Utilizing StatsBomb Football Performance Data to Optimize Training and Performance in Football

Analysis, Design, and Implementation Report

By

Victor Odoh
AI & Data Analytics Professional | MSc Applied Artificial Intelligence

Abstract

Football science faces a constant dual challenge: maximizing athletic performance and minimizing injury risk. This project tackles these issues head-on through a meticulous data-driven analysis of Atlético de Madrid's 2015/2016 La Liga season, utilizing StatsBomb's rich performance data. The foundation of this research lies in the rigorous collection and cleaning of the StatsBomb dataset. Focusing specifically on Atlético de Madrid's matches ensures data consistency and avoids potential biases that might arise from analyzing multiple teams with varying playing styles. This meticulous data preparation step is crucial, as it guarantees the accuracy and reliability of the subsequent analysis. Next, the project delves into the data using a potent blend of business intelligence and machine learning techniques. Business intelligence tools enable the researcher to explore trends, patterns, and relationships within the dataset. Machine learning algorithms, on the other hand, can uncover more nuanced insights and identify hidden patterns that might escape traditional statistical analysis. This combined approach empowers the research to extract the maximum value from the data.

Through this in-depth analysis, the project unveils critical performance indicators (KPIs). These KPIs are not just generic metrics; they are carefully chosen, data-driven insights that provide invaluable information for stakeholders like coaches, analysts, and team managers. By monitoring these KPIs, stakeholders can gain a deeper understanding of individual and team performance, allowing them to make informed decisions for improvement. For instance, the research might identify a previously overlooked correlation between a player's passing speed and their shooting accuracy within specific game situations. This newfound knowledge could then inform targeted training drills to improve a player's decision-making under pressure. Similarly, the analysis might reveal patterns in player movement during possession phases, leading to adjustments in tactical strategies to optimize ball retention.

Ultimately, the project champions data-driven decision-making as the key to unlocking peak performance for football teams, while also mitigating injury risks. The insights gleaned from this research empower stakeholders to refine training regimens, optimize player selection, and implement preventative measures to keep players healthy. This data-driven approach has the potential to propel teams and athletes towards achieving a significant competitive edge.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

1.2 StatsBomb Football Performance Data: A Valuable Resource

Among the various sources of football performance data, StatsBomb stands out as a comprehensive and reliable provider. StatsBomb utilizes advanced tracking technology and expert analysis to capture a vast array of in-game and training data points, covering various aspects of player and team performance. This data encompasses detailed information on player movements, passing locations, shot locations, tackling attempts, and other critical performance indicators. The high accuracy and granularity of StatsBomb data make it a valuable resource for researchers and practitioners in the sports science industry seeking to leverage data analytics for performance optimization. It is trusted by over 100 elite teams globally to be their data provider of choice.

1.3 The Challenge: Extracting Insights and Optimizing Performance

While the availability of data like StatsBomb offers immense potential for improving football performance, effectively utilizing this data presents significant challenges. The sheer volume of data can be overwhelming, and extracting meaningful insights requires robust data analysis techniques and expertise. Furthermore, translating these insights into actionable recommendations for training and performance improvement necessitates a deep understanding of sports science principles and the complexities of the game itself.

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

1.5 Project Aim and Objectives

By bridging the gap between player-centric metrics and tactical decision-making, this project aims to address the challenge of optimizing training and performance in the sports science industry by leveraging the insights derived from StatsBomb football performance data. Specifically, the project seeks to enhance decision-making processes related to athletic training, injury prevention, and overall performance improvement through the application of business intelligence and machine learning techniques.

1.5.1 Objectives

To achieve the stated aim, the project will pursue the following specific objectives:

- Collect and clean StatsBomb football performance data to ensure its accuracy and consistency for subsequent analysis.

- Apply appropriate business intelligence and machine learning techniques to analyze the data, identifying key trends, patterns, and relationships.
- Based on the data analysis, identify key performance indicators (KPIs) that are most effective in assessing and monitoring individual and team performance.
- Develop data visualizations and interactive dashboards using Power BI to effectively communicate the identified KPIs and trends to stakeholders within the sports science industry.
- Evaluate the effectiveness of the proposed business intelligence and machine learning methods in optimizing training and performance, demonstrating their potential value and impact on the sports science industry.

1.6 Significance of the Project

This project addresses a critical need within the sports science industry by exploring the potential of data-driven approaches for optimizing training and performance in football. The findings and recommendations generated through this project hold significant value for various stakeholders, including:

i. Sports scientists and performance analysts: The project offers insights into training methods and performance indicators, aiding in better training design, injury prevention, and athlete performance. It also showcases effective BI and ML techniques, advancing the sports science field.
ii. Athletes and coaches: Findings can inform personalized training plans, injury prevention, and tactical adjustments, helping to optimize training regimens, reduce injury risks, and enhance performance.
iii. Sports organizations: Demonstrating the value of data-driven optimization can encourage investment in analytics, providing a competitive edge through improved player performance insights, talent recruitment, and strategic decision-making.
iv. Broader sports community: The project's success in football analytics can inspire similar efforts in other sports, improving training methods, performance strategies, and overall athletic success across various disciplines.

Overall, this project holds significant promise for revolutionizing the way athletes, coaches, sports scientists, and organizations approach training and performance optimization in football and potentially across the broader sports landscape.

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Chapter Three: Methodology

The project utilized a combination of tools to achieve its objectives effectively. Microsoft Power BI served as the primary platform for data visualization, analysis, and dashboard creation. All preprocessing, advanced manipulation and machine learning tasks were facilitated using Python, M language, and DAX. JupyterLab was employed for dataset extraction from the online data source, ensuring data integrity and quality. The integration of these tools enabled comprehensive data analysis and visualization, that can empower stakeholders to make informed decisions on personalized training regimes for athletes as well as tactical decisions for performance optimization and injury prevention.

3.1 Data Source

StatsBomb's Github repository houses all its sports events data which are of the JSON type. Users can download or directly stream StatsBomb data into Python or R using log in credentials for their API access. This access is for paying customers. However, some data are available for free access in their open-source repository: https://github.com/statsbomb/open-data

3.1.1 Routine Task for Streaming Events Data from Source

Currently, the most efficient method for retrieving desired event data from the StatsBomb open-source repository involves utilizing their custom R or Python package (**statsbombpy)**. It took approximately 50 minutes to stream into Python, all events data pertaining to the team of interest.

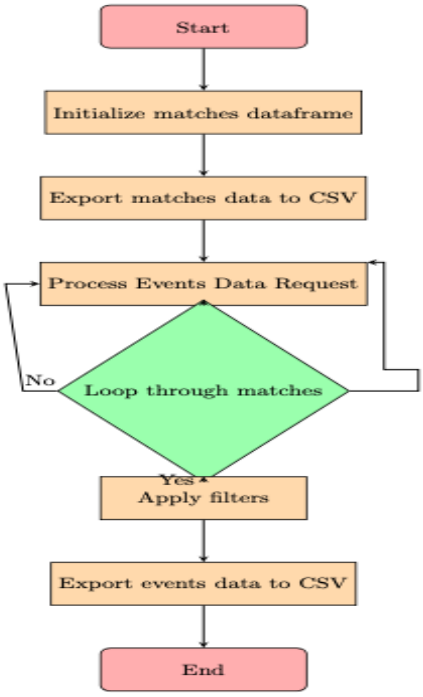


Figure 3.0 Routine to Fetch Dataset

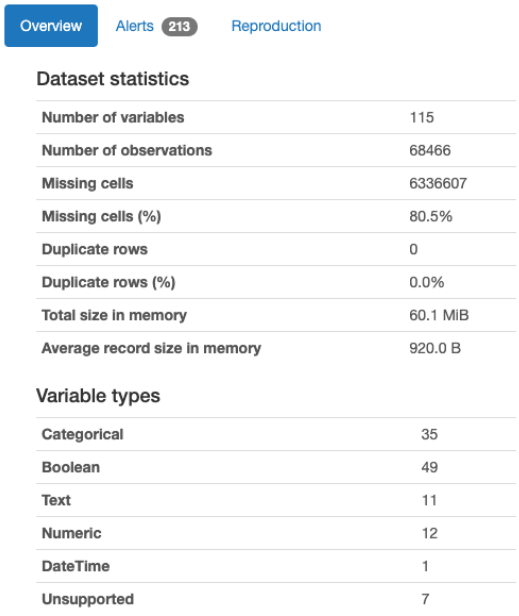
3.2 Dataset Description

Presently, among the extensive collection of football events data available in its GitHub repository, the most recent men's elite football clubs' competition data made available by StatsBomb for free public access dates back to the 2015/2016 football season. Specifically, this encompasses the La Liga 2015/2016 season. The dataset utilized for this project was derived from the events occurring within this League season. It comprises events associated with Atlético de Madrid, a notably competitive club that concluded the season in 3rd place with 88 points, trailing behind Real Madrid CF. The dataset includes two tables consisting of the following:

- Table 1: All events data across all matches played by Athletico Madrid during the 2015/2016 La Liga season.
- Table 2: Data on each match played in that season.

