

# Exploratory Data Analysis - Premier League

## 2023/2024



*Muhammad Rafa Kurnia*



Premier  
League

# Deskripsi Dataset



*Sumber Dataset:*

*Kaggle - English Premier League Matches 2023/2024 Season*

*Dataset ini berisi informasi pertandingan English Premier League musim 2023/2024 dengan total 760 entri. Setiap baris merepresentasikan performa satu tim dalam satu pertandingan, mencakup berbagai aspek seperti hasil pertandingan, jumlah gol, expected goals (xG), kepemilikan bola (possession), formasi, jumlah penonton, dan data lainnya. Dataset ini cocok untuk analisis performa tim, efektivitas strategi, serta visualisasi statistik pertandingan sepanjang musim.*

# Tujuan Exploratory Data Analysis



*Menemukan dan Membersihkan Missing Value*

*Serta menemukan Insight menarik dari  
Dataset tentang Premiere League*

# Exploratory Data Analysis

Library yang digunakan

```
# Import Library yang dibutuhkan
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

✓ 23.9s

```
data = pd.read_csv('matches.csv')
data.head()
✓ 0.0s
```

	Unnamed: 0	Date	Time	Comp	Round	Day	Venue	Result	GF	GA	...	Match Report	Notes	Sh	SoT	Dist	FK	PK	PKatt	Season	Team
0	1	2023-08-11	20:00	Premier League	Matchweek 1	Fri	Away	W	3	0	...	Match Report	NaN	17.0	8.0	13.9	0.0	0	0	2024	ManchesterCity
1	3	2023-08-19	20:00	Premier League	Matchweek 2	Sat	Home	W	1	0	...	Match Report	NaN	14.0	4.0	17.9	0.0	0	0	2024	ManchesterCity
2	4	2023-08-27	14:00	Premier League	Matchweek 3	Sun	Away	W	2	1	...	Match Report	NaN	29.0	9.0	17.3	2.0	0	1	2024	ManchesterCity
3	5	2023-09-02	15:00	Premier League	Matchweek 4	Sat	Home	W	5	1	...	Match Report	NaN	6.0	4.0	14.8	0.0	1	1	2024	ManchesterCity
4	6	2023-09-16	15:00	Premier League	Matchweek 5	Sat	Away	W	3	1	...	Match Report	NaN	29.0	13.0	16.4	1.0	0	0	2024	ManchesterCity

5 rows × 28 columns

```
data.tail()
✓ 0.0s
```

	Unnamed: 0	Date	Time	Comp	Round	Day	Venue	Result	GF	GA	...	Match Report	Notes	Sh	SoT	Dist	FK	PK	PKatt	Season	Team
755	36	2024-04-24	20:00	Premier League	Matchweek 29	Wed	Away	L	2	4	...	Match Report	NaN	10.0	4.0	17.8	1.0	0	0	2024	SheffieldUnited
756	37	2024-04-27	15:00	Premier League	Matchweek 35	Sat	Away	L	1	5	...	Match Report	NaN	15.0	4.0	13.5	0.0	0	0	2024	SheffieldUnited
757	38	2024-05-04	15:00	Premier League	Matchweek 36	Sat	Home	L	1	3	...	Match Report	NaN	16.0	4.0	18.0	0.0	1	1	2024	SheffieldUnited
758	39	2024-05-11	15:00	Premier League	Matchweek 37	Sat	Away	L	0	1	...	Match Report	NaN	13.0	1.0	21.0	0.0	0	0	2024	SheffieldUnited
759	40	2024-05-19	16:00	Premier League	Matchweek 38	Sun	Home	L	0	3	...	Match Report	NaN	6.0	1.0	18.0	1.0	0	0	2024	SheffieldUnited

5 rows × 28 columns

Tampilan Lima Data Awal dan Lima Data Terakhir

```
print(data.info())
0.0s

1 Date 760 non-null object
2 Time 760 non-null object
3 Comp 760 non-null object
4 Round 760 non-null object
5 Day 760 non-null object
6 Venue 760 non-null object
7 Result 760 non-null object
8 GF 760 non-null int64
9 GA 760 non-null int64
10 Opponent 760 non-null object
11 xG 760 non-null float64
12 xGA 760 non-null float64
13 Poss 760 non-null float64
14 Attendance 760 non-null float64
15 Captain 760 non-null object
16 Formation 760 non-null object
17 Referee 760 non-null object
18 Match Report 760 non-null object
19 Notes 0 non-null float64
20 Sh 760 non-null float64
21 SoT 760 non-null float64
22 Dist 760 non-null float64
23 FK 760 non-null float64
24 PK 760 non-null int64
25 PKatt 760 non-null int64
26 Season 760 non-null int64
27 Team 760 non-null object
dtypes: float64(9), int64(6), object(13)
memory usage: 166.4+ KB
None
```

## *Informasi dari data tersebut*

### Struktur Dataset:

Jumlah baris (data pertandingan): 760

Jumlah kolom (fitur): 28

### Tipe Data Kolom:

object (13 kolom): berisi data teks, seperti Team, Result, Venue, dll.

int64 (6 kolom): data numerik bulat, seperti GF (Goals For), PK, dll.

float64 (9 kolom): data numerik desimal, seperti xG, Poss, Attendance, dll

### Kolom Kosong:

Notes berisi 0 data (kosong total) → bisa dihapus.

### Kesimpulan Singkat:

Dataset sudah cukup rapi:

Tidak ada nilai kosong kecuali Notes

Tipe data sudah tepat, kecuali Date perlu dikonversi ke datetime

```
# Menampilkan jumlah nilai hilang per kolom
missing_data = data.isnull().sum()
print(missing_data)

[9]    ✓ 0.0s

... Unnamed: 0      0
Date          0
Time          0
Comp          0
Round         0
Day           0
Venue          0
Result         0
GF            0
GA            0
Opponent       0
xG            0
xGA           0
Poss           0
Attendance     0
Captain        0
Formation       0
Referee         0
Match Report    0
Notes          760
Sh             0
SoT            0
Dist           0
FK             0
PK             0
PKatt          0
Season          0
Team            0
dtype: int64
```

## Mengecek Missing Value

Dataset ini terdiri dari 28 kolom dan 760 baris data, dengan sebagian besar kolom terisi lengkap tanpa nilai kosong. Hanya satu kolom, yaitu **Notes**, yang seluruhnya kosong dan tidak mengandung informasi yang berguna untuk analisis. Oleh karena itu, kolom ini dapat dihapus dalam tahap pembersihan data. Secara keseluruhan, dataset ini dalam kondisi baik dan siap untuk dianalisis lebih lanjut.

*Menghapus kolom Notes karena seluruhnya kosong*

```
data.drop(columns=['Notes', 'Unnamed: 0'], inplace=True)

✓ 0.0s
```

## Mengecek Duplikat Data

```
# Mengecek jumlah baris yang duplikat
data.duplicated().sum()

[12]   ✓ 0.2s

... np.int64(0)
```

Menunjukkan bahwa hasil pengecekan duplikasi adalah nol, artinya tidak ditemukan baris data yang duplikat dalam dataset.

```
print(data.info())
✓ 0.0s

1 Date    760 non-null   object
2 Time    760 non-null   object
3 Comp    760 non-null   object
4 Round   760 non-null   object
5 Day     760 non-null   object
6 Venue   760 non-null   object
7 Result  760 non-null   object
8 GF      760 non-null   int64
9 GA      760 non-null   int64
10 Opponent 760 non-null   object
11 xG      760 non-null   float64
12 xGA     760 non-null   float64
13 Poss    760 non-null   float64
14 Attendance 760 non-null   float64
15 Captain 760 non-null   object
16 Formation 760 non-null   object
17 Referee 760 non-null   object
18 Match Report 760 non-null   object
19 Notes   0 non-null    float64
20 Sh      760 non-null   float64
21 SoT     760 non-null   float64
22 Dist    760 non-null   float64
23 FK      760 non-null   float64
24 PK      760 non-null   int64
25 PKatt   760 non-null   int64
26 Season  760 non-null   int64
27 Team    760 non-null   object
dtypes: float64(9), int64(6), object(13)
memory usage: 166.4+ KB
None
```

## Mengubah Type Data

Kolom Date saat ini bertipe object, yang berarti data ini disimpan sebagai teks biasa, bukan sebagai tanggal yang dapat diolah. Agar dapat melakukan analisis berbasis waktu (misalnya analisis tren berdasarkan bulan, musim, atau tahun), kolom ini perlu diubah menjadi tipe data datetime menggunakan kode dibawah

```
# Konversi Date ke type data yang benar
data['Date'] = pd.to_datetime(data['Date'])
```

✓ 0.3s

## *Ringkasan Numerik*

---

Dataset ini berisi data pertandingan Liga Premier Inggris 2023/2024 dengan 760 baris dan 28 kolom. Rata-rata gol yang dicetak (GF) dan kebobolan (GA) per pertandingan adalah sekitar 1.64. Nilai rata-rata ekspektasi gol (xG) dan kebobolan (xGA) juga sekitar 1.55. Penguasaan bola rata-rata adalah 50%, dengan rata-rata penonton sekitar 38,613. Tembakan per pertandingan rata-rata 13.66, dengan 4.67 tembakan tepat sasaran. Ada variasi dalam data, seperti penalti yang diberikan (PK) dan yang dicoba (PKatt).

```
# Ringkasan statistik untuk data numerik
data.describe()

[✓ 0.0s]

      Date        GF        GA       xG       xGA      Poss  Attendance        Sh        SoT       Dist       FK       PK     PKatt  Season
0   count      760  760.000000  760.000000  760.000000  760.000000  760.000000  760.000000  760.000000  760.000000  760.000000  760.000000  760.000000  760.0
1   mean  2024-01-05 22:10:06.315789312  1.639474  1.639474  1.550263  1.550263  50.0000  38613.313158  13.660526  4.668421  16.711842  0.372368  0.126316  0.140789  2024.0
2   min   2023-08-11 00:00:00  0.000000  0.000000  0.000000  0.000000  18.0000  10290.000000  1.000000  0.000000  7.000000  0.000000  0.000000  0.000000  0.000000  2024.0
3  25%   2023-10-28 18:00:00  1.000000  1.000000  0.900000  0.900000  40.0000  24444.750000  9.000000  3.000000  14.900000  0.000000  0.000000  0.000000  0.000000  2024.0
4  50%   2023-12-30 00:00:00  1.000000  1.000000  1.400000  1.400000  50.0000  38181.000000  13.000000  4.000000  16.700000  0.000000  0.000000  0.000000  0.000000  2024.0
5  75%   2024-03-30 00:00:00  2.000000  2.000000  2.100000  2.100000  60.0000  53371.750000  17.000000  6.000000  18.400000  1.000000  0.000000  0.000000  0.000000  2024.0
6   max   2024-05-19 00:00:00  8.000000  8.000000  7.000000  7.000000  82.0000  73612.000000  36.000000  15.000000  39.900000  3.000000  2.000000  2.000000  2.000000  2024.0
7    std      NaN  1.331297  1.331297  0.891314  0.891314  13.7495  17952.911952  6.058391  2.645612  2.856164  0.598218  0.359097  0.377104  0.0
```

## *Ringkasan Kategorikal*

---

Tabel ini memberikan ringkasan statistik untuk kolom kategori dalam dataset, seperti jumlah nilai unik, nilai yang paling sering muncul, dan frekuensi kemunculannya. Misalnya, pada kolom "Time," terdapat 15 waktu yang tercatat, dengan "15:00" muncul paling sering. Kolom "Comp" hanya memiliki satu nilai, yaitu "Premier League," sementara kolom "Result" memiliki 3 hasil, dengan "W" (menang) paling sering muncul.

```
# Ringkasan distribusi untuk kolom kategorikal
data.describe(include=['object'])

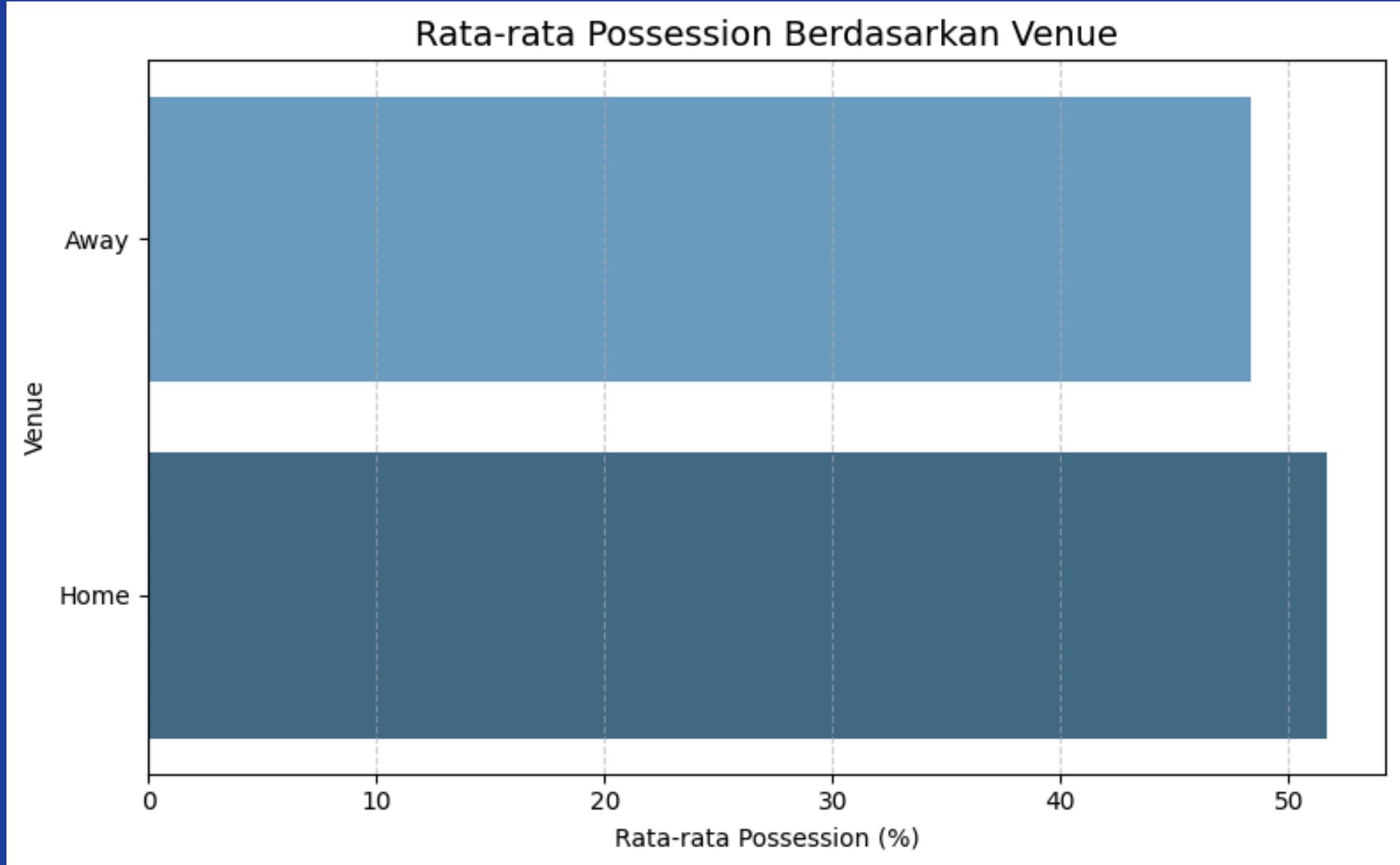
[✓ 0.1s]

   Time          Comp     Round      Day    Venue  Result Opponent   Captain Formation Referee Match Report        Team
count  760         760      760     760     760    760       760      760       760      760      760      760
unique   15           1       38       7       2       3       20       69       18       28       1       20
top    15:00  Premier League Matchweek 1      Sat     Away        W  Burnley Max Kilman 4-2-3-1 Anthony Taylor Match Report ManchesterCity
freq   264         760      20      390     380    298       38       37      296       54      760       38
```

# Data Analytics Visualization

*Setelah melalui tahap pembersihan dan eksplorasi data, kini saatnya menyajikan informasi dalam bentuk visual. Visualisasi data membantu kita memahami pola, tren, dan hubungan dalam dataset dengan lebih cepat dan intuitif. Dengan memanfaatkan grafik, diagram, dan chart, insight yang tersembunyi dapat diungkap secara lebih jelas, akurat, dan menarik. Tahap ini menjadi kunci dalam menyampaikan hasil analisis secara efektif kepada audiens non-teknis maupun pengambil keputusan. Berikut beberapa insight yang didapat dari Data Tersebut*

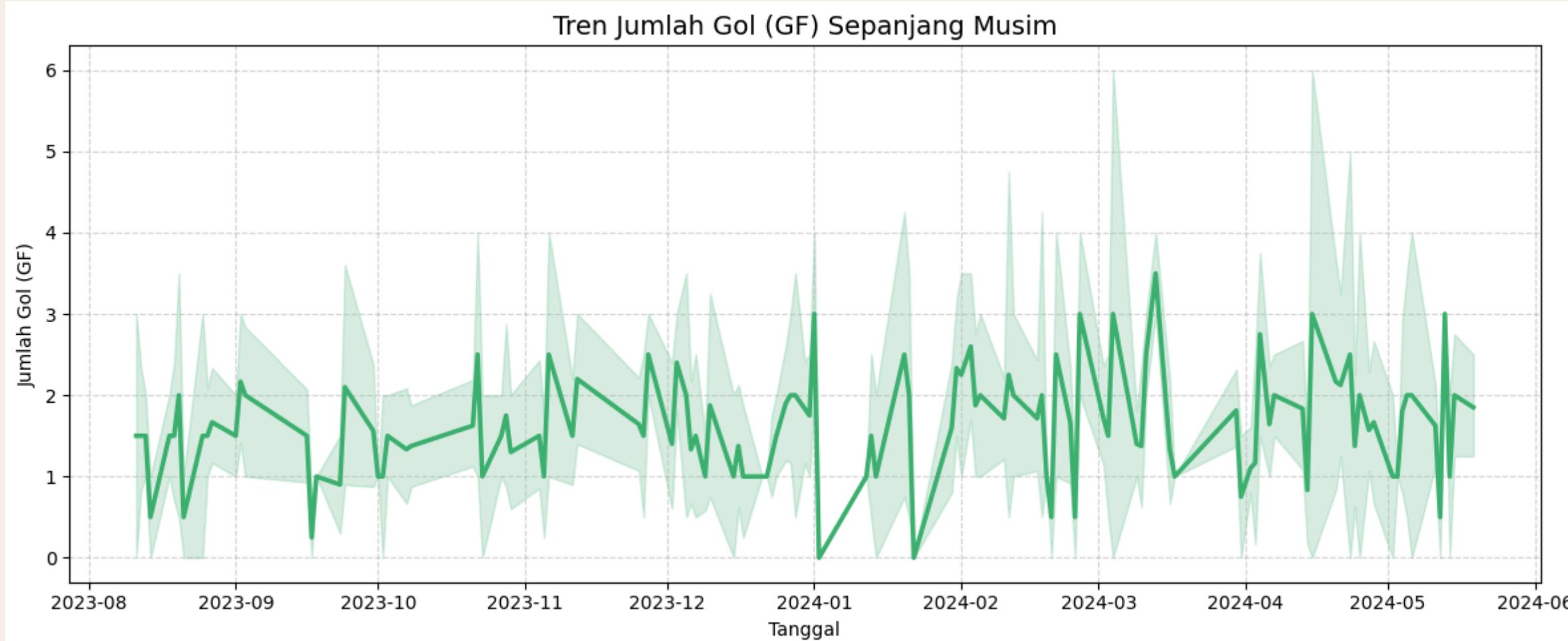
# Rata-rata Possession Berdasarkan Venue



Grafik ini menunjukkan rata-rata penguasaan bola (possession) berdasarkan venue (kandang vs tandang). Sumbu X menampilkan persentase penguasaan bola, sementara sumbu Y menunjukkan kategori venue. Grafik ini membantu menganalisis apakah tim lebih dominan dalam penguasaan bola saat bermain di kandang atau tandang.



# Tren Jumlah Gol (GF) Sepanjang Musim



Hasil dari visualisasi line chart ini menunjukkan tren jumlah gol (GF) sepanjang musim berdasarkan tanggal pertandingan. Dengan garis berwarna hijau yang mewakili jumlah gol, grafik ini memungkinkan kita untuk melihat bagaimana performa tim dalam mencetak gol berfluktuasi sepanjang waktu. Beberapa poin yang bisa diperhatikan dari grafik ini:

- **Puncak-puncak tinggi:** Menunjukkan pertandingan di mana tim mencetak banyak gol dalam satu pertandingan.
- **Penurunan:** Ketika jumlah gol menurun pada tanggal tertentu, yang dapat menunjukkan pertandingan dengan hasil kurang memuaskan.
- **Fluktuasi musiman:** Tren umum selama musim, apakah tim lebih konsisten mencetak gol di waktu tertentu atau ada peningkatan/perubahan yang signifikan di tengah musim.

# Kesimpulan

Dalam analisis eksplorasi data (EDA) ini, telah memeriksa dan membersihkan dataset sepak bola yang berisi informasi tentang pertandingan, gol, dan statistik lainnya. Data yang telah dianalisis mencakup penghitungan statistik deskriptif, pemrosesan tipe data, serta identifikasi missing values dan duplikat. Visualisasi seperti bar chart dan line chart memberikan wawasan tentang tren penguasaan bola dan jumlah gol sepanjang musim. EDA sangat penting untuk memahami struktur dan kualitas data sebelum melakukan analisis lebih lanjut, serta untuk mengidentifikasi pola dan anomali yang dapat memengaruhi hasil analisis. Semoga hasil analisis ini dapat membantu dalam memahami pola pertandingan, dan mohon maaf jika ada kekurangan dalam prosesnya.

---

# Thank You



---