In [1]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```python
pd.set_option('display.max_columns', None)
pd.set_option('display.precision', 2)
```

In [11]:

```python
tdf_riders = pd.read_csv("TDF_Riders_History.csv")
tdf_stages = pd.read_csv("TDF_Stages_History.csv")
tdff_riders = pd.read_csv("TDFF_Riders_History.csv")
tdff_stages = pd.read_csv("TDFF_Stages_History.csv")
```

In [13]:

```python
print("=== TDF Riders ===")
print(tdf_riders.head(), "\n")
print(tdf_riders.info(), "\n")
```

```
=== TDF Riders ===
   Unnamed: 0  Rank                  Rider  Rider No.          Team  \
0           0     1          MAURICE GARIN          1  TDF 1903 ***
1           1     2         LUCIEN POTHIER         37  TDF 1903 ***
2           2     3       FERNAND AUGEREAU         39  TDF 1903 ***
3           3     4        RODOLPHE MULLER         33  TDF 1903 ***
4           4     5  JEAN-BAPTISTE FISCHER         12  TDF 1903 ***

          Times             Gap    B    P  Year  Distance (km)  \
0  94h 33' 14''               -  NaN  NaN  1903           2428
1  97h 32' 35''  + 02h 59' 21''  NaN  NaN  1903           2428
2  99h 02' 38''  + 04h 29' 24''  NaN  NaN  1903           2428
3  99h 12' 44''  + 04h 39' 30''  NaN  NaN  1903           2428
4  99h 41' 58''  + 05h 08' 44''  NaN  NaN  1903           2428

   Number of stages  TotalSeconds  GapSeconds ResultType
0                 6        340394           0       time
1                 6        351155       10761       time
2                 6        356558       16164       time
3                 6        357164       16770       time
4                 6        358918       18524       time

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9878 entries, 0 to 9877
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Unnamed: 0        9878 non-null   int64 
 1   Rank              9878 non-null   int64 
 2   Rider             9878 non-null   object
 3   Rider No.         9878 non-null   int64 
 4   Team              9878 non-null   object
 5   Times             9878 non-null   object
 6   Gap               9878 non-null   object
 7   B                 222 non-null    object
 8   P                 56 non-null     object
 9   Year              9878 non-null   int64 
 10  Distance (km)     9878 non-null   int64 
 11  Number of stages  9878 non-null   int64 
 12  TotalSeconds      9878 non-null   int64 
 13  GapSeconds        9878 non-null   int64 
 14  ResultType        9804 non-null   object
dtypes: int64(8), object(7)
memory usage: 1.1+ MB
None
```

In [14]:

```
print("=== TDF Stages ===")
print(tdf_stages.head(), "\n")
print(tdf_stages.info(), "\n")
```

```
=== TDF Stages ===
   Unnamed: 0  Year  TotalTDFDistance                           Stage
0           0  1903              2428          Stage 1 : Paris > Lyon
1           1  1903              2428      Stage 2 : Lyon > Marseille
2           2  1903              2428  Stage 3 : Marseille > Toulouse
3           3  1903              2428   Stage 4 : Toulouse > Bordeaux
4           4  1903              2428     Stage 5 : Bordeaux > Nantes

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2365 entries, 0 to 2364
Data columns (total 4 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Unnamed: 0        2365 non-null   int64 
 1   Year              2365 non-null   int64 
 2   TotalTDFDistance  2365 non-null   int64 
 3   Stage             2365 non-null   object
dtypes: int64(3), object(1)
memory usage: 74.0+ KB
None
```

In [15]:

```
print("=== TDFF Riders ===")
print(tdff_riders.head(), "\n")
print(tdff_riders.info(), "\n")
```

```
=== TDFF Riders ===
   Unnamed: 0  Rank                 Rider  Rider No.  \
0           0     1  ANNEMIEK VAN VLEUTEN         11   
1           1     2        DEMI VOLLERING         21   
2           2     3  KATARZYNA NIEWIADOMA         61   
3           3     4       JULIETTE LABOUS         51   
4           4     5        SILVIA PERSICO        104   

