Columbia University

Applied Analytics Frameworks and Methods II MLS Analytics Final Project

Adeel Arif, Aaman Basra, Pallavi Gudipati, Sachin Otsuka Arjun

APANPS5205

Professor V. Lala

April 24, 2025

Introduction:

As passionate football (soccer) fans and devoted Manchester United supporters, we have witnessed endless debates on how a player's physical attributes affect their performance. Whether it is the criticism of Cristiano Ronaldo being "too old" to score effectively (Verschueren, 2024), discussions about Lisandro Martinez's height making him unfit to play as a center-back (CB) and needing to switch to left-back (LB) (Anka, 2024), or even claims that Romeulu Lukaku' weight negatively affects his game (Neville, 2018), media narratives often prioritize physical attributes like age, height, and weight over technical ability, intelligence, and awareness.

These narratives don't just appear in commentary but also shape how players are categorized and judged. Both football media and video games like FIFA use player archetypes, broad labels to describe players based on their contributions (e.g., "playmaker" or "swiss-army-knife"), to represent their style and overall contribution to the game (Nicholson, 2023). These archetypes go beyond traditional positions to capture how players actually impact the match, helping fans and analysts alike make better comparisons across different playing styles.

Yet history has repeatedly shown that football greatness transcends physical stereotypes. The assumption that a player's physical traits determine their effectiveness shapes scouting, key coaching/management decisions, and public perception. The goal of this study is to use a data-driven approach to analyze whether physical attributes like age, height, and weight influence a player's performance, and the reasons behind players outperforming their archetype-expected performance. By shedding light on this issue, we hope to reshape the conversation around player evaluation and contribute to a more informed understanding of what truly makes a footballer great.

Data Collection:

After some preliminary research, we decided to source individual player-level performance and demographic data from the American Soccer Analysis API, which allowed us to get detailed information about Major League Soccer (MLS) players. While our initial motivation did stem from media narratives around Manchester United players, using the MLS league as a whole would help build a more robust model to objectively compare metrics with player performance. Further, using the MLS gave us access to a clean, structured, API-accessible dataset.

We used two public-facing APIs from our source, with the first API source providing performance metrics like goals, shot data, and seasons played, while the second API source provided player demographic data like height, weight, position, and debut season. These APIs were pulled using the GET() function from R's httr package, which was then parsed from JSON() and stored in a structured data frame. After this, the raw data was retrieved, and the steps below were performed to optimally clean the data.

Firstly, we flattened the nested JSON in order to provide a clean data table for use. Height was converted from feet and inches to cm, and missing height/weight columns were imputed using position-wise imputation. Missing positions were inferred using literature-inferred relations between player demographics and positions. Missing debut seasons were imputed with the 2026 season, as we assumed these players were yet to make their debut. String fields like player names were cleaned by providing the "Unknown" label for missing names, and finally, both datasets were joined using player ID into one large table. This cleaning was critical as it allowed us to construct a complete data table ready for exploration, modeling, and anomaly detection techniques.

Key Definitions:

1) Goals – A goal is scored when the whole of the ball passes over the goal line.
2) Assists – The final pass that leads to a goal.
3) Points Added – It measures how all of a player's actions impact team performance, showing their total value over a season. (Muller, 2014)

Literature Review:

We examined several research works to assess how factors like age, weight, height, and position influence a player's performance, especially in the context of position-specific roles and age-performance trends.

According to a study done by Nicolas Brito, it was found that forwards usually have a peak performance window from around 25 to 30 years of age, whereas defenders and goalkeepers tend to peak later in their careers, around 27-32 years of age. This shows that forwards maximally optimize their goal-scoring opportunistic skill set at a young age because it requires quick sprints to goal-scoring positions, whereas defenders and goalkeepers peak much later, because they can afford to have a physical decline by compensating with experience, reflexes, and anticipation (Brito, n.d.).

Research by Bongiovanni et al. and Leão et al studied the varying body composition by football playing positions and consistently found that height and weight differ significantly by role and positions. Goalkeepers were the tallest players as their position required the biggest reach for shot-stopping, aerial dominance, and commanding the penalty area. Central defenders were also taller and heavier, which is consistent with the need for physicality to defend in duels and aerial challenges (Leão, 2019). Midfielders were generally the shortest with the least body mass, which is contrary to other positions because their role required them to be more agile and quick to change directions, which is essential for playmaking, pressing, and making attacking transitions (Bongiovanni, 2020). Another systematic review that was conducted supported this concept and emphasized the physical characteristics of football players, which concluded that goalkeepers and defenders were the tallest and heaviest, while midfielders were the shortest and lightest (Sebastiá-Rico, 2023).

These findings suggest that physical attributes are not random and typically align with positional demands in football. While existing studies address the demographic variables and how they impact a player's position and performance, only a few have combined this with modern football metrics like Points Added or explored whether any anomalies or outliers outperform or underperform their expected role.

This background study helped shape our research questions, providing us the opportunity to focus on the gaps of previous academic findings through actual MLS data to investigate age-performance trends and player archetypes.

Research Questions & Hypotheses:

Drawing from the gaps identified in existing literature and leveraging our cleaned MLS dataset, we formulated the following research questions:

1) Does position influence player performance in terms of goals, primary assists, and points added?

2) What types of performance-based player archetypes exist in MLS?

3) Do physical traits explain those archetypes, and can physical traits predict on-field performance?

4) Which MLS players significantly deviate from their position-specific performance and physical norms, and what underlying traits explain these anomalies?

Position Analysis

To determine whether player position influences offensive contribution, we compared *Goals + Assists per Season* and *Points Added per Season* between offensive and defensive position groups. Players were grouped into these categories based on their primary playing roles: forwards and midfielders were categorized as offensive, while defenders and goalkeepers were categorized as defensive.

1. Visual Analysis

Boxplots and density plots revealed clear differences in performance between the two groups. Offensive players exhibited distinct results (see appendix, Figures 1-4), as outlined below:

- Significantly higher medians and upper quartiles for both goals + assists per season and points added/season. (Figure 1 & Figure 2)
- Greater variability in contribution, with numerous high-performing outliers. (Figure 1 & Figure 2)
- Density distributions skewed heavily right, indicating that while many offensive players contribute moderately, a small subset delivers exceptional output. (Figure 3 & Figure 4)

Conversely, defensive players clustered tightly near zero for both metrics, with limited variance and very few outliers. (Figure 3 & Figure 4)

2. Statistical Significance (Welch's T-Tests)

To validate whether these observed differences were statistically significant, we ran Welch's two-sample t-tests, which are robust to unequal variances.

- Goals + Assists per Season
    - Mean (Offensive): 2.96 vs. Defensive: 0.76
    - $t = -11.8$, $p < 2.2e-16$
    - 95% Confidence Interval: [-2.57, -1.83]
- Points Added per Season
    - Mean (Offensive): 0.88 vs. Defensive: 0.17
    - $t = -10.92$, $p < 2.2e-16$
    - 95% Confidence Interval: [-0.83, -0.58]

Both tests yielded highly significant results, confirming that offensive players significantly outperform defensive players in both metrics.

3. Mean Comparison with Confidence Intervals

Bar charts with 95% confidence intervals further supported the findings (see appendix, Figures 5 & 6):

- Offensive players' mean contributions were not only higher, but their confidence intervals did not overlap with those of defensive players. (Figure 5 and Figure 6)
    - This further reinforces that the difference in means is not due to sampling variability but reflects a true underlying positional effect.

These analyses *strongly support* the decision to isolate offensive players for deeper modeling (e.g., archetype clustering, regression). The significant differences in scoring and value-adding metrics across positions highlight that offensive roles operate under fundamentally different performance dynamics. By segmenting players this way, we avoid confounding variables and ensure that performance models are fair, interpretable, and role-specific.

Archetype Analysis

To identify and understand different types of offensive players in the MLS, we performed K-means clustering on scaled performance metrics, including goals, assists, key passes, shots, minutes played, and points added. Based on the Elbow Method, we selected K = 3 as the optimal number of clusters (Figure 7), which allowed us to group midfielders and forwards into three distinct player archetypes: Low Impact Players, Reliable Starters, and Elite Playmakers (Figure 8). Although we did not explicitly run Principal Component Analysis (PCA), the fviz_cluster() function from the factoextra R package applied PCA automatically to reduce our six-dimensional data into two dimensions for visualization (Kassambara, 2017). Dimension 1 (82.8%) captured the majority of variation, while Dimension 2 (11.2%) accounted for additional variance, together explaining approximately 94% of the total variability in player performance (Figure 8). This means the 2D plot preserves almost all the key information from the original dataset, making it a reliable visual summary of player groupings. Players positioned closer together on the plot exhibit more similar offensive profiles.

Table 1 summarizes the average per-season performance for each player archetype. As shown, Elite Playmakers significantly outperform other groups across all metrics, averaging nearly 7 goals and over 3.5 assists per season. Reliable Starters contribute solidly across key stats, while Low Impact Players offer limited offensive output. To confirm these differences, we ran a one-way ANOVA, which showed that all performance metrics were statistically significant across groups ($p < 0.001$) (Figure 9). To further investigate which specific archetypes differed, we conducted pairwise t-tests (Figure 10). As shown in Figure 10 and Table 1, all comparisons across groups were statistically significant for every performance metric ($p < 0.05$), reinforcing the validity of our cluster distinctions. Importantly, the distribution of players across these archetypes is also telling: the majority fall into the Low Impact group (n = 377), a moderate number into Reliable Starters (n = 82), and only a small handful into the Elite Playmakers (n = 24) (Table 1). This pattern makes intuitive sense as it reflects the reality that top-tier performers are rare, while average or lower-impact players are far more common. Overall, this clustering framework offers a more nuanced understanding of player roles and value beyond just position or name recognition and lays the foundation for further analysis exploring how physical traits may influence or align with these archetypes.

