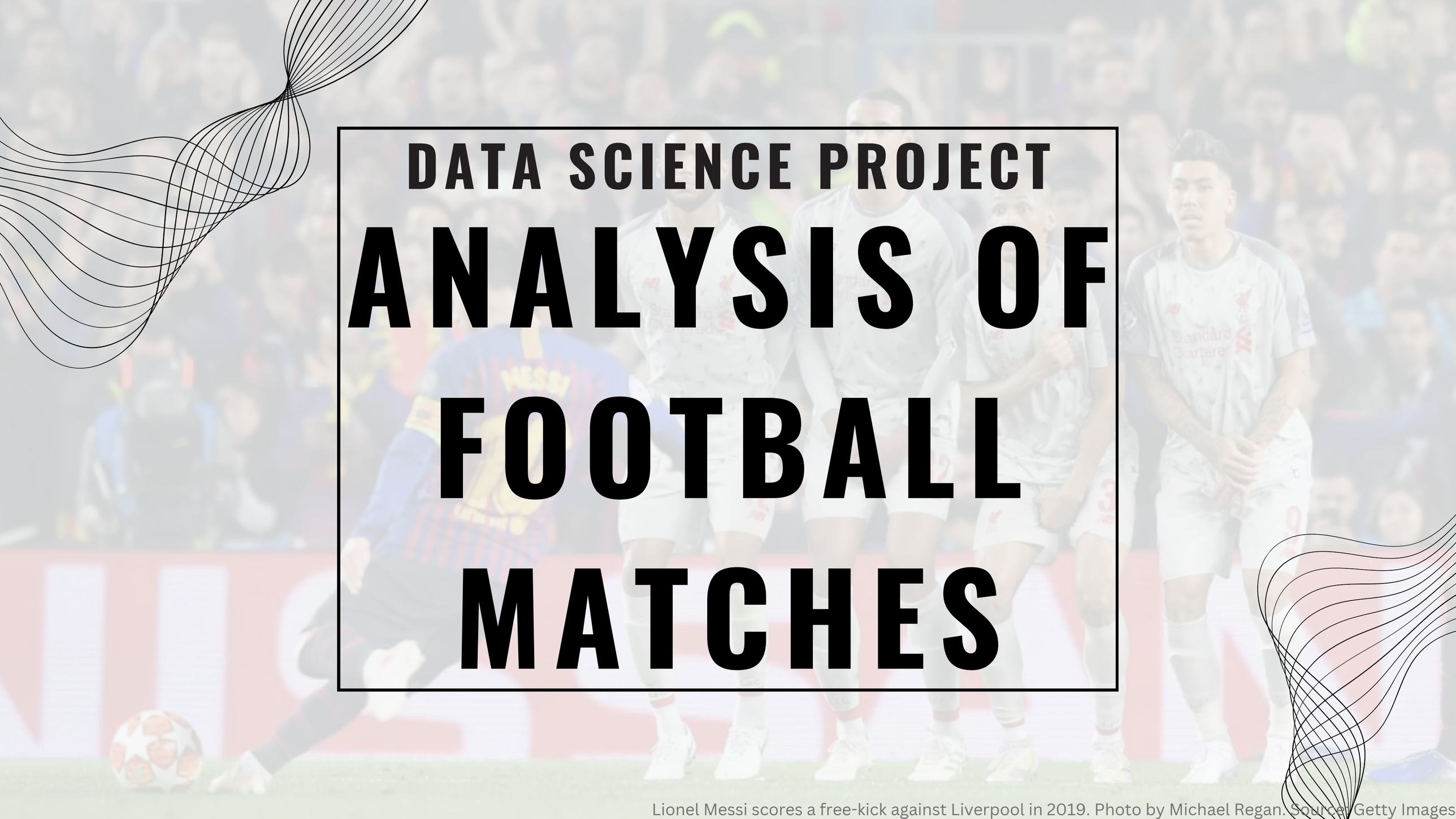
DATA SCIENCE PROJECT

ANALYSIS OF FOOTBALL MATCHES



Lionel Messi scores a free-kick against Liverpool in 2019. Photo by Michael Regan. Source: Getty Images

CONTENT

MOTIVATION

Our motivation to work on the football dataset came from our love for the sport and because it has a global fan base.

Football data can be used to answer a variety of questions about the sport. This includes questions about player performance, team strategy, and the overall dynamics of the game.

PSG Win over Milan - PSG Official Website - https://en.psg.fr

OUR RESEARCH QUESTION

Primary: *Does having more left-footers affect the outcome of football matches in the English Premier League?*

Secondary: Do other physical properties of (a) player/s affect the outcome of a match?

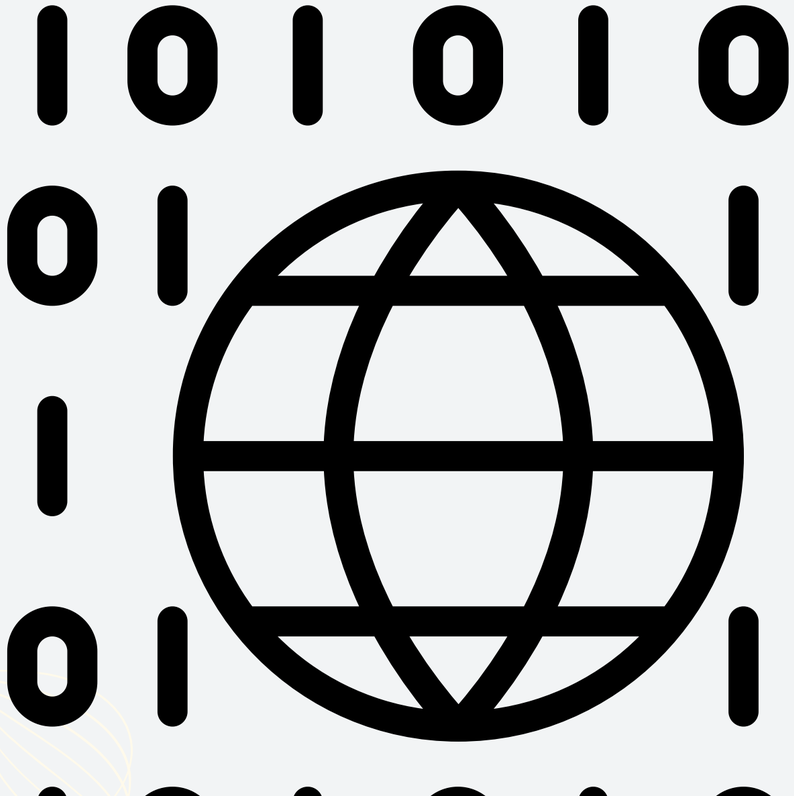
- *Height of players (min/max/avg)*
- *Weight of players (min/max/avg)*

Null Hypothesis *- The number of left-footed players does not impact the outcome of the match*

ABOUT THE DATA

The data was from publicly available (spatio-temporal) football data (study published in Nature, Scientific Data) which has metrics from the top five European matches and the World Cup.

Pappalardo, L., Cintia, P., Rossi, A. et al. A public data set of spatio-temporal match events in soccer competitions. Sci Data 6, 236 (2019). https://doi.org/10.1038/s41597-019-0247-7



ABOUT THE DATA

The data was arranged as Matches, Events, and Individual data.

In Complex nested JSON and CSV files

~40 M

Datapoints

ABOUT THE DATA

Matches

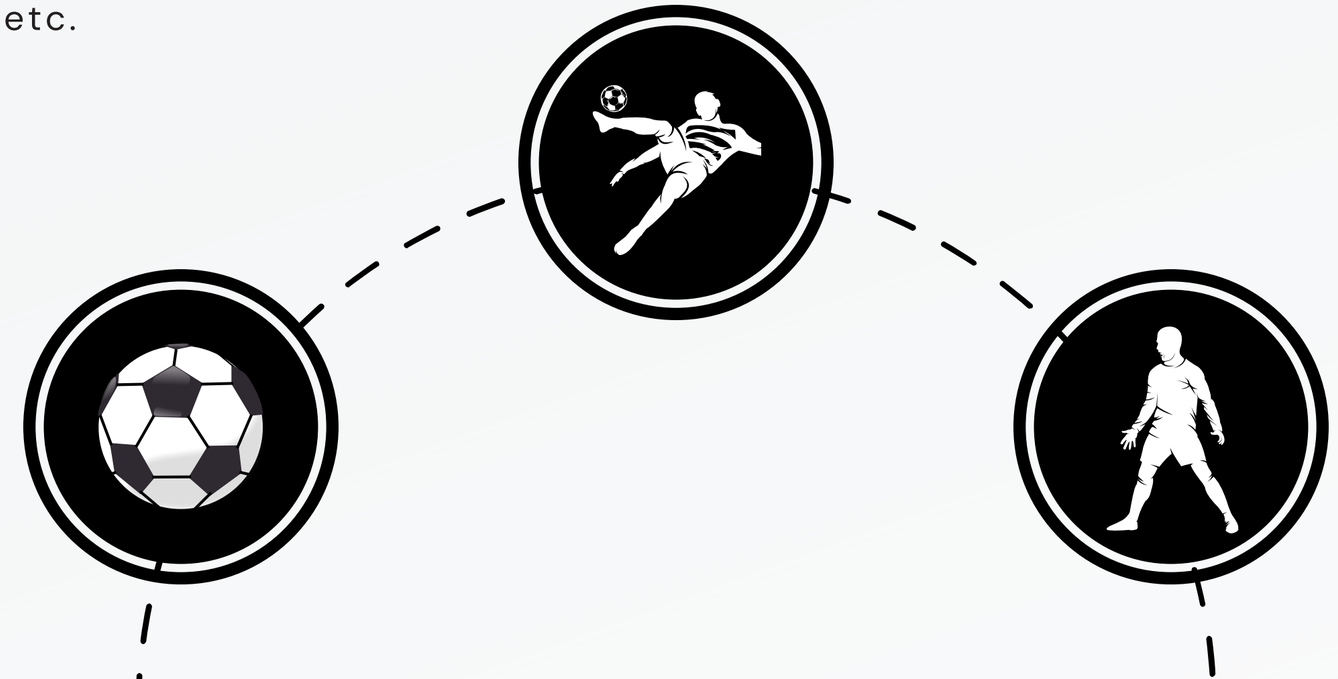
each match Included season, teams, players for each team, their match highlights, referees details etc.

Events

Spatiotemporal data including passes, throw-ins, free kicks etc.

Indivdual

Details about players, coaches, referees, teams etc.



01

EXTRACT

Data had to be **manually extracted** as it was in the form of complex nested JSON and CSV files. Special considerations were needed as **some data was not collected in some countries and tournaments**

02

CLEAN

Somewhat clean. We needed to consider some “0” values and “null” (str) values.

03

ANALYSE

Analysis was done in python with the help of specific libraries.

