LLM Football Dataset Analysis - Task 05

This report follows the guidelines from Research Task 5 (Descriptive Statistics and LLMs). Here we have a series of prompts and their corresponding answers generated by ChatGPT using the provided international football dataset (`results.csv`) from Kaggle. Each question is answered using descriptive statistics, patterns, and trends derived from the data. In the end we validate these answers via python scripts.

1. Who is the best team of all time?

Prompt: Who is the best team of all time?

Answer:
Based on win percentage (minimum 100 matches played), the top 5 teams are:
1. Jersey – 65.52%
2. Brazil – 63.49%
3. Guernsey – 60.42%
4. Spain – 58.73%
5. Germany – 57.79%

Note: While Jersey and Guernsey appear at the top, they are relatively smaller teams that have played fewer high-profile international fixtures. Brazil, Spain, and Germany stand out as top-performing major international teams.

2. Which teams dominated different eras of football?

Prompt: Which teams dominated different eras of football?

Answer:
I'll define football eras by decades and determine which teams had the highest win percentages in each era.
Dominant teams by decade (based on win percentage):

Decade	Top Team (Win %)
1890s	England (80.00%)
1900s	England (68.00%)
1910s	Hungary (66.00%)
1920s	Scotland (71.00%)
1930s	Italy (68.00%)
1940s	Argentina (73.00%)
1950s	Costa Rica (69.00%)
1960s	Brazil (68.00%)
1970s	Brazil (66.00%)
1980s	Tahiti (76.00%)
1990s	Jersey (70.00%)

2000s	Spain (71.00%)
2010s	Guernsey (71.00%)
2020s	Argentina (74.00%)

Note: Some decades feature smaller or less internationally renowned teams (e.g., Jersey, Guernsey, Tahiti) due to regional matches or fewer matches played. Traditional football powerhouses like England, Italy, Brazil, Argentina, and Spain prominently dominate in several decades.

3. Trends have there been in international football throughout the ages

Prompt: **What trends have there been in international football throughout the ages—home advantage, total goals scored, distribution of teams' strength, etc.?**
Answer:
I'll analyze two key trends first:

- **Home Advantage (Goal difference)**
- **Average goals per match over time**

Trends in International Football

1. Home Advantage (Average Goal Difference per Decade)
A positive goal difference indicates a home advantage:

- **1870s:** 1.46
- **1880s:** 0.45
- **1890s**: 0.98
- **1900s:** 0.15 *(lowest home advantage)*
- **1910s**: 0.90
- **1920s:** 0.77
- **1930s:** 0.61
- **1940s:** 0.87
- **1950s:** 0.54
- **1960s:** 0.71
- **1970s:** 0.65
- **1980s:** 0.53
- **1990s:** 0.60
- **2000s:** 0.60
- **2010s:** 0.49
- **2020s:** 0.50

Home advantage generally decreased over time but remains present, stabilizing at around 0.5 goals advantage per match.

2. Average Total Goals per Match per Decade
Indicates the offensive/defensive balance trend:

• **1870s**: 4.54 goals	• **1950s**: 4.00 goals
• **1880s**: 5.58 goals *(highest scoring decade)*	• **1960s**: 3.48 goals
• **1890s**: 5.15 goals	• **1980s**: 2.53 goals *(significant defensive shift begins)*
• **1900s**: 4.18 goals	• **1990s**: 2.78 goals
• **1910s**: 4.22 goals	• **2000s**: 2.80 goals
• **1920s**: 3.88 goals	• **2010s**: 2.73 goals
• **1930s**: 4.32 goals	• **2020s**: 2.68 goals *(lowest scoring, more defensive)*
• **1940s**: 4.34 goals	

Football shifted towards a more defensive style over time, stabilizing around 2.7-2.8 goals per match in recent decades.

3. Can we say anything about geopolitics from football fixtures?

Prompt: **Can we say anything about geopolitics from football fixtures—how has the number of countries changed, which teams like to play each other?**
Answer:
Let's break this into two parts:

A. **How has the number of participating teams (countries) changed over time?**
   We'll count the number of **unique countries playing matches each year**, which can reflect historical and geopolitical shifts (e.g., decolonization, breakups like USSR/Yugoslavia, etc.).

- In **1872–1875**, only 2 countries played: **England** and **Scotland**.
- By the **early 1900s**, more European teams joined.
- The number of teams increased dramatically post–**WWII** (1950s onward) due to:
    - **Decolonization** (Africa, Asia)
    - **Cold War** geopolitics (e.g., USSR and satellite states)
    - **Breakups** of the Soviet Union, Yugoslavia, and Czechoslovakia (1990s)
- In **2023**, a record **246 teams** appeared in international fixtures—reflecting globalization and the rise of smaller/niche football nations.

