Introduction:

Football is not just a game; it is a global phenomenon that combines athletic skill, tactical strategy, and moments of individual brilliance. Central to the excitement of football is the act of scoring a goal—the culmination of a series of decisions, movements, and actions on the field. Understanding how **goals are created**, the **factors that influence their success**, and the **characteristics of players and teams** that consistently convert opportunities into goals offers profound insights into the dynamics of the sport.

This project, **"The Anatomy of a Goal"**, examines detailed event-level football data to uncover patterns and determinants of goal-scoring. The dataset captures shots, assists, player attributes, and situational context across matches, allowing for an in-depth exploration of how goals occur. By analyzing variables such as shot location, body part used, assist method, match situation, and player skill, this study aims to go beyond mere goal counts to understand the mechanisms behind goal conversion.

The findings from this project have relevance not only for coaches and analysts seeking to **optimize team performance** but also for **fans, media, and sports strategists** interested in the finer details of goal creation. With the increasing availability of granular football data, this analysis leverages **statistical modeling and visualization** to dissect the anatomy of a goal.

Objectives:

The primary objective of this project is to perform a comprehensive data-driven and statistical analysis of goal-scoring in football matches. The specific aims of the study include:

- **Explore and visualize** the overall distribution of shots and goals.
- Investigate the **influence** of shot characteristics, including location on the pitch, body part used, and assist type, on goal conversion rates.
- Examine how **situational factors**, such as set pieces, corners, or open play, impact the likelihood of scoring.
- Identify **elite players and teams** with high goal conversion rates and analyze their performance across different conditions.
- Analyze **multivariate relationships**, such as body part and location or assist method and situation, to uncover patterns of successful goal-scoring combinations.
- Build **predictive models,** including **logistic regression**, to estimate the probability of a shot resulting in a goal.
- Evaluate **model performance** using metrics such as **ROC curves, confusion matrices, and pseudo R-squared** values.
- Provide **actionable insights** for football analysts, coaches, and teams to understand and enhance **goal-scoring efficiency**.

Data Description:

The analysis is based on a detailed dataset capturing event-level information from professional football matches. The dataset provides a granular view of **9,074 games,** totaling **941,009 events** from the biggest five European leagues—**England, Spain, Germany, Italy, and France**—spanning the **2011/2012 season to the 2016/2017 season** (as of 25.01.2017). Each row represents a single event, with attributes describing the player, team, situation, and outcome.

Key variables include:

- **time:** The minute in the match when the shot occurred.
- **side:** Indicates whether the shot was taken by the home or away team. Encoded as "Home" or "Away".
- **player:** The name of the player who attempted the shot.
- **event_team:** The team that attempted the shot.
- **is_goal:** A binary indicator representing whether the shot resulted in a goal (1 = goal, 0 = no goal).
- **location:** The area on the pitch from which the shot was taken. This captures the tactical positioning of the shot.

1. Attacking half
2. Defensive half
3. Centre of the box
4. Left wing
5. Right wing
6. Difficult angle and long range
7. Difficult angle on the left
8. Difficult angle on the right
9. Left side of the box
10. Left side of the six yard box
11. Right side of the box
12. Right side of the six yard box
13. Very close range
14. Penalty spot
15. Outside the box
16. Long range
17. More than 35 yards
18. More than 40 yards
19. Not recorded

- **bodypart:** The body part used to take the shot, categorized as "Right Foot," "Left Foot," or "Head."
- **assist_method:** The type of pass or setup leading to the shot, classified as "None," "Pass," "Cross," "Headed Pass," or "Through Ball."
- **situation:** The match situation during which the shot occurred, such as "Open Play," "Set Piece," "Corner," or "Free Kick."

This dataset allows for a granular examination of goal-scoring patterns at both the player and team levels. By combining situational, positional, and player-level attributes, it provides the foundation for both descriptive analysis and predictive modeling of goal conversion.

The dataset was loaded in R and further statistical analysis were performed accordingly.

```
[1] 941009     22
'data.frame': 941009 obs. of  22 variables:
 $ id_odsp      : chr  "UFot0hit/" "UFot0hit/" "UFot0hit/" "UFot0hit/" ...
 $ id_event     : chr  "UFot0hit1" "UFot0hit2" "UFot0hit3" "UFot0hit4" ...
 $ sort_order   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ time         : int  2 4 4 7 7 9 10 11 11 13 ...
 $ text         : chr  "Attempt missed. Mladen Petric (Hamburg) left footed shot from the left side of the box
 $ event_type   : int  1 2 2 3 8 10 2 8 3 3 ...
 $ event_type2  : int  12 NA NA NA NA NA NA NA NA NA ...
 $ side         : int  2 1 1 1 2 2 2 1 2 2 ...
 $ event_team   : chr  "Hamburg SV" "Borussia Dortmund" "Borussia Dortmund" "Borussia Dortmund" ...
 $ opponent     : chr  "Borussia Dortmund" "Hamburg SV" "Hamburg SV" "Hamburg SV" ...
```

```
 $ player       : chr  "mladen petric" "dennis diekmeier" "heiko westermann" "sven bender" ...
 $ player2      : chr  "gokhan tore" "dennis diekmeier" "heiko westermann" NA ...
 $ player_in    : chr  NA NA NA NA ...
 $ player_out   : chr  NA NA NA NA ...
 $ shot_place   : int  6 NA NA NA NA NA NA NA NA NA ...
 $ shot_outcome : int  2 NA NA NA NA NA NA NA NA NA ...
 $ is_goal      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ location     : int  9 NA NA NA 2 NA NA 2 NA NA ...
 $ bodypart     : int  2 NA NA NA NA NA NA NA NA NA ...
 $ assist_method: int  1 0 0 0 0 0 0 0 0 0 ...
 $ situation    : int  1 NA NA NA NA NA NA NA NA NA ...
 $ fast_break   : int  0 0 0 0 0 0 0 0 0 0 ...
```

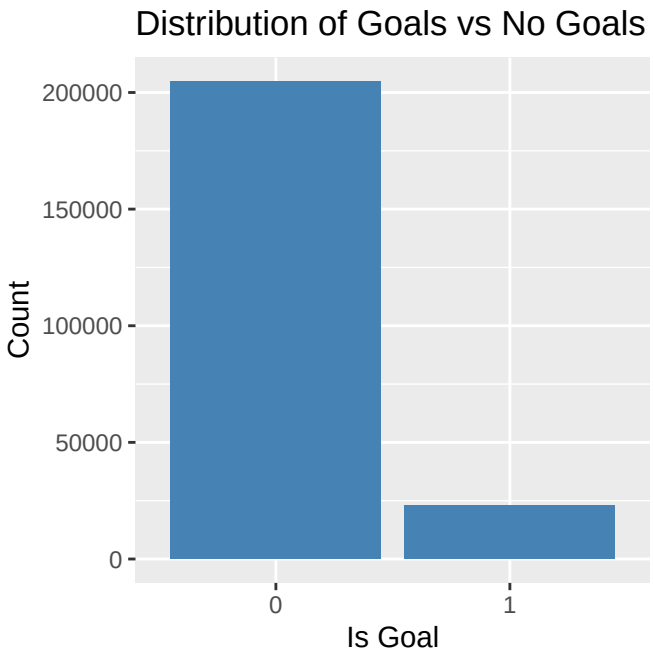
Feature Engineering:

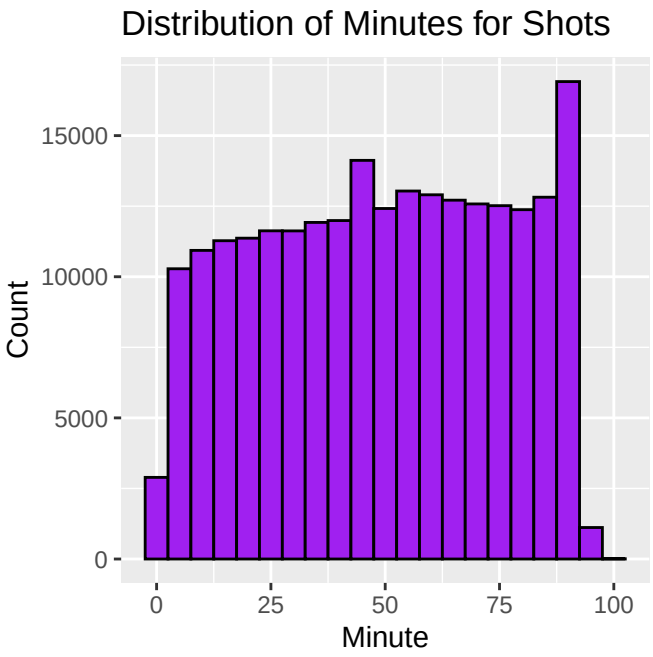
- **elite:** A derived binary variable indicating whether the player is among elite performers (top conversion rate with minimum 450 shots attempted).
- **Additional variables:** The dataset initially contained auxiliary columns like player substitutions, event IDs, and textual descriptions, which were removed for analysis to focus on relevant shot and goal attributes.
- Converting to suitable factors with suitable levels of various categorical variables.