VISUALIZATION

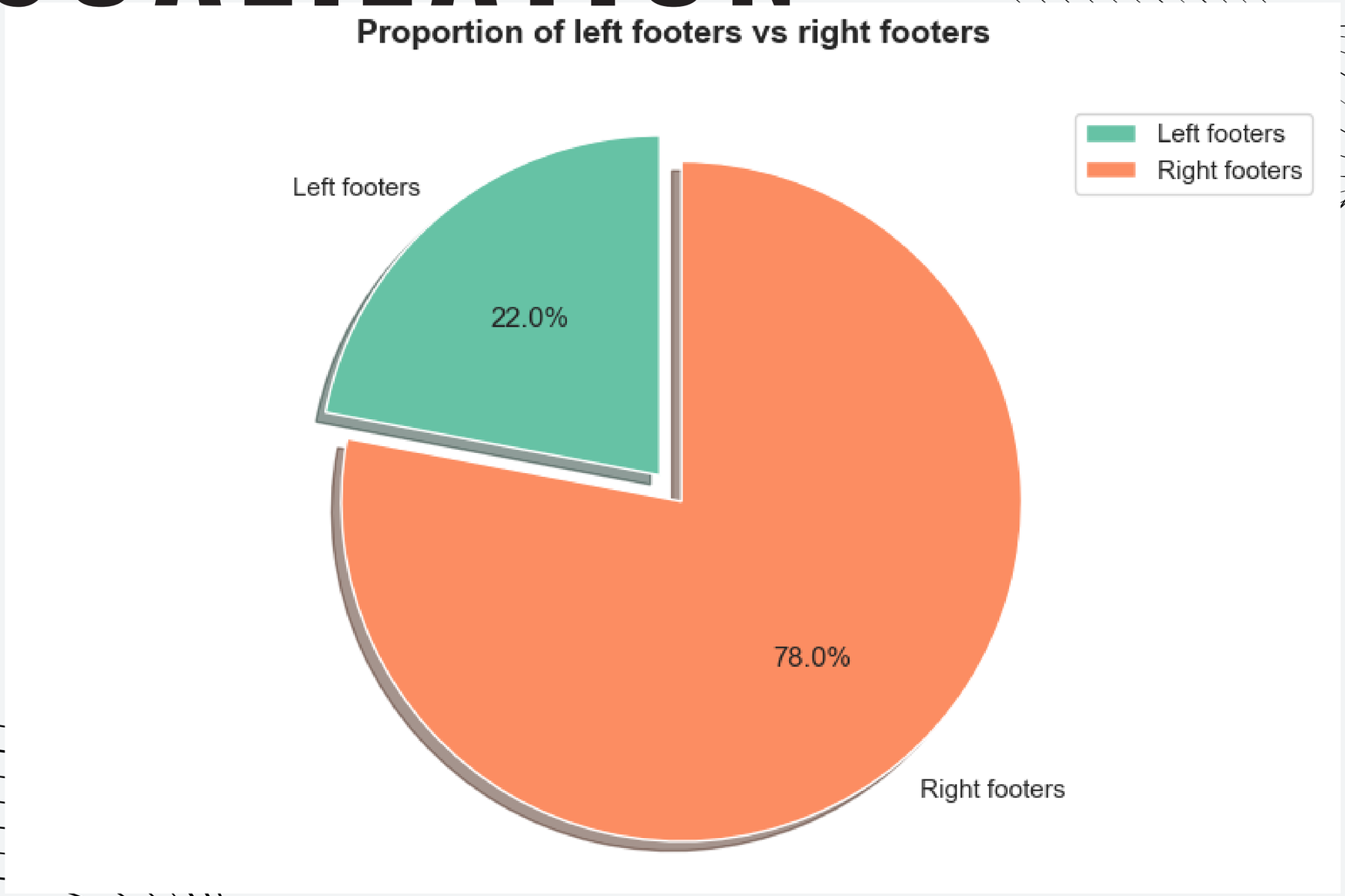
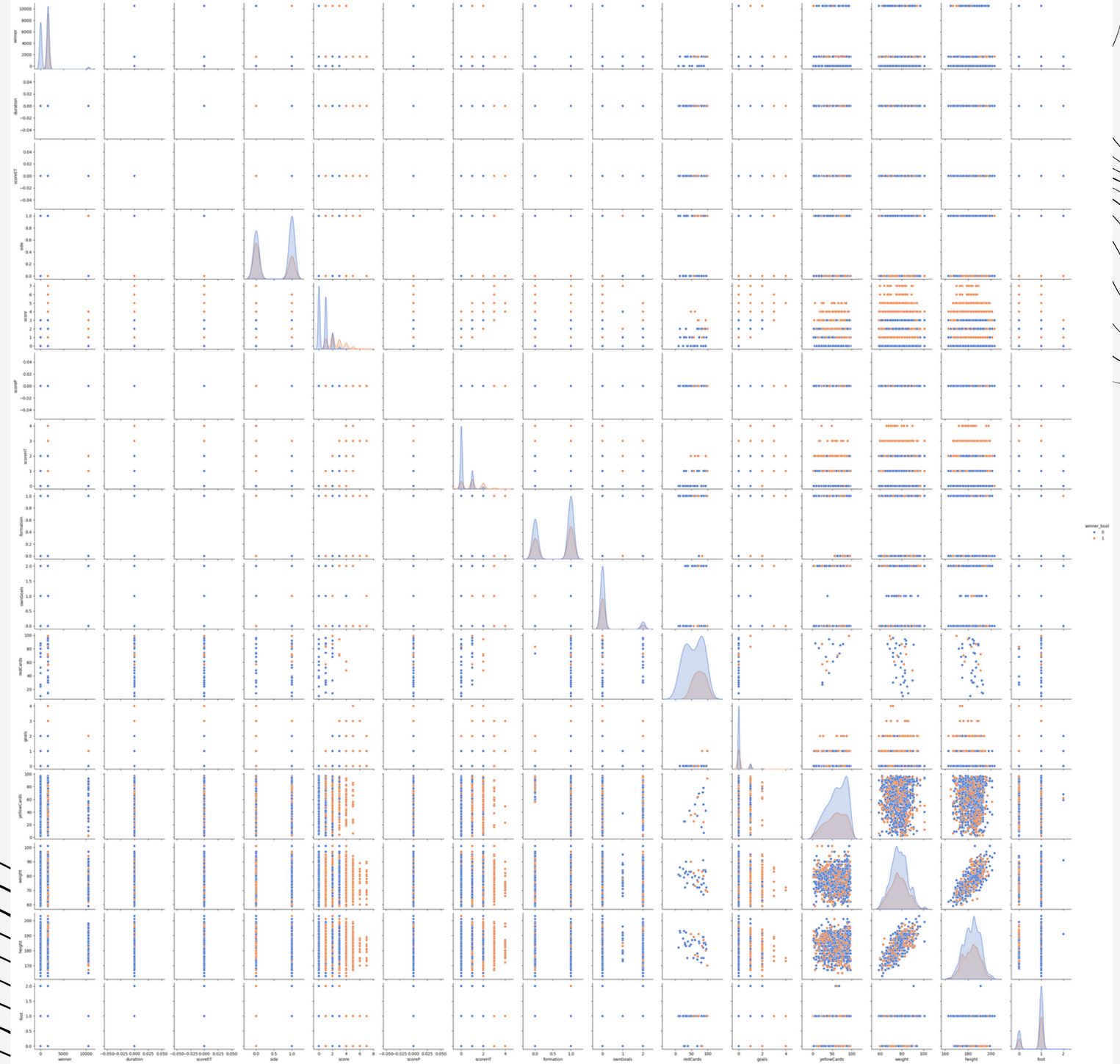
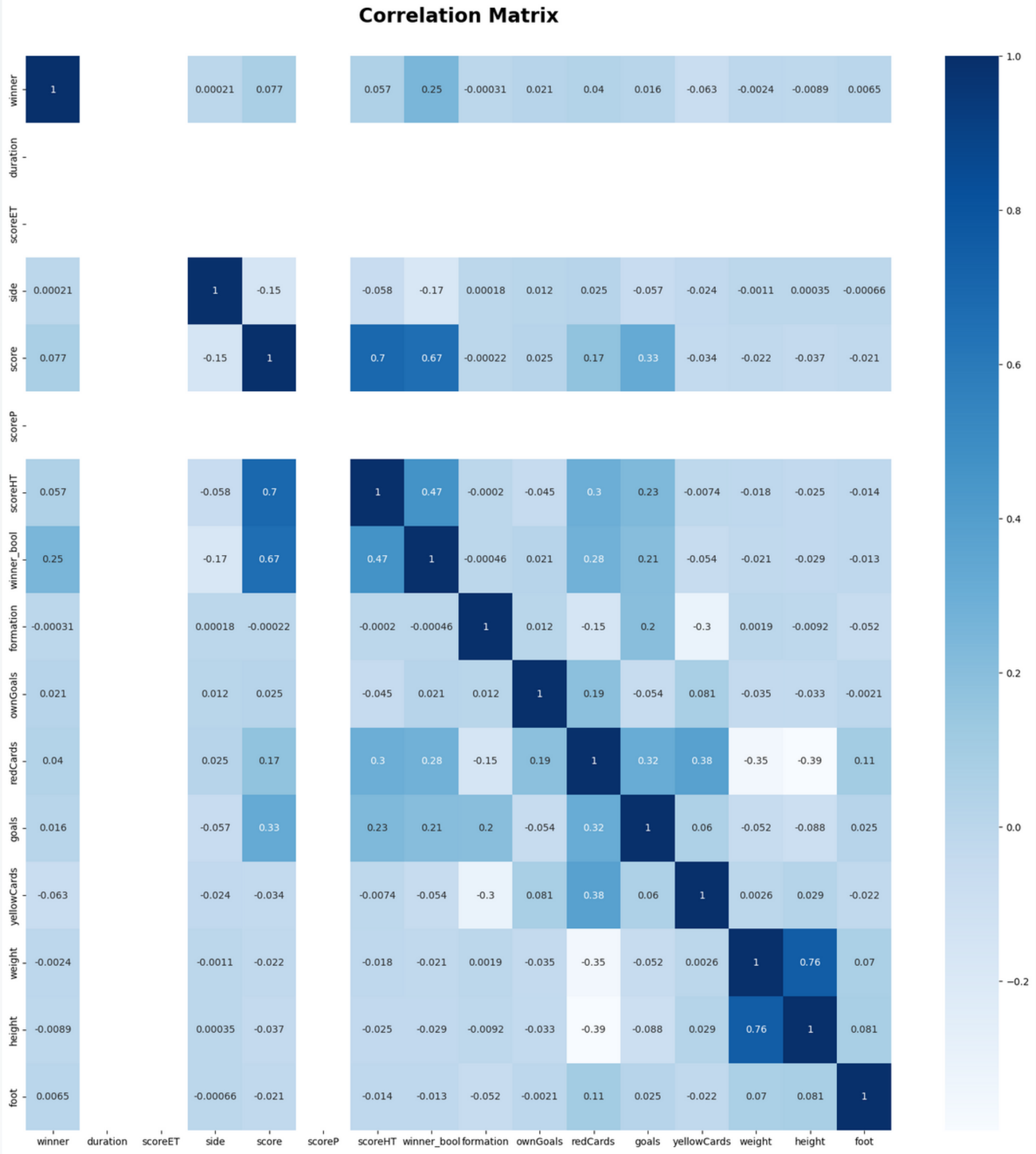


