FIFA 22 PLAYERS ANALYSIS

FINAL PROJECT DATA 101

Anish Bhurtyal, Avanish Chaulagai, Nikita Gerzhgorin, Joseph Donohoe

INTRODUCTION

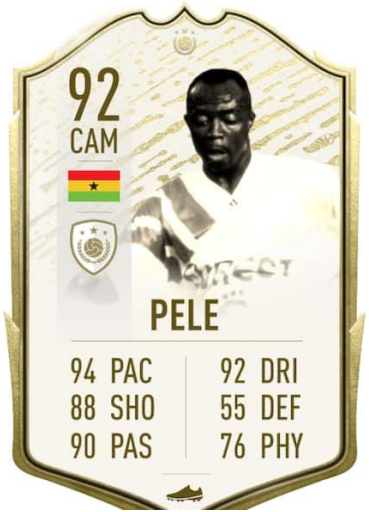
- With the season of FIFA upon us, it was natural that we as a group would pick this topic as our data set.
- It was founded in 1904 to promote unity among national soccer associations, and it now has 209 members.
- FIFA was created by seven national associations: Belgium, Denmark, France, the Netherlands, Spain, Sweden, and Switzerland to promote the organization of football matches at all levels.

INTRO CONT...

- In this presentation, we examine the relationship between a variety of player characteristics, including as age, nationality, value, positions, endurance, agility, response speed, clubs, and skills, and the player's overall field performance in FIFA matches.

WHY PLAYERS ANALYSIS?

- The player ratings are one of the components of FIFA.
- This helps team **coach/managers** pick who to select on any particular game
- During game whom to replace players



Player Stat of Abedi Pele

In this project, we take the analysis of the player and model each attribute and investigate how they affect overall performances.

HYPOTHESIS

- We have two hypothesis and we created two models to test our hypothesis:=
  1. Hypothesis 1: To test whether in-field attributes like shooting, passing, pace, dribbling determine a player's salary
  2. Hypothesis 2: To test whether Fifa was biased in assigning overall ratings based on the player's attributes
  3. Hypothesis 3: To test whether cluster analysis gives a representative playing team positions

DATASET OVERVIEW

- Our dataset consists of information about players.
- Among other things like, height, weight, overall performance.
- We got this dataset from kaggle.com.
  - Link: FIFA 22 complete player dataset | Kaggle
- Some columns consist of null values, which are handled later with either overall mean/median.

```r
players <- read.csv("players_22.csv")
dim(players)
```

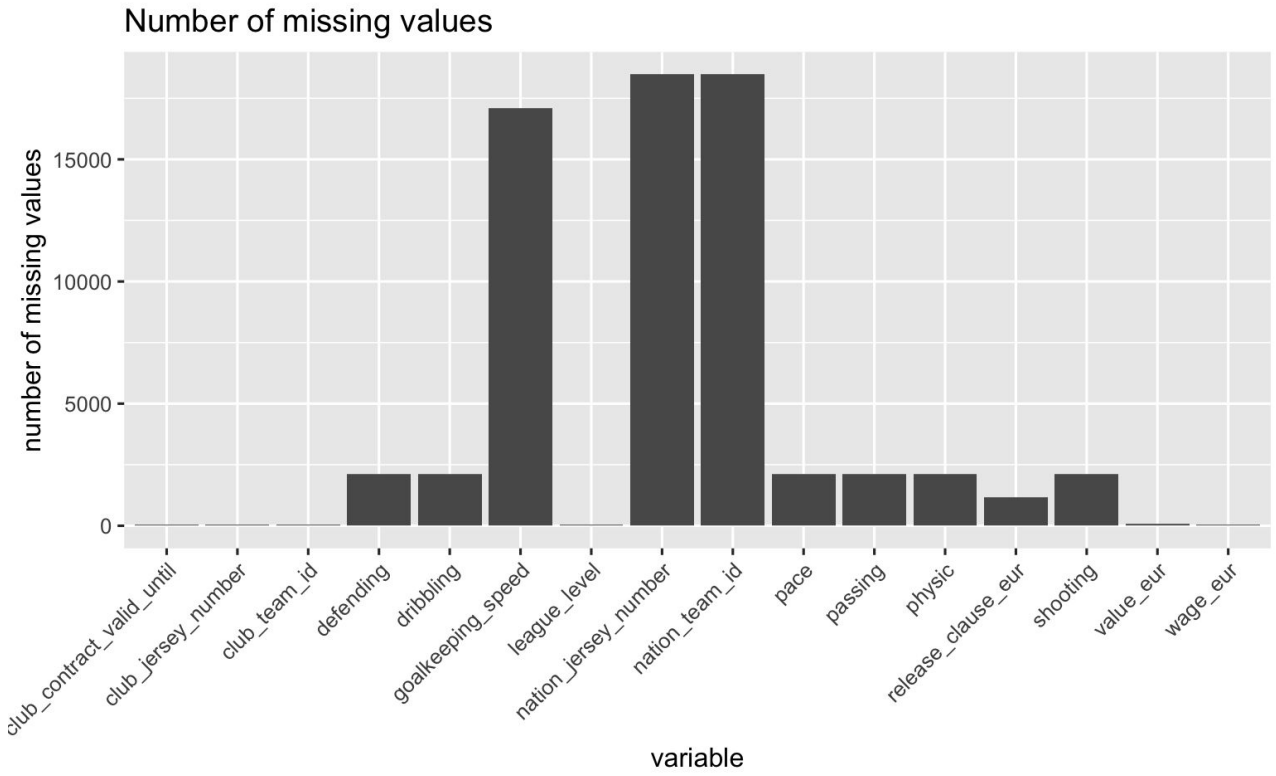
```
[1] 19239   110
```

COLUMNS/ATTRIBUTES

- Each column is made of information about each individual.
- These **110** attributes are what define each player.
- This provides a rich description when visualized properly on a graph.

```
$ pace                       : int  85 78 87 91 76 NA 97 NA NA 70 ...
$ shooting                   : int  92 92 94 83 86 NA 88 NA NA 91 ...
$ passing                    : int  91 79 80 86 93 NA 80 NA NA 83 ...
$ dribbling                  : int  95 86 88 94 88 NA 92 NA NA 83 ...
$ defending                  : int  34 44 34 37 64 NA 36 NA NA 47 ...
$ physic                     : int  65 82 75 63 78 NA 77 NA NA 83 ...
$ attacking_crossing         : int  85 71 87 85 94 13 78 15 18 80 ...
$ attacking_finishing        : int  95 95 95 83 82 11 93 13 14 94 ...
$ attacking_heading_accuracy : int  70 90 90 63 55 15 72 25 11 86 ...
$ attacking_short_passing    : int  91 85 80 86 94 43 85 60 61 85 ...
$ attacking_volleys          : int  88 89 86 86 82 13 83 11 14 88 ...
$ skill_dribbling            : int  96 85 88 95 88 12 93 30 21 83 ...
$ skill_curve                : int  93 79 81 88 85 13 80 14 18 83 ...
$ skill_fk_accuracy          : int  94 85 84 87 83 14 69 11 12 65 ...
$ skill_long_passing         : int  91 70 77 81 93 40 71 68 63 86 ...
$ skill_ball_control         : int  96 88 88 95 91 30 91 46 30 85 ...
```

ANYONE MISSING ?