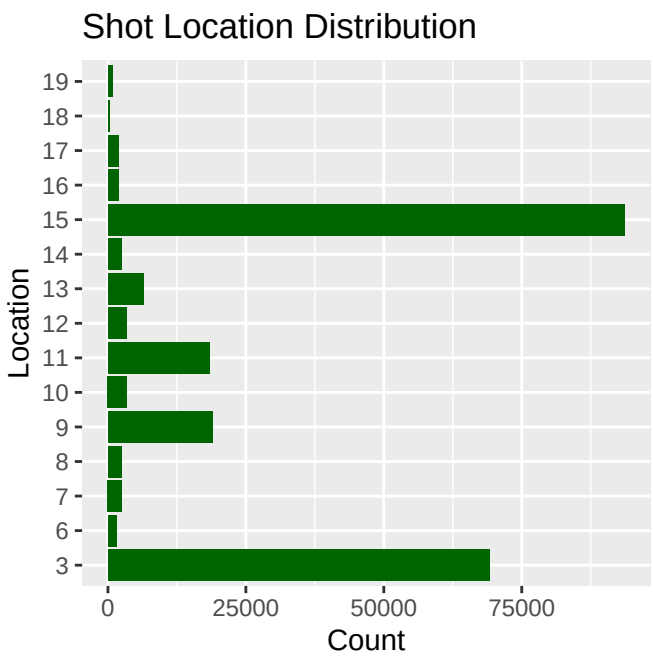
```
'data.frame': 227452 obs. of  14 variables:
 $ time         : int  2 14 17 19 20 25 25 26 28 29 ...
 $ side         : Factor w/ 2 levels "Home","Away": 2 1 1 1 2 1 1 1 1 1 ...
 $ event_team   : chr  "Hamburg SV" "Borussia Dortmund" "Borussia Dortmund" "Borussia Dortmund" ...
 $ opponent     : chr  "Borussia Dortmund" "Hamburg SV" "Hamburg SV" "Hamburg SV" ...
 $ player       : chr  "mladen petric" "shinji kagawa" "kevin grosskreutz" "mats hummels" ...
 $ shot_place   : int  6 13 4 2 2 7 2 5 9 5 ...
 $ shot_outcome : int  2 2 1 3 3 4 3 1 2 1 ...
 $ is_goal      : int  0 0 1 0 0 0 0 0 0 1 ...
 $ location     : int  9 15 9 15 15 3 15 3 9 3 ...
 $ bodypart     : Factor w/ 3 levels "Right Foot","Left Foot",..: 2 1 2 1 1 1 2 3 1 1 ...
 $ assist_method: Factor w/ 5 levels "None","Pass",..: 2 2 2 1 1 2 1 3 2 2 ...
 $ situation    : Factor w/ 4 levels "Open Play","Set Piece",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ fast_break   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ elite        : num  0 0 0 0 0 0 0 0 0 0 ...
 - attr(*, "na.action")= 'omit' Named int [1:713557] 2 3 4 5 6 7 8 9 10 11 ...
  ..- attr(*, "names")= chr [1:713557] "2" "3" "4" "5" ...
```

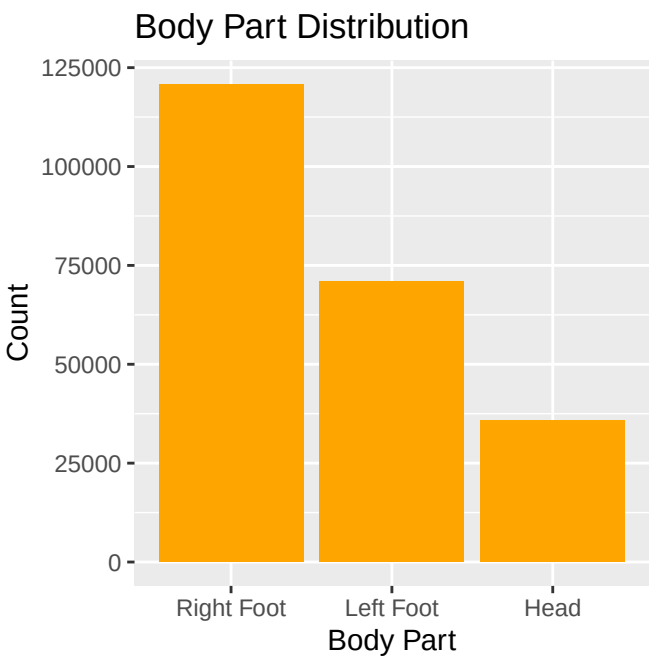
Exploratory Data Analysis (EDA):

Some Basic Plots and Takeways:





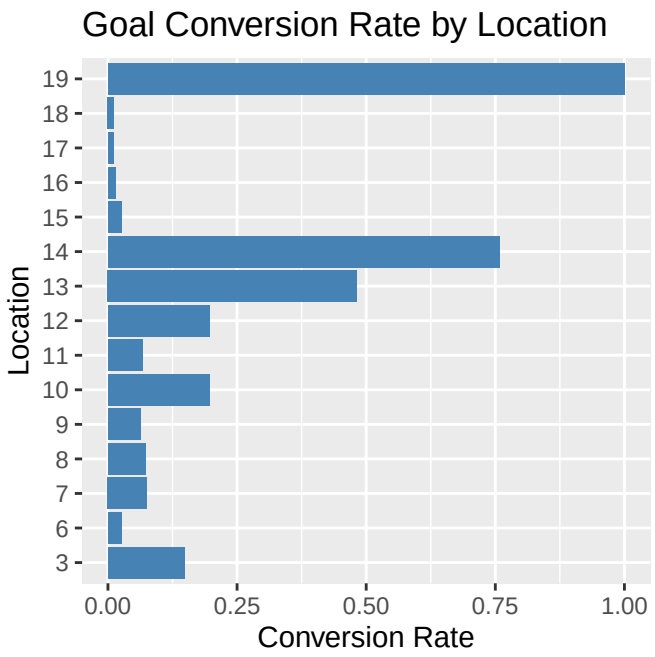


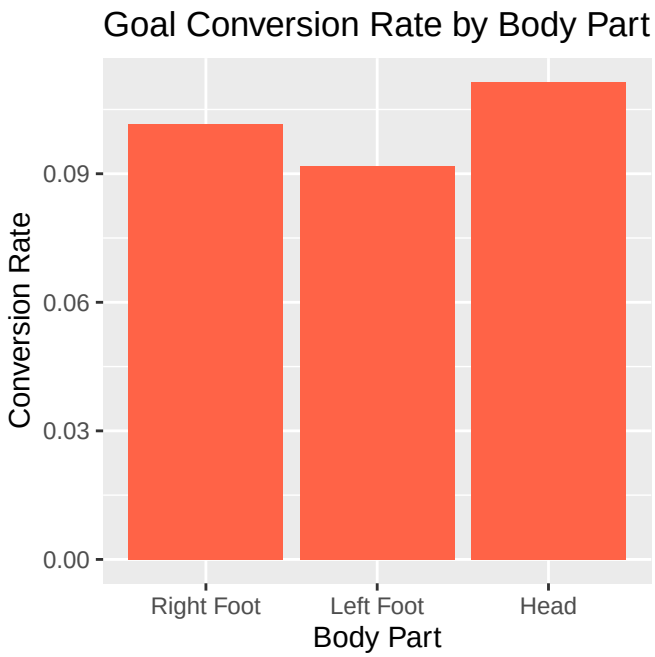


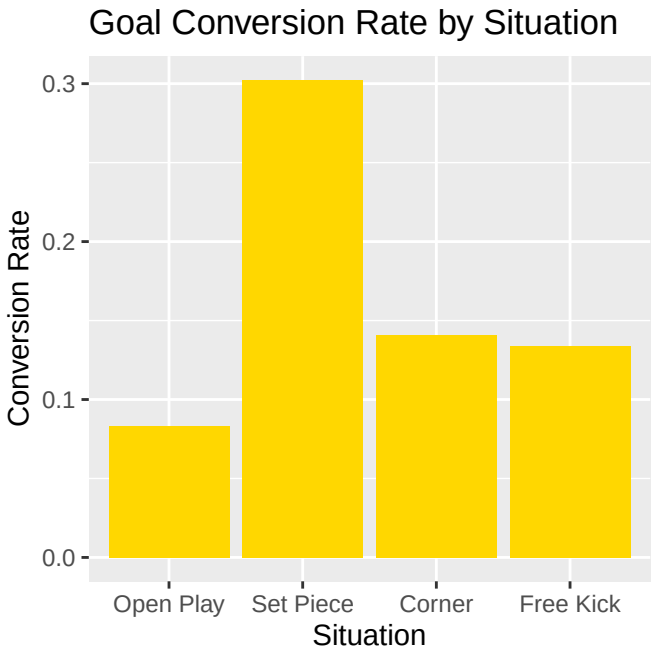
Key Takeaways:

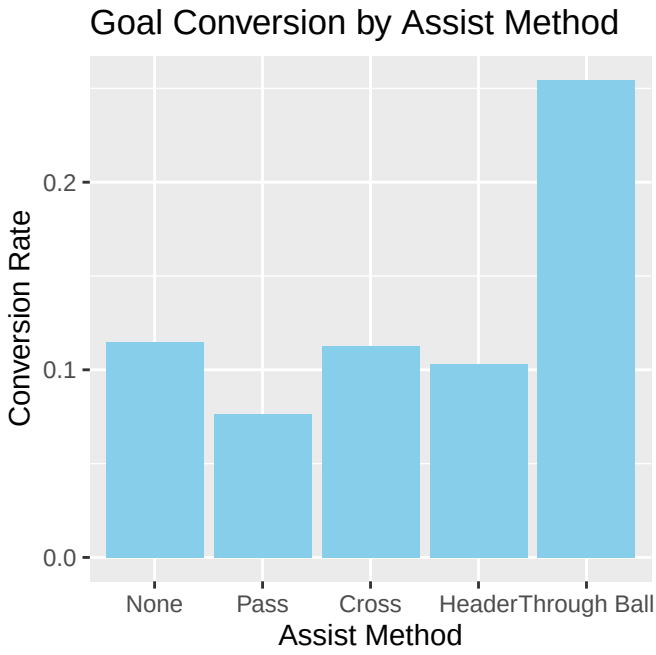
- **Goals are rare gems** – Only a small fraction of shots result in goals, highlighting the difficulty of scoring.
- **Late drama is real** – Shots peak towards the final minutes, showing how matches heat up near the end.
- **Hot zones matter** – A few specific shot locations dominate attempts, suggesting players repeatedly target high-probability areas.
- **Right foot rules** – Most shots come from the right foot, with the left foot and headers trailing behind.

