Analysis of "*FIFA - Team* and *Player Data*" Utility for Modelling Real Matches

Authors: Hamid Pirnia, Arijit Mondal, Kartik Tiwari, Herraj Luhano

Introduction & motivation:

As avid football fans and passionate supporters of various teams, our group took an interest in trying to search for how much our beloved FIFA game reflects real life football and its idiosyncrasies. While to the naked eye, ratings bestowed to players and teams in the game do more or less reflect what we see on the television or in stadiums, do statistical analysis techniques of the games statistics tell us the same story? As a result we have set out to utilize 10 statistical techniques to find correlations and make inferences about data utilizing data from the FIFA video game (Fifa releases a new game every year and we are using data from the 2008-2016 versions of the game) and comparing that to how events in football transpired in relation.

Characteristics of Data Used: Source of Data: Kaggle

https://www.kaggle.com/hugomathien/soccer

The dataset found consisted of data on all the players on the FIFA game between 2008-2016 with data on their teams and country of origin. In addition to that, EA Sports (creator of the game), rate individual attributes of the players like ball control, stamina, attacking work rate, defensive work rate, speed based on what they see every year and combine them with factual statistics such as preferred foot, height, weight etc. to create overall ratings for each player and team. The perspective based statistics change every year with the annual game released based on the events and results of the past year.

Data on the teams that each player in the game plays for with ratings formed based on the ratings of the players in the team was found on Kaggle.com. The Database itself consists of +25,000 matches and over 10,000 players. The csv files also include league information for 11 European Leagues (each league unique to a specific country) with individual teams unique to them. This is integral because most of the games that teams play are usually against teams within their leagues with the exception of continental competition where any team from any league can play anyone(the frequency of those matches are far lesser).

This dataset is interesting because, it combines perspective based ratings with factual detailed match events. Each player and team statistics from real life such as goal types, possession, number of corners, number of crosses, fouls, cards etc, have been used by EA sports to cross validate (to a certain extent) the ratings that they have attributed respectively based on their perception.

Note: *An important theme of our analysis is the understanding that soccer consists of different positions and each position has different requirements. Therefore, it means that while a defender may not have the same statistics as that of an attacker, all attributes are not of the same importance to different players in different positions.*

Data Cleanup and Alterations

Physical Raw Data Cleanup:

The original format of the raw data was in SQLite. This was then transformed into a csv format for convenience. Further, the data was filled with blank spaces for the Non-Applicable (NA) spots which were then eliminated utilising the *na.omit()* function. In addition, Excel manipulations were done using the *=VLOOKUP()* function to add player and team names to the teams, rather than keep the original *player_api_id* and *team_api_id* (Used *=VLOOKUP()* to extract the names from *Team.csv* and *Player.csv* to update *teamAttributes.csv* and *playerAttributes.csv* respectively). This helped ease an exhaustive and time consuming process of continually referring to api_id.

Data Alterations for Analysis purposes:

Part of the challenge of this project was to analyse what was found interesting from the topic and quite often it was difficult due to the lack of organization of the data. Hence methods were devised to combine data from different csv files into singular tables/relations that could help us analyze data more directly and efficiently from one source rather than writing more complex code due to sources of data being different. SQL was used through library(sqldf) as a method to combine columns from different relations for convenience. A particular model that was created was that of the overall rating of teams. Due to the data not containing an overall rating, it was difficult to carry out particular analysis with relation to teams. This would restrict the scope of the research therefore a ratio of total goals scored to total goals conceded for each team was used

as a metric instead. It would be slightly inefficient to use this due to the fact a team may win after scoring 8 goals while lose in-spite of just conceding just a goal. However, pertaining to the point of the game, scoring more goals and conceding the least is always a good indicator of a good team. Further, after making the manipulations and creating the model, it was found that FC Barcelona of Spain had the best ratio and it is commonly know that they were the most successful team in terms of trophies and perspective based best style of play. Further, this validates the data set's perception as the highest rated players in the game seemed to belong to the team.

Statistical Analysis:

1) **Regression Tree (Cross Validated (20 Fold))**

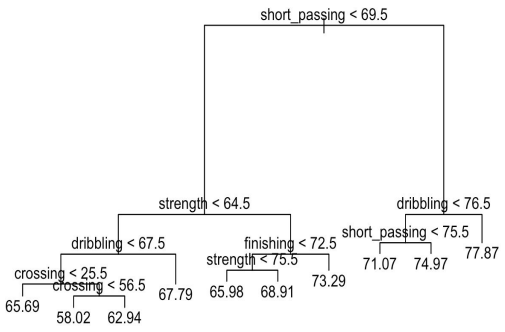
Objective: Researching the impact of all perspective-based attributes assigned by EA sports to players on the overall rating of a player.

