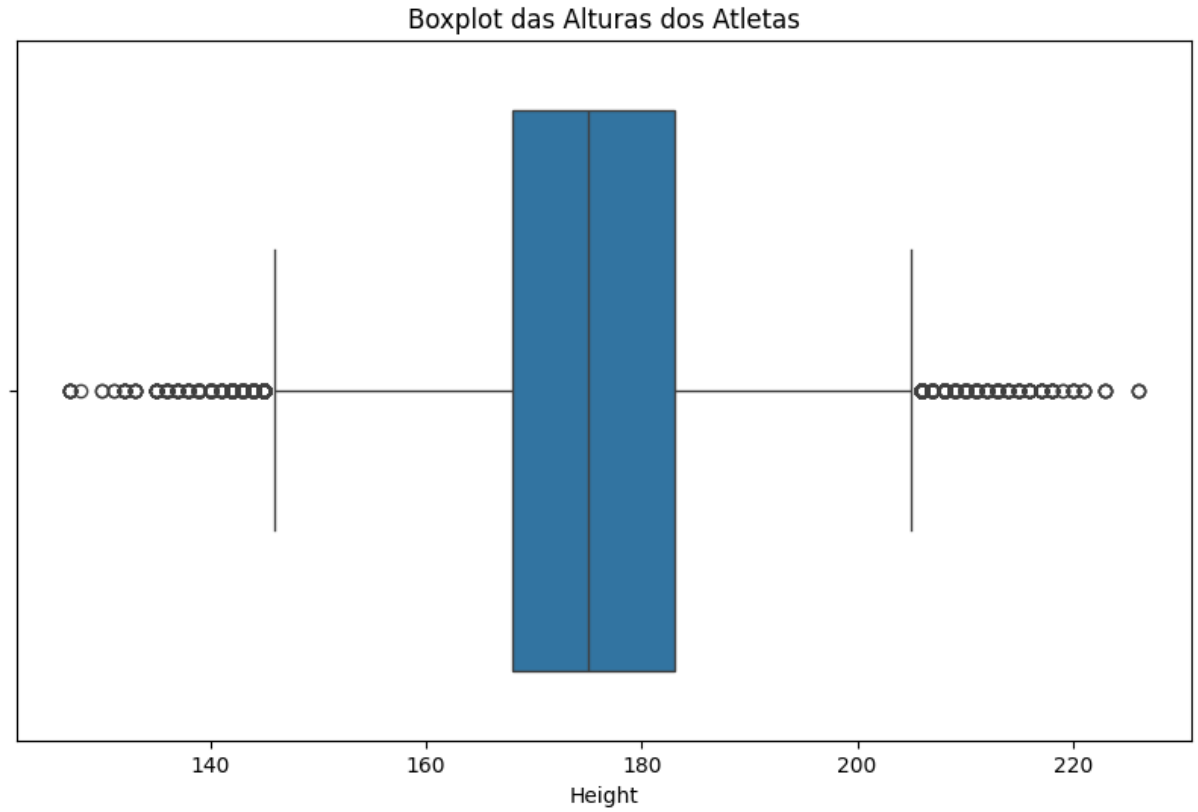

Figura 17: Boxplot da altura dos atletas olímpicos

O gráfico de distribuição da variável altura é apresentado a seguir:

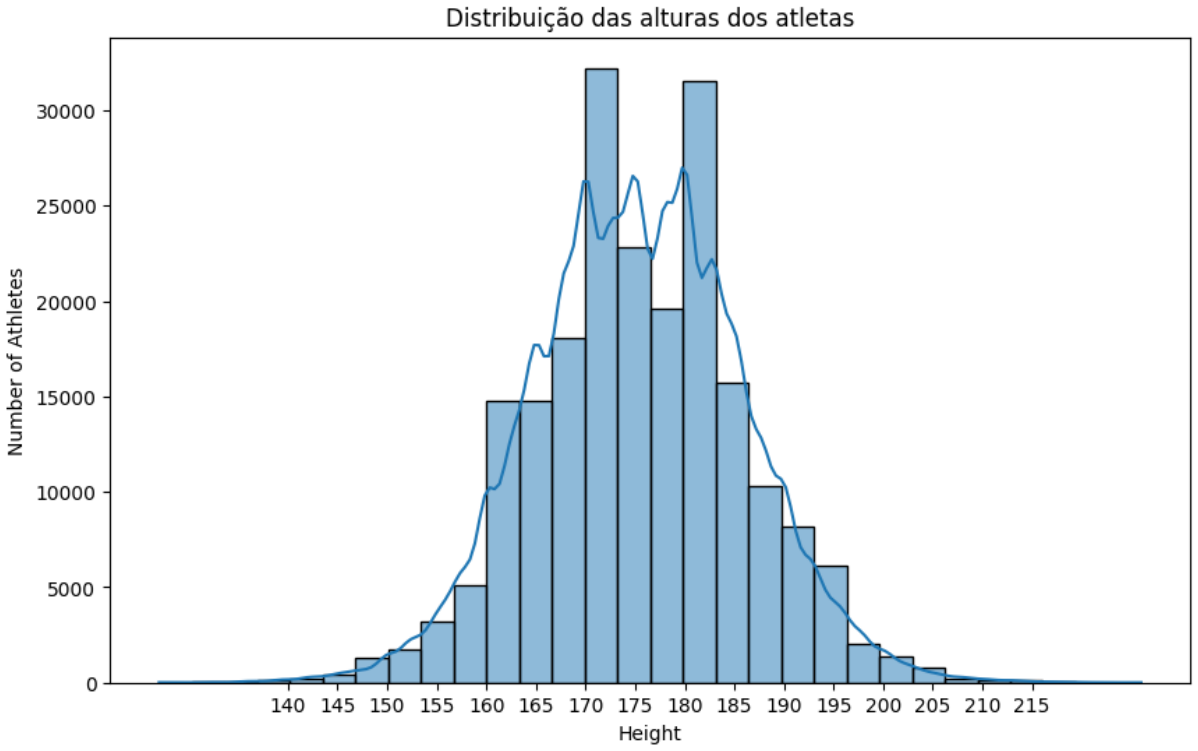

Figura 18: Distribuição da altura dos atletas olímpicos

A distribuição da altura dos atletas se aproxima de uma **distribuição normal**, com leve concentração em torno da média, o que é compatível com a diversidade esportiva e os critérios físicos típicos das modalidades olímpicas.

No gráfico a seguir, observa-se a variação da altura média dos atletas ao longo do tempo:

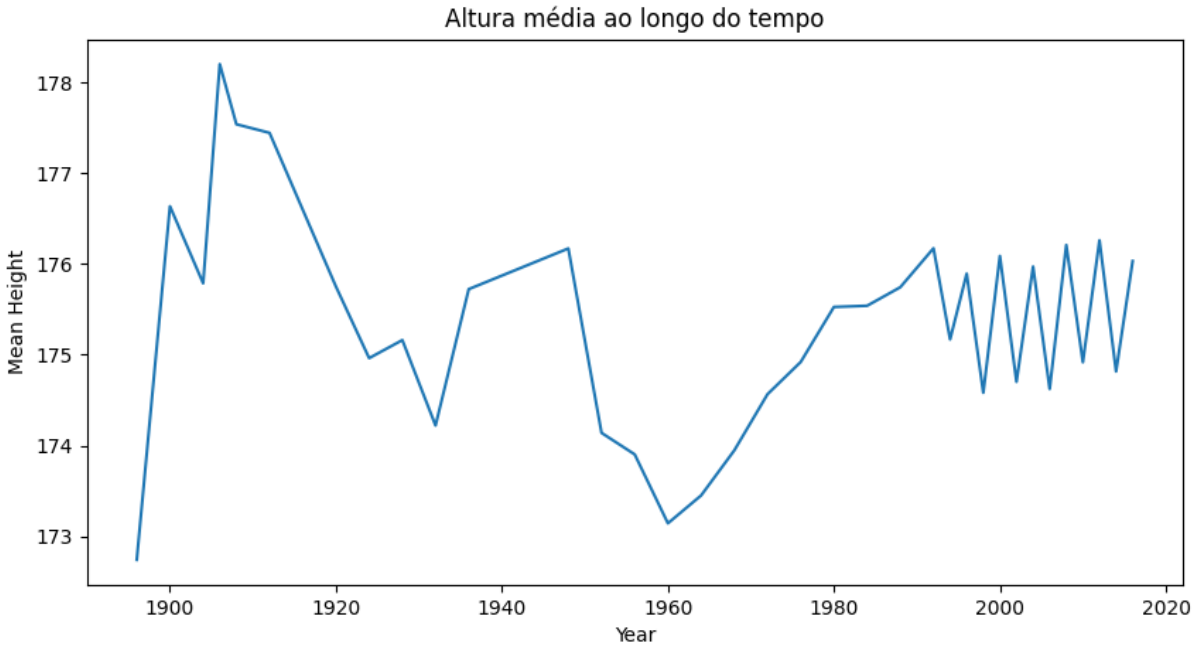

Figura 19: Altura média dos atletas ao longo do tempo

É notável que, a partir de 1994, quando os Jogos de Inverno passaram a ocorrer em anos separados dos Jogos de Verão, a altura média dos atletas passou a apresentar oscilações mais perceptíveis. Isso se deve ao fato de que as modalidades de inverno, em geral, envolvem atletas com altura média menor, o que influencia a média global nos anos em que esses jogos ocorrem isoladamente.

A seguir, temos a média de altura dos atletas por modalidade esportiva:

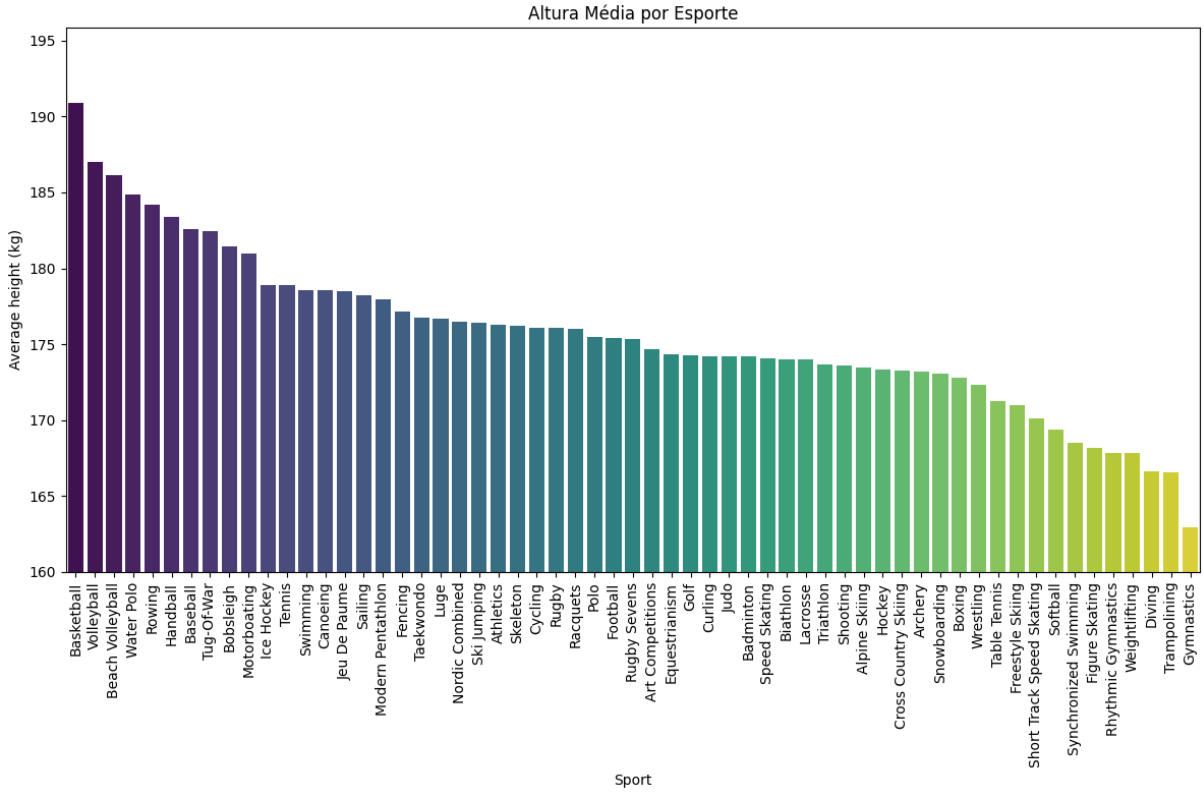

Figura 20: Altura média dos atletas por esporte

O esporte com maior altura média é o **Basquete (Basketball)**, com média de aproximadamente 190 cm, o que reflete a demanda física da modalidade por estatura elevada. Por outro lado, a **Ginástica Artística (Gymnastics)** apresenta a menor média, com cerca de 161 cm, característica comum em esportes que exigem agilidade, controle corporal e baixo centro de gravidade.

Por fim, o gráfico a seguir apresenta a **altura média dos atletas medalhistas por esporte**, permitindo observar se há diferenças relevantes em relação às médias gerais:

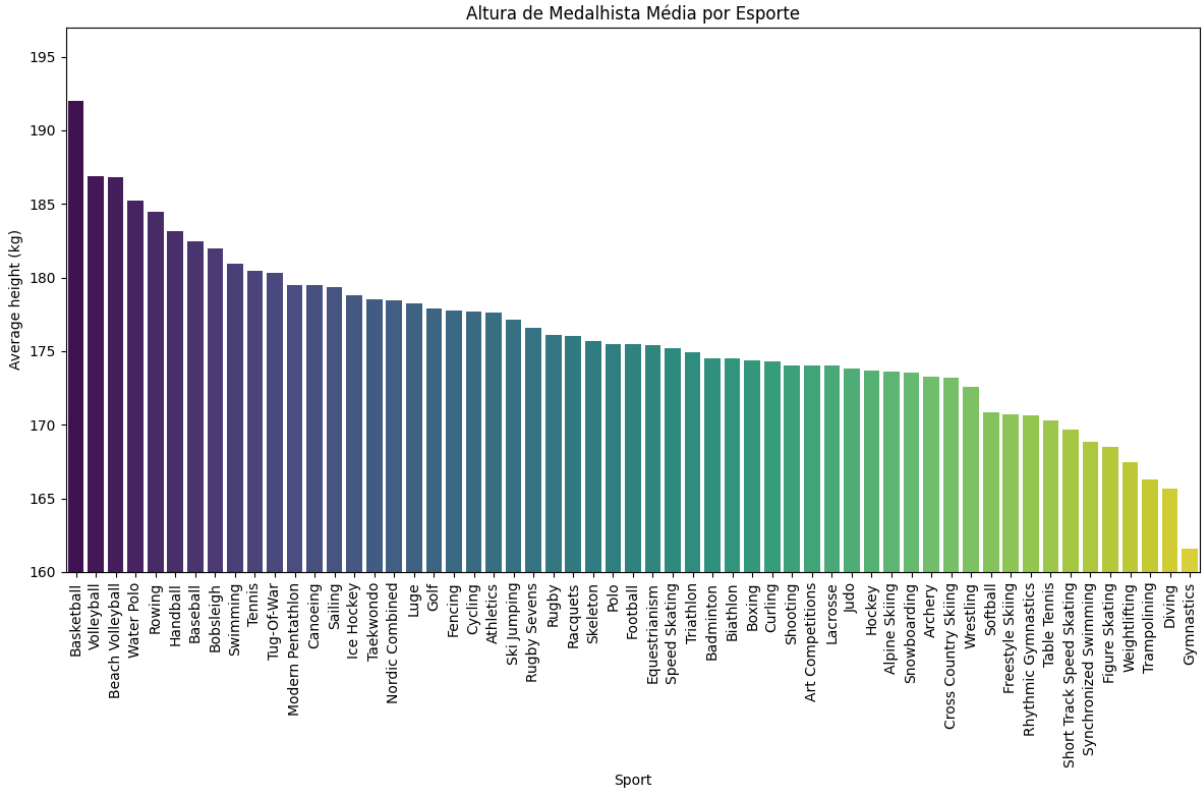

Figura 21: Altura média de atletas medalhistas por esporte

De maneira geral, a altura dos medalhistas por esporte acompanha as médias observadas na população total de atletas. No entanto, nos esportes com valores extremos, como **Basquete** e **Ginástica**, nota-se uma leve intensificação da tendência: a média de altura dos medalhistas em *Basketball* sobe para 191 cm, enquanto na *Gymnastics* diminui para 160 cm — reforçando o papel determinante do perfil físico ideal no desempenho esportivo de alto nível.

5. Aplicação da Técnica Estatística ou Preditiva

Nesta seção, foi realizada uma abordagem preditiva com o objetivo de estimar a **probabilidade de um atleta olímpico conquistar uma medalha**, a partir de suas características físicas e da modalidade em que compete. As variáveis utilizadas como preditoras foram: `sexo`, `idade`, `peso`, `altura` e `esporte`.

Dois modelos foram treinados e avaliados: o **Naive Bayes Gaussiano** e o **Random Forest** com `n_estimators=1000`. A variável alvo foi binária, indicando se o atleta havia ou não conquistado ao menos uma medalha.

5.1 Modelo: Naive Bayes Gaussiano

O primeiro modelo testado foi o Naive Bayes com suposição de distribuição gaussiana para as variáveis contínuas. Os resultados obtidos são apresentados abaixo:

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.99	0.92	35190
1	0.28	0.02	0.03	6043
accuracy			0.85	41233
macro avg	0.57	0.50	0.48	41233
weighted avg	0.77	0.85	0.79	41233

Matriz de Confusão (Naive Bayes):

