

Olympics Dataset

Equipe: Liga Olímpica

Integrantes:

RA 213374 - Áureo Henrique e Silva Marques

RA 176566 - José Alexandre dos Santos Barros

RA 220407 - Lindon Jonathan Sanley dos Santos Pereira Monroe

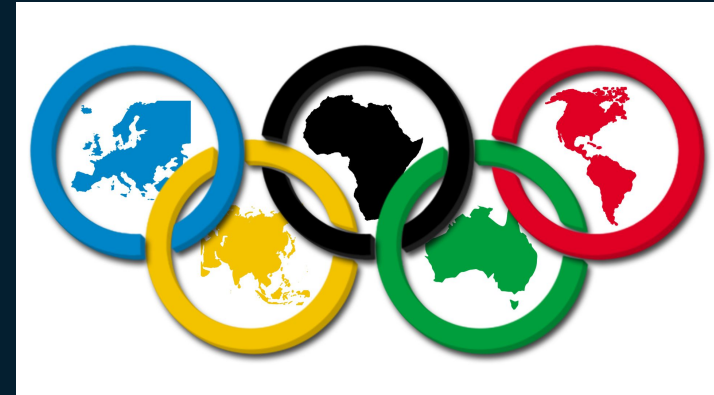
Tema

- Histórico dos Jogos Olímpicos nos últimos anos.
- Os Jogos Olímpicos, ou Olimpíadas, são o maior evento esportivo do mundo e, de 4 em 4 anos, reúnem milhares de atletas de vários países. Embora suas origens sejam da Grécia Antiga, as primeiras Olimpíadas ocorreram oficialmente em 1896, organizadas pelo Comitê Olímpico Internacional (COI) e, portanto, desse ano até hoje, tratam-se de mais de 30 edições dos jogos olímpicos.

