A group of men playing a game of football

Description generated with high confidence

**Analysis on Soccer Player Attributes**

**By: Sau Kha**

**April 2018**



# Executive Summary

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/).

Research Questions:

The research questions for this project are:

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

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. For analysis, 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 Analysis:

PCA 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 is a total of 68.5% of the variance in the final dataset of soccer players. Loading scores of all player attributes in 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. Scatter plots, distribution plots, joinplots and correlation coefficient analysis explained the clusters and substantiated the answers to the research questions.

Conclusions:

1. Consider all players, reactions attribute has a strong positive linear correlation (coefficient = 0.7248) with overall rating. The higher the reactions score, the higher the overall rating of the player.
2. However, after grouping player attributes into categories, total goalkeeping score has a very strong positive linear correlation (coefficient = 0.9783) with overall rating of the goal keepers. The higher the total goalkeeping score, the higher the overall rating of the goalies. However, no such correlation was established for the remaining players.
3. The following attributes set players apart into subgroups: gk\_diving, gk\_kicking, gk\_reflexes, gk\_handling, gk\_positioning, df\_marking, df\_sliding\_tackle, df\_standing\_tackle and mt\_interceptions. First five attributes pertain to goalkeeper position while the rest relate to defensive maneuvers.

Appendix A presents all the Python scripts used in this study.

# Table of Contents

[Executive Summary i](#_Toc514354021)

[Introduction 1](#_Toc514354022)

[Research questions 1](#_Toc514354023)

[Intended Audience 1](#_Toc514354024)

[Reference 1](#_Toc514354025)

[Methodology and Analysis Approach 2](#_Toc514354026)

[Findings 3](#_Toc514354027)

[Principal Component Analysis 3](#_Toc514354028)

[Visualization 4](#_Toc514354029)

[What attributes set players apart? 4](#_Toc514354030)

[Ball Control and Goalkeeping Attributes 9](#_Toc514354031)

[Which player’s attribute contributes most to a player’s overall rating? 12](#_Toc514354032)

[What category of player attributes contributes most to a player’s overall rating? 13](#_Toc514354033)

[Conclusion 15](#_Toc514354034)

[Answer to Research Question 1 15](#_Toc514354035)

[Answer to Research Question 2 15](#_Toc514354036)

[Answer to Research Question 3 16](#_Toc514354037)

[Conclusion 16](#_Toc514354038)

[APPENDIX A: Source of Data 17](#_Toc514354039)

[APPENDIX B: Jupyter Notebook - Python Codes and Outputs 19](#_Toc514354040)

List of Tables

Table 1. Loading Scores of Player Attributes in the First Principal Component, PC1 4

Table 2. Correlation Coefficient between Overall Rating and All Numeric Attributes 12

Table 3. Player’s Total Score Per Attribute Category 13

**List of Figures**

Figure 1. Scree Plot: Percentage of Explained Variance 3

Figure 2. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk\_diving 4

Figure 3. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk\_handling 5

Figure 4. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk\_positioning 5

Figure 5. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk\_reflexes 5

Figure 6. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk\_kicking 6

Figure 7. Distribution Plots of Soccer Players by Goalkeeping Attributes: Diving, Handling, Kicking, Positioning and Reflexes 6

Figure 8. Distribution Plots of Soccer Players by Marking and Sliding Tackle Attributes 7

Figure 9. Distribution Plots of Soccer Players by Standing Tackle and Interceptions Attributes 7

Figure 10. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by Marking 7

Figure 11. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by Sliding Tackle 8

Figure 12. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by Standing Tackle 8

Figure 13. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by Interceptions 8

Figure 14. PC1 Versus PC2 and PC1 Versus PC3 – Colored by Ball Control 9

Figure 15. Jointplot: Ball Control Attribute Versus gk\_Diving Attribute 9

Figure 16. Jointplot: Ball Control Attribute Versus gk\_Handling Attribute 10

Figure 17. Jointplot: Ball Control Attribute Versus gk\_Positioning Attribute 10

Figure 18. Jointplot: Ball Control Attribute Versus gk\_Reflexes Attribute 11

Figure 19. Jointplot: Ball Control Attribute Versus gk\_Kicking Attribute 11

Figure 20. Scatter Plot of Reactions Attribute Against Overall Rating 12

Figure 21. Scatter Plot Matrix of all Attribute Categories and Overall Rating 13

Figure 22. Scatter Plot of Total Goalkeeping Attribute against Overall Rating 14

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

## 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:
   1. <http://football-data.mx-api.enetscores.com/> : scores, lineup, team formation and events
   2. <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: <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: 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>

# 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 read\_sql and to\_csv, IO tools from pandas library, 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:
   1. I was shock 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 left those fields blank.
   2. 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:
   1. 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.
   2. I created a series of charts using a programming loop, which allowed exploration systematically on each attribute.
   3. 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 easy on human eyes when compared to adding a z-axis for the third attribute.
   4. 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.
   5. 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 evaluation.
8. Calculated correlation coefficient matrix for all attributes. Located attributes with strong correlation coefficient with overall rating.
9. Grouped attributes into Total Basic Skill, Total Attack, Total Movement, Total Power, Total Mentality, Total Defending and Total Goalkeeping categories. Calculated total scores for each attribute category.
10. Conducted further analysis based on total scores per category and overall rating. Created scatter plot matrix and calculated correlation coefficient between attributes.
11. 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.

A close up of a logo

Description generated with very high confidence

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 three principal components show that no one single player attribute contributes significantly more than any other attributes to these three principal components.

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

### What attributes set players apart?

Scatter plots were created using data from the first three principal components: PC1 versus PC2 and PC1 versus PC3. Two distinct clusters were noted. Largest variation is seen in the first principal component dimension, PC1. Player attributes were visualized in the plots by adding hue by different player attributes (the higher the score the lighter the hue). Figures 2 and 6 display various scatter plots, which indicate that goalkeeper attributes play a part in separating the players into the two clusters.

A close up of a tree

Description generated with high confidence

Figure 2. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk\_diving

A close up of a tree

Description generated with high confidence

Figure 3. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk\_handling

A close up of a tree

Description generated with high confidence

Figure 4. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk\_positioning

A screenshot of a cell phone

Description generated with high confidence

Figure 5. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk\_reflexes

A close up of a tree

Description generated with high confidence

Figure 6. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by gk\_kicking

Figure 7 present the distribution plots of the soccer players by the five goalkeeping attributes: gk\_diving, gk\_handling, gk\_kicking, gk\_positioning and gk\_reflexes. A small subgroup of players is shown distinctively on the far right of the plots, showing the subgroup with very high scores in these goalkeeping attributes. This suggests that the goalkeeping attributes set the soccer players (goalkeepers) apart from the rest of players. This agrees with the plots in Figures 2 through 6, which shows a smaller subgroup in light green color (the higher the score the lighter the hue).

Additionally, it is noticeable that some players in the larger subgroup in Figure 6 (colored by gk\_kicking) were presented in lighter hue, suggesting that those players received relatively higher scores than the rest in the larger subgroup. This also agrees with the distribution plot for gk\_kicking attribute below, which shows a slightly bigger subgroup on the right (see red brackets).

A screen shot of a social media post

Description generated with high confidence

Figure 7. Distribution Plots of Soccer Players by Goalkeeping Attributes:   
Diving, Handling, Kicking, Positioning and Reflexes

The distribution plots shown in Figures 8 and 9 present bimodal distribution, suggesting that players may be divided into two subgroups by interceptions, standing tackle, sliding tackle and marking attributes. These attributes, however, have very low loading scores, 0.071469, 0.065230, 0.059237 and 0.052163, respectively, in the PC1 dimension. This explains why the distinct large and small (goalkeepers) subgroups in the PC1 versus PC2 and PC1 versus PC3 plots in Figures 10 through 13 are distinguished by dark and light hue for interceptions, standing tackle, sliding tackle and marking attributes, unlike Figures 2 through 6.

A close up of a logo

Description generated with high confidence A close up of a logo

Description generated with very high confidence

Figure 8. Distribution Plots of Soccer Players by Marking and Sliding Tackle Attributes

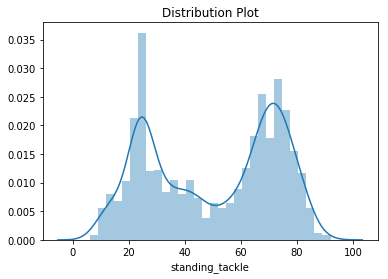
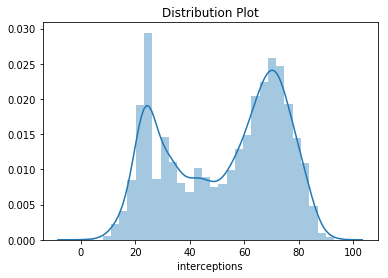
 

Figure 9. Distribution Plots of Soccer Players by Standing Tackle and Interceptions Attributes

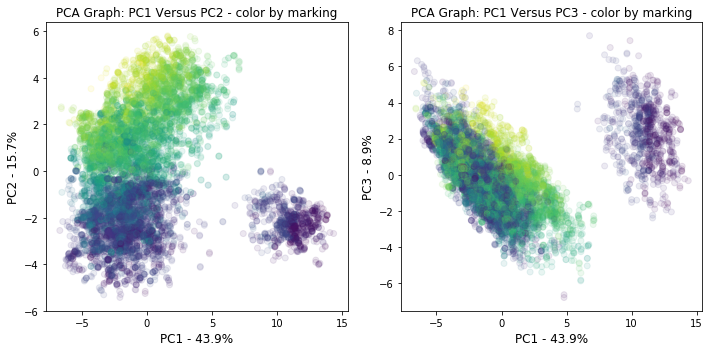


Figure 10. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by Marking

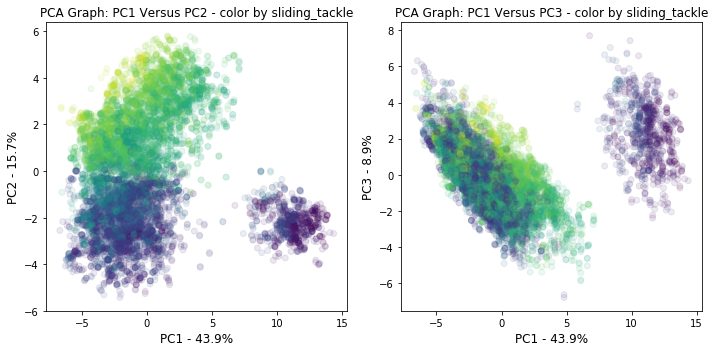


Figure 11. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by Sliding Tackle

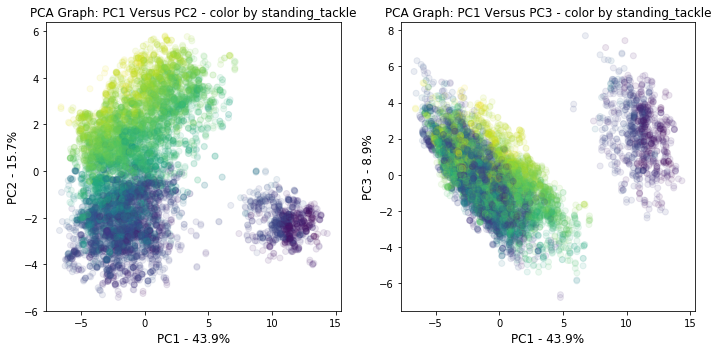


Figure 12. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by Standing Tackle

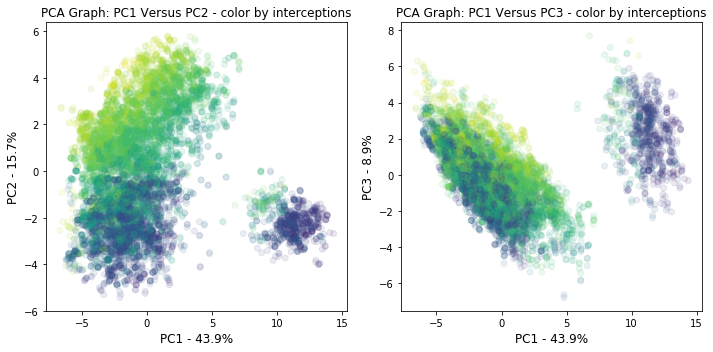


Figure 13. Scatter Plots of PC1 Versus PC2 and PC1 Versus PC3 – Colored by Interceptions

On the other hand, the ball control attribute has the highest loading score of 0.2339 in the PC1 dimension and explains the most variance for all players as a group. Figure 14 below suggests that the small goalkeeper subgroup on the right scores lower in the ball control attribute than the larger subgroup on the left (the higher the score the lighter the hue).

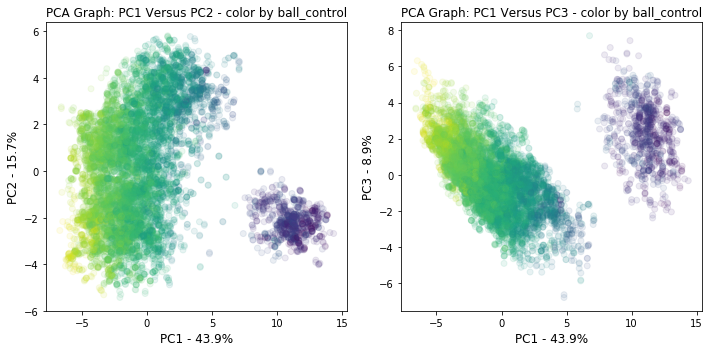


Figure 14. PC1 Versus PC2 and PC1 Versus PC3 – Colored by Ball Control

### Visualizing Ball Control Attribute Versus Goalkeeping Attribute

With the findings from above, jointplots were created using ball control attribute against the five goalkeeping attributes (see Figures 15 through 19). These scatter plots joint with distribution plots visualize goalkeepers as the smaller subgroup (top) and non-goalkeepers as the larger subgroup (bottom). Noted in Figure 19, there is a subgroup of non-goalkeepers who scored higher in gk\_kicking attribute that is positively correlated to scores for ball control attribute. This is not surprising as there may be some common techniques important to both attributes. Similar logic goes for other attribute pairs, for example: acceleration and sprint speed attributes (correlation coefficient = 0.9129), and ball control and dribbling (0.9273).

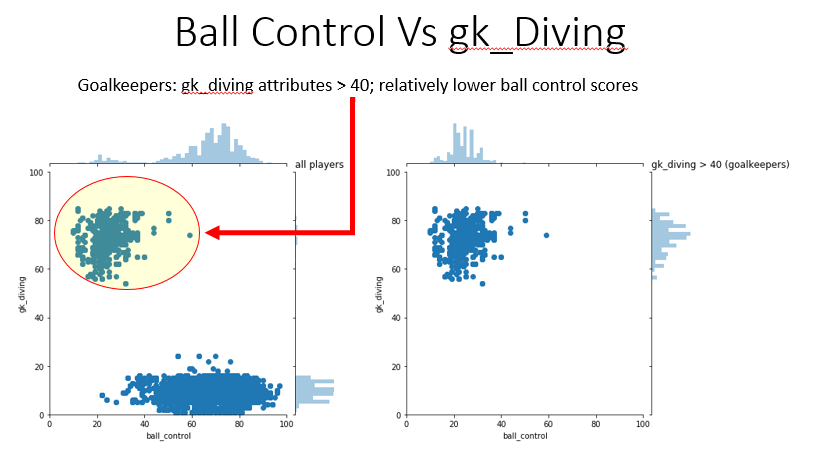


Figure 15. Jointplot: Ball Control Attribute Versus gk\_Diving Attribute

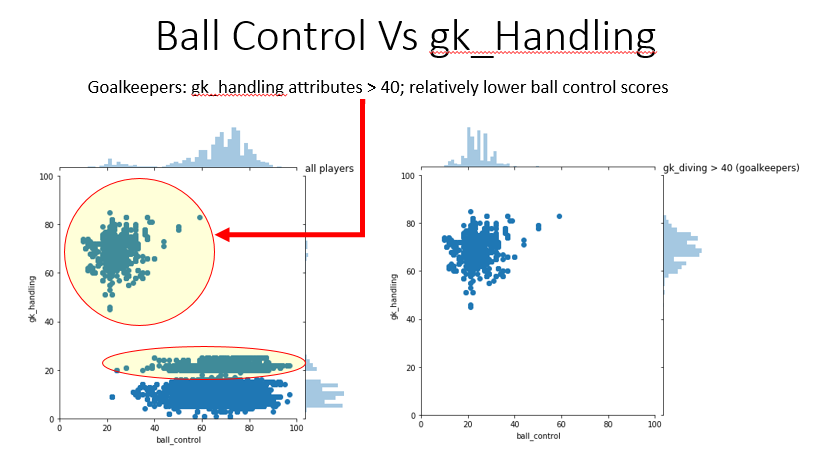


Figure 16. Jointplot: Ball Control Attribute Versus gk\_Handling Attribute

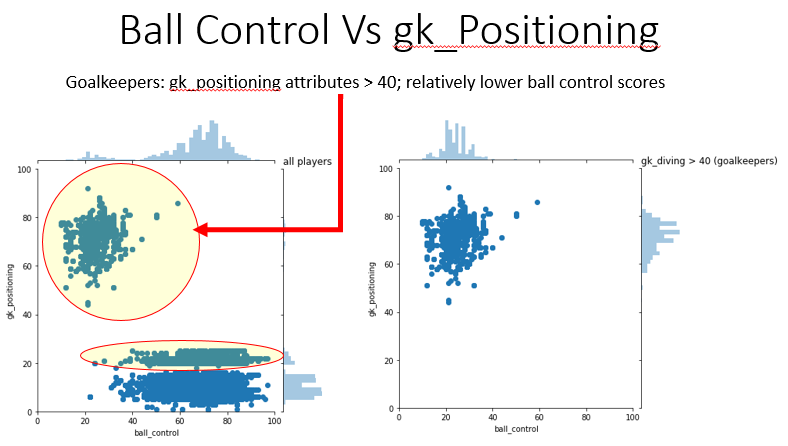


Figure 17. Jointplot: Ball Control Attribute Versus gk\_Positioning Attribute

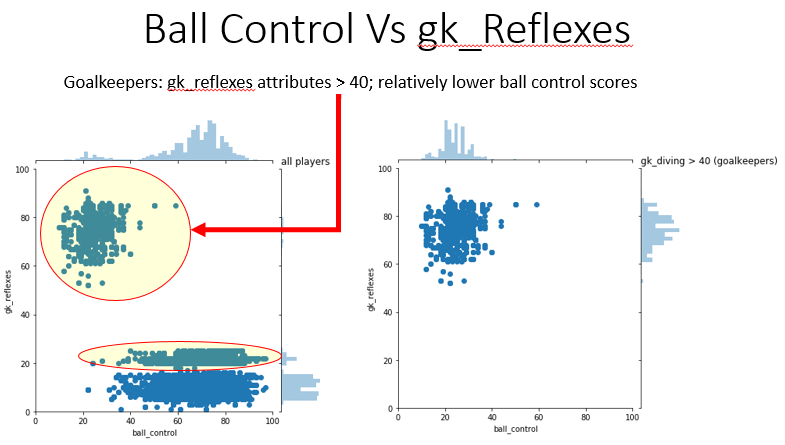


Figure 18. Jointplot: Ball Control Attribute Versus gk\_Reflexes Attribute

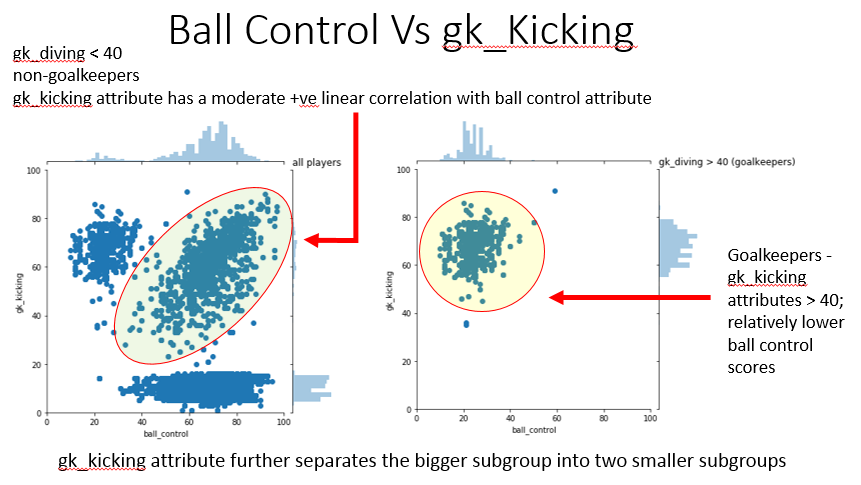


Figure 19. Jointplot: Ball Control Attribute Versus gk\_Kicking Attribute

### Which player’s attribute contributes most to a player’s overall rating?

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. The highest coefficient, 0.7840, was between overall rating and potential. Since “potential” was queried from the Player table instead of the Player Attributes table and may be considered as some sort of rating, the second highest correlation coefficient is taken into account.

Figure 20 shows that reactions has a strong linear positive correlation (coefficient = 0.7248) with player’s overall rating. Figure 21 presents the relationship with hue for 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.

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 |  |  |

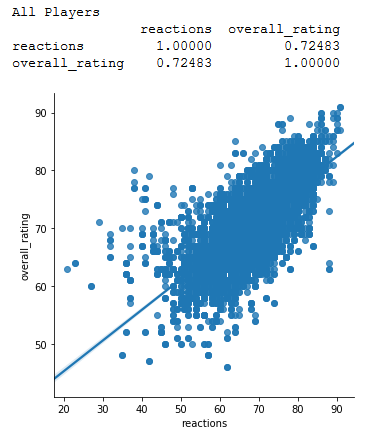


Figure 20. Scatter Plot of Reactions Attribute Against Overall Rating

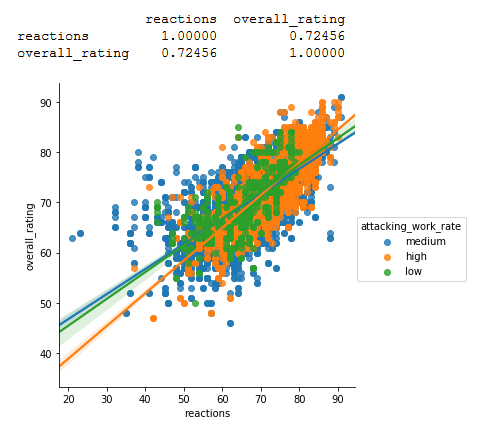
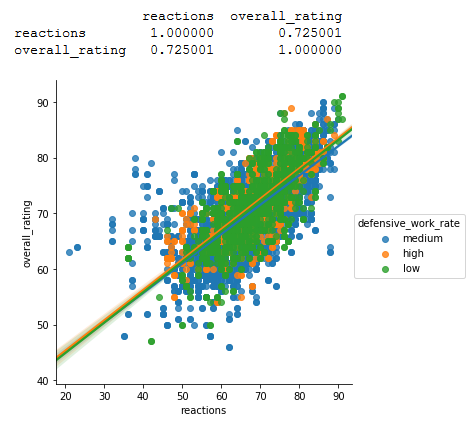


Figure 21. Scatter Plot of Reactions Attribute Against Overall Rating   
with Hue by Defensive Work Rate and Attacking Work Rate

Since there are two distinct subgroups in the dataset of players, the correlation of the reactions attribute with overall rating is further evaluated in each subgroup of goalkeepers and non\_goalkeepers. Tables 3 and 4 list the correlation coefficient between overall rating and all numeric attributes for the goalkeeper subgroup and non\_goalkeeper subgroup, respectively. Highlighted numbers indicate strong positive linear 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 |  |  |

Figure 22 shows that the correlation in the goalkeeper subgroup (Left) is only moderate, while the correlation is strong for the rest of the players (Right).

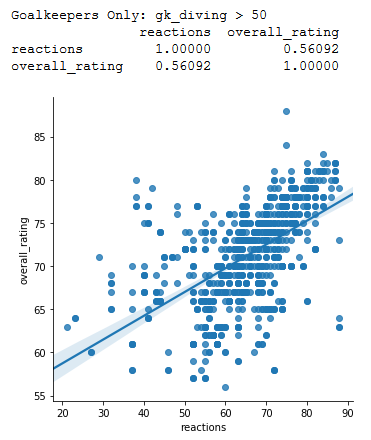
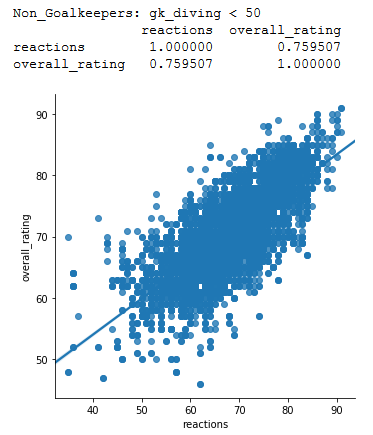
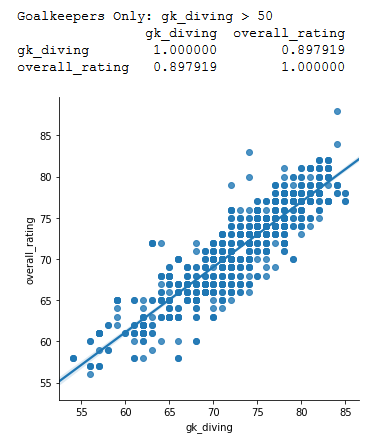
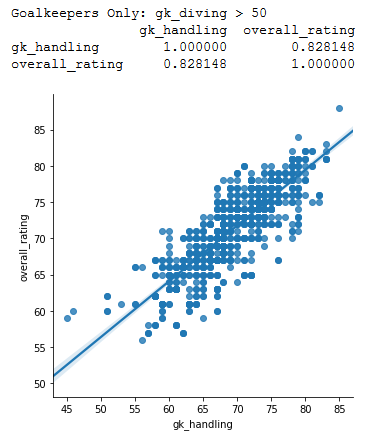
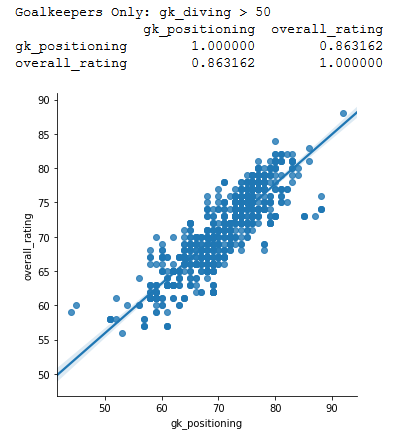
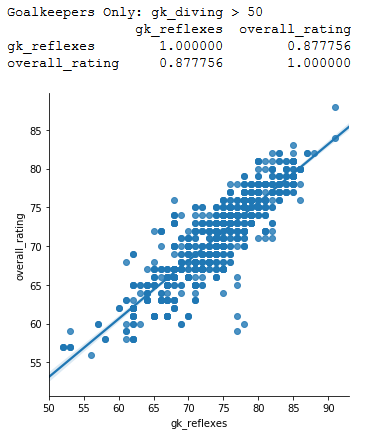
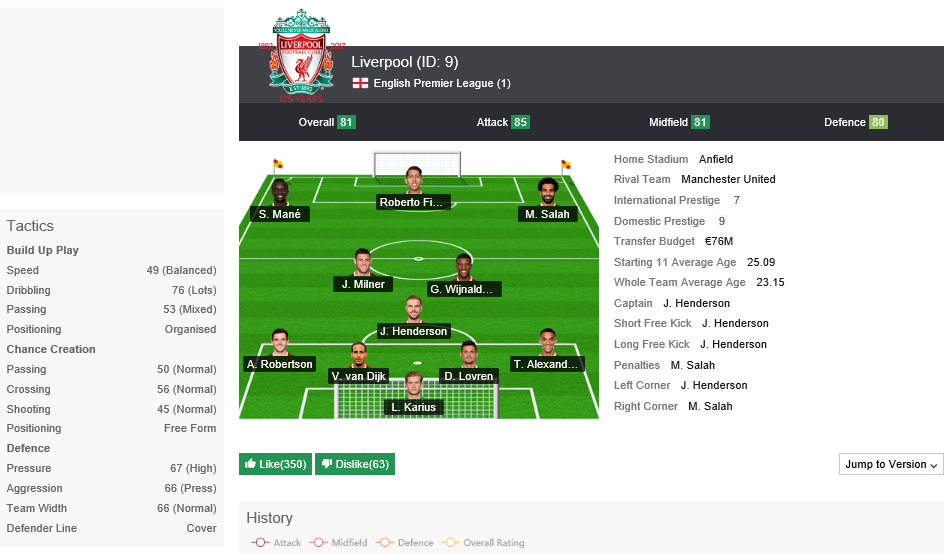
 

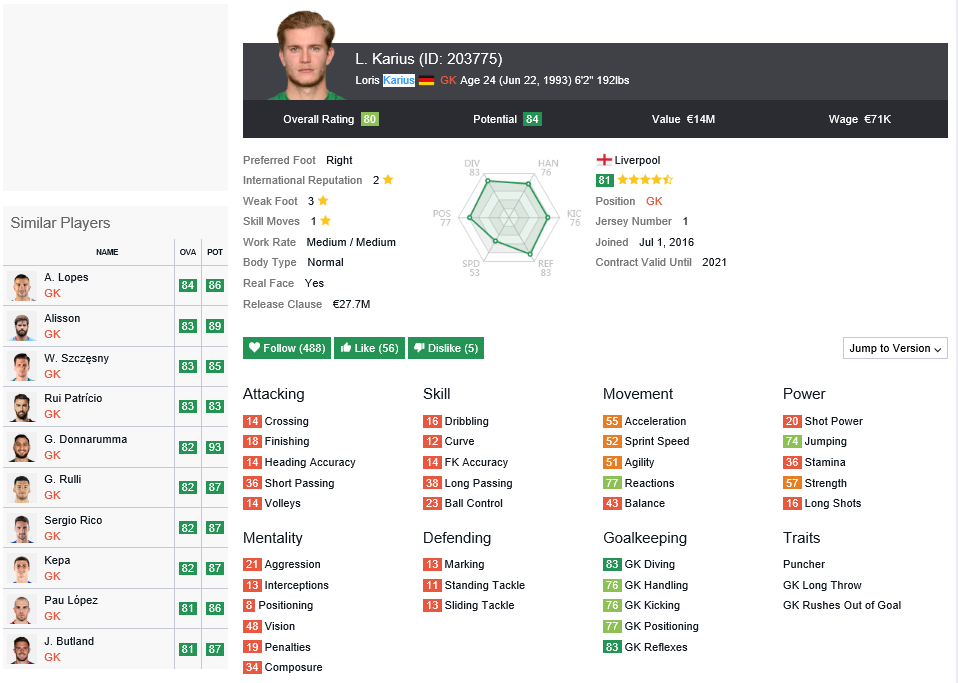
Figure 22. Scatter Plot of Reactions Attribute Against Overall Rating:  
Goalkeepers Subroup (Left) and Non\_Gaolkeepers Subgroup (Right)

# Conclusion





## Answer to Research Question 1

## Answer to Research Question 2

## Answer to Research Question 3

# Conclusion

In short, analysis of the dataset provides answers to all three research questions. Affluent consumers are the biggest users of electricity. And within the Affluent category, Acorn Type E, Career Climbers, use the most energy. Additional information concerning characteristics and lifestyles of Career Climbers may be obtained from the Acorn User Guide (Reference 1).

People’s career opportunity, family situation, financial obligations and living area and housing styles could very well explain the high energy usage of this group of people. Besides, participants of the Smart Meter pilot study did not acquire the meter at the same time. So, comparing the total usage from among Acorn categories or types may not be comparing apples to apples. Participants who started in the winter months would have logged more usage in the meter than those who started in the summer months the following year. Likewise, those who participated in the early phase of the pilot study will have a greater total usage. Use of more sophisticated programming tools will be able to assist with more precise data analysis. The total number of individuals in Acorn categories and types will certainly affect the total usage for the group. It is better to the calculate the average usage by dividing the total usage by the number of participating household. The number of family members in the household, the region and the type of housing, and so on, may have an impact on energy usage. However, this type of information is not available for the study.

As far as the peak usage in relation the weather in certain months of the year, it was pretty much predicted. Future work to continue this study includes investigating the use of API to collect weather data.

The major weakness of the study is the lack of proficiency in Python packages, such as sqlite3 or SQLAlchemy. These tools will allow queries to be run to gather relational information from different tables. In the case of the current study, information from different data blocks, may be assembled according to Acorn Category, Type, year, month or time of day and analyzed.

# 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