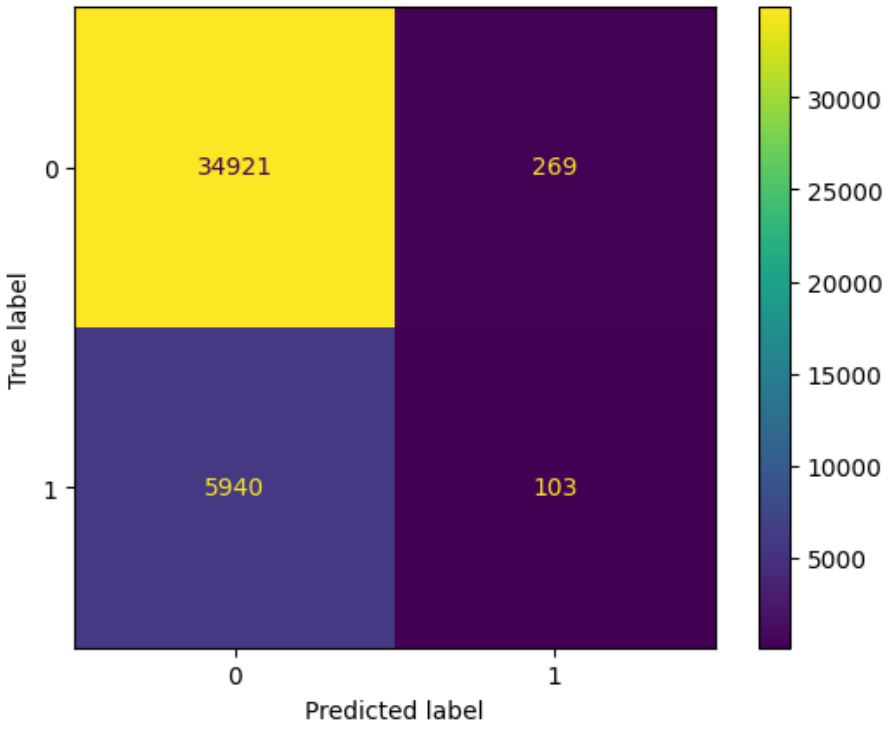

Figura 22: Matriz de confusão – Naive Bayes

AUC-ROC: 0.5947

Embora a acurácia global aparente seja alta (85%), ela é inflada pelo desbalanceamento das classes. O modelo praticamente não identifica corretamente os medalhistas, com *recall* de apenas 2%. O valor de AUC-ROC de 0,5947 indica que o modelo é apenas ligeiramente melhor do que o acaso (cuja AUC-ROC seria 0,5), sugerindo capacidade discriminativa bastante limitada.

5.2 Modelo: Random Forest

Em seguida, foi testado um modelo de **Random Forest** com 1000 estimadores. Os resultados foram consideravelmente superiores:

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.94	0.90	35190
1	0.34	0.17	0.23	6043
accuracy			0.83	41233
macro avg	0.60	0.56	0.57	41233
weighted avg	0.79	0.83	0.80	41233

Matriz de Confusão (Random Forest):

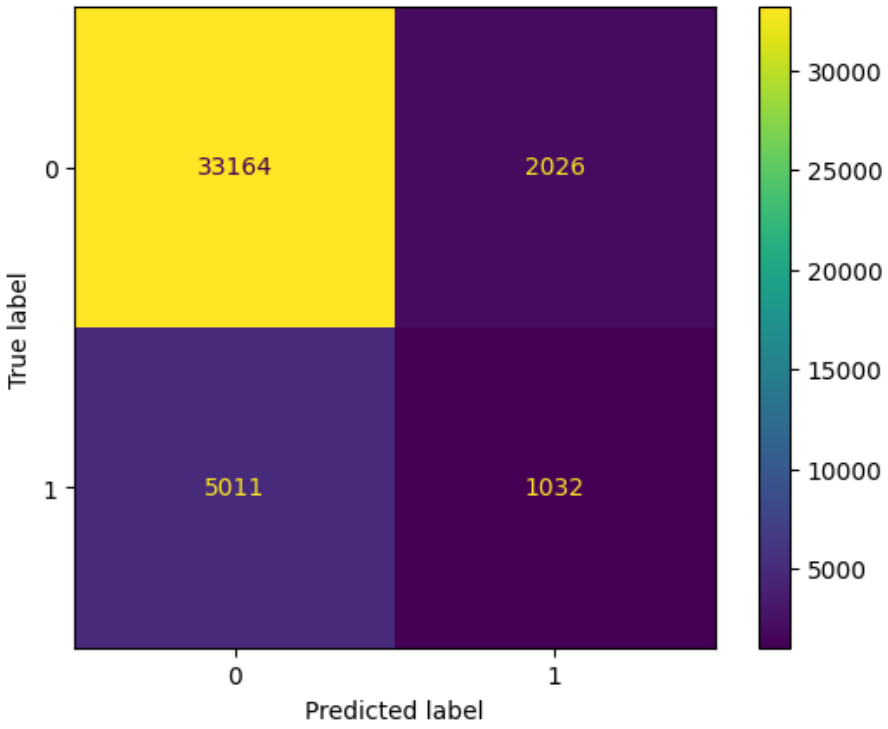

Figura 23: Matriz de confusão – Random Forest

AUC-ROC: 0.6805

Comparado ao Naive Bayes, o Random Forest apresentou ganhos consideráveis, especialmente na identificação de medalhistas, com *recall* subindo para 17%. O AUC-ROC de 0,6805 mostra que o modelo tem uma capacidade de discriminação entre medalhistas e não medalhistas **substancialmente superior ao acaso**, embora ainda longe do ideal para aplicações críticas.