Bivariate Relationships:







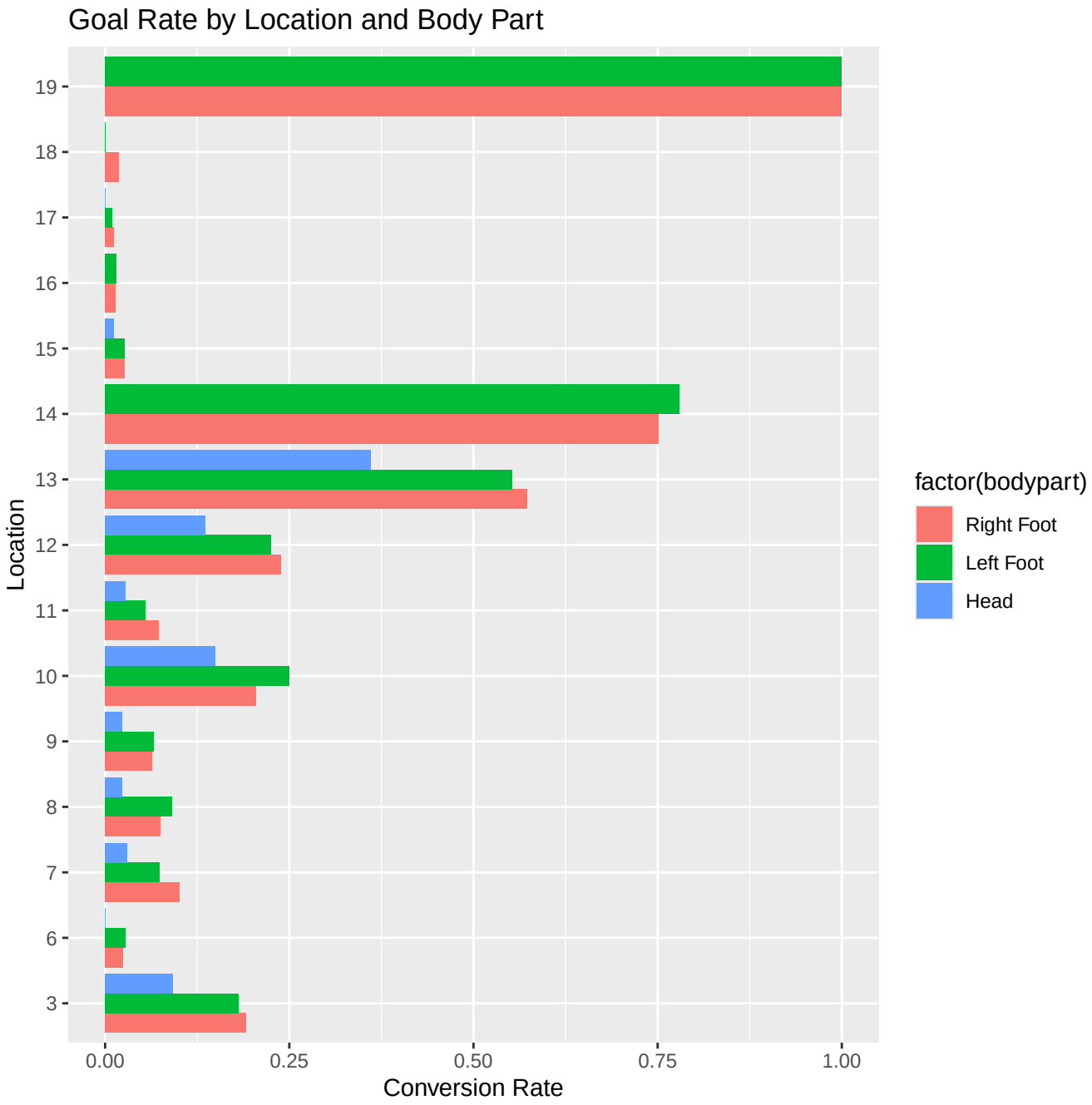


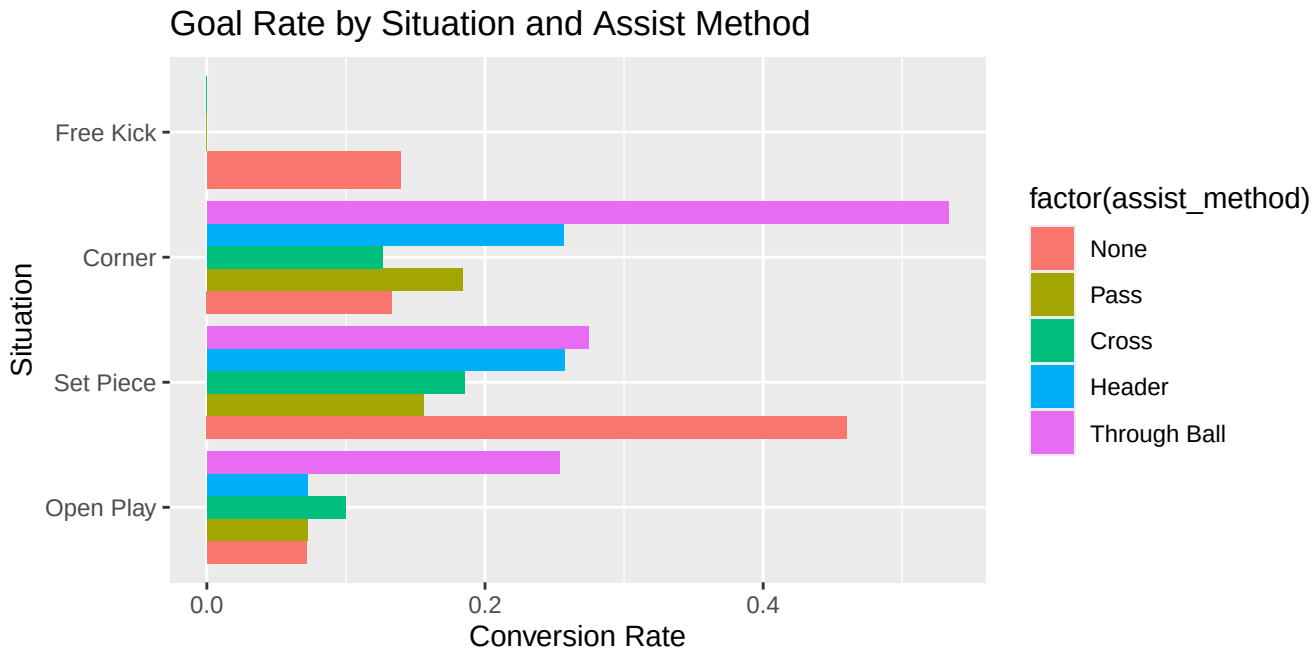
Key Takeaways:

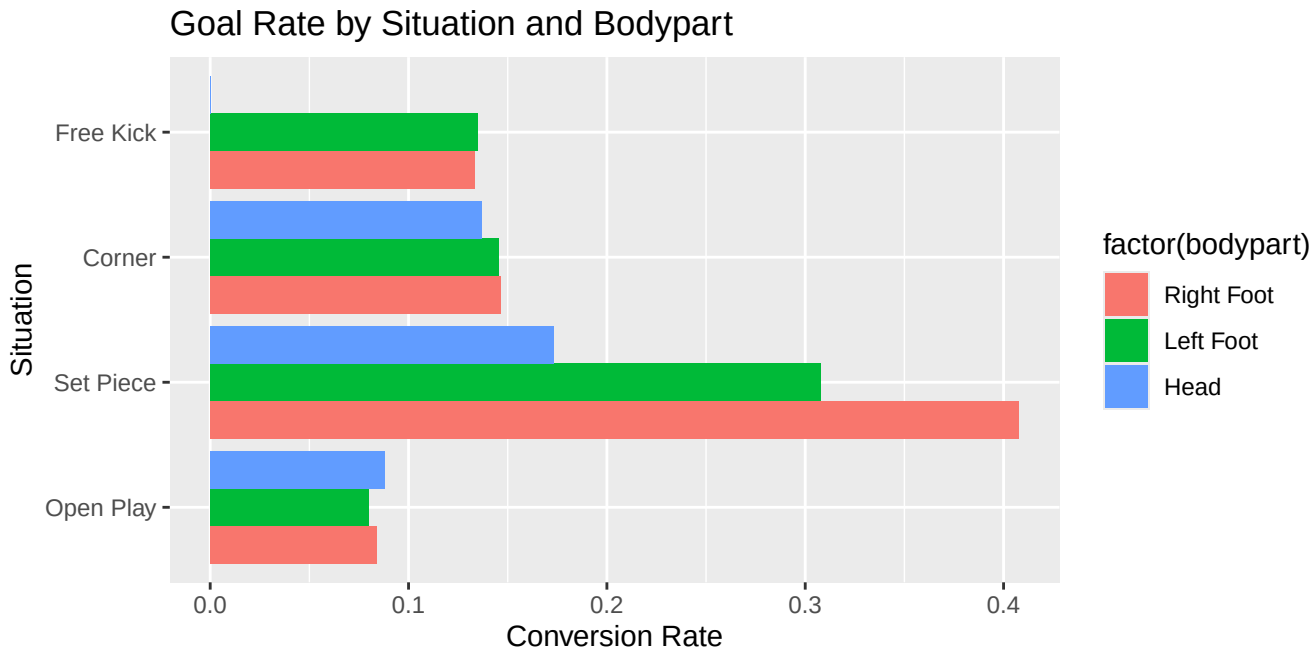
- **Sweet spots decide games** – Certain shot locations (like 14 and 19) boast conversion rates close to or above 75%, showing they're golden zones for scoring.
- **Headers pack a punch** – Despite fewer attempts, headers have the highest conversion rate, proving aerial strength is a deadly weapon.
- **Set pieces are gold mines** – With the highest conversion rate among situations, set pieces stand out as the most reliable path to goals.
- **Open play struggles** – Shots in open play convert less often, underlining the importance of structured attacking patterns.