Physical Traits:

To explore whether physical attributes align with on-field performance, we compared the average age, height, and weight across the three player archetypes (Table 2). The summary table (Table 2) and box plots (Figures 11-13) show that Elite Playmakers tend to be slightly older and lighter than the other groups. We conducted pairwise t-tests to assess the statistical significance of these differences (Figure 14). The results showed that age was the only trait with a statistically significant difference, specifically between Elite Playmakers and Reliable Starters ($p = 0.0229$), as highlighted in the age distribution plot. This suggests that more experienced players may be more likely to achieve elite performance. In contrast, height and weight differences were not significant ($p > 0.05$), indicating that these physical traits do not meaningfully distinguish between player archetypes (Figure 14).

Predictive Modeling

Although physical traits like height and weight did not significantly distinguish between archetypes, the age trend suggested a potential relationship with performance. Therefore, to predict offensive performance, we ran two types of models: decision trees and linear regression. Goals + Assists per season and points added/season were used as separate outcome variables. Dependent variables included age, height, and weight. For each outcome, we built a dedicated decision tree and regression model.

Across both decision trees, age emerged as the most important predictor of offensive performance (Figure 15 & Figure 16). Players aged 26 and older generally outperformed younger players, especially when paired with certain physical traits. In the Goals + Assists per season tree (Figure 15), the top performers were older than 26 and lighter in weight, with some subgroups averaging up to 24 goals + assists per season. In the Points Added/season tree (Figure 16), the highest values were seen in players over 26, lighter in weight, and at least 186 cm tall, averaging around 8 points added per season. Variable importance scores confirmed this trend, indicating that age ranked highest in both models, followed by weight and then height.

While both models provided valuable insights, we focused more heavily on the linear regression approach due to its ability to generate an explicit predictive formula. This formula allows us not only to quantify how much each trait impacts performance but also to apply the model for anomaly detection, flagging players whose performance is significantly above or below what would be expected based on their physical attributes. The regression models confirmed that age was a significant positive predictor for both goals + assists per season and points added/season. In the goals + assists per season model, age had the strongest effect ($\beta = 0.15$, $p < 0.01$), while weight had a small negative impact ($\beta = -0.04$), and height was not significant (Figure 17). In the points added/season model, age again showed a positive effect ($\beta = 0.05$, $p < 0.05$), with height and weight remaining non-significant (Figure 17).

Overall, our analysis highlights age as the most consistent and impactful predictor of offensive contribution in the MLS. Weight, while less consistent, also showed potential as a meaningful predictor. With these predictors identified, we next turned to a practical application of our regression model: using it to detect performance anomalies.

Anomaly Detection:

After identifying key variables in the multiple variable regression model, we wanted to use this model to predict a player's expected output in terms of goals + assists per season, as well as points added per season using the physical traits of age, height, and weight. This would allow us to generate baseline predictions that could then be compared to actual performance.

As seen in Figure 18 in the Appendix, in order to predict goals plus assists per season, a regression formula was built. Based on the output of our model, this equation was

$$Predicted\, goals\, +\, assists/season\; =\; 4.10\; +\; 0.153\, (Age)\; +\; 0.0006\, (Height)\; -\; 0.0368\, (Weight)$$

The R^2 value of 6.8% suggested that physical traits can provide a baseline, but a lot of player performance is also hinged on external factors like tactics and player chemistry.

As seen in Figure 19, we also ran a separate regression to predict points added per season, giving the following equation:

$$Predicted\, points\, added/season\; =\; 0.18\; +\; 0.0459\, (Age)\; +\; 0.0032\, (Height)\; -\; 0.0079\, (Weight)$$

This model's R^2 value of 4.7% suggested again that physical traits can provide a baseline, but a lot of player performance in terms of points added per season is also hinged on external factors like tactics and player chemistry (Figure 19).

These formulas helped us calculate residuals, the difference between a player's actual performance and the value predicted by the formula based on age, height, and weight. For example, in the case of goals + assists per season, the formula is:

$$Residual\; =\; Actual\, Goals\; +\; Assists\, per\, season\; -\; Predicted\, Goals\; +\; Assists\, per\, season$$

These residuals were calculated for both output variables and became the foundation of our anomaly detection process, with players with positive residuals being considered overperformers and players with negative residuals being considered underperformers. We initially planned to standardize residuals to calculate z-scores, allowing us to identify players who were statistically significant outliers based on predetermined thresholds. However, we realized that these thresholds were extremely restrictive. Thus, rather than using z-score thresholds, we used a rank-based approach, selecting the top 2 overperformers and underperformers for each player archetype, allowing us to identify standout examples in each archetype. We then visualized these anomalies using a bar plot, as well as descriptive tables, as seen in the appendix. This data-driven approach, incorporating clustered player archetypes and a multivariate linear regression model, can help scouts reevaluate player expectations and identify overlooked talents.

An example of a case of an anomaly is David Villa, a player clustered as a "low impact player" archetype based on his physical attributes of 68 kg weight, 175 cm height, and older age of 35 years old. Due to these demographic variables, the model predicted David Villa to only score 5.2 goals + assists per season, and contribute to adding only 1.5 points per season for his team. However, his real performance highly exceeded expectations, as noted in the table below.

Metric	Actual	Predicted	Residual
Goals+Assists per season	24.75	5.20	+19.55
Points added per season	9.48	1.51	+7.97

This makes David Villa a positive anomaly, or overperformer, in this dataset. Despite his age and weight, he consistently provided both goals + assists/season and points added/season to the team. Some potential real-world explanations for David Villa's overperformance include his extensive experience, having played at the highest levels as a World Cup and Euro Champion (Fernandes, 2024). This veteran background likely gave him a competitive edge in reading the game and staying composed under pressure. Additionally, his leadership role within the team was crucial. As a Designated Player and team captain, the system was often built around him, allowing him to maximize his impact on the field (Fernandes, 2024).

Another example of a performance anomaly is Juninho Pernambucano. Based on his physical attributes, the model predicted that he would contribute 4.33 goals + assists per season and 3.39 points added per season. However, his actual performance fell well below those expectations, with only 1.85 goals + assists and 1.75 points added per season. This resulted in negative residuals of -2.48 and -1.64, respectively, making Juninho a negative anomaly, or underperformer, in the dataset. Two key factors may explain this discrepancy. First, Juninho experienced a positional pivot, being deployed deeper in midfield as a playmaker rather than in an advanced attacking role, which naturally limited his opportunities to generate direct offensive output (Metro, 2018). Second, chemistry issues may have further impacted his performance. Reports suggest that disagreements with high-profile teammates like Thierry Henry disrupted team cohesion (MLSsoccer, 2013). This case illustrates how off-field dynamics and tactical usage can shape a player's performance beyond what physical attributes alone would predict.

Metric	Actual	Predicted	Residual
Goals+Assists per season	1.85	4.33	-2.48
Points added per season	1.75	3.39	-1.64

These examples highlight that while data can reveal valuable trends and outliers, it doesn't tell the full story. Context, like a player's role, experience, leadership, and team dynamics, plays a huge part in performance. Therefore, an amalgamation of data and context is critical to truly understanding player impact.

Conclusion:

This study explored whether physical traits – specifically age, height, and weight – influence offensive player performance in the MLS, using a combination of clustering, regression, and anomaly detection techniques. The analysis focused on offensive players only, as confirmed by statistically significant differences in performance metrics between offensive and defensive groups.

Our analysis showed that age is the most consistent predictor of offensive performance. Older players (particularly those over 26) tend to contribute more in terms of goals + assists per season, and points added per season, as seen in Figure 15 and Figure 16. These insights can be particularly useful for scouts and the technical audience, the target audience, when evaluating how effective a player may be.

As seen through the Pairwise T-tests in Figure 14, statistical comparisons revealed that height and weight showed limited or no significant influence, indicating that body size is a poor standalone predictor of offensive effectiveness. However, both metrics must be viewed separately and not combined to avoid multicollinearity, as height and weight affect goals, assists, and points in different ways. Combining them would oversimplify important performance dynamics. This suggests that focusing highly on body size may be misleading, and that scouts should instead focus more on technical ability when making player contract decisions.

As seen in Figure 8 and Table 1, clustering analysis grouped players into clear performance archetypes (Low Impact Players, Reliable Starters, Elite Playmakers), with only a small fraction of players reaching elite status. Recognizing these archetypes will allow scouts to find players who fit their team's needs most effectively.

Anomaly detection identified standout players who significantly overperform or underperform based on their physical traits, offering a data-driven lens for talent identification and scouting. As seen in Figure 20, players like David Villa who are expected to be low impact players, outperform expectations, while players like Roger Ezpinosa, as seen in Figure 21, who are expected to be elite playmakers underperform greatly, pulling back into the belief that experience, tactical role, and team dynamics often play a much more important role in shaping player performance. The combination of data and context is critical to truly understanding player impact. While our regression models, Figure 18 and Figure 19, offer a great baseline for evaluating the offensive output of players, the low R^2 values (6.8% for goals+assists per season, 4.7% for points added per season) indicate the pressing need to combine statistical insights with traditional scouting for the most informed decisions.

Limitations & Future Recommendations:

Our study did have some limitations, which guide our future recommendations.

Our dataset is focused solely on MLS data, a league that differs significantly from other leagues in terms of tactics and physical demands. This makes our current model and insights difficult to generalize to other, more prominent leagues like the Spanish La Liga or the English Premier League. Our recommendation for a future study would be to expand the dataset to incorporate multiple leagues to allow cross-continent comparisons across different leagues with differing playing styles and tactics.

