

# Trực quan hóa dữ liệu Premier League mùa giải 2022 - 2023

## Mô tả bộ dữ liệu: results overall

Thuộc tính	Mô tả
<b>Rk (Rank)</b>	Số thứ tự hoặc xếp hạng của đội bóng trong bảng xếp hạng
<b>Squad</b>	Tên chính thức của đội bóng tham gia Premier League.
<b>MP (Matches Played)</b>	Số trận đấu mà đội bóng đã tham gia trong mùa giải.
<b>W (Wins)</b>	Số lượng trận thắng mà đội bóng đã giành được trong mùa giải.
<b>D (Draws)</b>	Số lượng trận hòa mà đội bóng đã có trong mùa giải.
<b>L (Losses)</b>	Số lượng trận thua mà đội bóng đã phải chịu trong mùa giải.
<b>GF (Goals For)</b>	Tổng số bàn thắng mà đội bóng đã ghi được trong mùa giải.
<b>GA (Goals Against)</b>	Tổng số bàn thua mà đội bóng đã phải nhận trong mùa giải.
<b>GD (Goal Difference)</b>	Chênh lệch giữa số bàn thắng (GF) và số bàn thua (GA). Được tính bằng cách trừ số bàn thua từ số bàn thắng (GD = GF - GA).
<b>Pts (Points)</b>	Tổng số điểm mà đội bóng đã thu được trong mùa giải. Điểm số được tính dựa trên số trận thắng (W), số trận hòa (D), và quy tắc điểm của giải đấu.
<b>Pts/MP (Points per Match)</b>	Số điểm trung bình mà đội bóng kiếm được trên mỗi trận đấu. Được tính bằng cách chia tổng số điểm (Pts) cho tổng số trận đấu (MP).
<b>xG (Expected Goals)</b>	Dự đoán số bàn thắng mà đội bóng nên ghi dựa trên xác suất và hiệu suất trong các tình huống tấn công.
<b>xGA (Expected Goals Against)</b>	Dự đoán số bàn thua mà đội bóng nên nhận dựa trên xác suất và hiệu suất trong các tình huống phòng ngự.
<b>xGD (Expected Goal Difference)</b>	Chênh lệch giữa số bàn thắng dự kiến (xG) và số bàn thua dự kiến (xGA). Tương tự như GD, nhưng dựa trên dự đoán xác suất.
<b>xGD/90 (Expected Goal Difference per 90 Minutes)</b>	Chênh lệch giữa số bàn thắng dự kiến (xG) và số bàn thua dự kiến (xGA) trung bình trên mỗi 90 phút thi đấu.
<b>Attendance</b>	Số lượng khán giả trung bình tham dự trận đấu của đội bóng

## Trực quan hóa dữ liệu

In [3]:

```
!pip install plotly
```

```
Requirement already satisfied: plotly in c:\users\lenovo\anaconda3\lib\site-packages (5.17.0)
Requirement already satisfied: tenacity>=6.2.0 in c:\users\lenovo\anaconda3\lib\site-packages (from plotly) (8.2.3)
Requirement already satisfied: packaging in c:\users\lenovo\anaconda3\lib\site-packages (from plotly) (20.9)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\lenovo\anaconda3\lib\site-packages (from packaging->plotly) (2.4.7)
```

DEPRECATION: pyodbc 4.0.0-unsupported has a non-standard version number. pip 23.3 will enforce this behaviour change. A possible replacement is to upgrade to a newer version of pyodbc or contact the author to suggest that they release a version with a conforming version number. Discussion can be found at <https://github.com/pypa/pip/issues/12063>

## Sử dụng plotly cho phép tạo biểu đồ tương tác, màu mè hơn so với matplotlib (biểu đồ tĩnh)

In [4]:

```
import pandas as pd
import numpy as np
import plotly.express as px

df = pd.read_csv("results2022-202391_overall.csv")
df.head()
```

Out[4]:

	Unnamed: 0	Rk	Squad	MP	W	D	L	GF	GA	GD	Pts	Pts/MP	xG	xGA	xDG	xDG/90	At
0	0	1	Manchester City	38	28	5	5	94	33	61	89	2.34	78.7	32.1	46.6	1.23	
1	1	2	Arsenal	38	26	6	6	88	43	45	84	2.21	71.9	42.0	29.9	0.79	
2	2	3	Manchester Utd	38	23	6	9	58	43	15	75	1.97	67.7	50.4	17.3	0.45	
3	3	4	Newcastle Utd	38	19	14	5	68	33	35	71	1.87	72.0	39.6	32.4	0.85	
4	4	5	Liverpool	38	19	10	9	75	47	28	67	1.76	72.6	50.9	21.7	0.57	