- **Through balls unlock defenses** – They dominate the assist methods with the highest success rate, confirming their role as game-changers.
- **Simple passes lag behind** – Regular passes or crosses convert less, suggesting teams need sharper creativity to break through.

Multivariate Relationships:



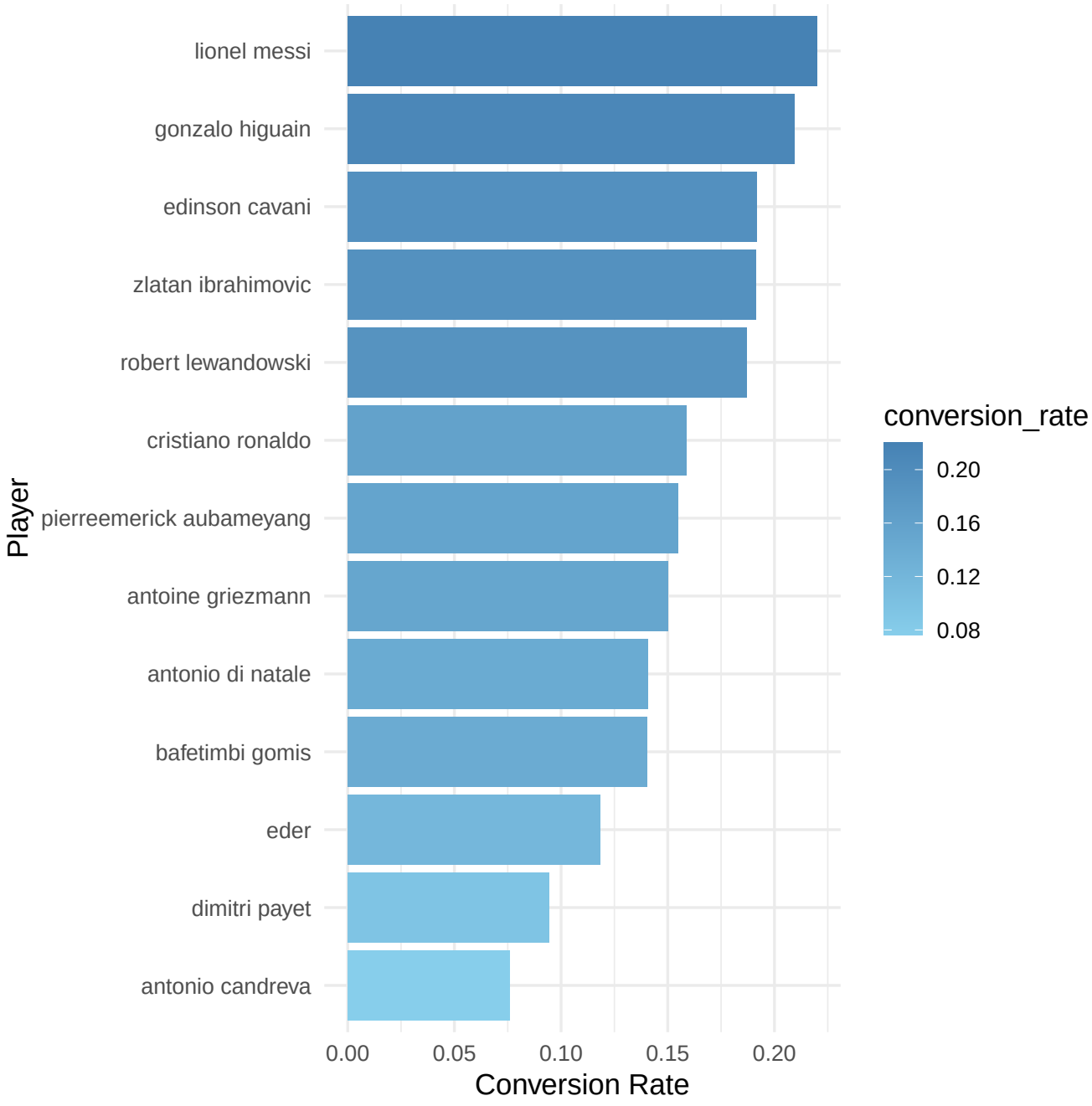




Key Takeaways:

- **Location trumps all** – Certain zones (like 14 and 19) dominate conversion regardless of body part, making them the most lethal scoring areas.
- **Through balls are deadly in open play** – They massively boost conversion rates compared to simple passes or crosses.
- **Set pieces + right foot = gold combo** – Set pieces finished with the right foot show the highest efficiency, making them a key tactical weapon.
- **Headers shine in corners** – Aerial duels during corners consistently yield strong goal rates, highlighting the power of good delivery.
- **Open play lags** – Goals from open play remain harder to come by.

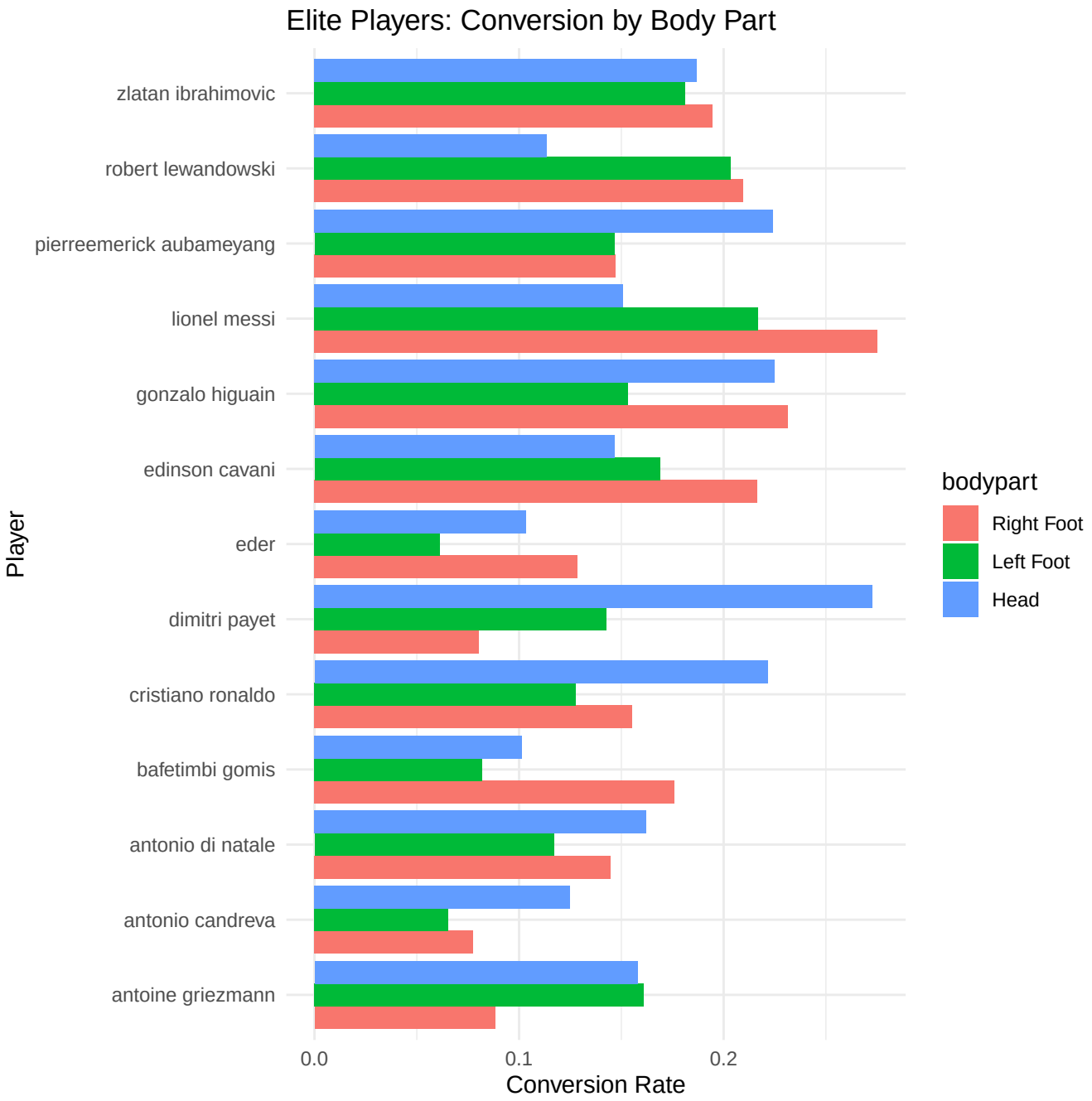
Which Players Boast the Best Goal Conversion Rates? (min 450 shots attempted)