A detailed description of the dataset, which is obtained from the StatsBomb data documentation file, has been attached in the Appendix section of this report due to the volume of information contained therein. Table A3 contains information on the tactical positions.

3.2.1 Exploratory Analysis of the Dataset


Overview Alerts 213 Reproduction

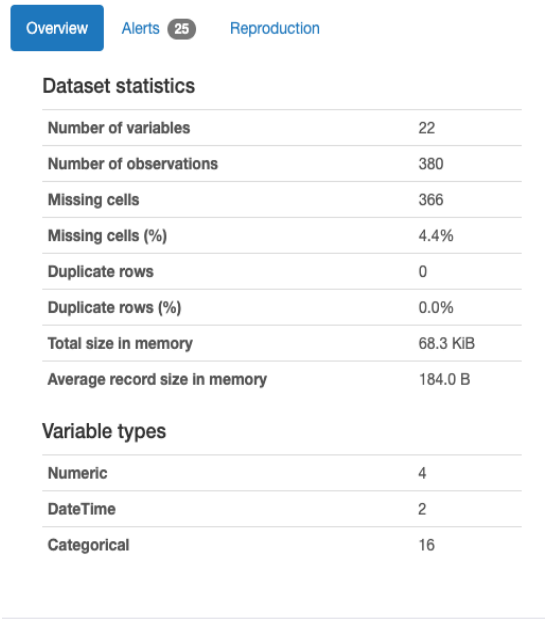
Dataset statistics

Number of variables	115
Number of observations	68466
Missing cells	6336607
Missing cells (%)	80.5%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	60.1 MiB
Average record size in memory	920.0 B

Variable types

Categorical	35
Boolean	49
Text	11
Numeric	12
DateTime	1
Unsupported	7

Figure 3.1a Events Data Overview


Overview Alerts 25 Reproduction

Dataset statistics

Number of variables	22
Number of observations	380
Missing cells	366
Missing cells (%)	4.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	68.3 KiB
Average record size in memory	184.0 B

Variable types

Numeric	4
DateTime	2
Categorical	16

Figure 3.1b Matches Data Overview

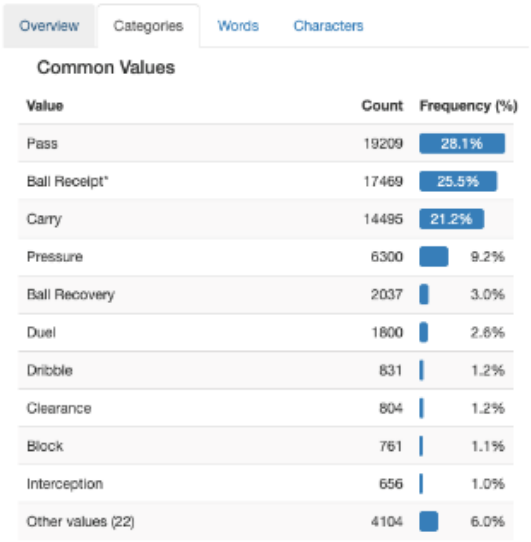


Figure 3.1c Event Type Category Distribution

Further checks were carried out to guide the preprocessing of the data.

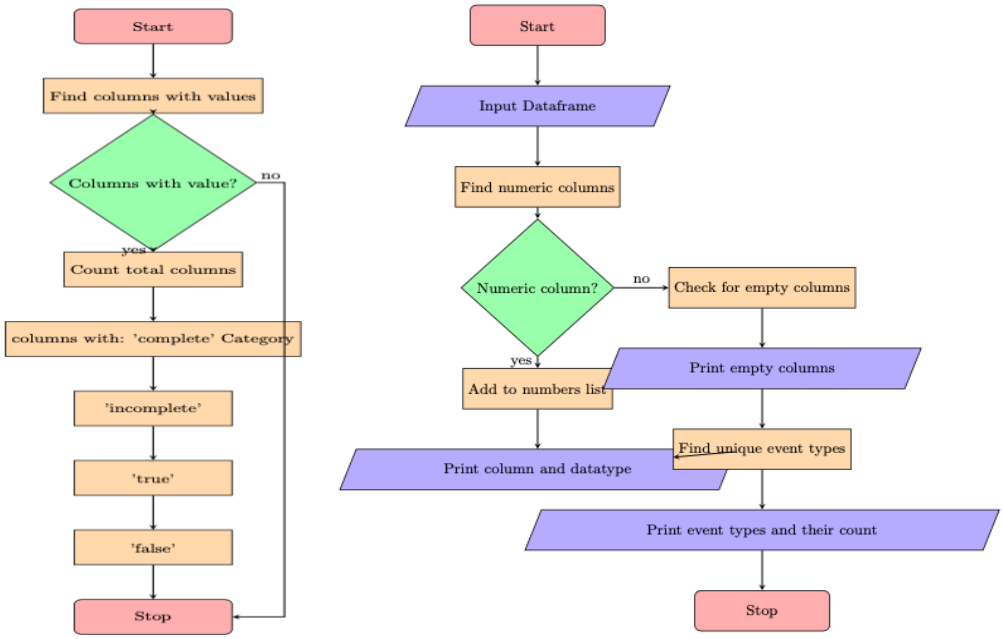


Figure 3.2 Pre-cleaning Checks

Key Observations:

- The Events Data table contains roughly 68,500 rows and 115 columns. While no errors were found upon loading, several data quality issues require attention.
- Missing Values: High occurrence (80.5%) due to the nature of categorical data ("type" with 32 event types). Nested event details further increase missing values.

- Inconsistent Labeling: Many variables (52 columns) lack complete labeling (e.g., missing "FALSE" for "TRUE"). Pass types and techniques lack crucial categories.
- Unnecessary Characters: Special characters like "*" appear in some category names.
- Sorting: Data needs sorting by Match Date, ID, Period, and timestamp.

These issues hinder analysis and require cleaning before drawing meaningful insights.

Matches Data:

- 22 columns and 380 rows
- 4 Numeric, 16 Categorical Variables, and 2 DateTime Variables
- No Duplicate rows
- No Blank Variables
- Only 4.4% of Data Missing
- No errors were detected upon loading into Power BI

3.3 Data Preprocessing

The data cleaning workflow involved removing irrelevant columns, handling missing values (imputed or replaced), converting timestamps, and sorting by match ID and time. Missing labels were addressed, and data types were standardized. Further cleaning occurred in Power Query and Power BI.

3.3.1 Variable Creation

3.3.3.1 Python Columns and Table

The following columns were created in the events data table using Python, most of which were involved in the creation of the data for the clustering algorithm:

- *timestamp_min*: *timestamp* column values (HH:MM:SS) represented in minutes to facilitate calculations involving playtime.

- *last_activity_mark*: Count to identify the instance where each player's last activity was recorded for each match. (Also used for appearance count).

- *playtime_p90*: playtime per match (min) for each player, recorded in the instance where the player's last match activity was recorded.

