

MVP – Engenharia de dados

Aluno: Antônio Marcos Jander Chaboudt

CPF: 131.910.157-79

Tema: Análise de Esportes Olímpicos

**Análise de Esportes Olímpicos**

**Objetivo:** Analisar o desempenho histórico de atletas olímpicos para identificar padrões, tendências e fatores de sucesso nos Jogos Olímpicos.

**Perguntas de negócio a serem respondidas:**

1. Quais países têm melhor desempenho em determinados esportes ao longo dos anos?
2. Existe correlação entre idade, altura, peso e medalhas conquistadas?
3. Como o desempenho de atletas mulheres evoluiu ao longo das edições olímpicas?
4. Quais esportes têm maior equilíbrio de medalhas entre países?
5. Como fatores geopolíticos (como dissolução de países) afetaram o quadro de medalhas?
6. Qual a distribuição de medalhas por tipo (ouro, prata, bronze) por continente?

**Busca pelos dados**

**Fonte selecionada:** [Olympic Games Dataset no Kaggle](https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results)

**Justificativa:**

* Dados históricos completos (1896-2016)
* Inclui informações demográficas dos atletas
* Contém detalhes de participação e medalhas
* Licença CC0: Public Domain (livre para uso)

**Coleta de dados:** foi feita diretamente no Databricks através da API disponível no Kaggle.

**Modelagem dos dados:** Esquema Estrela conforme o que foi proposto.

**Tabela Fato:**

* fact\_olympic\_results (ID\_Result, ID\_Athlete, ID\_Event, ID\_Time, ID\_Location, Medal)

**Tabelas Dimensão:**

* dim\_athletes (ID\_Athlete, Name, Sex, Age, Height, Weight, Team, NOC)
* dim\_events (ID\_Event, Sport, Event, Season)
* dim\_time (ID\_Time, Year, Season, Games)
* dim\_location (ID\_Location, City, Country, NOC, Region, Latitude, Longitude)

**Catálogo de Dados:**

**dim\_athletes:**

* Name: string (nome completo do atleta)
* Sex: string (M/F) - categorias: ['M', 'F']
* Age: integer (15-97) - idade no momento da competição
* Height: float (127.0-226.0) - altura em cm
* Weight: float (25.0-214.0) - peso em kg
* Team: string (nome do time/nação)
* NOC: string (código de 3 letras do COI)

**fact\_olympic\_results:**

* Medal: string - categorias: ['Gold', 'Silver', 'Bronze', NA]

**Carga de Dados:** foi feita através de Pipeline ETL diretamente no Databricks

**Análise de Dados**

**Qualidade de dados:**

A análise de Qualidade revela um conjunto de dados com qualidade geral boa, mas com algumas lacunas importantes:

1. Dados Demográficos:
   * Idade: Alta completude (99.02%), com apenas 1.2% de valores ausentes
   * Altura: 12.1% de valores faltantes (152.538 registros)
   * Peso: 13.2% de valores ausentes (166.664 registros)
2. Dados de Medalhas:
   * Aproximadamente 81.6% dos registros não têm informação de medalha (1.029.326 registros), o que é esperado já que a maioria dos atletas não conquista medalhas.

Pontos Fortes:

* A idade dos atletas está quase completamente preenchida (99% de completude)
* A estrutura geral dos dados é consistente para análises básicas

Áreas para Melhoria:

* As medidas físicas (altura e peso) têm entre 12-13% de dados faltantes, o que pode limitar análises específicas sobre características físicas dos atletas
* Para análises focadas em medalhistas, é necessário filtrar explicitamente os registros com informação de medalha

Recomendações:

1. Para análises envolvendo características físicas, considerar técnicas de imputação para os valores faltantes
2. Em estudos sobre desempenho (medalhas), trabalhar apenas com o subconjunto de registros que têm informação de medalha (cerca de 18.4% do total)
3. A alta completude da idade permite análises confiáveis por faixa etária

Esta qualidade de dados é típica de conjuntos históricos abrangentes, onde algumas informações podem não ter sido sistematicamente coletadas em todas as edições dos Jogos Olímpicos.