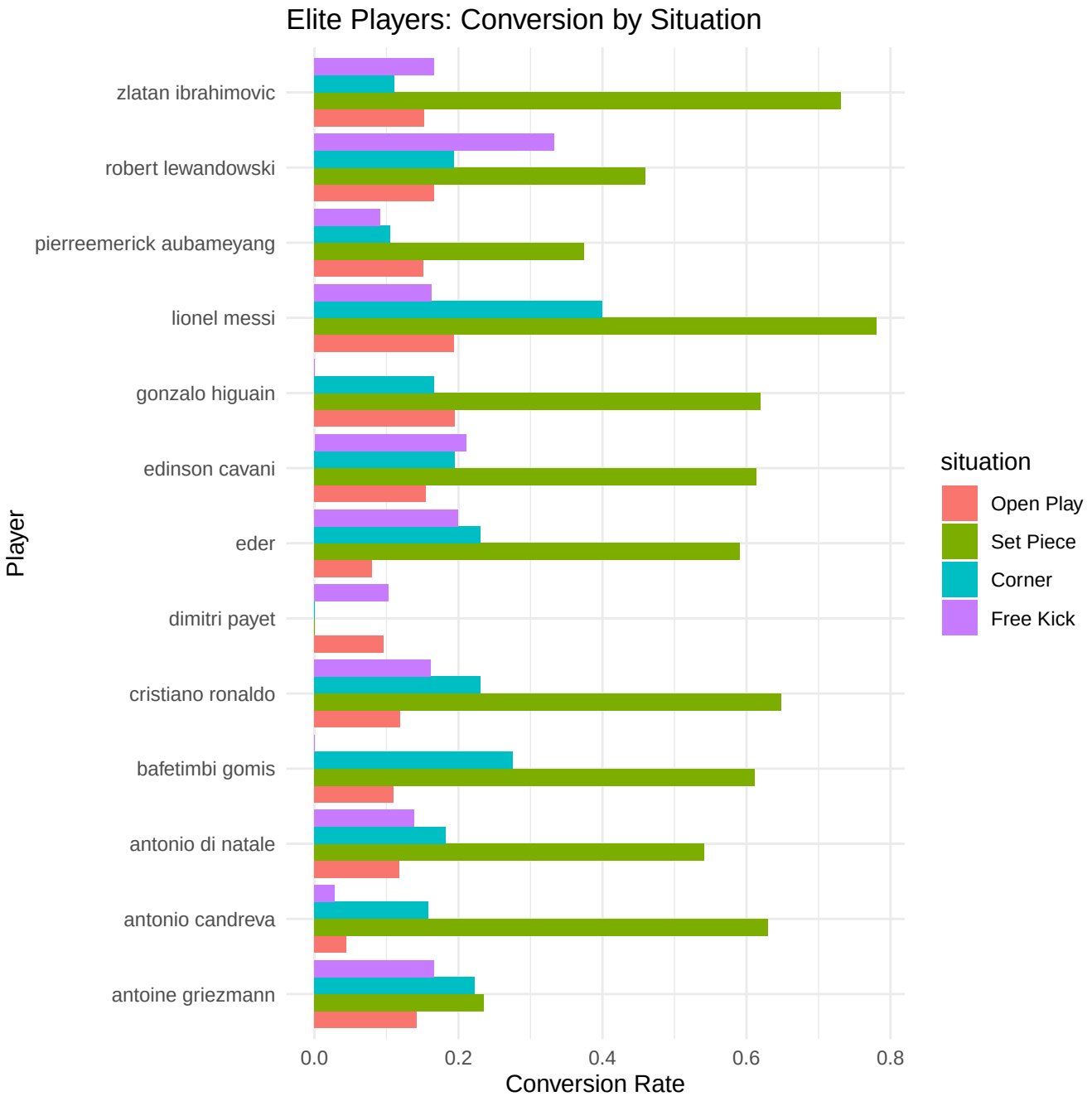


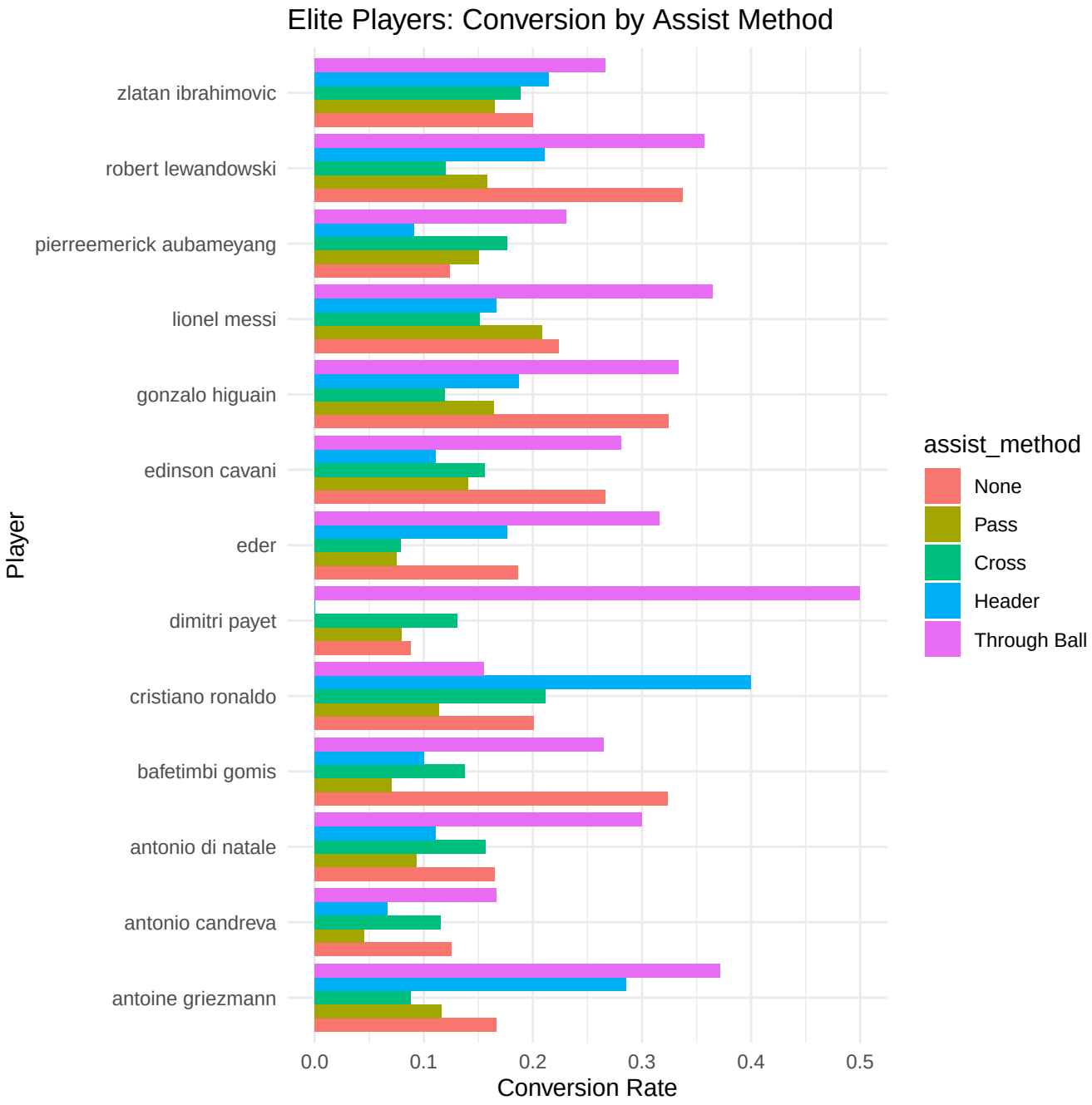
Key Takeaways:

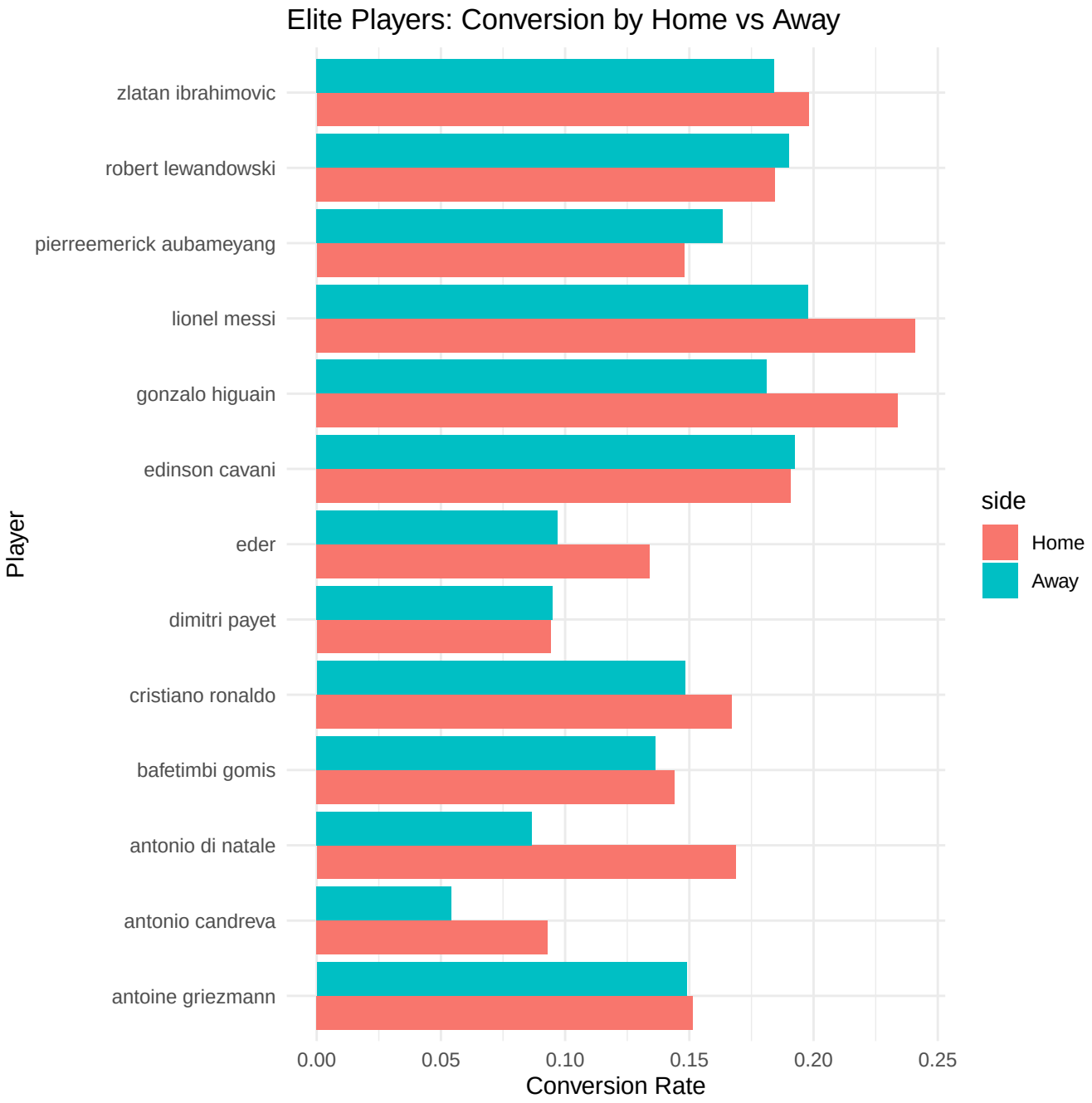
- **Messi leads the pack** – Lionel Messi tops the list with the highest goal conversion rate.
- **Clinical strikers dominate** – Higuaín, Cavani, and Ibrahimović sit right behind, showing elite efficiency.
- **Lewandowski and Ronaldo differ** – Both score plenty, but Lewandowski edges Ronaldo in conversion sharpness.
- **Speedsters deliver** – Aubameyang and Griezmann turn chances into goals at a strong rate.
- **Midfield Magic**– Players like Candreva and Payet although primarily midfielders make it into the top 10, highlighting the clinical ability of creators.