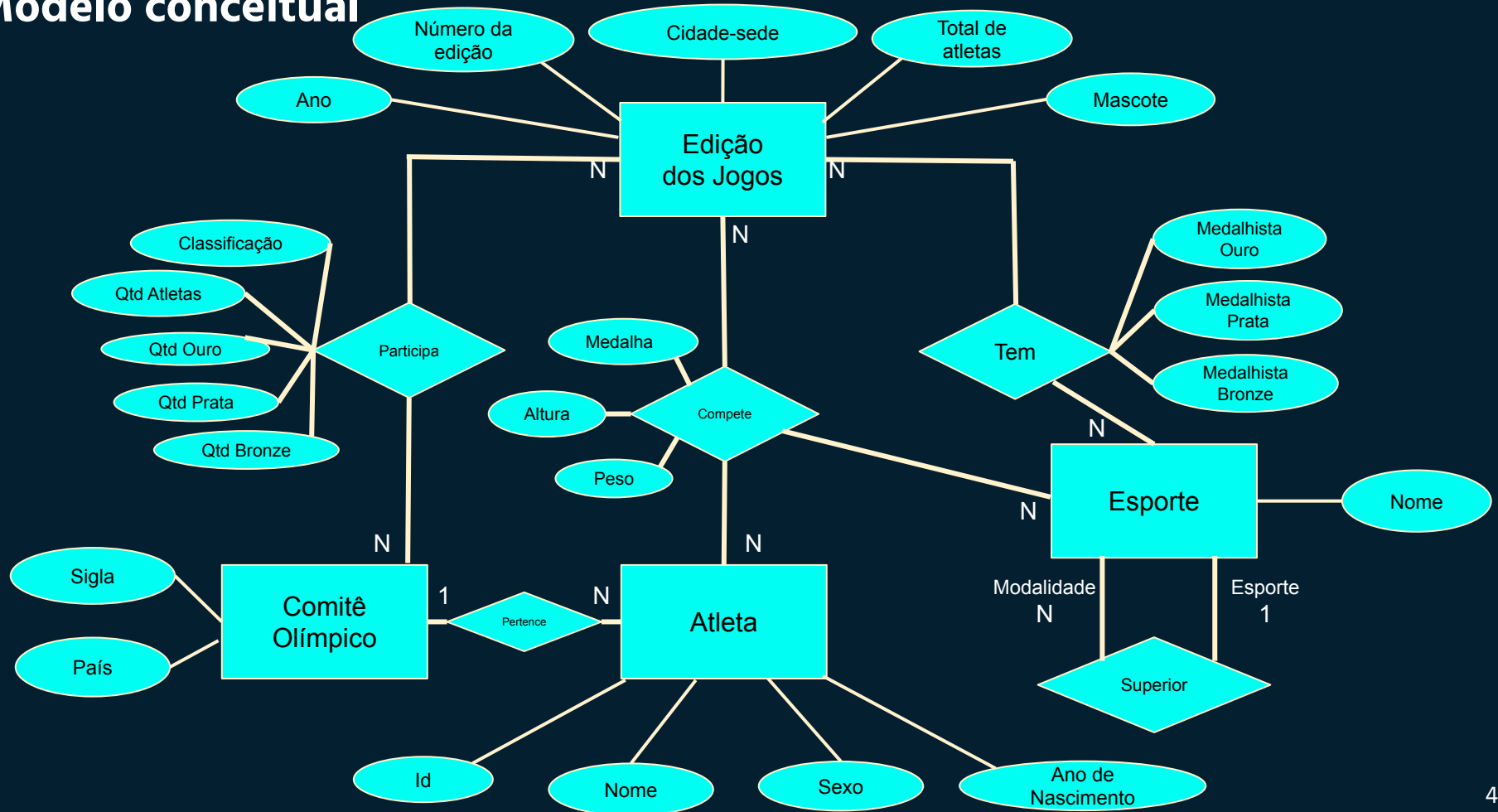


Tema

- Grande quantidade de informações sobre: atletas, países participantes, cidades-sede, medalhas, esportes, vencedores de cada modalidade, entre outros.
- É difícil encontrar bancos de dados atuais que possuam essas informações de forma centralizada e bem organizada.
- **Objetivo:** construir um dataset sobre os Jogos Olímpicos dos últimos anos que seja organizado e bem estruturado, permitindo diversos tipos de análises sobre o tema.



Modelo conceitual



Modelos lógicos

- **Modelo relacional:** O modelo lógico relacional servirá para tratamento e melhor organização dos dados obtidos nas tabelas das fontes. Além disso, diversas análises estatísticas e comparativas são possíveis utilizando dados estruturados.
- **Modelo de documentos:** O modelo de documentos ajudará a encapsular melhor as informações sobre cada edição, através da hierarquia de elementos. Com ele, será possível obter informações mais diretas sobre um atleta ou um esporte em uma determinada olimpíada.

Modelo lógico: Relacional

EdicaoDosJogos (Ano, NumeroDaEdicao, CidadeSede, TotalDeAtletas, Mascote)

Atleta (Id, Nome, AnoDeNascimento, Sexo)

ComiteOlimpico(Sigla, País)

EsporteModalidade(Id, Nome, EsportePai)

Modelo lógico: Relacional

ParticipacaoComites(IdComite, AnoEdicao, QtdAtletas, QtdOuro , QtdPrata , QtdBronze, Classificacao)

- Chaves estrangeiras: IdComite -> Comiteloimpico(Sigla), AnoEdicao -> EdicaoDosJogos(Ano)

ParticipacaoAtletas(IdAtleta, AnoEdicao, IdModalidade, Altura, Peso, Medalha)

- Chaves estrangeiras: IdAtleta -> Atleta(Id), AnoEdicao -> EdicaoDosJogos(Ano),
IdModalidade -> EsporteModalidade(Id)

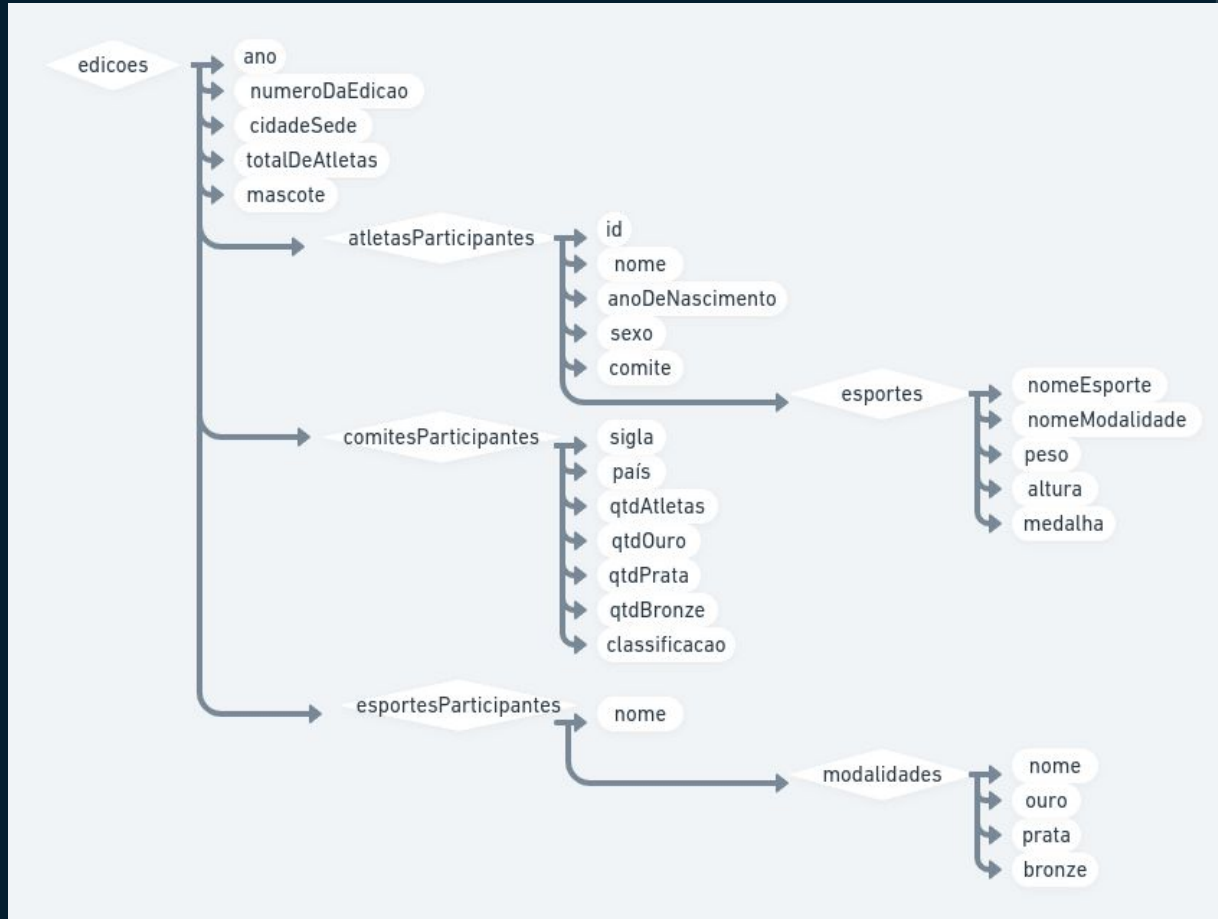
ComiteDosAtletas(IdComite, IdAtleta)