HANDLING MISSING VALUES

```
fifa <- fifa %>%
  mutate(passing = ifelse(is.na(passing),
                          median(passing, na.rm = T),
                          passing))%>%
  mutate(dribbling = ifelse(is.na(dribbling),
                          median(dribbling, na.rm = T),
                          dribbling))%>%
  mutate(pace = ifelse(is.na(pace),
                          median(pace, na.rm = T),
                          pace)) %>%
  mutate(defending = ifelse(is.na(defending),
                          median(defending, na.rm = T),
                          defending))
```

HANDLING DUPLICATES

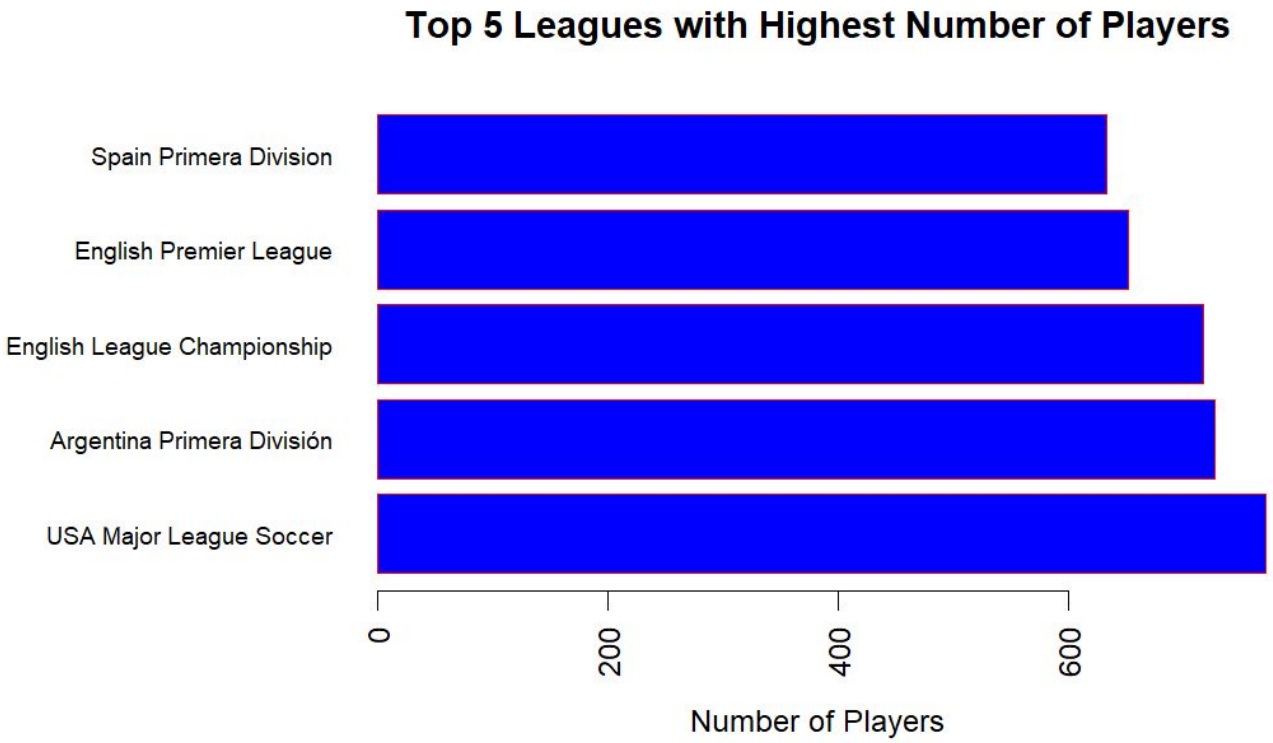
```
sprintf ("Number of duplicates columns:")
sum(duplicated(fifa$long_name) == TRUE)
sprintf ("Dimension  before")
dim(fifa)
```

```
"Number of duplicates columns:"
20
"Dimension  before"
19239   110
```

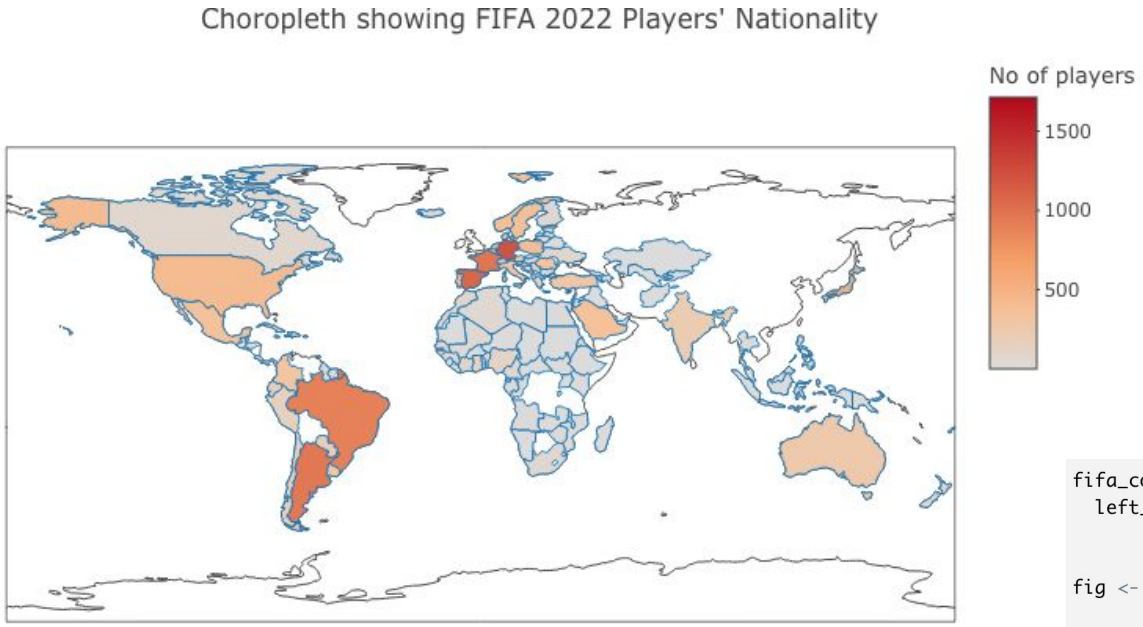
```
fifa <- fifa%>%distinct(long_name,
                        .keep_all = TRUE)
sprintf ("Dimension  after")
dim(fifa)
```

```
"Dimension  after"
19219   110
```

