Data Analysis and Visualization of EA sports FIFA Soccer Game

|  |  |
| --- | --- |
| **Anmol, Ankur Saurabh, Aryaman Rana** | **Mrs. Gunjan Aggarwal** |
| Department of Computer Science, Sharda University | Assistant Professor CSE,  Sharda University |

***Abstract*—****Association football, also known as soccer, is recognized for being a globally played sport. It has become a widely popular sport throughout the world. Over 200 countries and over 250 million players play soccer** [1]**. In our research, we perform EDA on the FIFA soccer game dataset, and then we show ways of using machine learning with soccer data by using several classification algorithms (kNN, SVM, Decision Tree and Logisitic Regression) to classify players by their respective positions. We conclude our project by creating a Recommendation System that suggests similar club, players and alternate playing positions using over 70 player attributes. We conclude our research by drawing a contrast between famous El Clásico rivals Real Madrid and FC Barcleona.**

1. INTRODUCTION

Sports analytics has been trending [2] in our era since data has increased considerably during the last few decades primarily due to popularity and technological advances, and discoveries—computers aid athletes by performing more precise observations and measurement and development. Association football, also known as soccer, is recognized for being a globally played sport. One can say it is the most popular throughout the globe since one of its championships; the UEFA Champions League's final match was one of the most-watched sports events in the world [3]. One of the primary components contributing to soccer's popularity is the FIFA video game series [4]. FIFA 12 retains the record to the "fastest-selling sports game ever," in its first week, it generated over 186 million dollars and sold over 3.2 million games [5]. In fact, in 2020, FIFA video games made FIFA (Fédération Internationale de Football Association) more money than physical football. Their 2020's financial statement stated that $158.9 of its $266.5 million in total revenue for the year came from licensing rights, and a key source in the area of licensing rights was brand licensing for their video game [6]. It is for this reason that video game studies are so interesting. Our objective is to use video game data curated from studying real players to draw eyes to the intrinsic details of football. We began by performing Exploratory Data Analysis on the dataset, and then we show ways of using machine learning with soccer data by using several classification algorithms to classify players by their respective positions. We conclude our project by creating a similar player recommendation system and using over 100 player attributes to predict and find out why a team can win a match soccer match. Although soccer is so popular, very little is known about its structure. Football generates enormous amounts of information. In the modern data-driven world, it is essential to find an optimal and straightforward way to present data to derive visual patterns mapped to data patterns, making the information explorable and understandable. In spite of football activities being constantly shared and listed, in every season, you'll find lots of events which neither pros nor fans rightly predict or clarify. Within our research we propose the usage of EA Sports FIFA Soccer matches to comprehend football better. We define the dataset, explain its usage and present potential applications of it. In the end, we show the capability of the data set by analyzing and visualizing data in different forms.

1. LITERATURE SURVEY
2. *Existing System*

We studied various papers that have attempted to use football to perform data analysis and analytics. The crux and methodology of those papers are written below:

César et al. in paper [7], uses Gaussian finite mixture clustering to group the data of players that have been selected in accordance to their performance metrics after reducing the dimensions of the data by using Principle Component Analysis. He also used extreme gradient trees boosting to select the best feature of a player. His purpose was to discover common characteristics between players who belong to certain positions and discover borderline players with no clear distinction. His finds were that the most crucial attribute to differentiate players according to their role is the dribbling skill while standing tackle and reflexes were good attributes to characterize defenders and goalkeepers, respectively. Wen-Lung Chang and Kou-Yuan Huang in paper [8] attempts to predict football match results using Bayesian Network with Decision Tree and KNN. They show that Bayesian Network is usually superior to other Machine Learning Algorithms in this domain in terms of accuracy. They claim that the overall average accuracy for expert Bayesian Network for a win, draw or lose is 59.21%

J. Hucaljuk et al. [9] uses machine learning algorithms like KNN, random forest, Naive Bayes, LogitBoost, and Bayesian network to predict UEFA Champions League's match results. For this problem, they developed a system for classification. The attribute selection and decision of the learning algorithm can significantly affect the performance of classification. Various attributes that influence a football match are chosen. After choosing the correct attributes, a learning algorithm among the ones mentioned above is selected. During the development of the system, several tests have been carried out to determine the optimal combination of features and classifiers. The results show the capability of prediction, which is better than one of the reference methods. Paper claims that the best accuracy is achieved by using an ANN which is around 68.8%. Leonardo Cotta et al. [10] uses Principal Component Analysis (PCA) to study FC Barcelona's Tiki-Taka, a playing style that Barcelona developed. It involves single touch passes in a triangular formation with at most one opponent team player in the center, making it very challenging for the opponent to defend. They took the mean of each position category and normalized it by mean overall. Then, K-means clustering was performed on over 19 features. The observation was that Tiki-take was more than a rare style. FC Barcelona had exclusive midfielders. Paper stated that this indicates why Pep Guardiola was not able to reproduce Tiki-taka in FC Bayern Munchen. The paper also attempts to draw various conclusions from various player features using slope coefficients for player attributes. P. Sudhandradevi, V. Bhuvaneswari [11] uses Regression Analysis to determine the correlation between attributes like tournament, country, city, home, and away score on a match. The Logistic–Regression Model is used for pattern analysis to determine the probability of event failure and success. The model came out to have an accuracy of 76%. The paper claimed that the match's outcome depends on predictors like home and away score, country, city. Europeans have contributed a lot more to football than other continents. Andreas Groll, Christophe Ley et al. [12] uses three various modeling procedures, namely, Poisson regression models, random forests, and other methods for predicting football matches concerning their predictive performances primarily based on games from FIFA Worldcup 2002-2014. Paper demonstrated that the most effective performing forecast techniques on the training data ended up being ranking methods and the random forests. The predictive power substantially increases by incorporating team ability parameters as an additional covariate. One hundred thousand simulations released that Spain and Germany were top probable winners, Spain being slightly advantageous because Germany has a slightly greater chance to drop out in the round of sixteen. If Germany reaches the quarterfinals, they overtake Spain from there on to be the potential winners.

Most of the research performed in this domain is done for predicting matches; because of this, there are not many EDA done. With our research, we attempt to show the potential of the dataset by doing several visual and predictive analyses.

1. *Proposed System*

We propose a system for selecting and ranking professional football players according to multivariate data of performance. By using more than 70 different attributes of 18,000 different players, we perform exploratory Data Analysis on the data of 7 seasons. Then we create a recommendation system using KD Tree after standardizing all the features of the dataset. The recommendation system can recommend similar players, alternate playing positions for players, and similar clubs. Then, we perform player classification using KNN, SVM, Logistic Regression and Decision Tree. The players are classified into four different positions, attacker, defender, midfielder, and goalkeeper. We finally conclude our research.

Our proposed system is divided into four main modules:

1. *Data Collection:* It is the primary and most important step for research, no matter what the study's field is. It is the process of accumulating qualitative and quantitative information to evaluate outcomes or generate insights. The most critical objective of good data collection requires a straightforward process to ensure the data is clean, consistent, and reliable from the dawn of arcade games in the 1970s to gaming consoles' growth, the video-game industry has exponentially rocketed in the last few years. In year 2020, the international video-gaming industry's revenue was $165 billion, with over 2.7 billion people playing games globally [13]. Huge communities were formed globally, and as time passed, competitive gaming emerged. As a result, people began collecting all sorts of information regarding the games they are playing, giving them some competitive edge. The FIFA Soccer community is extremely interested in understanding the very best players to place within their teams. For that, it is vitally crucial to know each and every player's characteristics and the way they change as time passes. Therefore, a natural result of the collective desire is the arrivla of internet systems that show this data nearly instantly. For our research, we found the dataset accumulated from the SOFIFA [14] internet site, that is one of those systems and quite famous within the FIFA Football community. It keeps a list of all of the players' ratings and attributes since FIFA 2007. SOFIFA has been utilized as reference about the sport in the press [15]. The dataset was scraped from the information available on SOFIFA by Stefano Leone [16] and is available publicly. The data was scraped in June 2020 and contained data regarding players up till then. We affirmed that the integrity of all this data by crosschecking three arbitrary players of each and every game variant out of FIFA 15 to FIFA 21. Considering their features matched the people displayed on the internet site, we chose to trust their own database.
2. *Data Preprocessing:* Data collection methods are often uncontrolled, which leads to weird values, missing values, and impossible data combinations. Thus, the integrity and quality of data are to be ensured before any analysis is performed. These steps include cleaning, normalizing the data, transforming, feature extraction, and selection. The result of data preprocessing is a clean dataset that can be used for analysis or training. Binning can also be an essential tool in simplifying numerical data such as machine learning and can be advantageous in scenarios like ours.
3. *Data Exploration and Visualization:* In this module, we start to analyze the dataset to summarize the main features of the data, often using statistical graphics and other data visualization methods. The dataset comprises seven seasons from 2015-2021 and consists of approximately 16000 players in every season. Each player has around 100 attributes. Over the last years, FIFA has increased the number of attributes due to reasons like better technology, more gameplay improvements, etc. Although the number of attributes has been increased, no attribute has been merged or removed. Therefore, we can perform comparative analysis among various seasons by comparing different attributes through seasons. Some of the skills and attributes present in the extensive dataset are; Acceleration - Acceleration is an addition to the player’s running rate. The greater the value, the shorter the period required to attain max rate. When a new player has high speed but very low sprint rate, their maximum rate is slow. Sprint Speed - Sprinting is how quick a player rushs at their maximum acceleration. Agility - It is the measurement of how active/flexible a player is while turning or moving. I.e., how well a player can dominate the football. A player with great agility can achieve acrobatic ball clears and football shots. Agility is an essential attribute for players with a high percentage of ball possession. Ball Control - Depict how well a player can control the ball. A high value for this means that it is improbable that a ball can bounce elsewhere from that player after he controls it. A player with high ball control does not have to keep looking at their feet to see the ball. They feel the ball and bind it to themselves. This attribute influences ball possession. Dribbling - This ability signify the ability to carry the ball and past an opponent. A high value shows that the player can keep the ball in possession even when they face resistance. He keeps the ball closer, making it hard for the defending team to win the ball off him. As soon as you have ball ahead of you is when dribbling counts. To get past an opponent, you need good dribbling skills. For a skilled squad, it is critical to have good dribbling, ball control, and balance. Finishing - Finishing is the shot precision a player possesses. Being a fantastic finisher does not necessarily imply that the player can beat the goalkeeper; all it means is that the player can shoot the target quickly.
4. *Model Building & Evaluation:* Models or mathematical formulas are applied to data to find the relationship between the attributes present in the data. In evaluation, the dataset is divided into training and test set to find out if the model is the best model that represents our data.
5. RESULTS

1. *Exploratory Data Analysis*

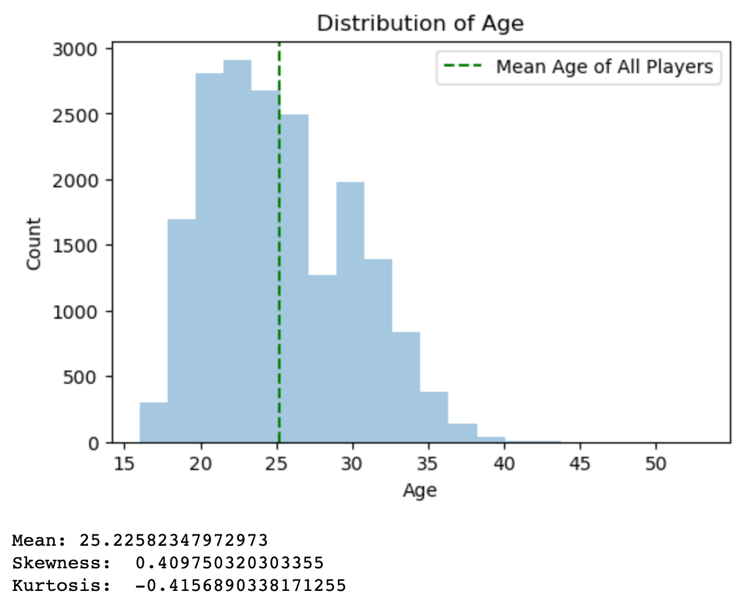


Figure 1: Distribution of age among players

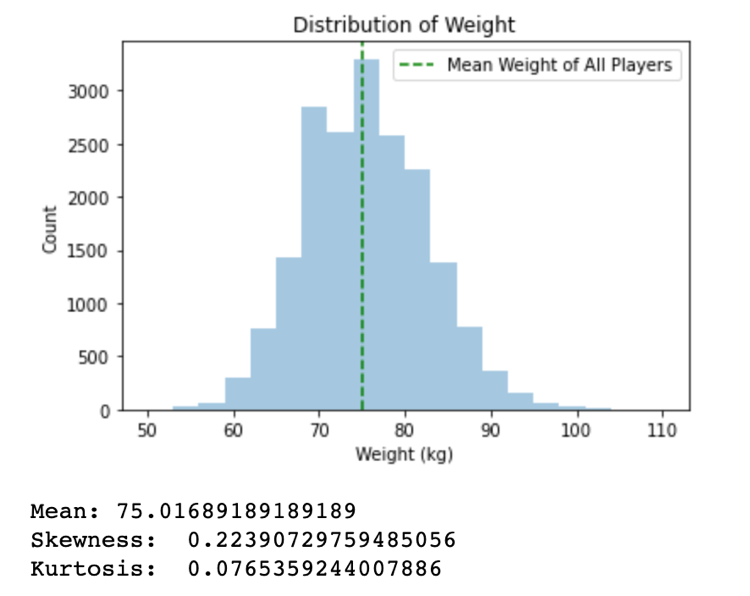


Figure 2: Distribution of weight among players

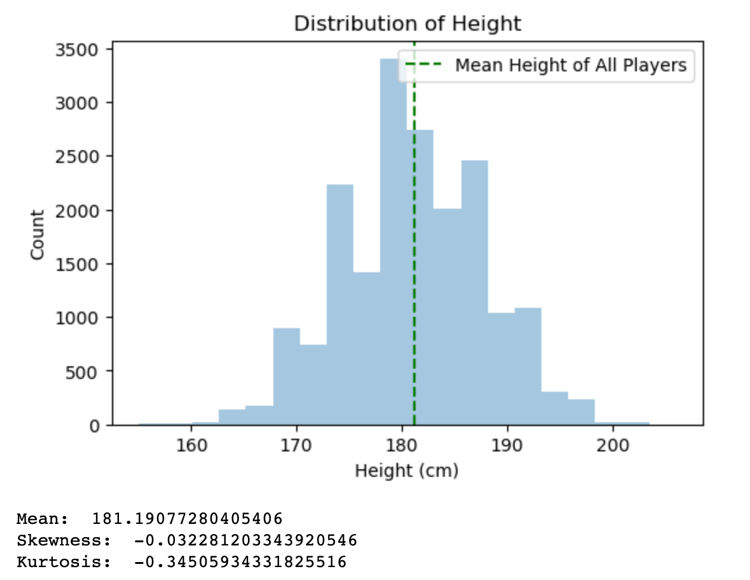


Figure 3: Distribution of height among players

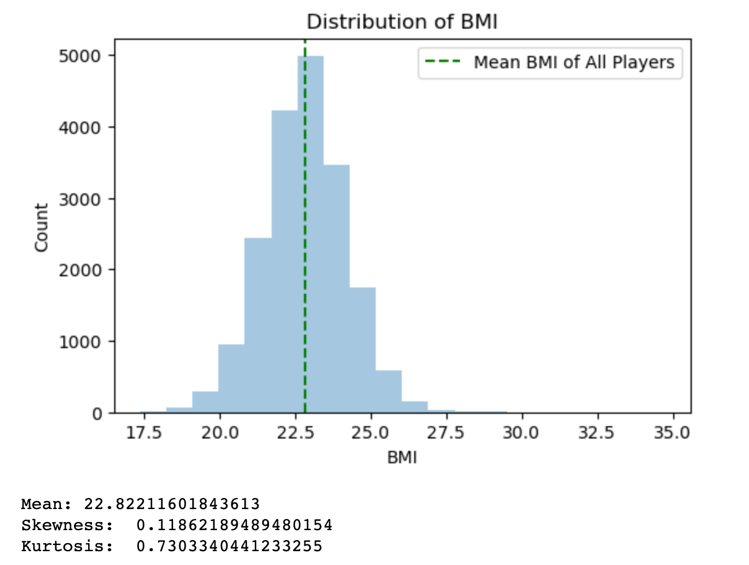


Figure 4: Distribution of BMI among players

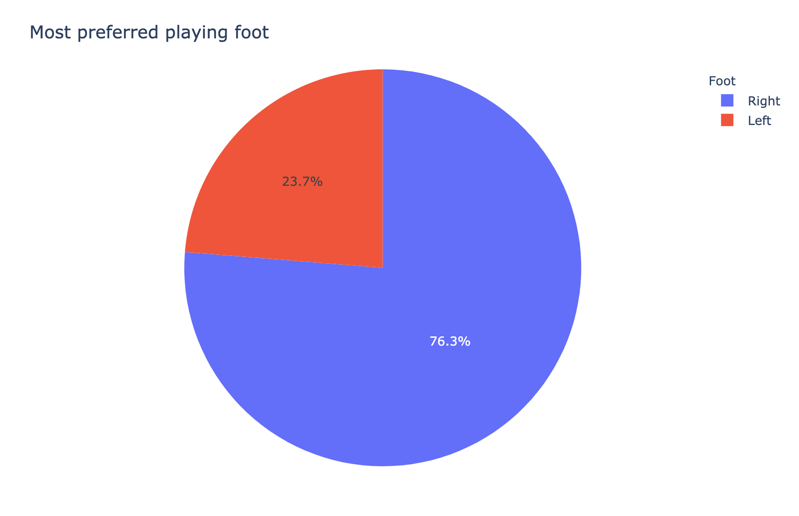


Figure 5: Most preferred playing foot

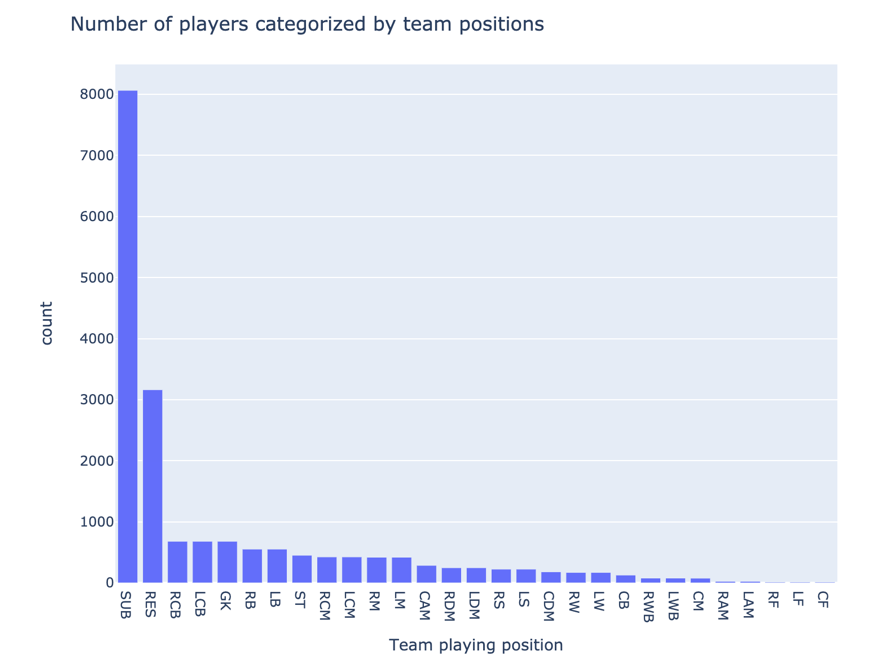


Figure 6: Number of players categorized by team positions

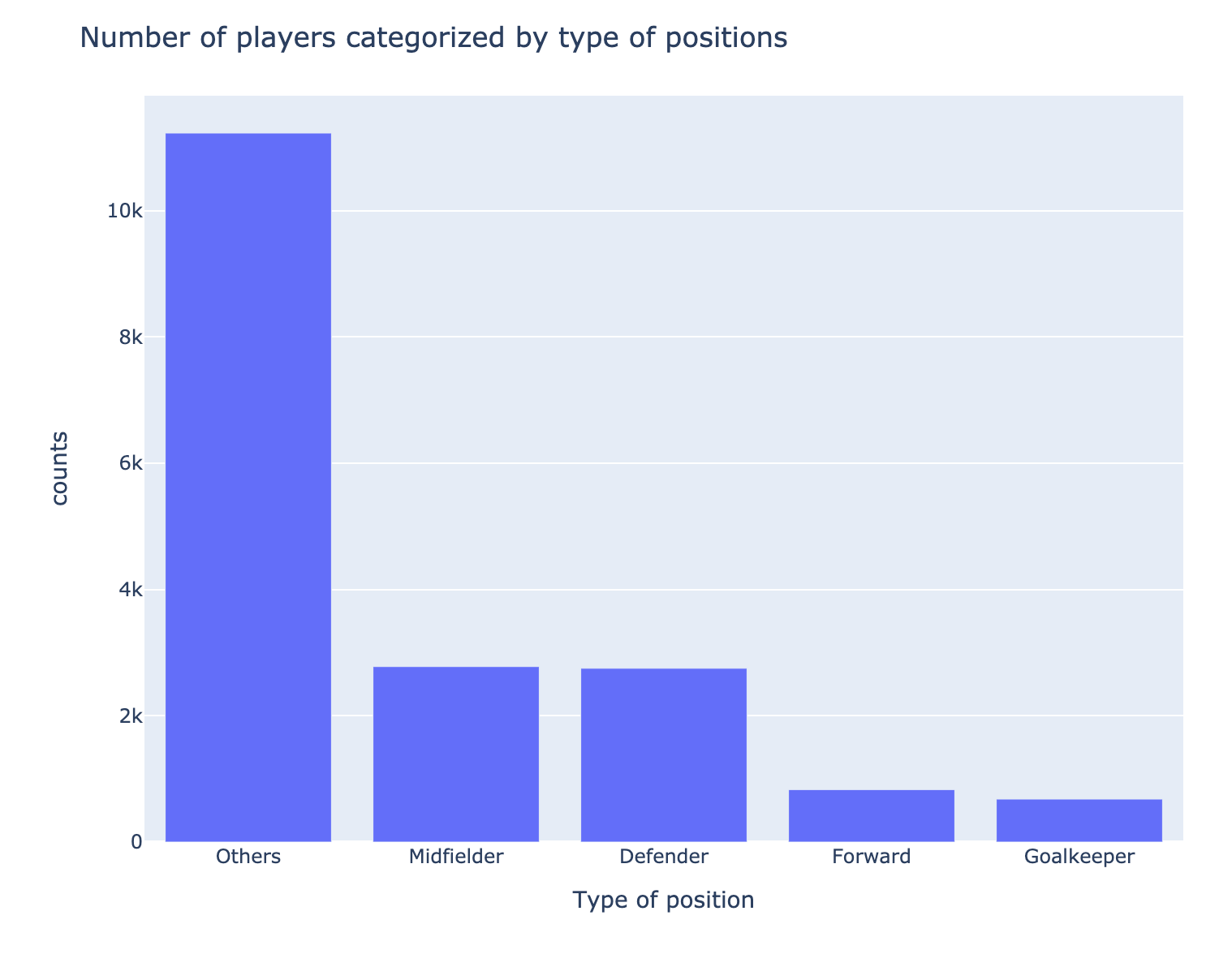


Figure 7: Number of players categorized by type of positions

Most of the players are either substitutes or reserves (Others). Most numbers of the players that play in the field are Midfielders, followed by Defenders.

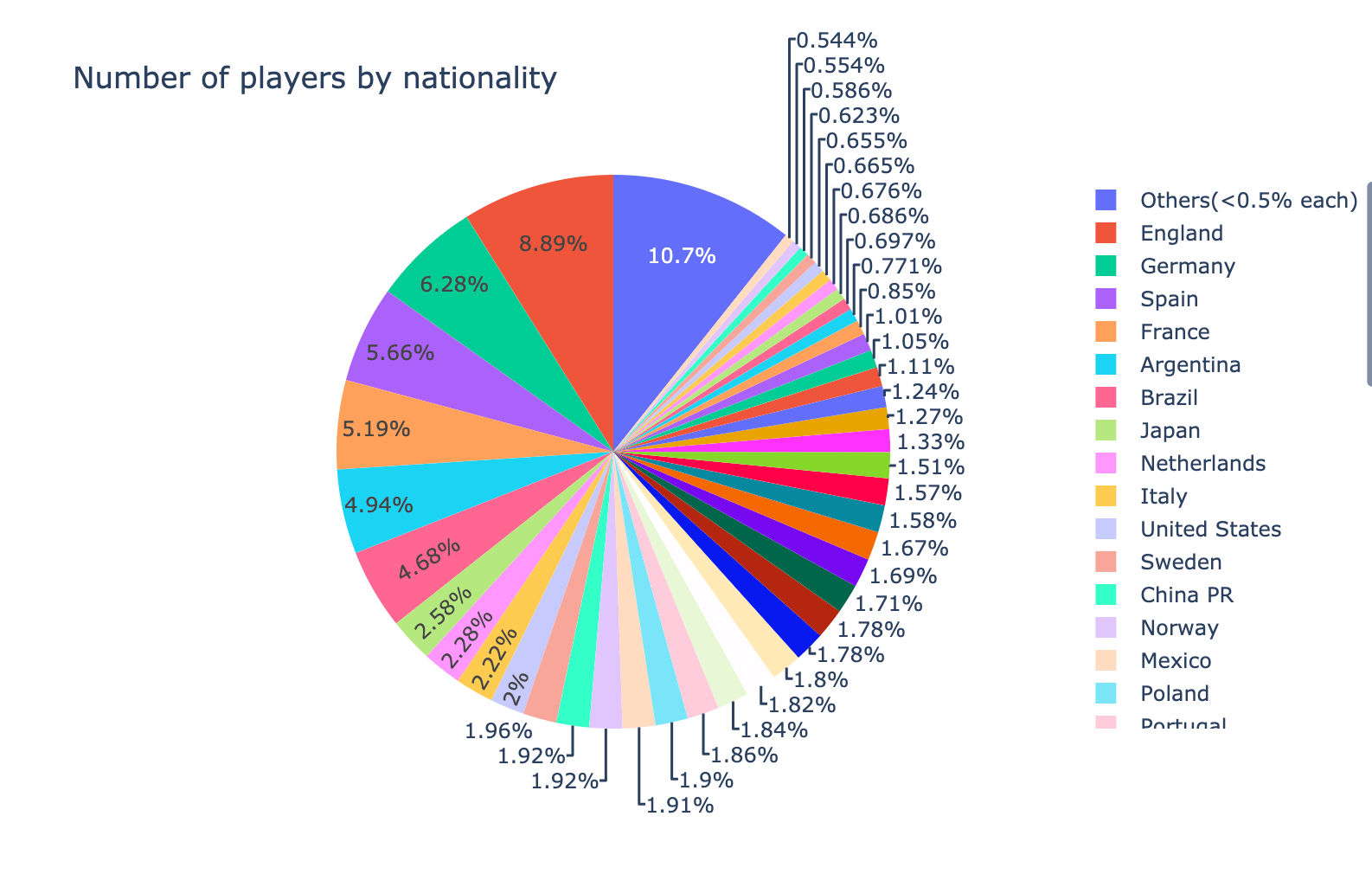


Figure 8: Number of players by nationality

The maximum number of players belongs to England (8.89%,) Germany (6.28%,) Spain (5.66%.) This observation is in tandem with the research conducted by V. Sudhandradevi's in paper [11] which stated that Europe is the biggest football contributor.



While 11 players from a team play in the field at one time (starring 11); It turns out a team has 27 players on average. This means a team has 16 substitutes/reserves

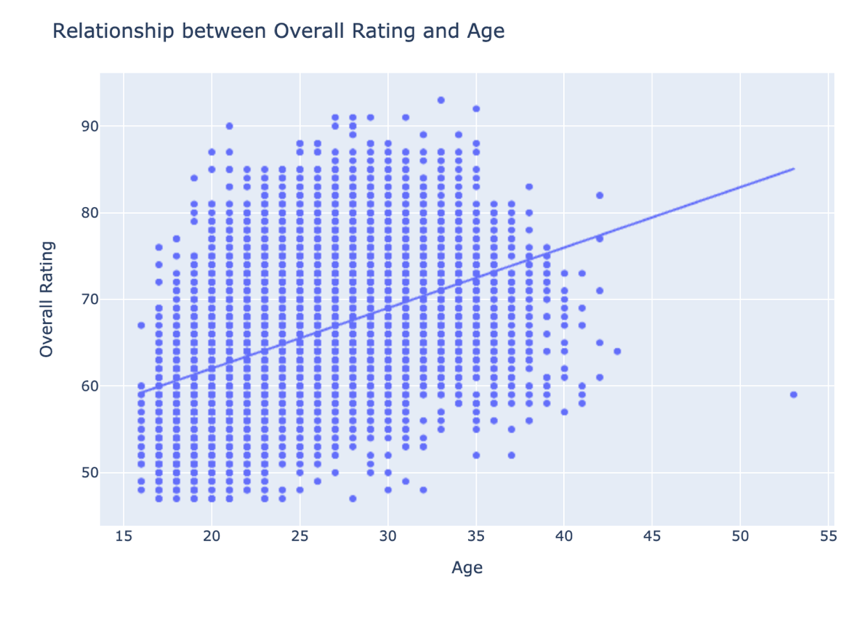


Figure 9: Relationship between Overall Rating and Age



Figure 10: Potential, Overall Rating and Age Trend Throughout Seasons

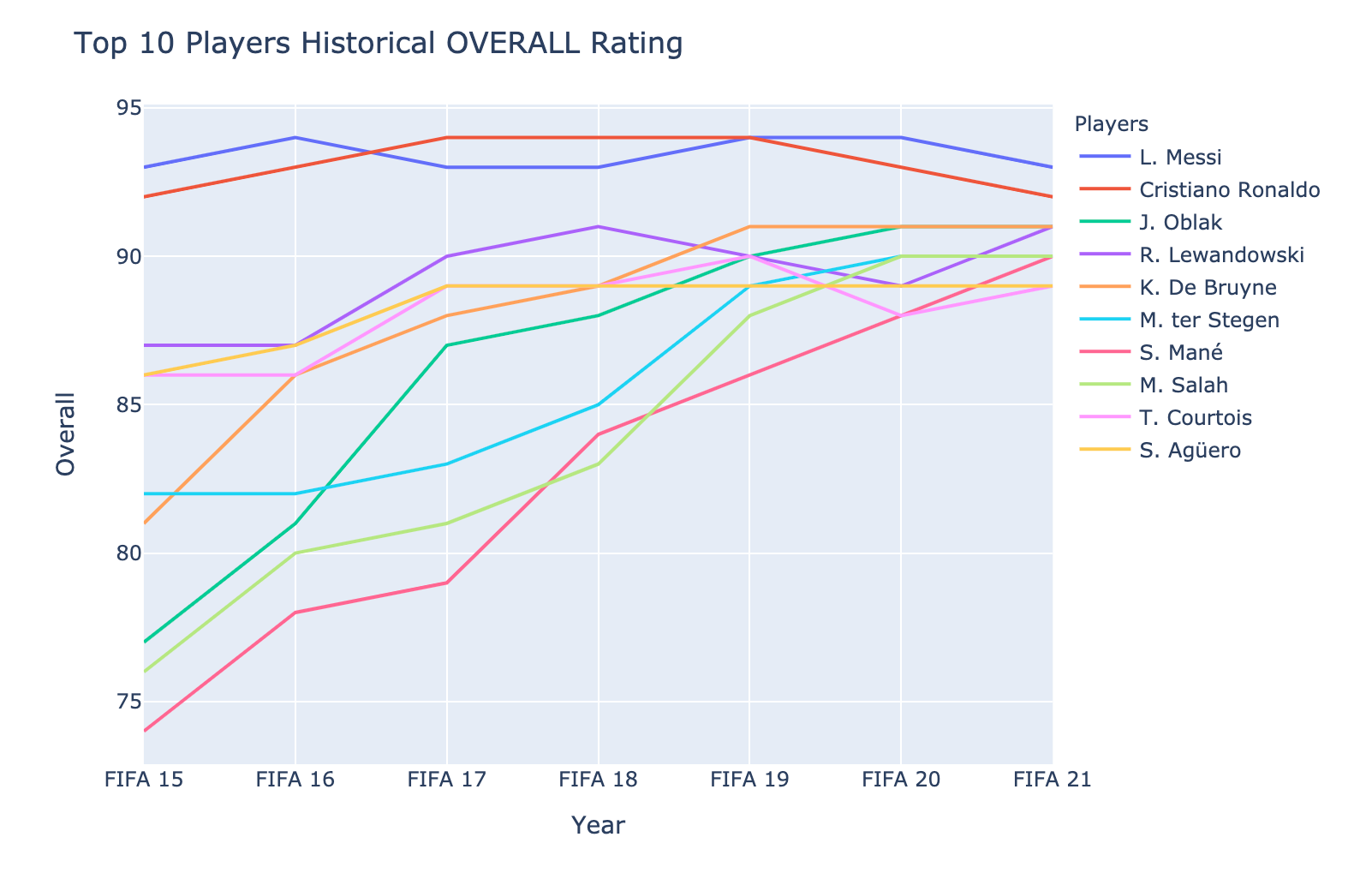


Figure 11: Top 10 Players Historial Overall Rating

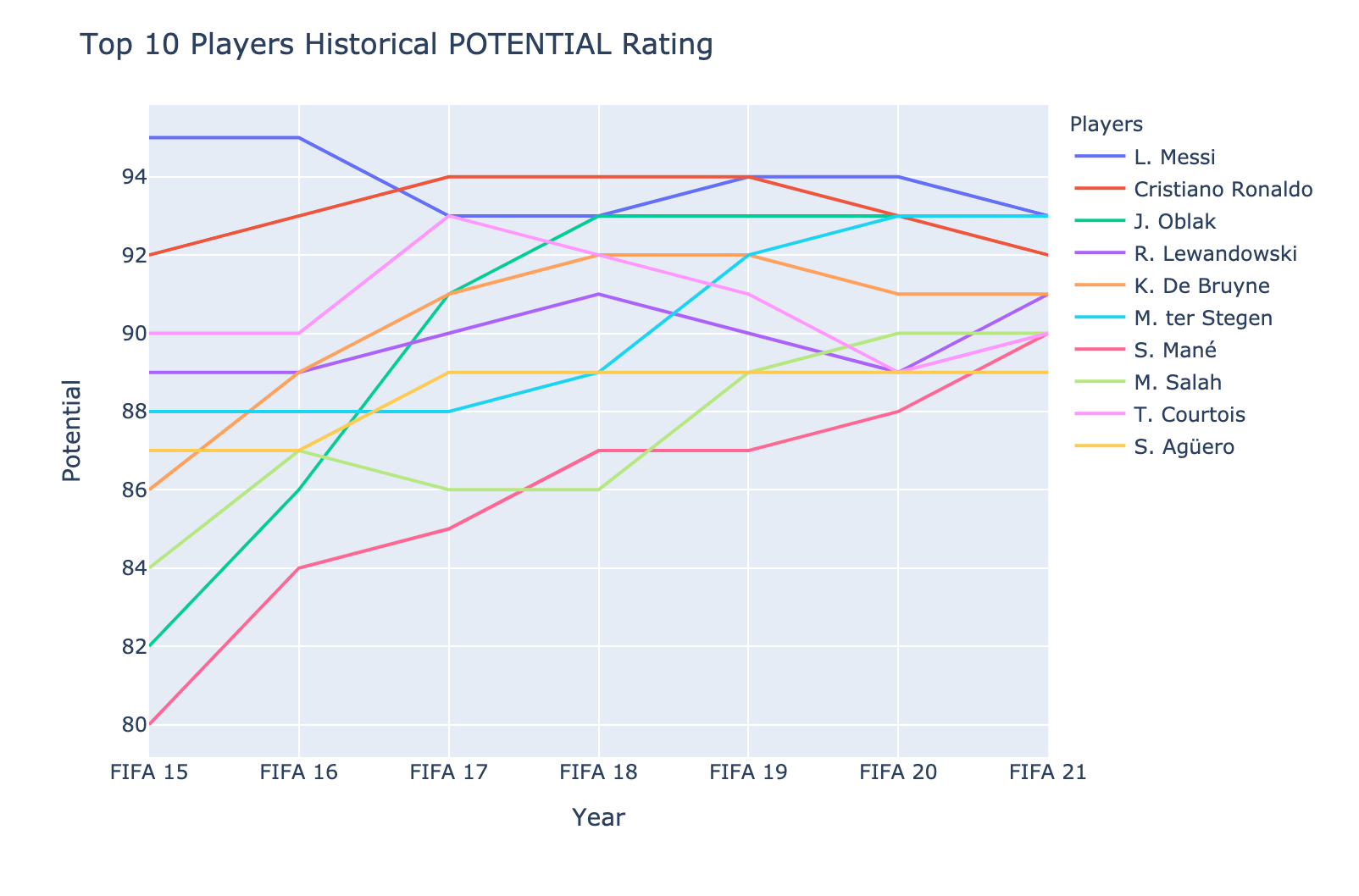


Figure 12: Top 10 Players Historical Potential

|  |  |
| --- | --- |
|  |  |
| Figure 13: Top 10 Players Season-wise |  |

|  |
| --- |
|  |
| Figure 14: FIFA 15 & 21 Dream Teams |

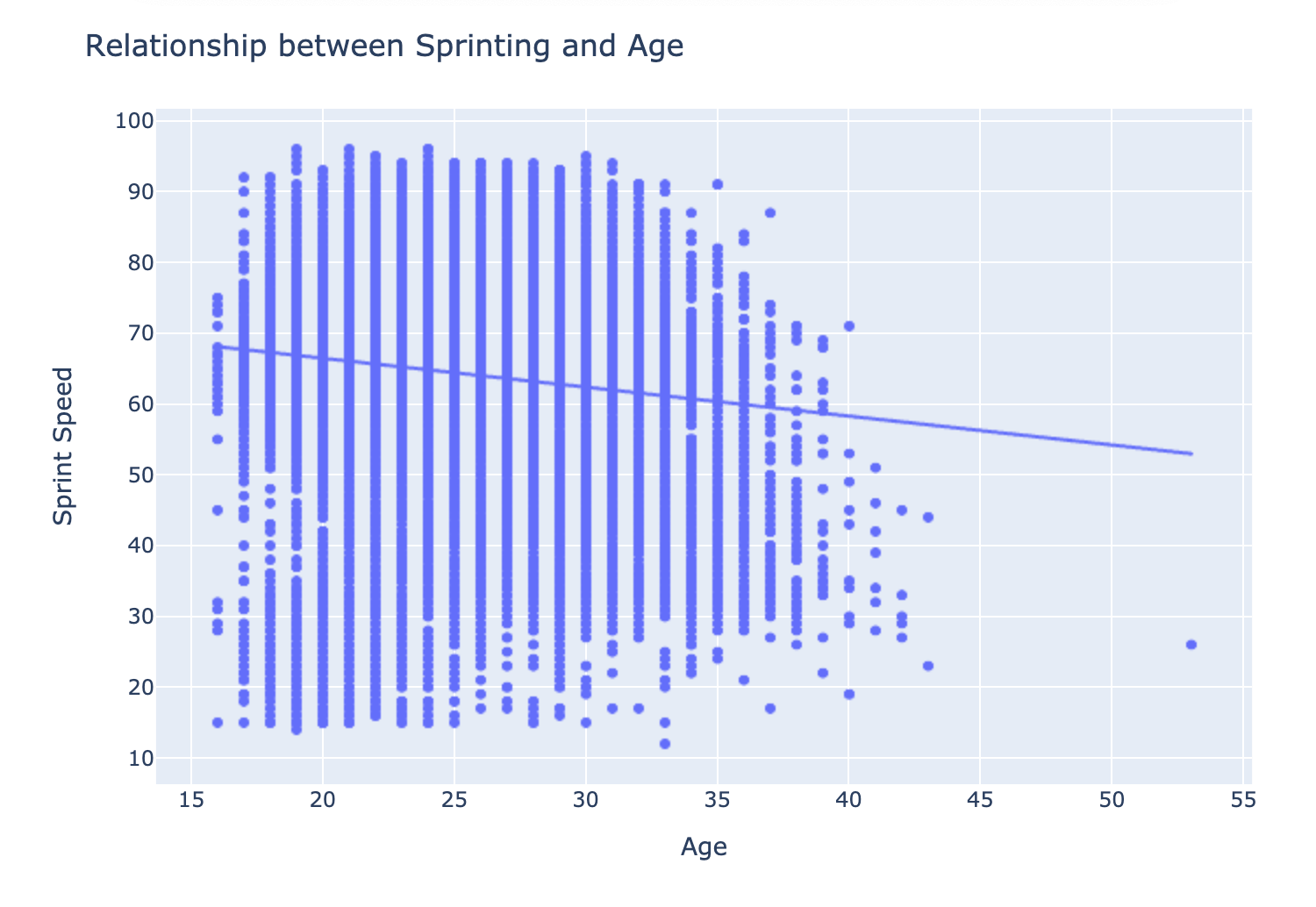


Figure 15: Relationship between Sprint and Age

From the graph above, we can see that as age increases, the sprinting speed decreases. This is true since it is scientifically proven that running speed decreases with age [17]



Figure 16: Top 3 Prominent Attributes for a Position

The result above tells us a lot about what it takes to play at a certain position in a match. We see what the most common skillset a certain team position requires are. Midfielders require balance, stamina, agility, short passing, acceleration, basically mostly movement, and power-based attributes. Defenders require strength, acceleration, aggression, and a good sprinting speed; again, mostly power and movement-based attributes. In contrast, attackers require to have high Shot power, dribbling, balance, acceleration, and agility; power, and then movement type attributes. We notice that all players require to have good movement.



Figure 17: Relationship between Potential and Overall Rating

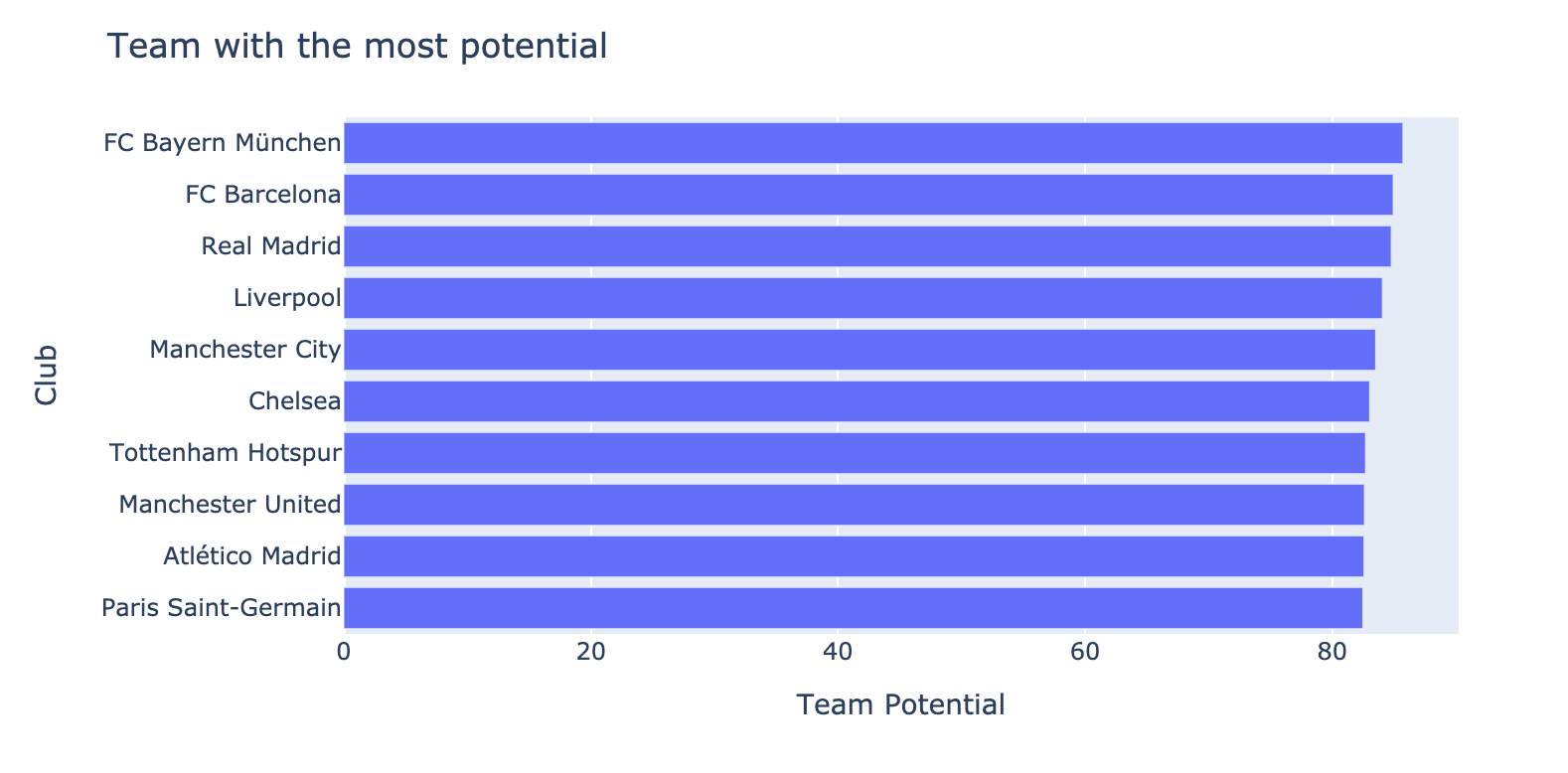


Figure 18: Teams with the Highest Potential

Figure 19: Relationship between Playing Positions and Weekly Wages

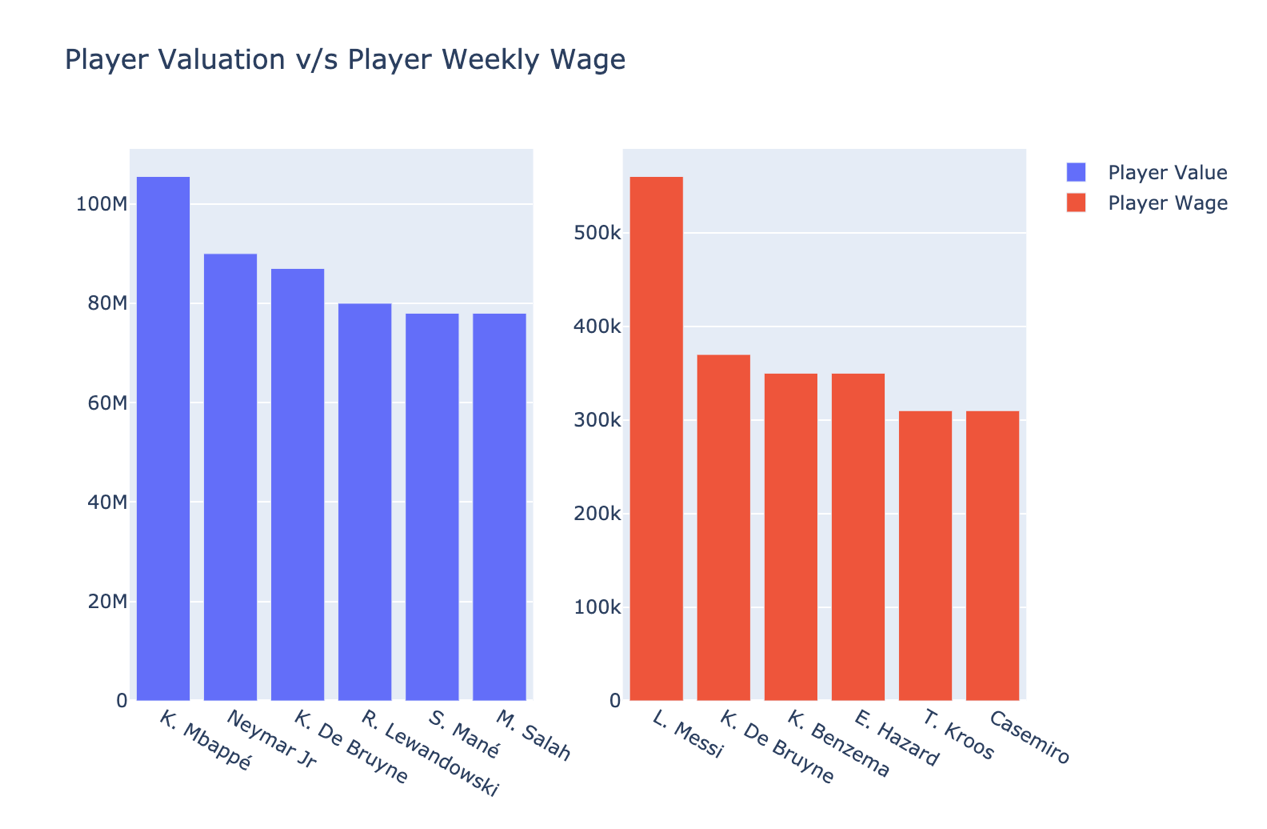


Figure 20: Player Valuation & Player Wage

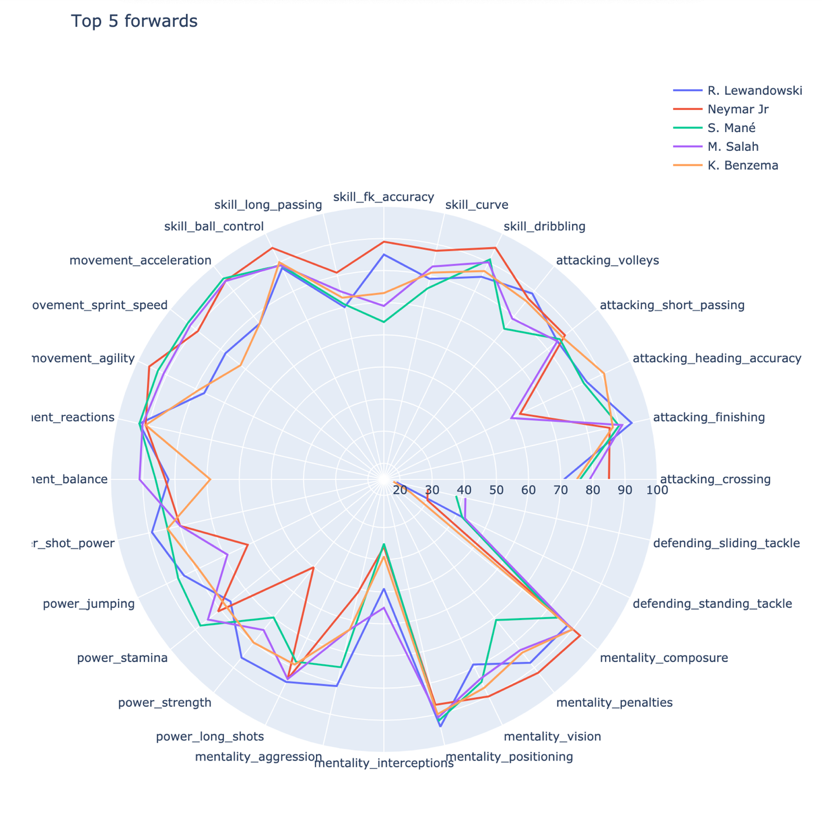


Figure 21: Radar Chart of Attributes of Top 5 Forwards

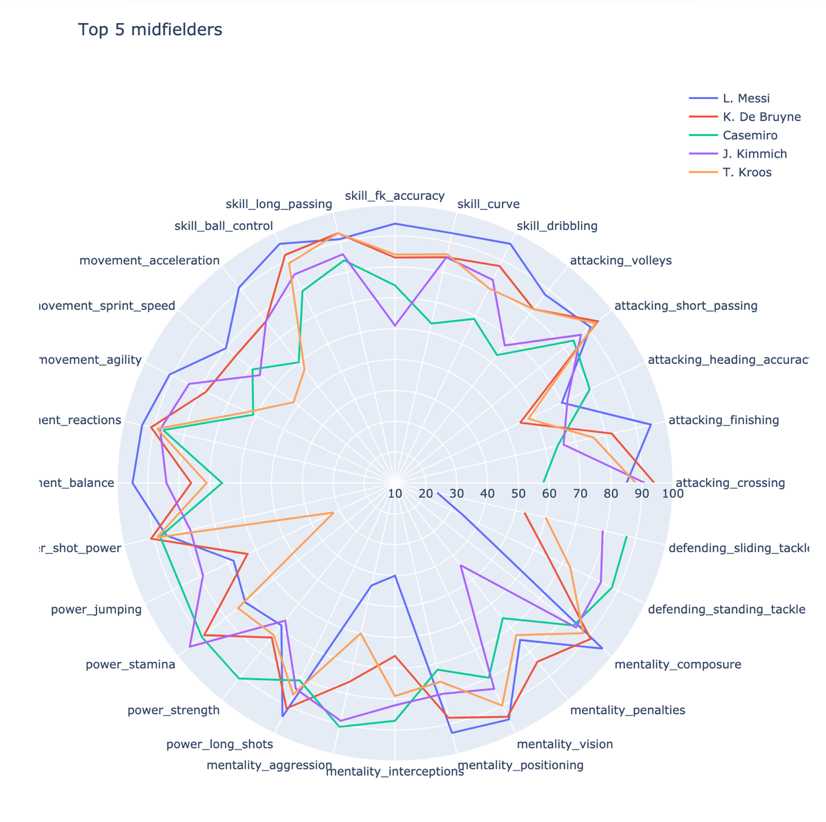


Figure 22: Radar Chart of Attributes of Top 5 Midfielders

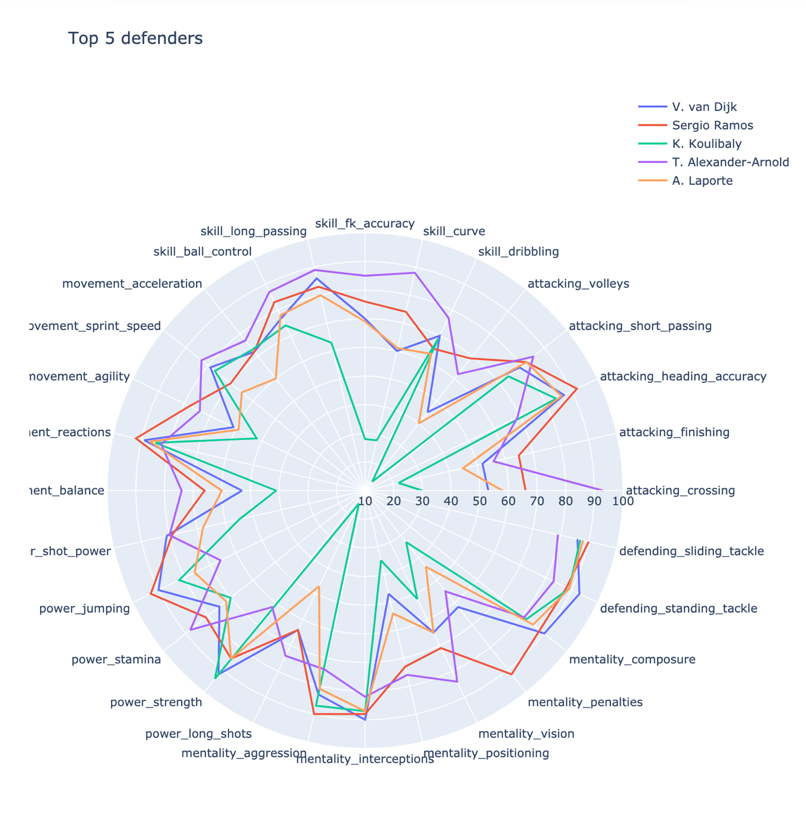


Figure 23: Radar Chart of Attributes of Top 5 Defenders

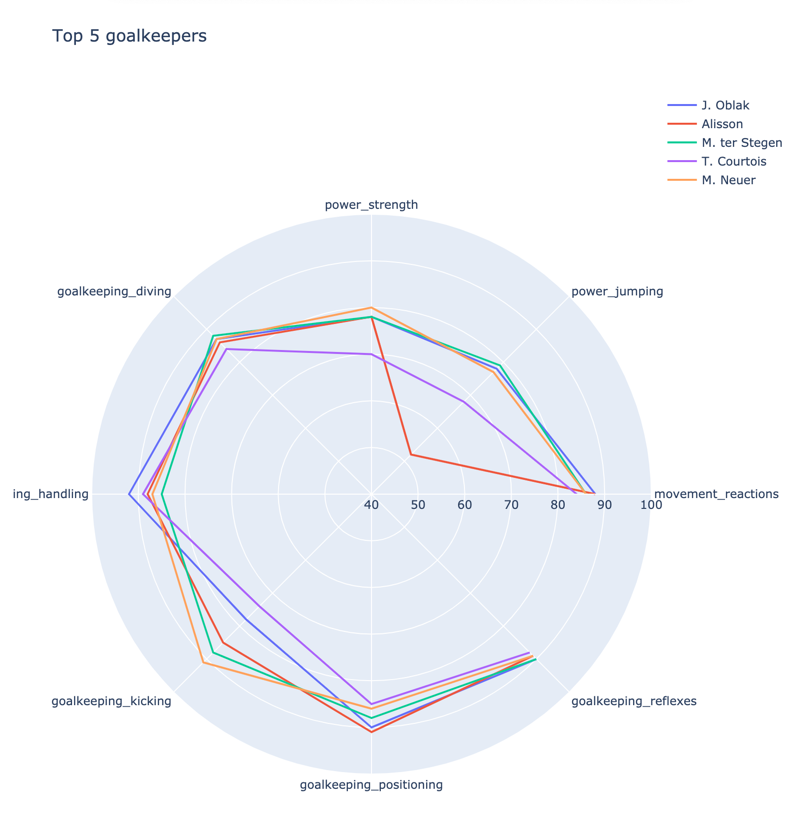


Figure 24: Radar Chart of Attributes of Top 5 Goalkeepers



Figure 25: All Attributes' Correlation With Each Other

Above, we plot a heatmap showing the correlation of different attributes with each other. At a glance, we can see that goalkeeper-related skills have a strong correlation with each other but not so good with other skills. This will help with classification.

1. *Player Classification*

|  |  |
| --- | --- |
| Figure 26: k-Nearest Neighbors Confusion Matrix | Figure 27: Decision Tree Confusion Matrix |
| Figure 28: Logistic Regression Confusion Matrix | Figure 29: Support Vector Machine Confusion Matrix |

|  |  |
| --- | --- |
| Figure 30: Decision Tree Classification Report | Figure 31: Logistic Regression Classification Report |
| Figure 32: k-Nearest Neighbors Classification Report | Figure 33: Support Vector Machine Classification Report |

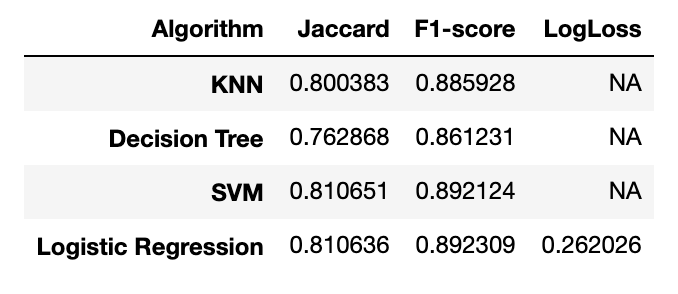


Figure 34: Jaccard, F1-score and LogLoss for all Algorithms

|  |
| --- |
| **Accuracy Scores:** |
| SVM: 0.8925837951966218  LR : 0.8923198733174981  KNN: 0.8857218263394036  DT : 0.8622327790973872 |

By far, the Support Vector Machine seems to be the algorithm that determined player position most accurately. When it comes to selecting which algorithm is the best performer; When the costs of False Positive are high. When high cost is associated with false-negative, natural is a goof model metric. If the goal is to find a balance between precision and recall, then F-score will be a better model metric.

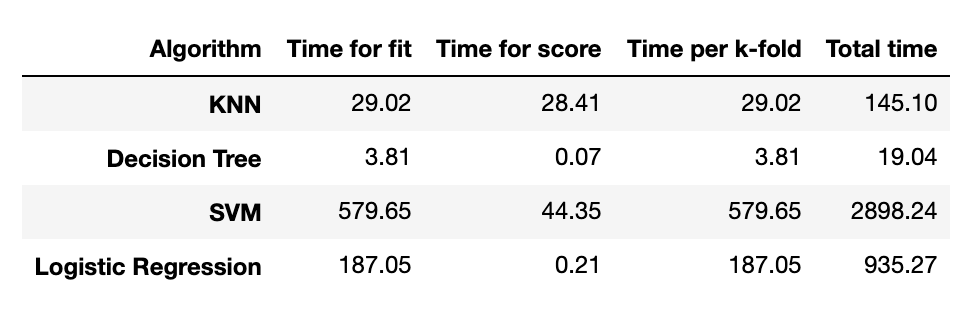


Figure 35: Time for fit, score, per k-fold and Total time for all Algorithms

We usually want more accuracy for our model. But it is crucial to consider the time the model takes for a small increment in accuracy. As we can see, SVM had 0.026% better accuracy than Logistic Regression, but it took SVM approximately 3.1 more times than Logistic Regression. It is best to consider time accuracy metrics. In the end, you have to decide if a slight increase in accuracy makes sense after considering the attributes of your data.

1. *Analysis of El Clásico*

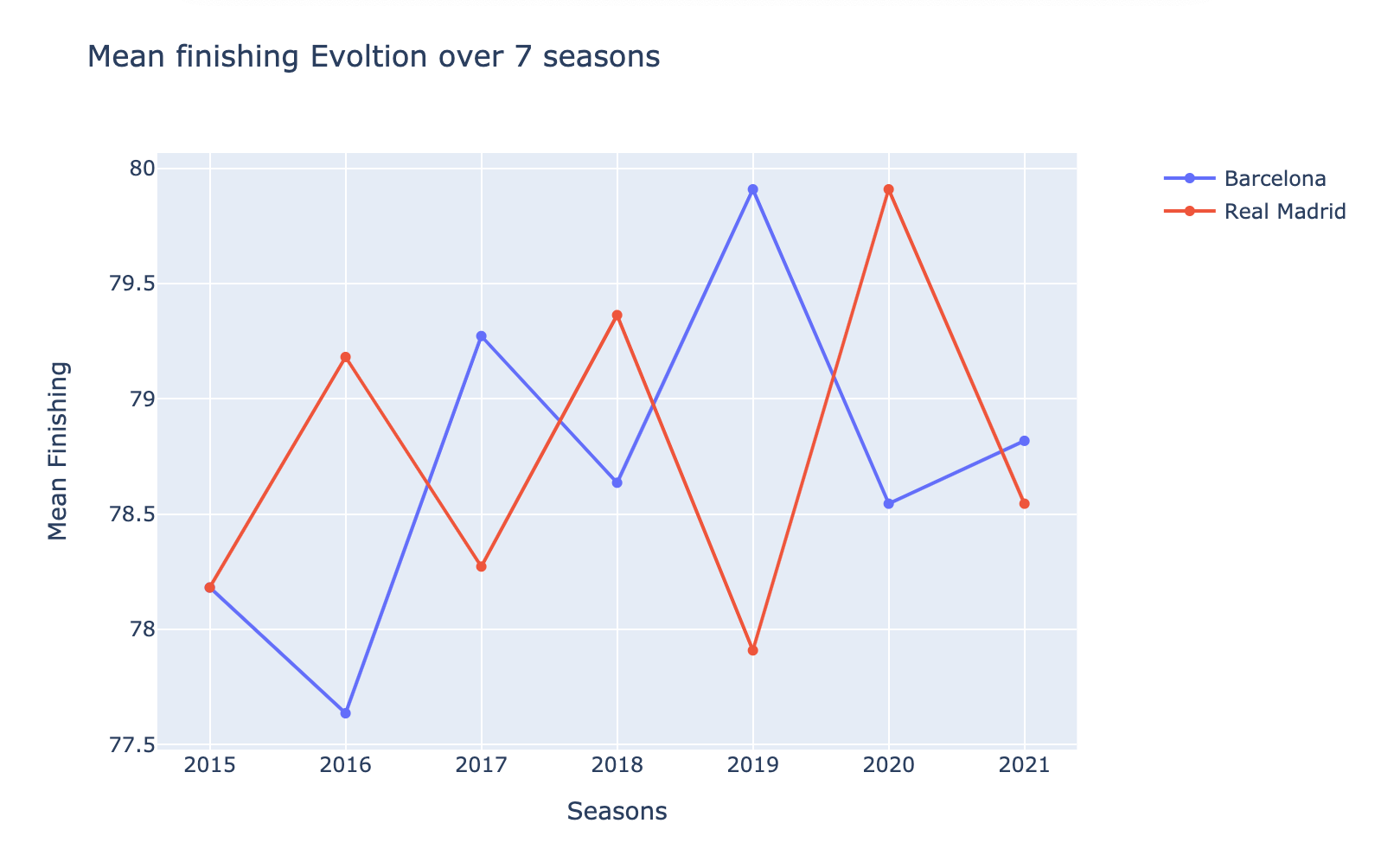


Figure 36: Mean Finishing Evolution of El Classico Rivals Over 7 Seasons

From above, we see that in the year 2015, Barcelona and Real Madrid were in equal standing. In 2016, Real Madrid had better finishing than Barcelona if we look at real matches; in 2016, Real Madrid won El Classico and the second 2016 match was a draw. In 2017, Barcelona had better finishing, and in reality, Barcelona won both matches. In 2019 Barcelona won the El Classico match in real life, and above, we can see that Barcelona had better finishing in 2019. For the year 2020, Real Madrid won both the matches in reality and above; we can see that the FIFA dataset also shows that Real Madrid had better finishing than Barcelona in the 2020 season.

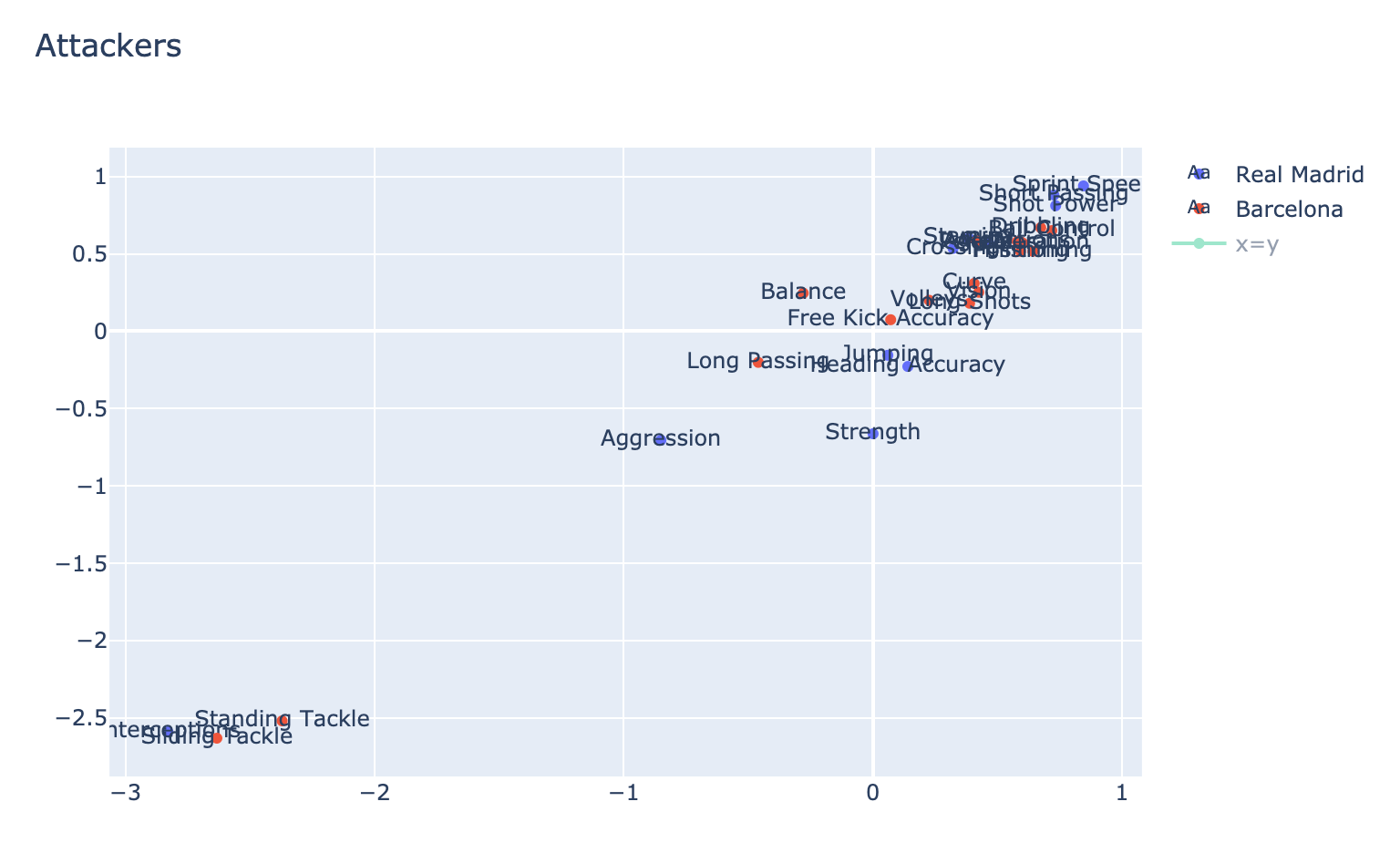
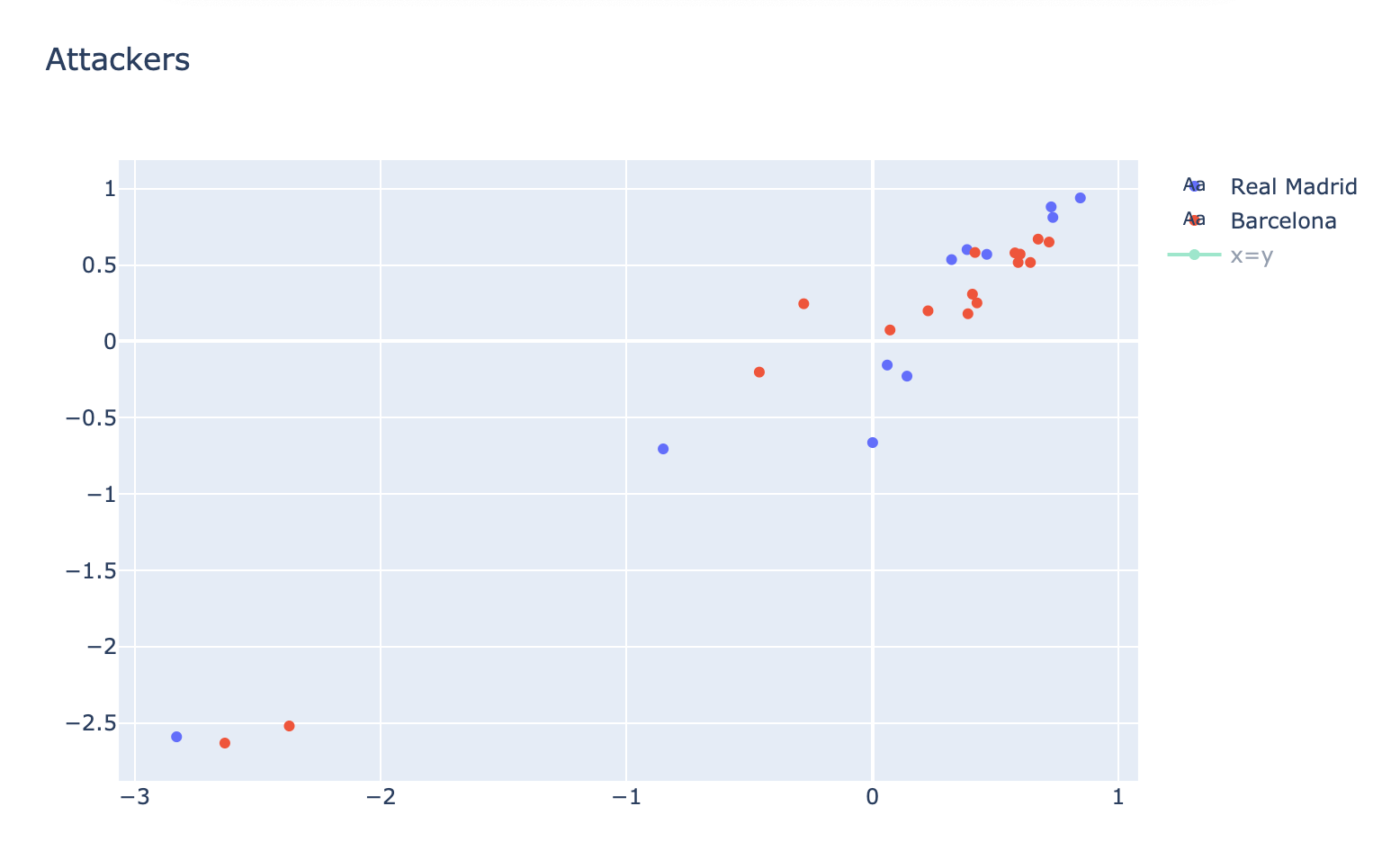


Figure 37: Real Madrid and Barcelona Attackers' Attributes Contrast

From above, we see that in terms of skills, Real Madrid is better than Barcelona in 11 different attributes while Barcelona is better in 16 different attributes when compared to Real Madrid. All of those are listed below: *Real Madrid*: Crossing, Heading Accuracy, Short Passing, Sprint Speed, Shot Power, Jumping, Stamina, Strength, Aggression, Interceptions, Penalties. *Barcelona*: Finishing, Volleys, Dribbling, Curve, Free Kick Accuracy, Long Passing, Ball Control, Acceleration, Agility, Reactions, Balance, Long Shots, Positioning, Vision, Standing Tackle, Sliding Tackle.



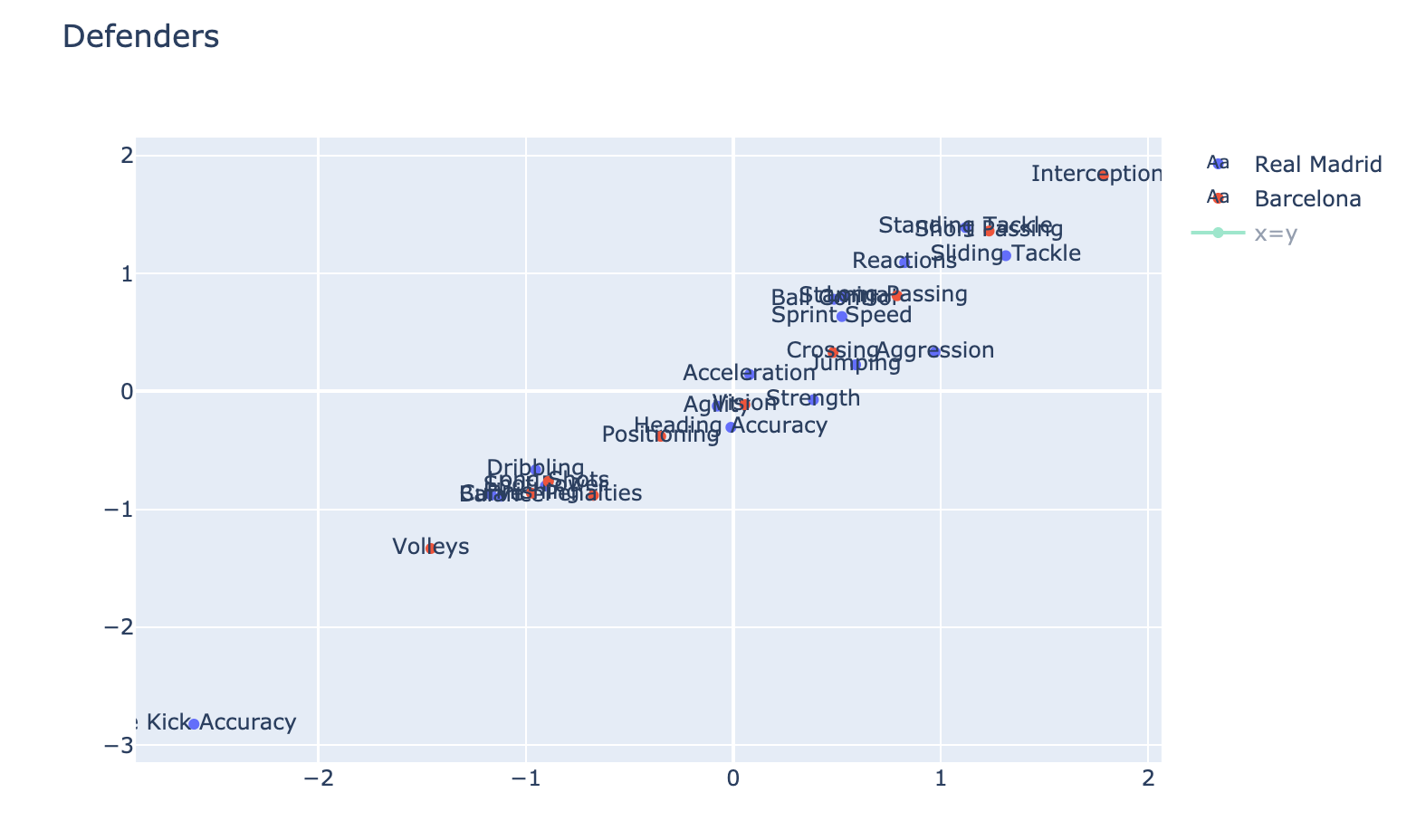
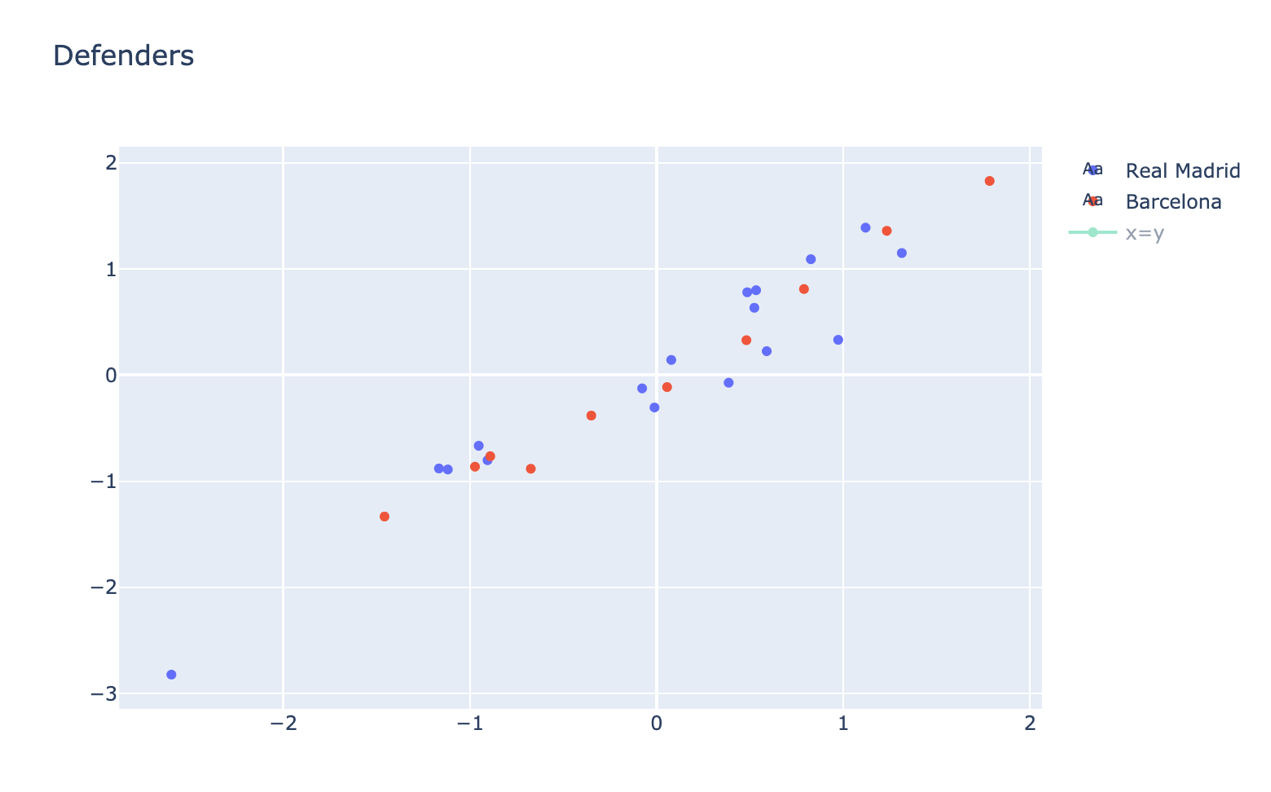


Figure 38: Real Madrid and Barcelona Defenders' Attributes Contrast

Real Madrid Defender’ are better than Barcelona in 17 different attributes while Barcelona is better in 10 different attributes when compared to Real Madrid. All of those are listed below: Real Madrid: Heading Accuracy, Dribbling, Curve, Free Kick Accuracy, Ball Control, Acceleration, Sprint Speed, Agility, Reactions, Balance, Shot Power, Jumping, Stamina, Strength, Aggression, Standing Tackle, Sliding tackle. Barcelona: Crossing, Finishing, Short Passing, Volleys, Long Passing, Long Shots, Interceptions, Positioning, Vision, Penalties



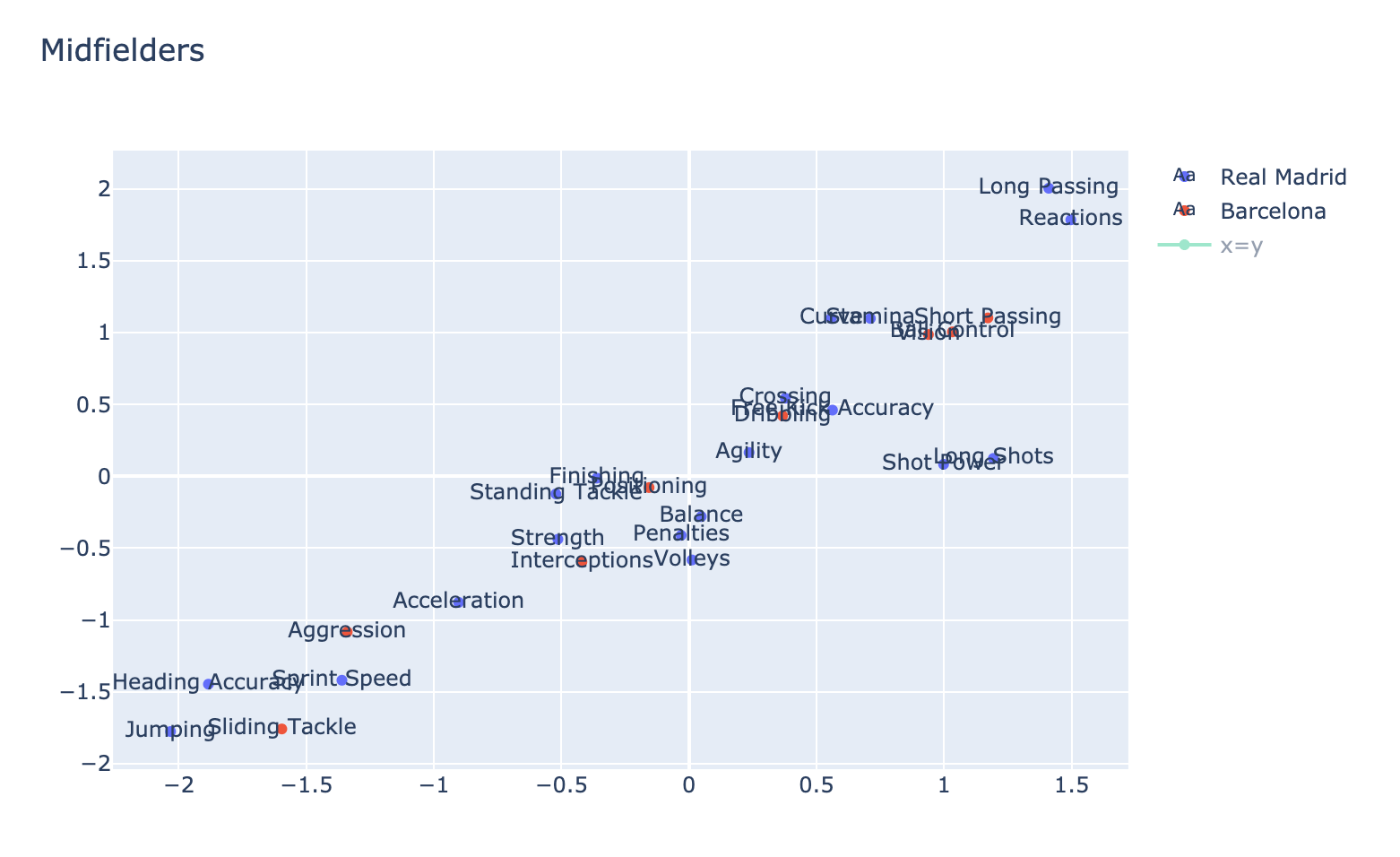
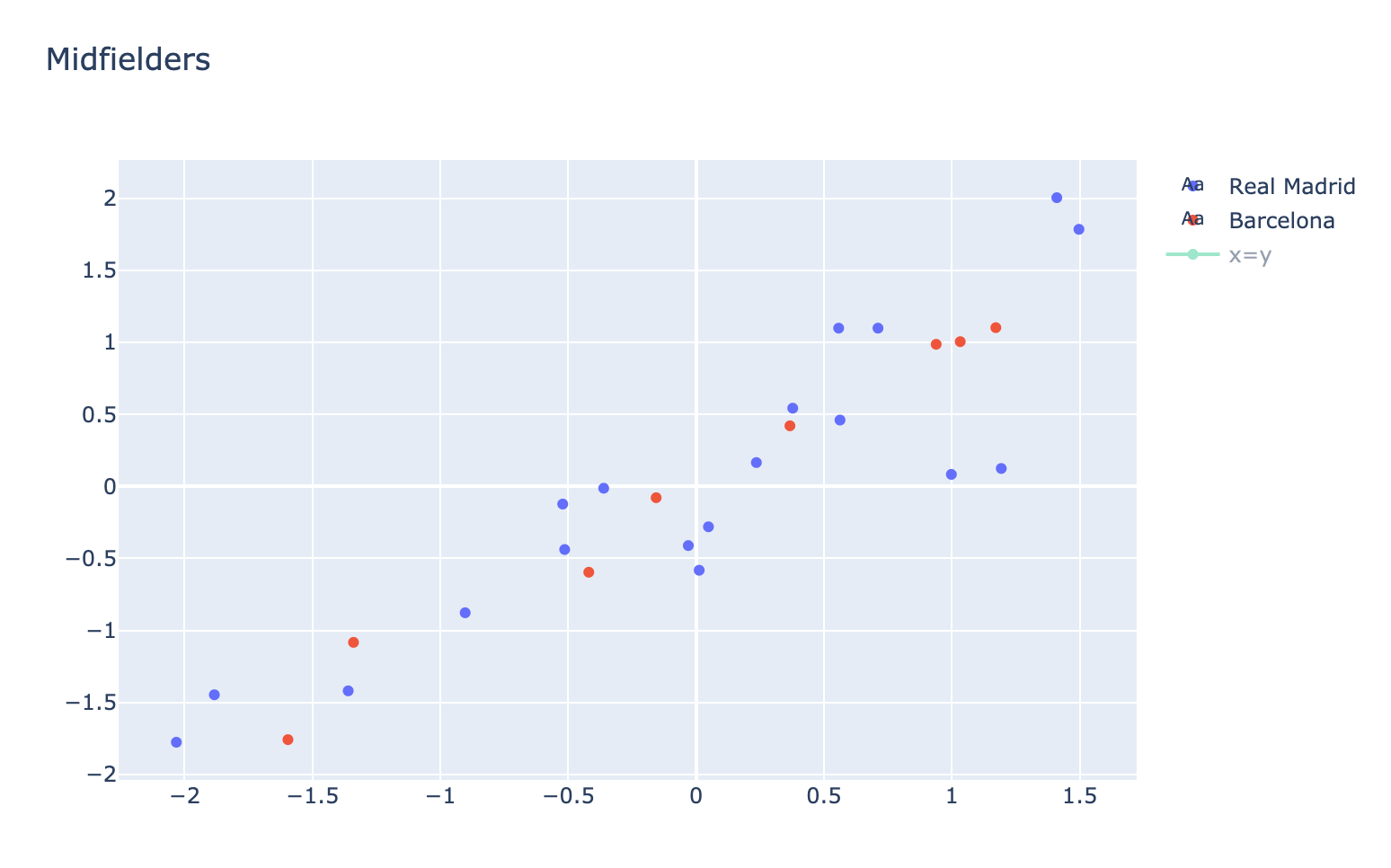


Figure 4: Real Madrid and Barcelona Midfielders' Attributes Contrast

From above, we see that in terms of skills, Real Madrid is better than Barcelona in 19 different attributes while Barcelona is better in 8 different attributes when compared to Real Madrid. All of those are listed below: Real Madrid: Crossing, Finishing, Heading Accuracy, Volleys, Curve, Free Kick Accuracy, Long Passing, Acceleration, Sprint Speed, Agility, Reactions, Balance, Shot Power, Jumping, Stamina, Strength, Long Shots, Penalties, Standing Tackle Barcelona: Short Passing, Dribbling, Ball Control, Aggression, Interceptions, Positioning, Vision, Sliding Tackle



If we consider all playing positions on average, it seems that Real Madrid is better in 15 different attributes, while Barcelona is better in 11 different attributes. We believe above we've presented very useful and actionable advice to both coaches and players on areas where they should focus on. If each team can become better at the attributes the other team is better in, they can significantly increase their chances of winning.

1. CONCLUSION & FUTURE WORK

From our study, we manage to understand football trends and events better. We did various Exploratory Data Analyses and used machine learning classification algorithms like Support Vector Machine, Decision Tree, K-nearest Neighbors, and Logistic regression to classify and recommend players. We also analyzed a widely discussed event (El Classico) and depicted how our dataset can provide precious insights. This paper backs the idea that the players' performance indicators heavily influence a players' performance. From this analysis, non-visual experts, amateurs football players who want to know what to expect and what attributes they should train for, gamers who want to make the best possible team in FIFA, scouts who want to look for players with the most potential can all indeed benefit. Hence, we can conclude that our research helps us understand football and its structure a lot better, and we also showed the potential the FIFA dataset holds and a plethora of analysis it can be used for. With our work, we have barely scratched the surface of endless possibilities. There is a treasure of valuable insights that can still be gained from the FIFA dataset. In the future, we would prefer to use more data from seasons older than 2015 so that we can have more affluent and more conclusive results. We would also like to use real-life world cup data to make prediction tools to predict real-life match results from the model made from the FIFA dataset and real-life World Cup data. We would also like to draw more attention to differences between national and club teams, use World Cup data to make prediction tools, and launch the research on a web app to reach a wider audience. The analysis of football data is still at the early stage, and it can only go up from here on out. With time, we can safely say that this field will gain more and more traction, and we believe that our paper can act as fuel and lead to other relevant research.

1. REFERENCES

|  |  |
| --- | --- |
| [1] | "FIFA Survey: approximately 250 million footballers worldwide," 2001. [Online]. Available: https://www.fifa.com/who-we-are/news/fifa-survey-approximately-250-million-footballers-worldwide-88048. |
| [2] | "How analytics is changing the world of sports," [Online]. Available: https://www.analyticsinsight.net/how-analytics-is-changing-the-world-of-sports/. |
| [3] | "Champions League final tops Super Bowl for TV market," [Online]. Available: http://news.bbc.co.uk/sport1/hi/football/europe/8490351.stm. |
| [4] | A. S. Markovits and A. I. Green, "FIFA , the video game: a major vehicle for soccer’s popularization in the United States," https://www.researchgate.net/publication/299403448\_FIFA\_the\_video\_game\_a\_major\_vehicle\_for\_soccer's\_popularization\_in\_the\_United\_States, 2016. |
| [5] | "FIFA 12 sells 3.2 million in one week," [Online]. Available: https://www.gamespot.com/articles/fifa-12-sells-32-million-in-one-week-ea/1100-6338168. |
| [6] | D. Owen, "Sign of the times: FIFA looks to have generated more income from video gaming than football in 2020," [Online]. Available: https://www.insidethegames.biz/articles/1105726/fifa-more-revenue-from-video-gaming. |
| [7] | C. Soto-Valero, "A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system," *RICYDE. Revista Internacional de Ciencias del Deporte,* vol. XIII, no. 49, pp. 244-259, 2017. |
| [8] | K.-Y. H. Wen-Lung Chang, "A neural network method for prediction of 2006 world cup football game," *The 2010 International Joint Conference on Neural Networks (IJCNN),* pp. 1-8, 2010. |
| [9] | A. R. J. Hucaljuk, "Predicting football scores using machine learning techniques," *MIPRO, 2011 Proceedings of the 34th International Convention,* pp. 1623-1627, 2011. |
| [10] | P. O. V. d. M. Leonardo Cotta, "Using FIFA Soccer video game data for soccer analytics," 2016. [Online]. Available: https://www.semanticscholar.org/paper/Using-FIFA-Soccer-video-game-data-for-soccer-Cotta/9ca3469f4de59c160289d3925ebfbf4b9ef4c0e7. |
| [11] | V. B. P. Sudhandradevi, "Sports Analysis of FIFA Football World Cup Tournament using Logistic Regression," *International Journal of Computational Intelligence and Informatics,* vol. 8, no. 2, September 2018. |
| [12] | C. L. G. S. Andreas Groll, "Prediction of the FIFA World Cup 2018 – A random forest approach with an emphasis on estimated team ability parameters," Big Data Analytics in sports (BDSports), 2018. [Online]. Available: https://www.researchgate.net/publication/325683270\_Prediction\_of\_the\_FIFA\_World\_Cup\_2018\_-\_A\_random\_forest\_approach\_with\_an\_emphasis\_on\_estimated\_team\_ability\_parameters. |
| [13] | O. Wallach, "The history of the gaming industry in one chart," [Online]. Available: https://www.weforum.org/agenda/2020/11/gaming-games-consels-xbox-play-station-fun. |
| [14] | "SOFIFA," [Online]. Available: https://sofifa.com. |
| [15] | D. Dubas-Fisher, "FIFA ratings vs real life: How does the video game measure up to actual Premier League stats?," 2014. [Online]. Available: https://www.mirror.co.uk/sport/football/news/fifa-ratings-vs-real-life-4027925. |
| [16] | S. Leone. [Online]. Available: https://www.linkedin.com/in/stefano-leone-ie/. |
| [17] | "Aging Doesn’t Slow Your Running Down That Much," [Online]. Available: https://www.runnersworld.com/health-injuries/a23845530/running-speed-decline-aging. |
| [18] | "3 most popular formations in football today," [Online]. Available: https://www.sportskeeda.com/football/3-most-popular-football-formations-today. |