                        Team         Times             Gap    B    P  Year  \
0        MOVISTAR TEAM WOMEN  26h 55' 44''               -  23'  NaN  2022   
1               TEAM SD WORX  26h 59' 32''  + 00h 03' 48''  14'  NaN  2022   
2      CANYON // SRAM RACING  27h 02' 19''  + 00h 06' 35''  06'  NaN  2022   
3                   TEAM DSM  27h 03' 12''  + 00h 07' 28''  NaN  NaN  2022   
4  VALCAR - TRAVEL & SERVICE  27h 03' 44''  + 00h 08' 00''  10'  NaN  2022   

   Distance (km)  Number of stages  TotalSeconds  GapSeconds ResultType  
0           1029                 8         96944           0       time  
1           1029                 8         97172         228       time  
2           1029                 8         97339         395       time  
3           1029                 8         97392         448       time  
4           1029                 8         97424         480       time   

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 232 entries, 0 to 231
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Unnamed: 0        232 non-null    int64 
 1   Rank              232 non-null    int64 
 2   Rider             232 non-null    object
 3   Rider No.         232 non-null    int64 
 4   Team              232 non-null    object
 5   Times             232 non-null    object
 6   Gap               232 non-null    object
 7   B                 32 non-null     object
 8   P                 4 non-null      object
 9   Year              232 non-null    int64 
 10  Distance (km)     232 non-null    int64 
 11  Number of stages  232 non-null    int64 
 12  TotalSeconds      232 non-null    int64 
 13  GapSeconds        232 non-null    int64 
 14  ResultType        232 non-null    object
dtypes: int64(8), object(7)
memory usage: 27.3+ KB
None
```

```
In [16]: print("=== TDFF Stages ===")
         print(tdff_stages.head(), "\n")
         print(tdff_stages.info(), "\n")
```

```
=== TDFF Stages ===
   Unnamed: 0  Year  TotalTDFDistance  \
0           0  2022              1029   
1           1  2022              1029   
2           2  2022              1029   
3           3  2022              1029   
4           4  2022              1029   

                                          Stage  
0  Stage 1 : Paris Tour Eiffel > Champs-Élysées  
1                   Stage 2 : Meaux > Provins  
2                   Stage 3 : Reims > Épernay  
3             Stage 4 : Troyes > Bar-sur-Aube  
4   Stage 5 : Bar-le-Duc > Saint-Dié-des-Vosges  

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16 entries, 0 to 15
Data columns (total 4 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Unnamed: 0        16 non-null     int64 
 1   Year              16 non-null     int64 
 2   TotalTDFDistance  16 non-null     int64 
 3   Stage             16 non-null     object
dtypes: int64(3), object(1)
memory usage: 640.0+ bytes
None
```

In [17]:

```python
print("Dimensiones:")
print(f"TDF Riders: {tdf_riders.shape}")
print(f"TDF Stages: {tdf_stages.shape}")
print(f"TDFF Riders: {tdff_riders.shape}")
print(f"TDFF Stages: {tdff_stages.shape}\n")
```

```
Dimensiones:
TDF Riders: (9878, 15)
TDF Stages: (2365, 4)
TDFF Riders: (232, 15)
TDFF Stages: (16, 4)
```

In [18]:

```python
print("Valores nulos por dataset:")
print("TDF Riders:\n", tdf_riders.isnull().sum(), "\n")
print("TDF Stages:\n", tdf_stages.isnull().sum(), "\n")
print("TDFF Riders:\n", tdff_riders.isnull().sum(), "\n")
print("TDFF Stages:\n", tdff_stages.isnull().sum(), "\n")
```

```
Valores nulos por dataset:
TDF Riders:
 Unnamed: 0             0
Rank                   0
Rider                  0
Rider No.              0
Team                   0
Times                  0
Gap                    0
B                   9656
P                   9822
Year                   0
Distance (km)          0
Number of stages       0
TotalSeconds           0
GapSeconds             0
ResultType            74
dtype: int64

TDF Stages:
 Unnamed: 0          0
Year                0
TotalTDFDistance    0
Stage               0
dtype: int64