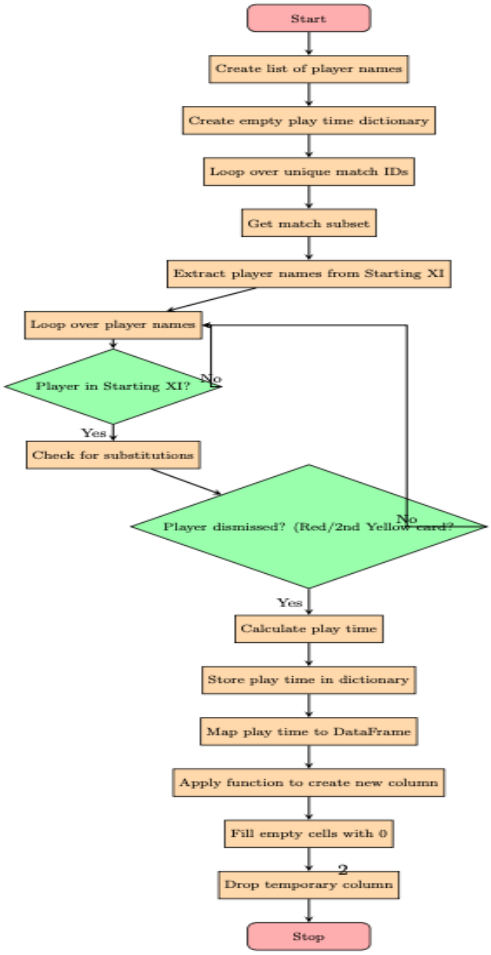


Figure 3.3 Flowchart for *playtime_p90*

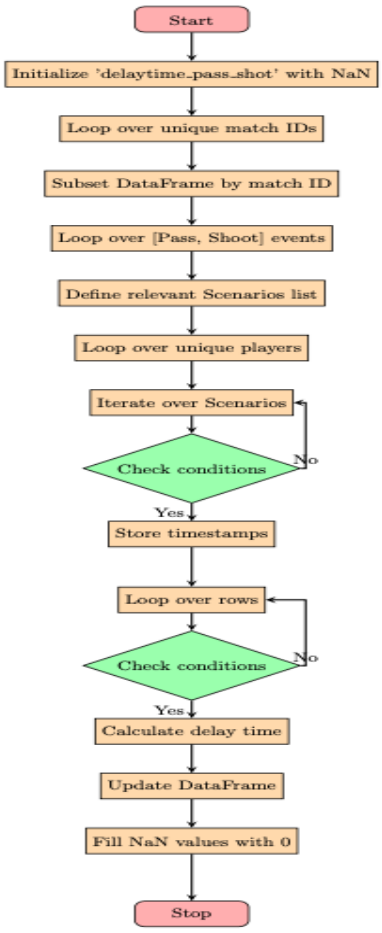


Figure 3.4 Flowchart for *delaytime_pass*_shot

- *delaytime_pass_shot*: time (sec) taken to make a pass or shot after possessing the ball (sec). It considers all scenarios or conditions by which the ball possession was achieved (ie. Either through a *50/50*, *Ball Receipt*, *Ball Recovery*, *Block*, *Duel*, or *Interception* event). For a *Pass* event type, the delay time is recorded in the instance where that pass event type is applicable, and likewise for a *Shot* event type.
- *pressure_influence_index_p90*: This metric quantifies each player's influence per 90 minutes (per match), in applying pressure during the game relative to the entire team. A higher value indicates a greater impact on pressurizing opponents.
- *ball_protection_index_p90:* A measure of how a player guards the ball against any potential tackle by an opponent, without attempting a dribble.
  carry_length: The distance covered (yd.) for each time a player moves the ball.
- *key_role*: For each player, the position played in most matches (Natural position), created to facilitate filtering actions based on this selection.
- *event_outcome*: Specifies if an event action (where applicable) led to a positive or negative outcome, created to facilitate calculations involving positive event outcomes. Eg. Efficiency.

Event Type	Categories (Positive Outcomes)
50_50	{'outcome': {'id': 4, 'name': 'Won'}}, {'outcome': {'id': 3, 'name': 'Success To Team'}}
ball_receipt_outcome	Complete
ball_recovery_recovery_failure	FALSE
type	Block, Carry, Clearance, Foul Won, Pressure, Shield
dribble_outcome	Complete
duel_outcome	Success, Success In Play, Success Out, Won
goalkeeper_outcome	Claim, Clear, Collected Twice, In Play, In Play Safe, Saved Twice, Success, Touched Out, Won, Success In Play, Punched out
interception_outcome	Success, Success In Play, Success Out, Won
pass_outcome	Complete, Injury Clearance
shot_outcome	Goal, Post, Saved, Saved to Post

Table 3.4 Event Types/Categories Assigned a Positive Outcome value

In Table 3.4, all specified categories (separated with commas) of each applicable event type were assigned 'Positive' outcome values based on careful consideration and interpretation of the dataset documentation. 10 event types (half start/end, referee actions, team changes, own goals, player substitutions) lacked positive/negative outcomes as they don't reflect player performance or direct match events. Excluding them with a 'N/A' value removes irrelevant data, focusing analysis on actions impacting performance and match outcomes.

The ***possession_management_data*** table was created for analysis on clusters based on some possession management related metrics generated. It consists of the following columns/metrics:

- *player*: Identifiers of individual players.
- *playtime*: Total amount of play time in minutes, of each player across all matches.
- *appearances:* Indicates the number of matches each player has participated.
- *ball_recovery_efficiency:* a measure of a player's effectiveness in recovering a loose ball.
- *duel_efficiency*: a player's effectiveness in winning duels or one-on-one battles for the ball.
- *interception_efficiency:* a measure a player's ability to intercept passes or gain possession through interceptions.
- *dribble_efficiency*: measure a player's effectiveness in dribbling past opponents.
- *pass_efficiency*: a measure that evaluates a player's accuracy and effectiveness in passing the ball to teammates.
- *pass_reception_efficiency*: a player's effectiveness in receiving a pass.

- *Ball_protection_index_p90*
- *pressure_influence_index_p90*: Average value of a player's Pressure Influence Index across all matches in which they participated.
- *cluster*: This column contains the cluster assignments generated by a machine learning model (Kmeans Model). These clusters identify groups of players with similar characteristics related to possession retention/recovery strengths, based on the provided metrics.

3.3.3.2 DAX Columns and Table

The following columns were created using DAX:

- *general_role*: created to facilitate filtering actions based on general playing roles (Goalkeeper, Defender, Midfielder, or Forward).
- *location:* Location of match (Home/Away)
- *opponent:* (All Opponents played during the selected time frame)
- *match_outcome:* Either a |Win, Loss, or Draw

The *position_mapping* table was created to facilitate the creation of the *general_role* column for filtering actions.

3.4 Data Modelling: Star Schema

A Star Schema model was chosen for data visualization. The central "events_data" table connects to three dimensions: "matches_data" (linked by match ID), "possession_management_data" (linked by player ID), and "PositionMapping" (linked by player position). This structure allows for easy analysis of events within matches, for specific players, and by position. All relationships are many-to-one, enabling filters in dimension tables to impact the central fact table.

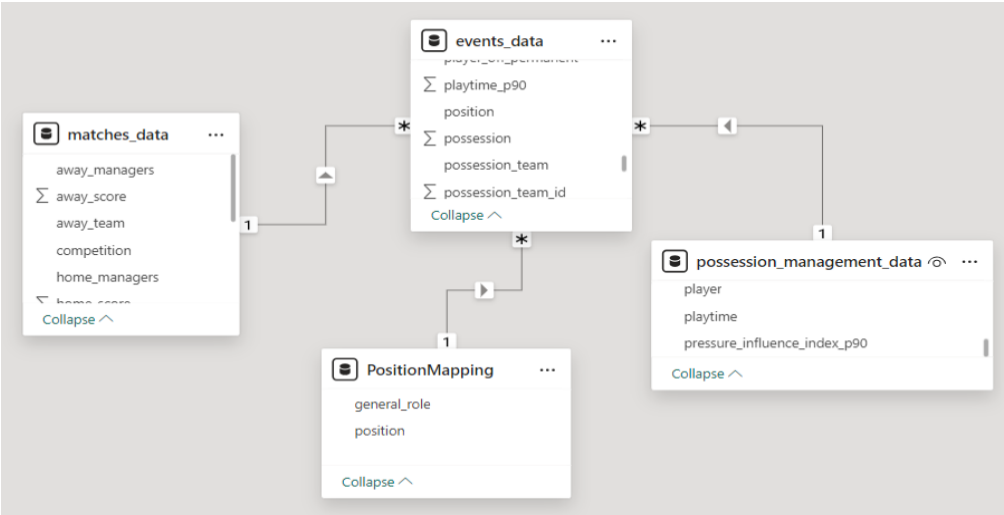


Figure 3.5 Star Schema Model

3.5 Dashboard Design

The dashboard consists of the following pages:

- Home Page
- Team Performance Overview
- Decision Making Analysis
- Possession Management Analysis
- Injury Risk Assessment
- Player Management

3.5.1 Team Performance Overview:

This was designed to give insights into the overall performance of the team. The following metrics were formulated:

- Performance Efficiency P90 (Team/Player): A measure of efficiency per 90 Minutes (Per Match) in terms of the number of positive outcomes for all event types. It is the Sum of a player's positive event outcomes divided by the sum of all instances of the player. For a team, it represents the average quality of the Team players per match.
- Contribution Index P90 (Team/Player): A measure of the extent to which a player contributes positively to the team relative to the team's total positive event outcomes. For a team, it represents the average the average Contribution Index of a typical player of the team value of the team players.
- Play time: Sum of in-play time in minutes for each player across all matches.