Fig: Footedness of the players in the tournament (England)


winner
duration
scoreET
side
score
scoreP
scoreHT
formation
ownGoals
redCards
goals
yellowCards
weight
height
foot
winner_bool
0
1


Correlation Matrix
winner
duration
scoreET
side
score
scoreP
scoreHT
winner_bool
formation
ownGoals
redCards
goals
yellowCards
weight
height
foot
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

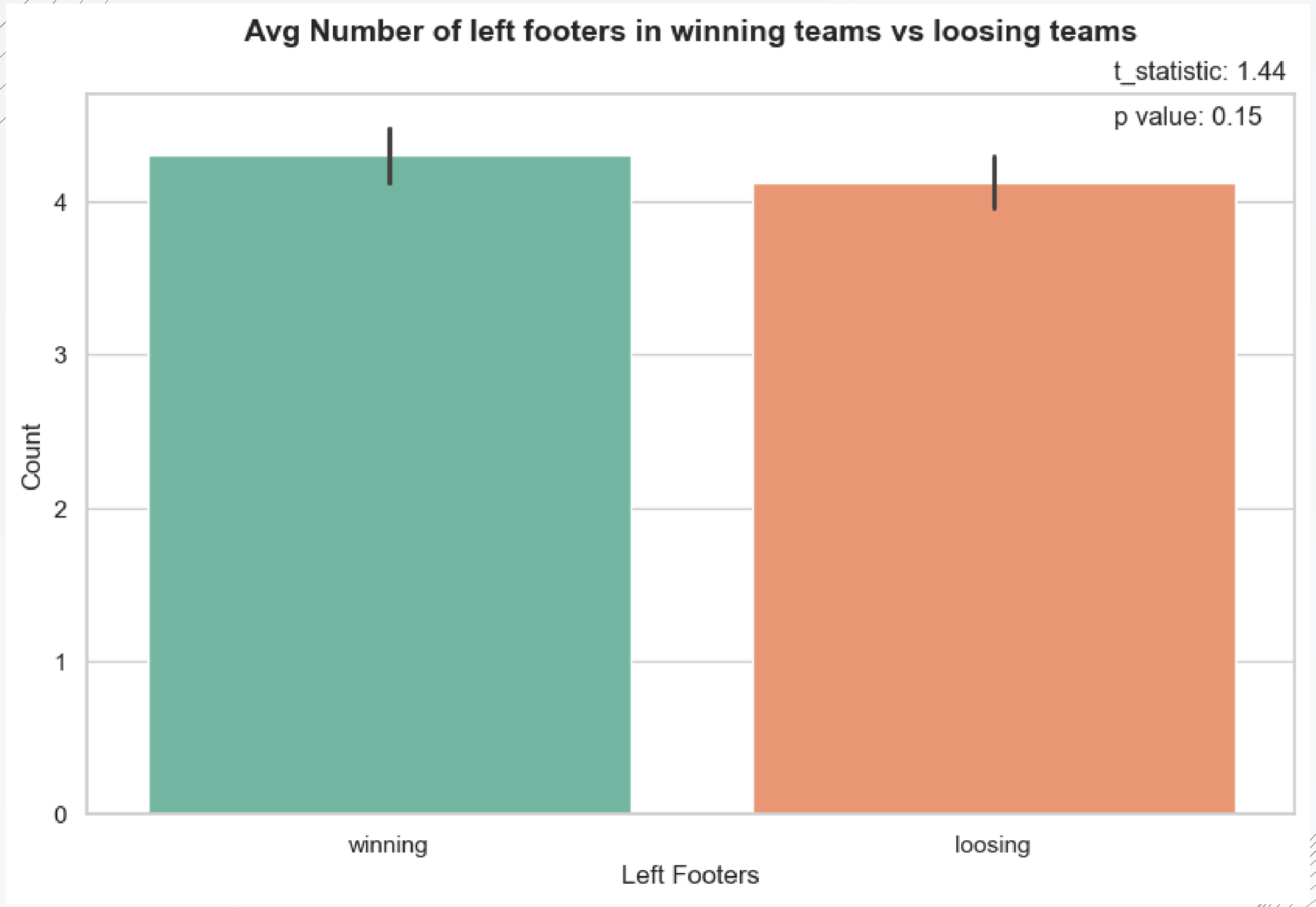


Fig: Average number of left footers in winning vs loosing team

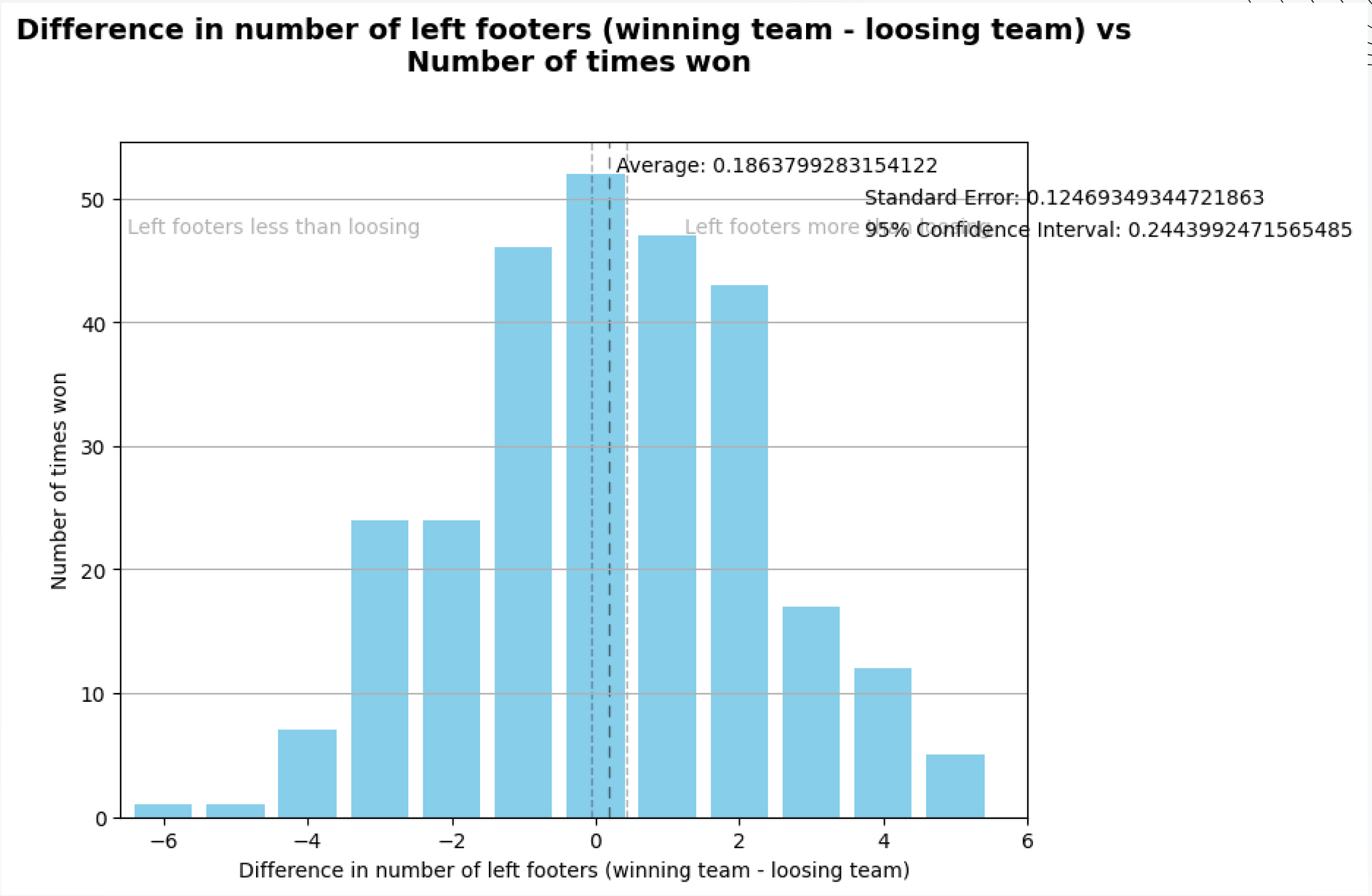


Fig: Distribution between left-footed players and winning

No of Left Footers	Wins	Winning Probability
0	0	0.0
1	9	0.375
2	28	0.358974
3	53	0.355705
4	65	0.338542
5	55	0.335366
6	44	0.435644
7	22	0.478261
8	3	0.6

Table: Winning probability with number of left foot players

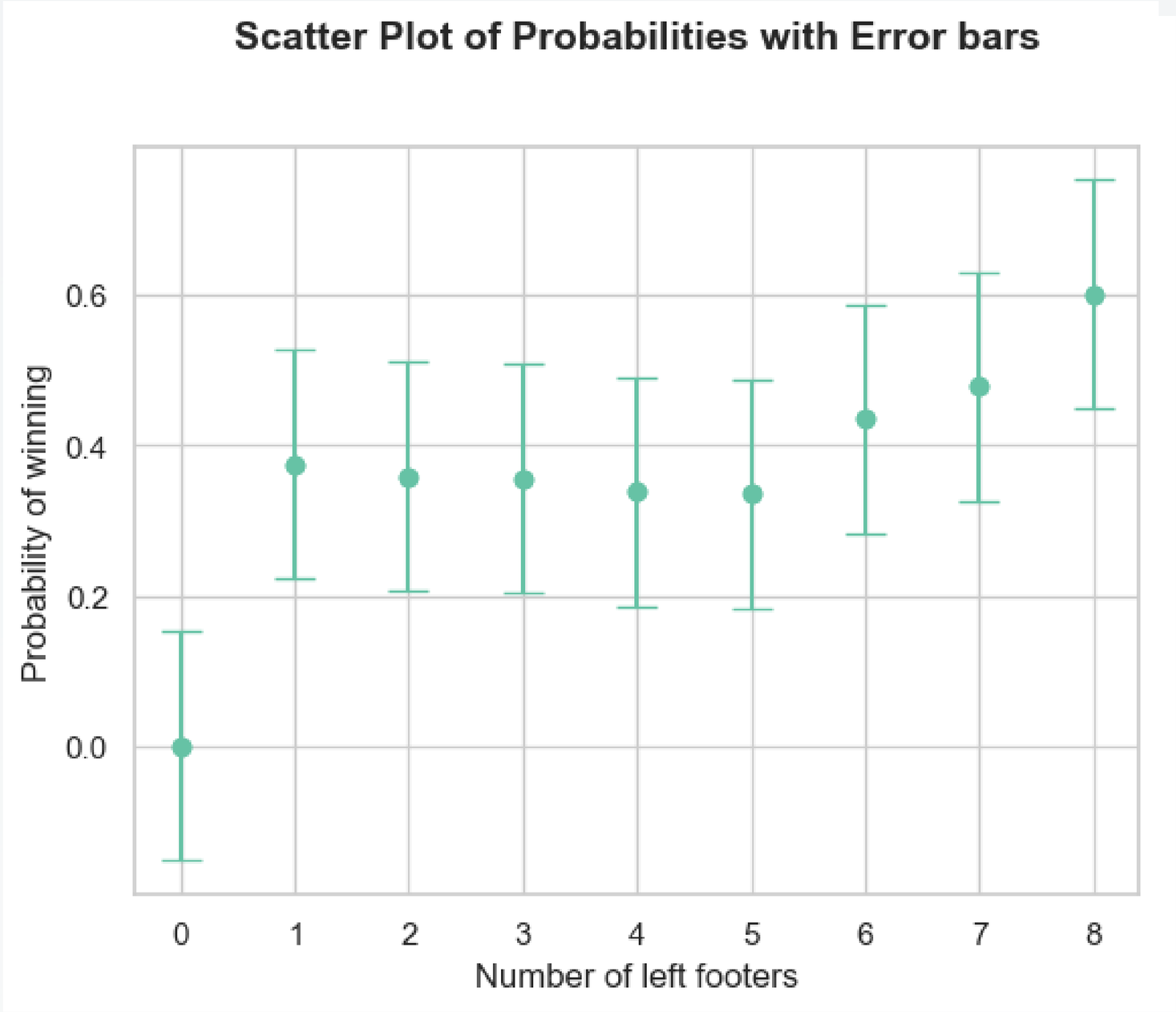


Fig: Scatter plot of winning proberbility with number of left-footed players

	Winning	Loosing
Left footers	1201	1149
right footers	3807	3847

t- statistic score: 1.28

p-value: 0.20

Table: Contingency table, left & right footed players played matchup in winning and losing teams