TDFF Riders:
 Unnamed: 0            0
Rank                  0
Rider                 0
Rider No.             0
Team                  0
Times                 0
Gap                   0
B                   200
P                   228
Year                  0
Distance (km)         0
Number of stages      0
TotalSeconds          0
GapSeconds            0
ResultType            0
dtype: int64

TDFF Stages:
 Unnamed: 0          0
Year                0
TotalTDFDistance    0
Stage               0
dtype: int64
```

In [20]:

```
print("Tipos de datos TDF Riders:\n", tdf_riders.dtypes, "\n")
```

```
Tipos de datos TDF Riders:
 Unnamed: 0           int64
Rank                  int64
Rider                object
Rider No.             int64
Team                 object
Times                object
Gap                  object
B                    object
P                    object
Year                  int64
Distance (km)         int64
Number of stages      int64
TotalSeconds          int64
GapSeconds            int64
ResultType           object
dtype: object
```

```
In [21]: print("Tipos de datos TDFF Riders:\n", tdff_riders.dtypes, "\n")
```

```
Tipos de datos TDFF Riders:
 Unnamed: 0           int64
Rank                  int64
Rider                object
Rider No.             int64
Team                 object
Times                object
Gap                  object
B                    object
P                    object
Year                  int64
Distance (km)         int64
Number of stages      int64
TotalSeconds          int64
GapSeconds            int64
ResultType           object
dtype: object
```

```
In [22]: print("Resumen numérico TDF Riders:\n", tdf_riders.describe(), "\n")
```

```
Resumen numérico TDF Riders:
        Unnamed: 0     Rank  Rider No.     Year  Distance (km)  \
count     9878.00  9878.00    9878.00  9878.00        9878.00   
mean      4938.50    58.18      89.98  1982.89        3943.07   
std       2851.68    41.22      60.61    30.26         571.17   
min          0.00     1.00       1.00  1903.00        2428.00   
25%       2469.25    24.00      37.00  1964.00        3504.00   
50%       4938.50    50.00      84.00  1989.00        3809.00   
75%       7407.75    87.00     134.00  2008.00        4254.00   
max       9877.00   174.00     321.00  2023.00        5745.00   

       Number of stages  TotalSeconds  GapSeconds  
count           9878.00      9.88e+03     9878.00  
mean              22.13      4.14e+05     9520.31  
std                2.95      1.87e+05    13527.12  
min                6.00      0.00e+00        0.00  
25%               21.00      3.22e+05     3938.00  
50%               22.00      3.53e+05     7494.00  
75%               24.00      4.29e+05    10922.50  
max               31.00      2.15e+06   178645.00  
```

In [23]:

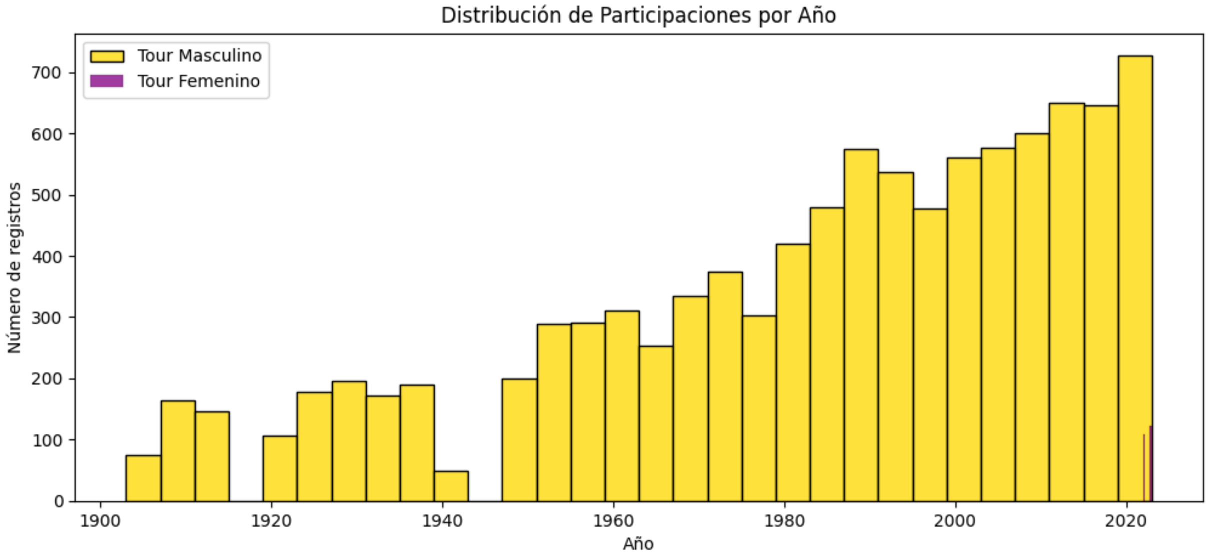
```python
print("Resumen numérico TDFF Riders:\n", tdff_riders.describe(), "\n")
```