- Chaves estrangeiras: IdComite -> Comiteloimpico(Sigla), IdAtleta -> Atleta(Id)

EsportesDasEdicoes(AnoEdicao, IdModalidade, Ouro, Prata, Bronze)

- Chaves estrangeiras: AnoEdicao -> EdicaoDosJogos(Ano), IdModalidade -> EsporteModalidade(Id)

Modelo lógico: Documentos



Fontes de dados

As fontes de dados que utilizamos no dataset são as seguintes:

- “120 years of Olympic history: athletes and results”

<https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results/discussion/69221>

Dataset histórico, uma tabela com dados das olimpíadas de 1896 a 2016. Foi criado a partir de dados do site www.sports-reference.com.

- 2021 Olympics in Tokyo


<https://www.kaggle.com/arjunprasadsarkhel/2021-olympics-in-tokyo>

Dataset que consiste em uma tabela com dados específicos das olimpíadas de Tóquio em 2021.

- <https://olympics.com>


Site oficial do Comitê Olímpico Internacional (IOC) contendo uma base extensa de dados, notícias e informações sobre os Jogos Olímpicos e seus envolvidos, em geral.


Name	Sex	Age	Height	Weight	Team	NOC									
134732 unique values	M	73%	23	8%	NA	22%	NA	23%	United States	7%	USA				
	F	27%	24	8%	180	5%	70	4%	France	4%	FRA				
			Other (227521)		84%	Other (198453)		73%	Other (198816)		73%	Other (241281)		89%	Other (239)
A. Dijiang	M		24		188		88		China		CHN				
A. Lamusi	M		23		178		68		China		CHN				
Gunnar Nielsen Aaby	M		24		NA		NA		Denmark		DEN				


 Olympic Games Athletes Sports News • Olympic Channel


Athletes


Search for an athlete



SIMONE BILES
USA, ARTISTIC G...



MICHAEL PHELPS
USA, SWIMMING


PUSARLA VENKA...
IND, BADMINTON


USAIN BOLT
JAM, ATHLETICS


NAOMI OSAKA
JPN, TENNIS


KATIE LEDECKY
USA, SWIMMING


YUI HIRONAKA
JPN, F...

Operações aplicadas aos bancos

- **Extração:** Usada para complementação dos dados dos datasets estruturados, extraindo dados do site <http://olympics.com>.
- **Integração:** O dataset integra dados das nossas 3 fontes principais e de mais algumas auxiliares.
- **Tratamento:** Foram tomadas medidas em relação a dados faltantes, como anos de nascimento e sexo dos atletas.
- **Transformação:** Os dados foram transformados de forma a obtermos as tabelas especificadas no modelo lógico a partir dos dados brutos, tornando a análise mais prática e eficiente.