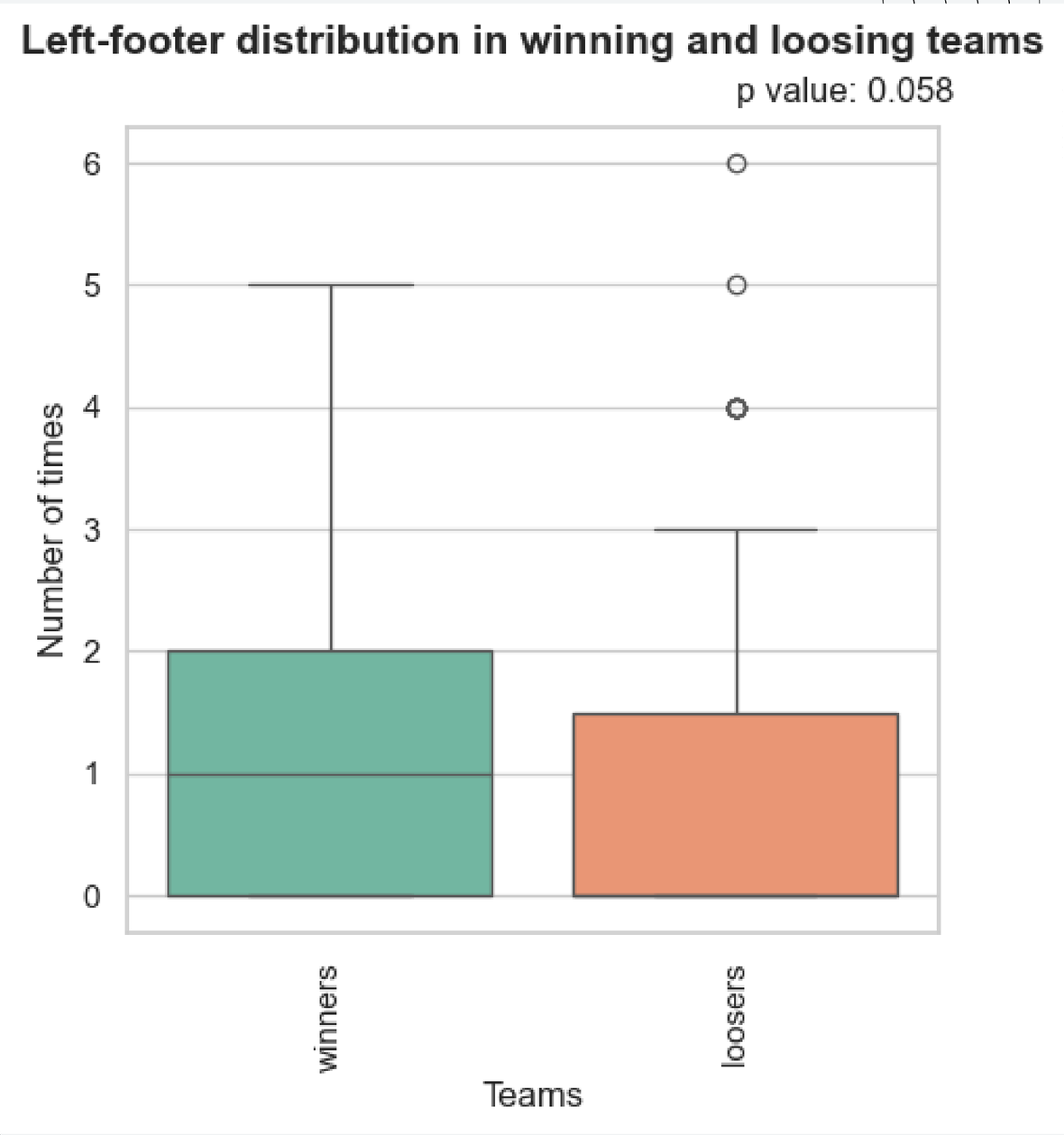
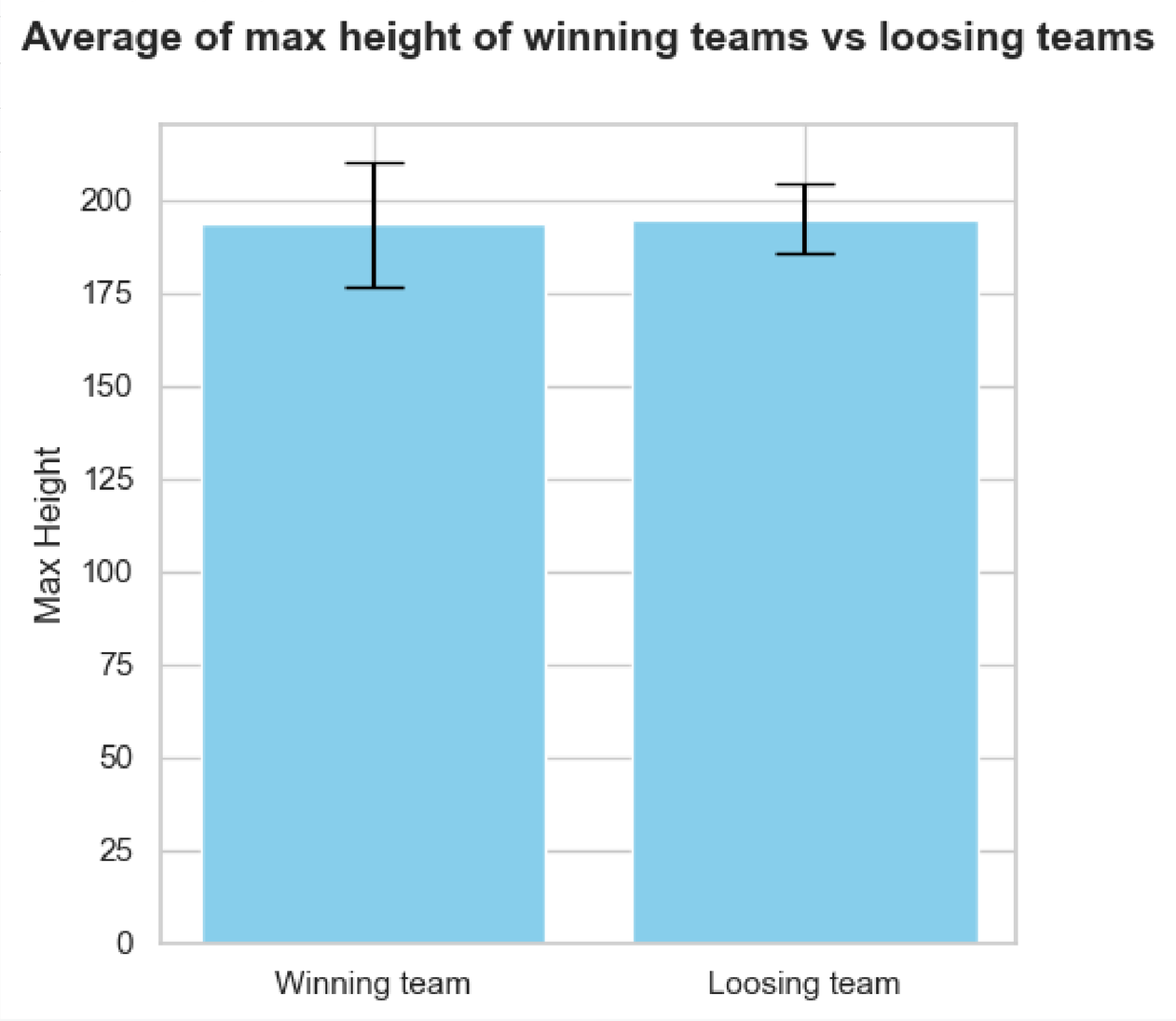


Fig: Distribution of left-footed players across winning and loosing



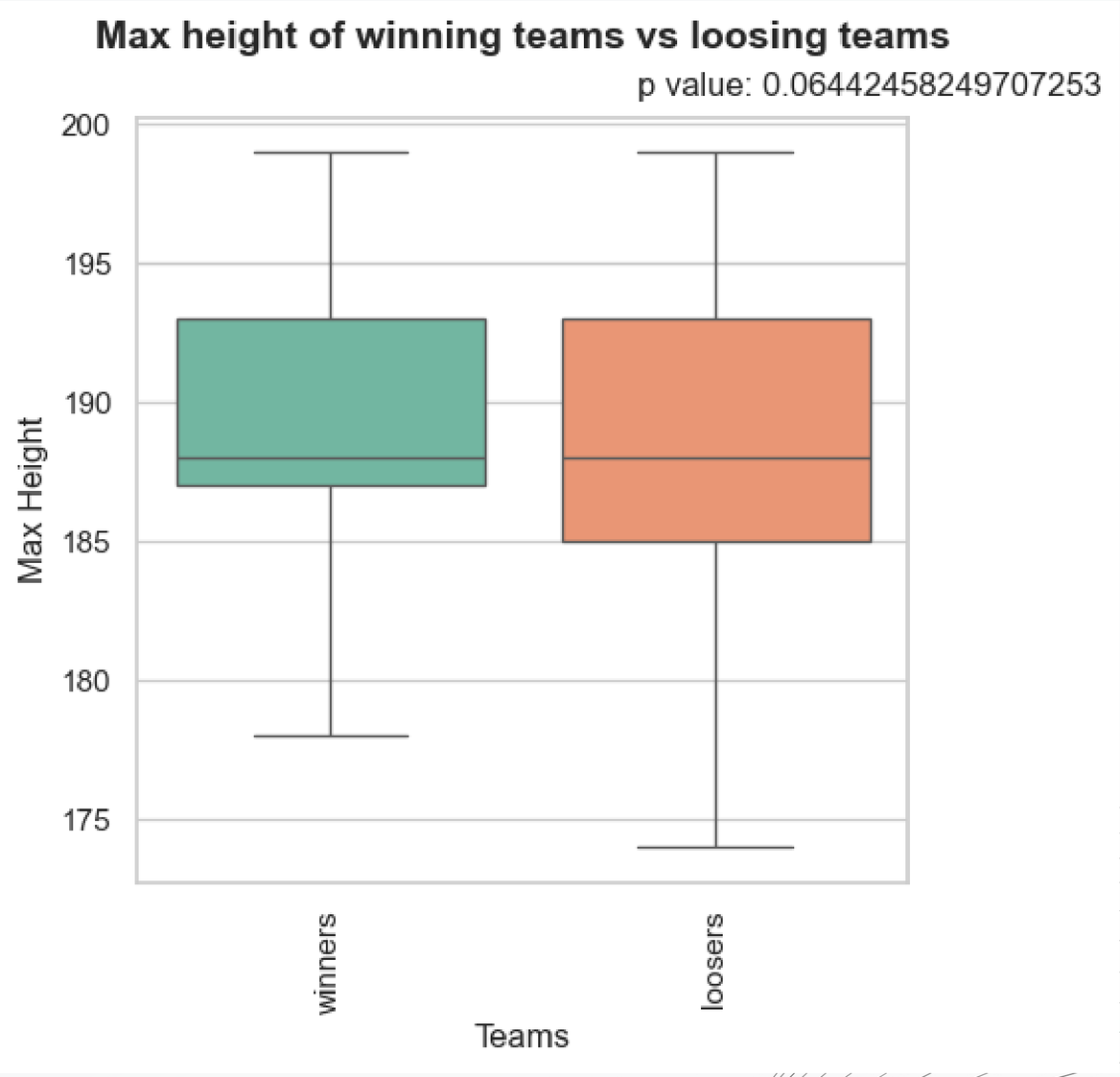
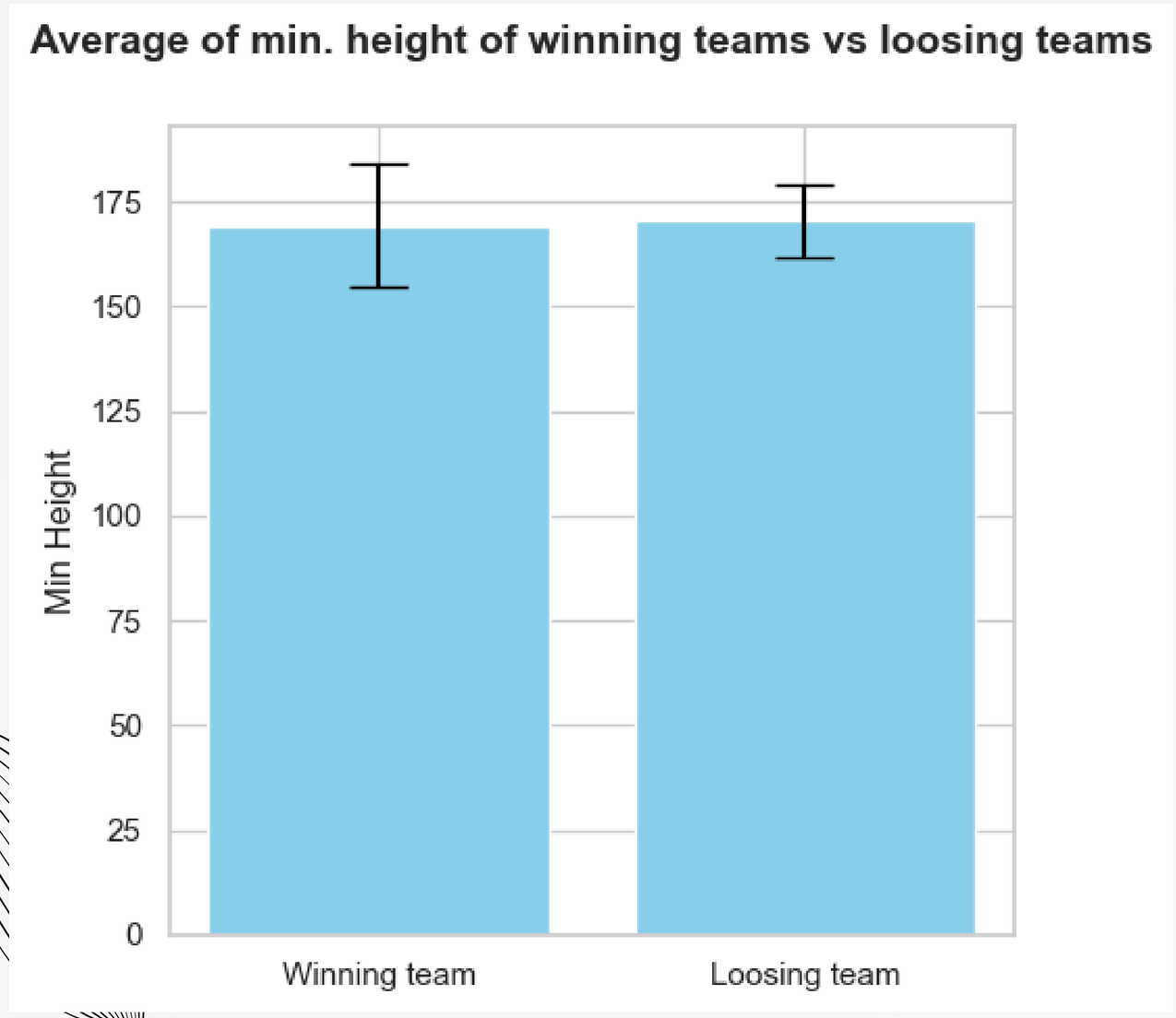


Fig: Footedness of the players in the tournament (England)



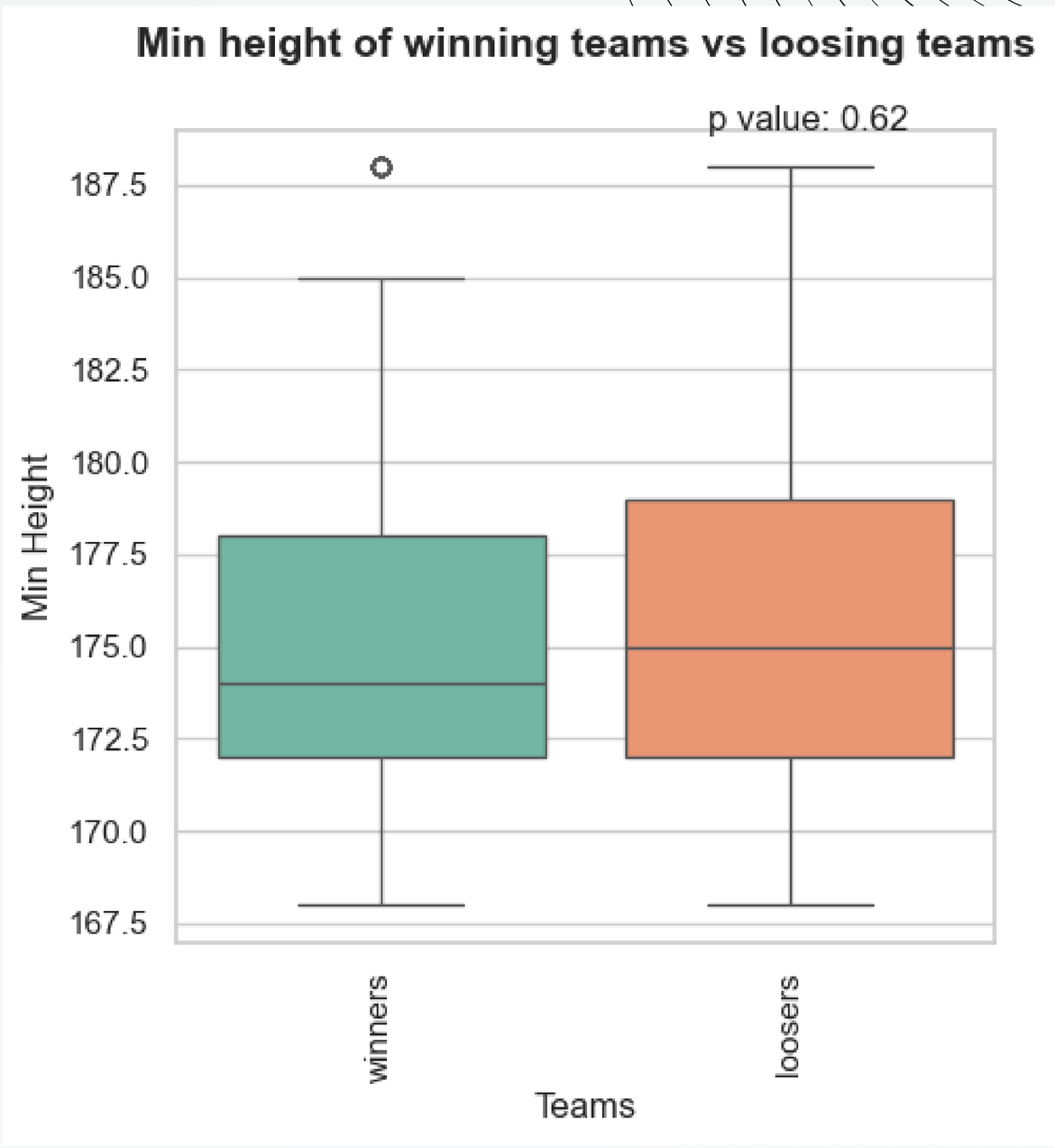
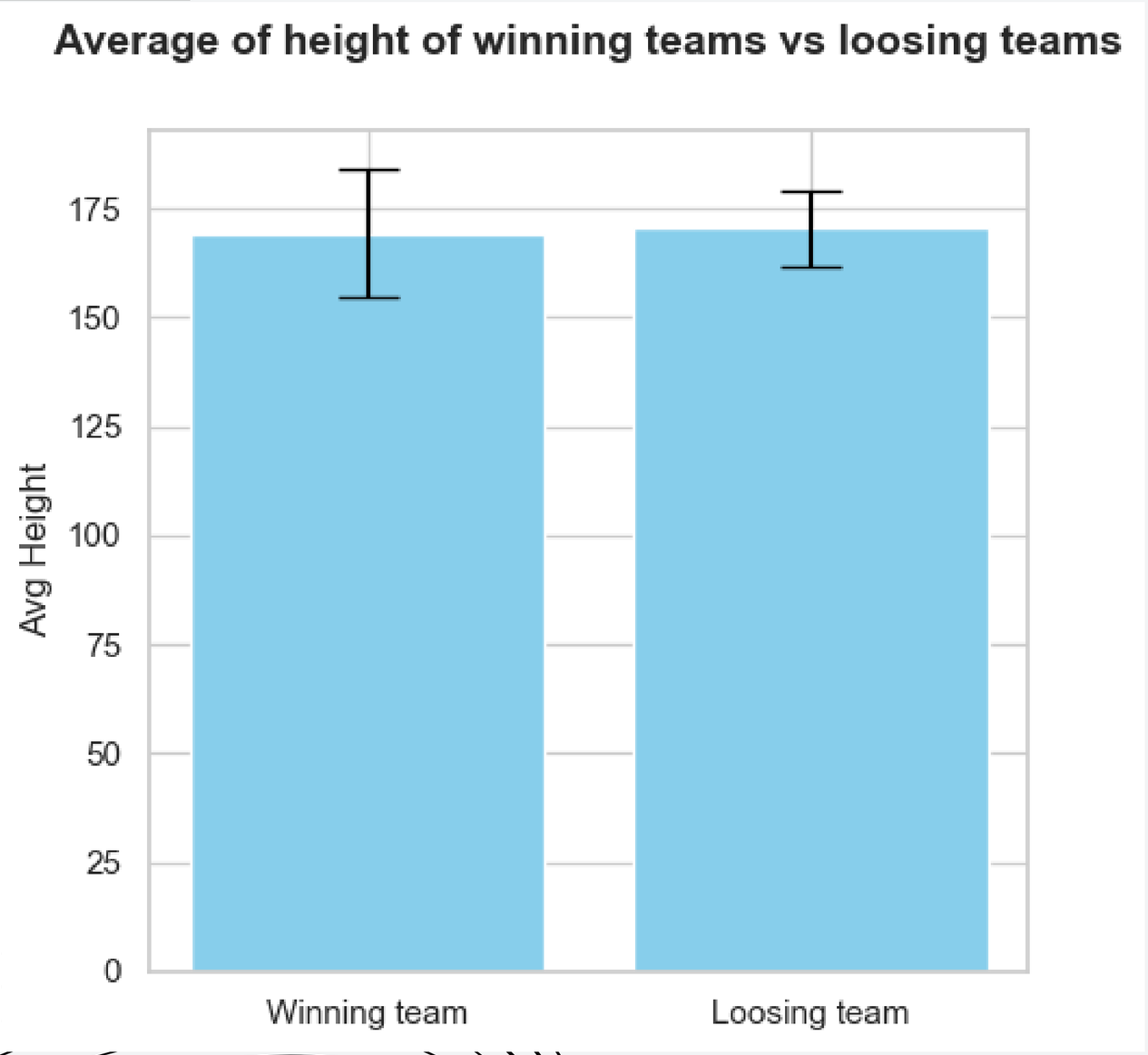


Fig: Footedness of the players in the tournament (England)



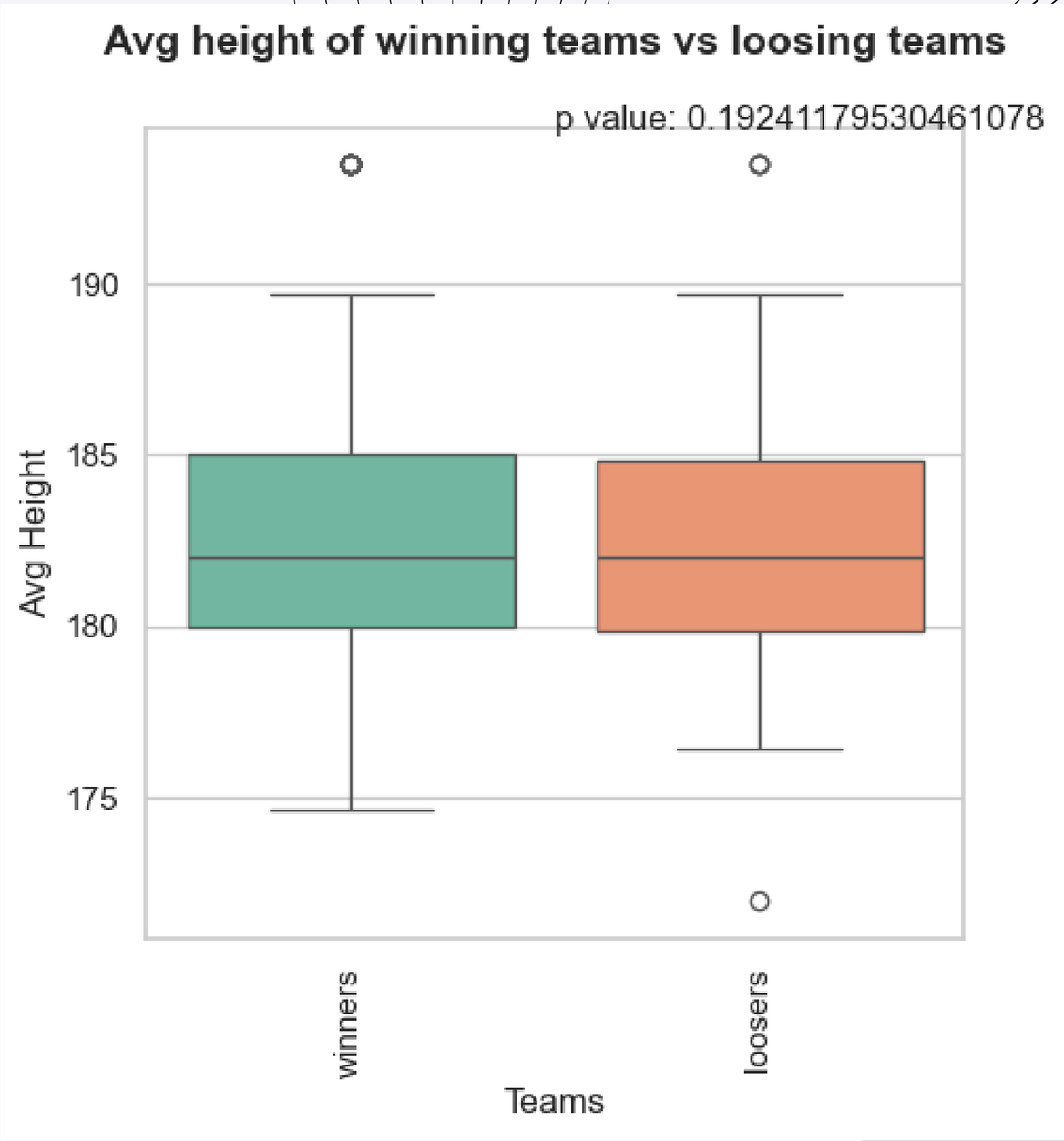
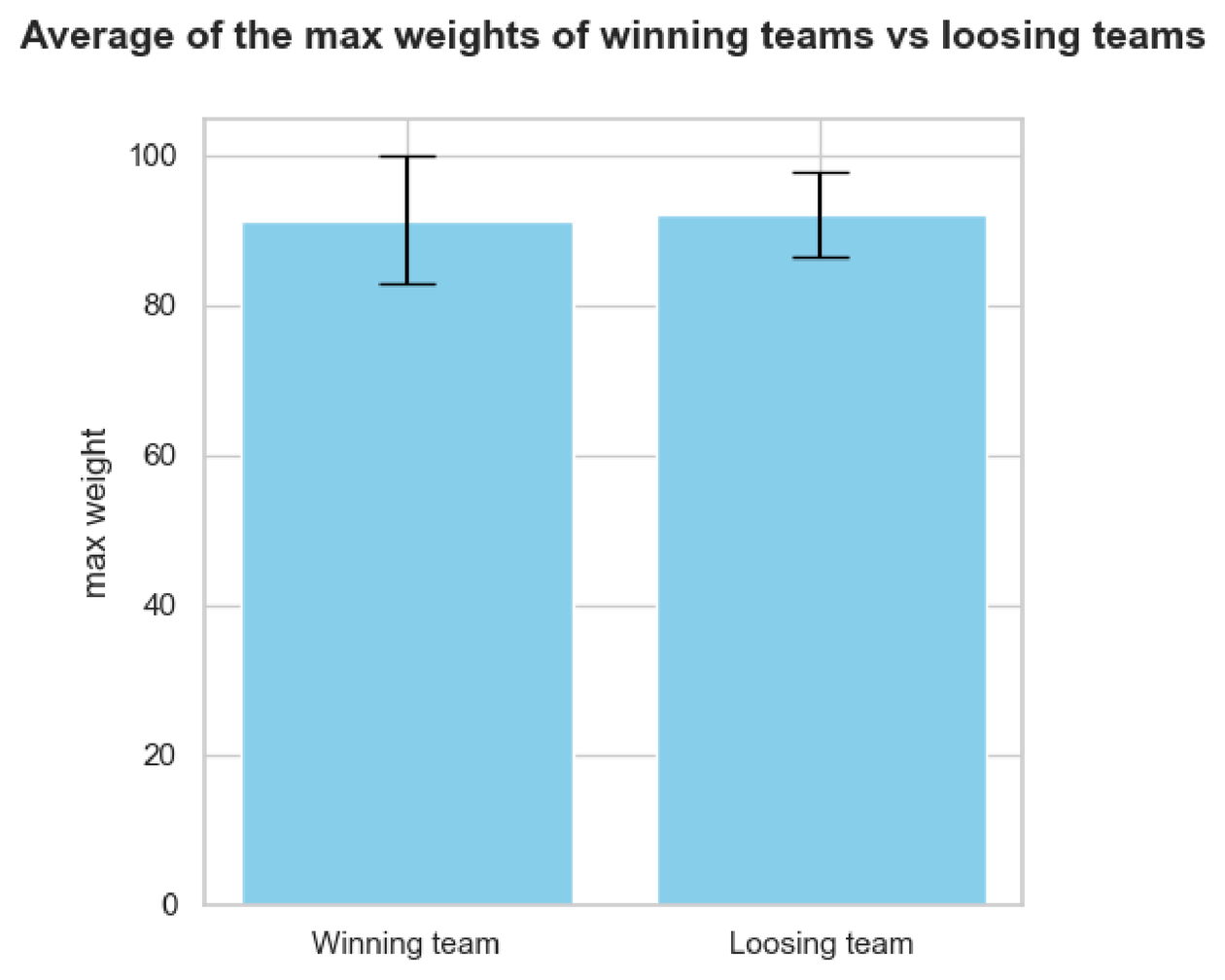


Fig: Footedness of the players in the tournament (England)



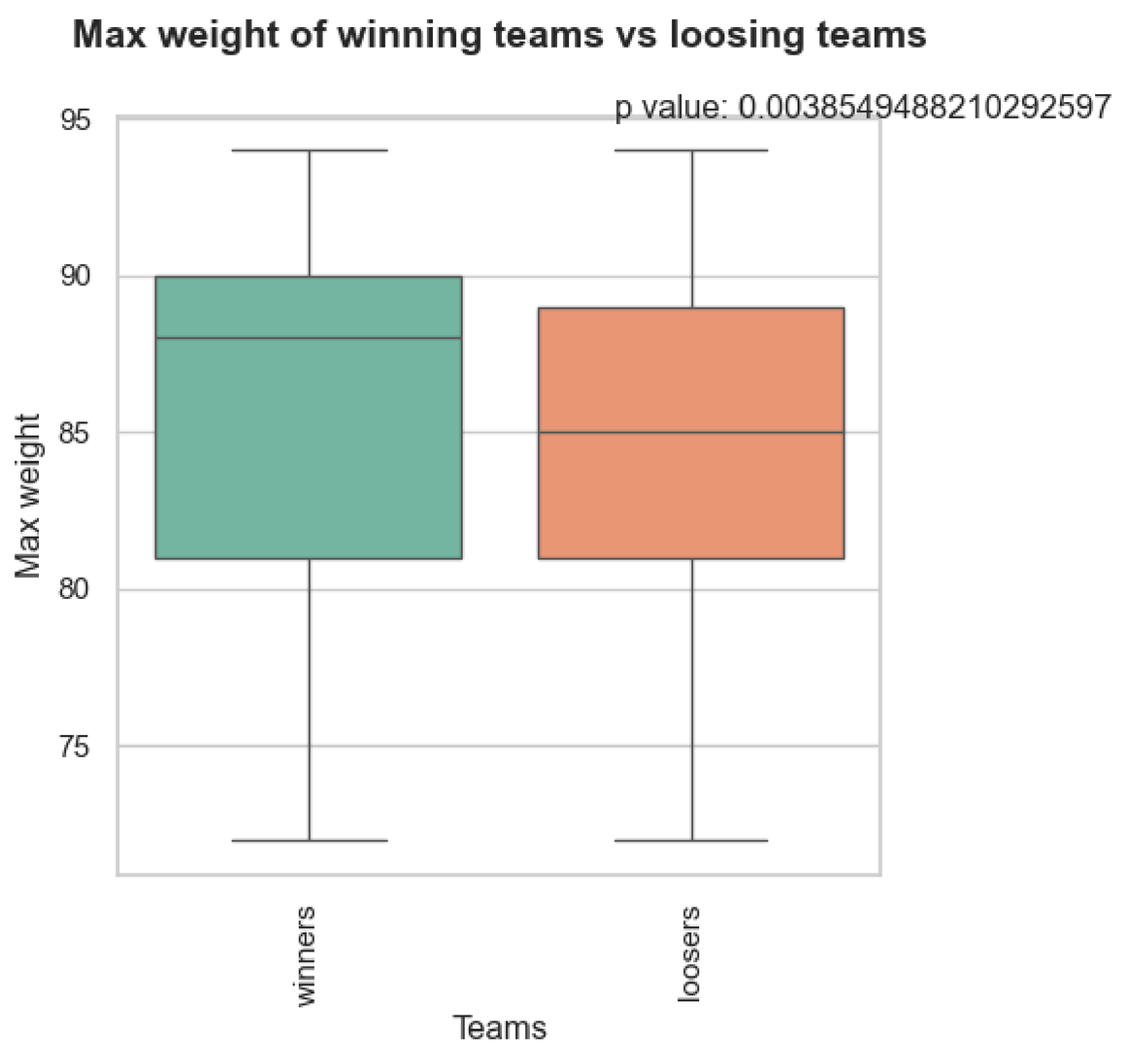
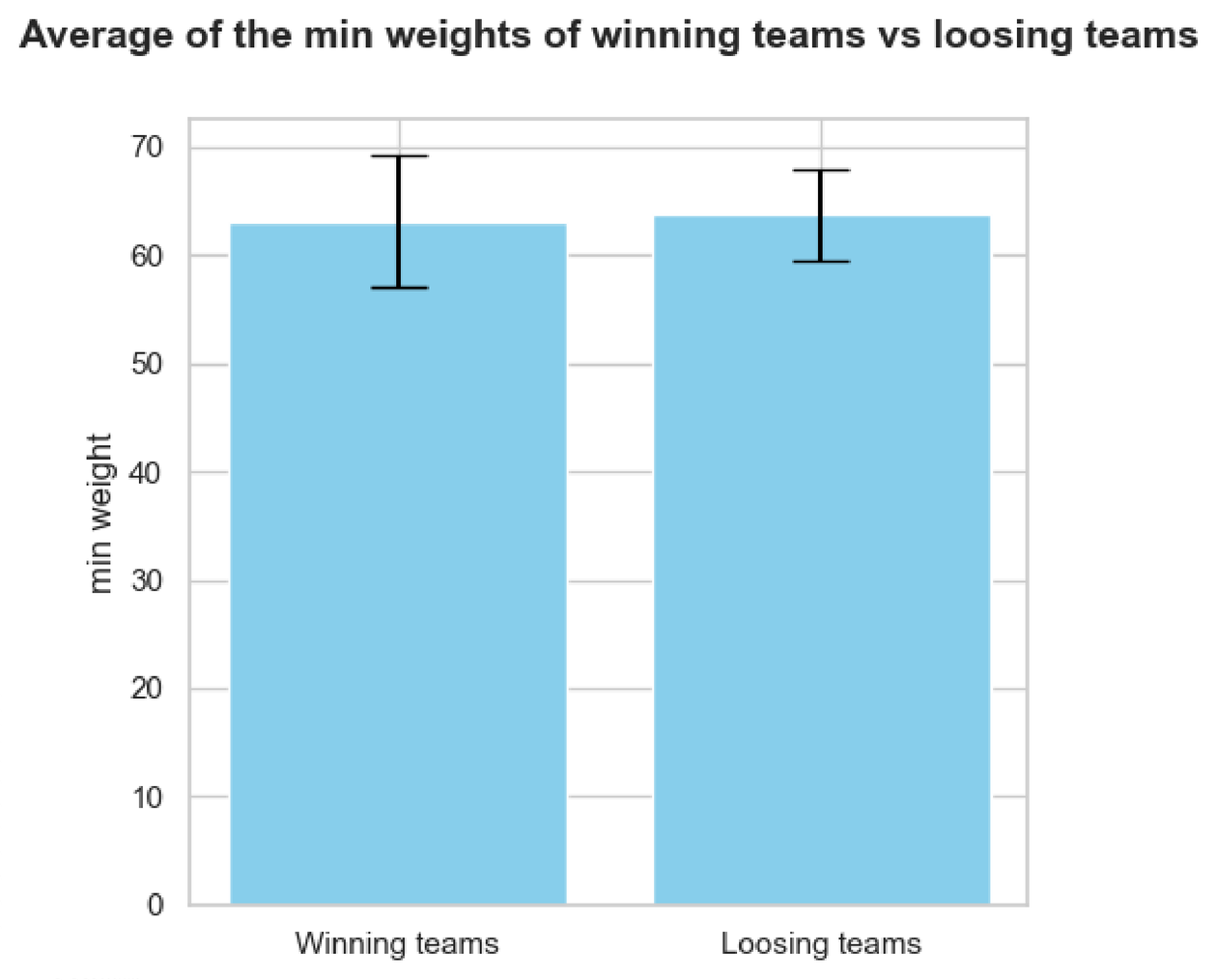


Fig: Footedness of the players in the tournament (England)



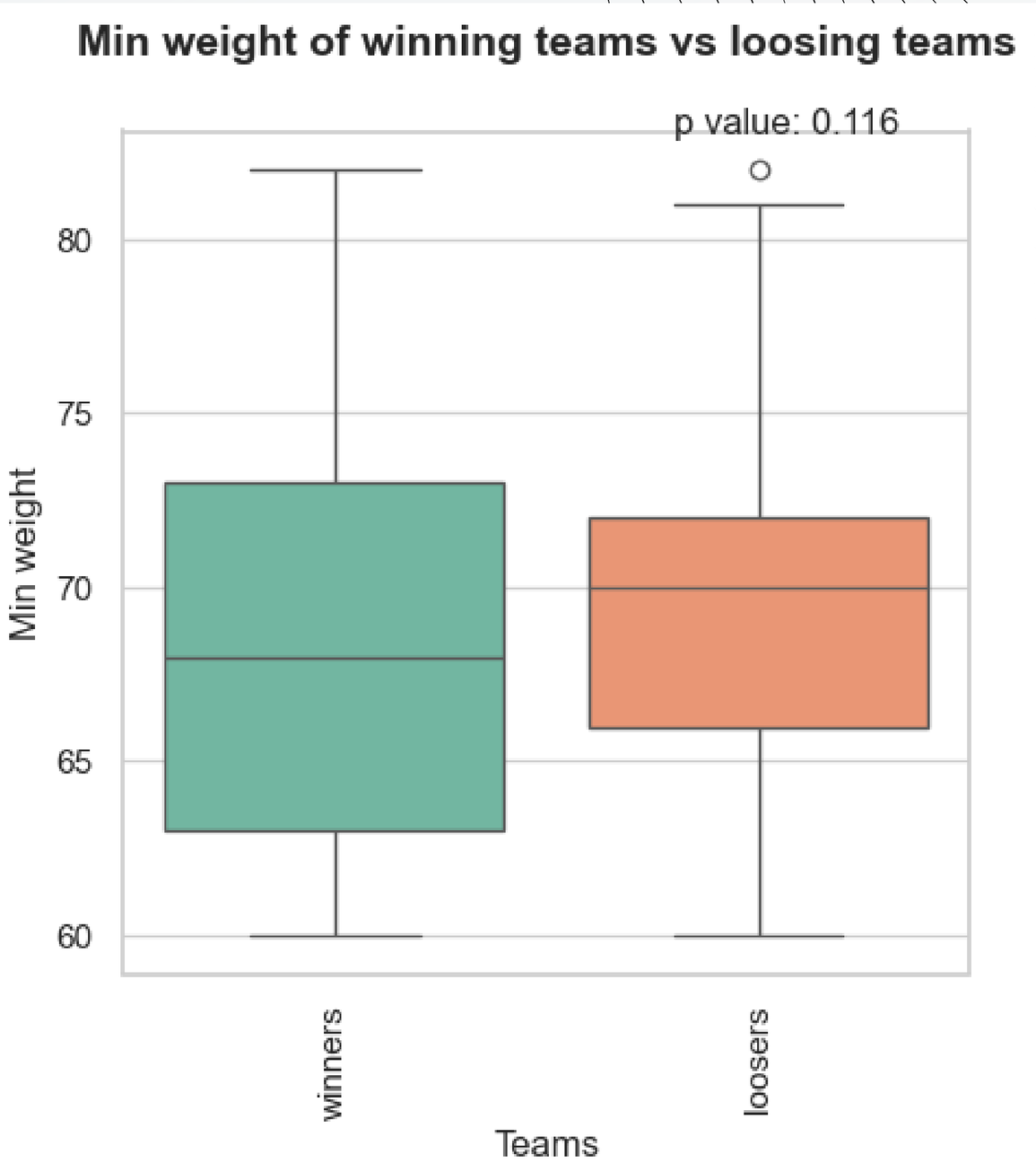
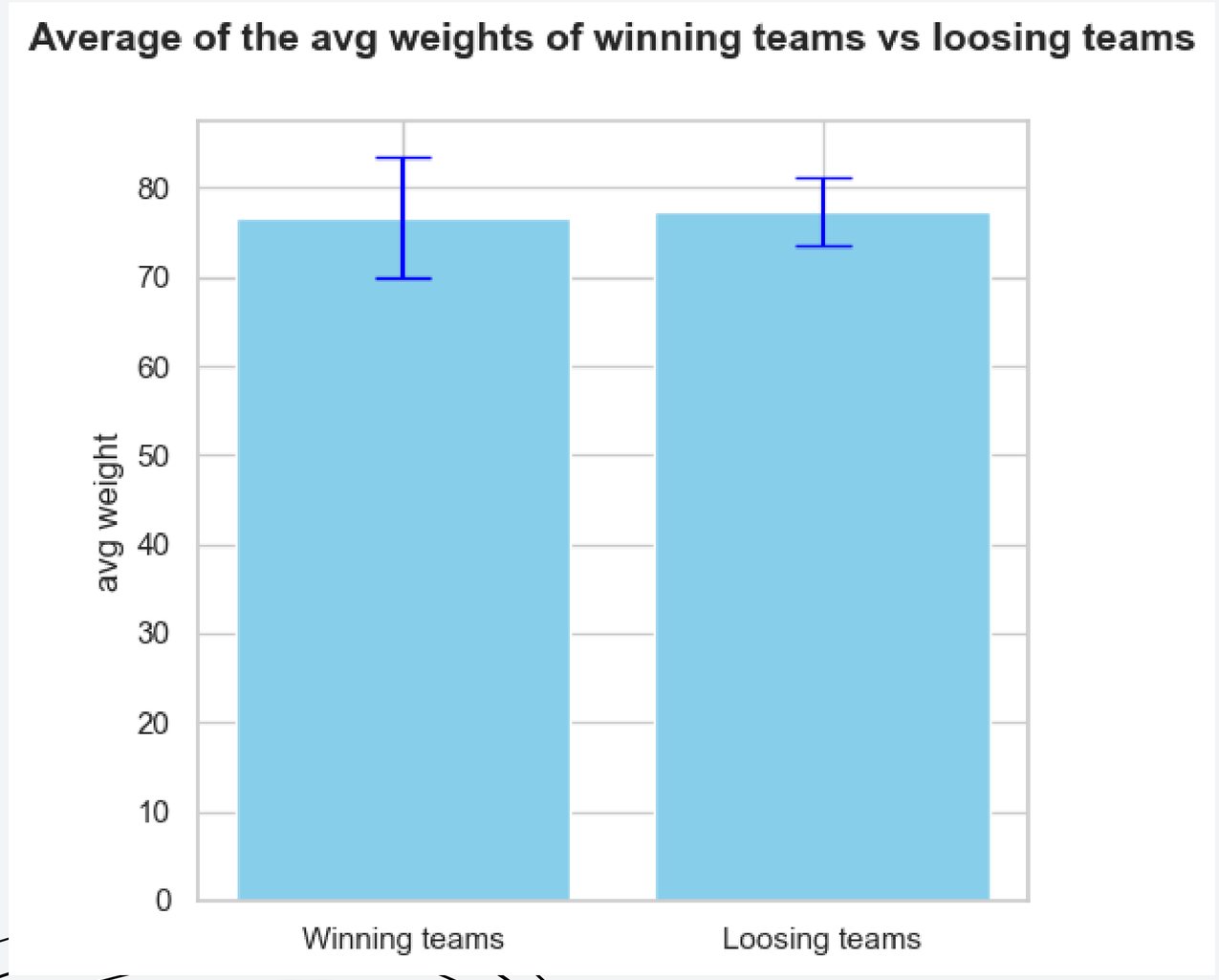


Fig: Footedness of the players in the tournament (England)



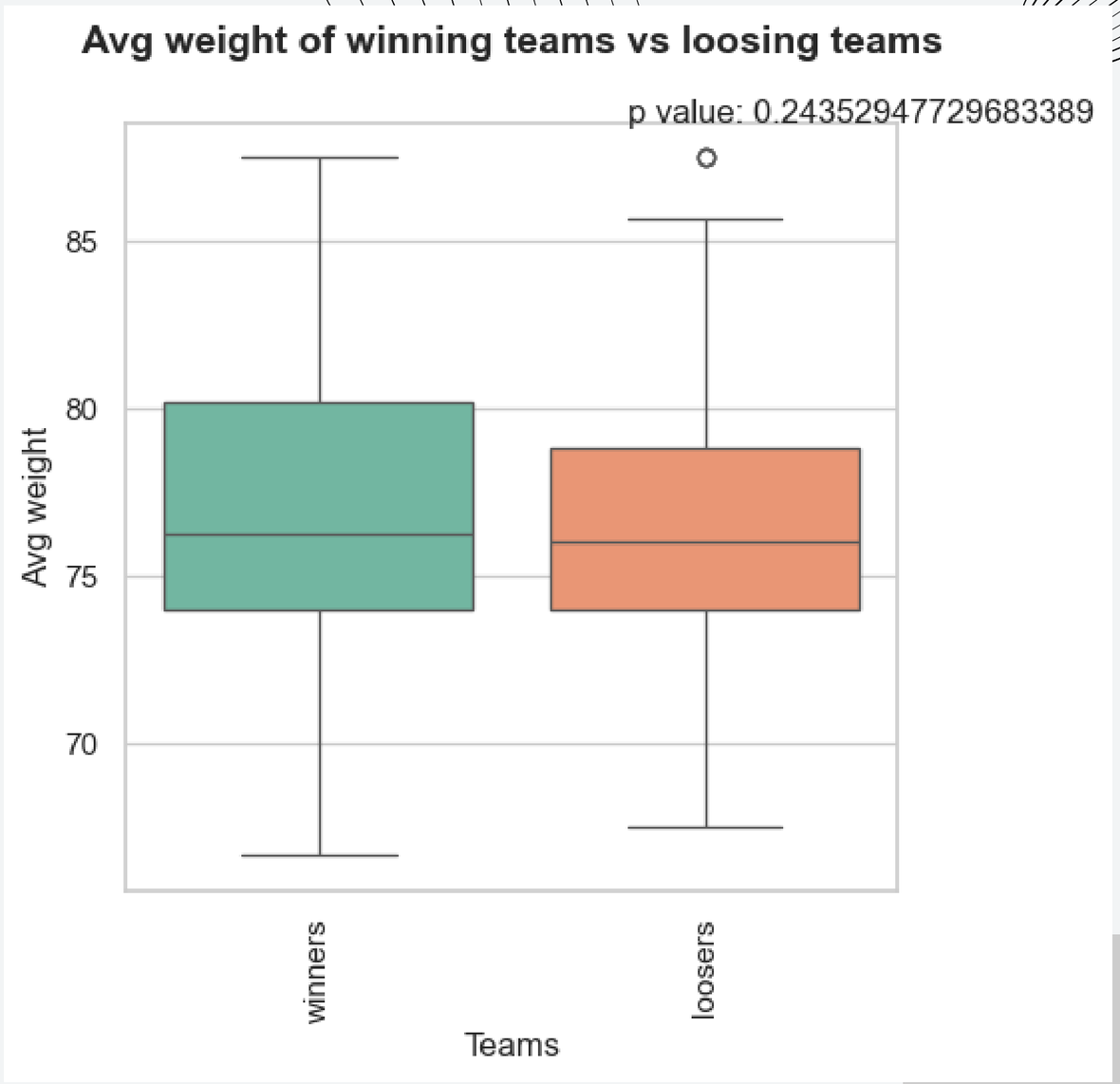


Fig: Footedness of the players in the tournament (England)

CONCLUSION & DISCUSSION

We were unable to disprove the null hypothesis that the number of left-footed players does not impact the outcome of the match!

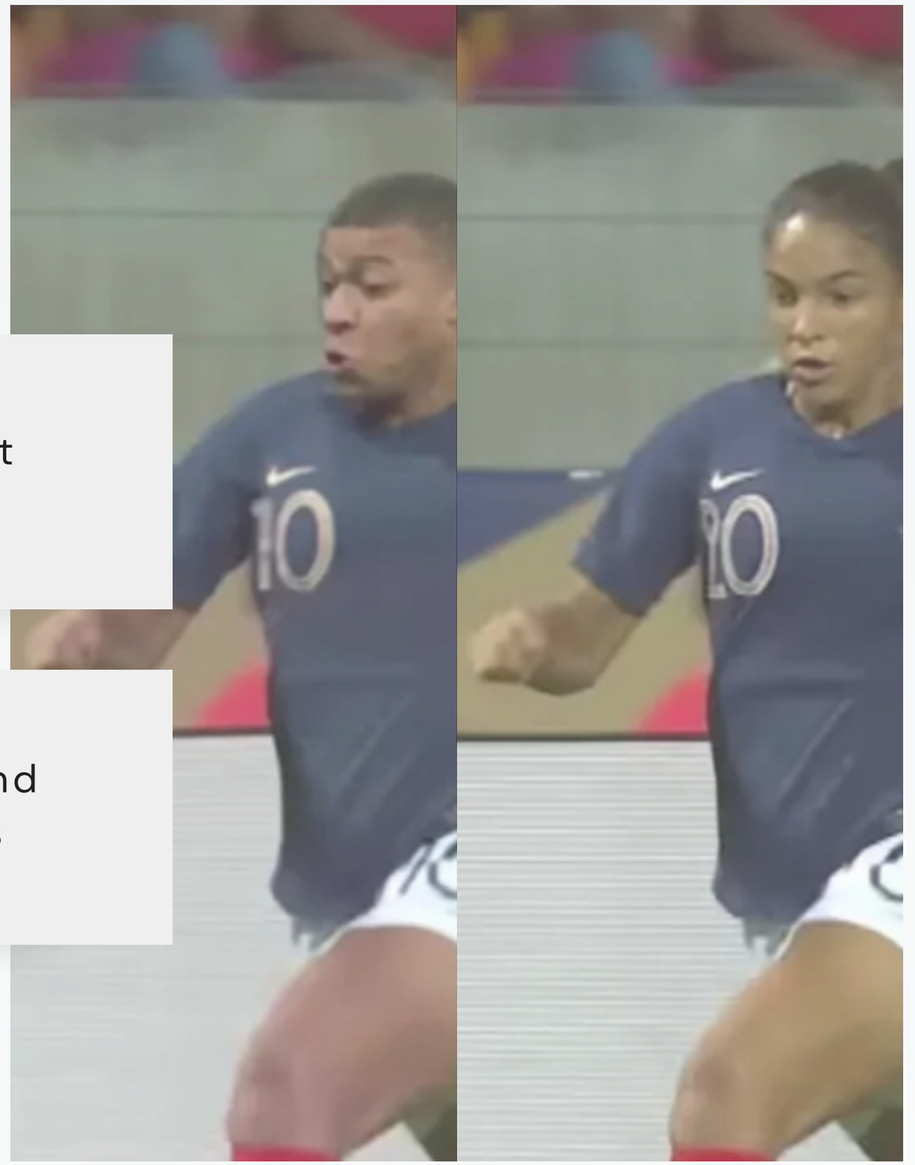
We Couldn’t find any significant data to say there was an association between players' height and the outcome of the game.

There was a significant difference between the winning teams' heaviest players and the losing teams' heaviest players. This might be an interesting topic to explore

ASSUMPTIONS

- Benched players were considered as if they were substituted into the match at some point.

- Own goals were considered mistakes and not included in individual players' goals



Off to a Disappointing Start - Euro Sports - https://www.eurosport.fr

LIMITATIONS

- Time in pitch for the players was not considered.
- Matches/teams/players were analysed in respective tournaments, not as a whole.
- Spatio-temporal data was not utilized.

Off to a Disappointing Start - Hungary Today - https://hungarytoday.hu

NEXT UP

Things that will be **neat** to try out

Take the playing time of each player into consideration when analysing the data

STEP 1

Explore the other tournaments, and take all of them as one and analyse the data.

STEP 2

Use the Spatio-Temporal data to make a network model and analyse the resulting data

STEP 3

THANK YOU FOR LISTENING