The demographic variables available to us were also limited to age, weight, and height, excluding numerous other demographic variables that could be key predictors of performance. With more demographic data like sprint speed, injury history, and tactical intelligence, a more robust model could be created to determine on-field output. Our recommendation for a future study would be to find richer datasets that capture much deeper demographic and technical information, rather than surface-level data available in the dataset used for this study.

After cleaning our dataset, we also noticed its limited sample size, with the Elite Playmaker archetype consisting of only 24 players, which would limit the statistical power of our findings. This study also viewed player performance statistics as a snapshot, rather than a longitudinal dataset that evolves over time. Our recommendation for a future study would be to use a larger dataset that tracks player performance longitudinally over time, allowing us to get more insights into how physical attributes and performance attributes evolve as players age

As mentioned previously in the report, a big limitation of this study is the absence of context for the data, whether that be chemistry issues, personal issues, or tactical issues. These factors heavily impact the output of players, and our model does not account for these factors, which explains the relatively low R^2 values in our regression models. A great future recommendation would be to incorporate team-level metrics like possession or other coaching style variables like formations and average position on the pitch to create more distinctions within each archetype

The comparatively low R^2 values (6.8% for goals+assists per season, 4.7% for points added per season) imply that the multiple linear regression models may oversimplify the complexity of football performance. In order to capture better effects and patterns to understand the intricacies in football performance, we could model our predictions using non-linear techniques like Random Forests, XGBoost, or Neural Networks, and create interaction between variables like height and weight, which could highlight subtle relationships that could have been overlooked using linear models.

Our K-means clustering method successfully divided all players naturally into three different groups based on the elbow method. We overlooked the fact that some players could display hybrid traits and show versatility, for example, a midfielder could switch positions mid-season as a forward. A good future recommendation would be to take into account such versatile players to discover a more fluid range of archetypes by incorporating dynamic clustering and techniques such as Hierarchical Clustering, DBSCAN, or Gaussian Mixture Models.

In our analysis, we solely focused on offensive players (Midfielders and Forwards) and left out Defenders and Goalkeepers. However, for a team's success, we will have to look into the overall team contribution from players across every position. To supplement our analysis, a future recommendation would be to have a parallel study focusing on defensive players (Defenders and Goalkeepers) by combining data that includes defensive metrics like clearances, interceptions, and save percentage.

References

Anka, C. (2024, October 23). Does Lisandro Martinez at left-back work as more than a short-term fix for Manchester United? *The New York Times*. https://www.nytimes.com/athletic/5849741/2024/10/23/lisandro-martinez-manchester-united-left-back/

Bongiovanni, T., Trecroci, A., Cavaggioni, L., Rossi, A., Perri, E., Pasta, G., Iaia, F. M., & Alberti, G. (2020). Importance of anthropometric features to predict physical performance in elite youth soccer: a machine learning approach. *Research in Sports Medicine*, 1–12. https://doi.org/10.1080/15438627.2020.1809410

Brito, N. (n.d.). *The Relationship Between Age and Performance Across Professional Athletes Playing in a Forward Position in the Premier League Item Type Senior Project*. https://soar.suny.edu/bitstream/handle/20.500.12648/11818/6456_Nicolas_Brito.pdf?sequence=1

Fernandes, A. (2024, October 29). *How David Villa Helped the MLS Grow*. MLS Multiplex. https://mlsmultiplex.com/how-david-villa-helped-the-mls-grow-01jbbrrfrrt3

Kassambara, A. (2017). *Multivariate Analysis I Practical Guide To Cluster Analysis in R Unsupervised Machine Learning*. https://xsliulab.github.io/Workshop/2021/week10/r-cluster-book.pdf

Leão, C., Camões, M., Clemente, F. M., Nikolaidis, P. T., Lima, R., Bezerra, P., Rosemann, T., & Knechtle, B. (2019). Anthropometric Profile of Soccer Players as a Determinant of Position Specificity and Methodological Issues of Body Composition Estimation. *International Journal of Environmental Research and Public Health*, *16*(13), 2386. https://doi.org/10.3390/ijerph16132386

Metro, O. a. (2018, April 8). *Juninho Pernambucano is taking his shots from the press box now*. Once a Metro. https://www.onceametro.com/2018/4/8/17213668/juninho-pernambucano-is-taking-his-shots-from-the-press-box-now-flamengo-new-york-red-bills

mlssoccer. (2013). *New York Red Bulls part ways with Brazilian midfielder Juninho | MLSSoccer.com*. Mlssoccer. https://www.mlssoccer.com/news/new-york-red-bulls-part-ways-brazilian-midfielder-juninho?utm_source=chatgpt.com

Muller, J. (2014). *American Soccer Analysis*. American Soccer Analysis. https://www.americansocceranalysis.com/what-are-goals-added

Neville, G. (2018, December 10). *Gary Neville on Romelu Lukaku: Verdict on striker's weight issue*. Sky Sports. https://www.skysports.com/football/news/29326/11577547/gary-neville-on-romelu-lukaku-verdict-on-strikers-weight-issue

Nicholson, J. (2023, August 21). *EA Sports FC 24 Archetypes - All acceleration types*. Gfinityesports.com. https://www.gfinityesports.com/article/ea-sports-fc-24-archetypes

Sebastiá-Rico, J., Soriano, J. M., González-Gálvez, N., & Martínez-Sanz, J. M. (2023). Body Composition of Male Professional Soccer Players Using Different Measurement Methods: A Systematic Review and Meta-Analysis. *Nutrients*, *15*(5), 1160. https://doi.org/10.3390/nu15051160

Verschueren, G. (2024). *Bayern Munich President: Cristiano Ronaldo Is "Too Old" Amid Rumoured Interest*. Bleacherreport.com.

https://bleacherreport.com/articles/2875281-bayern-munich-president-cristiano-ronaldo-is-too-old-amid-rumoured-interest

APPENDIX:

Key Figures:

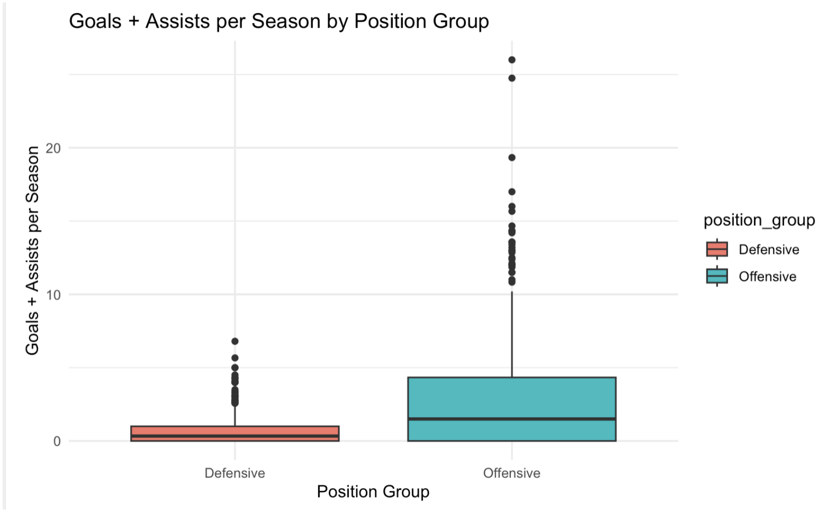


Figure 1: Boxplot - Goals + assists per season by position group

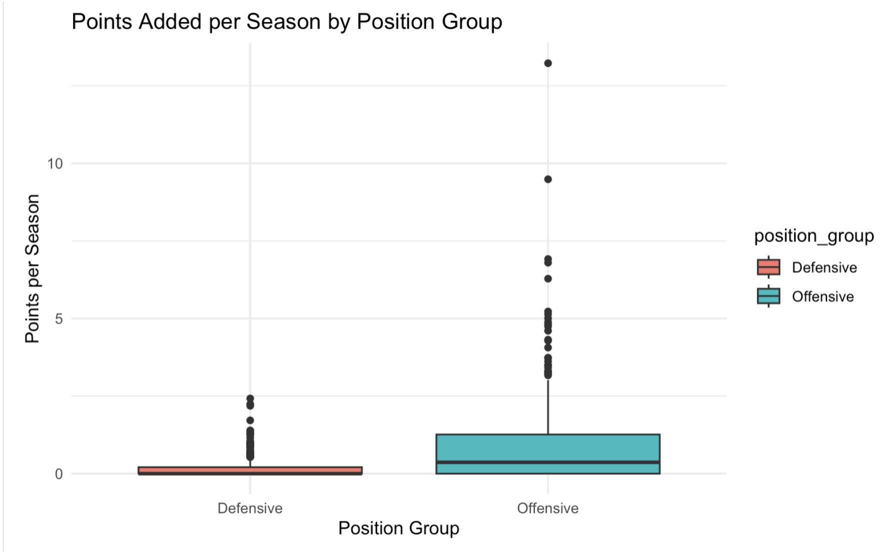


Figure 2: Boxplot - Points added per season by position group

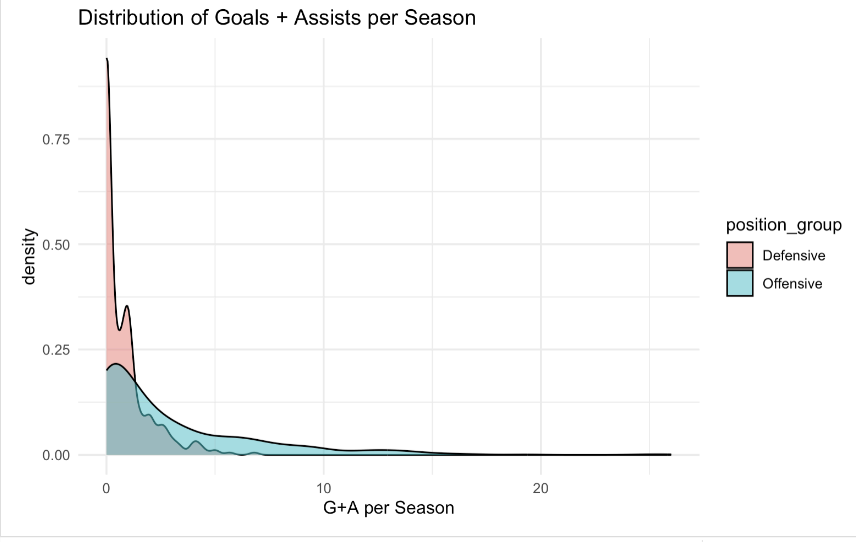


Figure 3: Density plot - Goals & assists per season by position group

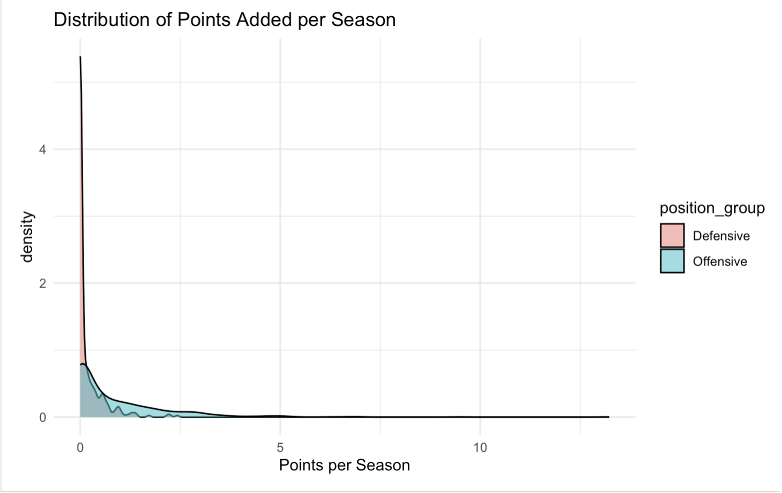


Figure 4: Density plot - Points added per season by position group

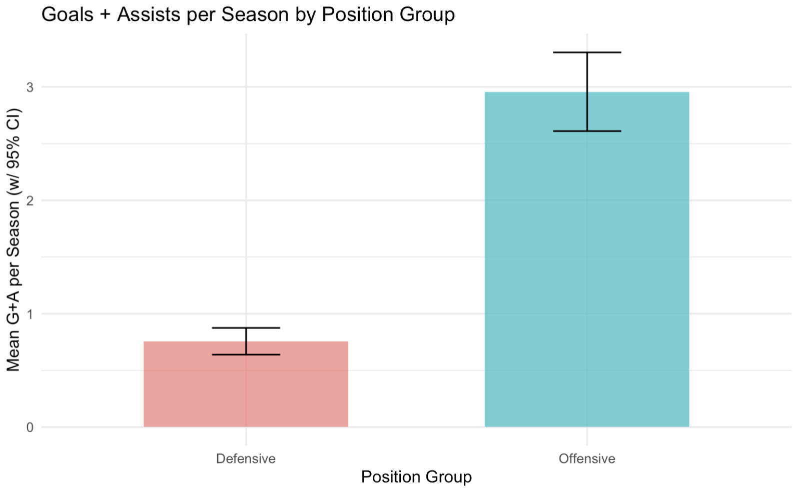


Figure 5: Bar chart - Goals & assists per season by position group with 95% Confidence Interval

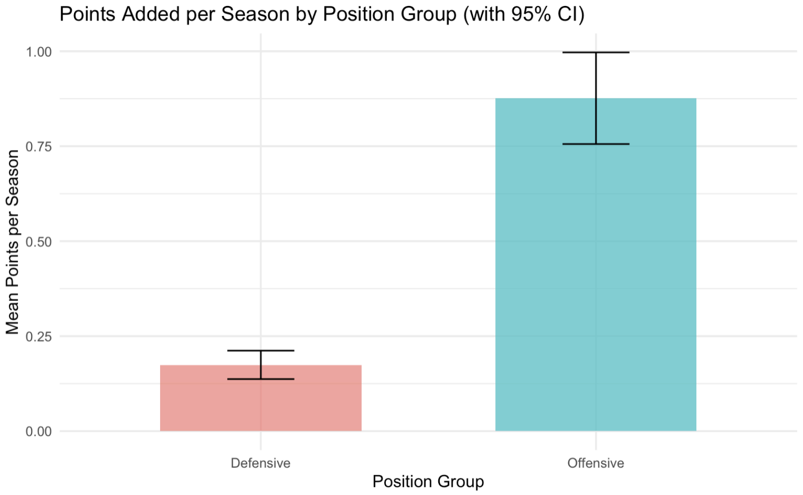


Figure 6: Bar chart - Points added per season by position group with 95% Confidence Interval

