Analysis on Soccer Player Attributes

By: Sau Kha
April 2018

Executive Summary

The purpose of this project is to study player attributes of a European soccer database from Kaggle.com. The database consists of eight tables, sqlite_sequence, Player_Attributes, Player, Match, League, Country, Team and Team Attributes. The database was populated with more than 25,0000 matches from 2008 to 2016 seasons, 10,000 players, and 11 European countries with their lead championship. Players' and teams' attributes are sourced from EA Sports' FIFA video game series (https://sofifa.com/).

Research Questions:
The research questions for this project are:

1. What attributes set the players apart?
2. Which player attribute contributes most to a player's overall rating?

Data Cleaning:
Data analysis was conducted in Python using various libraries/modules such as, sqlite3, pandas, principal component analysis from scikit-learn, matplotlib, and seaborn. Data from the Player and the Player_Attributes tables were queried with an inner join query on the player identification number from the FIFA application programming interface (API). Queried data comprises of 183929 rows and 50 columns. Data was then cleaned so that records with missing or null data in any column was removed. Duplicated data, if any, was removed. Final dataset consists of 10898 rows and 50 columns. 38 columns of attributes with numeric data were identified for principal component and further analysis.

Data Exploration and Analysis:
Principal component analysis was conducted. Results shows that the first three principal components out of 38 dimensions, PC1, PC2 and PC2, explain 43.9%, 15.7% and 8.9%, respectively, which only explained a total of 68.5% of the variance in the dataset. Loading scores of all player attributes in each of the three principal components show that no one single player attribute contributes significantly more than any other attributes to these three principal components. Various charts indicate that there are two distinct (large and small) subgroups of players in the dataset. Created scatter plots, distribution plots, joinplots and correlation coefficient analysis explained the clusters and substantiated the answers to the research questions.

Conclusions:

1. The following attributes set players apart into subgroups: gk_diving, gk_reflexes, gk_handling, gk_positioning, and ball control. The first four attributes pertain to goalkeeper position while ball control relate to more to non_goalkeepers. Goalkeepers score relatively high in the goalkeeping attributes, but low in ball control, while the rest of players score in the opposite way. Thus, players were set into one small subgroup of goalkeepers and one large group of the rest of the players. Gk_kicking, a goalkeeping attribute further divided the larger subgroup into two: some players in the back field had high scores in gk_kicking that had a moderate positive linear correlation to their ball control scores.
2. Considering all players, reactions attribute had a strong positive linear correlation with overall rating, thus contributes most to the rating. The higher the reactions score, the higher the overall rating of the player.
3. However, at the subgroup level, for the goalkeeper subgroup, gk_diving, gk_handling, gk_positioning and gk_reflexes attributes had very strong positive linear correlation with overall rating and contribute most to the players' rating. For the rest of the players, reactions attribute had a strong correlation and contributed most to their overall rating.

Appendix A shows information and sources of data. Appendix B presents all the Python scripts used for data queries, cleaning, exploration and analysis and their outputs.

Table of Contents

List of Tables

List of Figures

Introduction

This purpose of this project is to study player attributes of a European soccer database from Kaggle.com. The database consists of eight tables, sqlite_sequence, Player_Attributes, Player, Match, League, Country, Team and Team Attributes. The database was populated with more than 25,0000 matches from 2008 to 2016 seasons, 10,000 players, and 11 European countries with their lead championship. Players' and teams' attributes are sourced from EA Sports' FIFA video game series (https://sofifa.com/).

To be attain great team performance, individual soccer player typically undertakes intensive drills to master specific skills for his dedicated position(s). In addition to basic skills, specialized training may focus in attack, defensive, mentality, movement, power or goalkeeping attributes. When highly skillful soccer players work together as a team in a strategic team formation, the likelihood to win a game may increase significantly.

In the dataset, each soccer player was given an overall rating and potential score, along with scores for the following numeric attributes: crossing, finishing, heading accuracy, short passing, volleys, dribbling, curve, free kick accuracy, long passing, ball control, acceleration, sprint speed, agility, reactions, balance, shot power, jumping, stamina, strength, long shots, aggression, interceptions, positioning, vision, penalties, marking, standing tackle, sliding tackle, gk_diving, gk_handling, gk_kicking, gk_positioning and gk_reflexes.

Research questions

The research questions for this project are:

1. What attributes set players apart?
2. Which player's attribute contributes most to a player's overall rating?

Intended Audience

The intended audience for this project are: soccer fans and FIFA video game players.

Reference

1. Database source: https://www.kaggle.com/hugomathien/soccer/data
2. Original data sources:
    a. http://football-data.mx-api.enetscores.com/ : scores, lineup, team formation and events
    b. http://sofifa.com/ : players and teams attributes from EA Sports FIFA games. *FIFA series and all FIFA assets property of EA Sports*
3. Merriam-Webster – Visual Dictionary Online: "player positions" http://www.visualdictionaryonline.com/sports-games/ball-sports/soccer/player-positions.php
4. Video from Bing.com: "Soccer Goalie Diving Drills": https://www.bing.com/videos/search?q=soccer+goalie+diving+drills&view=detail&mid=95069C368876787EAEBE95069C368876787EAEBE&FORM=VIRE
5. Wikipedia.org: "Association Football Positions" https://en.wikipedia.org/wiki/Association_football_positions
6. YouTube video series: "StatQuest: Principal Component Analysis (PCA), Step-by-Step" https://www.youtube.com/playlist?list=PLblh5JKOoLUIcdlgu78MnlATeyx4cEVeR
7. YouTube video: "Top 10 BEST Longest Goals by Goalkeepers" https://www.youtube.com/watch?v=QUp6OyrHoMc

8. footy4Kids: “Goalkeeper coaching for U12 and up” http://www.footy4kids.co.uk/soccer-drills/goalkeeping/goalkeeper-coaching-for-u12-and-up/

Methodology and Analysis Approach

The methodology for this study basically is first to explore, clean up, analyze, visualize the data and then draw conclusions and attempt to find answers to the research questions. This study was conducted in Python using various libraries/modules such as, sqlite3, pandas, principal component analysis from scikit-learn, matplotlib, and seaborn. The steps below were taken to reach the findings and conclusions:

1. Database was downloaded from https://www.kaggle.com/hugomathien/soccer/data
2. Table structures and field datatypes were explored. Tables from the downloaded sqlite database were exported and previewed.
3. Using IO tools from pandas library, read_sql and to_csv, the Player and the Player_Attributes tables were queried with an inner join SQL on the player identification number from the FIFA application programming interface (API). Queried data was saved to a csv file.
4. Queried data comprises of 183929 rows and 50 columns. Data was then cleaned so that records with missing or null data in any column was removed. Duplicated data, if any, was removed. Final dataset consists of 10898 rows and 50 columns.
   Lessons Learned:
   a. I was surprise to find that a huge number of rows (players) were dropped till I read from the data source webpage and noted that the administrator of the database was not able to source quite some players’ attributes from FIFA and have those fields blank.
   b. Data does not lie.
5. Identified thirty-eight (38) columns of players’ attributes in numeric format for principal component analysis (PCA). Ran PCA to identify players’ attributes that explained the most variance, to help reduce dimensions before further analysis. .
   Lessons Learned:
   a. Loading scores from PCA did not support dimension reduction. At this time, a decision had to be made to how to proceed. Additional data exploration was needed to decide where to zoom in for further analysis. Since it did not go as what I expected, I turned to other tools to use and approach to take.
   b. I created a series of charts using a programming loop, which allowed exploration systematically on each attribute.
   c. Additionally, adding color in the scatter plots for each point added a third dimension and presented a third attribute in the charts. The added information was easier on human eyes when compared to adding a z-axis for the third attribute.
   d. Joint plots provided distribution of player variables that were presented in a scatter plot. These techniques provided valuable information as how to proceed with data analysis to get to the bottom to find answers to the research questions.
   e. Scatter plot matrix contains all the pairwise scatter plots of the variables on a single page in a matrix format. This was very useful to help roughly determine whether there was a linear correlation between multiple variables.
6. Conducted further data exploration and visualization with scatter plots matrix, scatter plots, distribution plots and joint plots.
7. Identified clusters with further analysis.

8. Calculated correlation coefficient matrix for all attributes. Located attributes with strong correlation coefficient with overall rating. This was repeated for the goalkeeper subgroup and for the rest of the players. Correlated attributes were identified. Created scatter plots for the correlated attributes for visual confirmation.
9. Draw conclusions based on findings.

Findings

Principal Component Analysis

Figure 1 presents the Scree Plot of all the principal components from the 38 numeric attributes of the final, cleanup dataset of soccer players. PCA results shows that the first three principal components, PC1, PC2 and P32, out of 38 dimensions explain 43.9%, 15.7% and 8.9% of variance, respectively. This is a total of 68.5% of all variance.