Construção do modelo relacional

Tratamento do Dataset:

“120 years of Olympic history: athletes and results”

1. Leitura do CSV do dataset.
2. Seleção apenas dos jogos de verão.

```
import pandas as pd
```

```
1 atletas_120 = pd.read_csv('athlete_events.csv')
```

```
2 atletas_120=atletas_120.loc[atletas_120["Season"] != 'Winter']
```

```
atletas_120=atletas_120.reset_index(drop=True)
```

```
atletas_120["Id"] = atletas_120.index
```

```
display(atletas_120)
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer

"120 years of Olympic history: athletes and results"

Tabela 1 - Edição dos Jogos				
Ano	NumeroDaEdicao	CidadeSede	TotalDeAtletas	Mascote

Podemos ver o padrão de criação das tabelas:

1. Seleção das colunas desejadas que estão no dataset fonte.
2. Renomeação adequada das colunas.
3. Criação de novas colunas vazias.
4. Preenchimento de dados através de função e lógica, nesse caso, o número de atletas em uma certa edição é calculado selecionando-se os atletas de um certo ano e eliminando os atletas repetidos.
5. Preenchimento de dados nulos com hífen.

```
edicaoDosJogos=pd.DataFrame(data=atletas_120[['Year', 'City']])
edicaoDosJogos=edicaoDosJogos.drop_duplicates("Year")
edicaoDosJogos=edicaoDosJogos.sort_values(by=["Year"])
edicaoDosJogos = edicaoDosJogos.reset_index(drop=True) #Reseta os index da tab
edicaoDosJogos = edicaoDosJogos.rename({"Year": "Ano"}, axis=1)
edicaoDosJogos = edicaoDosJogos.rename({"City": "Cidade"}, axis=1)
edicaoDosJogos["NumerodaEdicao"] = None
edicaoDosJogos["TotaldeAtletas"] = None
```

```
for i in range(len(edicaoDosJogos.index)):
    edicaoDosJogos.at[i, 'NumerodaEdicao']=1+i
```

```
edicaoDosJogos["Mascote"]=[ ' - ', ' - ', ' - ', ' - ', ' - ', ' - ', ' - ', ' - ', ' - ', ' - ', ' - ', ' - '
```

```
# Função que ajuda a encontrar a qtd de atletas
def setNumeroDeAtletas (row):
    atletasDeUmAno=pd.DataFrame(data=atletas_120[['Year', 'Name']])
    atletasDeUmAno=atletasDeUmAno.loc[atletasDeUmAno["Year"] ==row.Ano]
    atletasDeUmAno=atletasDeUmAno.drop_duplicates("Name")
    row.TotalAtletas=len(atletasDeUmAno.index)
    return row
```

```
edicaoDosJogos = edicaoDosJogos.apply(lambda x: setNumeroDeAtletas(x),axis=1)
anos = edicaoDosJogos['Ano'].tolist()
```

```
edicaoDosJogos=edicaoDosJogos.fillna("-")  
display(edicaoDosJogos)
```

Tratamento do Dataset: "2021 Olympics in Tokyo"

- Fornece dados básicos sobre os atletas e sobre as medalhas de cada comitê.
- Dataset com menos informações. Não possui modalidades dos esportes nem informações como ano de nascimento, sexo, altura e peso dos atletas.
- Necessidade de deixá-lo como um dataset separado.

Atletas

```
#Comando para carregar o dataset. Se o seu dataset for em  
atletas_original = pd.read_excel('Athletes.xlsx')  
  
#Comando 'head' mostra as 5 primeiras linhas da tabela  
atletas_original.head()
```

	Name	NOC	Discipline
0	AALERUD Katrine	Norway	Cycling Road
1	ABAD Nestor	Spain	Artistic Gymnastics
2	ABAGNALE Giovanni	Italy	Rowing
3	ABALDE Alberto	Spain	Basketball
4	ABALDE Tamara	Spain	Basketball

```
[194] medalhas1 = pd.read_excel("Medals.xlsx")  
medalhas1.head()
```

	Rank	Team/NOC	Gold	Silver	Bronze	Total	Rank by Total
0	1	United States of America	39	41	33	113	1
1	2	People's Republic of China	38	32	18	88	2
2	3	Japan	27	14	17	58	5
3	4	Great Britain	22	21	22	65	4
4	5	ROC	20	28	23	71	3



SIMONE BILES



United States of America



Artistic Gymnastics

Olympic Medals

4

G

1

S

2

B

Games participations

2

First Olympic Games

Rio 2016

Year of Birth

1997

Social Media



Olympic Results



Site "olympics.com"

RESULTS	EVENT	SPORT
Tokyo 2020		
#n/a	Women's All-Around	Artistic Gymnastics
B	Women's Balance Beam	Artistic Gymnastics
#n/a	Women's Floor Exercise	Artistic Gymnastics
S	Women's Team	Artistic Gymnastics
#n/a	Women's Uneven Bars	Artistic Gymnastics
#n/a	Women's Vault	Artistic Gymnastics
Rio 2016		
B	Balance Beam	Artistic Gymnastics
G	Floor Exercise	Artistic Gymnastics
G	Horse Vault	Artistic Gymnastics


```
# Trecho de código que realiza requisições ao site "olympics.com" e obtém  
# ano de nascimento do atleta e modalidades em que participou  
for i in range(len(atletas)):
```