A seguir, apresenta-se o gráfico com a **importância das variáveis** segundo o modelo Random Forest:

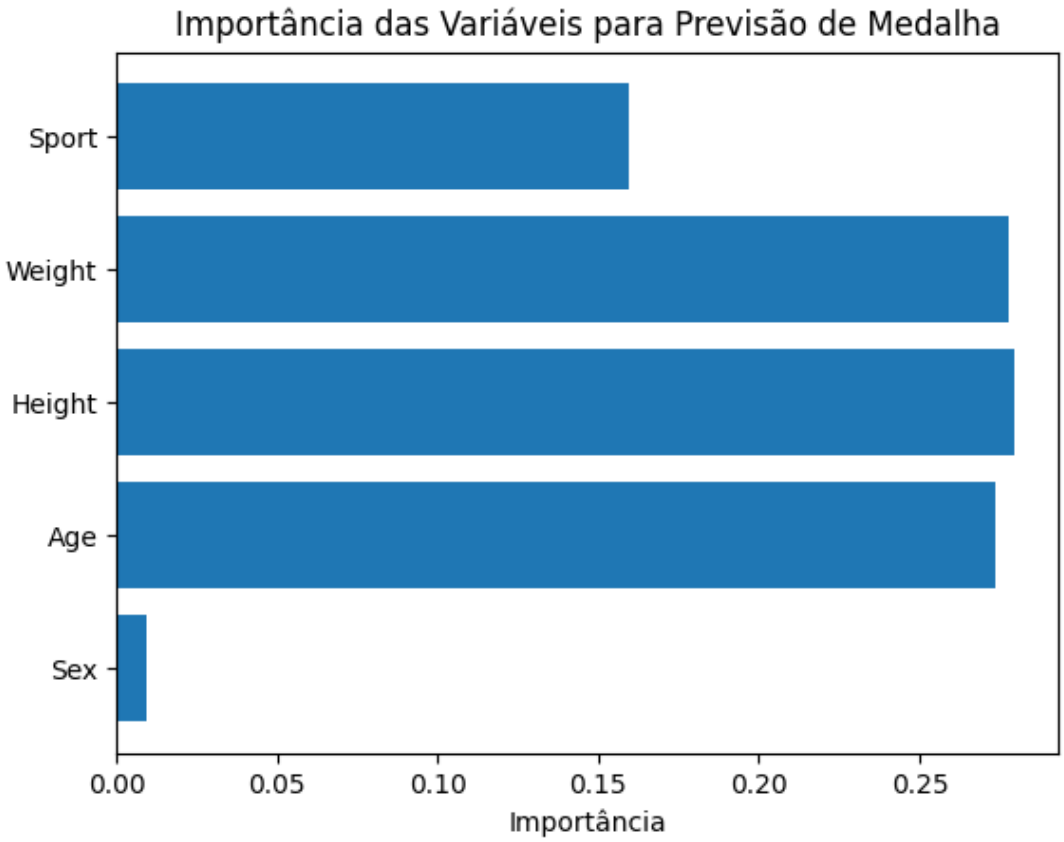

Figura 24: Importância das variáveis na predição de medalhas (Random Forest)

Nota-se que as variáveis físicas (`altura`, `peso` e `idade`) têm maior influência no modelo, seguidas pela variável `esporte`, que incorpora aspectos mais complexos e específicos de cada modalidade. A variável `sexo` teve a menor importância relativa entre as cinco consideradas.

