Final Project Report: 2021-2022 European Leagues Player Stats

Anusha Bhat, Nandini Neralagi, Noelani Phillips, Marcos Pi Marrero

---

Dataset Description

The dataset is sourced from the 2021-2022 European Leagues Player Stats and provides a comprehensive overview of 2,921 male soccer players across prominent leagues, including the Premier League, Ligue 1, Bundesliga, Serie A, and La Liga, during the 2021-2022 season. There are a total of 143 variables. These variables encompass player demographic information, match statistics, passing accuracy, offensive and defensive actions, as well as detailed metrics on aerial duels, tackles, dribbles, and other key aspects of individual and team performance in soccer. The dataset provides comprehensive insights for detailed analysis abd a holistic view of player performances.

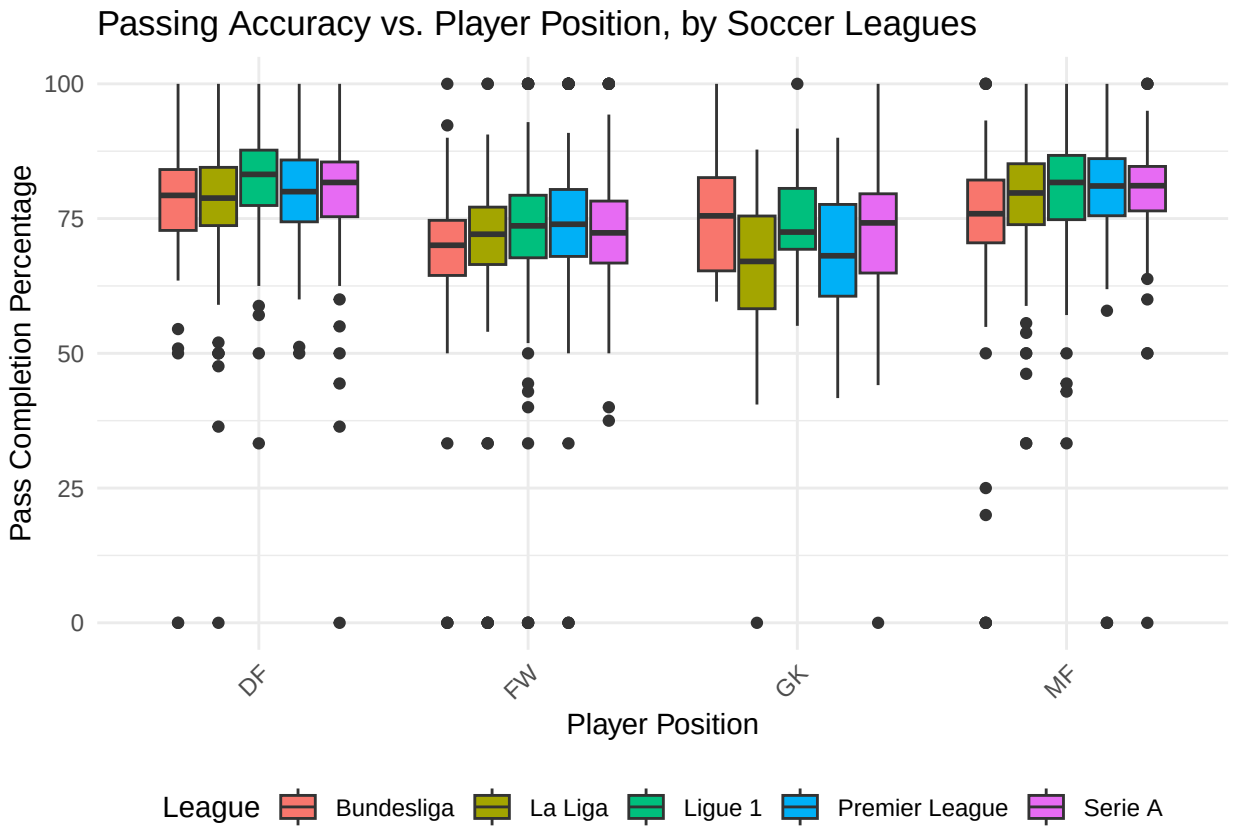
The research questions that we will be exploring are:

1. How are player demographics related to leagues and positions?
2. Offense: How does passing performance differ across player position and league, and what are predictors of goals?
3. Defense: How do metrics related to defensive performance differ across player position?

How are player demographics related to leagues and positions?

[add]

Offense: How does passing performance differ across player position and league?



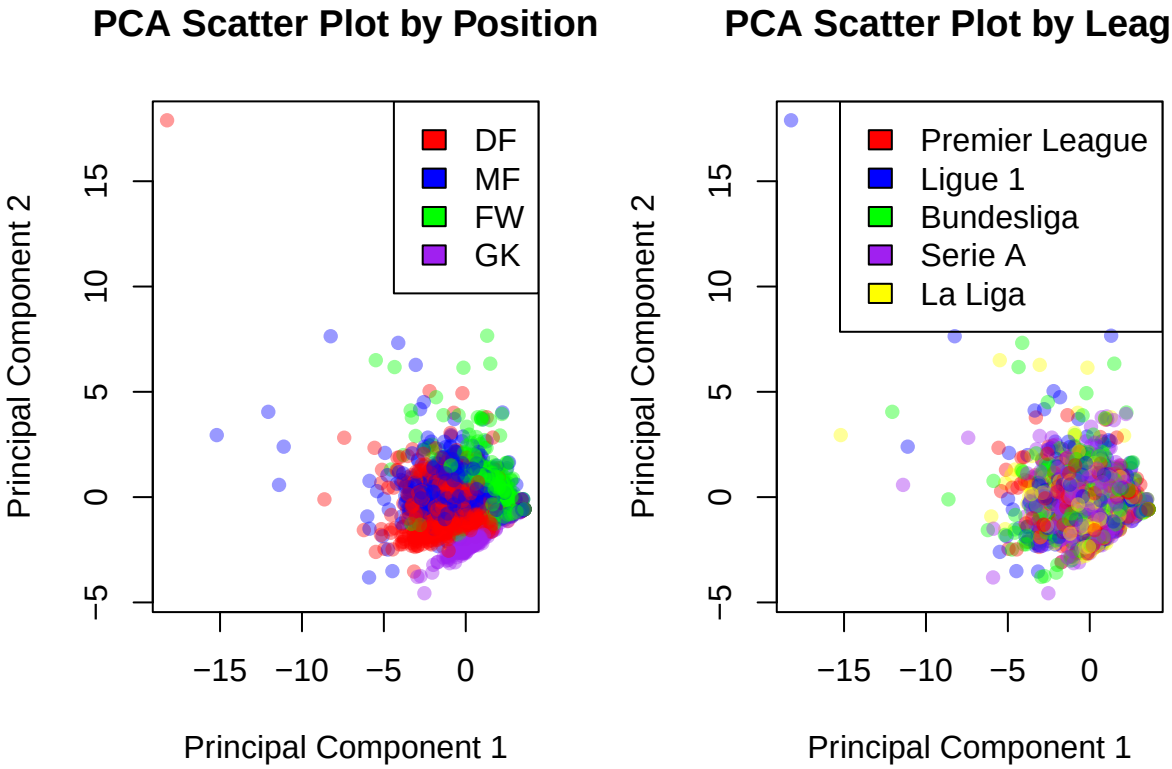
The box plot displays the distribution of passing completion percentages (`PasTotCmp%`) across different player positions, with the data color-coded by the league to show variations between leagues. The plot shows how passing completion percentages seems the highest for defenders, whereas forwards seem to have lower passing completion percentages. This might be because of defenders often engage in more controlled passes, while forwards may attempt riskier passes for goal-scoring opportunities.

The boxplots for different positions tend to overlap. Also, there does not seem to be a significant difference between the leagues for the passing accuracy by position, as all boxplots overlap for a given position. This plot seems to be relatively informative as we are easily able to compare the range of differences in passing accuracy by position and league that can help answer the question; however, there are some limitations. There appears to be many outliers, especially with lower passing percentages, for all positions other than goal keeper.

```
##                       Df Sum Sq Mean Sq F value   Pr(>F)
## Pos_simplified         3  41498   13833  94.623  < 2e-16 ***
## Comp                   4   6916    1729  11.827 1.57e-09 ***
## Pos_simplified:Comp   12   4830     403   2.754 0.000999 ***
## Residuals           2901 424093     146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After checking the Q-Q plot and residuals vs. fitted plot, the data seem to reasonably meet the assumptions for an ANOVA test. The ANOVA results above indicate significant differences in passing accuracy among different player positions (`Pos_simplified`), as indicated by the F-statistic of 94.623 (p-value $<$ 2e-16), at the

standard alpha level of 0.05. The factor `Comp` (representing soccer leagues) also significantly influences passing accuracy, with an F-statistic of 11.827 (p-value 1.57e-09), suggesting variations in passing performance across leagues. Furthermore, there is a significant interaction effect between player position and league (`Pos_simplified:Comp`), denoted by an F-statistic of 2.754 (p-value 0.000999), indicating that the impact of player position on passing accuracy differs across leagues. In summary, the findings suggest that both player position and league affiliation significantly contribute to differences in passing accuracy, and the interaction between these factors should be considered for a comprehensive understanding of the variations.



Using 7 passing variables- `PasTotCmp`, `PasTotDist`, `PasTotPrgDist`, `PasAss`, `PasProg`, `PasInt`, and `PasBlocks`- PCA projected the data into the lower-dimensional space defined by the first two principal components. There appears to be some grouping by player positions, suggesting that the principal components may capture patterns related to passing performance across different positions. However, there appears to limited distinction by league in the projection, suggesting that the variability explained may not strongly differentiate passing styles between the considered leagues.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5    PC6     PC7
## Standard deviation     1.7327 1.2618 0.9368 0.82631 0.69654 0.5748 0.17217
## Proportion of Variance 0.4289 0.2274 0.1254 0.09754 0.06931 0.0472 0.00423
## Cumulative Proportion  0.4289 0.6563 0.7817 0.87926 0.94857 0.9958 1.00000
```

As shown above, the first two principle components explain 65.63% of the variability in the data. This indicates majority of the original information is captured in these two components, suggesting an effective reduction in dimensionality while retaining a considerable amount of the variability present in the seven passing variables.

Offense: What are predictors of goals?

[add]

Defense: How do metrics related to defensive performance differ across player position?

[add]

Conclusion and Main Takeaways

[add]

Sources

- Kaggle