```
Resumen numérico TDFF Riders:
        Unnamed: 0    Rank  Rider No.     Year  Distance (km)  \
count      232.00  232.00     232.00   232.00         232.00   
mean       115.50   58.71     110.23  2022.53         990.30   
std         67.12   33.92      67.29     0.50          36.51   
min          0.00    1.00       1.00  2022.00         956.00   
25%         57.75   29.75      51.75  2022.00         956.00   
50%        115.50   58.50     111.00  2023.00         956.00   
75%        173.25   87.25     168.00  2023.00        1029.00   
max        231.00  123.00     236.00  2023.00        1029.00   

       Number of stages  TotalSeconds  GapSeconds  
count             232.0        232.00      232.00  
mean                8.0      97211.87     3390.05  
std                 0.0       3403.13     1687.37  
min                 8.0      91055.00        0.00  
25%                 8.0      94356.25     2215.50  
50%                 8.0      97196.50     3460.50  
75%                 8.0     100430.50     4562.75  
max                 8.0     103790.00     7820.00  
```

In [24]:

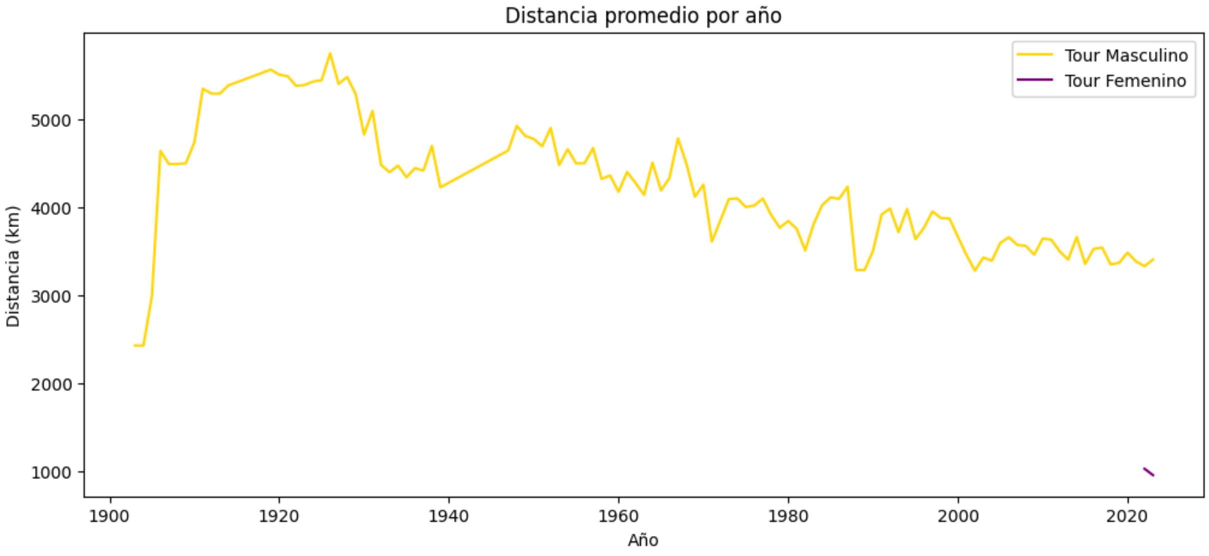
```python
plt.figure(figsize=(12,5))
sns.histplot(tdf_riders['Year'], bins=30, color="gold", label="Tour Masculir
sns.histplot(tdff_riders['Year'], bins=5, color="purple", label="Tour Femeni
plt.title("Distribución de Participaciones por Año")
plt.xlabel("Año")
plt.ylabel("Número de registros")
plt.legend()
plt.show()
```