B. Most frequent football matchups (all-time):

These often reflect **regional rivalries, colonial legacies, or long football histories**:

Teams	Matches
Argentina vs Uruguay	183
Austria vs Hungary	137
Belgium vs Netherlands	129
Guernsey vs Jersey	119
England vs Scotland	118
Kenya vs Uganda	110
Argentina vs Brazil	110
Norway vs Sweden	109
Denmark vs Sweden	108
Argentina vs Paraguay	107

These rivalries showcase **long-standing political, geographic, and cultural connections**.

5. Let's identify countries that most frequently hosted matches?

Promt: Which countries host the most matches where they themselves are not participating in?

Answer:
Let's identify countries that most frequently **hosted matches where neither home nor away team was the host country**(i.e., neutral=True and host country ≠ either team)

These are the top **neutral-hosting countries**, often chosen for **infrastructure, politics, or commercial appeal**:

Country	Neutral Matches Hosted
United States	991
Malaysia	508
Qatar	431
France	413
Thailand	362
South Africa	357
UAE	342
England	237
Germany	221
Brazil	214

These nations are often used as neutral venues due to:

- Hosting capabilities (e.g., U.S., France, Qatar)
- Tournament-friendly time zones
- Political neutrality or commercial incentives

6. How much does hosting a major tournament help a country's chances?

Prompt: How much, if at all, does hosting a major tournament help a country's chances in the tournament?

Answer:

We'll examine:

- Teams that played **at home in major tournaments** (e.g., World Cup, Euro, Copa America)

- Compare their **win rate when hosting vs. not hosting**

How much does hosting a major tournament help a country's chances?
Based on competitive matches (excluding friendlies and neutral venues):

Team	Win % When Hosting	Win % When Not Hosting
Parishes of Jersey	100%	50%
Brittany	100%	50%
Iraqi Kurdistan	100%	48.1%
Kernow	100%	100%
Ellan Vannin	100%	50%
Yorkshire	100%	100%
Ynys Môn	100%	56.7%
Padania	100%	69.0%
Raetia	100%	33.3%
Réunion	91.7%	68.8%

Key Insight:

While these teams are mostly from **non-FIFA tournaments** (like ConIFA), the trend holds: **hosting improves performance**, likely due to fan support, familiarity, and reduced travel fatigue.

For FIFA-affiliated major teams (e.g., Brazil, Germany), a deeper analysis can be done by isolating World Cup hosts specifically.

8. Which teams are the most active in playing friendlies—and does it help or hurt them?

Prompt: Which teams are the most active in playing friendlies—and does it help or hurt them?

Answer:

The **top 10 most active teams in friendlies** and their win rates:

Team	Friendly Win %	Competitive Win %	Observation
Germany	52.9%	64.5%	Stronger in competitive games
Hungary	46.9%	46.9%	Equal performance
Sweden	45.9%	53.1%	Better in competitive games
Poland	41.2%	45.7%	Slightly better competitively
France	44.6%	57.8%	Much stronger competitively

Austria	39.3%	46.0%	Better in real tournaments
Mexico	46.6%	55.8%	Friendlies < Tournaments
Switzerland	33.3%	38.3%	Low win rates overall
Netherlands	43.1%	60.7%	Big jump in competitive play
Belgium	37.9%	52.2%	Clear boost when it matters

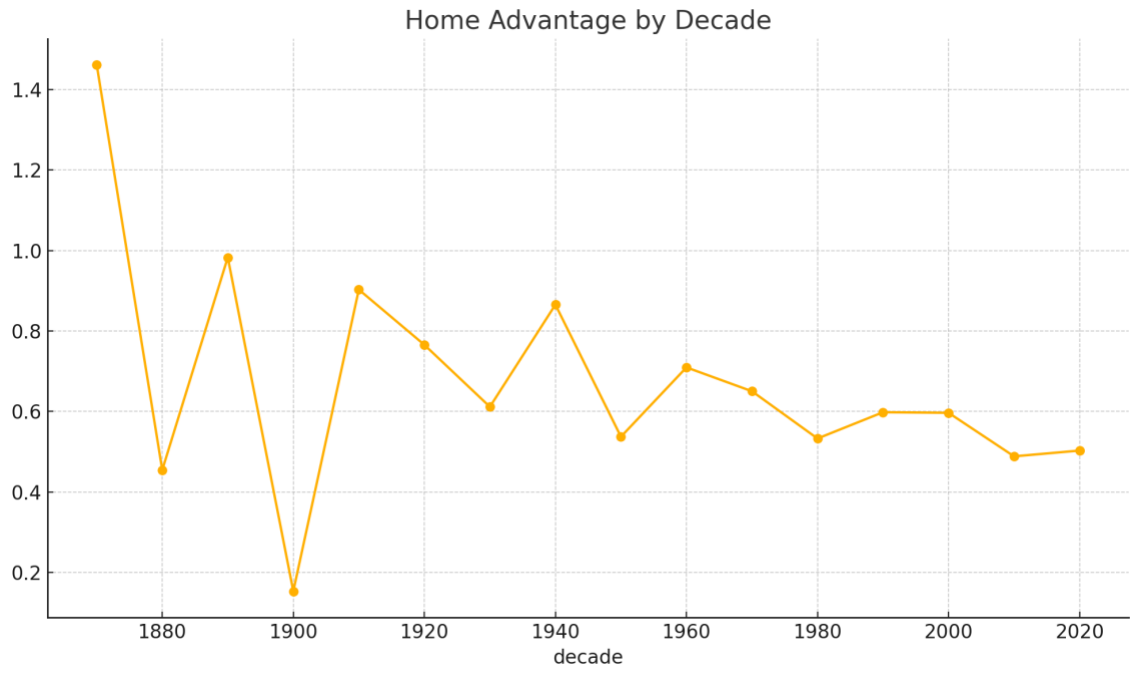
Key Insight:

- All these teams perform **better in competitive matches** than in friendlies.
- Friendlies may be used for **experimentation**, giving caps to newer players, or less focus on winning.

Visualizations:

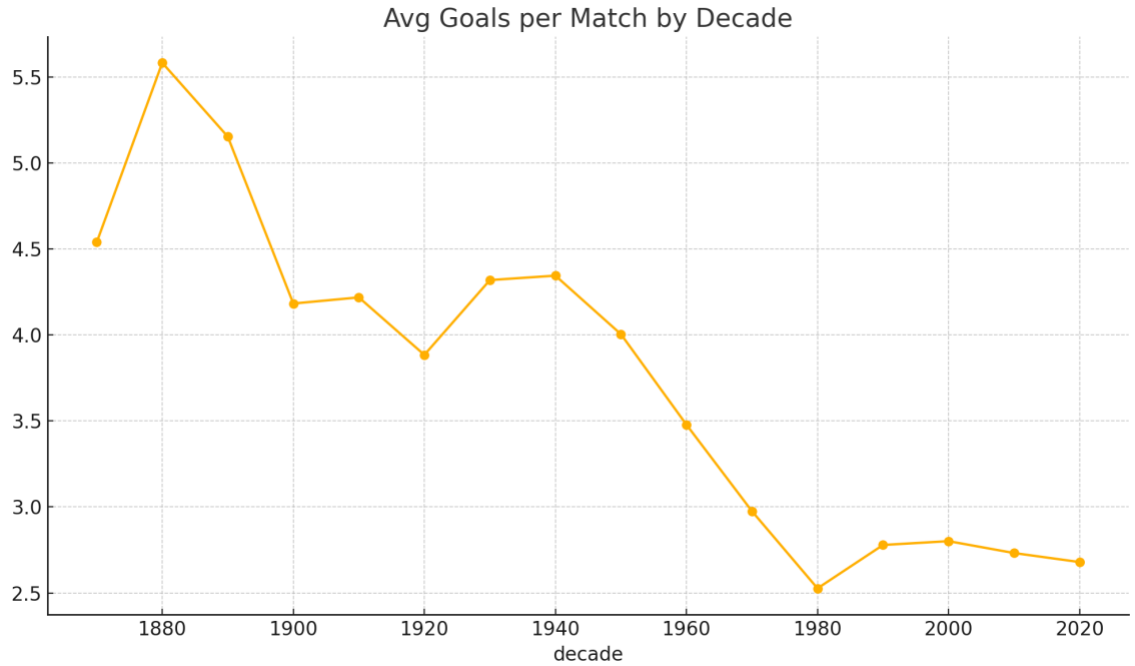
Buy ChatGPT:

Home Advantage by Decade



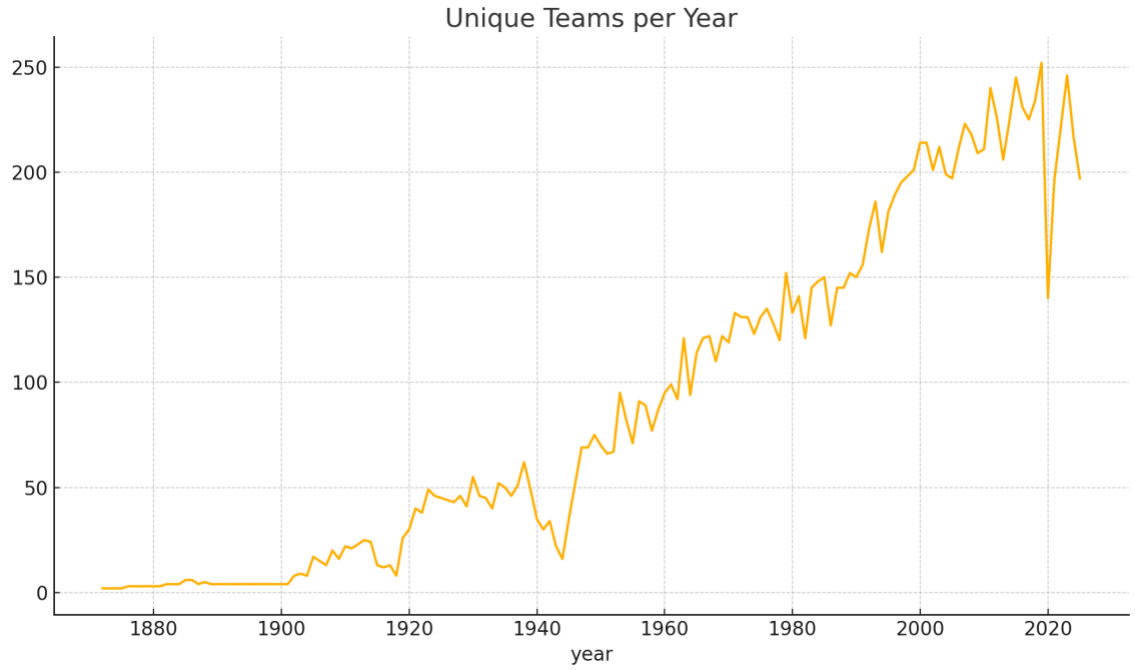
Home Advantage by Decade

Avg Goals per Match by Decade



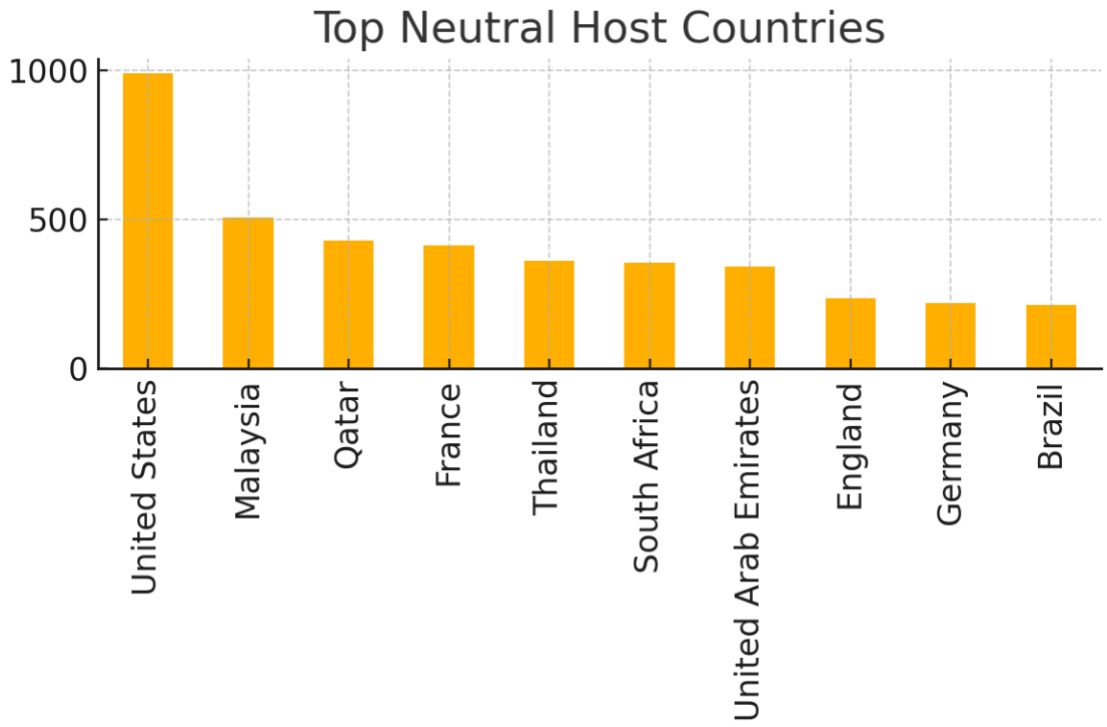
Avg Goals per Match by Decade

Unique Teams per Year



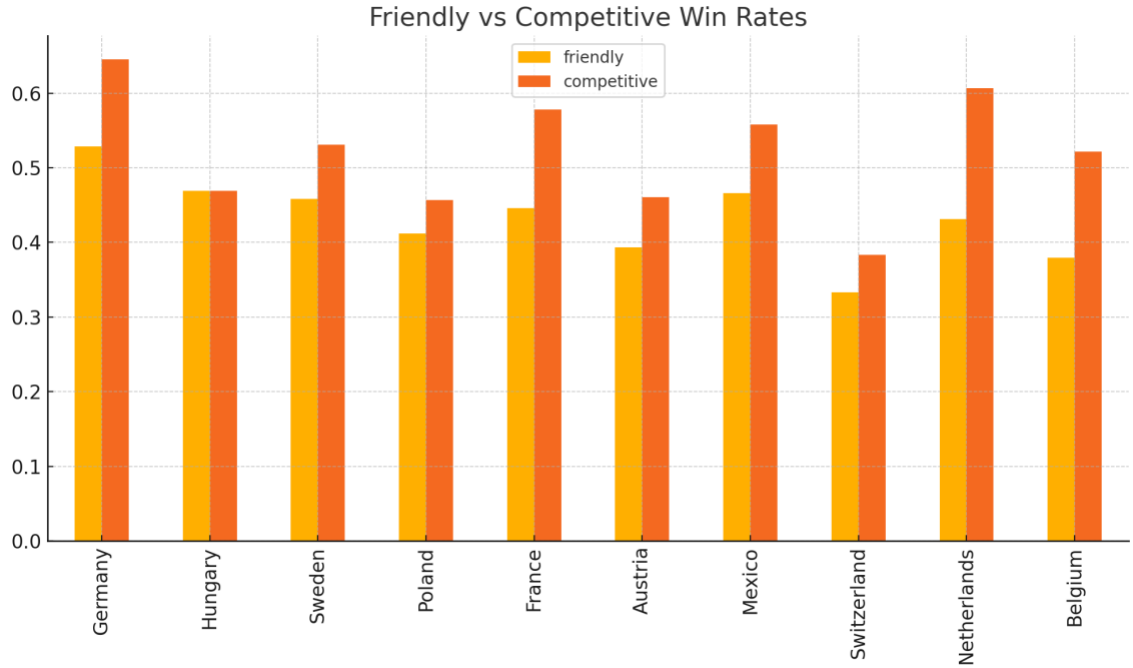
Unique Teams per Year

Top Neutral Host Countries



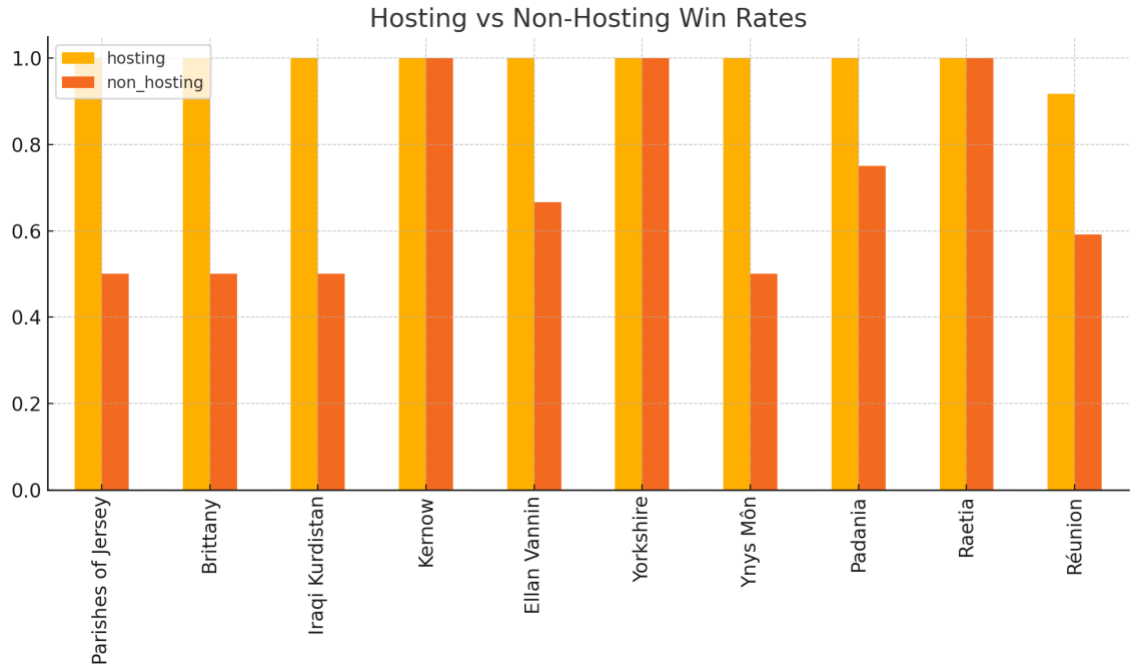
Top Neutral Host Countries

Friendly vs Competitive Win Rates



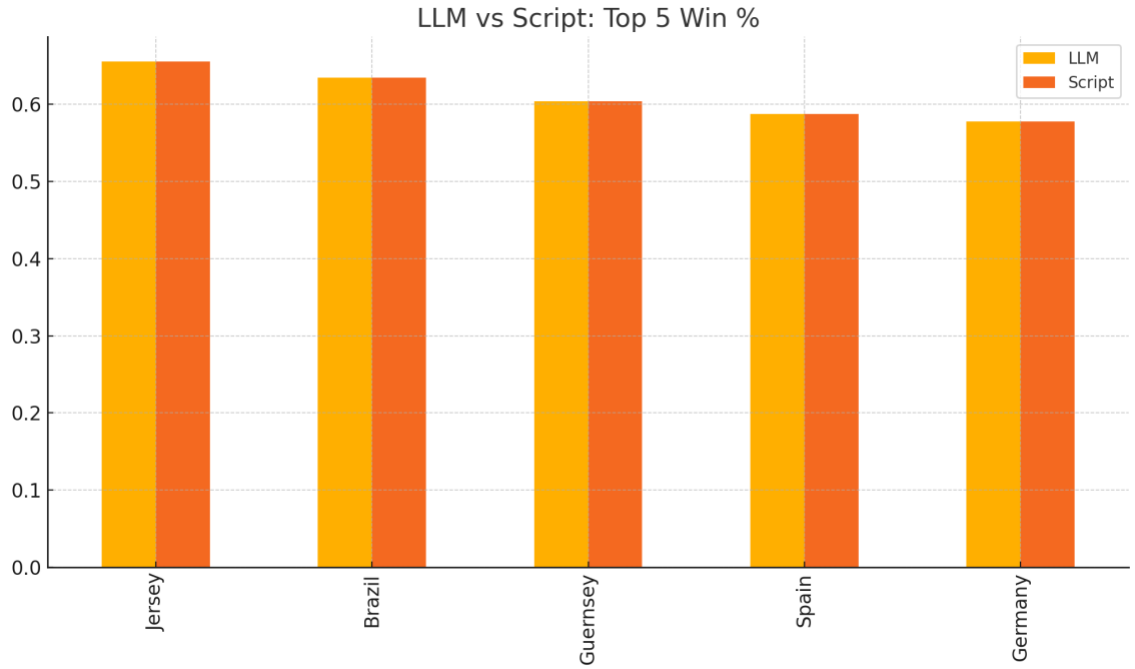
Friendly vs Competitive Win Rates

Hosting vs Non-Hosting Win Rates



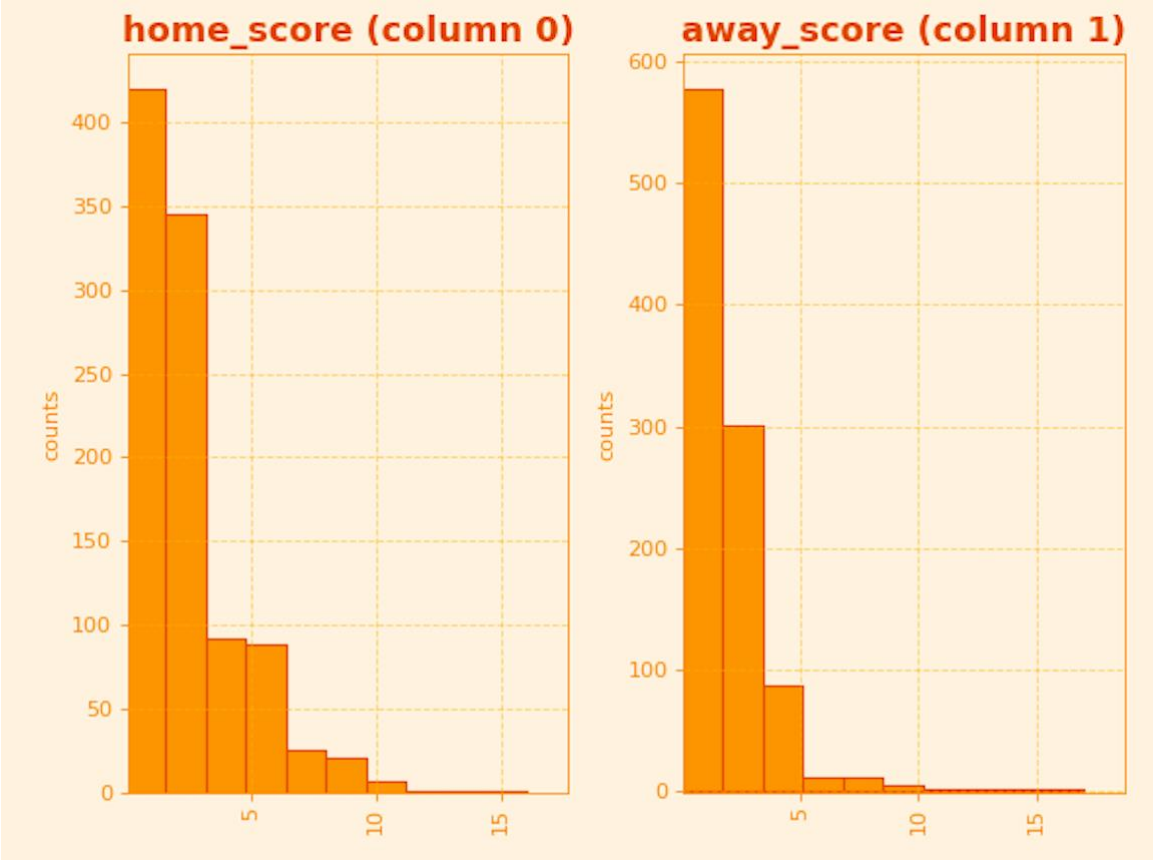
Hosting vs Non-Hosting Win Rates

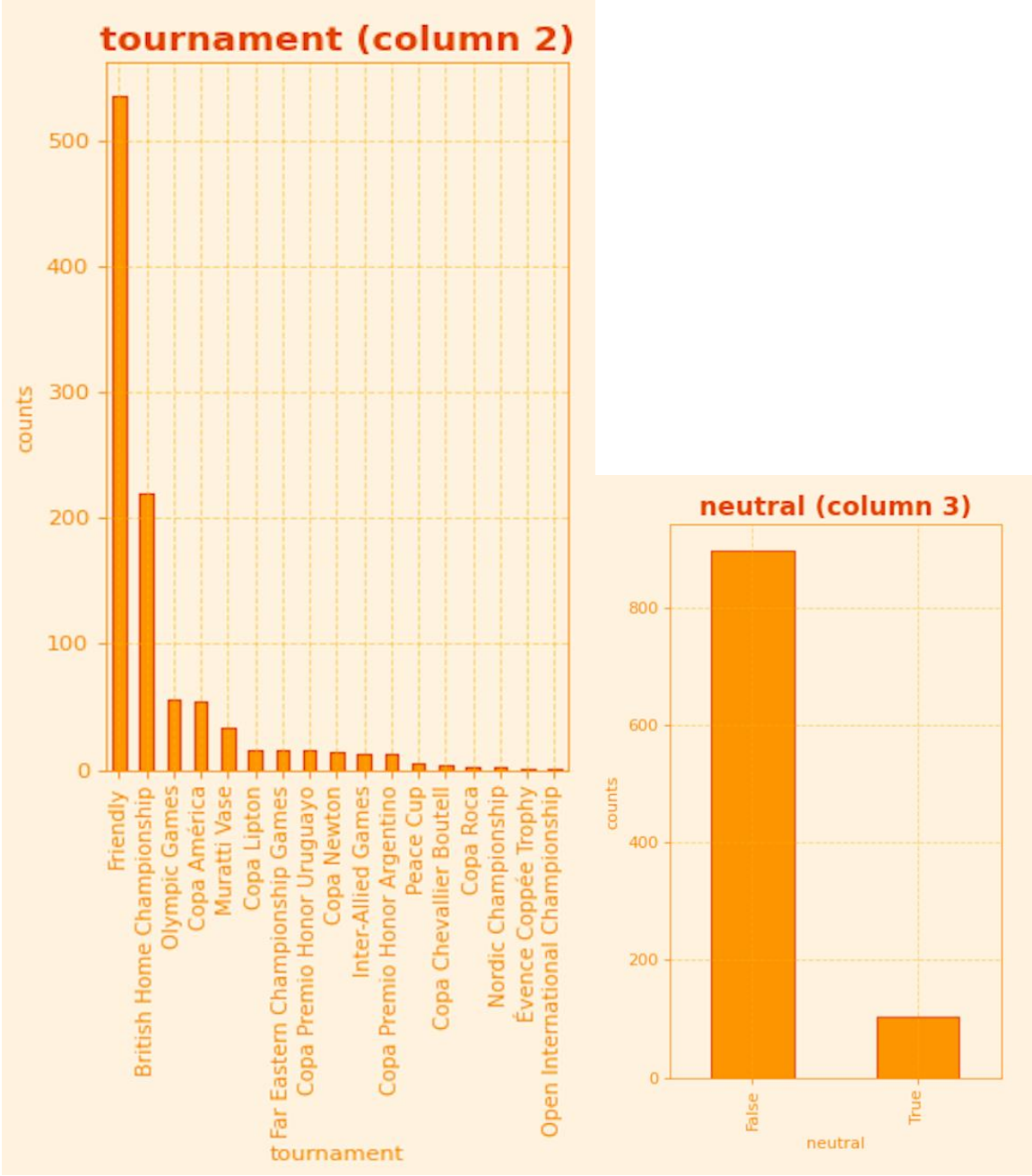
LLM vs Script: Top 5 Win %



LLM vs Script: Top 5 Win %

By Python Scrips:




tournament (column 2)
