

✓ The big dataset of ultra-marathon running

+ Código

+ Texto

RE:0. Justificativa do projeto:

Recentemente, desenvolvi um interesse crescente por ultramaratonas, inspirado pelo desafio físico e mental que essas corridas representam. Após concluir minha primeira prova de 10 km, percebi o quanto eventos de longa distância exigem não apenas preparo físico, mas também um entendimento profundo de diversos fatores, como clima, localização e perfil dos participantes. Esse projeto surgiu como uma oportunidade de explorar esse universo, unindo minha curiosidade sobre ultramaratonas com uma análise detalhada de como diferentes variáveis influenciam o desempenho dos atletas em corridas de 50 km.

Além disso, este projeto é uma extensão prática dos meus estudos em análise de dados e ciência de dados, áreas em que estou me especializando. Através dele, busco aplicar conceitos fundamentais como exploração de dados, visualização e identificação de padrões, utilizando ferramentas e técnicas essenciais para a análise de grandes datasets. Ao combinar meu interesse pessoal com objetivos acadêmicos e profissionais, este projeto se torna uma experiência enriquecedora que conecta aprendizado teórico com aplicações reais e alinhadas aos meus objetivos de carreira.

Inspiração: <https://www.youtube.com/watch?v=4sZFkPw87ng&t=4s> Data-set: <https://www.kaggle.com/datasets/aiaiaidavid/the-big-dataset-of-ultra-marathon-running/discussion/420633>

✓ 1. Importando as bibliotecas que preciso:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from google.colab import files
```

✓ 2. Carregando o meu data-set:

```
!pip install kaggle -q
```

```
files.upload()
```

Escolher arquivos kaggle.json

- **kaggle.json**(application/json) - 69 bytes, last modified: 26/11/2024 - 100% done

Saving kaggle.json to kaggle (3).json

```
!mkdir -p ~/.kaggle
```

```
!cp kaggle.json ~/.kaggle
```

```
!chmod 600 ~/.kaggle/kaggle.json
```

```
!kaggle datasets list
```

```
401 - Unauthorized - Unauthenticated
```

```
!kaggle datasets download aiaiaidavid/the-big-dataset-of-ultra-marathon-running
```

Dataset URL: <https://www.kaggle.com/datasets/aiaiaidavid/the-big-dataset-of-ultra-marathon-running>
 License(s): CC0-1.0
 the-big-dataset-of-ultra-marathon-running.zip: Skipping, found more recently modified local copy (use --force to force download)


```
!unzip /content/the-big-dataset-of-ultra-marathon-running.zip
```

Archive: /content/the-big-dataset-of-ultra-marathon-running.zip
 replace TWO_CENTURIES_OF_UM_RACES? [y]es, [n]o, [A]ll, [N]one, [r]ename: N

```
run = pd.read_csv('/content/TWO_CENTURIES_OF_UM_RACES.csv')
```

<ipython-input-58-cec05c552dc1>:1: DtypeWarning: Columns (11) have mixed types. Specify dtype option on import or set low_memory=False
 run = pd.read_csv('/content/TWO_CENTURIES_OF_UM_RACES.csv')

run.head(10)




	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club	Athlete country	Athlete year of birth	Athlete gender	Athlete age category	Athlete average speed	Athlete ID
0	2018	06.01.2018	Selva Costera (CHI)	50km	22	4:51:39 h	Tnfr	CHI	1978.0	M	M35	10.286	
1	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:15:45 h	Roberto Echeverría	CHI	1981.0	M	M35	9.501	
2	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:16:44 h	Puro Trail Osorno	CHI	1987.0	M	M23	9.472	
3	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:34:13 h	Columbia	ARG	1976.0	M	M40	8.976	
4	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:54:14 h	Baguales Trail	CHI	1992.0	M	M23	8.469	
5	2018	06.01.2018	Selva Costera (CHI)	50km	22	6:25:01 h	NaN	ARG	1974.0	M	M40	7.792	
6	2018	06.01.2018	Selva Costera (CHI)	50km	22	6:28:00 h	Los Patagones	ARG	1979.0	F	W35	7.732	
7	2018	06.01.2018	Selva Costera (CHI)	50km	22	6:32:24 h	Reactiva Chile	CHI	1967.0	F	W50	7.645	
8	2018	06.01.2018	Selva Costera (CHI)	50km	22	6:39:08 h	Puro Trail Osorno	CHI	1985.0	M	M23	7.516	
9	2018	06.01.2018	Selva Costera (CHI)	50km	22	6:45:11 h	Marlene Flores Team	CHI	1976.0	M	M40	7.404	

run.shape



(7461195, 13)

run.dtypes



	0
Year of event	int64
Event dates	object
Event name	object
Event distance/length	object
Event number of finishers	int64
Athlete performance	object
Athlete club	object
Athlete country	object
Athlete year of birth	float64
Athlete gender	object
Athlete age category	object
Athlete average speed	object
Athlete ID	int64


3. Limpando e modificando os dados:

run.shape

 (7461195, 13)

Limpando os valores nulos:

```
run.isna().sum()
```




	0
Year of event	0
Event dates	0
Event name	0
Event distance/length	1053
Event number of finishers	0
Athlete performance	2
Athlete club	2826524
Athlete country	3
Athlete year of birth	588161
Athlete gender	7
Athlete age category	584938
Athlete average speed	224
Athlete ID	0



```
run = run.dropna()
```

Resentando o index:

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



	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club	Athlete country	Athlete year of birth	Athlete gender	Athlete age category	Athl average speed
0	2018	06.01.2018	Selva Costera (CHI)	50km	22	4:51:39 h	Tnfr	CHI	1978.0	M	M35	10.
1	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:15:45 h	Roberto Echeverría	CHI	1981.0	M	M35	9.
2	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:16:44 h	Puro Trail Osorno	CHI	1987.0	M	M23	9.
3	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:34:13 h	Columbia	ARG	1976.0	M	M40	8.
4	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:54:14 h	Baguales Trail	CHI	1992.0	M	M23	8.
...
4419678	1995	07.01.1995	Centenary Lakes 50 Km Track Run (AUS)	50km	6	4:47:39 h	*QLD	AUS	1939.0	M	M55	1042
4419679	1995	07.01.1995	Centenary Lakes 50 Km Track Run (AUS)	50km	6	5:58:16 h	*QLD	AUS	1938.0	F	W55	837
4419680	1995	00.00.1995	Szombathely 24 hours running Race (HUN)	24h	3	241.000 km	*Budapest	HUN	1950.0	M	M40	1004



Consertando os tipos dos dados:

```
run.dtypes
```

	0
Year of event	int64
Event dates	object
Event name	object
Event distance/length	object
Event number of finishers	int64
Athlete performance	object
Athlete club	object
Athlete country	object
Athlete year of birth	float64
Athlete gender	object
Athlete age category	object
Athlete average speed	object
Athlete ID	int64

```
# Convertendo a coluna de datas para o formato datetime
run['Event dates'] = pd.to_datetime(run['Event dates'], errors='coerce', dayfirst=True)
```

4. Modificando o dataset:

run

	Year of event	Event dates	Event name	distance/length	Event number of finishers	Athlete performance	Athlete club	Athlete country	Athlete year of birth	Athlete gender	Athlete age category	Athlete average speed
0	2018	2018-01-06	Selva Costera (CHI)	50km	22	4:51:39 h	Tnfrc	CHI	1978.0	M	M35	10.286
1	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:15:45 h	Roberto Echeverría	CHI	1981.0	M	M35	9.501
2	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:16:44 h	Puro Trail Osorno	CHI	1987.0	M	M23	9.472
3	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:34:13 h	Columbia	ARG	1976.0	M	M40	8.976
4	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:54:14 h	Baguales Trail	CHI	1992.0	M	M23	8.469
...
7461092	1995	1995-01-07	Centenary Lakes 50 Km Track Run (AUS)	50km	6	4:47:39 h	*QLD	AUS	1939.0	M	M55	10429.0
7461093	1995	1995-01-07	Centenary Lakes 50 Km Track Run (AUS)	50km	6	5:58:16 h	*QLD	AUS	1938.0	F	W55	8374.0
7461192	1995	NaT	Szombathely 24 hours running Race (HUN)	24h	3	241.000 km	*Budapest	HUN	1950.0	M	M40	10042.0

```
# Criando colunas que podem ser interessantes:

# Idade do atleta:

# ano atual - 'Athlete year of birth'

import datetime
```

```
hoje = datetime.date.today().year
run['Idade do atleta'] = hoje - run['Athlete year of birth']