◀ ▶

In [5]:

```
# Số Lượng mẫu và thuộc tính
print("Số bản ghi : " +str(df.shape[0]))
print("Số trường : " +str(df.shape[1]))
```

Số bản ghi : 20  
 Số trường : 20

In [6]:

```
#Kiểm tra dữ liệu khuyết thiếu
miss_value = df.isnull().sum()
miss_value
```

```
Out[6]: Unnamed: 0          0
Rk              0
Squad           0
MP              0
W               0
D               0
L               0
GF              0
GA              0
GD              0
Pts             0
Pts/MP          0
xG              0
xGA             0
xGD             0
xGD/90          0
Attendance       0
Top Team Scorer 0
Goalkeeper      0
Notes            10
dtype: int64
```

**Thiếu dữ liệu ở cột Notes => không đáng quan ngại**

### Tiền xử lý

```
In [7]: # Sử dụng drop để xóa cột 'Unnamed'
df = df.drop('Unnamed: 0', axis=1)
```

```
In [8]: # Hiển thị kiểu dữ liệu của các thuộc tính
df.dtypes
```

```
Out[8]: Rk                  int64
Squad               object
MP                  int64
W                   int64
D                   int64
L                   int64
GF                  int64
GA                  int64
GD                  int64
Pts                 int64
Pts/MP              float64
xG                  float64
xGA                 float64
xGD                 float64
xGD/90              float64
Attendance          object
Top Team Scorer     object
Goalkeeper          object
Notes                object
dtype: object
```

```
In [9]: # Đổi dữ liệu cột "Attendance" về int
df.Attendance = df.Attendance.str.replace(',', '').astype(int)
```

```
In [10]: df.head(5)
```

Out[10]:

	Rk	Squad	MP	W	D	L	GF	GA	GD	Pts	Pts/MP	xG	xGA	xGD	xGD/90	Attendance
0	1	Manchester City	38	28	5	5	94	33	61	89	2.34	78.7	32.1	46.6	1.23	53249
1	2	Arsenal	38	26	6	6	88	43	45	84	2.21	71.9	42.0	29.9	0.79	60191
2	3	Manchester Utd	38	23	6	9	58	43	15	75	1.97	67.7	50.4	17.3	0.45	73671 F
3	4	Newcastle Utd	38	19	14	5	68	33	35	71	1.87	72.0	39.6	32.4	0.85	52127
4	5	Liverpool	38	19	10	9	75	47	28	67	1.76	72.6	50.9	21.7	0.57	53163

## 1. Thống kê số điểm của các đội Ngoại hạng Anh (EPL) mùa giải 2022–2023.

In [11]:

```
fig = px.bar(df,
              x=df.Squad,
              y=df.Pts,
              title="Total Points : EPL 2022 - 2023",
              color=df.Pts,
              text=df.Pts,
              color_continuous_scale="blues",
              height=600
)
fig.show()
```

MC vị trí dẫn đầu => vô địch mùa giải 2022-2023, top 4 gồm: MC, Arsenal, MU, Newcastle Utd, trong đó thể hiện rõ mức chênh lệch giữa top 2 và top 4 . Hai đội đứng đầu MC và Arsenal sẽ tự động được tham dự vòng bảng của UEFA Champions League.

Top 3 từ dưới lên: Leeds, Leicester và Southampton là đội phải chịu cảnh xuống hạng

## 2. Thống kê tổng số trận đấu đã chơi của các đội (so sánh giữa số trận đấu đã chơi và số điểm mà mỗi đội kiếm được)

In [12]:

```
fig = px.bar(df,
              x='Squad',
              y=['MP', 'Pts'],
              barmode='group',
              labels={'value': 'Count'},
              title='Matches Played (MP) and Points (Pts) : EPL 2022 - 2023',
              text_auto=True,
              color_discrete_sequence=["gray", "deepskyblue"])
fig.show()
```

### 3. Trực quan số người tham dự trung bình ở mỗi sân vận động

In [13]:

```
fig = px.bar(df,
              x=df.Squad,
              y=df.Attendance,
              title="Average Attendance : EPL 2022 - 2023",
              text=df.Attendance,
              color=df.Attendance,
              color_continuous_scale="greens",
              height=600
)
fig.show()
```

Rõ ràng trung bình khán giả tham dự tại sân của MU đông nhất: 73671 người (điều này có thể được lý giải do Mu có lịch sử dài và truyền thống mạnh mẽ trong giới bóng đá thu hút cộng đồng người hâm mộ đông đảo, sức chứa của Old Trafford thuộc câu lạc bộ lớn nhất Vương quốc Anh ) thứ hai là Tottenham: 61585 người (White Hart Lane - sân vận động bóng đá lớn thứ hai sau Old Trafford), thấp nhất là Bournemouth: 10362