VISUALIZING AMOUNTS



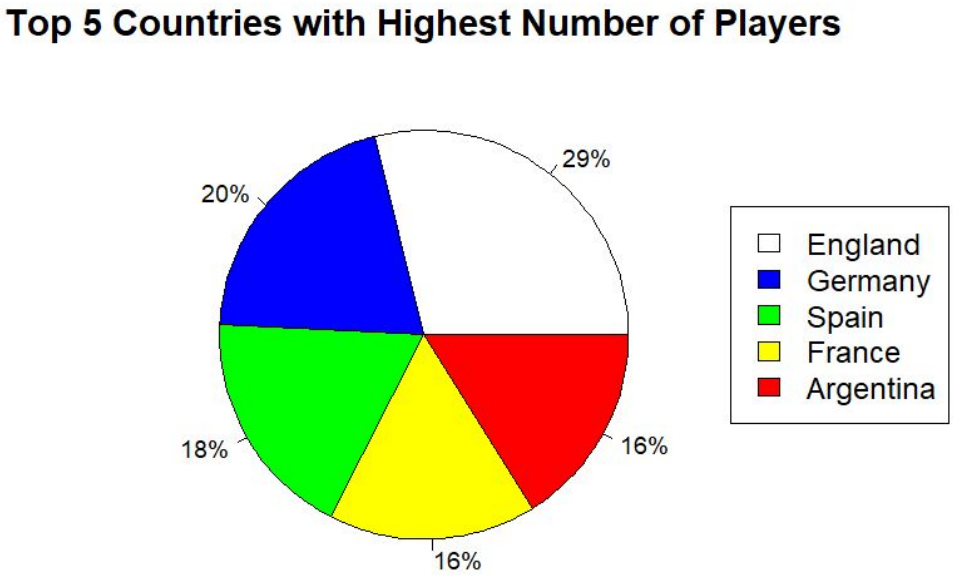
Where are the players from?



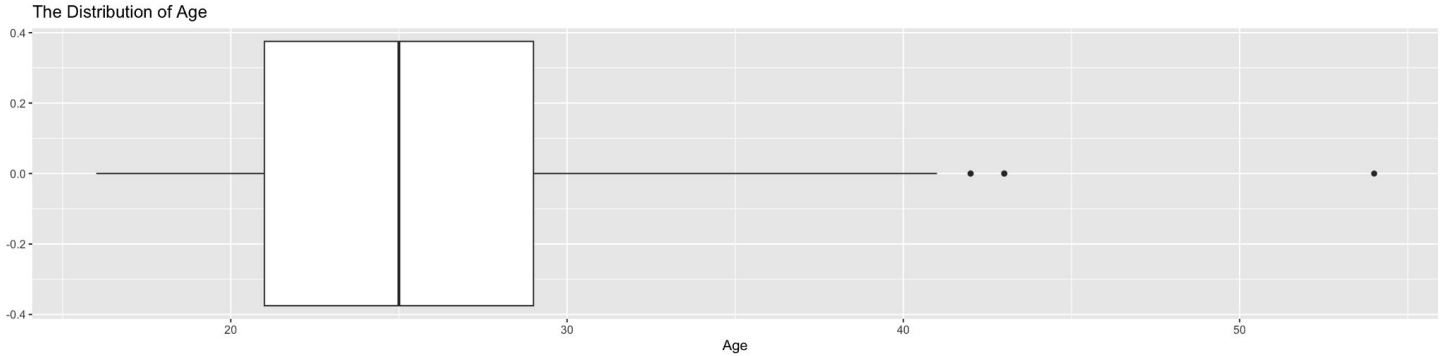
Country_with_code dataset: Link

```
fifa_country_count_with_code <- fifa_country_count %>%
  left_join( country_with_code,
           by=c('nationality_name' = 'name'))

fig <- plot_ly(fifa_country_count_with_code, type='choropleth',
               locations=fifa_country_count_with_code$alpha.3,
               z=fifa_country_count_with_code$Freq,
               text=fifa_country_count_with_code$nationality_name,
               colorscale="ice")
```

VISUALIZING PROPORTIONS



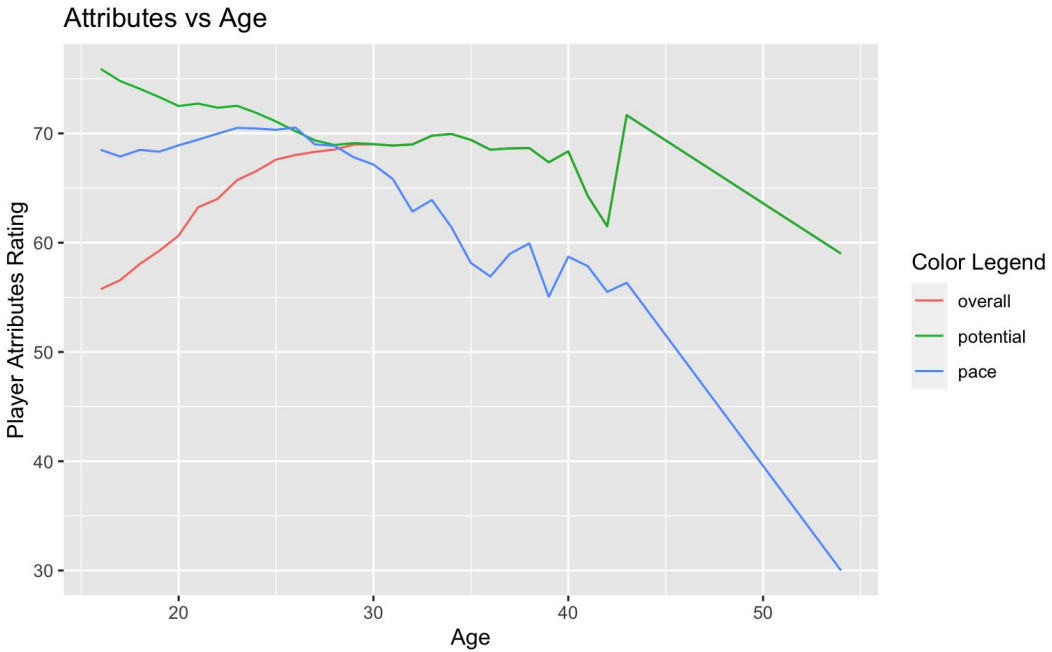
VISUALIZING AGE DISTRIBUTIONS



	short_name <chr>	nationality_name <chr>	age <int>
1	K. Miura	Japan	54
2	G. Buffon	Italy	43
3	C. Lucchetti	Argentina	43
4	S. Nakamura	Japan	43
5	D. Vaca	Bolivia	42