Key Visuals in this layout:

- Scatter Plot for Performance P90, Contribution Index P90 and Play Time (Min) by player.
- Clustered Bars for Inefficiency Chart for all event types
- Line Charts for insights on the Defence, Midfield, and Attack positions of the team.
- Multi Card KPIs visual to display key metrics and KPIs.
- Slicers: To filter across General Roles, Location, Opponents, and Match Outcome.

3.5.2 Decision-Making Analysis

This page was designed to provide insights into players' decision-making strengths relative to how fast and precise the decision was taken to make a pass or take a shot during game time. It considers all successful passes and shots on target for both under-pressure and regular play (all) conditions. The following metrics were formulated:

- Passing Speed (yd./s): The Speed of a pass in yards/sec. A higher passing speed indicates quicker execution of passes.

- Shooting Speed (yd./s): Speed of a shot taken by a player. A lower shooting delay implies faster decision-making and execution in shooting scenarios.
- Passing Delay: Time in seconds, taken to execute a pass to a teammate, calculated for both successful, and all passes. A lower passing delay indicates a quicker execution in passing situations.
- Shooting Delay: Time (sec) taken to execute, calculated for both successful (shots on target), and all shooting events. A lower shooting delay implies faster execution in shooting scenarios.
- Passing Decision Index (PDI): Passing Delay (all passes), divided by Passing Delay for successful passes. Passing Decision Index above 1 indicates that the player makes successful passes more quickly than unsuccessful passes.
- Shooting Decision Index (SDI): Shooting Delay (all shots), divided by Shooting Delay for shots on target. A Shooting Decision Index above 1 indicates that the player makes shots on target more quickly than they attempt off-target shots.

The **Decision Efficiency Index (DEI)** KPI, was created to provide an overall measure of a player's decision-making efficiency in passing and shooting situations, considering the PDI, and SDI. (Weighted Equally)

$$DEI = \frac{Passing\ Decision\ Index + Shooting\ Decision\ Index}{2}$$

Therefore, the DEI can be interpreted as follows:

- Higher values of DEI indicate better decision-making efficiency, in both passing and shooting situations.
- Lower values of DEI suggest that the player takes longer to make decisions or execute successful passes and shots, which may indicate less efficient decision-making in passing and shooting situations.

Key Visuals in this layout:

- Scatter Plots for:
  - Delay Per Successful Pass & PDI
  - Delay Per Shots on Target & SDI
  - Decision Making Score & PDI by SDI
- Slicer to filter based on play conditions (Under Pressure/ Normal Conditions)

3.5.3 Possession Management Optimization: Machine Learning Application

Analyzing possession goes beyond just holding the ball. It includes how effectively players win it back (recovery) and keep it (retention). By using machine learning, players are categorized based on these skills, revealing team strengths and weaknesses. This allows for targeted training to address skill gaps and optimize overall possession management, leading to better on-field performance.

3.5.3.1 Possession Management Factors: Metrics and KPIs

Two (2) KPIs were created to provide insights into the possession management strengths of a player/team.

The Possession Recovery Index (PRC) evaluates a player's ability to win back the ball. It combines weighted metrics:

- Ball Recovery Efficiency (weight: 3) - how often they reclaim loose balls.
- Duel Efficiency (weight: 2.6) - their success in head-to-head battles.
- Interception Efficiency (weight: 1) - their skill at reading and stopping passes.
- Pressure Influence Index (weight: 1) - their impact on disrupting opponent possession.
- Higher weights reflect greater importance for regaining control.

The weight assignment is based on the distribution of each type of activity during gameplay (Figure 3.1c) to reflect the relative importance of each metric in contributing to possession recovery. The Weights are multiplied by the corresponding efficiencies, and the sum is divided by the total weight (15.8) to normalize the index to a scale of 0 to 1.

$$PRC = \frac{3(Ball\ Recovery\ \%) + 2.6(Duel\ \%) + (Interception\ \%) + (PII)}{7.6}$$

The weighted average of these metrics ensures that each aspect of possession recovery is appropriately represented in the final index, enabling coaches and analysts to evaluate player performance and strategize accordingly.

A higher PRC indicates that the player is highly efficient in winning back possession for their team. This could imply that the player is proactive in disrupting opponent play and initiating counter-attacks. Conversely, a lower PRI suggests that the player may struggle in regaining possession or exerting pressure on opponents effectively.

The **Possession Retention Index (PRT)** goes beyond passing success, evaluating a player's overall ball control. It combines weighted metrics like passing (28.1%), receiving (25.5%), dribbling (1.2%), and inverted ball protection index- BPI (1.0) to create a single, comprehensive score. This reflects a player's effectiveness in maintaining possession during a match.

$$PRT = \frac{28.1(Passing\ \%) + 25.5(Pass\ Reception\ \%) + 1.2(Dribble\ \%) - (1 - BPI)}{55.8}$$

A higher PRT indicates that the player excels in retaining possession by executing accurate passes, receiving the ball effectively, and protecting it from opponents' challenges.

On the other hand, a lower PRT suggests that the player may face challenges in maintaining possession, either due to inaccurate passing, poor ball control, or vulnerability to opponent pressure.

Interpreting these indices (PRC, PRT) allows coaches, analysts, and team managers to assess player performance in crucial aspects of possession play. Players with high PRC and PRT values are likely to contribute significantly to their team's overall possession game and may be considered key assets in maintaining possession and regaining possession when lost. Also, by leveraging these indices, coaches can develop personalized training plans that address specific areas of improvement for each player, ultimately enhancing overall team performance on the field.

3.5.3.2 The K-Means Clustering Model for Possession Management Analysis

Below is the workflow:

- Feature Engineering: The *possession_management_data* table (Section 3.3.3.1) was specifically created for this task. It was created from the cleaned events data and all the required possession retention and possession recovery metrics for the model are represented as columns of this table: *ball_recovery_efficiency*, *duel_efficiency*, *interception_efficiency*, *dribble_efficiency*, *pass_efficiency*, *pass_reception_efficiency*, *Ball_protection_index_p90*, and *pressure_influence_index_p90*.
- Scaling the Data: Applied standardization to the selected columns to ensure uniform scaling. Utilized the **StandardScaler** to scale the data, maintaining consistency across variables.
- Exploring Optimal Cluster Number: Utilized the **Elbow Method** for Cluster Selection. The curve was analyzed to identify the point where the rate of decrease in the within-cluster sum of squares (WCSS) for each clustering solution starts to slow down significantly.