```
/Users/carloslopez/Library/Python/3.9/lib/python/site-packages/seaborn/_oldc
ore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be r
emoved in a future version. Convert inf values to NaN before operating inste
ad.
  with pd.option_context('mode.use_inf_as_na', True):
/Users/carloslopez/Library/Python/3.9/lib/python/site-packages/seaborn/_oldc
ore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be r
emoved in a future version. Convert inf values to NaN before operating inste
ad.
  with pd.option_context('mode.use_inf_as_na', True):
```

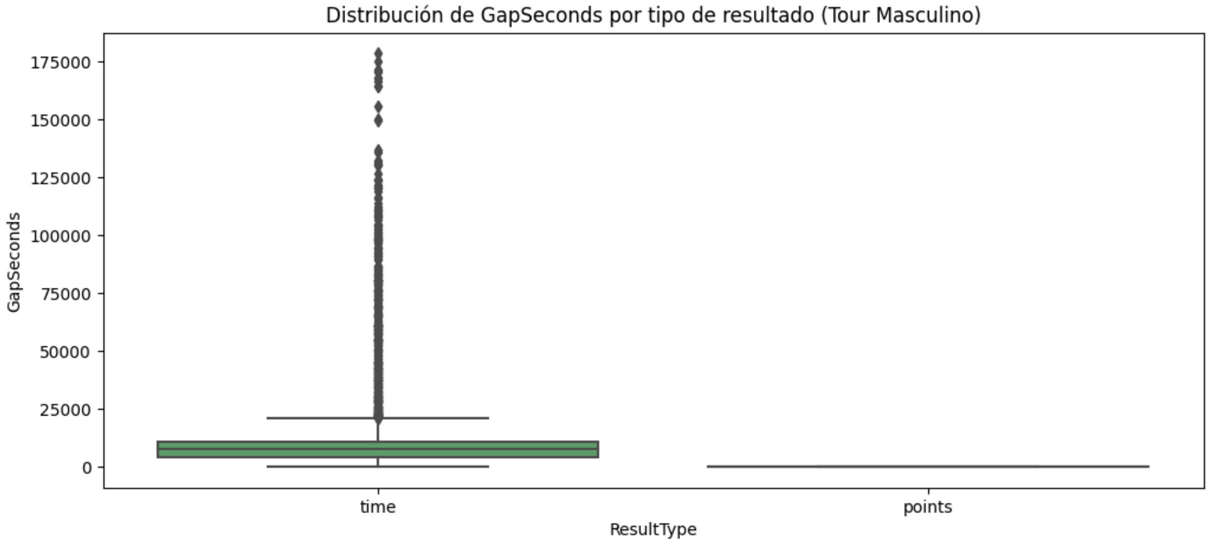


In [26]:

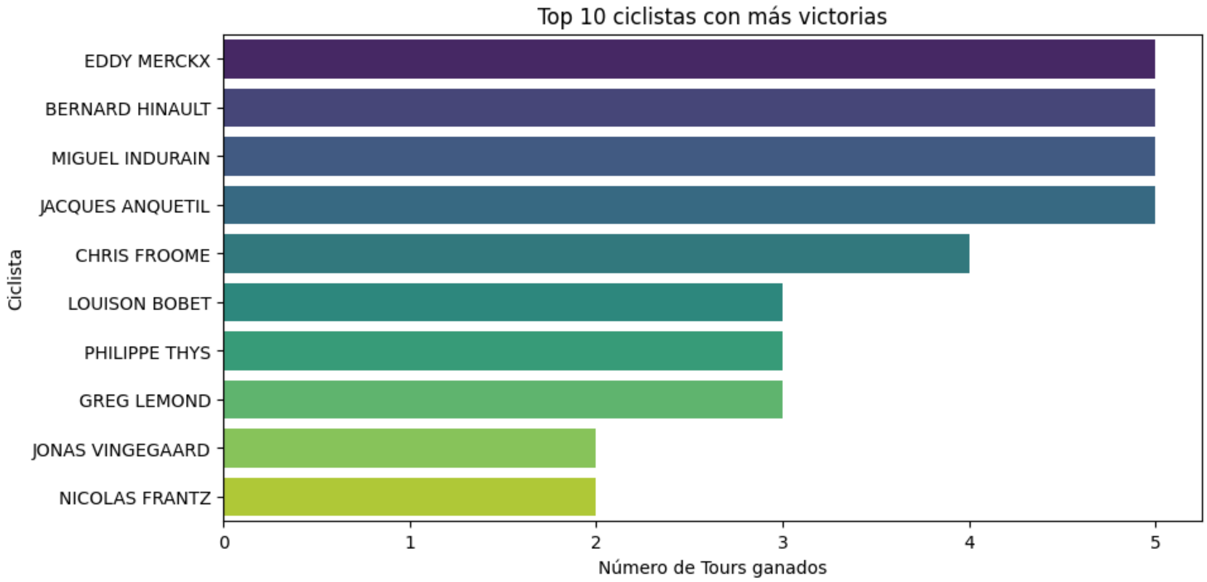
```
tdf_distances = tdf_riders.groupby("Year")["Distance (km)"].mean().reset_inc
tdff_distances = tdff_riders.groupby("Year")["Distance (km)"].mean().reset_i