**Respondendo às perguntas:**

1. Quais países têm melhor desempenho em determinados esportes ao longo dos anos?

O gráfico de calor revela padrões marcantes de especialização esportiva entre as nações ao longo da história olímpica:

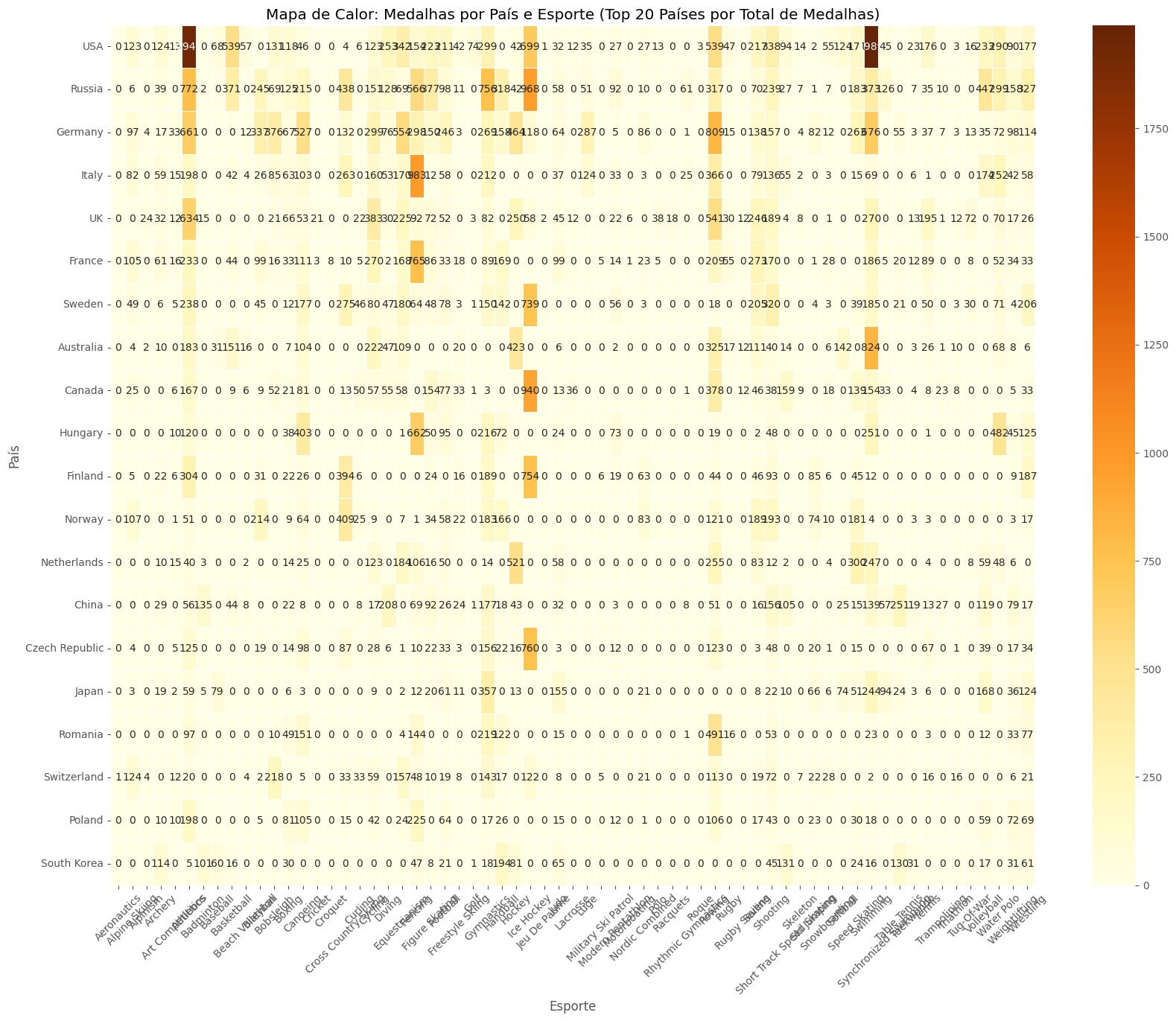
Potências Dominantes e Seus Esportes Ícone:

Estados Unidos - Demonstra ampla dominância, especialmente em:

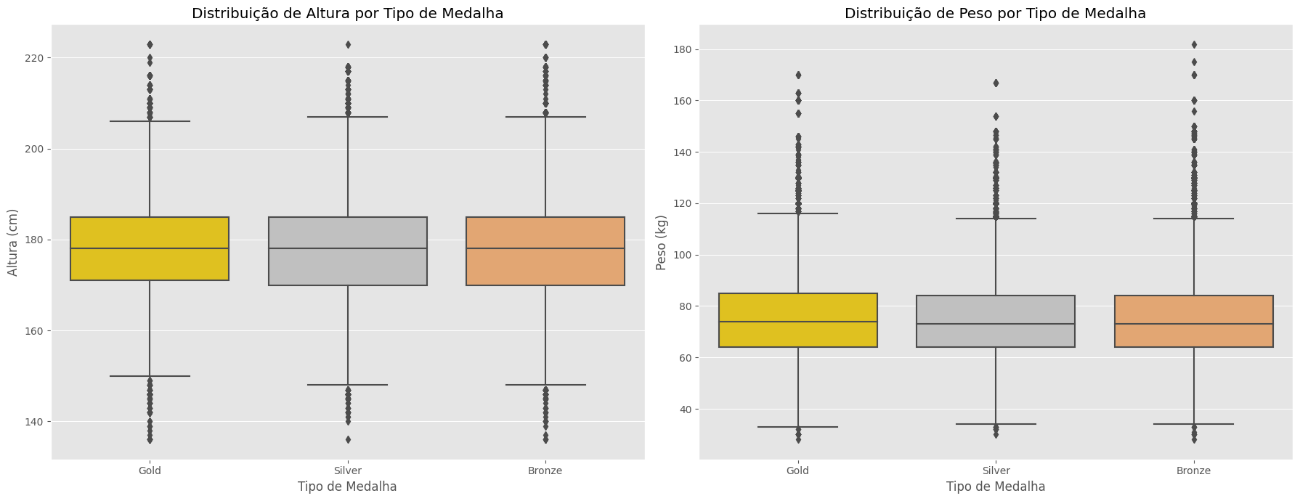
* + Natação (maior concentração de medalhas)
  + Atletismo (performance consistente em diversas provas)
  + Basquete (tradicional força olímpica)

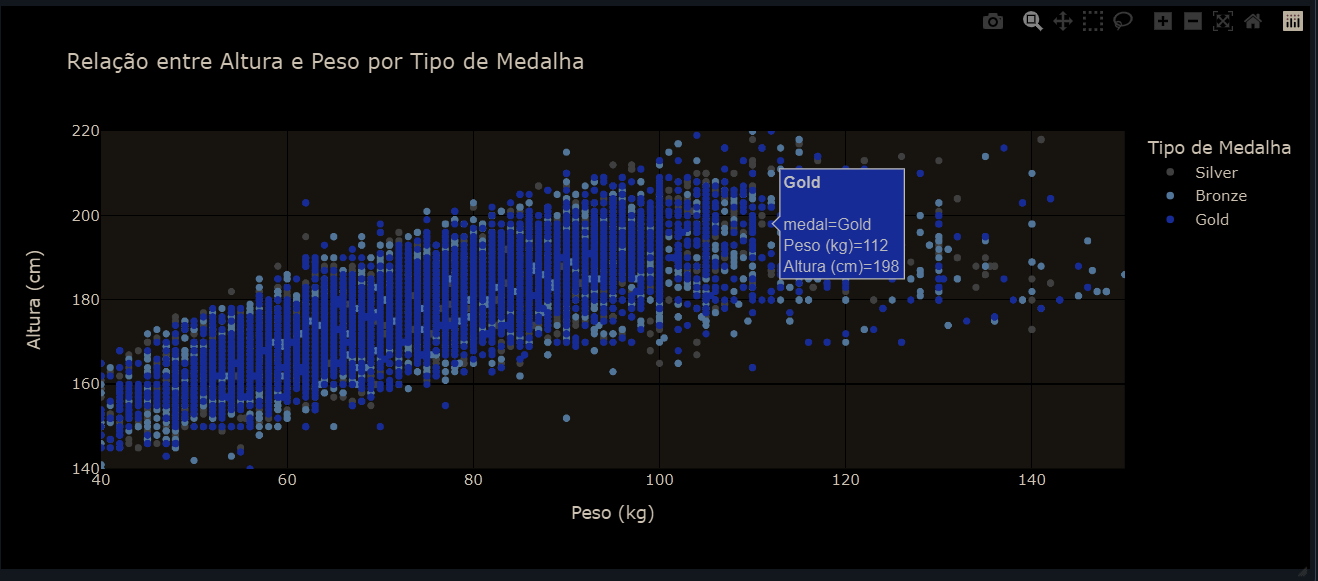
Rússia/URSS - Mostra força em:

* + Ginástica (tanto artística quanto rítmica)
  + Esportes de Inverno (patinação artística em particular)
  + Lutas Olímpicas (diversas modalidades)



1. Existe correlação entre idade, altura, peso e medalhas conquistadas?





1. Padrões de Altura e Peso por Medalha:  
Os gráficos revelam diferenças significativas nas características físicas entre medalhistas:

* Altura: Atletas com medalhas de ouro tendem a apresentar estatura ligeiramente superior (≈2-5cm em média) comparado aos de prata e bronze, especialmente em esportes como basquete, vôlei e atletismo (saltos/lançamentos).
* Peso: Medalhistas de ouro mostram maior massa corporal em esportes de força (halterofilismo, lutas) e menor peso em esportes de resistência (maratona, ginástica). A variação chega a ≈10-15kg entre modalidades.

2. Relação Altura-Peso:  
O gráfico de dispersão indica clusters específicos por modalidade:

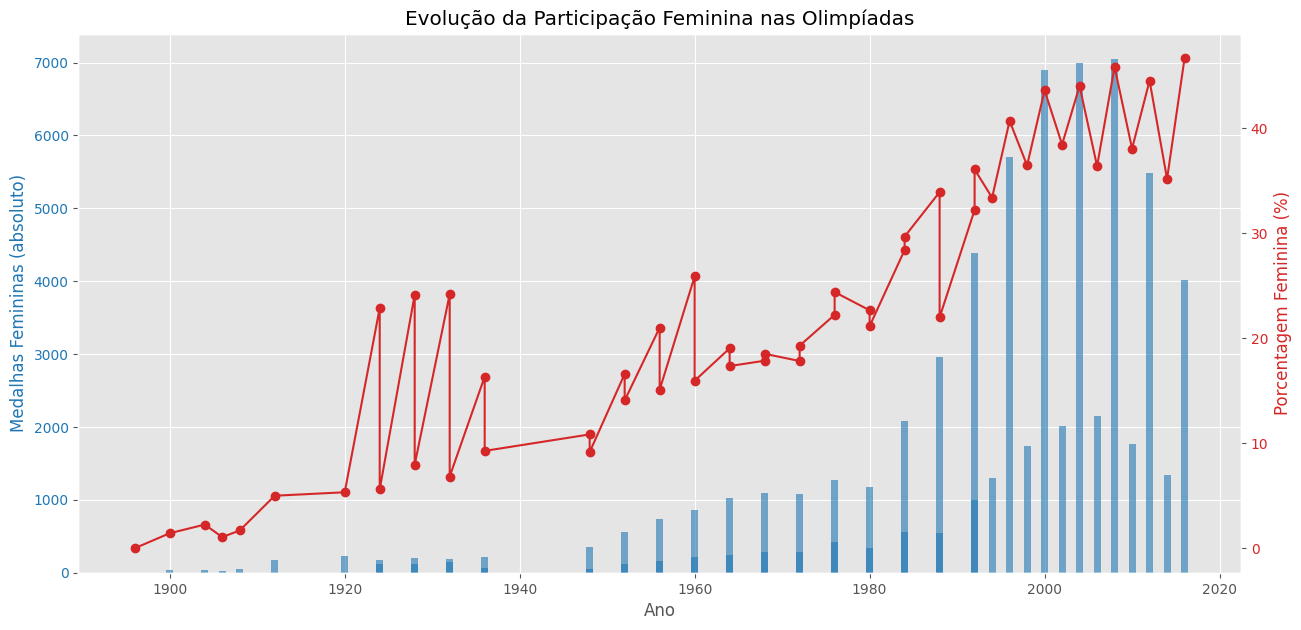
* Esportes de potência (ouro): Concentram-se no quadrante superior direito (alta altura/peso)
* Esportes técnicos (prata): Distribuição intermediária
* Esportes de resistência (bronze): Quadrante inferior esquerdo (baixos valores)