# Criando uma coluna apenas com os países:
run['pais'] = run['Event name'].str.extract(r'\\((.*?)\\)')

# Criando uma coluna com a estação do ano:

def estacao_do_ano(data):
    dia_mes = (data.month, data.day)
    if (3, 21) <= dia_mes < (6, 21): # Outono
        return 'Outono'
    elif (6, 21) <= dia_mes < (9, 23): # Inverno
        return 'Inverno'
    elif (9, 23) <= dia_mes < (12, 21): # Primavera
        return 'Primavera'
    else: # Verão
        return 'Verão'

# Aplicando a função e criando uma nova coluna
run['Estação'] = run['Event dates'].apply(estacao_do_ano)

# Retirando algumas colunas que não são interessantes:

# Athlete Club, Athlete year of birth, Athlete age category!

run = run.drop(['Athlete club', 'Athlete year of birth', 'Athlete age category'], axis = 1)

# Mudando idade do atleta para inteiro:
run['Idade do atleta'] = run['Idade do atleta'].astype(int)

# Renomeando as colunas que sobramram para pt-br:

run = run.rename(columns = {'Year of event':'Ano da corrida',
                             'Event dates':'Data',
                             'Event name':'Nome do evento',
                             'Event distance/length':'Distância',
                             'Event number of finishers':'Número de finalistas',
                             'Athlete performance':'Tempo do atleta',
                             'Athlete country':'Pais do atleta',
                             'Athlete gender':'Genero do atleta',
                             'Athlete average speed': 'Velocidade média do atleta',
                             'Athlete ID':'ID do atleta',
                             'country':'Pais da corrida'
                            })

run
```

	Ano da corrida	Data	Nome do evento	Distância	Número de finalistas	Tempo do atleta	Pais do atleta	Genero do atleta	Velocidade média do atleta	ID do atleta	Idade do atleta	pais	Estação
0	2018	2018-01-06	Selva Costera (CHI)	50km	22	4:51:39 h	CHI	M	10.286	0	46	CHI	Verão
1	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:15:45 h	CHI	M	9.501	1	43	CHI	Verão
2	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:16:44 h	CHI	M	9.472	2	37	CHI	Verão
3	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:34:13 h	ARG	M	8.976	3	48	CHI	Verão
4	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:54:14 h	CHI	M	8.469	4	32	CHI	Verão
...
7461092	1995	1995-01-07	Centenary Lakes 50 Km Track Run (AUS)	50km	6	4:47:39 h	AUS	M	10429.0	1082443	85	AUS	Verão
7461093	1995	1995-01-07	Centenary Lakes 50 Km Track Run (AUS)	50km	6	5:58:16 h	AUS	F	8374.0	1082581	86	AUS	Verão

5. Definição do problema que iremos analisar:


O objetivo desta análise é explorar diferenças de desempenho em corridas de 50 km no Brasil, considerando variáveis como gênero, faixa etária e sazonalidade (e para tal, vamos fazer esses recortes). Ao investigar esses fatores, buscamos entender padrões relevantes que podem influenciar a performance dos atletas, como diferenças de velocidade média entre homens e mulheres, quais grupos etários têm melhor ou pior desempenho e como as estações do ano podem impactar os resultados.

Essa análise pode ser útil para organizadores de eventos esportivos, treinadores e os próprios atletas, oferecendo insights sobre como características demográficas e condições ambientais podem afetar o desempenho. Também pode ajudar na criação de estratégias personalizadas para treinos ou na organização de eventos em períodos mais favoráveis ao melhor desempenho dos competidores.

Com base nesses objetivos, formulamos questões específicas para responder ao problema e obter conclusões significativas que contribuam para um entendimento mais profundo das dinâmicas em corridas de 50 km no Brasil.

6. Realizando o recorte para os dados que precisamos:

```
# Apenas os 50 km:
run_50 = run[run['Distância'] == '50km']
run_50
```



	Ano da corrida	Data	Nome do evento	Distância	Número de finalistas	Tempo do atleta	Pais do atleta	Gênero do atleta	Velocidade média do atleta	ID do atleta	Idade do atleta	país	Estação
0	2018	2018-01-06	Selva Costera (CHI)	50km	22	4:51:39 h	CHI	M	10.286	0	46	CHI	Verão
1	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:15:45 h	CHI	M	9.501	1	43	CHI	Verão
2	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:16:44 h	CHI	M	9.472	2	37	CHI	Verão
3	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:34:13 h	ARG	M	8.976	3	48	CHI	Verão
4	2018	2018-01-06	Selva Costera (CHI)	50km	22	5:54:14 h	CHI	M	8.469	4	32	CHI	Verão
...
7461089	1995	1995-01-07	Centenary Lakes 50 Km Track Run (AUS)	50km	6	4:19:56 h	AUS	F	11541.0	1046326	68	AUS	Verão
7461090	1995	1995-01-07	Centenary Lakes 50 Km Track Run (AUS)	50km	6	4:28:57 h	AUS	M	11154.0	1070007	70	AUS	Verão

```
# Agora apenas no BR:
run_BRA = run_50[run_50['país'] == 'BRA']
run_BRA
```

	Ano da corrida	Data	Nome do evento	Distância	Número de finalistas	Tempo do atleta	Pais do atleta	Genero do atleta	Velocidade média do atleta	ID do atleta	Idade do atleta	pais	Estação
4677	2018	2018-03-18	Supermaratona Cidade do Rio Grande 50km (BRA)	50km	227	3:11:30 h	BRA	M	15.666	4049	50	BRA	Verão
4678	2018	2018-03-18	Supermaratona Cidade do Rio Grande 50km (BRA)	50km	227	3:21:32 h	BRA	M	14.886	4050	37	BRA	Verão
4679	2018	2018-03-18	Supermaratona Cidade do Rio Grande 50km (BRA)	50km	227	3:25:50 h	BRA	M	14.575	4051	49	BRA	Verão
4680	2018	2018-03-18	Supermaratona Cidade do Rio Grande 50km (BRA)	50km	227	3:27:03 h	BRA	M	14.489	4052	45	BRA	Verão
4682	2018	2018-03-18	Supermaratona Cidade do Rio Grande 50km (BRA)	50km	227	3:31:48 h	BRA	M	14.164	4054	66	BRA	Verão
...
6773893	2015	2015-11-07	Costa Esmeralda Ultra Trail 50km (BRA)	50km	207	10:39:34 h	BRA	F	4.691	1321570	69	BRA	Primavera

Próximas etapas:

Gerar código com run_BRA

☒ Ver gráficos recomendados

New interactive sheet

run_BRA.size

35581

run_BRA é portanto, o conjunto que iremos estudar (apenas 50 km realizadas no BR)!

7. EDA - Analise expoloratória dos dados

run_BRA.dtypes

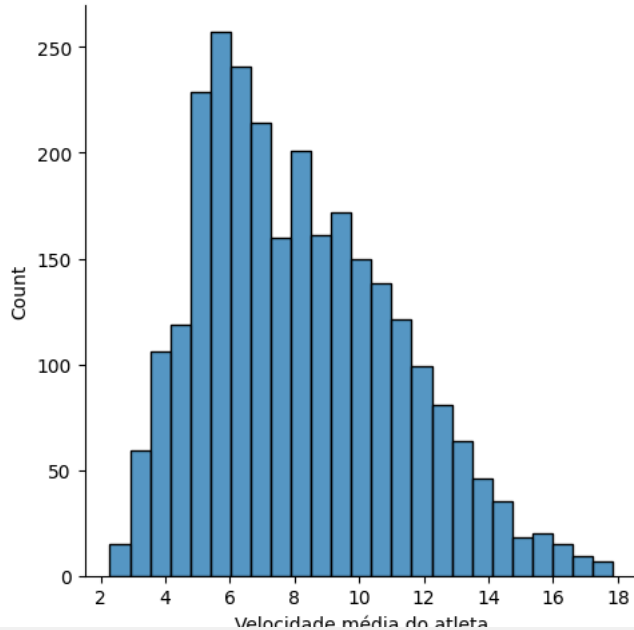
	0
Ano da corrida	int64
Data	datetime64[ns]
Nome do evento	object
Distância	object
Número de finalistas	int64
Tempo do atleta	object
Pais do atleta	object
Genero do atleta	object
Velocidade média do atleta	object
ID do atleta	int64
Idade do atleta	int64
pais	object
Estação	object

run_BRA['Velocidade média do atleta'] = run_BRA['Velocidade média do atleta'].astype(float)

<ipython-input-103-9c0ffde64952>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: [```
sns.displot\(run_BRA\[run_BRA\['Distância'\] == '50km'\]\['Velocidade média do atleta'\]\)
```](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-run_BRA['Velocidade média do atleta'] = run_BRA['Velocidade média do atleta'].astype(float)</a></p>
</div>
<div data-bbox=)

```
<seaborn.axisgrid.FacetGrid at 0x7e117b620e50>
```



```
run_BRA['Velocidade média do atleta'].mean()
```