#### 4.Thống kê số bàn thắng ghi được (Số bàn thắng cho — GF) và số bàn thua thủng lưới (Số bàn thua vào lưới — GA)

In [14]:

```
fig = px.bar(df,
              x='Squad',
              y=['GF', 'GA'],
              color_discrete_sequence=["darkgreen", "tomato"],
              barmode='stack',
              text_auto=True,
              labels={'value': 'Count'},
              title='Goals For, Goals Against : EPL 2022 - 2023',
              height=600
)
fig.show()
```

## 5. Thống kê hiệu số bàn thắng bại của mỗi đội

In [16]:

```
fig = px.bar(df,
              x=df.Squad,
              y=df.GD,
              title="Goal Difference : EPL 2022 - 2023",
              text=df.GD,
              color=df.GD,
              color_continuous_scale="ylgn",
              height=600
)
fig.show()
```

Dễ thấy, MC dẫn đầu với mức hênh lệch giữa số bàn thắng (GF) và số bàn thua (GA): +61 bàn - một thành tích bỏ xa các đối thủ trong bản xếp hạng cả mùa 22-23, thành tích tệ nhất thuộc về Southampton: -37 bàn

## 6. Thống kê về số trận thắng, trận hòa và trận thua của mỗi đội

In [19]:

```
fig = px.bar(df,
              x='Squad',
              y=['W', 'D', 'L'],
              barmode='group',
              title='Wins, Draws, Losses : EPL 2022 - 2023',
              color_discrete_sequence=["green","gold","tomato"],
              text_auto=True,
              height=600
)
fig.show()
```

**Thành tích tốt nhất vẫn thuộc về MC trong tổng số 38 trận đấu xuyên suốt mùa giải, thắng 28, hòa 5, thua 5. Top 2 thuộc về Arsenal với hiệu số lần lượt là 26, 6, 6**

## 6. Thống kê hiệu quả của mỗi đội về số cơ hội được tạo ra (bàn thắng dự kiến) và cơ hội bị thủng lưới (bàn thua dự kiến)

In [20]:

```
fig = px.scatter(df,
                  x='xG',
                  y='xGA',
                  size='xG',
                  color='Squad',
                  title="Team Efficiency: Chances Created Vs Chanced Conceded : EPL 2022",
                  text=df['Squad'].str[:4],
                  height=600,
                  )

fig.update_layout(
    xaxis=dict(title='Expected Goals (xG): Every goal-scoring chance, and the likelihood'),
    yaxis=dict(title='Expected Goals Against Team: (xGA)')
)
fig.show()
```

- xG:Dự đoán số bàn thắng mà đội bóng nên ghi dựa trên xác suất và hiệu suất trong các tình huống tấn công
- xGA:Dự đoán số bàn thua mà đội bóng nên phải nhận dựa trên xác suất và hiệu suất trong các tình huống phòng ngự

=> MC đạt hiệu suất tốt nhất, trong top5 gồm: Arsenal, Newcastle, Brighton, Liverpool. Điều bất ngờ khi MU lại chỉ nằm ở top 6, cho thấy có sự bất ổn về phong độ của đội bóng (Về mặt thực tế MU cũng không đạt được phong độ cao trong xuyên suốt mùa giải 2022-2023)

## 7. Trực quan hóa cầu thủ ghi nhiều bàn thắng nhất cho mỗi đội

In [21]:

```
df[['Player', 'Goals']] = df['Top Team Scorer'].str.split(' - ', expand=True)
df.Goals = df.Goals.astype(int)

fig = px.bar(df,
              y='Squad',
              x='Goals',
              color='Goals',
              color_continuous_scale="ylorrd",
              text='Top Team Scorer',
              title='Top Scorer for Each Team : EPL 2022 - 2023',
              height=600
)
fig.show()
```

**Erling Haaland** là cái tên dẫn đầu danh sách ghi bàn xuyên suốt mùa giải với 36 bàn (về mặt thực tế đây cũng là cầu thủ đạt phong độ rất cao trong thời gian gần đây, được đánh giá là cầu thủ đắt giá nhất thế giới hiện tại: T6 - 2023 Theo định giá mới nhất của chuyên trang Transfermarkt, Erling Haaland có giá 180 triệu Euro (tăng 10 triệu Euro) để trở thành cầu thủ đắt giá nhất thế giới.) => phá vỡ kỷ lục về số bàn thắng ghi được trong EPL

**Top 2** là Harry Kane với thành tích 30 bàn, thực tế Harry Kane giữ được điểm rơi phong độ rất tốt trong mùa giải 22/23

In [ ]: