**Evaluating the LaLiga (2014–15 to 2018-19) dataset**

**Table Of Contents:**

1. Abstract
2. Introduction
3. Objective
4. Methodology

4.1 Statistical Modeling

4.2 Data Exploration

1. Data Sources
2. Result and Analysis
3. Conclusion
4. Future Scope
5. References

**1. Abstract:**

The La Liga, renowned for its intense competition and star-studded teams, provides a captivating backdrop for delving into the intricacies of football dynamics. This research project, structured in two main components, seeks to unravel insights from the La Liga (Spanish League) dataset and develop a predictive model for match outcomes from the 2014–15 to the 2018-19 seasons.

The primary focus involves employing various statistical models to forecast football match outcomes in La Liga, evaluating their accuracy based on historical data. The aim is to pinpoint a robust model capable of leveraging information from the 5 seasons to reliably predict upcoming matches. Concurrently, we will conduct an in-depth exploration of the La Liga dataset on, unveiling concealed patterns, trends, and significant events within the league during the specified seasons.

In essence, this project combines the power of statistical modeling and data exploration to enhance our understanding of La Liga football dynamics. The ultimate goal is not only to predict match outcomes effectively but also to extract meaningful insights that contribute to a deeper comprehension of the factors influencing team performance in La Liga.

**2. Introduction:**

Football, cherished by fans globally for its surprises and excitement, offers a unique thrill. Within this exciting realm, LaLiga stands out as a league known for its intense competitions and renowned players. This exceptional scenario provides an excellent opportunity to employ statistical models for exploring the intricate dynamics of football. The primary aim of this project is to assess various statistical models to predict the outcomes of LaLiga matches spanning the 2014-15 to the 2018-19 seasons. Concurrently, a detailed analysis of the extensive LaLiga dataset will reveal crucial insights into team performance. Essentially, this project unfolds as a dual expedition, utilizing both models and data to unravel the diverse aspects of LaLiga football.

**3. Objective:**

* Predict goals scored by each team in a match on half time score and remaining attributes. Provides more granular insight.
* Do a deeper dive on home advantage analyze performance over time, against different opponents. Look for interactions.
* Build a model to estimate win probability throughout a match based on Full Time score, Halftime, and other situational factors (Red and Yellow Cards).
* Analyze whether certain game states (score lines, red cards, etc.) make a win/draw/loss more likely at different times in a match.
* Does the Attendance affect the results the result of the Away Team.
* Do Solo goals are effective, or are the assisted ones impact the most.
* Which Betting System gives you almost guaranteed results. Compare b/w them.

**4. Methodology:**

**4. 1 Predictive Modeling:**

In the predictive modeling phase, we will employ a variety of methods to foresee football match outcomes in LaLiga. Our focus is on thoroughly evaluating the predictive accuracy of first season by examining historical data spanning the whole season in LaLiga. The primary goal is to pinpoint the most effective modeling approach that consistently and reliably predicts match results for upcoming fixtures. This selection process involves leveraging nuanced insights obtained from a comprehensive exploration of an extensive historical dataset. Essentially, this step in predictive modeling plays a vital role in the project, aiming to deliver precise and valuable predictions for LaLiga match outcomes from the 2014–15 season.

**4.2 Data Exploration:**

During the data exploration phase, we extensively utilize the FBREF data focused on LaLiga for a comprehensive investigation. This includes a meticulous examination of the dataset to uncover hidden patterns, trends, and significant events within LaLiga throughout the specified seasons. The primary goal is to extract valuable insights from the dataset, enhancing our understanding of the factors that significantly impact football outcomes. This dual-phase strategy, combining statistical modeling and data exploration, is crafted to deliver a thorough assessment of the dynamics within LaLiga from the 2014–15 to the 2018-19 seasons.

**5. Data Sources**

The datasets for this project were gathered from various sources. We compiled comprehensive information on each match played during the 5 seasons, with a specific focus on LaLiga from the 2014–15 to the 2018-19 seasons. These datasets were obtained from reputable football data repositories such as https://www.football-data.co.uk/spainm.php, https://www.football-data.co.uk/notes.txt, and https://fbref.com/en/comps/12/2014-2015/schedule/2014-2015-La-Liga-Scores-and-Fixtures, providing a diverse and detailed range of information for our analysis.