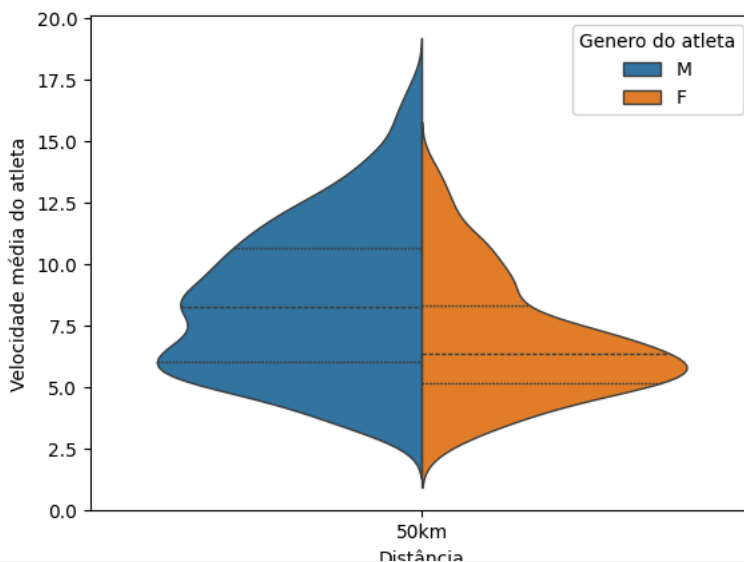
```
8.169839970770918
```

O gráfico mostra a distribuição da velocidade média dos atletas que participaram de corridas de 50 km no Brasil. A média calculada para essas velocidades foi de aproximadamente 8,17 km/h, indicando que a maior parte dos atletas se concentra em um desempenho relativamente moderado para longas distâncias, o que é esperado, dado o caráter extenuante desse tipo de prova. O pico no gráfico ocorre na faixa de 6 a 8 km/h, representando a maior frequência de velocidades médias registradas. Essa concentração sugere que essa faixa é o "ritmo padrão" para a maioria dos competidores.

A cauda do gráfico, que se estende até velocidades mais altas (12 a 18 km/h), indica que alguns atletas apresentam um desempenho excepcional, conseguindo manter velocidades médias mais altas ao longo dos 50 km. Por outro lado, o lado esquerdo do gráfico, com velocidades menores que 6 km/h, mostra uma menor frequência de atletas, que podem estar associados a participantes menos preparados ou que enfrentaram dificuldades durante a prova. Esse padrão reforça que a maior parte dos competidores apresenta velocidades consistentes em torno da média, com alguns destaques tanto acima quanto abaixo desse intervalo.

```
sns.violinplot(data = run_BRA, x = 'Distância', y = 'Velocidade média do atleta', hue = 'Genero do atleta', split = True, inner = 'quartiles')
```

```
<Axes: xlabel='Distância', ylabel='Velocidade média do atleta'>
```

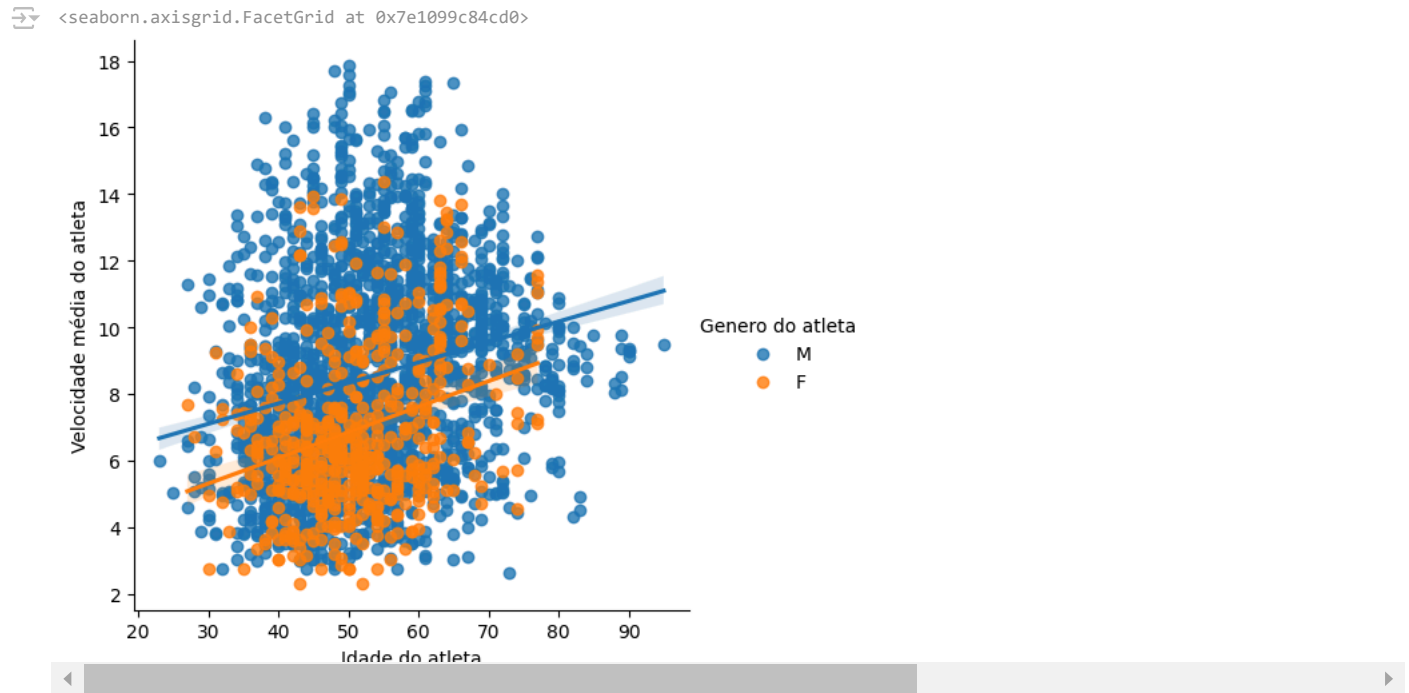




O gráfico violino apresenta a distribuição das velocidades médias dos atletas que participaram de corridas de 50 km no Brasil, segmentadas por gênero. A distribuição masculina (azul) e feminina (laranja) mostram comportamentos similares, mas com algumas diferenças importantes. A largura do violino indica a densidade dos atletas em cada faixa de velocidade média, enquanto as linhas internas representam os quartis, incluindo a mediana.

Os homens apresentam uma leve tendência a atingir velocidades médias mais altas, conforme evidenciado pela extensão superior do violino azul em comparação ao laranja, que atinge velocidades próximas a 17,5 km/h. No entanto, a maior densidade de ambos os gêneros está concentrada na faixa entre 6 e 10 km/h, sugerindo que esta é a faixa de ritmo mais comum para a maioria dos participantes, independentemente do gênero. A diferença de performance entre os gêneros é visível nas velocidades extremas, onde há uma densidade maior de homens nas velocidades mais altas. Isso pode estar relacionado a diferenças fisiológicas ou à representatividade proporcional dos gêneros na corrida.

```
sns.lmplot(data = run_BRA, x = 'Idade do atleta', y = 'Velocidade média do atleta', hue = 'Genero do atleta')
```



O gráfico de dispersão com linhas de regressão mostra a relação entre a idade e a velocidade média dos atletas em corridas de 50 km no Brasil, segmentado por gênero. A maior concentração de atletas está entre os 30 e 60 anos, com ambos os gêneros apresentando uma variação significativa na velocidade dentro dessa faixa etária. As linhas de regressão sugerem uma leve tendência de aumento na velocidade média com a idade até certo ponto, estabilizando ou diminuindo levemente em idades mais avançadas. Homens (linha azul) tendem a manter velocidades médias ligeiramente superiores às mulheres (linha laranja) em quase todas as idades, especialmente nas faixas etárias mais altas.

Para atletas acima de 70 anos, observa-se uma diminuição na densidade de participantes, mas com velocidades ainda competitivas para essa faixa etária. Esse comportamento indica que o treinamento e a experiência desempenham um papel importante no desempenho, permitindo que muitos atletas mantenham boas velocidades em idades avançadas. O gráfico reforça que a idade influencia o desempenho em corridas longas, mas de forma não linear, com diferenças consistentes entre os gêneros.

### Qual é a diferença de velocidade média entre homens e mulheres em corridas de 50 km no Brasil?

Avaliaremos a disparidade de desempenho por gênero, utilizando a velocidade média como métrica principal.

```
run_BRA.dtypes
```

|                            |                |
|----------------------------|----------------|
|                            | 0              |
| Ano da corrida             | int64          |
| Data                       | datetime64[ns] |
| Nome do evento             | object         |
| Distância                  | object         |
| Número de finalistas       | int64          |
| Tempo do atleta            | object         |
| Pais do atleta             | object         |
| Genero do atleta           | object         |
| Velocidade média do atleta | float64        |
| ID do atleta               | int64          |
| Idade do atleta            | int64          |
| pais                       | object         |
| Estação                    | object         |

```
Separando os grupos e calculando as médias:

run_BRA.groupby(['Distância', 'Genero do atleta'])['Velocidade média do atleta'].mean()
```

|           |                  |                            |
|-----------|------------------|----------------------------|
|           |                  | Velocidade média do atleta |
| Distância | Genero do atleta |                            |
| 50km      | F                | 6.899980                   |
|           | M                | 8.496492                   |

Os dados revelam que, em corridas de 50 km no Brasil, os homens têm uma velocidade média de 8,50 km/h, enquanto as mulheres apresentam uma média de 6,90 km/h, indicando uma diferença de aproximadamente 1,60 km/h. Essa variação pode estar relacionada a fatores fisiológicos, como a capacidade aeróbica e muscular, que geralmente favorecem os homens em esportes de longa distância. Além disso, aspectos como preparação, experiência e proporção de participantes por gênero podem também influenciar esse resultado.

Quais faixas etárias apresentam o melhor desempenho em corridas de 50 km no Brasil, considerando apenas grupos com pelo menos 20 participações?

Identificaremos as idades com maior velocidade média entre os atletas.

```
run_BRA.query('Distância == "50km"]').groupby('Idade do atleta')['Velocidade média do atleta'].agg(['mean', 'count']).sort_values('mean',
```

1 to 39 of 39 entries

Filter

?

| Idade do atleta | mean               | count |
|-----------------|--------------------|-------|
| 63              | 9.967259259259258  | 54    |
| 70              | 9.937904761904761  | 21    |
| 66              | 9.930046511627907  | 43    |
| 72              | 9.769961538461539  | 26    |
| 60              | 9.720870129870129  | 77    |
| 59              | 9.507505263157894  | 95    |
| 64              | 9.44582142857143   | 56    |
| 69              | 9.406829268292682  | 41    |
| 55              | 9.305572916666668  | 96    |
| 71              | 9.024583333333334  | 24    |
| 58              | 8.907677966101694  | 59    |
| 50              | 8.861642857142858  | 84    |
| 61              | 8.848117647058825  | 51    |
| 68              | 8.790521739130435  | 23    |
| 62              | 8.748682539682541  | 63    |
| 49              | 8.725704347826087  | 115   |
| 56              | 8.712922222222222  | 90    |
| 65              | 8.647121951219512  | 41    |
| 53              | 8.514888888888889  | 81    |
| 51              | 8.153816513761468  | 109   |
| 67              | 8.134961538461539  | 26    |
| 54              | 8.077119047619046  | 84    |
| 57              | 8.065828125        | 64    |
| 34              | 7.95184            | 25    |
| 41              | 7.727780487804878  | 82    |
| 42              | 7.50715            | 80    |
| 45              | 7.471925925925926  | 108   |
| 39              | 7.46073076923077   | 52    |
| 48              | 7.357990740740741  | 108   |
| 52              | 7.249402173913044  | 92    |
| 46              | 7.236575221238938  | 113   |
| 43              | 7.199231884057971  | 69    |
| 37              | 7.0338918918918925 | 37    |
| 38              | 6.972814814814814  | 54    |
| 36              | 6.943489361702127  | 47    |
| 44              | 6.912425925925926  | 108   |
| 47              | 6.798075949367089  | 79    |
| 35              | 6.722809523809524  | 21    |
| 40              | 6.486586206896551  | 58    |

Show 100 per page

Os dados mostram que atletas na faixa dos 60 a 70 anos têm as maiores velocidades médias em corridas de 50 km no Brasil, com destaque para os 63 anos (9,97 km/h, 54 participações). Isso indica que a experiência desempenha um papel importante no desempenho. Em contraste, atletas mais jovens, como os de 35 anos (6,72 km/h, 21 participações), apresentam velocidades médias mais baixas, sugerindo que a preparação e a estratégia superam a vantagem física em provas de longa distância.

Quais faixas etárias apresentam o pior desempenho em corridas de 50 km no Brasil, considerando apenas grupos com pelo menos 10 participações?

Analisaremos as idades com menor velocidade média, para entender padrões de baixo desempenho.

```
run_BRA.query('Distância == "50km"]').groupby('Idade do atleta')['Velocidade média do atleta'].agg(['mean', 'count']).sort_values('mean',
```

↺

1 to 25 of 45 entries Filter  ?

| Idade do atleta | mean               | count |
|-----------------|--------------------|-------|
| 30              | 6.361153846153846  | 13    |
| 40              | 6.486586206896551  | 58    |
| 35              | 6.722809523809524  | 21    |
| 47              | 6.798075949367089  | 79    |
| 44              | 6.912425925925926  | 108   |
| 36              | 6.943489361702127  | 47    |
| 38              | 6.972814814814814  | 54    |
| 37              | 7.0338918918918925 | 37    |
| 43              | 7.199231884057971  | 69    |
| 46              | 7.236575221238938  | 113   |
| 52              | 7.249402173913044  | 92    |
| 48              | 7.357990740740741  | 108   |
| 39              | 7.46073076923077   | 52    |
| 45              | 7.471925925925926  | 108   |
| 42              | 7.50715            | 80    |
| 41              | 7.727780487804878  | 82    |
| 33              | 7.882727272727272  | 11    |
| 34              | 7.95184            | 25    |
| 57              | 8.065828125        | 64    |
| 54              | 8.077119047619046  | 84    |
| 67              | 8.134961538461539  | 26    |
| 51              | 8.153816513761468  | 109   |
| 74              | 8.392882352941177  | 17    |
| 53              | 8.514888888888889  | 81    |
| 65              | 8.647121951219512  | 41    |

Show 

25

 per page

1

2

Os dados indicam que as maiores velocidades médias em corridas de 50 km no Brasil são alcançadas por atletas mais velhos, com destaque para a idade 77 anos, que apresenta a maior média de 10,15 km/h (18 participações), seguida por 63 anos (9,97 km/h, 54 participações) e 70 anos (9,94 km/h, 21 participações). Isso reforça o papel da experiência no desempenho em corridas longas. Por outro lado, atletas mais jovens, como os de 30 anos (6,36 km/h, 13 participações), têm velocidades médias significativamente mais baixas, indicando que, em provas de resistência, treinamento e estratégia parecem ser mais determinantes do que a juventude.

↵ Analisar as idades com menor velocidade média, para entender padrões de baixo desempenho.

O desempenho em corridas de 50 km no Brasil é mais lento no verão do que no inverno?

```
run_BRA.groupby('Estação')['Velocidade média do atleta'].agg(['mean', 'count']).sort_values('mean', ascending = False)
```