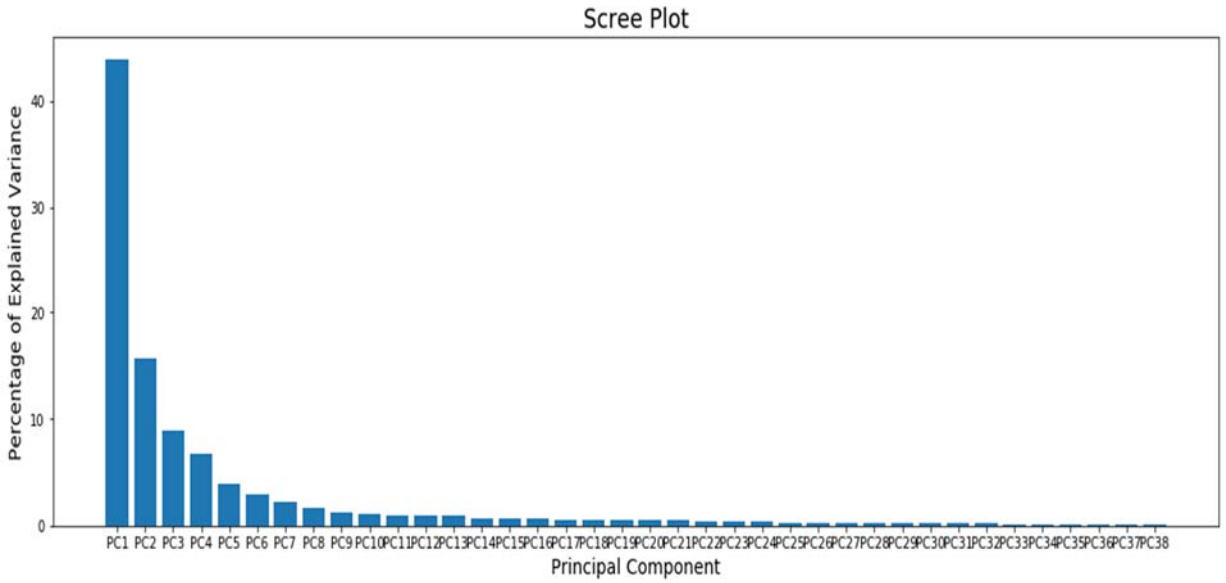

Figure 1. Scree Plot: Percentage of Explained Variance

Table 1 below lists the loading scores of player attributes in the first principal component. Loading scores of all player attributes in the first three principal components show that no one single player attribute contributes significantly more than any other attributes to the variance. None the less, the ball Control attribute was the leading attribute that explained the most variance. Considering the players as one group, overall rating has a low loading score in the PC1 dimension.

Table 1. Loading Scores of Player Attributes in the First Principal Component, PC1

Player Attributes	Loading Scores	Player Attributes	Loading Scores	Player Attributes	Loading Scores
ball_control	***0.233895***	gk_positioning	0.193662	gk_kicking	0.121982
dribbling	0.226609	volleys	0.193077	reactions	0.105162
short_passing	0.220168	finishing	0.191773	aggression	0.087644
crossing	0.213529	acceleration	0.185480	***overall_rating***	***0.083033***
Curve	0.211442	penalties	0.183370	potential	0.080174
long_shots	0.211257	sprint_speed	0.182038	interceptions	0.071469
positioning	0.204851	long_passing	0.177636	standing_tackle	0.065230
shot_power	0.199655	agility	0.168734	sliding_tackle	0.059237
vision	0.197614	stamina	0.158548	marking	0.052163
gk_diving	0.196600	balance	0.153811	strength	0.028138
free_kick_accuracy	0.196366	height	0.134852	age	0.005573
gk_reflexes	0.195234	heading_accuracy	0.132158	jumping	0.001561
gk_handling	0.195011	weight	0.125473		

Visualization

Research Question: What attributes set the players apart?

From the findings above, it was decided not to reduce any dimension for further analysis. Scatter plots were created using data from the first three principal components: PC1 versus PC2 and PC1 versus PC3. Two distinct clusters were noted. The largest variation is seen in the first principal component dimension, PC1.

To find out what player attribute(s) plays a part in distinguishing the players into the subgroups, color was added in the plots for one player attribute at a time for all 38 attributes (lighter hue indicates higher scores). Scatter plots in Figures 2 through 6 indicate that goalkeeper attributes play a part in separating the players into the two clusters. It is reasonable to believe that the smaller subgroup shown on the right of the scatter plots consist of goalkeepers, who score higher in goalkeepers attributes (lighter hue). There are two reasons: a) Each team usually designates only one player as the goalkeeper, hence makes goalkeepers a small subgroup in the dataset. b) Goalkeepers, who focus their training in goalkeeping skills, should score higher in these attributes.

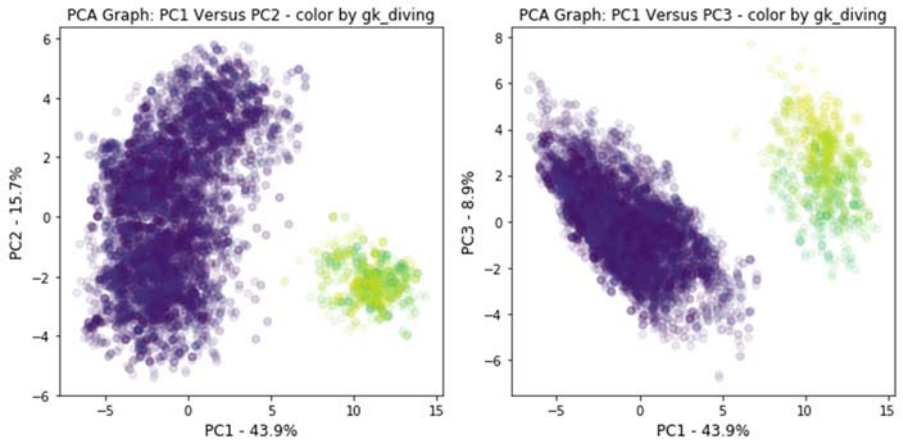

Figure 2. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk_diving

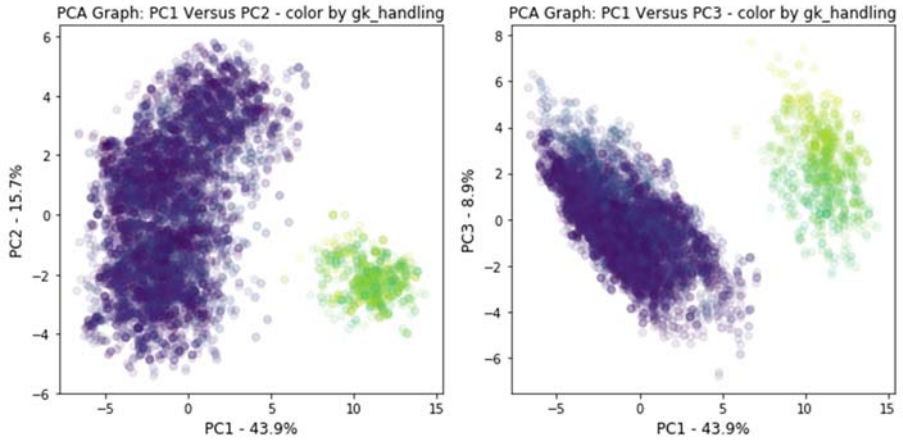

Figure 3. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk_handling

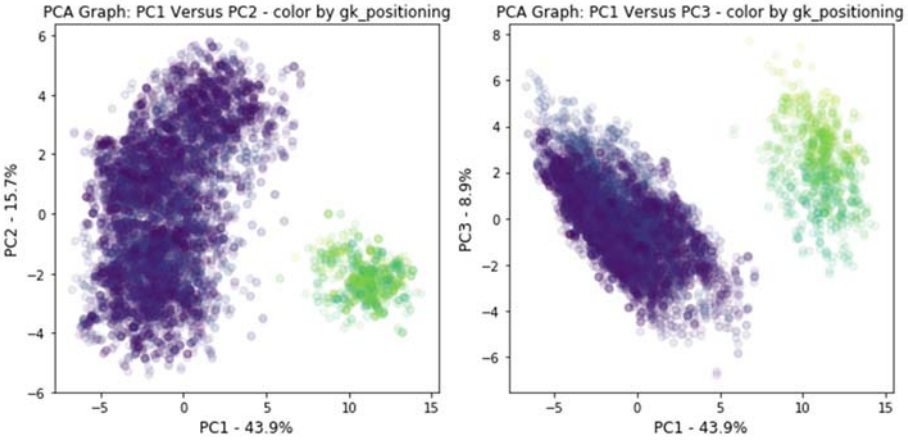

Figure 4. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk_positioning

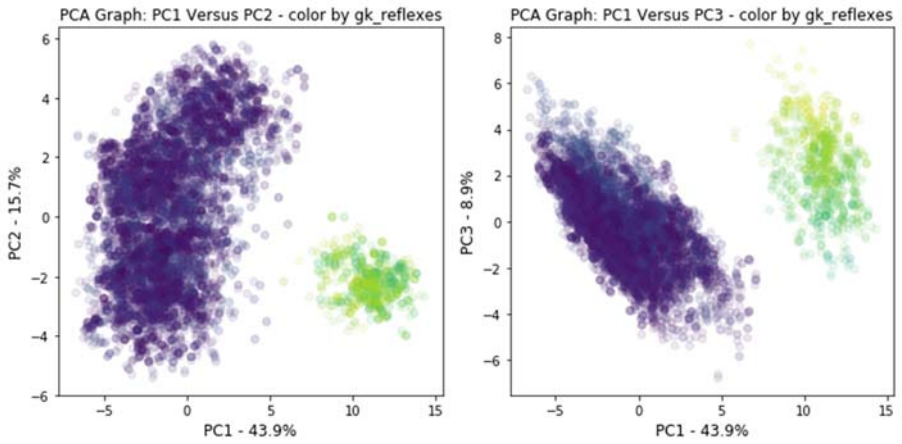

Figure 5. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk_reflexes

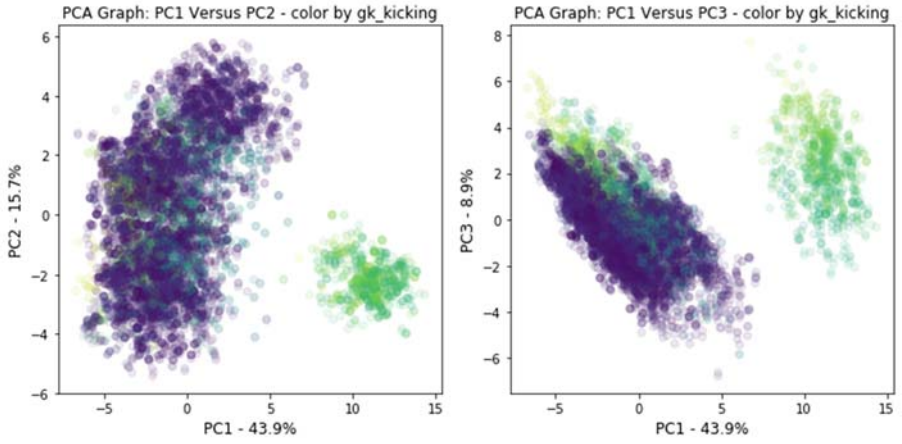

Figure 6. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk_kicking

Ball control attribute, with the highest loading score of 0.2339 in the first principal component, explains the most variance in the x-axis, PC1 dimension. Figure 7 below indicates that the small subgroup of goalkeepers on the right scores lower (darker hue) in the ball control attribute than the rest of the players (larger subgroup on the left). This attribute also set the players apart.

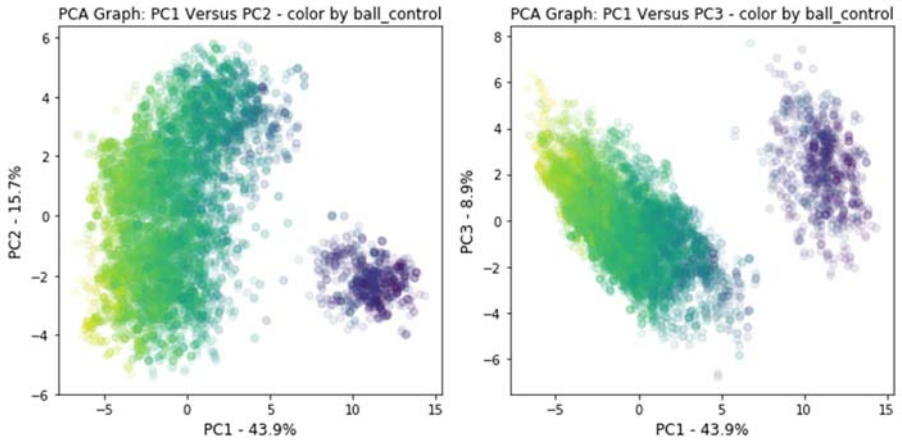

Figure 7. PC1 Versus PC2 and PC1 Versus PC3 – Colored by Ball Control

Figure 8 shows the distribution plots of all players attributes: age, height, weight, overall_rating, potential, crossing, finishing, heading_accuracy, short_passing, volleys, dribbling, curve, free_kick_accuracy, long_passing, ball_control, acceleration, sprint_speed, agility, reactions, balance, shot_power, jumping, stamina, strength, long_shots, aggression, interceptions, positioning, vision, penalties, marking, standing_tackle, sliding_tackle, gk_diving, gk_handling, gk_kicking, gk_positioning, gk_reflexes. It presents further evidence of separation of the players into subgroups by a few player attributes.

The distribution plots below show a small subgroup of players with relatively higher scores in gk_diving, gk_handling, gk_kicking, gk_positioning and gk_reflexes goalkeeping attributes (yellow highlight below and Figure 9) but with relatively low scores in ball control attribute (red highlight below and Figure 10). This reaffirms that goalkeeping and ball control attributes set the soccer players (goalkeepers) apart from the rest of players. This conclusion agrees with findings from the scatter plots in PC1 versus PC2 and PC1 versus PC3, in which a smaller subgroup is associated with lighter hue for relatively higher scores in goalkeeping attributes (Figures 2 through 6) and with darker hue for lower scores in ball control attribute (Figure 7).

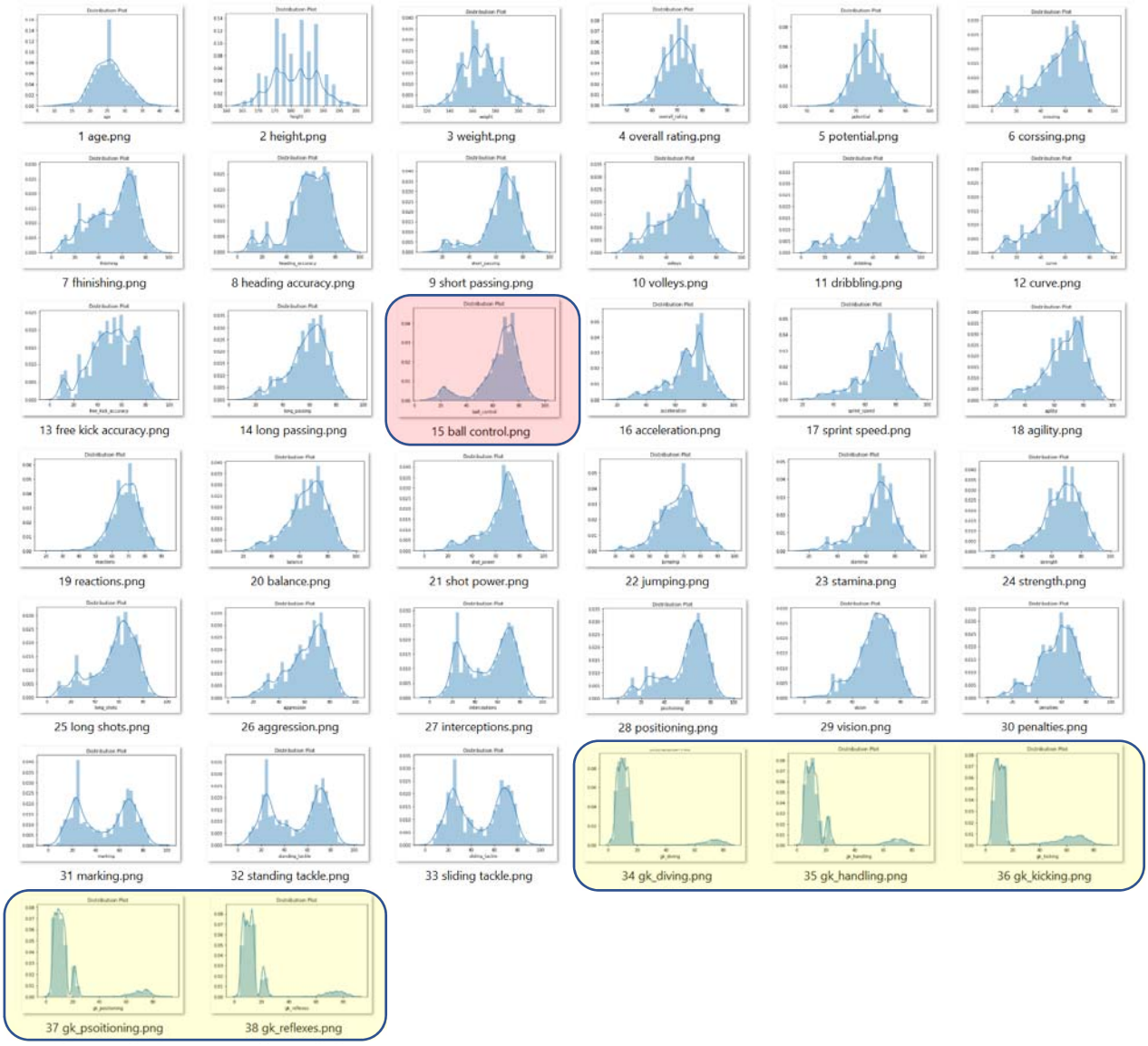

Figure 8. Distribution Plots of All Player Attributes

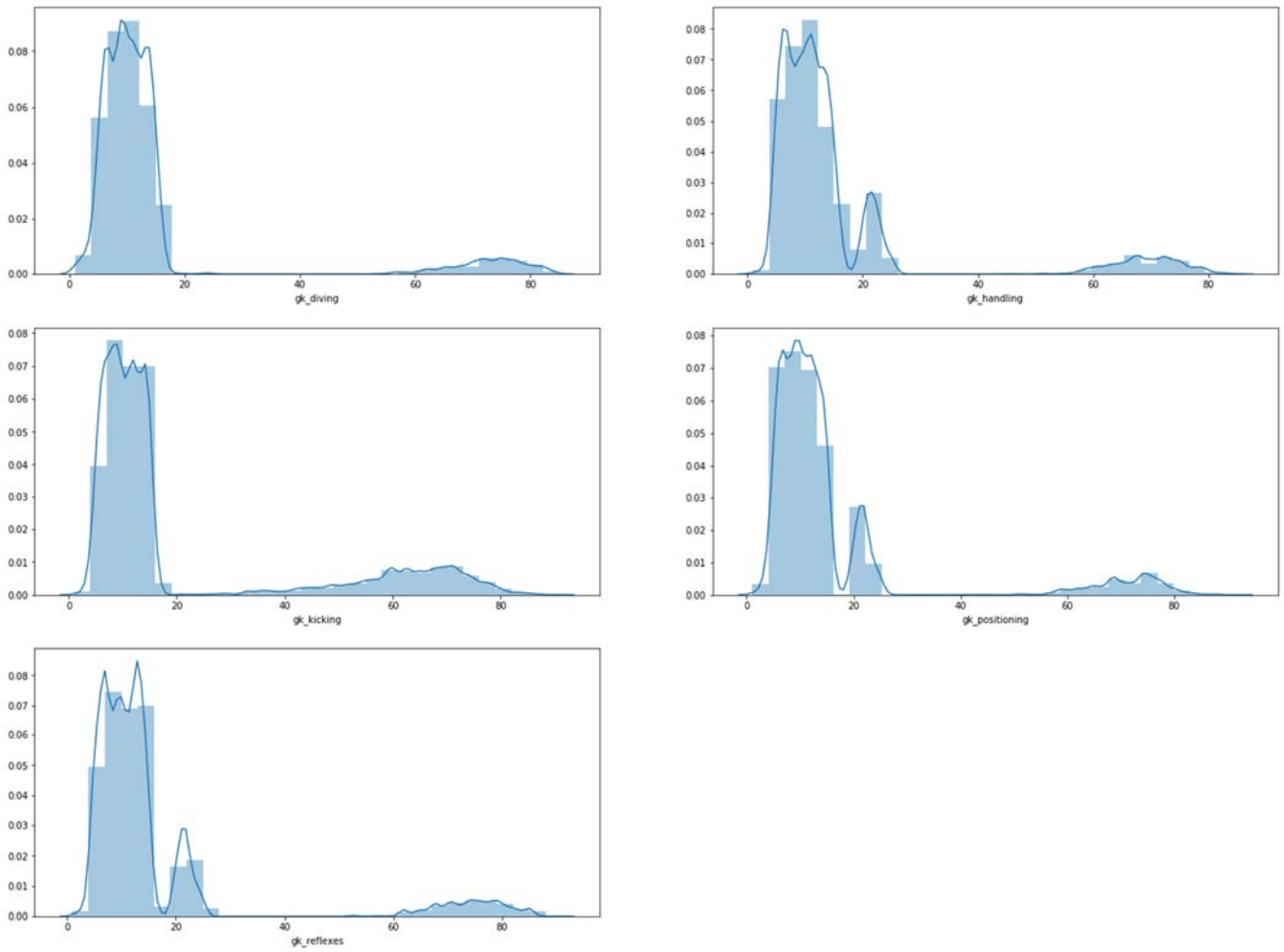

Figure 9. Distribution Plots of Soccer Players by Goalkeeping Attributes: Diving, Handling, Kicking, Positioning and Reflexes

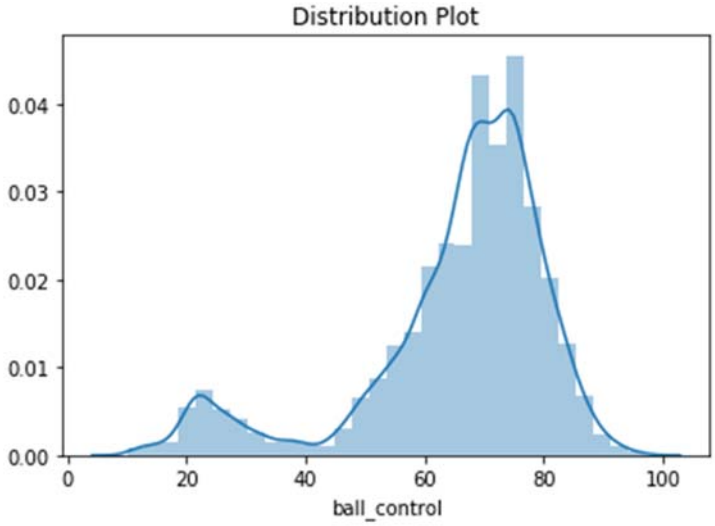

Figure 10. Distribution Plots of Soccer Players by Ball Control Attribute

Visualization in Ball Control Attribute Versus Goalkeeping Attributes

With the findings from above, jointplots were created using ball control attribute against the five goalkeeping attributes (see Figures 11 through 15). These scatter plots joint with distribution plots visualize goalkeepers as the smaller subgroup (top left) and non-goalkeepers as the larger subgroup (bottom).

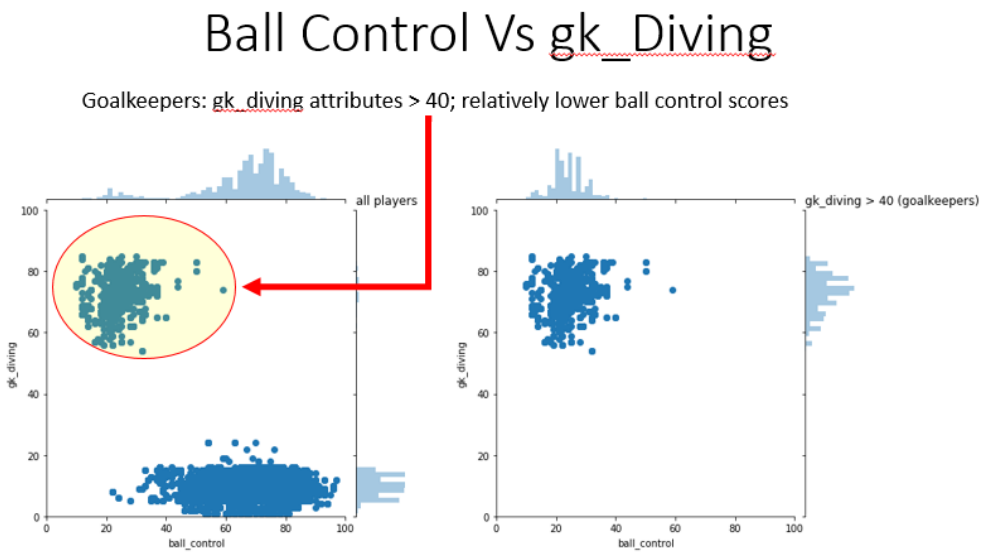

Figure 11. Jointplot: Ball Control Attribute Versus gk_Diving Attribute

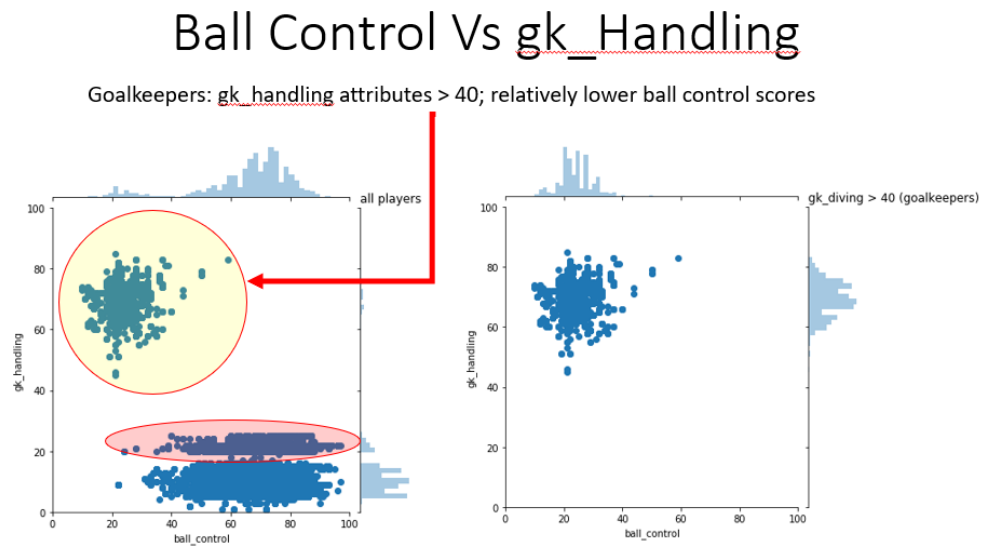

Figure 12. Jointplot: Ball Control Attribute Versus gk_Handling Attribute

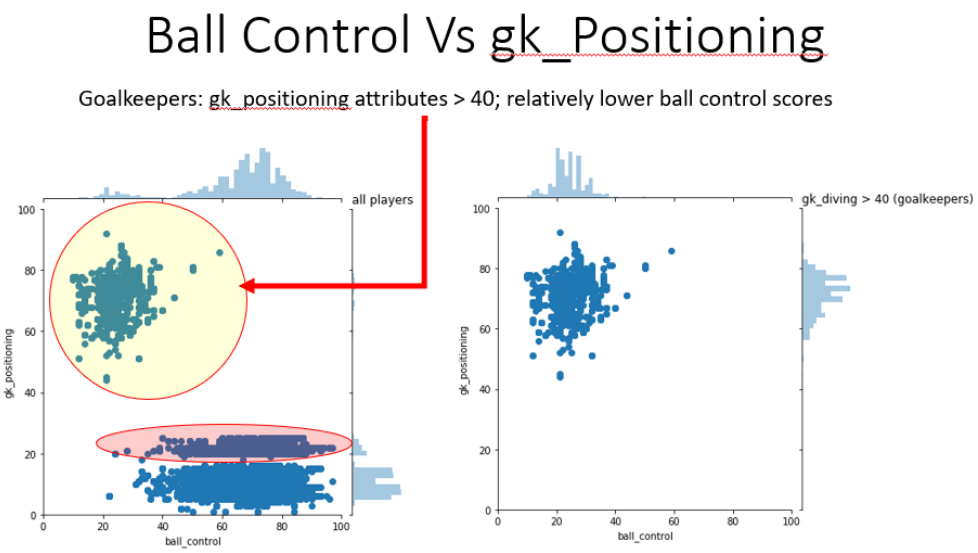

Figure 13. Jointplot: Ball Control Attribute Versus gk_Positioning Attribute

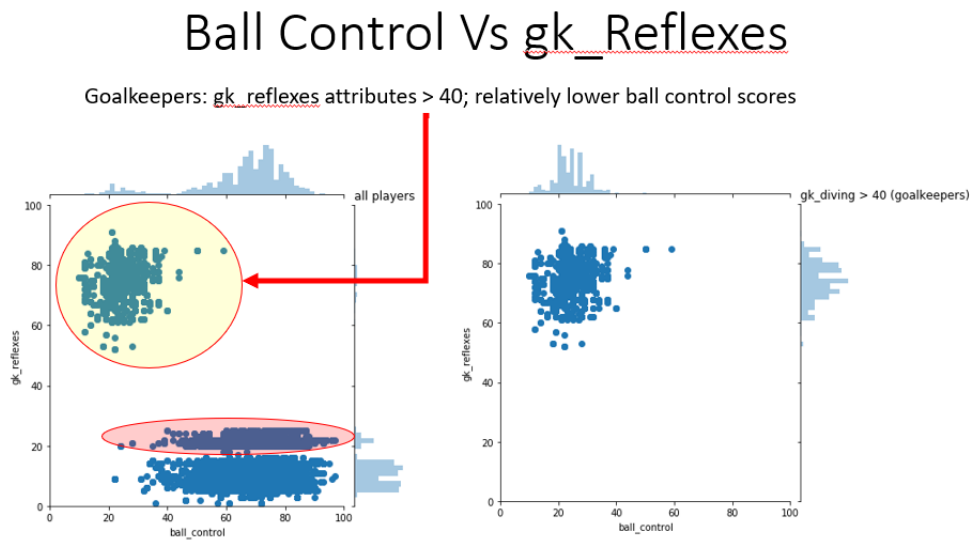

Figure 14. Jointplot: Ball Control Attribute Versus gk_Reflexes Attribute

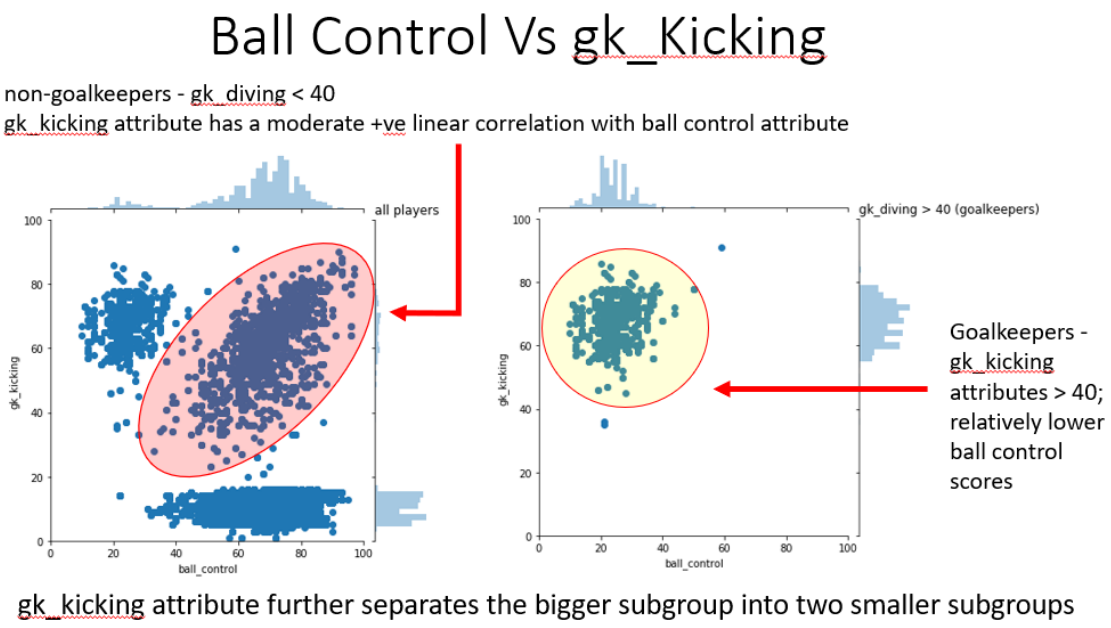

Figure 15. Jointplot: Ball Control Attribute Versus gk_Kicking Attribute

Side Notes: gk_Kicking Attribute

Noted in Figure 15, non-goalkeepers were further split apart into two subgroups by gk_kicking attribute. The group highlighted in red had a moderate positive linear correlation between their scores in gk_kicking attribute and their scores in ball control attribute. This is not surprising as there may be some common techniques that are pertinent to both attributes that are important for certain player positions. Similar logic goes for other attribute pairs, such as: acceleration and sprint speed attributes (correlation coefficient = 0.9129), and ball control and dribbling (0.9273). It would be interesting to conduct further analysis to find out whether the players highlighted red in Figures 12 through 14 are the same group of players highlighted red in Figure 15.

Research Question: Which player attribute contributes most to a player's overall rating?

All Players

Correlation coefficient matrix was generated for all 38 player attributes. Table 2 below lists the correlation coefficients between overall rating and all other player attributes. Overall rating and potential has the highest correlation coefficient, 0.7840. Since the potential score may be considered some sort of rating similar to overall rating, the second highest correlation coefficient is considered. Overall rating has a strong positive correlation coefficient of 0.7248 with the reactions attribute (see highlighted).

Table 2. Correlation Coefficient between Overall Rating and All Numeric Attributes

	overall_rating		**overall_rating**		**overall_rating**
age	0.3826	long_passing	0.4300	interceptions	0.2373
height	0.0259	ball_control	0.3726	positioning	0.2782
weight	0.0351	acceleration	0.2137	vision	0.3992
overall_rating	1.0000	sprint_speed	0.2184	penalties	0.3373
potential	0.7840	agility	0.2211	marking	0.1106
crossing	0.2911	reactions	0.7248	standing_tackle	0.1456
finishing	0.2694	balance	0.1094	sliding_tackle	0.1159
heading_accuracy	0.2403	shot_power	0.3703	gk_diving	0.0501
short_passing	0.4161	jumping	0.2112	gk_handling	0.0373
volleys	0.2908	stamina	0.2487	gk_kicking	0.0795
dribbling	0.2991	strength	0.2404	gk_positioning	0.0361
curve	0.2987	long_shots	0.3267	gk_reflexes	0.0437
free_kick_accuracy	0.3048	aggression	0.2615		

For a visual confirmation, a scatter plot of the reactions attribute against overall rating with trendline was created. Figure 20 shows a strong positive linear correlation. Figure 21 presents the correlation relationship with hue for two categorical attributes, defensive work rate and attacking work rate. Both plots show strong positive linear correlation within each subcategory by defensive work rate and by attacking work rate.

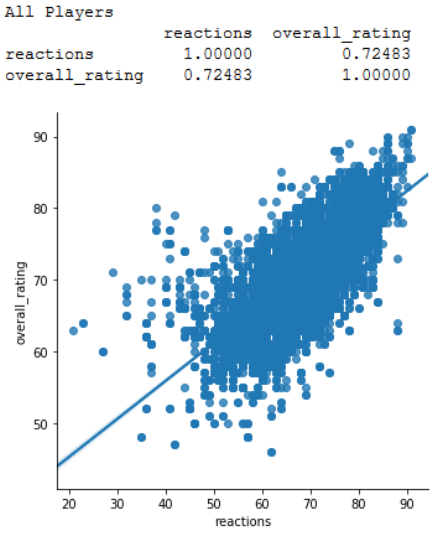

Figure 20. Scatter Plot of Reactions Attribute Against Overall Rating

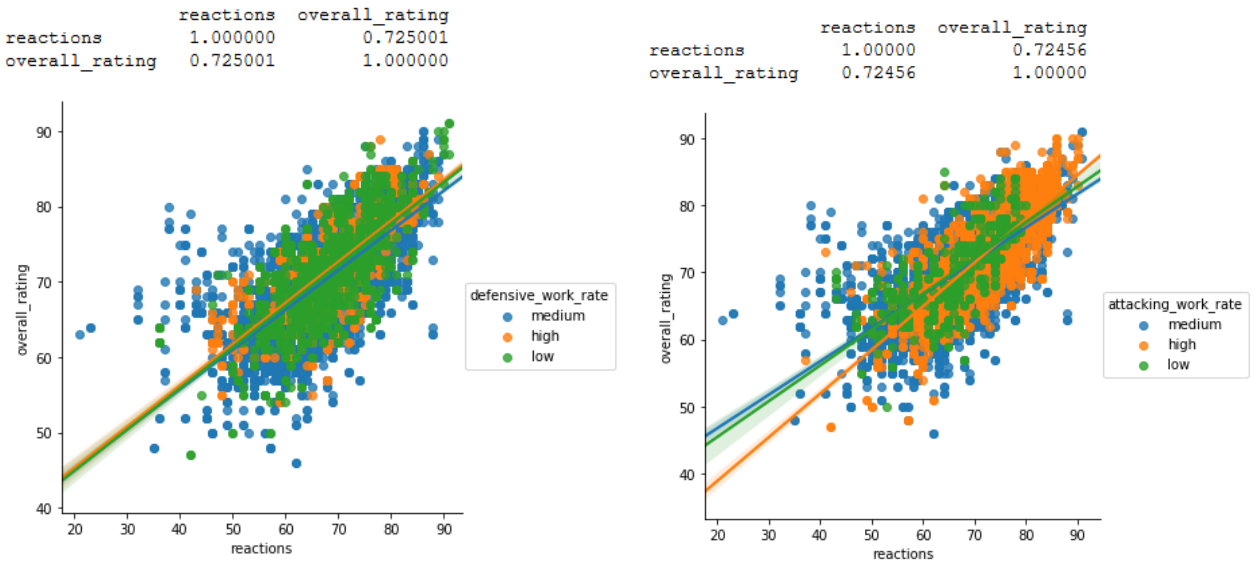

Figure 21. Scatter Plot of Reactions Attribute Against Overall Rating with Hue by Defensive Work Rate and Attacking Work Rate

Correlation Coefficients in Subgroups

However, there are two distinct subgroups in the dataset (goalkeepers and the rest players). Hence, the correlation of overall rating with player attributes is further evaluated in each subgroup separately. Tables 3 and 4 list the correlation coefficients between overall rating and all numeric attributes for the goalkeeper subgroup and the rest players, respectively. Highlighted numbers indicate strong positive correlation.

Table 3. Correlation Coefficient between Overall Rating and All Numeric Attributes - Goalkeeper Subgroup

	overall_rating		**overall_rating**		**overall_rating**
age	0.4274	long_passing	0.0919	interceptions	0.0655
height	-0.0283	ball_control	0.2050	positioning	-0.0385
weight	-0.0251	acceleration	0.3000	vision	0.1784
overall_rating	1.0000	sprint_speed	0.2066	penalties	-0.0169
potential	0.7805	agility	0.3076	marking	0.0523
crossing	0.0837	reactions	0.5609	standing_tackle	0.1062
finishing	0.0037	balance	-0.0955	sliding_tackle	0.1638
heading_accuracy	-0.0007	shot_power	0.1054	gk_diving	0.8979
short_passing	0.1165	jumping	0.4408	gk_handling	0.8281
volleys	0.1428	stamina	0.0990	gk_kicking	0.6697
dribbling	0.0882	strength	0.2180	gk_positioning	0.8632
curve	0.0373	long_shots	0.0251	gk_reflexes	0.8778
free_kick_accuracy	0.0166	aggression	0.1304		

Table 4. Correlation Coefficient between Overall Rating and All Numeric Attributes - Non_Goalkeeper Subgroup

	overall_rating		overall_rating		overall_rating
age	0.3814	long_passing	0.5553	interceptions	0.2728
height	0.0257	ball_control	0.6467	positioning	0.3837
weight	0.0368	acceleration	0.2624	vision	0.5061
overall_rating	1.0000	sprint_speed	0.2825	penalties	0.4523
potential	0.7848	agility	0.2420	marking	0.1299
crossing	0.4089	reactions	0.7595	standing_tackle	0.1708
finishing	0.3494	balance	0.1471	sliding_tackle	0.1333
heading_accuracy	0.3741	shot_power	0.5300	gk_diving	0.0717
short_passing	0.6655	jumping	0.1947	gk_handling	0.0059
volleys	0.3779	stamina	0.3278	gk_kicking	0.0744
dribbling	0.4587	strength	0.2426	gk_positioning	-0.0088
curve	0.4103	long_shots	0.4491	gk_reflexes	0.0254
free_kick_accuracy	0.4056	aggression	0.3122		

Visual Confirmation of Correlation in Subgroups

Scatter plots with trendlines were created for visual confirmation of correlation relationship between overall rating and correlated variables for each subgroup separately.

Non_Goalkeeper Subgroup

Figure 22 shows a stronger positive linear correlation between player reactions attribute and overall rating for non-goalkeeper subgroup alone, thus confirms that reactions attribute contributes most to the players' overall rating.

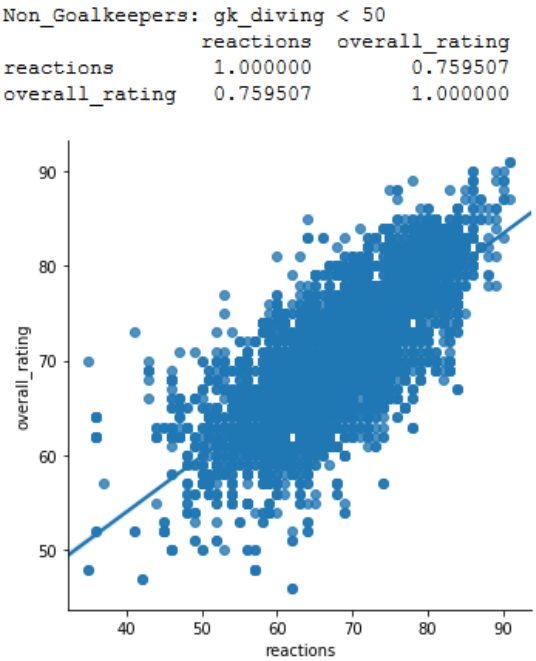

Figure 22. Non_Goalkeepers Subgroup: Scatter Plot of Reactions Attribute Against Overall Rating

Goalkeeper Subgroup

Figures 23 and 24 present the visual confirmation that gk_diving, gk_reflexes, gk_positioning and gk_handling attributes have a strong positive linear correlation with overall rating. The last goalkeeping attribute, gk_kicking, and reactions attribute only had a moderate positive linear correlation with overall rating (Figure 25).

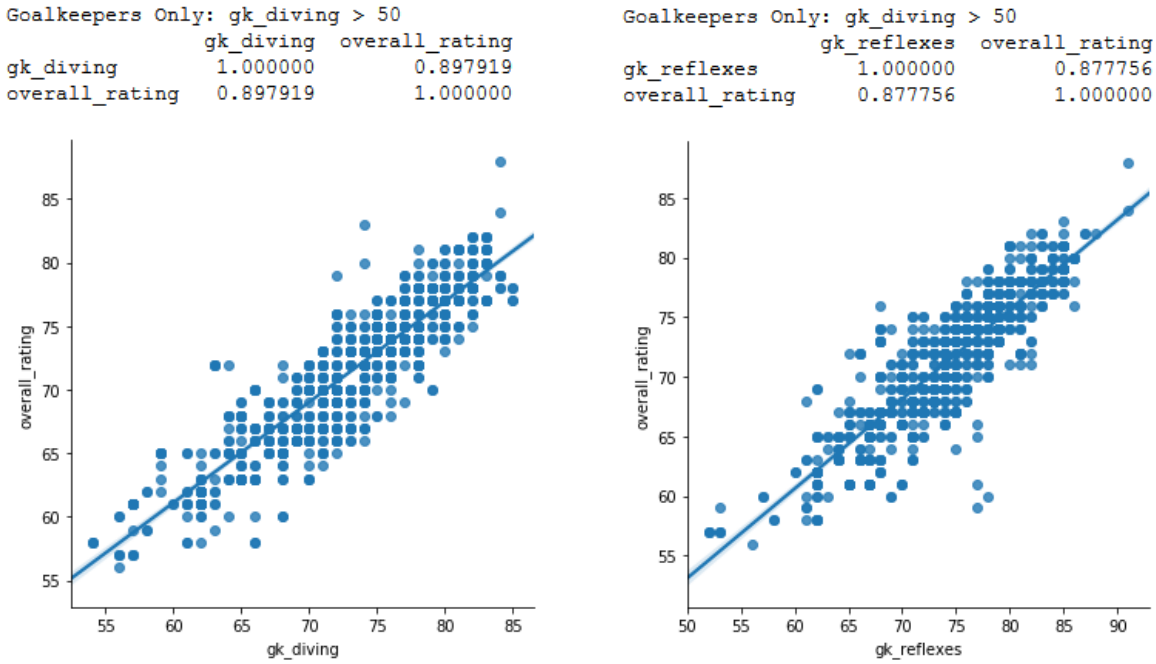

Figure 23. Goalkeeper Subgroup: Scatter Plots of gk_Diving (Left) and gk_Reflexes (Right) Attributes Against Overall Rating

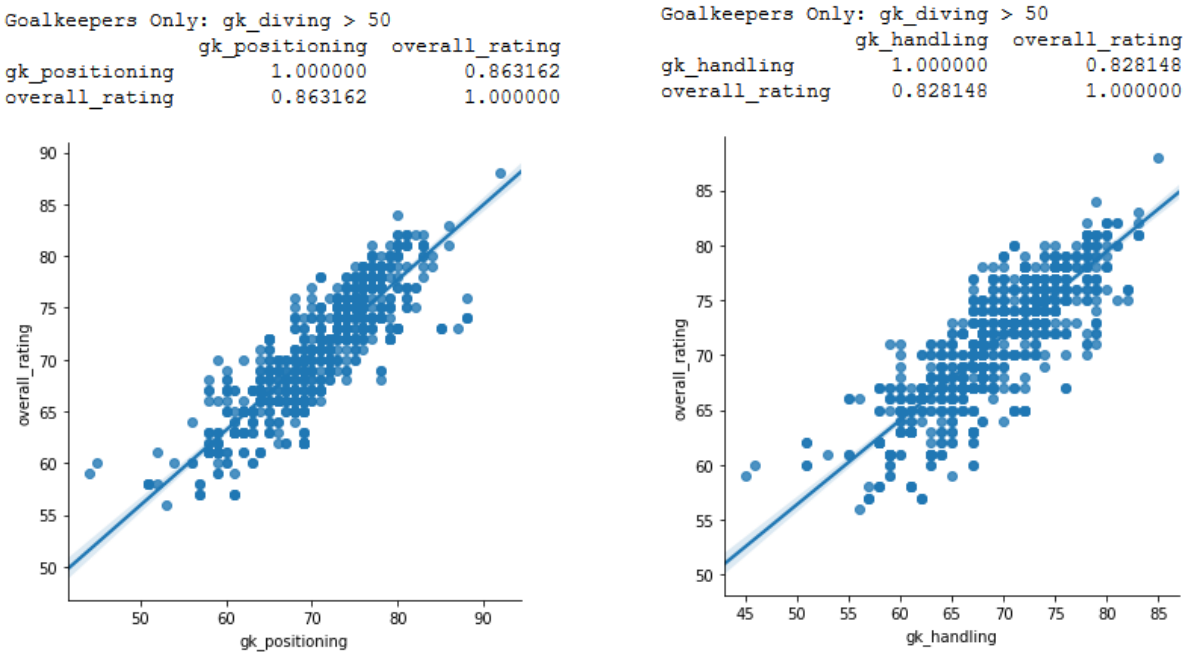

Figure 24. Goalkeeper Subgroup: Scatter Plots of gk_Positioning (Left) and gk_Handling (Right) Attributes Against Overall Rating

```
Goalkeepers Only: gk_diving > 50
                gk_kicking  overall_rating
gk_kicking        1.000000        0.669727
overall_rating    0.669727        1.000000
```

```
Goalkeepers Only: gk_diving > 50
                reactions  overall_rating
reactions         1.00000         0.56092
overall_rating    0.56092         1.00000
```

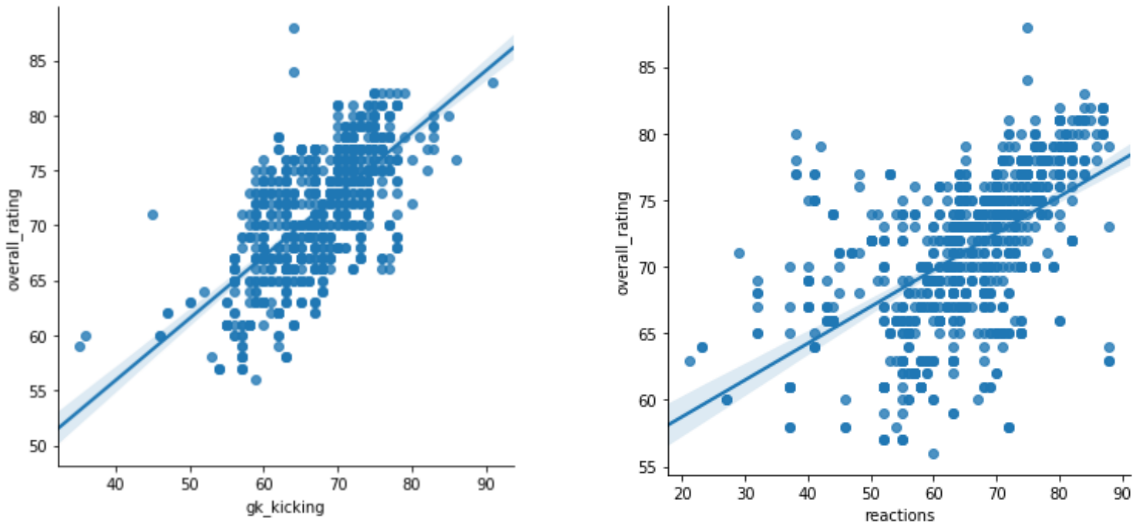

Figure 25. Goalkeeper Subgroup: Scatter Plots of gk_Kicking (Left) and Reactions (Right) Attributes Against Overall Rating

Conclusions and Answers to Research Question

Research Question: What attributes set the players apart?

Four of the five goalkeeping attributes and ball control attribute separated the players apart:

a) a small subgroup of 930 goalkeepers, who have relatively high scores in goalkeeping attributes, gk_diving, gk_handling, gk_positioning, and gk_reflexes, but low in ball control, and
b) a larger subgroup of the rest of the 9968 players, whose scores relatively low in the goalkeeping attributes, but high in ball control.

With gk_kicking attribute and ball control attribute, players were, however, divided into three apparent subgroups:

a) one small subgroup of 930 goalkeepers, who have relatively high scores in gk_kicking, but low in ball control,
b) one slightly larger subgroup of 1413 players, whose scores in gk_kicking (>= 20) had a moderate positive linear correlation to their ball control scores, and
c) one much larger subgroup of the rest of 8556 players, whose score in gk_kicking (< 20) had no correlation with their ball control scores.

Typically, soccer players concentrate their drills on the skills they need for their positions. Drills for goalkeepers are very different from other players because their position requires them to use primarily their hands to block the soccer ball from entering the goal. They typically divert the direction of the soccer away from the goal or take control of the soccer with their hands. Goalkeepers drills typically specialize in collecting balls, jumping to divert or collect overhead balls, goal positioning to the

attacking angle, diving for balls, ball distribution once collected, and so on. Specialized training brings special attributes that separate them apart from other players.

According to soccer game rules, if the offensive team kicks the ball out of the field, play is restarted with a goal kick. The goal kick can be taken by any player, rather than limited to the goalkeeper. So, it is not uncommon to have full back teammates to take the goal kick. This explains why there is a slightly larger subgroup of players who score well in gk_kicking and ball control.

Research Question: Which player attribute contributes most to a player's overall rating?

For all players in the dataset, reactions attribute with the strongest correlation coefficient (0.724830) contributes most to a player's overall rating.

When evaluating in subgroup levels, the answer is different for each group. For goalkeeper subgroup, gk_diving (corr coeff. = 0.897919), gk_reflexes (corr coeff. = 0.877756), gk_positioning (corr coeff. = 0.863162) and gk_handling (corr coeff. = 0.828148) goalkeeping attributes have very strong positive linear correlation with and contribute most to the players overall rating. For the rest of the players as the larger subgroup, reactions (corr coeff. = 0.759507) has a strong positive linear correlation with and contributes most to players' overall rating.

Limitation and Future Study

The limitation of this project is that the database does not contain player's position or team formation information for any game. So, to validate the answers to the research questions, such information will have to be obtained from other source, such as sofifa.com. Figure 26 shows the players and team formation of the soccer team, Liverpool, for 2018. Figure 27 presents attributes of a player from the team and similar players of other teams and their overall rating and potential scores.

As shown in Figure 28 there are various modern soccer team formations. Gray, blue, yellow and red dots represent players in goalkeeping, defensive, midfield, and attacking positions at the goal and in the back field, midfield and forefront. The number of players in the back, midfield and forward varies in different formations. But there is always only one goalkeeper. The drills of players may vary slightly based on their positions in the team formation. These team formation factors will affect the scores of various attributes for each player. However, there is a tactical theory in football, Total Football (Dutch: *totaalvoetbal*), in which any outfield player can take over the role of any other player in a team. The theory requires players to be comfortable in multiple positions; hence, it places high technical and physical demands on them. For teams that embrace this theory, players are likely to have attributes across positions. If player position and team formation information is available, the study may be bring to a whole new level for the future.

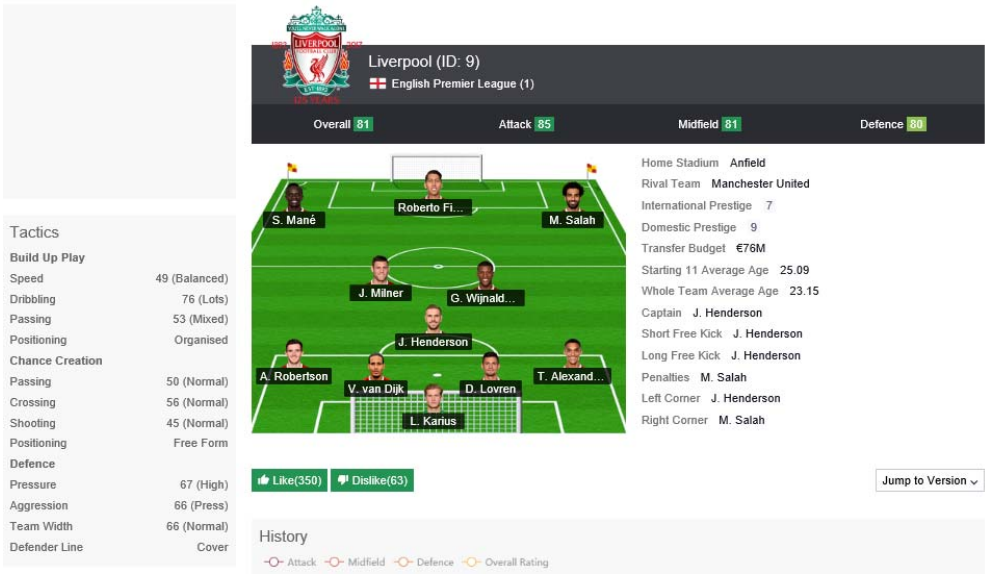

Figure 26. Players and Team Formation of Soccer Team, Liverpool, for 2018.

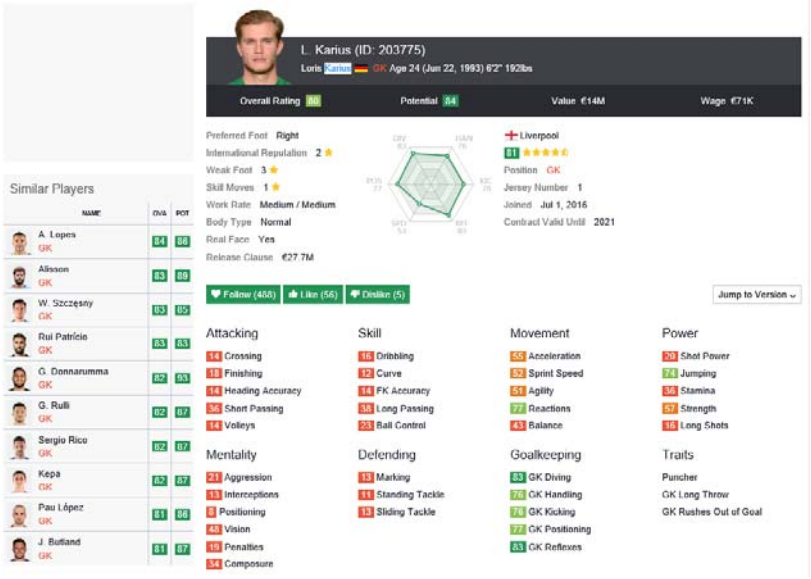

Figure 27. Attributes of a Player from the Soccer Team, Liverpool and Similar Players with Overall Rating and Potential

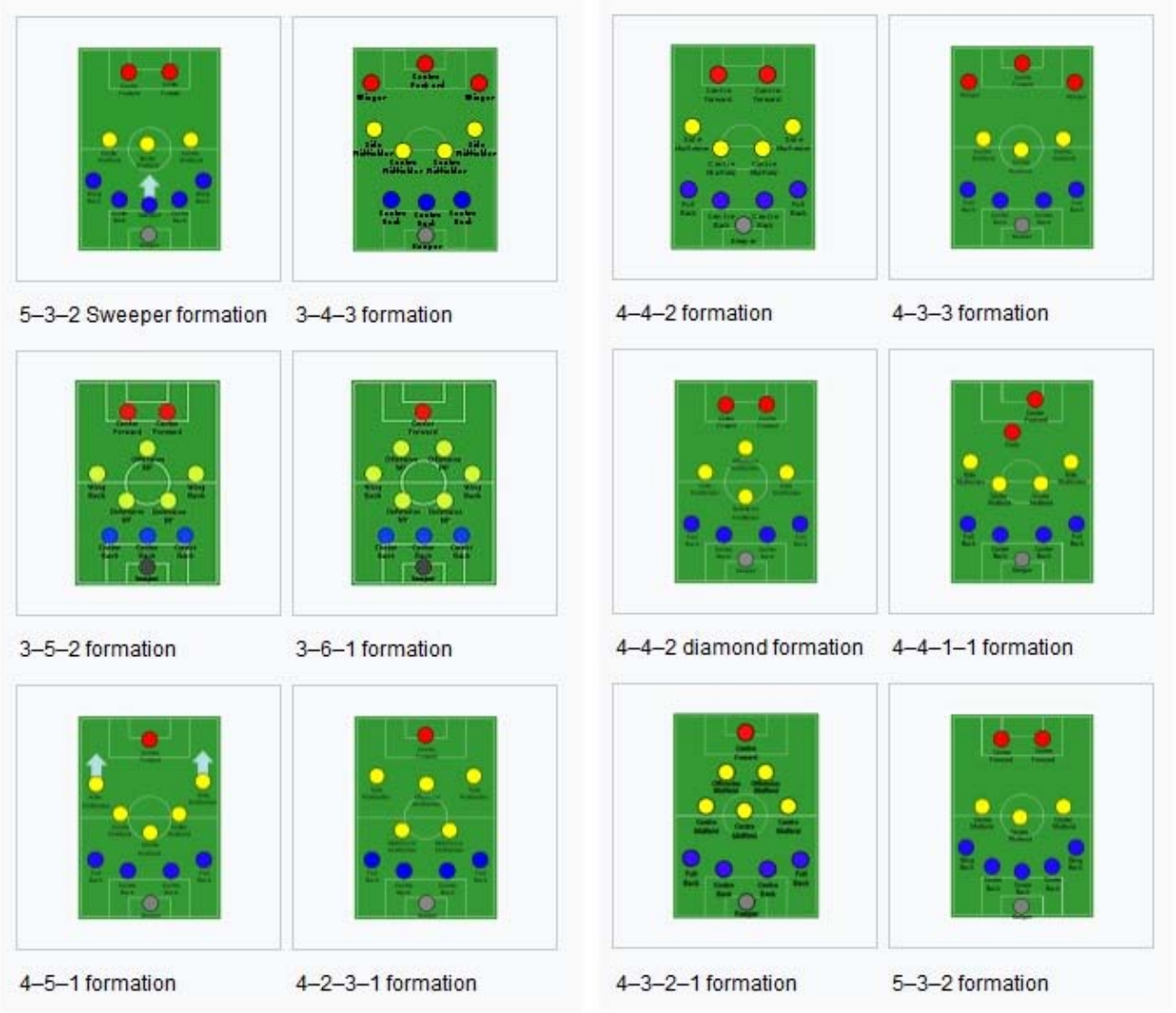

Figure 28. Examples of Modern Soccer Team Formations

APPENDIX A: Source of Data

Data for this project was downloaded from this website: https://www.kaggle.com/hugomathien/soccer

The following presents brief information of the database in SQLite3 format:

- +25,000 matches
- +10,000 players
- 11 European Countries with their lead championship
- Seasons 2008 to 2016
- Players and Teams attributes sourced from EA Sports FIFA video game series #
- Team line up with squad formation (X, Y coordinates)
- Betting odds from up to 10 providers
- Detailed match events (goal types, possession, corner, cross, fouls, cards etc...) for +10,000 matches

**16th Oct 2016: New table containing teams attributes from FIF !*

Original data source for players and teams attributes tables: http://sofifa.com/ : players and teams attributes from EA Sports FIFA games. *FIFA series and all FIFA assets property of EA Sports.* Foreign keys "api_id" for players and matches are the same as the original data sources.

APPENDIX B: Jupyter Notebook - Python Codes and Outputs