	short_name <chr>	nationality_name <chr>	age <int>
1	Gavi	Spain	16
2	V. Barco	Argentina	16
3	A. Kalogeropoulos	Greece	16
4	Yayo	Spain	16
5	R. van den Berg	Netherlands	16

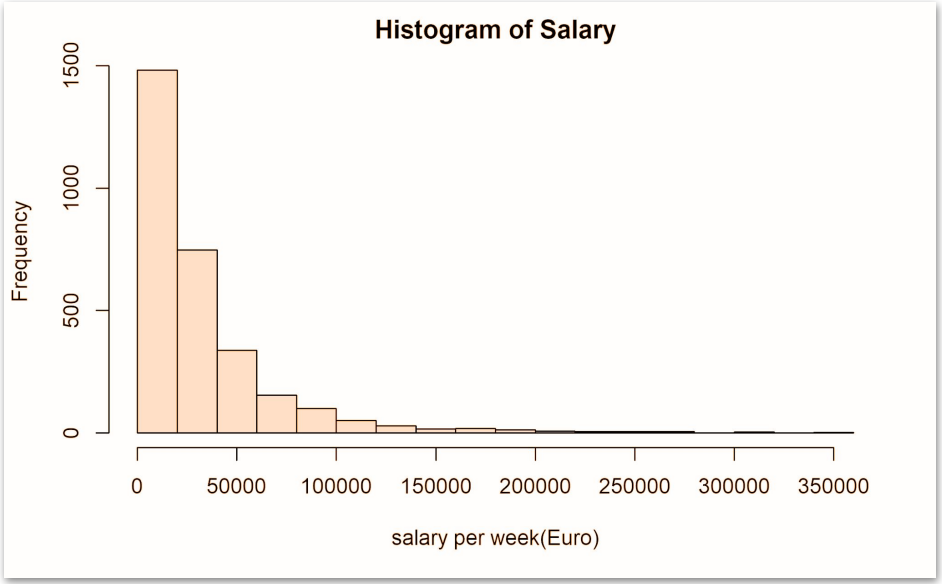
ATTRIBUTES OVER AGE

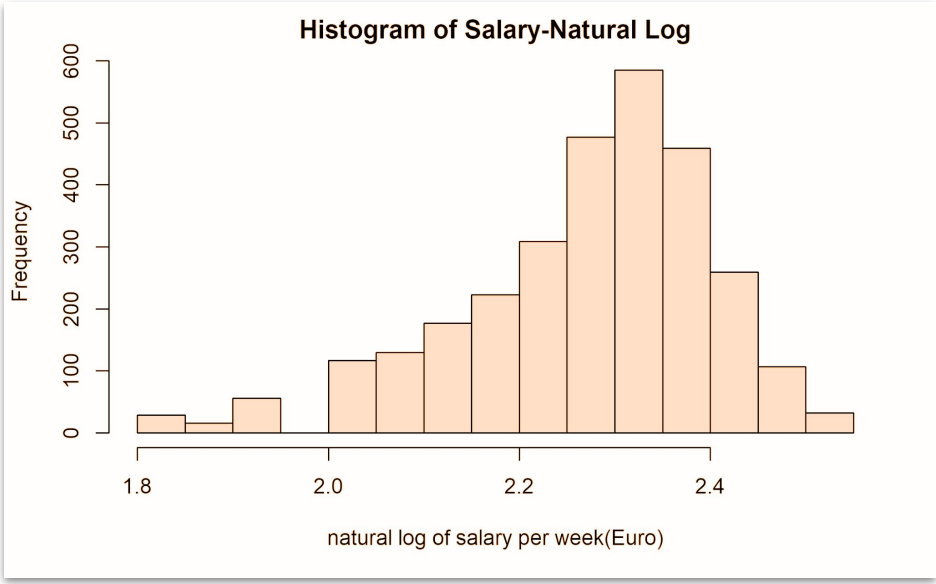



```r
filtered_attributes <- fifa%>%
  group_by(age) %>%
  summarise_at(vars(overall,potential,pace),
               list( mean))
filtered_attributes

library("reshape2")
data_long <- melt(filtered_attributes, id="age")

ggplot(data_long,
               aes(x = age,
                   y = value,
                   color = variable)) +
        geom_line()+
        labs(
      y = "Player Atrributes Rating",
      x = "Age",
      color = "Color Legend",
      title = "Attributes vs Age",
    )
```

ATTRIBUTES OVER AGE

DATA MODELLING - SALARY DISTRIBUTION





DATA MODELLING - CODE

```{r}

hist(data$overall, xlab="Overall Rating", main="Histogram of Overall Rating")
data$overall = log10(data$overall)

#data$overall = log(data$overall)
hist(data$overall, xlab="Log10 of Overall Rating", main="Histogram of Overall Rating- Log10")


wage_individualskills <- lm(overall ~ pace + shooting + passing + dribbling + defending +
physic + age + preferred_foot + attacking_crossing + attacking_finishing +
attacking_heading_accuracy + attacking_short_passing + attacking_volleys + skill_dribbling
+skill_curve +skill_fk_accuracy +skill_long_passing +skill_ball_control
+movement_acceleration +movement_sprint_speed +movement_agility +movement_reactions
+movement_balance +power_shot_power +power_jumping+ power_stamina +power_strength
+power_long_shots +mentality_aggression +mentality_interceptions +mentality_positioning
+mentality_vision +mentality_penalties +mentality_composure +defending_marking_awareness
+defending_standing_tackle +defending_sliding_tackle, data = data)

summary(wage_individualskills)