↺

|           | mean     | count |
|-----------|----------|-------|
| Estação   |          |       |
| Verão     | 9.183759 | 1222  |
| Outono    | 8.111480 | 498   |
| Inverno   | 7.830668 | 382   |
| Primavera | 6.469452 | 635   |

```
Agrupando os dados por estação e calculando a média e a contagem
estacao_agg = run_BRA.groupby('Estação')['Velocidade média do atleta'].agg(['mean', 'count']).sort_values('mean', ascending=False)

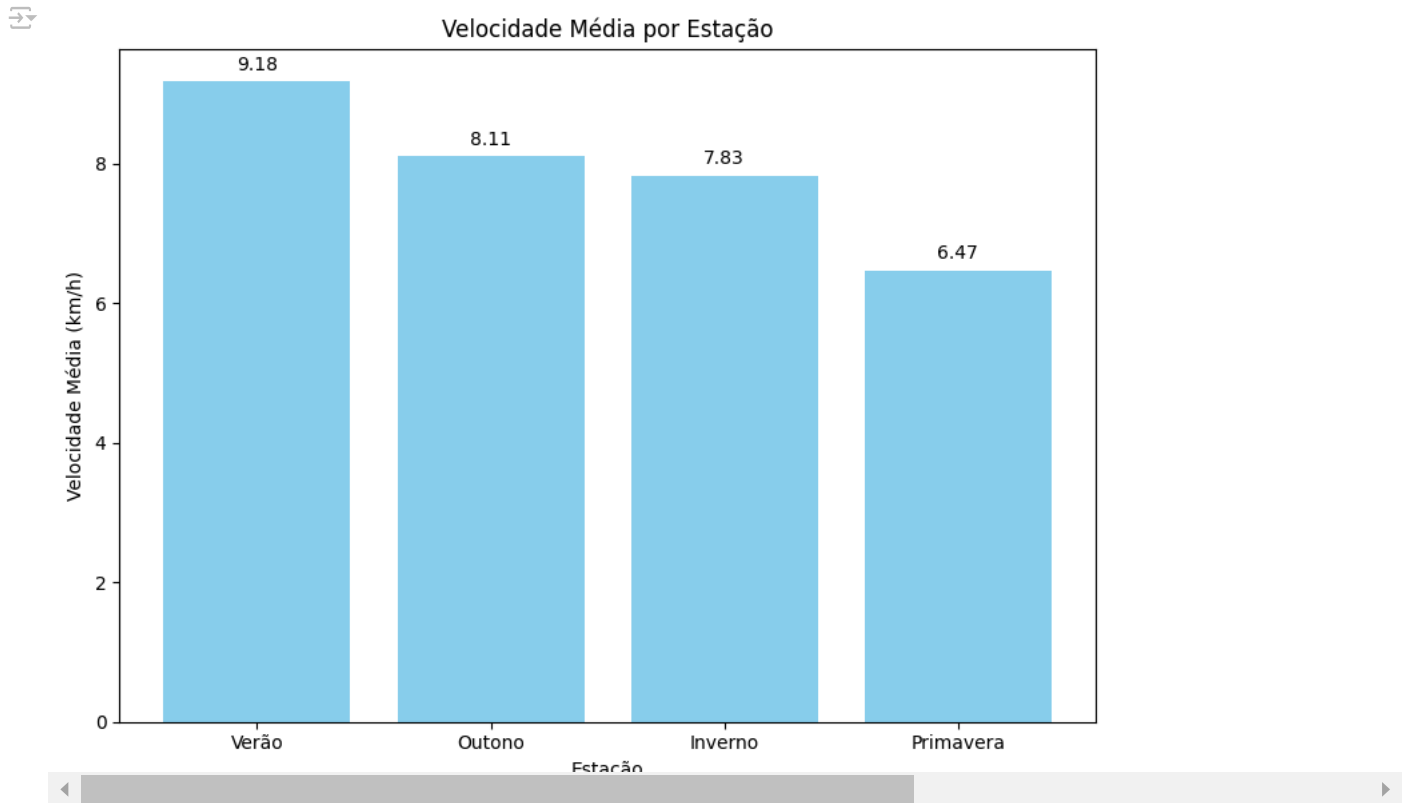
Configurando o gráfico
fig, ax1 = plt.subplots(figsize=(8, 6))

Gráfico de barras para a média de velocidade por estação
ax1.bar(estacao_agg.index, estacao_agg['mean'], color='skyblue', label='Média de Velocidade')
ax1.set_ylabel('Velocidade Média (km/h)')
ax1.set_xlabel('Estação')
ax1.set_title('Velocidade Média por Estação')

Adicionando os valores acima das barras
for i, v in enumerate(estacao_agg['mean']):
 ax1.text(i, v + 0.1, f'{v:.2f}', ha='center', va='bottom')

Exibindo o gráfico
```

```
plt.tight_layout()
plt.show()
```



O gráfico e a tabela indicam as velocidades médias dos atletas em corridas de 50 km no Brasil, distribuídas por estação do ano. A maior velocidade média ocorre no verão (9,18 km/h, com 1222 participações), seguida pelo outono (8,11 km/h, 498 participações), inverno (7,83 km/h, 382 participações) e, finalmente, pela primavera, que apresenta a menor média (6,47 km/h, 635 participações). A maior quantidade de participações também ocorre no verão, o que pode sugerir que a estação tem características favoráveis para essas provas.

O aumento da velocidade média no verão pode estar relacionado ao clima mais quente, que pode estimular maior esforço físico por parte dos atletas, especialmente em corridas de longa distância, para concluir a prova mais rapidamente e evitar exposição prolongada ao calor. Por outro lado, a primavera, com temperaturas amenas e potencialmente condições mais úmidas, pode favorecer desempenhos mais moderados, justificando a menor velocidade média observada. Esses padrões indicam que a sazonalidade exerce influência no desempenho dos atletas.

### ✓ Existe alguma diferença no desempenho entre homens e mulheres em diferentes estações do ano?

Analisaremos se as estações afetam igualmente ambos os gêneros ou se há diferenças.

```
desempenho_estacao_genero = (
 run_BRA.groupby(['Estação', 'Genero do atleta'])['Velocidade média do atleta']
 .agg(['mean', 'count'])
 .reset_index()
 .sort_values(['Estação', 'Genero do atleta'])
)

Exibindo os dados agrupados
print(desempenho_estacao_genero)

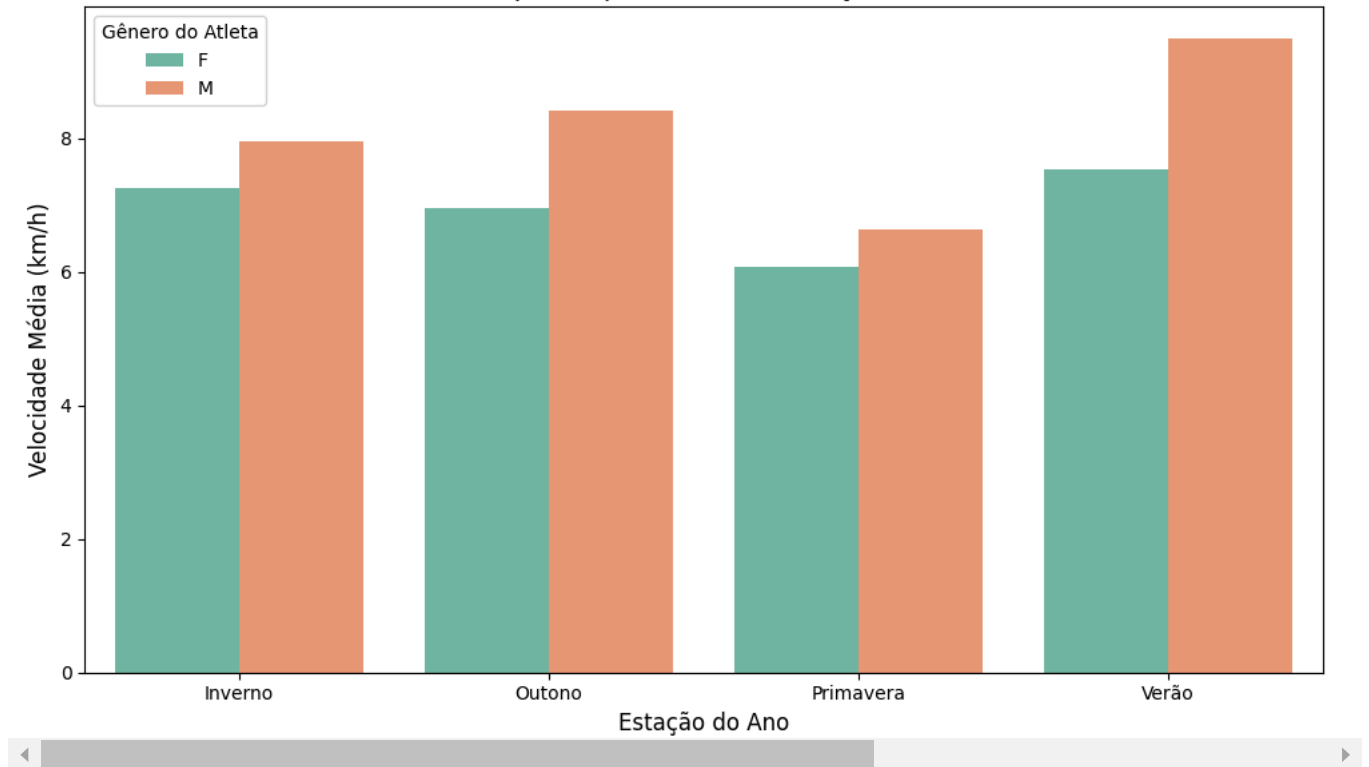
Configurando o gráfico
plt.figure(figsize=(10, 6))
sns.barplot(
 data=desempenho_estacao_genero,
 x='Estação',
 y='mean',
 hue='Genero do atleta',
 palette='Set2'
)

Adicionando título e rótulos
plt.title('Desempenho por Gênero e Estação do Ano', fontsize=14)
plt.ylabel('Velocidade Média (km/h)', fontsize=12)
plt.xlabel('Estação do Ano', fontsize=12)
plt.legend(title='Gênero do Atleta')
plt.tight_layout()
```

```
Exibindo o gráfico
plt.show()
```

|   | Estação   | Genero do atleta | mean     | count |
|---|-----------|------------------|----------|-------|
| 0 | Inverno   | F                | 7.262696 | 69    |
| 1 | Inverno   | M                | 7.955875 | 313   |
| 2 | Outono    | F                | 6.955485 | 103   |
| 3 | Outono    | M                | 8.412916 | 395   |
| 4 | Primavera | F                | 6.080286 | 192   |
| 5 | Primavera | M                | 6.636686 | 443   |
| 6 | Verão     | F                | 7.546087 | 196   |
| 7 | Verão     | M                | 9.496608 | 1026  |

