# The Beautiful Game

## Comprehending the Effect of Style of Play on Team Performance in the Top Three Leagues


**Bruno Fernandes** m20200111 | **Frederico Rodrigues** m20200583 | **Miguel Zina** m20200582 | **Pedro Fonseca** m20201037


## 01-Understanding the Problem

From Catenaccio to Gegenpress and Tiki-Taka, **football tactics** have been in constant evolution since its beginning. This evolution is promoted by the constant pursuit of **outperforming the competiton,** leading to victories and titles.

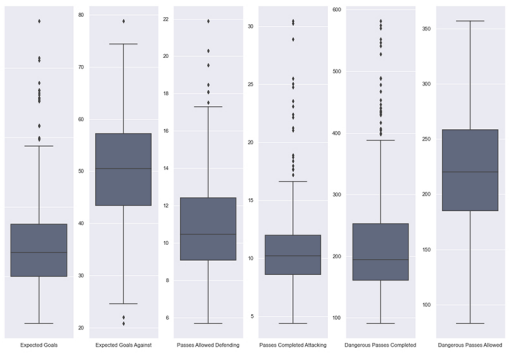
The objective of this project is to understand how **the style of play** of teams in the best leagues in the world affects, not only, their performance but also the amount of points obtained in each season.

For this analysis, data from **2014 to 2020** was gathered and processed.

**Data Description:**
- 3 Leagues
- 6 Season
- 24 Variables
- 84 Teams

**Techniques:**
- Decision Tree
- K-means
- Correspondence Analysis
- Pearson Correlation
- Hierarchical Clustering
- Significance Tests ( T-test & F-Test )
- Breusch-Pagan Test
- Multicollinearity Test
- Linear Regression

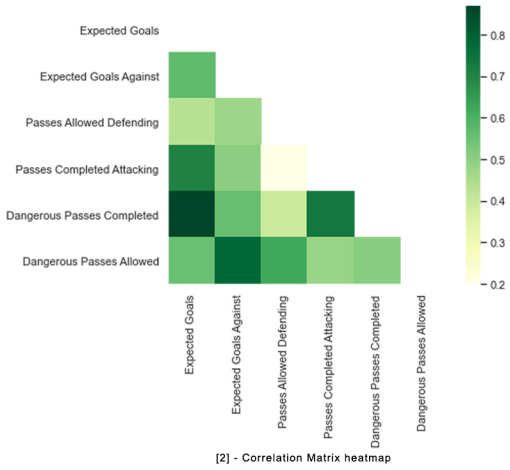
## 02-Data Preparation



[1] - Box plot with variable distributions

For variable interpretation, box plot visualizations were used, given that they handle large amount of observations easily and give a clear **summary of the data.**
The presence of **outliers is notable**, which can be understood as a great predominance of a team over the other.

As expected, a severe correlation in the **possession and exchange of passes** with the expected goals, both in attacking and defensive movements, was found. The **Pearson Correlation Matrix** emphasizes the **linkage** between the variables Dangerous Passes Completed and Expected Goals.




[2] - Correlation Matrix heatmap

## 03-Variable Analysis

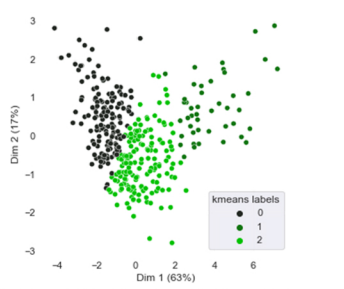
| | Coefficient | Std. Error | p-value |
|---|---|---|---|
| Expected Goals | 0.9253 | 0.062 | 0.000 |
| Expected Goals Against | -0.4197 | 0.069 | 0.000 |
| Passes Allowed Defending | 0.4203 | 0.235 | 0.075 |
| Passes Completed Attacking | 0.4045 | 0.204 | 0.048 |
| Dangerous Passes Completed | 0.0129 | 0.011 | 0.242 |
| Dangerous Passes Allowed | 0.0583 | 0.017 | 0.001 |

[3] - Linear Regression

Having now a more stable dataset and with the final variables defined and normalized, it is now time to analyse the results of the application of the linear regression. From this table we can observe the **coefficients**, the **standard errors** and the **p-values** (measuring how much evidence we have against the Null Hypothesis). For the p-values that are highlighted we fail to reject the Null Hypothesis, using a **significance level of 5%**.

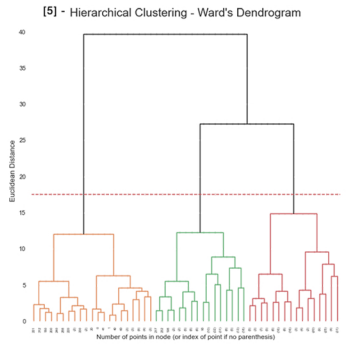
## 04-Clustering

The objective consisted in characterizing the dataset according to the variables that infuence the **style of play** of each team. A **cluster analysis** followed, using the variables that we concluded to be significant.
The Dendogram presented below results from a **hierarchical clustering** with the ward linkage, since it was the one with the best results. Finally, the PCA results below were obtained with the **k-means algorithm.**




[4] - PCA visualization 2 dimensions KMeans (Playing Style clusters)

[5] - Hierarchical Clustering - Ward's Dendrogram




## 05-Conclusion

In order to check for the significance of the model and see if the mean values are significant we have to see if the LR is **BLUE.**
Thus we will check for **Linearity, Exogenity, Multicolinearity** and **Homostedacity.**

**We tested this data with two models:**
**Model 1 -** Measures the impact of playing style throughout the season. LR was made with the dummies of the playing style.
**Model 2 -** Measures the impact of changing from the preferred playing style for a specific game.

**Homoscedasticity [Breusch-Pagan]**
**Model 1 -**
H0 : Heteroscedasticity
H1: Reject H0
p-value = 0.017856
We Have Homoscedasticity

**Model 2 -**
H0 : Heteroscedasticity
H1: Reject H0
p-value = 0.000000
We Have Homoscedasticity

**Exogeneity [E(u|X) = 0]**
**Model 1 =** 7.293949563477603e-3
**Model 2 =** 3.3781406790663384e-14

**Linearity**
Both models' parameteres are linear

**Multicolinearity**
**[VIF (Variable Inflation Factors)]**
In neither of the models the VIF's of the parameters in question had a value bigger than 5. Thus, it respect the assumptions

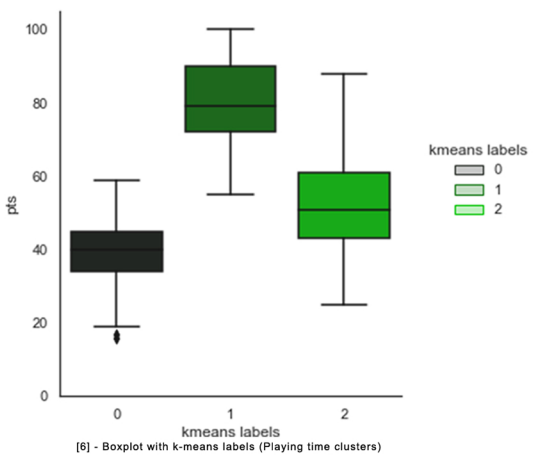
**Conclusion**
Our models are BLUE! This means that we can test the significance of the variables.
In both cases we confirm our initial hypotesis!
Thus we know that the style of play is significant in predicting the final points.
We can also understand that changing playing style in preparation for a game is beneficial for the team in question.

**Cluster 0 -** normally plays with a defensive mindset most of their games.
These teams are characterized by having few to none chances to score, presenting no initiative to control the ball on the opposition half and depending on counter attacks for success. On the other hand, they allow their adversaries to invade dangerous areas often.

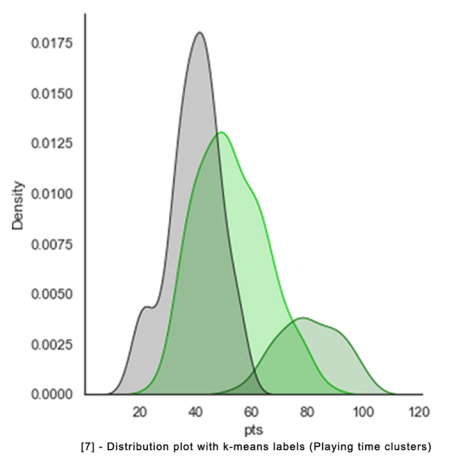
**Cluster 1** - has attacking tactics and dominant performances.
These teams are mostly based in the opposition half for the entire game, where they retain the ball in dangerous areas leading to a greater number of Expected Goals (xG). They don´t allow many opportunities nor do they allow to be controlled.
They are less likely to change their tactics and mindset to best suit their opponent.

**Cluster 2 -** adapts the tactics depending on the game/ opponent.
These teams dispute their games in the middle-field, consequently they don´t create nor allow many chances.
It´s expected that they transform the style of play to be more dominant with inferior teams and more defensive with superior teams.

**Cluster 0 -** **Cluster 1 -** **Cluster 2 -**




[6] - Boxplot with k-means labels (Playing time clusters)




[7] - Distribution plot with k-means labels (Playing time clusters)

**Data:** Football Data: Expected Goals and Other Metrics (Kaggle)