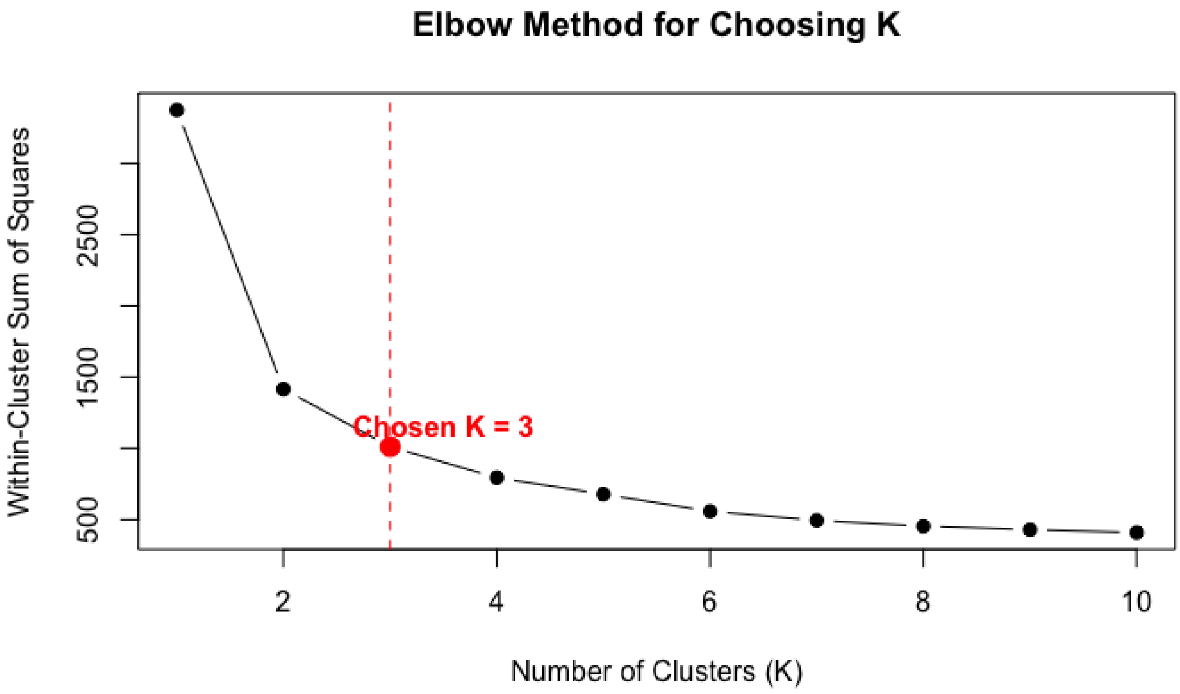


Figure 7: Elbow Method for Clustering

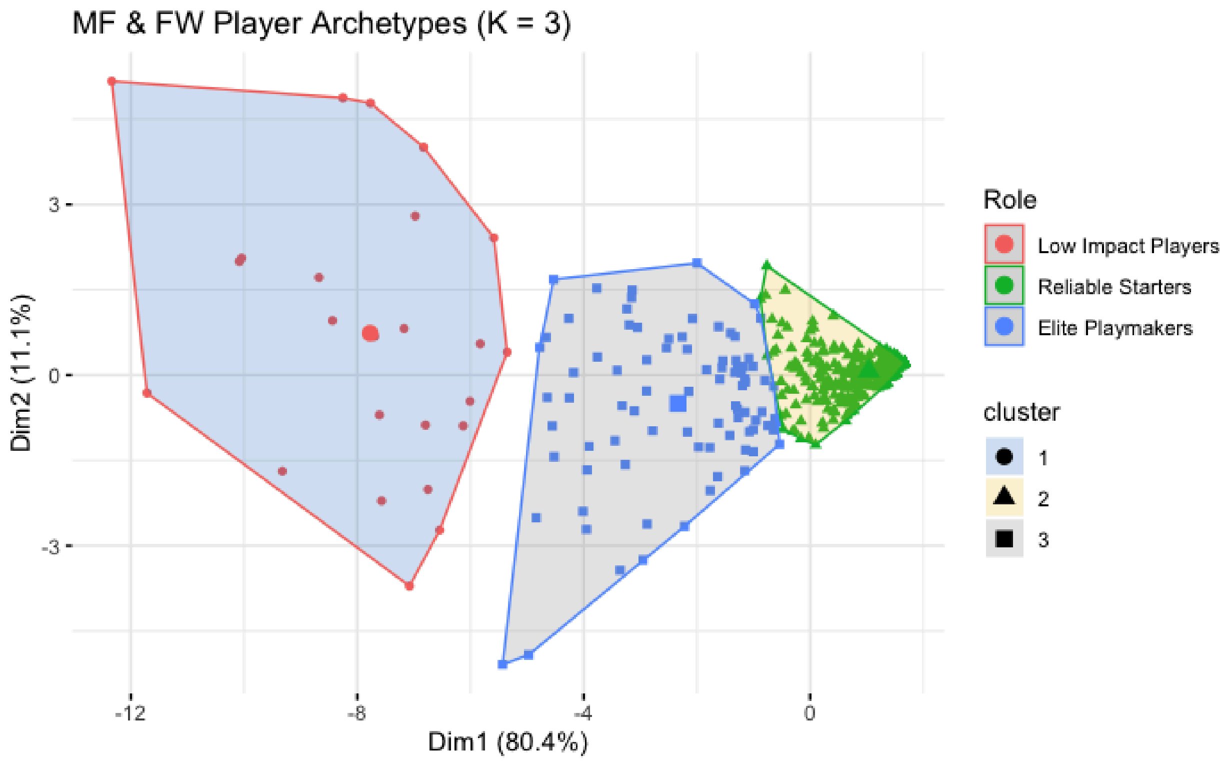
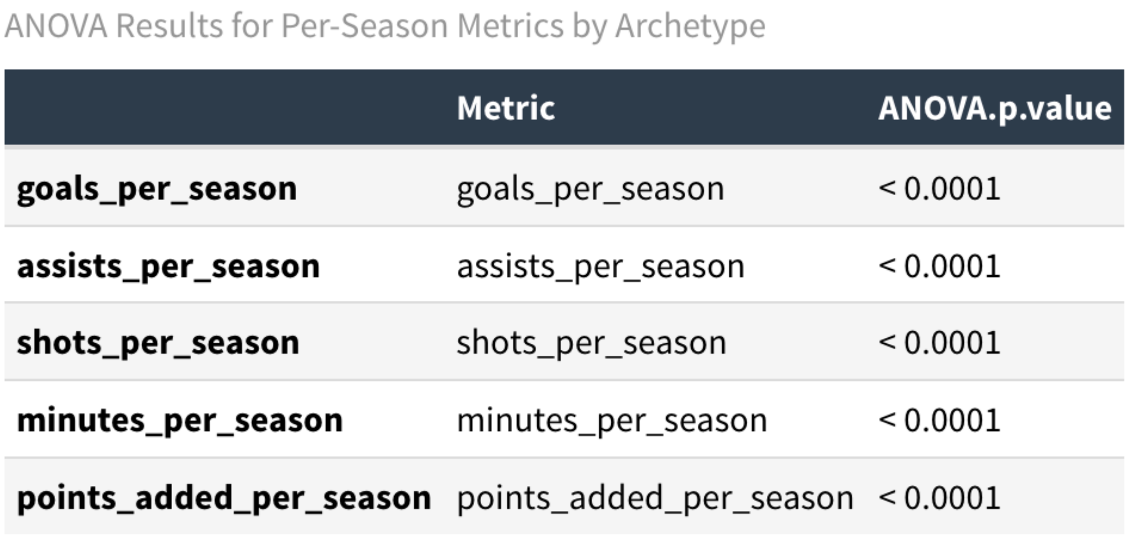


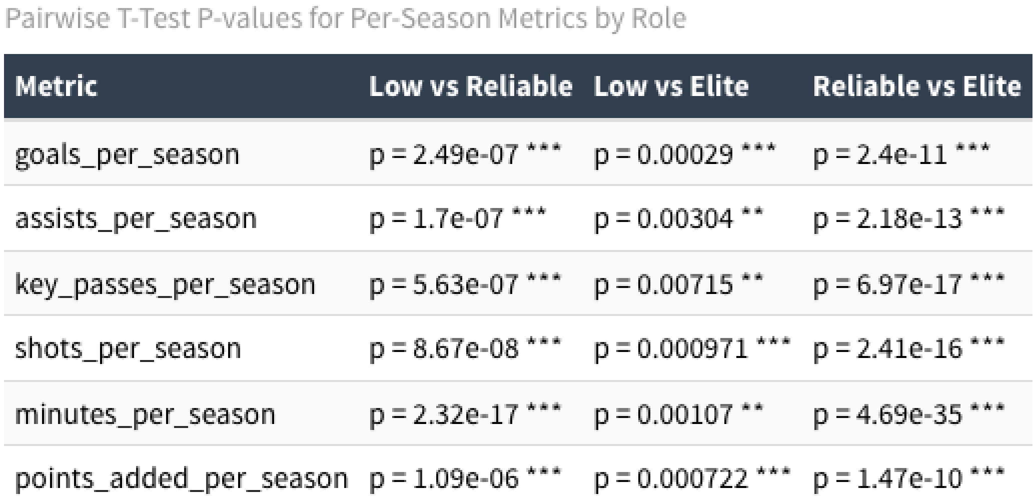
Figure 8: Clustering & PCA Using fviz_cluster() Function



ANOVA Results for Per-Season Metrics by Archetype

	Metric	ANOVA.p.value
goals_per_season	goals_per_season	< 0.0001
assists_per_season	assists_per_season	< 0.0001
shots_per_season	shots_per_season	< 0.0001
minutes_per_season	minutes_per_season	< 0.0001
points_added_per_season	points_added_per_season	< 0.0001

Figure 9: One-Way ANOVA for Archetype Performance Metrics



Pairwise T-Test P-values for Per-Season Metrics by Role

Metric	Low vs Reliable	Low vs Elite	Reliable vs Elite
goals_per_season	p = 2.49e-07 ***	p = 0.00029 ***	p = 2.4e-11 ***
assists_per_season	p = 1.7e-07 ***	p = 0.00304 **	p = 2.18e-13 ***
key_passes_per_season	p = 5.63e-07 ***	p = 0.00715 **	p = 6.97e-17 ***
shots_per_season	p = 8.67e-08 ***	p = 0.000971 ***	p = 2.41e-16 ***
minutes_per_season	p = 2.32e-17 ***	p = 0.00107 **	p = 4.69e-35 ***
points_added_per_season	p = 1.09e-06 ***	p = 0.000722 ***	p = 1.47e-10 ***