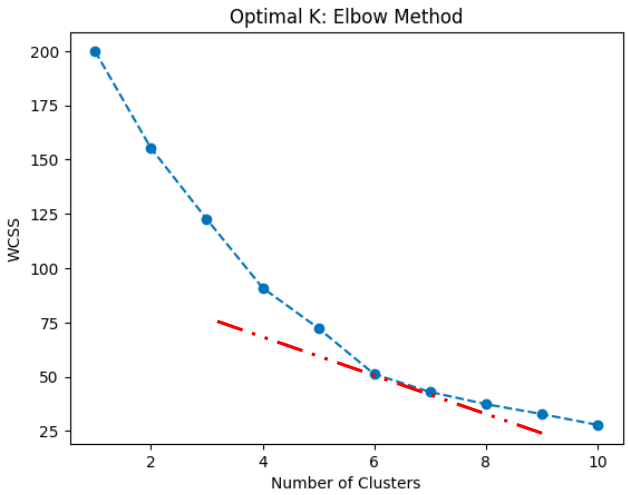


Figure 3.6 Elbow Point for Optimal K

- K-means Clustering: Executed K-means clustering on the scaled data, utilizing the optimal cluster number determined earlier from the Elbow Method. Each player was assigned to a cluster based on their efficiency metrics.
- Further evaluation was done to analyze the quality of clusters.
- The clusters were visualized to gain insights into player performance.

Key Visuals in this layout:

- Radar Chart for PRC and PRT by Player
- Bubble Chart by Akvelon for Cluster Distribution
- Matrix table: PMI, PRC, PRT, Total Playtime, Key Role

3.5.4 Injury Risk Assessment

The following Injury Risk factors (calculated metrics) were considered for this assessment:

- Total Duels: A duel is a 50-50 contest between two players of opposing sides in the match.
- Total Fouls: considers both total fouls won and fouls committed.
- Ball Time (min): Total time spent while active with the ball.
- Carry Distance (In yd.): Total distance covered while controlling the ball at the player's feet or standing.
- Average Carry Speed (yd./sec): Average Speed during Carry events of the player.
- Pressure Influence Index: A measure of the contribution of a player in pressuring the opposing team relative to the player's team (Total pressure instances of player Divided by Total pressure instances of the team).
- Injury Resilience Index: This measures how resilient or resistant a player is to injuries based on their performance or appearance frequence.

A KPI for Injury Risk, **Player Vulnerability Score** (PVS), was created as a weighted composite score that combines the above metrics to provide a comprehensive assessment of a player's injury susceptibility. The rationale behind the weight distribution is as follows:

- Total Duels: Duels are physical contests that can lead to injuries if not managed properly. A significant weight of **20%** was assigned to this metric as it reflects the player's involvement in physical battles.
- Total Fouls: Fouls, both won and committed, indicate the player's aggression and involvement in contentious situations. A moderate weight of **15%** was also assigned to this metric.
- Ball Time: Players who spend more time on the ball are more likely to be involved in gameplay, increasing their exposure to potential injury situations. Hence, a more significant weight of **20%** was assigned.
- Carry Distance: Players covering longer distances with the ball may experience fatigue, leading to potential injury risks. Hence, **15%** was assigned.

- Average Carry Speed: Higher average carry speed may increase the risk of injury due to the intensity of movements. A moderate weight of **10%** was assigned to this metric.
- Pressure Influence Index (PII): Players contributing more to pressuring the opposing team may engage in more physical play, increasing injury risk. **Weight: 15%**
- Injury Resilience Index (IRI): Players with a lower injury rate per match demonstrate better injury management and resilience. This metric was inverted and assigned a lower weight of **5%** since it represents resilience rather than susceptibility.

The scale of the final value of the PVS is adjusted based on the player's match experience to reflect the player's injury risk relative to their exposure to matches. It is divided by a scaling factor of 10 for easier interpretation. A higher PVS score implies a higher risk of injury for the player, and vice versa.

$$PVS = \frac{0.2(Total\ Duels + Ball\ Time) + 0.15(Total\ Fouls + Carry\ Distance) + 0.1(Avg.Carry\ Speed) + 0.15(PII) + 0.05(1 - IRR)}{10(Appearances)}$$

Key visuals in this layout:

- Funnel Chart for Player Vulnerability Score
- Pulse Chart for Weekly Player Vulnerability Score
- Multi Card KPIs to display key metrics.

3.5.5 Player Management

This page provides a summary on individual player performance metrics.

Key Visuals in this layout:

- Scatter Plot to give insights into Player Decision-making and Possession Management.
- Radar Chart for Performance by Key Events
- LineDot Chart for Weekly Performance P90.
- Line Chart for Player Performance Forecast.

Chapter Four: Findings and Evaluation

4.1 Team Performance Overview Page

The scatter plot depicting Performance Efficiency P90, Contribution Index P90, and Play Time (Min) by player provides a comprehensive overview of individual player performance metrics, identifying performing and poorly performing players, and allowing stakeholders to gain insights into various levels at which both performing and underperforming players contribute to the teams overall performance.

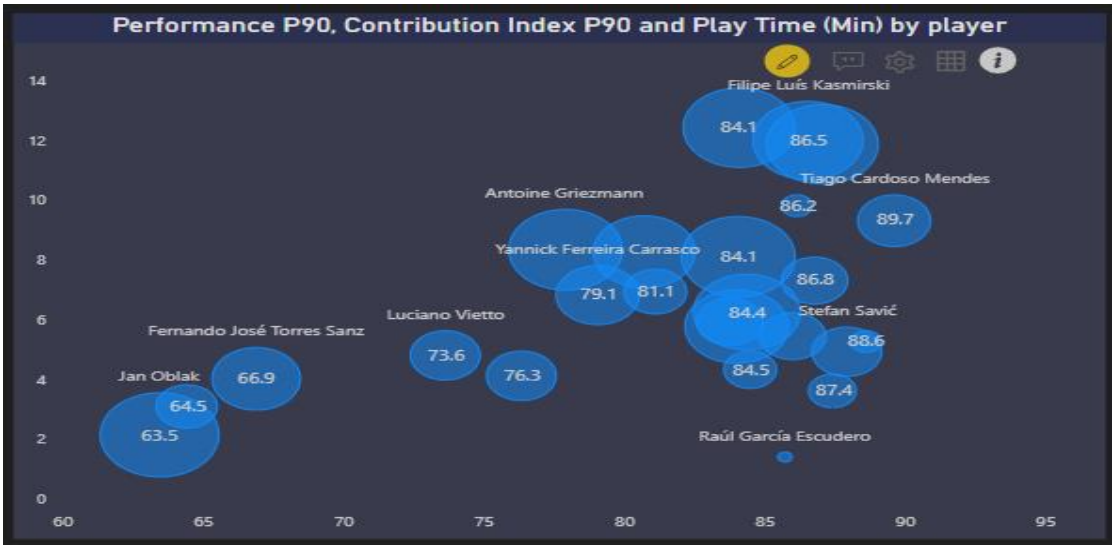


Figure 4.1 Performance P90 and Contribution Index P90 by Playtime

The chart reveals several insights:

- Many team members exhibit high performance efficiency, surpassing 80%.
- Despite Jan Oblak's significant playing time as the goalkeeper, his contribution index of 2.3 places him as the second lowest contributor to the team's overall performance, just after Raul Garcia Escudero. This could be because Oblak spends most of his time in the goalpost, facing fewer shot attempts from opponents. However, Oblak also has the lowest individual performance in the team, with a rating of 63.5%.
- Conversely, Raul Garcia Escudero boasts a commendable performance efficiency of 85.7% but ranks as the least contributing player per 90 minutes of play. This discrepancy may stem from his limited playing time compared to other team members.

Stakeholders at Athletico De Madrid may want to consider providing more playing opportunities to players with high performance efficiency, such as Raul Garcia Escudero, as they have the potential to significantly enhance team performance.

The clustered bar charts depicting inefficiencies across various event types provide valuable insights into areas where performance is lacking, facilitating targeted interventions for improvement. Upon analysis, it was noted that the team exhibits subpar quality in certain event types, particularly in 50/50 contests, duels, goalkeeping, and shooting.

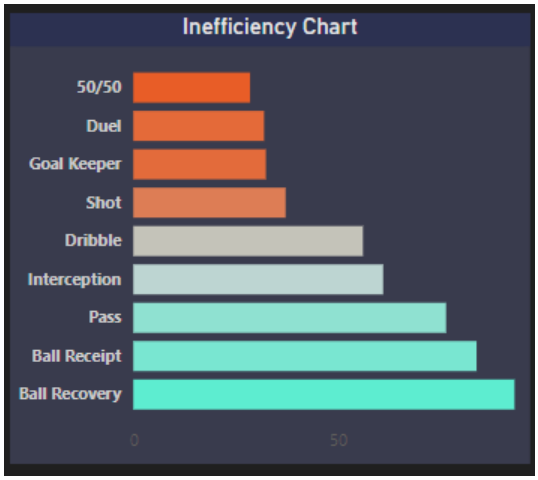


Figure 4.2 Inefficiency Chart

The notably poor performance in shooting events indicates that the team registers a higher number of shots off target compared to those on target. This observation serves as an indication to the manager of Athletico De Madrid that the team may be facing a skill deficiency in these specific areas.