Desempenho por Gênero e Estação do Ano



O gráfico apresenta o desempenho médio dos atletas em corridas de 50 km no Brasil, analisado por gênero (feminino e masculino) e distribuído por estação do ano. Observa-se que, em todas as estações, os homens (barras laranja) apresentam velocidades médias consistentemente superiores às mulheres (barras verde). Essa diferença pode estar relacionada a fatores fisiológicos, como maior capacidade aeróbica e resistência muscular masculina.

Além disso, a maior velocidade média para ambos os gêneros é registrada no verão, indicando que o calor pode estimular os atletas a manterem ritmos mais rápidos para reduzir o tempo de exposição às condições climáticas adversas. Em contrapartida, a primavera apresenta as menores velocidades médias, o que pode estar associado a condições climáticas menos favoráveis ou outros fatores que afetam o desempenho. Esses padrões mostram que tanto o gênero quanto a sazonalidade têm impacto significativo no desempenho dos atletas.

## ✓ Como o número de participantes influencia a velocidade média?

Avaliaremos se eventos com mais participantes têm uma velocidade média maior ou menor, indicando possíveis impactos da competitividade ou condições do evento.

```
evento_agg = (
 run_BRA.groupby('Nome do evento')['Velocidade média do atleta']
 .agg(['mean', 'count'])
 .reset_index()
 .rename(columns={'mean': 'Velocidade média', 'count': 'Número de participantes'})
)
```

```
Exibindo os dados agregados
print(evento_agg.head())
```

```
Criando um gráfico de dispersão para visualizar a relação
plt.figure(figsize=(10, 6))
sns.scatterplot(
 data=evento_agg,
 x='Número de participantes',
 y='Velocidade média',
 color='blue',
 alpha=0.7
```

)

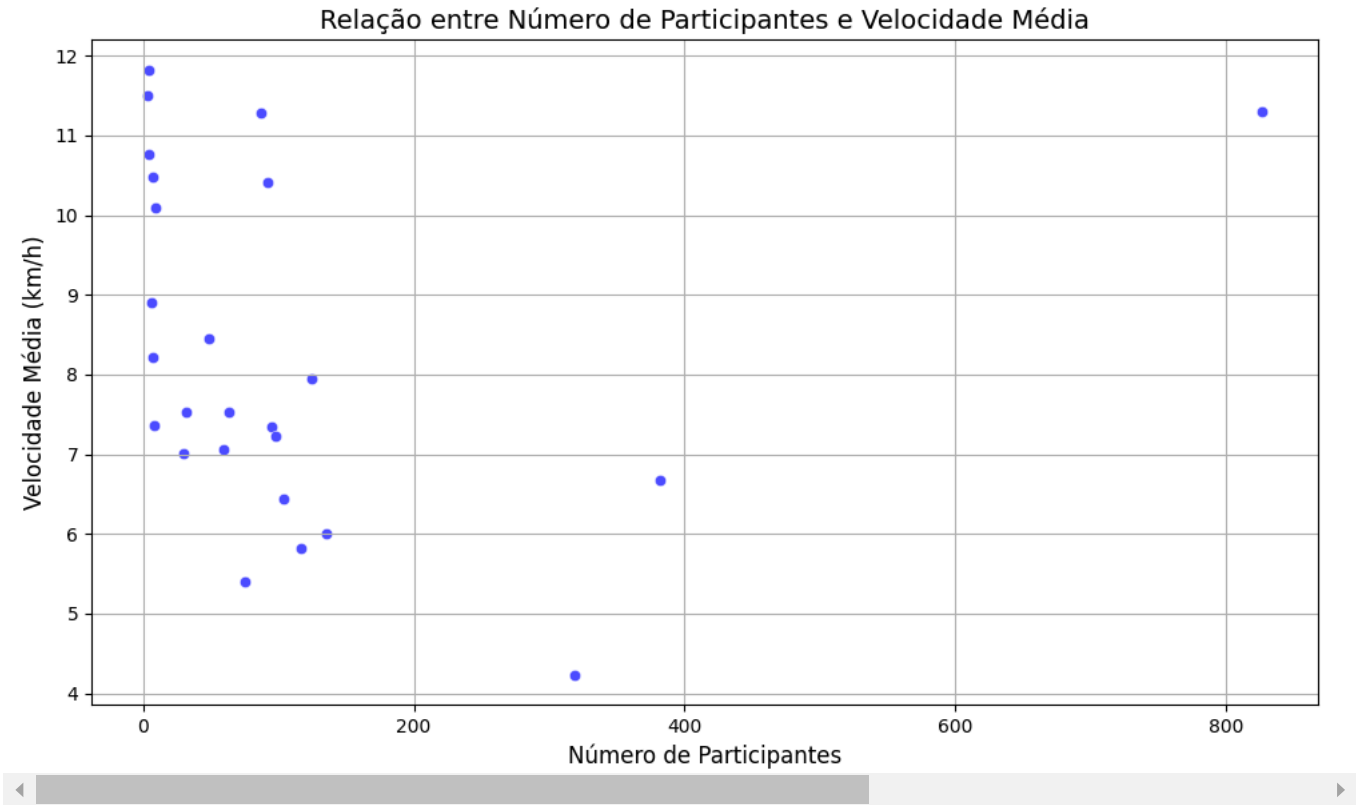
```
Ajustando o gráfico
plt.title('Relação entre Número de Participantes e Velocidade Média', fontsize=14)
plt.xlabel('Número de Participantes', fontsize=12)
plt.ylabel('Velocidade Média (km/h)', fontsize=12)
plt.grid(True)
plt.tight_layout()

Exibindo o gráfico
plt.show()
```

|   | Nome do evento                             | Velocidade média \ |
|---|--------------------------------------------|--------------------|
| 0 | 50km da Praia Grande (BRA)                 | 7.350189           |
| 1 | Campos do Jordão Ultra Trail (BRA)         | 7.055271           |
| 2 | Copa Brasil Caixa de Marcha Atletica (BRA) | 10.471143          |
| 3 | Copa Brasil de Marcha Atletica (BRA)       | 10.759000          |
| 4 | Costa Esmeralda Ultra Trail 50km (BRA)     | 6.675809           |

| Número de participantes |     |
|-------------------------|-----|
| 0                       | 95  |
| 1                       | 59  |
| 2                       | 7   |
| 3                       | 4   |
| 4                       | 382 |



O gráfico de dispersão mostra a relação entre o número de participantes por evento e a velocidade média dos atletas em corridas de 50 km. Observa-se que eventos com poucos participantes (< 100) apresentam uma maior variação de velocidades médias, incluindo as mais altas (acima de 11 km/h). Isso pode indicar que eventos menores, com menos atletas, tendem a ter maior competitividade ou condições mais favoráveis que permitem aos participantes manter um ritmo mais elevado.

Por outro lado, eventos com maior número de participantes (> 400) mostram velocidades médias mais concentradas entre 5 e 7 km/h, sugerindo que, à medida que o número de participantes aumenta, a média de velocidade tende a diminuir. Isso pode ser explicado por fatores como aumento da diversidade de habilidades dos atletas ou maior dificuldade de logística e organização em eventos maiores. Essa relação sugere que a competitividade e o perfil dos participantes variam significativamente com o tamanho dos eventos.

✓ Há diferença no desempenho de atletas de diferentes países em corridas realizadas no Brasil?