Figure 10: Pairwise T-Test P-values for Per-Season Metrics by Role

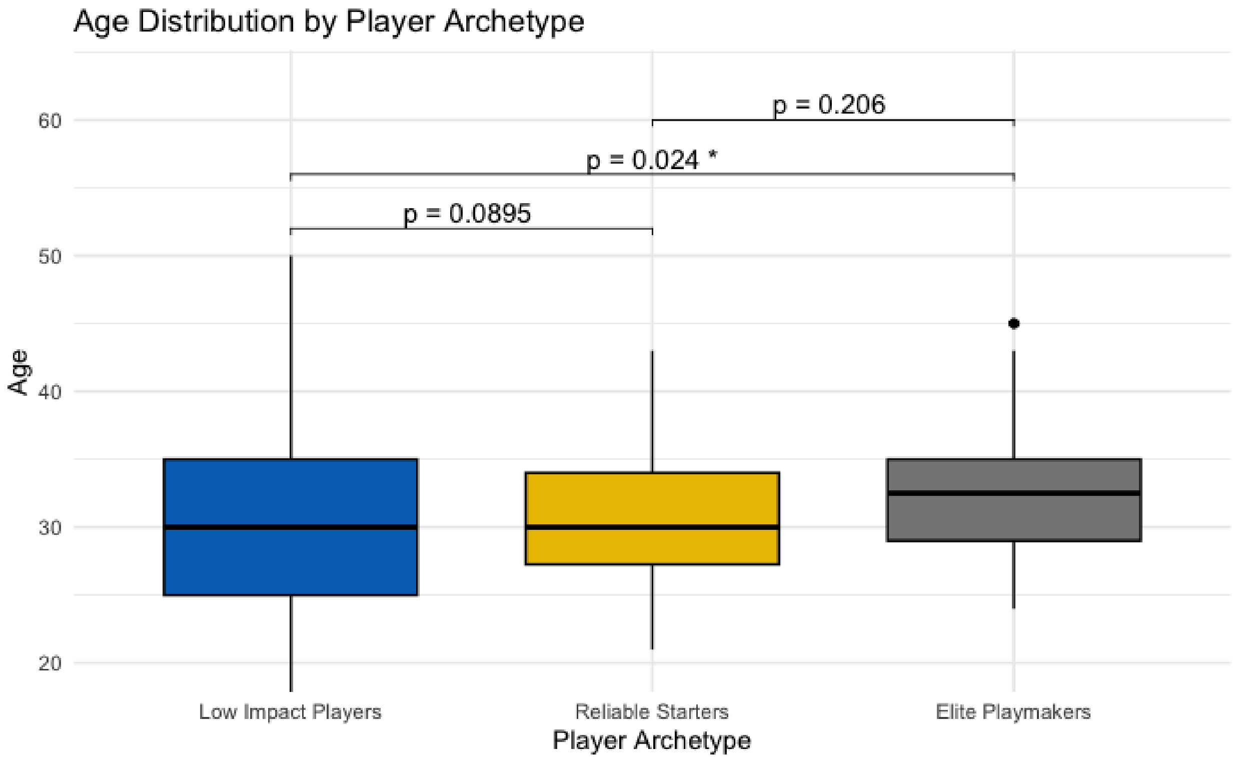


Figure 11: Age Distribution by Player Archetype

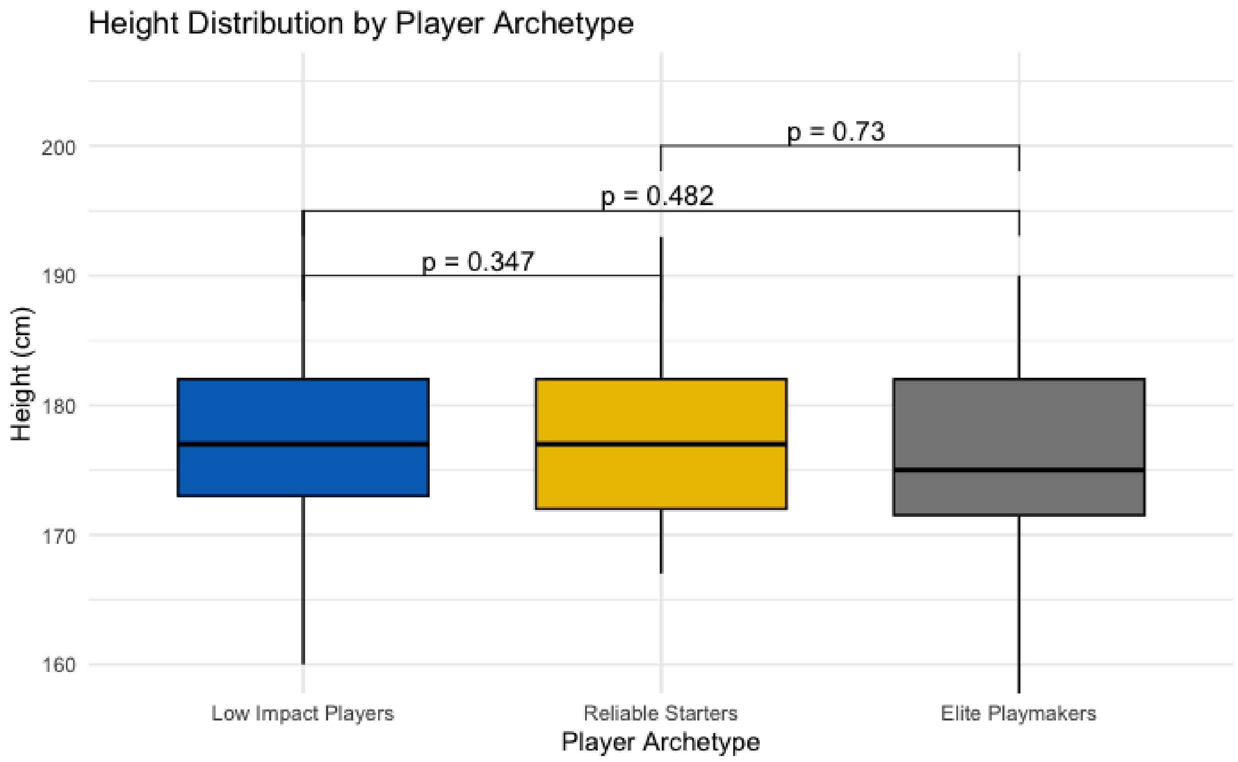


Figure 12: Height Distribution by Player Archetype

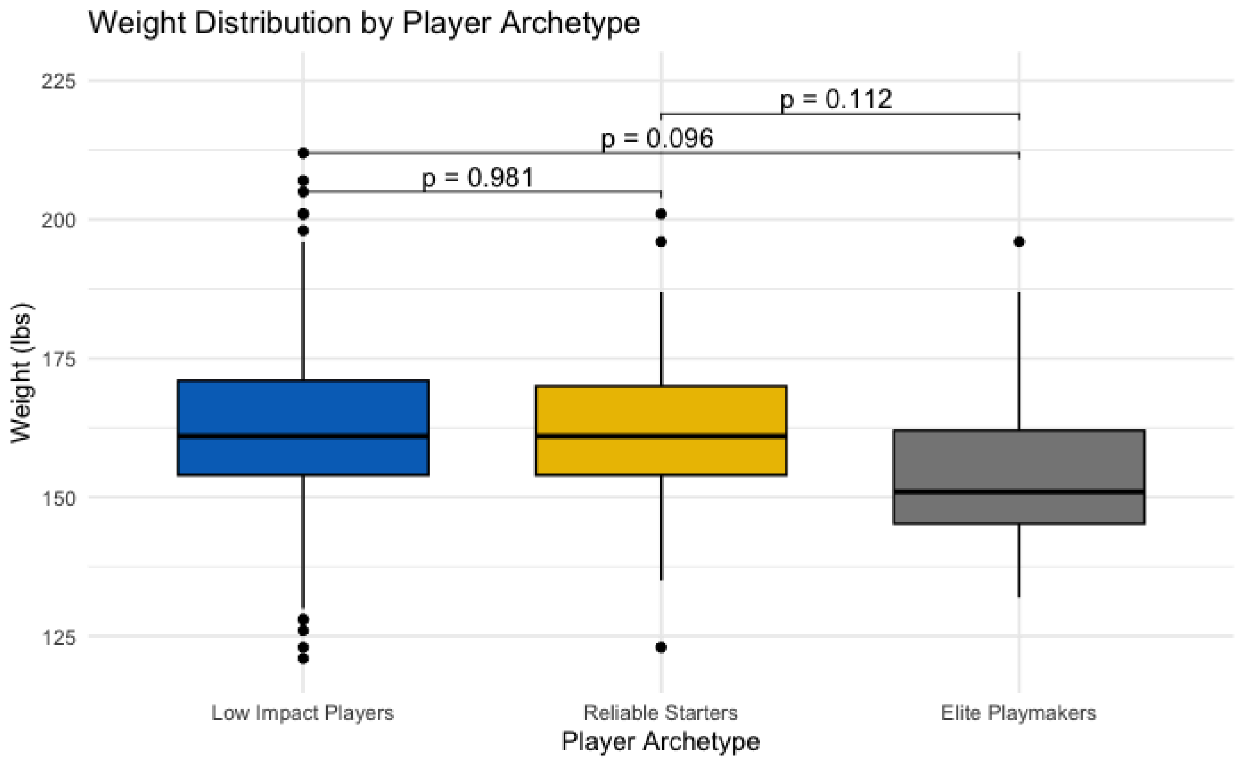
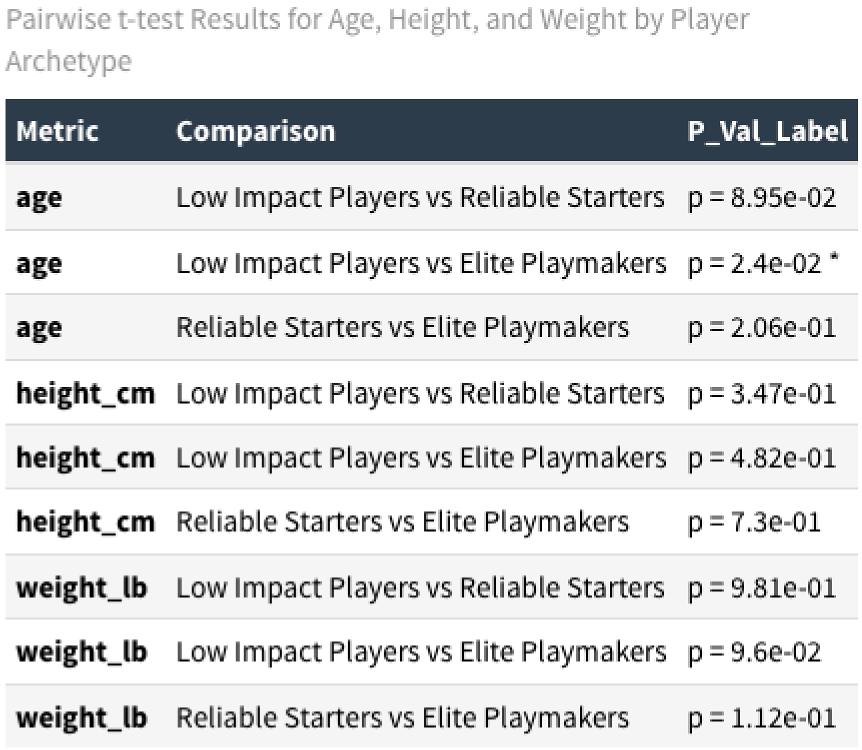


Figure 13: Weight Distribution by Player Archetype



Pairwise t-test Results for Age, Height, and Weight by Player Archetype

Metric	Comparison	P_Val_Label
age	Low Impact Players vs Reliable Starters	p = 8.95e-02
age	Low Impact Players vs Elite Playmakers	p = 2.4e-02 *
age	Reliable Starters vs Elite Playmakers	p = 2.06e-01
height_cm	Low Impact Players vs Reliable Starters	p = 3.47e-01
height_cm	Low Impact Players vs Elite Playmakers	p = 4.82e-01
height_cm	Reliable Starters vs Elite Playmakers	p = 7.3e-01
weight_lb	Low Impact Players vs Reliable Starters	p = 9.81e-01
weight_lb	Low Impact Players vs Elite Playmakers	p = 9.6e-02
weight_lb	Reliable Starters vs Elite Playmakers	p = 1.12e-01

Figure 14: Pairwise t-test Results for Age, Height, and Weight by Player Archetype