plt.figure(figsize=(12,5))
plt.plot(tdf_distances["Year"], tdf_distances["Distance (km)"], label="Tour 
plt.plot(tdff_distances["Year"], tdff_distances["Distance (km)"], label="Tou
plt.title("Distancia promedio por año")
plt.xlabel("Año")
plt.ylabel("Distancia (km)")
plt.legend()
plt.show()
```



In [27]:
```
plt.figure(figsize=(12,5))
sns.boxplot(x="ResultType", y="GapSeconds", data=tdf_riders, palette="summer
plt.title("Distribución de GapSeconds por tipo de resultado (Tour Masculino)
plt.show()
```



In [28]:
```
top_winners = tdf_riders[tdf_riders["Rank"] == 1]["Rider"].value_counts().he
plt.figure(figsize=(10,5))
sns.barplot(x=top_winners.values, y=top_winners.index, palette="viridis")
plt.title("Top 10 ciclistas con más victorias")
plt.xlabel("Número de Tours ganados")
plt.ylabel("Ciclista")
plt.show()
```



Conclusiones

1. **Cobertura y volumen de datos**

   - El Tour masculino (TDF) tiene 9,878 registros de ciclistas y 2,365 registros de etapas, cubriendo de 1903 a 2023.
   - El Tour femenino (TDFF) solo cuenta con 232 registros de ciclistas y 16 etapas, iniciando en 2022.
   - Esto implica que el análisis comparativo a largo plazo será posible solo para el Tour masculino.

2. **Valores nulos y calidad de datos**

   - En TDF Riders, las columnas B y P presentan más del 97% de valores nulos, y ResultType tiene 74 registros vacíos.
   - En TDFF Riders, las columnas B y P también presentan alta ausencia de datos.
   - Estos campos requerirán tratamiento en la fase de preparación de datos.

3. **Participaciones por año**

   - El número de ciclistas ha aumentado con el tiempo, con caídas notorias en periodos de guerra (1914–1918, 1940–1946).
   - El Tour femenino muestra un volumen de participación mucho menor y reciente.

4. **Distancia promedio**

   - En las primeras décadas del Tour masculino, las distancias superaban los 5,000 km; en décadas recientes se estabilizaron en torno a 3,500–4,000 km.
   - El Tour femenino mantiene distancias cercanas a 1,000 km.

5. **Gaps y rendimiento**

   - La distribución de GapSeconds en el Tour masculino para time presenta una gran dispersión y outliers significativos, lo que indica diferencias marcadas en el rendimiento.
   - Los resultados por puntos (points) muestran gaps cercanos a cero.

6. **Ganadores históricos**

   - Eddy Merckx, Bernard Hinault, Miguel Indurain y Jacques Anquetil lideran con 5 victorias cada uno.
   - Existe un patrón histórico de dominio por pocos corredores en determinadas épocas.