



# PROJECT WEEK 10:

## API AND WEB DATA SCRAPING

Ironhack Data Analytics Bootcamp

# INTRODUCTION

- Why?
- Personal football journey
- Modern football was codified in 1863 in England when The Football Association (FA) was formed
- Web Scraping: Wikipedia Premier League Table 2018-2025
  - Wikipedia: 160 rows × 13 columns
- Kaggle API: dataframes 2018-2024
  - 2018-2020: 2280 rows × 20 columns
  - 2020-2023: 2280 rows × 28 columns
  - 2023-2024: 20 rows × 10 columns

# KEY PHRASES

- GA, GF, GD, GA/GF Ratio
- Results: W/D/L
- Top 3 / Non-Top 3
- Formations
- Venue: Home / Away
- Possession (poss)
- SH, SOT
- PTS

# BUSINESS QUESTIONS

- **1.A. Goals For & Against vs League Winners**

- Q: Does a higher GF/GA ratio predict finishing top 3?

- Null Hypothesis ( $H_0$ ): There is no correlation between the teams with the highest GF/GA ratio and finishing top 3.

- **1.B. Formation vs GF/GA ratio vs League Winners**

- Q: Are formations with more in attack (e.g. 4-3-3 or 3-4-3) linked to a higher GF/GA ratio and is there a correlation with winning the league?

- Null Hypothesis ( $H_0$ ): There is no correlation between teams with more attacking formations and higher GF/GA ratio.

- **2.A. Possession vs League Winners**

- Question: Is possession percentage correlated with Premier League points?

- Null Hypothesis ( $H_0$ ): There is no or weak correlation with possession rate and higher points

- **2.B. Possession Home vs Away**

- Question: Is possession percentage higher at home or away games?

- Null Hypothesis ( $H_0$ ): There is no correlation between possession rate and playing games at home.

- **3. Shots & Shots on Target**

- Question: Do shots on target predict winning match outcomes better than total shots?

- Null Hypothesis ( $H_0$ ): Shots on target are not more strongly correlated with points per game than total shots.

# RECOMMENDATIONS

- Aim for higher GF/GD ratio
  - Focus on attacking formations for higher GF/GD ratio
  - Improve on Possession to finish higher in the league and gain more points
  - Replicate Home Success for Away Games
  - Focus on Quality rather than Quantity of Shots
- 
- **Takeaways:**
  - Teams with higher GF/GA ratios consistently earn more league points, scoring more while conceding less is a strong predictor of success.
  - GF/GA ratio higher for Top 3 vs Non-Top 3.
  - Formation: there is a correlation between more attacking formations (4-3-3 or 3-4-3) and higher GF/GA ratio.
  - Higher Average Possession = Higher Points
  - Spearman's: Monotonic relationship. P-value < 0.05, the relationship is statistically significant.
  - Teams usually have similar possession at home and away, but on average, they keep the ball a little more when playing at home.
  - Shots volume matters less than accuracy (SOT).

# QUESTIONS

Thank You!