```
    ###  
    # Tratamento do nome do atleta  
    ###
```

```
    |  
    try:  
        r = requests.get("https://olympics.com/en/athletes/"+nomeUrl)  
        tree = html.fromstring(r.content)
```

Requisição ao site para
obter dados de um
atleta específico

```
        details = tree.xpath('//ul[@class="detail_list"]/li/div/text()')  
        b = details.index("Year of Birth")  
        year = details[b+1]  
        atletas.loc[i, "Ano"] = year
```

```
        tabela_part = tree.xpath('//table[@class="sm-mb6 has-header"]/tbody/tr')
```

```
        nomeOlimp = tabela_part[0].xpath('./h2/text()')
```

```
        if len(nomeOlimp)>0:
```

```
            if nomeOlimp[0] == "Tokyo 2020":
```

```
                for tr in tabela_part:
```

```
                    td = tr.xpath('./td')
```

```
                    medalha = td[1].xpath('./div')[0].text_content().replace("'", '').replace("\n", "").replace("\r", "").replace(" ", "")
```

```
                    modalidade = td[2].text_content()
```

```
                    esporte = td[3].text_content()
```

```
                    atletasEsportes.loc[len(atletasEsportes)] = [nomeReal, nomeUrl, esporte, modalidade, medalha]
```

```
except:
```

```
    print(nomeReal)
```

Uso de xpath para
identificação de
elementos no html

Informações obtidas do site

	Nome	Esporte	Modalidade	Medalha
100	Valentina Acosta Giraldo	Archery	Mixed Team	#26
101	Valentina Acosta Giraldo	Archery	Women's Individual	#=33
102	Yenny Acuna	Football	Women	#11
103	Kazuya Adachi	Canoe Slalom	Men's Kayak	#16
104	Seiya Adachi	Water Polo	Men	#10
105	Amal Adam	Archery	Mixed Team	#29
106	Amal Adam	Archery	Women's Individual	#=33
107	Constantin Adam	Rowing	Men's Eight	#7
108	Klaudia Adamek	Athletics	Women's 4 x 100m Relay	#n/a
109	Patrycja Adamkiewicz	Taekwondo	Women -57kg	#=11
110	Liam Adams	Athletics	Men's Marathon	#24
111	Paul Adams	Shooting	Skeet Men	#21
112	Taeyanna Adams	Swimming	Women's 100m Breaststroke	#n/a
113	Yasemin Adar	Wrestling	Women's Freestyle 76kg	B

Id	Nome	Ano	Sexo
0	Katrine Aalerud	1994	F
1	Nestor Abad	1993	M
2	Giovanni Abagnale	1995	M
3	Alberto Abalde	NaN	None
4	Tamara Abalde	1989	F
5	Luc Abalo	1984	M
6	Cesar Abaroa	1996	M
7	Abobakr Abass	1998	M
8	Hamideh Abbasali	1990	F
9	Islam Abbasov	1996	M
10	Lois Abbingh	1992	F
11	Emily Abbot	1997	None
12	Monica Abbott	1985	None
13	Abubaker Haydar Abdalla	1996	M
14	Maryam Abdalla	NaN	None

Construção do modelo hierárquico

Método de criação:

- Utilização dos 8 arquivos .csv obtidos na criação das 8 tabelas do modelo relacional.
- Iterações sobre cada tabela, buscando os dados e incluindo-os em cada camada do modelo hierárquico.

```
atletas = pd.read_csv("atletas.csv")
edicoes = pd.read_csv("edicoes.csv")
paises = pd.read_csv("comites.csv")
esportesModalidades = pd.read_csv("esportes.csv")
participacaoComites = pd.read_csv("participacaoComites.csv")
participacaoAtletas = pd.read_csv("participacaoAtletas.csv")
comiteDosAtletas = pd.read_csv("comiteDosAtletas.csv")
esportesDasEdicoes = pd.read_csv("esportesDasEdicoes.csv")

edicoesInfo=[]
for k in range(len(edicoes)):
    ano = edicoes.loc[k].Ano

    edicao = edicoes.loc[edicoes.Ano==ano]
    participacaoAtletasEdicao = participacaoAtletas.loc[participacaoAtletas.AnoEdicao==ano]
    participacaoComitesEdicao = participacaoComites.loc[participacaoComites.AnoEdicao==ano]
    esportesEdicao = esportesDasEdicoes.loc[esportesDasEdicoes.AnoEdicao==ano]
```

Perguntas do modelo relacional

Pergunta 1: Para um determinado esporte, existe algum país que constantemente é medalha de ouro?

/*Pergunta 1*/

