FIFA 2022 World Cup Data Analysis

Author: Dessailly Kikuku
Goal: Turn raw World Cup 2022 match stats into clear insights for readers who may be new to **Python** and **soccer analytics**.

What you'll learn

- How to read and explore tabular data with **pandas**
- Soccer basics: goals, attempts, possession, cards, etc.
- How to reshape matches into **team-level** records for analysis
- How to build clean, interpretable charts and write takeaways
- How to export a shareable HTML/PDF report

> This notebook blends **education** (what each step means) with a **professional narrative** (why it matters and what we learned).

Setup: Libraries & Style

We load Python libraries:

- **pandas** for tables and grouping
- **numpy** for simple numeric helpers
- **matplotlib / seaborn** for charts

We also set global style options so every chart looks consistent and readable. If you see "FutureWarning" messages (version hints), we can hide them for a cleaner, report-ready notebook.

> **Why hide warnings?**
> Warnings are helpful for developers, but they distract non-technical readers.
> For a teaching/report version, we suppress them globally so charts stand out.

Load the Dataset

We load the CSV file that has **64 World Cup 2022 matches** with detailed stats. Each row = one match (Team 1 vs Team 2) with columns such as:

- `number of goals team1` / `team2` — goals scored
- `possession team1` / `team2` — % of ball control
- `total attempts ...` and `on target attempts ...` — shots and shots on goal
- passes completed, cards (yellow/red), corners, fouls, etc.

After loading, we preview a few rows to confirm columns and values look correct.

`.info()` shows data types and missing values.
`.describe()` gives quick stats like averages and ranges.
This helps us see if the data is clean.

Understand the Data: Columns & Types

Before analysis, we answer:

- What columns exist and what do they mean?
- Are columns numeric or text?
- Are there missing values?

This step helps us avoid surprises and choose the right operations for each column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 88 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   team1                                                 64 non-null     object
 1   team2                                                 64 non-null     object
 2   possession team1                                      64 non-null     object
 3   possession team2                                      64 non-null     object
 4   possession in contest                                 64 non-null     object
 5   number of goals team1                                 64 non-null     int64
 6   number of goals team2                                 64 non-null     int64
 7   date                                                  64 non-null     object
 8   hour                                                  64 non-null     object
 9   category                                              64 non-null     object
 10  total attempts team1                                  64 non-null     int64
 11  total attempts team2                                  64 non-null     int64
 12  conceded team1                                        64 non-null     int64
 13  conceded team2                                        64 non-null     int64
 14  goal inside the penalty area team1                    64 non-null     int64
 15  goal inside the penalty area team2                    64 non-null     int64
 16  goal outside the penalty area team1                   64 non-null     int64
 17  goal outside the penalty area team2                   64 non-null     int64
 18  assists team1                                         64 non-null     int64
 19  assists team2                                         64 non-null     int64
 20  on target attempts team1                              64 non-null     int64
 21  on target attempts team2                              64 non-null     int64
 22  off target attempts team1                             64 non-null     int64
 23  off target attempts team2                             64 non-null     int64
 24  attempts inside the penalty area team1                64 non-null     int64
 25  attempts inside the penalty area  team2               64 non-null     int64
 26  attempts outside the penalty area  team1              64 non-null     int64
 27  attempts outside the penalty area  team2              64 non-null     int64
 28  left channel team1                                    64 non-null     int64
 29  left channel team2                                    64 non-null     int64
 30  left inside channel team1                             64 non-null     int64
 31  left inside channel team2                             64 non-null     int64
 32  central channel team1                                 64 non-null     int64
 33  central channel team2                                 64 non-null     int64
 34  right inside channel team1                            64 non-null     int64
 35  right inside channel team2                            64 non-null     int64
 36  right channel team1                                   64 non-null     int64
 37  right channel team2                                   64 non-null     int64
 38  total offers to receive team1                         64 non-null     int64
 39  total offers to receive team2                         64 non-null     int64
 40  inbehind offers to receive team1                      64 non-null     int64
 41  inbehind offers to receive team2                      64 non-null     int64
 42  inbetween offers to receive team1                     64 non-null     int64
 43  inbetween offers to receive team2                     64 non-null     int64
 44  infront offers to receive team1                       64 non-null     int64
 45  infront offers to receive team2                       64 non-null     int64
 46  receptions between midfield and defensive lines team1  64 non-null     int64
 47  receptions between midfield and defensive lines team2  64 non-null     int64
 48  attempted line breaks team1                           64 non-null     int64
 49  attempted line breaks team2                           64 non-null     int64
```

```
 50  completed line breaksteam1                    64 non-null     int64
 51  completed line breaks team2                   64 non-null     int64
 52  attempted defensive line breaks team1         64 non-null     int64
 53  attempted defensive line breaks team2         64 non-null     int64
 54  completed defensive line breaksteam1          64 non-null     int64
 55  completed defensive line breaks team2         64 non-null     int64
 56  yellow cards team1                            64 non-null     int64
 57  yellow cards team2                            64 non-null     int64
 58  red cards team1                               64 non-null     int64
 59  red cards team2                               64 non-null     int64
 60  fouls against team1                           64 non-null     int64
 61  fouls against team2                           64 non-null     int64
 62  offsides team1                                64 non-null     int64
 63  offsides team2                                64 non-null     int64
 64  passes team1                                  64 non-null     int64
 65  passes team2                                  64 non-null     int64
 66  passes completed team1                        64 non-null     int64
 67  passes completed team2                        64 non-null     int64
 68  crosses team1                                 64 non-null     int64
 69  crosses team2                                 64 non-null     int64
 70  crosses completed team1                       64 non-null     int64
 71  crosses completed team2                       64 non-null     int64
 72  switches of play completed team1              64 non-null     int64
 73  switches of play completed team2              64 non-null     int64
 74  corners team1                                 64 non-null     int64
 75  corners team2                                 64 non-null     int64
 76  free kicks team1                              64 non-null     int64
 77  free kicks team2                              64 non-null     int64
 78  penalties scored team1                        64 non-null     int64
 79  penalties scored team2                        64 non-null     int64
 80  goal preventions team1                        64 non-null     int64
 81  goal preventions team2                        64 non-null     int64
 82  own goals team1                               64 non-null     int64
 83  own goals team2                               64 non-null     int64
 84  forced turnovers team1                        64 non-null     int64
 85  forced turnovers team2                        64 non-null     int64
 86  defensive pressures applied team1             64 non-null     int64
 87  defensive pressures applied team2             64 non-null     int64
dtypes: int64(80), object(8)
memory usage: 44.1+ KB
```

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
team1	64	32	ARGENTINA	5	NaN	NaN	NaN	NaN	NaN	NaN	NaN
team2	64	32	MOROCCO	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN
possession team1	64	32	51%	5	NaN	NaN	NaN	NaN	NaN	NaN	NaN
possession team2	64	34	35%	5	NaN	NaN	NaN	NaN	NaN	NaN	NaN
possession in contest	64	11	13%	12	NaN	NaN	NaN	NaN	NaN	NaN	NaN
number of goals team1	64.0	NaN	NaN	NaN	1.578125	1.551289	0.0	0.0	1.0	2.0	7.0
number of goals team2	64.0	NaN	NaN	NaN	1.109375	1.055856	0.0	0.0	1.0	2.0	4.0
date	64	23	22 NOV 2022	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN
hour	64	5	20 : 00	24	NaN	NaN	NaN	NaN	NaN	NaN	NaN
category	64	13	Round of 16	8	NaN	NaN	NaN	NaN	NaN	NaN	NaN
total attempts team1	64.0	NaN	NaN	NaN	11.140625	4.972519	2.0	8.0	10.0	14.0	25.0
total attempts team2	64.0	NaN	NaN	NaN	11.28125	5.807682	0.0	7.75	10.0	14.0	32.0
conceded team1	64.0	NaN	NaN	NaN	1.109375	1.055856	0.0	0.0	1.0	2.0	4.0
conceded team2	64.0	NaN	NaN	NaN	1.578125	1.551289	0.0	0.0	1.0	2.0	7.0
goal inside the penalty area team1	64.0	NaN	NaN	NaN	1.46875	1.563155	0.0	0.0	1.0	2.0	7.0
goal inside the penalty area team2	64.0	NaN	NaN	NaN	0.984375	0.999876	0.0	0.0	1.0	2.0	4.0
goal outside the penalty area team1	64.0	NaN	NaN	NaN	0.09375	0.293785	0.0	0.0	0.0	0.0	1.0
goal outside the	64.0	NaN	NaN	NaN	0.109375	0.314576	0.0	0.0	0.0	0.0	1.0

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
penalty area team2											
assists team1	64.0	NaN	NaN	NaN	1.171875	1.363407	0.0	0.0	1.0	2.0	6.0
assists team2	64.0	NaN	NaN	NaN	0.734375	0.895176	0.0	0.0	1.0	1.0	4.0

	team1	team2	possession team1	possession team2	possession in contest	number of goals team1	number of goals team2	date	hour	cate
0	QATAR	ECUADOR	42%	50%	8%	0	2	20 NOV 2022	17 : 00	Gro
1	ENGLAND	IRAN	72%	19%	9%	6	2	21 NOV 2022	14 : 00	Gro
2	SENEGAL	NETHERLANDS	44%	45%	11%	0	2	21 NOV 2022	17 : 00	Gro
3	UNITED STATES	WALES	51%	39%	10%	1	1	21 NOV 2022	20 : 00	Gro
4	ARGENTINA	SAUDI ARABIA	64%	24%	12%	1	2	22 NOV 2022	11 : 00	Gro

Reshape: From Match Rows to Team Rows

Each match row has two teams. To compare teams fairly, we create **one row per team per match**:

- Rename columns so that the current team's stats are always called `goals_for`, `attempts_team`, etc.
- Build two DataFrames (team1's view and team2's view), then stack them.
- Add a `result` label: **W** (win), **D** (draw), **L** (loss) based on goals.

This transformation lets us group by team later (e.g., total goals in the tournament).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 18 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   date                      64 non-null     object
 1   category                  64 non-null     object
 2   team1                     64 non-null     object
 3   team2                     64 non-null     object
 4   number of goals team1     64 non-null     int64
 5   number of goals team2     64 non-null     int64
 6   possession team1          64 non-null     object
 7   possession team2          64 non-null     object
 8   total attempts team1      64 non-null     int64
 9   total attempts team2      64 non-null     int64
 10  on target attempts team1  64 non-null     int64
 11  on target attempts team2  64 non-null     int64
 12  yellow cards team1        64 non-null     int64
 13  yellow cards team2        64 non-null     int64
 14  red cards team1           64 non-null     int64
 15  red cards team2           64 non-null     int64
 16  passes completed team1    64 non-null     int64
 17  passes completed team2    64 non-null     int64
dtypes: int64(12), object(6)
memory usage: 9.1+ KB
```

	date	category	team1	team2	number of goals team1	number of goals team2	possession team1	possession team2	total attempts team1	at
0	20 NOV 2022	Group A	QATAR	ECUADOR	0	2	42%	50%	5	
1	21 NOV 2022	Group B	ENGLAND	IRAN	6	2	72%	19%	13	
2	21 NOV 2022	Group A	SENEGAL	NETHERLANDS	0	2	44%	45%	14	
3	21 NOV 2022	Group B	UNITED STATES	WALES	1	1	51%	39%	6	
4	22 NOV 2022	Group C	ARGENTINA	SAUDI ARABIA	1	2	64%	24%	14	

	date	category	team	opponent	goals_for	goals_against	possession_team	possession_o
0	20 NOV 2022	Group A	QATAR	ECUADOR	0	2	42%	5
1	21 NOV 2022	Group B	ENGLAND	IRAN	6	2	72%	1
2	21 NOV 2022	Group A	SENEGAL	NETHERLANDS	0	2	44%	4
3	21 NOV 2022	Group B	UNITED STATES	WALES	1	1	51%	3
4	22 NOV 2022	Group C	ARGENTINA	SAUDI ARABIA	1	2	64%	2