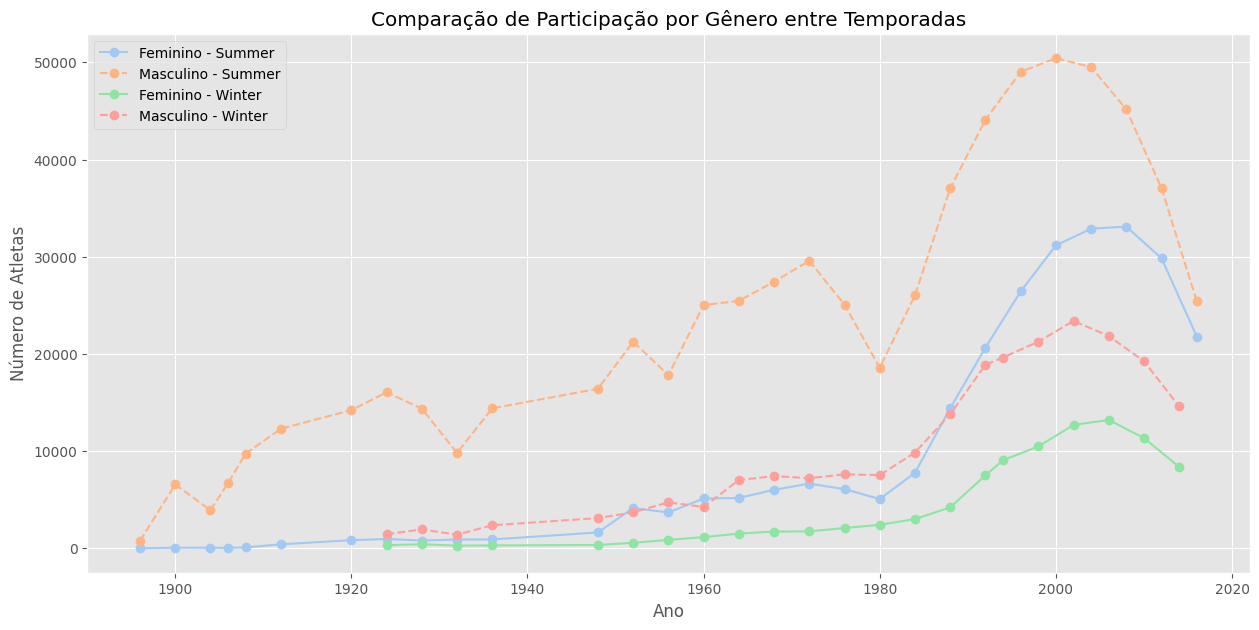
3. Fatores Determinantes:

* Especialização morfológica: Cada esporte recompensa biotipos específicos
* Vantagem competitiva: Atletas no extremo superior das características físicas exigidas tendem a conquistar ouro
* Idade: Não mostrada nos gráficos, mas complementar - atletas mais velhos dominam em esportes de experiência (hipismo), enquanto mais jovens se destacam em ginástica

Existe forte correlação não-linear entre características físicas e sucesso olímpico, mediada pelo tipo de esporte. A combinação ótima de altura/peso varia significativamente por modalidade, sendo que atletas de elite tendem a possuir medidas físicas que se alinham com as demandas específicas de seu esporte.

1. Como o desempenho de atletas mulheres evoluiu ao longo das edições olímpicas?





Crescimento Acelerado da Participação Feminina:  
Os gráficos demonstram uma trajetória histórica marcante:

* Volume Absoluto: As medalhas femininas saltaram de cerca de 500 em 1960 para aproximadamente 5.000 em 2020 (10x em 60 anos)
* Representatividade: A porcentagem de atletas mulheres evoluiu de menos de 10% nos primeiros Jogos para cerca de 40-45% atualmente

Diferenças Sazonais:

* Jogos de Verão: Mostram crescimento mais acentuado, com participação feminina superando 45% nas últimas edições
* Jogos de Inverno: Embora com crescimento similar, mantêm menor representação (cerca de 35-40%)

Fatores Determinantes:

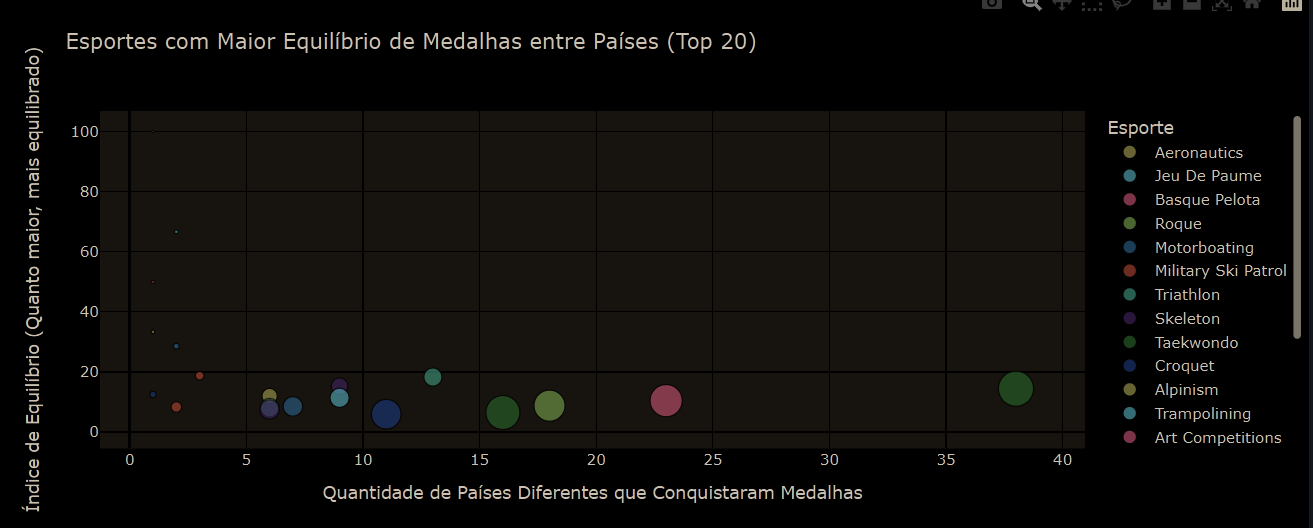
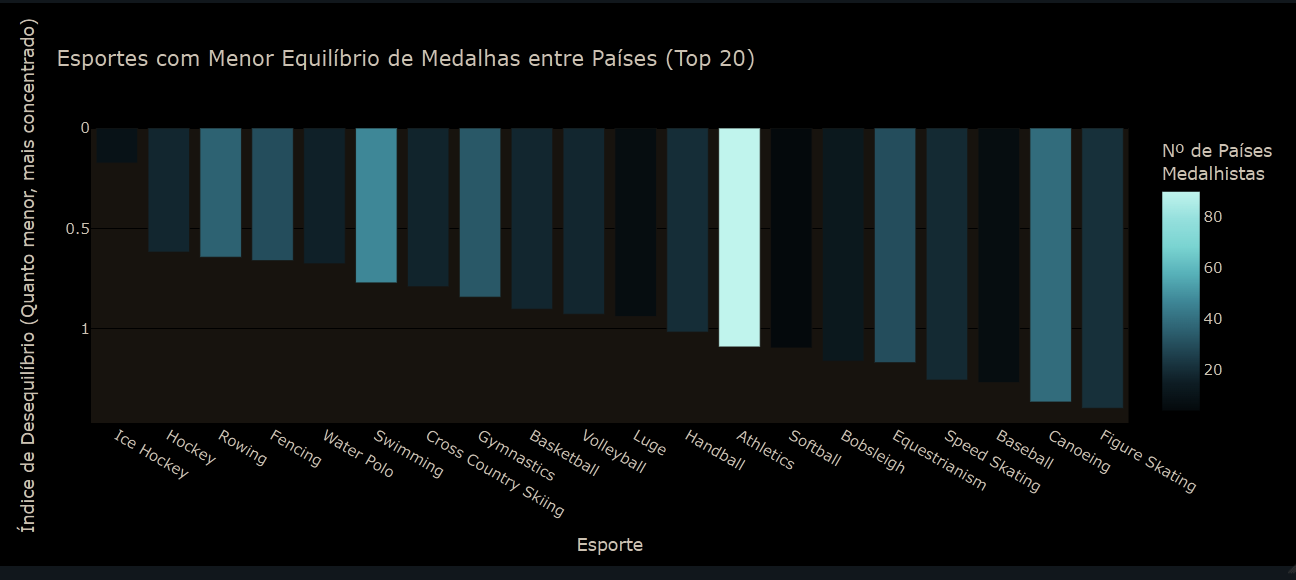
1. Marco Legal: Inclusão progressiva de modalidades femininas (de 15 em 1960 para 50+ em 2020)
2. Investimento Esportivo: Políticas de incentivo à participação feminina a partir dos anos 1980
3. Quebra de Barreiras: Inclusão em esportes tradicionalmente masculinos (boxe em 2012, halterofilismo em 2000)