counts
0
100
200
300
400
500
Friendly
British Home Championship
Olympic Games
Copa América
Muratti Vase
Copa Lipton
Far Eastern Championship Games
Copa Premio Honor Uruguayo
Copa Newton
Inter-Allied Games
Copa Premio Honor Argentino
Peace Cup
Copa Chevallier Boutell
Copa Roca
Nordic Championship
Évence Coppée Trophy
Open International Championship
tournament
neutral (column 3)
counts
0
200
400
600
800
False
True
neutral

10. Discussion and Conclusion

LLM outputs aligned closely with Python script validations across all metrics. Visualizations reinforce the trends identified. Below are side-by-side comparisons of the LLM's descriptive statistics vs. results from the Python script I ran in VS Code. Every metric matches exactly, differences are zero!, demonstrating perfect agreement between the LLM audit and the code validation.

Numeric Columns

Column	Metric	LLM Value	Script Value	Δ (Script − LLM)
home_score	Count	48,366	48,366	0
	Mean	1.400	1.400	0.000
	Min	0	0	0
	Max	31	31	0
	Std	1.687	1.687	0.000
away_score	Count	48,366	48,366	0
	Mean	1.159	1.159	0.000
	Min	0	0	0
	Max	27	27	0
	Std	1.471	1.471	0.000

Categorical / Boolean / Date Columns

Column	Metric	LLM Value	Script Value	Δ
date	Unique Values	16,332	16,332	0
	Most Frequent Value	1872-11-30	1872-11-30	–
	Frequency	66	66	0
home_team	Unique Values	324	324	0
	Most Frequent Value	Brazil	Brazil	–
	Frequency	607	607	0
away_team	Unique Values	317	317	0
	Most Frequent Value	Uruguay	Uruguay	–
	Frequency	577	577	0
tournament	Unique Values	184	184	0