The line charts offer insights into performance trends across Defense, Midfield, and Attack positions. A quick glance reveals that the collective performance of the team's defense and midfield has shown a gradual decline over time, whereas the team's attacking performance has remained consistent. This trend suggests that there may be a need for additional training programs focused on improving the performance of the defense and midfield. Stakeholders at Athletico De Madrid may consider allocating more resources to address this decline and enhance overall team performance.

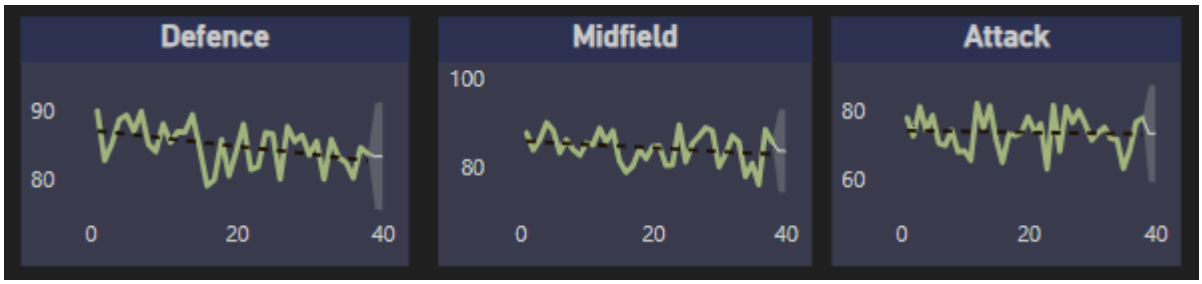


Figure 4. 3 Line Charts on Defence, Midfield, Attack

The multi-card KPIs visualization presents key metrics and KPIs in a concise format, facilitating quick assessment of team performance indicators

Figure 4.4 Team Performance Page

4.2 Decision Making Analysis Page

The visuals on this page provide crucial insights into players' decision-making strengths during gameplay, particularly in passing and shooting scenarios. By analyzing metrics such as Passing Speed, Shooting Speed, Passing Delay, and Shooting Delay, alongside derived indices like the Passing Decision Index (PDI) and Shooting Decision Index (SDI), the Decision-Making Analysis page offers a comprehensive view of players' decision-making efficiency. The Decision Efficiency Index (DEI) serves as a key performance indicator, representing an overall measure of a player's decision-making efficiency.

The scatter plots on this page provide a visual representation of the relationships between various decision-making metrics, facilitating the identification of patterns and areas for improvement. Furthermore, the slicer feature enables stakeholders to filter data based on play conditions, allowing for more nuanced analysis and actionable insights.

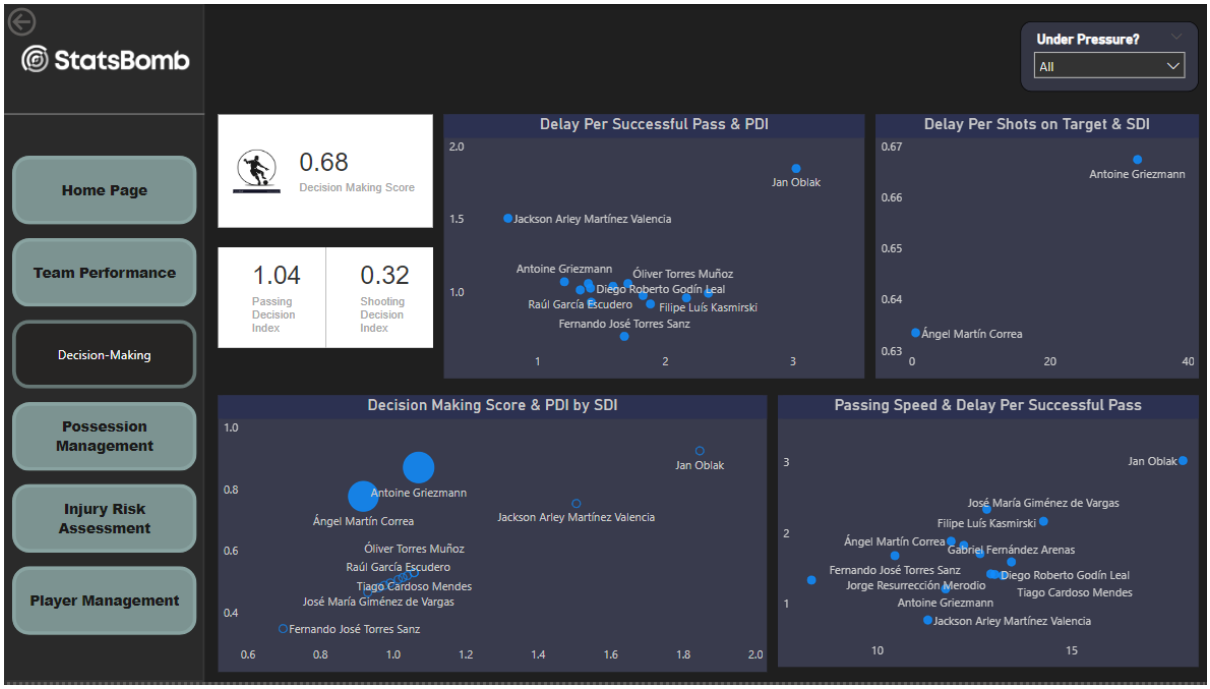


Figure 4.5 Decision Making Analysis Page

Observations from the data analysis reveal several key insights:
For normal playing conditions:

- The average Decision-making Index of a player in the team is 0.68, indicating a relatively balanced performance in decision-making. However, the average shooting Decision Index is notably low at 0.32, suggesting suboptimal decision-making in shooting scenarios. Conversely, the average Passing Decision Index is relatively high at 1.04, indicating quicker and more successful decision-making in passing situations. This discrepancy sheds light on the team's poor shooting efficiency observed in the Team Performance Page of the Dashboard.
- Antoine Griezeman emerges as a standout performer with the highest Passing and Shooting Decision Indices (1.07 and 0.67, respectively), along with the highest overall Decision-making score of 0.87. This highlights Griezeman's exceptional ability to execute successful passes and shots on target swiftly, making him a crucial asset to the team.
- In contrast, Fernando Jose Torres Sanz exhibits slower decision-making tendencies, with a Passing Decision Index of 0.7 and a Decision-making score of 0.35, indicating potential areas for improvement in his performance.

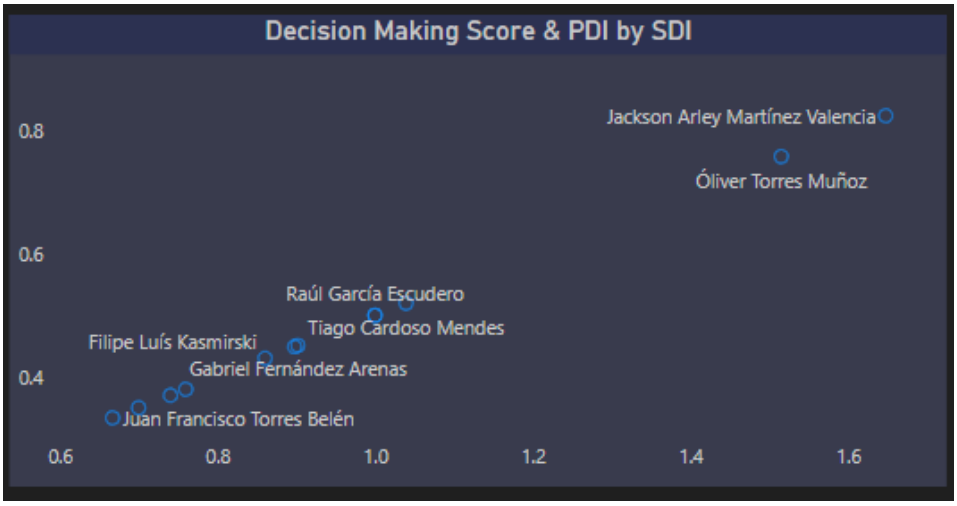


Figure 4.6 Decision making Score & PDI by SDI Filtered for ‘under pressure’ conditions.

For conditions in which passes and shots were executed under pressure:

- The team's average Decision-making score and Passing Decision Index decrease to 0.46 and 0.92, respectively. Additionally, it was observed that the average team player fails to hit the target when taking shots under pressure.
- Jackson Arley Martinez Valencia emerges as the standout decision-maker under pressure. With a Decision-Making Score of 0.82 and a Passing Decision Index of 1.65, the Left Centre Forward proves to be instrumental in the team's attacking prowess, showcasing exceptional skills in making quick decisions during the team's offensive buildup, potentially leading to goals or scoring opportunities.
- Conversely, Juan Francisco Torres Belen appears to be the slowest decision-maker when executing successful passes under pressure, with a Passing Decision Index of 0.67 and a subpar Decision-Making Score of 0.33.

These insights can inform strategic decisions aimed at optimizing player performance and enhancing overall team effectiveness. The insights gleaned from this dashboard offer valuable guidance for coaches, managers, or training teams at Athletico De Madri, suggesting tailored training programs aimed at bolstering athletes' cognitive abilities. These programs can target key areas such as situational awareness, anticipation, reaction time, and decision-making, particularly in shooting, and high-pressure scenarios.

4.3 Possession Management Analysis Page

These visuals play a pivotal role in bridging the gap between player-centric metrics and tactical decision-making, providing actionable insights into various aspects of possession management, including possession recovery and possession retention. The Possession Recovery Index (PRC) and Possession Retention Index (PRT) serve as key performance indicators (KPIs) that enable

coaches, analysts, and team managers to evaluate comprehensively, a player ability to regain, and maintain possession during gametime. These indices allow for a nuanced understanding of the players' ability to maintain, and recover possession during gametime, identifying skill shortages and areas for improvement.

Additionally, the analysis of player clusters aids in informing tactical decisions related to player selection and formation strategies, with a focus on optimizing team possession during matches. Coaches can identify players who either bolster or weaken team performance in terms of possession control, leveraging this information to tailor tactical setups and formations accordingly. For example, the clustering process, which generated six distinct clusters based on key possession management metrics, revealed that Cluster 4 attained the highest score in the Possession Management Index, followed by Cluster 3. This analysis offers valuable guidance to coaches and managers in making informed decisions regarding player selection and tactical formations geared towards enhancing possession gameplay.

4.3.1 Cluster Model Evaluation

The optimal number of clusters derived from the elbow point plot was 6 (Figure 3.6). Below is the cluster distribution chart resulting from the clustering Model.

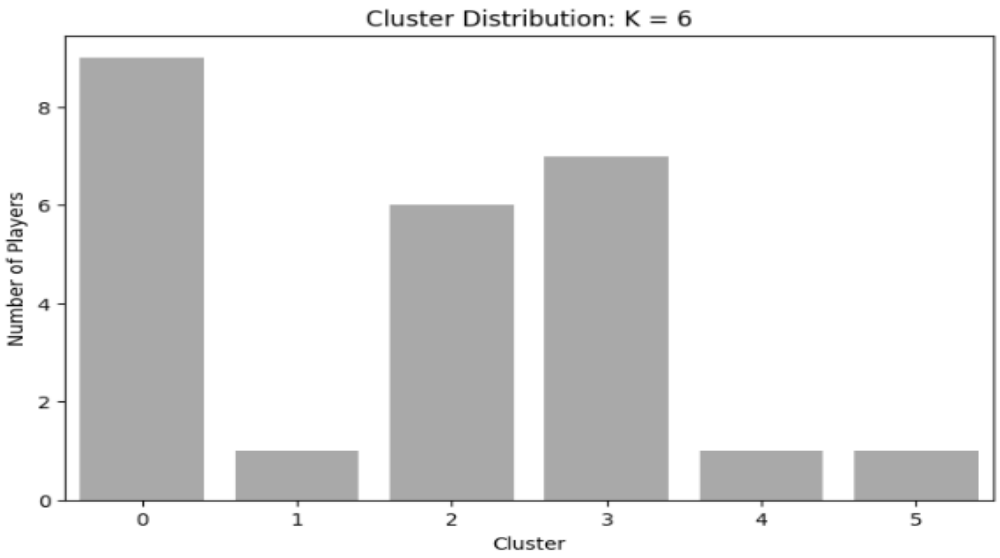


Figure 4.7 Cluster Distribution

The quality of this cluster distribution was further investigated using the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. These evaluation metrics provided insights into the overall quality of the clusters generated by the K-means algorithm:

- Silhouette Score: This assesses how well players are assigned to their clusters. Scores closer to 1 indicate a good fit (players similar within their cluster and dissimilar to others). Scores near 0 suggest overlapping clusters, where players might be better

suited elsewhere.. In this case, the score of 0.24 (Figure 4.7) indicates a fair separation between clusters.

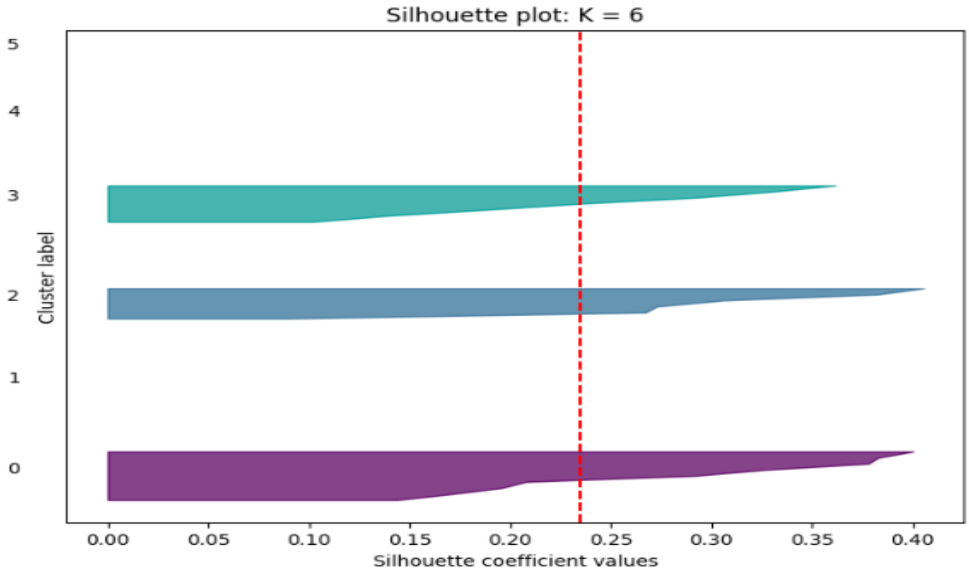


Figure 4.8 Silhouette Plot for Cluster Evaluation

- cc

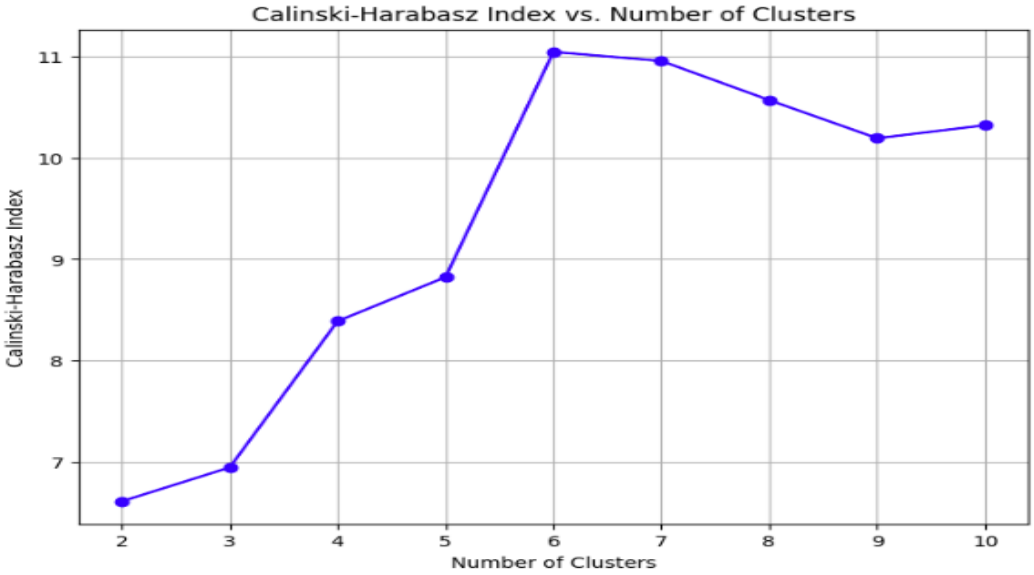


Figure 4.9 Calinski-Harabasz Index for Cluster Evaluation

- Davies-Bouldin Index (0.74): Lower values indicate better separation between player clusters. This score suggests a relatively good separation.