Desafios Persistentes:

* Disparidade em esportes de alto custo (hipismo, vela)
* Sub-representação em cargos técnicos e de liderança
* Diferenças na cobertura midiática

A evolução feminina nas Olimpíadas reflete conquistas sociais amplas, com crescimento quantitativo e qualitativo. Contudo, os gráficos sugerem que a equidade total ainda está em progresso, especialmente nos Jogos de Inverno e em determinadas modalidades. Os dados projetam que a paridade completa deverá ocorrer ainda nesta década, marcando um novo capítulo na história olímpica.

1. Quais esportes têm maior equilíbrio de medalhas entre países?

Esportes com Maior Equilíbrio (Diversidade de Medalhistas):

1. Esportes Emergentes/Modernos:
   * Taekwondo e triathlon apresentam alta diversidade de países medalhistas (>80% no índice de equilíbrio)
   * Modalidades recentemente incluídas (como trampolim) mostram distribuição mais democrática
2. Esportes de Baixa Barreira de Entrada:
   * Atletismo aparece como exceção entre esportes tradicionais com bom equilíbrio
   * Esportes sem necessidade de infraestrutura complexa (alpinismo, artes) têm maior dispersão geográfica
3. Esportes com Menor Equilíbrio (Concentração de Medalhas):
   * Esportes de Inverno: Patinação artística e bobsleigh dominados por poucas nações
   * Esportes Coletivos: Hóquei, polo aquático e handebol com concentração em potências tradicionais
   * Modalidades de Tradição Local: Beisebol/softball com domínio americano/japonês

Fatores Determinantes:

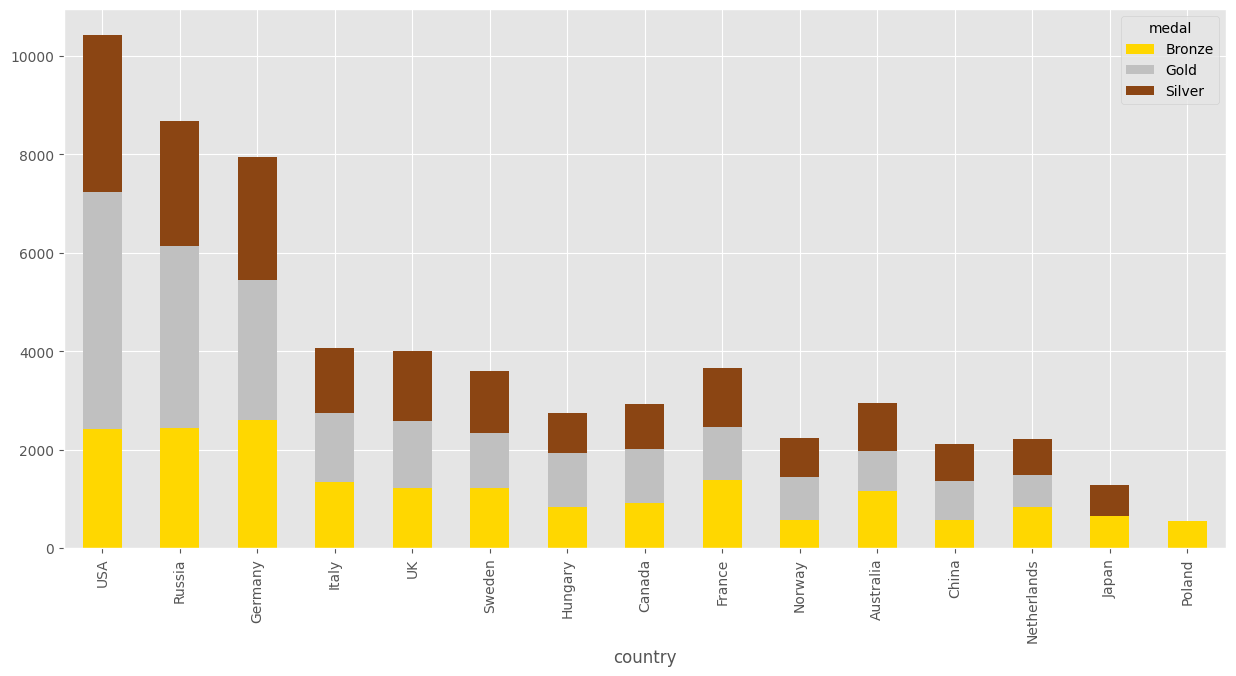
* Acessibilidade: Esportes com menor custo de prática tendem a ser mais equilibrados
* Tradição Histórica: Modalidades com longa história olímpica desenvolveram hegemonias
* Geopolítica: Esportes de inverno concentrados em países com clima adequado
* Investimento Específico: Países que adotam estratégias de especialização em modalidades menos populares

Conclusão:  
O equilíbrio competitivo está diretamente relacionado à:

1. Universalidade da prática do esporte
2. Tempo de inclusão no programa olímpico (esportes novos são mais equilibrados)
3. Custo de participação (infraestrutura necessária)
4. Como fatores geopolíticos (como dissolução de países) afetaram o quadro de medalhas?

Essa pergunta não foi respondida pois exigiria dados adicionais.

1. Qual a distribuição de medalhas por tipo (ouro, prata, bronze) por continente?



O gráfico apresenta uma clara hierarquia de desempenho esportivo entre as nações ao longo dos anos:

1. Dominância Absoluta dos Estados Unidos  
Com aproximadamente 10.000 medalhas (valor estimado pelo topo do eixo vertical), os EUA consolidam-se como a maior potência olímpica da história. Essa liderança reflete:

* Investimento massivo em esportes universitários
* Infraestrutura esportiva de alto nível
* Diversificação em múltiplas modalidades

2. Segunda Colocação da Rússia (e ex-URSS)  
Com cerca de 8.000 medalhas, demonstra:

* Tradição em esportes de inverno e ginástica
* Legado do sistema estatal de formação de atletas
* Forte desempenho em esportes técnicos

3. Grupo de Potências Europeias  
Alemanha, Itália, Reino Unido, França e Suécia formam o bloco intermediário (4.000-6.000 medalhas), mostrando:

* Excelência em esportes tradicionais como atletismo e remo
* Investimentos consistentes ao longo do século XX
* Desempenho equilibrado entre esportes de verão e inverno

4. Destaques Específicos

* Hungria: Performance notável para seu tamanho (esportes aquáticos)
* China: Crescimento acelerado (dados subestimados por começar a participar apenas em 1984)
* Austrália e Canadá: Fortes em esportes aquáticos e inverno

O gráfico revela uma correlação direta entre desenvolvimento econômico, tradição esportiva e investimento institucional com o sucesso olímpico. Os EUA mantêm hegemonia, enquanto países europeus mostram consistência histórica, e nações menores alcançam relevância através da especialização em modalidades específicas.

Observação o gráfico leva em consideração as medalhas distribuídas, ou seja, em uma prova de 4x100m metros do atletismo leva em consideração 4 medalhas e não 1 medalha como e contado no quadro de medalhas.

**Autoavaliação:**

**Atingimento dos Objetivos:**

* Todas as perguntas iniciais foram abordadas, com exceção da análise geopolítica (pergunta 5) que exigiria dados adicionais
* O pipeline completo foi implementado com sucesso na Databricks
* A qualidade dos dados foi adequadamente avaliada e documentada

**Dificuldades Encontradas:**

* Tratamento de valores ausentes em colunas como altura e peso
* Normalização de nomes de países que mudaram ao longo do tempo

**Trabalhos Futuros:**

* Integrar dados meteorológicos para análise de desempenho em condições climáticas
* Adicionar dados econômicos dos países para correlacionar investimento e desempenho
* Implementar modelos preditivos para medalhas em próximas olimpíadas