```

DATA MODELLING - SIGNIFICANCE SALARY

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -80503.139   1690.446 -47.622  < 2e-16 ***
pace                          121.772     45.289   2.689 0.007177 **
shooting                      -57.075    146.937  -0.388 0.697699
passing                      -210.410     86.741  -2.426 0.015287 *
dribbling                     309.577    122.807   2.521 0.011716 *
defending                     558.270    121.362   4.600 4.25e-06 ***
physic                        -40.448     60.816  -0.665 0.506004
age                          -527.895     31.351 -16.838  < 2e-16 ***
preferred_footRight            -6.481    278.387  -0.023 0.981426
attacking_crossing            105.856     23.586   4.488 7.23e-06 ***
attacking_finishing            62.332     68.116   0.915 0.360157
attacking_heading_accuracy    108.774     19.494   5.580 2.44e-08 ***
attacking_short_passing       162.186     37.422   4.334 1.47e-05 ***
attacking_volleys              96.738     19.222   5.033 4.88e-07 ***
skill_dribbling              -119.262     59.957  -1.989 0.046702 *
skill_curve                    50.176     17.839   2.813 0.004918 **
skill_fk_accuracy               9.490     15.844   0.599 0.549203
skill_long_passing             24.411     24.997   0.977 0.328797
skill_ball_control             97.273     43.946   2.213 0.026875 *
movement_acceleration         -22.313     28.961  -0.770 0.441058
movement_sprint_speed          56.451     30.944   1.824 0.068122 .
movement_agility             -111.605     20.461  -5.455 4.97e-08 ***
movement_reactions            760.495     23.455  32.424  < 2e-16 ***
movement_balance               20.432     16.387   1.247 0.212473
power_shot_power               45.528     34.831   1.307 0.191184
power_jumping                  20.420     12.730   1.604 0.108722
power_stamina                 -83.142     20.756  -4.006 6.21e-05 ***
power_strength                 24.093     29.140   0.827 0.408360
power_long_shots             -128.797     34.935  -3.687 0.000228 ***
mentality_aggression            8.391     18.179   0.462 0.644382
mentality_interceptions      -143.811     32.196  -4.467 7.99e-06 ***
mentality_positioning         -44.475     21.203  -2.098 0.035952 *
mentality_vision               81.719     21.731   3.760 0.000170 ***
mentality_penalties            25.361     18.081   1.403 0.160741
mentality_composure           197.401     19.705  10.018  < 2e-16 ***
defending_marking_awareness  -100.887     38.983  -2.588 0.009661 **
defending_standing_tackle     -74.295     46.549  -1.596 0.110494
defending_sliding_tackle     -100.377     31.034  -3.234 0.001221 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

DATA MODELLING - SALARY PARTITION

```r
dt = sort(sample(nrow(data),nrow(data)*.7))
train <- data[dt,]
test <- data[-dt,]


train<- train %>% filter (!is.na(wage_eur))
test <- test %>% filter (!is.na(wage_eur))
data2 <- data.frame(actual= test$wage_eur, predicted = predict(wage_individualskills, test))
dim(test)
dim(train)
data2
colSums(is.na(data2))
```

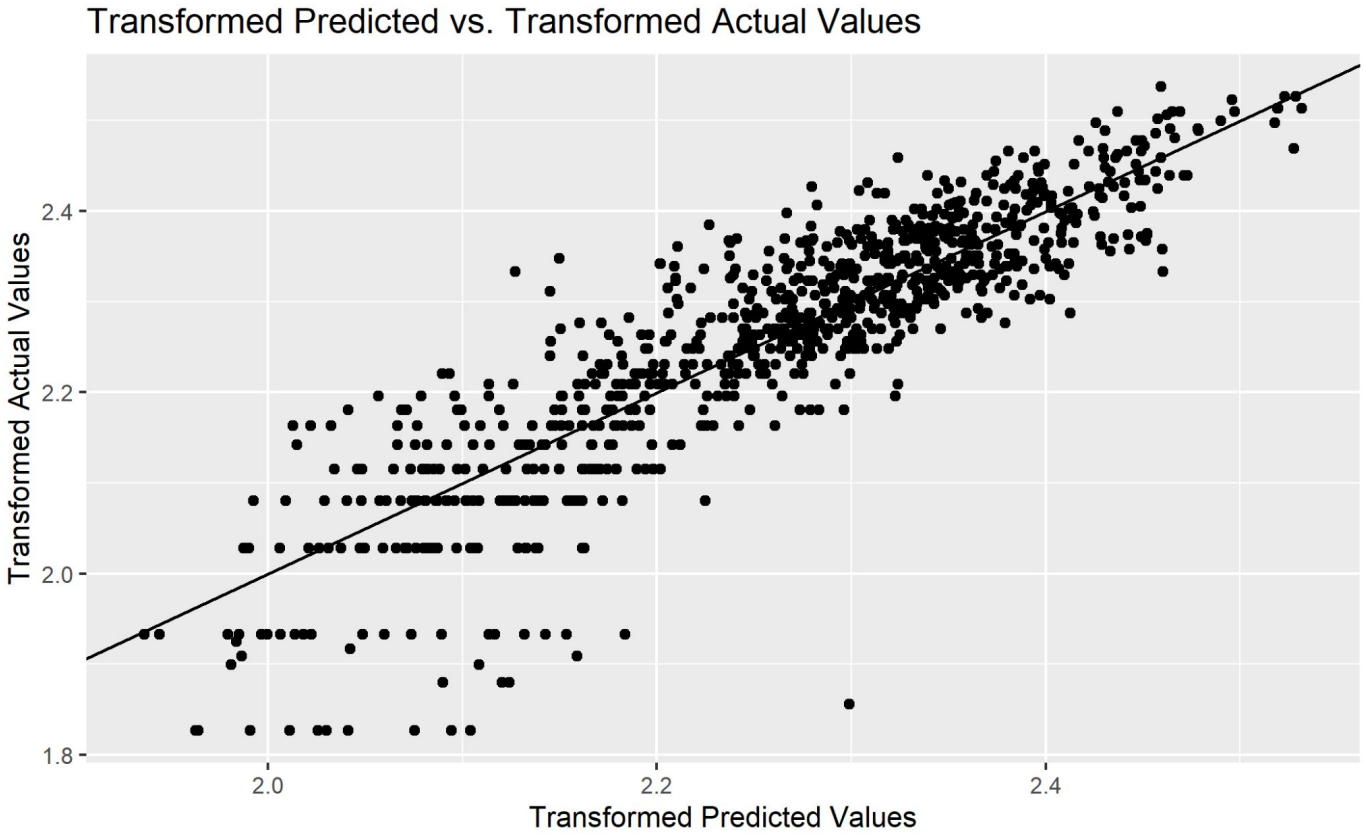
MODELLING ACCURACY- MEAN SQUARE ERROR

```
Residual standard error: 0.06401 on 2938 degrees of freedom
Multiple R-squared:  0.7662,    Adjusted R-squared:  0.7633
F-statistic: 260.3 on 37 and 2938 DF,  p-value: < 2.2e-16
```