Team Summary: Wins, Goals, Points

Now we condense each team's performance across the tournament:

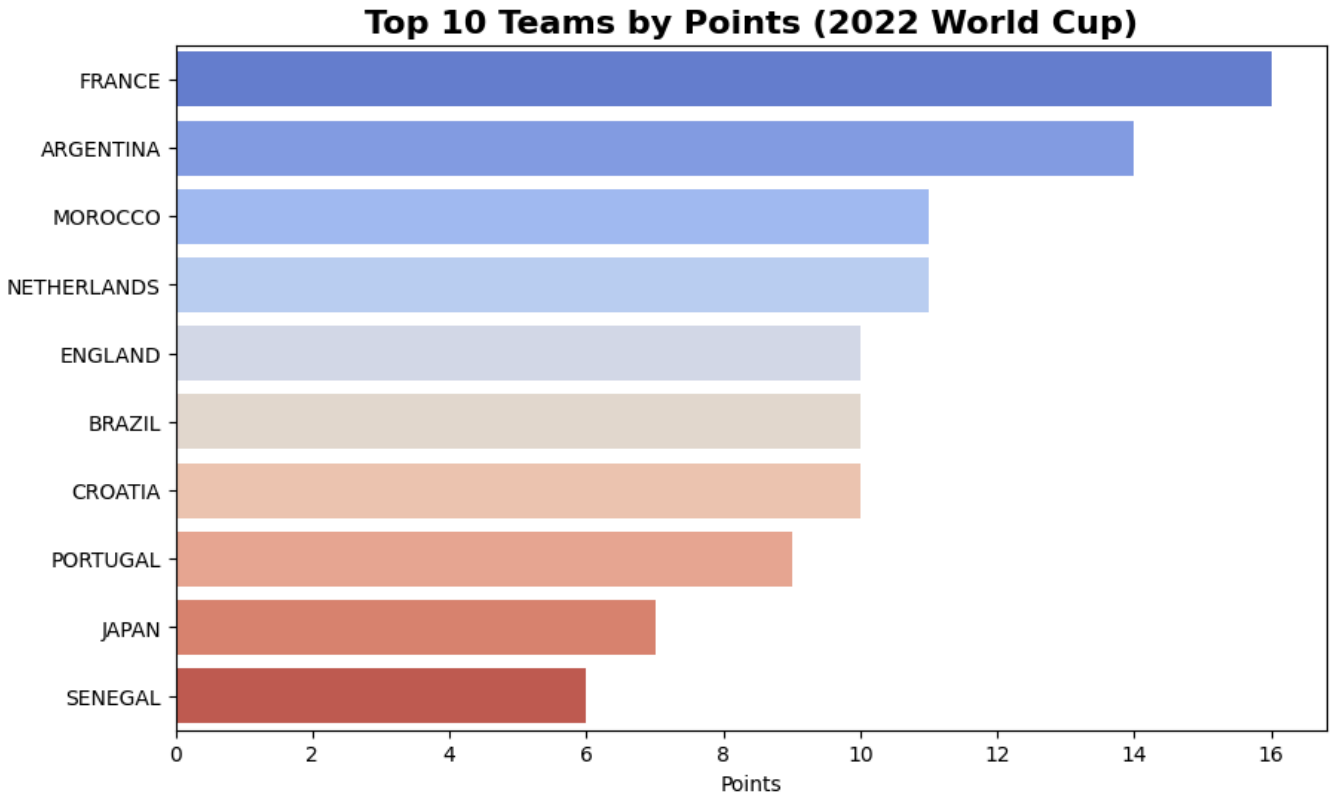
- **Games** played, **Wins/Draws/Losses**
- **Goals For/Against** and **Goal Difference**
- **Points** (3 per win, 1 per draw)
- Average **Attempts** and **Passes** per match
- A possession value we'll convert to % later

This acts like a mini "league table" for the whole World Cup.

	team	games	wins	draws	losses	goals_for	goals_against	attempts	passes	posse
11	FRANCE	7	5	1	1	16	8	14.428571	456.000000	
0	ARGENTINA	7	4	2	1	15	8	14.857143	548.714286	
19	NETHERLANDS	5	3	2	0	10	4	8.000000	488.400000	
18	MOROCCO	7	3	2	2	6	5	8.714286	317.000000	
10	ENGLAND	5	3	1	1	13	4	12.000000	544.600000	
3	BRAZIL	5	3	1	1	8	3	18.000000	539.200000	
7	CROATIA	7	2	4	1	8	7	11.571429	532.000000	
21	PORTUGAL	5	3	0	2	12	6	12.400000	523.000000	
15	JAPAN	4	2	1	1	5	4	10.500000	324.500000	
24	SENEGAL	4	2	0	2	5	7	12.750000	319.000000	

Who Performed Best? — Top 10 Teams by Points

Why this matters: Points summarize consistency in soccer (result-driven performance).
How to read: Bars to the right = more tournament points.