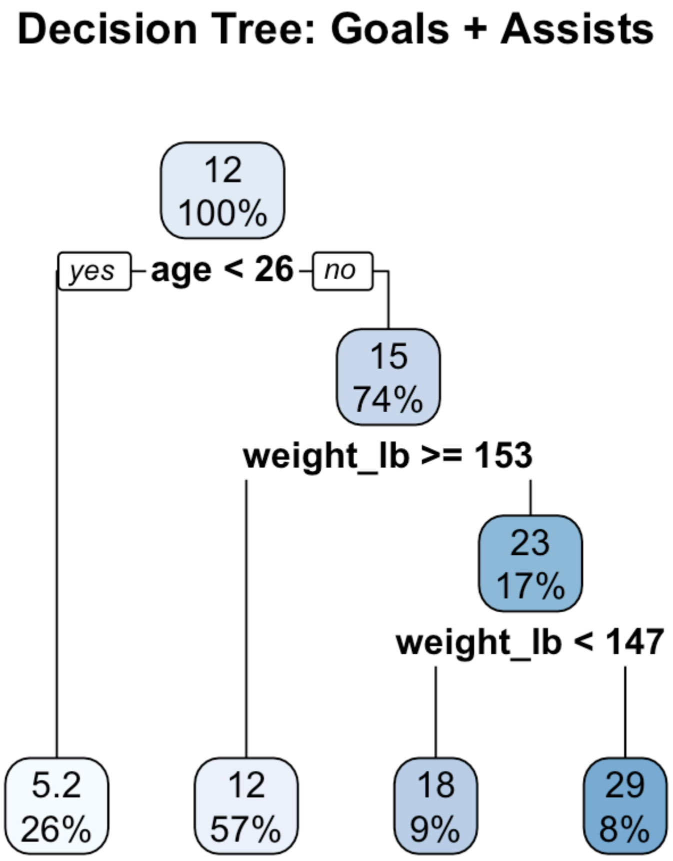


Figure 15: Decision Tree – Goals + Assists per Season

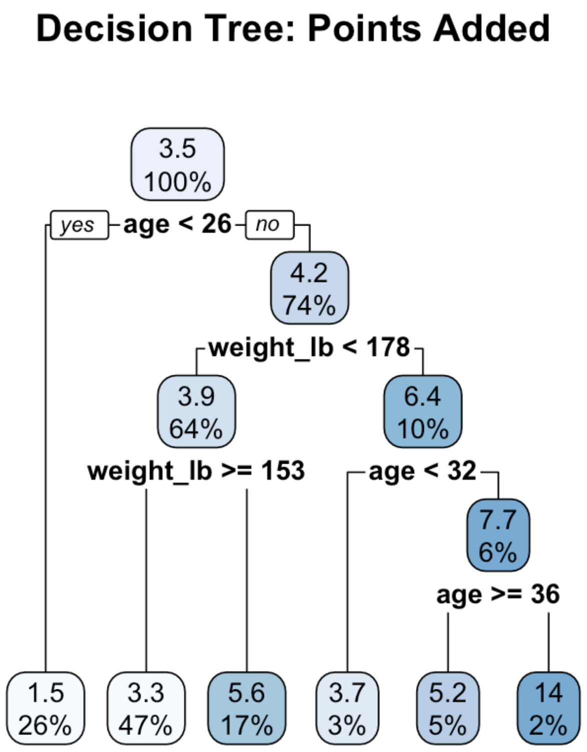


Figure 16: Decision Tree – Points Added/Season

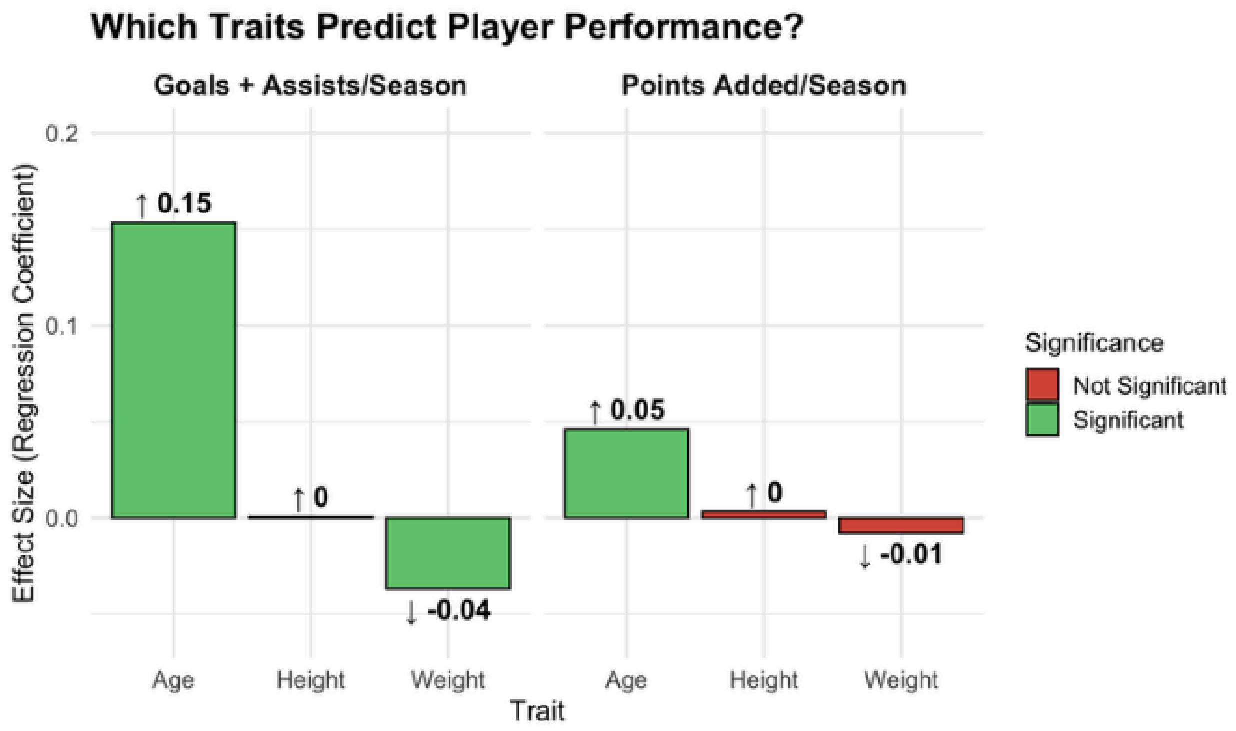
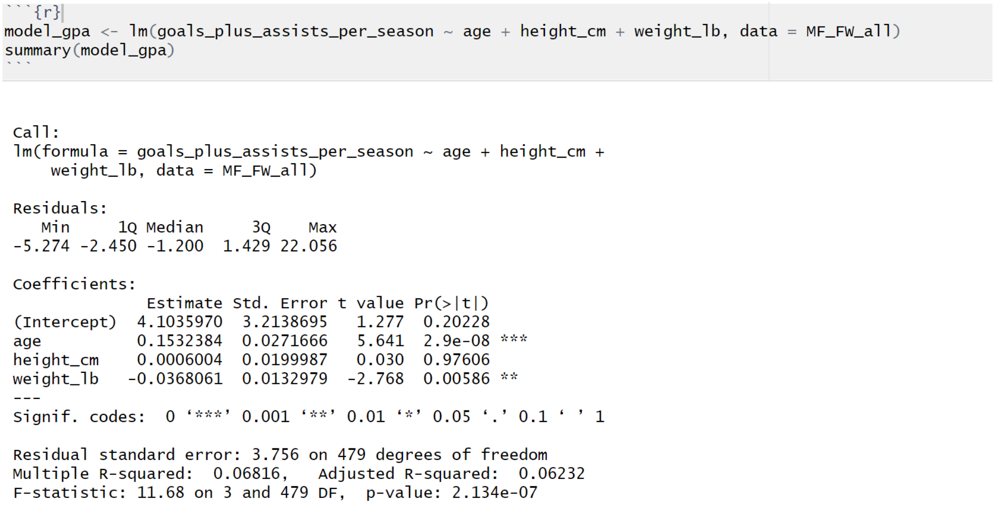


Figure 17: Linear Regression Results



~~~
```{r}
model_gpa <- lm(goals_plus_assists_per_season ~ age + height_cm + weight_lb, data = MF_FW_all)
summary(model_gpa)
```
~~~

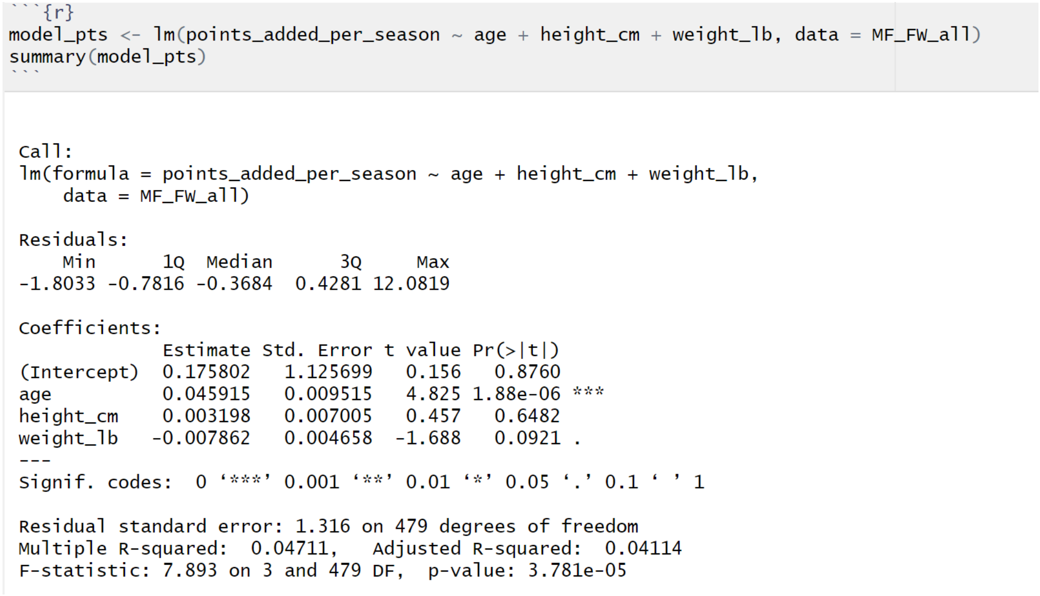
```
Call:
lm(formula = goals_plus_assists_per_season ~ age + height_cm + 
    weight_lb, data = MF_FW_all)

Residuals:
   Min     1Q Median     3Q    Max 
-5.274 -2.450 -1.200  1.429 22.056 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.1035970  3.2138695   1.277  0.20228    
age          0.1532384  0.0271666   5.641  2.9e-08 ***
height_cm    0.0006004  0.0199987   0.030  0.97606    
weight_lb   -0.0368061  0.0132979  -2.768  0.00586 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.756 on 479 degrees of freedom
Multiple R-squared:  0.06816,	Adjusted R-squared:  0.06232 
F-statistic: 11.68 on 3 and 479 DF,  p-value: 2.134e-07
```

Figure 18: Linear Regression Formula for Goals + Assists per Season



~~~
```{r}
model_pts <- lm(points_added_per_season ~ age + height_cm + weight_lb, data = MF_FW_all)
summary(model_pts)
```
~~~

```
Call:
lm(formula = points_added_per_season ~ age + height_cm + weight_lb, 
    data = MF_FW_all)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8033 -0.7816 -0.3684  0.4281 12.0819 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.175802   1.125699   0.156   0.8760    
age          0.045915   0.009515   4.825 1.88e-06 ***
height_cm    0.003198   0.007005   0.457   0.6482    
weight_lb   -0.007862   0.004658  -1.688   0.0921 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.316 on 479 degrees of freedom
Multiple R-squared:  0.04711,	Adjusted R-squared:  0.04114 
F-statistic: 7.893 on 3 and 479 DF,  p-value: 3.781e-05
```

Figure 19: Linear Regression Formula for Points Added/Season

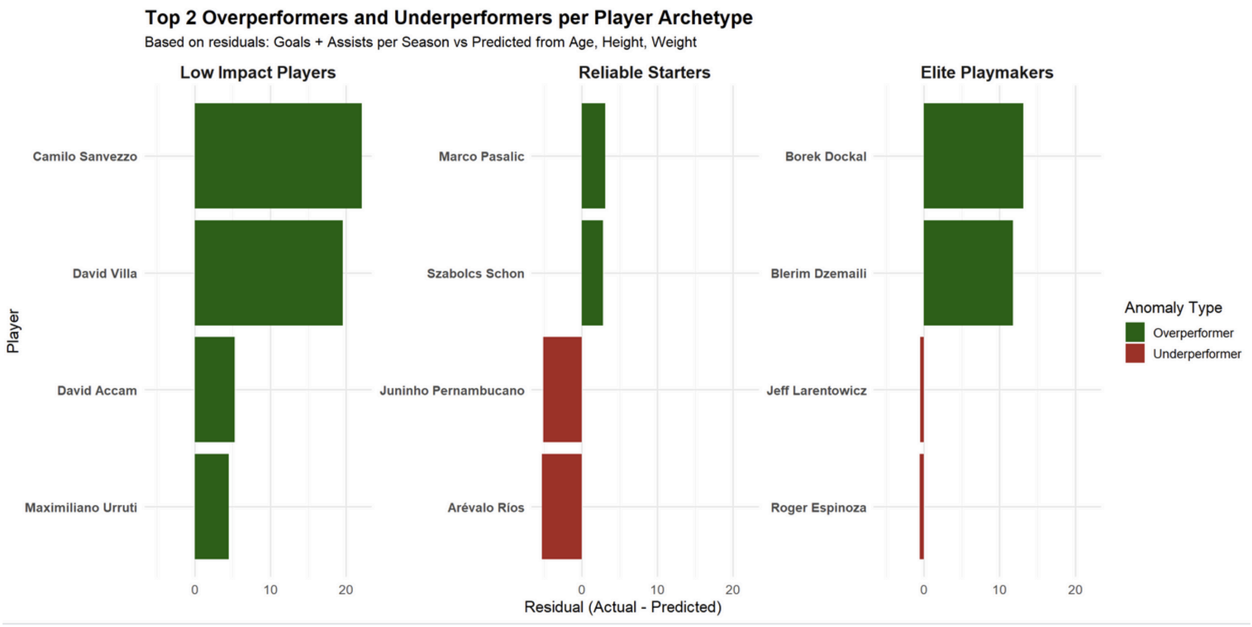


Figure 20: Overperformers and Underperformers for Goals + Assists per Season

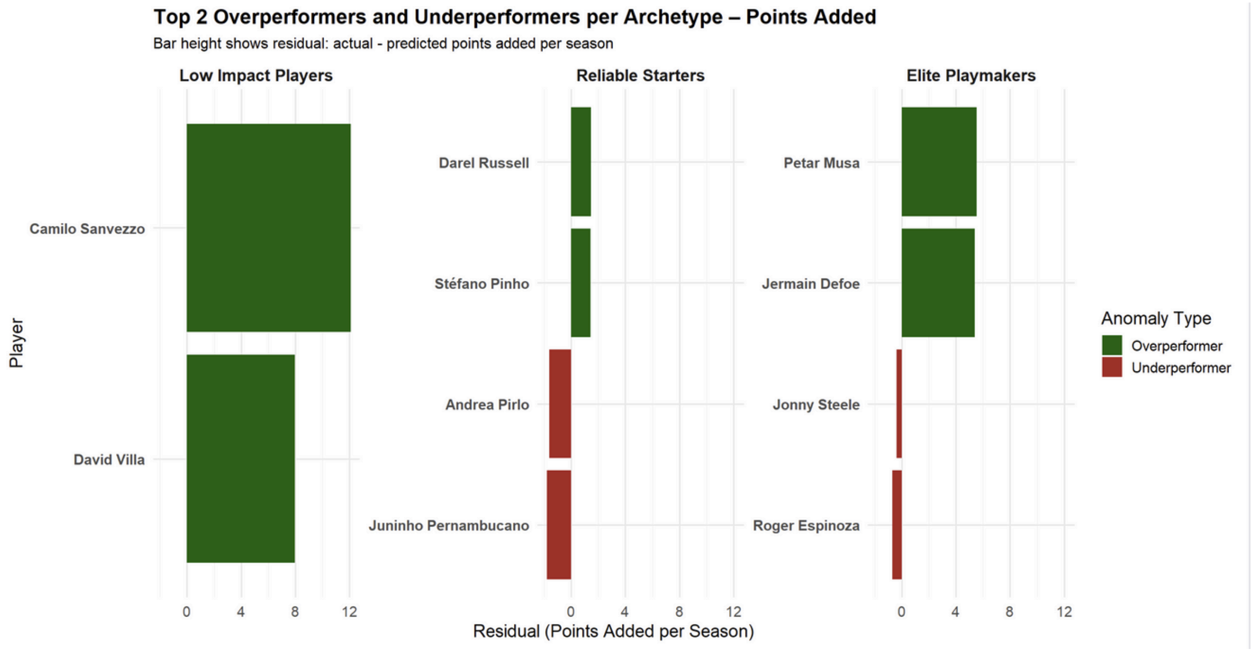
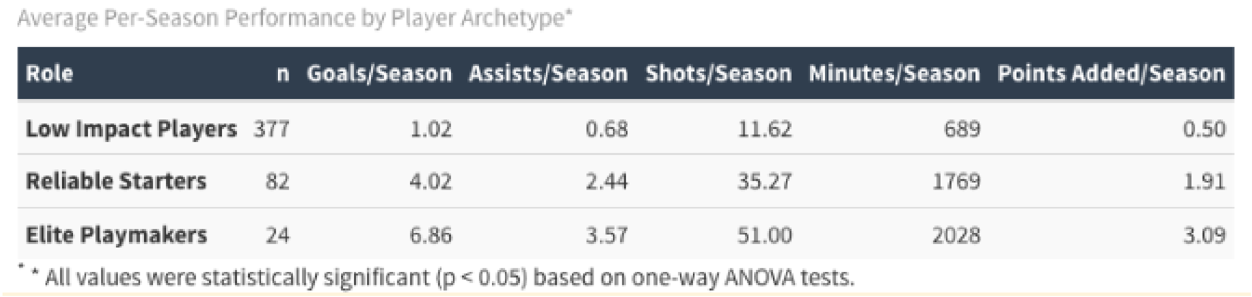


Figure 21: Overperformers and Underperformers for Points Added/Season

Key Tables:

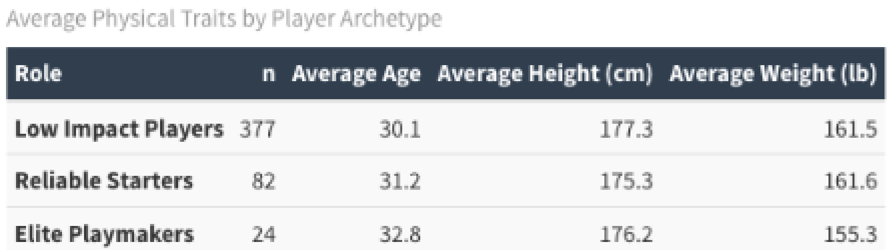


Average Per-Season Performance by Player Archetype*

Role	n	Goals/Season	Assists/Season	Shots/Season	Minutes/Season	Points Added/Season
Low Impact Players	377	1.02	0.68	11.62	689	0.50
Reliable Starters	82	4.02	2.44	35.27	1769	1.91
Elite Playmakers	24	6.86	3.57	51.00	2028	3.09

* * All values were statistically significant ($p < 0.05$) based on one-way ANOVA tests.

Table 1: Average Per-Season Performance by Player-Archetype



Average Physical Traits by Player Archetype

Role	n	Average Age	Average Height (cm)	Average Weight (lb)
Low Impact Players	377	30.1	177.3	161.5
Reliable Starters	82	31.2	175.3	161.6
Elite Playmakers	24	32.8	176.2	155.3

Table 2: Average Physical Traits by Player Archetype