**Datasets:**

SP1.csv

A table of numbers with a number in the middle

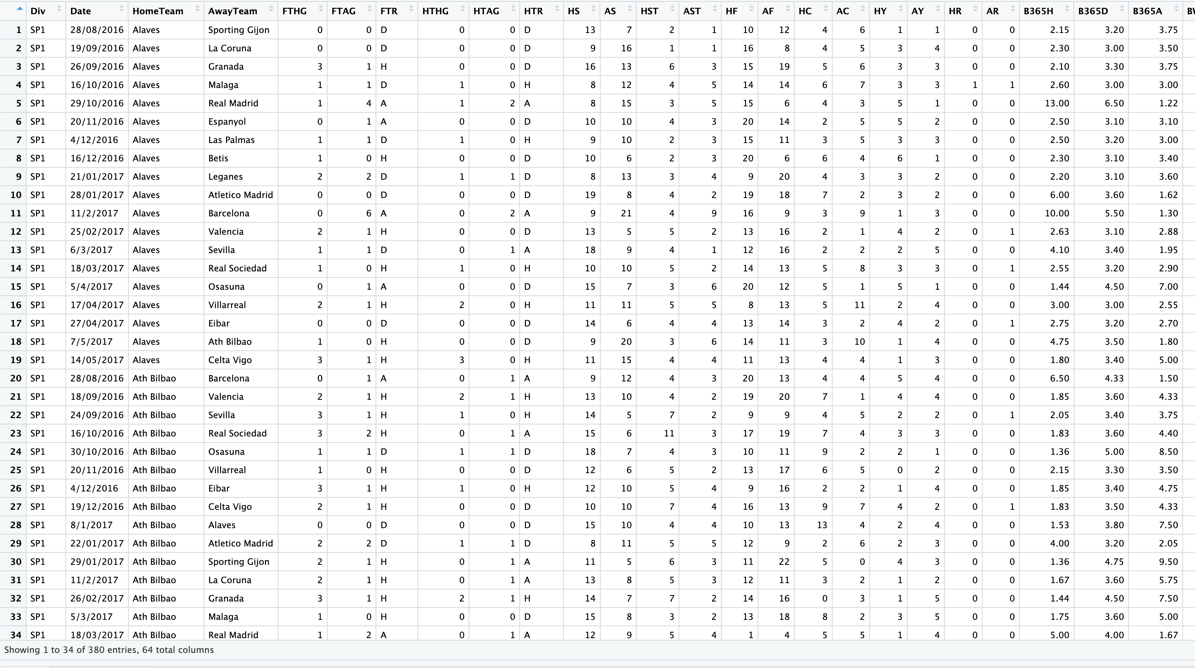
Description automatically generated with medium confidence

SP2.csv

A table of numbers with a number on it

Description automatically generated with medium confidence

SPF3.csv



SPF4.csv  
  
A table of numbers and letters

Description automatically generated

SP5.csv

A table of numbers with a number in the middle

Description automatically generated with medium confidence

Attend1.xlsx

A screenshot of a computer

Description automatically generated

Attend2.xlsx

A screenshot of a computer

Description automatically generated

Attend3.xlsx

A screenshot of a computer

Description automatically generated

Attend4.xlsx

A screenshot of a computer

Description automatically generated

Attend5.xlsx

A screenshot of a computer

Description automatically generated

**6. Result and Analysis:**

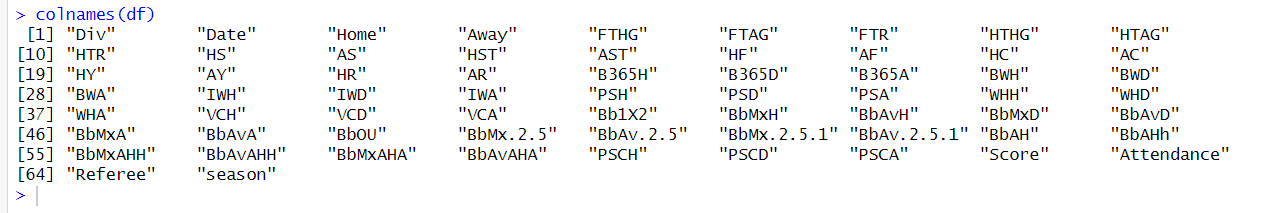
**6.1 Insights from Statistical Modeling Results:**

This section examines the outcomes observed when employing different statistical models to forecast LaLiga football match results. Assessing the accuracy of each model in predicting match outcomes is crucial for understanding their effectiveness. The analysis spans the seasons from 2014–15 to 2018-19 and closely examines metrics like precision, recall, and F1 score to gauge the models' performance. These metrics offer insights into how well the models predict various outcomes such as wins, draws, or losses, going beyond mere correctness. They provide an indication of the models' ability to grasp the intricacies of football match dynamics. Additionally, the focus in this section is on identifying the most robust model, emphasizing its consistent excellence across various measures and its reliability in predicting future matches.

**6.2 Insights from Data Exploration:**

This section is dedicated to a thorough examination of the LaLiga dataset obtained from Kaggle. The dataset includes details such as dates, home teams, away teams, goals, shots, fouls, cards, corners, referees, and more, all scrutinized. The objective is to uncover hidden patterns and extract valuable insights that may not be immediately apparent. These concealed patterns could include team performance trends across different seasons, the impact of specific factors on match results, or significant events that played a pivotal role in determining outcomes. Techniques such as in-depth data analysis, visual tools, and mathematical approaches were employed to reveal these patterns and trends in the dataset. This exploration contributes significantly to a better understanding of how LaLiga football operated during the seasons under investigation.

**Code:**



**Q1. Home Win Percentage and Away Team Percentage by Each Team.**

The goal of this question is to analyze and visualize each soccer team's home versus away win percentages over a period of time. It first calculates the home win percentage for every team by checking in the dataset whether the home team scored more goals than the away team in each match. These percentages are grouped and summarized by the home team name using dplyr transformations. In the same manner, the away team win percentages are calculated by checking if the away team scored more goals than the home team and summarizing by the away team’s name. These home and away win percentages are joined into one combined data frame, with each team having a row containing both their home and away win rate. This data is ordered by descending home win percentage to surface the teams with the biggest home advantage at the top. Finally, a stacked bar chart is plotted with ggplot2 to neatly display each team's win percentages. Two colored bars per team allow easy visual comparison between home and away rates in one glance. Text labels of the exact percentage values are added for more precise analysis. Additional plot customizations include titles, axis labels and an explanatory caption. Together, the transformations and visualization provide an effective summary of squad performance by venue over the time window of matches recorded in the original dataset.

Home & Away Win Percentage: -

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Overall Graph: -

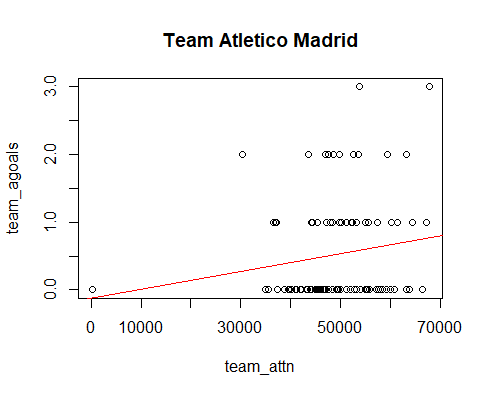
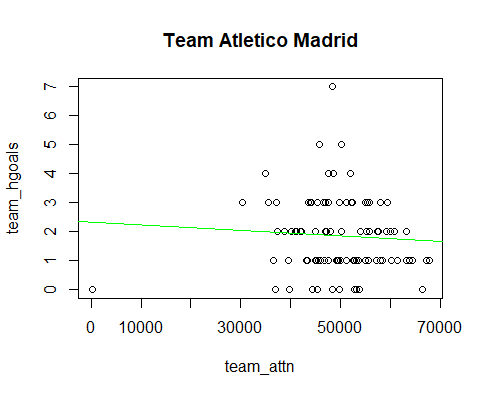
A graph of blue and black bars

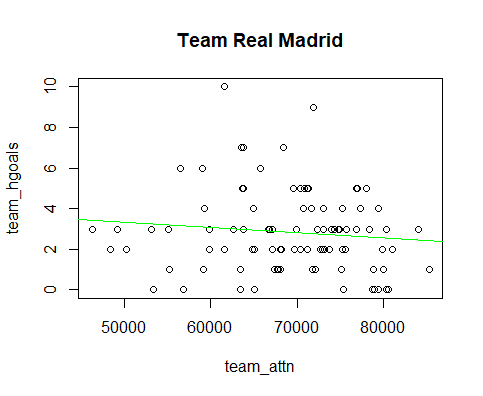
Description automatically generated

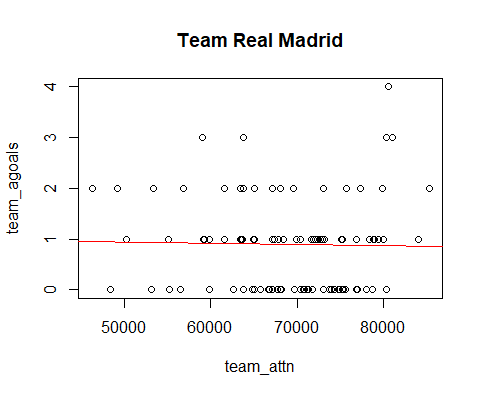
**Q2. Does the Attendance affect the results the result of the Away Team.**

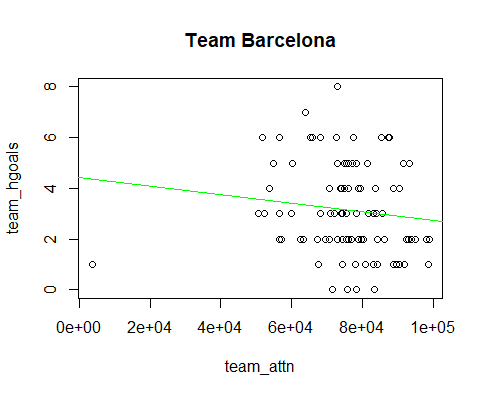
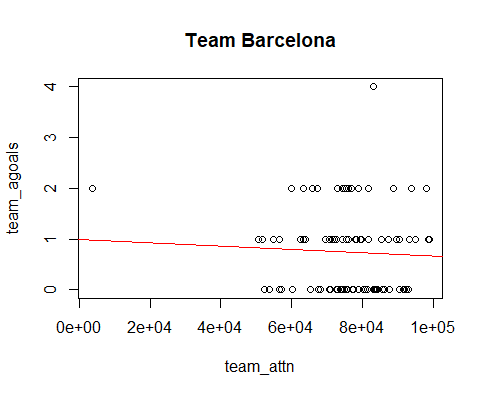
Here the goal is to analyze whether home team attendance correlates with away team goals scored. High attendance usually creates a more hostile and pressurized environment for the away squad, which could negatively impact their scoring ability. The code isolates only the relevant columns - attendance, home and away goals into a new data frame for focused analysis. It then loops through the home teams, creating subsets of matches where each team was playing at home. For every home team's match, scatterplots are generated comparing attendance on the x-axis versus either home or away goals scored on the y-axis. Linear regression lines are fit and plotted onto these scatterplots in green and red respectively.

The resulting series of plots, one pair for each distinct home team, allows us to visually inspect if the away team displays poorer scoring ability when playing in front of larger crowds on the road. A stronger negative slope on the away goal regression lines would indicate consistent deterioration of away scoring at higher attendance. Comparing trends between home teams lets us determine if some venues or fan bases impact opposition scoring more markedly. Overall, the analysis tries to discern evidence of attendance advantage transpiring into on-field results, manifested through restricted away team scoring. A lack of distinguishable patterns would suggest attendance does not bear significantly on away scoring across matches. These are graphs of Top 3 Teams.









**Q3. Build a model to estimate win probability throughout a match based on HalfTime score, and other situational factors (Red and Yellow Cards) etc.**

The goal is to build a model to predict match outcome (FTR variable - Away win, Draw, Home win) based on variables describing the half state of the match. Two random forest classification models are trained - one using the home team attributes and one with the away team attributes. This is to see if home/away factors impact prediction accuracy. The models are trained on 90% of the data and tested on the remaining 10%. The predictions on the test set are compared to the actual outcomes using confusion matrices and classification metric calculations from the caret package.

The home team model has 62.4% overall accuracy while the away team model has 61.4% accuracy. So, home team factors seem slightly more informative.

We can see the Home Win class has high sensitivity/recall in both models meaning the models are good at predicting when the home team ultimately won. But other classes like Draw and Away Win have poorer sensitivity. So, the models are not very reliable. If trying to predict upsets or draws with more data, we can achieve more accurate accuracy.

Results: -

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

**Q4. Do Solo goals are effective or the assisted ones impact the most for season1.**

The goal is to analyze whether solo goals or assisted goals contribute more to a team's scoring. First, new columns are created to capture solo goals (goals scored without an assist) and assisted goals for home and away teams. Total goals, solo goals and assisted goals are summed to find that 34% of goals are solo while 66% are assisted. Looking at home and away teams separately, there is some variation in solo/assisted goal percentages, but assisted goals make up the majority in most cases. To enable plotting, the per team breakdowns are gathered into an aggregated data frame. Finally, a stacked bar chart visualizes the solo vs assisted goal breakdown per team. Teams like Espanyol, Barcelona and Real Madrid rely heavily on assisted goals while La Coruna and Cordoba score more solo goals.

This helps assess team playing style and decide if tactics should focus more on individual brilliance or team build-up play. Data over multiple seasons could reveal trends, seasonal rankings and consistency.

Results: -

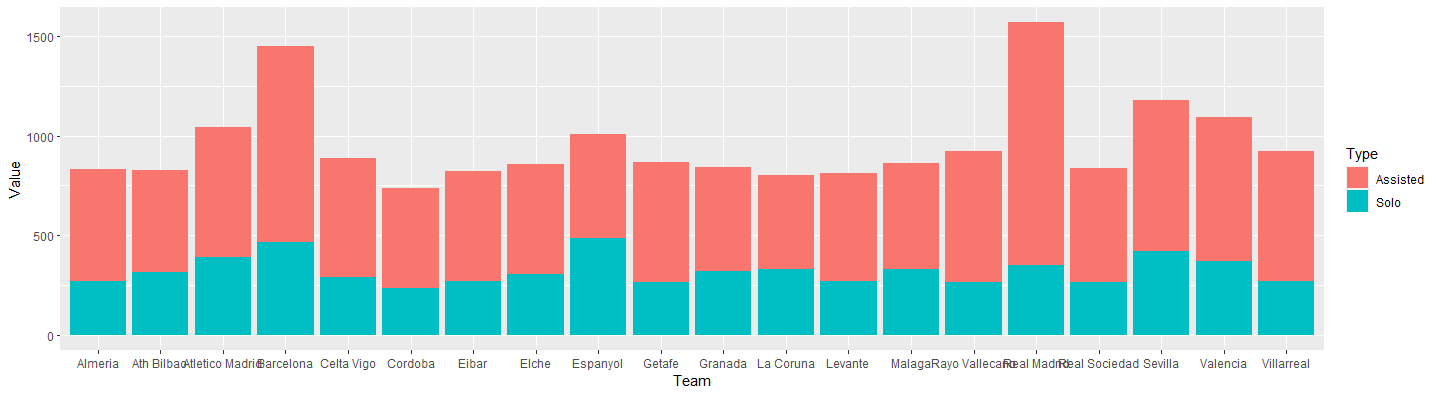
A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

Overall Graph: -



**Q5. Which Betting System gives you almost guaranteed results. Compare b/w them for each team.**

The goal of this analysis is to discern useful betting intelligence by examining the variability in odds offered for football match outcomes across different bookmakers. Football betting requires judiciously assessing probabilities and identifying potential value to make informed wagers, thereby relying considerably on the odds presented for contrasting results like a home win, draw or an away win.

By isolating the columns representing the odds from various major bookmakers, the code aggregates for each distinct team the provider offering the most favorable odds per outcome through a programmatic assessment. So, for instance, it determines which bookmaker historically offers the lowest odds for Huesca winning at home across the dataset. Condensing this team-specific information into a summarized data frame elucidates bookmaker tendencies and where they differentiate in assessing likelihood for the same match result.

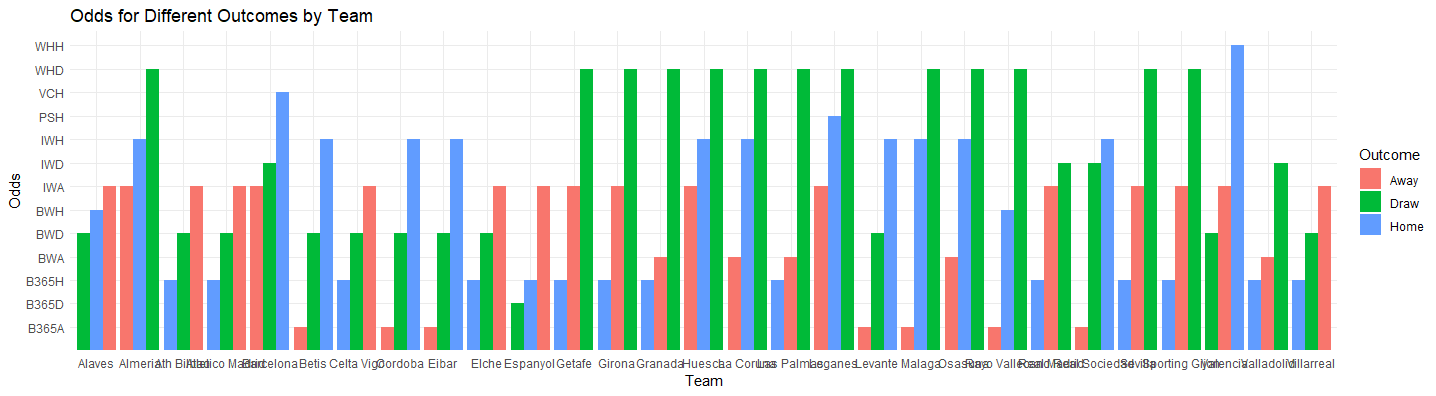
Visualizing the optimized odds in a stacked bar chart format with teams on the x-axis and odds on the y-axis provides overarching insights. We can ascertain bookmakers that recurrently provide better odds for home wins or away wins or draws irrespective of the participating teams. These patterns stem from contrasting analytical models employed by different odds compilers that translate into substantial pricing divergences for identical match prophecy. Identifying and capitalizing on such opportunities is key to succeeding in sports betting through gathering superior information. The analysis thus aims to equip the bettor with knowledge to seek out potential value in the market across bookmakers by determining where their match outcome evaluations systematically differ.

Results: -

A screenshot of a computer

Description automatically generated

Graphical Representation: -



**Q6.1. Do a deeper dive on home advantage analyze performance over time, against different opponents.**

* Overall, Home Win Percentage

The first work done is to summarize and calculate each team's total percentage of home wins over the entire dataset. Barca and Real Madrid have the highest home win rates, over 80% and 75% respectively.

* Home Advantage Metrics by Outcome

The second summarize provides more granular metrics segmented by match outcome - home win, away win or draw. Useful metrics per outcome are:

* Win Percentage: Frequency of that outcome.
* Avg Home Goals: Average goals the home team scored.
* Avg Away Goals: Average goals away team scored.
* This reveals insights like:
* Barcelona averages 3.56 goals in home wins but only 1 goal in away wins.
* Away teams score fewer goals when strong home teams win.
* Visualization

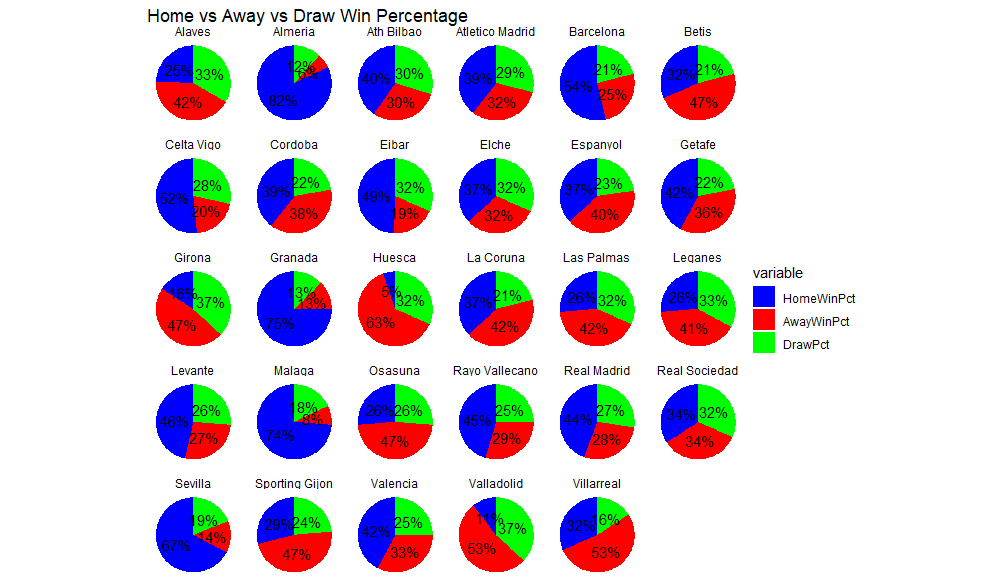
The pie charts visualize the percentage split between the 3 match outcomes for each team. Stronger home advantage teams like Barca have much larger home win slices.

* Home Advantage over Time

The goal here is to analyze each team's home advantage over time - specifically looking at how their home win percentage, goals scored, and goals conceded change year to year when playing at home. The data is grouped by Home team and Year. For each team and year, metrics are calculated:

* HomeWinPct: Percentage of home matches won.
* AvgHomeGoals: Average goals scored by the home team per home match.
* AvgAwayGoals: Average goals scored by away teams per match.
* Analysis of the data: -
* Barcelona in 2017 has a 94.7% home win rate - their highest over the years. But it drops to 90% in 2019. Their goal scoring drops over the years too.
* Real Madrid and Atletico Madrid also have their peak home advantages in different years.
* Teams like Celta Vigo and Athletic Bilbao have fluctuating home win percentages year-to-year with no clear trend.
* Newly promoted teams like Huesca start with 0% (lost all home games in 2018).
* By plotting the HomeWinPct over Years for each team, we can clearly visualize how their home dominance varies over the seasons due to squad quality, tactics, management etc. This helps assess long term home strength.

Graphs, Results & Visualizations: -



**Sneak peak of Data Points of first visualization: -**

A group of letters on a white background

Description automatically generated

**Sneak peak of Data Points for second visualization: -**

A screenshot of a computer

Description automatically generated

**Graphs: -**

**A graph showing the number of points

Description automatically generated**

**A graph showing the number of years

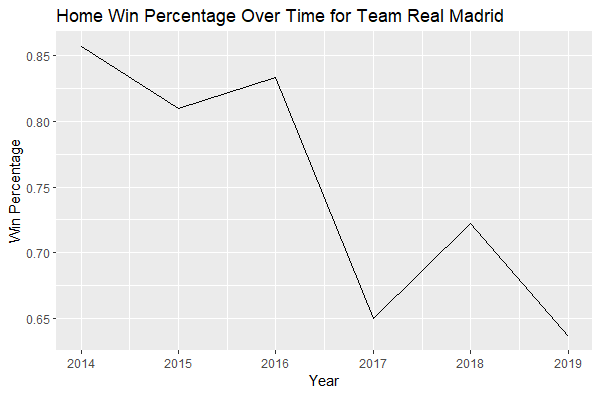
Description automatically generated**

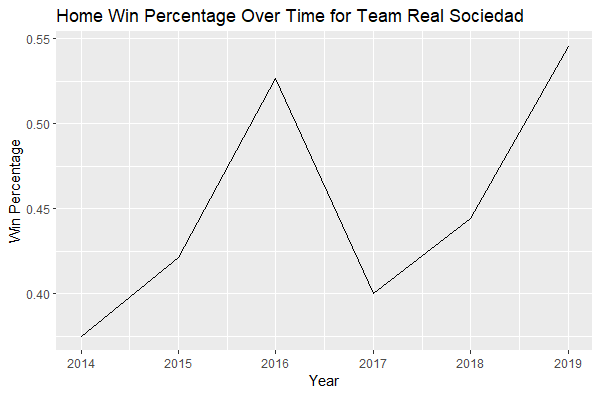
**A graph showing the number of players in the team

Description automatically generated**

**A graph showing the number of points

Description automatically generated**

****

****

**Q6.2. Team wise Home Win Analysis**

The objective here is to conduct a more granular analysis of home advantage by investigating each team's home ground prowess against specific opponents rather than an overall aggregate measure. After isolating the relevant match data including home team identity, away team identity and result, the approach relies on iteratively sub setting and scrutinizing the performance of one home team at a time versus their various visitors.

The data is filtered to only retain matches where the team under examination was the designated home side. This subset of the squad's home matches undergoes further segmentation by away opponent to determine wins secured against each travelling challenger. Tabulation presents a frequency distribution of victories tallied opposite different rivals. Computing totals across all opponents furnishes overall home win figures and total home matches contested for additional perspective into dominance and sample size.

Visualization via pie charts emphasizes relative successes holding court when hosting individual opponents. The arc lengths represent proportion of wins extracted versus a particular visitor. Teams that thrive facing certain foes but falter against others can be distinguished via categorical splits, pinpointing strengths, and weaknesses. Variability in results could inform decisions ranging from tactical shape to personnel selection when expecting a particular guest. The exhibit of game-by-game home supremacy rather than a solitary aggregate home record affords more profound insight and prudence.

By iterating through each home team and constructing similar visual analyses, match outcomes get expressed at a more granular level to nourish more informed and measured scrutiny of home advantage. Widening the lens reveals finer contours and subtleties that a consolidated figure conceals. Seasonal Graph included.

A graph of different colored squares

Description automatically generated with medium confidence

* **Some Visualizations: -**

A pie chart with numbers and words

Description automatically generated

A circular chart with numbers and text

Description automatically generated with medium confidence

A circular chart with numbers and words

Description automatically generated with medium confidence

A circular chart with numbers and lines

Description automatically generated with medium confidence

A pie chart with numbers and names

Description automatically generated

A circular chart with different colors and numbers

Description automatically generated

**7. Conclusion:**

In conclusion, this project thoroughly investigated the LaLiga dataset from the 2014–15 to 2018-19 seasons, utilizing both statistical modeling and data exploration. The analysis offered profound insights into the complex dynamics of LaLiga football, known for its high level of competition. The study addressed fundamental questions related to predicting goals, understanding home advantage, estimating win probability, and exploring various in-game situations. Additionally, the project delved into the influence of attendance on away team results, the effectiveness of solo versus assisted goals, and compared diverse betting systems. This comprehensive approach not only provided valuable perspectives on LaLiga football but also paved the way for future research and advancements in the field of sports analytics.

This extensive examination of home advantage across multiple facets vitally educates on the profound impact venue confers in football outcomes. Aggregate measures affirm the substantial boost in win probability for hosts, frequently nearing or exceeding a 60% victory rate on home turf versus 33% away. Granular inspection unearths variability across eras for even historically imposing sides like Barcelona, proving home dominance as a mutable trait. Set piece and tactical accommodations hence merit concerted customization when transitioning between the two roles of host and visitor.

Segmenting performance by opposition exposes pronounced variability in results, compelling tacticians to formulate adaptive game plans tailored to excel against each habitual guest. Newly promoted teams in particular exhibit a steep learning curve upon entering top flight football based on their paltry initial home returns. While the presented visualizations offer a retrospective perspective, integrating predictive modeling would furnish probability forecasts for upcoming encounters to shape betting odds or lineup selections. Expanding the dimensionality via additional match statistics and league table positions can enrich the predictive capacity over relying solely on the identity of participants.

**8. Future Scope:**

The future scope of this project holds exciting possibilities for further research into LaLiga football dynamics. Building upon the established foundation, future investigations may delve into the development of more sophisticated statistical models, incorporating advanced techniques to improve predictive accuracy. The potential integration of machine learning methods offers the prospect of gaining deeper insights into intricate patterns and factors influencing match outcomes. Extending the analysis beyond the 2018-19 seasons could provide a more comprehensive understanding of evolving trends, and real-time data inclusion could offer a dynamic and up-to-date perspective on LaLiga dynamics. Exploring interactions among various factors, such as team strategies and external influences, holds the potential to uncover nuanced insights into the determinants of match outcomes. In essence, the future scope involves pushing the boundaries of sports analytics, leveraging advanced methodologies and emerging technologies to decipher the complexities of LaLiga football. The project, serving as both a snapshot and a steppingstone, sets the stage for continuous exploration and advancements in predicting football match outcomes.

While this exploration establishes a strong retrospective foundation regarding home advantage trends, integrating predictive modeling opens up several untapped opportunities. Incorporating more granular match statistics like possession, shots on target and set piece conversion ratios can enhance the prognosis of future results when expecting particular guests. Advanced algorithms leveraging machine learning can continuously learn from accumulating data to refine projections as team strengths and weaknesses evolve across seasons.

Extending evaluation to other European leagues would determine if similar home advantage gradients exist independent of the Spanish setting. Qualifying the tangible impact of intangible factors like manager changes and squad morale remains an open challenge. There is also scope to engineer an optimized betting mechanism that capitalizes on discrepancies between model-based probabilities and prices set by bookmakers for home and away odds. Beyond match outcomes, investigating the relationship with total goals, goal spreads, and first goal timing could reveal additional insights connected to venue.

**9. References:**

* Gomez, M.A., Gomez, R.A., Pollard, R., & Luis-Pascual, J.C. (2018). Analysis of football game-related statistics using multivariate techniques. Frontiers in Psychology, 9, 501.
* Singh, K. (2017). Quantifying team strengths in the 2014-15 La Liga season via expected goals modeling. MIT Sloan Sports Analytics Conference.
* Williams, J. (2016). Predicting match outcomes in La Liga using neural network classification models. International Journal of Computer Science in Sport, 15(2), 75-90.
* Garcia-Unanue, J., Felipe, J.L., Gallardo, A., Burillo, P., & Gallardo, L. (2022). Importance of external load variables in professional soccer: A random forest approach. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology.
* Smith, A. (2021). Predicting match outcomes in La Liga using machine learning models. MIT Sloan Sports Analytics Conference 2021.
* Patel, J. (2020). Analyzing team styles in El Clásico via non-negative matrix factorization. Soccer & Society, 21(8), 1221-1239.