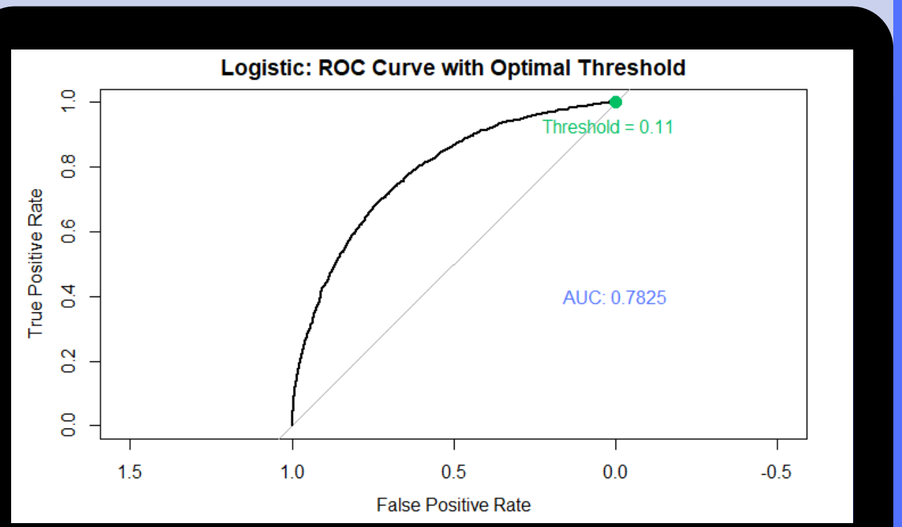
Occhio malocchio

Can Statistical Learning help more than superstitious rituals to understand football?

Contents

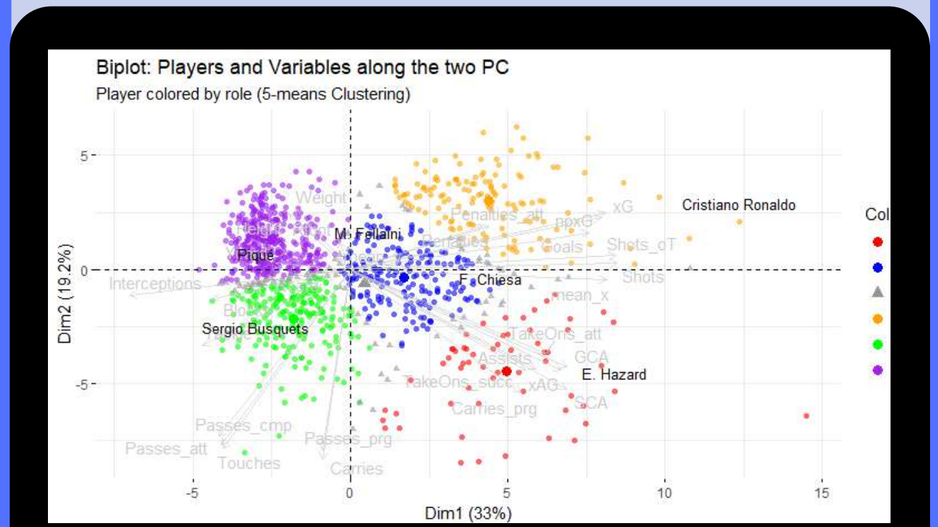
xG model

Predicting scoring probability of a shot based on shot's information



Role detector

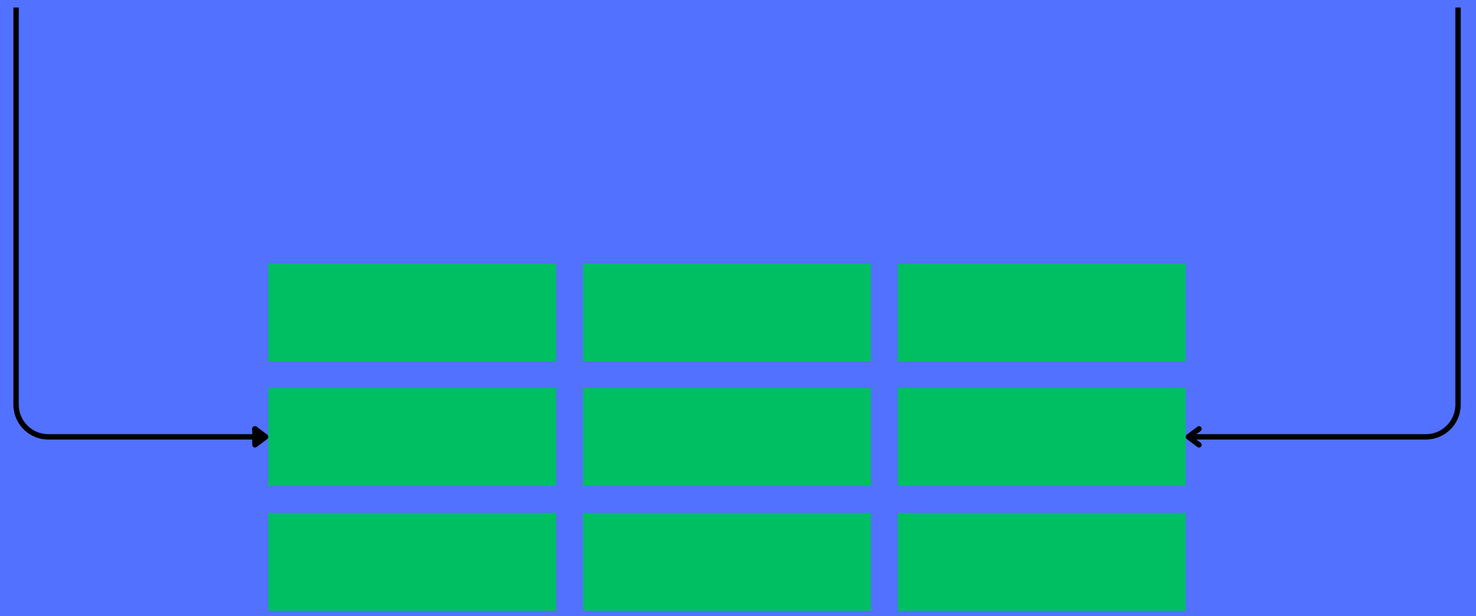
Predicting players' role based on their actions via unsupervised techniques.



Dataset

wyscout

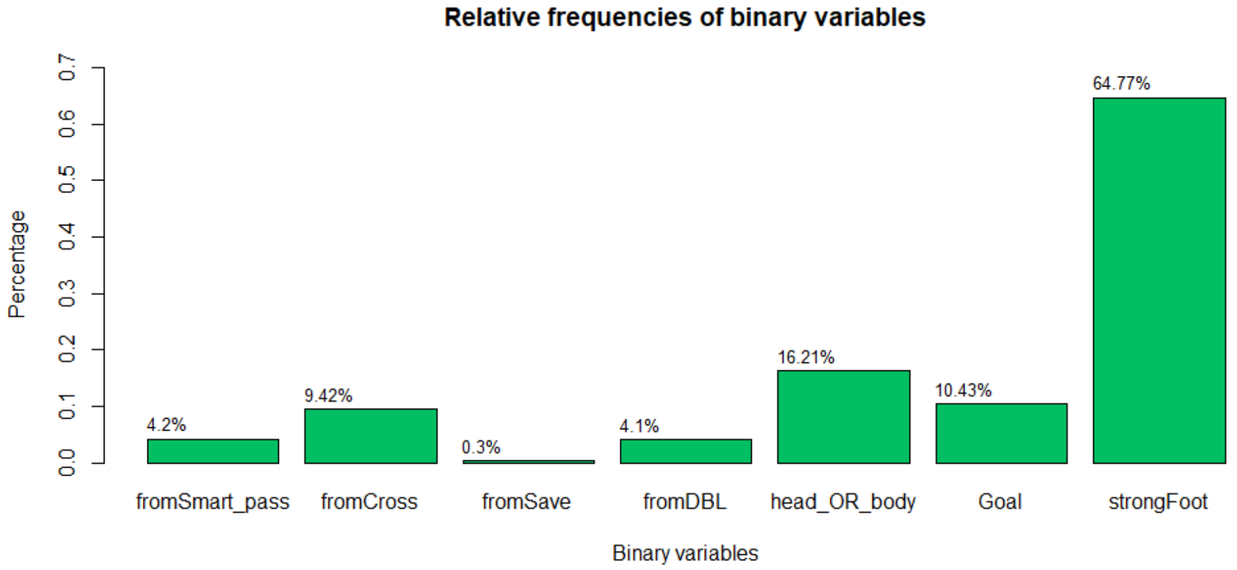
FBREF

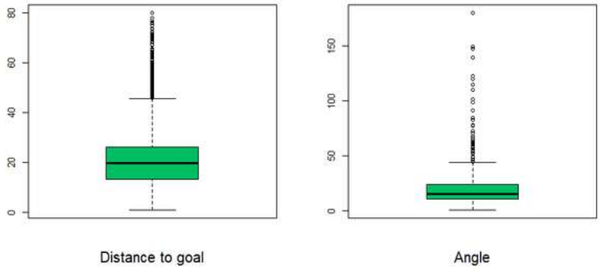


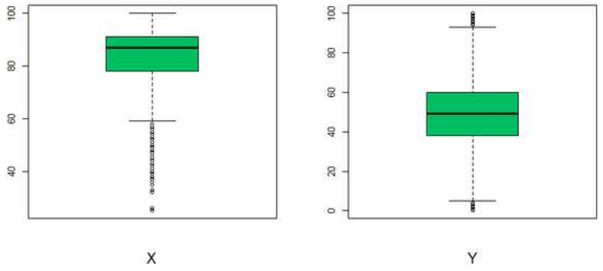
xG model

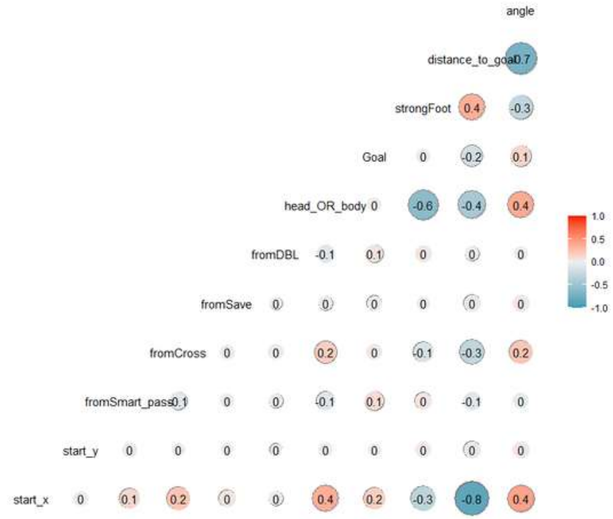
Descriptive

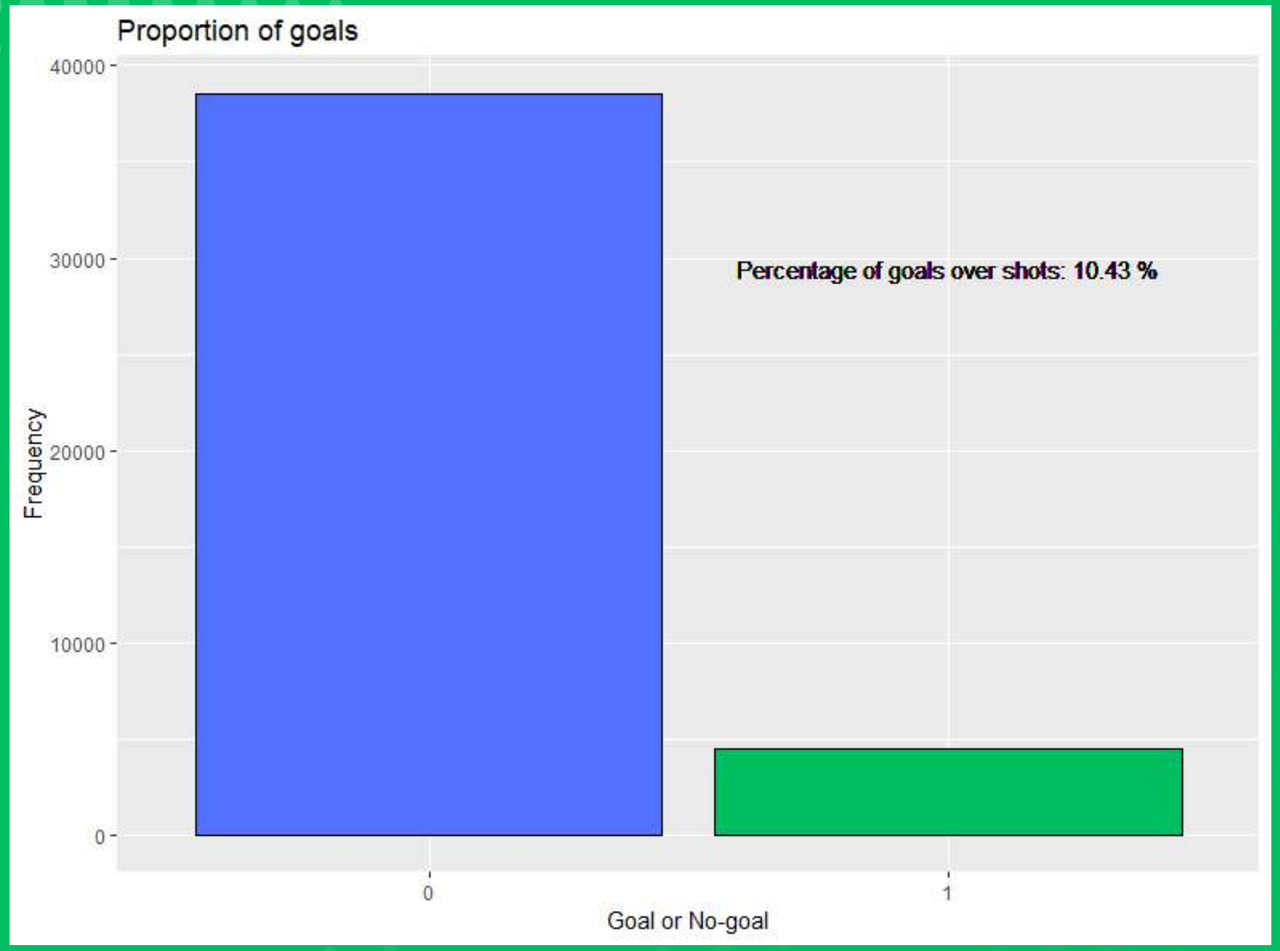
Boxplots of the continuous variables showed an anomaly in shots X coordinate, thus the shots with X smaller than 20 (i.e. made approximately inside their own penalty box) were dropped.










Proportion of goals
40000
30000
20000
10000
0
Frequency
Percentage of goals over shots: 10.43 %
0
1
Goal or No-goal

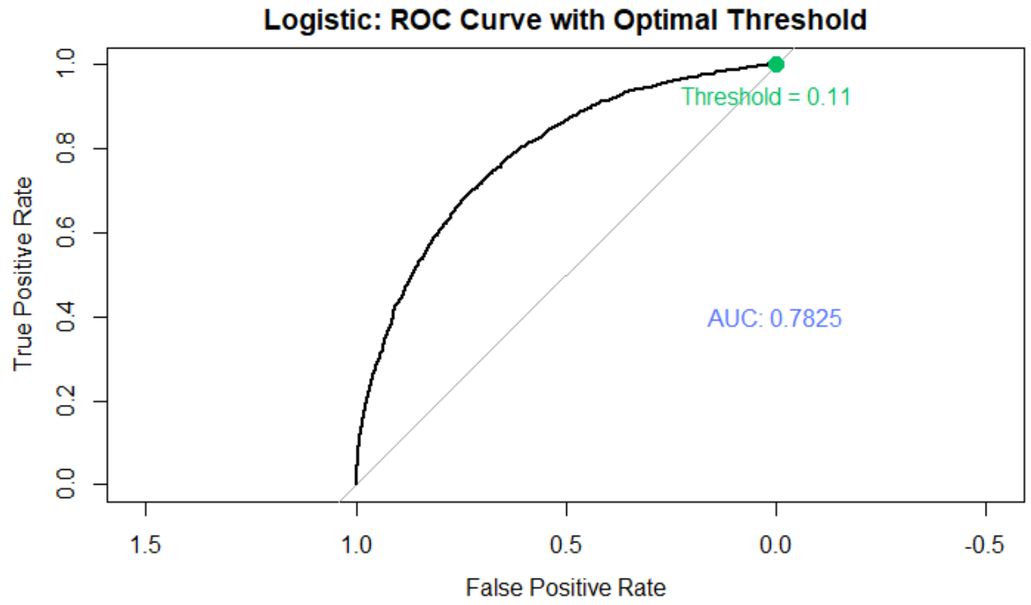
Logistic model

Table 2: Logistic Evaluation Metrics

Metric	Accuracy	Precision	Recall	MSE	Deviance
Value	0.725	0.232	0.698	0.082	16921

Table 1: Logistic Regression Coefficients

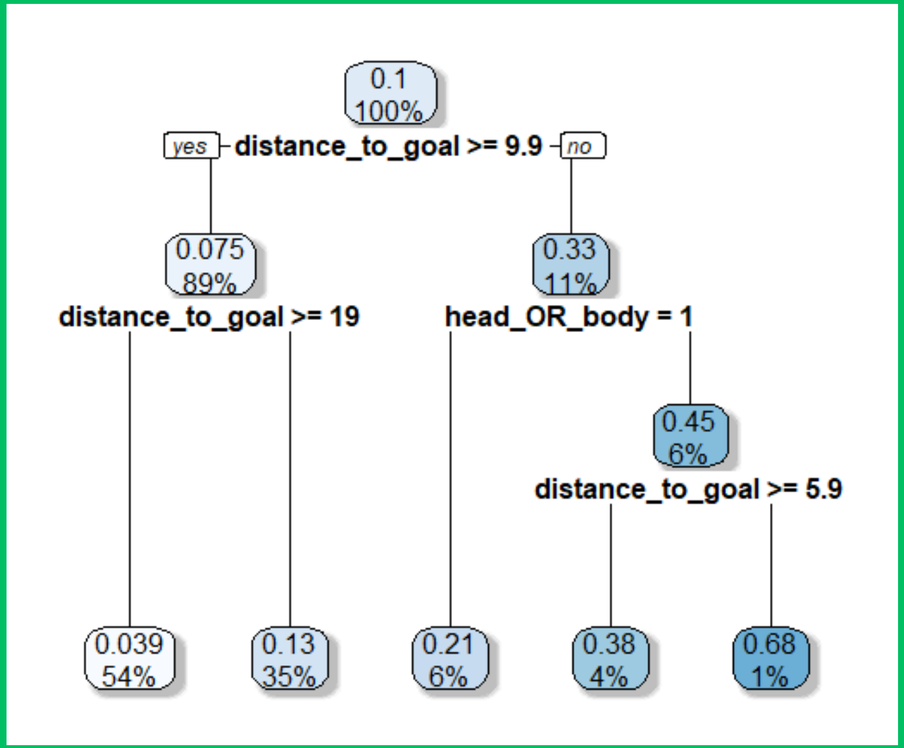
Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.217	0.126	-1.730	0.084
fromSmart_pass	0.557	0.079	7.011	< 0.001 ***
fromCross	-0.003	0.058	-0.049	0.961
fromSave	0.520	0.256	2.030	0.042 *
fromDBL	0.885	0.080	11.040	< 0.001 ***
head_OR_body	-0.824	0.067	-12.311	< 0.001 ***
strongFoot	0.140	0.053	2.670	0.008 **
distance_to_goal	-0.134	0.005	-26.358	< 0.001 ***
angle	0.013	0.002	7.344	< 0.001 ***
Null deviance: 20083 on 30137 degrees of freedom				
Residual deviance: 16921 on 30129 degrees of freedom				
AIC: 16939				

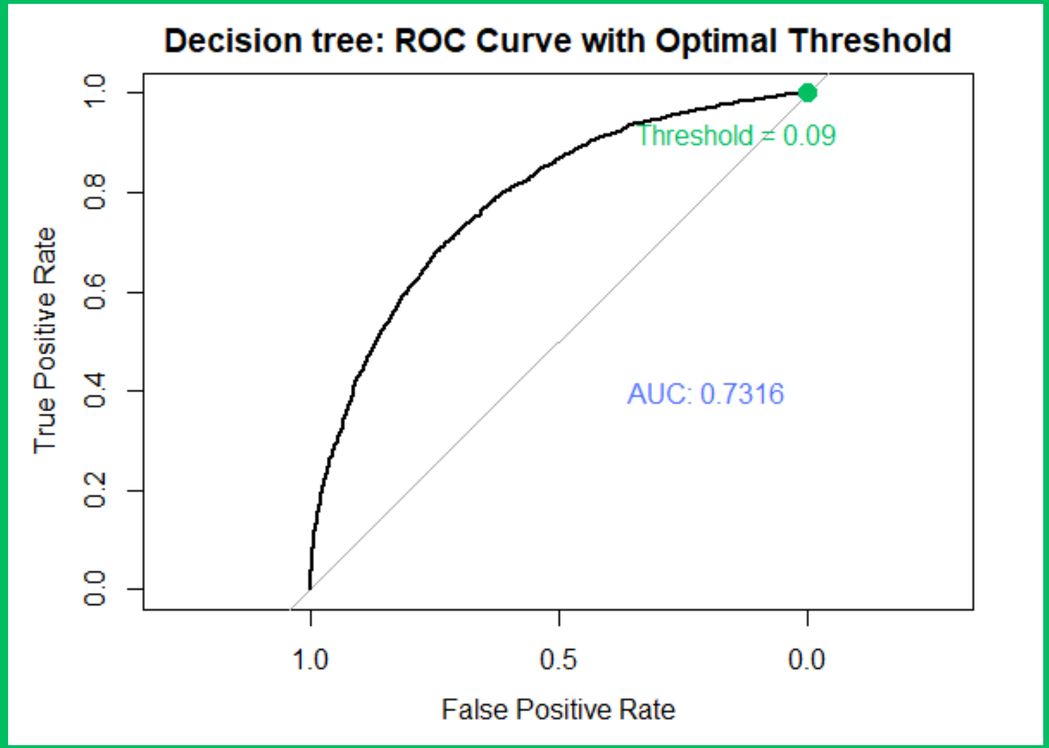


Decision tree

Table 4: Decision tree Evaluation Metrics

Metric	Accuracy	Precision	Recall	MSE
Value	0.604	0.183	0.792	0.085

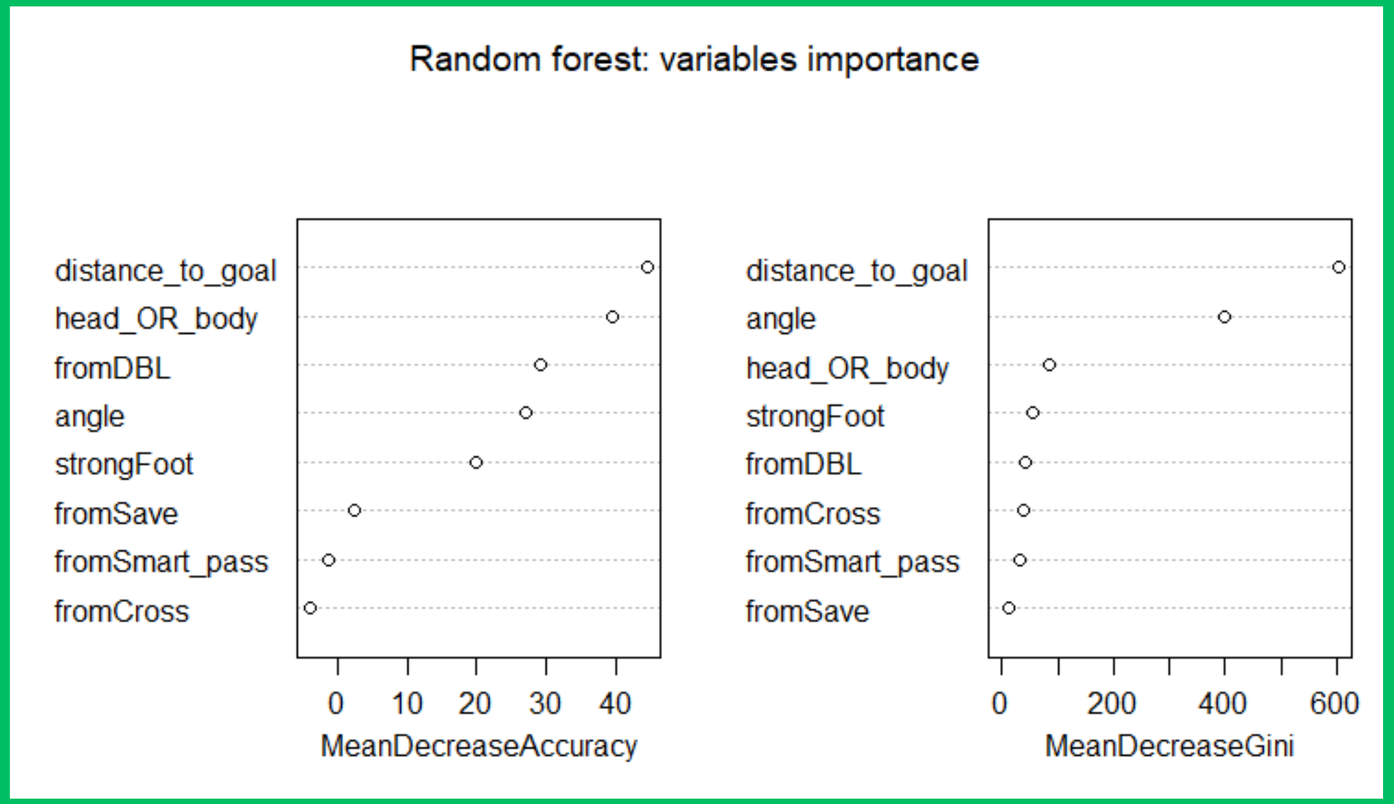


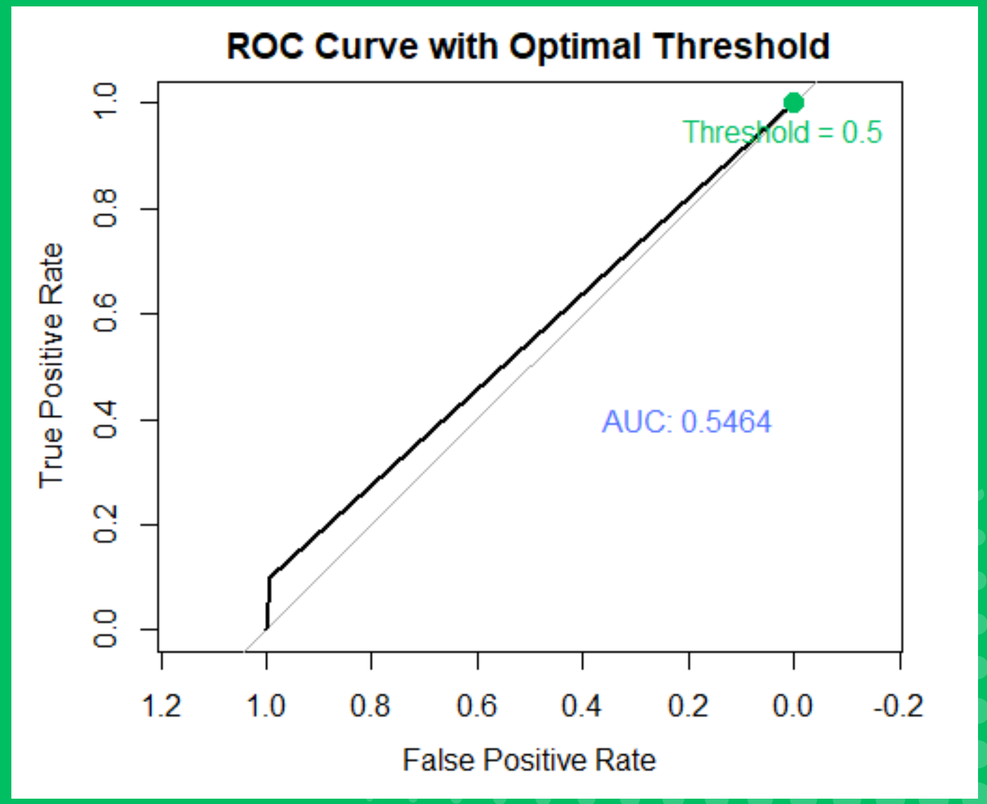


Random forest

Table 6: Random forest Evaluation Metrics

Metric	Accuracy	Precision	Recall	MSE
Value	0.898	0.603	0.100	0.102

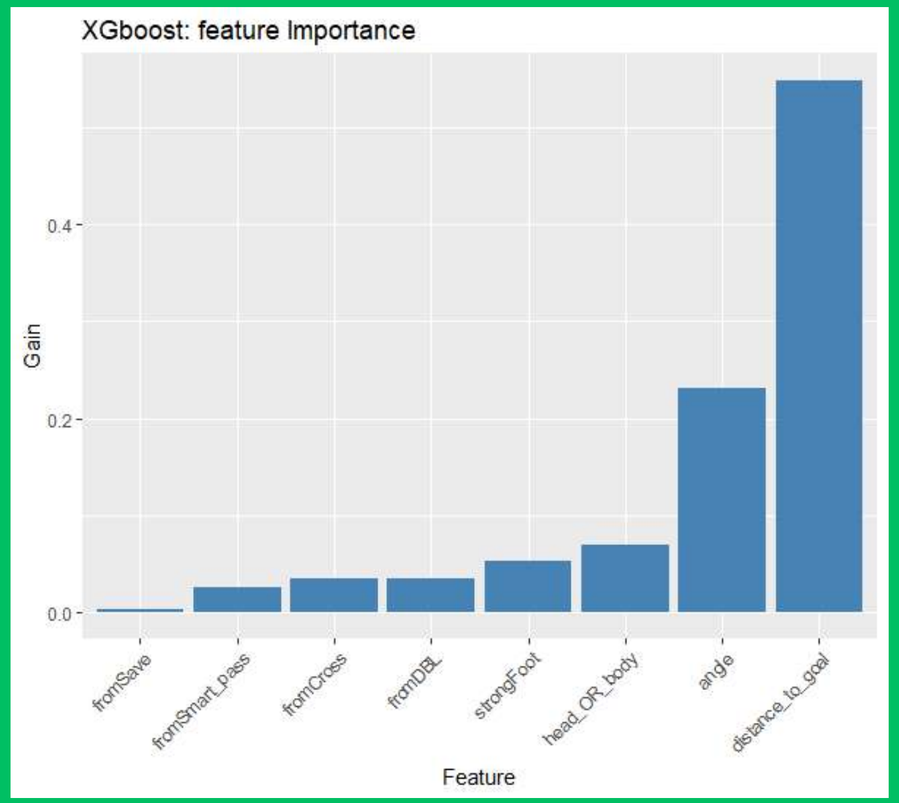


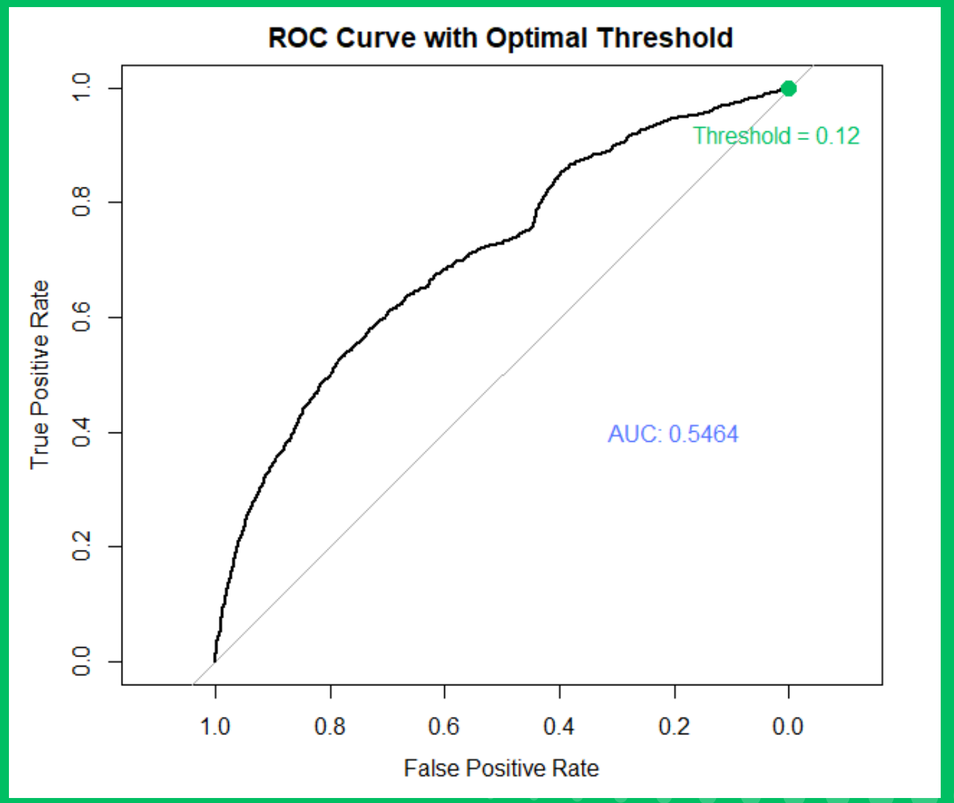


XGBoost

Table 8: XGBoost Evaluation Metrics

Metric	Accuracy	Precision	Recall	MSE
Value	0.754	0.222	0.532	0.095



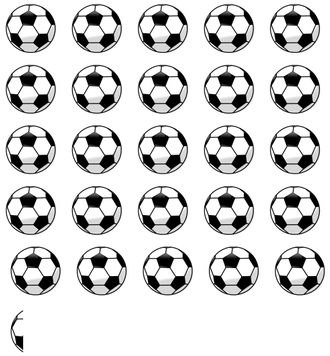


BEST xSCORERS

The top three of 2017/2018 for xG computed by the logistic model.

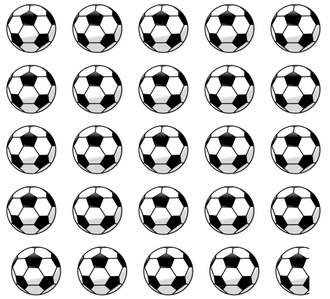
Cristiano Ronaldo

25.2 xG



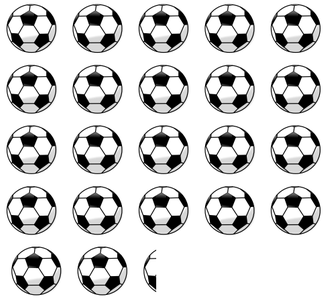
Harry Kane

24.7 xG



Robert Lewandowski

22.2 xG

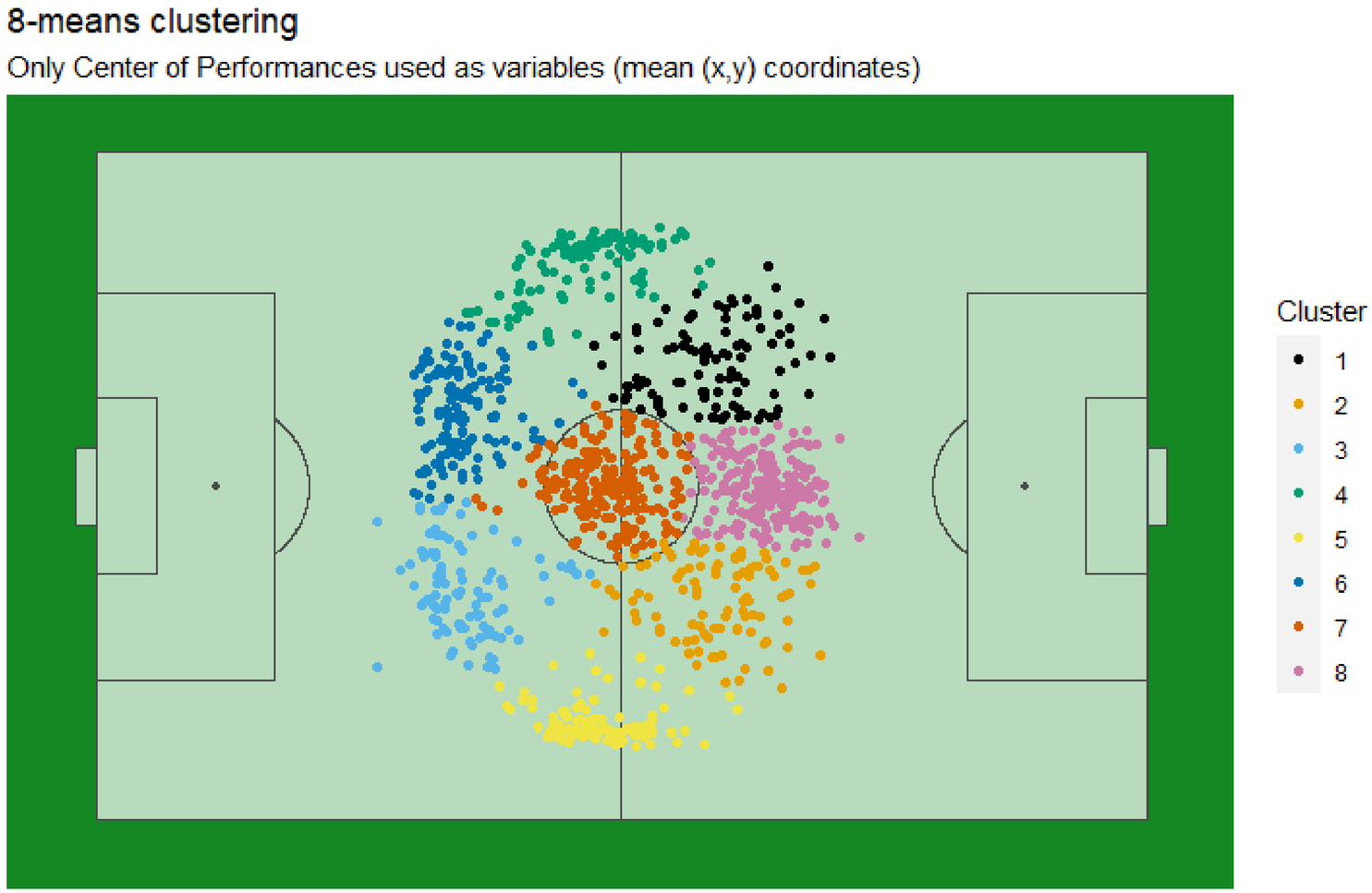


Role detector

Literature baseline

Cintia and Pappalardo's work replication.

Data come from: Pappalardo, L., Cintia, P., Rossi, A. et al. they were subset taking Italy, Spain and England first series events and joined with aggregated performance metrics by FBref.com.
The resulting dataset is a 1056x29 dataframe.



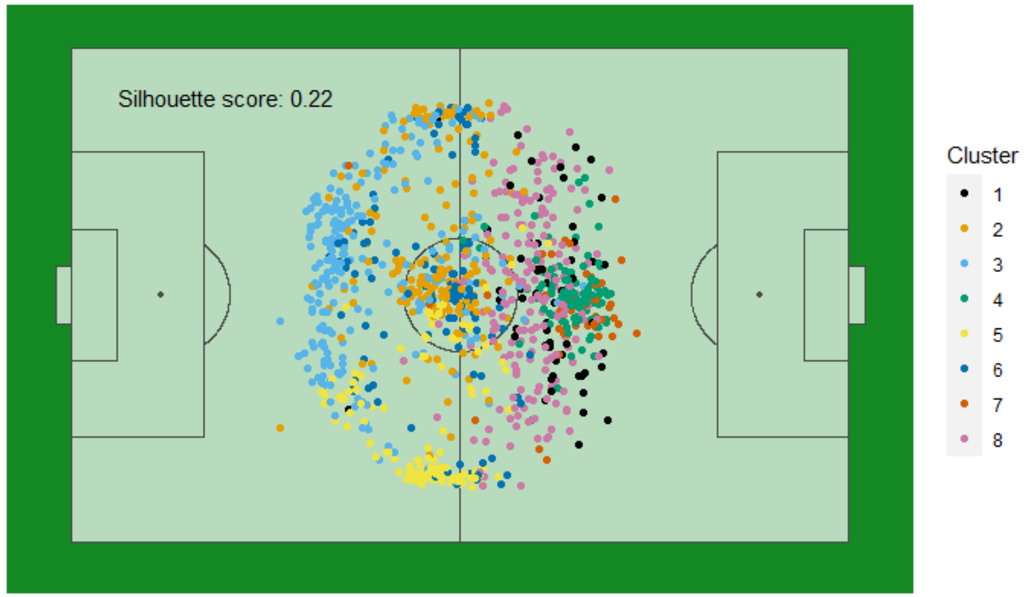
K-means

The first clustering approach is an 8-means over the whole dataset with poor results.

The second step is the hyperparameter tuning that results in K = 5

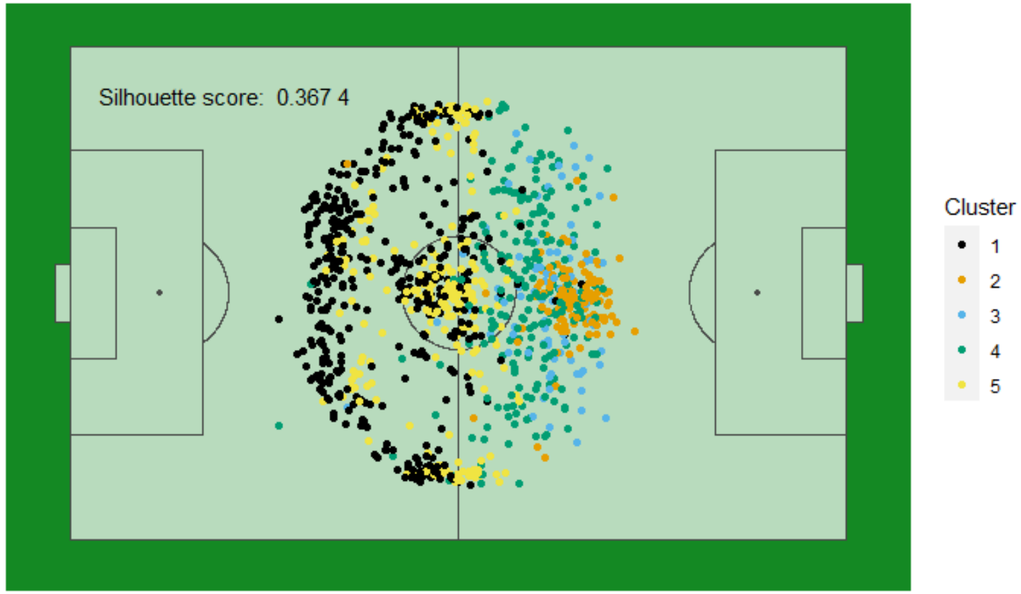
8-means clustering

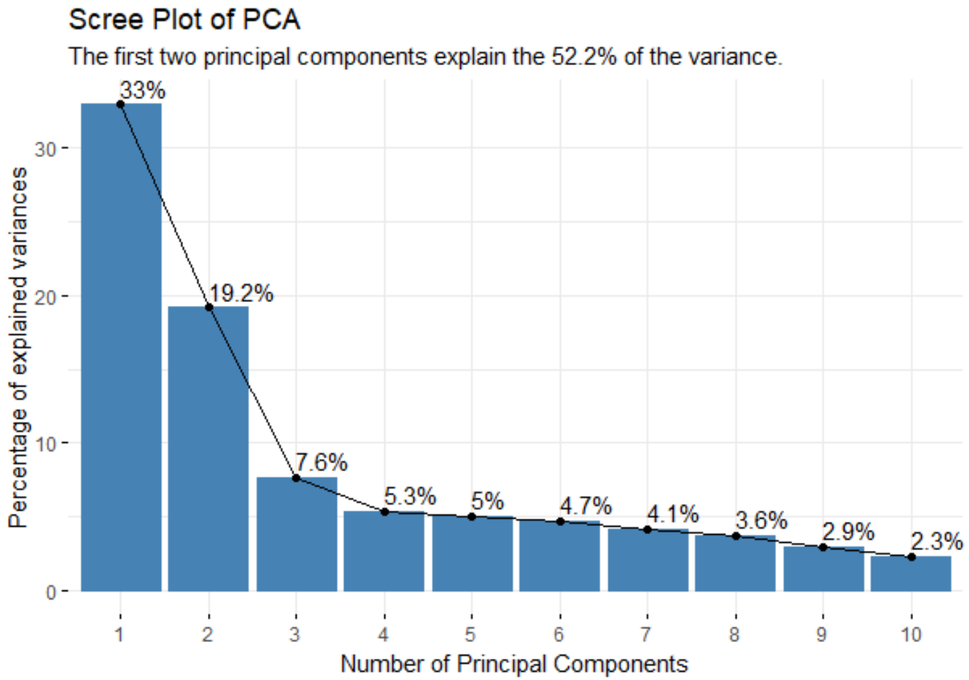
Using all the performance variables, the positional representation becomes messier.



5-means clustering

Using K = 5, silhouette score improves but not the interpretation

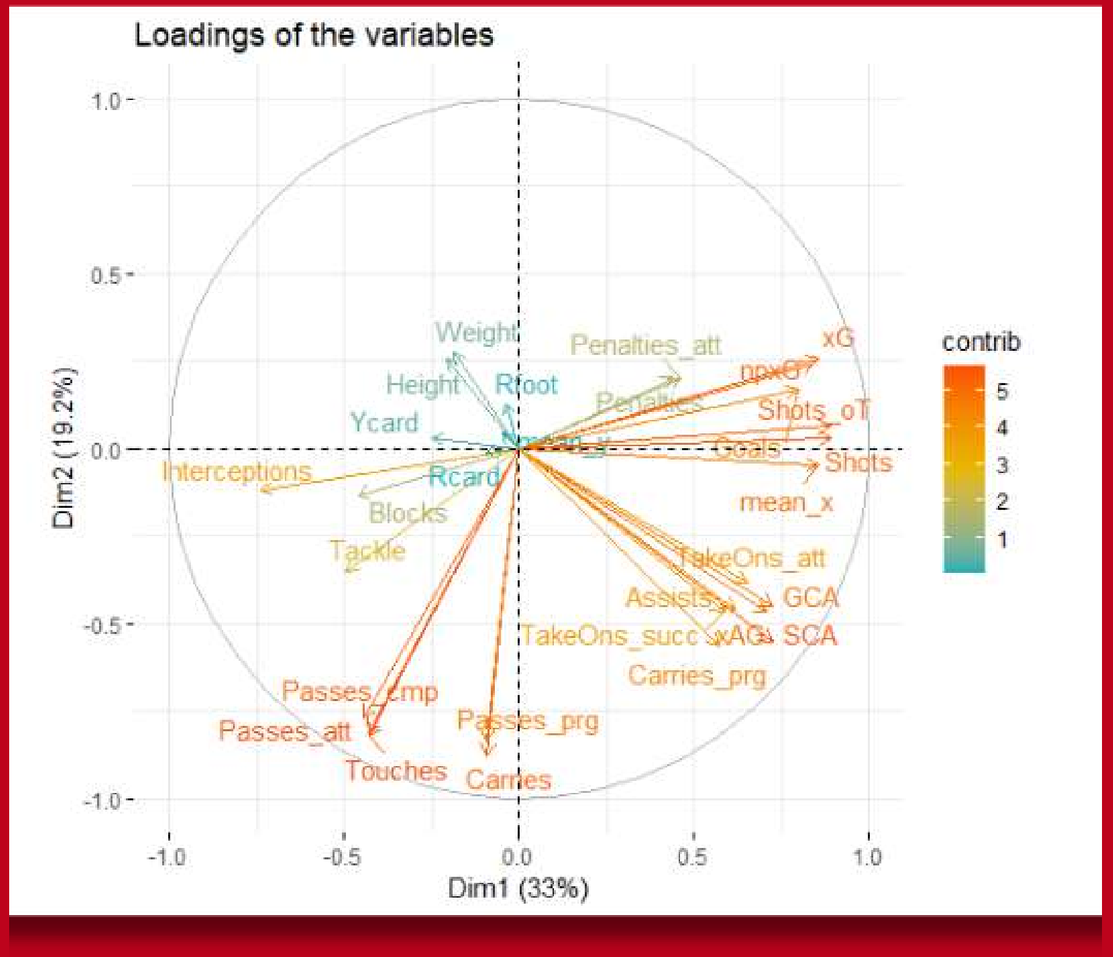


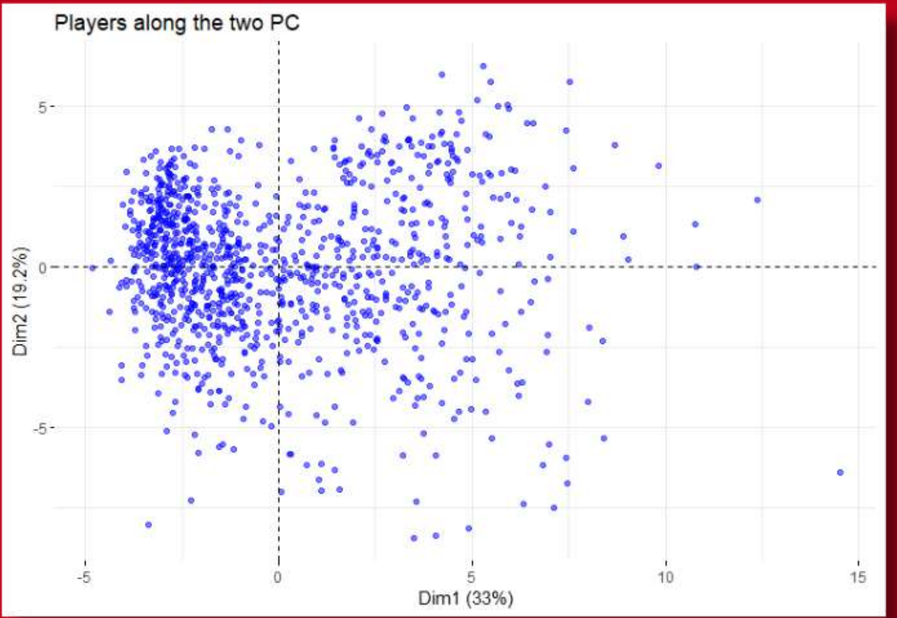


PCA

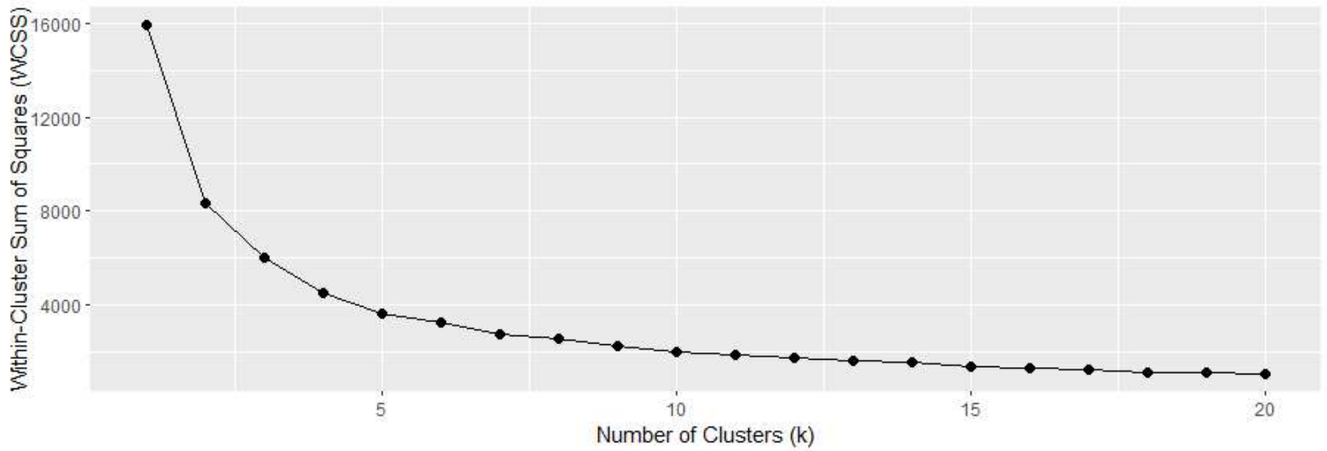
Dimensionality reduction is performed in order to **visualize roles' features** and extract **insights** from the variables.

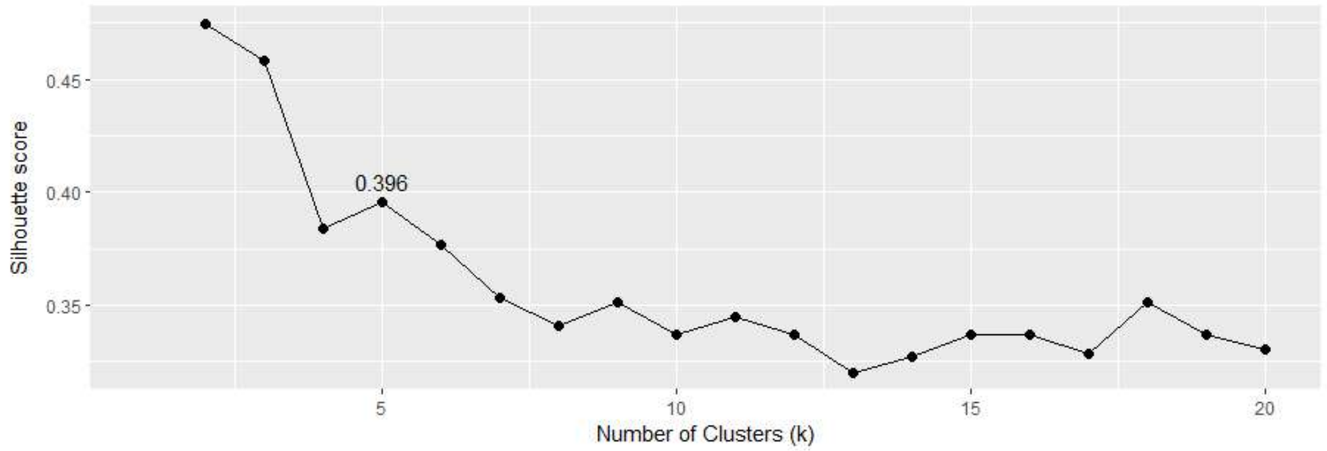
Two components are chosen, explaining the 52.2 % of the variance in the data.


Loadings of the variables
Dim2 (19.2%)
Dim1 (33%)
contrib
5
4
3
2
1
Weight
Penalties_att
xG
Height
Rfoot
npxG
Ycard
Penalties
Shots_oT
Goals
Shots
Interceptions
Rcard
mean_x
Blocks
Tackle
TakeOns_att
Assists
GCA
TakeOns_succ
xAG
SCA
Carries_prg
Passes_cmp
Passes_prg
Passes_att
Touches
Carries


Players along the two PC
Dim2 (19.2%)
Dim1 (33%)

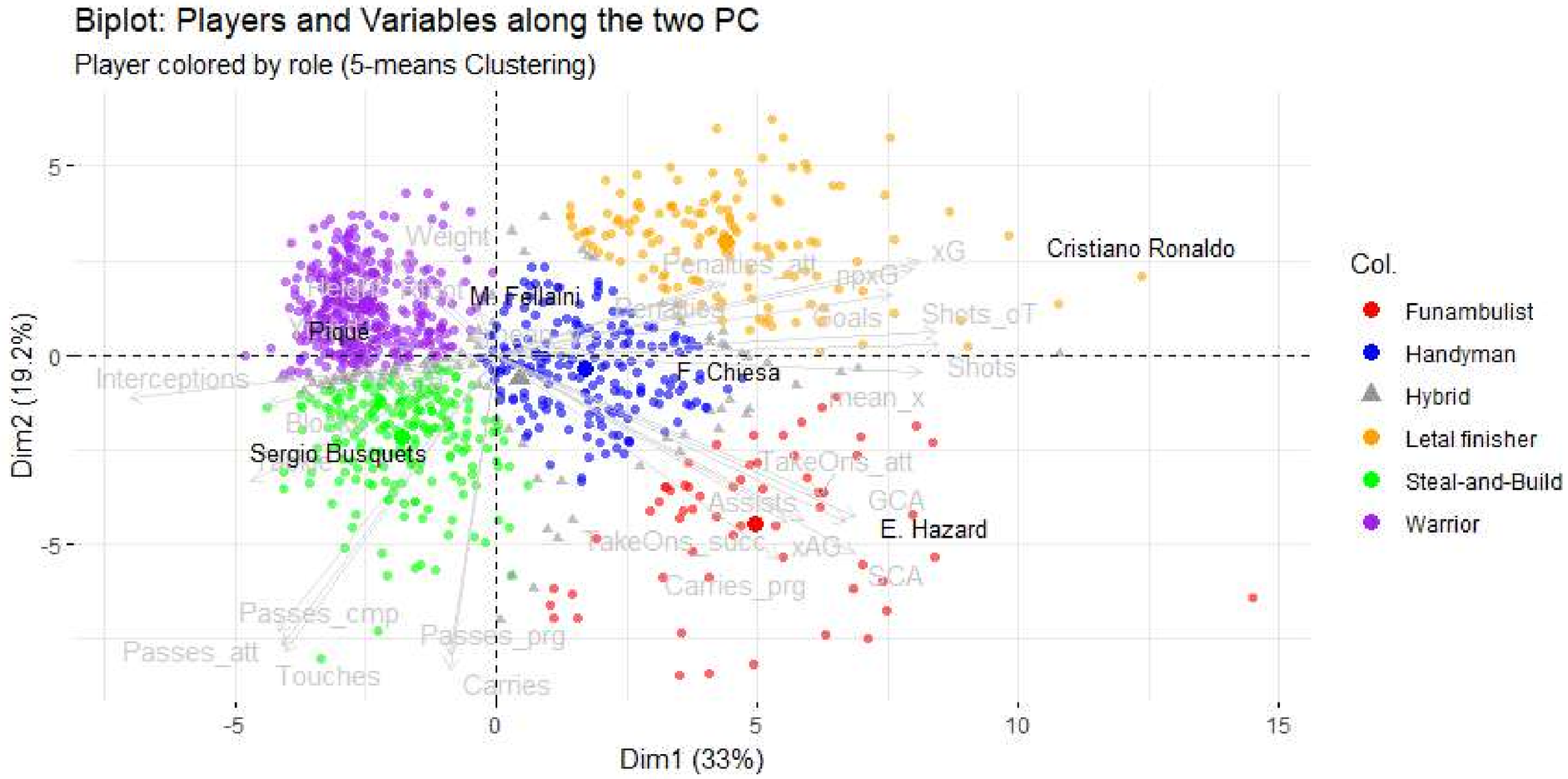
Kmeans clustering across projections onto PCs





*When dropping the binary variable *Rfoot* the best silhouette score is 0.373. Cluster assignment stays exactly the same for K=5.

(5+1)means clustering across projections onto PCs

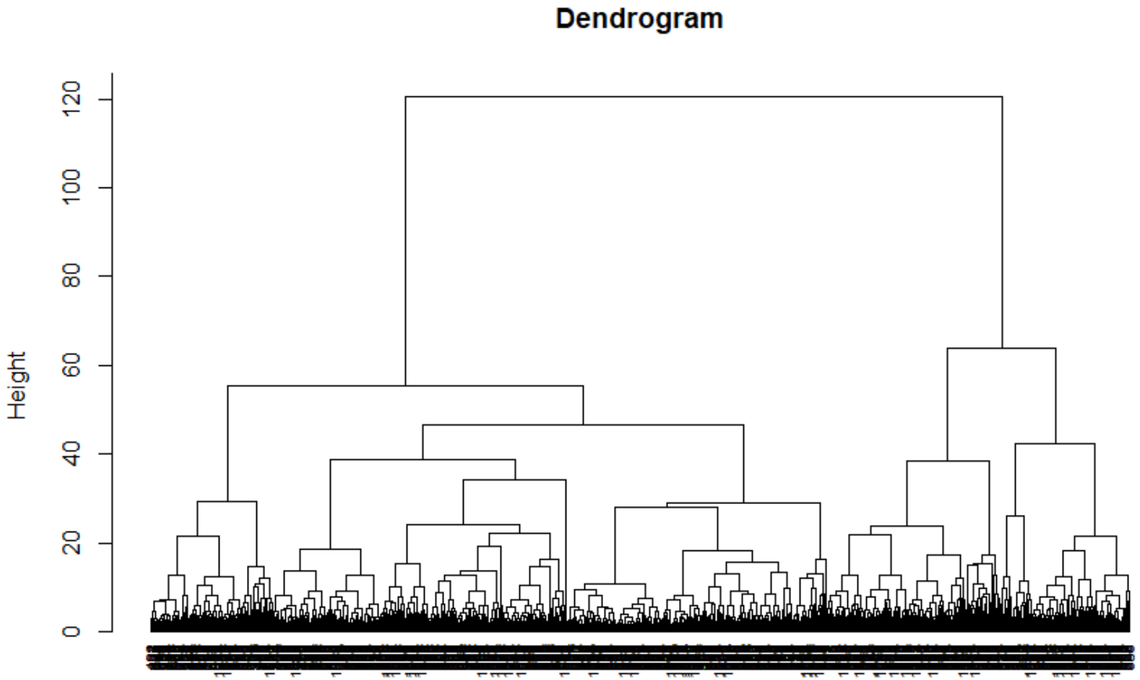


Hierarchical clustering

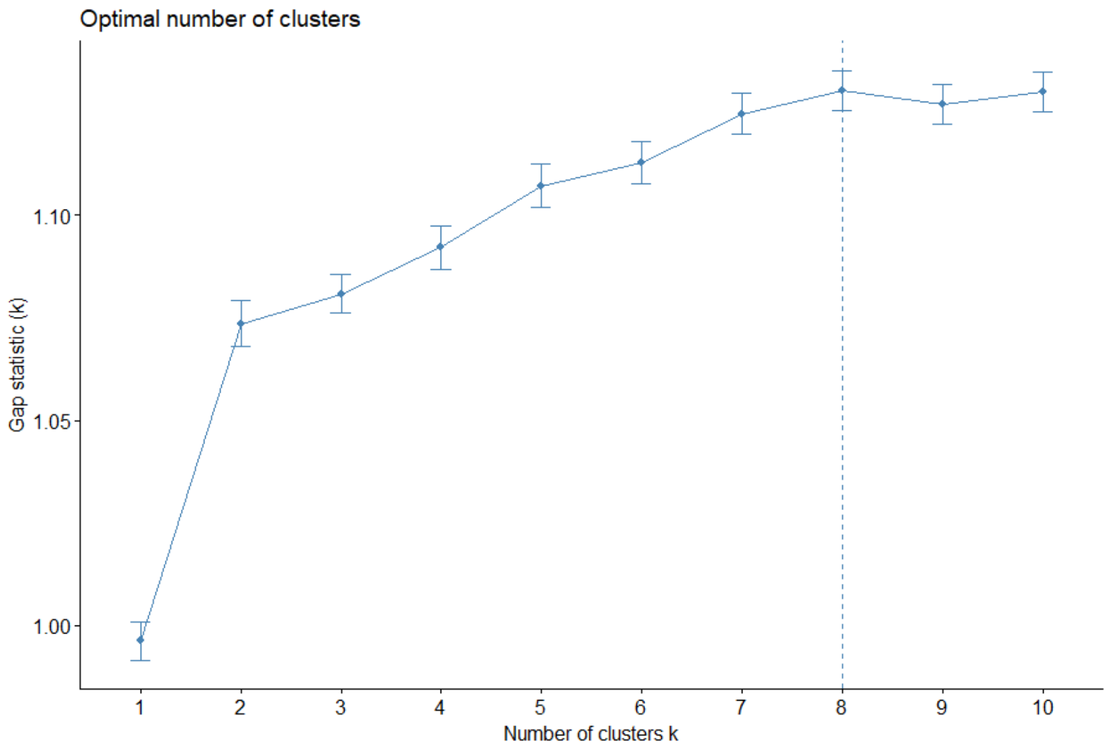
The third approach tries to explore a different clustering technique to spot new kinds of relationships using **different distance functions**.

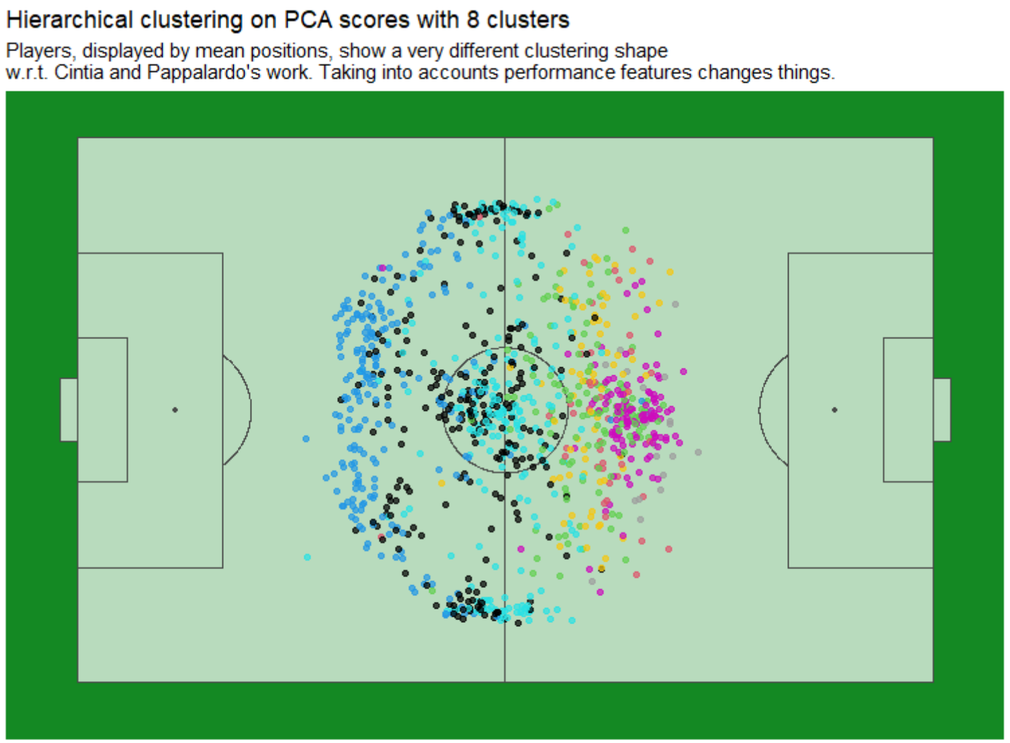
In this case **Ward's linkage** was the best to optimize Agglomerative coefficient. It minimizes the squared deviation from the centroid.

Linkage type	Average	Single	Complete	Ward
Agglomerative coefficient	0.884	0.882	0.904	**0.976**



Hierarchical clustering with Ward's linkage





Selection of K via Gap statistic

Tibshirani, R., Walther, G. and Hastie, T. (2001), Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63: 411-423. https://doi.org/10.1111/1467-9868.00293

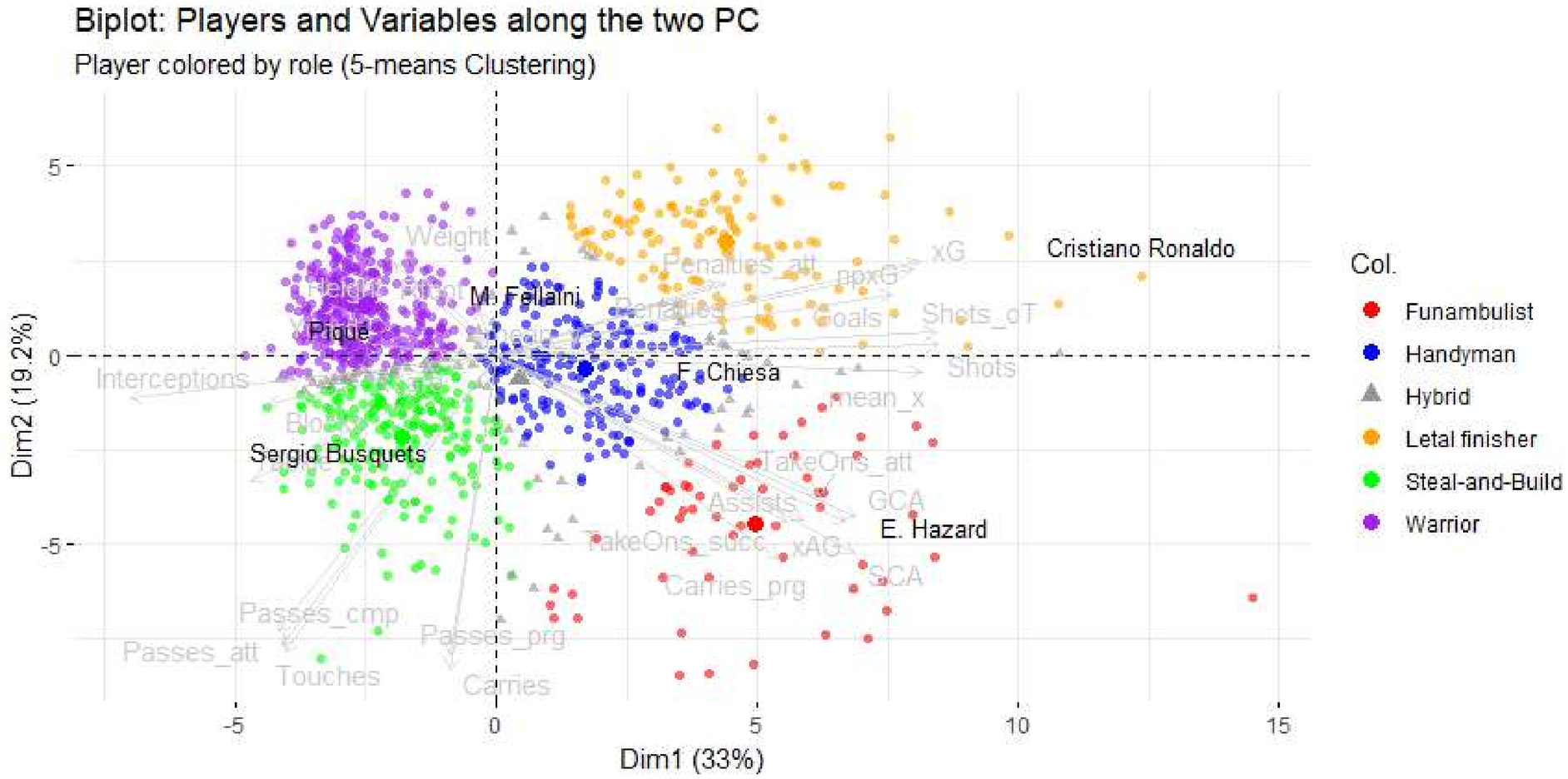
K = 8 Hierarchical clustering

5means on PCs

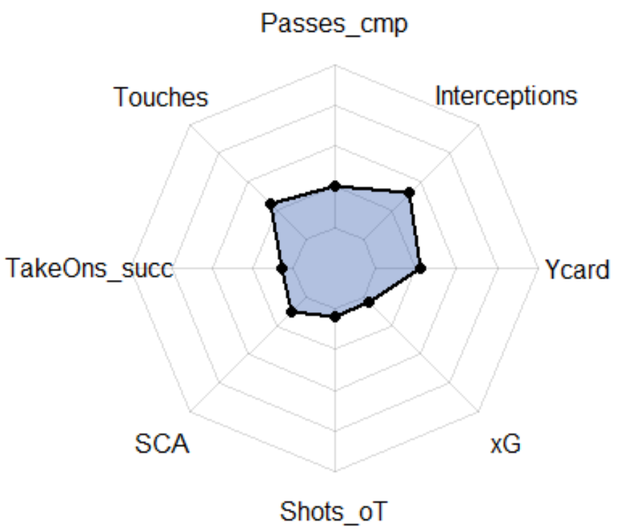
	1	2	3	4	5	6	7	8
1	0	0	22	2	33	0	100	1
2	0	1	98	97	2	2	3	0
3	104	48	0	54	0	45	0	0
4	9	0	29	0	0	0	1	36
5	18	67	0	48	0	234	2	0

Comparing Kmeans and a Hierarchical clustering

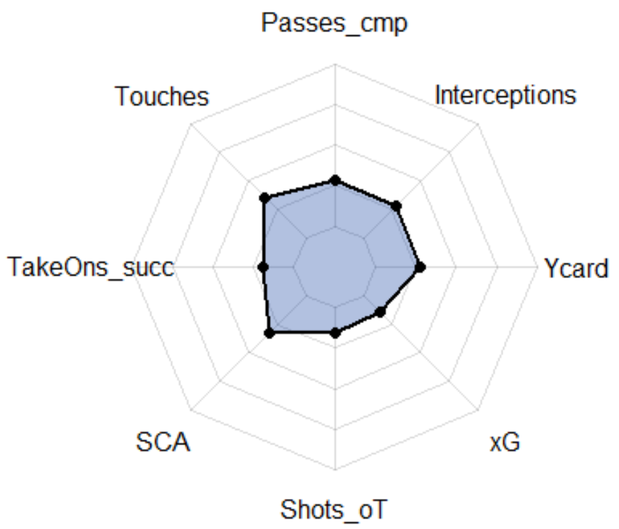
(5+1)means clustering across projections onto PCs



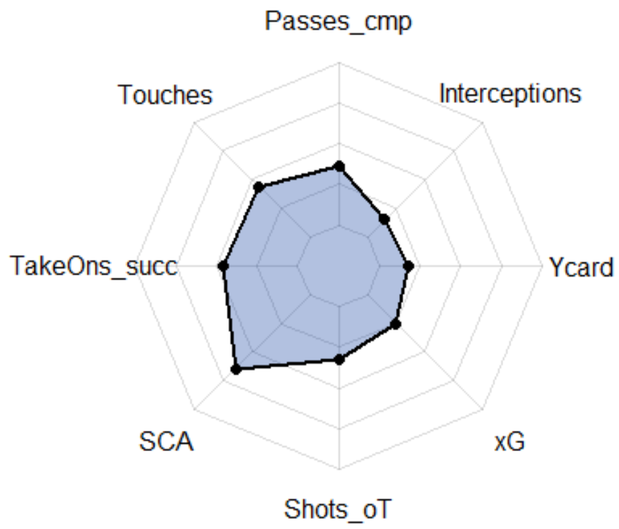
Warrior



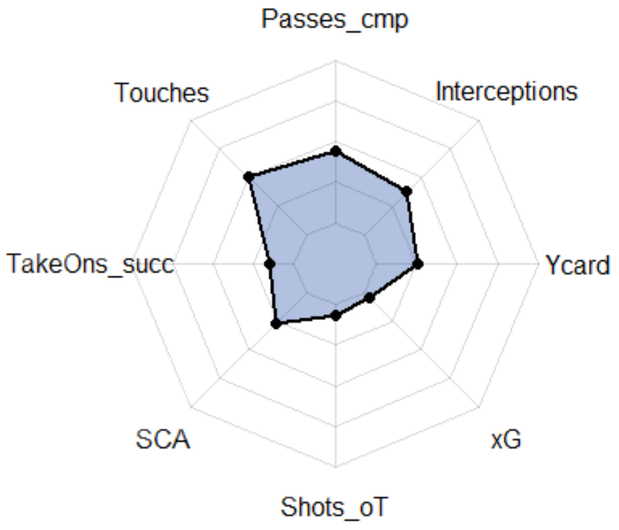
Hybrid



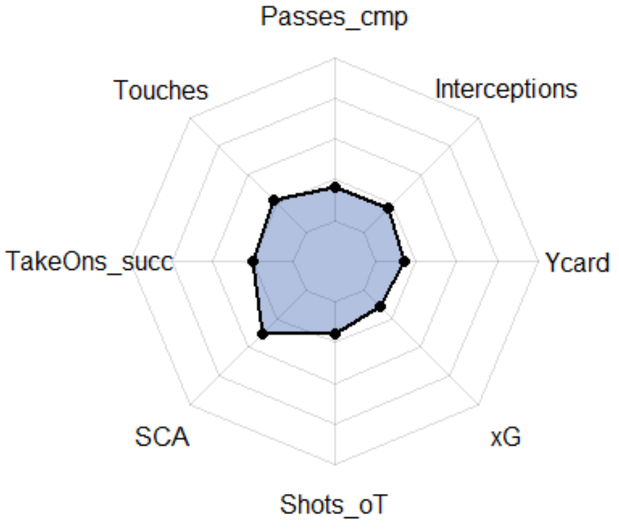
Funambulist



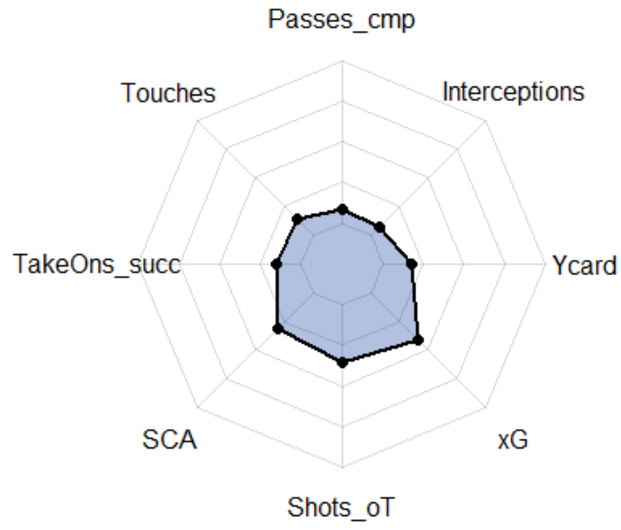
Steal-and-Build



Handyman

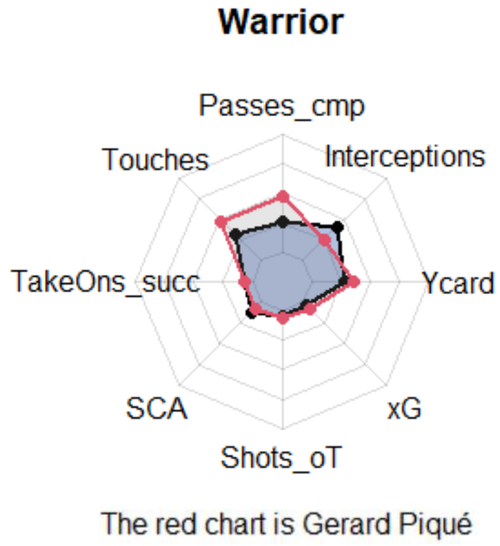


Fatal finisher

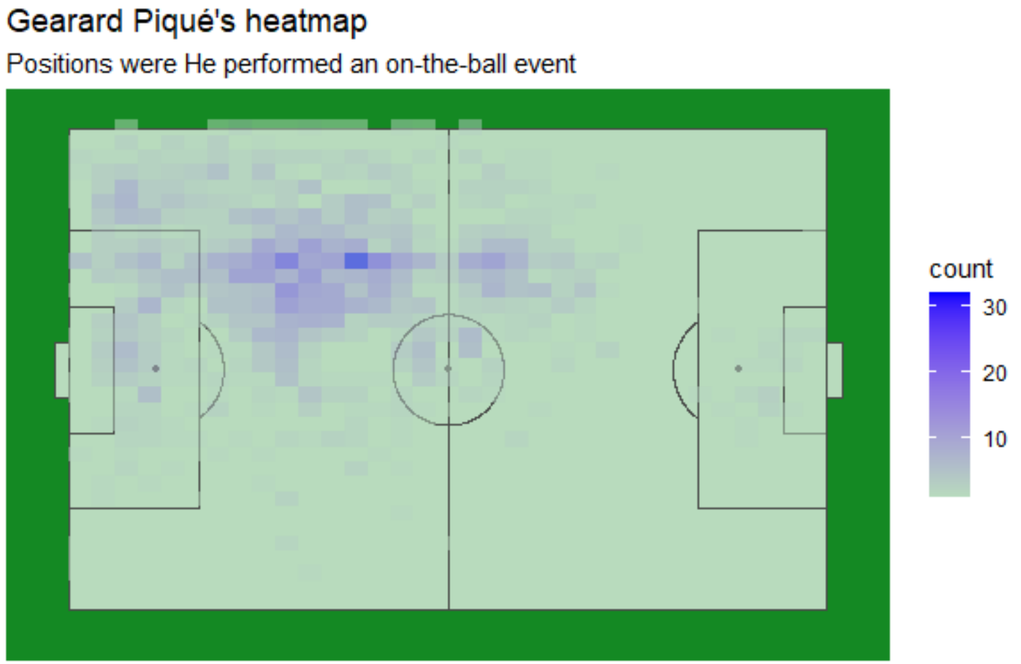


Warrior

Gerard Piqué

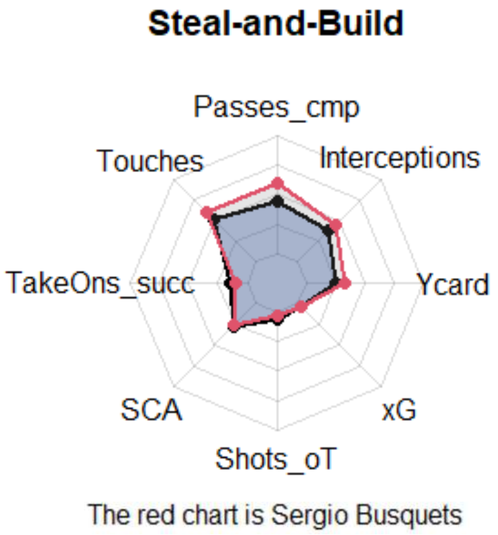


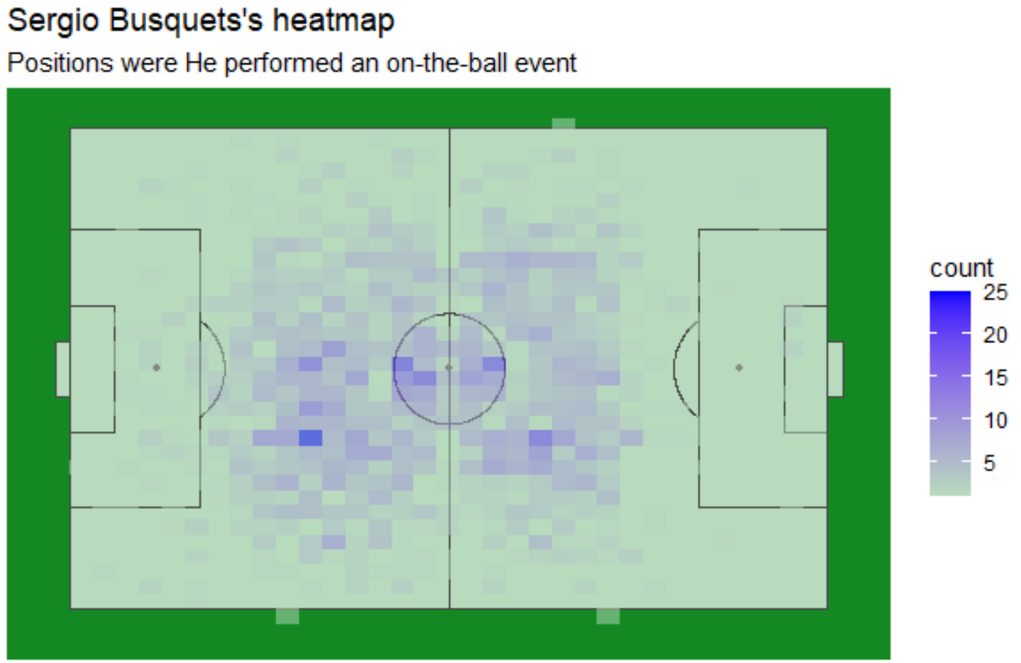
The red chart is Gerard Piqué



Steal-and-Build

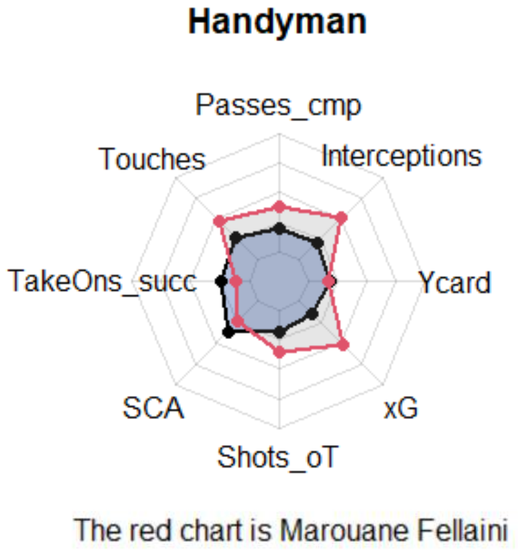
Sergio Busquets

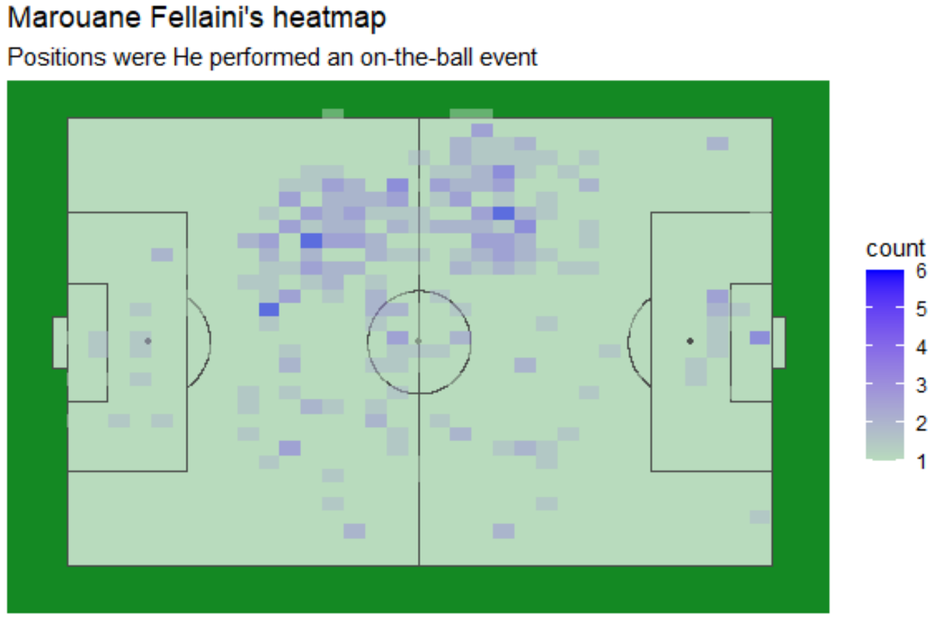




Handyman

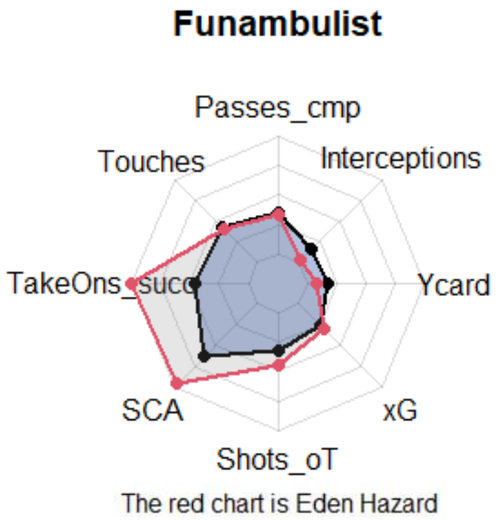
Marouane Fellaini

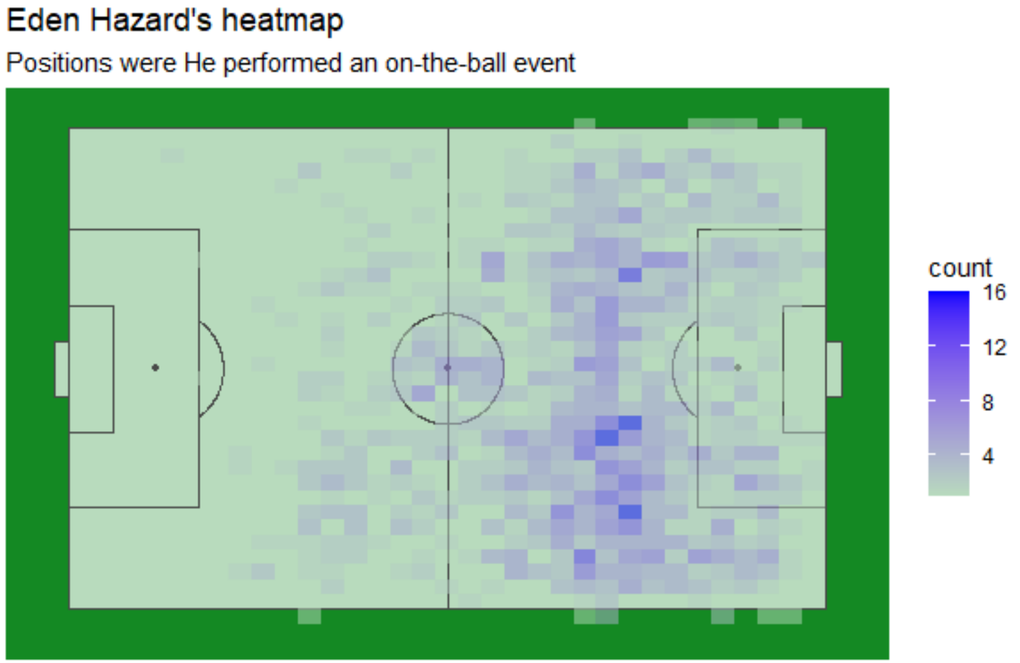




Funambulist

Eden Hazard

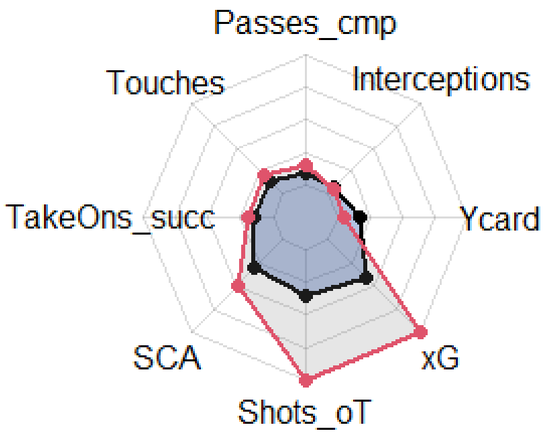




Fatal finisher

Cristiano Ronaldo

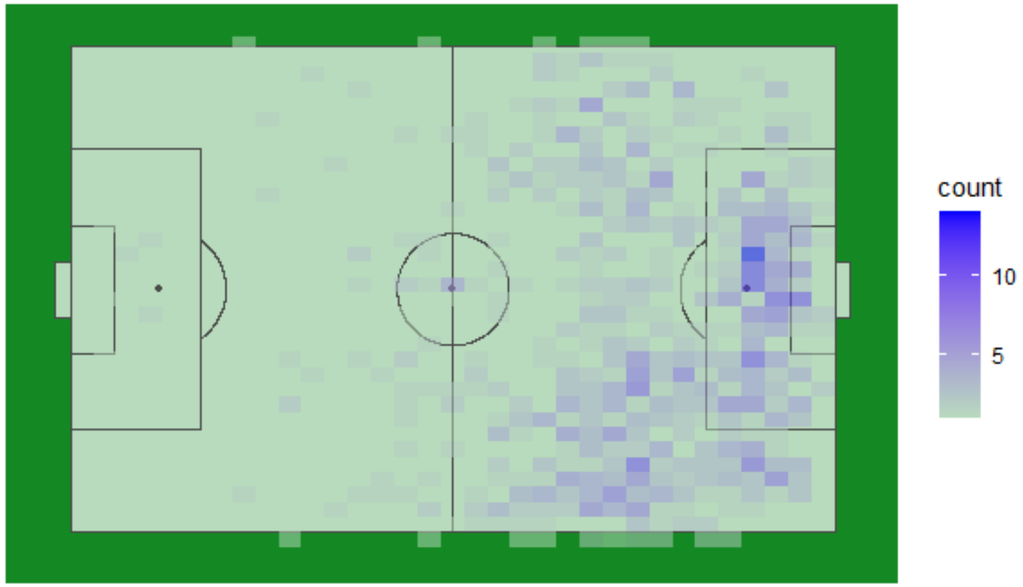
Letal finisher



The red chart is Cristiano Ronaldo

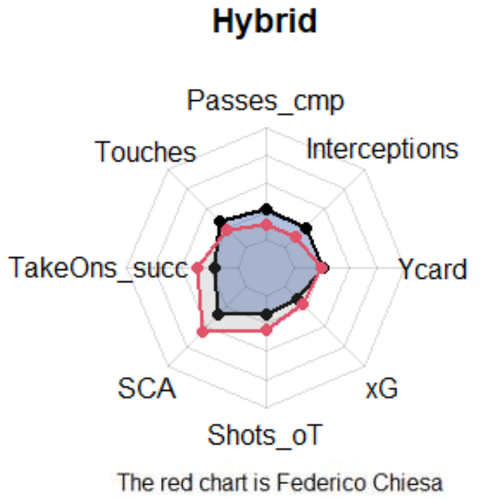
Cristiano Ronaldo's heatmap

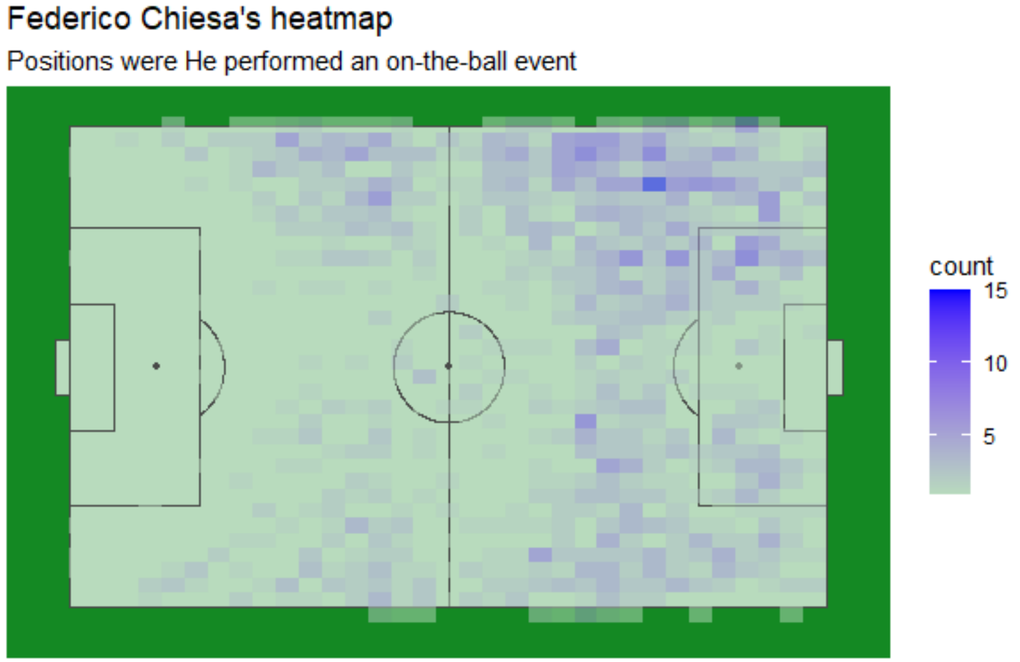
Positions were He performed an on-the-ball event



Hybrid

Federico Chiesa





Thank you for listening!