5.3 Considerações sobre os modelos

Os resultados mostram que prever o sucesso olímpico com base em características físicas básicas é uma tarefa desafiadora. Fatores como histórico de treinos, suporte institucional, performance em competições anteriores e aspectos psicológicos não estão disponíveis no dataset, mas impactam diretamente o desempenho esportivo.

Mesmo assim, o modelo Random Forest demonstrou melhor desempenho e maior robustez frente à simplicidade do Naive Bayes. O uso de técnicas de balanceamento de classes, modelos mais sofisticados e enriquecimento do conjunto de dados com variáveis externas poderia melhorar significativamente o poder preditivo dos modelos em trabalhos futuros.

6. Discussão dos Resultados Obtidos

A análise exploratória revelou padrões consistentes no perfil físico e demográfico dos atletas olímpicos ao longo do tempo. Verificou-se que a maioria dos atletas possui idade entre 20 e 30 anos, peso entre 60 kg e 80 kg e altura entre 165 cm e 185 cm. Além disso, esportes

como atletismo, natação e ginástica se destacaram tanto em volume de participantes quanto em diversidade de características físicas dos atletas.

Apesar da evolução da participação feminina desde meados do século XX, especialmente após 1994, ainda há predominância de atletas do sexo masculino em muitas modalidades. No que diz respeito às medalhas, observou-se leve predominância de medalhas de ouro em relação às de bronze e prata, o que pode ser atribuído a empates e particularidades históricas.

As análises por esporte mostraram que o perfil físico ideal varia significativamente entre modalidades. Esportes como basquete, levantamento de peso e remo concentram atletas com maiores médias de altura e peso, enquanto ginástica e esportes técnicos concentram atletas mais leves e baixos.

Na abordagem preditiva, os modelos estatísticos enfrentaram dificuldades para classificar corretamente atletas medalhistas. O modelo de *Naive Bayes Gaussiano* apresentou baixa sensibilidade e desempenho próximo ao acaso. O modelo de *Random Forest*, por outro lado, apresentou melhor desempenho, com AUC-ROC de aproximadamente 0,68 e destaque para as variáveis altura, peso e idade como as mais relevantes.

Esses resultados sugerem que, embora exista um perfil físico comum entre atletas de elite, o simples conhecimento de características como sexo, idade, peso, altura e esporte não é suficiente para prever com precisão a conquista de medalhas. Fatores externos — como treinamento, experiência, investimento esportivo e condições competitivas — não estão representados no conjunto de dados e têm grande influência sobre o resultado.

7. Considerações Finais

O presente trabalho teve como objetivo explorar estatisticamente os dados históricos dos Jogos Olímpicos, tanto de forma descritiva quanto preditiva, com foco no desempenho e perfil físico dos atletas. A análise descritiva permitiu identificar padrões relevantes relacionados ao gênero, modalidades esportivas e evolução histórica das características físicas dos atletas.

A tentativa de prever a conquista de medalhas por meio de modelos de classificação demonstrou os limites de abordagens baseadas exclusivamente em atributos físicos e esportivos. Embora o modelo de *Random Forest* tenha apresentado desempenho razoável, o problema é altamente influenciado por variáveis contextuais ausentes no dataset.

Em síntese, o trabalho evidenciou o potencial e os limites da estatística descritiva e preditiva na análise de dados esportivos. A combinação dessas abordagens permitiu não apenas observar tendências históricas relevantes, mas também refletir sobre a complexidade de prever o sucesso no esporte de alto rendimento com base em dados disponíveis.

Trabalhos futuros podem explorar outras fontes de informação, aplicar técnicas de balanceamento de classes, incorporar dados temporais (como performance por edição dos Jogos) ou utilizar modelos mais sofisticados, como redes neurais e embeddings para variáveis categóricas, com o objetivo de melhorar a capacidade preditiva das análises.