Player Conversion Rate across factors:




Elite Players: Conversion by Situation
zlatan ibrahimovic
robert lewandowski
pierreemerick aubameyang
lionel messi
gonzalo higuain
edinson cavani
eder
dimitri payet
cristiano ronaldo
bafetimbi gomis
antonio di natale
antonio candreva
antoine griezmann
Player
0.0
0.2
0.4
0.6
0.8
Conversion Rate
situation
Open Play
Set Piece
Corner
Free Kick


Elite Players: Conversion by Assist Method
zlatan ibrahimovic
robert lewandowski
pierreemerick aubameyang
lionel messi
gonzalo higuain
edinson cavani
eder
dimitri payet
cristiano ronaldo
bafetimbi gomis
antonio di natale
antonio candreva
antoine griezmann
Player
0.0
0.1
0.2
0.3
0.4
0.5
Conversion Rate
assist_method
None
Pass
Cross
Header
Through Ball


Elite Players: Conversion by Home vs Away
zlatan ibrahimovic
robert lewandowski
pierreemerick aubameyang
lionel messi
gonzalo higuain
edinson cavani
eder
dimitri payet
cristiano ronaldo
bafetimbi gomis
antonio di natale
antonio candreva
antoine griezmann
Player
0.00
0.05
0.10
0.15
0.20
0.25
Conversion Rate
side
Home
Away

Key Takeaways:

1. Overall Conversion Efficiency

- Messi leads with exceptional consistency in converting shots across all zones, showing efficiency not just in volume but also in accuracy.
- Lewandowski and Aubameyang also post high returns, though with greater reliance on specific situations.

2. Shot Location Patterns

- Messi and Higuaín convert heavily from central areas inside the box, thriving in tight spaces.
- Ronaldo maintains strong output from both inside and outside the box, offering unpredictability from distance.

3. Body Part Dependence

- Messi and Higuaín are right-foot dominant, with over 70% of their goals coming from the preferred side.
- Aubameyang and Lewandowski show balanced output, while Ronaldo and Ibrahimović remain aerial specialists.

4. Aerial Threats

- Ronaldo, Aubameyang, and Ibrahimović provide major heading threats, giving tactical width more value when they play.
- Their presence forces defenses to adjust, freeing space for second-line runners.

5. Assist Method Impact

- Through balls significantly raise conversion for Messi, Higuaín, and Aubameyang, showing the value of vertical passing.
- Crosses fuel Ronaldo's and Ibrahimović's efficiency, reinforcing the need for strong wide service.

6. Set-Piece Influence

- Cavani and Lewandowski excel in dead-ball phases, particularly corners and free-kick rebounds.
- Ronaldo's set-piece output adds a second dimension to his open-play threat.

7. Open Play Mastery

- Messi stands out for open-play finishing, rarely requiring structured setups.
- Griezmann's consistency in open play makes him an adaptable forward in fast-flowing systems.

8. Home vs Away Consistency

- Messi, Higuaín, and Ibrahimović post stronger home numbers, thriving with crowd momentum.
- Aubameyang and Griezmann remain steady across venues, showing mental resilience and adaptability.

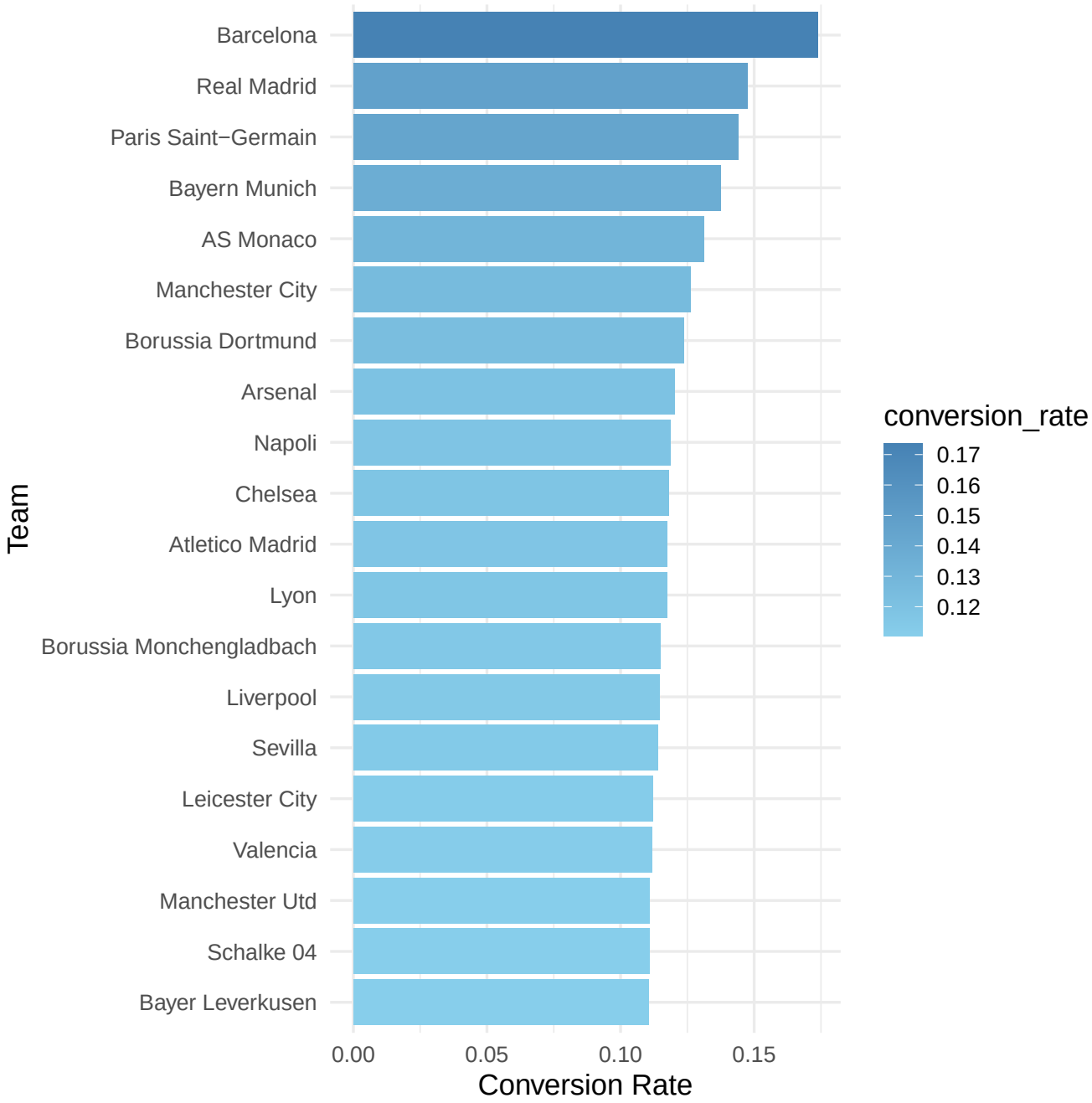
9. Adaptability to Supply

- Ronaldo adapts seamlessly to both direct crosses and quick counters.
- Ibrahimović remains versatile across all assist types, showcasing technical adaptability despite age.

10. Tactical Implication for Teams

- Messi/Higuaín: thrive with through-ball driven central attacks.
- Ronaldo/Aubameyang: demand wide service and aerial delivery.
- Cavani/Lewandowski: optimize set-piece and structured build-up situations.

Which Teams Boast the Best Goal Conversion Rates?



- **Barcelona – Masters of Conversion:** With the highest conversion rate, Barcelona showcase unmatched attacking efficiency, turning chances into goals at a rate no other team matches.
- **Real, PSG, Bayern** — Ruthless Finishers: Real Madrid, PSG, and Bayern Munich sit just behind, proving why these clubs dominate both domestically and in Europe—they don't just create chances, they finish them.
- **Monaco & Dortmund Break the Elite Barrier:** AS Monaco and Dortmund rival Europe's giants, highlighting systems that prioritize attacking fluidity and precision in the final third.
- **Premier League Giants Still Wasteful:** English clubs like Arsenal, Chelsea, and Liverpool cluster mid-table in conversion rate, showing attacking strength but also inefficiency compared to Spain's top two.
- **Small Margins, Big Titles:** From Atletico Madrid downward, conversion rates dip noticeably, underlining the fine margins at the top: a few percentage points in finishing separates title contenders from nearly-there sides.