```
SELECT E.Ouro, COUNT(*) QtdOuro
FROM EsportesDasEdicoes E, EsporteModalidade M
WHERE M.Id=E.IdModalidade and M.Nome='Athletics Men's 100 metres'
GROUP BY E.Ouro;
```

index	OURO	QTDOURO
0	RSA	1
1	USA	17
2	GER	1
3	URS	1
4	GBR	3
5	JAM	3
6	TTO	1
7	CAN	2

Athletics Men's 100 metres

index	OURO	QTDOURO
0	BRA	3
1	USA	3
2	URS	3
3	SCG	1
4	JPN	1
5	POL	1
6	NED	1
7	RUS	1

Volleyball Men's Volleyball

index	OURO	QTDOURO
0	CHN	1
1	USA	2
2	URS	1
3	KOR	8

Archery Women's Individual

Pergunta 2: Existe alguma relação entre altura do atleta e esporte praticado por ele? E em relação ao peso?

/*Pergunta 2*/

```
SELECT AnoEdicao, ROUND(AVG(CAST (Altura as float)),1) Media_altura
FROM ParticipacaoAtletas
WHERE AnoEdicao =2016 and Altura<>'-'
GROUP BY AnoEdicao;

SELECT *
FROM (
  SELECT E.EsportePai, E.Nome, ROUND(AVG(CAST (Altura as float)),1) Media_Altura
  FROM ParticipacaoAtletas P, EsporteModalidade E
  WHERE P.IdModalidade=E.Id and AnoEdicao=2016 and Altura<>'-'
  GROUP BY IdModalidade
)
ORDER BY MEDIA_ALTURA DESC
```

index	Key	Value
0	ANOEDICAO	2016
1	MEDIA_ALTURA	176

index	ESPORTEPAI	NOME	MEDIA_ALTURA
0	Basketball	Basketball Men's Basketball	200.5
1	Volleyball	Volleyball Men's Volleyball	196.9
2	Athletics	Athletics Men's Discus Throw	195.8
3	Beach Volleyball	Beach Volleyball Men's Beach Volleyball	194.7
4	Taekwondo	Taekwondo Men's Heavyweight	194.6

301	Gymnastics	Gymnastics Women's Uneven Bars	154.9
302	Weightlifting	Weightlifting Women's Featherweight	154.8
303	Gymnastics	Gymnastics Women's Team All-Around	154.1
304	Gymnastics	Gymnastics Women's Horse Vault	153.3
305	Weightlifting	Weightlifting Women's Flyweight	151.7

Pergunta 2: Existe alguma relação entre altura do atleta e esporte praticado por ele? E em relação ao peso?

/*PESO*/

```
SELECT AnoEdicao, ROUND(AVG(CAST (Peso as float)),1) Media_peso
FROM ParticipacaoAtletas
WHERE AnoEdicao =2016 and Peso<>'-'
GROUP BY AnoEdicao;

SELECT *
FROM (
  SELECT E.EsportePai, E.Nome, ROUND(AVG(CAST (Peso as float)),1) Media_peso
  FROM ParticipacaoAtletas P, EsporteModalidade E
  WHERE P.IdModalidade=E.Id and AnoEdicao=2016 and Peso<>'-'
  GROUP BY IdModalidade
)
ORDER BY MEDIA_peso DESC
```

index	Key	Value
0	ANOEDICAO	2016
1	MEDIA_PESO	71

index	ESPORTEPAI	NOME	MEDIA_PESO
0	Weightlifting	Weightlifting Men's Super-Heavyweight	140.3
1	Athletics	Athletics Men's Shot Put	127.6
2	Judo	Judo Men's Heavyweight	126.9
3	Wrestling	Wrestling Men's Super-Heavyweight, Greco-Roman	125.2
4	Wrestling	Wrestling Men's Super-Heavyweight, Freestyle	122.1

301	Gymnastics	Gymnastics Women's Uneven Bars	48.7
302	Rhythmic Gymnastics	Rhythmic Gymnastics Women's Individual	48.6
303	Weightlifting	Weightlifting Women's Flyweight	47.8
304	Gymnastics	Gymnastics Women's Horse Vault	47.7
305	Gymnastics	Gymnastics Women's Team All-Around	47.4

Pergunta 3: Qual a média de idade dos atletas nas primeiras Olimpíadas? E nas últimas?

```
/*Pergunta 3*/  
DROP TABLE IF EXISTS IDADE;  
CREATE VIEW Idade AS  
SELECT P.IdAtleta, P.AnoEdicao, (P.AnoEdicao-cast (A.AnoDeNascimento as float)) AS idade  
FROM Atleta A, ParticipacaoAtletas P  
WHERE A.Id=P.IdAtleta and A.AnoDeNascimento<>'-' ;  
  
SELECT AnoEdicao, ROUND(AVG(idade),0) Media_Idade  
FROM Idade  
GROUP BY AnoEdicao  
ORDER BY AnoEdicao
```

index	ANOEDICAO	MEDIA_IDADE
0	1896	23
1	1900	29
2	1904	27
3	1906	26
4	1908	27

23	1996	25
24	2000	25
25	2004	26
26	2008	26
27	2012	26
28	2016	26

Pergunta 4: No período da Guerra Fria, é possível ver o predomínio das duas grandes potências nos pódios das Olimpíadas?

*/*Perguntas 4*/*

```
SELECT AnoEdicao, IdComite, Classificacao  
FROM ParticipacaoComites  
WHERE (IdComite='USA' OR IdComite='URS') AND AnoEdicao>1947 AND AnoEdicao< 1989;
```

index	ANOEDICAO	IDCOMITE	CLASSIFICACAO
0	1948	USA	1
1	1952	USA	1
2	1952	URS	2
3	1956	URS	1
4	1956	USA	2
5	1960	URS	1
6	1960	USA	2
7	1964	USA	1
8	1964	URS	2
9	1968	USA	1
10	1968	URS	2
11	1972	URS	1
12	1972	USA	2
13	1976	URS	1
14	1976	USA	3
15	1980	URS	1
16	1984	USA	1
17	1988	URS	1
18	1988	USA	3

Pergunta 5: Qual a proporção de atletas do sexo masculino e do sexo feminino participando nos Jogos Olímpicos?

```
/*Pergunta 5*/  
SELECT P.ANOEDICAO, A.SEXO, COUNT(*) TOTAL  
FROM ATLETA A, PARTICIPACAOATLETAS P  
WHERE A.ID=P.IDATLETA  
GROUP BY P.ANOEDICAO, A.SEXO  
ORDER BY P.ANOEDICAO
```

0	1896	M	380
1	1900	F	33
2	1900	M	1903
3	1904	F	16
4	1904	M	1285
5	1906	M	1722
6	1906	F	11
7	1908	F	47
8	1908	M	3054
9	1912	F	87
10	1912	M	3953
11	1920	F	134
12	1920	M	4158

45	1996	M	8776
46	1996	F	5004
47	2000	F	5436
48	2000	M	8385
49	2004	F	5544
50	2004	M	7899
51	2008	F	5823
52	2008	M	7779
53	2012	F	5812
54	2012	M	7108
55	2016	F	6221
56	2016	M	7467

Potenciais perguntas não implementadas

- Quais os países que mais ganharam medalhas e os países que menos ganharam medalhas em uma determinada Olimpíada?
- Qual o número médio de medalhas de um país nas Olimpíadas que ele participou?
- Em quantas Olimpíadas um determinado atleta participou e quantas medalhas ele ganhou?
- Quais países que mais trazem atletas para os Jogos Olímpicos nas últimas edições?
- Para um determinado país, há uma tendência de piora ou melhora no desempenho, nas Olimpíadas em que participou?
- Na Olimpíada de 1936, como foi o desempenho da Alemanha Nazista?

Perguntas do modelo hierárquico

Pergunta 1: Quais foram os países ganhadores de medalha de ouro no esporte X nas últimas 5 Olimpíadas?

```
#Selecionando o esporte (Através do nome):
esporte="Volleyball"
modalidade="Volleyball Men's Volleyball"
edicoes=["2016", "2012", "2008", "2004", "2000"]

for i in range (5):
    ano=edicoes[i]
    for j in range (len(dados["edicoes"])-1, -1, -1):
        if (dados["edicoes"][j]["ano"]==ano):
            for k in range (len(dados["edicoes"][j]["esportesParticipantes"])):
                if (dados["edicoes"][j]["esportesParticipantes"][k]["nome"]==esporte):
                    for l in range (len(dados["edicoes"][j]["esportesParticipantes"][k]["modalidades"])):
                        if (dados["edicoes"][j]["esportesParticipantes"][k]["modalidades"][l]["nome"]==modalidade):
                            print(dados["edicoes"][j]["ano"]+": "+dados["edicoes"][j]["esportesParticipantes"][k]["modalidades"][l]["ouro"])
                            break
```

2016: BRA
2012: RUS
2008: USA
2004: BRA
2000: SCG

Men's Volleyball

2016: USA
2012: USA
2008: ITA
2004: ROU
2000: AUS

Swimming Women's 200 metres Freestyle

2016: FRA
2012: CUB
2008: CHN
2004: JPN
2000: CHN

Judo Women's Heavyweight

Pergunta 2: Quais as modalidades realizadas pelo atleta X na Olimpíada X?

```
#Pergunta: Quais as modalidades realizadas pelo atleta X na olimpíada X?

#Selecionando a olimpíada (Através do ano):
ano="2012"
#Selecionando o atleta (Através do nome):
atleta="Michael Fred Phelps, II"

for j in range (len(dados["edicoes"])):
    if (dados["edicoes"][j]["ano"]==ano):
        numero=j
        break

for k in range (len(dados["edicoes"][numero]["atletasParticipantes"])):
    if (dados["edicoes"][numero]["atletasParticipantes"][k]["nome"]==atleta):
        for l in range (len(dados["edicoes"][numero]["atletasParticipantes"][k]["esportes"])):
            print(dados["edicoes"][numero]["atletasParticipantes"][k]["esportes"][l]["nomeModalidade"])
```

Swimming Men's 4 x 100 metres Freestyle Relay
Swimming Men's 4 x 200 metres Freestyle Relay
Swimming Men's 100 metres Butterfly
Swimming Men's 200 metres Butterfly
Swimming Men's 200 metres Individual Medley
Swimming Men's 400 metres Individual Medley
Swimming Men's 4 x 100 metres Medley Relay

Michael Fred Phelps, II
Ano: 2012

Gymnastics Women's Individual All-Around
Gymnastics Women's Team All-Around
Gymnastics Women's Floor Exercise
Gymnastics Women's Horse Vault
Gymnastics Women's Uneven Bars

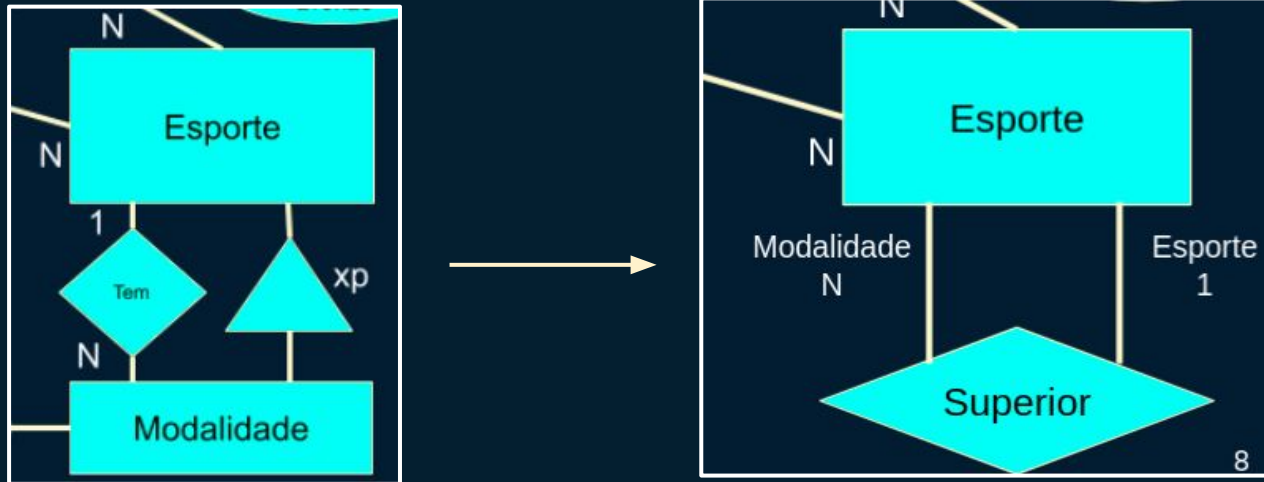
Daiane Garcia dos Santos
Ano: 2004

Outras perguntas implementadas

- Qual o número de atletas por país em uma determinada Olimpíada?
- Quais os atletas participantes de uma determinada Olimpíada e suas respectivas informações?
- Quais os comitês participantes de uma determinada Olimpíada e seus respectivos desempenhos?
- Quais os esportes de uma determinada Olimpíada e os resultados de pódio?

Evolução do Projeto

- Modelo conceitual



Evolução do Projeto

- Abrangência das edições

2000 - 2020



Todas

Evolução do Projeto

- Problemas enfrentados

- No dataset de Tokyo, houve busca de informações em outra fonte (olympics.com) através de web scraping
- Como não há o nome correto de todos os atletas e alguns atletas possuem informações faltantes no site Olympics, a tabela ficou com alguns campos incompletos.
- Para alguns dados, realmente não havia fontes de onde retirar a informação, como altura e peso dos atletas.

```
Id      0
Nome    0
Ano     1009
Sexo    1950
dtype: int64
```

Evolução do Projeto

- Lições aprendidas

- Tratar e integrar dados foi um trabalho um pouco mais complexo do que imaginamos.
- Fazer boas modelagens conceituais e lógicas não é trivial, e influencia diretamente no resultado que o banco de dados vai alcançar.
- Foi necessária bastante análise sobre o tema dos jogos olímpicos para entender como construir modelos organizados e bem estruturados de modo a atingir o nosso objetivo inicial.

The slide features a dark navy blue background. In the top-left and bottom-left corners, there are clusters of overlapping, semi-transparent geometric shapes in shades of green, blue, orange, and pink. Similarly, in the top-right and bottom-right corners, there are clusters of overlapping, semi-transparent geometric shapes in shades of green, blue, purple, and orange. The word "Obrigado!" is centered in the middle of the slide in a white, bold, sans-serif font.

Obrigado!