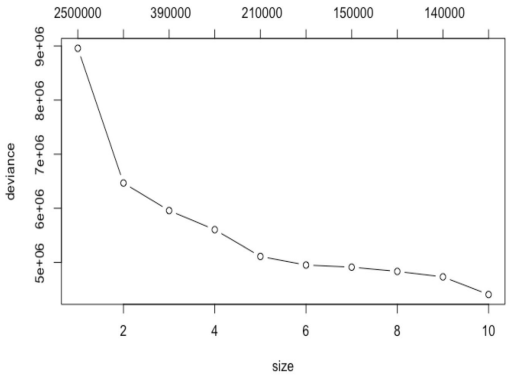
Statistical Technique used: Regression Tree.

Response Variable: Overall Ratings

Category:

Results & Analysis: For every run, the Regression Tree split at the attribute of short passing. As one goes along the Regression Tree, it is evident that the attribute of short passing has a massive impact on overall rating. High MSE: 30.42217. This particular run of cross-validation suggests a 10 terminal node tree





```
Regression tree:
tree(formula = overall_rating ~ (crossing + finishing + short_passing +
    volleys + dribbling + stamina + strength + sprint_speed),
    data = playatt)
Variables actually used in tree construction:
[1] "short_passing" "strength"      "dribbling"     "crossing"      "finishing"
Number of terminal nodes:  10
Residual mean deviance:  24.13 = 4375000 / 181300
Distribution of residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-29.91000  -2.97500   0.02452   0.00000   3.02500  25.31000
[1] 10
```

The tree shows how overall player ratings are affected most by the players' ability for short passing. It then breaks down into respective categories and as it can be seen, the ratings increase from left to right of the tree. However, an inconsistency at the left hand side of this tree where crossing response below 25.5 is greater than crossing response below 56.5 can be noticed. This is also reflective by our MSE which is relatively high. This however can be related to the positions of different players, as not all players require the same attributes to have a similar rating, for example an attacker and a goalkeeper may both have great passing, but an attacker is better at crossing the ball than a goalkeeper.

2) Linear and Multiple Regression

Objective: Finding the correlation between different attributes and overall ratings of players.

Multiple regressions have been done for different positions. Which attributes contribute the best to particular positions for ratings have been explored. Which position the player plays in can be found by doing this which is something that was not provided in the original dataset.

Response Variable: Overall Ratings

Category:

Results & Analysis: Run the 311 project.rmd, for MLR results:

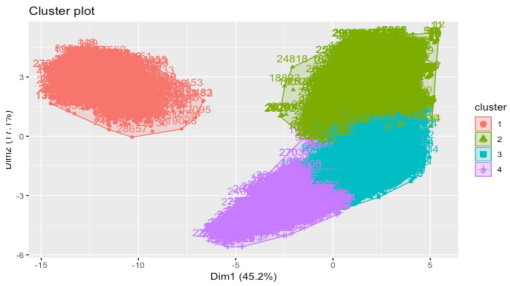
Using backward selection, attributes and columns that most affect the player's overall rating are found. The above results shows the attributes for midfielders, defenders, attackers, and goalkeepers going from top left to bottom right. The average mean squared error for these observations was 6.278. This is relatively low, as the player ratings are out of 99. These qualities also overlap between different playing positions, which assumably causes an increase in error and not all players necessarily play the same style. These attributes helped us create a third version of player attributes.csv file which included player positions. In turn the third version was used to do the K-Means clustering.

3) **K-Means Cluster**

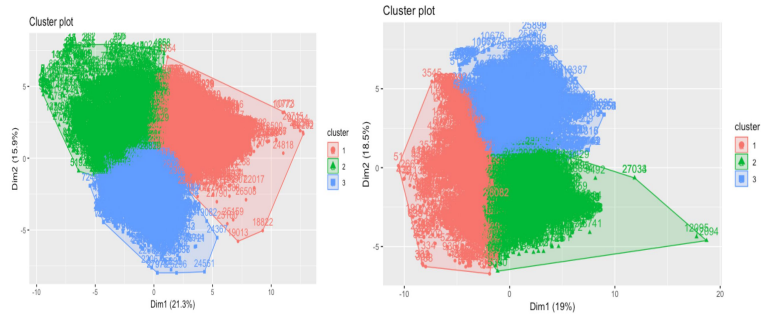
Objective: To find different groups of player positions based on player attributes(positions of players have not been provided on the Data set). Using Player Attributes 3.csv

Category: Clustering

Results & Analysis:

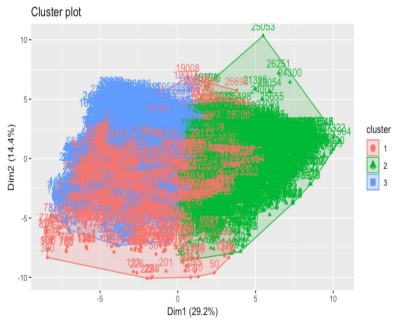


i) *Overall Player Ratings Cluster*: Here Two distinct groups and four groups in total are noticed. The Red Cluster can be deciphered to be that of the goalkeeper position. This is because goalkeepers have the most distinct attributes compared to that of the other positions. The other clusters are that of the attacking, midfield and defensive positions. They overlap slightly because while each position has its distinct attributes which is fairly evident in the clustering, some traits such as speed, stamina, ball control and passing (analysis shown in the multiple regressions previously) overlap between positions.

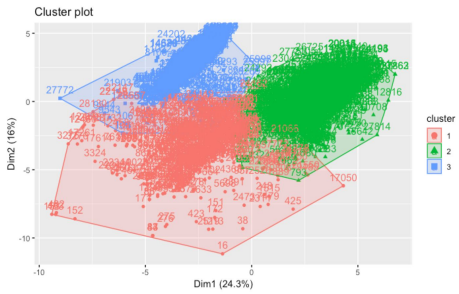


ii) *Attacking Cluster*: Represents different types of Attacking players, eg, Tall Players who rely on attributes such as heading to score more rather than speed merchants who rely on getting behind defenders and outpacing others etc. Shows how EA allows for great flexibility of attributes contributing to attacking ratings.

iii) *Midfield Cluster*. Similar to the attacking clusters. Midfield positions are the ones with most flexibility of attributes and its evident with the distinctness of the clusters. The red cluster's shape appears to be slightly different to the other two clusters for midfield as one of them maybe a defensive midfield group who require the possession of slightly different qualities as compared to other types of midfielders.



iv) *Defensive cluster*: Represents different types of Defensive players. It is fairly noticeable that most of these clusters are very overlapping and not very distinct for each other as compared to the attacking player cluster. This is because while it is easy to find multiple different ways to attack and score goals, the art of defending is fairly a one way street with slight room for flexibility. There maybe some outliers with defensive players who are genuinely better at passing but this cluster shows how the creators of the video game perceive the defensive ratings and account for lesser player attributes for these positions than the attacking positions.



v) *Goalkeeping luster*: Shows the different types of goalkeepers. They have a major overlap of attributes as basic duty of goalkeepers to make saves but are also distinct because of the different types of goalkeepers they are such as ball playing, or tradition long ball kicking.

4) **Principal Component Analysis/Linear Regression/ KNN**

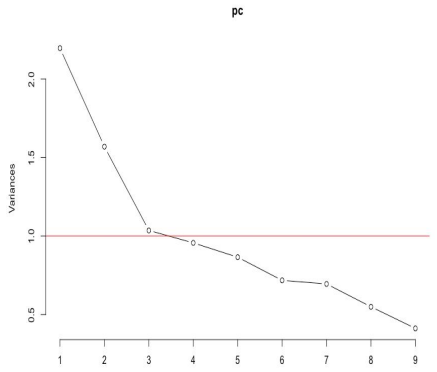
Objective: Study the data in Team Attributes and develop a predictive model for Goal Ratios.

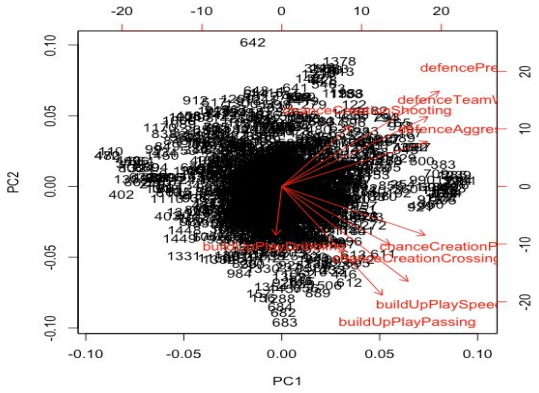
Response Variable: PC, PCA1, PCA2

Category: Regression and Dimension Reduction

Results & Analysis:

Though only 53.36% of variance is accounted by picking only the first 3 PC's, Kaiser Criterion and the scree plot suggest to do so (Top left diagram). In the bi-plot below we see the influence of scaled attributes on PC1 and PC2 and the rotations summary below confirms our observations. PC1 describes teams with all-round higher metrics whereas PC2 describes teams with weak build-up and chance creation metrics but keeps up the defensive metrics. Finally, PC3 describes teams opposite PC2, lower defensive metrics however better chance creation and build-up metrics. We pick two of these three PC's to perform a linear regression with cross validation (Since Cross-Validation suggests that 2 PC's have the least MSE).





Linear Model MSE (Cross validated): 0.5159629

KNN Model MSE (Cross validated): 0.0196

The response variable was the goal ratios of real teams from real games. The accuracy of the system demonstrates the reliability of FIFA game data for development of predictive models for the real game as most team_attributes predictors were from the game.

5) Quadratic Discriminant Analysis (Cross-Validated)

Objective: Prediction of team width classes (Narrow, wide, normal) based on defense Pressure and Aggression.

Response Variable: Team width class

Category: Classification

Results & Analysis:

```
> ct <- table(defenceTeamWidthClass, tdwqda$class)
> diag(prop.table(ct, 1))
    Narrow     Normal       Wide
0.09836066 0.96967341 0.49549550
> ct

defenceTeamWidthClass Narrow Normal Wide
               Narrow      6     54    1
               Normal     11   1247   28
               Wide        0     56   55
> # Misclassification Rate from built in CV (jacknife LOOCV)
> mis.cv<- 1- sum(diag(prop.table(ct)))
> mis.cv
[1] 0.1028807
>
> F1_Score(as.numeric(tdwqda$class),as.numeric(defenceTeamWidthClass))
[1] 0.1538462
>
> LogLoss <- function(pred, res){
+   #From https://rstudio-pubs-static.s3.amazonaws.com/157427_74913a13c3254d128bc69937434fbfa8.html
+   (-1/length(pred)) * sum (res * log(pred) + (1-res)*log(1-pred))
+ }
>
> #Log loss score
> LogLoss(abs((as.numeric(tdwqda$class)-0.01))/sum(as.numeric(tdwqda$class)),
+         as.numeric(defenceTeamWidthClass)/sum(as.numeric(defenceTeamWidthClass)))
[1] 0.005681369
```

The results showcase how most teams are predicted and found to have normal width class based on team aggression and pressure. It makes sense considering it is the most balanced approach. Soccer teams have opposition analysis teams today and prediction of other systems relies heavily on a very perspective based approach on aggression and pressure so it would be riskier to predict lesser balanced defence systems on other teams for one's own team. The balanced nature of the width class makes teams use it consistently and hence its own prediction and true nature makes sense. (1286 predicted Normal width classes and 1247 which are actually true).

Diagnostics: F1 score: 0.153, Log-Loss:0.00568(relatively low), Misclassification rate~10%

6) Hierarchical Clustering

Objective: Group teams with similar playing tactics by splitting the tactics in three major-groups: Buildup play, Chance creation, Defence strategy

Response Variable: N/A

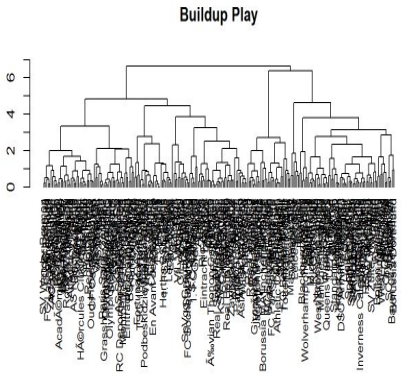
Category: Clustering

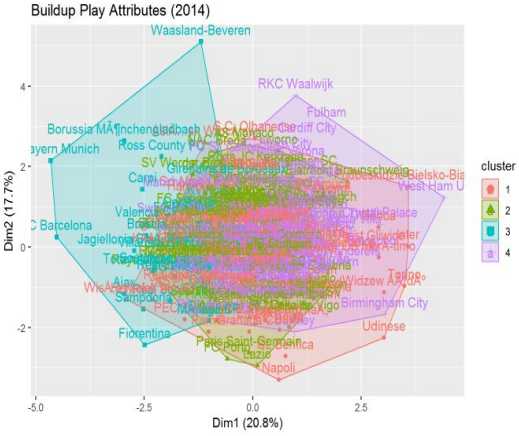
Methodology: R-Studio (Libraries used: factoextra, dplyr)

Results & Analysis:

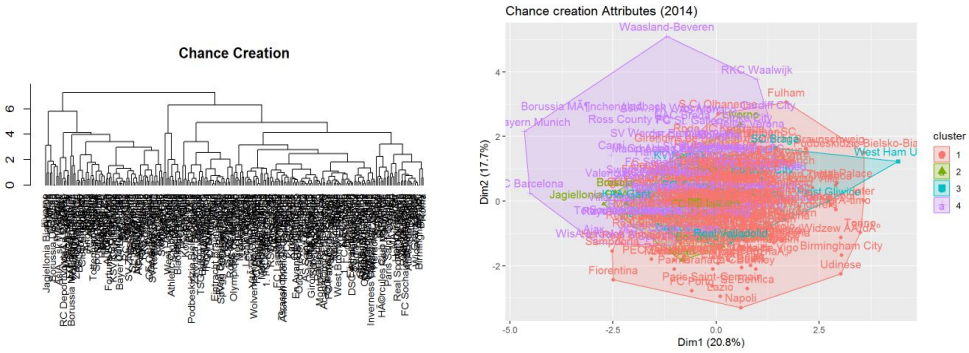
This was really good analysis because it provides an insight on how the soccer is played by European club teams. The way the teams play in Europe differs when compared to other continents but when looked at the tactics closely, all the teams around the globe use the same fundamental tactics with some tweaks of their own.

This is why Hierarchical Clustering was chosen as an analysis for this clustering problem. As seen below, the plots show major groups and then the smaller sub-groups. The major groups here for each plot, tells us about the fundamental tactic group the teams fall under. The sub-groups tell us how those teams have implemented or tweaked those tactics according to their liking.The research demands to know which major tactic are employed by teams and this analysis does a great job of dividing them into the major groups and the corresponding sub-groups.

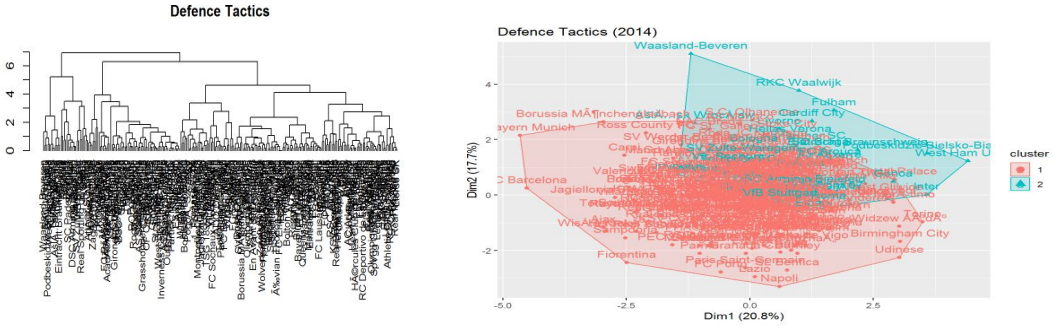




Buildup Play: These tactics correspond to when a team has the ball possession and the style or techniques they use to progress further up on the pitch, into the opponent's half. The dendrogram shows two major groups and within those, two major sub-groups. The two major groups here are: Slow Build up, Fast build up. The sub-groups tell a story about the Passing type(short or long) and Positioning type(fixed, free-form). The cluster plot shows four clusters, but it can clearly be seen that two major groups exist.



Chance Creation tactics: These tactics refer to the techniques used by a team to create opportunities to get a shot on target and score a goal. Its seen how, there are two major groups in both plots. These two groups represent Risky chance creation, or Safe chance creation. To provide an overview, risky chance creation involves long-passes, running into spaces behind defenders, taking long-shots e.t.c. With Risky chance creation, a team has a higher probability of losing the ball. The Safe chance creation is the exact opposite of risky tactics.



Defence Tactics: These are also categorized in two major groups, namely High-Press/Offside Trap and Deep/Cover. High-press tactics involve pressing the opponent team higher up the pitch in their half, and Deep defending involves the defending team to sit back and defend close to their own goal.

To conclude this analysis, I would like you to take note of two teams: FC Barcelona (Spain) and Bayern Munich (Germany). These teams can be seen in the same cluster for all three tactics: Buildup (cluster 3), Chance creation (cluster 4) and Defence (cluster 1). The fact that

these two teams have almost identical playing styles solidifies the reliability of this analysis. FC Barcelona is known for the possession based short passing style. In 2014, Bayern Munich hired a new coach who is a former Barcelona player and coach. Upon his arrival, he implemented the same tactics for his new team, Bayern Munich. Hence, both teams being in the same cluster.

7) Random Forest to Perform Bagging

Objective: Exploring the relationship between sprint speed as the response variable and how its affected by acceleration, dribbling and stamina (cornerstones of attacking football). A comparison of MSE to a cross validated (20 fold) regression tree has also been done.

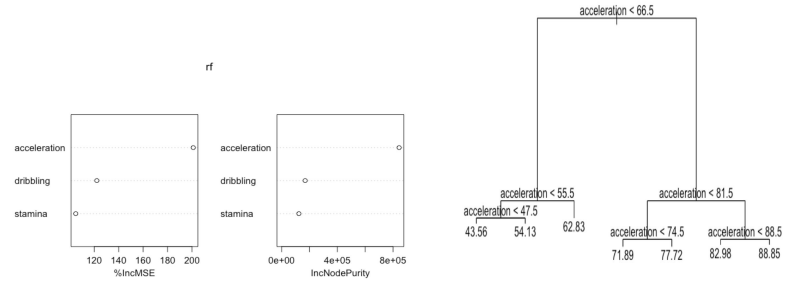
Response Variable: Sprint Speed

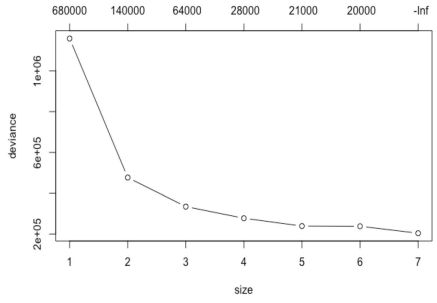
Category: Random Forest

Results & Analysis: The results rank acceleration as the most influential pertaining to on the ball and the off the ball sprinting (Some assume that off the ball sprinting would be more affected by dribbling). The MSE is ~ 8.124. The regression tree shows similar results.

```
Call:
 randomForest(formula = sprint_speed ~ (acceleration + stamina +      dribbling), data = play22, mtry = 2, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 8.12452
                    % Var explained: 94.7
```





This particular run of cross-validation suggests a 7 terminal node tree

Drawbacks and errors:

- H-Clust for different leagues was not done due to inexperience with data wrangling.
- Team rating was only a ratio of goals scored to conceded.
- Match.csv was not used much.
- Match outcomes were not predicted due to time constraints.
- Had to determine the positions ourselves due to which the K-Clusters may contain higher error than if it was provided by EA Sports.
- The team attributes and team csv files did not consist of players therefore it was not possible to relate the player and team data and carry out analysis in relation.

Conclusion:

In the beginning of the project, our group set out to search how accurately the Fifa game represented the real life football landscape. We faced multiple obstacles in our search for this correlation. This began with data cleanup and wrangling. We progressed on to carry out the analysis starting with a regression tree to see what predictors affected the overall player ratings. This was followed by carrying out multiple linear regression which helped us figure out the game positions for each player and carry out the K means clusters. Using the goal ratios, we figured out team ratings and carried out the H-Cluster analysis, in this we could see different attributes of the team affecting the team rating and we were successfully able to identify patterns about soccer in real life which had gone unnoticed by us. QDA was performed to make and inference between defensive width and defensive pressure/aggression, this was intended to provide us valuable intelligence that could be useful when planning a game strategy and player positions. QDA proved to be a success with a misclassification rate of only 10%. To develop a strong predictive model for the teams' goals ratios we resided to use PCA for dimension reduction. After we picked the PC's to use in our LDA we got a satisfying model. To conclude, Fifa ratings are relatively accurate and do provide very reliable data. This data was refined and analyzed by us in turn increasing our understanding of football, Fifa metrics and how to make more data driven decisions in the area. This project has taught us various ways of determining results and providing analysis of different datasets and has improved our knowledge of RStudio, SQL, and Excel and Statistical Learning.