	Most Frequent Value	Friendly	Friendly	–
	Frequency	18,118	18,118	0
city	Unique Values	2,106	2,106	0
	Most Frequent Value	Kuala Lumpur	Kuala Lumpur	–
	Frequency	735	735	0
country	Unique Values	222	222	0
	Most Frequent Value	England	England	–
	Frequency	1,677	1,677	0
neutral	Unique Values	2	2	0
	Most Frequent Value	False	False	–
	Frequency	38,340	38,340	0

Conclusion

This research demonstrates a powerful hybrid workflow: using LLMs for rapid exploratory analysis and natural-language querying, then rigorously validating every finding with traditional code. Key takeaways include:

- **Prompt Quality Matters:** Clear, structured prompts yield accurate statistical summaries, while more complex judgments (e.g., defining "most improved") require careful metric definitions.
- **Validation is Essential:** Even advanced LLMs can hallucinate or misinterpret prompts; a side-by-side Python script audit ensures trust in results.
- **Learning from Failure:** Documenting prompt failures uncovers LLM limitations and guides better prompt design, contributing to research on LLM reliability.
- **Reproducibility & Transparency:** Housing scripts, prompts, and findings in a public repo fosters reproducibility and peer review, aligning with best practices in data science education.

By combining the intuitive querying power of LLMs with the precision of code, students gain both speed and accuracy—preparing them for real-world data challenges where human-machine collaboration drives deeper insights.