An Analysis of Virtual Player Data from FIFA 21

University of Louisiana at Lafayette
INFX 512

Dr. Mehmet Tozal

Damian O'Boyle
(C00481724)

30 April 2020

Table of Contents

Dataset

Description

This dataset is derived from the popular football simulation video game FIFA 21 which is available to users on both console and PC. The data contained within is used within the game itself to provide statistical attributes for each individual player. These attributes determine the individual player's usefulness and skill level within the game and cover a broad range of abilities and skill sets from prowess at heading to simple competence at passing of the ball. It also includes different overall ratings for each facet of the game like defending and attacking.

There are a total of 106 different and unique variable attributes (columns) in this dataset, offering insight into the 18,944 individual unique players (rows).

Data Origin

The data was first discovered and accessed on the statistical dataset website Kaggle, on the 16th of March, 2021. The data itself was scraped from the website sofifa.com which is updated in real time with data directly from the FIFA servers typically used to update elements of the game itself. The exact version of the dataset used here was provided by Stefano Leone, who originally uploaded the dataset to Kaggle on 9 October, 2020. Coinciding with the physical release of the game, which took place three days earlier on 6 October, 2020. The dataset was downloaded in .CSV format and cleaned for use in R on 15 April 2021.

Data Cleaning

The original dataset consisted of a total of 106 different variables. This total was obviously far beyond the requirements of this project, not to mention the fact that some of the variables were redundant in terms of their analysis here. Many of these redundant variables have been integrated into other higher order variables in specific attribute categories. Cleaning was completed using Microsoft Excel and mainly involved deleting many of the variables deemed unnecessary to conduct substantial statistical analysis. Many variables were also renamed to provide ease of use within R. A selection of variables were altered in form or amalgamated to create terms more better equipped for statistical use. The final configuration reduced the original 106 variables down to just 34, representing the same 18,944 observations.

The author didn't provide any detailed description of the different variables, but was kind enough to alter the variable names from those provided directly from the originally scrapped source in order to make them more easily understood.

Variable Descriptions

Name	Mode	Description
name	character	The player's name as it appears in the game
age	integer	The age of the player measured in integer years
height	integer	The height of the player measured in centimetres
weight	integer	The weight of the player measured in kilograms
nationality	factor	The country for which the player has declared to play for the national team in international competition
club	factor	The full name of the club to which the player is contracted

league	factor	The full name of the league in which the player's club participates
overall	integer	A rating between 1-100 signifying how good the player is currently
potential	integer	A rating between 1-100 denoting the overall rating a player is likely to achieve
value	integer	How much the player is valued at on the transfer market in euro
wage	integer	How much the player is paid per week in euro
position	factor	The players position on the pitch
foot	factor	The player's preferred foot (Right, Left)
weakfoot	integer	A value between 1 and 5 indicating a player's prowess with their weak foot
reputation	integer	A value between 1 and 5 indicating a player's reputation within their national team
attack_wr	factor	The player's work rate when on attack (High, Medium, Low)
defend_wr	factor	The player's work rate when on defence (High, Medium, Low)
clause	integer	The cost to the club to release the player from their current contract
contract	integer	The year in which the player's current contract expires
loan	factor	Whether or not the player is currently on loan at a different club
jersey	integer	The number that the player wears on their jersey
pace	integer	A rating between 1-100 of a player's pace
shooting	integer	A rating between 1-100 of a player's shooting ability
passing	integer	A rating between 1-100 of a player's passing ability
dribbling	integer	A rating between 1-100 of a player's dribbling ability
defending	integer	A rating between 1-100 of a player's defending ability
physical	integer	A rating between 1-100 of a player's physicality
attacking	numeric	A rating between 1-100 of a player's attacking ability
skill	numeric	A rating between 1-100 of a player's skill, averaged from five specific skill related abilities
movement	numeric	A rating between 1-100 of a player's movement ability, averaged from five specific movement related abilities
power	numeric	A rating between 1-100 of a player's power, averaged from five specific power related abilities
mentality	numeric	A rating between 1-100 of a player's mentality, averaged from six specific mentality related abilities
tackling	numeric	A rating between 1-100 of a player's tackling ability, averaged from two specific tackling related abilities
goalkeeping	numeric	A rating between 1-100 of a player's goalkeeping ability, averaged from five specific goalkeeping related abilities

Data Loading

The cleaned dataset was loaded into R and structured using the following commands;

```
> dataset <- read.csv("C:\\...\\FIFA21.csv")
```

The structure of the dataset is shown;

```
> str(dataset)
```

```
'data.frame':  18944 obs. of  34 variables:
$ name          : chr  "L. Messi" "Cristiano Ronaldo" "J. Oblak"...
$ age           : int  33 35 27 31 28 29 21 28 28 27 ...
$ height        : int  170 187 188 184 175 181 178 187 193 191 ...
$ weight        : int  72 83 87 80 68 70 73 85 92 91 ...
$ nationality   : chr  "Argentina" "Portugal" "Slovenia" ...
$ club          : chr  "FC Barcelona" "Juventus" ...
$ league        : chr  "Spain Primera Division" ...
$ overall       : int  93 92 91 91 91 91 90 90 90 90 ...
$ potential     : int  93 92 93 91 91 91 95 93 91 91 ...
$ value         : int  67500000 46000000 75000000 80000000 ...
$ wage          : int  560000 220000 125000 240000 270000 ...
$ position      : chr  "RW" "ST" "GK" "ST" ...
$ foot          : chr  "Left" "Right" "Right" "Right" ...
$ reputation    : int  5 5 3 4 5 4 3 3 3 3 ...
$ weakfoot      : int  4 4 3 4 5 5 4 4 3 3 ...
$ attack_wr     : chr  "Medium" "High" "Medium" "High" ...
$ defend_wr     : chr  "Low" "Low" "Medium" "Medium" ...
$ clause        : int  138400000 75900000 159400000 132000000 ...
$ jersey        : int  10 7 13 9 10 17 7 1 4 1 ...
$ loan          : chr  "No" "No" "No" "No" ...
$ contract      : int  2021 2022 2023 2023 2022 2023 2022 2022 ...
$ pace          : int  85 89 NA 78 91 76 96 NA 76 NA ...
$ shooting      : int  92 93 NA 91 85 86 86 NA 60 NA ...
$ passing       : int  91 81 NA 78 86 93 78 NA 71 NA ...
$ dribbling     : int  95 89 NA 85 94 88 91 NA 71 NA ...
$ defending     : int  38 35 NA 43 36 64 39 NA 91 NA ...
$ physical      : int  65 77 NA 82 59 78 76 NA 86 NA ...
$ attacking     : num  85.8 87.4 19 84.6 81.6 81.4 81.6 23.6 ...
$ skill         : num  94 82.8 21.8 81.4 89.6 88.2 78.8 28.8 ...
$ movement      : num  90.2 86.2 61.4 81.4 90.6 79.6 91.6 50.8 ...
$ power         : num  77.8 88.8 53.6 84 71.4 81.6 80.8 53.6 ...
$ mentality     : num  73.8 74.7 34.7 79.8 74.8 83.2 70.8 40.2 ...
$ tackling      : num  29.5 28 15 30.5 29.5 59 33 11.5 89.5 ...
$ goalkeeping   : num  10.8 11.6 87.4 10.2 11.8 11.2 8.4 87.8 ...
```

The highlighted attributes need to be converted to factor data types. The follow commands were used to achieve this;

```
> dataset$nationality <- as.factor(dataset$nationality)
> dataset$club <- as.factor(dataset$club)
> dataset$league <- as.factor(dataset$league)
> dataset$position <- as.factor(dataset$position)
> dataset$foot <- as.factor(dataset$foot)
> dataset$reputation <- as.factor(dataset$reputation)
> dataset$weakfoot <- as.factor(dataset$weakfoot)
> dataset$attack_wr <- as.factor(dataset$attack_wr)
> dataset$defend_wr <- as.factor(dataset$defend_wr)
> dataset$jersey <- as.factor(dataset$jersey)
> dataset$loan <- as.factor(dataset$loan)
```

The head command was used as follows to display the first six columns of the dataset.

```
> head(dataset)
```

```
              name age height weight nationality                club
1          L. Messi  33    170     72   Argentina         FC Barcelona
2 Cristiano Ronaldo  35    187     83    Portugal             Juventus
3          J. Oblak  27    188     87    Slovenia      Atlético Madrid
4    R. Lewandowski  31    184     80      Poland    FC Bayern München
5         Neymar Jr  28    175     68      Brazil Paris Saint-Germain
6      K. De Bruyne  29    181     70     Belgium      Manchester City

                  league overall potential    value   wage position
1 Spain Primera Division      93        93 67500000 560000       RW
2        Italian Serie A      92        92 46000000 220000       ST
3 Spain Primera Division      91        93 75000000 125000       GK
4   German 1. Bundesliga      91        91 80000000 240000       ST
5         French Ligue 1      91        91 90000000 270000       LW
6 English Premier League      91        91 87000000 370000      CAM

   foot reputation weakfoot attack_wr defend_wr    clause jersey
1  Left          5        4    Medium       Low 138400000     10
2 Right          5        4      High       Low  75900000      7
3 Right          3        3    Medium    Medium 159400000     13
4 Right          4        4      High    Medium 132000000      9
5 Right          5        5      High    Medium 166500000     10
6 Right          4        5      High      High 161000000     17

  loan contract pace shooting passing dribbling defending physical
1   No     2021   85       92      91        95        38       65
2   No     2022   89       93      81        89        35       77
3   No     2023   NA       NA      NA        NA        NA       NA
4   No     2023   78       91      78        85        43       82
5   No     2022   91       85      86        94        36       59
6   No     2023   76       86      93        88        64       78

  attacking skill movement power mentality tackling goalkeeping
1      85.8  94.0     90.2  77.8      73.8     29.5        10.8
2      87.4  82.8     86.2  88.8      74.7     28.0        11.6
3      19.0  21.8     61.4  53.6      34.7     15.0        87.4
4      84.6  81.4     81.4  84.0      79.8     30.5        10.2
5      81.6  89.6     90.6  71.4      74.8     29.5        11.8
6      81.4  88.2     79.6  81.6      83.2     59.0        11.2
```

Expectations

First and foremost it is presumed that players with higher ratings in the different available attribute areas will perform better than the lesser rated players in game and thus achieve and maintain a higher transfer value and overall rating/rank within the game.

It will be interesting to determine what effects different variables have on each other exactly, for example are players paid more depending on a particular league or is this value purely determined by ability/overall ratings or even based simply on transfer value.

How do physical characteristics play into this such as age, height and weight compared to a player's nationality or the reputation within that nation? Can it be seen that the intangible abilities that have been quantified for game play purposes have a greater bearing on these value metrics.

The first expectation is that there will be correlations between a player's fiscal attributes, transfer value, weekly wage, release clause and their overall and potential ratings within the game. Coupled with this it is expected that the player's specific attribute ratings will correlate with the overall and potential rating as well as the fiscal attributes to differing degrees of significance.

A normal distribution should appear in the dataset population in regards to some attributes somewhere. Possibly related to the recorded physical attributes height, weight etc.

It is suspected that certain player positions might disproportionately account for specific ranges within variable densities. For example do attacking players get paid more than defenders or goalkeepers?

The majority of players are expected to have high to medium workrates in both the attacking and defending categories. This assumption is based purely on the fact that these players are professionals and get paid very well to play a game they are passionate about and have likely worked their entire lives to get to where they are today.

International reputation scores should account for higher overall ratings and values/wages.

It is expected that european nations of origin will account for the bulk of the players that exist within the game. Football is played more competitively in Europe than anywhere else on earth, this should be reflected in the games dataset as this is also where the majority of players of this game are typically from.

Some exploratory regression analysis would be interesting to undertake and observe. Perhaps trying to determine if left or right footed players are better on average? Exploring some interaction effects between correlated variables would return so statistically relevant information.

In terms of predictive analysis, the obvious metrics to attempt to predict would be the overall rating and value variables. For Overall rating the specific skill attributes would serve as very good predictors. Value may be a grader variable to predict or at least select predictors for.

Analysis
Numeric Variables

Correlations

The first task as with any new dataset was to determine if there were any correlations between the numeric variables of which there were quite a few. As initially indicated in the expectations higher overall player ratings are expected to be correlated to higher wage, transfer value and release clause values. It will be interesting to see what other less obvious correlations appear from this initial assessment.

```
> cor(dataset[c(2:4, 8:11, 18, 22:34)], use = "complete.obs")
```

	age	height	weight	overall	potential
age	1.00000000	0.075203596	0.22671625	0.47149861	-0.267886568
height	0.07520360	1.000000000	0.75720168	0.04849949	0.001548314
weight	0.22671625	0.757201685	1.00000000	0.15931314	-0.015523715
overall	0.47149861	0.048499489	0.15931314	1.00000000	0.632467597
potential	-0.26788657	0.001548314	-0.01552371	0.63246760	1.000000000
value	0.09242522	0.014177136	0.05273066	0.64223759	0.580923019
wage	0.16626319	0.036761770	0.07435097	0.58429922	0.481815671
clause	0.07395582	0.013015793	0.04773722	0.63326227	0.588536006
pace	-0.18083819	-0.410827378	-0.36869250	0.20265353	0.272341933
shooting	0.24489689	-0.194649884	-0.08916546	0.49137541	0.291248896
passing	0.34181796	-0.269299988	-0.16775267	0.71519743	0.446621643
dribbling	0.18719986	-0.378459100	-0.27513381	0.64408556	0.483593052
defending	0.25953859	0.213207073	0.21151784	0.36223599	0.178301237
physical	0.44632173	0.489759277	0.57758819	0.53385147	0.167945319
attacking	0.37470383	-0.129494212	-0.02018391	0.72095425	0.421935618
skill	0.30634589	-0.294074443	-0.18985463	0.67224275	0.432908292
movement	-0.01562652	-0.575180492	-0.47833198	0.39677656	0.360495579
power	0.45888371	0.123331504	0.25835720	0.70648239	0.325481585
mentality	0.49153289	-0.042820327	0.06521251	0.80751510	0.441011096
tackling	0.18371437	0.168073496	0.15652988	0.27800784	0.151713209
goalkeeping	0.15433699	0.016576023	0.04228228	0.08910300	-0.051594199

	value	wage	clause	pace	shooting
age	0.09242522	0.16626319	0.07395582	-0.18083819	0.24489689
height	0.01417714	0.03676177	0.01301579	-0.41082738	-0.19464988
weight	0.05273066	0.07435097	0.04773722	-0.36869250	-0.08916546
overall	0.64223759	0.58429922	0.63326227	0.20265353	0.49137541
potential	0.58092302	0.48181567	0.58853601	0.27234193	0.29124890
value	1.00000000	0.83948620	0.99391476	0.20527526	0.34394287
wage	0.83948620	1.00000000	0.83509684	0.13122826	0.30709531
clause	0.99391476	0.83509684	1.00000000	0.20505497	0.33623019
pace	0.20527526	0.13122826	0.20505497	1.00000000	0.35172007
shooting	0.34394287	0.30709531	0.33623019	0.35172007	1.00000000
passing	0.46724217	0.42674548	0.45920337	0.29546587	0.65861160
dribbling	0.45091305	0.39587166	0.44530649	0.54068874	0.77118786
defending	0.17351631	0.18172443	0.17047776	-0.28756807	-0.39599738
physical	0.25146008	0.24269726	0.24163902	-0.18141628	0.02384569
attacking	0.47468119	0.43734232	0.46479989	0.32095834	0.89501111
skill	0.44752398	0.40790256	0.43984596	0.35127194	0.75909588
movement	0.32388523	0.25201274	0.32047277	0.88653845	0.51116280
power	0.40652947	0.37360177	0.39414849	0.17640332	0.67774405
mentality	0.50262138	0.47697372	0.49155269	0.12285654	0.59383150
tackling	0.12869685	0.13759846	0.12724125	-0.27154931	-0.44911450
goalkeeping	0.01577790	0.02305929	0.01490433	-0.01801744	0.06039936

	passing	dribbling	defending	physical
age	0.34181796	0.1871998593	0.25953859	0.4463217267
height	-0.26929999	-0.3784590996	0.21320707	0.4897592765
weight	-0.16775267	-0.2751338135	0.21151784	0.5775881906
overall	0.71519743	0.6440855611	0.36223599	0.5338514714
potential	0.44662164	0.4835930518	0.17830124	0.1679453194
value	0.46724217	0.4509130529	0.17351631	0.2514600809
wage	0.42674548	0.3958716570	0.18172443	0.2426972606
clause	0.45920337	0.4453064857	0.17047776	0.2416390201
pace	0.29546587	0.5406887386	-0.28756807	-0.1814162795
shooting	0.65861160	0.7711878626	-0.39599738	0.0238456908
passing	1.00000000	0.8364507848	0.17575910	0.1711094299
dribbling	0.83645078	1.0000000000	-0.13793583	0.0002765352
defending	0.17575910	-0.1379358304	1.00000000	0.5568896045
physical	0.17110943	0.0002765352	0.55688960	1.0000000000
attacking	0.82239399	0.8251498990	-0.11066260	0.2116214639
skill	0.94288760	0.8957529846	0.01922348	0.0924080282
movement	0.54792069	0.7567337057	-0.20078251	-0.1587053469
power	0.59120229	0.5450401515	0.14890302	0.6717616416
mentality	0.82861726	0.6775489438	0.39519025	0.5159587763
tackling	0.13711810	-0.1699501762	0.97640266	0.4766625715
goalkeeping	0.06933206	0.0390423087	0.04463926	0.0900871384

	attacking	skill	movement	power	mentality
age	0.37470383	0.30634589	-0.015626523	0.45888371	0.49153289
height	-0.12949421	-0.29407444	-0.575180492	0.12333150	-0.04282033
weight	-0.02018391	-0.18985463	-0.478331978	0.25835720	0.06521251
overall	0.72095425	0.67224275	0.396776556	0.70648239	0.80751510
potential	0.42193562	0.43290829	0.360495579	0.32548158	0.44101110
value	0.47468119	0.44752398	0.323885226	0.40652947	0.50262138
wage	0.43734232	0.40790256	0.252012737	0.37360177	0.47697372
clause	0.46479989	0.43984596	0.320472769	0.39414849	0.49155269
pace	0.32095834	0.35127194	0.886538451	0.17640332	0.12285654
shooting	0.89501111	0.75909588	0.511162804	0.67774405	0.59383150
passing	0.82239399	0.94288760	0.547920695	0.59120229	0.82861726
dribbling	0.82514990	0.89575298	0.756733706	0.54504015	0.67754894
defending	-0.11066260	0.01922348	-0.200782511	0.14890302	0.39519025
physical	0.21162146	0.09240803	-0.158705347	0.67176164	0.51595878
attacking	1.00000000	0.84756376	0.518368663	0.73378378	0.76516555
skill	0.84756376	1.00000000	0.593189894	0.60448375	0.78176840
movement	0.51836866	0.59318989	1.000000000	0.31820785	0.37263045
power	0.73378378	0.60448375	0.318207845	1.00000000	0.74280375
mentality	0.76516555	0.78176840	0.372630449	0.74280375	1.00000000
tackling	-0.18154018	-0.01723227	-0.201655086	0.06097070	0.31956126
goalkeeping	0.07424141	0.05458311	0.007472144	0.09600512	0.09500852

	tackling	goalkeeping
age	0.18371437	0.154336995
height	0.16807350	0.016576023
weight	0.15652988	0.042282283
overall	0.27800784	0.089103004
potential	0.15171321	-0.051594199
value	0.12869685	0.015777905
wage	0.13759846	0.023059293
clause	0.12724125	0.014904332
pace	-0.27154931	-0.018017444
shooting	-0.44911450	0.060399365
passing	0.13711810	0.069332061
dribbling	-0.16995018	0.039042309
defending	0.97640266	0.044639256
physical	0.47666257	0.090087138

	tackling	goalkeeping
attacking	-0.18154018	0.074241409
skill	-0.01723227	0.054583110
movement	-0.20165509	0.007472144
power	0.06097070	0.096005121
mentality	0.31956126	0.095008519
tackling	1.00000000	0.027272511
goalkeeping	0.02727251	1.000000000

The values with a level of significance higher than 0.7 have been highlighted for easy viewing (Green: greater than ±0.9, yellow: greater than ±0.8 and orange: greater than ±0.7).

Three highly significant values above the 0.9 threshold were observed, namely;

value/clause	0.99391476
defending/tackling	0.97640266
skill/passing	0.94288760

Above the 0.8 threshold there were eleven significant correlations observed;

skill/dribbling	0.8957529846
attacking/shooting	0.89501111
movement/pace	0.88653845
attacking/skill	0.84756376
wage/value	0.83948620
passing/dribbling	0.83645078
wage/clause	0.83509684
passing/mentality	0.82861726
attacking/dribbling	0.8251498990
attacking/passing	0.82239399
overall/mentality	0.80751510

A similar selection of correlations were observed for values calculated above 0.7.

There were a selection of variables that achieved five or more threshold correlations with other variables in the dataset. These were, passing, dribbling, skill, mentality and attacking which achieved seven threshold correlations. These variables along with the other highly significant values found here, will prove useful when conducting more in-depth regression analysis on the dataset and its variables throughout the project.

The `use = "complete.obs"` instruction had to be used in this call because the variables clause, pace, shooting, passing, dribbling, defending and physical all contain NA values specifically for players in the GK (Goalkeeper) position, except in the case of clause, which had its own NA's. These were base statistical values presented in the dataset and goalkeeper versions of some of these metrics were presented but deleted and/or merged into a higher order variable due to their sparse use compared to the other variables.

Interestingly the overall rating variable didn't achieve any threshold correlations with the three fiscal variables, value, wage and clause as was expected. Nor did it have any correlations in the upper tier with any other variable in the dataset. In fact there were only three correlations above the 0.7 threshold set.

Further comment will be made and testing undertaken after these variables correlation coefficients are visualised.

Correlation Matrices

```
> pairs(dataset[c(2:4,8:11,18,22:34)])
```

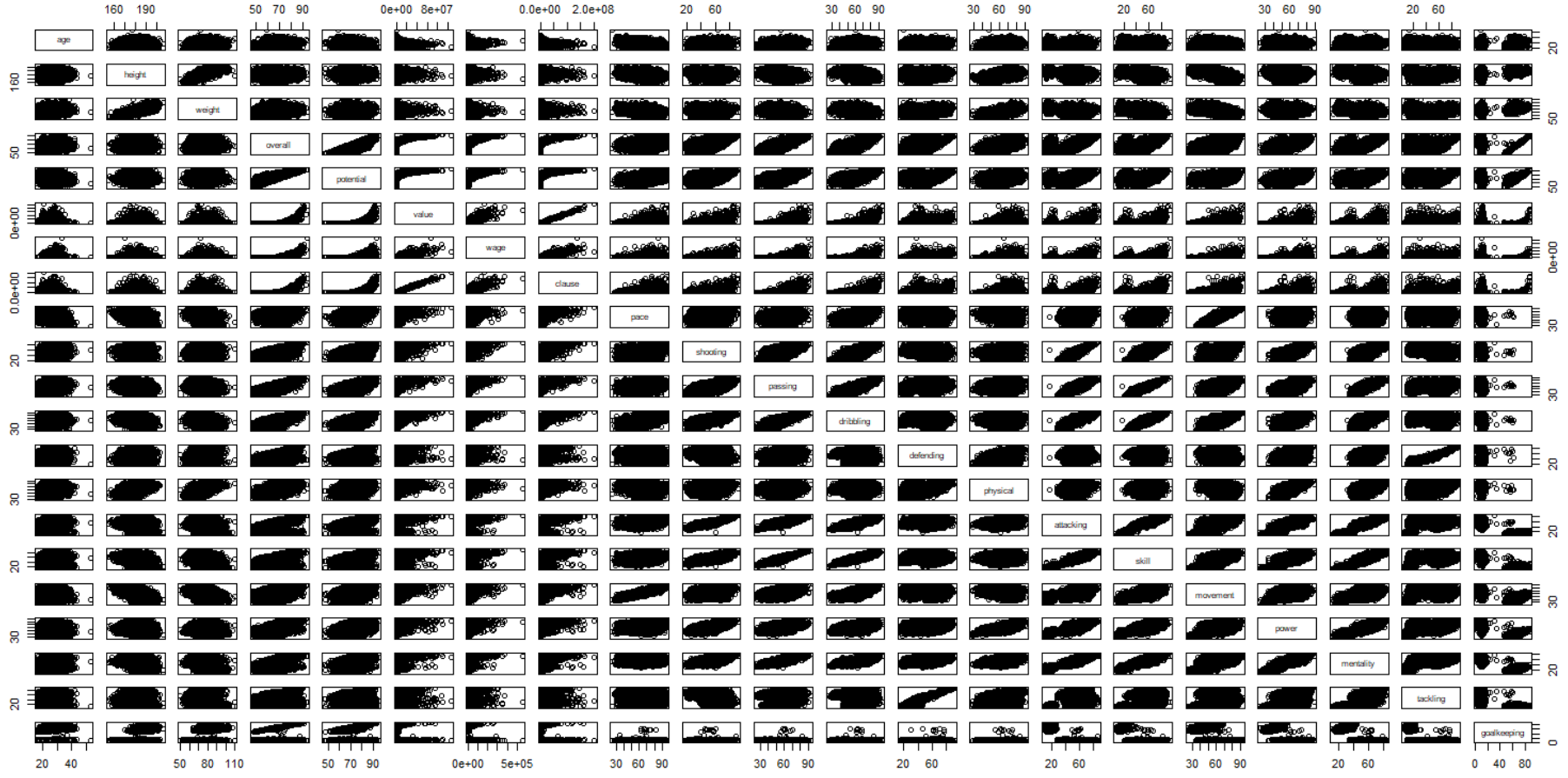

Figure 1 - Numerical variable correlations visualised using the pairs function

```
> pairs(dataset[c(8:11,18)])
```

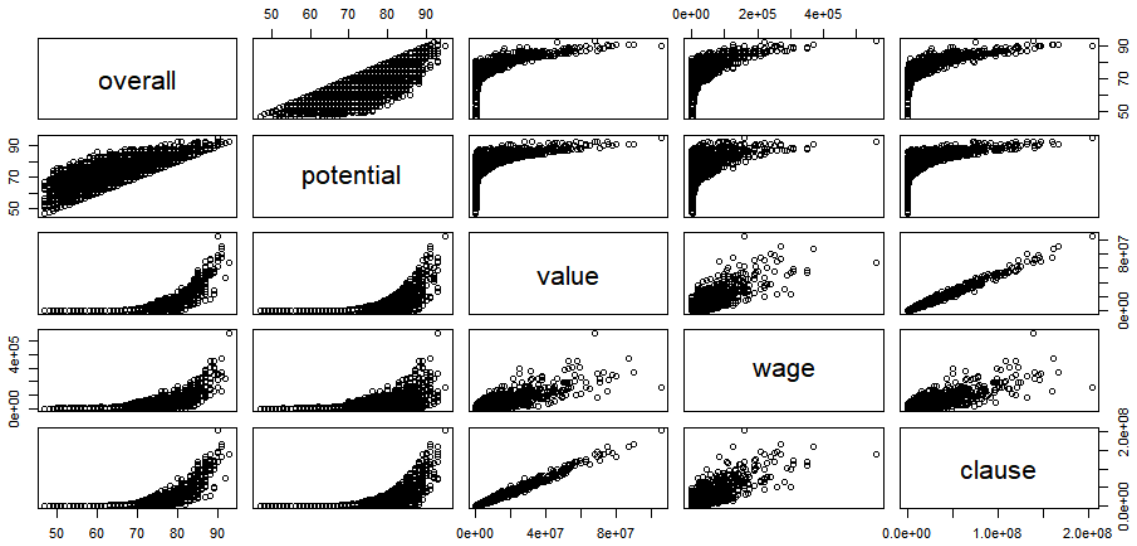

Figure 2 - A closer look at a section from the original pairs function call

This particular section of the pairs plot appeared interesting from an initial glance. Upon closer inspection the highest correlation in the dataset is found in this section between clause and value. With a supporting value from the `cor()` function, 0.99391476 and an R-squared value that suggests 98.8% of the variance in clause can be accounted for by the variance in value.

```
> r2 <- lm(value ~ clause, data = dataset)
> summary(r2)$r.squared

  [1] 0.9880128
```

Some other interesting observations from these specific variable correlations, which were noted as likely being statistically relevant to determining the most significant difference between observations (rows) within the dataset. Both overall and potential show seemingly exponential graphs when compared with the fiscal variables of value and clause. Indicating that these values are nominal for the majority of player ratings across the board, but players with particularly high ratings (above 70) immediately become more valuable in an exponential fashion. The same appears to be the case for overall and potential when compared to wage, just to a lesser extent.

Value and wage seem to be correlated from the plot and this interpretation is supported by the figure 0.83948620 found from using the `cor()` function call. With an R-squared value suggesting that 70.58% of the variance in one can account for the variance in the other.

```
> r2 <- lm(value ~ wage, data = dataset)
> summary(r2)$r.squared

  [1] 0.7058118
```

Finally the overall and potential attributes create some interesting plots when they are compared against each other. Visually one may be inclined to state that these values are correlated, but when inspected with the `cor()` function these variables return a value of 0.636366207, suggesting correlation but not exceedingly strong. The straight lines appearing on the plots giving this impression can be attributed to the fact that players cannot possibly have an overall rating higher than that of their potential rating. Thus for a lot of players who have reached their potential, they run along this limit line, the sagging effect is a representation of all those players that have not yet reached their potential.

```
> r2 <- lm(overall ~ potential, data = dataset)
> summary(r2)$r.squared

  [1] 0.4049619
```

The R-Squared error doesn't support any indication that this correlation coefficient is accurate or strong at all. The error value only estimates that 40.5% of the variance in a player's overall rating can be accounted for by the variance in their potential rating.

Summary Statistics

```
> summary(dataset)
```

```
   name               age             height          weight
Length:18944       Min.   :16.00   Min.   :155.0   Min.   : 50.00
Class :character   1st Qu.:21.00   1st Qu.:176.0   1st Qu.: 70.00
Mode  :character   Median :25.00   Median :181.0   Median : 75.00
                   Mean   :25.23   Mean   :181.2   Mean   : 75.02
                   3rd Qu.:29.00   3rd Qu.:186.0   3rd Qu.: 80.00
                   Max.   :53.00   Max.   :206.0   Max.   :110.00

   nationality                   club
England  : 1685                    :  225
Germany  : 1189   1. FSV Mainz 05 :   33
Spain    : 1072   Angers SCO      :   33
France   :  984   Arsenal         :   33
Argentina:  936   AS Monaco       :   33
Brazil   :  887   AS Saint-Étienne:   33
(Other)  :12191   (Other)         :18554

                        league         overall         potential
English League Championship:  709   Min.   :47.00   Min.   :47.00
USA Major League Soccer    :  701   1st Qu.:61.00   1st Qu.:67.00
Argentina Primera DivisiÃ³n:  659   Median :66.00   Median :71.00
English Premier League     :  654   Mean   :65.68   Mean   :71.09
Italian Serie A            :  645   3rd Qu.:70.00   3rd Qu.:75.00
Spain Primera Division     :  645   Max.   :93.00   Max.   :95.00
(Other)                    :14931

    value               wage            clause              position
Min.   :        0   Min.   :     0   Min.   :     9000   CB     :3252
1st Qu.:   300000   1st Qu.:  1000   1st Qu.:   525000   ST     :2645
Median :   650000   Median :  3000   Median :  1100000   CM     :2241
Mean   :  2224813   Mean   :  8676   Mean   :  4296353   GK     :2084
3rd Qu.:  1800000   3rd Qu.:  7000   3rd Qu.:  3200000   CDM    :1514
Max.   :105500000   Max.   :560000   Max.   :203100000   LB     :1402
                                     NA's   :995         (Other):5806
```

```
   foot        reputation weakfoot   attack_wr      defend_wr
Left : 4496   1:17593    1:  135   High  : 5272   High  : 3275
Right:14448   2: 1017    2: 4176   Low   :  981   Low   : 1752
              3:  285    3:11661   Medium:12691   Medium:13917
              4:   43    4: 2699
              5:    6    5:  273

    jersey       loan          contract         pace
Min.   : 1.00   No :18186   Min.   :2020   Min.   :25.00
1st Qu.: 9.00   Yes:  758   1st Qu.:2021   1st Qu.:62.00
Median :18.00               Median :2022   Median :68.00
Mean   :20.59               Mean   :2022   Mean   :67.67
3rd Qu.:27.00               3rd Qu.:2023   3rd Qu.:75.00
Max.   :99.00               Max.   :2028   Max.   :96.00
NA's   :225                 NA's   :225    NA's   :2083

   shooting        passing        dribbling       defending
Min.   :16.00   Min.   :25.00   Min.   :25.00   Min.   :15.00
1st Qu.:42.00   1st Qu.:50.00   1st Qu.:57.00   1st Qu.:36.00
Median :54.00   Median :58.00   Median :64.00   Median :56.00
Mean   :52.27   Mean   :57.14   Mean   :62.46   Mean   :51.32
3rd Qu.:63.00   3rd Qu.:64.00   3rd Qu.:69.00   3rd Qu.:64.00
Max.   :93.00   Max.   :93.00   Max.   :95.00   Max.   :91.00
NA's   :2083    NA's   :2083    NA's   :2083    NA's   :2083

   physical       attacking         skill          movement
Min.   :28.00   Min.   : 8.40   Min.   : 8.00   Min.   :24.40
1st Qu.:58.00   1st Qu.:44.40   1st Qu.:44.20   1st Qu.:57.80
Median :66.00   Median :52.40   Median :53.80   Median :65.20
Mean   :64.46   Mean   :49.73   Mean   :51.25   Mean   :63.49
3rd Qu.:72.00   3rd Qu.:59.40   3rd Qu.:62.00   3rd Qu.:71.20
Max.   :91.00   Max.   :87.40   Max.   :94.00   Max.   :92.80
NA's   :2083

    power          mentality        tackling      goalkeeping
Min.   :24.40   Min.   :10.50   Min.   : 6.00   Min.   : 1.00
1st Qu.:52.80   1st Qu.:46.50   1st Qu.:26.00   1st Qu.: 9.60
Median :60.40   Median :53.50   Median :53.50   Median :10.60
Mean   :59.29   Mean   :51.98   Mean   :46.56   Mean   :16.31
3rd Qu.:66.80   3rd Qu.:60.30   3rd Qu.:64.50   3rd Qu.:11.80
Max.   :88.80   Max.   :83.70   Max.   :89.50   Max.   :88.00
```

The first immediately noticeable point of information to be taken from this dataset summary is the number of NA values for the previously mentioned base player attributes, which are all matched in value. This value of 2083 gives us the exact number of goalkeepers that exist in the dataset. The same logic can be used to identify that 995 players do not currently have a release clause as part of their contract. While only 225 players do not currently have a contract at all which can be inferred from the number of NA values presented for contract and jersey variables.

The mean and median values, as well as the quartiles also offer some great insight into the rating attributes presented. The third quartile for player overall ratings is 70, which matches the observation from the correlation plots which begins its upward trajectory in respect to value/wage around the rating of 70.

Similar information can be derived from the fiscal variables. Value, wage and clause the third quartile values are actually lower in value than the mean values. Further supporting the postulation that nominal values are maintained for these values until players achieve higher ratings or reputations at which point their worth and compensation grows exponentially.

Barplots

Viewing the correlation scatter plots generated using the `pairs()` function, the financially numerical variables, value, wage and clause, which are correlated to each other to varying degrees, appear to be normally distributed about the physical player attributes age, height and weight. While this might be logical for age in that a player's ability is likely to get better as they get older from youth and to then decline as they age, and so they may be paid less as their ratings reduce. This is a little odd to see occur based on a player's height and weight.

These variables are more likely to be normally distributed individually considering this is a measurable population of players within the game based on actual persons.

```
> ggplot(data = dataset, aes(age))    + geom_bar(fill = "dark
blue")
> ggplot(data = dataset, aes(height)) + geom_bar(fill = "dark
blue")
> ggplot(data = dataset, aes(weight)) + geom_bar(fill = "dark
blue")
```

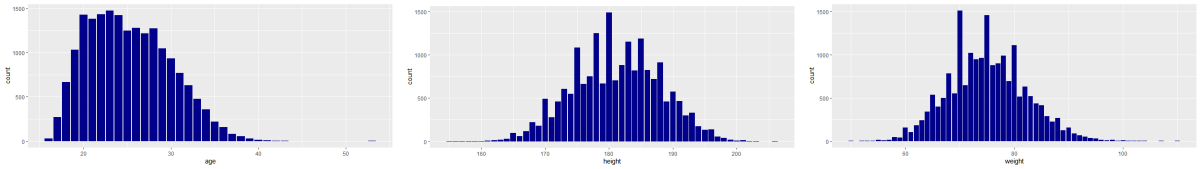

Figure 3 - Barplots of player's physical attributes

While these generated barplots might appear to be normally distributed to the eye, the Shapiro-Wilks test for normality can be performed to determine normality in a statistical manner.

```
> shapiro.test(table(dataset$age))

    Shapiro-Wilk normality test

    data:  table(dataset$age)
    W = 0.85305, p-value = 0.000879

> shapiro.test(table(dataset$height))

    Shapiro-Wilk normality test

    data:  table(dataset$height)
    W = 0.85286, p-value = 1.85e-05

> shapiro.test(table(dataset$weight))

    Shapiro-Wilk normality test

    data:  table(dataset$weight)
    W = 0.81462, p-value = 6.446e-07
```

The null-hypothesis for this test states that the population is normally distributed for the relevant attribute. Each of the p-values presented from these tests show a certainty level well below the 0.01 alpha. Therefore the null hypothesis is rejected and there is evidence that the data tested are not normally distributed.

```
> barplot(table(dataset$value),  col = "red")
> barplot(table(dataset$wage),   col = "red")
> barplot(table(dataset$clause), col = "red")
```

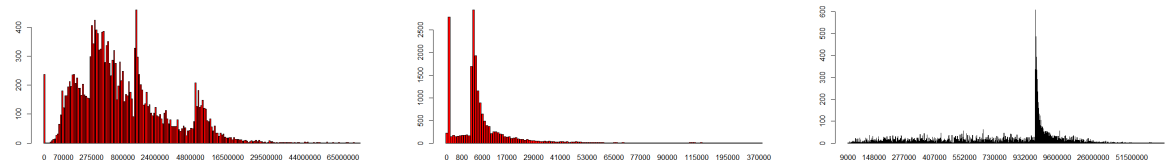

Figure 4 - Barplots for fiscal attributes value, wage and release clause

The fiscal variables on the other hand have absolutely no resemblance to normally distributed curves. This might indicate that the correlation plots for each of these variables against the attributes age, height and weight could be used to identify the "perfect" physical dimensionality for a player.

In order to determine which characteristic values for age, height and weight are optimal for a player to be considered good in this game. The scatter plots of each of the variables compared will be overlaid with a smoothing distribution curve and the apex of the point will be calculated and used as the optimal value in each case. Averaging these characteristic values across the respective fiscal values will offer an idea as to the optimal measures.

```
> p1 <- ggplot(data = dataset, aes(age, value)) + geom_point() +
  stat_smooth(method = "gam", formula = y ~ s(x), size = 1)
> p2 <- ggplot(data = dataset, aes(age, wage)) + geom_point() +
  stat_smooth(method = "gam", formula = y ~ s(x), size = 1)
> p3 <- ggplot(data = dataset, aes(age, clause)) + geom_point() +
  stat_smooth(method = "gam", formula = y ~ s(x), size = 1)

> gb1 <- ggplot_build(p1)
> gb2 <- ggplot_build(p2)
> gb3 <- ggplot_build(p3)

> apex1 <- gb1$data[[2]]$x[which.max(gb1$data[[2]]$y)]
> apex2 <- gb2$data[[2]]$x[which.max(gb2$data[[2]]$y)]
> apex3 <- gb3$data[[2]]$x[which.max(gb3$data[[2]]$y)]

> p1 + geom_vline(xintercept = apex1, linetype = "dashed",
  color = "red")
> p2 + geom_vline(xintercept = apex2, linetype = "dashed",
  color = "red")
> p3 + geom_vline(xintercept = apex3, linetype = "dashed",
  color = "red")
```

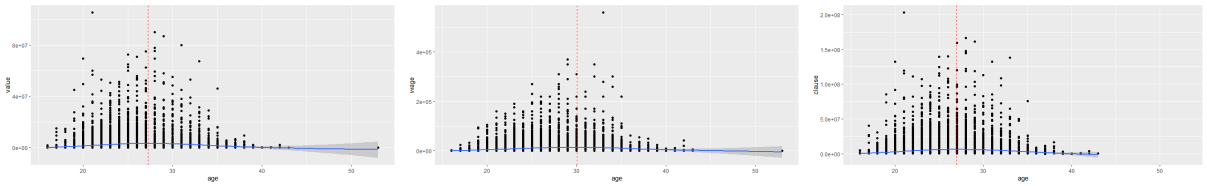

Figure 5 - Scatterplots of age vs the fiscal values with optimal age calculated

```
> apex1          > apex2          > apex3
[1] 27.24051     [1] 30.05063     [1] 26.93671
```

The average and thus optimal age for a good player is calculated to be 28 (28.07595).

Unfortunately due to the large density of observations along the x-axis in each instance, indicating that many players have lower values, wages and clauses, the smoothing curve produced cuts right through the data, likely skewing it away from the desired shape.

It was hoped that this curve when generated would follow along the visible upper limit of the data points within each plot. That way the apex point could be used as the marker along the x-axis to produce the relative physical attribute value.

This method only worked effectively for the age characteristic. For both height and weight the smooth line produced was almost flat and indicated an apex point far off to the right hand side of the grid. The scatterplots for height and weight will be displayed and the mean point of the characteristic displayed using a point and judged by the eye to determine if it aligns with apex of the model well, if not a vertical line will be applied to the guessed value on the x-axis.)

```
> ggplot(data = dataset, aes(height, value)) + geom_point()
  + geom_vline(xintercept = mean(dataset$height),
  linetype = "dashed", color = "red")

> ggplot(data = dataset, aes(height, wage)) + geom_point()
  + geom_vline(xintercept = mean(dataset$height),
  linetype = "dashed", color = "red")

> ggplot(data = dataset, aes(height, clause)) + geom_point()
  + geom_vline(xintercept = mean(dataset$height),
  linetype = "dashed", color = "red")
```

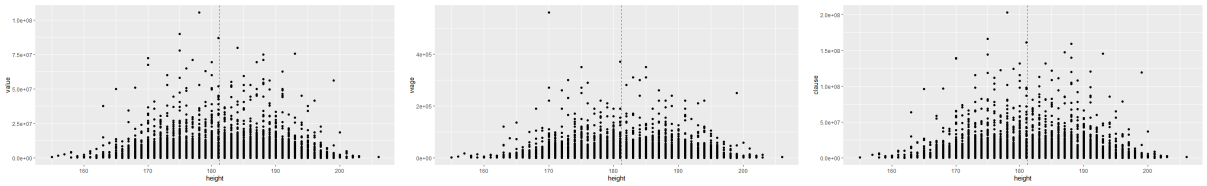

Figure 6 - Scatterplots of height versus fiscal variables with mean height

```
> ggplot(data = dataset, aes(weight, value)) + geom_point()
  + geom_vline(xintercept = mean(dataset$weight),
  linetype = "dashed", color = "red")

> ggplot(data = dataset, aes(weight, wage)) + geom_point()
  + geom_vline(xintercept = mean(dataset$weight),
  linetype = "dashed", color = "red")

> ggplot(data = dataset, aes(weight, clause)) + geom_point()
  + geom_vline(xintercept = mean(dataset$weight),
  linetype = "dashed", color = "red")
```

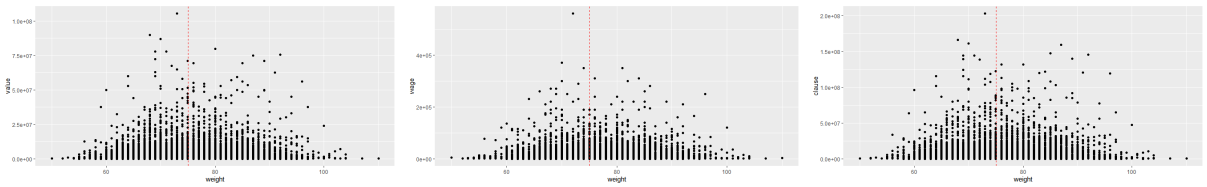

Figure 7 - Scatterplot of weight against the fiscal variables with mean weight

The mean average values, denoted by the dashed red lines appear to mark the apex fairly well for all of the plots for height and weight. With this it can be concluded that the optimal measure for these characteristics are in fact their mean values.

The optimal physical characteristics for a player are therefore believed to be;

Age = 28.07595 Height = 181.1908 cm Weight = 75.01689 kg

Density Maps

A more appropriate method to visually isolate typical physical characteristics for players is through the use of density graphs. The fiscal values do not offer any great insight in this regard as the high density areas are around the median physical attribute values but, as was seen with the exponential like graphs for value versus overall value, most players are likely to be valued or paid poorly and so the density will only appear in a spread area above the x-axis. Here we can determine a range for average player attributes.

```
> ggplot(data = dataset, aes(age, overall))
+ stat_density2d(geom="tile", contour=FALSE, aes(fill =
..density..)) + scale_fill_distiller(palette = 'Spectral')

> ggplot(data = dataset, aes(height, overall))
+ stat_density2d(geom="tile", contour=FALSE, aes(fill =
..density..)) + scale_fill_distiller(palette = 'Spectral')

> ggplot(data = dataset, aes(weight, overall))
+ stat_density2d(geom="tile", contour=FALSE, aes(fill =
..density..)) + scale_fill_distiller(palette = 'Spectral')
```

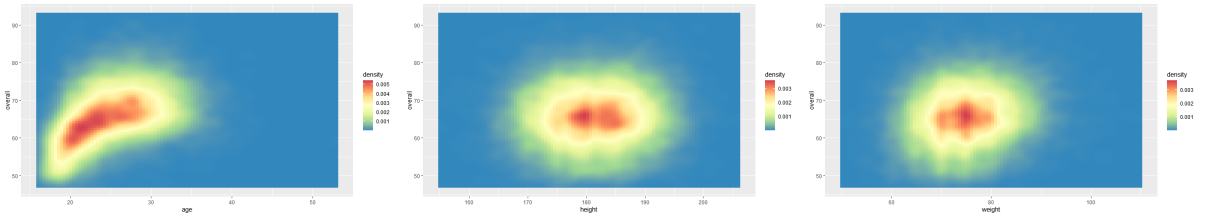

Figure 8 - Density graphs showing average player age, height and weight

Both the height and weight attributes are distributed in a generally central manner. Age however is skewed to the left of the graph indicating that more players tend to be younger rather than older.

Continuous vs Categorical Variables

Clustering

Clustering by player position should provide an insight into which position(s) are valued most in the game if any. Plotting the primary rating variable attributes overall and value against each other and differentiating by position returns the following.

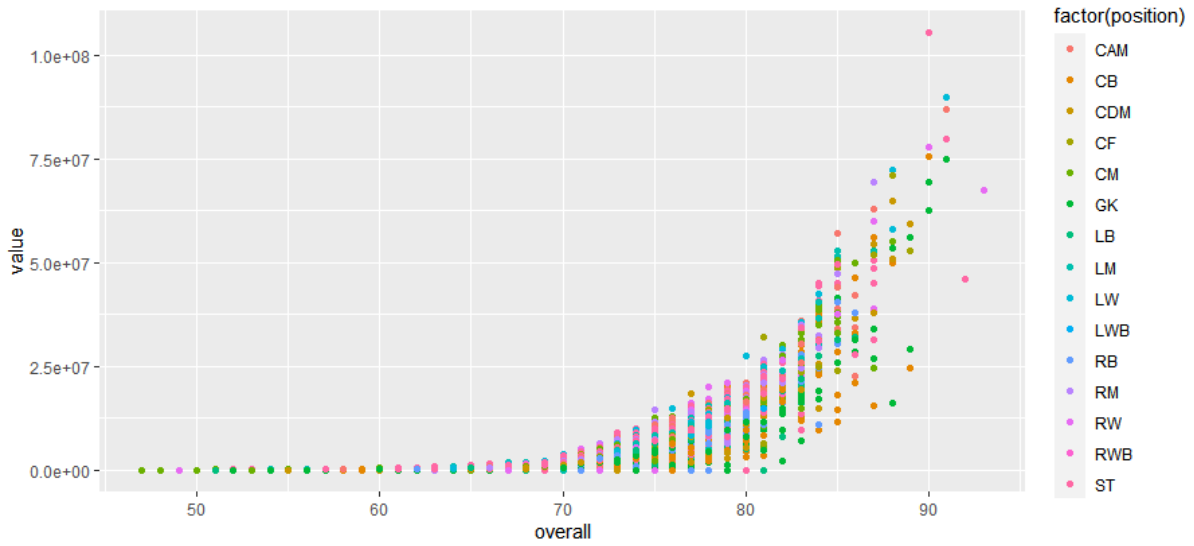

Figure 9 - Clustered plot of overall/value by position

Nothing in particular can be gleaned as statistically relevant from this visualisation. Other iterations such as wage/value, clause/value and height/weight were observed in an attempt to discover a clustering effect based upon player position, all returning similar negative results.

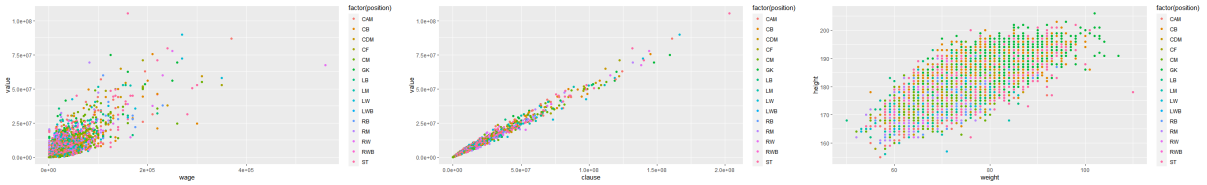

Figure 10 - Other attempted iterations at position clustering

It would be appropriate to conclude that there are no particularly significant interaction effects disproportionately affecting any specific playing position.

Congruent/Proportionality Barplots

```
> ggplot(workrate, aes(x = Category, y = Players, fill = WorkRate))
+ geom_bar(position = "dodge",  stat = "identity")

> ggplot(workrate, aes(x = WorkRate, y = Players, fill = Category))
+ geom_bar(position = "dodge",  stat = "identity")
```

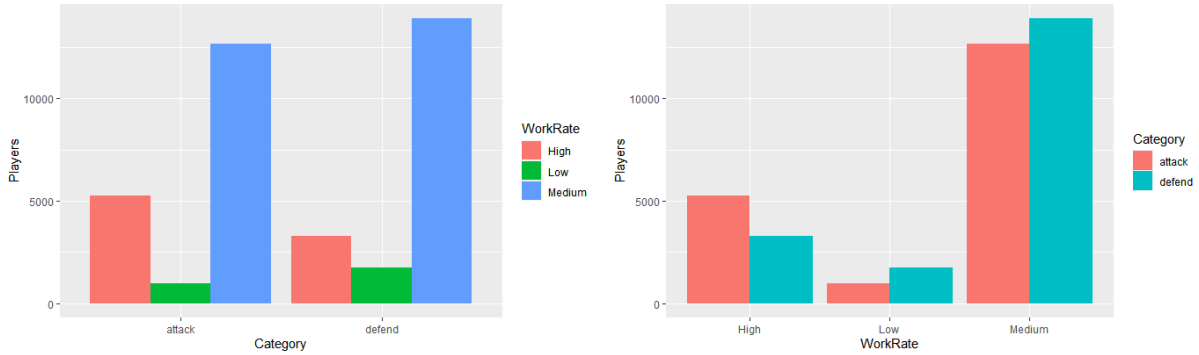

Figure 11 - Player work rate representations

These barplots show the representation of players within each of the work rate categories, from High to Low for both attack and defense.

The proportionality shows that the overwhelming majority of players in both categories have a medium level workrate, with less than half of this proportionality of players having a high workrate. This high workrate proportionality is higher for attacking players than defensive players. Very few players have a low workrate, showing a proportionality of less than half that was shown for players with a high workrate, but with this being more prevalent on defense than attack.

Pie Charts

Pie charts are used to show these proportionalities better in a percentage style rather than in total numbers as the barplots have shown.

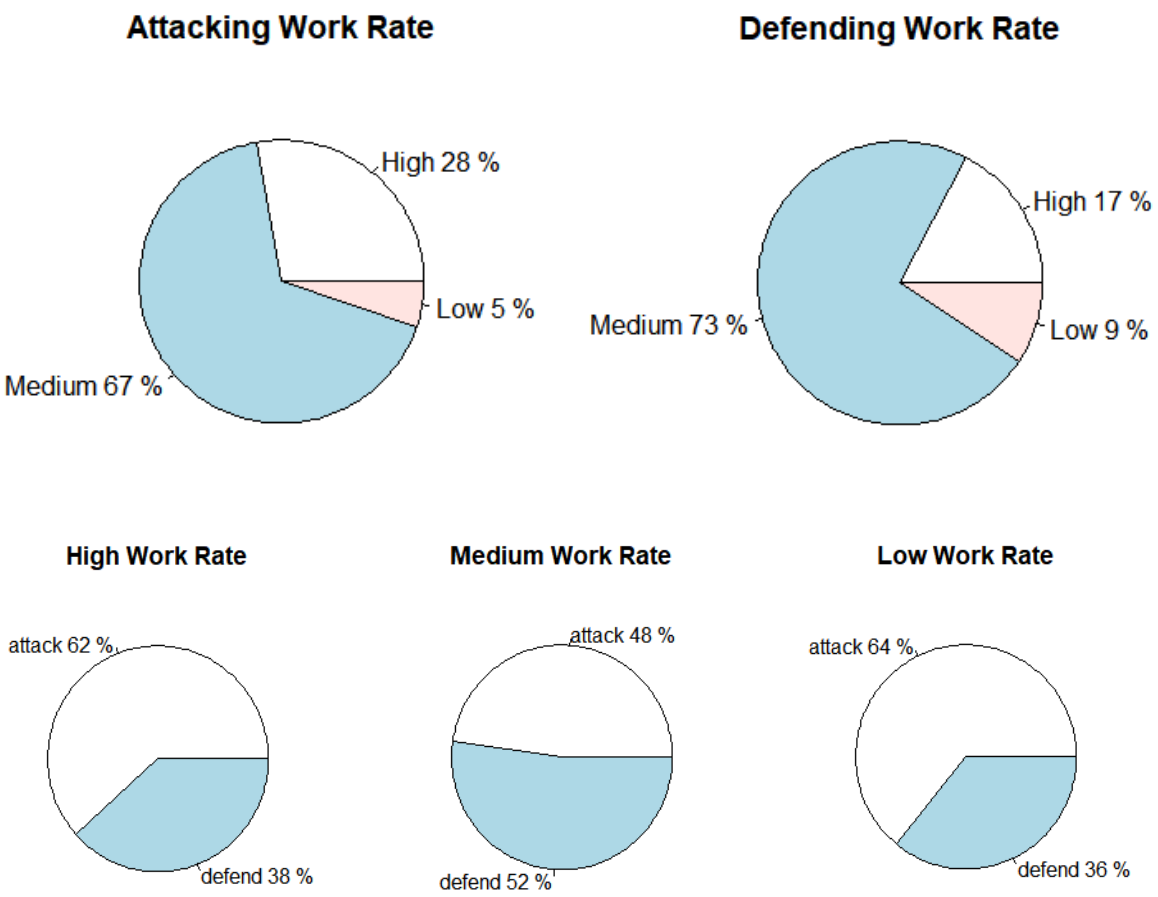

Figure 12 - Pie Chart percentages between work rate categories

The pie charts further support the interpretations offered from the barplots above. The majority of players have a medium work rate on both attack and defence. Medium workrate is in fact divided evening with a near 50:50 split on attack and defense, further supporting this level to be typical for most players. Interestingly both high and low workrates are more prevalent on the attacking side. This can likely be explained by the gulf in motivation between attacking players and defending players. Low workrates are uncommon with less than 10% of the population in both categories. This is as would be expected any professional football player would typically work hard at the job they love and get paid very well to do.

Boxplots

Exploring the reputation variable offers insight into the effects of a player's reputation on their value and overall rating or vise versa. Boxplots provide the best opportunity to visualise the differences and commonalities between categorical and numerical variables.

```
> ggplot(dataset, aes(reputation, overall))+ geom_boxplot()
+ stat_summary(fun = mean, geom = "point", color = "red", size =
3)

> ggplot(dataset, aes(reputation, value)) + geom_boxplot()
+ stat_summary(fun = mean, geom = "point", color = "red", size =
3)
```

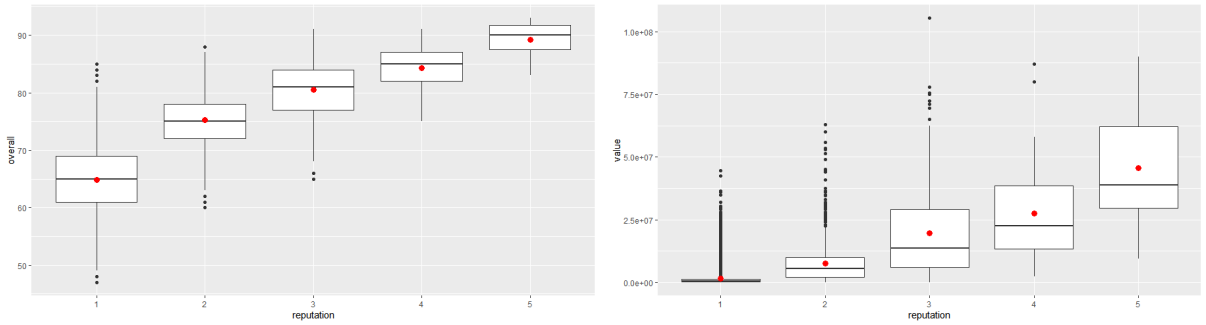

Figure 13 - boxplots of reputation vs overall and value respectively

Overall rating when compared to international reputation can be seen to increase in an almost linear manner on the average point. This is to be expected as the greater one's international reputation the higher their overall rating should be. It is interesting to see the upper boundaries of these reputation levels compared though. While on average there is definitely an increase based on reputation level, this visualisation suggests that there are players in lower reputation levels with higher overall ratings than some of their counterparts in higher levels. For example the upper boundary of level three is higher than the lower boundary of level five.

In contrast to the reputation versus overall boxplot this interpretation versus value offers a different perspective. Where the highest valued player that presents as an outlier appears in the level three of reputation over 10 million euros more than the highest valued player in the level five of reputation class. This supports the less obvious inconsistencies postulated from the previous plot and gives more insight into the different reputation requirements for each country. Countries with a history of great players appear to have higher standards to be considered within a particular reputation class, whereas countries with less available players or a less illustrious history are more likely to rate players higher in reputation.

Categorical Variables

Contingency Table

Continuing along the same vein of thinking, two logically associated variables are a player's international reputation with their nationality. Some useful techniques to observe this data are through the use of a contingency table followed by heatmap visualisation.

```
> contingency <- table(dataset$reputation, dataset$nationality)

  Afghanistan Albania Algeria Andorra Angola Antigua & Barbuda
1           2      45      40       1     15                 4
2           0       4       9       0      1                 0
3           0       0       3       0      0                 0
4           0       0       0       0      0                 0
5           0       0       0       0      0                 0

  Argentina Armenia Aruba Australia Austria Azerbaijan Barbados
1       853       3     1       239     306          6        1
2        58       0     0         2      12          0        0
3        20       1     0         0       2          0        0
4         4       0     0         0       1          0        0
5         1       0     0         0       0          0        0

  Belarus Belgium Belize Benin Bermuda Bolivia Bosnia Herzegovina
1       3     265      2    10       1     145                 53
2       0      18      0     1       0       1                 10
3       0      12      0     0       0       0                  3
```

```
4       0       3      0     0       0       0                  0
5       0       0      0     0       0       0                  0

  Brazil Bulgaria Burkina Faso Burundi Cameroon Canada Cape Verde
1    809       37           17       6       66     76          22
2     51        1            1       1        7      2           1
3     21        0            1       0        4      0           0
4      5        0            0       0        0      0           0
5      1        0            0       0        0      0           0

  Central African Republic Chad Chile China PR Chinese Taipei
1                        3    1   171      359              2
2                        1    0    16        5              0
3                        0    0     2        0              0
4                        0    0     2        0              0
5                        0    0     0        0              0

  Colombia Comoros Congo Costa Rica Croatia Cuba Curacao Cyprus
1      317      12    15         25     116    5      13      7
2       17       0     1          3       9    0       0      0
3        3       0     0          1       5    0       0      0
4        1       0     0          0       2    0       0      0
5        0       0     0          0       0    0       0      0

  Czech Republic Denmark Dominican Republic DR Congo Ecuador Egypt
1            103     282                  4       60     244    25
2              8      16                  0        5       8     2
3              0       2                  0        0       0     2
4              0       0                  0        0       0     0
5              0       0                  0        0       0     0

  El Salvador England Equatorial Guinea Eritrea Estonia Ethiopia
1           4    1609                 5       2       4        2
2           0      61                 1       0       1        0
3           0      14                 0       0       0        0
4           0       1                 0       0       0        0
5           0       0                 0       0       0        0

  Faroe Islands Finland France Gabon Gambia Georgia Germany Ghana
1             5      62    867    14     24      22    1084   109
2             0       3     87     1      0       0      81     7
3             0       0     25     1      0       0      17     2
4             0       0      5     0      0       0       6     0
5             0       0      0     0      0       0       1     0

  Greece Grenada Guam Guinea Guinea Bissau Guyana Haiti Honduras
1     91       4    1     26            15     10    10       14
2      9       0    0      4             0      0     0        2
3      3       0    0      0             0      0     0        0
4      0       0    0      0             0      0     0        0
5      0       0    0      0             0      0     0        0

  Hong Kong Hungary Iceland India Indonesia Iran Iraq Israel Italy
1         2      34      49    23         1   14    4     17   329
2         0       5       4     0         0    1    0      2    76
3         0       0       1     0         0    0    0      0    14
4         0       0       0     0         0    0    0      0     2
5         0       0       0     0         0    0    0      0     0

  Ivory Coast Jamaica Japan Jordan Kazakhstan Kenya Korea DPR
1          92      23   476      2          6     9         1
2          12       1    10      0          0     1         0
```

3	1	0	3	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0

	Korea Republic	Kosovo	Latvia	Lebanon	Liberia	Libya	Liechtenstein
1	324	44	7	3	4	4	5
2	15	2	0	0	0	0	0
3	2	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0

	Lithuania	Luxembourg	Macau	Madagascar	Malawi	Malaysia	Mali	Malta
1	10	11	1	8	1	1	49	1
2	0	0	0	1	0	0	1	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0

	Mauritania	Mexico	Moldova	Montenegro	Montserrat	Morocco
1	7	341	11	27	3	68
2	0	8	1	2	0	11
3	0	13	0	1	0	1
4	0	0	0	0	0	0
5	0	0	0	0	0	0

	Mozambique	Namibia	Netherlands	New Caledonia	New Zealand
1	4	3	375	1	43
2	1	0	42	0	1
3	0	0	15	0	0
4	0	0	0	0	0
5	0	0	0	0	0

	Nicaragua	Niger	Nigeria	North Macedonia	Northern Ireland	Norway
1	1	4	117	25	76	357
2	0	0	7	1	4	5
3	0	0	2	0	0	1
4	0	0	0	0	0	0
5	0	0	0	0	0	0

	Palestine	Panama	Papua New Guinea	Paraguay	Peru	Philippines
1	2	11	1	227	157	3
2	0	0	0	6	4	0
3	0	0	0	2	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

	Poland	Portugal	Puerto Rico	Ireland	Romania	Russia	Rwanda
1	341	305	1	328	318	67	1
2	12	30	0	10	6	8	0
3	6	16	0	0	0	2	0
4	1	1	0	0	0	0	0
5	0	1	0	0	0	0	0

	São Tomé & Príncipe	Saint Kitts and Nevis	Saint Lucia	Saudi Arabia
1	1	2	1	316
2	0	0	0	1
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0

	Scotland	Senegal	Serbia	Sierra Leone	Slovakia	Slovenia
1	282	116	111	9	64	43
2	4	9	16	0	2	6

```
3         1        3       3             0         2         3
4         0        0       0             0         0         0
5         0        0       0             0         0         0

  South Africa South Sudan Spain Sudan Sweden Switzerland Syria
1           71           2   948     4    356         192     3
2            1           0    85     0     14          14     1
3            0           0    31     0      1           5     0
4            0           0     8     0      0           0     0
5            0           0     0     0      1           0     0

  Tanzania Thailand Togo Trinidad & Tobago Tunisia Turkey Uganda
1        1        4   13                 6      33    328      9
2        0        0    1                 0       0     12      0
3        0        0    0                 0       0      4      0
4        0        0    0                 0       0      0      0
5        0        0    0                 0       0      0      0

  Ukraine UAE USA Uruguay Uzbekistan Venezuela Wales Zambia Zimbabwe
1      59  22 368     335          7       193   116     11        11
2       5   0   8      10          0         5     5      0         0
3       1   0   2       3          0         0     2      0         0
4       0   0   0       0          0         0     1      0         0
5       0   0   0       1          0         0     0      0         0
```

From viewing the contingency table it appears that some countries have higher numbers of players in certain reputation levels that do others. This seems to imply that the number of players in a reputation level is dependent upon the country in question. This can be tested through the use of a Chi Squared test.

H_0: The two categorical variables reputation and nationality are independent

H_1: The categorical variables reputation and nationality are dependent

```
> chisq.test(contingency)

    Pearson's Chi-squared test

    data:  contingency
    X-squared = 837.15, df = 644, p-value = 3.902e-07
```

The small p-value observed allows for the rejection of the null hypothesis, suggesting that the variables in this contingency table are dependent upon each other at a significance level far less than 0.01.

This makes sense logically as a country with more available players or greater international reputation is more likely to have better players on the national team so of course this would be dependent upon each specific country.

Heatmaps

Having proven this categorical dependency it is now worth while to visualise the data that was previously presented in the contingency table. This can be done through the use of heatmaps.

```
> heatmap(con, Colv = NA, Rowv = NA)
```

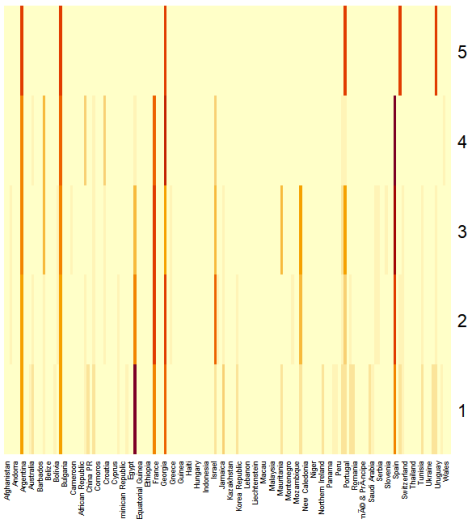

Figure 14 - Normalised heatmap visualisation of reputation/nationality contingency data

While the native heatmap function is limited in it's scaling and therefore viewing capabilities making it difficult to interpret the data visually due to the number of country variables that exist in this instance. It does offer a normalisation feature across the rows which gives a deeper insight into how the numbers of players from a certain county match up in respect to the number of players with a specific reputation level.

```
> ggplot(contingency, aes(Country, Reputation, fill = Freq)) + geom_tile() +
theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.4)) +
   scale_fill_distiller(palette = "YlOrRd", direction = 1)
```

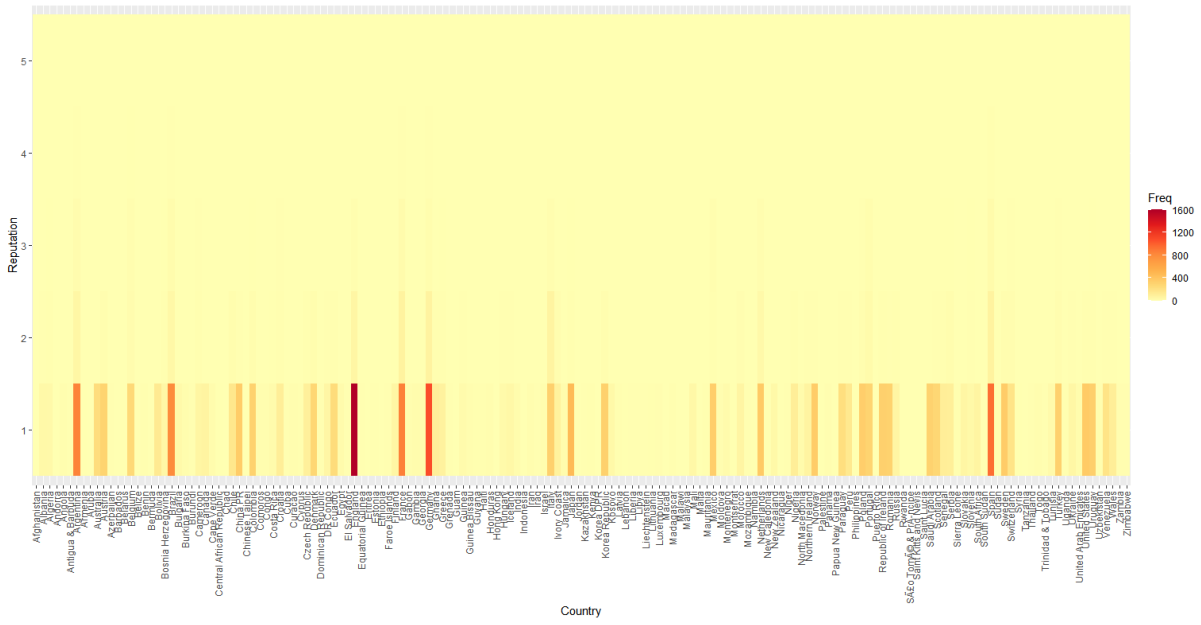

Figure 15 - Heatmap using ggplot without normalisation

This ggplot2 version using `geom_tile()` scales better, showing each country within its own space but there is no easy way to provide normalisation across the reputation levels such was the case with the native `heatmap()` function, without hard coding the math into the function call. It is however interesting and informative to see the data in an unnormalised manner.

The contingency table isn't pleasant viewing but it does provide greater depth of information than a heatmap.

Geomap

Populating a map of the world with the number of players in the dataset from each country should provide some interesting insight into where the game is most popular throughout the world, and where the best players originate from.

```
> library(ggplot2)
> library(dplyr)
> require(maps)

> players <- as.data.frame(table(dataset$nationality))

> player <- aggregate(players$Freq,
  by = list(region = players$Var), FUN = sum)

> player_map <- left_join(world_map, player, by = "region")

> ggplot(player_map, aes(long, lat, group = group))
 +  geom_polygon(aes(fill = x), color = "white")
```

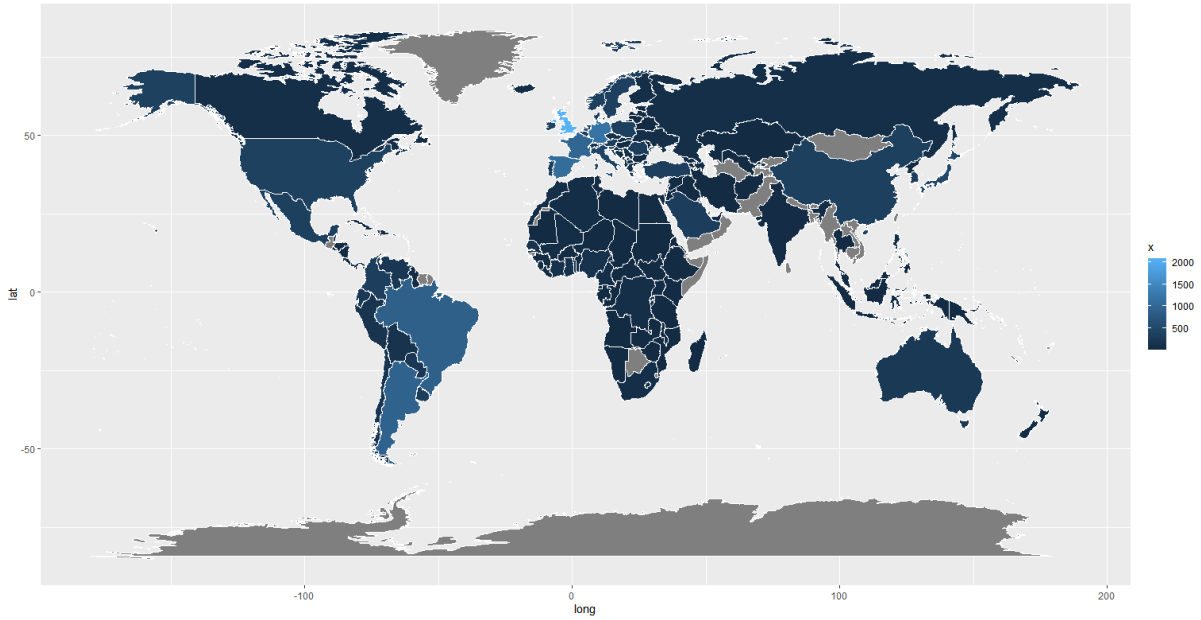

Figure 16 - Heatmap of player origins

It is immediately obvious from viewing this heatmap that the UK far out performs every other country in the world in respect to its number and production of football players, but they are followed by a handful of other well known footballing nations such as Spain, France, Germany, Brazil and Argentina. Ths US pales in comparison to these countries, but due to its sheer size not so much as do other smaller nations. A fact that can easily be explained by overwhelming interest in homegrown sports, such as American Football and Baseball, sports that are sparsely contested elsewhere in the world. This graph really brings truth to the phrase "Football is the World's Game." With less than ten nations not accounted for with zero players represented in the game.

This visualisation when coupled with the contingency reputation data offers some really great information and makes much more sense. The higher rated players in the 4 and 5 levels of reputation in particular can be traced to these highly populated footballing nations.

Simple Linear Regression

Hypothesis: Left footed players will have higher overall ratings on average.
(This will be considered the alternate hypothesis, the null hypothesis will be the inverse)

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

```
> ovrl.foot <- lm(overall ~ foot, data = dataset)
> summary(ovrl.foot)

  Call:
  lm(formula = overall ~ foot, data = dataset)

  Residuals:
       Min        1Q    Median        3Q       Max
```

```
-19.3372  -4.4726  -0.3372   4.5274  26.6628

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.3372     0.1043  636.09  < 2e-16 ***
footRight    -0.8646     0.1194   -7.24 4.66e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.993 on 18942 degrees of freedom
Multiple R-squared:  0.00276,    Adjusted R-squared:  0.002707
F-statistic: 52.42 on 1 and 18942 DF,  p-value: 4.655e-13
```

The output shows the average overall rating for a left footed player is estimated to be 66.3372, whereas right footed players are estimated to have an average overall rating of 65.4716 (66.3372 - 0.8646). The p-value calculated for the variable footRight is very significant at 4.66^{-13}, suggesting that there is statistical evidence of a difference in average overall rating between players who are left footed over players who are right footed, even if only marginally.

The p-value associated with the F-statistic is also similarly significant at 4.655^{-13}, indicating that the estimated β_1 is not equal to 0. This allows for the rejection of the null hypothesis and the conclusion that a relationship does indeed exist between left footed players and higher average overall ratings.

In an attempt to visualise this higher average overall rating based on preferred foot a boxplot is generated.

```
> ggplot(data = dataset, aes(foot, overall)) + geom_boxplot() +
  stat_summary(fun = mean, geom = "point", color = "red", size =
  3)
```

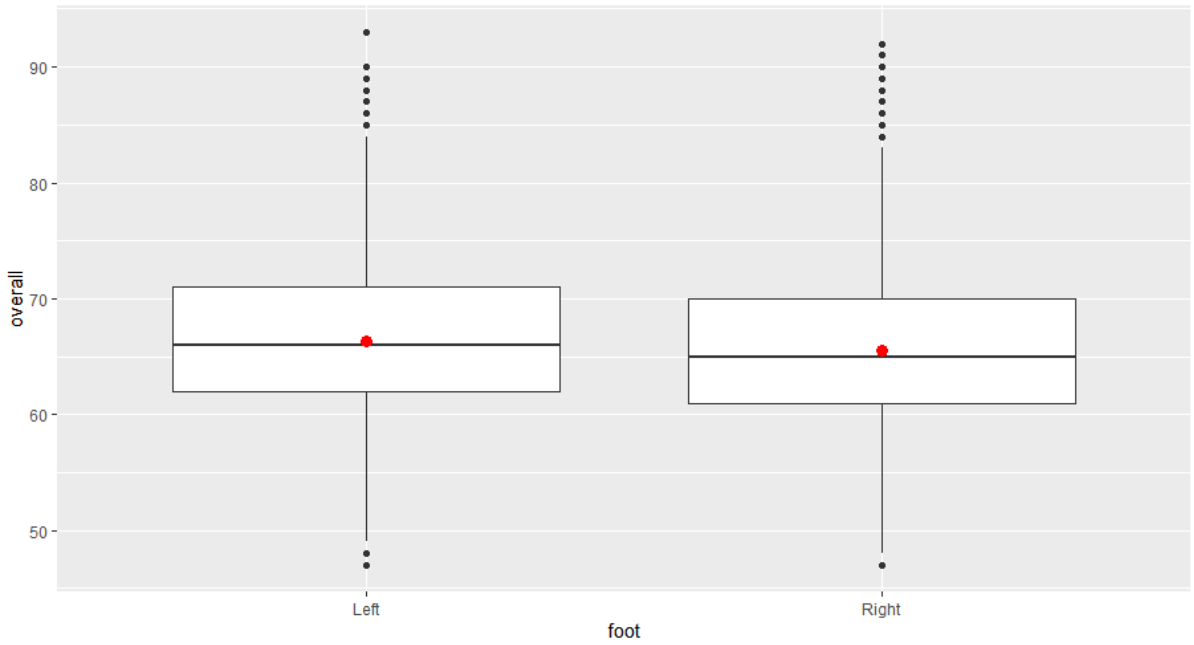

Figure 17 - Boxplot used to visualise average overall player rating based on preferred foot

Again we can see that left footed players on average have a higher overall rating. This visualisation also shows us deviation limits and outliers which suggest that in the grand scheme of things a player's preferred foot is fairly well matched for both left and right. However it is worth noting that the greatest outlier, the average point as well as the upper and lower deviation limits for left are all marginally greater than right, following the trend that left footed players typically have a slightly higher overall rating on average.

Multiplied Linear Regressions

Player value has two correlated coefficients, wage and clause these to values are themselves correlated and so are likely to have an interaction effect, this shall test this using a linear regression.

Hypothesis: There is a statistically significant synergistic interaction relationship between the response value and the multiplied variables wage and clause.

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

$H_1: \beta_1 = \beta_2 = \beta_3 \neq 0$

```
> multi <- lm(value ~ wage*clause, dataset)
> summary(multi)

 Call:
 lm(formula = value ~ wage * clause, data = dataset)

 Residuals:
       Min         1Q     Median         3Q        Max
 -10157076   -101590    -29637     73844  12303418

 Coefficients:
               Estimate Std. Error t value Pr(>|t|)
 (Intercept)  3.436e+04  4.968e+03   6.917 4.78e-12 ***
 wage         9.467e+00  4.306e-01  21.984  < 2e-16 ***
 clause       5.013e-01  7.954e-04 630.237  < 2e-16 ***
 wage:clause -2.468e-08  4.574e-09  -5.396 6.89e-08 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 561500 on 17945 degrees of freedom
   (995 observations deleted due to missingness)
 Multiple R-squared:  0.9883,  Adjusted R-squared:  0.9883
 F-statistic: 5.073e+05 on 3 and 17945 DF,  p-value: < 2.2e-16
```

The p-values presented show that the interaction between wage and clause is statistically significant as 6.89^{-8} is well below the upper threshold for significance of 0.05 which allows for rejection of the null hypothesis. With multiple predictors at play the significance of the F-statistic must be evaluated.

The p-value associated with the F-statistic is less than 2.2^{-16}, again indicating that there is a synergistic effect between these variables.

--

The player attribute overall has four correlated coefficients over the set threshold value of 0.7. These fours variables among them have five interrelated correlation coefficients over the same threshold of a possible six. Two of these five interrelations are above 0.8, these two will be used to assess interaction effects in a multiple linear regression.

Hypothesis: There is a statistically significant synergistic relationship between the response variable and the multiplied interaction variables.

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

$$H_1: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 \neq 0$$

```
> multi.2 <- lm(overall ~ passing*attacking + passing*mentality, dataset)
> summary(multi.2)

Call:
lm(formula = overall ~ passing * attacking + passing * mentality,
    data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-13.0191  -2.6225  -0.1123   2.4676  19.9953

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       29.6269943  0.9533824  31.076  < 2e-16 ***
passing           -0.0480477  0.0165745  -2.899  0.00375 **
attacking         -0.6187514  0.0279008 -22.177  < 2e-16 ***
mentality          1.2507236  0.0341625  36.611  < 2e-16 ***
passing:attacking  0.0143541  0.0004845  29.627  < 2e-16 ***
passing:mentality -0.0130632  0.0005739 -22.764  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.843 on 16855 degrees of freedom
  (2083 observations deleted due to missingness)
Multiple R-squared:  0.6906,  Adjusted R-squared:  0.6905
F-statistic:  7523 on 5 and 16855 DF,  p-value: < 2.2e-16
```

Each p-value presented in this summary is less than 2^{-16}. With the expeciont of the passing variable when considered individually, which has a p-value of 0.00375, still well below the significance threshold of 0.05. All of these variable coefficients therefore reject the null hypothesis claiming zero interaction effect and support the postulation that there are significant interaction effects between these variables individually and multiplied.

The F-statistic has a p-value of 2.2^{-16} and therefore suggests that there are in fact synergistic relationships between the predictors and the response.

Numerical Predictive Analysis

Multiple Linear Regression

While it is fair to assume that the more basic and specific player attributes will contribute to the overall player rating or perhaps their value/wage etc. It would be interesting to understand

which of these variables are most involved in the determination of a player's overall rating. To discover this we will attempt to use the individual player attribute ratings to predict a player's overall rating and in doing so determine the set of attributes best accomplished to determine or calculate a player's overall rating.

Hypothesis: There is a statistically significant relationship between the response variable and the selected predictor variables.

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = 0$

$H_1: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} \neq 0$

```
> overall.lm <- lm(overall ~ pace + shooting + passing + dribbling
+
  defending + physical + attacking + skill + movement + power +
  mentality + tackling + goalkeeping, data = dataset)

> summary(overall.lm)

  Call:
  lm(formula = overall ~ pace + shooting + passing + dribbling +
      defending + physical + attacking + skill + movement + power +
      mentality + tackling + goalkeeping, data = dataset)

  Residuals:
       Min       1Q   Median       3Q      Max
  -11.5115  -2.1604  -0.1911   1.9531  16.2085

  Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept) 12.508789   0.305923  40.889  < 2e-16 ***
  pace        -0.028196   0.005918  -4.764 1.91e-06 ***
  shooting    -0.179769   0.007849 -22.904  < 2e-16 ***
  passing     -0.073074   0.009217  -7.928 2.37e-15 ***
  dribbling    0.237433   0.007644  31.061  < 2e-16 ***
  defending    0.226129   0.008760  25.815  < 2e-16 ***
  physical     0.145549   0.007356  19.786  < 2e-16 ***
  attacking    0.482879   0.009373  51.516  < 2e-16 ***
  skill        0.001757   0.009083   0.193   0.8466
  movement     0.081188   0.009732   8.343  < 2e-16 ***
  power        0.050786   0.010145   5.006 5.62e-07 ***
  mentality    0.080305   0.009380   8.561  < 2e-16 ***
  tackling    -0.127019   0.006945 -18.289  < 2e-16 ***
  goalkeeping  0.031088   0.013394   2.321   0.0203 *
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 3.153 on 16847 degrees of freedom
    (2083 observations deleted due to missingness)
  Multiple R-squared:  0.7918,  Adjusted R-squared:  0.7917
  F-statistic:  4929 on 13 and 16847 DF,  p-value: < 2.2e-16
```

Seemingly every rating attribute within the game, except for skill curiously, can be considered an effective predictor of the overall rating as a response variable. They each have small p-values allowing for rejection of the null hypothesis, indicating this is unlikely that the coefficients calculated have occurred due to chance. As this model uses multiple predictors

however the p-value associated with the F-Statistic must be checked in order to determine if there is at least one statistically relevant variable in the bunch.

The small p-value associated with the F-Statistic indicates that there is at least one statistical variable and that there is likely a synergy existing between these predictors. The R-squared value estimates that 79% of variance in the model and it's predictor variables can be attributed to variance in the response variable further supporting statistical relevance.

Outliers

```
> plot(overall.lm)
```

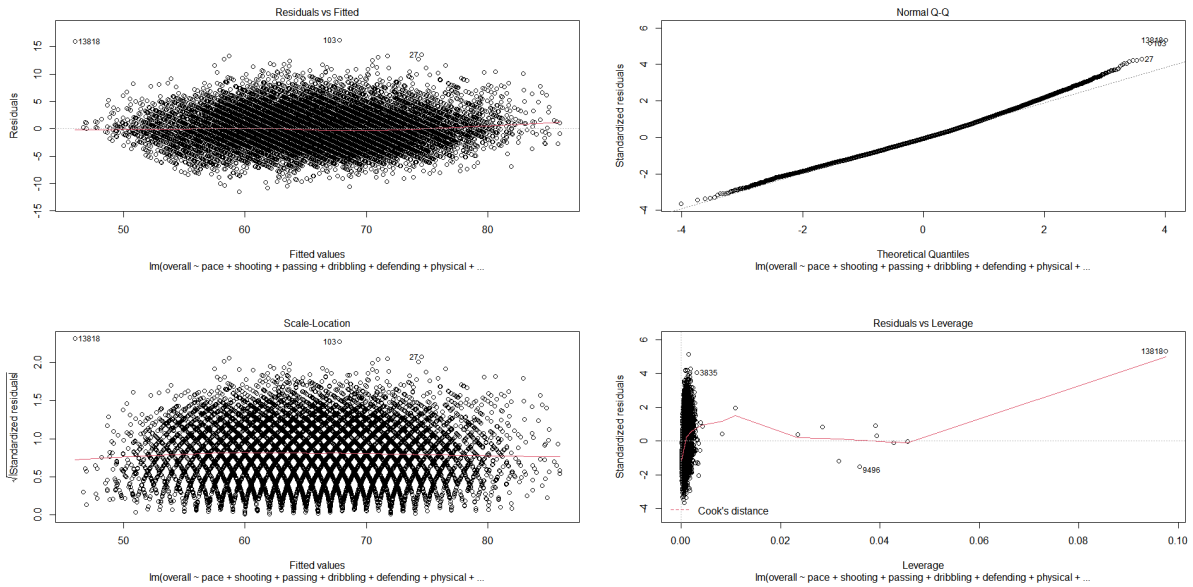

Figure 18 - Residual, fitted, error and leverage plots

It is immediately obvious from the upper left plot of 'Residuals vs Fitted,' that this model has absolutely no issue with fit when it comes to linearity, which will allow us to move forward with this linear method while performing predictive analysis. It also appears upon observing this plot that there may be one or two outliers in this dataset, observations 13818 and 103 specifically. Viewing the lower right graph of 'Residuals vs Leverage,' shows us the high leverage points within the dataset which again somewhat supports the understanding that there are outliers within the dataset that would benefit the model by being removed. Observation 13818 is immediately identifiable again in this instance because it is causing distortion shown by the red mean line. The other points while a little high points are not causing any significant distortion to the overall model.

On inspection it was discovered that the player at observation index 13818 was listed as a goalkeeper, yet had attribute rating in the general player variables. This was likely a player who played outfield as well as in goal, or there may have been some other error in the way the data was cleaned for the position variable, but due to the higher leverage this is seemingly unlikely, Therefore the point will be removed from the dataset for this prediction analysis.

```
> ovr.predictors <- dataset[-13818, c(8, 22:34)]
> overall.lm <- lm(overall ~ ., data = ovr.predictors)
> summary(overall.lm)

  Call:
```

```
lm(formula = overall ~ ., data = ovr.predictors)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4667  -2.1612  -0.1877   1.9525  16.2853

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.604536   0.306199  41.164  < 2e-16 ***
pace        -0.029017   0.005916  -4.905 9.42e-07 ***
shooting    -0.186821   0.007953 -23.492  < 2e-16 ***
passing     -0.080868   0.009325  -8.672  < 2e-16 ***
dribbling    0.233793   0.007668  30.489  < 2e-16 ***
defending    0.220625   0.008813  25.034  < 2e-16 ***
physical     0.142258   0.007376  19.287  < 2e-16 ***
attacking    0.491516   0.009504  51.715  < 2e-16 ***
skill        0.007817   0.009146   0.855    0.393
movement     0.082931   0.009729   8.524  < 2e-16 ***
power        0.055492   0.010175   5.454 5.00e-08 ***
mentality    0.085121   0.009415   9.041  < 2e-16 ***
tackling    -0.123387   0.006973 -17.696  < 2e-16 ***
goalkeeping  0.015590   0.013695   1.138    0.255
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.15 on 16846 degrees of freedom
  (2083 observations deleted due to missingness)
Multiple R-squared:  0.7922,  Adjusted R-squared:  0.792
F-statistic:  4939 on 13 and 16846 DF,  p-value: < 2.2e-16
```

The effect of this refitting is immediately clear to see, with the goalkeeping variable losing its significance.

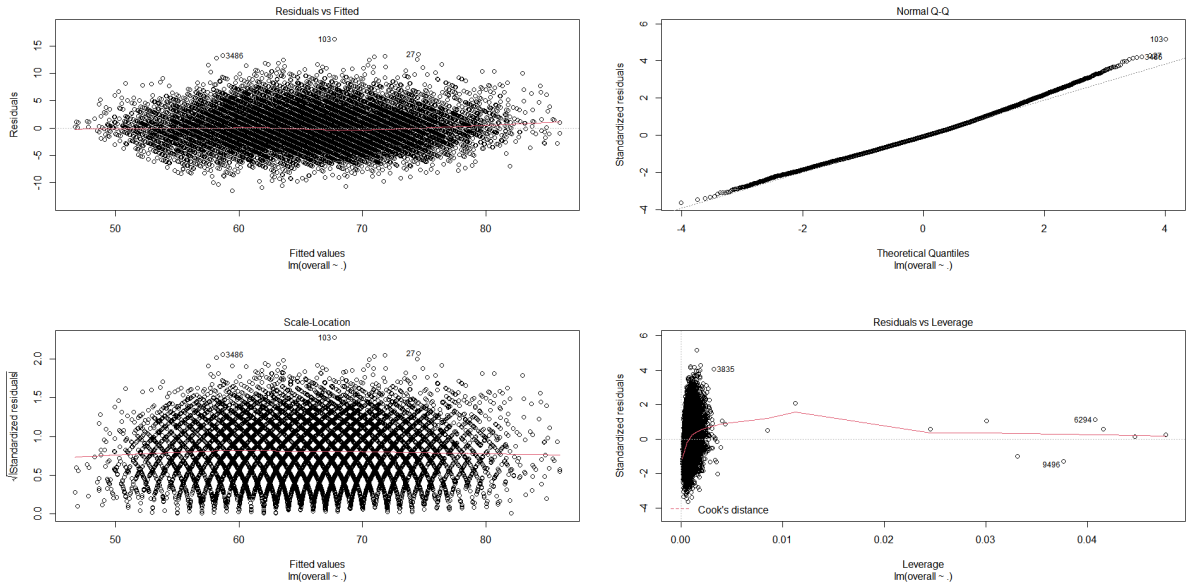

Figure 19 - Residual, fitted, error and leverage plots (refitted)

The high leverage point which was skewing the model is no longer present.

Best Subset Selection

```
> regfit <- regsubsets(overall ~ ., ovr.predictors)
> summary(regfit)

    Subset selection object
    Call: regsubsets.formula(overall ~ ., ovr.predictors)
    13 Variables  (and intercept)

                Forced in Forced out
    pace            FALSE      FALSE
    shooting        FALSE      FALSE
    passing         FALSE      FALSE
    dribbling       FALSE      FALSE
    defending       FALSE      FALSE
    physical        FALSE      FALSE
    attacking       FALSE      FALSE
    skill           FALSE      FALSE
    movement        FALSE      FALSE
    power           FALSE      FALSE
    mentality       FALSE      FALSE
    tackling        FALSE      FALSE
    goalkeeping     FALSE      FALSE

    1 subsets of each size up to 8
    Selection Algorithm: exhaustive

             pace shooting passing dribbling defending
    1  ( 1 ) " "  " "      " "     " "       " "
    2  ( 1 ) " "  " "      " "     " "       "*"
    3  ( 1 ) " "  " "      " "     " "       "*"
    4  ( 1 ) " "  " "      " "     "*"       "*"
    5  ( 1 ) " "  "*"      " "     "*"       "*"
    6  ( 1 ) " "  "*"      " "     "*"       "*"
    7  ( 1 ) " "  "*"      " "     "*"       "*"
    8  ( 1 ) " "  "*"      "*"     "*"       "*"

             physical attacking skill movement power
    1  ( 1 ) " "      " "       " "   " "      " "
    2  ( 1 ) " "      "*"       " "   " "      " "
    3  ( 1 ) "*"      "*"       " "   " "      " "
    4  ( 1 ) "*"      "*"       " "   " "      " "
    5  ( 1 ) "*"      "*"       " "   " "      " "
    6  ( 1 ) "*"      "*"       " "   " "      " "
    7  ( 1 ) "*"      "*"       " "   "*"      " "
    8  ( 1 ) "*"      "*"       " "   "*"      " "

             mentality tackling goalkeeping
    1  ( 1 ) "*"       " "      " "
    2  ( 1 ) " "       " "      " "
    3  ( 1 ) " "       " "      " "
    4  ( 1 ) " "       " "      " "
    5  ( 1 ) " "       " "      " "
    6  ( 1 ) " "       "*"      " "
    7  ( 1 ) " "       "*"      " "
    8  ( 1 ) " "       "*"      " "
> summary(regfit)$rsq

    [1] 0.6476070 0.7154149 0.7394371 0.7742743
    [5] 0.7826657 0.7881839 0.7897168 0.7902774
```

```
> plot(summary(regfit)$rss,    xlab="# Preds", ylab="RSS",    type = "b")
> plot(summary(regfit)$cp,     xlab="# Preds", ylab="Cp",     type = "b")
> plot(summary(regfit)$bic,    xlab="# Preds", ylab="BIC",    type = "b")
> plot(summary(regfit)$adjr2, xlab="# Preds", ylab="Adj R2", type = "b")
```

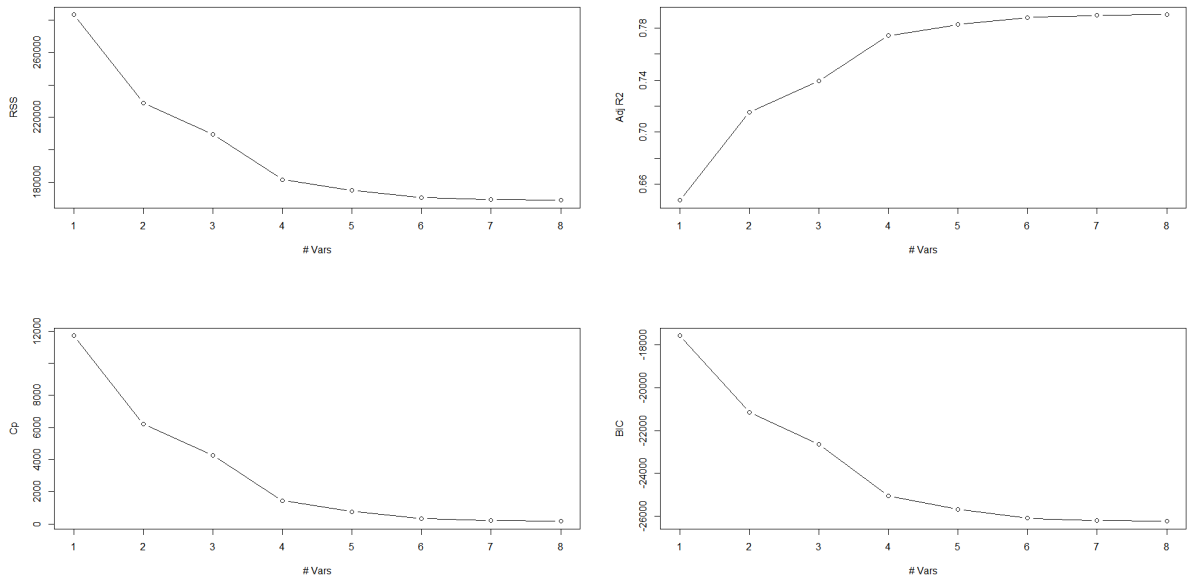

Figure 20 - Best subset selection confidence plots

```
> which.min(summary(regfit)$rss)
  [1] 8
> which.max(summary(regfit)$adjr2)
  [1] 8
> which.min(summary(regfit)$cp)
  [1] 8
> which.min(summary(regfit)$bic)
  [1] 8
```

It's abundantly obvious that the subset containing eight variable components is the version to go with here with all five error tests agreeing.

The variables within the chosen dataset are, shooting, passing, dribbling, defending, physical, attacking, movement and tackling. This is very pleasant to observe as this is a nice all around selection of attributes contributing to an overall rating which doesn't unproportionately or unfairly award any one particular attribute higher overall ratings on average.

Hypothesis: There is a statistically significant relationship between the response variable and the selected predictor variables.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$

$$H_1: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 \neq 0$$

```
> overall <- lm(overall ~ shooting + passing + dribbling +
  defending + physical + attacking + movement + tackling,
  data = ovr.predictors)

> summary(overall)
```

```
Call:
lm(formula = overall ~ shooting + passing + dribbling +
defending + physical + attacking + movement + tackling,
data = ovr.predictors)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5977  -2.1666  -0.1893   1.9677  15.5519

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.766224   0.276895  46.105  < 2e-16 ***
shooting    -0.134198   0.005323 -25.212  < 2e-16 ***
passing     -0.041660   0.006208  -6.711 1.99e-11 ***
dribbling    0.246126   0.006863  35.860  < 2e-16 ***
defending    0.261336   0.008174  31.972  < 2e-16 ***
physical     0.184676   0.003595  51.366  < 2e-16 ***
attacking    0.474259   0.009253  51.256  < 2e-16 ***
movement     0.048839   0.004643  10.519  < 2e-16 ***
tackling    -0.139413   0.006876 -20.274  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.164 on 16851 degrees of freedom
  (2083 observations deleted due to missingness)
Multiple R-squared:  0.7903,  Adjusted R-squared:  0.7902
F-statistic:  7937 on 8 and 16851 DF,  p-value: < 2.2e-16
```

The p-values for each attribute are very small indicating that there is statistical significance between them and the response variable through the rejection of the null hypothesis which states that no significance exists between any predictor and the response.

As this is a multiple linear regression with multiple predictors the F-Statistic must be validated in order to determine a true relationship between the predictors and the response. The p-value for the F-Statistic too is very small at 2.2^{-16} indicating that at least one of these predictors is relevant to the response.

Formula

This formula has been generated using the predictor coefficients calculated in the linear regression above. It can be used to predict the value of overall based on the values presented.

```
overall = 12.766224 - 0.134198(shooting)  - 0.041660(passing)   +
                      0.246126(dribbling) + 0.261336(defending) +
                      0.184676(physical)  + 0.474259(attacking) +
                      0.048839(movement)  - 0.139413(tackling)
```

Validation Set Approach

The first step in completing cross validation is to remove all of the goalkeepers from the dataset, until now the summary statistics generated have omitted these observations due to the NA values present in some of the rating attributes. But this becomes more of an issue here due to the use of manual calculations.

```
> no.gk <- dataset[dataset$position != "GK", ]
```

```
> ovr.predictors <- no.gk[, c(8, 22:34)]

> library(caret)
> ovr.part <- createDataPartition(y = ovr.predictors$overall,
  p = 0.7, list = FALSE)
> training <- ovr.predictors[ovr.part, ]
> test <- ovr.predictors[-ovr.part, ]

> lm.fit <- lm(overall ~ shooting + passing + dribbling +
defending + physical + attacking + movement + tackling, data =
training)

> summary(lm.fit)$r.squared

  [1] 0.7918071
```

This indicates that 79% of the variability from predictors in the training model can be accounted for by the variance in the response.

```
> mean((test$overall - predict(lm.fit, test))^2)

  [1] 10.03967
```

This is an error margin of ten points which is phenomenal for a prediction model. It indicates that the model can predict a player's overall rating with near certainty.
It can be run again on a different validation set to show further accuracy.

```
> sqrt(10.03967)

  [1] 3.168544
```

Rooted Mean Squared Error shows a more neutral value for mean error. At three rating points difference between observed and predicted overall rating this shows that the model training model produced is quite accurate.

New partition and therefore new validation and training sets are generated;

```
> ovr.part <- createDataPartition(y = ovr.predictors$overall,
  p = 0.7, list = FALSE)
> training <- ovr.predictors[ovr.part, ]
> test <- ovr.predictors[-ovr.part, ]
> lm.fit <- lm(overall ~ shooting + passing + dribbling + defending
+
  physical + attacking + movement + tackling, data = training

> summary(lm.fit)$r.squared

  [1] 0.7900919

> mean((test$overall - predict(lm.fit, test))^2)

  [1] 10.18449

> sqrt(10.18449)

  [1] 3.191315
```

Running this testing method multiple times returned similar values from each iteration, further proving the accuracy of the model.

Averages were calculated based on multiple iterations with;

RMSE = 3.18 and R-squared = 0.79

Leave-One-Out Cross-Validation

```
> ovr.loocv <- train(overall ~ shooting + passing + dribbling +
  defending + physical + attacking + movement + tackling,
  data = ovr.predictors, method = "lm", trControl = loocv)

> print(ovr.loocv)

  Linear Regression

  16860 samples
      8 predictor

  No pre-processing
  Resampling: Leave-One-Out Cross-Validation
  Summary of sample sizes: 16859, 16859, 16859, 16859, 16859, ...
  Resampling results:

    RMSE      Rsquared   MAE
    3.164706  0.7900355  2.493834

  Tuning parameter 'intercept' was held constant at a value of TRUE
```

Root mean squared error (RMSE) measures average differences between the predictions made by the model and the actual observations. The lower the RMSE, the more closely a model can predict the actual observations.

R-squared is a measure of correlation between predictions made by the model and the actual observations. The higher the R-squared, the more closely a model can predict actual observations.

Mean absolute error (MAE) is the average absolute difference between the predictions made by the model and the actual observations. The lower the MAE, the more closely a model can predict the actual observations.

The values presented here are similar to those estimated in the validation set approach above. Three rating points for RMSE and accuracy probability of 0.79 from R-squared, comfortably above the common threshold of 0.7. All of this points toward accuracy of the model at predicting overall player rating based on these specific predictor variables.

```
> library(glmnet)
> train.mtrx <- model.matrix(overall ~ ., data = training)
> test.mtrx <- model.matrix(overall ~ ., data = test)
```

Ridge Regression

```
> cv.ridge <- cv.glmnet(train.mtrx, training$overall, alpha = 0,
  lambda = grid, thresh = 1e-12)
```

```
> ridge <- cv.ridge$lambda.min

  [1] 0.01

> ridge.fit <- glmnet(train.mtrx, training$overall, alpha = 0,
lambda = grid, thresh = 1e-12)

> predict.ridge <- predict(ridge.fit, s = ridge, newx = test.mtrx)

> predict(ridge.fit, s = ridge, type = "coefficients")

  15 x 1 sparse Matrix of class "dgCMatrix"
                          1
  (Intercept) 12.492078218
  (Intercept)  .
  pace        -0.026025468
  shooting    -0.182275474
  passing     -0.076430813
  dribbling    0.229759902
  defending    0.210627462
  physical     0.142182075
  attacking    0.478544863
  skill        0.008024041
  movement     0.079503432
  power        0.062290630
  mentality    0.087582950
  tackling    -0.115218020
  goalkeeping  0.028739208
```

Formula

```
Overall = 12.492078218 - 0.026025468(pace) - 0.182275474(shooting) -
                         0.076430813(passing) + 0.229759902(dribbling) +
                         0.210627462(defending) + 0.142182075(physical)
                         0.478544863(attacking) + 0.008024041(skill)
                         0.079503432(movement) + 0.062290630(power)
                         0.087582950(mentality) - 0.115218020(tackling) +
                         0.028739208(goalkeeping)
```

```
> mean((test$overall - predict.ridge)^2)

  [1] 10.10868

> sqrt(10.10868)

  [1] 3.179415
```

The MSE and RMSE values are very similar to the values cross validated using the linear model and best subset selection. The values from the ridge model are slightly higher however and therefore less accurate in a statistical sense.

Lasso

```
 > cv.lasso <- cv.glmnet(train.mtrx, training$overall, alpha = 1,
   lambda = grid, thresh = 1e-12)

 > lasso <- cv.lasso1$lambda.min

   [1] 0.01

 > lasso.fit <- glmnet(train.mtrx, training$overall, alpha = 1,
```

```
   lambda = grid, thresh = 1e-12)

> predict.lasso <- predict(lasso.fit, s = lasso, newx = test.mtrx)

> predict(lasso.fit, s = lasso, type = "coefficients")

 15 x 1 sparse Matrix of class "dgCMatrix"
                       1
 (Intercept) 12.71488596
 (Intercept)  .
 pace        -0.01689496
 shooting    -0.16716935
 passing     -0.05838096
 dribbling    0.22739254
 defending    0.19864904
 physical     0.15017853
 attacking    0.46712653
 skill        .
 movement     0.06678587
 power        0.05126424
 mentality    0.07967496
 tackling    -0.10239236
 goalkeeping  0.02417459
```

Formula

```
overall = 12.71488596 - 0.01689496(pace) - 0.16716935(shooting) -
                        0.05838096(passing) + 0.22739254(dribbling) +
                        0.19864904(physical) + 0.46712653(attacking) +
                        0.06678587(movement) + 0.05126424(power) +
                        0.07967496(mentality) - 0.10239236(tackling) +
                        0.02417459(goalkeeping)
```

```
> mean((test$overall - predict.lasso)^2)

  [1] 10.12081

> sqrt(10.12081)

  [1] 3.181322
```

The lasso method makes use of far more of the variables then were used in the best subset selection. Only truly reducing one of the variables to zero, skill which was statistically insignificant anyway, along with an intercept. With the additional other variables only producing coefficients below 0.05 (pace and goalkeeping). Each method used has produced very similar MSE's suggesting that no method is particularly better than any other.

The best subset selection method for this linear regression appears to offer the greatest accuracy and as it uses fewer variables than both the lasso and ridge this helps the model avoid the issue of overfitting. It produced the lowest RMSE value indicated through the use of LOOCV and had a supporting R-squared probability of 0.79. This model should be used when predicting player overall rating coefficients.

Categorical Predictive Analysis

Considering that this dataset was scraped and released at the same time as the release of the game, it would be interesting to try and predict which players in the game would be likely to be loaned out throughout the season and as such predict which different attributes or variables are likely to affect such an occurrence.

The subset loan.predictors will lose some of the variables from the total 34 that exist in the no.gk subset that should prove to have little effect on the predictive analysis of this data. The name, jersey number, contract end date, goalkeeping statistics have all been removed.

```
> loan.predictors <- no.gk[,-c(1, 19, 21, 34)]
```

Best Subset Selection

Initial attempts to run the best subset validation on the 30 variable loan.predictors subset ran into issues relating to computational time and workload. After several attempts (one which was allowed to run for 24hrs), the method was abandoned due to the scale of the variables being accessed and the computational time required to complete.

The data subset was further reduced after realising that the factor values nationality, club and league were adding hundreds upon hundreds of factors variables to the analysis and so were making the selection process exceedingly difficult. These large factor values were completely removed because while it may have been interesting to understand whether or not a certain league or club was more or less likely to make a player to become loaned. This information could much more easily be interpreted from the use of a contingency table or some other fair less computational heavy analysis method.

```
> loan.predictors <- loan.predictors[,-c(4:6)]
> library(leaps)
> regfit.2 <- regsubsets(loan ~ ., loan.predictors, really.big =
T)
> summary(regfit.2)

 Subset selection object
 Call: regsubsets.formula(loan ~ ., loan.predictors, really.big = T)
 47 Variables  (and intercept)
                Forced in Forced out
 age                FALSE      FALSE
 height             FALSE      FALSE
 weight             FALSE      FALSE
 overall            FALSE      FALSE
 potential          FALSE      FALSE
 value              FALSE      FALSE
 wage               FALSE      FALSE
 positionCB         FALSE      FALSE
 positionCDM        FALSE      FALSE
 positionCF         FALSE      FALSE
 positionCM         FALSE      FALSE
 positionLB         FALSE      FALSE
 positionLM         FALSE      FALSE
 positionLW         FALSE      FALSE
 positionLWB        FALSE      FALSE
 positionRB         FALSE      FALSE
 positionRM         FALSE      FALSE
 positionRW         FALSE      FALSE
 positionRWB        FALSE      FALSE
 positionST         FALSE      FALSE
 footRight          FALSE      FALSE
 reputation2        FALSE      FALSE
 reputation3        FALSE      FALSE
 reputation4        FALSE      FALSE
 reputation5        FALSE      FALSE
 weakfoot2          FALSE      FALSE
 weakfoot3          FALSE      FALSE
 weakfoot4          FALSE      FALSE
```

```
weakfoot5           FALSE      FALSE
attack_wrLow        FALSE      FALSE
attack_wrMedium     FALSE      FALSE
defend_wrLow        FALSE      FALSE
defend_wrMedium     FALSE      FALSE
clause              FALSE      FALSE
pace                FALSE      FALSE
shooting            FALSE      FALSE
passing             FALSE      FALSE
dribbling           FALSE      FALSE
defending           FALSE      FALSE
physical            FALSE      FALSE
attacking           FALSE      FALSE
skill               FALSE      FALSE
movement            FALSE      FALSE
power               FALSE      FALSE
mentality           FALSE      FALSE
tackling            FALSE      FALSE
positionGK          FALSE      FALSE
1 subsets of each size up to 9
Selection Algorithm: exhaustive
         age height weight overall potential value wage
1  ( 1 ) "*" " "    " "    " "     " "       " "   " "
2  ( 1 ) "*" "*"    " "    " "     " "       " "   " "
3  ( 1 ) "*" "*"    "*"    " "     " "       " "   " "
4  ( 1 ) "*" "*"    "*"    "*"     " "       " "   " "
5  ( 1 ) "*" "*"    "*"    "*"     "*"       " "   " "
6  ( 1 ) "*" "*"    "*"    "*"     "*"       "*"   " "
7  ( 1 ) "*" "*"    "*"    "*"     "*"       "*"   "*"
8  ( 1 ) "*" "*"    "*"    "*"     "*"       "*"   "*"
9  ( 1 ) "*" "*"    "*"    "*"     "*"       "*"   "*"
         positionCB positionCDM positionCF positionCM positionGK
1  ( 1 ) " "        " "         " "        " "        " "
2  ( 1 ) " "        " "         " "        " "        " "
3  ( 1 ) " "        " "         " "        " "        " "
4  ( 1 ) " "        " "         " "        " "        " "
5  ( 1 ) " "        " "         " "        " "        " "
6  ( 1 ) " "        " "         " "        " "        " "
7  ( 1 ) " "        " "         " "        " "        " "
8  ( 1 ) "*"        " "         " "        " "        " "
9  ( 1 ) "*"        "*"         " "        " "        " "
         positionLB positionLM positionLW positionLWB positionRB
1  ( 1 ) " "        " "        " "        " "         " "
2  ( 1 ) " "        " "        " "        " "         " "
3  ( 1 ) " "        " "        " "        " "         " "
4  ( 1 ) " "        " "        " "        " "         " "
5  ( 1 ) " "        " "        " "        " "         " "
6  ( 1 ) " "        " "        " "        " "         " "
7  ( 1 ) " "        " "        " "        " "         " "
8  ( 1 ) " "        " "        " "        " "         " "
9  ( 1 ) " "        " "        " "        " "         " "
         positionRM positionRW positionRWB positionST footRight
1  ( 1 ) " "        " "        " "         " "        " "
2  ( 1 ) " "        " "        " "         " "        " "
3  ( 1 ) " "        " "        " "         " "        " "
4  ( 1 ) " "        " "        " "         " "        " "
5  ( 1 ) " "        " "        " "         " "        " "
6  ( 1 ) " "        " "        " "         " "        " "
7  ( 1 ) " "        " "        " "         " "        " "
8  ( 1 ) " "        " "        " "         " "        " "
9  ( 1 ) " "        " "        " "         " "        " "
```

```
          reputation2 reputation3 reputation4 reputation5
1  ( 1 ) " "         " "         " "         " "
2  ( 1 ) " "         " "         " "         " "
3  ( 1 ) " "         " "         " "         " "
4  ( 1 ) " "         " "         " "         " "
5  ( 1 ) " "         " "         " "         " "
6  ( 1 ) " "         " "         " "         " "
7  ( 1 ) " "         " "         " "         " "
8  ( 1 ) " "         " "         " "         " "
9  ( 1 ) " "         " "         " "         " "
          weakfoot2 weakfoot3 weakfoot4 weakfoot5 attack_wrLow
1  ( 1 ) " "       " "       " "       " "       " "
2  ( 1 ) " "       " "       " "       " "       " "
3  ( 1 ) " "       " "       " "       " "       " "
4  ( 1 ) " "       " "       " "       " "       " "
5  ( 1 ) " "       " "       " "       " "       " "
6  ( 1 ) " "       " "       " "       " "       " "
7  ( 1 ) " "       " "       " "       " "       " "
8  ( 1 ) " "       " "       " "       " "       " "
9  ( 1 ) " "       " "       " "       " "       " "
          attack_wrMedium defend_wrLow defend_wrMedium clause pace
1  ( 1 ) " "             " "          " "             " "    " "
2  ( 1 ) " "             " "          " "             " "    " "
3  ( 1 ) " "             " "          " "             " "    " "
4  ( 1 ) " "             " "          " "             " "    " "
5  ( 1 ) " "             " "          " "             " "    " "
6  ( 1 ) " "             " "          " "             " "    " "
7  ( 1 ) " "             " "          " "             " "    " "
8  ( 1 ) " "             " "          " "             " "    " "
9  ( 1 ) " "             " "          " "             " "    " "
          shooting passing dribbling defending physical attacking
1  ( 1 ) " "      " "     " "       " "       " "      " "
2  ( 1 ) " "      " "     " "       " "       " "      " "
3  ( 1 ) " "      " "     " "       " "       " "      " "
4  ( 1 ) " "      " "     " "       " "       " "      " "
5  ( 1 ) " "      " "     " "       " "       " "      " "
6  ( 1 ) " "      " "     " "       " "       " "      " "
7  ( 1 ) " "      " "     " "       " "       " "      " "
8  ( 1 ) " "      " "     " "       " "       " "      " "
9  ( 1 ) " "      " "     " "       " "       " "      " "

 skill movement power mentality tackling
1  ( 1 ) " "   " "      " "   " "       " "
2  ( 1 ) " "   " "      " "   " "       " "
3  ( 1 ) " "   " "      " "   " "       " "
4  ( 1 ) " "   " "      " "   " "       " "
5  ( 1 ) " "   " "      " "   " "       " "
6  ( 1 ) " "   " "      " "   " "       " "
7  ( 1 ) " "   " "      " "   " "       " "
8  ( 1 ) " "   " "      " "   " "       " "
9  ( 1 ) " "   " "      " "   " "       " "
```

```
> summary(regfit.2)$rsq

 [1] NaN NaN NaN NaN NaN NaN NaN NaN NaN

> plot(summary(regfit.2)$rss,   xlab="# Preds", ylab="RSS",
type = "b")
```

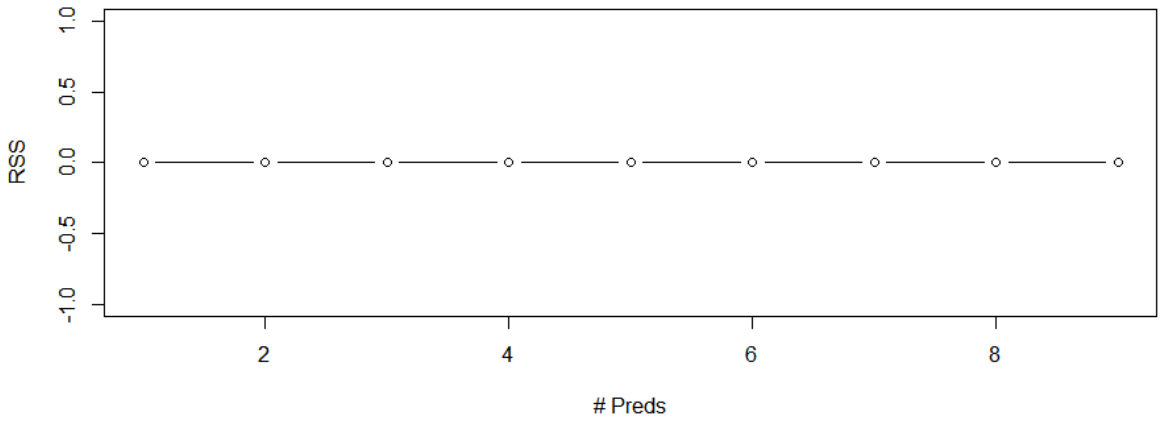

Figure 21 - RSS vs Number of Predictors (inconclusive)

```
> which.min(summary(regfit.2)$rss)

  [1] 1
```

```
> plot(summary(regfit.2)$cp,    xlab="# Preds", ylab="Cp",     type = "b")
> plot(summary(regfit.2)$bic,   xlab="# Preds", ylab="BIC",    type = "b")
> plot(summary(regfit.2)$adjr2, xlab="# Preds", ylab="Adj R2", type = "b")

Error in plot.window(...) : need finite 'ylim' values
In addition: Warning messages:
1: In min(x) : no non-missing arguments to min; returning Inf
2: In max(x) : no non-missing arguments to max; returning -Inf
```

```
> which.max(summary(regfit.2)$adjr2)

  integer(0)
```

```
> which.min(summary(regfit.2)$cp)

  integer(0)
```

```
> which.min(summary(regfit.2)$bic)

  integer(0)
```

It's clear that this method of selecting a subset of significant interaction variables has failed to return any usable data. Instead the lasso method will be implemented on the entire subset that had originally hoped to be assessed using this method (minus the clause variable due to the presence of over 600 NA values). This shrinkage method should return a smaller subset of interrelated variables with which to move forward with analysis.

The ridge regression will not be used for this analysis due to the sheer volume of variables being tested and the number of values that would be returned in order to generate a usable formula to calculate the probability of a player being loaned, because ridge regression does not shrink variables to zero as does the lasso.

Lasso

```
> library(glmnet)
```

```
> mtrx <- model.matrix(loan ~ ., data = loan.predictors)
> grid <- 10 ^ seq(4, -2, length = 100)

> cv.lasso.full <- cv.glmnet(mtrx, loan.predictors$loan, alpha =
1,
  lambda = grid, thresh = 1e-12, family = "binomial")

> lasso.full <- cv.lasso.full$lambda.min

> lasso.fit.full <- glmnet(mtrx, loan.predictors$loan, alpha = 1,
  lambda = grid, thresh = 1e-12, family = "binomial")

> predict(lasso.fit.full, s = lasso.full, type = "coefficients")
```

Due to the number of variables used in this lasso, only the variables that generated a statistically relevant interaction value are shown. All other variables there were shrunk to zero are omitted. The full list can be viewed in the appendix.

```
942 x 1 sparse Matrix of class "dgCMatrix"
                                                 1
(Intercept)                           -3.93264448
(Intercept)                            .
age                                   -0.03241422
clubDefensa y Justicia                 0.39245484
clubSangju Sangmu FC                   3.31612397
leagueSpanish Segunda DivisiÃ³n        0.30836288
potential                              0.02171299
```

The use of lasso here was far more effective than was the best subset selection method. This was as expected. The lasso returned a very interesting set of interaction variables. Which was also expected. This subset of statistically relevant variables contained factors from the club and league variables, with only two numerical variables, age and potential. Two seemingly obvious, but important to confirm, variables. Age with a negative interaction indicating that the younger a player is the more likely they are to get loaned out and the potential rating without the presence of the overall rating indicating that players with high potentials but lower relative current overall ratings that need playing time to develop are more likely to be loaned out to other teams.

The most interesting observation from this lasso is that one club and one league in particular are a huge indication of a player being loaned out, Sangju Sangmu FC with a coefficient value of 3.24818630, and Spanish Segunda División with a value of 0.30836288, compared to the values of below one for each of the other variables returned.

For ease of analysis these variables will be used as the "best" subset between different methods of regression analysis. This will help in determining which model provides the greatest level of accuracy.

Probability Formula

P(loan) =

$$\frac{e^{-3.9326448 - 0.03241422(age) + 0.39245484(club1) + 0.31612397(club2) + 0.30836288(league) + 0.02171299(potential)}}{1 + e^{-3.9326448 - 0.03241422(age) + 0.39245484(club1) + 0.31612397(club2) + 0.30836288(league) + 0.02171299(potential)}}$$

Validation

```
> prob.lasso.full <- predict(cv.lasso.full, newx = test.mtrx,
  s = lasso.full, type = "response")
> pred.lasso.full <- rep("No", nrow(test))
> pred.lasso.full[pred.lasso.full > .5] <- "Yes"
> table(pred.lasso.full, test$loan)

 pred.lasso.full   No  Yes
            No  4849  206
            Yes    0    2

> mean(pred.lasso.full != test$loan)

 [1] 0.04073561
```

This suggests an accurate model with only 4% test error and therefore 96% accuracy. If the confusion matrix is observed the "Yes" column from the test data shows that 206 instances were incorrectly predicted and only two were correctly predicted. This would suggest that the accuracy probability is incorrect or more appropriate is only accurate at predicting "No" instances. Other methods of fitting and validation will be used to test this further, but the lasso model appears to have failed at producing an accurate prediction model.

Dummy variables are now generated in a new dataframe for the purposes of logistic, linear discriminant and quadratic discriminant analysis.

```
> dummy.club1 <- as.numeric(loan.pred$club == "Defensa y
Justicia")
> dummy.club2 <- as.numeric(loan.pred$club == "Sangju Sangmu FC")
> dummy.league <- as.numeric(loan.pred$league ==
  "Spanish Segunda DivisiÃ³n")

> loan <- cbind(loan.predictors$age, dummy.club1, dummy.club2,
  dummy.league, loan.predictors$potential, loan.predictors$loan)

> str(loan)

 'data.frame': 16860 obs. of  6 variables:
  $ age         : num  33 35 31 28 29 21 28 28 28 32 ...
  $ dummy.club1 : num  0 0 0 0 0 0 0 0 0 0 ...
  $ dummy.club2 : num  0 0 0 0 0 0 0 0 0 0 ...
  $ dummy.league: num  0 0 0 0 0 0 0 0 0 0 ...
  $ potential   : num  93 92 91 91 91 95 91 90 90 89 ...
  $ loan        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 ...
```

Logistic Regression

```
> loan.glm <- glm(loan ~ ., loan, family = "binomial")
> summary(loan.glm)

 Call:
 glm(formula = loan ~ ., family = "binomial", data = loan)

 Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -1.6488  -0.3170  -0.2545  -0.2004   3.0532
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.080187   0.603378  -8.420  < 2e-16 ***
age          -0.087741   0.010122  -8.669  < 2e-16 ***
dummy.club1   2.476299   0.441023   5.615 1.97e-08 ***
dummy.club2   4.881825   0.477686  10.220  < 2e-16 ***
dummy.league  1.131606   0.144915   7.809 5.78e-15 ***
potential     0.054494   0.006744   8.080 6.46e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5793.4  on 16859  degrees of freedom
Residual deviance: 5424.4  on 16854  degrees of freedom
AIC: 5436.4

Number of Fisher Scoring iterations: 6
```

The statistical summary of the logistic regression fit on the loan dataset indicates small p-values for each of the predictor variables. Allowing for rejection of the null hypothesis, which states that no variable has statistical relevance to the response. This offers a conclusion that there is in fact statistical relevance between these variables and the response.

Probability Formula

$$P(loan) = \frac{e^{-5.080187 - 0.087741(age) + 2.476299(club1) + 4.881825(club2) + 1.131606(league) + 0.054494(potential)}}{1 + e^{-5.080187 - 0.087741(age) + 2.476299(club1) + 4.881825(club2) + 1.131606(league) + 0.054494(potential)}}$$

Dataset Partition

```
> library(caret)
> loan.part <- createDataPartition(y = loan$loan, p = 0.7,
  list = FALSE)

> training <- loan.predictors[loan.part, ]
> test <- loan.predictors[-loan.part, ]
```

This method is repeated two more times with different variable names in order to generate three different validation and training subset in order to offer more than one interpretation of this validation method and show greater accuracy in probability estimates. These training and validation sets will be used for each regression model logistic, lad and qda in this analysis.

Validation Set Approach

```
> glm.prob <- predict(loan.glm, test, type = "response")
> glm.pred <- rep("No", length(glm.prob))
> glm.pred[glm.prob > 0.5] <- "Yes"

> table(glm.pred, test$loan)

glm.pred   No  Yes
     No  4848  201
     Yes    1    7
```

```
> mean(glm.pred != test$loan)
[1] 0.03994463
```

The logistic regression model has returned a mean test error of 3.99%, a phenomenal level of accuracy.

Repeating this validation set approach two more times with a different validation set each time by partitioning the data again, will allow us to cross validate with greater accuracy and attempt to show that the accuracy shown in the first iteration wasn't purely due to chance.

```
glm.pred.2   No  Yes
       No  4846  200
       Yes    3    8

[1] 0.04014238
```

```
glm.pred.3   No  Yes
       No  4847  202
       Yes    2    6

[1] 0.04034012
```

The mean of the test errors calculated is 0.04014237667 or 4%, suggesting a high level of accuracy at approximately 96% on average. This proves that the results are repeatable.

It is worth noting however that the number of "No" observations far outweighs the "Yes" alternative. The model appears to be good at predicting "No" observations, but not so good at predicting "Yes" observations correctly. Observing the confusion matrices, of the "Yes" variables the model only correctly predicted six, seven or eight out of the 208 possible values. In each iteration.

Leave-One-Out Cross-Validation

```
> library(e1071)
> model <- train(loan ~ ., data = loan, family = "binomial",
  method = "glm", trControl = trainControl(method = "LOOCV"))

> print(model)

Generalized Linear Model

16860 samples
    5 predictor
    2 classes: 'No', 'Yes'

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 16859, 16859, 16859, 16859, 16859, 16859, ...
Resampling results:

  Accuracy   Kappa
  0.9594899  0.04987987
```

The accuracy rating from the Leave-One-Out Cross-Validation indicates that this model is 95.95% accurate at predicting a player's loan status. Which implies a test error rate of 4.05% which is almost exactly the same as the test error estimate produced using the validation set method for this model.

The Cohen's Kappa value however is troubling, it is very small, suggesting only 4.5% prediction accuracy of the model. This suggests that this high level of accuracy is likely due to chance.

Reviewing the confusion matrices produced in the validation set approach further, it appears that this may be explainable due to the fact that a very large proportion of the observations are not going to contain a "Yes" value in their loan variable column. From earlier analysis it is known that loaned players make up only 17% of the data subset. The model appears to be very good at predicting if a player will not be loaned based on its large majority of the subset observations. Actually paying attention to the confusion matrices it can be seen that the model only predicted less than ten players correctly loaned in each iteration. Whereas it incorrectly indicated that 200 or more players were not loaned when they actually were. The Cohen's Kappa value in this way has disproved the accuracy of this model.

Linear Discriminant Analysis

```
> loan.lda <- lda(loan ~ ., loan)

 Call:
 lda(loan ~ ., data = loan)

 Prior probabilities of groups:
        No        Yes
 0.95877817 0.04122183

 Group means:
          age dummy.club1  dummy.club2 dummy.league potential
 No  25.17897 0.001051655 0.0003711723   0.03031240  71.12403
 Yes 23.22158 0.011510791 0.0273381295   0.08633094  73.63741

 Coefficients of linear discriminants:
                       LD1
 age          -0.07797270
 dummy.club1   7.33630049
 dummy.club2  19.37315960
 dummy.league  1.83973819
 potential     0.05595955
```

Validation Set Approach

```
> lda.pred <- predict(loan.lda, test)
> table(lda.pred$class, test$loan)

         No  Yes
   No  4844  199
   Yes    5    9

> mean(lda.pred$class != test$loan)

 [1] 0.04034012
```

The above approach is repeated to generate three different versions of validation test error data. This should help offer a higher level of accuracy through averaging for this approach.

```
         No  Yes
   No  4840  196
   Yes    9   12

   [1] 0.04053787
```

```
         No  Yes
   No  4842  200
   Yes    7    8

   [1] 0.04093336
```

The mean test error error rate is only 4% indicating that this model like the Logistic model before it is extremely accurate at predicting a player's likelihood to be loaned out.

Upon closer examination of the confusion matrices again they are suggesting that this model is accurate at predicting correctly for instances of "No" in the loan variable but not so accurate at predicting observations with instances of "Yes."

Leave-One-Out Cross-Validation

```
> model.2 <- train(loan ~ ., data = loan, method = "lda",
trControl = trainControl(method = "LOOCV"))
> print(model.2)

 Linear Discriminant Analysis

 16860 samples
     5 predictor
     2 classes: 'No', 'Yes'

 No pre-processing
 Resampling: Leave-One-Out Cross-Validation
 Summary of sample sizes: 16859, 16859, 16859, 16859, 16859, ...
 Resampling results:

   Accuracy   Kappa
   0.9590154  0.0673226
```

Similar results to the logistic regression, a high level of accuracy not supported by the Kappa value which suggests the exact opposite that the model only has an accuracy rate of 6.7% not 95.9%.

Quadratic Discriminant Analysis

```
> loan.qda <- qda(loan ~ ., training)

 Call:
 qda(loan ~ ., data = loan)

 Prior probabilities of groups:
         No        Yes
 0.95877817 0.04122183

 Group means:
           age dummy.club1  dummy.club2 dummy.league potential
 No  25.17897 0.001051655 0.0003711723   0.03031240  71.12403
 Yes 23.22158 0.011510791 0.0273381295   0.08633094  73.63741
```

Validation Set Approach

```
> qda.pred <- predict(loan.qda, test)
> table(qda.pred$class, test$loan)

          No  Yes
   No   4716  181
   Yes   133   27

> mean(qda.pred$class != test$loan)
```

```
[1] 0.06209215
```

The above is repeated two more times to generate three unique validation test error rates in order to offer a greater degree of accuracy to this method.

```
        No  Yes                          No  Yes
  No  4688  177                    No  4708  177
  Yes  161   31                    Yes  141   31

  [1] 0.06683805                   [1] 0.06288313
```

The QDA has calculated a mean test error of 6.4% again a phenomenal accuracy rate but not as accurate as the LDA or Logistic models. Again it appears that the model is incorrectly predicting "Yes" status loan players compared to "No" instances.

```
> model.3 <- train(loan ~ ., data = loan, method = "qda",
trControl = trainControl(method = "LOOCV"))

 Warning message:
 model fit failed for Fold01664: parameter=none Error :
 cannot allocate vector of size 329 Kb

> print(model.3)

 Quadratic Discriminant Analysis

 16860 samples
     5 predictor
     2 classes: 'No', 'Yes'

 No pre-processing
 Resampling: Leave-One-Out Cross-Validation
 Summary of sample sizes: 16859, 16859, 16859, 16859, 16859, ...
 Resampling results:

  Accuracy   Kappa
  0.9341598  0.1015331
```

While the accuracy probability value suggests that this may be the worst model to fit the dataset with when attempting to predict if a player has been loaned. It actually boasts the highest Kappa value which suggests that it is in fact the most accurate of the three models fitted here. This makes sense as the QDA should theoretically be more accurate than the LDA in this case, not being constrained by covariance.

Despite this due to lack of accuracy observed by the Kappa value we fail to reject the null hypothesis. Therefore all three regression models have to be rejected and cannot be used to properly predict loan status due to their lack of true accuracy.

Summary

Findings

Each of the listed expectations were tested for validity with the addition of some unlisted exploratory analysis elements and a categorical predictive analysis. These additions were made on-the-fly based on the data being returned and interpreted from the original expectations.

First of all it was surprising to find that there were not greater levels of correlation between the primary player attributes, namely overall, potential, value, wage etc. The only significant values produced from this group of variables was between value and wage. The ratings values, overall and potential, generate some interesting correlation plots, with visualisations similar to exponential curves. Which makes more sense in hindsight that players that are rated higher would disproportionately be paid way more than their lower rated counterparts. As there is far more interest and capital generated in the bigger league and around the bigger teams than the run of the will mid-tier alternatives.

Again it was surprising to find that the physical player attributes were not normally distributed despite appearing so. Considering the large population size of close to 17,000 observations this was expected to be a given. However it was interesting to attempt to understand which values for the physical attributes were "optimal," for the concept of a good player. The technique used provided a great understanding of optimal age, suprisingly the mean average for height and weight appeared to meet the requirements for these variables.

Further to this, visualising the average ranges for physical characteristics via density graphs provided vital information into where a player should expect their characteristic values to fall in order to become a professional football at all.

The original suggestion that certain positions might disproportionately affect specific variables, value, wage, overall etc. seemed like a fair point to make before any analysis was conducted. The analysis conducted in relation to this suggestion however returned less than favourable results, forcing the conclusion that the player population is in fact well distributed in this game across the entire spectrum of position levels.

Exploratory analysis into the workrate variable factors, returned results pretty much inline with what was expected. While the high workrate level was expected to make up a higher percentage than it actually did. This lower value can likely be explained due to normalisation of this information in a population sense. These levels are likely relative to football players rather than the general population and so the overwhelming majority of medium workrate values actually make sense in this context.

The geo map offered some of the best information of any other analysis conducted in this project. It was expected that the european nations would vastly outnumber the rest of the world in terms of their players per capita and this expectation was mostly proven to be true. Some South American nations with long and proud footballing traditions delivered similar numbers as the mainland european countries, but the most interesting piece of information discovered from this data visualisation was the immensity at which the UK appear to control the game of football in a player count context.

Analysing the foot variable through simple linear regression offered insight into how left footed players are on average rating slightly higher than right footed players, by about one rating point. It was silently hoped that this might be the case as left footed layers make up a far smaller proportion of the player population. Therefore they tend to be better than their counterparts on account of their rareness and natural ability to confuse opponents who expect ball control and movement to be guided by the right side of the body.

The multiple linear regressions conducted were able to show statistical significance of interactions between certain predictor variables on chosen responses. There were no expectations made for this element of analysis as the variables were selected after observing the correlation information.

It was very satisfying to generate an accurate prediction model for a player's overall rating attribute. Similarly accurate models were fitted using Best Subset Selection, Lasso and Ridge Regression. The Best Subset Selection model was decided to be most accurate based on the test error values calculated from different validation methods. Such a model could be used to calculate or generate overall variable values for new virtual players added to the game as the physical player in the real world iterates through seasons in different game modes.

Finally the categorical predictive analysis was the most disappointing aspect of this whole project. Despite initially having thought that an extremely accurate predictive model had been fitted. It was saddening to understand upon deeper validation analysis that this high level of accuracy was itself inaccurate and thus disproved each of the model's effectiveness. Ultimately each model had to be discarded and the conclusion made that prediction of a player loan status was not viable based on the information available with the dataset.

Weaknesses

First point of issue from analysis of this dataset in different manners was the overall null effect of players in the GK or goalkeeper position, of which there were a total of 2083 entries, on much of the analysis conducted. These players lacked any rating values in basic metric statistical attributes namely, pace, shooting, passing, dribbling, defending and physical. The NA values that appear for players whose position is goalkeeper ultimately made their effect on the analysis mute. For this reason it would be pertinent going forward, or in hindsight to have structured or cleaned the database in a manner that was more inclusive of all of the data entries that were available. While the effect was not large as the GK position level only accounted for 2083/18944 = 11% of the total available entries, the elimination of an entire group of players can either be considered detrimental or in this case it may not have hampered and actually helped the overall understanding and analysis that was conducted. Goalkeepers are typically limited in their ratings and ability due to the fact their role on the team is unlike any other, they primarily use their hands rather then their feet while playing and like most positions their skill sets and abilities needs are different from other positions but goalkeepers are so far from every other position that their non-inclusion can actually be effective whether a conscious decision or not.

The presence of NA values in different variables made their use in predictive methods unfeasible at times. Clause for example had to be removed in order to use the lasso method.

There were multiple issues experienced with regard to the sheer scale of the dataset. While attempting to conduct some basic overarching exploratory analysis in that the system would be forced to hang because of the size, in memory, of the data being processed. This happened with the pairs() function call at the very beginning where the entire plot took well over an hour to compute and visualise. This occurred again to a lesser scale on some of the first numerical predictive analysis, but was again a big issue during the categorical predictive analysis. The Best Subset selection regsubsets() call in the categorical predictive analysis took each variable into account and put quite a bit of strain on the machine. The original version was allowed to run for 24 hours at one point and still had not completed its calculations and so was abandoned for a small subset with less factor variables present to reduce the exponential growth of possible variable interactions. This newer smaller version still took over an hour to complete its computation. The LOOCV conducted for each of the Logistic, LDA and QDA models also took several hours to compute.

Future

Further cleaning and optimisation of the dataset in the future would be imperative to producing more statistically significant information. To have to do the assignment over, reducing the total number of variables in the dataset would be suggested, as some of these

were still unnecessary and went unused throughout much the project and when they were used it was not in any meaningful manner other than a check to see if there was significance for sake of posterity.

It would really aid the dataset and the predictive analysis methods that could be conducted, by using different values such as zero, if possible instead of NA. NA variables cannot be analysed properly often being removed from the calculations automatically or requiring manual deletion.

It would have been very interesting to have taken datasets from previous iterations of the game and performed some time series analysis. While the data is readily available it would have been much more difficult to clean and align data from different iterations of the game, especially some older versions where the available attribute variables begin to reduce in number and are less effective at providing useful information in relation to how that player will control in game versus others, due to the limitations in technology at the time of development.

Appendix

Categorial Lasso Shrinkage Coefficient Matrix

```
942 x 1 sparse Matrix of class "dgCMatrix"
                                                 1
(Intercept)                                -3.93264448
(Intercept)                                 .
age                                        -0.03241422
height                                      .
weight                                      .
nationalityAlbania                          .
nationalityAlgeria                          .
nationalityAndorra                          .
nationalityAngola                           .
nationalityAntigua & Barbuda                .
nationalityArgentina                        .
nationalityArmenia                          .
nationalityAruba                            .
nationalityAustralia                        .
nationalityAustria                          .
nationalityAzerbaijan                       .
nationalityBarbados                         .
nationalityBelarus                          .
nationalityBelgium                          .
nationalityBelize                           .
nationalityBenin                            .
nationalityBermuda                          .
nationalityBolivia                          .
nationalityBosnia Herzegovina               .
nationalityBrazil                           .
nationalityBulgaria                         .
nationalityBurkina Faso                     .
nationalityBurundi                          .
nationalityCameroon                         .
nationalityCanada                           .
nationalityCape Verde                       .
nationalityCentral African Republic         .
nationalityChad                             .
nationalityChile                            .
nationalityChina PR                         .
```

```
nationalityChinese Taipei                    .
nationalityColombia                          .
nationalityComoros                           .
nationalityCongo                             .
nationalityCosta Rica                        .
nationalityCroatia                           .
nationalityCuba                              .
nationalityCuracao                           .
nationalityCyprus                            .
nationalityCzech Republic                    .
nationalityDenmark                           .
nationalityDominican Republic                .
nationalityDR Congo                          .
nationalityEcuador                           .
nationalityEgypt                             .
nationalityEl Salvador                       .
nationalityEngland                           .
nationalityEquatorial Guinea                 .
nationalityEritrea                           .
nationalityEstonia                           .
nationalityEthiopia                          .
nationalityFaroe Islands                     .
nationalityFinland                           .
nationalityFrance                            .
nationalityGabon                             .
nationalityGambia                            .
nationalityGeorgia                           .
nationalityGermany                           .
nationalityGhana                             .
nationalityGreece                            .
nationalityGrenada                           .
nationalityGuam                              .
nationalityGuinea                            .
nationalityGuinea Bissau                     .
nationalityGuyana                            .
nationalityHaiti                             .
nationalityHonduras                          .
nationalityHong Kong                         .
nationalityHungary                           .
nationalityIceland                           .
nationalityIndia                             .
nationalityIndonesia                         .
nationalityIran                              .
nationalityIraq                              .
nationalityIsrael                            .
nationalityItaly                             .
nationalityIvory Coast                       .
nationalityJamaica                           .
nationalityJapan                             .
nationalityJordan                            .
nationalityKazakhstan                        .
nationalityKenya                             .
nationalityKorea DPR                         .
nationalityKorea Republic                    .
nationalityKosovo                            .
nationalityLatvia                            .
nationalityLebanon                           .
nationalityLiberia                           .
nationalityLibya                             .
nationalityLiechtenstein                     .
nationalityLithuania                         .
```

```
nationalityLuxembourg                    .
nationalityMacau                         .
nationalityMadagascar                    .
nationalityMalawi                        .
nationalityMalaysia                      .
nationalityMali                          .
nationalityMalta                         .
nationalityMauritania                    .
nationalityMexico                        .
nationalityMoldova                       .
nationalityMontenegro                    .
nationalityMontserrat                    .
nationalityMorocco                       .
nationalityMozambique                    .
nationalityNamibia                       .
nationalityNetherlands                   .
nationalityNew Caledonia                 .
nationalityNew Zealand                   .
nationalityNicaragua                     .
nationalityNiger                         .
nationalityNigeria                       .
nationalityNorth Macedonia               .
nationalityNorthern Ireland              .
nationalityNorway                        .
nationalityPalestine                     .
nationalityPanama                        .
nationalityPapua New Guinea              .
nationalityParaguay                      .
nationalityPeru                          .
nationalityPhilippines                   .
nationalityPoland                        .
nationalityPortugal                      .
nationalityPuerto Rico                   .
nationalityRepublic of Ireland           .
nationalityRomania                       .
nationalityRussia                        .
nationalityRwanda                        .
nationalitySÃ£o TomÃ© & PrÃncipe         .
nationalitySaint Kitts and Nevis         .
nationalitySaint Lucia                   .
nationalitySaudi Arabia                  .
nationalityScotland                      .
nationalitySenegal                       .
nationalitySerbia                        .
nationalitySierra Leone                  .
nationalitySlovakia                      .
nationalitySlovenia                      .
nationalitySouth Africa                  .
nationalitySouth Sudan                   .
nationalitySpain                         .
nationalitySudan                         .
nationalitySweden                        .
nationalitySwitzerland                   .
nationalitySyria                         .
nationalityTanzania                      .
nationalityThailand                      .
nationalityTogo                          .
nationalityTrinidad & Tobago             .
nationalityTunisia                       .
nationalityTurkey                        .
nationalityUganda                        .
```

nationalityUkraine .
nationalityUnited Arab Emirates .
nationalityUnited States .
nationalityUruguay .
nationalityUzbekistan .
nationalityVenezuela .
nationalityWales .
nationalityZambia .
nationalityZimbabwe .
club1. FC Heidenheim 1846 .
club1. FC KÃ¶ln .
club1. FC Kaiserslautern .
club1. FC Magdeburg .
club1. FC NÃ¼rnberg .
club1. FC SaarbrÃ¼cken .
club1. FC Union Berlin .
club1. FSV Mainz 05 .
clubÃ–rebro SK .
clubÃ–stersunds FK .
clubÃ‡aykur Rizespor .
clubAalborg BK .
clubAalesunds FK .
clubAarhus GF .
clubAberdeen .
clubAbha Club .
clubAC Ajaccio .
clubAC Horsens .
clubAC Mineros de Guayana .
clubAC Monza .
clubAcademica Clinceni .
clubAccrington Stanley .
clubAD AlcorcÃ³n .
clubAdelaide United .
clubADO Den Haag .
clubAEK Athens .
clubAFC Wimbledon .
clubAIK .
clubAJ Auxerre .
clubAjax .
clubAl Adalah .
clubAl Ahli .
clubAl Ain FC .
clubAl Faisaly .
clubAl Fateh .
clubAl Fayha .
clubAl Hazem .
clubAl Hilal .
clubAl Ittihad .
clubAl Nassr .
clubAl Raed .
clubAl Shabab .
clubAl Taawoun .
clubAl Wehda .
clubAlanyaspor .
clubAlbacete BP .
clubAlianza Lima .
clubAlways Ready .
clubAmÃ©rica de Cali .
clubAmiens SC .
clubAngers SCO .
clubAntalyaspor .

clubAragua FC .
clubArgentinos Juniors .
clubArsenal .
clubArsenal de SarandÃ .
clubAS Monaco .
clubAS Nancy Lorraine .
clubAS Saint-Ã‰tienne .
clubÅšlÄ…sk WrocÅ‚aw .
clubAston Villa .
clubAstra Giurgiu .
clubAtalanta .
clubAthletic Club de Bilbao .
clubAtiker Konyaspor .
clubAtlÃ©tico Clube Goianiense .
clubAtlÃ©tico de San Luis .
clubAtlÃ©tico Madrid .
clubAtlÃ©tico Mineiro .
clubAtlÃ©tico Nacional .
clubAtlÃ©tico TucumÃ¡n .
clubAtlanta United .
clubAudax Italiano .
clubAZ Alkmaar .
clubBahia .
clubBarcelona Sporting Club .
clubBarnsley .
clubBarrow .
clubBayer 04 Leverkusen .
clubBayern MÃ¼nchen II .
clubBB Erzurumspor .
clubBeÅŸiktaÅŸ JK .
clubBeerschot AC .
clubBeijing Sinobo Guoan FC .
clubBenevento .
clubBirmingham City .
clubBK HÃ¤cken .
clubBlackburn Rovers .
clubBlackpool .
clubBoavista FC .
clubBoca Juniors .
clubBohemian FC .
clubBologna .
clubBolton Wanderers .
clubBorussia Dortmund .
clubBorussia MÃ¶nchengladbach .
clubBotafogo .
clubBournemouth .
clubBrÃ¸ndby IF .
clubBradford City .
clubBrentford .
clubBrescia .
clubBrighton & Hove Albion .
clubBrisbane Roar .
clubBristol City .
clubBristol Rovers .
clubBSC Young Boys .
clubBurnley .
clubBurton Albion .
clubBusan IPark .
clubC.D. CastellÃ³n .
clubCA Osasuna .
clubCÃ¡diz CF .

clubCagliari .
clubCambridge United .
clubCaracas FC .
clubCardiff City .
clubCarlisle United .
clubCD Huachipato .
clubCD LeganÃ©s .
clubCD Lugo .
clubCD MirandÃ©s .
clubCD Nacional .
clubCD Tenerife .
clubCD Tondela .
clubCE Sabadell FC .
clubCearÃ¡ Sporting Club .
clubCeltic .
clubCentral CÃ³rdoba .
clubCentral Coast Mariners .
clubCentro AtlÃ©tico FÃ©nix .
clubCerezo Osaka .
clubCF Fuenlabrada .
clubCFR Cluj .
clubChamois Niortais Football Club .
clubCharlton Athletic .
clubChelsea .
clubCheltenham Town .
clubChicago Fire .
clubChievo Verona .
clubChindia TÃ¢rgoviÅŸte .
clubChongqing Dangdai Lifan FC SWM Team .
clubClermont Foot 63 .
clubClub AmÃ©rica .
clubClub Athletico Paranaense .
clubClub AtlÃ©tico Aldosivi .
clubClub AtlÃ©tico Banfield .
clubClub AtlÃ©tico ColÃ³n .
clubClub AtlÃ©tico Grau .
clubClub AtlÃ©tico HuracÃ¡n .
clubClub AtlÃ©tico LanÃºs .
clubClub AtlÃ©tico Talleres .
clubClub AtlÃ©tico Tigre .
clubClub Atlas .
clubClub Blooming .
clubClub BolÃvar .
clubClub Brugge KV .
clubClub GuaranÃ .
clubClub LeÃ³n .
clubClub Libertad .
clubClub Necaxa .
clubClub Plaza Colonia .
clubClub Tijuana .
clubClube Sport MarÃtimo .
clubColchester United .
clubColo-Colo .
clubColorado Rapids .
clubColumbus Crew SC .
clubCoquimbo Unido .
clubCoritiba .
clubCork City .
clubCoventry City .
clubCracovia .
clubCrawley Town .

clubCrewe Alexandra .
clubCrotone .
clubCruz Azul .
clubCrystal Palace .
clubCusco FC .
clubDaegu FC .
clubDalian YiFang FC .
clubDamac FC .
clubDC United .
clubDefensa y Justicia 0.39245484
clubDelfÃn SC .
clubDenizlispor .
clubDeportivo AlavÃ©s .
clubDeportivo Binacional .
clubDeportivo Cali .
clubDeportivo Pasto .
clubDeportivo Toluca .
clubDerby County .
clubDerry City .
clubDijon FCO .
clubDinamo BucureÅŸti .
clubDinamo Zagreb .
clubDjurgÃ¥rdens IF .
clubDoncaster Rovers .
clubDSC Arminia Bielefeld .
clubDundalk .
clubDundee United .
clubDynamo Kyiv .
clubEintracht Braunschweig .
clubEintracht Frankfurt .
clubEl Nacional .
clubElche CF .
clubEmelec .
clubEmpoli .
clubEn Avant de Guingamp .
clubESTAC Troyes .
clubEstudiantes de La Plata .
clubEstudiantes de MÃ©rida .
clubEttifaq FC .
clubEverton .
clubExeter City .
clubFalkenbergs FF .
clubFamalicÃ£o .
clubFarense .
clubFatih KaragÃ¼mrÃ¼k S.K. .
clubFC Admira Wacker MÃ¶dling .
clubFC ArgeÈ™ .
clubFC Augsburg .
clubFC Barcelona .
clubFC Basel 1893 .
clubFC Bayern MÃ¼nchen .
clubFC BotoÅŸani .
clubFC Cartagena .
clubFC Chambly Oise .
clubFC Cincinnati .
clubFC Dallas .
clubFC Emmen .
clubFC Erzgebirge Aue .
clubFC Girondins de Bordeaux .
clubFC Groningen .
clubFC Hansa Rostock .

clubFC Hermannstadt .
clubFC Ingolstadt 04 .
clubFC JuÃ¡rez .
clubFC KÃ¸benhavn .
clubFC Lausanne-Sport .
clubFC Lorient .
clubFC Lugano .
clubFC Luzern .
clubFC Metz .
clubFC Midtjylland .
clubFC Nantes .
clubFC NordsjÃ¦lland .
clubFC PaÃ§os de Ferreira .
clubFC Porto .
clubFC Red Bull Salzburg .
clubFC Schalke 04 .
clubFC Seoul .
clubFC Sion .
clubFC Sochaux-MontbÃ©liard .
clubFC St. Gallen .
clubFC St. Pauli .
clubFC Tokyo .
clubFC Twente .
clubFC Utrecht .
clubFC Vaduz .
clubFC Viitorul .
clubFC Voluntari .
clubFC WÃ¼rzburger Kickers .
clubFC ZÃ¼rich .
clubFCSB (Steaua) .
clubFenerbahÃ§e SK .
clubFeyenoord .
clubFinn Harps .
clubFiorentina .
clubFK Austria Wien .
clubFK BodÃ¸/Glimt .
clubFK Haugesund .
clubFlamengo .
clubFleetwood Town .
clubFluminense .
clubForest Green Rovers .
clubFortaleza .
clubFortuna DÃ¼sseldorf .
clubFortuna Sittard .
clubFSV Zwickau .
clubFulham .
clubGÃ¶ztepe SK .
clubGÃ³rnik Zabrze .
clubGalatasaray SK .
clubGamba Osaka .
clubGangwon FC .
clubGaz Metan MediaÅŸ .
clubGaziÅŸehir Gaziantep F.K. .
clubGenÃ§lerbirliÄŸi SK .
clubGenoa .
clubGetafe CF .
clubGil Vicente FC .
clubGillingham .
clubGimnasia y Esgrima La Plata .
clubGirona FC .
clubGodoy Cruz .

clubGoiÃ¡s .
clubGrÃªmio .
clubGranada CF .
clubGrenoble Foot 38 .
clubGrimsby Town .
clubGuadalajara .
clubGuangzhou Evergrande Taobao FC .
clubGuangzhou R&F FC .
clubGwangJu FC .
clubHallescher FC .
clubHamburger SV .
clubHamilton Academical FC .
clubHammarby IF .
clubHannover 96 .
clubHarrogate Town .
clubHatayspor .
clubHebei China Fortune FC .
clubHellas Verona .
clubHelsingborgs IF .
clubHenan Jianye FC .
clubHeracles Almelo .
clubHertha BSC .
clubHibernian .
clubHJK Helsinki .
clubHokkaido Consadole Sapporo .
clubHolstein Kiel .
clubHouston Dynamo .
clubHuddersfield Town .
clubHull City .
clubIF Elfsborg .
clubIFK GÃ¶teborg .
clubIFK NorrkÃ¶ping .
clubIK Sirius .
clubIK Start .
clubIncheon United FC .
clubIndependiente .
clubIndependiente del Valle .
clubIndependiente MedellÃn .
clubInter .
clubInter Miami .
clubInternacional .
clubIpswich Town .
clubJagiellonia BiaÅ,ystok .
clubJeonbuk Hyundai Motors .
clubJiangsu Suning FC .
clubJorge Wilstermann .
clubJunior FC .
clubJuventus .
clubKAA Gent .
clubKaizer Chiefs .
clubKalmar FF .
clubKarlsruher SC .
clubKAS Eupen .
clubKashima Antlers .
clubKashiwa Reysol .
clubKasimpaÅŸa SK .
clubKawasaki Frontale .
clubKayserispor .
clubKFC Uerdingen 05 .
clubKilmarnock .
clubKRC Genk .

clubKristiansund BK .
clubKSV Cercle Brugge .
clubKV Kortrijk .
clubKV Mechelen .
clubKV Oostende .
clubLa Berrichonne de ChÃ¢teauroux .
clubLA Galaxy .
clubLASK Linz .
clubLazio .
clubLDU Quito .
clubLe Havre AC .
clubLecce .
clubLech PoznaÅ„ .
clubLechia GdaÅ„sk .
clubLeeds United .
clubLegia Warszawa .
clubLeicester City .
clubLevante UD .
clubLeyton Orient .
clubLincoln City .
clubLiverpool .
clubLiverpool FÃºtbol Club .
clubLivingston FC .
clubLlaneros de Guanare .
clubLokomotiv Moscow .
clubLos Angeles FC .
clubLOSC Lille .
clubLuton Town .
clubLyngby BK .
clubMÃ¡laga CF .
clubMacarthur FC .
clubMalmÃ¶ FF .
clubManchester City .
clubManchester United .
clubMansfield Town .
clubMazatlÃ¡n FC .
clubMedipol BaÅŸakÅŸehir FK .
clubMelbourne City FC .
clubMelbourne Victory .
clubMelgar FBC .
clubMiddlesbrough .
clubMilan .
clubMillonarios FC .
clubMillwall .
clubMilton Keynes Dons .
clubMinnesota United FC .
clubMjÃ¸ndalen IF .
clubMjÃ¤llby AIF .
clubMKE AnkaragÃ¼cÃ¼ .
clubMolde FK .
clubMonterrey .
clubMontpellier HSC .
clubMontreal Impact .
clubMorecambe .
clubMoreirense FC .
clubMotherwell .
clubMSV Duisburg .
clubNÃ®mes Olympique .
clubNacional AsunciÃ³n .
clubNacional de Montevideo .
clubNacional PotosÃ .

clubNagoya Grampus .
clubNapoli .
clubNashville SC .
clubNew England Revolution .
clubNew York City FC .
clubNew York Red Bulls .
clubNewcastle Jets .
clubNewcastle United .
clubNewell's Old Boys .
clubNewport County .
clubNorthampton Town .
clubNorwich City .
clubNottingham Forest .
clubOceÃ¢nico FC .
clubOdds BK .
clubOdense Boldklub .
clubOGC Nice .
clubOita Trinita .
clubOldham Athletic .
clubOlimpia AsunciÃ³n .
clubOlympiacos CFP .
clubOlympique de Marseille .
clubOlympique Lyonnais .
clubOriente Petrolero .
clubOrlando City SC .
clubOrlando Pirates .
clubOs Belenenses .
clubOud-Heverlee Leuven .
clubOxford United .
clubPachuca .
clubPalmeiras .
clubPanathinaikos FC .
clubPAOK .
clubParis FC .
clubParis Saint-Germain .
clubParma .
clubPatronato .
clubPau FC .
clubPeÃ±arol .
clubPEC Zwolle .
clubPerth Glory .
clubPeterborough United .
clubPFC CSKA Moscow .
clubPhiladelphia Union .
clubPiast Gliwice .
clubPlymouth Argyle .
clubPodbeskidzie Bielsko-BiaÅ‚a .
clubPogoÅ„ Szczecin .
clubPohang Steelers .
clubPolitehnica IaÅŸi .
clubPort Vale .
clubPortimonense SC .
clubPortland Timbers .
clubPortsmouth .
clubPreston North End .
clubPSV .
clubPuebla FC .
clubQingdao Huanghai F.C. .
clubQueens Park Rangers .
clubQuerÃ©taro .
clubRacing Club .

clubRacing Club de Lens	.
clubRakÃ³w CzÄ™stochowa	.
clubRanders FC	.
clubRangers FC	.
clubRayo Vallecano	.
clubRB Leipzig	.
clubRC Celta	.
clubRC Strasbourg Alsace	.
clubRCD Espanyol	.
clubRCD Mallorca	.
clubReading	.
clubReal Betis	.
clubReal Madrid	.
clubReal Oviedo	.
clubReal Salt Lake	.
clubReal Sociedad	.
clubReal Sporting de GijÃ³n	.
clubReal Valladolid CF	.
clubReal Zaragoza	.
clubRio Ave FC	.
clubRiver Plate	.
clubRiver Plate AsunciÃ³n	.
clubRiver Plate Montevideo	.
clubRKC Waalwijk	.
clubRochdale	.
clubRodez Aveyron Football	.
clubRoma	.
clubRosario Central	.
clubRosenborg BK	.
clubRoss County FC	.
clubRotherham United	.
clubRoyal Antwerp FC	.
clubRoyal Excel Mouscron	.
clubRSC Anderlecht	.
clubSÃ¸nderjyskE	.
clubSÃ£o Paulo	.
clubSagan Tosu	.
clubSalford City	.
clubSampdoria	.
clubSan Jose Earthquakes	.
clubSan Lorenzo de Almagro	.
clubSandefjord Fotball	.
clubSanfrecce Hiroshima	.
clubSangju Sangmu FC	3.31612397
clubSanta Clara	.
clubSantos	.
clubSantos Laguna	.
clubSarpsborg 08 FF	.
clubSassuolo	.
clubSC Braga	.
clubSC Freiburg	.
clubSC Heerenveen	.
clubSC Paderborn 07	.
clubSC Verl	.
clubSCR Altach	.
clubScunthorpe United	.
clubSD Aucas	.
clubSD Eibar	.
clubSD Huesca	.
clubSD Ponferradina	.
clubSeattle Sounders FC	.

clubSeongnam FC .
clubSepsi OSK .
clubServette FC .
clubSevilla FC .
clubSG Dynamo Dresden .
clubShakhtar Donetsk .
clubShamrock Rovers .
clubShandong Luneng TaiShan FC .
clubShanghai Greenland Shenhua FC .
clubShanghai SIPG FC .
clubSheffield United .
clubSheffield Wednesday .
clubShelbourne FC .
clubShenzhen FC .
clubShijiazhuang Ever Bright F.C. .
clubShimizu S-Pulse .
clubShonan Bellmare .
clubShrewsbury .
clubSint-Truidense VV .
clubSivasspor .
clubSK Brann .
clubSK Rapid Wien .
clubSK Slavia Praha .
clubSK Sturm Graz .
clubSKN St. PÃ¶lten .
clubSL Benfica .
clubSligo Rovers .
clubSol de AmÃ©rica .
clubSouthampton .
clubSouthend United .
clubSPAL .
clubSparta Praha .
clubSparta Rotterdam .
clubSpartak Moscow .
clubSpezia .
clubSport Huancayo .
clubSporting CP .
clubSporting de Charleroi .
clubSporting Kansas City .
clubSportivo LuqueÃ±o .
clubSpVgg Greuther FÃ¼rth .
clubSpVgg Unterhaching .
clubSSV Jahn Regensburg .
clubSt. Johnstone FC .
clubSt. Mirren .
clubSt. Patrick's Athletic .
clubStabÃ¦k Fotball .
clubStade Brestois 29 .
clubStade de Reims .
clubStade Malherbe Caen .
clubStade Rennais FC .
clubStal Mielec .
clubStandard de LiÃ¨ge .
clubStevenage .
clubStoke City .
clubStrÃ¸msgodset IF .
clubSunderland .
clubSuwon Samsung Bluewings .
clubSV Darmstadt 98 .
clubSV Meppen .
clubSV Ried .

clubSV Sandhausen .
clubSV Waldhof Mannheim .
clubSV Wehen Wiesbaden .
clubSV Werder Bremen .
clubSV Zulte-Waregem .
clubSwansea City .
clubSwindon Town .
clubSydney FC .
clubTÃ¼rkgÃ¼cÃ¼ MÃ¼nchen .
clubTianjin TEDA FC .
clubTigres U.A.N.L. .
clubTorino .
clubToronto FC .
clubTottenham Hotspur .
clubToulouse Football Club .
clubTrabzonspor .
clubTranmere Rovers .
clubTSG 1899 Hoffenheim .
clubTSV 1860 MÃ¼nchen .
clubTSV Hartberg .
clubU.N.A.M. .
clubUD AlmerÃa .
clubUD Las Palmas .
clubUD LogroÃ±Ã©s .
clubUdinese .
clubUlsan Hyundai FC .
clubUniÃ³n de Santa Fe .
clubUniÃ³n La Calera .
clubUniversidad CatÃ³lica .
clubUniversidad CatÃ³lica del Ecuador .
clubUniversitatea Craiova .
clubUrawa Red Diamonds .
clubUSL Dunkerque .
clubUTA Arad .
clubVÃ¥lerenga Fotball .
clubVÃ©lez Sarsfield .
clubValencia CF .
clubValenciennes FC .
clubVancouver Whitecaps FC .
clubVarbergs BoIS .
clubVasco da Gama .
clubVegalta Sendai .
clubVejle Boldklub .
clubVfB LÃ¼beck .
clubVfB Stuttgart .
clubVfL Bochum 1848 .
clubVfL OsnabrÃ¼ck .
clubVfL Wolfsburg .
clubViking FK .
clubViktoria KÃ¶ln .
clubViktoria PlzeÅˆ .
clubVillarreal CF .
clubVissel Kobe .
clubVitÃ³ria GuimarÃ£es .
clubVitesse .
clubVVV-Venlo .
clubWaasland-Beveren .
clubWalsall .
clubWarta PoznaÅ„ .
clubWaterford FC .
clubWatford .

clubWellington Phoenix .
clubWest Bromwich Albion .
clubWest Ham United .
clubWestern Sydney Wanderers .
clubWestern United FC .
clubWigan Athletic .
clubWillem II .
clubWisÅ‚a KrakÃ³w .
clubWisÅ‚a PÅ‚ock .
clubWolfsberger AC .
clubWolverhampton Wanderers .
clubWSG Tirol .
clubWuhan Zall .
clubWycombe Wanderers .
clubYeni Malatyaspor .
clubYokohama F. Marinos .
clubYokohama FC .
clubZagÅ‚Ä™bie Lubin .
clubZamora FC .
leagueArgentina Primera DivisiÃ³n .
leagueArgentinian Primera B Nacional .
leagueAustralian Hyundai A-League .
leagueAustrian Football Bundesliga .
leagueBelgian Jupiler Pro League .
leagueCampeonato Brasileiro SÃ©rie A .
leagueChilian Campeonato Nacional .
leagueChinese Super League .
leagueColombian Liga PostobÃ³n .
leagueCroatian Prva HNL .
leagueCzech Republic Gambrinus Liga .
leagueDanish Superliga .
leagueEcuadorian Serie A .
leagueEnglish League Championship .
leagueEnglish League One .
leagueEnglish League Two .
leagueEnglish Premier League .
leagueFinnish Veikkausliiga .
leagueFrench Ligue 1 .
leagueFrench Ligue 2 .
leagueGerman 1. Bundesliga .
leagueGerman 2. Bundesliga .
leagueGerman 3. Bundesliga .
leagueGreek Super League .
leagueHolland Eredivisie .
leagueItalian Serie A .
leagueItalian Serie B .
leagueJapanese J. League Division 1 .
leagueKorean K League Classic .
leagueLiga de FÃºtbol Profesional Boliviano .
leagueMexican Liga MX .
leagueNorwegian Eliteserien .
leagueParaguayan Primera DivisiÃ³n .
leaguePeruvian Primera DivisiÃ³n .
leaguePolish T-Mobile Ekstraklasa .
leaguePortuguese Liga ZON SAGRES .
leagueRep. Ireland Airtricity League .
leagueRomanian Liga I .
leagueRussian Premier League .
leagueSaudi Abdul L. Jameel League .
leagueScottish Premiership .
leagueSouth African Premier Division .

leagueSpain Primera Division	.
leagueSpanish Segunda DivisiÃ³n	0.30836288
leagueSwedish Allsvenskan	.
leagueSwiss Super League	.
leagueTurkish SÃ¼per Lig	.
leagueUAE Arabian Gulf League	.
leagueUkrainian Premier League	.
leagueUruguayan Primera DivisiÃ³n	.
leagueUSA Major League Soccer	.
leagueVenezuelan Primera DivisiÃ³n	.
overall	.
potential	0.02171299
value	.
wage	.
positionCB	.
positionCDM	.
positionCF	.
positionCM	.
positionGK	.
positionLB	.
positionLM	.
positionLW	.
positionLWB	.
positionRB	.
positionRM	.
positionRW	.
positionRWB	.
positionST	.
footRight	.
reputation2	.
reputation3	.
reputation4	.
reputation5	.
weakfoot2	.
weakfoot3	.
weakfoot4	.
weakfoot5	.
attack_wrLow	.
attack_wrMedium	.
defend_wrLow	.
defend_wrMedium	.
pace	.
shooting	.
passing	.
dribbling	.
defending	.
physical	.
attacking	.
skill	.
movement	.
power	.
mentality	.
tackling	.