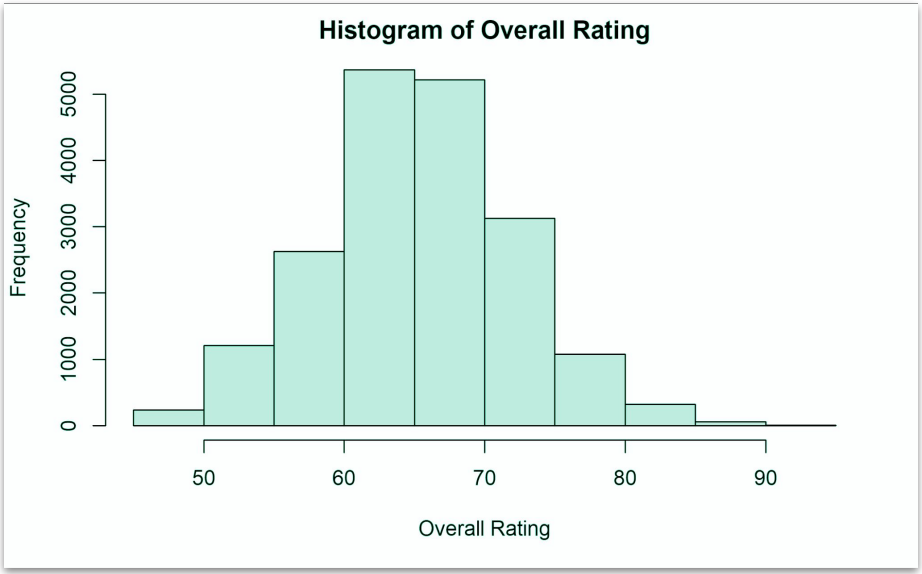
```
mean(( data2$actual- data2$predicted)^2)
```

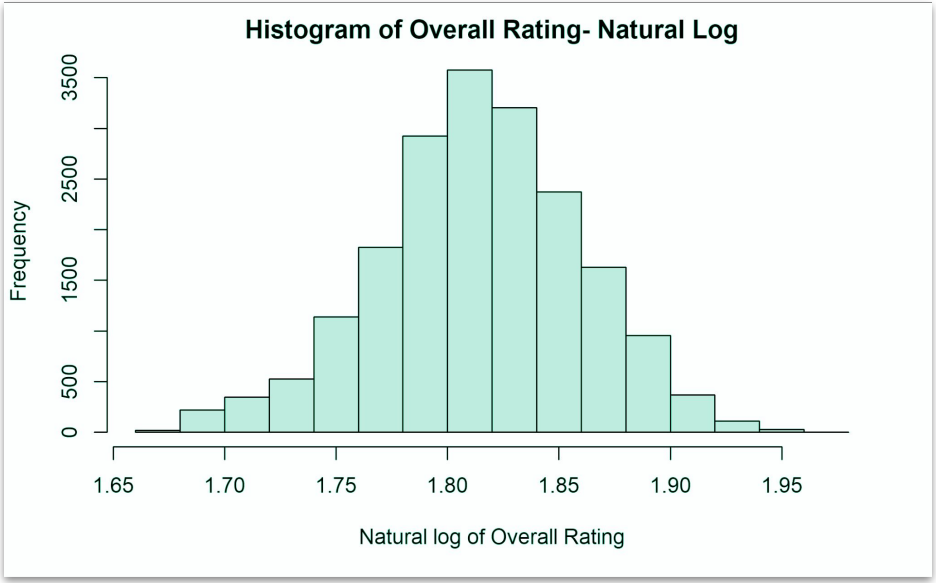
```
[1] 0.003964325
```

MODELLING ACCURACY- SALARY GRAPH



DATA MODELLING- OVERALL DISTRIBUTION





DATA MODELLING- OVERALL SIGNIFICANCE

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -0.0538191  0.0495974  -1.085 0.277888
pace                       -0.0030619  0.0013367  -2.291 0.022004 *
shooting                   -0.0169516  0.0043905  -3.861 0.000113 ***
passing                    -0.0018266  0.0025804  -0.708 0.479037
dribbling                   0.0187117  0.0036707   5.098 3.49e-07 ***
defending                   0.0263420  0.0036312   7.254 4.26e-13 ***
physic                     -0.0022711  0.0018051  -1.258 0.208359
age                        -0.0085944  0.0009188  -9.354  < 2e-16 ***
preferred_footRight        -0.0166559  0.0082164  -2.027 0.042666 *
attacking_crossing          0.0027059  0.0006967   3.884 0.000103 ***
attacking_finishing         0.0067074  0.0020333   3.299 0.000973 ***
attacking_heading_accuracy  0.0037963  0.0005764   6.587 4.67e-11 ***
attacking_short_passing     0.0023179  0.0011186   2.072 0.038262 *
attacking_volleys           0.0029147  0.0005687   5.125 3.02e-07 ***
skill_dribbling            -0.0029281  0.0017855  -1.640 0.101053
skill_curve                 0.0021980  0.0005223   4.208 2.59e-05 ***
skill_fk_accuracy          -0.0020740  0.0004672  -4.439 9.11e-06 ***
skill_long_passing         -0.0002938  0.0007373  -0.399 0.690251
skill_ball_control          0.0009281  0.0013133   0.707 0.479756
movement_acceleration       0.0026368  0.0008514   3.097 0.001958 **
movement_sprint_speed       0.0042153  0.0009140   4.612 4.03e-06 ***
movement_agility           -0.0014098  0.0006083  -2.318 0.020487 *
movement_reactions          0.0229011  0.0006888  33.248  < 2e-16 ***
movement_balance           -0.0021278  0.0004832  -4.404 1.07e-05 ***
power_shot_power            0.0093756  0.0010373   9.038  < 2e-16 ***
power_jumping               0.0012343  0.0003744   3.297 0.000980 ***
power_stamina              -0.0009368  0.0006137  -1.526 0.126922
power_strength              0.0021199  0.0008649   2.451 0.014261 *
power_long_shots           -0.0005828  0.0010418  -0.559 0.575896
mentality_aggression        0.0012883  0.0005400   2.386 0.017050 *
mentality_interceptions    -0.0070151  0.0009574  -7.327 2.49e-13 ***
mentality_positioning      -0.0025240  0.0006256  -4.035 5.50e-05 ***
mentality_vision            0.0005344  0.0006494   0.823 0.410583
mentality_penalties         0.0016514  0.0005381   3.069 0.002151 **
mentality_composure         0.0081468  0.0005769  14.123  < 2e-16 ***
defending_marking_awareness -0.0052978  0.0011593  -4.570 4.92e-06 ***
defending_standing_tackle   -0.0073121  0.0013835  -5.285 1.28e-07 ***
defending_sliding_tackle    -0.0019457  0.0009183  -2.119 0.034125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