Compararemos a velocidade média dos atletas estrangeiros com a dos brasileiros em eventos realizados no Brasil.

```
Agrupando os dados por nacionalidade, calculando a média e o número de participantes

Garantir que a coluna 'Velocidade média do atleta' está no formato numérico
run_50['Velocidade média do atleta'] = pd.to_numeric(run_50['Velocidade média do atleta'], errors='coerce')

Remover valores nulos após a conversão
```

```
run_50 = run_50.dropna(subset=['Velocidade média do atleta'])

Agrupando os dados por nacionalidade, calculando a média e o número de participantes
desempenho_pais = (
 run_50.groupby('Pais do atleta')['Velocidade média do atleta']
 .agg(['mean', 'count'])
 .reset_index()
 .rename(columns={'mean': 'Velocidade média', 'count': 'Número de participantes'})
)

Filtrando países com pelo menos 20 participantes para evitar distorções
desempenho_pais_filtrado = desempenho_pais.query('`Número de participantes` >= 20')

Exibindo os dados agregados
print(desempenho_pais_filtrado)

Criando o gráfico
plt.figure(figsize=(12, 6))
sns.barplot(
 data=desempenho_pais_filtrado.sort_values('Velocidade média', ascending=False),
 x='Pais do atleta',
 y='Velocidade média',
 palette='viridis'
)

Ajustando o gráfico
plt.title('Comparação de Velocidade Média por Nacionalidade', fontsize=14)
plt.ylabel('Velocidade Média (km/h)', fontsize=12)
plt.xlabel('Nacionalidade', fontsize=12)
plt.xticks(rotation=45, fontsize = 5)
plt.tight_layout()

Exibindo o gráfico
plt.show()
```



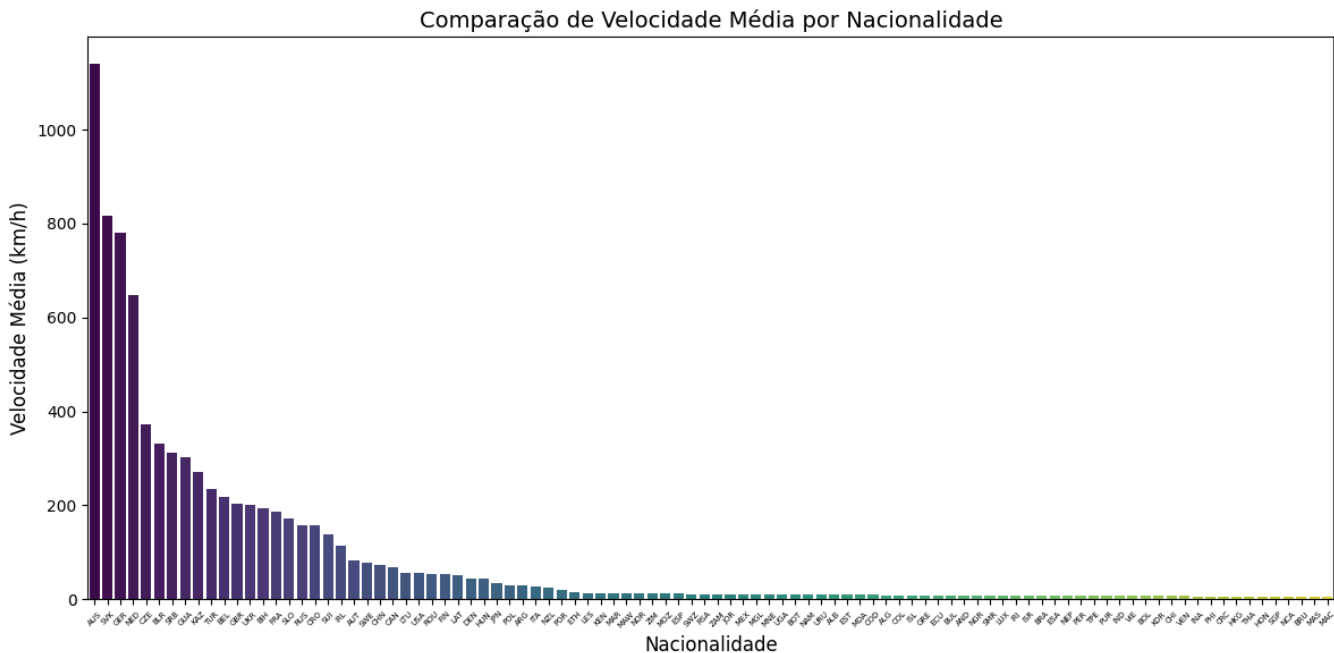
|     | Pais do atleta | Velocidade média | Número de participantes |
|-----|----------------|------------------|-------------------------|
| 1   | ALB            | 9.420661         | 56                      |
| 2   | ALG            | 9.119571         | 42                      |
| 3   | AND            | 8.478615         | 26                      |
| 5   | ARG            | 29.773004        | 2531                    |
| 7   | AUS            | 1140.154384      | 6505                    |
| ..  | ...            | ...              | ...                     |
| 151 | USA            | 55.553256        | 476436                  |
| 153 | VEN            | 7.077600         | 90                      |
| 154 | VIE            | 7.461611         | 36                      |
| 156 | ZAM            | 10.953946        | 37                      |
| 157 | ZIM            | 12.241095        | 599                     |

[96 rows x 3 columns]

<ipython-input-137-6c153d919caa>:25: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `le

sns.barplot(



# Adicionando uma cor de destaque para o Brasil

highlight\_color = 'red'

default\_color = 'gray'

# Criando uma lista de cores: 'red' para Brasil e 'gray' para os demais

colors = [

highlight\_color if pais == 'BRA' else default\_color

for pais in desempenho\_pais\_filtrado['Pais do atleta']

]

# Criando o gráfico com destaque para o Brasil

plt.figure(figsize=(12, 6))

sns.barplot(

data=desempenho\_pais\_filtrado.sort\_values('Velocidade média', ascending=False),

x='Pais do atleta',

y='Velocidade média',

palette=colors

)

# Ajustando o gráfico

plt.title('Comparação de Velocidade Média por Nacionalidade', fontsize=14)

plt.ylabel('Velocidade Média (km/h)', fontsize=12)

plt.xlabel('Nacionalidade', fontsize=12)

plt.xticks(rotation=45, fontsize=5) # Reduzindo o tamanho da fonte

plt.tight\_layout()

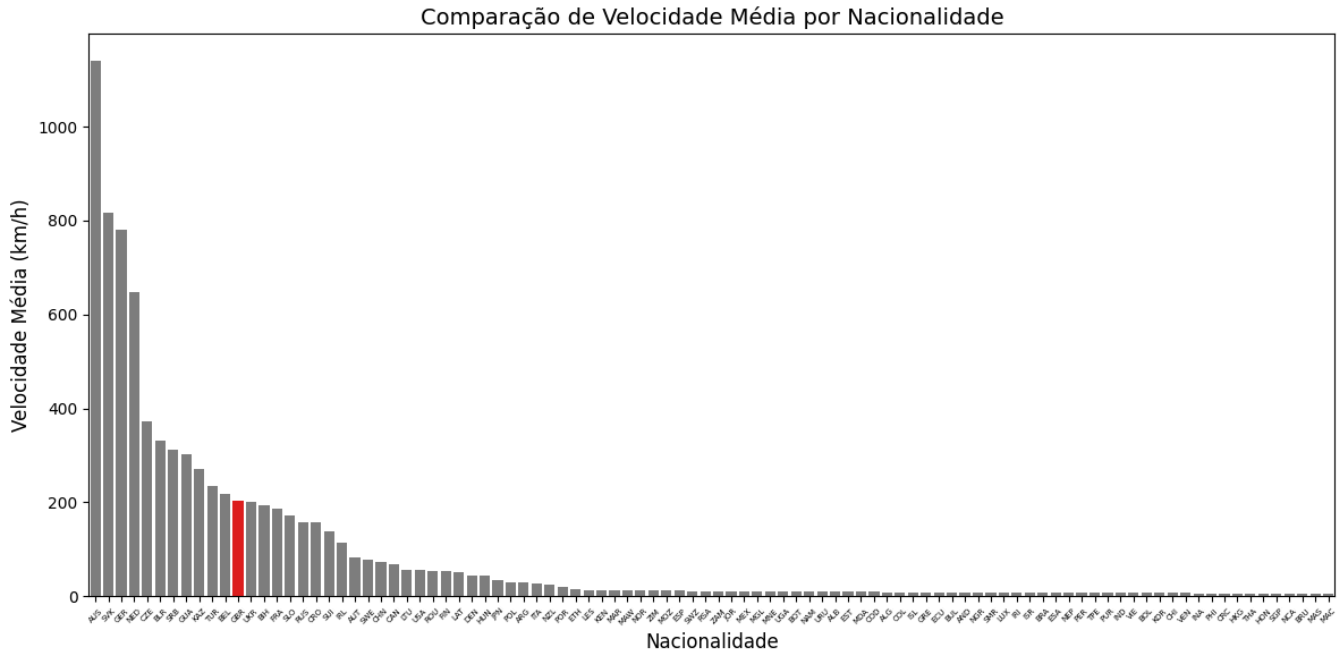
# Exibindo o gráfico

plt.show()

```

<ipython-input-140-0f384b77aacf>:13: FutureWarning:
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `le
sns.barplot(

```



O primeiro gráfico apresenta a comparação da velocidade média dos atletas em corridas de 50 km, por nacionalidade, sem destaque visual. Observa-se que há uma grande variabilidade na velocidade média entre os países, com algumas nacionalidades apresentando um desempenho médio mais elevado. O gráfico também reflete a diversidade de participantes, evidenciada pela presença de muitas nacionalidades, mas dificulta a identificação de países específicos devido à falta de destaque.

O segundo gráfico adiciona uma cor de destaque para o Brasil (em vermelho), facilitando a identificação do desempenho dos atletas brasileiros em relação a outras nacionalidades. Isso permite perceber que, embora o Brasil não esteja entre as nacionalidades com as maiores velocidades médias, ele se encontra em uma posição competitiva intermediária, com uma quantidade significativa de participantes. O destaque ajuda a focar a análise no desempenho nacional, enquanto mantém a perspectiva global de comparação com outros países.

## ✓ Como o desempenho dos atletas muda ao longo dos anos?

Verificaremos se há uma tendência de aumento ou diminuição na velocidade média dos atletas ao longo das décadas.

```

Agrupando os dados por ano e calculando a velocidade média
desempenho_ano = (
 run_BRA.groupby('Ano da corrida')['Velocidade média do atleta']
 .mean()
 .reset_index()
 .rename(columns={'Velocidade média do atleta': 'Velocidade média'})
)

Exibindo os dados agregados
print(desempenho_ano)

Criando o gráfico de linha para observar a tendência ao longo dos anos
plt.figure(figsize=(10, 6))
sns.lineplot(
 data=desempenho_ano,
 x='Ano da corrida',
 y='Velocidade média',
 marker='o',
 color='blue'
)

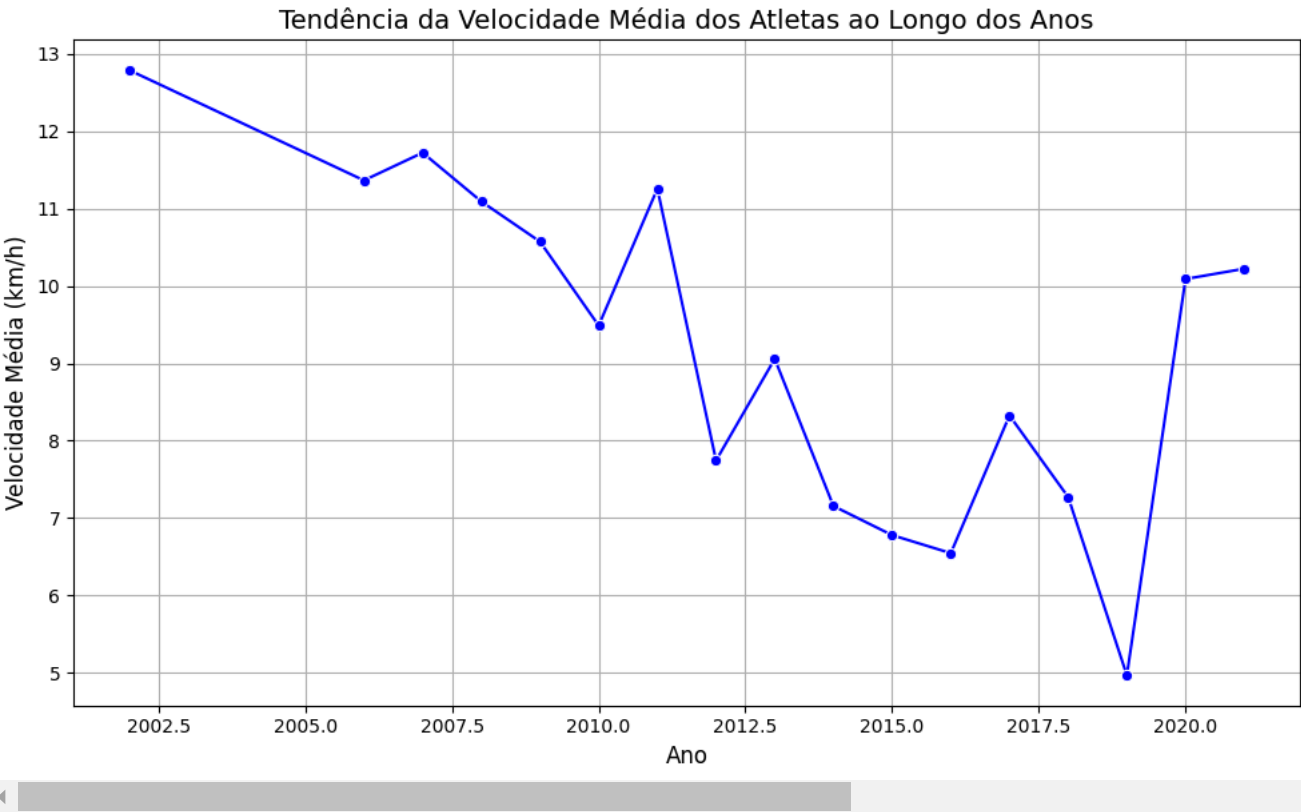
Ajustando o gráfico
plt.title('Tendência da Velocidade Média dos Atletas ao Longo dos Anos', fontsize=14)
plt.xlabel('Ano', fontsize=12)
plt.ylabel('Velocidade Média (km/h)', fontsize=12)
plt.grid(True)

```

```
plt.tight_layout()
```

```
Exibindo o gráfico
plt.show()
```

|    | Ano da corrida | Velocidade média |
|----|----------------|------------------|
| 0  | 2002           | 12.791514        |
| 1  | 2006           | 11.365393        |
| 2  | 2007           | 11.723351        |
| 3  | 2008           | 11.092897        |
| 4  | 2009           | 10.573283        |
| 5  | 2010           | 9.487845         |
| 6  | 2011           | 11.257393        |
| 7  | 2012           | 7.743460         |
| 8  | 2013           | 9.057333         |
| 9  | 2014           | 7.154052         |
| 10 | 2015           | 6.777803         |
| 11 | 2016           | 6.544318         |
| 12 | 2017           | 8.325000         |
| 13 | 2018           | 7.270399         |
| 14 | 2019           | 4.965818         |
| 15 | 2020           | 10.090111        |
| 16 | 2021           | 10.223000        |




O gráfico mostra a tendência da velocidade média dos atletas ao longo dos anos em corridas de 50 km. Inicialmente, é possível observar um declínio gradual no desempenho médio, especialmente entre 2002 e 2012, onde a velocidade média caiu de cerca de 13 km/h para 10 km/h. Essa redução pode ser atribuída a mudanças nas condições dos eventos, maior diversidade de participantes ou outros fatores externos que impactaram o desempenho.

Após 2012, a velocidade média apresenta uma oscilação mais evidente, com períodos de recuperação e declínio até 2020. É importante destacar o ponto mais baixo, registrado em torno de 2018, com uma velocidade média próxima a 5 km/h, seguido por uma recuperação significativa nos anos subsequentes. Esse padrão sugere que fatores como a organização dos eventos, condições climáticas ou mudanças na demografia dos participantes podem ter influenciado o desempenho geral. Analisar esses fatores em detalhes poderia oferecer mais insights sobre as variações observadas.

Quais são os eventos com as maiores velocidades médias globais e quais características eles têm em comum?

Identificaremos padrões entre eventos com alto desempenho, como localização, estação do ano ou características demográficas dos atletas.

```
run_BRA.dtypes
```



|                      |  | 0              |
|----------------------|--|----------------|
| Ano da corrida       |  | int64          |
| Data                 |  | datetime64[ns] |
| Nome do evento       |  | object         |
| Distância            |  | object         |
| Número de finalistas |  | int64          |
| Tempo do atleta      |  | object         |
| Pais do atleta       |  | object         |