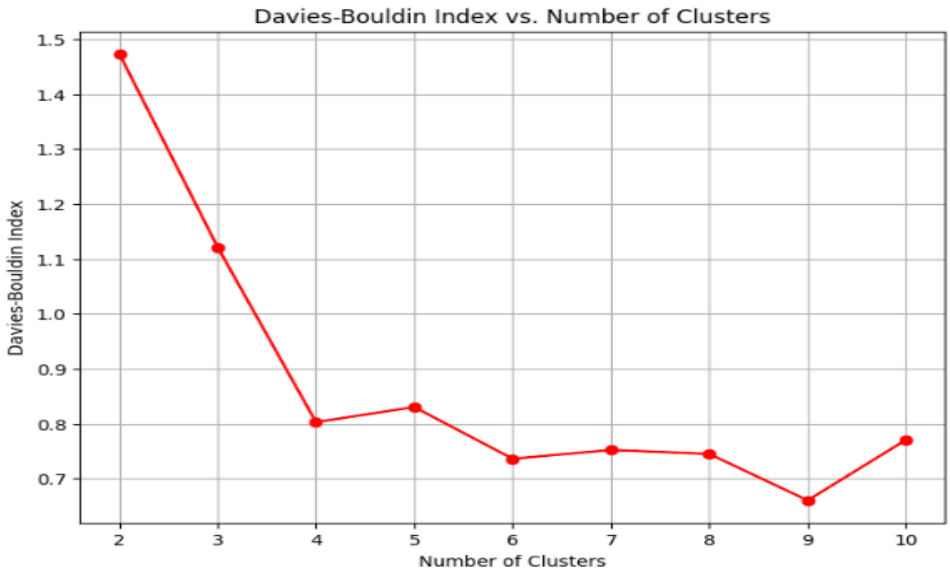


Figure 4.10 Davies-Bouldin Index for Cluster Evaluation

Overall, these metrics suggest that the 6-cluster solution derived from the Elbow method provides reasonable cluster separation and cohesion.

Figure 4.11 Possession Management Analysis Page

4.4 Injury Risk Assessment Page

The funnel chart for the Player Vulnerability Score (PVS) provides a concise summary of each player's injury susceptibility, allowing stakeholders to quickly identify high-risk individuals

who may require special attention or tailored training programs. This visualization serves as a valuable decision-making tool for coaches, managers, and medical staff to prioritize injury prevention strategies and allocate resources effectively. Additionally, the pulse chart for the Weekly Player Vulnerability Score offers a dynamic view of how players' injury risks fluctuate over time, enabling proactive monitoring and intervention to mitigate potential injury threats. The multi-card KPIs provide detailed insights into the contributing factors of the PVS, helping stakeholders understand the underlying metrics influencing players' injury susceptibility and guiding targeted interventions to address specific risk areas. Overall, these visuals empower stakeholders with actionable information to reduce injury incidence in professional football.

Figure 4.12 Injury Risk Assessment Page

4.5 Player Management Page

The Scatter Plot provides insights into player decision-making and possession management, offering a visual representation of key metrics such as passing accuracy, dribble efficiency, and defensive actions. This allows the management team to identify players who excel in specific areas and those who may require additional training or support. The Radar Chart provides a comprehensive overview of player performance across key events, enabling stakeholders to quickly assess strengths and weaknesses in various aspects of play. The LineDot Chart depicting weekly performance P90 allows for tracking player progress over time, highlighting trends and identifying areas for improvement or further investigation.

Figure 4.13 Player Management Overview Page

Collectively, these visuals play a crucial role in enhancing decision-making processes related to athletic training, injury prevention, and overall performance improvement of players/team. Insights generated from this dashboard can inform coaches and training teams to subject athletes to undergo training programs that focus on enhancing cognitive skills such as situational awareness, anticipation, reaction time, and decision-making under pressure.

Chapter Five: Discussion and Conclusion

5.1 Challenges and Limitations:

While designing the visualizations and metrics for the dashboard, it became evident that certain requirements, particularly those involving complex calculations and preprocessing tasks, would be computationally demanding if implemented solely using DAX within Power BI. To address this challenge and ensure efficient processing, Python was utilized alongside its extensive library collections, offering greater flexibility.

The calculation of metrics such as the Injury Frequency Ratio and Injury Resilience Index posed limitations due to unavailable data on minor injuries sustained during gameplay that did not result in player substitution. Consequently, the metrics consider only injuries leading to player substitution, potentially underrepresenting the true frequency of injuries.

Furthermore, the dashboard lacks a dedicated page exploring the assessment of players' technical abilities on a body part level. Insights into how players utilize different body parts during various game actions could provide valuable insights into performance optimization. For example, analyzing outcomes based on the best/worst foot for passing and shooting or the preferred hand for goalkeepers could enhance training program customization and strategic decision-making. Exploring these uncharted territories has the potential to uncover valuable insights to further refine athlete training programs.

5.2 Ethical Considerations

Ethical Compliance and Data Sourcing:
The research project diligently adhered to StatsBomb's terms and conditions governing the use of their football performance data. This included proper attribution of StatsBomb as the data source and incorporation of their logo in the project, ensuring compliance and transparency in data usage and dissemination.

Public Accessibility of Data:
The football performance data sourced from StatsBomb originates from publicly available channels, including broadcast feeds and direct inputs from professional football teams. Given its public nature and association with athletes operating within the public domain, the data lacks the personal expectation of privacy typically afforded to private individuals. Consequently, data privacy settings were appropriately configured to "Public," facilitating seamless data manipulation and preprocessing using Python scripts within Power BI and ensuring visibility for all users on the Power BI Service.

Alignment with Legitimate Interests:
Processing of this publicly available data aligns with the legitimate interests pursued by the research project, particularly in advancing sports science, performance optimization, and

evidence-based decision-making. This alignment with legitimate interests is recognized by GDPR as a lawful basis for processing personal data. The project strictly utilized essential athlete-identifiable information necessary for achieving its defined objectives, excluding unnecessary details to minimize the scope of personal data processing.

Purpose-Limited Usage of Data:
The research project commits to strict purpose limitation, ensuring that person-identifiable data is utilized solely for the specific purposes outlined in the project. This measure safeguards against any unauthorized or incompatible usage of the data, preserving the integrity and ethical conduct of the research endeavor.

5.3 Conclusion and Future Works

This project demonstrably addresses the challenge of optimizing training and performance in football by leveraging the power of StatsBomb data. By applying a blend of business intelligence and machine learning techniques, the research has yielded valuable insights to enhance decision-making in athletic training, injury prevention, and overall performance improvement. The analysis of player-centric metrics and tactical choices has unveiled crucial information about individual and team performance, empowering stakeholders to make data-driven decisions for success.

This research opens doors for exciting advancements. The integration of real-time data streams could provide even more immediate and accurate performance analysis. Additionally, incorporating more granular data on player technical abilities and injury risk factors could offer deeper insights into skill development and preventative strategies. Furthermore, exploring advanced machine learning algorithms holds promise for predictive analytics, allowing teams to anticipate performance trends and injury risks. By continuously refining data analysis techniques, we can ensure sports science practices remain at the forefront, contributing to the ever-evolving world of performance optimization in professional football.

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX