FIFA 22 Analysis

Md Shamsul Hoque Khan

2023-10-01

Introduction

This analysis is intended to explore various aspects of FIFA 22 game data, focusing on player attributes, their positions, and other relevant details that define their in-game performance. By utilizing R, we aim to derive insights and make conclusions about player characteristics, abilities, and their representation within the game.

Data Preparation and Exploration

Necessary packages and dateset for this work:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(writexl)
library(ggplot2)
library(randomForest)
```

```
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
file_path <- "FIFA22_official_data.csv"
player_df <- read.csv(file_path)
```

Now we will find out what are the columns we need for only player stats analysis.

```
colnames(player_df)
```

```
##  [1] "ID"                       "Name"
##  [3] "Age"                      "Photo"
##  [5] "Nationality"              "Flag"
##  [7] "Overall"                  "Potential"
##  [9] "Club"                     "Club.Logo"
## [11] "Value"                    "Wage"
## [13] "Special"                  "Preferred.Foot"
## [15] "International.Reputation" "Weak.Foot"
## [17] "Skill.Moves"              "Work.Rate"
## [19] "Body.Type"                "Real.Face"
## [21] "Position"                 "Jersey.Number"
## [23] "Joined"                   "Loaned.From"
## [25] "Contract.Valid.Until"     "Height"
## [27] "Weight"                   "Crossing"
## [29] "Finishing"                "HeadingAccuracy"
## [31] "ShortPassing"             "Volleys"
## [33] "Dribbling"                "Curve"
## [35] "FKAccuracy"               "LongPassing"
## [37] "BallControl"              "Acceleration"
## [39] "SprintSpeed"              "Agility"
## [41] "Reactions"                "Balance"
## [43] "ShotPower"                "Jumping"
## [45] "Stamina"                  "Strength"
## [47] "LongShots"                "Aggression"
## [49] "Interceptions"            "Positioning"
## [51] "Vision"                   "Penalties"
## [53] "Composure"                "Marking"
## [55] "StandingTackle"           "SlidingTackle"
## [57] "GKDiving"                 "GKHandling"
## [59] "GKKicking"                "GKPositioning"
## [61] "GKReflexes"               "Best.Position"
## [63] "Best.Overall.Rating"      "Release.Clause"
## [65] "DefensiveAwareness"
```

```
unique(player_df$Position)
```

```
##  [1] "<span class=\"pos pos18\">CAM" "<span class=\"pos pos11\">LDM"
##  [3] "<span class=\"pos pos24\">RS"  "<span class=\"pos pos13\">RCM"
##  [5] "<span class=\"pos pos7\">LB"   "<span class=\"pos pos9\">RDM"
##  [7] "<span class=\"pos pos15\">LCM" "<span class=\"pos pos28\">SUB"
##  [9] "<span class=\"pos pos26\">LS"  "<span class=\"pos pos12\">RM"
## [11] "<span class=\"pos pos6\">LCB"  "<span class=\"pos pos16\">LM"
## [13] "<span class=\"pos pos3\">RB"   "<span class=\"pos pos10\">CDM"
## [15] "<span class=\"pos pos23\">RW"  "<span class=\"pos pos27\">LW"
## [17] "<span class=\"pos pos25\">ST"  "<span class=\"pos pos14\">CM"
## [19] "<span class=\"pos pos20\">RF"  "<span class=\"pos pos8\">LWB"
## [21] "<span class=\"pos pos17\">RAM" "<span class=\"pos pos21\">CF"
## [23] "<span class=\"pos pos29\">RES" "<span class=\"pos pos22\">LF"
## [25] "<span class=\"pos pos2\">RWB"  "<span class=\"pos pos5\">CB"
## [27] "<span class=\"pos pos4\">RCB"  "nan"
## [29] "<span class=\"pos pos19\">LAM" "<span class=\"pos pos0\">GK"
```

```
class(player_df$Position)
```

```
## [1] "character"
```

It seems the Position column has weird values instead of player positions. The player positions seem to be at the end of the character values. We will only keep the player position and delete rest of the character.

```
player_df[, "Position"] <- str_replace_all(player_df$Position, ".*>", "")

unique(player_df$Position)
```

```
##  [1] "CAM" "LDM" "RS"  "RCM" "LB"  "RDM" "LCM" "SUB" "LS"  "RM"  "LCB" "LM"
## [13] "RB"  "CDM" "RW"  "LW"  "ST"  "CM"  "RF"  "LWB" "RAM" "CF"  "RES" "LF"
## [25] "RWB" "CB"  "RCB" "nan" "LAM" "GK"
```

```
unique(player_df$Best.Position)
```

```
##  [1] "CAM" "CM"  "ST"  "LB"  "CDM" "CB"  "RB"  "LM"  "RW"  "LW"  "CF"  "LWB"
## [13] "RM"  "RWB" "GK"
```

I just realized that Position column also include Sub (SUB) and reserve (RES) as position definitions. In such as, I think it's better to use the best position to define position for a particular player. We can remove Postion column. Also, weight and Body.Type doesn't add any value to a players overall ranking in FIFA.

```
unnecessary_column <- c("Photo", "Flag", "Potential", "Club.Logo",
                        "Special", "International.Reputation",
                        "Weak.Foot", "Body.Type", "Real.Face", "Jersey.Number",
                        "Joined", "Loaned.From", "Contract.Valid.Until",
                        "Weight", "Best.Overall.Rating", "Release.Clause", "Position")

player_stats <- select(player_df, -unnecessary_column)

summary(player_stats)
```

```
##        ID             Name                Age        Nationality
##  Min.   :    27   Length:16710       Min.   :16.00   Length:16710
##  1st Qu.:203891   Class :character   1st Qu.:22.00   Class :character
##  Median :229253   Mode  :character   Median :25.00   Mode  :character
##  Mean   :220560                      Mean   :25.73
##  3rd Qu.:245369                      3rd Qu.:29.00
##  Max.   :264704                      Max.   :54.00
##
##     Overall          Club              Value               Wage
##  Min.   :28.00   Length:16710       Length:16710       Length:16710
##  1st Qu.:63.00   Class :character   Class :character   Class :character
##  Median :68.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :67.65
##  3rd Qu.:72.00
##  Max.   :93.00
##
##  Preferred.Foot      Skill.Moves     Work.Rate            Height
##  Length:16710       Min.   :1.000   Length:16710       Length:16710
##  Class :character   1st Qu.:2.000   Class :character   Class :character
##  Mode  :character   Median :2.000   Mode  :character   Mode  :character
##                     Mean   :2.475
##                     3rd Qu.:3.000
##                     Max.   :5.000
##
##     Crossing       Finishing     HeadingAccuracy  ShortPassing
##  Min.   : 7.00   Min.   : 3.00   Min.   : 5.00   Min.   : 8.00
##  1st Qu.:42.00   1st Qu.:34.00   1st Qu.:46.00   1st Qu.:57.00
##  Median :57.00   Median :53.00   Median :57.00   Median :64.00
##  Mean   :52.21   Mean   :48.73   Mean   :54.12   Mean   :61.31
##  3rd Qu.:65.00   3rd Qu.:64.00   3rd Qu.:66.00   3rd Qu.:70.00
##  Max.   :94.00   Max.   :95.00   Max.   :93.00   Max.   :94.00
##
##     Volleys        Dribbling        Curve        FKAccuracy
##  Min.   : 4.00   Min.   : 5.00   Min.   : 6.00   Min.   : 4.00
##  1st Qu.:33.00   1st Qu.:54.00   1st Qu.:39.00   1st Qu.:33.00
##  Median :48.00   Median :63.50   Median :53.00   Median :44.00
##  Mean   :45.65   Mean   :58.59   Mean   :50.57   Mean   :45.07
##  3rd Qu.:59.00   3rd Qu.:70.00   3rd Qu.:64.00   3rd Qu.:59.00
##  Max.   :90.00   Max.   :96.00   Max.   :94.00   Max.   :94.00
##  NA's   :37                      NA's   :37
##   LongPassing     BallControl     Acceleration    SprintSpeed
##  Min.   : 9.00   Min.   : 8.00   Min.   :13.00   Min.   :15.00
##  1st Qu.:47.00   1st Qu.:58.00   1st Qu.:58.00   1st Qu.:59.00
##  Median :58.00   Median :65.00   Median :68.00   Median :68.00
##  Mean   :55.45   Mean   :61.39   Mean   :65.68   Mean   :65.78
##  3rd Qu.:66.00   3rd Qu.:71.00   3rd Qu.:76.00   3rd Qu.:76.00
##  Max.   :93.00   Max.   :96.00   Max.   :97.00   Max.   :97.00
##
##     Agility        Reactions        Balance        ShotPower        Jumping
##  Min.   :18.00   Min.   :28.00   Min.   :19.00   Min.   :12.00   Min.   :22.0
##  1st Qu.:57.00   1st Qu.:58.00   1st Qu.:57.00   1st Qu.:52.00   1st Qu.:59.0
##  Median :68.00   Median :64.00   Median :67.00   Median :62.00   Median :67.0
##  Mean   :65.22   Mean   :63.59   Mean   :64.94   Mean   :60.64   Mean   :65.9
##  3rd Qu.:75.00   3rd Qu.:69.00   3rd Qu.:75.00   3rd Qu.:70.00   3rd Qu.:74.0
##  Max.   :96.00   Max.   :96.00   Max.   :96.00   Max.   :95.00   Max.   :95.0
##  NA's   :37                      NA's   :37                      NA's   :37
##     Stamina        Strength       LongShots        Aggression    Interceptions
##  Min.   :13.0   Min.   :18.00   Min.   : 4.00   Min.   :11.00   Min.   : 4.00
##  1st Qu.:57.0   1st Qu.:59.00   1st Qu.:37.00   1st Qu.:47.00   1st Qu.:28.00
##  Median :67.0   Median :68.00   Median :55.00   Median :61.00   Median :55.00
##  Mean   :64.2   Mean   :66.27   Mean   :49.99   Mean   :57.96   Mean   :48.06
##  3rd Qu.:75.0   3rd Qu.:75.00   3rd Qu.:65.00   3rd Qu.:71.00   3rd Qu.:65.00
##  Max.   :97.0   Max.   :97.00   Max.   :94.00   Max.   :95.00   Max.   :95.00
##                                                                 NA's   :8
##   Positioning        Vision       Penalties      Composure      Marking
##  Min.   : 3.00   Min.   :10.00   Min.   : 7.00   Min.   :12    Min.   : 4.00
##  1st Qu.:44.00   1st Qu.:48.00   1st Qu.:41.00   1st Qu.:55    1st Qu.:27.00
##  Median :58.00   Median :58.00   Median :52.00   Median :62    Median :53.00
##  Mean   :53.15   Mean   :56.38   Mean   :50.62   Mean   :61    Mean   :48.02
##  3rd Qu.:67.00   3rd Qu.:66.00   3rd Qu.:62.00   3rd Qu.:69    3rd Qu.:67.00
##  Max.   :96.00   Max.   :95.00   Max.   :96.00   Max.   :96    Max.   :94.00
##  NA’s   :8       NA’s   :37                      NA’s   :251   NA’s   :15818
##  StandingTackle  SlidingTackle     GKDiving       GKHandling
##  Min.   : 3.00   Min.   : 6.00   Min.   : 1.00   Min.   : 1.00
##  1st Qu.:29.00   1st Qu.:26.00   1st Qu.: 8.00   1st Qu.: 8.00
##  Median :57.00   Median :54.00   Median :11.00   Median :11.00
##  Mean   :49.16   Mean   :46.83   Mean   :15.68   Mean   :15.52
##  3rd Qu.:67.00   3rd Qu.:65.00   3rd Qu.:14.00   3rd Qu.:14.00
##  Max.   :93.00   Max.   :95.00   Max.   :91.00   Max.   :92.00
##                  NA’s   :37
##    GKKicking     GKPositioning    GKReflexes     Best.Position
##  Min.   : 1.00   Min.   : 1.00   Min.   : 1.00   Length:16710
##  1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.00   Class :character
##  Median :11.00   Median :11.00   Median :11.00   Mode  :character
##  Mean   :15.46   Mean   :15.58   Mean   :15.79
##  3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00
##  Max.   :93.00   Max.   :93.00   Max.   :90.00
##
##  DefensiveAwareness
##  Min.   : 3.00
##  1st Qu.:30.00
##  Median :54.00
##  Mean   :48.02
##  3rd Qu.:65.00
##  Max.   :93.00
##  NA’s   :892
```

It seems there are lots of NA values. I don’t want to drop all these values because there are some really good players, such as Bruno Fernandes and L. Goretzka, who have missing values. They are the face of the game. To solve this, I want to categorize players based on their position and overall rating. Then I will take the median Scoring of a variable for that particular position within that overall rating range. I want to create a separate table to avoid confusion.

```
# Create a new column that groups 'Overall' into ranges

player_stats <- player_stats %>%mutate(Overall_group = case_when(
    Overall >= 80 ~ "80-99",
    Overall >= 70 ~ "70-79",
    Overall >= 60 ~ "60-69",
    TRUE ~ "<60"))


# Compute median scores for 'Marking' and 'Volleys' columns in each group

medains <- player_stats %>% group_by(Best.Position, Overall_group) %>%
  summarize(Marking_median = median(Marking, na.rm = TRUE),
            Volleys_median = median(Volleys, na.rm = TRUE),
            Curve_median = median(Curve, na.rm = TRUE),
            Agility_median = median(Agility, na.rm = TRUE),
            Balance_median = median(Balance, na.rm = TRUE),
            Jumping_median = median(Jumping, na.rm = TRUE),
            Interceptions_median = median(Interceptions, na.rm = TRUE),
            Positioning_median = median(Positioning, na.rm = TRUE),
            Vision_median = median(Vision, na.rm = TRUE),
            Composure_median = median(Composure, na.rm = TRUE),
            SlidingTackle_median = median(SlidingTackle, na.rm = TRUE),
            DefensiveAwareness_median = median(DefensiveAwareness, na.rm = TRUE))
```

```
## ‘summarise()‘ has grouped output by ’Best.Position’. You can override using the
## ‘.groups‘ argument.
```

```
player_stats <- left_join(player_stats, medains, by = c("Best.Position", "Overall_group"))


# Now Replace the NaN values with the medians

player_stats_fixed <- player_stats %>%
  mutate("Marking" = ifelse(is.na(Marking), Marking_median, Marking),
         "Volleys" = ifelse(is.na(Volleys), Volleys_median, Volleys),
         "Curve" = ifelse(is.na(Curve), Curve_median, Curve),
         "Agility" = ifelse(is.na(Agility), Agility_median, Agility),
         "Balance" = ifelse(is.na(Balance), Balance_median, Balance),
         "Jumping" = ifelse(is.na(Jumping), Jumping_median, Jumping),
         "Interceptions" = ifelse(is.na(Interceptions), Interceptions_median, Interceptions),
         "Positioning" = ifelse(is.na(Positioning), Positioning_median, Positioning),
         "Composure" = ifelse(is.na(Composure), Composure_median, Composure),
         "SlidingTackle" = ifelse(is.na(SlidingTackle), SlidingTackle_median, SlidingTackle),
         "DefensiveAwareness" = ifelse(is.na(DefensiveAwareness), DefensiveAwareness_median, DefensiveAw
         "Vision" = ifelse(is.na(Vision), Vision_median, Vision))


summary(player_stats_fixed)
```

```
##        ID               Name                Age        Nationality
##  Min.   :    27   Length:16710       Min.   :16.00   Length:16710
##  1st Qu.:203891   Class :character   1st Qu.:22.00   Class :character
##  Median :229253   Mode  :character   Median :25.00   Mode  :character
##  Mean   :220560                      Mean   :25.73
##  3rd Qu.:245369                      3rd Qu.:29.00
##  Max.   :264704                      Max.   :54.00
##
##     Overall          Club              Value               Wage
##  Min.   :28.00   Length:16710       Length:16710       Length:16710
##  1st Qu.:63.00   Class :character   Class :character   Class :character
##  Median :68.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :67.65
##  3rd Qu.:72.00
##  Max.   :93.00
##
##  Preferred.Foot      Skill.Moves     Work.Rate            Height
##  Length:16710       Min.   :1.000   Length:16710       Length:16710
##  Class :character   1st Qu.:2.000   Class :character   Class :character
##  Mode  :character   Median :2.000   Mode  :character   Mode  :character
##                     Mean   :2.475
##                     3rd Qu.:3.000
##                     Max.   :5.000
##
##     Crossing       Finishing     HeadingAccuracy  ShortPassing
##  Min.   : 7.00   Min.   : 3.00   Min.   : 5.00   Min.   : 8.00
##  1st Qu.:42.00   1st Qu.:34.00   1st Qu.:46.00   1st Qu.:57.00
##  Median :57.00   Median :53.00   Median :57.00   Median :64.00
##  Mean   :52.21   Mean   :48.73   Mean   :54.12   Mean   :61.31
##  3rd Qu.:65.00   3rd Qu.:64.00   3rd Qu.:66.00   3rd Qu.:70.00
##  Max.   :94.00   Max.   :95.00   Max.   :93.00   Max.   :94.00
##
##     Volleys        Dribbling         Curve         FKAccuracy
##  Min.   : 4.00   Min.   : 5.00   Min.   : 6.00   Min.   : 4.00
##  1st Qu.:33.00   1st Qu.:54.00   1st Qu.:39.00   1st Qu.:33.00
##  Median :48.00   Median :63.50   Median :53.00   Median :44.00
##  Mean   :45.68   Mean   :58.59   Mean   :50.59   Mean   :45.07
##  3rd Qu.:59.00   3rd Qu.:70.00   3rd Qu.:64.00   3rd Qu.:59.00
##  Max.   :90.00   Max.   :96.00   Max.   :94.00   Max.   :94.00
##
##   LongPassing     BallControl     Acceleration    SprintSpeed
##  Min.   : 9.00   Min.   : 8.00   Min.   :13.00   Min.   :15.00
##  1st Qu.:47.00   1st Qu.:58.00   1st Qu.:58.00   1st Qu.:59.00
##  Median :58.00   Median :65.00   Median :68.00   Median :68.00
##  Mean   :55.45   Mean   :61.39   Mean   :65.68   Mean   :65.78
##  3rd Qu.:66.00   3rd Qu.:71.00   3rd Qu.:76.00   3rd Qu.:76.00
##  Max.   :93.00   Max.   :96.00   Max.   :97.00   Max.   :97.00
##
##     Agility        Reactions        Balance        ShotPower
##  Min.   :18.00   Min.   :28.00   Min.   :19.00   Min.   :12.00
##  1st Qu.:57.00   1st Qu.:58.00   1st Qu.:57.00   1st Qu.:52.00
##  Median :68.00   Median :64.00   Median :67.00   Median :62.00
##  Mean   :65.23   Mean   :63.59   Mean   :64.95   Mean   :60.64
##  3rd Qu.:75.00   3rd Qu.:69.00   3rd Qu.:75.00   3rd Qu.:70.00
##  Max.   :96.00   Max.   :96.00   Max.   :96.00   Max.   :95.00
##
##     Jumping         Stamina        Strength       LongShots       Aggression
##  Min.   :22.00   Min.   :13.0   Min.   :18.00   Min.   : 4.00   Min.   :11.00
##  1st Qu.:59.00   1st Qu.:57.0   1st Qu.:59.00   1st Qu.:37.00   1st Qu.:47.00
##  Median :67.00   Median :67.0   Median :68.00   Median :55.00   Median :61.00
##  Mean   :65.91   Mean   :64.2   Mean   :66.27   Mean   :49.99   Mean   :57.96
##  3rd Qu.:74.00   3rd Qu.:75.0   3rd Qu.:75.00   3rd Qu.:65.00   3rd Qu.:71.00
##  Max.   :95.00   Max.   :97.0   Max.   :97.00   Max.   :94.00   Max.   :95.00
##
##  Interceptions    Positioning        Vision        Penalties
##  Min.   : 4.00   Min.   : 3.00   Min.   :10.00   Min.   : 7.00
##  1st Qu.:28.00   1st Qu.:44.00   1st Qu.:48.00   1st Qu.:41.00
##  Median :55.00   Median :58.00   Median :58.00   Median :52.00
##  Mean   :48.06   Mean   :53.15   Mean   :56.39   Mean   :50.62
##  3rd Qu.:65.00   3rd Qu.:67.00   3rd Qu.:66.00   3rd Qu.:62.00
##  Max.   :95.00   Max.   :96.00   Max.   :95.00   Max.   :96.00
```

```
##
##    Composure        Marking      StandingTackle  SlidingTackle
##  Min.   :12.00   Min.   : 4.00   Min.   : 3.00   Min.   : 6.00
##  1st Qu.:55.00   1st Qu.:29.00   1st Qu.:29.00   1st Qu.:26.00
##  Median :62.00   Median :47.50   Median :57.00   Median :54.00
##  Mean   :61.08   Mean   :46.92   Mean   :49.16   Mean   :46.82
##  3rd Qu.:69.00   3rd Qu.:64.00   3rd Qu.:67.00   3rd Qu.:65.00
##  Max.   :96.00   Max.   :94.00   Max.   :93.00   Max.   :95.00
##                  NA's   :132
##     GKDiving       GKHandling      GKKicking     GKPositioning
##  Min.   : 1.00   Min.   : 1.00   Min.   : 1.00   Min.   : 1.00
##  1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.00
##  Median :11.00   Median :11.00   Median :11.00   Median :11.00
##  Mean   :15.68   Mean   :15.52   Mean   :15.46   Mean   :15.58
##  3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00
##  Max.   :91.00   Max.   :92.00   Max.   :93.00   Max.   :93.00
##
##    GKReflexes    Best.Position      DefensiveAwareness Overall_group
##  Min.   : 1.00   Length:16710       Min.   : 3.00      Length:16710
##  1st Qu.: 8.00   Class :character   1st Qu.:31.00      Class :character
##  Median :11.00   Mode  :character   Median :54.00      Mode  :character
##  Mean   :15.79                      Mean   :48.08
##  3rd Qu.:14.00                      3rd Qu.:65.00
##  Max.   :90.00                      Max.   :93.00
##
##  Marking_median  Volleys_median   Curve_median   Agility_median
##  Min.   :12.00   Min.   : 8.00   Min.   :12.00   Min.   :29.00
##  1st Qu.:29.00   1st Qu.:35.00   1st Qu.:41.00   1st Qu.:57.00
##  Median :47.50   Median :48.00   Median :52.00   Median :69.00
##  Mean   :46.88   Mean   :45.31   Mean   :50.65   Mean   :65.62
##  3rd Qu.:64.00   3rd Qu.:59.00   3rd Qu.:61.00   3rd Qu.:74.00
##  Max.   :86.00   Max.   :81.00   Max.   :84.00   Max.   :86.00
##  NA's   :132
##  Balance_median  Jumping_median  Interceptions_median Positioning_median
##  Min.   :39.00   Min.   :32.00   Min.   :11.00        Min.   : 7.00
##  1st Qu.:56.00   1st Qu.:62.00   1st Qu.:30.00        1st Qu.:41.00
##  Median :68.00   Median :66.00   Median :46.00        Median :60.00
##  Mean   :65.47   Mean   :66.68   Mean   :47.67        Mean   :53.24
##  3rd Qu.:74.00   3rd Qu.:72.00   3rd Qu.:64.00        3rd Qu.:65.00
##  Max.   :84.00   Max.   :82.00   Max.   :83.00        Max.   :85.00
##
##  Vision_median   Composure_median SlidingTackle_median
##  Min.   :32.00   Min.   :31.50    Min.   :12.00
##  1st Qu.:50.00   1st Qu.:58.00    1st Qu.:27.00
##  Median :60.00   Median :60.00    Median :45.50
##  Mean   :56.64   Mean   :61.38    Mean   :46.25
##  3rd Qu.:65.00   3rd Qu.:68.00    3rd Qu.:64.00
##  Max.   :84.00   Max.   :83.00    Max.   :82.00
##
##  DefensiveAwareness_median
##  Min.   : 9.00
##  1st Qu.:33.00
##  Median :46.00
##  Mean   :47.81
##  3rd Qu.:64.00
##  Max.   :84.00
##
```

It seems like we still have 132 NA values in Marking. In this case, I will inspect whether these players has good overall. If so, I will not drop the NAs. But if not, then I will drop the values because our data set has a lot of observations to draw meaningful insight with these values.

```
inspection_marking <- player_stats_fixed %>%
  filter(is.na(Marking) & Overall > 75) %>%
  select(ID, Name, Best.Position, Overall, Marking)
```

A close inspection of the inspection_marking data frame shows that we can't drop the NA values becasue there are some big names, such as Neymar Jr.

```
unique(inspection_marking$Best.Position)
```

```
## [1] "LM" "LW" "RM"
```

Really interesting thing is that all NA value players are wingers, LM, LW, and RM. For this positions, Marking attribute is not important and doesn't contribute to their overall rating. For details: https://fifaforums.easports.com/en/discussion/277545/how-player-rating-is-calculated-it-is-a-total-mess

So, we can give them a random low score of 40 in Marking.

```
player_stats_fixed$Marking[is.na(player_stats_fixed$Marking)] <- 40

sum(is.na(player_stats_fixed$Marking))
```

```
## [1] 0
```

Now we need to separate Goal-Keepers from the outfield players because Goal-Keepers have different attributes, and we are only interested in the Out-field players.

```
outfield_player <- player_stats_fixed %>% filter(Best.Position != "GK") %>%
  select(-c("GKDiving", "GKHandling", "GKKicking", "GKPositioning", "GKReflexes"))

outfield_player <- outfield_player[,-((ncol(outfield_player) - 11) : ncol(outfield_player))]
```

The Work.Rate column has Attacking and Defensive work rate of a player. I want to have two different columns for each.

```
# Replace "N/A/ N/A" with "Medium/ Medium"
outfield_player$Work.Rate[outfield_player$Work.Rate == "N/A/ N/A"] <- "Medium/ Medium"

# Separate Work.Rate into two new columns
outfield_player <- outfield_player %>%
  separate(Work.Rate, into = c("Attack.work.rate", "Defensive.work.rate"), sep = "/")
```

As I was exploring the data, I realized the Name column has some Values that start with number, such as “12 Roberto Carlos”.

```
outfield_player$Name <- str_remove(outfield_player$Name, "^\\d+\\s")

colnames(outfield_player)
```

```
##  [1] "ID"                  "Name"                "Age"
##  [4] "Nationality"         "Overall"             "Club"
##  [7] "Value"               "Wage"                "Preferred.Foot"
## [10] "Skill.Moves"         "Attack.work.rate"    "Defensive.work.rate"
## [13] "Height"              "Crossing"            "Finishing"
## [16] "HeadingAccuracy"     "ShortPassing"        "Volleys"
## [19] "Dribbling"           "Curve"               "FKAccuracy"
## [22] "LongPassing"         "BallControl"         "Acceleration"
## [25] "SprintSpeed"         "Agility"             "Reactions"
## [28] "Balance"             "ShotPower"           "Jumping"
## [31] "Stamina"             "Strength"            "LongShots"
## [34] "Aggression"          "Interceptions"       "Positioning"
## [37] "Vision"              "Penalties"           "Composure"
## [40] "Marking"             "StandingTackle"      "SlidingTackle"
## [43] "Best.Position"       "DefensiveAwareness"  "Overall_group"
```

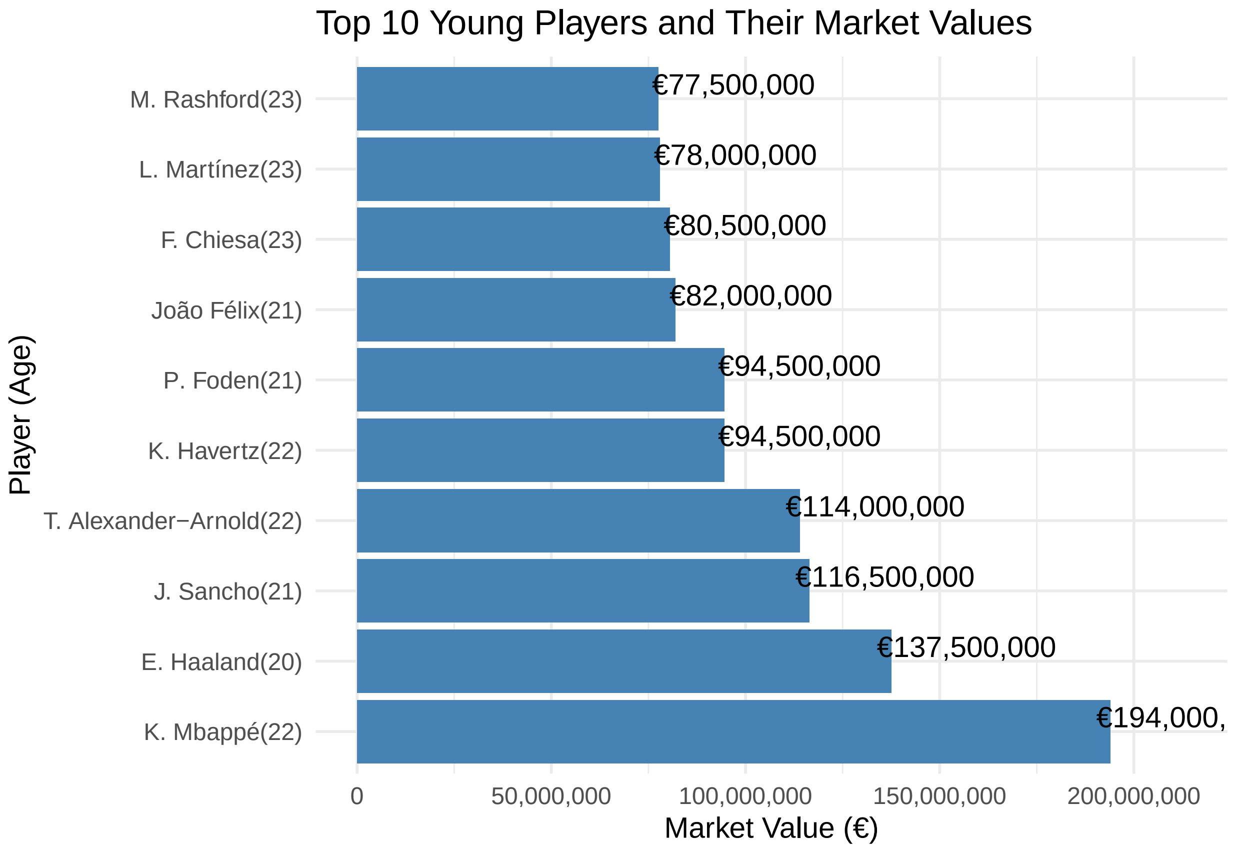
Now the data is clean!!! I can have the clean dataset in excel format:

```
write_xlsx(outfield_player, "outfield_player.1.xlsx")
```

Detailed Analysis

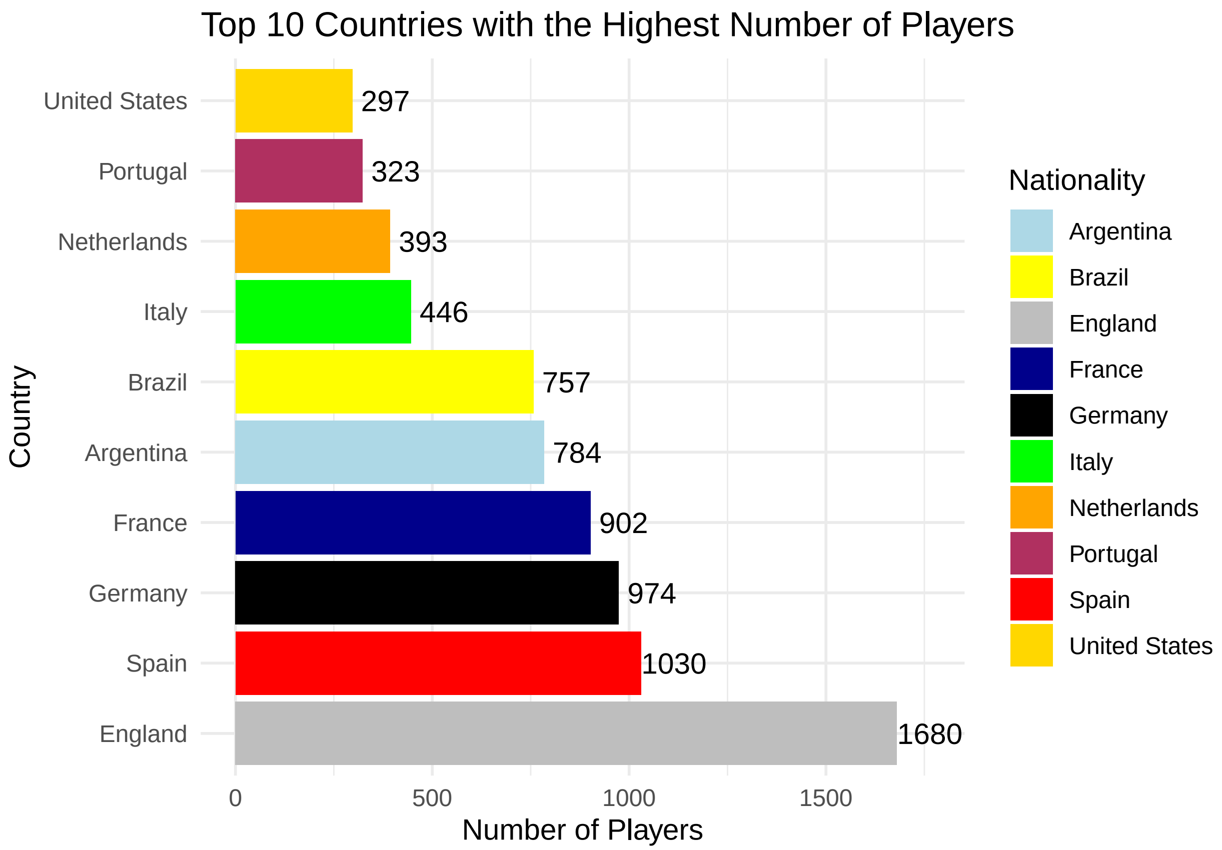
Now I will explore the data to gather different insight from it. Firstly, I want to find out the best young players and their market values.

```
young_players <- outfield_player %>%
  filter(Age <= 23) %>%
  mutate(Value = gsub("€", "", Value),  # remove Euro symbol
         Value = gsub("M", "e6", Value),  # replace M with e6
         Value = gsub("K", "e3", Value),  # replace K with e3
         Value = as.numeric(gsub("e6", "",
                                 ifelse(grepl("e6", Value), as.character(
                                   as.numeric(gsub(
                                     "e6", "", Value)) * 1e6), Value))),# for millions
         Value = as.numeric(gsub(
           "e3", "", ifelse(grepl("e3", Value), as.character(
             as.numeric(gsub("e3", "", Value)) * 1e3), Value)))) %>% # for thousands
  arrange(desc(Value)) %>%
  head(10) %>%
  mutate(label = paste0("€", scales::comma(Value)),
         player_age = paste(Name, "(", Age, ")", sep = ""))



ggplot(young_players, aes(x = reorder(player_age, -Value), y = Value)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = label), vjust = -0.2, nudge_y = max(
    young_players$Value) * 0.1) + # Adjusting the label position here
  coord_flip() +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Top 10 Young Players and Their Market Values",
       x = "Player (Age)",
       y = "Market Value (€)") +
  theme_minimal()
```



Let's see the top 10 countries with highest number of players.

```
# Group by Nationality and count the number of players from each country
top_countries <- outfield_player %>%
  group_by(Nationality) %>%
  summarize(Num_of_Players = n()) %>%
  arrange(desc(Num_of_Players)) %>%
  head(10)

# Define the colors for each country
country_colors <- c(
  "England" = "gray",
  "Spain" = "red",
  "Germany" = "black",
  "France" = "darkblue",
  "Argentina" = "lightblue",
  "Brazil" = "yellow",
  "Italy" = "green",
  "Netherlands" = "orange",
  "Portugal" = "maroon",
  "United States" = "gold"
)

ggplot(top_countries, aes(x = reorder(Nationality, -Num_of_Players), y = Num_of_Players)) +
  geom_bar(stat = "identity", aes(fill = Nationality)) +
  geom_text(aes(label = Num_of_Players, y = Num_of_Players + max(top_countries$Num_of_Players) * 0.05))
  coord_flip() +
  scale_fill_manual(values = country_colors) +  # Apply custom color mapping here
  labs(title = "Top 10 Countries with the Highest Number of Players",
       x = "Country",
       y = "Number of Players") +
  theme_minimal()
```



There is no surprise that England is leading the chart followed by Spain. The reason is England and Spain have more league divisions in the game than any other countries.

Let's find out the details about top 10 players in the game.

```
# Arrange the dataframe by Overall Ratings and pick the top 10
top_players <- outfield_player %>%
  arrange(desc(Overall)) %>%
  head(10)

# View the result
```

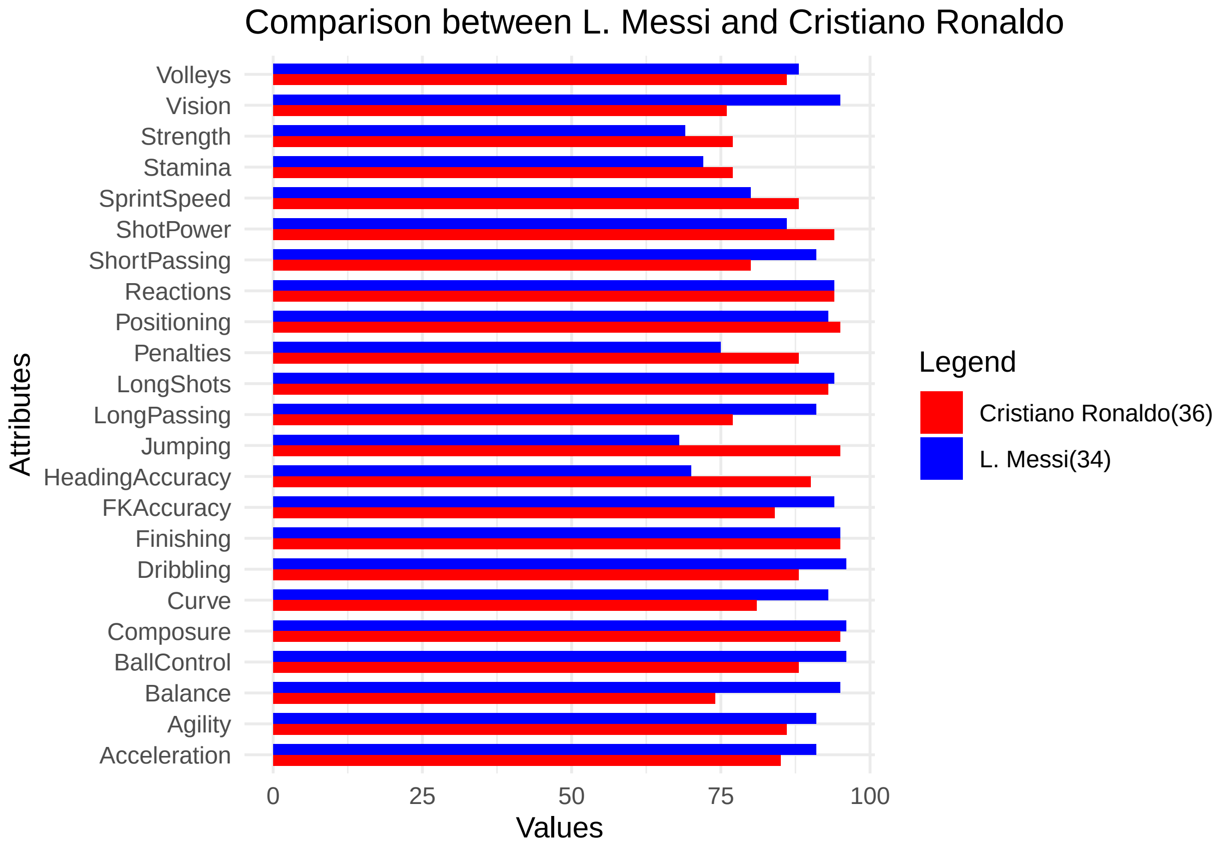
```r
print(top_players[, c("Name", "Age", "Nationality", "Club", "Overall")])
```

```
##                  Name Age Nationality                Club Overall
## 1            L. Messi  34   Argentina Paris Saint-Germain      93
## 2      R. Lewandowski  32      Poland   FC Bayern München      92
## 3        K. De Bruyne  30     Belgium     Manchester City      91
## 4   Cristiano Ronaldo  36    Portugal   Manchester United      91
## 5           Neymar Jr  29      Brazil Paris Saint-Germain      91
## 6           K. Mbappé  22      France Paris Saint-Germain      91
## 7             H. Kane  27     England   Tottenham Hotspur      90
## 8            N. Kanté  30      France             Chelsea      90
## 9          J. Kimmich  26     Germany   FC Bayern München      89
## 10           Casemiro  29      Brazil      Real Madrid CF      89
```

Looks like L. Messi is leading the chart with overall rating of 93. Important point to observe here is that the top 4 players are over the age of 30 or 30. The youngest amoung the top 10 is K. Mbappé.

A Fifa 22 player will most likely want to see comparison between L. Messi and Cristiano Ronaldo.

```r
# Selecting Players
players_comparison <- outfield_player %>%
  filter(Name %in% c('L. Messi', 'Cristiano Ronaldo'))

# Adding Age to their Names for Legend
players_comparison <- players_comparison %>%
  mutate(Legend = paste(Name, "(", Age, ")", sep = ""))

# Selecting relevant columns
relevant_columns <- c("Legend", "Finishing", "HeadingAccuracy", "Volleys", "Dribbling",
                      "FKAccuracy", "BallControl", "Acceleration", "SprintSpeed",
                      "Agility", "Reactions", "Balance", "ShotPower", "Jumping",
                      "Stamina", "Strength", "LongShots", "Positioning",
                      "Vision", "Penalties", "Composure", "Curve", "LongPassing",
                      "ShortPassing")

# Filtering to selected columns and transforming the data
players_long <- players_comparison[ , relevant_columns] %>%
  pivot_longer(cols = -Legend, names_to = "Attribute", values_to = "Value")

# Assign colors to each Legend
unique_legends <- unique(players_comparison$Legend)
colors <- setNames(c('blue', 'red'), unique_legends)

# Plotting
ggplot(players_long, aes(x = Attribute, y = Value, fill = Legend)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  coord_flip() +
  labs(title = "Comparison between L. Messi and Cristiano Ronaldo",
       x = "Attributes",
       y = "Values") +
  scale_fill_manual(values = colors) +
  theme_minimal()
```

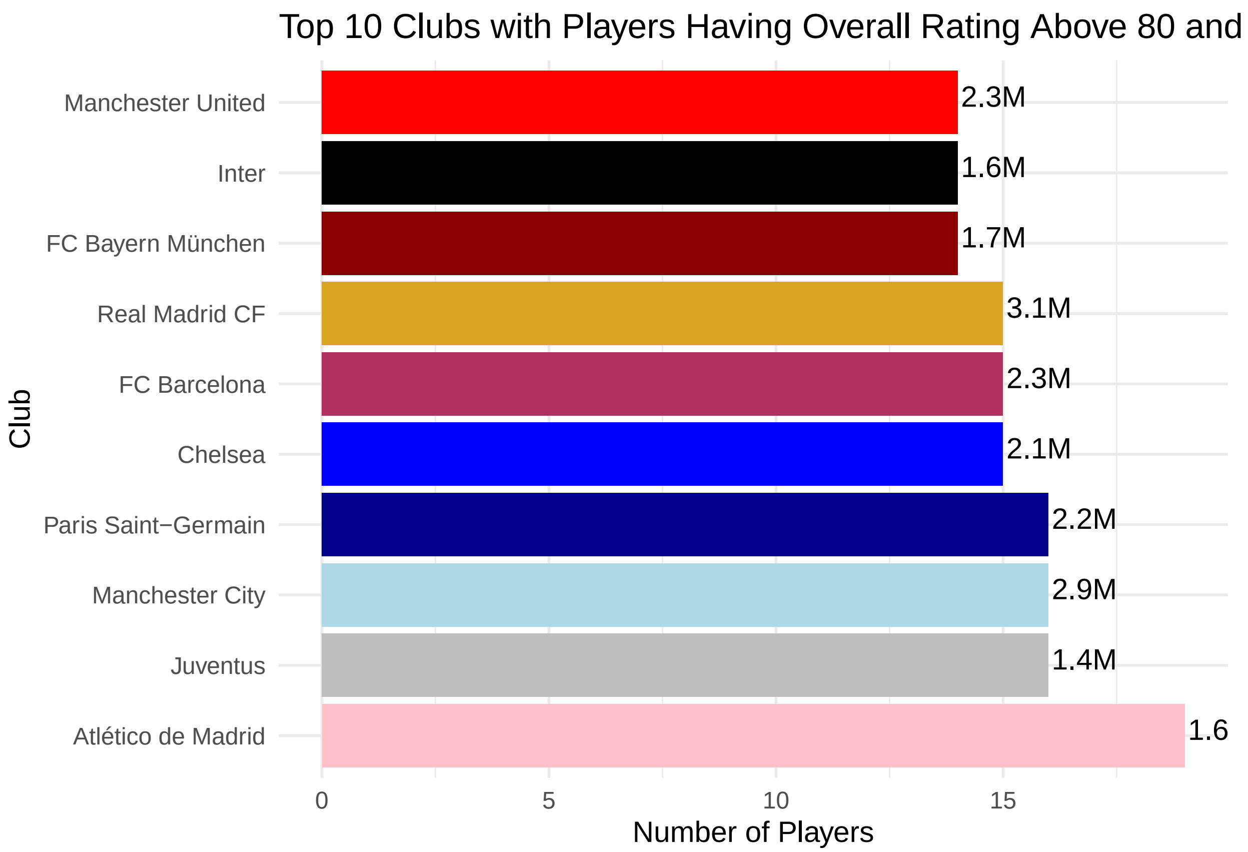


I can see that the image you provided is a comparison between Lionel Messi and Cristiano Ronaldo. Ronaldo has a higher rating in the following attributes: Heading Accuracy, Strength, Stamina, Jumping, Sprint Speed, Penalties, Positioning, Shot Power. They are equal in Finishing and Reactions. Overall, Messi has a higher rating than Ronaldo in 8 out of 24 attributes. However, it is important to note that these are just individual attributes. It is also important to consider the players' overall playing style and their ability to fit into a particular team.

Messi is a more creative player who is known for his vision, passing, and dribbling skills. Ronaldo is a more physical player who is known for his pace, heading ability, and finishing ability.

Lastly, let's find out the top 10 Clubs with Players having overall rating above 80 and how much the clubs spent on these players' wage.

```
# Transforming the Wage column
high_rated_players <- outfield_player %>%
  filter(Overall > 80) %>%
  mutate(Wage = gsub("€", "", Wage),  # remove Euro symbol
         Wage = gsub("M", "e6", Wage),  # replace M with e6 to convert to numeric later
         Wage = gsub("K", "e3", Wage),  # replace K with e3 to convert to numeric later
         Wage = as.numeric(gsub("e6", "", ifelse(grepl(
           "e6", Wage), as.character(as.numeric(gsub(
             "e6", "", Wage)) * 1e6), Wage))),  # for millions
         Wage = as.numeric(gsub("e3", "", ifelse(grepl(
           "e3", Wage), as.character(as.numeric(gsub(
             "e3", "", Wage)) * 1e3), Wage)))) %>%
  group_by(Club) %>%
  summarize(Total_Wage = sum(Wage, na.rm = TRUE), Num_of_Players = n()) %>%
  arrange(desc(Num_of_Players)) %>%
  head(10)

# Defining the colors for each club
club_colors <- c(
  "Atlético de Madrid" = "pink",
  "Juventus" = "gray",
  "Manchester City" = "lightblue",
  "Paris Saint-Germain" = "darkblue",
  "Chelsea" = "blue",
  "FC Barcelona" = "maroon",
  "Real Madrid CF" = "goldenrod",
  "FC Bayern München" = "darkred",
  "Inter" = "black",
  "Manchester United" = "red"
)

# Plotting
ggplot(high_rated_players, aes(x = reorder(Club, -Num_of_Players),
                               y = Num_of_Players, fill = Club)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  geom_text(aes(label = paste0(round(Total_Wage / 1e6, 1), "M")),
            vjust = 0.2, hjust = -0.05) +
  scale_fill_manual(values = club_colors) +
  coord_flip() +
  labs(title = "Top 10 Clubs with Players Having Overall Rating Above 80 and Spent on Wage",
       x = "Club",
       y = "Number of Players") +
  theme_minimal()
```



The chart shows Atlético de Madrid has the highest number of players (19 players) over overall rating 80, a fact which could be surprising to any FIFA 22 game player because one would think that clubs like Manchester City, Real Madrid CF, or FC Barcelona should be leading this chart. Although, Real Madrid CF has only 15 players over the overall rating of 80, the club spent 3.1M on wage, which is the highest among the 10 clubs.

What leads to higher market value for a player in forward position?

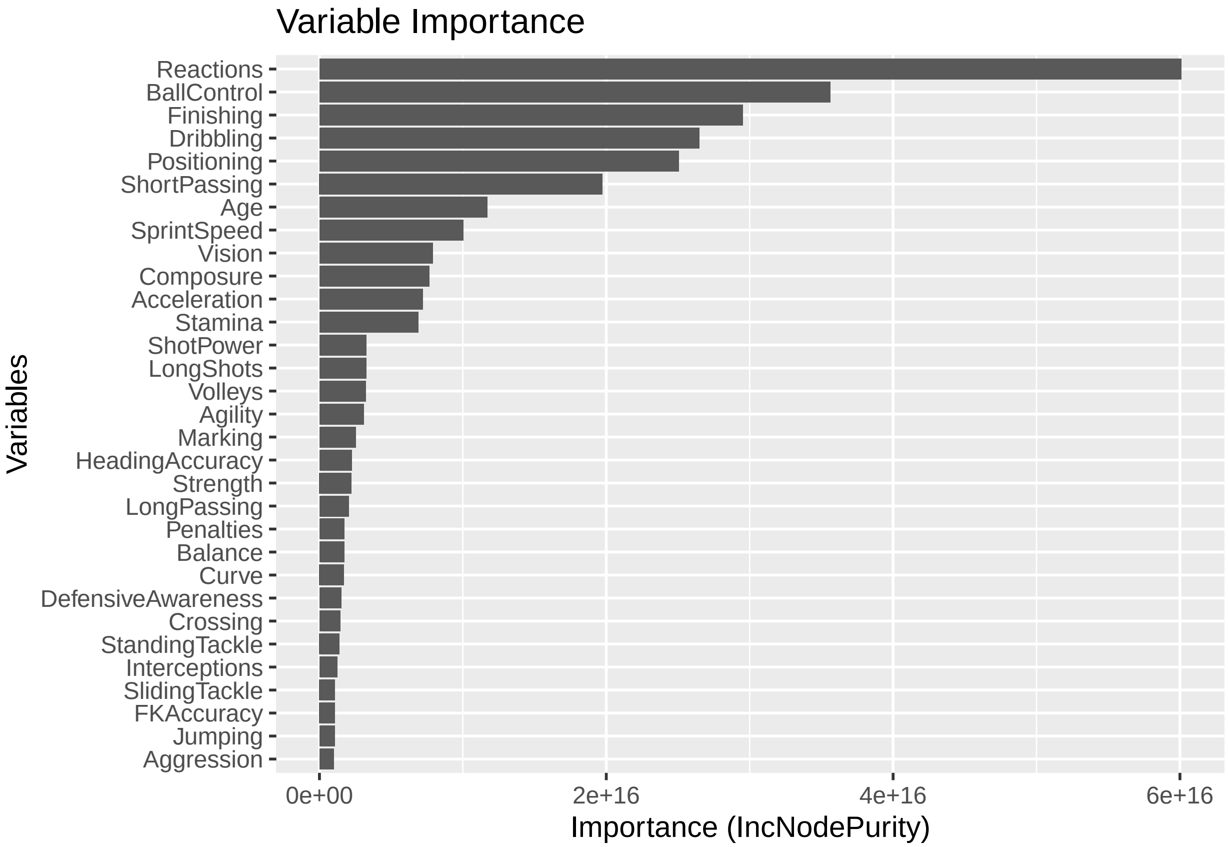
I am curious about what are the attributes the make a player highly valued than other players in the game. For this analysis, I will consider players in the forward position: "ST", "RW", "LW", and "CF". I am selecting only the forward players because it will save computational time.

Based on my experience with the game, I am assuming that the top attributes are finishing, reaction, dribbling, and positioning. There can be many more.

```
# Filter the dataframe to contain only players with ST, RW, LW, CF positions
selected_positions <- c("ST", "RW", "LW", "CF")
filtered_player <- outfield_player %>%
  filter(Best.Position %in% selected_positions)

# Convert the Value column to numeric
filtered_player <- filtered_player %>%
  mutate(Value_numeric = case_when(
    str_detect(Value, "M") ~ as.numeric(str_extract(Value, "[0-9.]+")) * 1e6, # Convert M to million
    str_detect(Value, "K") ~ as.numeric(str_extract(Value, "[0-9.]+")) * 1e3, # Convert K to thousand
    TRUE ~ as.numeric(str_replace(Value, "€", "")) # Assume the rest are in numeric format and remove E
  ))

# Remove columns that are not necessary for our analysis
filtered_player <- filtered_player %>%
  select(-c(1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 43, 45))
```

I am picking Random Forest model to capture the relationships.

```
rf_model <- randomForest(Value_numeric ~ ., data = filtered_player)
importance_df <- as.data.frame(importance(rf_model))
importance_df$variable <- rownames(importance_df)
ggplot(importance_df, aes(x = reorder(variable, IncNodePurity), y = IncNodePurity)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Variable Importance", x = "Variables", y = "Importance (IncNodePurity)")
```



The graph shows that “Reactions” is by far the most important attribute that influences market price of a player. This attribute is followed by “BallControl”, “Finishing”, “Dribbling”, “Positioning”, “ShortPassing”, and “Age”. Having Short Passing attribute in top 5 is a surprise for me.

It seems like the market value of a creative forward player is higher because these are the attributes of such player.

Lastly, it makes sense to have Interceptions, Defensive awarness, Sliding tackle, Jumping, and Aggression in the bottom as they are attributes of a defensive player and it doesn’t make sense to value a attacking player based on their defensive skill.

Insights and Observations

High Rated Players and Club Strategy:

The findings about Atlético de Madrid having the highest number of players with overall ratings above 80 could mean that the club has a strategy of acquiring high-rated players, potentially to boost its competitive edge.

Wage Expenditure Analysis:

Real Madrid's higher wage expenditure, despite having fewer high-rated players, could reflect a strategy of investing in "marquee" players, potentially to enhance the club's brand value and commercial revenue.

Market Value of Forwards:

The insight that 'Reactions' is the most significant attribute influencing the market value of forwards is valuable. It may suggest that the ability to quickly respond to game situations is highly valued for offensive players, potentially more than goal-scoring attributes like Finishing.

Defensive Attributes and Market Value:

The finding that defensive attributes like Interceptions and Sliding Tackle are less influential in determining market value implies a potential undervaluation of defensive players, which could be an opportunity for clubs to acquire quality defensive players at a lower cost.