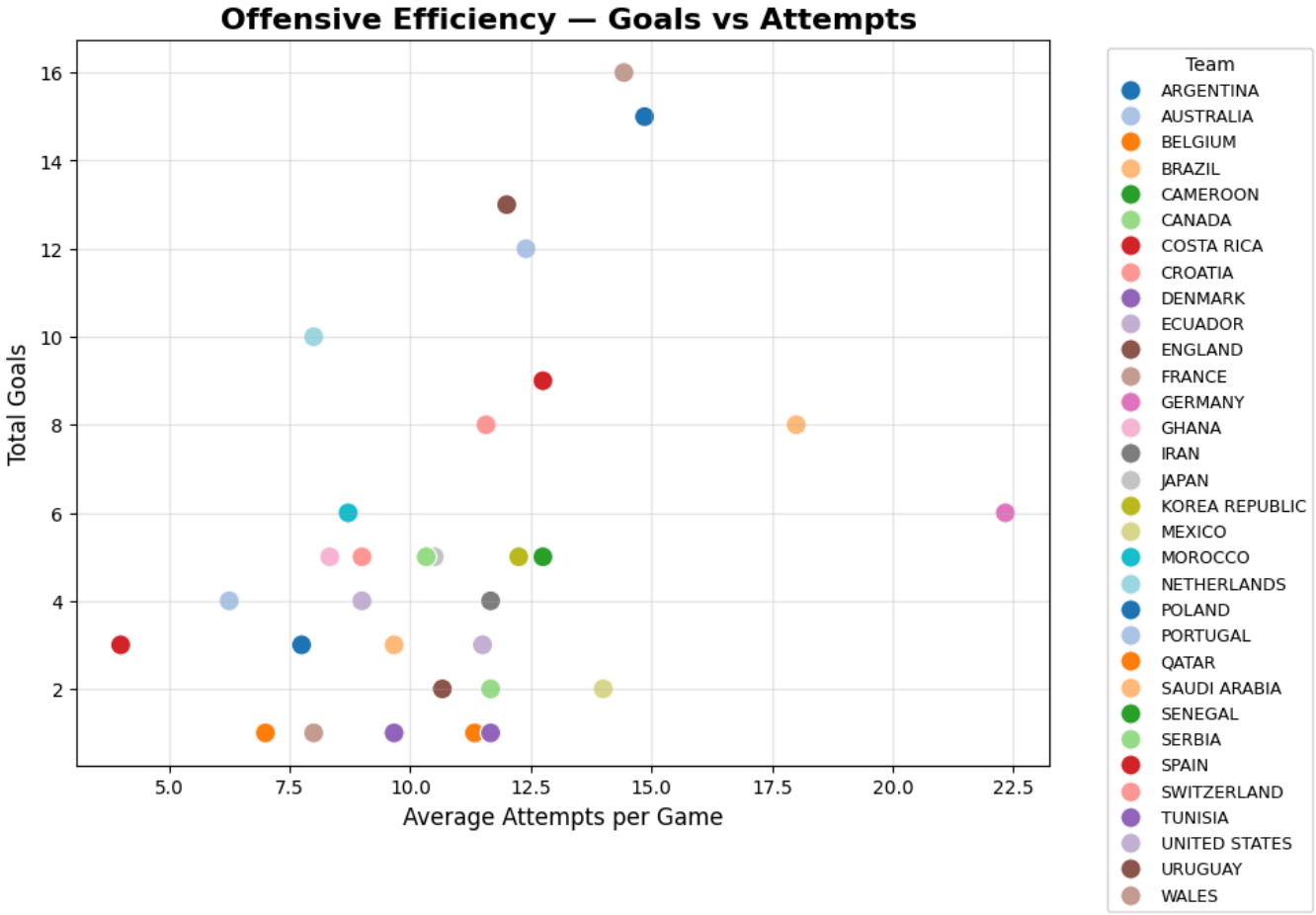


What we see:

- The leaders reflect strong group stages and knockouts.
- Look for surprises (e.g., a team with fewer total goals still ranking high due to tight wins).

Can You Turn Chances into Goals? — Offensive Efficiency

Shots (attempts) are opportunities; goals are the outcome.
A team high on goals with **fewer** attempts = **efficient finishing**;
a team high on attempts but low on goals may struggle with conversion.



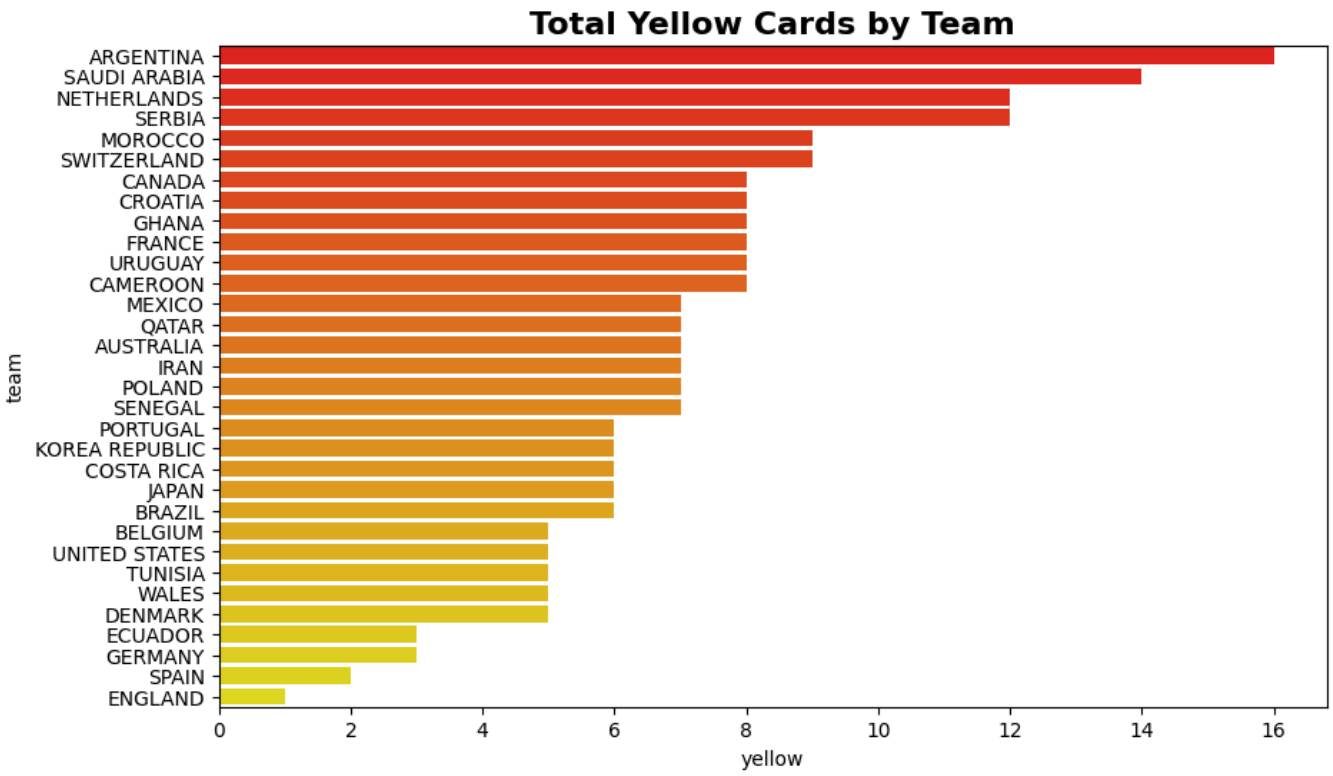
Takeaway:
Which teams sit **above** the general trend? Those are converting chances well.
Teams far right but not very high may need better shot quality/finishing.

Discipline — Yellow and Red Cards

Cards indicate physicality and risk:

- **Yellow** = caution for unsporting behavior
- **Red** = player sent off (team plays with fewer players)

Higher card totals can reflect a physical style or matches under pressure.



Interpretation:

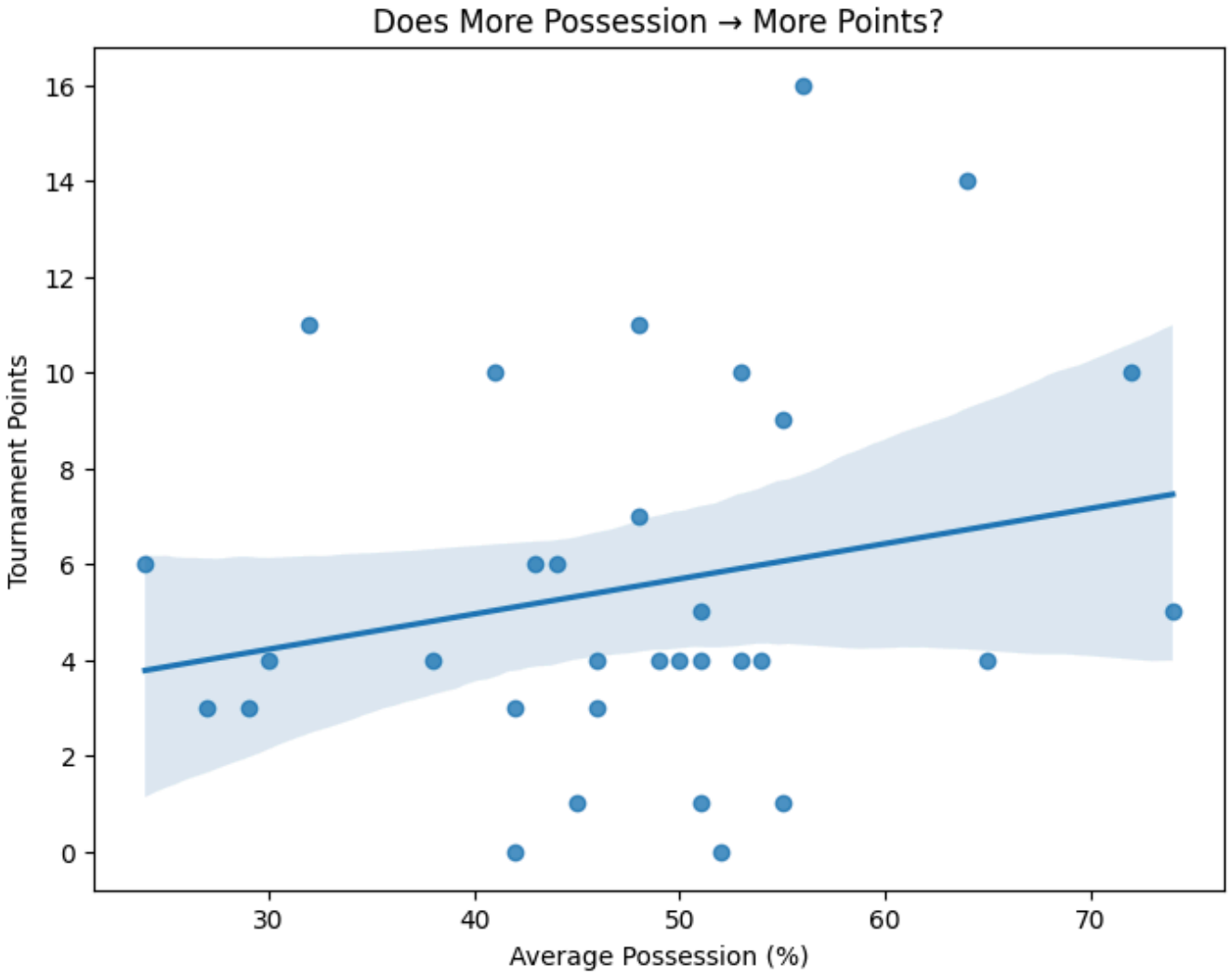
Compare high-card teams to their results. Did discipline issues correlate with fewer points or key match incidents?

Does Possession Translate to Success?

Possession measures time on the ball.

Some teams dominate the ball; others prefer **compact defense + counter-attacks**.

We compare **average possession** with **total tournament points**.



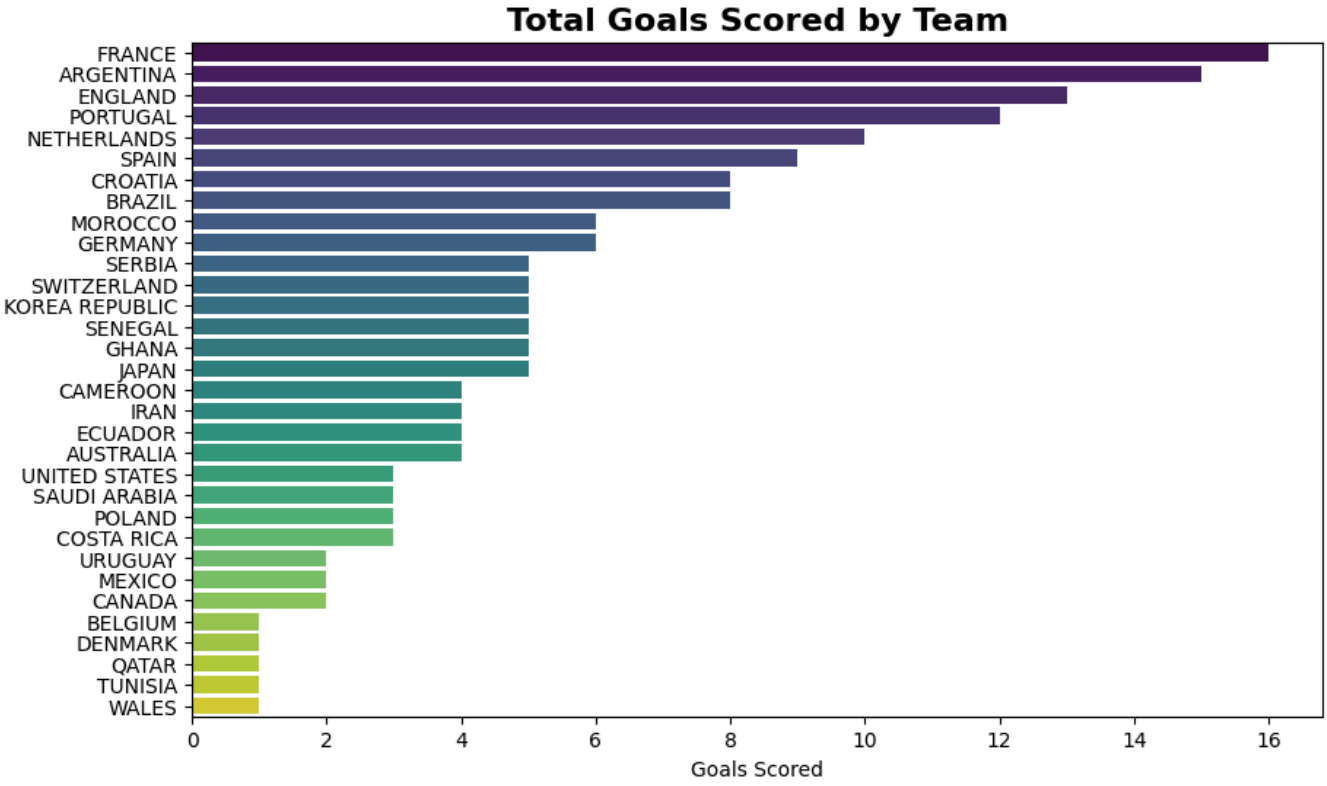
Insight:

Possession alone isn't destiny. Efficient transitions and finishing matter.

Teams can earn results with less possession if they create better chances.

Which teams scored the most goals?

This reveals which countries contributed most to scoring. But does scoring more goals means success?



What This Chart Tells Us

- Teams at the top of the chart—such as **France**, **England**, and **Argentina**—scored the highest number of goals.
  This often reflects strong attacking lines, efficient finishing, and deep tournament runs.
- Mid-table teams may have balanced scoring, showing tactical approaches that combine offense and structure.
- Nations near the bottom typically struggled to create or finish chances, or faced stronger opponents early in the tournament.

Extra Insight

A team's goal total doesn't tell the full story (e.g., defensive teams may advance with fewer goals), but it **does** reveal which sides posed the biggest attacking threats throughout the competition.

If paired with "Goals Conceded" or "Goal Difference," you get an even clearer picture of overall strength.

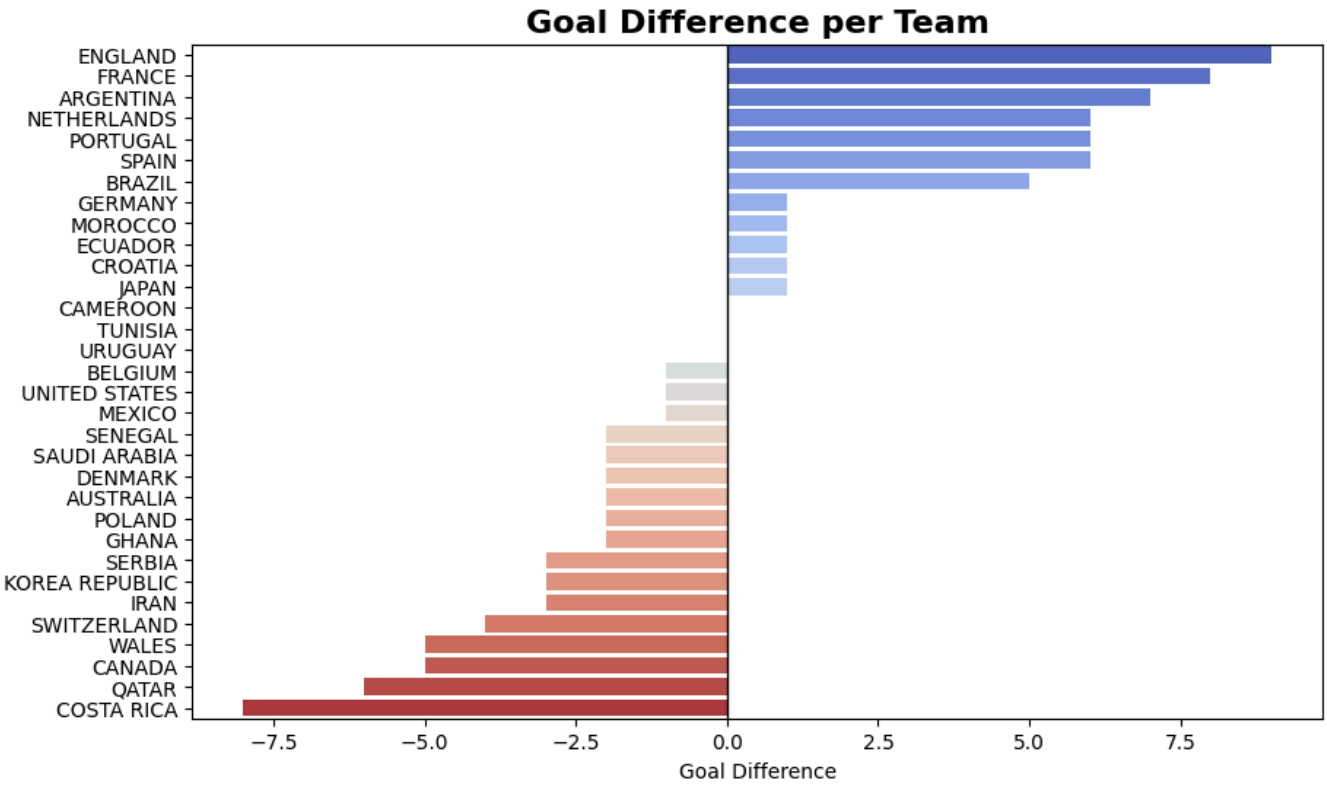
Goal Difference — Measuring Net Team Performance

Goal Difference (GD) = *Goals Scored − Goals Conceded*

It's one of the simplest but most powerful soccer indicators.
A **positive GD** means the team scored more than they conceded — a sign of both strong attack and

solid defense.
A **negative GD** suggests defensive weaknesses or struggles in finishing chances.

This chart compares the *net performance* of each team to show who dominated and who struggled during the tournament.



What This Shows

- Teams **above the center line (positive GD)** consistently outscored opponents — they controlled matches both offensively and defensively.
- Teams **below zero (negative GD)** likely struggled to convert chances or defend against stronger opponents.

In the 2022 World Cup, **France**, **Argentina**, and **England** stood out with high goal differences, reflecting both attacking firepower and defensive structure.
Meanwhile, early exits with negative GD often came from teams that conceded heavily or lost multiple close matches.

> **Interpretation Tip:**
> Goal Difference summarizes efficiency — a few big wins can inflate it, but steady small margins often show true consistency.

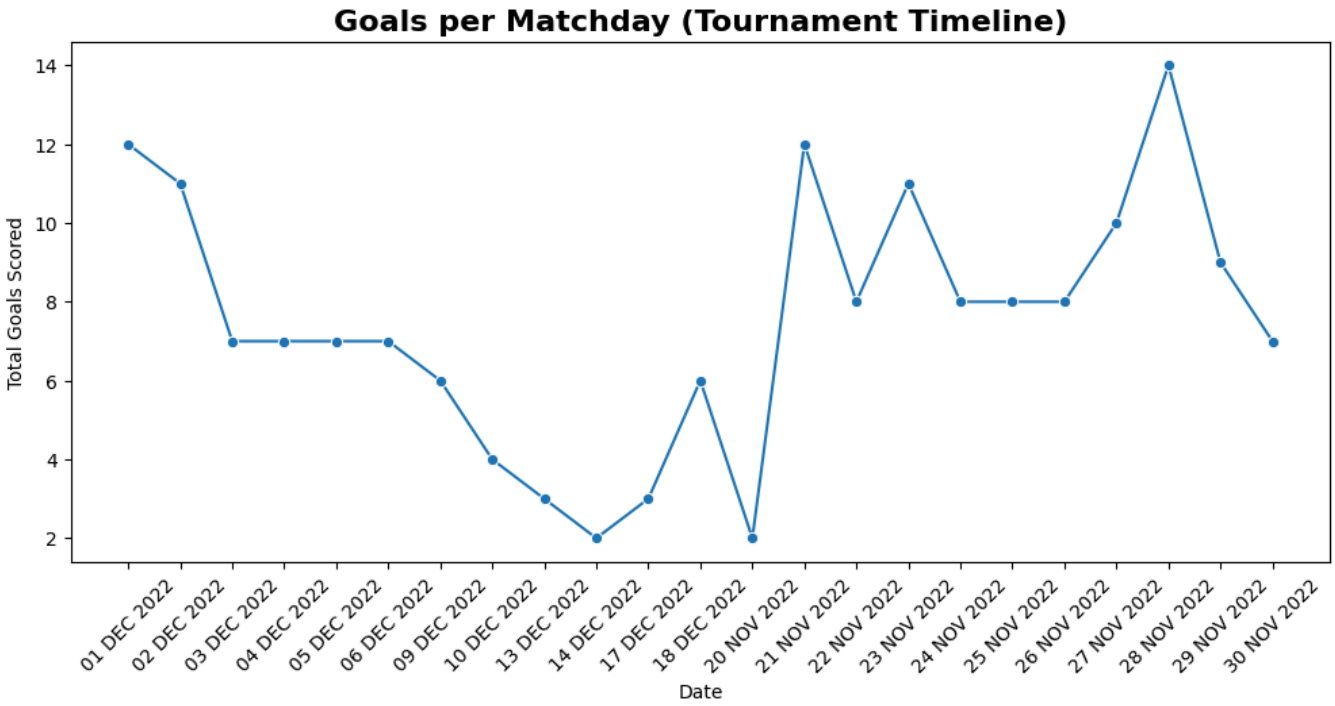
Goals Timeline — Which Matchdays Were the Most Explosive?

Tournaments often have moments where the action peaks — days with many goals, dramatic matches, or decisive group outcomes.
A **timeline visualization** helps us understand:

- How goal scoring changed throughout the competition
- Whether early group-stage matches or knockout rounds had more goals
- Which specific days were the most exciting for fans

This also helps explain shifts in team behavior:
early matches may be cautious, while later stages push teams to attack aggressively.

By plotting total goals per date, we can see the rhythm of the tournament and identify high-impact matchdays.



What This Timeline Reveals

- **Peaks** in the chart represent matchdays with multiple high-scoring games.
  These are usually group-stage days with 3–4 matches played.

- **Dips** in goal totals often occur during the knockout rounds,
  where teams play more cautiously because a single mistake can eliminate them.

- The trend helps us understand the competition's flow:
  early excitement, tactical adjustments, and high-intensity elimination matches.

Why This Matters

Seeing goals over time tells a *story* about how the tournament evolved — from the chaotic, attacking early days to the tightly contested matches near the end.

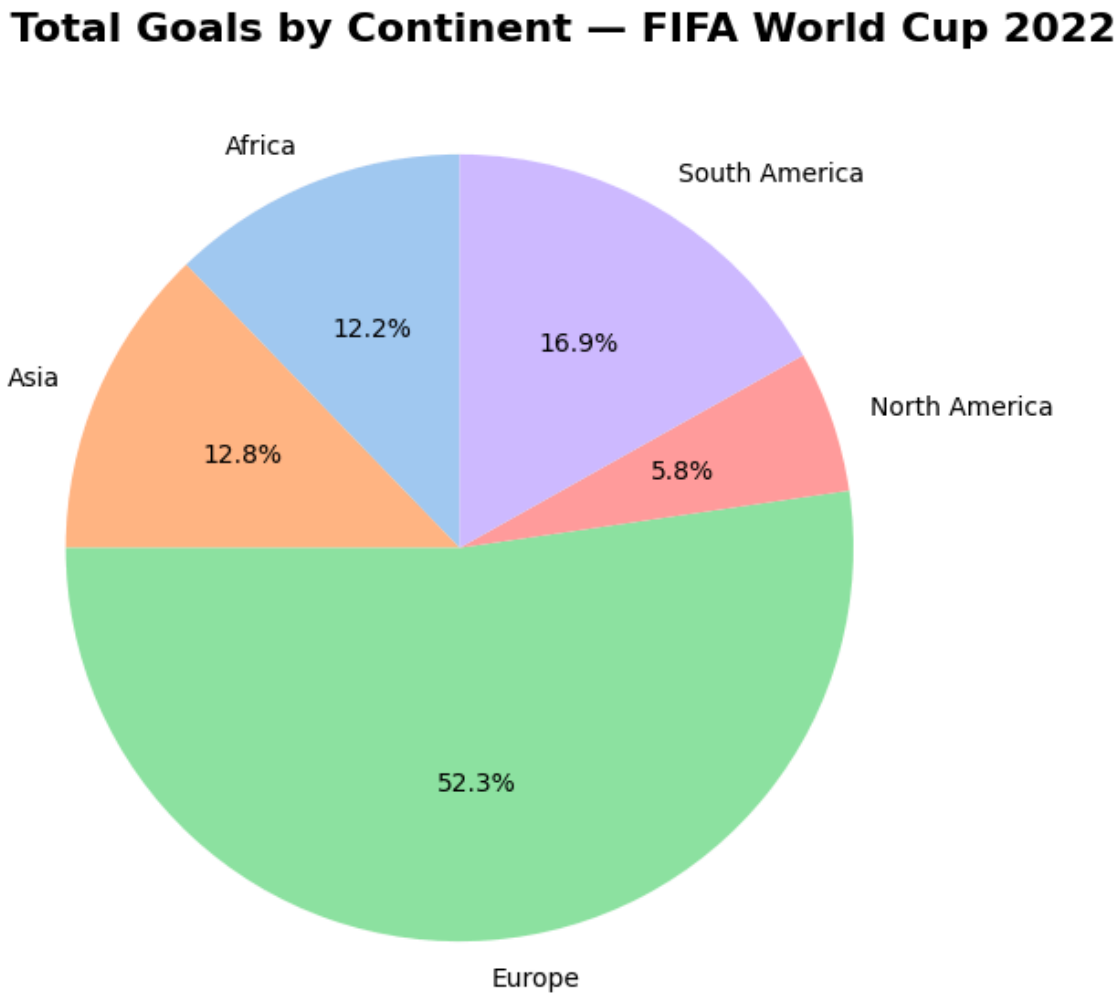
This timeline also helps pinpoint which days were the most memorable for fans and analysts tracking the drama of the World Cup.

```
Teams with missing region info:
Series([], Name: team, dtype: object)
```

	region	goals_for
0	Africa	21
1	Asia	22
2	Europe	90
3	North America	10
4	South America	29

Where Did the Goals Come From? — Continents

We assign each nation to a **continent/region** and sum total goals per region.
This reveals which parts of the world contributed most to scoring.



Reading the pie:
Regions with larger slices scored more goals overall.
This doesn't prove "better teams," but shows scoring distribution by confederation.

Summary & Next Steps

You learned how to:

- load and clean soccer data
- summarize wins, goals, and possession
- visualize team and continent performance

Next Challenges

- Add player-level data (goals, assists)
- Predict winners using machine learning
- Compare 2022 vs 2018 World Cup

```
[NbConvertApp] Converting notebook /content/drive/MyDrive/Colab Notebooks/Fifa_2022_Analysis.
ipynb to html
[NbConvertApp] WARNING | Alternative text is missing on 8 image(s).
[NbConvertApp] Writing 899598 bytes to /content/Fifa_2022_Report.html
```