Logistic Regression:

We use logistic regression to model the probability of scoring a goal:

$$logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 C X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon_i$$

- Y = is_goal (1 if goal scored, 0 otherwise)
- X_1 = location (shot location on the pitch)
- X_2 = bodypart (foot, head, etc.)
- X_3 = situation (open play, set-piece, penalty, etc.)
- X_4 = assist_method (cross, through ball, no assist, etc.)
- X_5 = side (Home/Away)
- X_6 = time (minute of the match when shot was taken)
- X_7 = elite (indicator for elite vs. non-elite player)
- ε = Random error

Now we fit the logistic regression model to our given data. We estimate the parameters using Maximum Likelihood Estimation (MLE). In R, we use the glm() function with the family set to binomial to facilitate the calculations.

```
Call:
glm(formula = is_goal ~ location + bodypart + situation + assist_method +
    side + time + elite, family = binomial, data = df)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.312e+00  4.335e-02 -30.268  < 2e-16 ***
location6                 -1.946e+00  1.633e-01 -11.912  < 2e-16 ***
location7                 -1.066e+00  7.694e-02 -13.860  < 2e-16 ***
location8                 -1.157e+00  8.017e-02 -14.436  < 2e-16 ***
location9                 -1.249e+00  3.275e-02 -38.127  < 2e-16 ***
location10                 3.779e-01  4.527e-02   8.349  < 2e-16 ***
location11                -1.213e+00  3.244e-02 -37.382  < 2e-16 ***
location12                 3.656e-01  4.576e-02   7.991 1.34e-15 ***
location13                 1.751e+00  2.832e-02  61.819  < 2e-16 ***
location14                 2.002e+00  5.862e-02  34.160  < 2e-16 ***
location15                -2.050e+00  2.438e-02 -84.107  < 2e-16 ***
location16                -2.633e+00  1.910e-01 -13.784  < 2e-16 ***
location17                -2.880e+00  2.150e-01 -13.398  < 2e-16 ***
location18                -2.735e+00  5.815e-01  -4.704 2.55e-06 ***
location19                 3.296e+01  1.515e+02   0.218  0.82780
bodypartLeft Foot         -4.852e-02  1.849e-02  -2.624  0.00868 **
bodypartHead              -6.734e-01  2.697e-02 -24.965  < 2e-16 ***
situationSet Piece         7.246e-01  3.219e-02  22.507  < 2e-16 ***
situationCorner            4.441e-01  2.701e-02  16.443  < 2e-16 ***
situationFree Kick        -1.389e+01  5.714e+01  -0.243  0.80797
assist_methodPass          1.168e-01  2.149e-02   5.437 5.41e-08 ***
assist_methodCross        -3.333e-01  2.810e-02 -11.862  < 2e-16 ***
assist_methodHeader       -2.417e-01  4.466e-02  -5.412 6.23e-08 ***
assist_methodThrough Ball  1.069e+00  3.405e-02  31.411  < 2e-16 ***
sideAway                  -3.904e-02  1.573e-02  -2.482  0.01307 *
```

```
time                        1.054e-03  2.955e-04   3.568  0.00036 ***
eliteNo                    -3.334e-01  3.596e-02  -9.272  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 147990  on 227451  degrees of freedom
Residual deviance: 117128  on 227425  degrees of freedom
AIC: 117182

Number of Fisher Scoring iterations: 16
```

We interpret the following from the R output:

Null deviance = 147,990 on 227,451 df; Residual deviance = 117,128 on 227,425 df; AIC = 117,182:

- The large drop in deviance suggests the model explains a substantial amount of variation in whether a shot results in a goal.
- The lower AIC indicates a reasonably good fit.

Intercept = -1.312, p-value < 2e-16:

- When all predictors are at their baseline, the log-odds of scoring are -1.312.
- This corresponds to a low baseline probability of a goal. The effect is statistically significant.

Location effects:

- Several locations show strong and significant effects.
- Example: Location 13 (Estimate = 1.751, p < 2e-16) means shots from this location have much higher odds of scoring compared to the baseline.
- Conversely, Location 15 (Estimate = -2.050, p < 2e-16) indicates far lower odds of scoring from this area.

Body part:

- Head (Estimate = -0.673, p < 2e-16) significantly decreases the odds of scoring compared to the baseline (Right Foot).
- Left Foot (Estimate = -0.0485, p = 0.00868) has a small but statistically significant negative effect.

Situation:

- Set Piece (Estimate = 0.7246, p < 2e-16) and Corner (Estimate = 0.444, p < 2e-16) significantly increase the odds of scoring compared to open play.
- Free Kick (Estimate = -13.89, p = 0.808) shows no significant effect, likely due to very sparse data.

Assist method:

- Through Ball (Estimate = 1.069, p < 2e-16) substantially increases the odds of scoring.
- Cross (Estimate = -0.333, p < 2e-16) and Header Assist (Estimate = -0.242, p < 0.001) significantly reduce scoring odds relative to the baseline.
- Pass (Estimate = 0.117, p < 0.001) shows a modest positive effect.

Side (Away = -0.039, p = 0.013): Playing away slightly reduces the odds of scoring, and this effect is statistically significant.

Time (Estimate = 0.00105, p < 0.001): Each additional minute into the game slightly increases the odds of scoring, a statistically significant but small effect.

Elite (No = -0.333, p < 2e-16): Non-elite teams have significantly lower odds of scoring compared to elite teams.

Now we shall look to calculate **McFadden's Pseudo R^2** and its **p-value** for our model.

```
[1] "Log-likelihood of Null Model: -73995.2486"
[1] "Log-likelihood of Proposed Model: -58564.0534"
[1] "McFadden's Pseudo R-squared: 0.2085"
[1] "p-value for McFadden's Pseudo R-squared: 0"
```

Interpretation of Model Fit:

- The logistic regression model demonstrates a **substantial improvement** over the null model.
- The log-likelihood of the null model was estimated at –73995.25, whereas the proposed model achieved a higher log-likelihood of –58564.05, indicating a **markedly better fit**.
- The McFadden's pseudo $\mathbf{R}^2$ value of 0.2085 suggests that the model explains approximately 21% of the variation in the likelihood function, which falls within the range typically considered as **strong explanatory power** for logistic regression models.
- Furthermore, the likelihood ratio test yielded a p-value < 0.001, confirming that the improvement in fit is highly **statistically significant**.
- These results provide strong evidence that the chosen predictors—shot location, body part, situation, assist method, side, time, and elite status—collectively **contribute meaningfully** to explaining the probability of a shot resulting in a goal.

Model Performance Evaluation:

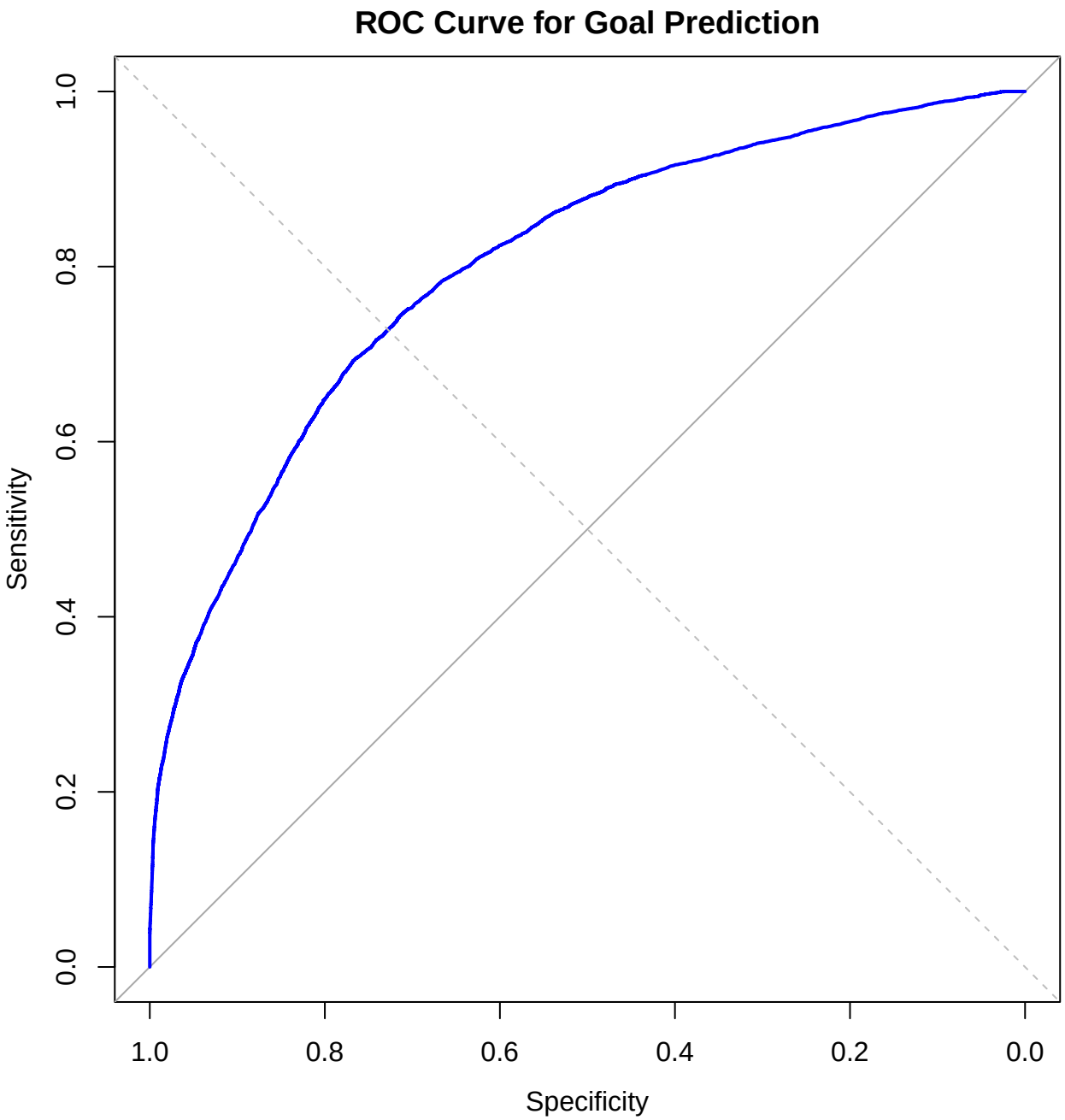
The steps we will follow to evaluate the performance of our logistic regression model are as follows:

- Split the dataset into training (70%) and testing (30%) sets using stratified sampling
- Fit a logistic regression model on the training set with predictors such as location, body part, situation, assist method, side, time, and elite status
- Predict the probability of a goal on the test set and classify outcomes as "Yes" (goal) or "No" (not goal)
- Evaluate classification performance using a confusion matrix
- Assess model discrimination using ROC curve and compute the Area Under the Curve (AUC)
- Visualize the ROC curve to compare model performance against random chance

```
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  60934  5408
       Yes   547  1346

               Accuracy : 0.9127
                 95% CI : (0.9106, 0.9148)
    No Information Rate : 0.901
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2801

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.19929
            Specificity : 0.99110
         Pos Pred Value : 0.71104
         Neg Pred Value : 0.91848
             Prevalence : 0.09898
         Detection Rate : 0.01973
   Detection Prevalence : 0.02774
      Balanced Accuracy : 0.59520

       'Positive' Class : Yes
```

Let us now calculate the AUC and plot the ROC curve.

```
Setting levels:  control = No, case = Yes
Setting direction:  controls < cases

[1] "The AUC value is:  0.800496176845254"
```

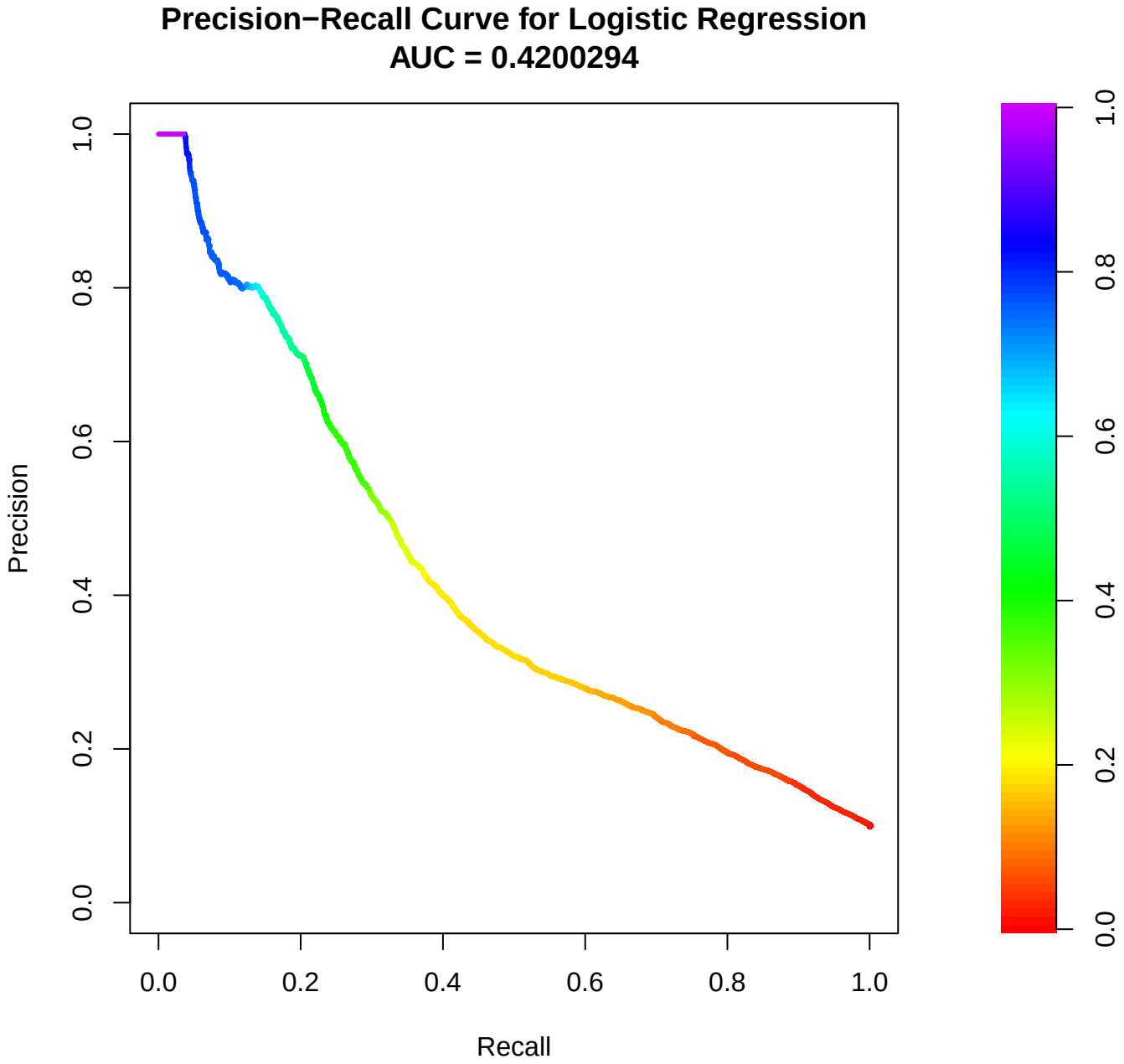


Key Takeaways:

- The logistic regression model for goal prediction achieved an overall accuracy of 91.3%, largely driven by its ability to **correctly identify** "No Goal" events (specificity = 99.1%).
- However, its sensitivity for detecting actual goals was **much lower** at 19.9%, indicating that the model often misses true scoring moments. When a goal is predicted, the **model is correct** about 71.1% of the time, while the balanced accuracy (which accounts for class imbalance) stands at 59.5%.
- The moderate Kappa value (0.28) suggests **fair agreement** beyond chance.
- Importantly, the ROC curve highlights the model's **strong discrimination power**, with an AUC of 0.80, meaning it can correctly rank a true goal over a non-goal **80%** of the time.

In summary, the model is a **reliable "defender"** in avoiding false alarms but struggles to identify goals consistently. This makes it highly **effective** for ruling out **non-scoring plays** but **less effective** for capturing **exciting goal events.**

Let us now look at the Precision-Recall curve.



The **Precision–Recall curve** provides a closer look at model performance in handling imbalanced data, where goals are much rarer than non-goals.

- **AUC (PR curve) = 0.42** → This is substantially lower than the ROC AUC (0.80), highlighting that while the model separates classes reasonably well overall, its ability to capture the **minority "goal" class is weaker**.
- **Precision trend:** Starts very high (close to 1.0) at low recall, meaning the few goals it predicts with high certainty are often correct. However, as recall increases, precision drops quickly, showing the trade-off between detecting more goals and maintaining accuracy.
- **Recall:** The model **struggles to cover** a large fraction of true goals without sacrificing precision.

While the ROC curve suggested strong discrimination, the PR curve reveals the real challenge with rare events like goals: the model is good at saying **"this is probably not a goal,"** but when it tries to catch more goals, it becomes much **less reliable.**

Decoding Football's Finishing Secrets:

Analyzing over **941,000 events** across **Europe's top five leagues** (2011–2017), this study uncovers what truly drives goal-scoring.

- Goals remain rare gems, with hot zones (penalty spot & six-yard box) boasting >70% conversion, while headers deliver the highest efficiency despite lower frequency.
- Set pieces emerge as gold mines, especially when finished with the right foot, and through balls outshine crosses or simple passes as the most effective assist method. Open play lags in efficiency, underlining the need for structured attacking creativity.
- Among players, Messi leads with unmatched consistency, excelling in open play and tight spaces, while Ronaldo and Ibrahimović dominate aerially, Lewandowski and Cavani thrive in set pieces, and Aubameyang and Griezmann excel with speed and through balls.
- At the team level, Barcelona, Real Madrid, PSG, and Bayern Munich convert chances with ruthless precision, leaving English giants comparatively wasteful.
- A logistic regression model achieved 91.3% accuracy, 99.1% specificity, 19.9% sensitivity, and an AUC of 0.80, proving excellent at ruling out non-goals but weaker in detecting actual scoring moments.
- The insights highlight that finishing is not just talent but a product of zones, supply chains, and tactical context, offering clubs a roadmap: exploit golden zones, design lethal set-piece routines, and match striker profiles with the right supply for maximum efficiency.

Future Statistical Work:

- **Expand Predictors Beyond On-Pitch Events:** Incorporate external factors such as team form, opponent strength, and league context into predictive models for richer probability estimates.
- **Causal Impact Evaluation:** Apply Difference-in-Differences or Synthetic Control to measure how tactical changes (e.g., formation shifts, key player signings, managerial changes) affect goal conversion rates.
- **Survival Analysis of Goal Opportunities:** Model the "lifespan" of an attacking move from build-up to shot, identifying factors that prolong or kill scoring chances.
- **Advanced Machine Learning Models:** Apply Random Forests, Gradient Boosting, or Neural Networks to capture non-linear interactions between shot location, assist type, and player skill for superior goal prediction.

Conclusion:

This project successfully dissects the **anatomy of a goal**, revealing how shot location, assist method, body part, situation, and player quality interact to determine scoring efficiency. By applying **logistic regression, ROC/AUC** analysis, and conversion metrics, it establishes a strong quantitative foundation for understanding finishing dynamics.

However, expanding the scope to include **tactical, psychological, and contextual variables**, along with more advanced modeling approaches, would unlock even deeper insights. These findings not only advance academic understanding but also provide practical pathways for clubs, coaches, and analysts to optimize attacking strategies, maximize conversion, and align player strengths with tactical supply chains.