DATA MODELLING- OVERALL PARTITION

```r
dt = sort(sample(nrow(data),nrow(data)*.7))
train <- data[dt,]
test <- data[-dt,]


train<- train %>% filter (!is.na(overall))
test <- test %>% filter (!is.na(overall))
data2 <- data.frame(actual= test$overall, predicted = predict(OverallRatingPrediction, test))
dim(test)
dim(train)
data2
colSums(is.na(data2))
```

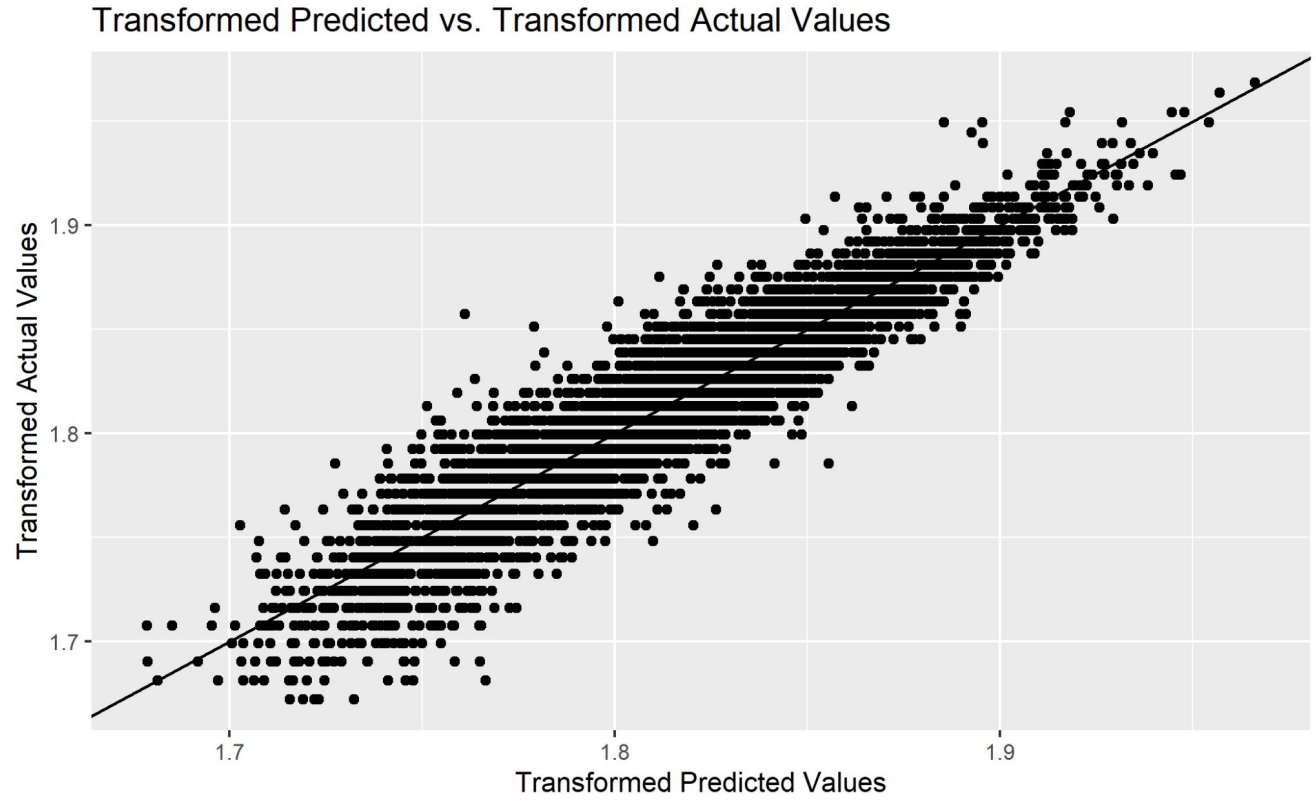
MODELLING ACCURACY- MEAN SQUARE ERROR

```
Residual standard error: 0.01661 on 19201 degrees of freedom
Multiple R-squared:  0.8692,    Adjusted R-squared:  0.869
F-statistic:  3449 on 37 and 19201 DF,  p-value: < 2.2e-16
```

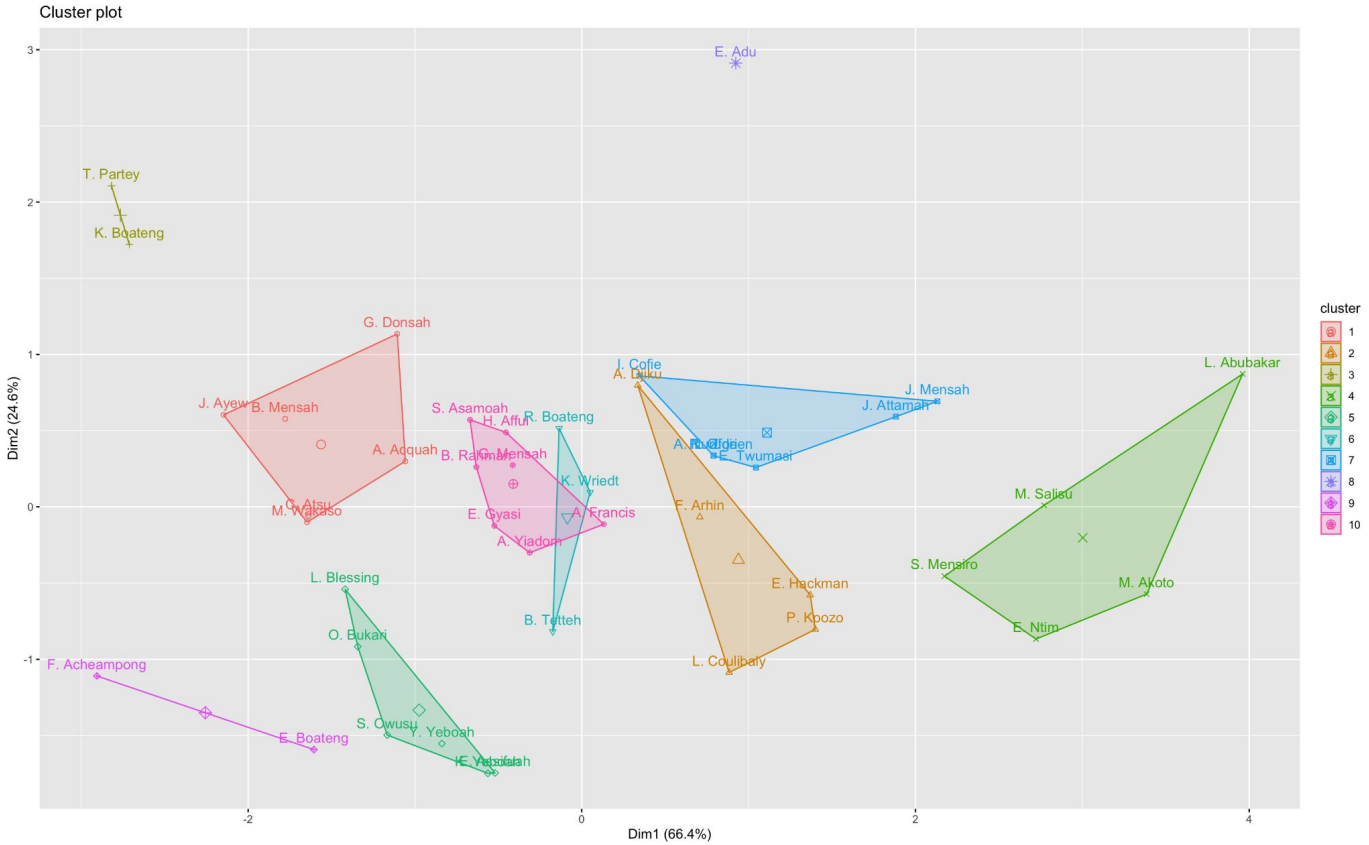
```
mean((data2$actual- data2$predicted)^2)
```

```
[1] 0.0002739077
```

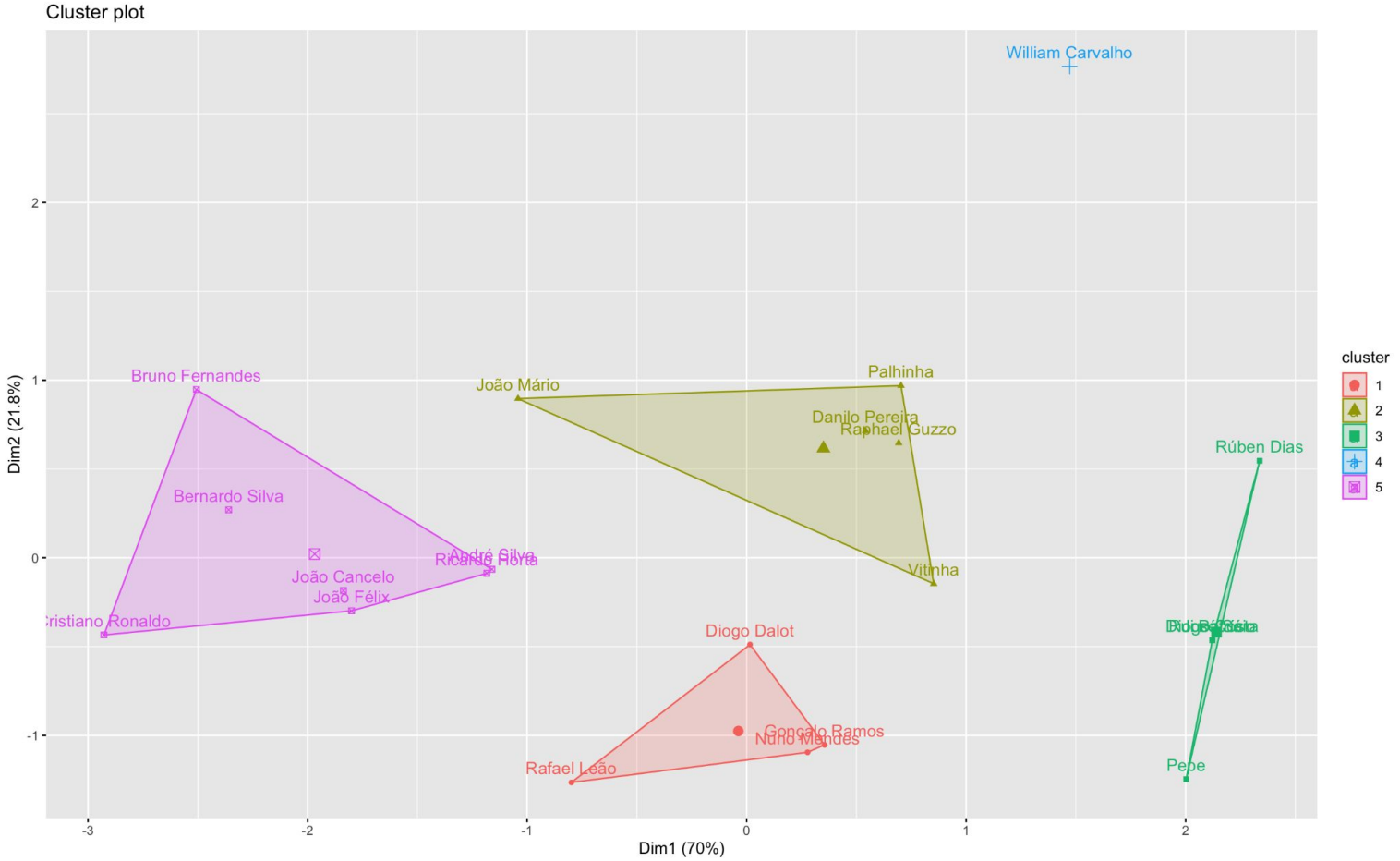
DATA MODELLING- OVERALL GRAPH



CLUSTER ANALYSIS - GHANA PLAYERS



PORTUGAL PLAYERS



CONCLUSION

- Football is very popular in Europe compared to other countries.
- A player's peak performance occurs at the ages of 20-30 years old. Pace, stamina, and value decrease after that.

- Hypothesis 1: In-field attributes like shooting and dribbling do determine a player's salary but there are various other things that are to be considered as well(player's agent, how emotionally is the player attached to the team, etc)
- Hypothesis 2: FIFA is not biased in assigning the overall rating of players based on their attributes
- Hypothesis 3: The cluster analysis classified a team into various cluster based on a player's different attribute. That could be helpful to coach/manager to pick their playing 11 and also find best replacement for a player. Different clusters represented different possible playing positions of players.